Inferencia Estadística
Clase 2

MIA
Magíster en
Inteligencia Artificial

IMT3860 – Introducción a Data Science
Profesor: Alejandro Cataldo Cornejo
correo: aecatald@uc.cl

AGENDA DE HOY

01 Presentación del Proyecto Aplicado.

02 Creación de Grupos

03 Inferencia Estadística

04 Taller 1

+ BREAK 20:00 (10-15')

Motivación

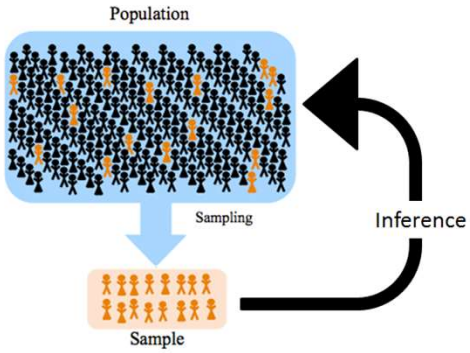
- Los procesos en nuestras vida y entorno generan datos ➔ los datos representan trazos de procesos que ocurren en el mundo real.
  - ¿cómo describimos y explicamos estos procesos para resolver los problemas que abordamos?

- Hay dos fuentes de aleatoriedad e incerteza:
  - La aleatoriedad e incerteza propia del proceso
  - La incerteza asociada a los métodos de recolección de datos

➔ Necesitamos procedimientos, métodos y teoremas que nos permitan extraer significado e información a partir de data generada por procesos estocásticos (aleatorios).

Motivación

- En DS, buscamos comúnmente generar conclusiones sobre una población, a partir de una muestra (ruidosa).



- Ejemplos:
  - ¿Quién va a ganar las elecciones?
  - Pronósticos meteorológicos
  - Comportamiento de consumidores o mercados
  - ¿Son precisas nuestras estimaciones sobre una cierta población?
  - ¿Cuál es el impacto de una determinada política?¿Hay un efecto real?

➔**Algunas dificultades**:

- ¿Es la muestra representativa de la población sobre la cual queremos inferir conclusiones?
- ¿Hay variables (no)observadas/(des)conocidas que contaminan nuestras conclusiones?
- ¿Hay sesgos (bias) sistemáticos debidos a datos faltantes, o diseño del estudio?
- ¿Qué aleatoriedad existe en los datos, y cómo las consideramos?
- ¿Estamos tratando de estimar un modelo subyacente del fenómeno estudiado?

Poblaciones y Muestras

- **Población:** cualquier universo de objetos o entidades (personas, imágenes, mensajes, tweets, etc.)

  N ➔ número de observaciones en la población

- **Muestra:** un subconjunto de la población, que usaremos para inferir conclusiones acerca de la población.

  n ➔ tamaño de la muestra

  ¿Cómo seleccionamos los n elementos de la muestra, de manera que las conclusiones obtenidas a partir de ella no sean erróneas y/o distorsionadas?

- En la era de los datos: ¿Observamos ***todo***? ¿necesitamos ***muestrear***?
  - Consideraciones de ingeniería: ¿almacenar todos los datos?
  - Bias: aún con grandes cantidades de data, sólo podemos hacer inferencias acerca de una cierta población, y se requiere conocer el contexto de los datos.
  - Cada proceso de muestreo introduce una incerteza.
  - Distintos tipos de datos requieren distintas estrategias de muestreo.

O'Neil & Schutt 2013, Doing Data Science..

Poblaciones y Muestras

- Es un error pensar que con conjuntos masivos de datos ("Big Data"),
  - entonces *N=all,*
  - o que es suficiente identificar correlaciones, e innecesario plantear modelos.

- ¿Qué entendemos por ***modelos***?
  - Una construcción o representación de la realidad a través de un lente particular (arquitectónico, biológico, matemático, etc.), en la cual se extraen o abstraen detalles "irrelevantes".

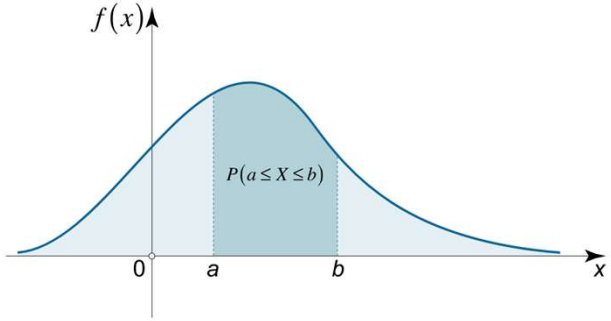
- Modelo estadístico:
  - ¿Cuál puede ser una aproximación al proceso subyacente a nuestros datos? ¿qué tiene influencia / causalidad sobre qué?
    - Expresiones matemáticas (suficientemente generales, de manera que incluyen parámetros cuyos valores son aún desconocidos), diagramas, etc.
  - ¿Cómo construimos el modelo? EDA, estadística descriptiva (cuantitativa/gráfica), construcción de funciones crecientemente complejas.

Variables Aleatorias

Conceptos Fundamentales de Probabilidades

Distribución de probabilidad (función de densidad)

- Los procesos naturales permiten generar mediciones que tienen asociada una incertidumbre ➔ *variables aleatorias*.
- Notación:
  - X → variable aleatoria
  - x → un valor particular de la variable aleatoria X



