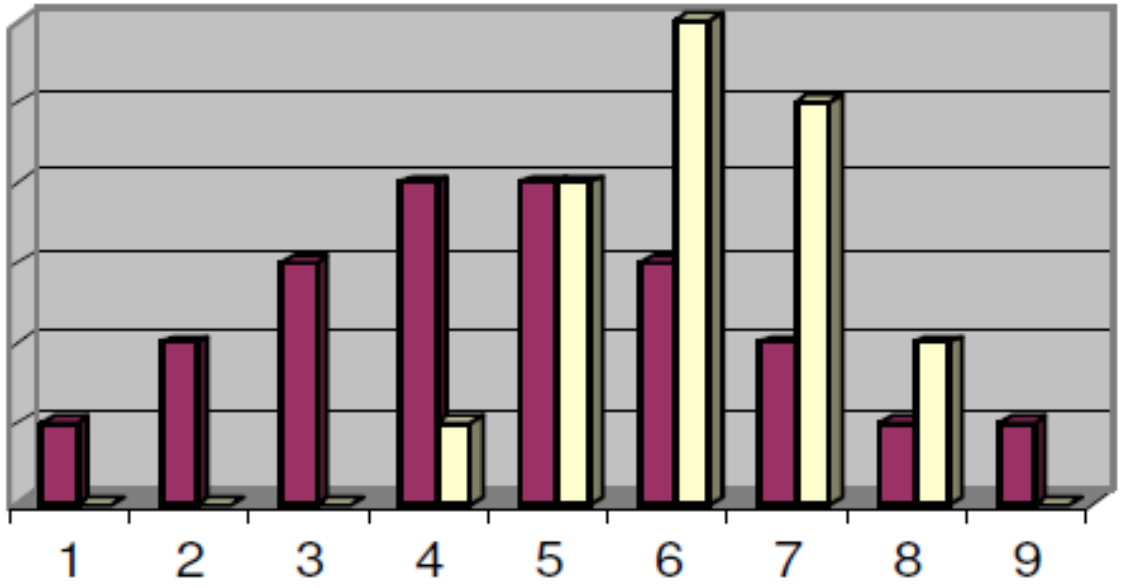
Clase 2: Estadística

Los datos

- Todo conjunto de datos tiene al menos dos características principales:

CENTRO Y DISPERSIÓN

- Los gráficos de barra, histogramas, de puntos, entre otros, nos dan cierta idea sobre ellos.



Estadísticos

- Los estadísticos son resúmenes de los datos muestrales. Describen una distribución según como se comporta el centro, su dispersión y su forma. Se agrupan en estadísticos de:

Tendencia central

Posición

Dispersión

Forma

- **Estadísticos de tendencia central**: Se ubican al centro de la distribución de los datos.
- Media aritmética (centro de gravedad de los datos)
- Moda (valor de la variable con mayor frecuencia)
- Mediana (valor central en el 50%)

Formato de una Tabla de Frecuencias

Valor (x_i)	n_i	N_i	f_i	F_i
x_1	n_1	$N_1 = n_1$	$f_1 = n_1/N$	$F_1 = f_1$
x_2	n_2	$N_2 = N_1+n_2$	$f_2 = n_2/N$	$F_2 = F_1+f_2$
x_3	n_3	$N_3 = N_2+n_3$	f_3	$F_3 = F_2+f_3$
⋮	⋮	⋮	⋮	⋮
x_k	n_k	$N_k = N$	$f_k = n_2/N$	$F_k = 1$
Total	N	——	1	——

donde n_i es la frecuencia absoluta, N_i frecuencia acumulada, f_i frecuencia relativa y F_i frecuencia relativa acumulada de la i-ésima categoría (clase), respectivamente.

…Formato de una Tabla de Frecuencias

Intervalo	Centro y_i	n_i	N_i	f_i	F_i
$(a_0,a_1]$	$y_1 = ½(a_0+a_1)$	n_1	N_1	f_1	F_1
$(a_1,a_2]$	$y_2 = ½(a_1+a_2)$	n_2	N_2	f_2	F_2
$(a_2,a_3]$	y_3	n_3	N_3	f_3	F_3
⋮	⋮	⋮	⋮	⋮	
$(a_k,a_{k+1}]$	y_k	n_k	$N_k = N$	f_k	$F_k = 1$
Total	——	N	——	1	——

donde y_i es la marca (punto medio) de la i-ésima clase.

Media aritmética

- **En datos sin tabular:**

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

donde x_i es el i-ésimo dato y n es el tamaño de la muestra.

- **En datos tabulados:**

$$\overline{x} = \frac{\sum_{i=1}^{k} y_i n_i}{n}$$

donde y_i es la marca de la i-ésima clase (o categoría), n_i la frecuencia absoluta de la i-ésima clase y k es el número de categorías.

Mediana

- **En datos sin tabular**: los datos se ordenan de menor a mayor y se ubica el valor central. Si hay dos valores centrales, entonces se promedian.
- **En datos tabulados**:

$$M_d = L_i + c\left(\frac{\frac{n}{2} - N_{i-1}}{n_i}\right)$$

la mediana se encuentra dentro de la clase (categoría) que contiene a la posición n/2. Donde L_i es el límite inferior de esta clase, c es la amplitud de esta clase, N_{i-1} es la frecuencia acumulada anterior a esta clase y n_i es la frecuencia absoluta.

Moda

- **En datos sin tabular**: es el valor de la variable con mayor frecuencia.
- **En datos tabulados:**

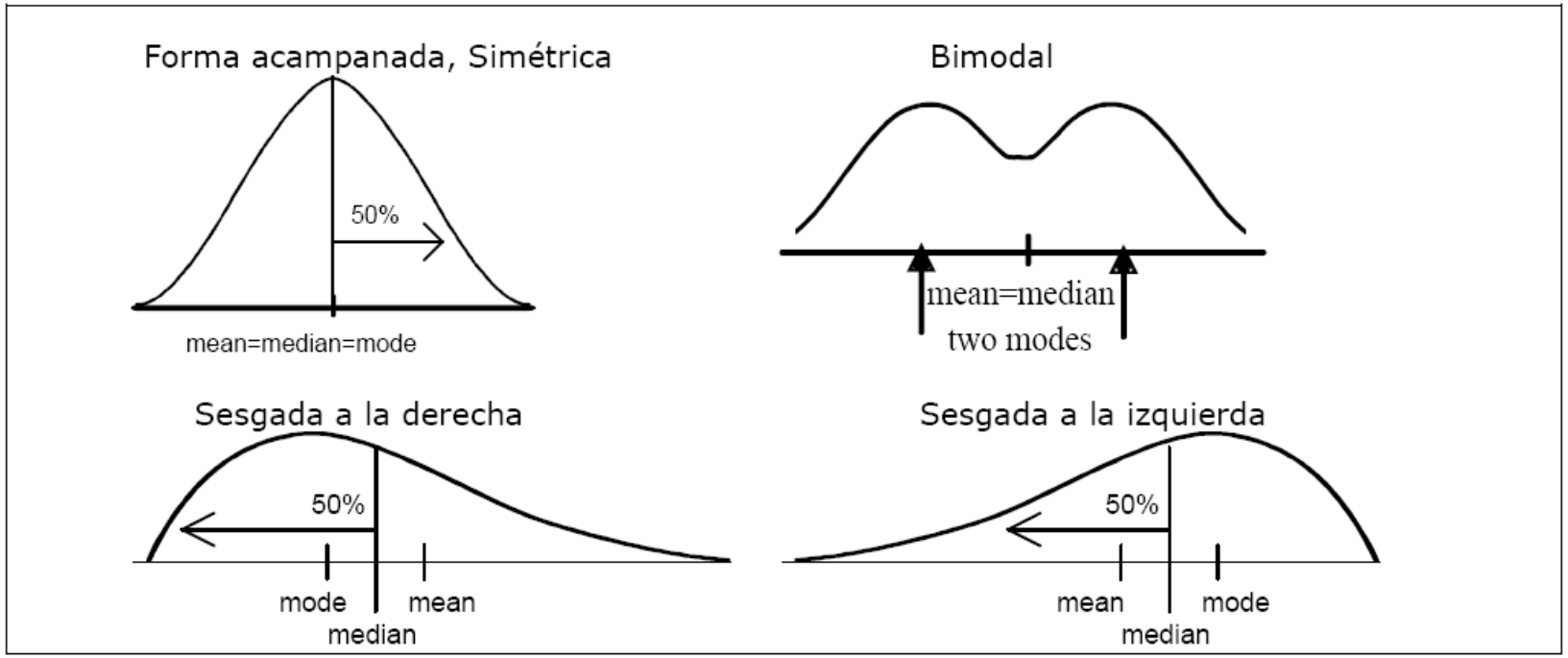
$$M_d = L_i + c\left(\frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})}\right)$$

donde n_i es la frecuencia absoluta mayor.

