UDS

Mi Universidad

LIBRO

Estadística Descriptiva

Licenciatura en Psicología

Segundo Cuatrimestre

Enero - Abril

Marco Estratégico de Referencia

Antecedentes históricos

Nuestra Universidad tiene sus antecedentes de formación en el año de 1979 con el inicio de actividades de la normal de educadoras "Edgar Robledo Santiago", que en su momento marcó un nuevo rumbo para la educación de Comitán y del estado de Chiapas. Nuestra escuela fue fundada por el Profesor Manuel Albores Salazar con la idea de traer educación a Comitán, ya que esto representaba una forma de apoyar a muchas familias de la región para que siguieran estudiando.

En el año 1984 inicia actividades el CBTiS Moctezuma Ilhuicamina, que fue el primer bachillerato tecnológico particular del estado de Chiapas, manteniendo con esto la visión en grande de traer educación a nuestro municipio, esta institución fue creada para que la gente que trabajaba por la mañana tuviera la opción de estudiar por las tardes.

La Maestra Martha Ruth Alcázar Mellanes es la madre de los tres integrantes de la familia Albores Alcázar que se fueron integrando poco a poco a la escuela formada por su padre, el Profesor Manuel Albores Salazar; Víctor Manuel Albores Alcázar en julio de 1996 como chofer de transporte escolar, Karla Fabiola Albores Alcázar se integró en la docencia en 1998, Martha Patricia Albores Alcázar en el departamento de cobranza en 1999.

En el año 2002, Víctor Manuel Albores Alcázar formó el Grupo Educativo Albores Alcázar S.C. para darle un nuevo rumbo y sentido empresarial al negocio familiar y en el año 2004 funda la Universidad Del Sureste.

La formación de nuestra Universidad se da principalmente porque en Comitán y en toda la región no existía una verdadera oferta Educativa, por lo que se veía urgente la creación de una institución de Educación superior, pero que estuviera a la altura de las exigencias de los

jóvenes que tenían intención de seguir estudiando o de los profesionistas para seguir preparándose a través de estudios de posgrado.

Nuestra Universidad inició sus actividades el 18 de agosto del 2004 en las instalaciones de la 4ª avenida oriente sur no. 24, con la licenciatura en Puericultura, contando con dos grupos de cuarenta alumnos cada uno. En el año 2005 nos trasladamos a nuestras propias instalaciones en la carretera Comitán – Tzimol km. 57 donde actualmente se encuentra el campus Comitán y el corporativo UDS, este último, es el encargado de estandarizar y controlar todos los procesos operativos y educativos de los diferentes campus, así como de crear los diferentes planes estratégicos de expansión de la marca.

Misión

Satisfacer la necesidad de Educación que promueva el espíritu emprendedor, aplicando altos estándares de calidad académica, que propicien el desarrollo de nuestros alumnos, Profesores, colaboradores y la sociedad, a través de la incorporación de tecnologías en el proceso de enseñanza-aprendizaje.

Visión

Ser la mejor oferta académica en cada región de influencia, y a través de nuestra plataforma virtual tener una cobertura global, con un crecimiento sostenible y las ofertas académicas innovadoras con pertinencia para la sociedad.

Valores

- Disciplina
- Honestidad
- Equidad
- Libertad

Escudo

El escudo del Grupo Educativo Albores Alcázar S.C. está constituido por tres líneas curvas que nacen de izquierda a derecha formando los escalones al éxito. En la parte superior está situado un cuadro motivo de la abstracción de la forma de un libro abierto.

Eslogan

"Mi Universidad"

ALBORES

Es nuestra mascota, un Jaguar. Su piel es negra y se distingue por ser líder, trabaja en equipo y obtiene lo que desea. El ímpetu, extremo valor y fortaleza son los rasgos que distinguen.

Estadística Descriptiva

Objetivo de la materia:

Conocer y aplicar correctamente los procedimientos de análisis de datos que más habitualmente son utilizados en el proceso de obtención de información científica en el ámbito de la Psicología.

Criterios de evaluación:

No	Concepto	Porcentaje
1	Trabajos Escritos	10%
2	Actividades Áulicas	20%
3	Trabajos en plataforma Educativa	20%
4	Examen	50%
Total de Criterios de evaluación		100%

INDICE

Unidad 1

INTRODUCCION A LA ESTADISTICA APLICADA A LA PSICOLOGIA.

1.1. La Estadística

1.2. El método científico y la Estadística

1.3. ¿Por qué la Estadística en el grado de Psicología?

1.4. Algunos conceptos básicos de Estadística

1.5. Metodologías de investigación y Estadística

1.6. Estadística descriptiva y estadística inferencial

1.6.1. Población y muestra

1.6.2. Parámetros y estadísticos

Unidad 2

ORGANIZACIÓN Y REPRESENTACIÓN GRÁFICA DE LOS DATOS

2.1. La distribución de frecuencias

2.2. La representación gráfica de una distribución de frecuencias

2.3. Propiedades de las distribuciones de frecuencias

2.4. Estadísticos de posición grupal

2.4.1. Variables categóricas: la moda

2.4.2. Variables ordinales: la mediana, el mínimo y el máximo, los cuantíales

2.4.3. Variables cuantitativas: la media y sus alternativas robustas

2.5. Estadísticos de dispersión

2.5.1. Variables categóricas: la razón de variación y el índice de variación cualitativa

2.5.2. Variables ordinales: el rango y el rango intercuartil

2.5.3. Variables cuantitativas: la varianza, la desviación típica y el coeficiente de variación.

Unidad 3

ESTADÍSTICOS DE FORMA DE LA DISTRIBUCIÓN

3.1 Asimetría

3.2. Apuntamiento

3.3. Estadísticos de posición individual

3.1.1. Los porcentajes acumulados

3.1.2. Las puntuaciones típicas

3.1.3. Las escalas derivadas

3.4. Organización y representación gráfica de datos multivariados

3.4.1. La distribución conjunta multivariada

3.4.2. La tabla de contingencia

3.4.3. Representaciones gráficas

3.4.3.1. El caso de dos variables categóricas

3.4.3.2. El caso de dos variables cuantitativas

3.4.3.3. El caso de una variable categórica y una variable cuantitativa

Unidad 4

ESTADÍSTICOS DE ASOCIACIÓN ENTRE VARIABLES

4.1. Concepto de asociación entre variables

4.2. Midiendo la asociación entre dos variables

4.2.1. El caso de dos variables categóricas

4.2.2. El caso de una variable categórica y una cuantitativa

4.2.3. El caso de dos variables cuantitativas

4.5. El modelo de regresión lineal

4.5.1. Conceptos básicos sobre el análisis de regresión lineal

4.5.2. Ajuste de la recta de regresión

4.5.3. Bondad de ajuste del modelo de regresión

4.6 .La estadística inferencial: algunos conceptos previos

4.6.1. Teoría de la probabilidad

4.6.2. Variables aleatorias

4.6.3. Modelos teóricos de distribución de probabilidad

4.6.3.1. La distribución binomial

4.6.3.2. La distribución o curva normal

4.6.4. La selección de la muestra

UNIDAD I

1.1.- Estadística descriptiva:

Describe, analiza y representa un grupo de datos utilizando métodos numéricos y gráficos que resumen y presentan la información contenida en ellos. Se puede definirse como aquel método que contiene la recolección, organización, presentación y resumen de una serie de datos. El mencionado resumen puede ser tabular, gráfico o numérico. El análisis que se realiza se limita en sí mismo a los datos recolectados y no se puede realizar inferencia alguna o generalizaciones algunas, acerca de la población de donde provienen esos datos estadísticos.

Una de las ramas de la Estadística más accesible a la mayoría de la población es la Descriptiva. Esta se dedica única y exclusivamente al ordenamiento y tratamiento mecánico de la información para su presentación por medio de tablas y de representaciones gráficas, así como de la obtención de algunos parámetros útiles para la explicación de la información.

La estadística descriptiva analiza, estudia y describe a la totalidad de los individuos de una población, su finalidad es obtener información, analizarla, elaborarla y simplificarla lo necesario para que pueda ser interpretada cómoda y rápidamente y, por tanto, pueda utilizarse eficazmente para el fin que se desee.

El proceso que sigue la estadística descriptiva para el estudio de una cierta población consta de los siguientes pasos:

1 Selección de caracteres dignos de ser estudiados.

2 Mediante encuesta o medición, obtención del valor de cada individuo en los caracteres seleccionados.

3 Elaboración de tablas de frecuencias, mediante la adecuada clasificación de los individuos dentro de cada carácter.

4 Representación gráfica de los resultados (elaboración de gráficos estadísticos).

5 Obtención de parámetros estadísticos, números que sintetizan los aspectos más relevantes de una distribución estadística.

Estadística inferencial:

Es aquella rama de la estadística que apoyándose en el cálculo de probabilidades y a partir de datos muéstrales, efectúa estimaciones, decisiones, predicciones u otras generalizaciones sobre un conjunto mayor de datos. Puede definirse como aquella rama de la estadística que hace posible la estimación de una característica de una población o la toma de una decisión referente a una población, fundamentándose sólo en los resultados de la muestra.

La estadística Inferencial, por otro lado, se refiere a la rama de la estadística que trata de los procesos inferenciales, la que a su vez vislumbra la teoría de estimación y prueba de hipótesis. Uno de los primordiales aspectos de la inferencia estadística es el proceso que radica en utilizar estadísticos muéstrales para adquirir conclusiones sobre los verdaderos parámetros de la población.

Los requerimientos de los métodos de la inferencia estadística se originan de la necesidad del muestreo. Al tornarse muy grande una población, comúnmente resulta demasiado costoso, prolongado en el tiempo y complicado obtener información de la población completa. Las decisiones con respecto a las características de la población se deben basar en la información contenida en una muestra de esa población. La teoría de la probabilidad suministra él vínculo, determinando la probabilidad de que los resultados provenientes de la muestra reflejen los resultados que se obtendrían de la población.

La fidelidad de cualquier estimación tiene una importancia enorme. Esta precisión depende en gran parte de la forma de tomar la muestra y de la atención que se ponga en que esta muestra suministre una imagen fiable de la población, pero casi nunca la muestra representa la población en toda su plenitud, y de ello resultará un error maestral.

1.2.- Finalidad de la estadística

La estadística es una ciencia o método científico que en la actualidad es considerada como un poderoso auxiliar en las investigaciones científicas, que le permite a ésta aprovechar el material cuantitativo.

Historia de la estadística

Desde el inicio de la civilización han existido formas sencillas de estadística, puesto que en la antigüedad se utilizaban representaciones gráficas y otros símbolos en pieles, rocas, palos de madera y paredes de cuevas para contar el número de personas, animales o ciertas cosas que eran de importancia en aquellas civilizaciones. El término estadístico es ampliamente percibido y pronunciado a diario desde diversos sectores activos de la sociedad. No obstante, hay una gran diferencia entre el sentido del término cuando se utiliza en el lenguaje corriente, generalmente al anteceder una citación de carácter numérico, y lo que la estadística significa como ciencia.

La razón o razones que motivaron al hombre en un momento de su desarrollo a tomar en cuenta datos con propósitos estadísticos, posiblemente se encuentra si se toma en cuenta que es difícil suponer un organismo social, sea cual fuere la época, sin la necesidad, casi instintiva, de recoger aquellos hechos que aparecen como actos esenciales de la vida; y así, al ubicarnos en una etapa del desarrollo de la estadística podemos especular que se convirtió en una aritmética estatal para asistir al gobernante que necesitaba conocer la riqueza y el número de los súbditos entre otros, con el objeto de recaudar impuestos o presupuestar la guerra.

Desde los comienzos de la civilización han existido formas sencillas de estadística, pues ya se utilizaban representaciones gráficas y otros símbolos en pieles, rocas, palos de madera y paredes de cuevas para contar el número de personas, animales o cosas. Hacia el año 3000 a.C. los babilonios usaban pequeñas tablillas de arcilla para recopilar datos sobre la producción agrícola y sobre las especies vendidas o cambiadas mediante trueque.

1.3.- conceptos básicos

Universo:

En estadística es el nombre especifico que recibe particularmente en la investigación social la operación dentro de la delimitación del campo de investigación que tienen por objeto la determinación del conjunto de unidades de observaciones del conjunto de unidades de observación que van a ser investigadas. Para muchos investigadores él termino universo y población son sinónima. En general, el universo es la totalidad de elementos o características que conforman el ámbito de un estudio o investigación.

Población:

En estadística el concepto de población va más allá de lo que comúnmente se conoce como tal. En términos estadísticos, población es un conjunto finito o infinito de personas, animales o cosas que presentan características comunes, sobre los cuales se quiere efectuar un estudio determinado. En otras palabras, la población se define como la totalidad de los valores posibles (mediciones o conteos) de una característica particular de un grupo especificado de personas, animales o cosas que se desean estudiar en un momento determinado. Así, se puede hablar de la población de habitantes de un país, de la población de estudiantes universitarios de la zona sur del Estado Anzoátegui, de la población de casas de la Urbanización Los Ríos de la ciudad de El Tigre, el rendimiento académico de los estudiantes del IUTJAA, el número de carros marca Corola de la ciudad de El Tigre, la estatura de un grupo alumnos del IUTJAA, la talla, etc.

Muestra:

La muestra es un subconjunto de la población, seleccionado de tal forma, que sea representativo de la población en estudio, obteniéndose con el fin de investigar alguna o algunas de las propiedades de la población de la cual procede. En otras palabras es una parte de la población que sirve para representarla. Según el DRAE, es una parte o porción extraída de un conjunto por métodos que permiten considerarla como representativa del mismo.

Entonces, una muestra no es más que una parte de la población que sirve para representarla. La muestra debe obtenerse de la población que se desea estudiar; una muestra debe ser definida sobre la base de la población determinada, y las conclusiones que se obtengan de dicha muestra sólo podrán referirse a la población en referencia.

Muestreo:

Es el procedimiento mediante el cual se obtiene una o más muestras de una población determinada. Existen dos tipos de muestreos a saber:

Los Parámetros:

Son cualquiera característica que se pueda medir y cuya medición se lleve a cabo sobre todos los elementos que integran una población determinada, los mismos suelen representarse con letras griegas. El valor de un parámetro poblacional es un valor fijo en un momento dado. Ejemplo: La media Aritmética = m (miu), La desviación Típica = s, (Sigma) etcétera.

Dato estadístico:

Es un conjunto de valores numéricos que tienen relación significativa entre sí. Los mismos pueden ser comparados, analizados e interpretados en una investigación cualquiera. Se puede afirmar que son las expresiones numéricas obtenidas como consecuencia de observar un individuo de la población; por lo tanto, son las características que se han tomado en cuenta de cualquiera población para una investigación determinada.

Frecuencia:

La frecuencia es el número de veces que se repite (aparece) el mismo dato estadístico en un conjunto de observaciones de una investigación determinada, las frecuencias se les designan con las letras fi, y por lo general se les llaman frecuencias absolutas.

Distribución de Frecuencia:

En estadística existe una relación con cantidades, números agrupados o no, los cuales poseen entre sí características similares. Existen investigaciones relacionadas con los precios de los productos de la dieta diaria, la estatura y el peso de un grupo de individuos, los salarios de los empleados, los grados de temperatura del medio ambiente, las calificaciones de los estudiantes, etc., que pueden adquirir diferentes valores gracias a una unidad apropiada, que recibe el nombre de variable. La representación numérica de las variables se denomina dato estadístico. La distribución de frecuencia es una disposición tabular de datos estadísticos, ordenados ascendente o descendentemente, con la frecuencia (fi) de cada dato. Las distribuciones de frecuencias pueden ser para datos no agrupados y para datos agrupados o de intervalos de clase.

Distribución de frecuencia para datos no Agrupados:

Es aquella distribución que indica las frecuencias con que aparecen los datos estadísticos, desde el menor de ellos hasta el mayor de ese conjunto sin que se haya hecho ninguna modificación al tamaño de las unidades originales. En estas distribuciones cada dato mantiene su propia identidad después que la distribución de frecuencia se ha elaborado. En estas distribuciones los valores de cada variable han sido solamente reagrupados, siguiendo un orden lógico con sus respectivas frecuencias.

Distribución de frecuencia de clase o de datos Agrupados:

Es aquella distribución en la que las disposiciones tabulares de los datos estadísticos se encuentran ordenados en clases y con la frecuencia de cada clase; es decir, los datos originales de varios valores adyacentes del conjunto se combinan para formar un intervalo de clase. No existen normas establecidas para determinar cuándo es apropiado utilizar datos agrupados o datos no agrupados; sin embargo, se sugiere que cuando el número total de datos (N) es igual o superior 50 y además el rango o recorrido de la serie de datos es mayor de 20, entonces, se utilizará la distribución de frecuencia para datos agrupados, también se utilizará este tipo de distribución cuando se requiera elaborar gráficos lineales como el histograma, el polígono de frecuencia o la ojiva.

La razón fundamental para utilizar la distribución de frecuencia de clases es proporcionar mejor comunicación acerca del patrón establecido en los datos y facilitar la manipulación de los mismos. Los datos se agrupan en clases con el fin de sintetizar, resumir, condensar o hacer que la información obtenida de una investigación sea manejable con mayor facilidad.

Componentes de una distribución de frecuencia de clase

1.- Rango o Amplitud total (recorrido).- Es el límite dentro del cual están comprendidos todos los valores de la serie de datos, en otras palabras, es el número de diferentes valores que toma la variable en un estudio o investigación dada. Es la diferencia entre el valor máximo de una variable y el valor mínimo que ésta toma en una investigación cualquiera. El rango es el tamaño del intervalo en el cual se ubican todos los valores que pueden tomar los diferentes datos de la serie de valores, desde el menor de ellos hasta el valor mayor estando incluidos ambos extremos. El rango de una distribución de frecuencia se designa con la letra R.

2.- Clase o Intervalo de clase.- Son divisiones o categorías en las cuales se agrupan un conjunto de datos ordenados con características comunes. En otras palabras, son fraccionamientos del rango o recorrido de la serie de valores para reunir los datos que presentan valores comprendidos entre dos límites. Para organizar los valores de la serie de datos hay que determinar un número de clases que sea conveniente. En otras palabras, que ese número de intervalos no origine un número pequeño de clases ni muy grande. Un número de clases pequeño puede ocultar la naturaleza natural de los valores y un número muy alto puede provocar demasiados detalles como para observar alguna información de gran utilidad en la investigación.

Tamaño de los Intervalos de Clase

Los intervalos de clase pueden ser de tres tipos, según el tamaño que estos presenten en una distribución de frecuencia:

a) Clases de igual tamaño, b) clases desiguales de tamaño y c) clases abiertas.

3.- Amplitud de Clase, Longitud o Ancho de una Clase

La amplitud o longitud de una clase es el número de valores o variables que concurren a una clase determinada. La amplitud de clase se designa con las letras lc. Existen diversos criterios para determinar la amplitud de clases, ante esa diversidad de criterios, se ha considerado que lo más importante es dar un ancho o longitud de clase a todos los intervalos de tal manera que respondan a la naturaleza de los datos y al objetivo que se persigue y esto se logra con la práctica.

4.-Punto medio o Marca de clase

El centro de la clase, es el volar de los datos que se ubica en la posición central de la clase y representa todos los demás valores de esa clase. Este valor se utiliza para el cálculo de la media aritmética.

5.-Frecuencia de clase

La frecuencia de clase se le denomina frecuencia absoluta y se le designa con las letras fi. Es el número total de valores de las variables que se encuentran presente en una clase determinada, de una distribución de frecuencia de clase.

6.- Frecuencia Relativa

La frecuencia relativa es aquella que resulta de dividir cada uno de los fi de las clases de una distribución de frecuencia de clase entre el número total de datos(N) de la serie de valores. Estas frecuencias se designan con las letras fr; si cada fr se multiplica por 100 se obtiene la frecuencia relativa porcentual (fr %).

7.-Frecuencias acumuladas

Las frecuencias acumuladas de una distribución de frecuencias son aquellas que se obtienen de las sumas sucesivas de las fi que integran cada una de las clases de una distribución de frecuencia de clase, esto se logra cuando la acumulación de las frecuencias se realiza tomando en cuenta la primera clase hasta alcanzar la ultima. Las frecuencias acumuladas se designan con las letras fa. Las frecuencias acumuladas pueden ser menor que (fa< que) y frecuencias acumuladas mayor que (fa>que).

8.- Frecuencia acumulada relativa

La frecuencia acumulada relativa es aquella que resulta de dividir cada una de las fa de las diferentes clases que integran una distribución de frecuencia de clase entre el número total de datos (N) de la serie de valores, estas frecuencias se designan con las letras far. Si las frace multiplican por 100 se obtienen las frecuencias acumuladas relativas porcentuales y las mismas se designan así: far %.

La mediana

La mediana (Md) es una medida de posición que divide a la serie de valores en dos partes iguales, un cincuenta por ciento que es mayor o igual a esta y otro cincuenta por ciento que es menor o igual que ella. Es por lo tanto, un parámetro que esta en el medio del ordenamiento o arreglo de los datos organizados, entonces, la mediana divide la distribución en una forma tal que a cada lado de la misma queda un número igual de datos.

