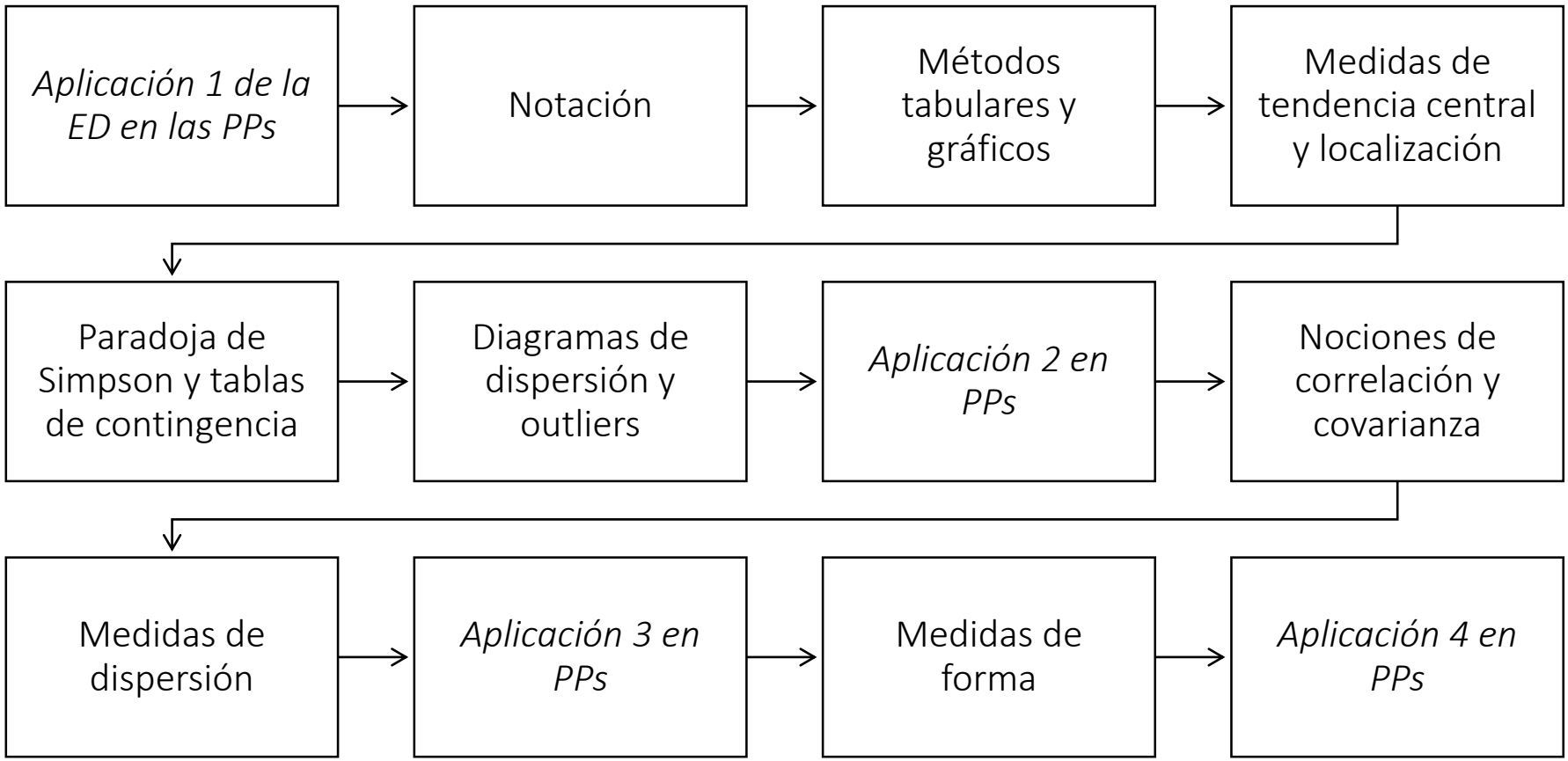
Estadística Descriptiva

Sesión 2

Dr. Carlos J. Vilalta (DAP)
Estadística - Curso propedéutico MAPP 2016
Centro de Investigación y Docencia económicas (CIDE)
Contacto: carlos.vilalta@cide.edu

Lo que vamos a ver hoy

- Asumiendo que leyeron



Utilidad en Políticas Públicas

- Merino, M. y Vilalta, C. 2014. La desigualdad de trato en el diseño del gasto público federal mexicano. Conapred y CIDE.
  - Medir la desigualdad de trato presupuestal por 3 vías: Inequidad, invisibilidad y exclusión presupuestal. Revisión del PEF en 40 sus ramos. Enfoque en poblaciones vulnerables.
  - Liga: http://www.conapred.org.mx/documentos_cedoc/La_desigualdad_Dtrato_CIDE_INACSS.pdf

- Mucha gente piensa que un incremento igual al presupuesto es lo más justo
  - Porque el incremento es igual para todos
  - Porque lo más sencillo es un incremento % igual para todos
  - Si tu presupuesto sube un 5%, puedes subir el gasto en 5%...

- Pregunta: ¿un incremento porcentualmente igual para todos es lo mejor?
- ¿Aumenta, mantiene o reduce las desigualdades?
- Veamos qué nos dice la estadística descriptiva…

Utilidad en Políticas Públicas

- Mediciones de desigualdad:

	SEP: recursos ejercidos en 2012 (millones)
Niñas y niños	18,565
Indígenas	3,032
Mujeres	272
Personas con discapacidad	30
Total	21,899
Media aritmética	5,475
Rango	18,535
Varianza	78,011,861
Desv. Est.	8,832

Mediciones de desigualdad

- Ahora apliquemos el 5% de incremento igualitario para todos…

Utilidad en Políticas Públicas

- Resultados del incremento "igualitario" para todos:

	SEP: recursos ejercidos en 2012 (millones)	Incremento porcentual	Ahora (millones)	Incremento absoluto (millones)	Incremento porcentual
Niñas y niños	18,565	5.0%	19,493	928	-
Indígenas	3,032	5.0%	3,184	152	-
Mujeres	272	5.0%	286	14	-
Personas con discapacidad	30	5.0%	32	2	-
Total	21,899	5.0%	22,994	1,095	-
Media aritmética	5,475	-	5,748	274	5.0%
Rango	18,535	-	19,462	927	5.0%
Varianza	78,011,861	-	81,912,454	3,900,593	5.0%
Desv. Est.	8,832	-	9,274	442	5.0%

- El incremento "igualitario" aumentó la desigualdad entre grupos
- Los niños y niñas se llevaron el 84.8% del total del incremento y las personas con discapacidad el 0.1% del incremento
- Y la desigualdad de trato se incrementó en 5%...

