Estadística Descriptiva

La descripción de los datos: Medidas de tendencia central, dispersión y posición

Rodrigo Asun Inostroza

Temas a tratar

- Principales medidas de tendencia central de datos nominales, ordinales y de intervalo
  - Promedio, Promedio recortado, Mediana y Moda.
- Principales medidas de dispersión de los datos
  - Rango, Varianza, Desviación típica, Coeficiente de Variación.
- Principales medidas de posición de los datos.
  - Distribuciones de frecuencias absoluta, relativa y acumulada. Medidas de posición no central: los cuantiles.
- Representaciones gráficas de estas medidas.

Propiedades de un conjunto de datos

- Ejemplos de conjuntos de datos sociológicamente relevantes:
  - Sueldos de los chilenos y chilenas.
  - Conciencia de clase de obreros y obreras.
  - Grado de acuerdo con ley Naim-Retamal.
  - Identidad étnica de estudiantes de la UFRO.
- Cualquier análisis depende del nivel de medición de las variables.
  - Ese nivel de medición depende de la naturaleza de la variable y como ha sido medida.

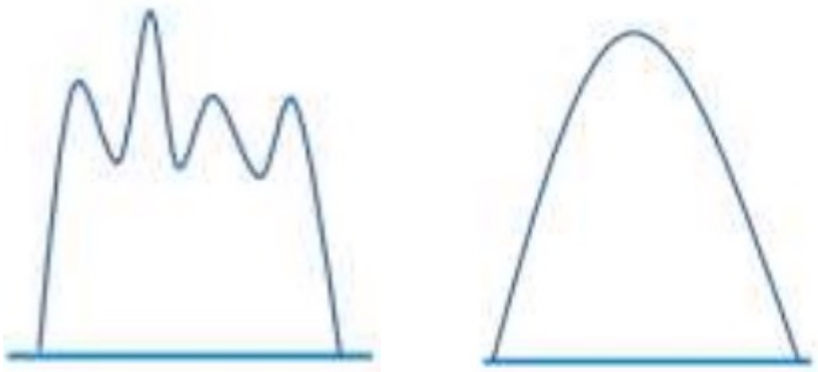
Propiedades de un conjunto de datos: La Tendencia Central

- Tendencia Central según nivel de medición:
  - Para variables de intervalo: Promedio.
  - Para variables ordinales: Mediana.
  - Para variables nominales: Moda.

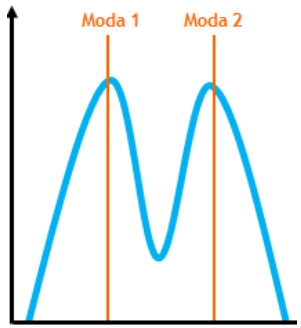
$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Propiedades y alternativas al promedio:
  - Es afectado por distancias (requiere nivel de intervalo) de cada dato a la tendencia central. Eso implica que influyen más en determinar el promedio los casos más extremos.
  - Se puede calcular el promedio ponderado en que cada Xi vale distinto.
  - Se puede evitar los impactos muy fuertes de los valores extremos con el Promedio Recortado (usualmente al 5% o 10%) que quita el 5% o 10% de los valores más extremos (% sumando todo lo eliminado).

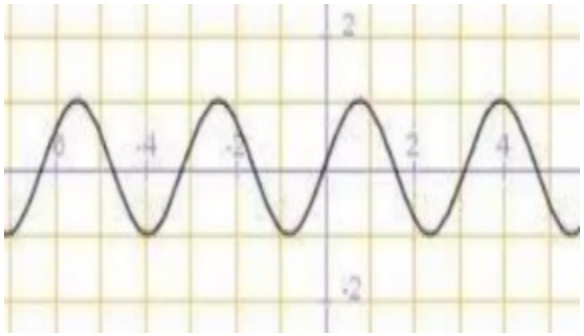
- Definición y propiedades de la Mediana:
  - Es aquel valor de la base de datos que deja a la mitad de los datos de la base sobre él (son mayores que él) y a la otra mitad bajo él (son menores que él).
  - No es afectado por los valores extremos.
  - Se puede usar con variables ordinales y de intervalo.
- Definición y propiedades de la Moda:
  - Es aquel valor más repetido de un conjunto de datos.
  - En un conjunto de datos puede haber una o varias modas.
  - Se puede usar con variables nominales, ordinales y de intervalo.



