# Master en Big Data. Fundamentos matemáticos del análisis de datos.

## Sesión 3: Probabilidad

Fernando San Segundo

Curso 2021-22. Última actualización: 2021-08-25

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI
DEPARTAMENTO DE MATEMÁTICA APLICADA

# RECUERDA: ANTES DE SEGUIR

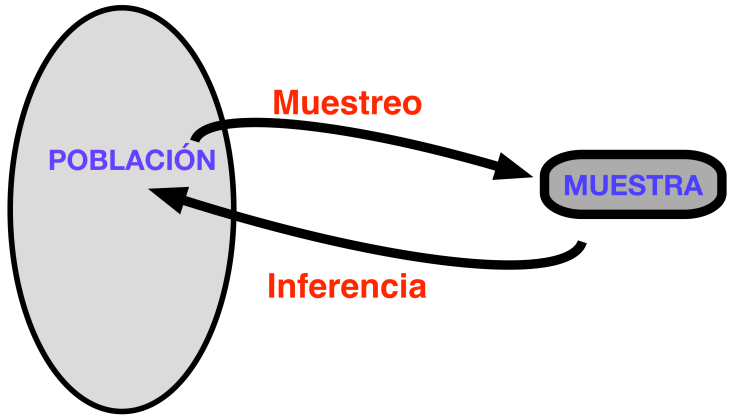
# EJECUTA GIT PULL EN EL REPOSITORIO FMAD2122

Sección 1

# Población y muestra.

# Inferencia Estadística.

- El objetivo central de la Estadística es obtener información fiable sobre las características de una **población** a partir de **muestras**. Ese término significa aquí un conjunto de entidades individuales (individuos), no necesariamente seres vivos. La población pueden ser los vehículos matriculados en 2015 o las órdenes de compra recibidas por una empresa cierto mes o las especies de colibrí que visitan un comedero en Costa Rica en los últimos 10 años, etc.
- Muchas veces estudiar toda la población es demasiado difícil, indeseable o imposible. Entonces surge la pregunta de si podemos usar las muestras para *inferir*, o *predecir* las características de la población. ¿Hasta qué punto los datos de la muestra son *representativos*de la población?
- La *Inferencia Estadística* es el núcleo de la Estadística porque da sentido a estas preguntas, las formaliza y responde.

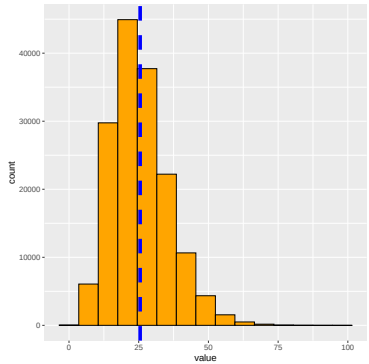


# Poblaciones y muestras aleatorias simples con vectores usando R.

- Al estudiar una población nos interesan determinadas características individuales, que pueden cambiar de un individuo a otro y que constituyen las *variables de interés*. Cuando tomamos una muestra obtenemos los valores de esas variables en algunos individuos de la población.
- Para que la muestra sea representativa lo mejor es que sea una **muestra aleatoria simple**: elegimos a los individuos al azar y con remplazamiento (podemos incluir al mismo individuo más de una vez en la muestra). Tomemos esta población.

```
set.seed(2019)
N = 158000
poblacion = as.integer(2 * rchisq(N, df = 13), 0)
```

- Para entenderlo mejor haremos un experimento con R. En este caso vamos a suponer una población de $N = 1.58 \times 10^5$ individuos. Por ejemplo, los viajeros que pasan por un aeropuerto en un día y sea la variable de interés su edad. El código de esta sesión construye un vector `poblacion` con las edades de los viajeros. Vamos a hacer una pequeña trampa y mostraremos el histograma de las edades de la población, que en un caso real desconoceríamos. La línea de puntos indica *la media poblacional de la edad.* ¿Cuál crees que es?