Más sencillo…

- Un incremento salarial en el CIDE a 2 personas

	Salario	Incremento porcentual	Ahora	Incremento absoluto	Incremento porcentual
1 Profesor	20,000	3.1%	20,620	620	-
1 Empleado en serv. gral.	6,000	3.1%	6,186	186	-
Total	26,000	3.1%	26,806	806	-
Media aritmética	13,000	-	13,403	403	3.1%
Rango	14,000	-	14,434	434	3.1%
Varianza	98,000,000	-	101,038,000	3,038,000	3.1%
Desv. Est.	9,899	-	10,206	307	3.1%

- Con un incremento inflacionario “igualitario” de 3.1%, ahora la diferencia salarial entre ambos empleados es mayor…
- Esto sólo lo pudimos ver a través de las medidas de dispersión o desigualdad

Notación

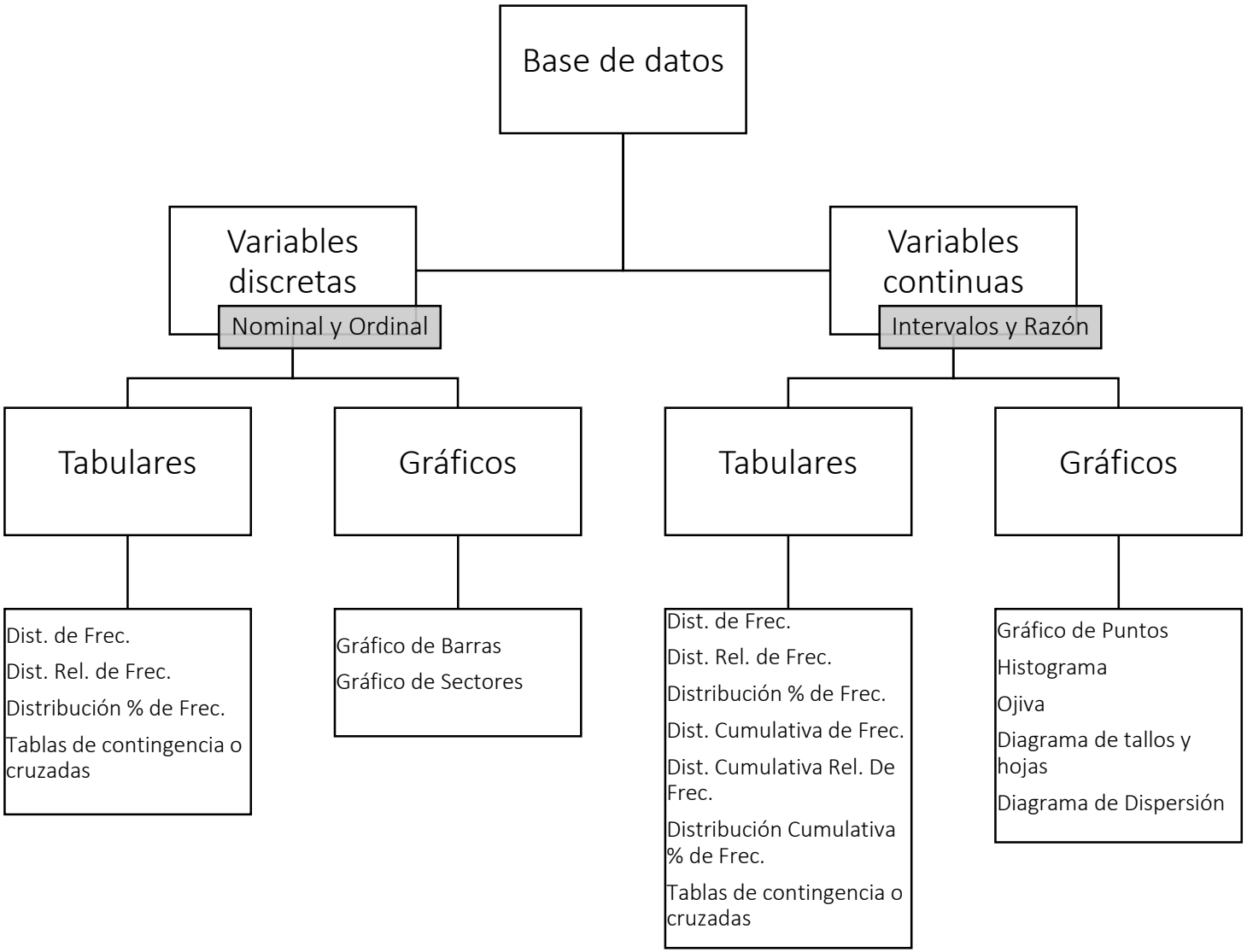
- Población vs. muestra

Población	vs.	Muestra
Alfabeto griego		Alfabeto romano/latino
μ, σ^2, σ, β		M, s^2, s, b

