# Técnicas de Simulación

03. Errores tipo I y tipo II

Miguel A. Castellanos

## Contents



---

> _Te recomiendo que empieces a leer el documento por el último punto *Generación de documentos con Rmarkdown* porque la última tarea requiere reescribir los resultados obtenidos en un fichero .pdf usando Rmarkdown. Es posible que te resulte más cómodo empezar por ahí para escribir directamente las tareas y sus explicaciones en ese formato._En esta práctica, además del código es importante las explicaciones e interpretaciones de los resultados obtenidos, incluye resultados, tablas y las gráficas que sean necesarias para ejemplificar vuestras conclusiones.

Los conceptos con los que vamos a trabajar en este estudio de caso provienen de la estadística inferencial. Aunque aquí se mencionarán brevemente no es posible explicar los conceptos en profundidad, por lo que el alumno que lo requiera deberá acudir a textos básicos de estadística. Algunos de los más habituales son:

Amón (1996), Field (2013), Howell (2014), Pardo y Ruiz (2009) y Pardo y San Martín (1998). Las referencias se encuentran al final del documento.

Además, junto con este documento, os dejo un pdf que contiene el capítulo 8 del libro Análisis de datos en Psicología de Martínez Arias, Chacón y Castellanos. Es el capítulo que más se centra en los conceptos tratados y es posible que os resulte de utilidad.

# 1 Estudio y manipulación de la normalidad de una variable

Para estudiar la normalidad de una variable se suelen utilizar dos estrategias, la primera consiste en describir la variable y la segunda en realizar algún contraste de hipótesis. Los descriptivos básicos usados son los momentos de Pearson (media, desviación asimetría y apuntamiento), los histogramas, las figuras de caja y bigote y, lo más útil, los qqplot. Para contrastar si una variable es normal se pueden usar los contrastes de Kolmogorov-Smirnov o de Shapiro-Wilk. Vamos a verlo con R:

```
# usamos el paquete moments para calcular asimetria y apuntamiento
if (!require("moments")) install.packages("moments")
```

```
## Loading required package: moments
```

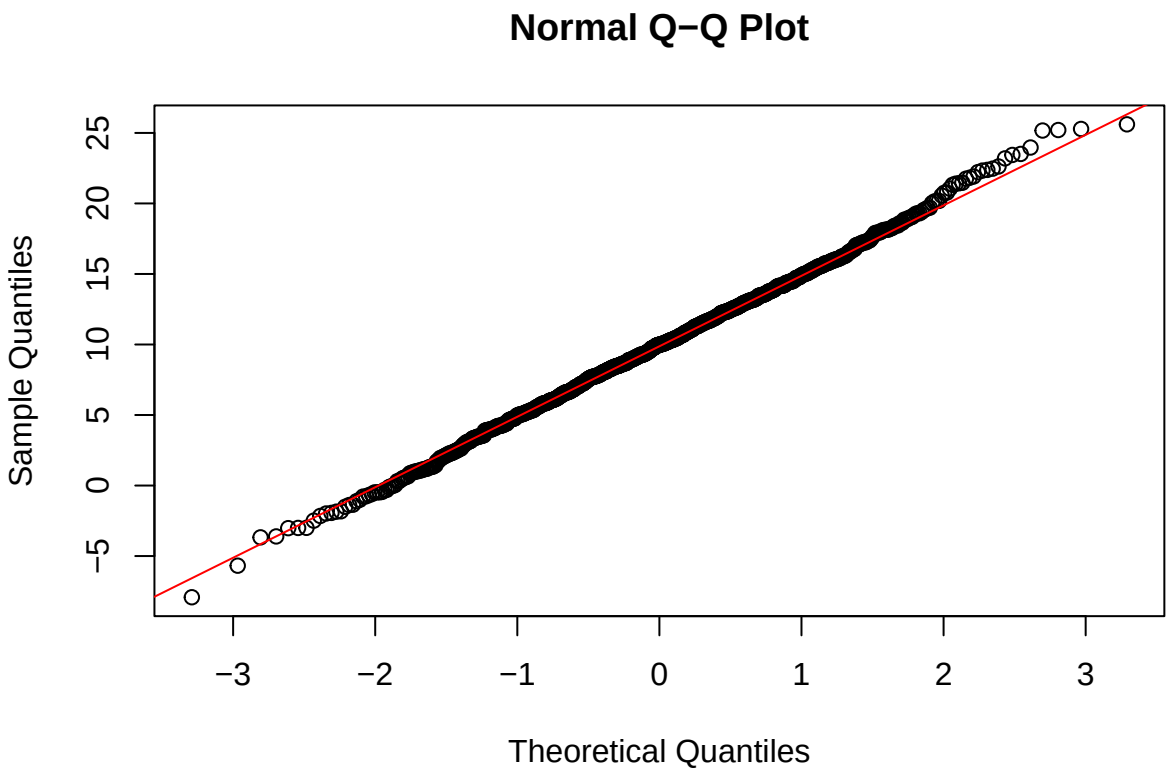
```
d <- rnorm(1000, 10, 5)

skewness(d)
```

```
## [1] 0.01706649
```

```
kurtosis(d)
```

```
## [1] 3.132645
```

```
# Ojo: si los valores no estan normalizados suelen ser de 0 y 3 para la normal en asimetria y apuntamie

# un qqplot
qqnorm(d)
qqline(d, col = "red")
```




```
# El test de Shapiro sobre normalidad
shapiro.test(d)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d
## W = 0.99884, p-value = 0.7776
```

Cuando una variable no es normal se puede intentar normalizar usando algún tipo de transformación como la de Box-Cox que funciona razonablemente bien. Una segunda situación ligada a esta es justo la inversa: mi distribución es normal pero quiero que tenga la forma que yo desee, con los 4 momentos (media, desviación, asimetría y apuntamiento) definidos por mí. Esto es lo que me permite hacer la distribución de Johnson, implementada en el paquete SuppDists. Fíjate en el siguiente código:

```
# funcion de johnson
if (!require("SuppDists")) install.packages("SuppDists")
```

