Introducción a la Estadística y Probabilidad

Tema 2. Descripción de datos

MANUEL MONGE, Ph.D.

Departamento de Economía Aplicada y Métodos Cuantitativos
Facultad de Derecho, Economía y Gobierno
Universidad Francisco de Vitoria

Contenido

1. Contenidos de este tema

1. Contenidos de este tema

- Distribuciones de frecuencias: atributos y variables cuantitativas.
- Representación gráfica: atributos y variables cuantitativas.
- Tratamiento estadístico
- Medidas de posición: medias, mediana, moda y cuantiles.
- Medidas de dispersión: absolutas y relativas.
- Medidas de forma: asimetría y curtosis.
- Variable tipificada.

2. Tabla de frecuencias

2. Tabla de frecuencias

Para organizar y resumir datos, las **tablas de frecuencias** son una herramienta útil.

Si la variable de estudio es una **variable cualitativa** o **cuantitativa discreta** y los datos se recogen a partir de una muestra de tamaño n, la tabla tiene:

- Una primera columna (llamada **clases** o **grupos**) que contiene todas las posibles respuestas de dicha variable.
- A continuación, se le añade la columna de **frecuencia absoluta (f_i)** que contiene el número de observaciones correspondientes a cada clase.
- La siguiente columna recoge las **frecuencias absolutas acumuladas (F_i)**, que se calculan como suma de las frecuencias absolutas de las clases anteriores.
- La siguiente columna contiene las **frecuencias relativas (h_i)**, que se calculan como cociente entre las frecuencias absolutas y el tamaño de muestra **($h_i = f_i/n$)**.
- A continuación, se añaden las **frecuencias relativas acumuladas (H_i)**, calculadas como la suma de las frecuencias relativas de las clases anteriores.
- Por último, la tabla concluye con las frecuencias relativas y relativas acumuladas en tantos por cien ($h_i\,\%$ y $H_i\,\%$).

2. Tabla de frecuencias

Tabla 1. Tabla de Frecuencias para variable cualitativa o cuantitativa discreta

Clases	f_i	F_i	h_i	H_i	$h_i\,\%$	$H_i\,\%$
		n		1		100
Totales	n		1		100	

Ejemplo - Para variable cualitativa o cuantitativa discreta

Se preguntó a un total de 2.000 viajeros frecuentes (de negocios) qué ciudad de la región central de Estados Unidos preferirían: Indianápolis, St. Louis, Chicago o Milwaukee. De ellos, 100 contestaron que Indianápolis; 450, St. Louis; 1.300, Chicago; y el resto dijo que Milwaukee. Elabore una tabla de frecuencias para resumir esta información.

Solución del ejemplo

Clases	f_i	F_i	h_i	H_i	$h_i\,\%$	$H_i\,\%$
Indianápolis	100	100	0.05	0.05	5 %	5 %
St. Louis	450	550	0.225	0.275	22.5 %	27.5 %
Chicago	1300	1850	0.65	0.925	65 %	92.5 %
Milwaukee	150	2000	0.075	1	7.5 %	100
Totales	2000		1		100	

2. Tabla de frecuencias

Si la variable de estudio es **cuantitativa continua** y los datos se recogen a partir de una muestra de tamaño n, para construir la tabla de frecuencias es necesario decidir en cuántos intervalos distribuiremos los datos (clases o grupos).

Para ello vamos a ver cómo definimos la **amplitud de los intervalos**.

2. Tabla de frecuencias

Amplitud de los intervalos

1. Hay que decidir el número, k, de intervalos (clases).
2. Los intervalos (clases) deben ser de la misma amplitud, w:

$$w = \textit{amplitud de los intervalos} = \frac{(\textit{núm.mayor} - \textit{núm.menor})}{\textit{núm.de intervalos}}$$

Tanto k como w deben redondearse al alza, posiblemente al siguiente número entero mayor.

3. Los intervalos (clases) deben ser inclusivos y no solaparse.

En definitiva, la tabla de frecuencias tendría una estructura similar a la de una variable cualitativa, pero con una columna más, que recogería el punto medio de los intervalos. Esta columna recibe el nombre de **marca de clase** (x_i).

2. Tabla de frecuencias

Tabla 2. Tabla de Frecuencias para variable cuantitativa continua

Clases	x_i	f_i	F_i	h_i	H_i	$h_i\,\%$	$H_i\,\%$
			n		1		100
Totales		n		1		100	

Ejemplo - Para variable cuantitativa continua

Se desea estudiar el índice de obesidad de los jóvenes de entre 18 y 25 años. Para ello se extrae una muestra representativa de la población de tamaño 500 y se anota el peso de cada individuo.

Individuo	1	2	3	4	...	499	500
Peso (kg)	50.1	62.3	58.1	49.5	...	72.4	103.3

Solución del ejemplo

Vamos a construir la **tabla de frecuencias para variables cuantitativas continuas** con los datos anteriores.