Si una distribución muestra dos valores modales, indicaría la posibilidad que dos poblaciones se encuentren mezcladas y sea necesario separarlas.

Relación entre Media, Mediana y Moda

- Si media=moda=mediana, la distribución es simétrica
- Si media > mediana, la distribución es asimétrica con cola a la derecha (sesgada a la derecha).
- Si media < mediana, la distribución es asimétrica con cola a la izquierda (sesgada a la izquierda).



…Media vs. Mediana

- La media es un estadístico sensible a valores extremos. Basta que algún dato dentro de la muestra sea muy alto o muy bajo, el promedio se verá alterado.
- La mediana, en cambio, es un estadístico robusto. Aunque los extremos de los datos se vean alterados, la mediana permanece invariable.
- El famoso trío - **media**, **mediana** y **moda** – representan tres métodos diferentes para encontrar el valor del **centro**. Estos tres valores pueden ser un mismo valor pero a menudo son distintos. Cuando son distintos, pueden servir para diferentes interpretaciones de los datos que queremos resumir. Considere el ingreso mensual de cinco familias en un barrio:

  $120 000 $120 000 $300 000 $900 000 $1 000 000

  ¿Cuál es el ingreso típico de este grupo?

  El ingreso mensual promedio es:

  La mediana del ingreso mensual es:

  La moda del ingreso mensual es:

  Si tú estás tratando de promover el barrio, ¿Qué medida usarías?

  Si tú estás tratando que bajen las contribuciones, ¿Qué medida usarías?

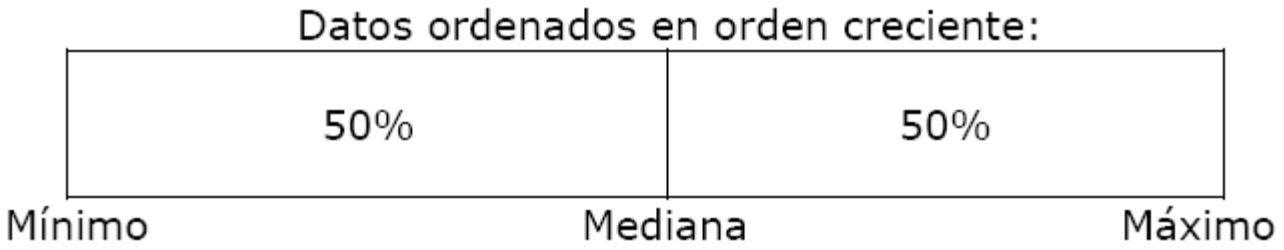
Estadísticos de Posición

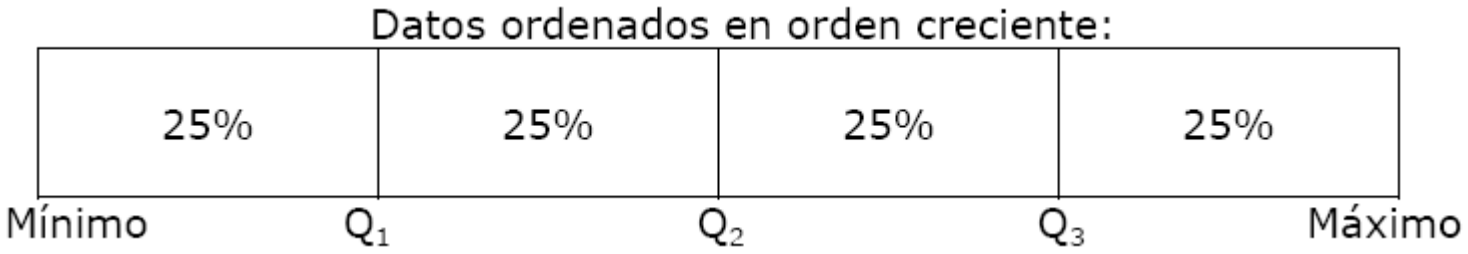
- Son valores de la variable que dividen a la muestra en partes de igual porcentaje. Los percentiles separan la muestra en grupos de 1% cada uno (son 99).

Cuartiles: agrupan 25% cada uno (son 3)

Quintiles: agrupan 20% cada uno (son 4)

Deciles: agrupan 10% cada uno (son 9)





Percentiles

- **En datos sin tabular:**
- Primero se ordenan de menor a mayor los n datos.
- Calcular el valor

$$A = \frac{n \times k}{100}$$

1. Si A es entero, entonces el percentil k corresponde al valor medio de las observaciones ubicadas en las posiciones A y A+1.
2. Si A no es un entero, el percentil k corresponde a la observación ubicada en la posición entera siguiente, es decir, [A+1].

…Ejemplos de percentiles

- Determinar los percentiles 25 y 60 de los siguientes datos:

  3, 5, 5, 8, 12, 15, 21, 23, 25, 26, 29, 35

- P_{25}: A= 12 x 25 /100 = 3

  Aqui, resulta un entero, por tanto el P25 corresponde al promedio de las observaciones en las posiciones 3º y 4º, es decir, P_{25}= (5+8)/2 = 6.5

- P_{60}: A = 12 x 60 / 100 = 7.2

  En este caso A no es un entero, nos movemos al entero siguiente. Es decir, P_{60} = 23 (observación en la 8ª posición).

…Percentiles

- **En datos agrupados:**

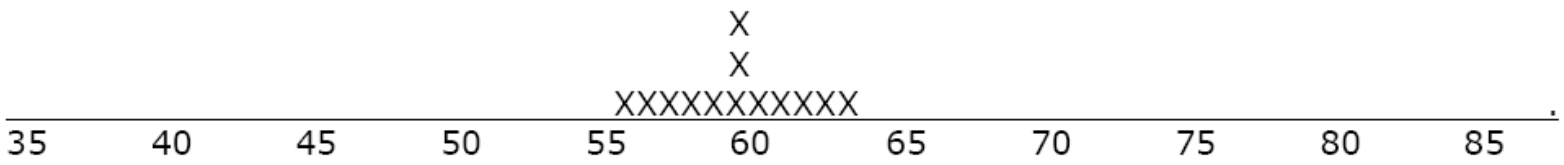
$$P_j = L_i + c\left(\frac{\frac{n \times j}{100} - N_{i-1}}{n_i}\right)$$

donde j es el porcentaje hasta donde se desea acumular, L_i es el límite inferior de la clase del percentil, N_{i-1} es la frecuencia acumulada anterior a esta clase y n_i la frecuencia absoluta.

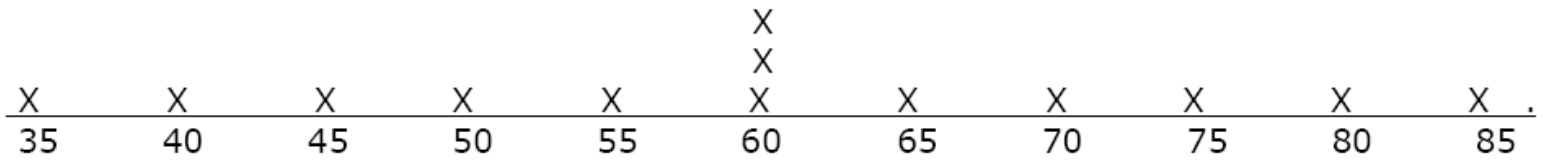
Estadísticos de Dispersión

- Las medidas de tendencia central son útiles pero nos dan una interpretación parcial de los datos. Consideremos los dos siguientes conjuntos de datos:

Datos 1: 55, 56, 57, 58, 59, 60, 60, 60, 61, 62, 63, 64, 65



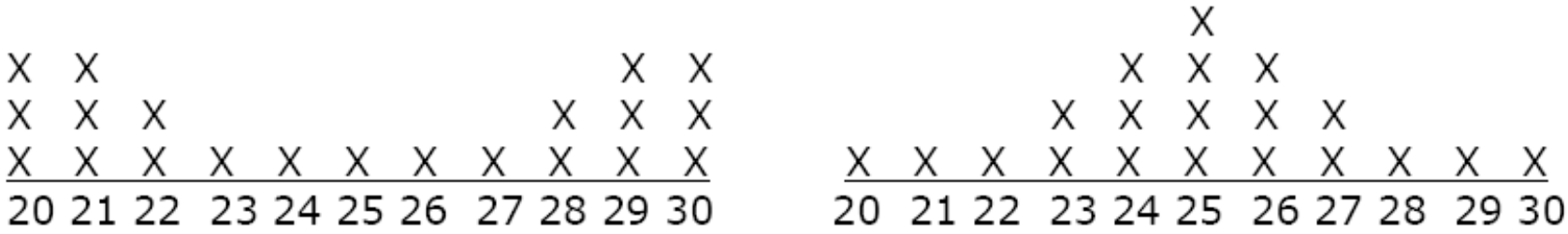
Datos 2: 35, 40, 45, 50, 55, 60, 60, 60, 65, 70, 75, 80, 85



