Estadística (repaso)

Camilo Vargas Walteros

La Estadística

En el mundo antiguo

- *FUENTE: www.franceinter.fr/emissions/la-marche-de-l-histoire/la-marche-de-l-histoire-13-novembre-2014*

La Estadística

- La **"Estadística"** era utilizada en el mundo antiguo por parte de los gobernantes para registrar datos de la población (censos).
- La descripción de la información en forma resumida hace parte de la **"Estadística Descriptiva"**.
- Mientras que en las ciencias naturales se utilizan **"datos experimentales"** en las ciencias sociales se emplean **"datos no experimentales"**.
- Se utilizan diferentes tipos de información como series de tiempo, corte transversal y datos combinados.

1. Introducción

1.1 Series de tiempo

Tipos de información

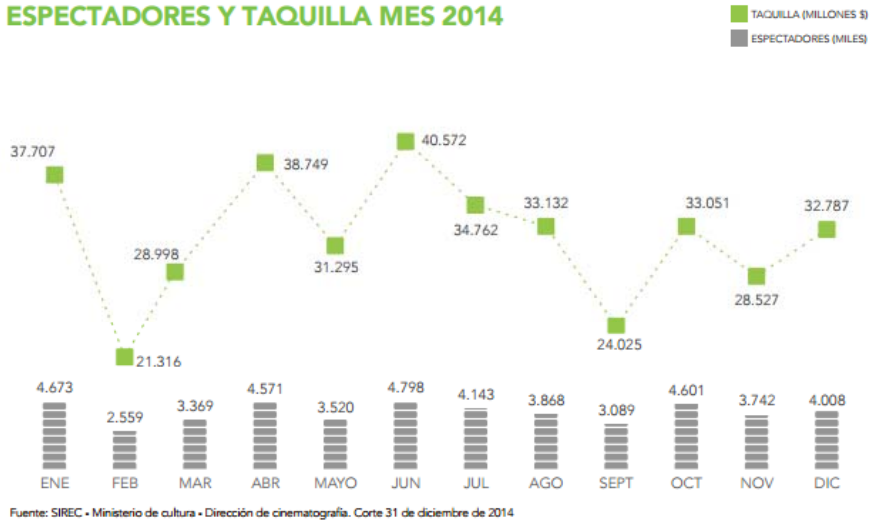
Series de tiempo

- El orden de los datos es importante.
- La historia y el ciclo de cada variable son importantes.
- Cada variable tiene una frecuencia específica.
- Los datos pueden ser **"cuantitativos"** (tipo de cambio) o **"cualitativos"** (días de la semana).

Variable	Frecuencia	Entidad
Precio de acciones	Continua	BVC
Agregados Monetarios y DTF	Semanal	Banco de la República
Inflación y Desempleo	Mensual	DANE
PIB y Déficit Fiscal	Trimestral	DANE / Min Hacienda
Población	Anual	DANE

Tipos de información

Serie de tiempo (taquilla y espectadores 2014)



Fuente: SIREC - Ministerio de cultura - Dirección de cinematografía. Corte 31 de diciembre de 2014

- *FUENTE: www.mincultura.gov.co (Anuario Estadístico Cine Colombiano 2014, P16)*

Tipos de información

Serie de tiempo (valor de mercado 2010 – 2018)

- *FUENTE: www.transfermarkt.co.uk/james-rodriguez/marktwertverlauf/spieler/88103*

Tipos de información

Serie de tiempo

obsno	año	prommin	coverprom	desempl	PNB
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
⋮	⋮	⋮	⋮	⋮	⋮
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

- *FUENTE: Wooldridge (2010), Introducción a la Econometría, P 9, Tabla 1.3*
- *"prommin" es el salario mínimo promedio" de ese año.*
- *"coverprom" es el porcentaje de trabajadores protegidos por la ley del salario mínimo.*

1. Introducción

1.2 Corte transversal

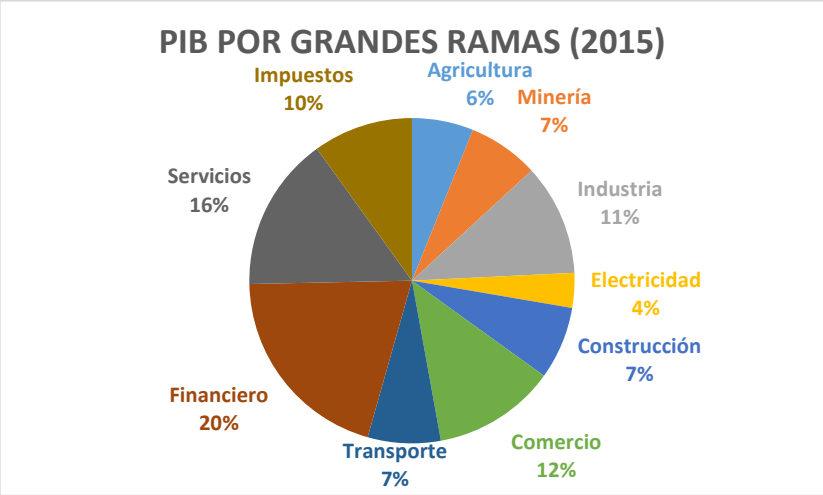
Tipos de información

Corte transversal (sección cruzada)

- Estas variables registran información para un momento del tiempo.
- Por ejemplo, las calificaciones del primer parcial de una materia.
- El orden de los datos **"no"** es importante.
- Una serie de tiempo es una **"película"** mientras que un corte transversal es una **"foto"**.

Tipos de información

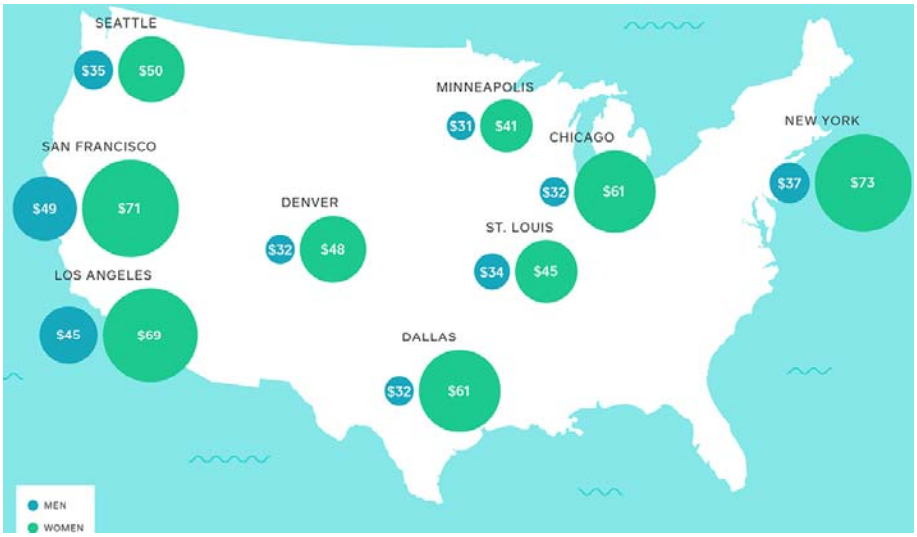
Corte transversal (composición del PIB en Colombia)



- *FUENTE: Elaboración propia. Datos DANE.*

Tipos de información

Corte Transversal (Precio promedio de un corte de peluquería USA)



- *FUENTE: www.usnews.com/news/blogs/data-mine/2014/02/28/what-america-pays-for-a-haircut*

Tipos de información

Corte transversal

obsno	salario (wage)	educ	exper	mujer (female)	casado (married)
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
⋮	⋮	⋮	⋮	⋮	⋮
525	11.56	16	5	0	1
526	3.50	14	5	1	0

- *FUENTE: Wooldridge (2010), Introducción a la Econometría, P 7, Tabla 1.1*

1. Introducción

1.3 Datos combinados

Tipos de información

Datos combinados

- Tiene dimensiones de serie de tiempo y corte transversal.
- Los **"datos panel"** son un caso especial en el cual se realiza seguimiento a las **"mismas unidades"** a lo largo del tiempo.
- Un ejemplo de datos panel es la inflación de todos los países de América Latina en 2010 y 2015 (los países no se crean ni se destruyen).
- Un ejemplo de datos combinados son estadísticas empresariales (entran y salen empresas).

