2013

INTRODUCCIÓN AL ANÁLISIS UNIVARIADO DE DATOS

Trabajo presentado a la escuela de Psicología de la Universidad Central de Venezuela como parte del programa de formación para el escalafón docente de asistente.

Dimas Sulbarán

22-9-2013

ÍNDICE

1. Elementos del análisis de Datos 3
1.1. Estadística y Método Científico. 4
1.2. Niveles de medición y Estadística. 7
1.2.1. Nominal. 8
1.2.2. Ordinal. 9
1.2.3. Intervalo. 9
1.2.4. Razón. 10
2. Medición y representación de datos 13
2.1. Organización y representación de datos numéricos 15
2.1.1. Construcción de bases de datos. 16
2.1.2. Tabulación de frecuencias. 17
2.1.2.1. Tabulación de frecuencias por clases. 19
2.1.3. Representación gráfica de datos 19
2.1.3.1. Representación de caracteres cualitativos 20
2.1.3.2. Representación de caracteres cuantitativos 20
Diagramas diferenciales. *20*
Diagramas integrales. *20*
Diagramas exploratorios. *20*
2.1. Análisis de datos cualitativos 22
2.1.1. Tablas de contingencia 23
Notación de tablas de contingencia. 24
Cálculo de porcentajes 25
2.2. Indicadores sumarios de posición, centralidad, variabilidad y forma de la distribución (asimetría y curtosis) 26
2.2.1. Cuantiles, percentiles y cuartiles. 27
2.2.2. Estadísticos de tendencia central. 29
2.2.2.1. Moda. 30
2.2.2.2. Mediana. 30
2.2.2.3. Media. 30
Propiedades de la media: *30*
2.2.3. Estadísticos de variabilidad. 31
Coeficiente de variación de Pearson 31

2.2.4. Estadísticos de forma de la distribución. 31
2.2.4.1. Asimetría 32
2.2.4.2. Curtosis. 33
2.3. La distribución normal 34
2.3.1. Estandarización de variables aleatorias normales 35
2.3.1.1. Ajuste de la Binomial a la normal: Teorema de Laplace - DeMoivre 35
2.3.1.2. Teorema de Tchebycheff 36
2.3.1.3. La distribución normal bivariada. 36
2.4. Transformaciones de Puntajes 37
2.4.1. Transformaciones lineales y no lineales. 38
2.4.1.1. Escalas directas y derivadas. 38
Bibliografía 41

1. Elementos del análisis de Datos

La naturaleza del dato es el objeto de reflexión fundamental del análisis estadístico de datos. La necesidad de procurar observaciones de calidad de variables estocásticas es la situación que históricamente vincula la estadística y el método científico. Esta necesidad por atender variables que superaban las posibilidades de los modelos matemáticos y las teorías de la medición clásica dio lugar al surgimiento de nuevas teorías de la medición, tales como la operacional y la representacional que permitieron traducir a valores numéricos casi cualquier objeto de la naturaleza. Esta ampliación de los alcances de la medición se vio fundamentada en el desarrollo de una teoría acerca de los niveles de medición, atribuible al Psicólogo norteamericano Stanley Stevens. La construcción de datos cuantitativos condujo al cuestionamiento con relación a las ventajas de asignar significados numéricos a nuestro objeto de estudio y como un caso particular en Psicología; lo cual, se ampararía en una argumentación pertinente con respecto a los supuestos de la medición en el estudio de la conducta. Como ciencia empírica, la Psicología cuenta con una impronta metodológica importante del positivismo y entre sus máximas la operacionalización de los constructos; lo cual, le ha permitido a esta rama de la ciencia alcanzar el estudio de los procesos mentales que no pueden ser medidos directamente, pero cuya existencia puede ser inferida de la conducta. Definimos estadísticamente el dato como la observación sistemática y rigurosa de las variables estocásticas. Finalmente, se expondrá una clasificación y notación de variables con base en los criterios estadísticos y metodológicos.

1.1. Estadística y Método Científico.

La relación entre método científico y estadística, está fundada en la reflexión epistemológica o, en otras palabras, en las disertaciones que emanan desde la filosofía del conocimiento y que nos inspiran a entender el mundo como un universo objetivo, susceptible de ser estudiado, descrito y explicado desde la cuantificación de sus elementos. (Pagano, 2011). El tratamiento analítico de la naturaleza se inserta en la tradición epistemológica iniciada con los pitagóricos y continuada durante el renacimiento y la modernidad con los aportes de pensadores de la talla de Kepler (1571-1630) y Galileo Galilei (1564-1642), Francis Bacon (1561-1626), René Descartes (1596-1650), Isaac Newton (1642-1727), Locke (1632-1704), Leibniz (1646-1716) y August Comte (1798 – 1857), Hans Reichenbach (1891-1953), Rudolph Carnap (1891-1970) y los padres de la *estadística moderna*, Karl Pearson (1857-1936), Ronald Fisher (1890-1962), Jerzey

Neyman (1894-1981), entre muchos más, quienes contribuyeron a desarrollar y establecer a las matemáticas como el sistema lógico por excelencia para traducir el comportamiento de la naturaleza. Este supuesto fundamental, se materializa en el método científico fundamental para la tradición empírico-positivista, a saber: la observación de los hechos que son susceptibles de ser cuantificados. De este modo, la realidad hace referencia y tiene sentido para el científico cuando hace referencia a elementos que pueden ser medibles y la medición permite identificar las leyes que rigen el comportamiento de todas las cosas en la naturaleza.

Por extensión, con el surgimiento del interés por el estudio de los fenómenos sociales, en este contexto paradigmático dominante, autores como Emile Durkheim favorecieron que el estudio de lo social se asociara con el uso de la estadística y el método hipotético deductivo heredado de las ciencias naturales. El estudio de Durkheim de la *tasa de suicidio anual* dio lugar a su obra cumbre: el suicidio. Por su parte, fueron los trabajos de Quetelet los que contribuyeron a la construcción del hombre medio. Mientras que Friedrich Gauss nos hablaba de una variabilidad en las medidas del comportamiento de los cuerpos celestes, cuya solución fue el reconocimiento de un error que tendía a comportarse de manera normal y establecía la *ley de los mínimos cuadrados*, concepto fundamental para el estudio de las variables estocásticas. Por su parte, Karl Pearson desde sus estudios biométricos apoyaba a Francys Galton en la declaración de las leyes de la Psicología diferencial y la regresión estadística. (Sáiz & Sáiz, 2009).

Las críticas al positivismo derivadas de la crítica a las teorías de la física clásica, lideradas por las posturas indeterministas de autores como Heissenberg (1927) socavaron las bases de la precisión matemática y dieron lugar a la necesidad de mecanismos que permitieran lidiar con la indeterminación, en otros términos, con las variables estocásticas caracterizadas por poseer una varianza de error inherente. Con el establecimiento del positivismo lógico y el circulo de Viena, la estadística, en especial la estadística inferencial con su impronta teórica matemática, se alzará como el recurso más importante para construir modelos que permitan representar la naturaleza y "reducir los niveles de incertidumbre" (Rivadulla, 1991), en un mundo signado por la aleatoriedad y la capacidad finita del hombre para captar las medidas de la naturaleza y por consecuencia sus relaciones.

Esta necesidad del hombre de ciencia por dominar el azar, dio origen a la popularización de conceptos estadísticos que marcaron un importante impacto en el quehacer científico del siglo XX, tales como distribución de probabilidad (Bernoulli, Laplace, De Moivre, Gauss, entre otros),

mínimos cuadrados (Gauss, Karl Pearson) y contraste de hipótesis (Fisher y Neyman-Pearson). Entre estos debemos señalar el avance en los estudios correlaciónales y el diseño de experimentos. Así como los distintos modelos de la inferencia estadística, teorías de la probabilidad como el enfoque clásico, frecuentista y la Bayesiana. De este modo, gran parte de la producción científica del S. XX y XXI, especialmente en ciencias sociales tales como la Biología, Psicología, Sociología y Economía, se fundamenta en un riguroso tratamiento estadístico de los datos, inspirado en los trabajos de autores como Sir Ronald Fisher, Gosset (Student), Jerzey Neyman, Karl y Egon Pearson, Campbell y Stanley, entre otros.

A la par que la estadística se alzaba como alternativa al método científico positivista, las críticas al positivismo lógico no fueron tangenciales a la teoría estadística. En este sentido, cabe destacar las críticas a las teorías de la medición, así como a la inferencia estadística. Con relación a esto último no podemos descartar los trabajos de Karl Popper, quien en su obra de (1962), *La lógica de la investigación científica*, se establece como el máximo exponente del positivismo crítico y expone sus ataques al verificacionismo por acumulación, lo cual rescata la tesis del escepticismo de David Hume. A partir de este punto, el establecimiento de teorías por acumulación de casos pierde sustento lógico. Para Popper, el método de la ciencia debe ser crítico de aquel que permite confrontar las teorías existentes a evidencias que las contradigan. De modo que un conocimiento será válido en la medida que logre superar a las pruebas de falsación.

Vemos entonces como la relación entre método científico y estadística se cristaliza en la práctica científica cotidiana, de una comunidad de investigadores de orientación epistemológica cuantitativa. En primer lugar, el trabajo se ha de enmarcar en teorías de orden positivista, entendidas como aquellas que conciben la naturaleza como un orden lógico, sujeto a leyes que permiten construir hipótesis descriptivas, relacionales o causales-explicativas. En segundo lugar, supone que esta realidad está sometida a leyes naturales que la tornan objetiva, empírica, variable, controlable y susceptible de ser cuantificada; es decir, operacionalizada, por tanto, medible. En tercer lugar, se busca la aprehensión de la variabilidad en la forma de mediciones, lo cual favorece que la estadística se concrete en un conjunto de operaciones aritméticas que permiten determinar estadígrafos, a partir de una muestra de observaciones, para la descripción de las distribuciones de las variables e inferir el comportamiento de poblaciones. Finalmente, se usan los insumos anteriores para tomar decisiones con relación a las hipótesis planteadas y, por consecuente, con relación a las teorías.

Una lectura interesante que cabe agregar en este punto es compartida con Sánchez (2001), quien aporta ideas interesantes acerca de la trascendencia metodológica de la estadística e introduce el carácter histórico y político de este recurso retórico, ya que como el autor menciona en la presentación de su obra: se muestra el papel que tiene la estadística a la hora de naturalizar el orden social, como paso previo a su aceptación como orden político. (p. 33). Nos basta como ejemplo el caso de las contiendas de encuestadoras durante el proceso electoral para la elección del presidente de la república durante el año 2012. Como apoyo a la ciencia, la estadística también se ha establecido como un importante recurso retórico en la constitución de modelos políticos, así la historia da cuenta de los estudios en eugenesia que se derivaron de los laboratorios de Galton y Pearson. Pero también a las importantes contribuciones a las demandas de numerosos grupos sociales, ecológicos, feministas, laborales, de los derechos humanos, etc.