- **Rango**: Es la medida de variabilidad o dispersión más simple. Se calcula tomando la diferencia entre el valor máximo y el mínimo observado. **Rango = Máximo – Mínimo.**

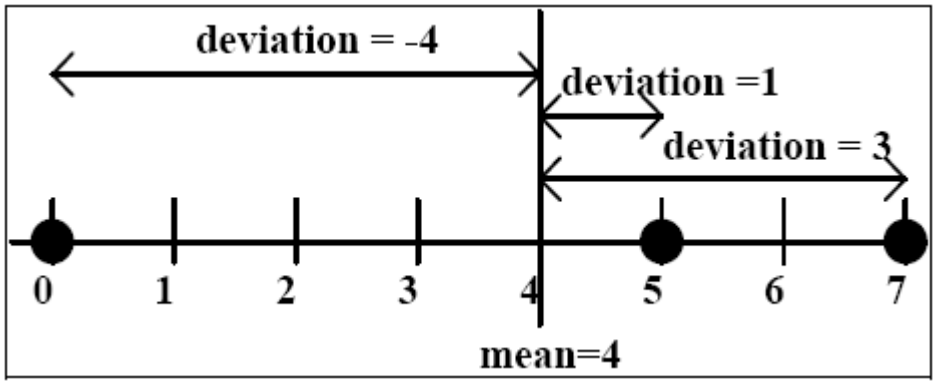
Desviación estándar

- Analizar cuáles podrían ser las ventajas y desventajas del rango como medida de variabilidad.



- **Desviación estándar**

  Es una medida de la dispersión de las observaciones a la media. Es un **promedio de la distancia de las observaciones a la media**.



…Varianza muestral

Observación x	Desviación $x - \bar{x}$	Desviación al cuadrado $(x - \bar{x})^2$
0	0 – 4 = - 4	16
5	5 – 4 = 1	1
7	7 – 4 = 3	9
Promedio = 4	Suma = 0	Suma = 26

La **varianza muestral** está definida como la suma de las desviaciones al cuadrado divididas por el tamaño muestral menos 1, es decir, dividas por n – 1 .

$$\text{varianza muestral} = \frac{(-4)^2 + (1)^2 + (3)^2}{3-1} = \frac{16+1+9}{2} = \frac{26}{2} = 13$$

$$\text{desviación estándar muestral} = \sqrt{13} \approx 3.6$$

En Resumen

Pensemos la desviación estándar como aproximadamente un *promedio de las distancias* de las observaciones a la media.

Si todas las observaciones son iguales, entonces la desviación estándar es cero.

La desviación estándar es positiva y mientras más alejados están los valores del promedio, mayor será la desviación estándar.

…Varianza muestral

- **En datos sin tabular:** Si x_1, x_2, …, x_n denota una muestra con n observaciones, la varianza muestral se denota por:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

- La desviación estándar muestral, denotada por s, es la raíz cuadrada de la varianza

$$s = \sqrt{s^2}$$

- La varianza y la desviación estándar no son medidas de variabilidad distintas, debido a que la última no puede determinarse a menos que se conozca la primera.

…Varianza muestral

- A menudo se prefiere la desviación estándar en relación con la varianza, porque se expresa en las mismas unidades físicas de las observaciones.
- **Si los datos están tabulados:**

$$s = \sqrt{\frac{\sum_{i=1}^{k}(y_i - \overline{x})^2 n_i}{n-1}}$$

- donde y_i es la marca de clase de la categoría i-ésima, n_i la frecuencia absoluta de la i-ésima clase y k es el número de categorías.

…Rango entre Cuartiles

- Así como el promedio es una medida de tendencia central que no es resistente a las observaciones extremas, la desviación estándar, que usa el promedio en su definición, tampoco es una medida de dispersión resistente a valores extremos.
- Tenemos argumentos estadísticos para demostrar por qué dividimos por n − 1 en vez de n en el denominador de la varianza muestral.
- **Rango entre cuartiles**

  La diferencia entre el tercer cuartil y el primer cuartil se llama rango entre cuartiles, denotado por $\mathbf{RQ=Q_3-Q_1}$. El rango entre cuartiles mide la variabilidad de la mitad central de los datos.

Variabilidad

¿Qué es Variabilidad?

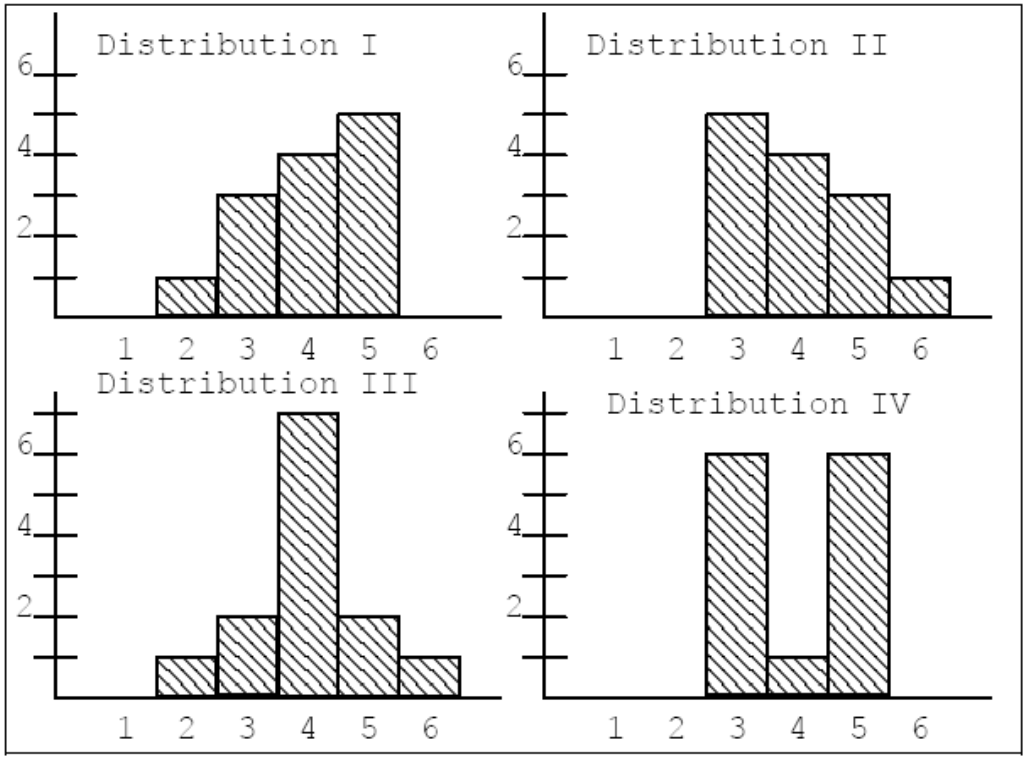
Considere los 4 conjuntos de datos siguientes y sus histogramas:

Datos I:
2 3 3 3 4 4 4 4 5 5 5
5 5

Datos II:
3 3 3 3 3 4 4 4 4 5 5
5 6

Datos III:
2 3 3 4 4 4 4 4 4 4 5
5 6

Datos IV:
3 3 3 3 3 3 4 5 5 5 5
5 5



Medidas de variabilidad	I	II	III	IV
Rango				
Rango entre cuartiles				
Desviación Estándar				

¿Qué es variabilidad?

- Algunas personas asocian variabilidad con rango mientras que otras asocian variabilidad con cómo difieren los valores de la media. Hay muchas medidas de variabilidad, y la desviación estándar es la más usada. Pero recuerden que una distribución con la menor desviación estándar no es necesariamente la distribución que es menos variable con respecto a otras definiciones de variabilidad.
- **Resumen**: Cuando queremos describir una variable usamos alguna medida de posición central y una medida de dispersión. El par de medidas más comúnmente usado es la media aritmética y la desviación estándar. Pero vimos que cuando la distribución de las observaciones es sesgada, la media no es una buena medida de posición central y preferimos la mediana. La mediana en general va acompañada del rango como medida de dispersión. Pero cuando observamos valores extraños (extremos) el rango se ve muy afectado, por lo que preferimos usar el rango entre cuartiles.

…Resumen

Medida de tendencia central	Medida de dispersión	Uso en Distribuciones	Ventajas	Desventajas
Promedio	Desviación estándar	Simétricas	Buenas propiedades, muy usados.	Sensible a valores extremos.
Mediana	Rango	Sesgadas, sin valores extremos	Mediana robusta a valores extremos. Rango muy conocido, fácil de entender.	Rango sensible a valores extremos.
Mediana	Rango entre cuartiles	Sesgadas con valores extremos	Medidas robustas a valores extremos.	El rango entre cuartiles no es muy conocido.

¿Qué son los outliers?

