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

Sesión 5. Introducción a la Inferencia Estadística.

Fernando San Segundo

Curso 2021-22. Última actualización: 2021-09-18

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI
DEPARTAMENTO DE MATEMÁTICA APLICADA

1. El Teorema Central del Límite.
2. Intervalos de confianza para la media.
3. Intervalos de confianza para la varianza.
4. Evaluación de la normalidad.

Sección 1

El Teorema Central del Límite.

Medias muestrales.

- En temas anteriores hemos visto de manera informal y mediante simulaciones que la distribución muestral de la media producía una curva normal. Ahora que sabemos más sobre la normal vamos a expresar ese resultado de forma más precisa y lo usaremos para empezar a hacer Inferencia.
- Queremos estudiar la distribución de una variable aleatoria cuantitativa X definida en los individuos de cierta población. En particular, la variable X tendrá una media μ y una varianza σ^2.
- Vamos a **estimar** el valor de μ usando **muestras** de la población. Si tenemos una **muestra aleatoria simple** formada por n valores como $x_1, x_2, \ldots, x_n$ (elegidos al azar y con remplazamiento) podemos usar la media muestral

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

para estimar la media poblacional μ.

El espacio muestral.

- Es el conjunto de todas las muestras aleatorias simples posibles de tamaño n que llamaremos Ω^n. Como ya vimos, al pasar de la población original al espacio muestral en general estamos pasando a un espacio muchísimo más grande.
- **Ejemplo:** si tenemos una población de tamaño 1000, ¿cuántas muestras aleatorias simples de tamaño 7 podemos construir? Es fácil ver que son

$$1000^7 = 1000000000000000000000$$

  muestras distintas.
- Entre todas esas muestras hay *muestras buenas* (en las que $\bar{x} \approx \mu$) y *muestras malas*, con un valor de $\bar{x}$ poco representativo. Si elegimos la muestra al azar, *¿cómo de probable es que nos toque una muestra buena?*
- Para responder necesitamos información sobre la distribución de los valores de $\bar{X}$ entre todas las muestras posibles (en Ω^n).

Distribución muestral de la media: teorema central del límite (TCL).

- Sea X una v.a. con media μ_X y varianza σ^2. Sea $\bar{X}$ la media muestral construida a partir de una muestra aleatoria simple $X_1, X_2, \cdots, X_n$ de tamaño n. Es decir:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

donde las X_i son *copias independientes entre sí de* X.

Teorema Central del Límite.
Cuando consideramos valores **suficientemente grandes** del tamaño muestral n, la distribución de la media muestral en el espacio muestral Ω^n se aproxima a una variable normal, cuya media y varianza son:

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma}{\sqrt{n}}\right)$$

- ¿Cuánto es *suficientemente grande*? Depende de la población inicial. Por ejemplo, si la población es normal, n puede ser arbitrariamente pequeño (incluso $n = 1$). Pero si la población es, por ejemplo, muy asimétrica, entonces puede que necesitemos n bastante grande.

Sección 2

Intervalos de confianza para la media.

Estimación en forma de intervalo.

- Empezamos pensando en el caso más sencillo: suponemos que la variable X es (aproximadamente) normal, pero desconocemos su media μ y queremos estimarla usando muestras.
- Este caso es bastante frecuente porque hay muchas magnitudes en la naturaleza cuya distribución es (aproximadamente) normal.
- Si X es normal el TCL es válido para cualquier tamaño muestral n. Podemos tomar una muestra aleatoria simple y usar la estimación $\mu \approx \bar{X}$. Naturalmente esto significa;

$$\mu = \bar{X} + \text{error}$$

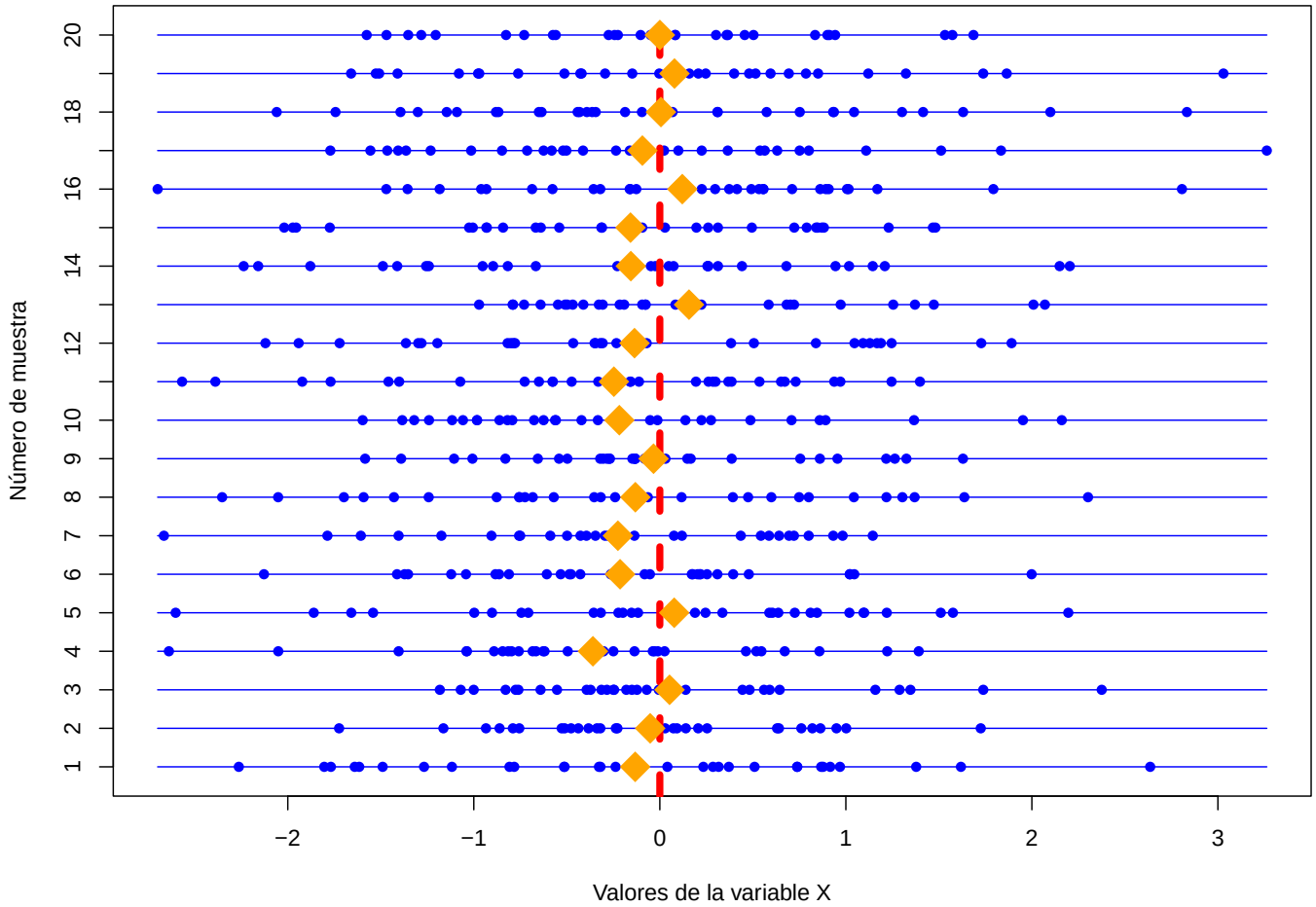
  Es muy importante entender que **el error es aleatorio**.
- Para que esto tenga alguna utilidad científica es imprescindible cuantificar ese error. Si descubrimos que el tamaño del error es menor que δ (piensa en un número pequeño) entonces podremos decir que:

$$\bar{X} - \delta < \mu < \bar{X} + \delta$$

  y nuestra estimación de μ será **en forma de intervalo** $(a, b) = (\bar{X} - \delta, \bar{X} + \delta)$. Como veremos el TCL nos ayuda a (obtener δ y) construir esos intervalos.

El error es aleatorio porque la muestra es aleatoria.

- En esta figura (mira el código que la ha generado) hemos obtenido 20 muestras de tamaño $n = 30$. La marca roja indica la media de la población, que es $\mu = 0$. Los puntos de cada muestra (puntos azules) están todos a la misma altura y se señala la media de esa muestra con un rombo naranja. Como ves, el error es aleatorio. Recuerda que en un caso real no sabemos donde está la línea roja.



