# Fundamentos para al Análisis de Datos y la Investigación

## Tema 4 - Muestreo y distribuciones de datos

**José María Sarabia**

Máster Universitario en Ciencia de Datos
CUNEF Universidad

# Contents

# 1 Introducción

- En este tema estudiaremos aspectos relevantes en la ciencia de datos
- Estos aspectos incluyen el muestreo y los modelos de distribuciones
- La distribución normal es el modelo principal en ciencia de datos
- Estudiaremos el método Naive Bayes de clasificación

# Contents

## 2 Muestreo aleatorio y sesgo de la muestra

- En la era de los *macrodatos o big data*, un error muy extendido es que esto supone el fin de la necesidad de hacer muestreos.
- Sin embargo, la proliferación de datos de distinta calidad y relevancia, hace más importante al muestreo como herramienta para trabajar con datos de manera más eficiente y minimizar el sesgo.
- Existen dos aspectos básicos con los que se trabaja: *población y muestra*.
- En este sentido es importante distinguir entre *paramétros de la muestra y de la población*. Por ejemplo, la media muestral $\bar{X}$ y la media poblacional $\mu$.

## 2 Muestreo aleatorio y sesgo de la muestra

- Una muestra (*sample*) es un subconjunto de datos elegido de un conjunto más grande.
- El muestreo aleatorio (*random sampling*): cada elemento de la población tiene la misma probabilidad de ser elegido.
- La calidad de los datos es más importante que la cantidad. La calidad de los datos supone: integridad, coherencia del formato, limpieza y precisión de los datos individuales.
- La estadística incorpora el término de *representatividad*.
- En R se seleccionan las muestras mediante la función `sample`. Sintaxis:

```
sample(x, size, replace = FALSE, prob = NULL)
```

## 2 Muestreo aleatorio y sesgo de la muestra

- Sesgo: es el error sistemático. El sesgo se da cuando las mediciones u observaciones son erróneas de forma sistemática porque no son representativas de toda la población.
- Sesgo de elección: ocurre a partir de la forma en que se seleccionan las observaciones.
- Selección aleatoria. Todas las observaciones tienen la misma probabilidad de ser elegidas.
- Tamaño frente a calidad. La calidad de los datos suele ser más importante que los mismos y el muestreo aleatorio puede reducir el sesgo y dar lugar a cálculos más precisos.

# Contents

# 3 Distribución Normal

Cuando hay una distribución que no es normal, hay que transformarla a la normal, una manera es el log

- La *distribución normal, Gaussiana o campana de Gauss* es la distribución más importante en estadística. Un gran número de fenómenos naturales, sociales y económicos se ajustan a una distribución normal.
- La justificación es por medio del teorema central del límite, de modo que si un fenómeno aleatorio es suma de una serie de causas independientes, se puede modelizar por medio de una distribución normal.
- La función de densidad, media y varianza son:

$$f(x) = \frac{\exp(-(x-\mu)^2/2\sigma^2)}{\sigma\sqrt{2\pi}}; \; E(X) = \mu; \; var(X) = \sigma^2$$

- En el intervalo $(\mu - 2\sigma, \mu + 2\sigma)$ se encuentran el 95 % de los datos, y en el intervalo $(\mu - 3\sigma, \mu + 3\sigma)$ más del 99 %

# 3 Distribución Normal

HERRAMIENTAS( GRAFICA[QQ], CONTRASTE [SATIRP-WICK])
Y=log(x)
y=r2x

- Ejemplos:
  - Distribuciones de carácter biométrico como la altura, peso de individuos adultos, etc.
  - Características psicológicas como puntuaciones en test, ci, etc.
  - Distribución de gastos
  - Errores de medición etc.
- La distribución normal estándard con $\mu = 0$ y $\sigma = 1$ se denota por $Z$. Una normal cualquiera $X \sim N(\mu, \sigma)$ se puede convertir en normal estándard mediante la transformación $Z = \frac{X-\mu}{\sigma}$
- Comandos en R:
  - `dnorm(x,media,dt)` densidad
  - `pnorm(x,media,dt` probabilidad
  - `qnorm(prob,media,dt)` cuantil
  - `rnorm(nsample,media,dt)` simular, obtener datos al azar

p(Altura<=183)=pnorm(183,m=175,s=5)

calcula la probabilidad de que la altura sea menor de 183cm.

