

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

Tema 5: Introducción a la Inferencia Estadística.

Fernando San Segundo

Curso 2019-20. Última actualización: 2019-09-19



Sección 1

El Teorema Central del Límite.

- 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, \dots, x_n (elegidos al azar y con remplazamiento) podemos usar la media muestral

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

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, \dots, X_n de tamaño n . Es decir:

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

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**.

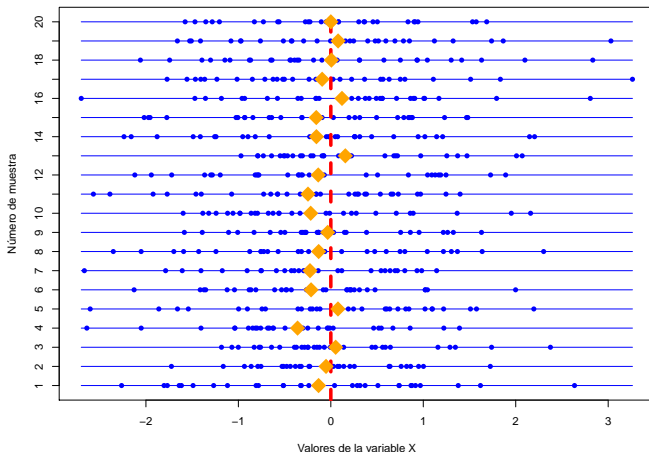
- Para que esto tenga alguna utilidad científica es imprescindible cuantificar ese error. Si descubrimos que el tamaño del error es menor que δ (piensa en un número pequeño) entonces podremos decir que:

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

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 con una “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 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.

- 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:

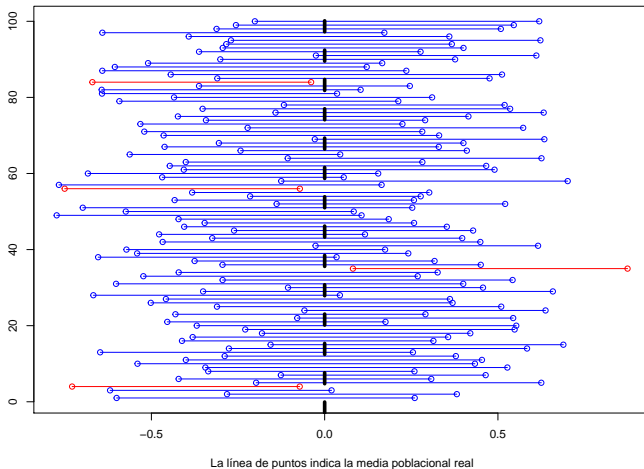
(1) La **anchura** del intervalo (a, b) mide la **precisión** (o el error) con la que estimamos el valor de μ . Cuanto más estrecho sea el intervalo, mejor.

(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.



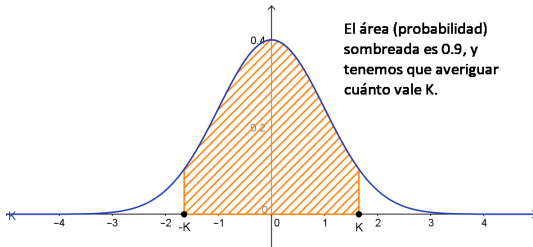
El papel del TCL en la construcción de intervalos de confianza.

- Para una población normal el TCL garantiza que

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

Eso significa que $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ 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:

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

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*:

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

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:

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

¿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:

```
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 [06-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:

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

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 $s = 1.3\text{mm}$. 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?*

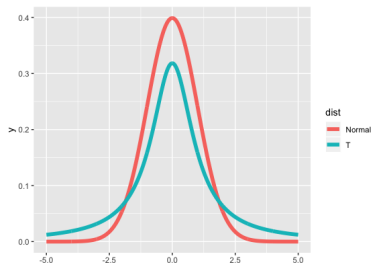
Se desea una precisión $\delta = 0.1\text{mm}$. Al ser $nc = 0.99$, tenemos $\frac{\alpha}{2} = 0.005$, y $z_{\alpha/2} = z_{0.1} \approx 2.58$. Sustituyendo

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

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

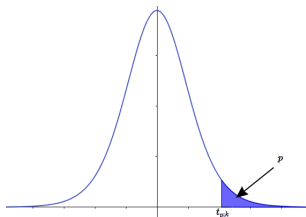
Muestras pequeñas en poblaciones normales.

- 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. El código de este tema usa la librería `manipulate` de RStudio para explorar como cambia t con n .



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 t_c 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] 0.4187084 0.7392641 1.2306202 -1.7145000 -0.6126917
```

```
## [6] 1.1330675 -0.7748450 0.3466095
```

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:**

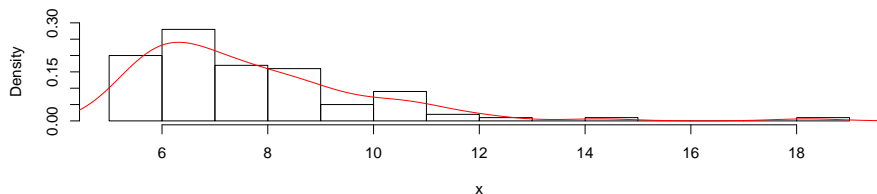
El TCL permite usar la fórmula previa con t para el intervalo de confianza. Enseguida discutiremos que significa ser aproximadamente normal.

- **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 directamente la URL):

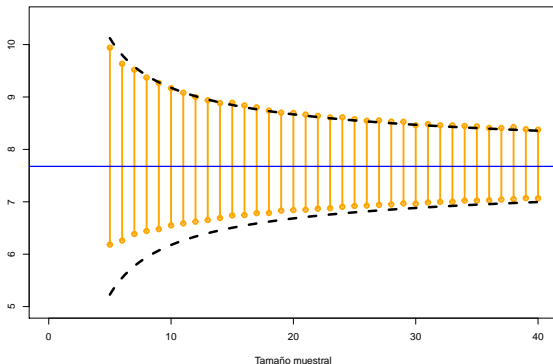
```
x = read.table(file = "http://www.postdata-statistics.com/docs/skewdata.csv",  
              header = TRUE)[, 1]
```

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 muy sencillo y lo usaremos para introducir el uso de los bucles `for`. Al final de este tema hay una pequeña introducción a los bucles y estructuras de control en R.

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 y en rojo representamos 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.



Código R del bootstrap.

```
# 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** $\mu - \text{bar}X$. 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:

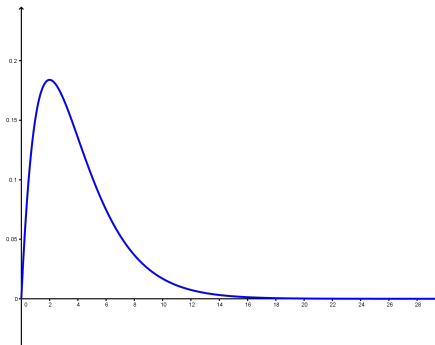
$$(n-1) \frac{s^2}{\sigma^2} \sim \chi_{n-1}^2$$

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

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

La distribución χ_k^2 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_4^2 tiene este aspecto:

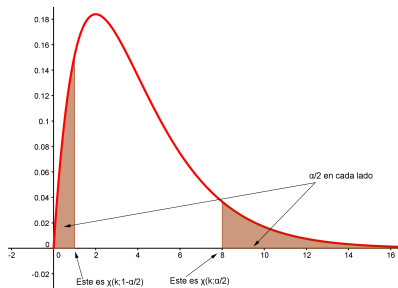


La asimetría, como veremos, afecta al proceso de construcción de intervalos de confianza basados en esta distribución.

- En R disponemos de las funciones `pchisq`, `qchisq` y `rchisq` con los significados previsibles. El código de este tema usa `manipulate` para explorar esta distribución.

Intervalos de confianza para la varianza.

- La novedad en este caso es que por la asimetría de χ_k^2 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 $Y = \chi_k^2$ se cumple $P(Y > \chi_{k,p}^2) = p$.

Intervalo de confianza para σ^2 en poblaciones normales.

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

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
```