Intervalos de confianza para la media.

- Si nos toca una muestra *"buena"* el error será pequeño, pero si damos con una muestra *"mala"* puede ser bastante grande. El TCL garantiza que cuando n aumenta las muestras buenas son mucho más abundantes que las malas.
- Recuerda que el muestreo es aleatorio: podemos *hacerlo todo bien* y obtener una estimación errónea por azar. Buscamos garantizar que es *poco probable* que nos pase eso. Por eso los intervalos de estimación que construimos tienen forma probabilística:

Intervalos de confianza.
Dado un **nivel de confianza** nc, un intervalo (a, b) tal que

$$P(a < \mu < b) = nc$$

es un **intervalo de confianza al nivel** nc para la media μ.

La probabilidad aquí se mide **sobre el conjunto (normalmente enorme) de todas las muestras aleatorias simples** de tamaño n y nc, el **nivel de confianza**, es la probabilidad de que nos toque una muestra *"buena"*. Siempre tomará valores cercanos a uno, como 0.90, 0.95 o 0.99.

Comentarios sobre la definición de intervalo de confianza.

- *La probabilidad nc no se refiere a un intervalo concreto sino al método de construcción de intervalos a partir de muestras*. Se puede entender así:

  (Si estimas μ usando este método) hay una probabilidad del 95 % de que (te toque una muestra buena y) μ esté dentro del intervalo (a, b).

  Las partes entre paréntesis suelen omitirse pero están implícitas.

- En particular los valores de a y b son aleatorios y **dependen de la muestra que nos toque**.

- Es importante además entender que en la construcción del intervalo entran en juego dos fuentes distintas de incertidumbre:

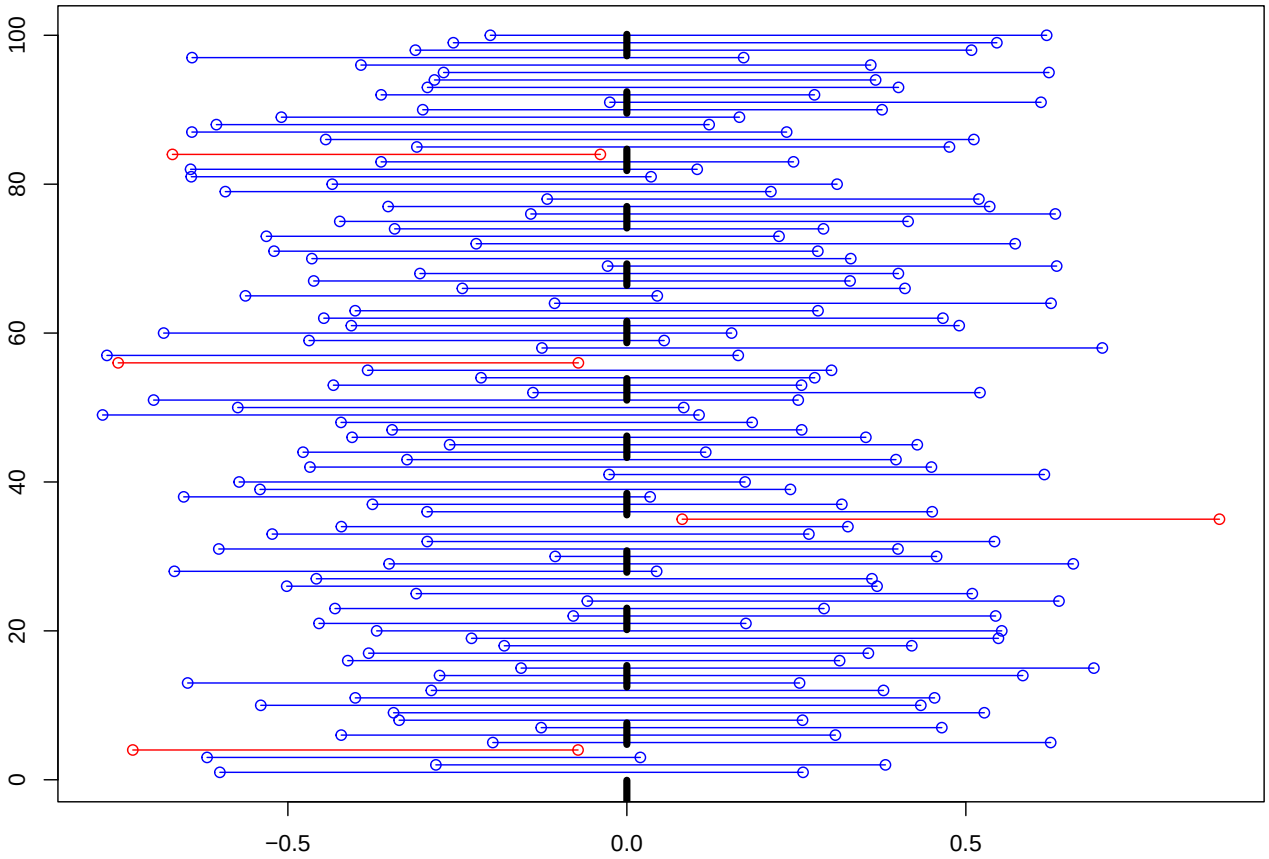
  (1) La **anchura** del intervalo (a, b) mide la **precisión** (o el error) con la que estimamos el valor de μ. Cuanto más estrecho sea el intervalo, mejor.

  (2) pero el nivel de confianza *nc* mide la **probabilidad muestral** de esa estimación, que depende de que hayamos tenido suerte con la muestra. Cuanto más cerca de 1 esté *nc*, mejor.

  Pero la precisión y la incertidumbre no son independientes, y en la práctica es necesario establecer un equilibrio entre las dos.

Interpretación probabilística de los intervalos de confianza.

- La construcción del intervalo parte de una muestra aleatoria y ya que hay muestras buenas y malas, **a veces el intervalo puede errar por completo** y μ no pertenece a ese intervalo. Eso no significa que hayamos hecho nada mal, hemos tenido mala suerte. La figura (¡ver código!) ilustra esto con 100 intervalos a partir de sendas muestras.



La línea de puntos indica la media poblacional real

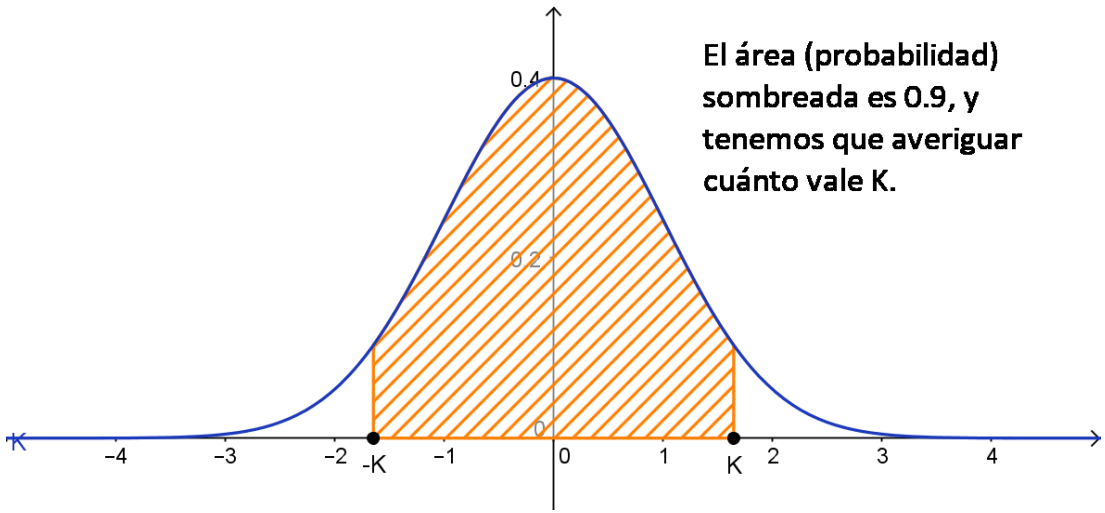
El papel del TCL en la construcción de intervalos de confianza.

- Para una población normal el TCL garantiza que

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma}{\sqrt{n}}\right)$$

Eso significa que $Z = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ es una normal estándar $N(0, 1)$.

