Análisis y Visualización de Datos

Clase 2

Conceptos de Estadística Descriptiva

Docentes : **Soledad Palacios(UNLP)**

Milagro Teruel (UNC)

Repaso de la clase anterior

- Estudiamos en profundidad el concepto de Probabilidad, estudiando sus tres variantes: probabilidad clásica, probabilidad frecuentista y probabilidad subjetiva
- Examinamos la clasificación de los datos: variables cuantitativas y variables categóricas
- Revisamos el concepto de variable aleatoria
- Analizamos las distribuciones de las variables aleatorias

Definición de probabilidad

La **probabilidad** es una medida de la certidumbre asociada a un suceso o evento futuro y suele expresarse como un número entre 0 y 1 (o entre 0 % y 100 %).
Una forma tradicional de estimar algunas probabilidades sería obtener la frecuencia de un acontecimiento determinado mediante la realización de experimentos aleatorios, de los que se conocen todos los resultados posibles, bajo condiciones *suficientemente* estables. Un suceso puede ser improbable (con probabilidad cercana a cero), probable (probabilidad intermedia) o seguro (con probabilidad uno).

Estadistica

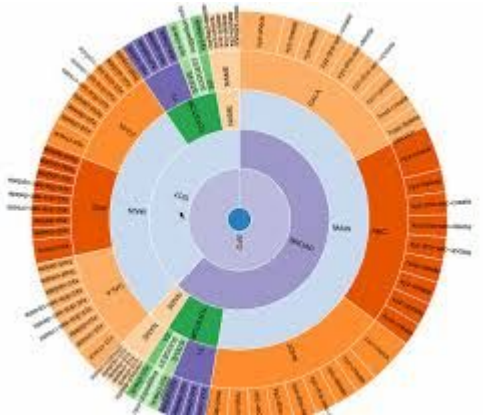
La estadística (la forma femenina del término alemán Statistik, derivado a su vez del italiano statista, "hombre de Estado") es una rama de las matemáticas y una herramienta que estudia usos y análisis provenientes de una muestra representativa de datos, que busca explicar las correlaciones y dependencias de un fenómeno físico o natural, de ocurrencia en forma aleatoria o condicional.
El campo de la estadística tiene que ver con la recopilación, organización, análisis y uso de datos para tomar decisiones razonables basadas en tal análisis.

Muestra y población

Cuando recogemos los datos muchas veces es imposible relevar la característica de interés de todos el grupo entero (***población***) o universo, se examina una pequeña parte del grupo, llamada muestra.

Estadística Descriptiva

La parte de la estadística que estudia la muestra sin inferir alguna conclusión sobre la población es la estadística descriptiva.
En particular la estadística descriptiva trata sobre los métodos para recolectar, organizar y resumir datos.
El análisis se limita en sí mismo a los datos coleccionados y no se realiza inferencia alguna o generalizaciones acerca de la totalidad de dónde provienen esas observaciones.



Analicemos el dataset del Titanic

Vamos a trabajar con el dataSet del Titanic.

Veamos que datos tenemos

Miramos que datos tiene

les pedimos una descripción total a panda

In [5]:

Slide Type

```
titanic.describe()
```

Out[5]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

¿Qué hacemos con estos datos?

Distribución de frecuencias para variables cuantitativas

Un tema íntimamente relacionado con los histogramas son las tablas de distribución de frecuencia, en definitiva los histogramas no son más que gráficos de tablas de distribución de frecuencia. La distribución de frecuencia de una variable cuantitativa consiste en un resumen de la ocurrencia de un dato dentro de una colección de categorías que no se superponen. Estas categorías las vamos a poder armar según nuestra conveniencia y lo que queramos analizar.

Distribución de frecuencias

In [27]:

Slide Type

```
# Distribución de frecuencia.
# 1ro creamos un rango para las categorías.
contenedores = np.arange(0, 81., 10)

# luego cortamos los datos en cada contenedor
frec = pd.cut(titanic['Age'], contenedores)

# por último hacemos el recuento de los contenedores
# para armar la tabla de frecuencia.
tabla_frec = pd.value_counts(frec)
tabla_frec
```

Out[27]:

```
(20.0, 30.0]    407
(30.0, 40.0]    155
(10.0, 20.0]    115
(40.0, 50.0]     86
(0.0, 10.0]      64
(50.0, 60.0]     42
(60.0, 70.0]     17
(70.0, 80.0]      5
Name: Age, dtype: int64
```

