UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Capítulo 8:

Intervalos de confianza

Erik Cobo, Belchin Kostov, Jordi Cortés, José Antonio González y Pilar Muñoz

Hector Rufino, Rosario Peláez, Marta Vilaró y Nerea Bielsa

Septiembre 2014

Departament d'Estadística i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA

equator network

UPC

ELSEVIER DOYMA

MEDICINA CLINICA

TRIALS

Intervalos de confianza

* Indica tema más avanzado que no es crucial para los ejercicios, aunque el lector debe recordar que aquí lo tiene —cuando lo necesite.

Presentación

El Intervalo de Confianza (IC) proporciona los valores del parámetro más compatibles con la información muestral. Para obtenerlos, tomaremos de R los valores de 2 nuevas distribuciones: la *t de Student* y la χ^2 (Ji Cuadrado).

Como el parámetro es un valor poblacional, se pretende conocer verdades absolutas y dar respuestas universales. Verdades universales, aunque reducidas a la población objetivo, con sus condiciones y criterios. En la perspectiva que presentamos, antes de hacer el estudio, cualquier valor del parámetro es teóricamente posible. Pero después del estudio, los contenidos en el IC son los más verosímiles. En resumen, los IC cuantifican el conocimiento, tanto sobre el auténtico valor, como sobre la incertidumbre que sobre él tenemos: mayor amplitud del intervalo, mayor imprecisión.

No es necesario que recuerde o aplique las fórmulas, pero SÍ que compruebe que sabe obtener con R los resultados e interpretar su significado.

Como siempre, no es necesario que entre a fondo en los puntos marcados con asterisco; pero SÍ que conviene que recuerde que aquí tiene la solución a ese problema por si alguna vez se le presenta.

Contribuciones: (1) la versión original de 2013 descansa en el libro de Bioestadística para No estadísticos de Elsevier de EC, JAG y PM y en el material de la asignatura de PE de la FIB (UPC); fue editada por BK y EC y revisada por RP y JC; (2) la de enero de 2014 fue revisada por JAG, RP, HR y MV para incorporar mejoras y sugerencias anónimas; y (3) la de septiembre de 2104 por NB y EC.

1. IC de μ con σ conocida*

En el capítulo anterior propusimos usar el valor de la media muestral como estimador puntual del parámetro poblacional, lo que venía avalado por ser la media muestral un estimador insesgado. Además, el error típico informaba sobre la oscilación o imprecisión (el "ruido") de la información (la "señal") aportada por la media muestral. Al final, con la ayuda de la distribución Normal, construimos un intervalo que contenía el 95% de las medias muestrales.

Pero a nivel práctico, conocemos X y queremos estimar μ. Es decir, la pregunta de interés es: conocido el estimador muestral media (X), ¿qué sabemos sobre la esperanza poblacional E(X) = μ?

Queremos un intervalo que informe, con una certeza cuantificable, dónde se encuentra el valor del parámetro. Para construirlo, recuperamos los valores $\pm Z_{\alpha\ 2} \cdot \sigma \quad \overline{n}$ que poníamos alrededor de μ; y cambiamos μ por X.

Nota: La Figura1.1 muestra gráficamente el efecto de sumar y restar la distancia $\pm Z_{\alpha\ 2} \cdot \sigma \quad \overline{n}$

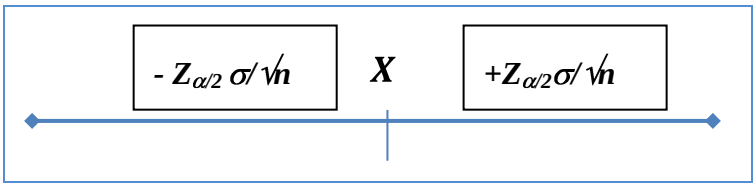


Figura1.1. Representación gráfica del IC.

Nota: A nivel práctico se puede coger tanto $Z_{\alpha/2}$ como $Z_{1-\alpha/2}$ dado la simetría de la distribución Normal. En el caso de un α del 5%, $Z_{0.025} = -1.96 = -Z_{0.975}$

La Figura 1.2 muestra el resultado de añadir esta distancia alrededor de 7 posibles medias muestrales X_i. Los intervalos de las medias 1 a 5 (X_1 a X_5), incluyen el valor μ del parámetro (línea vertical), es decir, aciertan, tal y como también lo harían todos los intervalos sobre medias contenidas entre los límites L_1 y L_2, que delimitan, precisamente, el 95% central de las medias muestrales.

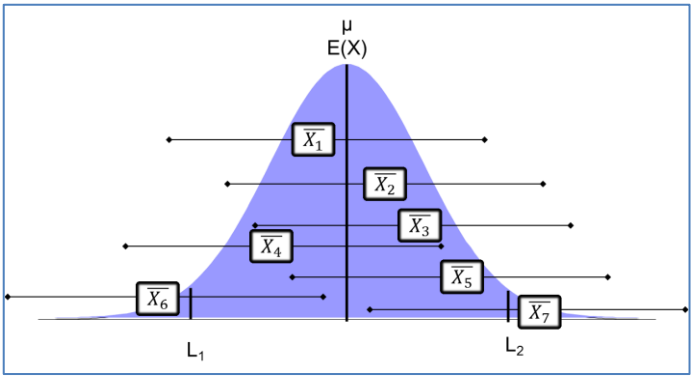


Figura1.2 Siete posibles medias muestrales y sus respectivos ICs.

En cambio, los intervalos de las medias de las muestras 6 y 7 (X_6 y X_7) no contienen el parámetro. Representan a ese α = 5% de posibles muestras que fallarían.

El intervalo así construido tiene, por tanto, un 95% de posibilidades de contener el parámetro poblacional, por lo que recibe el nombre de Intervalo de Confianza del 95% ($IC_{95\%}$).

Nota: Un 95% de confianza significa que (cabe esperar que), cada 20 estudios que se realicen, 19 contengan el parámetro de interés y 1 no lo haga**.**

Nota: Si desea aumentar la cobertura al 99% (α = 1%) o al 99.9% (α = 0.1%), simplemente se trata de sustituir el $Z_{0.975} = 1.96$ por los correspondientes cuantiles ($Z_{0.995} = 2.58$ y $Z_{0.9995} = 3.29$).

Recuerde

Este método requiere conocer la dispersión poblacional σ y por tanto es poco usado.

Ejemplo 1.1 (Prestado del control de calidad y de la vida misma): La asociación de usuarios (ASU) sospecha que las gasolineras no sirven la cantidad pactada. Por ley, se acepta que el dispensador tenga un error σ=10cc por cada litro que expende. En una muestra de n=100 pedidos de 1 litro (¡qué poco suspicaz el dependiente!), la media observada ha sido X = 995cc. El $IC_{95\%}$ de μ vale:

$$IC_{95\%}\ \mu\ = X \pm Z_{\alpha\ 2} \cdot \sigma \quad \overline{n} = 995 \pm 1.96 \cdot 10 \quad \overline{100} =$$

$$= 995 + 1.96 = \ 993.04\ , 996.96$$

Por tanto, se cree con una confianza del 95% que la auténtica media poblacional (μ) de esta máquina está entre 993cc y 997cc.

Ejemplo 1.2: La glicemia en mmol/L tiene una desviación típica igual a 1. En una muestra de 9 pacientes, la media ha sido de 5.

$$\mathrm{IC}_{95\%}\ \mu\ = X \pm Z_{\alpha\ 2} \cdot \sigma \quad \overline{n} = 5 \pm 1.96 \cdot 1 \quad \overline{9} \approx$$

$$\approx 5 + 2\ \ 3 = \ 4.34\ , 5.67$$

Se cree, con una "fuerza" del 95% que el auténtico valor poblacional se encuentra entre estos límites.

Esta fórmula para calcular el $IC_{95\%}$ de μ utiliza σ, lo que implica que, para poder estimar la media poblacional necesita conocer previamente la varianza de la variable. Esta situación es casi excepcional.

Ejemplo 1.3: La distribución de cierto parámetro sanguíneo sigue una N(μ,σ). Por un cambio del procedimiento analítico, se incrementan sus valores en una cierta constante K y se puede asumir que el nuevo valor sigue una N(μ',σ), que tenga una media desconocida y una varianza conocida.

Recuerde

El IC de μ conocida σ introduce el tema. Sólo se usa para predeterminar 'n'.

2. IC de μ con σ desconocida

¿Qué ocurre si σ es desconocida? De hecho, esta es la situación habitual. Ahora, para construir los intervalos de confianza, ya no usaremos esta versión del estadístico señal/ruido

$$Z = \frac{X - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N\ 0,1$$

sino en esta otra:

$$t = \frac{X - \mu}{\sqrt{\frac{S^2}{n}}}$$

Nota: Sustituir el parámetro σ por el estadístico *S* implica sustituir una constante, que tiene un único valor, por una variable aleatoria, que tiene toda una distribución de valores.

Cambiar σ por *S* tiene el precio de recurrir a una nueva distribución: la t de Student.

2.1. Distribución χ^2 (Ji o Chi cuadrado)

Antes de la distribución t de Student, necesitamos otra distribución, la χ^2.

Si X es N(0,1), su cuadrado, X^2, sigue una distribución de **Ji cuadrado** con 1 grado de libertad (GdL): $X^2 \sim \chi_1^2$

Al ser un cuadrado, todos sus valores son positivos.

Ejemplo 2.1: Sea *X* una v.a. N(0,1),

sabemos que P (X > 1.96) = P (X < -1.96) = 0.025

o también, que P (|X| > 1.96) = P (X > 1.96) + P (X < -1.96) =0.05

Por tanto P ($|X|^2$>1.96^2) = P (X^2> 3.84) =P (χ_1^2> 3.84)= 0.05

Ejemplo de R

```
# Cálculo de Fx: P(X<3.84) si X es una χ² con 1 GdL
> pchisq(1.96^2,df=1)
[1] 0.9500042
# Cálculo de x: P(X<=x)=0.2 si X es una χ²chi^2 con 1 GdL
> qchisq(0.2,1)
[1] 0.06418475
```

Sean ahora *n* variables aleatorias independientes idénticamente distribuidas (v.a.i.i.d) con distribución Normal centrada (μ=0) y reducida (σ=1):

$$X_1, X_1, \dots, X_n \sim N(0,1) \text{v.a.i.i.d}$$

entonces, la suma de sus cuadrados sigue una distribución de Ji cuadrado con *n* grados de libertad (GdL):

$$\sum_{i=1}^{n} X_i^2 = X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi_n^2$$

Esta distribución tiene una forma asimétrica que se reduce cuando aumenta el número de GdL, tal y como muestra la **Figura 2.1**.

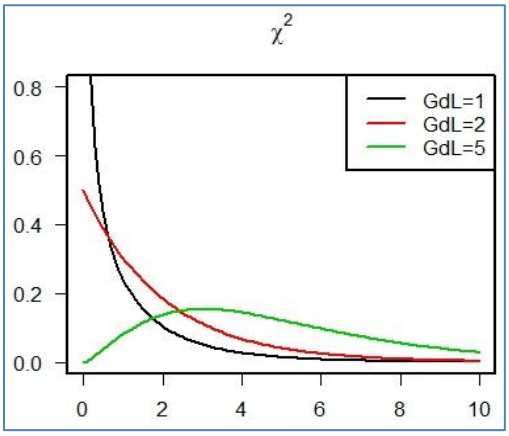


Figura 2.1. Distribuciones de χ_n^2 con 1, 2 y 5 GdL.

Ejercicio 2.1

Calcule con R las probabilidades P(X≤1), P(X≤3) y P(1≤X≤3) si X ~χ_3^2

2.2. Distribución T de student

A partir de las distribuciones Normal y Ji-Cuadrado, se puede obtener la distribución t de Student. $Si\ Z\ es\ N\ 0,1\ ,\ Y_n$ es χ_n^2 independientes, entonces $T = \frac{Z}{\sqrt{Y\ n}} \sim t_n$ y se dice que T sigue una distribución **t de Student** con *n* GdL.

Ejemplo 2.2. Sea t una v.a. con distribución t de Student con 14 GdL (t ~ t_{14}). La probabilidad de que t pueda tomar valores inferiores a -2.5 es P (t<-2.5)= 0.012. Asimismo, P (t>2.5)= 0.012. Y el valor de t que deja por debajo una probabilidad de 0.025 es -2.14.

