Análisis e Interpretación de Datos

MÁSTER UNIVERSITARIO EN ANÁLISIS Y VISUALIZACIÓN DE DATOS MASIVOS / VISUAL ANALYTICS AND BIG DATA

Miller Janny Ariza Garzón

Tema 3. Medidas que resumen la información II

unir LA UNIVERSIDAD EN INTERNET

Universidad Internacional de La Rioja

Tabla de contenido

Tabla de contenido

Medidas resumen de la información

Tendencia central

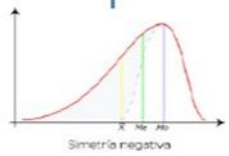
- Mediana = $x_{central}$
- Media
  - Media aritmética: $\frac{\sum x_i}{n}$
  - Media ponderada: $\frac{\sum w_i x_i}{\sum w_i}$
- Moda= n_{max}

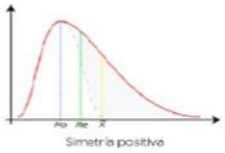
Dispersión

- Varianza: $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- Rango = $x_{max} - x_{min}$
- Coeficiente de variación: $cv = \frac{s}{\overline{x}} \times 100\%$
- Desviación típica: $s = \sqrt{s^2}$

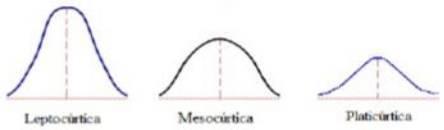
Posición y forma

- Asimetría



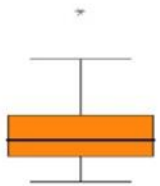


- Cuartiles y percentiles: Q_1, Q_2, Q_3 ; $P_1 \dots P_{99}$
- Tipificación: $z_i = \frac{x_i - \overline{x}}{s}$
- Apuntamiento



Valores atípicos

Caja y bigotes



Medidas que resumen la información

3.2 Medidas de tendencia central

3.3 Medidas de tendencia central robustas

} representantes de los datos

3.4 Medidas de dispersión

3.5 Medidas de dispersión robustas

} desviaciones de los datos

3.6 Medidas de posición y forma

} posición respecto a la media

3.7 Gráficos de caja

3.8 Datos atípicos y análisis exploratorio de datos

} analizar valores extremos

sd_trim() in R

- De la clase anterior.

```
> sd_trim(Datalc$dti_n,trim=0.05)
[1] 7.817512
Warning message:
In sd_trim(Datalc$dti_n, trim = 0.05) :
  Did you specify the correct consistency constant for
trimming?
> sd_trim(Datalc$dti_n,trim=0.05, const=FALSE) #
const=TRUE,incluye factor de consistencia solo para distr.
normal, disponible solo para trim=0.1 o 0.2.
[1] 7.817512
```