- Además, dado un nivel de confianza nc como 0.9 sabemos construir un intervalo simétrico $(-K, K)$ tal que $P(-K < Z < K) = nc$ como en la figura:



Sustituyendo la anterior expresión de Z aquí y despejando μ obtenemos la fórmula del intervalo de confianza.

Fórmula preliminar del intervalo del confianza.

- Pero antes vamos a darle un nombre a K. La zona sombreada de la anterior figura tiene probabilidad nc. Queda una probabilidad

$$\alpha = 1 - nc$$

para repartir *entre las dos colas*. Así, *cada una de las dos colas* que son iguales por simetría tiene una probabilidad igual a $\frac{\alpha}{2}$.

- Dada una probabilidad p, el **valor crítico** z_p es el valor de la normal estándar que deja **a su derecha** esa probabilidad p. Es decir, $P(Z > z_p) = p$. Y por tanto, $K = z_{\alpha/2}$.
- Una *versión preliminar* de la fórmula del intervalo de confianza es:

Un intervalo de confianza (a, b) al nivel nc es:

$$a = \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad b = \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Que se resume así:

$$\mu = \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

¿Por qué preliminar? Fíjate en que aquí aparece σ, que es desconocido.

La aproximación de las muestras grandes.

- ¿Y si no conocemos σ entonces qué hacemos? Hay un remedio sencillo **siempre que la variable X sea normal en la población y además la muestra sea suficientemente grande**.
- En esos casos podemos cambiar σ por *desviación típica muestral* s en la primera fórmula utilizable del intervalo.

Intervalo de confianza al nivel *nc*, población normal y muestra grande.

$$\mu = \bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

- ¿Qué es una muestra grande? $n = 30$ puede servir, pero recomendamos $n > 100$.
- **Ejemplo:** una muestra de una población normal tiene estos *valores muestrales*:

$$n = 100, \qquad \bar{X} = 7.34, \qquad s = 0.31$$

Sea $nc = 0.95$ (luego $\alpha = 0.05$). Sabiendo que $z_{\alpha/2} \approx 1.96$ el intervalo de confianza al 95 % que se obtiene es:

$$\mu = \bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} \approx 7.34 + 1.96 \frac{0.31}{\sqrt{100}} = (7.279, 7.401).$$

¿Cómo hemos llegado a ese valor de $z_{\alpha/2} \approx 1.96$?

Valores críticos e intervalos de confianza con R.

- El cálculo de $z_{\alpha/2}$ para cualquier α (y cualquier *nc*) se realiza en R con `qnorm`. ¡Pero cuidado!, por defecto R trabaja con la cola izquierda.
- Usando por ejemplo el nivel de confianza $nc = 0.95$ calculemos el correspondiente valor crítico $z_{0.025}$, que guardaremos en la variable `zc`:

```
nc = 0.95
alfa = 1 - nc
(zc = qnorm(alfa / 2, lower.tail = FALSE)) # Atención, cola derecha
```

```
## [1] 1.959964
```

- A partir de aquí obtener el intervalo partiendo de los valores muestrales es muy fácil:

```
## Intervalos de confianza con R.
n = 100
barX = 7.34
s = 0.31
(intervalo = barX + c(-1, 1) * zc * s / sqrt(n))
```

```
## [1] 7.279241 7.400759
```

- Partiendo de un fichero csv con la muestra, como 05-IntervConfNormalGrande.csv: (*a*) Leemos los datos con `read.table`. (*b*) Calculamos *n*, $\bar{X}$ y *s* con `length`, `mean`, `sd`, respectivamente. (*c*) Procedemos como antes.
- **Ejercicio:** con los datos de ese fichero calcula un intervalo de confianza para la media.

Cálculo del tamaño muestral necesario.

- En la primera fórmula vimos que la **semianchura del intervalo** es $\delta = z_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}}$. Esta cantidad es la que define la **precisión** del intervalo. Para conseguir una precisión δ dada, por ejemplo 0.0001, podemos tratar de despejar en esta fórmula n,el tamaño muestral necesario:

$$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \delta \qquad \Rightarrow \qquad n = \left(z_{\alpha/2} \cdot \frac{\sigma}{\delta} \right)^2$$

  Pero de nuevo, desconocemos σ. La solución es hacer un *estudio piloto* con una muestra pequeña para estimar con s la desviación típica σ.

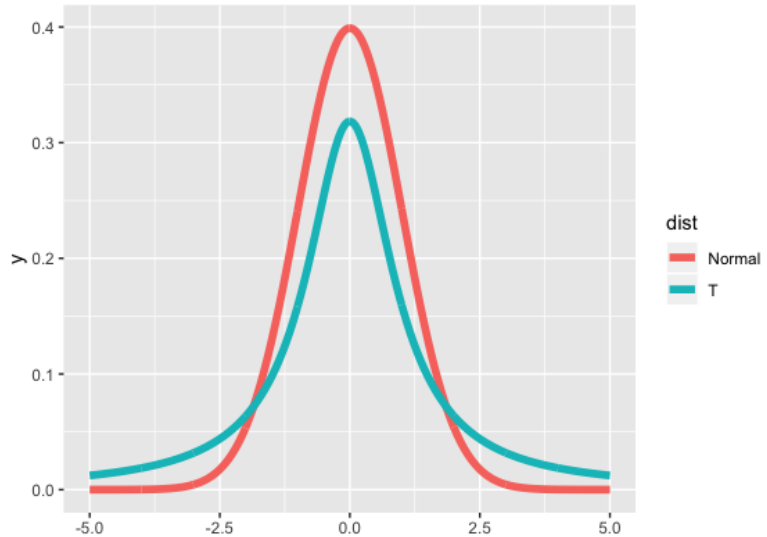
- **Ejemplo.** *Una empresa produce unas piezas y desea estimar su diámetro medio (que sigue una distribución normal). Una muestra piloto tuvo una desviación típica* $s = 1.3mm$. *La empresa quiere una medida del diámetro con un error no mayor de* $0.1mm$ *y un nivel de confianza del* 99 %. *¿Qué tamaño de muestra debe utilizarse para conseguir ese objetivo?*
  Se desea una precisión $\delta = 0.1$mm. Al ser $nc = 0.99$, tenemos $\frac{\alpha}{2} = 0.005$, y $z_{\alpha/2} = z_{0.1} \approx 2.58$. Sustituyendo

$$n = \left(z_{\alpha/2} \cdot \frac{\sigma_X}{\delta} \right)^2 \approx \left(2.58 \cdot \frac{1.3}{0.1} \right)^2 \approx 1121.3$$

  Usaríamos una muestra de tamaño 1122 *al menos* (conviene ser precavidos y redondear al alza).

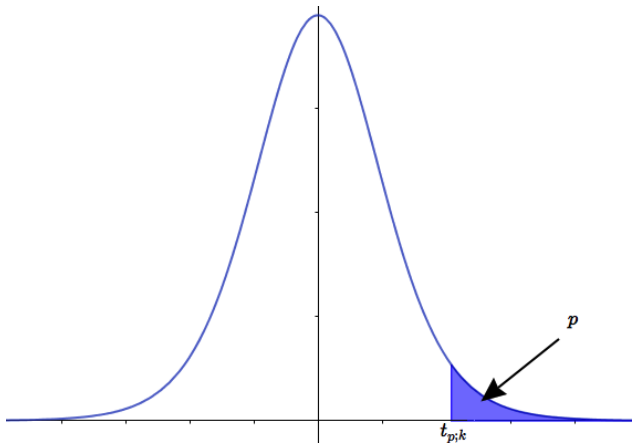
Muestras pequeñas en poblaciones normales.

- Los resultados anteriores sirven *para poblaciones normales y muestras grandes*. ¿Qué sucede si sabemos que **la variable X tiene una distribución normal** en la población, pero sólo disponemos de una **muestra pequeña** (con $n < 30$)?
- Si la muestra es pequeña disponemos de menos información sobre la variable X. Eso debe traducirse, necesariamente, en un intervalo de confianza más ancho. Student (que en realidad se llamaba William S. Gosset) se dio cuenta de que en este tipo de problemas no se podía usar Z directamente y descubrió un sustituto, la distribución t de Student.
- Esa distribución tiene las *colas más pesadas* (con más probabilidad) que Z. En realidad hay una t distinta para cada tamaño muestral. La siguiente figura compara Z con la distribución t con $df = 2$ (muestras de tamaño 3).