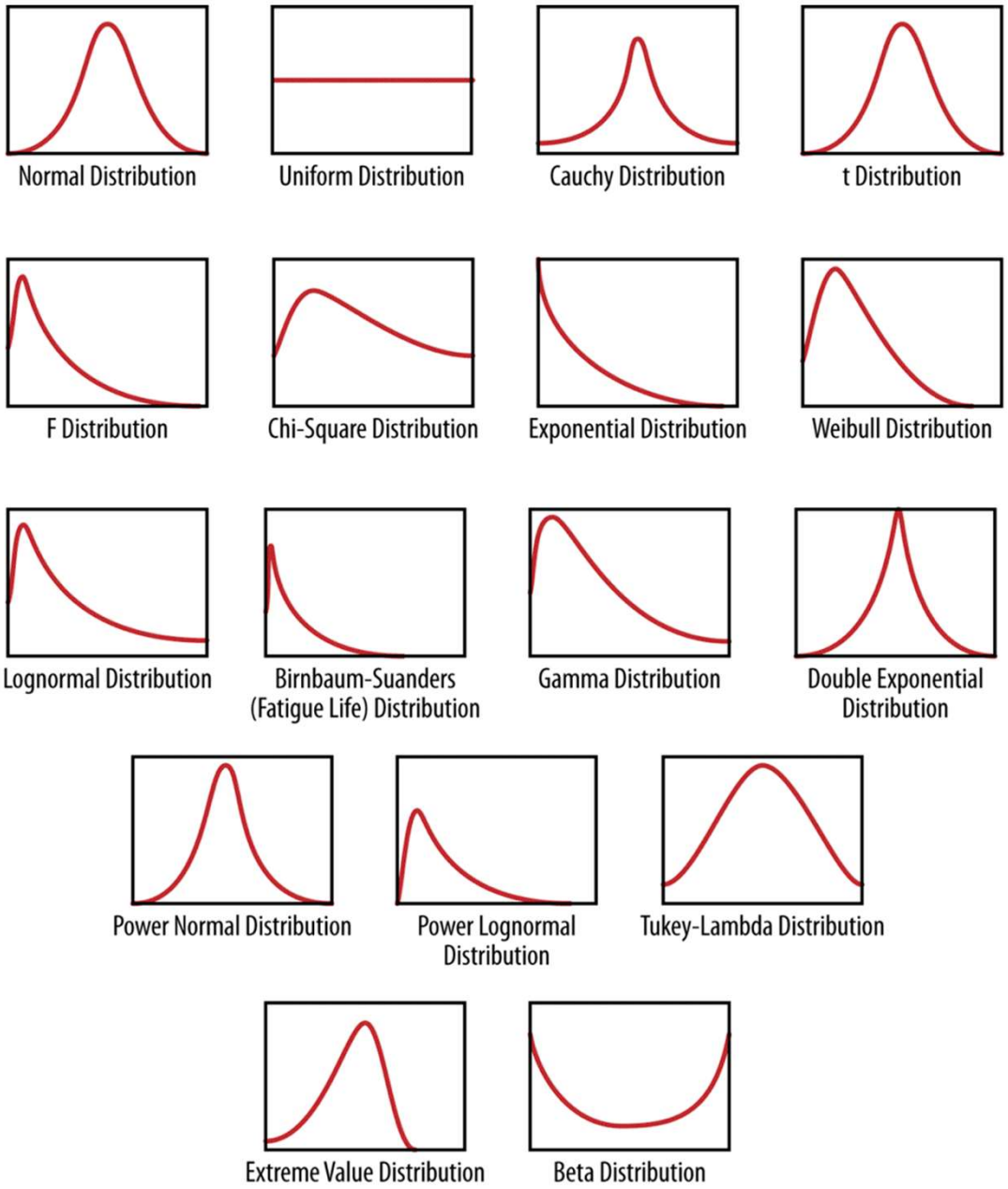
- Para una variable aleatoria x, la función de densidad de probabilidad (pdf) $f_X(x)$ de una variable aleatoria es una función que asigna a cada valor o evento de la variable, la probabilidad de que dicho valor ocurra.

$$f_X(x) = P(X = x)$$

Distribuciones de Probabilidad

- Los procesos naturales tienden a generar mediciones cuya forma empírica puede ser aproximada por funciones matemáticas con un conjunto reducido de parámetros, que pueden ser estimados a partir de datos.
  - Ej: Gauss, Poisson, Weibull, Gamma, exponencial.

- Estas funciones pueden utilizarse como elementos base de modelos estadísticos.

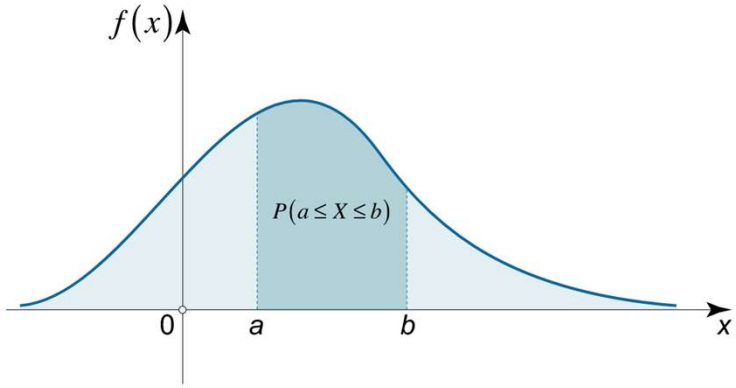


Conceptos Fundamentales de Probabilidades

Distribución de probabilidad

- El área bajo la curva entre dos valores a y b entrega la probabilidad de que la variable X tome un valor en ese rango:

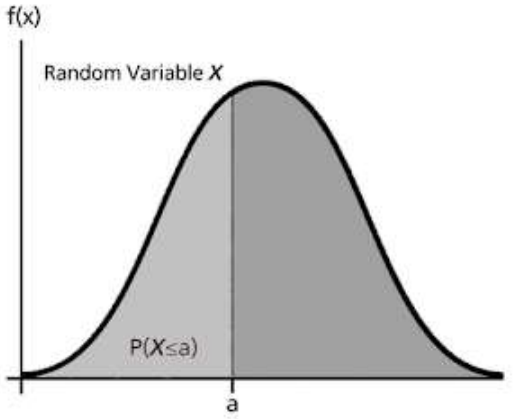
$$P[a < X \leq b] = \int_a^b f_X(x)\,dx$$



- Si integramos la función de densidad entre $-\infty$ y x, obtenemos la probabilidad de que X sea menor o igual a x ➔ función de distribución acumulada (cdf), $F_X(x)$

$$P[X \leq x] = \int_{z=-\infty}^{x} p_X(z)\,dz$$

$$= \sum_{z=-\infty}^{x} p_X(z)$$

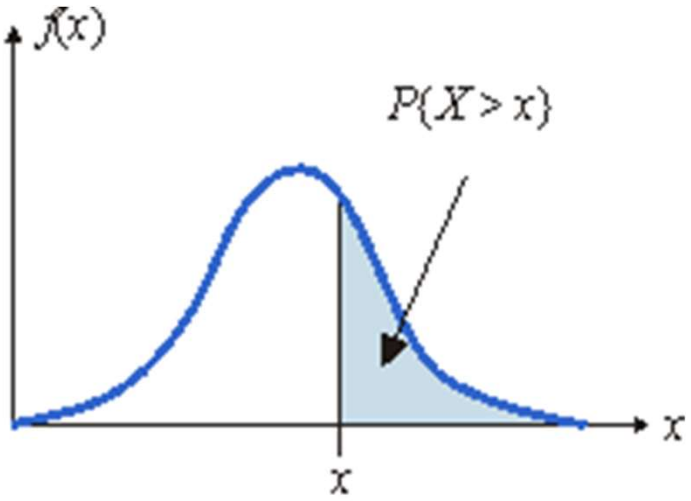


Conceptos Fundamentales de Probabilidades

Distribución de probabilidad

- Similarmente, la integral entre x y $+\infty$ entrega la probabilidad de exceder el valor $X = x$ ➔ probabilidad de excedencia (supervivencia)

$$P(X > x) = \int_{x}^{\infty} f_X(z)dz$$

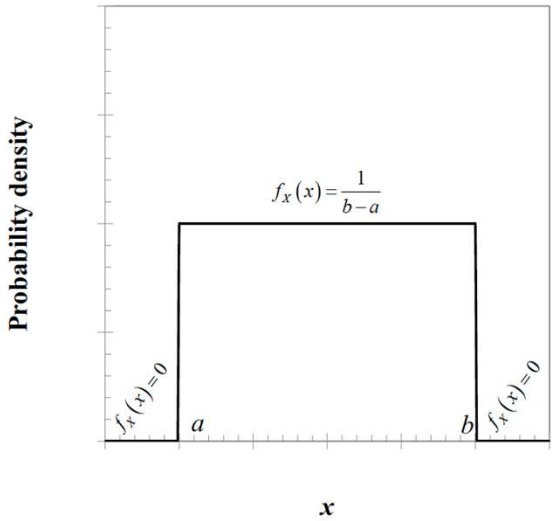


- Frecuentemente, es conveniente escribir las funciones de densidad de probabilidad como funciones paramétricas.
  - Ej: distribución uniforme, normal y lognormal.