- **Valores extremos o anómalos (outliers):** son observaciones que se alejan del conjunto der datos.
- Una regla para determinar si un dato es outliers es:
  - Si un dato es < Q1 – 1.5(Q3-Q1)
  - Si un dato es > Q3 + 1.5(Q3-Q1)
- Los valores extremos por lo general son atribuibles a una de las siguientes causas:
- La observación se registra incorrectamente.
- La observación proviene de una población distinta.
- La observación es correcta pero representa un suceso poco común (fortuito).

…Ejemplo

- Analizar si los siguientes datos poseen valores outliers. Se trata de las edades de un grupo de pacientes de un médico:

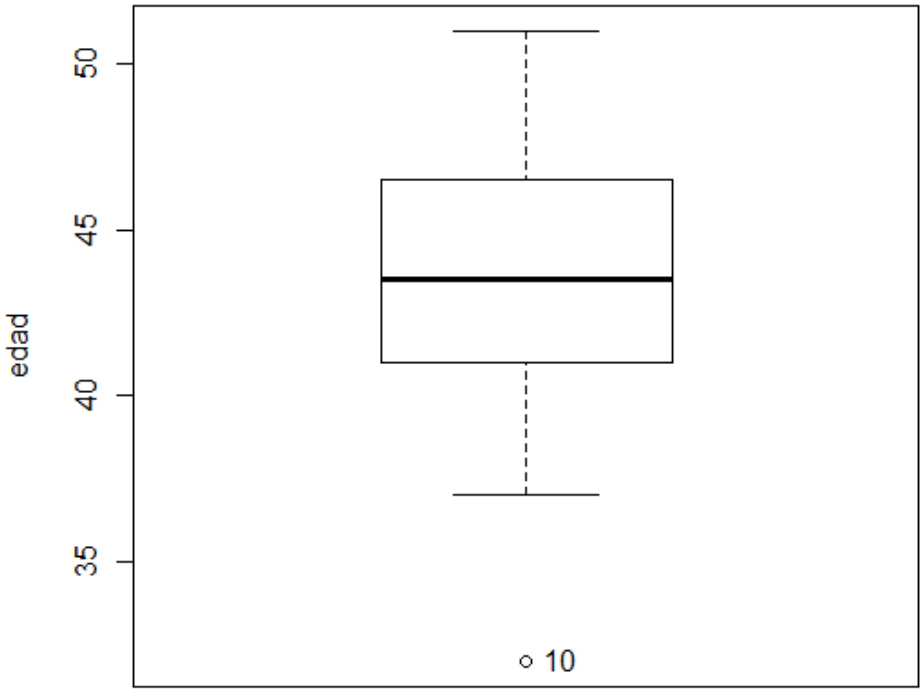
  45 41 51 46 47 42 43 50 39 32 41 44 47 49 45 42 41 40 45 37
- Primero ordenamos la muestra:

  32 37 39 40 41 41 41 42 42 43 44 45 45 45 46 47 47 49 50 51
- Calcular los cuartiles: $Q1=P_{25}=41$, $Q2=P_{50}=43.5$ y $Q3=P_{75}=46.5$
- Rango entre cuartiles: Q3-Q1=46.5-41=5.5
- límite inferior: 41-1.5x5.5= 32.75
- Límite superior: 46.5+1.5x5.5= 54.75
- Por lo tanto queda una observación fuera del límite inferior: 32 (la décima observación de la base de datos original).

Boxplot

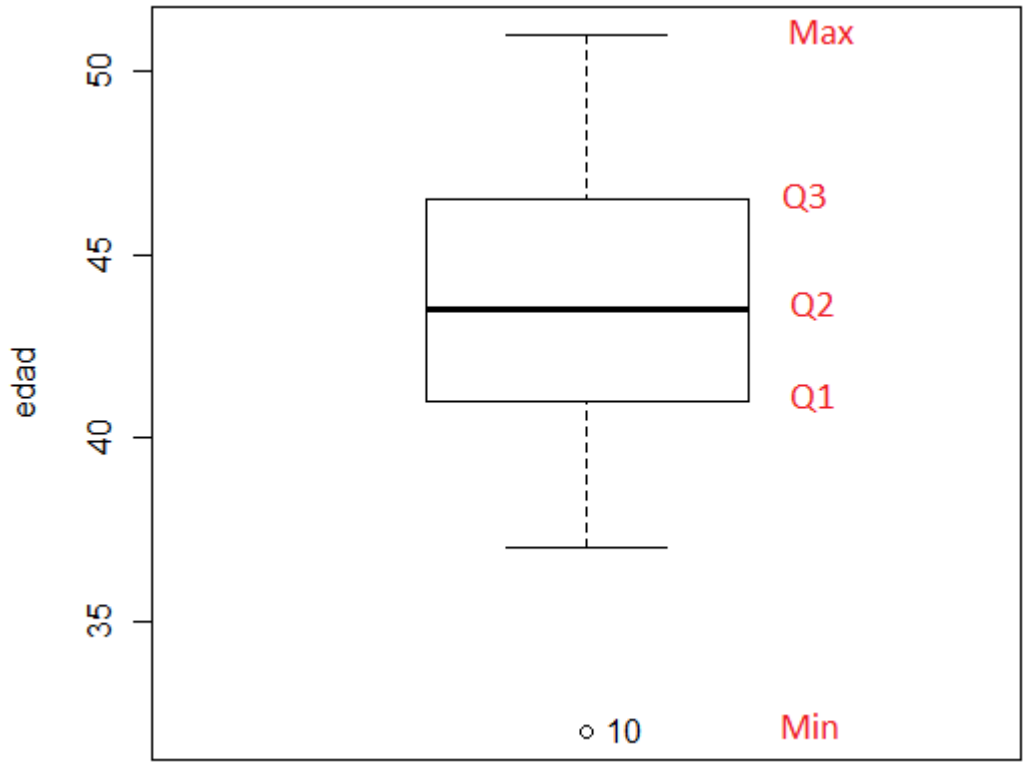
El diagrama de cajas de construye de la siguiente forma:

- Dibujar la caja que empieza en el primer cuartil y termina en el tercer cuartil.
- Dibujar la mediana con una línea dentro de la caja.
- Por último, se extienden las líneas (bigotes) saliendo de la caja hasta el mínimo y el máximo (salvo en la presencia de outliers).

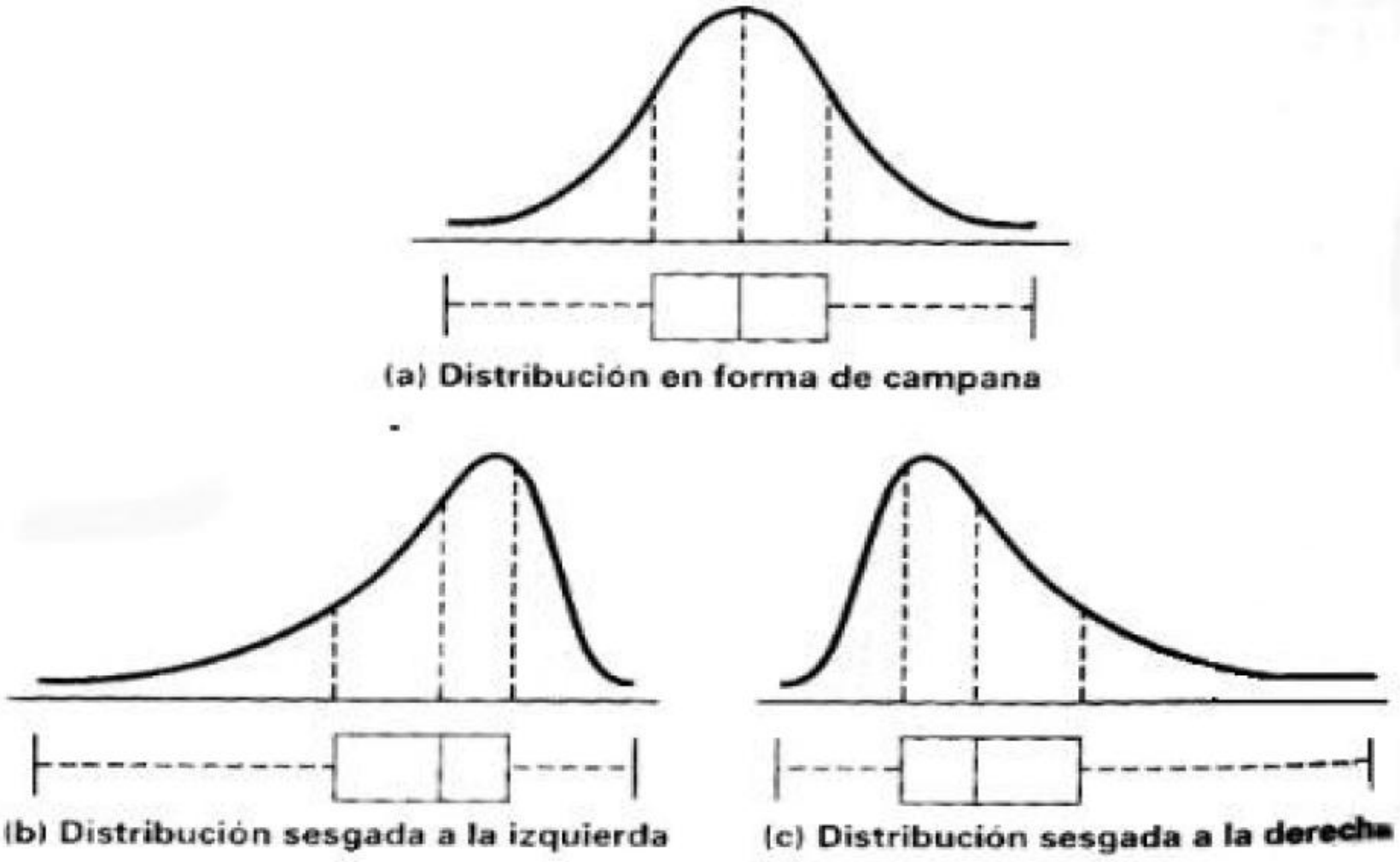


… Boxplot

- En la presencia de outliers, los bigotes se extienden hasta el valor observado anterior al valor extremo. La distancia entre la mediana y los cuartiles es aproximadamente la misma, lo que nos hace pensar que la distribución de los datos es más o menos simétrica.

