Estadística: estadística
Grado en Relaciones Laborales | Curso 2019-2020
Tema 1. Análisis descriptivo de una variable

Alejandro Saavedra Nieves

Introducción

estadística.
(Del al. Statistik).
1. f. Estudio de los datos cuantitativos de la población, de los recursos naturales e industriales, del tráfico o de cualquier otra manifestación de las sociedades humanas.
2. f. Conjunto de estos datos.
3. f. Rama de la matemática que utiliza grandes conjuntos de datos numéricos para obtener inferencias basadas en el cálculo de probabilidades.

Diccionario de la lengua española. Real Academia Española

Introducción

La estadística es una ciencia con base matemática referente a la recolección, análisis e interpretación de datos, que busca explicar condiciones regulares en fenómenos de tipo aleatorio.
Es transversal a una amplia variedad de disciplinas, desde la física hasta las ciencias sociales, desde las ciencias de la salud hasta el control de calidad, y es usada para la toma de decisiones en áreas de negocios e instituciones gubernamentales.

Wikipedia

Introducción

La solución de problemas de la vida real responden a razonamientos de tipo "inductivo". En general, se pretende extender a un todo las conclusiones obtenidas en una parte. Así, se hacen continuamente afirmaciones acerca de un grupo de individuos habiendo observado en realidad sólo una parte (a veces pequeña) de ellos. El método científico adecuado para validar tales generalizaciones es el método estadístico.

- Un psicólogo podrá caracterizar a sus pacientes, diagnosticar síndromes, recomendar tratamientos...
- Un sociólogo será capaz de orientar a los políticos, interpretar estados sociales, adelantarse a las crisis...
- Un médico conocerá los riesgos de determinados medicamentos, de las peculiaridades de ciertos pacientes en relación con algunos fármacos...
- Un economista ayudará a la empresa a prevenir problemas, a identificar riesgos, a diseñar campañas atendiendo a los perfiles de los clientes...

Introducción

La repetición de los experimentos en condiciones idénticas da lugar a resultados distintos, debido a factores o causas como pueden ser los errores por la manipulación del experimentador o por el aparato de medida; pero además tenemos la “variabilidad”de los individuos objeto de estudio: dos seres vivos nunca son iguales, ni un individuo es igual a sí mismo en diferentes etapas de su vida.

Tenemos entonces fenómenos que son esencialmente impredecibles en sus resultados, y las afirmaciones acerca de ellos sólo pueden hacerse en términos de probabilidad o posibilidad *(experimentos aleatorios)*. El modo de obtener resultados científicos válidos a partir de datos que son fundamentalmente impredecibles es a través de las técnicas estadísticas, pues son capaces de tener en cuenta la variabilidad aludida.

Introducción

En toda investigación experimental pueden distinguirse tres etapas:

1. Diseño
2. Recopilación de datos
3. Análisis de los resultados y obtención de las conclusiones

* En las tres etapas, la Estadística es fundamental.

Ejemplo: El servicio municipal de transportes, con el propósito de mejorar su funcionamiento, desea conocer el tiempo de espera de sus pasajeros antes de subirse a un autobús. ¿Cómo debe proceder?

Conceptos básicos

Población: Es un conjunto de objetos, personas, entidades de la más diversa índole, que constituyen el objetivo de nuestro estudio. Es el universo de individuos al cual se refiere el estudio que se pretende realizar.

Variable: Rasgo o característica de los elementos de la población que se pretende analizar.

Muestra: Subconjunto de la población cuyos valores de la variable que se pretende analizar son conocidos. En nuestro contexto, en el procedimiento de extracción va a intervenir el azar. Por tanto, la muestra consistirá en un conjunto de realizaciones de un experimento aleatorio.

Tamaño muestral: Número de individuos que componen la muestra. Lo representamos por n.

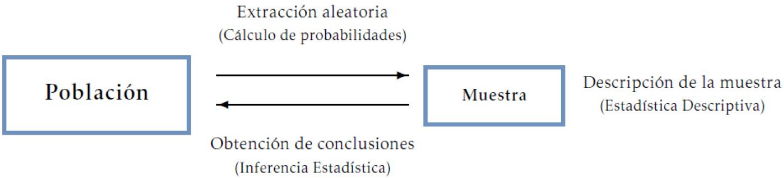
Estadística

Clasificamos las tareas vinculadas a la Estadística en tres grandes disciplinas:

Estadística Descriptiva. Se ocupa de recoger, clasificar y resumir la información contenida en la muestra.

Cálculo de Probabilidades. Es una parte de la matemática teórica que estudia las leyes que rigen los mecanismos aleatorios.

Inferencia Estadística. Pretende extraer conclusiones para la población a partir del resultado observado en la muestra.



La Inferencia Estadística tiene un objetivo más ambicioso que el de la mera descripción de la muestra (Estadística Descriptiva). Dado que la muestra se obtiene mediante procedimientos aleatorios, el Cálculo de Probabilidades es una herramienta esencial de la Inferencia Estadística.

Tipos de Variables

Variables cualitativas: No aparecen en forma numérica, sino como categorías o atributos.

- el sexo
- color de ojos
- nivel de estudios
- deporte favorito
- ...

Variables cuantitativas: Toman valores numéricos porque son frecuentemente el resultado de una medición.

- edad (m) de una persona
- el peso (kg.) de una persona
- número de hijos
- número de empleados en empresas
- ...

Las variables cualitativas, también llamadas *atributos o variables categóricas* pueden clasificarse a su vez en:

- Cualitativas nominales: Miden características que no toman valores numéricos (sin orden). A estas características se les llama modalidades.
  - el sexo (hombre o mujer)
  - creencias religiosas
  - color de ojos
  - ...
- Cualitativas ordinales: Sus posibles valores admiten una relación de orden.
  - máximo curso en el que se está matriculado
  - categoría hotelera
  - hábitos de consumo de tabaco
  - ...

Es muy común asignar códigos numéricos a las categorías de los datos cualitativos. Esto no los convierte en datos cuantitativos: esos códigos numéricos son meros símbolos que representan a las categorías.