Conceptos Fundamentales de Probabilidades

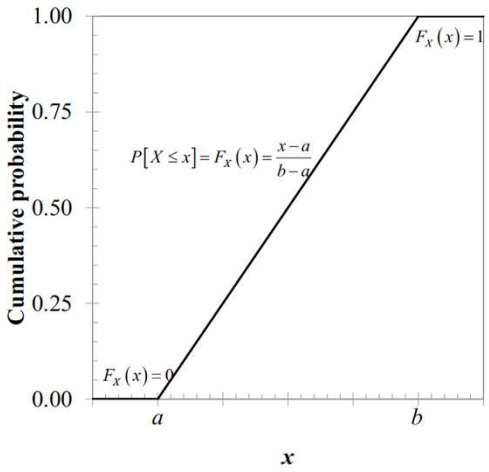
Distribución uniforme

- Una variable con distribución uniforme entre valores a y b puede tomar cualquier valor en ese rango con igual probabilidad.



- Función de densidad de probabilidad:

$$f_X(x)=0 \qquad x<a$$
$$=\frac{1}{b-a} \qquad a\le x\le b$$
$$=0 \qquad x>b$$



- Distribución acumulada:

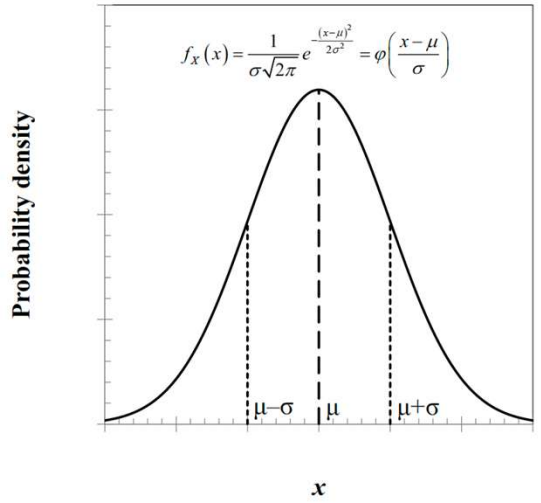
$$F_X(x)=0 \qquad x<a$$
$$=\frac{x-a}{b-a} \qquad a\le x\le b$$
$$=1 \qquad x>b$$

Conceptos Fundamentales de Probabilidades

Distribución normal (Gaussiana)

- Permite modelar numerosos fenómenos naturales, sociales y psicológicos.

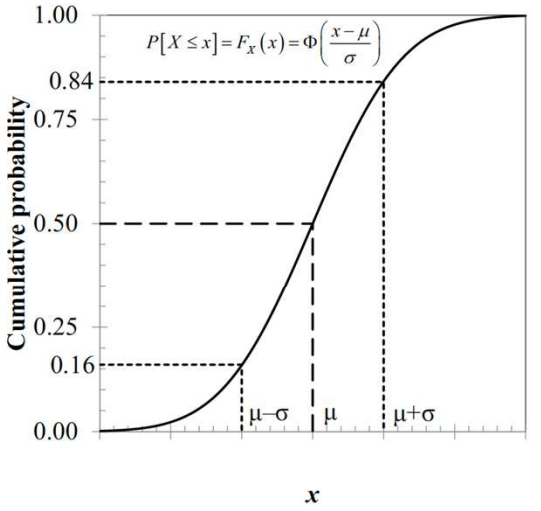
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \varphi\left(\frac{x-\mu}{\sigma}\right)$$



- Parámetros:

μ ➔ valor medio esperado, σ ➔ desviación estándar.

  - X puede tomar cualquier valor entre $-\infty$ y $+\infty$
  - μ puede tomar cualquier valor real
  - σ sólo puede tomar valores positivos. A mayor σ, mayor incerteza.



- Distribución acumulada (cdf): $P[X \le x] = F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz = \Phi\left(\frac{x-\mu}{\sigma}\right)$

Conceptos Fundamentales de Probabilidades

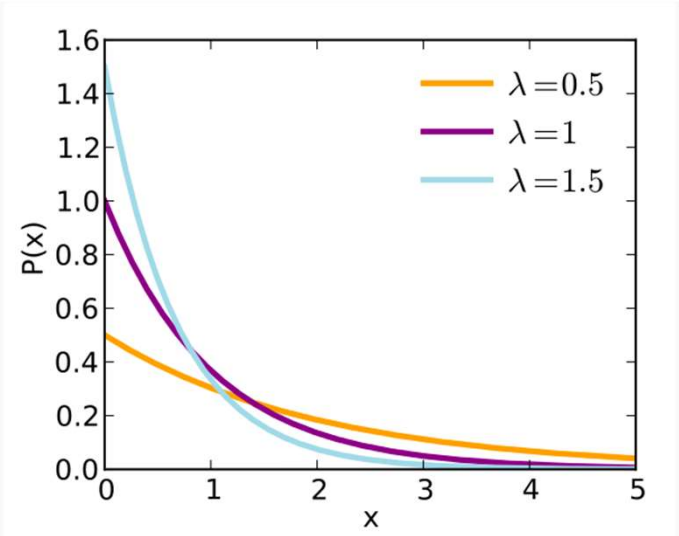
Distribución Exponencial

- Se utiliza típicamente para modelar tiempos de espera ante la ocurrencia de un cierto evento.

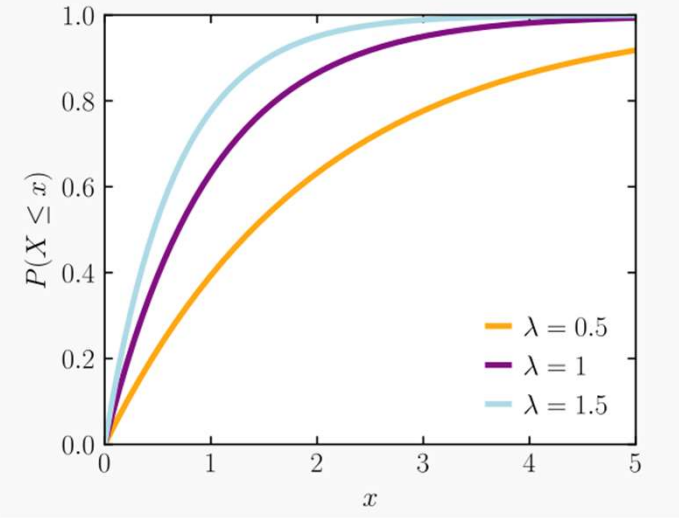
$$f(x) = \lambda e^{-\lambda x} \quad x \geq 0$$

- Distribución acumulada (cdf):

$$F(x) = 1 - e^{-\lambda x} \quad x \geq 0$$



Función de densidad de probabilidad



....etc

Conceptos Fundamentales de Probabilidades

Probabilidad condicional

- Dados dos sucesos A y B, con $P(B) > 0$, definimos la probabilidad de A condicionada a B, $\boldsymbol{P(A|B)}$, como la probabilidad de que ocurra A, sabiendo que ha ocurrido B.