Variables	vs.	Constantes
Últimas letras		Primeras letras
... x, y, z		a, b, c…

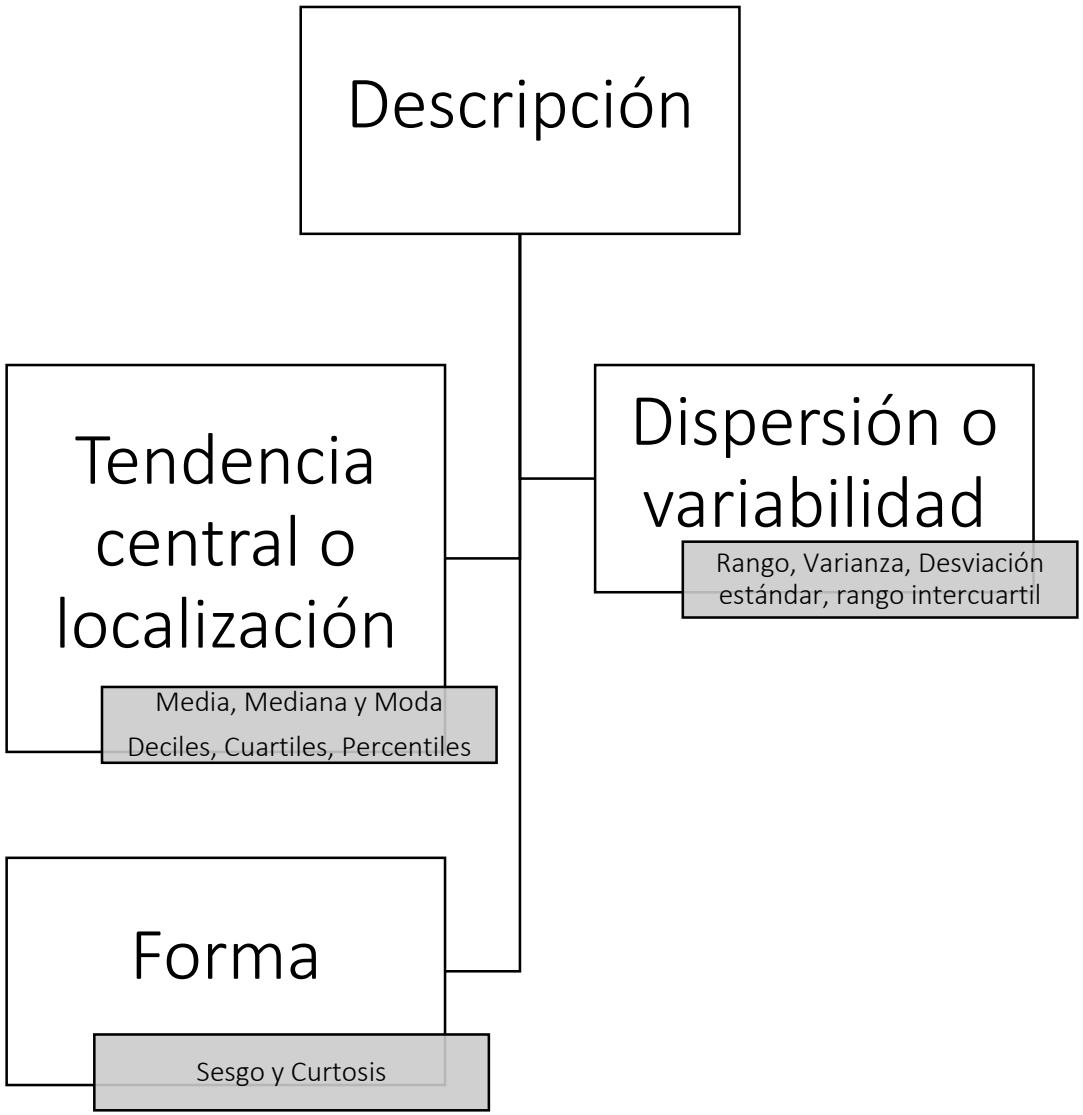
Resumen

- Métodos tabulares y gráficos para transformar datos en información



Resumen

- Estadísticos descriptivos



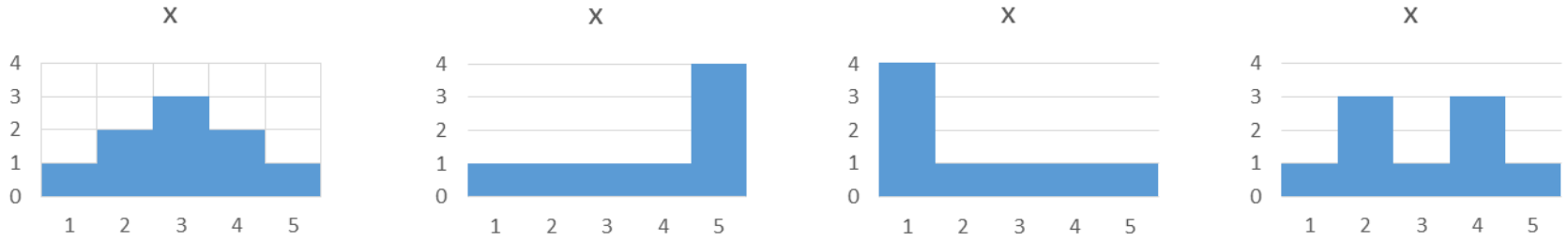
1. Medidas de tendencia central o localización (MTC)

Medidas de tendencia central

- Media aritmética: lo que todos tendríamos si todos fuéramos iguales
  - Ventajas: fácil de calcular, fácil de entender, representa el valor más cercano a todos y es ampliamente utilizada en estadística inferencial paramétrica
  - Desventajas: sensible a *outliers*
- Mediana: lo que tiene el que está a la mitad del grupo
  - Ventajas: fácil de calcular, fácil de entender y no sensible a *outliers, c*onveniente en distribuciones asimétricas
  - Desventajas: no pondera cada valor por el número de veces que aparece repetido y no usa todos los datos con lo que se pierde información
- Moda: el valor(es) más frecuente
  - Ventajas: fácil de calcular, fácil de entender y no sensible a outliers
  - Desventajas: no siempre existe y no utiliza todos los datos
- Cuartil: uno de los 3 puntos en que se divide una distribución para visualizar 4 grupos al interior de la misma (1er cuartil=25% de los datos etc.)
  - 2do cuartil = mediana
  - 5to decil = mediana
  - 50 percentil = mediana