Tipos de Variables. Variables cuantitativas

Se clasifican a su vez en:

- Cuantitativas discretas: Toman un número discreto de valores (en el conjunto de números naturales). Sus posibles valores están separados entre sí.
  - número de multas en un año
  - número de hijos
  - número de pasajeros en vuelos nacionales
  -
- Cuantitativas continuas: Toman valores numéricos dentro de un intervalo real.
  - el peso
  - la edad
  - nivel de glucosa en sangre
  - salario bruto anual
  - ...

Ejemplo

El servicio médico de una empresa recibe la visita de ocho de sus empleados con dolor lumbar a durante una semana. Todos los datos se encuentran resumidos en la siguiente tabla. Clasifica las variables recogidas (sexo, peso, estatura, temperatura, número de visitas previas al servicio y dolor).

Sexo	Peso (kg.)	Estatura (m.)	Temperatura (°C)	Visitas	Dolor
M	63	1.74	38	0	Leve
M	58	1.63	36.5	2	Intenso
H	84	1.86	37.2	0	Intenso
M	47	1.53	38.3	0	Moderado
M	70	1.75	37.1	1	Intenso
M	57	1.68	36.8	0	Leve
H	87	1.82	38.4	1	Leve
M	55	1.46	36.6	1	Intenso

Descripción de variables cualitativas y cuantitativas discretas

Supongamos que los n valores que puede tomar una variable X son: $x_1, x_2, \ldots, x_m$.

Frecuencia absoluta: Se denota por n_i y representa el número de veces que ocurre el resultado x_i.

Frecuencia relativa: Se denota por f_i y representa la proporción de datos en cada una de las clases,

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

Frecuencia absoluta acumulada. Es el número de veces que se ha observado el resultado x_i o valores anteriores. La denotamos por

$$N_i = n_1 + n_2 + n_3 + \cdots + n_i = \sum_{x_j \leq x_i} n_j$$

Frecuencia relativa acumulada. Es la frecuencia absoluta acumulada dividida por el tamaño muestral. La denotamos por

$$F_i = \frac{N_i}{n} = f_1 + f_2 + f_3 + \cdots + f_i = \sum_{x_j \leq x_i} f_j$$

Descripción de variables cualitativas y cuantitativas discretas

Las frecuencias se pueden escribir ordenadamente mediante una **tabla de frecuencias**, que adopta esta forma:

x_i	n_i	f_i	N_i	F_i
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	n_m	f_m	N_m	F_m

Descripción de variables cualitativas y cuantitativas discretas

Las frecuencias se pueden escribir ordenadamente mediante una **tabla de frecuencias**, que adopta esta forma:

x_i	n_i	f_i	N_i	F_i
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	n_m	f_m	N_m	F_m

Propiedades:

Frecuencias absolutas	$0 \leq n_i \leq n$	$\sum_{i=1}^{m} n_i = n$
Frecuencias relativas	$0 \leq f_i \leq 1$	$\sum_{i=1}^{m} f_i = 1$
Frecuencias absolutas acumuladas	$0 \leq N_i \leq n$	$N_m = n$
Frecuencias relativas acumuladas	$0 \leq F_i \leq 1$	$F_m = 1$

No calculamos las frecuencias acumuladas si la variable es cualitativa nominal.

- **Variables cualitativas**:
  - Diagrama de barras.
    Consiste en levantar sobre cada valor o modalidad de la variable una barra (segmento de recta o rectángulo) de altura igual o proporcional a la correspondiente frecuencia absoluta (n_i) o relativa (f_i).
  - Polígono de frecuencias.
    Se obtiene uniendo mediante segmentos de recta los puntos (x_i, n_i) o (x_i, f_i) para todo $i = 1, \ldots, k$.
  - Gráfico de sectores.
    Se divide un círculo en sectores circulares, uno por cada valor o modalidad de la variable, de forma que el ángulo de cada sector sea $\alpha_i = 360 \times f_i$ grados.

Ejemplo. Variable cualitativa nominal. Procedencia

En una muestra de 50 turistas de Vigo se estudia su "procedencia" y se ha obtenido la información detallada a continuación:

- 35 personas residen fuera de la UE,
- 10 personas residen en la UE (no en España),
- 3 personas son españolas y
- 2 personas son gallegas.

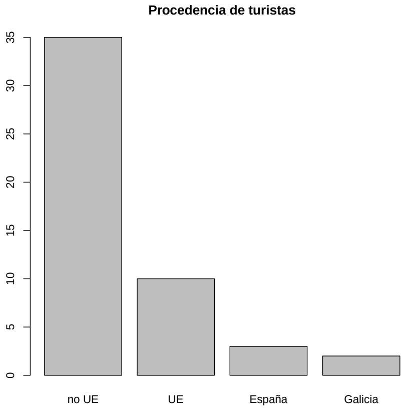
Ejemplo. Variable cualitativa nominal. Procedencia

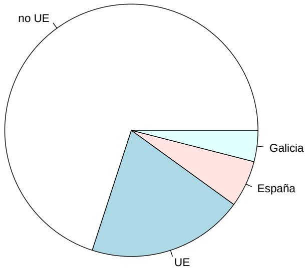
Tamaño muestral: $n = 50$

Procedencia (x_i)	n_i	f_i
Fuera de la UE	35	0'7
UE (no España)	10	0'2
España	3	0'06
Galicia	2	0'04

Nótese que no calculamos las frecuencias acumuladas pues la variable Procedencia es nominal.

Representaciones gráficas. Variable cualitativa nominal. Procedencia





Ejemplo. Variable cualitativa ordinal. Nivel máximo de Estudios

En una muestra de 30 candidatos a una oferta de empleo se está estudiando "máximo nivel de estudios", y se han obtenido los siguientes resultados:

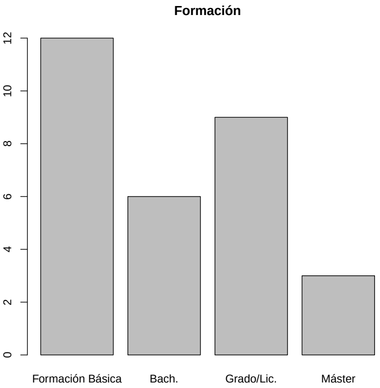
- 12 personas tienen formación básica,
- 6 personas han acabado bachillerato,
- 9 personas han estudiado una carrera y
- 3 personas han realizado un máster.

Ejemplo. Variable cualitativa ordinal. Nivel máximo de estudios

Tamaño muestral: $n = 30$.