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

- Si A y B son eventos independientes ➔ $\mathrm{P}(A \cap B) = P(A) \cdot P(B) \Rightarrow \mathrm{P}(A|B) = P(A)$

Conceptos Fundamentales de Probabilidades

Ejemplo: Probabilidad condicional

Consideremos una población en la que cada individuo es clasificado según dos criterios: es o no portador de un determinado virus, y pertenece o no al grupo de riesgo R (Ej.: mayores de 75 años).

La correspodiente tabla de probabilidad es:

	Portador (A)	**No portador (A^c)**	
Pertenece a R (B)	0.003	0.017	0.020
No pertenece a R (B^c)	0.003	0.977	0.980
	0.006	0.994	1.000

Dado que una persona seleccionada al azar pertenece al grupo de riesgo R, ¿Cuál es la probabilidad de que sea positivo para el virus?

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.003}{0.020} = 0.150$$

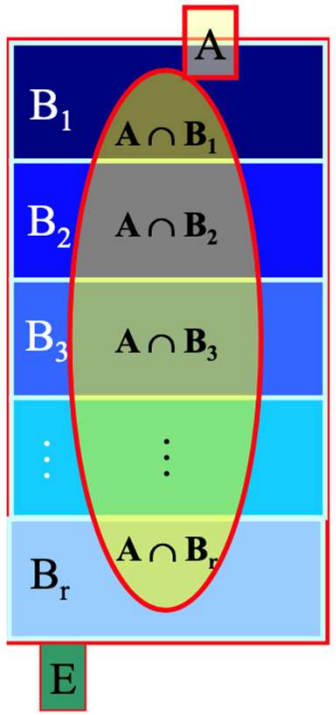
Conceptos Fundamentales de Probabilidades

Probabilidad total

- Sea $A_1, A_2, \ldots, A_n$ una partición sobre el espacio muestral y sea B un suceso cualquiera del que se conocen las probabilidades condicionales $\boldsymbol{P(B|A_i)}$, entonces la probabilidad del suceso B, $\boldsymbol{P(B)}$, viene dada por la expresión:

$$P(B) = P(A_1) \cdot P(B|A_1) + P(A_2) \cdot P(B|A_2) + \cdots + P(B|A_n)P(A_n)$$

$$= \sum_{i=1}^{n} P(A_i) \cdot P(B|A_i)$$



Distribuciones de Probabilidad y Modelos.

- Distribución conjunta: $\mathbf{p(x,y)}$
- Distribución condicional: $\mathbf{p(x|y)}$ ➔ distribución de probabilidad de x, dado un cierto valor de y.
  - En el contexto de datos, equivale a estudiar un subconjunto

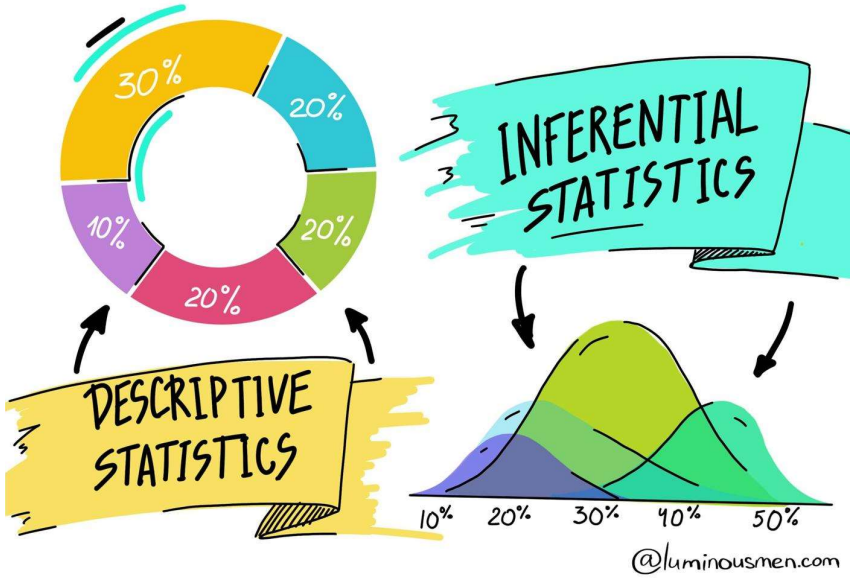
- **Ajustar un modelo:** estimar los parámetros de un modelo usando datos observados.
  - Usamos la data como evidencia para aproximar el proceso matemático real que genera los datos.
  - El ajuste de un modelo implica el uso de métodos y algoritmos de optimización como: min(Chi^2), máximum likelihood, etc.
  - Cuando estimamos los parámetros, éstos pasan a ser "estimadores" (a su vez son funciones de los datos).
  - **Sobreajuste (overfitting):** cuando usamos un dataset para estimar los parámetros de un modelo, pero el modelo no captura la realidad más allá de los datos de muestra

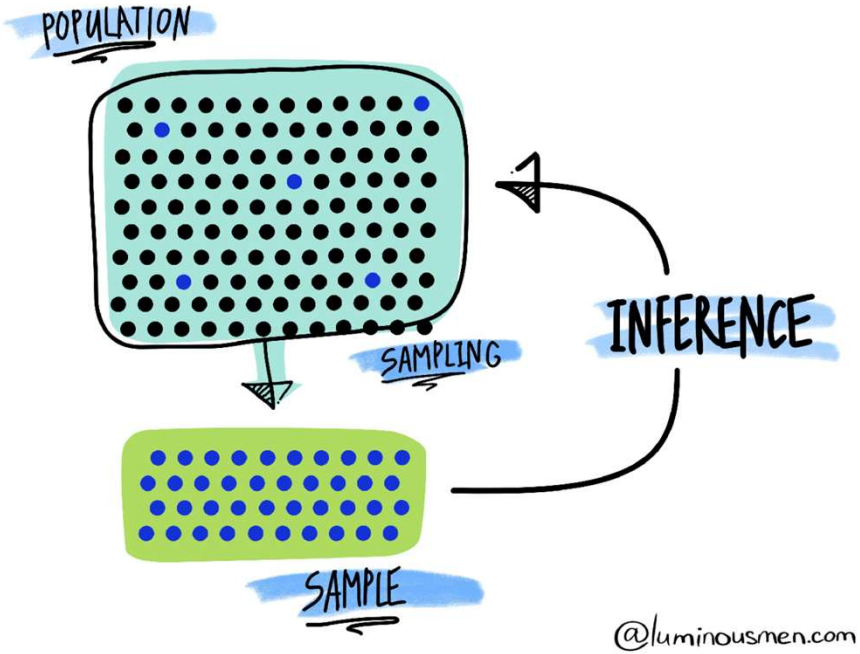
    ➔ no predice bien los valores o etiquetas de un conjunto de datos de validación o de la población.