Tipos de información

Datos combinados

obsno	año	preciov	impprov	piescuadr	recs	baños
1	1993	85500	42	1600	3	2.0
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮
250	1993	243600	41	2600	4	3.0
251	1995	65000	16	1250	2	1.0
252	1995	182400	20	2200	4	2.0
253	1995	97500	15	1540	3	2.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
520	1995	57200	16	1100	2	1.5

- *FUENTE: Wooldridge (2010), Introducción a la Econometría, P 10, Tabla 1.4*

Tipos de información

Datos panel

obsno	ciudad	año	homicidios	población	desempl	policía
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
⋮	⋮	⋮	⋮	⋮	⋮	⋮
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

- *FUENTE: Wooldridge (2010), Introducción a la Econometría, P 11, Tabla 1.5*

2. La probabilidad

Edad	Mujer	Estatura
19	0	170
20	0	170
20	1	160
18	1	163
18	0	179
20	0	193
19	0	170
23	0	185
24	1	153
19	1	172
22	0	177
18	1	160
17	1	159
24	0	183
20	0	175

La probabilidad

- La **"frecuencia absoluta"** cuenta el número de elementos pertenecientes a un valor especifico de una variable aleatoria.
- Una probabilidad se puede expresar como una **"frecuencia relativa" (FR)**:

$$FR = P(A) = \frac{n_A}{n}$$

- La FR muestra el número de elementos como proporción del total de observaciones.

La probabilidad

Edad	Frecuencia Absoluta	Frecuencia Relativa
17		
18		
19		
20		
22		
23		
24		
Total		

- *Utiliza los valores de la variable aleatoria "edad" para completar la tabla.*

La probabilidad

Edad	Frecuencia Absoluta	Frecuencia Relativa
17	1	6,7%
18	3	20%
19	3	20%
20	4	26,7%
22	1	6,7%
23	1	6,7%
24	2	13,3%
Total	**15**	**100%**

$$P(X = 18) = \frac{3}{15} = 0{,}20$$

- El 20% de las personas de la muestra tienen 20 años.

Ingresa a la siguiente página:

- *www.100people.org*

Ingresa a:

- *Statistics/100people*

3. Funciones de probabilidad

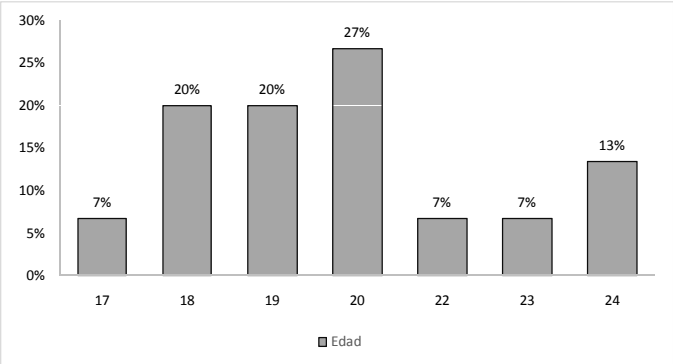
Funciones de probabilidad

- La variable aleatoria "edad" es generada a través de un proceso o "**función de probabilidad**".

$$f(X=b)=P(X=b)=\frac{n_b}{n}$$

- *Utilizando la variable aleatoria "edad" grafica la distribución de probabilidad colocando las frecuencias relativas en el eje vertical y los valores de la variable aleatoria en el eje horizontal.*

Funciones de probabilidad

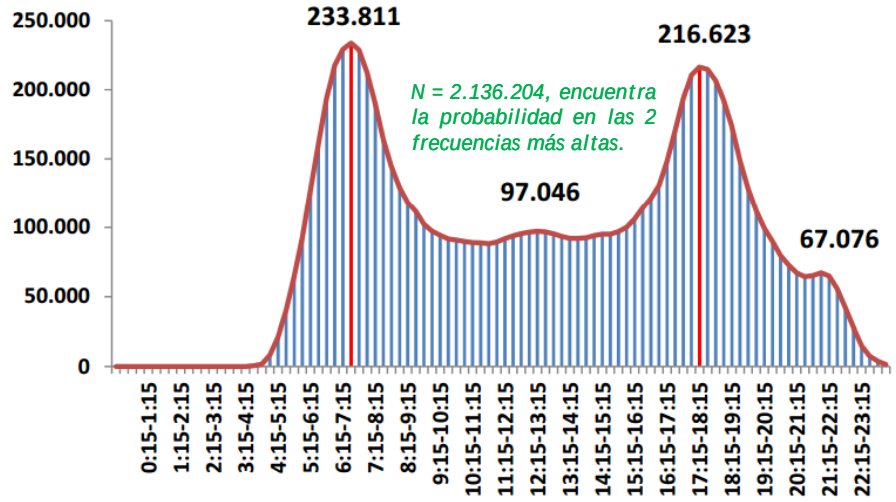


- *Interpreta las probabilidades de la función de probabilidad.*
- *¿Las probabilidades son positivas o negativas?, ¿Cuánto es su suma?*

$$0 \le f(X_i) \le 1 \qquad \sum_{i=1}^{n} f(X_i) = 1$$

Funciones de probabilidad

Perfil de demanda para un día típico en Transmilenio (06-11-13)



- *FUENTE: Alcaldía Mayor Bogotá (2013). Capítulo para la caracterización de la demanda de transporte del SITP con la inclusión de nuevos proyectos de infraestructura, Grafico 1, P 10*

Funciones de probabilidad

- La "**función de probabilidad acumulada**" se puede expresar como:

$$F(X) = P(X \leq b)$$

- Esta función estima las frecuencias acumuladas sumando todos los elementos para los cuales la variable aleatoria adquiere un valor especifico o inferior.

Funciones de probabilidad

Edad	Frecuencia Relativa	Frecuencia Acumulada
17	6,7%	
18	20%	
19	20%	
20	26,7%	
22	6,7%	
23	6,7%	
24	13,3%	
Total	**100%**	

- *Utiliza los valores de la frecuencia relativa de la variable aleatoria "edad" para completar la tabla.*

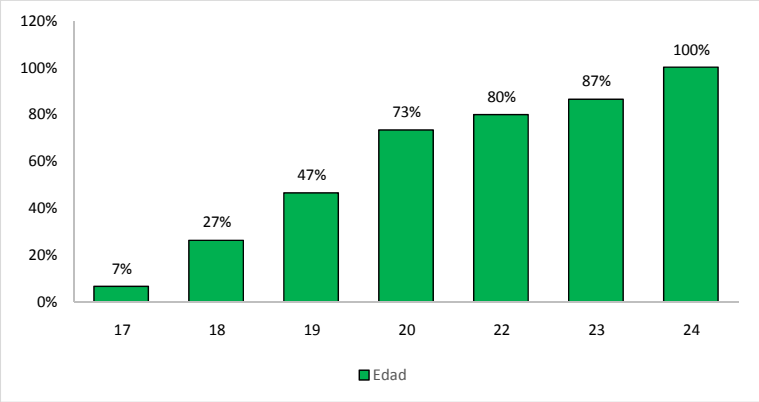
Funciones de probabilidad

Edad	Frecuencia Relativa	Frecuencia Acumulada
17	6,7%	6,7%
18	20%	27%
19	20%	47%
20	26,7%	73%
22	6,7%	80%
23	6,7%	87%
24	13,3%	100%
Total	**100%**	

$$P(X \leq 19) = 6{,}7\% + 20\% + 20\% = 47\%$$

- El 47% de las personas de la muestra tienen al menos 19 años.

Funciones de probabilidad



Funciones de probabilidad

IPC; variación mensual, año corrido y anual en Colombia

Año(aaaa)-	IPC	Variación mensual	Variación año corrido	Variación anual
2017-01	134,76594	1,02%	1,02%	5,47%
2017-02	136,12133	1,01%	2,04%	5,18%
2017-03	136,75543	0,47%	2,52%	4,69%
2017-04	137,40327	0,47%	3,00%	4,66%
2017-05	137,71286	0,23%	3,23%	4,37%
2017-06	137,87074	0,11%	3,35%	3,99%
2017-07	137,80022	-0,05%	3,30%	3,40%
2017-08	137,99321	0,14%	3,44%	3,87%
2017-09	138,04879	0,04%	3,49%	3,97%
2017-10	138,07187	0,02%	3,50%	4,05%
2017-11	138,32156	0,18%	3,69%	4,12%
2017-12	138,85399	0,38%	4,09%	4,09%
2018-01	139,72469	0,63%	0,63%	3,68%
2018-02	140,71151	0,71%	1,34%	3,37%

- *FUENTE: www.banrep.gov.co/es/ipc*

Funciones de probabilidad

- Una **"función de probabilidad conjunta"** incluye dos o más variables aleatorias.

$$f(X,Y) = P(X = b, Y = c)$$

- Esta distribución cumple las propiedades:

$$0 \leq f(X_i, Y_j) \leq 1$$

$$\sum_{i=1}^{n} \sum_{j=1}^{m} f(X_i, Y_j) = 1$$

Funciones de probabilidad

Edad	Mujer	Estatura
19	0	170
20	0	170
20	1	160
18	1	163
18	0	179
20	0	193
19	0	170
23	0	185
24	1	153
19	1	172
22	0	177
18	1	160
17	1	159
24	0	183
20	0	175

Funciones de probabilidad

	Mujer (Y)		
Edad (X)	Y = 0	Y = 1	**Total**
X = 17			
X = 18			
X = 19			
X = 20			
X = 22			
X = 23			
X = 24			
Total			