Finalmente, comparto la tesis de autores como (Brown & Ghiselli, 1969), apoyan la idea de que con la creciente aplicación de la metodología científica al estudio de la conducta se generó una creciente demanda de técnicas que permitieran el estudio cuantitativo y, por tanto, matemático de los datos psicológicos. Aunque advierten que, las variables psicológicas tienen propiedades únicas que hacen peligroso utilizar a ciegas los procedimientos analíticos de la matemática y la estadística. No obstante, es fácil reconocer los importantes aportes que la estadística ha ofrecido al desarrollo de la psicología como ciencia.

1.2. Escalas de medición y Estadística.

También conocido como escalas de medición, este concepto atiende a una de las teorías más influyentes de la medición en ciencias sociales, iniciada por los trabajos de Stanley S. Stevens, Psicólogo de la Universidad de Harvard, quien en el año de 1946 publicara en la revista Science, el clásico artículo titulado: “On the theory of scales and measurement”, en el cual describe las reglas, propiedades numéricas y operaciones estadísticas aplicables para los tipos de escalas. En resumen, el autor propone cuatro tipos de escalas, a saber: nominal, ordinal, intervalo y razón. Estos tipos de escalas son definidos por: a) su relación con las propiedades de los números reales, y b) cómo son afectadas por las transformaciones (operaciones matemáticas). En resumen: las relaciones de las propiedades entre los cuatro tipos de escalas pueden ser demostradas como:

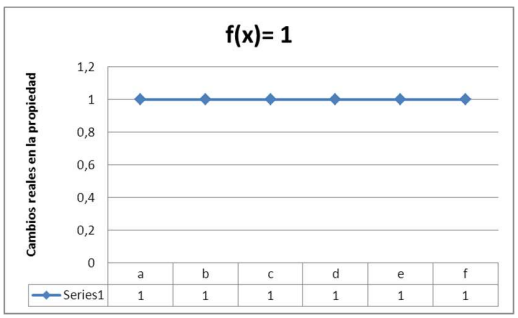
$$\left.\begin{array}{l} f(X_1) = f(X_2) \text{ iff } X_1 = X_2 \;\} \text{nominal} \\ f(X_1) > f(X_2) \text{ iff } X_1 > X_2 \end{array}\right\} \text{ordinal}$$
$$\left.\begin{array}{l} \text{ordinal} \\ f(X_1) - f(X_2) = b\left[g(X_1) - g(X_2)\right] = c\left(X_1 - X_2\right) \end{array}\right\} \text{interval}$$
$$\left.\begin{array}{l} \text{interval} \\ f(X_i) = 0 \text{ for all } X_i = 0 \end{array}\right\} \text{ratio}$$

Cada tipo de escala sucesiva incorpora las características de las anteriores. Veamos el desarrollo de dichas propuestas en términos de sus posibilidades de representación gráfica. Para ello, se partirá de un sistema de coordenadas cartesiano, donde el eje de las abscisas representará los cambios en el sistema de medición propuesto. Por su parte, el eje de las ordenadas estará asociado a los cambios observables o presumibles en la propiedad en cuestión.

Un asunto importante en la teoría de la estimación estadística es el asunto del ajuste de las escalas de medición al estadístico empleado (Hand, 1996).

1.2.1. Nominal.

Piense en el caso de una variable de naturaleza meramente cualitativa; nominal en los términos de Stevens (1946). Supongamos que se trata de un registro de la variabilidad asociada al tipo de bebida consumida por las personas en un local de comidas rápidas cualquiera. El investigador encontrará fácilmente comprensible que la atribución de valores numéricos a las propiedades de la variable en cuestión es, particularmente, arbitraria. En términos estrictamente empíricos, la propiedad asociada a la variable en estudio se remite a dos opciones: presencia / ausencia, para cualquier valor de x.



Entre los estadísticos permisibles se cuentan los análisis de frecuencias y derivados como la moda y los análisis de contingencia, a nivel inferencial es posible trabajar con pruebas para el

contraste de porciones como la binomial y la multinomial; así como de independencia asociado a la chí cuadrado.

1.2.2. Ordinal.

En el caso de las variables ordinales, se puede sostener que los valores asumidos por el sistema de medición considerado permiten capturar en forma fundamental la progresión u orden entre los valores en consonancia con el *principio de transitividad*. Por este motivo algunos autores coinciden en aseverar que el mínimo nivel de medición que puede asumir una variable para ser calificada dentro del esquema propiamente cuantitativo es el de ordinal. Esta variable se define analíticamente como $f(x_i) = a + bx_i \pm e$, donde en esencia se trata de una función lineal con una medida de intervalo que se ve alterada por la presencia de un menor nivel de control para el error que en las escalas reales de intervalo. Bajo esta premisa se plantea la siguiente teoría para el tratamiento de variables ordinales como escalas de intervalo.

Además de los análisis de frecuencias, las variables con nivel de medición ordinal admiten el cálculo de estadísticos de rango, tales como percentiles; por tanto, hace viable el estudio de posición y tendencia central como la mediana y de dispersión como el rango total e intercuartílico. A nivel correlacional se pueden calcular los coeficientes de correlación tau de kendall o spearman para rangos ordenados. A nivel del contraste de hipótesis, según el diseño empleado, son admisibles el uso de pruebas no paramétricas como la U de Mann-Whitney, la Wilconxon y la prueba de rangos ordenados de Kruskal-Wallis.

1.2.3. Intervalo.

Las variables con nivel de medida de intervalo suponen la posibilidad de identificar la diferencia entre dos puntos a lo largo del mismo continuo de la variable. Esto básicamente por el hecho de que la medición favorece un mayor control del error correspondiente a la función decategorización. Las operaciones posibles son todas las de escalas anteriores, más las operaciones aritméticas de suma y resta.

Este tipo de medida tiene como valor agregado, con relación a las anteriores, que las diferencias entre dos valores continuos del objeto representan cambios equivalentes en el atributo.

Estas variables nombran, ordenan y presentan igualdad de magnitud. Por lo tanto, operaciones tales como la adición y la sustracción tienen significado. Sin embargo, el cero es un valor meramente referencial y no representa ausencia de la propiedad medida, dado que el rango de la función de categorización para los valores de x=0 es distinto a cero, esto es $f(x) \neq 0, \forall\, x = 0$. La función de categorización que describe este tipo de escalas se define como: $f(x) = a + bx \pm e, \forall\, x \in \mathbb{R}$

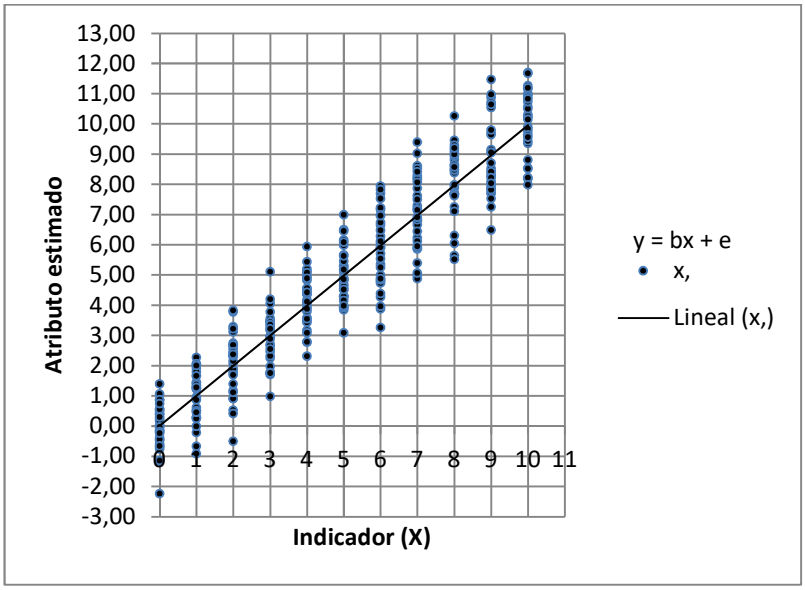
Donde *a* representa el punto de origen para el rango de la función de categorización que corresponde a un punto de la magnitud para la propiedad medida que es diferente de cero.

Por su parte, b representa la pendiente de la función de regresión que da cuenta de la relación entre los valores del objeto y la propiedad o atributo medida. Sin embargo, la medición de intervalo mantiene cuestionamientos con respecto al isomorfismo o simetria (Luce, 1997).

Además de las anteriores, vinculadas con las medidas orinales y nominales, las medidas en escala de intervalo admiten el uso de estadísticos de tendencia central como la media, aunque en términos estrictos sería válido cuestionar el uso del promedio aritmético con medida de intervalo; así como la desviación típica para la descripción de la variabilidad y de correlación como los estadísticos de rangos ordenados de Spearman y la producto-momento de Pearson. A nivel inferencial, según el diseño pertinente, admite estadísticos para el contraste de hipótesis como la prueba t de Student para el contraste de medias y la prueba de ANOVA para el análisis de varianza.

1.2.4. Razón.

Las variables con nivel de medida de razón suponen, además de la posibilidad de identificar la diferencia entre dos puntos a lo largo del mismo continuo de la variable, identificar el punto de origen para las magnitudes del fenómeno. De modo que la función de categorización en este tipo de variables se define como: $f(x) = bx, \forall\, x \in \mathbb{R}$. Esto básicamente por el hecho de que en esta situación la medición favorece el más alto control del error correspondiente a la función de categorización. Matemáticamente se puede comprobar que las variables con nivel de medición de razón admiten legítimamente todas las operaciones posibles asociadas a las escalas anteriores, más las operaciones de multiplicación y división.



Como se puede ilustrar en la gráfica anterior, este tipo de medida tiene como valor agregado, con relación a las anteriores, que las diferencias entre dos valores continuos del objeto representan cambios *proporcionalmente equivalentes* en el atributo. Estas variables nombran, ordenan y presentan igualdad de cambios en la magnitud con respecto al punto de origen de la propiedad. Por lo tanto, operaciones tales como la adición y la sustracción, además de la multiplicación y división tienen significado real. (Hand, 1996). En este punto, el cero deja de ser un valor meramente referencial y representa la ausencia real de la propiedad medida, dado que el rango de la función de categorización para los valores de x=0 es igual a cero, esto es $f(x) = 0, \forall\, x = 0$. La función de categorización que describe este tipo de escalas se define como: $f(x) = bx \pm e, \forall\, x \in \mathbb{R}$

Al igual que en los casos anteriores, *b* representa la pendiente de la función de regresión que da cuenta de la relación entre los valores del objeto y la propiedad o atributo medida. No obstante, a diferencia de todas las escalas anteriores, la medición de goza del apoyo irrestricto de la comunidad científica con respecto al isomorfismo o simetría (Luce, 1997).

