Fundamentos del *big data*: tratamiento de los datos

PID_00247341

Antoni Morell

Tiempo mínimo de dedicación recomendado: 2 horas

Universitat Oberta de Catalunya

Índice

Introducción

Como hemos visto en el material anterior («Fundamentos del *big data*: arquitectura del sistema»), tres pasos fundamentales en cualquier problema que trabaje sobre datos, sea o no a gran escala, son (por orden):

1) capturar los datos;
2) limpiar los datos y
3) analizar los datos para responder a una o varias cuestiones planteadas.

Vamos a centrarnos ahora en las herramientas clásicas para abordar el tercer punto de análisis de los datos. Es importante tener presentes estas herramientas y conceptos, puesto que constituyen la base de los sistemas de *big data*, además de que pueden resultar útiles por sí mismas en la fase de desarrollo.

Este material se plantea en su vertiente más práctica sobre el lenguaje de programación R, puesto que es ampliamente usado en el campo de la estadística clásica, y existen multitud de paquetes desarrollados por la comunidad y abiertos a los usuarios. Cabe decir que no se limita únicamente al tratamiento de problemas convencionales (de tamaño moderado), ya que hoy día constituye una de las posibilidades que hay que considerar también en *big data* a través de las herramientas de integración con plataformas específicas, como por ejemplo Spark. Empezaremos tratando las herramientas de visualización de datos, puesto que pueden ser de gran utilidad para comprender mejor nuestro problema y orientarnos hacia la solución. En este sentido, R dispone de buenos paquetes que nos facilitan el trabajo.

En la segunda parte, ya nos centraremos en las herramientas estadísticas clásicas. Hablaremos aquí de los conceptos básicos de probabilidad, regiones de confianza y tests de hipótesis, los cuales nos servirán para ver si una hipótesis de partida es soportada por los datos obtenidos. Por ejemplo, imaginemos que queremos saber si un fármaco A produce un determinado efecto Y. El procedimiento que hay que seguir será tomar una población de test y administrar A a parte de la población, y placebo al resto, y después medir los efectos producidos y decidir con fundamento estadístico si realmente hay diferencias entre los fármacos o no.

1. Explorar los datos

1.1. Motivación

Como ya se ha apuntado en la introducción, la visualización de los datos de forma gráfica es esencial por distintos motivos:

1) permite entender mejor los datos;
2) la simple exploración visual permite ver posibles patrones;
3) también permite intuir qué estrategias para modelar los datos nos serán más útiles;
4) nos permite comprobar hipótesis y modelos; y
5) nos sirve como una muy buena herramienta para comunicar resultados.

Cuando el objetivo perseguido se corresponde con los puntos 1) - 4) anteriores, entonces hablaremos de *gráficos exploratorios*, y son gráficos que:

1) se hacen de forma rápida;

2) elaboraremos varios para un mismo problema;

3) el objetivo es entender;

4) no son definitivos (por ejemplo, las leyendas o ejes son solo para nuestra información y se podrán pulir en caso de usar dichos gráficos para comunicar resultados); y

5) usaremos atributos como el color o tamaño de los elementos para distinguir grupos y proporcionar información.

A continuación, centraremos el estudio en el lenguaje de programación R, aunque no se pretende hacer una descripción completa del lenguaje y sus funcionalidades, pero sí una demostración en un ámbito conceptual de las distintas herramientas de las que disponemos. Lo importante de esta parte del material es tomar conciencia de la utilidad de las herramientas gráficas.

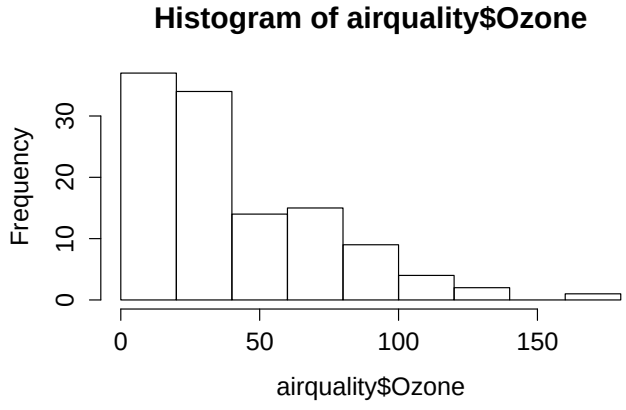
1.2. Herramientas gráficas en R

R tiene distintos paquetes gráficos con los que podemos visualizar los datos. Nos permiten visualizar por pantalla, o bien generar un archivo gráfico con el dibujo en cuestión. Los tres paquetes principales que tiene R para tratar con gráficos son el paquete básico, Lattice y Ggplot2. A continuación, se mostrarán algunos ejemplos de gráficos

elaborados sobre el paquete básico, cuya filosofía de funcionamiento es semejante a la de Matlab. Empezamos por ejecutar el comando correspondiente al tipo de gráfico que queremos generar (con una serie de atributos que ya se pueden especificar de entrada, como por ejemplo el título del gráfico). A partir de este momento, sucesivos comandos servirán para ir modificando el gráfico en cuestión a nuestro gusto. Los otros paquetes gráficos nos permitirán hacer gráficos más elaborados en un ámbito de diseño. Cada uno tiene su forma de trabajar. Por ejemplo, en Lattice los gráficos se hacen siempre con un único comando. En cambio, en el sistema Ggplot2 crearemos un objeto que describe el propio gráfico a partir de los datos, y a continuación podremos añadir cambios estéticos (color, tamaño), objetos geométricos, facetas para gráficos múltiples y otros elementos. Este será nuestro paquete de referencia si queremos generar gráficos para publicar.