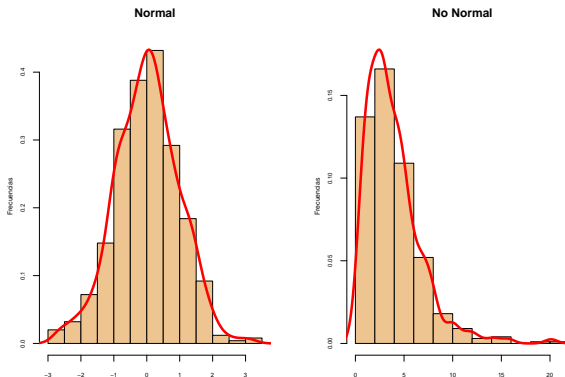
Fíjate en que el valor crítico de cola derecha se usa en el extremo izquierdo del intervalo y viceversa. Y si queremos un intervalo para σ simplemente calculamos la raíz cuadrada. `sqrt(intervalo)` produce el intervalo (5.074, 17.34) para σ .

Sección 4

Evaluación de la normalidad.

¿Cómo podemos analizar la normalidad de una población?

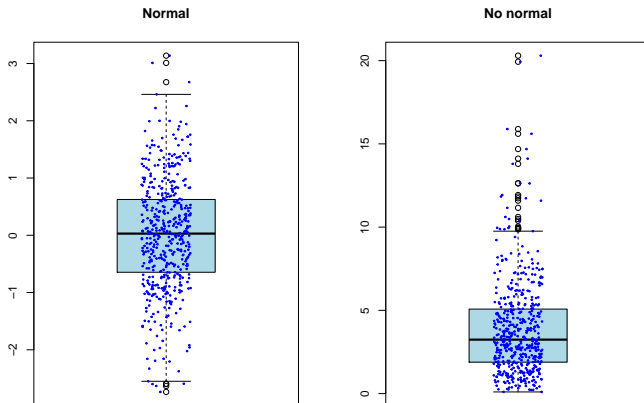
- Los métodos de los apartados anteriores requieren evaluar si la variable de interés es (al menos aproximadamente) normal. En muestras grandes examinaremos *histogramas* y *curvas de densidad*. La figura muestra a la izquierda una muestra de datos normales y a la derecha datos no normales, con $n =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

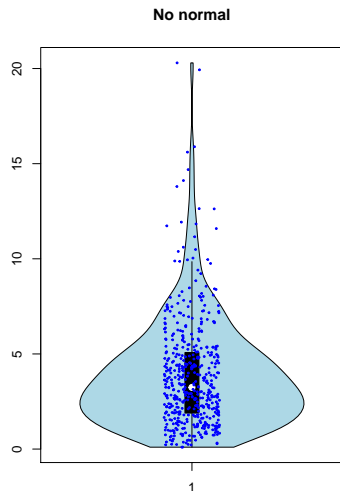
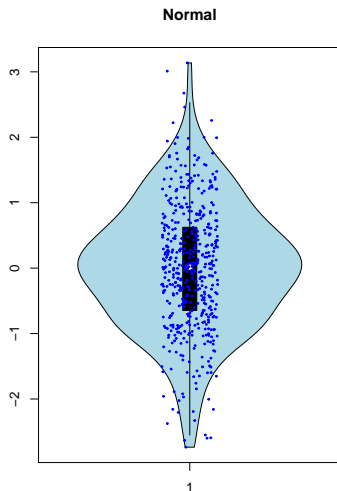
Boxplots para analizar la simetría.

- 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 (¡puede ser buena idea añadir los puntos de la muestra!).



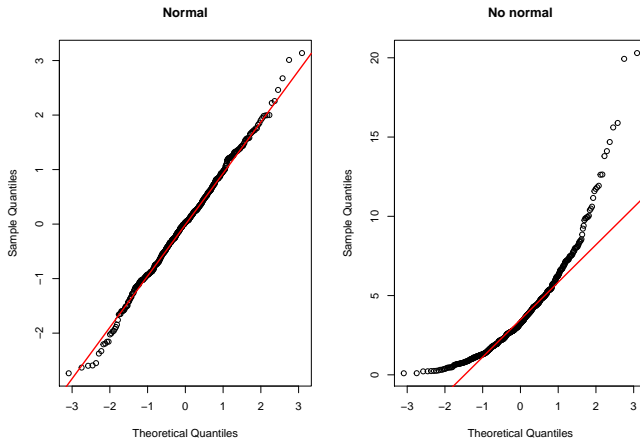
Violinplot.

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



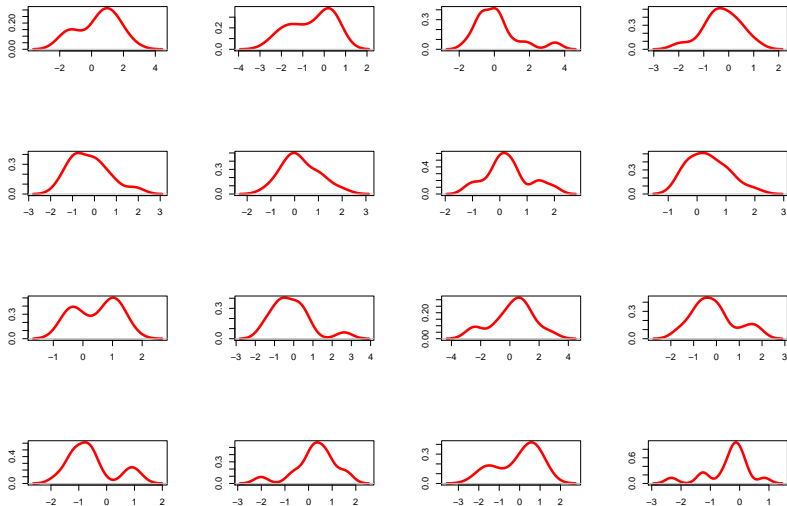
QQplots.

- 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.



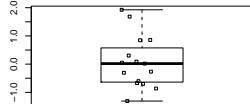
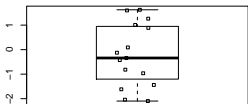
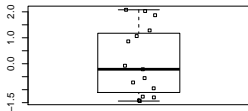
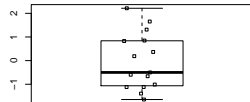
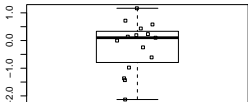
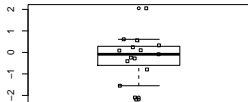
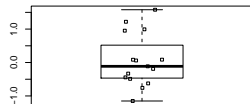
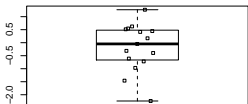
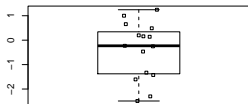
¡¡Precaución con las muestras pequeñas!!

- 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 **proviene de poblaciones normales**.



Con los boxplots sucede algo parecido.

- Todos estos boxplots son de muestras normales con $n = 15$.



Sección 5

Contrastes de Hipótesis.

Ejemplo inicial: una discusión científica.