Para encontrar la mediana en una serie de datos no agrupados, lo primero que se hace es ordenar los datos en una forma creciente o decreciente y luego se ubica la posición que esta ocupa en esa serie de datos; para ello hay que determinar si la serie de datos es par o impar, luego el número que se obtiene indica el lugar o posición que ocupa la mediana en la serie de valores, luego la mediana será el número que ocupe el lugar de lo posición encontrada.

La moda

La moda es la medida de posición que indica la magnitud del valor que se presenta con más frecuencia en una serie de datos; es pues, el valor de la variable que más se repite en un conjunto de datos. De las medias de posición la moda es la que se determina con mayor facilidad, ya que se puede obtener por una simple observación de los datos en estudio, puesto que la moda es el dato que se observa con mayor frecuencia. La moda se designa con las letras Mo.

Desviación típica o estándar

Es la medida de dispersión más utilizada en las investigaciones por ser la más estable de todas, ya que para su cálculo se utilizan todos los desvíos con respecto a la media aritmética de las observaciones, y además, se toman en cuenta los signos de esos desvíos. Se le designa con la letra castellana S cuando se trabaja con una muestra y con la letra griega minúscula s (Sigma) cuando se trabaja con una población. Es importante destacar que cuando se hace referencia a la población él número de datos se expresa con N y cuando se refiere a la muestra él número de datos se expresa con n. La desviación típica se define como:

Interpretación de la desviación estándar

La desviación típica como medida absoluta de dispersión, es la que mejor nos proporciona la variación de los datos con respecto a la media aritmética, su valor se encuentra en relación directa con la dispersión de los datos, a mayor dispersión de ellos, mayor desviación típica, y a menor dispersión, menor desviación típica.

Varianza

Es otra de las variaciones absolutas y la misma se define como el cuadrado de la desviación típica; viene expresada con las mismas letras de la desviación típica pero elevada al cuadrado, así S2 y s2. Las fórmulas para calcular la varianza son las mismas utilizadas por la desviación típica, exceptuando las respectivas raíces, las cuales desaparecen al estar elevados el primer miembro al cuadrado

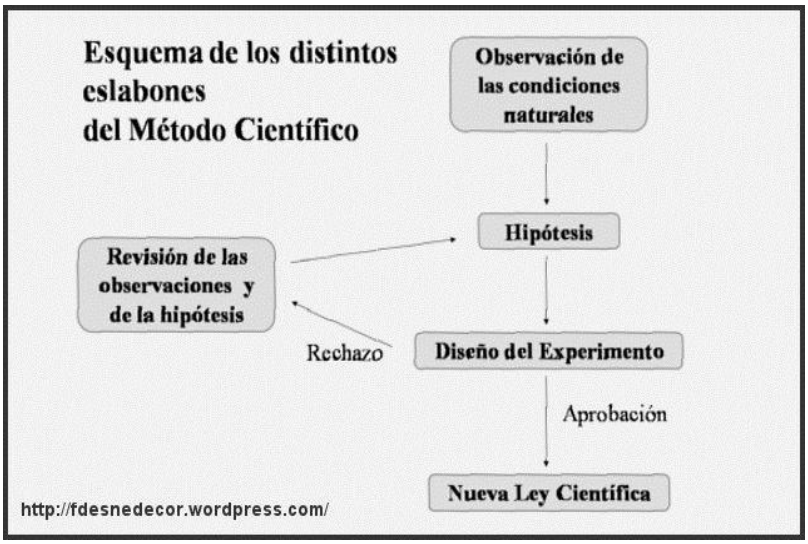
La Estadística dentro del Método Científico

La estadística no se puede utilizar como una caja mágica para extraer certezas, donde se introducen datos y se extraen leyes. La estadística, en el contexto de probabilidades y técnicas de inferencia, es incapaz por sí misma de suplantar al Método Científico, sólo es un gran apoyo.

¿Cómo ayuda la estadística en el Método Científico?

Definimos el Método Científico como un método o conjunto sistematizado de procesos en los que se basa la ciencia para explicar cualquier fenómeno y las leyes que los administran.

En la siguiente imagen os muestro, muy esquematizado, el proceso que se sigue al aplicar el Método Científico.



La estadística descriptiva es la herramienta más útil en la etapa de **observación**, ya que nos permite extraer información para realizar nuestras hipótesis fundadas en estos resultados. También es utilizada para valorar los resultados del **experimento**.

La estadística analítica se utiliza a partir de la **observación**, ya que dependiendo de los datos observados, se utilizará una técnica u otra, y por supuesto en el proceso del **experimento**, ya que su diseño dependerá en cierta medida de las técnicas estadísticas más apropiadas, además, la estadística analítica es el primer y principal razonamiento válido.

Como vemos, la estadística proporciona un gran apoyo al Método Científico en las fases de observación y experimentación, pero en el proceso de hipótesis y en el de la obtención de una ley científica son otras las bases

1.3.- ¿Por qué es útil la estadística en Psicología?

Hay una asignatura que suele llamar poderosamente la atención en aquellos que empiezan la carrera/grado de psicología. Se trata de la estadística. Pensando que se han dejado atrás los números, de pronto aparece esta inquietante materia. Pero, ¿para qué sirve la estadística? ¿Por qué es útil para un psicólogo o para alguien que se interese por la psicología?

Cursando el grado en psicología nos vamos a encontrar con asignaturas como: "Fundamentos de investigación", "Análisis de datos", "Diseños de investigación", "Psicometría" o similares. Hablamos de asignaturas cuyo plan se asienta directamente en la estadística. Por otro lado, este tipo de asignaturas no suelen ser las más apreciadas de la carrera ya que ningún estudiante la comienza con la motivación de cursar las mismas.

En este artículo trataremos de responder a aquellas preguntas que muchos alumnos y curiosos se hacen al encontrarse con ellas. Para desarrollar estas respuestas, hablaremos acerca de la Psicología como ciencia y la utilidad de la estadística en Psicología.

La investigación en psicología y su metodología

Recordemos que la psicología es una ciencia. Todas las conclusiones que se derivan de esta disciplina proceden (o deberían proceder) de la aplicación de un sistema exhaustivo y fiable denominado método científico. Este método se basa en la acumulación progresiva de evidencia utilizando diferentes recursos matemáticos.

Concepto de Metodología Estadística

La metodología y la estadística son dos ramas de conocimientos importantes para llevar a cabo un estudio científico. La metodología permite diseñar el estudio con un soporte en sus características e importancia; establece cuales son las variables que influyen en el estudio, las técnicas de control, tipos y planteamientos de la investigación. La estadística también es importante en la conducta del individuo, permite organizar, resumir, recopilar, analizar y

representar los datos y la preparación de conclusiones válidas, además proporciona tomar decisiones lógicas fundamentadas en el análisis estadístico.

Metodología Estadística

Los métodos estadísticos posibilitan ejecutar la investigación y conseguir conclusiones con base de cierta disciplina en los estudios psicológicos, siendo importantes en las habilidades del psicólogo sin interesar cual sea la dirección o campo de estudio.

Mientras la tecnología evoluciona el profesional se hace frente con las informaciones tipo descriptiva pero más cuantitativa. En tal sentido la estadística en un método importante para manejar y organizar los estudios cuantitativos con el objetivo de poder explicar debidamente los resultados obtenidos. La planificación, elaboración e interpretación de un estudio va de la mano con la metodología estadística.

Objetivos de la Metodología Estadística.

◦Plantear: en todo tipo de investigación cuantitativa es principal, plantear detalladamente el estudio.

◦Debatir: se refiere a opinar sobre cualquier tema dichos anteriormente, o sobre otro tema interesante.

◦Solucionar: proponer una solución específica que ayude a cualquier tema en estudio.

◦Unir: el fundamento principal de la estadística es la cuantificación de los elementos desde una muestra o cifra de observaciones.

También la metodología estadística se refiere a la aplicación de procedimientos estadísticos dentro de una población determinada. Una población con cualquier número de elementos puede ser centro de observación. Con el objetivo de describir el grupo de datos obtenidos y determinar las características de las observaciones de la investigación.

La metodología estadística también es aplicable en campos tales como las ciencias físicas y naturales, la ingeniería, en el proceso de control de la calidad, en las ciencias sociales, ciencias de la salud y en las organizaciones gubernamentales. La aplicación es muy amplia por la habilidad de poder manejar elevados conjuntos de datos numéricos, por la utilización de la tecnología.

La Metodología Estadística se divide en dos áreas:

◦Estadística descriptiva: se encarga de representar, observar y analizar las características de un grupo de datos que se pueden desarrollar a través de tablas, gráficos o valores numéricos por ejemplo, cuál es la tasa de pobreza y cuántas personas viven en un país.

◦Estadística inferencial o inductiva: se utiliza para sacar conclusiones basándose en los datos obtenidos de una muestra estudiada.

Para que la estadística se lleve a cabo realiza operaciones numéricas en que la población, la muestra, el muestreo y la observación son fundamentales.

1.- La población en esta rama de la matemática es el conjunto de personas que habitan en un país, región o ciudad. También se refiere a objetos que tengan características similares.

2.- La muestra tiene como significado la selección de una parte de la población para realizar el estudio y el resultado que se obtenga sólo se referirá a la muestra seleccionada. Una de las características que tiene la muestra es que es un procedimiento más fácil de tener conclusiones que un estudio de una población mayor.

3.- El muestreo es un método que permite conseguir uno o más muestras de una población. El muestreo consta de dos procedimientos para la elección de las muestras: el muestreo aleatorio y no aleatorio.

1.6.- Estadística descriptiva y estadística inferencial

La estadística descriptiva e inferencial forman parte de las dos ramas fundamentales en las que se divide la estadística, la ciencia exacta que se encarga de extraer información de diversas variables, midiéndolas, controlándolas y comunicándolas en caso de que haya incertidumbre.

De esta manera, la estadística tiene como objetivo cuantificar y controlar comportamientos y eventos tanto científicos como sociales.

Ramas de la estadística

La estadística descriptiva se encarga de resumir la información derivada de los datos relativos a una población o muestra. Su objetivo es sintetizar dicha información de forma precisa, sencilla, clara y ordenada (Santillán, 2016).

Es así como la estadística descriptiva puede señalar los elementos más representativos de un grupo de datos, conocidos como datos estadísticos. En pocas palabras, este tipo de estadística se encarga de hacer descripciones de dichos datos. Por su parte, la estadística inferencial se encarga de hacer inferencias sobre los datos recogidos. Arroja conclusiones diferentes a lo mostrado por los datos en sí. Este tipo de estadística va más allá de la simple recopilación de información, relacionando cada dato con fenómenos que pueden alterar su comportamiento. La estadística inferencial llega a conclusiones relevantes sobre una población a partir de análisis de una muestra. Por lo tanto, siempre debe calcular un margen de error dentro de sus conclusiones.

Estadística descriptiva

Es la rama de la estadística más popular y conocida. Su principal objetivo es el de analizar variables y posteriormente describir los resultados obtenidos de dicho análisis.

La estadística descriptiva busca describir un grupo de datos con el objetivo de señalar de forma precisa las características que definen a dicho grupo.

Se puede decir que esta rama de la estadística es la responsable de ordenar, resumir y clasificar los datos resultantes del análisis de la información derivada de un grupo.

Algunos ejemplos de la estadística descriptiva pueden incluir los censos de población de un país en un año determinado o el número de personas que fueron recibidas en un hospital dentro de un margen de tiempo determinado.

Categorías

Existen ciertos conceptos y categorías que forman parte exclusivamente del campo de la estadística descriptiva. Algunos se listan a continuación:

– Dispersión: es la diferencia que existe entre los valores incluidos dentro de una misma variable. La dispersión también incluye el promedio de dichos valores.

– Promedio: es el valor que resulta de la sumatoria de todos los valores incluidos en una misma variable y la posterior división del resultado por el número de datos incluidos en la sumatoria. Se define como la tendencia central de una variable.

– Sesgo o curtosis: es la medida que indica qué tan inclinada es una curva. Es el valor que indica la cantidad de elementos que se encuentran más próximos al promedio. Existen tres tipos diferentes de sesgo (leptocúrtica, mesocúrtica y platicúrtica), cada uno de ellos indica qué tan alta es la concentración de datos alrededor del promedio.

– Gráficos: son la representación gráfica de los datos obtenidos del análisis. Usualmente, son utilizados diferentes tipos de gráficos estadísticos, incluidos los de barras, circulares, lineales, poligonales, entre otros,

– Asimetría: es el valor que muestra la manera como los valores de una misma variable se encuentran repartidos con relación al promedio. Puede ser negativa, simétrica o positiva (Formulas, 2017).}

Estadística inferencial

Es el método de análisis utilizado para hacer inferencias sobre una población, teniendo en cuenta los datos arrojados por la estadística descriptiva sobre un segmento de la misma muestra. Dicho segmento debe ser elegido bajo criterios rigurosos.

La estadística inferencial se vale del uso de herramientas especiales que le permiten hacer afirmaciones globales sobre la población, a partir de la observación de una muestra.

Los cálculos llevados a cabo por este tipo de estadística son aritméticos y siempre dan cabida a un margen de error, cosa que no sucede con la estadística descriptiva, que se encarga de analizar a la totalidad de la población.

Por tal motivo, la estadística inferencial requiere de hacer uso de modelos de probabilidades que le permiten inferir conclusiones sobre una población amplia basándose únicamente en lo que una parte de ella

Según la estadística descriptiva es posible obtener datos de una población general a partir del análisis de una muestra conformada por individuos seleccionados de forma aleatoria.

Categorías

La estadística inferencial puede ser clasificada en dos grandes categorías descritas a continuación:

– Pruebas de hipótesis: como su nombre lo indica, consiste en poner a prueba aquello que se concluyó sobre una población a partir de los datos arrojados por la muestra.

– Intervalos de confianza: estos son los rangos de valores señalados dentro de la muestra de una población para identificar una característica relevante y desconocida (Minitab Inc., 2017). Por su naturaleza aleatoria, son los que permiten reconocer un margen de error dentro de cualquier análisis estadístico inferencial.

Diferencias entre la estadística descriptiva y la inferencial

La principal diferencia entre la estadística descriptiva y la inferencial radica en que la primera busca ordenar, resumir y clasificar los datos derivados del análisis de variables.

Por su parte, la estadística inferencial, lleva a cabo deducciones con base a los datos previamente obtenidos.

Por otro lado, la estadística inferencial depende del trabajo de la estadística descriptiva para llevar a cabo sus inferencias.

De este modo, la estadística descriptiva constituye la base sobre la que posteriormente la estadística inferencial llevará a cabo su trabajo.

También es importante señalar que la estadística descriptiva se utiliza para analizar tanto poblaciones (grupos numerosos) como muestras (subconjuntos de las poblaciones).

Mientras que la estadística inferencial se encarga de estudiar muestras a partir de las cuales busca llegar a conclusiones sobre la población general.

Otra diferencia entre estos dos tipos de estadística radica en que la estadística descriptiva únicamente se centra en la descripción de los datos obtenidos, sin asumir que estos tengan ninguna propiedad relevante.

Ésta no va más allá de lo que los mismos datos obtenidos puedan señalar. Por su parte, la estadística inferencial cree que todos los datos derivados de cualquier análisis estadístico dependen de fenómenos externos y aleatorios que pueden alterar su valor.

Población y muestra

Las estadísticas de por sí no tienen sentido si no se consideran o se relacionan dentro del contexto con que se trabajan. Por lo tanto es necesario entender los conceptos de población y de muestra para lograr comprender mejor su significado en la investigación educativa o social que se lleva a cabo.

POBLACIÓN - es el conjunto total de individuos, objetos o medidas que poseen algunas características comunes observables en un lugar y en un momento determinado. Cuando se vaya a llevar a cabo alguna investigación debe de tenerse en cuenta algunas características esenciales al seleccionarse la población bajo estudio.

Entre éstas tenemos:

1. Homogeneidad - que todos los miembros de la población tengan las mismas características según las variables que se vayan a considerar en el estudio o investigación.
2. Tiempo - se refiere al período de tiempo donde se ubicaría la población de interés. Determinar si el estudio es del momento presente o si se va a estudiar a una población de cinco años atrás o si se van a entrevistar personas de diferentes generaciones.
3. Espacio - se refiere al lugar donde se ubica la población de interés. Un estudio no puede ser muy abarcador y por falta de tiempo y recursos hay que limitarlo a un área o comunidad en específico.

4. Cantidad - se refiere al tamaño de la población. El tamaño de la población es sumamente importante porque ello determina o afecta al tamaño de la muestra que se vaya a

seleccionar, además que la falta de recursos y tiempo también nos limita la extensión de la población que se vaya a investigar.

MUESTRA - la muestra es un subconjunto fielmente representativo de la población.
Hay diferentes tipos de muestreo. El tipo de muestra que se seleccione dependerá de la calidad y cuán representativo se quiera sea el estudio de la población.

1. ALEATORIA - cuando se selecciona al azar y cada miembro tiene igual oportunidad de ser incluido.
2. ESTRATIFICADA - cuando se subdivide en estratos o subgrupos según las variables o características que se pretenden investigar. Cada estrato debe corresponder proporcionalmente a la población.
3. SISTEMÁTICA - cuando se establece un patrón o criterio al seleccionar la muestra. Ejemplo: se entrevistará una familia por cada diez que se detecten.

El muestreo es indispensable para el investigador ya que es imposible entrevistar a todos los miembros de una población debido a problemas de tiempo, recursos y esfuerzo. Al seleccionar una muestra lo que se hace es estudiar una parte o un subconjunto de la población, pero que la misma sea lo suficientemente representativa de ésta para que luego pueda generalizarse con seguridad de ellas a la población.

El tamaño de la muestra depende de la precisión con que el investigador desea llevar a cabo su estudio, pero por regla general se debe usar una muestra tan grande como sea posible de acuerdo a los recursos que haya disponibles. Entre más grande la muestra mayor posibilidad de ser más representativa de la población.
En la investigación experimental, por su naturaleza y por la necesidad de tener control sobre las variables, se recomienda muestras pequeñas que suelen ser de por lo menos 30 sujetos.
En la investigación descriptiva se emplean muestras grandes y algunas veces se recomienda seleccionar de un 10 a un 20 por ciento de la población accesible.

Las razones para estudiar muestras en lugar de poblaciones son diversas y entre ellas podemos señalar

a. Ahorrar tiempo. Estudiar a menos individuos es evidente que lleva menos tiempo.
b. Como consecuencia del punto anterior ahorraremos costes.
c. Estudiar la totalidad de los pacientes o personas con una característica determinada en muchas ocasiones puede ser una tarea inaccesible o imposible de realizar.
d. Aumentar la calidad del estudio. Al disponer de más tiempo y recursos, las observaciones y mediciones realizadas a un reducido número de individuos pueden ser más exactas y plurales que si las tuviésemos que realizar a una población.
e. La selección de muestras específicas nos permitirá reducir la heterogeneidad de una población al indicar los criterios de inclusión y/o exclusión.

1.6.2.- **PARAMETROS ESTADISTICOS**

Un parámetro estadístico es un número que se obtiene a partir de los datos de una distribución estadística.

Los parámetros estadísticos sirven para sintetizar la información dada por una tabla o por una gráfica.

Tipos de parámetros estadísticos

Hay tres tipos parámetros estadísticos:

De centralización

De posición

De dispersión

Medidas de centralización

Nos indican en torno a qué valor (centro) se distribuyen los datos.

Las medidas de centralización son:

Media aritmética

La media es el valor promedio de la distribución.

Mediana

La mediana es la puntación de la escala que separa la mitad superior de la distribución y la inferior, es decir divide la serie de datos en dos partes iguales.

Moda

La moda es el valor que más se repite en una distribución.

Medidas de posición

Las medidas de posición dividen un conjunto de datos en grupos con el mismo número de individuos.

Para calcular las medidas de posición es necesario que los datos estén ordenados de menor a mayor.

Las medidas de posición son:

Cuartiles

Los cuartiles dividen la serie de datos en cuatro partes iguales.

Deciles

Los deciles dividen la serie de datos en diez partes iguales.

Percentiles

Los percentiles dividen la serie de datos en cien partes iguales.

Medidas de dispersión

Las medidas de dispersión nos informan sobre cuanto se alejan del centro los valores de la distribución.

Las medidas de dispersión son:

Rango o recorrido

El rango es la diferencia entre el mayor y el menor de los valores de una distribución estadística.

Desviación media

La desviación media es la media aritmética de los valores absolutos de las desviaciones respecto a la media.

Varianza

La varianza es la media aritmética del cuadrado de las desviaciones respecto a la media.

Desviación típica

La desviación típica es la raíz cuadrada de la varianza

UNIDAD II

Una distribución de frecuencias o tabla de frecuencias es una ordenación en forma de tabla de los datos estadísticos, asignando a cada dato su frecuencia correspondiente.

2.1.-Frecuencia absoluta

La frecuencia absoluta es el número de veces que aparece un determinado valor en un estudio estadístico.

Se representa por f_i.