# Medias muestrales

- Ese es justo el tipo de preguntas que esperamos que responda la Estadística. Aunque en este caso disponemos del vector completo de edades debes tener claro que en los problemas reales no será así. Así que recurrimos a las muestras aleatorias (con remplazamiento), en inglés *random sample (with replacement)*. Por ejemplo, de tamaño 20. En R construimos una de esas muestras así:

```
n = 20
(muestra = sample(poblacion, n, replace = TRUE))
```

```
 [1] 20 10 18 39 36 29 55 25 30 40 18 44 12 30 18 15 12 22 10 19
```

Esas son las 20 edades $x_1, \ldots, x_{20}$ de los viajeros de la muestra. Para *estimar* la edad media de *todos los viajeros* a partir de estos valores calcularíamos la **media muestral**.

$$\bar{x} = \frac{x_1 + \cdots + x_{20}}{n} = \frac{20 + 10 + \cdots + 19}{20} \approx 25.1 = \text{mean(muestra) en R}$$

- Naturalmente, si tomas otra muestra, su media muestral puede ser otra:

```
(muestra2 = sample(poblacion, n, replace = TRUE))
```

```
 [1] 16 28 38 18 28 46 18 32 27 16 15 23 18 30 48 23 30 14 23 31
```

```
mean(muestra2)
```

```
[1] 26.1
```

# Muestras buenas y malas.

- Hemos visto que cada muestra produce una media muestral y que esas medias muestrales pueden ser distintas. ¿Cuántas muestras distintas hay? Hay una cantidad inimaginablemente grande:

$$158000^{20} = 9.4003005 \times 10^{103}$$

  Para ponerlo en perspectiva, se estima que en el universo hay menos de $10^{40}$ estrellas. Esta cantidad enorme de muestras, de las que solo hemos visto 2, forman lo que se llama el **espacio muestral** (de tamaño $n = 20$) de este problema.
- Entre esas muestras hay muestras *buenas* y muestras *malas*. ¿Qué queremos decir con esto? Para seguir con nuestro experimento vamos a ordenar *la población completa* por edad y tomemos los 20 primeros valores:

```
(muestra3 = sort(poblacion)[1:20])
```

```
 [1] 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3
```

  Hemos llamado `muestra3` a ese vector porque es una más de las muchísimas muestras posibles que podríamos haber obtenido al elegir al azar 20 viajeros. Y si usáramos esta muestra para estimar la media de la población obtendríamos

```
mean(muestra3)
```

```
[1] 2.5
```

  Eso es lo que llamamos una *muestra mala*, poco representativa.