Nos centramos a continuación en el paquete básico, y tomaremos como datos de ejemplo los del paquete *datasets* de R, por lo que en la consola de Rstudio habrá que introducir `library(datasets)`. En concreto, usaremos el dataframe *airquality*, que recoge distintas medidas relacionadas con la calidad del aire en Nueva York entre mayo y septiembre de 1973. Si queremos ver su contenido, tecleamos `str(airquality)`. En general, las funciones más empleadas para el análisis exploratorio de datos son el histograma, si queremos analizar una sola variable, y los *scatterplots* (o también los *boxplots*) si lo que queremos es analizar la interacción entre dos variables. Habitualmente trabajaremos con gráficos en 2D, ya que son más fáciles de interpretar visualmente que los gráficos en 3D. En cualquier caso, y dado que la limitación está en el 3D, ya que no podemos hacer gráficos *n*-dimensionales, la solución para el análisis de un conjunto de variables pasa por hacer varios gráficos 2D que muestren las interacciones entre las variables dos a dos.

Figura 1. Histograma de la concentración de ozono



Ejecutando `hist(airquality$Ozone)` obtenemos el resultado de la figura 1, en la que apreciamos la distribución de la concentración de ozono entre todas las medidas tomadas. Si queremos afinar más y ver, por ejemplo, la evolución de esta concentración mes a mes, entonces podemos emplear un *boxplot*. Si escribimos en la consola

```
boxplot(Ozone ~ Month, airquality, xlab = "Mes", ylab = " Ozono (ppb) ")
```

obtendremos un resumen estadístico, mes a mes, de los datos recolectados. El resultado se puede apreciar en la figura 2. Dentro de cada una de las cajas, vemos:

1) en trazo más grueso, la mediana o segundo cuartil (no valor medio) de la muestra;

2) el límite superior marca el tercer cuartil (solo el $25\,\%$ de las muestras tienen un valor superior); y

3) el límite inferior marca el primer cuartil (el $25\,\%$ de las muestras tienen un valor inferior).

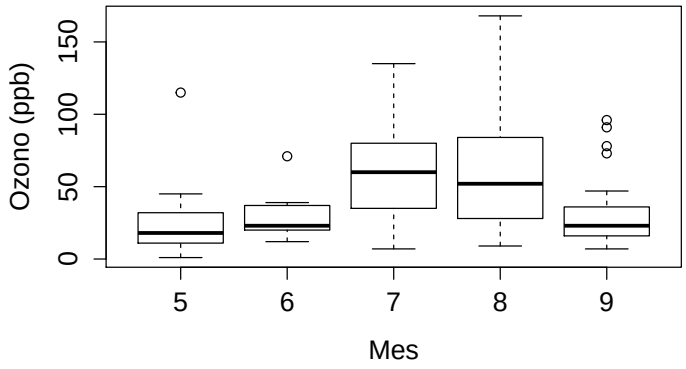
Fuera de las cajas, nos encontramos con:

1) los círculos que nos marcan los *outliers* y

2) los bigotes que nos marcan los valores mínimo y máximo de los datos.

Cabe aclarar que, una vez recogidos los datos y obtenida su mediana y cuartiles, se definen unas barreras para detectar posibles *outliers*. Típicamente, para ello se extienden el primer y tercer cuartil una distancia 1,5 veces la distancia entre los dos. Por lo tanto, los bigotes son los valores máximo y mínimo, sin considerar los *outliers*. Observemos también que en la llamada a la función `boxplot` se han especificado ya las etiquetas de los ejes.

Figura 2. *Boxplot* (diagrama de caja) de la concentración de ozono en función del mes



Por último, veremos algunos ejemplos de cómo trabajar con los *scatterplots* o nubes de puntos. No obstante, antes vamos a introducir una serie de funciones importantes del sistema básico de gráficos de R, que son `plot` (para dibujar); `lines` (añadir líneas a un dibujo); `points` (añadir puntos); `text` (añadir etiquetas de texto); `title` (añadir título); `mtext` (añadir texto a los márgenes); y `axis` (modificar lo relativo a los ejes). Podemos obtener información de cada una de las funciones a través de la ayuda de R, por ejemplo `help("points")`. En cuanto a los parámetros que emplean estas funciones, enumeramos a continuación algunos de los más importantes, que son `pch` (el símbolo de dibujo, un círculo por defecto); `lty` (el tipo de línea,

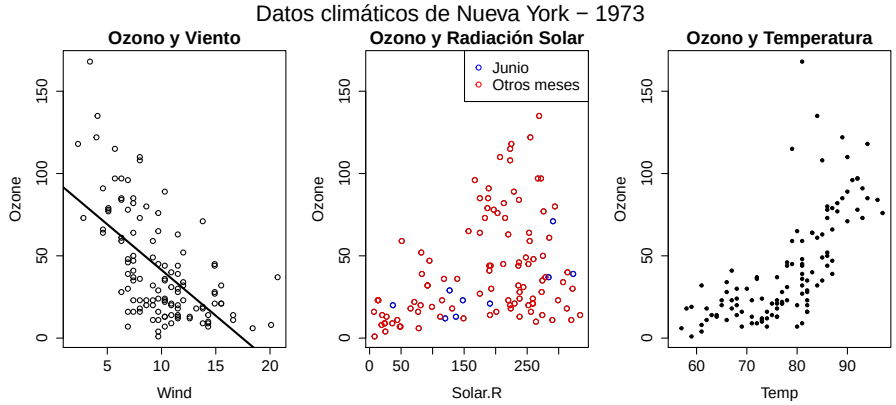
sólida por defecto); `lwd` (grosor de línea); `col` (color); `xlab, ylab` (etiquetas de los ejes); `bg` (color de fondo); `mar` (márgenes interiores); `oma` (márgenes exteriores); o `mfrow` (número de gráficos por fila, llenado en sentido de las filas). Teniendo esto en cuenta, podemos ejecutar los comandos que aparecen en la figura 3, con los que obtendremos el gráfico de la figura 4.