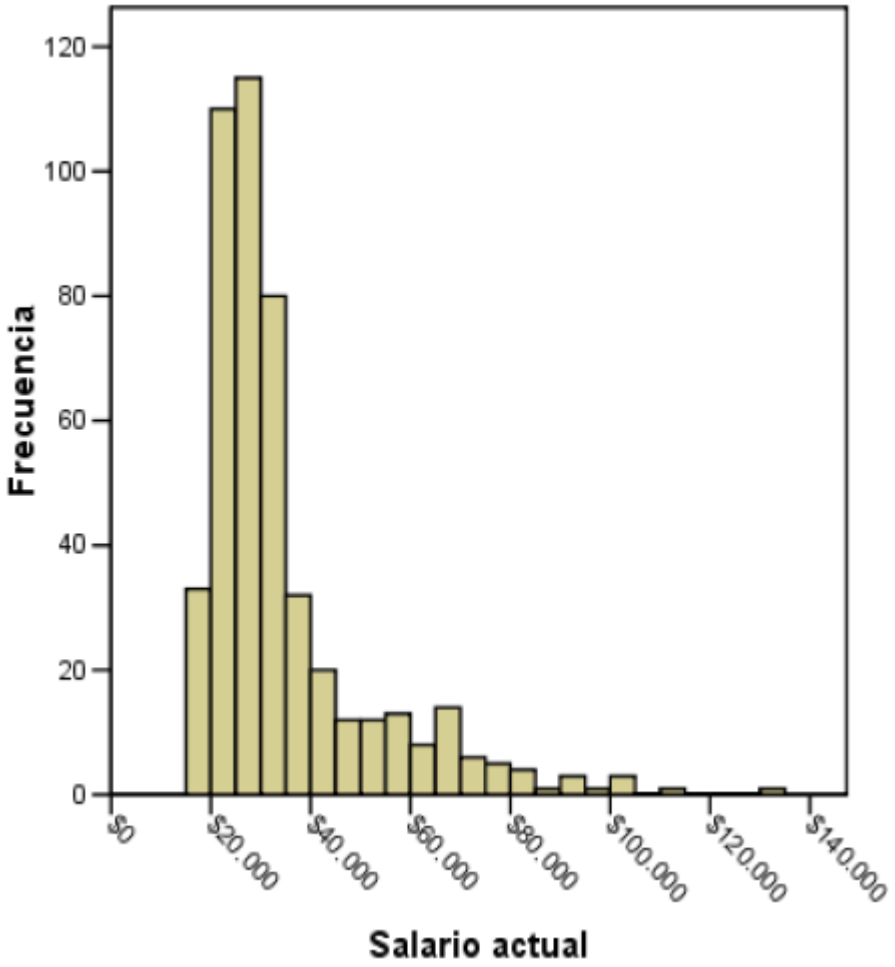


…Boxplot


(a) Distribución en forma de campana
(b) Distribución sesgada a la izquierda
(c) Distribución sesgada a la derecha

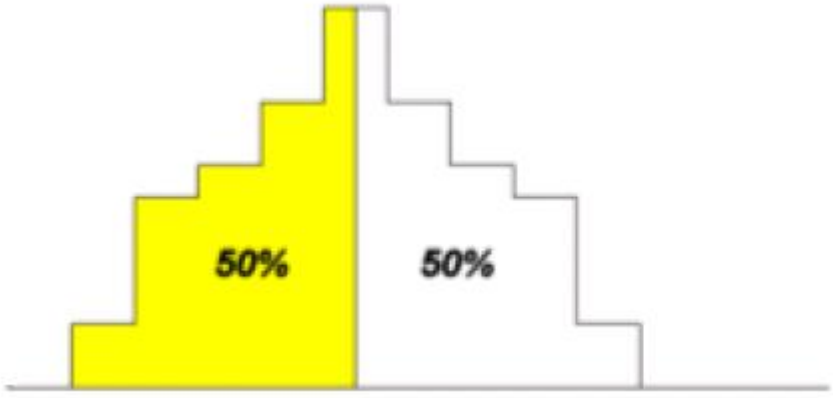
Estadísticos de Forma

¿Qué nos dice la forma de la distribución de la variable salario actual que se muestra en el siguiente histograma?



Asimetría

- La simetría de una distribución de frecuencias hace referencia al grado en que valores de la variable, equidistantes a un valor que se considere centro de la distribución, poseen frecuencias similares.
- Es un concepto más intuitivo a nivel visual, especialmente, si se observa una representación gráfica (diagrama de barras, histograma…) de la distribución de frecuencias. Ésta será simétrica si la mitad izquierda de la distribución es la imagen especular de la mitad derecha.

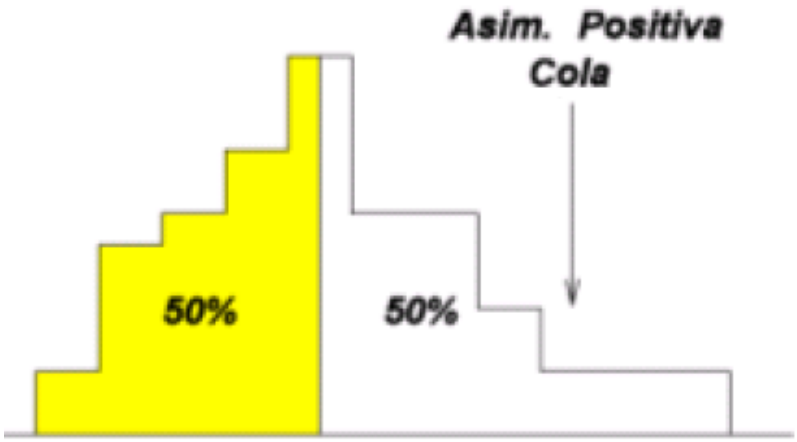


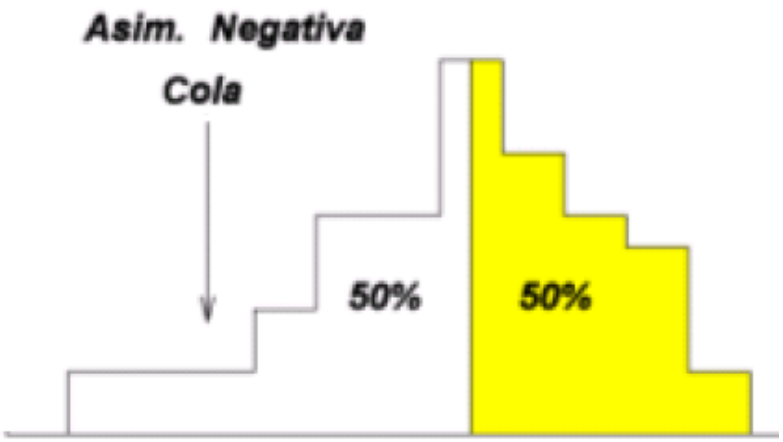
…Asimetría

- Media y mediana coinciden en las distribuciones simétricas. Si sólo hay una moda (distribución unimodal), el valor de ésta también será igual a las dos anteriores.
- En distribuciones unimodales, el nivel de simetría se suele describir de acuerdo a tres grandes categorías: distribuciones simétricas, distribuciones asimétricas positivas (o sesgada a la derecha) y distribuciones asimétricas negativas (o sesgada a la izquierda). Tomando como eje de referencia a la moda, estas categorías de asimetría vienen definidas por el diferente grado de dispersión de los datos a ambos lados (colas) de ese eje virtual. La cola más dispersa en el lado de los valores altos de la variable caracteriza a la asimetría positiva; si en el lado de los más bajos, a la asimetría negativa; y si la dispersión es igual o muy similar a ambos lados, a una distribución de frecuencias simétrica.