En términos estrictos, sólo las variables con nivel de medición de razón admiten los análisis de varianza. La razón es muy sencilla, sólo las variables con nivel de medida de razón tienen una referencia clara con relación al punto de origen para la medición de las magnitudes de los atributos en cuestión. Al respecto, autores como (Brown & Ghiselli, 1969) aseveran lo siguiente: no se

justifica que comparemos dos objetos o procesos si no los medimos a partir del mismo punto inicial. (p. 139).

Además de las anteriores, vinculadas con las medidas nominales, ordinales y de intervalo las medidas en escala de razón admiten el uso de estadísticos de tendencia central como la media; la cual, en términos estrictos ya no es posible cuestionar como con el caso de las medidas de intervalo; la desviación típica para la descripción de la variabilidad y de correlación como el estadístico producto-momento de Pearson. A nivel inferencial, según el diseño pertinente, admite los mismos estadísticos para el contraste de hipótesis con medidas de intervalo, tales como la prueba t de Student para el contraste de medias y la prueba de ANOVA para el análisis de varianza.

En su exposición, Stevens (1946) emparejó ciertos métodos estadísticos a los tipos de escalas; punto que ha sido ampliamente elaborado como se puede apreciar en el trabajo de Michell (1986), “Measurement scales and statistics: a clash of paradigms”. Citando a este autor, salvando los elaborados argumentos técnicos, el concepto básico es que “la interpretación de los análisis estadísticos no puede superar la complejidad de los datos”. A decir de Lord (1953), en su célebre trabajo: “On the statistical treatment of football numbers”, los tipos de escalas no prohíben llevar a cabo ningún tipo de análisis estadístico particular, porque; a fin de cuentas, números son números, pero el tipo de escala es importante para la interpretación de los resultados de estos análisis. Lo que resulte podrá ser definido como datos con o sin sentido y esa es la verdadera diferencia.

2. Medición y representación de datos

La estadística es fundamentalmente el estudio científico de la variabilidad; la misma que, aunque no exclusivamente, ha estado muy vinculada con los complejos fenómenos sociales. Por tanto, se establece como un cuerpo de conocimientos teórico-prácticos con el fin de sistematizar la recolección, procesamiento, análisis y representación de grandes volúmenes de datos, que dan cuenta de las características de una determinada población. Teóricamente, la estadística tiene su asiento en la matemática. En la práctica se orienta por el método científico y se utiliza en ramas tan diversas como la ingeniería, la salud y las ciencias sociales. Así, la estadística se presenta con un alcance tan amplio como sus propios orígenes en estudios astronómicos, biológicos, sociológicos, económicos y psicológicos, entre otros.

El estudio de la variabilidad ha inspirado la construcción de dos grandes ramas de la estadística, a saber: descriptiva e inferencial. Este apartado se dedica, fundamentalmente, al estudio de los conceptos con relación a la *estadística descriptiva*. Específicamente, revisaremos los conceptos de organización y representación de datos (cuantitativos o numéricos), indicadores de resumen para posición, centralidad, variabilidad y forma de la distribución. Para finalizar con la distribución normal y la transformación de puntajes.

La estadística moderna y, en especial, la estadística descriptiva en ciencias sociales debe sus orígenes a los importantes aportes de investigadores como Adolphe Quetelet (1796-1874), quien a principios del siglos XIX contribuyó a la construcción de métodos estadísticos para realizar observaciones con el fin de determinar empíricamente el comportamiento de los fenómenos sociales. En su trabajo en estadística, como en ciencias naturales, Quetelet hizo gran hincapié en la necesidad de uniformidad en los métodos de recolección, tabulación y representación de datos. Para principios de los años 30 el trabajo de Quetelet se volcó de forma específica a los asuntos relativos a los seres humanos. Los trabajos de este autor en temas como el crímen y características antropométricas, como el peso y la altura, le permitieron construir el gran aporte de la obra de Quetelet como lo fue la propuesta del hombre medio o “homme moyen” (Landau & Lazarsfeld, 1978).

Después de los trabajos inspiradores de Quetelet, muchos han sido los autores que han contribuido a enriquecer la construcción de métodos estandarizados que definen el discurso estadístico. De esta manera, hoy en día podemos contar con una amplia oferta de recursos para la recolección, tabulación y representación de datos cuantitativos para el estudio de los fenómenos sociales. Desde las tablas y gráficas de frecuencias, hasta estadísticos de resumen que mantienen

el uso del valor medio como uno de los estimadores más apreciados para la descripción de las características poblacionales.

Hasta ahora nos hemos concentrado en aquellos elementos que permiten describir las características, estrictamente, de los datos observados. Una de las distribuciones teóricas más populares de la estadística hace su incursión con los trabajos de Gauss-Laplace y se conoce con el nombre de distribución normal o campana de Gauss. La influencia de estos autores en el trabajo de Quetelet lo convirtieron en el primer investigador en usar la curva normal como una forma de la ley de los errores y proponer su uso para determinar una medida promedio para una serie de fenómenos idénticos. (Landau & Lazarsfeld, 1978). El uso de la curva normal con fines descriptivos será el objeto de este primer acercamiento.

Es importante recordar que toda descripción en la ciencia es fundamentalmente un proceso de comparación. Sin embargo, para comparar es necesario que se cumplan algunas condiciones fundamentales como la conmensurabilidad de los objetos. Para cumplir con este principio y favorecer las comparaciones de nuestras observaciones, a nivel intra e intergrupal el investigador apelará, no en pocas ocasiones a la transformación de las escalas de sus datos. Aunque son múltiples las operaciones concretas de transformación que existen para llevar nuestras escalas observadas a otras escalas de interés, estas se reducen a dos categorías fundamentales: lineales y no lineales. Los detalles al respecto se exponen en el punto pertinente.

Debido a las limitaciones de esta exposición, no se espera que esta sea una disertación extensa con relación a los puntos en cuestión. Sin embargo, se hará el esfuerzo porque al final de este capítulo el lector tenga una impresión clara, a nivel práctico e intuitivo, con relación a los recursos claves del análisis descriptivo de datos, esto es, las posibilidades para la recolección, tabulación, ordenamiento, análisis y representación de grandes volúmenes de datos, con el fin de representar, en términos matemáticos, las características de una determinada población.

2.1. Organización y representación de datos numéricos

Por organización y representación de datos numéricos se entiende el conjunto de estrategias dirigidas al registro, ordenamiento y representación de las observaciones llevadas a cabo en la investigación. Superado el asunto de la medición, el investigador deberá lidiar con una cantidad más o menos importante de observaciones. Sin embargo, este dato es sólo información en potencia

en tanto el investigador no sea capaz de procesarlo efectivamente con el fin de construir conocimiento con ellos.

El procesamiento de los datos comienza con una efectiva estrategia de almacenamiento. El almacenamiento se fundamenta en la construcción de bases de datos. Asuntos como la codificación y la discriminación entre las distintas variantes de bases de datos son la clave de la tabulación. El investigador deberá distinguir entre las bases de datos primarias, referidas a las observaciones en bruto y las bases de datos secundarias basadas en un procesamiento previo de los datos.

Tabulados los datos, el investigador buscará la forma de develar la información oculta en el caos de los datos almacenados en bruto. Una forma preparatoria de procesamiento y análisis de los datos consiste en ordenarlos. Aunque básica, el ordenamiento de los datos le permite al investigador identificar los valores para estadísticos como el rango total y generar una primera impresión con relación a los límites de la distribución.

No obstante, muchas preguntas quedan abiertas tras el ordenamiento de los datos. Preguntas del tipo ¿cuáles son los valores que más se repiten?, ¿existen valores atípicos?, ¿cuáles son los vacíos en la distribución?, entre otras con respecto a la tendencia central, la variabilidad y el sesgo. Este tipo de preguntas comienzan a encontrar referencias claras para responderlas con los análisis de las frecuencias. Tal como veremos más adelante, los análisis de frecuencia nos permiten responder con precisión las preguntas con relación a cuantas veces aparecen los distintos valores observados en la data.

Los análisis de frecuencias representan una de las formas más ricas de obtener información de los datos con relación a nuestros intereses de investigación. No obstante, las tablas de frecuencias pueden resultar un poco complicadas de manejar para algunas personas que prefieren moverse por un discurso icónico. Para este tipo de personas que confía en aquél dicho: más dice una imagen que mil palabras, la representación gráfica de las dsitribuciones de los datos resulta un recurso de inmenso valor. Sin embargo, a diferencia de las tablas de frecuencia, la representación gráfica encierra una serie de compromisos con relación a la pertinencia que hacen clave la necesidad de distinguir, efectivamente, la aplicabilidad de uno u otro modelo según criterios como el nivel de medición de las variables y el objetivo del análisis: univariado o bivariado.

2.1.1. Construcción de bases de datos.

El término “base de datos” se usa para indicar que la matriz (una organización de datos en forma de filas y columnas) con la que estamos tratando organiza los datos provenientes de la realidad, de determinados aspectos medidos en la realidad, y no se trata de números generados al azar. Es decir, toda base de datos es una matriz, pero no toda matriz es una base de datos.

Presentaremos en este punto algunos tipos de bases de datos, clasificados según la naturaleza de los datos que contienen. Este apartado es apenas introductorio, y no pretende ser exhaustivo.

Siguiendo el esquema propuesto por (Grande & Abascal, 1989), debemos diferenciar las bases de datos en función de criterios como: a) el nivel de procesamiento de los datos, b) la naturaleza de los datos (cuantitativa, cualitativa y mixta) y c) el nivel de complejidad de la respuesta (simple, múltiple o mixta).

Nuestra clasificación comienza por distinguir entre *bases de datos primarias* y *bases de datos secundarias*. Las “Bases de Datos Primarias” consisten en la transcripción y ordenación de los datos, sin ningún procesamiento ulterior y las “Bases de Datos Secundarias” (conocidas generalmente como “tablas”), constituyen un procesamiento efectuado sobre las tablas de datos primarias.

De este primer grupo, bases de datos primarias, debemos diferenciar entre: a) bases de datos para datos cuantitativos (variables métricas y de series temporales), b) bases de datos para datos cualitativos (variables ordinales y de preferencias, de modalidades, binarias o de disyuntiva completa) y c) mixtas.

Por su parte, las bases de datos secundarias se generan luego del procesamiento de bases de datos primarias. Son llamadas “tablas” y, aunque pueden considerarse como “resultados” o “análisis” de información que ya fue transcrita en forma de bases de datos primarias, en ocasiones representan insumos para nuevos análisis estadísticos. Consideraremos dos: las Tablas de Contingencia y las Tablas de Proximidades y Distancias. Finalmente, el nivel de complejidad de la respuesta: a) simple, b) múltiple y c) múltiple.

2.1.2. Tabulación de frecuencias.

A nivel descriptivo, una de las prácticas heurísticamente más ricas corresponde al análisis de la estructura de los datos; los cuales, se clasifican de acuerdo con las mediciones realizadas y se ordenan, anotando sus resultados en una tabla.

El registro, ordenamiento y agrupación de los datos es lo que llamaremos tabulación de frecuencias.

