

# UDS

## ANTOLOGIA

*Estadística Descriptiva*

*Licenciatura en Contaduría Pública*

*3° Cuatrimestre*

*Mayo 2022*

---

## 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 – Tzitol 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.

---

## ESTADISTICA DESCRIPTIVA

---

**Objetivo de la materia:**

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

**Criterios de evaluación:**

No	Concepto	Porcentaje
<b>1</b>	<b>Actividades web escolar</b>	<b>50%</b>
a)	Primera actividad	25%
b)	Segunda Actividad	25%
<b>2</b>	<b>Examen</b>	<b>50%</b>
Total de Criterios de evaluación		100%

## ÍNDICE

### **Unidad I INTRODUCCION A LA ESTADISTICA APLICADA A LA PSICOLOGIA.**

1.1 La Estadística	10
1.2 Historia General	12
1.3 Conceptos básicos de la estadística.	13
1.4 Componentes de una distribución de frecuencia de clase:	15
1.5 La Media y la Moda.	18
1.6 El método científico y la Estadística	19
1.7 ¿Por qué la Estadística en el grado de Psicología?	20
1.8 Metodologías de investigación y Estadística	21
1.9 Estadística descriptiva	23
1.10 Estadística inferencial	25
1.11 Población y muestra	27
1.12. Parámetros y estadísticos	29

### **Unidad 2 ORGANIZACIÓN Y REPRESENTACIÓN GRÁFICA DE LOS DATOS**

2.1. La distribución de Frecuencias	31
2.1.1 Tipos de Frecuencias	34
2.2. La representación gráfica de una distribución de frecuencias	39
2.3. Propiedades de las distribuciones de frecuencias	44
2.4 Estadísticos de posición grupal	50
2.5 Variables categóricas: la moda	51
2.6. Variables ordinales: la mediana, el mínimo y el máximo, los cuantíales	52
2.7 Variables cuantitativas: la media y sus alternativas robustas	60
2.8. Estadísticos de dispersión	70
2.9. Variables categóricas: la razón de variación y el índice de variación cualitativa	71
2.10. Variables ordinales: el rango y el rango intercuartil	73
2.11. Variables cuantitativas: la varianza, la desviación típica y el coeficiente de variación.	75

### **Unidad 3 ESTADÍSTICOS DE FORMA DE LA DISTRIBUCIÓN**

3.1 Asimetría	77
---------------	----



3.2. Apuntamiento	78
3.3. Estadísticos de posición individual	80
3.4. Los porcentajes acumulados	80
3.5. Las puntuaciones típicas	83
3.6. Medidas de Posición	84
3.7. Las escalas derivadas	89
3.8. Organización y representación gráfica de datos multivariados	90
3.9. La distribución conjunta multivariada	92
3.10. La tabla de contingencia	93
3.11. Cálculo para una distribución de Frecuencia	95
3.12. Representaciones gráficas	101
3.13. El caso de dos variables categóricas	102
3.14. El caso de dos variables cuantitativas	104
3.15. El caso de una variable categórica y una variable cuantitativa	106

#### **Unidad 4 ESTADÍSTICOS DE ASOCIACIÓN ENTRE VARIABLES**

4.1. Concepto de asociación entre variables	107
4.2. Midiendo la asociación El caso de dos variables categóricas	109
4.3. Midiendo la asociación El caso de una variable categórica y una cuantitativa	111
4.4. Midiendo la asociación El caso de dos variables cuantitativas	113
4.5. El modelo de regresión lineal	127
4.5.1 Conceptos básicos sobre el análisis de regresión lineal	127
4.6 Hipótesis del modelo de regresión lineal clásico	129
4.7 Tipos de modelos de regresión lineal:	130
4.8 Fórmulas de regresión no lineales	131
4.9. Estimadores	132
4.10 Regresión Lineal Por Covarianza	133
4.11. Regresión Lineal Por Mínimos Cuadrados	136

## UNIDAD I

### INTRODUCCION A LA ESTADISTICA APLICADA A LA PSICOLOGIA.

#### I.1.- La Estadística

Describe, analiza y representa un grupo de datos utilizando métodos numéricos y gráficos que resumen y presentan la información contenida en ellos. Se puede definir 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:

- Selección de caracteres dignos de ser estudiados.

- Mediante encuesta o medición, obtención del valor de cada individuo en los caracteres seleccionados.
- Elaboración de tablas de frecuencias, mediante la adecuada clasificación de los individuos dentro de cada carácter.
- Representación gráfica de los resultados (elaboración de gráficos estadísticos).
- 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 muestrales, 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 muestrales 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 el 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.

### **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.

## **1.2 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 que incurren en la Estadística

**Universo:** En estadística es el nombre específico 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 el término universo y población son sinónimos. 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 de 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 =  $\mu$  (miu), La desviación Típica =  $\sigma$ , (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  $f_i$ , 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 ( $f_i$ ) 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.

#### **1.4 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  $I_c$ . 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 valor 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  $f_i$ . 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  $f_i$  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  $f_r$ ; si cada  $f_r$  se multiplica por 100 se obtiene la frecuencia relativa porcentual ( $f_r \%$ ).

#### 7.-Frecuencias acumuladas

Las frecuencias acumuladas de una distribución de frecuencias son aquellas que se obtienen de las sumas sucesivas de las  $f_i$  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  $f_a$ . Las frecuencias acumuladas pueden ser menor que ( $f_a < que$ ) y frecuencias acumuladas mayor que ( $f_a > que$ ).

#### 8.- Frecuencia acumulada relativa

La frecuencia acumulada relativa es aquella que resulta de dividir cada una de las  $f_a$  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  $f_{ar}$ . Si las  $f_{ar}$  se multiplican por 100 se obtienen las frecuencias acumuladas relativas porcentuales y las mismas se designan así:  $f_{ar} \%$ .

## 1.5 La Media y la Moda

**La mediana :** La mediana ( $M_d$ ) 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 está 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  $M_o$ .

**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.

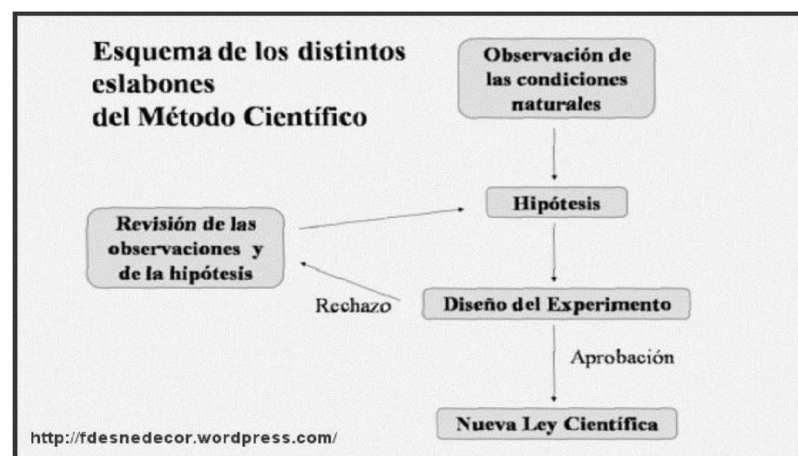
**Varianza:** Es otra de las variaciones absolutas y la misma se define como el cuadrado de la desviación típica; viene expresada con las mismas letras de la desviación típica pero elevada al cuadrado, así  $S^2$  y  $s^2$ . Las fórmulas para calcular la varianza son las mismas utilizadas por la desviación típica, exceptuando las respectivas raíces, las cuales desaparecen al estar elevados el primer miembro al cuadrado

## 1.6 El Método Científico y la Estadística

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.

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

### **1.7.- ¿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.

### **I.8 Metodología de investigación y 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 es 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.

### **1.9.- Estadística descriptiva**

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).}

## **1.10 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.

### 1.11 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 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.

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.

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.

- I. **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 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.

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 puntuación 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**

### **ORGANIZACIÓN Y REPRESENTACIÓN GRÁFICA DE LOS DATOS**

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.

La recogida de datos asociada a la realización de un estudio suele representar la obtención de un conjunto más o menos numeroso de datos, ahora bien, la interpretación de los mismos a simple vista suele resultar poco inteligible en la mayoría de los casos. La estadística descriptiva nos ofrece herramientas para organizar y resumir los datos que hayamos recogido, de modo que pueda ser extraída e interpretada la información contenida en los mismos que sea de nuestro interés

#### **2.1.-La distribución de Frecuencias.**

La distribución de frecuencias constituye una de las formas más intuitiva de organizar los datos de una variable: se basa en el conteo del número de entidades (casos, sujetos) que tienen cada

uno de los valores con que la variable se ha manifestado (modalidades). Es una técnica estadística básica pero, a la vez, muy informativa y relevante en la práctica del análisis de datos.

- El número de veces que aparece una determinada modalidad de una variable (X) es lo que se conoce como la frecuencia absoluta ( $n_i$ ) de esa modalidad o valor.

- Derivadas de las frecuencias absolutas se pueden obtener las frecuencias relativas o proporciones ( $p_i$ ):

$$p_i = n_i / n$$

- Las frecuencias relativas también pueden expresarse como porcentajes ( $\%i$ ) con tan sólo multiplicar su valor por 100:  $\%i = p_i \cdot 100$

Ejemplo para la variable categórica “Estado civil” (X), habiendo sido recogidos datos para una muestra de 50 personas de la ciudad de Castellón ( $n = 50$ ): X: {0, 0, 1, 2, 2, 0, 1, 3, 2, 0, 1, 0, 1, 2, 0, 2, 1, 1, 0, 1, 0...} Codificación: [0: soltero/a; 1: casado/a; 2: separado/a o divorciado/a; 3: viudo/a]

$X_i$	Frec. absoluta ( $n_i$ )	Frec. relativa ( $p_i$ )	Porcentaje ( $\%i$ )
0	15	0,3	30
1	20	0,4	40
2	11	0,22	22
3	4	0,08	8
	50	1,00	100

En el caso de las variables cuantitativas y las cuasi-cuantitativas , además de lo anterior, se puede obtener también la siguiente información para cada una de las modalidades: - las frecuencias absolutas acumuladas ( $n_a$ ), - las frecuencias relativas acumuladas ( $p_a$ ), - y los porcentajes acumulados ( $\%a$ ).

Ejemplo para la variable cuantitativa “Nº de hijos/as” (X), con datos para una muestra de 20



familias del barrio de Velluters de la ciudad de Valencia: X: {2, 1, 0, 3, 2, 2, 3, 1, 1, 0, 1, 2, 1, 2, 0, 2, 4, 2, 3, 1}

$X_i$	Frec. absoluta ( $n_i$ )	Frec. relativa ( $p_i$ )	Porcentaje ( $\%_i$ )	Frec. absoluta acumulada ( $n_a$ )	Frec. relativa acumulada ( $p_a$ )	Porcentaje acumulado ( $\%_a$ )
0	3	0,15	15	3	0,15	15
1	6	0,30	30	9	0,45	45
2	7	0,35	35	16	0,80	80
3	3	0,15	15	19	0,95	95
4	1	0,05	5	20	1,00	100
	20	1	100			

• Algunas anotaciones acerca de las distribuciones de frecuencias:

(1) Es costumbre situar los valores correspondientes a la columna de las modalidades de la variable X en sentido creciente de arriba hacia abajo

(2) Para los valores de la variable que no haya ningún caso es costumbre no dedicar ninguna fila en la tabla de la distribución de frecuencias a fin de que ésta ocupe menos espacio. (

3) Las frecuencias relativas o proporciones se caracterizan por: tomar valores entre 0 y 1; ser la suma de todas ellas igual a la unidad.

Ejercicio 1: Las siguientes datos son de los estudiantes de una clase en la que un observador, durante el tiempo que ha durado una sesión de clase de 2 horas, ha anotado el número de veces que ha participado cada uno de los estudiantes dirigiéndose a todo el grupo en voz alta.

2 2 3 0 3 1 8 0 3 9 1 1 0 4 0 2 9 5 0 1 9 8 Obtener la distribución de frecuencias completa y contestar las siguientes preguntas (utilizar dos decimales en los cálculos y a la hora de presentar resultados): a) ¿Qué proporción de estudiantes participaron en menos de 2 ocasiones en la sesión de clase?, ¿cuántos estudiantes son? b) ¿Qué porcentaje de estudiantes participaron 5 veces? ¿Cuántos son? c) ¿Qué proporción de estudiantes participaron 4 veces o menos? ¿Cuántos son? d) ¿Qué porcentaje de estudiantes participaron

más de 4 veces? ¿Cuántos son? e) ¿Qué proporción de estudiantes participaron al menos una vez? ¿Cuántos son? f) ¿Qué porcentaje de estudiantes participaron entre 2 y 5 veces, ambas inclusive? ¿Cuántos son? g) ¿Qué porcentaje de estudiantes participaron 8 ó 9 veces? ¿Cuántos son? h) ¿Qué proporción de estudiantes participaron 4 veces o más? ¿Cuántos son?

(4) En el caso de las variables cuantitativas continuas dado que, si la medida de la variable se realiza con cierta precisión, se puede obtener un número amplio de datos diferentes, es práctica habitual que en la columna de las modalidades ( $X_i$ ) los valores representen a intervalos de valores de igual amplitud.