Relación entre Media arit., Mediana y Moda

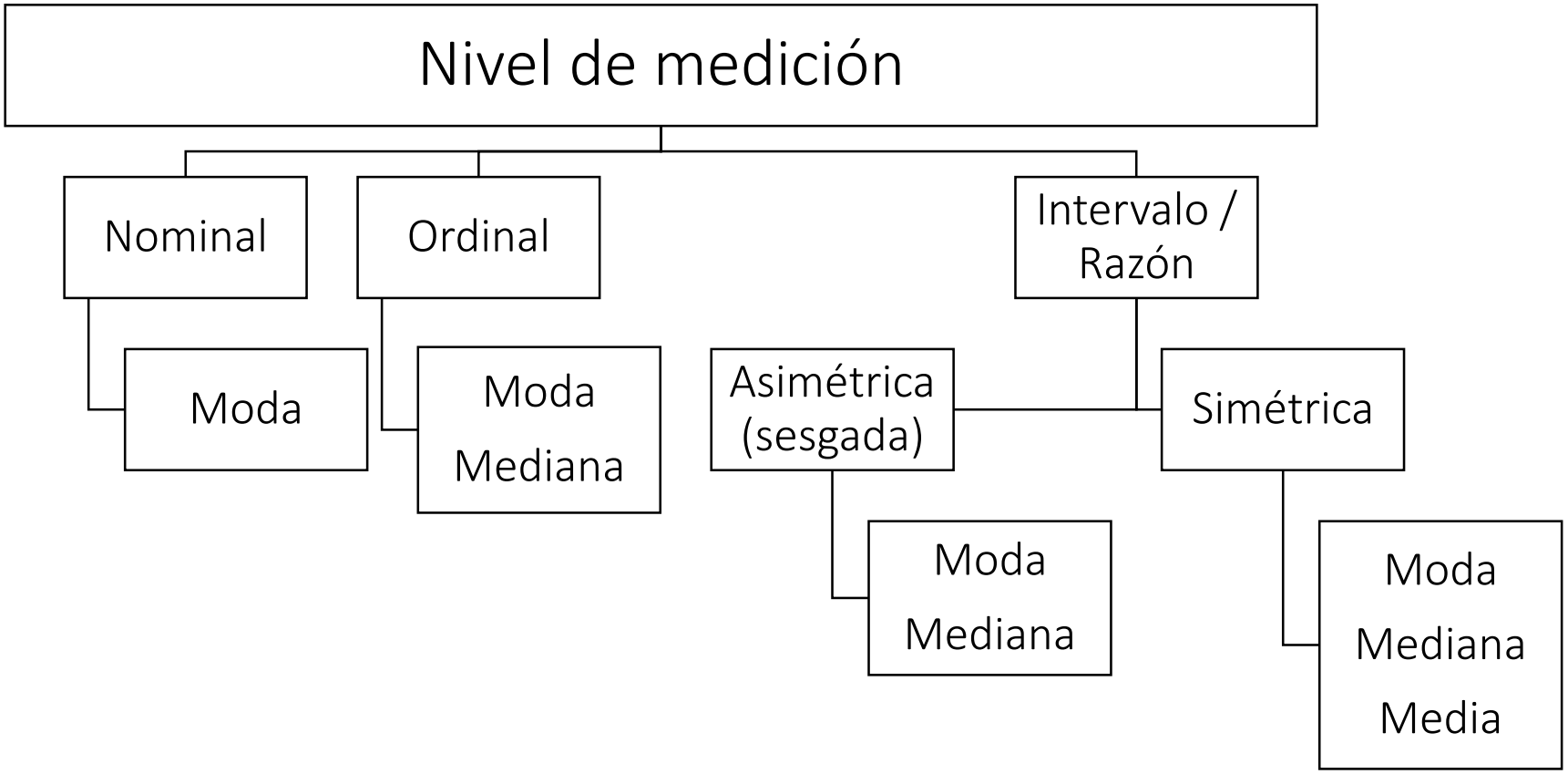
- Distribución simétrica: son iguales
- Distribución sesgada a la derecha: media > mediana
- Distribución sesgada a la izquierda: mediana > media



x	Simétrica f_i	Sesgada izqu. f_i	Sesgada der. f_i	Bimodal f_i
1	1	1	5	1
2	2	1	1	3
3	3	1	1	1
4	2	1	1	3
5	1	5	1	1
Media	**3.0**	**3.9**	**2.1**	**3.0**
Mediana	**3**	**5**	**1**	**3**
Moda*	**3**	**5**	**1**	**2**
Media geométrica	**2.7**	**3.5**	**1.7**	**2.7**

*Pueden haber varias modas. Por convención se reporta la menor e igualmente se reporta (por ejemplo con un asterisco) que hay otras modas en la distribución

Cómo elegir una MTC



MTC y niveles de medición

- A cada nivel de medición corresponde una medida de tendencia central
- Mediciones de igualdad o similitud: cómo se concentran los datos

Nivel de medición / tipo de estadístico descriptivo	Nominal	Ordinal	Intervalo o razón*
Media aritmética	No	No	Si
Media geométrica	No	No	Si
Mediana	No	Si	Si
Moda	Si	Si	Si

*Revisar si la distribución es asimétrica o sesgada

¿La media de una variable ordinal?

- Caso: Opinión/evaluación de “X” tema en una escala de 1 a 5
  - ¿Qué significan o cómo puedo interpretar las fracciones de algo en una escala discreta?
  - Al no haber “cero” ¿puedo/debo decir que una evaluación de “4” es el doble mejor que una evaluación de “2”? ¿Es “Mal” la mitad de “Bien”? ¿1 “Bien” equivale a 2 “Mal”?

