Análisis y Visualización de Datos

Diplomatura CDAAyA 2021

Primero: ¿cuál es el problema?

¿Cuánto voy a cobrar?

Implementar un sistema que, dadas las características de una persona, devuelva el sueldo que puede esperar cobrar como programador(e)

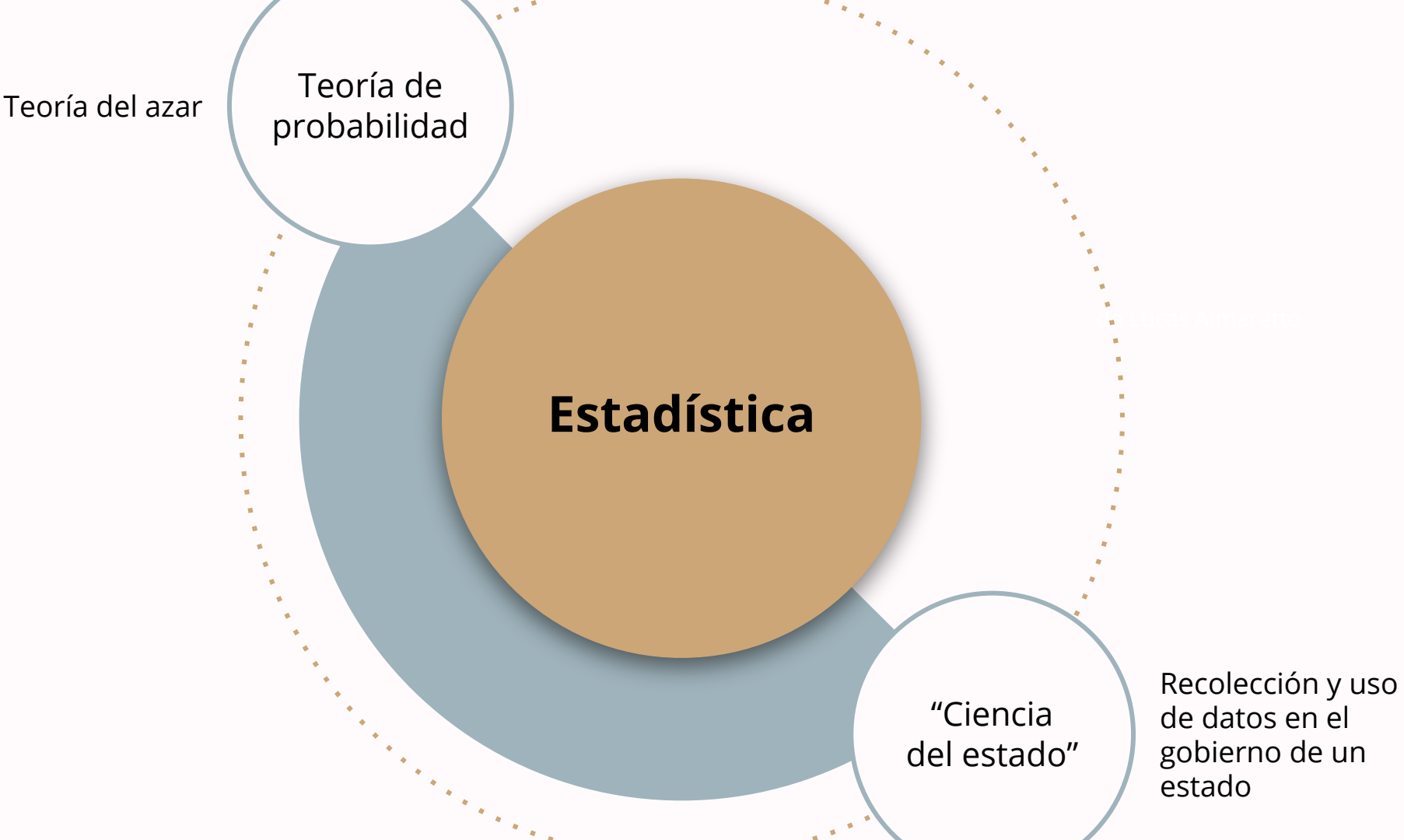
Encuesta Sysarmy

- Encuesta personal y voluntaria que busca relevar información sobre salarios y condiciones de trabajo de programadores, que se realiza anualmente.
- Usaremos sólo los datos provenientes de Argentina
- Link a los datos

Escapen... Yo ya estoy hace muchos años... Estoy jodido.