Unimodales



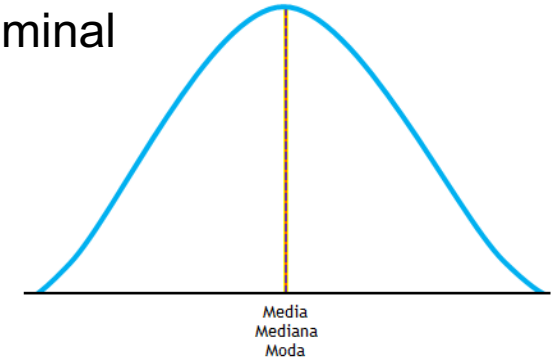
Bimodal



Multimodal

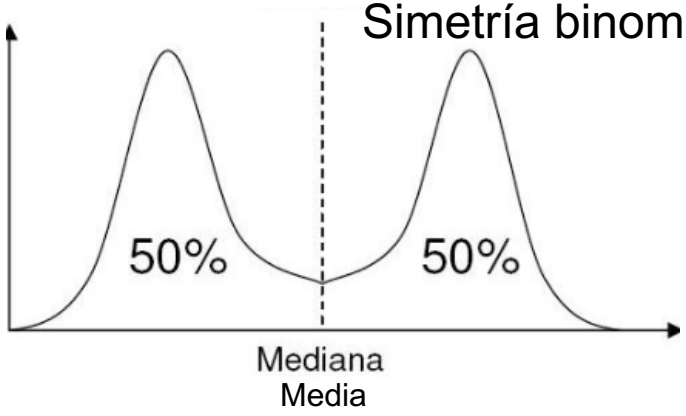
¿Qué implica que la Media, Mediana y Moda sean el mismo valor?

Simetría uninominal

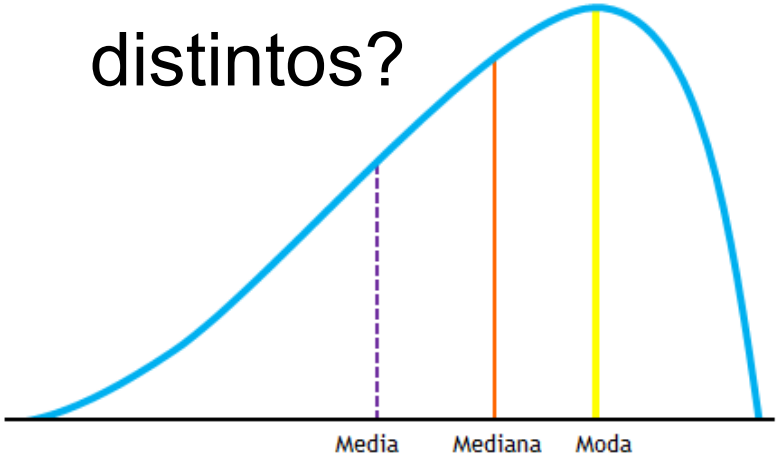


¿Y si sólo Media y Mediana son el mismo valor?

Simetría binominal



- ¿Y si los tres son valores distintos?



Distribución de Ingresos en Chile

Propiedades de un conjunto de datos: Dispersión o heterogeneidad

- Dispersión según nivel de medición:
  - Para variables de intervalo: Varianza o desviación estándar.
  - Para variables ordinales: Rango.
  - Para variables nominales: Nº de Categorías / Distribución de Frecuencias.
- Propiedades de la varianza y desviación estándar:

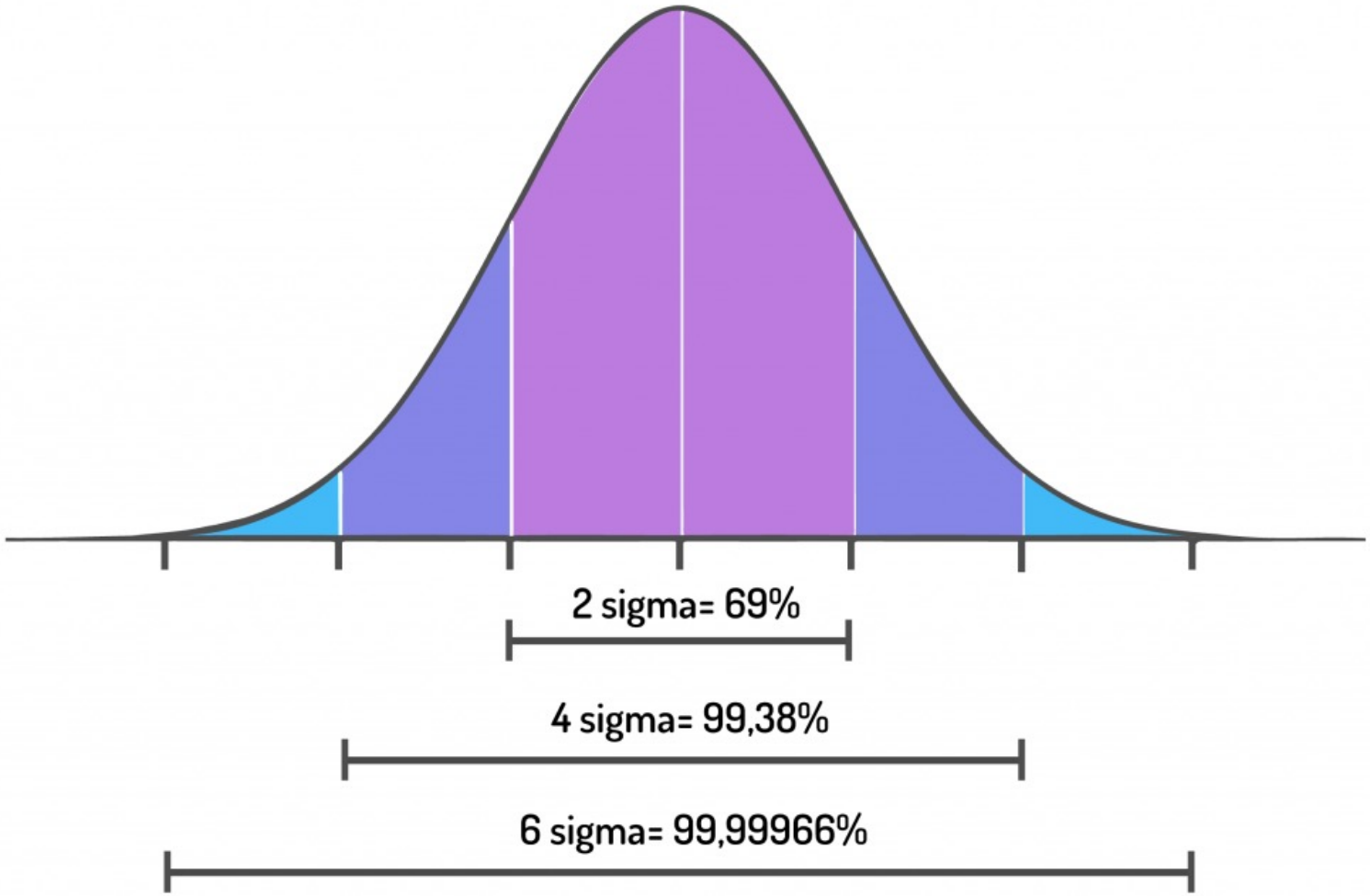
$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \qquad s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Varianza Desviación Estándar

  - Una es el cuadrado de la otra.
  - Se emplean en distintas operaciones estadísticas, por eso existen ambas.

- ¿Qué son la varianza y la desviación estándar?
  - La varianza: Es el promedio de las diferencias al cuadrado de los casos a la media.
  - La desviación estándar: es la raíz cuadrada del promedio de las diferencias al cuadrado de los casos a la media, es decir, la raíz cuadrada de la varianza.
- ¿Cómo se puede interpretar una varianza o desviación estándar?
  - Si una distribución de datos es normal, aproximadamente el 69% de los casos estarán entre una desviación estándar bajo la media y sobre la media.
  - Si una distribución de datos es normal, aproximadamente el 99% de los casos estarán entre dos desviaciones estándar bajo la media y dos sobre la media.
  - Si la distribución no es normal, es más compleja la interpretación, pero siempre se puede usar comparativamente. Si se dispone de dos conjuntos de datos, será más homogéneo el que tenga menor desviación estándar.

Distribución Normal y sus áreas bajo la curva



***Sigma (σ) = Forma poblacional de hablar de s

- Si la distribución no es normal, es más compleja la interpretación, pero siempre se puede usar comparativamente.
- Si se dispone de dos conjuntos de datos, será más homogéneo el que tenga menor desviación estándar.