Ejemplos de R

```
# Sea X una t de Student con 14 GdL
# P(X<-2.5)
> pt(q=-2.5,df=14)
[1] 0.01273333
# P(X>2.5)
> pt(q=2.5,df=14,lower.tail=FALSE)
[1] 0.01273333
# P(X<x)=0.025
> qt(p=0.025,df=14,lower.tail=TRUE)
[1] -2.144787
```

La t de Student es simétrica alrededor de cero, muy parecida a la normal, especialmente para valores grandes de GdL.

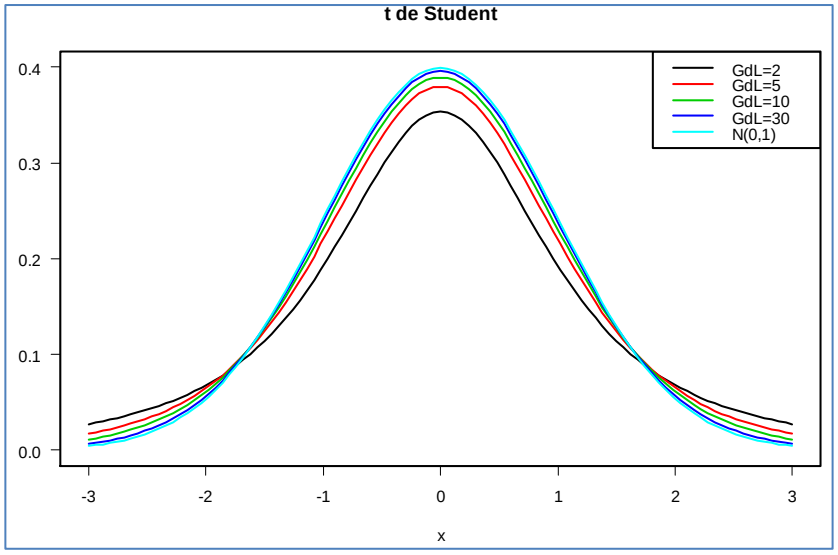


Figura2.2. La distribución Normal y t de student con diferentes GdL (2,5,10 y 30)

La distribución 't' se aplana y se aleja más de la distribución Normal cuanto más pequeña sea la muestra.

Nota: Gosset era el responsable de calidad de la cervecera Guiness. Para detectar los lotes que no cumplían con las especificaciones deseadas, él había aceptado el coste de rechazar un 5% de los que sí que las cumplieran, para lo que utilizaba los límites –1.96, +1.96. Pronto sospechó que desechaba demasiados: fuera de estos límites había más del 5% de los lotes correctos. Cayó en la cuenta de que S era un estadístico y no un parámetro y propuso una distribución algo más aplanada que la Normal, en la que observó que rechazaba el α% deseado de lotes correctos. Recibe este nombre porque lo **firmó** con el seudónimo de "estudiante" —dicen que porque Guiness no quería que se supiera que estudiaban su calidad..

Ejercicio 2.2.

Sea t una variable aleatoria con distribución t de Student con 12 grados de libertad ($t \sim t_{12}$). Encuentre la probabilidad de P (t > 1.796).

La simetría de la t de Student permitirá trabajar de forma simétrica.

Ejemplo de R

```
# Para calcular t19,0.025 y t19,0.975 en R
> qt(p=0.025,df=19)
[1] -2.093024
> qt(p=0.975,df=19)
[1] 2.093024
```

2.3. IC de μ usando S

La t de Student permite construir IC para μ desconociendo σ^2.

Fórmula

El **Intervalo de Confianza de (1-α)% de μ** , sin conocer σ es:

$$IC_{1-\alpha}\ \mu\ = X \pm t_{n-1,\alpha\ 2} \cdot \frac{S}{\sqrt{n}}$$

Ejemplo 2.3: El tiempo utilizado en la atención al paciente sigue una distribución Normal. Para conocer el tiempo medio empleado en este servicio, se han recogido 20 observaciones que han tardado, en minutos, $\bar{X} = 34$ y S=2.3.

$$IC\ \ 0.95, \mu\ \ = x \pm t_{19,0.025} \cdot \frac{S}{\sqrt{n}} = 34 \pm 2.093 \cdot \frac{2.3}{\sqrt{20}} \approx 34 \pm 1.08 = \ \ 32.92,35.08$$

Se cree, con una confianza del 95%, que la media poblacional del tiempo de atención se sitúa entre 32.92 y 35.08 minutos.

Recuerde

La amplitud del IC valora la ignorancia o incertidumbre sobre el único y auténtico valor de la esperanza μ. No indica que μ oscile ni que tenga más de un valor.

Ejercicio 2.3

Sin cambiar la confianza, ¿cómo podría reducir el intervalo del Ejemplo 1.2 a la mitad?

Ejercicio 2.4

Con los datos del Ejemplo 1.2, calcule el IC para una confianza del 99%.

Ejercicio2.5

Al final, ¿el $IC_{95\%}$ contiene o no contiene μ?.

Ejercicio2.6

El $IC_{99\%}$ (elija una):

a) incluye el 99% de las medias poblacionales

b) incluye el 99% de las medias muestrales

c) incluye la media poblacional el 99% de las ocasiones

d) incluye la media muestral el 99% de las ocasiones

Ejercicio 2.7

Con un $IC_{95\%}$ ($1-\alpha=95\%$) de μ podemos afirmar que (elija una):

a) el 95% de los casos están dentro del intervalo.

b) si se repitiera el proceso, el 95% de los casos estarían dentro del intervalo.

c) hay una probabilidad del 5% de que el parámetro μ no esté en el intervalo.

d) hay una confianza del 95% de que el parámetro μ esté en el intervalo.

Ejercicio 2.8

Asumiendo que la desviación típica poblacional de las GOT (Transaminasa Glutámico Oxalacética) es de 120 u, ¿cuántos casos se necesitan para...

...tener un error típico de estimación de μ ($\sigma/\sqrt{n}$) igual a 12 u?

...tener una semi-amplitud del $IC_{95\%}$ de μ ($Z_{0.975}\sigma/\sqrt{n}$) igual a 12 u.?

...tener una amplitud total del $IC_{95\%}$ de μ ($\pm Z_{0.975}\sigma/\sqrt{n}$) igual a 12 u.?

Nota técnica: En la estadística clásica, no bayesiana, el parámetro es una constante, no una variable aleatoria. Por ello, se evita hablar de un intervalo de probabilidad del parámetro y se usa el término de confianza. Desde esta perspectiva sólo puede usarse probabilidad en lugar de confianza si queda claro que las variables aleatorias son los extremos del intervalo. En otras palabras, no decir que entre los límites a y b del intervalo se encuentre un parámetro "flotante" con alta probabilidad, como si a y b fueran fijos, sino que el procedimiento del IC garantiza con alta probabilidad que el parámetro esté entre los dos valores aleatorios a y b.

2.3.1. Premisas para estimar μ sin conocer σ

Nota: Para referirse al término inglés *assumptions*, diferentes autores utilizan diferentes vocablos: asunciones, hipótesis previas necesarias, requisitos, condiciones de aplicación... Como dijimos en el capítulo 1, usamos "*premisas*" para resaltar su papel secundario y diferenciarlas de las hipótesis, que aunque también son supuestos, reflejan el objetivo del estudio.

Para poder afirmar que el estadístico t sigue una t de Student con n-1 GdL, la premisa necesaria es que la variable en estudio X siga una distribución Normal. Ahora bien, aunque no sea Normal, si el tamaño muestral crece, la estimación S^2 de σ^2 mejora, acercándose al valor real, por lo que la sustitución de σ^2 por S^2 tiene menores implicaciones. Por esta razón, aunque la variable estudiada no sea Normal, en estudios grandes puede usarse la Normal.

Recuerde

La fórmula requiere: o bien que X sea Normal; o bien que $n \geq 30$.

Nota: ¿Qué significa tamaño grande? ¿Por qué unos autores dicen 20, otros 30 y otros 100? ¿Hay algún número mágico que cambie tanto la forma de la distribución? No, se trata de una aproximación sucesiva y se necesitará menos muestra cuanto más se asemeje X a la Normal.

Así pues, se sabe cómo inferir los resultados de la muestra a la población si se dispone de una variable Normal; o bien si la muestra es suficientemente grande. Estas fórmulas deben servir para solucionar la gran mayoría de las situaciones.

Ejercicio 2.9

En una muestra de 100 pacientes con infarto, se ha valorado la Transaminasa Glutámico Oxalacética (GOT) a las 12 horas. La media ha sido de 80 y la desviación típica de 120. Haga un $IC_{95\%}$ de la media.

Nota: Se pide un tamaño muestral mayor que 30 para poder usar una fórmula estadística. Pero en un estudio clínico, el tamaño muestral debe fijarse por la cantidad de información que se desea disponer.

Lectura: En el caso de que no disponga de una muestra grande ni de una variable con distribución Normal se puede recurrir a dos grandes grupos de soluciones: 1) métodos estadísticos que no requieren esta distribución (cálculos exactos o por re-muestreo, principalmente); y 2) transformar la variable para conseguir su Normalidad. Existen varias transformaciones que funcionan muy bien en la práctica. Para variables positivas (como "el tiempo hasta..." o "el nivel de GOT") la transformación logarítmica suele corregir la habitual asimetría y conseguir distribuciones muy parecidas a la Normal. Por otro lado, si se dispone de un recuento de fenómenos raros, de baja probabilidad, que suelen seguir una distribución de Poisson, la transformación raíz cuadrada suele funcionar bien.

Ejemplo de R

```
# Dada una muestra X, con t.test se obtiene el IC de la
# de la media poblacional:
> X <- c(110,100,115,105,104)
> t.test(x=X,conf.level=0.95)
              One Sample t-test
data:  X
t = 41.1378, df = 4, p-value = 2.087e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
99.59193 114.00807
sample estimates:
mean of x
106.8
#Se cree, con una "fuerza" del 95%, que el auténtico valor poblacional
se encuentra entre [99.59 , 114.01].
```

2.3.2. Estimación auto-suficiente (*bootstrap*) *

La informática permite sistemas alternativos de estimación que descansan en menos premisas. El más conocido consiste en generar sub-muestras al azar de la muestra obtenida y, a partir de ella, deducir la distribución del estimador muestral.

Ejemplo 2.4: Un estudio sobre acupuntura emplea el índice BDI-II (Inventario de Depresión de Beck-II) para evaluar la gravedad de la depresión. Como esta variable no se ajusta bien a la normal, decide calcular el IC(μ) con Bootstrap .

Ejemplo de R

```
#Instalar paquete
>install.packages("bootstrap")
#Cargar paquete
>library("bootstrap")
##-- IC para una media  (BDI-II)
#Semilla
>set.seed(123)
#Tamaño de la muestra
>n<-755
```

```
#Generación de la muestra (BDI-II)
>x <- runif(n,0,65)
#Parámetro para el que quiere calcular el IC
>theta <- function(x){mean(x)}
#Bootsrap con 1000 repeticiones
>results <- bootstrap(x,1000,theta)
# Cálculo del IC
>IC <- quantile(results$thetastar,c(0.025,0.975))
>IC
    2.5%    97.5%
31.12718 33.93205
#Siendo la media observada en la muestra 32.5
```

Por tanto, la interpretación será: "mediante un método de bootstrap, libre de premisas sobre la forma de la distribución de la variable, la estimación puntual de la media poblacional es 32.5, con una incertidumbre ($IC_{95\%}$) desde 31.1. a 33.9.

Dado que este método genera submuestras al azar, diferentes ejecuciones, pueden originar diferentes resultados. Para garantizar que no se ha escogido el resultado más conveniente (una variante del "outcome selection bias"), conviene especificar en el protocolo la semilla que generará las sub-muestras y el programa para obtener y analizar los datos.

Ejemplo 2.4 (cont): Veamos ahora cómo calcular el IC del coeficiente de correlación de, por ejemplo, el índice BDI-II y la edad a la que el individuo sufrió el mayor episodio de depresión.