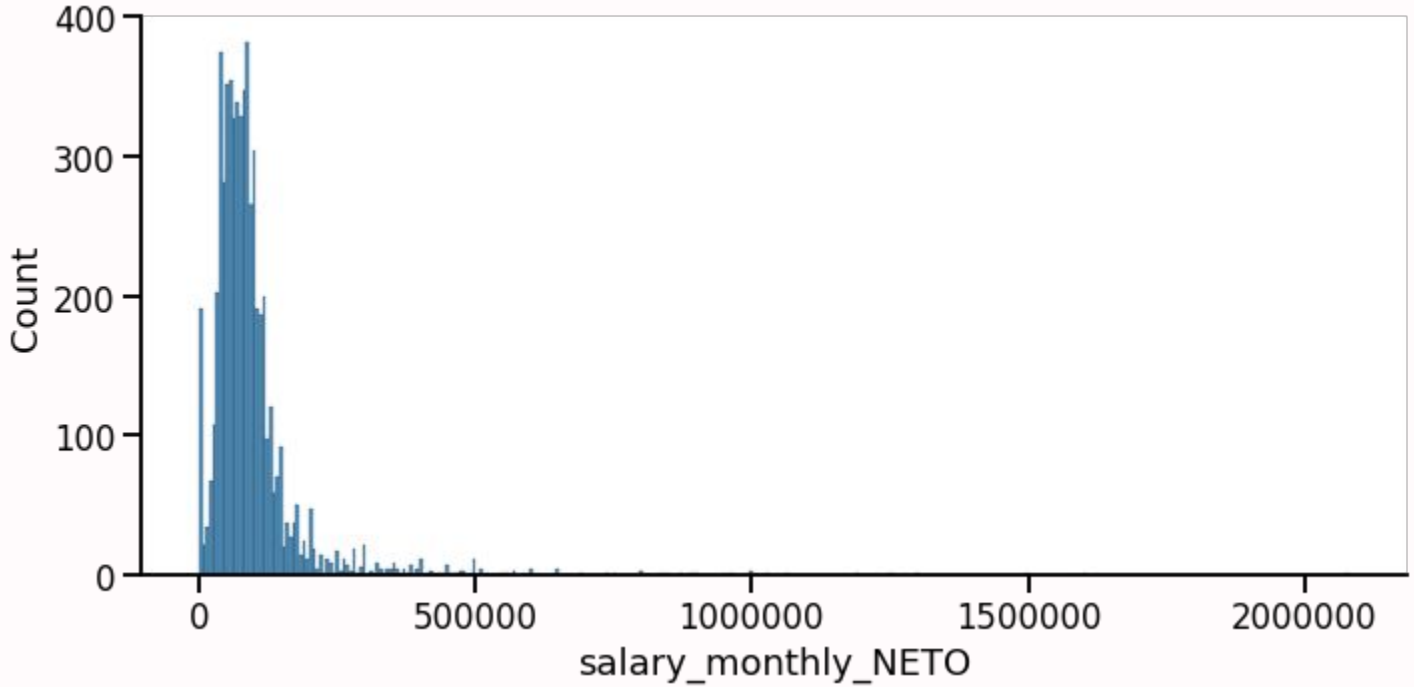
Demo con Notebook

01 Probabilidad.ipynb


Teoría del azar
Teoría de probabilidad
Estadística
“Ciencia del estado”
Recolección y uso de datos en el gobierno de un estado

Utilidad de la Estadística

- **Descripción de datos**
- **Análisis de muestras**
- Contrastación de Hipótesis, toma de decisiones
- Medición de relaciones
- Inferencia
- Predicción



¿Cuál es el concepto matemático que usamos para modelar la columna salary_monthly_NETO?

Variable Aleatoria

Una **variable aleatoria (v.a.)** X es una función

$$X: \Omega \rightarrow R_X$$

donde Ω es un conjunto llamado **Espacio de estados** y R_X es un conjunto de valores que toma la variable llamado **Rango.**

Una **variable aleatoria** **(v.a.)** X es una función

$$X: \Omega \rightarrow R_X$$

donde Ω es un conjunto llamado **Espacio de estados** y R_X es un conjunto de valores que toma la variable llamado **Rango.**

X es la v.a. salary_monthly_NETO, que toma una persona que programador en Argentina en 2020 que hizo la encuesta y devuelve su salario mensual neto.