Donde:	5 = Muy bien; 4 = Bien; 3 = Regular; 2 = Mal; 1 = Muy mal

	Evaluación	Significado
A	5	Muy bien
B	5	Muy bien
C	5	Muy bien
D	3	Regular
E	1	Muy mal

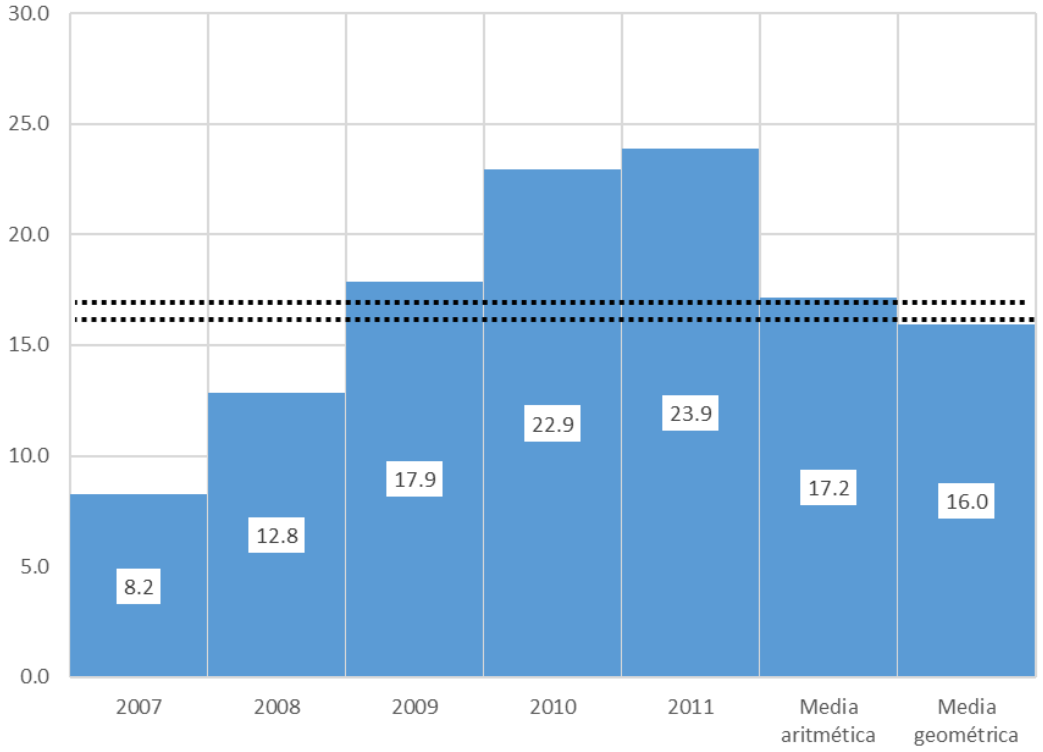
Media	3.8	Entre regular y bien, tirando a bien… (¿y si redondeo? Ojo: nadie dio un 4)
Mediana	5	Muy bien
Moda	5	Muy bien

Media geométrica

- Media geométrica: raíz del producto de los valores vs. su suma
  - Ventajas: en el análisis de tendencias es menos sensible a los outliers (observaciones atípicas o valores extremos)
  - Desventajas: No aplicable con valores negativos o un valor de “0”

Media aritmética ≥ Media geométrica

	México: tasa de homicidios	Cambio anual
2007	8.2	
2008	12.8	55.7%
2009	17.9	39.4%
2010	22.9	28.2%
2011	23.9	4.1%
Media aritmética	17.2	31.9%
Media geométrica	16.0	22.5%
Rango	15.6	51.6%
Desv. Est.	6.7	21.6%



Paradoja de Simpson y tablas cruzadas

- Caso especial del efecto de variable “omitida” en un modelo causal
- Una consideración básica al realizar comparaciones es saber si una tercera variable explica el patrón observado en los datos

ROBO * genero Crosstabulation

% within genero

		genero		
		masculino	femenino	Total
ROBO	Robo simple	45.4%	57.8%	46.6%
	Robo con violencia	54.6%	42.2%	53.4%
Total		100.0%	100.0%	100.0%

ROBO * genero * estado Crosstabulation

% within genero

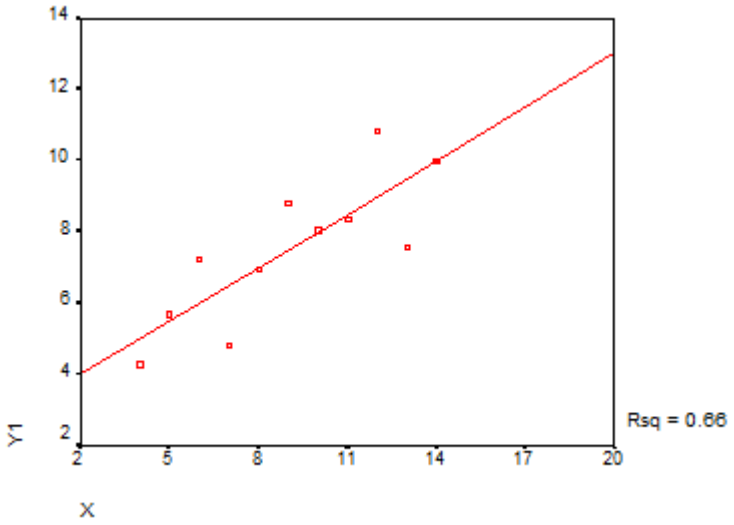
			genero		
estado			masculino	femenino	Total
distrito federal	ROBO	Robo simple	56.5%	67.6%	57.8%
		Robo con violencia	43.5%	32.4%	42.2%
	Total		100.0%	100.0%	100.0%
estado de mexico	ROBO	Robo simple	30.4%	38.6%	31.1%
		Robo con violencia	69.6%	61.4%	68.9%
	Total		100.0%	100.0%	100.0%