Ejemplo de R

```
##-- IC para una la correlación
# Tamaño muestral
n <- 755
#Semilla
set.seed(123)
#Generación de y1 (BDI-II)
y1 <-  runif(n,0,65)
#Generación de y2 (Edad de mayor episodio de depresión)
y2 <-  rnorm(n,22.5,12.28)
#Unimos y1 e y2 en un data.frame
xdata <- matrix(c(y1,y2),ncol=2)
#Parámetro para el que se quiere calcular el IC (en este caso, coef, de
correlación)
theta <- function(x,xdata){cor(xdata[x,1],xdata[x,2])}
# Bootstrap con 1000 repeticiones
results <- bootstrap(x=1:n ,1000,theta,xdata)
#Cálculo del IC
```

```
IC <- quantile(results$thetastar,c(0.025,0.975))
IC
      2.5%       97.5%
-0.05202905  0.08796262
```

Recuerde:

Si no se cumplen las premisas, valore emplear el método *bootstrap*.

2.3.3. Interpretación y uso de la transformación logarítmica *

Algunas variables sólo pueden tomar valores positivos y son muy asimétricas.

Ejemplo 2.5: El salario, que por ahora aún no es negativo, cumple el modelo de Pareto: "el 80% de Italia está en manos del 20% de los italianos". Los aumentos de sueldo no se negocian de forma aditiva o lineal (100€ más para todos), sino multiplicativa: un "5% más" significa multiplicar por 1.05. Y, en matemáticas, las multiplicaciones 'piden' logaritmos.

Definimos Y como la transformación logarítmica (natural, neperiana o de base e) de la variable X. Es decir, Y=log(X). Obtendremos los estadísticos de Y, haremos su IC y, a partir de él, obtendremos el IC de X mediante la operación inversa.

Nota técnica: $\exp\{Y\} = e^Y$ indica el número e = 2.7183 elevado al número Y. La operación matemática EXP y log son inversas: $e^{\ln(y)} = Y$; $\ln(e^Y) = Y$. El lector no debe desanimarse por la aparición de unos logaritmos a los que no está habituado. Piense que son tan solo un instrumento para dar simetría a las variables. Recuerde que el pH no tiene secretos para Vd: Es cómodo valorar la acidez con el pH, aunque sea el logaritmo de la concentración de hidrogeniones.

Definición:

Sea *Y*=log(*X*)

$$IC_{1-\alpha}\ \mu_Y\ = Y \pm t_{n-1,\alpha/2} \cdot S\quad \overline{n}$$

$$IC_{1-\alpha}\ \mu_X\ = exp\ IC_{1-\alpha}\ \mu_Y$$

Ejemplo 2.6 (cont. del Ejemplo 2.3): La media del logaritmo (Y) del tiempo utilizado en la atención al paciente (en la muestra de 20 pacientes) es de $\bar{y}$ = 3.55 y su desviación estándar S=0.069. Como *Y* sigue razonablemente bien la Normal, el $IC_{95\%}$ de μ_Y es:

$$IC_{95\%}\ \mu_Y\ \approx y \pm t_{n-1,\alpha\ 2} \cdot \frac{S}{n} \approx 3.55 \pm 2.09 \cdot \frac{0.07}{20} \approx 3.55 \pm 0.03 \approx\ 3.52{,}3.58$$

Para facilitar la interpretación se deshace el logaritmo mediante la función exponencial. La estimación puntual de μ_X es $e^{3.55}$= 34.81 y por intervalo:

$$IC_{95\%}\ \mu_X = \exp\ IC_{95\%}\ \mu_Y = e^{3.52}, e^{3.58} = 33.71, 35.95$$

Los resultados son muy similares a los originales, $IC_{95\%}$= [32.92, 35.08]. Es bueno que, independientemente de las premisas de salida, obtengamos conclusiones similares. Ahora la simetría ocurre en una escala multiplicativa: 35.95=34.81·1.03; y 33.71=34.81/1.03. Es decir, la imprecisión obliga a multiplicar y dividir por 1.03.

3. IC de σ^2*

El IC se basa en que, si X es N, S^2 multiplicada por (n-1) y dividida por la varianza poblacional sigue una distribución Ji cuadrado: $\frac{S^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$

Fórmula

El **Intervalo de Confianza (1-α)% de σ^2** es:

$$IC_{1-\alpha}\ \sigma^2 = \frac{S^2\ n-1}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{S^2\ n-1}{\chi^2_{n-1,\frac{\alpha}{2}}}$$

Premisa: X~N

Ejemplo 3.1: El tiempo observado hasta la desaparición de un signo en 25 pacientes ha mostrado una variabilidad $S^2=8^2$ min^2. ¿Qué sabemos sobre el auténtico valor de la varianza poblacional?

$$IC_{95\%}\ \sigma^2 = \frac{64(25-1)}{\chi^2_{n-1,1-\alpha/2}}, \frac{64(25-1)}{\chi^2_{n-1,\alpha/2}} = \frac{1536}{39.36}, \frac{1536}{12.4} = 38.98, 123.87$$

Por tanto, habiendo observado una varianza muestral S^2=64 min^2, sabemos sobre la varianza poblacional σ^2 que, con una confianza del 95%, es alguno de los valores comprendidos entre 38.98 min^2 y 123.87 min^2. Dos aspectos resaltan: la asimetría del intervalo alrededor de la estimación puntual (64) y su gran magnitud: aunque la muestra no es muy pequeña (n=25), el grado de incertidumbre parece notable. Para evitar tener que interpretar "minutos cuadrados", haremos su raíz:

$$IC_{95\%}(\sigma) \approx [6.24, 11.13]$$

El intervalo sigue siendo asimétrico alrededor de la estimación puntual, que era 8. Y sigue pareciendo grande (el extremo superior casi dobla al inferior). Pero esta impresión ya no es tan exagerada. Lo que no hay duda es que ahora, sin cuadrados, es más fácil interpretarlo: con una confianza del 95%, la desviación típica poblacional es algún valor comprendido entre 6.24 min y 11.13 min.

Ejemplo de R

```
# R no dispone de ninguna función específica para calcular este
# intervalo. Podemos crearla nosotros
> IC_var <- function(x,confidence){
S2 <- var(x)           # Varianza muestral
  alfa <- 1-confidence  # Nivel de significación
  n <- length(x)        # Tamaño muestral
  X1 <- qchisq(p=1-alfa/2,df=n-1)  # Valor de Ji cuadrado 1
  X2 <- qchisq(p=alfa/2,df=n-1)  # Valor de Ji cuadrado 2
  LI <- (S2*(n-1))/X1     # Limite Inferior
  LS <- (S2*(n-1))/X2    # Limite Superior
  return(c(LI,LS))      # Retorna el Intervalo
}
# Ejemplo con una muestra de 5 valores
> PAS <- c(128,102,126,116,100)
> conf <- 0.95
> IC_var(PAS,0.95)
[1]   61.31046 1410.35059
# Y el intervalo de confianza de la desviación típica :
> sqrt(IC_var(PAS,0.95))
[1]  7.83010 37.55463
```

Ejercicio 3.1.

Preguntados por el nº de asignaturas matriculadas, 4 alumnos han contestado: 2, 3, 4 y 5. Con la función de R anterior, calcule S^2 y S y estime σ^2 y σ.

Nota técnica: Los GdL o la información "neta" de una muestra vienen dados por el número de observaciones (independientes) menos las preguntas que previamente ha debido contestar. Por ejemplo, si para calcular S^2 en una muestra de n casos primero se ha debido estimar 1 parámetro μ mediante x, los

GdL que tiene esta estimación de la varianza son "n-1". Más formalmente, un sistema de *n* ecuaciones (piezas de información) con k incógnitas tiene n-k GdL.

4. IC de la diferencia de 2 medias

4.1. Muestras independientes

Fórmula

El **Intervalo de Confianza (1-α)% de $\mu_1 - \mu_2$** en muestras independientes es:

$$IC_{1-\alpha}\ \mu_1 - \mu_2 \ = \ y_1 - y_2 \ \pm t_{n-2,\alpha\ 2} \cdot \ \sqrt{S^2\ \frac{1}{n_1} + \frac{1}{n_2}}$$

Tenemos sólo una S^2 porque asumimos igualdad de varianzas (homoscedasticidad: $\sigma^2_1 = \sigma^2_2 = \sigma^2$); y entonces, S^2_1 y S^2_2 estiman el mismo parámetro σ^2 mediante la ponderación de S^2_1 y S^2_2 según sus GdL.

Fórmula

La **estimación conjunta ("*pooled*") de la varianza** en 2 muestras se calcula:

$$S^2 = \frac{n_1 - 1\ S_1^2 + \ n_2 - 1\ S_2^2}{n_1 + n_2 - 2} = \frac{\sum_{i=1}^{n_1} y_{1i} - y_1 \ + \ \sum_{i=1}^{n_2} y_{2i} - y_2}{n_1 + n_2 - 2}$$

Nota: observe que esta ponderación acaba siendo la fórmula de siempre de la varianza: la suma de todas las distancias a su propia media, dividida por sus GdL.

Recuerde

La fórmula del $IC_{1-\alpha}\ \mu_1 - \mu_2$ requiere:

(i) MAS independientes

(ii) Varianzas (desconocidas) iguales: "*homoscedasticidad*"

(iii) $Y_1 \sim N$; $Y_2 \sim N$

Ejemplo 4.1: Para comparar 2 intervenciones, usamos el tiempo medio hasta la desaparición de un signo en 2 grupos de pacientes en condiciones independientes (n_1=50 y n_2=100). Los resultados son: $y_1 = 24\ y\ y_2 = 21$, siendo S_1=8 y S_2=6. Suponiendo MAS independientes y con varianzas poblacionales iguales, encuentre el $IC_{95\%}$ de $\mu_1 - \mu_2$.

$$S^2 = \frac{n_1 - 1\ S_1^2 + \ n_2 - 1\ S_2^2}{n_1 + n_2 - 2} = \frac{49 \cdot 8^2 + 99 \cdot 6^2}{148} \approx 44$$

$$IC_{95\%}\ \mu_1 - \mu_2 \ = \ y_1 - y_2 \ \pm t_{148,0.025} \cdot \ \sqrt{S^2\ \frac{1}{n_1} + \frac{1}{n_2}} \ \approx 3 \pm 1.976 \cdot 1.149 \approx$$

$$\approx 3 \pm 2.27 \approx \ 0.73,5.27$$

Ejemplo de R

```
# Cálculo del valor de t con 148 GdL
> qt(p=0.025,df=148,lower.tail=FALSE)
[1] 1.976122
```

Nota: La homoscedasticidad o estabilidad de las varianzas aparece cuando el efecto se concentra en los valores medios: lo que sucede cuando el cambio de tratamiento produce el mismo efecto en todos los casos y hace relevante a todas las unidades el efecto poblacional medio. Aunque la igualdad de varianzas poblacional no es directamente observable, sí lo es el nivel de similitud de los valores muestrales.

Ejemplo de R

```
# Dadas dos muestras indep. x e y, la función t.test da el IC de μ1−μ2

> x <- c(1,5,6,8,10)
> y <- c(2,7,11,1,12,3,4)
> t.test(x,y,var.equal=TRUE)
Two Sample t-test
data:  x and y
t = 0.1214, df = 10, p-value = 0.9057
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.956190  5.527618
sample estimates:
mean of x mean of y
6.000000  5.714286
```

4.2. Muestras apareadas

En ocasiones, las unidades de las dos muestras para las cuales quiere calcular la diferencia de medias se encuentran emparejadas por algún factor. El caso más habitual podría ser el de un conjunto de pacientes en qué se mide una variable en el momento basal del estudio y en una visita posterior. En este caso, tenemos las 2 muestras (basal y visita posterior) emparejadas por cada paciente. Para el cálculo del IC en muestras apareadas, se calcula primero la variable diferencia $D_i = Y_{iA} - Y_{iB}$ y luego se aplica el método del cálculo del IC de μ para una muestra.