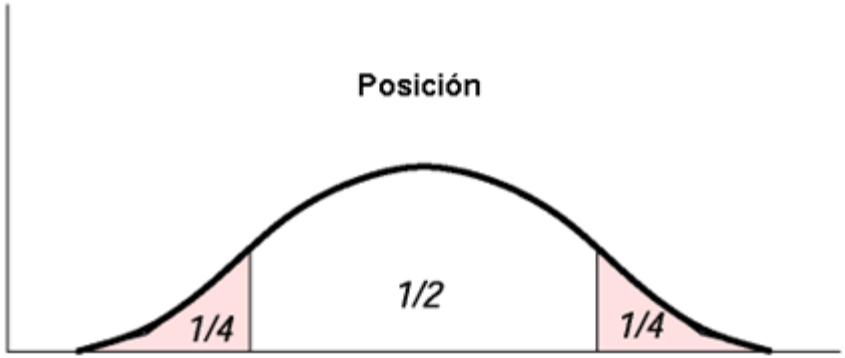
Medidas que resumen la información

Medidas de Localización

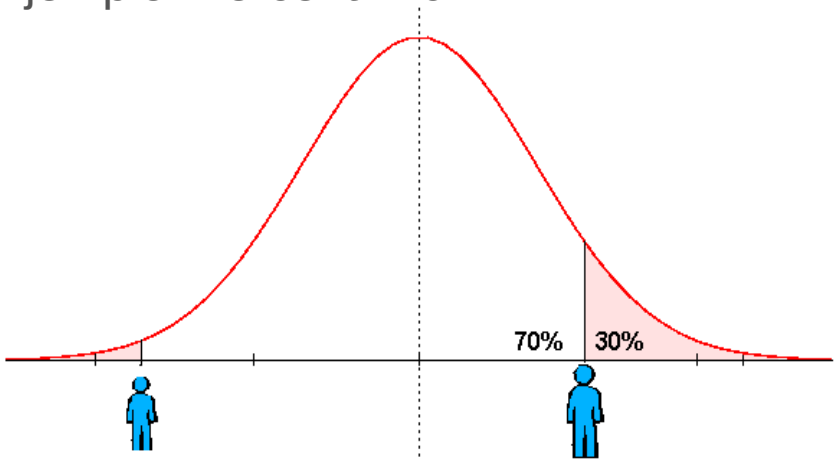
Cuantiles, cuartiles, deciles o percentiles

Cuantil: es la expresión más general de medidas de posición y comprende a todas las otras; el valor que tome el cuantil “X” es el valor que deja por debajo de sí al “X” % de los datos

Casos particulares son los percentiles, cuartiles, deciles, quintiles,...

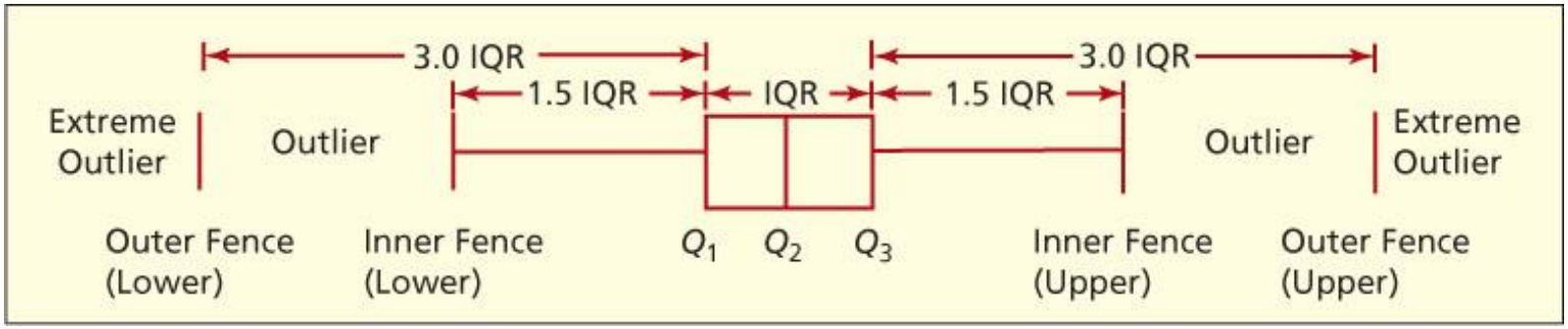


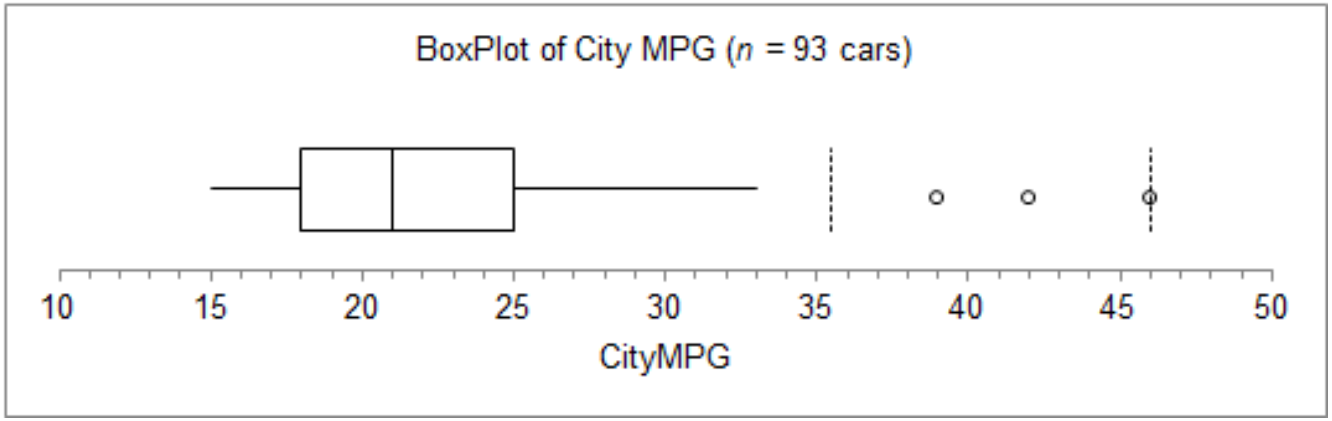
Ejemplo. Percentil 70.



