MEDIDAS RESUMEN: Numéricas y Gráficas.

Ejemplo. “Admítelo una salchicha no es una zanahoria”. Así decía la revista ”El Consumidor” en un comentario sobre la baja calidad nutricional de las salchichas.

Hay tres tipos de salchichas:

i. carne vacuna,
ii. mezcla (carne porcina, vacuna y de pollo)
iii. pollo.

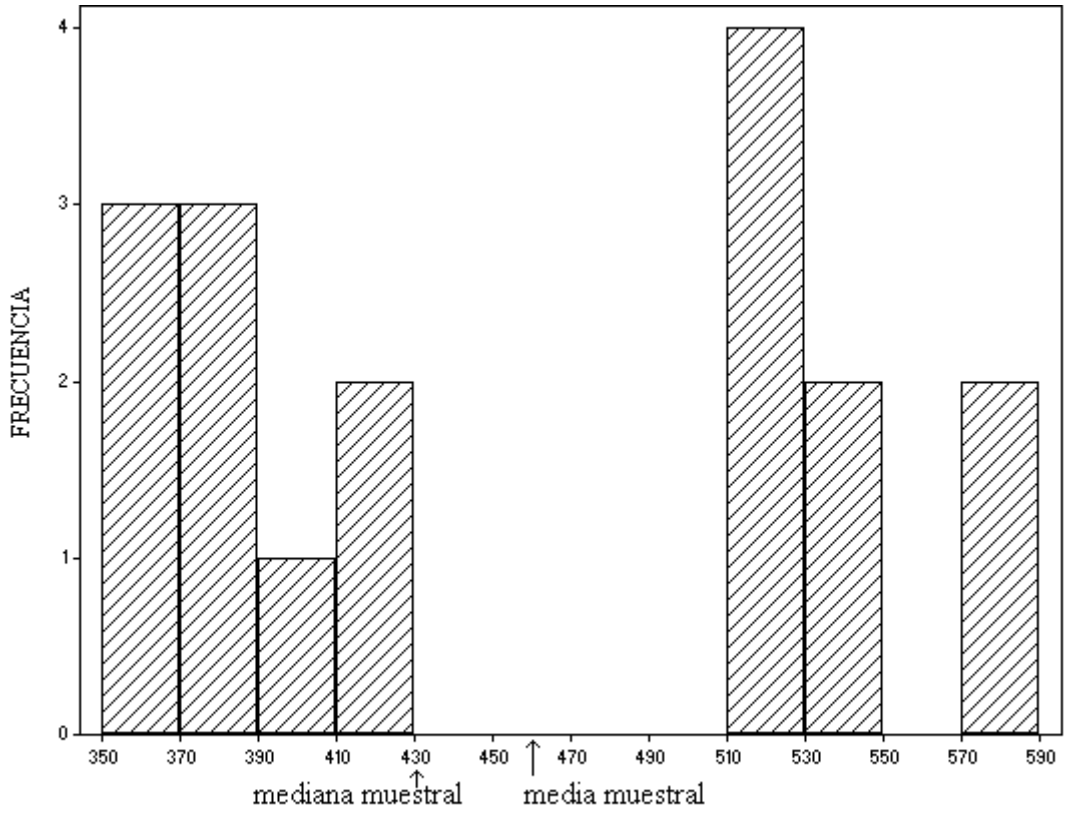
¿Existe alguna diferencia sistemática entre estos tres tipos de salchichas, en estas dos variables?

Calorías y sodio en salchichas por tipo

Vacuno		Mezcla		Pollo	
Calorías	Sodio	Calorías	Sodio	Calorías	Sodio
186	495	173	458	129	430
181	477	191	506	132	375
176	425	182	473	102	396
149	322	190	545	106	383
184	482	172	496	94	387
190	587	147	360	102	542
158	370	146	387	87	359
139	322	139	386	99	357
175	479	175	507	170	528
148	375	136	393	113	513
152	330	179	405	135	426
111	300	153	372	142	513
141	386	107	344	86	358
153	401	195	511	143	581
190	645	135	405	152	588
157	440	140	428	146	522
131	317	138	339	144	545
149	319				
135	296				
132	253				

Nos interesa resumir las características más importantes del conjunto de datos en una pequeña cantidad de números que sean fácilmente interpretables.

