Material didáctico de estadística y sistemas de información educativa en ciencias de la educación

*Versión Actualizada

Autores:
Jorge Lorenzo
Manuel Giovine

Tutorial de Estadística

Cátedra:

Estadística y Sistemas de Información Educativa Escuela de Ciencias de la Educación

Facultad de Filosofía y Humanidades Universidad Nacional de Córdoba

Jorge Lorenzo
Manuel Giovine

Material didáctico de estadística y sistemas de información educativa en ciencias de la educación: Tutorial de Estadística / Jorge Lorenzo; Manuel Giovine. – Versión Actualizada para el alumno - Córdoba, año 2021.

Libro digital, PDF

Archivo Digital: descarga
ISBN 978-987-86-9322-4

1. Estadísticas. 2. Ciencias de la Educación.
CDD 378.0028

Contenido

Contenido ... 3
Introducción ... 7
Capítulo 1 ... 12
Variables y Unidad de análisis ... 12
Sistemas de Medición ... 15
Otro uso del término escala ... 23
Conceptos y Constructos ... 25
Guía de Actividades N°1: ... 27
Actividades en PSPP N° 1: ... 30
Capítulo 2 ... 33
Medidas de tendencia central ... 33
Modo ... 33
Mediana ... 37
Media Aritmética o Promedio ... 39
Varianza y Desviación Estándar ... 41
Coeficiente de variación de Pearson ... 43
Procedimiento para el cálculo del promedio, la varianza y la desviación estándar ... 44
Guía de Actividades N°2: ... 48
Actividades en PSPP N° 2: ... 51
Capítulo 3 ... 55
Tablas de Frecuencia ... 55
Tabla de frecuencia: sistema de medición nominal ... 55
Utilidad de la tabla de frecuencia a través de un ejemplo ... 59
Tabla de frecuencia: sistema de medición ordinal ... 58
Tablas de frecuencias para variables métricas ... 62
Representación gráfica de las tablas de frecuencias ... 65
Diagrama de barras ... 65
Diagrama de sectores ... 66
Diagrama de barras agrupadas ... 66
Histograma ... 67
Polígonos de frecuencia ... 69
Gráfico de ojiva ... 71
Otros tipos de gráficas ... 71

Pirámides de población ... 72
Pictogramas ... 75
Mapas ... 75
Gráficas Polares ... 76
Box Plot o Diagrama de Cajas ... 77
Diagrama de cajas y medidas de posición: cuartiles ... 78
Diagrama de cajas – valores atípicos o extremos ... 79
Diagrama de cajas – restricción de la variabilidad ... 81
Diagrama de cajas en distribuciones sesgadas ... 82
Guía de Actividades N°3 ... 85
Actividades en PSPP N° 3: ... 91
Capítulo 4 ... 97
Técnicas de muestreo ... 97
Muestra y Población ... 97
Población ... 98
Muestra e Inferencia ... 99
Muestreo Probabilístico ... 101
Muestreo aleatorio simple ... 102
Muestreo aleatorio sistemático ... 104
Muestreo aleatorio estratificado ... 106
Muestreo por conglomerados ... 109
Muestreo No Probabilístico ... 111
Muestreo accidental ... 111
Muestreo por cuotas ... 111
Muestreo intencional (por conveniencia u orientado) ... 112
Bola de nieve (o en cascada) ... 112
Comentarios Finales ... 112
Misceláneas ... 113
Guía de Actividades N°4: ... 116
Actividades en PSPP N° 4: ... 119
Capítulo 5 ... 126
La distribución normal ... 126
La distribución normal como modelo matemático: la normal estandarizada ... 129
La distribución normal estandarizada y la proporción de casos ... 132
Áreas bajo la curva como distribución de probabilidades ... 134
Misceláneas ... 140

Distribucionesasimétricas 141
Medidas de asimetría y curtosis de una distribución 143
Guía de Actividades N° 5 147
Actividades en PSPP N° 5: 149
Capítulo 6 157
Estimación de parámetros 157
Estimación de una media poblacional 157
Estimación de una proporción poblacional 159
Comparación de proporciones 161
Comparación de proporciones muestrales 162
Estimación de parámetros: conceptos teóricos 164
Estimación puntual 164
Estimación por intervalos. 164
Guía de Actividades N° 6 165
Actividades en PSPP N° 6: 168
Capítulo 7 175
Prueba de hipótesis sobre la media paramétrica 175
La media paramétrica y muestral 176
Los datos y el modelo estadístico 177
La distribución normal como modelo estadístico 182
Comparación de distintos promedios con el modelo estadístico de la distribución normal estándar 186
Prueba de hipótesis sobre media paramétrica 187
Guía de Actividades N° 7 194
Actividades en PSPP N° 7: 196
Capítulo 8 203
Prueba de hipótesis y el modelo estadístico χ^2 203
Cálculo del estadístico χ^2 a partir de la tabla de contingencia 204
Reglas de decisión basadas en χ^2 y errores de decisión 207
El grado de asociación entre las variables 209
Coeficiente de contingencia C de Pearson 209
Coeficiente V de Cramer 210
Ejemplo de aplicación 210
Las frecuencias esperadas como eventos independientes 212
Los grados de libertad en tablas de contingencia 214

Distribución χ^2 215
Recomendaciones para el uso de la prueba χ^2 216
La prueba de la mediana 216
La mediana combinada 217
Análisis de una muestra simple: prueba χ^2 para la bondad de ajuste 219
Guía de Actividades N° 8 222
Actividades en PSPP N° 8: 224
Capítulo 9 228
Correlación de variables 228
El diagrama de dispersión 229
Coeficiente de correlación r de Pearson 233
Coeficiente de correlación r_S de Spearman 236
Coeficiente de correlación y prueba de hipótesis 237
Algunas precisiones sobre el coeficiente de correlación 241
Guía de Actividades N° 9 243
Actividades en PSPP N° 9: 247
Capítulo 10 251
¿Qué son los indicadores? 251
Indicadores en la Educación Argentina 252
Importancia de los indicadores educativos 254
Clasificación de indicadores 256
Sistema de indicadores: ¿Todo sistema que brinda información es un sistema de indicadores? 258
Criterios y características de los sistemas de indicadores 259
Sistema Nacional de Indicadores 259
Indicadores seleccionados 263
Analizar una política pública desde indicadores. Las Evaluaciones Estandarizadas 271
Guía de Actividades N°10: Indicadores, sistema de indicadores y sistema nacional de indicadores 273
Bibliografía 276
Reflexiones finales 277
Bibliografía sugerida 279

La estadística permite dar a conocer procesos tales como aquellos que conducen a la eliminación diferencial de los alumnos de diferentes orígenes, y que en su complejidad presentan una regularidad tal que, para describirla, incitan a tomar prestadas metáforas mecanicistas (Bourdieu, 2013: 14-15[1]).

Introducción

La estadística es una rama de las matemáticas y como tal, nos ofrece distintos modelos de análisis de datos cuantitativos. Puede ser vista como una herramienta para el resumen de grandes volúmenes de información y un puntal para la investigación científica en educación. El lenguaje estadístico establece puentes que conectan diversas áreas, convirtiéndose por ello en un elemento que promueve la visión interdisciplinaria de los fenómenos educativos. En otras palabras, la estadística se encuentra articulando la educación, la psicología, la antropología, la economía, la sociología, la geografía, etc. Todas estas disciplinas cuando convergen en la educación, pueden conectarse desde el lenguaje que proporciona el método científico y cuyo principal auxiliar es la estadística.

Bourdieu, Chamboredon y Passeron en *El oficio de sociólogo* proponían a la estadística como un instrumento poderoso de ruptura contra las pre-nociones del sentido común que permite poner en duda las primeras impresiones que siempre tenemos. “La influencia de las nociones comunes es tan fuerte que todas las técnicas de objetivación deben ser aplicadas para realizar efectivamente una ruptura, más a menudo anunciada que efectuada. Así los resultados de la medición estadística pueden, por lo menos, tener la virtud negativa de desconcertar las primeras impresiones.” (Bourdieu et al., 2002: 28)

Siendo la estadística derivada de las matemáticas comparte el lenguaje de los números, pero difiere en la manera en que los números cobran sentido. Buena parte de la estadística puede entenderse desde las operaciones básicas y un álgebra elemental. Lo verdaderamente creativo en la estadística, es el modo en que tales operaciones se utilizan para responder preguntas concretas y direccionar acciones en la vida real. Para aprender en esta disciplina, lo más importante es el modo en que puede encararse un problema usando una lógica y un razonamiento particular.

Una vez que se logra comprender el fundamento de un modelo estadístico, se adquiere una manera de pensar la realidad. Permite extender los límites de lo que

1 Bourdieu, P. (2013) *La nobleza de estado educación de elite y espíritu de cuerpo.* Buenos Aires: Siglo XXI.

tenemos inmediatamente próximo y enfocarse en lo general; de este modo es factible ver el comportamiento de los sistemas, establecer series de tiempo evolutivas o históricas, y lo más importante, en la investigación científica nos permite evaluar la verosimilitud de hipótesis teóricas e identificar los alcances y límites de las hipótesis empíricas.

Esta visión de conjunto está plasmada en la etimología de la palabra estadística, la cual deriva de la voz latina "Estado" (*status*). En tal sentido, se sabe que el Imperio Romano llevaba un minucioso compendio de los movimientos poblacionales y las riquezas de las regiones y sus habitantes. Anteriormente, en Grecia, Sócrates hablaba de lo necesario que es para el gobierno, el conocimiento de la población y la riqueza de las ciudades. En Egipto, la organización administrativa del Estado se basaba en observaciones sistemáticas y periódicas de la población y las producciones agrícolas de las distintas regiones, que se volcaron a tablas que han sobrevivido hasta nuestro tiempo. La estadística aparece en textos bíblicos, tal el caso de *Los Números*, donde se cuenta que Moisés ordenó el primer censo de los israelitas dispersos por el desierto. También se sabe del censo ordenado por David, por el relato en el segundo "Libro de Los Reyes". En China el rey Yao (3000 a.C.), ya contaba con un censo de población y producción agrícola.

La caída del Imperio Romano y su disolución dejó escasos documentos estadísticos y censales, y se produjo una merma notable en la recolección y análisis de información. El rey Guillermo de Inglaterra reconoció la importancia de los censos y ordenó que se censara la población en todo su dominio en el año 1000 aproximadamente. La iglesia, en el Concilio de Trento, introduce en forma obligatoria el registro de nacimientos, matrimonios y defunciones. Se reconoce nuevamente la importancia de los registros numéricos de población para las cuestiones estatales, y se debe a German Cönning las primeras ideas sobre registros y análisis de datos estadísticos, disciplina que llamó Estadística Universitaria. El trabajo de Cönning es lo que hoy forma el apartado de la estadística descriptiva. En Inglaterra John Graunt comienza a utilizar los registros estadísticos más allá de la simple descripción. Su profesión de demógrafo lo lleva a fundar las bases de la bioestadística y se lo considera el precursor de la Epidemiología. Aunque sus investigaciones estuvieron alejadas del método científico actual, le dio a la estadística un uso que amplió notablemente sus horizontes.

Por aquella época destacados matemáticos se habían interesado en el problema del azar y el cálculo de probabilidades, pero fue Jaques Bernouilli quien introduce los conceptos de certeza y probabilidad en problemas sociales. Los aportes de Abraham de Moivre y Pierre Simon Laplace al cálculo de probabilidades, le permitió a Karl Friedrich Gauss desarrollar su teoría sobre la distribución de los errores y proponer el modelo estadístico más reconocido: la curva de Gauss o distribución normal. Adolfo Quetelet aplica el modelo aportado por Gauss a estudios sociales y antropológicos y se lo considera el padre de la Antropometría. El mismo modelo de distribución normal de Gauss fue utilizado por

Francis Galton en sus estudios sobre herencia e inteligencia, y aparece con él la disciplina de la Psicometría. La estadística sigue un curso que se imbrica cada vez más con los modernos métodos de la estimación y la investigación científica, y son hombres interesados en diferentes ramas de la ciencia los que siguen dando impulso a los nuevos métodos de análisis estadísticos. Nombres reconocidos de la moderna estadística son Karl Pearson, Sir Ronald Fisher, George Udny Yule, William Sealy Gosset, Charles Spearman, entre otros.

En este breve compendio histórico puede verse que la estadística tiene siglos de historia y se ha ganado un lugar de preponderancia en muchas disciplinas sociales. En la actualidad y con los modernos métodos de cálculo apoyados en el uso de computadoras, pueden obtenerse indicadores estadísticos casi al instante. Pero estos cálculos por si mismos no tienen demasiado sentido si no hay personas preparadas para darles una interpretación racional. Un curso de estadística, por breve que sea, debe introducir no solo las nociones básicas, también debe fortalecer la capacidad analítica del estudiante. Actualmente los informes estadísticos se encuentran disponibles en publicaciones con diversos soportes y están abiertas al público general. De hecho, es práctica corriente la difusión de informes estadísticos sobre el estado de la educación por parte de los ministerios de casi todos los países del mundo. Comprender los informes ejecutivos elaborados por estas entidades requiere de una comprensión particular de los modelos que subyacen a los análisis informados en ellos. La alfabetización estadística es ya una herramienta del pensamiento profesional que deben dominar los estudiantes de humanidades, especialmente aquellos que se interesan por la educación, puesto que los sistemas internacionales de indicadores educativos son estadísticos.

La Estadística provee técnicas para el trabajo con datos que surgen en su mayor parte de la necesidad de fundamentar la toma de decisiones en condiciones donde prevalece la incertidumbre. La noción de incertidumbre estadística se relaciona con la idea vulgar de indecisión de la vida cotidiana, pero se diferencia fundamentalmente en que, bajo los modelos estadísticos, la incertidumbre es un aspecto mensurable y por tanto conocido, de la realidad.

Por ejemplo, al leer textos informativos, académicos o de divulgación, aparecen las nociones estadísticas sobre las cuales se basan los juicios valorativos. Sirvámonos de un ejemplo: la versión electrónica del diario El Cronista, publica el 25 de abril de 2016[1] un informe del Ministerio de Seguridad sobre el mapa de delitos en la República Argentina, destacándose que hubo una caída del delito entre los años 2014 a 2015, aunque las tasas se mantienen por encima de las de 2008. Refiriéndose específicamente a delitos contra las personas y la propiedad, se subraya que la tasa nacional de este tipo de delitos cayó un 3% entre los años mencionados, aunque se mantiene un 10,2% por encima del 2008.

Vemos en este escueto ejemplo la manera en que se intenta responder a la pregunta sobre la inseguridad utilizando valores porcentuales para mensurar el impacto sobre la sociedad. Las interpretaciones de estas cifras de seguro van a cambiar de un analista a otro, pero esa discusión excede este trabajo. Lo que destacamos es que la estadística construye sus datos, y en este sentido el interés del analista o investigador no puede prescindir del pensamiento teórico. En otras palabras, la estadística por sí misma no es suficiente para comprender cabalmente los problemas de los que se ocupa. Son los investigadores los que se valen de ella para enfrentar los problemas y tomar medidas.

La estadística define un lenguaje propio de la disciplina. Así, términos como variables, unidad de análisis, medición, variabilidad, estimación y otros, se conjugan en lo que constituye la semántica propia del análisis estadístico. Es necesario que los estudiantes comprendan paulatinamente que las afirmaciones de tipo estadístico, tales como chance o probabilidad, se ajustan a modelos matemáticos que los investigadores utilizan para describir la realidad, y que se apartan del sentido común dado a esos términos (Lorenzo y Giovine, 2019[2]).

Los datos que se "recogen" siempre refieren a unidades de análisis y los estadísticos calculados son formas de caracterizarlas. De este modo, las tasas de repitencia de un determinado nivel del sistema educativo, refiere a unidades de análisis identificadas como alumnos. Pero el foco de este análisis está puesto sobre una propiedad que trasciende al alumno y sirve para caracterizar un colectivo o grupo agregado de unidades individuales. Dicho en otras palabras, si un alumno no es promovido al año siguiente superior, se cuenta entre aquellos que repiten el grado. Del conteo de estos alumnos surge una cifra que es la cantidad de repitentes, que al ser ponderada por la cantidad total de alumnos se transforma en una proporción que se interpreta como tasa de repitencia. Este indicador, nos habla de la situación de un colegio, una jurisdicción, una provincia o el país.

Las técnicas estadísticas en su posibilidad de agregar y desagregar datos nos muestran el comportamiento y dinámica de variables propias de una población (estudiantes para este ejemplo). El potencial del análisis estadístico se aprecia cuando es posible combinar sus métodos con otras formas de análisis. En educación es común que se hable de un universo "micro" si el interés recae en la escuela, el grado o el alumno, luego, el universo "macro" caracteriza los colectivos agregados como los alumnos o las escuelas. La frontera entre estos términos es arbitraria y subjetiva; de modo que los investigadores siempre tienen que hacer explícitas sus técnicas de análisis y en ese instante es donde pueden darse intersecciones que operen rupturas con los velos que impone un saber cotidiano y no cuestionado, lo dado y aceptado, en fin, con los prejuicios y las prenociones.

Este tutorial ha sido redactado de manera que pueda ser usado como material introductorio a los principales temas de la estadística para ciencias de la educación y

---

2 Lorenzo, J. y Giovine, M. (2019) Probabilidad para Ciencias de la Educación. Córdoba: Ansenuza. Disponible en: https://ansenuza.unc.edu.ar/comunidades/bitstream/handle/11086.1/1341/Probabilidad%20para%20educaci%C3%B3n-Lorenzo%20Giovine.pdf?sequence=1

como soporte para el análisis de los Sistemas de Información Estadística Disponibles. Se verá aquí que el contenido matemático ha sido reducido al mínimo, aunque no se prescinde de algunas fórmulas conceptuales y ciertas rutinas de cálculo. El uso de estas fórmulas tiene por finalidad que los estudiantes comprendan los conceptos que subyacen al estadístico que se calcula, mediante la simbología matemática. Por otra parte, se ha incorporado el uso de una herramienta para realizar análisis estadístico de manera automatizada, tal el caso del software PSPP. Junto a la resolución de las actividades, se encontrará una guía para la aplicación de tecnologías de la información, que contribuye a facilitar la aplicación de los conceptos estadísticos.

Capítulo 1

Variables y Unidad de análisis

En este capítulo repasaremos conceptos introductorios tales como el de variables, unidad de análisis y sistemas de medición. El objetivo es brindar una comprensión acabada de la interrelación entre tales conceptos dado que todo lo que se conoce como análisis de datos, implica utilizar un modelo estadístico que interconecta estas ideas. Como idea rectora de este apartado conviene pensar en ese sentido sistémico, reconociendo la manera en que, al definir un concepto, tal el caso de unidad de análisis, también se definen los otros.

Las variables son la piedra angular de un análisis estadístico. Constituyen básicamente lo que el investigador resalta como importante en un estudio, y determina el modelo de análisis que podrá aplicar para encontrar patrones dentro de los datos, someter a prueba hipótesis, estimar parámetros, etc. El ajuste de estos modelos depende además de cómo se haya realizado la medición, de ello resulta que existen diferentes sistemas de medición aplicables a distintos tipos de variables. Finalmente, las variables pueden seleccionarse por los atributos directamente observables en la unidad de análisis, o pueden definirse teóricamente. De esto último, deriva la idea de rasgo latente, que identifica aquellas variables que guardan un isomorfismo entre el atributo y la medición.

En este capítulo abordaremos el tema de variables, unidad de análisis y sistemas de medición. Estos tres tópicos conforman un conjunto que se definen mutuamente y que son la base para el análisis y resumen de datos. En otras palabras, dependiendo de cuáles sean las variables de interés, podremos escoger el sistema de medición más conveniente a los objetivos de un estudio. Determinar qué vamos a medir y cómo lo mediremos, implica definir la unidad de análisis correspondiente. Como veremos aquí, Variable – Unidad de Análisis – Sistema de Medición, son tres elementos fundamentales para cualquier operación con datos. En estadística es importante saber cuál es el dato que estamos buscando para cumplir con el propósito de un estudio. El dato, que es el pilar fundamental donde se asienta el análisis de cualquier problemática, depende de lo que hemos definido previamente como variable. La variable, en términos sencillos, nos indica qué es lo que estamos midiendo y la unidad de análisis nos indica donde medirlo.

Hay que tener en cuenta que la variable importa en la medida en que luego nos indica algo acerca de la unidad de análisis, por lo que ambas no pueden existir de manera independiente. Las técnicas y modelos estadísticos son diferentes según el tipo de datos con que se cuente, es decir, la variable que se analiza. Existen diferentes clasificaciones para los datos, pero aquí intentaremos simplificarlas. En principio los datos pueden clasificarse en dos grandes categorías que son:

a) datos cuantitativos, que consisten en números que resultan de conteos o mediciones,

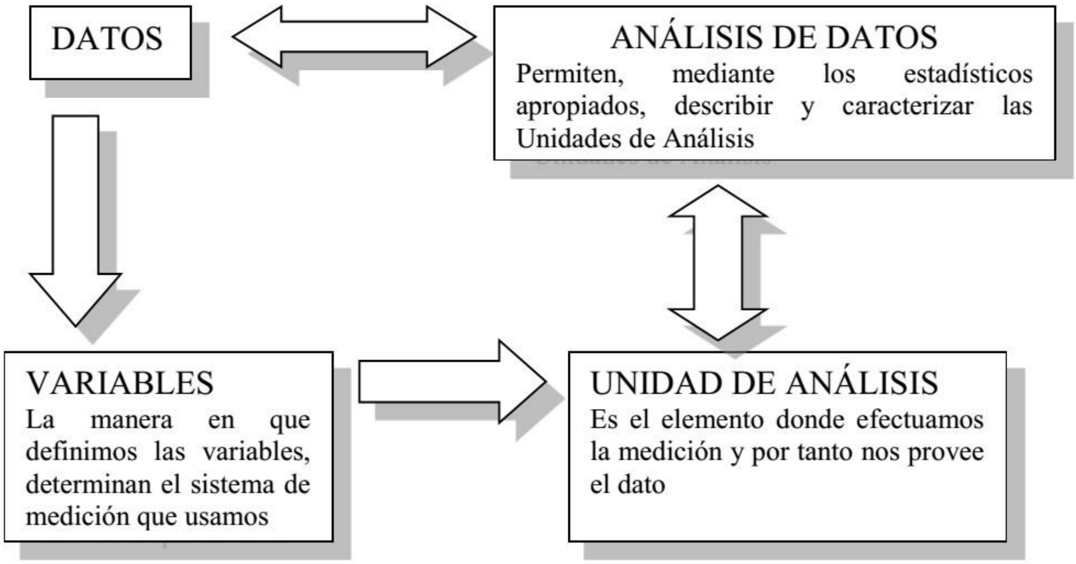
b) datos cualitativos que consisten en atributos o categorías a las que pueden asignársele valor numérico, pero no puede operarse matemáticamente con ellas.

En este punto conviene discriminar los términos cuantitativo y cualitativo cuando son aplicados a los datos, de las veces que se usan para referirse a la metodología de la investigación. Veamos la diferencia utilizando un ejemplo ficticio. Supongamos que estamos relevando información de la cantidad de profesores con título de Doctor en diferentes disciplinas; en este estudio, también recogemos datos de la edad a la que se doctoró la persona. La edad es un dato cuantitativo, y como tal es tabulado numéricamente. La disciplina en la que recibió el título de Doctor, es un dato cualitativo y será tabulado mediante una etiqueta, por ejemplo, Doctor en Filosofía, Doctor en Letras, Doctor en Ciencias de la Educación, etc. Promediando las edades y ordenándolas por disciplinas, podríamos obtener la siguiente tabla:

Disciplina en la que se doctoró y edad promedio de obtención del título máximo

Disciplina	**Edad Promedio (en años)**
Filosofía	34
Letras	36
Cs. Educación	37
Historia	41
Cs. Sociales	42
Piscología	44

Nota. En la tabla se observa que los profesores que obtienen su grado de doctor a menor edad, pertenecen en su mayoría a la disciplina Filosofía, mientras que los que se gradúan más tardíamente con el título máximo, tienden a estar agrupados en la disciplina Psicología. La tabla también separa el dato cualitativo en la primera columna, y el dato cuantitativo en la segunda. Para comprender mejor la interrelación entre datos, variables y unidad de análisis, presentamos el esquema que sigue a continuación:



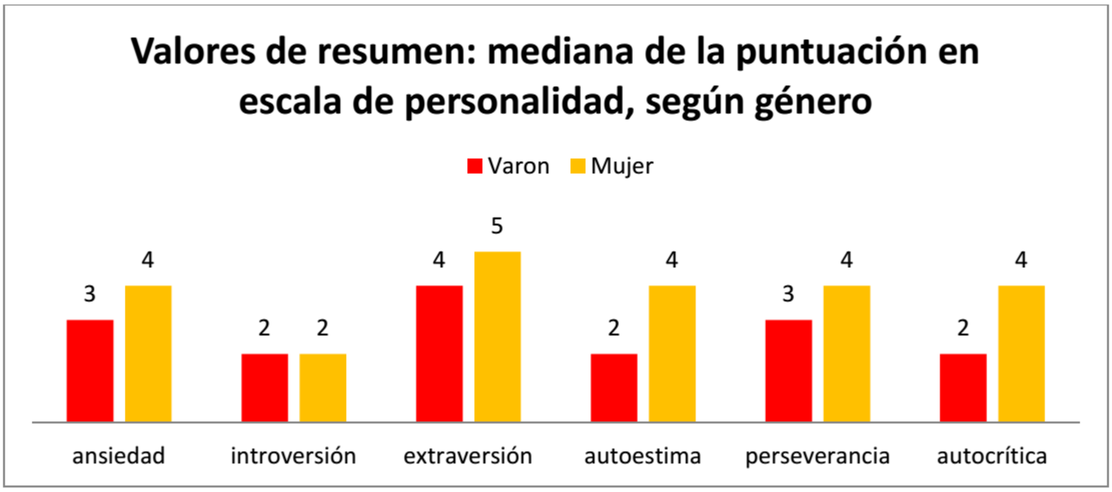
Veamos ahora con unos sencillos ejemplos la forma en que variable y unidad de análisis quedan identificadas en el contexto de un problema: Un investigador se encuentra realizando un estudio sobre personalidad en estudiantes de secundario de una escuela ficticia, le pongamos por nombre "Manuel Belgrano". Para ello, selecciona una muestra accidental de 100 estudiantes (50 mujeres y 50 varones) de esa escuela, a quienes aplica un cuestionario de personalidad, cuyos resultados se expresan de manera cuantitativa con puntajes de 1 a 5 (1=poco; 5=mucho) en los siguientes factores: a) ansiedad, b) introversión, c) extraversión, d) autoestima, e) perseverancia, f) autocrítica. El estudio pretende determinar si existe un perfil de personalidad diferente según el género de los estudiantes.

Primero identificamos la variable que está en estudio, en este caso se trata de un perfil de personalidad definida por los factores descriptos desde a) hasta f). Esta variable se mide con un instrumento que es un cuestionario, el cual se aplica a una muestra de estudiantes de la escuela, diferenciados según el género. Por lo tanto, tenemos que:

Unidad de Análisis: estudiantes de la escuela Manuel Belgrano.

Variables: a) Género del estudiante, b) Factores del cuestionario de personalidad.

En este ejemplo ficticio, podríamos obtener el perfil de personalidad diferenciado por género, tomando como estadístico de resumen la mediana de las puntuaciones en el cuestionario de personalidad, lo cual podría resumirse en el siguiente gráfico.



Como se aprecia en el gráfico, las mujeres tienden a mostrarse más ansiosas que los varones. En cuanto a la introversión, no hay diferencias por género y la tendencia muestra que no se perciben introvertidos. En extroversión se ve que ambos sexos se perciben muy extrovertidos, aventajando las mujeres a los varones. En autoestima y autocrítica, las mujeres aventajan a los varones, percibiéndose con autoestima alta y mucha autocrítica. En perseverancia, ambos sexos se ven perseverantes, con una ventaja de las mujeres sobre los varones.

En este punto no llevaremos más lejos las interpretaciones dado que todavía no hemos desarrollado las medidas de tendencia central y las técnicas de muestreo, además no se trata de una investigación real. No obstante, el ejemplo pretende hacer notar que una vez que hemos definido la variable, su escala de medición y hemos obtenido los datos, éstos se resumen con un estadístico, y la información se presenta en un gráfico desde el cuál hacemos su lectura. Es importante notar desde ya, que la conclusión aporta información sobre las unidades de análisis, y esa información está resumida mediante estadísticos.

Sistemas de Medición

Siempre que definimos una variable de interés para un estudio, es porque vamos a efectuar una medición sobre ella. La definición de medición que vamos a dar aquí es de tipo general, de manera que nos permita entender que existen diferentes sistemas de medición cada uno con características particulares. Entonces, efectuamos una medición cada vez que asignamos un valor a una variable siguiendo una regla. Debe notarse aquí que la palabra valor, no alude necesariamente a un valor numérico y que es el propio investigador quien puede determinar cuál es la regla de asignación. Una vez establecida la regla, no puede ser cambiada en el curso de la recogida de datos, por lo cual debe asegurarse que la regla pueda contener a todas las unidades de la muestra o población. Es decir, el modo en que asignamos valor a la variable en estudio debe abarcar a todas las unidades de análisis. Habiendo definido lo que se entenderá de aquí en más por medir, corresponde ahora referirnos a los sistemas de medición. Una aclaración importante: en el contexto de esta obra, sistema de medición y escala de medición pueden tomarse como sinónimos sin que ello represente mayores inconvenientes, pero desde el punto de vista estadístico existen diferencias entre estos términos. Básicamente, un sistema de medición determina cómo se construye la escala. De esta manera, todo sistema de medición nominal, dará como resultado escalas con las mismas características. Lo mismo aplica a para los otros sistemas de medición que se describen a continuación.

Sistema Nominal: el sistema de medición nominal es el más simple y extendido de los sistemas de medición, consiste en crear categorías para clasificar a las unidades de análisis. Este sistema puede admitir valores numéricos, pero solo a título de etiquetas. El ejemplo más sencillo para captar el nivel de medición nominal, son las categorías a) SI, b) NO y c) NS/NC. Por ejemplo, se podría hacer un sondeo de opinión sobre un determinado curso que se ha dictado a los estudiantes con la siguiente pregunta: *¿Cree Ud. que el curso de prevención contra enfermedades contagiosas ha sido de utilidad?* Podríamos usar las categorías de respuestas para clasificar a los encuestados, o suplantarlas por números de la siguiente manera: 1= SI; 2= NO; 3= NS/NC. Nótese que los números solo designan el tipo de respuesta, y no es lícito suponer que la respuesta SI es menor que la respuesta NO, por tener ésta última el valor 2.

Suponiendo que existan en la oferta tres métodos de enseñanza de gramática con distintas estrategias pedagógicas y los mismos contenidos, se podría crear una escala nominal con las siguientes categorías: método 1, método 2 y método 3. En este caso la pregunta en el sondeo de opinión sería *¿Qué método le parece más efectivo?* Nótese que, en este caso, el docente que responde podría reemplazar la etiqueta método 1, simplemente por el valor numérico que lo identifica, es decir, 1; lo mismo resulta válido para los otros métodos.

Si la elección de la persona encuestada no se inclina por una de las categorías presentes en la escala, su respuesta puede clasificarse en una categoría particular, por caso: Otros. Por ejemplo, si estamos relvando el tipo de deportes que practican asiduamente los estudiantes, la cantidad de categorías que debería incluir la escala se hace muy extensa. En tal caso, convendría enumerar los deportes más comunes y agregar la categoría Otros y No Practica deportes, para que todos los encuestados tengan opción de respuesta. Esto último es una regla de asignación que deberá ser respetada cada vez que se clasifique la respuesta del encuestado.

Muchas características de las personas suponen un sistema de clasificación nominal, por ejemplo: a) lugar de procedencia, b) profesión, c) género, d) situación laboral, etc. Sugerimos al estudiante consultar el "Cuadernillo del estudiante" del operativo APRENDER, donde se presentan muchas variables de este tipo (el operativo APRENDER es el dispositivo nacional de evaluación de los aprendizajes de los estudiantes y de sistematización de información acerca de algunas condiciones en las que ellos se desarrollan y puede consultarse en el siguiente enlace: http://aprenderdatos.educacion.gob.ar/aprender/pdfs/2018/Cuestionario%20al%20Estudiante%202018.pdf).

Un sistema de clasificación nominal exige casi siempre que las categorías sean exhaustivas y mutuamente excluyentes. Es decir, cada vez que usamos este sistema se espera que la persona quede registrada solo en un valor de la escala; por ejemplo, no se espera que un individuo en la variable *Último nivel educativo que cursa o cursó*, sea clasificado al mismo tiempo como Secundario y Universitario. De repetirse esto, generaría un doble conteo de unidades de análisis y la base de datos contendría muchas más respuestas que individuos[3].

Para sintetizar diremos que, si tomamos a los individuos como unidad de análisis, las categorías de las variables nominales son mutuamente excluyentes; eso significa que a un individuo debe atribuírsele sí o sí una categoría y sólo una. Un ejemplo es cuando registramos la condición final de un estudiante de la materia, este puede ser libre, vocacional, regular o promocional, pero no dos al mismo tiempo.

Sistema Ordinal: este sistema es idéntico al anterior salvo en que los valores preservan una relación de orden, aunque su naturaleza no es métrica. Dicho en otros términos, en un nivel de medición ordinal, los datos se agrupan de acuerdo a un orden expresado con los operadores "más que", "mayor que" (o su contrario implícito "menos que", "menor que"). Los sistemas ordinales son muy utilizados en los casos en que deben establecerse juicios de orden en la variable, pero no es posible establecer iguales distancias entre las valoraciones.

---

3 Sin embargo, se puede encontrar situaciones en las que es necesario clasificar las respuestas como categorías. En este caso, se dice que el sistema clasificatorio es múltiple, y lo que se toma como unidad de análisis son las respuestas dadas y no a los individuos. Un ejemplo de ello es cuando se pregunta en sondeos de opinión cuál o cuáles son las preferencias por objetos, lugares o consumos. Siguiendo este razonamiento, supongamos que preguntamos a una muestra aleatoria de 200 estudiantes de la Facultad de Filosofía y Humanidades, cuáles son las actividades recreativas preferidas los fines de semana. Nótese que en principio no hay categorías previas de clasificación, sino que estas quedarán establecidas a partir de las respuestas recogidas. De este modo, un individuo puede optar por el deporte y la vida al aire libre, mientras que otro puede optar por el deporte, el cine y los recitales, otro incluso podría optar por un paseo de compras, etc. Si bien este sistema de clasificación es nominal, se diferencia de una escala nominal en que: a) se clasifican las respuestas y no los individuos, b) pueden no tener un sistema clasificatorio establecido de antemano, c) los individuos pueden dar una o más respuestas, e) es exhaustivo, pero las respuestas no son mutuamente excluyentes. Cuando la unidad de análisis son las respuestas de los individuos no se satisface necesariamente esta condición.

El nivel de medición ordinal también puede usarse para asignar rangos a las categorías, práctica muy común en las ciencias sociales. Los rangos ordenados pueden extender el número de categorías y aproximarse a una escala métrica. Veamos un ejemplo: se intenta conocer la opinión de los profesores de una escuela sobre la calidad de la gestión del cuerpo directivo. Para ello se realiza una encuesta anónima pidiendo la valoración de dicha gestión utilizando el rango 1 a 10, donde 1 indica una mala gestión y 10 una excelente gestión. La afirmación que deberán puntuar los profesores sería la siguiente: *"Valore de 1 a 10 en qué medida considera provechosa para la escuela la actual gestión del cuerpo directivo"*.

Cabe destacar que muchas encuestas usan este tipo de afirmaciones como disparadores para valorar las respuestas de las personas, dichas afirmaciones se denominan "reactivos", que no es otra cosa que la oración a la que deberá otorgar un puntaje el encuestado. Técnicamente, cada pregunta, afirmación o ítem de un test o cuestionario, recibe el nombre de reactivo en tanto es el estímulo al cual debe responder el individuo.

Continuemos con el ejemplo y supongamos que respondieron 100 docentes. Es necesario ahora reordenar los datos para describir el resultado obtenido. Esto puede hacerse mediante una tabla de frecuencia que indique que porcentaje de docentes han respondido en cada grupo construido. El resultado quedaría expresado de la siguiente manera.

Distribución porcentual de la valoración de la gestión escolar

1 – 2	7%
3 – 4	12%
5 – 6	9%
7 – 8	44%
9 – 10	28%

Como se aprecia, los valores de la tabla han sido agrupado de a dos para reducir su tamaño, además se puede observar que la mayoría de los docentes que respondieron la pregunta asignan a la gestión de la escuela una alta valoración. Por ahora no avanzaremos más en la interpretación, puesto que tablas de frecuencia es tema que abordaremos en el capítulo siguiente.

La asignación de rangos puede aplicarse a casos en que el atributo medido pueda oscilar en dos extremos opuestos. Imaginemos la siguiente situación: se propone una escala de 1 a 5 para valorar el estado de satisfacción de un alumno con una asignatura. El reactivo usado podría ser el siguiente: *"Valore de 1 a 5 en qué medida Usted se siente satisfecho con la asignatura Estadística"*. La única conclusión que podríamos sacar de aquellas personas que asignan en el valor 1, es que no están satisfechas, pero no podríamos saber nada de su grado de insatisfacción, pues esa información no es captada ni por la escala, ni por el reactivo. Para poder llegar a esa información, será necesario modificar ambos. Primero agregaríamos una categoría neutra, el cero, y luego valores positivos y negativos. La ponderación quedará expresada en una escala como la que se muestra a continuación:

-5 -4 -3 -2 -1 0 +1 +2 +3 +4 +5

Ahora deberíamos modificar el reactivo de la siguiente manera: *"Valore su grado de satisfacción en relación con la asignatura Estadística, siendo 0=neutro; -5 muy insatisfecho; +5 muy satisfecho"*. Los rangos ordenados nos proporcionan ahora mucha más información, y sabremos que aquellos individuos que se categorizan en el extremo positivo de la escala son quienes experimentan satisfacción, al contrario, los que se categorizan en el extremo negativo manifiestan insatisfacción. El cero es el punto de referencia de la escala de rangos ordenados.

Algo que no debemos olvidar en este tipo de escalas, es que las distancias entre los valores no tienen intervalos iguales, por lo tanto, una persona que puntúa 4 en la escala anterior, no está doblemente satisfecho en comparación con otra persona que puntúa 2. Simplemente diremos que el primer individuo está "más satisfecho" que el segundo.

En ciencias sociales un aporte fundamental a este tipo de escalas de medición fue realizado por Rensis Likert, quien fue educador y psicólogo organizacional. Como parte de su trabajo en este último campo desarrolló la metodología de medición que se conoce como escala Likert. Este tipo de escala posee los atributos que hemos mencionado anteriormente, y resultan fundamentales para valorar objetivamente cuestiones que tienen un carácter subjetivo. Rensis Likert desarrolló escalas basándose en la asignación de valores numéricos para medir las motivaciones, los estilos de liderazgo, las preferencias, los acuerdos, el agrado o desagrado con ciertas afirmaciones, las inclinaciones por un método, modelo o candidato, las intenciones, etc. Actualmente, se reconoce su aporte cuando se crea una escala ordinal con asignación numérica a la valoración, dado que se la llama escala tipo Likert.

Hasta ahora hemos descripto escalas ordinales numéricamente valoradas, pero esta propiedad no es una condición necesaria. Esto quiere decir que podemos utilizar etiquetas de valor para realizar las escalas. Las etiquetas indicarán en cada caso las diferencias entre las categorías. Veamos un ejemplo: se consulta a una muestra aleatoria de 200 docentes de distintas Facultades de la UNC, sobre la implementación del examen de ingreso obligatorio. La valoración se realiza para la siguiente afirmación: *"en qué medida acuerda usted con implementar un examen de ingreso a la Universidad Nacional"*. Los resultados de esta hipotética encuesta se resumen en la siguiente tabla:

Nivel de acuerdo de docentes universitarios con el examen de ingreso

Acuerdo completamente	11%
Acuerdo moderadamente	4%
Desacuerdo moderadamente	38%
Desacuerdo completamente	47%

Las cuatro etiquetas de valor pueden ser reemplazadas por números, pero para este ejemplo vemos que conviene más conservarlas, dado que describen mejor el resultado obtenido. Por otro lado, aunque no se utilicen números se observa que las categorías están ordenadas en la medida en que establecen un continuo entre aquellos que están completamente de acuerdo con el examen de ingreso, y quienes desacuerdan completamente. La interpretación de este resultado indicaría que la mayoría de los docentes se manifiesta en desacuerdo con el examen de ingreso a la Universidad Nacional.

Existe una regla empírica a la hora de crear una escala ordinal para una investigación, y es que si la escala puede reducirse a unas pocas categorías (no más de cinco), conviene utilizar etiquetas de valor. Si la escala debe captar más información, y por tanto se necesitan más categorías, conviene valorarla numéricamente. Esta regla deriva del hecho de que es más fácil utilizar descriptores verbales cuando las categorías son pocas, pero si se usan muchas categorías la situación que se complica notablemente.

Sistema Métrico: en este sistema además de establecerse una relación de orden, se verifica que las diferencias entre las magnitudes se corresponden puntualmente con las diferencias de los objetos de medición. Esto permite establecer diferencias numéricamente ponderadas entre objetos; un ejemplo al que estamos acostumbrados y que ilustra lo antes dicho, son las diferencias en kilogramos en el peso corporal de los estudiantes.

Dentro del sistema métrico los valores que asumen las variables son cuantitativos. Existen dos grandes grupos de variables con valores cuantitativos, las **variables cuantitativas discretas** que se caracterizan por no presentar valores intermedios entre dos valores consecutivos de la variable. Por ejemplo, si la cantidad de estudiantes de una sección es de 30, no habrá ningún arreglo de ese conjunto que arroje como valor de la escala 29.5 estudiantes. Otra manera de expresar lo anterior es diciendo que las variables cuantitativas discretas no admiten valores decimales.

El segundo grupo se caracteriza por las **variables cuantitativas continuas,** donde existen infinitos valores intermedios entre dos valores consecutivos de la variable, como por ejemplo la altura del alumnado. De acuerdo con los instrumentos de medición que contemos podremos determinar la altura de dos alumnos con el valor de 1.75 metros, pero si tuviéramos un instrumento con mayor precisión, podríamos llegar a valores tan precisos que permitan diferenciarlos, como por ejemplo que el primero tenga 1.750 metros y el segundo 1.759 metros. Todos los patrones de medida son sistemas métricos, y solo en este caso es posible operar matemáticamente con las magnitudes obtenidas. Dentro de los sistemas métricos pueden distinguirse dos tipos de escalas:

Escalas Intervalares: en un nivel de medición intervalar los datos se ordenan en intervalos iguales; aunque se puede operar matemáticamente con los valores de la escala, no es posible establecer razones dado que no se cuenta con un cero absoluto. El mejor ejemplo para ilustrar este nivel de medición son las escalas de temperatura. En la escala Celsius, el valor cero está dado por el punto de congelación del agua y el valor 100 por su punto de ebullición (de allí que se denomine escala centígrada). Cada intervalo entre cero y cien es idéntico. Se comprende entonces que decir cero grados no implica ausencia de temperatura, es simplemente el referente empírico en donde se sitúa el valor cero. El sistema centígrado no es el único sistema para medir temperatura y, de hecho, en el sistema internacional la temperatura se mide en grados Kelvin.

El sistema centígrado y el sistema Kelvin emplean intervalos iguales cada uno, pero el tamaño del intervalo no es el mismo en las dos escalas y el punto cero tampoco. Por lo tanto, para establecer una relación entre ambas escalas, es necesario emplear una ecuación que represente la transformación entre escalas. En otras palabras, estos dos sistemas son intercambiables mediante la siguiente ecuación:

$$\boxed{\begin{array}{c} C° = K° - 273.15 \\ K° = C° + 273.15 \end{array}}$$

Supongamos que estamos interesados en encontrar el equivalente en grados Kelvin de una temperatura de 21° Celsius, para ello reemplazamos los valores en la ecuación y tenemos que: K°= 21°+273.15°=294,15°. Inversamente 294,15°-273.15°=21° Celsius. La transformación entre escalas de temperatura es posible porque los intervalos en cada una de ellas son iguales.

Como ya mencionamos, lo que diferencia una escala intervalar de una proporcional es que el punto cero es arbitrario. Dado este hecho es incorrecto hablar de razón en estas escalas, lo cual puede mostrarse fácilmente con el siguiente ejemplo. Imaginemos que tenemos dos escalas graduadas del 0 al 9 para medir algo, y que por comodidad desplazamos el cero una unidad en una de ellas, tendríamos dos escalas casi idénticas, salvo por la posición del cero, situación que se expone en el siguiente esquema:

A)	0 1 2 3 4 5 6 7 8 9
B)	0 1 2 3 4 5 6 7 8 9

En la escala A entenderíamos que 4 es el doble de 2 y que 8 es el doble de 4. Al ser escalas de intervalos iguales, podríamos proyectar la relación de la escala A en la escala B, y se produciría una inconsistencia: el valor 4 de la escala A, coincide con el valor 3 de la escala B, pero no resulta verdadero que 3 sea el doble de 1. Lo mismo ocurre con el valor 8 de la escala A, que al ser proyectado sobre la escala B, coincide con el valor 7, de lo cual resulta también falso que 7 sea el doble de 3. Es decir, al ser el origen arbitrario, no es posible equiparar sin más ambas escalas, aun cuando posean intervalos iguales.

¿Es un problema no contar con un punto de origen fijo? En estadística esto no es realmente un problema, pues las escalas de intervalo permiten operar matemáticamente y por tanto es posible calcular todos los estadísticos. Además, muchas escalas de medición de uso común, tienen la propiedad de intervalos iguales y cero arbitrario, por ejemplo, la altura desde el nivel del mar, la distancia de una localidad respecto de un punto de referencia, el índice de masa corporal, etc. Por ello, los paquetes informáticos actuales no hacen distinción entre las escalas de intervalo y de razón, y simplemente se refieren a ellas como escalas métricas.

Escalas Proporcionales o de Razón: este nivel de medición es similar al nivel anterior, pero con la propiedad de que el cero tiene valor absoluto. Es decir, allí donde se registre cero, será igual a ausencia de atributo en la variable. Durante mucho tiempo se ha trabajado para que todas las unidades del sistema internacional estén expresadas en escala de razón. Aunque no se ha logrado completamente, se avanzó mucho en este sentido; así la longitud se mide en metros, la masa en kilogramos, el tiempo en segundos, etc. De este modo, un elemento que haya recorrido 0 metro, indica que no se ha movido de lugar, un elemento con 0 kg de masa, no posee masa, y el instante 0 segundo, es el momento inicial.

En dos escalas de razón existe la misma relación entre los intervalos y además existe la misma relación entre dos puntos de la escala. Por esta propiedad, las transformaciones entre escalas de razón son monotónicas y quedan expresadas por una función, que en notación matemática es y= *f*(x). Esto quiere decir que cualquier valor de **y** puede calcularse a partir de los valores de **x**, mediante la función *f*. Si *x* e *y* son dos escalas de razón, tendremos correspondencias entre ambas para cualquier valor dado.

Como ya mencionamos, en estadística no es tan importante la distinción entre escalas de intervalo o de razón, ya que ambas serán tratadas como escalas métricas. Lo importante es recordar que solo en este tipo de escalas pueden emplearse operaciones matemáticas para los estadísticos que las requieran, como por ejemplo el promedio o la varianza. Asimismo, en ciencias sociales las escalas métricas se emplean cuando:

a) es posible medir con un patrón, el caso de la altura o el peso de una persona;

b) se pueden realizar conteos, por ejemplo, cantidad de materias aprobadas, cantidad de alumnos inscriptos al comienzo del año lectivo, y

c) cuando se emplean pruebas estandarizadas, por ejemplo, el puntaje de la evaluación del Programa Internacional para la Evaluación de Estudiantes (PISA por sus siglas en inglés: *Programme for International Student Assessment*).

Otra cuestión a destacar es que las escalas métricas pueden reducirse convenientemente a escalas ordinales, pero una escala ordinal no podrá ser tratada como métrica. Para ilustrar este punto tomemos un ejemplo sencillo. Supongamos que medimos la altura de cien alumnos de una escuela. Cada medición quedará expresada por un valor en centímetros y así habremos obtenido cien mediciones que corresponden a una escala métrica. Podría darse el caso de que las cien mediciones efectuadas fueran diferentes, aunque tales diferencias fueran mínimas; por ejemplo: el alumno A tiene una altura de 1,71 cm, el alumno B tiene una altura de 1,65 y el alumno C una altura de 1,66. Bajo estas circunstancias, podríamos reducir la escala métrica original a una escala ordinal con las siguientes categorías; a) Alumno de estatura baja: 1,60 cm o menos; b) Alumno de estatura media: 1,61 cm a 1,70 cm; y c) Alumno de estatura alta: 1,71 cm o más. La escala métrica original quedó expresada en una escala ordinal con tres categorías: Alumno de estatura Alta – Media – Baja.

Aunque ahora hay menos categorías y la escala es más simple de manipular, se ha perdido información de la escala métrica original, pues si decimos que el alumno D es de estatura Alta, mientras que el alumno E es de estatura Media, solo sabremos que D mide 1,71cm o más y que E tiene una estatura comprendida entre 1,61 y 1,70 cm. Si los cien alumnos hubieran sido clasificadas en una escala ordinal como la que creamos, sería imposible reconstruir la escala métrica original, y esto es lo que se quiere expresar cuando afirmamos que una escala ordinal no podrá ser tratada como métrica.

Otro uso del término escala

La clasificación de las escalas de medición que hemos ofrecido en los párrafos anteriores, corresponde a la propuesta de S. Stevens, quien en 1946 publicó un artículo de amplia difusión en la revista Science (N°103, pp: 677-680), titulado *On the theory of Scales Measurement* (Sobre la teoría de las escalas de medición). Desde entonces se ha mantenido por conveniencia la diferenciación presentada.

Sin embargo, en la moderna investigación social, han aparecido escalas de medición que no fueron contempladas en ese artículo. Mencionaremos solo tres que son las más conocidas:

a) Escala porcentual: el porcentaje consiste en tomar cualquier valor y reducirlo a una base cien, de allí que sea posible establecer cambios porcentuales, incrementos o decrementos. De esto se derivó que los porcentajes puedan ser tratados como escalas de medición de tipo métricas y calcularse diferentes estadísticos a partir ellos, de hecho, el promedio para una escala porcentual se denomina media geométrica;

b) Escala autograduada: esta fue una invención muy útil en la medición del dolor, cuestión que durante mucho tiempo no contó con un procedimiento metódico. La escala fue creada combinando las propiedades de la escala de intervalo y ordinal. El procedimiento consistió en darle al paciente una regla graduada en centímetros del 0 al 10; se le informaba que cero es ausencia del dolor y diez es un dolor imposible de soportar sin analgésicos. Los pacientes se familiarizaban con el tamaño de la regla, pero luego debían sostenerla de manera tal que no pudieran ver los números. Así, debían indicar cuánto dolor experimentaban, señalando con el dedo en el dorso de la regla. Sorprendentemente se encontró en los reportes que el umbral del dolor soportable sin analgesia estaba entre el 6 y el 7. La familiaridad de las personas con una regla de 10 cm., hizo que su ponderación del dolor se volviera regular. A partir de este hecho, muchas escalas han aprovechado la propiedad de mediciones conocidas por las personas para ponderar aspectos muy subjetivos como por ejemplo las preferencias estéticas;

c) Pruebas estandarizadas: el auge de las pruebas estandarizadas se dio entre el fin del siglo XIX y principios del siglo XX. La escala estandarizada más conocida es la desarrollada en 1904 por Alfred Binet para medir la inteligencia. La teoría de la medición psicométrica, encontró por aquella época muchos seguidores y su método pronto se extendió a otros campos de las ciencias humanas. Los aportes fundamentales de Chales Spearman, Edward Thordinke, Lewis Therman y Robert Woodworth, sirvieron para desarrollar numerosas escalas para medir no solo aspectos psicológicos, sino también factores sociales y antropológicos. Existen dos cuestiones fundamentales que definen a una escala estandarizada; la primera es que se cuenta con valores de referencia para una población o grupo etario. La segunda es su validez y confiabilidad. El primer aspecto se refiere a que, al utilizar una prueba estandarizada en un conjunto de individuos, resulta factible realizar una comparación con un patrón de referencia, cuyo nombre técnico es baremo. La segunda cuestión se refiere a que la prueba ha demostrado que efectivamente mide lo que se propone medir, y que lo hace de manera consistente. Más adelante discutiremos estos puntos en el apartado variables de constructo. Por el momento basta con recordar que las pruebas estandarizadas contienen en su mayoría escalas de tipo métricas.

Conceptos y Constructos

Para una revisión pormenorizada de estos términos en la investigación científica, sugerimos la lectura de la obra de Kerlinger, F. (1988). "*Investigación del Comportamiento*". Segunda Edición. México: McGraw-Hill. Siguiendo la distinción de este autor, los términos "concepto" y "constructo" tienen significados similares, aunque no idénticos. Un concepto expresa una abstracción formada por generalizaciones sustraídas de casos particulares, por ejemplo, los objetos tienen una propiedad intrínseca a la que se ha denominado "peso", sin embargo, el peso no es una propiedad natural del objeto, sino que surge de la comparación de objetos entre sí. De tal modo, en la antigüedad, era común justipreciar el peso de dos objetos diciendo: "el objeto A es más pesado que el objeto B". El concepto de peso se independiza de los objetos particulares, cuando es posible crear una escala en la que todo objeto pueda ser medido con la misma regla de asignación de valores numéricos. De este modo, el concepto peso pasa a ser una propiedad medible y con la cual es posible comparar dos objetos cualesquiera. Surgen así diferentes escalas de medición. Las convenciones internacionales escogieron como patrón, el kilogramo y el gramo para expresar el peso, pero es frecuente que algunos países no adopten dicho patrón, como por ejemplo los países anglosajones, que expresas los pesos en onzas.

Un constructo es un concepto que incorpora un sentido adicional, el de haber sido inventado o adoptado de manera deliberada y consciente para un propósito científico especial. Tomemos como ejemplo el concepto de "aprovechamiento escolar"; en primer lugar, es necesario reconocer que tal concepto no es más que una abstracción, que surge de la síntesis de la observación de ciertas conductas infantiles. Tales conductas no pertenecen a cualquier dominio, sino a uno muy puntual: el aprendizaje. Por lo tanto, el aprovechamiento es el concepto que describe de manera general comportamientos observables en tareas escolares, tales como la lectura, la resolución de problemas aritméticos, la expresión mediante dibujos, etc. Las distintas conductas observadas que pertenezcan a un mismo dominio pueden ser sintetizadas bajo un mismo concepto, en tal caso es posible expresarse acerca de los individuos mediante conceptos tales como la inteligencia, la agresividad, la conformidad, la honestidad. En la investigación científica de la conducta humana, los conceptos son tratados como constructos en dos sentidos: a) como parte de los esquemas teóricos que interrelacionan diversos constructos. Por ejemplo, el aprovechamiento escolar es en parte una función de la inteligencia y la motivación; b) por otro lado, aprovechamiento debe definirse de manera tan específica como sea posible, para que puede ser observada y medida. En sentido metodológico, el punto a) corresponde con lo que se denomina definición conceptual del constructo, mientras que el punto b) es la definición operacional del mismo. Así, es posible crear una serie de condiciones específicas en las que una conducta que pertenece al constructo sea observada y medida. Por ejemplo, la lectura en voz alta pertenece al constructo aprovechamiento escolar, luego, es posible escoger diversos textos de complejidad creciente y verificar si la niña o el niño es capaz de leerlo sin cometer errores. De este modo, el nivel de lectura queda definido por el número de errores cometido durante la lectura de un texto con un grado de dificultad definido. De esta manera es posible crear una escala de medición que pondere la capacidad de lectura de los escolares, la cual puede expresarse numéricamente. Sintetizando, las variables de constructo son aquellas que se cuantifican mediante indicadores medidos en formas de comportamiento.

Una cuestión importante a destacar es que existen ramas de las ciencias humanas dedicadas a la construcción de pruebas estandarizadas para medir conceptos que han logrado alcanzar el nivel de constructos y ser tratadas como variables. La psicometría, edumetría, econometría, antropometría, sociometría, y otras, son las disciplinas que se ocupan de obtener escalas de medición confiables y válidas para variables de constructo. Por lo dicho, de aquí en adelante cuando se haga mención al término prueba estandarizada, nos estaremos refiriendo a este tipo de pruebas, cuyos puntajes pueden ser tratados como variables métricas[4].

[4] En el siguiente artículo puede encontrarse una discusión pormenorizada sobre los constructos en investigación educative: Soler Cárdenas y Silvio Faustino. LOS CONSTRUCTOS EN LAS INVESTIGACIONES PEDAGÓGICAS: CUANTIFICACIÓN Y TRATAMIENTO ESTADÍSTICO. Atenas, vol. 3, núm. 23, 2013, pp. 84-101

Guía de Actividades N°1:

Variable, Unidad de Análisis y Sistema de medición.

Objetivos de la actividad:

a) aplicar el concepto de variable a situaciones concretas,

b) reconocer y utilizar los distintos sistemas de medición

Actividad 1: En los siguientes ejemplos señalar la/s variable/s que se desea estudiar y la unidad de análisis:

1. El Ministerio de Educación decide lanzar un plan para mejorar la calidad edilicia de las escuelas. Dentro del plan se contempla la construcción de nuevas aulas en aquellos establecimientos que tengan sobrepoblación escolar. Atendiendo a ese objetivo, el Ministerio se propone sistematizar la información sobre la cantidad de alumnos que atiende cada establecimiento.
2. El Ministerio de Educación de la Provincia de Córdoba se encuentra realizando un estudio sobre abandono escolar, en el cual se pretende calcular un índice de riesgo de abandono escolar. Dicho índice se presenta como una medida de probabilidad expresada en porcentaje, según la cual por debajo de 10% el riesgo es muy bajo y por encima del 70% el riesgo es muy alto. En la primera etapa del estudio se han seleccionado dos muestras aleatorias de estudiantes: a) alumnos con necesidades básicas insatisfechas[5] y b) alumnos que tuvieron NBI y ya no las tienen.
3. En un estudio educativo se analizó si el rendimiento en la asignatura Lenguaje Oral y Escrito, era diferente según el turno de cursado, a lo largo de todo el nivel inicial.
4. En un estudio educativo se aplicó un cuestionario de evaluación de rendimiento a una muestra accidental de estudiantes de la Licenciatura en Ciencias de la Educación. El objetivo de la investigación era determinar si el nivel educativo de los padres incidía en el promedio de las calificaciones de sus hijos o no.
5. En una investigación se desea saber si existe asociación entre el género (<varón>,<mujer>, <otro>), el compromiso religioso (<bajo>, <alto>) y la libertad de expresión en términos de sexualidad (<liberal> <conservador>), en estudiantes ingresantes a la Facultad de Filosofía y Humanidades. Para ello se toma una muestra aleatoria sistemática de estudiantes de las diferentes carreras de esa

5 Se considera Población con Necesidades Básicas Insatisfechas (NBI) a la que se reúne alguna de las siguientes condiciones: más de tres personas viviendo en una misma habitación; alojamiento en viviendas precarias o de inquilinato; falta en la vivienda de retrete con descarga de agua; que en la familia exista algún niño entre 6 a 12 años que no vaya a la escuela.

Facultad y se aplica un cuestionario a tales efectos.

Actividad 2: En la siguiente tabla se presenta una serie de variables. Mencione el tipo de escala (Nominal, Ordinal, Métrica), que a su juicio mejor se ajusta para la medición. Justifique su respuesta.

Variable	Escala de medición
Número de hijos por familia.	
Tiempo (meses) en situación de desempleo.	
Cantidad de veces en situación de desempleo.	
Nivel socioeconómico: medio, alto y bajo.	
Nivel de estudios: primarios, secundarios, terciarios.	
Horas semanales dedicadas al estudio.	
Edad cronológica en años.	
Calificaciones escolares: bueno, muy bueno, excelente, etc.	
Calificaciones escolares con valores de 0 – 10.	
Clase social definida como ingresos en $.	
Distancia recorrida en metros.	
Puntajes en test estandarizados de CI, con valores de 0 – 100.	
Especialidad elegida por los alumnos en el ciclo polimodal.	
Nivel de satisfacción con el producto: alta, media, baja.	
Carrera que cursa en FFyH: filosofía, letras, etc.	
Situación laboral: activo, retirado, desocupado, etc.	
Motivación para la tarea: mucha, poca, ninguna.	
Escala de motivación con valores de 0 – 100.	
Provincia de procedencia.	
Peso en kilogramos.	
Altura en centímetros.	
Tipo de tarjeta de crédito: visa, naranja, mastercard, etc.	
Cantidad de tarjetas: 1, 2, + de 2.	
Ingreso familiar: hasta $500, de $500 a $1000, + de $1000	
Escala de satisfacción laboral con rango 0 – 10.	
Tiempos de reacción a un estímulo visual medido en segundos.	
Grupo religioso al que pertenece: católico, judío, protestante, otro.	
Cantidad de errores en una prueba de dictado.	
Puntaje obtenido por los equipos de fútbol al final del torneo.	

Cantidad de profesores con dedicación exclusiva en la U.N.C.	
Variable	**Escala de medición**
Género de los beneficiarios de un programa de asistencia	
Respuestas tabuladas como: a) de acuerdo, b) en desacuerdo, c) NS/NC	

Actividad 3: Utilice un criterio para clasificar las siguientes variables métricas como continuas o discretas

Variable	Clasificación
Edad de ingreso a la universidad (solo años).	
Tiempo requerido para completar una tarea.	
Puntajes en una prueba de lectura con escala de 0 – 10.	
Peso en gramos al momento del nacimiento.	
Presión arterial en personas con más de 60 años.	
Temperatura pronosticada para el día de la fecha.	
Cantidad de alumnos repitentes en una escuela.	
Altura de los jugadores de un equipo de básquet (en cm)	

Actividad 4: En los siguientes ejemplos deberá indicar cuál es la variable o variables en estudio, su sistema de medición y sus correspondientes unidades de análisis.

a) Un investigador seleccionó una muestra de 20 estudiantes de un profesorado universitario, a quienes aplicó un cuestionario sobre los siguientes rasgos de carácter: a) ansiedad, b) introversión, c) extraversión, d) autoestima, e) perseverancia, f) autocrítica. Cada rasgo está calificado de manera cuantitativa con puntajes de 1 a 5 (1=escaso; 5=excesivo).

b) Un sociólogo ha relevado la situación laboral de una muestra accidental de hombres y mujeres que asisten a hospitales públicos y ha utilizado las siguientes categorías de clasificación: EMPLEADO – DESEMPLEADO – JUBILADO – BENEFICIARIO DE PLAN SOCIAL.

c) En una encuesta realizada a docentes del Ciclo Básico de Nivel Secundario de Córdoba Capital, se relevó información sobre los libros de texto solicitados para la asignatura Matemáticas: título, editorial, año de publicación, ciudad de publicación y cantidad de páginas.

Actividades en PSPP N° 1:

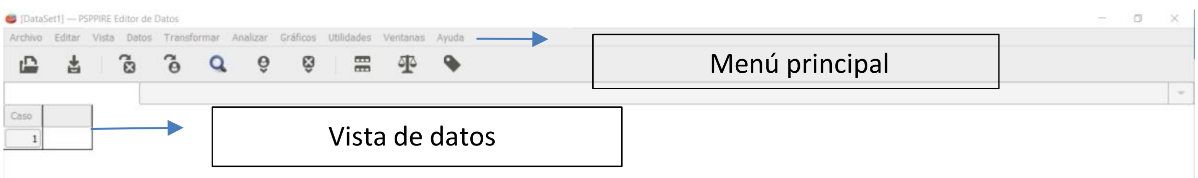
En la siguiente actividad práctica instalará el programa PSPP, un software específico para el análisis de datos. Vamos a realizar un conjunto de actividades prácticas que le permitirá ejercitar lo que ha aprendido en el capítulo. De este modo, usted estará preparado para construir bases de datos y hacer procesamientos estadísticos de datos primarios (construidos por usted) o de secundarios (que han construido otros).

Hay a disposición muchos paquetes de análisis estadístico. Para el análisis de encuestas se destacan por su popularidad el SPSS (*Statistical Package for Social Sciences*), diseñado por la Universidad de Chicago en la década de 1950, el STATA (que viene de las palabras estadística, *statistics,* y datos, *data*) creado por StataCorp en 1985 y R, que es un lenguaje de programación con especificidad en el análisis estadístico que apareció el 1993.

Todos estos programas permiten realizar análisis de datos en los que intervienen gran cantidad de variables y de casos. Partiendo de una filosofía del Software libre (GNU) se ha desarrollado el proyecto PSPP, que son las siglas del SPSS invirtiendo las S por P y viceversa. El software libre está caracterizado por la distribución gratuita y por el aporte de distintos programadores en colaboración. Debido a que el funcionamiento es similar al SPSS, en este curso vamos a utilizar la herramienta gratuita PSPP, pero usted podrá realizar las mismas operaciones, de un modo similar, en el programa SPSS.

Para la instalación del programa deberá descargarlo de algunas de las copias disponibles en internet, de acuerdo con su sistema operativo: Windows, Linux o Mac. Una de las opciones es descargarlo desde la página web del mismo proyecto PSPP, http://www.gnu.org/software/pspp/get.html, que nos dirige a los links de descarga para cada sistema operativo en sus versiones 32 y 64 bits.

Una vez descargado e instalado el programa de PSPP, procedemos a instalarlo y abrirlo. Nos vamos a encontrar con una interfaz que tiene un menú principal y debajo una pequeña ventana que muestra los casos y las variables, cómo se ve a continuación:

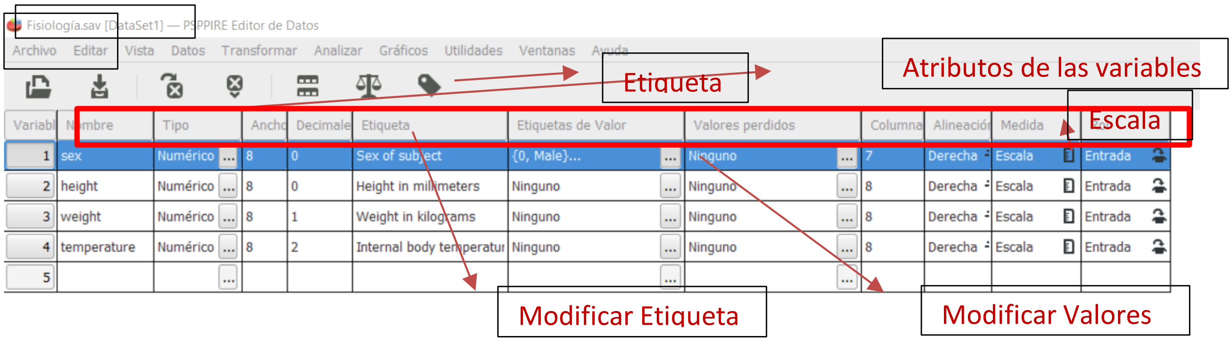


Seleccionamos el menú Archivo/File, Abrir/Open y el menú Ejemplos/Examples. De los ejemplos que trae el programa, seleccionamos el archivo que dice physiology.sav. Como puede ver los archivos ".sav" son de bases de datos. Posteriormente, va a poder ver una base de datos que registra 40 casos y 4 variables.

Describa el tipo de variable, para cada una de ellas, su nivel de medición y la escala:……………………………………………………………………………………………………
……
……
……
……
……
…………………………………………………………………………………..

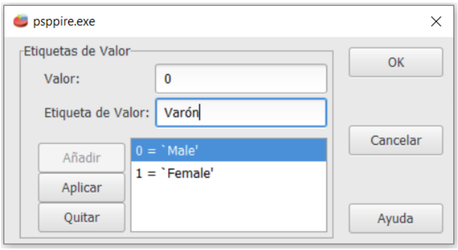
Ahora vamos a modificar la base de datos para poner los títulos en español. Lo primero será guardar la base con otro nombre, por ejemplo fisiología.sav, en la carpeta de Documentos donde puede crear una subcarpeta que se llame Estadística, o en la carpeta que usted desee. Para ello aplicamos el comando Guardar como/Save as, seleccionamos la carpeta y el programa nos permitirá cambiar el nombre de la base de datos, que llamaremos "Fisiología". Una vez realizado el cambio, veremos que en el margen superior izquierdo del programa dirá Fisiología.sav.

El paso siguiente será modificar los nombres de las variables en inglés por los correspondientes en español: Sexo, Altura, Peso y Temperatura. Para ello, en el margen inferior izquierdo verán una solapa que dice "Vista de Datos" y otra que dice "Vista de Variables". Seleccionen la opción de "Vista de Variables" y allí van a encontrar los atributos de cada una de las variables de la base como sigue:



Haciendo clic sobre los nombres de las variables podrá editarlos. Lo que puede notar también es que el tipo de variable es numérico para todas ellas. Esto se debe a que las variables están codificadas con valores, por ejemplo, para sexo 0=Male y 1=Female. Esto podrá verlo volviendo a la "Vista de Datos" y en el menú "Ver" poner "Ver etiquetas de valor", o con el ícono de etiqueta que está debajo del menú. De este modo los 0 cambiarán por Male y los 1 por Female.

Para modificar estas etiquetas al español vamos a la "Vista de Variables" y en la columna Etiquetas de Valor, seleccionamos los "…" que están en la casilla de la variable Sexo. Luego seleccionamos donde dice 0=`Male´ y arriba en Etiqueta de Valor modificamos Male por Varón, presionamos el botón Aplicar y lo mismo hacemos para Female por Mujer. Finalmente seleccionamos OK.



Una vez que hemos realizado esta tarea vamos a Vista de Datos y observamos como las etiquetas han quedado modificadas por Varón y Mujer.

Posteriormente, cambiaremos las etiquetas de las variables en la columna Etiqueta, pasándolas al español. En la columna de Medida usted pondrá si la variable es Nominal, Ordinal o Escala=Métrica. Por ahora, no vamos a modificar ninguna de las restantes columnas.

Finalmente, vamos a revisar los datos, por ejemplo, el primer paciente es varón, tiene 1799 milímetros, pesa 90.9 kilos y registra una temperatura al momento del estudio de 37.53 grados centígrados. Si revisan todos los pacientes, verán que el número 30 registra una altura de 179 milímetros, apenas más que un bolígrafo. Esto es lo que uno denomina error de carga de datos. En este caso se le puede asignar un valor estimativo en función de las otras variables o el valor promedio de la altura sin contar este caso, pero nosotros sólo agregaremos un 0, quedando 1790. Esto puede hacerse posicionándose sobre la celda y editar el dato. Ahora grabamos la base. Es conveniente grabar la base regularmente por si el programa se detiene o tilda.

Para terminar se les solicita que describan 3 datos de la base que estamos estudiando, como se hizo para el primer caso en el párrafo anterior, indicando UUAA, variable, valor, para cada una de las 4 variables:

....
....
....
....
....
....

Capítulo 2

Medidas de tendencia central

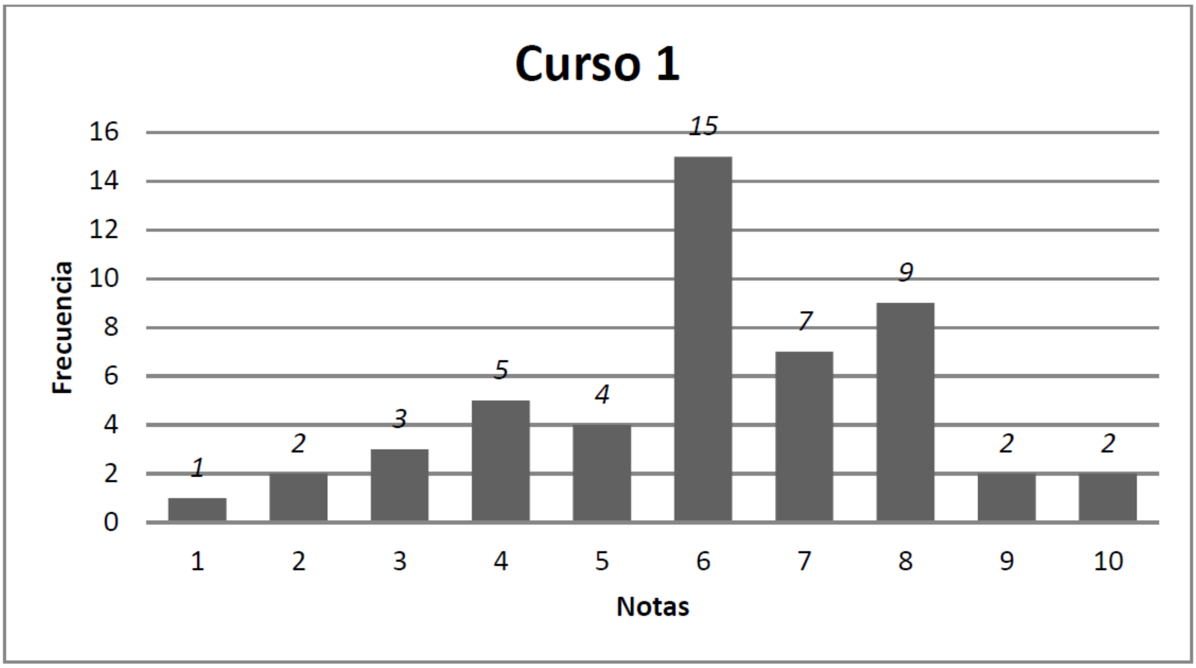
Las medidas de tendencia central son números sintéticos, esto es, valores únicos que permiten resumir la información de un conjunto más amplio de datos. Resultan fundamentales por dos razones; la primera es que unos pocos números específicos revelan el comportamiento de variables. La segunda razón es que sirven de base para el cálculo de estadísticos más complejos, especialmente los involucrados en la prueba de hipótesis en investigación.

Si se dispone de un conjunto de datos, es posible utilizar algunas medidas que permitan resumir en unos cuantos números significativos, las principales características del conjunto completo. Estos números, cuando son obtenidos de una muestra reciben el nombre de **estadísticos** y cuando se calculan en una población se denominan **parámetros.** Una cuestión importante a tener en cuenta desde ya es que los parámetros son difíciles de obtener para poblaciones muy grandes. Por lo tanto, los estadísticos obtenidos de muestras representativas de la población, sirven para estimarlos. Algunos estadísticos se basan en la centralidad del conjunto de datos, por lo cual se reciben el nombre de medidas de tendencia central. Repasaremos aquí las medidas más comúnmente utilizadas.

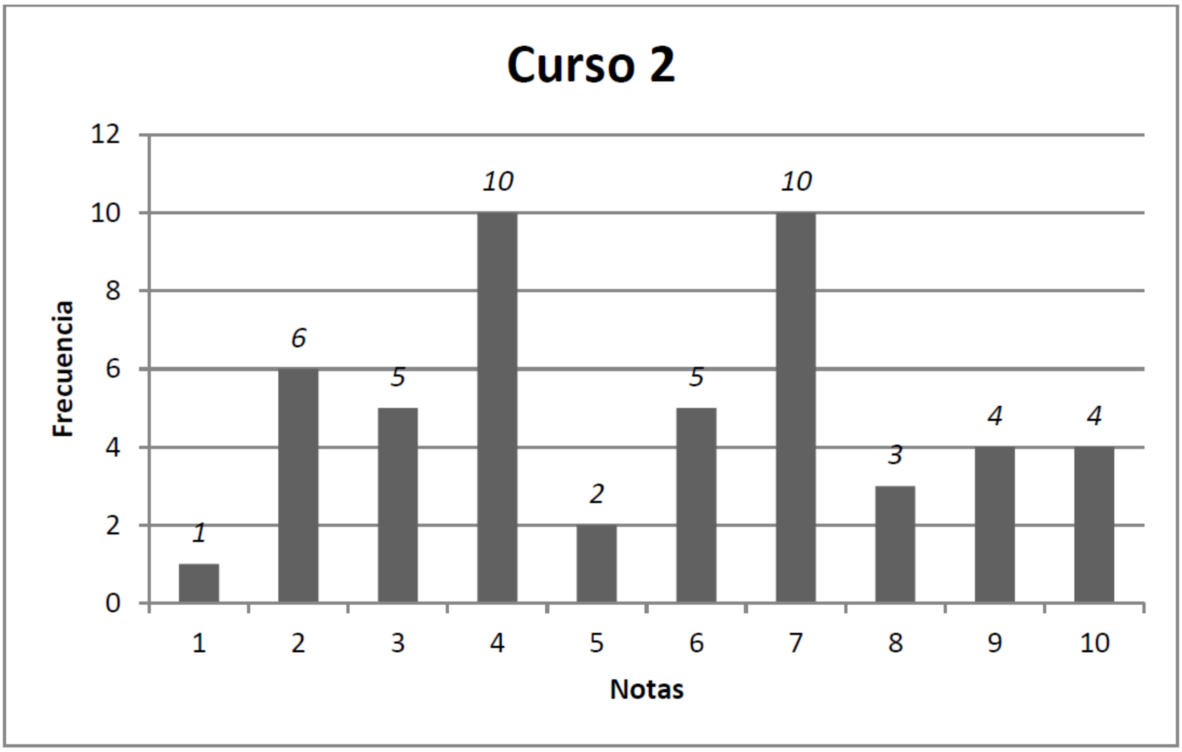
Modo

También recibe el nombre de moda, y se define como la categoría de la variable que concentra la mayor frecuencia. Si en un conjunto de datos, existe un valor que se repite más veces que el resto, se lo denomina modo; en tal caso se dice que la distribución de esos datos es unimodal. La mayoría de las distribuciones de datos son unimodales, pero en ocasiones puede que existan dos valores que se han repetido igual número de veces, siendo estos los más frecuentes. En tal caso, la distribución tiene dos modos, y por tanto es bimodal. En el caso de que existan varios valores que concentran las mayores frecuencias de la distribución de datos, se tiene una distribución multimodal. El modo es la medida más sencilla de caracterizar una distribución de datos.

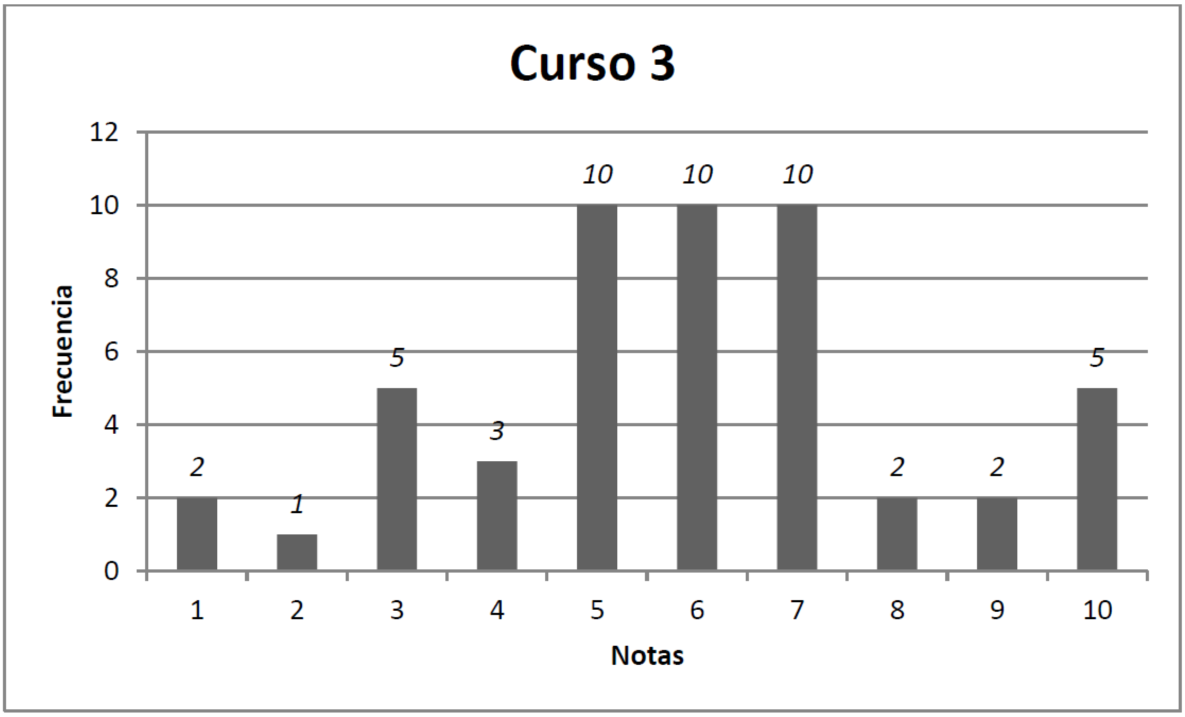
Ejemplo: un docente toma un examen de matemáticas en tres cursos de 50 alumnos. Luego realiza una gráfica con la distribución de los resultados, identificando en cada caso el valor modal.



Lo que se observa en el gráfico del resultado del examen de matemáticas del curso 1, es que la mayoría obtuvo una nota de 6 (15 de los 50 alumnos, esto es el 30% de ellos). Siendo el valor modal el que concentra la mayor frecuencia, se tiene que la nota 6 es el modo de esta distribución.



En este gráfico se muestra la distribución de frecuencias del curso 2. Observamos que los alumnos que obtuvieron como nota 4 y 7 tienen igual frecuencia, y que en conjunto concentran la mayor cantidad de casos de toda la distribución. Esto significa que el 40% de los alumnos se han sacado notas de 4 y 7 respectivamente, en igual proporción. La distribución en este caso se dice que es bimodal, siendo los modos la nota 4 y la 7.



La distribución de las frecuencias de las notas del curso 3 se diferencia de las otras en que se observa que las frecuencias más altas se agrupan en las notas 5, 6 y 7. Se trata entonces de una distribución multimodal (es decir, que contiene más de un modo); el 60% de los casos se encuentra agrupada en las notas 5, 6 y 7.

Un aspecto destacado del modo como medida de tendencia central es que puede aplicarse a variables que no han sido medidas en escala métrica, dado que se basa en identificar la categoría de la variable que concentra la mayor frecuencia.

Veamos un ejemplo: En una librería se realiza un conteo de los libros vendidos en el último mes, discriminando su género literario. El resultado se expresa en la siguiente tabla de frecuencias:

Género de la publicación	**Cantidad de ventas**
Novela	198
Ensayo	212
Manual	128

Género de la publicación	Cantidad de ventas
Técnico	406
Autoayuda	449
Infantil	228
Audiolibro	41

Como puede apreciarse en esta tabla, el género literario más vendido es el de autoayuda, cuya frecuencia es de 449. De esto se desprende que la categoría modal de venta es autoayuda.