Por defecto muestra el intervalo con más ocurrencias en primer lugar

Histograma

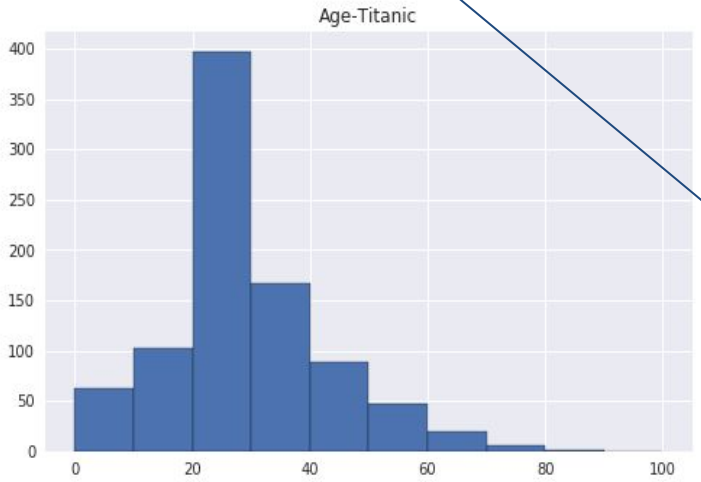
In [30]:

```
plt.title('Age-Titanic')
plt.hist(titanic.Age, edgeColor='black', bins=[0, 10, 20, 30, 40, 50, 60, 70, 80, 90,100])
```

Slide Type

Out[30]:

```
(array([ 62., 102., 397., 167.,  89.,  48.,  19.,   6.,   1.,   0.]),
 array([  0,  10,  20,  30,  40,  50,  60,  70,  80,  90, 100]),
 <a list of 10 Patch objects>)
```



Con este parámetro seteamos el valor del intervalo a graficar

Con este parámetro seteamos la separación entre las barras

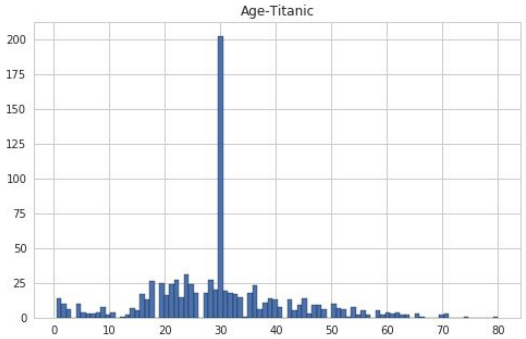
Propiedades del histograma

- Cada uno de los subintervalos representados por las barras son llamadas **clases.**
- En general se deben graficar entre 5 y 20 barras.
- El punto medio de cada clase es la marca de clase.
- La longitud de cada intervalo de clase es el ancho de clase
- el área de la barra debe ser proporcional a la frecuencia de la clase.

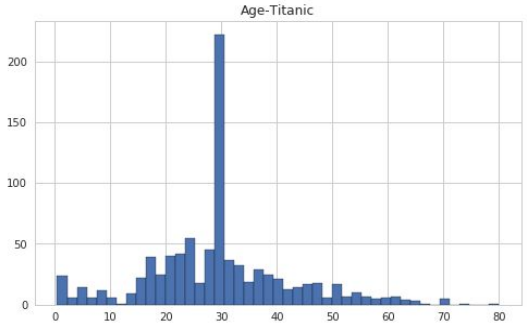
Propiedades del histograma

Los histogramas son útiles al proporcionar una impresión visual del aspecto que tiene la distribución de las mediciones, así como información sobre la dispersión de los datos. Al construir una tabla de frecuencias se pierde información, sin embargo esa pérdida de información es a menudo pequeña si se le compara con la facilidad de interpretación ganada al utilizar la distribución de frecuencias y el histograma.