A continuación se presentarán los conceptos, tipos y estrategias fundamentales para la construcción de tablas de frecuencias.

Dado un conjunto de observaciones para una variable (x) cualquiera, se llama *frecuencia* a la cantidad de veces (n) que se repite un determinado valor de la variable (x_i). En términos analíticos, se traduce en la ecuación: $f_{(x_i)} = \sum_{i=1}^{I} n_{xi}$

En términos generales, se pueden clasificar las distribuciones de frecuencias en función de dos criterios: a) la escala (absoluta y relativa) y b) según el nivel de acumulación de los datos (acumuladas y no acumuladas). Como se trata de categorías independientes, se pueden dar cualquiera de las condiciones de sus combinaciones, como se muestra en la tabla siguiente:

		Nivel de acumulación	
		Acumulada	No acumulada
Escala	Absoluta	Absolutas acumuladas	Absolutas No acumuladas
	Relativa	Relativas acumuladas	Relativas No acumuladas

La *frecuencia absoluta* f_i, de un determinado valor de X (x_i), corresponde con el número de veces n que aparece un determinado valor (n_{xi}) en un conjunto de observaciones N. se corresponde con la forma más básica del cálculo de frecuencias, de modo que: $f_{(x_i)} = \sum_{i=xi}^{k} n_{xi}$

Se llama *frecuencia relativa* h_i, de la modalidad x_i, al cociente de dividir el número de veces que aparece un determinado valor n_{xi}, entre el número total de observaciones N. en su forma fundamental, se trata de una porción, de modo que, su cálculo se resume a la siguiente ecuación: $p_{(x_i)} = \frac{\sum_{i=xi}^{k} n_{xi}}{N}$, esto da lugar a la forma más popular de la frecuencia relativa, es decir, el porcentaje (P_{x_i}), para su cálculo sólo se requiere multiplicar la porción o frecuencia relativa simple por cien (100). De modo que: $P_{x_i} = \left(p_{(x_i)} * 100\right) = \left(\frac{\sum_{i=xi}^{k} n_{xi}}{N}\right) * 100$

La *frecuencia acumulada* es la suma de las frecuencias (absolutas o relativas) de todos los valores inferiores o iguales al valor considerado. Se representa por Fa. De modo que, su cálculo se resume a la siguiente ecuación: $Fa_i = \sum_{i=1}^{k} n_i$

2.1.2.1. Tabulación de frecuencias por clases

En ocasiones la tabulación de los datos por intervalos simples, es decir, creando una categoría por cada uno de los valores observados, resulta en una forma poco económica para su análisis por lo que el investigador buscará alternativas parsimoniosas que permitan preservar la información de los datos en una versión resumida de los mimos. A esta estrategia se le conoce como *tabulación de frecuencias por clases* y se resume en las siguientes tareas:

1) Determinar la amplitud total del rango de la distribución. Se calcula como: Rgo= (max-min)+1
2) Definir el intervalo de agrupamiento de las clases. Aunque algunos autores han propuesto algunas alternativas matemáticas a este problema, en la práctica es recurrente el uso de un recurso subjetivo que permita construir, de forma pertinente, un número entre 5 y 15 intervalos.
3) Determinar los límites de los intervalos de clases.
4) Efectuar la tabulación con los datos agrupados por clases.

2.1.3. Representación gráfica de datos

Si bien, una pertinente tabulación de los datos permite comunicar eficientemente la información con relación al comportamiento de la distribución de una variable observada cualquiera, reza el dicho: una imagen vale más que mil palabras. Un despliegue gráfico es una de las herramientas más valiosa de la estadística con la que cuenta el investigador para comunicar las tendencias en términos de centralidad y variabilidad de los datos. Existen un número cada día mayor de alternativas gráficas para representar el comportamiento de los datos. En este punto no haremos una exposición exhaustiva de estos, en su lugar, nos limitaremos a presentar los paradigmas gráficos más emblemáticos en la escena.

Metodológicamente, la clasificación de los gráficos responde fundamentalmente a los siguientes criterios: a) nivel de medición de la variable y b) el número de variables implicadas.

2.1.3.1. Representación de caracteres cualitativos

Diagrama de barras.
Diagrama de sectores o (gráfico de torta).
Perfiles.

2.1.3.2. Representación de caracteres cuantitativos

Diagramas diferenciales.

- Histogramas.
- Polígono de frecuencias simples.
- Curva de frecuencias.

Diagramas integrales.

- Diagrama de frecuencias acumuladas.
- Polígono de frecuencias acumuladas.
- Curva acumulativa de frecuencias u ojiva.

Diagramas exploratorios.

El análisis exploratorio es una de las herencias más significativas de los trabajos de Tukey (1977 c.p. Vargas Sabadías, 1995), quien propuso un complejo sistema de análisis con el fin de atender a la descripción de la estructura de los datos en términos de:

- Localizar los valores de tendencia central.
- Conocer la dispersión con respecto a los valores centrales.
- Obtener una visión panorámica de los niveles de sesgo y curtosis.
- Descubrir lagunas (vacíos) en la distribución de los datos.
- Detectar posibles anomalías o valores atípicos.

- Encontrar valores de uso frecuente.

Diagrama de tallo y hojas.

El diagrama "tallo y hojas" permite ofrecer de forma simultánea una distribución de frecuencias de la variable y su representación gráfica. Para construirlo basta con separar en cada dato el último dígito de la derecha (que constituye la hoja) del bloque de cifras restantes (que formará el tallo).

Diagramas de caja y bigotes.

Los diagramas de caja corresponden a un modelo de gráficos, basado fundamentalmente en la representación de los cuartiles (Tukey, 1977 c.p. Vargas Sabadías, 1995, págs. 133-135), mediante el cual se visualiza la distribución de un conjunto de datos. Está compuesto por un rectángulo, la "caja", y dos brazos, los "bigotes".

Para dibujar los bigotes (líneas que se extienden desde la caja hasta los extremos), hay que calcular los límites: inferior (Li) y superior (Ls), que se identifiquen con los valores típicos. Para ello se calculan los puntos críticos a partir de los cuales los valores se consideran atípicos o muy atípicos. Llamaremos α_1 y α_2 a los puntos críticos, inferior y superior, para los valores típicos, que consiste en restarle o sumarle a los rangos inferior y superior, respectivamente una y media (1.5) veces la amplitud de las cajas y β_1 y β_2 a los puntos críticos, inferior y superior, para los valores muy atípicos. De modo que llamaremos atípicos al conjunto de los valores definido por el siguiente rango:

$$Q_1 - (1.5 * R_{Q3-Q1}) > \boldsymbol{atípicos} > Q_3 + (1.5 * R_{Q3-Q1})$$

Por su parte, llamaremos muy atípicos al conjunto de los valores definido por el siguiente rango:

$$Q_1 - (3 * R_{Q3-Q1}) > \boldsymbol{muy\ atípicos} > Q_3 + (3 * R_{Q3-Q1})$$

Donde: R_{Q3-Q1} es la amplitud de las cajas.

2.1. Análisis de datos cualitativos

Entenderemos por datos cualitativos el caso de datos cuantitativos con variables no métricas o categóricas. En la investigación social es muy frecuente la necesidad de lidiar con el procesamiento de datos para variables cualitativas o, en otros términos, variables categóricas o aquellas cuyo nivel de medición permite clasificarlas como: nominales. (Stevens, 1946). Ejemplo de estas son el sexo, la clase social, el lugar de procedencia, el estado civil, etc. Son variables que, en cualquier caso, sólo permiten el conteo de la cantidad de veces que aparecen los distintos valores o categorías que las constituyen como tales. En otras palabras, las variables cualitativas son, básicamente, aquellas que por su naturaleza no numérica sólo permiten el análisis de frecuencias.

Además del análisis de frecuencias, el cual es la forma más básica de análisis univariado de datos, el análisis cualitativo puede extenderse al análisis bivariado por la vía de los análisis de contingencia. Un análisis de continencia responde de forma fundamental a una dsitribución bivariada de datos, definidos en una matriz de doble entrada. Con el análisis de contingencia, las posibilidades heurísticas del investigador con relación a los datos se extiende para incluir, además de los análisis de frecuencia para cada variable por separado, los análisis de las frecuencias conjuntas o aquellos en los cuales el carácter de la unidad de análisis está definido por la combinación de dos valores (x,y).

Uno de los grandes recursos estadísticos para el análisis de datos cualitativos es la prueba de chí cuadrado. La misma es uno de los productos más importantes de la obra de Karl Pearson, con un rango de aplicaciones mucho mayor que el problema específico para el cual fue creado. (Walker, 1978). La prueba de chí cuadrado se fundamenta en el análisis de las frecuencias de los datos observados, es por esta razón, que su aplicación es viable incluso para variables con un nivel de medición tan bajo como el nominal. Los principales contextos para el uso de la prueba de chí cuadrado se reducen a los siguientes: análisis de la bondad de ajuste y el contraste de independencia de las variables.

En cualquier caso, la hipótesis nula que da sentido al uso de la prueba de chí cuadrado es que las frecuencias observadas se comportan de la misma manera que las frecuencias determinadas para una distribución teórica de referencia. En el caso de los análisis univariados sirve de base para la inferencia del contraste asociado con la proporción. La hipótesis nula clásica, para los análisis

de contingencia, sostiene que las n para las frecuencias de las distintas casillas definidas por un carácter i y j, son equivalentes más allá del error aleatorio.

2.1.1. Tablas de contingencia

Una distribución conjunta de dos variables Cuando se trabaja el cruce de variables categóricas se requiere, en principio, la construcción de bases de datos con por lo menos dos entradas, en las que cada entrada representa las variaciones de una determinada variable categórica. Como resultado de esta distribución se genera una presentación de los valores de cada una de las variables implicadas en filas y columnas, colocando en cada casilla el número de casos que cumple con ambos valores. De forma que las frecuencias hacen referencia a la presencia conjunta de los valores de las variables implicadas, en las distintas unidades de análisis y, por tanto, de la relación entre las variables involucradas. A estas tablas de frecuencia se les conoce como tablas de contingencia. (Pardo & Ruíz, 2005).

La tabla x, es un ejemplo de tabla de contingencia para el caso de una investigación en la cual el investigador se propuso estudiar la posible relación entre el bienestar psicológico y la disposición a fluir en el trabajo de una muestra de 200 empleados en el área de la salud.

Tabla x. *tabla de contingencia para la relación entre el bienestar psicológico y la disposición a fluir en el trabajo.*

		Disposición a Fluir en el Trabajo			
		Bajo	Medio	Alto	Total
Bienestar Psicológico General	Bajo	25	3	10	38
	Medio	65	49	22	136
	Alto	1	0	25	26
Total		91	52	57	200

Fuente: propia.

Tal como se mencionara en el párrafo anterior, en lugar de utilizar sólo dos variables o criterios de clasificación para generar una tabla de contingencia bidimensional, también se podría haber utilizado tres o más criterios, lo que llevaría a obtener tablas multidimensionales.