- Vamos a decidir el número de intervalos. Para ello utilizaremos el término **amplitud**.
- La variable de estudio es una variable cuantitativa continua. Como el tamaño muestral es n = 500, deberíamos distribuir los datos entre 8 y 10 intervalos.
- En nuestro caso vamos a considerar 10 intervalos. Imaginemos que la persona con menor peso ha sido un individuo con 42,7 kg y la persona con mayor peso pesa 112,5 kg. Redondeando, consideramos que el rango de peso oscila entre 40 y 120 kg.

Así, la amplitud de los intervalos sería:

$$w = \frac{120-40}{10} = 8$$

2. Distribución de frecuencias

La tabla de frecuencias quedaría del siguiente modo:

Peso (kg)	x_i	f_i	F_i	h_i	H_i	$h_i\,\%$	$H_i\,\%$
[40, 48[	44	26	26	0.052	0.052	5.2 %	5.2 %
[48, 56[	52	200	226	0.400	0.452	40 %	45.2 %
[56, 64[	60	150	376	0.300	0.752	30 %	75.2 %
[64, 72[	68	50	426	0.100	0.852	10 %	85.2 %
[72, 80[	76	20	446	0.040	0.892	4 %	89.2 %
[80, 88[	84	21	467	0.042	0.934	4.2 %	93.4 %
[88, 96[	92	20	487	0.040	0.974	4 %	97.4 %
[96, 104[	100	10	497	0.020	0.994	2 %	99.4 %
[104, 112[	108	0	497	0	0.994	0 %	99.4 %
[112, 120[	116	3	500	0.006	1	0.6 %	100 %
Total		500		1		100	

3. Representación gráfica

3. Representación gráfica

Una **representación gráfica** nos permite ver de una manera rápida los resultados obtenidos. Es una herramienta para **representar la distribución de frecuencias**.

3.1. Representación gráfica para describir variables cualitativas y cuantitativas discretas

DIAGRAMA DE BARRAS

El diagrama de barras es utilizado cuando el objetivo es llamar la atención sobre la frecuencia de cada categoría, donde la altura de cada rectángulo representa la frecuencia.

3.1. Representación gráfica para describir variables cualitativas y cuantitativas discretas

Ejemplo de Diagrama de Barras

La revista Forbes enumera anualmente las principales empresas del mundo. La siguiente tabla muestra la distribución de frecuencias de las cinco empresas que tienen mayor capitalización del mundo en 2018.

Empresas	Capitalización de mercado f_i	H_i %
ICBC	311.01	0.19
China Construction Bank	261.17	0.16
JP Morgan Chase	387.67	0.24
Berkshire Hathaway	491.89	0.30
Agricultural Bank of China	184.13	0.11