Fórmula

El **Intervalo de Confianza de (1-α)% de μ_1 - μ_2** en muestras apareadas es:

$$IC_{1-\alpha}\ \mu_1-\mu_2\ = D \pm t_{n-1,\alpha\ 2}\cdot \frac{\overline{S_D}}{n}$$

Recuerde

La fórmula requiere:

(ii) MAS apareadas

(iii) D~N

Ejemplo 4.2: Las 2 intervenciones anteriores, A y B, se han probado en los 6 mismos pacientes y los tiempos hasta la desaparición del síntoma han sido:

	Pac. 1	Pac. 2	Pac. 3	Pac. 4	Pac. 5	Pac. 6	y_j	S_j^2	S^2
Y_{iA}	23.05	39.06	21.72	24.47	28.56	27.58	27.406	39.428	42.009
Y_{iB}	20.91	37.21	19.29	19.95	25.32	24.07	24.460	44.591	
$D_i = Y_{iA}.Y_{iB}$	2.13	1.85	2.43	4.51	3.24	3.51	2.946	0.996	

Si consideramos las 2 muestras como independientes (solución errónea) el $IC_{95\%}$ es:

$$IC_{95\%}\ \mu_1-\mu_2\ =\ y_1-y_2\ \pm t_{\alpha\ 2,n-2}\cdot \sqrt{S^2\ \frac{1}{n_1}+\frac{1}{n_2}}\ =$$

$$=\ 27.41-24.46\ \pm 2.23\cdot \sqrt{42.01\cdot \frac{1}{6}+\frac{1}{6}}\ = 2.95 \pm 8.34 =\ -5.39,11.28$$

En cambio, el $IC_{95\%}$ correcto, considerando las 2 muestras apareadas es:

$$IC_{95\%}\ \mu_1-\mu_2\ = D \pm t_{n-1,\alpha\ 2}\cdot \frac{\overline{S_D}}{n} = 2.95 \pm 2.57 \cdot 0.41 = 2.95 \pm 1.05 =\ 1.90\ ,4.00$$

Así, el cálculo erróneo previo provoca una estimación demasiado alta de la imprecisión, y daba un IC con el valor 0 de no diferencias en su interior.

Observe que la varianza de la variable diferencia (0.996) es muy inferior a la "*pooled*" (42.009), indicando el beneficio de hacer un diseño con datos apareados. Enseguida explicaremos sus razones, que no son más que eliminar la variabilidad compartida.

Ejemplo de R

```
# Dadas dos muestras apareadas x,y t.test y paired=TRUE
# dan el IC de la diferencia de μ en muestras apareadas
> ?sleep
> data(sleep)
> t.test(extra~group,data=sleep,paired=TRUE)
   Paired t-test
data:  extra by group
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean of the differences
-1.58
```

Ejercicio 4.1.

Calcular, con R, el IC de la diferencia de las medias de Y_A y Y_B

Y_A = 23.05,39.06,21.72,24.47,28.56,27.58

Y_B = 20.91,37.21,19.29,19.95,25.32,24.07

(i) Considerando que son muestras independientes.

(ii) Considerando que son muestras apareadas.

(iii) Compare los errores típicos de ambos e interprete.

Si no se puede asumir que las varianzas sean iguales aparecen dos dificultades. La primera es práctica: la diferencia de las medias ya no representa un efecto común para atribuir a cada caso. La segunda es técnica: el estadístico ya no sigue una t de Student. Encontrar una transformación de *Y*, en que las varianzas sean iguales y la distribución normal soluciona ambos problemas.

5. IC del coeficiente de correlación de Pearson (ρ) *

La covarianza y la correlación indican la relación entre 2 variables numéricas X, Y:

	Población	Muestra
Covarianza	σ_{XY}	S_{XY}
Correlación	ρ_{XY}	r_{XY}

Tabla 5.1. Nomenclatura para covarianza y correlación.

La covarianza indica el grado de variación conjunta entre las 2 variables. A nivel muestral, la covarianza se calcula de forma muy similar a la varianza:

$$S_{XY} = \frac{\sum_{1=1}^{n} x_i - x \quad y_i - y}{n-1}$$

Ejercicio 5.1.

Imagine la covarianza de una variable X consigo misma. ¿En qué se convierte la formula anterior de la covarianza al aplicarla a X con X: S_{XX}?

La covarianza tiene las unidades de medida de ambas variables, por lo que conviene definir un coeficiente que pueda ser interpretado de la misma forma para cualquier unidad de medida. El coeficiente de correlación lineal "tipifica" la covarianza dividiéndola por sus desviaciones típicas. A nivel muestral, se calcula:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{1=1}^{n} x_i - x \quad y_i - y \quad n-1}{S_X S_Y} = \frac{\sum_{1=1}^{n} x_i - x \quad y_i - y}{\sum_{1=1}^{n} x_i - x \ ^2 \quad \sum_{1=1}^{n} y_i - y \ ^2}$$

La correlación varía entre -1 y +1, donde el signo indica la dirección de la relación: directa (si es positivo) o inversa (si es negativo). La magnitud mide la intensidad de la relación. $r_{XY} = 0$ indica ausencia de relación lineal. En cambio, $r_{XY} = 1$ o $r_{XY} = -1$ indica una relación lineal 'perfecta' que se puede representar mediante una recta $Y = a + bX$ (Figura 5.1).

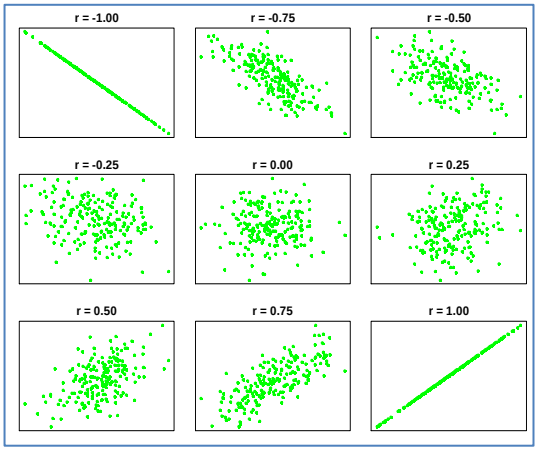


Figura 5.1. Ejemplos de diferentes grados de correlación entre dos variables X y Y

Ejercicio 5.2

Recupere los datos del capítulo 3 sobre peso del recién nacido y semana de gestación.

A) Mirando sus gráficos y la figura anterior, ¿qué correlación adivina entre ambas variables? (a qué figura se parece más?)

B) Suponga que ha decidido estudiar sólo los partos a término (≥ 38 semanas): ¿cuál cree que es ahora el valor de la correlación?

C) Busque en R el comando para obtener el coeficiente de correlación y obténgalo para las preguntas A y B (recuerde que puede seleccionar casos mediante, por ejemplo, el comando subset(data.frame, concidición lógica), en este caso subset(births,births$gestwks>=38)).

D) ¿Por qué cree que han dado diferente las correlaciones para las 2 situaciones anteriores?

Recuerde

Si reduce la "ventana" de su estudio restringiendo una variable, disminuirá su variabilidad y las posibilidades de observar relación con otras variables.

El $IC_{95\%}$ del coeficiente de correlación lineal se puede estimar de diferentes maneras aunque lo más habitual es hacerlo mediante la transformación de Fisher.

Recuerde

El Intervalo de Confianza de (1-α)% del coeficiente de correlación (ρ) se obtiene mediante una fórmula de la que sólo debe recordar que genera intervalos asimétricos y permite usar la D. Normal..

Nota: la transformación de Fisher es: $tanh\ \ arctanh\ r\ -\frac{Z_{\alpha\ 2}}{n-3}\ \ ,tanh\ \ arctanh\ r\ +\frac{Z_{\alpha\ 2}}{n-3}$

Usaremos R para obtener e interpretar los resultados.

Ejemplo de R

```
# Consideramos las dos variables X y Y
> X<-c(23.05,39.06,21.72,24.47,28.56,27.58)
> Y<-c(20.91,37.21,19.29,19.95,25.32,24.07)
# Coeficiente de correlación y su IC95%
> cor.test(X,Y)
```

```
        Pearson's product-moment correlation
data:  X and Y
t = 14.0386, df = 4, p-value = 0.0001494
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9078685 0.9989555
sample estimates:
cor
0.9900039
```

Ejemplo 5.1 (cont. del ejemplo anterior de R): Hemos obtenido una estimación puntual del coeficiente de correlación muy alta, r=0.990. Además, bajo la premisas de MAS, sabemos que ρ, el auténtico coeficiente poblacional de correlación, es algún valor comprendido entre 0.908 y 0.999. Nótese la gran asimetría del intervalo alrededor de 0.990.

Ejercicio 5.3

A) Calcule, con R, el IC del coeficiente de correlación para las muestras:

Y_A = 23.1 39.3 21.3 24.5 28.6 25.4

Y_B = 20.6 37.2 19.4 18.5 24.9 24.1

B) Dibuje su gráfico bivariante según las instrucciones vistas en el capítulo 3.

5.1. Variabilidad compartida: correlación intraclase*

En el caso de datos apareados, ambas variables están en la misma escala y puede recurrirse al Coeficiente de Correlación Intra-clase (ICC), Se asume que las unidades tienen dos tipos de variabilidades. Una que comparten ambas determinaciones y que diferencia unos individuos de otros: variabilidad entre-casos (σ^2_E) —o también, idiosincrasia: aquello que es propio de una unidad. La otra, la no compartida, contiene lo que no se repite, como podrían ser los errores de medida o las variaciones temporales dentro del individuo, muchas veces denominada, variabilidad intra-caso (σ^2_I). ICC es simplemente la proporción de variabilidad compartida:

$$ICC = \frac{\sigma_E^2}{\sigma_E^2 + \sigma_I^2}$$

Recuerde

ICC distingue 2 fuentes de variabilidad.

A diferencia de la correlación r de Pearson, ICC solo puede tomar valores entre 0 y 1.

Nota: En el caso de datos apareados, tiene sentido rechazar correlaciones negativas, en las que, al repetirse la determinación, un caso se parecería menos a sí mismo que a los otros: para 2 determinaciones de una misma variable en la misma escala ambos coeficientes coinciden.

En el caso de sólo 2 repeticiones, ICC puede obtenerse a partir de la correlación r de Pearson.

Ejemplo 5.2 (cont del Ejemplo 4.1): Obtuvo una estimación puntual del coeficiente de correlación muy alta, r=0.990. Al haber sólo 2 repeticiones, puede interpretar como ICC. Existe una gran repetibilidad de los valores. El análisis de datos apareados, al hacer la diferencia entre ambas variables, elimina la variabilidad compartida, entre-casos, σ^2_E, y el análisis de datos apareados será más preciso, con un error típico e estimación mucho menor.

Lectura: extendido a más de 2 determinaciones, el ICC valora el grado de similitud entre los k casos pertenecientes a un grupo.

Recuerde

ICC extiende el coeficiente r a más de 2 determinaciones.

6. IC de la probabilidad π

Una variable dicotómica, se puede resumir como el hecho de padecer o no cierto acontecimiento adverso (AA), definida mediante la proporción P de pacientes que lo han experimentado. La proporción P de la muestra estima la probabilidad poblacional π de que un nuevo paciente de las mismas características presente dicho AA.

Población	Muestra
Probabilidad	Proporción
π	P

Tabla 6.1. Nomenclatura para probabilidad y proporción

Nota: P es un estimador insesgado de π: E(P) = π. Y es convergente, ya que su varianza disminuye al aumentar el tamaño muestral: V(P) = π·(1-π)/n**.**

6.1. Método para muestras grandes

Si el tamaño muestral lo justifica, es cómodo recurrir a la aproximación a la Normal (mediante la binomial) de la distribución del estimador P, P~ N (π, π·(1-π)/n)

Definición

El **error típico del estimador P** cuantifica su distancia esperada al parámetro π y vale $\sqrt{\frac{\pi\ 1-\pi}{n}}$.

Ejercicio de Navegación

Observe que la aproximación de la Binomial a la Normal es tanto mejor cuanto mayor es el número de observaciones y más alejado de 0 y de 1 está el valor de π.

Nota: Observe que, en una binomial, dará los mismos resultados estimar la probabilidad π de éxito, que su complementario, la probabilidad 1-π de fracaso. O de la proporción poblacional de hombres y mujeres. Por ello, π y 1-π tienen un papel simétrico, por lo que la condición de que π no sea muy pequeña también aplica a 1-π.

Utilizando la Normal, el cálculo del IC es casi idéntico al de μ.

Fórmula

El **IC (1-α)% de una probabilidad (π)** es:

$$IC_{1-\alpha}\ \pi\ = \mathrm{P} \pm Z_{\alpha\ 2} \cdot \sqrt{\frac{\pi\ 1-\pi}{n}}$$

Recuerde

Se aceptan como **condiciones de aplicación** de la aproximación Normal que el tamaño muestral sea grande y las probabilidades *π* y *1-π* no extremas:

π·n ≥ 5 y (1-π)·n ≥ 5

Note la situación circular: ¡para estimar el intervalo de *π* es necesario conocer *π*! Hay dos posibles soluciones. La primera viene de que el producto π·(1-π) tiene un máximo cuando π = 0.5 = 1-π (**Tabla 6.2**).