Medidas que resumen la información

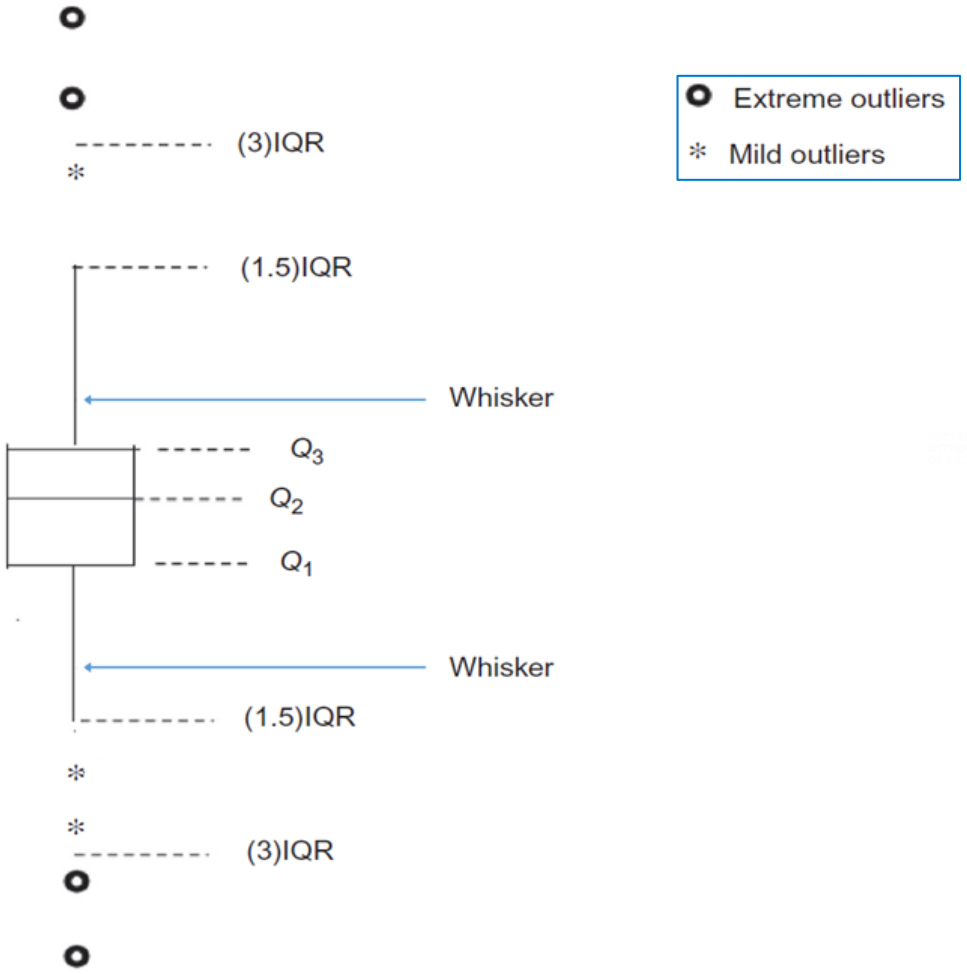
BOX-PLOT (Tukey J)