La suma de las frecuencias absolutas es igual al número total de datos, que se representa por N.

$$f_1 + f_2 + f_3 + \ldots + f_n = N$$

Para indicar resumidamente estas sumas se utiliza la letra griega Σ (sigma mayúscula) que se lee suma o sumatoria.

$$\sum_{i=1}^{i=n} f_i = N$$

Frecuencia relativa

La frecuencia relativa es el cociente entre la frecuencia absoluta de un determinado valor y el número total de datos.

Se puede expresar en tantos por ciento y se representa por n_i.

$$n_i = \frac{f_i}{N}$$

La suma de las frecuencias relativas es igual a 1.

Frecuencia acumulada

La frecuencia acumulada es la suma de las frecuencias absolutas de todos los valores inferiores o iguales al valor considerado.

Se representa por F_i.

Frecuencia relativa acumulada

La frecuencia relativa acumulada es el cociente entre la frecuencia acumulada de un determinado valor y el número total de datos. Se puede expresar en tantos por ciento.

Ejemplo

Durante el mes de julio, en una ciudad se han registrado las siguientes temperaturas máximas:

32, 31, 28, 29, 33, 32, 31, 30, 31, 31, 27, 28, 29, 30, 32, 31, 31, 30, 30, 29, 29, 30, 30, 31, 30, 31, 34, 33, 33, 29, 29.

En la primera columna de la tabla colocamos la variable ordenada de menor a mayor, en la segunda hacemos el recuento y en la tercera anotamos la frecuencia absoluta.

x_i	Recuento	f_i	F_i	n_i	N_i
27	I	1	1	0.032	0.032
28	II	2	3	0.065	0.097
29	𝍸 I	6	9	0.194	0.290
30	𝍸 II	7	16	0.226	0.516
31	𝍸 III	8	24	0.258	0.774
32	III	3	27	0.097	0.871
33	III	3	30	0.097	0.968
34	I	1	31	0.032	1

		31		1	

Este tipo de tablas de frecuencias se utiliza con variables discretas.

Distribución de frecuencias agrupadas

La distribución de frecuencias agrupadas o tabla con datos agrupados se emplea si las variables toman un número grande de valores o la variable es continua.

Se agrupan los valores en intervalos que tengan la misma amplitud denominados clases. A cada clase se le asigna su frecuencia correspondiente.

Límites de la clase

Cada clase está delimitada por el límite inferior de la clase y el límite superior de la clase.

Amplitud de la clase

La amplitud de la clase es la diferencia entre el límite superior e inferior de la clase.

Marca de clase

La marca de clase es el punto medio de cada intervalo y es el valor que representa a todo el intervalo para el cálculo de algunos parámetros.

Construcción de una tabla de datos agrupados

3, 15, 24, 28, 33, 35, 38, 42, 43, 38, 36, 34, 29, 25, 17, 7, 34, 36, 39, 44, 31, 26, 20, 11, 13, 22, 27, 47, 39, 37, 34, 32, 35, 28, 38, 41, 48, 15, 32, 13.

1° se localizan los valores menor y mayor de la distribución. En este caso son 3 y 48.

2° Se restan y se busca un número entero un poco mayor que la diferencia y que sea divisible por el número de intervalos de queramos poner. Es conveniente que el número de intervalos

oscile entre 6 y 15.En este caso, 48 - 3 = 45, incrementamos el número hasta 50 : 5 = 10 intervalos.

Se forman los intervalos teniendo presente que el límite inferior de una clase pertenece al intervalo, pero el límite superior no pertenece intervalo, se cuenta en el siguiente intervalo.

	c_i	f_i	F_i	n_i	N_i
[0, 5)	2.5	1	1	0.025	0.025
[5, 10)	7.5	1	2	0.025	0.050
[10, 15)	12.5	3	5	0.075	0.125
[15, 20)	17.5	3	8	0.075	0.200
[20, 25)	22.5	3	11	0.075	0.275
[25, 30)	27.5	6	17	0.150	0.425
[30, 35)	32.5	7	24	0.175	0.600
[35, 40)	37.5	10	34	0.250	0.850
[40, 45)	42.5	4	38	0.100	0.950
[45, 50)	47.5	2	40	0.050	1
		40		1	

2.2 La representación gráfica de una distribución de frecuencias

REPRESENTACIÓN GRÁFICA DE DATOS ESTADÍSTICOS

En los análisis estadísticos, es frecuente utilizar representaciones visuales complementarias de las tablas que resumen los datos de estudio. Con estas representaciones, adaptadas en cada caso a la finalidad informativa que se persigue, se transmiten los resultados de los análisis de forma rápida, directa y comprensible para un conjunto amplio de personas.

Tipos de representaciones gráficas

Cuando se muestran los datos estadísticos a través de representaciones gráficas, se ha de adaptar el contenido a la información visual que se pretende transmitir. Para ello, se barajan múltiples formas de representación:

- Diagramas de barras: muestran los valores de las frecuencias absolutas sobre un sistema de ejes cartesianos, cuando la variable es discreta o cualitativa.
- Histogramas: formas especiales de diagramas de barras para distribuciones cuantitativas continuas.
- Polígonos de frecuencias: formados por líneas poligonales abiertas sobre un sistema de ejes cartesianos.
- Gráficos de sectores: circulares o de tarta, dividen un círculo en porciones proporcionales según el valor de las frecuencias relativas.
- Pictogramas: o representaciones visuales figurativas. En realidad son diagramas de barras en los que las barras se sustituyen con dibujos alusivos a la variable.
- Cartogramas: expresiones gráficas a modo de mapa.
- Pirámides de población: para clasificaciones de grupos de población por sexo y edad.

Diagramas de barras e histogramas

Los diagramas de barras se usan para representar gráficamente series estadísticas de valores en un sistema de ejes cartesianos, de manera que en las abscisas se indica el valor de la variable estadística y en las ordenadas se señala su frecuencia absoluta.

Estos gráficos se usan en representación de caracteres cualitativos y cuantitativos discretos. En variables cuantitativas continuas, se emplea una variante de los mismos llamada histograma.

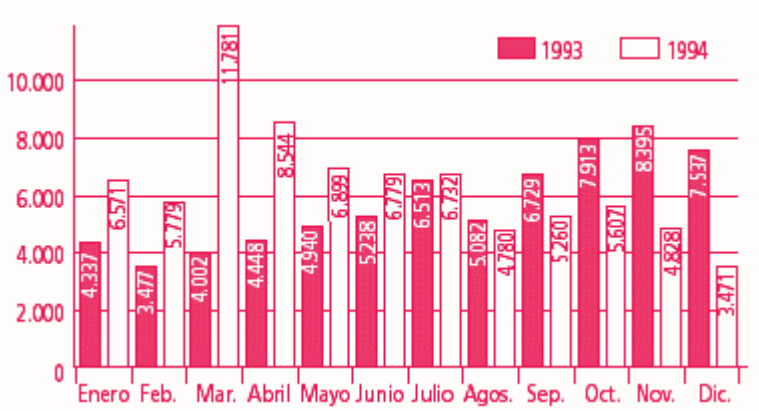
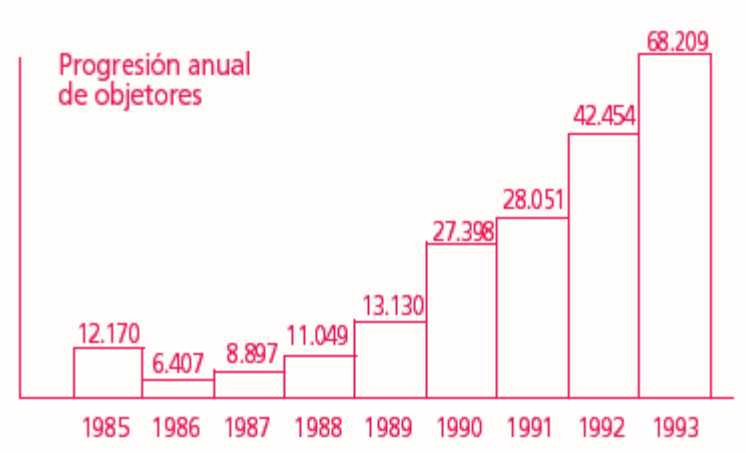


Diagrama de barras.



Histograma.

Polígonos de frecuencias

Para construir polígonos de frecuencias, se trazan las frecuencias absolutas o relativas de los valores de la variable en un sistema de ejes cartesianos y se unen los puntos resultantes mediante trazos rectos. Con ello se obtiene una forma de línea poligonal abierta.

Los polígonos de frecuencias se utilizan preferentemente en la presentación de caracteres cuantitativos, y tienen especial interés cuando se indican frecuencias acumulativas. Se usan en la expresión de fenómenos que varían con el tiempo, como la densidad de población, el precio o la temperatura.

Gráficos de sectores

En los diagramas de sectores, también llamados circulares o de tarta, se muestra el valor de la frecuencia de la variable señalada como un sector circular dentro de un círculo completo. Por ello, resultan útiles particularmente para mostrar comparaciones entre datos, sobre todo en forma de frecuencias relativas de las variables expresadas en forma de porcentaje.

Pictogramas y cartogramas

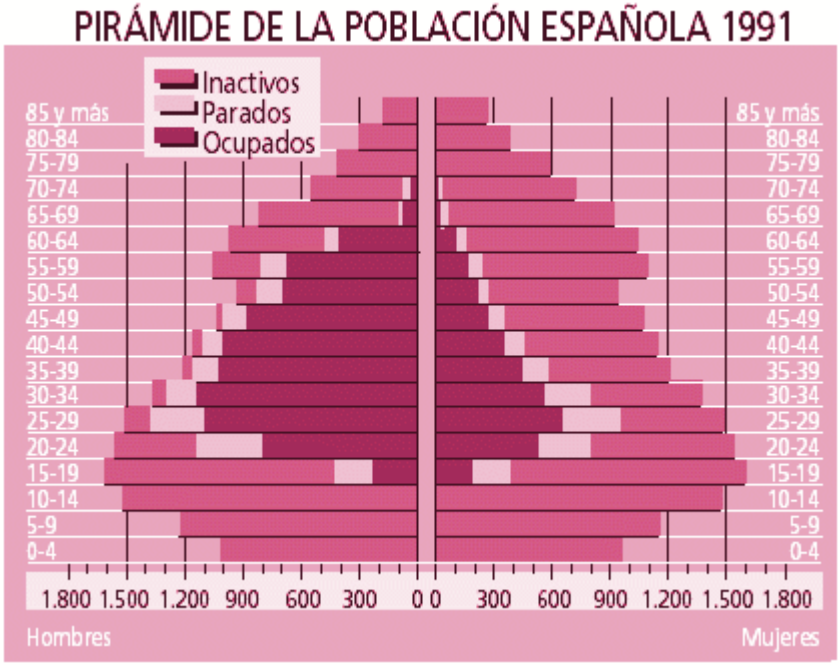
Para aligerar la presentación de datos estadísticos, con frecuencia se recurre a imágenes pictóricas representativas del valor de las variables. Dos formas comunes de expresión gráfica de los datos son:

- Los pictogramas, que muestran diagramas figurativos con figuras o motivos que aluden a la distribución estadística analizada (por ejemplo, una imagen antropomórfica para indicar tamaños, alturas u otros).
- Los cartogramas, basados en mapas geográficos que utilizan distintas tramas, colores o intensidades para remarcar las diferencias entre los datos.

Pirámide de población

Otra forma corriente de presentación visual de datos estadísticos es la llamada pirámide de población.

Las pirámides de población se utilizan en la expresión de informaciones demográficas, económicas o sociales, y en ellas se clasifican comúnmente los datos de la población del grupo de muestra considerado en diferentes escalas de edad y diferenciada por sexo.



Propiedades de la distribución de frecuencias

Definimos la distribución de frecuencias como la reunión de unos datos en categorías excluyentes entre ellas. La forma de una distribución se caracteriza por cuatro propiedades básicas de la distribución de frecuencias que definiremos en este artículo de Psicología-Online: la tendencia central, la variabilidad, el sesgo o la asimetría y la curtosis o apuntamiento.

También te puede interesar: Teoría de respuesta al ítem - Aplicaciones y Test

Propiedades de la distribución de frecuencias

Tendencia central

Es un valor de la variable que se encuentra hacia el centro de la distribución de frecuencias. A este valor se le denomina promedio y es un valor que sintetiza a todos los valores de la distribución.

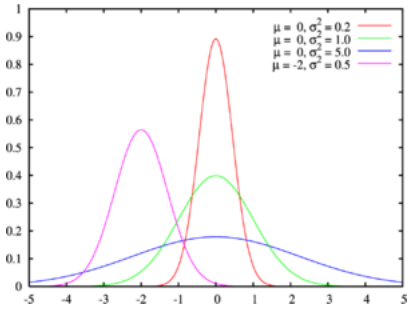
Variabilidad

Es un índice o medida que resume el grado de concentración de los valores de una distribución en torno a un promedio. Si los valores están muy cercanos al promedio se habla de una distribución homogénea; si, por el contrario, los valores se alejan del promedio se habla de distribución heterogénea.

Sesgo o asimetría

Se refiere al grado de simetría o asimetría de una distribución de frecuencias. Si hay un número de observaciones similar por debajo y por encima del promedio se dice que la distribución es simétrica. Si hay una mayor frecuencia de valores bajos que de valores altos se dice que la distribución es asimétrica positiva. Si hay una mayor frecuencia de valores altos que bajos, se dice que la distribución es asimétrica negativa. Las distribuciones asimétricas positivas son propias de tareas o tests difíciles, al contrario que las distribuciones asimétricas negativas, que suelen ser de tareas fáciles.

Enfoque descriptivo[editar]



Gráficas de distribuciones normales para distintos valores de sus dos parámetros.

Un parámetro estadístico es una medida poblacional. Este enfoque es el tradicional de la estadística descriptiva.[567] En este sentido, su acepción se acerca a la de medida o valor que se compara con otros, tomando una unidad de una determinada magnitud como referencia.

Por su parte, la facción más formal de la estadística, la estadística matemática y también la inferencia estadística utilizan el concepto de parámetro en su acepción matemática más pura, esto es, como variable que define una familia de objetos matemáticos en determinados modelos. Así se habla, por ejemplo, de una distribución normal de parámetros μ y σ como de una determinada familia de distribuciones con una distribución de probabilidad de expresión conocida, en la que tales parámetros definen aspectos concretos como la esperanza, la varianza, la curtosis, etc. Otro ejemplo común en este sentido es el de la distribución de Poisson, determinada por un parámetro, λ; o la distribución binomial, determinada por dos parámetros, n y p. Desde el punto de vista de la estadística matemática, el hecho de que estas distribuciones describan situaciones reales y los citados parámetros signifiquen un resumen de determinado conjunto de datos es indiferente.

Propiedades deseables en un parámetro[editar]

Según Yule[8] un parámetro estadístico es deseable que tenga las siguientes propiedades:

- Se define de manera objetiva, es decir, es posible calcularlo sin ambigüedades, generalmente mediante una fórmula matemática. Por ejemplo, la media aritmética se

define como la suma de todos los datos, dividida por el número de datos. No hay ambigüedad: si se realiza ese cálculo, se obtiene la media; si se realiza otro cálculo, se obtiene otra cosa. Sin embargo, la definición de moda como el "valor más frecuente", puede dar lugar a confusión cuando la mayor frecuencia la presentan varios valores distintos.

- No desperdicia, a priori, ninguna de las observaciones. Con carácter general, un parámetro será más representativo de una determinada población, cuántos más valores de la variable estén implicados en su cálculo. Por ejemplo, para medir la dispersión puede calcularse el recorrido, que sólo usa dos valores de la variable objeto de estudio, los extremos; o la desviación típica, en cuyo cálculo intervienen todos los datos del eventual estudio.
- Es interpretable, significa algo. La mediana, por ejemplo, deja por debajo de su valor a la mitad de los datos, está justo en medio de todos ellos cuando están ordenados. Esta es una interpretación clara de su significado.
- Es sencillo de calcular y se presta con facilidad a manipulaciones algebraicas. Se verá más abajo que una medida de la dispersión es la desviación media. Sin embargo, al estar definida mediante un valor absoluto, función definida a trozos y no derivable, no es útil para gran parte de los cálculos en los que estuviera implicada, aunque su interpretación sea muy clara.
- Es poco sensible a las fluctuaciones muestrales. Si pequeñas variaciones en una muestra de datos estadísticos influyen en gran medida en un determinado parámetro, es porque tal parámetro no representa con fiabilidad a la población. Así pues es deseable que el valor de un parámetro con esta propiedad se mantenga estable ante las pequeñas oscilaciones que con frecuencia pueden presentar las distintas muestras estadísticas. Esta propiedad es más interesante en el caso de la estimación de parámetros. Por otra parte, los parámetros que no varían con los cambios de origen y escala o cuya variación está controlada algebraicamente, son apropiados en determinadas circunstancias como la tipificación.

Principales parámetros

Artículo principal: Estadístico maestral

Habitualmente se agrupan los parámetros en las siguientes categorías:

Medidas de posición.[9]

Se trata de valores de la variable estadística que se caracterizan por la posición que ocupan dentro del rango de valores posibles de esta. Entre ellos se distinguen:

- Las medidas de tendencia central: medias, moda y mediana.
- Las medidas de posición no central: cuantiles (cuartiles, deciles y percentiles).

Medidas de dispersión.[10]

Resumen la heterogeneidad de los datos, lo separados que estos están entre sí. Hay dos tipos, básicamente:

- Medidas de dispersión absolutas, que vienen dadas en las mismas unidades en las que se mide la variable: recorridos, desviaciones medias, varianza, y desviación típica.
- Medidas de dispersión relativa, que informan de la dispersión en términos relativos, como un porcentaje. Se incluyen entre estas el coeficiente de variación, el coeficiente de apertura, los recorridos relativos y el índice de desviación respecto de la mediana.

Medidas de forma.[11]

Su valor informa sobre el aspecto que tiene la gráfica de la distribución. Entre ellas están los coeficientes de asimetría y los de curtosis.

Otros parámetros.

Además, y con propósitos más específicos, existen otros parámetros de uso en situaciones muy concretas, como son las proporciones, los números índice, las tasas y el coeficiente de Gini.

Medidas de tendencia central o centralización[editar]

Artículo principal: *Medidas de tendencia central*

Son valores que suelen situarse cerca del centro de la distribución de datos. Los más destacados son las medias o promedios (incluyendo la media aritmética, la media geométrica y la media armónica), la mediana y la moda.

Media aritmética o promedio

La estatura media como resumen de una población homogénea (abajo) o heterogénea (arriba).

Artículo principal: *Media aritmética*

La media muestral o media aritmética es, probablemente, uno de los parámetros estadísticos más extendidos.[12] Sus propiedades son:[13]

- Su cálculo es muy sencillo y en él intervienen todos los datos.

- Se interpreta como "punto de equilibrio" o "centro de masas" del conjunto de datos, ya que tiene la propiedad de equilibrar las desviaciones de los datos respecto de su propio valor:

- Minimiza las desviaciones cuadráticas de los datos respecto de cualquier valor

  prefijado, esto es, el valor de es mínimo cuando . Este resultado se conoce como Teorema de König. Esta propiedad permite interpretar uno de los parámetros de dispersión más importantes: la varianza.

- Se ve afectada por transformaciones afines (cambios de origen y escala), esto es, si

  Entonces , donde es la media aritmética de los , para $i = 1, ..., n$ y a y b números reales.

Este parámetro, aun teniendo múltiples propiedades que aconsejan su uso en situaciones muy diversas, tiene también algunos inconvenientes, como son:

- Para datos agrupados en intervalos (variables continuas), su valor oscila en función de la cantidad y amplitud de los intervalos que se consideren.

- Es una medida a cuyo significado afecta sobremanera la dispersión, de modo que cuanto menos homogéneos son los datos, menos información proporciona. Dicho de otro modo, poblaciones muy distintas en su composición pueden tener la misma media.[14] Por ejemplo, un equipo de baloncesto con cinco jugadores de igual estatura, 1,95, pongamos por caso, tendría una estatura media de 1,95, evidentemente, valor que representa fielmente a esta homogénea población. Sin embargo, un equipo de estaturas más heterogéneas, 2,20, 2,15, 1,95, 1,75 y 1,70, por ejemplo, tendría también, como puede comprobarse, una estatura media de 1,95, valor que no representa a casi ninguno de sus componentes.

- Es muy sensible a los valores extremos de la variable. Por ejemplo, en el cálculo del salario medio de una empresa, el salario de un alto directivo que gane 1.000.000 de €

tiene tanto peso como el de mil empleados "normales" que ganen 1.000 €, siendo la media de aproximadamente 2.000 €.

Moda

Artículo principal: Moda (estadística)

La moda es el dato más repetido, el valor de la variable con mayor frecuencia absoluta.[15] En cierto sentido se corresponde su definición matemática con la locución "estar de moda", esto es, ser lo que más se lleva.

Su cálculo es extremadamente sencillo, pues sólo necesita de 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.