- Hemos desarrollado un nuevo fármaco, *Pildorín Complex*, para tratar la depresión severa en el Canguro Rojo australiano. Pensamos que es tan bueno que, después de administrárselo, los pacientes darán saltos de alegría. De hecho, afirmamos que “*la altura (media) de esos saltos será mayor de lo que era antes del tratamiento*”. Esta es nuestra **hipótesis**.
- Para obtener datos relacionados con nuestra afirmación, hemos tomado una muestra de $n = 100$ canguros depresivos, a los que administramos el medicamento. Y nos ponemos muy contentos, porque la altura media de sus saltos, después de usar *Pildorín*, es mayor que antes de tratarlos.
- Concretamente, un experto en canguros nos dice que la altura (en metros) de los saltos **se distribuye como una normal**, con media

$$\mu_0 = 2.5$$

(en metros). Pero en la muestra de 100 canguros depresivos tratados con *Pildorín Complex* hemos observado una altura de salto media (muestral)

$$\bar{X} = 2.65$$

(en metros), con desviación típica muestral $s = 0.5$.

- Pero ese no es el final de la historia. . .

Hipótesis nula y alternativa.

- El laboratorio de la competencia, que lleva años vendiendo su medicamento *Saltaplus Forte*, dirá que nuestro medicamento tiene **efecto nulo** y que los saltos que hemos observado en nuestros canguros depresivos son, simplemente, sus saltos habituales, que los canguros a veces saltan más y a veces menos, y que nuestras medidas son simplemente **fruto del azar**. Tenemos así dos afirmaciones o hipótesis enfrentadas.
- La hipótesis de la competencia, que llamaremos **hipótesis nula** H_0 (porque dice que el efecto es nulo), sostiene que la media no ha aumentado con el tratamiento.
- Nuestra hipótesis, que dice que la media sí ha aumentado. A esta la llamaremos **hipótesis alternativa** H_a .
- Un **contraste de hipótesis** puede entenderse como la forma científica de resolver esta discusión, usando los datos y la teoría sobre Probabilidad que hemos aprendido.

- Vamos a usar la siguiente notación, y es **muy importante** entenderla bien desde el principio:
- Llamaremos siempre μ a la **media real** de la población de la que hemos tomado la muestra (en el ejemplo, los canguros tratados). Ni los defensores de H_0 ni los de H_a conocen (ni es posible que conozcan) este valor.
- Además en la discusión ha aparecido un **valor de referencia** $\mu_0 = 2.5$, que compararemos con μ mediante muestras. Este valor se utiliza para formular claramente las dos hipótesis contrapuestas.
- Los dos valores μ y μ_0 son *valores teóricos*, no observados. Por último, tenemos el valor de la media muestral, \bar{X} , que es un *valor empírico* y procede de las observaciones. Pero es el valor fundamental para decidir a cuál de las dos hipótesis damos más credibilidad.
- Además es importante entender que ambas partes aceptan el valor de \bar{X} ; ese valor no se discute (sería otra discusión). Recuerda que usamos \bar{X} para **estimar** μ . Así que si \bar{X} es muy grande, ¿a quién parecen darle la razón los datos?

Formalizando el contraste.

- La utilidad de la notación es que podemos usarla para escribir las dos hipótesis con más precisión:

(a) La **hipótesis alternativa** H_a sostiene que la media de la población (recuerda, la población es *tratada*) es mayor que el valor de referencia.

$$H_a = \{\mu > \mu_0\}$$

(b) La **hipótesis nula** H_0 dice justo lo contrario: que la media de la población (recuerda, la población es *tratada*) es menor o igual que el valor de referencia.

$$H_0 = \{\mu \leq \mu_0\}$$

Fíjate en que ponemos el igual en H_0 porque si la media es igual, seguirá teniendo razón en que no ha habido **efecto** del tratamiento. Hay autores que siempre usan $=$ en la hipótesis nula y ponen $H_0 = \{\mu = \mu_0\}$.

- En el ejemplo:** Recuerda que era $\mu_0 = 2.5$. Por lo tanto la hipótesis alternativa es

$$H_a = \{\mu > 2.5\}$$

mientras que la nula es:

$$H_0 = \{\mu > 2.5\} \quad (\text{o bien } H_0 = \{\mu = 2.5\})$$

¡Atención! es un error incluir la media muestral $\bar{X} = 2.65$ en las hipótesis.

La idea clave.

- El punto de partida es este: dado que la muestra procede de la población a examen, debe ser $\bar{X} \approx \mu$ y, por lo tanto, si \bar{X} es mayor que μ_0 , eso parece darle la razón a H_a .
- Pero recuerda que hay “*muestras malas*”. Así que el partidario de H_0 dirá ese valor de \bar{X} se debe a que **por azar** nos ha tocado una muestra mala. Naturalmente, cuanto más grande sea el valor de \bar{X} , **menos probable** es que nos haya tocado por azar **una muestra así de mala**.
- **En el ejemplo:** el partidario de $H_0 = \{\mu \leq 2.5\}$ puede entonces decir que el valor $\bar{X} = 2.65$ se debe al azar y a una muestra desafortunada. Pero si el valor muestral hubiera sido $\bar{X} = 5$, ese argumento de que la muestra es mala pierde mucho peso porque **es muy poco probable que nos toque una muestra tan mala**.
- Para hacer de esto una discusión precisa: ¿podemos calcular esa probabilidad? Es decir, ¿podemos calcular la probabilidad de muestras tan malas como esa?

Teorema Central del Límite y p-valor.

- Lo que vamos a hacer es esto: supondremos, provisionalmente, que H_0 es cierta. De hecho, admitiremos como correcto el valor de μ que más le conviene al partidario de H_0 (luego volvemos sobre esto). Ese valor es:

$$\mu = \mu_0$$

- Y ahora usaremos esa suposición provisional para calcular la probabilidad de una muestra *tan mala o peor para H_0* como la nuestra. La probabilidad que vamos a calcular es el **p-valor** del contraste de hipótesis.
- Al suponer (provisionalmente) que $\mu = \mu_0$, podemos usar el TCL para decir que la distribución de la media muestral es:

$$\bar{X} \sim N\left(\frac{\mu_0}{\frac{s}{\sqrt{n}}}\right)$$

O, lo que es lo mismo, que

$$\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \sim Z$$

Y esto nos permite calcular la probabilidad que buscamos, el p-valor, usando la normal estándar.

Cálculo del p-valor en el ejemplo.

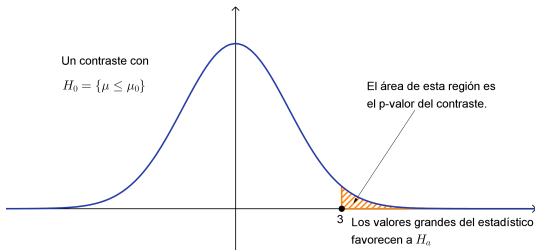
- Recuerda que teníamos:

$$\mu_0 = 2.5, \quad n = 100, \quad \bar{X} = 2.65, s = 0.5$$

Así que el valor de Z que obtenemos es:

$$\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{2.65 - 2.5}{\frac{0.5}{\sqrt{100}}} = 3$$

Y la siguiente figura ilustra que el p-valor es la probabilidad de la cola derecha de 3 en Z .



- En R, obtenemos `pValor = pnorm(3, lower.tail = FALSE) \approx 0.00135`

Interpretación del resultado.

- Lo que hemos hecho se puede resumir así: *suponiendo que la hipótesis nula fuera cierta* y usando $\mu = \mu_0$, la probabilidad de obtener un valor muestral tan grande o más que \bar{X} es de tan sólo 0.00135.
- El partidario de H_0 puede insistir en que es fruto del azar, pero ahora sabemos cuantificarlo. Para que el partidario de H_0 tenga razón nos debería haber tocado una muestra tan mala que sólo hay una así en cada mil.
- Imagínate que el valor de \bar{X} hubiera sido 2.7, más alejado de $\mu_0 = 2.5$ (y por tanto más favorable a H_a). Puedes comprobar que el correspondiente valor de Z sería

$$\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{2.7 - 2.5}{\frac{0.5}{\sqrt{100}}} = 4$$

y entonces el p-valor habría sido aún más pequeño: $1 - \text{pnorm}(4) \approx 3.167 \times 10^{-5}$. En ese caso al partidario de H_0 *le costaría mucho más hacernos creer que todo es fruto del azar*.