- *Utiliza los valores de las variables aleatorias "edad" y "mujer" para completar la tabla (frecuencias absolutas).*
- *La variable "mujer" toma el valor de uno si la persona es mujer y cero en caso de ser hombre.*

	Mujer (Y)		
Edad (X)	Y = 0	Y = 1	**Total**
X = 17	0	1	**1**
X = 18	1	2	**3**
X = 19	2	1	**3**
X = 20	3	1	**4**
X = 22	1	0	**1**
X = 23	1	0	**1**
X = 24	1	1	**2**
Total	**9**	**6**	**15**

- *Esta es la tabla de frecuencias absolutas.*

- La celda (X = 20, Y = 0) muestra cuantos hombres tienen 20 años. **(3 hombres)**
- La suma de la columna (Y = 1) señala el número total de mujeres (sin importar la edad). **(6 mujeres)**
- La suma del renglón (X = 20) indica el número de personas con 20 años (sin importar el sexo). **(4 personas)**

	Mujer (Y)		
Edad (X)	Y = 0	Y = 1	**Total**
X = 17	0%	7%	**7%**
X = 18	7%	13%	**20%**
X = 19	13%	7%	**20%**
X = 20	20%	7%	**27%**
X = 22	7%	0%	**7%**
X = 23	7%	0%	**7%**
X = 24	7%	7%	**13%**
Total	**60%**	**40%**	**100%**

- *Esta es la tabla de frecuencias relativas.*

$$P(X = 20, Y = 0) = \frac{3}{15} = 0{,}2$$

- El 20% de las personas de la muestra son hombres y tienen 20 años.

Funciones de probabilidad

- La **"Función de probabilidad marginal de X":**

$$f(X) = \sum_{j=1}^{m} f(X, Y)$$

Muestra la probabilidad para un valor especifico de la variable aleatoria X considerando **"todos los valores"** de la variable aleatoria Y.

- La **"Función de probabilidad marginal de Y"** es:

$$f(Y) = \sum_{i=1}^{n} f(X, Y)$$

	Mujer (Y)		
Edad (X)	Y = 0	Y = 1	**Total**
X = 17	0%	7%	**7%**
X = 18	7%	13%	**20%**
X = 19	13%	7%	**20%**
X = 20	20%	7%	**27%**
X = 22	7%	0%	**7%**
X = 23	7%	0%	**7%**
X = 24	7%	7%	**13%**
Total	**60%**	**40%**	**100%**

- La marginal (X = 20) es la probabilidad de tener 20 años sin importar el sexo (suma de probabilidades para el renglón X = 20).

$$f(X = 20) = 20\% + 7\% = 27\%$$

	Mujer (Y)		
Edad (X)	Y = 0	Y = 1	**Total**
X = 17	0%	7%	**7%**
X = 18	7%	13%	**20%**
X = 19	13%	7%	**20%**
X = 20	20%	7%	**27%**
X = 22	7%	0%	**7%**
X = 23	7%	0%	**7%**
X = 24	7%	7%	**13%**
Total	**60%**	**40%**	**100%**

- La marginal (Y = 1) es la probabilidad de ser mujer sin importar la edad (suma de probabilidades para la columna Y = 1).

$$f(Y = 1) = 7\% + 13\% + \ldots + 7\% = 40\%$$

Funciones de probabilidad

- La **"Función de probabilidad condicionada de X"**:

$$f(Y|X) = \frac{f(X = b, Y = c)}{f(X = b)}$$

Muestra la probabilidad de un elemento de la función de probabilidad conjunta considerando **"un valor especifico"** de la variable aleatoria X.

- La **"Función de probabilidad condicionada de Y"**:

$$f(X|Y) = \frac{f(X = b, Y = c)}{f(Y = c)}$$

	Mujer (Y)		
Edad (X)	Y = 0	Y = 1	**Total**
X = 17	0%	7%	**7%**
X = 18	7%	13%	**20%**
X = 19	13%	7%	**20%**
X = 20	20%	7%	**27%**
X = 22	7%	0%	**7%**
X = 23	7%	0%	**7%**
X = 24	7%	7%	**13%**
Total	**60%**	**40%**	**100%**

- Dentro de las personas que tienen 19 años la probabilidad de que en ese grupo se encuentre una mujer es 35%.

$$f(Y|X) = \frac{f(X = 19, Y = 1)}{f(X = 19)} = \frac{7\%}{20\%} = 35\%$$

	Mujer (Y)		
Edad (X)	Y = 0	Y = 1	**Total**
X = 17	0%	7%	**7%**
X = 18	7%	13%	**20%**
X = 19	13%	7%	**20%**
X = 20	20%	7%	**27%**
X = 22	7%	0%	**7%**
X = 23	7%	0%	**7%**
X = 24	7%	7%	**13%**
Total	**60%**	**40%**	**100%**

- Dentro de los hombres la probabilidad de que en ese grupo se encuentre una persona con 19 años es 22%.

$$f(X|Y) = \frac{f(X=19, Y=0)}{f(Y=0)} = \frac{13\%}{60\%} = 22\%$$

Funciones de probabilidad

- La edad es considerada una variables aleatoria **"discreta"** porque adquiere valores enteros (17 años).
- Una persona en cada momento registra aumentos de su edad (17,3 años) y en estos casos son más precisas las variables aleatorias **"continuas".**
- Las sumatorias se reemplazan por **"integrales"**.

4. Características de las funciones de probabilidad

Características de las funciones

- Una **"función de probabilidad"** se puede caracterizar mediante sus momentos.

Primer momento: Valor Esperado

- Es una "**media ponderada**" de una variable aleatoria (medida de tendencia central).

$$\mu_X = E(X) = \sum_{i=1}^{n} X_i \left[f\left(X_i\right) \right]$$

- *E(X) es el valor esperado, f(X) es la frecuencia relativa y "X" es el valor de la variable aleatoria.*
- *Utiliza los valores de la frecuencia relativa de la variable aleatoria "edad" para encontrar el valor esperado.*

Características de las funciones

Edad	Frecuencia Relativa
17	6,7%
18	20%
19	20%
20	26,7%
22	6,7%
23	6,7%
24	13,3%
Total	**100%**

$$E(X) = (17)(0{,}067) + (18)(0{,}20) + \ldots + (24)(0{,}133) = 20{,}07$$

Ingresa a la siguiente página:

- *www.youtube.com/watch?v=ZFNstNKgEDI*

Responde lo siguiente:

- *¿Por qué se sobre venden algunos servicios?*
- *¿Cuál es la probabilidad de viajar en avión?*
- *¿Cuál es el ingreso esperado por pasajero?*
- *¿Cuánto es el ingreso esperado de 195 pasajes?*

Características de las funciones

- Cuando se tiene una muestra se utiliza el estimador del valor esperado (**"media muestral"**):

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- La **"mediana"** es el dato que se encuentran en la mitad de la muestra mientras que la **"moda"** es el dato que más se repite.
- *Calcula la media muestral (promedio) mediana (organiza los datos) y moda para la variable "edad".*

Edad
19
20
20
18
18
20
19
23
24
19
22
18
17
24
20

Edad
19
20
20
18
18
20
19
23
24
19
22
18
17
24
20

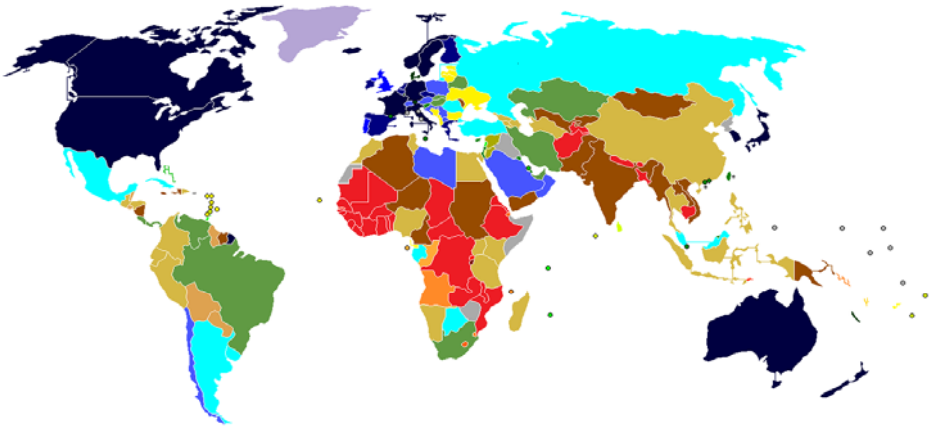
Edad
17
18
18
18
19
19
19
20
20
20
20
22
23
24
24

- **Media** = 20,07
- **Mediana** = 20
- **Moda** = 20

$$\overline{X} = \frac{19+20+20+18+18+\ldots+17+24+20}{15} = 20{,}07$$

Características de las funciones

Media Muestral: Producto Interno Bruto per cápita (2012)



- *FUENTE: http://es.wikipedia.org/wiki/Renta_per_c%C3%A1pita#mediaviewer/File:Capita20122.png (PIB es PIB per cápita en dólares)*

Características de las funciones

Segundo Momento: Varianza

- Muestra que tan distantes están los valores de su media en una distribución (su variabilidad).

$$\sigma_X^2 = \mathrm{Var}(X) = \sum_{i=1}^{n} (X_i - \mu_X)^2 \, f(X_i)$$

- En Finanzas las medidas de "dispersión" muestran el "riesgo" de un activo.