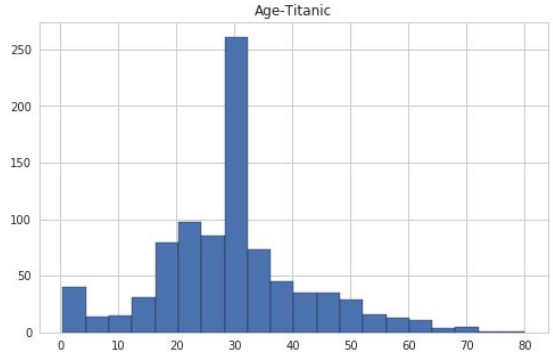
Variación gráfica respecto las clases




```
plt.title('Age-Titanic')
plt.hist(titanic.Age,
edgeColor='black', bins=90)
```




```
plt.title('Age-Titanic')
plt.hist(titanic.Age,
edgeColor='black', bins=45)
```




```
plt.title('Age-Titanic')
plt.hist(titanic.Age,
edgeColor='black',
bins=20)
```

Histogramas de variables categóricas

Las tablas de frecuencia y los histogramas también pueden emplearse en datos cualitativos o categóricos , es decir la muestra no consiste de valores numéricos (datos cuantitativos) sino que los datos se ordenan en categorías y se registra cuántas observaciones caen en cada categoría (las categorías pueden ser masculino , femenino o fumador, no fumador o clasificar según nivel educativo: primario, secundario, terciario, universitario, ninguno). Cuando los datos son categóricos las clases se dibujan con el mismo ancho.

Ejemplo

En la Provincia de Buenos Aires

INTENCIÓN DE VOTO A SENADOR NACIONAL

Esteban Bullrich - Cambiemos

33,9%

Cristina Kirchner - Fte. de Unidad Ciudadana

29,6%

Sergio Massa - 1País

11,7%

Florencio Randazzo - Frente Justicialista

5,6%

Néstor Pitrola - FIT

1,7%

En la Provincia de Buenos Aires

INTENCIÓN DE VOTO A SENADOR NACIONAL

Esteban Bullrich - Cambiemos

33,9%

Cristina Kirchner - Fte. de Unidad Ciudadana

29,6%

Sergio Massa - 1País

11,7%

Florencio Randazzo - Frente Justicialista

5,6%

Néstor Pitrola - FIT

3,7%

Medidas Descriptivas

Del mismo modo que las gráficas pueden mejorar la presentación de los datos, las descripciones numéricas también tienen gran valor. Se presentan varias medidas numéricas importantes para describir las características de los datos.

Una característica importante de un conjunto de números es su localización o tendencia central .

Medidas de localización - Media

La medida más común de localización o centro de un grupo de datos es el promedio aritmético ordinario o media. Ya que casi siempre se considera a los datos como una muestra, la media aritmética se conoce como media muestral.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Propiedades de la media

- Su cálculo es muy sencillo y en él intervienen todos los datos.
- Su valor es único para una serie de datos dada.
- Se usa con frecuencia para comparar poblaciones, aunque es más apropiado acompañarla de una medida de dispersión.
- Se interpreta como "punto de equilibrio" o "centro de masas" del conjunto de datos, ya que tiene la propiedad de equilibrar las desviaciones de los datos respecto de su propio valor:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n}\overline{x} = \overline{x} - \overline{x} = 0$$

- Minimiza las desviaciones cuadráticas de los datos respecto de cualquier valor prefijado, esto es, el valor de $\frac{1}{n}\sum_{i=1}^{n}(x_i - k)^2$ es mínimo cuando $k = \overline{x}$

Inconvenientes

- Para datos agrupados en intervalos (variables continuas) su valor oscila en función de la cantidad y amplitud de los intervalos que se consideren.
- Es una medida a cuyo significado afecta sobremanera la dispersión, de modo que cuanto menos homogéneos sean los datos, menos información proporciona
- En el cálculo de la media no todos los valores contribuyen de la misma manera. Los valores altos tienen más peso que los valores cercanos a cero

Mediana

Representa el valor de la variable de posición central en un conjunto de datos **ordenados.**

Existen dos métodos para el cálculo de la mediana:

1. Considerando los datos en forma individual, sin agruparlos.
2. Utilizando los datos agrupados en intervalos de clase.

Propiedades de la mediana

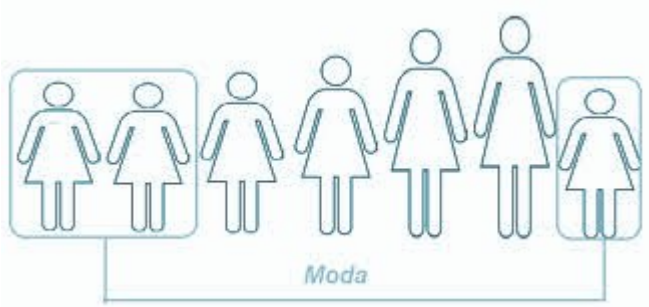
- Fácil de calcular si el número de observaciones no es muy grande.
- No se ve influenciada por valores extremos, ya que solo influyen los valores centrales.
- Fácil de entender.
- Se puede calcular para cualquier tipos de datos cuantitativos, incluso los datos con clase de extremo abierto.
- Es la medida de tendencia central más representativa en el caso de variables que solo admiten la escala ordinal.