¿Se puede ser más específico?

Una **variable aleatoria** **(v.a.)** X es una función

$$X: \Omega \rightarrow R_X$$

donde Ω es un conjunto llamado **Espacio de estados** y R_X es un conjunto de valores que toma la variable llamado **Rango.**

El espacio de estados Ω es el conjunto de valores posibles (personas) que podríamos haber encontrado en nuestra encuesta.

Ω = {ω / ω es una persona viva que trabaja en Argentina}

Puede tener más de una definición:

Ω = {ω / ω es una persona viva que trabaja en Argentina como desarrollador(e)}

Una **variable aleatoria** **(v.a.)** X es una función

$$X: \Omega \rightarrow R_X$$

donde Ω es un conjunto llamado **Espacio de estados** y R_X es un conjunto de valores que toma la variable llamado **Rango.**

El rango R_x es el conjunto de valores posibles de salary_monthly_NETO.

$R_x = \mathbb{R}$? (conjunto de números reales)

$R_x = \mathbb{N}$? (conjunto de números naturales)

¿Cómo podemos calcular el rango de R_x en la encuesta?

Una **variable aleatoria (v.a.)** X es una función

$$X: \Omega \rightarrow R_X$$

donde Ω es un conjunto llamado **Espacio de estados** y R_X es un conjunto de valores que toma la variable llamado **Rango.**

ω = Persona que respondió primero

X(ω) = 43000.0

X(ω) se denomina **realización** de la v.a. X

Variable Aleatoria - Otros ejemplos

X	Ω (universo en el cual vamos a estar midiendo cosas)	R_X
horas diarias que trabaja	personas que son programadores …	1 - 24
cantidad de globulos rojos en sangre	personas	valores en numeros reales.

Tipos de variables aleatorias

Las variables aleatorias pueden ser de distinto tipo, de acuerdo a los valores presentes en el Rango y su interpretación.

- Numéricas
  - Continuas
  - Discretas (un conjunto finito o infinito numerable de valores posibles)
- Categóricas
- Ordinales

Determinar los tipos de datos/variable que estamos usando nos permite seleccionar las herramientas adecuadas para obtener información a partir de ellos

Hagamos una pregunta interesante: ¿Tener más años de experiencia significa que se cobra más?

¿Cómo hacer este análisis?

Si no planteamos una hipótesis primero, es difícil determinar qué pasos hay que seguir para poder hacer el análisis

Una vez que la hipótesis está definida, hay que determinar QUÉ hay que medir para poder comprobarla.

Una vez que está definido qué medir, se seleccionan las herramientas para medirlo.

¿Cómo hacer este análisis?

Tener más años de experiencia significa que se cobra más

`salary_monthly_NETO`
`profile_years_experience`

????

Teoría de probabilidad

¿Por qué podemos usar probabilidad?

Cuando hablamos de hacer ciencia de datos, lo que estamos buscando es poder **razonar** sobre fenómenos reales. E.T. Jaynes resume esta necesidad como:

1. Representar los grados de plausibilidad de los fenómenos usando números.
2. Correspondencia cualitativa con el sentido común.
3. Consistencia.

Teoría de probabilidad

Conjunto de herramientas matemáticas que nos permite razonar sobre **experimentos aleatorios**, que debe cumplir:

1. Puede repetirse infinitas veces con la misma configuración experimental.
2. Tiene un conjunto fijo de posibles resultados
3. Antes de realizarlo, no se puede predecir el resultado que va a obtenerse.

¿Se puede modelar esta encuesta como un experimento aleatorio? ¿Cómo?

Teoría de probabilidad

Los experimentos aleatorios se modelan utilizando el concepto de **espacio probabilístico**, compuesto de:

1. Un conjunto Ω llamado espacio muestral con todos los resultados posibles.
2. Un conjunto $\mathcal{F}$ de eventos que contiene los resultados efectivamente observados
3. Una función P: $\mathcal{F} \rightarrow$ [0, 1] que asigna a cada evento su **probabilidad**

Teoría de probabilidad

¿Podemos pensar en ejemplos de espacios probabilísticos en este y otros experimentos aleatorios?