Características de las funciones

Edad	FR	VE	Edad - VE	(Edad - VE)^2	FR*(Edad - VE)^2
17	6,7%	20,07			
18	20%	20,07			
19	20%	20,07			
20	26,7%	20,07			
22	6,7%	20,07			
23	6,7%	20,07			
24	13,3%	20,07			

- *Completa la tabla y encuentra la varianza poblacional de la variable aleatoria "edad".*
- *VE es el Valor Esperado.*

Características de las funciones

Edad	FR	VE	Edad - VE	(Edad - VE)^2	FR*(Edad - VE)^2
17	6,7%	20,07	-3,07	9,42	0,63
18	20%	20,07	-2,07	4,28	0,86
19	20%	20,07	-1,07	1,14	0,23
20	26,7%	20,07	-0,07	0,00	0,00
22	6,7%	20,07	1,93	3,72	0,25
23	6,7%	20,07	2,93	8,58	0,57
24	13,3%	20,07	3,93	15,44	2,06
				Varianza P	**4,60**

- *Varianza P es la varianza poblacional.*
- *Una varianza toma en cuenta las distancias de cada observación en relación al Valor Esperado. Si son positivas están por encima. Si son negativas están por debajo.*

Características de las funciones

- Cuando se tiene una muestra se calcula un estimador de la varianza (varianza muestral):

$$S_X^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$$

- *Calcula la varianza muestral, el valor mínimo y máximo además del rango para la variable aleatoria "edad".*

- La desviación estándar muestral es:

$$S_X = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}}$$

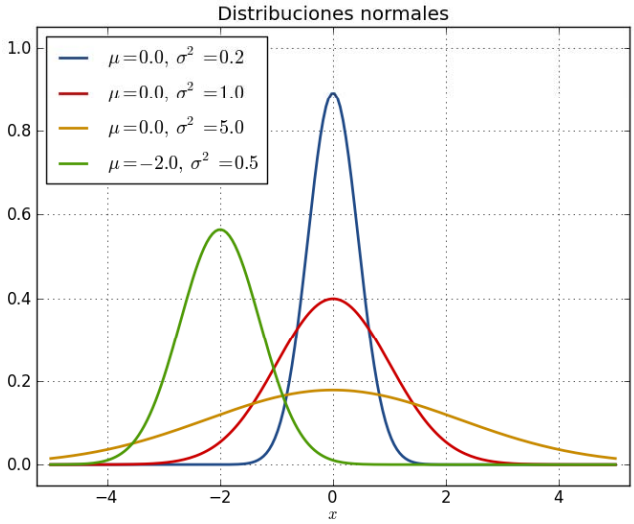
Edad	Media	Edad - Media	(Edad - Media)^2
19	20,07	-1,07	1,14
20	20,07	-0,07	0,005
20	20,07	-0,07	0,005
18	20,07	-2,07	4,28
18	20,07	-2,07	4,28
20	20,07	-0,07	0,005
19	20,07	-1,07	1,14
23	20,07	2,93	8,58
24	20,07	3,93	15,44
19	20,07	-1,07	1,14
22	20,07	1,93	3,72
18	20,07	-2,07	4,28
17	20,07	-3,07	9,42
24	20,07	3,93	15,44
20	20,07	-0,07	0,005
		68,93	**Suma**
		4,92	**Varianza Muestral**
		2,22	**Desviación Estándar**

- **Mínimo** = 19
- **Máximo** = 24
- **Rango** = 5

- Dada una media (20,07 años) y desviación estándar (2,22 años) aproximadamente el rango de edad fluctúa entre 17,9 años (20,1 – 2,2) y 22,3 años (20,1 + 2,2).

Características de las funciones

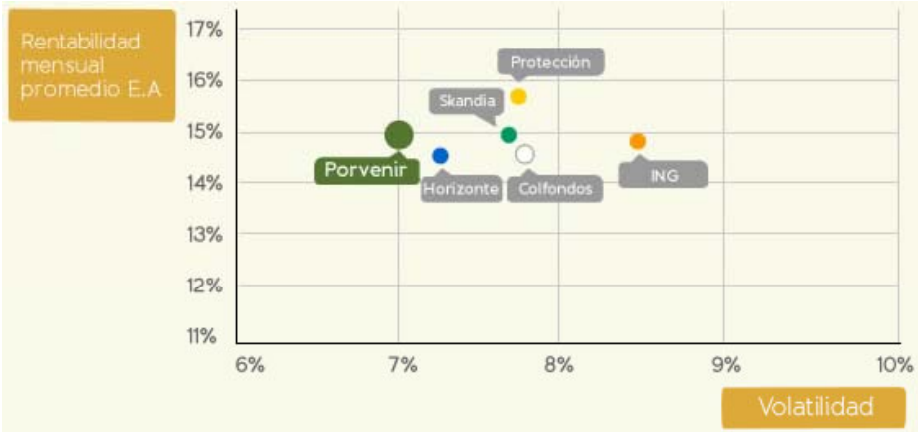
Media y varianza para varias distribuciones hipotéticas



- *FUENTE: http://pybonacci.org/2012/04/21/estadistica-en-python-con-scipy/*

Características de las funciones

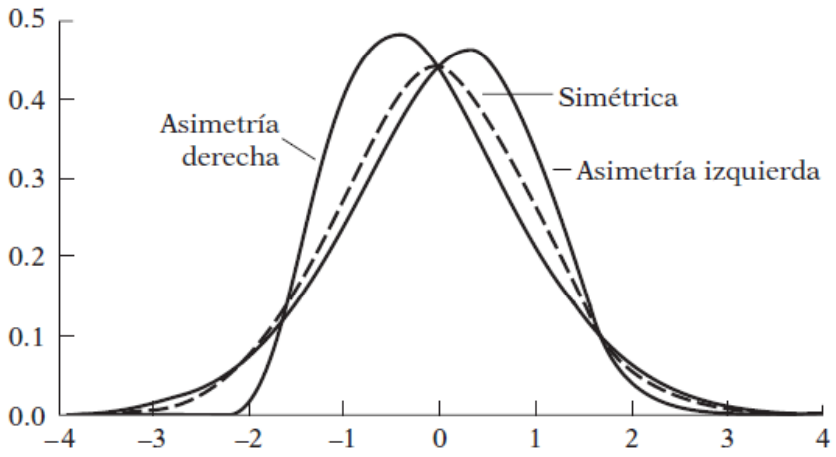
Fondos de pensiones – rentabilidad (media) y volatilidad (DE)



- *FUENTE: www.porvenir.com.co/CentroNoticias/Paginas/conozca_rentabilidades_fondomoderado-home-PO.aspx*
- *Datos Superintendencia Financiera de Colombia. DE es Desviación Estándar.*

Características de las funciones

Tercer momento (asimetría o sesgo)



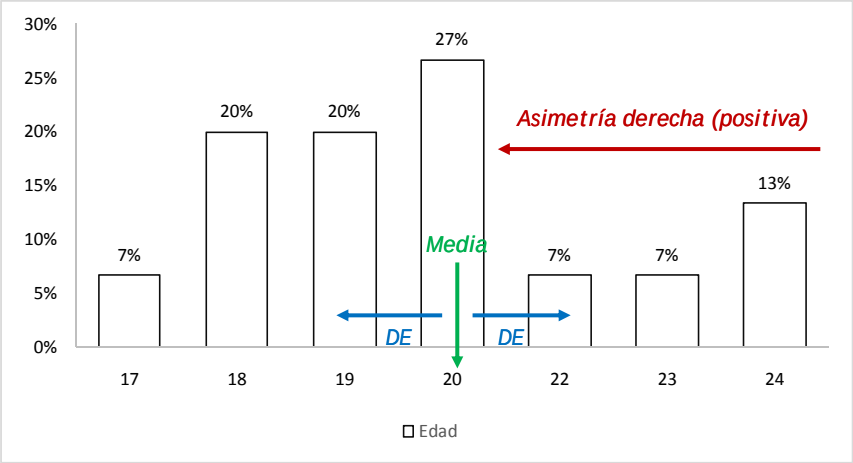
- *FUENTE: Gujarati (2010), Econometría, P 816, Tabla A.3 (a)*