Intervalos de confianza usando la t de Student.

- **Grados de libertad:** Sea X una variable normal en la población y supongamos que el tamaño n de la muestra es pequeño. Diremos que $k = n - 1$ son los grados de libertad (en inglés, *degrees of freedom*) de esa muestra.
- **Valores críticos de t:**si T es una variable t de Student con k grados de libertad, el valor $t_{k;p}$ verifica $P(T_k > t_{k;p}) = p$ (su cola derecha tiene probabilidad p).



- Con esta terminología podemos dar la fórmula para el intervalo de confianza para μ usando t:

Intervalo de confianza al nivel nc, población normal, muestra pequeña.

$$\mu = \bar{X} \pm t_{k;\alpha/2} \frac{s}{\sqrt{n}}$$

La distribución t en R.

- La función `pt` es análoga a `pnorm` y sirve para el *cálculo directo de probabilidad*. Por ejemplo, para calcular $P(T_{17} > 2.5)$ (que es una cola derecha) usaríamos:

```
1 - pt(2.5, df = 17)
```

```
## [1] 0.0114739
```

  Fíjate en que se indican los grados de libertad con `df` (degrees of freedom).

- `qt`, como `qnorm`, hace cálculos inversos de probabilidad; dada una probabilidad buscamos *el valor* que deja esa probabilidad en su cola izquierda o derecha. Por ejemplo, para calcular el valor crítico `tc` para un nivel de confianza `nc` cualquiera haríamos:

```
n = 20
nc = 0.95
alfa = 1 - nc
df = n - 1
(tc = qt(alfa / 2, df, lower.tail = FALSE)) # Atención, cola derecha
```

```
## [1] 2.093024
```

- La función `rt` sirve para simular valores aleatorios de una variable t de Student.

```
rt(8, df = 19)
```

```
## [1] -0.5787343 -0.3383073 -0.2600134 -0.8701595  0.4070822
## [6]  0.2702574 -0.3835929  0.7091232
```

Ejemplo de cálculo de intervalo de confianza con la t de Student.

- **Ejemplo:** *Se sospecha que en las aguas de un embalse las concentraciones de nitritos superan el umbral tolerable por los peces, que es de 0.03 mg NO2/l o menos. Para verificar esta sospecha se midieron los niveles de nitritos en diez puntos aleatorios del embalse, obteniendo estos valores:*

  `0.04, 0.05, 0.03, 0.06, 0.04, 0.06, 0.07, 0.03, 0.06, 0.02`

  Calculemos un intervalo de confianza al 95 % para el nivel medio de nitritos en las aguas del embalse.

```
datos = c(0.04, 0.05, 0.03, 0.06, 0.04, 0.06, 0.07, 0.03, 0.06, 0.02)
n = length(datos)
barX = mean(datos)
s = sd(datos)
nc = 0.95
alfa = 1 - nc
tc = qt(1 - alfa/2, df = n - 1)
(intervalo = barX + c(-1, 1) * tc * s / sqrt(n))
```

```
## [1] 0.03422133 0.05777867
```

¿Cuál es la conclusión?

Resumen de intervalos de confianza para la media μ.

- **Variable X normal y muestra grande ($n > 100$)**:

$$\mu = \bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

  En raras ocasiones usaremos aquí σ en lugar de s.

- **Variable X normal pero muestra pequeña**:

$$\mu = \bar{X} \pm t_{\alpha/2;k} \frac{s}{\sqrt{n}}$$

  con $k = n - 1$, los grados de libertad.

- **Variable X *aproximadamente normal* y muestra grande:**

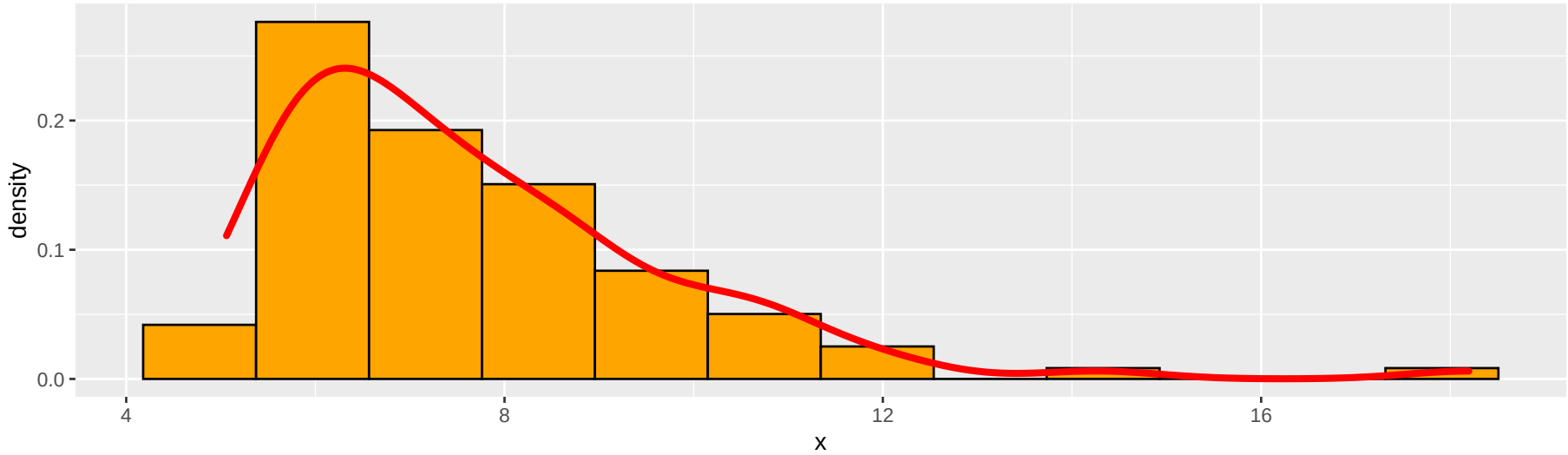
  El TCL permite usar la fórmula previa con t para el intervalo de confianza. Enseguida discutiremos que significa ser aproximadamente normal.

- **Variable posiblemente no normal:**

  En este caso los métodos que hemos visto no sirven para obtener un intervalo de confianza para la media.

Intervalos de confianza por bootstrap.

- Muchos métodos de la Estadística clásica (intervalos de confianza, contrastes de hipótesis) asumen que las variables son al menos aproximadamente normales. Entre otras cosas, eso implica que los intervalos de confianza para la media son simétricos respecto a la media muestral. Pero a menudo encontramos muestras muy asimétricas, que no justifican la simetría del intervalo.
- El aumento de la capacidad de cómputo ha propiciado el desarrollo de **métodos no paramétricos** para los intervalos de confianza basados en el **remuestreo**, como el **bootstrap**. Vamos a usar ese método para obtener un intervalo de confianza de los datos contenidos en el fichero skewdata.csv (basado en un ejemplo de (Crawley 2005, pág. 47)). La figura ilustra la asimetría de esos datos:



Esquema del método.

- Empezamos leyendo esos datos (fíjate en que usamos la url directamente):

```
url =
  "https://raw.githubusercontent.com/mbdfmad/fmad2122/main/data/skewdata.csv"
x = pull(read_csv(file = url), 1)
```

  Ejercicio: ¿Qué hace la función `pull` en este código? ¿Es necesaria, hay otras formas de hacerlo?

- Ahora vamos a explorar los tamaños muestrales entre $n = 5$ y $n = 40$:

  (a) Para cada tamaño construiremos 10000 remuestreos aleatorios con remplazamiento de esa muestra.