V.A. X	Espacio muestral Ω	$\mathcal{F}$ conjunto de eventos

¿Probabilidad? - Interpretación axiomática

P es una **medida de Probabilidad** en el **espacio muestral Ω** si para cada subconjunto A de **Ω** , **P**(A) es un número tal que:

- **0** ≤ **P**(A) ≤ 1
- **P**(**Ω**) = 1
- **P**(A U B) = **P**(A) + **P**(B), para A y B disjuntos (o excluyente)
- **P**(U_i A_i) = $\sum_i$ **P**(A_i) para A_1, A_2,... disjuntos

¿Cómo se calcula?

Nuestro **Ω** son todas las respuestas de la encuesta, cada ω_i es una respuesta, y el conjunto A son las respuestas en la que el fenómeno ocurre.

Si cada una de nuestros eventos es **independiente e idénticamente distribuido**, es decir, que $P(\{\omega_i\}) = 1/k$, entonces la probabilidad de un conjunto $A \subset \Omega$ es la proporción de eventos en A.

$$P(\{\omega_i\}) = 1/k \implies |A|/k$$

Situaciones más complejas

Si hay dos situaciones a estudiar, entonces se modela el problema usando las columnas `salary_monthly_NETO` y `profile_years_experience` para crear conjuntos de eventos y comprobar si existe una relación entre ellos.

Los conjuntos que se eligen son los que determinan el **experimento**

- $A = \{ \omega_i : \texttt{salary\_monthly\_NETO} > \text{avg}(\texttt{salary\_monthly\_NETO}) \}$
- $B = \{ \omega_i : \texttt{profile\_years\_experience} > 5 \}$

Situaciones más complejas

A = { ω_i : `salary_monthly_NETO` > avg }

B = { ω_i : `profile_years_experience` > 5 }

La **probabilidad conjunta** de que ocurran ambos eventos al mismo tiempo se modela usando la intersección de los conjuntos:

$$P(A \cap B)$$

Situaciones más complejas

A = { ω_i : `salary_monthly_NETO` > avg }

B = { ω_i : `profile_years_experience` > 5 }

La **probabilidad condicional** de que el salario esté por encima del promedio, suponiendo que ocurre el evento de tener más de 5 años de experiencia, se calcula como:

$$P(B) \neq 0 \implies P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Situaciones más complejas

A = { ω_i : `salary_monthly_NETO` > avg }

B = { ω_i : `profile_years_experience` > 5 }

A y B se dicen conjuntos **independientes** si

$$P(A \cap B) = P(A)P(B)$$

$$P(B) \neq 0 \implies P(A|B) = P(A)$$

¿Si uno tiene más de 5 años de experiencia, la probabilidad de cobrar más que el promedio aumenta? ¿Estos eventos, son independientes?

Ejercicio + recreo

Volvemos a las: 19:55

¿Son independientes o no?

Teorema de Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Tiene muchas aplicaciones en la ciencia de datos, incluyendo el aprendizaje bayesiano, pero no profundizamos en este tema porque lo van a ver con mucho más detalle en materias siguientes, cuando vean el clasificador Naive Bayes.

Estadística Descriptiva

Estadística Descriptiva

Herramientas para resumir un conjunto de datos (modelados como realizaciones de una variable aleatoria), a través de ciertas medidas numéricas.

Representa la información de una manera distinta para facilitar su interpretación, pero no permite realizar predicciones o inferencias

Análisis de frecuencias

¿Cuánto ocurre cada uno de los valores de una v.a.?