- En resumen, un p-valor pequeño le quita la razón al partidario de H_0 y nos llevaría a **rechazar la hipótesis nula**.

- La idea del p-valor es medir **cómo de extraños, inexplicables o sorprendentes le parecen los resultados de la muestra a alguien que cree en la hipótesis nula**. Simbólicamente:

$$\text{p-valor} = P(\text{datos} \mid H_0 \text{ es cierta})$$

- En muchos casos la hipótesis nula representa el conocimiento establecido o aceptado. Y por eso, en general, debemos estar muy convencidos antes de rechazar la hipótesis nula. Por eso la hipótesis nula *“juega con ventaja”*.
- Por eso el valor de $\mu = \mu_0$ es el más ventajoso para la hipótesis nula, Si al calcular el p-valor usáramos otro valor de μ menor que μ_0 el p-valor habría sido aún más pequeño. Así que usamos μ_0 para darle a H_0 todas las ventajas.
- Y por eso incluimos todos los valores de la cola derecha. Es la misma idea: si tomáramos sólo una parte de esa cola la probabilidad (el p-valor) sería aún menor, así que usamos toda la cola.
- Hemos usado el TCL para calcular el p-valor, pero también se puede hacer mediante remuestreo, como en el bootstrap. Esa es una opción muy interesante, que cada vez gana más peso en las aplicaciones. Mira el código de este tema.

Rechazando la hipótesis alternativa.

- Volviendo al ejemplo que venimos usando, supongamos que hubiéramos obtenido $\bar{X} = 2.51$, manteniendo todos los demás valores iguales. Entonces

$$\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{2.51 - 2.5}{\frac{0.5}{\sqrt{100}}} = 0.2$$

y el p-valor correspondiente sería: $1 - \text{pnorm}(0.2) \approx 0.4207$.

Para alguien que cree que la hipótesis nula es cierta eso significa que el valor de \bar{X} que hemos obtenido no es, en absoluto, una sorpresa (¡la probabilidad es el 42 %!). Así que no hay evidencia, usando estos datos, para rechazar H_0 y, en su lugar, rechazamos la hipótesis alternativa H_a .

- Hay otra situación que a veces causa confusión al principio. Desde luego, si hubiésemos obtenido una media muestral como $\bar{X} = 2.45$, que es menor que $\mu_0 = 2.5$ **no necesitaríamos siquiera calcular el p-valor para rechazar H_a** . Recuerda que creemos que $\mu \sim \bar{X}$ y, por tanto, tratar de convencer a alguien de que $\mu > 2.5$ enseñándole el valor $\bar{X} = 2.45$ es una pérdida de tiempo. Pero por supuesto puedes calcular el valor de Z y el p-valor, que son respectivamente:

$$\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{2.45 - 2.5}{\frac{0.5}{\sqrt{100}}} = -1, \quad \text{p-valor} \approx 0.8413$$

Regla de decisión y p-valor. Nivel de significación.

- Recuerda siempre que:
 - (a) con un **p-valor suficientemente pequeño rechazamos la hipótesis nula.**
 - (b) con un **p-valor grande rechazamos la hipótesis alternativa**
- ¿Qué es un p-valor pequeño? Para que la decisión sea más objetiva y simple se suele utilizar un umbral de corte predeterminado, llamado el **nivel de significación** ns . Entonces, los p-valores más pequeños que

$$\alpha = 1 - ns$$

se consideran suficientemente pequeños. Los valores más frecuentes de ns coinciden con los que usamos como nivel de confianza en los intervalos, y son 0.90, 0.95 y 0.99. ¡No los confundas, son cosas distintas! Los correspondientes valores de $\alpha = 1 - ns$ son 0.10, 0.05 y 0.01.

- Por lo tanto, si hacemos un contraste de hipótesis usando un nivel de significación del 95 % y obtenemos un p-valor = 0.004, puesto que $1 - ns = 0.05$, teniendo en cuenta que

$$\text{p-valor} = 0.004 < 0.05 = \alpha = 1 - ns$$

diremos que el p-valor es suficientemente pequeño y rechazamos H_0 . Si obtuviéramos, por ejemplo, un p-valor = $0.07 > 0.05 = \alpha$ rechazaríamos H_a .

- Al realizar un contraste de hipótesis podemos cometer dos tipos de errores **por la naturaleza aleatoria del proceso de muestreo**.

	¿Qué hipótesis es cierta?	
	H_a (alternativa) es cierta	H_0 (nula) es cierta
Rechazar H_0	Decisión correcta	Error tipo I (α)
Rechazar H_a	Error tipo II (β)	Decisión correcta

- Esta situación recuerda a la de las pruebas diagnósticas y, de hecho, ese lenguaje se aplica también aquí hasta cierto punto.
- En muchos casos los errores de tipo I se consideran los más importantes. ¿Cuál es la probabilidad de cometer un error de tipo I, cuando usamos un nivel de significación ns (y el correspondiente $\alpha = 1 - ns$)? Sería:

$$P(\text{rechazar } H_0 \mid H_0 \text{ es cierta})$$

Pero eso ocurre precisamente si la muestra que hemos usado es una de esas muestras malas cuyo p-valor es menor que α . Así que cuando pensamos en todas ellas vemos que **la probabilidad de cometer un error de tipo I usando ns es precisamente α** .

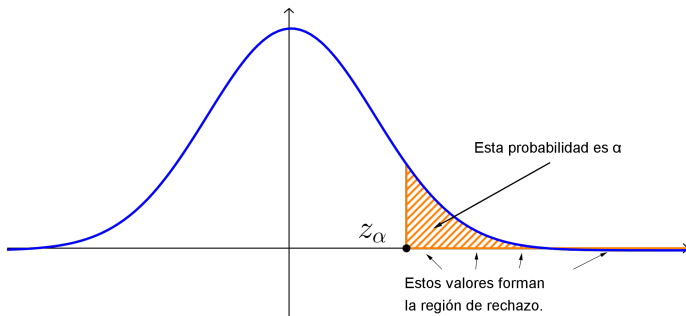
Estadístico del contraste. Región de rechazo.

- En el ejemplo que hemos utilizado hemos organizado el cálculo del p-valor y la decisión del contraste este diagrama:

$$H_0 \text{ y } H_a \longrightarrow \text{datos muestra } n, \bar{X}, s \longrightarrow \text{estadístico } Z = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \longrightarrow$$

p-valor usando $Z \longrightarrow \text{¿p-valor} < \alpha = 1 - ns? \longrightarrow \text{rechazar } H_0 \text{ o rechazar } H_a$

- Al hacer esto vemos que existe un valor z_α tal que si el *estadístico* Z calculado en la muestra cumple $Z > z_\alpha$, entonces rechazamos H_0 . Esos valores del estadístico forman la **región de rechazo** del contraste.



Ejemplo de región de rechazo y uso de los valores muestrales.

- En el caso del ejemplo que venimos usando, si queremos trabajar a un nivel de significación del 95 % entonces $\alpha = 0.05$ y el valor $z_\alpha = z_{0.05}$ es:

$$\text{qnorm}(1 - 0.05) \approx 1.645$$

Por lo tanto la región de rechazo la forman los valores de $Z > 1.645$.

- Pero también podemos expresar el valor Z a partir de los valores muestrales de ese ejemplo y escribir esa condición en términos de \bar{X} :

$$\frac{\bar{X} - 2.5}{\frac{0.5}{\sqrt{100}}} > 1.645$$

Despejando de aquí el valor de \bar{X} obtenemos esta condición:

$$\bar{X} > 2.582$$

que nos indica a partir de que valores de la media muestral rechazaríamos H_0 . Pero cuidado, esto se debe interpretar con prudencia, porque cada muestra produce su propio valor de s .