Figura 3. Ejemplo de gráfico múltiple

```
library(datasets)
par(mfrow = c(1, 3), mar = c(4, 5, 2, 1), oma = c(0, 0, 2, 0))
model <- lm(Ozone ~ Wind, airquality)
with(airquality, {
  plot(Wind, Ozone, main = "Ozono y Viento")
  abline(model, lwd = 2)
  plot(Solar.R, Ozone, main = "Ozono y Radiación Solar")
  with(subset(airquality, Month == 6),
       points(Solar.R, Ozone, col = "blue"))
  with(subset(airquality, Month != 6),
       points(Solar.R, Ozone, col = "red"))
  legend("topright", pch = 1, col = c("blue", "red"),
         legend = c("Junio", "Otros meses"))
  plot(Temp, Ozone, pch = 20, main = "Ozono y Temperatura")
  mtext("Datos climáticos de Nueva York - 1973", outer = TRUE)
})
```

En este caso, lo primero que hacemos es definir un gráfico múltiple de 1 fila y 3 columnas. También definimos los márgenes internos y externos con `mar` y `oma`. Hay que tener en cuenta que los márgenes se especifican en número de líneas de texto, se define primero el margen inferior y luego vamos dando la vuelta en el sentido de las agujas del reloj. Así pues, `mar = c(4, 4, 2, 1)` define un margen interior de 4 líneas de texto por debajo y 4 más a la izquierda, lo que le da gran holgura para las etiquetas de los ejes. Respecto al margen exterior, dejamos 2 líneas en la parte superior para el título. A continuación, calculamos el modelo de regresión lineal de la concentración de ozono con respecto al viento (lo veremos en el apartado 2.3) y ya procedemos a dibujar las distintas nubes de puntos. En el primer caso (ozono y viento), se superpone la recta de regresión calculada. En el segundo caso (ozono y radiación), se distingue con colores distintos los puntos que corresponden al mes de junio del resto. Para esto, hemos empleado la función `subset` de R, que nos permite tomar el subconjunto de interés de nuestro conjunto de datos o *dataframe*.

Figura 4. Múltiples gráficos 2D exploratorios



En estos ejemplos, hemos visto la utilidad de las herramientas gráficas a la hora de hacer una primera exploración de los datos y guiarnos en el análisis posterior. A simple vista, somos capaces de ver, entre otros aspectos, si existe relación entre variables o si un factor (como por ejemplo el mes en el que se han tomado los datos en el caso anterior) tiene relevancia o no.

2. Análisis estadístico básico

En esta parte del material, vamos a empezar tratando los conceptos básicos de probabilidad y variables aleatorias para luego ya centrarnos en aspectos prácticos del análisis estadístico. Existen multitud de referencias básica de estadística [Devore (2011), Casella y Berger (2001), Rosner (2011)] y, en concreto, el contenido de este material está basado en [Devore (2011)]. Nuestro objetivo básico será poder comprobar si nuestras hipótesis o modelos son respaldados por los datos que tenemos, y hasta qué punto.

2.1. Probabilidad

La probabilidad asociada a un suceso futuro A es una medida, de valor entre 0 y 1, que nos dice lo fácil o difícil que es que el evento ocurra. Asociado al concepto de probabilidad está siempre el de *variable aleatoria* (v.a.), que denotamos $X(A)$ y que se trata de una variable cuyo valor resulta de la medición del suceso aleatorio A, una vez llevado a cabo el experimento. Por ejemplo, si el experimento es tirar una moneda, $A = \{\text{cara}, \text{cruz}\}$ podemos considerar $X(\text{cara}) = 0$ y $X(\text{cruz}) = 1$. No obstante, habitualmente se omite el suceso aleatorio y escribimos simplemente X. La v.a. puede tomar un valor dentro de un conjunto finito de posibilidades, como en el caso de la moneda (variable aleatoria discreta), o bien dentro de un conjunto infinito de posibilidades (variable aleatoria continua), como por ejemplo sería la medición del consumo eléctrico de un hogar de 1 día.

En v.a. continuas, la máxima información que podemos obtener de una v.a. es a través de su *función densidad de probabilidad** $f_X(X)$. El valor de la pdf en un punto $X = x_0$ indica, en términos de densidad de probabilidad, lo fácil o difícil que es que este ocurra. Hablamos en términos de densidad de probabilidad, porque la probabilidad de un único valor es en realidad cero, y solo cuando consideramos un intervalo de posibles valores (por pequeño que sea) tiene sentido hablar de probabilidad. Así pues, la probabilidad de que el resultado de un experimento dé lugar a $X \in [A, B]$ se calcula como

*** fdp o *pdf* en inglés.**

$$P(A \leq X \leq B) = \int_A^B f_X(X)\, dX,$$

y la probabilidad de todos los eventos posibles es

$$P(-\infty \leq X \leq +\infty) = \int_{-\infty}^{+\infty} f_X(X)\, dX = 1.$$

En v.a. discretas el concepto es el mismo, y llamamos a la función también $f_X(X)$, pero en este caso sí que asociamos probabilidades a valores discretos de X. Se puede hablar entonces de *función de masa de probabilidad* (pmf). En este caso, si la v.a. puede tomar N posibles valores $x_1, \ldots, x_N$, la probabilidad de que sucedan por ejemplo tres de estos eventos (digamos x_1, x_5 y x_8) sería

$$P(X \in \{x_1, x_5, x_8\}) = f_X(x_1) + f_X(x_5) + f_X(x_8)$$

y, al igual que en el caso anterior,

$$\sum_{i=1}^{N} f_X(x_i) = 1,$$

es decir, las probabilidades de todos los eventos posibles tienen que sumar la unidad, como es lógico.