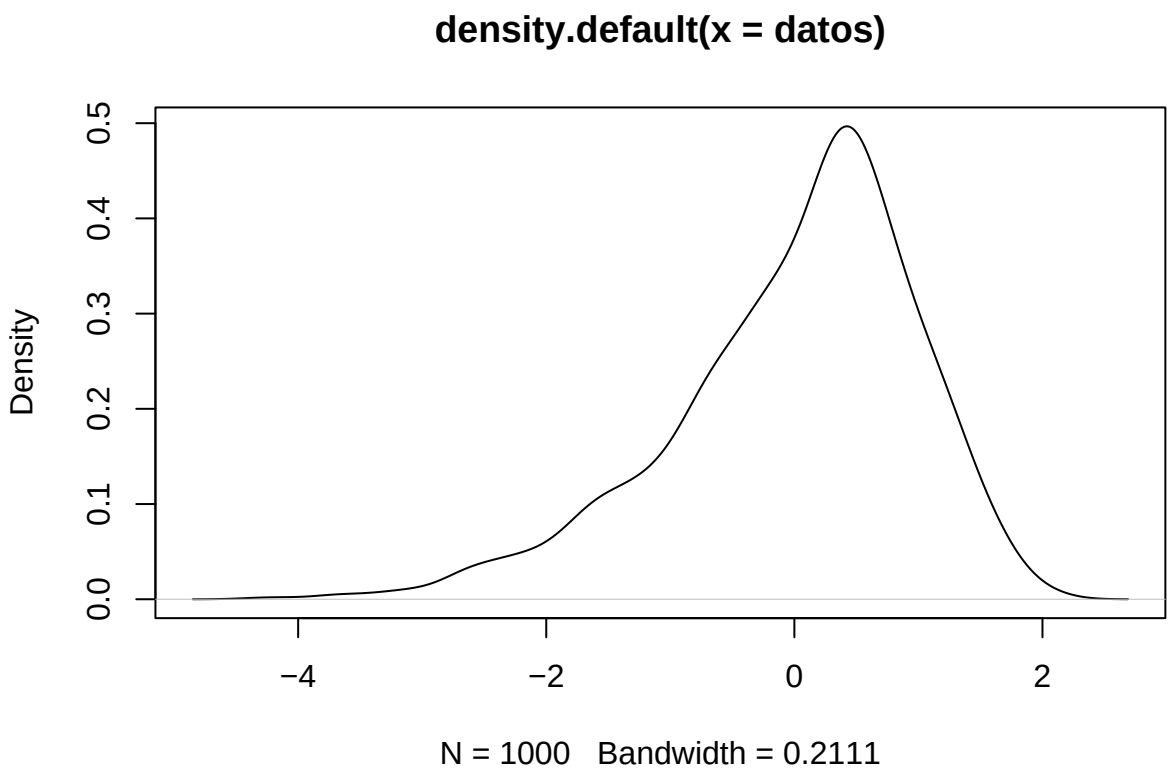
```
## Loading required package: SuppDists
```

```
# media, sd, asimetria y apuntamiento c(0, 1, -.5, 4)
parms<-JohnsonFit(c(0, 1, -.8, 4), moment="use")
sJohnson(parms)
```

```
## $title
## [1] "Johnson Distribution"
##
## $gamma
```

```
## [1] -4.910355
##
## $delta
## [1] 2.603182
##
## $xi
## [1] -18.92311
##
## $lambda
## [1] 21.94518
##
## $type
## SB
##  4
##
## $Mean
## [1] 0.00085953
##
## $Median
## [1] 0.1325832
##
## $Mode
## [1] 0.410716
##
## $Variance
## [1] 0.9755719
##
## $SD
## [1] 0.9877105
##
## $ThirdCentralMoment
## [1] -0.7174966
##
## $FourthCentralMoment
## [1] 3.434918
##
## $PearsonsSkewness...mean.minus.mode.div.SD
## [1] -0.4149561
##
## $Skewness...sqrtB1
## [1] -0.7446135
##
## $Kurtosis...B2.minus.3
## [1] 0.6090906
```

```
datos <- rJohnson(1000, parms)
plot(density(datos))
```




Hemos definido la forma de la función (una asimétrica negativa) pasándole los momentos que hemos deseado, para posteriormente generar datos que provienen de esa distribución. Esto aunque parezca irrelevante nos va a ser muy útil en las simulaciones a partir de ahora.

> **Un aviso muy importante**: No todas las combinaciones de parámetros son posibles, a veces no se pueden dibujar y a veces no producen exactamente el resultado deseado. Esto es debido a que matemáticamente no se pueden definir inconsistencias entre los momentos de una distribución. Cuando esto ocurra simplemente probamos configuraciones de parámetros más razonables.

## 1.1 Tarea 1

Utilizando la distribución de Johnson calcula por fuerza bruta qué combinaciones de parámetros generan una distribución válida. Puedes fijar los valores de la media y la desviación típica a 0 y 1 y probar con un doble bucle. Valores razonables de asimetría y apuntamiento para probar pueden ser [-2, 2] y [0,6] respectivamente. Recuerda que las asimetría y el apuntamiento (tercer y cuarto momento de pearson) de la normal son 0 y 3 respectivamente.

Cuando una función falla en R se detiene el procesamiento pero eso es necesario utilizar funciones como tryCatch. Busca ejemplos de su uso.

## 1.2 Tarea 2

Utilizando la distribución de Johnson dibuja las siguientes distribuciones:

- Una normal
- Una platicúrtica
- Una mesocúrtica
- Una asimétrica positiva

- Una asimétrica negativa

## 2 Distribución muestral de la media

Una distribución muestral de un estadístico es la distribución que adopta el estadístico al extraerse infinitas muestras de tamaño n de una población. Por ejemplo, para la media:

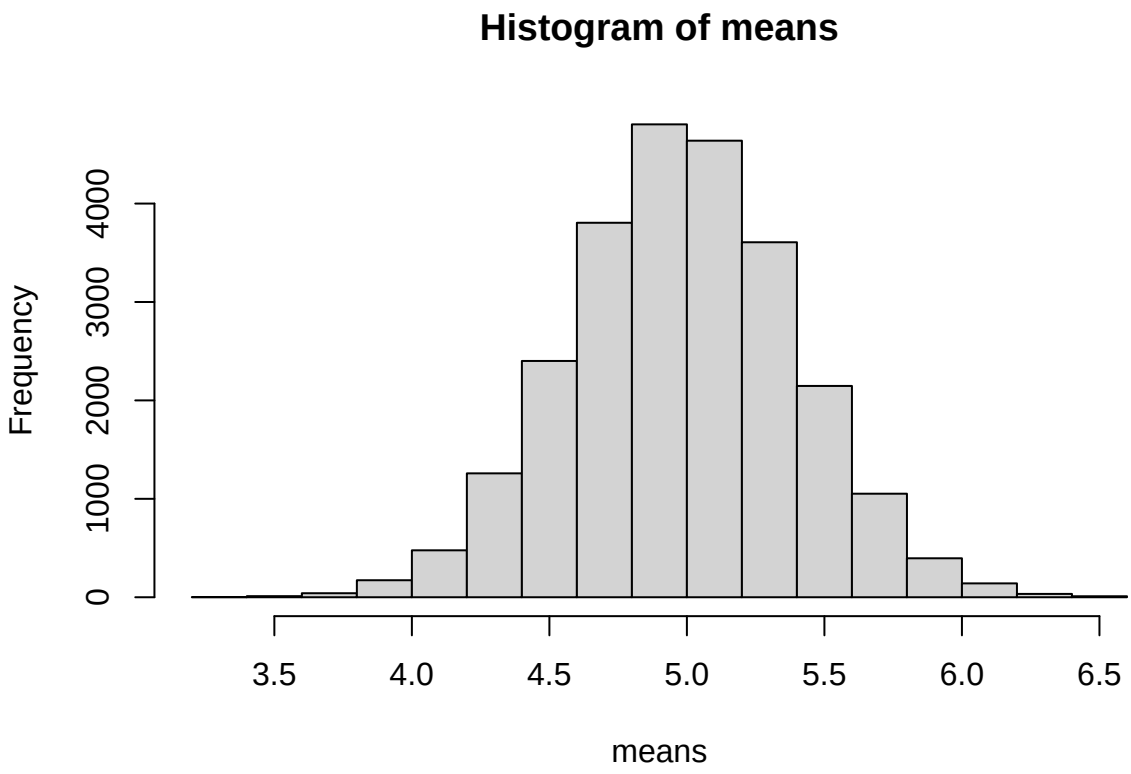
- Tenemos una población de 1000 elementos (N=1000).
- Extraigo de ella 25000 muestras (k=25000) de 25 elementos (n=25) , obviamente con reposición.
- Con esas k muestras calculo la media.
- Con esas k medias dibujo la distribución (histograma).

Si hacemos esto, la teoría nos dice que la distribución resultante, llamada **distribución muestral de la media**, cumple las siguientes condiciones:

- Tendrá forma normal, es decir $N(\mu, \sigma/\sqrt(n))$
- Su media es la media poblacional $\mu_{\overline{X}} = \mu$
- Su desviación típica es igual $\sigma_{\overline{X}} = \sigma/\sqrt(n)$

Esto lo podemos ver fácilmente con R

```
set.seed(1)

N=1000   # tamano de la poblacion
n = 25  # tamano de las muestras
k = 25000

# creamos una poblacion de N numeros aleatorios con media 5 y sd =2
poblacion <- rnorm(N, 5, 2)

# comprobamos media y sd porque al ser aleatorios no tienen que ser exactamente 5 y 2
mu.pob <- mean(poblacion)
sd.pob <- sd(poblacion)

# extraigo k muestras
# no es el codigo mas eficiente pero al menos reservo la memoria al principio
means <- vector(length=k)

for (i in 1:k){
  muestra <- sample(poblacion, n)
  means[i] <- mean(muestra)
}

# dibujamos esa distribucion de medias
# plot(density(means))  mas bonito pero menos preciso
hist(means)
```




```
# Deben coincidir las afirmaciones teoricas
m.teo <- mu.pob
sd.teo <- sd.pob/sqrt(n)

m.emp <- mean(means)
sd.emp <- sd(means)

sprintf("Media teoría: %.2f -- Media empírica : %.2f", m.teo, m.emp)
```

```
## [1] "Media teoría: 4.98 -- Media empírica : 4.98"
```

```
sprintf("Sd teoría: %.2f -- Sd empírica : %.2f", sd.teo, sd.emp)
```

```
## [1] "Sd teoría: 0.41 -- Sd empírica : 0.41"
```

Como se puede comprobar las medias y las desviaciones típicas coinciden prácticamente.

## 2.1 Tarea 3

Muestra que esto se cumple, aunque la población de origen no sea normal, prueba con la distribución uniforme y con una muy sesgada positiva como la ji-cuadrado

## 2.2 Tarea 4

Muestra que la distribución muestral de la proporción es también $N(\mu, \sigma/\sqrt{n})$ con muestras de aproximadamente $n \geq 30$, pero no lo es para muestras más pequeñas (por ejemplo de $n = 5$). Recuerda que la población que tienes que crear ahora es binomial (ceros y unos)

### 2.3 Tarea 5

La distribución muestral de la varianza es más compleja ya que no sigue una distribución normal. Sin embargo tiene algunas propiedades muy interesantes:

- La esperanza de la varianza es:

$$\mu_{S^2} = \frac{n-1}{n}\sigma^2$$

- La esperanza de la varianza insesgada es:

$$S^2_{n-1} = \frac{\sum(X - \overline{X})^2}{n-1}$$

$$\mu_{S^2_{n-1}} = \sigma^2$$

- La varianza de $nS^2/\sigma^2$ es 2(n-1)

$$\sigma^2_{nS^2/\sigma^2} = 2(n-1)$$

- La distribución de $nS^2/\sigma^2$ es una ji-cuadrado con n-1 grados de libertad.

$$nS^2/\sigma^2 \Rightarrow \chi^2_{n-1}$$

Muestra con R las afirmaciones anteriores.

## 3 Los intervalos de confianza

Si la distribución de la media es como hemos dicho $N(\mu, \sigma/\sqrt{n})$ podemos construir un intervalo de confianza siguiendo la regla:

$$IC : \mu \pm Z_{\alpha/2}\sigma/\sqrt{n}$$

Donde $Z_{\alpha/2}$ son los valores de 1.96 y 2.58 para valores de alfa de 0.05 y 0.01 respectivamente.

### 3.1 Tarea 6

Un intervalo construido con la fórmula anterior cumple, por lógica, la siguiente afirmación:

> Con un $\alpha = 0.05$ el 95% de las muestras extraídas de una población tendrán una media que caerá dentro de los límites del intervalo, y con un $\alpha = 0.01$ caerán el 99% de las medias.

Compruebe con R la afirmación anterior.