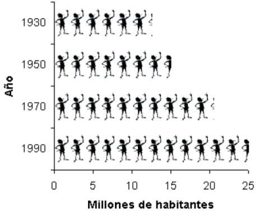
Nivel máximo de estudios	n_i	f_i	N_i	F_i
Formación Básica	12	$0'4$	12	$0'4$
Bachillerato	6	$0'2$	18	$0'6$
Grado/Licenciatura	9	$0'3$	27	$0'9$
Máster	3	$0'1$	30	1

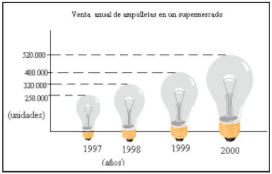
Representaciones gráficas. Variable cualitativa nominal. Nivel máximo de estudios



Representaciones gráficas. Variable cuantitativas

- **Variables cuantitativas discretas**: diagrama de barras o también el de sectores (cuando los valores que toma X son pocos)
- **Variables cuantitativas continuas agrupadas**: histograma, diagrama de tallos y hojas.
- Otros:
  - Pictograma. Se sustituye la típica barra por un dibujo relacionado con la variable que se representa.





Nº de bombillas vendidas en un supermercado

  - Cartograma. Se representa la variable sobre un mapa.

Ejemplo. Variable cuantitativa discreta. Número de empleados

Consideremos una muestra de 80 empresas, en las que observamos el número de empleados. Los datos que se han obtenido son:

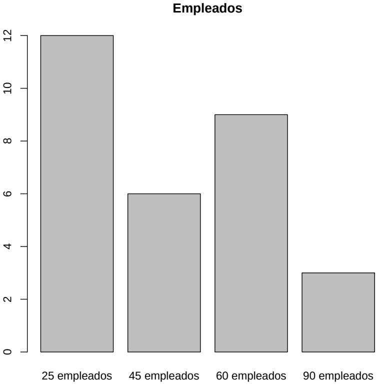
- 8 empresas con 25 empleados,
- 20 empresas con 45 empleados,
- 28 empresas con 60 empleados y
- 24 empresas con 90 empleados.

Ejemplo. Variable cuantitativa discreta. Número de empleados

Tamaño muestral: $n = 80$

x_i	n_i	f_i	N_i	F_i
25	8	0.1	8	0.1
45	20	0.25	28	0.35
60	28	0.35	56	0.7
90	24	0.3	80	1

Representaciones gráficas. Variable cuantitativa discreta. Número de empleados



Construcción de un histograma

- Para construir las frecuencias es habitual agrupar los valores que puede tomar la variable en intervalos. De este modo contamos el número de veces que la variable cae en cada intervalo
- A cada uno de estos intervalos le llamamos **intervalo de clase (Clase)**: $(L_{i-1}, L_i]$:
  - Su punto medio es la **marca de clase**: $c_i = \dfrac{L_{i-1} + L_i}{2}$,
  - la **amplitud** del intervalo es $a_i = L_i - L_{i-1}$ y
  - la **densidad de datos** del intervalo: $d_i = \dfrac{n_i}{a_i}$.
- Por tanto, para la definición de las frecuencias y la construcción de la tabla de frecuencias sustituiremos los valores x_i por los intervalos de clase y las marcas de clase.

Descripción de variables cuantitativas continuas

Algunas consideraciones a tener en cuenta:

- *Número de intervalos a considerar:*
    - Cuantos menos intevalos tomemos, menos información se recoge.
    - Cuantos más intervalos tomemos, más difícil es manejar las frecuencias.

  Se suele tomar como número de intervalos el entero más próximo a $\sqrt{n}$.

- *Amplitud de cada intervalo:* Lo más común, salvo justificación en su contra, es tomar todos los intervalos de igual longitud.

- *Posición de los intervalos:* Los intervalos deben situarse allí donde se encuentran las observaciones y de forma contigua.

Ejemplo. Variable cuantitativa continua

Se considera una muestra de 10 personas, y se les pregunta la edad (en años) en la que firmaron su primer contrato indefinido. Las respuestas fueron las siguientes:

$$52, 47, 51, 28, 64, 31, 22, 53, 29, 23$$

¿Cómo resumimos la información contenida en los datos de la variable Edad?

Ejemplo. Variable cuantitativa continua

Tabla de frecuencias con estos datos:

- Muestra ordenada: 22, 23, 28, 29, 31, 47, 51, 52, 53, 64.
- Recorrido$= 64 - 22 = 42$.
- Número de intervalos$\simeq \sqrt{10} \simeq 3'162 \simeq 3$.
- Como $42/3 = 14$, tomaremos 15 como amplitud de cada intervalo. Así conseguimos contener toda la muestra y los extremos de los intervalos resultan manejables.

Ejemplo. Variable cuantitativa continua

Tabla de frecuencias con estos datos:

- Muestra ordenada: 22, 23, 28, 29, 31, 47, 51, 52, 53, 64.
- Recorrido$= 64 - 22 = 42$.
- Número de intervalos$\simeq \sqrt{10} \simeq 3'162 \simeq 3$.
- Como $42/3 = 14$, tomaremos 15 como amplitud de cada intervalo. Así conseguimos contener toda la muestra y los extremos de los intervalos resultan manejables.

Intervalo de clase $(L_{i-1}, L_i]$	Marca de clase c_i	n_i	f_i	N_i	F_i	Densidad de frecuencia $d_i = n_i / (L_i - L_{i-1})$
$(20, 35]$	$27'5$	5	$0'5$	5	$0'5$	$5/15$
$(35, 50]$	$42'5$	1	$0'1$	6	$0'6$	$1/15$
$(50, 65]$	$57'5$	4	$0'4$	10	1	$4/15$