# La distribución de las medias muestrales.

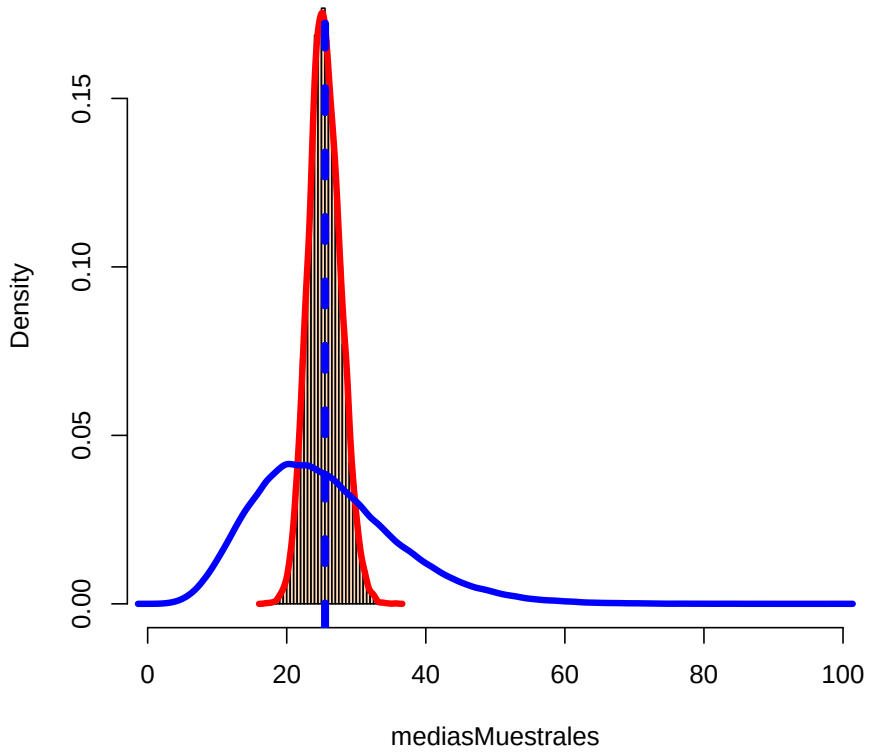
- La última muestra que hemos examinado era muy poco representativa. Pero la pregunta esencial para la estadística es ¿cuál es la relación entre muestras buenas y malas? Al elegir una muestra al azar, ¿cómo de probable es que nos toque una muestra tan mala en lugar de una buena?
- Podemos hacer otro pequeño experimento para explorar el espacio muestral. No podemos repasar todas las muestras una por una para clasificarlas en buenas o malas (eso sería demasiado incluso para R) pero podemos tomar *muchas* muestras aleatorias (pongamos $k = 10000$) y ver como de buenas o malas son (hacemos una *muestra de muestras*). En R es muy fácil hacer esto usando la función `replicate`:

```
k = 10000
# replicate repite k veces los comandos entre llaves y guarda el resultado
# del último comando en el vector mediasMuestrales
mediasMuestrales = replicate(k, {
  muestra = sample(poblacion, n, replace = TRUE)
  mean(muestra)
})
head(mediasMuestrales, 10)
```

```
 [1] 25.00 28.70 24.85 26.05 25.75 27.15 28.05 25.15 28.40 28.40
```

Se muestran las primeras 10 de las 10000 medias muestrales que hemos obtenido.

- En lugar de examinar una a una esas 10000 medias muestrales vamos a representarlas en un histograma y una curva de densidad. Además, aprovechándonos de que en este caso tenemos acceso a la población completa hemos añadido su curva de densidad:




- Este es posiblemente **el gráfico más importante del curso**. Fíjate en tres cosas:
  - La *media de las medias muestrales* coincide con la media de la población.
  - Prácticamente no hay *muestras malas*. Es *extremadamente improbable* que una muestra elegida al azar sea muy mala.
  - La distribución de las medias muestrales tiene forma de campana (y es muy estrecha).

  Para entender bien estas ideas *necesitaremos aprender más sobre Probabilidad*.

# Otra población, mismos resultados.

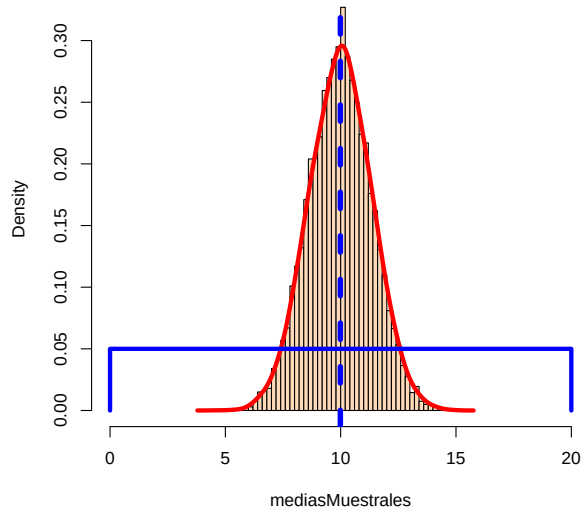
- Pero antes de lanzarnos a la probabilidad vamos a asegurarnos de algo. Puede que te preguntes si la población con la que hemos empezado tenía algo especial. Probemos con otra muy distinta. La población la forman 20000 números elegidos al azar del 0 al 20, siendo todos los valores igual de probables (su curva de densidad es horizontal).

```
poblacion = sample(0:20, 20000, replace = TRUE)
```