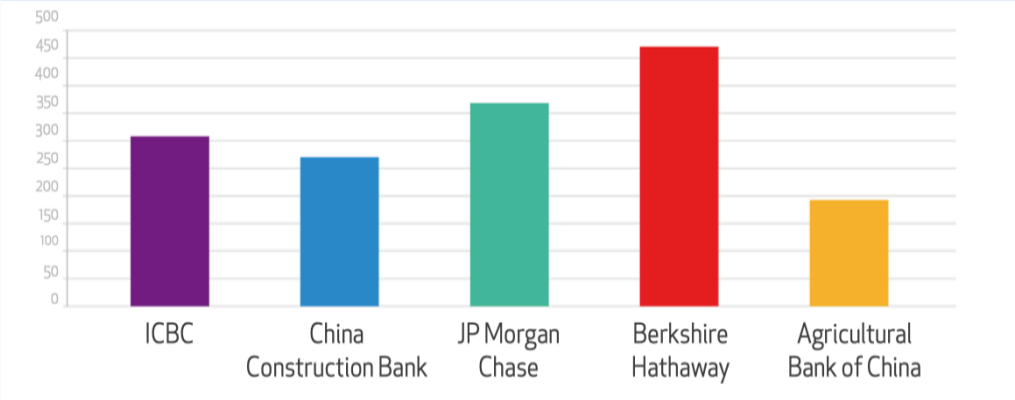


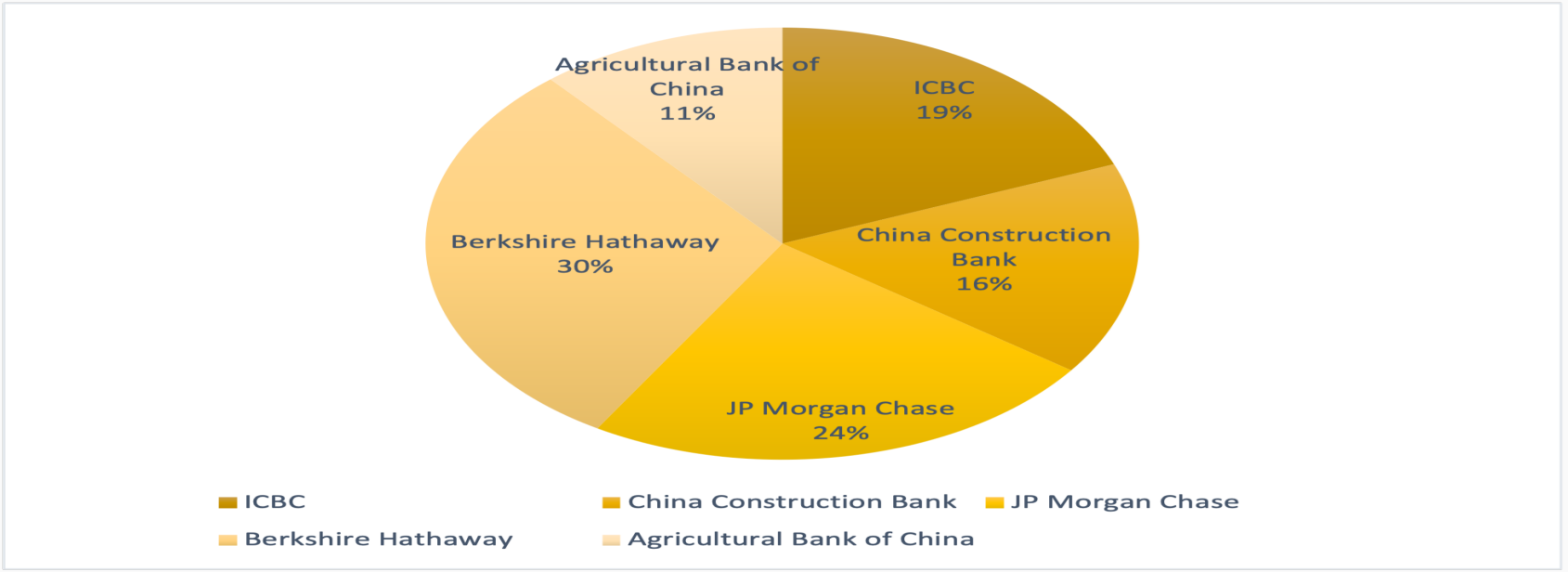
DIAGRAMA DE SECTORES o GRÁFICO DE TARTA

El diagrama de sectores o gráfico de tarta trata de visualizar la proporción de las frecuencias de cada categoría. En este caso, el círculo representa la proporción total y los sectores que lo parten tienen un área proporcional a la frecuencia de cada categoría.

3.1. Representación gráfica para describir variables cualitativas y cuantitativas discretas

Ejemplo diagrama de sectores o gráfico de tarta

El siguiente gráfico representa la información del ejemplo anterior:



3.2. Representación gráfica para describir variables continuas

HISTOGRAMA

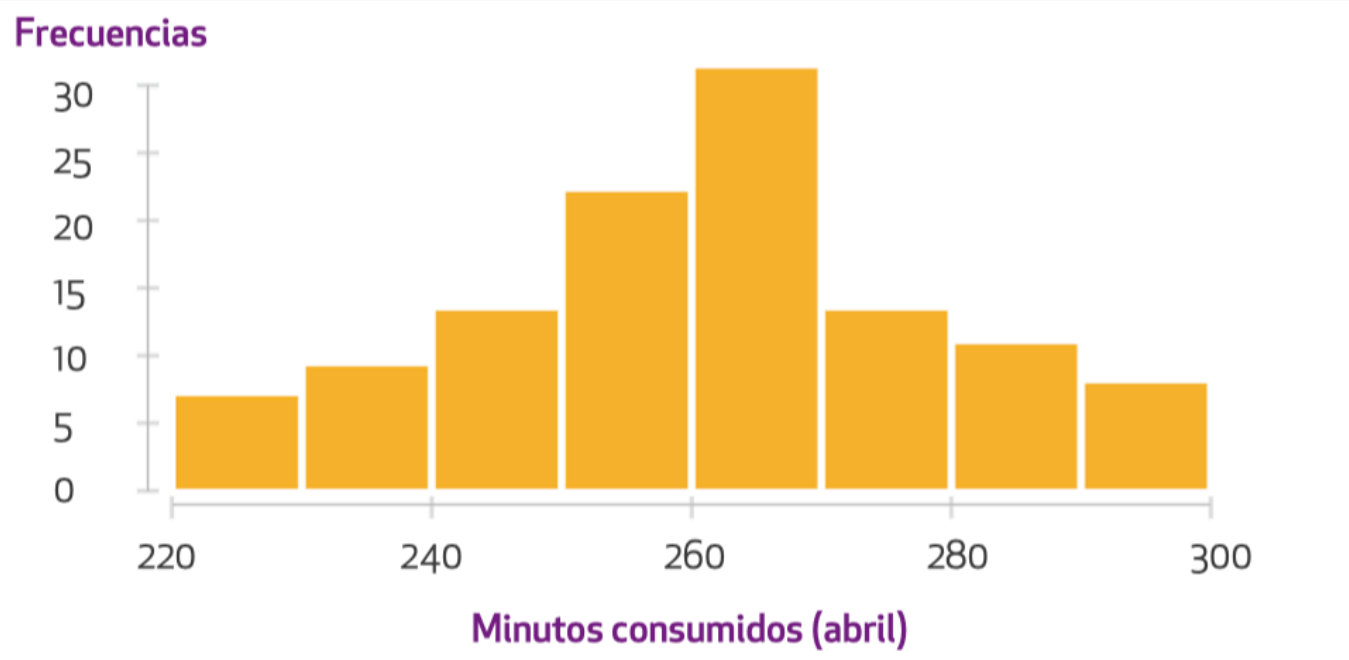
- El **histograma** es un gráfico formado por un **conjunto de rectángulos** (que reciben el nombre de **intervalos de clase**), donde cada uno representa un intervalo de agrupación o clase. Los **intervalos** corresponden a los construidos en una tabla de frecuencias, donde la altura de cada barra es proporcional al número de observaciones que hay en ese intervalo.
- El diagrama de barras está pensado, sobre todo, para variables ordinales, mientras que el histograma está concebido para **variables que siguen una escala numérica de razón** (cuantitativas, idealmente continuas).
- Lo importante de un histograma son las áreas de los rectángulos, no sus alturas.