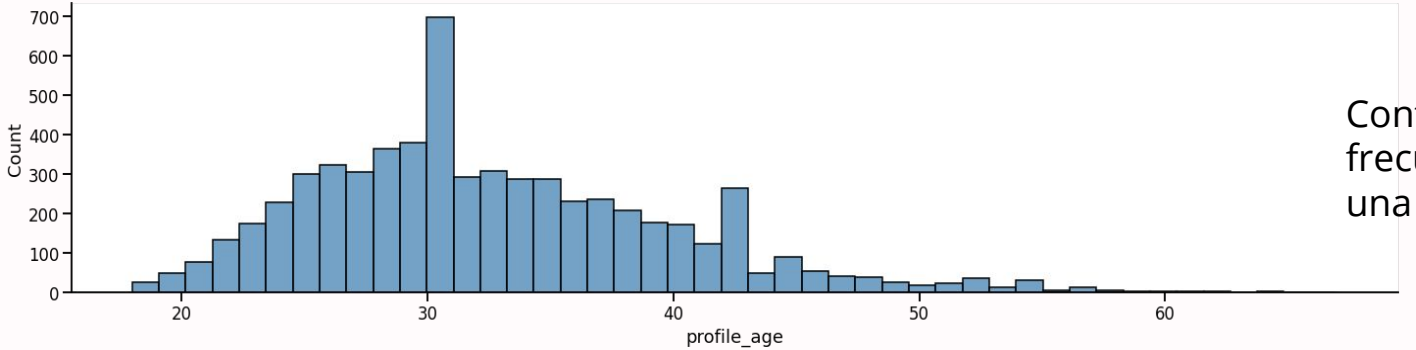
Medidas de tendencia central

¿Cuál es el valor más representativo de una v.a.?

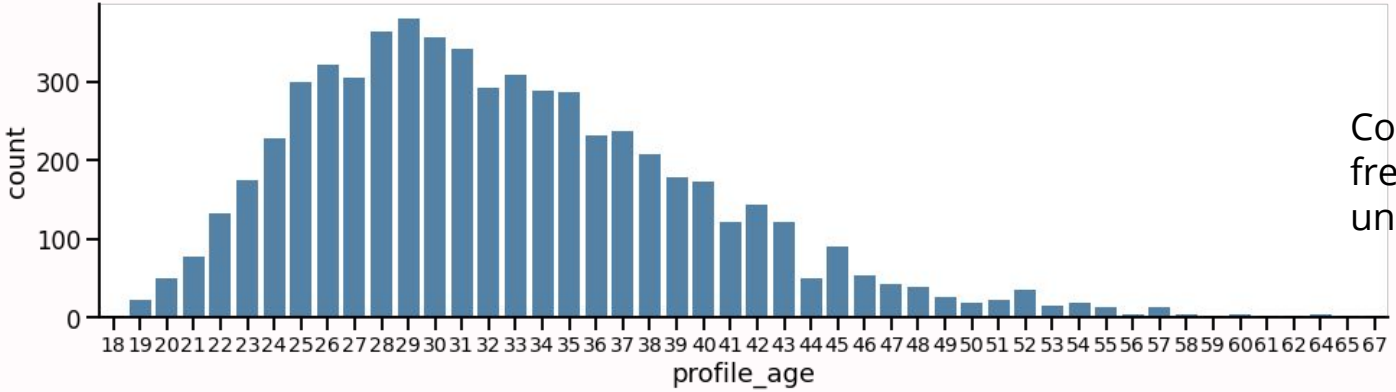
Medidas de dispersión

¿Qué tan alejados están los datos de la tendencia central?

Análisis de frecuencias



Conteo de frecuencias como una v.a. continua



Conteo de frecuencias como una v.a. categórica

En la página de opengube hay muchos análisis de frecuencias para poder ver

Medidas de tendencia central

Dada X una v.a **numérica** y un conjunto de realizaciones $x = \{x_1, x_2, ...\}$ donde $x_i = X(\omega)$ para algún $\omega \in \Omega$, y $N = |x|$

La **media muestral** (aritmética) o promedio se calcula como:

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

Existen otros tipos de media, pero esta es la más usada.

Medidas de tendencia central

Dada X una v.a **numérica** y un conjunto de realizaciones

$x = \{x_1, x_2, ...\}$

donde $x_i = X(\omega)$ para algún $\omega \in \Omega$, y $N = |x|$

La **mediana** se calcula como:

1. Ordenar las realizaciones de menor a mayor
2. Si N es impar, la mediana es el valor central:

$$median = x_{N/2}$$

3. Si N es par, la mediana es el promedio de los dos valores centrales:

$$median = \frac{1}{2}(x_{\lfloor N/2 \rfloor} + x_{\lfloor N/2 \rfloor + 1})$$

Medidas de tendencia central

Dada X una v.a **categórica** y un conjunto de realizaciones $x = \{x_1, x_2, ...\}$ donde $x_i = X(\omega)$ para algún $\omega \in \Omega$, y $N = |x|$

La **moda** son los valores con mayor frecuencia, es decir, los que más se repite.

Sólo hay más de una moda cuando el conteo de dos valores es igual.