Ejemplo de la distribución de frecuencias elaborada a partir de los datos de la variable “Peso (kg)” ( $X$ ) de los 420 jugadores inscritos en la liga profesional masculina de balonmano en la temporada 2008/09:  $X: \{82,5; 91,1; 90,6; 83,8; 92,1; 88,3; 93,6; 101,4; 91,7; 80,2; \dots\}$

## 2.1.1 Tipos de Frecuencia.

### a) 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  $N$ .

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

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

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

### b) 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.

### c) 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  $F_i$ .

### d) 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	HHH I	6	9	0.194	0.290
30	HHH II	7	16	0.226	0.516
31	HHH 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	I
		31		I	

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 queremos 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

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

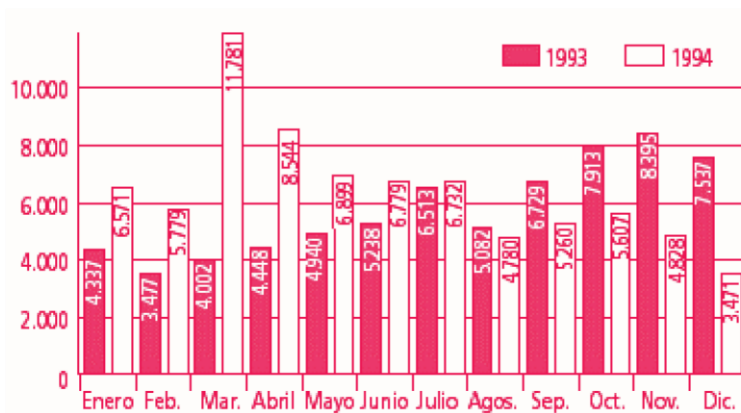
ejes cartesianos.

- 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

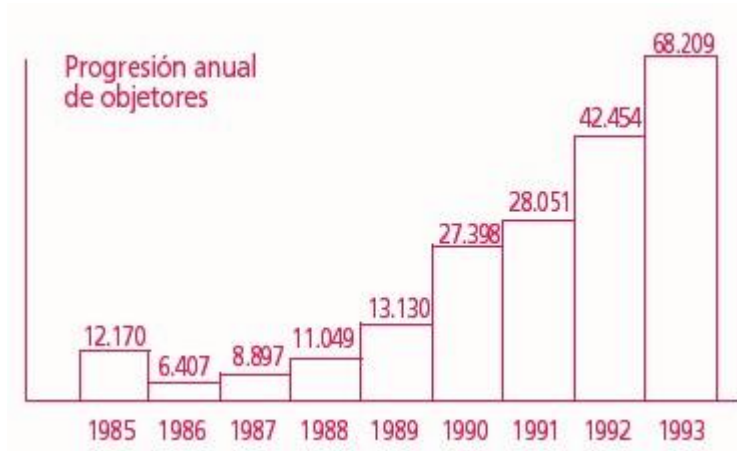
Los diagramas de barras se usan para representar gráficamente series estadísticas de valores en un sistema de ejes cartesianos, de manera que en las abscisas se indica el valor de la variable estadística y en las ordenadas se señala su frecuencia absoluta.

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.



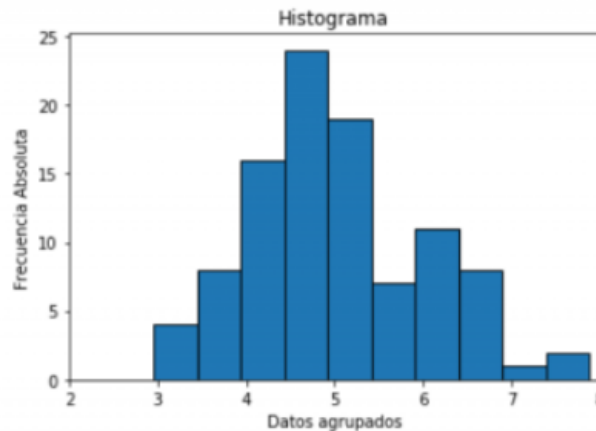


**Diagrama de barras.**



**Histograma.**

Un histograma es la representación gráfica en forma de barras, que simboliza la distribución de un conjunto de datos. Sirven para obtener una "primera vista" general, o panorama, de la distribución de la población, o de la muestra, respecto a una característica, cuantitativa y continua. En un histograma el eje de las  $x$  (o abscisas) consiste del rango en el cual se encuentran los datos. Ahora, las bases de los rectángulos consisten de los intervalos en los cuales agrupamos dichos datos.



## 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:

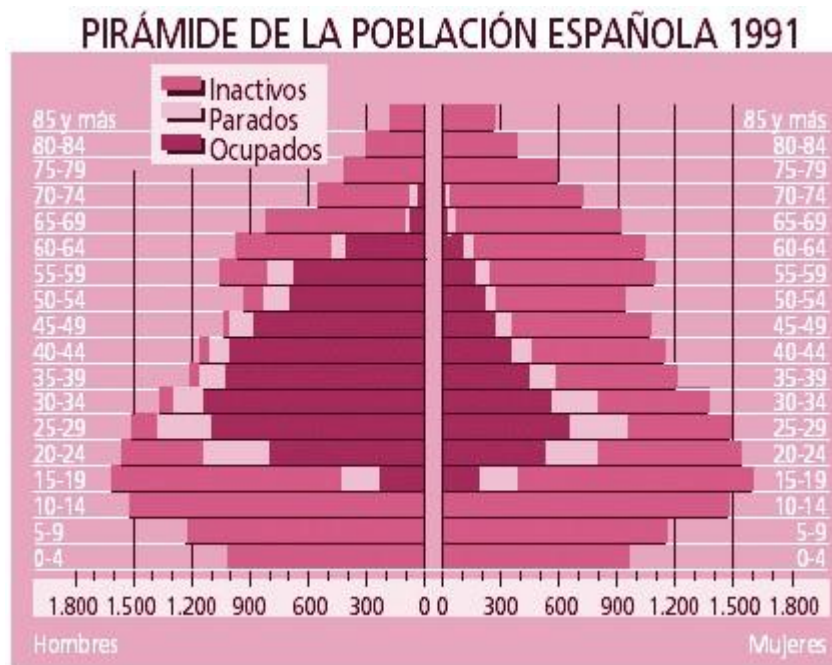
- Los pictogramas, que muestran diagramas figurativos con figuras o motivos que aluden a la distribución estadística analizada (por ejemplo, una imagen antropomórfica para indicar tamaños, alturas u otros).
- Los cartogramas, basados en mapas geográficos que utilizan distintas tramas,

colores o intensidades para remarcar las diferencias entre los datos.

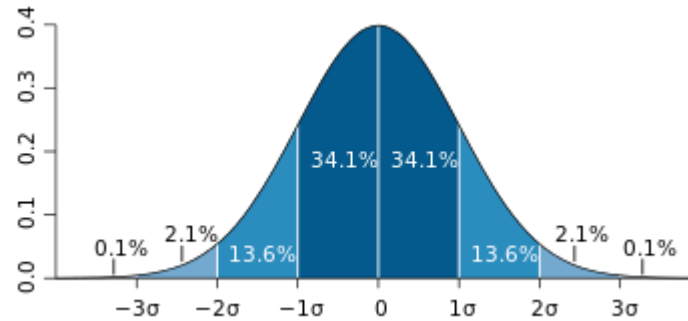
### Pirámide de población

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 PsicologíaOnline: 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

## 2.3 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.

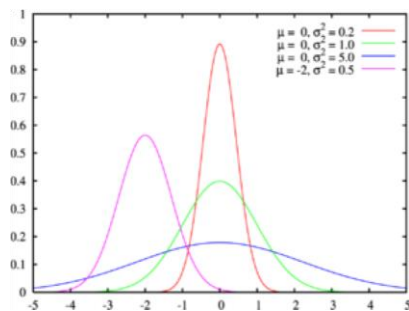
### Variabilidad:

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



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.<sup>567</sup> 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  $\mu$  y  $\sigma$  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,  $\lambda$ ; 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.

Según Yule 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, cuantos 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 maestro Habitualmente se agrupan los parámetros en las siguientes categorías:

#### **Medidas de posición.**

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.

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, varianzas, 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:

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.

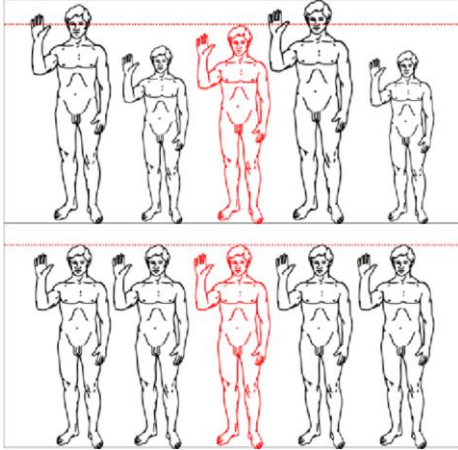
Además, y con propósitos más específicos, existen otros parámetros de uso en situaciones muy concretas, como son las proporciones, los números índice, las tasas y el coeficiente de Gini.

Medidas de tendencia central o centralización

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.<sup>12</sup> Sus propiedades son:<sup>13</sup>

- 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  $\bar{x}$  es mínimo cuando  $\bar{x}$ . 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  $\bar{x}$  es la media aritmética de los  $x_i$ , para  $i = 1, \dots, n$  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.<sup>14</sup> 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 €.

## 2.4 Estadísticos de Posición Grupal

Para una variable discreta, se define el percentil de orden  $k$ , como la observación,  $P_k$ , que deja por debajo de sí el  $k\%$  de la muestra. Esta definición es semejante a la de la mediana, pues como consecuencia de la definición, es evidente que  $Med = P_{50}$ .

Por su propia naturaleza, el percentil puede estar situado en cualquier lugar de la distribución, por

lo que no se puede considerar como una medida de tendencia central, sino más bien de *posición*.

Los cuartiles,  $Q_i$ , son un caso particular de los percentiles. Hay 3, y se definen como:

**Primer Cuartil:**  $Q_1 = P_{25}$ .

**Segundo Cuartil:**  $Q_2 = P_{50}$ . Es equivalente a la Mediana.

**Tercer Cuartil:**  $Q_3 = P_{75}$ .

## 2.5 Variables Categóricas: La Moda

Artículo principal: Moda (estadística)

La moda es el dato más repetido, el valor de la variable con mayor frecuencia absoluta.<sup>15</sup> 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".<sup>16</sup>

**Inconvenientes:**

- Su valor es independiente de la mayor parte de los datos, lo que la hace muy sensible a variaciones muestrales. 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).

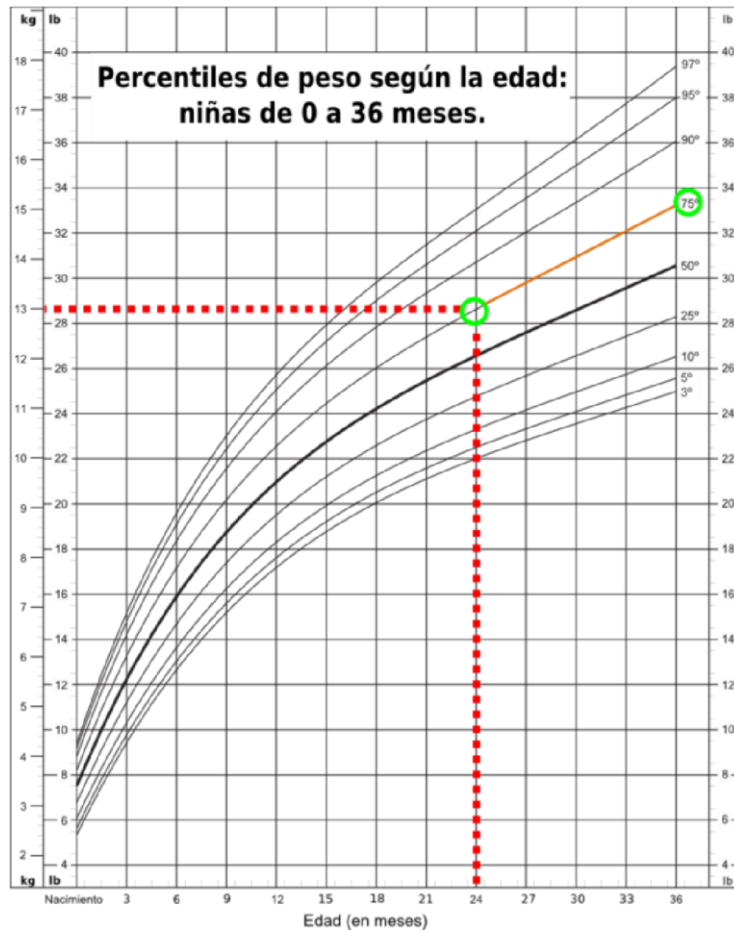
**2.6- Variables ordinales: la Mediana, el mínimo, y el máximo los cuantíales.**

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.<sup>17</sup> 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:<sup>18</sup>

- 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

#### 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.<sup>19</sup> 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.<sup>20</sup>

### 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.<sup>21</sup> 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:<sup>22</sup>

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)*

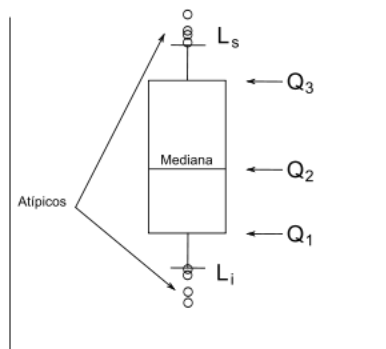


Diagrama de caja que muestra la dispersión gráficamente, usando los cuartiles como referencia.