Inconvenientes

- No utiliza en su “cálculo” toda la información disponible.
- No pondera cada valor por el número de veces que se ha repetido.
- Hay que ordenar los datos antes de determinarla.

Moda

La moda es la observación que se presenta con mayor frecuencia en la muestra



Propiedades de la moda

- No requiere cálculos.
- Puede usarse para datos tanto cuantitativos como cualitativos.
- Fácil de interpretar.
- No se ve influenciada por valores extremos.
- Se puede calcular en clases de extremo abierto.

Inconvenientes

- Para conjuntos pequeños de datos su valor no tiene casi utilidad, si es que de hecho existe. Solo tiene significado en el caso de una gran cantidad de datos.
- No utiliza toda la información disponible.
- No siempre existe, si los datos no se repiten.
- En ocasiones, el azar hace que una sola observación no represente el valor más frecuente del conjunto de datos.
- Difícil de interpretar si los datos tiene 3 o más modas.

El engañoso término medio

Martín Gardner en su libro “¡Aja! Paradojas” nos cuenta la historia de la fábrica del Sr. Artilugio (PRODILUGIO S.A.) y su nuevo empleado Félix, quien fuera víctima de los engaños de los parámetros estadísticos de posición. La dirección de PRODILUGIO está a cargo del Sr. Art ilugio, su hermano y seis parientes. La fuerza laboral consiste en cinco encargados y diez operarios. Los negocios van bien y la fábrica precisa un operario más. El Sr. Artilugio entrevista a Félix, candidato al puesto, y le explica que su empresa paga muy bien, ya que el salario medio es de $600 semanales. Al cabo de unos cuantos días, Félix quiso ver al jefe. Este fue su diálogo: Félix: -¡Me ha engañado usted! He hablado con los otros operarios y ninguno gana más de $200 a la semana. ¿Cómo puede ser que el salario medio sea de $600? Sr. Artilugio: -Vamos Félix, no se excite. El salario medio es de $600 y se lo voy a demostrar. Vea esta nómina por favor.

Empleado	Sueldo semanal
Sr. Artilugio	$4800
Hermano del Sr. Artilugio	$2000
6 parientes	$500 (c/u)
5 capataces	$400 (c/u)
10 operarios	$200 (c/u)
Total 23 empleados:	**$13.800**

El sueldo promedio **resulta entonces:** $\frac{\$13800}{23} = \$600$

Félix consideró que, si bien era cierto que el sueldo medio resultaba de $600, de todas formas lo habían engañado.

Pero el Sr. Artilugio insistió y le propuso ordenar los sueldos de mayor a menor:

4800, 2000, 500, 500, 500, 500, 500, 500, 400, 400, 400, **400**, 400, 200, 200, 200, 200, 200, 200, 200, 200, 200, 200.

El valor 400 indicado en rojo, ocupa **el lugar central** en esa sucesión (existe la misma cantidad de valores de la variable sueldo por encima que por debajo de él). El Sr. Artilugio le explicó a Félix que esa era la **mediana**.

Pero Félix continuaba insatisfecho con estos argumentos y volvió a preguntar ya un poco alterado:

Félix: -**¿Y qué significan entonces los $200?**

Sr. Artilugio: -Eso, muchacho, se llama **moda. Y es el salario ganado por el mayor número de personas.**

Sin duda, después de la experiencia de Félix, la representatividad de la media, la moda y la mediana para caracterizar una situación (en este caso, la situación salarial de los empleados del Sr. Artilugio) será un motivo de análisis y discusión.

Medidas de variabilidad

La localización o tendencia central no necesariamente proporciona información suficiente para describir datos de manera adecuada.

Las medidas de variabilidad nos informan sobre el grado de concentración o dispersión que presentan los datos respecto a su promedio.