Media aritmética	Mediana
Muy afectada por valores extremos	Poco afectada por valores extremos
Sólo es adecuada para las variables ordinales donde la distancia entre valores es lineal	Puede aplicarse a variables numéricas y a todas las variables ordinales
El cálculo tiene complejidad lineal, y puede aplicarse a grandes datos	El cálculo tiene complejidad n*log(n), ya que los datos deben ordenarse primero

Demo con Notebook

02 Estadística.ipynb

[Opcional] ¿Por qué la media se ve tan afectada por valores extremos?

Porque minimiza las desviaciones cuadráticas.

$$dc(k) = \frac{1}{N}\sum_{i=0}^{N}(x_i - k)^2 \qquad \min dc = \bar{x}$$

Si x_i es muy grande, entonces $(x_i - \text{media})^2$ tendrá mucho peso dentro de la función de desviaciones cuadráticas dc.

La mediana se ve menos afectada por los valores extremos, ya que siempre elige alguno de los valores cercanos al centro de la distribución.

Medidas de posición

Dada X una v.a **numérica** y un conjunto de realizaciones

$x = \{x_1, x_2, ...\}$

donde $x_i = X(\omega)$ para algún $\omega \in \Omega$, y $N = |x|$

El **percentil-k** de una conjunto x es el valor x_i tal que el k% de los valores de la muestra son menores a x_i.

No hay una única fórmula para calcular los percentiles, pero en general:

1. Ordenar las realizaciones tal que $x_j \leq x_{j+1}$
2. Seleccionar el elemento de la serie en la posición

$$n = \left\lceil \frac{P}{100} \times N \right\rceil .$$

Medidas de dispersión

Dada X una v.a **numérica** y un conjunto de realizaciones $x = \{x_1, x_2, ...\}$ donde $x_i = X(\omega)$ para algún $\omega \in \Omega$, y $N = |x|$

La **varianza muestral** mide la variación de los datos a través de la distancia cuadrada a la media muestral.

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

La **desviación estándar** es la raíz cuadrada de la varianza. Está en la misma unidad que los datos.

El **coeficiente de variación** es la desviación estándar dividida la media muestral. Es comparable entre distintas v.a.

Medidas de dispersión

Dada X una v.a

numérica y un conjunto

de realizaciones

$x = \{x_1, x_2, ...\}$

donde $x_i = X(\omega)$ para

algún $\omega \in \Omega$, y $N = |x|$

El rango y el rango intercuartílico miden en qué intervalo se encuentran un cierto porcentaje de los datos.

Rango:

percentil-100 - percentil-0

Rango intercuartílico:

percentil-75 - percentil-25

Q3 - Q1

Usos de los percentiles y rangos

- En el caso de la mediana (percentil-50), medir la tendencia central
- Contextualizar el valor de una realización con respecto a los otros

  Una persona de sexo femenino de 6 años mide 95cm

  Está en el 10% de personas con menor estatura del mismo grupo. [Curva]

- Identificación y eliminación de valores extremos

Interpretación de los estadísticos

Lo más común es interpretar las medidas o estadísticos más comunes (media y la desviación estándar) como si los datos siguieran una distribución normal, es decir, como si su histograma se pareciera a una campana de gauss.

Es necesario **seleccionar, comunicar e interpretar** los estadísticos teniendo en cuenta la forma de la distribución (asimetría y apuntalamiento). No es necesario tener en cuenta la cantidad de datos.

Asimetría

El **coeficiente de asimetría de Fisher** CA_F evalúa la proximidad de los datos a su media x. Cuanto mayor sea la suma $\sum(x_i - x)^3$, mayor será la asimetría. Sea la muestra $x_1, x_2, \ldots, x_N$, entonces la fórmula de la asimetría de Fisher es:

$$CA_F = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^3}{N \cdot S_x^3}$$