Una cuestión que debe tenerse en cuenta cuando se identifica el modo en una distribución de frecuencias es que el modo no es la frecuencia más alta, sino la categoría donde ésta se concentra. En este sentido, y volviendo al ejemplo de la librería, el modo es autoayuda y no 449 que es su frecuencia.

Habíamos dicho que el modo es una medida de tendencia central muy fácil de calcular, pero resulta por lo mismo, muy inestable. Por ello conviene tener en cuenta a la hora de interpretar su sentido, lo que se denomina variación en torno al modo, que representa la manera en que se distribuyen las frecuencias próximas al valor modal. Para entender este concepto veamos el siguiente ejemplo: existen tres manuales especializados para la enseñanza de la matemática en el nivel primario. Se realiza una encuesta a una muestra representativa de docentes para saber cuál resulta de su preferencia a la hora de planificar sus actividades áulicas. Por razones de síntesis diremos que los manuales son A, B y C.

Podríamos encontrarnos con tres situaciones diferentes, las cuales se muestran en las siguientes tablas:

Primera situación posible

Tipo de manual	Preferido por los docentes en %
Manual A	20%
Manual B	10%
Manual C	70%

En esta primera situación, claramente el valor modal es Manual C. Se deduce entonces que este es el que más se utiliza para la planificación de actividades en el área de las matemáticas en el nivel primario.

Segunda situación posible

Tipo de manual	Preferido por los docentes en %
Manual A	40%
Manual B	35%
Manual C	25%

En esta segunda situación hipotética nos encontramos que el modo es Manual A (40%). Pero nótese que el Manual B concentra el 35% de las preferencias, esto es, se separa del modo en solo cinco puntos porcentuales. Esto equivale a decir que existe poca variación en torno al modo cuando se consideran juntos los manuales A y B. La conclusión sería que el manual de preferencia de los docentes de matemáticas de primaria es el Manual A (valor modal), pero que el Manual B también es un texto que tiene alta preferencia (por su cercanía con el valor modal).

Tercera situación posible

Tipo de manual	Preferido por los docentes en %
Manual A	34 %
Manual B	33%
Manual C	33%

En esta tercera situación nos encontramos con que el modo es Manual A (34%). Sin embargo, debemos notar que la variación en torno al modo es de un punto porcentual respecto a los manuales B y C (33% cada uno). Es decir, existe muy poca variación entre el valor modal y los restantes valores de frecuencias. La conclusión lógica es que los docentes de matemáticas de nivel primario, estarían escogiendo casi con igual frecuencia los tres tipos de manuales.

En conclusión, al identificar el valor modal de una distribución de datos, debemos ver en qué medida dicho valor se aparta del resto de las frecuencias representadas en las restantes categorías de la variable. Cuanto mayor la distancia, mayor variación en torno al modo y por tanto más representativo será éste.

Mediana

La mediana representa el valor de la distribución que deja por encima y por debajo la misma cantidad de casos, otra manera de expresarlo es: la mediana es un valor teórico que permite determinar, sobre el recorrido observado de la variable, el 50% de los casos que quedan en una posición superior e inferioralamisma. Para ver como se emplea esta medida, tomemos un conjunto ficticio de datos, que llamaremos notas del parcial de Estadística de la comisión A. El conjunto está compuesto por 19 valores que van entre 2 y 10. Al ordenarlos de menor a mayor encontramos que el valor 5 es el punto en que se puede separar al conjunto de datos en dos mitades iguales. Dicho en otras palabras, el valor 5 permite separar los casos en dos mitades de nueve casos.

2, 2, 2, 2, 4, 4, 4, 5, 5, 5, 5, 6, 10, 10, 10, 10, 10, 10, 10

En este punto es necesario reparar en que el valor de mediana se calcula sobre los valores observados de la variable, es por ello que situados en el valor 5 es posible separar dos conjuntos que contienen al 50% de los casos. Debe notarse que el valor de mediana es un valor posicional. Esto significa que el valor cinco señalado más atrás, es el que divide toda la distribución de notas en dos conjuntos iguales. En la distribución que se muestra resulta sencillo observar el valor que se posiciona "en medio" del conjunto total de valores.

En la distribución que hemos presentado el valor 5 coincide con un valor efectivamente observado, sin embargo, si la distribución de casos fuera par, deberíamos encontrar un valor de mediana utilizando los valores empíricos del centro del conjunto de datos. Continuando con el ejemplo, digamos que en una segunda instancia rinden el mismo parcial un conjunto de diez estudiantes cuyas notas son las siguientes:

4, 4, 5, 5, {6, 7}, 8, 8, 9, 10

Aquí se aprecia que la mediana es un valor que debe estar situado entre dos valores dados, y por tanto será el promedio entre ellos: Mdn={6 + 7}/2= 6,5.

4, 4, 5, 5, 6, 6,5, 7, 8, 8, 9, 10

Ahora, el valor 6,5 es aquel que separa el conjunto de datos en dos mitades iguales, aunque ese mismo valor no pertenece a la distribución original de datos.

La pregunta que sigue es cómo interpretar la mediana de una distribución, y para ello debemos tener en cuenta varios aspectos. Primero, debemos tener en cuenta los valores completos de la escala de medición y aquellos valores efectivamente observados. Centrémonos en el segundo ejemplo, si las notas observadas pertenecen a las calificaciones de un examen, cuya escaña es 1 a 10, vemos que no se han registrado todos valores de la misma. Es decir, no hubo alumnos o alumnas con notas de 1, 2 o 3. Esto es importante porque la mediana se calcula sobre los valores efectivamente observados. Entonces hay que contar los casos que obtuvieron 4, 5, 6, 7, 8, 9 y 10. Una vez ordenados los casos de menor a mayor, se calcula el valor de mediana. De esto se deduce que, si el número de observaciones es impar, existe un valor que separa en dos mitades iguales la distribución de datos; pero en el caso en que la distribución contenga un numero par de valores, la mediana se calcula como el promedio de los valores centrales.

Es importante retener el concepto de la mediana, dado que en raras ocasiones vamos a tener distribuciones tan poco numerosas de casos que permitan encontrar la mediana con cálculos tan sencillos como los mostrados más atrás. Entonces, la mediana es el valor que permite dividir una distribución de casos en dos mitades iguales, es decir, se corresponde con el 50% de la distribución. Entonces, al trabajar la mediana (sea como valor empírico o teórico), estamos utilizando este estadístico para entender la posición de los casos en relación a los valores observados de la variable.

En procedimientos para visualizar distribución de casos, la mediana es el estadístico de referencia. El más común de ellos es la gráfica de cajas que veremos más adelante. Por otro lado, la mediana es una medida que puede aplicarse a variables ordinales numéricamente valoradas, en tanto se trata de una medida que indica el "centro" de la distribución.

En este punto es importante despejar un error común en la interpretación del concepto de mediana, y es que éste estadístico debe ser un valor único ubicado en un punto equidistante de un conjunto heterogéneo de datos. Para ejemplificar esto veamos el siguiente ejemplo: supóngase que se tienen dos distribuciones de calificaciones de nueve alumnos cada una. La distribución 1, es más heterogénea que la distribución 2, la cual se ha visto restringida en su variación. Ambas distribuciones pueden tener la misma mediana, aunque la interpretación es diferente en cada caso.

Distribución 1 (n=9): 2, 2, 3, 4, 5, 6, 7, 8, 9
Distribución 2 (n=9): 4, 4, 4, 4, 5, 10, 10, 10, 10

Ambas distribuciones tienen la misma mediana, pero se desprende de la inspección visual de los datos que en la distribución 1 la mediana se ubica en un recorrido mayor de valores observados, en comparación con la distribución 2, que tiene un recorrido menor.

Media Aritmética o Promedio

La media aritmética es la medida que más se utiliza para resumir información. Puede definirse como el valor equidistante de un conjunto de valores y por esta propiedad sería el dato que mejor representa a dicho conjunto. Veamos un ejemplo de uso del promedio. En la siguiente tabla se muestra un resumen del resultado de la aplicación de tres métodos diferentes de aprendizaje de la lectura a escolares de primaria. Cada método ha recibido una denominación: A, B y C; la tabla expresa la cantidad de palabras que los niños fueron capaces de leer correctamente, luego de aplicarse un test de lectura.

Método	Método B	Método
22	35	44
18	26	49
19	30	51
30	30	52
25	12	39
12	17	43
15	28	50

Una manera de determinar cuál ha sido el método que da mejores resultados, es mediante la comparación del rendimiento promedio de cada uno de los grupos. Matemáticamente el promedio se define como la sumatoria de los valores de la variable, dividido el total de casos. La fórmula conceptual que resume el cálculo es la siguiente:

$$\bar{x} = \frac{\sum_{i=1}^{n} x}{n}$$

$\overline{x}$: media o promedio aritmético

$\sum_{i=1}^{n}$ x: indica la sumatoria de todos los valores de la variable
n: total de casos

El cálculo del promedio para los tres grupos en la prueba de lectura se muestra a continuación:

Método A	Método B	Método C
20,14	25,42	46,85

El promedio estaría reflejando el rendimiento general en lectura de los escolares que aprendieron bajo tres métodos diferentes. Nótese que el promedio es un único valor que debe expresar la centralidad de un conjunto de valores. Por tanto, en la distribución original, algunos de ellos quedarán por encima del promedio y otros quedarán por debajo de él. Por ejemplo, en el caso de la distribución de valores del método A, se tiene que el valor 30 se encuentra por encima del promedio, mientras que el valor 15 se encuentra por debajo. Adviértase que si el conjunto de datos no hubiera mostrado variabilidad (es decir, todos los escolares hubieran leído la misma cantidad de palabras), no haría falta calcular un promedio.

La variación de los valores en torno al promedio se denomina dispersión, y es posible obtener medidas de esta propiedad de los datos.

Varianza y Desviación Estándar

Veamos gráficamente que se quiere expresar con variaciones en torno al promedio. En el siguiente renglón se han ordenado los valores originales del método A, incluyendo el valor promedio:

12 – 15 – 18 – 19 - 20,14 - 22 – 25 – 30

Es de notar que el promedio es un valor que representa al conjunto de valores, los cuales pueden estar próximos o distantes de él. Mientras más cercanos entre si los valores originales, mayor será la proximidad de los mismos a su promedio y viceversa.

La medida que expresa la distancia del conjunto de valores al promedio, se denomina varianza. Conceptualmente se define como el promedio de las diferencias cuadráticas entre la media y los valores originales. La fórmula que lo expresa es la siguiente:

$$Var = \frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n}$$

$\sum_{i=1}^{n}(x-\bar{x})^2$: sumatoria de las diferencias de cada valore de *x* respecto a la media

n= cantidad de casos.

La fórmula de cálculo es similar a la del promedio, solo que en este caso el numerador indica que se debe realizar la sumatoria de las diferencias entre cada valor individual y el promedio. Dado que hay valores que se encuentran por encima del promedio y otros que se encuentran por debajo de él, el resultado debe elevarse al cuadrado. De otro modo, el resultado final sería siempre cero.

Es importante comprender que la varianza está expresando la dispersión de los valores originales respecto del promedio, y por tanto indica que tan homogéneos son esos valores. Dado que esta medida de dispersión esta expresada por valores elevados al cuadrado, se reemplaza por otra medida llamada **Desvío Estándar**. El desvío estándar es la raíz cuadrada de la varianza y su interpretación es equivalente en tanto medida de dispersión, pero con la particularidad de que los valores están en la misma unidad que la media.

$$de = \sqrt{\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n}}$$

Veamos cómo se presentan las varianzas y los desvíos estándar de los tres métodos de enseñanza de la lectura.

	Método A	**Método B**	**Método C**
Media	20,14	25,42	46,85
Varianza	37,14	65,28	23,8
D. E.	6,09	8,07	4,87

Dijimos al principio que necesitábamos una aproximación para determinar cuál método de aprendizaje de la lectura era el más eficiente, o al menos el que producía mejores resultados. Estas medidas en conjunto nos permiten aproximarnos con bastante precisión a tal respuesta:

a) El método de lectura C, es el que muestra el promedio más alto y la menor varianza, por tanto, es el que ha dado los mejores resultados.

b) El método B produce mejores resultados que el método A, puesto que su promedio es más alto. Pero la dispersión observada es mayor, de modo que el método A produce resultados más bajos en comparación con el método B, pero es más homogéneo en su composición.

c) El método A es el que produce los resultados más bajos.

La lectura del promedio junto con la desviación estándar nos estaría permitiendo ver que aquellos escolares que aprendieron con el método C, elevaron su rendimiento de manera pareja. En cambio, aquellos que aprendieron con el método B, solo algunos mejoraron su rendimiento, y finalmente los que aprendieron con el método A tuvieron un rendimiento más parejo, pero inferior al obtenido por el uso de los otros métodos de aprendizaje.

En el ejemplo anterior, los datos están disponibles y son solo siete casos, lo cual facilita la lectura de los estadísticos y los casos simultáneamente. Pero, el promedio, la varianza y la desviación estándar, se calculan para resumir conjuntos grandes de datos. Veamos el siguiente ejemplo: en la tabla que se muestra más abajo se presentan los resultados de un muestreo sobre ausentismo en el 3º ciclo de EGB y el Ciclo de Especialización. La variable ha sido medida en dos momentos diferentes sobre las mismas unidades de análisis (alumnos). Para interpretar el conjunto de datos recogido se han resumido en los estadísticos de media, varianza y desviación estándar.

	Primer Cuatrimestre		Segundo Cuatrimestre	
	3º EGB	CE	3º	CE
Media	16	15,83	9,33	16,33
Varianza	56	60,56	13,46	4,667
DE	7,48	7,78	3,66	2,16

Los datos que se presentan muestran que, en el primer cuatrimestre, tanto en el 3° ciclo de EGB como en el Ciclo de Especialización, la media y la dispersión en la cantidad de faltas de los alumnos es similar. Sin embargo, en el segundo cuatrimestre es factible observar que el promedio de faltas del 3° ciclo de EGB es más bajo, mientras que en Ciclo de Especialización el promedio se mantiene muy similar. Paralelamente, la varianza del 3° ciclo de EGB es menor en el segundo cuatrimestre, lo cual refleja que, en conjunto, los escolares están faltando menos. Para el Ciclo de Especialización, la varianza también es menor, pero dado que el promedio sigue siendo prácticamente muy similar al del primer cuatrimestre, es factible deducir que el grupo tiende a mostrar un comportamiento particular y es que aquellos escolares que faltaban poco a la escuela, ahora faltan más y aquellos que faltaban mucho, ahora faltan menos.

Conociendo la media y el desvío estándar, se puede determinar la posición relativa de un individuo en relación al grupo. Así, por ejemplo, de un alumno que obtiene una calificación de 6 en matemáticas y de 8 en inglés, nos inclinaríamos a pensar que le fue mejor en ésta última materia. Sin embargo, si consideramos la media y el desvío estándar del grupo en ambas materias, esta primera impresión resulta injustificada.

Materia	**Media**	**D. E.**
Matemáticas	4.0	1.68
Inglés	8.55	1.34

Nótese ahora que una calificación de 6 en matemáticas está por encima de la media en más de una desviación estándar, mientras que una calificación de 8 en inglés está a menos de una desviación estándar por debajo de la media. Ampliaremos esta lógica de análisis cuando veamos distribución normal.

Lo antes dicho nos lleva a enfatizar que todas las medidas de tendencia central deben acompañarse de medidas de dispersión, ambas nos dan una aproximación más certera del comportamiento de la variable en el conjunto de unidades de análisis estudiadas.

Coeficiente de variación de Pearson

El coeficiente de variación V de Pearson, viene a resolver el problema de comparar la variabilidad de diferentes conjuntos de datos, que han sido medidos con distintas escalas. Se define como el cociente entre la desviación estándar de la distribución y su media aritmética. Como se deduce, el coeficiente es una manera de determinar cuántas veces la desviación estándar contiene a la media, y por tanto nos informa de la dispersión general. Dado que no depende de la escala en que ha sido medida la variable, lo hace apto para comparar distintas distribuciones. Una medida general del coeficiente se logra multiplicándolo por cien, y en este caso es una medida porcentual de variación. Su formulación matemática es la siguiente:

$$v = \frac{\sigma}{\bar{x}}$$

Veamos el siguiente ejemplo para ilustrar su utilidad: Un psicólogo educacional tiene datos de tres muestras de individuos que han sido evaluados con diferentes pruebas de inteligencia. Desea determinar el nivel de dispersión para cada una de ellas y verificar cuál es la que tiene menos dispersión. En esta situación la desviación estándar no es la medida óptima pues al usarse pruebas diferentes de inteligencia, no se puede comparar entre sí. Para ello aplica el coeficiente de variación *v* de Pearson, que para este ejemplo se obtienen los siguientes resultados.

	Media	**Desviación Estándar**	**Coeficiente de Variación (%)**
Muestra A	8	5,51	68,87
Muestra B	125	46	36,8
Muestra C	510	212	41,57

Independientemente de la prueba de inteligencia utilizada, se observa que la muestra que menor dispersión evidencia es la B.

Procedimiento para el cálculo del promedio, la varianza y la desviación estándar

Ya no se calculan a mano los estadísticos que hemos repasado hasta aquí, puesto que existen programas como el PSPP con esas funciones, incluso una calculadora científica puede hacer ese trabajo. El procedimiento que mostramos a continuación tiene la finalidad de ilustrar la aplicación de la fórmula de cálculo de la media, la varianza y la desviación estándar. Se basa en la construcción de una tabla que ayuda en la serie de cálculos que estas medidas requieren.

A continuación, se presenta una serie de puntajes obtenidos por siete alumnos en una prueba de ortografía, donde se contabilizó la cantidad de palabras escritas correctamente al dictado.

Alumno	Cantidad de palabras correctas
1	28
2	36
3	40
4	30
5	40
6	25
7	35

Para el cálculo de los estadísticos requeridos, primero es necesario agregar algunas columnas a la tabla; estas columnas son:

$|x - \bar{x}|$ representa la diferencia de cada valor individual respecto a la media, sin considerar el signo de la sustracción, esto es, el valor absoluto de la diferencia.

$(x - \bar{x})^2$ elevar al cuadrado el resultado de la sustracción anterior.

Alumno	**Cantidad de palabras correctas**	$\|x - \bar{x}\|$	$(x - \bar{x})^2$
1	28	5,42	29,37
2	36	2,58	6,65
3	40	6,58	43,29
4	30	3,42	11,69
5	40	6,58	43,29
6	25	8,42	70,89
7	35	1,58	2.49
n=7	**∑=234**		**∑=205,18**

Con las dos primeras columnas de la tabla y aplicando la fórmula presentada anteriormente, se obtiene el valor de la media: $\bar{x}$= 234/7=33,42.

Lo que sigue es restar este valor a cada observación para completar la siguiente columna. De este modo, para el alumno 1 se tendrá: 28 - 33.42= -5,42, puesto que no consideramos el signo de la resta utilizaremos el valor absoluto: 5,42. La siguiente columna se completa elevando al cuadrado el resultado obtenido: (5,42)2=29,37. Realizando esta operación para cada uno de los alumnos, obtendremos la columna final, cuya sumatoria nos permitirá obtener la varianza:

Var= 205,18/7= 29,31.

La desviación estándar, es la raíz cuadrada de la varianza, por lo tanto:

DE= $\sqrt{29,31}$=5,41.

Así, para el conjunto de datos analizados se tiene que

Promedio	**Varianza**	**Desviación Estándar**
33,42	29,31	5,41

Las medidas de dispersión que hemos presentados, son útiles cuando las variables están expresadas en sistemas de medición métrico. En el caso en que las variables se expresen en sistemas de medición ordinal numéricamente valoradas y, además, dicha variables posea una variación no menor a diez puntos, puede utilizarse una medida simple de dispersión conocida como **rango de variación**, o rango, para abreviar. El rango también puede ser utilizado en escalas métricas, pero su uso aporta mucho menos información que la varianza y la desviación estándar. Veamos un ejemplo: a un grupo de docentes se les pide valorar la siguiente afirmación: *"creo que es provechoso el uso de los teléfonos celulares en el aula"*. La escala de valoración es la siguiente; 1= completamente de acuerdo – 10=completamente en desacuerdo. En este caso, las y los docentes deberán puntuar la afirmación dentro de una escala con recorrido de 1 a 10. Seleccionamos diez docentes y los dividimos en dos grupos, a los que denominaremos A y B para simplificar, los cuales se representan en la siguiente tabla.

Grupo A	**Grupo B**
7	6
6	5
2	7
1	3
9	4

Como puede observarse a simple vista, la variabilidad de los datos aportados por las y los docentes del grupo A es mayor que la del grupo B. En caso en que necesitemos una única medida sintética de variabilidad, utilizamos el rango. La operación consiste en a) identificar los valores mínimo y máximo en la distribución de datos, y b) restar el mayor valor observado al menor valor observado. Para el grupo A, el valor máximo es 9; el valor mínimo es 1; por lo tanto, su rango de variabilidad es 9 – 1=8. Para el grupo B, el valor máximo es 7; el valor mínimo es 3; por lo tanto, su rango de variabilidad es

7 - 3=4. La diferencia en los rangos observados en el grupo A y B indica que se

hallan representados más valores de la escala teórica en el grupo A; otra manera de decirlo es que en el grupo B se registra menos variabilidad que en el grupo A.

Como resulta evidente, una desventaja del rango es que solo utiliza dos valores del conjunto total de datos (el valor máximo y el mínimo), por lo cual, es una medida inestable de variabilidad. Es decir, no representa adecuadamente todo el conjunto de valores. Su ventaja radica en que es muy fácil de calcular y nos da una primera aproximación de la dispersión cuando la escala es de tipo métrica, y es la única que se puede utilizar en escalas de tipo ordinales numéricamente valoradas. Cuando se usa el rango en escalas ordinales, es necesario destacar que su interpretación se restringe solo al recorrido efectivamente observado respecto de los valores teóricos de la escala. En el ejemplo que hemos dado más atrás, la escala teórica tiene un rango de 9 puntos (R= 10 – 1=9). En el grupo A, los valores extremos de la escala han sido observados puesto que su rango está muy próximo al valor teórico. Pero en el grupo B, no se registran valores tan extremos como en el grupo A, dado que su rango es solo de 4.

El uso e interpretación de los rangos en escala ordinales no va más allá de esta interpretación, dado que en teoría no existen intervalos iguales en la escala, por lo tanto, la comparación entre dos rangos en escala ordinal no debe interpretarse como valores absolutos.

Guía de Actividades N°2:

Análisis descriptivos-Medidas de tendencia central- Medidas de dispersión.

Objetivos de la actividad:

a) calcular y utilizar diferentes medidas de tendencia central y de posición para caracterizar un conjunto de datos.

b) calcular y utilizar diferentes medidas de variabilidad para caracterizar un conjunto de datos.

Actividad 1: Lea atentamente las siguientes explicaciones referidas a las escalas de medición y diga si es Verdadera o Falsa y justifique su respuesta.

a) La única medida de tendencia central que puede calcularse para las escalas nominales es el Modo.

b) En una distribución de datos nominales siempre existirá un único valor modal.

c) El Modo sólo puede calcularse cuando se han atribuido números a la escala nominal.

d) En escalas donde el atributo está ordenado, puede calcularse la Media y el Desvío Estándar como medidas de tendencia central y variabilidad respectivamente.

e) Para calcular el modo en las escalas ordinales, éstas no deben contener el valor cero.

f) En las escalas ordinales no es posible calcular el rango.

g) Sólo en las escalas métricas puede calcularse la Varianza.

h) Una escala métrica puede ser reducida a una escala ordinal.

i) No es posible utilizar los cuartiles en las escalas métricas.

Actividad 2: La tabla que se presenta a continuación, representa la cantidad de palabras escritas correctamente por un grupo de escolares en una prueba de ortografía. La prueba utilizada es un instrumento estandarizado y toma los siguientes criterios de rendimiento:

Rendimiento bueno	**Rendimiento regular**	**Rendimiento insatisfactorio**
29 errores o menos	entre 30 y 44 errores	45 errores o más

Con los datos presentados se pide:

a) Identifique Variable/s y Unidad de Análisis,

b) Identifique el intervalo modal, y los intervalos que contienen los Cuartiles en la tabla.

c) Teniendo en cuenta los criterios de la prueba, intente caracterizar en sus propias palabras el rendimiento de los escolares.

Rendimiento en la prueba de ortografía

Cant. Palabras	Casos
20 – 24	10
25 – 29	12
30 – 34	12
35 – 39	10
40 – 44	14
45 – 49	20
50 – 54	16
55 – 59	6
Total	

Actividad 3: En una escuela se realizó una evaluación del nivel de conocimiento de matemáticas de los alumnos de sexto grado de ese establecimiento. Los resultados obtenidos se expresaron mediante algunas medidas de tendencia central y de variabilidad, tal como se muestra en la tabla que sigue.

Medidas de tendencia central de una muestra de escolares

	M_x	Minimo	Maximo	Q1	Mdn	Q3	Rango
Puntaje en matemát	34.58	17.0	59.0	28.0	34.5	40.0	42.0

Posterior a la recolección de estos datos, ingresa al establecimiento un nuevo grupo de cinco niños, quienes al ser evaluados en matemáticas mostraron los siguientes puntajes:

1) 49; 2) 37; 3) 19; 4) 25; 5) 32.

Dados los puntajes de estos escolares, se solicita que establezca su posición en función de la media, la mediana, el máximo y el mínimo.

Actividad 4: En un nuevo proyecto dentro de la escuela se aplicaron tres métodos diferentes de enseñanza de la lectura a un total de 15 alumnos de sexto grado. Los resultados, luego de aplicar la misma prueba de rendimiento lector, se muestran en la siguiente tabla.

Puntajes en prueba de lectura según método

Método A	Método B	Método C
22	35	44
26	26	49
19	30	51
30	41	52
25	52	39

Con los datos ofrecidos:

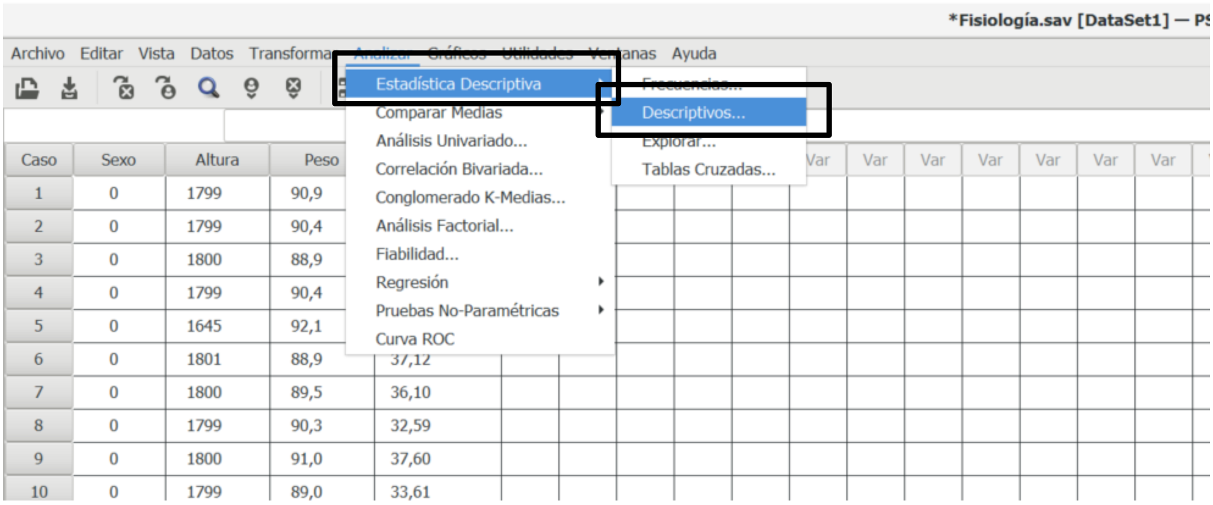
a) Calcule la Media, Varianza y Desvío Estándar de los puntajes.

b) Analice las diferencias entre cada uno de los métodos de acuerdo a los estadísticos obtenidos.

Actividades en PSPP N° 2:

En la guía de actividades del capítulo 1 hemos abierto una base de datos, que trae como ejemplo el programa PSPP, que se llama Phisiology.sav y lo renombramos por el archivo Fisiología.sav, al que le cambiamos el nombre de las variables en inglés por su equivalente en español, modificamos las etiquetas de la variable sexo, originalmente en inglés, por las etiquetas en español. Finalmente hemos corregido un error de carga de datos en la variable altura.

En esta ocasión vamos a procesar las variables de esta base de datos que hemos adecuado previamente. Para ello, lo primero que hará será dirigirnos al menú principal y seleccionar el comando Analizar, dentro de él se ubicará en Estadísticos descriptivos y posteriormente en el comando Descriptivos…



Una vez allí, vamos a seleccionar la/s variable/s que queremos procesar. En este caso comenzaremos con las variables Altura y Peso, que son variables que tienen un nivel de medición……………………………….. y por lo tanto podremos pedirle todas las medidas estadísticas de posición y de dispersión. Noten que podemos desplazarnos por el cuadro de Estadísticos para tildar las medidas que corresponden a esta variable.

Una vez seleccionadas las variables en la casilla de Estadísticos: vamos a pedir la media (o promedio), la Desviación estándar, el Mínimo, el Máximo, el Intervalo (o rango), la Suma y la Varianza. El resultado será la siguiente tabla:

Casos válidos = 40; casos con valor(es) perdido(s) = 0.

Variable	*N*	*Media*	*Desv Std*	*Varianza*	*Intervalo*	*Mínimo*	*Máximo*	*Suma*
Altura en milímetros	40	1717,40	101,07	10215,28	305,00	1598,00	1903,00	68695,94
Peso en kilogramos	40	72,12	26,71	713,17	147,67	-55,60	92,07	2884,83

De ella podemos decir que la variable Altura en milímetros tiene 40 casos, con una media en 1717.4 milímetros, es decir aproximadamente 1.72 metros. Que el mínimo de la variable es 1598 milímetros, es decir quién menor altura tiene, y el máximo 1903 milímetros, el más alto. Por lo tanto, la diferencia entre el mínimo y el máximo, que es el Intervalo, es de 305 milímetros, aproximadamente 30 centímetros.

También nos aporta el valor de la suma de todos los casos, que en esta ocasión no tendrá aplicación. Además, vamos a disponer de la varianza ya calculada por el programa, la desviación estándar y podemos calcular el coeficiente de variación, dividiendo la desviación estándar con el promedio, que en el caso de la Altura en milímetros dará 101.07/1717.40=0.0589, que multiplicado por 100 nos dará el porcentaje de un 5.89% de variación respecto del promedio.

Ahora se le pide que usted describa la variable Peso en kilogramos del mismo modo en que hemos descripto la variable Altura:

………

………

………

………

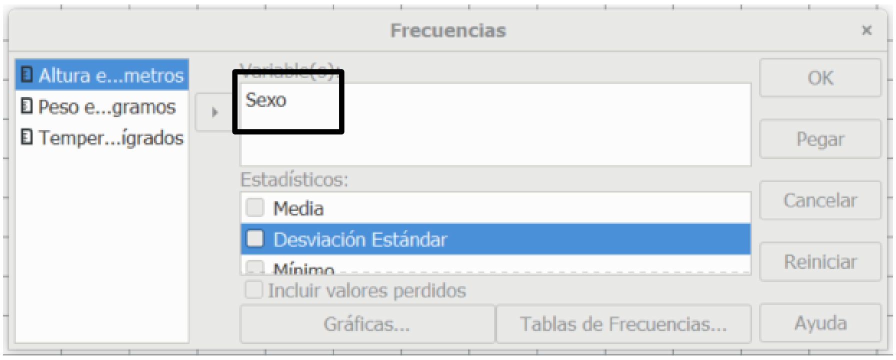
………

……
……
……
……
……

Ahora le pedimos que usted, por su cuenta, realice estas operaciones en el programa PSPP para la variable Temperatura, pegue la tabla y describa los resultados obtenidos.
……
……
……
……
……
……
……
……
……
……

Finalmente, se puede calcular la proporción de varones y mujeres de la variable Sexo, que tiene un nivel de medición ………………………………………… Por medio de la opción Analizar, Estadística Descriptiva, Frecuencias…

Una vez allí, dentro del menú de Frecuencias, vamos a pedir cuántas veces se repite cada categoría de la variable sexo entre los 40 casos. Para ello seleccionamos Sexo y lo desplazamos hacia la derecha en el cuadro de variables y verificamos que en Estadísticos no esté ninguno cargado:



Le damos OK y tendremos la siguiente salida:

Sexo del sujeto

Etiqueta de Valor	*Valor*	*Frecuencia*	*Porcentaje*	*Porcentaje Válido*	*Porcentaje Acumulado*
Varón	0	22	55,00	55,00	55,00
Mujer	1	18	45,00	45,00	100,00
	Total	40	100,0	100,0	

Lo primero que podemos ver, es que a la izquierda de la tabla aparecen las dos categorías de la variable Sexo: Varón y Mujer. Luego podemos ver los valores con los que éstas han sido codificadas y la frecuencia para cada una de ellas, por ejemplo, en este caso tenemos 22 varones y 18 mujeres. La columna que sigue a la derecha indica el porcentaje que da un 55% para los varones y un 45% para las mujeres. La columna siguiente, Porcentaje Válido, es útil en caso que haya valores perdidos (más adelante veremos qué significan), en este caso replica la columna anterior. Finalmente está el porcentaje acumulado, que al ser sólo dos categorías en la segunda línea acumula el 100%. El porcentaje acumulado sólo tendrá sentido en las variables con nivel de medición ordinal y métrico, es decir, donde podemos acumular.

Si bien Tablas de frecuencia es el tema que veremos en el capítulo siguiente, en esta oportunidad, y con sólo dos categorías, como ha sido codificada la variable sexo aquí (actualmente en los relevamientos esta variable debe incluir al menos la categoría otros), determinaremos que el valor modal es varón, porque es la categoría de la variable que acumula la mayor cantidad de casos, en esta ocasión 22 individuos.

Capítulo 3

Tablas de Frecuencia

En este capítulo veremos las tablas de frecuencia, que representan una manera sencilla de resumir información sobre una (o varias) variables y se basan en el conteo de la cantidad de unidades que quedan comprendidas en sus diferentes niveles de medición. Las tablas casi siempre acompañan a los textos y son las referencias que se toman cuando es necesario mostrar comparaciones o destacar algún fenómeno puntual. Una tabla de frecuencia se construye a partir de filas y columnas, y por conveniencia, se utilizan las primeras para los valores de la variable y las segundas para los distintos tipos de frecuencia que es posible calcular. En los apartados que siguen se mostrarán ejemplos de tablas de frecuencia para los sistemas de medición nominal, ordinal y métrico. Además, dedicaremos una gran parte del capítulo a ver gráficos asociados a las tablas de frecuencia, como gráficos de barras, sectores, histogramas, diagramas de cajas, entre otros.

Tabla de frecuencia: sistema de medición nominal

Supongamos que una investigadora ha realizado una medición de la variable estado civil, en el conjunto de todos los empleados de una facultad, donde N=345. Los niveles de medición de la variable estado civil son: a) Solteros, b) Casados, c) Separados, d) Viudos, e) Otras. Para construir una tabla de frecuencia, es necesario determinar por conteo cuántas personas se encuentran comprendidas en cada una de las categorías; la tabla de frecuencia para este ejemplo podría tener la siguiente distribución de casos:

Estado civil del personal de la facultad (N=345)

Variable	**Frecuencia**
Solteros	122
Casados	107
Separados	110
Viudos	1
Otras	5
Total	345

Nótese que en la columna frecuencia absoluta realizamos el conteo de casos en cada una de las categorías de la variable, y así establecemos que en la facultad existen 122 personas solteras, 107 de ellas están casadas, etc. Puesto que el procedimiento consiste en contar unidades (personas en este caso), en cada uno de los niveles de la variable, la suma de la columna frecuencia absoluta debe ser igual al total de casos.

Las tablas de frecuencia también pueden contener las frecuencias relativas y estas se pueden transformar a porcentajes. Trabajar con porcentajes es más operativo que hacerlo con las frecuencias absolutas, dado que éstos nos permiten realizar comparaciones directas en caso de tratar con otra tabla que contenga información sobre la misma variable, pero con distinto número de casos. La frecuencia relativa se obtiene del cociente entre la cantidad de casos en cada uno de los niveles de la variable, sobre el total de casos. El porcentaje, se obtiene de multiplicar la frecuencia relativa por 100. En la siguiente tabla se grafica el procedimiento, y luego se presenta resumido:

Estado civil del personal de la facultad (n=345)

Variable Estado Civil	**Frecuencia Relativa**	**Porcentaje**
Solteros	122/345=0.3536	0.3536x100=35.36%
Casados	107/345=0.3101	0.3101x100=31.01%
Separados	110/345=0.2898	0.2898x100=28.98%
Viudos	1/345=0,0028	0,0028=0.28%
Otras	5/345=0,0144	0,0144x100=1.44%

Resumen

Variable	**Frecuencia Relativa**	**Porcentaje**
Solteros	0.3536	35.36%
Casados	0.3101	31.01%
Separados	0.2898	28.98%
Viudos	0,0028	0.28%
Otras	0,0144	1.44%
Total	0.9707≈1	97.07%≈100%

Vamos a tomar los valores de la segunda tabla, donde se encuentra resumido el procedimiento. Nótese que la frecuencia relativa y el porcentaje representan la misma información, pero dado que el porcentaje es un valor más familiar podríamos concluir que el 35.36% del personal de la facultad está compuesto por personas solteras, que el 31.01% de ellos son casados, etc. Como se aprecia, la información absoluta puede expresarse de una manera diferente sin perder valor.

Dado que para obtener las frecuencias relativas debemos dividir por el total de casos, siempre obtendremos un número decimal. En caso de obtener un valor decimal no periódico, se deberán utilizar solo algunos decimales; por ejemplo, el valor real de la frecuencia relativa para solteros es de: 0,35362318840579710144927536231884, un número con más de treinta cifras en su valor decimal. Puesto que tanta precisión no es necesaria, se suele usar dos, tres o cuatro decimales, tal como se hizo en la tabla. En tal caso, y si no se ha redondeado el valor decimal, la suma a través de la columna de frecuencias relativas dará un número muy próximo a 1. Puesto que el porcentaje es la frecuencia relativa multiplicada por cien, la suma a través de la columna porcentaje se aproximará a 100.

Una manera de reducir la cantidad de decimales es el procedimiento conocido como redondeo. Para aplicar este procedimiento se debe: a) determinar el número de decimales que necesitamos, es decir, la precisión con la que queremos trabajar, b) ubicar el número decimal a redondear siguiendo la regla: 5 o mayor se cambia el valor decimal inmediato superior; menor a 5, no se cambia el valor decimal inmediato superior. Un ejemplo no permitirá ver lo sencillo del procedimiento. Vamos a tomar el valor completo de la frecuencia relativa de la categoría Solteros de la tabla:

Solteros: f r = 0,35362318840579710144927536231884

Si queremos redondear el valor a tres decimales debemos contar cuatro lugares luego de la coma y verificar el número que queda en esa posición:

f r = 0,35362318840579710144927536231884

Puesto que el valor observado es mayor que cinco, el decimal inmediato anterior se redondea sumando una unidad, en tal caso la frecuencia relativa redondeada quedará expresada de la siguiente manera:

f r = 0,354

Si hubiésemos optado por redondea a dos cifras decimales, el procedimiento sería el siguiente:

f r = 0,35362318840579710144927536231884

En este caso, el valor decimal en la tercera posición es menor a cinco, por lo tanto, no sumamos una unidad al decimal inmediato anterior, y el valor redondeado queda expresado de la siguiente manera.

f r = 0,35

Otra forma de trabajar con una tabla de frecuencia es a través de sus frecuencias acumuladas. Para obtener las frecuencias acumuladas, se deben sumar de manera descendente los valores de frecuencia en cada una de las celdas de la categoría de la variable. En las siguientes tablas se presenta el procedimiento y su resumen, utilizando como base la frecuencia absoluta y el porcentaje:

Estado civil del personal de la facultad (N=345). Frecuencia absoluta acumulada

Variable Estado Civil	**Frecuencia Absoluta**	**Frecuencia Absoluta Acumulada**
Solteros	122	122
Casados	107	122+107=229
Separados	110	229+110=339
Viudos	1	339+1=340
Otras	5	340+5=345
Total	345	

Porcentaje acumulado

Variable Estado Civil	**Porcentaje**	**Porcentaje Acumulado**
Solteros	35.36%	35.36%
Casados	31.01%	35.36+31.01=66.37%
Separados	28.98%	66.37+28.98=95.35%
Viudos	0.28%	95.35+0.28=95.63%
Otras	1.44%	95.63+1.44=97.07%
Total	97.07%□100%	

El procedimiento para obtener la frecuencia acumulada es el mismo, tanto si se trata de la frecuencia absoluta o del porcentaje. Véase la tabla de frecuencia absoluta; en la categoría de la variable solteros, se cuentan 122 individuos; puesto que no hay ninguna categoría por encima, en este punto solo se cuentan esos individuos. La siguiente categoría es casados, y en ella se cuentan 107 individuos, que sumados a los 122 de la categoría solteros dan un total de 229 personas; por lo tanto, la frecuencia absoluta acumulada al contar solteros y casados es de 229 personas. Si a estas le sumamos la categoría separados, debemos añadir a las 229 personas contabilizadas, otras 110, que es la frecuencia para la categoría soleteros, y entonces tendremos 339 personas contabilizadas en las tres primeras categorías de la variable. El procedimiento se repite para las dos categorías restantes, y en la última categoría que sumemos, obtendremos el total de casos. Con el porcentaje se ha procedido de la misma manera, y en resumen la tabla se presentaría de la siguiente forma:

Estado civil del personal de la facultad (n=345). Resumen

Variable Estado Civil	**Frecuencia Absoluta Acumulada**	**Porcentaje Acumulado**
Solteros	122	35.36%
Casados	229	66.37%
Separados	339	95.35%
Viudos	340	95.63%
Otras	345	97.07%

Una conclusión que puede derivarse de la observación de los porcentajes acumulados es que más de la mitad del personal de la planta son personas solteras y casadas, que junto con los individuos separados conforman la mayoría del personal.

Utilidad de la tabla de frecuencia a través de un ejemplo

En una escuela con más de dos mil alumnos se desea adaptar la cobertura social de sus trabajadores, permitiendo la opción de contratar el servicio en distintas prestadoras de salud y de acuerdo a las reales necesidades del trabajador. Para cumplir con este objetivo, s e le solicita a un investigador social, que sugiera algunas recomendaciones a los efectos de que la propuesta pueda ser bien recibida por los docentes. Dadas las circunstancias, el investigador supone que la principal preocupación del docente en relación a su cobertura social, sería la prestación extendida al grupo familiar. Por ende, en una primera etapa, realiza un relevamiento del estado civil de los 345 docentes. Los resultados obtenidos se presentan en la siguiente tabla, de la cual se pueden sacar las conclusiones que siguen:

Resumen

Estado Civil	**Frecuencia Absoluta**	**Porcentaje**	**Porcentaje Acumulado**
Solteros	122	35.36%	35.36%
Casados	107	31.01%	66.37%
Separados	110	28.98%	95.35%
Viudos	1	0.28%	95.63%
Otras	5	1.44%	97.07%
Total	345	100%	

a) De los planes de salud al menos el 31% de ellos, debería incluir una ampliación para el cónyuge.

b) Aproximadamente el 60% de los planes de salud deberá incluir en el grupo familiar a uno o más hijos. Contando con que el 31% ya incluya a un cónyuge.

c) Al menos el 35% de los planes de salud podría incluir a un solo individuo como beneficiario.

Tabla de frecuencia: sistema de medición ordinal

Cuando se trabaja con una variable de tipo ordinal con las categorías expresadas como etiquetas, el procedimiento para construir una tabla de frecuencia es idéntico al que se usa para una variable medida en escala nominal. Si la variable se expresa como una escala numérica, cada valor es equivalente a la etiqueta de la variable.

Veamos el siguiente ejemplo: un sociólogo desea medir el grado de cohesión de un grupo de estudiantes y además el sentido con que identifica las interacciones entre individuos. Para ello, desarrolla dos escalas likert a partir de los siguientes reactivos:

a) Exprese su grado de acuerdo con la siguiente afirmación: *"el grupo al que pertenezco se encuentra muy unido"*. Utilice según su criterio una escala de 1 a 10, siendo: 1=completamente en desacuerdo; 10= completamente de acuerdo.

b) Juzgue de qué manera Usted percibe las interacciones entre los miembros del grupo. Utilice según su criterio una escala de -5 a +5, siendo:

-5= los individuos interactúan de modo completamente agresivo;

+5= los individuos interactúan de modo completamente colaborativo;

Nótese que, en este ejemplo, se han utilizado dos tipos diferentes de escalas, pero siguiendo un criterio similar. En el segundo caso, el signo positivo o negativo indica el polo al que tienden las interacciones, agresivas o colaborativas. Suponiendo que el investigador planea realizar una serie de tareas con la finalidad de afianzar la pertenencia grupal de los estudiantes, necesitará al menos dos mediciones para verificar si este aspecto del grupo se ha modificado. Tales mediciones se tomarán en antes y después de las tareas de integración. Continuando con el ejemplo, la tabla que se muestran a continuación refleja lo sucedido:

Grado de cohesión del grupo n=55

	Antes		Después	
Escala de medición	**Frecuencia Absoluta**	**Porcentaje**	**Frecuencia Absoluta**	**Porcentaje**
1	1	1.81	0	0
2	3	5.45	2	3.63
3	2	3.63	3	5.45
4	8	14.54	5	9.09
5	14	25.45	12	21.81
6	12	21.81	11	20
7	10	18.18	19	34.54
8	2	3.63	1	1.81
9	2	3.63	1	1.81
10	1	1.81	1	1.81
Total	**55**	**99,94**	**55**	**99,95**

Vamos a proponer una posible interpretación de la anterior tabla frecuencias: en primer lugar, se observa que las frecuencias tienden a concentrarse en los valores medios de la escala propuesta, la cual refleja el grado de cohesión percibido por los miembros del grupo de estudiantes. En este sentido, las frecuencias están en los valores 5, 6 y 7; al estar estos valores en el rango medio, es posible suponer que el grupo en general se percibe medianamente cohesionado.

Ahora bien, se proponen una serie de actividades que se espera que fortalezcan la cohesión del grupo de estudiantes, por lo tanto, un cambio en la variable, debería quedar reflejada en una distribución de frecuencias concentrada en los valores altos de la escala. Partiendo de esta interpretación, la columna de la tabla de frecuencias que refleja lo sucedido luego de la intervención del sociólogo y sus actividades grupales, muestra que las frecuencias altas siguen concentradas en los valores medios de la escala, apareciendo como diferencia entre la situación antes y después, la acumulación de casos en el valor 7 de la escala luego de la intervención (34.5%).

En general, las distribuciones de frecuencias para la situación antes y después no presentan variaciones notables. Esta distribución de datos, sirve como disparador de preguntas de investigación que podrían ser las siguientes:

a) ¿han sido adecuadas las intervenciones planeadas?
b) ¿el tiempo transcurrido entre las mediciones antes – después ha sido el suficiente como para apreciar un cambio en el grupo
c) ¿se trata de un grupo reactivo al cambio?
d) ¿los alumnos han logrado identificar adecuadamente las acciones que tienden a la cohesión?
e) ¿el instrumento de medición es el apropiado para captar un cambio sutil pero significativo en la cohesión grupal de estudiantes?

Supóngase que, en la pregunta sobre el modo de interacción, veinte de los estudiantes encuestados no dan respuesta. Bajo esta circunstancia, una escala de once categorías ordenadas como la que se propone (nótese que aquí se proponen cinco valores positivos, cinco negativos y un cero para el neutro), puede que contenga varias celdas vacías o con frecuencia cero. Bajo estas circunstancias, sería conveniente para el investigador reducir las categorías de la escala; una forma de hacerlo es agrupando en categorías los valores originales. La tabla que sigue es un posible ejemplo de cómo podría hacerse.

Categoría	Las interacciones son en su mayoría agresivas	A veces las interaccio-nes son agresivas	Las interaccio-nes son neutras	A veces las interaccio-nes son colaborativas	Las interacciones son en su mayoría colaborativas
Puntaje Original	-5 -4	-3 -2 -1	0	1 2 3	4 5

Al reducir de este modo la escala se evitaría trabajar con categorías con frecuencia cero. Como se aprecia, el procedimiento de reducción de la escala consiste en agrupar los valores de ésta en etiquetas que reflejen los valores originales, es decir, que respete el sentido del ordenamiento de la escala. El resultado de tal agrupamiento produce una escala con cinco categorías ordenadas, cuya regla de asignación es contar los individuos que puntuaron en la escala con -5, -4 en la categoría: Las interacciones son en su mayoría agresivas, luego los individuos que puntuaron en la escala -3, -2 y -1 se agrupan en la categoría: A veces las interacciones son agresivas; el procedimiento se repite para todos los valores observados.

Con la escala de medición reducida, es posible construir una tabla de frecuencias que resulte visualmente manejable y posibilite una aproximación más sencilla a su interpretación. En la siguiente tabla, el investigador presenta los datos totales del grupo de estudiantes, y en columnas donde se han separado las respuestas por su género:

Percepción de las interacciones del grupo de estudiantes n=25

Las interacciones en el grupo son:	**Total**	**Varones**	**Mujeres**
En su mayoría agresivas	1	1	0
A veces son agresivas	7	6	1
Son neutras	6	2	4
A veces son colaborativas	7	1	6
En su mayoría colaborativas	4	2	2
Total	25	12	13

Para interpretar la información de la tabla debemos considerar primeramente la pérdida de respuestas, dado que involucra un número importante de individuos. Deberíamos concentrarnos en determinar si la pregunta ha sido apropiada, si hay variables externas al estudio que estuvieran determinando que las personas no respondan, etc.

Una línea interpretativa, podría llevarnos a comparar la distribución de frecuencias del total, con la que aparece segmentada por género. En este caso se aprecia que las frecuencias más altas están en el rango medio de la escala, con una tendencia hacia el polo colaborativo de esta. Véase la distribución de frecuencias del total y tenemos que prácticamente la misma cantidad de individuos han percibido las interacciones grupales como: A veces son agresivas, Son neutras, A veces son colaborativas. En esas categorías se contabilizan 20 individuos; de los 5 restantes que forman el total, 4 han puntuado en la categoría En su mayoría colaborativas.

Si se proyecta esa tendencia sobre las frecuencias discriminadas por género, aparecen diferencias que merecen ser atendidas: a) la frecuencia más alta en el grupo de varones se encuentra en la categoría A veces son agresivas, b) en el grupo de mujeres las frecuencias tienden a agruparse en las categorías Son neutras, A veces son colaborativas. Algunas preguntas importantes que se desprenderían de estos datos podrían ser:

a) ¿qué factores habrían influido en la falta de respuesta a la pregunta?

b) ¿existe verdaderamente diferencias de género en las respuestas, o simplemente se deben a la escasa cantidad de personas que respondieron?

c) ¿las personas que respondieron son las que más interacciones con el resto del grupo han mostrado?

d) ¿las personas que no han respondido son quienes desplegaron menos interacciones con el resto del grupo?

Tablas de frecuencias para variables métricas

Las variables métricas utilizan un continuo de valores y tienen propiedades en las que los números admiten las operaciones matemáticas. Conviene recordar que este tipo de variables pueden ser continuas si admiten cualquier valor, o bien discretas, en cuyo caso solo se admiten números enteros. Por lo tanto, las tablas de frecuencia de estas variables suelen agruparse en intervalos dado que, entre dos valores cualesquiera, pueden encontrarse otros tantos valores fraccionarios, o pueden tener un rango de valores muy amplios aun siendo discretas.

Ejemplo: en una escuela se aplicó una prueba de madurez lectora (Prueba de figuras inversas de Edfeldt), para determinar el nivel de aprestamiento de los alumnos al comienzo del ciclo lectivo en el primer grado. La prueba evalúa el nivel de madurez visoespacial del alumno para el aprendizaje de las letras del alfabeto. Se evaluó una muestra de 75 niños, y el resultado de dicha evaluación se muestra en la siguiente tabla:

Valor Real del Intervalo	Valor Aparente del Intervalo	Frecuencia	Frecuencia acumulada	Porcentaje	Porcentaje acumulado
66,0<x<=68,0	66 – 68	5	5	6,66	6,66
69,0<x<=71,0	69 – 71	8	13	10,66	17.32
72,0<x<=74,0	72 – 74	10	23	13,33	30,65
75,0<x<=77,0	75 – 77	11	34	14,66	45,31
78,0<x<=80,0	78 – 80	15	49	20	65,31
81,0<x<=83,0	81 – 83	19	68	25,33	90,64
84,0<x<=86,0	84 – 86	7	75	9.33	99,97≃100
		75			

La interpretación de los valores de frecuencia y porcentaje son los mismos que para variables medidas en escalas nominales y ordinales. Lo que se modifica en este caso es la manera de presentar los valores de la variable.

Supongamos que los 75 escolares evaluados hubieran obtenido un valor diferente en la prueba; si en tales circunstancias construyéramos una tabla de frecuencia con cada valor individual, ésta contendría 75 renglones con valor de frecuencia igual a 1. Una tabla de estas características sería inútil. Por ello, siempre que se trabaja con variables métricas es conveniente agrupar los valores en intervalos.

La tabla que estamos presentando contiene intervalos de valor 1, esto es a los fines didácticos pues en tablas reales los intervalos son más amplios. Los valores de la prueba Edfeldt son discretos, es decir, la puntuación de prueba no contiene decimales. Para construir una tabla de frecuencia comenzamos con la primera columna que contiene los valores reales del intervalo, allí se cuentan todas las puntuaciones con valor 66 hasta 68 inclusive. Por tratarse de una variable discreta sabemos que ese intervalo solo contiene lo siguientes valores: 66, 67, 68. Si estuviéramos trabajando con variables de tipo continuas, el intervalo podría contener cualquier valor entre 66,0… y 68,9… (volveremos sobre esto más adelante). La siguiente columna representa el valor aparente del intervalo, que no es otra cosa que el valor real sin los decimales. En este caso es sencillo ver los límites del intervalo por tratarse de una variable discreta. Las columnas que siguen contienen información que ya conocemos.

Cuando se publica una tabla de frecuencia de una variable métrica (sea continua o discreta), se emplea la columna de valor aparente del intervalo puesto que es más fácil de interpretar. Tomemos el tercer intervalo cuyo valor aparente es 72 – 74, su frecuencia es 10 lo cual indica que hay diez escolares que han obtenido un puntaje que está entre el valor 72 y el valor 74.

Veamos ahora algunas precisiones contenidas en la columna valor real del intervalo; existen diez escolares que han obtenido un puntaje 74 o mayor, y menor o igual que 76, esto esta expresado por el limite real de intervalo 74,0<x<=76,0, donde x indica el valor obtenido por cualquiera de los diez escolares. Cuando los valores de una variable métrica se agrupan en intervalos, no es posible identificar el verdadero valor individual. Dicho en otras palabras, al agrupar los valores de la variable en intervalos se logra reducir la información para que pueda caber en una tabla que sea fácilmente manejable, el costo de ello es que se debe sacrificar ciertas precisiones en los valores individuales. Por lo tanto, la cantidad de intervalos y su amplitud, debe ser tal que permita un óptimo manejo de los datos en la tabla.

En el siguiente cuadro se muestran algunas referencias para la prueba, los cuales sirven en general para interpretar el valor obtenido por un individuo particular. Por ejemplo, si un niño obtiene un puntaje igual o menor a 60, indica que sus aptitudes visoespaciales para la lectura aún no se encuentran maduras. En el otro extremo, un escolar que obtuviera un puntaje igual o mayor a 75, mostraría que su capacidad visoespacial se encuentra madura para iniciar el aprendizaje de la lectura.

Puntajes de prueba	Interpretación
60 o menos	El niño no posee aún las aptitudes suficientes para el aprendizaje del alfabeto
74 o menos	El niño posee aptitudes limitadas para el aprendizaje del alfabeto
75 o más	El niño posee aptitudes suficientes para el aprendizaje del alfabeto

Nota. Las interpretaciones ofrecidas son una simplificación de los estándares de la prueba

Según muestra la tabla de frecuencia, ningún escolar obtiene un puntaje menor de 60, lo cual indica que todos poseen aptitudes para el aprendizaje de alfabeto. Luego, tenemos que 23 escolares obtienen puntajes entre 66 y 74, valor que resulta del conteo de los tres primeros intervalos de la variable, y que se verifica en la columna de frecuencia acumulada. Podemos decir, según los valores de referencia de la prueba, que un 30,65% de la muestra de escolares tiene aptitudes limitadas para el aprendizaje del alfabeto, por lo cual hay que trabajar en este grupo reforzando el aprendizaje visual de los grafemas. El 69,35% de los niños posee aptitudes suficientes para el aprendizaje del alfabeto. Con esos datos concluimos que es posible focalizar diferencialmente el proceso de enseñanza para aquellos niños que aún no han desarrollado suficientemente sus capacidades visoespaciales.

Veamos ahora un ejemplo con una variable métrica continua y las diferencias en los intervalos de confianza. En una escuela se les tomó un examen a los alumnos y se los calificó con una nota, también se contabilizó el tiempo que emplearon para realizarla. En la siguiente tabla se muestra el tiempo empleado por los alumnos agrupados en intervalos. Una aclaración para esta tabla es que el tiempo está contado en minutos.

Valor Real del Intervalo	**Valor Aparente del Intervalo**	**Frecuencia**	**Frecuencia acumulada**	**Porcentaje**	**Porcentaje acumulado**
60,0<x<=65,59	60 – 65	2	2	2,78	2,78
66,0<x<=71,59	66 – 71	12	14	16,67	19,45
72,0<x<=77,59	72 – 77	16	30	22,22	41,67
78,0<x<=83,59	78 – 83	21	51	29,17	70,84
84,0<x<=89,59	84 – 89	14	65	19,44	90,28
90,0<x<=95,59	90 – 95	7	72	9,72	100
		72		100	

La interpretación que podemos ofrecer de estos datos sigue la misma lógica que la aplicada al ejemplo anterior. Aquí nos detendremos en la diferencia entre valor real de intervalo y valor aparente. Si tomamos el valor del primer intervalo vemos que dos individuos emplearon entre 60 y poco más de 66 minutos en realizar la prueba. El valor exacto del tiempo de prueba de estos dos individuos lo desconocemos, pues es información que se pierde al agrupar en intervalos. Luego sabemos que 12 escolares tardaron entre 66 y poco más de 71 minutos en finalizar la prueba.

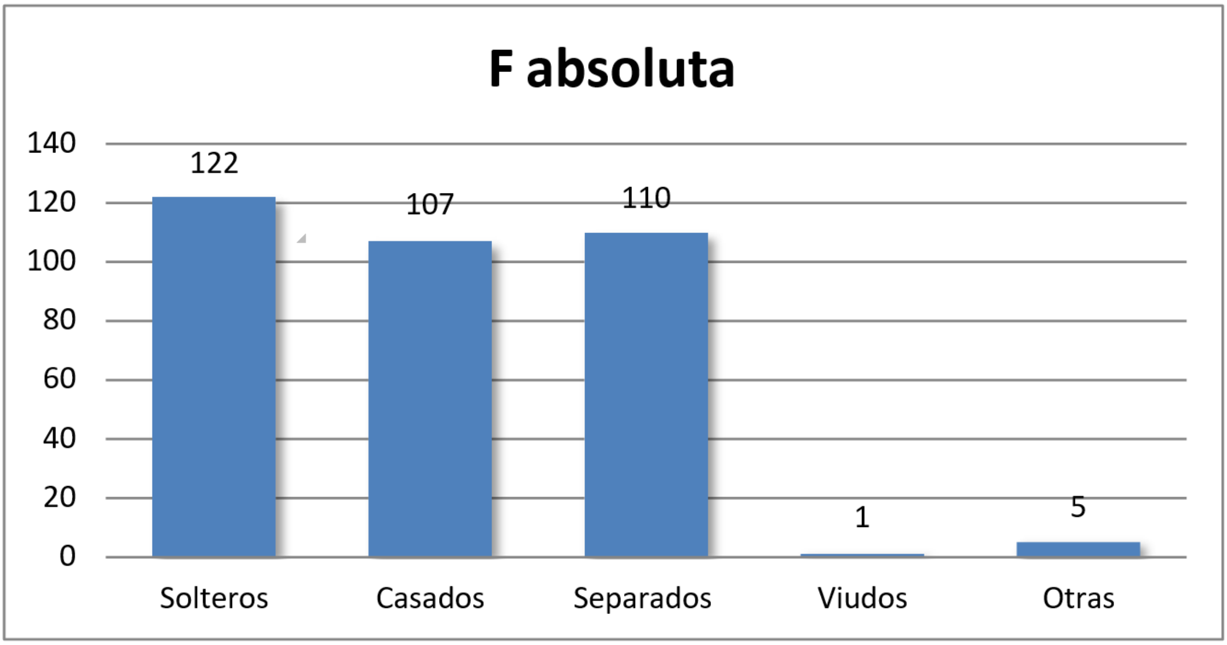
Los intervalos primero y segundo en su valor real se representan como 60,0<x<=65,59 y 66,0<x<=71,59 lo que indica que el primero abarca los tiempos de 60 minutos hasta 65 minutos, 59 segundos. Dado que la tabla se construyó con el tiempo en minutos, transcurridos los 65 minutos con 59 segundos, pasamos al siguiente intervalo que comienza en 66 minutos y culmina en 71 minutos y 59 segundos. Así continuamos para cada uno de los intervalos. Este modo de presentar la información es muy preciso pero engorroso, por lo cual simplificamos mucho la escritura de una conclusión si solo empleamos el valor aparente del intervalo, que sería 60 – 65 y 66 – 71 para el primer y segundo intervalo respectivamente. De este modo, podríamos redactar una conclusión diciendo que dos alumnos fueron los que emplearon el menor tiempo para realizar la prueba, que estuvo entre 60 y 65 minutos. Quienes más demoraron en realizarla (siete alumnos) tardaron entre 90 y 95 minutos. La mayoría empleó un tiempo de 78 a 83 minutos (21 alumnos).

Representación gráfica de las tablas de frecuencias

En las páginas que siguen se mostrarán algunos de los gráficos más usados para representar los datos contenidos en las tablas de frecuencias. A modo de resumen, se usarán los mismos ejemplos dados anteriormente.

Diagrama de barras

El siguiente diagrama representa la primera columna de la tabla que pertenece a la frecuencia absoluta de la variable estado civil de los docentes.

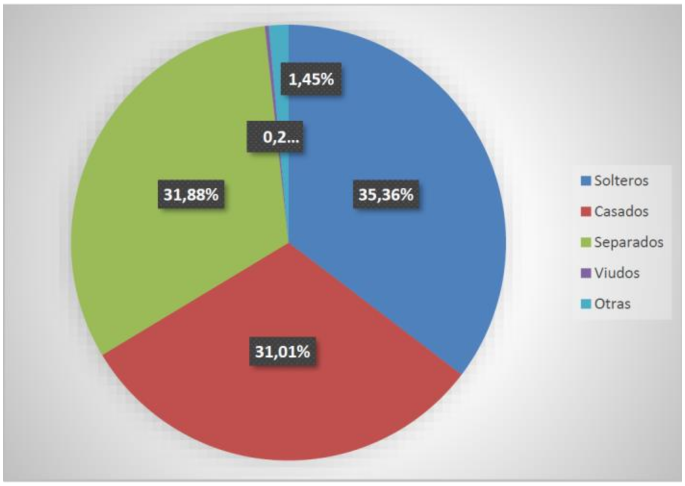


Los diagramas de este tipo se representan sobre un plano delimitado por dos vectores llamados ejes cartesianos. Uno de esos vectores es conocido como eje *x* o abscisa, y es la horizontal del plano; el eje *y*, conocido como ordenada, es el eje vertical del plano. En todo gráfico de esta naturaleza, los ejes cartesianos deben estar debidamente señalados. En este caso se observa que las categorías de las variables (solteros, casados, etc.) se han puesto sobre el eje *x*. El eje de las *y* contiene los valores de frecuencia absoluta. De este modo, la altura de la barra en un punto dado, representa la cantidad de casos que se han contado en la categoría.

Este arreglo visual de la tabla de frecuencia, llamado diagrama de barras, es particularmente útil para visualizar la distribución de frecuencia entre las categorías. Dado que se trata de una variable de tipo nominal, las barras están separadas ente si, denotando que no existe continuidad entre las categorías. Se observa entonces que la población analizada consta mayoritariamente de individuos solteros, casados y separados.

Diagrama de sectores

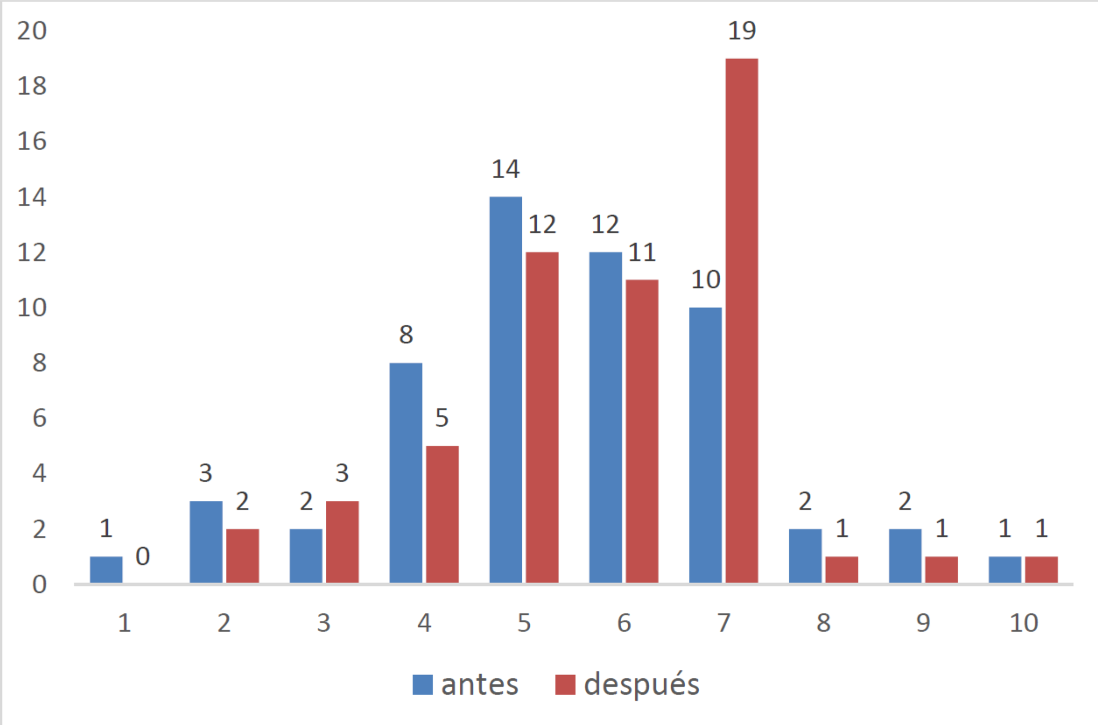
El diagrama de sectores o gráfico de tortas es una alternativa al diagrama de barras, cuando se utilizan porcentajes. En este caso, el área total de una circunferencia, o bien los 360 ° de giro sobre la misma, representa el 100% de los casos. Los porcentajes propios de cada una de las categorías de la variable, se equiparan al tamaño de una porción o sector del gráfico. El gráfico correspondiente a la columna porcentaje de la variable estado civil, se presenta de la siguiente forma:



Nótese que el diagrama de sectores prescinde de los ejes cartesianos, dado que la información principal que se usa para el análisis es el tamaño relativo de cada porción de la circunferencia, la cual expresa el porcentaje correspondiente a la categoría de la variable.

Diagrama de barras agrupadas

Este tipo de gráfico tiene las mismas propiedades que el gráfico de barras, pero en él es posible agrupar barras que representen distintas mediciones de una misma variable. El siguiente gráfico ejemplifica los datos de la tabla donde se midió mediante una escala ordinal, el nivel de cohesión del grupo, antes y después de haber aplicado una serie de actividades que supuestamente la favorecían. Nótese que el eje vertical o eje *y*, contiene la frecuencia absoluta, pero ahora el eje horizontal o eje de las *x*, contiene las categorías de la variable ordenada. La medida en que las actividades propuestas han contribuido a aumentar la cohesión grupal, se debe interpretar a partir de las diferencias encontradas en la altura de las barras, en las dos condiciones en que se realizó la intervención: antes – después. Conviene señalar que cada vez que se realizan comparaciones, en mediciones nominales u ordinales, se utiliza preferentemente este tipo de gráficos.



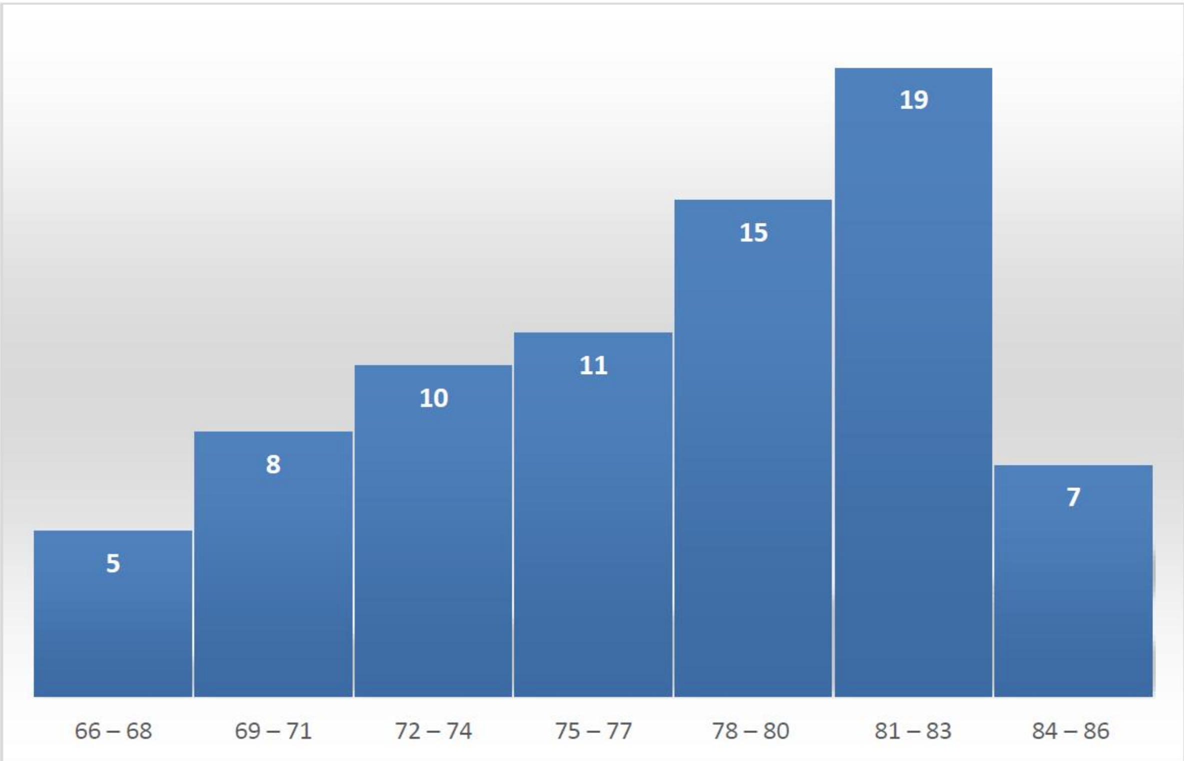
Según se esbozó en la interpretación dada para la tabla de frecuencias, se aprecia que las actividades tendientes a favorecer la cohesión grupal, produjeron como resultado una acumulación de casos en la categoría 7 de la escala, sin modificar sustantivamente la tendencia general de la distribución de frecuencias.

Histograma

Los datos que se muestran en la tabla que corresponden a una variable de tipo métrica, se representan con gráficos similares a los anteriores (diagrama de barras), pero en los que las columnas son adyacentes entre sí. Ello evidencia que los intervalos de la variable son categorías de un conjunto de datos continuo o discreto. Generalmente, una primera indicación del tipo de datos que se está presentando, viene dado por el tipo de gráfico que se utiliza, el histograma y otros gráficos similares indican claramente el uso de variables métricas. El ancho de las barras de un histograma y su cantidad dependen de la amplitud del intervalo utilizado. La cantidad de intervalos usados para la representación gráfica deben ser suficientes para que el histograma refleje la tendencia general de la distribución de frecuencias.

En este sentido, si se usan intervalos muy amplios el recorrido de los valores de la variable queda contenido en muy pocos intervalos, y el histograma resultante tendrá pocas barras, lo cual puede dificultar la observación de la forma de la distribución de frecuencias. Lo mismo ocurre si el intervalo es estrecho y quedan varios de ellos con frecuencia igual a cero. El histograma para los puntajes de la prueba de madurez lectora, utiliza un intervalo estrecho, pero suficiente para graficar la tendencia general de los datos dado que contiene solo un intervalo con frecuencia cero al comienzo de la distribución.

Prueba de madurez visoespacial: Histograma:

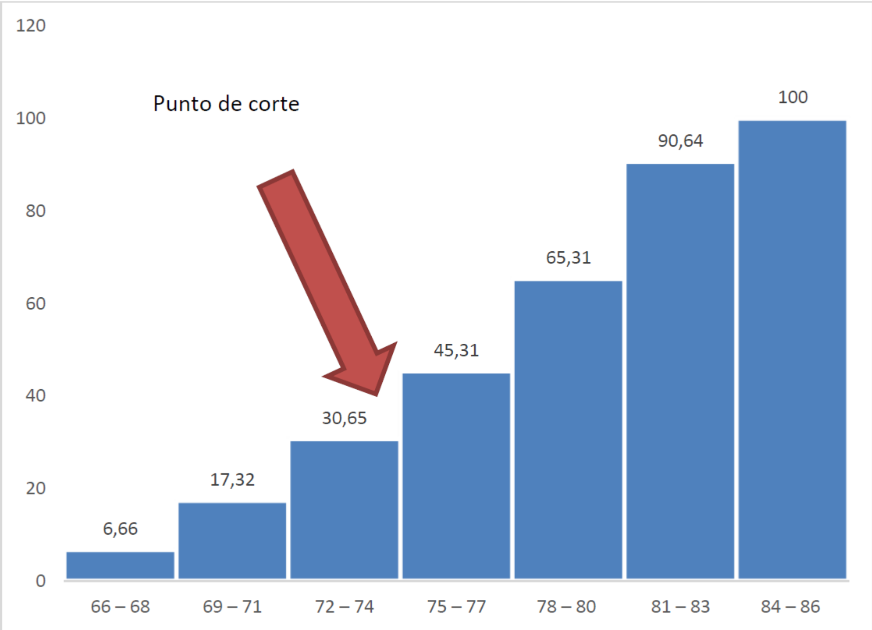


Aquí también se tiene que el eje de las *y* (vertical), representa la frecuencia absoluta, que en este caso se consignan en cada barra del gráfico; el eje de las *x* (horizontal), representa los intervalos de los valores en los que se agruparon las ediciones originales. Por lo tanto, al igual que en el diagrama de barras, la altura de las mismas corresponde a la frecuencia observada en ese intervalo.

Dado que esta es una variable de tipo métrica, es posible realizar un histograma con los valores de frecuencia absoluta o porcentual acumulados. En ambos casos, el histograma representa la progresiva acumulación de casos a través de las categorías de la variable, y ello tiene importancia cuando es posible establecer algún punto de corte teórico para la variable que se está analizando. Recuérdese que, en el ejemplo de aplicación de la prueba de madurez lectora, se trazaron límites en los puntajes para favorecer una interpretación cualitativa de la misma, que se reproducen en el siguiente cuadro:

Puntajes de prueba	Interpretación
60 o menos	El niño no posee aún las aptitudes suficientes para el aprendizaje del alfabeto
74 o menos	El niño posee aptitudes limitadas para el aprendizaje del
75 o más	El niño posee aptitudes suficientes para el aprendizaje del

Estos valores teóricos pueden insertarse en el gráfico para subrayar su importancia al momento de la interpretación.



Siendo un puntaje de 74 el valor por encima del cual se considera que el niño ha desarrollado adecuadamente las habilidades visoespaciales para la lectura, el histograma de porcentajes acumulados muestra que por debajo de ese valor se encuentran el algo más del 30 % de la muestra (23 niños).

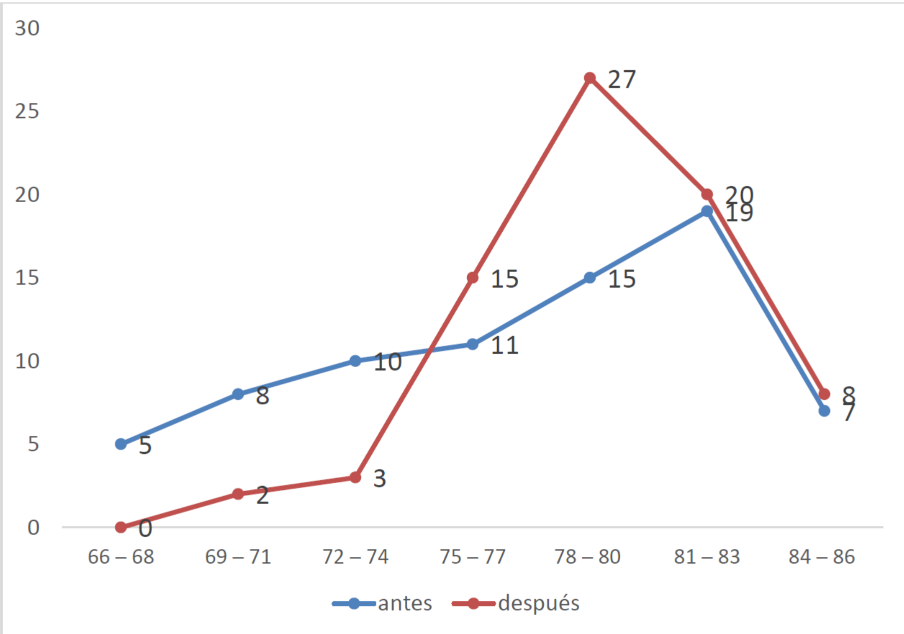
Es frecuente que, en algunos manuales de estadística descriptiva, se presente el histograma de frecuencias acumulares como Grafica de Pareto, en referencia a Vilfredo Pareto. La ventaja de estos gráficos es que transcribe visualmente la columna de frecuencias acumuladas (absolutas, relativas o porcentajes), y hace más intuitiva la progresión acumulada.

Polígonos de frecuencia

Anteriormente vimos que podía usarse el diagrama de barras agrupado para hacer comparaciones. Con los histogramas una comparación de ese tipo se hace dificultosa cuando se tiene varios grupos. Sin embargo, la altura de la barra puede reemplazarse por un punto (recuérdese que la altura representa el valor de frecuencia de la categoría), y luego los puntos obtenidos pueden unirse entre sí para generar un polígono. Al reemplazar el histograma por el polígono, es posible comparar la distribución de frecuencia de dos o más grupos en una misma variable métrica.

Para este caso, ampliaremos el ejemplo anterior sobre la prueba de madurez lectora. Supongamos que la muestra de escolares a las que se les aplicó la prueba de madurez visoespacial, realizó actividades tendientes a estimular el reconocimiento de las letras del alfabeto. Los maestros deciden reevaluar a los niños tres meses después y verificar si hubieron cambios en los puntajes de prueba. Explícitamente, se espera que aquellos niños con menor puntajes alcancen una puntuación mayor, pero no se anticipan cambios en aquellos escolares que ya poseían la madurez visoespacial suficiente. Un resultado de este hipotético experimento se muestra en el siguiente polígono de frecuencia. Nos vamos a centrar en el puntaje 74 que es el punto de corte que separa los escolares que poseen aptitud suficiente de aquellos que tienen aptitudes limitadas.

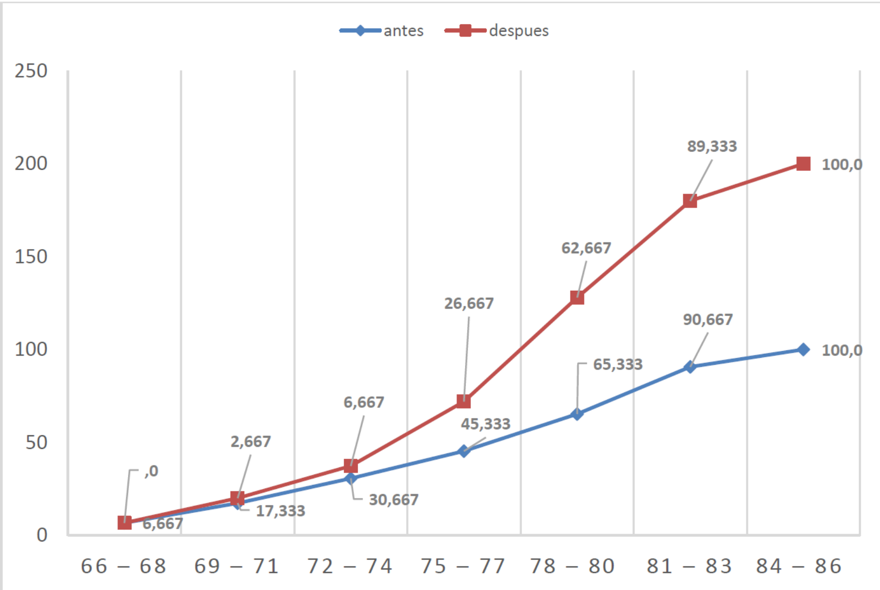
Polígono de frecuencia: prueba de madurez visoespacial para la lectura



Lo primero que apreciamos en estos datos es que, tras la intervención de los docentes, no se verifican casos con puntajes en el rango 66 – 68. La cantidad de casos con puntajes iguales o menores a 74 es ahora de 5 escolares (es la cantidad de alumnos en los intervalos anteriores a 75-77), cuando antes era de 23. Con esos sencillos datos es posible afirmar que la intervención realizada ha resultado beneficiosa para aquellos escolares con aptitudes visoespaciales menos desarrolladas al comienzo del aprendizaje de la lectura. Como era de esperar, aquellos alumnos que alcanzaron los mayores puntajes de prueba antes de la intervención, no modificaron sustantivamente su posición en la prueba tres meses después. Esto era de esperar debido a que las actividades desarrolladas no se proponían estimular a aquellos con las aptitudes visoespaciales suficientes, sino a aquellos con menores niveles de esta habilidad.

Gráfico de ojiva

El gráfico que se presenta a continuación corresponde a los valores observados de la variable puntajes obtenidos en una prueba de lectura. En él, cada punto representa el valor observado de la variable y la frecuencia que se acumula con cada valor. Este tipo de gráfico es utilizado para representar el total acumulado a través de los valores ascendentes de la variable.