Clasificación

Medidas dimensionales

- Rango
- Rango intercuartílico
- Varianza
- Desviación típica
- covarianza

Medidas adimensionales

- Coeficiente de Pearson
- Rango intercuartílico
- Varianza
- Desviación típica
- Covarianza

Rango de la muestra y rango intercuartílico

Una medida muy sencilla de variabilidad es el rango de la muestra , definido como la diferencia entre las observaciones más grande y más pequeña. Es decir

$$R = Max\ x_i - Min\ x_i$$

Inconvenientes del rango

El rango ignora toda la información que hay en la muestra entre las observaciones más chica y más grande.
Además es muy sensible a los datos de los extremos por lo cual en algunas ocasiones es recomendable usar el rango intercuartílico. Éste se construye encontrando los cuartiles de la muestra: q1, q2 y q3. En q1 tendremos el 25% de los datos más pequeños, luego de q2 el 50%, etc. Es así que el rango intercuartílico, representado por q3-q1 deja afuera de su cálculo esos valores extremos, siendo menos sensible a éstos

Desviación estándar y Varianza Muestral

La desviación estándar mide cuán lejos se encuentran los datos de la media muestral. Un modo de medir la variabilidad de los datos de una muestra sería tomar algún valor central, por ejemplo la media, y calcular el promedio de las distancias a ella. Mientras mayor sea este promedio, más dispersión deberían presentar los datos.

Sin embargo, esta idea no resulta útil, ya que las observaciones que se encuentran a la derecha de la media tendrán distancias (o desviaciones) positivas, en tanto que las observaciones menores que la media tendrán distancias negativas y la suma de las distancias a la media será inevitablemente igual a cero. Un modo de evitar este inconveniente es elevar las distancias al cuadrado y de este modo tener todos sumandos positivos.

Desviación estándar y Varianza Muestral

Definimos la varianza de una muestra de observaciones X_1 , X_2 , ..., X_n , cuya media es X , como

$$S^2 = \frac{\sum_i (Xi - \bar{X})^2}{n}$$

Monitorear la varianza es esencial en las industrias de manufactura y control de calidad, porque con la reducción de la varianza del proceso aumenta la precisión y disminuye el número de defectos

También se define con la siguiente ecuación

Corrección de Bessel

$$S^2 = \frac{\sum_{i=1}^{n} (X_j - \overline{X})^2}{n-1}$$

La razón para usar (n – 1) y no n en el denominador de la varianza muestral tiene que ver con el hecho de que el valor de s^2 obtenido en una muestra, se usa para estimar la varianza poblacional σ^2 . Definida con (n – 1) en el denominador la varianza muestral posee una propiedad deseable, resulta ser insesgado , esto es, en promedio no subestima ni sobrestima el valor de la varianza poblacional. Revisaremos esto en la 3° clase.

Desviación estándar

La varianza muestral puede pensarse como “promedio” de las distancias a la media al cuadrado.
Sin embargo, la varianza no tiene las mismas unidades que los datos. Para salvar este inconveniente, definimos la desviación estándar muestral como la raíz cuadrada positiva de la varianza

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

Volviendo al problema de los sueldos

Recordemos como se distribuían los sueldos en la empresa del Sr. Artilugio:

Empleado	Sueldo semanal
Sr. Artilugio	$4800
Hermano del Sr. Artilugio	$2000
6 parientes	$500 (c/u)
5 capataces	$400 (c/u)
10 operarios	$200 (c/u)
Total 23 empleados:	**$13.800**

Allí, el salario promedio era de $600, sin embargo, Félix (y nosotros) vimos que no era representativo dada la gran dispersión de valores de sueldo que existe con respecto a esa media. Ahora vamos a calcular esa dispersión

$$\sigma^2 = \frac{(4800-600)^2 + (2000-600)^2 + (500-600)^2.6 + (400-600)^2.5 + (200-600)^2.10}{23}$$

$$\sigma^2 \cong 933.043{,}5 \Rightarrow \sigma = \sqrt{933.043{,}5} \Rightarrow \sigma = 965.94$$

Este valor tan grande de la desviación estándar (incluso superior al promedio) pone aun más en evidencia la falta de representatividad del sueldo medio de $600 en la empresa del Sr. Artilugio. Cualquier distribución de sueldos que usted proponga con menor dispersión de valores con respecto a la media (esto es con un desvío estándar menor) tendrá una media más representativa de la situación que la de esta empresa.

Correlación

En muchas ocasiones es necesario estudiar conjuntamente dos características de un fenómeno aleatorio, es decir, el comportamiento conjunto de dos variables aleatorias, intentando explicar la posible relación existente entre ellas.

La correlación trata de establecer la relación o dependencia que existe entre las dos variables que intervienen en una distribución bidimensional.

Cuando se realiza un análisis de información una de las herramientas más potentes para poder extraer conclusiones es realizar correlaciones.