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.

## 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 1º y 3º, respectivamente, y  $R_s$  representa la mitad del recorrido o rango intercuartílico, también conocido como recorrido semiintercuartílico.

### Desviaciones medias

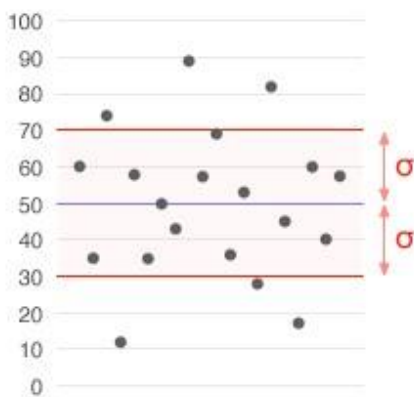
## Artículo principal: Desviación media

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  $\bar{x}$  entonces  $\bar{|x_i - c|}$ . De este modo se definen la desviación media respecto de la media o la desviación media respecto de la mediana ( $c = \text{mediana}$ ), cuya interpretación es sencilla en virtud del significado de la media aritmética.<sup>23</sup> 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

### Artículo principal: Varianza



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:<sup>25</sup> o sea, la media de los cuadrados de las desviaciones respecto de la media. La desviación típica,  $\sigma$ , se define como la raíz cuadrada de la varianza, esto es,  $\sigma = \sqrt{s^2}$ . Para variables agrupadas en intervalos, se usan las marcas de clase (un valor apropiado del interior de cada intervalo) en estos cálculos.

Propiedades:

- 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  $\bar{x} \pm k\sigma$  se encuentran, al menos, el  $\frac{100}{k^2}$  % de las observaciones (véase Desigualdad de Tchebyshev).

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  $\bar{x} \pm 2\sigma$  están, al menos, el 75% de los datos.
- En el intervalo  $\bar{x} \pm 3\sigma$  están, al menos, el 89% de los datos.

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

## 2.7- Variables Cuantitativas: la media y sus alternativas robustas.

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.

### 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.

### Coeficiente de variación de Pearson

Artículo principal:

**Coeficiente de variación:** Se define como, donde  $\sigma$  es la desviación típica y  $\bar{x}$  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:

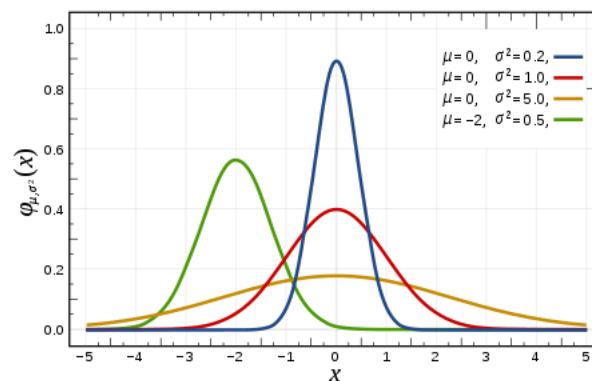
Se define como el cociente entre los valores extremos de la distribución de datos, esto es, dada una distribución de datos estadísticos  $x_1, x_2, \dots, x_n$ , su coeficiente de apertura,  $C_A$ . Se usa para comparar salarios de empresas.

### 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<sup>29</sup> 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.

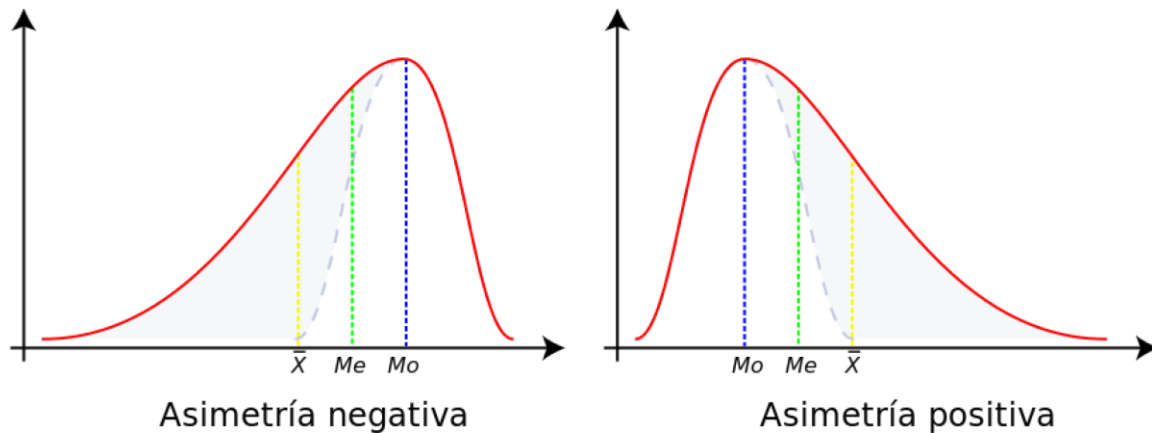
Las medidas de forma caracterizan la forma de la gráfica de una distribución de datos estadísticos. La mayoría de estos parámetros tiene un valor que suele compararse con la campana de Gauss, esto es, la gráfica de la distribución normal, una de las que con más frecuencia se ajusta a fenómenos reales.

## 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:



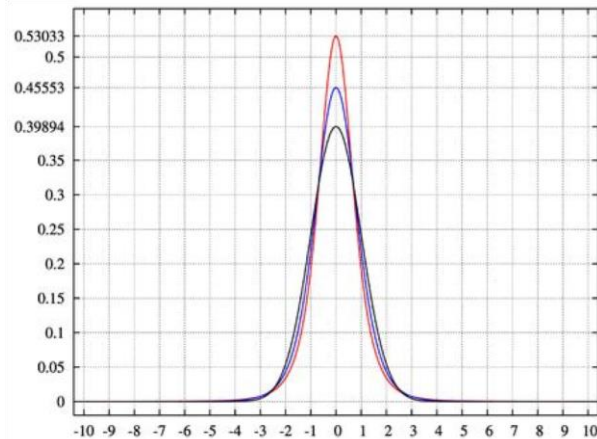
Ello puede demostrarse fácilmente si se tiene en cuenta la atracción que la media aritmética siente por los valores extremos, que ya se ha comentado más arriba y las definiciones de mediana (justo en el centro de la distribución, tomando el eje de abscisas como referencia) y moda (valor que presenta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



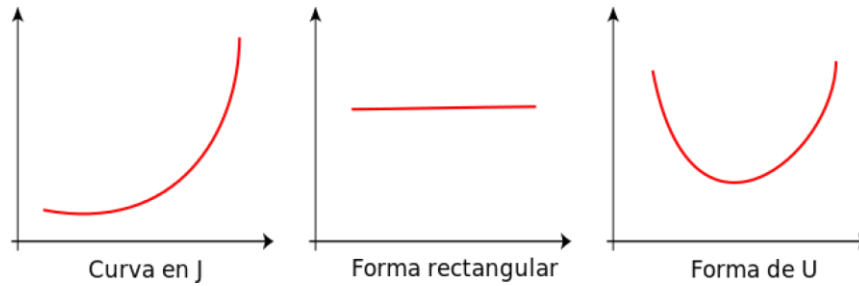
Tres distribuciones con distintos grados de apuntamiento.

Artículo principal: [Curtosis](#)

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.<sup>30</sup>

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

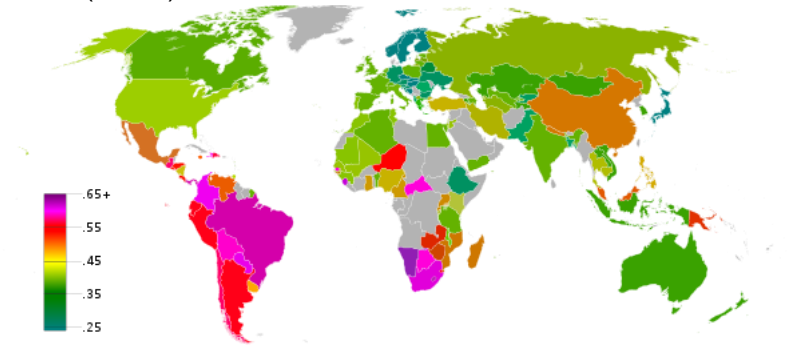
Un número índice es una medida estadística que permite estudiar las fluctuaciones o variaciones de una magnitud o de más de una en relación al tiempo o al espacio. Los



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<sup>31</sup>

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.<sup>31</sup> 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.

## 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  $x_1, x_2, \dots, x_n$ , 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.<sup>33</sup> 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.<sup>35</sup>

## 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  $i = 1, \dots, n$ , 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 ( $\bar{x}$ ) o la desviación típica de los pesos ( $\sigma_y$ ). 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  $Me_{y/x=x_k}$ .

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

$\bar{x}$  y  $\bar{y}$  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

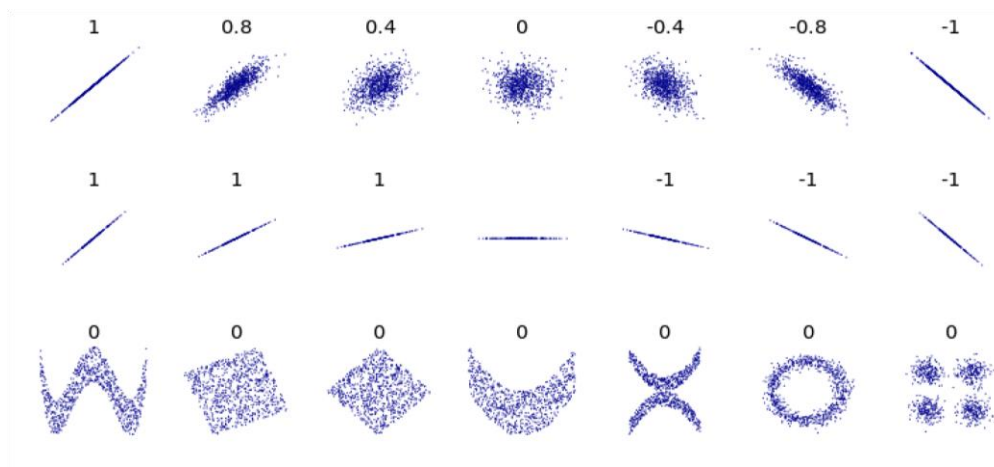
La covarianza o varianza conjunta de una distribución bidimensional se define como:

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.<sup>38</sup> 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 ocurriría 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  $\sigma_{xy}$  es la covarianza y  $\sigma_x$  y  $\sigma_y$ , 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  $r$ .<sup>39</sup>

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 maestra.

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.<sup>40</sup>

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 maestra, la varianza maestra 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  $\theta$  es un conjunto de parámetros de la distribución, se definiría la media maestra  $n$ -ésimo como:

En el caso concreto de la varianza maestra, 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.

### 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.

## 2.8 Estadísticos de Dispersión

La dispersión estadística es el grado en que una distribución de datos se aleja o se acerca, en valor absoluto a la media aritmética, como estadístico de posición central.

**Importancia de la dispersión estadística:** Cuando queremos llevar a cabo un análisis

descriptivo, en primer lugar calculamos las medidas resumen de posición. Las más habituales son la media, mediana, moda o cuartiles, deciles, quintiles o percentiles. Además, necesitamos saber la dispersión estadística, Las medidas de dispersión: aportan información muy relevante. Si la dispersión es muy elevada afecta a la media y esta deja de ser representativa del grupo como medida resumen. Por eso, normalmente ambos datos van juntos.

Medidas de dispersión Estadística: Existen medidas de dispersión que permiten su medición-

Rango; No es más que la diferencia entre el menor y el mayor del valor de distribución,

Desviación Media: sería el equivalente al promedio de las diferentes desviaciones de cada dato respecto a la media.

Variación y desviación Típica: Son las medidas de dispersión conocidas, suelen utilizarse la segunda que es más fácil de calcular ( raíz de la varianza ) y de interpretar, se expresan en valores absolutos.

Coeficiente de variación: en este caso se calcula la desviación típica y la media y sirve para comparar ya que se expresa en valores relativos (%)

## **2.9 Variables categóricas: la razón de la variación y el índice de variación cualitativa**

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. 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

categorica.

Una variable categorica que puede tomar dos valores se denomina una variable binaria o una variable dicotomica; un caso especial importante es la variable de Bernoulli. Las variables categoricas con más de dos valores posibles se denominan variables politomicas; las variables categoricas a menudo se supone que son politomicas menos que se especifique lo contrario.