En esta gráfica vemos el porcentaje acumulado de casos según los puntajes en la prueba de madurez visoespacial. Se puede ver claramente que el gráfico de ojiva es una variante del polígono de frecuencia. En la situación “antes”, cuando no se ha realizado ninguna intervención, vemos que la ojiva tiene una pendiente menos pronunciada, comparada con la situación “después”, momento en el que se realizó la intervención con actividades para estimular el aprestamiento visoespacial. La diferencia entre ambas curvas u ojivas se debe a la celeridad con la que se acumulan los casos. Cuantos más casos se acumulan en cada intervalo, más pronunciada la pendiente, la distancia entre ambas curvas marca la diferencia entre lo ocurrido antes de la intervención y posterior a la misma.

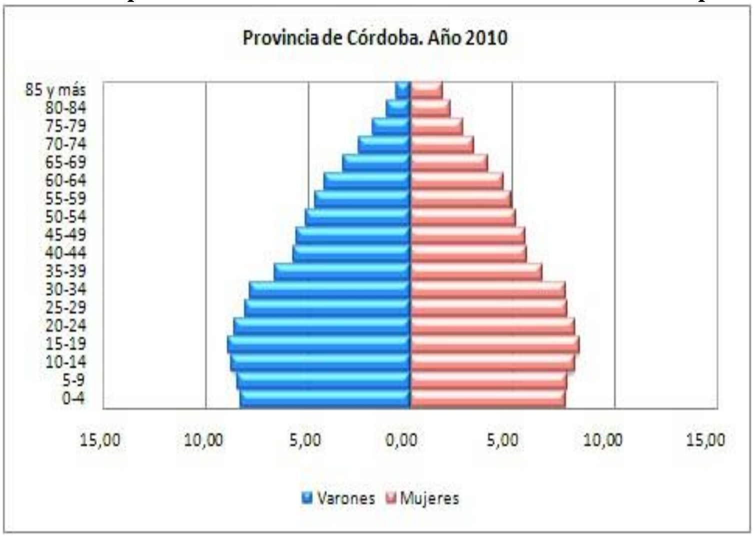
Otros tipos de gráficas

Los gráficos mostrados hasta aquí se corresponden con el tipo de variables que hemos analizado, pero en estadística existen muchos tipos de gráficas diferentes dependiendo del tipo de información que se quiera transmitir, del énfasis que se quiera dar a ciertos aspectos de los datos, si se utilizará con fines científicos o de divulgación. Muchas publicaciones especializadas requieren que los autores utilicen ciertos estilos y formatos en la presentación de datos. Por lo dicho, en los apartados siguientes mostraremos algunos gráficos frecuentemente utilizados en reportes estadísticos.

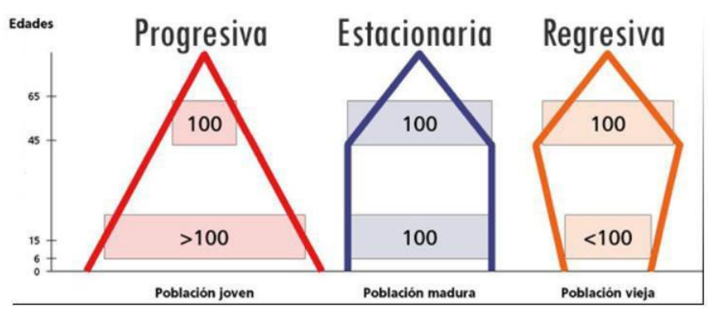
Pirámides de población

La pirámide de población es una gráfica muy utilizada en epidemiología y en estudios demográficos. Se trata de la superposición de dos histogramas, uno que corresponde a varones y el otro a mujeres. En el eje de ordenadas se representan por intervalos las edades de los individuos, y en el eje de la abscisa el porcentaje de población en ese rango de edad. Dado que la cantidad de población disminuye con la edad, lo común es ver este gráfico en forma de pirámide. Cuando esto sucede, el exponente de crecimiento poblacional es positivo, dado que hay más nacimientos que defunciones, y la población joven sobrepasa a la añosa. Pero suele suceder que, en algunas regiones o países, esa pirámide se invierte y el pico es mayor que la base; en tal caso el exponente de crecimiento poblacional es negativo. El gráfico que sigue corresponde a la población de la Provincia de Córdoba; en él se aprecia que el exponente de crecimiento poblacional es positivo por tanto la pirámide es más ancha en su base. En el eje de las *x* la referencia está dada por la cantidad de varones y mujeres en miles de personas. En el eje de las y se hace referencia a la franja de edad correspondiente. Bajo condiciones de crecimiento normal vegetativo de la población, la distribución de la población de varones y mujeres tiende a ser simétrica. Un dato relevante es la pérdida de simetría del gráfico a partir de la franja etaria de 65 a 69 años, que mostraría que las mujeres tienden a ser más longevas que los varones.

Los estudios de poblaciones diferencian al menos tres tipos de pirámides

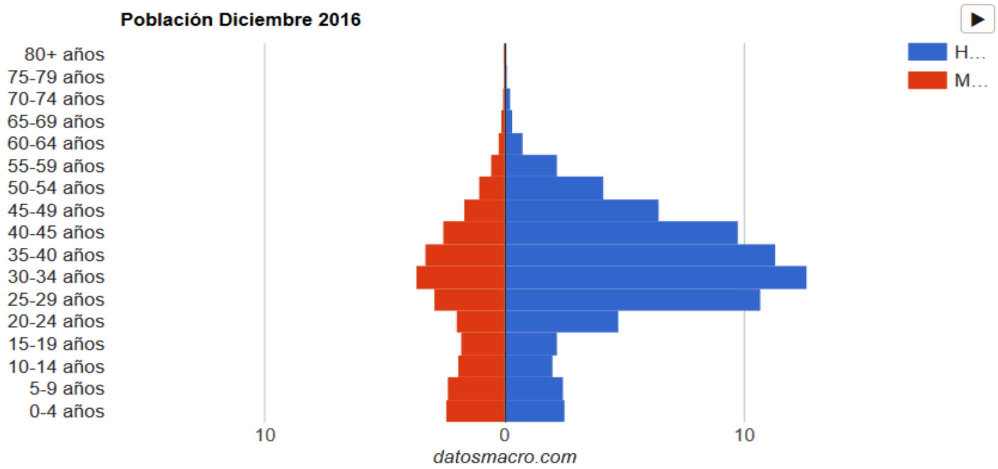


de población. Un tipo de pirámide llamada progresiva, tiene base ancha y se va reduciendo paulatinamente hacia la cúspide. Es consecuencia de una alta tasa de natalidad y con población mayormente joven. Muchos países latinoamericanos presentan este tipo de crecimiento poblacional. También existen las pirámides estacionarias donde se manifiesta que la tasa de nacimientos es menor y está equilibrada con la población joven, esto da cuenta que la natalidad y mortalidad se han mantenido sin variaciones significativas durante un periodo de tiempo largo. Esta pirámide se considera el paso intermedio entre la pirámide progresiva y la regresiva. La pirámide regresiva es más ancha en los grupos superiores que en la base, debido al descenso en la natalidad y al envejecimiento de su población. A medida que descienden los nacimientos y la población se hace añosa, la pirámide toma un aspecto invertido. Actualmente, en varios países europeos se verifica este tipo de crecimiento poblacional. En el siguiente gráfico podemos apreciar las diferencias mencionadas.



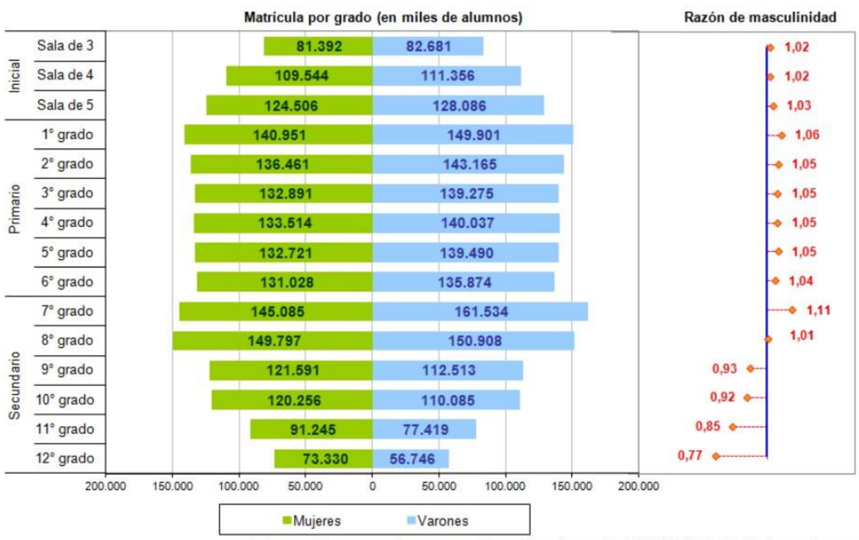
Cabe destacar que muchos fenómenos sociales pueden intervenir para que la forma esperada de la distribución de la población pierda su armonía entre varones y mujeres. Durante el siglo XX las dos guerras mundiales alteraron la relación entre varones y mujeres en las pirámides de población europeas. Actualmente, son las migraciones en masa las que están produciendo efectos de distorsión en el crecimiento poblacional. En el gráfico que se muestra más abajo se aprecia una cantidad mucho mayor de hombres jóvenes con relación a mujeres en los Emiratos Árabes Unidos, especialmente en su capital Abu Dabi. El gráfico que se muestra a continuación es del total del país y la causa del crecimiento de población masculina se debe a la migración para los trabajos de construcción en la capital del país.

Pirámide de población para Emiratos Árabes Unidos



Las pirámides de población también pueden utilizarse en educación para mostrar la cantidad de estudiantes por nivel según el sexo. En la siguiente gráfica se muestra la tasa neta de escolarización de la provincia de Buenos Aires para el año 2010. Un dato que suele incluirse en estas gráficas es la razón de masculinidad, la cual resulta de dividir la cantidad de varones por mujeres. Si se hubiera realizado la operación inversa se tendría la razón de femineidad. Los que se observa en este gráfico es que, para el año reportado, la razón de masculinidad se invierte en el tramo final del trayecto educativo. La razón es simplemente que para esa franja etaria hay más mujeres que varones.

Matrícula por edad simple y condición de edad, por sexo. Educación común. Año 2010. En miles de alumnos

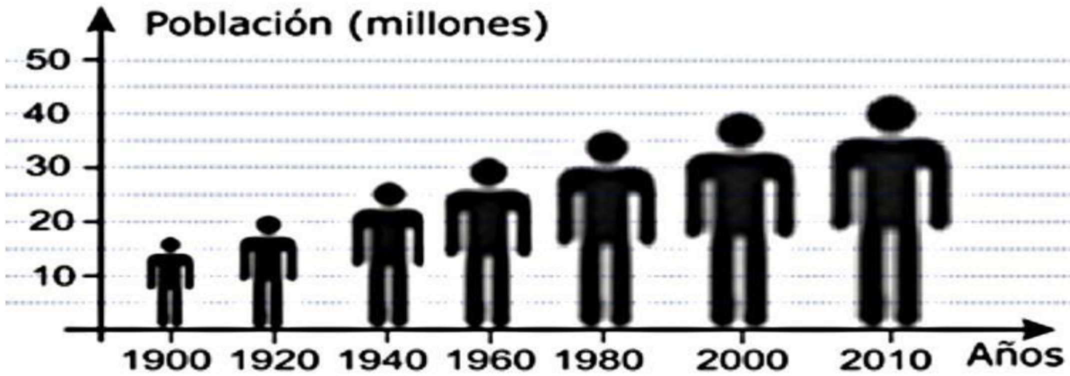


Fuente: Procesamientos propios sobre datos de DiNIECE-ME. Relevamiento Anual de Matrícula y Cargos. Anuario estadístico 2010

Pictogramas

El pictograma no es otra cosa que un diagrama de barras, que incorpora figuras que simbolizan a la variable en estudio. No suele presentar mayor información que el diagrama de barras, pero en ocasiones es más sencillo de interpretar por su recurso iconográfico. Nótese cómo en el gráfico que presentamos a continuación, aparece claramente expresado el crecimiento de la población argentina mediante el uso de una figura humana en reemplazo de una barra. Una cuestión que siempre debe tenerse en cuenta a la hora de interpretar este tipo de gráfica, es la cantidad de individuos (o unidades de análisis) que se representan en cada pictograma.

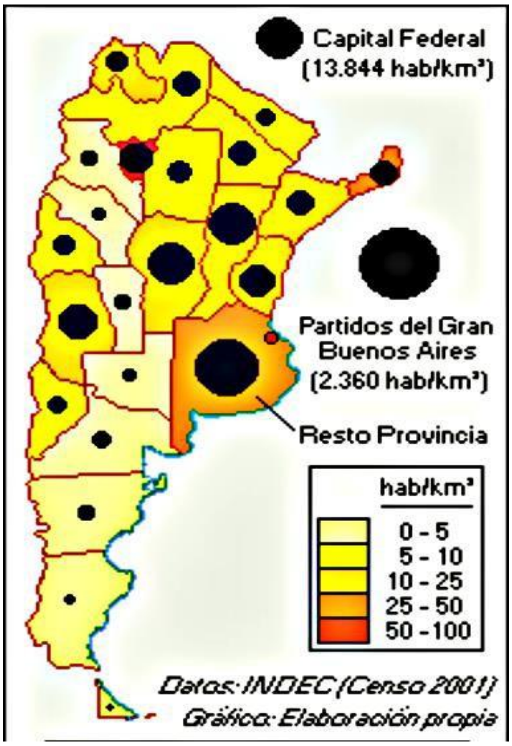
En este caso cada figura representa a millones de persona



Nota. En el año 2010 en argentina se contabilizaron 40.091.359 personas

Mapas

Como se habrá notado en el pictograma, la representación gráfica es aproximativa y sus valores no están rigurosamente escalados. Cuando se presenta esta dificultad y se requiere precisar información, es posible utilizar varios recursos simultáneos, tal como se muestra en el siguiente gráfico.



Este gráfico utiliza tres recursos: a) el mapa del territorio argentino, b) puntos cuyo tamaño representa la cantidad de población concentrada en cada provincia y c) una paleta de colores que representa la densidad poblacional. Combinando los tres recursos es posible ver que la provincia más densamente poblada es Buenos Aires, pero la que concentra mayor cantidad de habitantes por Km2 es la provincia de Tucumán.

Gráficas Polares

La gráfica polar fue una ingeniosa idea para agregar información al diagrama de sectores, ideada por la enfermera Florence Nightingale durante la guerra de Crimea (1853-1856). En su trabajo en los hospitales de campaña, Florence Nightingale recopiló datos que demostraban que las heridas de guerra eran una fracción menor de muerte de los combatientes, en comparación con enfermedades como el tifus, el cólera y la disentería. Estas últimas eran las principales causas de muerte, debidas principalmente a las deficientes condiciones de higiene hospitalaria. El desafío que enfrentó Florence Nightingale, fue mostrar de manera convincente a las autoridades, la necesidad de cambiar de manera drástica y urgente las condiciones de prácticas quirúrgicas y médicas. El diagrama que se muestra a continuación fue el que cambio el curso de las atenciones recibidas por los soldados durante la guerra.

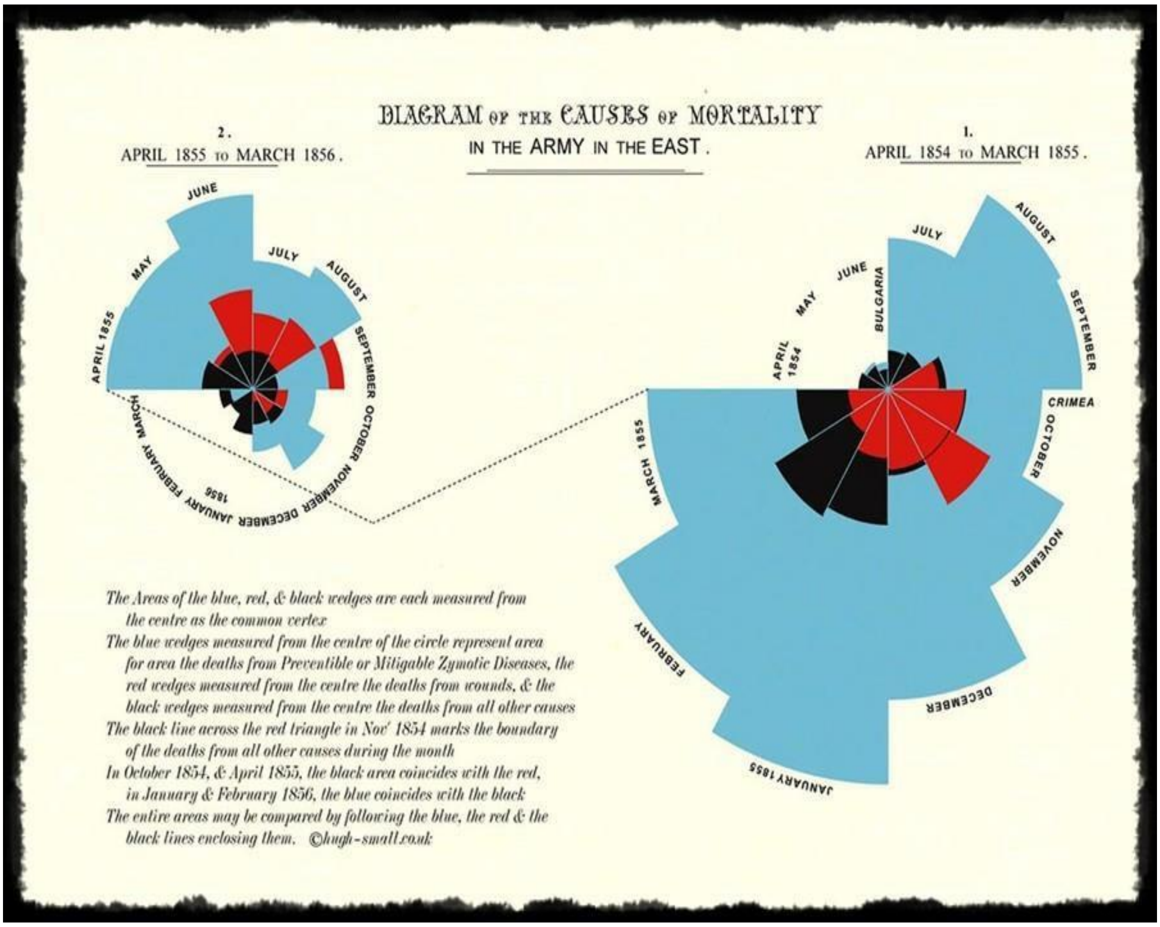
Interesa destacar en este caso que cada sector del gráfico se corresponde a un mes. Luego, los colores indican las causas de muerte de los soldados:

Azul: Muertes por enfermedades infecciosas, desde prevenibles hasta mitigables

Rojo: Muertes por heridas de guerra

Negras: Muertes por otras causas

Nótese cómo el sector azul crece mucho más que el sector rojo alcanzando su pico en enero de 1855; en julio de 1854 ya es posible ver que las enfermedades aventajan en causa de muerte a las heridas en combate. La información presentada por Nightingale fue decisiva para promover el cambio en las prácticas hospitalarias, y su enfoque estadístico convenció a las autoridades militares, al parlamento y a la reina Victoria, para llevar a cabo la reforma. Gracias a ella se mejoraron las condiciones de sanidad y se consiguió reducir la proporción de muerte de sus pacientes. En febrero de 1855 la tasa de mortalidad había descendido del 60% al 42,7% y en primavera ya era del 2,2%.

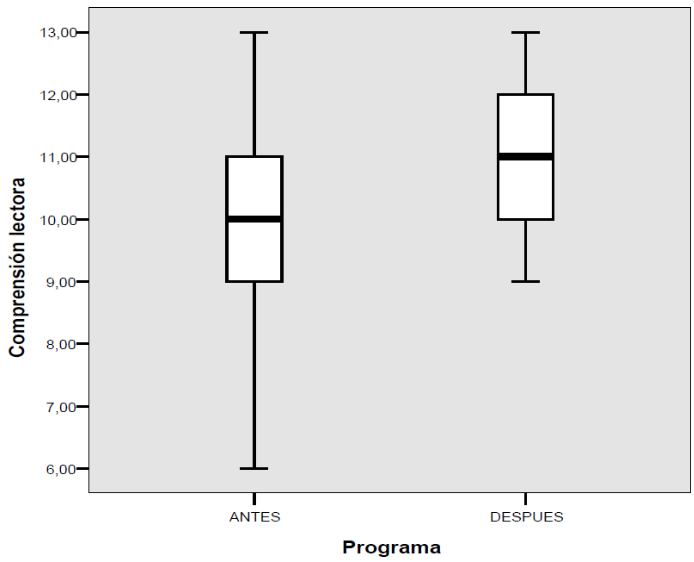


Box Plot o Diagrama de Cajas

El box-plot o diagrama de cajas es una sencilla representación de la dispersión de datos para una variable que permite comparar fácilmente el comportamiento de dicha variable en uno o más grupos. El diagrama exhibe como puntos de referencia un rectángulo (caja), intersectado por una línea. Generalmente, estos puntos representan la posición de los cuartiles de la distribución de datos (eventualmente, la línea media en la caja puede representar el promedio del grupo). De cada extremo de la caja se extienden líneas perpendiculares a la misma y representan en cada caso, la distribución de datos hacia el máximo y el mínimo valor registrado en la distribución de valores de la variable. La interpretación de un diagrama de cajas depende de la posición de los puntos de referencia del diagrama respecto de la distribución de datos observada, es por ello que el diagrama siempre se representa en el marco de la intersección de ejes cartesianos.

Diagrama de cajas y medidas de posición: cuartiles

Para comprender mejor la utilidad de este tipo de diagramas, proponemos el siguiente ejemplo: Una investigadora desarrollo un programa de lectura compartida para aplicar en el aula. Según ella el programa estimula la lectura y mejora la calidad de la comprensión. Para verificar la eficacia del programa, seleccionó 196 escolares, en quienes evaluó la comprensión lectora antes de la aplicación del programa, y después de dos meses de aplicación del mismo. La comprensión lectora se evaluó mediante una prueba estandarizada, con valores que van de 0 a 15, siendo este último el máximo puntaje de la prueba. En el siguiente gráfico se ofrece un diagrama de caja con los resultados obtenidos.



Para comprender los diagramas de cajas primeramente se debe atender a la referencia que proporcionan los ejes cartesianos. En este caso el eje horizontal indica la medición de la variable en dos situaciones diferentes: antes de aplicar el programa y después de aplicado el mismo. El eje vertical es la referencia para la distribución de valores en la variable comprensión lectora. Así, es posible observar que se ha producido un cambio en la variable una vez aplicado el programa, dado que la posición de ambas cajas es diferente.

Una primera interpretación comienza con la posición que ocupa el valor máximo y mínimo en cada distribución; de este modo se observa una diferencia en el valor mínimo de cada distribución. Tal diferencia es de tres puntos en la variable, e indica un incremento del valor mínimo luego de la aplicación del programa. Por otro lado, se observa que el valor máximo no se ha modificado.

La longitud de la línea inferior que alcanza el rectángulo recibe el nombre de bigote (al igual que la línea superior). Indican la posición en los valores de la variable para ¼ de los casos. En otras palabras, la longitud del bigote indica qué valores de la variable han obtenido aquellos casos que se encuentra entre el mínimo valor registrado y el primer cuartil. Como se aprecia la distribución es diferente antes y después de aplicar el programa; antes de aplicarlo ¼ de los casos obtuvo valores de 6 a 9 en comprensión lectora, luego de aplicado el programa ¼ de la distribución se concentra en los valores 9 y 10. Por tanto se registra un incremento en la comprensión lectora de aquellos individuos que antes habían mostrado el menor rendimiento.

La parte inferior del rectángulo y la línea media que lo divide indica la distancia entre el primer cuartil y el segundo cuartil o mediana. La distancia entre ambos indica los valores obtenidos en la variable por ¼ de la distribución. Nótese que la cantidad de casos acumulados entre el mínimo y la mediana corresponde a la mitad de los casos de la muestra. En el diagrama, se observa que la dispersión en ambos grupos, entre el primer y segundo cuartil, es la misma, sin embargo, los valores de la variable son superiores luego de la aplicación del programa. Dicho en otras palabras, antes de la aplicación del programa ¼ de la distribución de casos obtuvo valores ente 9 y 10; luego de aplicar el programa ¼ de la distribución alcanza valores entre 10 y 11. La situación es similar para todos los casos que se encuentran entre el segundo y tercer cuartil (que corresponde también a ¼ de los casos).

Finalmente, y dado que no se ha modificado el máximo valor de la variable, se observa que luego de la aplicación del programa la dispersión es menor. En otras palabras, antes de la aplicación del programa el cuarto de los casos con mayor rendimiento en comprensión lectora alcanzaba puntaje de prueba entre 11 y 13; luego de la aplicación este grupo alcanza valores concentrados en 12 y 13. En conclusión, la comparación mediante el diagrama de cajas de la situación antes y después de la aplicación del programa, muestra que este resultó efectivo en la muestra analizada, para mejorar el nivel de comprensión lectora.

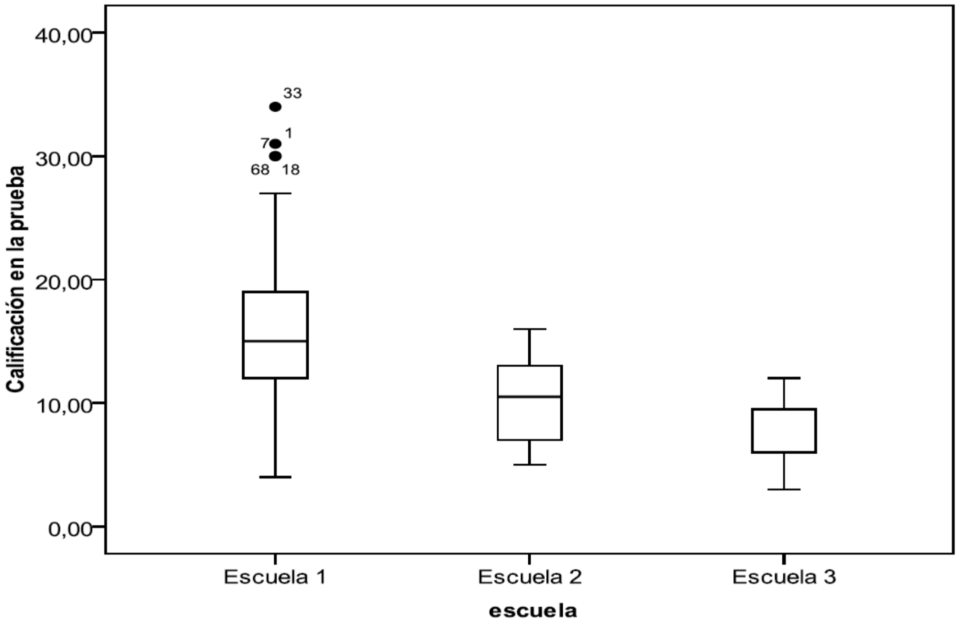
Diagrama de cajas – valores atípicos o extremos

Para determinar que una distribución de datos contiene valores extremos, se toma en consideración una medida proporcional del rango intercuartilar, mediante la cual se determina la extensión de los máximos y mínimos (bigotes de la caja). El rango intercuartilar se puede calcular de diferentes maneras, aquí consideraremos las dos más usadas:

a) Rango intercuartilar (RI): Q3-Q1

b) Rango semiintercuartilar (RSI): Q3-Q1/2

Para desarrollar este tema proponemos el siguiente ejemplo: aplicando una prueba estandarizada para evaluar razonamiento numérico, se compararon los puntajes obtenidos por tres muestras de escolares del segundo ciclo de EGB de tres escuelas diferentes. El siguiente diagrama de cajas muestra los resultados obtenidos en las evaluaciones realizadas.



En este ejemplo se tiene las referencias de cada caja en el eje horizontal (escuela), y la distribución de los valores de la variable en el eje vertical (puntaje en la prueba de razonamiento numérico). En lo esencial la interpretación del diagrama es el mismo que hemos desarrollado anteriormente, con la excepción de que los bigotes no se extienden del máximo al mínimo puntaje, sino que representan una proporción del RSI. Eso hace que en la escuela 1 se observen casos atípicos. En esa muestra existen cinco casos atípicos acumulados en el extremo superior de la distribución de la variable.

Considere la posición de Q1 y Q3 en la distribución de valores de la variable razonamiento numérico de la escuela 1 en el gráfico, se tiene que Q1=11 y Q3=19 (los valores son aproximados). Por tanto, se tiene que RI= 8; RSI= 4. Los casos extremos se identifican cuando muestran valores que exceden la longitud de los bigotes, y esta longitud es una proporción de RI o de RSI. Digamos que si tomamos 1.5 del valor de RSI, el recorrido de los bigotes se extiende desde 19 a 25 en el límite superior, y de 5 a 11 en el límite inferior. Estos valores surgen de los siguientes cálculos:

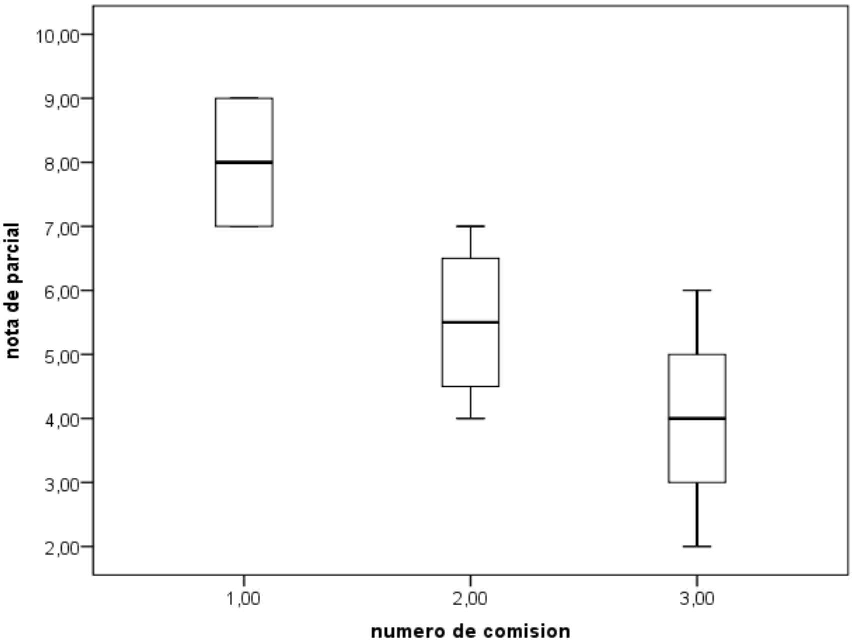
$$1.5xRSI= 1.5x4= 6$$
$$Q3+(1.5xRSI)=19 + 6= 25$$
$$Q1-(1.5xRSI)= 11-6= 5$$

Se aprecia que en una distribución de datos cuyos límites máximos y mínimos son una proyección de un valor teórico (en este ejemplo es 1.5 del valor de RSI), pueden existir valores que exceden ese límite. A tales casos se los denomina atípicos o extremos. En las gráficas estos casos aparecen marcados con puntos y apropiadamente identificados. De este modo, sabemos que los casos 68 y 18 son atípicos puesto que obtuvieron un puntaje de 30; del mismo modo, el caso 7 y 1 obtuvieron un valor de 31 y el caso 33 obtuvo un valor de 35 aproximadamente. Nótese que los casos atípicos aparecen solo en la escuela 1, la cual es la que muestra la mayor dispersión en los puntajes de la prueba de razonamiento numérico.

Otro aspecto interesante presente en el gráfico aparece en la distribución de la escuela 3, en la cual se observa que no hay variación entre Q1 y Q3, lo cual queda reflejado visualmente en que no existe separación entre la parte inferior del rectángulo y la intersección que marca la mediana. En otras palabras, la mediana (Q2) y el primer cuartil (Q1) tienen el mismo valor, indicando que 1/4 de la muestra no evidencia variación en el puntaje de la prueba de razonamiento numérico (todos los casos obtuvieron una calificación de 8).

Diagrama de cajas – restricción de la variabilidad

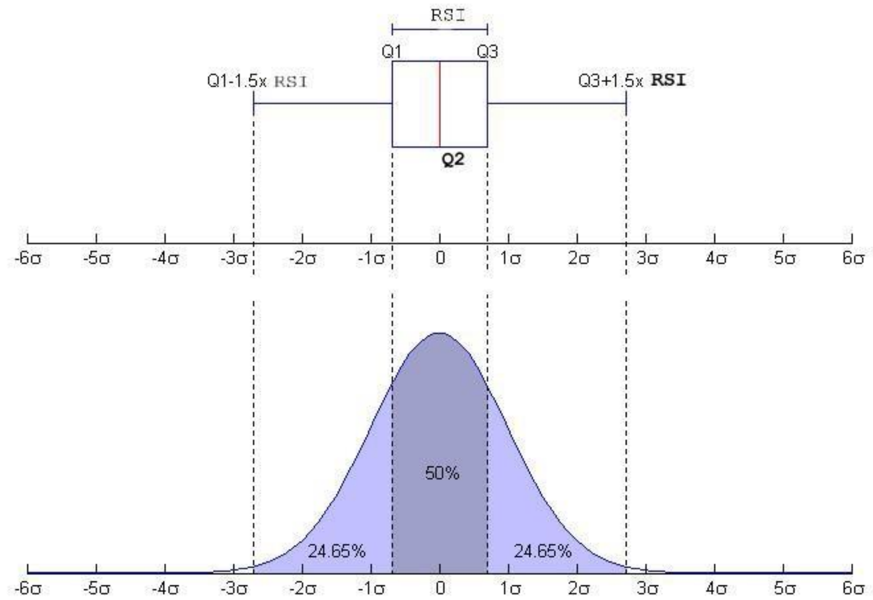
Así cómo es posible identificar casos extremos en diagrama de cajas, también se puede identificar la situación en que los casos se concentran en un valor determinado de la variable. Anteriormente vimos que la restricción de la variabilidad hace que coincida el valor de mediana con el primer cuartil. En el gráfico que sigue, la restricción de la variabilidad se expresa en la concurrencia del valor máximo y mínimo con el tercer y el primer cuartil respectivamente. Veamos el siguiente ejemplo: en la materia Sociología se dividieron a los alumnos en tres comisiones diferentes. Los docentes supusieron que el orden de las comisiones del parcial no alteraría el rendimiento de los estudiantes. Atendiendo a los resultados de los exámenes se elaboró el siguiente diagrama de cajas.



Tal como se observa, el rendimiento de los alumnos muestra un patrón que los diferencia claramente. Sobresalen dos aspectos, el primero es que el rendimiento es menor a medida que se pasa de la comisión 1 a la 3, el segundo es que, siguiendo ese mismo orden, los resultados de los exámenes muestran mayor variabilidad. En la comisión 1 se observa una particularidad que es la restricción en la variabilidad. En esta comisión los alumnos obtuvieron 7, 8 y 9. Más precisamente, el diagrama muestra que un tercio del total de alumnos obtuvo un siete, otro tercio obtuvo un ocho y otro más un nueve, completándose así el total de la muestra de alumnos de esa comisión. En una situación como la descripta, el máximo y el mínimo de la distribución no se diferencian del primer y el tercer cuartil. En otras palabras, no existe variabilidad más allá de los cuartiles.

Diagrama de cajas en distribuciones sesgadas

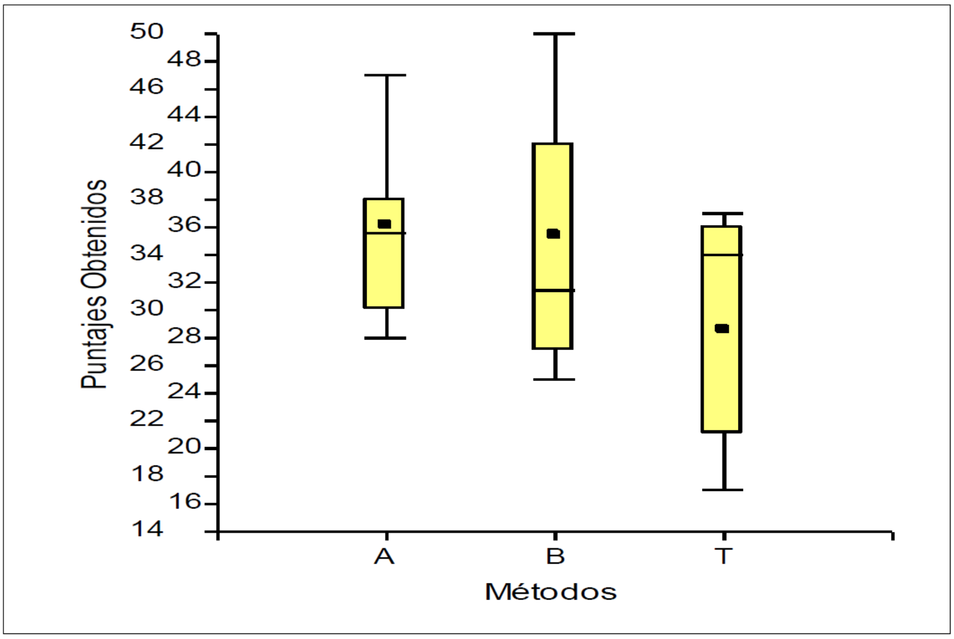
Más adelante veremos una distribución muy utilizada en ciencias sociales que permite describir el comportamiento de muchas variables, se trata de la distribución normal. Una de las principales características de la distribución normal teórica es su simetría, y esta propiedad es importante porque es posible determinar (numéricamente o gráficamente) cuándo existen sesgos en una distribución empírica de datos. Los sesgos son indicadores de que los valores de la variable tienden a acumularse hacia alguno de los extremos de todos los valores posibles de la variable en cuestión. No abundaremos en más detalles pues profundizaremos este tema en el capítulo dedicado a distribución normal. En este apartado destacamos que existe una relación entre la forma de una distribución normal teórica y el diagrama de cajas. La siguiente figura muestra esa relación.



El diagrama de cajas que le corresponde a una distribución normal teórica, muestra que las extensiones de los bigotes de la caja son proporcionales a las extensiones de la posición de los cuartiles. Dicho en otras palabras, la distancia entre el valor mínimo y Q1 es la misma que entre el valor máximo y Q3; además la distancia entre Q2 y Q1, es la misma que la que existe entre Q2 y Q3. La distribución descripta es "ideal", y sirve de punto de referencia teórico para comparar distribuciones empíricas.

Por lo dicho, para saber si una distribución de datos empíricos muestra algún sesgo, solo basta agregar a un diagrama de cajas una medida complementaria de la mediana, que es la media aritmética y verificar además las distancias entre las distintas medidas de posición. Entonces, si la media aritmética y la mediana se apartan entre sí lo mismo que las medidas de posición, nos informa de la tendencia del sesgo de una distribución. Una aclaración importante en este punto es que no debe entenderse que las distribuciones segadas son "malas distribuciones" puesto que la forma del sesgo informa cómo están ordenados los datos de una variable empírica.

Para completar lo que acabamos de mencionar, veamos un ejemplo: en una escuela se aplicaron tres métodos diferentes para la enseñanza de las matemáticas. Luego de dos meses, los niños que aprendieron con cada método fueron evaluados con una prueba estandarizada cuyo puntaje máximo alcanza 50 puntos. Los resultados de los grupos en la evaluación se muestran en el siguiente gráfico de cajas. En esta gráfica se aprecia que la mediana está representada por una línea transversal a la caja y la media aritmética por un punto en el interior de la caja.



Para este ejemplo utilizaremos las siguientes referencias: el método A y B son los métodos nuevos, los cuales son comparados con el método tradicional (T: Método Tradicional). No analizaremos en detalle la complejidad de los gráficos solo su asimetría. El método A muestra la particularidad de que Q2 y Q3 están más próximos entre sí, que Q2 y Q1, por tanto, un tercio de los casos se ha concentrado en los valores 36, 37 y 38. Por otro lado, los bigotes son diferentes, el inferior es considerablemente más pequeño que el superior, por lo tanto, hay mayor dispersión de casos en los valores altos de la variable. El método B ofrece una distribución diferente y se observa que el sesgo está marcado por los valores que están por encima de la mediana. Esto lo sabemos porque la media aritmética está por encima de la mediana, la distancia entre Q2 y Q3 es mayor que entre Q2 y Q1 y porque el bigote superior es más largo que el inferior. Considerando todos estos aspectos del gráfico concluimos que los que participaron en este método tienden a dispersarse más hacia los valores altos de la variable. Por último, el método tradicional es una imagen casi especular del método B, solo que la tendencia es hacia los valores bajos de la variable. En conclusión, es posible afirmar que los métodos que mejores resultados ofrecen son el A y el B en comparación con el método tradicional. Comparando los métodos A y B respectivamente, el primero muestra una ventaja relativa respecto del segundo.

Guía de Actividades N°3

Distribuciones de frecuencias-Representación gráfica de distribuciones de frecuencias-Interpretación de gráficos

Objetivos de la actividad:

a) Estimular la capacidad producir y utilizar tablas de frecuencia.
b) Fomentar la construcción e interpretar diferentes tipos de gráficos.

Actividad 1: Los datos que se muestran en la siguiente tabla, corresponden a las notas obtenidas en la asignatura lenguaje oral y escrito, de un grupo de escolares de 6º grado. Las columnas de la tabla indican la nota obtenida por los alumnos. Actualmente en la escuela existe un debate sobre la nota que debería utilizarse como criterio para determinar que el alumno alcanzó un conocimiento suficiente en esa asignatura para ser promovido. Un grupo de profesores estima que la nota de corte apropiada es de 6, mientras otros profesores estiman que la nota de corte debería ser 4.

Distribución de notas de 40 alumnos de 6º grado en la asignatura lenguaje oral y escrito:

Notas	**Frecuencia**	**F. Relativa**	**Porcentaje**	**P.**
3	12			
4	16			
5	21			
6	11			
7	14			
8	9			
9	5			
10	4			
Total	**92**			

Con los datos de la tabla se solicita:

a) Identifique la unidad de análisis y la variable en estudio
b) Construya una tabla de frecuencias que contenga: Frecuencia Relativa, Porcentajes, Porcentaje Acumulado.
c) Construya un gráfico de barras con los datos de la columna porcentaje.
d) Determine las diferencias en la distribución de alumnos promovidos, teniendo en cuenta ambos criterios propuestos por los docentes.

e) Argumente a favor del criterio que a su juicio considera más conveniente como nota de corte.

Actividad 2: Las tablas que se muestran a continuación contienen las frecuencias de los promedios de alumnos que formaron parte de una muestra de egresados de la carrera de psicología en el año 2005 y de una muestra del mismo tipo, egresados en el año 2010 y 2015.

Frecuencias de los promedios obtenidos por alumnos de la carrera de psicología

Año 2005	
Nota	**F absoluta**
5	30
6	40
7	140
8	190
9	80
10	50
TOTAL	

Año 2010	
Nota	**F absoluta**
5	120
6	360
7	210
8	80
9	70
10	20
TOTAL	

Año 2015	
Nota	**F absoluta**
4	20
5	155
6	160
7	110
8	60
9	30
10	12
TOTAL	

Con la información ofrecida se solicita:

a) Identifique la unidad de análisis y la variable en estudio.
b) Complete la tabla con las frecuencias relativas y frecuencia relativa expresada en porcentajes, para las cohortes.
c) Construya un gráfico adecuado con los porcentajes obtenidos y analice su distribución.
d) Compare las tablas y los gráficos, y ensaye una breve conclusión.

Actividad 3: En un estudio longitudinal se analizó la tasa de repitencia en escuelas del interior y de la Capital de la Provincia de Córdoba. En las siguientes tablas se presentan las frecuencias relativas tomadas de una serie de evaluaciones cada cinco años, durante el período 1985 a 2010. Estas frecuencias corresponden a las tasas de repitencia de alumnos del nivel medio.

Tasa de repitencia de alumnos del Nivel Medio Escuelas del Interior y Capital; Provincia de Córdoba

Años	**Tasa de repitencia Escuelas del Interior**	**Tasa de repitencia**
1985	17 %	10 %
1990	22 %	14 %
1995	23 %	23 %
2000	19 %	24 %
2005	17 %	27 %
2010	16%	29%

A partir de los datos presentados:

a) Identifique la unidad de análisis y la variable en estudio.
b) Grafique en ejes cartesianos la distribución de los porcentajes en los años estudiados.
c) Describa la tendencia observada.
d) Esboce una posible explicación de la misma.

Actividad 4: En las siguientes tablas se muestran dos distribuciones de frecuencias agrupadas correspondiente a una prueba de ortografía. La tabla 1, corresponde a una muestra de alumnos que aprendieron con el método tradicional en donde se cuenta la cantidad de palabras escritas correctamente. La tabla 2 muestra la misma distribución, pero para alumnos que aprendieron con un método nuevo.

Tabla 1: Método Tradicional

Cant. Palabras	**f**	**fa**	**f´**	**f´a**
20 – 24	5			
25 – 29	9			
30 – 34	8			
35 – 39	5			
40 – 44	4			
45 – 49	9			
50 – 54	7			
55 – 59	3			
Total	50			

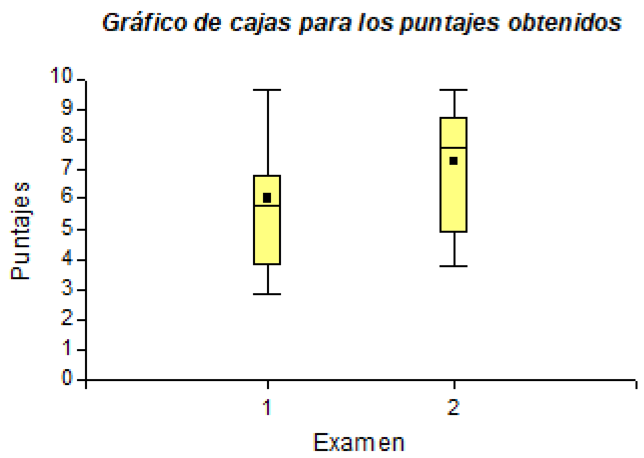
Tabla 2: Método Nuevo

Cant. Palabras	**f**	**fa**	**f´**	**f´a**
20 – 24	5			
25 – 29	3			
30 – 34	4			
35 – 39	5			
40 – 44	10			
45 – 49	11			
50 – 54	9			
55 – 59	3			
Total	50			

A partir de los datos presentados:

a) Identifique la unidad de análisis y la variable en estudio.
b) Calcule las frecuencias acumuladas, relativas y relativas acumuladas.
c) Construya un histograma de la distribución de frecuencias absolutas para ambas tablas.
d) Describa la tendencia observada.
e) Esboce una posible explicación de la misma.

Actividad 5: Un grupo de niños fueron evaluados en cuanto a su desempeño en Lengua con dos exámenes distintos (forma 1 y forma 2). Los docentes encontraron diferencias en los resultados y los expresaron en un informe de donde se extrajeron las siguientes afirmaciones. Léalas atentamente y diga a que examen se refiere cada una. Nota: el punto dentro del Box Plot identifica el promedio, no lo tenga en cuenta para la respuesta.



a) El rendimiento más bajo en el examen (____) se encuentra entre los puntajes 3 y 4.

b) Los alumnos que rindieron con el examen (____) obtuvieron notas entre 5 y 9 según lo muestra la posición de los cuartiles Q1 y Q3.
c) La mediana del examen (____), fue de 8 puntos.
d) En el examen (____) se verifica que la distancia entre Q2 y Q3 está comprendida entre las notas 6 y 7.

Actividad 6: En una escuela se tomó una prueba de madurez visoespacial para la lectura a los niños que ingresaban al primer grado. Según los estándares de la prueba, el punto de corte para establecer que el niño posee una madurez suficiente es de 74 puntos. En la siguiente tabla se muestra la distribución de alumnos evaluados según hayan sido asignados al turno mañana o tarde.

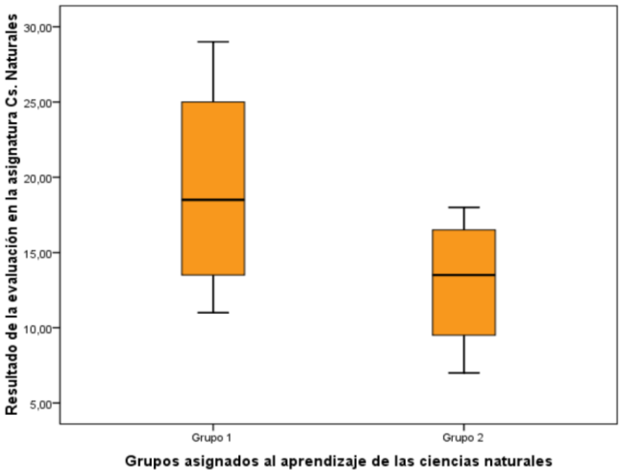
Prueba de madurez lectora alumnos de turno mañana y tarde.

Intervalo	**Turno Mañana**	**Turno Tarde**
66 – 68	0	3
69 – 71	1	3
72 – 74	3	15
75 – 77	4	8
78 – 80	3	5
81 –83	6	4
84 – 86	15	3
87 – 89	8	3
90 – 92	4	0
Total	44	44

Con la información ofrecida se solicita:

a) construya un histograma con la distribución porcentual de los puntajes en cada uno de los turnos,
b) establezca el porcentaje de niños que obtuvieron un puntaje igual o mayor al punto de corte estandarizado, en cada uno de los turnos.

Actividad 7: El diagrama de cajas que se presenta a continuación caracteriza el rendimiento, evaluado mediante una prueba estandarizada diseñada ad hoc, de dos muestras de escolares identificados como grupo 1 y 2. El grupo 1 participó a lo largo de un semestre de la enseñanza de ciencias naturales a través del desarrollo de un proyecto grupal con la participación de investigadores invitados a las clases que oficiaban de tutores. El grupo 2, sólo aprendió la asignatura con el dictado de clases magistrales por parte del profesor. Mediante la inspección del gráfico, identifique las principales diferencias en el rendimiento de ambos grupos, utilizando como referencia las medidas de posición de cada uno de ellos.


30,00
25,00
20,00
15,00
10,00
5,00
Resultado de la evaluación en la asignatura Cs. Naturales
Grupo 1
Grupo 2
Grupos asignados al aprendizaje de las ciencias naturales

Actividades en PSPP N° 3:

En la guía de actividades del capítulo 1 y 2 hemos abierto una base de datos que trae como ejemplo el programa PSPP, llamada Phisiology.sav, a la que le cambiamos el nombre de las variables en inglés por su equivalente en español, modificamos las etiquetas de la variable Sexo, originalmente en inglés, por las etiquetas en español. Finalmente, hemos corregido un error de carga de datos en la variable Altura.

Además, hemos pedido las medidas de tendencia central y de posición de las variables Temperatura… y Peso. Finalmente, hemos pedido una tabla de frecuencias de la variable Sexo y determinamos la moda.

En esta oportunidad, vamos a hacer algo similar a la unidad dos con una segunda base que trae el programa entre los ejemplos. Es una encuesta de satisfacción realizada en un hotel. La base se llama hotel.sav y se encuentra en la carpeta ejemplos/examples en la que hallamos la base anterior. Una vez que abrimos la base, veremos que es necesario colocar las etiquetas de los valores, como hicimos en la unidad anterior, para entender qué significa cada código. Además, rápidamente podremos ver que la base cuenta con 17 casos y 5 variables.

El paso que sigue es modificar la base para llevar las etiquetas de las variables al español y las etiquetas de los valores también. Para ello, lo primero será guardar la base con otro nombre en nuestra carpeta de actividades prácticas. La guardaremos con el nombre Hotel (con mayúscula) para no confundirla con la base "hotel" en inglés.

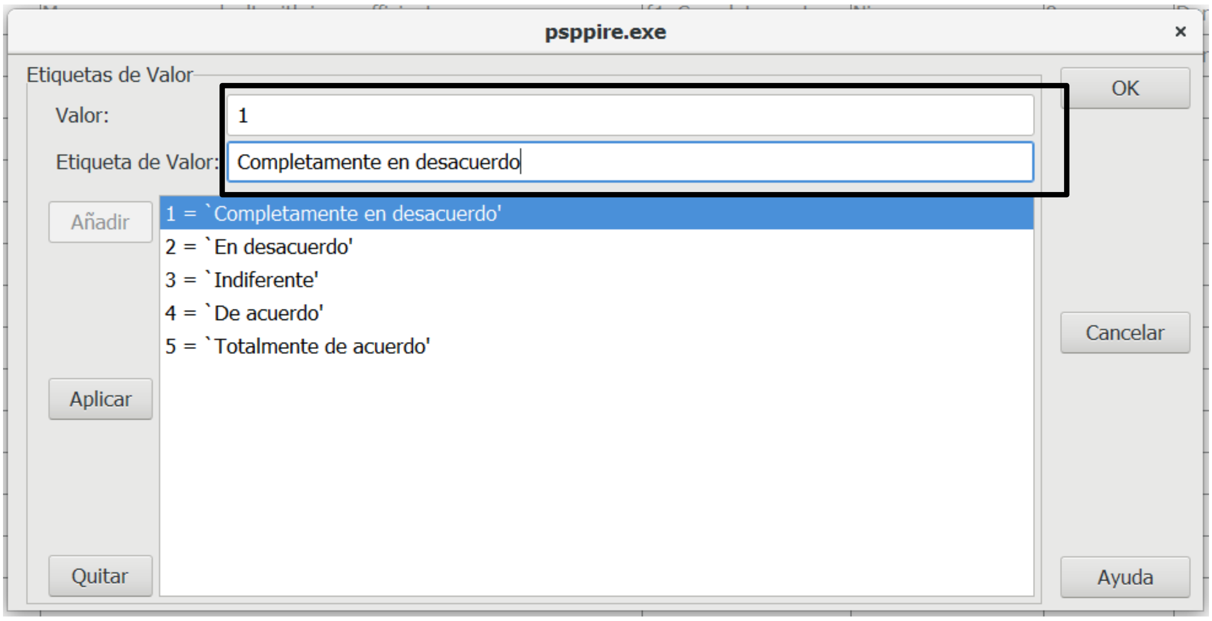
Luego vamos a la vista de variables y modificamos el nombre de las variables como sigue:

Variables	Etiqueta vieja	Etiqueta Nueva
V1	I am satisfied with the level of service	Estoy satisfecho con el nivel del servicio
V2	The value for money was good	La relación calidad precio fue buena
V3	The staff were slow in responding	El personal tardó en responder
V4	My concerns were dealt with in an efficient manner	Mis demandas fueron tratadas de un modo eficiente
V5	There was too much noise in the rooms	Había mucho ruido en las habitaciones.

Una vez que hemos renombrado las variables en español, procederemos a renombrar las categorías de cada una de las variables:

Variables	Etiqueta vieja	Etiqueta Nueva
C1	Strongly Disagree	Completamente desacuerdo
C2	Disagree	Desacuerdo
C3	No Opinion	Indiferente (vamos a usar esta, aunque no sea literal, para que la variable sea ordinal)
C4	Agree	De acuerdo
C5	Strongly Agree	Totalmente de acuerdo

Luego que hemos modificado las categorías para la primera variable, será necesario modificar las categorías, una por una, para todas las demás variables.



Usted debe indicar ahora el nivel de medición de las variables en cuestión:
………………………………………………………………………………………………………
………………………………………………………………………………………………………
………………………………………………………………………………………………………
………………………………………………………………………………………………………
………………………………………………………………………………………………………
………………………………………………………………………………………………………
…………………………………………..

Ahora vamos a pedir al programa que nos brinde una tabla de frecuencia para cada una de las variables estudiadas. Para ello vaya al menú Analizar, Estadística Descriptiva, Frecuencias… y seleccione las cinco variables. En esta ocasión no solicitamos al programa ninguno de los estadísticos, por ser variables con un nivel de medición …………………………………………………..

Estoy satisfecho con el nivel del servicio

Etiqueta de Valor	Valor	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Completamente en desacuerdo	1	5	29,41	29,41	29,41
En desacuerdo	2	4	23,53	23,53	52,94
Indiferente	3	3	17,65	17,65	70,59
De acuerdo	4	4	23,53	23,53	94,12
Totalmente de acuerdo	5	1	5,88	5,88	100,00
	Total	17	100,0	100,0	

La relación calidad precio fue buena

Etiqueta de Valor	Valor	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Completamente en desacuerdo	1	5	29,41	29,41	29,41
En desacuerdo	2	7	41,18	41,18	70,59
Indiferente	3	2	11,76	11,76	82,35
De acuerdo	4	1	5,88	5,88	88,24
Totalmente de acuerdo	5	2	11,76	11,76	100,00
	Total	17	100,0	100,0	

El personal tardó en responder

Etiqueta de Valor	Valor	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Completamente en desacuerdo	1	2	11,76	11,76	11,76
En desacuerdo	2	2	11,76	11,76	23,53
Indiferente	3	4	23,53	23,53	47,06
De acuerdo	4	4	23,53	23,53	70,59
Totalmente de acuerdo	5	5	29,41	29,41	100,00
	Total	17	100,0	100,0	

My concerns were dealt with in an efficient manner

Etiqueta de Valor	Valor	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Completamente en desacuerdo	1	5	29,41	29,41	29,41
En desacuerdo	2	3	17,65	17,65	47,06
Indiferente	3	6	35,29	35,29	82,35
De acuerdo	4	3	17,65	17,65	100,00
	Total	17	100,0	100,0	

Había mucho ruido en las habitaciones

Etiqueta de Valor	Valor	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Completamente en desacuerdo	1	7	41,18	41,18	41,18
En desacuerdo	2	6	35,29	35,29	76,47
Indiferente	3	2	11,76	11,76	88,24
De acuerdo	4	1	5,88	5,88	94,12
Totalmente de acuerdo	5	1	5,88	5,88	100,00
	Total	17	100,0	100,0	

Ahora se solicita que describa los resultados obtenidos para cada una de las variables, teniendo en cuenta los datos ofrecidos por la tabla. También solicitamos que indique para cada una de las distribuciones de datos en qué es está la moda, cuál es la categoría que corresponde al primer, segundo y tercer cuartil y cuál es el valor de la mediana. Tómese todo el espacio que necesite.

………

………

………

………

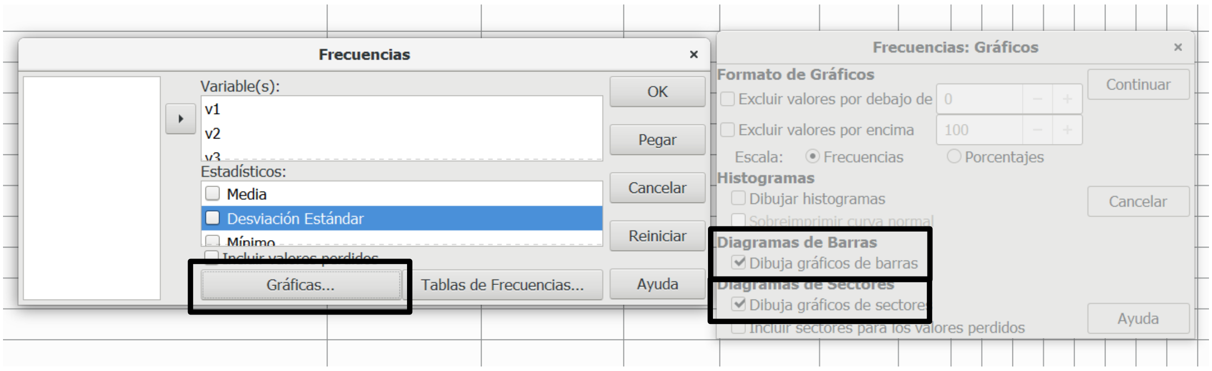
………

………

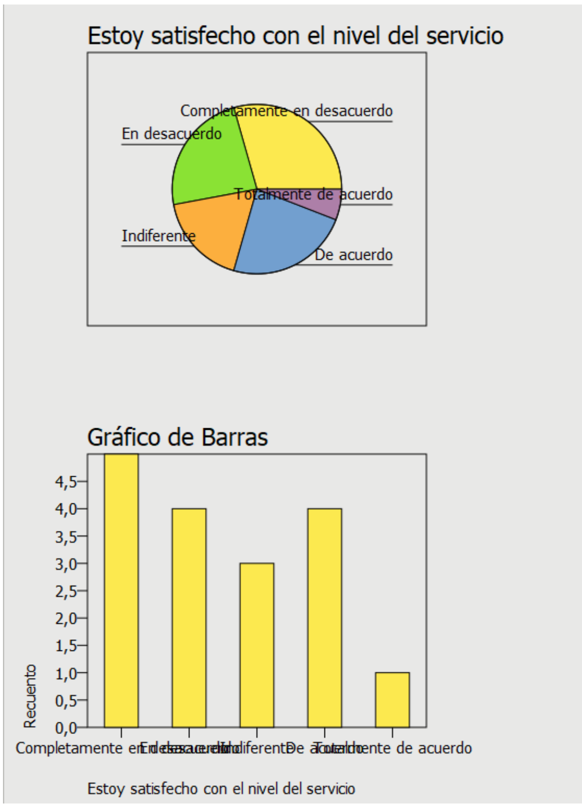
………

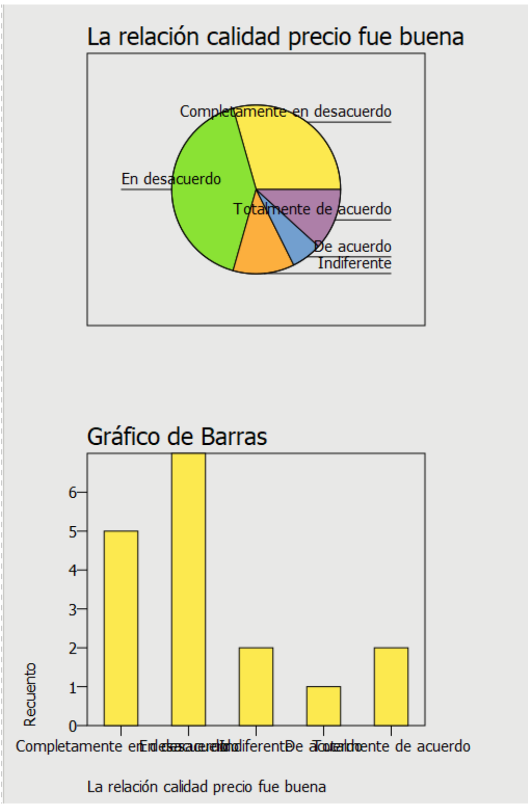
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……………………………………………………………

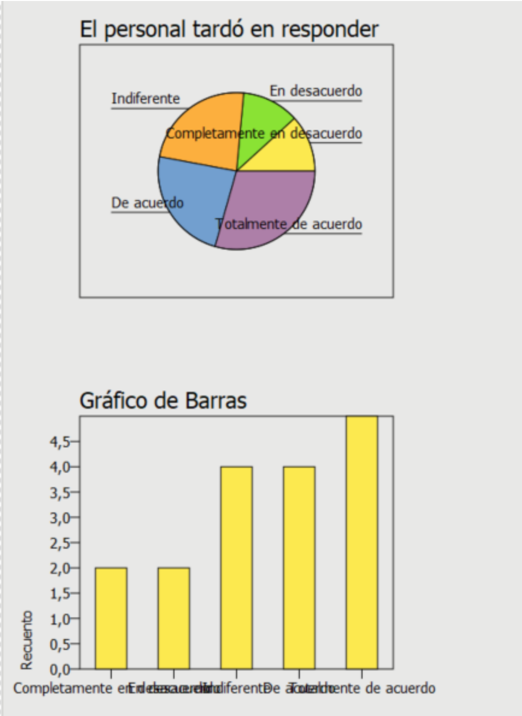
Ahora se le solicita que confeccione gráficos de barras para todas las variables. Para ello realizamos la misma operación que hicimos para las tablas de frecuencias y, en el momento de seleccionar las variables, daremos click en el botón de gráficos, una vez allí le pedimos que nos dé la opción de diagrama de barras y diagrama de sectores.

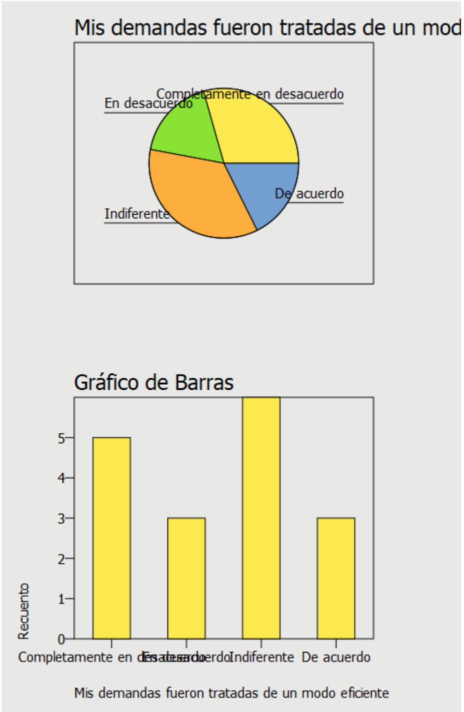


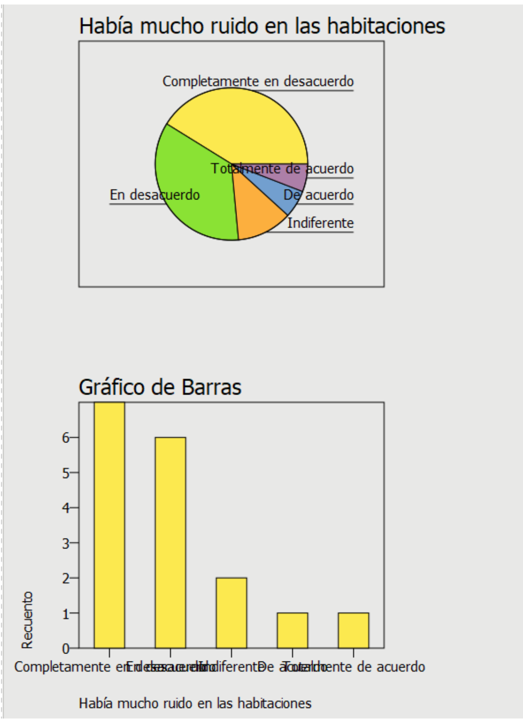
Se les pide además que describan los resultados que deben aparecer como los siguientes:











Como pude observar, no son de muy buena calidad los diagramas del PSPP, pero siempre tendrán la opción de copiar la tabla que brinda por resultado el programa a Excel y graficar con este último programa. Más adelante le explicaremos como se hace esto.

Capítulo 4

Técnicas de muestreo

¡Para muestra vale un botón! Habrán escuchado esta frase más de una vez y lo que sugiere es que un botón nos permite inferir las características del conjunto de botones que está en el cajón. En esta ocasión nos vamos a avocar a estudiar cómo se construyen las muestras y qué tipo de muestras podemos extraer de una población que queremos investigar, en función de los objetivos que nos hayamos propuesto. El muestreo es una técnica fundamental por muchas razones, pero quizás la más importante es que nos brinda información acerca de un conjunto de individuos según determinadas condiciones y para un conjunto de variables que estamos estudiando. Existen diferentes técnicas de muestreo y estas diferencias están condicionadas por la información que tengamos de la población, las posibilidades para seleccionar una muestra aleatoria y la necesidad de expandir los resultados a la población o no.

La estadística cuenta con diversas herramientas para estimar en una población, los parámetros que deseamos conocer. La estimación dependerá en buena medida de una apropiada definición operacional de lo que entendemos como población. Definir una población con precisión implica un trabajo detallado y minucioso de sus límites, de modo que sea factible obtener su muestra representativa. Si estamos interesados en estimar un parámetro (como por ejemplo el promedio de las notas en una asignatura), tendremos que optar por un sistema de muestro que garantice la confianza en la estimación. Es por ello que apelamos a las técnicas de muestreo probabilístico.

Junto al muestreo probabilístico, existen otras técnicas menos rigurosas para obtener muestras de una población, los muestreos no probabilísticos. Aunque no se emplean en la estimación de parámetros, los tipos de muestreo no probabilísticos se aplican en casos puntuales aportando valiosa información del comportamiento de una variable de interés. En este apartado veremos la manera en que se define teórica y operacionalmente una población, y cómo se aplican las distintas técnicas de muestreo, tanto probabilístico como no probabilístico.

Muestra y Población

Supongamos que impartimos un curso de Reanimación Cardiopulmonar (RCP) en alumnos de quinto y sexto año de una escuela; pasado cierto tiempo deseamos saber qué conocimientos han fijado los estudiantes del curso. Entonces, seleccionamos al azar 10 de quinto y 10 de sexto y les pedimos que apliquen la técnica. De acuerdo al despeño de los 20 estudiantes, tenemos una representación de lo aprendido por todos los alumnos de quinto y sexto, lo que equivale a pensar que los demás también han fijado un contenido similar acerca de la aplicación de la técnica de RCP.

Si bien en estadística raras veces nos vamos a ocupar de este tipo de problemas, este simple ejemplo nos ilustra acerca de tres aspectos importantes: a) Población, b) Muestra, c) Inferencia. Si consideramos los estudiantes de quinto y sexto como la población de interés, la selección aleatoria de 20 alumnos correspondería a una muestra de la población. Al determinar que los alumnos seleccionados están en condiciones de aplicar la técnica de RCP, inferimos que el resto de sus compañeros deberían estar en la misma situación.

Población

En el lenguaje cotidiano la palabra población nos recuerda a individuos viviendo en una misma región, pero en estadística todo conjunto completo de elementos es una población. Así, todos aquellos individuos que habitan en suelo argentino, conforman la población de Argentina. Podemos extender los límites de este conjunto y decir que la Argentina es un país de Sudamérica y, por tanto, es un subconjunto de una población mayor, la de todos los habitantes de América del Sur. Sin embargo, las poblaciones de las que se ocupa la estadística van mucho más allá de las personas. Podemos definir como población cualquier entidad que pueda formar un conjunto. Por ejemplo, la cantidad de escuelas municipales de la ciudad de Córdoba también corresponde a un conjunto, y puede considerarse una población.

En estadística se dice que una población es finita cuando pueden enumerarse o listarse todos los elementos que forman el conjunto. Por ejemplo, si se define como población a todas las personas que se encuentran en condiciones de asistir a un determinado establecimiento escolar en calidad de alumnos, se tiene que la población está listada en lo que se conoce como la matrícula del colegio. Si no podemos establecer el número total de los elementos que contiene la población, decimos que es infinita.

Dado que el término finito e infinito se presta a confusión, se está reemplazando paulatinamente por el de límite de la población. Entonces, habrá poblaciones que tendrán un límite acotado, mientras que otras tendrán un límite no acotado. En estadística el límite de la población se toma en un sentido práctico y siempre acorde a los objetivos de la investigación. En este sentido, la primera cuestión que debe resolver el investigador es si tendrá acceso a la población que ha definido teóricamente. Se desprende de ello que la definición de la población determinará el tipo de muestreo posible.

Los sistemas de información estadísticos son las fuentes más consultadas cuando se quiere definir el límite de poblaciones a estudiar, especialmente por los datos que allí se recogen sobre población e instituciones. Por ejemplo, si estamos interesados en estudiar algún aspecto relevante de las escuelas municipales de Córdoba Capital, nuestra principal fuente de datos será el Ministerio de Educación. En esa institución están listadas todas las escuelas que conformarían la población de interés.

Existen poblaciones que permanecen no acotadas a pesar de los esfuerzos que se hacen por acotarlas, son las llamadas poblaciones de fracción desconocida. Este es un término técnico, pero con un ejemplo puede entenderse de que se trata. Imaginemos que estamos relevando en Córdoba Capital, la cantidad de personas que están en situación de Educación Hospitalaria, cuya proporción en la población es relativamente baja. En primer lugar, consultamos los datos del Ministerio de educación de la provincia para saber cuántos casos existen en la actualidad, y ese sería el límite de la población. Pero debemos suponer que hay más casos que no están escolarizados en esa modalidad, dado que muchos de ellos no han ingresado al sistema.

En criminología, se habla de una cifra negra de la criminalidad en tanto se sabe que muchos hechos delictivos no se denuncian, por ende, las estadísticas solo alcanzan a cubrir aquellos casos que efectivamente han sido denunciados. Aquí las estadísticas continuas son de una invalorable ayuda, dado que es posible estimar y hacer proyecciones a futuro a partir de datos conocidos.

Tomemos como ejemplo algunas situaciones donde las estadísticas son inexactas, pero de inestimable valor en la escuela secundaria: cantidad de casos de aborto, uso de drogas ilegales, abusos, violencia doméstica, etc.

Muestra e Inferencia

Cada vez que se seleccione una muestra o una población, está implícito que se desea medir o conocer alguna característica o propiedad importante; ya hemos visto que a estas propiedades las denominamos variables. La pregunta recae la mayoría de las veces en conocer cómo se comportan esas variables en la población. La inferencia es un procedimiento mediante el cual se extraen conclusiones de una población, a partir de una muestra representativa de la misma. Los modelos estadísticos utilizados para el análisis de datos, nos enseñan que una muestra representativa de una población es suficiente para conocer las principales características de ella. Entonces, al utilizar una muestra para determinar una propiedad de la población, estamos realizando una inferencia. Hay una rama de la estadística que recibe el nombre de estadística inferencial, justamente porque desarrolla y aplica métodos para que el proceso de extrapolación de la información que se obtiene de una muestra, se ajuste lo más fielmente posible a la población.

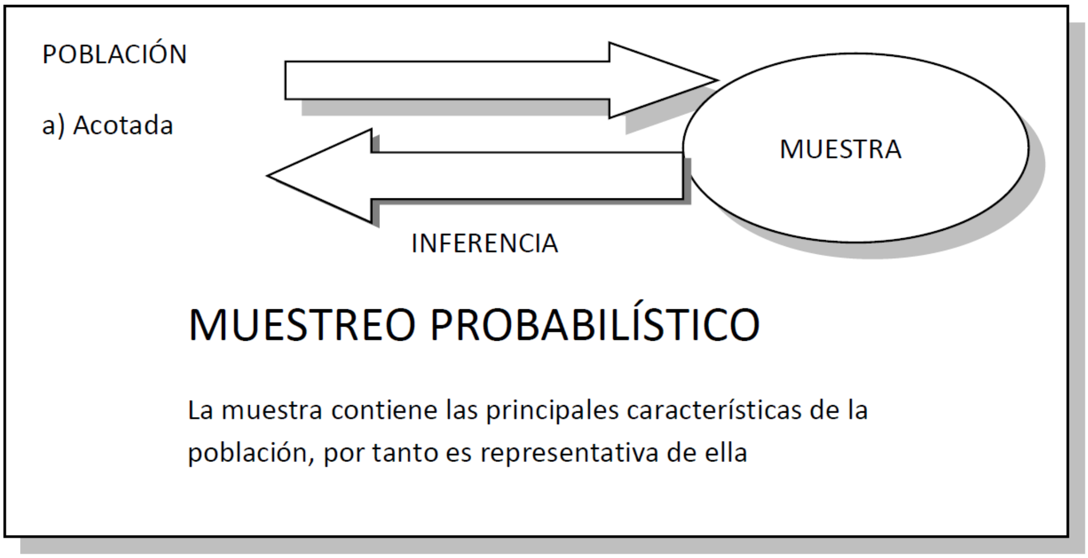
La inferencia será siempre inexacta, dado que tomamos una parte (muestra), para conocer el todo (población), aun así, es posible conocer el grado de incertidumbre con el que se trabaja, y además es factible fijar el nivel de error durante el proceso en términos probabilísticos. Para que esto sea posible, la definición de población debe ser precisa para que no se produzcan sesgos desde la propia definición. Luego, si la muestra es representativa de la población, contendrá las mismas características que ella. Es conveniente aclarar en este punto, que la representatividad de la muestra estará dada en buena medida por la técnica de muestreo utilizada.

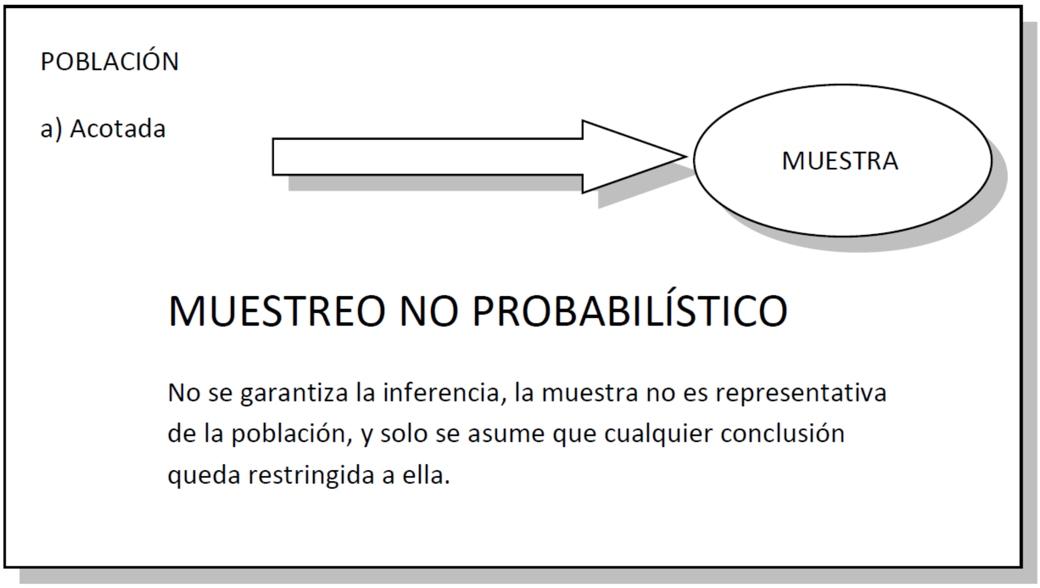
Es importante hacer notar que toda vez que calculemos un indicador sobre la población, lo denominaremos **parámetro**. Puesto que hay poblaciones que son muy difíciles de abarcar en su totalidad, utilizamos el muestreo para estimar ese parámetro; en tal caso el indicador será un **estadístico**. La estadística inferencial permite la estimación confiable de parámetros toda vez que se obtienen muestras representativas de una población; y estas solo se obtienen cuando el proceso de muestreo es de tipo probabilístico.

Sin embargo, no siempre se busca obtener una muestra representativa de la población para realizar una inferencia. En muchos casos, interesa tener una aproximación a la población pero no es finalidad de la investigación la estimación de un parámetro. Por ejemplo, en las etapas previas a la preparación de un instrumento de recolección de datos, éstos se prueban sobre muestras pequeñas de la población que no son representativas del total. Lo que se pretende es simplemente comprobar la confiabilidad del instrumento. Otras veces, se desea recolectar una muestra intencionalmente construida y sacar conclusiones sobre ella, sin extrapolar el resultado a la población. Un ejemplo de esto sería el trabajo que se realiza en unidades carcelarias, donde las conclusiones del estudio quedan confinadas a la muestra estudiada.

Finalmente, existen situaciones donde es muy complejo y costoso obtener una muestra representativa de la población y es necesario trabajar sobre los casos que se disponen. Por ejemplo, los estudios que se enfocan sobre características particulares que solo están presentes en una fracción muy pequeña de la población, tal el caso de enfermedades poco comunes como las encefalopatías virales y otros padecimientos similares. En situaciones donde no se pretende realizar una inferencia o cuando no es posible hacerlo, el muestreo es de tipo no probabilístico.

El siguiente esquema resume los principales aspectos del muestreo y sus propiedades.





Muestreo Probabilístico

Una cuestión fundamental para comprender el muestreo es que éste se planea y se ejecuta siempre en el contexto de un estudio con objetivos definidos. Tales estudios pueden ser de índole científica o de mercado, pero en todos los casos la finalidad del mismo determina el tipo de muestra requerida. En este apartado, realizaremos una breve descripción de los tipos de muestro conocidos como probabilísticos, es decir aquellos que son recomendados si se quiere garantizar una apropiada inferencia estadística.

Muestreo aleatorio simple

En este tipo de muestreo, cada uno de los elementos de la población tiene la misma probabilidad de estar en la muestra seleccionada. Supongamos que definimos como población a todos los alumnos que cursan la carrera de enfermería en la Universidad Nacional de Córdoba. Mediante el registro de matrícula, sabemos que se encuentran anotados 875 estudiantes. Para denotar el tamaño de la población utilizaremos la letra **N**, en mayúscula, para diferenciarlo de la muestra, para la cual utilizaremos la letra **n**, en minúscula. Así el tamaño de la población de interés es N=875.

De esta población, queremos tomar una parte, una muestra, que sea representativa el total; para ello realizamos un muestreo aleatorio simple. Nótese que bajo estas circunstancias es sencillo realizar este tipo de muestreo porque las unidades de muestreo se encuentran todas listadas de antemano (registro de alumnos matriculados). Ahora bien, para obtener de esta población una muestra, debemos definir el tamaño de la misma. Así, tendríamos que:

N=875 (población)
n=100 (muestra)

En este ejemplo, tamaño muestral lo hemos fijado de antemano, pero en una investigación real éste deberá calcularse según dos criterios que son: el tamaño del efecto y la potencia estadística. Estos criterios exceden la explicación que podemos dar en este contexto, pero cabe destacar que, en las publicaciones científicas, existe casi siempre una referencia a ellos. Además, en los programas estadísticos para poder estimar el tamaño muestral necesitamos establecer una serie de parámetros tales como el intervalo de confianza de la estimación. Estos temas no serán tratados en este tutorial, dado su carácter introductorio.

Continuando con el ejemplo, supongamos que el estudio que se lleva a cabo indaga acerca de las perspectivas laborales que tienen los estudiantes. Para tal fin se aplica un cuestionario que contiene la siguiente pregunta: *"De acuerdo a su opinión, la profesión de enfermero tiene una demanda laboral: a) elevada, b) moderada, c) escasa, d) casi nula"*.

Como ya se dijo, para obtener una aproximación a la opinión general de los estudiantes, es necesario que la muestra sea representativa del total de la población, y para ello se opta por un muestreo aleatorio simple. Una condición de este tipo de muestreo es que cada uno de los miembros de la población tenga la misma chance de ser seleccionado, por lo tanto, para obtener la muestra de 100 estudiantes procederíamos de la siguiente manera:

a) numeramos a los estudiantes desde el 1 al 875,
b) seleccionamos uno al azar, ese será el primer estudiante de la muestra,

c) reponemos el estudiante seleccionado a la población,

d) seleccionamos un segundo estudiante para la muestra,

e) el procedimiento se repite hasta completar los 100 estudiantes.