- Aplicando medias y desviaciones estándar:
  - Si les dijera que dividí el curso en cuatro grupos, cuyas notas tuvieron las siguientes propiedades…
    - Grupo A: Media de notas: 4.8, desviación estándar: 0,3.
    - Grupo B: Media de notas: 5,5, desviación estándar 2,5.
    - Grupo C: Media de notas: 4,7, desviación estándar 4,2.
    - Grupo D: Media de notas: 4,2, desviación estándar 0,5.
  - ¿En que grupo le gustaría estar, en cuál no y por qué?
- El Coeficiente de Variación:
  - Relación entre la desviación estándar y el promedio

$$CV = \frac{\sigma_x}{|\bar{X}|}$$

- Interpretación del Coeficiente de Variación:
  - Usualmente se lo multiplica por 100 para expresarlo en porcentaje.
  - Indica que porcentaje de heterogeneidad tiene un conjunto de datos en relación a su promedio.
  - Se usa para comparar la heterogeneidad de grupos de datos (por ejemplo sueldos de diferentes países) eliminando las distorsiones que produce que en un grupo de datos el promedio sea muy grande y en otro muy pequeño.
- El Rango: Distancia entre el valor más alto y el más bajo.
  - Valor máximo – valor mínimo = Rango
- Cantidad de categorías: indicador de heterogeneidad para variables nominales. Ejemplos:
  - Diferencias de género en encuestas.
  - Grupos étnicos en Chile o Sudáfrica.
  - Idiomas hablados en Nueva Zelanda

- Distribución de Frecuencias:
  - Distribución de frecuencias absolutas.
  - Distribución de frecuencias (proporción o porcentaje) relativas.
  - Distribución de frecuencias acumuladas (absolutas o relativas).

TRANS tipo de transporte

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	1 metro	53	46,5	46,5	46,5
	2 bus	29	25,4	25,4	71,9
	3 tren	13	11,4	11,4	83,3
	4 coche	11	9,6	9,6	93,0
	5 moto	3	2,6	2,6	95,6
	7 otros	5	4,4	4,4	100,0
	Total	114	100,0	100,0	

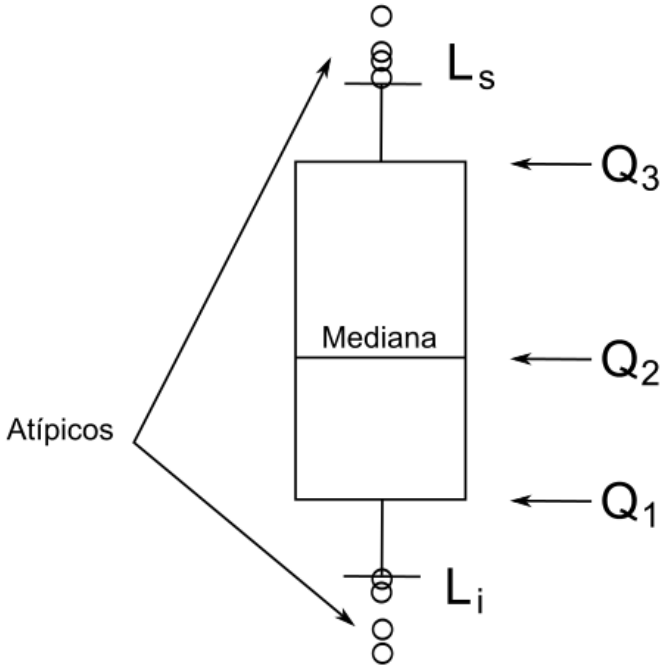
Para variables nominales, no tiene mucho sentido el porcentaje acumulado.