Recuerde que estamos calculando los intervalos utilizando la distribución normal, porque en nuestras simulaciones creamos nosotros las poblaciones, y por tanto $\mu$ y $\sigma$ (los parámetros poblacionales) son conocidos. En general, con poblaciones reales, estos parámetros son desconocidos por lo que sustituimos $Z_{\alpha/2}$ por $t_{n-1,\alpha/2}$. En este caso los intervalos de confianza para la media se calculan de la siguiente manera:

$$IC : \overline{X} \pm t_{n-1,\alpha/2}S_x/\sqrt{n}$$

A efectos prácticos, ambas distribuciones son tan similares con muestras mayores de 30 (lo estudiamos en semanas anteriores) que podemos simplificar y usar los valores convencionales de 1.96 y 2.58 para valores de alfa de 0.05 y 0.01 respectivamente.

# 4 Contrastes de hipótesis

Siguiendo mi adorada wikipedia podemos leer:

> Dentro de la inferencia estadística, un contraste de hipótesis (también denominado test de hipótesis o prueba de significación) es un procedimiento para juzgar si una propiedad que se supone en una población estadística es compatible con lo observado en una muestra de dicha población. Fue iniciada por Ronald Fisher y fundamentada posteriormente por Jerzy Neyman y Karl Pearson.
>
> Mediante esta teoría, se aborda el problema estadístico considerando una hipótesis determinada $H_0$, y una hipótesis alternativa $H_1$, y se intenta dirimir cuál de las dos es la hipótesis verdadera, tras aplicar el problema estadístico a un cierto número de experimentos.
>
> Está fuertemente asociada al concepto estadístico de potencia y a los conceptos de errores de tipo I y II, que definen respectivamente, la posibilidad de tomar un suceso falso como verdadero, o uno verdadero como falso.

No voy a explicar los conceptos de contraste, si es necesario os recomiendo que reviséis vuestros materiales de estadística de la carrera o alguno de los libros que he recomendado al principio del documento. En R existen diferentes funciones que tienen implementados los contrastes de hipótesis, por ejemplo la función *t.test* realiza contrastes sobre la media. Para poner a prueba la hipótesis nula de que la población tiene un media de 10 ($H_0 : \mu = 10$) aplicaríamos la siguiente sintaxis:

```
set.seed(1)
d <- rnorm(50, 10, 10)
t.test(d, mu=10)
```

```
##
##  One Sample t-test
##
## data:  d
## t = 0.85432, df = 49, p-value = 0.3971
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   8.641687 13.367278
## sample estimates:
## mean of x
##  11.00448
```

Obteniendo un pvalor de 0.3971 y por tanto aceptando la $H_0$. Si por otro lado hubiese probado que la media de la población es 14 habría rechazado la $H_0$ con un pvalor de 0.014.

```
t.test(d, mu=14)
```

```
##
##  One Sample t-test
##
## data:  d
## t = -2.5477, df = 49, p-value = 0.01403
## alternative hypothesis: true mean is not equal to 14
## 95 percent confidence interval:
##   8.641687 13.367278
## sample estimates:
## mean of x
##  11.00448
```

Se pueden cometer dos errores en un contraste de hipótesis:

- Errores tipo I: $\alpha$ , nivel de significación. Probabilidad de rechazar la $H_0$ cuando esta es verdadera

- Errores tipo II: $\beta$. Probabilidad de aceptar la $H_0$ cuando esta es falsa

veámoslos:

# 5 Errores tipo I

Los errores tipo I son prefijados antes del contraste por el investigador, y los llamamos $\alpha$ teóricos ($\alpha_{teo}$), sin embargo es posible que ciertas circunstancias, como el incumplimiento de los supuestos de la prueba, lleve a que estemos cometiendo más error del establecido a priori, es lo que vamos a llamar el $\alpha$ empírico ($\alpha_{emp}$).

> En realidad, la comprobación de este $\alpha_emp$ requiere más conocimiento de lo que vamos a utilizar en esta práctica, es decir, lo que vamos a hacer no es del todo correcto, solo una simplificación, pero nos valdrá para poner a prueba la idea principal.

Si extraemos k muestras de una población que es normal $N(\mu_0, \sigma)$ y realizamos un contraste para una media en el que ponemos a prueba la $H_0 : \mu = media$ deberíamos encontrarnos con que rechazaríamos la $H_0$ el 5% de esas k veces (si hemos establecido $\alpha_{teo} = 0.05$). Es decir, partimos de un $\alpha_{teo} = 0.05$ y queremos comprobar si realmente nuestro $\alpha_{emp}$ es también igual a 0.05. Vamos a verlo con R:

```
set.seed(1)
N=1000
n = 25

k = 500

#poblacion <- sample(1:10, N, replace=T)
poblacion <- rnorm(N, 10, 10)
mu.pob <- mean(poblacion)
sd.pob <- sd(poblacion)


p <- vector(length=k)
for (i in 1:k){
  muestra <- poblacion[sample(1:N, n)]
  p[i] <- t.test(muestra, mu = mu.pob)$p.value
}

a_teo = 0.05
a_emp = length(p[p<a_teo])/k
sprintf("alpha_teo = %.3f <-> alpha_emp = %.3f", a_teo, a_emp)
```

```
## [1] "alpha_teo = 0.050 <-> alpha_emp = 0.050"
```

El valor de ambos es idéntico, eso indica que con esas condiciones no hay diferencia entre el $\alpha$ empírico y teórico.