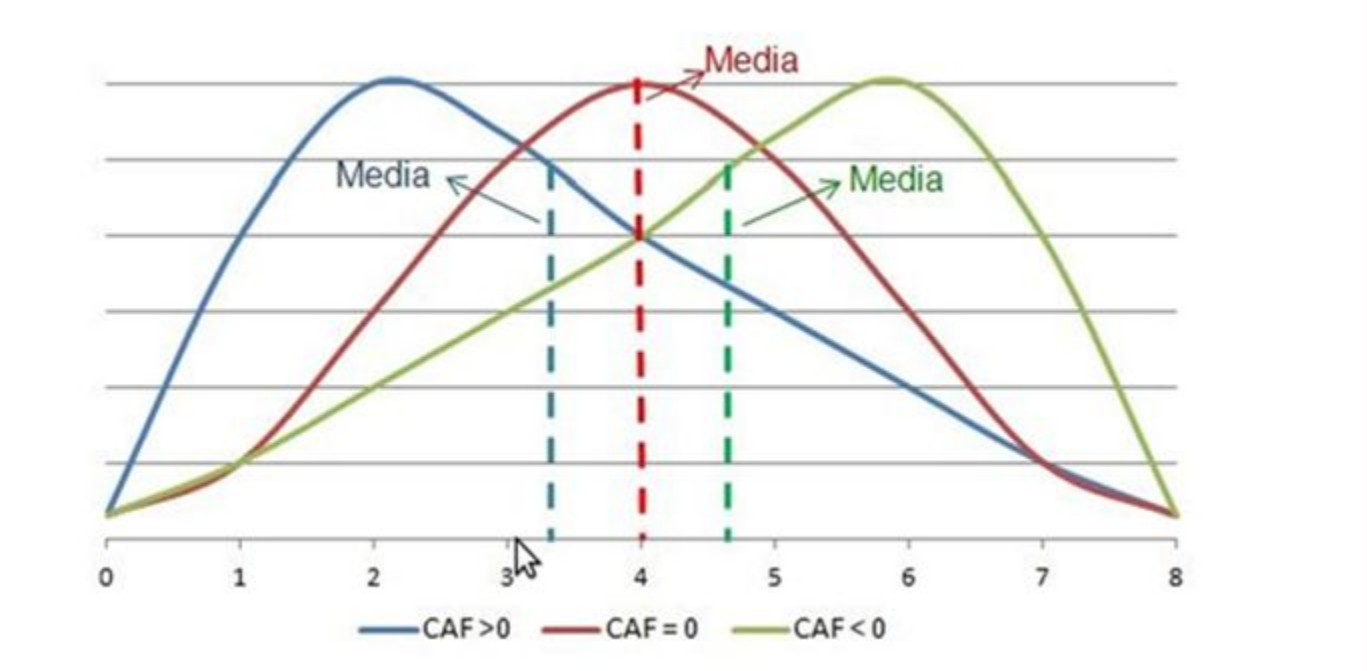
siendo $\overline{x}$ la media y S_x la desviación típica

Asimetría

La asimetría es la medida que indica la simetría de la distribución de una variable respecto a la media aritmética, sin necesidad de hacer la representación gráfica. Los coeficientes de asimetría indican si hay el mismo número de elementos a izquierda y derecha de la media.

- **Asimetría negativa**: la cola pronunciada a la izquierda
- **Simétrica**: colas balanceadas. Coinciden la media, la mediana y la moda. Distribución similar a la campana de Gauss, o distribución normal.
- **Asimetría positiva**: cola pronunciada a la derecha.

Asimetría



Curtosis o apuntalamiento

La **curtosis** de una variable estadística/aleatoria es una característica de forma de su distribución de frecuencias/probabilidad.
Una curtosis grande implica una mayor concentración de valores de la variable tanto muy cerca de la media de la distribución (pico) como muy lejos de ella (colas), al tiempo que existe una relativamente menor frecuencia de valores intermedios. Esto explica una forma de la distribución de frecuencias/probabilidad con colas más gruesas, con un centro más apuntado y una menor proporción de valores intermedios entre el pico y colas.