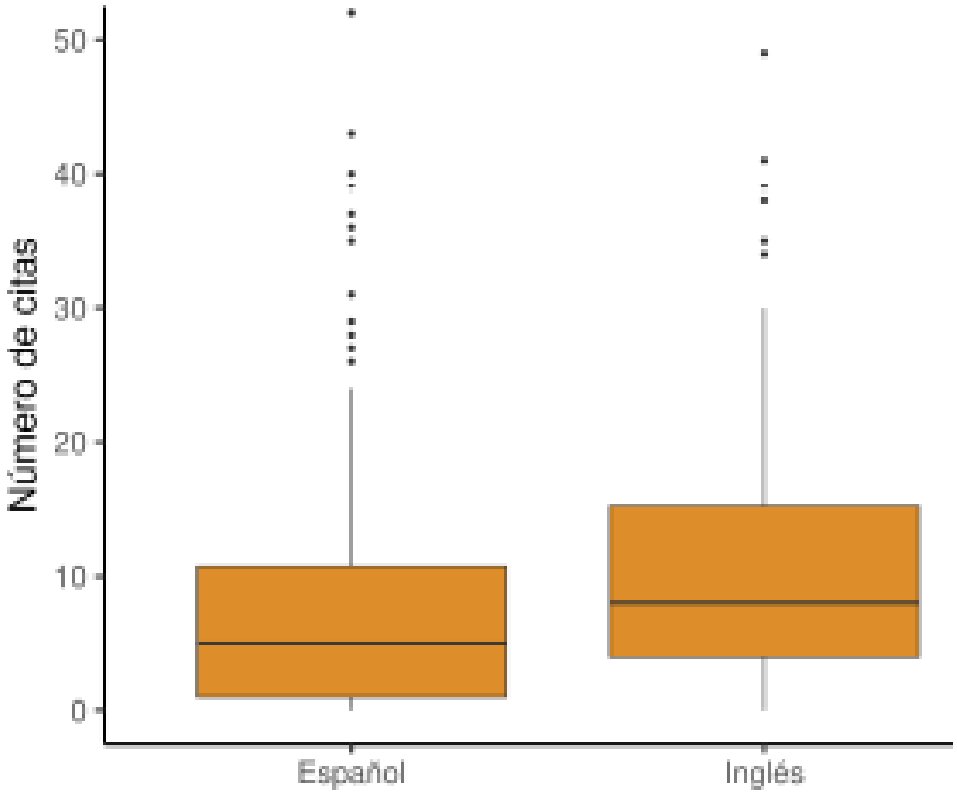
número de teléfonos móviles por hogar familiar	número de familias con esa cantidad	frecuencia absoluta acumulada		frecuencia relativa acumulada		porcentaje acumulado
x_i	f_i	F_i	h_i	H_i	%	% acu.
2	12	12	0,30	0,30	30	30
3	16	28	0,40	0,70	40	70
4	8	36	0,20	0,90	20	90
5	3	39	0,08	0,98	7,5	97,5
6	1	40	0,03	1,00	2,5	100
N =	40					

Acá tiene más sentido el porcentaje acumulado.

Forma de representar la heterogeneidad de una variable (para variables de intervalo)

- Gráfico de caja:

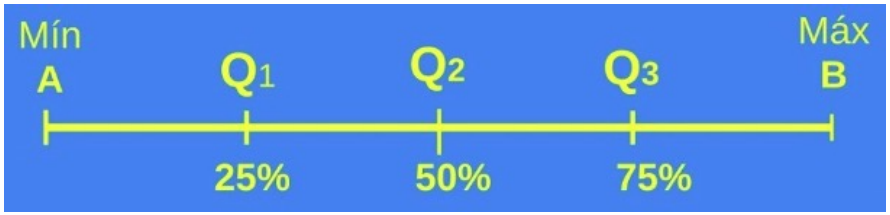




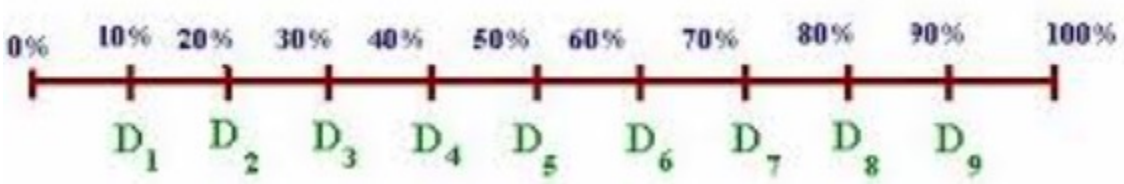
Propiedades de un conjunto de datos: Posición

- Las medidas de posición señalan en que "lugar" de una distribución se encuentra un dato o un conjunto de datos en relación al resto.
- Sólo tienen sentido en variables ordinales y de intervalo (y razón, claro).
- En general se los llama "Cuantiles" y los principales son:
  - Cuartiles.
  - Deciles.
  - Percentiles.

