Tema 1. Introducción a los métodos de remuestreo con soporte computacional en R

basado en

B. Efron, R. Tibshirani (1993). An Introduction to the bootstrap.

O. Kirchkamp (2019). Resampling methods.

Curso 2022/2023

Introducción

- La Estadística es una ciencia puramente experimental que aprende de la experiencia, especialmente de la que aparece de manera *secuencial*.

- Se puede considerar que comienza alrededor del siglo XVII cuando Graunt (UK) empezó a usar **tablas de mortalidad**.

- Actualmente se aplica en todas las ciencias que requieren tratamiento de la información, desde ciencias biomédicas, psicología, educación y economía, hasta el estudio de las partículas en física cuántica, o de galaxias extremadamente distantes.

Introducción

- Pero la mayoría de las personas y los dispositivos **no** son muy eficientes para encontrar patrones en un mar de datos con *ruido*.
- Es decir, tenemos tendencia a contemplar patrones inexistentes fuera que suceden para satisfacer nuestros propósitos como se hace en astrología...
- La teoría estadística intenta proporcionar métodos óptimos para la búsqueda de señales reales en ambientes ruidosos y también proporciona controles estrictos contra la *sobre-interpretación* de patrones al azar.

Introducción

La teoría estadística trata de responder a tres preguntas básicas:

I. ¿Cómo debo recoger o muestrear mis datos?

II. ¿Cómo debo analizar y resumir los datos que he recogido?

III. ¿Cómo son de exactos los resúmenes de mis datos?

El punto 3 constituye parte del proceso conocido como *Inferencia Estadística*.

Introducción

- Las técnicas de remuestreo son técnicas desarrolladas hace pocos años para calcular estadísticos, basándose en técnicas computacionales intensivas que evitan los cálculos complejos de la teoría estadística tradicional.

- Aunque el uso de las técnicas de remuestreo implican el uso de conceptos tradicionales de inferencia estadística, cambia radicalmente su implementación.

- Mediante el uso de computación intensiva se aplican las técnicas de manera flexible, fácil y con un mínimo de aparato matemático.

Ejemplo

- Los tres conceptos básicos: recogida de datos, resumen de los mismos e inferencia se ilustran en la noticia aparecida en el *New York Times*.
- Se hizo un estudio sobre si pequeñas dosis de aspirina podrían prevenir los ataques al corazón en personas de mediana edad.
- Los datos del estudios se tomaron de manera eficiente mediante un estudio controlado, aleatorizado y *doble ciego*.

HEART ATTACK RISK FOUND TO BE CUT BY TAKING ASPIRIN

LIFESAVING EFFECTS SEEN

Study Finds Benefit of Tablet Every Other Day Is Much Greater Than Expected

By HAROLD M. SCHMECK Jr.

A major nationwide study shows that a single aspirin tablet every other day can sharply reduce a man's risk of heart attack and death from heart attack.

The lifesaving effects were so dramatic that the study was halted in mid-December so that the results could be reported as soon as possible to the participants and to the medical profession in general.

The magnitude of the beneficial effect was far greater than expected, Dr. Charles H. Hennekens of Harvard, principal investigator in the research, said in a telephone interview. The risk of myocardial infarction, the technical name for heart attack, was cut almost in half.

'Extreme Beneficial Effect'

A special report said the results showed "a statistically extreme beneficial effect" from the use of aspirin. The report is to be published Thursday in The New England Journal of Medicine.

In recent years smaller studies have demonstrated that a person who has had one heart attack can reduce the risk of a second by taking aspirin, but there had been no proof that the beneficial effect would extend to the general male population.

Dr. Claude Lenfant, the director of the National Heart Lung and Blood Institute, said the findings were "extremely important," but he said the general public should not take the report as an indication that everyone should start taking aspirin.