Estadística Descriptiva e Inferencial

- **Estadística Descriptiva:** provee información acerca de los datos que ya hemos obtenido.

- **Inferencia Estadística:** nos permite obtener conclusiones acerca de la población de la cual proviene la muestra de datos.
  - Estimación de parámetros
  - Cálculo de intervalos de confiabilidad
  - Testeo de hipótesis.





Estadística Descriptiva

Análisis Exploratorio de Datos

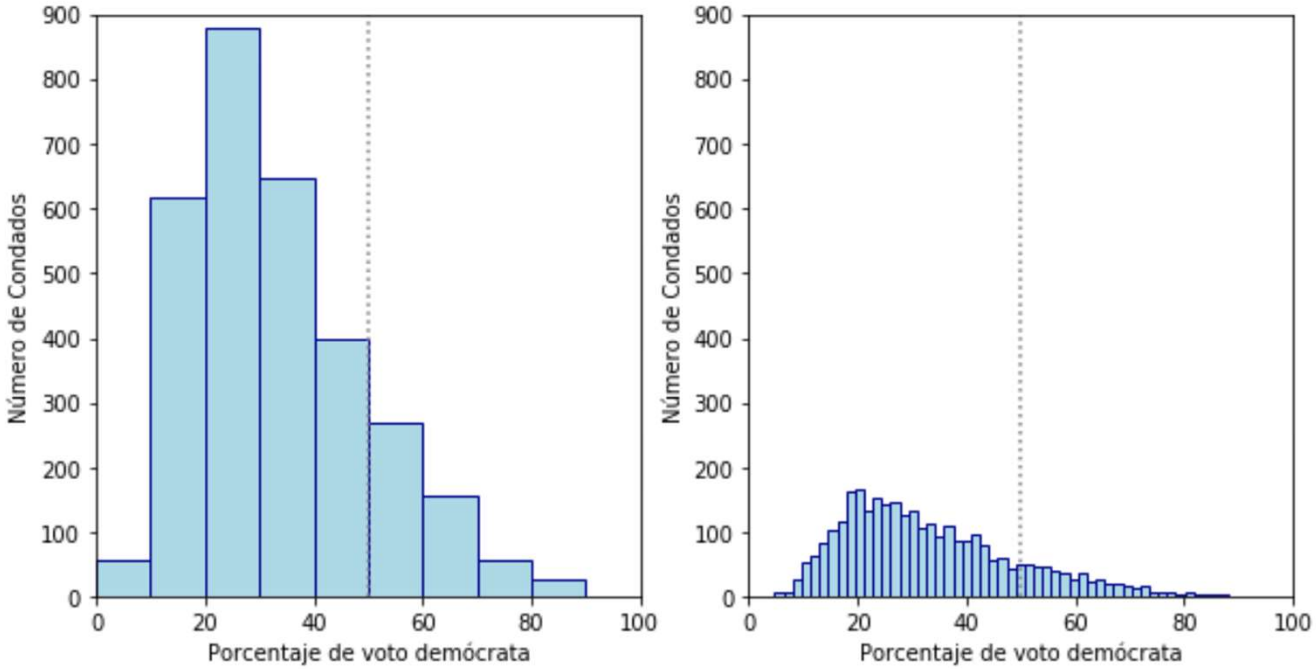
- Es lo contrario al análisis "confirmatorio" de datos: no hay hipótesis ni modelo.
- Exploratorio ➔ nuestra comprensión del problema cambia a medida que lo desarrollamos.

- Herramientas básicas:
  - Plots, gráficos, estadísticas de resumen

- Permite:
  - Ganar intuición acerca de los datos
  - Hacer comparaciones entre distribuciones
  - Chequeos de sanidad (escalas, formatos, etc,)
  - Identificar datos faltantes o outliers.
  - Resumir los datos.
  - Identificar posibles correlaciones

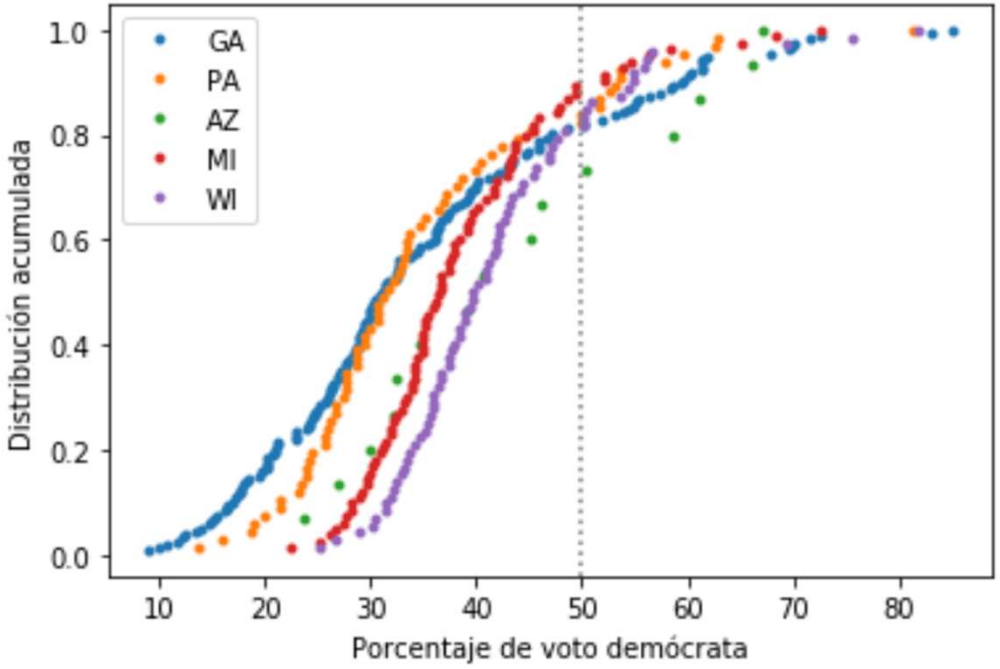
Análisis Exploratorio de Datos: Gráficos