Sus principales propiedades son:

- Cálculo sencillo.
- Interpretación muy clara.
- Al depender sólo de las frecuencias, puede calcularse para variables cualitativas. Es por ello el parámetro más utilizado cuando al resumir una población no es posible realizar otros cálculos, por ejemplo, cuando se enumeran en medios periodísticos las características más frecuentes de determinado sector social. Esto se conoce informalmente como "retrato robot".[16]

Inconvenientes:

- Su valor es independiente de la mayor parte de los datos, lo que la hace muy sensible a variaciones muéstrales. Por otra parte, en variables agrupadas en intervalos, su valor depende excesivamente del número de intervalos y de su amplitud.
- Usa muy pocas observaciones, de tal modo que grandes variaciones en los datos fuera de la moda, no afectan en modo alguno a su valor.
- No siempre se sitúa hacia el centro de la distribución.

- Puede haber más de una moda en el caso en que dos o más valores de la variable presenten la misma frecuencia (distribuciones bimodales o multimodales).

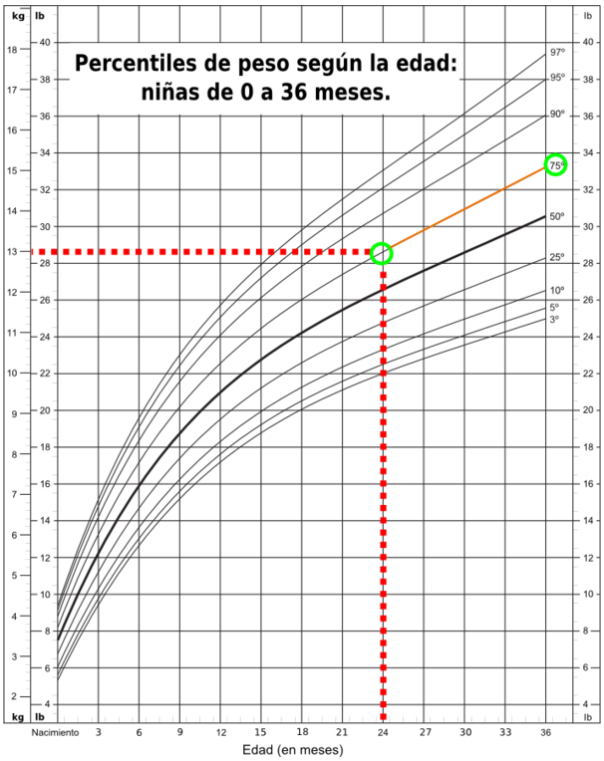
Mediana

Artículo principal: Mediana (estadística)

La mediana es un valor de la variable que deja por debajo de sí a la mitad de los datos, una vez que estos están ordenados de menor a mayor.[17] Por ejemplo, la mediana del número de hijos de un conjunto de trece familias, cuyos respectivos hijos son: 3, 4, 2, 3, 2, 1, 1, 2, 1, 1, 2, 1 y 1, es 2, puesto que, una vez ordenados los datos: 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4, el que ocupa la posición central es 2:

En caso de un número par de datos, la mediana no correspondería a ningún valor de la variable, por lo que se conviene en tomar como mediana el valor intermedio entre los dos valores centrales. Por ejemplo, en el caso de doce datos como los anteriores:

Se toma como mediana



En este ejemplo basado en una tabla real de percentiles usada en pediatría, puede comprobarse que una niña de 24 meses con un peso de 13 kg estaría en el percentil 75°, esto es, su peso es superior al 75% de las niñas de su edad. La mediana correspondería, aproximadamente, a 12 kg (intersección de la línea curva más oscura con la línea horizontal correspondiente al valor 12 en el eje vertical, para esa misma edad).

Existen métodos de cálculo más rápidos para datos más numerosos (véase el artículo principal dedicado a este parámetro). Del mismo modo, para valores agrupados en intervalos, se halla el "intervalo mediano" y, dentro de este, se obtiene un valor concreto por interpolación.

Propiedades de la mediana como parámetro estadístico:[18]

- Es menos sensible que la media a oscilaciones de los valores de la variable. Un error de transcripción en la serie del ejemplo anterior en, pongamos por caso, el último número, deja a la mediana inalterada.
- Como se ha comentado, puede calcularse para datos agrupados en intervalos, incluso cuando alguno de ellos no está acotado.
- No se ve afectada por la dispersión. De hecho, es más representativa que la media aritmética cuando la población es bastante heterogénea. Suele darse esta circunstancia cuando se resume la información sobre los salarios de un país o una empresa. Hay unos pocos salarios muy altos que elevan la media aritmética haciendo que pierda representatividad respecto al grueso de la población. Sin embargo, alguien con el salario "mediano" sabría que hay tanta gente que gana más dinero que él, como que gana menos.

Sus principales inconvenientes son que en el caso de datos agrupados en intervalos, su valor varía en función de la amplitud de estos. Por otra parte, no se presta a cálculos algebraicos tan bien como la media aritmética.

Medidas de posición no central [editar]

Artículo principal: Medidas de posición no central

Directamente relacionados con la anterior, se encuentran las medidas de posición no central, también conocidas como cuantiles. Se trata de valores de la variable estadística que dejan por debajo de sí determinada cantidad de los datos. Son, en definitiva, una generalización del concepto de la mediana. Mientras que ésta deja por debajo de sí al 50% de la distribución, los cuantiles pueden hacerlo con cualquier otro porcentaje.[19] Se denominan medidas de posición porque informan, precisamente, de la posición que ocupa un valor dentro de la distribución de datos.

Tradicionalmente se distingue entre cuartiles, si se divide la cantidad de datos en cuatro partes antes de proceder al cálculo de los valores que ocupan cada posición; deciles, si se divide los datos en diez partes; o percentiles, que dividen la población en cien partes.

Ejemplos: si se dice que una persona, tras un test de inteligencia, ocupa el percentil 75, ello supone que el 75% de la población tiene un cociente intelectual con un valor inferior al de esa persona. Este criterio se usa por las asociaciones de superdotados, que limitan su conjunto de miembros a aquellas que alcanzan determinado percentil (igual o superior a 98 en la mayoría de los casos).

El ejemplo que se muestra en la imagen de la derecha es el correspondiente al cálculo inverso, esto es, cuando se desea conocer el percentil correspondiente a un valor de la variable, en lugar del valor que corresponde a un determinado percentil.

Otras medidas de posición central son la media geométrica y la media armónica que, aunque tienen determinadas propiedades algebraicas que podrían hacerlas útiles en determinadas circunstancias, su interpretación no es tan intuitiva como la de los parámetros anteriores.[20]

Comentarios sobre las medidas de posición

Este tipo de parámetros no tienen por qué coincidir con un valor exacto de la variable y, por tanto, tampoco pueden usarse con carácter general para hacer pronósticos. Por ejemplo, si se dice que la media aritmética de los hijos de las familias de un país es de 1,2, no es posible encontrar familias con ese valor en concreto. Un segundo ejemplo: a ninguna fábrica de zapatos se le ocurriría fabricar los suyos con tallas únicamente correspondientes al valor promedio, ni siquiera tienen por qué ser estas tallas las más fabricadas, pues en tal caso sería más apropiado atender a la moda de la distribución de tallas de los eventuales clientes.

La elección de uno u otro parámetro dependerá de cada caso particular, de los valores de la variable y de los propósitos del estudio. Su uso indiscriminado puede ser deliberadamente tendencioso o involuntariamente sesgado, convirtiéndose, de hecho, en un abuso.[21] Puede pensarse, por ejemplo, en la siguiente situación: un empresario publica que el salario medio en su empresa es de 1.600 €. A este dato, que en determinadas circunstancias podría considerarse muy bueno, podría llegarse si la empresa tuviese cuatro empleados con salarios de 1.000 € mensuales y el salario del jefe, incluido en la media, fuese de 4.000 € al mes:[22]

Con carácter general y a modo de resumen podría decirse que la media aritmética es un parámetro representativo cuando la población sigue una distribución normal o es bastante homogénea; en otras situaciones de fuerte dispersión, habría que decantarse por la mediana. La moda es el último recurso (y el único) cuando de describir variables cualitativas se trata.

Medidas de dispersión

Artículo principal: *Dispersión (matemática)*

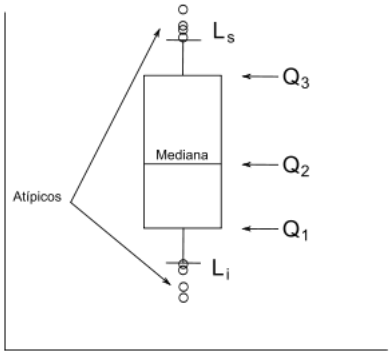


Diagrama de caja que muestra la dispersión gráficamente, usando los cuartiles como referencia. Entre Q_1 y Q_3 (rango intercuartílico) se encuentran el 50% de las observaciones.

Las medidas de posición resumen la distribución de datos, pero resultan insuficientes y simplifican excesivamente la información. Estas medidas adquieren verdadero significado cuando van acompañadas de otras que informen sobre la heterogeneidad de los datos. Los parámetros de dispersión miden eso precisamente, generalmente, calculando en qué medida los datos se agrupan en torno a un valor central. Indican, de un modo bien definido, lo homogéneos que estos datos son. Hay medidas de dispersión absolutas, entre las cuales se encuentran la varianza, la desviación típica o la desviación media, aunque también existen otras menos utilizadas como los recorridos o la meda; y medidas de dispersión relativas, como el coeficiente de variación, el coeficiente de apertura o los recorridos relativos. En muchas ocasiones las medidas de dispersión se ofrecen acompañando a un parámetro de posición central para indicar en qué medida los datos se agrupan en torno de él.[23]

Medidas de dispersión absolutas

Recorridos

El recorrido o rango de una variable estadística es la diferencia entre el mayor y el menor valor que toma la misma. Es la medida de dispersión más sencilla de calcular, aunque es algo burda porque sólo toma en consideración un par de observaciones. Basta con que uno de estos dos datos varíe para que el parámetro también lo haga, aunque el resto de la distribución siga siendo, esencialmente, la misma.

Existen otros parámetros dentro de esta categoría, como los recorridos o rangos intercuantílicos, que tienen en cuenta más datos y, por tanto, permiten afinar en la dispersión. Entre los más usados está el rango intercuartílico, que se define como la diferencia entre el cuartil tercero y el cuartil primero. En ese rango están, por la propia definición de los cuartiles, el 50% de las observaciones. Este tipo de medidas también se usa para determinar valores atípicos.

En el diagrama de caja que aparece a la derecha se marcan como valores atípicos todos aquellos que caen fuera del intervalo $[L_i, L_s] = [Q_1 - 1{,}5 \cdot R_s, Q_3 + 1{,}5 \cdot R_s]$, donde Q_1 y Q_3 son los cuartiles 1° y 3°, respectivamente, y R_s representa la mitad del recorrido o rango intercuartílico, también conocido como recorrido semiintercuartílico.[24]

Desviaciones medias

Artículo principal: Desviación media

Dada una variable estadística X y un parámetro de tendencia central, c, se llama desviación de un valor de la variable, x_i, respecto de c, al número $|x_i - c|$. Este número mide lo lejos que está cada dato del valor central c, por lo que una media de esas medidas podría resumir el conjunto de desviaciones de todos los datos.

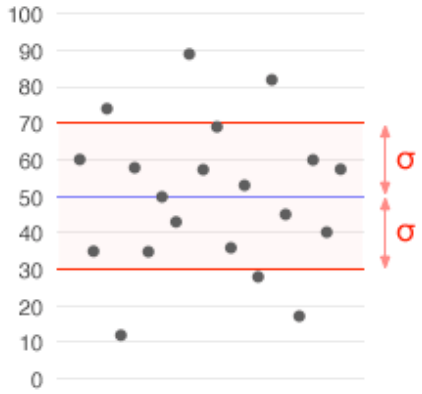
Así pues, se denomina desviación media de la variable X respecto de c a la media aritmética de las desviaciones de los valores de la variable respecto de c, esto es, si

Entonces

De este modo se definen la desviación media respecto de la media o la desviación media

respecto de la mediana ($c =$), cuya interpretación es sencilla en virtud del significado de la media aritmética.[23] Sin embargo, el uso de valores absolutos impide determinados cálculos algebraicos que obligan a desechar estos parámetros, a pesar de su clara interpretación, en favor de los siguientes.

Varianza y desviación típica



Conjunto de datos estadísticos de media aritmética 50 (línea azul) y desviación típica 20 (líneas rojas).

Como se vio más arriba, la suma de todas las desviaciones respecto al parámetro más utilizado, la media aritmética, es cero. Por tanto si se desea una medida de la dispersión sin los inconvenientes para el cálculo que tienen las desviaciones medias, una solución es elevar al cuadrado tales desviaciones antes de calcular el promedio. Así, se define la varianza como:[25] o sea, la media de los cuadrados de las desviaciones respecto de la media. La desviación típica, σ, se define como la raíz cuadrada de la varianza, esto es,

Para variables agrupadas en intervalos, se usan las marcas de clase (un valor apropiado del interior de cada intervalo) en estos cálculos.

Propiedades:[25]

- Ambos parámetros no se alteran con los cambios de origen.
- Si todos los valores de la variable se multiplican por una constante, *b*, la varianza queda multiplicada por b^2.
- En el intervalo se encuentran, al menos, el de las observaciones (*véase* Desigualdad de Tchebyschev).

Esta última propiedad muestra la potencia del uso conjunto de la media y la desviación típica como parámetros estadísticos, ya que para valores de *k* iguales a 2 y 3, respectivamente, se obtiene que:

- En el intervalo están, al menos, el 75% de los datos.
- En el intervalo están, al menos, el 89% de los datos.

Se cumple la siguiente relación entre los parámetros de dispersión: donde , y son, respectivamente, la desviación media respecto de la mediana, la desviación media respecto de la media y la desviación típica (*véase* Desviación media).

La media. Es una medida de dispersión que tiene, por su propia definición, las mismas propiedades que la mediana. Por ejemplo, no se ve afectada por valores extremos o atípicos.[27]

Medidas de dispersión relativa

Son parámetros que miden la dispersión en términos relativos, un porcentaje o una proporción, por ejemplo, de modo que permiten una sencilla comparación entre la dispersión de distintas distribuciones.[28]

Coeficiente de variación de Pearson

Artículo principal:
Coeficiente de variación

Se define como , donde σ es la desviación típica y es la media aritmética. Se interpreta como el número de veces que la media está contenida en la desviación típica. Suele darse su valor en tanto por ciento, multiplicando el resultado anterior por 100. De este modo se obtiene un porcentaje de la variabilidad.

Su principal inconveniente es que en el caso de distribuciones cuya media se acerca a cero, su valor tiende a infinito e incluso resulta imposible de calcular cuando la media es cero. Por ello no puede usarse para variables tipificadas.

Coeficiente de apertura

Se define como el cociente entre los valores extremos de la distribución de datos, esto es, dada una distribución de datos estadísticos x_1, x_2, ..., x_n, su coeficiente de apertura, C_A .Se usa para comparar salarios de empresas.

Recorridos relativos.

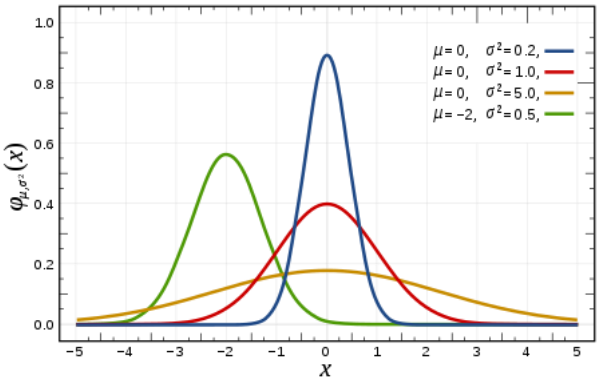
Dado R_e, el recorrido de una distribución de datos estadísticos, el recorrido relativo, R_R es, donde es la media aritmética de la distribución. Dada una distribución de datos estadísticos con cuartiles Q_1, Q_2 y Q_3, el recorrido intercuartílico relativo, R_{IQR} se define como[29] por otra

parte, se define el recorrido semi intercuartílico relativo, R_{SIR}, como

Índice de desviación respecto a la mediana

Se define como , donde D_{Me} es la desviación media respecto de la mediana y *Me* es la mediana de una distribución de datos estadísticos dada.

Medidas de forma



La campana de Gauss, curva que sirve de modelo para el estudio de la forma de una distribución.

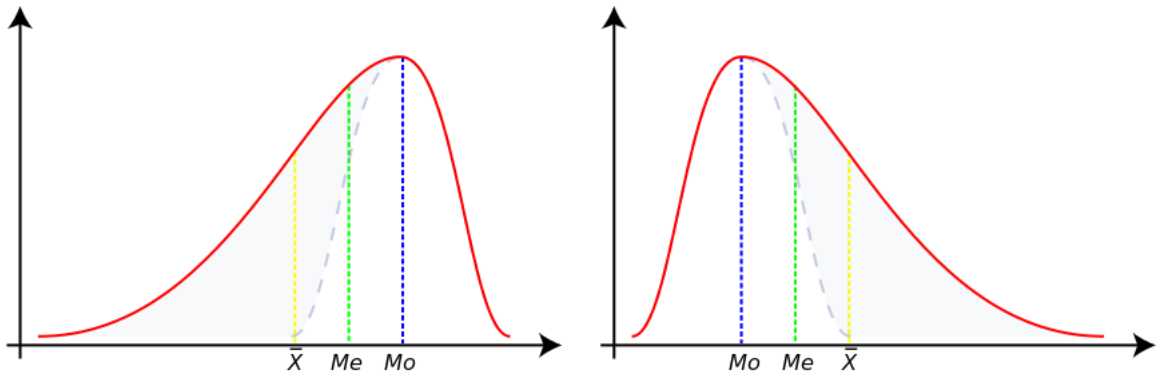
Las medidas de forma caracterizan la forma de la gráfica de una distribución de datos estadísticos. La mayoría de estos parámetros tiene un valor que suele compararse con la campana de Gauss, esto es, la gráfica de la distribución normal, una de las que con más frecuencia se ajusta a fenómenos reales.

Medidas de asimetría

Artículo principal: Asimetría estadística

Se dice que una distribución de datos estadísticos es simétrica cuando la línea vertical que pasa por su media, divide a su representación gráfica en dos partes simétricas. Ello equivale a decir que los valores equidistantes de la media, a uno u otro lado, presentan la misma frecuencia.

En las distribuciones simétricas los parámetros media, mediana y moda coinciden, mientras que si una distribución presenta cierta asimetría, de un tipo o de otro, los parámetros se sitúan como muestra el siguiente gráfico:

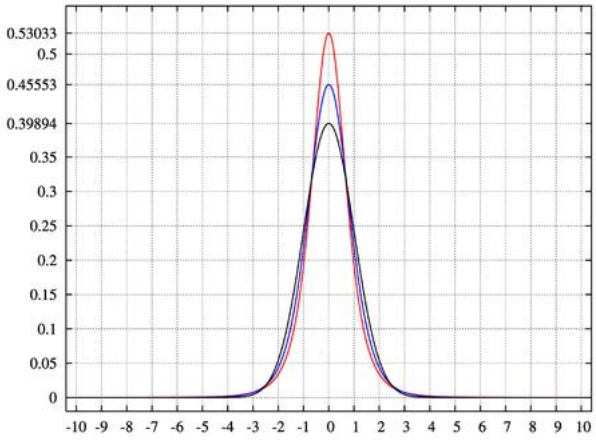


Ello puede demostrarse fácilmente si se tiene en cuenta la atracción que la media aritmética siente por los valores extremos, que ya se ha comentado más arriba y las definiciones de mediana (justo en el centro de la distribución, tomando el eje de abscisas como referencia) y moda (valor que presenta una ordenada más alta).

Por consiguiente, la posición relativa de los parámetros de centralización pueden servir como una primera medida de la simetría de una distribución.

Otras medidas más precisas son el coeficiente de asimetría de Fisher, el coeficiente de asimetría de Bowley y el coeficiente de asimetría de Pearson.

Medidas de apuntamiento o curtosis

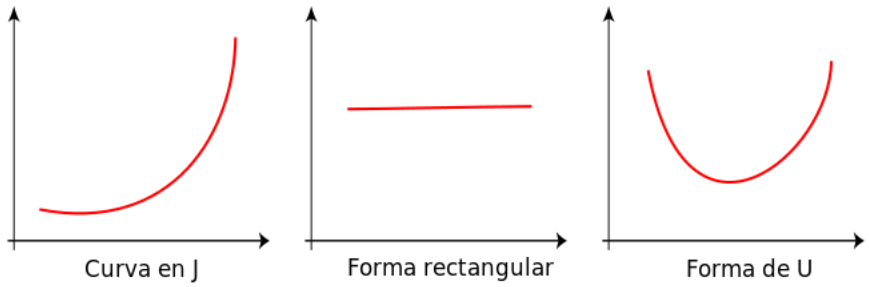


Tres distribuciones con distintos grados de apuntamiento.

Con estos parámetros se pretende medir cómo se reparten las frecuencias relativas de los datos entre el centro y los extremos, tomando como comparación la campana de Gauss.