Notación de tablas de contingencia

Una vez familiarizados con lo que es una distribución conjunta y el marginal de una tabla, podemos pasar a emplear una notación para referirnos a cada uno de los elementos que la conforman. A los valores de la variable puestos en el eje de las filas se le denota como *i* y a los puestos en el eje de las columnas como *j*, por lo que a la frecuencia conjunta se le conoce como n*ij* . Al máximo de *i* se le denota como *I*, y al máximo de *j* como *J*, de forma que en este ejemplo *I*=2 y *J*=3. Para referirnos a la dimensión de la tabla, se estila multiplicar el número de filas por los de columnas del modo *I x J*, en este caso sería de 2 (filas) x 3 (columnas), por lo que nos referimos a esta como una tabla 2x3, lo que da un total de 6 casillas de frecuencias conjuntas. La Tabla 1 permite ilustrar la forma convencional de notación empleada para las tablas de contingencia. El resultado, es el siguiente:

Tabla x. *notación de una tabla 2x3.*

		Variable B		
Variable A	j=1	j=2	j=3	$n_{i\cdot} = \sum_{j=1}^{J} n_{ij}$
i= 1	n_{11}	n_{12}	n_{13}	$n_{1\cdot} = \sum_{i=1}^{J} n_{1j}$
i= 2	n_{21}	n_{22}	n_{23}	$n_{2\cdot} = \sum_{i=1}^{J} n_{2j}$
$n_{\cdot j} = \sum_{i=1}^{I} n_{ij}$	$n_{\cdot 1} = \sum_{i=1}^{I} n_{i1}$	$n_{\cdot 2} = \sum_{i=1}^{I} n_{i2}$	$n_{\cdot 3} = \sum_{i=1}^{I} n_{i3}$	$N = n_{\cdot\cdot} = \sum_{i=1}^{I}\sum_{i=1}^{j} n_{ij}$

Aclarada la notación para las frecuencias conjuntas, veamos ahora la notación para los marginales. Como ya vimos, los marginales se calculan sumando en el sentido en el que está la variable. En nuestro ejemplo, la variable B conforma las columnas, y por tanto corresponde con la letra *j*. Si estamos interesados en calcular la cantidad total de elementos en B (que se corresponde con j=1), en términos de notación debemos sumar n11+n21, es decir, el subíndice *i* cambia, pero el subíndice *j* se mantiene igual (j=1). Para expresar este cambio de la i y esta constancia de la j

diremos que el marginal es n·j, en este caso más en concreto n·1=n11+n21. Esto amerita que recordemos que el · significa que estamos sumando las filas (i) en el sentido de las columna que estamos calculando (j).

Hasta ahora se ha presentado la notación para hacer referencia a la distribución conjunta de variables, en términos de las frecuencias absolutas de cada casilla (en este caso, el número de personas). Sin embargo, en algunas ocasiones el investigador puede preferir otras opciones a las frecuencias absolutas, como es el caso de los porcentajes para poder comparar tanto poblaciones de distinto tamaño como las casillas de una misma tabla en la que los marginales son distintos. (Pardo & Ruíz, 2005). Lo habitual en ciencias sociales es que el resultado quede redondeado a un decimal, a ninguno (menos frecuente en el ámbito académico) o quizá hasta la centésima (en un prurito inútil de precisión), pero no más.

Cálculo de porcentajes

Podemos recordar que el procedimiento general para calcular cualquier tipo de porcentajes es sencillo: basta con dividir la frecuencia de la casilla que nos interesa sobre el total (marginal) que corresponda, y multiplicar por 100. A diferencia del cálculo de los porcentajes para distribuciones univariadas, en el caso de las distribuciones conjuntas se nos presentan las opciones de calcular el porcentaje de las casillas con base en los tres totales que evidenciamos en la sección anterior, a saber: a) sobre la base del total general, b) sobre la base del total de las filas, y c) sobre la base del total de las columnas.

Tomemos el caso *a* cálculo de los porcentajes con base al total general, si deseamos calcular el porcentaje sobre la base de éste total, lo que habrá que hacer es dividir la frecuencia de cada casilla (n_{ij}) sobre el total de la tabla (N). La fórmula, por tanto, sería:

$$p_{ij}^{T} = \frac{n_{ij}}{N} * 100$$

Donde p_{ij}^{T} hace referencia al porcentaje para una determinada casilla *ij* con base al total (*T*), n_{ij} es el número de elementos que se encuentran en la casilla *ij* y N representa la sumatoria de todos los elementos presentes en la tabla, para todas las filas y columnas.

Para el caso *b* cálculo de los porcentajes con base al total de las filas, si deseamos calcular el porcentaje sobre la base de éste total, lo que habrá que hacer es dividir la frecuencia de cada casilla (n_{ij}) sobre el total de la fila en cuestión ($n_{i.}$). La fórmula, por tanto, sería:

$$p_{ij}^{F} = \frac{n_{ij}}{n_{i.}} * 100$$

Donde p_{ij}^{F} hace referencia al porcentaje para una determinada casilla *ij* con base al total de la fila correspondiente ($n_{i.}$), n_{ij} es el número de elementos que se encuentran en la casilla *ij* y $n_{i.}$ representa la sumatoria de todos los elementos presentes en la fila *i*, para todas las columnas.

Para el caso *c* cálculo de los porcentajes con base al total de las columnas, si deseamos calcular el porcentaje sobre la base de éste total, lo que habrá que hacer es dividir la frecuencia de cada casilla (n_{ij}) sobre el total de la columna en cuestión ($n_{.j}$). La fórmula, por tanto, sería:

$$p_{ij}^{C} = \frac{n_{ij}}{n_{.j}} * 100$$

Donde p_{ij}^{C} hace referencia al porcentaje para una determinada casilla *ij* con base al total de la columna correspondiente ($n_{.j}$), n_{ij} es el número de elementos que se encuentran en la casilla *ij* y $n_{.j}$ representa la sumatoria de todos los elementos presentes en la columna *j*, para todas las filas.

2.2. Indicadores sumarios de posición, centralidad, variabilidad y forma de la distribución (asimetría y curtosis)

Hemos visto cómo se puede resumir el volumen de los datos a versiones más económicas que llamamos distribuciones de frecuencias. Sin embargo, en la mayoría de los casos el investigador requerirá de valores que permitan sintetizar la información general a unos pocos valores característicos que llamamos estadísticos o estadígrafos. En muchas ocasiones éstos suelen funcionar como estimadores de los parámetros de la población. El paso de la descripción a la inferencia es objeto de otro capítulo, en este punto nos ocuparemos de la conceptualización y construcción de los principales estadígrafos. Entre estos se cuentan tres clases: a) medidas de posición y tendencia central, b) medida de dispersión y c) medida de forma (asimetría y curtosis).

Una de las formas privilegiadas para la representación de los datos implica el uso de estadísticos de resumen. Los estadísticos de resumen permiten obtener información valiosa en forma eficiente con relación a las preguntas fundamentales para la descripción de los datos, a saber:

en torno a qué datos se tienden a agrupar los datos y cuán concentrados están los datos en torno a los valores centrales. En este sentido, los estadísticos del tipo: cuantiles, percentiles, media, varianza, entre otros, serán presentados y estudiados. Para complementar la información que nos proporcionan los estadísticos de tendencia central y variabilidad, el investigador cuenta con los estadísticos de forma de la distribución. Con estos, el investigador cuenta con indicadores estandarizados para dar cuenta del sesgo y consistencia de los datos.

2.2.1. Cuantiles, percentiles y cuartiles.

El nombre genérico de estadístico de posición hace referencia al punto de corte en la distribución de valores de x asociada con una determinada frecuencia acumulada; lo cual da lugar a la forma más simple del estadístico de posición el **cuantil** y el mismo se define como el valor bajo el cual se encuentra una determinada proporción de los valores de una distribución. Por su parte, se entiende por **percentil** aquellos valores que dividen a un conjunto de datos ordenados en cien partes iguales. Se representan por P_1, P_2 …, P_{99}. Por **cuartil** se entiende los valores que dividen en cuatro partes iguales a un conjunto de datos ordenados. Se representan por Q_1, Q_2, y Q_3. Para cerrar los **deciles** designan a los valores que dividen en diez partes iguales a un conjunto de datos ordenados. Se representan por D_1, D_2, D_3, …D_9.

1. Indicar para qué sirve el conocimiento de los cuantiles.
2. Presentar las fórmulas para el cálculo de cuantiles, tanto para datos agrupados como no agrupados.

   Para datos no agrupados:
   Se comienza por ordenar los datos de manera ascendente.
   Luego se calcula el índice: $i = \left(\frac{p}{100}\right) * n$

Si "i" es entero, el *p-ésimo* percentil es el promedio de los valores de los datos ubicados en los lugares "i" e "i +1".
Si "i" no es entero, se redondea. El valor entero inmediato mayor que "i" indica la posición del *p-ésimo* percentil.

Para datos agrupados por intervalos simples:

Interpolación Lineal

En el artículo de Lezama (2011), podemos encontrar apoyo a la idea de estimar los valores percentilares para un determinado valor de x, cuando se conoce la línea de regresión o función definida por la relación entre los valores de x y los valores de y $f_{(x_i)} = y_i$. Partiendo del supuesto de que el comportamiento de los valores percentilares con respecto a los valores de x obedecen a una razón constante o, en otras palabras, a una distribución uniforme. De modo que, la relación entre las distancias de x y y está dada por la siguiente ecuación: $\frac{\overline{AC}}{\overline{AB}} = \frac{\overline{CE}}{\overline{BD}}$, tal que $\overline{BD} = \frac{\overline{AB}\,\overline{CE}}{\overline{AC}}$; Donde:

$$\overline{AC} = x_2 - x_1$$
$$\overline{AB} = x - x_1$$
$$\overline{CE} = y_2 - y_1$$
$$\overline{BD} = y - y_1$$

Sustituyendo: $\overline{BD} = \frac{\overline{AB}}{\overline{AC}}\overline{CE}$ es igual a escribir: $y = \left[\frac{(x-x_1)(y_2-y_1)}{(x_2-x_1)}\right] + y_1$

Supongamos el siguiente caso:

x = Puntajes	y = Percentiles
x_2= 39	y_2= 15
x= 38	y= ?
x_1= 37	y_1= 4

Procedemos a sustituir: $y = \left[\frac{(38-37)(15-4)}{(39-37)}\right] + 4 = y = \left(\frac{1*11}{2}\right)$+4= 9.5 ≈ 10

Con estos resultados se puede afirmar que para un valor de x igual a 38 corresponde un valor percentilar de 10. En otras palabras, con un puntaje de 38 se ha superado aproximadamente al 10% de las observaciones.