La discretización es el tratamiento de los datos continuos como si fuera categorica. La dicotomización es el tratamiento de los datos continuos o variables politomicos 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 categoricas 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

### 2.10- Variables ordinales: el rango y el rango intercuartil

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:

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

**2.1 I - Variables cuantitativas; la Varianza, la desviación típica y el coeficiente de variación.**

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

También es la media aritmética de los cuadrados de las desviaciones, por lo que se calcula elevando al cuadrado cada desviación, sumando esos cuadrados y dividiendo entre n; en nuestro ejemplo resulta 58 la suma de cuadrados de las desviaciones, con lo que la varianza es  $58/13 = 4.46$  ; el cuadrado es matemáticamente mucho más manejable que el valor absoluto, lo que hace de la varianza la reina de los parámetros de dispersión desde un punto de vista teórico. Sin embargo, el hecho de que carezca de interpretación intuitiva y que sus unidades sean cuadradas (¿hermanos cuadrados?) hace que en la práctica se use mucho más su raíz cuadrada,

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:

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

La fórmula que simplifica el cálculo de la varianza es:

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

**Desviación standard** o **Desviación típica**, con mucho la más usada de las medidas de dispersión, y que en nuestro ejemplo valdría 2.11, con lo que el informe más habitual para nuestros datos daría una media de 2.00 y la desviación típica de 2.11 como parámetros más informativos. Por motivos difíciles de explicar aquí, relacionados con cuestiones de inferencia estadística, es más recomendable usar el denominador  $n-1$  en lugar del  $n$  al calcular la varianza y la desviación típica de una muestra, quedándose el  $n$  para el caso en que se conoce toda la población; en nuestro ejemplo, pues, sería mejor calcular como varianza  $58/12 = 4.83$  y como desviación standard su raíz cuadrada 2.20 (estos últimos serían la varianza muestral o quasivarianza y la desviación típica muestral y serían los utilizados en la práctica, aunque la definición teórica sea con denominador  $n$  por ser la varianza una "media"). A efectos comparativos entre distintas muestras e incluso entre distintas variables.

El **Coefficiente de variación**, que es el cociente, a menudo expresado en tanto por ciento, entre la desviación típica y la media de una distribución. Es una especie de desviación típica "relativa", y en nuestro ejemplo valdría  $2.20/2.00 = 1.100055$  ó bien 110.05% (nótese que no es un verdadero porcentaje, porque puede valer más del 100%); este resultado indicaría mucha dispersión en los datos del ejemplo en relación con la media.

## Unidad 3

### ESTADÍSTICOS DE FORMA DE LA DISTRIBUCIÓN

#### 3.1 Asimetría

Las medidas de asimetría son indicadores que permiten establecer el *grado de simetría* (o asimetría) que presenta una distribución de probabilidad de una variable aleatoria sin tener que hacer su representación gráfica. Como eje de simetría consideramos una recta paralela al eje de ordenadas que pasa por la media de la distribución. Si una distribución es simétrica, existe el mismo número de valores a la derecha que a la izquierda de la media, por tanto, el mismo número de desviaciones con signo positivo que con signo negativo. Decimos que hay asimetría positiva (o a la derecha) si la "cola" a la derecha de la media es más larga que la de la izquierda, es decir, si hay valores más separados de la media a la derecha. Diremos que hay asimetría negativa (o a la izquierda) si la "cola" a la izquierda de la media es más larga que la de la derecha, es decir, si hay valores más separados de la media a la izquierda.

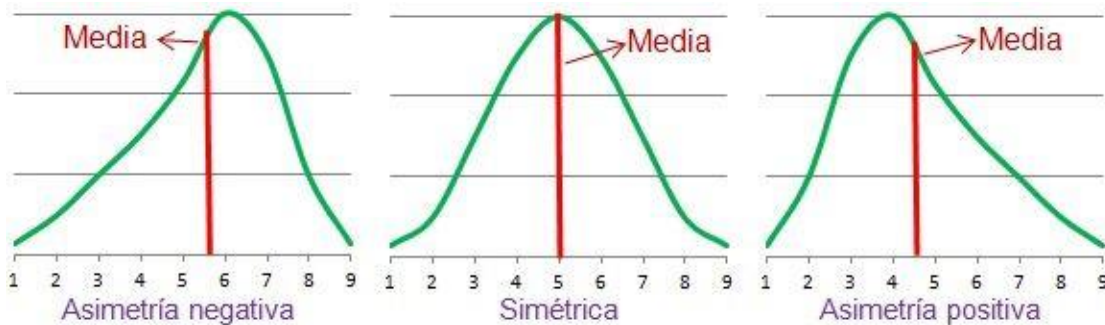
La asimetría resulta útil en muchos campos. Muchos modelos simplistas asumen una distribución normal, esto es, simétrica en torno a la media. La distribución normal tiene una asimetría cero. Cuando el tamaño de la muestra aumenta cualquier población tiende a volverse simétrica. Una asimetría positiva implica que hay más valores distintos a la derecha de la media.

Las medidas de asimetría, sobre todo el coeficiente de asimetría de Fisher, junto con las medidas de apuntamiento o curtosis se utilizan para contrastar si se puede aceptar que una distribución estadística sigue la distribución normal. Esto es necesario para realizar numerosos contrastes estadísticos en la teoría de inferencia estadística.

Existen tres tipos de curva de distribución según su asimetría:

- **Asimetría negativa:** la cola de la distribución se alarga para valores inferiores a la media.

- **Simétrica:** hay el mismo número de elementos a izquierda y derecha de la media. En este caso, coinciden la media, la mediana y la moda. La distribución se adapta a la forma de la campana de Gauss, o distribución normal.
- **Asimetría Positiva:** La cola de la Distribución se alarga ( a la derecha) para valores superiores a la media.



### 3.2 Apuntamiento o curtosis

La curtosis es una medida estadística que determina el grado de concentración que presentan los valores de una variable alrededor de la zona central de la distribución de frecuencias. También es conocida como medida de apuntamiento.

Cuando medimos una variable aleatoria, por lo general, los resultados que tienen una mayor frecuencia son los que se sitúan en torno a la media de la distribución. Imaginemos la altura de los alumnos de una clase. Si la altura media de la clase es 1,72 cm, lo más normal es que las alturas del resto de los alumnos estén en torno a este valor (con cierto grado de variabilidad, pero sin ser esta demasiado grande). Si esto sucede, se considera que la distribución de la variable aleatoria se distribuye con normalidad. Pero dada la infinidad de variables que se pueden medir, esto no siempre sucede así.

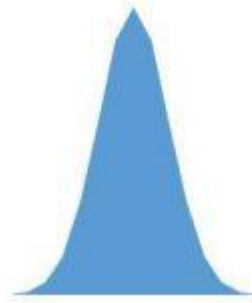
Existen algunas variables que presentan un mayor grado de concentración (menor dispersión) de los valores en torno a su media y otras, por el contrario, presentan un menor grado de concentración (mayor dispersión) de sus valores en torno a su valor central. Por tanto, la

curtosis nos informa de lo apuntada (mayor concentración) o lo achatada (menor concentración) que es una distribución

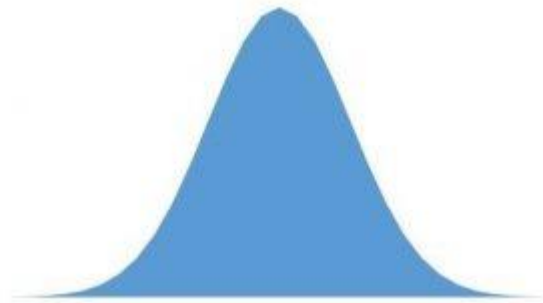
### Tipos de curtosis:

Dependiendo del grado de curtosis, tenemos tres tipos de distribuciones:

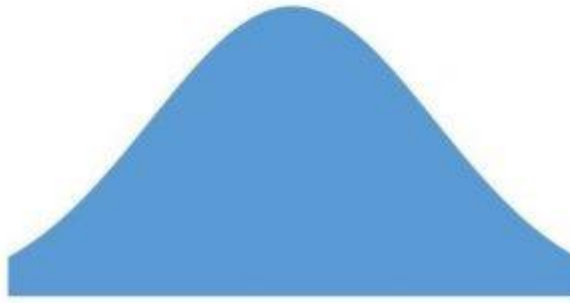
1. **Leptocúrtica:** Existe una gran concentración de los valores en torno a su media ( $g_2 > 3$ )



2. **Mesocúrtica:** Existe una concentración normal de los valores en torno a su media ( $g_2 = 3$ ).



3. **Platicúrtica:** Existe una baja concentración de los valores en torno a su media ( $g_2 < 3$ ).



### 3.3 Estadísticos de posición Individual

Hasta ahora se ha abordado la descripción de los datos de conjuntos de observaciones (variables); en este capítulo, por el contrario, se centra la atención en la descripción de observaciones particulares, en concreto, en estadísticos que nos van a ofrecer información sobre un valor concreto en relación a la posición que ocupa dentro de un conjunto de valores observados.

En cuanto que se trata de estadísticos que ofrecen información sobre la posición de un valor respecto a un grupo de referencia, nos van a permitir establecer una interpretación relativa de los valores observados.

Ejemplo: Nos dice un amigo que les han pasado a todos los trabajadores de su empresa un test de aptitudes verbales y que ha obtenido una puntuación igual a 134; a continuación, sin más detalles, nos pregunta que si esa puntuación significa que es bueno o malo en aptitudes verbales. ¿Qué información adicional podría sernos de utilidad a fin de poder ofrecerle algún tipo de valoración de esa puntuación?

### 3.4 Los porcentajes Acumulados.

El porcentaje acumulado (%a) de un valor concreto de una variable es el porcentaje de casos que obtienen un valor inferior o igual a ese en la variable en cuestión, información que puede obtenerse directamente a partir de la distribución de frecuencias correspondiente a esa



variable.

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.

A estos porcentajes acumulados se les suele denominar más habitualmente, aunque de un modo equívoco, como percentiles (término ya utilizado en el contexto de los estadísticos de posición grupal). Así, es más común escuchar expresiones tales como “Va a ser muy alto, está en el percentil 80 de altura” o “No ha llegado al percentil 5 en la prueba de coordinación óculo-manual”.

Ejemplo de obtención de porcentajes acumulados o percentiles: sea la siguiente distribución de frecuencias de las puntuaciones en un test de inteligencia administrado a una muestra de 250 personas. - ¿Cuál sería el porcentaje acumulado correspondiente a una puntuación de 97 en ese test?, ¿cómo la interpretaríamos? - ¿Y si la puntuación fuese igual a 103? - ¿Y si fuese igual a 91? => interpolación del %a

CI	$n_i$	$n_a$	$p_i$	$p_a$	$\%_a$
89	1	1	0,004	0,004	0,4
90	2	3	0,008	0,012	1,2
92	3	6	0,012	0,024	2,4
93	5	11	0,02	0,044	4,4
94	8	19	0,032	0,076	7,6
95	10	29	0,04	0,116	11,6
96	14	43	0,056	0,172	17,2
97	17	60	0,068	0,24	24
98	24	84	0,096	0,336	33,6
99	29	113	0,116	0,452	45,2
100	36	149	0,144	0,596	59,6
101	33	182	0,132	0,728	72,8
102	26	208	0,104	0,832	83,2
103	19	227	0,076	0,908	90,8
104	12	239	0,048	0,956	95,6
105	7	246	0,028	0,984	98,4
107	2	248	0,008	0,992	99,2
110	1	249	0,004	0,996	99,6
114	1	250	0,004	1	100

250

Los %a o percentiles son muy utilizados en la interpretación de las puntuaciones de los tests. Así, una tabla con la correspondencia entre los posibles valores observados (puntuaciones) en un test y los correspondientes porcentajes acumulados constituye lo que se conoce como el baremo de un test. Un baremo suele ser elaborado a partir de una muestra representativa de la población a la que pertenezcan los sujetos a los que a posteriori se tiene intención de aplicar ese test.

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

### 3.5 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  $X_i$  le restamos la media de su grupo obtenemos una puntuación diferencial o de diferencia, que representamos por  $x_i$  (minúscula):

$$x_i = X_i - \bar{X}$$

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.

### 3.6 Medidas de Posición

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 ".

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 tercer 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.

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

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.

$L_i$ ,  $f_{aa}$ ,  $f_i$ ,  $I_c$  : 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.

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

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

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

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 tercer 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:

$\frac{1 * n}{4}$  Si tenemos una serie de valores X1, X2, X3 ... Xn, se localiza el primer cuartil como el valor

$\frac{3(n+1)}{4}$  cuando n es par, y  $\frac{1(n+1)}{4}$  cuando n es impar. Para el tercer cuartil será



$$\frac{3 * n}{4} \quad (n \text{ par}); (n \text{ impar}).$$

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  $\frac{A(n + 1)}{10}$

$$\frac{A * n}{100} \text{ o } \frac{A(n + 1)}{100} .$$

siendo A el número del decil; y para los percentiles

Ejemplo:

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

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

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

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

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° .

### 3.7 Las Escalas Derivadas:

Una dificultad con las puntuaciones típicas se puede plantear a la hora de comunicar resultados debido a las posiciones decimales y valores negativos inherentes a las mismas. Es por ello que se han propuesto algunas transformaciones lineales de las puntuaciones típicas que pretenden hacerlas más intuitivamente interpretables

- Todas estas escalas derivadas de la escala de las puntuaciones típicas se basan en una transformación genérica del tipo:

$$D_i = a z_i + b,$$

Consecuencia inmediata de este tipo de transformación, las nuevas puntuaciones  $D$  pasarán de tener una media 0 y una desviación típica 1, a una media  $b$  y una desviación típica  $a$ .