Ejemplo

- La mitad de las personas recibió una sustancia placebo y las personas se asignaron de manera aleatoria a los tratamientos.
- Tanto los sujetos como las personas que trabajaban en el estudio no sabían tanto el tipo de sustancia que recibía o daba a los pacientes.
- Los estadísticos de resumen del artículo eran muy simples:

	Ataques Corazón	**No Ataque Corazón**
Aspirina	104	11037
Placebo	189	11034

Ejemplo

- A simple vista parece que hay ***menos*** ataques de corazón en el grupo de aspirina.

- La *razón de odds* (la razón entre ambos ratios de ataques la corazón) es

$$\widehat{\theta} = \frac{104/11037}{189/11034} = 0{,}55$$

  Ver

  `https://es.wikipedia.org/wiki/Odds_ratio`

- Según este estudio, las personas que toman aspirinas tienen casi la **mitad de riesgo** de sufrir un ataque al corazón.

Ejemplo

- Pero $\widehat{\theta}$ es solo un estimador del valor **poblacional** desconocido.
- La muestra parece suficientemente grande en el estudio: 22071, aunque la conclusión de que la aspirina funciona bien se basa en **solo** 293 casos observados de ataques al corazón.
- ¿Se puede asegurar que se obtendría el mismo resultado si tomamos otra muestra distinta?
- Según la estadística clásica, aproximando por la distribución normal, se tiene que un intervalo al 95 % para el valor poblacional θ es

$$0{,}43 < \theta < 0{,}70$$

- Parece obvio que nunca excede el valor de 1, es decir que la aspirina en cualquier caso **no** es significativamente mala para la salud.

Razón de odds y distribución asintótica

- Dada una tabla de contingencia 2×2

	C	**D**
A	n_{11}	n_{12}
B	n_{21}	n_{22}

- El estimador que se utiliza para la razón de odds es

$$\widehat{\theta} = \frac{n_{11} n_{22}}{n_{12} n_{21}}$$

- La distribución de este estimador es muy **asimétrica** por lo que para considerar una aproximación a la normal es mejor tomar la transformación $\log(\widehat{\theta})$

Razón de odds y distribución asintótica

- Aplicando el método **delta**

  `https://en.wikipedia.org/wiki/Delta_method`

  Una estimación del error estándar de $\log(\widehat{\theta})$ es

$$\widehat{\sigma}_{\log(\widehat{\theta})} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

  de modo que el correspondiente intervalo de Wald es

$$\log(\widehat{\theta}) \pm z_{\frac{\alpha}{2}} \widehat{\sigma}_{\log(\widehat{\theta})}$$

- Si se toman la función exponencial (*antilogaritmo*) de los extremos se obtiene el intervalo correspondiente para θ.

- El test es algo ***conservador*** (la probabilidad de cubrimiento es algo mayor que el nivel nominal).

Ejemplo sobre aspirina

```
(teta = (104/11037)/(189/11034))
```

```
[1] 0.550115
```

```
SE = sqrt(1/104 + 1/11037 + 1/189 + 1/11034)
LSup = log(teta) + qnorm(1-0.05/2)*SE
LInf = log(teta) - qnorm(1-0.05/2)*SE

exp(LInf)
exp(LSup)
```

```
[1] 0.4324113
[1] 0.699858
```

Otro ejemplo relacionado con aspirinas

- En otro estudio sobre accidentes cerebrovasculares (*ictus*) se observó

	Ictus	**No Ictus**
Aspirina	119	11037
Placebo	98	11034

$$\widehat{\theta} = \frac{119/11037}{98/11034} = 1{,}21$$

- de modo que un intervalo al 95 % para el verdadero valor es θ es $0{,}93 < \theta < 1{,}59$

- Este intervalo incluye al valor *neutro* 1, deduciéndose que la aspirina no es significativamente ni mejor ni peor que el placebo.

Bootstrap

- Se podría decir que la aspirina es *significativamente* beneficiosa para prevenir ataques al corazón pero no es significativamente perjudicial para causar accidentes cerebrovasculares. Pero es un tema que sigue siendo controvertido.

- El bootstrap es un método de simulación que se puede usar en casos como el anterior.