Es necesario reponer el estudiante seleccionado a la población porque si así no se hiciera, el primer estudiante tiene una probabilidad de ser seleccionado de 1/875, pero el estudiante número dos tiene una probabilidad de 1/874, que es diferente a la del primero. El muestreo aleatorio simple, implica siempre reponer la unidad seleccionada a la población si queremos que cada individuo seleccionado tenga la misma probabilidad de ser escogido, aun a riesgo de volverlo a seleccionar. Se desprende de lo dicho que este tipo de muestreo es aplicable cuando:

a) la población no es muy grande,

b) se tiene acceso a todos los individuos,

c) es posible hacer un listado exhaustivo de la población,

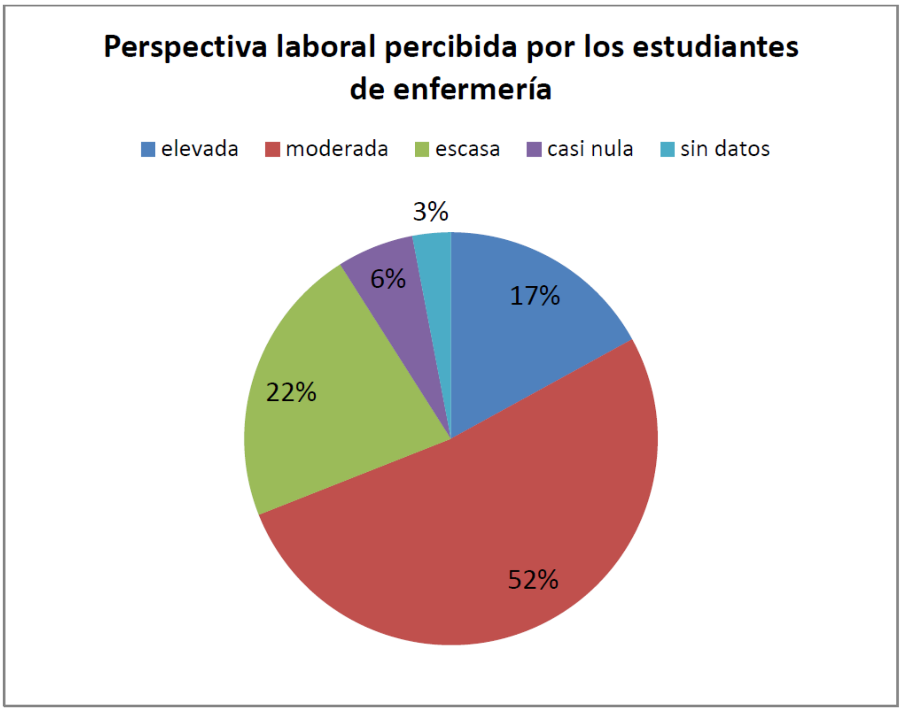
d) se puede repetir el proceso de muestreo cuantas veces sea necesario.

Supongamos entonces que la encuesta arrojó los siguientes resultados:

Perspectiva laboral percibida en estudiantes de la carrera de enfermería

elevada	**moderada**	**escasa**	**casi nula**	**sin datos**
17	52	22	6	3

Los mismos datos pueden representarse en porcentajes en un gráfico de sectores, tal como se muestra a continuación.



Podemos concluir que la mayoría (52%) estima que su título de grado le garantiza una inserción laboral moderada, en un porcentaje menor (22%) se encuentran aquellos que tiene una opinión poco optimista de tal inserción, aunque los que consideran que las oportunidades de trabajo son elevadas se aproximan a aquello que consideran nula esta posibilidad (17%).

Dado que el procedimiento para recoger esta muestra ha sido aleatorio simple, es posible afirmar que, de haber encuestado a todos los estudiantes, los resultados estarían muy próximos a los reportados. El sustento de tal inferencia es posible porque la muestra es representativa del total de estudiantes de la carrera de enfermería. Una conclusión inicial a la que lleva estos datos, es que casi tres cuartos de los estudiantes, estima que sus posibilidades laborales no estarían garantizadas con la obtención del título.

Muestreo aleatorio sistemático

El muestreo aleatorio sistemático es una variante del muestreo aleatorio simple para ser aplicado a poblaciones mucho más grandes. En este caso los elementos de la muestra se eligen a intervalos del listado completo de la población. Supongamos que el Sindicato Argentino de Docentes Privados (SADOP), desea realizar un sondeo de opinión sobre las dificultades para articular medidas de fuerza en las escuelas privadas. Según su registro (ficticio), existen actualmente 6600 docentes en ese sindicato; entre ellos se desea recolectar una muestra de 450 docentes del sector privado, a quienes se le solicitará su opinión para la siguiente pregunta: *"En qué medida se ha sentido coaccionado por acatar medidas de fuerza en su institución"*; las opciones son: a) completamente, b) severamente, c) moderadamente, d) levemente.

Dado que el tamaño de la población es grande, se opta por realizar un muestreo de tipo aleatorio sistemático. Para ello, se debe escoger el valor de arranque para la selección de los participantes. Este valor se obtiene del cociente entre el total de la población y el tamaño de la muestra; esto es:

$$\boxed{N/n = 6600/450 = 14{,}66}$$

Dado que 14,66 es un número decimal lo redondeamos al entero más próximo, esto es 15. Esto nos dice que, al dividir la población sobre la muestra, sabemos que aquella puede ser parcelada 450 veces si tomamos grupos de aproximadamente 15 personas.

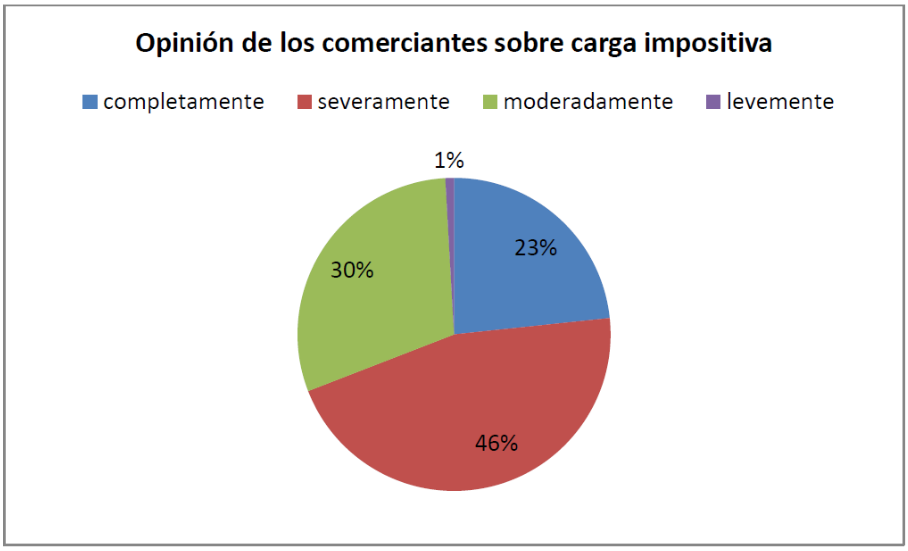
El muestreo sistemático se basa en la elección aleatoria de un docente cada 15, para esto escogemos un valor por sorteo comprendido entre 1 y 15. Suponiendo que el valor sorteado es el número 7, la primera unidad de muestreo será el séptimo docente de la lista. El siguiente docente seleccionado será el que caiga en el lugar 7+15=22, esto es, el docente vigesimosegundo de la lista es la siguiente unidad de muestreo; se deduce entonces que la tercera unidad de muestreo es el docente en el puesto 22+15=37. Mediante este procedimiento se escogen todas las unidades de muestreo a intervalos regulares del listado original.

Este procedimiento requiere el cuidado de mezclar los elementos de la lista de docentes (población), para que no contenga ningún sesgo. Nótese que el muestreo sistemático puede aplicarse a poblaciones más grandes al presentar menos complicaciones que un muestreo aleatorio simple; sin embargo, si este último es aplicable es el método que debe escogerse.

Continuando con el ejemplo que esbozamos, el resultado de la encuesta puede resumirse en la siguiente tabla y gráfico:

Opinión de los docentes de instituciones privadas sobre las coacciones por medidas de fuerza.

completam	severam	moderadam	leveme
105	206	135	4



De acuerdo a los resultados obtenidos en la muestra seleccionada, es posible afirmar que las coacciones tienen un peso importante sobre la capacidad de realizar o adherir a medidas de fuerza, dado que una amplia mayoría de los docentes de instituciones privadas (69%) reparten su opinión en las categorías completamente y moderadamente.

Muestreo aleatorio estratificado

Como su nombre lo indica, en este tipo de muestreo es necesario definir los estratos de los cuales se van a obtener las muestras. En términos sencillos, un estrato se determina a partir de cualquier propiedad de la población que permita dividirla, tal división posibilita realizar un muestreo aleatorio dentro de cada estrato, dado que siempre serán más pequeños que la población.

Pareciera relativamente sencillo determinar qué puede definirse como estrato para una población, pero no es tarea fácil dado que cada estrato debe ser homogéneo en su interior, y además deben ser heterogéneos entre ellos. Homogeneidad y heterogeneidad deben entenderse en el contexto de la variable que se está estudiando. Por ejemplo, la matrícula escolar, es una variable con la que pueden segmentarse las escuelas por tamaño, de tal modo, las escuelas quedan clasificadas como grandes, medianas y pequeñas. El tamaño de la escuela en este caso es el estrato y por tanto todas aquellas escuelas clasificadas como pequeñas, se espera que sean más parecidas entre sí, respecto de escuelas grandes. Siguiendo este razonamiento, podemos suponer que, en otro estrato, digamos el de directoras/es de grandes escuelas, su opinión será diferente respecto de las/os directoras/es de escuelas de menor tamaño y se evidenciará heterogeneidad en la opinión a través de los estratos estudiados. Finalmente, es de mencionar que los estratos, una vez definidos, deben ser exhaustivos para la población.

El tipo de gestión de una escuela es un atributo que puede funcionar como estrato para la población, así es posible realizar un listado completo de los alumnos que asisten a la escuela media en una ciudad dada, considerando como estratos a las escuelas de gestión estatal y de gestión privada. Si se está considerando el alumnado, las categorías descriptas funcionan adecuadamente como estratos, dado que un alumno puede asistir sólo a una escuela.

Tomemos por caso una variable ficticia que definimos como consumos culturales recreativos (v.g. música, lecturas, etc.). Podemos suponer en este ejemplo que los consumos culturales recreativos de los alumnos que asisten a escuelas de gestión estatal serán distintos de aquellos que asisten a escuelas de gestión privada. Los estratos definidos permitirían captar el grado de heterogeneidad entre ellos. Asimismo, la gestión de la escuela permite definir un grupo más o menos homogéneo en esa variable, cuando la analizamos al interior del estrato.

Dado que, en general, los estratos contienen diferente número de unidades de muestreo, en ocasiones el muestreo estratificado se toma proporcional al tamaño del estrato. En nuestro ejemplo, seguramente habrá menos alumnos matriculados en escuelas privadas que en escuelas estatales, por lo tanto, será necesario tener en la muestra más alumnos de este último tipo de escuelas. Sin embargo, en ocasiones no es necesario respetar la proporcionalidad de los estratos, de modo que una vez que los hemos definido, se pueden extraer muestras de igual tamaño de cada uno de ellos. La necesidad de tener muestras proporcionales al estrato, dependerá del tipo de variable que se analice. En general, cuanta menor correlación tiene la variable de interés con el estrato, menos importante es mantener la proporcionalidad del muestreo.

Veamos ahora un ejemplo de muestreo estratificado: el Organismo Consultivo Nacional para la Educación Superior, lanzó una encuesta de opinión sobre las condiciones laborales en las distintas unidades académicas del país. Se contaron 3042 docentes con dedicación exclusiva, de los cuales se decidió tomar una muestra de 600 docentes. Para realizar el muestreo se dividió la población en los siguientes estratos de acuerdo a la orientación del título dado por la unidad académica o facultad:

Ingeniería	**Ciencias de la Salud**	**Ciencias Sociales**	**Artes y Humanidades**	**Matemática Física y Astronomía**
660	677	578	758	369
21.7%	22.25%	19%	24.9%	12.13%
130	134	114	149	73

La segunda fila de la tabla contiene la cantidad de docentes por estrato y la tercera fila representa el porcentaje del total de la población. Si se quisiera realizar un muestreo proporcional al estrato, utilizamos ese porcentaje para conocer cuántos docentes (por estrato) deberían seleccionarse para la muestra (cuarta fila). Si la muestra requerida es de 600 docentes, multiplicamos: Proporción de docentes en el estrato, por, tamaño muestral, y así tenemos la cantidad de docentes que se requieren para ese estrato.

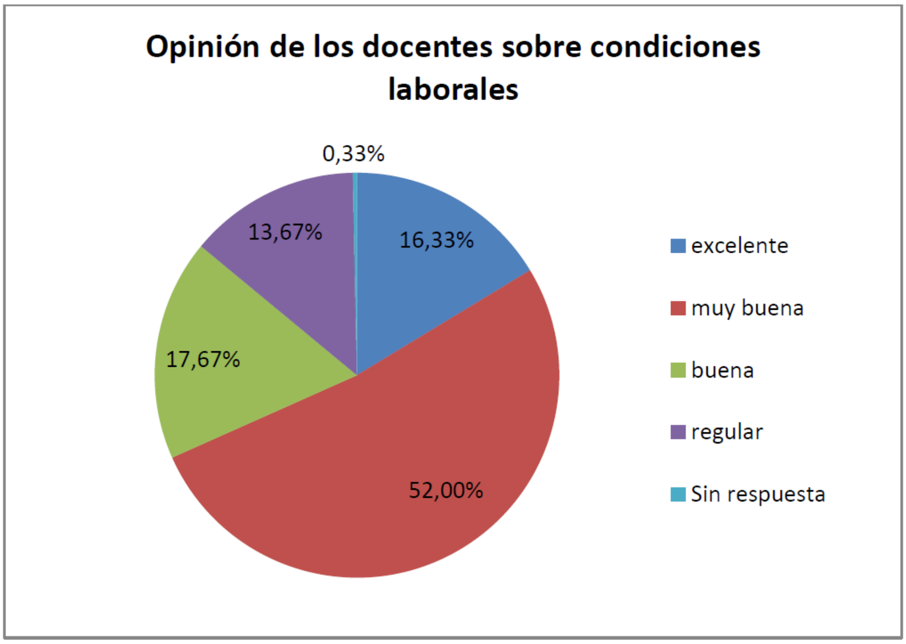
Así, en ingeniería la cantidad de docentes a ser encuestados serían: 0,217x600= 130,2 que podemos aproximar a ciento treinta docentes. Este procedimiento se repite para cada estrato y de este modo, obtenemos la muestra necesaria, proporcionada al tamaño del estrato.

El procedimiento nos da una aproximación al número de unidades de muestreo requeridas, pero existen otros modos de ponderación más exactos. Nótese que la multiplicación se realiza empleando la frecuencia relativa y no el porcentaje (más adelante veremos que el porcentaje se deriva directamente de la frecuencia relativa).

Continuando con el ejemplo, diremos que en el sondeo de opinión se realiza la siguiente pregunta: *"Considera Ud. que las condiciones laborales de su unidad académica son: a) excelentes, b) muy buenas, c) buenas, d) regulares"*. Las categorías de respuestas se resumen en la siguiente tabla y gráfico:

Opinión sobre las condiciones laborales de los docentes de educación superior

excelente	muy buena	buena	regular	Sin respuesta
98	312	106	82	2



Nota. Se observa que poco más de la mitad de los docentes encuestados considera que las condiciones laborales de su unidad académica son muy buenas. El resto de las opiniones muestran porcentajes similares, pero en categorías disímiles en su contenido. Dado que, en este tipo de muestreo, cada estrato puede estudiarse como una unidad en sí misma, puede desagregarse la información por unidad académica, es decir en los diferentes estratos, lo cual permitiría analizarlos y compararlos.

En la tabla que se muestra a continuación se ha realizado esto, favoreciendo la comparación pretendida.

	excelente	muy buena	buena	regular	sin respuesta
Ingeniería	40	71	7	11	0
Ciencias de la	18	79	27	19	1
Ciencias	12	59	17	15	0
Artes y	11	88	45	34	1
Matemática, Física y Astronomía	17	15	10	3	0

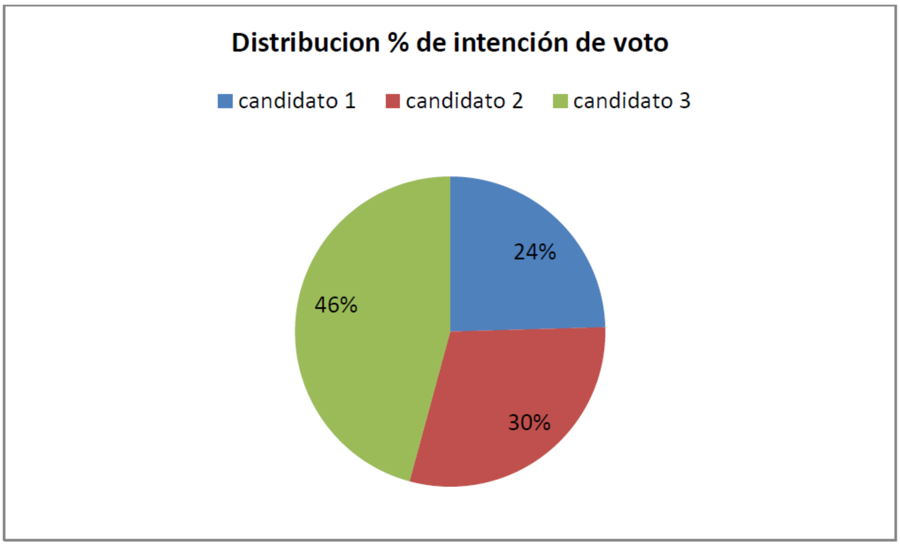
Observando cada estrato vemos que hay diferencias de opinión de los docentes de las distintas unidades académicas: por ejemplo, los docentes de la facultad de Matemática, Astronomía y Física tienden a una opinión positiva en relación a sus condiciones laborales (la mayoría de las respuestas corresponden a las categorías excelente y muy buena) mientras que más de la mitad de los docentes de Artes y Humanidades tienen una opinión positiva y el resto de las categorías más representadas son las correspondiente a buena y regular. Otra conclusión en este mismo sentido sería que, los docentes de las Facultades de Ingeniería y Matemática, Astronomía y Física, tienen opiniones más parecidas entre sí; lo mismo que las facultades de Artes y Humanidades y Ciencias de la Salud.

Muestreo por conglomerados

En el caso en que no pueda definirse el límite de la población, o bien ésta sea demasiado grande como para obtener un listado de las unidades de muestreo, es posible trazar conglomerados dentro de los cuales realizar el muestreo. El término conglomerado se ha mantenido en la literatura puesto que proviene de la planificación para obtener grandes muestras de la población y se aplica a las parcelas geográficas o demográficas en que ésta puede dividirse. De acuerdo a la variable de interés, se intenta que los conglomerados definidos sean, en general, heterogéneos en su interior ya que deben conservar la diversidad que existe en la población, y homogéneos entre sí ya que unos pocos conglomerados deberían alcanzar para representar toda la población.

El muestreo por conglomerados se aplica cuando las unidades de muestreo no están listadas; por tanto, es necesario trasladarse hasta donde se encuentran estas unidades. Como se mencionó, el conglomerado suele coincidir con una parcela geográfica. Como ejemplo, puede tomarse los diferentes distritos electorales en los sondeos de opinión. La zona de cobertura de una escuela también puede ser tomada como un conglomerado.

Para comprender mejor este procedimiento, tomemos otro ejemplo ficticio. Una autoridad del Ministerio de educación realiza en período preelectoral un sondeo de opinión de los tres candidatos principales a sustituirla. Para este estudio, en el ejido municipal se identificaron cinco conglomerados. Se realiza un muestreo aleatorio de escuelas dentro de cada una y se recogió información sobre la preferencia de los candidatos. El resultado se expresa en el siguiente gráfico:



Como se aprecia, el candidato 3 es el que tiene una clara ventaja en la preferencia, mientras que el candidato 1 y 2 están más próximos entre sí. La información podría completarse analizando la composición de la preferencia por conglomerado; nótese que este procedimiento es similar al realizado con los estratos. Continuando con el ejemplo, supongamos que la composición de la preferencia por candidato, queda expresada en la siguiente tabla:

Distribución de la preferencia por candidato

	Conglomerado A	Conglomerado B	Conglomerado C	Conglomerado D	Conglomerado E	Total	Distribución %
Candidato 1	84	322	86	90	400	982	24,55
Candidato 2	400	463	89	5	230	1187	29,675
Candidato 3	316	15	625	705	170	1831	45,775
Total	800	800	800	800	800	4000	

La tabla contiene valiosa información para comprender e interpretar los porcentajes, entre la que se cuenta la cantidad de encuestas realizadas, y el número total de respuestas positivas para cada candidato. Si analizamos la información comparando los distintos conglomerados, podemos observar que el candidato 3 tiene mayor preferencia en la tabla general, pero no así en los conglomerados B y E, donde los otros candidatos tienen mayor aceptación. Esta observación puede hacerse extensiva para analizar la composición de la preferencia por conglomerado de los otros candidatos. Pero, a diferencia del muestreo estratificado, aquí suponemos que los conglomerados son parecidos entre sí, por lo tanto, al analizar la composición de la preferencia del candidato 3, sería erróneo suponer que los conglomerados B y E son distintos al resto.

Muestreo No Probabilístico

En ocasiones no es necesario realizar un muestreo aleatorio o probabilístico, dado que no es el objetivo de la investigación la inferencia a la población. Otras veces, se puede usar este tipo de muestreo en estudios orientados a una parcela específica de una población bien definida. En cualquier caso, el resultado del muestreo no se orienta a obtener datos generalizables.

Muestreo accidental

En el muestreo accidental se procede a tomar para la muestra aquellos individuos que desean participar del estudio. Un ejemplo son las encuestas callejeras, donde los encuestadores solicitan respuestas a transeúntes que prestan su colaboración. Los muestreos accidentales tienen muy poca validez y raramente se desprenden generalizaciones de ellos, aunque son excelentes para la prueba de las herramientas de testeo, tales como los cuestionarios de opinión.

En la actualidad, es muy frecuente ver en los medios de comunicación este tipo de encuestas, y la información es utilizada para valorar la opinión del consumidor. Por ejemplo, en los periódicos por internet, se suele pedir al lector que valore el interés por una nota periodística, en otros sitios que ofrecen ayuda en línea, se pide que se valoren en qué medida le fue útil la ayuda ofrecida, etc. Estos son sondeos basados en muestras accidentales, que no tienen ninguna posibilidad de generalización.

Muestreo por cuotas

La metodología del muestreo por cuotas es similar a la que se utiliza en el muestreo accidental, pero en este caso se debe cumplir con una cuota en las unidades de muestreo. El término cuota se aplica a una característica que define claramente a la población que se desea muestrear. Por ejemplo, si estamos interesados en los alumnos de una escuela, éstos pueden dividirse entre varones y mujeres; si además sabemos que la proporción es 3/1 para mujeres, en el muestreo por cuotas deberemos recolectar una muestra que tenga una mayor proporción de mujeres que de hombres, definiendo la cuota mayor para las mujeres.

En este ejemplo, si se desea obtener una muestra de 40 alumnos, 30 de ellos deberán ser mujeres, que sería la cuota para el sexo femenino. Como se dijo, el procedimiento del muestreo es accidental, de modo que participarán las primeras 30 mujeres, y los primeros 10 varones que den su consentimiento. Ahora bien, por tratarse de un muestreo no probabilístico, puede que no se quiera obtener un muestreo proporcional, por lo cual puede fijarse una cuota similar para los alumnos de ambos sexos. En tal caso, participarán 20 mujeres y 20 varones en la muestra.

Muestreo intencional (por conveniencia u orientado)

En este tipo de muestreo interesa la unidad de muestreo, más que el proceso por el cual se obtiene la muestra. Por ejemplo, es posible que estemos interesados en recoger una muestra de directores de escuelas primarias con más de quince años de trayectoria, o mujeres que ejerzan en cargos ejecutivos. En cualquiera de los casos, el principal interés reside en la información que aporta la unidad de muestreo en un conjunto de variables, luego puede fijarse la cantidad de unidades de muestreo que quieran seleccionarse en función de su disponibilidad.

Bola de nieve (o en cascada)

El muestreo comienza localizando a unos pocos individuos, y mediante estos se intenta localizar otros que permitan completar el tamaño de la muestra establecida. Este tipo de muestreo se emplea muy frecuentemente cuando se hacen estudios con poblaciones a las que resulta difícil acceder o que son muy poco comunes. Por ejemplo, aquellos investigadores interesados en sectas religiosas, encontrarán muy complicado realizar un muestreo probabilístico, pero puede facilitar su tarea de campo si logra contactar un miembro de la secta, y a través de él, llegar a otros. El estudio de los sociolectos es otro ejemplo donde se utiliza este tipo de muestreo. Primero se contacta a un miembro del grupo de interés y a través de él se arma una red de contactos que sean potenciales integrantes de la muestra. La conformación de grupos focales es otro ejemplo en donde utiliza este muestreo.

Comentarios Finales

En los apartados sobre la metodología empleada para obtener una muestra probabilística de la población, suele presentarse el error de muestreo como un estadístico. Este término no debe confundirse con un muestreo erróneo, sino que se refiere al grado de precisión en que la muestra representa la población. El error de muestreo suele expresarse en porcentaje, y se espera que este no supere el 5%. En tal caso, estaríamos seguros en un 95% de que el estadístico calculado en la muestra, coincide con el parámetro poblacional.

El término proviene de la siguiente comprobación empírica: si medimos una variable y promediamos su valor en varias muestras iguales de una misma población, extraídas mediante un muestreo probabilístico; la distribución de los promedios sigue un modelo normal. Tomando como base ese modelo, es factible aproximarse con bastante fiabilidad al nivel de error cometido en la muestra obtenida. El modelo de distribución normal se verá más adelante.

Vale decir entonces que el procedimiento de muestreo probabilístico garantiza que ese error sea el menor posible, y sólo en este tipo de muestreos es posible calcularlo. Por otra parte, los errores en la recogida de una muestra se denominan sesgos muestrales. Existen principalmente dos razones por las cuales se producen sesgos en la muestra, una de las cuales es haber definido mal la población hacia la cual se quiere dirigir la inferencia.

Recuérdese que existen poblaciones acotadas y no acotadas, por lo tanto, una mala definición de la población puede repercutir en que se trabaje sobre límites ficticios de ella. Por ejemplo, si se define como población de interés aquellas personas que han sufrido asalto a mano armada en su vivienda y se toma como base las denuncias realizadas a la policía, deberá tenerse en cuenta que dichas denuncias no contienen el total de la población, dado que varios de esos hechos delictivos no se denuncian. Además, puede que la variable de interés quede moderada por otra variable que no se ha tenido en cuenta, por ejemplo, que las denuncias de esos hechos delictivos sean más frecuentes en ciertas zonas de la ciudad. Pretender una generalización fiable bajo estas circunstancias es riesgoso, puesto que estaríamos cometiendo un error no reconocido en la inferencia.

Otra razón que conduce a sesgos en las muestras, afecta a la definición de estrato y conglomerado. Los estratos deben ser homogéneos en su interior y heterogéneos entre ellos. Por ejemplo, estudiantes de primaria y de secundaria podrían definir dos estratos que cumplirían con ese requerimiento, dado que la diferencia de edad garantiza que tengan una composición desigual entre ellos y sean parecidos en su interior. Pero no ocurriría lo mismo si definiéramos dos estratos como alumnos de tercero y cuarto grado. Es muy probable que, en conjunto, los alumnos tengan más similitudes que diferencias, y en tal caso, ello invalidaría la definición de estrato.

Algo similar ocurre cuando se trazan los conglomerados, se supone que estos pueden ser diferentes en su interior y tener características similares entre ellos. Por ejemplo, si se divide el mapa de Córdoba en nueve conglomerados, siete de los cuales tienen composiciones de familias de clase baja, y los dos restantes contienen el resto de la población. Si se decide tomar una muestra sobre sólo tres conglomerados, lo más probable es que se favorezcan sectores de clase social desfavorecida. En una situación como la descrita, será necesario redefinir los límites de los conglomerados.

Misceláneas

La encuesta del Literary Digest: En 1936 se llevaron a cabo elecciones en EE.UU. y los candidatos a presidentes eran Alf Landon y Franklin D. Roosevelt. La revista Literary Digest llevó a cabo una encuesta y predijo que el nuevo presidente sería Alf Landon. Los resultados de la elección dieron por ganador a Franklin D. Roosevelt por una gran mayoría. ¿Qué fue lo que falló? Como luego se evidenció, la revista envió la encuesta por correo a 10 millones de ciudadanos estadounidenses, la cual es una muestra lo suficientemente grande como para ser representativa, pero solo contestaron la encuesta 2.3 millones de votantes, lo cual representa solo el 23% de los encuestados.

Esta cifra es demasiado pequeña como para avanzar alguna predicción fiable del resultado de la votación. Aun así, los directivos del Literary Digest se aventuraron a predecir un resultado electoral, en el que se daba por ganador al candidato Alf Landon. La pérdida de credibilidad de la revista y la repercusión del fracaso en las encuestas electorales, condujo a que se modificaran sustantivamente las técnicas de muestreo en los sondeos de opinión electoral. Una cuestión que quedó claro fue que: no es el tamaño de la muestra lo que garantiza su representatividad, sino el método empleado para recogerla.

La paradoja del candidato: hay importantes cuestiones sobre las elecciones por sistemas de votación que han sido estudiadas por matemáticos y estadísticos, tales como Richard, G. Nemi, William H. Riker, Donald Saari y Allyn Jackson. Una de ellas se refiere a la paradoja del candidato la cual expondremos brevemente. Supongamos que hacemos competir en una elección presidencial al partido del CENTRO, el de DERECHA y el de IZQUIERDA; ¿es posible que de antemano arreglemos la elección para que gane uno de ellos? En teoría no sería posible porqué depende del caudal de votos que pueda captar cada partido, pero analicemos esta situación hipotética. Se les pide a los votantes que ordenen a los candidatos en relación a su preferencia, y de ello resulta que:

a) 1/3 de la población prefiere C – D – I,
b) 1/3 de la población prefiere D – I – C, y
c) c) 1/3 de la población prefiere I – C – D.

En el caso a) decimos que, un tercio de la población prefiere al candidato de centro, por sobre el candidato de derecha y a este por sobre el candidato de izquierda. Similar interpretación se aplica en los casos b) y c). Supongamos que deseamos que el candidato favorecido sea el de izquierda, entonces en primera vuelta hacemos competir al candidato de centro y de derecha. De esta elección resulta favorecido el candidato del centro con 2/3 de los votos, y al candidato de derecha lo elige solo 1/3 de la población. En segunda vuelta hacemos que compitan el candidato de centro con el de izquierda, y tenemos que 2/3 de la población prefiere al candidato de izquierda por encima del de centro, mientras que solo 1/3 de la población prefiere al candidato del centro por sobre el de izquierda. El resultado de estas dos vueltas electorales ¡consagra ganador al candidato de izquierda! Lo interesante de este juego es que es posible hacer lo mismo con los otros candidatos.

Esta paradoja es posible porque en realidad no existe una relación transitiva entre los candidatos de centro, izquierda y derecha. Las relaciones transitivas solo se verifican cuando es posible ordenar atributos o personas con cuantificadores tales como mayor que/menor que o similares. En tal caso, si decimos que A>B (A es mayor que B) y luego B>C (B mayor que C), entonces se deduce que A es mayor que C por propiedad transitiva. De ello se desprende que esta propiedad no puede aplicarse cuando la relación entre atributos es del tipo, “prefiere a”, dado que este no es un cuantificador. Es decir, si solo se trata de ordenar por preferencia, se puede dar que alguien prefiera A sobre B, B sobre C, pero a C sobre A. La paradoja en este caso se conoce como Paradoja de Arrow, en honor a Kenneth J. Arrow, quien postulo que un sistema de votación democrática perfecto es imposible.

Guía de Actividades N°4:

Población Muestra y tipos de muestreo

Objetivos de la actividad:
a) comprender conceptos básicos sobre población y muestra,
b) identificar los distintos tipos de muestreo posibles sobre una población

Actividad 1: Para los siguientes ejemplos, deberá identificar la/s variable/s, su escala de medición, la unidad de análisis en estudio y el muestreo utilizado:

a) Mediante un relevamiento muestral en Córdoba Capital, y como parte de la Encuesta Permanente de Hogares (EPH), se ha recolectado información del ingreso mensual de los hogares, el Gran Córdoba. Dicha variable se midió en monto total de dinero en pesos, percibido por el Jefe de Hogar.

Variable/s: Escala:
UA:
Muestreo:

b) Un sindicato desea conocer el nivel de satisfacción que tienen sus afiliados con los servicios de su secretaría de deportes. Para ello solicita a un conjunto de personas que asiste regularmente a las instalaciones recreativas, que puntúen de 1 a 10 los servicios de esa secretaría.

Variable/s: Escala:
UA:
Muestreo:

c) El servicio de salud mental del Instituto Provincial de Alcoholismo y Drogadicción de Córdoba, desea conocer las siguientes características de la población que atiende diariamente: 1) edad en años; 2) género; 3) motivo de consulta; 4) profesión y 5) lugar de residencia. Utiliza como muestra, las diez primeras personas que asisten a consulta los días hábiles de la semana.

Variable/s: Escala:
UA:
Muestreo:

Actividad 2: En las siguientes situaciones se identifica una variable de interés juntamente con su escala de medición. Al mismo tiempo, se indica de donde se extrae la muestra. Se pide que proponga un tipo de muestreo posible para cada una de las situaciones, de acuerdo a la variable que se intenta medir:

a) Se desea conocer el consumo de gaseosas medido en monto de la compra en pesos. Se consulta a las personas que compran en un hipermercado, seleccionando 10 individuos en cada uno de los siguientes rangos de edad: 1) 25 – 35; 2) 35 – 45; 3) 45 – 55; 4) 55 – 65.

Variable	**Escala de medición de la variable**	**Tipo de muestreo posible**

b) Se necesita conocer el índice de estrés laboral de operarios de la industria textil, medido en una escala de 1 al 10 (siendo 1 muy estresado y 10 sin estrés). Se seleccionan aleatoriamente operarios de los siguientes rangos de antigüedad en el puesto: 1) 0 – 5, 2) 5– 10, 3) 10 o más años.

Variable	**Escala de medición de la variable**	**Tipo de muestreo posible**

c) La municipalidad de Carlos Paz quiere saber la opinión de los vecinos sobre la calidad del servicio de recolección de basura, medido mediante la siguiente escala: Malo – Regular – Bueno –Muy Bueno – Excelente. Se realiza un muestreo aleatorio de vecinos de barrios con diferentes ubicaciones geográficas, de la Ciudad de Villa Carlos Paz.

Variable	**Escala de medición de la variable**	**Tipo de muestreo posible**

Actividad 3: En todas las situaciones presentadas se supone que está garantizada la generalización de los datos a la población de interés de acuerdo a la variable en estudio. Usted deberá decir si se justifica o no esta posición. Argumente de acuerdo a lo que conoce sobre muestreo:

a) Se desea conocer la altura en centímetros en la población de niños entre

4 y 6 años residentes en la Ciudad de Córdoba. Mediante un procedimiento de muestreo aleatorio sistemático, se extrae una muestra de 400 niños de 20 jardines de infantes de esa ciudad.

b) Se desea conocer la opinión de los directores de IPEM de la Provincia de Córdoba con 10 o más años de antigüedad en el cargo, respecto de la reforma a aplicarse en el nivel medio. Mediante un procedimiento de muestreo intencional, se seleccionan un total de 40 directivos con esas características.

c) Se desea conocer el estado general de las aulas de los establecimientos educativos de nivel primario de gestión oficial de la ciudad de Córdoba. Se seleccionaron 550 aulas para observar, en un procedimiento por etapas que consistió de; etapa 1: selección aleatoria de los establecimientos; etapa 2: selección aleatoria de las aulas de los establecimientos previamente identificados.

d) Se desea conocer los tiempos de reacción en individuos luego de haber ingerido bebidas alcohólicas. Se seleccionaron 40 personas en un control de alcoholemia cuyo nivel de alcohol en sangre superaba 0,2 en la prueba de alcoholemia. Estos individuos participaron de una prueba de discriminación de estímulos visuales.

e) Se desea conocer el nivel educativo de los padres de los ingresantes a la facultad de Ciencias Químicas de la UNC según los estudiantes provengan de la provincia de Córdoba, la región noroeste y la región noreste. Según las regiones geográficas se realizó un muestreo aleatorio sistemático entre los estudiantes de cada una.

f) Se desea conocer aspectos relacionados a la motivación de la elección de la carrera docente en los alumnos de 1° año del profesorado de matemáticas del Instituto del Profesorado Simón Bolívar de la ciudad de Córdoba. Se sabe que en dicho profesorado cursan 250 alumnos, de los cuales 50 alumnos participaron voluntariamente de una encuesta diseñada a tal fin.

Actividades en PSPP N° 4:

Vamos a abrir nuevamente el programa PSPP, pero en esta ocasión vamos a trabajar con una base de datos que descargaremos del Instituto Nacional de Estadísticas y Censos (INDEC). Para ello ingresamos a su página web, https://www.indec.gob.ar/. Una vez dentro de la página web podrá identificar, al final de la página, un enlace que dice "Bases de datos" y dentro de él una sección que se titula "Encuesta de Actividades de niños, niñas y adolescentes 2016-2017".

Dentro de esta sección puede a ingresar a "Documentos metodológicos" y allí proceder a descargar "EANNA urbana. Manual de uso de las bases de datos" que es el documento metodológico que utilizaremos en las sucesivas actividades.

Luego ingresamos a "Bases de microdatos de la EANNA urbana" allí a "Base individual (cuestionario 2)", donde vamos a descargar el archivo "Formato spss (.sav)" como se ve en la siguiente imagen. Se recomienda descargar ambos archivos, el manual y la base, en una misma carpeta de ejercicios de la materia.

Encuesta de Actividades de Niños, niñas y adolescentes 2016-2017

La Secretaría de Gobierno de Trabajo y Empleo (Ministerio de Producción y Trabajo), a través de la Dirección General de Estudios Macroeconómicos y Estadísticas Laborales, conjuntamente con el Instituto Nacional de Estadística y Censos (INDEC) realizaron la Encuesta de Actividades de Niños, Niñas y Adolescentes (EANNA) 2016-2017. Se trata de la segunda encuesta de este tipo aplicada en el país y la primera de carácter nacional, ya que cubre a toda la población de la Argentina, tanto la residente en zonas urbanas como en áreas rurales. A continuación se presentan las bases de microdatos de ambos dominios. Las mismas tienen un carácter preliminar en cuanto a la cantidad de variables presentadas, ya que contienen variables básicas referidas a la vivienda, integrantes del hogar y características educativas y de realización de actividades productivas de los niños, niñas y adolescentes de 5 a 17 años (NNyA). Adicionalmente, se pone a disposición de los usuarios una base de pesos replicados, para poder calcular las estimaciones de los errores de muestreo correspondientes a los resultados de la EANNA urbana y EANNA rural.

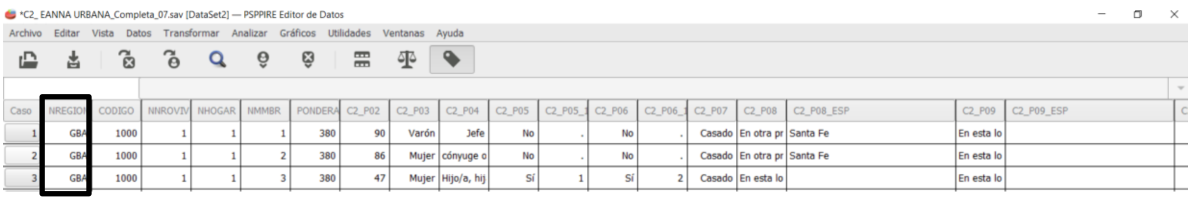
Documentos metodológicos ▾

- EANNA urbana. Manual de uso de las bases de datos
  - Formato pdf
  - Formato xls
- EANNA rural. Manual de uso de las bases de datos
  - Formato pdf
  - Formato xls
- Nota técnica EANNA. Factores de expansión, estimación y cálculo de los errores por muestra para el dominio urbano
- Factores de expansión, estimación y cálculo de los errores por muestra para el dominio rural
- Cuestionarios de la EANNA

Bases de microdatos de la EANNA urbana ▾

- Base de hogares (cuestionario 1)
  - Formato spss (.sav)
  - Formato csv
- Base individual (cuestionario 2)
  - Formato spss (.sav)
  - Formato csv

Luego procedemos a descomprimir la base de datos (automáticamente la descarga en formato ZIP) y a abrirla con el programa PSPP. Como hemos aprendido, aplicamos la función etiquetas de los valores y comparamos las columnas con las respuestas en el instructivo:



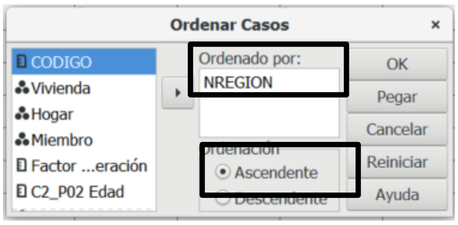
Caso	NREGION	CODIGO	NNROVIV	NHOGAR	NMMBR	PONDERA	C2_P02	C2_P03	C2_P04	C2_P05	C2_P05_1	C2_P06	C2_P06_1	C2_P07	C2_P08	C2_P08_ESP	C2_P09	C2_P09_ESP
1	GBA	1000	1	1	1	380	90	Varón	Jefe	No	.	No	.	Casado	En otra pr	Santa Fe	En esta lo	
2	GBA	1000	1	1	2	380	86	Mujer	cónyuge o	No	.	No	.	Casado	En otra pr	Santa Fe	En esta lo	
3	GBA	1000	1	1	3	380	47	Mujer	Hijo/a, hij	Sí	1	Sí	2	Casado	En esta lo		En esta lo	

De este modo, podemos observar que la primera columna responde a la primera pregunta del cuestionario 2, que está codificada con los valores 1, 40, 41, 42, 43, y 44. Para cada valor corresponde una etiqueta, por ejemplo, para el valor 42 corresponde la Región Cuyo:

CUESTIONARIO 2

Nombre de la variable	Descripción		Universo
	Código de región		
	1	Región GBA	
	40	Región NOA	
NREGIÓN	41	Región NEA	
	42	Región Cuyo	
	43	Región Pampeana	
	44	Región Patagonia	

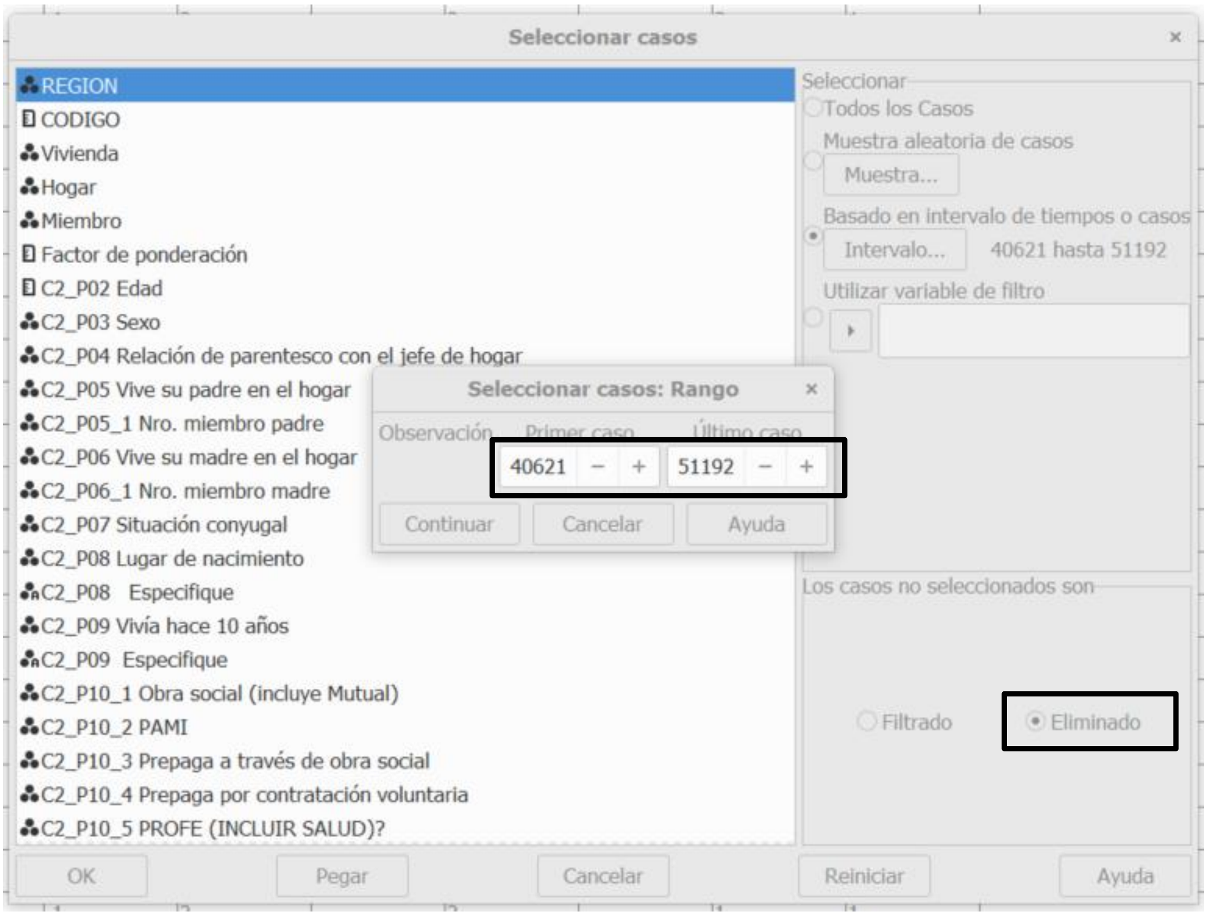
Dado que nos interesa trabajar con la región en la que se encuentra Córdoba, vamos a recortar (dejar sólo los casos) la base de datos para esa región. Primero ordenaremos la base según la columna NREGION por orden alfabético. Para ello nos dirigirnos al comando Datos, y seleccionamos la opción Ordenar casos. En el cuadro que aparecen las variables para ordenar la base, seleccionamos NREGIÓN y la llevamos hacia la derecha con la flecha, de este modo indicamos al programa que ordene Ascendente según esa variable.



Le damos a OK y la base estará ordenada de menor a mayor según la variable NREGION. Una vez ordenada la variable, vamos a proceder a cortar los casos para la región 42 que, como dijimos, es el código de la región Centro.

Para ello debemos identificar la fila en que comienza el valor 42, que en este caso es la 40621, y la fila donde termina este valor, que es la 51192.

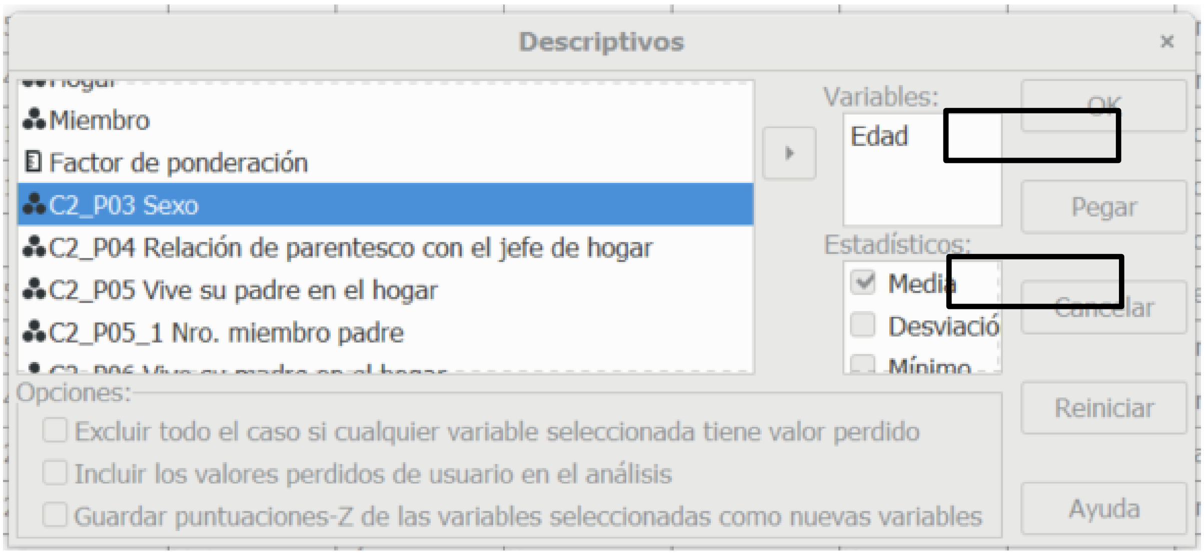
Una vez que hemos identificado estas filas procedemos a abrir el menú Datos, vamos a la opción Seleccionar casos, y seleccionamos la opción Basado en intervalo de tiempo o casos, allí seleccionamos donde dice Intervalo… y en Primer caso colocaremos 40621 y en Último caso, colocaremos 51192, seleccionamos Continuar y en el cuadro En los casos no seleccionados son… ponemos Eliminado, finalmente damos sobre OK.



Ya tenemos nuestra base de datos para la región Centro. Podemos verificarlo con observar que la variable NREGION ahora dejó de ser una variable y pasó a ser una constante, donde el único valor es 42. Grabaremos nuestra nueva base recortada con el nombre C2_EANNA URBANA_CENTRO.SAV. ¡No olviden grabar regularmente!

Por ahora vamos a *suponer* que todos estos datos son nuestra población de estudio. Decimos *suponer* porque en realidad estos datos son producto de una muestra representativa de los aglomerados urbanos de la región Centro, y no consideran en resto de la región. Pero para este ejercicio vamos a hacer como si estos casos fueran todos los datos que queremos analizar, por lo tanto, nuestra población, para evitar operaciones más complejas que veremos más adelante.

Entonces, pediremos el promedio de la edad. Para ello volvemos al manual y como pueden apreciar la variable que buscamos se llama C2_P02. Una vez identificada en la base de datos, iremos a Vista de variables y modificaremos su nombre por el de “Edad”. Luego iremos al comando Analizar, Descriptivos, seleccionamos la variable C2_P02 Edad y pedimos la Media:



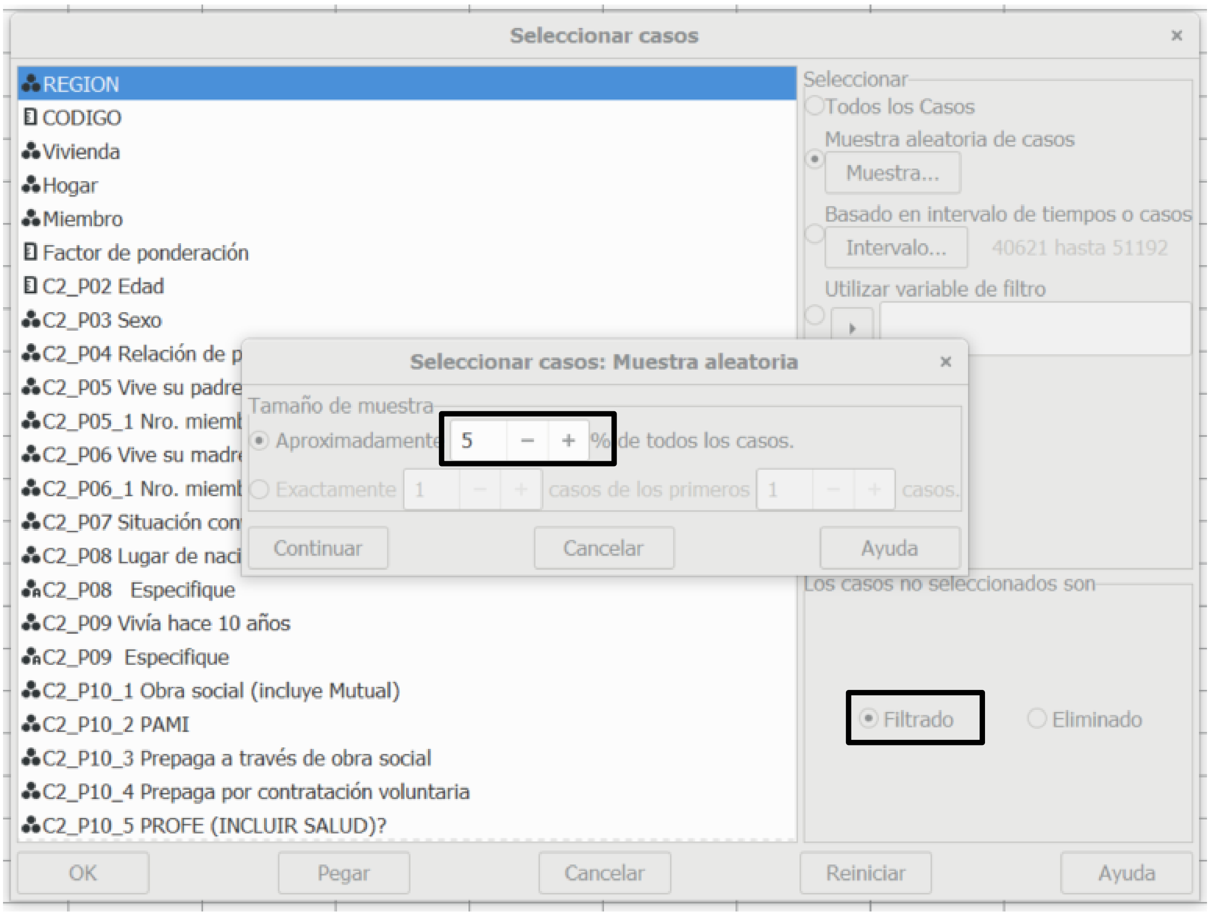
Damos OK y veremos en la ventana de resultados un cuadro que dice que la variable es C2_P02 Edad, la cantidad de casos N (nuestra población) es 10572 y la Media es 35.28 años.

Variable	*N*	*Media*
C2_P02 Edad	10572	35,28

Esta es la Media Aritmética o Promedio de la variable Edad de nuestra población. Nuestro próximo objetivo será seleccionar una muestra aleatoria de esta población. Buscamos estimar la Edad de lo que acordamos es nuestra población a partir de un conjunto menor de casos. De este modo podremos ver qué tan próximo está un estimador de una muestra seleccionada por nosotros del parámetro poblacional. Recordemos que un estimador y un parámetro deben aplicar la misma medida resumen, como por ejemplo el promedio o media aritmética, que cuando se obtiene de la población se llama parámetro y cuando se calcula de la muestra se denomina estimador; porque el promedio de la muestra, si está es representativa de la población, nos permite estimar el parámetro poblacional, o dicho en otros términos con el promedio de la muestra podremos aproximarnos a la edad promedio de toda la población.

Para ello vamos a Datos, marcamos Los casos no seleccionados son… Filtrado (si no hacemos esto el programa eliminará los casos que no correspondan a la muestra), marcamos Seleccionar casos, seleccionamos el botón Muestra, seleccionamos Aproximadamente y colocamos un 5% de los casos.

Luego damos a OK:



Ya tenemos una muestra que ha sido seleccionada por un algoritmo de modo aleatorio en nuestra base de datos, que ha dado por resultado que queden activos sólo el 5% de los casos. Nuevamente vamos al comando Analizar, Estadística descriptiva, seleccionamos Edad; ya estará seleccionado por defecto Media, damos OK y veremos los resultados sobre la misma variable con una cantidad de casos mucho menor, 523 casos contra 10572, con una media de 35.24 años.

Variable	N	Media
C2_P02 Edad	523	35,34

No se preocupen si les da un promedio levemente diferente, dado que la selección se realiza de modo aleatorio, siempre las muestras serán diferentes en cada muestreo y el promedio puede variar un poco en cada procesamiento, pero podemos ver que nuestro estimador de la población, a partir de la selección aleatoria de 535 casos de un total de 10572, está bastante próximo al valor de parámetro poblacional, 35.34 años contra 35.28 años respectivamente.

La diferencia entre estos dos es lo que se llama error del estimador de la muestra. La diferencia en este procesamiento dio de 0.07 años, aproximadamente 26 días. Imagine que, si tienen que realizar un relevamiento de 10572 casos, el costo será ampliamente mayor al de 535 casos, por lo que ustedes como investigadores deberán evaluar si el margen de error justifica tal inversión.

No hemos trabajado aquí como calcular el error de estimación del parámetro poblacional, ni cuántos casos son suficientes para una buena estimación. Todo ello puede ser calculado perfectamente, pero excede este manual.

De todos modos, ustedes han podido realizar una muestra de un conjunto de datos y comparar un estimador con un parámetro poblacional de manera intuitiva.

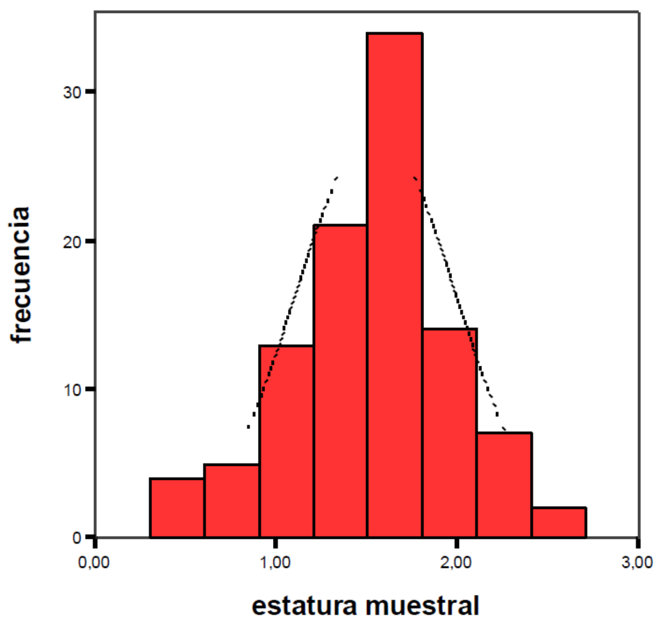
Capítulo 5

La distribución normal

En nuestra vida cotidiana estamos utilizando casi todo el tiempo el concepto de normalidad estadística, aunque no somos conscientes de ello. Tal concepto nos permite hacer juicios de valor y establecer generalidades acerca de aspectos de la vida que son los más frecuentes de observar, y nos permite también comprender fácilmente a qué nos referimos cuando valoramos un evento o acontecimiento como muy poco probable.

En este capítulo vamos a trabajar el modelo teórico matemático de distribución normal estandarizada, la vamos a analizar a partir de ejemplos de distribuciones empíricas que se comportan de acuerdo a este modelo y estudiaremos cómo podemos pasar de los valores observados en las distribuciones empíricas a los valores propios del modelo de distribución normal estandarizada que se denominan valores en puntuación Z. También veremos las formas diferentes que toman estas distribuciones empíricas en función de su simetría/asimetría y de su forma más o menos puntiaguda.

Vamos a tomar una variable aleatoria en la población, tal como es la estatura de las personas. Decimos en este caso que la variable es aleatoria porque en cada medición que realicemos, el valor puede variar libremente. Sin embargo, la mayoría de las medidas tomadas en la población, tenderán a centrarse en un determinado valor, el cual aparecerá como el más frecuente. Para este ejemplo, tomamos datos de salud pública del Ministerio de Salud de la Nación del año 2010. Seleccionamos accidentalmente 100 personas. Con estos datos construimos el siguiente histograma.



Además, calculamos algunos estadísticos sobre la muestra que se presentan en la siguiente tabla.

Estadístico de la variable estatura muestral:

Media	**Mediana**	**Modo**	**Desviación estándar**	**Varianza**	**Máximo**	**Mínimo**
1,54	1,64	1,70	0,44	0,19	2,53	0,6

Nota. Los valores están expresados en metros.

En este ejemplo el valor modal de la distribución es 1,70 mts. que está representado por el punto más alto de la distribución, luego vemos que las estaturas bajas son menos frecuentes, lo mismo que las estaturas altas. El centro de la distribución está próximo al modo ya que el promedio es de 1,54 mts. para esta muestra.

Basándonos en la información que tenemos y en el gráfico diremos que aquellos individuos por debajo de 1,7 mts. deberían considerarse bajos, mientras que aquellos por encima de esa estatura se consideran individuos altos. En la tabla de frecuencia original, aparecen tres casos con una estatura de 0,60 mts. que consideramos demasiado bajos. También aparece un caso con una estatura de 2,53 mts., que consideramos como demasiado alto. Estadísticamente diremos que estos casos son atípicos o extremos dentro de la distribución que estamos analizando.

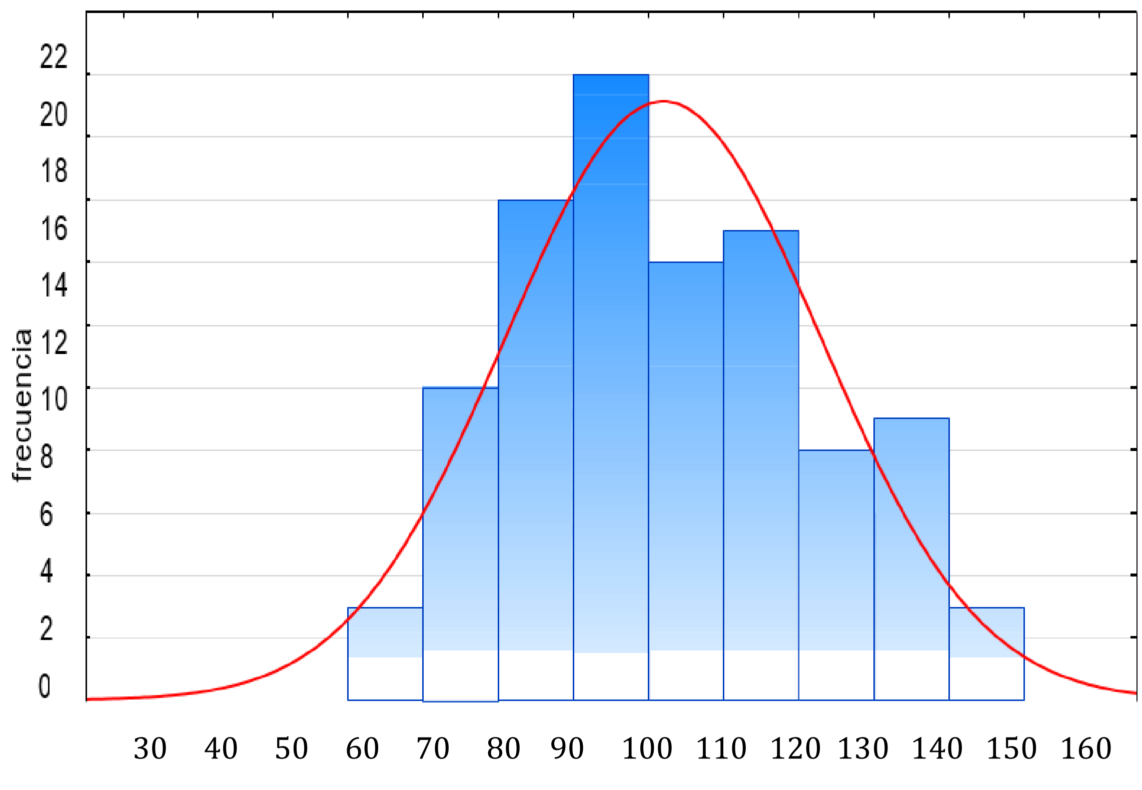
Sobre el histograma hemos superpuesto una curva con forma de campana, conocida como distribución normal. Esa distribución corresponde a una curva teórica en donde las estaturas se acumulan proporcionalmente en torno a una media de 1,54 mts. y una desviación estándar de 0,44. Vemos que la superposición entre la curva normal y el histograma coinciden en varios puntos, por lo cual es factible utilizar el modelo de distribución normal para la variable estatura en esta muestra.

Resumiendo lo dicho hasta aquí, tenemos que lo que nos informaría esta curva para la variable estatura es que la mayoría de las personas están centradas en el valor 1,70 mts, luego los más altos se acumularían hacia la derecha de la distribución y los más bajos hacia la izquierda. Como los individuos altos y muy altos (como así también los bajos y muy bajos), no son comunes en la población, tenemos que su frecuencia disminuye a medida que nos alejamos del valor central, lo que confiere a la distribución su típica forma acampanada. Debemos tener siempre presente que la altura de la curva indica frecuencia.

Volvamos ahora sobre los casos atípicos o extremos: ¿cómo explicar que en una población existan casos de individuos con 60 centímetros de estatura y otro con 2,53 metros? Recordemos que la muestra es accidental y obtenida de un listado de salud pública del Ministerio de Salud de la Nación del año 2010. Ese listado incluye persona de todas las edades y al cotejar la edad de los individuos de 60 cm. de estatura encontramos que son menores de un año de edad. Por lo tanto, no pueden considerarse como casos atípicos.

¿Qué hay de una persona de 2,53 mts? En el listado original no había ninguna persona de esa estatura, con la intención de mostrar un caso verdaderamente atípico incluimos la estatura de Trijnje Keever, quien se supone fue la mujer más alta de los Países Bajos, y según se reporta vivió solo 17 años. En el listado original la estatura promedio de quienes tienen 17 y 18 años es de 1,69 mts. por lo cual encontrar una mujer de 17 años de 2,53 sí puede considerarse como un caso excepcional o atípico.

La distribución normal tiene importantes propiedades para la estadística descriptiva e inferencial, dado que puede utilizarse como modelo matemático para analizar variables aleatorias relevantes para la investigación educacional. El siguiente gráfico puede ilustrar lo dicho dado que se trata de un conjunto de valores generados aleatoriamente tomando como referencia una media igual a 100 y varianza igual a 500. Lo interesante de esta simulación es que el patrón de valores individuales fue generado en función de los resultados de la prueba PISA aplicada en 2012 para matemáticas.



Nótese que este simple ejercicio genera un patrón de datos que puede ser analizado bajo los postulados de la distribución normal estandarizada. En el apartado que sigue nos ocuparemos de esta propiedad de la distribución normal.

La distribución normal como modelo matemático: la normal estandarizada

Antes de comenzar este apartado es importante reconocer que las propiedades del modelo matemático implícito en la distribución normal estandarizada, son aplicables a una distribución empírica solo cuando ésta última se aproxima a la normal. Aunque no es objeto de desarrollo aquí, existen procedimientos de transformación de las puntuaciones originales de una distribución empírica para que se empareje con una distribución normal y así utilizar sus propiedades. Un ejemplo de esto lo vimos anteriormente cuando simulamos una muestra aleatoria de las puntuaciones PISA con una media de 100 y varianza de 500.

La distribución normal es unimodal y simétrica, lo cual significa que la media, la mediana y el modo coinciden en el mismo valor, que es el centro de la distribución. La simetría indica que la distribución es exactamente igual hacia ambos lados de la media.

La distribución normal es continua y asintótica (esto es, que se extiende indefinidamente en sus extremos), esto quiere decir que cualquier valor de la variable estaría contenido dentro de la distribución normal estandarizada. Las variables empíricas, como sabemos, no contienen infinitos valores, por lo cual, dentro de ciertos límites una distribución normal contendrá el 100% de los valores de esa distribución empírica. Esto quiere decir que la varianza de una distribución empírica es finita; otra manera de expresarlo es diciendo que todo conjunto de datos tiene un valor máximo y un mínimo. El Teorema de Chebyshev proporciona una regla empírica que determina que dado un conjunto de valores que se asemeje a una distribución normal, aproximadamente el 90% de los valores de esa distribución estará comprendida entre dos desviaciones estándar por encima y por debajo de la media. Los detalles de este teorema están desarrollados en la desigualdad de Chebyshev, pero no lo desarrollaremos en este tutorial.

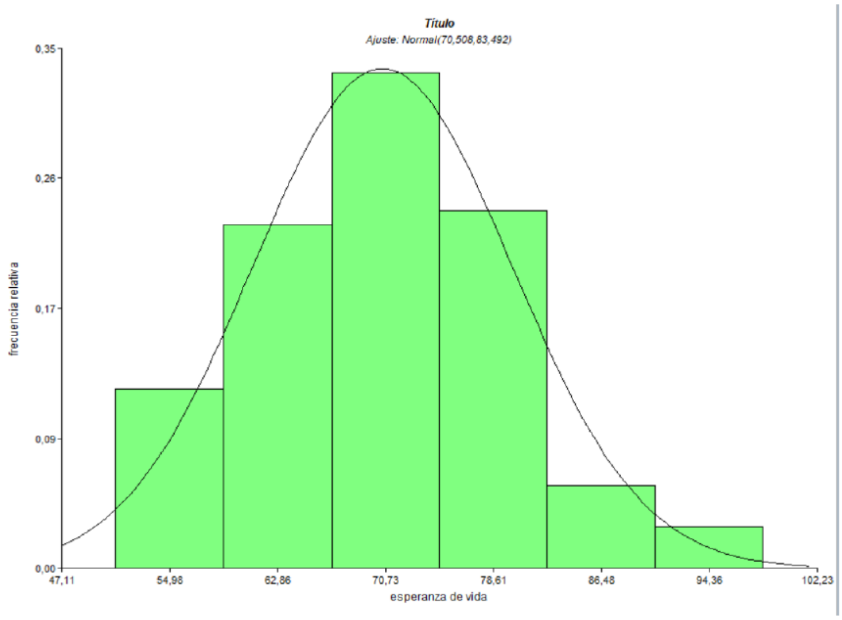
Los valores de media y desviación estándar en una distribución normal estandarizada están expresados en puntuación Z. De esto se deriva que la media de una distribución normal estándar será siempre cero y la desviación estándar será siempre uno. Esta es una propiedad muy importante dado que cualquier distribución empírica puede ser estandarizada, transformando sus puntuaciones originales en puntuación Z aplicando la siguiente fórmula:

$$Z = \frac{x - \bar{x}}{DE}$$

Veamos como ejemplo la siguiente tabla extraída del ministerio de Salud de la Nación del año 2010, en la que se obtuvo una muestra aleatoria de 109 casos donde se registró la esperanza de vida. La distribución obtenida se aproxima a la distribución normal, por lo tanto, es lícito realizar la transformación Z.

Esperanza de vida en la población

Casos	media	DE	mínimo	máximo
109	70.15	10.57	43	82



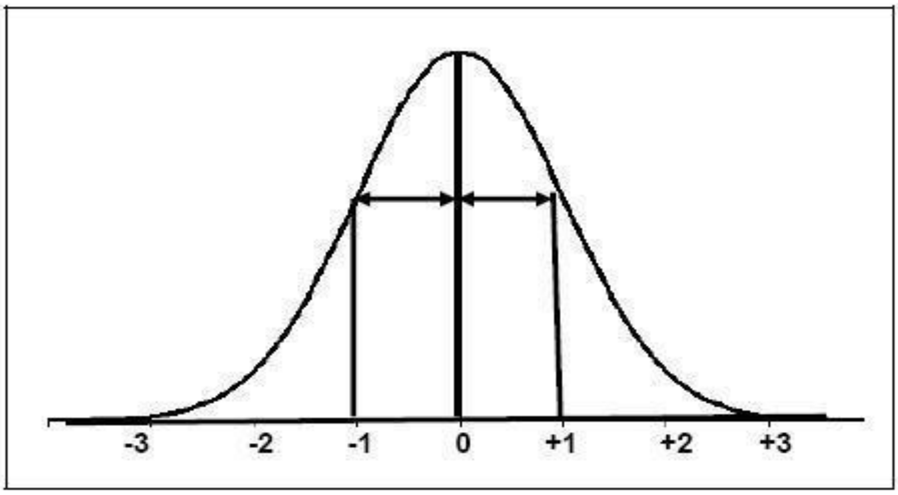
Como se aprecia en el gráfico, la distribución empírica se adapta a una distribución normal estándar. Del total de casos analizado elegimos tres y transformamos sus puntajes a puntuación Z, tal como lo muestra la siguiente tabla

Puntaje original	Puntaje Z
79,00	0,84
80,00	0,93
53,00	-1,62

Los puntajes representados en la tabla son equivalentes, la diferencia radica en que el puntaje original está centrado en la media de 70,15 y la transformación Z arroja una media de 0. Veamos ahora cómo queda la tabla si calculamos los mismos estadísticos sobre la distribución esperanza de vida a valores Z.

Casos	media	DE	mínimo	máximo
109	70.15	10.57	43	82
109	0.00	1.00	-2.57	1.12

De esta forma la distribución de puntuaciones originales transformadas de una distribución empírica, queda asimilada en la siguiente distribución, cualquiera sea la puntuación original. Como vimos, la única condición para esto es que la variable empírica se aproxime con fidelidad a una distribución normal teórica.



Volvamos un instante al ejemplo de la distribución de las estaturas. Haciendo un recorte de la muestra de estaturas para las personas que tiene 17 y 18 años, tenemos que la media es de 1,69 mts. y la desviación estándar es de 0,22 mts. ¿Qué tan diferente resulta la estatura de Trijnje Keever? Si observamos el gráfico anterior vemos que la mayoría de las personas se encuentran entre ±1 desviación estándar de la media, más infrecuentes son las que se apartan ±2 desviaciones estándar y aquellas a ±3 desviaciones estándar son sin duda casos atípicos, o muy raros. La estatura de Trijnje Keever se aparta 3.81 desviaciones estándar de la media.

$$Z = \frac{2{,}53 - 1{,}69}{0{,}22} = 3.81$$

Trijnje Keever murió a los 17 años en 1633 y hasta el presente no se ha registrado una altura igual en una mujer.

Sobre la distribución estandarizada podemos trabajar a la inversa. Es decir, conociendo los valores en puntuación Z, podemos descomponer el valor original. Recordemos del ejemplo de la esperanza de vida de la población que su media era de 70,15 años y su desviación estándar de 10,57 años. Para obtener los puntajes originales tendríamos que emplear la siguiente ecuación:

$$Po = \bar{x} + (Z * de)$$

Siendo Po la puntuación original, la tabla a continuación muestra el cálculo de transformación de puntaje Z a puntuación original.

Puntaje Z		Po
0,84	70,15+(0,84*10,57)	79,0
0,93	70,15+(0,93*10,57)	80,0
-1,62	70,15+(-	53,0

La distribución normal estandarizada y la proporción de casos

Hasta aquí hemos podido ver que, mediante la estandarización de una distribución empírica, hacemos razonable y comprensible los términos de normal o típico, o bien, extremo o atípico. También podemos trabajar la proporción de casos que son típicos y atípicos, mediante la utilización de las áreas bajo la curva normal en término de proporciones.

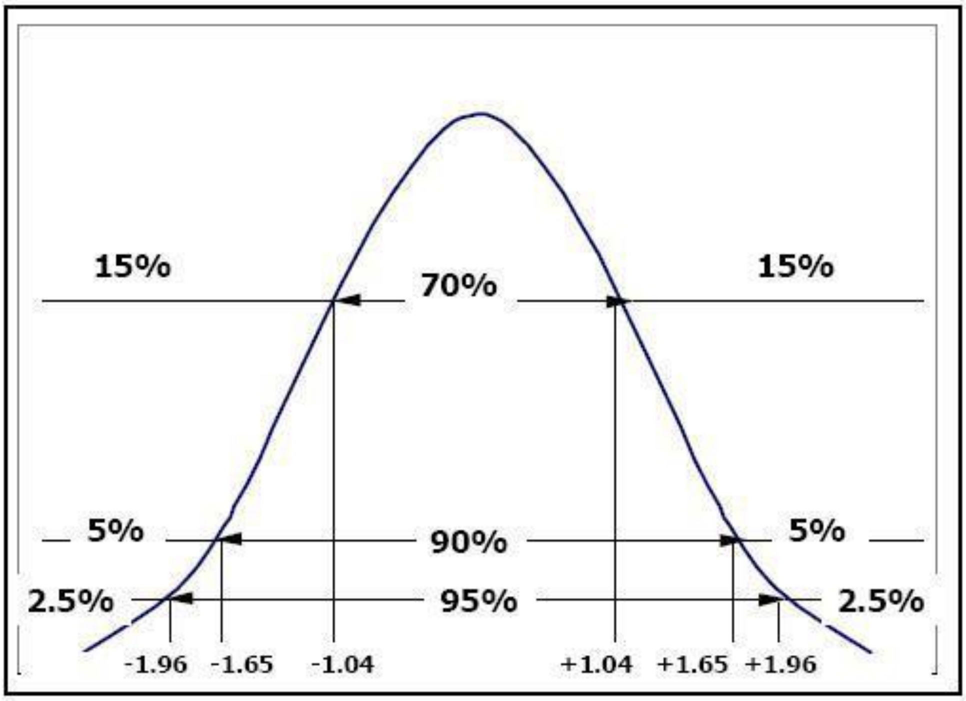
La siguiente figura nos muestra en el eje de las abscisas las puntuaciones estandarizadas Z, las cuales se proyectan sobre la curva y en el punto donde se intersectan, se define un área bajo la curva. De este modo, si tomamos el segmento comprendido entre los valores Z= ±1.04, tenemos que se ha cubierto el 70% del área bajo la curva. Esto tiene una utilidad fundamental para comprender el comportamiento de variables empíricas en poblaciones e intentaremos mostrarlo mediante un ejemplo.

Supongamos que se ha tomado una prueba de rendimiento en lectura en la provincia de Córdoba, cuya distribución de valores se aproxima a una distribución normal. La media en dicha prueba es de 550, y la desviación estándar es de 125. La curva que se muestra en la figura que sigue podría utilizarse como modelo dimensional de esa variable y mediante algunos simples cálculos podríamos contestar las siguientes preguntas:

1 ¿Qué puntuaciones originales de la prueba comprenden el 70% de la población, considerada como rendimiento medio?

2 ¿Qué puntuación le corresponde al 5% de la población con el rendimiento más bajo?

3 ¿Qué puntuación puede considerarse excepcionalmente alta en la prueba?



Nota. Figura publicada en: Morales Vallejo, Pedro (2008) Estadística aplicada a las Ciencias Sociales. Madrid: Universidad Pontificia Comillas

Para responder a la primera pregunta debemos aplicar la fórmula mediante la cual transformamos la puntuación Z en puntuación original. Los valores Z que nos informan qué valores contienen al 70% de la población con puntuación en los valores medios de la distribución son, como vimos ±1,04. Pasándolos a puntuación original tenemos que el 70% de la población con puntuación en los valores medios de la prueba han obtenido puntajes de 420 como mínimo y 680 como máximo.

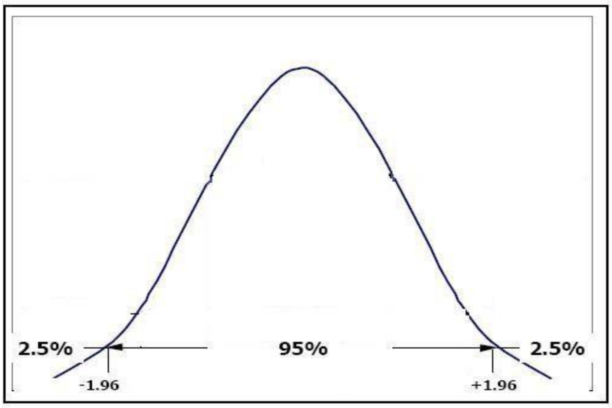
Para responder a la segunda pregunta debemos considerar el valor Z que deja por debajo el 5% del total de valores contenidos en la distribución. Dicho de otro modo, debemos buscar el valor correspondiente al punto de corte para el 5% hacia la izquierda de la distribución; ese valor es -1.65. Nuevamente, transformando ese valor a puntuación original tenemos que la puntuación original de prueba obtenemos el valor 343,75≃344 (por redondeo). Por lo tanto, todos aquellos alumnos que hubieran obtenido esa puntuación o menor, estarían en el 5% de la población con más bajo rendimiento.

Por último, una puntuación excepcionalmente alta podría considerarse aquella que corresponde al 1% de la población con más alto rendimiento. Repitiendo el procedimiento, tenemos que el valor Z de 2,34 deja por debajo de al 99,04% de la distribución. Es decir que por encima de ese valor se encuentra el 1% de la población aproximadamente (este valor no consta en la figura, se extrajo de una tabla que describe los valores Z y el porcentaje de área correspondiente). Transformando ese valor Z a puntuación original tenemos que aquellos individuos con un puntaje de

842,5 o más en la prueba de lectura, serían los que ocuparían el 1% de la distribución con rendimiento excepcionalmente alto.

Áreas bajo la curva como distribución de probabilidades

Más adelante veremos cómo la distribución de probabilidades sirve a distintos propósitos, especialmente para la prueba de hipótesis, por esto merece una referencia el porcentaje de área bajo la distribución normal como distribución de probabilidades. Para ello vamos a prestar especial atención al valor Z de la curva ±1,96. Veamos la siguiente figura:



Vemos que los valores ±1,96 contienen el 95% de casos de la distribución, y dejan fuera el 5% de los restantes casos. Ese 5% se reparte entre aquellos que ocupan ambos extremos de la distribución, por tanto, se cuentan 2,5% en la parte izquierda (aquellos que están por debajo del valor Z -1,96), y un 2,5% en la parte derecha (aquellos que están por encima del valor Z +1,96).

Si interpretamos esos porcentajes en término de probabilidades, y tenemos que escoger un valor al azar del total de valores de la variable en esa distribución, es muy probable que el valor escogido sea mayor o igual que -1,96 y menor o igual que + 1,96. Deducimos esto porque el 95% de los valores de la distribución se halla contenido entre esos dos puntajes Z. Esto equivale a decir que obtener por azar un valor menor que -1,96, o mayor que + 1,96 es muy improbable. Esto se debe a que por debajo y por encima de esos valores solo se halla contenida el 5% de la distribución.

Veamos con un sencillo ejemplo cómo esta propiedad de la distribución normal puede ayudarnos a esclarecer algunos aspectos en las conclusiones o afirmaciones de un estudio (más adelante veremos que es el procedimiento usual en la prueba de hipótesis). Supongamos que la evaluación provincial de lectura, arroja una media de 550. Supongamos ahora que la media de Córdoba Capital en esa evaluación es de 643 con una desviación estándar de 104. Al ver la diferencia del promedio provincial con el de Capital, podríamos afirmar que el rendimiento lector de los escolares cordobeses que residen en la capital es significativamente diferente del resto de la provincia. Entendamos el término significativamente como una diferencia sistemática suficientemente amplia que debe interpretarse como que los escolares residentes en Capital leen mejor que el resto.

Podemos utilizar la distribución normal para testear si esa afirmación puede sostenerse empíricamente o solo debe tomarse como una afirmación basada en una simple diferencia. Dijimos que la distribución de las puntuaciones de esa evaluación seguía un modelo normal, por lo cual dichas puntuaciones pueden transformarse a puntaje Z. Si los escolares de Capital tienen un rendimiento significativamente superior al resto, la puntuación Z de esa localidad debería aparecer en la distribución como un valor atípico. Dicho en otras palabras, su puntaje Z debería corresponder a la zona delimitada por el 2,5% superior de la distribución. Como ya sabemos, ese puntaje es Z=1,96.