2.1.1. Parámetros estadísticos

Si bien la pdf o pmf nos dan mucha información acerca de un suceso aleatorio, no dejan de ser funciones. Si queremos comparar dos situaciones, no será obvio comparar directamente las funciones. Entonces, es más conveniente para el análisis comparar parámetros (números) que resuman la información de la pdf/pmf correspondiente. Los parámetros estadísticos más importantes son la media y la varianza.

Empecemos por el caso discreto. La media se calcula como

$$\bar{X} = E\{X\} = \sum_{i=1}^{N} x_i f_X(x_i),$$

y simplemente hacemos un promedio de todos los valores x_i que puede tomar la variable, ponderados por las correspondientes probabilidades. Esto es lo que hace el operador esperanza $E\{\cdot\}$, es decir, promediar según probabilidades lo que pongamos dentro, que en el caso de la media es directamente el valor de la v.a. Por otro lado, la varianza se calcula como

$$\sigma_X^2 = E\left\{(X - \bar{X})^2\right\} = \sum_{i=1}^{N} (x_i - \bar{X})^2 f_X(x_i),$$

y en este caso promediamos según probabilidades lo que se aleja cada uno de los posibles valores que toma la variable respecto a la media (en términos de distancia al cuadrado). En definitiva, nos proporciona una idea de cuánto están de dispersos los valores respecto a su media. Si la varianza es muy baja, esperaremos que todos los valores estén cerca del valor medio. Si es muy alta, habrá valores alejados de la media

(supondremos a derecha e izquierda, aunque esta información desaparece al elevar al cuadrado).

En el caso continuo, los conceptos son los mismos, solo varía el cálculo y pasan de sumatorios a integrales. En este caso, la media se calcula como

$$\bar{X} = E\{X\} = \int_{-\infty}^{+\infty} X f_X(X)\, dX$$

y la varianza según

$$\sigma_X^2 = E\left\{(X - \bar{X})^2\right\} = \int_{-\infty}^{+\infty} (X - \bar{X})^2 f_X(X)\, dX.$$

En muchas ocasiones, hablamos de desviación típica o estándar σ_X en vez de esperanza, para ponerla en términos de distancia y no de distancia al cuadrado, así que simplemente $\sigma_X = \sqrt{\sigma_X^2}$.

2.1.2. Promedios de muestreo o empíricos

En análisis estadístico trabajamos con el concepto de probabilidad, pero desconocemos la pdf/pmf que hay detrás. Lo que sí tendremos son muestras del resultado de los sucesos. Por ejemplo, si nuestra v.a. es el consumo eléctrico de un hogar en un periodo de 24 h, lo que tendremos es, pasados los días, meses o años, varias lecturas (muestras) de cuál ha sido el consumo cada día. No sabremos la pdf, pero sí tenemos un conocimiento indirecto de la misma. De hecho, a partir de las muestras de una v.a., nos podemos hacer una idea tanto de su pdf como de su media, varianza o desviación estándar. La pdf se puede ver dibujando su histograma, y a mayor número de muestras, mejor la aproximaremos. Observad que los valores con menor densidad de probabilidad ocurren con baja frecuencia y por tanto, si tenemos pocas muestras, entonces lo más probable es que no estén representados (que no aparezcan en el histograma). En cambio, cuando tengamos muchas muestras, entonces alguna de ellas sí podrá representar estos valores (con la baja frecuencia que le corresponda).

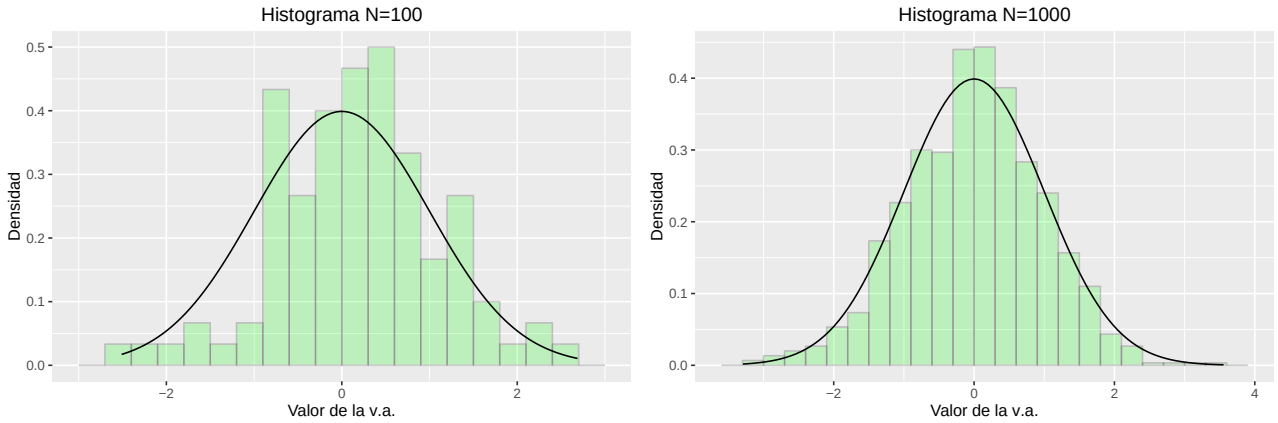
La estimación de media, varianza y desviación típica a partir de N muestras de X que llamaremos $s_1, \ldots, s_N$ se calcula como:

- Media empírica: $\hat{\bar{X}} = \sum_{i=1}^{N} s_i \frac{1}{N}$
- Varianza empírica: $S_X^2 = \frac{\sum_{i=1}^{N} (s_i - \hat{\bar{X}})^2}{N-1}$
- Desviación típica empírica: $S_X = \sqrt{S_X^2}$

Observad que en el cálculo de la media lo que estamos haciendo es, en realidad, algo equivalente al operador esperanza, asignando a cada muestra un peso de $1/N$. Por lo

tanto, el peso final de un determinado valor s_i vendrá dado por el número de veces que se repita entre las N muestras, con lo que este proceso lo que hace es una estimación implícita de la probabilidad de dicho valor. Lo mismo sucede en los casos de varianza y desviación típica, aunque por razones que escapan de este texto usamos $N-1$ en vez de N (solo tiene importancia para valores pequeños de N).