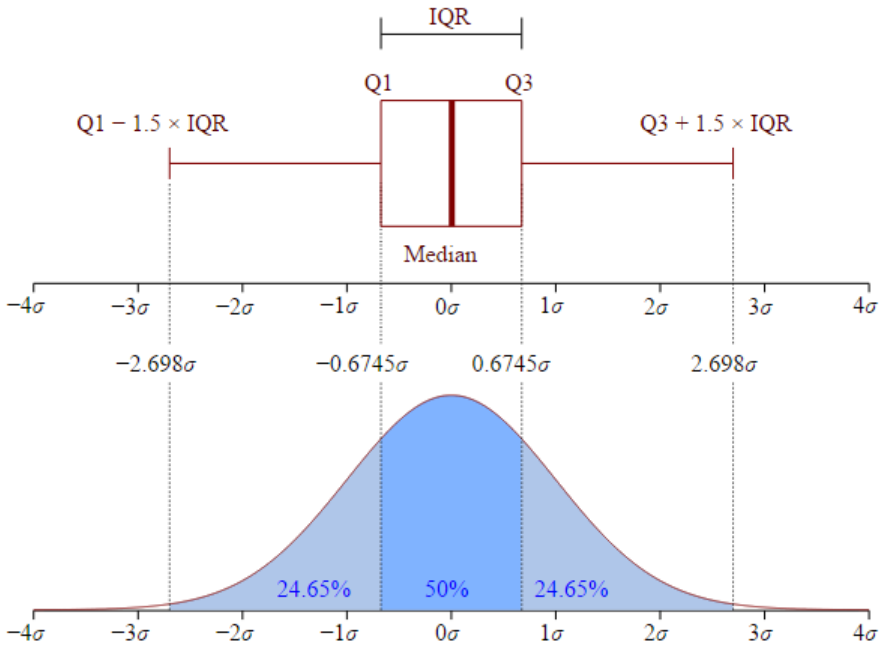
Medidas que resumen la información

BOX-PLOT (Tukey J)



Medidas que resumen la información

BOX-PLOT (Tukey J)



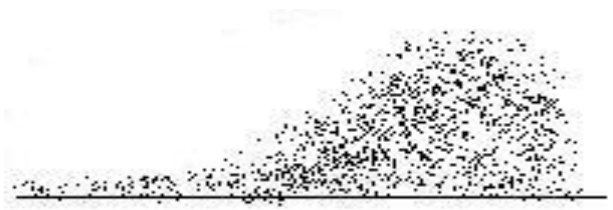
Medidas que resumen la información

Algunos ejemplos de estadísticos de Forma.

Coeficiente de asimetría o sesgo: Establece el sesgo y orientación de la distribución de los datos. Relacionado con el tercer momento de una variable. K=3.

$$\mu_k = E[(X-\mu)^k]$$

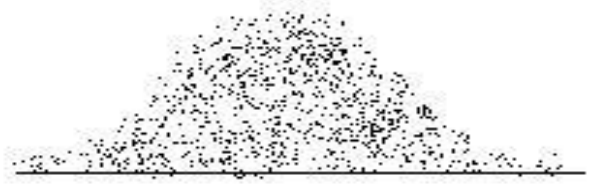
$$a = \left[\frac{n}{(n-1)(n-2)}\right]\left[\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{S^3}\right]$$