π	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1-π	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
π(1-π)	**0.09**	**0.16**	**0.21**	**0.24**	**0.25**	**0.24**	**0.21**	**0.16**	**0.09**

Tabla 6.2 . Ilustración de que el máximo de π·(1-π) es para π = 0.5.

Se puede, por tanto, adoptar una actitud conservadora y decir que, en una muestra de tamaño n, la dispersión del estadístico P vale, como mucho:

$$\sigma_P = \sqrt{\frac{\pi\ 1-\pi}{n}} = \sqrt{\frac{0.5\ 1-0.5}{n}} = \frac{0.5}{\sqrt{n}}$$

Por lo que el cálculo del $IC_{1-\alpha}$ de π es:

Fórmula

$$IC_{1-\alpha}\ \pi\ = \mathrm{P} \pm Z_{\alpha\ 2} \cdot \sigma_P = \mathrm{P} \pm Z_{\alpha\ 2} \cdot \sqrt{\frac{0.5\ 1-0.5}{n}} = \mathrm{P} \pm Z_{\alpha\ 2} \cdot \frac{0.5}{\sqrt{n}}$$

La segunda solución consiste en sustituir π por p, tal como se hizo con σ^2 por S^2. Ahora, el cálculo del $IC_{1-\alpha}$ de π es:

Fórmula

$$IC_{1-\alpha}\ \pi\ = \mathrm{P} \pm Z_{\alpha\ 2} \cdot \sigma_P = \mathrm{P} \pm Z_{\alpha\ 2} \cdot \sqrt{\frac{\mathrm{P}\ 1-\mathrm{P}}{n}}$$

Recuerde

En el $IC_{95\%}$ de π, en lugar de π, se emplea, o bien 0.5, o bien P:

$IC_{1-\alpha}\ (\pi) = \mathrm{P} \pm Z_{\alpha/2}\ \sigma_p = \mathrm{P} \pm Z_{\alpha/2}\ \sqrt{[0.5 \cdot (1-0.5)/n]}$

$IC_{1-\alpha}\ (\pi) = \mathrm{P} \pm Z_{\alpha/2}\ \sigma_p = \mathrm{P} \pm Z_{\alpha/2}\ \sqrt{[\mathrm{P} \cdot (1-\mathrm{P})/n]}$

Ejemplo 6.1: Se lanza 100 veces una moneda al aire y se observan 56 caras.

Según el primer método:

$$IC_{95\%}\ \pi\ = \mathrm{P} \pm Z_{\alpha\ 2} \cdot \sqrt{\frac{0.5 \cdot 0.5}{n}} = 0.56 \pm 1.96 \cdot \frac{0.5}{\sqrt{100}} \approx 0.56 \pm 0.10 = \ 0.46\ ,0.66$$

Y de acuerdo con el segundo:

$$IC_{95\%}\ \pi\ = \mathrm{P} \pm Z_{\alpha\ 2} \cdot \sqrt{\frac{p \cdot\ 1-p}{n}} = 0.56 \pm 1.96 \cdot \frac{\sqrt{0.56 \cdot 0.44}}{\sqrt{100}} \approx 0.56 \pm 0.10 = \ 0.46\ ,0.66$$

Ambos métodos conducen a un intervalo muy similar (idéntico hasta el segundo decimal). Interpretamos que, con una confianza del 95%, la probabilidad de cara es uno de los valores comprendidos entre 0.46 y 0.56.

Nota: Se da esta coincidencia de resultados porque, en este ejemplo, p se encuentra muy cerca de 0.5, su máximo. Si se estuviera estimando un fenómeno más raro, con una π alejada de 0.5, la concordancia entre ambos procedimientos sería menor.

Nota: Puede decirse que $\sqrt{(0.5 \cdot 0.5/n)} = 0.5/\sqrt{n}$ es el valor del error típico de *p* en la situación de máxima indeterminación. Tiene la ventaja de que, dado cierto tamaño muestral, se dispone del mismo valor para cualquier variable dicotómica que desee estimar. Por lo tanto, en una encuesta con muchas preguntas o en una variable con varias categorías (por ejemplo, en la intención de voto) puede usar el mismo valor de σ_P para cada una de ellas.

Ejemplo de R

```
# La instrucción prop.test propociona el IC para pi
> prop.test(56,100)
1-sample proportions test with continuity correction
data:  56 out of 100, null probability 0.5
X-squared = 1.21, df = 1, p-value = 0.2713
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4573588 0.6579781
sample estimates:
p
0.56
```

Nota: Hay una pequeña diferencia entre la fórmula que emplea R y el cálculo anterior que no debe preocupar al estudiante. Los ejercicios de e-status dan ambas respuestas como buenas. El método clásico (la fórmula explicada) sólo funciona para 'n' muy grande, mientras que el método que proporciona R (prop.test, basado en el "***Wilson score method***") funciona bien en general, incluso para tamaños de pocas decenas.

Ejercicio 6.1

Dispone de una población, pongamos que infinita, de preguntas tipo test. Para un examen se seleccionan al azar 30 preguntas y un alumno contesta bien 18 de ellas. Como el interés del evaluador es conocer la proporción de preguntas de la población conocidas por este alumno (no de esta muestra de 30 preguntas)¿qué sabe sobre la proporción poblacional de preguntas que conoce el alumno?

Ejercicio 6.2

En un mega-ensayo, de los primeros 160 pacientes incluidos, 34 presentan una infracción mayor del protocolo en la primera visita. Calcule, con R, el $IC_{95\%}$ de la probabilidad de que un paciente tenga esta condición.

Ejercicio 6.3

¿Qué amplitud máxima tiene el $IC_{95\%}(\pi)$ de la proporción de pacientes con AA si n=100? ¿Y si n=400? ¿Y si n=2500? ¿Y si n=10000?

Ejercicio 6.4

¿Qué relación hay entre la amplitud del $IC_{95\%}$de π y el tamaño muestral n? Si quiere reducir la amplitud del intervalo a la mitad, ¿cuánto debe aumentar 'n'?

Ejercicio 6.5

De un total de 100 médicos, 40 prescriben cierto fármaco. Calcule el $IC_{95\%}$ de la proporción poblacional de médicos que lo prescriben. ¿Algún comentario sobre cómo deberían haber sido seleccionados estos médicos?

Ejercicio 6.6

Situándonos en el caso de mayor variabilidad o incertidumbre (Π=1-Π=0.5), ¿cuántos casos se necesitan para...

... estimar una proporción con un error típico de 0.05?

... estimar una proporción con un $IC_{95\%}$ de amplitud total de 0.05?

6.2. Método para muestras pequeñas

También se puede calcular el $IC_{95\%}(\pi)$ mediante un cálculo exacto basado en la Binomial.

Recuerde

También en muestras pequeñas puede obtener de R el $IC_{95\%}$ de π.

Ejemplo 6.2: Auditando la calidad de la documentación de las historias clínicas, observamos 8 de 10 programas que sí que cumplían con todas las normas de calidad. ¿Qué sabemos sobre la auténtica probabilidad π de que la historia clínica esté bien documentada? No hacen falta muchos cálculos para saber que π no puede ser 0. Ni tampoco 1. Veamos qué otros valores pueden ser razonables y cuáles no. Si asumimos que π=0.8, la probabilidad de observar X=8 en una muestra de n=10 vale:

$$P[X=8|X\sim B(10,0.8)] = \binom{10}{8} \pi^8 \ 1-\pi\ ^2 = 0.302 \quad [\text{dbinom}(8,10,0.8)]$$

Por tanto, π=0.8 parece un valor razonable. Ahora bien, si π fuera 0.3:

P[X=8|X~B(10,0.3)] = $\binom{10}{8} \pi^8 \ 1-\pi\ ^2$ =0.001 [dbinom(8,10,0.3)]

Y la de observar 8 o más sería:

P X≥8 X~B 10,0.3 = $\binom{10}{8} \pi^8 \ 1-\pi\ ^2 + \binom{10}{9} \pi^9 \ 1-\pi\ ^1 + \binom{10}{10} \pi^{10} \ 1-\pi\ ^0 =$ =

0.002 [1 - pbinom(7,10,0.3)]

Por lo tanto, π=0.3 no es un valor razonable.

Podemos proponer como valores poco 'razonables' aquellos para los cuales la probabilidad de observar 8 o más observaciones NO alcanza el valor α deseado. Por ejemplo:

Límite Inferior del $IC_{95\%}\ \pi\ = \pi_L$ tal que cumpla que:

$$P\ X \geq 8|X \sim B\ 10, \pi\ = 0.025 \Rightarrow \pi_L = 0.444$$

Límite Superior del $IC_{95\%}\ \pi\ = \pi_U$ tal que cumpla que:

$$P\ X \leq 8|X \sim B\ 10, \pi\ = 0.025 \Rightarrow \pi_U = 0.975$$

Es decir, 0.444 y 0.975 son valores del parámetro π que hacen poco probables (<0.05) muestras con 8 observaciones (o más extremas). Por tanto, el $IC_{95\%}$ del parámetro π va de 0.444 a 0.975:

$IC_{95\%}(\pi)$ = [0.444, 0.975]

En otras palabras: habiendo observado 8 de 10 historias con una documentación perfecta, lo único que podemos garantizar (con un riesgo α=0.05) es que la auténtica probabilidad de que una historia de este programa esté bien documentada es algún valor entre 0.444 y 0.975.

Notemos la gran amplitud de este intervalo, resultado de un tamaño muestral pequeño para una variable dicotómica. Lo que hace más relevante el $IC_{95\%}$.

Recuerde

En muestras pequeñas aún es más importante reflejar la incertidumbre y proporcionar el $IC_{95\%}$ de π.

Ejemplo de R

```
# IC95% exacto para πcon 8 éxitos de 10 observaciones
> binom.test(8,10,conf.level = 0.95)
        Exact binomial test
data:  8 and 10
number of successes = 8, number of trials = 10, p-value = 0.1094
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4439045 0.9747893
```

Nota: El método de R "binom.exact" es apropiado para cualquier 'n' (¡incluso para n=2!) pero es costoso en tiempo de ejecución para 'n' grandes. Por lo que en ese caso es mejor usar el método "prop.test" (***Wilson score method***).

Ejercicio 6.7.

Suponiendo en el ejemplo 6.2 que de los 10 programas estudiados, sólo 2 cumplían con las normas de calidad, encontrar el $IC_{95\%}$ para π mediante un cálculo exacto basado en la Binomial. Comparar con el anterior e interpretar: ¿son complementarios?

7. IC de medidas de riesgo en tablas 2x2

7.1. Diferencia de proporciones (Riesgos)*

Se definió la diferencia de riesgos como la diferencia entre la probabilidad de que un caso expuesto al factor desarrolle la enfermedad y la misma probabilidad en un caso no expuesto al factor (diferencia de riesgo entre expuestos y no expuestos).

Ejemplo 7.1: Recuerde la siguiente tabla en la que la estimación muestral *p* de la probabilidad en los expuestos era 5.3% [$P(Y+|X+) = 7 / 132 \approx 0.053$] mientras que en los no expuestos era 0.9% [$P(Y+|X-) = 8 / 868 \approx 0.009$].

	Y+	Y-	Total
X+	7	125	132
X-	8	860	868
Total	15	985	1000

Tabla 7.1 Presencia de la enfermedad Y y el factor de riesgo X en 1000 casos.

La diferencia entre 0.053 y 0.009 es 0.044, es decir, expresado en porcentajes, un 4.4%.

Fórmula

El **Intervalo de Confianza de (1-α)% de la DR** es:

$$IC_{1-\alpha}\ DR\ = DR \pm Z_{\alpha\ 2} \cdot \sigma_{DR} = DR \pm Z_{\alpha\ 2} \cdot \sqrt{\frac{P_1\ 1-P_1}{n_1} + \frac{P_2\ 1-P_2}{n_2}}$$

Nota: Como en el caso de la diferencia de medias en muestras independientes, la imprecisión de la diferencia de las proporciones es la suma de las imprecisiones de ambas proporciones.

El **requisito** para poder aplicar esta fórmula es que el tamaño muestral sea grande. Por dar unas cifras "mágicas" de referencia, las frecuencias de las celdas de la tabla 2x2 deberían ser superiores a 3 y el tamaño total de la tabla, a 100.