Figura 5. Histograma de la v.a. gaussiana con $N = 100$ y $N = 1000$



En la figura 5, podemos ver los histogramas de 100 y 1000 muestras correspondientes a una pdf de tipo gaussiana de media 0 y varianza 1, junto con el dibujo de la propia pdf. La pdf gaussiana responde a la expresión $f_X(X) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{(X-\bar{X})^2}{2\sigma_X^2}}$, donde $\bar{X}$ y σ_X^2 son media y varianza, respectivamente. Para las muestras generadas, la estimación de media y varianza vale 0.17 y 0.89 para $N = 100$, -0.006 y 0.94 para $N = 1000$.

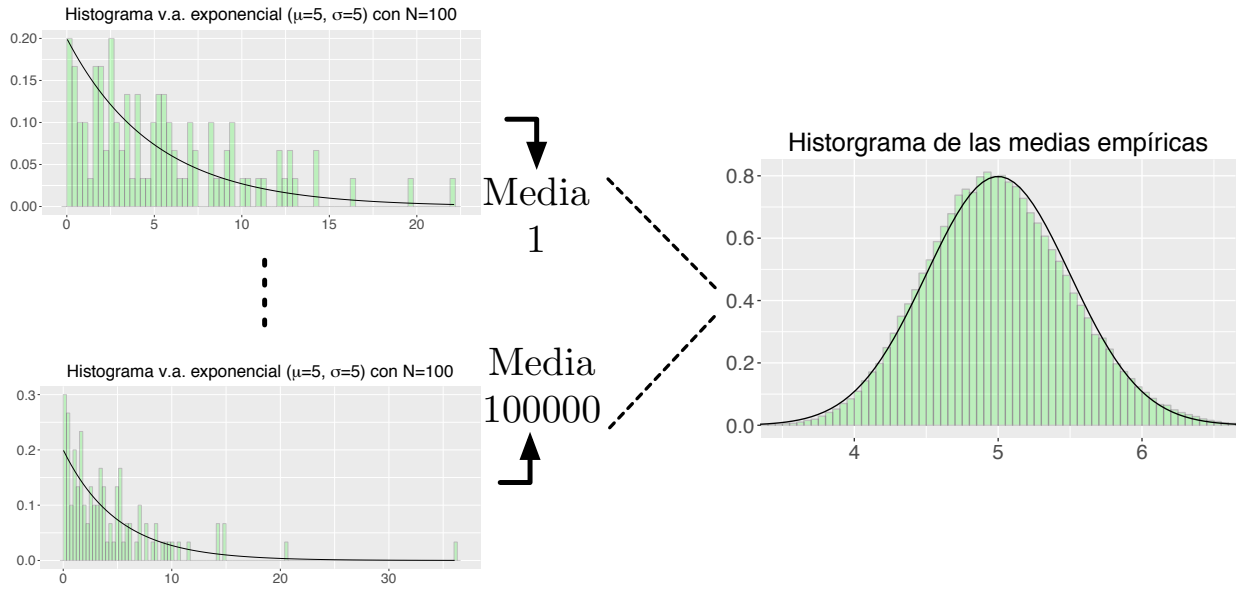
2.1.3. Ley de los grandes números

Por último, si nos fijamos ahora en la media empírica, se sabe que su varianza es para cualquier pdf $\sigma_{\hat{X}}^2 = \frac{\sigma_X^2}{N}$, como veremos en el siguiente apartado. Aplicado al caso anterior, por ejemplo, en el que hemos elegido $\sigma_X = 1$, este resultado nos dice que si repetimos el experimento de tomar, por ejemplo, $N = 100$ muestras de una v.a. gaussiana muchas veces y dibujamos su histograma, la dispersión de las medias estimadas corresponderá a $1/100$. En cambio, si lo hacemos con 1000 muestras acertaremos más, ya que esta dispersión bajará a $1/1000$. Estamos hablando en todo momento de muestras que se generan cada una independientemente de las anteriores, y siguiendo una misma distribución (lo que se conoce como iid o independientes e idénticamente distribuidas), y el resultado obtenido es consistente con la ley de los grandes números, que nos dice que la media empírica converge a la media real a medida que N va creciendo. Esto será la base de razonamiento de los dos puntos que veremos a continuación, en los que el objetivo será sacar evidencia estadística de un determinado acontecimiento a partir de muestras del mismo.

2.2. Teorema central del límite, distribución t de Gosset e intervalos de confianza

El teorema central del límite (CLT) es uno de los resultados más importantes en estadística y, enfocado a nuestros intereses aquí, nos dice que para una distribución cualquiera de media $\bar{X} = \mu$ y $\sigma_X^2 = \sigma^2$, la media empírica $\hat{\bar{X}}$ es aproximadamente gaussiana de media μ y varianza σ^2/N para un número de muestras N suficientemente grande. La figura 6 ilustra estas ideas. En ella, vemos cómo la distribución de la media empírica de 100 muestras de muchas v.a. exponenciales de media 5 y desviación típica 5 es aproximadamente gaussiana de media 5 y desviación típica $5/\sqrt{100} = 0{,}5$.