Diversas propuestas de escala derivadas han sido planteadas sijn que haya generalizado el uso concreto de ninguna de ellas. Entre las que han tenido más repercusión las siguientes:

$$\text{- La escala } T \rightarrow T_i = 10 \cdot z_i + 50 \quad (\bar{T} = 50 \quad s_T = 10)$$

$$\text{- La escala } S \text{ o de estatinos} \rightarrow S_i = 2 \cdot z_i + 5 \quad (\bar{S} = 5 \quad s_S = 2)$$

$$\text{- La escala } CI \rightarrow CI_i = 15 \cdot z_i + 100 \quad (\bar{CI} = 100 \quad s_{CI} = 15)$$

### 3.8 Organización y Representación Gráfica de Datos Multivariados

Tras abordar en temas previos el tratamiento individualizado (univariado) de las variables, en éste y temas sucesivos se describen una serie de procedimientos asociados al tratamiento conjunto de dos o más variables, los cuales van a permitir extraer diversas facetas de la información compartida por esas variables. En bastantes momentos se va a ceñir esta exposición al caso bivariado (dos variables) por ser más sencillo en su presentación y por tratarse, con frecuencia, del caso particular más simple del modo general de abordar el problema a nivel multivariado (dos o más variables)

- I. La distribución conjunta multivariada , La tabla de contingencia. Representaciones gráficas.. El caso de dos variables categóricas.. El caso de dos variables cuantitativas.3. El caso de una variable categórica y una variable cuantitativa Tras abordar en temas previos el tratamiento individualizado de las variables (estadística univariada), en éste y temas sucesivos se describen

una serie de procedimientos asociados al tratamiento conjunto de dos o más variables. Estos procedimientos estadísticos van a permitir extraer diversas facetas de la información compartida por esas variables. En bastantes momentos se va a ceñir esta exposición al caso bivariado (dos variables) por ser más sencillo en su presentación y por tratarse, con frecuencia, del caso particular más simple del modo general de abordar el problema a nivel multivariado (dos o más variables).. La distribución conjunta multivariada De modo análogo a lo que se planteó para el caso univariado en el tema, un resumen básico de la información de un grupo de o más variables consiste en la distribución conjunta de frecuencias de las mismas, la cual se basa en el conteo del número de casos (frecuencias) que presentan las distintas combinaciones de valores que a nivel empírico se hayan dado para esas variables. Las modalidades de una distribución conjunta de frecuencias consisten, no en los valores de una variable concreta, sino en todas las posibles combinaciones de los valores de las variables que se Prof. J. Gabriel Molina y María F. Rodrigo.

2. consideren excepto aquellas combinaciones que no se hayan presentado a nivel empírico y que por tanto no tiene sentido incluir en la distribución de frecuencias. Ejemplo: La siguiente tabla de datos procede de un estudio sobre las relaciones de pareja en que se obtuvo información en una muestra de 7 sujetos de las 3 variables siguientes: Sexo (: Hombre; : Mujer); N° de parejas estables a lo largo de los últimos 5 años; y Situación emocional actual (: Satisfactoria; : Ni satisfactoria ni insatisfactoria; 3: Insatisfactoria). ID Sexo Num\_parejas Sit\_actual La organización de los datos de la anterior tabla en forma de distribución conjunta de frecuencias absolutas quedaría como sigue, donde  $X = \text{Sexo}$ ;  $Y = \text{Num\_parejas}$  y  $Z = \text{Sit\_actual}$ :  $n_{i X i ; Y i ; Z i ; ; ; 3 ; ; 3 ; ; ; 8 ; ; 3 6 ; ; 5 ; ; ; 3 ; ; ; ; 3 ; 7$  La distribución conjunta de frecuencias relativas o proporciones ( $p_{i i}$ ) y la de porcentajes ( $\% i i$ ) pueden obtenerse a partir de las frecuencias absolutas dividiendo cada frecuencia absoluta entre el n° de casos (n) y multiplicando las frecuencias relativas por cien, respectivamente. Prof. J. Gabriel Molina y María F. Rodrigo
3. El ordenamiento de las modalidades en una distribución conjunta de frecuencias carece de sentido, si bien, se suelen situar en orden alfabético/numérico creciente a fin de poder localizar más fácilmente cualquier combinación de valores de las variables. La obtención de las frecuencias acumuladas, ya sean absolutas, relativas o porcentajes, carece también aquí de sentido dado que las modalidades de la distribución no representan un continuo -al igual que ocurría con las distribuciones de frecuencias de las variables categóricas. Por ello mismo, los índices de posición y de dispersión descritos para las variables categóricas podrían ser aplicados también en una distribución conjunta de frecuencias. Inconvenientes: Si el número de variables es amplio o si alguna de las variables tiene muchos valores, el número de combinaciones de valores posibles puede llegar a ser muy numeroso, tanto que la visualización de la distribución de frecuencias resulte poco ventajosa en su propósito de resumir la información de los datos.

Existen algunas alternativas que pueden ayudar a resolver este problema en algunas situaciones:

() En el caso de una variable (o más) con muchos posibles valores (como es lo más habitual con variables cuantitativas), una opción es colapsar esos valores en intervalos. De este modo se pierde en precisión de la información, si bien, se hace factible la reducción drástica el número de combinaciones de valores posibles. Por ejemplo, supongamos que tenemos dos variables, cada una con los tiempos (seg.) empleados por un grupo de personas en ejecutar dos tareas procedentes de un test de aptitudes mecánicas. Suponiendo un rango de valores en ambas variables de entre y seg., una posible agrupación de los mismos a la hora de crear una distribución conjunta de frecuencias podría quedar así:  $n_i$   $X_i$  (seg.),  $Y_i$  (seg.) -5 ; ; ; ; ; ; ; ( Cuántas filas tendrá la anterior distribución conjunta de frecuencias?, Cuántas filas tendría, en teoría, la distribución conjunta de frecuencias sin agrupar?)

### 3.9 La Distribución Conjunta Multivariada.

De modo análogo al caso univariado, un resumen básico de la información de un grupo de 2 o más variables consiste en la distribución conjunta de frecuencias de las mismas, la cual se basa en el conteo del número de casos (frecuencias) que presentan las distintas combinaciones de valores que a nivel empírico se hayan dado para esas variables. Las modalidades ( $X_i, Y_i, Z_i \dots$ ) de una distribución conjunta representarán, no a los valores de una variable concreta, sino a todas las posibles combinaciones de los valores de las variables que se consideren –excepto aquellas combinaciones que no se hayan presentado a nivel empírico y que por tanto no tiene sentido incluir en la distribución de frecuencias

Ejemplo: La siguiente tabla de datos procede de un estudio sobre las relaciones de pareja en que se obtuvo información en una muestra de 71 sujetos de las 3 variables siguientes: sexo (1: hombre; 2: mujer); nº de parejas estables a lo largo de los últimos 5 años; y situación emocional actual (1: satisfactoria; 2: ni satisfactoria ni insatisfactoria; 3: insatisfactoria).

ID	Sexo	Num parejas	Sit actual
1	1	1	3
2	1	4	2
3	2	1	1
4	2	2	1
5	2	1	3
6	1	0	1
7	2	3	2
...	...	...	...
71	1	1	1

La organización de los datos de la anterior tabla en forma de distribución conjunta de frecuencias (absolutas) quedaría como sigue, donde  $X = \text{Sexo}$ ;  $Y = \text{Num\_parejas}$  y  $Z = \text{Sit\_actual}$

$X_b Y_b Z_i$	$n_i$
1,0,1	4
1,0,2	3
1,0,3	2
1,1,1	12
1,1,2	8
1,1,3	6
1,2,1	5
1,2,2	1
1,2,3	2
1,4,2	1
2,0,1	6
...	...
2,3,2	1
	71

### 3.10. La tabla de contingencia

En el caso de dos variables, una forma muy conveniente de visualizar la distribución conjunta de ambas es en forma de tabla de contingencia, esto es, una tabla de doble entrada en que cada

lado de la tabla contiene las modalidades de una variable. En las casillas interiores de la tabla aparecen las frecuencias conjuntas (ya sean absolutas, relativas o porcentajes) de la combinación de los valores fila y columna correspondientes.

• Ejemplo: se llevó a cabo un estudio para evaluar si el estado de ánimo de los mayores de 65 años podía verse influido por el hecho de vivir en una residencia geriátrica o no. Se recogieron datos de una muestra de 500 personas de las variables “Estado de ánimo”: negativo (-); neutro ( $\pm$ ); positivo (+) y “Vivir en residencia” (Sí; No). La distribución conjunta de frecuencias es la siguiente:

	-	$\pm$	+
<b>Sí</b>	48	42	60
<b>No</b>	70	105	175

¿Cómo se ha construido esa tabla de contingencia? Realizando, a partir de la matriz de datos original, un recuento del n° de casos que presentan cada combinación de par de valores.

<b>Caso</b>	<b>Residencia</b>	<b>Estado ánimo</b>
1	S	-
2	N	$\pm$
3	S	-
4	S	+
...	...	...
500	N	$\pm$

• También a partir de esa tabla de datos, como ya se ha visto, es posible obtener la distribución de cada variable por separado (= distribuciones marginales):

Residencia (X)

$X_i$	$n_i$	$p_i$
Sí	150	0.30
No	350	0.70
	500	1

Estado ánimo (Y)

$Y_i$	$n_i$	$p_i$
-	118	0.236
±	147	0.294
+	235	0.470
	500	1

- o La distribución conjunta de ambas variables:

$X_i, Y_i$	$n_i$	$p_i$
Sí $\cap$ -	48	0.096
Sí $\cap$ ±	42	0.084
Sí $\cap$ +	60	0.120
No $\cap$ -	70	0.140
No $\cap$ ±	105	0.210
No $\cap$ +	175	0.350
	500	1

En las tablas de contingencia es habitual incluir en los laterales las sumas de las celdas de filas y columnas => distribuciones marginales (= distribución de cada variable por separado

	-	±	+	Total
Sí	48	42	60	150
No	70	105	175	350
Total	118	147	235	500

### 3.1 I-Cálculo 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:

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

Donde:

li: límite inferior de la clase que lo contiene: valor que representa la posición de la medida. fi: la frecuencia de la clase que contiene la medida solicitada.

fa-l: 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



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

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

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

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

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

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

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

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

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

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

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

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 está 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:

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

Donde:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$Q_3 = 51,5 + \frac{49,5 - 65}{1} * 6 = -41,5$$

### 3.12 Representación gráfica en el Análisis de Datos

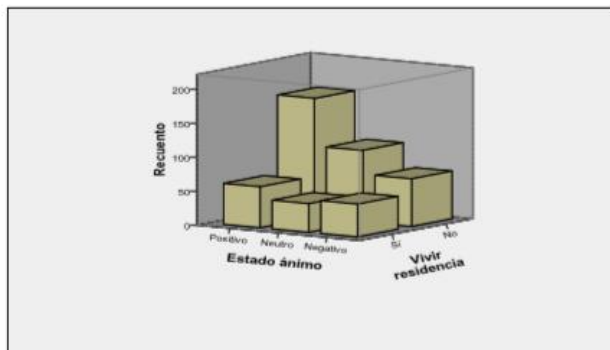
Se presentan a continuación una serie de gráficos orientados a mostrar datos multivariados, si bien, la mayoría se ciñen al caso de 2 variables. Se diferencia su presentación en función del tipo de variables al que van dirigidos. No se van a presentar gráficos específicos para las variables ordinales, si bien, puede ser utilizado cualquiera de los orientados a variables categóricas o, si se asume naturaleza cuantitativa para las mismas, los orientados a este tipo de variables.

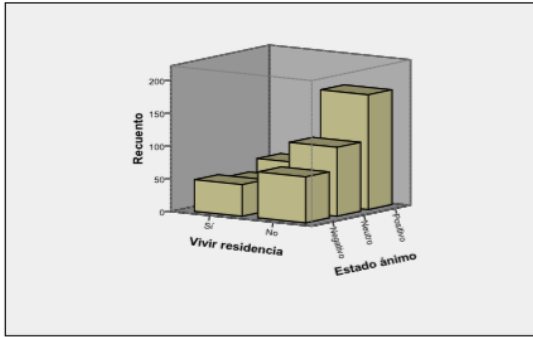
La realización de los estudios clínico-epidemiológicos implica finalmente emitir unos resultados cuantificables de dicho estudio o experimento. La claridad de dicha presentación es de vital importancia para la comprensión de los resultados y la interpretación de los mismos. A la hora de representar los 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.

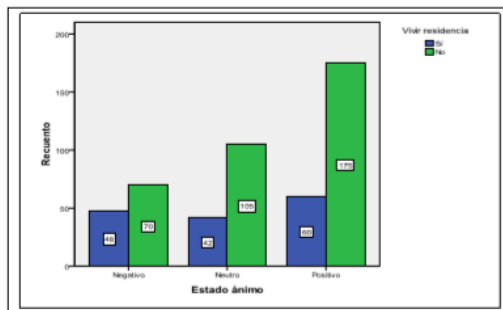
### 3.13 El caso de dos variables Categóricas

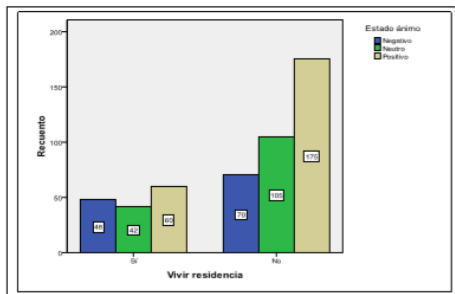
El diagrama de barras tridimensional o 3-D Ejemplos de diagrama de barras 3-D con la distribución conjunta de frecuencias absolutas de “Estado de ánimo” y “Vivir residencia”, intercambiando la posición de ambas variables.