Y ahora repetimos el proceso de construcción de medias muestrales usando replicate

```
k = 10000
mediasMuestrales = replicate(k, {
  muestra = sample(poblacion, n, replace = TRUE)
  mean(muestra)
})
```

El gráfico del resultado muestra el mismo comportamiento de las medias muestrales, lo que se conoce como **Teorema Central del Límite**:



Sección 2

# Probabilidad básica.

- Para entender resultados como el Teorema Central del Límite tenemos que aprender el mínimo vocabulario necesario para poder hablar con precisión sobre la Probabilidad.
- Lo primero de lo que hay que ser conscientes es de que nuestra intuición en materia de probabilidad suele ser muy pobre. Vamos a empezar usando ejemplos de juegos de azar (dados, naipes, etc.) para poder desarrollar el lenguaje, igual que sucedió históricamente.

# Experimentos del Caballero de Méré.

- ¿Qué es más probable?
  ($a$) obtener al menos un seis en cuatro tiradas de un dado, o
  ($b$) obtener al menos un seis doble en 24 tiradas de dos dados?
- Los jugadores que en el siglo XVIII se planteaban esta pregunta pensaban así:
  ($a$) La probabilidad de obtener un seis en cada tirada es $\frac{1}{6}$. Por lo tanto, en cuatro tiradas es $\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{2}{3}$.
  ($b$) La probabilidad de un doble seis en cada tirada de dos dados es $\frac{1}{36}$, (hay 36 resultados distintos) y todos aparecen con la misma frecuencia. Por lo tanto, en veinticuatro tiradas será $\frac{24}{36}=\frac{2}{3}$.
  Así que en principio ambas apuestas parecen iguales,
- Vamos a usar R para jugar a estos dos juegos sin tener que jugarnos el dinero. Descarga este fichero de código y ejecútalo.

# La paradoja del cumpleaños.

- Otro experimento que puede servir para afianzar la idea de que la probabilidad es poco intuitiva. Si en una sala hay 1000 personas entonces es seguro que hay dos que cumplen años el mismo día. De hecho basta con que haya 367 personas. Si hay menos de ese número, la probabilidad de que dos cumpleaños coincidan disminuye. ¿Cuál es el *menor número de personas* que nos garantiza una probabilidad mayor del 50 % de coincidencia?
- Usemos R para averiguar ese número. Repite el experimento varias veces para convencerte..

```
## La paradoja del cumpleaños.
n = 366 # Número de personas en la sala

# Vamos a repetir el experimento N veces (N salas de n personas)
N = 10000
pruebas = replicate(N, {
fechas = sort(sample(1:366, n, replace=TRUE))
max(table(fechas)) # si el máximo es mayor que 1 es que 2 fechas coinciden
})
mean(pruebas > 1) # ¿qué proporción de salas tienen coincidencias?
```

```
[1] 1
```

# Regla de Laplace.

- Fue históricamente el primer resultado que hizo posible calcular probabilidades de una manera sistemática, aunque como veremos no está libre de problemas.
- Vamos a fijar el lenguaje necesario para entender esa regla.
  ($a$) Estudiamos un experimento aleatorio con *n resultados elementales* posibles (no simultáneos) que además son *equiprobables*; es decir, sus frecuencias relativas son iguales cuando el experimento se repite muchas veces.:

$$\{a_1, a_2, \ldots, a_n, \}$$

  ($b$) El *suceso aleatorio* $A$ es un {subconjunto del conjunto de resultados elementales}. Por ejemplo, si lanzamos un dado, $A$ puede ser: obtener un número par.
  ($c$) Los resultados elementales que forman $A$ son los {resultados favorables} a $A$. Por ejemplo, si lanzamos un dado, los resultados favorables al suceso

$$A = \text{obtener un número par}$$

  son $\{2, 4, 6\}$.
- **Regla de Laplace:** En esas condiciones la probabilidad de $A$ es

$$P(A) = \frac{\text{número de sucesos elementales favorables a } A}{\text{n (número de sucesos elementales posibles)}}$$