Curtosis o apuntalamiento

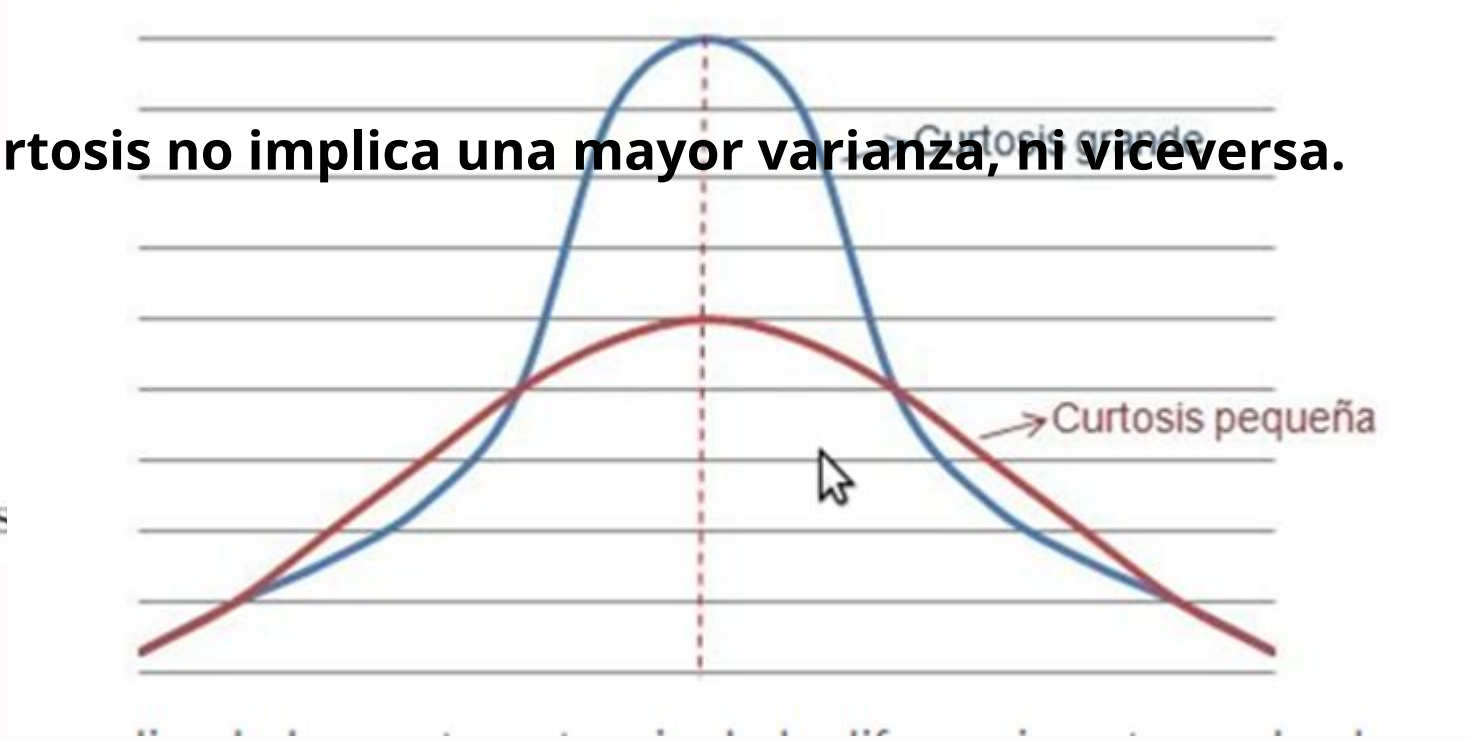
Un coeficiente de apuntamiento o de curtosis es el cuarto momento con respecto a la media estandarizado que se define como:

$$Curtosis = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^4}{N \cdot S_x^4} - 3$$

siendo $\overline{x}$ la media y S_x la desviación típica

Curtosis o apuntalamiento

Una mayor curtosis no implica una mayor varianza, ni viceversa.



En **una frase**:
¿Cuánto cobran les programadores en Argentina?

¿Respuestas?
¿Qué pregunta respondimos en realidad?

Población

El grupo completo de estados Ω
que se busca estudiar

¿Cuál es nuestra población?

Muestra

Un subconjunto de Ω elegido
para un experimento particular

¿Cuál es nuestra muestra?

Muestras sesgadas

Al estimar la medida de probabilidad como una proporción, estamos asumiendo que cada evento es independiente e idénticamente distribuido.

El proceso de selección de los eventos para un experimento determina las características de la muestra obtenida.

- Muestras convenientes (los que “estaban a mano”)
- Muestras de respuestas voluntarias

Muestras sesgadas

¿Qué sesgos tenemos en esta muestra?

- ...

¿Influyen en nuestra variable de estudio (el salario)?

¿Cómo afecta el nivel de estudios en el salario de les programadores en Argentina?

1. Ejercicio grupal, ingresar al meet designado
2. Seguir el proceso de análisis propuesto:
   a. Hipótesis
   b. Análisis de v.a.
   c. Experimento
3. Volvemos a las: ...