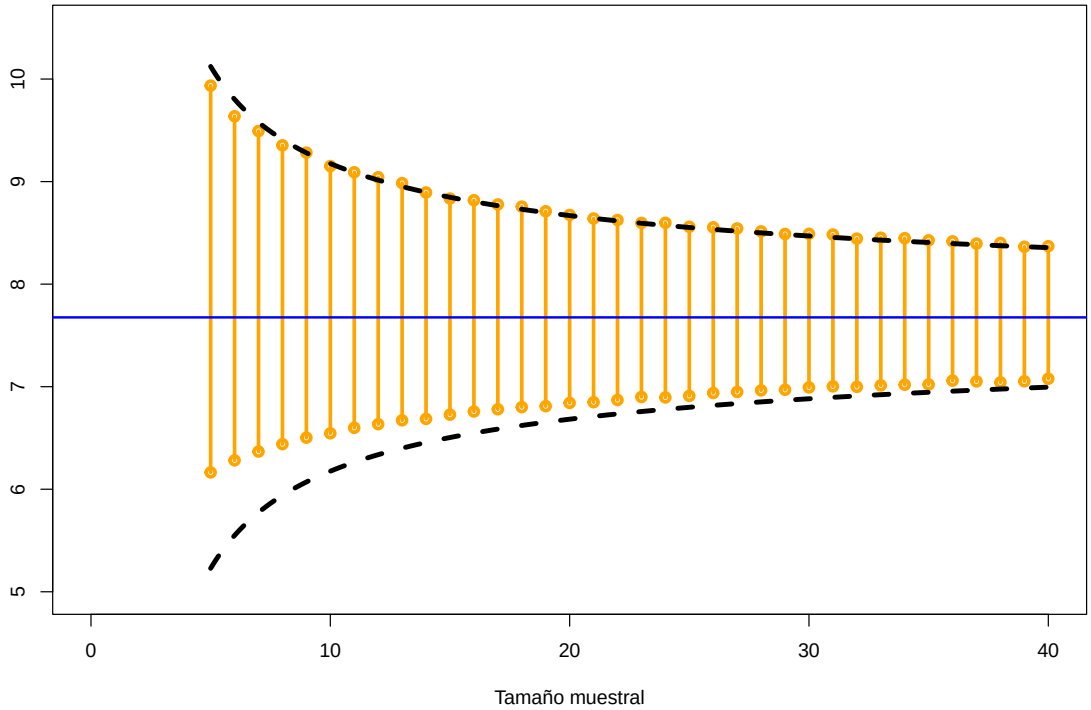
  (b) En cada remuestreo calculamos la media obteniendo así 10000 medias muestrales.

  (c) Dibujamos el intervalo que va del primer al tercer cuartil de esas 10000 medias (todas de muestras de tamaño n).

- El código R correspondiente a este esquema es un ejemplo muy sencillo de uso de los bucles `for` que ya conocemos.

Representación gráfica de los intervalos bootstrap.

- En la gráfica el eje horizontal es el tamaño de la muestra y el vertical los valores de X. La media de X se indica con una línea horizontal azul.
- Los intervalos bootstrap se muestran como segmentos verticales en naranja, la media en azul las líneas de trazos negras representan los intervalos *clásicos* usando la t de Student. Fíjate en que para muestras grandes no hay apenas diferencia. Pero en muestras pequeñas el intervalo bootstrap refleja mucho mejor la asimetría de los datos y, en particular, los intervalos no son simétricos respecto a la media.



Código R del bootstrap.

- También puedes verlo (y copiarlo) en el fichero Rmarkdown de la sesión.

```
# Creamos la "caja" del gráfico.
plot(c(0, 40), c(5,10.5), type="n", xlab="Tamaño muestral", ylab="")

for (k in seq(5, 40, 1)){ # Este bucle recorre los tamaños muestrales
  a =  numeric(10000) # el vector a almacenará las medias muestrales
  for (i in 1:10000){ # este es el bucle de remuestreo (bootstrap)
  # generamos un remuestreo con reemp. y calculamos su media
    a[i] = mean(sample(x, k, replace=T))
    }
  # dibujo del intervalo bootstrap de este tamaño muestral
  points(c(k,k), quantile(a, c(.025,.975)), type="o",
         col = "orange", lwd= 3)
}

# el siguiente bloque de código genera una banda con
# los intervalos clásicos correspondientes a esas muestras.
xv = seq(5, 40, 0.1)
yv = mean(x) - qt(0.975, xv) * sqrt(var(x) / xv)
lines(xv, yv, lty = 2, col = "black", lwd = 4)
yv = mean(x) + qt(.975, xv) * sqrt(var(x) / xv)
lines(xv, yv, lty = 2, col = "black", lwd = 4)

# añadimos una línea horizontal en la media
abline(h = mean(x), col="blue", lwd=2)
```

Sección 3

Intervalos de confianza para la varianza.

Distribución muestral de s^2 y la distribución χ^2 (chi cuadrado).

- Después de μ, lo natural es calcular intervalos de confianza para σ^2.
- Sea X de tipo $N(\mu, \sigma)$. Lo idea natural es aproximar σ^2 mediante s^2. Para que la idea necesitamos algo como el TCL: información que relacione σ^2 con la distribución de s^2 en el conjunto de todas las n-muestras posibles (espacio muestral).
- Importante: la media es una medida central y por eso era interesante analizar la **diferencia** $\mu - barX$. Pero la varianza es una medida de dispersión y por eso los **cocientes** son más útiles que las diferencias.
- El resultado que necesitamos es este:

Distribución muestral de σ^2 en poblaciones normales.
Si X es una variable aleatoria de tipo $N(\mu; \sigma)$, y se utilizan muestras aleatorias de tamaño n, entonces:

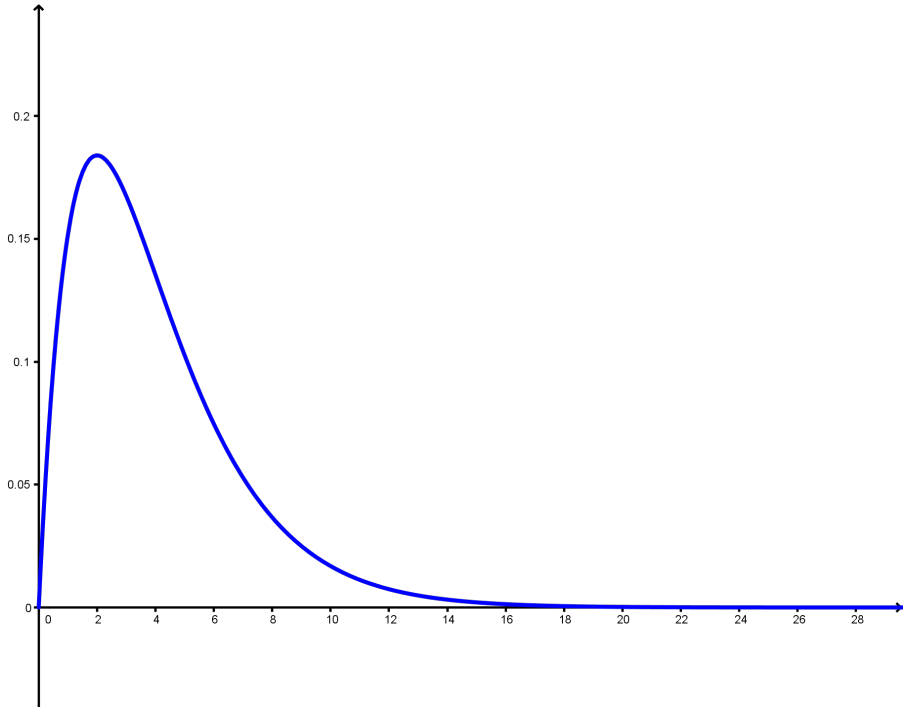
$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}$$

siendo χ^2_{n-1} la **distribución chi cuadrado con $n-1$ grados de libertad,**

Veamos como es esa distribución χ^2_{n-1}.

La distribución χ^2_k y funciones de R.

- Esta distribución *sólo toma valores positivos* y además es *asimétrica*, a diferencia de la Z o la t de Student. Por ejemplo, la distribución χ^2_4 tiene este aspecto:

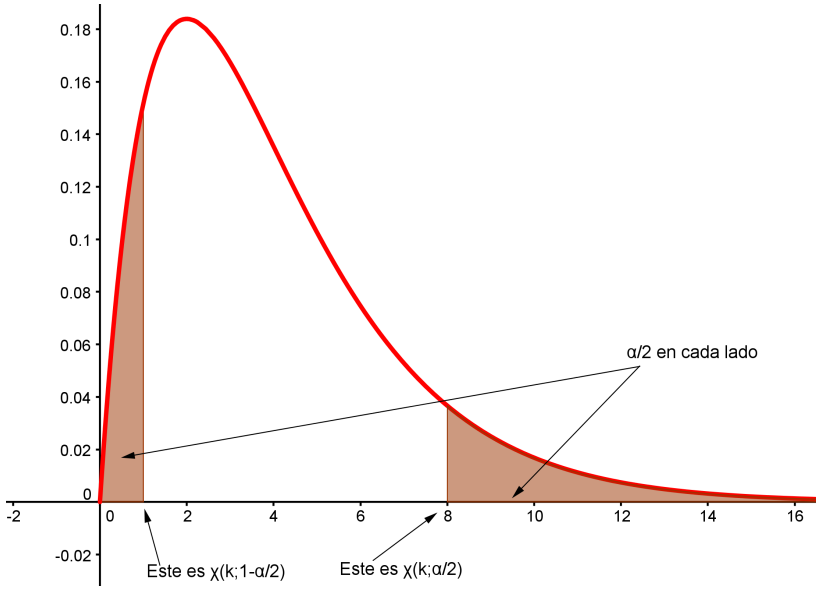


  La asimetría, como veremos, afecta al proceso de construcción de intervalos de confianza basados en esta distribución.