Edad	Media	Edad - Media	(Edad - Media)^3
19	20,07	-1,07	-1,23
20	20,07	-0,07	-0,0003
20	20,07	-0,07	0,00
18	20,07	-2,07	-8,87
18	20,07	-2,07	-8,87
20	20,07	-0,07	-0,0003
19	20,07	-1,07	-1,23
23	20,07	2,93	25,15
24	20,07	3,93	60,70
19	20,07	-1,07	-1,23
22	20,07	1,93	7,19
18	20,07	-2,07	-8,87
17	20,07	-3,07	-28,93
24	20,07	3,93	60,70
20	20,07	-0,07	-0,0003
		94,52	**Suma**
		2,22	**Desviación Estándar**
		0,58	**Asimetría**

$$A_X = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{nS_X^3}$$

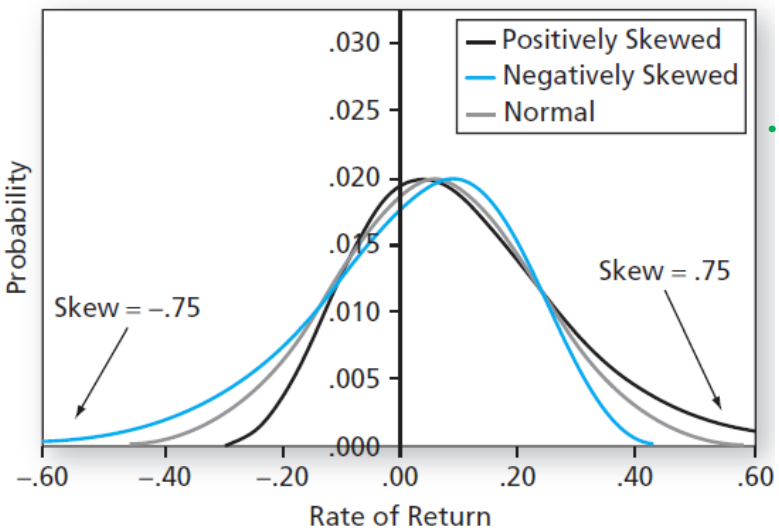
- La suma de los datos por encima de la media (números positivos) es más grande que la suma de los datos por debajo de la media (negativos).

Características de las funciones



Características de las funciones

Tercer momento (asimetría o sesgo)

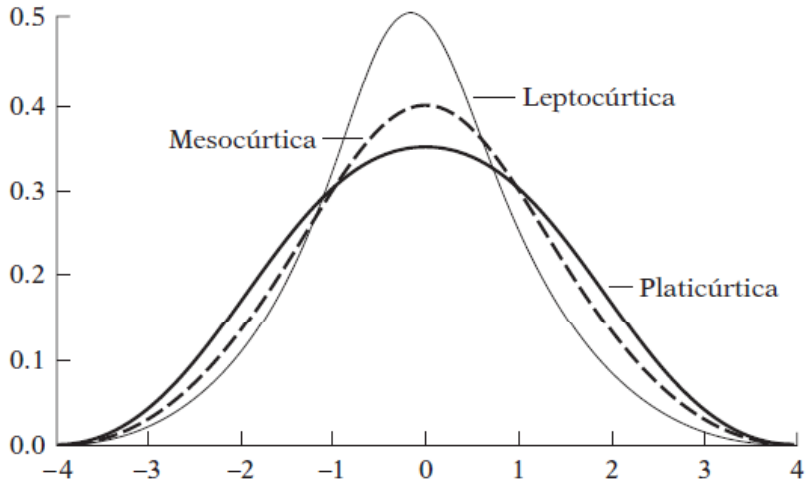


- *Cuando la asimetría es positiva los valores positivos dominan la distribución (sesgo derecho).*

- *FUENTE: Bodie, Kane y Marcus (2014), Investments, P 138, Figure 5.5.A*

Características de las funciones

Cuarto momento (curtosis)



- *FUENTE: Gujarati (2010), Econometría, P 816, Tabla A.3 (b)*

5. Relaciones entre variables

Relaciones entre variables

Covarianza

- Mide el grado de asociación **"lineal"** entre dos variables aleatorias:

$$\mathrm{Cov}(X,Y) = \sigma_{XY} = \sum_{i=1}^{n}\sum_{j=1}^{m}(X_i - \mu_X)(Y_j - \mu_Y)f(X,Y)$$

- Cuando se tiene una muestra se calcula un estimador de la covarianza (muestral):

$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

Relaciones entre variables

		1			2	
Edad	Media	Edad - Media	Estatura	Media	Estatura - Media	(1)*(2)
19	20,07		170	171,26		
20	20,07		170	171,26		
20	20,07		160	171,26		
18	20,07		163	171,26		
18	20,07		179	171,26		
20	20,07		193	171,26		
19	20,07		170	171,26		
23	20,07		185	171,26		
24	20,07		153	171,26		
19	20,07		172	171,26		
22	20,07		177	171,26		
18	20,07		160	171,26		
17	20,07		159	171,26		
24	20,07		183	171,26		
20	20,07		175	171,26		

- *Completa la tabla y encuentra la covarianza muestral de las variables aleatorias "edad" y "estatura".*
- *La media de la estatura se obtiene de la formula "media muestral".*

Relaciones entre variables

		1			2	
Edad	Media	Edad - Media	Estatura	Media	Estatura - Media	(1)*(2)
19	20,07	-1,07	170	171,26	-1,26	1,35
20	20,07	-0,07	170	171,26	-1,26	0,09
20	20,07	-0,07	160	171,26	-11,26	0,79
18	20,07	-2,07	163	171,26	-8,26	17,10
18	20,07	-2,07	179	171,26	7,74	-16,02
20	20,07	-0,07	193	171,26	21,74	-1,52
19	20,07	-1,07	170	171,26	-1,26	1,35
23	20,07	2,93	185	171,26	13,74	40,26
24	20,07	3,93	153	171,26	-18,26	-71,76
19	20,07	-1,07	172	171,26	0,74	-0,79
22	20,07	1,93	177	171,26	5,74	11,08
18	20,07	-2,07	160	171,26	-11,26	23,31
17	20,07	-3,07	159	171,26	-12,26	37,64
24	20,07	3,93	183	171,26	11,74	46,14
20	20,07	-0,07	175	171,26	3,74	-0,26
						88,73 Suma
						6,34 Covarianza Muestral

- La relación entre la variable edad y estatura es directamente proporcional (covarianza positiva).

Relaciones entre variables

Coeficiente de Correlación

- **"Cuantifica"** la covarianza entre dos variables aleatorias:

$$corr(X,Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- Cuando se tiene una muestra se calcula el coeficiente de correlación muestral:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

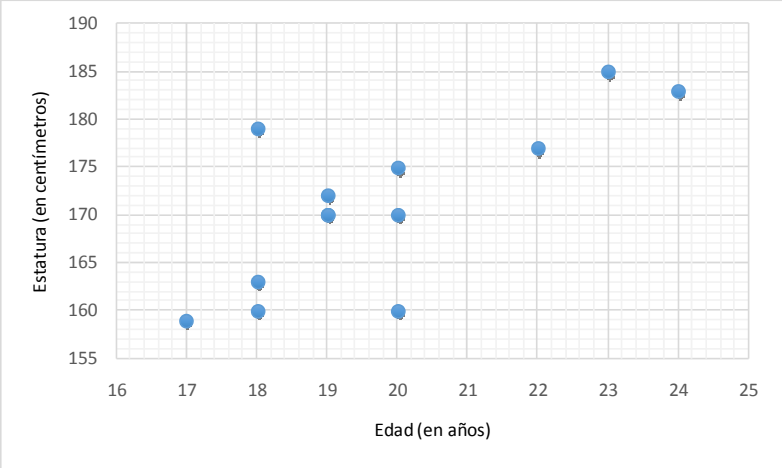
- *Calcula el coeficiente de correlación muestral para las variables "edad" y "estatura".*