Ejemplo de Histograma

Uso del teléfono móvil (en minutos)	f_i	H_i %
[220, 230[	5	4.5
[230, 240[	8	11.8
[240, 250[	13	23.6
[250, 260[	22	43.6
[260, 270[	32	72.7
[270, 280[	13	84.5
[280, 290[	10	93.6
[290, 300[	7	100

3.2. Representación gráfica para describir variables continuas

Ejemplo de Histograma



Interpretar un Histograma

- Para interpretar el **significado de los rectángulos**, vamos a fijarnos detenidamente en la figura de la diapositiva anterior.
- Si nos preguntaran cuántas personas de nuestra muestra hablan entre 220 y 230 minutos, nosotros podríamos responder que un 5 %.
- Si nos preguntaran cuántas personas de nuestra muestra hablan entre 260 y 270 minutos, responderíamos que aproximadamente un 32 %, siendo esta última la más frecuente.

3.2. Representación gráfica para describir variables continuas

Representación gráfica. Otros gráficos. Tallos-hojas

- Un diagrama de tallo y hojas (stem and leaf en inglés) es un gráfico alternativo al histograma que **combina la representación gráfica con la información proporcionada por las cifras**.
- En este gráfico, los datos se agrupan de acuerdo con sus primeros dígitos, llamados **tallos**, y se hace un listado de los últimos dígitos, llamados **hojas**, de cada miembro de su clase. Las hojas se muestran individualmente en orden ascendente después de cada uno de los tallos.
- La ventaja es que el rectángulo está relleno de los propios valores numéricos, pero **se evita la repetición de los primeros dígitos de cada cifra**. Se puede elegir su amplitud, aunque siempre es preferible que las amplitudes sean de 5 o de 10 unidades.
- De un vistazo aparece el histograma. No hay más que girar la figura mentalmente 90 grados hacia la izquierda.

Ejemplo Tallos-hojas

Construyamos un diagrama de tallo y hojas de las horas que dedican 20 estudiantes a estudiar para un examen de estadística:

3.5	2.8	4.5	6.2	4.8	2.3	2.6	3.9	4.4	5.5
5.2	6.7	3.0	2.4	5.0	3.6	2.9	1.0	2.8	3.6

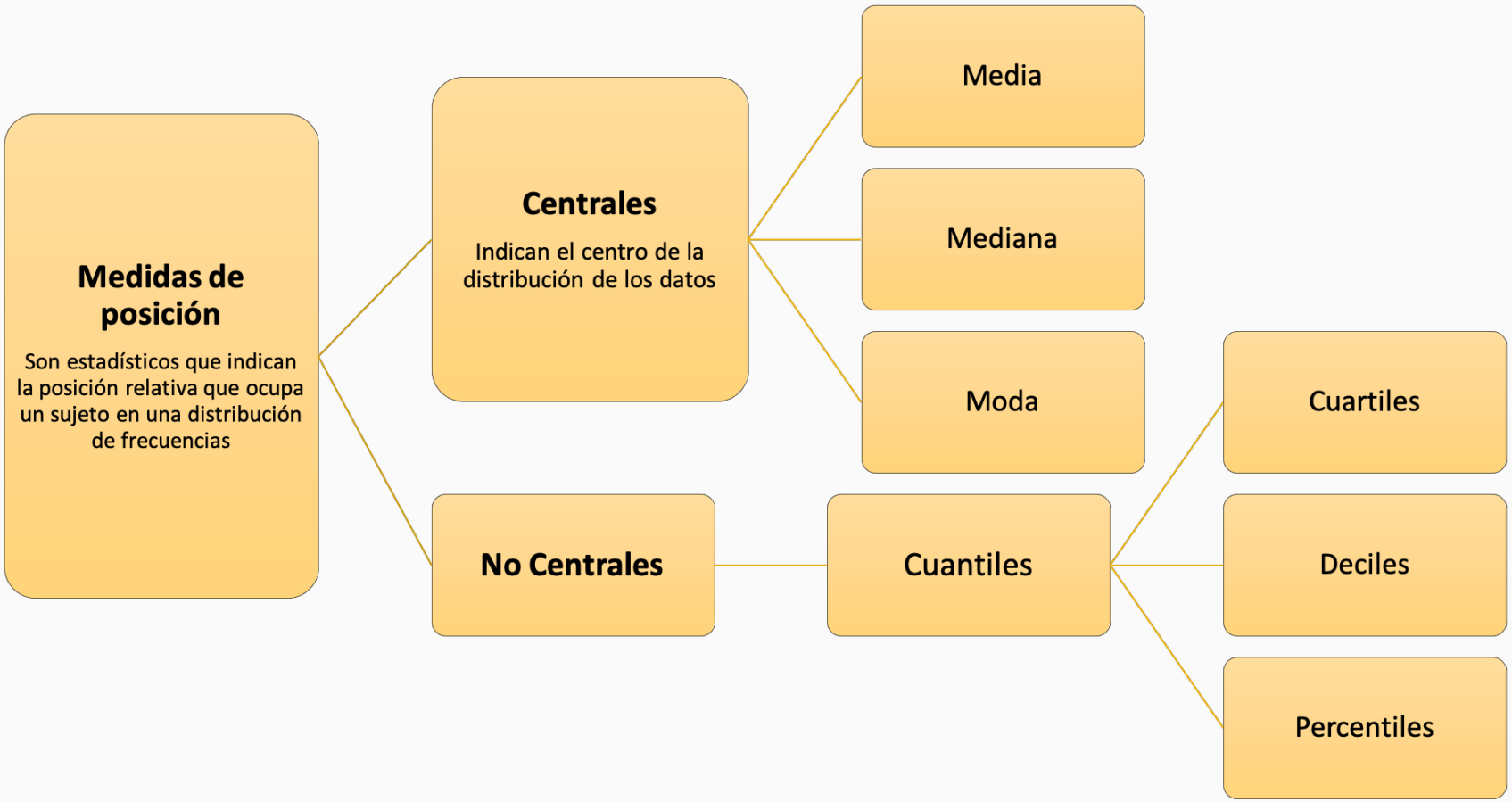
Solución ejemplo Tallos-hojas

Para construir el diagrama, es necesario ordenar los datos de menor a mayor:

1|0
2|346889
3|05669
4|458
5|025
6|27

4. Medidas de posición

4. Medidas de posición



4.1. Medidas de centralización

Para obtener **información numérica sobre una observación** típica de los datos, utilizamos medidas de centralización. En este apartado analizamos la media, la moda y la simetría de los datos.

Media

Dado un conjunto de datos numéricos $(x_1, ..., x_n)$, se define la media aritmética por

$$\overline{x} = \frac{x_1 + ... + x_n}{n} = \frac{\sum x_i}{n}$$

donde el símbolo $\sum$, que se denomina sumatorio, quiere decir que debemos sumar todos los valores de la variable.

Para datos discretos agrupados (por ejemplo tabla de la diapositiva 7), el cálculo de la media se efectúa teniendo en cuenta los valores distintos de la variable (x_j) y sus frecuencias relativas $(fr(x_j))$:

$$\overline{x} = \frac{\sum x_j \cdot fr(x_j)}{n}$$

Mediana

Ordenada la distribución, de menor a mayor, la mediana es el valore que divide los datos en dos partes iguales (en cuanto al número de observaciones se refiere), es decir, deja a ambos lados de la distribución el mismo número de frecuencias. (si hay un número par de datos entonces se hace la media aritmética entre los dos valores centrales). Viene expresada en las mismas unidades que la variable de estudio.

Moda

La moda, si existe, es el valor que aparece con más frecuencia.

4.2. Medidas de posición no central

Cuantiles

Son aquellos valores de la variable, que ordenados de menor a mayor, dividen a la distribución en partes, de tal manera que cada una de ellas contiene el mismo número de frecuencias. Se dividen en:

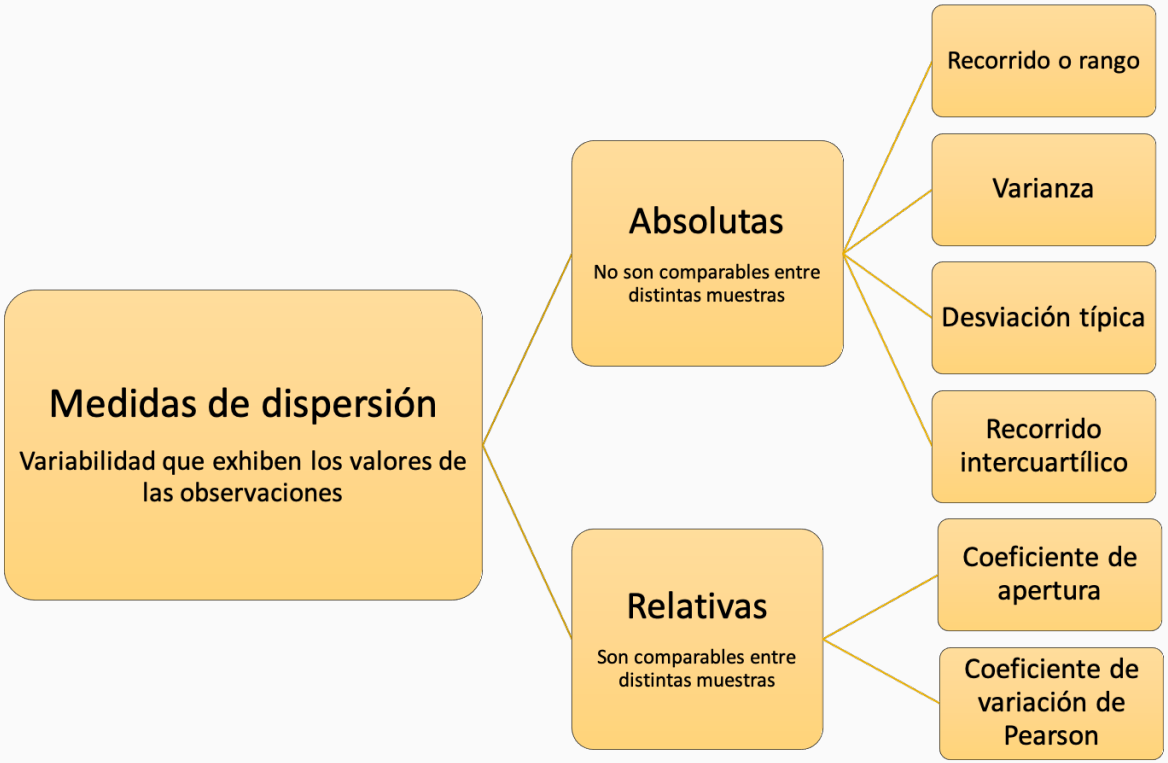
- **Cuartiles**: Dividen a la distribución en cuatro partes iguales.
- **Deciles**: Dividen a la distribución en 10 partes iguales.
- **Percentiles**: Dividen a la distribución en 100 partes iguales.

5. Medidas de dispersión

5. Medidas de dispersión

La media no es por sí sola una descripción completa o suficiente de los datos. En este apartado veremos descriptivos que miden la variabilidad o dispersión de las observaciones con respecto a la media.

5. Medidas de dispersión



Rango

El rango es la diferencia entre la observación mayor y la menor.

$$\textbf{Rango} = \textbf{Valor máximo - Valor mínimo} = X_{máx} - X_{min}$$

Rango intercuartílico (RIC)

El rango intercuartílico (RIC) mide la dispersión que hay en el 50 por ciento central de los datos; es la diferencia entre las observación Q_3 (**tercer cuartil** o **percentil 75**) y la observación Q_1 (**primer cuartil** o **percentil 25**).

$$RIC = Q_3 - Q_1$$

donde Q_3 y Q_1 se encuentran situados en la posición $0{,}75(n+1)$ y $0{,}25(n+1)$, respectivamente, cuando los datos se encuentran ordenados en sentido ascendente.

5. Medidas de dispersión

Varianza[1]

Definimos **varianza** como la media aritmética de las desviaciones de la media elevada al cuadrado.

$$Varianza = \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Desviación estándar poblacional

La desviación estándar poblacional describe la dispersión media en torno a la media.

$$Desviación\ estándar = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

[1] Los artículos en revistas científicas suelen utilizar la desviación estándar poblacional (σ) y la varianza poblacional (σ^2), dejando a un lado la desviación estándar muestral (s) y la varianza muestral (s^2).

Coeficiente de variación (CV)

Es una medida de la dispersión relativa que expresa la desviación típica en porcentaje respecto a la media (siempre que la media sea positiva).

$$CV = \frac{\sigma}{\mu} x 100\,\% \text{ si } \mu > 0$$

Con el coeficiente de variación, podemos comparar la dispersión de dos variables que se miden en distintas escalas.

Ejemplo

Ejemplo

Se considera una muestra aleatoria de ocho empresas tailandesas. Los beneficios por acción de cada empresa han experimentado este año las siguientes variaciones porcentuales en comparación con el año pasado: 0 %, 0 %, 8.1 %, 13.6 %, 19.4 %, 20.7 %, 10 % y 14.2 %. Realicemos un **análisis descriptivo** de los datos.

Ejemplo

Solución del Ejemplo

La **variación porcentual media** de los beneficios por acción de esta muestra es la siguiente:

$$\overline{x} = \frac{0+0+8{,}1+13{,}6+19{,}4+20{,}7+10+14{,}2}{8} = 10{,}75 \approx 10{,}75\,\%$$

Para calcular la **variación porcentual mediana** de los beneficios por acción, se ordenan los valores de forma ascendente:

$$0\,\%,\ 0\,\%,\ 8.1\,\%,\ 10\,\%,\ 13.6\,\%,\ 14.2\,\%,\ 19.4\,\%,\ 20.7\,\%$$

$$Me = \frac{10+13{,}6}{2} = 11{,}8\,\%$$

Podemos fijarnos que la **tasa porcentual modal** ($\text{Mo} = 0\,\%$) no es un buen representante del centro de estos datos.

Ahora calculamos la **varianza** y la **desviación típica poblacional**:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} = 66{,}15$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}} = 8{,}13$$

6. Medidas de Asimetría y curtosis

6.1. Medidas de Asimetría y curtosis

Estas medidas informan sobre dos aspectos importantes de la forma de la distribución:

- Por un lado, miden su grado de asimetría.
- Por otro, su grado de homogeneidad.