- El diagrama de barras Ejemplos de diagrama de barras de la distribución conjunta de frecuencias absolutas de “Estado de ánimo” y “Vivir residencia” intercambiando la posición de ambas variables. Para diferenciar verbalmente ambos, haremos referencia al primero como diagrama de barras de frecuencias absolutas de la variable “Estado de ánimo” agrupada en función de “Vivir residencia”, mientras que al segundo como diagrama de barras de la variable “Vivir residencia” agrupada en función de “Estado de ánimo”. En ambos gráficos se representan frecuencias absolutas, por lo que las barras en ambos deben sumar el total del tamaño de la muestra ( $n = 500$ ).



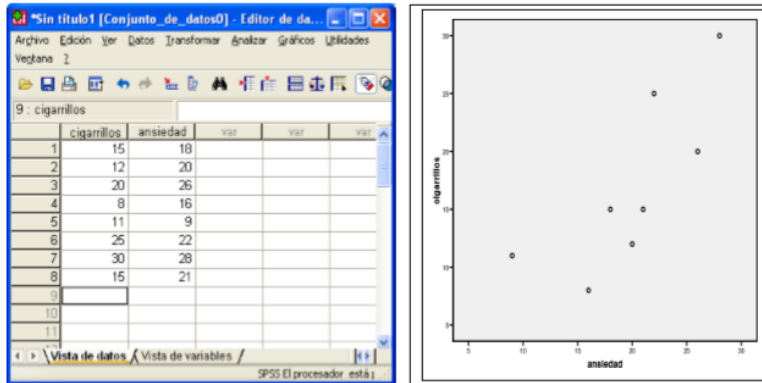


Este tipo de diagrama de barras de la distribución conjunta de dos variables sólo puede ser representado en SPSS para frecuencias absolutas, pero no para porcentajes. Cuando se solicita que se representen los porcentajes (la opción de frecuencias relativas no se ofrece en SPSS), lo que se representa no son los porcentajes en sí (ver gráfico de ejemplo a continuación), sino un tipo de porcentaje condicional que resulta útil para un determinado objetivo sobre el que se tratará en el próximo capítulo. Puede comprobarse como las barras de este gráfico no suman 100.

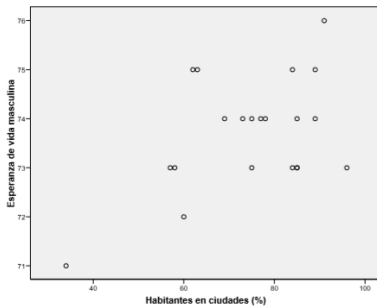
### 3.14 El caso de dos variables cuantitativas

El diagrama de dispersión (bivariado) Ejemplo de diagrama de dispersión obtenido a partir de los datos de una muestra de 8 fumadores en las variables “Nº de cigarrillos que, en promedio, se fuma al día” y “Puntuación en un test de ansiedad [0, ..., 30]”. Se muestran también los datos a partir de los que ha sido obtenido el mismo con el programa SPSS:

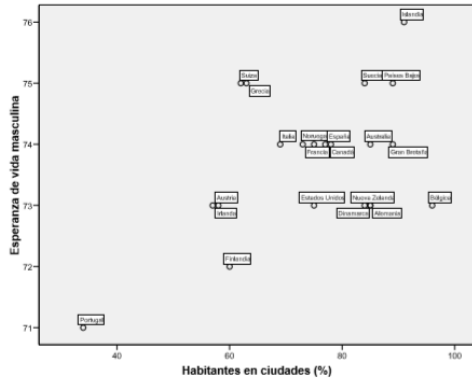




Otro ejemplo de diagrama de dispersión obtenido, en este caso, a partir de los datos de esperanza de vida masculina y el porcentaje de población que viven en ciudades en un conjunto de países de la OCDE en el año 1995:



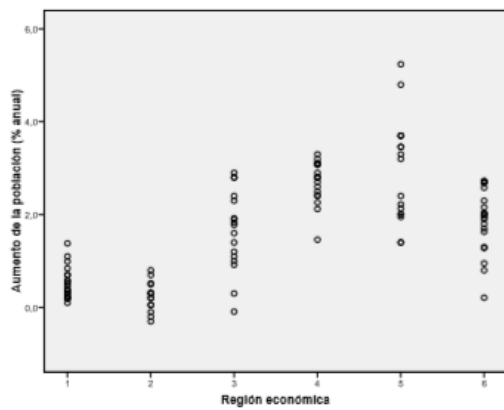
A continuación se muestra el mismo diagrama de dispersión con los puntos etiquetados, a fin de facilitar la posición de los casos (en este ejemplo, países) en la representación gráfica. Esta estrategia sólo resulta conveniente en el caso en que el número de casos en el archivo de datos no sea muy numeroso pues, en caso contrario, resulta complicado visualizar la información presentada:



### 3.15. El caso de una variable categórica y una variable cuantitativa

El diagrama de dispersión también puede ser aplicado en la representación conjunta de la distribución de frecuencias absolutas de una variable categórica y una variable cuantitativa. A este tipo de gráfico se le denomina en algunos textos como diagrama de puntos y es habitual que aparezca representada la variable categórica en el eje de abscisas y la variable cuantitativa en el eje de ordenadas.

Ejemplo de diagrama de dispersión de la distribución conjunta de las variables “Región económica” [1:OCDE; 2: Europa oriental; 3: Asia/Pacífico; 4: África; 5: Oriente Medio; 6: America latina]” y “% anual de crecimiento de la población” obtenida a partir de los datos recogidos para un total de 109 países de todo el mundo (N = 109):



Piénsese acerca de la tabla de datos a partir de la que se ha obtenido este gráfico: ¿cuántas

filas?; ¿cuántas columnas?; ¿qué tipo de variables?... Fragmento de la tabla de datos original:

1	Azerbaijón	Oriente Medio	1,4
2	Afganistán	Asia / Pacífico	2,8
3	Alemania	OCDE	,4
4	Arabia Saudí	Oriente Medio	3,2
5	Argentina	América Latina	1,3
6	Armenia	Oriente Medio	1,4
7	Australia	OCDE	1,4
8	Austria	OCDE	,2
9	Bahrein	Oriente Medio	2,4
10	Bangladesh	Asia / Pacífico	2,4
11	Barbados	América Latina	,2
12	Bélgica	OCDE	,2
13	Bielorusia	Europa Oriental	,3
14	Bolivia	América Latina	2,7
15	Bosnia	Europa Oriental	,7
16	Botswana	África	2,7
17	Brasil	América Latina	1,3
18	Bulgaria	Europa Oriental	-,2
19	Burkina Faso	África	2,8

## Unidad 4

### ESTADÍSTICOS DE ASOCIACIÓN ENTRE VARIABLES

#### 4.1. Concepto de asociación entre variables

El análisis estadístico de la asociación (relación, covarianza, correlación) entre variables representa una parte básica del análisis de datos en cuanto que muchas de las preguntas e hipótesis que se plantean en los estudios que se llevan a cabo en la práctica implican analizar la existencia de relación entre variables.

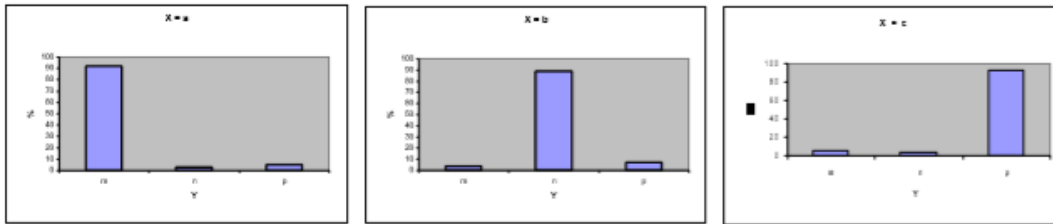
- La existencia de algún tipo de asociación entre dos o más variables representa la presencia de algún tipo de tendencia o patrón de emparejamiento entre los distintos valores de esas variables.

A modo de ejemplo esquemático, si tenemos una variable  $X$  [ $a, b, c$ ] y otra variable  $Y$  [ $m, n, p$ ], de modo que los datos empíricos evidencian que las entidades que en  $X$  son  $a$  en  $Y$  tienden a ser  $n$  (o viceversa), que las que son  $b$  tienden a ser  $p$ , y que las que son  $c$  tienden a ser  $m$ , se pone de manifiesto que existe cierta asociación entre ambas variables.

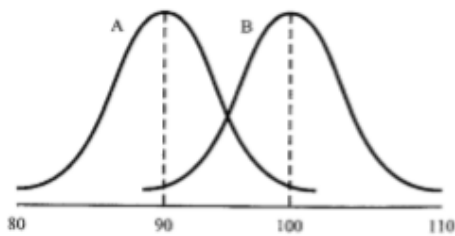
- Más formal que ésta, Solanas et al. (2005) ofrecen otra propuesta de definición general de lo que significa la asociación entre 2 variables: la existencia de asociación entre dos variables

indicaría que la distribución de los valores de una de las dos variables difiere en función de los valores de la otra

Sean el caso para nuestro ejemplo anterior, de las distribuciones de frecuencias (expresadas en %) de la variable Y para aquellos casos que en la variable X son 'a', 'b' y 'c', respectivamente:



Otro ejemplo de la presencia de asociación entre 2 variables, una, la puntuación en un test de aptitud lingüística [0 a 150] y, la otra, la variable sexo [A: hombre; B: mujer]. Véase a continuación, para un conjunto de datos de estas dos variables, la diferencia existente entre las distribuciones de frecuencias de la variable “Aptitud lingüística” condicionada a la variable “Sexo”:



- Complementariamente, se habla de independencia entre variables cuando no existe tal patrón de relación entre los valores de las mismas.

## 4.2. Midiendo la asociación en el caso de dos variables categóricas

- ¿Qué se puede decir acerca de la asociación entre las dos variables de la tabla de contingencia (“Estado de ánimo” y “Vivir en residencia”)?

	-	±	+	Total
Sí	48	42	60	150
No	70	105	175	350
Total	118	147	235	500

Para evaluar si ambas variables están relacionadas hay que observar si la distribución de los valores de una de las variables difiere en función de los valores de la otra, esto es, hay que comparar las distribuciones condicionadas de una de las dos variables agrupada en función de los valores de la otra. Si no hay relación entre las variables estas distribuciones deberían ser iguales. Por ejemplo, podemos comparar las distribuciones de frecuencias absolutas de “Estado de ánimo” condicionadas a vivir en una residencia (48, 42, 60) y a no vivir en una residencia (70, 105, 175).

- Si nos fijamos en las distribuciones de frecuencias absolutas de “Estado de ánimo” condicionadas a los valores de “Vivir en residencia”, se observa que estas distribuciones no son iguales, sin embargo, esto puede ser debido a que hay más sujetos que no viven en una residencia (350) que sujetos que sí viven en ella (150). En conclusión, no se deben comparar las distribuciones condicionadas en frecuencias absolutas si el número de casos difiere en las categorías de la variable condicionante o agrupadora.
- La asociación entre dos variables categóricas aparece más explícita en una tabla de frecuencias relativas condicionadas, pues de ese modo se relativiza el posible diferente tamaño de los subgrupos definidos por cualquiera de las dos variables. Este tipo de tabla se puede obtener de 2 formas alternativas, bien dividiendo las celdas de cada fila entre el respectivo marginal (total) de fila, bien cada columna entre el total de columna. Ambas tablas permitirán llegar al mismo

tipo de conclusiones respecto a la asociación entre las 2 variables.

- Si la relación entre las variables es asimétrica, la variable agrupadora o condicionante sería la que sea considerada la variable explicativa (predictora, independiente). Por ejemplo, en un estudio en que se evalúa la influencia del “Nivel de estudios” [primarios, secundarios, superiores] sobre la “Percepción de la influencia de la ciencia en la sociedad” [negativa, indiferente, positiva], dado que el nivel de estudios sería la variable explicativa, deberíamos comparar las distribuciones de la percepción de la influencia de la ciencia condicionadas al nivel de estudios, es decir, en cada categoría de nivel de estudios. En nuestro ejemplo sobre “Estado de ánimo” y “Vivir en residencia”, dado que la relación es asimétrica y la variable explicativa es “Vivir en residencia” debemos comparar las distribuciones de “Estado de ánimo” condicionadas a “Vivir en residencia”:

	-	±	+	Total
Sí	0,32 (48/150)	0,28 (42/150)	0,40 (60/150)	1
No	0,20 (70/350)	0,30 (105/350)	0,50 (175/350)	1
Total	0,236 (118/500)	0,294 (147/500)	0,470 (235/500)	1

En la tabla anterior, la comparación de las distribuciones de frecuencias relativas condicionales con la distribución marginal de la variable de respuesta, nos permitirá comprobar la existencia de asociación entre las dos variables y, en el caso de que exista, la naturaleza de la misma.

A modo de ejemplo, si no hubiera relación entre ambas variables, las distribuciones de frecuencias relativas de “Estado de ánimo” condicionadas a “Vivir en residencia” serían iguales a la distribución marginal de la variable “Estado de ánimo”, esto es:

	-	±	+	Total
<b>Si</b>	0,236	0,294	0,470	1
<b>No</b>	0,236	0,294	0,470	1
Total	0,236	0,294	0,470	1

Como se puede comprobar, las distribuciones de frecuencias relativas de “Estado de ánimo” condicionadas a “Vivir en residencia” difieren bastante de las de la tabla anterior. Así, por ejemplo, se observa que la proporción de sujetos que tienen un estado de ánimo negativo entre los que viven en una residencia (0,32) es superior a la que cabría esperar si no hubiera relación entre ambas variables (0,236). Esto parece indicar que sí existe una relación entre ambas variables.