Para datos agrupados por intervalos de clase:

a. Se aplica la fórmula: $p = Li + \left[\frac{\left\{\left(N*\frac{p}{100}\right) - F_i^{-1}\right\}}{f_i} * a\right]$

Para aplicar la fórmula, los pasos son:

1. Ubicar el resultado de:$\left\{n * \frac{p}{100}\right\}$, donde: "n" es el número de casos y "p" es el percentil que se desea trabajar.
2. Si no está el valor, se pasa al inmediato superior.
3. Al ubicar el valor de F_i determinamos el intervalo de donde se obtendrán los datos para sustituir en la ecuación.

Nota: si se trabajan con límites reales e imaginarios, se toman los reales.

Donde:
p= percentil que se desea conocer.
L_i= límite inferior del intervalo.
F_i^{-1}= frecuencia absoluta acumulada hasta el intervalo anterior al que contiene al percentil en cuestión.
f_i= frecuencia absoluta simple del intervalo percentilar.
a= amplitud del intervalo de clase = $(l_s - l_i) + 1$.

Nota: el facilitador tendrá presente explicar la razón por la cual los resultados obtenidos de la aplicación del cálculo de los cuantiles para datos no agrupados difiere en la mayoría de las ocasiones de los resultados obtenidos de la aplicación de la fórmula para datos agrupados. Para ello refrescará la idea que sugiere que la fórmula para el cálculo de los cuantiles con datos agrupados asume el supuesto matemático de que todos los valores dentro del intervalo de clase poseen cargas homogéneas de frecuencias, lo cual es algo que no ocurre en la mayoría de las distribuciones estadísticas. Ante este escenario es lógicamente admisible el hecho de que la cantidad de casos acumulados hasta un determinado valor de la distribución no coincida al comparar los cálculos por ambas fórmulas. El ejemplo a continuación ilustra lo anterior.

RANGO PERCENTILAR

Es una expresión mediante la cual podemos hallar el porcentaje, asociado a un valor de la variable. Dicha expresión se obtiene al despejar "$p_{(x)}$" en la fórmula de percentiles para datos agrupados, lo que deviene en la siguiente ecuación:

$$p_{(x)} = \frac{F_i^{-1} + \frac{(x - x_{li}) * f_i}{a}}{N} * 100$$

Donde:
$p_{(x)}$= Percentil correspondiente al valor de x_i.
N= número total de observaciones.
x_{li}= límite inferior del intervalo que contiene a x_i.
F_i^{-1}= frecuencia absoluta acumulada hasta el intervalo anterior al que contiene al valor de x en cuestión.
f_i= frecuencia absoluta simple del intervalo que contiene a x.
a= amplitud del intervalo de clase = $(l_s - l_i) + 1$.

En lo sucesivo el proceso para hallar el rango percentilar es:

1. Ubicar el valor de la variable (x) que nos dan, en el intervalo que le corresponda.
2. Una vez ubicado, podemos determinar l_i, f_i, etc., para sustituir en la fórmula.

2.2.2. Estadísticos de tendencia central.

Al describir grupos de observaciones, con frecuencia es conveniente resumir la información con un solo número. Este número que, para tal fin, suele situarse hacia el centro de la distribución de datos se denomina estadígrafo de tendencia central o de centralización. Entre las medidas de tendencia central tenemos: a) moda, b) mediana y c) media.

2.2.2.1. Moda.

La moda se denota Mo y refiere a la categoría más repetida en la distribución, en otras palabras, el valor de la variable con mayor frecuencia absoluta. En cierto sentido la definición matemática corresponde con la expresión cotidiana "estar de moda", esto es, ser lo que más se lleva.

Su cálculo es extremadamente sencillo, pues sólo necesita un recuento. En variables continuas, expresadas en intervalos, existe el denominado intervalo modal o, en su defecto, si es necesario obtener un valor concreto de la variable, se recurre a la interpolación.

2.2.2.2. Mediana.

La mediana se denota como Md y puede definirse como el valor que en una distribución de valores ordenados es asociado con el percentil cincuenta en un conjunto de observaciones. En otras palabras, es el valor que divide a la distribución de valores ordenados en dos partes con cargas iguales de casos.

2.2.2.3. Media.

La media puede definirse como el valor que se obtiene al dividir la sumatoria de un conjunto de valores observados entre el número total de observaciones realizadas. Es aquella que representa el promedio aritmético de una medición, la misma actúa como un punto de equilibrio, de manera que las observaciones menores equilibran a las mayores.

Propiedades de la media:

Se definen como propiedades de la media:

a) La media aritmética está comprendida entre el valor máximo y el valor mínimo del conjunto de los datos.
b) La suma de las desviaciones de todos los valores a la media es cero.
c) Si a todos los valores de la variable se le suma o resta una constante, la media aritmética queda aumentada o disminuida en dicha cantidad.
d) Si todos los valores de la variable se multiplican o dividen por una constante la media aritmética queda multiplicada o dividida por dicha constante.

e) La media no es un dato confiable cuando hay datos extremos que toman valores muy altos o muy bajos.

2.2.3. Estadísticos de variabilidad.

Las medidas de variabilidad, también llamadas medidas de dispersión, muestran la variabilidad de una distribución, indicando por medio de un número, cuán diferentes son las puntuaciones de una variable con respecto a un valor central o promedio. Cuanto mayor sea ese valor, mayor será la variabilidad, cuanto menor sea, más homogénea serán las observaciones. En resumen, la dispersión nos informa cuán variados son los individuos observados. Se incluyen: a) el rango o recorrido, b) la desviación típica, asociada a los trabajos pioneros de Karl Pearson (Walker, 1978) y c) las medidas de dispersión relativas (coeficiente de variación de Pearson).

Coeficiente de variación de Pearson

Se define el *coeficiente de variación de Pearson* (CV) para todo x distinto de 0, como: $CV = \frac{S_x}{\bar{X}}$

2.2.4. Estadísticos de forma de la distribución.

Un interés subyacente al análisis descriptivo de los datos es evaluar la calidad de estos como estimadores. Hasta ahora hemos visto una serie de estadísticos que nos permiten atender las demandas fundamentales del análisis descriptivo de los datos, me refiero a los criterios: tendencia central y variabilidad. Sin embargo, hemos visto también que estos indicadores se quedan cortos a la hora definir objetivamente cuán normal es la distribución en cuestión. La normalidad es un requisito clave en el cálculo de probabilidades y, por consecuente, en la inferencia estadística.

Evaluar los niveles de normalidad de la distribución nos permitirá evaluar la capacidad de nuestros estadígrafos para representar el comportamiento de la población. Esto se basa en dos ideas claves: la primera, cuan amplia es la diferencia entre la media y la mediana de los datos, es decir, cuán sesgada está la media para dar cuenta de todos los distintos valores observados. Segundo, cuál es el nivel de consistencia de la media; es decir, cuán parecidas son las distintas observaciones

con respecto a la media. Estas dos preguntas se responden con los estadísticos de forma de la distribución, básicamente: asimetría y curtosis.

2.2.4.1. Asimetría

Las medidas de asimetría son estadísticos que permiten determinar el grado de simetría (o asimetría) observado en una distribución para una variable aleatoria cualquiera.

El estadístico de asimetría corresponde, fundamentalmente, con el tercer momento central de la distribución. Por tanto, toma como eje de simetría una recta paralela al eje de las ordenadas que pasa por la media de la distribución.

El coeficiente de asimetría cumple con las siguientes propiedades: a) si una distribución es simétrica, existe el mismo número de valores a la derecha que a la izquierda de la media, por tanto, el mismo número de desviaciones con signo positivo que con signo negativo, b) la asimetría es positiva (o a la derecha) si la "cola" a la derecha de la media es más larga que la de la izquierda, es decir, si hay valores más separados de la media a la derecha y c) la asimetría es negativa (o a la izquierda) si la "cola" a la izquierda de la media es más larga que la de la derecha, es decir, si hay valores más separados de la media a la izquierda. (Glass & Stanley, 1974).

Hemos hablado de la asimetría como el tercer momento central de la distribución. Sin embargo, esta es una definición particularmente sesgada asociada al concepto fisheriano de asimetría. El cálculo de la asimetría ha recibido aportes de tres autores claves: Fisher, Pearson y Bowley. Los detalles con relación a las fórmulas para cada caso se exponen a continuación.

Coeficiente de asimetría de Fisher:

$$Asimetría = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{NS_x^3} = \frac{1}{N}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{S_x}\right]^3 = \frac{\sum_{i=1}^{n}(z)^3}{N} = \overline{z^3}$$

Coeficiente de asimetría de Pearson:

$$Asimetría = \frac{3(\bar{X} - Md)}{S_x}$$

Coeficiente de asimetría de Bowley:

$$Asimetría = \frac{Q_3 + Q_1 - 2Md}{Q_3 - Q_1}$$

2.2.4.2. Curtosis.

Las medidas de curtosis son estadísticos que permiten determinar el grado de apuntamiento (o aplanamiento) observado en una distribución para una variable aleatoria cualquiera.

El estadístico de curtosis corresponde, fundamentalmente, con el cuarto momento central de la distribución. Por tanto, toma como eje de apuntamiento o pendiente una recta paralela al eje de las ordenadas que pasa por la media de la distribución.

El coeficiente de curtosis cumple con las siguientes propiedades: a) si una distribución es más alta en el centro y tiene colas más anchas que la normal es leptocúrtica, b) la distribución es mesocurtica si la curtosis es moderada, por tanto, la distribución tiene una altura en el centro y colas tan anchas como una distribución normal y c) la distribución es platicurtica si la curtosis es baja, por tanto, la distribución es menos alta en el centro y tiene colas menos anchas que una distribución normal. (Glass & Stanley, 1974).

El cálculo de la curtosis se asocia fundamentalmente con la siguiente fórmula:

$$Curtosis = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{NS_x^4} = \frac{1}{N}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{S_x}\right]^4 = \frac{\sum_{i=1}^{n}(z)^4}{N} = \overline{z^4}$$

De este modo, se cumple:

Curtosis > 3, la distribución es leptocurtica.

Curtosis = 3, la distribución es mesocurtica.

Curtosis < 3, la distribución es platicurtica.

Algunos autores sugieren hacer un ajuste al resultado del cálculo de la curtosis, con el fin de equiparar a las distribuciones mesocurticas con el punto de origen. El ajuste consiste en restar 3 puntos al resultado de la ecuación.

$$Curtosis = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{NS_x^4}\right] - 3$$

De este modo, se cumple:

Curtosis > 0, la distribución es leptocurtica.

Curtosis = 0, la distribución es mesocurtica.

Curtosis < 0, la distribución es platicurtica.

2.3. La distribución normal

En el vocabulario estadístico y, especialmente, en el de la estadística matemática y de la probabilidad, la distribución normal o gaussiana, conocida por la famosa campana de Gauss representa uno de los fenómenos más significativos para la estimación de parámetros. La misma hace referencia al comportamiento de variables continuas *estocásticas*, a partir de la definición de una de las funciones teóricas de distribución de probabilidad que se presume mejor representa el comportamiento de las distribuciones asociadas con fenómenos reales, en el contexto de la aleatoriedad. En el contexto de la probabilidad, la distribución normal aparece como el límite de varias distribuciones para variables continuas y discretas, tales como la binomial, la función *t, F* y chí cuadrado.