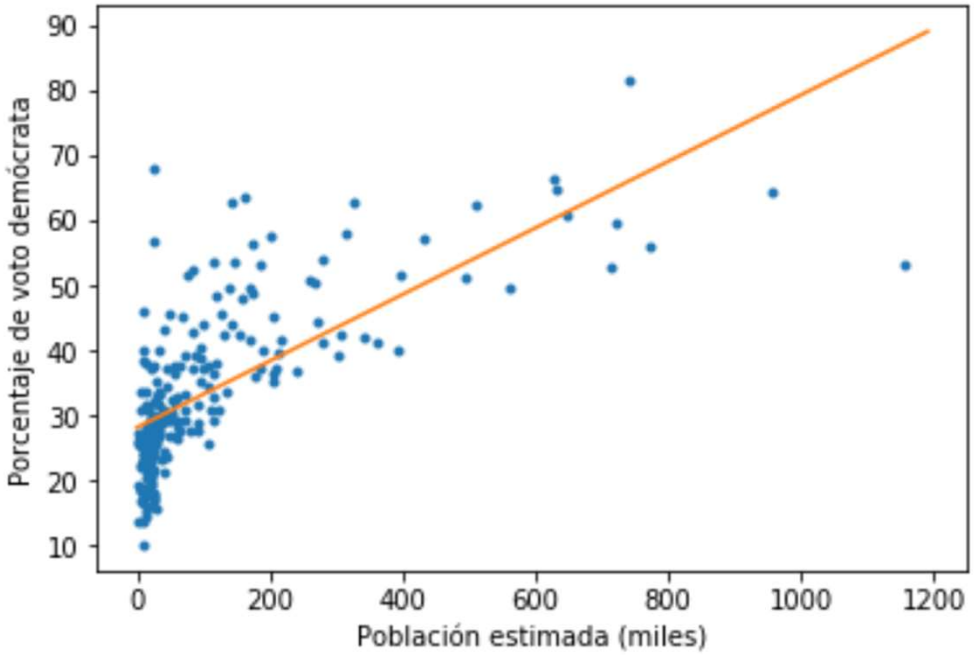
- Análisis gráfico:
  - Histograma
    - Binnig bias
  - Función de distribución acumulada empírica (ECDF)
  - Gráficos de dispersión
  - Diagramas de caja



Análisis Exploratorio de Datos: Gráficos

- Análisis gráfico:
  - Histograma
    - Binning bias
  - Función de distribución acumulada empírica (ECDF)
  - Gráficos de dispersión
  - Diagramas de caja





Análisis Exploratorio de Datos: Estadísticas de Resumen

- **Media:** es la suma de todos los valores, dividida por el número de puntos.

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

- **Mediana:** es el valor medio de un conjunto de datos. Es inmune a valores extremos o outliers. Para calcularla, se ordenan los datos y se elige el valor que queda en la mitad.
- **Percentiles:** el percentil p, corresponde al valor que es mayor al p% de los datos.

- **Varianza:** promedio de la distancia cuadrática de los datos a la media. Es una medida de la dispersión de los datos.

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

- **Desviación Estándar:** es la raíz cuadrada de la varianza. Está en la misma escala de unidades que los datos.

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

Análisis Exploratorio de Datos: Estadísticas de Resumen

- **Covarianza:** es una medida de cómo dos cantidades varían juntas. Es la media del producto entre las diferencias de los valores respecto a la media.

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

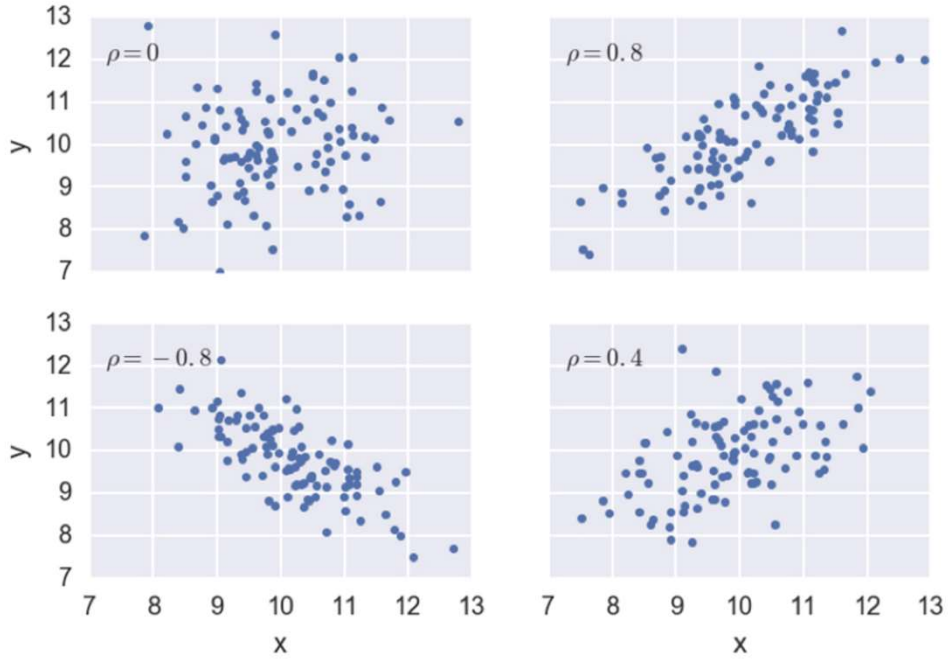
  - Si x e y tienden a estar ambas arriba, o ambas abajo de la media, la covarianza es positiva.
  - Esto quiere decir que hay una correlación positiva: cuando x es alta, y es alta.
  - Por el contrario, si x es alta cuando y es baja, la covarianza es negativa y los datos están anticorrelacionados.

- **Coeficiente de Pearson** (ρ)**:** para tener una medida más general y aplicable de la correlación entre dos variables, necesitamos que sea adimensional. Por lo tanto dividimos la covarianza por las desviaciones estándar de x e y.

$$\rho = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

  - Es la comparación de la variabilidad en los datos debido a una codependencia (covarianza), con la variabilidad inherente de cada variable (sus desviaciones estándar).
  - Un valor 0 indica que no hay correlación, valor 1 indica alta correlación.

Análisis Exploratorio de Datos: Estadísticas de Resumen



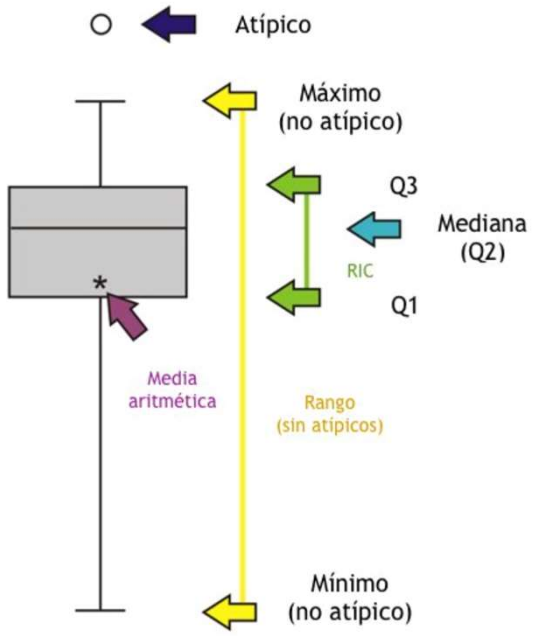
- **Coeficiente de Pearson** (ρ): para tener una medida más general y aplicable de la correlación entre dos variables, necesitamos que sea adimensional. Por lo tanto dividimos la covarianza por las desviaciones estándar de x e y.