- Los Cuartiles (Q) dividen a un conjunto de datos en 4 grupos en función de su magnitud u orden:
  - En el primer cuartil está el 25% de los valores menores, en el segundo cuartil el siguiente 25% y así…



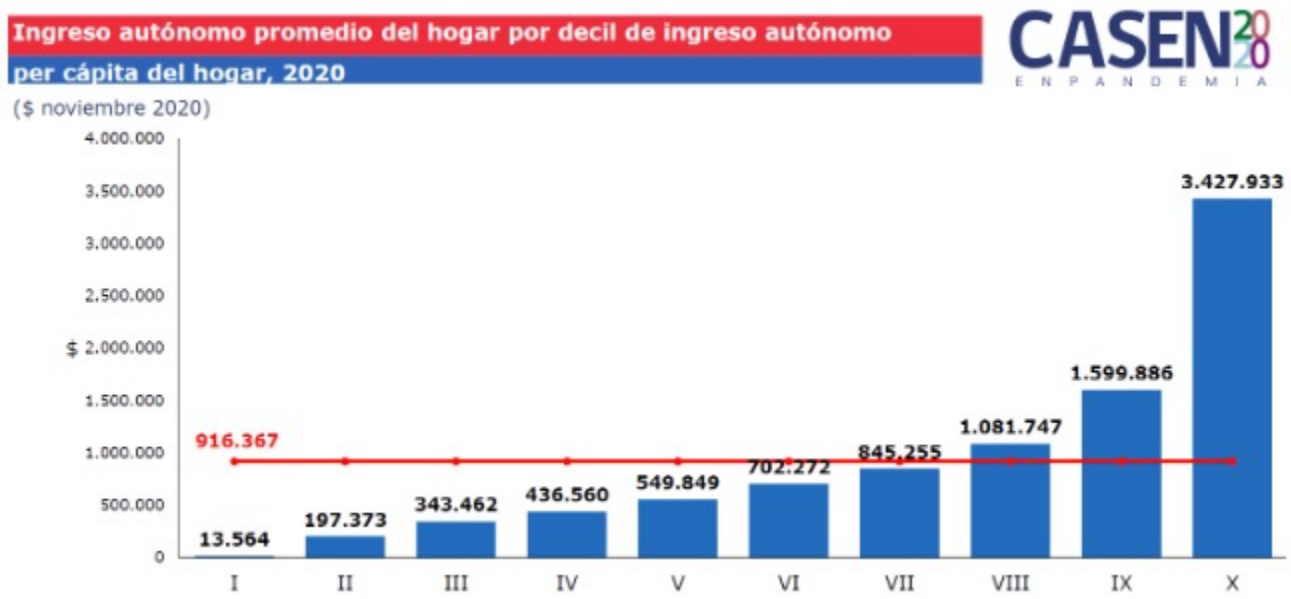
  - Q1 es el valor que separa el 25% menor del siguiente.
  - Q2 es la mediana.
  - Q3 es el valor que separa el 75% más bajo del 25% más alto.
- Los Deciles (D) hacen lo mismo pero dividiendo los datos en 10 grupos.



- Los percentiles (P) dividen a un conjunto de datos en 100 grupos en función de su magnitud u orden.
  - Ejemplo: el percentil 67 (P67) es el valor que divide al 67% de valores que obtuvo menor puntaje, del 34% que obtuvo mejor puntaje.
- Note que siempre hay un cuantil menos que el número de grupos que se quiere construir:
  - Ejemplo: 3 Cuartiles (Q) dividen la muestra en 4 grupos o 99 Percentiles (P) dividen la muestra en 100 grupos.
- ¿Para que sirven los Cuantiles?
  - Para dividir un grupo de datos en grupos de igual tamaño y ordenados.
  - Para determinar en que posición está un dato concreto.

- Ejemplos de uso:
  - Percentiles: Suponga que usted en una prueba de inteligencia queda en el P21, o en el P98… ¿es igual?, ¿cuál lo alegra más?
  - Cuartiles: Imagine que usted quiere dividir un curso en 4 grupos de distinto rendimiento para evaluar el impacto que tiene en cada uno de esos grupos distintas estrategias de enseñanza de la estadística.

Deciles de ingreso en Chile 2020



  - Cuantiles en general: imagine que un país decide dar educación superior gratuita al 60% más pobre del país. ¿Qué requeriría?

FIN