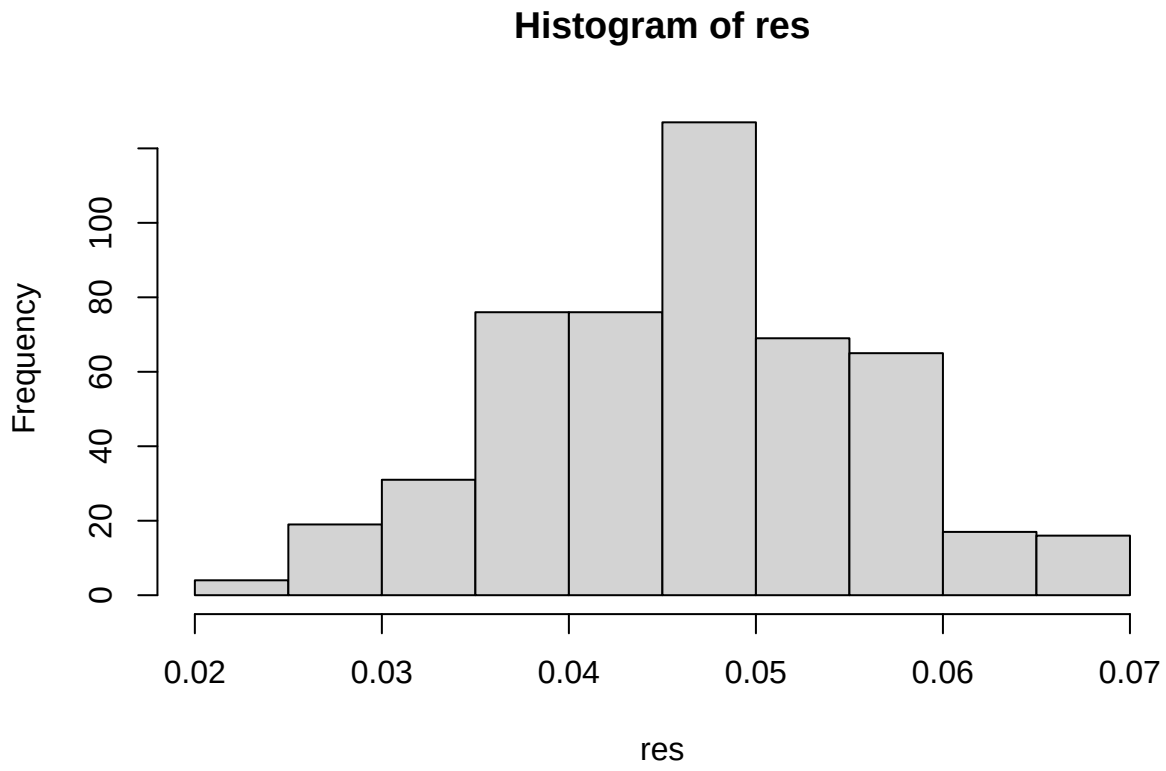
Sin embargo, siendo justos, lo que hemos obtenido es un valor que corresponde a un único experimento, es decir, en nuestro experimento:

- Hemos creado una población
- Hemos extraído muestras
- Hemos hecho una prueba t
- Hemos comprobado la cantidad de aciertos y fallos que hemos cometido en nuestra decisión (alfa)

Pero todos esos pasos juntos son un único experimento, si repitiésemos el experimento más veces (una nueva población, extracción, etc.) podríamos obtener resultados diferentes. **Compruébalo, ejecuta varias el código anterior y mira cuál es el valor del alfa empírico** (quitando el set.seed(1), obviamente). Como ves estás obteniendo valores del alfa empírico diferentes.

Es decir, como resultado del experimento has obtenido un valor que pertenece a una distribución de posibles resultados. Para poder caracterizar esa distribución debes repetir el experimento un número razonable de veces.

```
expto1 <- function(N,n,k){

  #poblacion <- sample(1:10, N, replace=T)
  poblacion <- rnorm(N, 10, 10)
  mu.pob <- mean(poblacion)
  sd.pob <- sd(poblacion)


  p <- vector(length=k)
  for (i in 1:k){
    muestra <- poblacion[sample(1:N, n)]
    p[i] <- t.test(muestra, mu = mu.pob)$p.value
  }
  a_emp = length(p[p<a_teo])/k
  return(a_emp)
}

set.seed(1)
rep <- 500
res <- vector(length=rep)
for (i in 1:rep) res[i] <- expto1(1000, 25, 500)
hist(res)
```



```
mean(res)
```

```
## [1] 0.047084
```

```
sd(res)
```

```
## [1] 0.009296763
```

Ahora el tiempo de cálculo aumenta significativamente. La distribución de los alfa empíricos tiene forma de normal, sin embargo su media está ligeramente desplazada, según estos datos parece que al realizar el contraste con estas condiciones cometemos ligeramente menos errores tipo I de los indicados por la teoría. Sin embargo es posible encontrar circunstancias en las que esto no es así, como por ejemplo si:

- Los datos no son normales
- Los datos no son estrictamente continuos
- Los datos tienen un tamaño insuficiente o hay una gran cantidad de datos *missing*

  Importante: crear una población

### 5.1 Tarea 7

Realiza simulaciones para comprobar si se mantienen iguales ambos alfas (teórico y empírico) cuando partimos de variables de rango limitado (y por tanto no normales). Crea un gráfico en el que se pueda ver si cambia el alfa empírico desde un rango de valores de: 2, 3, 4 y 5.

### 5.2 Tarea 8

Realiza simulaciones para comprobar cómo es la relación entre los dos alfas al aumentar el tamaño de la muestra (n = 2, 10, 15, 20). Haz un gráfico.

### 5.3 Tarea 9

Utilizando la distribución de Johnson estudia cómo afecta al contraste de una media que la distribución se vaya haciendo más y más asimétrica (utiliza dos o tres niveles de asimetría).

## 6 Errores tipo II

La evaluación de los errores tipo II ($\beta$) es más compleja, ya que dependen de tres elementos:

- El alfa teórico
- El tamaño del efecto
- El error típico de medida

Para la evaluación de $\beta$, o lo que es equivalente, la potencia, necesitamos dejar todo lo constantes que podamos estos tres elementos. Veamos cómo podemos estimar (recordamos que es una simplificación) una $\beta$ empírica ($\beta_e mp$).

En un contraste de una media, el tamaño del efecto se calcula simplemente como:

$$d = \frac{\mu_0 - media}{S_X}$$

Así que podemos fijar a priori un $\mu_0$, por ejemplo $\mu_0 = 0$ y generar k muestras que provengan de una distribución N(media, sd) donde la media es diferente a $\mu_0$, por ejemplo N(3, 2). Posteriormente, con cada una de las k muestras generadas podemos poner a prueba la $H_0 : \mu = \mu_0$ y, encontrar la beta empírica simplemente contando las veces que hemos aceptado la $H_0$. Para calcular la beta teórica podemos usar la librería pwr y seguir estas instrucciones pwr. Por ejemplo, en nuestros datos para conocer la potencia teórica

usaríamos la función *pwr.t.test()*. Como la función nos da el valor de la potencia para saber el valor de beta simplemente: $\beta = 1 - potencia$. Vamos a verlo con R:

```
# si no esta instalado pwr se instala
if (!require("pwr")) install.packages("pwr")
```

```
## Loading required package: pwr
```