Al ser medidas de forma, no dependen de las unidades de medida de los datos.

6.1. Coeficiente de Asimetría

Coeficiente de asimetría

Se expresa de la siguiente forma:

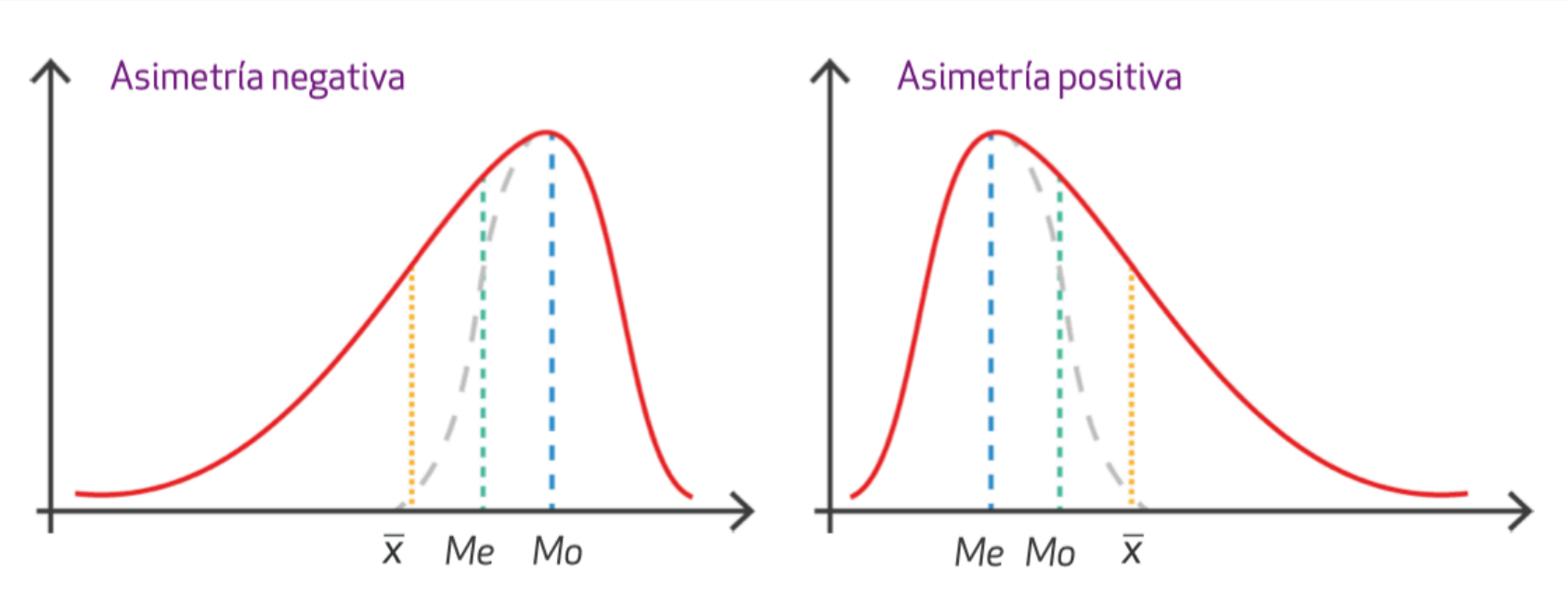
$$CA = \frac{\sum (x_i - \overline{x})^3}{n\sigma^3}$$

Para saber la forma de la distribución, debemos fijarnos en el signo del coeficiente de asimetría. En este sentido, las distribuciones pueden ser simétricas o asimétricas:

- Son **simétricas** cuando las dos colas de su histograma (derecha e izquierda) tienen la misma longitud.
- Son **asimétricas** cuando las colas tienen distinta longitud. En el caso de la asimetría, podemos encontrar que puede ser positiva o negativa:
  - Si el coeficiente (sesgo) es negativo, la distribución para valores inferiores a la media se alarga. En este caso, la cola de la izquierda será más larga.
  - Si el coeficiente (sesgo) es positivo, la cola de la distribución se extiende para valores superiores a la media. En este caso, la cola de la derecha es más prolongada.

6.1. Coeficiente de Asimetría

Los conceptos de la diapositiva anterior se ilustran en la siguiente figura:



6.1. Coeficiente de Asimetría

Otra medida de asimetría poco utilizada es la siguiente (también adimensional):

$$\frac{\overline{x} - mediana}{\sigma}$$

Para concluir...

Lo ideal para muchos procedimientos estadísticos es que la asimetría no sea grande y que el coeficiente de asimetría esté lo más próximo posible a 0.