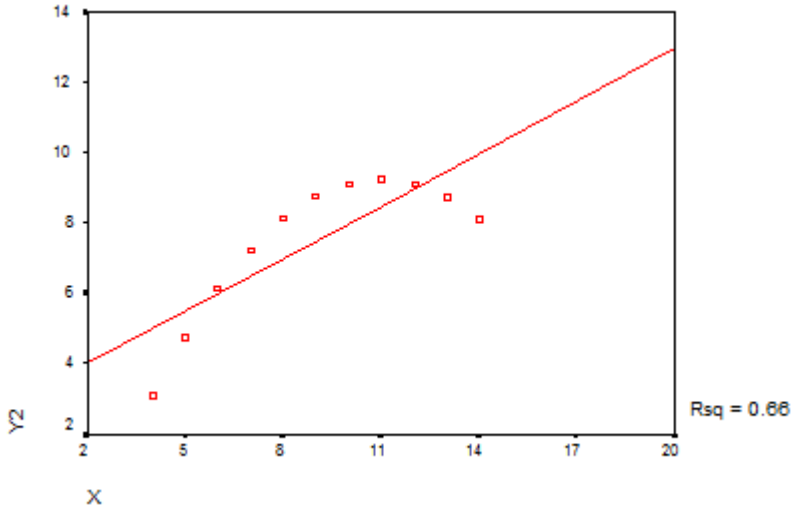
Detección de outliers

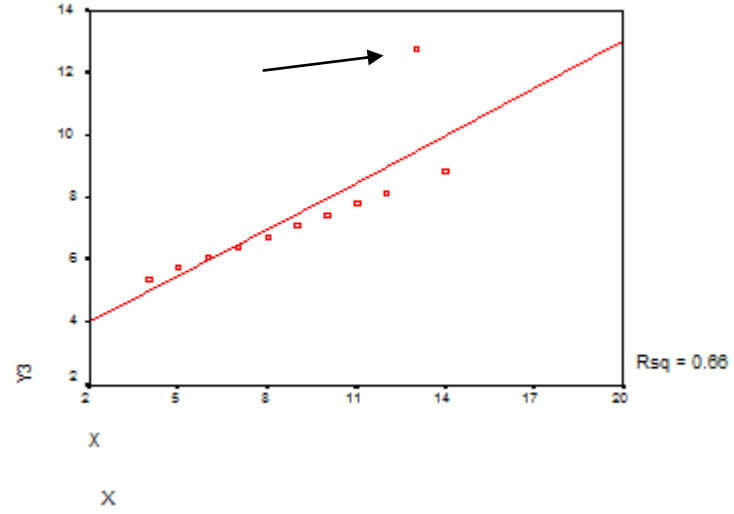
- Hay varios motivos por los que existen:
  - Error de captura
  - Error de encuesta: pertenece a otra Población no objetivo de la encuesta
  - Observación legítima por lo que merece análisis aparte
- Hay varias formas de detectarlos:
  - Valores Z y aplicando el teorema de Chebyshev (previa vista de la distribución)
    - Puntuaciones Z o estándar: se sustrae la media de cada observación y se divide entre la desv. Est.
    - Nos dice a qué distancia en desviaciones estándar se encuentra cada observación
  - Pruebas de valores extremos: G de Grubb

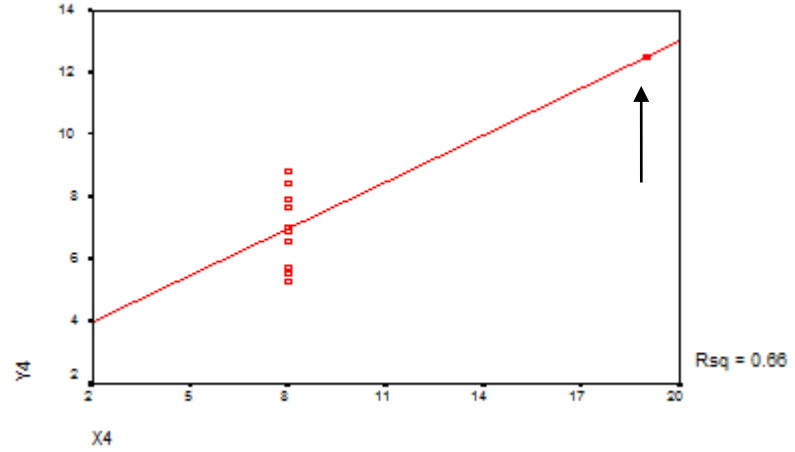
Diagramas de dispersión y outliers

- El efecto de los outliers (observaciones atípicas) y relaciones NO lineales









Nociones de covarianza y correlación

- Covarianza: mide la fuerza de la relación entre 2 variables
  - Desventajas: afectada por outliers y no es estandariza porque usa las unidades (métrica) originales de medición lo que no permite la comparabilidad (varía entre $-\infty$ e ∞)
- Coeficiente de correlación: mide la fuerza de la relación entre 2 variables. Es la covarianza divida entre el producto de la desviación estándar de las 2 variables
  - Ventaja: estandariza las unidades de las 2 variables y varía entre -1 y 1 (r de Pearson y Rho de Speaman solamente)
  - Desventaja: difícil de calcular y es afectada por outliers