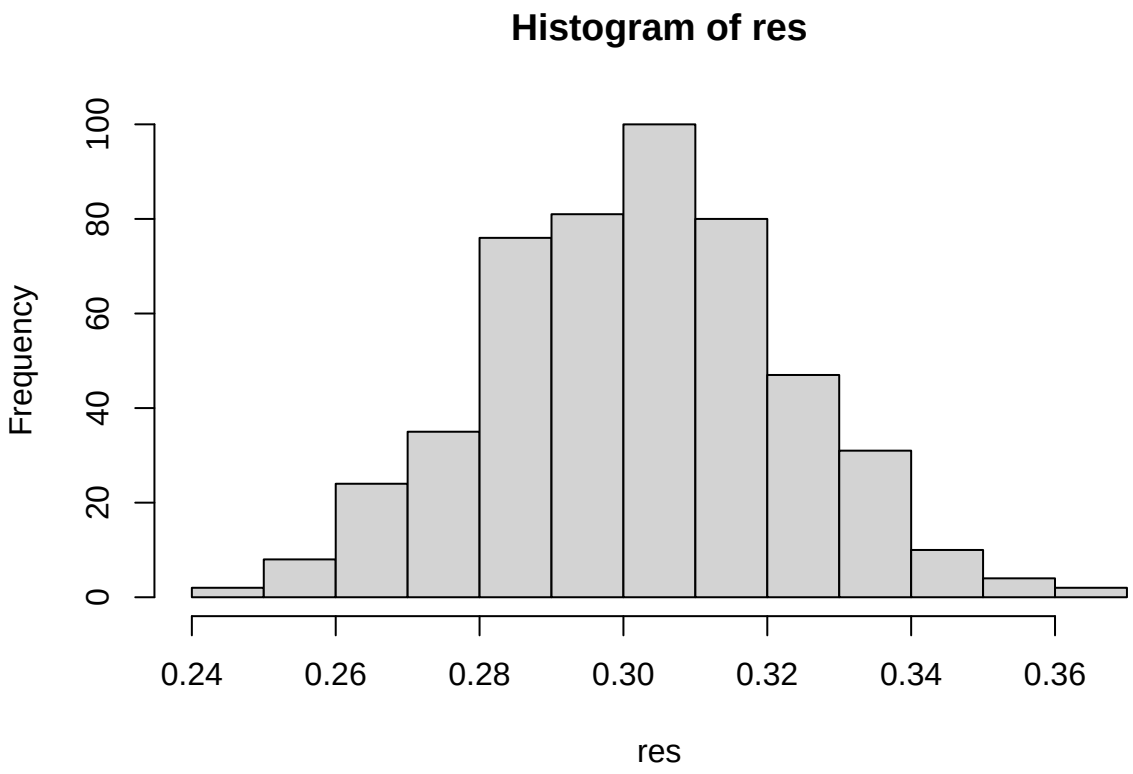
```
# parametros
N=1000
n = 25
k = 500
a_teo = 0.05

# creamos la poblacion
# set.seed(1)
poblacion <- rnorm(N, 10, 10)
m.pob <- mean(poblacion)
sd.pob <- sd(poblacion)

# simulamos los datos y recogemos el valor de pvalue
p <- vector(length=k)
for (i in 1:k){
  muestra <- poblacion[sample(1:N, n)]
  p[i] <- t.test(muestra, mu=15)$p.value
}

# calculamos beta empirica
beta_emp <- length(p[p>a_teo])/k

# calculamos beta teorica
d=(10 - 15) / sd.pob
beta_teo <- 1 - pwr.t.test(n,
                           d=d,
                           sig.level=a_teo,
                           type="one.sample")$power

# como las betas teóricas cambian al generarse nuevas poblaciones calculamos en nuestro experimento la
sprintf("beta_teo = %.3f -- beta_emp = %.3f -- Diff = %.3f", beta_teo, beta_emp, beta_teo-beta_emp)
```

```
## [1] "beta_teo = 0.351 -- beta_emp = 0.366 -- Diff = -0.015"
```

Y al igual que en el caso de los errores tipo I, lo que hemos obtenido es el un valor concreto resultado de un experimento. Si queremos caracterizar la distribución de resultados debemos repetirlo un número significativo de veces.

```
expto2 <- function(N,n,k){

  #poblacion <- sample(1:10, N, replace=T)
  poblacion <- rnorm(N, 10, 10)
  mu.pob <- mean(poblacion)
  sd.pob <- sd(poblacion)


  # se elige un valor cualquiera de mu pero que genere valores razonables de potencia
  # Si esta muy proximo a la media poblacional la potencia será 0 y si está muy lejos
```

```r
  # será 1 y no nos interesan esos casos.
  # Una estrategia es hacer que mu = mean(poblacion)+0.8*sd(poblacion)


  p <- vector(length=k)
  for (i in 1:k){
    muestra <- poblacion[sample(1:N, n)]
    p[i] <- t.test(muestra, mu=mu.pob+0.8*sd.pob)$p.value
  }

  # calculamos beta empirica
  beta_emp <- length(p[p>a_teo])/k

  # calculamos beta teorica
  d=(10 - 15) / sd.pob
  beta_teo <- 1 - pwr.t.test(n,
                             d=d,
                             sig.level=a_teo,
                             type="one.sample")$power
  return (beta_teo-beta_emp)
}

set.seed(1)
rep <- 500
res <- vector(length=rep)
for (i in 1:rep) res[i] <- expto2(1000, 25, 500)
hist(res)
```




```
mean(res)
```

```
## [1] 0.3017927
```

```
sd(res)
```

```
## [1] 0.0210436
```

Al igual que con los errores tipo I obtenemos una distribución normal centrada en el cero y con una desviación típica razonablemente alta. Los datos encontrados reflejan la distribución de diferencias entre los betas para unas "buenas" condiciones (cumplimiento de los supuestos de la prueba), ahora podemos estudiar cómo su incumplimiento a esa diferencia entre las betas empíricas y teóricas.

### 6.1 Tarea 10

Realiza simulaciones para comprobar si mantienen iguales ambos betas (teórica data por pwr y empírica) cuando partimos de variables de rango limitado (y por tanto no normales). Crea un gráfico en el que se pueda ver si cambia la diferencia entre las betas para un rango valores de 2, 3, 4 y 5.