Nos resta transformar la puntuación original de Capital a valor Z y comprobar donde queda situada en la distribución. Para esto empleamos la fórmula ya conocida y tenemos que:

$$Z_{Capital} = \frac{643 - 550}{104} = 0{,}89$$

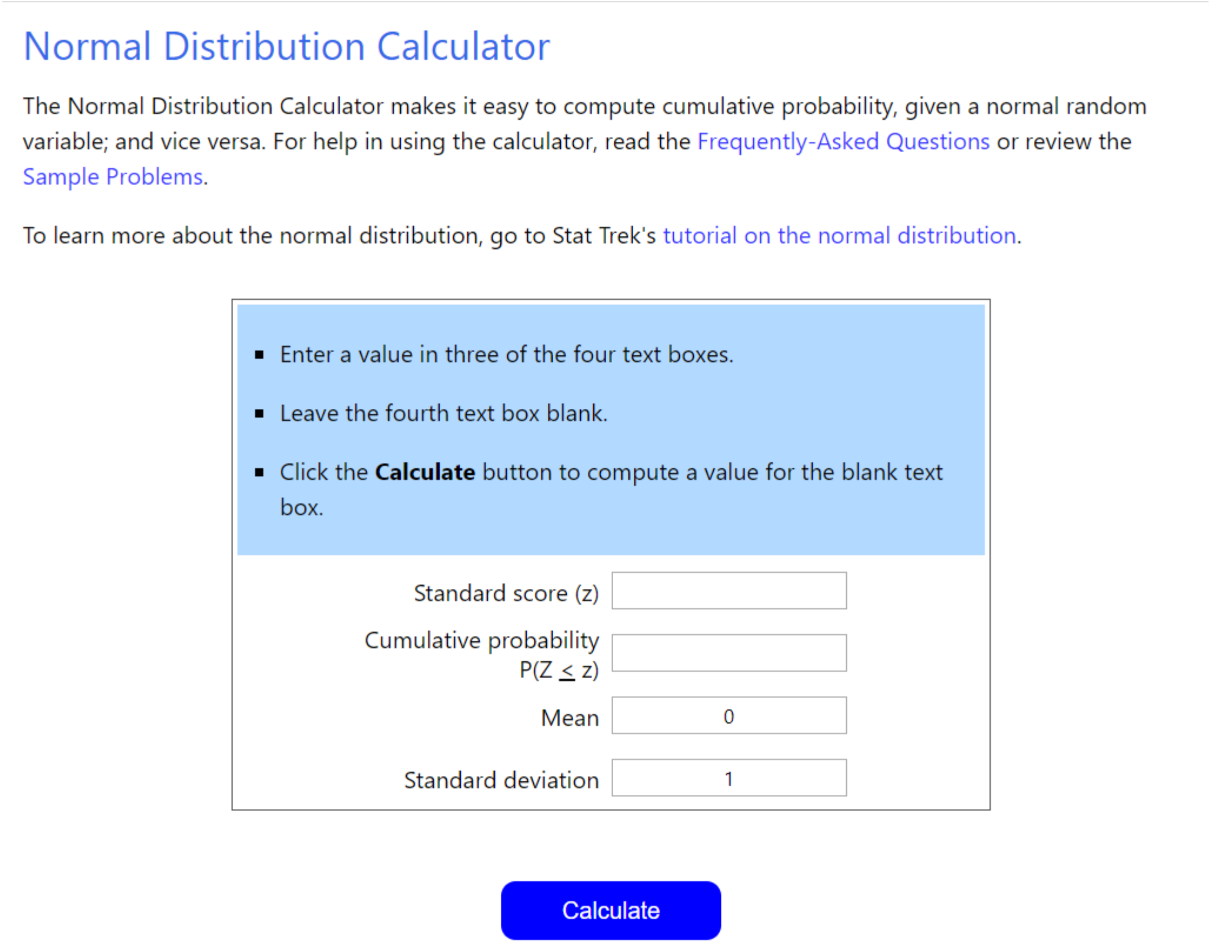
Como se aprecia el valor Z de Capital en la distribución es menor que el valor crítico 1,96 sobre el cual estarían los valores correspondientes al 2,5% de mayor rendimiento. Es factible por tanto concluir que, aunque el promedio de Capital es mayor que el del resto de la provincia, tal diferencia no es suficiente para afirmar que sea significativa.

La posibilidad de utilizar la distribución normal como distribución de probabilidades, resulta muy útil en el estudio de poblaciones, especialmente cuando nuestro foco de atención son variables que se aproximan a esa distribución teórica. Puesto que la distribución de cualquier variable con distribución normal puede ser estandarizada; esto es, transformada a valor Z, es factible calcular la probabilidad de que un valor en la variable sea mayor o menor a un valor Z dado.

Existen maneras de calcular los valores Z mediante las tablas de distribución normal que se publican en los textos de estadística. Aquí recurriremos a una función de un programa estadístico gratuito que se encuentra en el siguiente enlace.

https://stattrek.com/online-calculator/normal.aspx

Al ingresar al sitio encontraremos una pantalla como la siguiente:



Aunque la página está en inglés, nos interesa solo los comandos que aparecen señalados como:

Standard score (z): Puntaje estandarizado Z.

Cumulative Probability P(Z≤z): (Probabilidad acumulada hasta un valor Z dado).

Mean=0: (Media= 0)

Standard deviation=1: (desviación estándar= 1)

La función puntaje estandarizado Z, nos permite ingresar un valor Z conocido, y nos devuelve en la función Probabilidad acumulada hasta ese valor Z. Las restantes funciones indican que estamos trabajando sobre una distribución normal estandarizada con media igual a cero y desviación estándar igual a uno.

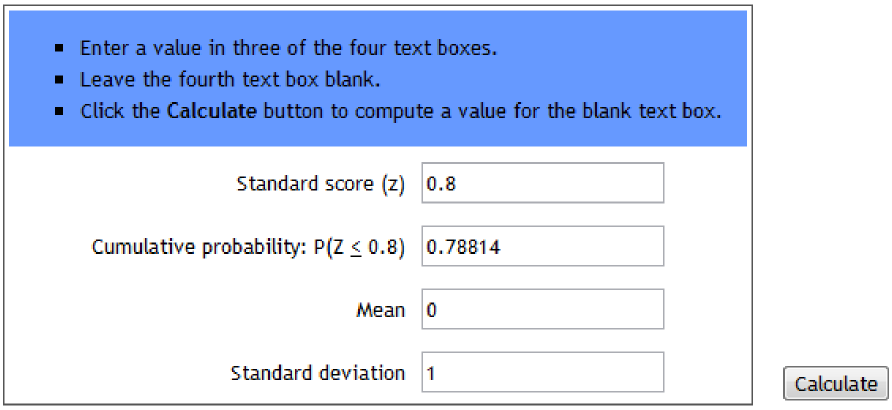
Retomemos ahora el ejemplo de la prueba de rendimiento en lectura en la provincia de Córdoba, cuya distribución de valores se aproxima a una distribución normal. Habíamos dicho que la media en dicha prueba es de 550, y la desviación estándar es de 125. En base a estos datos vamos a plantear dos preguntas:

1. ¿Cuál es la probabilidad seleccionar al azar un individuo de esa población, cuya puntuación en la prueba de lectura sea de 650 o más puntos que la media?

Para resolver esta cuestión debemos transformar el puntaje original en puntuación Z mediante la siguiente ecuación:

$$Z = \frac{650 - 550}{125} = 0{,}8$$

Al ingresar ese valor Z en el cuadro de diálogo, nos da como resultado que la probabilidad acumulada hasta Z=0,8 es de 0,78814, tal como indica la figura siguiente.



Dado que la pregunta es sobre la probabilidad de hallar un valor igual o mayor que 650, debemos restar ese valor a 1, que sería la probabilidad acumulada total. Por lo tanto, tendremos que:

$$P_{z \leq 650} = 1 - 0{,}78814 = 0{,}21186$$

Dado que las probabilidades pueden interpretarse como porcentaje, estamos en condiciones de afirmar que aproximadamente el 21% de la población obtendrá un rendimiento en la prueba de lectura de 650 puntos o más.

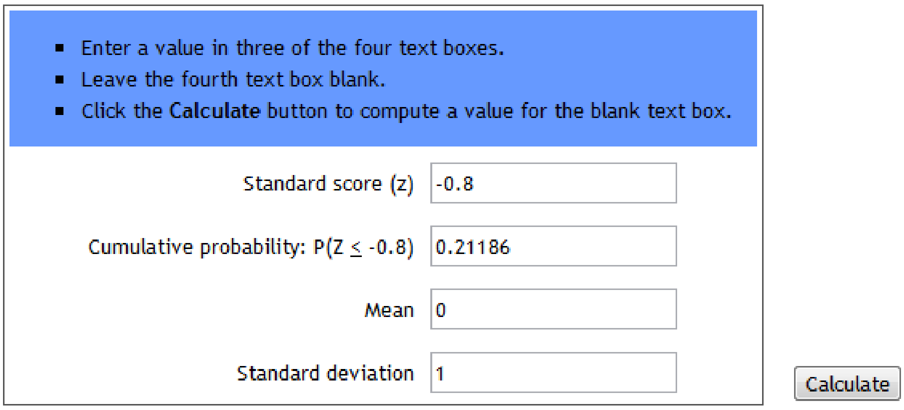
Esta propiedad de la distribución normal nos permite realizar evaluaciones a grandes conjuntos de la población y luego tipificar sus valores en términos porcentuales. Las pruebas estandarizadas de rendimiento más conocidas, muestran sus valores poblacionales en percentiles, que es una escala Z transformada a valores que van de cero a 100. Esta es una herramienta muy útil para establecer comparaciones entre diversas poblaciones utilizando solo un instrumento de evaluación.

Otra pregunta que podríamos plantearnos es la siguiente, dados dos puntajes de prueba:

2. ¿Qué porcentaje de la población obtiene puntajes en la prueba, comprendidos entre 450 y 650 puntos?

Esta pregunta se plantea en el contexto de este ejemplo como una derivación de lo que hemos esbozado en la respuesta anterior. Nótese que estamos intentando establecer un porcentaje de población entre dos valores dados que son cien puntos por debajo y por encima de la media poblacional. Aquí no podríamos utilizar simplemente la desviación estándar dado que es de 125. En este caso procedemos a calcular los valores Z correspondientes a las puntuaciones originales. Ya sabemos que la puntuación Z para el puntaje 650 es de 0,8; deducimos entonces que el valor Z correspondiente a un valor de 450 es de -0,8. La deducción se basa en que ambos puntajes se apartan de la media en 100 puntos, y en el hecho de que una distribución normal es simétrica.

Asimismo, ya sabemos que el valor de probabilidad acumulada hasta el puntaje Z=0,8 es de 0,78814. Nos resta saber cuál es la probabilidad acumulada hasta el valor Z= -0,8. Utilizando nuevamente el programa tenemos que ese valor es igual a 0,21186. Como ya lo habrá notado ese es valor de probabilidad por encima de Z=0,8, solo que ahora estamos trabajando sobre la otra mitad de la curva. Podríamos haber llegado a ese valor por deducción, pero interesa subrayar el carácter simétrico de la distribución y la aplicación de esta propiedad.



Dado que el valor p=0,78814 es la probabilidad acumulada hasta el puntaje Z=0,8, debemos ahora sustraer el valor p=0,21186, que es la probabilidad acumulada hasta el valor Z=-0,8. Esta resta la hacemos a los fines de centrarnos en el porcentaje comprendido entre esos valores; así tenemos que la resta de ambas probabilidades nos da 0,57628, lo cual transformado a porcentaje nos brinda la respuesta buscada. Aproximadamente el 57,6% de la población, obtendrá puntajes en la prueba de rendimiento lector comprendida entre 450 y 650 puntos.

Una de las propiedades de la distribución normal y la transformación Z es que se pueden hacer comparables dos variables que no están en la misma unidad de medida. Dicho en otros términos, la puntuación Z es una medida estandarizada independiente de la unidad de medición.

En el apartado de medidas de tendencia central, nos encontramos un problema en el cual un alumno obtuvo un 6 en matemáticas y un 8 en inglés. Utilizando la media y el desvío estándar concluimos que la nota 6 era una mejor puntuación en términos relativos. Si las mediciones provinieran de una distribución normal de datos, hubiera sido posible transformar los puntajes de matemática e inglés a valores Z mediante la siguiente fórmula:

$$Z = \frac{(x - \bar{x})}{DE}$$

Y así obtendríamos que:

$$Z_{matemáticas} = \frac{(6 - 4)}{1,68} = 1,19$$

$$Z_{inglés} = \frac{(8 - 8,55)}{1,34} = -0,41$$

Ahora vemos cómo el rendimiento en matemáticas es mayor que el rendimiento en inglés, mediante la comparación de sus respectivos puntajes
Z: -0,41<1.19.

Las puntuaciones Z, al ser medidas estandarizadas, permiten comparaciones aun cuando la variable hubiera sido medida en la misma escala. Esto facilita la interpretación de la posición relativa de cada unidad de análisis. Por ejemplo, si quisiéramos ubicar la posición relativa de cuatro provincias respecto al promedio nacional en las evaluaciones nacionales sobre rendimiento en matemáticas, se podría utilizar el puntaje Z de la siguiente manera.

	Media	**DE**
Total País	59	25

Teniendo la media y la desviación estándar para el total del país, es posible posicionar cada provincia mediante su puntaje Z; así para Córdoba esta puntuación será:

$$Z_{Córdoba} = \frac{(61 - 59)}{25} = 0.08$$

Si procedemos de la misma manera con cada provincia, obtenemos la siguiente tabla:

Provincia	Media	Puntaje Z
Córdoba	61	0.08
Rio Negro	68	0.36
Buenos	63	0.16
Formosa	41	-

Entonces, mediante la puntuación Z es posible establecer cuáles provincias han obtenidos buenos resultados en la evaluación, y cuáles han obtenido una puntuación menos favorable.

Misceláneas

1. La distribución normal y sus propiedades han sido objeto de estudio de matemáticos que han dejado una profunda huella en la historia, tal el caso de Abraham De Moivre, Pierre Simon Laplace y Carl Friedrich Gauss. Sin embargo, el uso y la aplicación extendida al estudio de fenómenos sociales se lo debemos a Lambert Quetelet y Francis Galton. Pero fue finalmente Karl Pearson quien popularizó el término de curva normal.

2. La curva de distribución normal es un modelo matemático aplicado a la distribución de variables empíricas. La función de densidad de la distribución está dada por la siguiente expresión.

$$f(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

Parece una fórmula muy compleja, pero contiene solo dos parámetros que son variables: μ (mu) y σ (sigma), donde:

μ: media de la distribución poblacional

σ: desviación estándar de la distribución poblacional

1. Cuando tratamos con la distribución normal estándar, tenemos que μ=0 y σ=1. En estadística matemática suele encontrarse que la distribución normal estándar se expresa como: $X \sim N(0,1)$.

2. La utilización de la distribución de las áreas bajo la curva normal que hemos visto en apartados anteriores se lo debemos principalmente al trabajo de Pafnuti Chebyshev, quien propuso su Teorema donde:

Si $X \sim N(\mu,\sigma)$, entonces:

a) $p(\mu-\sigma<X<\mu+\sigma) = 0{,}68$

b) $p(\mu-2\sigma<X<\mu+2\sigma) = 0{,}95$

c) $p(\mu-3\sigma<X<\mu+3\sigma) = 0{,}997$

Lo que equivale a decir que el 68% (aproximadamente) de los valores que tome la variable x estarán situados a una distancia de la media inferior a una desviación estándar. De la misma manera, el 95% de los valores estarán situados a menos de 2 veces la desviación estándar, y un 99,7% de dichos valores se encontrarán dentro de 3 desviaciones estándar. Por lo tanto, para una variable empírica que se aproxime a la distribución normal, la mayor parte de los valores quedan comprendidos a tres desviaciones estándar de la media. Este teorema se volvió fundamental en la investigación y la prueba de hipótesis.

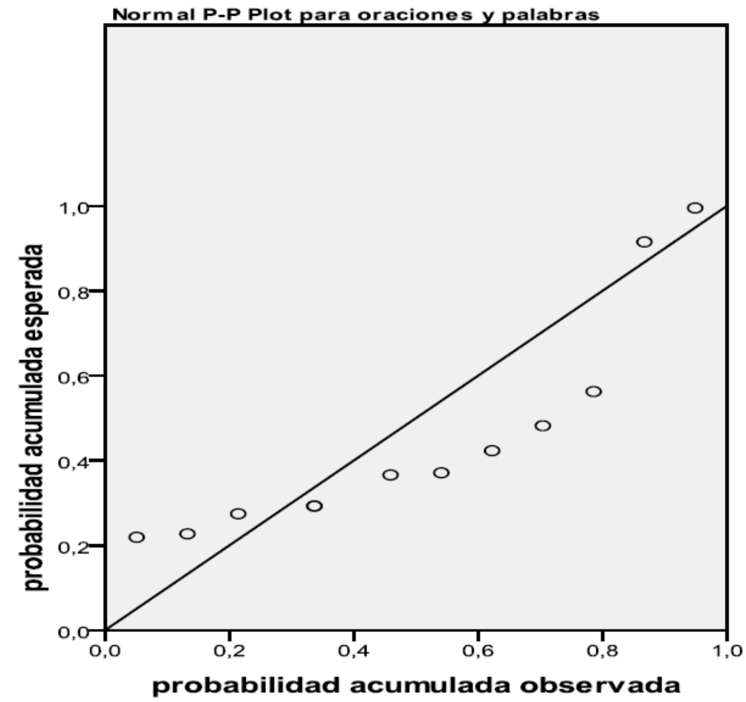
Distribuciones asimétricas

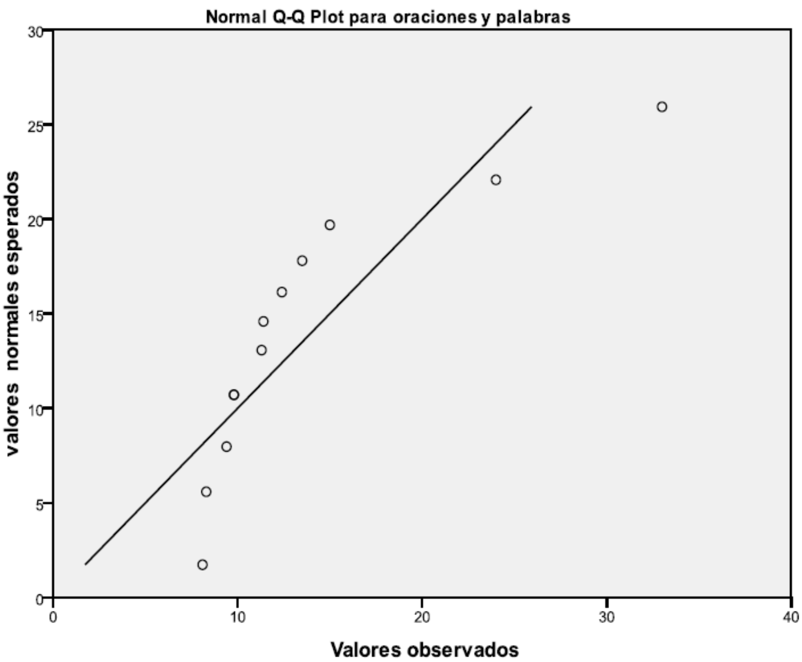
La distribución normal como modelo es una herramienta poderosa para el análisis de datos, pero resulta esencial verificar que una variable empírica se comporta como tal, dado que no podríamos aplicar ninguna de las técnicas estadísticas descriptas si ese no fuera el caso. Las más simples y útiles de las verificaciones son los gráficos de la distribución de los valores de la variable empírica, tales como el histograma o el diagrama de cajas. En ellos, una inspección visual nos informa si la variable sobre la que estamos trabajando se aproxima o no a una distribución normal. Aquí es importante tener en cuenta el término aproximación dado que nunca una variable empírica resultará exactamente una curva normal. Por lo tanto, existen medidas, llamadas coeficientes de asimetría y curtosis que nos dicen qué tan diferente es una distribución empírica de una normal. Solo por citar uno, el coeficiente de asimetría y de curtosis de Fisher nos informa dentro de qué límites podremos considerar a una distribución empírica como una distribución normal.

En los programas estadísticos existen gráficos y pruebas de normalidad que

nos permiten verificar rápidamente si una distribución empírica puede tratarse como una normal. Un ejemplo son los gráficos P-P, donde se representan en ejes cartesianos las proporciones acumuladas de la variable empírica, con las de una distribución normal teórica tomada de esos mismos valores. En el gráfico aparece a continuación aparece una variable real que se denomina oraciones y palabras. Se evalúa la capacidad de retención mnémica. Es una variable métrica y se ha aplicado a una muestra de alumnos de primaria. Para saber si la variable sigue una distribución normal bastaría ver su histograma, pero el gráfico P-P estandariza la variable para que quede comprendida entre los valores 0 y 1, luego se grafica la proporción de casos acumularos si la variable tuviera una distribución normal perfecta. En el gráfico se representa como una recta y es la probabilidad esperada ideal para la distribución normal. Entonces, si la distribución empírica está próxima a la normal los casos efectivamente observados estarían muy próximos a la recta (los puntos deberían parearse con la distribución teórica). En el diagrama que vemos más abajo, se aprecia que los casos se apartan de la recta, y por tanto estamos en presencia de una distribución que no se aproxima de manera fiable a una distribución normal. El gráfico recibe el nombre P-P en tanto produce una transformación de probabilidad acumulada basada en el modelo normal.

Otros gráficos que siguen la misma lógica son conocidos como gráficos Q-Q donde se representan los cuantiles observados en la distribución empírica respecto a los cuantiles que se esperaría observar si los datos provinieran de una distribución normal. Los cuantiles son una estandarización, pero que se toma en base 100.

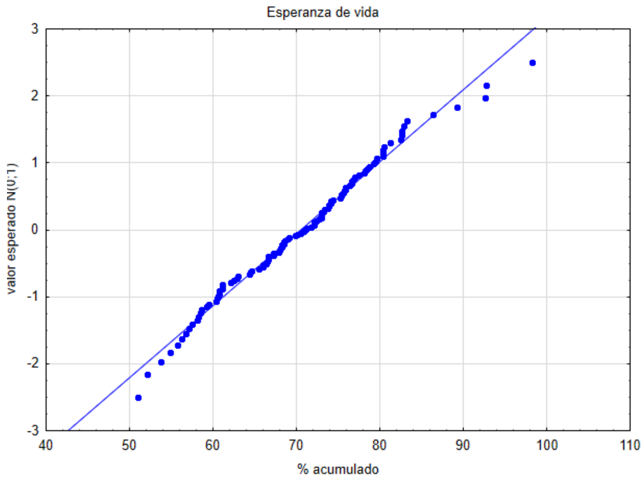




La interpretación de este diagrama es la misma que para los gráficos P-P, pero basada en los percentiles.

Además de permitir valorar la desviación de la normalidad, estos gráficos permiten determinar si las curvas observadas muestran desviaciones atendibles de la simetría (curvas en forma de U), o la curtosis (curvas en forma de S).

Los ejemplos que vimos anteriormente representan curvas que se apartan de la normalidad estadística. Pero, hemos trabajado con una variable llamada esperanza de vida en la población que tiene una gran proximidad con la distribución normal. En una gráfica Q-Q esta gráfica se vería de la siguiente manera.



Nótese que, en esta gráfica, casi todos los puntos caen sobre la recta que representa los valores acumulados de una distribución normal teórica estandarizada.

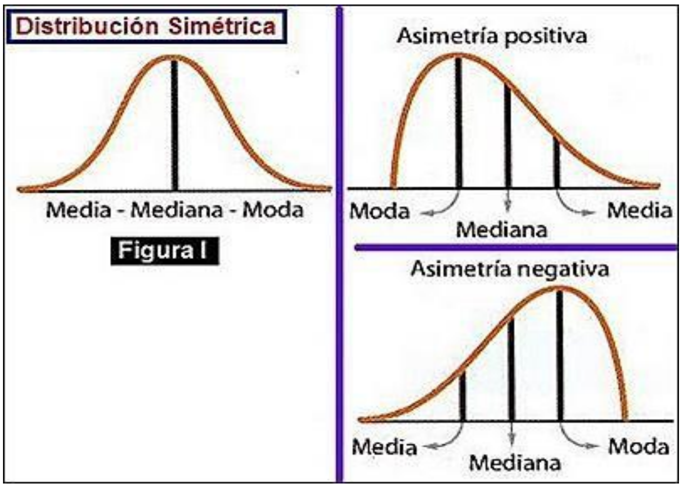
Medidas de asimetría y curtosis de una distribución

Mediante la exploración gráfica, es posible visualizar si las distribuciones se apartan de la normal. El histograma y el polígono de frecuencia son los gráficos más sencillos para ver los sesgos de la distribución, y como vimos anteriormente, existen las gráficas P-P y Q-Q que cuantifican puntualmente la asimetría. Otras medidas de asimetría se calculan directamente de la relación entre media, mediana y modo. Siendo la distribución normal simétrica (y por ello tanto el modo, la mediana y la media coinciden en la parte central de la distribución), la resta de las medidas mencionadas será cero. De esta propiedad se derivan dos coeficientes de asimetría conocidos como:

a) Coeficiente de Asimetría de Pearson, y

b) Coeficiente de Asimetría de Fisher.

En ambos casos, cuando el coeficiente de asimetría sea cero, indicará que el ajuste de la distribución observada es perfecto. En cambio, si el coeficiente de asimetría es mayor que cero (Ca>0), indica que la distribución muestra un sesgo positivo. Al contrario, si el coeficiente de asimetría es menor que cero (Ca<0), indica que la distribución muestra un sesgo negativo.

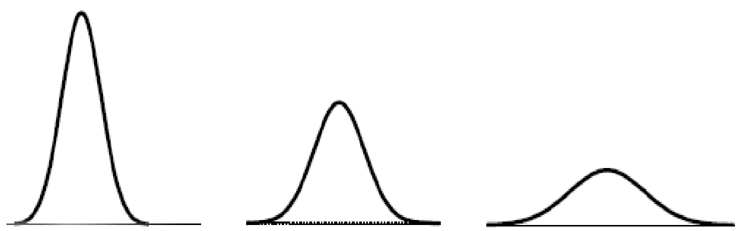


Otros coeficientes de asimetría toman en consideración la distribución de casos en cada uno de los cuartiles. En una distribución normal, la distribución de casos a lo largo de todos los valores de la variable será proporcional en cada uno de los cuartiles. Sobre esta base se calcula el coeficiente de asimetría de Bowley:

$$\mathrm{Ab} = \frac{(\mathrm{Q3} + \mathrm{Q1} - 2\mathrm{Me})}{(\mathrm{Q3} + \mathrm{Q1})}$$

En ambas expresiones Q son los cuartiles, y Me es la mediana de la distribución. En este caso también se cumple que, si la distribución se ajusta a la normal, el valor del coeficiente es cero. Luego, será mayor que cero en el caso en que la distribución muestre un sesgo positivo, y menor que cero en el caso en que la distribución tenga asimetría negativa.

Otra medida que acompaña a las de simetría es la curtosis, y se refiere a la acumulación de casos en torno a la media en una distribución. También toma como modelo a la distribución normal, cuyo valor de referencia en este estadístico es cero. La distribución normal por definición es mesocúrtica, por lo tanto, el coeficiente de curtosis será cero para toda distribución que se empareje a la normal. Si la distribución empírica bajo análisis acumula la mayoría de los casos próximos a la media, su forma aparece como más puntiaguda que la normal. En tal caso se dice que la distribución es leptocúrtica y su coeficiente de curtosis será mayor que cero. Al contrario, si la distribución observada muestra que los casos se reparten más dispersos en torno a la media, se trata de una distribución platocúrtica y su coeficiente de curtosis será menor que cero. En el siguiente gráfico se muestran los distintos tipos de distribuciones.



Los coeficientes de asimetría y curtosis, nos indican el grado en que una distribución empírica de datos se aparta de una distribución normal. Cuando existen variaciones menores, evidenciada por estos coeficientes y los gráficos correspondientes, no existen restricciones para asimilar la distribución empírica y la normal estándar bajo el modelo matemático de esta última. Pero si las desviaciones aparecen marcadas es necesario recurrir a diferentes pruebas de normalidad estadística que no solo indican cuánto se aparta una distribución empírica de tal modelo, sino que tipo de correcciones deben hacerse. Estas pruebas son conocidas también como Bondad de Ajuste.

Pruebas de bondad de ajuste

Test de Shapiro–Wilk	Se usa para contrastar la normalidad de un conjunto de datos. Fue publicado por Samuel Shapiro y Martin Wilk. Se considera uno de los test más potentes para el contraste de normalidad, sobre todo para muestras pequeñas (n<50).
Prueba de Kolmogórov- Smirnov	Es una prueba no paramétrica que determina la bondad de ajuste de dos distribuciones de probabilidad entre sí. Es una alternativa para verificar la normalidad de una distribución.

Prueba de Anderson- Darling	La prueba asume que no existen parámetros a estimar en la distribución que se está probando, en cuyo caso la prueba y su conjunto de valores críticos siguen una distribución libre. Cuando se aplica para probar si una distribución normal describe adecuadamente un conjunto de datos, es una de las herramientas estadísticas más potentes para la detección de la mayoría de las desviaciones de la normalidad.
Test de Jarque-Bera	En estadística, el test de Jarque-Bera es una prueba de bondad de ajuste para comprobar si una muestra de datos tiene la asimetría y la curtosis de una distribución normal. El test recibe el nombre de Carlos Jarque y Anil K. Bera.

Puesto que el tema sobrepasa el desarrollo en este trabajo, solo mencionaremos algunas de tales pruebas. Todas se basan en el mismo principio: el valor de cada prueba cuantifica la probabilidad de que la distribución empírica observada provenga de una distribución normal. Si la probabilidad es alta, el modelo normal puede aplicarse, pero si tal probabilidad es baja, se descarta esa posibilidad y la distribución debe ser tratada bajo otro modelo matemático distinto que la normal estándar.

Guía de Actividades N° 5

Modelo de distribución normal - Puntuación Z.

Objetivos de la actividad:

a) comprender los conceptos básicos de distribución de valores de variables bajo del modelo de distribución normal y normal estándar.

b) interpretar los valores Z como formas de transformación de puntajes originales.

Actividad 1: Un criterio para seleccionar personas de “alto coeficiente intelectual”, consiste en elegir sólo el 10% más inteligente de los individuos que se someten a una prueba de inteligencia. El punto z crítico en la distribución normal estándar es 1.28. Se sabe que las puntuaciones de una prueba de selección se comportan de acuerdo a una distribución normal con media 500 y desviación estándar de 100. Se extraen cinco individuos al azar con las siguientes puntuaciones:

Sujeto	Puntuación
1	487
2	567
3	635
4	442
5	630

¿En dónde se ubica cada sujeto respecto al punto z crítico?

¿Qué puntuación será el límite para rechazar o aceptar un candidato?

Actividad 2: Se sabe que las notas de matemáticas y de lengua para un grupo de alumnos, se comporta de acuerdo a una distribución normal con media 55 y varianza 121 para matemáticas y media 70 con varianza 81 para lengua. De todos los alumnos, se extrajeron cinco al azar cuyo rendimiento fue el siguiente:

Alumno	Matemáticas	Lengua
1	65	43
2	40	89
3	35	55
4	67	75
5	35	52

a) Ubique la posición de cada nota en la distribución que corresponda.

b) Describa el desempeño de cada uno de estos 5 alumnos en relación al grupo.

Actividad 3: Se sabe que las notas de los alumnos que acaban de ingresar a la Lic. en Ciencias de la Educación se comportan de acuerdo a una distribución normal con media 60 y varianza 144 para Problemáticas Filosóficas y Educación y media 70 con varianza 64 para Historia de la Educación Argentina. De todos los alumnos, se extrajeron cinco al azar cuyo rendimiento fue el siguiente:

Alumno	Problemáticas Filosóficas y Educación	Historia de la Educación Argentina
1	81	43
2	40	89
3	30	75
4	85	82
5	55	52

a) Ubique la posición de cada nota en la distribución que corresponda.

b) Describa el desempeño de cada uno de estos 5 alumnos en relación al grupo.

c) El 10% de los alumnos con mejores resultados obtuvieron un valor z de 1.28, determine quienes superaron este resultado en Problemáticas, en Historia o en ambas.

Actividad 4: Utilizando pruebas estandarizadas, en una escuela se realizó una evaluación de Matemáticas y Lengua, de todos los alumnos que finalizaron el nivel medio. Se encontró que la distribución de los puntajes en esas evaluaciones sigue una distribución normal. En matemáticas resulto que la media fue de 45 y la varianza de 121. Para lengua la media fue de 70 con varianza igual a 81. De todos los alumnos, se extrajeron cinco al azar cuyo rendimiento fue el siguiente:

Alumno	Matemáticas	Z Matemáticas	Lengua	Z Lengua
1		1.81		-3
2		-0.45		2.11
3		-0.91		-1.67
4		2		0.56
5		-0.73		-2

Con los datos ofrecidos, se pide: calcule la puntuación original de los alumnos en las materias y ubique el/los alumnos que hayan puntuado una desviación estándar o más por encima de la media en ambas asignaturas.

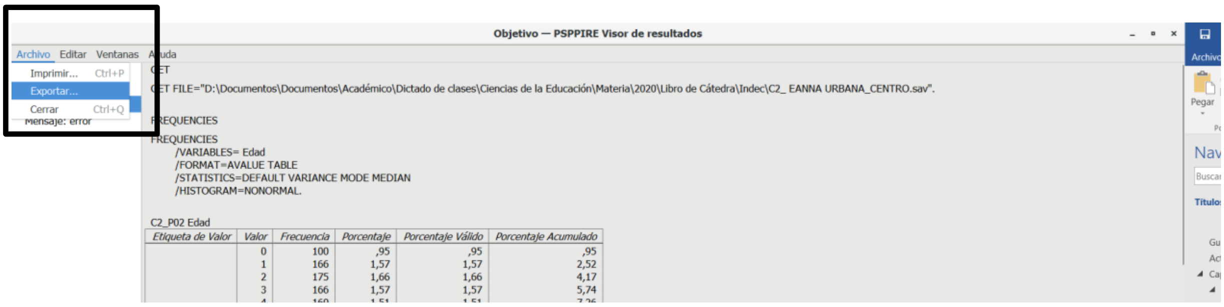
Actividad 5: el peso de una población de estudiantes se distribuye normalmente con una media de 72kg y una desviación estándar de 9kg. De una población de 200 personas se sabe que 5 tienen un peso elevado de 85, 90, 95, 99 y 102kg. Además, se ha determinado que 3 o más desviaciones estándar de la media es considerado obesidad infantil. Indique cuáles de estos casos entran dentro de la categoría de obesidad infantil.

Actividades en PSPP N° 5:

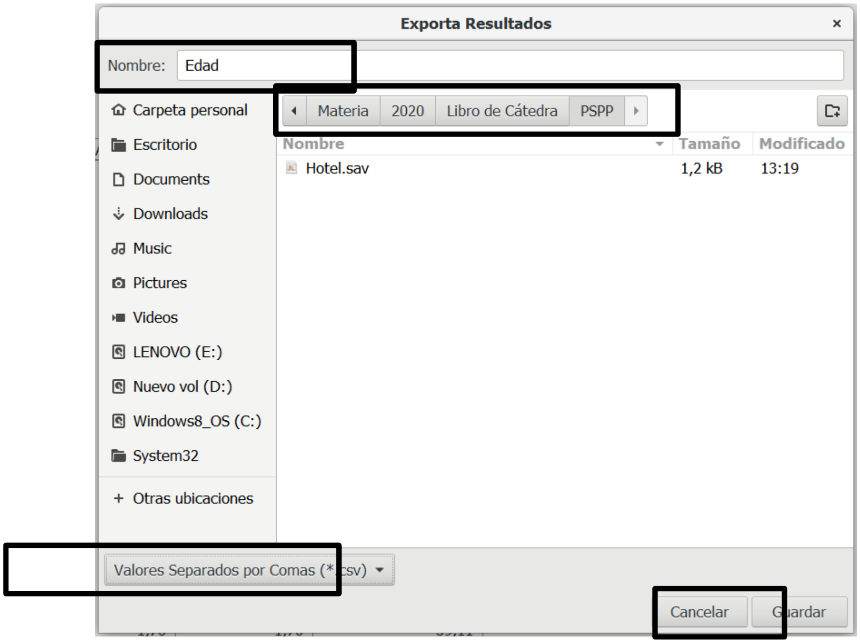
Vamos a abrir nuevamente el programa PSPP, y en esta ocasión trabajaremos por segunda vez con la base de datos que descargamos del Instituto Nacional de Estadísticas y Censos (INDEC). Para ello abrimos el documento que guardamos con el nombre C2_EANNA URBANA_CENTRO.SAV.

Una vez abierta la base de datos, nos concentraremos en el análisis de la variable edad. Nos interesa conocer la distribución de la edad de los casos para la región Centro, por lo que vamos a pedir al programa que realice una tabla de frecuencias, que nos dé las medidas resumen de la variable y que realice un Histograma de frecuencias.

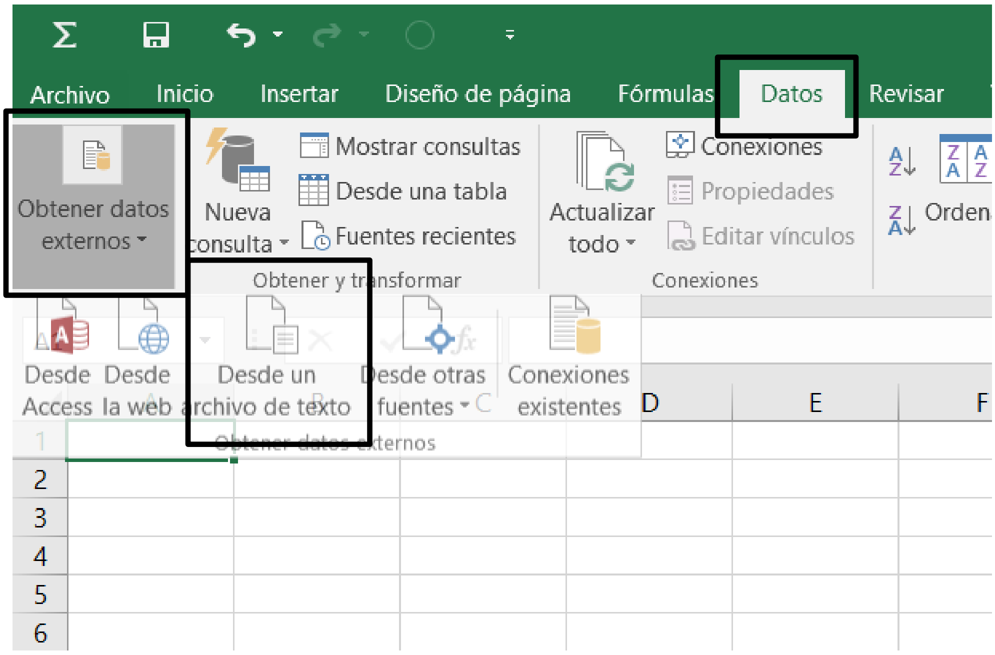
Ustedes van a poder ver los resultados en una tabla muy extensa y difícil de manejar. Esto es producto de ser una variable métrica discreta con muchos valores, dado que la edad presenta valores que van de 0 a 100 años, o incluso más. Para hacer visible esta tabla la podemos exportar a Excel. Para ello, en el visor de resultados del programa, vamos a Archivo y seleccionamos Exportar.



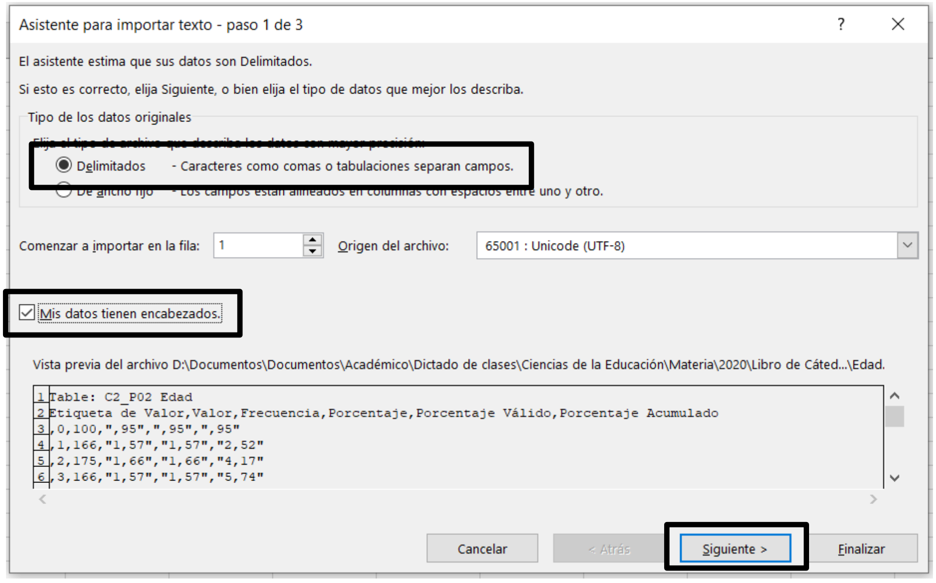
Colocamos el nombre del archivo de destino, que será Edad, y seleccionamos la carpeta donde tenemos la base de datos como ruta para guardarlo. Finalmente, donde dice Inferir tipo de archivo por la extensión, seleccionamos la extensión CSV.



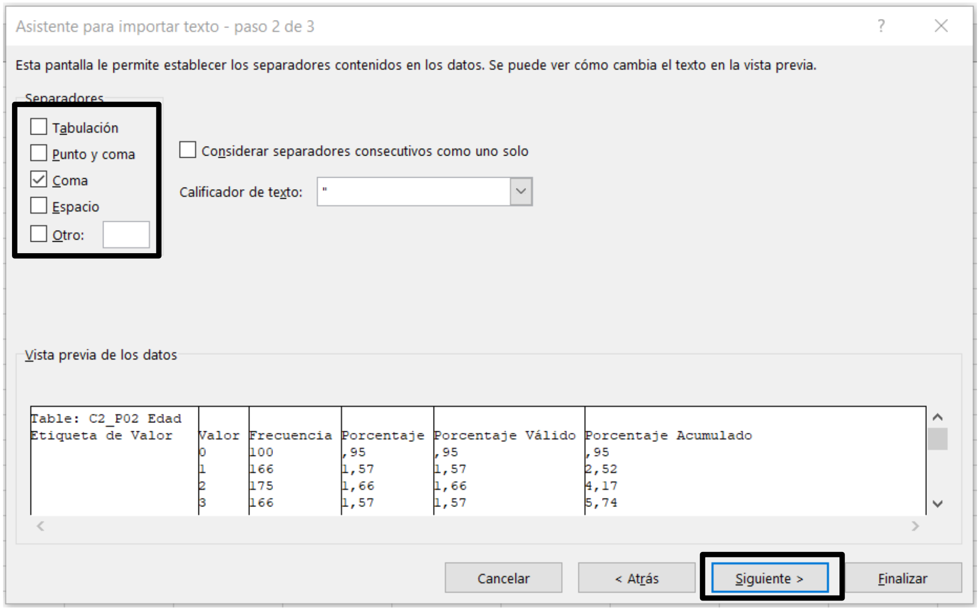
Luego abrimos una planilla de Excel en blanco, vamos a Datos, Obtener datos externos, y ponemos Desde un archivo de texto:



Seleccionamos el archivo que hemos guardado con el nombre “Edad.CSV” y dejamos la opción que está delimitado por Caracteres como comas o tabulaciones separan campos, también tildamos que Mis datos tienen encabezados, y damos Siguiente:



Destildamos la opción Tabulador y tildamos por Coma, todo lo demás queda como está y damos siguiente:



Luego indicaremos qué tipo de datos tiene cada columna. En este caso podemos dejar general a todas las columnas y damos a finalizar.

Excel nos va preguntar si coloca los datos en la hoja de cálculo existente o crea una nueva, lo dejamos en la existente y finalmente tenemos nuestros datos en Excel o su equivalente en Libre Office:

Table: C2_P02 Edad					
Etiqueta de Valor	Valor	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
	0	100	0,95	0,95	0,95
	1	166	1,57	1,57	2,52
	2	175	1,66	1,66	4,17
	3	166	1,57	1,57	5,74
	4	160	1,51	1,51	7,26
	5	158	1,49	1,49	8,75
	6	139	1,31	1,31	10,06
	7	168	1,59	1,59	11,65
	8	153	1,45	1,45	13,1
	9	170	1,61	1,61	14,71
	10	162	1,53	1,53	16,24
	11	138	1,31	1,31	17,55
	12	175	1,66	1,66	19,2
	13	144	1,36	1,36	20,56
	14	163	1,54	1,54	22,11
	15	172	1,63	1,63	23,73
	16	171	1,62	1,62	25,35
	17	169	1,6	1,6	26,95
	18	212	2,01	2,01	28,95
	19	181	1,71	1,71	30,67
	20	162	1,53	1,53	32,2
	21	197	1,86	1,86	34,06
	22	190	1,8	1,8	35,86
	23	164	1,55	1,55	37,41
	24	180	1,7	1,7	39,11
	25	193	1,83	1,83	40,94
	26	149	1,41	1,41	42,35
	27	132	1,25	1,25	43,6
	28	155	1,47	1,47	45,06
	29	137	1,3	1,3	46,36
	30	164	1,55	1,55	47,91
	31	122	1,15	1,15	49,06
	32	152	1,44	1,44	50,5
	33	131	1,24	1,24	51,74
	34	117	1,11	1,11	52,85
	35	167	1,58	1,58	54,43
	36	124	1,17	1,17	55,6
	37	128	1,21	1,21	56,81
	38	134	1,27	1,27	58,08
	39	117	1,11	1,11	59,18
	40	161	1,52	1,52	60,71
	41	118	1,12	1,12	61,82
	42	154	1,46	1,46	63,28
	43	136	1,29	1,29	64,57
	44	93	0,88	0,88	65,45
	45	130	1,23	1,23	66,68
	46	117	1,11	1,11	67,78
	47	92	0,87	0,87	68,65
	48	104	0,98	0,98	69,64
	49	82	0,78	0,78	70,41
	50	143	1,35	1,35	71,77
	51	108	1,02	1,02	72,79
	52	131	1,24	1,24	74,03
	53	116	1,1	1,1	75,12
	54	97	0,92	0,92	76,04
	55	114	1,08	1,08	77,12
	56	100	0,95	0,95	78,06
	57	77	0,73	0,73	78,79
	58	128	1,21	1,21	80
	59	97	0,92	0,92	80,92
	60	152	1,44	1,44	82,36
	61	97	0,92	0,92	83,28
	62	120	1,14	1,14	84,41
	63	111	1,05	1,05	85,46
	64	81	0,77	0,77	86,23
	65	128	1,21	1,21	87,44
	66	112	1,06	1,06	88,5
	67	79	0,75	0,75	89,25
	68	107	1,01	1,01	90,26
	69	82	0,78	0,78	91,03
	70	116	1,1	1,1	92,13
	71	56	0,53	0,53	92,66
	72	98	0,93	0,93	93,59
	73	74	0,7	0,7	94,29
	74	51	0,48	0,48	94,77
	75	67	0,63	0,63	95,4
	76	40	0,38	0,38	95,78
	77	47	0,44	0,44	96,23
	78	50	0,47	0,47	96,7
	79	43	0,41	0,41	97,11
	80	61	0,58	0,58	97,68
	81	33	0,31	0,31	97,99
	82	41	0,39	0,39	98,38
	83	25	0,24	0,24	98,62
	84	25	0,24	0,24	98,86
	85	29	0,27	0,27	99,13
	86	21	0,2	0,2	99,33
	87	21	0,2	0,2	99,53
	88	12	0,11	0,11	99,64
	89	8	0,08	0,08	99,72
	90	8	0,08	0,08	99,79
	91	3	0,03	0,03	99,82
	92	4	0,04	0,04	99,86
	93	6	0,06	0,06	99,91
	94	4	0,04	0,04	99,95
	95	2	0,02	0,02	99,97
	96	1	0,01	0,01	99,98
	97	1	0,01	0,01	99,99
	98	1	0,01	0,01	100
Total		10572	100	100	

Table: C2_P02 Edad		
N	Válido	10572
	Perdidos	0
Media		35,28
Modo		18
Desv Std		22,86
Varianza		522,47
Mínimo		0
Máximo		98
	50 (Mediana)	32

Ya hemos aprendido a pasar los resultados de PSPP a Excel. Esta operación en SPSS es muchísimo más sencilla y se resuelve sólo con los comandos copiar y pegar, pero, dado que es muy importante que usted pueda llevar los resultados del PSPP a Excel, hemos decidido realizar este instructivo para PSPP.

Como podemos ver, luego de llevarlo a Excel sigue siendo una distribución tan extensa que no tiene mucho sentido presentarla toda. Tampoco tiene mucho sentido el valor modal, que en este caso que será 18 años con un 2.01% de los casos.

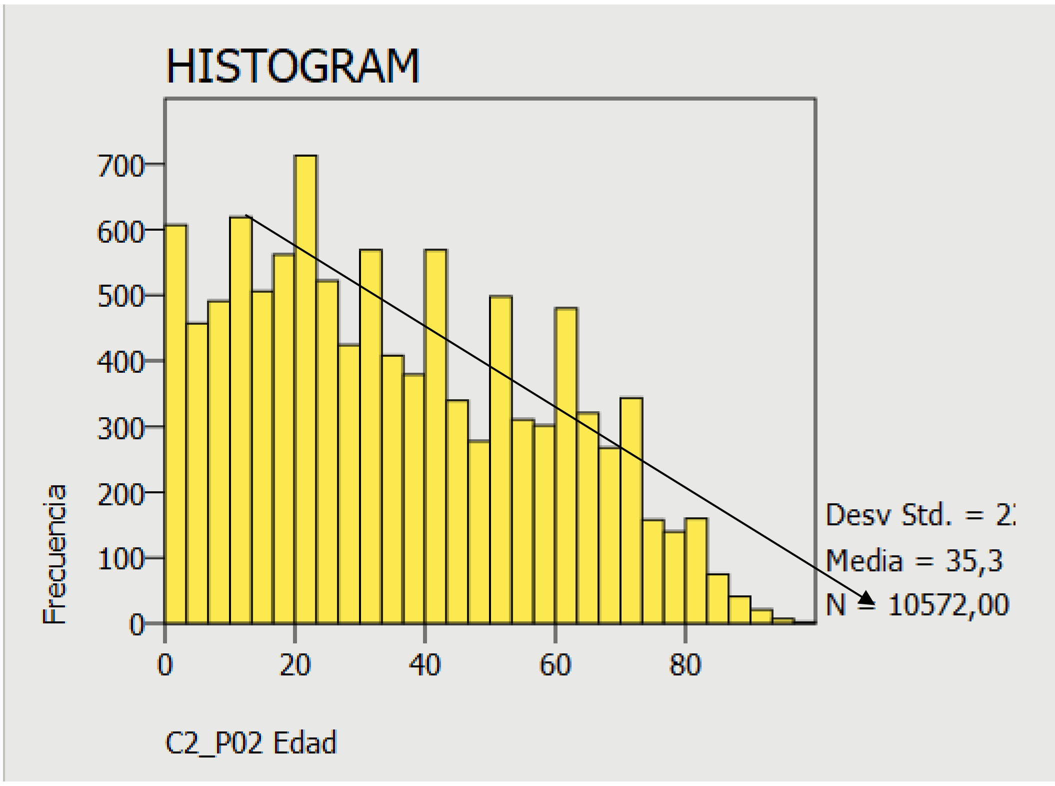
Sí será valioso obtener los valores de los cuartiles y de la mediana. También la media, la varianza, el mínimo, el máximo, el rango y la Desviación estándar. Lo invitamos a que consigne esos datos considerando el porcentaje acumulado:

………
………
………
………
………
………
………
………
………
………
………
………
………

………
………
…………..

Con el objetivo de resumir los datos, es conveniente presentar la misma variable reorganizada por intervalos de edad, como por ejemplo por quinquenios de edad o lustros (0-4, 5-9, …). A esto lo realizaremos más adelante.

No obstante la larga cantidad de valores para el gráfico, podemos ver que la distribución de los datos son reagrupados automáticamente en intervalos. También podremos observar que la edad dista de ser parecida a la de una curva normal, que comienza con pocos casos, tiene un máximo y luego cae en la cantidad de casos nuevamente. Más bien se parece a una línea inclinada hacia abajo.



En esta ocasión se le pide que describa los resultados teniendo en cuenta el Histograma:

……………………………………………………………………………………………………
………
………
………
………
………
………
………
………

……
……
……
……
………………………………………………………………………………………………

Ahora sí, vamos a analizar una distribución que se asemeja a una distribución normal. Para ello vamos a tomar la variable C2_P28 Horas de trabajo en la ocupación principal y vamos a pedir la tabla de frecuencias y las medidas resumen. Los resultados tendrían ser los siguientes:

C2_P28 Horas de trabajo ocupación principal

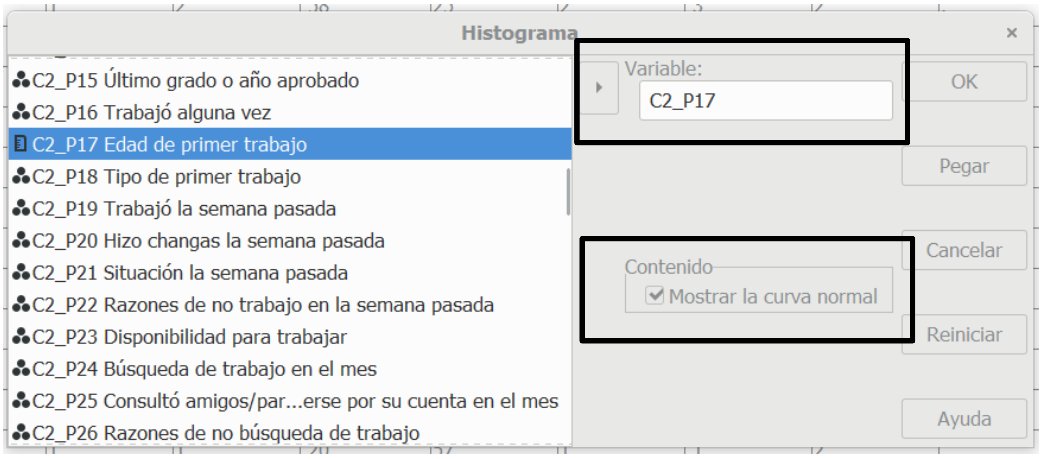
N	*Válido*	2255
	Perdidos	8317
Media		40,17
Modo		40,00
Desv Std		16,68
Varianza		278,10
Mínimo		1,00
Máximo		140,00
	50 (Mediana)	40

Se le pide que describa los resultados para la variable.

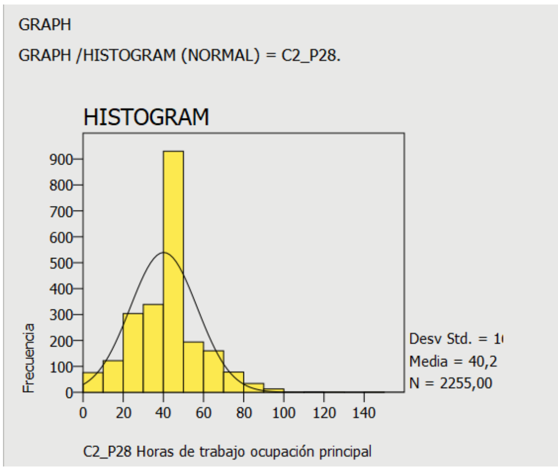
………
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……

……
……
……
………………………………………………………

Queremos comprobar que la distribución es similar a la de una curva normal. Si bien existen instrumentos estadísticos para hacer esta comprobación, en esta ocasión lo haremos de un modo visual. Para ello vamos a Gráficos, Histograma y seleccionamos la variable en estudio, además le pedimos que nos muestre la curva normal para poder comparar:



El resultado será:



Como ustedes pueden comprobar en el gráfico, la distribución crece hasta el intervalo modal que está entre 40 y 49 horas y luego decrece gradualmente hasta llegar a cero. Se puede apreciar que esta variable tiene un comportamiento más parecido al de una curva normal.

En esta ocasión se le solicita que, con el valor de la media y la desviación estándar que usted calcule con el programa, realice el grafico de la curva normal estandarizada de la variable C2_p28 y ubique en el gráfico los casos 11, 37, 57, 58, 115 y 154 de acuerdo a su valor en dicha variable, para ello deberá transformarlos a puntuación Z.

Capítulo 6

Estimación de parámetros

Como hemos mencionado más arriba, estudiar una población completa suele ser muy costoso, en tiempo, recursos materiales y humanos. Por ello, mediante un muestreo probabilístico de la población de interés, es posible estimar cualquier parámetro poblacional que estemos interesados conocer.

Cuando calculamos un índice en una población, lo denominamos parámetro.

Cuando solo contamos con una muestra aleatoria de la población, y calculamos el mismo índice nos referimos a él como un estadístico. Si hemos asegurado un muestreo aleatorio de la población, es posible utilizar el índice calculado para estimar, con cierta confiabilidad, el parámetro poblacional.

Como se desprende de lo dicho, la estadística nos sirve en este caso para conocer aspectos fundamentales de la población a un costo mucho menor. El proceso mediante el cual estimamos un parámetro a partir de una muestra se denomina **inferencia estadística**. Es por ello que para distinguir un parámetro de un estadístico se reserva el uso de las letras del alfabeto griego para referirnos a parámetros y del alfabeto latino para los estadísticos.

En este capítulo vamos a estimar parámetros de una media y una proporción poblacional a partir de los datos de una muestra. También vamos a comparar las proporciones y medias en poblaciones y muestras diferentes y aprenderemos a realizar estimaciones puntuales y por intervalos de parámetros poblacionales.

Estimación de una media poblacional

Suponemos que una variable cualquiera en la población tiene una distribución, que puede ser conocida o no. Mediante el teorema del límite central sabemos que sucesivas muestras del mismo tamaño extraídas de esa población, se aproxima a la distribución normal. Es decir, si procedemos a extraer muestras aleatorias de igual tamaño de la población (procedimiento que teóricamente puede repetirse infinitas veces), encontraremos que la distribución del estadístico en esa distribución sigue siempre un modelo normal, como el que vimos en la anterior unidad.

Supongamos que deseamos conocer el promedio de años de estudios alcanzados en la población comprendida entre 30 y 40 años de edad residente en Córdoba Capital. El parámetro en el que estamos interesados es un promedio al que vamos a denominar $\boldsymbol{\mu}$ (letra griega que se pronuncia mu), por tratarse de la población. Si pudiéramos extraer sucesivas muestras de esa población y calculamos el promedio de años de estudio en ella, encontraríamos que el estadístico sigue una distribución normal. Es decir, la mayoría de las muestras arrojarán un promedio o media, que se agruparán en torno a un valor particular. Lo que demuestra el teorema del límite central es que el valor del estadístico, en este caso la media, estará próximo al verdadero valor de μ, siempre que la muestra sea aleatoria.

Nótese que la media estará próxima pero no será exactamente igual a μ. Es decir, la estimación de una media poblacional mediante una media muestral estará sujeta a error. Lo interesante del teorema del límite central es que especifica la distribución de ese error, llamado error estándar de la media, y es posible calcularlo mediante la siguiente ecuación:

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

Conociendo las propiedades del teorema del límite central y la distribución de muestreo del estadístico, es posible estimar el parámetro estableciendo un límite de confianza, basado en las propiedades de la distribución normal estandarizada. Este proceso es el que se denomina estimación de parámetros.

Continuando con el ejemplo anterior, supongamos que se realiza un estudio tomando una muestra aleatoria de la población de 200 personas. Los resultados muestran que el promedio de años de estudio es 15,2 con una desviación estándar de 9.

Dado que se cuenta con una muestra representativa de la población de interés, se decide estimar el parámetro poblacional años promedio de estudios en la población adulta de 30 – 40 años, de la Ciudad de Córdoba Capital.

Datos:

Promedio de años de estudio: 15,2

Desvío Estándar: 9

Tamaño de la muestra: n=200

Para establecer los límites de confianza de la estimación, recurrimos a las áreas bajo la curva normal estandarizada. Sabemos que estas áreas se expresan en puntuación Z, dado que la distribución tiene como propiedad distintiva que su media es cero y su desviación estándar es igual a uno. En primer lugar, necesitamos establecer el porcentaje de confiabilidad para la estimación. Supondremos para este estudio que queremos una confianza elevada para la estimación, esto es, del 99%. El paso siguiente es traducir ese porcentaje a valor Z. Por las propiedades de la distribución normal estándar sabemos que un valor Z=± 2,58 cubre el 99% del área bajo la curva, de modo que ese dato es el que necesitamos para ajustar la estimación.

Puesto que solo estamos utilizando una sola muestra para estimar el parámetro, es necesario conocer el error estándar de la media para el intervalo de confianza. En la ecuación original, el error estándar se calcula a partir de la desviación estándar de la población que es σ (sigma). Como este es un parámetro y por tanto resulta desconocido, reemplazamos en la fórmula original estos términos por la desviación estándar muestral y a N por n-1 (tamaño muestral menos uno). Así el error estándar para la media de este ejemplo será:

$$EE_M = \frac{DE}{\sqrt{n-1}} = \frac{9}{\sqrt{199}} = 0{,}637$$

Ahora debemos calcular el Intervalo de Confianza para la media. El intervalo establece dos límites, uno superior y otro inferior, dentro de los cuales se supone que se encuentra el parámetro poblacional. La ecuación para calcular ese intervalo es la siguiente:

$$LI = M \pm (Z \times EE_M)$$

En nuestro ejemplo estos límites serán:
Superior: 15,2+(2,58*0,637)=16,84
Inferior: 15,2-(2,58*0,637)=13,55

El resultado se expresa de la siguiente manera: 13,55 ≤ μ ≤ 16,84. Así, es posible afirmar que la cantidad de años de estudio promedio de la población de Córdoba Capital, entre 30 y 40 años, se encuentra comprendida entre los 14 y los 16 años aproximadamente (con una confianza del 99%).

Estimación de una proporción poblacional

De la misma manera en que se puede estimar una media paramétrica, también es posible estimar otros parámetros en la población, como por ejemplo una proporción. La distribución de muestreo de una proporción sigue los mismos principios que la distribución para la media, es decir, que se pueden explicar mediante el teorema del límite central, y por tanto la distribución de referencia será la normal estandarizada.

Veamos mediante un ejemplo cómo se estima una proporción paramétrica. Supongamos que se desea estimar la tasa de promoción entre primaria y secundaria en la ciudad de Córdoba Capital. Para ello, se toma una muestra de 500 alumnos de diez escuelas de la ciudad. Se encuentra que en la muestra, 439 alumnos son efectivamente promovidos. La tasa resultante es 439/500=0.87. Con estos datos se estima que la tasa de promoción en la población escolar en la ciudad de Córdoba Capital es del 87%.

Datos:

Proporción de alumnos promovidos: p=0.87 (87%)

Tamaño de la muestra: 500

Como en el caso anterior, es necesario fijar un nivel de precisión para la estimación. En este caso el nivel de precisión será del 95%. Aquí nuevamente consultamos la distribución normal estandarizada y obtenemos el valor Z que corresponde al 95% del área bajo la curva que es igual a ± 1,96.

Con los datos obtenidos debemos calcular el error estándar de la proporción, de la misma manera en que calculamos el error estándar de la media en el ejercicio anterior. La ecuación de cálculo es la siguiente:

$$EE_p = \sqrt{\frac{p \times (1-p)}{n}}$$

Reemplazando los términos en la fórmula tenernos que el error estándar de la proporción es:

$$EE_p = \sqrt{\frac{0.87 \times (1-0.87)}{500}} = 0{,}015$$

Para calcular el Intervalo de Confianza para la proporción, utilizamos la misma ecuación que la empleada para el intervalo de confianza de la media, reemplazando en ella los términos correspondientes. Así tenemos que en este caso la ecuación se define como:

$$LI = p \pm (Z \times EE_p)$$

Así, los intervalos de confianza buscados son:

Superior: 0,87+(1,96*0,015)= 0,8994

Inferior: 0,87-(1,96*0,015)= 0,855

Obtenemos como resultado lo siguiente; 85,5% ≤ p ≤ 89,94%. Ello se interpreta de la siguiente manera: la tasa de promoción entre primaria y secundaria en la ciudad de Córdoba Capital se encuentra comprendida entre el 90% y el 85,5% aproximadamente (con un 95% de confianza en la estimación).

Comparación de proporciones

Con frecuencia ocurre que se toman estadísticas continuas y es posible obtener datos poblacionales históricos que nos sirven de referencia para indagar en qué medida la situación reflejada en ellos pudiera haber cambiado. Para explayarnos sobre esta situación sirvámonos del siguiente ejemplo: De acuerdo a datos tomados en el año 2001, se constató que el 67% de los puestos directivos de las principales instituciones escolares del país estaba ocupado por varones. En el año 2010, se realiza un nuevo estudio tomando una muestra aleatoria de 249 puestos directivos de las principales instituciones educativas del país, de los cuales la proporción de varones es de 64,66%.

Un investigador social se pregunta si existen diferencias significativas entre las proporciones encontradas en estos dos periodos respecto al género y el puesto dentro de las instituciones educativas. Para resolver estadísticamente este problema, primero debemos reducir el problema a proporciones, tal como se definen para el estudio realizado en 2010, en tal caso tenemos que:

Proporción de varones en la muestra: p=0.6466 (64,66%)
Proporción de mujeres en la muestra: q= 1-p= 1- 0.6466= 0.3534 (35,34%)

Se trata de ver ahora si la proporción encontrada en 2001 es diferente a la encontrada en 2010. La prueba de proporciones consiste en comparar un valor Z empírico, con un valor Z teórico definido a partir de la probabilidad asignada a la diferencia. El valor de Z empírico se obtiene aplicando la siguiente fórmula:

$$Z_{empirico} = \frac{p_{muestra} - p_{población}}{\sqrt{\frac{p_{población} - q_{población}}{n_{muestra}}}}$$

Reemplazando los valores del problema en la fórmula tenemos que:

$$Z_{empirico} = \frac{0{,}6466 - 0{,}67}{\sqrt{\frac{0{,}67 - 0{,}33}{249}}} = -0{,}65$$

Para el ejemplo que estamos desarrollando, el valor de Z empírico es -0.65. El valor Z teórico definido a partir de la probabilidad asignada a la diferencia, es el equivalente al intervalo de confianza para la estimación del parámetro. En este caso el valor Z teórico se asigna con anterioridad al desarrollo de la investigación, pero a los fines de este ejemplo lo asignamos en este punto.

Digamos que deseamos tener una seguridad del 99% para determinar si existe una diferencia entre las proporciones. Por lo ya visto, sabemos que el valor Z teórico correspondiente a ese valor de área bajo la curva de distribución normal es igual a Z±2,58. Por lo tanto para aseverar que la diferencia entre las proporciones es significativa, el valor de Z empírico, deberá ser mayor que 2,58 o menor que -2,58.

Solo aquellos valores que se ubiquen en los extremos definidos por el valor Z teórico dado, serán considerados suficientes como para determinar que la diferencia es estadísticamente significativa. En este ejemplo tenemos que $Z_{empirico}$-0.65>$Z_{teórico}$-2,58, por lo cual no es posible aceptar que existe una diferencia de proporciones y que la diferencia observada se debe a la aleatoriedad de los datos.

Comparación de proporciones muestrales

En el caso anterior se contaba con una proporción encontrada en una población y se la comparó con la proporción encontrada en una muestra. Cuando solo se cuenta con proporciones extraídas de muestras, igualmente se puede determinar si una diferencia de proporciones resulta significativa. En tal caso se utiliza una distribución teórica muestral de diferencia de pares de proporciones de muestreo, pertenecientes a una misma población. Esta distribución, tal como hemos visto hasta aquí, también sigue un modelo normal, y en este caso, un supuesto central para esta distribución teórica es que la media de las diferencias de proporciones es igual a cero. Bajo el mencionado supuesto, el error estándar de la distribución muestral de diferencias de proporciones será igual a:

$$\sigma_{dif} = \sqrt{\frac{p_1 \times q_1}{n_1} + \frac{p_2 \times q_2}{n_2}}$$

Veamos un ejemplo de la manera en que procederíamos al comparar dos proporciones muestrales. Supongamos que, mediante un muestreo aleatorio de la población, se cuenta con dos proporciones referidas a la proporción de egresados de carreras docentes que no cuentan con empleo como docentes en las ciudades de Paraná y Concordia, tomando como referencia que ambas forman parte de la población de la Provincia de Entre Ríos. La muestra 1 (n_1=500) tomada de Paraná muestra que la proporción de egresados que no trabajan como docentes es de 15% (p_1). La muestra 2 tomada en Concordia (n_2=280) muestra la proporción de egresados que no trabajan como docentes es del 25% (p_2). Nos interesa determinar si las dos proporciones encontradas difieren significativamente, o si la diferencia observada es debida al azar. Para ello debemos establecer un intervalo de confianza donde situar la diferencia encontrada. Como ya sabemos que la diferencia de pares de proporciones de muestreo pertenecientes a una misma población sigue una distribución normal, el valor del intervalo lo establecemos utilizando los valores de área bajo la curva normal estandarizada. Para este ejemplo, vamos a establecer un intervalo de confianza del 95%, por lo tanto, el valor crítico de Z será ±1.96. Ahora procedemos a calcular el error estándar de las diferencias de proporciones:

$$\sigma_{dif} = \sqrt{\frac{0{,}15 \times 0{,}85}{500} + \frac{0{,}25 \times 0{,}75}{280}} = 0{,}0179$$

Necesitamos ahora calcular el valor Z empírico, con el cual compara el valor Z crítico establecido en el intervalo de confianza elegido. Para esto debemos resolver la siguiente ecuación:

$$Z_{empírico} = \frac{p_1 - p_2}{\sigma_{dif}}$$

Con los datos que se tienen, el valor Z empírico es igual a

$$Z_{empírico} = \frac{0{,}15 - 0{,}25}{0{,}0179} = -5{,}58$$

Dado que -5,58<-1.96 se puede concluir que la diferencia de proporciones es significativamente diferente. Es decir, no son producto de las variaciones debidas al azar o error de muestreo.

En este ejemplo debemos tener en cuenta dos cuestiones importantes que, por la simplicidad de la presentación, hemos dejado de lado. Primero; lo esbozado en los párrafos precedentes tiene sentido cuando se comparan proporciones extraídas de la misma población, dado que este es un supuesto central para aplicar la distribución de muestreo de diferencia de proporciones. En este caso, la población de referencia de donde se extraen las muestras es la provincia de Entre Ríos. Segundo; es necesario definir con anterioridad el sentido de la comparación, dado que, en la segunda ecuación, debemos calcular la diferencia de proporciones según lo que hayamos establecido previamente como p_1 y p_2. Generalmente, se parte de algún supuesto de base para establecer qué proporción es cada una, de lo contrario perdería todo sentido la comparación.

Estimación de parámetros: conceptos teóricos

Estimación puntual: en la estimación puntual se da como valor estimado de μ el valor de la media muestral. De acuerdo al teorema del límite central el valor de la media muestral será igual o estará muy próximo a μ. La diferencia entre la media muestral y μ, corresponde al error propio del muestreo, pero tal diferencia será pequeña. Entonces, en una estimación puntual se dirá que el valor de μ estará próximo a la media muestral, o bien que la diferencia entre μ y la media muestral será pequeña.

Estimación por intervalos: en la estimación por intervalos, se establecen límites de confianza entre los cuales es más probable que se encuentre la media poblacional. En otras palabras, los límites determinan la probabilidad dentro de los cuales μ esté contenido. Esta probabilidad se llama probabilidad de confianza y el intervalo marcado por los límites se llama intervalo de confianza.

Dado que la distribución de muestreo de las medias sigue el modelo normal, se sostiene que, si la media poblacional es μ, existe una probabilidad del 95% de que la media de la muestra aleatoria extraída de esa población tome un valor μ-(1.96*σ_M) y μ+(1.96*σ_M). Dicho de otra manera; el 95% de las muestras aleatorias extraídas de la población, tendrán una media comprendida entre μ-(1.96*σ_M) y μ+(1.96*σ_M).

Si se cuenta con una muestra aleatoria de la población, el intervalo de confianza para μ, se establece de la siguiente manera: $\bar{x}$-(1.96*EEM) y $\bar{x}$+(1.96*EEM).

Ello implica que:

a) El intervalo de confianza está centrado en la media observada en una muestra aleatoria extraída de la población.

b) Existe un 5% de probabilidades de que ese intervalo no contenga al parámetro μ. Dado que la probabilidad de ocurrencia es baja (5 de cada 100), estaremos en condiciones de afirmar que en la mayoría de los casos el intervalo de confianza contiene el valor de μ.

c) En raras ocasiones se conocerá el error estándar de la media o σ_M, por lo que también deberá ser estimado a partir de la muestra. Para ello, se utiliza la desviación estándar de la muestra, tal como mostró en los ejemplos.

Guía de Actividades N° 6

Estimación de parámetros

Objetivos de la actividad:

a) Comprender los conceptos básicos de estimación de parámetros poblacionales a partir de estimadores muestrales.
b) Determinar si un parámetro muestral corresponde a una población determinada y si dos muestras pertenecen a una misma población.

Actividad 1: Se realiza un estudio para la provincia de Córdoba tomando como muestra un aleatoria de la población a un total 1200 personas que cuentan con empleo en instituciones educativas. Se quiere conocer antigüedad promedio medida en años que llevan en su labor. Los resultados muestran que el promedio en años de trabajo es de 14,5 años y la desviación estándar es de 7 años.

Dado que se cuenta con una muestra representativa de la población de interés, se decide estimar el parámetro poblacional años promedio de antigüedad en la población de docentes, de la provincia de Córdoba.

Promedio de años de estudio: 14,5

Desvío Estándar: 7

Tamaño de la muestra: n=1200

Establezca los límites de confianza para la estimación del parámetro poblacional con un 99%. Considere que el valor de Z=± 2,58 cubre el 99% del área bajo la curva. Para ello calcule el intervalo de confianza.

Actividad 2: Se realiza una evaluación de la promoción de los alumnos en el Nuevo Plan de Estudios, particularmente en las escuelas PROA. Para ello, se toma una muestra de 300 alumnos de 14 escuelas de la provincia. Se encuentra que en la muestra, 239 alumnos son efectivamente promovidos sin ningún tipo de aplazo. La tasa resultante es 239/300=0.80. Con estos datos se estima que la tasa de promoción en las escuelas PROA en la provincia de Córdoba es del 80%.

Proporción de alumnos promovidos: p=0.80 (80%)

Tamaño de la muestra: 300

En este caso se pide fijar el nivel de precisión de la estimación en el 95%. Este valor en una distribución normal estandarizada corresponde a un valor de Z de ± 1,96.

A partir de estos datos se le pide que calcule el error estándar de la proporción en la estimación de la proporción población con el nivel de precisión presentado. Además, se le pide que calcule los límites superior e inferior.

Actividad 3: Considerando un conjunto de datos tomados pro SADOP en el año 2010 (ficticio), se constató en un censo que el 69% de los puestos cargos docentes de las principales instituciones escolares del país estaba interino. En el año 2019, se realiza un nuevo estudio tomando una muestra aleatoria de 350 puestos docentes de las principales instituciones educativas del país, de los cuales la proporción de interinos es de 63.66%.

Proporción de interinos en la muestra: p=0.6366 (63,66%)

Proporción de no interinos en la muestra: q= 1-p= 1- 0.6366= 0.3634 (36,34%)

Un investigador social se pregunta si existen diferencias significativas entre las proporciones encontradas en estos dos periodos dentro de las instituciones educativas, es decir si ha disminuido significativamente la proporción de interinos en las instituciones principales instituciones privadas del país con un 99% de confianza. ¿Es posible aceptar que existe una diferencia entre proporciones y que la diferencia se debe a la aleatoriedad de los datos?

Actividad 4: Mediante un muestreo aleatorio de la población, que busca conocer los docentes en situación de desplazarse a más de 5km de su hogar para cumplir sus funciones se cuenta con dos proporciones referidas a la proporción de docentes que viven a más de 5km de su trabajo en las ciudades de Córdoba y Río Cuarto, tomando como referencia que ambas forman parte de la población de la Provincia de Córdoba. La muestra 1 (n_1=250) tomada de Córdoba muestra que la proporción de docentes a más de 5km de su lugar de trabajo es de 65% (p_1). La muestra 2 tomada en Río Cuarto (n_2=180) es del 48% (p_2).

Queremos saber si las dos proporciones encontradas difieren significativamente, o si la diferencia observada es debida al azar. Con un intervalo de confianza del 99%, por lo tanto, el valor crítico de Z será ±2.58.

Actividad 5: Un profesor de educación física ha registrado el promedio de tiempo en segundos en que sus alumnos realizan una prueba de atletismo. Cuando fue a compararlo con los resultados notó que se han mezclado sus papeles y no reconoce cuál de los 7 promedios guardados es del curso en cuestión. ¿Le ayudamos a reconocerlo? El resultado de la prueba actual es de 183 segundos, con una desviación estándar de 10 segundos y el grupo que realizó la tarea son 30 estudiantes.

Promedio de años de estudio: 183s

Desvío Estándar: 10s

Tamaño de la muestra: n=30

Establezca los límites de confianza para la estimación del parámetro poblacional con un 99% de cada una de los siguientes registros. Considere que el valor de Z=± 2,58 cubre el 99% del área bajo la curva. Para ello calcule el intervalo de confianza. Luego elija el/los que pueden contener al registro actual y justifique su respuesta.

Registros anteriores:

Registros en el cuaderno	**Media**	**DS**	**n**
1	160	15	28
2	172	18	32
3	182	25	26
4	210	7	35
5	200	10	30
6	198	16	29
7	199	22	31

Actividades en PSPP N° 6:

Vamos a abrir nuevamente el programa PSPP. En esta ocasión trabajaremos nuevamente con la base de datos que descargamos del Instituto Nacional de Estadísticas y Censos (INDEC). Para ello abrimos el documento que guardamos con el nombre C2_EANNA URBANA_CENTRO.SAV.

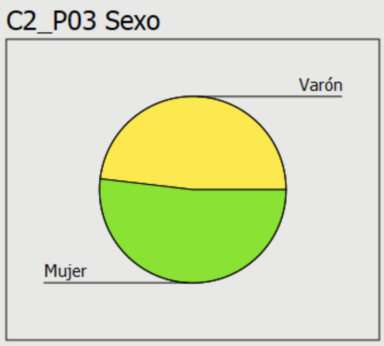
Una vez abierta la base de datos vamos a calcular nuevamente la proporción de mujeres en la región Centro. Para ello vamos al menú Analizar, opción Estadística descriptiva, pedimos Frecuencias y una vez abierto el cuadro seleccionamos la variable C2_P03 Sexo y la llevamos a la derecha. Si hubiera alguna variable de un procesamiento anterior en el cuadro de variables proceda a quitarla.

Además, proceda a destildar todos los estadísticos que estén marcados. También se le solicita que en gráficos pida un diagrama de sectores. No vamos a describir este procedimiento porque lo hemos realizado en unidades anteriores.

El resultado debería ser:

C2_P03 Sexo

Etiqueta de Valor	*Valor*	*Frecuencia*	*Porcentaje*	*Porcentaje Válido*	*Porcentaje Acumulado*
Varón	1	5090	48,15	48,15	48,15
Mujer	2	5482	51,85	51,85	100,00
	Total	10572	100,0	100,0	



Describa los resultados obtenidos en el cuadro y el gráfico considerando todas las columnas. Simultáneamente, a partir de estos datos, se le pide que calcule el error estándar de la proporción de mujeres en la estimación de la proporción poblacional de mujeres con un nivel de precisión de un 95% y se le pide que establezca los límites superior e inferior del intervalo:

………

………

………

………

………

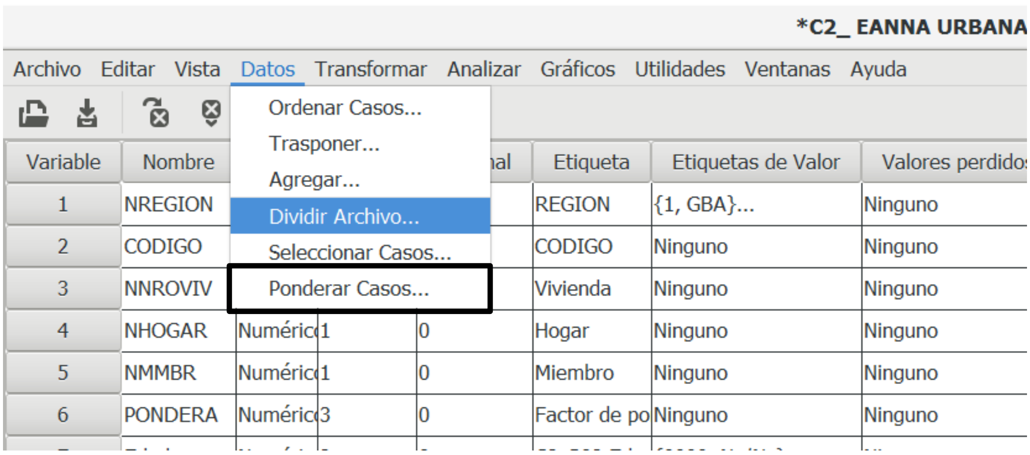
………

………
………
………
………
………
………
………
………
………
………
……

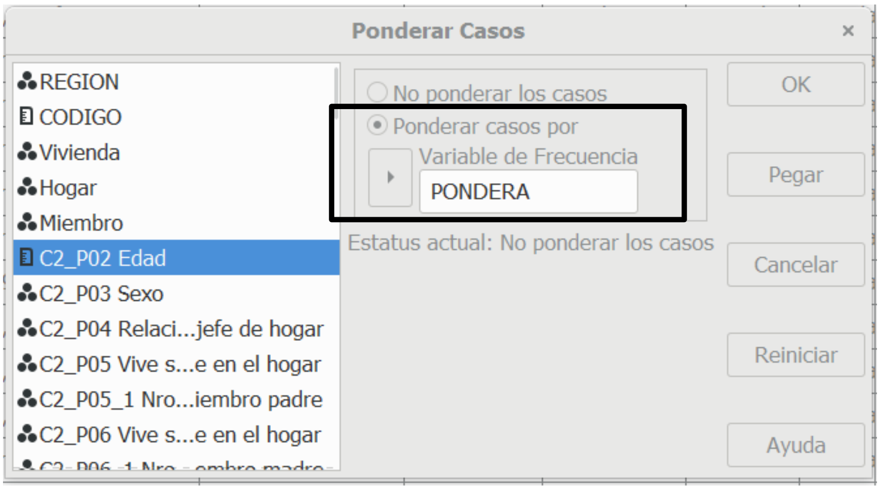
Una vez realizado esto, vamos a ponderar la base. La ponderación es una operación expansión de la base de datos que se realiza sobre cada caso para asignarle un valor equivalente al de la población. Cómo habíamos dicho en unidades anteriores, esta es una muestra representativa y el ponderador expande los casos a la población que la muestra representa.