…Asimetría

- En caso de asimetría, los valores de la media, mediana y moda difieren. En concreto si la asimetría es positiva: media>mediana>moda. Si la asimetría es negativa: media<mediana<moda.





…Asimetría

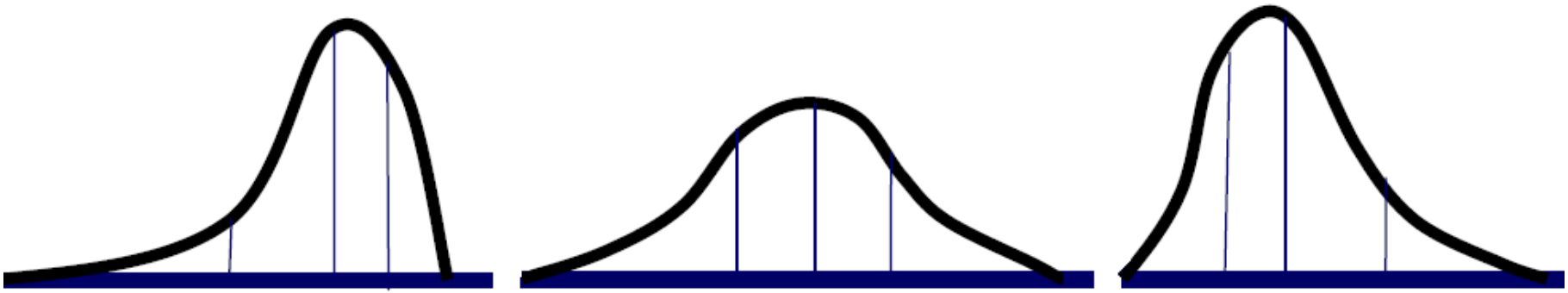
- A continuación se presentan diferentes índices estadísticos que permiten cuantificar el nivel de asimetría de una variable. Destacar antes que para variables nominales no tiene sentido el plantear este tipo de índices, dado que no existe un orden intrínseco a los valores de la variable.
- **Índice de asimetría para variables ordinales:**

  Se basa en las distancias entre los cuartiles a fin de establecer un resumen de la asimetría de la distribución.

$$As = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

Nota: oscila entre -1 y 1 lo cual facilita la comprensión.

… Asimetría



- **Índice de asimetría para variables cuantitativas:**

  Primer coeficiente de Pearson: se basa en la relación existente entre la media y la moda en distribuciones unimodales asimétricas.

$$As = \frac{\overline{x} - M_o}{s}$$

…Asimetría

Interpretación del coeficiente de Pearson: los valores menores que 0 indican asimetría negativa; los mayores, asimetría positiva y cuando sea cero, o muy próximo a cero, simétrica. No está limitado a un rango de valores.

- Coeficiente de asimetría de Fisher: se basa en las desviaciones de los valores observados respecto a la media. La interpretación de los resultados proporcionados por este coeficiente es igual a la del primer coeficiente de Pearson.

$$As = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{ns^3}$$

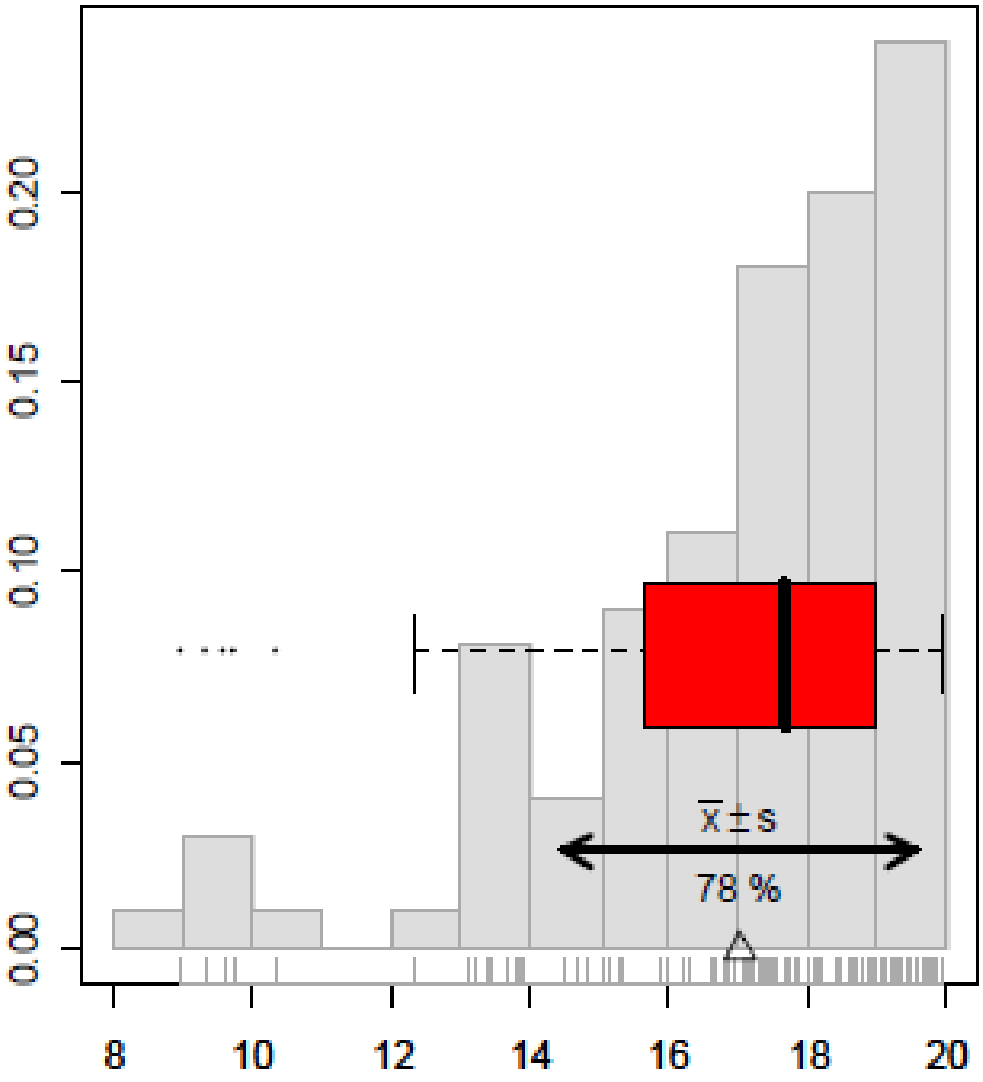
…Coeficiente de asimetría de Fisher

- Y para el caso de datos tabulados:

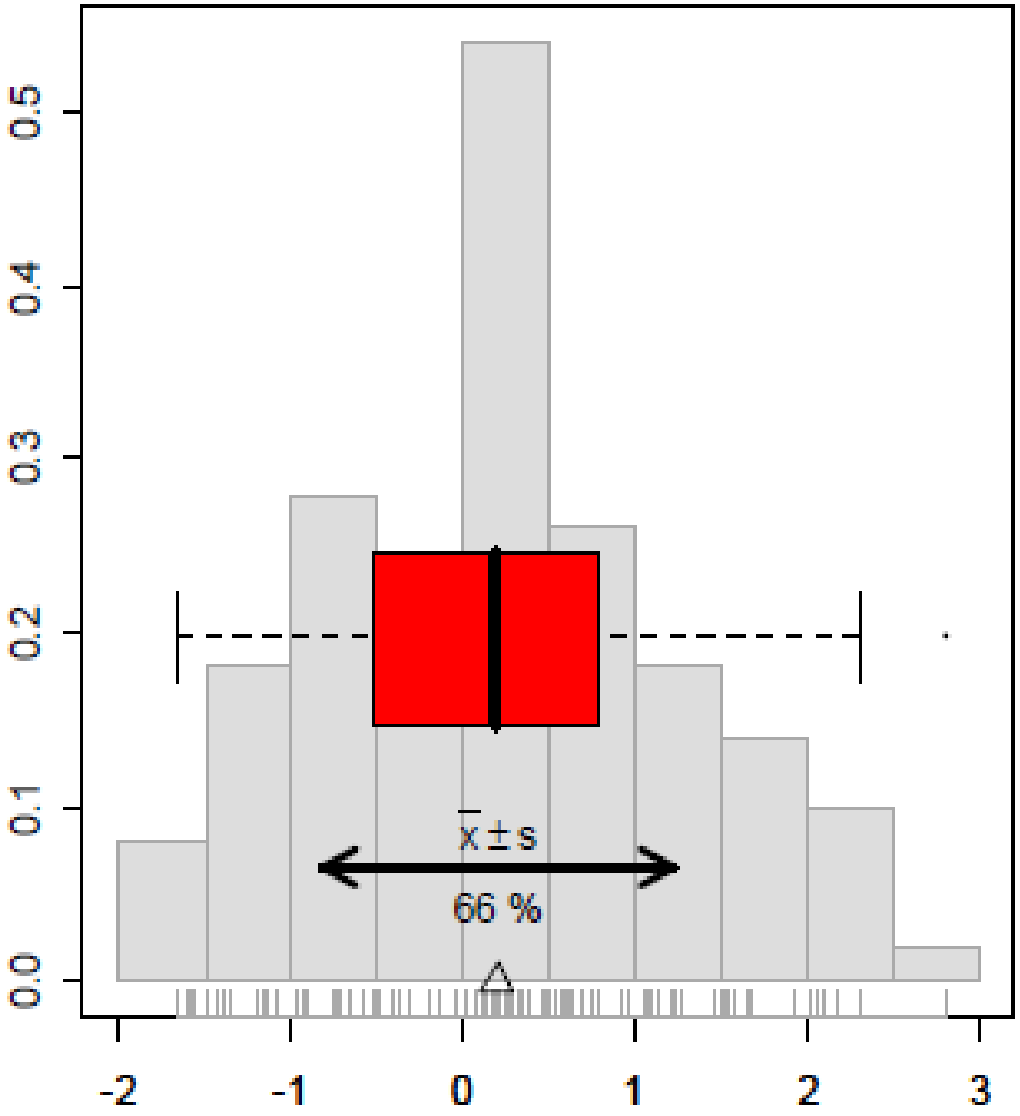
$$A_s = \frac{\sum_{i=1}^{k}(y_i - \overline{x})^3 n_i}{n s^3}$$