# Aplicaciones y limitaciones de la Regla de Laplace.

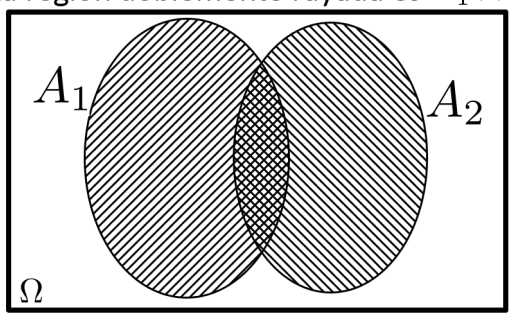
- Con la Regla de Laplace y un poco de Combinatoria (don't panic!) es posible responder a preguntas como estas:
  - ¿Cual es la probabilidad de que la suma de los resultados al lanzar dos dados sea igual a siete?
  - ¿Cuál es la probabilidad de que al tirar tres dados aparezca el seis en uno de los dados (no importa cual), pero sólo en uno de ellos?
  - En un paquete hay 20 tarjetas numeradas del 1 al 20. Se escogen al azar dos tarjetas. ¿Cuál es la probabilidad de que las dos que se han elegido sean la número 1 y la número 20? ¿Hay alguna diferencia entre sacar las dos tarjetas a la vez, o sacarlas consecutivamente sin remplazamiento? ¿Y si es con remplazamiento?
- Pero es necesario entender que la Regla de Laplace *no es una definición de Probabilidad*. En primer lugar porque sería una definición circular. Y en segundo lugar porque no sirve para responder a preguntas sencillas que tienen respuestas intuitivamente obvias como esta:
  - Si elegimos al azar un número real $x$ en el intervalo $[0, 1]$, ¿cuál es la probabilidad de que sea $1/3 \leq x \leq 2/3$? ¿Qué dice (a gritos) la intuición? Y ahora trata de pensar en este problema usando la regla de Laplace. ¿Cuántos casos posibles (valores de $x$) hay? ¿Cuántos son los casos favorables? Experimenta con este fichero de código R.

  La Regla de Laplace no se diseñó para tratar con valores continuos, como el $x$ de este ejemplo. Necesitamos una noción de Probabilidad más general.

# Teoría Axiomática de la Probabilidad.

- Los detalles técnicos son complicados pero, simplificando mucho hay tres ingredientes:
  a. Tenemos un *espacio muestral* $\Omega$ que es el conjunto de todos los posibles resultados de un experimento.
  b. Un *suceso aleatorio* es (casi) cualquier subconjunto de $\Omega$ (no *demasiado raro*).
  c. Una *función probabilidad* que representaremos con la letra $P$ que asigna un número $P(A)$ a cada suceso aleatorio $A$ del espacio muestral $\Omega$. La función probabilidad debe cumplir tres propiedades:
     1. $P(\Omega) = 1$.
     2. Sea cual sea el suceso aleatorio $A$, se tiene $0 \leq P(A) \leq 1$.
     3. Si $A_1$ y $A_2$ son dos sucesos aleatorios entonces
$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$
- Aquí $A_1 \cup A_2$ es la *unión* de sucesos y $A_1 \cap A_2$ la *intersección*, como ilustra el diagrama de Venn (cambia probabilidades por áreas e imagina que el área del rectángulo es 1).