### 6.2 Tarea 11

Realiza simulaciones para comprobar cómo es la diferencia entre las dos betas al aumentar el tamaño de la muestra (n = 2, 10, 15, 20). Haz un gráfico e interprétalo.

### 6.3 Tarea 12

Utilizando la distribución de Johnson estudia como afecta al contraste de una media que la distribución se vaya haciendo más y más asimétrica (utiliza dos o tres niveles de asimetría).

# 7 Generación de documentos con Rmarkdown

Cuando trabajamos en Ciencia de Datos es muy fácil encontrarse en uno de estos dos escenarios:

- Necesito escribir documentación técnica, con ecuaciones, resultados, gráficas y explicaciones de una manera rápida y efectiva, con calidad para ser enviada a publicaciones científicas o empresas. Por ejemplo, realizo simulaciones y quiero que los resultados se integren de forma automática en el *artículo* que voy a enviar a publicar.
- Necesito escribir masivamente informes con resultados de análisis para ser enviados a sus propietarios. Por ejemplo realizo evaluaciones psicológicas a alumnos de un centro escolar y necesito crear informes personalizados para cada uno de ellos.

Por suerte en R disponemos de herramientas para llevar a cabo ambas tareas de una manera muy eficaz. Respecto al segundo escenario disponemos de librerías que nos generan automáticamente documentos en pdf, word, excel, powerpoint, etc (por ejemplo reporteRs) o sweave, y el alumno puede leer por su cuenta y aprender a utilizar estos sistemas si lo necesita. Ahora, en este punto del curso, nos vamos a centrar en el primer escenario, en el que necesitamos generar informes de calidad con nuestros análisis.

En cuanto empiezas a trabajar con documentos científicos y técnicos te das cuenta de que las soluciones comerciales como word generan muchos problemas, haciéndonos perder una gran cantidad de tiempo y obteniéndose resultados de calidad baja. Por supuesto que hay sistemas de altísima calidad, como las distintas aplicaciones de edición profesional de adobe, pero conlleva mucho tiempo usarlas, son muy caras y no ayudan a la hora de trabajar con datos.

Desde el mundo de la ciencia la solución tradicional ha sido trabajar con un lenguaje llamado LaTeX, que permite definir tus documentos con código y obtener resultados de altísima calidad para imágenes y fórmulas. Por desgracia el aprendizaje de Latex es lento y complejo, por lo que no ha llegado a extenderse demasiado fuera de las ciencias duras (matemáticas, físicas, ingenierías, etc.).

Desde hace unos años (2004) se ha desarrollado un segundo lenguaje, creado por John Gruber, cuyo objetivo es ser muy sencillo de usar, de alta compatibilidad y calidad ,y que sea útil para una gran diversidad de objetivos, este lenguaje es Markdown. Markdown es muy sencillo de usar y se integra con muchas herramientas dedicadas al análisis de datos. Si Latex era el lenguaje de los físicos, Markdown es el lenguaje de los informáticos y los científicos de datos.

Aprender markdown es sencillo, y existen infinidad de documentos en internet que explica su sintaxis, por ejemplo, en la Figure 1 puede verse la diferencia entre escribir en html y hacerlo en markdown y se puede apreciar la simplicidad del código.

Este lenguaje tiene distintas adaptaciones dependiendo de dónde se va a utilizar. Lo que vamos a ver ahora es la versión de markdown creada por la empresa Rstudio para el trabajo con R, cuyo nombre es Rmarkdown. Rmarkdown es el lenguaje que estoy utilizando ahora mismo para escribir este documento.

Lo primero que hay que saber es que por defecto Rstudio permite guardar la ejecución de un fichero de script como un documento de html, pdf o word. Esto se hace a través de una herramienta llamada Knitr y es tan sencillo como pinchar en knit y elegir el tipo de resultado que deseamos, mira la Figure 2.

De esta forma podremos exportar nuestros resultados en el formato que queramos, pero para poder integrarlo con markdown tenemos que hacer algunas cosas más:

1. Crea un fichero nuevo y llámalo con la extensión *.Rmd*
2. Copia la siguiente cabecera en el principio del fichero

```
---
title: "Técnicas de Simulación"
subtitle: "03. Errores tipo I y tipo II"
author: "Miguel A. Castellanos"
linkcolor: "blue"
output:
```