La distribución de la cantidad de sodio en las salchichas de pollo muestra dos grupos distintivos. Este tipo de distribuciones no estará bien representada por las medidas resumen.



Cantidad de sodio en salchichas de pollo

Los resúmenes pueden ser muy útiles pero no son los detalles. Generalmente los detalles agregan poco, pero es importante estar preparados para las ocasiones en que sí agregan mucho.

Medidas Resumen.

Las medidas resumen clásicas, para resumir un conjunto de datos de *n* observaciones, $x_1, x_2, \ldots, x_n$,utilizan solamente operaciones aritméticas simples (+, *, raíz cuadrada).

La *media muestral* $\bar{x}$, como medida de la posición del centro de los datos,

$$\bar{x} = \frac{x_1 + \quad + x_n}{n},$$

la *varianza muestral*,

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

ó el desvió estándar

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

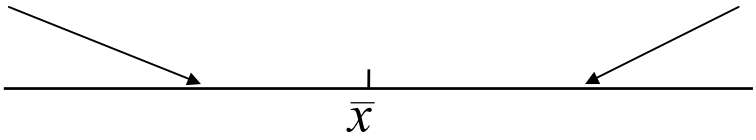
como medida de variabilidad o dispersión.

Observación: el desvío estándar (DS) tiene las mismas unidades que las observaciones.

Desviación de la media

$x_i - \bar{x}$ *desviación i-ésima respecto de la media*.

Desvío negativo: $x_i < \bar{x}$ Desvío positivo: $x_i > \bar{x}$



Si

- todas las diferencias son pequeñas en valor absoluto:
- las observaciones x_i están cerca de $\bar{x}$ ∴ los datos presentan poca variabilidad,
- algún $x_i - \bar{x}$ es grande en valor absoluto se tiene mayor variabilidad.

Es fácil ver que $\sum(x_i - \bar{x}) = 0$.

La varianza muestral mide la desviación cuadrática de los datos respecto de su media, es la medida clásica de variabilidad. Esto se debe a que es más fácil realizar cálculos con desvós cuadráticos, $(x_i - \bar{x})^2$,

que con desvíos absolutos, $|x_i - \bar{x}|$.

Versiones poblacionales, para poblaciones finitas

Si datos son poblacionales tendremos:

- como medida de posición, la *media poblacional* μ que se calcula como

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

- como medida de dispersión, la *varianza poblacional* σ^2

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

ó la raíz cuadrada de σ^2, σ , que llamaremos *desvío estándar*.

.

Población ocupada, República Argentina, Octubre de 1994. Síntesis 3, INDEC, 1995

Aglomerado Urbano	Pobl. Ocup.	Aglomerado Urbano	Pobl. Ocup.
Gran Buenos Aires	4300500	Gran Tucumán y Tafí Viejo	197809
Gran Córdoba	440558	Neuquén	66506
Gran Mendoza	294768	Paraná	66604
Gran Rosario	401203	Santa Rosa - Toay	32286

La cantidad media de ocupados por aglomerado urbano (n=8) es 725029 y su desvío estándar es 1359044.
Si excluimos Gran Buenos Aires (n=7) tendremos media = 214248 y desvío estándar = 155692.
Una sola observación ha modificado fuertemente los resultados.
Las medidas resumen deberían ser resistentes (varíen poco en presencia de un cambio arbitrario de una pequeña parte del lote).

Un único dato aberrante puede producir un importante efecto adverso tanto en la media muestral como la varianza muestral

Medidas resistentes a datos extremos o aberrantes.

Ordenamos los datos, $x_1, x_2, \ldots, x_n$, en orden ascendente y obtenemos la muestra ordenada:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)};$$

Podemos contar desde el más pequeño hacia el más grande, *rango ascendente*, ó desde el más grande hacia el más pequeño, *rango descendente*.

Definición: La *profundidad* de un dato en la muestra es el menor de los rangos ascendente y descendente.