La región doblemente rayada es $A_1 \cap A_2$



# Propiedades adicionales.

- En la próxima sesión veremos ejemplos concretos y útiles de como construir esas funciones de probabilidad tanto en casos discretos como continuos.
- La probabilidad del *suceso vacío* $\emptyset$ es 0; es decir $P(\emptyset) = 0$.
- Dos sucesos $A_1$ y $A_2$ se llaman *incompatibles* o *disjuntos* si su intersección es vacía; es decir, no pueden ocurrir a la vez. En tal caso:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

- Dado un suceso aleatorio $A$, el *suceso complementario* $A^c$ se define como "no ocurre $A$' '. Y siempre se cumple que

$$P(A^c) = 1 - P(A).$$

- Si $A \subset B$ (se lee: si $A$ es un subconjunto de $B$) entonces

$$P(A) \leq P(B)$$

- **Ejercicio:** Calcular la probabilidad de que un número de cuatro cifras tenga alguna repetida. Extra: diseña una simulación con R para comprobar tu resultado.

# Probabilidad condicionada.

- El concepto de probabilidad condicionada trata de reflejar los cambios que se producen en el valor de probabilidad $P(A)$ de un suceso cuando tenemos alguna *información adicional (pero parcial)* sobre el resultado de un experimento aleatorio.
- *Ejemplo.* ¿Cuál es la probabilidad de que al lanzar un dado obtengamos un número par? Está claro que es 0.5. Pero y si te dijera, sin revelarte el resultado, que al lanzar el dado hemos obtenido un número estrictamente mayor que 3. ¿Seguirías pensando que esa probabilidad es 0.5?
- Lo que ocurre en situaciones como esa es que queremos calcular la probabilidad de un suceso $A$ *sabiendo con certeza que* ha ocurrido otro suceso $B$, lo que se denomina probabilidad de $A$ condicionada por $B$ y se representa mediante $P(A \mid B)$. La definición, que se puede justificar con la Regla de Laplace, es:

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

- *Ejemplo (continuación):*

$$P(\text{dado par}|\text{sabiendo que dado} > 3) = \frac{P(\text{dado par y a la vez dado} > 3)}{P(\text{dado} > 3)} = \frac{2/6}{3/6} = \frac{2}{3}$$

- El denominado Problema de Monty Hall es un ejemplo famoso de como la información adicional altera nuestra estimación de probabilidades (ver también).

# Sucesos independientes.

- El suceso $A$ es independiente del suceso $B$ si el hecho de saber que el suceso $B$ ha ocurrido no afecta a nuestra cálculo de la probabilidad de que ocurra $A$. Es decir, la independencia significa que $P(A \mid B) = P(A)$. Hay una manera equivalente de escribir esto que deja claro que la independencia es simétrica:

$$A \text{ y } B \text{ son independientes significa que } P(A \cap B) = P(A)\,P(B)$$

- **Nunca confundas sucesos independientes e incompatibles** Los sucesos incompatibles no pueden ser independientes.
- Esta noción de independencia es una abstracción matemática, que raras veces coincidirá en la práctica con nuestra noción intuitiva de que dos fenómenos son independientes. Más adelante tendremos ocasión de profundizar en esta discusión y hablaremos de como medir en casos reales la independencia.