Otro ejemplo de contraste de hipótesis.

- Vamos a pensar en otro ejemplo:

La inspección de consumo está examinando un envío de latas de conserva, de las que el fabricante afirma que el peso medio son 1000 gramos. Al examinar una muestra aleatoria de 100 latas, un inspector obtuvo un peso medio muestral de 998.5 gramos, con una varianza muestral de $s^2 = 36.1$ (gramos²). Con esos datos, el inspector se pregunta si el peso medio de las latas será en realidad menor que el enunciado por el fabricante. Al nivel de confianza 95 %, ¿qué responderías a la pregunta del inspector? Queremos, además, obtener el p-valor de este contraste.

- Vamos a tomar como valor de referencia $\mu_0 = 1000\text{g}$, como afirma el fabricante. Es importante entender que el peso medio real μ no se conoce. La sospecha del inspector se puede traducir en forma de esta hipótesis alternativa:

$$H_a : \{\mu < \mu_0\}, \quad \text{con } \mu_0 = 1000$$

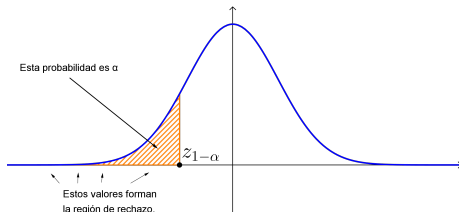
a la que corresponde la hipótesis nula:

$$H_0 : \{\mu \geq \mu_0\}$$

Las desigualdades en estas hipótesis tienen el sentido contrario al del ejemplo inicial con los canguros.

Cálculo del p-valor.

- El esquema es muy parecido al que hemos usado en el primer ejemplo, pero la figura de referencia ahora es esta:



Calculamos el estadístico que es, naturalmente, negativo:

$$\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{998.5 - 1000}{\sqrt{\frac{36.1}{100}}} \approx -2.497$$

- El p-valor ahora es la probabilidad de la cola izquierda del estadístico:

$$\text{pnorm}((998.5 - 1000)/\text{sqrt}(36.1/100)) \approx 0.006262$$

Comparamos el p-valor con $\alpha = 0.05$ y al ser $\text{p-valor} < \alpha$, rechazamos H_0 . Con esos datos rechazamos que el peso medio de las latas sea $\geq 1000\text{g}$.

Contraste bilateral.

- Pensemos el mismo problema desde la perspectiva del fabricante. Al inspector le preocupa que el peso de las latas pueda ser menor que 1000g, porque eso podría ser un fraude a los consumidores (pero si el fabricante decide envasar en cada lata más producto del que anuncia, el inspector no pondrá pegas).
- En cambio la decisión del fabricante es más complicada:
 - si envasa demasiado poco producto, el inspector le sancionará.
 - si, para evitar eso, envasa demasiado producto en cada lata, estará perdiendo dinero.
- ¿Cuál debe ser entonces su objetivo? Lo razonable es intentar que la cantidad de producto envasado se parezca mucho al objetivo marcado $\mu_0 = 1000$ gramos. Así que el fabricante tratará de cumplir la hipótesis nula (bilateral):

$$H_0 = \{\mu = \mu_0\}.$$

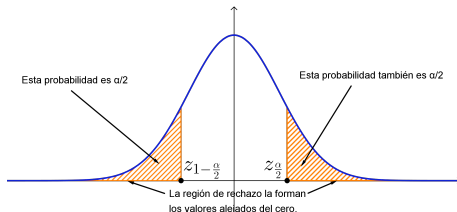
El departamento de control de calidad de la fábrica trabajará para contrastar esta hipótesis frente a la hipótesis alternativa

$$H_a = \{\mu \neq \mu_0\}.$$

- En un contraste bilateral es más difícil rechazar H_0 . Por eso, si hay “*presunción de inocencia*” para H_0 se suelen usar contrastes bilaterales.

Cálculo del p-valor en el caso bilateral. ¡¡Cuidado!!

- En este caso al defensor de la hipótesis nula le preocupa por igual alejarse de μ_0 hacia valores más bajos o más altos. La figura que refleja esa situación es:



Por eso, al calcular el estadístico del contraste incluimos un valor absoluto:

$$\text{estadístico} = \left| \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \right| = \left| \frac{998.5 - 1000}{\sqrt{\frac{36.1}{100}}} \right| \approx 2.497$$

+ Al calcular el p-valor **sumamos la probabilidad de las dos colas**, multiplicando por 2:

$$2 * \text{pnorm}(\text{estadístico}, \text{lower.tail} = \text{FALSE}) \approx 0.01252$$

En este caso, con $ns = 0.95$ también rechazaríamos la H_0 .

Contrastes sobre la media con muestras pequeñas en variables normales.

- El contraste es análogo, cambiando Z por la t_k , siendo k el tamaño de la muestra.
- **Ejemplo:** Vamos a hacer el contraste:

$$H_0 = \{\mu \geq 4\}, \quad H_a = \{\mu < 4\}$$

con $ns = 99\%$ y estos datos muestrales:

$$n = 21, \quad \bar{X} = 3.6, \quad s = 0.6$$

- El valor de referencia es $\mu_0 = 4$, así que el estadístico es:

$$T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{3.6 - 4}{\frac{0.6}{\sqrt{21}}} \approx -3.055$$

Aunque la fórmula es la misma, lo llamamos T porque usamos la t_k de Student (con $k = n - 1 = 20$) para calcular el p-valor:

$$\text{pValor} = \text{pt}(\text{estadistico}, \text{df} = n - 1) \approx 0.003125$$

¿Cuál es la decisión?

- Los otros tipos de contrastes son similares, cambiando Z por la t_k .

La función t.test de R.

- Usar siempre pnorm para hacer contrastes no es cómodo ni eficiente. Por eso no existe t.test.
- **Ejemplo:** Vamos a usar t.test con los datos de la variable cty en la tabla mpg (librería tidyverse) para contrastar la hipótesis alternativa:

$$H_a = \{\mu \neq 16\}$$

```
library(tidyverse)
(testCty = t.test(mpg$cty, mu = 16,
                  alternative = "two.sided", conf.level = 0.95))
```

```
##
##      One Sample t-test
##
## data:  mpg$cty
## t = 3.0874, df = 233, p-value = 0.002264
## alternative hypothesis: true mean is not equal to 16
## 95 percent confidence interval:
##  16.31083 17.40712
## sample estimates:
## mean of x
##  16.85897
```

Las H_a para contrastes unilaterales se indican con less y greater.

Detalles adicionales sobre t.test

- Asignar el resultado de `t.test` a una variable permite acceder a componentes de la respuesta. Por ejemplo, el p-valor:

```
testCty$p.value
```

```
## [1] 0.002263908
```

- Además `t.test` también calcula un intervalo de confianza para la media:

```
testCty$conf.int
```

```
## [1] 16.31083 17.40712
```

```
## attr(,"conf.level")
```

```
## [1] 0.95
```

Si el contraste es unilateral R produce un intervalo de confianza *no acotado*. Por ejemplo, si para la variable `displ` de `mpg` contrastamos $H_a = \{\mu > 3.4\}$

```
testDispl = t.test(mpg$displ, mu = 3.4,  
                   alternative = "greater", conf.level = 0.95)
```

```
testDispl$conf.int
```

```
## [1] 3.332319      Inf
```

```
## attr(,"conf.level")
```

```
## [1] 0.95
```

El intervalo (al 95 %) para μ es $(3.332, +\infty)$

Contrastes de hipótesis para la varianza.