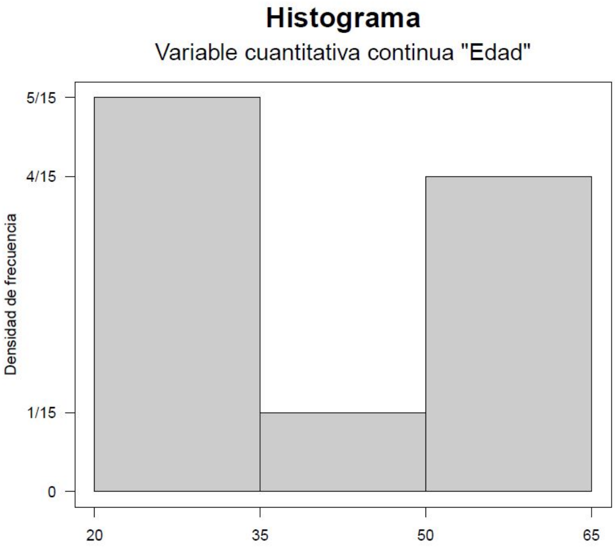
La distribución de frecuencias de una variable continua se representa mediante el llamado **histograma.**

El área (y no la altura) de los rectángulos debe ser proporcional a la frecuencia. Así, el eje de ordenadas no refleja la frecuencia, sino que la altura de cada rectángulo representa la **densidad de frecuencia** sobre ese intervalo, definida como:

$$\text{Densidad de frecuencia} = \frac{\text{frecuencia absoluta}}{\text{amplitud}}$$

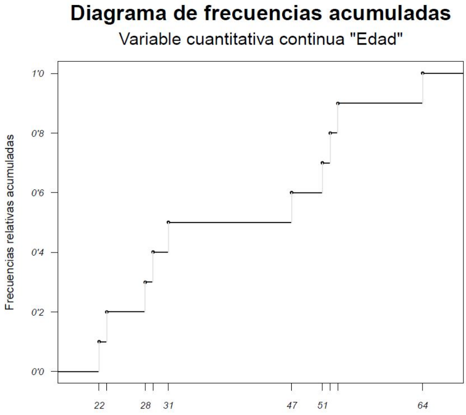
Sólo cuando todos los intervalos tengan la misma amplitud, será equivalente representar la frecuencia o la densidad de frecuencia.

Representaciones gráficas. Variable cuantitativa continua. Edad



Representaciones gráficas. Variable cuantitativa continua. Edad

En el caso de disponer de todos los datos, representaremos las frecuencias acumuladas mediante el **diagrama de frecuencias acumuladas** o **diagrama escalonado**, que en este caso tiene la peculiaridad de que todos los saltos son de amplitud $\frac{1}{n}$. Los saltos se producen en los datos de la muestra, de modo que en las zonas donde hay más datos la escalera crece más rápidamente.



Representaciones gráficas. Variable cuantitativa continua. Edad

En caso de agrupación en intervalos, las frecuencias acumuladas se representan mediante el **polígono de frecuencias acumuladas**. Como no se conoce el lugar exacto en el que se encuentra cada individuo de la muestra, se reparte la frecuencia de cada intervalo de manera uniforme dentro del intervalo, lo cual resulta en segmentos cuya pendiente es la densidad de frecuencia en cada intervalo.

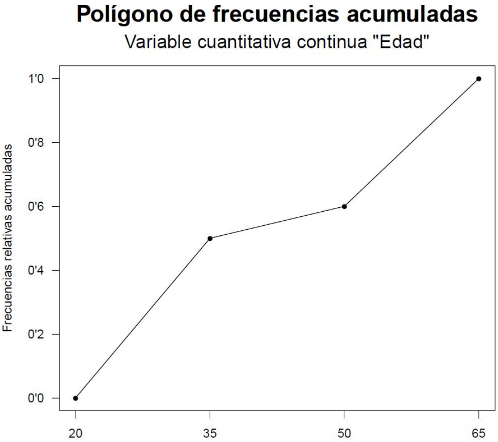


Diagrama de tallos y hojas

- Permite obtener simultáneamente una distribución de frecuencias de la variable y su representación gráfica.
- Para su construcción, basta separar en cada dato el último dígito de la derecha (la hoja) del bloque de cifras restantes (el tallo).
- Una vez construido, permite la recuperación de los datos originales.
- Proporciona una visualización de la distribución de frecuencias como el histograma.

Diagrama de tallos y hojas: ejemplo

Los siguientes datos corresponden a los precios de la libra de cobre en la Bolsa de Metales de Londres en Enero de 2000.

Día	Precio	Día	Precio	Día	Precio
1		12	82.7	23	
2		13	84.2	24	84.9
3		14	83.8	25	84.1
4	83.1	15		26	83.6
5	82.5	16		27	82.5
6	83.1	17	83.7	28	83.5
7	83.1	18	83.7	29	
8		19	85.0	30	
9		20	86.1	31	82.2
10	83.0	21	85.6		
11	82.5	22			

Tallos	Hojas
82.	2 5 5 5 7
83.	0 1 1 1 5 5 6 7 7 8
84.	1 2 3
85.	0 6
86.	1

Ejercicio

El servicio médico de una empresa recibe la visita de ocho de sus empleados con dolor lumbar a durante una semana. Todos los datos se encuentran resumidos en la siguiente tabla. Clasifica las variables recogidas (sexo, peso, estatura, temperatura, número de visitas previas al servicio y dolor).

Sexo	Peso (kg.)	Estatura (m.)	Temperatura (°C)	Visitas	Dolor
M	63	1.74	38	0	Leve
M	58	1.63	36.5	2	Intenso
H	84	1.86	37.2	0	Intenso
M	47	1.53	38.3	0	Moderado
M	70	1.75	37.1	1	Intenso
M	57	1.68	36.8	0	Leve
H	87	1.82	38.4	1	Leve
M	55	1.46	36.6	1	Intenso

Resume la información contenida en los datos de las diferentes variables.

Medidas características: Medidas de posición, de dispersión y de forma

Por **medida** entendemos un número que se calcula sobre la muestra y que refleja cierta cualidad de la misma. Parece claro que el cálculo de estas medidas requiere la posibilidad de efectuar operaciones con los valores que toma la variable. Por este motivo, en lo que resta del tema tratamos sólo con variables cuantitativas.

Medidas características: Medidas de posición, de dispersión y de forma

Por **medida** entendemos un número que se calcula sobre la muestra y que refleja cierta cualidad de la misma. Parece claro que el cálculo de estas medidas requiere la posibilidad de efectuar operaciones con los valores que toma la variable. Por este motivo, en lo que resta del tema tratamos sólo con variables cuantitativas.