Debido al error muestral, las estimaciones de totales en una encuesta por muestreo varían de muestra a muestra, cómo indica INDEC en su documento metodológico sobre estimación en la EPH. En una encuesta a individuos u hogares se pueden estimar más o menos personas con nivel universitario, más o menos los ocupados, etc., dependiendo de la composición de los hogares que residen en las viviendas. Según el tamaño de la muestra, y lo que estemos estudiando, las fluctuaciones tendrán mayor o menor impacto en las estimaciones. Para ello, se usan los ponderadores.

Ahora vamos a realizar una nueva operación para solicitar la proporción de varones y de mujeres, pero habiendo ponderado la base. Para ello, vamos a la pestaña de Datos y seleccionamos Ponderar casos:



Seleccionamos como variable de ponderación la que tiene la etiqueta Factor de ponderación, que luego de pasarla al box de ponderación aparecerá con el nombre de PONDERA.

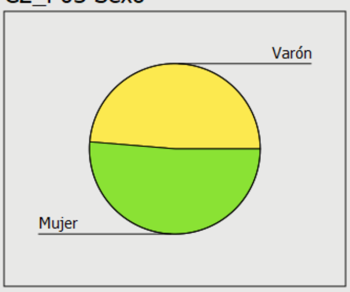


Como veremos la proporción ha cambiado levemente, pero la cantidad de casos ha pasado de 10572 a 2.622.429 personas. Ello es porque el ponderador ha expandido la muestra a un número aproximado al de la población que se quiere estudiar.

C2_P03 Sexo

Etiqueta de Valor	*Valor*	*Frecuencia*	*Porcentaje*	*Porcentaje Válido*	*Porcentaje Acumulado*
Varón	1	1276957	48,69	48,69	48,69
Mujer	2	1345472	51,31	51,31	100,00
	Total	2622429	100,0	100,0	

C2_P03 Sexo



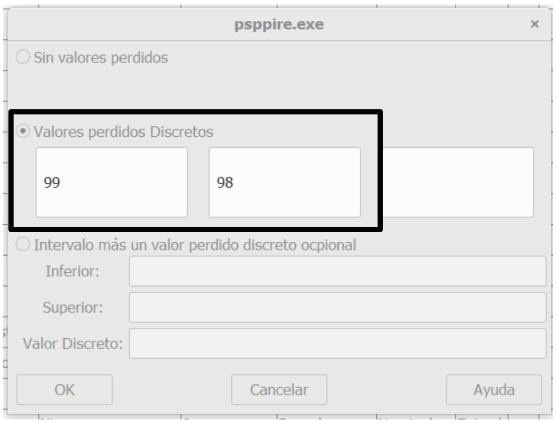
Usted puede notar que el porcentaje de Mujeres es del 51,31%, mientras que antes de ponderar era del 51,85%. La ponderación ha afectado levemente la proporción de mujeres.

Veamos qué sucede si tomamos otras variables, como el Nivel educativo más alto, la variable C2_P13. Cuando pedimos una tabla de Frecuencias el resultado será:

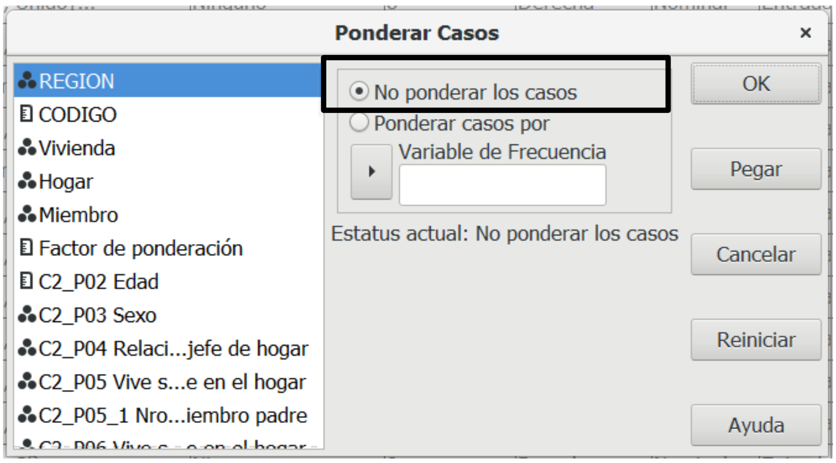
C2_P13 Nivel educativo más alto

Etiqueta de Valor	*Valor*	*Frecuencia*	*Porcentaje*	*Porcentaje Válido*	*Porcentaje Acumulado*
Jardín/Preescolar	1	2479	,09	,26	,26
EGB (1 a 9 año)	2	13809	,53	1,45	1,72
Primario (1 a 6/7 grado)	3	199891	7,62	21,06	22,77
Polimodal (1 a 3/4 año)	4	19455	,74	2,05	24,82
Secundario (1 a 5 ó 6/7 año)	5	452412	17,25	47,66	72,48
Superior no universitario	6	100384	3,83	10,57	83,05
Universitario	7	151699	5,78	15,98	99,03
Posgrado universitario	8	4989	,19	,53	99,56
Educación especial	98	2500	,10	,26	99,82
Ns/Nr	99	1698	,06	,18	100,00
	.	1673113	63,80	Perdidos	
	Total	2622429	100,0	100,0	

Lo primero que haremos es dar por perdidos los casos Ns/Nr. Para ello vamos a la Vista de variables, buscamos la variable C2_P13 y damos por perdido en la columna Valores perdidos el Valor discreto 99 y el 98, como pueden ver a continuación:



Luego de ello, vamos a quitar el ponderador del siguiente modo, vamos a Datos, Ponderar casos y seleccionamos No ponderar los casos:



Procedemos a pedir nuevamente la tabla de Frecuencias:

C2_P13 Nivel educativo más alto

Etiqueta de Valor	*Valor*	*Frecuencia*	*Porcentaje*	*Porcentaje Válido*	*Porcentaje Acumulado*
Jardín/Preescolar	1	11	,10	,30	,30
EGB (1 a 9 año)	2	45	,43	1,24	1,54
Primario (1 a 6/7 grado)	3	861	8,14	23,71	25,25
Polimodal (1 a 3/4 año)	4	67	,63	1,85	27,10
Secundario (1 a 5 ó 6/7 año)	5	1792	16,95	49,35	76,45
Superior no universitario	6	337	3,19	9,28	85,73
Universitario	7	507	4,80	13,96	99,70
Posgrado universitario	8	11	,10	,30	100,00
	.	6926	65,51	Perdidos	
Educación especial	98	9	,09	Perdidos	
Ns/Nr	99	6	,06	Perdidos	
	Total	10572	100,0	100,0	

Lo primero que usted puede notar que ahora volvemos a tener 10.572 casos. Por otra parte, hay 15 casos que han sido dado por perdidos, 9 correspondientes a Educación especial y 6 a la categoría NS/NR. Hemos dado por perdido estos casos porque así la variable permite discriminar por nivel, antes, no podíamos precisar si estos casos pertenecían a un nivel u otro.

Ahora podremos calcular la proporción de individuos que han accedido al nivel Superior universitario (Superior no universitario, Universitario y Posgrado Universitario), que son el total válido menos un 76.45%, que declaran niveles menores al superior no universitario. Esto da por resultado que un 23.54% de la población ha accedido al nivel Superior Universitario. Se le pide que calcule con estos datos el EE_M y los límites inferior y superior del intervalo.

………
………
………
………
………
………
………
………
………
……………………………………………………………………

Ahora volvemos a ponderar y se les solicita indicar si el valor ponderado para la población cae dentro del rango de estimación de la muestra. Para ello vamos a cotejar el valor de la proporción de individuos con acceso al nivel superior universitario en relación con la muestra no ponderada.

C2_P13 Nivel educativo más alto					
Etiqueta de Valor	*Valor*	*Frecuencia*	*Porcentaje*	*Porcentaje Válido*	*Porcentaje Acumulado*
Jardín/Preescolar	1	11	,10	,30	,30
EGB (1 a 9 año)	2	45	,43	1,24	1,54
Primario (1 a 6/7 grado)	3	861	8,14	23,71	25,25
Polimodal (1 a 3/4 año)	4	67	,63	1,85	27,10
Secundario (1 a 5 ó 6/7 año)	5	1792	16,95	49,35	76,45
Superior no universitario	6	337	3,19	9,28	85,73
Universitario	7	507	4,80	13,96	99,70
Posgrado universitario	8	11	,10	,30	100,00
	.	6926	65,51	Perdidos	
Educación especial	98	9	,09	Perdidos	
Ns/Nr	99	6	,06	Perdidos	
	Total	10572	100,0	100,0	

……
……….
……….
……….
……….
……….
……….
……….
……….
…………………………………………………………………………

Finalmente, vamos a quitar por última vez el ponderador y procedemos a obtener una submuestra de los casos a un 10%. Para ello vamos a Datos, Seleccionar casos y le indicamos que vamos a seleccionar una muestra aleatoria de un 10% de los casos, como hemos realizado anteriormente. Luego de ello, volvemos a pedir la tabla de frecuencias de la variable C2_P13.

C2_P13 Nivel educativo más alto					
Etiqueta de Valor	*Valor*	*Frecuencia*	*Porcentaje*	*Porcentaje Válido*	*Porcentaje Acumulado*
Jardín/Preescolar	1	2	,19	,55	,55
EGB (1 a 9 año)	2	4	,37	1,10	1,65
Primario (1 a 6/7 grado)	3	91	8,53	25,07	26,72
Polimodal (1 a 3/4 año)	4	6	,56	1,65	28,37
Secundario (1 a 5 ó 6/7 año)	5	179	16,78	49,31	77,69
Superior no universitario	6	25	2,34	6,89	84,57
Universitario	7	55	5,15	15,15	99,72
Posgrado universitario	8	1	,09	,28	100,00
	.	702	65,79	Perdidos	
Educación especial	98	1	,09	Perdidos	
Ns/Nr	99	1	,09	Perdidos	
	Total	1067	100,0	100,0	

Vamos a considerar que esta es una muestra de nuestra población. Por lo tanto, vamos a extraer la proporción maestral correspondiente al nivel Superior universitario, que da 22.31%. Se le pide que calcule el EE_M de la muestra y los límites inferior y superior con un nivel de precisión del 99%. Se le pide además que indique si el promedio de nuestra "población" está dentro de los errores de estimación.

……
………
………
………
………
………
………
………
………
……………………………………………………………………

Capítulo 7

Prueba de hipótesis sobre la media paramétrica

La investigación científica se inscribe fuertemente en lo que se denomina prueba de hipótesis. Es un tema que abarca numerosos capítulos de los manuales de metodología de la investigación, por lo que en este apartado solo haremos referencia a la manera de plantear una hipótesis de investigación que pueda ser puesta a prueba con un modelo estadístico. Es este otro aspecto de lo que denominamos estadística inferencial, que se refiere a la manera en que los datos empíricos pueden ser analizados con diferentes modelos matemáticos. De los muchos modelos estadísticos utilizados para la prueba de hipótesis nos centraremos en dos: el modelo normal y el modelo chi cuadrado. Ambos tienen la propiedad de que son excelentes ejemplos para generalizar a cualquier otro modelo estadístico.

En este capítulo vamos a ver cómo se realiza una prueba de hipótesis. En este marco procederemos a distinguir la media paramétrica y la media maestral y comprender el valor de un modelo estadístico para relacionarlas. Luego vamos a conocer como diseñar un modelo estadístico de prueba de hipótesis por medio de la distribución normal y vamos a comparar medias con este modelo. Finalmente, vamos a aprender a realizar prueba de hipótesis para la media paramétrica.

Comencemos con lo básico que es comprender la estructura de una hipótesis. Una hipótesis se parece a lo que conocemos como una conjetura o suposición, pero tiene la propiedad de explicitar en qué medida es posible demostrar con cierto grado de rigor su veracidad o falsedad. Las hipótesis estadísticas son un caso especial de lo que mencionamos anteriormente ya que siempre es posible expresarlas de manera tal que, mediante una colección de datos, podamos ponerlas a prueba.

Veamos un ejemplo: un grupo de investigadoras verifica que el promedio en matemáticas de las pruebas APRENDER 2019 es más alto que el obtenido en las pruebas APRENDER 2016 para los alumnos de Córdoba Capital. En base a esta información pueden suponer que el promedio del operativo de evaluación APRENDER 2022 será también más alto. Esta suposición puede tomar la forma de una hipótesis estadística si la redactamos de la siguiente manera: el promedio de los alumnos de Córdoba Capital en el operativo nacional de evaluación APRENDER 2022, será mayor que el obtenido en el APRENDER 2019. También podemos expresarlo estadísticamente:

$$\boxed{H_1: \mu_{\text{APRENDER-2022}} > \mu_{\text{APRENDER-2019}}}$$

El planteo estadístico de la hipótesis de investigación se lee del siguiente modo: la media poblacional en el operativo de evaluación APRENDER 2022, será mayor a la media poblacional obtenida en APRENDER 2019. Esta hipótesis es válida para la población de escolares de Córdoba Capital y para el área de matemáticas.

Las investigadoras también podrían haber hecho otra suposición si los datos fueran diferentes. Tomemos el caso de que los promedios obtenidos por los escolares de Córdoba Capital para las evaluaciones de matemáticas de ONE 2013 sean un poco más bajas que las obtenidas en el APRENDER 2016. Las investigadoras ahora podrían suponer que habrá diferencias más grandes entre el APRENDER 2019 y el operativo APRENDER 2022. El problema que enfrenan es que no pueden decidir el sentido de la diferencia, tal como se planteaba en la hipótesis anterior. En tal caso la hipótesis queda redactada de la siguiente manera: el promedio de los alumnos de Córdoba Capital en el operativo nacional de evaluación Aprender 2022, será diferente que el obtenido en el APRENDER 2019. También podemos expresarlo estadísticamente:

$$\boxed{\text{H1: } \mu \text{ aprender-2022} \neq \mu \text{ APRENDER-2019}}$$

El planteo estadístico de la hipótesis de investigación se lee de la siguiente manera: la media poblacional en el operativo de evaluación APRENDER 2022, será diferente a la media poblacional obtenida en APRENDER 2019. Nuevamente esta hipótesis es válida para la población de escolares de Córdoba Capital y para el área de matemáticas.

Podemos establecer cualquier hipótesis respecto de las diferencias en los promedios o proporciones. Incluso podemos establecer hipótesis sobre la magnitud de las correlaciones y asociaciones entre variables. En estadística, las hipótesis plantadas siempre descansan en un modelo matemático que permite ponderar las diferencias o magnitudes y así decidir sobre la validez de la hipótesis.

La media paramétrica y muestral

Como se habrá notado en el apartado anterior, la comprobación de las hipótesis planteadas se corroborará una vez obtenidos los datos del operativo APRENDER de 2022. En general, la prueba de hipótesis puede plantearse en diseños que no requieren parámetros, es decir, datos poblacionales, y la consecuente dilación en el tiempo que esto supone. El procedimiento por el cual obtenemos muestras representativas de una población, nos permite obtener datos con los que realizar comparaciones, estableciendo previamente hipótesis de investigación adecuadas. Como ya se ha mencionado, todos los escolares de una escuela pueden representar una población bien definida. En lo que sigue vamos a proponer ejemplos, a partir de los cuales se pueden plantear hipótesis como las vistas anteriormente. Supongamos que un grupo de docentes decide cambiar el método de evaluación en el área de las ciencias naturales, considerando que el método usado en la actualidad no refleja apropiadamente todas las capacidades y competencias de los estudiantes. La hipótesis implícita en este caso es que el método tradicional produce una subestimación de las calificaciones, las cuales quedarían mejor plasmadas por el nuevo método de evaluación. Tomando en consideración el rendimiento promedio en el área de ciencias naturales, la hipótesis comparativa quedaría expresada de la siguiente manera:

$$H_1 = \bar{x}_{nuevo} > \bar{x}_{tradicional}$$

La hipótesis refleja la suposición que el rendimiento promedio de los escolares evaluados en el área de ciencias naturales con el nuevo método, sea mayor que el obtenido si la evaluación se realiza mediante el método tradicional.

Una cuestión a tener en cuenta en este punto es que la prueba de hipótesis se realiza sobre lo que se conoce como hipótesis nula, y en este caso, solo se presenta la hipótesis de investigación. Más adelante se verá la relación entre ambas.

Los datos y el modelo estadístico

Antes de continuar con el planteo de hipótesis, debemos adentrarnos al tema de los modelos estadísticos. Intentaremos hacerlo apelando a la mayor simplicidad matemática sin perder la rigurosidad.

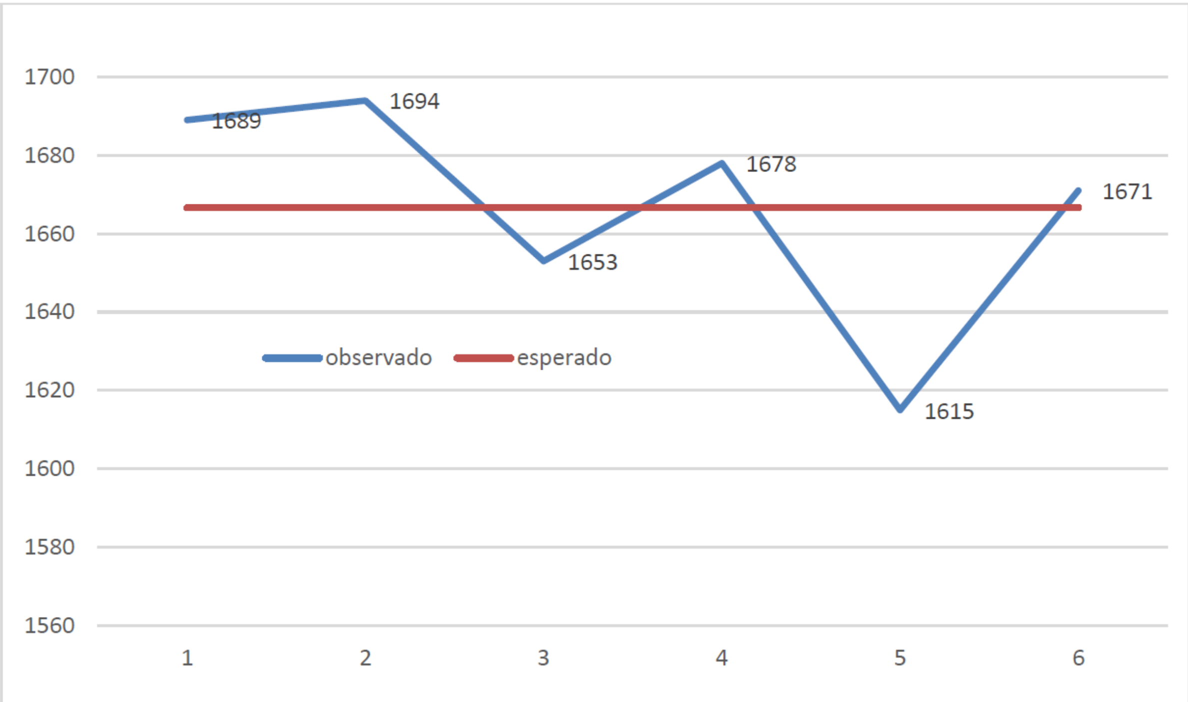
Comencemos con una distribución de datos teórica que proviene del lanzamiento de un dado. El lanzamiento del dado puede considerarse como un espacio de posibilidades perfectamente conocido, pues sabemos que el resultado variará entre 1 y 6. Suponiendo que el dado no tiene ninguna falla, si lanzamos una serie infinita de veces el dado el valor de probabilidad para cada cara del dado es igual a 1/6 o bien, p(x)=0,16, siendo x cualquier valor entre 1 y 6. En la siguiente tabla vemos el valor de x que varía entre 1 y 6, luego la probabilidad de que se obtenga un valor dado en un lanzamiento del dado para un espacio muestra igual a: E{1,2,3,4,5,6}.

Sabemos de antemano que esa posibilidad es la probabilidad de x, que es igual a p(x)=1/6. La columna f(x) contiene la frecuencia de una simulación de diez mil tiradas de un dado bajo el espacio muestral E{1,2,3,4,5,6}, con una probabilidad de 1/6 para cada evento. La columna f(x)p(x) es la frecuencia esperada en una distribución de 10.000 tiradas con la probabilidad exacta de que cada resultado ocurra 1/6 de las veces.

Valor	p(x)	f(x)p(x)	f(x)
1	1/6	1666.66	1639
2	1/6	1666.66	1584
3	1/6	1666.66	1633
4	1/6	1666.66	1638
5	1/6	1666.66	1577
6	1/6	1666.66	1621

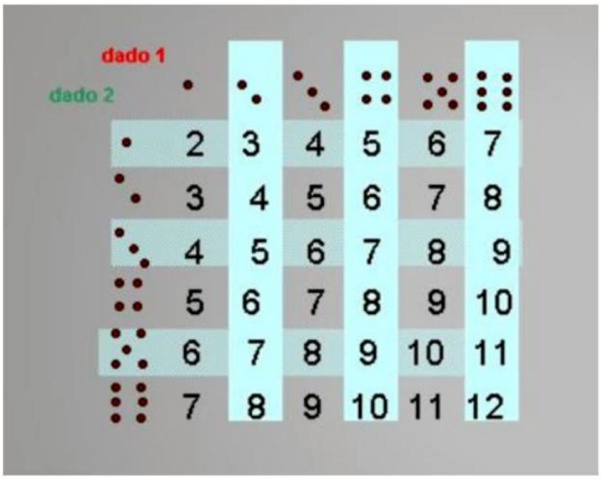
Entonces, si pudiéramos repetir el evento una cantidad infinita de veces tendríamos una distribución de datos perfectamente uniforme, pues sabemos que la ley de los grandes números determina que todo proceso aleatorio (en este caso la tirada de un dado), se empareja a su probabilidad de ocurrencia (en este caso, que cada valor aparezca 1/6 de las veces). Si tiramos el dado unas diez mil veces esperamos que cada valor de E{1,2,3,4,5,6}, se repita unas 1666 veces.

Sin embargo, las distribuciones empíricas (en este caso los resultados de la simulación), se aproximan a los valores esperados, pero no se igualan perfectamente. En la siguiente gráfica representamos los valores de f(x) y f(x)p(x); f(x) representa las frecuencias observadas en la modelización para una secuencia de 10.000 tiradas de un dado, f(x)p(x) es el modelo teórico de las frecuencias esperadas para una secuencia de 10.000 tiradas de un dado.



Vemos que existen diferencias entre el valor esperado y el observado, y a esto se lo llama fluctuación aleatoria. La fluctuación aleatoria ocurre porque la simulación solo toma un número finito de observaciones, en este ejemplo fueron 10.000, y el valor esperado se cumple cuando el valor es infinito. Es de notarse que, a pesar de las fluctuaciones en torno al valor esperado, éstas no son tan diferentes; todas están próximas a 1.666. Sobre los datos que hemos presentado podemos afirmar que los valores observados pertenecen a una distribución uniforme del espacio muestral E{1,2,3,4,5,6} con una probabilidad 1/6 para cada evento.

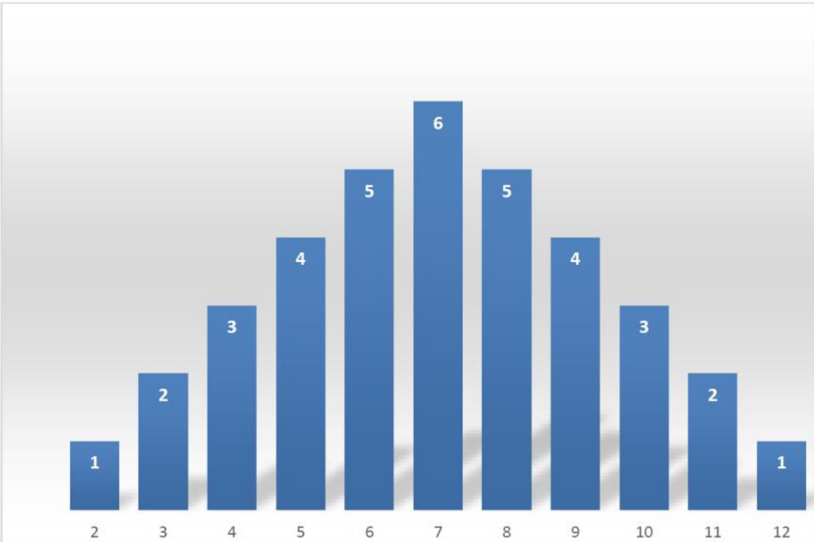
Ahora, repetiremos todo el proceso anterior, pero teniendo en cuenta el resultado del lanzamiento de dos dados. El espacio muestral sobre el que ahora trabajamos resulta de la suma de los valores de las dos caras del dado, por lo tanto, tenemos una variación de 2 a 12. Es decir, E{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}. Ahora, cada uno de los valores de E no tiene la misma posibilidad de ocurrir; solo hay una forma en de obtener un 2 en la suma de las caras del dado, y hay seis formas de obtener un 7 en la suma de las caras del dado. Para ilustrar la combinatoria de resultados posibles véase la siguiente figura.



En la primera fila están las seis caras del dado 1, en la primera columna las seis caras del dado 2. Lo que buscamos es el resultado de la suma de los valores de las caras de ambos dados, por lo cual existen 36 resultados posibles. Si observamos la cuadricula, vemos que solo hay una forma de obtener un 2, que es cuando se obtiene 1 en ambas caras de los dados. Luego vemos que hay dos formas de obtener un 3 que es cuando obtenemos 1 en la cara del primer dado y obtenemos 2 en la cara del segundo dado, y a la inversa. Siguiendo este razonamiento vemos que hay tres formas de obtener un 4 y así para los demás resultados. Por lo tanto, es factible construir una tabla de frecuencias con las probabilidades de ocurrencia de cada resultado de la suma de las caras del dado, que denominaremos *x*.

Valor de *x*	f(*x*)	p(*x*)
2	1	1/36
3	2	2/36
4	3	3/36
5	4	4/36
6	5	5/36
7	6	6/36
8	5	5/36
9	4	4/36
10	3	3/36
11	2	2/36
12	1	1/36

La gráfica resultante de esas frecuencias se muestra a continuación y en ella podemos ver qué valores de la suma de las dos caras de un dado tienen mayores chances de ocurrir. Dicho de otro modo, la gráfica nos muestra que hay seis combinaciones posibles para obtener un 7, hay cinco combinaciones posibles para un 6 y 8, y así para los otros resultados tenemos que existen menos combinaciones posibles.

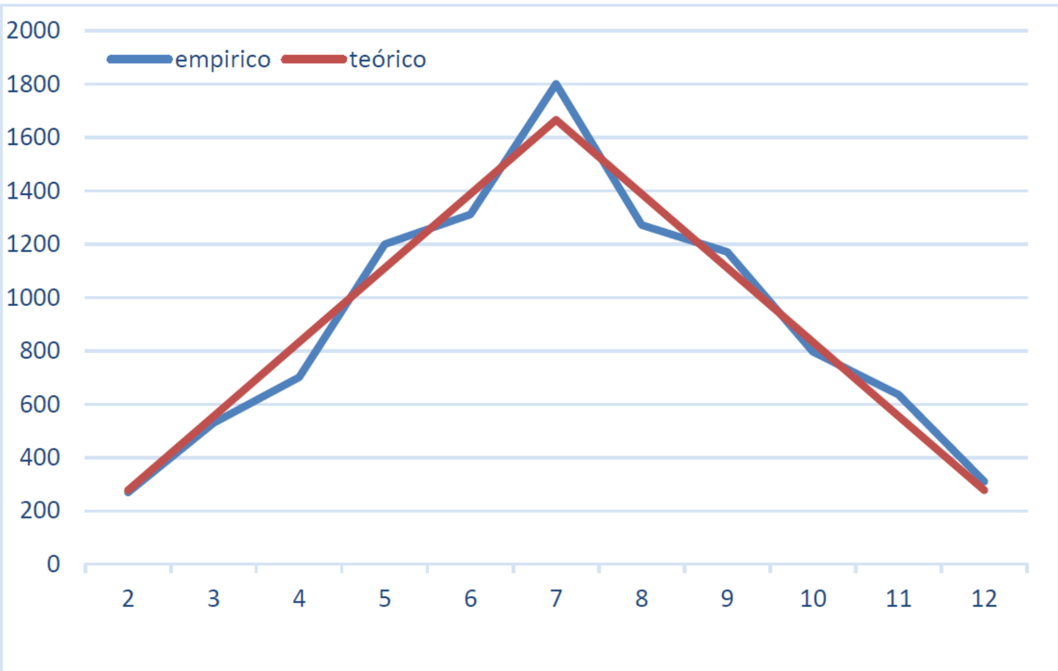


Anteriormente, conociendo la probabilidad de ocurrencia del valor de la tirada de un dado obtuvimos una distribución uniforme que empleamos para simular los resultados de una tirada de 10.000 veces un dado. Podemos repetir el mismo proceso de simulación para la tirada de dos dados, basándonos en las probabilidades teóricas de obtener cualquier valor de E{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}. En la tabla, esas probabilidades se representan por p(*x*) como vimos anteriormente, ese el modelo probabilístico esperado de la distribución de frecuencias de la suma de las dos caras de un dado en una tirada de diez mil veces.

Luego f(x)p(x) son las frecuencias efectivamente observadas en la simulación de la suma de las dos caras de un dado en una tirada de diez mil veces.

	p(x)	**f(x)p(x)**
2	277,7	270
3	555,5	530
4	833,3	801
5	1111,1	1200
6	1388,8	1312
7	1666,6	1801
8	1388,8	1272
9	1111,1	1171
10	833,3	797
11	555,5	536
12	277,7	310

Al graficar los valores de la tabla vemos que el modelo teórico es una distribución perfectamente simétrica con su valor máximo en el resultado x=7. La distribución empírica se aproxima bien a ese modelo teórico a pesar de las fluctuaciones aleatorias.



Lo que hemos visto hasta aquí fueron dos simulaciones de distribuciones de datos basadas en un modelo probabilístico conocido. Las distribuciones observadas se aproximaban mucho a los modelos utilizados, y esta diferencia entre el modelo teórico y los datos empíricos le llamamos fluctuación aleatoria. Otra manera de expresar lo dichos es afirmando que una distribución de datos empírica puede modelizarse mediante una distribución de probabilidades conocida. Cuando ello ocurre decimos que la distribución de datos empírica se ajusta a un modelo teórico y que las fluctuaciones aleatorias son mínimas.

Es así que las propiedades matemáticas del modelo teórico pueden utilizarse para describir el conjunto de datos empíricos. Existen muchos modelos estadísticos que sirven a ese propósito, es por ello que para cada conjunto de datos se estima una bondad de ajuste a un modelo dado.

La distribución normal como modelo estadístico

En los ejemplos anteriores establecimos previamente un modelo probabilístico y procedimos a simular un conjunto de datos tomando como referencia el modelo propuesto. En el apartado en que describimos la distribución normal, se mencionó que ésta podía ser utilizada como modelo estadístico y que es uno de los más usados en estadística.

Comencemos describiendo un conjunto de datos E{2, 4, 6}. Para este conjunto calculamos las medidas de tendencia central y las de dispersión, que son las siguientes:

promedio: 4

Varianza: 2.66

Desviación Estándar: 1.63

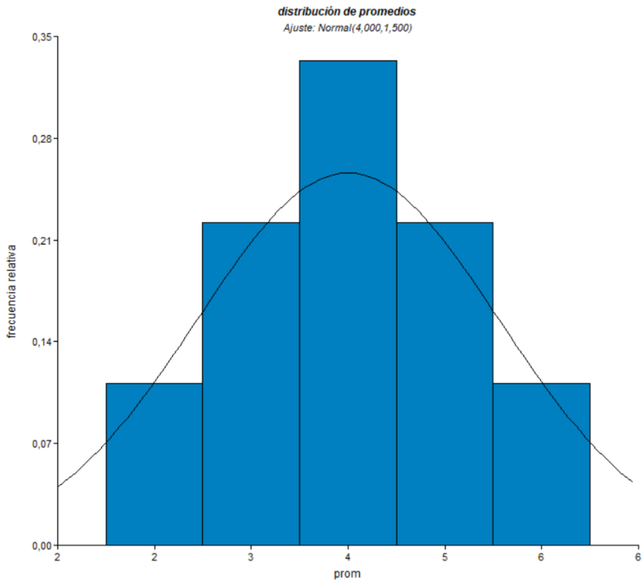
Ahora, vamos a calcular todos promedios del conjunto E{2, 4, 6}, pero tomando valores de a pares. Es decir, vamos a generar una distribución de promedios a partir del conjunto E{2, 4, 6}. Como en el ejemplo de los dados, es posible pasar los valores a una tabla cuyas filas y columnas sean los valores del conjunto E{2, 4, 6}, y las celdas contengan los promedios de los pares de valores.

Distribución de promedios	**2**	**4**	**6**
2	2	3	4
4	3	4	5
6	4	5	6

Esta información la pasamos a una tabla que contenga la frecuencia de con la que ocurre cada promedio y la probabilidad de obtener cada uno de ellos.

Promedio	**f(x)**	**p(x)**
2	1	1/9
3	2	2/9
4	3	3/9
5	2	2/9
6	1	1/9

Ahora podemos trazar un histograma de la distribución de promedios y comprobar empíricamente a qué tipo de distribución probabilística podemos asimilarla.



Vemos que la distribución de los promedios del conjunto E{2, 4, 6} da como resultado un gráfico que se asemeja a una distribución normal. Es decir, la distribución de promedios del conjunto E{2, 4, 6} tomado de a dos valores, produce una distribución normal. Esta distribución está centrada en el valor promedio 4, siendo menos frecuentes los promedios 2 y 5, y menos frecuente aún los promedio 2 y 6.

Esto se cumple para cualquier distribución de promedios de una variable aleatoria y ha sido descripto como teorema del límite central. Este teorema tiene toda una formulación matemática, pero intentaremos explicarlo de manera sencilla. Nuestro conjunto de datos original es E{2, 4, 6} cuyo promedio es 4. Si procediéramos a seleccionar aleatoriamente dos valores cualquiera de ese conjunto vemos que es más probable escoger algunos valores que otros.

Para simplificar esto, diremos que p(x) es la probabilidad de obtener un promedio dado y lo expresamos en valor porcentual. Luego tabulamos la distribución teórica dado los valores probabilísticos asignados de obtener cualquier promedio del conjunto E{2, 4, 6} cuyos valores son tomados de a pares en una simulación de un muestro aleatorio de 10.000 ensayos. El resultado se muestra en la siguiente tabla:

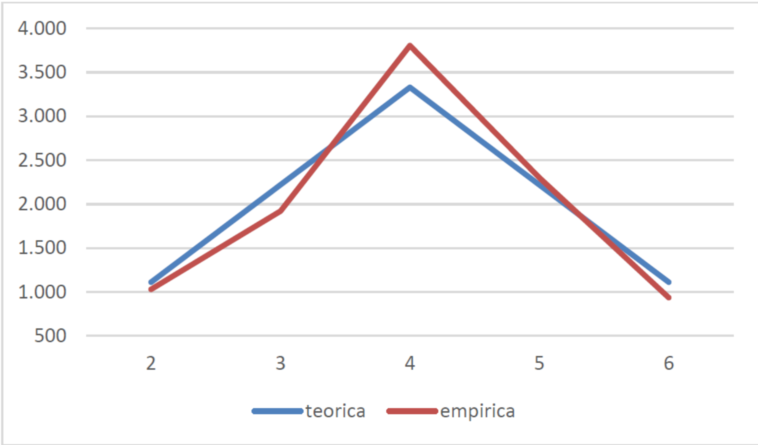
Promedio	$\boldsymbol{p(x)}$	$\boldsymbol{f(x)}$ **teórica**	**f(*x*)**
2	11%	1.110	1030
3	22%	2.220	1920
4	33%	3.330	3805
5	22%	2.220	2301
6	11%	1.110	934

Según se lee en la tabla, la probabilidad de obtener un promedio de 2 en este experimento ficticio es solo del 11%, mientras que obtener un promedio de 4 tiene una chance del 33%. Teóricamente en un ensayo de 10.000 repeticiones, obtendríamos 1.100 muestras con promedio de 2 y unas 3.300 muestras con promedio de cuatro.

Existe una sola posibilidad en que podamos obtener un promedio de dos, y es que en el par de valores extraído del conjunto E{2, 4, 6}, ambos tengan valor 2. Pero la combinación de pares para obtener un promedio de 4 es mayor, dado que esto ocurre cuando la muestra de pares de valores de E se compone de 6+2, 2+6 y 4+4. Entonces, el proceso de repetir el muestreo aleatorio de promediar pares de valores extraídos del conjunto E tiende a centrarse en 4, y los promedio 2 y 6 que solo pueden ocurrir por la combinación de dos valores, tienen a ser memos frecuentes y ocupar los extremos de la distribución.

Es por ello que el modelo estadístico que mejor refleja el procedimiento es la distribución normal. La columna final tiene los datos de una simulación de 10.000 promedios basados en la combinación de dos valores del conjunto E{2, 4, 6}. Como vemos, se cumple que en el largo plazo los valores empíricos se aproximan a los valores empíricos con bastante fiabilidad.

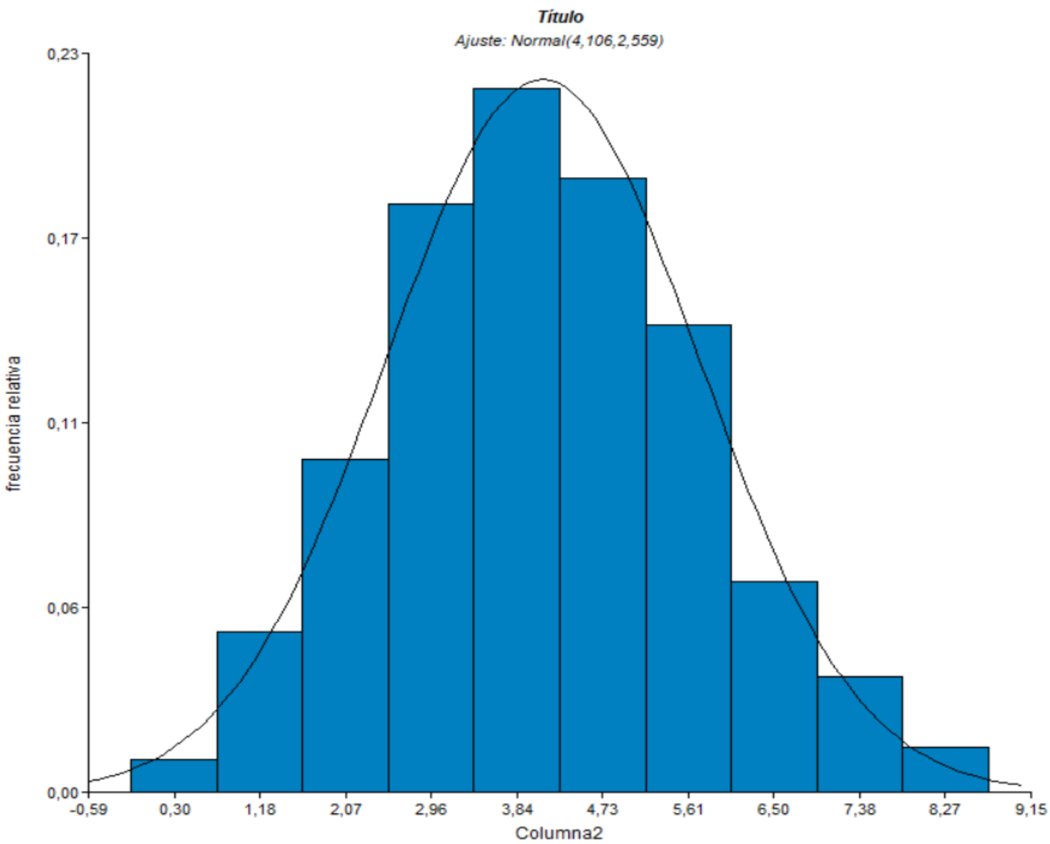
En el siguiente gráfico se muestran las curvas resultantes del modelo teórico de distribución de frecuencias, y el resultado empírico de una simulación de 10.000 promedio del conjunto E{2, 4, 6} basados en p(x). Nuevamente vemos la semejanza de las distribuciones y las fluctuaciones aleatorias entre el valor teórico y el empírico.



Sigamos con nuestro hipotético conjunto de datos, E{2, 4, 6}, y supongamos que fue extraído por medio de un muestreo aleatorio de una población con los siguientes parámetros:

μ	σ
4,11	1,6

Sabiendo que el conjunto E{2, 4, 6} tiene un promedio de 4, concluimos que se encuentra muy próximo al centro de la distribución y no hay razón para dudar de ello. Si observamos el gráfico de la población de origen del conjunto E{2, 4, 6} comprobaremos de manera intuitiva lo dicho anteriormente.



Comparación de distintos promedios con el modelo estadístico de la distribución normal estándar

Vamos a tomar ahora tres conjuntos de datos e intentaremos decidir si pertenecen a una población con parámetros

μ	σ
4,11	1,6

Los conjuntos son los siguientes:

Conjunto A	Conjunto B	Conjunto C	Conjunto D
3,28	2,77	6,93	-3,5
6,7	8,05	7,05	-7,2
2,11	4,99	8,09	6,8
$\bar{x}$=4,03	$\bar{x}$=5,27	$\bar{x}$=7,36	$\bar{x}$=-1,3

Ordenando los conjuntos por el valor de su promedio, vemos que, de A a D, se alejan cada vez más del valor paramétrico. Ahora bien, si consideramos que en la distribución normal estandarizada los limites Z=±1,96 cubren el 95% de la distribución, es posible suponer que cualquier valor de Z por encima o por debajo de Z=±1,96 puede considerarse poco probable, y si este es el caso, es lícito suponer que esa muestra: a) no pertenece a la distribución original, o b) es una muestra posible pero atípica para la distribución original.

Para considerar la posición de cada conjunto en la muestra original, vamos a transformar los valores promedios a puntuación Z y comprobar si se encuentran dentro del área comprendida por los valores Z=±1,96.

Conjunto A	Conjunto B	Conjunto C	Conjunto D
$\bar{x}$=4,03	$\bar{x}$=5,27	$\bar{x}$=7,36	$\bar{x}$=-1,3
Z=-0,05	Z=0,725	Z=2,03	Z=-3,38
p=0,961	p =0,469	p =0,043	p =0,001

Como puede observarse, el conjunto A y B se encuentran dentro de los límites del valor Z=±1,96. Por lo tanto no hay razón para dudar de que pertenecen al conjunto original de datos con parámetro μ=4,11. En cambio, el conjunto C y D están fuera de los límites de Z=±1,96. Recordemos una vez más que en una distribución normal estándar, el valor Z=±1,96 cubre el 95% de la distribución, por lo tanto, los promedios de los conjuntos C y D están por fuera de esa área. En otras palabras, los consideramos ahora como conjuntos atípicos para una población con μ=4,11, y es lícito suponer que no pertenezcan a esa población.

Para afirmar lo dicho anteriormente hemos agregado a la tabla la probabilidad de ocurrencia de una muestra con los promedios de los conjuntos A, B, C y D, si consideramos que efectivamente todos ellos pertenecen a una población con μ =4,11. Para una interpretación más directa de la probabilidad, transformamos los valores a porcentajes y tenemos que para el conjunto A la probabilidad de pertenecer a la población de origen es de 96,1%, para el conjunto B la probabilidad de pertenecer a la población de origen es 46,9%. Ahora la probabilidad del conjunto C de pertenecer a la población de origen es de 4,3% y la del conjunto D solo del 0,1%.

Supongamos ahora que tenemos que responder a la siguiente pregunta: ¿dado los valores del conjunto C, es posible afirmar que pertenecen a una población con μ=4,11? Dado que la probabilidad de ocurrencia de una muestra de tales características es menor al 5%, es posible afirmar que ese conjunto de valores pertenece a otra población.

Prueba de hipótesis sobre media paramétrica

Ahora vamos a utilizar todos los conceptos descriptos anteriormente en un planteo de prueba de hipótesis utilizando el modelo de la distribución normal. Veamos el siguiente ejemplo: Una investigadora analizó los promedios finales de la carrera de ingeniería de la UNC. Los docentes de esa Facultad emplean una evaluación en cada materia de puntuaciones de 1 al 10. Las estadísticas universitarias hasta el año 2012 indican que el promedio final de la carrera sigue una distribución normal con $\bar{x}$=6,33 y de=1,8. Esta investigadora sospecha que durante los últimos años el promedio puede haber variado. Por ello, se decide hacer un estudio, registrando el promedio al final de la carrera de una muestra aleatoria de 200 alumnos.

La sospecha de la investigadora tiene que redactarse como una hipótesis de investigación y, además, plantearse en términos estadísticos.

H_1: el promedio al final de la carrera de ingeniería de la UNC es diferente de los valores registrados hasta el año 2012.
H_0: el promedio al final de la carrera de ingeniería de la UNC no es diferente de los valores registrados hasta el año 2012.

H_1: $\bar{x}$ 2012$\neq\bar{x}$ actual
H_0: $\bar{x}$ 2012=$\bar{x}$ actual

Nótese que en este ejemplo hemos introducido una nueva hipótesis que recibe el nombre de Hipótesis Nula, que designamos como H_0 y que es la negación de la Hipótesis de Investigación, que designamos como H_1, y que es sobre la que hemos estado trabajando hasta el momento.

La pregunta es ¿por qué dos hipótesis? La respuesta es simple, porque la hipótesis que someteremos a prueba con los datos disponibles es la Hipótesis Nula. Hacemos esto puesto que H_0 es exacta y, por tanto, se ajusta a una distribución de probabilidad conocida. Para este ejemplo la distribución de probabilidad está especificada en la población de origen y es el modelo normal. Estadísticamente ambas hipótesis quedan redactadas de la siguiente manera.

$$\boxed{\begin{array}{l} H_0: \mu = 6{,}33 \\ H_1: \mu \neq 6{,}33 \end{array}}$$

Las hipótesis planteadas hasta aquí deben entenderse de la siguiente manera. La investigadora supone que el promedio de los estudiantes ya no es el mismo que aquel registrado hasta el año 2012, por lo tanto, los datos tomados de aquella población no representan adecuadamente a los estudiantes actuales. Sin embargo, no se puede decidir si el nuevo parámetro poblacional es mayor o menor que $\mu = 6{,}33$. Es por ello que la hipótesis de investigación solo especifica una diferencia.

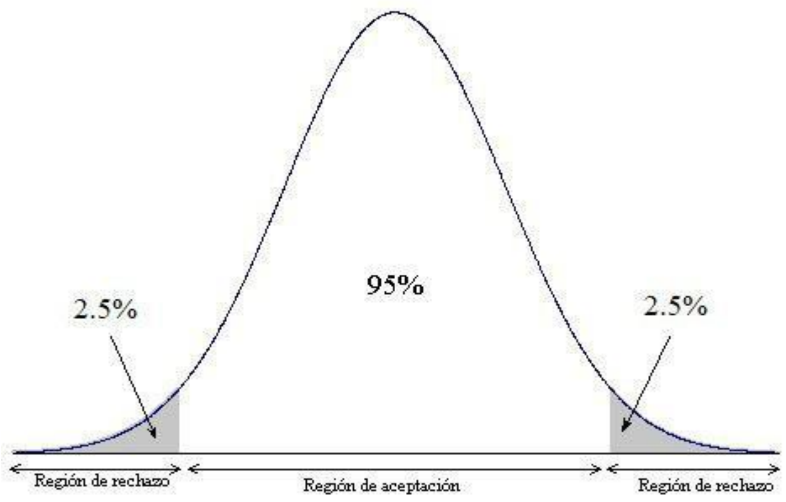
A este tipo de hipótesis se las denomina bidireccionales en tanto no anticipan la dirección de la diferencia. De esta manera, si el valor de μ encontrado en una nueva muestra de estudiantes tomada en el presente año está próxima al valor promedio 6,33, no hay razón para suponer que el valor del parámetro ha cambiado y por tanto se descarta la hipótesis de investigación. Al hacerlo se sostiene la hipótesis nula y en tal caso, la sospecha de la investigadora se resolvería concluyendo que no han existido cambios sustantivos en el promedio final de los egresados de la carrera de ingeniería y por tanto es lícito suponer que el parámetro para esa población es $\mu=6{,}33$.

Hemos mencionado que no descartaremos la hipótesis nula siempre que el valor promedio al final de la carrera tomado de estudiantes en la actualidad, esté próximo a $\mu=6{,}33$. Esto quiere decir que esperamos encontrar alguna fluctuación aleatoria en el valor de μ, pero que se encuentre dentro de ciertos límites de probabilidad. Estadísticamente, debemos establecer un límite dentro del cual aceptamos que el valor de μ sea diferente de 6,33 y que tal diferencia se deba a las fluctuaciones aleatorias del conjunto de datos. Una vez establecido este límite, este servirá para determinar que las variaciones observadas no pueden ser explicadas por simples fluctuaciones aleatorias y por tanto no es factible sostener que el valor del parámetro es el que se especifica en la H_0.

Como vimos anteriormente la distribución normal nos ofrece un modelo para evaluar la validez de la H_0. Según hemos visto por el teorema del límite central, cualquier promedio de una muestra aleatoria extraída de una población tiende a ubicarse en el centro de esa distribución, de tal modo que: $\mu \sim \bar{x}$. Esto quiere decir que un conjunto de muestras tomadas aleatoriamente de esa población se ajusta a una distribución normal.

Para determinar qué tan diferente es el valor de la muestra obtenida respecto del parámetro, solo basta comparar la desviación de μ respecto de $\bar{x}$. Para hacerlo utilizamos la puntuación Z correspondiente al modelo de la distribución normal estándar. Según hemos visto, el valor Z=±1,96 se corresponde con un área que cubre el 95% de la distribución. De este modo podemos utilizar ese valor de referencia para determinar cuándo un promedio muestral se ha apartado lo suficiente del valor paramétrico para considerar que tal diferencia ya no puede atribuirse a simples variaciones aleatoria, sino que el valor del promedio muestral pertenece a una distribución con un valor paramétrico diferente.

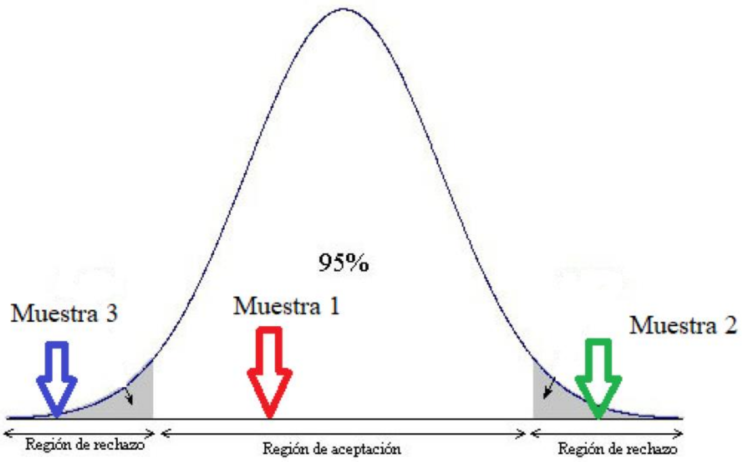
Bajo tales condiciones diremos que rechazaremos la H_0 planteada toda vez que encontremos un valor que sea mayor o menor que Z=±1,96. Siendo así, hemos definido dos áreas dentro de la distribución normal, una en la que especifica los valores dentro de los cuales la H_0 debe ser aceptada, y otra en la que H_0 debe ser rechazada. Gráficamente esto se representa de la siguiente manera:



Tomando el valor crítico Z=±1,96 tenemos que dentro de esos límites existe un 95% de posibilidades de que una muestra aleatoria de la población tenga un valor próximo al parámetro cuando es cierta H_0. Rechazaremos la H_0 cuando su posibilidad de certeza sea baja, que según se muestra en la gráfica, ello ocurre cuando el valor muestral se aparta del centro de la distribución hacia alguno de los extremos. En tal caso, se espera que H_0 sea cierta solo el 5% de las veces. En el caso de las hipótesis bidireccionales, al no especificar el sentido de la diferencia respecto al parámetro, se toman los dos extremos de la distribución, es por eso que la zona de rechazo de H_0 se encuentra repartida en un 2.5% en cada extremo de la distribución.

Vamos a introducir ahora un nuevo término para explicar cómo interpretar las zonas de aceptación y rechazo de H_0. Esto términos son el Error tipo I y el Error tipo II. Cometemos un Error Tipo I toda vez que rechazamos una H_0 que es verdadera, y cometemos el Error tipo II toda vez que no rechazamos una H_0 que es falsa. Ambos tipos de errores son complementarios y al ajustar uno de ellos también ajustamos el otro. Para simplificar la explicación, trabajaremos solo sobre el Error tipo I.

En el contexto de la prueba de hipótesis, queremos rechazar todas las H_0 que son falsas, pero eso es imposible debido a las fluctuaciones aleatoria que vimos anteriormente. Es por eso que el modelo probabilístico sobre el que trabajemos, nos proporciona valores críticos sobre los cuales establecer zonas de aceptación y rechazo de la H_0. En la figura que se muestra a continuación, se han situado los valores promedios de tres muestras tomadas de una población con parámetro μ. Recordemos que la H_0 establece que $\mu \sim \bar{x}$.



Por lo tanto, el promedio de la muestra 1 no es suficientemente diferente del parámetro como para rechazar la H_0. En cambio, los valores de los promedios de las muestras 2 y 3 sí son los suficientemente diferentes como para rechazar la H_0.

Definimos un valor crítico para la aceptación o rechazo de la H_0 representándolo con la letra α (alfa). Si queremos tener una certeza del 95% de que al rechazar la H_0 estamos haciendo lo correcto, el valor de alfa es igual a 0.05. El valor de α se expresa como una probabilidad, pero es fácil comprender que esa probabilidad puede expresarse en porcentaje al multiplicarla por 100. Por tanto, α=0.05 es la probabilidad de cometer el Error Tipo I el 5% de las veces al rechazar una H_0. Otra manera de decirlos es, que α=0.05 nos da una certeza de que el 95% de las veces no rechazaremos H_0 cuando es cierta. Por lo tanto, minimizamos el Error Tipo I, si rechazamos H_0 cuando el valor promedio obtenido fuera el de las muestras 2 y 3. En el caso de la muestra 1 lo mejor es no rechazar la H_0.

Volvamos ahora al ejemplo que planteamos al comienzo. Una investigadora analizó los promedios finales de la carrera de ingeniería de la UNC. Los docentes de esa Facultad emplean una evaluación en cada materia de puntuaciones de 1 al 10. Las estadísticas universitarias hasta el año 2012 indican que el promedio final de la carrera sigue una distribución normal con M=6,33 y de=1,8. Esta investigadora sospecha que durante los últimos años el promedio puede haber variado. Por ello, se decide hacer un estudio, registrando el promedio al final de la carrera de una muestra aleatoria de 200 alumnos. La hipótesis planteada y su modelo estadístico es el siguiente:

H_1: el promedio al final de la carrera de ingeniería de la UNC es diferente de los valores registrados hasta el año 2012.

H_0: el promedio al final de la carrera de ingeniería de la UNC no es diferente de los valores registrados hasta el año 2012.

H_1: $\bar{x}_{2012} \neq \bar{x}_{actual}$ H_1: $\mu \neq 6.33$

H_0: $\bar{x}_{2012} = \bar{x}_{actual}$ H_0: $\mu = 6.33$

Dado que la investigadora no predice una diferencia en ninguna dirección H_1 es Bidireccional, por lo tanto, estaremos usando los dos extremos del modelo de distribución normal estándar. Es decir, la prueba será de dos extremos.

Ahora debemos establecer el valor crítico de rechazo de H_0, es decir, debemos darle un valor a α. Para continuar con lo argumentado hasta aquí, establecemos α=0.05. Siendo este el valor establecido, sabemos ahora que los valores de Z críticos que establecen las zonas de aceptación y rechazo de H_0 corresponden a los valores Z± 1.96.

Para finalizar el ejemplo, supongamos que los resultados del estudio realizado por la investigadora, determinaron que el promedio registrado en la muestra es de $\bar{x}$=5.75 con una *de*= 1,6.

Para obtener el valor Z observado que compararemos con el valor Z crítico, necesitamos obtener el error estándar de la media. La distribución del error estándar de la media es el mismo que utilizamos para la estimación de parámetros, salvo que el denominador utiliza el valor del tamaño muestral.

Continuando con el ejemplo, diremos que para el estudio que realiza la investigadora, se utiliza una muestra aleatoria de tamaño *n*=200. De los cual resulta que el error estándar de la media para ese tamaño muestral es de 0,113. La siguiente ecuación resume lo dicho para la distribución de datos del ejemplo:

$$EE_{\bar{x}} = \frac{DE}{\sqrt{n}} = \frac{1.6}{\sqrt{199}} = 0{,}113$$

Sabemos que el valor Z crítico para este ejemplo se estableció en Z±1.96. Debemos, encontrar el valor de Z observado para realizar la comparación. Sabemos que para encontrar una puntuación Z debemos proceder con la siguiente ecuación:

$$Z = \frac{x - \bar{x}}{DE}$$

Para la prueba de hipótesis, debemos reemplazar en la ecuación el promedio muestral por el parámetro, y la desviación estándar por el $EE\bar{x}$, de modo que el valor Z observado se resuelve de la siguiente manera.

$$Z_{observado} = \frac{\bar{x} - \mu}{EE_{\bar{x}}}$$

Asumiendo que, en la investigación realizada, se encontró que el promedio final de los estudiantes de ingeniería es de 6,86, el cálculo del valor Z observado da como resultado:

$$Z_{observado} = \frac{6.86 - 6.33}{0.113} = 4.6$$

La zona de rechazo de H_0 se encuentra en algún valor entre Z±1.96. Encontramos que el valor de Z observado supera largamente el valor de Z crítico establecido para rechazar H_0. Asumimos entonces que, siendo $Z_{observado} > Z_{crítico}$ para el valor de $\alpha=0.05$, se rechaza H_0. Por lo tanto, la hipótesis de la investigadora es correcta al suponer que el parámetro ha cambiado.

Habiendo rechazado la H_0 nos queda como válida H_1, y en este caso corresponde recalibrar el parámetro μ para el promedio de los estudiantes de ingeniería al final de la carrera realizando para ello una nueva estimación de este.

La hipótesis sobre la que realizamos todo el planteo es de tipo bidireccional en tanto no anticipamos que el nuevo parámetro estuviera por encima o por debajo del parámetro actual. Sin embargo, se podría haber previsto esa diferencia tal como se hizo con la hipótesis planteada al principio de este apartado sobre los operativos de evaluación.

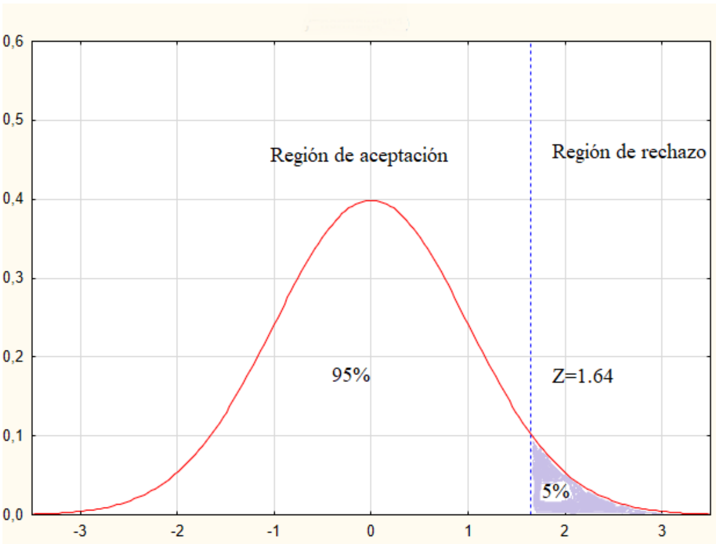
Retomemos el ejemplo sobre el que estamos trabajando y reformulemos H_1. Una investigadora analizó los promedios finales de la carrera de ingeniería de la UNC. Los docentes de esa Facultad emplean una evaluación en cada materia de puntuaciones de 1 al 10. Las estadísticas universitarias hasta el año 2012 indican que el promedio final de la carrera sigue una distribución normal con M=6,33 y de=1,8. Esta investigadora sospecha que durante los últimos años el promedio ha variado y en la actualidad es más alto. Por ello, se decide hacer un estudio, registrando el promedio al final de la carrera de una muestra aleatoria de 200 alumnos.

El modelo estadístico de la hipótesis planteada es el siguiente:

H_1: $\bar{x}_{2012} < \bar{x}_{actual}$ H_1: $\mu > 6{,}33$
H_0: $\bar{x}_{2012} = \bar{x}_{actual}$ H_0: $\mu = 6{,}33$

Vemos que el planteo supone ahora que la diferencia entre el parámetro establecido a partir de los datos de 2012 es menor al actual. Por su parte H_0 sigue siendo la misma en tanto se trata de negar a la H1.

De todo este planteo, lo único que se modifica es la zona en donde estableceremos el punto de aceptación o rechazo de la hipótesis nula. Si examinamos la H_1 vemos que es unidireccional en tanto determina que el nuevo parámetro es mayor, de modo que el valor del punto de corte Z debe situarse sobre el extremo derecho de la curva. Este nuevo punto de corte debe establecer las zonas de aceptación y rechazo de H_0. Si no cambiamos el ejemplo más que en el punto mencionado, el nuevo valor de Z critico se establece es $Z_{critico}$=1,64. Gráficamente ello se representa como se muestra a continuación.



Vemos que toda la zona de aceptación queda delimitada por el área de la izquierda de la curva y el valor Z crítico se sitúa en el extremo derecho de la distribución, a partir del valor 1,64.

En el ejemplo, habíamos encontrado que el valor de Z observado= 4.6 que es mayor que 1.64. De este modo, bajo este nuevo planteo de hipótesis también rechazaríamos H_0.

Finalmente, cabe destacar que, para las hipótesis unidireccionales, se utilizan los extremos de la distribución normal, de acuerdo al modo en que se plantea la hipótesis de investigación, es decir, de acuerdo a la manera en que anticipamos en que ocurrirá la diferencia. Solo para las hipótesis bidireccionales utilizamos ambos extremos de la distribución simultáneamente.

Guía de Actividades N° 7

Prueba de hipótesis y distribución normal como modelo estadístico.

Objetivos de la actividad:

a) realizar planteos estadísticos y la prueba de hipótesis correspondiente, a partir del planteo de una hipótesis de investigación y una hipótesis nula

b) aplicar la distribución norma como modelo estadístico para la resolución de problemas de prueba de hipótesis.

Actividad 1: El baremo de un test de rendimiento académico aplicado cada año a estudiantes universitarios de la Universidad Autónoma de Madrid arroja como media un puntaje de 50. Algunos investigadores sospechan que con el correr de los años la media se ha modificado, y plantean una hipótesis al respecto. Para verificar tal suposición se tomó una muestra de 200 estudiantes de esa misma universidad. Los datos indican que la media muestral resultante fue de 47,5, con un desvío estándar de 10.

Con los datos ofrecidos se pide:

a) Plantee la expresión estadística de la hipótesis nula y de la hipótesis de investigación,

b) Determine si la hipótesis de investigación es: UNIDIRECCIONAL – BIDIRECCIONAL, UN EXTREMO – DOS EXTREMOS,

c) Determine el valor de Z crítico, para una probabilidad de error tipo I con α=0.05,

d) Determine si es plausible rechazar la hipótesis nula y esboce una conclusión.

Valores críticos Puntuaciones Z

	α= 0.05	α= 0.01
1	1.65	2.33
2	1.96	2.58

Actividad 2: Un equipo de psicólogos ha comprobado que, en cierta población infantil con dificultades motrices, el tiempo (en minutos) empleado en realizar actividades manuales pautadas, sigue un modelo de distribución normal con media 17 minutos. Como parte de las actividades de un programa terapéutico, un grupo de 36 niños seleccionados en dicha población, asistió a sesiones especiales de juegos que permiten desarrollar las capacidades motoras por medio de ejercicios específicos. Estos niños, realizaron la actividad manual en un tiempo medio de 14 minutos, con un desvió estándar 8 minutos.

A partir de esta información, y planteando la prueba estadística para un nivel de significación del 1% (α=0,01) ¿podríamos rechazar la hipótesis de que el tiempo medio, tras haber sido entrenados en ejercicios de coordinación motora, resultará sistemáticamente inferior a 17 minutos?

Actividad 3: En una muestra aleatoria de estudiantes de secundaria de quinto y sexto año (*n*=600), se encontró que el 63,4% tenía acceso a internet domiciliario. En comparación con los datos proporcionados por la última encuesta de estudiantes y docentes realizada por el Ministerio de Educación de Córdoba, se observa una diferencia; para los datos censales dicha proporción es de 41,6%. Un grupo de docentes sospecha que éstos últimos datos se encuentran desactualizados.

Según lo esbozado, se solicita: a) Plantee la expresión estadística de la hipótesis nula y de la hipótesis de investigación, b) Determine el valor de Z crítico para la comparación de proporciones con una probabilidad de error tipo I con α=0.05, c) Determine si es plausible rechazar la hipótesis nula y esboce una conclusión.

Actividades en PSPP N° 7:

Ejecutamos el programa PSPP y en esta ocasión vamos a trabajar nuevamente con la base de datos que descargamos del Instituto Nacional de Estadísticas y Censos (INDEC). Para ello abrimos el documento que guardamos con el nombre C2_EANNA URBANA_CENTRO.SAV.

Una vez abierta la base de datos, vamos a retomar la variable que trabajamos anteriormente. Para ello, vamos primeramente a ponderar y luego a pedir los estadísticos descriptivos de C2_P13 Nivel educativo más alto. Ya hemos dado por perdidos los valores 98 y 99 anteriormente. En caso que no haya quedado grabado, usted deberá hacerlo nuevamente. De este modo, podrá pedir una tabla de frecuencias con los estadísticos respectivos, como la siguiente:

C2_P13 Nivel educativo más alto

Etiqueta de Valor	*Valor*	*Frecuencia*	*Porcentaje*	*Porcentaje Válido*	*Porcentaje Acumulado*
Jardín/Preescolar	1	2479	,09	,26	,26
EGB (1 a 9 año)	2	13809	,53	1,46	1,72
Primario (1 a 6/7 grado)	3	199891	7,62	21,15	22,87
Polimodal (1 a 3/4 año)	4	19455	,74	2,06	24,93
Secundario (1 a 5 ó 6/7 año)	5	452412	17,25	47,87	72,80
Superior no universitario	6	100384	3,83	10,62	83,42
Universitario	7	151699	5,78	16,05	99,47
Posgrado universitario	8	4989	,19	,53	100,00
	.	1673113	63,80	Perdidos	
Educación especial	98	2500	,10	Perdidos	
Ns/Nr	99	1698	,06	Perdidos	
	Total	2622429	100,0	100,0	

C2_P13 Nivel educativo más alto

N	*Válido*	945118
	Perdidos	1677311
Media		4,95
Desv Std		1,35
Mínimo		1,00
Máximo		8,00

Como puede observar, esta es una variable que no debe ser considerada como ordinal, porque a las categorías de la variable no se les puede atribuir un orden de menor a mayor. De hecho, el nivel EGB corresponde a otro sistema educativo y toma los primeros 9 años de escolaridad, por lo que una parte de este grupo debería asistir al secundario (los últimos 3 años). Estas son algunas de las dificultades que encontramos cuando trabajamos con bases de datos secundarias (realizadas por otros), que casi nunca se adecúan con exactitud a nuestros propósitos. No obstante, vale la pena buscar la forma de readecuarla, en la medida en que es prácticamente imposible para nosotros realizar una encuesta a familias en la región centro con más de diez mil casos.

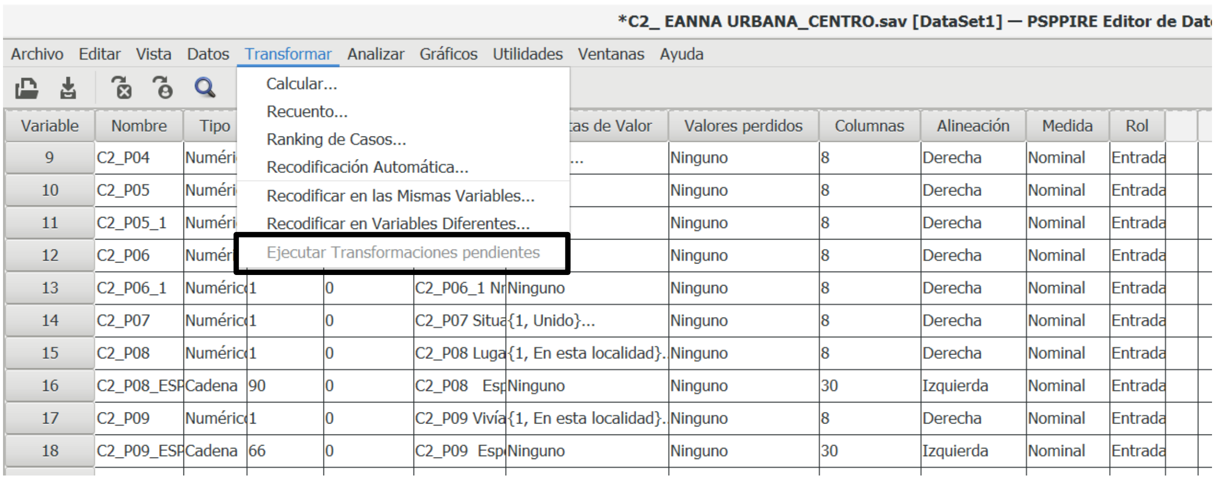
No sucede lo mismo con los que están en el viejo nivel Polimodal, que sí podrían asociarse a los que están en el secundario, dado que el polimodal correspondía a los últimos 3/4 años del secundario. Sin embargo, en tanto para este estudio nos interesa el nivel Superior universitario (Superior no universitario, Universitario y Posgrado), podremos transformar la variable en una ordinal resumiendo todos los otros niveles (Jardín, EGB, Primario, Polimodal y Secundario) en otra categoría que se llame Hasta Secundario. De este modo la variable pasará a ser ordinal, con cuatro categorías: Hasta secundario, Superior no universitario, Universitario y Posgrado (universitario).

Esto es posible en el marco de los objetivos de nuestra investigación. Es muy importante, a la hora de analizar los datos, tener presente los objetivos de nuestra investigación, para determinar si corresponde utilizar y/o reagrupar una variable de un modo u otro.

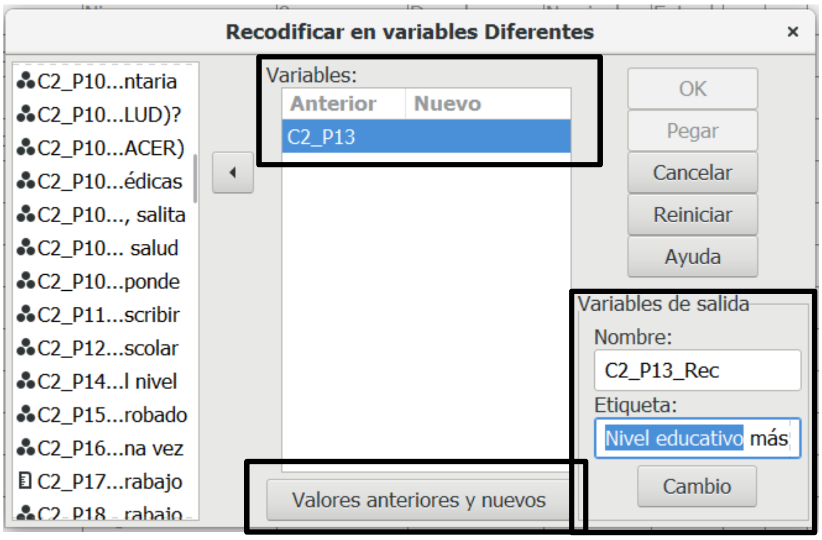
Para ello, vamos a recodificar la variable. Primero, vamos a juntar las categorías Jardín/Preescolar, EGB, Primario, Polimodal y Secundario en una categoría que se llame Hasta secundario. Luego nos quedarán como categoría individual Superior no universitario, Universitario y Posgrado universitario. A las categorías 98 y 99 no las consideraremos en la nueva variable que vamos a crear.

Como habrán notado en el Manual de la EANNA_URBANA la población considerada para estas preguntas son los mayores de 18 años inclusive en hogares con niños, niñas y adolescentes que asisten o asistieron a algún establecimiento educativo. Tener en claro el universo de análisis también será primordial para ver si se adecúa a nuestros objetivos.

Para realizar la operación de transformación vamos a recurrir al comando Transformar y dentro de él al comando Recodificar en variables diferentes.

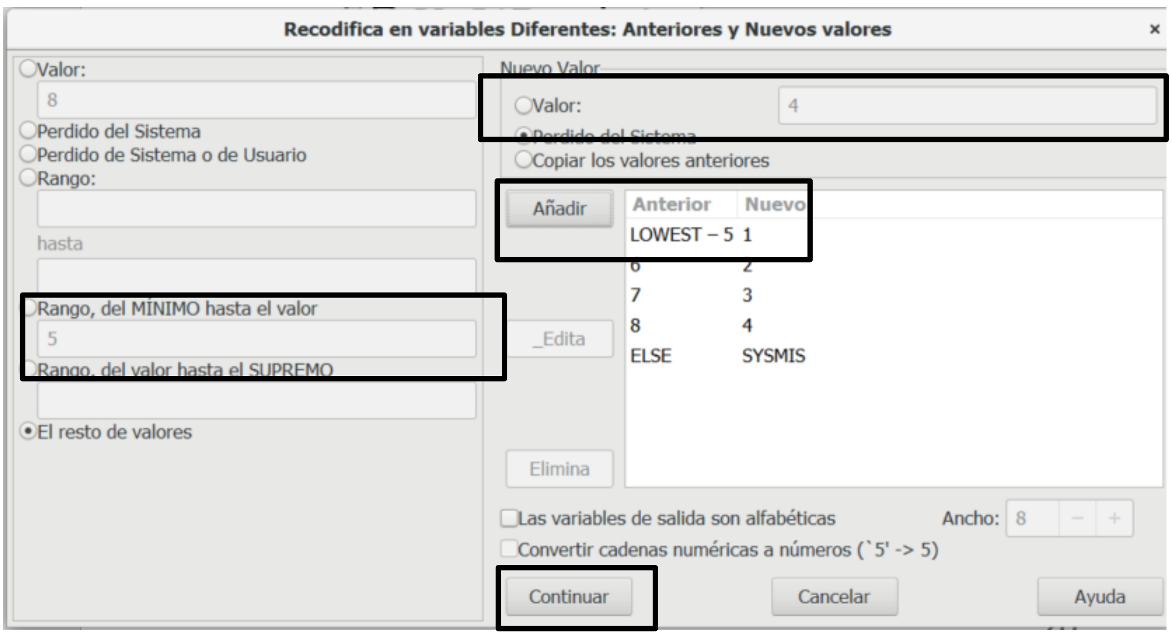


Una vez abierto el cuadro de Recodificar en variables diferentes, vamos a seleccionar la variable que queremos recodificar, C2_P13 y la vamos a renombrar con C2_P13_Rec y renombraremos la etiqueta de la variable por Nivel educativo más alto recodificada.



Luego vamos a Valores anteriores y nuevos en la parte inferior de cuadro y seleccionamos allí. Se abre un nuevo cuadro que dice Recodifica en variables Diferentes: Anteriores y Nuevos valores. En ese cuadro, buscamos a la izquierda donde dice Rango y pondremos Rango, del MÍNIMO hasta el valor: 5 y de la derecha, buscamos donde dice Valor, signaremos el Valor: 1 y damos al botón añadir.

Luego vamos a donde dice Valor: del lado de la izquierda del cuadro y tildamos el círculo de esta opción, ponemos el Valor 6 y del lado de la derecha le asignamos el Valor 2, ponemos Añadir nuevamente. A 7 le asignamos el valor 3, a 8 el valor 4 y nos vamos a la izquierda y tildamos donde dice: El resto de los valores, en la derecha tildamos Perdido del Sistema, damos a Añadir y el cuadro quedará así:



Le damos a Continuar y en el siguiente cuadro le damos al botón Cambio para que aparezca el nuevo nombre de la variable. Finalmente le damos a OK. Si se fijan ahora tendrán al final de la Vista de variables la nueva variable C2_P13_Rec creada. Ahora vamos a asignar una etiqueta a los valores de la nueva variable.

Para:

1 ➔ Hasta secundario

2 ➔ Superior no universitario

3 ➔ Universitario y

4 ➔ Posgrado.

Ahora para cerciorarnos que hemos realizado las operaciones correctamente pedimos una tabla de frecuencias de nuestra nueva variable C2_P13_Rec, que debería haber quedado así:

Nivel educativo más alto recodificada

Etiqueta de Valor	*Valor*	*Frecuencia*	*Porcentaje*	*Porcentaje Válido*	*Porcentaje Acumulado*
Hasta secundario	1,00	688046	26,24	72,80	72,80
Superior no universitario	2,00	100384	3,83	10,62	83,42
Universitario	3,00	151699	5,78	16,05	99,47
Posgrado	4,00	4989	,19	,53	100,00
	.	1677311	63,96	Perdidos	
	Total	2622429	100,0	100,0	

Ya tenemos una nueva variable, que ahora sí tiene nivel de medición ordinal. Sabemos además que el 16.55% de los encuestados tienen nivel Universitario o más, sumando las categorías de Universitario y Posgrado.

Nuestro investigador sospecha que este porcentaje ha cambiado con el paso de los años, por lo que consulta una muestra aleatoria de personas en la región Centro realizada por medio de la Encuesta Permanente de Hogares en el Tercer trimestre de 2019, con un total de casos de 4328 individuos, donde encontró que el 22,0% de los mayores de 18 años tenía acceso al nivel Universitario más el nivel Posgrado. En este caso es importante que tenga en cuenta que, para comparar las proporciones de las muestras, debe tomar el total de casos sin ponderar la base, que en esta oportunidad son 7511 casos.

Según lo visto hasta aquí se solicita: a) plantee la expresión estadística de la hipótesis nula y de la hipótesis de investigación de nuestro investigador, b) determine el valor de Z crítico para la comparación de proporciones, con una probabilidad de error tipo I con α=0.05, c) determine si es plausible rechazar la hipótesis nula y esboce una conclusión.

………

………

………

………

………

………

………

………

………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………………..

Ahora vamos a realizar una nueva estimación para una variable métrica. Para ello vamos a tomar un indicador de tiempo dedicado al empleo. Consultaremos la variable C2_P28 Horas de trabajo ocupación principal, pero primeramente volvemos a ponderar la base mediante la variable Factor de ponderación como lo hemos hecho con anterioridad:

C2_P28 Horas de trabajo ocupación principal

Etiqueta de Valor	*Valor*	*Frecuencia*	*Porcentaje*	*Porcentaje Válido*	*Porcentaje Acumulado*
	1	243	,01	,04	,04
	2	1278	,07	,21	,25
	3	859	,05	,14	,40
	4	5850	,33	,98	1,37
	5	2003	,11	,33	1,71
	6	2533	,14	,42	2,13
	7	674	,04	,11	2,24
	8	2973	,17	,50	2,74
	9	5771	,32	,96	3,70
	10	6081	,34	1,01	4,71
	84	5331	,30	,89	99,07
	86	1012	,06	,17	99,24
	90	492	,03	,08	99,32
	96	1233	,07	,21	99,52
Ns/Nr	99	1247	,07	,21	99,73
	105	301	,02	,05	99,78
	112	607	,03	,10	99,88
	120	263	,01	,04	99,93
	140	442	,02	,07	100,00
	.	1191602	66,52	Perdidos	
	Total	1791254	100,0	100,0	

C2_P28 Horas de trabajo ocupación principal

N	*Válido*	599652
	Perdidos	1191602
Media		40,78
Desv Std		17,26
Mínimo		1,00
Máximo		140,00

Cómo vemos es una variable que se comporta de un modo similar a una distribución normal, con una media de 40.78 horas y una desviación estándar de 17.26 horas. Dado que nuestro investigador también supone que ha disminuido la cantidad de horas promedio trabajadas de las personas en la Región Centro contabilizadas por semana, busca los resultados de esta variable en la segunda base que mencionamos y consulta los resultados, con los siguientes estadísticos:

Estadísticos

PP3E_TOT		
N	Válido	2523
	Perdidos	1805
Media		34,93
Desv. Desviación		17,425
Rango		112
Mínimo		0
Máximo		112

Según lo esbozado, se solicita: a) plantee la expresión estadística de la hipótesis nula y de la hipótesis de investigación, b) determine el valor de Z crítico para la comparación de proporciones con una probabilidad de error tipo I con α=0.05, c) determine si es plausible rechazar la hipótesis nula y esboce una conclusión.

……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
…………

Capítulo 8

Prueba de hipótesis y el modelo estadístico χ^2

En este capítulo vamos a ver cómo se realiza la prueba de hipótesis por medio de otro modelo estadístico: el modelo de Chi-Cuadrado. Comenzaremos con el cálculo de X^2 a partir de una tabla de contingencia y vamos a ver cómo tomar decisiones a partir de este estadístico y los tipos de errores asociados. También trabajaremos el grado de adecuación entre las variables y el coeficiente de contingencia C de Pearson y el V de Crammer, que nos permitirán evaluar la relación entre las variables.

También trabajaremos las frecuencias esperadas como eventos independientes y los grados de libertad de una tabla de contingencia. Vamos a trabajar específicamente la distribución X^2 y las recomendaciones para el uso de la prueba en cuestión. Veremos cómo se realiza la prueba de la mediana y la mediana combinada y el análisis de una muestra simple por medio de la prueba de X^2 para la bondad de ajuste de valores esperados.

Supongamos que un investigador dispone de una muestra de 100 profesores, algunos de ellos dictan clases en el último ciclo de primaria y otros en el primer año de Secundaria. A todos ellos se les hizo la siguiente pregunta: ¿Utiliza usted la evaluación como una instancia de aprendizaje? La respuesta requerida se clasificó como una variable dicotómica nominal: SI – NO. El investigador sospecha que aquellos profesores que dictan clases en los cursos de primaria tienden a usar las evaluaciones como instancias de aprendizaje, lo cual no ocurre con los profesores que dictan clases en la Secundaria.