Relaciones entre variables numéricas

La existencia de algún tipo de asociación entre dos o más variables representa la presencia de algún tipo de tendencia o patrón de emparejamiento entre los distintos valores de esas variables.

Complementariamente, se habla de independencia entre variables cuando no existe tal patrón de relación entre los valores de las mismas.

Tablas de Contingencia

Cuando se trabaja con variables categóricas, los datos suelen organizarse en tablas de doble entrada en las que cada entrada representa un criterio de clasificación (una variable categórica). Como resultado de esta clasificación, las frecuencias (el número o porcentaje de casos) aparecen organizadas en casillas que contienen información sobre la relación existente entre ambos criterios. A estas tablas de frecuencias se les llama **tablas de contingencia**.

La figura muestra el gráfico de barras agrupadas correspondiente a los datos de la tabla. Cada barra del gráfico se corresponde con una casilla de la tabla. Podemos trasladar más de una variable tanto a la lista filas como a la lista columnas. En ese caso, cada variable fila se cruza con cada variable columna para formar una tabla de contingencia distinta. Seleccionando, por ejemplo, dos variables fila y tres variables columna, obtendríamos seis tablas de contingencia diferente

Variables aleatorias independientes

Intuitivamente decimos que dos variables, X e Y, son independientes si el valor que toma una de ellas no influye de ninguna manera sobre el valor que toma la otra. Esto lo establecemos más formalmente:

Sea (X , Y) una variable aleatoria bidimensional discreta. Sea p (xi , y j) su fdp conjunta y p (x_i) y q (y_j) las correspondientes fdp marginales de X e Y. Decimos que X e Y son variables aleatorias independientes si y sólo si

$$p(x_i, y_j) = p(x_i)\, q(y_j) \;\forall\; (x_i, y_j) \in R_{XY}$$

Covarianza

Es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias. Es el dato básico para determinar si existe una dependencia entre ambas variables y además es el dato necesario para estimar otros parámetros básicos, como el coeficiente de correlación lineal o la recta de regresión.

$$Cov(X,Y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Propiedades e inconvenientes

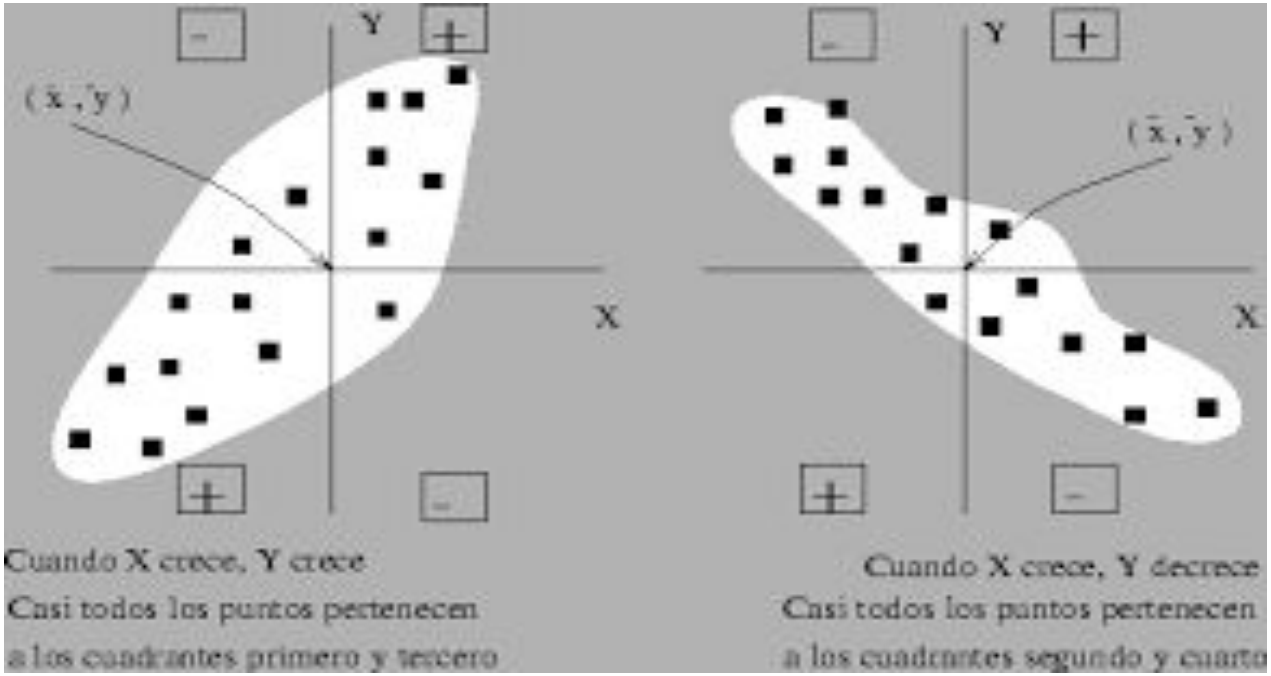
La covarianza permite estimar conceptos relativos a la correlación entre las dos variables