- Medidas de posición: son medidas que nos indican la posición que ocupa la muestra
- Medidas de dispersión: se utilizan para describir la variabilidad o esparcimiento de los datos de la muestra respecto a la posición central
- Medidas de forma: tratan de medir el grado de simetría y apuntamiento en los datos. Estas no las estudiaremos en detalle!

Medidas de posición

- Media aritmética
- Mediana
- Moda
- Cuantiles

Medidas de posición. Media aritmética

Se define la media aritmética (o simplemente media) como:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{o bien} \quad \bar{x} = \sum_{i=1}^{k} (c_i \text{ ó } x_i) f_i$$

donde la primera expresión se emplea cuando se dispone de todos los datos (sin agrupar), mientras que la segunda expresión se aplica a datos agrupados, empleando las frecuencias de cada valor diferente.

En el caso de una variable continua, tenemos dos opciones: o calculamos la media con todos los datos, que denotamos por x_i (los sumamos y dividimos por el tamaño muestral), o usamos la tabla de frecuencias considerando las marcas de clase (c_i en vez de x_i) y las frecuencias en cada clase.

Propiedades

1. La media se mide en las mismas unidades que los datos originales.

2. Es el centro de gravedad de los datos:

$\min\ x_i \leq \bar{x} \leq max\ x_i$

$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \min_{a \in \mathbb{R}} \sum_{i=1}^{n}(x_i - a)^2$

3. Si $y_i = a + bx_i$ entonces $\bar{y} = a + b\bar{x}$, esto es, si se multiplican por b las observaciones de X (por ejemplo, al cambiar de unidades) y se trasladan sumando una constante a, entonces la media de X cambia sus unidades y se traslada en la misma constante.

4. Si la distribución de frecuencias es simétrica respecto a un valor M, entonces $\bar{x} = M$.

Ejemplo. Medidas de posición. Media. Número de horas extra

La dirección de una empresa evalúa el número de horas extra de sus 20 empleados, detallados en la siguiente tabla.

N.º de horas extra

x_i	n_i	f_i	N_i	F_i	$x_i \cdot f_i$
0	5	$0'25$	5	$0'25$	0
1	8	$0'40$	13	$0'65$	$0'40$
2	4	$0'20$	17	$0'85$	$0'40$
3	2	$0'10$	19	$0'95$	$0'30$
4	1	$0'05$	20	1	$0'20$
SUMAS	20	1			$1'3$

Por lo tanto, la media es:

$\bar{x} = \sum_{i=1}^{k} x_i f_i = 1'3$ horas extra.

Ejemplo. Medidas de posición. Media. Edad

- Ya que disponemos de todos los datos, calcularemos la media de la variable edad con todos ellos:
  $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(52+47+51+\cdots+53+29+23) = \frac{400}{10} = 40$ años
- A continuación, calculamos la media de la variable edad a partir de los datos ya agrupados en intervalos de clase:

Intervalo de clase $(L_{i-1}, L_i]$	Marca de clase c_i	n_i	f_i	N_i	F_i	$c_i f_i$
$(20, 35]$	$27'5$	5	$0'5$	5	$0'5$	$13'75$
$(35, 50]$	$42'5$	1	$0'1$	6	$0'6$	$4'25$
$(50, 65]$	$57'5$	4	$0'4$	10	1	23
	SUMAS	20	1			41

En este caso:
$\bar{x} = \sum_{i=1}^{k} c_i f_i = 41$ años
Si queremos dar la edad en meses:
$y_i = 12 * x_i$ y $\bar{y} = 12 * \bar{x} = 12 * 40 = 480$ meses.

Media ponderada

En la media aritmética todos los valores tienen el mismo peso, pero puede interesarnos que haya datos con más peso (importancia) que otros, indicando su importancia relativa dentro del conjunto de datos. Su valor es:

$$\overline{x}^w = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_k w_k}{w_1 + w_2 + \ldots + w_k}$$

Ejemplo: cálculo de la nota final

Un opositor obtiene las puntuaciones de 8, 7 y 6 en tres pruebas sucesivas de dificultad creciente, cada una con doble valoración que la anterior. ¿Cual es su nota media en la oposición?

Asignamos los pesos $w_1 = 1$, $w_2 = 2$ y $w_3 = 4$. Entonces,

$$\overline{x}^w = \frac{8 \times 1 + 7 \times 2 + 6 \times 4}{1 + 2 + 4} = 6{,}5714$$

Media en subpoblaciones

La población está divida en L grupos de los cuales conocemos:

- N_j, cuántos individuos hay en cada uno de ellos y
- $\bar{x}_j$, la media de la variable dentro del grupo,

para cada grupo $j = 1, \ldots, L$.

La media total $\bar{x}$ es la media ponderada, mediante el número de observaciones, de las medias de las subpoblaciones, es decir:

$$\overline{x} = \frac{\overline{x}_1 N_1 + \overline{x}_2 N_2 + \ldots + \overline{x}_L N_L}{N_1 + N_2 + \ldots + N_L}$$

Ejercicio La tabla siguiente muestra información sobre la variable X = “Renta mensual del hogar en miles de euros” para 2000 hogares de Galicia:

	A Coruña	Lugo	Ourense	Pontevedra
Renta media por hogar	1.984	1.892	1.707	1.899
Número de hogares	821	258	264	657

Calcula la renta familiar de los hogares de Galicia.

Medidas de posición. Mediana

Una vez ordenados los datos de menor a mayor, se define la mediana como el valor más pequeño de la variable que deja a su izquierda, como mínimo, la mitad de los valores de dicha variable.

Si la variable está agrupada en intervalos de clase, buscamos sobre la tabla de frecuencias el primer intervalo cuya frecuencia relativa acumulada es mayor o igual que $\frac{1}{2}$, la clase mediana, y dentro de ella se puede obtener la mediana por interpolación lineal, pues suponemos (véase el polígono de frecuencias acumuladas) que los datos se distribuyen de manera uniforme dentro del intervalo.

Ejemplo. Medidas de posición. Mediana. Número de horas extra

N.º de horas extra

x_i	n_i	f_i	N_i	F_i
0	5	$0'25$	5	$0'25$
1	8	$0'40$	13	$0'65$
2	4	$0'20$	17	$0'85$
3	2	$0'10$	19	$0'95$
4	1	$0'05$	20	1
SUMAS	20	1		

Por lo tanto, la mediana es 1 hora extra.