# 3 Distribución Normal

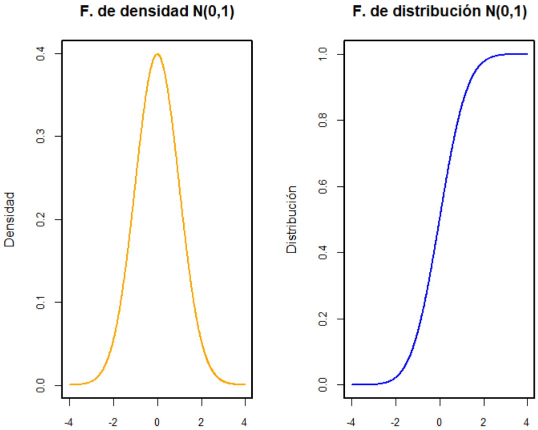


Figura: Funciones de densidad y de distribución de la N(0,1)

# 3 Distribución Normal: diagramas QQ y prueba SW

- El diagrama QQ sirve para identificar si unos datos siguen una distribución normal de forma gráfica
- La prueba de Shapiro-Wilk sirve para determinar si unos datos siguen una distribución normal de forma exacta (usando el $p$-value)
- Muchos conjuntos de datos mediante transformaciones sencillas ($y = \log(x)$, $y = x^2$, etc) se pueden transformar en una distribución normal

# 3 Distribuciones relacionadas

- Algunos modelos de distribuciones pueden parecerse a la normal, pero diferenciarse en las *colas*. Un ejemplo son las distribuciones de colas pesadas (datos de rendimientos, etc.)
- Las distribuciones $t$ se Student, chi-cuadrado de Pearson y $F$ de Snedecor se obtiene a partir de distribuciones normales.
- **Tips**: distribuciones son curtosis mayores que 3, asimétricas, o datos correspondientes a valores máximos o mínimos, no pueden seguir el modelo normal

# Contents

# 4 Distribuciones muestrales

- El término distribución muestral (*sampling distribution*) de un estadístico se refiere a la distribución del estadístico de la muestra a partir de muchas muestras extraidas de la misma población.
- Cuando se extrae una muestra el objetivo es medir algo (con el estadístico muestral) o bien modelar algo (con un modelo estadístico o de aprendizaje automático). Dado que la estimación se basa en una muestra, podría dar lugar a errores. Estamos interesados en cuanto de diferente puede ser. Para ello, obtenemos muchas muestras y obtenemos su distribución.
- Por tanto, es importante distinguir entre distribución de los datos (data distribution) y distribución de un estadístico muestral (sampling distribution)

# 4 Distribuciones muestrales

- Las dos formas de obtener distribuciones muestrales es mediante:
    - El Teorema central del Límite
    - El método Bootstrap

# 4 Distribuciones muestrales: TCL

- El Teorema central del límite asegura que las medias extraidas de varias muestras se asemejan a la distribución normal.
- Es decir:

$$\frac{X_1 + \cdots + X_n}{n} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

  donde $\mu$ y $\sigma$ se refieren a la media y dt de la población. El procedimiento se implementa mediante *simulación* Monte Carlo.
- El algoritmo consta de los siguientes pasos (caso de la media):
  1. Extraemos varias muestras (de tamaño n) de la población.
  2. Para cada muestra calculamos la media muestral
  3. Se repiten los pasos 1-2 $m$ veces
  4. Se usan las $m$ muestras para calcular la media, la desviación estándard y obtener un histograma

# 4 Distribuciones muestrales: Bootstrap

- Una manera sencilla y efectiva de estimar la distribución muestral de un estadístico es mediante el *método bootstrap*
- La idea es extraer muestras adicionales con reemplazamiento de la misma muestra, y volver a calcular el estadístico o modelo en cada muestra repetida
- Este método no requiere que los datos sean normales, y puede aplicarse a cualquier distribución muestral (media, mediana etc) o modelo estadístico.

# 4 Distribuciones muestrales: Bootstrap

- El algoritmo consta de los siguientes pasos (en el caso de la media):
  1. Extraemos un valor de la muestra, lo registramos y devolvemos (con reemplazamiento)
  2. Lo repetimos *n* veces
  3. Calculamos la media de los *n* valores muestrados
  4. Se repiten los pasos 1-3 *R* veces
  5. Se usan los resultados de *R* para: calcular la desviación estándard, obtener un histogrma o similar y un intervalo de confianza
- El bootstrap se puede usar para estimar la estabilidad (variabilidad) de los parámetros de un modelo o para mejorar la capacidad de pronóstico.
- Con los árboles de clasificación y regresión ejecutar varios árboles en muestras bootstrap y a continuación promediar sus pronósticos generalmente funciona mejor que usar un sólo árbol. Este método se denomina *bagging* (*bootstrap aggregating*)
- El paquete `boot` permite el uso del bootstrap de modo sencillo

# Contents

# 5 Anexo: Probabilidad: axiomas y consecuencias

- Cuando un experimento se repite muchas veces, la probabilidad de ese suceso corresponde a la frecuencia relativa:

$$P(A) = \frac{\text{número de veces que ocurre A}}{\text{número total de experimentos}}$$