El parámetro usado con más frecuencia para esta medida es el coeficiente de curtosis de Fisher, aunque hay otros como el coeficiente de curtosis de Kelley o el coeficiente de curtosis percentílico. La comparación con la distribución normal permite hablar de distribuciones platicúrticas o más aplastadas que la normal; distribuciones mesocúrticas, con igual apuntamiento que la normal; y distribuciones leptocúrticas, esto es, más apuntadas que la normal.[30]

Por último, existen otras medidas para decidir sobre la forma de una distribución con ajuste a modelos menos usuales como los que se muestran en las siguientes gráficas:



Otros parámetros

Se presentan en este apartado otros parámetros que tienen aplicación en situaciones muy concretas, por lo que no se incluyen entre los grupos anteriores, aunque tienen cabida en este artículo por su frecuente uso en medios de comunicación y su facultad de resumir grandes cantidades de datos, como ocurre con las medidas tratadas hasta ahora.

Proporción

Artículo principal: *Proporción*

La proporción de un dato estadístico es el número de veces que se presenta ese dato respecto al total de datos. Se conoce también como frecuencia relativa y es uno de los parámetros de cálculo más sencillo. Tiene la ventaja de que puede calcularse para variables cualitativas.

Por ejemplo, si se estudia el color de ojos de un grupo de 20 personas, donde 7 de ellas los tienen azules, la proporción de individuos con ojos azules es del 35% (= 7/20).

El dato con mayor proporción se conoce como moda.

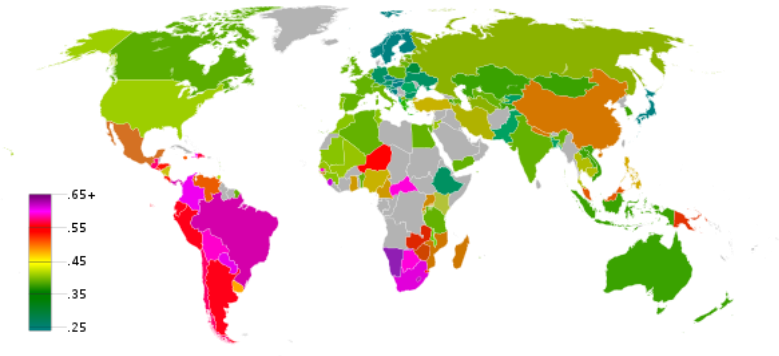
En inferencia estadística existen intervalos de confianza para la estimación de este parámetro.

Número índice

Un número índice es una medida estadística que permite estudiar las fluctuaciones o variaciones de una magnitud o de más de una en relación al tiempo o al espacio. Los índices más habituales son los que realizan las comparaciones en el tiempo. Algunos ejemplos de uso cotidiano de este parámetro son el índice de precios o el IPC[31]

Tasa

Artículo principal: Tasa (índice)



Coeficiente de Gini en el mundo

La tasa es un coeficiente que expresa la relación entre la cantidad y la frecuencia de un fenómeno o un grupo de fenómenos. Se utiliza para indicar la presencia de una situación que no puede ser medida en forma directa.[31] Esta razón se utiliza en ámbitos variados, como la demografía o la economía, donde se hace referencia a la tasa de interés.

Algunos de los más usados son: tasa de natalidad, tasa de mortalidad, tasa de crecimiento demográfico, tasa de fertilidad o tasa de desempleo.

Coeficiente de Gini

Artículo principal: *Coeficiente de Gini*

El índice de Gini o coeficiente de Gini es un parámetro de dispersión usado para medir desigualdades entre los datos de una variable o la mayor o menor concentración de los mismos.

Este coeficiente mide de qué forma está distribuida la suma total de los valores de la variable. Se suele usar para describir salarios. Los casos extremos de *concentración* serían aquel en los que una sola persona acapara el total del dinero disponible para salarios y aquel en el que este total está igualmente repartido entre todos los asalariados.[32]

Momentos

Artículos principales: Momento estándar y Momento centrado.

Los momentos son una forma de generalizar toda la teoría relativa a los parámetros estadísticos y guardan relación con una buena parte de ellos. Dada una distribución de datos estadísticos x_1, x_2, ..., x_n, se define el momento central o momento centrado de orden k como

Para variables continuas la definición cambia sumas discretas por integrales (suma continua), aunque la definición es, esencialmente, la misma.[33] De esta definición y las propiedades de los parámetros implicados que se han visto más arriba, se deduce inmediatamente que: y que. Se llama momento no centrado de orden k a la siguiente expresión:

De la definición se deduce que:

Usando el binomio de Newton, puede obtenerse la siguiente relación entre los momentos centrados y no centrados:
Los momentos de una distribución estadística la caracterizan unívocamente.[35]

Parámetros bidimensionales

Artículo principal: Estadística bidimensional

En estadística se estudian en ocasiones varias características de una población para compararlas, estudiar su dependencia o correlación o realizar cualquier otro estudio conjunto. El caso más común de dos variables se conoce como estadística bidimensional.

Un ejemplo típico es el de un estudio que recoja la estatura (denotémosla por *X*) y el peso (sea *Y*) de los *n* individuos de una determinada población. En tal caso, fruto de la recogida de datos, se obtendría una serie de parejas de datos (x_i, y_i), con $i = 1, ..., n$, cada una de las cuales estaría compuesta por la estatura y el peso del individuo *i*, respectivamente.

En los estudios bidimensionales, cada una de las dos variables que entran en juego, estudiadas individualmente, pueden resumirse mediante los parámetros que se han visto hasta ahora.

Así, tendría sentido hablar de la media de las estaturas ( ) o la desviación típica de los pesos (σ_Y). Incluso para un determinado valor de la primera variable, x_k, cabe hacer estudios condicionados. Por ejemplo, la mediana condicionada a la estatura x_k sería la mediana de los pesos de todos los individuos que tienen esa estatura. Se denota $Me/x=x_k$.

Sin embargo existen otros parámetros que resumen características de ambas distribuciones en su conjunto. Los más destacados son el centro de gravedad, la covarianza y el coeficiente de correlación lineal.

Centro de gravedad

Dadas dos variables estadísticas *X* e *Y*, se define el centro de gravedad como la pareja donde

y son, respectivamente, las medias aritméticas de las variables *X* e *Y*. El nombre de este parámetro proviene de que en una representación de las parejas del estudio en una nube de puntos, en la que cada punto tuviese un peso proporcional a su

Covarianza

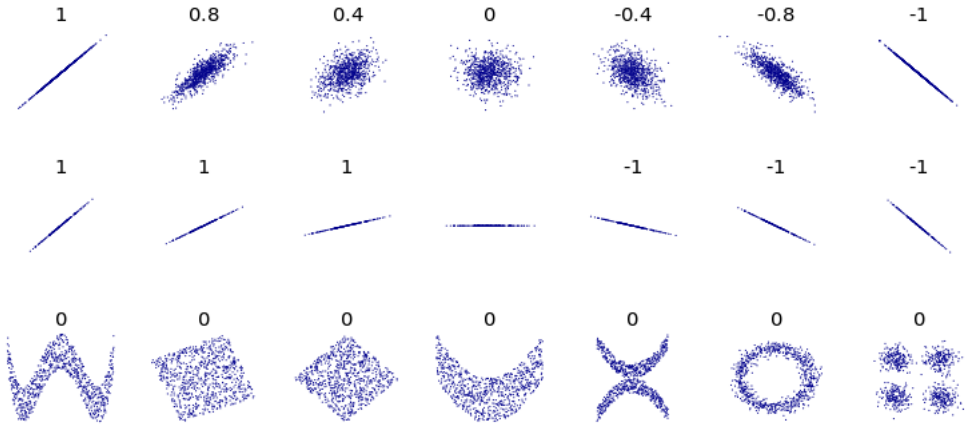
La covarianza o varianza conjunta de una distribución bidimensional se define como:

La interpretación de este parámetro tiene que ver con la eventual correlación lineal de las dos variables. Una covarianza positiva implica una correlación directa y una negativa, una correlación inversa.[38] Por otra parte, es un parámetro imprescindible para el cálculo del coeficiente de correlación lineal o los coeficientes de regresión, como se verá más abajo.

En su contra tiene que se ve excesivamente influenciada, al igual que ocurría con la media aritmética, por los valores extremos de las distribuciones y los cambios de escala.

Coeficiente de correlación lineal

Artículo principal: *Coeficiente de correlación*



Variación del coeficiente de correlación lineal en función de la nube de puntos asociada.

Se trata de un coeficiente que permite determinar la bondad del ajuste de la nube de puntos por una recta.

Se define como: donde σ_{xy} es la covarianza y σ_x y σ_y, las desviaciones típicas respectivas de las distribuciones implicadas. El coeficiente de correlación lineal toma valores entre -1 y 1. En esa escala, mide la correlación del siguiente modo:

- La correlación lineal es más fuerte cuanto más cerca esté de -1 o 1.
- La correlación lineal es más débil cuanto más próximo a cero sea *r*.[39]

El diagrama de la derecha ilustra cómo puede variar *r* en función de la nube de puntos asociada:

Otros parámetros bidimensionales son, el coeficiente de correlación de Spearman, los coeficientes de correlación no paramétricos, el coeficiente de determinación o los coeficientes de regresión lineal.

Al igual que con distribuciones unidimensionales, existe una forma equivalente de desarrollar la teoría relativa a los parámetros estadísticos bidimensionales usando los momentos.

Los parámetros en la inferencia estadísticas

Artículos principales: Estimación estadística y Estadístico maestral.

En ocasiones los parámetros de una determinada población no pueden conocerse con certeza. Generalmente esto ocurre porque es imposible el estudio de la población completa por cuestiones como que el proceso sea destructivo (p. e., vida media de una bombilla) o muy caro (p.e., audiencias de televisión). En tales situaciones se recurre a las técnicas de la inferencia estadística para realizar estimaciones de tales parámetros a partir de los valores obtenidos de una muestra de la población.[40]

Se distingue entonces entre parámetros y estadísticos. Mientras que un parámetro es una función de los datos de la población, el estadístico lo es de los datos de una muestra. De este modo pueden definirse la media maestral, la varianza maestral o cualquier otro parámetro de los vistos más arriba.

Por ejemplo, dada una muestra estadística de tamaño n, de una variable aleatoria X con distribución de probabilidad $F(x,\theta)$, donde θ es un conjunto de parámetros de la distribución, se definiría la media maestral n-enésimo como:

En el caso concreto de la varianza maestral, suele tomarse, por sus mejores propiedades como estimador, la siguiente: donde se ha tomado como denominador n-1, en lugar de n. A este parámetro también se le llama cuasivarianza.[41]

Controversias y malas interpretaciones

Como se ha dicho, los parámetros estadísticos, en el enfoque descriptivo que aquí se adopta, substituyen grandes cantidades de datos por unos pocos valores extraídos de aquellos a través de operaciones simples. Durante este proceso se pierde parte de la información ofrecida originalmente por todos los datos. Es por esta pérdida de datos por lo que la

estadística ha sido tildada en ocasiones de una falacia. Por ejemplo, si en un grupo de tres personas una de ellas ingiere tres helados, el parámetro que con más frecuencia se utiliza para resumir datos estadísticos, la media aritmética del número de helados ingeridos por el

grupo sería igual a 1 ( ), valor que no parece resumir fielmente la información. Ninguna de las personas se sentiría identificada con la frase resumen: "He ingerido un helado de media".

Un ejemplo menos conocido pero igual de ilustrativo acerca de la claridad de un parámetro es la distribución exponencial, que suele regir los tiempos medios entre determinados tipos de sucesos. Por ejemplo, si la vida media de una bombilla es de 8.000 horas, más del 50 por ciento de las veces no llegará a esa media. Igualmente, si un autobús pasa cada 10 minutos de media, hay una probabilidad mayor del 50 por ciento de que pase menos de 10 minutos entre un autobús y el siguiente.

Otro ejemplo que suele ofrecerse con frecuencia para argumentar en contra de la estadística y sus parámetros es que, estadísticamente hablando, la temperatura media de una persona con los pies en un horno y la cabeza en una nevera es ideal.

Variable categórica

En estadística, una variable categórica es una variable que puede tomar uno de un número limitado, y por lo general fijo, de posibles valores, asignando a cada unidad individual u otro tipo observación a un grupo en particular o categoría nominal sobre la base de alguna característica cualitativa.1 En informática y algunas ramas de las matemáticas, las variables categóricas se conocen como enumeraciones o tipos enumerados. Comúnmente (aunque no en este artículo), cada uno de los posibles valores de una variable categórica se conoce como un nivel. La distribución de probabilidad asociada con una variable categórica se llama una distribución categórica.

Una variable categórica que puede tomar dos valores se denomina una variable binaria o una variable dicotómica; un caso especial importante es la variable de Bernoulli.

Las variables categóricas con más de dos valores posibles se denominan variables politómicas; las variables categóricas a menudo se supone que son politómicas menos que se especifique lo contrario. La discretización es el tratamiento de los datos continuos como si fuera categórica. La dicotomización es el tratamiento de los datos continuos o variables politómicos como si fueran variables binarias. El análisis de regresión trata a menudo pertenencia a una categoría como cuantitativa variable ficticia. En una versión simplificada, las variables categóricas son datos que se pueden ver en un gráfico.

Variable cualitativa ordinal

☐El pensante ☐Matemáticas

Quizás lo mejor, antes de abordar la definición de Variable cualitativa ordinal, así como los distintos ejemplos que pueden exponerse respecto a ella, sea revisar de forma breve algunas definiciones, necesarias para entender este tipo de variable estadística dentro de su contexto teórico preciso.

Definiciones fundamentales

En este sentido, puede que sea pertinente entonces comenzar por pasar revista sobre la propia definición de Variable estadística, así como también será necesario abordar la definición de Variables cualitativas, a fin de tener presente la naturaleza de la variable a la que pertenece como subtipo la Variable cualitativa ordinal. A continuación, cada una de las definiciones:

Variable estadística

Con respecto a la noción de Variable estadística, esta es señalada en forma general por las diferentes fuentes teóricas como una característica, que tiende a la fluctuación y la variación, teniendo además la capacidad de adquirir en cada momento valores diferentes, los cuales dan lugar también a que se realicen mediciones u observaciones en base a ellos. Así mismo, la Ciencia estadística hace énfasis en que la Variable es capaz de asumir un valor específico, solo

cuando entra en relación con otras variables, esto es cuando junto con otras características de su misma naturaleza comienza a formar parte de una hipótesis. Por otro lado, la Estadística también distingue entre varios tipos de variables: Cualitativas (Nominal y Ordinal) y Cuantitativas (Continua y Discreta).

Variables Cualitativas

Por consiguiente, las Variables Cualitativas serán consideradas entonces como uno de los dos tipos de Variables estadísticas, las cuales son usadas dentro de la ciencia estadística como variables cuyo principal propósito es dar cuenta de las distintas cualidades o modalidades del objeto, población o sujeto al cual se hace referencia. En este sentido, la Variable Cualitativa constituye una herramienta esencial a la hora de clasificar los elementos de una población según sus atributos. Dentro de las Variable cualitativas se distinguen básicamente dos tipos o clases de ellas: las Variables cualitativas nominales y las Variables cualitativas ordinales

PARÁMETROS DE DISPERSIÓN.

Parámetros de dispersión. Son datos que informan de la concentración o dispersión de los datos respecto de los parámetros de centralización.

Por ejemplo, vamos a suponer que hemos realizado el mismo examen en dos grupos distintos. En uno, todos los alumnos han sacado la misma nota, un 5; en otro, la mitad de los alumnos ha sacado un 0 y la otra mitad un 10. ¿Cuál es la media en los dos casos? ¿Se pueden considerar los dos grupos iguales si la media coincide?

Parece entonces que no es suficiente con las medidas de centralización, hace falta otros parámetros que informen sobre la mayor o menor concentración de los datos.

Recorrido. Se define el recorrido como la diferencia entre el mayor y el menor de los valores de la variable. Se representa por R. Nos indica un intervalo en el que están

comprendido todos los datos.

A veces puede ocurrir que hay valores de la variable, excesivamente pequeños o grandes que hacen que la información que proporciona el recorrido sea equivocada, por ejemplo si en la estatura tenemos todos los alumnos y alumnas con una estatura normal y uno o una mide alrededor de dos metros. Para estos casos es más útil el siguiente parámetro.

Recorrido intercuartílico. Es la diferencia entre los cuartiles tercero y primero. Se representa por R_I ($R_I=C_3-C_1$) y representa la amplitud del intervalo en el que se encuentra el 50% central de los datos.

Desviación media. Al calcular la media, podemos ver la diferencia que hay entre este parámetro y cada valor de la variable, a la que llamaremos desviación. Podemos definir la desviación media como la media aritmética de todas las desviaciones, pero si la calculamos nos llevaremos la sorpresa de que vale 0. ¿Por qué?

Para evitar esta situación, se define la desviación media como la media aritmética de los valores absolutos de las desviaciones respecto de la media. La podremos calcular con la fórmula:

$$Dm = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots + |x_N - \bar{x}|}{N} = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}| \cdot f_i}{N}$$

En la siguiente escena se puede calcular la desviación media.

Escena 14. Cálculo de la desviación media.

Varianza. Se define la varianza como la media aritmética de los cuadrados de las desviaciones respecto de la media.

Para calcularla, aplicamos la fórmula:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot f_i}{N}$$

Si desarrollamos esta fórmula, podemos encontrar otra expresión más sencilla para el cálculo de la varianza:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot f_i}{N} = \frac{\sum_{i=1}^{n} (x_i^2 - 2x_i \cdot \bar{x} + \bar{x}^2) \cdot f_i}{N} = \frac{\sum_{i=1}^{n} x_i^2 \cdot f_i}{N} - 2 \cdot \frac{\sum_{i=1}^{n} x_i \cdot \bar{x} \cdot f_i}{N} + \frac{\sum_{i=1}^{n} \bar{x}^2 \cdot f_i}{N} =$$

$$= \frac{\sum_{i=1}^{n} x_i^2 \cdot f_i}{N} - 2\bar{x} \cdot \frac{\sum_{i=1}^{n} x_i \cdot f_i}{N} + \bar{x}^2 \cdot \frac{\sum_{i=1}^{n} f_i}{N} = \frac{\sum_{i=1}^{n} x_i^2 \cdot f_i}{N} - 2\bar{x} \cdot \bar{x} + \bar{x}^2 \cdot \frac{N}{N} = \frac{\sum_{i=1}^{n} x_i^2 \cdot f_i}{N} - 2\bar{x}^2 + \bar{x}^2 = \frac{\sum_{i=1}^{n}}{}$$

La fórmula que simplifica el cálculo de la varianza es:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 \cdot f_i}{N} - \bar{x}^2} \quad \sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 \cdot f_i}{N} - \bar{x}^2$$

3 era unidad

Los porcentajes acumulados

Puede incluir una columna o una fila en el informe que muestre un total acumulado. El total acumulado se puede expresar como un valor numérico o un porcentaje. En informes de Reporter, se puede calcular un total acumulado para más de una categoría.

Por ejemplo, puede crear un informe que muestre los ingresos de cada uno de los cuatro últimos trimestres. El total acumulado mostrará los ingresos totales al final de cada trimestre. Si añade un total acumulado como porcentaje del total vendido, podrá ver el porcentaje de ventas de todo el año conseguidas al final del trimestre.

Puede suprimir la categoría que representa el valor base del informe una vez creado el porcentaje acumulativo del valor base.

En modo Explorer, sólo puede seleccionar una categoría sobre la que calcular un total acumulado como porcentaje numérico o acumulativo del valor base.

No puede calcular el porcentaje acumulativo del cálculo base sobre una categoría de clasificación.

Nota: En modo Reporter, si la categoría seleccionada incluye un cálculo existente, el valor del cálculo se incluye en el total acumulado.

•Cálculo de los totales acumulados como valores numéricos

Se pueden mostrar totales acumulados como valores numéricos.

•Cálculo totales acumulados como valores de porcentaje se pueden mostrar totales acumulados como valores de porcentaje

Las puntuaciones típicas

Las puntuaciones directas (puntuaciones de un sujeto en un test, etc.) son los primeros datos de los que habitualmente disponemos pero la comparación de las puntuaciones directas de un mismo sujeto en dos variables puede llevarnos a confusión. De hecho, conocida una puntuación directa no sabemos si se trata de un valor alto o bajo porque esto depende del promedio del grupo.

Si a una puntuación directa Xi le restamos la media de su grupo obtenemos una puntuación diferencial o de diferencia, que representamos por xi (minúscula):

$$xi = Xi - \bar{X}$$

Las puntuaciones diferenciales nos indican si la puntuación coincide con la media de su grupo, es inferior o es superior a ella. Tienen las siguientes propiedades:

•Su media es cero: $\bar{x} = 0$

•La varianza de las puntuaciones diferenciales es igual a la varianza de las puntuaciones directas.

Por tanto, al restar a las puntuaciones directas su media hemos obtenido una nueva escala con media 0 y con idéntica varianza a las puntuaciones directas.