2. Medidas de dispersión o variabilidad

Medidas de dispersión

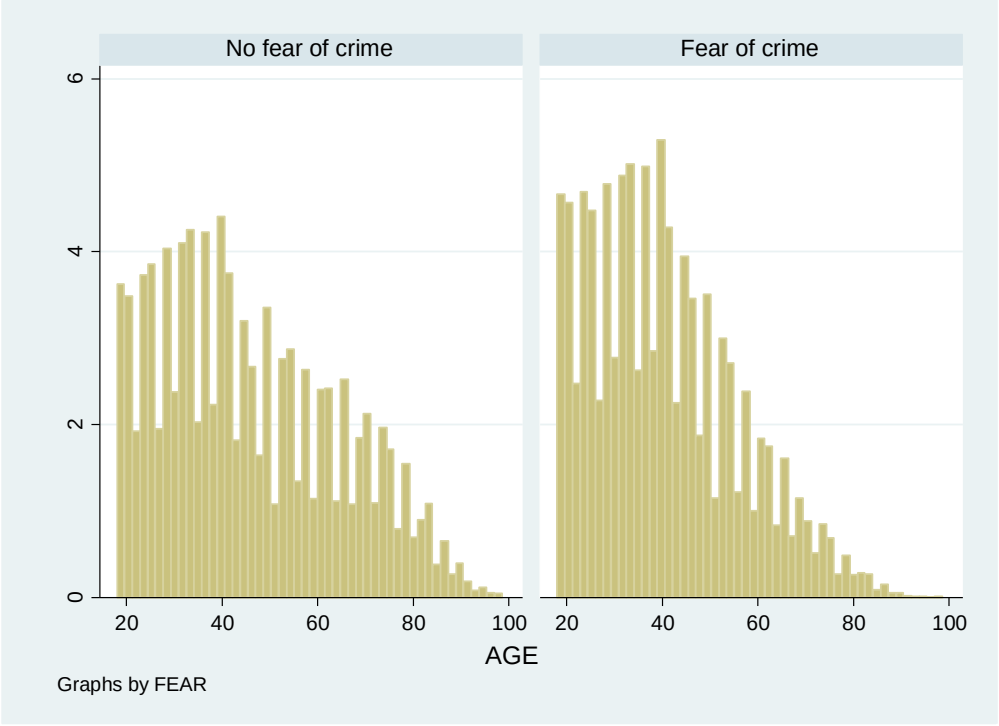
- Rango: la distancia entre el mayor y el menor valor
  - Ventajas: fácil de calcular y fácil de entender
  - Desventajas: sólo utiliza 2 observaciones por lo que pierde información y es sensible a outliers
- Varianza: la media aritmética de las diferencias cuadráticas a la media
  - Ventajas: utiliza todos los datos en la distribución
  - Desventajas: difícil de entender, no aplicable analíticamente porque tiene una escala cuadrática de la variable de interés y sensible a outliers
- Desviación estándar (σ, s):
  - Ventajas: fácil de calcular, fácil de entender, tiene la misma escala que la variable de interés, y ampliamente aplicable en estadística inferencial paramétrica
  - Desventajas: sensible a outliers
- Rango intercuartil: dice la porción central de la distribución (25% y 75% de los datos)
  - Ventajas: relativamente fácil de entender y no sensible a outliers
  - Desventajas: difícil de calcular y pierde los extremos de la distribución en su cálculo
- Coeficiente de variación: media de la “s” o variabilidad relativa (%)
  - Ventajas: para comparar variabilidad/dispersión de variables diferentes
  - Desventajas: no se puede calcular si M = 0

Inferir con la Desviación Estándar

- Teorema de Chebyshev:
  - Para cualquier variable aleatoria “X” con media “μ” y desviación estándar “σ” la probabilidad de que “X” tome un valor dentro de “k” desviaciones estándar de la media aritmética es cuando menos = $1 – 1/k^2$
  - Regla general:
    - Al menos el 75% de los datos se hallan entre +/- 2s de M
    - Al menos el 89% de los datos se hallan entre +/- 3s de M
    - Al menos el 94% de los datos se hallan entre +/- 4s de M
  - Para una distribución normal estándar:
    - Aproximadamente el 68% de los datos se hallan entre +/- 1s de M
    - Aproximadamente el 95% de los datos se hallan entre +/- 2s de M
    - Aproximadamente el 99% de los datos se hallan entre +/- 3s de M

Una aplicación en Política Pública

- Definir rango de edad más adecuado para promover una campaña de medios que reduzca la inseguridad en el transporte público
- ¿Qué forma del teorema de Chebyshev debería aplicar? ¿regla general o el aplicable a una distribución normal?



-> RISK = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
AGE	24289	45.86866	18.58248	18	97

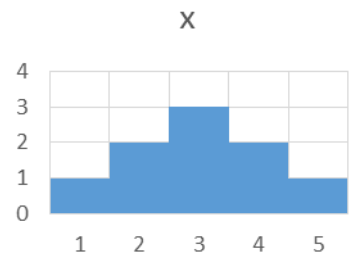
-> RISK = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
AGE	59620	40.37895	15.30448	18	97

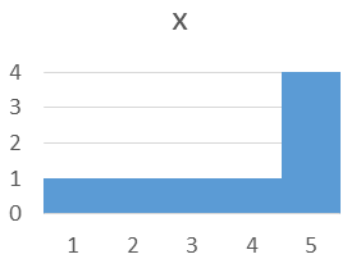
3. Medidas de forma: sesgo y curtosis

Sesgo

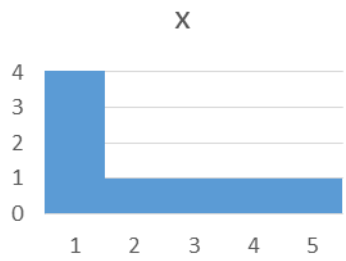
- Mide la asimetría de una distribución
- Valor positivo: sesgo a la derecha (mayoría por debajo de la media)



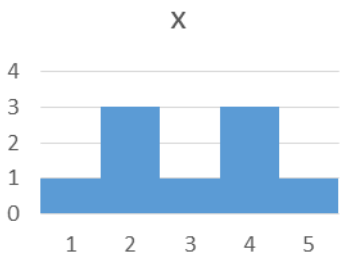
Sesgo = 0



Sesgo negativo



Sesgo positivo



Sesgo = 0

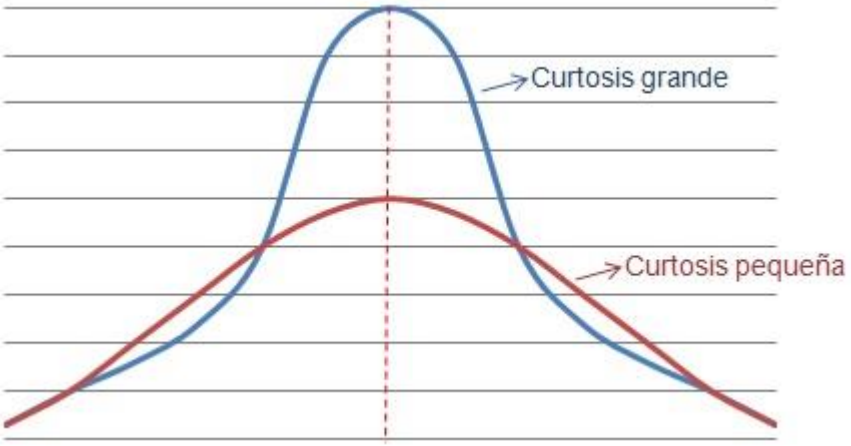
x	Simétrica f_i	Sesgada izqu. f_i	Sesgada der. f_i	Bimodal** f_i
1	1	1	5	1
2	2	1	1	3
3	3	1	1	1
4	2	1	1	3
5	1	5	1	1
Media	**3.0**	**3.9**	**2.1**	**3.0**
Mediana	**3.0**	**5.0**	**1**	**3.0**
Varianza	**1.5**	**2.4**	**2.4**	**1.8**
Desviación estándar	**1.22**	**1.54**	**1.54**	**1.32**
Sesgo	**0.00**	**-1.09**	**1.09**	**0.00**