- El término *bootstrap* deriva de la frase ***to pull oneself up by one's bootstrap***. Se basa en el libro del siglo XVIII *las aventuras del barón Munchausen* de Rudolph E. Raspe.

- El barón había caído en el fondo de un profundo lago y se le ocurrió escapar **tirando** de los cordones de sus propias botas...

Ideas de bootstrap en el ejemplo de aspirina

- En el caso del bootstrap, se consideran dos poblaciones:
  - la primera con 11037 observaciones contiene 119 *unos* y 10918 *ceros*
  - la segunda con 11034 observaciones contiene 98 *unos* y 10936 *ceros*.

- Se genera una muestra **con reemplazamiento** de 11037 elementos de la primera población, y una muestra de 11034 elementos de la segunda población.

- Se calcula la proporción respectiva de *unos* en ambas muestras:

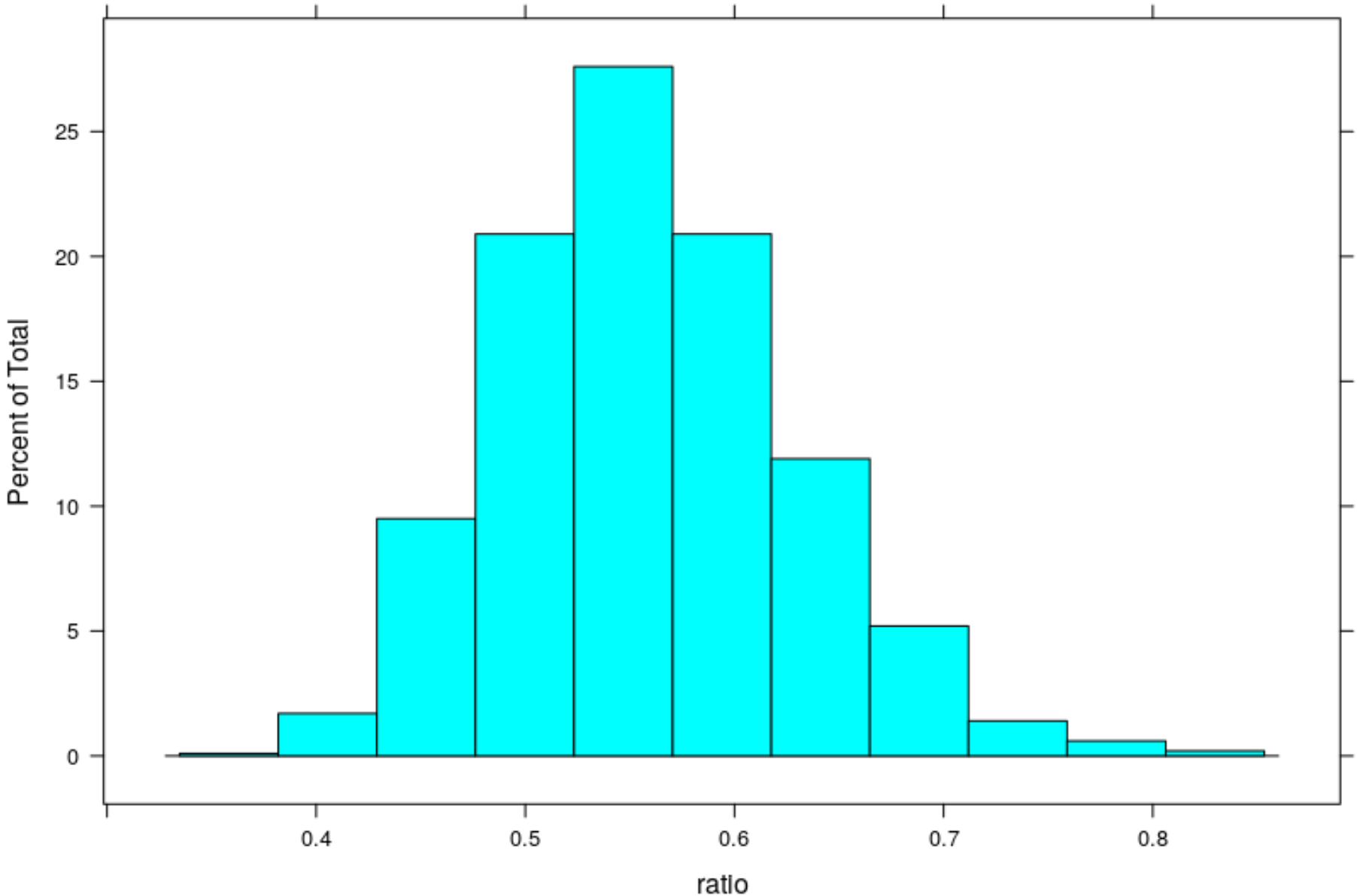
$$\widehat{\theta}^* = \frac{\text{Proporción de “unos” en la muestra 1}}{\text{Proporción de “unos” en la muestra 2}}$$