- En el caso de la varianza, que es una medida de **dispersión**, las comparaciones adecuadas utilizan **cocientes en lugar de diferencias**.
- Tipos de contraste: dos unilaterales y una bilateral:

$$H_0 = \{\sigma^2 \leq \sigma_0^2\}, \quad H_a = \{\sigma^2 > \sigma_0^2\}$$

$$H_0 = \{\sigma^2 \geq \sigma_0^2\}, \quad H_a = \{\sigma^2 < \sigma_0^2\}$$

$$H_0 = \{\sigma^2 = \sigma_0^2\}, \quad H_a = \{\sigma^2 \neq \sigma_0^2\}$$

- Si la variable es normal el estadístico adecuado es un cociente, cuya distribución es:

$$Y = (n-1) \frac{s^2}{\sigma^2} \sim \chi_k^2, \quad \text{con } k = n-1.$$

¡Cuidado! Cuando usemos este resultado cambiaremos σ^2 por σ_0^2 porque ese es el valor más favorable a la hipótesis nula.

- El **p-valor** se calcula como en el caso de la media: calculamos la probabilidad de la cola adecuada del estadístico en los casos unilaterales y la multiplicamos por dos en el bilateral.
- **Cuidado:** cuando los contrastes se plantean sobre la **desviación típica** hay que elevar al cuadrado o calcular la raíz cuadrada de los datos muestrales según sea preciso.

Ejemplo de contraste sobre la desviación típica.

- **Ejemplo:** *Un laboratorio farmacéutico garantiza que produce comprimidos de diámetro uniforme, porque la desviación típica de su diámetro es 0.5mm. Una muestra de 15 unidades dio una desviación típica $s = 0.7\text{mm}$. ¿Es aceptable la afirmación del laboratorio al nivel de significación del 5 %?*

- El valor de referencia es $\sigma_0^2 = 0.5^2$ y las hipótesis son:

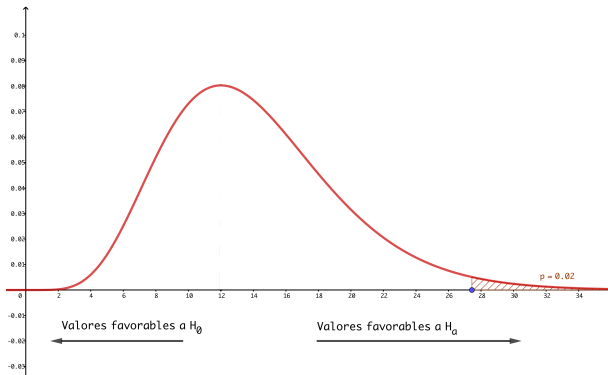
$$H_0 = \{\sigma^2 \leq \sigma_0^2\}, \quad H_a = \{\sigma^2 > \sigma_0^2\}$$

- Con los datos muestrales calculamos el estadístico:

$$Y = (n - 1) \frac{s^2}{\sigma_0^2} = (15 - 1) \frac{0.7^2}{0.5^2} \approx 27.44$$

- Para calcular el p-valor tienes que decidir si calculas la cola izquierda o derecha de este valor en χ_{14}^2 . **¡Es bueno pensar sobre un dibujo que ayude a elegir los valores más favorables a cada una de las hipótesis!**

- La figura para entender la situación es parecida a:



Por tanto para obtener el p-valor calculamos la cola derecha:

$$p\text{Valor} = 1 - \text{pchisq}(\text{estadistico}, \text{df} = n - 1) \approx 0.01687$$

con lo que a un nivel de significación del 95 % ($\alpha = 0.05$) podemos rechazar H_0 y concluir que los datos no permiten afirmar que la desviación típica sea menor o igual que 0.5.

Opcional: la función `sigma.test` de la librería `TeachingDemos`.

- Aunque R básico no incluye ninguna función para los contrastes de varianza en una única variable normal, la librería `TeachingDemos` proporciona `sigma.test`. Asegúrate de instalarla antes de ejecutar este código que ilustra el contraste de

$$H_a = \{\sigma^2 > 16\}$$

en la variable `cty` de `mpg`.

```
require(TeachingDemos)
```

```
## Loading required package: TeachingDemos
```

```
(varTestCty = sigma.test(mpg$cty, sigmasq = 16,  
  alternative = "greater", conf.level = 0.95))
```

```
##
```

```
##   One sample Chi-squared test for variance
```

```
##
```

```
## data:  mpg$cty
```

```
## X-squared = 263.77, df = 233, p-value = 0.0811
```

```
## alternative hypothesis: true variance is greater than 16
```

```
## 95 percent confidence interval:
```

```
##  15.65366      Inf
```

```
## sample estimates:
```

```
## var of mpg$cty
```

```
##      18.11307
```


Tamaño muestral y potencia del contraste.

- Recordemos:
 - Error de tipo I: rechazar H_0 cuando es cierta. $P(\text{error tipo I}) = \alpha$.
 - Error de tipo II: rechazar H_a cuando es cierta. $P(\text{error tipo II}) = \beta$.
- La **potencia** de un contraste es $1 - \beta$ y por tanto puedes pensar que:

$$\begin{aligned}\text{potencia} &= 1 - \beta = 1 - P(\text{error de tipo II}) = \\ &= 1 - P(\text{rechazar } H_a | H_a \text{ es cierta}) = P(\text{rechazar } H_0 | H_0 \text{ es falsa}).\end{aligned}$$

- La potencia del contraste mide cómo de bueno es detectando una H_0 falsa. En general nos gustaría tener a la vez α pequeño y potencia grande. Pero, como en otros casos, no se pueden tener las dos cosas a la vez.
- El cálculo de la potencia es en general complicado. **En el caso de un contraste para la media** es aproximadamente:

$$\text{potencia} = 1 - \beta = K \frac{\delta \sqrt{n} \alpha}{\sigma}$$

donde K es una constante de proporcionalidad, n , α y s son conocidos y δ es el **tamaño del efecto**. Es decir, la diferencia mínima entre μ y μ_0 que queremos que el contraste sea capaz de detectar para rechazar H_0 en tal caso.

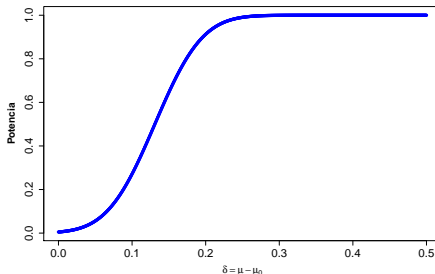
- Observa la ecuación de potencia que hemos obtenido:

$$\text{potencia} = 1 - \beta = K \frac{\delta \sqrt{n} \alpha}{\sigma}$$

- Aunque esta ecuación es sencilla, ilustra varias ideas importantes sobre la potencia:
 1. n es el tamaño de la muestra. A más muestra, más potencia.
 2. δ es la diferencia entre μ y μ_0 . Es decir, el *efecto* que esperamos detectar. Cuanto mayor sea el efecto, mayor potencia.
 3. α es el nivel de significación del contraste, que solemos fijar en 0.05. **No podemos tener a la vez α pequeño y potencia = $1 - \beta$ grande**
 4. σ es la desviación típica, que indica la dispersión de la población. La solemos estimar con s , a menudo procedente de estudios piloto o previos. Y a mayor dispersión, menor potencia.

Curvas de potencia.

- Las ecuaciones de potencia permiten dibujar las llamadas *curvas de potencia* que, para valores de n , α y s fijos, muestran como depende la potencia del tamaño del efecto (medido en “unidades” s). El aspecto típico de una de estas curvas es:



que confirma la idea de que a mayor tamaño del efecto, mayor es la potencia.