$$A_P = \frac{\bar{x} - Mo}{S_x}$$ **Coef. Asimetría de Pearson**

Sesgo a la izquierda: **a < 0**

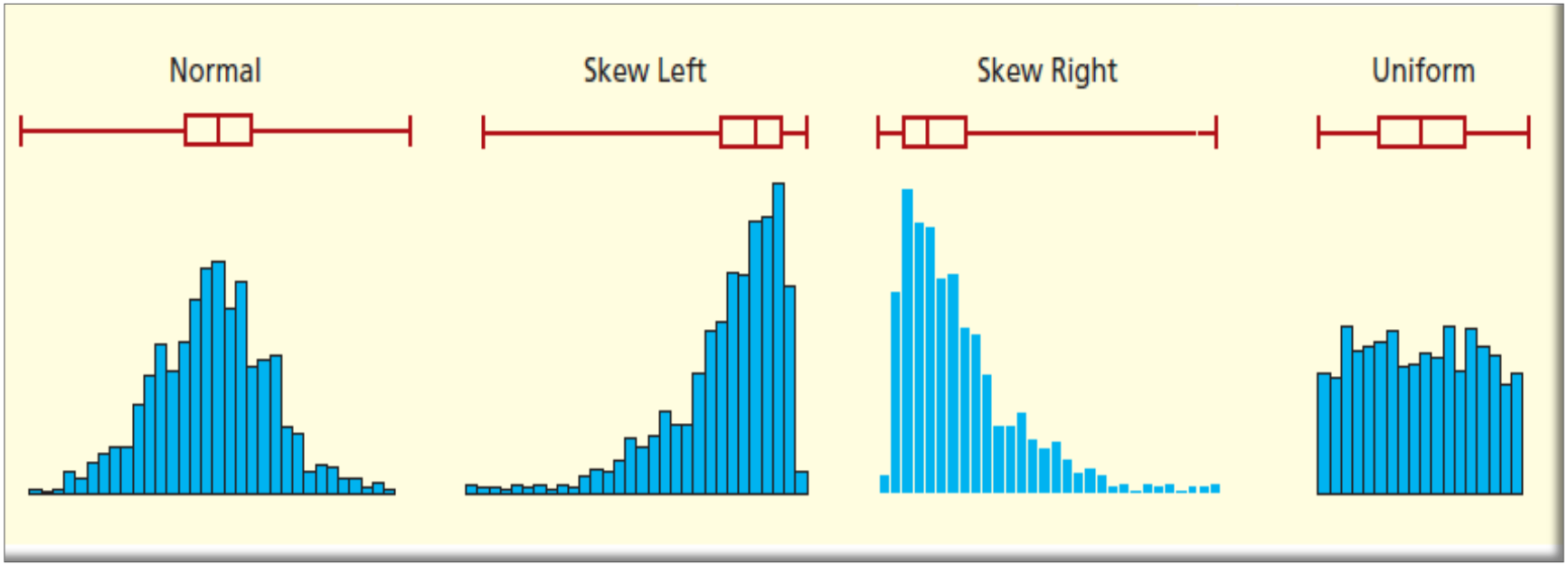
Distribución simétrica: **a = 0**



Sesgo a la derecha: **a > 0**

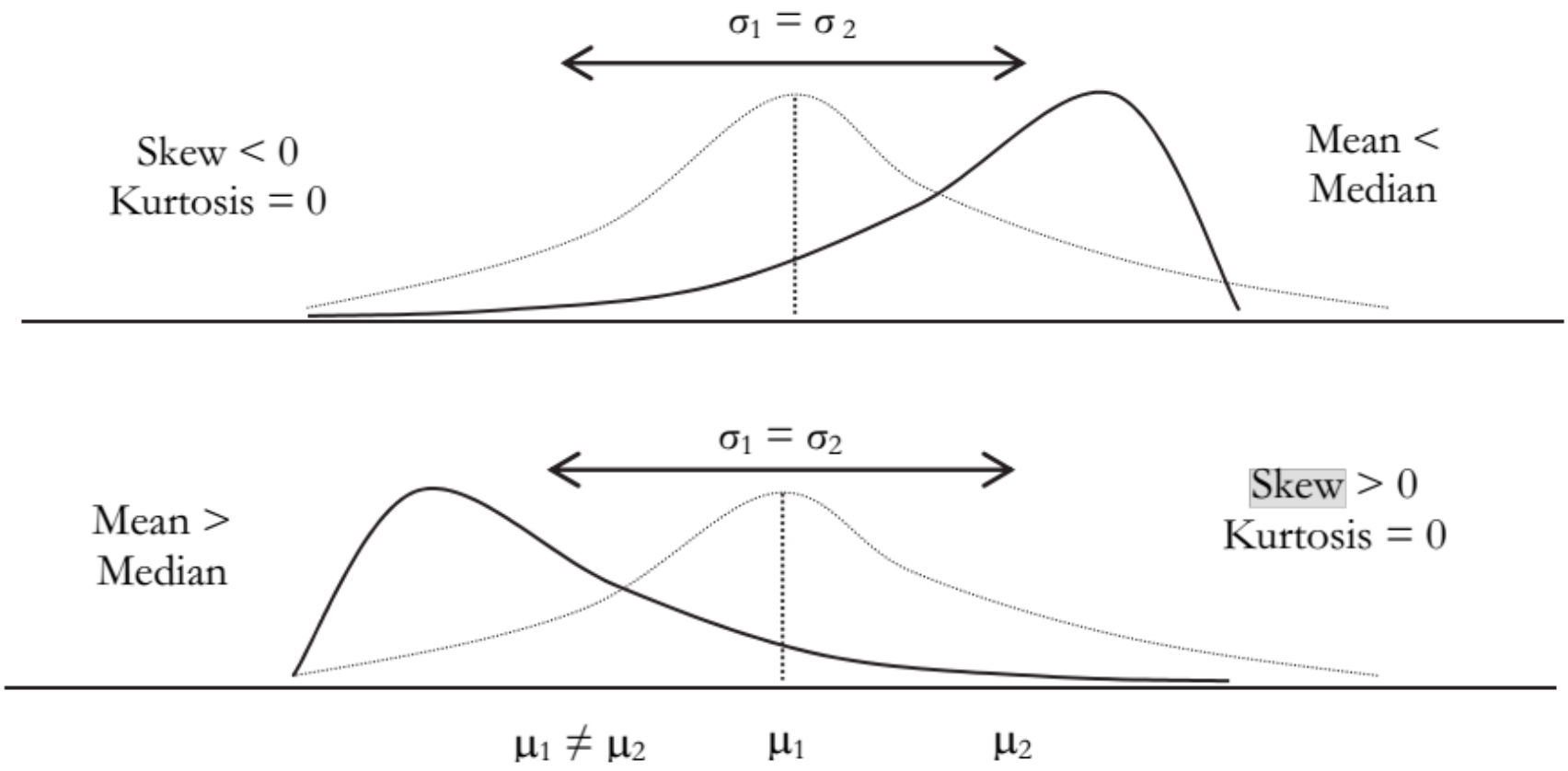
Medidas que resumen la información

Sesgo con el Box-Plot



Medidas que resumen la información

Sesgo con media y mediana



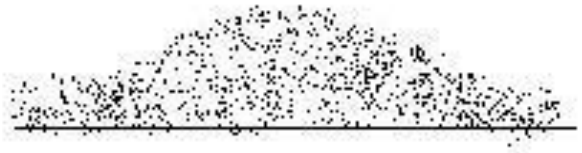
Medidas que resumen la información

Coeficiente de curtosis: El concepto de curtosis o apuntamiento de una distribución surge al comparar la forma de dicha distribución con la forma de la distribución Normal. De esta forma, clasificaremos las distribuciones según sean más o menos apuntadas que la distribución Normal, frente a su media, y a su vez el peso de sus colas, los dos componentes. Relacionado con el cuarto momento de una variable. K=4.

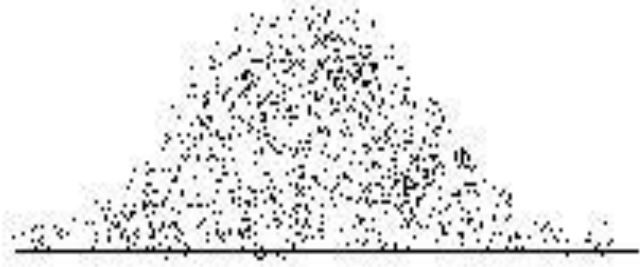
$$\mu_k = E[(X-\mu)^k]$$

Nos indica el grado de apuntamiento (aplastamiento) y pesos de las colas de una distribución con respecto a la distribución normal o gaussiana.