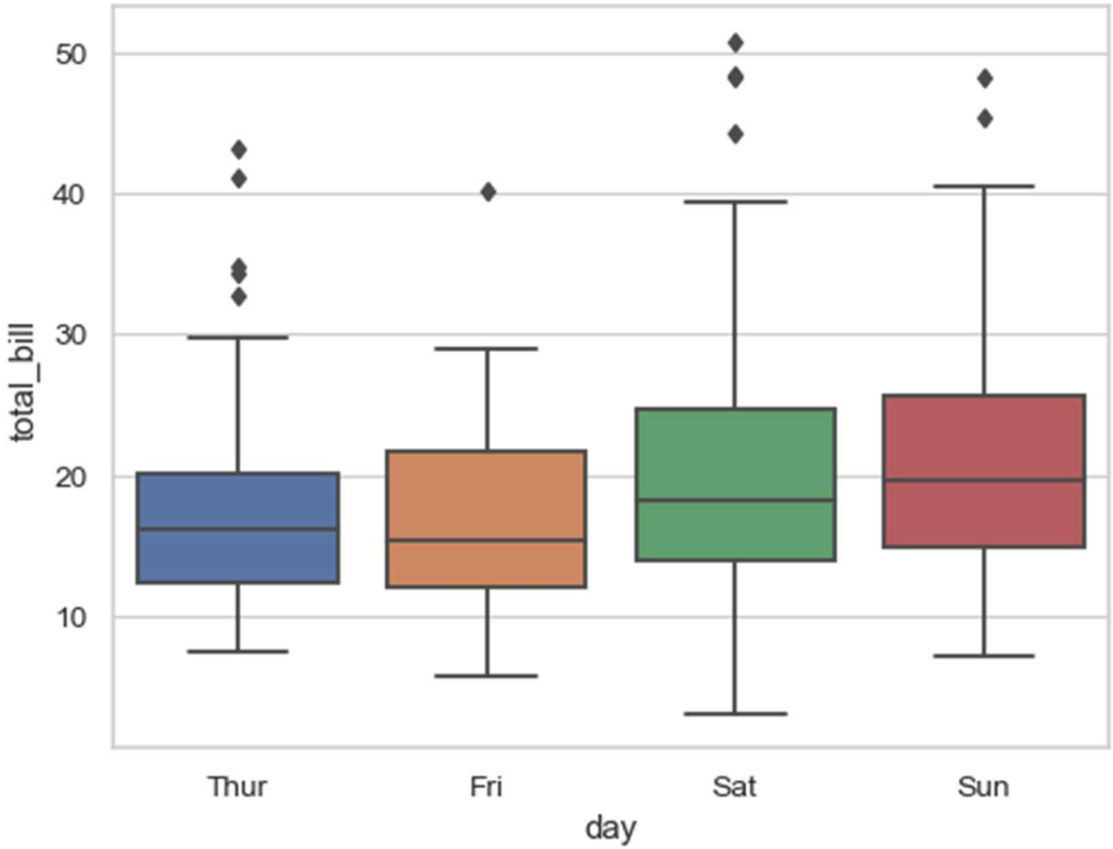
$$\rho = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

  - Es la comparación de la variabilidad en los datos debido a una codependencia (covarianza), con la variabilidad inherente de cada variable (sus desviaciones estándar).
  - Un valor 0 indica que no hay correlación, valor 1 indica alta correlación.

Análisis Exploratorio de Datos: Gráficos + Estadística

- **Diagramas de caja:** Los percentiles pueden utilizarse para representar los datos utilizando "gráficos de caja y bigotes" (boxplots), que resumen los valores de la mediana, cuartiles (percentiles 25 y 75) y valores extremos o outliers.





Inferencia Estadística

Inferencia Estadística y Bootstrapping

- Si tenemos un conjunto de mediciones, ¿cuál será el valor promedio del siguiente conjunto de mediciones?

  - No podemos predecir el valor exacto, pero sí describir lo que esperamos en términos probabilísticos.
  - Dado un conjunto de datos, la inferencia estadística nos permite describir probabilísticamente que esperamos para muestras de datos adquiridos una y otra vez, incluyendo una medida de incerteza.

- ¿Cómo repetimos una y otra vez la adquisición de datos?
  - Podemos simularlos computacionalmente, generando "réplicas" o "permutaciones" de nuestras observaciones.
  - Resampleamos los datos una y otra vez, calculamos estadísticas de resumen, parámetros óptimos y sus respectivas distribuciones de probabilidad.

- **Bootstrapping** → uso de datos remuestreados para realizar inferencia estadística sobre una población.

Bootstrapping

- Uso de datos resampleados para hacer inferencia estadística. Permite descripción probabilística de los datos.

  - **Muestras** (bootstrap sample): con reemplazo
    - Simple o de a pares
    - En Python: `np.random.choice()`
  - **Permutaciones** (bootstrap permutation)
  - **Réplicas** (bootstrap replicate): estadística calculada a partir de una muestra o permutación.
    - Ej.: media, PDA, parámetros, coeficiente de correlación, etc.

Datos:

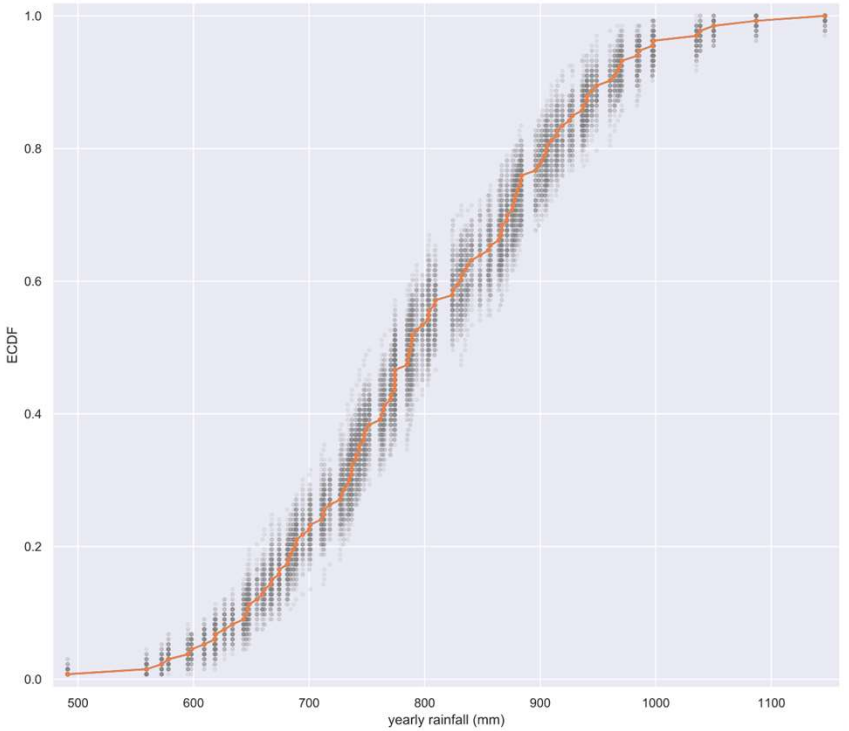
```
[23.3, 27.1, 24.3, 25.7, 26.0]
```

Mean = 25.2

Remuestreo:

```
[27.1, 26.0, 23.3, 25.7, 23.3]
```

Mean = 25.08



Bootstrap e Intervalos de Confiabilidad.

- Ajuste de un modelo paramétrico
  - Parámetros óptimos cambian si cambiamos las muestras.
  - Podemos calcular intervalos de confiabilidad para cada uno.

Ej: Bootstrap de a pares para regresión lineal.