Media y mediana. Comparación

La media y la mediana tendrán valores similares, salvo cuando haya valores atípicos (valores extremados o raros) o cuando la distribución sea muy asimétrica.

Ejemplo: Consideremos las observaciones siguientes: 4; 1; 3; 2.

Su media es 2'5 y su mediana es 2'5 (una vez ordenados los datos: 1; 2; 3; 4).

Ejercicio Supongamos ahora que tenemos una observación más, 22, que podríamos considerarla como un dato atípico. Calcula la media y la mediana y discute los resultados obtenidos.

Medidas de posición. Moda

- Se denotado por Mo y el intervalo con mayor frecuencia será la **clase modal.**
- Es el valor de la variable que se presenta con mayor frecuencia.
  - Datos no agrupados: valor de la variable de mayor frecuencia absoluta o relativa.
  - Datos agrupados: se busca el intervalo modal (i), el de mayor densidad de datos. Entonces,

$$Mo = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} a_i$$

    - Si $i = 1$, $d_{i-1} = 0$. Por tanto, $Mo = L_i$.
    - Si $i =$ nº de clases, $d_{i+1} = 0$. Por tanto, $Mo = L_{i-1}$.
- Puede ocurrir que haya una única moda, en cuyo caso hablamos de distribución de frecuencias **unimodal.** Si hay más de una moda, diremos que la distribución es **multimodal.**

Ejemplo. Medidas de posición. Moda. Número de horas extra

N.º de horas extra

x_i	n_i	f_i	N_i	F_i
0	5	$0'25$	5	$0'25$
1	8	$0'40$	13	$0'65$
2	4	$0'20$	17	$0'85$
3	2	$0'10$	19	$0'95$
4	1	$0'05$	20	1
SUMAS	20	1		

Ejemplo. Medidas de posición. Moda. Edad

Intervalo de clase $(L_{i-1}, L_i]$	Marca de clase c_i	n_i	f_i	N_i	F_i	$c_i f_i$
$(20, 35]$	$27'5$	5	$0'5$	5	$0'5$	$13'75$
$(35, 50]$	$42'5$	1	$0'1$	6	$0'6$	$4'25$
$(50, 65]$	$57'5$	4	$0'4$	10	1	23
	SUMAS	20	1			

Medidas de posición. Cuantiles

- Hemos visto que la mediana divide a los datos en dos partes iguales. Pero también tiene interés estudiar otros parámetros, llamados cuantiles, que dividen los datos de la distribución en partes iguales, es decir en intervalos que comprenden el mismo número de valores.
- Sea $p \in (0, 1)$. Una vez ordenados los datos de menor a mayor, se define el **cuantil p**, como el valor más pequeño de la variable que deja a su izquierda np observaciones. Lo que es lo mismo, la frecuencia relativa acumulada hasta el **cuantil p** es mayor o igual que **p**. Nótese que la mediana es el **cuantil 0'5**. Los cuantiles, al igual que la mediana, sólo se podrán calcular con variables que admitan un orden.
- Algunos órdenes de los cuantiles tienen nombres específicos. Así los **cuartiles** son los cuantiles de orden (0.25, 0.5, 0.75) y se representan por Q_1, Q_2, Q_3. Los cuartiles dividen la distribución en cuatro partes. Los **deciles** son los cuantiles de orden (0.1, 0.2,..., 0.9). Los **percentiles** son los cuantiles de orden $j/100$ donde j=1,2,...,99.

Medidas de posición. Cuantiles. Cálculo

Si la variable es discreta, o si es continua y disponemos de todos los datos, empezamos ordenando la muestra.
El **cuantil p** es el menor dato de la muestra (primero de la muestra ordenada) cuya frecuencia relativa acumulada es mayor o igual que **p**.

Para **datos no agrupados** se busca la primera frecuencia acumulada tal que $N_i \geq pn$:

1. Si $N_i > pn$, entonces $x_p = x_i$.
2. Si $N_i = pn$, entonces $x_p = \dfrac{x_i + x_{i+1}}{2}$.

Si la variable es continua y se encuentra **agrupada** en intervalos de clase, buscamos el primer intervalo cuya frecuencia relativa acumulada es mayor o igual que **p**, que se corresponde con el valor

$$x_p = L_{i-1} + \frac{pn - N_{i-1}}{n_i} a_i = L_{i-1} + \frac{p - F_{i-1}}{f_i} a_i$$

Ejemplo. Medidas de posición. Cuantiles. Jornada laboral

A continuación figuran las duraciones de la jornada laboral de dieciocho individuos:

$$6'56,\ 6'53,\ 6'50,\ 6'74,\ 6'55,\ 6'58,\ 6'75,\ 6'60,\ 6'51,\ 6'44,\ 6'67,\ 6'46,\ 6'71,\ 6'37,\ 6'81,\ 6'39,\ 6'14,\ 6'66$$

Calcula la mediana, cuartiles y percentiles.

Ejemplo. Medidas de posición. Cuantiles. Jornada laboral

A continuación figuran las duraciones de la jornada laboral de dieciocho individuos:

$$6'56,\ 6'53,\ 6'50,\ 6'74,\ 6'55,\ 6'58,\ 6'75,\ 6'60,\ 6'51,\ 6'44,\ 6'67,\ 6'46,\ 6'71,\ 6'37,\ 6'81,\ 6'39,\ 6'14,\ 6'66$$

Calcula la mediana, cuartiles y percentiles. Lo primero que tenemos que hacer es ordenar los datos de menor a mayor:

$$6'14,\ 6'37,\ 6'39,\ 6'44,\ 6'46,\ 6'50,\ 6'51,\ 6'53,\ 6'55,\ 6'56,\ 6'58,\ 6'60,\ 6'66,\ 6'67,\ 6'71,\ 6'74,\ 6'75,\ 6'81$$

La *mediana* es $\mathbf{m} = 6'555$
El *primer cuartil:* $\mathbf{Q_1} = 6'46$
El *tercer cuartil:* $\mathbf{Q_3} = 6'67$
El *cuantil* $0'10$ es: $6'37$
El *cuantil* $0'40$ es: $6'53$
El *percentil* 90 es: $6'75$