La función `power.t.test`

- Las ecuaciones de potencia permiten determinar el tamaño muestral necesario para poder detectar un efecto de tamaño dado, con niveles de significación y potencia dados. En R las funciones con nombres que empiezan por `power` sirven para esto.
- Ejemplo:** ¿Cuál es el tamaño muestral n necesario para un contraste unilateral de hipótesis con $nc = 0.99$ ($\alpha = 0.01$) y potencia $1 - \beta = 0.80$ que sea capaz de detectar un efecto (diferencia entre las medias μ y μ_0) mayor o igual a $\delta = 0.1$? En un estudio piloto obtuvimos $s = 0.5$.
- Usamos:

```
power.t.test(delta = 0.1, sd = 0.5, sig.level = 0.05,  
             power = 0.80, type="one.sample", alternative="one.sided")
```

```
##  
##      One-sample t test power calculation  
##  
##              n = 155.9257  
##          delta = 0.1  
##             sd = 0.5  
##    sig.level = 0.05  
##         power = 0.8  
## alternative = one.sided
```

- Ejercicio.** Aquí hemos calculado n , pero esta función puede usarse para determinar uno de los valores a partir de los restantes. Prueba

```
power.t.test(delta = 0.1, sd = 0.5, sig.level = 0.05, n = 300,  
             type="one.sample", alternative="one.sided")
```

Sección 6

Uso y abuso del p-valor.

P-VALUE

INTERPRETATION

0.001

0.01

0.02

0.03

HIGHLY SIGNIFICANT

0.04

0.049

SIGNIFICANT

0.050

OH CRAP. REDO
CALCULATIONS.

0.051

ON THE EDGE
OF SIGNIFICANCE

0.06

0.07

0.08

0.09

HIGHLY SUGGESTIVE,
SIGNIFICANT AT THE
 $P < 0.10$ LEVEL

0.099

≥ 0.1

HEY, LOOK AT
THIS INTERESTING
SUBGROUP ANALYSIS

Significación estadística vs relevancia científica.

- *Un fabricante garantiza que produce comprimidos de diámetro medio de 13mm. En una muestra de 50 unidades tiene un diámetro medio $\bar{X} = 13.05\text{mm}$, desviación típica $s = 0.6\text{mm}$. ¿Es aceptable esta afirmación al nivel de significación del 99 %?*
- La hipótesis nula es $H_0 : \{\mu = 13\}$, y la alternativa $H_a : \{\mu \neq 13\}$. El estadístico y p-valor (calculado con la t de Student) son:

$$T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{13.05 - 13}{\frac{0.6}{\sqrt{50}}} \approx 0.5893, \quad \text{p-valor: } 0.5584$$

- El contraste **no** es significativo, rechazamos H_a . Fíjate en que el efecto es $\delta = 0.05$. Pero ahora repetimos la cuenta con los mismos valores, salvo que aumentamos el tamaño muestral a $n = 5000$.

$$T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{13.05 - 13}{\frac{0.6}{\sqrt{5000}}} \approx 5.893, \quad \text{p-valor: } 4.052 \times 10^{-9}$$

- Ahora el p-valor es *muy* pequeño y rechazamos H_0 . Este ejemplo ilustra un principio general: **si se usan muestras suficientemente grandes, incluso un efecto $\delta = \mu - \mu_0$ muy pequeño (irrelevante) puede llegar a ser estadísticamente significativo.**

- ¿Como podemos entonces juzgar la **relevancia** del efecto observado? El primer consejo es que *los resultados de un contraste deberían ir siempre acompañados de estimaciones del tamaño del efecto*. Por ejemplo, usando intervalos de confianza.

- La **d de Cohen** es:

$$d = \frac{\bar{X} - \mu_0}{s}$$

Podemos usarla para hacernos una idea aproximada de la relevancia del efecto observado teniendo en cuenta estas indicaciones:

- Un valor $d < 0.2$ indica un efecto no relevante.
- Si es $d > 0.8$ es muy posible que la diferencia sea relevante.
- Cuando $0.2 < d < 0.8$ se necesita la **opinión de un experto** que juzgue la relevancia de los resultados.

- En el ejemplo anterior, con la muestra grande, se obtiene:

$$d = \frac{\bar{X} - \mu_0}{s} = \frac{13.05 - 13}{0.6} \approx 0.08333$$

así que parece que ese efecto $\delta = 0.05$ es, seguramente, irrelevante.

Sección 7

Uso y abuso del p-valor.

El problema de los contrastes múltiples.

- El contraste de hipótesis es el método habitual para confirmar un resultado científico o técnico a partir de los datos. Pero su uso puede prestarse a errores o incluso a manipulaciones mal intencionadas.
- Por ejemplo, ya sabemos que α es la probabilidad de cometer un error de tipo I (rechazar una H_0 cierta). Si por ejemplo $\alpha = 0.05$. Si realizamos 20 contrastes independientes de hipótesis nulas **todas ellas ciertas** ¿cuál es la probabilidad de que (nos toque alguna muestra mala y) rechacemos alguna H_0 ? Es fácil ver que la situación tiene todos los ingredientes de una binomial $B(20, \alpha)$ y si $X =$ (número de H_0 rechazadas) entonces

$$P(X > 0) = 1 - P(X = 0) = 1 - (1 - \alpha)^{20} \approx 0.641514$$

que también puedes calcular en R como: `1 - dbinom(0, size = 20, prob = 0.05)`.

- Eso significa que simplemente repitiendo el contraste 20 veces hay un 64 % de probabilidades de obtener un resultado *significativo jiy falso!!*.

Simulando contrastes múltiples con R.

- Vamos a usar R para confirmar la discusión anterior en una simulación.

```
set.seed(2019)
nTests = 20 # Haremos 20 contrastes
# y este vector los 20 p-valores
pValores = numeric(nTests)
# Ahora hacemos los contrastes y guardamos los p-valores
for(i in 1:nTests){
  muestra = c(rnorm(15))
  pValores[i] = t.test(muestra, alternative = "two.sided", mu = 0)$p.value
}
# ¿Cuál es el p-valor más pequeño?
min(pValores)

## [1] 0.03307146
```

Como puede verse, hay al menos un p-valor $< \alpha$, así que estaríamos rechazando incorrectamente al menos una de las H_0 .

- Una primera solución consiste en aplicar la [corrección de Bonferroni](#), que en esencia cambia el criterio de rechazo de H_0 de p-valor $< \alpha$ por el criterio p-valor $< \frac{\alpha}{m}$ siendo m el número de contrastes.

- La situación que acabamos de describir puede ocurrir por desconocimiento, pero también puede ser una estratagema de alguien tratando de obtener un resultado significativo a cualquier coste. Este tipo de manejos forman parte de lo que se denomina [p-hacking](#), o [data-dredging](#).
- Recomendamos leer esta nota breve de [Investigación y Ciencia](#) o aún mejor el [artículo de Nature](#) del que procede, o este [otro más extenso](#), de Regina Nuzzo, también en Nature.
- El control del error en contrastes repetidos es un tema ha sido objeto de estudio intenso recientemente. Por ejemplo en *Genómica* o en *Big Data* son comunes los casos en los que se contrastan decenas de miles de hipótesis a la vez (uno por gen, por ejemplo). En esos casos usar correcciones tipo Bonferroni sería demasiado drástico: rechazaríamos demasiado pocas H_0 . Una referencia elemental para empezar a entender este tema es [este artículo](#) de la Wikipedia.

Sección 8

Complementos de R.

- Como referencias para este apartado puedes usar (Boehmke 2016, capítulo 11), (Matloff 2011, capítulo 4).
- A diferencia de los vectores, las listas sirven para guardar elementos heterogéneos (incluidas sublistas). La forma más sencilla de crear una lista es usando `list`:

```
(planeta = list(nombre = "Marte", exterior = TRUE,  
               radio = 3389.5, satelites = list("Fobos", "Deimos")))
```

```
## $nombre  
## [1] "Marte"  
##  
## $exterior  
## [1] TRUE  
##  
## $radio  
## [1] 3389.5  
##  
## $satelites  
## $satelites[[1]]  
## [1] "Fobos"  
##  
## $satelites[[2]]  
## [1] "Deimos"
```

Accediendo a los elementos de una lista.