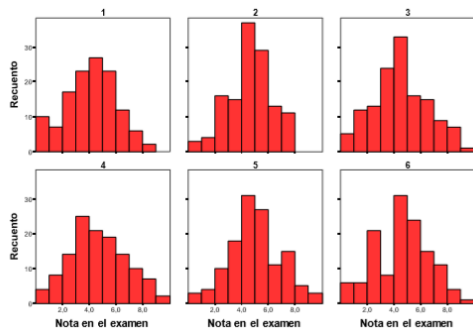
### 4.3. Midiendo la asociación El caso de una variable categórica y una cuantitativa

- De nuevo, el análisis de este tipo de asociación supone comparar las distribuciones condicionales de una variable para los distintos valores que toma la otra. Normalmente, se suele tomar como condicionada a la cuantitativa y como condicionante a la categórica, si bien, las conclusiones a las que llegaríamos serían las mismas si se hiciese al revés. Si no hay diferencias entre las distribuciones condicionales, ello indicará que no hay asociación entre ambas variables.

- Ejemplo del caso en que se quiera analizar la asociación entre las variables “Nota en un examen de una asignatura [0 a 10]” y “Grupo en el que se está matriculado [1 a 6]”, disponiéndose de los datos de un total de 768 estudiantes de 6 grupos:

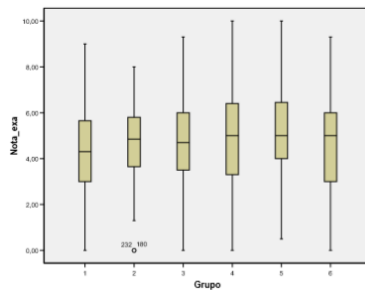
Grupo 1		Grupo 2		Grupo 3		Grupo 4		Grupo 5		Grupo 6	
$X_i$	$n_i$	$X_i$	$n_i$	$X_i$	$n_i$	$X_i$	$n_i$	$X_i$	$n_i$	$X_i$	$n_i$
0	2	0	3	0	2						
.3	3	1,3	1	.5	1						
.5	1	1,8	1	.8	2						
.8	3	2,0	2	1,3	2						
1,0	1	2,3	4	1,5	2						
1,3	1	2,5	2	1,8	1						
1,5	2	2,8	5	2,0	7						
1,8	2	2,9	1	2,3	3						
2,0	2	3,0	4	2,5	2						
2,3	3	3,3	3	2,6	3						
2,5	2	3,5	6	2,8	2						
2,6	1	3,8	3	2,9	1						
2,8	6	3,9	1	3,0	2						
3,0	5	4,0	2	3,3	3						
3,3	3	4,3	5	3,5	7						
3,5	5	4,5	6	3,8	9						
3,8	7	4,6	1	3,9	1						
4,0	8	4,7	8	4,0	4						
4,3	7	4,8	6	4,1	1						
4,5	5	4,9	6	4,3	3						
4,7	4	5,0	5	4,5	9						
4,8	3	5,1	1	4,7	6						
4,9	5	5,3	5	4,8	7						
5,0	3	5,4	1	4,9	4						
5,1	1	5,5	6	5,0	3						
5,3	6	5,6	1	5,3	4						
5,5	4	5,8	9	5,5	4						
5,8	7	5,9	1	5,6	1						
6,0	5	6,0	5	5,8	1						
6,1	1	6,3	2	5,9	2						
6,3	5	6,5	4	6,0	4						
6,5	3	6,8	4	6,3	3						
6,8	2	6,9	1	6,5	6						
7,0	1	7,0	2	6,8	2						
7,5	3	7,3	4	7,0	4						
7,6	1	7,5	3	7,1	1						
8,0	2	7,8	2	7,3	2						
9,0	2	8,0	2	7,5	5						
				8,0	1						
				8,3	2						
				8,4	1						
				9,0	4						
				9,3	1						

Dada la dificultad que puede representar comparar las distribuciones condicionales de una variable cuantitativa, se puede recurrir a representaciones gráficas que faciliten la realización de este tipo de comparación. A modo de ejemplo, las dos siguientes obtenidas para los datos anteriores con el paquete estadístico SPSS o, también, el diagrama de dispersión presentado en el capítulo anterior para estos mismos datos:





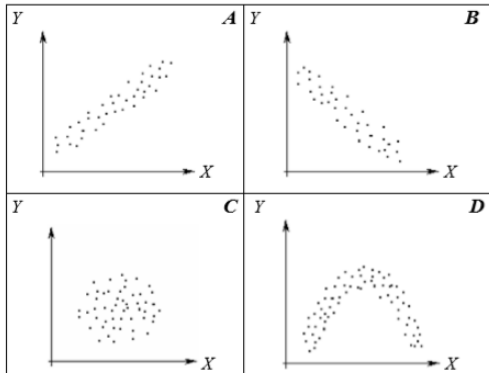
Ejemplo de diagrama de caja y bigotes con la distribución de la variable “Nota en un examen de una asignatura” condicionada a la variable “Grupo en el que se está matriculado”



Estos gráficos nos permiten comparar el grado de solapamiento (coincidencia) de las distribuciones condicionales. En general, cuanto mayor sea el solapamiento, menor será la intensidad de la asociación entre las dos variables y, viceversa, cuanto menor sea el solapamiento, mayor será el tamaño del efecto de la relación. En el ejemplo anterior existe bastante solapamiento entre las 6 distribuciones condicionales, poniendo de manifiesto una escasa relación entre ambas variables.

#### 4.4. Midiendo la asociación El caso de dos variables cuantitativas

- Al igual que en los casos anteriores, la existencia de correlación o asociación entre 2 variables cuantitativas viene determinada por la presencia de diferencias en las distribuciones condicionales de una variable para los distintos valores de la otra.
- Sin embargo, dado el número tan amplio de distribuciones condicionales que se pueden llegar a obtener en este caso, es más habitual analizar la asociación directamente sobre un diagrama de dispersión, observando la disposición de la nube de puntos que representa la distribución conjunta de ambas variables. Así, ¿qué podríamos decir acerca de la asociación entre los 4 pares de variables cuyos diagramas de dispersión se muestran a continuación?



Un aspecto relevante del análisis de la correlación entre dos variables cuantitativas es que la presencia de ésta se puede plantear de acuerdo a diferentes modelos o patrones de asociación, por ejemplo, en forma de línea recta, tal como en los ejemplos A (relación lineal directa o positiva) y B (relación lineal inversa o negativa) de arriba, o en forma curvilínea tal como en D (relación parabólica o cuadrática). Así, la forma de evaluar la intensidad de la correlación suele consistir en analizar el ajuste de la nube de puntos al modelo de asociación que se considere que representa más adecuadamente a la distribución conjunta de ambas variables.

- En la cuantificación de la asociación entre 2 variables cuantitativas nos vamos a ceñir al supuesto de que un modelo de relación lineal subyace a la asociación entre ambas. Subrayar que con frecuencia se obvia en los textos estadísticos que la relación que se analiza es en realidad una relación de tipo lineal. Los índices más utilizados en la práctica estadística a la hora de analizar la intensidad o tamaño del efecto de la relación lineal entre dos variables son los tres siguientes:

(1) La covarianza ( $S_{xy}$  o  $\sigma_{xy}$ ):

$$S_{xy} = \frac{\sum (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{n}$$

- Al numerador de esta expresión se le conoce en la literatura estadística como suma de

productos cruzados (SPXY), por lo que la anterior expresión queda como:  $s_{XY} = SPXY / n$  -  
Desarrollando algebraicamente la fórmula de la covarianza se puede llegar a una fórmula que se considera más conveniente cuando el cálculo de la misma se ha de realizar de forma manual

$$S_{xy} = \frac{\sum X_i Y_i}{n} - \bar{X}\bar{Y}$$

Ejemplo para las variables Calificaciones en música (X) y Calificaciones en matemáticas (Y) obtenidas por un grupo de 10 niños

X	Y	X*Y
5	6	30
7	8	56
8	7	56
5	6	30
9	10	90
4	5	20
5	5	25
5	7	35
7	6	42
8	9	72
$\bar{X} = 6,3$	$\bar{Y} = 6,9$	$\Sigma (X*Y) = 456$

$$S_{xy} = \frac{\sum X_i Y_i}{n} - \bar{X}\bar{Y} = \frac{456}{10} - (6,3 \cdot 6,9) = 2,13$$

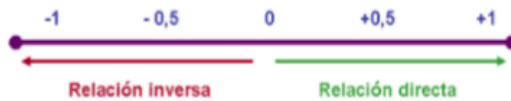
- La covarianza puede tomar valores tanto positivos como negativos. A nivel interpretativo, un mayor valor de la covarianza en valor absoluto indicará una relación lineal más intensa entre las dos variables. Un valor positivo pone de manifiesto una relación lineal directa; uno negativo, una relación lineal inversa; y si igual o muy próximo a 0, la inexistencia de relación lineal entre las dos variables.

(2) El coeficiente de correlación producto-momento de Pearson ( $r_{XY}$ )

- Los inconvenientes de la covarianza –por una parte, no tiene valores máximo y mínimo y, por otra parte, depende de las unidades de medida de las variables- se resuelven estandarizando este índice al dividirlo por el producto de las desviaciones típicas de ambas variables. Se obtiene así el conocido como coeficiente de correlación de Pearson:

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$

– El coeficiente de correlación de Pearson se interpreta de modo análogo a la covarianza pero, al oscilar entre -1 y 1 como máximo, la interpretación del mismo resulta más intuitiva a la vez que facilita el establecimiento de comparaciones entre los coeficientes obtenidos para conjuntos de datos distintos.



El coeficiente de correlación posee las siguientes características:

- El valor del coeficiente de correlación es independiente de cualquier unidad usada para medir las variables.
- 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 I. 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 I. Cálculo del Coeficiente de correlación de Pearson entre las variables talla y peso de 20 niños varones

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

$Y - \bar{Y} \quad (X - \bar{X}) * (Y - \bar{Y})$  Tabla I. 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

$$\sum 290.8$$

$X(\text{Media de } \bar{X} = 66.35)$

$Y(\text{Media de } \bar{Y} = 7.6)$

$$\text{Covarianza} = \frac{\sum(\bar{X} - X) * (\bar{Y} - Y)}{n - 1} = \frac{290.8}{19} = 15.30$$

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

$S_x = \text{Desviación típica } x = 8.087$

$S_y = \text{Desviación típica } y = 2.137$

## 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.

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

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).

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

Como quiera que  $r = 0.885 > 2.10 * 0.109 = 2.30$  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:

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

$L_n$  representa el logaritmo neperiano en la base  $e$

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

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

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

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

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

$$\frac{e^{2z_1} - 1}{e^{2z_1} + 1} \quad \alpha \quad \frac{e^{2z_2} - 1}{e^{2z_2} + 1}$$

Utilizando el ejemplo de la Tabla I, 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.

$$\frac{e^{2 \cdot 0.922} - 1}{e^{2 \cdot 0.922} + 1} \quad \alpha \quad \frac{e^{2 \cdot 1.873} - 1}{e^{2 \cdot 1.873} + 1}$$

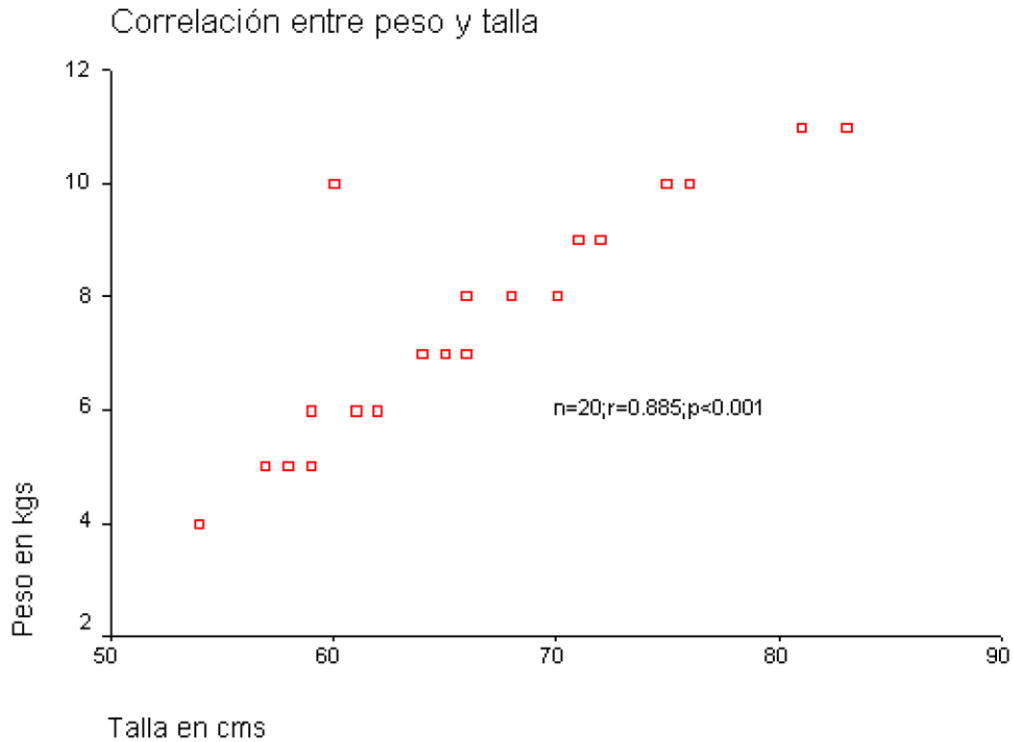
0.726 a 0.953 son los intervalos de confianza (95%) de r.