Bootstrap

- Se repite este proceso un número elevado de veces (digamos $N = 1000$) obteniéndose una muestra de 1000 valores de $\widehat{\theta}^*$.
- Este proceso es fácil de implementar con cualquier software estadístico como `R` o `Python`, por ejemplo.
- Esta muestra de 1000 valores de $\widehat{\theta}^*$ contiene información que se puede usar para realizar inferencia a partir de los datos reales.
- Por ejemplo, Efron y Tibshirani obtienen una desviación estándar muestral alrededor de 0,17 y un intervalo de confianza basado en los cuantiles muestrales alrededor de $(0{,}93; 1{,}60)$.

Bootstrap

- Basta tomar el elemento de la muestra que ocupa el lugar 25 y el que ocupa el lugar 975 una vez ordenada la muestra de los 1000 valores de $\widehat{\theta}^*$.
- Este resultado es similar al obtenido mediante métodos clásicos de inferencia (basados en la aproximación a la distribución normal).
- Pero aquí **NO** se ha usado ningún argumento de aproximación del tipo *Teorema Central del Límite* como en los métodos clásicos.
- Histograma de las estimaciones del ratio


Percent of Total
25
20
15
10
5
0
0.4
0.5
0.6
0.7
0.8
ratio

Ejemplo con Rcpp

- Se escribe un programa en C++ y se graba e.j. en el fichero denominado `boot_ratio2prop.cpp`:
- Se usa la interfaz de R con C++ mediante `Rcpp`.

```
library(Rcpp)

# Se compila el programa escrito en C++
sourceCpp("boot_ratio2prop.cpp")

# Se ejecuta el programa desde R
replica = 2000

sale = boot_ratio2prop(p1, p2, replica)

lattice::histogram(sale)
```

Estudio de la media

- Por ejemplo, se puede considerar el caso de la media muestral, estudiando su precisión como estimador de la media poblacional.

- **Ejemplo**: Ratones bajo tratamiento o no para prolongar su supervivencia después de cirugía invasiva.

Tratamiento	94	197	16	38	99	141	23		
Control	52	104	146	10	51	30	40	27	46

- **Método**: comparación de las medias de tiempos de vida entre el tratamiento y el control

- Las medias de cada grupo son
  (I) Tratamiento: 86.86
  (II) Control: 56.22

Ejemplo ratones

- La diferencia entre ambas medias $\boxed{\bar{x} - \bar{y} = 30{,}63}$ sugiere que hay **bastante efecto** del tratamiento.

- ¿Cómo son de precisos estos estimadores?
  Resulta que las muestras son muy *pequeñas*...