- La probabilidad proporciona una medida del grado de ocurrencia de un suceso.
- *Axiomas de la probabilidad*: Si $\Omega$ representa el espacio muestral correspondiente a un experimento aleatorio, la probabilidad de un suceso $A$ verifica:

$$\begin{aligned} 0 \leq P(A) &\leq 1 \\ P(\Omega) &= 1 \\ P(A_1 \cup \cdots \cup A_n) &= P(A_1) + \cdots + P(A_n) \end{aligned}$$

# 5 Anexo: Probabilidad: axiomas y consecuenciasd

Otras reglas importantes de la probabilidad son:

- La probabilidad del complementario de un suceso: $P(A^c) = 1 - P(A)$
- La probabilidad del suceso imposible es cero: $P(\emptyset) = 0$
- Si $A \subset B$ entonces: $P(A) \leq P(B)$
- La probabilidad de la unión de dos sucesos no necesariamente disjuntos viene dada por,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

# 5 Anexo: Probabilidad condicionada y sucesos independientes

- La *probabilidad condicionada* de un suceso $A$ dado que ha ocurrido un suceso $B$ es:

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

- Ejemplo: Se lanza un dado, probabilidad que salga un 6, sabiendo que se ha obtenido un número par.
- Ejemplo: Probabilidad de coger el paraguas, dado que está lloviendo.

# 5 Anexo: Probabilidad condicionada y sucesos independientes

- Dos sucesos $A$ y $B$ son *independientes* si la ocurrencia de uno de ellos no influye en la ocurrencia del otro
- En términos de probabilidades, dos sucesos son independientes si se verifica una de las siguientes condiciones:

$$\begin{aligned} P(A|B) &= P(A), \\ P(B|A) &= P(B) \\ P(A \cap B) &= P(A)P(B) \end{aligned}$$

- Si $A$ y $B$ son independientes, también son independientes las parejas: $\{A, B^c\}$, $\{A^c, B\}$ y $\{A^c, B^c\}$.

## 5 Anexo: Teorema de Bayes

- El Teorema de Bayes es un resultado básico en probabilidad
- Es posible una nueva interpretación que da lugar a la llamada *Estadística Bayesiana*
- Una de las aplicaciones más relevantes es en clasificación.

# 5 Anexo: Teorema de Bayes

- Como paso previo tenemos el *teorema de las probabilidades totales*. Si el espacio muestral $\Omega$ se puede dividir en una partición $B_i$, de modo que $\Omega = \cup_{i=1}^{n} B_i$ y $B_i \cap B_j = \emptyset$, con $i \neq j$, y si $A$ es un suceso cualquiera se cumple:

$$
\begin{aligned}
P(A) &= P(A \cap B_1) + \cdots + P(A \cap B_n) \\
&= P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n) \\
&= \sum_{j=1}^{n} P(A|B_j)P(B_j)
\end{aligned}
$$

- Con las mismas hipótesis del teorema de las probabilidades totales, el *teorema de Bayes* establece que si $A$ es nuevamente un suceso cualquiera se verifica

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}, \quad i = 1, 2, \ldots, n$$

# 5 Anexo: Teorema de Bayes

Interpretación:

- *Probabilidades a priori*:

$$P(B_1), P(B_2), \ldots, P(B_n)$$

  Son las probabilidades de pertenencia a los sucesos de la partición sin ningún tipo de evidencia previa

- *Evidencia o Verosimilitud*: Recibimos una información sobre un suceso ajeno $A$, y disponemos de

$$P(A|B_1), P(A|B_2), \ldots, P(A|B_n)$$

- *Probabilidades a posteriori*: La evidencia anterior nos lleva a corregir las probabilidades iniciales por el Teorema de Bayes:

$$P(B_1|A), P(B_2|A), \ldots, P(B_n|A)$$

- La principal aplicaciçon del Teorema de Bayes en Ciencias de datos es el *método de clasificación Naivë Bayes*.

# 5 Anexo: Otros Conceptos

- Si en un experimento aleatorio $A$ es un suceso tenemos:

  (a) La probabilidad:

  $$p = P(A)$$

  (b) El Odds ratio de $A$:

  $$\text{Odds}(A) = \frac{p}{1-p}$$

  (c) El logit de $A$:

  $$\text{logit}(A) = \log\left(\frac{p}{1-p}\right)$$

# 5 Anexo: Probabilidades por simulación

- Se pueden calcular probabilidades de sucesos mediante simulación haciendo uso de dos comandos:
  - El comando `sample(1:6,10,replace=TRUE)` permite simular 10 lanzamientos de un dado.
  - El comando `runif(10,0,1)` genera diez muestras en el intervalo $(0, 1)$. Si escribimos `runif(1)<p` obtenemos un valor lógico (`TRUE` o `FALSE`), que nos permite aproximar probabilidades.

# Fundamentos para al Análisis de Datos y la Investigación

# Tema 4 - Muestreo y distribuciones de datos

**José María Sarabia**

Máster Universitario en Ciencia de Datos
CUNEF Universidad