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) puede que  $X$  influya o cause  $Y$ , b) puede que influya 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:

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

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      Contenido de Nicotina por cigarrillo

(nmol/litro)	(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:

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

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

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

$$\begin{aligned} \sum r_x &= \sum r_y = 55 & \sum r_x^2 &= \sum r_y^2 = 385 \\ \sum r_x r_y &= 2(8) + 5(3) + 8(6) + \dots + 4(1) = 325 \end{aligned}$$

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

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.

#### 4.5. El modelo de regresión lineal

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,<sup>1</sup> y en dónde se incluía una versión del teorema de Gauss-Márkov.

##### 4.5.1 Conceptos básicos sobre Análisis de regresión Lineal.

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.<sup>2</sup> 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 ( $k = 1, \dots, K$ ), o cualquier transformación de éstas que generen un hiperplano de parámetros desconocidos: 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 ( $i = 1, \dots, l$ ) cualquiera, se registra el comportamiento simultáneo de la variable dependiente y las variables explícitas (las perturbaciones aleatorias se suponen no observables).

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



#### 4.6 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 uno 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 muestral no viene influenciado por los valores de las perturbaciones correspondientes a otras observaciones muestrales.
4. Regresores estocásticos.
5. Independencia lineal. No existen relaciones lineales exactas entre los regresores.
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:<sup>3</sup>
  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.

#### 4.7. Tipos de modelos de regresión lineal:

Podemos realizar 3 modelos de análisis distintos en función del número de variables y la forma de interactuar entre ellas:

- Modelo de regresión **lineal simple**
- Modelo de regresión **lineal múltiple**
- Modelo de regresión **no lineal**

- **Regresión lineal simple**

Sólo se maneja una variable independiente, por lo que sólo cuenta con dos parámetros. Son de la forma 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: Derivando respecto a ( $Y$ ) e igualando a cero, se obtiene: El análisis de regresión lineal simple es el más utilizado y el más sencillo de todos. Se trata de estudiar el efecto de una variable independiente sobre una única variable dependiente de la primera —o que al menos a nivel teórico hemos considerado que es dependiente—. Empleando esta ecuación de regresión lineal simple se puede realizar una estimación basándose en los datos obtenidos.

*Fórmula de la regresión lineal simple*

$$y = B_0 + B_1 x + \varepsilon$$

Donde  $B_0$  es el valor de la variable independiente,  $B_1$  es la variable dependiente y  $\varepsilon$  representa el residuo o error. La función de  $\varepsilon$  es explicar la posible variabilidad de los datos que no pueden explicarse a través de la relación lineal de la fórmula.

**Modelo de regresión lineal múltiple:** En el caso de la regresión lineal múltiple nos encontramos con un modelo que sencillamente cuenta con más de una variable independiente. Este modelo lo aplicaremos cuando tengamos razones para creer que hay más de un factor que afecta a la variable de estudio.

*Fórmula de la regresión lineal múltiple*

$$Y = 0 + B_1 \cdot X_1 + B_2 \cdot X_2 + \dots + B_n \cdot X_n + \varepsilon$$

Nuevamente, Y representa la variable dependiente que se está estudiando y  $B_1$ ,  $B_2$ ,  $B_n$  son todas las variables independientes que pueden afectar al valor de la variable dependiente Y. De igual modo  $\varepsilon$  sigue representando el posible error existente.

**Modelo de regresión no lineal:** Hay ocasiones en las que la relación que puede darse entre variables independientes y la variable dependiente no tenga un desarrollo lineal, sino que tengan, por ejemplo, un crecimiento exponencial. En esos casos, el modelo de regresión no lineal entra en juego y permite que obtengamos una aproximación de los valores de la variable dependiente en un entorno no lineal. Tengamos presente que el proceso de una regresión no lineal es más complejo, ya que puede no coincidir el número de parámetros con el de las variables independientes.

**4.8. Fórmulas de regresión no lineales:** En muchos casos es posible modificar un modelo no lineal para convertirlo en un modelo lineal. Por ejemplo aplicando logaritmos a su fórmula inicial.

- *Regresión exponencial*

$$y = a \cdot b^x$$

Esta fórmula puede transformarse en una lineal mediante el uso de logaritmos. Quedando de la siguiente manera la fórmula.

$$\log y = \log(a \cdot b^x) = \log a + x \log b$$

- *Regresión potencial*

$$y = a \cdot x^b$$

Si volvemos a aplicar logaritmos, transformamos en un modelo lineal la fórmula inicial.  
 $\log y = \log a + b \log x$

- *Regresión parabólica*

$$y^* = a_0 + a_1 x + a_2 x^2$$

#### 4.9. Estimadores:

Para abordar este tema, es importante recordar que la inferencia estadística es el procedimiento mediante el cual se trabaja con muestras y de acuerdo con los resultados arrojados se realizan las inferencias en las poblaciones.

La estadística inferencial se divide en dos áreas de estudio: los estimadores y las pruebas de hipótesis.

El estudio de los estimadores implica calcular a partir de los datos de la muestra, algún estadístico que permite una aproximación del parámetro correspondiente a la población de la cual se extrajo la muestra; para cada uno de los parámetros se pueden calcular dos diferentes tipos de estimadores: el puntual y el de intervalos.

El estimador puntual es un solo valor numérico empleado para estimar el parámetro correspondiente de la población; los estimadores de este tipo se forman por medio de fórmulas por ejemplo,

∑ si se desea calcular la media muestral, se emplea la fórmula

se emplea la fórmula  $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$  ;

Y al unísono se estima la media de la población, es decir,  $\mu$

Los estimadores cuentan con tres propiedades y la nomenclatura para representar un parámetro, es decir, el valor numérico que describe las características de la población, el cual es  $\theta$ ; y para expresar el estimador, es decir, el valor numérico que describe las características de la muestra, se emplea  $\hat{\theta}$ .

Los estimadores tienen las siguientes propiedades:

- Si el  $\lim_{n \rightarrow \infty} E(\hat{\theta}) = E(\theta)$
- $\lim_{n \rightarrow \infty} \text{var}(\hat{\theta}) = 0$ ; propiedad de los insesgados.
- Cuando  $n \rightarrow \infty$ ; y si  $n$  crece se vuelve insesgado.

#### 4.10 Regresión Lineal Por Covarianza

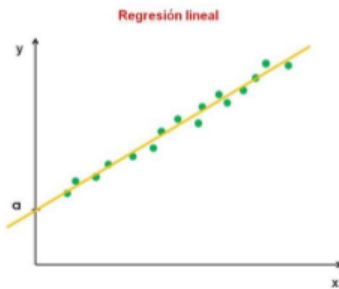
En múltiples ocasiones se requiere analizar la relación entre dos variables cuantitativas. Los dos objetivos fundamentales de este análisis son:

1. Determinar si dichas variables están asociadas y en qué sentido se da dicha asociación (es decir, si los valores de una de las variables tienden a aumentar –o disminuir- al aumentar los valores de la otra).
2. Estudiar si los valores de una variable pueden ser utilizados para predecir el valor de la otra.

La forma correcta de abordar el primer problema es recurriendo a coeficientes de correlación. Sin embargo, el estudio de la correlación es insuficiente para obtener una respuesta a la segunda pregunta, ya que se limita a indicar la fuerza de la asociación mediante un único número, tratando las variables de modo simétrico, mientras que lo que se busca es modelar

dicha relación y usar una de las variables para explicar la otra. Para tal propósito se recurrirá a la técnica de regresión<sup>2</sup>.

La regresión lineal permite definir la recta que mejor se ajusta a esta nube de puntos. Gráficamente:



La recta está definida por la siguiente expresión:

**$y = a + bx$**  ( donde  $y$  es la variable dependiente y  $x$  es la variable independiente. Sus coeficientes representan:

$b$ : determina la pendiente de la recta, es decir, su grado de inclinación. Se calcula como la covarianza de las dos variables, dividida por la varianza de la variable  $x$  :

$$b = \frac{\sigma_{xy}}{\sigma_x^2}$$

$a$  : es el valor que toma  $y$  cuando la variable independiente  $x$  vale cero. Es el punto donde la recta cruza el eje vertical, llamado ordenada al origen de la recta. Se calcula como la media de la variable  $y$  , menos la media de la variable  $x$  multiplicada por el parámetro  $b$  que se ha calculado:

$$a = \bar{y} - b \cdot \bar{x}$$

Ejemplo. Obtener y graficar la recta de regresión con los datos de estatura y peso de los 10 jugadores del equipo de fútbol americano de la UNAM del ejemplo anterior.

Solución. Los valores obtenidos de la covarianza y de la desviación estándar de las estaturas son:  $2510 \cdot xy = \sigma$

$0431850 \cdot x = \sigma$

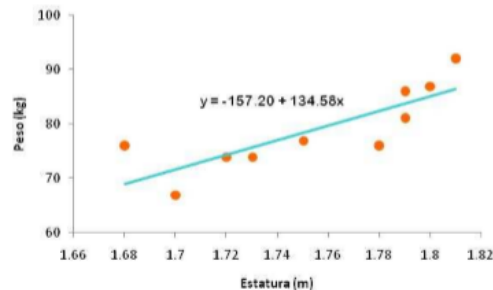
$$\Rightarrow b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{0.251}{(0.043185)^2} = 134.58$$

Considerando que  $\bar{x} = 1.755$  m. y que  $\bar{y} = 79$  kg:

$$a = \bar{y} - b \cdot \bar{x} = 79 - (134.58)(1.755) = -157.20$$

Por lo que la recta de regresión lineal es:  $y = -157.20 + 134.58x$

Su gráfica es:



Nótese como la gráfica es congruente con el coeficiente de correlación ( $r = 0.82030$ ). Muestra una pendiente positiva y se ajusta a una recta lo que ratifica que a mayor estatura de los jugadores, mayor es su peso.

## 4.11 Regresión Lineal Por Mínimos Cuadrados

Una estrategia adicional para ajustar adecuadamente el comportamiento o la tendencia general de los datos a través de una recta que minimice la suma de los cuadrados de las distancias verticales de los puntos a la recta. Este método se conoce como regresión por mínimos cuadrados.

Para obtener una recta de la forma:  $y = a + bx$ , este método se basa en la aplicación de las siguientes expresiones:

$$a = \frac{\sum_{i=1}^n x_i \cdot \sum_{i=1}^n x_i y_i - \sum_{i=1}^n y_i \cdot \sum_{i=1}^n (x_i)^2}{\left(\sum_{i=1}^n x_i\right)^2 - n \sum_{i=1}^n (x_i)^2}$$

$$b = \frac{n \cdot \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i\right)^2}$$

Ejemplo.

Aplicando el método de mínimos cuadrados, obtener la recta de regresión para los siguientes datos:

Estatura (m)	1.72	1.79	1.78	1.75	1.80	1.79	1.81	1.70	1.68	1.73
Peso (kg)	74	81	76	77	87	86	92	67	76	74



Solución.

Acomodando la tabla convenientemente y obteniendo  $x \cdot y$  y  $x^2$  se tiene:

Estatura $x$	Peso $y$	$x \cdot y$	$x^2$
1.72	74	127.28	2.9584
1.79	81	144.99	3.2041
1.78	76	135.28	3.1684
1.75	77	134.75	3.0625
1.8	87	156.6	3.24
1.79	86	153.94	3.2041
1.81	92	166.52	3.2761
1.7	67	113.9	2.89
1.68	76	127.68	2.8224
1.73	74	128.02	2.9929
17.55	790	1388.96	30.8189

$$a = \frac{\sum_{i=1}^{10} x_i \cdot \sum_{i=1}^{10} x_i y_i - \sum_{i=1}^{10} y_i \cdot \sum_{i=1}^{10} (x_i)^2}{\left(\sum_{i=1}^{10} x_i\right)^2 - 10 \sum_{i=1}^{10} (x_i)^2} = \frac{17.55(1388.96) - 790(30.8189)}{(17.55)^2 - 10(30.8189)} = -157.19$$

$$b = \frac{10 \cdot \sum_{i=1}^{10} x_i y_i - \sum_{i=1}^{10} x_i \sum_{i=1}^{10} y_i}{10 \sum_{i=1}^{10} (x_i)^2 - \left(\sum_{i=1}^{10} x_i\right)^2} = \frac{10(1388.96) - 17.55(790)}{10(30.8189) - (17.55)^2} = 134.58$$

Por lo que la recta de regresión lineal usando mínimos cuadrados es:  $y = -157.19 + 134.58x$   
 Puede advertirse que esta ecuación es prácticamente igual que la obtenida por medio del método de covarianza. Eso significa que ambos métodos son aplicables siempre que el coeficiente de correlación sea cercano a la unidad.

**Bibliografía básica:**

- “Estadística Descriptiva 2019,” Gongora Cortes, Jose Juan Hernandez Ramirez, Roberto, Editorial Trillas.
- Barón-López, J. (2005). Estadística descriptiva en Ciencias del Comportamiento. Editorial: Thompson.
- Freud, John y Gary A. Simon, Estadística elemental, México, Editorial Prentice-Hall, 2003