6.2. Coeficiente de Curtosis

Coeficiente de curtosis

Para hallar el coeficiente de **curtosis**, utilizaremos la siguiente fórmula:

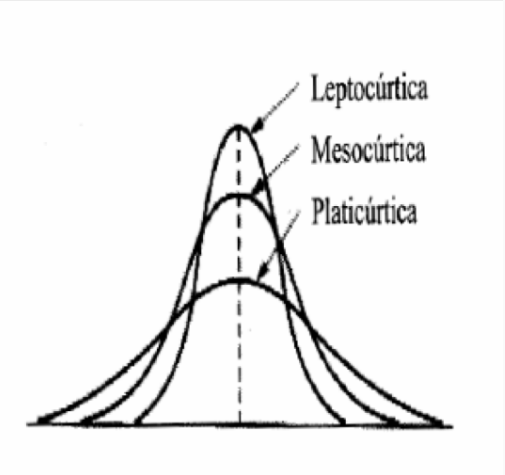
$$CC = \frac{\sum (x_i - \overline{x})^4}{n\sigma^4}$$

El coeficiente de curtosis es siempre mayor o igual que 1. Este coeficiente es importante porque nos informa respecto a la heterogeneidad de la distribución.

6.2. Coeficiente de Curtosis

Según el resultado obtenido mediante el coeficiente de curtosis, las formas que puede adoptar la curva son:

- **Mesocúrtica**: grado de concentración medio alrededor de los valores centrales de la variable.
- **Leptocúrtica**: grado de concentración elevado.
- **Platicúrtica**: grado de concentración reducido.



7. Diagrama de cajas

7. Diagrama de cajas

Diagrama de caja

Un **diagrama de caja** es una representación gráfica basada en cuartiles que que ayuda a presentar un conjunto de datos. Para construir un diagrama de caja solo se necesitan cinco estadísticos, que son valor mínimo, primer cuartil (Q_1), mediana, tercer cuartil (Q_3) y valor máximo.

7. Diagrama de cajas

Ejemplo

Alexander's Pizza ofrece entregas gratuitas de pizza a 24km a la rotonda. Alex, el propietario, desea información relacionada con el tiempo de entrega. ¿Cuánto tarda una entrega típica? ¿En qué margen de tiempo deben completarse la mayoría de las entregas? Alex recopiló la siguiente información de una muestra de 20 entregas:

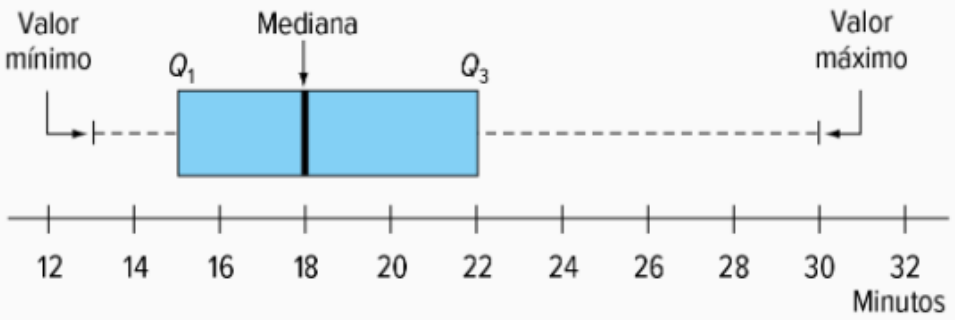
- Valor mínimo $= 13$ minutos.
- $Q_1 = 15$ minutos.
- Mediana $= 18$ minutos.
- $Q_3 = 22$ minutos.
- Valor máximo $= 30$ minutos.

Elabore un diagrama de caja de los tiempos de entrega. ¿Qué conclusiones tiene acerca de estos?

7. Diagrama de cajas

Solución del Ejemplo

- El primer paso para elaborar un diagrama de caja consiste en crear una escala adecuada a lo largo del eje horizontal.
- Luego, se debe dibujar una caja que inicie en Q_1 (15 minutos) y termine en Q_3 (22 minutos).
- Dentro de la caja trace una línea vertical para representar a la mediana (18 minutos).
- Por último, prolongue líneas horizontales a partir de la caja dirigidas al valor mínimo (13 minutos) y al valor máximo (30 minutos). Estas líneas horizontales que salen de la caja, a veces reciben el nombre de **bigotes**, en virtud de su virtud de su parecido a los bigotes de un gato.



Continuación solución del Ejemplo...

- El diagrama de caja también revela que la distribución de los tiempos de entrega tiene un sesgo positivo[2].
- ¿Cómo saber que esta distribución tiene un sesgo positivo? En este caso, hay dos tipos de información que lo sugieren.
- Primero, la línea punteada a la derecha de la caja, que va de 22 minutos (Q_3) al tiempo máximo de 30 minutos, es más larga que la línea punteada a la izquierda, que va de 15 minutos (Q_1) al valor mínimo de 13 minutos. En otras palabras, 25 % de los datos mayores que el tercer cuartil se encuentra más disperso que 25 % que es menor al primer cuartil.
- Segundo, la mediana no se encuentra en el centro de la caja. La distancia del primer cuartil a la mediana es menor que la distancia de la mediana al tercer cuartil. El número de tiempos de entrega entre 15 y 18 minutos es el mismo que el número de tiempos de entrega entre 18 y 22 minutos.

[2]Recuerde que el sesgo se define como la falta de simetría en un conjunto de datos.