Medidas de dispersión

- Recorrido o rango
- Recorrido intercuartílico:
- Varianza y desviación típica
- Cuasivarianza y cuasidesviación típica
- Coeficiente de variación

Medidas de dispersión

Las medidas de dispersión se utilizan para describir la variabilidad o esparcimiento de los datos de la muestra respecto a la posición central. A continuación describimos las más importantes:

- Recorrido o rango: $R = \max_i x_i - \min_i x_i$
- Recorrido intercuartílico: Diferencia entre el cuartil tercero y primero
- Varianza y desviación típica
- Cuasivarianza y cuasidesviación típica
- Coeficiente de variación

Medidas de dispersión. Recorrido y Recorrido intercuartílico

- Recorrido o rango: $R = \text{máx}_i\, x_i - \text{mín}_i\, x_i$
  En el ejemplo de la jornada laboral,
  recorrido $= 6'81 - 6'14 = 0'67$
- Recorrido intercuartílico: Diferencia entre el cuartil tercero y primero
  En el ejemplo de la jornada laboral,
  recorrido intercuartílico $= 6'67 - 6'46 = 0'21$

* llevan asociadas las unidades de medida.

Diagrama de caja (boxplot)

El diagrama de caja es una representación gráfica que se utiliza con variables continuas. Permite describir la dispersión y la simetría de la distribución de datos. El diagrama de caja está formado por:

- una caja delimitada por los cuartiles Q1 y Q3, y en cuyo interior se representa una línea horizontal a la altura de la mediana. Nótese que dentro de la caja se encontrará la mitad de las observaciones. Si la mediana no se encuentra en el centro de la caja, interpretamos que la distribución no es simétrica.
- una línea vertical desde el tercer cuartil hasta el valor mayor de la muestra que no sea un valor atípico,
- una línea vertical desde el primer cuartil hasta el valor menor de la muestra que no sea un valor atípico,
- círculos que representan los valores atípicos de la muestra.

Diagrama de caja (boxplot)

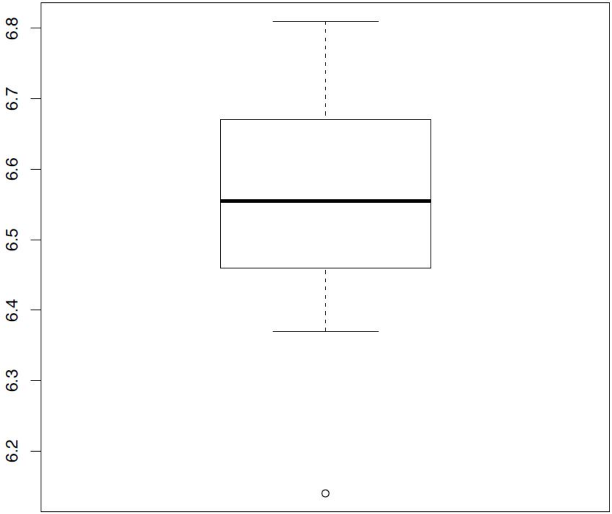
- Los segmentos horizontales inferior y superior (llamados bigotes whiskers) y que delimitan las líneas verticales discontinuas, como las que se muestran en el diagrama de caja de abajo, alcanzan a las últimas observaciones de la muestra que no son atípicas. Por tanto, el extremo inferior será la menor observación mayor o igual que $Q1 - 1,5 \cdot RIC$ y el extremo superior será la mayor observación menor o igual que $Q3 + 1,5 \cdot RIC$.

Nota Se considera que un dato x es **atípico** si está en alguna de estas dos circunstancias:

$$x < Q1 - 1'5 \cdot RIC \quad \text{o} \quad x > Q3 + 1'5 \cdot RIC$$

siendo RIC $=$ rango intercuartílico $= Q3 - Q1$.

Ejemplo. Diagrama de caja (boxplot). Jornada laboral



Medidas de dispersión. Varianza

La media se emplea como medida de posición. Entonces, parece razonable tomar como medida de dispersión algún criterio de discrepancia de los puntos respecto a la media.

Recuerda que la simple diferencia de los puntos a la media, al ponderarla, da cero. Por tanto, elevamos esas diferencias al cuadrado para que no se cancelen los sumandos positivos con los negativos. El resultado es la **varianza**:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \,; \qquad S^2 = \sum_{i}^{k}(x_i \text{ ó } c_i - \bar{x})^2 f_i$$

Propiedades

1.- $S^2_{a+X} = S^2_X$. La varianza no se ve afectada por cambios de localización.
2.- $S^2_{b\cdot X} = b^2 \cdot S^2_X$. La varianza se mide en el cuadrado de la escala de la variable.
Que una medida de dispersión no se vea afectada por cambios de localización, como ocurre con la varianza (propiedad 1), es una condición casi indispensable para admitirla como tal medida de dispersión. La dispersión de un conjunto de datos no se ve alterada por una mera traslación de los mismos.

La propiedad 2 nos da pie a calcular la raíz cuadrada de la varianza, obteniendo así una medida de dispersión que se expresa en la mismas unidades de la variable. Esta medida es la **desviación típica**, o *desviación estándar*, que en coherencia denotamos por S.

Ejemplo. Medidas de dispersión. Nº de horas extra

Calculemos la varianza y la desviación típica de la variable n.º de horas extra:

N.º de horas extra	n_i	f_i	N_i	F_i	$x_i f_i$	$(x_i - \bar{x})^2 f_i$
0	5	0,25	5	0,25	0	0,4225
1	8	0,40	13	0,65	0,40	0,0360
2	4	0,20	17	0,85	0,40	0,0980
3	2	0,10	19	0,95	0,30	0,2890
4	1	0,05	20	1	0,20	0,3645
SUMAS	20	1			1,3	1,21

Observemos que en el cálculo de la varianza y desviación típica necesitamos calcular previamente la media de la variable ($\bar{x} = 1{,}3$).
Por lo tanto, la varianza de la variable n.º de horas extra es:

$$S^2 = \sum_{i=1}^{k} (x_i - \bar{x})^2 f_i = \sum_{i=1}^{k} (x_i - 1{,}3)^2 f_i = 1{,}21$$

Y la desviación típica:
$S = \sqrt{1{,}21} = 1{,}1$ (aproximadamente, una hora extra)