Estatura	Media	Estatura - Media	(Estatura - Media)^2	
170	171,26	-1,26	1,59	
170	171,26	-1,26	1,588	
160	171,26	-11,26	126,788	
163	171,26	-8,26	68,23	
179	171,26	7,74	59,91	
193	171,26	21,74	472,628	
170	171,26	-1,26	1,59	
185	171,26	13,74	188,79	
153	171,26	-18,26	333,43	
172	171,26	0,74	0,55	
177	171,26	5,74	32,95	
160	171,26	-11,26	126,79	
159	171,26	-12,26	150,31	
183	171,26	11,74	137,83	
175	171,26	3,74	13,988	
			1716,93	**Suma**
			122,64	**Varianza Muestral**
			11,07	**Desviación Estándar**

$$r_{XY} = \frac{6,34}{(2,22)(11,07)} = 0,26$$

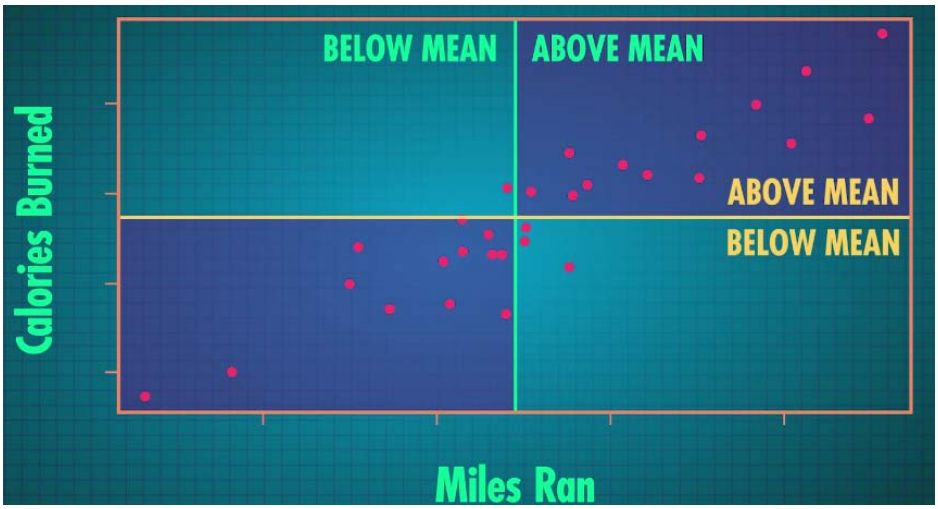
Relaciones entre variables

Edad y Estatura (correlación positiva: 26%)



Relaciones entre variables

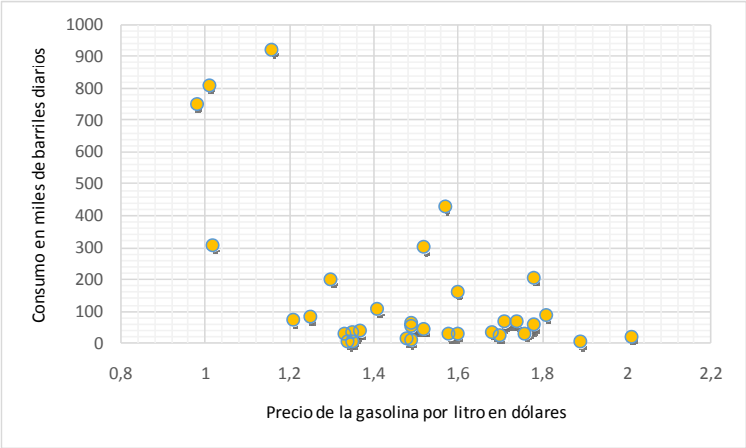
Correlación positiva



- *FUENTE: Crash Course Statistics (2018). Correlation doesn´t equal causation. [Archivo de video]. Recuperado de: www.youtube.com/watch?v=GtV-VYdNt_g&t=0s*

Relaciones entre variables

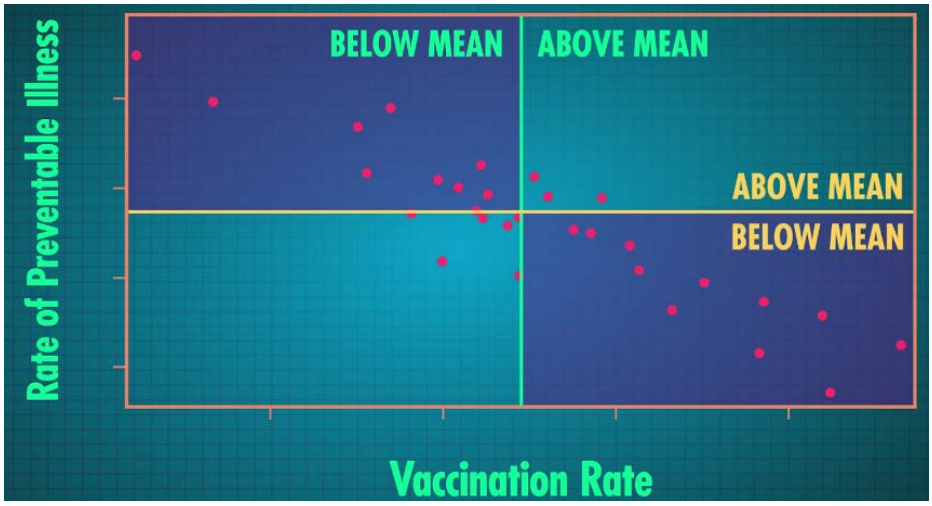
Consumo y precio de la gasolina (Correlación negativa: -56%)



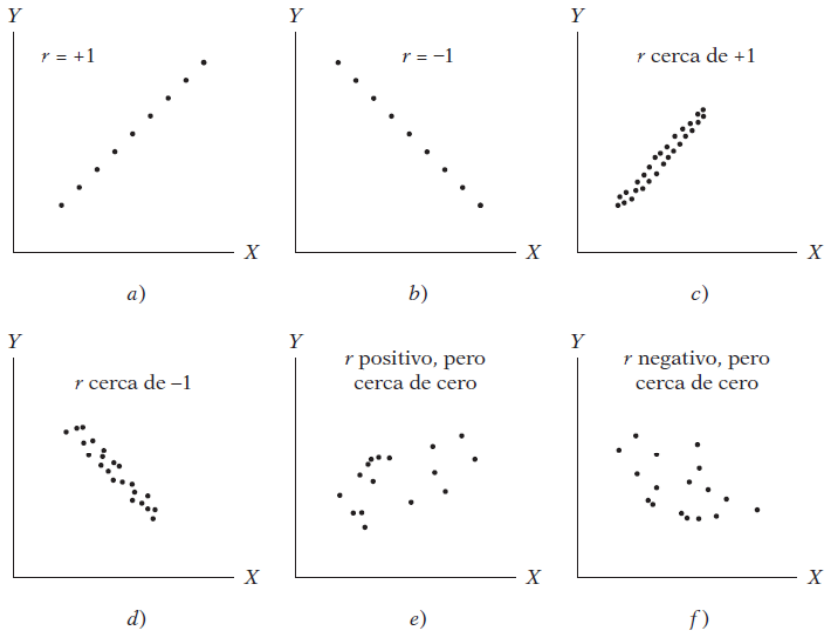
- *FUENTE: Elaboración propia.*
- *Datos 2012 (https://es.theglobaleconomy.com/rankings/gasoline_consumption/) y (https://es.globalpetrolprices.com/gasoline_prices/).*

Relaciones entre variables

Correlación negativa



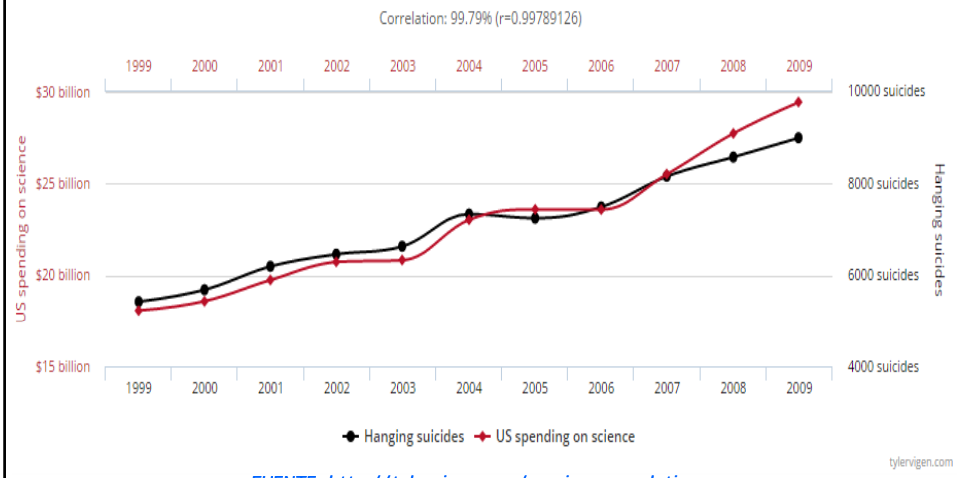
- *FUENTE: Crash Course Statistics (2018). Correlation doesn´t equal causation. [Archivo de video]. Recuperado de: www.youtube.com/watch?v=GtV-VYdNt_g&t=0s*