- Acorde al tipo de variable que nos ocupa, el histograma representa la mejor opción en la visualización de la asimetría de una variable, por otro lado, el diagrama de caja y bigotes (boxplot) también constituye una opción válida para tal fin. A continuación se presenta un ejemplo con ambos tipos de gráficos superpuestos, en que se muestran 3 variables que ilustran distribuciones con diferente nivel de asimetría:

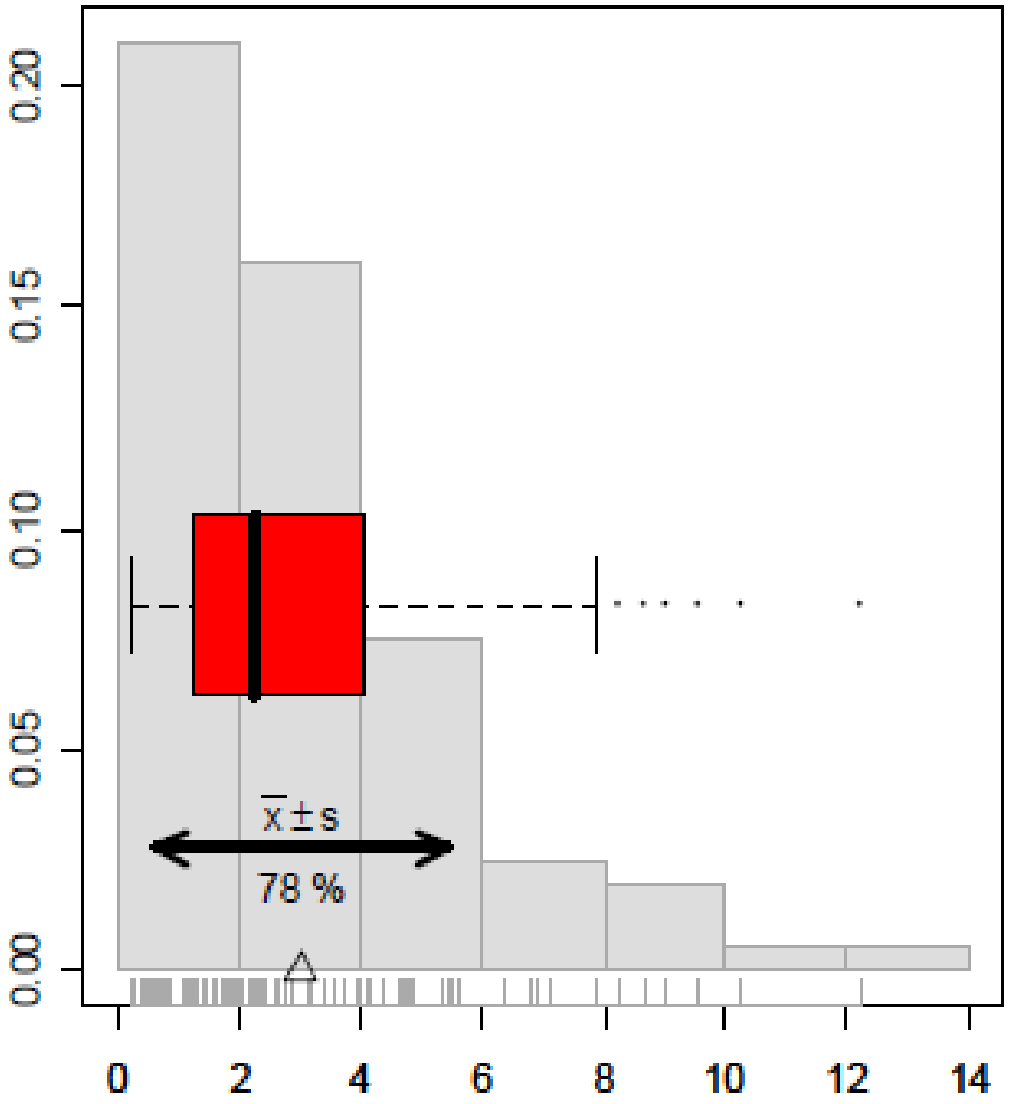
Asimetría negativa



Asimetría cercana a cero



Asimetría positiva



Apuntamiento (curtosis)

- El apuntamiento o curtosis de una distribución de frecuencias no tiene un referente natural como en el caso de la simetría, sino que se sustenta en la comparación respecto a una distribución de referencia, en concreto, la distribución normal o campana de Gauss. En consecuencia, su obtención sólo tendrá sentido en variables cuya distribución de frecuencias sea similar a la de la curva normal –en la práctica ello se reduce, básicamente, a que sea unimodal y más o menos simétrica.
- El apuntamiento expresa el grado en que una distribución acumula casos en sus colas en comparación con los casos acumulados en las colas de una distribución normal cuya dispersión sea equivalente. Así, de forma análoga a la asimetría, se diferencian **3 grandes categorías de apuntamiento**:

…Curtosis

- Distribución platicúrtica (apuntamiento negativo): indica que en sus colas hay más casos acumulados que en las colas de una distribución normal.
- Distribución leptocúrtica (apuntamiento positivo): justo lo contrario.
- Distribución mesocúrtica (apuntamiento normal): como en la distribución normal.
- Coeficiente de apuntamiento de Fisher para variables cuantitativas: se basa en las desviaciones de los valores observados respecto a la media.

$$K = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^4}{n s^4} - 3$$

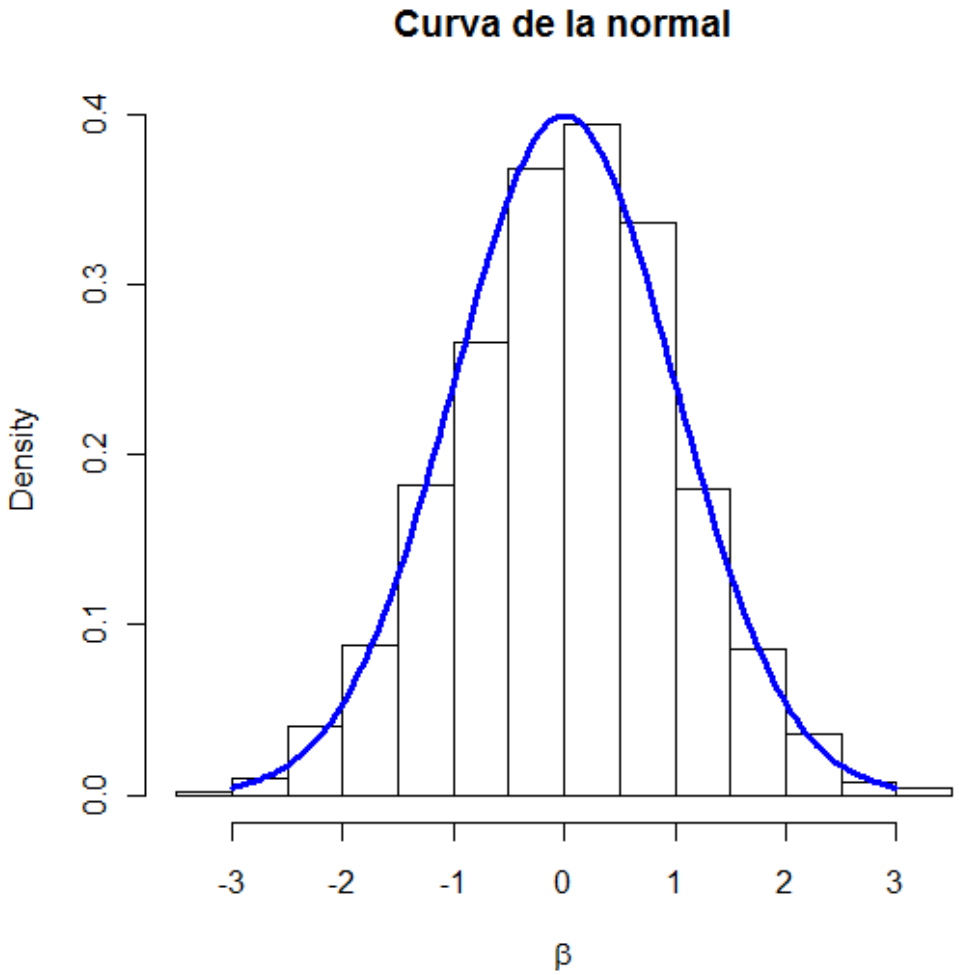
…Curtosis

- Y para el caso de datos tabulados:

$$K = \frac{\sum_{i=1}^{k}(y_i - \overline{x})^4 n_i}{n s^4} - 3$$

- Interpretación: el valor de este coeficiente para la distribución normal será igual a 0, o sea que cualquier distribución para la que se obtenga un valor de *K* igual o próximo a 0 significará que su nivel de apuntamiento es como el de la distribución normal (mesocúrtica). Valores mayores que 0, expresan que la distribución es leptocúrtica, mientras que si son menores que 0 ponen de manifiesto que la distribución es platicúrtica. No está limitado a un rango de valores.

Histograma de datos normales



La regla de Chebyshev

- Es una regla que pone un límite sobre la dispersión de la mayoría de los datos en torno de la media.
- **Teorema**. Para *cualquier* conjunto de datos, la proporción de datos que distan menos de *m* desviaciones estándar de la media es como mínimo.

$$1 - \frac{1}{m^2}$$

- Dice, por ejemplo, que por lo menos 75% de las observaciones están a menos de m=2 desviaciones estándar de la media y por lo menos, 88.88% de las observaciones están a menos de m=3 desviaciones estándar de la media.

…La regla de Chebyshev

- Ejemplo: Los siguientes datos son los números de crías nacidas conjuntamente para 18 parejas de ratones campestres.

  3 6 5 6 5 7 5 7 6 6 6 5 5 5 4 5 6 4

  Calculando la media 5.33 y la desviación estándar 1.03. Luego, la regla de Chebyshev dice que por los menos un 75% de los datos están contenidos en el intervalo (3.27, 7.39) y que el intervalo

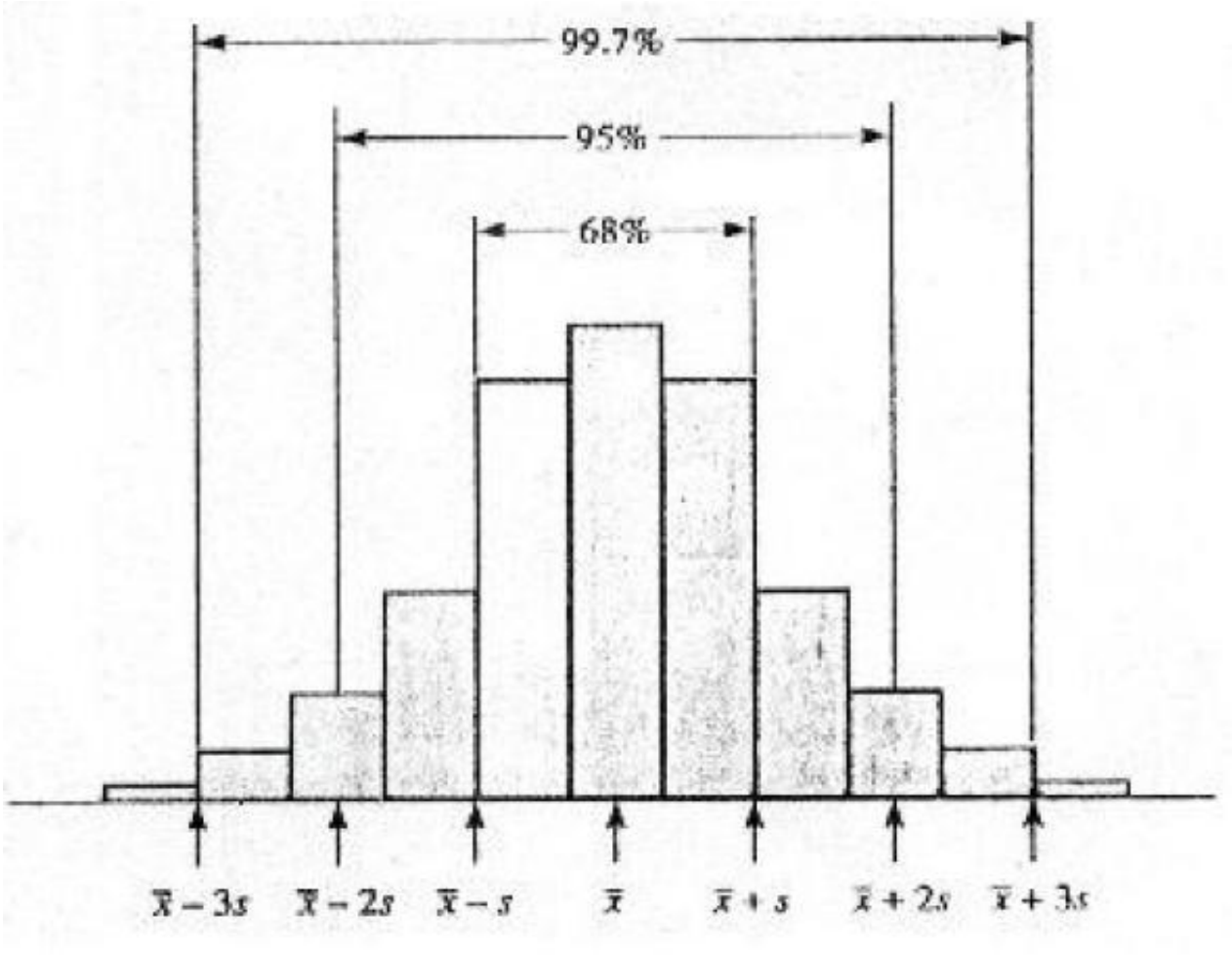
  $$5.33 \pm 3 \times 1.03 = (2.24, 8.42)$$

  contiene por lo menos un 88.88% de los datos.

Una regla empírica

- Una regla empírica dice que si la distribución de los datos es más o menos simétrica y unimodal, (es decir con una distribución normal) entonces aproximadamente un 68% de los datos caerán dentro de ±1 desviaciones estándar de la media, 95% dentro de ±2 desviaciones y 99.7% dentro de ±3 desviaciones estándar de la media.

…Una regla empírica



Fuente : Estadística Elemental. Johnson – Kuby pag 82

El Coeficiente de Variación

- Es otra medida de variabilidad que tiene la ventaja de ser sin unidades.

  Para una muestra de datos con media $\overline{x}$ y desviación estándar *s*, se define el coeficiente de variación como

$$CV = \frac{s}{|\overline{x}|}$$

  Si cambiamos la escala de medir en la variable, el coeficiente de variación no cambia. No obstante, si la media es igual a cero, el coeficiente de variación no existe.

Transformaciones

- En muchas ocasiones se quiere transformar los datos originales para que la distribución de la variable transformada tenga mejores propiedades de simetría etc., o para simplificar el análisis.
- Es interesante saber cómo cambian las características de la muestra como la media y desviación estándar.
- En general, no existe una fórmula sencilla para calcular la media de los datos transformados, salvo en el caso de que la transformación sea lineal.

… Transformaciones Lineales

- Teorema. Supongamos que tenemos una muestra $x_1, \ldots, x_n$ con media $\bar{x}$ y desviación estándar *s* y que hacemos una transformación lineal de los datos

$$y_i = \alpha + \beta x_i \quad \text{para } i = 1, \ldots, n$$

entonces la media, la varianza y desviación estándar de la muestra $y_1, \ldots, y_n$ son,

$$\begin{aligned} \bar{y} &= \alpha + \beta \bar{x} \\ s_y^2 &= \beta^2 s_x^2 \\ s_y &= \beta s_x \end{aligned}$$

respectivamente. Tarea demostrar estos resultados.

Estandarizando las observaciones

- Teorema. Dada la muestra $x_1, \ldots, x_n$ con media $\bar{x}$ y desviación estándar s_x , la distribución de las variables estandarizadas

$$y_i = \frac{x_i - \bar{x}}{s_x} \quad para\ i = 1, \ldots, n$$

tiene media 0 y desviación estándar 1.

Tarea demostrar este resultado.