Definición: La *mediana*, M es el valor que deja la misma cantidad de los datos ordenados de cada lado.

La mediana es una medida resistente de posición del centro de los datos.

La profundidad de la mediana es $p_M = \frac{n+1}{2}$.

La mediana se calcula como el valor central si n es impar y promedio de los dos valores centrales si n es par

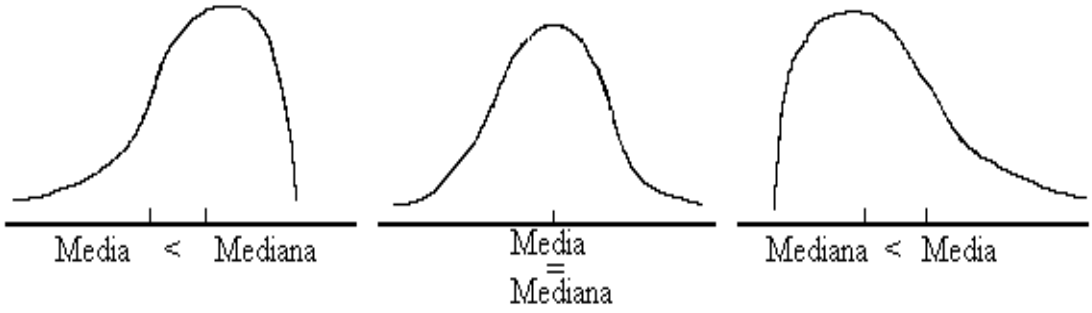
Ejemplo (continuación): La mediana es el dato con profundidad $\frac{59+1}{2} = 30$. M = 63.53.

PROF.	# hojas	TALLO	HOJAS
1	1	628	: 5
1	0	629	:
4	3	630	: 358
7	3	631	: 033
9	2	632	: 77
18	9	633	: 001446669
23	5	634	: 01335

La media, 63.589, es cercana a la mediana. Este hecho es coherente con la simetría que presentan los datos alrededor de

	10	635	: 0000113668	la mediana.
26	7	636	: 0013689	
19	2	637	: 88	Una profundidad identifica *dos* valores de los datos, uno por debajo y otro por encima de la mediana.
17	6	638	: 334668	
11	5	639	: 22223	
6	0	640	:	
6	1	641	: 2	
5	3	642	: 147	
2	0	643	:	
2	2	644	: 02	

Comparación de media y mediana para distintos tipos de distribuciones mediante histogramas suavizados.



Asimétrica a izquierda Simétrica Asimétrica a derecha

Media podada.

Ordene los datos, descarte las 100α% de las observaciones menores y el 100α% de las observaciones mayores; calcule el promedio de los datos restantes. Se recomienda tomar α entre 0.1 y 0.2:

$$\overline{x}_{\alpha} = \frac{x_{[n\alpha]+1} + \quad + x_{n-[n\alpha]}}{n - 2[n\alpha]},$$

Otras medidas de posición.

A la mediana y los extremos les agregamos otro par de valores resumen, los *cuartiles*, que dejan un cuarto y tres cuartos de las observaciones a cada lado.

$$\text{profundidad del cuartil} = \frac{n+1}{4}$$

En el ejemplo, la profundidad del cuartil es $\frac{59+1}{4} = 15$

Por lo tanto: Cuartil inferior=63.36 Cuartil superior=63.84

Otras medidas de dispersión de los datos.

- ***distancia intercuartil*** (d_Q) , o rango intercuartil,
  d_Q = Cuartil superior - Cuartil inferior

- ***rango***, la diferencia entre los valores extremos, también refleja la dispersión pero valores sueltos afectan tanto el rango que su resistencia es despreciable.

MAD: Desvio absoluto respecto de la Mediana: Es una versión resistente del desvío estándar basada en la mediana.

$$\mathbf{MAD} = \mathbf{mediana}\left(\left|x_i - \mathbf{M}\right|\right)$$

¿Cómo calculamos la MAD?