- *FUENTE: Gujarati (2010), Econometría, P 78, Figura 3.10*

Relaciones entre variables

Gasto en ciencia y suicidios por ahorcamiento

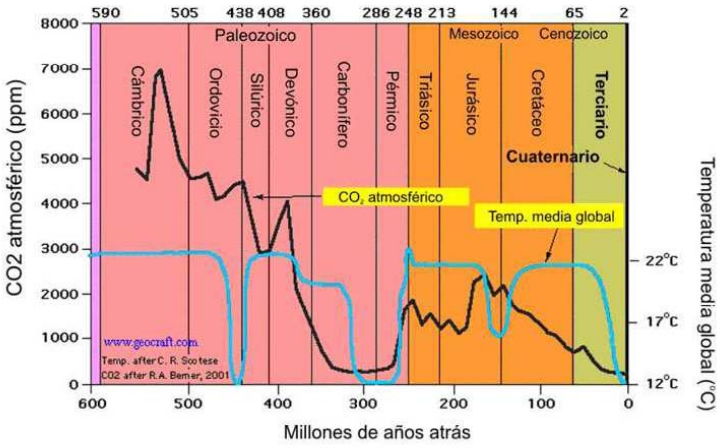


- *FUENTE: http://tylervigen.com/spurious-correlations*

- *¿Cuánto es la correlación entre las variables?, ¿Tiene sentido?*

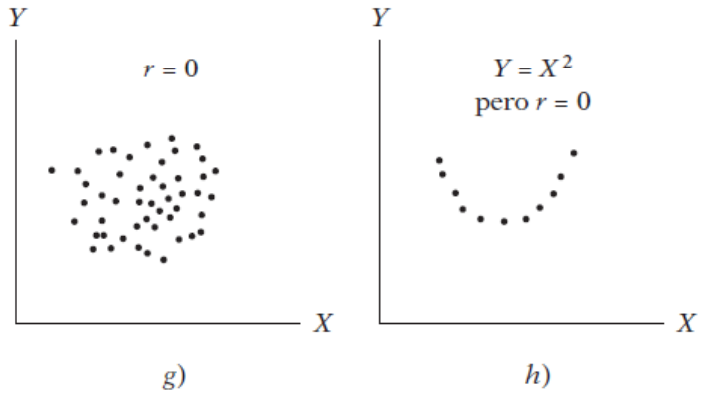
Relaciones entre variables

Cantidad de CO2 atmosférico y temperatura global



- *FUENTE: www.mitosyfraudes.org/Calen-5.html*

- *¿Las variables están relacionadas?, ¿De que otras variables depende la temperatura del planeta?*



- *FUENTE: Gujarati (2010), Econometría, P 78, Figura 3.10*

Relaciones entre variables

¿Por qué pueden estar correlacionadas una variable "X" y una variable "Y"?

- La variable X causa la variable Y.
- La variable Y causa la variable X.
- Otra variable Z causa a la variable X y la variable Y.
- Coincidencia.

6. Distribuciones de probabilidad continuas

Distribuciones Continuas

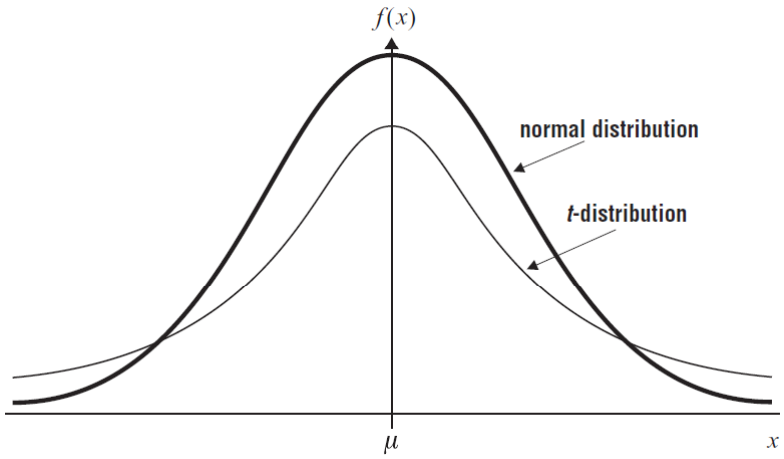
- En el caso de las **"variables aleatorias continuas"** las probabilidades se calculan para un rango numérico dado que la probabilidad de asumir un valor determinado es cero.

Ejemplos de distribuciones de probabilidad de variables continuas son:

- Distribución "Z" (normal).
- Distribución "t" (student).
- Distribución "chi cuadrada".
- Distribución "F" (de Fisher).

Distribuciones Continuas

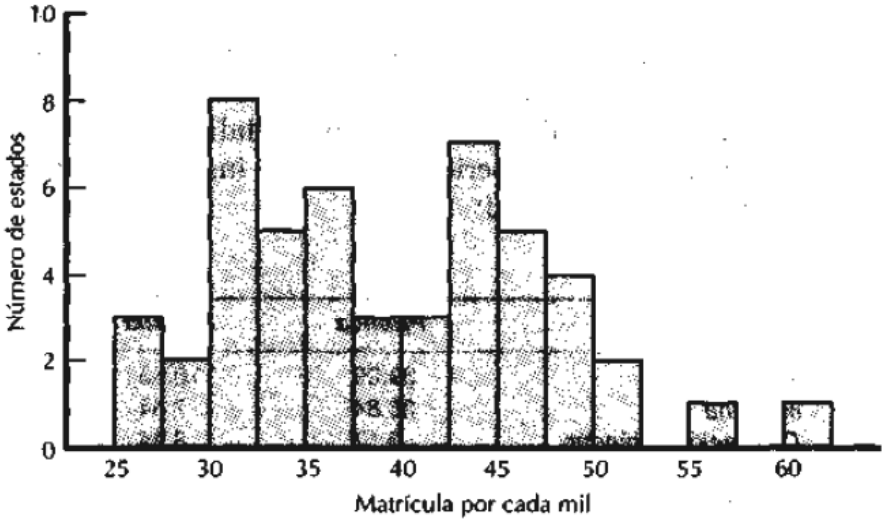
Distribución Normal y "t" student



- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 55, Figure 2.12*

Distribuciones Continuas

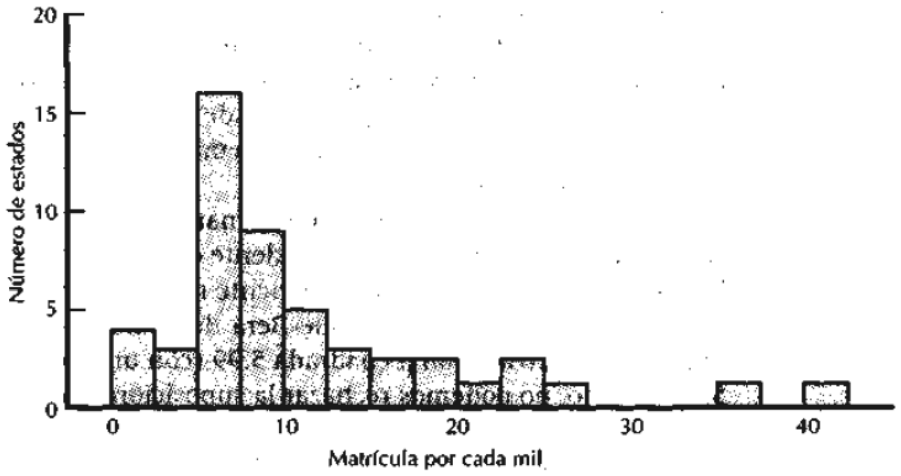
Histograma de matrículas (públicas)



- *FUENTE: Pindyck y Rubinfeld (2001), Econometría: Modelos y pronósticos, P 49, Figura 2.13.a*

Distribuciones Continuas

Histograma de matrículas (privadas)



- *FUENTE: Pindyck y Rubinfeld (2001), Econometría: Modelos y pronósticos, P 49, Figura 2.13.b*

Distribuciones Continuas

Estadísticas descriptivas (matrículas)

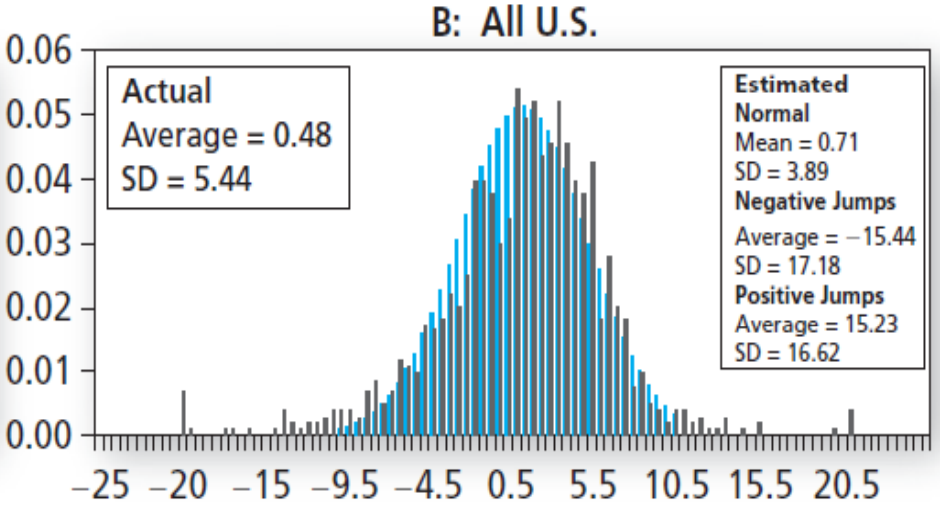
	Pública	***Privada***
Media	39.29	10.53
Mediana	38.84	7.84
Desviación estándar	8.17	8.24
Sesgo	0.38	1.78
Kurtosis	2.61	6.24
Jarque-Bera	1.54	48.26

- *FUENTE: Pindyck y Rubinfeld (2001), Econometría: Modelos y pronósticos, P 50*

- El histograma de escuelas públicas se asemeja más a la distribución **"normal"** porque su **"media"** y **"mediana"** son muy parecidas además su **"asimetría" (sesgo)** es cercana a cero.