Sin embargo, dos puntuaciones diferenciales idénticas pueden tener un significado muy diferente en función de la media y de la varianza de las distribuciones de las que proceden. Para eliminar este inconveniente se utilizan las puntuaciones típicas que nos permiten no sólo comparar las puntuaciones de un sujeto en dos variables distintas sino también comparar dos sujetos distintos en dos pruebas o variables distintas.

Al proceso de obtener puntuaciones típicas se llama tipificación, y las puntuaciones se denominan también "tipificadas".

Las puntuaciones típicas tienen las siguientes propiedades:

•Su media es cero

•Su varianza es igual a 1

Las puntuaciones típicas reflejan las relaciones entre las puntuaciones con independencia de la unidad de medida. Así, permiten hacer comparaciones entre distintos grupos e incluso entre distintas variables.

Introducción

Las medidas de posición nos facilitan información sobre la serie de datos que estamos analizando. La descripción de un conjunto de datos, incluye como un elemento de importancia la ubicación de éstos dentro de un contexto de valores posible. Una vez definidos los conceptos básicos en el estudio de una distribución de frecuencias de una variable, estudiaremos las distintas formas de resumir dichas distribuciones mediante medidas de posición (o de centralización), teniendo presente el error cometido en el resumen mediante las correspondientes medidas de dispersión.
Se trata de encontrar unas medidas que sinteticen las distribuciones de frecuencias. En vez de manejar todos los datos sobre las variables, tarea que puede ser pesada, podemos caracterizar su distribución de frecuencias mediante algunos valores numéricos, eligiendo como resumen de los datos un valor central alrededor del cual se encuentran distribuidos los valores de la variable Son medidas estadísticas cuyo valor representa el valor del dato que se encuentra en el centro de la distribución de frecuencia, por lo que también se les llama "Medidas de Tendencia Central ".

2. Medidas de Posición

Son indicadores usados para señalar que porcentaje de datos dentro de una distribución de frecuencias superan estas expresiones, cuyo valor representa el valor del dato que se encuentra en el centro de la distribución de frecuencia, por lo que también se les llama " Medidas de Tendencia Central ".

Pero estas medidas de posición de una distribución de frecuencias han de cumplir determinadas condiciones para que lean verdaderamente representativas de la variable a la que resumen. Toda síntesis de una distribución se considerara como operativa si intervienen en su determinación todos y cada uno de los valores de la distribución, siendo única para cada distribución de frecuencias y siendo siempre calculable y de fácil obtención. A continuación se describen las medidas de posición más comunes utilizadas en estadística, como lo son:

Cuartiles: Hay 3 cuartiles que dividen a una distribución en 4 partes iguales: primero, segundo y tecer cuartil.

Deciles: Hay 9 deciles que la dividen en 10 partes iguales: (primero al noveno decil).

Percentiles: Hay 99 percentiles que dividen a una serie en 100 partes iguales: (primero al noventa y nueve percentil).

Cuartiles (Q1, Q2, Q3)

Aquel valor de una serie que supera al 25% de los datos y es superado por el 75% restante.

Formula de Q1 para series de Datos Agrupados en Clase.

$$Q_1 = L_i + \frac{\frac{\sum f_i}{4} - f_{aa}}{f_i} * I_c$$

Donde:

$\frac{\sum f_i}{4}$: posición de Q1, la cual se localiza en la primera frecuencia acumulada que la contenga, siendo la clase de Q1, la correspondiente a tal frecuencia acumulada.

Li, faa, fi, Ic : idéntico a los conceptos vistos para Mediana pero referidos a la medida de la posición correspondiente.

Primer cuartil (Q1):

Segundo cuartil (Q2):

Coincide, es idéntico o similar al valor de la Mediana (Q2 = Md). Es decir, supera y es superado por el 50% de los valores de una Serie.

c) Tercer cuartil (Q3):

Aquel valor, termino o dato que supera al 75% y es superado por el 25% de los datos restantes de la Serie.
Formula de Q3 para series de Datos Agrupados en Clase.

$$Q_3 = L_i + \frac{\frac{3\sum f_i}{4} - f_{aa}}{f_i} * I_c$$

Donde:

$\frac{3\sum f_i}{4}$: posición de Q3, todo idéntico al calculo de la Mediana.

Deciles. (D1,D2,…D9) Primer Decil (D1), Quinto Decil (D5) y Noveno Decil (D9). El primer decil es aquel valor de una serie que supera a 1/10 parte de los datos y es superado por las 9/10 partes restantes (respectivamente, hablando en porcentajes, supera al 10% y es superado por el 90% restante),

$$D_1 = L_i + \frac{\frac{*\sum f_i}{10} - f_{aa}}{f_i} * I_c$$

$$D_5 = L_i + \frac{\frac{*5\sum f_i}{10} - f_{aa}}{f_i} * I_c = M_d$$

$$D_9 = L_i + \frac{\frac{*9\sum f_i}{10} - f_{aa}}{f_i} * I_c$$

El D9 (noveno decil) supera al 90% y es superado por el 10% restante.

Como se observa, son formulas parecidas a la del calculo de la Mediana, cambiando solamente la respectivas posiciones de las medidas.

Percentiles (P1,P2,…P99) Primer Percentil (P1), Percentil 50 (P50) y Percentil 99 (P99). El primer percentil supera al uno por ciento de los valores y es superado por el noventa y nueve por ciento restante.Fórmulas de P1, P50, P99 para series de Datos Agrupados en Clase.

$$P_1 = L_i + \frac{\frac{*\sum f_i}{100} - f_{aa}}{f_i} * I_c$$

$$P_{50} = L_i + \frac{\frac{*50\sum f_i}{100} - f_{aa}}{f_i} * I_c = M_d$$

$$p_{99} = L_i + \frac{\frac{*99\sum f_i}{100} - f_{aa}}{f_i} * I_c$$

El P99 (noventa y nueve percentil) supera al 99% de los datos y es superado a su vez por el 1% restante.

Idénticas formulas al calculo de la Mediana, cambiando obviamente las correspondientes posiciones de cada medida.

Para determinar estas medidas se aplicara el principio de la mediana; así, el primer cuartil cereal valor por debajo del cual se encuentra el 25 por ciento de los datos; bajo el tecer cuartil se encuentra el 75 por ciento; el 80 decil será el valor por encima del cual estará el 20

por ciento de los datos, etc. Como se observa, todas estas medidas no son sino casos particulares del percentil ya que el primer cuartil no es sino el 25° percentil, el tercer cuartil el 75° percentil, el cuarto decil el 40° percentil, etc.

Datos no agrupados: Se hace difícil calcular estas medidas, sin embargo, siguiendo los mismos principios mencionados para la Mediana, se pueden localizar en la forma siguiente:

Si tenemos una serie de valores X1, X2, X3 … Xn, se localiza el primer cuartil como el valor $\frac{1*n}{4}$ cuando n es par, y $\frac{1(n+1)}{4}$ cuando n es impar. Para el tercer cuartil será $\frac{3*n}{4}$ (n par); $\frac{3(n+1)}{4}$ (n impar).

En caso de los textiles será $\frac{A*n}{6}$ o $\frac{A(n+1)}{6}$ donde A representa el número del textil.

Para los deciles será $\frac{A*n}{10}$ o $\frac{A(n+1)}{10}$ siendo A el número del decil; y para los percentiles $\frac{A*n}{100}$ o $\frac{A(n+1)}{100}$.

Ejemplo:

En una serie de 32 términos se desea localizar el 4° sextil, 8° decil y el 95° percentil.

$$4^{\blacktriangle}sextil = \frac{4*32}{6} = 21$$

$$8^{\blacktriangle}decil = \frac{8*32}{10} = 25.6$$

$$95^{\blacktriangle}percentil = \frac{95*2}{100} = 30,4$$

Esto significa que el 4° textil se encuentra localizado en el término número 21, es decir, el que ocupa la 21° posición; el 8° decil se encuentra localizado entre el termino numero 25° y 26° ; y el 95° percentil entre la posición 30° y 31° .

Calculo para una distribución de frecuencia

Para el cálculo de esta medida en datos agrupados en una distribución de frecuencia, se utiliza el mismo procedimiento estudiado para el cálculo de la Mediana, e; cual es:

Se efectúa la columna de las frecuencias acumuladas.

Se determina la posición del término cuyo valor se pretende calcular, en caso de ser el primer cuartil será $\frac{1*\sum f_i}{4}$, si fuese el 95° centil $\frac{95*\sum f_i}{100}$ … etc.

Se verifica cual es la clase que lo contiene; para ello se utiliza la columna de las frecuencias acumuladas.

Se hace la diferencia entre el número que representa el orden de posición cuyo valor se pretende calcular y la frecuencia acumulada de la clase anterior a la que lo contiene.

Se calcula la medida solicitada de acuerdo a la siguiente fórmula:

$$P = 1_i + \frac{P - f_{a-1}}{f_i} * I_c$$

Donde:

1i: limite inferior de la clase que lo contiene.P: valor que representa la posición de la medida. fi: la frecuencia de la clase que contiene la medida solicitada. fa-1: frecuencia acumulada anterior a la que contiene la medida solicitada. Ic: intervalo de clase.

Ejemplo:

Determinación del primer cuartil, el cuartil textil, el séptimo decil y el 30° percentil.

Salarios (l. de Clases)	N° de empleados (fi)	fa
200 – 299	85	85
300 – 399	90	175
400 – 499	120	295
500 – 599	70	365
600 – 699	62	427
700 – 800	36	463

$\frac{463}{4} = 115,5$ $L_i = 300$ $115,5 - 85 = 30,75$

$f_i = 90$ $I_c = 100$ $Q_1 = 300 + \frac{30,75}{90} * 100 = 334$

$4^{\blacktriangle} sextil: posición = \frac{4(463)}{6} = \frac{1852}{6} = 308,66$

$308,66 - 295 = 13,66$ $f_i = 70$

$4^{\blacktriangle} S = 500 + \frac{13,66}{70} * 100 = 59,51$

$7^{\blacktriangle} decil: posición = \frac{7(463)}{10} = \frac{3241}{10} = 324.,1$

$324,1 - 295 = 29,1$ $f_i = 70$

$7^{\blacktriangle} D = 500 + \frac{29,1}{70} * 100 = 541,57$

$$30^{\blacktriangle} percentil: posición = \frac{30(463)}{100} = \frac{13890}{100} = 138,9$$

$$138,9 - 85 = 53,9 \quad f_i = 90$$

$$30^{\blacktriangle} P = 300 + \frac{53,9}{90} * 100 = 359,88$$

Estos resultados nos indican que el 25 por ciento de los empleados ganan salarios por debajo de Bs. 334; que sobre Bs. 519,51 ganan el 33,33 por ciento de los empleados; que bajo 541,57 gana el 57 por ciento de los empleados y sobre Bs. 359,88 gana el 70 por ciento de los empleados. Muchas veces necesitamos conocer el porcentaje de valores que esta por debajo o por encima de un valor dado; lo que representa un problema contrario al anterior, esto es, dado un cierto valor en la abscisa determinar en la ordenada el tanto por ciento de valores inferiores y superiores al valor dado. Operación que se resuelve utilizando la siguiente formula general:

$$P = \left[f_{a-1} + \frac{f_i(P - L_i)}{Ic} \right] \frac{100}{N}$$

Dónde:

P: lugar percentil que se busca. P: valor reconocido en la escala X.

fa-1: frecuencia acumulada de la clase anterior a la clase en que está incluida P.

fi: frecuencia de la clase que contiene a p.

Li: límite inferior de la clase que contiene a P.

Ic: intervalo de clase.

N: frecuencia total.

Ejemplo:

Utilizando la distribución anterior, determinar qué porcentaje de personas ganan salarios inferiores a Bs. 450,00

$$P = \left[175 + \frac{120(450 - 400)}{100}\right]\frac{100}{463} = 50{,}75$$

El 50,75 por ciento de las personas ganan salarios inferiores a Bs. 450.

Método gráfico para fraccionar la distribución Se pueden obtener en forma gráfica, a través de la curva de la frecuencia acumulada (ojiva). Para ello basta después de trazar la ojiva, llevar el orden de posición de la medida que se quiere sobre la ordenada, trazar por ese punto una perpendicular toca a la ojiva, baja una paralela a la ordenada hasta tocar la abscisa; en el punto donde toque a dicho eje, se encontrará el valor buscado. Obtención gráfica de las medidas de posición Similar o idéntico a la distribución grafica de la Mediana con la sola excepción de que se llevaría al eje vertical (frecuencias acumuladas) las especificas posiciones de cada indicador de posición en particular.

Ejemplo:

Forma de obtener los indicadores de posición (cuartiles, deciles y percentiles) para series de datos agrupados en clases: Supongamos la siguiente distribución de frecuencias referidas a las estaturas que representaban 40 alumnos de un curso.

(I. de Clases)	Estaturas (mts)	N° alumnos (fi)	fa
1,60	1,639	5	5
1,64	1,679	8	13
** 1,68	1,719	15	** 28
* 1,72	1,759	10	38 *
1,76	1,80	2	40

$\sum f_i = 40$

Q3=?

$$Q_3 = L_i + \frac{\frac{3\sum f_i}{4} - f_{aa}}{f_i} * I_c$$

$$posición * Q_3 = \frac{3\sum f_i}{4} = \frac{120}{4} = *30$$

La cual se ubica en la primera fa que la contenga

$$Q_3 = 1{,}72 + \frac{30 - 28}{10} * 0{,}04$$

$$Q_3 = 1{,}72 + 0{,}008 \approx 1{,}73 mts$$

Esta estatura de Q3 = 1,73 mts. Supera en la distribución de frecuencia al 75% de los alumnos del curso y es superada por el 25% de los mismos.

D8 = ?

$$D_8 = \frac{\frac{8\sum f_i}{10} - f_{aa}}{f_i} I_c$$

$$posición * D_8 = \frac{8\sum f_i}{10} = \frac{320}{10} = *32$$

$$D_8 = 1{,}72 + \frac{32 - 28}{10} * 0{,}04$$

$$D_8 = 1{,}72 + 0{,}016 \approx 1{,}736 mts$$

Supera esta estatura de 1,736 mts a 8/10 partes de curso y es superado por las 2/10 partes

restantes.

P55 = ?

$$P_{55} = L_i + \frac{\frac{55\sum f_i}{100} - f_{aa}}{f_i} * I_c$$

$$posición * P_{55} = \frac{(55)\sum f_i}{100} = **22$$

$$P_{55} = 1{,}68 + \frac{22 - 13}{15} * 0{,}04$$

$$P_{55} = 1{,}68 + 0{,}024 \approx 1{,}70 mts$$

Esta estatura supera al 55% de los alumnos del curso y es superada por el 45% restante.

Calcular de cada uno de los intervalos de clases cuartiles, deciles y percentiles.

Datos agrupados

l. de clases	fi	fa
10 – 15	10	10
16 – 21	18	28
22 – 27	10	38
28 – 33	8	46
34 – 39	9	55
40 – 45	7	62

46 – 51	3	65
52 – 57	1	66

n = 66

Cuartiles:

$$Q_3 = L_i + \frac{\frac{3\sum f_i}{4} - f_{ac}}{f_i} * I_c \quad L_i = L_{inf} - 0{,}5 = 10 - 0{,}5 = 9{,}5$$

$$Posic.Q_3 = \frac{3\sum f_i}{4} = \frac{3*66}{4} = \frac{198}{4} = 49{,}5$$

$$Q_3 = 9{,}5 + \frac{49{,}5 - 0}{10} * 6 = 39{,}2$$

$$Q_3 = L_i + \frac{\frac{3\sum f_i}{4} - f_{ac}}{f_i} * I_c \quad L_i = L_{inf} - 0{,}5 = 16 - 0{,}5 = 15{,}5$$

$$Posic.Q_3 = \frac{3\sum f_i}{4} = \frac{3*66}{4} = \frac{198}{4} = 49{,}5$$

$$Q_3 = 15{,}5 + \frac{49{,}5 - 10}{18} * 6 = 28{,}66$$

$$Q_3 = L_i + \frac{\frac{3\sum f_i}{4} - f_{ac}}{f_i} * I_c \quad L_i = L_{inf} - 0{,}5 = 22 - 0{,}5 = 21{,}5$$

$$Posic.Q_3 = \frac{3\sum f_i}{4} = \frac{3*66}{4} = \frac{198}{4} = 49{,}5$$

$$Q_3 = 21{,}5 + \frac{49{,}5 - 28}{10} * 6 = 34{,}4$$

$$Q_3 = L_i + \frac{\frac{3\sum f_i}{4} - f_{ac}}{f_i} * I_c \quad L_i = L_{inf} - 0{,}5 = 28 - 0{,}5 = 27{,}5$$

$$Posic.Q_3 = \frac{3\sum f_i}{4} = \frac{3*66}{4} = \frac{198}{4} = 49{,}5$$

$$Q_3 = 27{,}5 + \frac{49{,}5 - 38}{8} * 6 = 36{,}12$$

$$Q_3 = L_i + \frac{\frac{3\sum f_i}{4} - f_{ac}}{f_i} * I_c \quad L_i = L_{inf} - 0{,}5 = 34 - 0{,}5 = 33{,}5$$

$$Posic.Q_3 = \frac{3\sum f_i}{4} = \frac{3*66}{4} = \frac{198}{4} = 49{,}5$$

$$Q_3 = 33{,}5 + \frac{49{,}5 - 46}{9} * 6 = 35{,}83$$

$$Q_3 = L_i + \frac{\frac{3\sum f_i}{4} - f_{ac}}{f_i} * I_c \quad L_i = L_{inf} - 0{,}5 = 40 - 0{,}5 = 39{,}5$$

$$Posic.Q_3 = \frac{3\sum f_i}{4} = \frac{3*66}{4} = \frac{198}{4} = 49{,}5$$

$$Q_3 = 39{,}5 + \frac{49{,}5 - 55}{7} * 6 = 34{,}78$$

$$Q_3 = L_i + \frac{\frac{3\sum f_i}{4} - f_{ac}}{f_i} * I_c \quad L_i = L_{inf} - 0{,}5 = 46 - 0{,}5 = 45{,}5$$

$$Posic.Q_3 = \frac{3\sum f_i}{4} = \frac{3*66}{4} = \frac{198}{4} = 49{,}5$$

$$Q_3 = 45{,}5 + \frac{49{,}5 - 62}{3} * 6 = 20{,}5$$

$$Q_3 = L_i + \frac{\frac{3\sum f_i}{4} - f_{aa}}{f_i} * I_c \quad L_i = L_{inf} - 0{,}5 = 52 - 0{,}5 = 51{,}5$$

$$Posic.Q_3 = \frac{3\sum f_i}{4} = \frac{3*66}{4} = \frac{198}{4} = 49{,}5$$

$$Q_3 = 51{,}5 + \frac{49{,}5 - 65}{1} * 6 = -41.5$$

Deciles:

$$D_9 = L_i + \frac{\frac{*9\sum f_i}{10} - f_{aa}}{f_i} * I_c$$

$$Posic.D_9 = \frac{*9\sum f_i}{10} = \frac{9*66}{10} = 59{,}4$$

$$D_9 = 9{,}5 + \frac{59{,}4 - 0}{10} * 6 = 45{,}14$$

$$D_8 = L_i + \frac{\frac{*8\sum f_i}{10} - f_{aa}}{f_i} * I_c$$

$$Posic.D_8 = \frac{*8\sum f_i}{10} = \frac{8*66}{10} = 52{,}8$$

$$D_8 = 15{,}5 + \frac{52{,}8 - 10}{18} * 6 = 29{,}76$$

$$D_3 = L_i + \frac{*3\sum f_i}{\frac{10}{f_i}} - f_{aa} * I_c$$

$$Posic.D_3 = \frac{*3\sum f_i}{10} = \frac{3*66}{10} = 19{,}8$$

$$D_3 = 21{,}5 + \frac{19{,}8 - 28}{10} * 6 = 16{,}5$$

$$D_7 = L_i + \frac{*7\sum f_i}{\frac{10}{f_i}} - f_{aa} * I_c$$

$$Posic.D_7 = \frac{*7\sum f_i}{10} = \frac{7*66}{10} = 46{,}2$$

$$D_7 = 27{,}5 + \frac{46{,}2 - 38}{8} * 6 = 33{,}65$$

$$D_6 = L_i + \frac{*6\sum f_i}{\frac{10}{f_i}} - f_{aa} * I_c$$

$$Posic.D_6 = \frac{*6\sum f_i}{10} = \frac{6*66}{10} = 39{,}6$$

$$D_6 = 33{,}5 + \frac{39{,}6 - 46}{9} * 6 = 29{,}23$$

$$D_9 = L_i + \frac{*9\sum f_i}{\frac{10}{f_i}} - f_{aa} * I_c$$

$$Posic.D_9 = \frac{*9\sum f_i}{10} = \frac{9*66}{10} = 59{,}4$$

$$D_9 = 39{,}5 + \frac{59{,}4 - 55}{7} * 6 = 43{,}27$$

$$D_9 = L_i + \frac{\frac{*9\sum f_i}{10} - f_{aa}}{f_i} * I_c$$

$$Posic.D_9 = \frac{*9\sum f_i}{10} = \frac{9*66}{10} = 59{,}4$$

$$D_9 = 45{,}5 + \frac{59{,}4 - 62}{3} * 6 = 40{,}3$$

$$D_9 = L_i + \frac{\frac{*9\sum f_i}{10} - f_{aa}}{f_i} * I_c$$

$$Posic.D_9 = \frac{*9\sum f_i}{10} = \frac{9*66}{10} = 59{,}4$$

$$D_9 = 51{,}5 + \frac{59{,}4 - 65}{1} * 6 = 17{,}9$$

Percentiles:

$$P_{99} = L_i + \frac{\frac{*99\sum f_i}{100} - f_{aa}}{f_i} * I_c$$

$$Posic.P_{99} = \frac{*99\sum f_i}{100} = \frac{99*66}{100} = 65{,}34$$

$$P_{99} = 9{,}5 + \frac{65{,}34 - 0}{10} * 6 = 48{,}70$$

$$P_{55} = L_i + \frac{\frac{*55\sum f_i}{100} - f_{aa}}{f_i} * I_c$$

$$Posic.P_{55} = \frac{*55\sum f_i}{100} = \frac{55*66}{100} = 36,3$$

$$P_{55} = 15,5 + \frac{36,3 - 10}{18} * 6 = 24,26$$

$$P_{35} = L_i + \frac{\frac{*35\sum f_i}{100} - f_{aa}}{f_i} * I_c$$

$$Posic.P_{35} = \frac{*35\sum f_i}{100} = \frac{35*66}{100} = 23,1$$

$$P_{35} = 21,5 + \frac{23,1 - 28}{10} * 6 = 18,56$$

$$P_{39} = L_i + \frac{\frac{*39\sum f_i}{100} - f_{aa}}{f_i} * I_c$$

$$Posic.P_{39} = \frac{*39\sum f_i}{100} = \frac{39*66}{100} = 25,74$$

$$P_{39} = 27,5 + \frac{25,74 - 38}{8} * 6 = 18,30$$

$$P_{28} = L_i + \frac{\frac{*28\sum f_i}{100} - f_{aa}}{f_i} * I_c$$

$$Posic.P_{28} = \frac{*28\sum f_i}{100} = \frac{28*66}{100} = 18,48$$

$$P_{28} = 33{,}5 + \frac{18{,}48 - 46}{9} * 6 = 15{,}15$$

$$P_{20} = L_i + \frac{\frac{*20\sum f_i}{100} - f_{aa}}{f_i} * I_c$$

$$Posic.P_{20} = \frac{*20\sum f_i}{100} = \frac{20 * 66}{100} = 13{,}2$$

$$P_{20} = 39{,}5 + \frac{13{,}2 - 55}{7} * 6 = 3{,}67$$

Representación gráfica en el Análisis de Datos

Introducción

La realización de los estudios clínico-epidemiológicos implica finalmente emitir unos resultados cuantificables de dicho estudio o experimento. La claridad de dicha presentación es de vital importancia para la comprensión de los resultados y la interpretación de los mismos. A la hora de representar los resultados de un análisis estadístico de un modo adecuado, son varias las publicaciones que podemos consultar. Aunque se aconseja que la presentación de datos numéricos se haga habitualmente por medio de tablas, en ocasiones un diagrama o un gráfico pueden ayudarnos a representar de un modo más eficiente nuestros datos.

En este artículo se abordará la representación gráfica de los resultados de un estudio, constatando su utilidad en el proceso de análisis estadístico y la presentación de datos. Se describirán los distintos tipos de gráficos que podemos utilizar y su correspondencia con las distintas etapas del proceso de análisis.

Análisis descriptivo

Cuando se dispone de datos de una población, y antes de abordar análisis estadísticos más

complejos, un primer paso consiste en presentar esa información de forma que ésta se pueda visualizar de una manera más sistemática y resumida. Los datos que nos interesan dependen, en cada caso, del tipo de variables que estemos manejando.

Para variables categóricas, como el sexo, estadio TNM, profesión, etc., se quiere conocer la frecuencia y el porcentaje del total de casos que "caen" en cada categoría. Una forma muy sencilla de representar gráficamente estos resultados es mediante diagramas de barras o diagramas de sectores. En los gráficos de sectores, también conocidos como diagramas de "tartas", se divide un círculo en tantas porciones como clases tenga la variable, de modo que a cada clase le corresponde un arco de círculo proporcional a su frecuencia absoluta o relativa. Un ejemplo se muestra en la Figura 1. Como se puede observar, la información que se debe mostrar en cada sector hace referencia al número de casos dentro de cada categoría y al porcentaje del total que estos representan. Si el número de categorías es excesivamente grande, la imagen proporcionada por el gráfico de sectores no es lo suficientemente clara y por lo tanto la situación ideal es cuando hay alrededor de tres categorías. En este caso se pueden apreciar con claridad dichos subgrupos.

Los diagramas de barras son similares a los gráficos de sectores. Se representan tantas barras como categorías tiene la variable, de modo que la altura de cada una de ellas sea proporcional a la frecuencia o porcentaje de casos en cada clase (Figura 2). Estos mismos gráficos pueden utilizarse también para describir variables numéricas discretas que toman pocos valores (número de hijos, número de recidivas, etc.).

Para variables numéricas continuas, tales como la edad, la tensión arterial o el índice de masa corporal, el tipo de gráfico más utilizado es el histograma. Para construir un gráfico de este tipo, se divide el rango de valores de la variable en intervalos de igual amplitud, representando sobre cada intervalo un rectángulo que tiene a este segmento como base. El criterio para calcular la altura de cada rectángulo es el de mantener la proporcionalidad entre las frecuencias absolutas (o relativas) de los datos en cada intervalo y el área de los rectángulos. Como ejemplo, laTabla I muestra la distribución de frecuencias de la edad de 100 pacientes, comprendida entre los 18 y 42 años. Si se divide este rango en intervalos de

dos años, el primer tramo está comprendido entre los 18 y 19 años, entre los que se encuentra el 4/100=4% del total. Por lo tanto, la primera barra tendrá altura proporcional a 4. Procediendo así sucesivamente, se construye el histograma que se muestra en la Figura 3. Uniendo los puntos medios del extremo superior de las barras del histograma, se obtiene una imagen que se llama polígono de frecuencias. Dicha figura pretende mostrar, de la forma más simple, en qué rangos se encuentra la mayor parte de los datos. Un ejemplo, utilizando los datos anteriores, se presenta en la Figura 4.

Otro modo habitual, y muy útil, de resumir una variable de tipo numérico es utilizando el concepto de percentiles, mediante diagramas de cajas. La Figura 5 muestra un gráfico de cajas correspondiente a los datos de la Tabla 1. La caja central indica el rango en el que se concentra el 50% central de los datos. Sus extremos son, por lo tanto, el 1^{er} y 3^{er} cuartil de la distribución. La línea central en la caja es la mediana. De este modo, si la variable es simétrica, dicha línea se encontrará en el centro de la caja. Los extremos de los "bigotes" que salen de la caja son los valores que delimitan el 95% central de los datos, aunque en ocasiones coinciden con los valores extremos de la distribución. Se suelen también representar aquellas observaciones que caen fuera de este rango (outliers o valores extremos). Esto resulta especialmente útil para comprobar, gráficamente, posibles errores en nuestros datos. En general, los diagramas de cajas resultan más apropiados para representar variables que presenten una gran desviación de la distribución normal. Como se verá más adelante, resultan además de gran ayuda cuando se dispone de datos en distintos grupos de sujetos.

Por último, y en lo que respecta a la descripción de los datos, suele ser necesario, para posteriores análisis, comprobar la normalidad de alguna de las variables numéricas de las que se dispone. Un diagrama de cajas o un histograma son gráficos sencillos que permiten comprobar, de un modo puramente visual, la simetría y el "apuntamiento" de la distribución de una variable y, por lo tanto, valorar su desviación de la normalidad. Existen otros métodos gráficos específicos para este propósito, como son los gráficos P-P o Q-Q. En los primeros, se confrontan las proporciones acumuladas de una variable con las de una distribución normal. Si la variable seleccionada coincide con la distribución de prueba, los puntos se

concentran en torno a una línea recta. Los gráficos Q-Q se obtienen de modo análogo, esta vez representando los cuantiles de distribución de la variable respecto a los cuantiles de la distribución normal. En la Figura 6 se muestra el gráfico P-P correspondientes a los datos de la Tabla I que sugiere, al igual que el correspondiente histograma y el diagrama de cajas, que la distribución de la variable se aleja de la normalidad.

Comparación de dos o más grupos

Cuando se quieren comparar las observaciones tomadas en dos o más grupos de individuos una vez más el método estadístico a utilizar, así como los gráficos apropiados para visualizar esa relación, dependen del tipo de variables que estemos manejando.

Cuando se trabaja con dos variables cualitativas podemos seguir empleando gráficos de barras o de sectores. Podemos querer determinar, por ejemplo, si en una muestra dada, la frecuencia de sujetos que padecen una enfermedad coronaria es más frecuente en aquellos que tienen algún familiar con antecedentes cardiacos. A partir de dicha muestra podemos representar, como se hace en la Figura 7, dos grupos de barras: uno para los sujetos con antecedentes cardiacos familiares y otro para los que no tienen este tipo de antecedentes. En cada grupo, se dibujan dos barras representando el porcentaje de pacientes que tienen o no alguna enfermedad coronaria. No se debe olvidar que cuando los tamaños de las dos poblaciones son diferentes, es conveniente utilizar las frecuencias relativas, ya que en otro caso el gráfico podría resultar engañoso.

Por otro lado, la comparación de variables continuas en dos o más grupos se realiza habitualmente en términos de su valor medio, por medio del test t de Student, análisis de la varianza o métodos no paramétricos equivalentes, y así se ha de reflejar en el tipo de gráfico utilizado. En este caso resulta muy útil un diagrama de barras de error, como en la Figura 8. En él se compara el índice de masa corporal en una muestra de hombres y mujeres. Para cada grupo, se representa su valor medio, junto con su 95% intervalo de confianza. Conviene recordar que el hecho de que dichos intervalos no se solapen, no implica necesariamente que la diferencia entre ambos grupos pueda ser estadísticamente significativa, pero sí nos puede

servir para valorar la magnitud de la misma. Así mismo, para visualizar este tipo de asociaciones, pueden utilizarse dos diagramas de cajas, uno para cada grupo. Estos diagramas son especialmente útiles aquí: no sólo permiten ver si existe o no diferencia entre los grupos, sino que además nos permiten comprobar la normalidad y la variabilidad de cada una de las distribuciones. No olvidemos que las hipótesis de normalidad y homocedasticidad son condiciones necesarias para aplicar algunos de los procedimientos de análisis paramétricos.

Por último, señalar que también en esta situación pueden utilizarse los ya conocidos gráficos de barras, representando aquí como altura de cada barra el valor medio de la variable de interés. Los gráficos de líneas pueden resultar también especialmente interesantes, sobre todo cuando interesa estudiar tendencias a lo largo del tiempo (Figura 9). No son más que una serie de puntos conectados entre sí mediante rectas, donde cada punto puede representar distintas cosas según lo que nos interese en cada momento (el valor medio de una variable, porcentaje de casos en una categoría, el valor máximo en cada grupo, etc).

Relación entre dos variables numéricas.

Cuando lo que interesa es estudiar la relación entre dos variables continuas, el método de análisis adecuado es el estudio de la correlación. Los coeficientes de correlación (Pearson, Spearman, etc.) valoran hasta qué punto el valor de una de las variables aumenta o disminuye cuando crece el valor de la otra. Cuando se dispone de todos los datos, un modo sencillo de comprobar, gráficamente, si existe una correlación alta, es mediante diagramas de dispersión, donde se confronta, en el eje horizontal, el valor de una variable y en el eje vertical el valor de la otra. Un ejemplo sencillo de variables altamente correlacionados es la relación entre el peso y la talla de un sujeto. Partiendo de una muestra arbitraria, podemos construir el diagrama de dispersión de la Figura 10. En él puede observarse claramente como existe una relación directa entre ambas variables, y valorar hasta qué punto dicha relación puede modelizarse por la ecuación de una recta. Este tipo de gráficos son, por lo tanto,

especialmente útiles en la etapa de selección de variables cuando se ajusta un modelo de regresión lineal.

Otros gráficos

Los tipos de gráficos mostrados hasta aquí son los más sencillos que podemos manejar, pero ofrecen grandes posibilidades para la representación de datos y pueden ser utilizados en múltiples situaciones, incluso para representar los resultados obtenidos por métodos de análisis más complicados. Podemos utilizar, por ejemplo, dos diagramas de líneas superpuestos para visualizar los resultados de un análisis de la varianza con dos factores (Figura 11). Un diagrama de dispersión es el método adecuado para valorar el resultado de un modelo de regresión logística (Figura 12). Existen incluso algunos análisis concretos que están basados completamente en la representación gráfica. En particular, la elaboración de curvas ROC (Figura 13) y el cálculo del área bajo la curva constituyen el método más apropiado para valorar la exactitud de una prueba diagnóstica.

Hemos visto, por lo tanto, como la importancia y utilidad que las representaciones gráficas pueden alcanzar en el proceso de análisis de datos. La mayoría de los textos estadísticos y epidemiológicos hacen hincapié en los distintos tipos de gráficos que se pueden crear, como una herramienta imprescindible en la presentación de resultados y el proceso de análisis estadístico. No obstante, es difícil precisar cuándo es más apropiado utilizar un gráfico que una tabla. Más bien podremos considerarlos dos modos distintos pero complementarios de visualizar los mismos datos. La creciente utilización de distintos programas informáticos hace especialmente sencillo la obtención de las mismas. La mayoría de los paquetes estadísticos (SPSS, STATGRAPHICS, S-PLUS, EGRET,...) ofrecen grandes posibilidades en este sentido. Además de los gráficos vistos, es posible elaborar otros gráficos, incluso tridimensionales, permitiendo grandes cambios en su apariencia y facilidad de exportación a otros programas para presentar finalmente los resultados del estudio.

Figura 1. Ejemplo de gráfico de sectores. Distribución de una muestra de pacientes según el hábito de fumar.

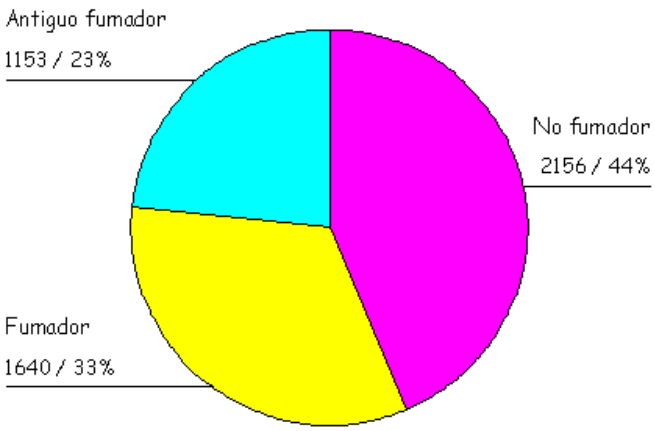


Figura 1

Figura 2. Ejemplo de gráfico de barras. Estadio TNM en el cáncer gástrico.

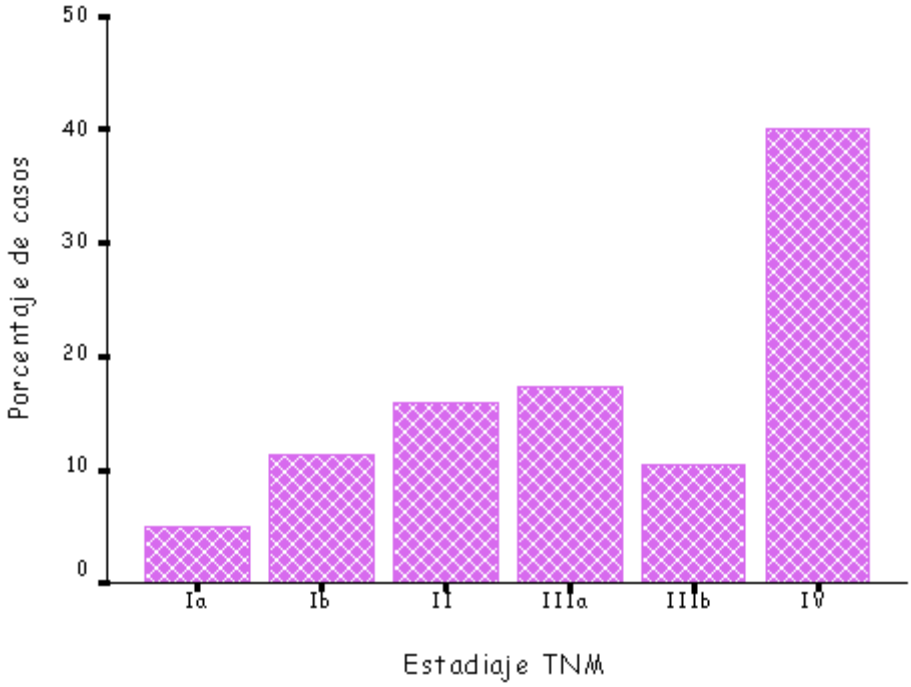

50
40
30
20
10
0
Porcentaje de casos
Ia
Ib
II
IIIa
IIIb
IV
Estadiaje TNM

Figura 2

Unidad IV

Relaciones entre variables

Introducción

En el análisis de los estudios clínico-epidemiológicos surge muy frecuentemente la necesidad de determinar la relación entre dos variables cuantitativas en un grupo de sujetos. Los objetivos de dicho análisis suelen ser:

a. Determinar si las dos variables están correlacionadas, es decir si los valores de una variable tienden a ser más altos o más bajos para valores más altos o más bajos de la otra variable.
b. Poder predecir el valor de una variable dado un valor determinado de la otra variable.
c. Valorar el nivel de concordancia entre los valores de las dos variables

Correlación

En este artículo trataremos de valorar la asociación entre dos variables cuantitativas estudiando el método conocido como correlación. Dicho cálculo es el primer paso para determinar la relación entre las variables. La predicción de una variable. La predicción de una variable dado un valor determinado de la otra precisa de la regresión lineal que abordaremos en otro artículo.

La cuantificación de la fuerza de la relación lineal entre dos variables cuantitativas, se estudia por medio del cálculo del coeficiente de correlación de Pearson. Dicho coeficiente oscila entre –1 y +1. Un valor de –1 indica una relación lineal o línea recta positiva perfecta. Una correlación próxima a cero indica que no hay relación lineal entre las dos variables.

El realizar la representación gráfica de los datos para demostrar la relación entre el valor del coeficiente de correlación y la forma de la gráfica es fundamental ya que existen relaciones no lineales.

El coeficiente de correlación posee las siguientes características:

a. El valor del coeficiente de correlación es independiente de cualquier unidad usada para medir las variables.
b. El valor del coeficiente de correlación se altera de forma importante ante la presencia de un valor extremo, como sucede con la desviación típica. Ante estas situaciones conviene realizar una transformación de datos que cambia la escala de medición y modera el efecto de valores extremos (como la transformación logarítmica).
c. El coeficiente de correlación mide solo la relación con una línea recta. Dos variables pueden tener una relación curvilínea fuerte, a pesar de que su correlación sea pequeña. Por tanto cuando analicemos las relaciones entre dos variables debemos representarlas gráficamente y posteriormente calcular el coeficiente de correlación.
d. El coeficiente de correlación no se debe extrapolar más allá del rango de valores observado de las variables a estudio ya que la relación existente entre X e Y puede cambiar fuera de dicho rango.
e. La correlación no implica causalidad. La causalidad es un juicio de valor que requiere más información que un simple valor cuantitativo de un coeficiente de correlación.

El coeficiente de correlación de Pearson (r) puede calcularse en cualquier grupo de datos, sin embargo la validez del test de hipótesis sobre la correlación entre las variables requiere en sentido estricto: a) que las dos variables procedan de una muestra aleatoria de individuos. b) que al menos una de las variables tenga una distribución normal en la población de la cual la muestra procede. Para el cálculo válido de un intervalo de confianza del coeficiente de correlación de r ambas variables deben tener una distribución normal. Si los datos no tienen una distribución normal, una o ambas variables se pueden transformar (transformación logarítmica) o si no se calcularía un coeficiente de correlación no paramétrico (coeficiente de

correlación de Spearman) que tiene el mismo significado que el coeficiente de correlación de Pearson y se calcula utilizando el rango de las observaciones.

El cálculo del coeficiente de correlación (r) entre peso y talla de 20 niños varones se muestra en la tabla 1. La covarianza, que en este ejemplo es el producto de peso (kg) por talla (cm), para que no tenga dimensión y sea un coeficiente, se divide por la desviación típica de X (talla) y por la desviación típica de Y (peso) con lo que obtenemos el coeficiente de correlación de Pearson que en este caso es de 0.885 e indica una importante correlación entre las dos variables. Es evidente que el hecho de que la correlación sea fuerte no implica causalidad. Si elevamos al cuadrado el coeficiente de correlación obtendremos el coeficiente de determinación (r^2=0.783) que nos indica que el 78.3% de la variabilidad en el peso se explica por la talla del niño. Por lo tanto existen otras variables que modifican y explican la variabilidad del peso de estos niños. La introducción de más variable con técnicas de análisis multivariado nos permitirá identificar la importancia de que otras variables pueden tener sobre el peso.