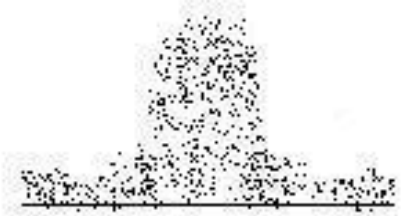
$$Curtosis = K = \left[\frac{n(n+1)}{(n-1)(n-2)(n-3)}\right]\left[\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^4}{S^4}\right]-\left[\frac{3(n-1)^2}{(n-2)(n-3)}\right]$$



Platicúrtica (aplanada): K < 0

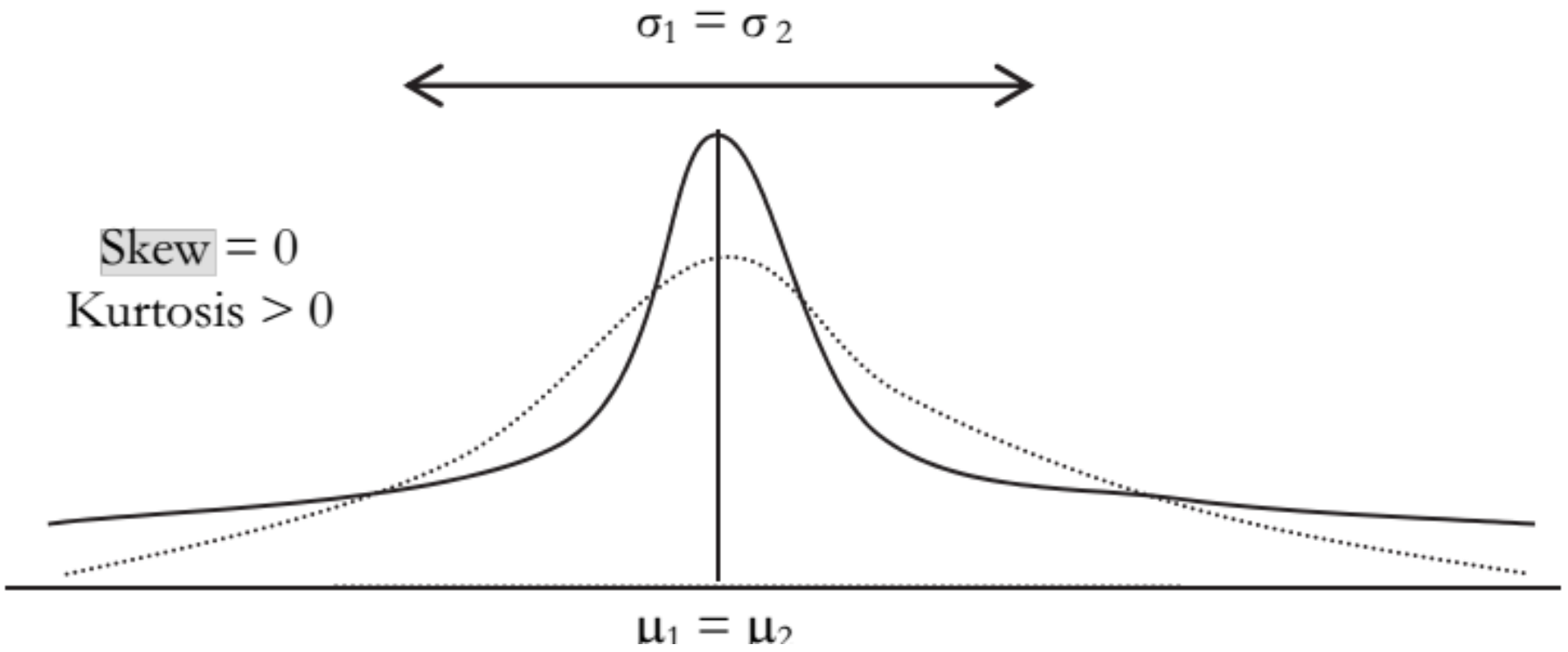


Mesocúrtica (como la normal): K = 0



Leptocúrtica (apuntada): K > 0

Medidas que resumen la información



In R

```
> median(Datalc$dti_n)
> quantile(Datalc$dti_n,probs = 0.50)
> (Q1=quantile(Datalc$dti_n,probs = 0.25))
> quantile(Datalc$dti_n,probs=c(0.25,0.5,0.75))
> boxplot(Datalc$int_rate)
> skewness(Datalc$dti_n)
> kurtosis(Datalc$annual_inc, type = 1)# definición típica
```

Próxima sesión

- Tema 4: Regresión y correlación

  - Correlación.
  - Regresión lineal.
  - Gráfico de residuos.

www.unir.net