Así, el investigador plantea la siguiente hipótesis:

H_1: Existe relación entre el nivel de escolaridad donde el profesor dicta sus clases y la utilización de la evaluación como instancia de aprendizaje.

Nótese entonces que la hipótesis de investigación es una formalización de lo que él intuye que ocurre con las evaluaciones en los dos ciclos de escolarización. La hipótesis de investigación siempre se acompaña de la hipótesis nula (H_0). Esta última toma la forma de una negación de la hipótesis de investigación; en otras palabras, es su contrario. Por lo tanto, la misma se formaliza como:

H_0: No existe relación entre el nivel de escolaridad donde el profesor dicta sus clases y la utilización de la evaluación como instancia de aprendizaje.

Si hay asociación entre las variables deberíamos esperar que las proporciones de aquellos profesores que responden afirmativamente y negativamente a la pregunta se concentren preferentemente en categorías opuestas, que para ese problema sería el nivel de escolaridad. Al contrario, si no existe asociación entre las variables deberíamos esperar que la respuesta de los profesores estuviera repartida proporcionalmente en cada una de las categorías de nivel de escolaridad. En otras palabras, si no existe asociación entre las variables, las frecuencias en las categorías SI y NO, estarían proporcionalmente repartidas entre los docentes que enseñan en primaria y en Secundaria. Al cruzar las variables nivel de escolaridad y respuestas de los profesores, es posible construir una tabla de contingencia de cuatro celdas, tal como se muestra a continuación.

	¿Utiliza la evaluación como instancia de aprendizaje?	
Nivel donde Dicta clases	SI	NO
EGB		
Secundaria		

Las celdas están allí donde se entrecruzan los niveles de las variables. Bajo el supuesto de la hipótesis nula, las frecuencias de las celdas deberían repartirse proporcionalmente de acuerdo a las frecuencias marginales. Cabe destacar que la sospecha del investigador es que los profesores del último ciclo de primaria son los que utilizan la evaluación como instancia de aprendizaje, pero esta situación no está expresada en la hipótesis planteada. Es decir, si la hipótesis debiera responder estrictamente a la sospecha del investigador, esta debería haberse planteado de modo unidireccional. Pero, por tratarse solo de una sospecha, nuestro investigador solo plantea una hipótesis bidireccional.

Siendo P_{si} la proporción de profesores que responden afirmativamente y P_{no} la proporción de profesores que responden negativamente a la pregunta del investigador, las hipótesis toman la siguiente forma estadística de acuerdo a la categoría: ciclo en el que dicta clase el profesor:

$$H_1: p_{si} \neq p_{no}: \chi^2 > 0$$
$$H_0: p_{si} = p_{no}: \chi^2 = 0$$

Cálculo del estadístico χ^2 a partir de la tabla de contingencia

Siguiendo con el ejemplo planteado, supondremos que el investigador cuenta con las respuestas dadas por los profesores, y vuelca los datos en una tabla de contingencia, colocando en cada una de las entradas las variables de interés, y arreglando las frecuencias observadas en cada una de las casillas. La frecuencia total que resulta de sumar las frecuencias en cada una de las casillas que forman las filas y las columnas se denominan marginales de fila y de columna respectivamente. De este modo, los datos se presentarían tal como lo muestra la siguiente tabla.

	¿Utiliza la evaluación como instancia		
Nivel donde dicta clases	**SI**	**NO**	**Marginal Fila**
Primaria	33	19	52
Secundaria	27	21	48
Marginal Columna	60	40	100

Esta tabla muestra que a la pregunta respondieron 52 profesores de primaria y 48 de secundaria; además se observa que 60 de estas preguntas fueron afirmativas y 40 negativas. Estos son los marginales por fila y columna de la tabla. En cada una de las celdas tenemos las proporciones de respuestas afirmativas y negativas de los profesores de Primaria y Secundaria. Estas son las frecuencias observadas, es decir, aquellas que se obtienen empíricamente del conteo de las unidades de análisis que caen en el cruce de cada una de las categorías de las variables.

Anteriormente se dijo que bajo la hipótesis nula debería esperarse una distribución proporcional de las frecuencias en cada una de las celdas. Conociendo las frecuencias marginales es posible calcular las frecuencias que cabría esperar en cada una de las celdas, si efectivamente no existiera asociación entre las variables. Estas frecuencias se llaman frecuencias esperadas porque son aquellas frecuencias teóricas que cabría esperar, de ser cierta la hipótesis nula. Su cálculo es muy sencillo, se trata de multiplicar el marginal de fila y columna correspondiente a cada celada y dividirlo por el total de casos. Veamos en detalle cómo se obtienen:

Frecuencias esperadas para:

a) Profesores de primaria que respondieron afirmativamente: 60*52/100= 31.2

b) Profesores de primaria que respondieron negativamente: 40*52/100= 20.8

c) Profesores de Secundaria que respondieron afirmativamente: 60*48/100= 28.8

d) Profesores de Secundaria que respondieron negativamente: 40*48/100= 19.2

Entonces, las frecuencias observadas son las que efectivamente se cuentan como cantidad de casos en cada una de las celdas que surgen del cruce entre las categorías de las variables. Las frecuencias esperadas son teóricas y se obtienen de la operación de multiplicar las frecuencias marginales correspondientes a cada celda y dividirla por el total de casos. Ahora, se puede arreglar las frecuencias observadas y esperadas en la misma tabla y compararlas. Cabe recordar en este punto que las frecuencias esperadas son las que se hubieran obtenido de ser cierta la hipótesis nula, y puesto que ésta representa la situación en donde no hay asociación entre las variables, la proporción de las frecuencias esperadas es la misma que la proporción en los marginales.

Así, por ejemplo, la proporción de profesores de primaria sobre profesores de Secundaria es de 52/48=1,083, la misma proporción entre los profesores que respondieron afirmativamente es de 31,2/28,8=1,083, y la proporción entre los profesores que respondieron negativamente es de 20,8/19,2=1,083 (el mismo cálculo puede efectuarse con los marginales de columnas). Entonces las frecuencias esperadas representan una distribución proporcional de casos en cada una de las celdas, dados los marginales de tabla.

Comparación de frecuencias observadas y esperadas

	¿Utiliza la evaluación como instancia de aprendizaje?		
Nivel donde dicta clases	SI	NO	Marginal Fila
EGB	33 (31.2)	19 (20.8)	52
Secundaria	27 (28.8)	21 (19.2)	48
Marginal Columna	60	40	100

Nota. Las frecuencias esperadas se muestran entre paréntesis

Según lo que se ha explicado, si las dos variables estuvieran relacionadas entre sí, la distribución de frecuencias no debería ser proporcional, dado que una variable ejercería alguna "acción" sobre la otra. En este ejemplo, se supone que los profesores que enseñan en distintos niveles, toman a la evaluación de manera diferente. Por lo tanto, si las frecuencias observadas están próximas o son iguales a las frecuencias esperadas, estaríamos en la situación planteada por la hipótesis nula. Al contrario, si las frecuencias observadas se apartan notablemente de las esperadas será necesario rechazar la hipótesis nula.

El estadístico χ^2 toma en consideración la distancia entre las frecuencias observadas y las esperadas, y sirve para estimar la probabilidad de que tales diferencias sean debidas al azar. Veamos primero la forma de cálculo de este estadístico y luego retomaremos este punto.

La fórmula de cálculo de χ^2 puede representarse de la siguiente manera:

$$\chi^2 = \frac{\sum(\text{fo} - \text{fe})^2}{\text{fe}}$$

El estadístico χ^2 es la sumatoria de las desviaciones cuadráticas entre las frecuencias observadas y esperadas, sobre las frecuencias esperadas. Dado el caso en que las frecuencias observadas sean las mismas que las frecuencias esperadas, el numerador de la fracción es igual a cero, por lo tanto, el estadístico también es igual a cero. Esta es la situación exacta que plantea la hipótesis nula. En otro caso tendremos que las frecuencias observadas serán distintas de las frecuencias esperadas y el estadístico se apartará progresivamente de cero. Por ende, cuanto mayor sea la diferencia entre las frecuencias observadas y las esperadas, más alejados estaremos de la situación planteada por la hipótesis nula.

En el ejemplo el estadístico χ^2 se calcula de la siguiente manera:

$$\boxed{\chi^2 = (33\text{-}31.2)^2/31.2 + (19\text{-}20.8)^2/20.8 + (27\text{-}28.8)^2/28.8 + (21\text{-}19.2)^2/19.2 = \mathbf{0.538}}$$

El valor del estadístico χ^2 debería servirnos para decidir si rechazamos o no la hipótesis nula. Para ello debemos tomar en cuenta la distribución del estadístico, y el valor de probabilidad asignado al Error Tipo I que vimos en la unidad anterior.

Reglas de decisión basadas en χ^2 y errores de decisión

Vimos que, cuando se plantea una hipótesis de investigación, tenemos también una hipótesis nula. En el ejemplo que estamos desarrollando, la hipótesis nula es analizada bajo el modelo del estadístico χ^2 para decidir si es posible rechazarla. Nótese que anteriormente planteamos que la hipótesis nula es una negación de la hipótesis de investigación, de modo que, si es posible rechazar la hipótesis nula, la hipótesis de investigación es la alternativa válida. En caso que no sea posible rechazar la hipótesis nula, debemos descartar la hipótesis de investigación. En otras palabras, al rechazar la hipótesis nula se asume que efectivamente existe asociación entre las variables y que tal asociación no es producto del azar.

Sabemos que al tomar la decisión de rechazar la hipótesis nula se puede cometer un error, rechazarla cuando esta es verdadera. Si procedemos de tal manera habremos cometido un error tipo I. En cambio, si aceptamos una hipótesis nula cuando es falsa, estamos cometiendo un error tipo II. No es posible disminuir al mínimo ambos tipos de errores al mismo tiempo, pero en un planteo de hipótesis es factible establecer un equilibrio entre ambos.

Vimos que una manera de hacerlo es establecer una condición donde la probabilidad de cometer el error tipo I sea la menor posible. Convencionalmente, en una investigación la probabilidad de cometer el error tipo I se establece a priori y se define como α (alfa), este valor se determina por convención en $\alpha=0.05$. Mediante el valor de α podemos establecer un valor crítico para χ^2, con el cual compararemos el valor χ^2 efectivamente observado. La regla de decisión quedará expresada de la siguiente forma:

Sea $\alpha=0.05$, se rechaza la hipótesis nula siempre y cuando:

$$\boxed{\chi^2 \text{ observado } \geq \chi^2 \text{ crítico}}$$

Es decir, si el valor de χ^2 observado es igual o mayor al valor de χ^2 crítico, rechazamos la hipótesis nula que establece que las variables no están asociadas. Si, al contrario, el valor de χ^2 observado no supera el valor de χ^2 crítico, no hay evidencia estadística suficiente para rechazar la hipótesis nula.

Ya se dijo que el valor de χ^2 crítico se obtiene al establecer el valor de α. Ahora bien, hay que tener en cuenta que la distribución de este estadístico conforma una familia de curvas, que depende del tamaño de la tabla de contingencia. Es decir, el valor de χ^2 crítico varía según los grados de libertad de la tabla. Aunque esto parece engorroso, en realidad es un concepto muy sencillo, digamos que para conocer cuántos grados de libertad tiene una tabla debemos resolver la siguiente fórmula:

$$\boxed{gl= (f-1)*(c-1)}$$

Entonces, los grados de libertad de la tabla se definen como la cantidad de filas menos 1, multiplicado por la cantidad de columnas menos 1. En la tabla de nuestro ejemplo tenemos 2 filas, una que corresponde a profesores de primaria y la otra corresponde a profesores de secundaria. También tenemos dos columnas, una que corresponde a todos los que respondieron SI y la otra la que corresponde a todos los que respondieron NO. Al aplicar la fórmula tenderemos que los grados de libertad de la tabla resultan, gl= (2-1)*(2-1)= 1.

El valor de χ^2 crítico que estamos buscando para resolver nuestro ejemplo, es aquel que corresponde a una tabla de contingencia con 1 grado de libertad, para un valor de $\alpha=0.05$.

Existen tablas donde se ha tabulado el valor de χ^2 crítico para diferentes grados de libertad, y para distintos valores de α. En nuestro caso el valor buscado de χ^2 crítico es de 3,841. Al aplicar la regla de decisión planteada tenemos que χ^2 observado < χ^2 crítico esto es, 0,538 < 3,841 y por lo tanto no es posible rechazar la hipótesis nula.

De acuerdo a lo observado en la tabla 9, tenemos que no existe asociación entre el nivel en que enseña el profesor y el hecho de utilizar la evaluación como una instancia de aprendizaje. Ahora el investigador sabe que no existe asociación entre esas variables, y que su sospecha posiblemente se debió a un sesgo en sus observaciones.

En el siguiente cuadro se ofrecen valores críticos de χ^2 que corresponde a valores de α=0,05 y α=0,01 y diferentes dimensiones de la tabla de contingencia.

Gl	ET I =1%	ET I= 5%
1	6,63	3,48
2	9,21	5,99
3	11,34	7,81
4	13,27	9,48

El grado de asociación entre las variables

En el ejemplo que hemos dado, nos encontramos con que no ha sido rechazada la hipótesis nula y por tanto se asume que no hay asociación entre las variables. Pero en el caso en que se haya rechazado la hipótesis nula, asumiremos que las variables están relacionadas y será necesario determinar el grado de asociación entre ellas. Basados en la distribución χ^2 existen varias alternativas para calcular la asociación entre las variables. En los apartados siguientes veremos dos estadísticos que determinan esa asociación y que son los más comúnmente usados.

Coeficiente de contingencia C de Pearson

Cuando se encuentra correlación entre las variables e interesa conocer el grado de asociación entre ellas, se emplea el coeficiente de contingencia C de Pearson, el cual se define como: la raíz cuadrada del cociente entre el valor de χ^2 y el valor de χ^2 más el número de casos N.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

El coeficiente de contingencia C de Pearson, (que no debe ser confundido con el coeficiente de correlación r de Pearson), estima la asociación entre variables en una tabla de contingencia. Cuando estas son independientes, el coeficiente es igual a cero, y se hace mayor a medida que aumenta la dependencia de las variables. El valor máximo del coeficiente depende del número de filas y columnas en la tabla de contingencia, pero es siempre menor que la unidad.

El hecho que este coeficiente no tenga un valor máximo establecido para cualquier tabla, dificulta su interpretación, pero al ser constante permite la comparación entre resultados obtenidos con tablas de igual tamaño.

Coeficiente V de Cramer

Otro coeficiente de asociación que puede utilizarse para estimar el grado de asociación entre los atributos es el coeficiente V de Cramer, cuya fórmula es la siguiente:

$$V = \sqrt{\frac{\chi^2}{\min(h-1; c-1) * N}}$$

En la fórmula la expresión min. (h–1; c–1) alude al número de hileras o columnas menos uno. Es decir, en el numerador entrará el menor valor de los dos. Entonces V se define como: la raíz cuadrada del cociente entre el valor χ^2 observado y el número de columnas o filas menos uno (dependiendo de cuál sea el menor), multiplicado por el número de casos. Este coeficiente puede aplicarse a tablas de cualquier tamaño y alcanza un máximo de uno.

Ejemplo de aplicación

Vamos a presentar un ejercicio de aplicación de la prueba χ^2 para el caso de una tabla de contingencia mayor a 2x2. Un investigador realiza una encuesta a un grupo de docentes de primaria, que enseñan en escuelas privadas y públicas, sobre la importancia de trabajar en equipos interdisciplinarios cuando en el aula existen niños con problemas de aprendizaje. El investigador desea comprobar si existe asociación ente los atributos, tipo de escuela y valoración del trabajo interdisciplinario. Plantea la hipótesis de que los docentes darían distinta valoración al trabajo en equipo, de acuerdo a la gestión de la escuela en que trabajan. Para comprobar la existencia de asociación entre las variables utiliza el estadístico χ^2.

De acuerdo a las hipótesis esbozadas se tiene que el planteo estadístico correspondiente es:

$$\boxed{\begin{array}{c} H_1: \chi^2>0 \\ H_0: \chi^2=0 \end{array}}$$

La tabla recoge la distribución de frecuencias observadas que resulta de la combinación de las variables gestión de la escuela y valoración del trabajo interdisciplinario.

Frecuencias Observadas

	Valoración del trabajo interdisciplinario		
Gestión de la escuela	**Es muy importante**	**Es importante pero no necesario**	**No es tan importante**
Privadas	102	106	21
Públicas	195	92	89

Corresponde ahora construir la tabla que hubiéramos encontrado de ser verdadera la hipótesis nula. Dicha tabla se construye con las frecuencias esperadas y refleja la situación en que el estadístico χ^2 es igual a 0.

Frecuencias Esperadas

	Valoración del trabajo interdisciplinario		
Gestión de la escuela	Es muy importante	Es importante pero no necesario	No es tan importante
Privadas	112,42	74,94	41,63
Públicas	184,58	123,05	68,36

La diferencia entre la distribución de frecuencias observadas y esperadas, especificará la magnitud del estadístico χ^2. Para ello debemos calcular la sumatoria de esas diferencias, tal como se muestra a continuación:

1	(102-
2	(195-
3	(106-74.94)2/74.94=12.87
4	(92-123.05)2/123.05=7.83
5	(21-41.63)2/41.63=10.22
6	(89-68.36)2/68.36=6.23
Σ	38.703

Hemos obtenido el valor de χ^2 observado y corresponde compararlo con el valor χ^2 crítico, para tomar la decisión de rechazar o no la hipótesis nula. El valor de χ^2 crítico para un error tipo I igual a α=0.05 y una tabla con 2 grados de libertad es de 5,99. Entonces tenemos que el valor de χ^2 observado es mayor que el valor de χ^2 crítico; de acuerdo a la regla de decisión corresponde rechazar la hipótesis nula.

Aceptando que los atributos de las variables están relacionados, corresponde determinar el grado de asociación entre las variables. Para ello aplicamos el coeficiente V de Cramer.

$$V = \sqrt{\frac{38{,}703}{605}} = 0{,}252$$

Encontramos que las variables están asociadas, pero la magnitud de dicha asociación es moderada a débil. ¿Cuál es la manera de interpretar esa asociación? Si repasamos la tabla vemos que la mayor diferencia se produce en el nivel de la variable valoración del trabajo interdisciplinario para escuelas de gestión privadas, en menor medida se registran discrepancias en las escuelas de gestión públicas.

Por lo tanto, es factible concluir que la asociación entre las variables está dada por la inclinación de los docentes de escuelas privadas a opinar que el trabajo interdisciplinario es importante pero no necesario y los docentes de escuelas públicas a opinar que no es tan importante. Aunque la asociación resulta de moderada a débil, tal inclinación resulta atendible en virtud de que el hemos rechazado la hipótesis de no asociación entre las variables.

Las frecuencias esperadas como eventos independientes

Hasta aquí hemos señalado que las frecuencias esperadas son las que ocurrirían de ser cierta la hipótesis nula y, además, es la distribución de frecuencias que hace que el estadístico χ^2 sea igual a cero. Esto, equivale a decir que cualquier evento que conste en las columnas será independiente de cualquier otro evento que conste en las filas.

Volvamos por un momento a la tabla precedente. Si no hay asociación esto quiere decir que los profesores que dictan clase en escuelas privadas pueden valorar de cualquier manera el trabajo interdisciplinario. Un razonamiento similar puede hacerse si se parte desde las filas. De modo que la frecuencia esperada, resulta de tomar como independientes las filas y las columnas y obteniendo de ello una probabilidad. La regla de la multiplicación de probabilidades de eventos independientes sirve para entender por qué las frecuencias esperadas se calculan de esta manera.

Por ejemplo, si quisiéramos conocer la probabilidad de obtener un resultado particular cuando se lanzan al aire dos veces la misma moneda, se aplica la regla multiplicativa de las probabilidades independientes, dado que suponemos que un evento (el primer lanzamiento de la moneda), no influye en el segundo evento (el segundo lanzamiento de la moneda). Llamamos A al primer evento y B al segundo; ambos son independientes. Entonces, la probabilidad de un resultado AB se expresa como p (AB)= p(A) x p(B). Para el caso de la moneda, en la cual podemos obtener solo dos resultados posibles, la probabilidad de A es igual a 1/2= 0.5 (la probabilidad de B será la misma que A). Según lo expresado en la fórmula, la probabilidad de dos veces cara será p2 caras = (0.5) x (0.5)= 0.25.

Si volvemos por un momento a nuestro ejemplo y considerando cierta la hipótesis nula: el tipo de gestión de la escuela donde dicta clases el profesor no se relaciona con la valoración del trabajo interdisciplinario, se puede buscar la probabilidad de los eventos combinados, siempre considerando que estos eventos son independientes.

Se puede calcular por un lado, la probabilidad de obtener de la muestra de profesores, un profesor que pertenezca al tipo de gestión privado. Para ello consideramos a todos los profesores en ese nivel de la variable que suman 229 profesores (este es el marginal de fila) y lo dividimos por el total 229/605=0,378.

Asimismo, podemos calcular la probabilidad de obtener de todos los profesores, uno que considere el trabajo interdisciplinario como muy importante. Para ello debemos considerar a todos los profesores de la primera columna que suman 297 (este es el marginal de columna) y dividimos por el total 297/605=0,49.

La probabilidad de ambos eventos combinados resulta de la multiplicación de los eventos independientes, esto es: 0,378x0.49= 0,18522.

El valor obtenido es la frecuencia relativa o valor de probabilidad, pero puesto que en la tabla se tienen las frecuencias absolutas se debe transformar la frecuencia obtenida. Por lo que ya sabemos, para pasar de una frecuencia relativa a una absoluta, hay que multiplicarla por N, entonces 0,185x605= 112,0581, es la frecuencia esperada buscada para la intersección de las variables Profesor de escuela privada que valora el trabajo interdisciplinario como muy importante (la diferencia en decimales entre el valor obtenido y el de tabla se debe al redondeo).

El proceso de multiplicar marginal de fila por marginal de columna y dividirlo por el total de casos es una simplificación de lo anteriormente dicho.

Resumiendo, las frecuencias esperadas representan la probabilidad de un evento combinado, cuando ambos eventos son independientes, y la independencia de los eventos es el postulado de la hipótesis nula. Asimismo, el estadístico χ^2 es una comparación entre las frecuencias observadas y las esperadas, bajo la hipótesis de la independencia de los eventos (para un desarrollo más detallado ver Lorenzo y Giovine, 2019[6]).

---

6 Lorenzo, J. y Giovine, M. (2019) *Probabilidad para las ciencias de la educación.* Córdoba: ANSENUZA. Link: https://ansenuza.unc.edu.ar/comunidades/bitstream/handle/11086.1/1341/Probabilidad%20para%20educaci%C3%B3n-Lorenzo%20Giovine.pdf?sequence=1

Los grados de libertad en tablas de contingencia

Sabemos que la frecuencia marginal de una fila se construye a partir de la suma de las frecuencias observadas en cada una de las celdas. Así, calculamos la frecuencia marginal de la fila correspondiente a profesores de EGB, sumando las frecuencias de celdas que son 33 y 19, por tanto esta frecuencia marginal es igual a 52 (ver tabla). Sabiendo ahora que el total de esa suma es 52, ¿cuántos valores pueden variar libremente si se necesitan dos sumandos para ese resultado? Para exponerlo gráficamente, estaríamos en esta situación:

__ + __ = 52

Si se completa el primer espacio en blanco, el segundo espacio queda determinado por defecto, así si en el primer espacio se coloca el número 1, en el segundo espacio se debe colocar el número 51 para que la suma sea igual a 52. Si en vez de colocar el número 1 se coloca el número 2, en el segundo casillero se debe colocar el número 50 para que la suma siga siendo igual a 52. Es decir, cuando tengo un resultado dado, que se obtiene por la suma de dos miembros, solo uno puede variar libremente.

En una tabla de contingencia de 2x2 cada uno de los marginales es el resultado de una suma de dos números, por lo tanto, solo un número en una de las cuatro celdas puede variar libremente.

Sigamos el procedimiento esbozado más arriba. La menor frecuencia marginal de la tabla es igual a 40 (que es la suma de las celdas de los profesores de EGB y Secundaria que respondieron negativamente). Sabiendo que ningún miembro de la suma puede ser mayor que el total, diremos que el límite para repartir libremente los valores dentro de la tabla es de 40. Construyamos entonces una primera tabla asignando el valor 1 a la casilla correspondiente a profesores de primaria que responden negativamente:

	¿Utiliza la evaluación como instancia de aprendizaje?		
Dicta clases	SI	NO	M. Fila
Primaria	51	1	52
Secundaria	9	39	48
M. Columna	60	40	100

Si determinamos ese valor de celda arbitrariamente, el resto de los valores queda establecido, ya que no hay otros valores posibles que den como resultado los marginales de tabla. Sigamos con ese procedimiento y asignemos el valor 2 a esa casilla. Tal como muestra la tabla, al asignar el número 2 los únicos valores de las restantes son los siguientes.

	¿Utiliza la evaluación como instancia de aprendizaje?		
Nivel donde Dicta clases	SI	NO	M. Filas
Primaria	50	2	52
Secundaria	10	38	48
M. Columnas	60	40	100

Como ya vimos, una tabla 2x2 esta tendrá 1 solo grado de libertad. Ahora bien, resulta que, si la tabla tiene más filas y más columnas, tendrá más grados de libertad.

Dados los marginales de la tabla, ¿cuántas tablas de contingencia diferentes pueden construirse? De hecho, sabemos que se pueden construir al menos 40 tablas distintas. Determinar la cantidad de tablas es un cálculo complicado, pero si sabemos que el menor valor de frecuencia marginal de la tabla es 40, y si además sabemos que ese valor se obtiene de la suma de dos celdas, y que determinado el valor de una de esas celdas, los valores de las restantes celdas quedan determinados por defecto, se concluye que se pueden construir varias tablas distintas con los marginales dados.

Ahora bien, de construir efectivamente esas tablas, sabríamos que algunas de ellas representarían una distribución proporcional de frecuencias. Es decir, algunas tablas estarían en consonancia con lo planteado por la hipótesis nula, mientras que otras presentarían valores en las celdas que harían insostenible esa hipótesis. Entonces: ¿qué probabilidad existe de haber encontrado una distribución de frecuencias como la observada en la tabla, bajo el supuesto que la hipótesis nula es verdadera? En tal caso, revisaríamos la probabilidad dada por la prueba χ^2 la cual establece que para una hipótesis bidireccional como la planteada, la probabilidad de ocurrencia de una tabla con una distribución de frecuencia observada, ocurriría con una probabilidad mayor al 50% (concretamente el 54,15%). Tal probabilidad de ocurrencia es alta y por tanto no se puede restar credibilidad a la hipótesis nula.

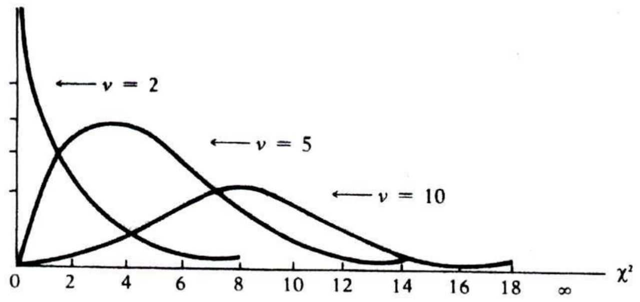
Distribución χ^2

Como ya se dijo antes, la distribución χ^2 es una distribución conocida conformada por una familia de curvas; estas diferentes curvas se basan en los grados de libertad, que como vimos, no es otra cosa que la cantidad de celdas que están libres de variar, cuando se han establecido los marginales de la tabla.

Como se aprecia en la siguiente figura, la distribución es asimétrica y se aproxima a una distribución normal cuando los grados de libertad son mayores que 10. En todos los casos la distribución muestra la probabilidad de ocurrencia de un valor χ^2, bajo el supuesto de que la hipótesis nula es verdadera.

Se observa entonces que, si el valor está próximo a 0, la probabilidad de que la hipótesis nula sea verdadera es alta y, al contrario, a medida que aumenta el valor de χ^2 esa probabilidad se hace cada vez menor.

El valor de χ^2 crítico se sitúa sobre el eje de las abscisas y resulta el punto de comparación con el valor de χ^2 observado. En este caso es posible utilizar la regla de decisión esbozada anteriormente o bien, determinar el valor exacto de ocurrencia del estadístico χ^2 obtenido.



Recomendaciones para el uso de la prueba χ^2

La prueba requiere que las frecuencias esperadas en cada celda no sean demasiado pequeñas, ya que, si este es el caso, la prueba pierde potencia. Las recomendaciones más citadas para la correcta aplicación son:

Tablas de 2x2; gl=1: cuando *n* es igual o menor que 20, se recomienda el uso de la prueba exacta de Fisher. Cuando *n* es mayor que 20 y menor que 40, la prueba χ^2 puede utilizarse si las frecuencias esperadas alcanzan un valor no menor que 5. Si existen frecuencias esperadas menores que 5, conviene utilizar la prueba exacta de Fisher.

Tablas de *n* x k; gl>1: puede utilizarse la prueba χ^2 si menos del 20% de las celdas registran valores menores que 5 en las frecuencias esperadas, y si no hay celdas con frecuencias esperadas menor que 1. Si estos requisitos no son cubiertos por los datos en la forma en que originalmente fueron recolectados, se recomienda combinar las categorías adyacentes para incrementar las frecuencias esperadas en las celdas. Esto ajusta los valores observados de χ^2 a la distribución teórica.

La prueba de la mediana

La distribución χ^2 ofrece la posibilidad de conocer si existen diferencias entre grupos utilizando la mediana de dos distribuciones de datos. Si bien existen diferentes pruebas estadísticas que sirven a este fin, un simple arreglo de una tabla de 2x2 permite utilizar la prueba χ^2 como una alternativa. Supongamos entonces que se tienen dos muestras independientes; entonces, se puede someter a prueba la hipótesis de que ambas provienen de una misma población, o bien que ambas provienen de poblaciones diferentes.

Si se trata de demostrar que existen diferencias entre los grupos, la hipótesis de investigación establecerá que las medianas de ambas muestras son representativas de poblaciones diferentes. Por ende, la hipótesis nula establecerá que las medianas de las muestras serán las mismas (o estarán muy próximas entre sí), dado que las muestras provienen de la misma población. La hipótesis de investigación y la hipótesis nula se formalizan de la siguiente manera, donde los subíndices 1 y 2 representan las muestras respectivamente:

$$H_1: Mdn_1 \neq Mdn_2$$
$$H_0: Mdn_1 = Mdn_2$$

En este caso, la hipótesis de investigación es bidireccional, pero si el investigador sospecha que la diferencia observada se manifestará en alguna dirección predecible, puede formalizarse de esta manera:

$$H_1: Mdn_1 > Mdn_2$$

Aquí el investigador sostiene que la mediana del grupo 1 resultara significativamente mayor que la mediana del grupo 2.

La mediana combinada

La prueba de la mediana consiste en calcular la mediana combinada de las dos muestras (Mdn1,2); esto se logra formando un solo grupo con todas las observaciones y calculando luego la mediana de ese grupo. Esa mediana se utiliza como punto de corte para separar los casos que caen por encima y por debajo de la mediana combinada, según prevengan del grupo 1 o del grupo 2. Este arreglo en una tabla 2x2 se representa de la siguiente manera:

	Grupo 1	**Grupo 2**
Valores por encima de la Mdn1,2.	A	B
Valores por debajo de la Mdn1,2.	C	D

La celda A corresponde a los casos que están por encima de la mediana combinada y que pertenecen al grupo 1, la celda B corresponde a los casos que están por encima de la mediana combinada y que pertenecen al grupo 2. Lo mismo para las celdas C y D, que son aquellos casos que están por debajo de la mediana, según pertenezcan al grupo 1 y 2.

Si se cumple la situación anticipada en la hipótesis nula, la mediana de los grupos será la misma o estarán próximas entre sí. Por lo tanto, los valores por encima y por debajo de la mediana estarán proporcionalmente repartidos en las celdas A, B, C y D.

Ahora bien, si la mediana de uno de los grupos es mayor que la mediana del otro, aquellos casos que muestren valores por encima de la mediana combinada tenderán a agruparse en la celda A o B. Bajo esta situación la distribución no será proporcional. En otras palabras, si hay diferencia entre los grupos se acumularán más casos en la celda correspondiente a la categoría valores por encima de la mediana combinada, y viceversa.

Así se apreciará la concentración de casos opuestos por el vértice, que es propia cuando los atributos están asociados. Bajo tales circunstancias, el estadístico χ^2 será significativo. Por lo tanto, si dos muestras difieren en cuanto a su mediana, el procedimiento descripto permite utilizar una tabla de 2x2 y calcular un valor de χ^2 que, si resulta significativo, se utiliza para rechazar la hipótesis nula que postula que las medianas son iguales por haber sido extraídas las muestras de la misma población.

Veamos ahora un ejemplo sencillo de la prueba de la mediana: en una escuela se toma un examen final de ciencias naturales a todos los alumnos de cuarto grado. Los maestros observan que en general, las notas de los alumnos del turno mañana son más altas que las del turno tarde y deciden averiguar si esta tendencia es producto del azar o existe alguna influencia que puede explicar las diferencias.

Puesto que el examen no ha sido estandarizado, no se tiene certeza de que el promedio sea el estadístico adecuado para comparar a los alumnos, por ello deciden utilizar la prueba de la mediana. Los alumnos de cuarto grado en la escuela son 78 en total, y de acuerdo al arreglo de la tabla presentada anteriormente, se definió a los alumnos del turno mañana como Grupo 1, y a los alumnos del turno tarde como Grupo 2. La tabla quedó conformada como se muestra a continuación:

	Grupo 1	**Grupo 2**	**Total**
Valores por encima de la Mdn1,2.	29	9	38
Valores por debajo de la Mdn1,2	4	36	40
Total	33	45	78

Como se observa a simple vista, los valores de las celdas están opuestos por el vértice, lo que denota en principio que las medianas de los grupos no son las mismas. En otras palabras, existe una mayoría de casos en el grupo 1 que están por encima de la mediana combinada, y existe una mayoría de casos del grupo 2 por debajo de la mediana combinada. Con un arreglo de celdas como este, una alternativa para el cálculo del estadístico χ^2, es utilizando la siguiente fórmula:

A	B	A+B
C	D	C+D
A+C	B+D	N

$$\chi 2 = \frac{N * (A * D - B * C)^2}{(A + B) * (C + D) * (A + C) * (B + D)}$$

Reemplazando los valores en la fórmula se tiene que:

$$\chi 2 = \frac{78 * (29 * 36 - 9 * 4)^2}{(38) * (40) * (33) * (45)} = 35.111$$

Consultando los valores tabulados de χ^2 se tiene que el valor observado tiene una probabilidad $p<0.05$, lo cual nos resultaría suficiente para desestimar la hipótesis nula de que ambas muestras provienen de la misma población.

Esto último, en el contexto del ejemplo que nos ocupa, se interpreta de la siguiente manera: si no existiera diferencia entre los alumnos de cuarto grado de ambos turnos, la proporción de casos por encima y por debajo de la media combinada debería haber sido proporcional, y por ende el valor del estadístico χ^2 debería haber estado próximo a 0.

Dado que las distribuciones de frecuencias al interior de la tabla no son proporcionales, el estadístico χ^2 se aparta notablemente de cero, superando su valor crítico con una probabilidad de ocurrencia, bajo la hipótesis nula, menor al 5%.

Por ende, es factible afirmar que los alumnos de cuarto grado del turno mañana muestran un rendimiento significativamente diferente a los alumnos de la tarde.

Análisis de una muestra simple: prueba χ^2 para la bondad de ajuste

La prueba χ^2 puede usarse para determinar si una distribución de frecuencias dadas se ajusta a los valores esperados de una distribución. En otras palabras, las diferencias entre las frecuencias observadas y las esperadas permiten cuantificar el grado en que una distribución empírica se aparta de una distribución teórica conocida.

En muchos casos, la utilidad de la prueba χ^2 estriba en que permite verificar si el número de unidades de análisis que se distribuyen en las categorías de una variable, es verdaderamente diferente del esperado por azar. Es decir, es posible probar que existe una diferencia estadísticamente significativa entre el número observado de unidades de análisis que caen en cada categoría, y el número esperado que debería observarse de ser cierta la hipótesis nula, que en este caso supone que las unidades de análisis se distribuyen aleatoriamente.

Al utilizar el estadístico χ^2 como una prueba de bondad de ajuste, las frecuencias esperadas necesarias para su cálculo se deducen del planteo de la hipótesis nula. Entonces, cuando se analiza una sola muestra distribuida en k categorías, las frecuencias esperadas se definen como:

$$\boxed{\text{fe} = \text{N}/k.}$$

Esta prueba tiene ciertas restricciones en su uso, que dependen del valor de las frecuencias esperadas, así cuando los gl=1 la frecuencia esperada debe ser igual o mayor que 5. Cuando gl>1, se debe comprobar que menos del 20% de las frecuencias esperadas sean menor que 5, y ninguna frecuencia esperada debe ser igual que 1.

Veamos un ejemplo sencillo: en una escuela se realizan evaluaciones periódicas de materiales didácticos con una serie de cinco listas de control de calidad. En un período de tiempo dado cada profesor debe tomar una de esas cinco listas al azar y realizar el control. Al final del mes se observó la cantidad de veces que se habían utilizado cada una de las listas, esperando una proporción cercana al azar en la cantidad de veces que se utilizaron. Teniendo en cuenta que durante ese mes se realizaron 45 controles de calidad, la frecuencia esperada para el uso de las cinco listas en 45 pruebas resulta en 45/5= 9. El número de veces que se utilizó cada lista se muestra en la tabla que sigue:

Listas	**f observada**	**f esperada**	**Residual (fo-fe)**
1	7	9	-2.0
2	16	9	7.0
3	6	9	-3.0
4	8	9	-1.0
5	8	9	-1.0
Total	45	45	

Teniendo en cuenta que la lista 2 se utilizó un número inesperadamente alto de veces, se sospechó que ella era la preferida por los profesores para realizar los controles sobre el material didáctico. Por ello se utilizó una prueba χ^2 para determinar si las frecuencias observadas podías ser representativas de una distribución por azar. El resultado se muestra en la siguiente tabla:

Variable Lista	
χ^2	7.111
gl	4
Sig (p)	0.130

El valor de significación crítico para χ^2 es de p=0,05. Puesto que el valor de significación observado supera al crítico, se infiere que el inusual aumento en el uso de la lista 2, está dentro de lo que corresponde esperar por azar. Nótese que en este caso no se realiza una comparación sobre valores observados y esperados de χ^2 puesto que la probabilidad de ocurrencia del estadístico es suficiente para determinar el ajuste a una distribución conocida.

Guía de Actividades N° 8

Distribución Chi cuadrado y Prueba de hipótesis

Objetivos de la actividad:

a) realizar planteos estadísticos y la prueba de hipótesis correspondiente, a partir del planteo de una hipótesis de investigación y una hipótesis nula

b) utilizar los conceptos básicos de la distribución χ 2 en el marco de la prueba de hipótesis.

Actividad 1: Un grupo de pedagogos sostiene que el tiempo de lectura que comparten los padres con los hijos, estimula el aprendizaje de esta habilidad. A partir de ello, se realizó una investigación donde se recabaron datos sobre el tiempo compartido de lectura y la capacidad lectora de los niños. Los datos recogidos mediante encuestas se clasificaron como se muestra en la siguiente tabla:

		Muy Bueno	**Bueno**	**Regular**
Tiempo de lectura compartida	**Mucho**	35	30	15
	Poco	20	20	35
	Muy Poco	5	10	30

Con los datos ofrecidos en tabla, se pide:

a) Plantee la hipótesis nula y alternativa.

b) Calcule $\chi_0^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$

c) Determine si existe asociación entre las variables a un nivel de significación del 1%.

Valores críticos de χ^2		
	α = 0.05	**α = 0.01**
1 gl	3.84	6.64
2 gl	5.99	9.21
3 gl	7.82	11.34
4 gl	9.49	13.28

d) Calcule el grado de asociación entre las mismas.

e) Esboce una conclusión.

Actividad 2: Se ha comprobado que los alumnos frente al fracaso de alguna instancia de evaluación, presentan tres tipos de comportamiento diferente: a) pérdida de interés por la asignatura, b) indiferencia, c) desánimo. Un docente desea comprobar si estas actitudes se relacionan con el género de los alumnos. Para ello toma una muestra de 110 alumnos que han desaprobado matemática. El resultado se expresa en la siguiente tabla:

	Tipo de Comportamiento		
	Pérdida de	**Indiferencia**	**Desánimo**
Varones	12	27	9
Mujeres	38	14	10

Con los datos ofrecidos en tabla, se pide:

a) Realice el planteo estadístico para determinar si existe asociación entre las variables, con un nivel de significación del 5% (α =0,05).

b) Calcule el grado de asociación entre las mismas.

c) Esboce una conclusión.

Actividad 3: en la Escuela de Ciencias de la Educación de la Universidad de Valverde se ha notado que en los promedios de los alumnos del Ciclo de complementación las notas en la materia Análisis de datos son más altas (aclaramos que es un caso ficticio) que los alumnos que están haciendo la Licenciatura en Ciencias de la Educación. Los profesores y deciden averiguar si esta tendencia es producto del azar o existe alguna influencia que puede explicar las diferencias.

Puesto que los alumnos no tienen la misma trayectoria, no se tiene certeza de que el promedio sea el estadístico adecuado para compararlos, por ello deciden utilizar la prueba de la mediana. Los alumnos de la materia son 186 en total, y de acuerdo al arreglo de la tabla presentada más adelante, se definió a los alumnos del Ciclo de complementación como Grupo 1, y a los alumnos de la Licenciatura como Grupo 2.

La tabla quedó conformada como se muestra a continuación:

	Grupo 1	**Grupo 2**	**Total**
Valores por encima de la Mdn1,2.	68	23	91
Valores por debajo de la Mdn1,2	9	86	95
Total	77	109	186

Consultando los valores tabulados de χ^2 y teniendo en cuenta la tabla y sabemos que si el valor observado de x2 es mayor a 3.84, tiene una probabilidad $p<0.05$. De acuerdo con este dato, determine si la hipótesis nula se puede falsear y consecuentemente si ambas muestras provienen de la misma población.

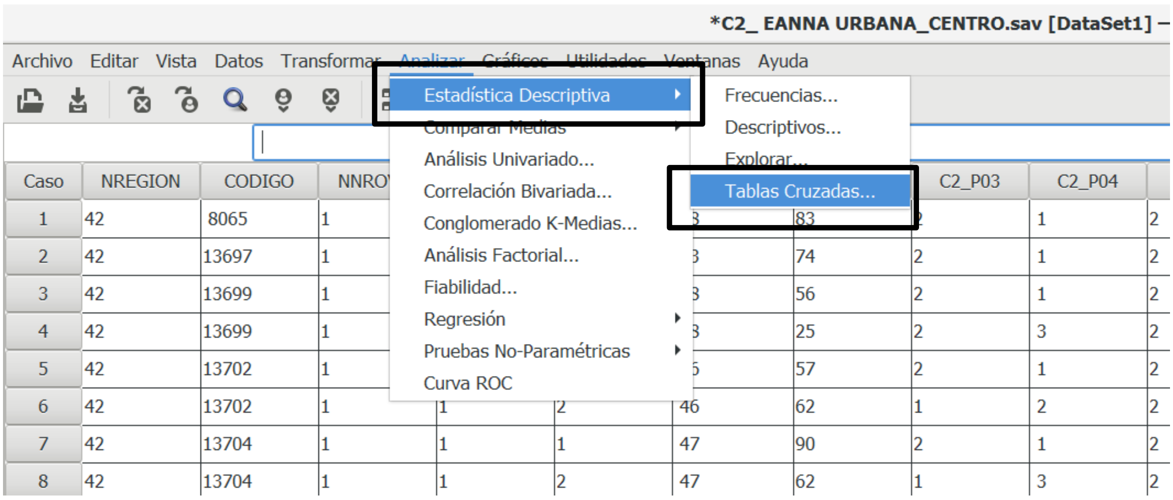
Actividades en PSPP N° 8:

Nuevamente abrimos el programa PSPP y vamos a trabajar con la base de datos que descargamos del Instituto Nacional de Estadísticas y Censos (INDEC). Para ello, abrimos el documento que guardamos con el nombre C2_EANNA URBANA_CENTRO.SAV.

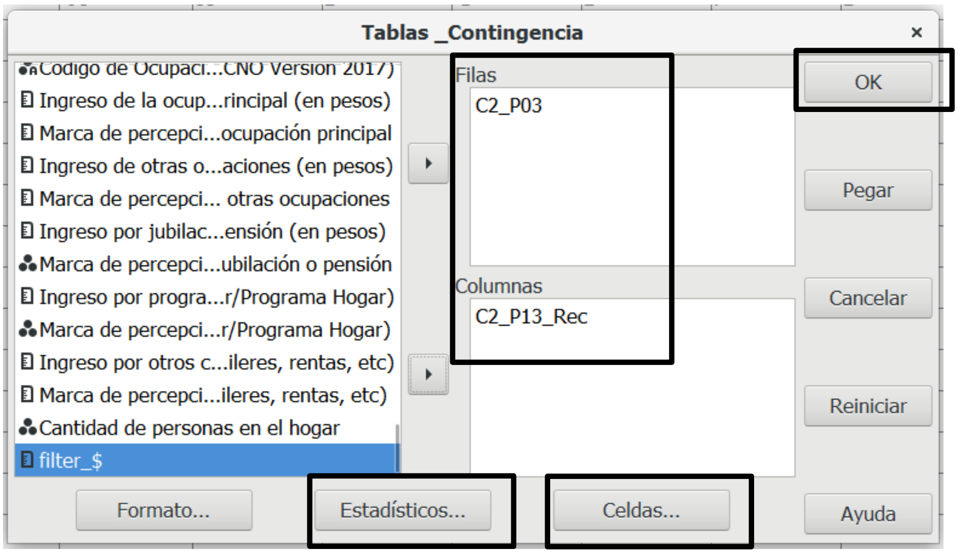
En esta oportunidad realizamos una prueba de hipótesis sobre la base del modelo estadístico chi-cuadrado, para ver la relación entre dos variables por medio del programa PSPP. Para ello, primeramente, vamos a intentarlo de modo manual y luego de modo informático. Aprenderemos a observar cómo se realizan estas solicitudes por medio del programa y de qué modo podemos leer los resultados.

Lo primero que haremos en esta oportunidad es una tabla de contingencia o una tabla cruzada con el programa PSPP, que nos permita ver como se distribuyen los casos en dos variables al mismo tiempo. Para ello, en primer lugar, vamos a ponderar la base según lo hemos realizado en unidades anteriores. Si usted ha guardado convenientemente la base encontrará que ya está ponderada. Si no, lo puede realizar nuevamente.

A posteriori, vamos a seleccionar el comando Analizar, Estadística descriptiva, Tablas cruzadas como se ve a continuación:



Una vez allí veremos un segundo cuadro de diálogo titulado Tablas_Contingencia, que a la izquierda tendrá todas las variables disponibles para realizar una tabla de contingencia y a la derecha dos cuadros, uno para las variables en filas y otro para las variables en columnas. Vamos a poner en filas a la variable C2_P02 Sexo y en las columnas a la variable que ya hemos creado C2_P13_Rec Nivel educativo más alto recodificada:



Completos estos campos, entramos en Estadísticos, destildamos todos los estadísticos y luego entramos en Celdas… y dejamos tildada sólo la casilla de recuento. Finalmente, damos a OK y tendremos automáticamente nuestra tabla de contingencia para los 915122 casos válidos.

Resumen.

	Casos					
	Válido		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
C2_P03 Sexo * Nivel educativo más alto recodificada	915122	51,1%	876132	48,9%	1791254	100,0%

C2_P03 Sexo * Nivel educativo más alto recodificada [recuento].

	Nivel educativo más alto recodificada				
C2_P03 Sexo	Hasta secundario	Superior no universitario	Universitario	Posgrado	Total
Varón	316340,0	34894,00	61416,00	2538,00	415188,0
Mujer	345052,0	64813,00	87618,00	2451,00	499934,0
Total	661392,0	99707,00	149034,0	4989,00	915122,0

Con estos resultados, se le pide que determine los grados de libertad de la tabla y que calcule si existe asociación entre las variables y el grado de asociación entre ellas:

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

………
………
………
………
………
……

También podemos pedir al programa calcule automáticamente los porcentajes en filas o los porcentajes en columnas. Para ello, vamos nuevamente a Tablas cruzadas y en Celdas destildamos Recuento y pedimos Fila y Columna -esto lo podemos hacer simultáneamente o en dos pedidos diferidos, que darán lugar a dos tablas diferentes. En este caso lo haremos simultáneamente:

Resumen.

	Casos					
	Válido		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
C2_P03 Sexo * Nivel educativo más alto recodificada	915122	51,1%	876132	48,9%	1791254	100,0%

C2_P03 Sexo * Nivel educativo más alto recodificada [fila %, columna %].

	Nivel educativo más alto recodificada				
C2_P03 Sexo	Hasta secundario	Superior no universitario	Universitario	Posgrado	Total
Varón	76,19% 47,83%	8,40% 35,00%	14,79% 41,21%	,61% 50,87%	100,00% 45,37%
Mujer	69,02% 52,17%	12,96% 65,00%	17,53% 58,79%	,49% 49,13%	100,00% 54,63%
Total	72,27% 100,00%	10,90% 100,00%	16,29% 100,00%	,55% 100,00%	100,00% 100,00%

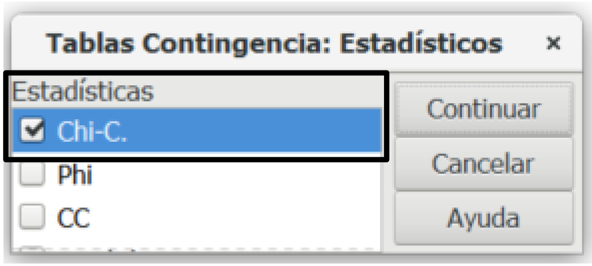
De este modo, usted puede leer la tabla en dos direcciones, por filas y por columnas, con sus totales respectivos, según considere el primer renglón de la casilla o el segundo. Se le pide que realice a continuación las dos lecturas, por filas y por columnas.

……
………
………
………
………
………
………
………
………
………
………
………
………
………

……

……

Posteriormente, usted podrá realizar el cálculo con el programa estadístico. Para esa tarea, vamos al comando Analizar, Estadística descriptiva, Tablas Cruzadas y allí vamos a estadísticos. Si se fija, tenemos muchas opciones para los estadísticos. En esta ocasión sólo seleccionaremos Chi-C, que corresponde a chi cuadrado.



Damos a continuar y obtendremos una tabla como la siguiente:

Pruebas Chi-cuadrado.

Estadístico	*Valor*	*df*	*Sig. Asint. (2-colas)*
Chi-cuadrado de Pearson	7044,76	3	,000
Razón de Semejanza	7137,34	3	,000
Asociación Lineal-by-Lineal	3495,44	1	,000
N de casos válidos	915122		

Como se puede observar en el cuadro el valor de Chi-cuadrado de Pearson es de 7044.76 con una cantidad de grados de libertad de 3 y una significación asintótica para dos colas, es decir a izquierda y derecha, que se corresponde con una hipótesis bidireccional, de 0.000, menor que 0.05 que es la que hemos puesto como máximo para rechazar H0, por lo tanto, se rechaza H0 y podemos decir que la relación entre las variables es estadísticamente significativa. Existe evidencia estadística para decir que el Sexo y el Nivel educativo más alto codificada están relacionadas.

Capítulo 9

Correlación de variables

En este capítulo vamos a estudiar específicamente la correlación entre las variables. Para ello vamos a presentar casos en diagramas de dispersión que nos permitan aproximarnos a modelos explicativos de la correlación entre ellas. Luego vamos a estudiar el coeficiente de correlación R de Pearson y el R_S de Spearman, para determinar si dos variables tienen relación entre sí según su escala de medición. Finalmente, vamos a aplicar lo estudiado sobre prueba de hipótesis al estudio de la correlación entre las variables, buscando establecer la magnitud y el signo de la relación.

Dos variables están relacionadas si los valores de una ellas se pueden predecir, en algún grado, de los valores observados en la otra. Esta idea puede comprenderse fácilmente si consideramos la altura y el peso de las personas; cuánto más alto es un individuo hay más probabilidades que sea más pesado. Esta variación conjunta entre la altura y el peso se puede estimar a partir de un coeficiente de correlación.

Determinar que existe correlación entre dos variables representa una importante fuente de conocimiento en materia de investigación, dado que podemos aprender mucho del comportamiento de una variable sobre la que sabemos muy poco, a partir del conocimiento que tenemos de otra variable de la que sabemos más.

Comenzaremos con un ejemplo sencillo de este concepto. Supongamos que queremos averiguar qué factores o variables se relación con el promedio al final del primer año de la carrera. Para ello seleccionaremos una muestra aleatoria de estudiantes de una carrera universitaria (para este ejemplo el tamaño de la muestra será igual a 30), y recogemos datos sobre el promedio obtenido al final del primer año de estudio.

En principio suponemos que mientras más horas de estudio por semana le dedique el estudiante, más alto será su promedio. Por ello, suponemos que, si el estudiante no puede disponer libremente del tiempo porque trabaja, estas horas de trabajo se restan a las horas de estudio, por lo tanto, a mayor cantidad de horas trabajadas menor el rendimiento académico y más bajo el promedio.

La historia académica del estudiante podría darnos alguna pista sobre el rendimiento, en tal caso tomamos el promedio obtenido al final de la secundaria como indicador de la trayectoria académica, suponiendo que aquellos estudiantes con mejor promedio al final del secundario, tendrán mejores promedios en la universidad.

Finalmente, la ansiedad ante los exámenes es un factor que suele disminuir la competencia académica de estudiantes capaces, pero con dificultades para enfrentarlos. Por lo tanto, suponemos que aquellos alumnos con mayores niveles de ansiedad como rasgo de personalidad, serán los que muestren promedios más bajos.

Nótese que cada vez que suponemos la existencia de una relación entre variables también suponemos una dirección para esa relación. Por ejemplo, mayor cantidad de horas dedicadas al estudio más alto el promedio, implica que el aumento en la magnitud de una variable, es seguido por el aumento en la magnitud de la otra variable. Bajo estas circunstancias se dice que la relación entre las variables es directa.

La relación también puede mostrar un sentido inverso, tal el caso en el que suponemos que altos niveles de ansiedad en el examen, dará como resultado un bajo rendimiento académico. Como se deduce, esto implica que el aumento en la magnitud de una variable, es seguido por la disminución en la magnitud de la otra variable.

Un coeficiente de correlación expresa mediante su signo esta cualidad de la relación entre variables. Así, un coeficiente de correlación positivo expresará una relación directa entre las variables, y un coeficiente negativo expresará una relación inversa.

Asimismo, las variables se relacionan entre sí con una determinada magnitud o

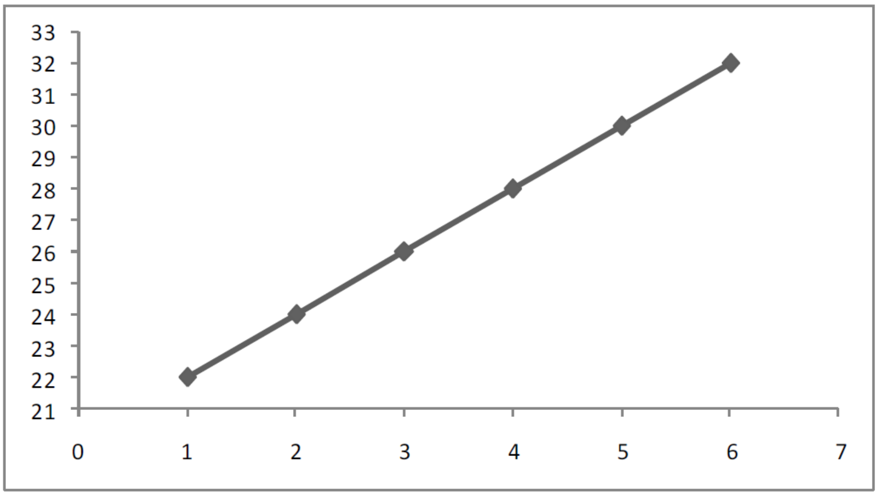
fuerza. Cuanto mayor sea la magnitud de la correlación, mayor determinación entre las variables. Por ejemplo, si el promedio al final del primer año dependiera exclusivamente de las horas de estudio, con solo saber cuántas horas le dedicó un individuo al estudio, sabríamos qué promedio obtuvo. Para determinar la magnitud y el tipo de relación conviene explorar mediante un gráfico denominado diagrama de dispersión.

El diagrama de dispersión

El diagrama de dispersión es un arreglo de las variables en ejes cartesianos que nos permite determinar qué tipo de relación existe entre ellas. En este caso nos ocuparemos de las relaciones lineales entre las variables. Comenzaremos analizando el tipo de relación entre dos variables ficticias x e y, definida por la relación y=m*x*+k, siendo m igual a 2 y k igual a 20.

X	***Y***
1	22
2	24
3	26
4	28
5	30

Como puede verse, para obtener un valor de *y*, todo lo que debemos hacer es sustituir en la ecuación el valor de *x*, por lo tanto, es sencillo construir la tabla para cualquier valor de *x*. La función y= m*x*+k, es la ecuación de una recta, y mediante ella todos los valores de *y* están perfectamente determinados por la función y por ende el diagrama cartesiano que pone en correspondencia *x* e *y* debería ser una recta, tal como se muestra en la siguiente figura.



En el eje horizontal (abscisa) se han representado los valores de *x*, y en el vertical (ordenadas) los valores de *y*. Como puede observarse la ecuación de la recta nos permite graficar justamente una recta. Lo importante para destacar aquí es que no estamos tratando con ninguna medición de un caso concreto, simplemente creamos una función de *y* cuyos valores dependen totalmente de *x*, por lo tanto, para saber cuál es el valor de *y* solo nos queda sustituir ese valor en la ecuación.

Bajo tales circunstancias el cálculo de un coeficiente de correlación sería igual a 1, lo que equivale a decir que:

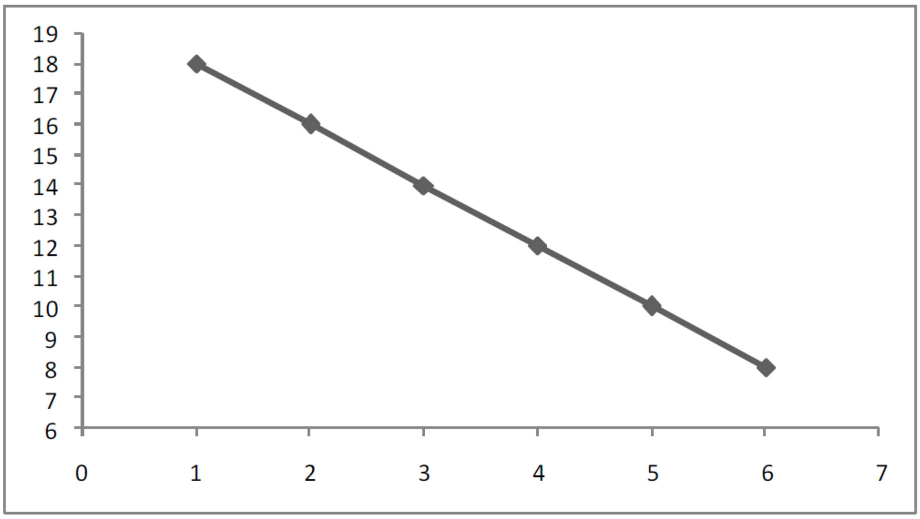
a) la correlación entre las variables es perfecta, lo que significa que, conociendo el valor de *x*, ya sabemos cuál es el valor de *y*,

b) el incremento en el valor de *x* es seguido por un incremento en el valor de *y*, puesto que el coeficiente es positivo.

Ahora vamos a utilizar los mismos valores de la tabla, pero la relación entre *x* e *y* estará dada por la función *y*= -m*x*+k, siendo m igual a -2 y k igual a 20.

X	***Y***
1	18
2	16
3	14
4	12
5	10

La única diferencia con la función anterior es que ahora la constante m es negativa, por lo tanto, la recta se oriente de manera inversa a la anterior, tal como lo muestra el siguiente gráfico:



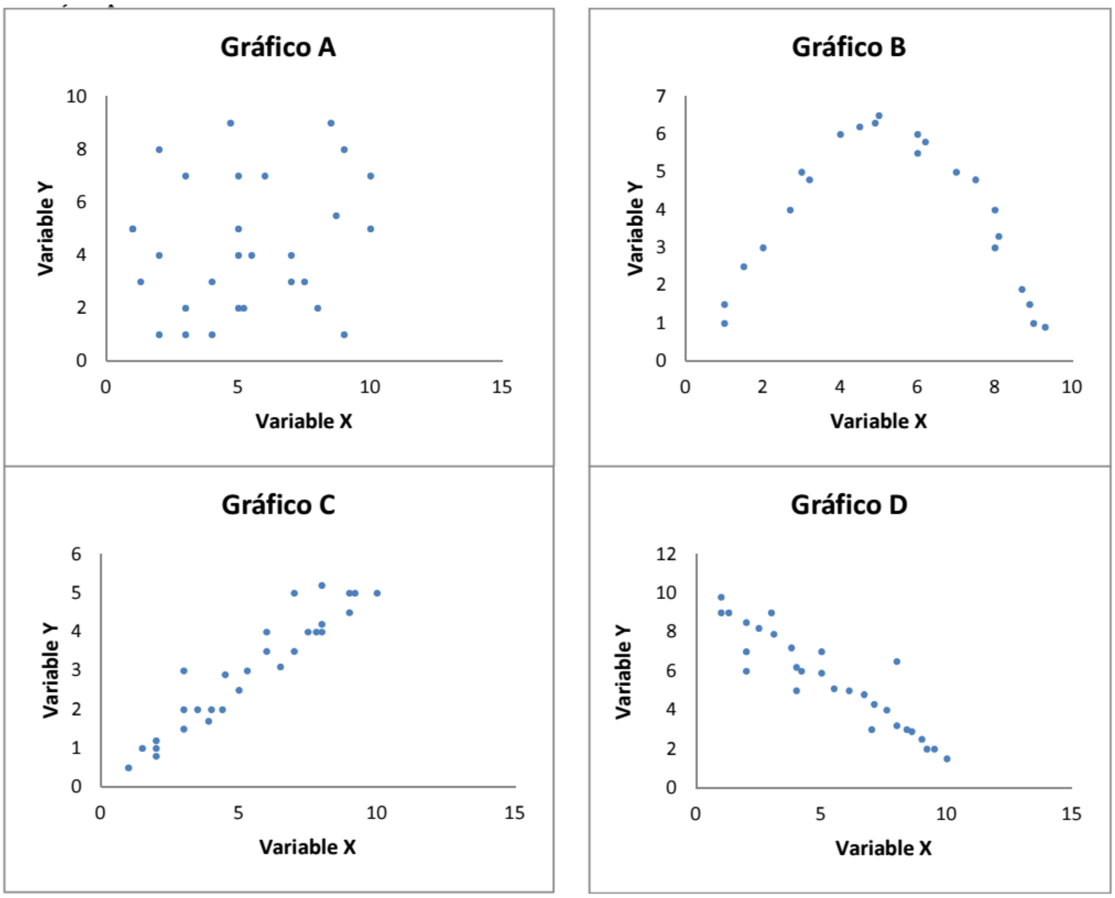
En este caso el cálculo de un coeficiente de correlación sería igual a -1, lo que equivale a decir que:

a) la correlación entre las variables es perfecta, puesto que, conociendo el valor de *x*, ya sabemos cuál es el valor de *y*,

b) el incremento en el valor de *x* es seguido por un decremento en el valor de *y*, puesto que el coeficiente es negativo.

El diagrama de dispersión nos brinda una importante ayuda para determinar si dos variables empíricas (producto de datos de la realidad) están relacionadas de manera lineal. Si lo están, el coeficiente de correlación nos indica el sentido de la relación mediante el signo, y la fuerza de la relación mediante su magnitud.

Puede suceder que las variables no se relacionen de ninguna manera, en cuyo caso diremos que ambas son independientes. También puede ocurrir que la relación no sea lineal y deba expresarse mediante algún otro modelo matemático que no sea la ecuación de la recta. En la siguiente imagen se muestra un conjunto de situaciones posibles.



Nota. En la gráfica A se observa que los puntos cartesianos o pares ordenados (*x; y*), están dispersos por todo el plano. Bajo estas circunstancias una predicción del comportamiento de la variable *x* a partir de los valores de la variable *y* es casi imposible, dicho en otras palabras, la predicción no es confiable.

En la gráfica B se observa que existe un patrón en la distribución de los puntos cartesianos, pero dicho patrón no es descripto por una recta. El comportamiento de los pares ordenados describe una figura conocida como U invertida, que es propia de las funciones cuadráticas ($y=x^2$) que describen una parábola en su gráfica (dependiendo de la ecuación, la parábola puede abrirse hacia arriba en U, o hacia abajo ∩). En ciencias sociales una de las curvas en U invertida más conocidas es la curva de Kuznets (basada en la hipótesis de la distribución propuesta por el economista Simon Kuznets), que puso en relación la distribución del ingreso (PBI) con la igualdad en su distribución.

Las gráficas C y D se asemejan a lo que ya vimos respecto de ecuaciones lineales, en donde C representa un modelo de relación positiva entre las variables y D un modelo de relación negativa. En ambos casos, la ecuación de la recta es la que mejor describe el comportamiento de las variables.

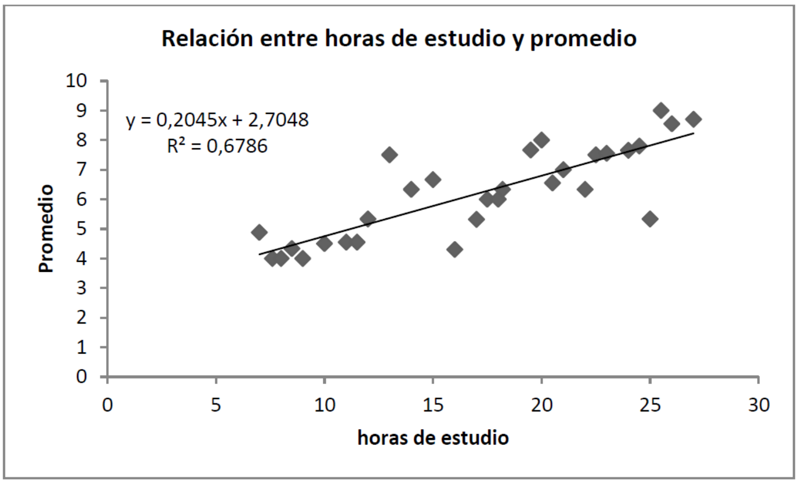
Coeficiente de correlación r de Pearson

Volvamos ahora a nuestro ejemplo, intentando determinar la relación entre las horas de estudio semanales y el rendimiento promedio. En primer lugar, creamos un diagrama de dispersión para verificar el tipo de relación que existe entre las variables.

En el ejemplo, se muestra un gráfico creado con Excel para las dos variables. Lo primero que se aprecia es que en el eje de las *y* se ha situado la variable promedio al final del primer año (promedio), en el eje de las *x* se ha situado la variable horas de estudio. Por cada valor de *y* existe un valor de *x* que ha sido representado en el plano cartesiano, dando origen a los puntos correspondientes a cada par (*x;y*.)

Estos puntos forman lo que se denomina nube de puntos y nos da una primera impresión del tipo de relación entre las variables. Recuérdese que en nuestro ejemplo suponemos una relación directa entre las variables: a más horas de estudio mejor será el promedio. Como se aprecia en la disposición de los puntos en la nube, la relación estaría en consonancia con lo supuesto, ya que a más horas de estudio más alto es el promedio.

Si la relación fuera perfecta, todos los puntos se ubicarían sobre una recta, pero este raramente es el caso con las variables empíricas. En el gráfico se ha dibujado una recta teórica, que se ajusta a la nube de puntos de manera que todos ellos queden equidistantes. Como se deduce, mientras más cercanos estén esos puntos a la línea recta, más próximo a la unidad estará el coeficiente de correlación.



En el ejemplo que nos ocupa el coeficiente de correlación r de Pearson, es r=0.823. Es un coeficiente positivo y próximo a uno, la relación entre las variables es directa y fuerte. Esto expresa que, a mayor cantidad de horas de estudio semanales, mejor será el promedio.

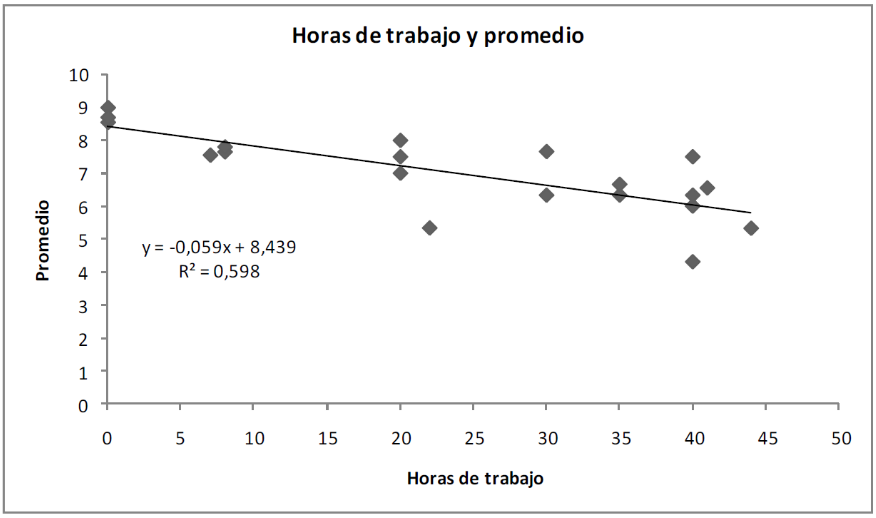
Puesto que la relación entre las variables es lineal, se puede expresar mediante la ecuación de la recta que en este caso es y=0.204*x*+2.704. Esta ecuación, denominada ecuación de regresión, permite calcular cualquier valor de *x*, pero debemos tener en cuenta que los valores de *x* calculados estarán sujetos a un cierto error de estimación, dado que la relación entre las variables no es igual a 1.

En el gráfico también se muestra otro coeficiente, llamado de determinación y que expresa el monto de varianza de *y* que puede explicarse a partir de *x*. Este coeficiente R^2 puede interpretarse a partir de su valor porcentual, que en nuestro ejemplo es de 67,8%.

Ahora sí estamos en condiciones de afirmar que aquellos alumnos que dedican más horas de estudio semanales tendrán un mejor promedio al final del año. Si nos preguntamos cuánto es lo que podemos explicar de la variación en los promedios de todos los individuos de nuestra muestra, utilizando las horas de estudio como variable independiente, podemos afirmar que aproximadamente el 68%. Finalmente, si nos preguntamos cuál sería aproximadamente el promedio de un individuo que estudiara 35 horas semanales, aplicamos la ecuación de regresión:

$$\boxed{y = 0.204x35 + 2.704 = 9{,}84}$$

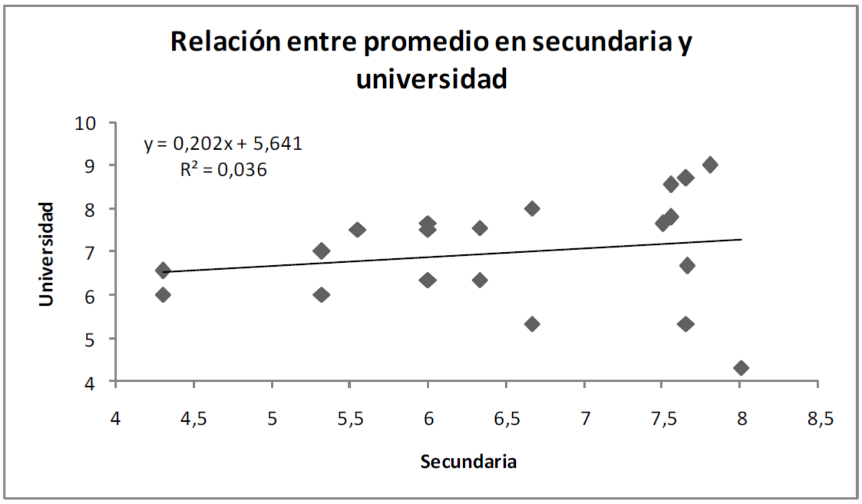
Podemos suponer también que, si el estudiante no puede disponer libremente del tiempo porque trabaja, las horas de trabajo se restan a las horas de estudio, por lo tanto, a mayor cantidad de horas trabajadas menor el rendimiento académico y más bajo el promedio. En la muestra de estudiantes encontramos algunos que no trabajan, otros que trabajan algunas horas semanales, y otros que lo hacen en jornadas de ocho o más horas semanales. En el diagrama de dispersión queda expresada la relación entre las variables y nos permite determinar cómo afecta el trabajo al rendimiento académico. Nuevamente, la relación se ha expresado mediante un gráfico de Excel.



El coeficiente de correlación de Pearson en este caso es r=-0.773. Es un coeficiente alto, pero de signo negativo, lo cual implica una relación inversa entre las variables. El coeficiente pone en evidencia que, a mayor cantidad de horas de trabajo, menor el promedio del estudiante. La cantidad de varianza que puede ser explicada del promedio, utilizando las horas de trabajo como variable independiente, es aproximadamente el 60%. Finalmente, si quisiéramos estimar cual sería el promedio de una persona que trabaja 25 horas semanales, utilizamos la ecuación de regresión:

$$\boxed{y=-0.059x25+8.439=6{,}964\simeq 7}$$

Otra de las variables que nos proponíamos analizar, era la historia académica del estudiante suponiendo que aquellos estudiantes con mejor promedio al final del secundario, tendrán mejores promedios en la universidad. Procediendo de la misma manera en que lo hicimos en los casos anteriores, ponemos las variables en correspondencia en un diagrama de dispersión, tal como se muestra a continuación.



El coeficiente de correlación en este caso es r=0.189, un coeficiente positivo de baja magnitud. Esto se interpreta como una relación muy débil entre las variables. El promedio al final del secundario nos informa muy poco del rendimiento académico de un estudiante en universidad, aun más, si quisiéramos explicar el rendimiento académico universitario del conjunto de datos a partir de la variable promedio en el secundario, solo alcanzaríamos a dar cuenta del 3,6% de la varianza.

En este punto analizaremos más detenidamente el gráfico. Recordemos que, en el caso de una asociación perfecta entre las variables, todos los puntos de la nube caen sobre la recta de regresión. A medida que el coeficiente se aleja de la unidad y se aproxima a cero la nube de puntos se hace cada vez más dispersa en torno a la recta. Esto significa que es cada vez más imprecisa alguna predicción de los valores de una variable a partir de los valores de la otra variable.

Veamos ahora el gráfico, y se notará que los puntos se dispersan cada vez más a medida que ascendemos en los valores de *x*, que en este caso es promedio en secundaria. Si tomamos el valor promedio 4,33 en secundaria, notaremos que hay dos valores próximos entre sí en la variable promedio en la universidad, estos son 6 y 6,55, pero si tomamos el valor 7,55 en la variable promedio en secundaria notaremos que hay valores de 5 hasta 8,5 aproximadamente. En tal caso, la predicción que pudiéramos hacer sería demasiado ambigua e inexacta, dada la dispersión de los valores. Esta es la razón por la cual el coeficiente de correlación entre las variables tiene una magnitud tan baja.

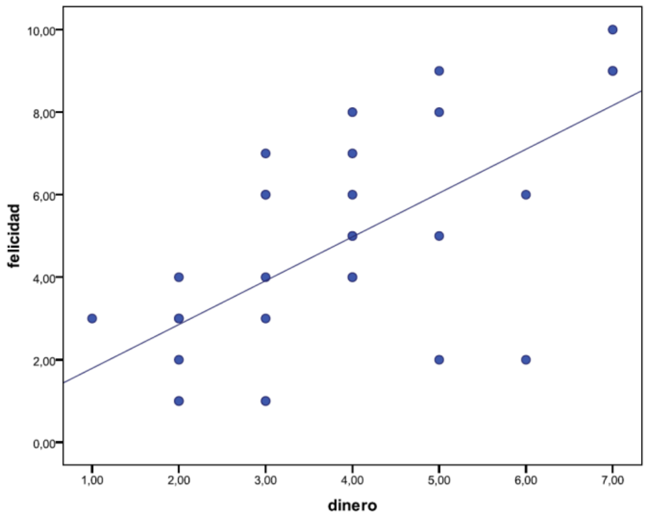
Coeficiente de correlación r_S de Spearman

Charles Spearman durante sus investigaciones sobre la inteligencia humana, desarrollo un coeficiente de correlación basado en rangos, que hoy lleva su nombre. Este coeficiente de correlación es una opción cuando una o ambas variables se han medido en escala ordinal. Es decir, es una medida que puede aplicarse a aquellos casos en que las escalas no son métricas. Este coeficiente resulta de mucha utilidad cuando se trabajan con escalas de tipo Llickert, donde los individuos juzgan un atributo en una escala ordinal especialmente diseñada para la ocasión.

Veamos un ejemplo: En un estudio se pidió a un grupo de personas que juzgaran simultáneamente su estado de ánimo actual y la importancia que ellas le otorgaban al dinero. El objetivo de la investigación consistió en determinar si existía relación entre el estado de ánimo y factores materiales. Para el estado de ánimo se utilizó una escala autoadministrada de felicidad en donde el individuo debía indicar en un orden de 0 a 10 cómo se sentía en ese momento, siendo 10 el nivel más alto y correspondiente a muy feliz. Igualmente, para la importancia del dinero, se utilizó un orden de 0 a 7 donde el máximo puntaje representaba muy importante. Dado que ambas variables están diseñadas sobre una escala ordinal y no poseen la misma cantidad de reactivos, se utilizó el coeficiente de correlación r_S para determinar el grado de correlación entre las variables.

Como en los casos anteriores analizaremos primero el diagrama de dispersión.

En este caso ha sido graficado sobre un total de 30 casos para cada variable y utilizando el programa SPSS.



Como puede observarse, el diagrama sugiere una correlación lineal positiva, de manera que aquellos individuos que juzgan el dinero como más importante, son los que tienden a mostrarse más felices. Nótese sin embargo que existen personas con altos niveles de felicidad, que no le dan al dinero demasiada importancia, juzgándolo en valores bajos. Estos individuos se apartan de la tendencia general. Además, se evidencia que hay individuos que le otorga un valor de importancia muy alto al dinero, pero no manifiesta un alto grado de felicidad. Con estos indicios es de suponer que el coeficiente de correlación tendrá un valor medio-alto. El cálculo del coeficiente nos muestra que efectivamente este alcanza el valor r_s=0.619.

El coeficiente de correlación ordinal de Spearman nos permite trabajar con variables ordinales y su interpretación es la misma que ya vimos para el coeficiente de correlación de Pearson. Sin embargo, no es conveniente con este tipo de coeficiente calcular una ecuación de regresión o el coeficiente de determinación R^2, dada la naturaleza ordinal de las variables.

Coeficiente de correlación y prueba de hipótesis

Se puede plantear como hipótesis que dos variables están relacionadas, en tal caso, el planteo implica determinar la magnitud y el signo de la correlación encontrada entre las variables con una distribución teórica que estipula que ambas variables no tienen relación. Esto último corresponde a la hipótesis nula.

Como hemos mencionado, la prueba de hipótesis se basa en un resultado sobre el cual rechazaremos o no esta hipótesis, basándonos en la probabilidad de encontrar un resultado específico, considerando cierta la hipótesis nula. Un ejemplo, nos permitirá clarificar los conceptos expuestos para este caso.

Basándose en estudios previos, un investigador supone que la madurez escolar está relacionada con las habilidades lingüísticas. Para probar esta afirmación evalúa una muestra de 12 escolares con dos pruebas específicas, una para cada variable. Las pruebas utilizadas están diseñadas de modo que sus puntuaciones responden a una escala métrica, por lo cual el investigador decide comprobar la relación utilizando el coeficiente de correlación de Pearson. Los datos recogidos se muestran en la siguiente tabla:

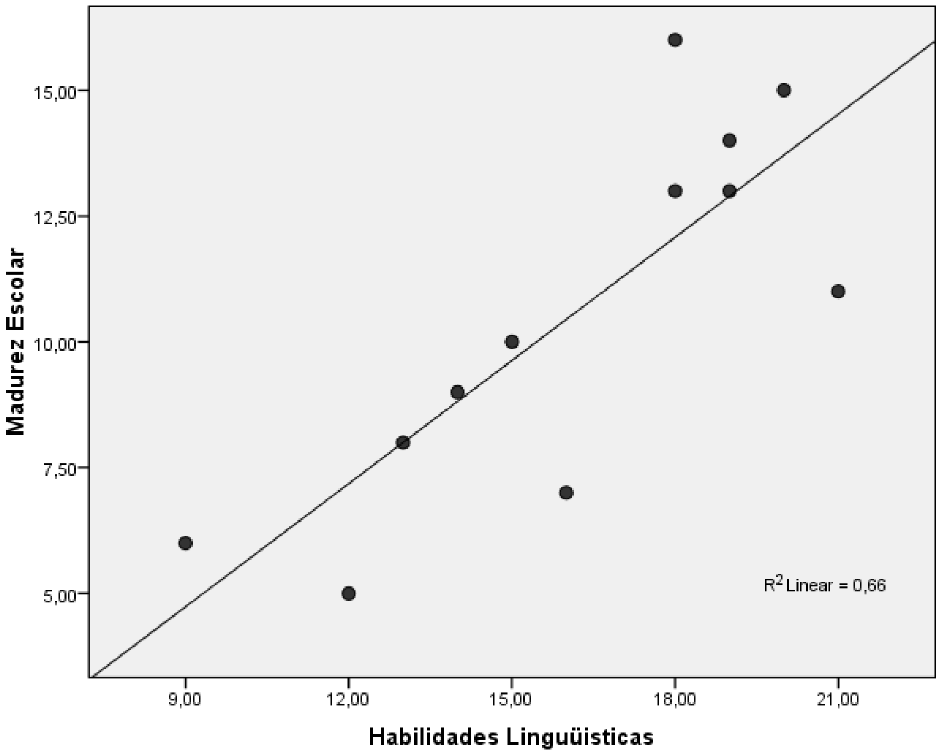
Sujetos	**Madurez Escolar (*x*)**	**Habilidades lingüísticas (*y*)**
1	10	15
2	14	19
3	9	14
4	8	13
5	13	18
6	13	19
7	16	18
8	6	9
9	11	21
10	7	16
11	5	12
12	15	20

La tabla muestra en la primera columna los sujetos que participaron en el estudio, la segunda columna muestra los puntajes obtenidos en madurez escolar que hemos denominado como *x*, la tercera columna muestra los puntajes en la prueba de habilidades lingüísticas que hemos denominado *y*. La hipótesis nula implica negar la hipótesis de investigación, por lo cual su planteo implica que no existe relación entre las variables. Ahora corresponde expresar en términos estadísticos ambas hipótesis:

$$H_0: r=0$$
$$H_1: r\neq 0$$

Nótese que la formulación estadística supone que el coeficiente de correlación será cero bajo el supuesto de la hipótesis nula. Al contrario, la hipótesis de investigación supone un coeficiente de correlación diferente de cero. Corresponde ahora determinar bajo qué circunstancias es factible rechazar la hipótesis nula. Pues bien, suponiendo que en la población el coeficiente de correlación para las variables mencionadas es cero, la probabilidad de encontrar un coeficiente de una magnitud distinta de cero resulta muy poco probable. Bajo estas condiciones establecemos el valor de α correspondiente a la probabilidad de cometer un error tipo I. Usualmente este valor es α=0.05.