La distribución normal es ampliamente conocida por la forma de la gráfica de su función de densidad, la cual se asocia con la figura de una campana; ésta característica está asociada con las claves teóricas de dicha distribución para la inferencia estadística. La principal característica asociada a la distribución normal es el de ser simétrica con respecto a la media o parámetro estadístico central. La siguiente propiedad es el carácter asintótico de la misma con relación al eje de las abscisas, en otras palabras, sus valores en los extremos de la distribución tienden a acercarse indefinidamente a cero.

Numerosos autores coinciden en afirmar que “la importancia de esta distribución radica en que permite modelar numerosos fenómenos naturales, sociales y psicológicos”. (Kerlinger & Lee, 2002). Su uso en las ciencias sociales es una de las herencias más significativas emanadas de los trabajos de Francys Galton y los estudios antropométricos. Los trabajos de Galton condujeron a la conclusión de que variables como el peso y la estatura se distribuye de manera normal, a lo cual le sucedieron los trabajos de psicólogos como Alfred Binet quienes a principios del S.XX determinaron como el comportamiento de las medidas de inteligencia se comporta de manera

normal. En lo sucesivo y hasta el día de hoy es incontable la cantidad de investigaciones que han fundado el contraste de sus hipótesis en la curva normal.

En el contexto de los estudios inferenciales, especialmente los bivariados, la distribución normal es importante por su relación con el comportamiento del error de estimación y la construcción de funciones por mínimos cuadrados, uno de los métodos de modelización más populares en estadística derivado de los trabajos de Frederich Gauss. La distribución normal es el fundamento de muchas pruebas estadísticas, particularmente de la familia de las pruebas paramétricas, las cuales están basadas en una supuesta "normalidad".

2.3.1. Estandarización de variables aleatorias normales

Partiendo del concepto de simetría es posible relacionar todas las variables aleatorias que se comportan de manera normal con la distribución normal estándar.

Si X ~ N(μ,σ), entonces: $z = \frac{x-\mu}{\sigma}$; $\forall\, x \in R$ es una variable aleatoria normal estándar: Z ~ N(0,1).

La transformación de una distribución X~N(μ, σ) en una N(0, 1), llamada transformación lineal directa, se conoce también como normalización, estandarización o tipificación de la variable X. Esta es una propiedad particularmente útil para la transformación lineal y la comparación de puntajes en distintas distribuciones. En el contexto de la investigación psicológica se asocia con la construcción de normas y perfiles en psicometría.

2.3.1.1. Ajuste de la Binomial a la normal: Teorema de Laplace - DeMoivre

Una de las bondades de la distribución normal es que sus propiedades pueden extenderse incluso a variables que no son continuas. Se atribuye a los matemáticos Abraham DeMoivre y Pierre Laplace, el desarrollo de la tesis de que dada una variable binomial cualquiera Bin(x), de parámetros n y p. donde n es igual al número de observaciones y p es la frecuencia relativa de aparición del evento favorable. Se dice que esta tiende a distribuirse de manera normal en la medida en que el número de observaciones aumenta, y los valores de p no son demasiado cercanos a 0 o 1, es decir Bin(x) ~N(np;npq). Por extensión, se dice que:

$$f_{np,npq}(x) = \frac{1}{npq\sqrt{2\pi}}\, e^{-\frac{1}{2}(\frac{x-np}{npq})^2};\ \forall\ X \in R$$

2.3.1.2. Teorema de Tchebycheff

Se atribuye al matemático ruso P. L. Tchebycheff determinar que la fracción de área entre cualesquiera dos valores simétricos alrededor de la media está relacionada con la desviación estándar. Considerando que el área bajo una curva de distribución de probabilidad, o de un histograma de probabilidad, suma 1, el área entre cualesquiera dos números es la probabilidad de que la variable aleatoria tome un valor entre estos números. De este modo, el área y por tanto la probabilidad de ocurrencia de un grupo de terminado de valores en torno de la media está comprendida por una serie de porciones, entre las más emblemáticas se cuentan:

- En el intervalo [μ - σ, μ + σ] se encuentra comprendida, aproximadamente, el 68,26% de la distribución;
- En el intervalo [μ - 2σ, μ + 2σ] se encuentra, aproximadamente, el 95,44% de la distribución;
- Por su parte, en el intervalo [μ -3σ, μ + 3σ] se encuentra comprendida, aproximadamente, el 99,74% de la distribución. Estas propiedades son de gran utilidad para el establecimiento de intervalos de confianza.

Otras áreas y sus magnitudes pueden ser calculadas con el uso de tablas estadísticas que presentan el área acumulada bajo la curva normal estandarizada.

2.3.1.3. La distribución normal bivariada.

Como se ha mencionado, una de las aplicaciones más importantes para la teoría de la distribución normal ha sido en el campo de la inferencia estadística, particularmente, en la estadística bivariada. Hasta ahora se ha hecho un estudio de la distribución normal univariada, en lo sucesivo atenderemos el caso en el cual la distribución obedece a una distribución bivariada.

Al igual que la distribución normal univariada, la distribución normal bivariada constituye una familia de superficies tridimensionales. (Glass & Stanley, 1974). De este modo, si X e Y están

normalmente distribuidas y son independientes, su distribución conjunta también está normalmente distribuida, es decir, el par (X, Y) debe tener una distribución normal bivariante. En cualquier caso, un par de variables aleatorias normalmente distribuidas no tienen por qué ser independientes al ser consideradas de forma conjunta. Todas las distribuciones normales bivariadas responden a las siguientes características:

- La distribución de las puntuaciones de X, de forma independiente a las puntuaciones de Y con las cuales se aparean, es una distribución normal.
- La distribución de las puntuaciones de Y, de forma independiente a las puntuaciones de X con las cuales se aparean, es una distribución normal.
- Para cada puntuación X, por ejemplo, las puntuaciones Y correspondientes en X, se distribuyen normalmente con una varianza $\sigma_{y,x}^2$
- Para cada puntuación Y, por ejemplo, las puntuaciones X correspondientes en Y, se distribuyen normalmente con una varianza $\sigma_{x,y}^2$
- Las medianas de las puntuaciones Y correspondientes a las X, conforman una recta.

La función de distribución para el caso de una distribución bivariada tipificada de parámetros $\mu = (\bar{x}_1, \bar{x}_2)$ y Σ= matriz de covarianza x,y. de modo que:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}-\frac{2\rho xy}{(\sigma_x\sigma_y)}\right)\right)$$

Donde ρ es el coeficiente de correlación entre X e Y.

2.4. Transformaciones de Puntajes

La esencia de la transformación de puntajes es facilitar la interpretación de los puntajes de una determinada variable al permitir las comparaciones dentro y entre distribuciones, a nivel intra e intergrupal. Se puede definir la transformación de puntajes como una estrategia estadística que consiste en una familia de operaciones que permiten convertir, a partir de reglas estandarizadas, los puntajes originales observados en una determinada variable a otros valores con fines de comparación. Aunque son innumerables las técnicas de transformación de puntajes, éstas se han agrupado en dos grandes categorías: lineales y no lineales.

2.4.1. Transformaciones lineales y no lineales.

Una transformación lineal implica cambiar la escala original, mientras se conserva idéntica la forma de la distribución original y, en consecuencia, las posiciones relativas de los individuos en esa distribución transformada. (Magnusson D. , 1975).

Nos referimos a puntajes transformados no lineales, cuando "la forma de la distribución de los puntajes originales es no normal, en cuyo caso la transformación realizada cambia la forma de esa distribución, `obligándola' a comportarse como una distribución normal" (Lezama, 2011).

2.4.1.1. Escalas directas y derivadas.

El término lineal se deriva del discurso matemático y hace referencia a las funciones cuyo exponente es igual a 1, en las variables y en los parámetros. (Gujarati & Porter, 2010). Este concepto es cónsono con la función de categorización $x' = a + bz_i$ que caracteriza a las variables con nivel de medición de intervalo, propio de la medición en ciencias sociales. Donde, empíricamente se establece que x' es el valor estimado de la variable medida, a corresponde con el punto de origen relativo para la función de categorización, asociado con la media de los valores estimados cuando el valor del indicador es igual a cero, b representa la magnitud del cambio para el valor estimado por cada unidad del indicador y z_i es el indicador definido como la variable observada con el fin de estimar los cambios en x. La esencia de la transformación lineal está en la manipulación de los parámetros que definen la ecuación de la recta, entiéndase *a* y *b*.

Las transformaciones lineales directas se fundamentan en las desviaciones típicas de la distribución. También se conocen como puntuaciones z y señalan la distancia en términos de las desviaciones típicas a las cuales se encuentra un puntaje con respecto a la media. (Lezama, 2011). Dichas puntuaciones se obtienen al despejar de la función de categorización para la medición $x' = a + bz_i$ los valores de z, dando lugar a la ecuación: $z_i = \frac{x'-a}{b}$. De esta manera, hemos independizado el cambio relativo en la variable observada, del intercepto y la pendiente estimada, de la función de categorización que define a la distribución de origen.

Por ejemplo: dado un puntaje de 20 puntos en una distribución normal con media igual a 10 y desviación típica de 5, en otras palabras N~(20,5), el valor de z es igual a 2. Dado que: $z =$

$\frac{20-10}{5} = \frac{10}{5}$.= 2. De este modo, se puede decir que la persona con un puntaje estimado de 20 puntos se encuentra a 2 desviaciones típicas con respecto a la media de su grupo normativo.

De esta misma manera, las puntuaciones lineales derivadas, se fundamentan en la ecuación de la recta con base en las puntuaciones típicas resultadas de las escalas directas. De este modo permite convertir la escala de valores tipificados a una escala determinada cuyos valores tienen como principal propiedad superar el punto medio igual a cero y permitir que todos los valores adquieran un valor positivo real. (Lezama, 2011).

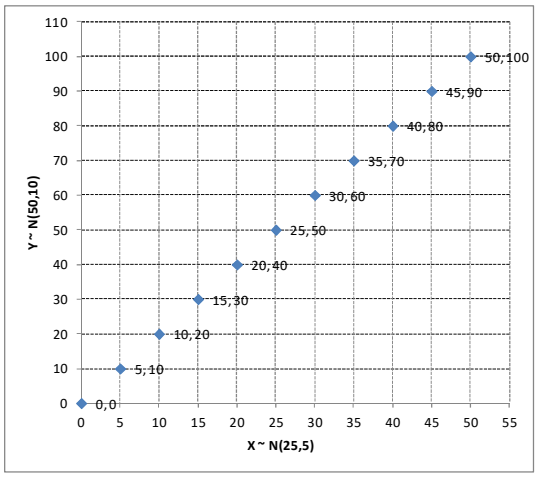
Reducida cualquier distribución a una escala con media cero y desviación típica uno y suponiendo que se comporta como una variable de intervalo, podemos someterla a la ecuación de transformación $y = a + bx_i$ o $\hat{y} = \bar{Y} + S_y z_i$. Empíricamente se puede evidenciar que la transformación mantiene idéntica la estructura de la variable original, a partir del estudio de la correlación de los puntajes directos y derivados. Así mismo ocurre con la gráfica de estos y la construcción de la línea de regresión. En cualquier caso nos indica como la ubicación relativa de un punto con respecto a la media de la escala de referencia se mantiene, más allá de los valores absolutos con los cuales se definan las escalas en contraste.