*Si hay varias modas, por convención se reporta la menor e igualmente se reporta (por ejemplo con un asterisco) que hay otras modas en la distribución

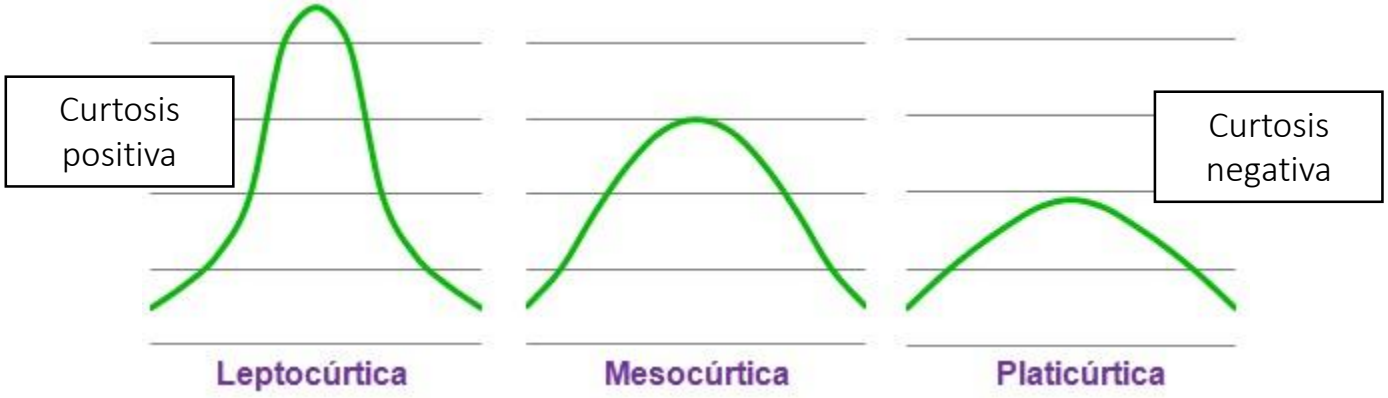
**Nota: en este ejemplo es simétrica, por ende sesgo = 0

Curtosis

- Mide cuán aguda o achatada está una distribución
  - Mayor curtosis: mayor concentración de valores alrededor de la media
  - Menor curtosis: menor concentración de valores alrededor de la media
  - Valores positivos y negativos: más aguda que una dist. normal estándar y viceversa

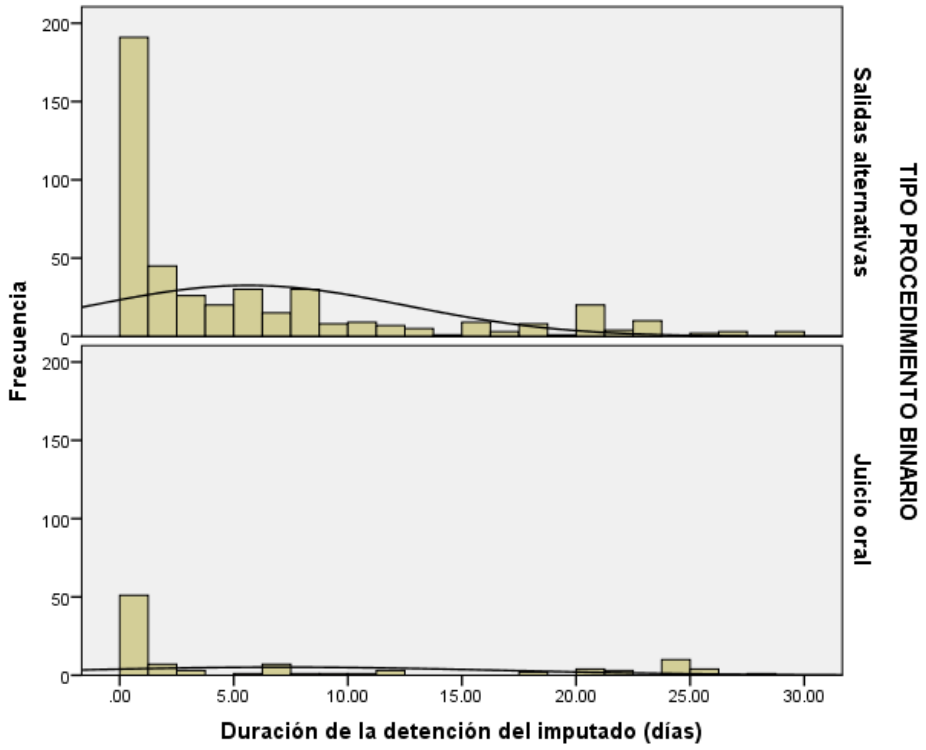


Ojo: al igual que el sesgo, es sensible a outliers



Una aplicación en Política Pública

- Duración de la detención: comparación entre juicios orales y salidas alternativas. P: ¿Quiénes pasan más días detenidos?



	Salidas alternativas	Juicio oral
Media	5.6	7.1
Mediana	2	1
Desviación estándar	6.9	9.6
Sesgo	1.57	1.01
Curtosis	1.47	-0.68

Juicio Oral:

- Mayor media
- Menor mediana
- Más variabilidad entre detenidos
- Menos sesgo
- Sesgada a la derecha

Pero…

- La mayoría por debajo de la media, y
- Pocos alrededor de la media

Tarea

- Anderson et al.: Resolver ejercicios (entregar a mano y con procedimientos):
  - Medidas de tendencia central (Cap. 3): 1 a 12
  - Medidas de dispersión (Cap. 3): 13 a 24
  - Suplementarios (Cap. 3): 58, 59, 61 y 62
- Investigar: ¿por qué el denominador de la varianza de la Muestra usa una corrección (n-1) por tamaño de muestra?
- Entrega: jueves en el laboratorio