- En el caso de la media, el error estándar estimado basado en n valores de una *m.a.s.* $x_1, x_2, \ldots, x_n$ es

$$\sqrt{\frac{s^2}{n}}$$

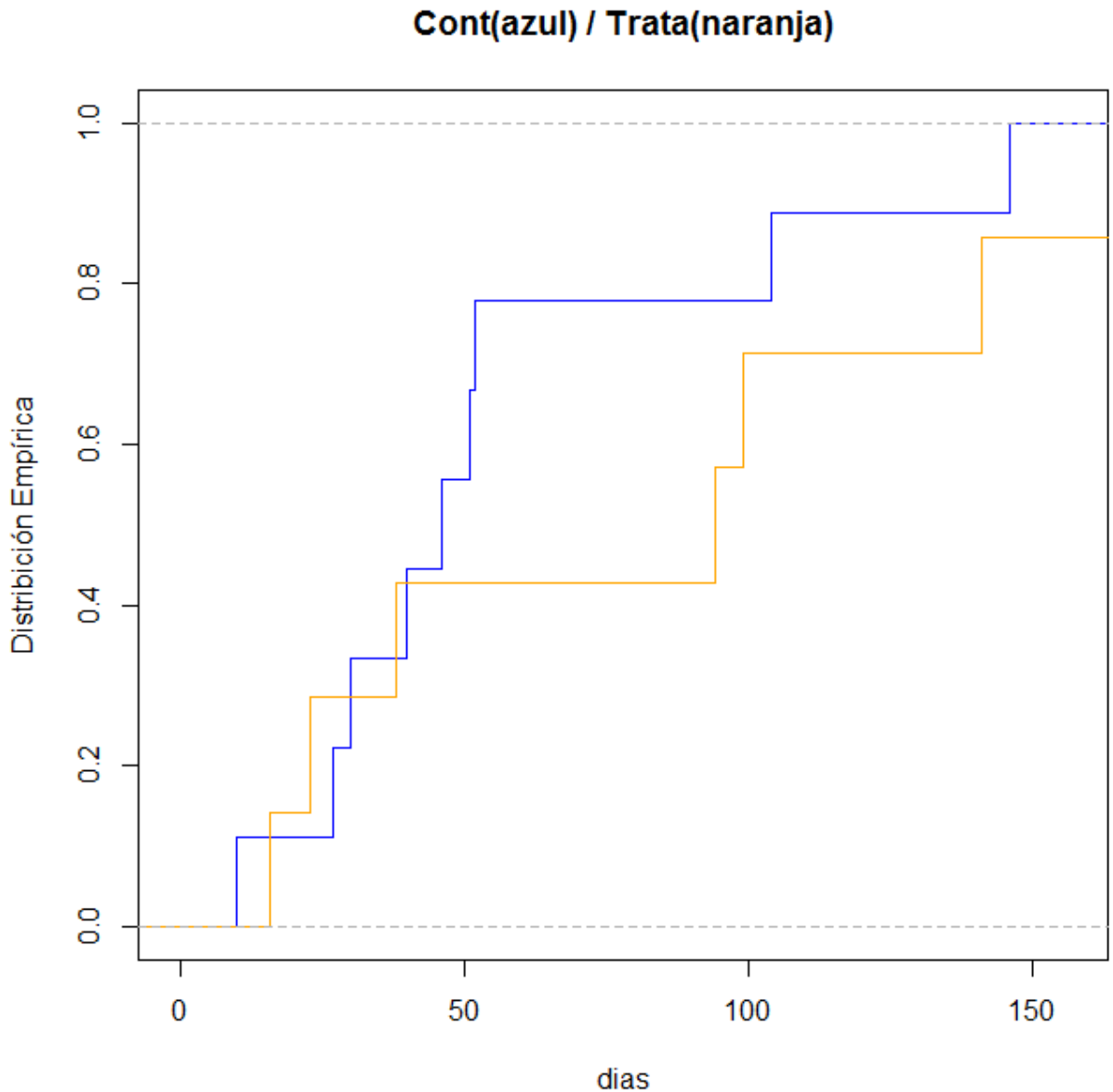
donde

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

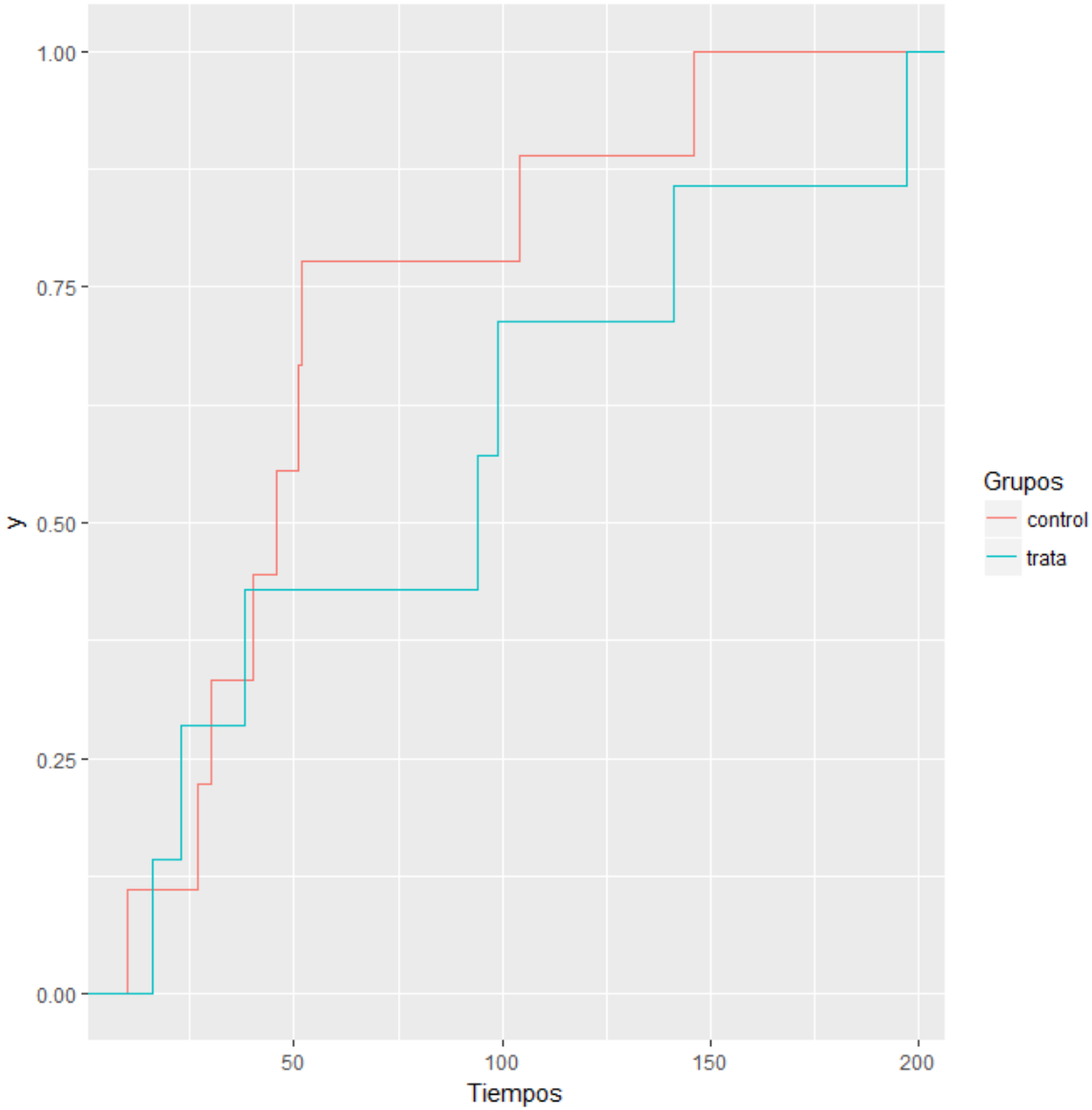
Ejemplo ratones

- Si el error estándar de las estimas de las medias es pequeño, entonces la diferencia entre las medias de supervivencia sería significativa.
- Los errores estándar de x e y son $SE_x = 25{,}24$ y $SE_y = 14{,}14$
- El error estándar para la diferencia $\bar{x} - \bar{y}$ es igual a $\sqrt{25{,}24^2 + 14{,}14^2} = 28{,}93$
- Pero la diferencia observada es 30.63. Si se calcula el estadístico de contraste: $30{,}63/28{,}93 = 1{,}05$, es decir es **mucho menor** que el valor 1.96 típico de la distribución normal.
- Esto implica que el tratamiento NO tiene efectos significativos en la supervivencia de los ratones.