- En R disponemos de las funciones `pchisq`, `qchisq` y `rchisq` con los significados previsibles.

Intervalos de confianza para la varianza.

- La novedad en este caso es que por la asimetría de χ^2_k hay que usar valores críticos distintos a derecha e izquierda. Cada uno de ellos deja una probabilidad $\alpha/2$ en la cola correspondiente.



donde si $Y = \chi^2_k$ se cumple $P(Y > \chi^2_{k,p}) = p$.

Intervalo de confianza para σ^2 en poblaciones normales.

$$\frac{(n-1)s^2}{\chi^2_{k,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{k,1-\alpha/2}}, \qquad \text{con } k = n-1$$

Construcción con R de intervalos de confianza para la varianza.

- **Ejemplo:** *La variable aleatoria X tiene una distribución normal. Una muestra aleatoria de 7 valores de X dio como resultado* $s^2 = 62$. *Vamos a construir con R un intervalo de confianza (nc = 95 %) para* σ^2.

```
# Estos son los valores muestrales y el nc deseado
varianza = 62 # cuidado si el dato muestral es s y no s^2
n = 7
nc = 0.95
(alfa = 1 - nc)
```

```
## [1] 0.05
```

```
# Calculamos dos valores críticos de chi cuadrado.
(chi1 = qchisq(alfa / 2, df = n - 1, lower.tail = FALSE)) # cola derecha
```

```
## [1] 14.44938
```

```
(chi2 = qchisq(alfa/2, df = n - 1)) # cola izquierda
```

```
## [1] 1.237344
```

```
# Construimos el intervalo
(intervalo = (n - 1) * varianza / c(chi1, chi2))
```

```
## [1]  25.74506 300.64390
```

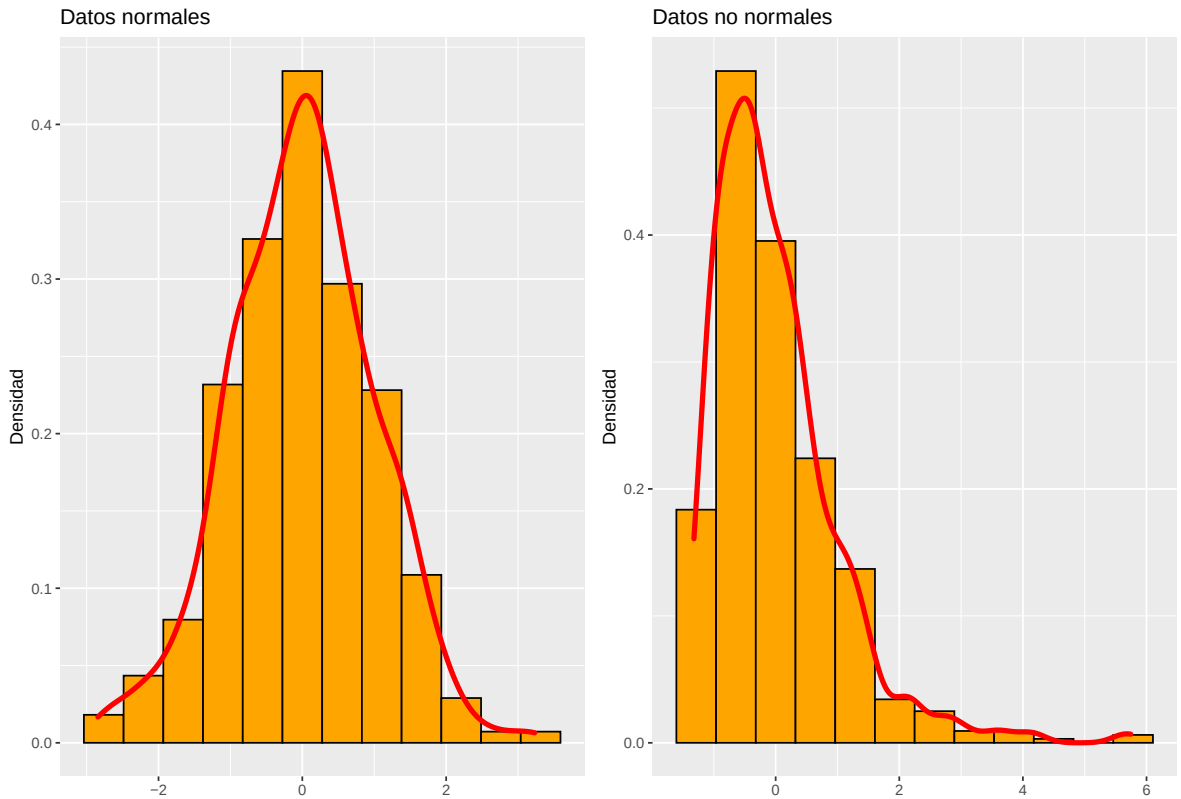
Fíjate en que el valor crítico de cola derecha se usa en el extremo izquierdo del intervalo y viceversa. Y si queremos un intervalo para σ simplemente calculamos la raíz cuadrada. `sqrt(intervalo)` produce el intervalo (5.074, 17.34) para σ.

Sección 4

Evaluación de la normalidad.

¿Cómo podemos analizar la normalidad de una población?

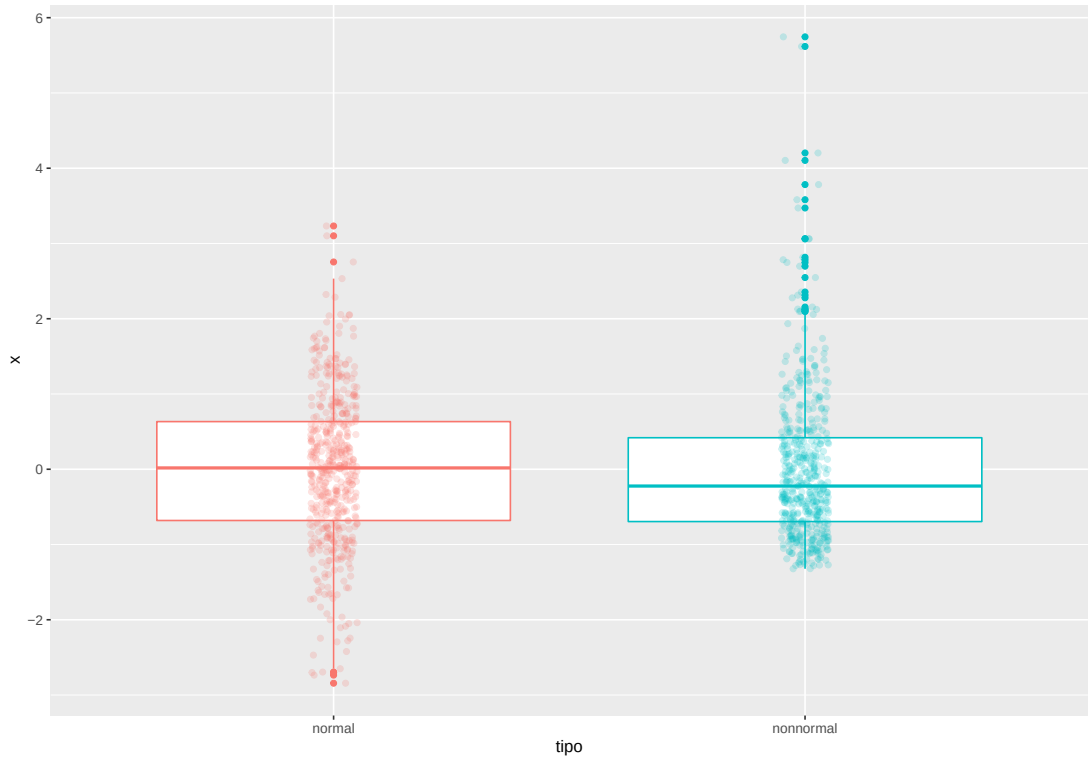
- Los métodos de los apartados anteriores requieren evaluar si la variable de interés es (al menos aproximadamente) normal. En muestras grandes examinaremos *histogramas* y *curvas de densidad*. La figura muestra a la izquierda una muestra de datos normales y a la derecha datos no normales, con $n = 500$ en ambos casos. Con muestras más pequeñas las cosas pueden estar menos claras.