- R usa \$ o doble corchete [[]] para identificar los elementos de la lista.

```
planeta[[1]]
```

```
## [1] "Marte"
```

```
planeta$exterior
```

```
## [1] TRUE
```

```
planeta$satelites[[1]]
```

```
## [1] "Fobos"
```

La salida es del tipo de objeto que hay en esa posición de la lista.

- Pero fíjate en la diferencia si usamos un único corchete:

```
planeta[1]
```

```
## $nombre
```

```
## [1] "Marte"
```

```
planeta["exterior"]
```

```
## $exterior
```

```
## [1] TRUE
```

En este caso la salida *siempre es una lista*.

Funciones list, append y c.

- Atención a esta diferencia:

```
(l1 = list("A", "B"))
```

```
## [[1]]  
## [1] "A"  
##  
## [[2]]  
## [1] "B"
```

```
(l2 = list(c("A", "B")))
```

```
## [[1]]  
## [1] "A" "B"
```

La función `list` siempre crea *listas anidadas*. Por ejemplo este comando (no se muestra la salida) crea una lista con dos componentes y el primero es `l2`:

```
(l3 = list(l2, "C"))
```

- Las funciones `append` y `c` *adjuntan* elementos. Estos comandos son equivalentes:

```
l4 = append(l2, "D")  
(l4 = c(l2, "D"))
```

```
## [[1]]  
## [1] "A" "B"  
##  
## [[2]]  
## [1] "D"
```

También se pueden añadir elementos por nombre, como en

```
planeta$distSol = 227.9
```

Otras propiedades y operaciones con listas.

- La función `length` produce el número de elementos de una lista. Y con `names` se obtienen los nombres de sus elementos (si se han dado nombres).
- **Ejercicio:** Prueba a usar `names` y `length` con varias de las listas que hemos creado. Ejecuta (`sesion = sessionInfo()`) para ver lo que hace esa función. Y luego explora como acceder a las componentes usando `sesion$`
- Para eliminar elementos de una lista basta con hacerlos `NULL`.

```
l4[3] = NULL
l4
```

```
## [[1]]
## [1] "A" "B"
##
## [[2]]
## [1] "D"
```

- La función `unlist` *aplana* una lista dando como resultado un vector:

```
unlist(l1)
```

```
## [1] "A" "B"
```

- **Ejercicio:** ¿qué se obtiene al aplicar `unlist` a la siguiente lista?
`lista = list(letters[1:3], matrix(1:12, nrow = 3), TRUE)`.

Estructuras de control en R. Bloques if/else.

- Como referencias para este apartado puedes usar (Boehmke 2016, capítulo 19), (Matloff 2011, capítulo 7).
- **Bloques if/else.** La estructura básica de estos bloques es:

```
if (condición) {  
  ...  
  sentencias que se ejecutan si condicion = TRUE  
  ...  
} else {  
  ...  
  sentencias que se ejecutan si condicion = FALSE  
  ...  
}
```

Si necesitas condiciones anidadas puedes cambiar else por else if y añadir a continuación otra condición para crear un nuevo nivel de la estructura.

- La estructura if está pensada para ejecutarse sobre una *única* condición que produzca un *único* valor TRUE/FALSE. Existe también una función vectorializada, llamada ifelse que se puede aplicar a un vector de condiciones. Un ejemplo:

```
ifelse(((1:5) < 3), yes = "A", no = "B")
```

```
## [1] "A" "A" "B" "B" "B"
```

Bucles for.

- El bucle for se utiliza cuando queremos repetir un bloque de comando y conocemos de antemano el número máximo de repeticiones. Su estructura básica es similar a esta:

```
for(k in valores_k) {  
  ...  
  cuerpo del bucle, se repite a lo sumo length(valores_k) veces  
  ...  
}
```

La variable *k* (el nombre es arbitrario) es el *contador* del bucle for. El vector *valores_k* contiene los valores que toma *k* en cada iteración.

- Si en alguna iteración queremos interrumpir el bucle cuando se cumple alguna condición (y no hacer ninguna iteración más), podemos combinar *if* con la función *break*. Si lo que queremos es solamente pasar a la siguiente iteración usamos *next* en lugar de *break*.
- A menudo se usa un bucle for para “rellenar” un objeto como un vector o matriz. Es importante recordar que R es poco eficiente haciendo “crecer” estos objetos. En esos casos es mucho mejor comenzar creando el objeto completo, con todas sus posiciones, e ir asignado valores a posiciones en cada iteración (R lo inicializa a 0).
- En general **es preferible usar operaciones vectorializadas en lugar de bucles for**. Pronto aprenderemos las funciones de la familia *apply* para hacer esto.

Ejemplo de bucle for con next y break.

- El siguiente código ilustra un bucle for con el uso de next y break. Ejecútalo varias veces para ver como se comporta según los valores de sus parámetros.

```
valores = numeric(10) # Creamos un vector del tamaño previsto
for (k in 1:10){
  sorteo = sample(1:20, 1)
  print(paste0("k = ", k, ", sorteo = ", sorteo))
  if (k %in% 5:6){
    next # saltamos dos valores
  } else if (sorteo == 1){
    print("Resultado del sorteo es 1, fin del bucle")
    break # paramos si un valor aleatorio es 1
  }
  valores[k] = k # se ejecuta cuando no se cumplan las condiciones
}
valores
```

- Ejercicio:** ¿Qué valores asigna R a los elementos de los vectores creados respectivamente con `x = logical(10)` y con `v = character(10)`?

Otros bucles: while y repeat.

- En R también existen estos dos tipos de bucles, comunes a muchos lenguajes. Conviene insistir en que suele ser más eficiente evitar el uso de bucles.

- El bucle `while` tiene esta estructura:

```
while (condición){  
    ...  
    cuerpo del bucle, que eventualmente debe hacer condición TRUE o usar break  
    ...  
}
```

- El bucle `repeat` tiene esta estructura:

```
repeat {  
    ...  
    cuerpo del bucle, que debe usar break  
    ...  
}
```

Insistimos: a diferencia de otros lenguajes, en R un bucle `repeat` debe usar explícitamente `break` para detenerse.

- El código de este tema contiene ejemplos de bucle `while` y `repeat` con `break`, que puedes ejecutar varias veces. Observa las diferencias en el comportamiento de ambos bucles.

Rmarkdown para la creación de documentos.

- En las sesiones del curso veremos algunos ejemplos sencillos para que puedas iniciarte en el manejo de Rmarkdown.
- Rmarkdown, creado por Yihui Xie, es una herramienta muy general para la creación de documentos, en especial documentos relacionados con el Análisis de Datos. A partir de los ficheros escritos con Rmarkdown se pueden obtener fácilmente salidas en formato pdf o HTML, pero también presentaciones, artículos e informes técnicos, entradas de blog, libros, dashboards con Shiny y la lista sigue creciendo.
- Hay dos fuentes de información online básicas para adentrarse en las posibilidades de RMarkdown:
Del propio Yihui Xie: [R Markdown: The Definitive Guide](#).
De RStudio: [R Markdown: Get Started](#).
- Recomendamos empezar leyendo una [introducción reciente a RMarkdon](#) de T. Goicoa (Univ. Pública de Navarra).

Enlaces

- [Código de esta sesión](#)

Bibliografía

Boehmke, Bradley C. 2016. *Data Wrangling with R*. Springer.
<https://doi.org/10.1007/978-3-319-45599-0>.

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

Matloff, Norman S. 2011. *The art of R programming : tour of statistical software design*. No Starch Press. <https://doi.org/10.1080/09332480.2012.685374>.