Figura 6. Teorema central del límite (CLT)



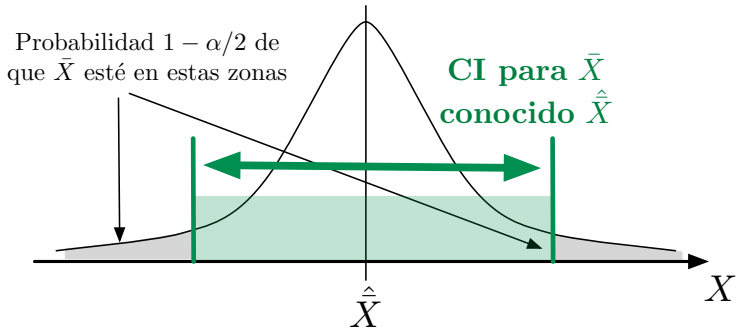
Relacionada con el CLT, encontramos la distribución t de Student, que fue publicada por el estadístico William S. Gosset bajo el seudónimo de Student mientras trabajaba para la destilería Guinness. Se puede entender como una versión del CLT que describe la distribución de la estimación de la media de N muestras de una v.a. gaussiana de media $\bar{X}$ y varianza σ_X^2. En este caso, no se requiere que N sea grande a cambio de imponer que la distribución sea gaussiana. No obstante, la distribución t-Student funciona bien con otras distribuciones, mientras sean de tipo campana. Para distribuciones no simétricas, es posible evaluar otros parámetros como la mediana; o bien hacer transformaciones como la logarítmica, para reducirlo al caso conocido; o bien estudiar cada distribución en particular; pero todo esto queda fuera del alcance de este material. La distribución t-Student tiene como parámetro el número de muestras N, en concreto lo que se conoce como grados de libertad de la distribución o df, siendo $df = N - 1$. Para valores pequeños de N, la forma de la distribución es de tipo campana, pero más plana que la gaussiana definida por el CLT, y a medida que N va creciendo, la distribución t-Student tiende a la gaussiana.

Tanto el CLT como la distribución t-Student nos sirven para, dadas N muestras de una v.a. y calculada su media empírica, determinar una región sobre la cual debe estar la media real $\bar{X}$. A esta región de incertidumbre la llamamos intervalo de confianza (CI), y se identifica por un valor α, de tal modo que la probabilidad $\Pr(\bar{X} \in \text{CI}) = 1 - \alpha$. Valores típicos de α son 0.05 y 0.01 para los CI del 95 % y 99 %, respectivamente.

2.2.1. Intervalos de confianza para una muestra

Ya sea trabajando con el CLT o con la distribución t-Student, el cálculo de CI implicará determinar los valores límite respecto a los cuales el área de la distribución es exactamente $1-\alpha$. En los dos casos, esta integración no se puede hacer de forma analítica, y tendremos que recurrir a tablas. Además, dada la simetría de las distribuciones, lo que se hace es buscar el valor de X a partir del cual el área bajo la curva sea $\alpha/2$. Por lo tanto, fijamos la probabilidad de pasar uno de los límites a $\alpha/2$. Por consiguiente, la probabilidad de pasar el otro límite será también $\alpha/2$, y la probabilidad de caer dentro de CI será $1-\alpha$, como se puede apreciar en la figura 7.

Figura 7. Intervalos de confianza



Ahora bien, dado que resulta poco práctico tabular las distribuciones para todos los posibles valores de $\hat{\bar{X}}$ y S^2, la alternativa es normalizarlas siempre a una media 0 y varianza 1. Consideremos primero el caso del CLT. Esto es equivalente a decir que la v.a. $Z = \frac{\hat{\bar{X}}-\bar{X}}{\sigma_X/\sqrt{N}}$ verifica $z \sim \mathcal{N}(0,1)$, o sea, es normal de media 0 y varianza unidad. Para finalmente calcular el CI de $1-\alpha$, tenemos que ver qué valor de $Z = z$ hace que $\int_z^{+\infty} f_Z(Z)\,dZ = \alpha/2$. Esta información la encontramos en las tablas o bien empleando los programas estadísticos. Por ejemplo, para $\alpha = 0{,}05$, $z_{\alpha/2} \approx 1{,}96$ y para $\alpha = 0{,}01$, $z_{\alpha/2} \approx 2{,}575$. Esto mismo lo vemos en R a través de la función `qnorm`, por ejemplo `qnorm(0.975)=1.959964`. Ahora, ya podemos definir el CI de la media como:

$$-z_{\alpha/2} \leq Z = \frac{\hat{\bar{X}} - \bar{X}}{\sigma_X/\sqrt{N}} \leq z_{\alpha/2} \quad \longrightarrow \quad \hat{\bar{X}} - z_{\alpha/2}\frac{\sigma_X}{\sqrt{N}} \leq \bar{X} \leq \hat{\bar{X}} + z_{\alpha/2}\frac{\sigma_X}{\sqrt{N}} \quad (1)$$

Observad que salvo que sepamos σ_X de antemano, no la podremos obtener de nuestra muestra, pero sí podemos obtener S_X. En el caso de que N sea suficientemente gran-

de, podemos usar S_X. Para N pequeño debemos recurrir a la t-Student. Del mismo modo, definimos $T = \frac{\hat{\bar{X}}-\bar{X}}{S/\sqrt{N}}$ y el CI de la media será:

$$\hat{\bar{X}} - t_{\alpha/2,n-1}\frac{\sigma_X}{\sqrt{N}} \leq \bar{X} \leq \hat{\bar{X}} + t_{\alpha/2,n-1}\frac{\sigma_X}{\sqrt{N}} \quad (2)$$

donde $t_{\alpha/2,n-1}$ será el valor t que hace $\int_t^{+\infty} f_T(T;\ N-1)\,dT = \alpha/2$, con $f_T(T;\ N-1)$ la distribución t-Student de $N-1$ grados de libertad. En R tenemos la función `qt` para calcular estos valores, la cual necesita un segundo parámetro para poder indicar los grados de libertad de la distribución.