Tabla 1. Cálculo del Coeficiente de correlación de Pearson entre las variables talla y peso de 20 niños varones

Y	X			
Peso (Kg)	Talla (cm)	$X-\overline{X}$	$Y-\overline{Y}$	$(X-\overline{X})*(Y-\overline{Y})$
9	72	5.65	1.4	7.91
10	76	9.65	2.4	23.16
6	59	-7.35	-1.6	11.76
8	68	1.65	0.4	0.66
10	60	-6.35	2.4	-15.24
5	58	-8.35	-2.6	21.71
8	70	3.65	0.4	1.46
7	65	-1.35	-0.6	0.81

Tabla 1. Cálculo del Coeficiente de correlación de Pearson entre las variables talla y peso de 20 niños varones

4	54	-12.35	-3.6	44.46
11	83	16.65	3.4	56.61
7	64	-2.35	-0.6	1.41
7	66	-0.35	-0.6	0.21
6	61	-5.35	-1.6	8.56
8	66	-0.35	0.4	-0.14
5	57	-9.35	-2.6	24.31
11	81	14.65	3.4	49.81
5	59	-7.35	-2.6	19.11
9	71	4.65	1.4	6.51
6	62	-4.35	-1.6	6.96
10	75	8.65	2.4	20.76
				$\sum$ 290.8

$X\left(Media\ de\ \overline{X} = 66.35\right)$

$Y\left(Media\ de\ \overline{Y} = 7.6\right)$

$$Covarianza = \frac{\sum\left(\overline{X} - X\right)*\left(\overline{Y} - Y\right)}{n-1} = \frac{290.8}{19} = 15.30$$

$$r = \frac{covarianza}{S_x * S_y} = \frac{15.30}{8.087 * 2.137} = 0.885$$

S_x = Desviación típica x = 8.087

S_y = Desviación típica y = 2.137

Test de hipótesis de r

Tras realizar el cálculo del coeficiente de correlación de Pearson (r) debemos determinar si dicho coeficiente es estadísticamente diferente de cero. Para dicho calculo se aplica un test basado en la distribución de la t de student.

$$Error\ estandard\ de\ r = \sqrt{\frac{1-r^2}{n-2}}$$

Si el valor del r calculado (en el ejemplo previo r = 0.885) supera al valor del error estándar multiplicado por la t de Student con n-2 grados de libertad, diremos que el coeficiente de correlación es significativo.

El nivel de significación viene dado por la decisión que adoptemos al buscar el valor en la tabla de la t de Student.

En el ejemplo previo con 20 niños, los grados de libertad son 18 y el valor de la tabla de la t de student para una seguridad del 95% es de 2.10 y para un 99% de seguridad el valor es 2.88. (Tabla 2).

$$Error\ estandard\ de\ r = \sqrt{\frac{1-0.885^2}{20-2}} = 0.109$$

Como quiera que r = 0.885 > a 2.10 * 0.109 = 2.30 podemos asegurar que el coeficiente de correlación es significativo ($p<0.05$). Si aplicamos el valor obtenido en la tabla de la t de Student para una seguridad del 99% (t = 2.88) observamos que como r = 0.885 sigue siendo > 2.88 * 0.109 = 0.313 podemos a su vez asegurar que el coeficiente es significativo ($p<0.001$). Este proceso de razonamiento es válido tanto para muestras pequeñas como para muestras grandes. En esta última situación podemos comprobar en la tabla de la t de student que para una seguridad del 95% el valor es 1.96 y para una seguridad del 99% el valor es 2.58.

Intervalo de confianza del coeficiente de correlación

La distribución del coeficiente de correlación de Pearson no es normal pero no se puede transformar r para conseguir un valor z que sigue una distribución normal (transformación de Fisher) y calcular a partir del valor z el intervalo de confianza.

La transformación es:

$$z = 1/2L_n \frac{1+r}{1-r}$$

L_n representa el logaritmo neperiano en la base e

$$El\ error\ standard\ de\ z\ es = \frac{1}{\sqrt{n-3}}$$

donde n representa el tamaño muestral. El 95% intervalo de confianza de z se calcula de la siguiente forma:

$$z_1(limite\ \text{inf}\ erior) = z - 1.96/\sqrt{n-3}$$

$$z_2(limite\ \text{sup}\ erior) = z + 1.96/\sqrt{n-3}$$

Tras calcular los intervalos de confianza con el valor z debemos volver a realizar el proceso inverso para calcular los intervalos del coeficiente r

$$\frac{e^{2z1}-1}{e^{2z1}+1} \quad a \quad \frac{e^{2z2}-1}{e^{2z2}+1}$$

Utilizando el ejemplo de la Tabla 1, obtenemos r = 0.885

$$z = 1/2L_n \frac{1+0.885}{1-0.885} = 1.398$$

95% intervalo de confianza de z

$$z_1 = 1.398 - 1.96/\sqrt{20-3} = 0.922$$

$$z_2 = 1.398 + 1.96/\sqrt{20-3} = 1.873$$

Tras calcular los intervalos de confianza de z debemos proceder a hacer el cálculo inverso para obtener los intervalos de confianza de coeficiente de correlación r que era lo que buscábamos en un principio antes de la transformación logarítmica.

$$\frac{e^{2*0.922}-1}{e^{2*0.922}+1} \quad a \quad \frac{e^{2*1.873}-1}{e^{2*1.873}+1}$$

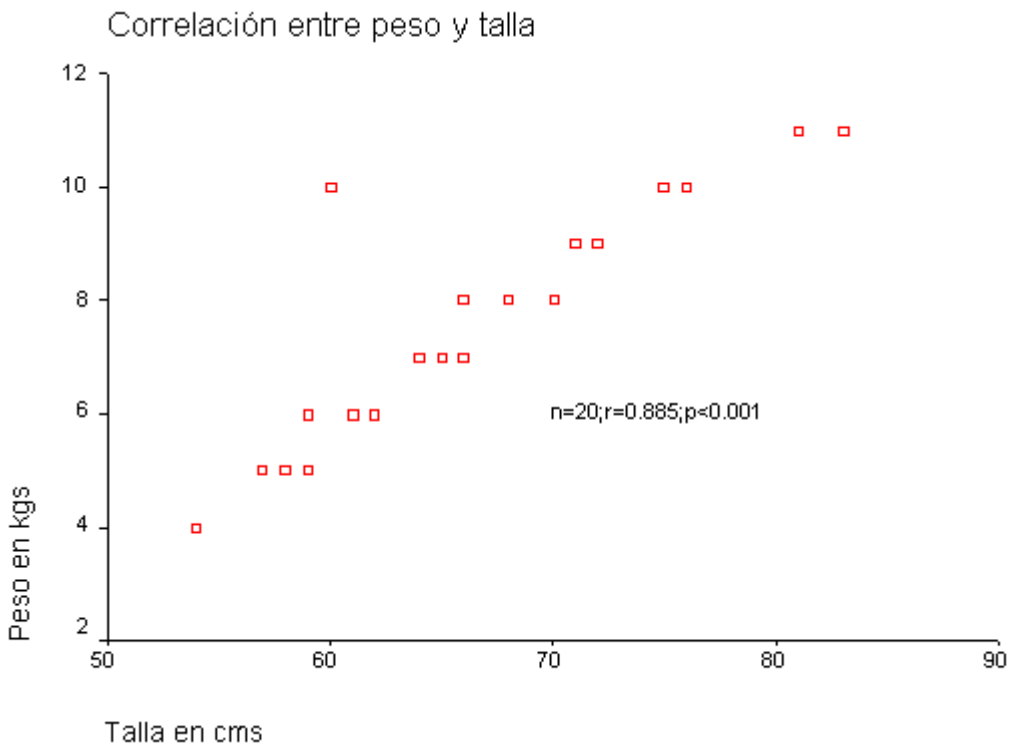
0.726 a 0.953 son los intervalos de confianza (95%) de r.

Presentación de la correlación

Se debe mostrar siempre que sea posible la gráfica que correlaciona las dos variables de estudio (Fig 1). El valor de r se debe mostrar con dos decimales junto con el valor de la p si el test de hipótesis se realizó para demostrar que r es estadísticamente diferente de cero. El número de observaciones debe a su vez estar indicado.

Figura 1. Correlación entre Peso y Talla

Figura 1. Correlación entre Peso y Talla



Interpretación de la correlación

El coeficiente de correlación como previamente se indicó oscila entre –1 y +1 encontrándose en medio el valor 0 que indica que no existe asociación lineal entre las dos variables a estudio. Un coeficiente de valor reducido no indica necesariamente que no exista correlación ya que las variables pueden presentar una relación no lineal como puede ser el peso del recién nacido y el tiempo de gestación. En este caso el r infraestima la asociación al medirse linealmente. Los métodos no paramétrico estarían mejor utilizados en este caso para mostrar si las variables tienden a elevarse conjuntamente o a moverse en direcciones diferentes.

La significancia estadística de un coeficiente debe tenerse en cuenta conjuntamente con la relevancia clínica del fenómeno que estudiamos ya que coeficientes de 0.5 a 0.7 tienden ya a ser significativos como muestras pequeñas. Es por ello muy útil calcular el intervalo de confianza del r ya que en muestras pequeñas tenderá a ser amplio.

La estimación del coeficiente de determinación (r^2) nos muestra el porcentaje de la variabilidad de los datos que se explica por la asociación entre las dos variables.

Como previamente se indicó la correlación elevada y estadísticamente significativa no tiene que asociarse a causalidad. Cuando objetivamos que dos variables están correlacionadas diversas razones pueden ser la causa de dicha correlación: a) pude que X influencie o cause Y, b) puede que influencie o cause X, c) X e Y pueden estar influenciadas por terceras variables que hace que se modifiquen ambas a la vez.

El coeficiente de correlación no debe utilizarse para comparar dos métodos que intentan medir el mismo evento, como por ejemplo dos instrumentos que miden la tensión arterial. El coeficiente de correlación mide el grado de asociación entre dos cantidades pero no mira el nivel de acuerdo o concordancia. Si los instrumentos de medida miden sistemáticamente cantidades diferentes uno del otro, la correlación puede ser 1 y su concordancia ser nula.

Coeficiente de correlación de los rangos de Spearman

Este coeficiente es una medida de asociación lineal que utiliza los rangos, números de orden, de cada grupo de sujetos y compara dichos rangos. Existen dos métodos para calcular el coeficiente de correlación de los rangos uno señalado por Spearman y otro por Kendall. El r de Spearman llamado también rho de Spearman es más fácil de calcular que el de Kendall. El coeficiente de correlación de Spearman es exactamente el mismo que el coeficiente de correlación de Pearson calculado sobre el rango de observaciones. En definitiva la correlación estimada entre X e Y se halla calculado el coeficiente de correlación de Pearson para el conjunto de rangos apareados. El coeficiente de correlación de Spearman es recomendable utilizarlo cuando los datos presentan valores externos ya que dichos valores afectan mucho el coeficiente de correlación de Pearson, o ante distribuciones no normales.

El cálculo del coeficiente viene dado por:

$$r_s = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)}$$

en donde $d_i = r_{xi} - r_{yi}$ es la diferencia entre los rangos de X e Y.

Los valores de los rangos se colocan según el orden numérico de los datos de la variable.

Ejemplo: Se realiza un estudio para determinar la asociación entre la concentración de nicotina en sangre de un individuo y el contenido en nicotina de un cigarrillo (los valores de los rangos están entre paréntesis) .

X	Y
Concentración de Nicotina en sangre (nmol/litro)	Contenido de Nicotina por cigarrillo (mg)
185.7 (2)	1.51 (8)
197.3 (5)	0.96 (3)
204.2 (8)	1.21 (6)
199.9 (7)	1.66 (10)
199.1 (6)	1.11 (4)
192.8 (6)	0.84 (2)
207.4 (9)	1.14 (5)
183.0 (1)	1.28 (7)
234.1 (10)	1.53 (9)
196.5 (4)	0.76 (1)

Si existiesen valores coincidentes se pondría el promedio de los rangos que hubiesen sido asignado si no hubiese coincidencias. Por ejemplo si en una de las variables X tenemos:

X (edad)	(Los rangos serían)
23	1.5
23	1.5
27	3.5
27	3.5
39	5
41	6
45	7
...	...

Para el cálculo del ejemplo anterior de nicotina obtendríamos el siguiente resultado:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6\left[(2-8)^2 + (5-3)^2 + (8-6)^2 + \ldots + (4-1)^2\right]}{10(10^2-1)} = 1 - \frac{6(120)}{10(99)} = 0.27$$

Si utilizamos la fórmula para calcular el coeficiente de correlación de Pearson de los rangos obtendríamos el mismo resultado

$$r_s = \frac{n\sum r_x r_y - \sum r_x \sum r_y}{\sqrt{\left[n\sum r_x^2 - \left(\sum r_x\right)^2\right]\left[n\sum r_y^2 - \left(\sum r_y\right)^2\right]}}$$

$$\sum r_x = \sum r_y = 55 \qquad \sum r_x^2 = \sum r_y^2 = 385$$

$$\sum r_x r_y = 2\,(8) + 5\,(3) + 8\,(6) + \ldots + 4\,(1) = 325$$

$$r_s = \frac{10(325) - 55(55)}{\sqrt{\left[10\,(385) - 55^2\right]\left[10(385) - 55^2\right]}} = 0.27$$

La interpretación del coeficiente r_s de Spearman es similar a la Pearson. Valores próximos a 1 indican una correlación fuerte y positiva. Valores próximos a –1 indican una correlación fuerte y negativa. Valores próximos a cero indican que no hay correlación lineal. Así mismo el r_s^2 tiene el mismo significado que el coeficiente de determinación de r^2. La distribución de r_s es similar a la r por tanto el cálculo de los intervalos de confianza de r_s se pueden realizar utilizando la misma metodología previamente explicada para el coeficiente de correlación de Pearson.

MEDIDAS DE ASOCIACIÓN ENTRE DOS VARIABLES

Las medidas de asociación tratan de estimar la magnitud con la que dos fenómenos se relacionan. Se emplean:

Covarianza: Es una medida de asociación entre dos variables y se calcula:

Muestral: $Sxy = \frac{\Sigma(xi - X)(yi - Y)}{n-1}$

Poblacional: $Sxy = \frac{\Sigma(xi - \mu x)(yi - \mu y)}{N}$

Coeficiente de correlación: Puede utilizarse para medir el grado de relación de dos variables siempre y cuando ambas sean cuantitativas.

Muestral: $rxy = \frac{Sxy}{Sx\,Sy}$

Poblacional: $Pxy = \frac{\tilde{O}xy}{\tilde{O}x\,\tilde{O}y}$

Coeficiente de regresión: Indica el número de unidades en que se modifica la variable dependiente "Y" por efecto del cambio de la variable independiente "X" o viceversa en una unidad de medida.

Clases de coeficiente de Regresión: El coeficiente de regresión puede ser: Positivo, Negativo y Nulo.

Es positivo cuando las variaciones de la variable independiente X son directamente proporcionales a las variaciones de la variable dependiente "Y".

Es negativo, cuando las variaciones de la variable independiente "X" son inversamente proporcionales a las variaciones de las variables dependientes "Y".

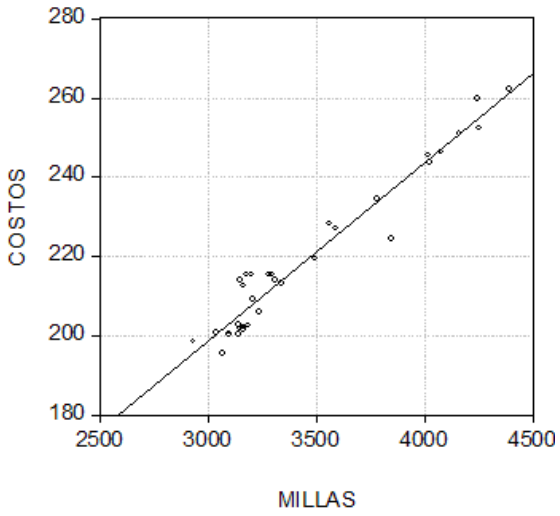
Es nulo o cero, cuando entre las variables dependientes "Y" e independientes "X" no existen relación alguna.

Se calcula:

$$y - Y = \frac{Sxy}{S^2 y}$$

$$Y - Yl = m\,(x - xi)$$

Gráfico de dispersión:



REGRESION LINEAL

Historia

La primera forma de regresión lineal documentada fue el método de los mínimos cuadrados que fue publicada por Legendre en 1805, Gauss publicó un trabajo en donde desarrollaba de manera más profunda el método de los mínimos cuadrados,[1] y en dónde se incluía una versión del teorema de Gauss-Márkov.

El término *regresión* se utilizó por primera vez en el estudio de variables antropométricas: al comparar la estatura de padres e hijos, donde resultó que los hijos cuyos padres tenían una estatura muy superior al valor medio, tendían a igualarse a éste, mientras que aquellos cuyos padres eran muy bajos tendían a reducir su diferencia respecto a la estatura media; es decir, "regresaban" al promedio.[2] La constatación empírica de esta propiedad se vio reforzada más tarde con la justificación teórica de ese fenómeno.

El término *lineal* se emplea para distinguirlo del resto de técnicas de regresión, que emplean modelos basados en cualquier clase de función matemática. Los modelos lineales son una explicación simplificada de la realidad, mucho más ágiles y con un soporte teórico mucho más extenso por parte de la matemática y la estadística.

Pero bien, como se ha dicho, se puede usar el término lineal para distinguir modelos basados en cualquier clase de aplicación.

El modelo de regresión lineal

El modelo lineal relaciona la variable dependiente Y con K variables explícitas $(k = 1,...K)$,

o cualquier transformación de éstas que generen un hiperplano de parámetros desconocidos:

(2) donde es la perturbación aleatoria que recoge todos aquellos factores de la realidad no controlables u observables y que por tanto se asocian con el azar, y es la que confiere al modelo su carácter estocástico. En el caso más sencillo, con una sola variable explícita, el hiperplano es una recta:

El problema de la regresión consiste en elegir unos valores determinados para los

parámetros desconocidos , de modo que la ecuación quede completamente especificada. Para ello se necesita un conjunto de observaciones. En una observación i-ésima *(i= 1,... I)* cualquiera, se registra el comportamiento simultáneo de la variable dependiente y las variables explícitas (las perturbaciones aleatorias se suponen no observables).

(4) Los valores escogidos como estimadores de los parámetros , son los coeficientes de regresión sin que se pueda garantizar que coincidan con parámetros reales del proceso generador. Por tanto, en

(5) Los valores son por su parte estimaciones o errores de la perturbación aleatoria.

Hipótesis del modelo de regresión lineal clásico

1. Esperanza matemática nula: Para cada valor de X la perturbación tomará distintos valores de forma aleatoria, pero no tomará sistemáticamente valores positivos o negativos, sino que se supone tomará algunos valores mayores que cero y otros menores que cero, de tal forma que su valor esperado sea cero.
2. Homocedasticidad: para todo *t*. Todos los términos de la perturbación tienen la

   misma varianza que es desconocida. La dispersión de cada en torno a su valor esperado es siempre la misma.
3. Incorrelación o independencia: para todo *t,s* con *t* distinto de *s*. Las covarianzas entre las distintas perturbaciones son nulas, lo que quiere decir que no están

correlacionadas. Esto implica que el valor de la perturbación para cualquier observación maestral no viene influenciado por los valores de las perturbaciones correspondientes a otras observaciones muéstrales.

4. Regresares estocásticos.
5. Independencia lineal. No existen relaciones lineales exactas entre los represores.
6. Suponemos que no existen errores de especificación en el modelo, ni errores de medida en las variables explicativas.
7. Normalidad de las perturbaciones: Supuestos del modelo de regresión lineal

Para poder crear un modelo de regresión lineal es necesario que se cumpla con los siguientes supuestos:[3]

1. Que la relación entre las variables sea lineal.
2. Que los errores en la medición de las variables explicativas sean independientes entre sí.
3. Que los errores tengan varianza constante. (Homocedasticidad)
4. Que los errores tengan una esperanza matemática igual a cero (los errores de una misma magnitud y distinto signo son equiprobables).
5. Que el error total sea la suma de todos los errores.

Tipos de modelos de regresión lineal

Existen diferentes tipos de regresión lineal que se clasifican de acuerdo a sus parámetros:

Regresión lineal simple

Sólo se maneja una variable independiente, por lo que sólo cuenta con dos parámetros. Son de la forma:[4]

(6) donde es el error asociado a la medición del valor y siguen los supuestos de modo que (media cero, varianza constante e igual a un y con Dado el

modelo de regresión simple anterior, si se calcula la esperanza (valor esperado) del valor *Y*, se obtiene:[5]

(7) Derivando respecto a y e igualando a cero, se obtiene:[5]

Bibliografía básica y complementaria:

Probabilidad y estadística de George Canavos

Estadística de Murray R. Spiegel