En el discurso estadístico, este punto de origen (a) se asocia con el primer momento de la distribución con respecto a la media $\bar{X}$, por esta razón el intercepto de la función afín que determina la transformación de puntajes derivados de las puntuaciones tipificadas toma como criterio la media de la escala a la cual se pretende convertir los puntajes originales. De este modo, un puntaje igual a cero en la escala original equivale a la media en la escala derivada.

Por su parte, la pendiente (b) obedece a la diferencia con respecto a la media por cada cambio en una unidad típica en la escala de puntajes directos. Esta unidad de cambio para los valores de los puntajes transformados será dada por la desviación típica de la distribución a la cual se convertirán los puntajes directos. Esto definirá una razón de cambio que determinará una distancia fija entre todos los puntos de la distribución.

Finalmente, x representa la variable independiente. Remplazaremos x por z, siendo z una variable que simplifica el comportamiento de cualquier distribución a su mínima expresión. Como se mencionó en el punto en cuestión, los puntajes z al restarle la media a los valores puntuales cancela cualquier distancia entre el punto de origen y los valores asociados a la media. Por su parte, la división de la diferencia entre los valores observados y la media, entre la desviación típica simplifica los desvíos observados con relación al desvío promedio.

Supongamos que hemos decidido convertir los puntajes de una distribución X~N(25;10) a una escala Y(50;10). Bastará con independizar el comportamiento de los puntajes de la distribución de origen de la escala original, reduciendo los valores a puntajes tipificados o z. luego esta unidad básica será usada en la ecuación siguiente: $\hat{y} = 50 + 10(z_i)$. Como se puede apreciar en el patrón de coordenadas de la gráfica, la media de origen se cruza con la media de la escala transformada, en lo sucesivo, un cambio en una desviación típica en la escala original se refleja en un cambio de una unidad típica en la escala transformada.



De acuerdo con los postulados de la medición expuestos por autores como (Stevens, 1946), (Dingle, 1950), (Gaito, 1980), (Hand, 1996), (Michell, 1986), entre otros. La transformación lineal, en los términos que han sido estudiados, sólo atiende a las variables con un nivel de medición de intervalo. Se descarta para este tipo de tareas a las variables con un nivel de medición inferior a intervalo, entiéndase ordinales o nominales. Por su parte, las variables con nivel de medición de razón y con un punto de origen real en 0 atenderán a la siguiente función de transformación: $x' = bz_i$

Bibliografía

Abelson, R. (1998). *La estadística razonada: reglas y principios.* Barcelona: Paidós.

American Psychological Association. (01 de Junio de 2010). *Ethical Principles of Psychologists and Code of Conduct.* Recuperado el 13 de Marzo de 2013, de http://www.apa.org/ethics/code/index.aspx#

American Psychological Association. (2010). *Manual de Publicaciones de la Amercan Psychological Assciation* (Tercera ed.). (M. Guerra Frias, Trad.) D.F., México: Manual Moderno.

Anastasi, A., & Urbina, S. (1998). *Test Psicológicos.* D.F., México: Prentice Hall.

Babbie, E. (1988). *Métodos de investigació por encuesta.* (J. Utrilla, Trad.) D.F., México: Fondo de Cultura Económica.

Balluerka, N., & Vergara, A. (2002). *Diseños de Investigación Experimental en Psicología.* Madrid, España: Prentice Hall.

Brown, C., & Ghiselli, E. (1969). *El método científico en Psicología.* (E. Prieto, Trad.) Buenos Aires, Argentina: Paidós.

Cohen, J. (1992). Cosas que he aprendido (hasta ahora). *Anales de Psicología, 8*(1-2), 3-17.

Comisión de Ética, Bioética y Biodiversidad. (Diciembre de 2010). Código de ética para la vida. (T. e. Ministerio del Poder Popular para la Ciencia, Ed.) Caracas, Venezuela: Ministerio del Poder Popular para la Ciencia, Tecnología e Industrias Intermedias.

Dingle, H. (1950). A theory of measurement. *The British Journal for the Philosophy of Science*, 5 - 26.

Federación de Psicólogos de Venezuela. (1981). Código de Ética Profesional del Psicólogo de Venezuela. *II Asamblea Nacional Ordinaria de la Federación de Psicólogos de Venezuela.* Barquisimeto: Autor.

Gaito, J. (1980). Measurement scales and statistics: resurgence of an old misconception. *Psychological Bulletin*, 564 - 567.

Glass, G., & Stanley, J. (1974). *Métodos Estadísticos Aplicados a las Ciencias Sociales*. (E. Galvis, & E. Guzman, Trads.) Madrid, España: Prentice Hall.

Grande, I., & Abascal, E. (1989). *Métodos de análisis multivariante para la investigación comercial.* Barcelona: Ariel.

Gujarati, D., & Porter, D. (2010). *Econometría* (Quinta ed.). (P. Carril Villareal, Trad.) D.F., México: Mc Graw Hill.

Hand, D. J. (1996). Statistics and the Theory of Measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 445 - 492.

Hernández, R., Fernández, C., & Baptista, P. (2010). *Metodología de la Investigación* (Quinta ed.). D.F., México: Mc Graw Hill.

Kerlinger, F., & Lee, H. (2002). *Investigación del comportamiento. Métodos de investigación en ciencias sociales. (4a Ed).* México: Mc Graw Hill.

Knapp, T. (1990). Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nurse Research*, 121 - 123.

Landau, D., & Lazarsfeld, P. (1978). Quetelet Adolphe. En W. Kruskal, & J. Tanur, *International Encyclopedia of Statistics* (Vol. 2, págs. 824-834). New York, USA: The Fee Press.

León, O., & Montero, I. (2003). *Métodos de Investigación en Psicología y Educación* (Tercera ed.). Madrid, España: Mc Graw Hill.

Lezama, L. (2011). Puntuaciones relacionadas con las normas. *Psicología*, 107-143.

Lord, F. (1953). On the statistical treatment of football numbers. *The American Psychologist*, 750 - 751.

Luce, R. D. (1997). Quantification and symmetry: Commentary on Michell Quanttative Science and the definition of measurement in Psychology. *British Joumat of Psychology*, 395 - 398.

Magnusson, D. (1975). *Teoría de los test.* México: Biblioteca Técnica de Psicología.

Marx, M., & Hillix, W. (1983). *Sistemas y teorías psicológicas contemporáneas* (3a ed.). Buenos Aires, Argentina: Paidós.

McGuigan, F. (1977). *Psicología Experimental: enfoque metodológico* (Segunda ed.). (A. Fabre, Trad.) D.F., México: Trillas.

Michell, J. (1986). Measurement scales and statistics: a clash of paradigms. *Psychological Bulletin, 100*(3), 398 - 407.

Montgomery, D. C. (1984). *Design adn Análysis of Experiments* (Second ed.). New York, United States of América: John Wiley & Sons, Inc.

Morris, C., & Maisto, A. (2009). *Psicología* (13a ed.). D.F., México: Pearson Educación.

Muñíz, J. (1998). La medición de lo Psicológico. *Psicothema*, 1 - 21.

Narens, L., & Luce, R. D. (1986). Measurement: The Theory of Numerical Assignments. *Psychological Bulletin*, 166 - 180.

Navarro, A. (1989). *La Psicología y sus múltiples objetos de estudio.* Caracas, Venezuela: Consejo de Desarrollo Científico y Humanístico.

Pagano, R. (2011). *Estadística para las ciencias del comportamiento.* (M. Torres, Trad.) D.F., México: CENGAGE Learning.

Pardo, A., & Ruíz, M. (2005). *Análisis de datos con SPSS 13 Base.* Madrid, España: Mc Graw Hill.

Recalde, L. C. (2009). Los axiomas de la cantidad de Hölder y la fundamentación del continuo lineal. *Matemáticas: Enseñanza Universitaria*, 101 - 121.

Rivadulla, A. (1991). *Probabilidad e inferencia científica.* Barcelona: Anthropos.

Saavedra, N. (2000). La axiomática de Kolmogorov: fundamentos de la teoría de la probabilidad. *Números, 43*, 185-190.

Sáchez Carrion, J. (2001). Estadística, orden natural y orden social. *Papers*, 33 - 46.

Sáiz Roca, M., de la Casa Rivas, G., Dolores Saíz, L., Ruiz, G., & Sánchez, N. (2009). Fundación y establecimiento de la Psicología Científica. En M. Sáiz Roca, *Historia de la Psicología* (págs. 55 - 150). Barcelona: UOC.

Sáiz, M. (2009). Los tiempos de reacción. La ecuación personal y el impulso nervioso. En M. Sáiz Roca, *Historia de la Psicología* (págs. 43 - 46). Barcelona: UOC.

Sáiz, M., & Sáiz, D. (2009). La Psicología Científica Británica. En M. Sáiz Roca, *Historia de la Psicología* (págs. 97 - 113). Barcelona: UOC.

Siegel, S., & Castellan, N. (1995). *Estadística No Paramétrica* (Cuarta ed.). (L. Aragón, & L. Fierros, Trads.) D.F., México: Trillas.

Stahl, S. (2006). The evolution of the normal distribution. *Mathematics Magazine*, 96 - 113.

Stevens, S. (Abril de 1935). The Operational Basis of Psychology. *The American Journal of Psychology, 47*(2), 323-330.

Stevens, S. (1935). The operational definition of psychological concepts. *Psychological Review, 42*(6), 517-527.

Stevens, S. (1946). On the theory of scales of measurement. *Science*, 677 - 680.

Thomas, H. (1982). IQ interval scales, and normal distribution. *Psychological Bulletin*, 198 - 202.

Vargas Sabadías, A. (1995). *Estadística Descriptiva e Inferencial.* Publicaciones de Universidad de Castilla-La Mancha.

Velleman, P., & Wilkinson, L. (1993). Nominal, ordinal, interval and ratio typologies are misleading. *The American Statistician*, 65 - 72.

Walker, H. (1978). Pearson, Karl. En W. Kruskal, & J. Tanur, *International Encyclopedia of Statistics* (Vol. 2, págs. 691-698). New York, U.S.A.: The Free Press.

Wiener, P. (1978). Peirce, Charles Sanders. En W. Kruskal, & J. Tanur, *International Encyclopedia of Statistics* (Vol. 2, págs. 698-702). New York, U.S.A.: The Free Press.