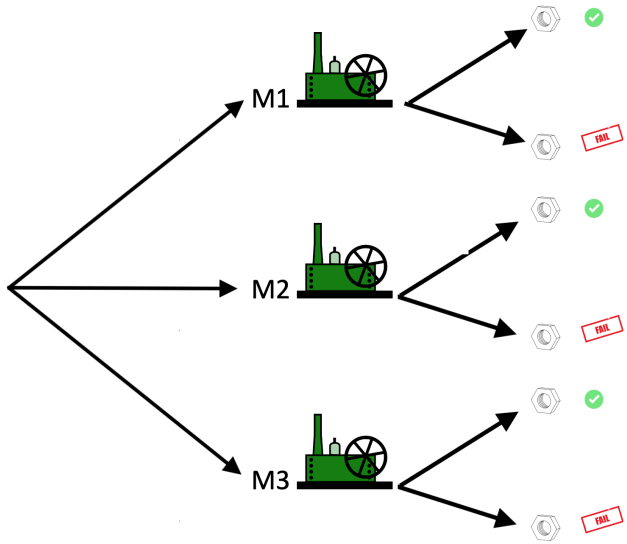
Sección 3

# Probabilidad total y Regla de Bayes.

# Teorema de la probabilidad total.

- Este resultado sirve para calcular la probabilidad de un suceso $A$ que puede ocurrir a través de uno de entre $k$ mecanismos excluyentes.
- *Ejemplo:* una fábrica produce un tipo de piezas usando tres máquinas. (a) Cada pieza proviene de una y una sola de esas máquinas. (b) Cada una de las máquinas produce una fracción conocida de las piezas y (c) tiene una tasa de piezas defectuosas también conocida. Con esa información queremos calcular la tasa total de piezas defectuosas.

• Sea $A$ el suceso *la pieza es defectuosa*; lo que queremos calcular es $P(A)$. Sean $M_1, M_2, M_3$ los sucesos *la pieza se fabrica en la máquina 1 o en la 2 o 3 respectivamente.*




- T conocemos las tres probabilidades condicionadas $P(A \mid M_1)$, $P(A \mid M_2)$ y $P(A \mid M_3)$.
- En problemas como este el Teorema de la Probabilidad Total afirma que:

$$P(A) = \underbrace{P(A \mid M_1)P(M_1) + P(A \mid M_2)P(M_2) + P(A \mid M_3)P(M_3)}_{\text{un término para cada máquina / camino}}$$

# Teorema de Bayes

- El Teorema de Bayes se usa en situaciones idénticas a la que acabamos de ver, pero sirve para hacer una *pregunta inversa*. *Sabiendo que la pieza es defectuosa*, ¿cuál es la probabilidad de que provenga de la máquina M1 (por ejemplo)? Se trata por tanto de calcular $P(M_1 \mid A)$. Y el resultado es:

$$P(M_1 \mid A) = \frac{P(A \mid M_1)P(M_1)}{P(A \mid M_1)P(M_1) + P(A \mid M_2)P(M_2) + P(A \mid M_3)P(M_3)}$$

  Fíjate en que el denominador es $P(A)$.

- *Ejemplos:*
  - responde a la pregunta con la que se abre esta página.
  - Lo más difícil al usar el Teorema de Bayes suele ser identificar los datos de forma correcta, Un ejemplo típico se ilustra en este problema: *Un hospital tiene dos quirófanos en funcionamiento. En el primero se han producido incidentes en el 20 % de sus operaciones y el segundo sólo en el 4 %. El número de operaciones es el mismo en ambos quirófanos. La inspección hospitalaria analiza el expediente de una operación, elegido al azar y observa que en esa operación se produjo un incidente. ¿Cuál es la probabilidad de que la operación se realizara en el primer quirófano?*

# Jugando con el Teorema de Bayes y R

- Vamos a usar una tabla de datos sobre spam en mensajes de correo electrónico. La tabla se llama `spam` y pertenece a la librería `kernlab`. Instala la librería y carga la tabla con `data(spam)`, La tabla contiene datos sobre varios miles de mensajes de correo. La última columna contiene la clasificación como spam o no spam. Las primeras 48 columnas indican el porcentaje de palabras del mensaje que coinciden con el título de la columna. Aquí se muestra una parte de la tabla:

```
library(kernlab)
data(spam)
spam[1:4, c(1:10, 58)]
```

```
  make address  all num3d  our over remove internet order mail type
1 0.00    0.64 0.64     0 0.32 0.00   0.00     0.00  0.00 0.00 spam
2 0.21    0.28 0.50     0 0.14 0.28   0.21     0.07  0.00 0.94 spam
3 0.06    0.00 0.71     0 1.23 0.19   0.19     0.12  0.64 0.25 spam
4 0.00    0.00 0.00     0 0.63 0.00   0.31     0.63  0.31 0.63 spam
```