Ejemplo. Medidas de dispersión. Edad

Calculemos la varianza y la desviación típica de la variable edad a partir de la tabla de frecuencias donde los datos han sido agrupados en intervalos de clase:

Intervalo de clase $(L_{i-1}, L_i]$	Marca de clase c_i	n_i	f_i	N_i	F_i	$c_i f_i$	$(c_i - \bar{x})^2 f_i$
$(20, 35]$	27,5	5	0,5	5	0,5	13,75	91,125
$(35, 50]$	42,5	1	0,1	6	0,6	4,25	0,225
$(50, 65]$	57,5	4	0,4	10	1	23	108,900
	SUMAS	20	1			41	200,25

La varianza y la desviación típica son, respectivamente
$S^2 = \sum_{i=1}^{k}(c_i - \bar{x})^2 f_i = \sum_{i=1}^{k}(c_i - 41)^2 f_i = 200{,}25$ y $S = \sqrt{200{,}25} \simeq 14{,}15$ años

• Para calcular la desviación típica con todos los datos:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 40 \text{ años}$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{10}\left[(52-40)^2 + (47-40)^2 + \cdots + (23-40)^2\right] = 201{,}8$$

$$S = \sqrt{201{,}8} \simeq 14{,}21 \text{ años}$$

Es muy habitual modificar ligeramente el cálculo de la varianza, dividiendo por $(n-1)$ en lugar de por n. De este modo obtenemos lo que se conoce como **cuasivarianza**:

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

Conociendo la varianza se puede calcular la cuasivarianza, y viceversa, pues $S_c^2 = n \cdot S^2/(n-1)$. Además, ambas medidas se expresan en la unidades de la variable al cuadrado, y presentan el mismo comportamiento frente a cambios de localización y escala.

La **cuasidesviación típica** es simplemente la raíz cuadrada de la cuasivarianza, y por tanto la denotamos por S_c.

Ejemplo. Medidas de dispersión. Cuasivarianza y cuasidesviación típica. Nº de horas extra. Edad

Nº horas extra

$S_c^2 = \dfrac{n}{n-1} S^2 = \dfrac{20}{19} \cdot 1{,}21 \simeq 1{,}27$

$S_c = \sqrt{1{,}27} \simeq 1{,}13$ (aproximadamente una hora extra)

Edad

$S_c^2 = \dfrac{n}{n-1} S^2 = \dfrac{10}{9} \cdot 201{,}8 \simeq 224{,}22$

$S_c = \sqrt{224{,}22} \simeq 14{,}97$ años

Hay situaciones en las que tenemos que comparar poblaciones en las que

- las unidades de medida son distintas

Ejemplo:

Peso de hormigas en gramos: ($s = 2{,}41$ gramos)

8.180881	10.503650	8.210198	13.096271	9.259044
15.540982	7.854185	12.010111	8.725924	11.712810

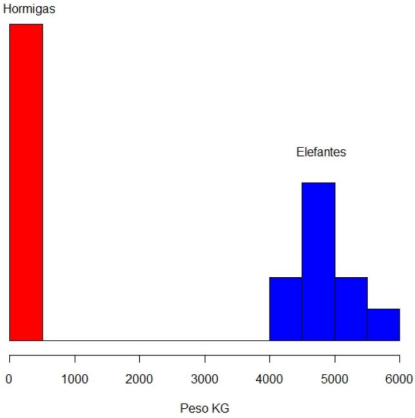
Peso de elefantes en kg: ($s = 320{,}0495$ kilos)

5100.636	4987.702	5035.441	5321.591	5502.833
4737.402	4537.105	4731.434	4742.981	4444.282

Medidas de dispersión. Coeficiente de variación

Hay situaciones en las que tenemos que comparar poblaciones en las que

- o que aún teniendo la misma unidad de medida difieren en sus magnitudes.



Para estos casos necesitamos una medida de la dispersión en la que no influyan las unidades, sería conveniente tener una medida adimensional.

Medidas de dispersión. Coeficiente de variación

Si queremos una medida de dispersión que no dependa de la escala y que, por tanto, permita una comparación de las dispersiones relativas de varias muestras, existen varias propuestas, pero nos quedamos con el **coeficiente de variación**, que se define así:

$$CV = \frac{S}{|\bar{x}|}$$

Nº de horas extra

$CV = \dfrac{1'1}{1'3} \simeq 0'846$ (84'6 %)

Edad

$CV = \dfrac{14'21}{40} \simeq 0'356$ (35'6 %)

Medidas de forma

- Coeficiente de asimetría de Fisher
- Coeficiente de apuntamiento o curtosis

Medidas de forma: asimetría y apuntamiento

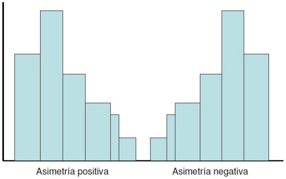
Coeficiente de asimetría de Fisher

Las medidas de forma se refieren, como su nombre indica, a la forma de la representación gráfica de los datos. Una de las medidas de forma trata de reflejar la simetría de los datos. Se define el coeficiente de asimetría como

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \overline{X})^3}{S_X^3}$$

Interpretación

- $g_1 > 0$: asimetría positiva o por la derecha.
- $g_1 < 0$: asimetría negativa o por la izquierda.
- $g_1 = 0$: la distribución es simétrica.



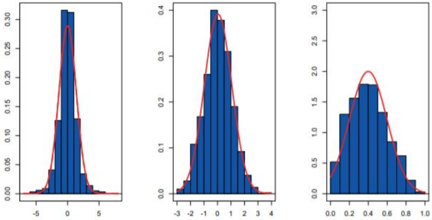
Coeficiente de apuntamiento o curtosis

Distribuciones simétricas pueden tener distinta forma dependiendo de como se repartan las frecuencias entre el centro y los extremos. Las medidas de apuntamiento se basan en la comparación de este valor con el de una distribución normal.

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \overline{X})^4}{S_X^4} - 3$$

Interpretación

- $g_2 > 0$: más apuntamiento que la distribución normal (leptocúrtica).
- $g_2 = 0$: apuntamiento equivalente a la distribución normal (mesocúrtica).
- $g_2 < 0$: menos apuntamiento que la distribución normal (platicúrtica).



Distribuciones leptocúrtica, mesocúrtica y platicúrtica