Distribuciones Empiricas



Distribuciones Empiricas



Cálculo de medianas

- El cálculo de los errores estándar es un método para estudiar la precisión de los estimadores. Pero, salvo en el caso de la media, no hay en general fórmulas concretas para estimarlos.

- Por ejemplo, si se calculan las medianas en el ejemplo de los ratones, se tiene que la mediana para el tratamiento es 94 y para el control es 46.

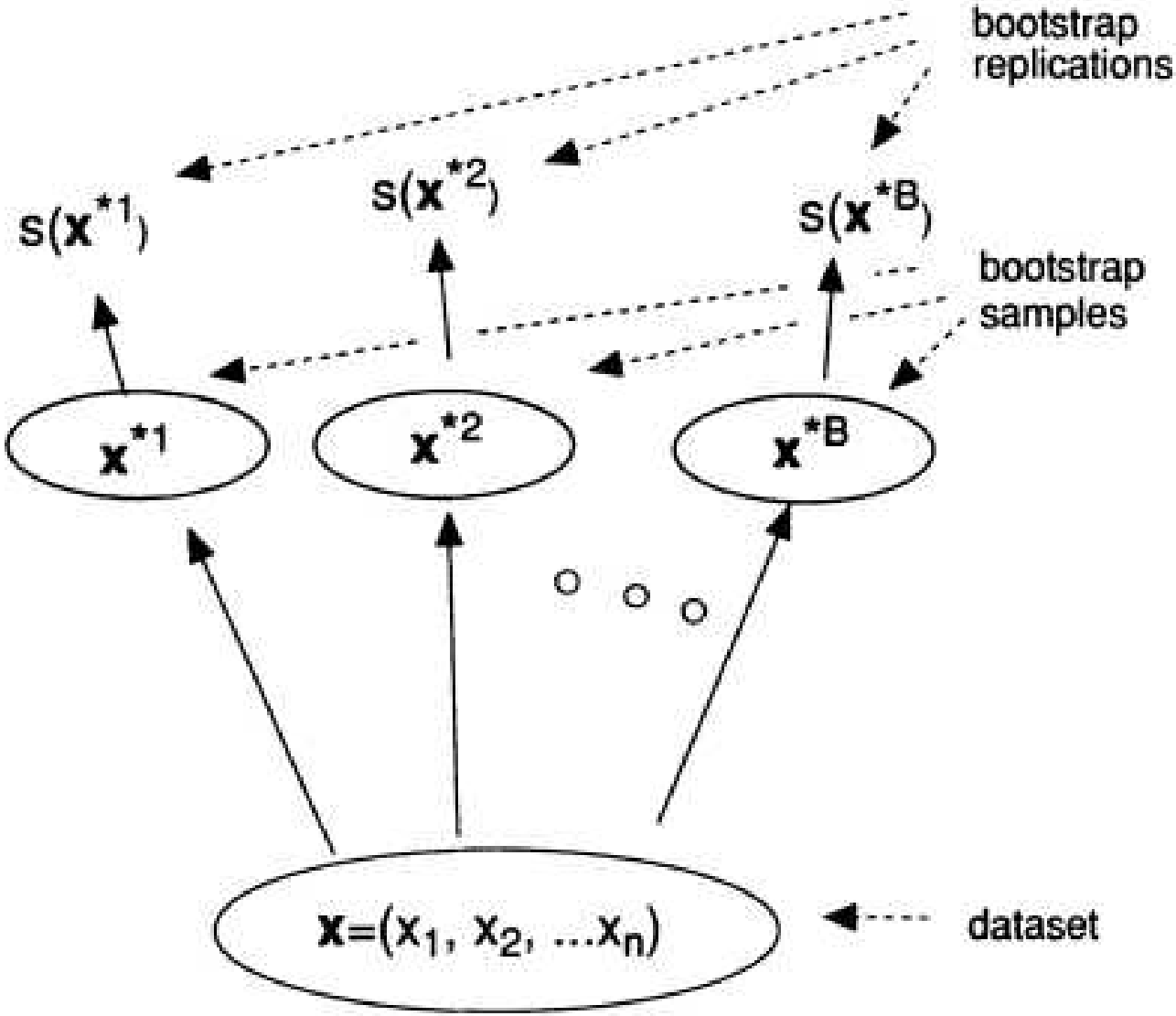
- La diferencia entre medianas es 48, mucho **mayor** que la diferencia entre medias.