El diagrama de dispersión y el cálculo del valor del coeficiente de correlación, junto a su probabilidad de ocurrencia bajo la hipótesis nula, se realizó utilizando el programa SPSS.



El cálculo del coeficiente de correlación de Pearson r=0,812, y la probabilidad asociada a este coeficiente bajo la hipótesis nula es *p*=0,001. Ahora bien, considerando que la probabilidad de haber obtenido un coeficiente de correlación como el reportado es muy baja tomando en cuenta la hipótesis nula, resulta que el supuesto de que ambas variables no tienen relación debe ser rechazado. En otras palabras, rechazamos la hipótesis nula que supone que no existe relación entre las variables, a favor de la hipótesis que sostiene que las variables están relacionadas.

Finalmente vemos que el coeficiente de determinación, tomando como variable independiente a las habilidades lingüísticas, es de 0,66. Como ya dijimos, esto equivale a que el 66% de la varianza en la madurez escolar puede explicarse mediante las habilidades lingüísticas del individuo.

Veamos otro ejemplo de prueba de hipótesis para este coeficiente de determinación, donde la hipótesis de investigación es más específica: bajo el supuesto que la memoria es un recurso directamente involucrado en el desarrollo de las capacidades verbales, en un experimento se analizó si la amplitud de la memoria verbal inmediata, se relacionaba con la capacidad de resolver el cierre de sentido de oraciones sintácticamente ambiguas.

Los investigadores utilizaron la prueba de retención de dígitos para medir la amplitud de memoria verbal, la cual consiste en repetir de manera correcta series de números no consecutivos que van 2 a 9 dígitos. Por su parte, la capacidad de resolución de ambigüedades se estudió mediante la aplicación de la prueba estandarizada diseñada a tal fin (Prueba de Cloze), en la cual se asigna un puntaje de cero a veinte, de acuerdo a la cantidad de respuestas correctas.

En la etapa inicial del estudio, se tomó una muestra de diez estudiantes universitarios y los resultados de la aplicación de estas pruebas se detallan en la siguiente tabla.

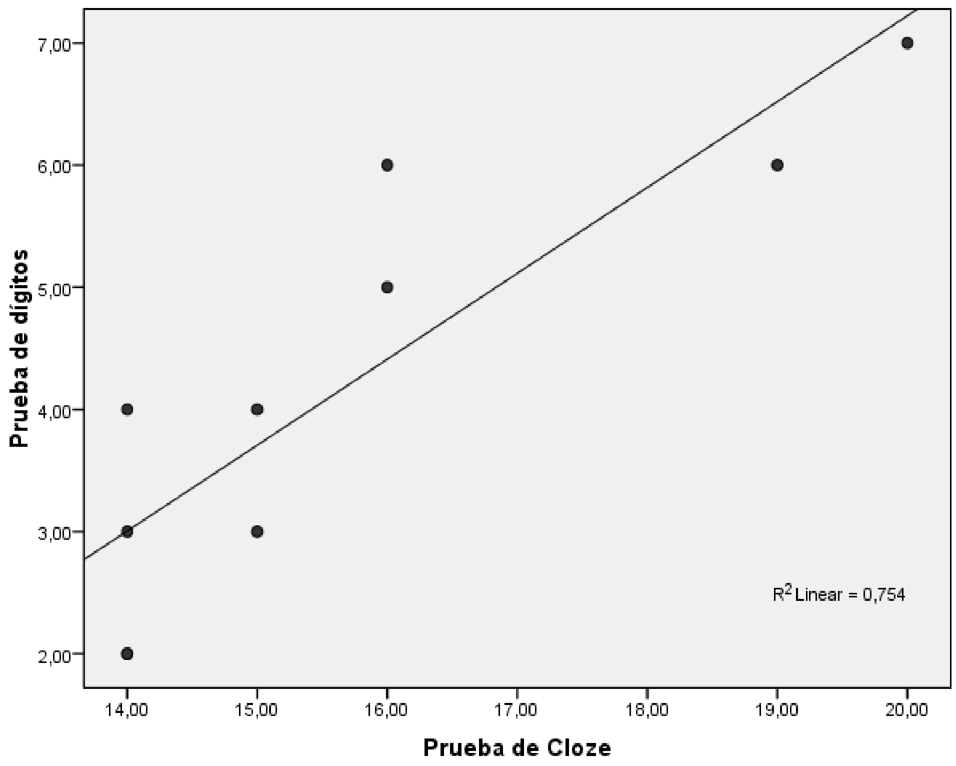
Si bien la prueba de Cloze es una prueba estándar, sus puntajes no se consideran de escala métrica, sino ordinal. Por lo tanto, el coeficiente de correlación que se emplea para la prueba de hipótesis es r_s.

Casos	**Prueba de dígitos (*x*)**	**Prueba de Cloze (*y*)**
1	2	17
2	3	17
3	2	14
4	4	18
5	3	13
6	4	14
7	5	15
8	6	16
9	6	15
10	7	15

Dado que los investigadores suponen que la memoria está directamente involucrada como recurso cognitivo en la prueba de resolución de ambigüedades, se espera encontrar una correlación directa (positiva) entre las variables. La formulación estadística de la hipótesis de investigación y la hipótesis nula sería:

$$H_0: r=0$$
$$H_1: r>0$$

Seguidamente analizamos el diagrama de dispersión y se calcula el valor del coeficiente de correlación, junto a su probabilidad de ocurrencia bajo la hipótesis nula. En este ejemplo también, utilizamos el programa SPSS para realizar esos cálculos.



El cálculo del coeficiente de correlación de Spearman r_s=0,889, y la probabilidad asociada a este coeficiente bajo la hipótesis nula es p<0,001. Aquí también resulta que la probabilidad de obtener un coeficiente de correlación como el reportado es muy baja, por lo tanto, se rechaza la hipótesis que ambas variables no tienen relación. En otras palabras, rechazamos la hipótesis nula. El coeficiente de determinación, tomando como variable independiente a la amplitud de memoria verbal, es de 0,754, lo cual representa una varianza explicada de 75,4% en la prueba de Cloze.

Algunas precisiones sobre el coeficiente de correlación

Un coeficiente de correlación indica la magnitud de una relación lineal, a mayor valor de r, mayor grado de correlación, por ende, si se compara una correlación de r=0.2 con una r=0.4, se concluye que esta última es mayor que la primera, pero no es correcto afirmar que es el doble. Si se desea comparar dos o más coeficientes se utiliza el cuadrado de la correlación r, esto es R^2 que, como ya vimos, representa la proporción de la varianza explicada en la variable dependiente; esto se denomina reducción proporcional del error.

De acuerdo a lo dicho, para comparar dos coeficientes de correlación como r_1=0.2, y r_2=0.4, las elevamos al cuadrado, y entonces tenemos que R^2_1= 0.04 y

R^2_2=0.16. En proporción, la magnitud de la segunda correlación es cuatro veces mayor que la primera y no dos veces como se deduciría si usáramos el coeficiente de correlación r.

Guía de Actividades N° 9

Correlación entre variables y prueba de hipótesis

Objetivos de la actividad:

a) Determinar si un conjunto de puntos, nube de datos, se aproxima a una relación entre las variables y distinguir el tipo de relación.

b) Calcular el coeficiente de relación entre las variables según sea el caso y establecer hipótesis de correlación.

Actividad 1: Un docente de matemáticas cree que el rendimiento de sus alumnos está relacionado con las horas que dedican a hacer sus tareas. Para ello, ha pedido a sus alumnos que cronometren minutos dedicados a su materia durante las dos semanas que ha durado la lección. Luego los ha evaluado convenientemente al finalizar la lección por medio de una prueba estandarizada con una escala proporcional de 1 a 10 y estos son los resultados:

Caso	Minutos de estudio	Calificación
1	79	10
2	90	10
3	120	10
4	90	8
5	66	8
6	80	7
7	59	7
8	100	7
9	60	6
10	82	6
11	25	5
12	45	5
13	84	5
14	70	4
15	55	4
16	15	3
17	21	3
18	0	2
19	50	2
20	45	2

Se le solicita que grafique y determine a través del gráfico si existe relación entre las variables y el tipo de relación que estas tienen. Además, se le solicita que asocie la distribución de los datos con algunas de los gráficos vistos durante la unidad.

Actividad 2: Volvamos sobre la tabla de la actividad 1. Luego de hacer algunas operaciones estadísticas el profesor ha estimado la regresión y ha calculado el coeficiente de correlación de Pearson. Los resultados que halló son los siguientes: $y=0.0665x+1.5921$ y el valor de R^2 = 0.5873. Se le solicita que interprete los resultados.

Actividad 3: Se pidió a un grupo de personas que juzgaran su estado de satisfacción con una evaluación de estadísticas que acababan de rendir, con una escala auto administrada de Likert de 1 a 7, conde 1 era completamente insatisfecho y 7 era completamente satisfecho.

El objetivo de la investigación consistió en determinar si existía relación entre el estado de satisfacción con su desempeño en la evaluación y el resultado obtenido. Para el resultado se utilizó una escala de 1 a 7 donde el máximo puntaje representaba la mejor resolución del examen y el mínimo la que se encontró más incompleta.
Los resultados son los siguientes:

Caso	Autopercepción	Calificación2
1	1	2
2	1	2
3	1	3
4	1	3
5	1	5
6	2	3
7	2	4
8	2	4
9	2	5
10	3	4
11	3	3
12	3	5
13	3	6
14	3	7
15	4	3
16	4	6
17	5	5
18	5	7
19	6	4
20	7	7

Realice el diagrama de dispersión. Identifique si éste sugiere una correlación lineal y su signo. Interprete los resultados.

Identifique qué coeficiente de correlación debe calcularse en este caso y considerando que el valor del coeficiente que corresponde es 0.6842 describa los resultados.

Actividad 4: Un equipo de investigación, en función de estudios previos, supone que el rendimiento escolar está relacionado con la cantidad de libros que el alumno posee en su casa. Para probar esta afirmación evalúa una muestra de 29 escolares con su promedio durante el año y una breve encuesta donde se le demanda que cuente la cantidad de libros de humanidades, ciencias o literatura que tiene en su hogar. Ambas variables están diseñadas de modo que sus puntuaciones responden a una escala métrica, por lo cual el investigador decide comprobar la relación utilizando el coeficiente de correlación de Pearson.

Los datos recogidos se muestran en la siguiente tabla:

Caso	**Calificación**	**Cantidad de libros**
1	6	16
2	8	10
3	4,625	4
4	3,5	2
5	4,25	6
6	9,5	47
7	1,875	3
8	3,25	0
9	3	4
10	5,75	8
11	6,25	14
12	4,5	10
13	6,5	23
14	3	5
15	4,5	6
16	6	15
17	3	2
18	5	9
19	8,5	62
20	4,375	5
21	6,5	15
22	7	9
23	4,25	7
24	4,375	7
25	5,125	7
26	4	3
27	7	11
28	5	5
29	9,5	122

Plantee las hipótesis del ejercicio y explicite bajo qué circunstancias es factible de rechazar la hipótesis nula con un valor de $\alpha=0.05$. Para realizar esta tarea le brindamos el coeficiente dos coeficientes de correlación:

Medidas simétricas

		Valor	Error estándar asintótico[a]	T aproximada[b]	Significación aproximada
Intervalo por intervalo	R de Pearson	,720	,065	5,393	,000[c]
Ordinal por ordinal	Correlación de Spearman	,881	,035	9,670	,000[c]
N de casos válidos		29			

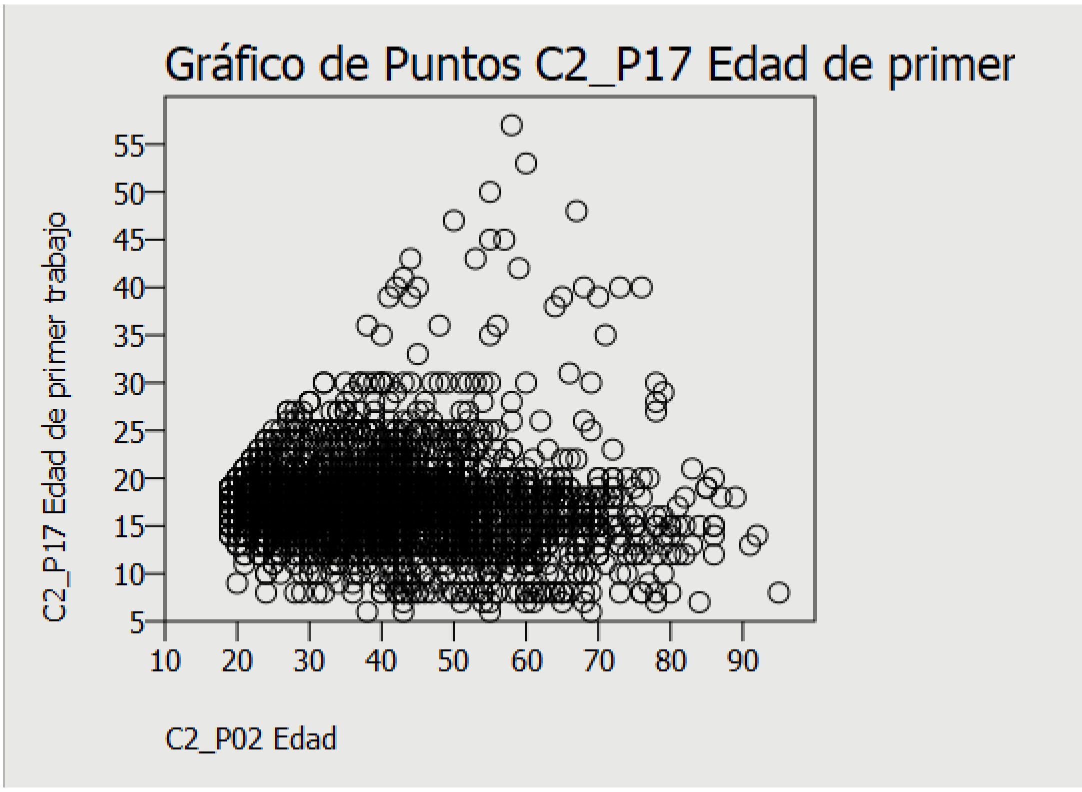
Usted debe seleccionar el adecuado e interpretar los resultados a la luz de los estadísticos propuestos.

………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
………
……

Actividades en PSPP N° 9:

En esta oportunidad vamos a calcular los coeficientes de correlación que hemos trabajado durante la unidad por medio del programa PSPP. Abrimos el programa PSPP y vamos a trabajar la base de datos que descargamos del Instituto Nacional de Estadísticas y Censos (INDEC). Para ello, abrimos el documento que guardamos con el nombre C2_EANNA URBANA_CENTRO.SAV.

Una vez abierta la base de datos y ponderada, lo primero que haremos es obtener una nube de datos, es decir, un diagrama de dispersión. Para ello vamos a tomar las variables C2_P02 Edad y la variable C2_P17 Edad del primer trabajo. Antes de realizar cualquier operación, damos por perdido el valor 99 de la variable C2_P17. Luego pedimos un gráfico de puntos. Para ello, vamos a Gráficos, Grafico de Puntos, y seleccionamos en el eje Y: C2_P17 y en el eje X: C2_P02. Damos a ok y tendremos:



Allí podemos observar un gráfico de puntos de dos variables con muchos casos, más precisamente los que ponderados dan 833822 casos.

Resumen.

	Casos					
	Válido		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
C2_P02 Edad * C2_P17 Edad de primer trabajo	833822	46,5%	957432	53,5%	1791254	100,0%

Luego he pedido una tabla cruzada para estas dos variables, indicando en Celdas que queremos sólo los recuentos y en Estadísticos sólo Chi-Cuadrado y Correlación (Corr), como hemos realizado con anterioridad.

Veremos que la tabla de por sí es inmanejable por sus dimensiones. Ustedes pueden comprobarlo sólo con ver esta esquina a continuación o realizando la operación por sí mismos, pero es probable que el programa no funcione correctamente.

C2_P02 Edad * C2_P17 Edad de primer trabajo [recuento].

C2_P02 Edad	6	7	8	9	10	11	12	13	14	15	16	17	
19	,00	,00	,00	,00	,00	,00	,00	,00	200,00	2071,00	1627,00	3208,00	64
20	,00	,00	,00	232,00	,00	,00	,00	1558,00	409,00	854,00	749,00	1412,00	31
21	,00	,00	,00	,00	,00	284,00	383,00	331,00	715,00	1918,00	986,00	2080,00	71
22	,00	,00	,00	,00	,00	,00	,00	,00	,00	460,00	2631,00	1701,00	43
23	,00	,00	,00	,00	,00	,00	296,00	941,00	819,00	1764,00	770,00	1618,00	27
24	,00	,00	267,00	,00	749,00	,00	,00	1501,00	1130,00	2281,00	2021,00	1731,00	63
25	,00	,00	,00	,00	,00	410,00	28,00	793,00	2686,00	2612,00	2644,00	2559,00	61
26	,00	,00	,00	,00	270,00	,00	1072,00	657,00	1834,00	1985,00	2428,00	2612,00	62
27	,00	,00	,00	,00	,00	,00	,00	1155,00	1037,00	2279,00	1907,00	1751,00	32
28	,00	,00	445,00	,00	,00	136,00	1078,00	2050,00	2579,00	2193,00	2887,00	1244,00	42
29	,00	,00	354,00	,00	433,00	,00	173,00	136,00	1433,00	3586,00	2410,00	2793,00	57
30	,00	,00	,00	,00	214,00	109,00	411,00	318,00	271,00	3887,00	2728,00	4016,00	48
31	,00	,00	214,00	,00	514,00	,00	890,00	123,00	1601,00	4301,00	2393,00	3694,00	56
32	,00	,00	98,00	,00	,00	,00	1046,00	1185,00	1205,00	1524,00	4851,00	4165,00	46
33	,00	,00	,00	,00	476,00	181,00	481,00	1331,00	3619,00	1896,00	1758,00	2259,00	75
34	,00	,00	,00	674,00	,00	490,00	894,00	1229,00	105,00	2614,00	1390,00	2528,00	55
35	,00	,00	,00	,00	,00	,00	56,00	1345,00	2232,00	5751,00	2828,00	2438,00	90
36	,00	,00	110,00	93,00	411,00	106,00	1180,00	258,00	684,00	1832,00	1660,00	3279,00	76

El tamaño de la tabla resultante es muy complejo para leer los datos, y mucho más para pensar si existe una relación entre las variables. No obstante, los estadísticos a continuación nos brindan información acerca de la relación entre las variables:

Pruebas Chi-cuadrado.

Estadístico	*Valor*	*df*	*Sig. Asint. (2-colas)*
Chi-cuadrado de Pearson	1493293,87	2880	,000
Razón de Semejanza	543133,30	2880	,000
Asociación Lineal-by-Lineal	2865,50	1	,000
N de casos válidos	833822		

Medidas simétricas.

Categoría	*Estadístico*	*Valor*	*Err. Est. Asint.*	*T Aproxim.*	*Sign. Aproxim.*
Ordinal según Ordinal	Correlación de Spearman	-,08	,00	-73,52	
Intervalo según Intervalo	R de Pearson	-,06	,00	-53,62	
N de casos válidos		833822			

Se le pide que interprete los resultados:

....

....

....

....

....

....

………
………
………
………
………
………
………
………
………
………
……

Ahora vamos a realizar el mismo tipo de análisis con variables ordinales. Para ello, nos dirigimos a Tablas cruzadas, y seleccionamos en las filas la variable C2_P39 Tamaño del establecimiento (en la que previamente debemos dar por perdido el valor 9) y en las columnas la variable C2_P13_Rec. En Estadísticos seleccionamos además de la opción Chi-Cuadrado, la opción Corr y en celdas porcentajes por columnas. El resultado debería dar 579999 casos válidos y la siguiente tabla:

Resumen.

	Casos					
	Válido		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
C2_P39 Tamaño del establecimiento * Nivel educativo más alto recodificada	579999	32,4%	1211255	67,6%	1791254	100,0%

C2_P39 Tamaño del establecimiento * Nivel educativo más alto recodificada [columna %].

	Nivel educativo más alto recodificada				
C2_P39 Tamaño del establecimiento	Hasta secundario	Superior no universitario	Universitario	Posgrado	Total
1 persona	27,02%	9,99%	11,77%	,00%	22,00%
2 a 5 personas	30,74%	23,86%	18,99%	43,13%	27,86%
6 a 15 personas	13,44%	15,55%	15,59%	,00%	13,98%
16 a 40 personas	11,74%	25,17%	15,93%	,00%	14,00%
41 a 200 personas	11,50%	13,65%	24,31%	14,48%	14,14%
más de 200 personas	5,57%	11,79%	13,41%	42,39%	8,02%
Total	100,00%	100,00%	100,00%	100,00%	100,00%

Como pueden apreciar la variable C2_P39 es ordinal. Usted podrá preguntarse ¿por qué esta la variable C2_P39 es ordinal? Esto responde a que la última categoría está abierta y entonces los intervalos no son proporcionales. Más allá de esto, se puede observar perfectamente en la tabla la relación entre las variables con sólo apreciar que los que tienen hasta secundario se concentran en establecimientos hasta 40 personas, con un 82.94% de los casos. Y los establecimientos de más de 200 personas tienen apenas un 5.57% de los casos. Esta relación se va invirtiendo gradualmente en la medida en que aumenta el nivel educativo.

No obstante, como esta observación estimada o "a ojo", carece de rigor estadístico y podemos verificarla por medio de la tabla de correlaciones:

Pruebas Chi-cuadrado.

Estadístico	*Valor*	*df*	*Sig. Asint. (2-colas)*
Chi-cuadrado de Pearson	53656,48	15	,000
Razón de Semejanza	51967,72	15	,000
Asociación Lineal-by-Lineal	37876,67	1	,000
N de casos válidos	579999		

Medidas simétricas.

Categoría	*Estadístico*	*Valor*	*Err. Est. Asint.*	*T Aproxim.*	*Sign. Aproxim.*
Ordinal según Ordinal	Correlación de Spearman	,26	,00	205,92	
Intervalo según Intervalo	R de Pearson	,26	,00	201,30	
N de casos válidos		579999			

Se le pide que interprete los resultados:

……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
…

Con este ejercicio hemos terminado las guías de actividades en PSPP. Si bien el libro se propone brindar los conocimientos básicos para manejar las nociones más importantes de la estadística aplicada a las Ciencias de la Educación, hemos diseñado esta guía de PSPP con la firme convicción que en la actualidad la mayoría de las operaciones que se muestran en este libro se realizan sobre las bases de datos con algún tipo de software estadístico.

Capítulo 10

Contribución de Prof.ª Mariana Alejandra Pino[7]

¿Qué son los indicadores?

Cuando hablamos de indicadores nos estamos refiriendo a un término polisémico y utilizado en diversos campos de análisis. Específicamente en ciencias sociales hablamos de indicadores producto de la operacionalización de conceptos, que se entiende como el proceso de definición de las variables de manera estratégica para la medición de conceptos y/o fenómenos de la realidad social que no poseen una medida determinada (Capuano,2014). Un indicador, es entonces un valor posible de observar, es decir, aquel atributo de una unidad de análisis que por su ausencia o por su grado de presencia, nos remite a hacernos preguntas que se pueden traducir en una investigación.

Si nos remitimos a sus orígenes, los primeros indicadores fueron creados en el campo económico, con el objetivo de sintetizar información y hacerla comprensible, describiendo y evaluando realidades complejas.

Seguramente has leído titulares tales como "*(...)El gasto público educativo en 2019 se situó en el 4,26% del PIB*" [8]. En este titular se utiliza el indicador económico de **producto bruto interno,** uno de los indicadores más utilizados en el campo de la economía, y también en diversos campos, para comparar el gasto público estatal.

En este sentido es importante recuperar que el uso de indicadores inicia en la década de 1960, tales como el PIB, renta per cápita, entre otros, que tomaron gran relevancia. Por sus características aportan confiabilidad para el uso en los análisis económicos.

Se podría identificar el uso de indicadores aproximadamente un siglo antes de lo mencionado. Actualmente los indicadores poseen gran relevancia dentro del campo de la investigación social, y se conciben como un instrumento que viene a "dar luz", a la realidad social. Recuperando los aportes de Durkheim (1897) se entiende a los indicadores como "aquello que inicia la indagación". Aquello que se expresa en un lenguaje común, tales como docentes, escuelas, planes de estudio, entre otros conceptos que no se asumen "medibles", por medio de la construcción de indicadores se transforman en conceptos científicos para las ciencias sociales y para la educación (Capuano, 2014).

En el siglo XX, el sociólogo Paul Lazarsfeld (1979) sostiene - desde una postura

---

[7] Capítulo realizado bajo la supervisión de los profesores Jorge Lorenzo y Manuel Giovine.

[8] Disponible en: https://www.magisnet.com/2021/03/el-gasto-publico-en-educacion-se-situo-en-2019-en-el-426-del-pib-frente-al-421-de-2018/#:~:text=El%20gasto%20p%C3%BAblico%20educativo%20en,5%2C04%25%20del%20PIB. Fecha: 3-Mar-2021

positivista - que ninguna ciencia tienen la posibilidad de analizar un objeto de estudio en su plenitud. Las ciencias en general determinan ciertas propiedades e intentan establecer relaciones entre ellas para explicar los fenómenos del modo más complejo posible. Para esto el Lazarsfeld propone una estrategia considerada empírica basada en la construcción de un lenguaje, el que denomina “lenguaje de las variables”, con el fin de que:

> “Todos los fenómenos de interés para la investigación social podrían ser conceptualizados y medidos al menos en cierto nivel, correlacionados y manipulados de diversas maneras por las técnicas de formales del análisis de variables, permitiendo formular y poner a prueba hipótesis” (Hughes y Sharrock, 1999; en Capuano,2014: 2).

A posteriori, Robert Merton, un teórico funcionalista se diferencia de los postulados de Lazarsfeld, dado que él sostenía que la visión de los fenómenos sociales en términos generales perdía vinculación entre la teoría y la investigación. Para obtener esa mayor vinculación, Merton pensaba que era necesaria la construcción de indicadores, es decir, brindarles la concepción de “concepto”, basados en aquello que es posible observar de los fenómenos y aquello que entre las variables se puede relacionar en términos empíricos. De esta manera buscaba realizar análisis de datos minuciosamente claros, con mayor precisión para la investigación social.

Finalmente, a inicios del siglo XXI, se establecen sistemas de indicadores por ranking, éstos sostienen la necesidad de establecer medidas estándares para conceptos tales como “nivel de vida”, “condiciones sociales”, “condiciones de desarrollo”, entre otros conceptos, que con el avanzado proceso de globalización se emplearan por los diversos organismos internacionales para la comparación entre las naciones.

Indicadores en la Educación Argentina.

Con este breve recorrido por la construcción conceptual de los indicadores, proponemos adentrarnos a pensar en los indicadores en el campo de la educación argentina. Vamos a realizar un breve recorrido histórico, con la pretensión de comprender ciertas dinámicas del contexto, la función de la enseñanza, los cambios y sus resultados.

En la década de 1990 ya se utilizaban los datos estadísticos para la elaboración de políticas públicas. Pero se presentaron dos problemáticas, por un lado, no se consideraba que los datos sean lo suficientemente representativos. Esto se debía a la descentralización del sistema educativo, que tuvo sus inicios durante la última dictadura militar en la década de la década del 1970.

Es por esto por lo que nace la necesidad de establecer un Sistema Nacional de Estadísticas que tiene la capacidad de ser adaptable a lo largo y ancho del país. Más adelante en este texto desarrollaremos, en qué consistes el Sistema Nacional de Indicadores, pero es importante comprender históricamente como se llegó a su definición.

En 1991 se establece la ley de Transferencia de los Servicios Educativos a las provincias, donde se descentraliza el nivel secundario y superior no universitario (el nivel primario había sido transferido a las provincias en la década de 1970), de esta forma se hace efectiva la descentralización de la totalidad del sistema educativo. En 1993 se establece la Ley Federal de Educación (N° 24.195), donde se organiza el sistema educativo en su totalidad, sus diferentes niveles, modalidades y se extiende su obligatoriedad a 10 años. También se crea la Red Federal de Información Educativa (RedFIE), dependencia del Ministerio de Educación.

Esta dirección pretende concentrar la información de todo el sistema educativo de forma homogénea. De esta manera se pretendía poseer la suficiente información para la elaboración de políticas públicas representativas y confiables para acompañar las transformaciones educativas de todo el país.

En este punto es importante recordar la gran influencia que poseían los organismos internacionales en la década de 1990:

> "(...)los bancos internacionales (principalmente el Banco Interamericano de Desarrollo y el Banco Mundial) que actuaban como fuente de financiamiento de los programas de reforma educativa, tuvieron también una fuerte presencia en la determinación de las prioridades y las estrategias elegidas, entre ellas, generar un sistema de información y de medición del rendimiento del sistema" (Pascual, 2016: 1)

En el año 94' se realiza el Censo Nacional de Docentes y Establecimientos educativos, primer dispositivo de recolección de datos del país. En él se podían obtener datos del sistema educativo como: matriculaciones, docentes, infraestructura de las instituciones, recursos, entre otros. No obstante, al ser un censo se realiza cada 10 años, y las transformaciones en el sistema educativo se producen como mucha celeridad. Es así como este recurso de recolección de datos dejaba de ser representativo y no permitía el seguimiento de las políticas públicas

En el año 1996 se desarrolla un sistema de recolección de datos estandarizado para todo el país denominado como "Relevamiento anual" (RA), que es utilizado hasta la actualidad. En él se recogen los datos anualmente de cada escuela, de cada provincia en todo el territorio nacional, y se establecen criterios metodológicos y procedimientos comunes para todas las jurisdicciones. De este modo, el Estado Nacional establece políticas públicas para todo el sistema educativo, y las jurisdicciones también pueden establecer políticas públicas específicas para su población siguiendo estos criterios comunes.

En el año 2016 con la creación de Ley de acceso a la información pública, nos permitió poder acceder a información relevante del sistema educativo. Esta iniciativa como otras en nuestro país, posee una fuerte influencia de los organismos internacionales, que permite que toda persona interesada pueda observar el funcionamiento del sistema educativo, no solo el Estado Nacional.

En la actualidad contamos con una amplia cantidad de recursos que nos brindan datos del sistema educativo, así como diversos agentes de producción de datos, tanto a nivel internacional, nacional, provincial, y municipal que nos permite un amplio campo de conocimiento para la creación de políticas públicas e investigación educativa. A continuación, les vamos a dejar una lista de los recursos de datos disponibles a nivel nacional:

1)Censo Nacional. Su realización es cada 10 año, el último fue en 2022.

2)Relevamiento Anual.

3)Censo Nacional de Docentes y Establecimiento educativos. Realizado en 1994.

4)Censo Docente. Realizado en 2004.

5)Relevamiento de Escuelas Rurales. Realizado en 2006 y 2009.

6) Padrón Oficial de Establecimientos Educativos. Con actualización de forma continua y automática desde 2011.

7) Sistema Integral de Información Digital Educativa (SInIDE). Inició en el año 2012, aportando información estadística de actualización constante por medio de un sitio web, incluyendo todos los niveles del sistema educativo y el nivel universitario. Siendo este un sistema de carga y consulta continua.

8) Censo Nacional del Personal de los Establecimientos Educativos - CENPE. Primer censo realizado en 2014, que mide la nueva estructura establecida con la Ley de Educación Nacional (2006). Se realiza cada 10 años con el objetivo de cuantificar el personal docente y no docente de todos los establecimientos educativos contemplados por la ley de todo el país.

Importancia de los indicadores educativos.

Habiendo dicho esto, se torna sumamente importante el desarrollo teórico-metodológico de los indicadores educativos, sus características, sus usos y el análisis del sistema nacional de indicadores.

Para iniciar es necesario comprender, en primer lugar, que los indicadores educativos nacionales fueron diseñados de forma metodológica que permita vislumbrar aspectos centrales del sistema educativo, con el fin de evaluar, planificar y diseñar políticas públicas, "(...)Los indicadores educativos suelen definirse como medidas estadísticas sobre aspectos que se consideran importantes del sistema educativo." (Morduchowicz,2006: 2).

Es en estos términos que un indicador educativo es un instrumento que nos posibilita dar cuenta de los fenómenos educativos. Nos brinda un panorama de la situación del sistema educativo. Y es importante comprender que como instrumento no es información en sí misma, si no que el indicador es la operacionalización, la construcción de aquello que consideramos observable; y hay que diferenciar entre indicadores y la producción de análisis estadísticos que responden a las necesidades establecidas para la elaboración de un informe.

Es de esta manera que comprendemos a los indicadores educativos como medidas estadísticas que enfatizan sobre un tema en particular, establecidos por medio de un consenso general, pero que se entienden variables a lo largo del tiempo (Morduchiwicz, 2006).

Con la presentación conceptual que se ha realizado hasta este punto consideramos que hemos podido hacer énfasis en lo importante, que es el desarrollo de indicadores educativos. En este punto definiremos: ¿Para qué son importantes?

Para responder a esta pregunta es necesario comprender que los indicadores deben reunir una serie de características específicas para que su utilidad sea precisa. Sus principales características, definidas y generales para todos los indicadores son:

- Que deben ser *fácilmente mensurables*, es decir que deben poder medirse. Sabemos que en el campo de la educación hay fenómenos que no son medibles. Pero realizando procesos de operacionalización y estableciendo criterios consensuados es posible identificar aquello que es "válido", observar y medir en particular de cada fenómeno.
- También son *definidos con relación al contexto.* Esto está ligado a la característica anterior, cada país estipula cuáles son las necesidades de su sistema educativo y de este modo definirá los indicadores en función de sus objetivos. Esto transforma al indicador en "preciso", útil como herramienta cuantificable y metódica.
- También debe establecerse con criterios claros, que permitiría ser *descriptivos* del sistema educativo.
- Es necesario que sean *medibles a lo largo del tiempo*. Esto los hace confiables.
- Tienen que *dar cuenta de la diversidad o desigualdades* que existen en el sistema educativo nacional.
- Deberán ser *transparentes* y *tener relación* entre el fenómeno de la realidad social y la definición del indicador.

Los indicadores poseen una gran complejidad, y como ya hemos podido ver su especificidad está centrada en el campo de lo cuantitativo. Por tal motivo es que para su comprensión es necesario conocer conceptos como el de Unidad de análisis. Con este concepto, por ejemplo, es posible identificar de acuerdo con el tipo de indicador seleccionado a qué o a quién refieren son los datos que nos brinda. Nos permite realizar los diferentes análisis necesarios y de esta manera medir el comportamiento del sistema educativo. Un ejemplo sería: Si tomamos un indicador del área de Recursos humanos, tal como podría ser “Máximo nivel alcanzado por los docentes”, podemos determinar que nuestra unidad de análisis “son los docentes”. A su vez podemos comprender que el tipo de variable ordinal y si tenemos datos sobre este indicador, podemos resumir la información, hacerla legible y por medio de diversas técnicas.

Por mencionar solo algunos de los contenidos que desarrollaremos a lo largo de la asignatura. Lo que nos brinda la posibilidad de analizarla, hacer inferencias, sintetizar y realizar conclusiones, para las resoluciones de las problemáticas o investigar lo que puedan presentarse, como necesidad, en forma de datos.

Clasificación de indicadores.

Así como los indicadores poseen una serie de características también pueden ser clasificados.

Su clasificación ha ganado un consenso generalizado en los diversos autores que desarrollan este tema, pero nos vamos a centrar en la clasificación definida por Alejandro Morduchowicz (2006).

El autor define una clasificación en seis tipos de indicadores que exponemos a continuación:

<table>
<tr><td colspan="4">Clasificación de indicadores según Alejandro Morduchowicz.[9]</td></tr>
<tr><td colspan="4">Clasificación de Indicadores</td></tr>
<tr><td rowspan="2">Descriptivos</td><td rowspan="2">Permiten graficar el estado situacional en cada momento, la forma en que funciona el sistema educativo. Y esta clasificación posee una serie de componentes:</td><td>Demanda</td><td>Mide el volumen, características centrales de la población y necesidades. Definiendo sus capacidades explícitas o implícitas. ¿Qué bienes y servicios requiere la comunidad?</td></tr>
<tr><td>Capacidad instalada</td><td>Identificación de variables descriptivas de la composición interna. Maneras en la que se distribuyen los recursos estructurales e insumos. ¿Como se satisfacen las necesidades? ¿Dónde? ¿Con qué estructura? ¿Quién es el responsable?</td></tr>
</table>

[9] Morduchowicz, A. “Los indicadores educativos y las dimensiones que lo integran.” Año 2006. UNESCO - Instituto Internacional de Planeamiento de la Educación. Buenos Aires. pág. 5

		Resultados programados	Permiten analizar metas y acciones estipulando los objetivos a alcanzar. ¿Qué productos o bienes se esperan obtener? ¿cuáles como resultado de la gestión? ¿Cuántos? ¿Con qué características?
		Resultados obtenidos	Nos posibilita la medición de los resultados o productos del servicio o actividad desarrollada. ¿Qué productos o bienes se obtuvieron como resultado de la gestión? ¿Cuántos? ¿Con qué características?
Explicativos	No solo permiten la descripción de la actividad, sino que también nos permite acercarnos a la identificación de los factores que determinan su estado y/o situación.		
Simples	Son observables, surgen de la cuantificación de las características o atributos que se pretenden describir.		
Elaborados o construidos	Medibles por medio de funciones que contienen más de una variable.		
Sincrónicos	Son medibles en un contexto y en un corte de tiempo estático. Los indicadores de eficiencia y eficacia se presentan como datos sincrónicos en el estado de situación y en el monitoreo.		
Diacrónicos	Son aquellos indicadores que manifiestan la evolución en el tiempo de un fenómeno determinado.		

Pero estas no son las únicas formas de clasificación, también están las clasificaciones entre indicadores:

- ✓ cuantitativos, aquellos estrictamente medibles expresados en números tales como “Alumnos por Unidad Educativa”;
- ✓ y cualitativos, que miden cualidades, juicios prácticos, entre otras opciones medibles específicas del sistema educativo, por ejemplo “Clima institucional”.

Esta es otra forma de clasificación que requiere de mayor especificación de lo que se desea medir por su dificultad de establecer una unidad de medida estándar, por lo que requiere de un concepto y una fundamentación conceptual muy precisa para que posean valor.

Sistema de indicadores: ¿Todo sistema que brinda información es un sistema de indicadores?

Para comprender qué es un sistema de indicadores, inicialmente hay que diferenciarlo de lo que es un sistema de información. Estos dos son complejos de ser diferenciados si no hacemos este análisis conceptual.

En primer lugar, los *sistemas de información* son reconocidos por la exhaustividad de la información, es decir se construyen y observan todos los datos obtenidos por el sistema educativo. Lo que hace que lleve mucho tiempo su construcción y no sea sencillo. Por lo tanto, es posible poder observar representados argumentos más que por datos; Imposibilitando el análisis de los fenómenos o hechos de la realidad social.

Otro inconveniente de los sistemas de información es que no se encuentran relacionados los datos entre sí. No poseen una orden y como consecuencia podrían brindar una información línea o simplista de los fenómenos o hechos del sistema educativo.

En primer lugar, el *sistema de indicadores* se construye desde una base teórica o un marco conceptual de lo que se necesita observar, anteponiendo los postulados epistemológicos y políticos de cómo se entiende la educación. Es de este modo que nos permite, inicialmente, discernir qué información se necesita observar, posibilitando relacionar los datos entre sí e interpretar la situaciones, fenómenos o hechos en el marco del sistema educativo. Por otro lado, nos permite el monitoreo, la comparación, evaluación y medición de la evolución del sistema educativo, de una política pública o de un proyecto de investigación.

Es importante destacar que un solo indicador, de forma independiente, nos ofrece información, pero no nos permite captar la complejidad de los hechos. Es en este punto que radica la importancia de su *carácter sistémico,* porque si bien un solo indicador nos sirve para identificar, por ejemplo, una problemática, va a ser el sistema de indicadores el que nos permita una visión más compleja y representativa de la situación problemática.

Los indicadores son agrupados de forma tal que nos permitan ver el sistema educativo en su complejidad, conocer las variables asociadas a determinados fenómenos, determinar tendencias posibles, realizar inferencias sobre las situaciones o hechos e incluso la resolución de problemas, pero también nos permiten hacer análisis parciales, para ver si se están cumpliendo con las metas u objetivos establecidos sobre el sistema educativo. Y como ya hemos dicho, los indicadores son en su génesis medidas cuantitativas, pero en su análisis conjunto, agrupados, habilitan a construir datos, información o explicar fenómenos que surgen de los datos cualitativos.

Criterios y características de los sistemas de indicadores.

Si bien muchas de las características de los sistemas de indicadores se expusieron, en este punto, serán sistematizados con el objetivo didáctico de su comprensión. Como futuros Licenciadxs en Ciencias de la Educación, el uso del sistema de indicadores posee tres formas principales, por un lado, la elaboración de políticas públicas o proyectos educativos, por otro la investigación educativa, y finalmente, el más común puede llegar a ser el de asesoramiento pedagógico y/o análisis educativo. Es en este sentido radica la importancia de conocer los criterios en los que se constituye la construcción de un sistema de indicadores; en particular el nacional.

Para seguir avanzado la construcción de un sistema de indicadores, recuperamos lo que plantea A. Tiana Ferrer (1997) "(...)el principal interés de los indicadores consiste en ofrecer una información sintética, relevante, y significativa sobre una parcela de la realidad(...)." (52)

Pero para que el funcionamiento del sistema de indicadores sea preciso tiene que ser construido sobre ciertos criterios; a) en primer lugar, deben estar construidos en base al *consenso y la aceptación* general, y en particular de los representantes políticos. b) es necesario que tenga una *tendencia integral,* con relación entre el diagnóstico y un indicador en particular, y c) también la visión individual de un indicador *debe hacer posible identificar el modelo* general del sistema de indicadores. En el caso del Sistema Nacional de indicadores el modelo que se utiliza es de los más sencillos y comunes, compuesto por cinco dimensiones: contexto-recursos-procesos-resultados- impacto.

d) Uno de los criterios importantes para un sistema de indicadores es que deben *reflejar la complejidad* de las relaciones entre los indicadores y el sistema educativo. e) Los sistemas de indicadores tiene que poder ser *comparables* con el de otros países f) y deberían tener la posibilidad de que esta comparación sea a *lo largo del tiempo*, así como también poseer la capacidad de *observar los resultados.*

g) Otra cuestión sobre la que existe consenso en los diversos autores consultados es que los sistemas de indicadores deben estar compuestos por la *menor cantidad de indicadores posible*, pero la selección de estos debe poder representar la complejidad del sistema educativo al que pertenecen. h) Finalmente, ligado a este criterio, debe existir *una relación entre los indicadores,* donde se haga visible también la interdependencia con las variables educativas, que haría que el sistema de indicadores sea relevante, consistente, válido y oportuno como herramienta de consulta y análisis.

Sistema Nacional de Indicadores.

Los sistemas educativos, en particular de Latinoamérica, surgen a partir del

Proyecto Regional de indicadores Educativos del año 2002. Donde se establecieron y consensuaron los criterios, características y construcción metodológica en las que cada país se centraría para la definición de su propio sistema de indicadores.

Hasta ahora hemos hecho un detallado desarrollo de estos consensos comunes a todos los sistemas de indicadores de la región, en esta instancia ingresamos a analizar el sistema nacional de indicadores de nuestro país en particular.

Lo primero que podemos decir es que, al igual que todos los sistemas de indicadores, el de nuestro país se sienta en las bases de valoración de *la calidad y la equidad educativa,* que no es solo una tarea de carácter técnico.

Al hablar de *calidad educativa*, nos enfrentamos a un concepto polisémico y que se ha ido transformando a lo largo de los años. En la actualidad, sintéticamente podríamos definir que se comprende como la capacidad de proporcionar a los estudiantes el dominio de los conocimientos básicos comunes, de los códigos culturales, de la adquisición de capacidad para ejercer la ciudadanía en democracia. (J. Mejía, 2015)

Así también nuestro sistema educativo sostiene como objetivo que todos los habitantes del territorio argentino tengan acceso a la educación como derecho humano. Con el objetivo de garantizar la igualdad de oportunidades a todos los sectores de la población. Pero también el de generar acciones políticas para que los diversos sectores de la sociedad argentina tengan la posibilidad de acceder, participar y alcanzar los resultados educativos esperados, definido esto como *equidad educativa.*

Estas categorías se ven reflejadas, de un modo más o menos borroso, en las políticas públicas y en los indicadores que lo conforman. De esta manera, las problemáticas o características observadas que nos permiten identificar y/o tomar en consideración serán aquellas que alteren la calidad y la equidad educativa. Así como, también, nuestro sistema de indicadores nos permitirá vislumbrar la complejidad del sistema educativo.

De esta manera el sistema Nacional de indicadores posibilita que habiliten criterios para la evaluación, seguimiento y monitoreo del sistema educativo.

Cada sistema de indicadores posee un modelo de clasificación. En el caso del sistema de indicadores definido en Argentina se comprende por cinco áreas, tal como podemos observar en la siguiente tabla[10], que desarrollaremos a continuación:

[10] Manual metodológico: Sistema Nacional de Indicadores Educativos. Laboratorio de Estadística. Red Federal de Información Educativa. Dirección Nacional de Información y Evaluación de la Calidad Educativa. Ministerio de Educación, Ciencia y Tecnología. Argentina. Pág. 11

- <u>Área de Contexto:</u> esta mide las condiciones del ambiente en el cual se desenvuelve el sistema educativo. Principalmente dejando en evidencia los patrones demográficos, socioeconómicos y su influencia en la educación y en los objetivos establecidos.

	Área	Dimensiòn
1	Contexto	Demográfico
		Socioeconómico
2	Recursos	Físicos
		Humanos
		Financieros
		Organizacionales
3	Proceso	Acceso
		Flujo
4	Resultado	Egreso
		Logros académicos
5	Impacto	Socio-econòmico

Tabla 1- Distribución de dimensiones por áreas del "Sistema Nacional de Indicadores

- <u>Área de Recursos:</u>

  Esta área describe los recursos con los que cuenta el sistema educativo, en cuatro dimensiones: físicos, relacionados a la infraestructura; humanos, contabilizando la cantidad y situación de los docentes que componen el sistema educativo; financieros, en los que se observa directamente con el presupuesto disponible y la distribución de los gastos y finalmente la dimensión organizacional, que establece la capacidad de cobertura del sistema educativo.
- <u>Área de Proceso:</u>

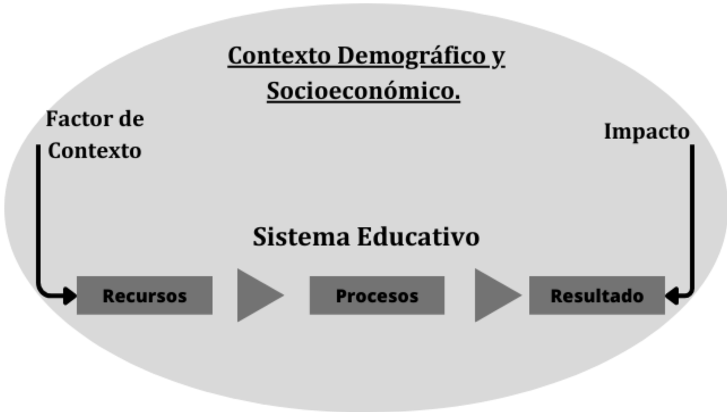
  El área de proceso se conforma por: la dimensión de acceso, la cual mide la posibilidad de ingresar a la educación de los habitantes del territorio argentino. Está ligada con la dimensión de flujo que analiza la continuidad de los estudiantes en el sistema educativo. De esta manera, esta área permite hacer notorias las características de los procesos de enseñanza y aprendizaje al nivel del aula.
- <u>Área de Resultados:</u>

  En esta área se miden los resultados obtenidos en función de las diversas políticas públicas que pretende resolver problemáticas en particulares y también sostener los objetivos generales ligados a la igualdad de oportunidades, acceso equitativo de los diversos sectores sociales de la nación, el desarrollo económico, la inserción laboral, el aprendizaje de los conocimientos prioritarios, entre otros. Esta dimensión se mide en relación con la cantidad de egresos y logros académicos y es lo que brinda un panorama para el monitoreo y evaluación del sistema educativo.
- <u>Área de Impacto:</u>

  Finalmente, la última dimensión del sistema Nacional de indicadores es la de impacto; conformada por una única área denominada socio-económica. Esta va a medir los efectos que la educación está teniendo en la sociedad argentina en relación con la formación integral, analizando el impacto social con relación a la inserción laboral, las capacidades adquiridas de los individuos y el desarrollo de conocimientos en las diversas áreas, principalmente en el área de la ciencia y la tecnología. También tiene lugar el impacto social de la baja de los índices de pobreza y de desigualdad social.

Pero ni las áreas, ni las dimensiones, ni los indicadores de forma individual funcionan, como ya hemos mencionado, todo sistema a de indicadores tienen la necesidad de ser observado de forma agrupada e interrelacionada.

Ilustración 1 Comportamiento de las diversas áreas y sus relaciones en el "Sistema Nacional de Indicadores Educativos"



Tal como podemos ver en esta imagen[11], cada dimensión actúa en relación con la otra al momento de establecer los objetivos e implementar las políticas públicas.

Como ya dejamos en evidencia el *contexto* demográfico y socioeconómico, revela las problemáticas y los factores de demanda de la comunidad educativa.

Una vez determinadas las necesidades, se analizan los *recursos* con los que cuenta el sistema educativo en sus cuatro dimensiones: capacidad física, por ejemplo: ¿Contamos con los establecimientos suficientes?; humanos ¿Los docentes cuentan con la formación necesaria para el desarrollo de un proyecto determinado?; financieros ¿Contamos con el financiamiento para el desarrollo del proyecto? y organizacional ¿Cómo y qué tipo de modalidad de escuelas serán prioritarias para el proyecto educativo?

Una vez que se determinan estas dos dimensiones necesitamos conocer la dimensión de *procesos.* En esta dimensión, vamos a observar las características y las posibilidades de los habitantes que acceden acceder a la educación y factores medibles

---

[11] Esta *imagen* se extrae del Manual metodológico: Sistema Nacional de Indicadores Educativos. Laboratorio de Estadística. Red Federal de Información Educativa. Dirección Nacional de Información y Evaluación de la Calidad Educativa. Ministerio de Educación, Ciencia y Tecnología. Argentina. Pág. 10

del aula que suceden en el nivel micro. Por ejemplo, el indicador de "esperanza de vida", nos permite medir en años el tiempo que se espera que un individuo estudie dentro del sistema educativo y se espera que tenga correlación con la tasa de escolarización en relación con la edad. De esta manera, se mide el flujo comprendido como el movimiento del sistema educativo.

Con relación a esta dimensión, y todavía observando hacia adentro del sistema educativo y su funcionamiento, se miden los resultados obtenidos de la política pública o del proyecto educativo. Los resultados se miden en la cantidad de egresados, el tiempo en que se egresan y cuando hablamos de logros académicos se tienen en consideración los resultados de las evaluaciones estandarizadas.

Finalmente, el área de Impacto deja pasa a observar hacia afuera del sistema educativo y mira el contexto, pero ahora con el objetivo de monitorear los resultados que el sistema educativo ha obtenido.

Una política pública para ser medida en términos de eficiencia debe generar un impacto en la sociedad, un movimiento en las otras dimensiones, es decir un cambio en sus indicadores. Por ejemplo, podemos observar la dimensión de flujo, la más comúnmente analizada. Luego de haber establecido una política pública que pretende reducir la tasa de abandono interanual, es esperable que dicha tasa se haya reducido en relación con cómo estaba antes de llevar adelante acciones sobre dicha situación.

Esto que he desarrollado es lo que se entiende por algunos autores como *matriz de relaciones* y por otros como el *carácter sistémico* de los indicadores. Es por esto por lo que con un solo indicador no se puede evaluar o monitorear el sistema educativo, se puede partir desde un indicador en particular, pero siempre será necesario pasar por todo el sistema de indicadores. Esto no quiere decir que haya que tomar la totalidad de los indicadores, sino que hay que seleccionar un conjunto de indicadores que resulten relevantes, para que en su conjunto permitan observar la complejidad de hechos o fenómenos del sistema educativo con el objetivo de comprenderlos, monitorearlos, analizarlos y evaluarlos.

Indicadores seleccionados.

La selección de indicadores de un sistema educativo refleja la importancia que el Estado Nacional brinda a ciertos aspectos, por ejemplo, en Guatemala el 53% de la población pertenece a uno de los 22 pueblos originarios del país[12], siendo el segundo país con mayor población nativa de la región. Y este factor de contexto demográfico hace que este país seleccione un indicador dentro del área de proceso que mide el "Idioma utilizado como medio de enseñanza". Este indicador está ligado a los criterios de selección, como una forma de representar la complejidad de su sistema educativo y una visión panorámica de este en particular. Este indicador, por ejemplo, no es común a otros, nuestro sistema educativo no mide el uso de lenguas

[12] Dato disponible: https://www.iwgia.org/es/guatemala.html Fecha: 28 de marzo de 2023.

nativas, por ende, no se desarrollan políticas públicas que abarquen la totalidad del país que aborden dicha temática.

Como ya hemos desarrollado, tampoco es suficiente este indicador para medir el sistema educativo de Guatemala, pero sí nos puede servir como punto de partida para diversos análisis, así como para compararlo con el nuestro.

Siguiendo con la selección de los indicadores de nuestro sistema educativo, los indicadores que funcionan como inspiración de los objetivos de las políticas públicas de nuestro país, son los expresados en el siguiente cuadro[13]:

Indicadores Seleccionados		
Área	**Dimensión**	**Indicador**
Contexto	Demográfico	Distribución porcentual de la población por edades simples hasta 24 años, por grupos quinquenales desde 25 años. Total, y para cada sexo.
		Tasa anual media de crecimiento poblacional.
		Porcentaje de población urbana.
	Económico	PBI per cápita / PBG per cápita[14].
Recursos	Físicos	Porcentaje de edificios escolares con materiales precarios.
		Porcentajes de edificios escolares con distribución interna de agua.
		Porcentaje de edificios escolares con provisión de energía eléctrica.
		Metros cuadrados edificados por alumno.
		Alumno por artefacto sanitario.
	Humanos	Distribución porcentual de los docentes por grupos quinquenales de edad. Total y para cada sexo.
		Máximo nivel de enseñanza alcanzado por los

[13] Esta *imagen* se extrae del Manual metodológico: Sistema Nacional de Indicadores Educativos. Laboratorio de Estadística. Red Federal de Información Educativa. Dirección Nacional de Información y Evaluación de la Calidad Educativa. Ministerio de Educación, Ciencia y Tecnología. Argentina. Pág. 13 y 14

[14] *Producto Bruto Geográfico (PBG):* Indicador macroeconómico que permite medir la suma de valores totales producidas por los bienes y servicios de las jurisdicciones en un año.

		docentes.
		Distribución porcentual de los docentes según antigüedad en la función. Total y para cada sexo.
		Porcentaje de docentes con formación pedagógica.
		Porcentaje de cargos Horas cátedra atendidos por suplentes
		Alumnos por docente.
	Financieros	Gasto total y per cápita en educación como porcentaje del PBI (PBG) total y per cápita.
		Gasto público en educación como porcentaje del gasto público.
		Gasto público corriente en educación como porcentaje del gasto en educación.
		Gasto público en educación como porcentaje del gasto en educación.
		Gasto público en educación por alumno, en establecimientos de sector de gasto estatal.
		Gasto privado en la educación como porcentaje del gasto en educación.
		Gasto público en educación por alumno, en establecimientos del sector de gestión estatal.
		Gasto público y privado en educación por alumno, en establecimientos de sector de gestión privado.
		Gasto público destinado a establecimientos privados (transferencias) como porcentaje del gasto público total en Educación.
	Organizacionales	Alumno por unidad educativa.
		Distribución porcentual de alumnos por sector de gestión.
		Porcentaje de secciones múltiples.

		Alumnos por sección.
		Distribución porcentual de los alumnos por modalidad/ orientación de nivel medio.
		Porcentaje de cargos docentes con función frente a alumnos.
		Porcentaje de cargos docentes / horas cátedras titulares y funciones frente a alumnos, del total de cargos docentes / horas cátedra con funciones frente a alumnos.
		Porcentaje de unidades educativas de personal único de nivel primario / EGB 12 del sector estatal.
		Cargos directivos con funciones frente a alumno (personal único) por cada 100 cargos de director / regente.
Proceso	Acceso	Tasa neta de escolarización por nivel de enseñanza.
		Tasa de escolarización por grupos de edad.
		Esperanza de vida escolar.
		Porcentaje de alumnos de 1er grado que han asistido al nivel inicial.
	Flujo	Tasa de promoción.
		Tasa de repetición.
		Tasa de reinscripción.
		Tasa de abandono interanual.
		Atraso escolar
Resultados	Egreso	Tasa de egreso
		Coeficiente de eficiencia.
		Duración media de la permanencia de los egresados.
		Duración media de la permanencia de los que abandonaron.

		Grado/año de estudio promedio de los no graduados.
	Logros Académicos	Porcentaje de aciertos correspondiente a las evaluaciones de Matemática y Lengua.
		Índice de comprensión lectora.
		Índice de resolución de problema
Impacto	Socio - Económico	Máximo nivel de enseñanza alcanzado por la población adulta. Total y PEA[15].
		Tasa de actividad refinada por nivel de enseñanza alcanzado.
		Tasa de ocupación, subocupación y desocupación por nivel de enseñanza alcanzado.
		Escolaridad media de población según quintiles de ingreso.

Estos indicadores, separados por dimensión y por área y con múltiples relaciones, se constituyen en criterios comunes a los de otros países, pero los indicadores no están estandarizados, sino que estos son particularmente los que representan al sistema educativo de Argentina, en tu totalidad y complejidad. Esto no quiere decir que no existan indicadores comunes en sistemas de países distintos, por ejemplo, así como con Guatemala tenemos indicadores muy diferentes, compartimos muchos que miden lo mismo y esto se da porque pertenecemos a la misma región, es decir, compartimos un contexto demográfico y socioeconómico similar, pero también se realizan acuerdos previos en los que definimos principios y objetivos similares.

Quizás esto no pasaría con países de otras regiones, por ejemplo, con Finlandia. Y esto es porque con Guatemala poseemos sistemas de indicadores que responden a bases de equidad, se comprende la equidad de la misma forma, los objetivos y principios establecidos son similares, en cambio Finlandia posee un sistema educativo diferente, con concepciones de equidad distinta en función de las posibilidades de acceso que ellos tienen.

Esto hace que se tenga una determinada selección y agrupamiento de cada sistema de indicadores de forma particular según el país, pero principalmente en la forma de entender la equidad y la eficiencia.

Muchas veces aquí radican las grandes resistencias o conflictos en el campo educativo, dado que la equidad y la eficiencia algunas veces parecen no poder ir a la

[15] Población Económicamente activa (PEA): Indicador macroeconómico que cuantifica la cantidad de personas que poseen un trabajo activamente o que lo están buscando.

par. Es en este sentido que se recomienda centrar la atención en las formas de ser medida y no en la dimensión valorativa.

Dicho esto, se puede medir la eficiencia en dos niveles. En primer lugar, se mide la eficiencia interna, mirando hacia adentro del sistema educativo, con el objetivo de poder captar la capacidad del sistema en tanto que los individuos accedan a la educación, analizar con qué nivel de fluidez se da el proceso de escolarización y finalmente si este proceso llega a la terminalidad del nivel educativo en cuestión. Es importante destacar que suele ser difícil definir los factores que producen la educación, por lo que, se utiliza un coeficiente definido como insumo-producto que pretende acercar lo máximo posible a la medida de eficiencia hacia adentro del sistema educativo.

En segundo lugar, está el abordaje de la eficiencia externa, que mide la capacidad de los individuos de ser productivos; y entendemos a una "persona productiva" en función de lo que se espera de la educación. Esto quiere decir que va a medir el impacto social de la educación ligada a la inserción laboral de los individuos que pasaron por el sistema educativo y la vida en sociedad en los valores de la democracia. Es improbable que en países complejos como Argentina se pueda medir la eficiencia externa del sistema educativo, en función de los niveles de inserción laboral o en la reducción de las desigualdades o el desarrollo económico, cuando sabemos que el contexto demográfico y socioeconómico es poco propicio. En este sentido es que, como analistas, hay que ser cuidadosos, la eficiencia no se debe medir con otros criterios, sino que se debe contemplar con los factores de contexto. Dado que la forma de medida de la eficiencia es una medida instrumental, comprendida en un contexto, en caso contrario podría establecer factores de mérito.

Analizar una política pública desde indicadores. Las Evaluaciones Estandarizadas

Como parte de la comunidad educativa, muchas veces y con mayor intensidad en los últimos años los resultados de las pruebas estandarizadas se comprenden como medida de la situación generalizada en la que se encuentra la educación en el país. En particular desde una perspectiva mediática.

Es así como en muchos casos se realizan especulaciones por profesionales de diversas áreas y con gran desconocimiento del sistema educativo, incluso del sistema nacional de indicadores educativos o de sus usos. De esta manera la calidad de la educación ingresa en discusión y la complejidad de los procesos de enseñanza y aprendizajes parecen reducidas a algunas cifras. Dando por resultado que, cuando los índices de calidad educativa no alcanzan, las responsabilidades recaen en los docentes, los alumnos y/o en los responsables políticos, anteponiendo factores de mérito en los argumentos analíticos que se exponen para la definición de el o los responsables.

Este apartado posee el objetivo de poner en tensión y brindar conocimiento sobre el modo en que las evaluaciones estandarizadas resultan un instrumento con gran potencial para la elaboración de políticas públicas y cómo se pueden utilizar como base de sistemas de monitoreo.

Es necesario en primera instancia comprender la raíz de esta discusión. ¿Qué entendemos por calidad educativa? En segunda instancia, analizar la influencia de las evaluaciones estandarizadas en la educación, y finalmente ¿qué es y para qué sirve el sistema de monitoreo para las políticas públicas, programas o actividades educativas?

En apartados anteriores hemos definido brevemente de qué se habla cuando se habla de calidad educativa, la intención ahora es poner esa definición en juego, no solo en términos teóricos sino en lo concreto y en su vínculo con los sistemas de indicadores. No vamos a detenernos a hacer una reconstrucción de la historia del término "calidad" porque no es el fin de este capítulo. Sin embargo, es importante destacar que en el marco de las reformas educativas de la década de 1990 la globalización, el avance tecnológico y con ello el aguaje de gestión empresarial para los servicios públicos, buscaban generar perfeccionamiento en las instituciones educativas.

En este marco, la calidad educativa se define en función de la eficiencia y la eficacia, características que corresponden a Sistemas educativos homogéneos, tales como los de Japón, Estados Unidos, China, Finlandia y algunos países de la Unión Europea. Quienes evalúan estos sistemas educativos serán los que imponen los estándares de calidad que posteriormente se reproducen por medio de los organismos internacionales. Esto llevó a un fuerte rechazo por la idea de Calidad Educativa.

En este marco histórico se ponen en discusión los rankings de sistemas educativos, que se realizaban en base a las Evaluaciones Estandarizadas. Los resultados de estas pruebas exponen fuertemente a sistemas educativos como el argentino y surgen los debates acerca de si es posible establecer medidas de calidad con criterios hegemónicos, como si todos los sistemas educativos tuvieran características homogéneas hacia adentro y hacia afuera. De estos debates surge la necesidad de establecer "competencias básicas" como medida de calidad. Es así como nacen las evaluaciones internacionales y los estándares tal como los conocemos, con una visión de equidad más fuerte, pero aun sosteniendo la importancia de garantizar una educación de calidad y eficiente, centrada en el interés en la educación y la evaluación como aquella que puede garantizar las competencias básicas para que cumpliera con el objetivo.

> "Al señalar la escuela básica como el instrumento más importante para la adquisición de esas competencias educacionales básicas, se abren unos caminos para diseñar los instrumentos que permitan evaluar los niveles de dominio de las diversas competencias como objetivos de aprendizaje y se inicia la realización de pruebas

estándares a nivel internacional de esas competencias, que en el primer momento tienen que ver con capacidades lógico-matemáticas y lecto escritura." (Mejía,2015;11)

En el área de impacto de nuestro sistema nacional de indicadores tenemos uno que mide justamente el logro académico como el *"Porcentaje de aciertos correspondiente a las evaluaciones de Matemática y Lengua"*.

Muchos sistemas educativos realizaron sus propias evaluaciones estandarizadas, no solo para medir el nivel de competencia básicas que se han alcanzado, sino para sostener un sistema de monitoreo.

Si bien podríamos continuar con los debates en torno a la polisemia de calidad educativa, lo que nos convoca es comprender de qué modo podemos medir la calidad educativa en el marco del sistema nacional de indicadores y cómo esto permite la construcción de políticas públicas.

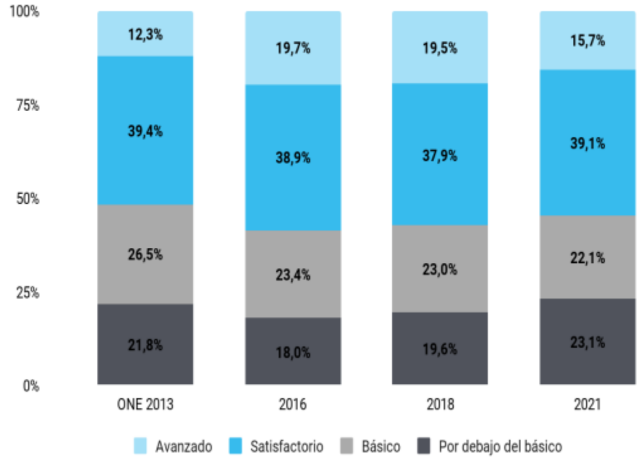
A. Morduchowicz (2018) entiende que un sistema de monitoreo cuenta con tres fases:

1) La primera fase es la *construcción de la base de datos,* en primera instancia y para lograr sistematizar la información, sin caer en un sistema de información que no nos permita analizarla. Es necesario incluir los indicadores que tengan relevancia para los criterios de evaluación que se pretende emplear. No toda la información dispone forma parte del análisis, si no que se seleccionan los datos que realmente nos sean significativos.

2) La segunda fase es el *proceso analítico*, donde los datos se transforman en información. Aunque suene casi mecánico, esta fase es la que más tiempo lleva a los especialistas, dado que requiere agregar, combinar y establecer índices que componen los datos. Esta fase es la que nos permite realizar inferencias sobre los potenciales resultados. Por ejemplo, si hacemos un breve análisis de los datos de las pruebas APRENDER 2021 de la muestra de estudiantes de 6to grado, podemos observar la siguiente información, que un grupo de especialistas se encargó de resumir y hacer fácil para su lectura por medio del siguiente gráfico [16]:

[16] Este gráfico se extrae del Informe Resultados Aprender 2021. pág. 12

Gráfico 7. Distribución de estudiantes por nivel de desempeño en Matemática. ONE 2013, Aprender 2016, 2018 y 2021.



Sabemos que estos datos no son suficientes para establecer conjeturas específicas o exactas, pero podemos observar ciertas tendencias sobre los resultados. Se observa que en la categoría "Por debajo del básico" el desempeño de matemática de los estudiantes de 6to grado es el más alto de los 8 años. Esto quiere decir que es mayor la diferencia con años anteriores de estudiantes que no han podido evidenciar el aprendizaje de los contenidos básicos.

Esto resulta preocupante e invita a pensar y realizar algunas hipótesis al respecto, por ejemplo, pensar si tiene impacto en estos resultados la pandemia.

La *comparación* es el último elemento de esta fase, en este caso como ejercicio didáctico, y se realiza entre las evaluaciones estandarizadas de los últimos ocho años. También podría ser entre países de la misma o de diferentes regiones, dependiendo que se desea comparar.

3) Se *toman decisiones* con relación a la información minuciosamente construida y analizada. La decisión inicial sería la de establecer dónde centramos la atención. Nos interesa que la mayor cantidad de estudiantes puedan aprobar con los conocimientos básicos, con una visión más equitativa, o hacemos foco en la excelencia y orientamos los recursos para que aumente el 15,7% de desempeño avanzado en las próximas pruebas APRENDER, quitando atención en los estudiantes que no llegaron al mínimo.

Esta toma de decisiones puede derivar en la construcción de políticas públicas, programas, actividades o proyectos educativos, que dependerá de las necesidades y objetivos que se hayan establecido. No obstante, no se puede minimizar el peso que tienen los medios de comunicación al respecto.

Al establecer cualquier proyecto de acción hay que tener además de los objetivos y la problemática a resolver, conocimiento sobre la organización, los recursos humanos, la capacitación de los recursos humanos, la infraestructura con la que se cuenta, el presupuesto y cómo se distribuye, entre otros factores determinantes para que dicho proyecto educativo o política funcionen. Información a la que, con un buen uso del sistema de indicadores, nos podemos aproximar.

Siguiendo el ejemplo de las pruebas APRENDER 2021 podemos decir que, al encontrarnos con los macros resultados en conocimientos básicos de matemática, se establecieron varias acciones para que este indicador en el año 2022 muestre una realidad diferente. Es así como la calidad educativa está íntimamente ligada con un indicador de impacto. Si bien aún no se encuentran los resultados de las pruebas APRENDER 2022, las acciones tales como cuadernillos para estudiantes, talleres y cuadernillos para docentes, donde se explica qué se espera para la resolución, en paralelo a un sistema de simulador web de pruebas estandarizadas donde puede acceder de manera libre toda la comunidad educativa para practicar y ver el tipo de conocimientos que se espera, pretenden generar un cambio en los resultados.

Es de esta manera que la media de calidad de impacto que podemos esperar no necesariamente será del 100% de los estudiantes de 6to grado del territorio argentino con resultados avanzados, si no que para 2022 los resultados de matemática en el nivel de desempeño “por debajo del básico” sea inferior a 23,2%. Aún no están disponibles los resultados de 2022, pero cuando obtengamos dichos resultados el sistema de monitoreo por medio de las evaluaciones estandarizadas establecerá nuevos esperables y se planteará nuevos objetivos para garantizar la calidad educativa, dentro de las bases de la igualdad, la equidad y la eficiencia.

Guía de Actividades N°10:

Indicadores, sistema de indicadores y sistema nacional de indicadores.

Objetivos de la actividad:

a) Comprender conceptualmente los núcleos centrales del capítulo.
b) Relacionar dichos núcleos conceptuales.
c) Utilizar el sistema de indicadores nacionales situaciones concretas diseñadas con fines pedagógicos.

Actividad 1: Se solicita la construcción de un esquema conceptual, ya sea de forma individual o en pequeños grupos donde puedan responder a las siguientes preguntas:

a) ¿Qué es un indicador? ¿Cuál es su finalidad?
b) ¿Cuáles son sus usos en el campo educativo?
c) ¿Para qué sirven los indicadores?
d) ¿Cómo se clasifican los indicadores?
e) ¿Qué es calidad y equidad educativa en el marco de los indicadores educativos?

Para la realización del diagrama pueden utilizar papel o pueden valerse de aplicaciones que faciliten su construcción, algunas que recomendamos son:

- xmind.works
- diagrams.net

Actividad 2: Explique brevemente cómo se relacionan las dimensiones del sistema nacional de indicadores.

a) ¿Por qué es un sistema?
b) ¿Con qué objetivo se realizan programas o políticas públicas siguiendo los criterios del sistema nacional de indicadores? ¿Qué se desea obtener?
c) ¿Podría describir cuáles son los criterios que el Estado argentino definió como importantes desde la lectura de los indicadores? Y mencionar ¿cuáles son los problemas para resolver?

Una pista: Este es un titular de diario del 4 de septiembre del 2019 “En **Córdoba**, sólo el 57% de los alumnos que empiezan primer año del secundario lo termina en los seis años previstos.”

Actividad 3: Con la construcción conceptual que se les ha propuesto, se les solicita hacer un análisis del “Programa PIT 14/17”. Debajo se les deja una breve reseña del programa:

> “El Programa de inclusión para la terminalidad de la educación secundaria y formación laboral para jóvenes de 14 a 17 años, posibilita a los adolescentes que hayan abandonado la escuela al menos un año antes de la puesta en marcha del programa (en el 2010) o que no la iniciaron, finalizar sus estudios a través de un servicio educativo, que, debido a su diseño más flexible, permite a cada estudiante construir su propio trayecto formativo.
>
> La propuesta –organizada como un bachillerato en ciencias sociales– reconoce los recorridos educativos previos –debidamente certificados– de los jóvenes y establece un sistema de promoción por asignaturas, en vez de por año, de manera tal de permitirles avanzar en su escolarización en función de un régimen de correlatividades.
>
> Los espacios curriculares se organizan en obligatorios –que reúnen los saberes básicos previstos para la educación secundaria, de duración anual o cuatrimestral– y complementarios, cuyo propósito es diversificar la oferta educativa.
>
> Asimismo, están previstas instancias de tutorías y de coordinación pedagógica para asesorar a los adolescentes en el diseño de su itinerario de cursado e introducirlos en el dominio de estrategias generales para la organización del tiempo y el manejo de los materiales de estudio.
>
> El programa se desarrollará, en principio, sólo por dos cohortes sucesivas (2010/2011), en 31 escuelas –20 en Capital y las restantes en ciudades del interior– que brindan actualmente educación secundaria y que, por su organización institucional y disposición, están en condiciones de llevar a cabo esta iniciativa.
>
> El cursado para el joven que carezca de trayectos escolares previos en el nivel se ha establecido que sea de 4 años, y de un tiempo menor –que dependerá de lo que pueda acreditar el alumno– para quienes sí los tengan.” (“Programa PIT 14/17 - Gobierno de Córdoba”)

a) ¿Qué dimensiones del sistema nacional de indicadores se tuvieron en cuenta para generar este programa?
b) ¿Qué indicadores nos permitirán observar los resultados de este programa?
c) Siguiendo los indicadores de impacto. ¿Qué conclusiones podríamos sacar sobre este programa?
d) Si este programa fue creado para solo dos cohortes (2010/2011), y han pasado 13 años de este. ¿Que sostiene que aun siga en vigencia?

e) ¿Que podríamos inferir que ha cambiado en los objetivos gubernamentales para que este programa hoy se encuentre con posibilidad de cierre o la inversión a este programa haya disminuido abruptamente? ¿Qué inferencias podrías sacar al respecto?

   i) Para él análisis es posible que se necesite recurrir a otros datos. Pueden hacer uso de información encontrada en el navegador, pero les dejo un informe que podría servirles también.

      1) https://rdu.unc.edu.ar/bitstream/handle/11086/18755/UNICEF__educacion_PIT_DIC_OKweb.pdf?sequence=1&isAllowed=y

Bibliografía:

Capuano, A. M. (2014) ¿Qué son los indicadores? Perspectivas y usos diferentes. Revista Argentina de Estadística Aplicada – RAEstA 1 - Año 1

Mejía J, M. R (2015) La calidad de la educación, una disputa polisémica por sus sentidos. En Debate sobre calidad educativa sobre grupos de Incidencia. Editorial: CEAAL. Serie: Miradas desde la educación popular. 7 - 24

Montoya, S. y Volman, V. (2014). ¿Cuáles son los determinantes de la calidad educativa en la Ciudad de Buenos Aires?: los índices de calidad y equidad porteños. Documento de trabajo No. 44 de la Escuela de Economía "Francisco Valsecchi". Facultad de Ciencias Económicas. Universidad Católica Argentina.

Morduchowicz, A. (2006). Los indicadores educativos y las dimensiones que los integran. Buenos Aires, diciembre. IIPE – UNESCO

Pascual, L (2016) Las estadísticas educativas y los desafíos futuros: un sistema de información por alumno. Revista Argentina de Estadística Aplicada – RAEstA 3 - Año 3

Sauvaget, C. (1999) Indicadores para la planificación de la educación: Guía Práctica. Paris, Francia. IIPE - UNESCO

Sistema Nacional de Indicadores Educativos. Laboratorio de Estadística. Red Federal de Información Educativa. Dirección Nacional de Información y Evaluación de la Calidad Educativa. Ministerio de Educación, Ciencia y Tecnología. Argentina.

Tiana Ferrer, A. (1997) Qué son y qué pretenden. Marzo - Nª 256 - Cuadernos de Pedagogía. 50 -53

Reflexiones finales

A lo largo del libro se han presentado las nociones de estadística que consideramos básicas y necesarias para la aplicación al campo de las Ciencias de la Educación. Tenemos la convicción de que la importancia que tienen estas nociones para comprender datos y obtener información del sistema estadístico nacional y de los sistemas estadísticos internacionales es fundamental para cualquier educador.

En la bibliografía es conocida la relevancia que tiene el análisis de datos estadísticos en diferentes dimensiones de la vida pública y en el diagnóstico de problemáticas propias del sistema educativo. Además, la información producida por el análisis de datos estadísticos ha mostrado una gran influencia en el diseño de políticas públicas en educación -del mismo modo que la tiene en otras áreas del Estado como la salud-, en sus diversas etapas como son la implementación, el seguimiento de dichas políticas y finalmente en la evaluación de los resultados obtenidos a partir de su aplicación.

Pero este no es el único dominio de la educación donde las estadísticas tienen algo que aportar, también tienen mucho para decir sobre el proceso educativo de la población, como se evidencia en el seguimiento de las cohortes de estudiantes, en los estudios sobre acceso, permanencia y egreso a determinado nivel del sistema educativo o en las investigaciones que toman un grupo particular de edad que se es de interés por motivos específicos, como pueden ser los niños de 3 a 5 años.

Las estadísticas educativas permiten, por otra parte, relaciones con otros sistemas estadísticos diferentes como los datos que surgen del estudio del empleo formal e informal y del mercado de trabajo. Un ejemplo podría ser el de la relación entre el sistema educativo obligatorio y la inserción en el mercado de trabajo formal e informal de los jóvenes entre 15 y 17 años.

Otro elemento de nodal importancia para el estudio de estadísticas aplicadas a las Ciencias de la Educación es la actualización de las bases, en función de las nuevas tecnologías de recolección y procesamiento de datos, hacia los relevamientos nominales o por estudiante. Estas bases toman como unidad de análisis a los estudiantes y permiten hacer un seguimiento de trayectoria individual por el sistema educativo. Este avance está relacionado con un gran cambio que estamos a punto de vivenciar en los sistemas estadísticos nacionales, que pasaran de capturar datos agregados a poder analizar trayectorias de individuales a lo largo del sistema educativo.

Sólo por mencionar tres aplicaciones más de las estadísticas a los sistemas educativos y a las Ciencias de la Educación, diremos que existen en la actualidad sistemas de medición internacionales del despeño de los estudiantes y de recursos de los establecimientos, que de un modo u otro buscan comparar los sistemas educativos nacionales entre sí. Esto habilita a establecer jerarquías entre los países y condenar o absolver, en series temporales que muestren la variación en estos indicadores, a determinados sistemas educativos. Está de más mencionar la influencia que tienen en la actualidad estos sistemas de medición internacionales en la definición de políticas públicas locales.

Otro aporte fundamental de la estadística, es el que brinda en el conocimiento de las condiciones estructurales en las que trabajan los docentes del país. Podemos calcular la relación de estudiantes por docente en los distintos sectores de gestión o para las escuelas de una zona en particular y conocer datos de la realidad de esos docentes: si trabajan en más de una institución, si trabajan muchas horas, el sueldo promedio, cuál es la distancia promedio respecto del lugar de trabajo. Estos ejemplos son sólo por mencionar algunas de las potencialidades de las estadísticas para este grupo.

Finalmente, pero a sabiendas que esta es una descripción acotada de la potencialidad de la estadística, están los beneficios que podemos obtener dentro de las instituciones educativas, comparando el desempeño de los alumnos en dos momentos de tiempo o de dos grupos distintos de alumnos en una asignatura a partir de la aplicación de una determinada estrategia pedagógica u otra. También puede servir para comparar el desempeño de un grupo que ha perdido más clases que otro.

Las aplicaciones parecen no más tener límites que los que les pongamos nosotros. Lo que sí debemos tener presente es que las lecturas y los análisis de los resultados estadísticos en educación deben ser realizadas por los especialistas en Ciencias de la Educación. Dejar en manos de otras disciplinas o de otros profesionales esta lectura es ceder un terreno que nos pertenece.

Bibliografía sugerida

En este apartado sugerimos algunos textos y monografías en los que podrán ahondar en los conceptos que hayan encontrado en este libro y que, por su especificidad, no han sido abordados en toda su complejidad.

Textos de estadística para estudiantes:

Clegg, Frances, (1984). Estadística fácil aplicada a las ciencias sociales. Editorial Crítica. Madrid.

Estadística y Educación

Ferguson, G. (1986). Análisis estadístico en educación y psicología. Editorial Anaya.

Bologna, E. (2013). Estadística para psicología y educación. Editorial Brujas. Buenos Aires, Argentina

Estadística y Sociología

Sierra Bravo, R. (1981). Ciencias sociales. Análisis estadístico y modelos matemáticos: teoría y ejercicios. Editorial Paraninfo.

Series Monográficas

Lorenzo, J. (2013). Nociones Básicas de Muestreo. Disponible en Aula Estadística: ANSENUZA – FFyH. URI: http://hdl.handle.net/11086.1/746

Lorenzo, J. (2019). Distribución Normal. Disponible en Aula Estadística: ANSENUZA – FFyH. URI: http://hdl.handle.net/11086.1/1344

Lorenzo, J. (2019). Inferencia Estadística. Disponible en Aula Estadística: ANSENUZA – FFyH. URI: http://hdl.handle.net/11086.1/1345

Lorenzo, J. (2019). Distribución de probabilidades. Disponible en Aula Estadística: ANSENUZA – FFyH. URI: http://hdl.handle.net/11086.1/1346