Distribuciones Continuas

Distribución de las frecuencias en las tasas de retorno en USA (1926-2012)



- *FUENTE: Bodie, Kane y Marcus (2014), Investments, P 144, Figure 5.6.B*

Edad	Media	Edad - Media	Edad estandarizada
19	20,07	-1,07	-0,48
20	20,07	-0,07	-0,03
20	20,07	-0,07	-0,03
18	20,07	-2,07	-0,93
18	20,07	-2,07	-0,93
20	20,07	-0,07	-0,03
19	20,07	-1,07	-0,48
23	20,07	2,93	1,32
24	20,07	3,93	1,77
19	20,07	-1,07	-0,48
22	20,07	1,93	0,87
18	20,07	-2,07	-0,93
17	20,07	-3,07	-1,38
24	20,07	3,93	1,77
20	20,07	-0,07	-0,03

$$Z = \frac{X_i - \overline{X}}{S_X}$$

- Una variable estandarizada se obtiene al restar la media de cada observación y dividir entre la desviación estándar.

- *Suponiendo que la muestra de estudiantes sigue la distribución normal, ¿Cuál es la probabilidad de encontrar una persona con una edad superior a 22 años?*
- *La media de la edad es 20,07 y la desviación estándar 2,22.*

$$P(X > 22)$$

$$P(X - \overline{X} > 22 - \overline{X})$$

$$P\left(\frac{X - \overline{X}}{S_X} > \frac{22 - \overline{X}}{S_X}\right)$$

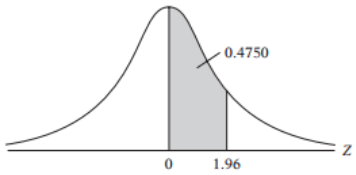
$$P\left(Z > \frac{22 - 20{,}07}{2{,}22}\right)$$

$$P(Z > 0{,}87)$$

Ejemplo

$\Pr(0 \leq Z \leq 1.96) = 0.4750$

$\Pr(Z \geq 1.96) = 0.5 - 0.4750 = 0.025$



Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4454	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817

- *FUENTE: Gujarati (2010), Econometría, P 878, Tabla D.1*

$$P(Z > 0{,}87)$$

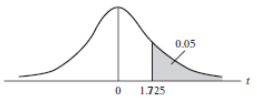
$$P(Z > 0{,}87) = 0{,}5 - P(Z < 0{,}87)$$

$$P(Z > 0{,}87) = 0{,}5 - 0{,}3078$$

$$P(Z > 0{,}87) = 0{,}19$$

- La variable aleatoria **"edad"** registró una media (20,07) similar a su mediana (20) adicionalmente la asimetría se encuentra cercana a cero (0,58).
- Por tanto la edad se asemeja a una distribución **"normal"** la cual al ser estandarizada (Z), muestra una probabilidad igual a 19% de encontrar una persona con una edad superior a 22 años.

Ejemplo
$\Pr(t > 2.086) = 0.025$
$\Pr(t > 1.725) = 0.05$ para gl = 20
$\Pr(|t| > 1.725) = 0.10$



Pr / gl	0.25 0.50	0.10 0.20	0.05 0.10	0.025 0.05	0.01 0.02	0.005 0.010	0.001 0.002
1	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	0.816	1.886	2.920	4.303	6.965	9.925	22.327
3	0.765	1.638	2.353	3.182	4.541	5.841	10.214
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232
120	0.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090

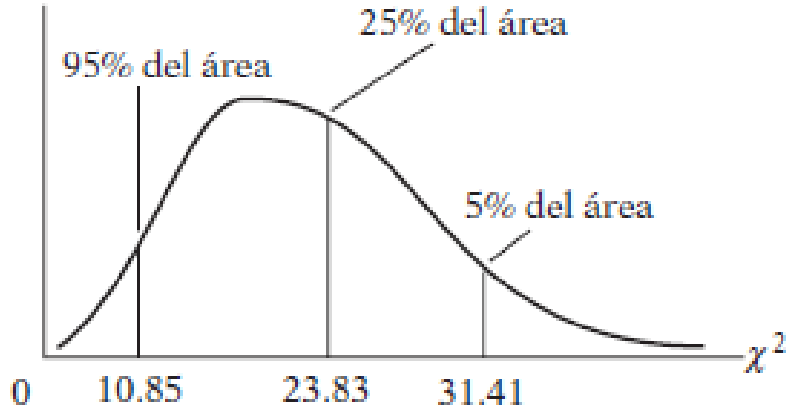
- *FUENTE: Gujarati (2010), Econometría, P 879, Tabla D.1*

$$t_{obs} = \frac{\overline{X} - \mu_X}{\left(\frac{S_X}{\sqrt{n}}\right)}$$

- La variable **"*t-student*"** utiliza la desviación estándar muestral para realizar una prueba de medias.

$$t_{CRT} \approx t_{n-1}$$

- La variable **"*chi cuadrado*"** (ji cuadrado) utiliza la desviación estándar muestral para realizar una prueba de varianzas.

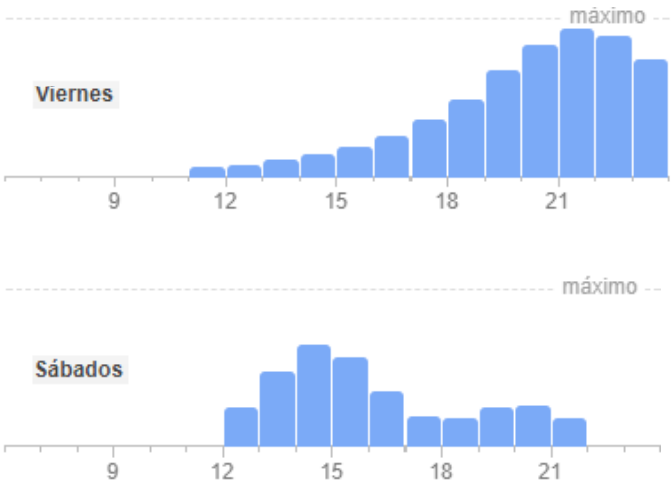


$P(\chi^2 > 10{,}85) = 0{,}95$ $P(\chi^2 > 23{,}83) = 0{,}25$ $P(\chi^2 > 31{,}41) = 0{,}05$

- *FUENTE: Gujarati (2010), Econometría, P 886, Tabla D.4 (se utilizan 20 grados de libertad)*

Repaso estadístico

Distribución de clientes en dos restaurantes y días diferentes

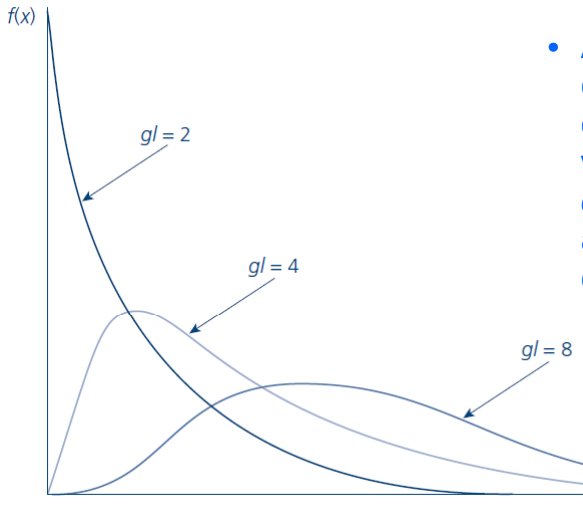


- *FUENTE: www.google.com*

Repaso estadístico

Distribución chi cuadrada

$$\chi^2_{CRT} \approx \chi^2_{n-1}$$



- A medida que aumenta el número de observaciones la variable **"*chi cuadrado*"** se aproxima a la distribución normal estándar.

$$\chi^2_{obs} = (n-1)\left(\frac{S_X^2}{\sigma_X^2}\right)$$

- *FUENTE: Wooldridge (2010), Introducción a la Econometría, P 742, Figura B.9*

Referencias

- Bodie, Z., Kane, A., y Marcus, A (2014). *Investments*. United States: McGraw Hill.
- Brooks, C (2008). *Introductory Econometrics for Finance*. USA: Cambridge University Press.
- Gujarati, D., y Porter, D (2010). *Econometría*. México: McGraw Hill.
- Pindyck, R., y Rubinfeld, D (2001). *Econometría: Modelos y pronósticos*. México: McGraw Hill.
- Wooldridge, G (2010). *Introducción a la Econometría*. México: Cengage Learning.