Recuerde

Para poder aplicar la fórmula se requiere:

(i) Celdas con más de 3 efectivos

(ii) Tamaño muestral superior a 100

Ejemplo 7.2: En los datos del ejemplo, el $IC_{95\%}(RA)$ es

$$IC_{1-\alpha}\ RA\ = RA \pm Z_{\alpha\ 2} \cdot \sqrt{\frac{P_1\ 1-P_1}{n_1} + \frac{P_2\ 1-P_2}{n_2}} =$$

$$= 0.044 \pm 1.96 \sqrt{\frac{0.053 \cdot 0.947}{132} + \frac{0.009 \cdot 0.991}{868}} \approx 0.044 \pm 1.96 \cdot 0.0198$$

$$= 0.044 \pm 0.038 = \ 0.0051, 0.0826\ \approx\ 0.5\%, 8.3\%$$

Y se concluye, por tanto, que los expuestos al factor tienen un riesgo entre 0.5% y 8.3% superior.

Nota: Para evitar el uso de frases con connotación causal, no hemos dicho "la exposición aumenta el riesgo entre un 0.5% y un 8.3%" .

	Y+	Y-	Total
X+	94	38	132
X-	215	653	868
Total	309	691	1000

Tabla 7.2 Datos para los ejercicios 7.1, 7.2 y 7.3.

Ejercicio 7.1

Con los datos de la **Tabla 7.2**, calcule el $IC_{95\%}$ (DR)

7.2. Riesgo relativo (RR)*

Se definió el riesgo relativo como el cociente entre las probabilidades de desarrollar la enfermedad; los expuestos dividida por la de los no expuestos (razón entre riesgo en expuestos y riesgo en no expuestos).

Ejemplo 7.3: Siguiendo con los datos de la **Tabla 7.2**, la razón entre 0.053 y 0.009 vale 5.7538, es decir, que el riesgo relativo observado es casi 6 veces superior en los expuestos.

Fórmula

El **Intervalo de Confianza de (1-α)% del RR (o cociente de probabilidad)** es:

$$IC_{1-\alpha}\ Log\ RR\ = Log\ RR\ \pm Z_{\alpha\ 2}\cdot\sigma_{LogRR}$$

$$= Log\ RR\ \pm Z_{\alpha\ 2}\cdot\sqrt{\frac{1-P_1}{n_1P_1}+\frac{1-P_2}{n_2P_2}}$$

Recuerde

El **requisito** para aplicar esta fórmula es, como antes, tamaño muestral grande.

Nota técnica: Este cálculo es ahora más complejo. Dada la asimetría del RR (que oscila entre 0 y 1 para riesgos inferiores en los expuestos y entre 1 e infinito para riesgos superiores) es preciso hacer previamente la transformación logarítmica natural (neperiana) para poder aprovechar la simetría resultante. La varianza del logaritmo del RR tiene ahora la misma interpretación en cualquier sentido.

Nota técnica: La fórmula de la varianza del logaritmo del RR no es inmediata. Es la suma de las varianzas de los logaritmos de las proporciones que son, a su vez, la varianza de la binomial dividida por el cuadrado de la proporción.

Ejemplo 7.4: En los datos del ejemplo, el RR=5.7538

Log(RR) = Log(5.7538) = 1.7499

$$IC_{95\%}\ Log\ RR\ = Log\ RR\ \pm Z_{0.025}\cdot\sqrt{\frac{1-P_1}{n_1P_1}+\frac{1-P_2}{n_2P_2}}$$

$$= 1.7499 \pm 1.96\cdot\sqrt{\frac{0.947}{132\cdot 0.053}+\frac{0.991}{868\cdot 0.009}} \approx 1.7499 \pm 1.96\cdot 0.5090$$

$$= 1.7499 \pm 0.9977 = \ 0.7521\ ,2.7476$$

Así, se puede afirmar que el valor de log(RR) aumenta entre 0.75 y 2.75, lo que resulta prácticamente imposible de interpretar: ¿Qué significa un aumento de log(RR) igual a 2.75? Para facilitar la interpretación se deshace el logaritmo:

$$IC_{95\%}\ RR\ = exp\ IC_{95\%}\ Log\ RR\ \ = e^{0.7521}, e^{2.7476} \approx 2.1, 15.6$$

Por lo que se concluye que los expuestos tienen un riesgo que es entre 2.1 y 15.6 veces superior: sea cual sea el riesgo en los no expuestos, en los expuestos, éste es entre 2.1 y 15.6 superior.

Nótese que el intervalo del RR es claramente asimétrico alrededor de la estimación puntual 5.75.

Nota: Una vez más para disminuir la connotación causal, hemos evitado en la frase verbos como 'aumenta' o 'multiplica': "la exposición al factor aumenta el riesgo entre 2.1 y 15.6 veces" o "el hecho de estar expuestos multiplica el riesgo entre 2.1 y 15.6 veces".

Ejercicio 7.2

Con los datos del Ejercicio 7.1 calcule el IC del RR

7.3. Odd ratio (OR)*

Se definió el *odds ratio* como el cociente entre las *odds* (o razones sí/no) de desarrollar la enfermedad entre los expuestos y los no-expuestos.

Ejemplo 7.5: Siguiendo con los datos de la **Tabla 7.2**, las *odds* son 0.056 y 0.009 y su razón vale 6.0200, es decir, que la razón enfermo/sano es 6 veces superior en los expuestos.

Como con el riesgo relativo, la asimetría del OR aconseja emplear la transformación logarítmica.

Fórmula

El **Intervalo de Confianza de (1-α)% del OR (o cociente de momios)** es:

$$IC_{1-\alpha}\ Log\ OR\ \ = Log\ OR\ \pm Z_{\alpha\ 2} \cdot \sigma_{LogOR}$$

$$= Log\ OR\ \pm Z_{\alpha\ 2} \cdot \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

Donde *a*, *b*, *c* y *d* representan los 4 valores de la tabla 2 x 2

Nota técnica: El IC del OR se obtiene asumiendo estimaciones de Poisson independientes en las 4 celdas.

Recuerde

El **requisito** para aplicar esta fórmula es, otra vez, tamaño muestral grande.

Ejemplo 7.6: En los datos del ejemplo, el OR= (7/125)/(8/860)= 6.0200

Log(OR) = Log(6.0200) = 1.7951

$$IC_{95\%}\ Log\ OR\ = Log\ OR\ \pm Z_{0.025}\cdot\ \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}=$$

$$=1.7951\pm 1.96\cdot\ \sqrt{\frac{1}{7}+\frac{1}{125}+\frac{1}{8}+\frac{1}{860}}\approx 1.7951\pm 1.96\cdot 0.5263$$

$$=1.7951\pm 1.0316=\ 0.763\ ,2.827$$

Y para facilitar la interpretación se deshace el logaritmo:

$$\boldsymbol{IC_{95\%}\ RR\ = exp\ IC_{95\%}\ Log\ RR\ \ = e^{0.7563}\ ,e^{2.827}\ \approx\ 2.1,16.9}$$

Por lo que se concluye que los expuestos tienen una razón enfermo/sano que es entre 2.1 y 16.9 veces superior.

Lectura: Como siempre, se ha evitado hablar de efecto causal con frases como "el factor multiplica la razón enfermo / sano entre 2.1 y 16.9 veces".

Nota: Observe que los IC del RR y del OR son muy similares. Recuerde que esto ocurre con eventos raros, como es el caso, donde la proporción de enfermos es muy baja en los 2 grupos.

Lectura: **Serra-Prat M**. Si agrupamos las distintas categorías de la variable origen en dos categorías (autóctonos e inmigrantes), observamos una asociación estadísticamente significativa entre el déficit de yodo y el origen; OR = 2.88; $IC_{95\%}$: [1.33 , 6.12].

Ejercicio 7.3

Con los datos del Ejercicio 7.1 calcule el $IC_{95\%}$ del OR

7.4. Cálculo con R de los IC de DR, RR y OR

Los IC de las 3 medidas de asociación más usuales para dicotomías se obtienen con R.

Ejemplo de R

```
# IC95% mediante la funcion epi2x2 del package epibasix
> install.packages('epibasix')
> library(epibasix)
> tabla <- matrix(c(7,125,8,860),2,2,byrow=T) # Tabla 7.2
> results <- epi2x2(tabla)
> attach(results)
```

```
# CIL=Confidence Interval Lower; CIU=Confidence Interval Upper
# rdCo=Risk Difference
# Estimación puntual e IC para la DR
> rdCo;rdCo.CIL;rdCo.CIU
[1] 0.04381371
[1] 0.0006959811
[1] 0.08693145

# Estimación puntual e IC para el RR
> RR;RR.CIL;RR.CIU
[1] 5.753788
[1] 2.121543
[1] 15.60471

# Estimación puntual e IC para el OR
> OR;OR.CIL;OR.CIU
[1] 6.02
[1] 2.145785
[1] 16.88911
> detach(results)
```

Lectura: Los intervalos de confianza son el método de inferencia más relevantes y fácilmente comunicables. Las revistas biomédicas más importantes aconsejan basar la presentación de los resultados del estudio en intervalos de confianza. En el ítem 17b de la guía CONSORT **(**Figura 7.1) puede encontrar con más detalle el porqué de la presentación de los resultados en intervalos de confianza. Este ítem recomienda reportar a la vez una medida basada en diferencias (el RA) y otra basada en cocientes (OR o RR) ya que ninguna por separado aporta una visión completa del efecto y sus implicaciones.

Table 7 | Example of reporting both absolute and relative effect sizes. (Adapted from table 3 of The OSIRIS Collaborative Group[242])

	Percentage (No)			
Primary outcome	Early administration (n=1344)	Delayed selective administration (n=1346)	Risk ratio (95% CI)	Risk difference (95% CI)
Death or oxygen dependence at "expected date of delivery"	31.9 (429)	38.2 (514)	0.84 (0.75 to 0.93)	−6.3 (−9.9 to −2.7)

Figura7.1. Modelo de Consort para presentar los resultados de dicotomías.

Ejercicio 7.4

Pongamos que se define el Fracaso Escolar (FE) como el hecho de no terminar los estudios dentro del plazo previsto más un año de margen (posibles valores: SÍ/NO). Se dispone de un posible predictor dicotómico de FE: notas de entrada superiores (S) o inferiores (I) a la media de dicho centro.

a) Invente una tabla 2x2 que muestre relación entre FE y notas.

b) Calcule las 3 medidas y sus $IC_{95\%}$ con R.

Ejercicio 7.5

El comité de cierta empresa solicita una compensación económica para los empleados que pasan mucho tiempo delante del ordenador, alegando que este hecho genera Enfermedades de la Columna Vertebral (ECV). Vd forma parte del equipo que debe pronunciarse sobre este tema. Han recogido información sobre ECV en todos los trabajadores de la empresa y comparan los datos de aquellos que pasan más de 25 horas a la semana delante del ordenador con los que pasan menos de 10 horas. Los datos figuran en la tabla siguiente:

	ECV+	**ECV-**
≥ 25	111	87
≤ 10	231	261

a) Vd debe elegir entre una medida de asociación para comparar los riesgos de ambos grupos. A partir de la nota técnica final del apartado 4.2, ¿qué implican los modelos aditivo y multiplicativo que subyacen detrás de la diferencia de riesgos y del riesgo relativo?

b) Calcule el RA.

c) Calcule el RR.

d) Finalmente han decidido utilizar la medida de asociación más habitual: el odds ratio. Calcúlelo con R junto con su $IC_{95\%}$.

e) Interprete el resultado anterior. En concreto, ¿se sostiene que la probabilidad de ECV es la misma en ambos grupos?

Ejercicio 7.6

En la tabla figuran datos de Bishop et al. sobre la promulgación de la pena de muerte (P: SI/NO) en función de la raza (Blanco/negro) del acusado (A) y de la víctima (V). Construya la tabla para estudiar la relación entre la pena y la raza de la víctima sin tener en cuenta la raza del acusado. Estime con R el IC95% del odds ratio. Interprete el resultado.

Pena de muerte: SÍ			Pena de muerte: NO		
	A:Blanco	A:Negro		A:Blanco	A:Negro
V:Blanco	19	11	V:Blanco	132	52
V:Negro	0	6	V:Negro	9	97

Soluciones a los ejercicios.