En esta y en las siguientes páginas, mira el código de este tema.

Boxplots para analizar la simetría.

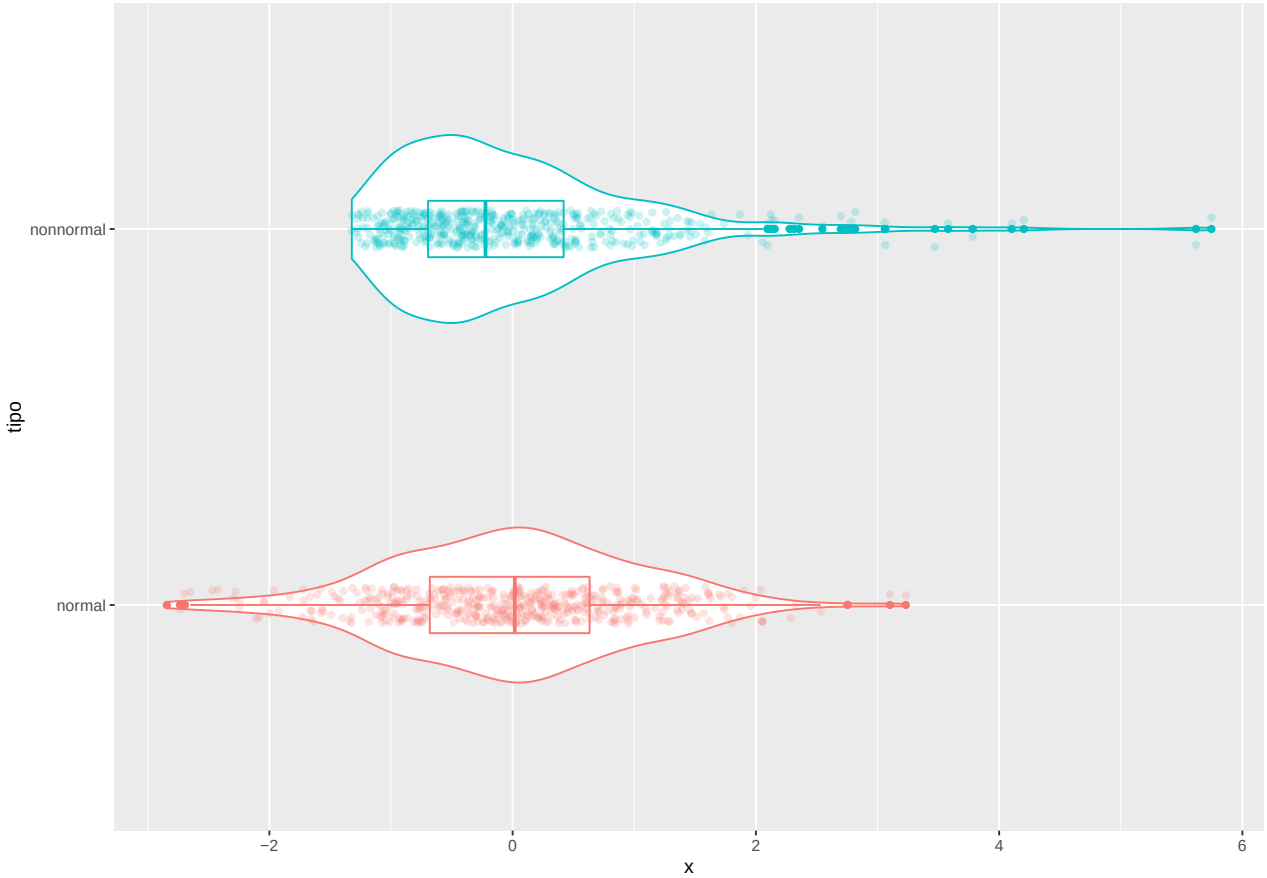
- A menudo la simetría es el requisito más importante para que los métodos de la Estadística (basados en el TCL) funcionen. Los boxplots son especialmente útiles para detectar la falta de simetría. Para las mismas dos muestras de antes:



Mira el código para ver cómo usamos `jitter` y `alpha` para evitar superposición de puntos y añadir transparencia.

Violinplot.

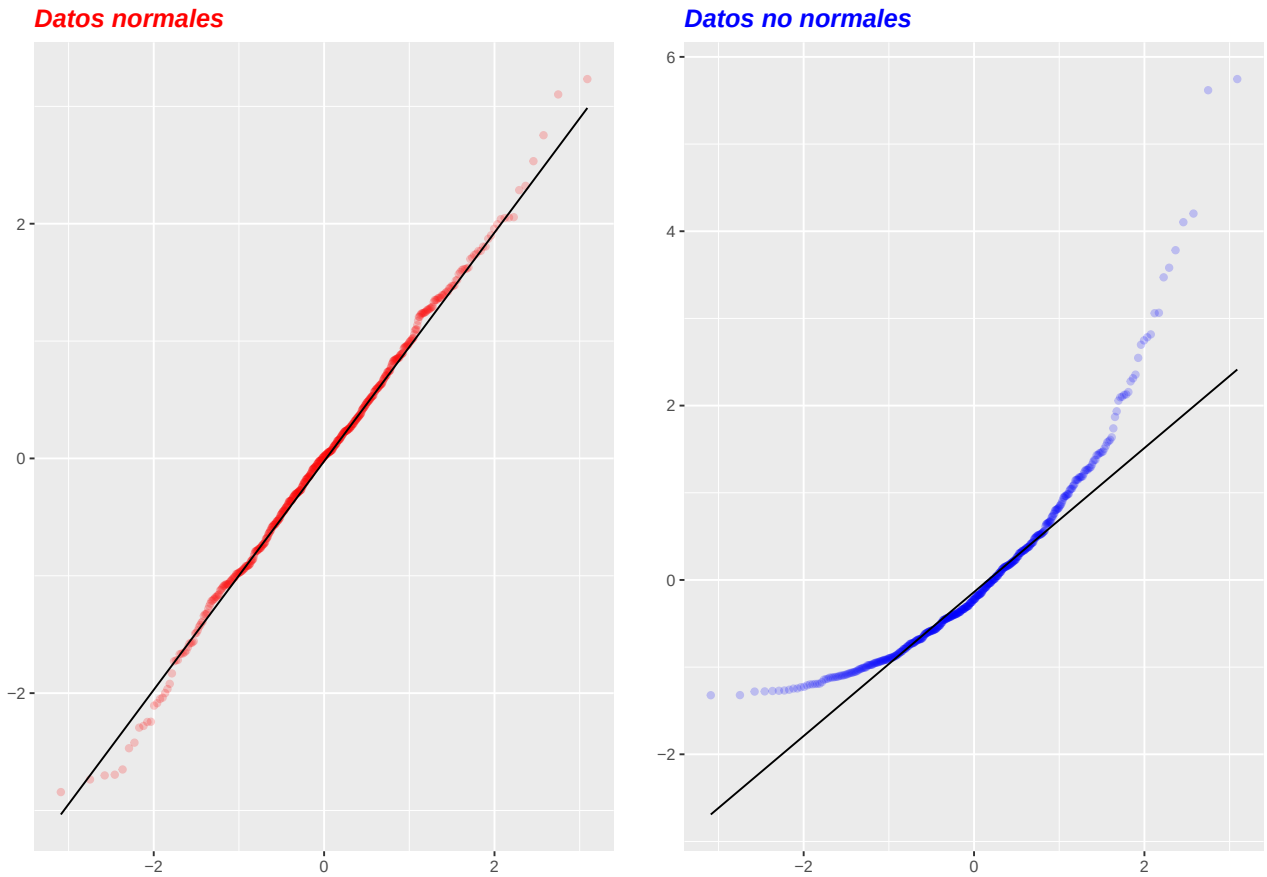
- Este tipo de gráfico son interesantes para combinar la curva de densidad con el boxplot. Y de nuevo, es posible, añadir los puntos de la muestra:



Mira el código para ver cómo hemos situado los gráficos en horizontal, para resaltar la forma de las curvas de densidad.