- Ordenamos los datos de menor a mayor.
- Calculamos la mediana, valor en la posición (n+1)/2.
- Calculamos los desvíos absolutos de cada dato repecto de la mediana (la distancia de cada dato a la mediana, sin signo).
- Ordenamos los desvios absolutos de menor a mayor.
- Calculamos la mediana de los desvios.

Observación: Si deseamos comparar la distancia intercuartil y la MAD con el desvío standard es conveniente dividirlas por constantes adecuadas. En ese caso se compara el *DS* con

$$\frac{MAD}{0.675} \qquad \frac{d_Q}{1.35}$$

Siguiendo con el ejemplo de los puntos de fusión de ceras naturales, obtenemos las siguientes medidas resumen:

DESCRIPTIVE STATISTICS

	FUSION
MEAN	63.589
SD	0.3472
MINIMUM	62.850
1ST QUARTI	63.360
MEDIAN	63.530
3RD QUARTI	63.840
MAXIMUM	64.420
MAD	0.2300

d_Q = Cuartil superior - Cuartil inferior = 63.84 - 63.36 = 0.48

$\frac{MAD}{0.675}$ = 0.23 / 0.675 = 0. 341

$\frac{d_Q}{1.35}$ = 0.48 / 1.35 = 0.356

SD = 0.3472

Veremos más adelante qué características deben presentar los datos para que las tres medidas de dispersión sean similares, como ocurre en el ejemplo.

Más medidas de posición: Percentiles.

La mediana de un conjunto de datos ordenados es el valor que los divide en dos partes iguales, tiene profundidad (n+1)*0.5. Es el *percentil* del 50% (100*0.5%).
El cuartil inferior, que deja a su izquierda al 25% de los datos y se encuentra en la posición (n+1)*0.25, es el percentil del 25% (100*0.25%). El cuartil superior, tiene la posición (n+1)*0.75.
Así, el valor que deja un 95% de los datos por debajo y un 5% por encima es el percentil del 95%.

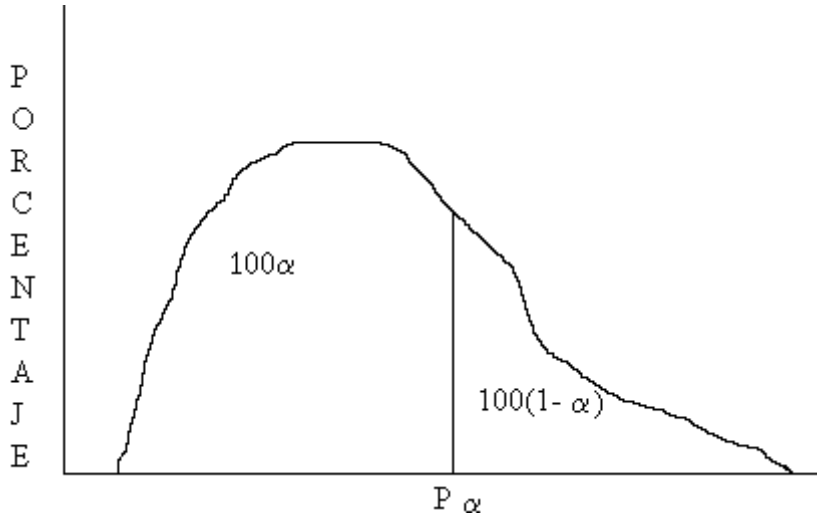


Gráfico de un percentil en un histograma suavizado.

El percentil del 100*α%, P_α, de un conjunto de datos ordenados, es el valor que deja un 100*α% de los datos por debajo y un 100*(1-α)% por encima se encuentra en la posición (n+1)* α. Cuando este valor no es entero se interpola.

Percentiles de la altura (cm) de mujeres y varones de 18 años (Crecimiento y Desarrollo. Sociedad Arg. de Pediatría. 1986)

Percentil	3%	10%	25%	50%	75%	90%	97%
Varón	1.60	1.64	1.68	1.72	1.77	1.81	1.85
Mujer	1.49	1.53	1.56	1.60	1.64	1.68	1.72

En distribuciones perfectamente simétricas los percentiles del 100*α% y del 100*(1-α)% equidistan de la mediana.