2.1. (Redondeamos al entero superior para obtener la amplitud deseada "o superior")

$X \sim \chi^2_3$. > `pchisq(1,df=3)`

```
[1] 0.198748
> pchisq(3,df=3)
[1] 0.6083748
```

P(1≤X≤3) = P(X ≤3) - P(X ≤1) = >`pchisq(3,df=3) - pchisq(1,df=3)`

```
[1] 0.4096268
```

2.2. $X \sim t_{12}$. P(X > 1.796) = 0.05

```
.>pt(q=1.796,df=12,lower.tail=FALSE)
[1] 0.04884788
```

2.3 La amplitud del intervalo es lo que en la fórmula va detrás del "±". Por ello, la amplitud depende de 3 valores: $Z_{\alpha/2}$, σ y n. Por el enunciado, no podemos cambiar la confianza y por tanto $Z_{\alpha/2}$ deberá quedar igual. Así pues, sólo disponemos de σ y de 'n' para hacer más estrecho el intervalo. Podríamos disminuir σ controlando sus fuentes de variación, pero por ahora centrémonos en 'n'. Como ésta dentro de una raíz cuadrada, para conseguir que el IC95% sea la mitad de amplio, hay que multiplicar por 4 el tamaño muestral.

2.4 Debemos cambiar el valor 1.96 por 2.576 obtenido de R:

```
>qnorm(p=0.995)
[1] 2.575829
```

$IC_{99\%}(\mu) = \bar{X} \pm Z_{0.995}\sigma/\sqrt{n} = 5 \pm 2.576 \cdot 1/\sqrt{9} = 5 \pm 2.576/3 \approx [4.14, 5.86]$

2.5 No puede saberse si uno concreto contiene μ. Si se repite indefinidamente el proceso, el (1-α)% de las ocasiones contendrá μ, pero no se puede saber para cada vez.

2.6 La respuesta correcta es la c), ya que el IC se hace alrededor de la media muestral observada $\bar{X}$ para tener una alta confianza de contener a la (única) media poblacional μ desconocida. ['a' es falsa porque sólo hay 1 media poblacional; 'b' porque sólo sería cierto si, por azar, $\bar{X}=\mu$, lo que tienen una probabilidad prácticamente nula (0 en caso de continuas); y 'd' porque siempre incluye a la media muestral en que se basa.]

2.7 El IC se no hace referencia a los casos, sino a los parámetros desconocidos, por ello, las respuestas posibles son la c) o la d), si bien es más correcto formalmente hablar de confianza que de probabilidad (lea la "nota" que sigue al ejercicio para más explicaciones).

2.8 a) Si $\sigma/\sqrt{n} = 12$ y $\sigma = 120 \rightarrow n = 100$

b) Si $Z_{0.975}\sigma/\sqrt{n} = 12$; $1.96 \cdot 120/\sqrt{n} = 12$; $\rightarrow n = (1.96 \cdot 120/12)^2 = 384.16 \rightarrow n = 385$

c) Si $\pm Z_{0.975}\sigma/\sqrt{n} = \pm 6$; $1.96 \cdot 120/\sqrt{n} = 6$; $\rightarrow n = (1.96 \cdot 120/6)^2 = 1536.64 \rightarrow n = 1537$

2.9 Dado que la muestra es de 100 casos, no es necesario preguntarse si GOT es Normal (lo que es una suerte, ya que GOT son positivas, por lo que una desviación típica mayor que la media implicaría valores negativos en una distribución simétrica como la Normal).

```
> qt(p=0.025,df=99)
[1] -1.984217
```

$$IC_{95\%}(\mu) = \bar{X} \pm t_{99,0.975}\ S/\sqrt{n} \approx 80 \pm 1.98 \cdot 120/\sqrt{100} \approx 80 \pm 24 \approx [56, 104]$$

3.1. Cálculo de S^2 y S:

#Con R, el intervalo de la varianza (σ^2) es

```
> muestra <- c(2,3,4,5)
>IC_var(muestra,0.95)
[1]  0.5348507 23.1701080
#Y, el intervalo de confianza de la desviación típica (σ) es
>sqrt(IC_var(muestra,0.95))
[1] 0.7313349 4.8135338
```

4.1. a) Muestras independientes → IC95% = [-5.39 , 11.29]

```
> YA<-c(23.05,39.06,21.72,24.47,28.56,27.58)
> YB<-c(20.91,37.21,19.29,19.95,25.32,24.07)
>t.test(YA,YB,var.equal = TRUE)
[…]
95 percent confidence interval:
 -5.39087 11.28754
[…]
```

b) Muestra sapareadas→ IC95% = [1.90 , 4.00]

```
>t.test(YA,YB,paired=TRUE)
[…]
95 percentconfidenceinterval:
 1.898994 3.997673
[…]
```

c) Comparación de los errores estándar. En el caso de muestras apareadas, el error estándar es mucho más pequeño (0.41 vs. 3.74)

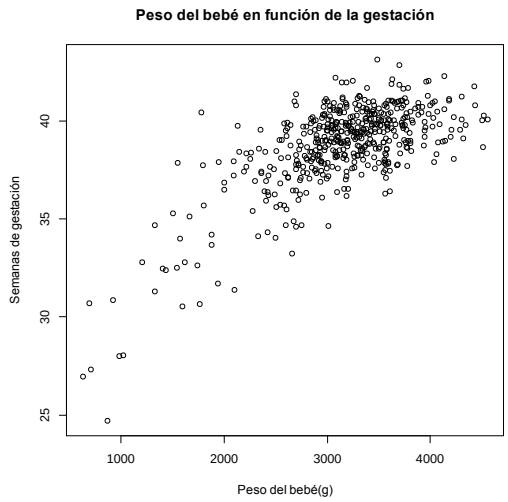
```
> # Error típico para muestras independientes
>var_pooled<-(var(YA)*5+var(YB)*5)/10
>errortip_ind<-sqrt(var_pooled*(1/6+1/6))
>errortip_ind
[1] 3.742676
> # Error típico para muestras apareadas
>var_apa<-var(YA-YB)
>errortip_apa<-sqrt(var_apa/6)
```

```
>errortip_apa
[1] 0.4082109
```

5.1 Se convierte en la varianza de X: al cambiar Y por X, la X aparece 2 veces y queda SX2.

5.2 a) `>install.packages('Epi')`

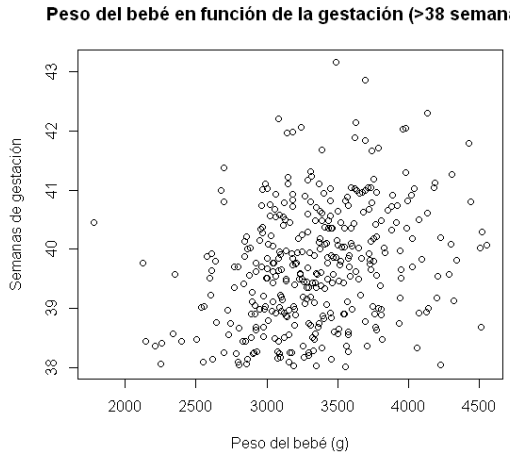
```
>library(Epi)
>data(births)
>plot(births$gestwks~births$bweight,  main="Peso del bebé en función de la gestación",
      xlab="Peso del bebé (g)",ylab="Semanas de gestación")
```



Observe que a la figura a la que más se asemeja es r=0.75

b) #Seleccionamos sólo los tiempos de gestación ≥38 semanas

```
>births2<-subset(births,births$gestwks>=38)
>plot(births2$gestwks~births2$bweight, main="Peso del bebé en función de la gestación (>38 semanas)",
      xlab="Peso del bebé (g)", ylab="Semanas de gestación")
```



Ahora la más parecida es r=0.25

c) El comando a utilizar es cor(x,y)

```
>cA<-cor(births$bweight,births$gestwks,use="pairwise.complete.obs")
> cA
[1] 0.7122162
>cB<-cor(births2$bweight,births2$gestwks, use="pairwise.complete.obs")
> cB
[1] 0.2896377
```

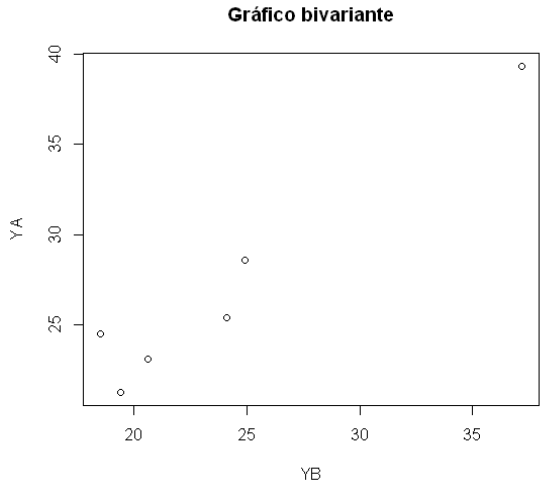
Nota: si pide que R le muestre el data.frame births, verá que hay algunas variables de interés (las utilizadas para el cálculo de correlación) que contienen NA's; con el argumento '`use="pairwise.complete.obs"`' le indicamos que calcule el coeficiente de correlación sólo con aquellos individuos que no contengan NA's en estas variables.

d) Observe en el gráfico que la impresión de relación viene sobre todo por los valores del cuadrante inferior izquierdo: son los bebés muy pre-término los que muestran un peso menor y marcan más la relación. Al eliminarlos, baja el valor de r. En el caso extremo que reduzcamos las semanas de gestación a un único valor, no tendríamos variabilidad en esta variable, no podríamos distinguir los casos por la duración de la gestación y no podríamos mirar si los de más semanas pesan más: su correlación sería 0.

5.3 a) El $IC_{95\%}$para ρ es [0.739, 0.997]

```
>YA<-c(23.1,39.3,21.3,24.5,28.6,25.4)
>YB<-c(20.6,37.2,19.4,18.5,24.9,24.1)
>cor.test(YA,YB)$conf.int
[1] 0.7389701 0.9967569
```

b) `> plot(YA~YB, main="Gráfico bivariante")`



6.1. $IC_{95\%}(\pi) = P \pm Z_{\alpha/2}\sqrt{[P(1-P)/n]} = P \pm Z_{\alpha/2}\sqrt{[0.6 \cdot 0.4/30]} \approx 0.60 \pm 0.18 = [0.42, 0.78]$

Parece que, con 30 preguntas, se sabe, de este alumno, menos de lo que parecía: sólo se sabe que la proporción poblacional de preguntas que conoce este alumno es algún valor entre el 42 y el 78%. Si descontamos la influencia del azar, podemos afirmar que este alumno sabe entre un 42 y un 78% de las preguntas. [Recuerde la premisa de

independencia de las observaciones: si algunas preguntas estuvieran relacionadas, el intervalo de incertidumbre sería aún mayor.]

[Condiciones de aplicación: 0.42·30 = 12.6>5 y (1-0.78)·30 = 6.6>5]

Como ya se ha dicho, el método de R garantiza mejor cobertura en muestras pequeñas

```
>prop.test(18,30)
```

[…]

```
95 percent confidence interval:
0.4075022 0.7677666
```

[…]

6.2. $IC_{95\%}(\pi)= P \pm Z_{\alpha/2}\sqrt{[P(1-P)/n]}=0.212\pm Z_{\alpha/2}\sqrt{[0.212\cdot0.788/160]} \approx 0.212\pm 0.0634\approx[0.1491, 0.2759]\approx[15\%, 28\%]$

[Condiciones de aplicación: 0.15·160 = 24 >5]

Con R:

```
>prop.test(34,160)
```

[…]

```
95 percent confidence interval:
0.1535181 0.2856165
```

6.3. Amplitud máxima $IC_{95\%}\pi\rightarrow\pm 1.96\sqrt{[0.5\cdot0.5/n]}$

a) n=100 $\rightarrow\pm 1.96\sqrt{[0.5\cdot0.5/100]} = \pm 1.96\cdot0.05 = \pm 0.098 \approx\pm 10\%$

b) n=400 $\rightarrow\pm 1.96\sqrt{[0.5\cdot0.5/400]} = \pm 1.96\cdot0.025 = \pm 0.049 \approx\pm 5\%$

c) n=2500 $\rightarrow\pm 1.96\sqrt{[0.5\cdot0.5/2500]} = \pm 1.96\cdot0.01 = \pm 0.0196 \approx\pm 2\%$

d) n=10000 $\rightarrow\pm 1.96\sqrt{[0.5\cdot0.5/10000]} = \pm 1.96\cdot0.005 = \pm 0.0098 \approx\pm 1\%$

6.4. La amplitud del intervalo es inversamente proporcional a la raíz del tamaño muestral. Como en el caso de la media muestral, para disminuir la incertidumbre a la mitad, es necesario aumentar el tamaño muestral cuatro veces.