2.2.2. Intervalos de confianza para dos muestras

De mayor interés que el análisis de una muestra es el de dos, ya que nos será útil, por ejemplo, para comprobar si un determinado tratamiento tiene efecto cuando es aplicado sobre el grupo que hay que tratar y otro grupo de control. Si no surge efecto alguno, las medias de lo que midamos deben ser muy parecidas. Si realmente surge efecto, debemos notar un sesgo o desviación entre los grupos, con cierto peso estadístico. Por lo tanto, representando un grupo como la v.a. X (N muestras), y el otro como la v.a. Y (M muestras), nos interesa analizar $\hat{\bar{X}} - \hat{\bar{Y}}$.

Con resultado parecido al CLT, sabemos que para N, M suficientemente grandes la distribución de $\hat{\bar{X}} - \hat{\bar{Y}}$ es aproximadamente normal de media $\bar{X} - \bar{Y}$ y varianza $\sigma_X^2/N + \sigma_Y^2/M$, y se puede sustituir la varianza real, por ejemplo σ_X^2, por la obtenida de las muestras correspondientes, es decir, S_X^2. Al igual que en el caso de una muestra definimos $Z = \frac{(\hat{\bar{X}}-\hat{\bar{Y}})-(\bar{X}-\bar{Y})}{\sqrt{S_X^2/N + S_Y^2/M}}$, quedando el CI de $\bar{X} - \bar{Y}$ y nivel de confianza $1-\alpha$ como

$$(\hat{\bar{X}} - \hat{\bar{Y}}) - z_{\alpha/2}\sqrt{\frac{S_X^2}{N} + \frac{S_Y^2}{M}} \leq \bar{X} - \bar{Y} \leq (\hat{\bar{X}} - \hat{\bar{Y}}) + z_{\alpha/2}\sqrt{\frac{S_X^2}{N} + \frac{S_Y^2}{M}} \quad (3)$$

También como en el caso de una muestra, si alguna de las muestras o las dos son pequeñas, entonces debemos recurrir a la t-Student y definir $T = \frac{(\hat{\bar{X}}-\hat{\bar{Y}})-(\bar{X}-\bar{Y})}{\sqrt{S_X^2/N + S_Y^2/M}}$. En este caso, los grados de libertad de la distribución dependen del tamaño de las muestras y de sus varianzas (de muestra), y se calculan como $df = \frac{(S_X^2/N + S_Y^2/M)^2}{\frac{(S_X^2/N)^2}{N-1} + \frac{(S_Y^2/M)^2}{M-1}}$. El CI es entonces

$$(\hat{\bar{X}} - \hat{\bar{Y}}) - t_{\alpha/2,df}\sqrt{\frac{S_X^2}{N} + \frac{S_Y^2}{M}} \leq \bar{X} - \bar{Y} \leq (\hat{\bar{X}} - \hat{\bar{Y}}) + t_{\alpha/2,df}\sqrt{\frac{S_X^2}{N} + \frac{S_Y^2}{M}} \quad (4)$$

Por lo tanto, en las tablas de la t-Student buscaremos los valores $t_{\alpha/2,df}$ en función de dos parámetros: nivel de confianza y grados de libertad. Por ejemplo, para un nivel de confianza del $95\,\%$ buscaremos en las tablas para $\alpha/2 = 0{,}025$ y el valor df calculado.

2.2.3. Otros casos de interés

Los intervalos que se han planteado aquí son de algún modo los más generalistas, pero existen también otros casos de interés que escapan a los contenidos de este material. Dos de estos casos son los datos agrupados o combinados (*pooled data*) y los datos emparejados (*paired data*). En los datos agrupados, suponemos que $\sigma_X = \sigma_Y$. Los CI cambian y se obtienen resultados más precisos siempre que la hipótesis de partida sea cierta. En los datos emparejados, las muestras X e Y no provienen de poblaciones distintas, sino que se trata, por ejemplo, de evaluar, respecto a una misma población, cuándo se le aplica un tratamiento y cuándo no. Aquí, lo que se hace es restar los pares y trabajar como en el caso de una sola muestra.

2.3. Tests de hipótesis y valores p

Muy relacionados con los intervalos, encontramos los tests de hipótesis. En un test de hipótesis, tenemos dos hipótesis: la hipótesis nula, a la que llamamos H_0, y la hipótesis alternativa (lo que no es H_0), y el objetivo es comprobar si los datos de los que disponemos soportan (o hay evidencia estadística de ello) una u otra hipótesis. A continuación, describiremos los tests de hipótesis para el último caso que hemos visto, es decir, dos muestras de poblaciones independientes que modelamos con la t-Student. No obstante, tanto el concepto como el procedimiento se pueden extender fácilmente al resto de los casos. La hipótesis nula será en este caso $H_0 : \bar{X} - \bar{Y} = \Delta_0$. Por ejemplo, si queremos comprobar el efecto de un determinado tratamiento, que aplicamos a la población de X y no a la población de Y, fijaremos $\Delta_0 = 0$ y observaremos si hay evidencia estadística de que no se aprecia diferencia alguna debida a la aplicación del tratamiento en una de las poblaciones. Debemos definir también la hipótesis alternativa. Pueden ser tres, según lo que queramos: $H_a : \bar{X} - \bar{Y} \neq \Delta_0$ (decimos que se trata de un test *two-tailed* o bilateral), $H_a : \bar{X} - \bar{Y} > \Delta_0$ (*test upper-tailed* o lateral superior), o bien $H_a : \bar{X} - \bar{Y} < \Delta_0$ (test *lower-tailed* o lateral inferior).