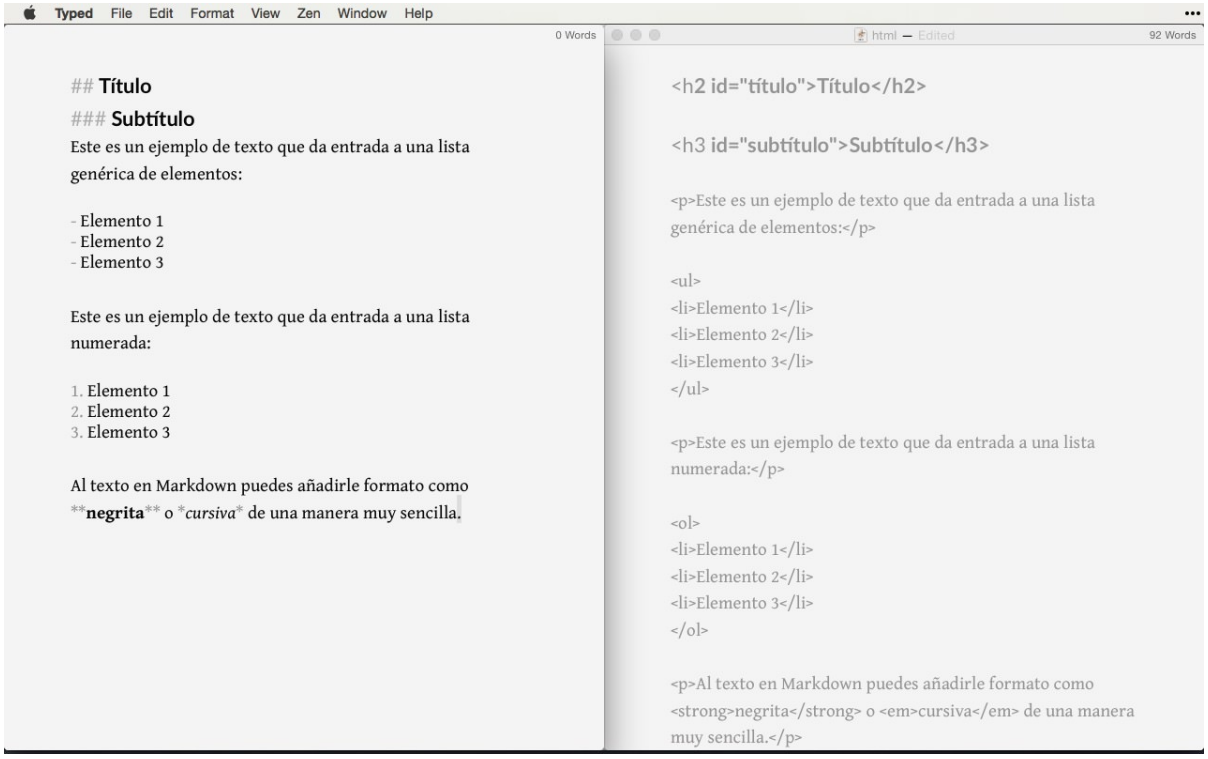



Figure 1: Makdown y html

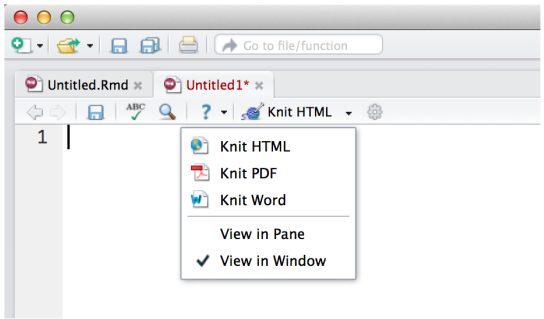



Figure 2: Knit en Rstudio

```
  pdf_document:
    toc: true
    toc_depth: 2
    number_sections: true
editor_options:
  chunk_output_type: console
---
```

3. Para que estén disponibles algunas opciones es posible que tengas que instalar una versión de LaTeXen tu ordenador. Si no, no podrás compilar y generar pdfs o tendrás problemas con las ecuaciones. Esta versión de Latex depende de tu sistema operativo pero en esencia las opciones están disponibles en la página de descarga del proyecto y son:

- TeX para Linux
- MacTeX para Mac OS
- MiKTeX para Windows

De las que yo he probado la que mejor funciona es la de Linux, la de Windows ocupa mucho espacio y tarda en instalarse, pero es el precio que hay que pagar por usar Windows.

4. Una vez instalado todo ya puedes generar tus propios informes en markdown, haciéndolos bonitos y rápidos. Para general el documento simplemente clica en el botón de **Knit** en Rstudio y el documento se generará automáticamente.

   consejo: Al principio, hasta que te acostumbras al formato de fórmula de LaTeXes cómodo usar un editor online

Os dejo junto con este documento un resumen con la sintaxis de Rmarkdown más habitual, también es posible descargarlo de internet en el siguiente enlace: Rmarkdown-cheatsheet.

### 7.1 Tarea 13

Reconvierte el documento que has preparado para esta práctica en un pdf con explicaciones utilizando Rmarkdown. Entrega dos documentos, uno con solo el código (en .R) y otro con el código integrado con los comentarios y explicaciones (en .pdf hecho con Rmarkdown)

Por supuesto no vas a poder instalar nada en los ordenadores públicos de la universidad, por lo que tendrás que hacer la instalación en tu propio ordenador, en cualquier caso si te vas a dedicar al análisis de datos es muy recomendable.

## 8 Referencias


Amón, J. (1996). Estadística para psicólogos II. Probabilidad. Estadística inferencial (9ª ed.) Madrid: Pirámide.

Field, A. (2009). Discovering statistics using SPSS (4th. Ed.), London: SAGE Publications.

Howell, D.C. (2014). Fundamental statistics for the behavioral sciences. Belmont, CA: Wadsworth.

Martínez-Arias, Chacón y Castellanos (2014). Análisis de datos en Psicología y Ciencias de la salud (volumen II). EOS universitaria.

Pardo, A., Ruíz, M. y San Martín, R. (2009). Análisis de datos en Ciencias Sociales y de la Salud I. Madrid: Síntesis.

Pardo, A. y San Martín, R. (1998). Análisis de datos en psicología. Vol. II. Madrid: Pirámide.

# 9 Pequeño glosario de términos

Alpha (). Nivel de significación o punto de corte que el investigador utiliza para rechazar la hipótesis nula. Probabilidad de error de Tipo I.

Asimetría negativa. Cualidad mostrada por una distribución de valores con gran concentración de valores a la derecha y cola larga y estrecha a la izquierda.

Asimetría positiva. Una distribución con gran concentración de puntuaciones en la zona de puntuaciones bajas y una cola larga a la derecha.

Contraste de hipótesis. Proceso por el que se toman decisiones con respecto a los valores de los parámetros.

Diseño de medidas repetidas. Diseño experimental en el que el mismo sujeto es observado bajo más de un tratamiento o condición.

Error de Tipo I. Error que se comete al rechazar una hipótesis nula que es verdadera.

Error de Tipo II. Error que se comete al no rechazar la hipótesis nula cuando es falsa.

Estimación por intervalos. Procedimiento de estimación que proporciona un rango o intervalo para el parámetro, asociado a una cierta probabilidad.

Hipótesis nula (H0). La hipótesis estadística que se pone a prueba por medio de un contraste estadístico. Con frecuencia establece la ausencia de diferencias o de relaciones.

Hipótesis. Una predicción o conjetura de la relación que existe entre las variables investigadas. También una proposición formal de la predicción del investigador de la relación que existe entre las variables bajo investigación.

Hipótesis alternativa (H1). Hipótesis estadística que propone que el parámetro en la población es un valor diferente del que establece la hipótesis nula. Es la hipótesis que se acepta cuando H0 es rechazada. Normalmente coincide con la hipótesis de investigación.

p-valor (p-value, en inglés). La probabilidad de que un resultado particular ocurra por azar si H0 es cierta. Es la probabilidad exacta de un error de Tipo I.

Potencia de un contraste (1-). Probabilidad de rechazar la hipótesis nula cuando es falsa o de aceptar la hipótesis alternativa cuando es cierta.

Tamaño de efecto. Una medida de la fuerza de una relación. Sirve para examinar la significación práctica. Indica el grado de separación de la hipótesis nula. Hay medidas basadas en diferencias de medias estandarizadas y en fuerza de la asociación (con valores entre 0 y 1).