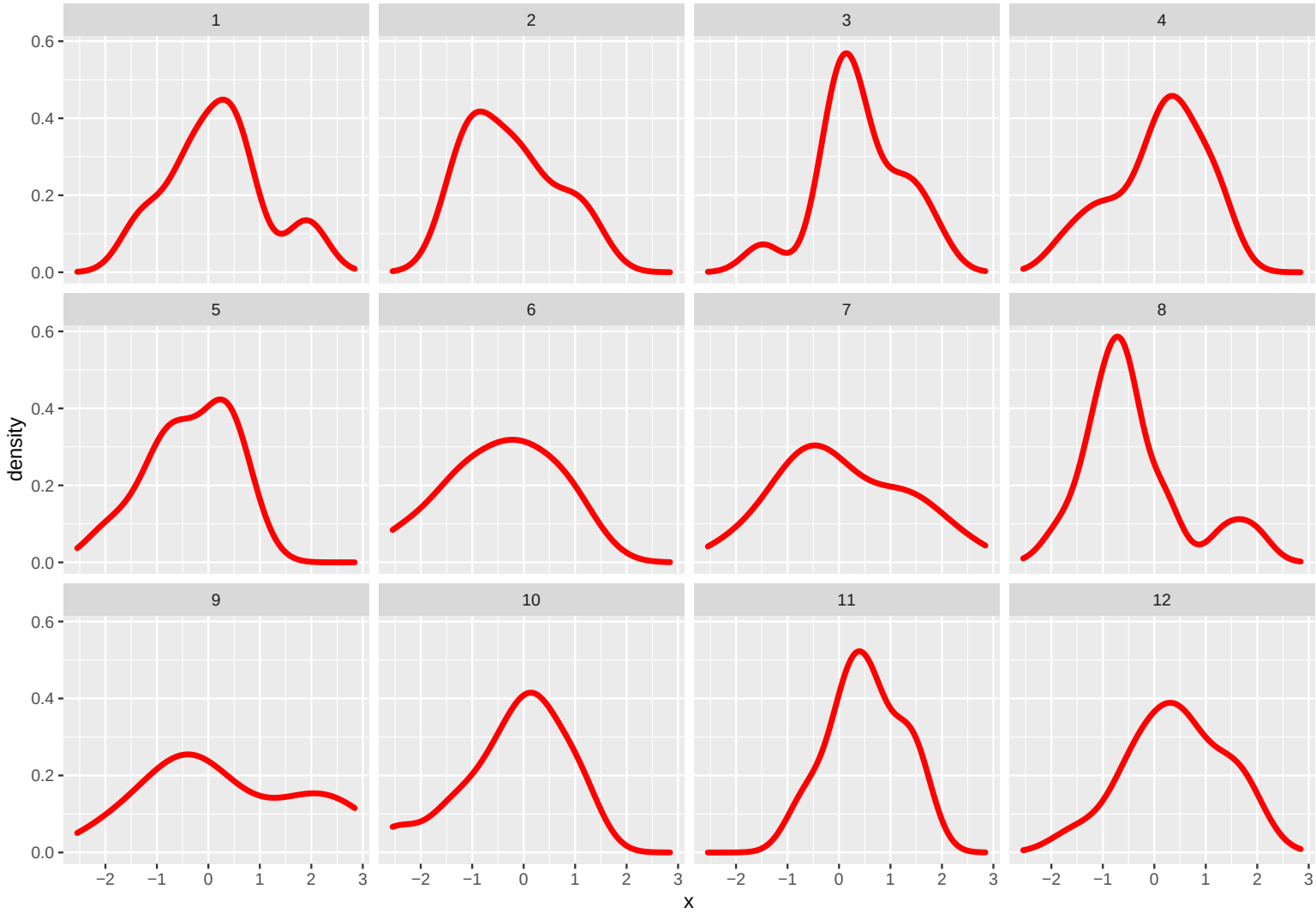
QQplots.

- El nombre proviene de *"quantile vs quantile"*, porque se representa en el eje horizontal los percentiles de una variable normal exacta y en el vertical los de la muestra a examen. Son el tipo de gráficos más utilizado para analizar la normalidad. Si la muestra procede de una variable normal, los puntos deben coincidir con la recta.



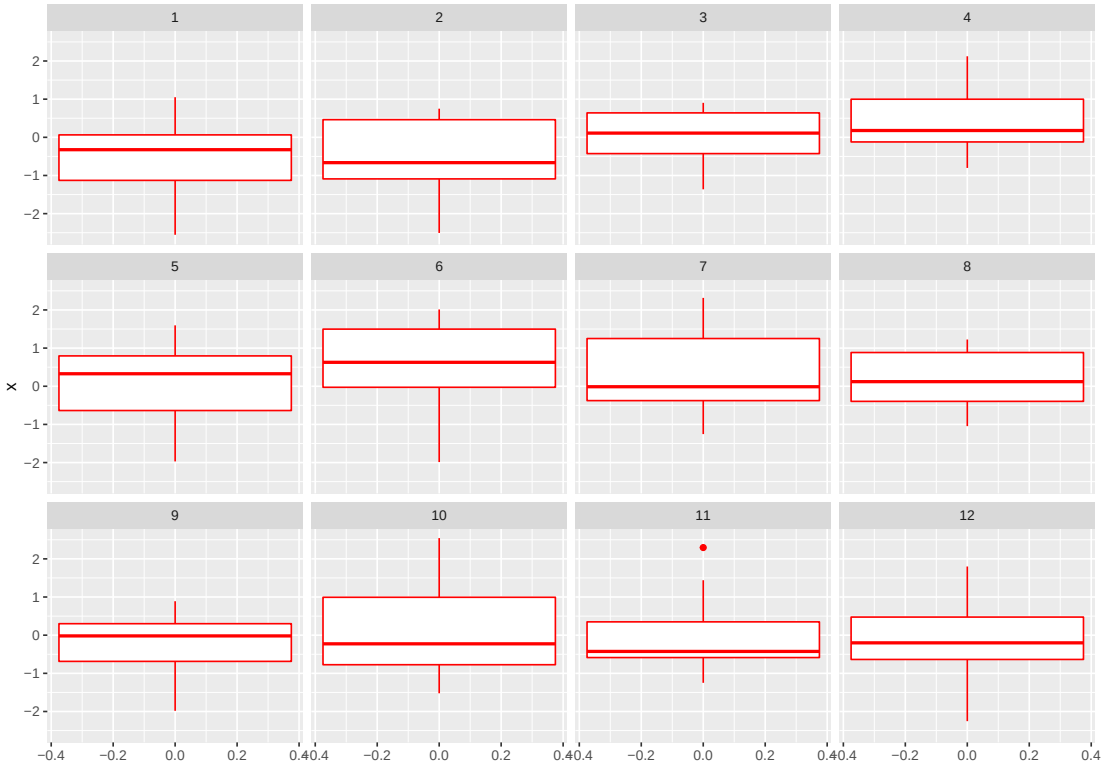
¡¡Precaución con las muestras pequeñas!!

- Los métodos que hemos descrito funcionan bien con muestras grandes. Para muestras pequeñas, las cosas se complican. Todas las figuras son curvas de densidad de muestras de tamaño 15 que **provienen de la misma población normal**.



Con los boxplots sucede algo parecido.

- Todos estos boxplots son también de muestras con $n = 15$ que **provienen de la misma población normal**.



La disparidad de formas y la variabilidad en la simetría en estas dos últimas páginas deben servir de advertencia: evaluar la normalidad en muestras pequeñas es complicado. Existen criterios formales para hacer esa evaluación, pero para muestras pequeñas esos criterios pueden proporcionar una falsa sensación de rigor con poca base real.

Referencias para la sesión

Enlaces

- Código de esta sesión

Bibliografía

Crawley, M. J. (2005). *Statistics: an introduction using R. 327 p.* John Wiley Sons.