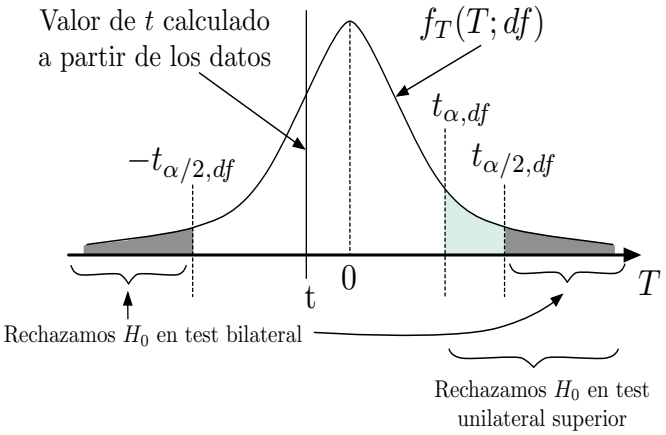
Lo primero que necesitamos es un estadístico de prueba (un valor calculado a partir de las muestras). Para el caso que nos ocupa, dicho estadístico sería:

$$t = \frac{(\hat{\bar{X}} - \hat{\bar{Y}}) - \Delta_0}{\sqrt{S_X^2/N + S_Y^2/M}} \tag{5}$$

Y coincide con la definición de T empleada para obtener (4), salvo que ahora fijamos la diferencia de medias (justamente lo que queremos comprobar). En otras palabras, seguimos trabajando con la distribución t-Student para dos conjuntos de muestras independientes, pero imponiendo ahora que la diferencia entre medias es justamente Δ_0. Podemos interpretar que buscamos si Δ_0 caería dentro del correspondiente CI de nivel de confianza $1 - \alpha$.

Empecemos por el caso $H_a : \bar{X} - \bar{Y} \neq \Delta_0$. Para quedar fuera del intervalo de confianza, es necesario o bien que $t \geq t_{\alpha/2,df}$, o bien que $t \leq -t_{\alpha/2,df}$. Si una de las dos cosas pasa, rechazamos la hipótesis H_0. En caso contrario, decimos que no podemos rechazar la hipótesis nula. Cuando la hipótesis alternativa es $H_a : \bar{X} - \bar{Y} > \Delta_0$, no nos preocupamos de la parte inferior del intervalo de confianza asociado, así que lo único que debemos comprobar es que no se sobrepase por arriba, o lo que es lo mismo, que $t \geq t_{\alpha,df}$. Observad ahora que pasamos de $\alpha/2$ a α, pues todo el error se concentra en una de las colas de la distribución. Por último, si $H_a : \bar{X} - \bar{Y} < \Delta_0$, entonces debemos comprobar si $t \leq -t_{\alpha,df}$. La figura 8 nos da una idea gráfica del funcionamiento de los tests de hipótesis.

Figura 8. Tests de hipótesis



Finalmente, relacionado con los test de hipótesis, tenemos los valores p (p-values). Consideremos primero un test lateral (*one-sided*). En este caso, el valor p nos da la probabilidad de obtener a partir de las muestras un valor tan extremo o más que el que hemos medido, y este es solo fruto de la casualidad. Por ejemplo, en el caso de la diferencia de medias de poblaciones independientes, imaginemos que obtenemos de las muestras $\hat{\bar{X}} - \hat{\bar{Y}} = 30$, y nuestra hipótesis de partida es $\Delta_0 = 10$. Entonces, suponiendo que realmente la media entre poblaciones es 10, ¿cuál es la probabilidad de haber obtenido una diferencia de medias de muestreo de 10, debido, digamos, a la mala suerte? Obviamente, dependerá también del tamaño de las muestras y de sus varianzas de muestreo. Para ello, evaluamos el valor de $t = \frac{(\hat{\bar{X}} - \hat{\bar{Y}}) - \Delta_0}{\sqrt{S_X^2/N + S_Y^2/M}}$, que nos fija un punto en la distribución de t-Student $f_T(T;\ df)$, y calculamos el área bajo la curva a partir de ese punto t (con tablas o usando algún software), es decir, $p = \int_t^\infty f_T(T;\ df)\, dT$. Es la misma idea que en los tests de hipótesis, pero ahora, en vez de buscar un valor $t_{\alpha,df}$ para un cierto nivel de confianza, lo que hacemos es buscar el valor de α que hace $t = t_{\alpha,df}$. Por lo tanto, dado el valor p, podemos hacer un test de hipótesis para cualquier valor de α: si el valor p es menor que α, entonces rechazamos H_0. Por ejemplo, si $t = 2{,}4$ y tenemos un valor de $df = 10$ en nuestras muestras, calcularíamos el valor de p en R ejecutando `pt(2.4, 10, lower.tail = FALSE)=0.01865`. De este modo, si la media real entre poblaciones es 10 y nosotros hemos medido 30, la probabilidad de haber obtenido este resultado fruto de la casualidad es 0.01865.

Para tests laterales inferiores la idea es la misma, pero ahora hay que obtener el área de la cola izquierda de la distribución. En tests bilaterales, se entiende que buscamos un valor tan extremo o más que el valor t calculado, pero entendido en magnitud. Es decir, si para un caso particular obtenemos $t = 2{,}4$, entonces debemos mirar el área bajo $f_T(T;\ df)$ tanto por encima de 2.4 como por debajo de -2.4. Por este motivo, en tests bilaterales se dobla el valor p respecto al test unilateral.

En general, los valores p son muy utilizados para dar el grado de certeza de un test. Si el valor está por debajo de 0.05, empieza a ser fiable y nos da aún más seguridad estadística si es inferior a 0.001. En lenguaje R, tenemos la función `t.test` para llevar a cabo tests de hipótesis basados en la distribución *t*-Student.

Bibliografía

Casella, G.; Berger, R. (2001). *Statistical Inference*. (2.a ed.). Duxbury Resource Center. ISBN 0-534-24312-6.

Devore, J. L. (2011). *Probability and Statistics for Engineering and the Sciences*. 8a ed.). Brooks/Cole. ISBN-13: 978-0-538-73352-6.

Rosner, B. (2011). *Fundamentals of Biostatistics*. 7th ed.). Brooks/Cole. ISBN-13: 978-0-538-73349-6.