- Con estos datos y usando funciones de R responde a estas preguntas:
  - ¿Cuál es la probabilidad de que un mensaje elegido al azar sea spam?
  - ¿Cuál es la probabilidad de que un mensaje elegido al azar contenga la palabra *order*?
  - *Sabiendo* que un mensaje es spam, ¿cuál es la probabilidad de que contenga la palabra *order*?
  - Y ahora, usando la fórmula de Bayes, vamos a construir el programa antispam más simple del mundo: sabiendo que un mensaje contiene la palabra *order*, ¿cuál es la probabilidad de que sea spam?
- Este método es muy rudimentario, pero cuando aprendas algoritmos de clasificación estudiarás el método Naive Bayes (Bayes ingenuo) que se basa en ideas similares.

Sección 4

# Tablas de Contingencia.

# Tablas de contingencia 2x2

- En el problema anterior nos hemos encontrado con una situación típica en la que hay dos factores binarios. Un factor S con valores *spam / no spam* y un factor O, con valores "contiene order/ no contiene order." Al combinarlos hay cuatro casos posibles que podemos representar en una tabla dos por dos.
- Primero usaremos `dplyr` para obtener una tabla en la que solo aparezcan esos dos factores, aprendiendo de paso alguna manipulación adicional:

```
library(tidyverse)
spam = spam %>%
  select(order, type) %>%
  mutate(hasOrder = factor(order > 0, # Creamos el factor hasOrder
                           levels = c(TRUE, FALSE),
                           labels = c("order", "no order")),
         type = relevel(type, ref = "spam"), # Reordenamos los niveles
         -order) # y elminamos el factor order original
```

Ahora podemos obtener la tabla con

```
table(spam$hasOrder, spam$type)
```

```
           spam nonspam
  order     555     218
  no order 1258    2570
```

# Vocabulario adicional para tablas de contingencia 2x2

- El lenguaje de las tablas de contingencia proviene en buena medida del contexto de las pruebas diagnósticas para enfermedades. Esas pruebas no son infalibles: a veces dan como resultado que una persona padece la enfermedad, cuando en realidad no es así. Es lo que se llama un *falso positivo* (FP). En otras ocasiones será al contrario. La prueba dirá que la persona está sana, aunque de hecho está enfermo. Eso es un *falso negativo* (FN). Los resultados correctos, que están en la diagonal principal de la tabla, son los TP (true positives) y los TN (true negatives).
- Por ejemplo, podemos tener una tabla como esta:

| | | **Padecen la enfermedad** | | |
|---|---|---|---|---|
| | | Enfermo | Sano | Total |
| **Resultado de la Prueba** | Positivo | TP = 192 | FP = 158 | 350 |
| | Negativo | FN = 4 | TN = 9646 | 9650 |
| | Total | 196 | 9804 | 10000 |

- Vamos a usar este script de R para aprender algo más de lenguaje sobre tablas de contingencia y de como manejarlas con R.
- Una prueba diagnóstica es un *clasificador* de pacientes. Más adelante vamos a encontrar muchos algoritmos clasificadores, porque clasificar es una de las tareas básicas en *Machine Learning*. Veremos entonces que en ese contexto se usa mucho el vocabulario de pruebas diagnósticas.

# Referencias para la sesión

## Enlaces

- Código de esta sesión
- Cookbook for R
- Página web de ggplot2, que contiene el resumen (chuleta) elaborado por RStudio.
- Resumen sobre importación de datos a R (chuleta) elaborado por RStudio.
- Web del libro PostData y los tutoriales asociados. Para esta sesión se recomienda el Capítulo 2.

## Bibliografía