I. Su signo indica el sentido de la correlación entre las variables.
- Si $Cov_{xy} > 0$, la correlación es directa.
- Si $Cov_{xy} < 0$, la correlación es inversa.

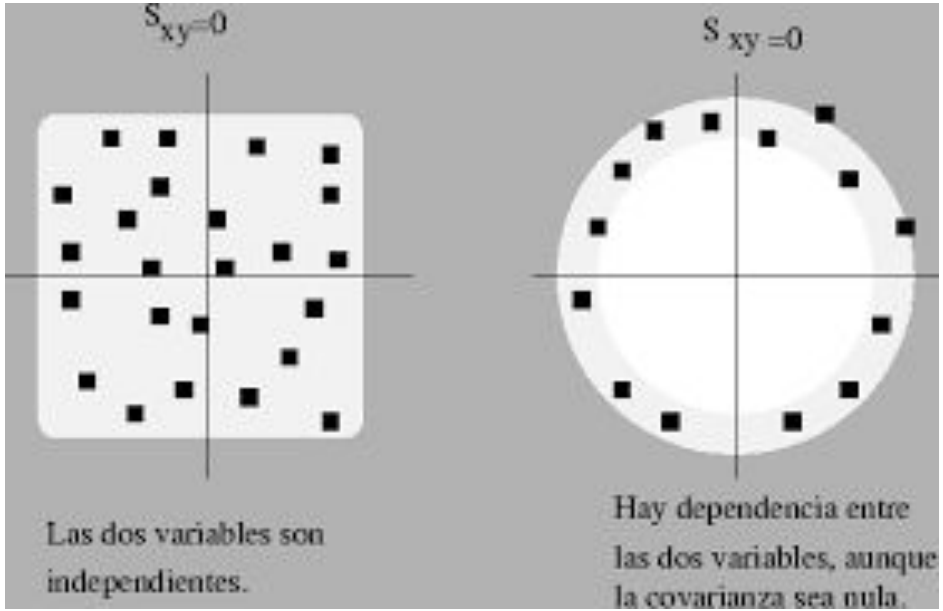
II. Un valor grande de Cov_{xy} advierte que la correlación entre las variables puede ser fuerte, pero no lo asegura, no siendo interesante la comparación de dos distribuciones por la covarianza.

Sólo da el sentido de la correlación: directa si es positiva e inversa si es negativo.

Gráficamente



Mas gráficos



Pearson

Sea (X , Y) una variable aleatoria bidimensional. Definimos el coeficiente de correlación lineal entre X e Y como

$$r = r_{xy} = \frac{\mathrm{Cov}(x, y)}{S_x \times S_y}$$

Este coeficiente nos da una idea del grado de asociación entre las variables aleatorias X e Y