6.5. $IC_{95\%}(\pi) = P \pm Z_{\alpha/2}\sqrt{[P(1-P)/n]} = 0.40 \pm Z_{\alpha/2}\sqrt{[0.40\cdot0.60/100]} \approx 0.40 \pm 0.096 \approx [0.304, 0.496] \approx [30\%, 50\%]$

[Condiciones de aplicación: 0.3·100 = 30>5]

Con R:

```
>prop.test(40,100)
```

[…]

```
95 percent confidence interval:
0.3047801 0.5029964
```

[…]

Debería ser una selección al azar. Y no lo ha dicho. Recuerde que el IC y el error típico de estimación sólo tienen en cuenta los errores aleatorios, pero no los sistemáticos. Si la muestra no fuera al azar, los autores deberían mencionar que, por la existencia de un sesgo impredecible, la incertidumbre es quizás mayor que la reflejada por el intervalo.

6.6. Si $\sigma_P = \sqrt{\pi(1-\pi)/n} = \sqrt{0.5\cdot0.5/n} =0.05 \rightarrow n=100$

Si $\pm Z_{0.975}\sigma_P = \pm 025$; $1.96\cdot\sqrt{0.5\cdot0.5/n} = 0.025$; $\rightarrow n = (1.96\cdot0.5/0.025)^2 =1536.64 \rightarrow n=1537$

6.7.

```
>binom.test(2,10,conf.level=0.95)$conf.int
```

```
[1] 0.02521073 0.55609546
```

El IC95% de [0.025, 0.556] es el complementario del hallado para 8 casos ya que 0.025=1-0.975 y 0.556=1-0.444.

7.1. $IC_{95\%}(RA) = RA \pm Z_{\alpha/2}\sqrt{[P_1\cdot(1-P_1)/n_1 + P_2\cdot(1-P_2)/n_2]} =$

$= 0.4644 \pm 1.96\sqrt{[(0.712\cdot0.288/132)+(0.248\cdot0.752/868)]} \approx$

$= 0.4644 \pm 1.96\cdot0.0420 = 0.4644 \pm 0{,}0824 = [0.3820, 0.5468] \approx [38{,}2\%, 54{,}7\%]$

Por lo que puede afirmarse que los expuestos al factor presentan entre un 38 y 55% más de riesgo.

7.2. RR=0.7121/0.2477=2.875→Log(RR) = 1.0560

$IC_{95\%}\log(RR) = \underline{Log(RR)} \pm Z_{\alpha/2}\sqrt{[(1-p_2)/n_2p_2 + (1-p_1)/n_1p_1]} =$

$= 1.0560 \pm 1.96\sqrt{[0.2879/132\cdot0.7121 + 0.7523/868\cdot0.2477]} \approx$

$= 1.0560 \pm 1.96\cdot0.0810 = 1.0560 \pm 0.1588 = [0.8973, 1.2148]$

$IC_{95\%}(RR) = \exp[IC_{95\%}\log(RR)] = [e^{0.8973}, e^{1.2148}] \approx [2.45 , 3.37]$

Por lo que se concluye que los expuestos tienen un riesgo que es entre 2.45 y 3.37 veces superior.

7.3. En los datos del ejemplo, el OR = (94/38)/(215/653)= 7.5131 →Log(OR) = 2.0166

$IC_{95\%}\log(OR) = \underline{Log(OR)} \pm Z_{\alpha/2}\sqrt{(1/a+1/b+1/c+1/d)} =$

$= 2.0166 \pm 1.96\sqrt{[1/94+1/38+1/215+1/653]} \approx$

$= 2.0166 \pm 1.96\cdot0.2077 = 2.0166 \pm 0.4071 = [1.6096, 2.4237]$

$IC_{95\%}OR = \exp[IC_{95\%}\log(OR)] = [e^{1.6096}, e^{2.4273}] \approx [5.0 , 11.3]$

Por lo que se concluye que los expuestos tienen una razón enfermo / sano que es entre 5.0 y 11.3 veces superior.

7.4. a) La tabla muestra un posible ejemplo.

	FE:NO	FE: SÍ
PAU:S	200	10

b) Puede hallar los resultados con R con el siguiente código:

```
>install.packages('epibasix')
> library(epibasix)
>tabla<- matrix(c(200,10,100,100),2,2,byrow=T)
> results <- epi2x2(tabla)
> attach(results)
# Estimación puntual e IC para el RA
>rdCo;rdCo.CIL;rdCo.CIU
# Estimación puntual e IC para el RR
> RR;RR.CIL;RR.CIU
# Estimación puntual e IC para el OR
> OR;OR.CIL;OR.CIU
> detach(results)
```

7.5. a) Ambos parten del principio de que una proporción de casos desarrollan la ECV, independientemente de su exposición al ordenador. Pero difieren en que la diferencia de riesgos considera que por el hecho de estar expuesto, aparecen nuevos casos, diferentes a los anteriores, que desarrollan también la enfermedad. En cambio, el riesgo

relativo considera que el hecho de estar expuesto aumenta, en una cierta persona, la probabilidad de desarrollar ECV. Es decir, en la diferencia de riesgos se 'suman' dos grupos de casos, mientras que en el relativo, lo que se modifica es la probabilidad de cada caso.

b) RA = (111/(111+87)) - (231/(231+261)) ≈ 0.091

c) RR = (111/(111+87)) / (231/(231+261)) ≈ 1.194

d) OR = 111·261/(87·231) ≈ 1,442

ln(OR) ≈ 0.366

V(ln(OR)) = (1/111)+(1/261)+(1/87)+(1/231)= 0.029

SE(ln(OR)) ≈ 0.1693

$IC_{95\%}$ln(OR) = ln(or) ± 1.96· SE(ln(or)) ≈ (0.034, 0.698)

$IC_{95\%}$ OR = exp(0.034, 0.698) ≈ (1.034 , 2.009)

e) No, dado que el IC excluye el valor de no relación, podemos rechazar la independencia entre el grado de exposición al ordenador y la presencia de ECV. Otro tema es la relación causal, ya que se trata de un estudio transversal y no puede distinguirse qué variable sigue a qué variable.

Puede hallar los resultados con R con el siguiente código:

```
>install.packages('epibasix')
> library(epibasix)
>tabla<- matrix(c(111,87,231,261),2,2,byrow=T)
> results <- epi2x2(tabla)
> attach(results)
#b) Estimación puntual (e IC) para el RA
>rdCo;rdCo.CIL;rdCo.CIU
#c) Estimación puntual (e IC) para el RR
> RR;RR.CIL;RR.CIU
#d) Estimación puntual e IC para el OR
> OR;OR.CIL;OR.CIU
#Forma logarítmica
>lnOR<-log(OR)
>varlnOR<-(1/111)+(1/261)+(1/87)+(1/231)
>SElnOR<-sqrt(varlnOR)
>LI<-log(OR)-1.96*SElnOR
>LS<-log(OR)+1.96*SElnOR
>IC<-c(exp(LI),exp(LS))
    > detach(results)
```

7.6. En los datos globales, sin tener en cuenta otras variables, la disparidad "PENA MUERTE = SÍ/PENA MUERTE = NO" es entre 1.16 y 7.15 superior cuando la víctima es de raza blanca que cuando lo es de raza negra.

Víctima	Blanco	Negro
Pena: SÍ	30	6
Pena:NO	184	106
OR=(30*106)/(184*6)=2.88		

$\ln(OR) = \ln(2.88) \cong 1.06$

$V(\ln(OR)) = a^{-1} + b^{-1} + c^{-1} + d^{-1} =$

$= 30^{-1} + 106^{-1} + 184^{-1} + 6^{-1} \cong 0.21$

$IC_{95\%}\ln(OR) \cong 1.06 \pm 1.96\sqrt{0.22} \cong 1.06 \pm 0.91 = [0.15, 1.97]$

$IC_{95\%}OR \cong [\exp(0.15), \exp(1.97)] \cong [1.16, 7.15]$

[Nótese la simetría del intervalo en la escala logarítmica y su asimetría en la escala natural].

Puede hallar los resultados con R con el siguiente código:

```
>install.packages('epibasix')
> library(epibasix)
>tabla<- matrix(c(30,6,184,106),2,2,byrow=T)
>results <- epi2x2(tabla)
>attach(results)
># Estimación puntual
>OR;OR.CIL;OR.CIU
#Forma logarítmica
>lnOR<-log(OR)
>varlnOR<-(1/30)+(1/184)+(1/6)+(1/106)
>SElnOR<-sqrt(varlnOR)
>LI<-log(OR)-1.96*SElnOR
>LS<-log(OR)+1.96*SElnOR
>IC<-c(exp(LI),exp(LS))
>detach(results)
```

Tabla salvadora

La siguiente tabla le recuerda las fórmulas y comandos de R que proporcionan los IC estudiados.

No debe recordarlos, pero sí saber interpretar sus resultados.

		Fórmula	R
IC de μ	σ desconocida	$IC_{1-\alpha}\ \mu\ = x \pm t_{n-1,\alpha\ 2} \cdot \frac{S}{\sqrt{n}}$	*t.test*
IC de σ^2		$IC_{1-\alpha}\ \sigma^2\ = \frac{S^2\ n-1}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{S^2\ n-1}{\chi^2_{n-1,\frac{\alpha}{2}}}$	*Función propia*
IC de $(\mu_1 - \mu_2)$	σ_1^2 y σ_2^2 desconocidas	$IC_{1-\alpha}\ \mu_1-\mu_2\ =\ y_1-y_2\ \pm t_{n-2,\alpha\ 2} \cdot \sqrt{S^2\ \frac{1}{n_1}+\frac{1}{n_2}}$ Dónde: $S^2 = \frac{n_1-1\ S_1^2+\ n_2-1\ S_2^2}{n_1+n_2-2}$	*t.test*
IC de π	Muestras grandes	$IC_{1-\alpha}\ \pi\ = \mathrm{P} \pm Z_{\alpha\ 2} \cdot \sigma_P = \mathrm{P} \pm Z_{\alpha\ 2} \cdot \sqrt{\frac{\mathrm{P}\ 1-\mathrm{P}}{n}}$	*prop.test*
	Muestras pequeñas	$IC_{1-\alpha}\ \pi\ = [P\ X \geq 8 \mid X \sim B\ \mathrm{n}, \pi\ = 0.025,$ $P\ X \leq 8 \mid X \sim B\ \mathrm{n}, \pi\ = 0.025]$	*binom.test*
IC del RA		$IC_{1-\alpha}\ RA\ = RA \pm Z_{\alpha\ 2} \cdot \sigma_{RA} =$ $= RA \pm Z_{\alpha\ 2} \cdot \sqrt{\frac{\mathrm{P}_1\ 1-\mathrm{P}_1}{n_1}+\frac{\mathrm{P}_2\ 1-\mathrm{P}_2}{n_2}}$	*epi2x2* (*'epibasix'*)
IC del RR		$IC_{1-\alpha}\ Log\ RR\ = Log\ RR\ \pm Z_{\alpha\ 2} \cdot \sigma_{LogRR} =$ $= Log\ RR\ \pm Z_{\alpha\ 2} \cdot \sqrt{\frac{1-\mathrm{P}_1}{n_1\mathrm{P}_1}+\frac{1-\mathrm{P}_2}{n_2\mathrm{P}_2}}$	*epi2x2* (*'epibasix'*)
IC del OR		$IC_{1-\alpha}\ Log\ OR\ = Log\ OR\ \pm Z_{\alpha\ 2} \cdot \sigma_{LogOR} =$ $= Log\ OR\ \pm Z_{\alpha\ 2} \cdot \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$ *a*, *b*, *c* y *d* representan los 4 valores de la tabla 2 x 2	*epi2x2* (*'epibasix'*)
Recuerde que cuando no se cumplen las premisas de normalidad puede ser útil utilizar métodos no paramétricos o de remuestreo, como por ejemplo el bootstrap			install.packages("bootstrap") library("bootstrap")

Tabla 7.3. Tabla resumen de las fórmulas vistas en este capítulo.