- Las muestras de datos tienen dos variables: (x,y)
- Resampleamos pares (x,y)
- Calculamos pendiente e intercepto para muestras.
- Calculamos intervalos de confiabilidad para cada parámetro óptimo.

Testeo de Hipótesis

- ¿Cómo evaluamos si es un modelo razonable?
  - Evaluamos qué tan razonables son los datos observados, asumiendo que una cierta hipótesis es verdadera.
  - **Hipótesis nula (H_o):** la hipótesis que estamos probando

- **Pasos:**
  1. Plantear la hipótesis nula.
  2. Definir la estadística de prueba.
  3. Generar muchos conjuntos de datos simulados asumiendo que la hipótesis nula es verdadera.
  4. Calcular estadísticas de prueba para cada set de datos simulados.
  5. Evaluar si la estadística observada, está dentro de lo esperable bajo la hipótesis nula.
     - Valor-p: fracción de la data simulada para la cual la estadística de prueba es tan extrema como el valor observado.

Testeo de Hipótesis

1. Plantear la hipótesis nula.

Testeo de Hipótesis: Estadísticas de Prueba

2. Definir una estadística de prueba.

Buscamos evaluar qué tan razonable es la data observada asumiendo que una hipótesis es verdadera:

- ¿Qué evaluamos, cómo cuantificamos la evaluación?

❖ Estadística de prueba:

- Un número que puede ser calculado de data observada, y a partir de la data que simulamos bajo la hipótesis nula.
- Sirve como comparación entre lo que la hipótesis predice, y lo que observamos.

- Tiene que ser algo pertinente a lo que tratamos de responder con el test de hipótesis.
  - Ej: si dos distribuciones son iguales, su media debe ser la misma.

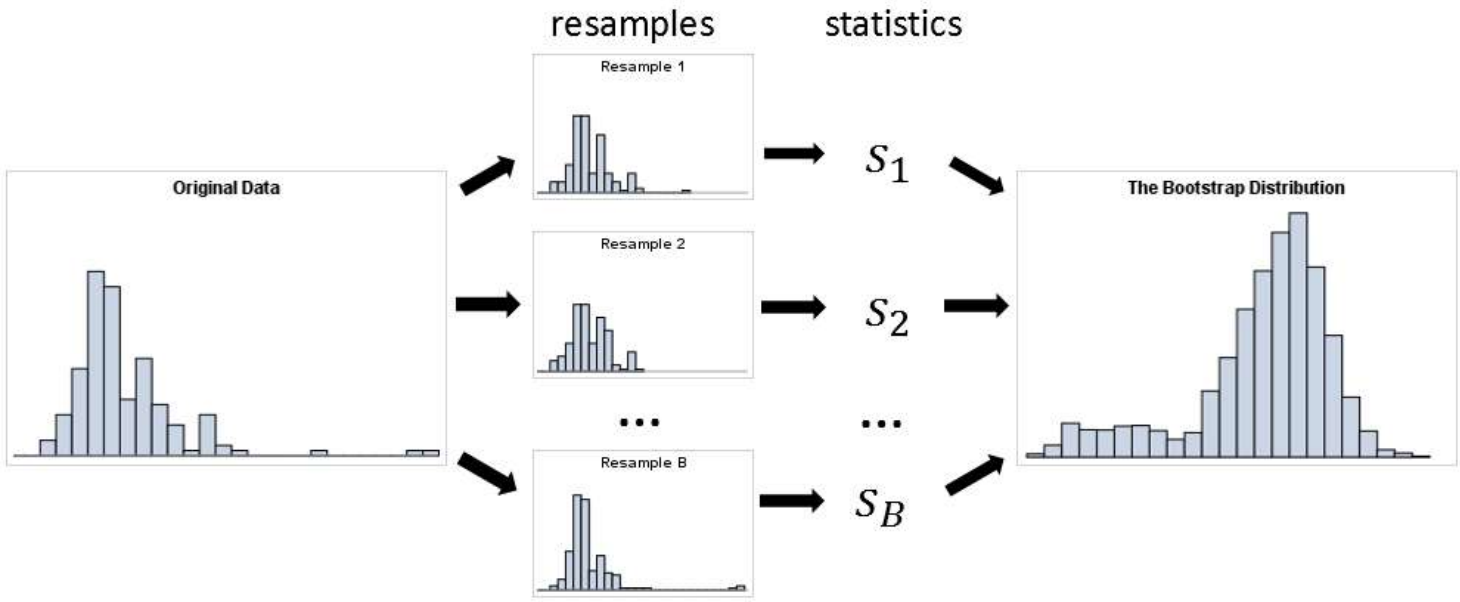
Testeo de Hipótesis: Simulaciones

3. Generar muchos conjuntos de datos simulados asumiendo que la hipótesis nula es verdadera.

- Resampleo: seleccionar muestras (con o sin reemplazo) a partir del conjunto original de datos.
  - Testeo de parámetros de una distribución

- Permutaciones: concatenar sets de datos y reordenar.
  - Ej: Comparación de dos distribuciones, ¿son iguales o diferentes?

4. Calc **nulados.**

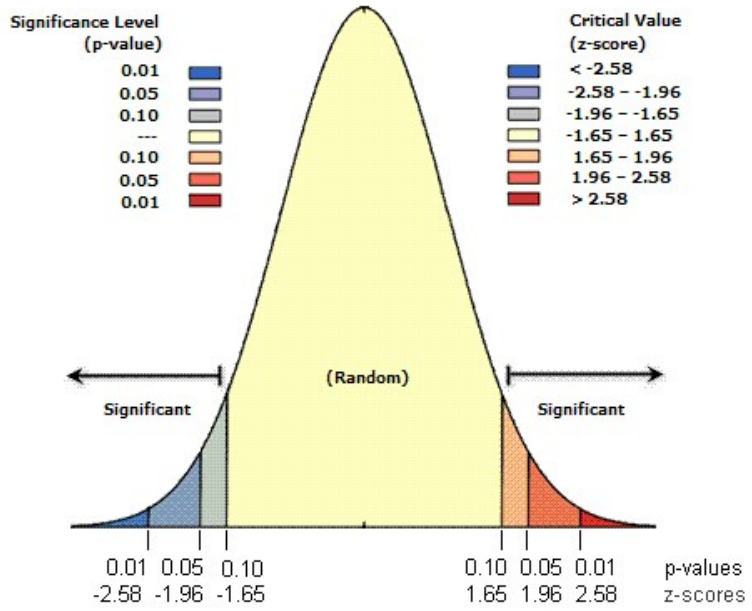


Testeo de Hipótesis: Evaluación

5. Evaluar si la estadística observada, está dentro de lo esperable bajo la hipótesis nula.

Valor p:

- Probabilidad de obtener un valor de la estadística de prueba que sea igual o más extrema que la observada, bajo la hipótesis nula.
- *No es la probabilidad de que la hipótesis sea verdadera*
- Valor-p pequeño ➔ data es estadísticamente significativamente diferente que lo que se observaría bajo la hipótesis nula.



Caso de Estudio: Elecciones Presidenciales EEUU 2020