Interpretación

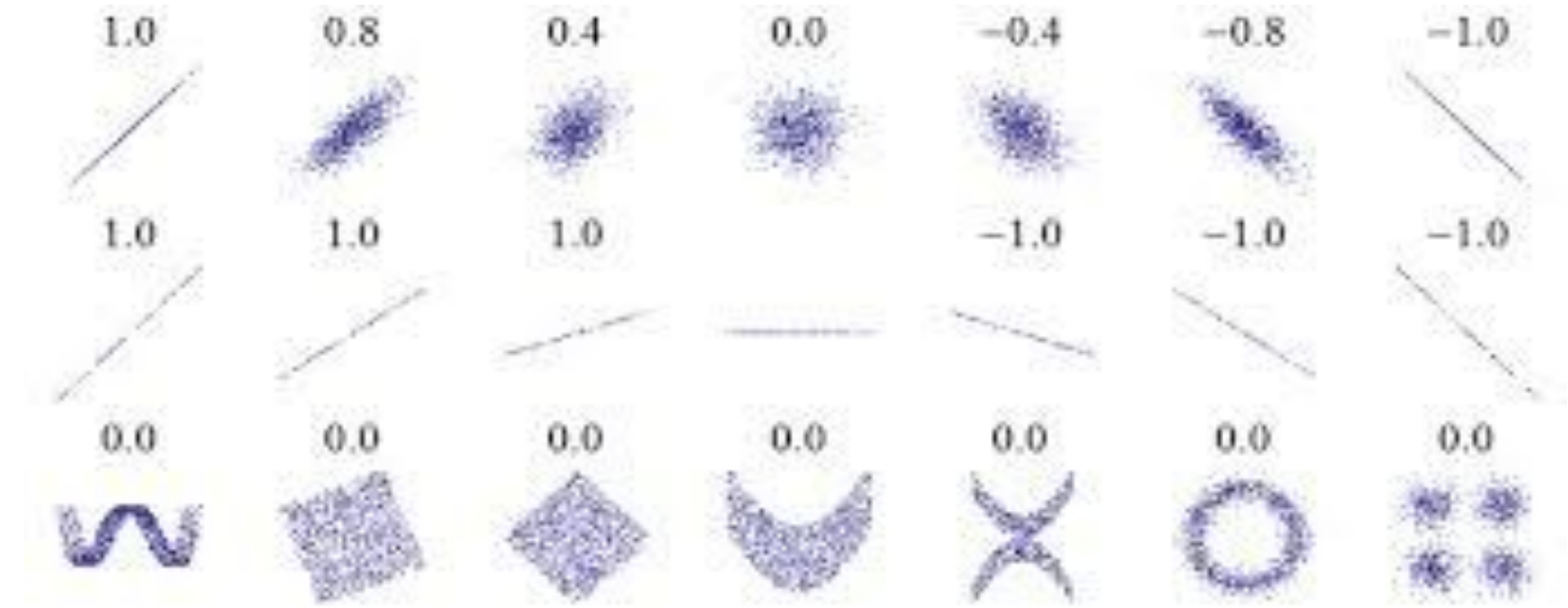
El valor del índice de correlación varía en el intervalo [-1,1], indicando el signo el sentido de la relación:

- Si $r = 1$, existe una correlación positiva perfecta. El índice indica una dependencia total entre las dos variables denominada *relación directa*: cuando una de ellas aumenta, la otra también lo hace en proporción constante.
- Si $0 < r < 1$, existe una correlación positiva.
- Si $r = 0$, no existe relación lineal. Pero esto no necesariamente implica que las variables son independientes: pueden existir todavía relaciones no lineales entre las dos variables.

Interpretación

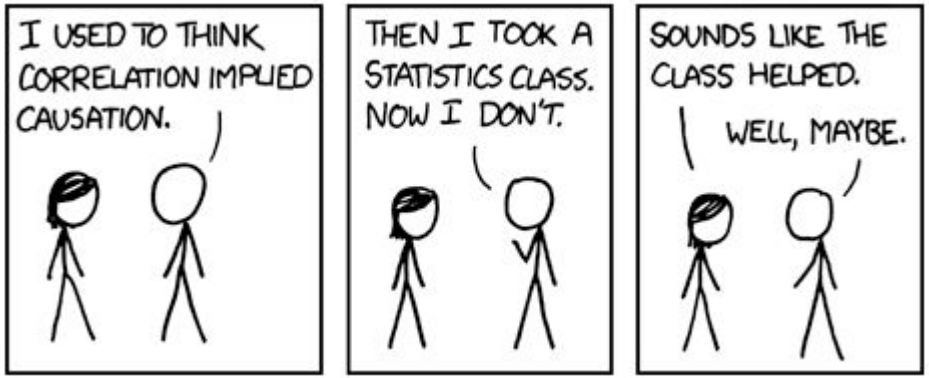
- Si $-1 < r < 0$, existe una correlación negativa.
- Si $r = -1$, existe una correlación negativa perfecta. El índice indica una dependencia total entre las dos variables llamada *relación inversa*: cuando una de ellas aumenta, la otra disminuye en proporción constante.

Gráficamente



No todo es magia

Recuerde que la correlación no implica causalidad. Por ejemplo, si las ventas de helados están correlacionadas positivamente con los ataques de los tiburones a los nadadores, eso no significa que el consumo de helados de alguna manera hace que los tiburones ataquen. Otra variable, como el clima cálido, puede provocar un aumento tanto en las ventas de helados como en las visitas a las playas.



¿Dudas?