- Pero, ¿cómo es de precisa esta estimación?
  La única manera de responder es usando bootstrap en este caso.

Esquema del Bootstrap

- Supongamos que se observa una muestra $\mathbf{x} = x_1, x_2, \ldots, x_n$ sobre la que se calcula un cierto estadístico $s(\mathbf{x})$.
- Por ejemplo, $\mathbf{x}$ es el grupo control de observaciones y $s(\mathbf{x})$ es la media muestral.
- En el caso del bootstrap, se define una *muestra bootstrap* $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ que se obtiene muestreando n veces **con reemplazamiento** a partir de los datos originales $(x_1, x_2, \ldots x_n)$
- Por ejemplo si $n = 7$ una posible muestra bootstrap podría ser

$$\mathbf{x}^* = (x_5, x_7, x_5, x_4, x_7, x_3, x_1)$$


bootstrap
replications
$s(x^{*1})$
$s(x^{*2})$
$s(x^{*B})$
bootstrap
samples
x^{*1}
x^{*2}
x^{*B}
$x=(x_1, x_2, \ldots x_n)$
dataset

Algoritmo del Bootstrap

- Se genera un número B elevado de muestras bootstrap $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$ cada una de tamaño n. Los tamaños típicos de B para errores estándar suelen estar entre 500 y 5000.
- Para cada muestra bootstrap $b = 1, \ldots, B$ se calcula el estadístico $s\left(\mathbf{x}^{*b}\right)$. Por ejemplo, la *mediana*.
- El estimador bootstrap del error estándar es la desviación estándar de las B muestras bootstrap

$$\widehat{se}_{Boot} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(s\left(\mathbf{x}^{*b}\right) - s(\cdot)\right)^2}$$

donde

$$s(\cdot) = \frac{1}{B}\sum_{b=1}^{B} s\left(\mathbf{x}^{*b}\right)$$

Ejemplo con la mediana

- Por ejemplo, Efron y Tibshirani obtienen en el grupo control un error estándar igual a 11.54 y en el grupo tratamiento, 36.35 basándose en $B = 100$ réplicas.
- Así, usando ese número de réplicas la diferencia de medianas observada igual a 48 se obtiene un error estándar estimado igual a

$$\sqrt{36{,}352 + 11{,}542} = 38{,}14$$

- Así el estadístico es $48/38{,}14 = 1{,}26$ tampoco es significativamente mayor que 0, aunque es mayor que en el caso de la media.
- Para la mayoría de los estadísticos **no existen** fórmulas explícitas que sirvan para calcular el error estándar, por ello se puede usar un procedimiento bootstrap.

Programas para el ejemplo de los ratones

- Se puede programar fácilmente el ejemplo de los ratones con las librerías `bootstrap` y `boot`.
- Como introducción al manejo de la librería `boot`, se pueden consultar las siguientes páginas Web:

```
www.mayin.org/ajayshah/KB/R/documents/boot.html

cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf

ww2.coastal.edu/kingw/statistics/R-tutorials/resample.html
```