Charles Edward Spearman (1863-1945, Londres, UK)

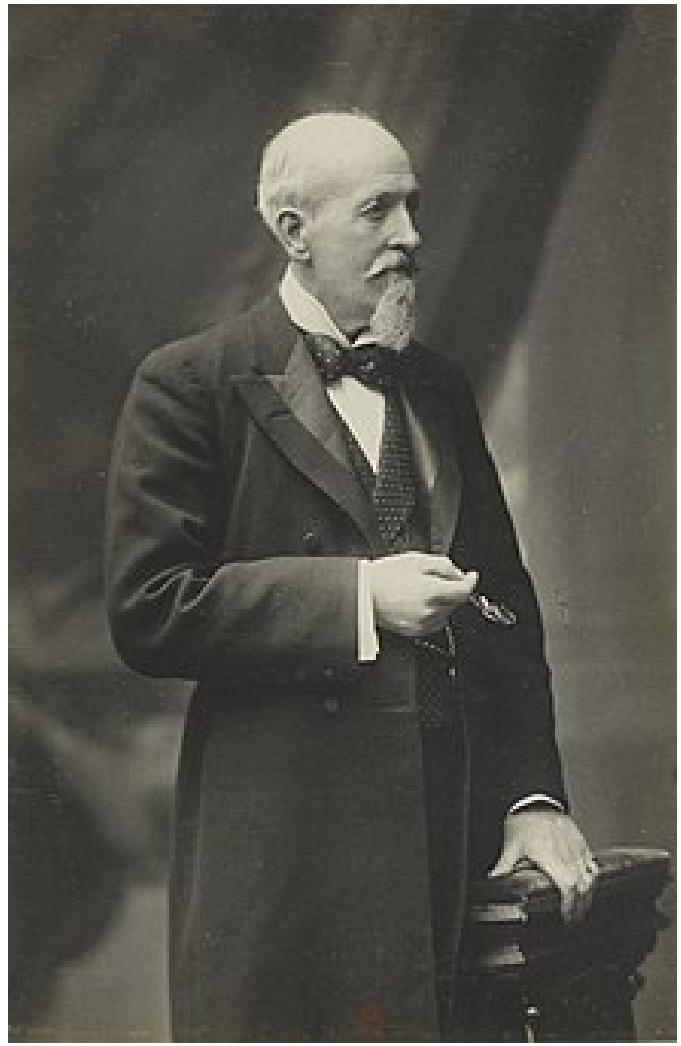


Figura: Spearman fue un psicólogo inglés. Contribuyó además en el estudio de la inteligencia humana

Clase 22 de Bioestadística

Test de aleatoriedad (Spearman)

Ernesto Mordecki

CMAT, Facultad de Ciencias, Universidad de la República.

Uruguay

Contenidos de la clase

Test de aleatoriedad

- Nuestro punto de partida en estadística son las muestras aleatorias simples (m.a.s.)
- Eso supone (i) independencia y (ii) equidistribución
- Queremos diseñar un test que nos indique desvíos de estas hipótesis
- Hay muchos tipos de desvíos.
- Hoy estudiamos la posibilidad de que los datos sean crecientes o decrecientes, en cuyo caso se viola la independencia
- ¿Cómo construímos un estadístico para este test?

Estadísticos de orden

- Existen varios procedimientos estadísticos útiles para trabajar con una muestra

$$X_1, X_2 \ldots, X_n$$

- Uno de ellos es es calcular los estadísticos de orden
- Se trata de los mismo datos pero considerados en orden creciente, se escribe:

$$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$$

- En particular

$$X_{(1)} = \text{mín}\{X_k : k = 1, \ldots, n\}$$

$$X_{(n)} = \text{máx}\{X_k : k = 1, \ldots, n\}$$

- Por ejemplo si

$$X_1, X_2 \ldots, X_5 = 1{,}04,\ 0{,}11,\ -1{,}75,\ 1{,}55,\ 0{,}31$$

- los estadísticos de orden son

$$X_{(1)}, X_{(2)}, \ldots, X_{(5)} = -1{,}75,\ 0{,}11,\ 0{,}31,\ 1{,}04,\ 1{,}55$$

- Estas muestras se generaron en R mediante `mas<-rnorm(5)` y se aplica el comando `sort` (ordenar)

```
> sort(mas)
[1] -1.75  0.11  0.31  1.04  1.55
```

Rangos

- El lugar que ocupa en la muestra ordenada cada dato es se define como su rango
- En nuestro ejemplo tenemos los rangos del 1 al 5:

Dato	1.04,	0.11,	-1.75,	1.55,	0.31
Rango	4	2	1	5	3

- Entonces el rango 1 corresponde al mínimo (el dato menor) mientras que el rango 5 corresponde al máximo (el dato mayor)

- La notación para los rangos de una muestra es

$$R_1, \ldots, R_n$$

- Para nuestra muestra entonces

$$R_1, \ldots, R_n = 4,\ 2,\ 1,\ 3,\ 5$$

- En R se utiliza el comando `rank`:

```
> rank(mas)
[1] 4 2 1 5 3
```

- Que nos da `rank(sort(mas))`? y `rank(-sort(mas))`?

Test de rangos de Spearman de una muestra

- Queremos distinguir si nuestros datos son i.i.d. o tienen algún patrón de crecimiento o decrecimiento
- Formulamos entonces

$$H_0 : X_1, \ldots, X_n \text{ son i.i.d}$$

contra

$$H_1 : X_1, \ldots, X_n \textbf{ no } \text{son i.i.d}$$

- Spearman propuso comparar los rangos de la muestra de estudio con los números 1,2,. . . ,n
- Si encontramos similitud (aunque no estén totalmente ordenados) se rechaza la hipótesis nula.

- Cómo medir la similitud entre dos n-plas de números?
- Spearman propuso utilizar la correlación de Pearson entre los rangos[1]
- Si denominamos $N = 1, 2, \ldots, n$ los números, el estadístico es entonces

$$\rho_s = \frac{cov(R, N)}{\sigma_R \sigma_N}$$

[1] Otro test consiste en calcular la correlación entre los datos directamente, sin tomar los rangos.

Tenemos

- $cov(R, N) = \frac{1}{n}\sum_{k=1}^{n} kR_k - \left(\frac{1}{n}\sum_{k=1}^{n} k\right)\left(\frac{1}{n}\sum_{k=1}^{n} k\right)$
- $\sigma_R^2 = \sigma_N^2 = \frac{n^2-1}{12}$
- Para calcular el estadístico se calcula primero la suma de las diferencias:

$$S = \sum_{k=1}^{n}(R_k - k)^2$$

- Se demuestra que

$$\rho_s = 1 - \frac{6S}{n(n^2-1)}$$

Ventajas y desventajas

- La mayor desventaja es que sólo detecta desvíos de la hipótesis nula consistentes en datos aproximadamente crecientes o datos aproximadamente decrecientes
- Como ventajas, es un test no paramétrico, en el sentido de que al tomar los rangos, el comportamiento estadístico no depende de la distribución

Determinación de la región crítica

- El estadístico del test siempre cumple

$$-1 \leq \rho_s \leq 1$$

  por ser una correlación

- Los valores cercanos a 1 (-1)indican datos aproximadamente crecientes (decrecientes)

- Bajo la hipótesis nula, el estadístico, es simétrico, y cumple

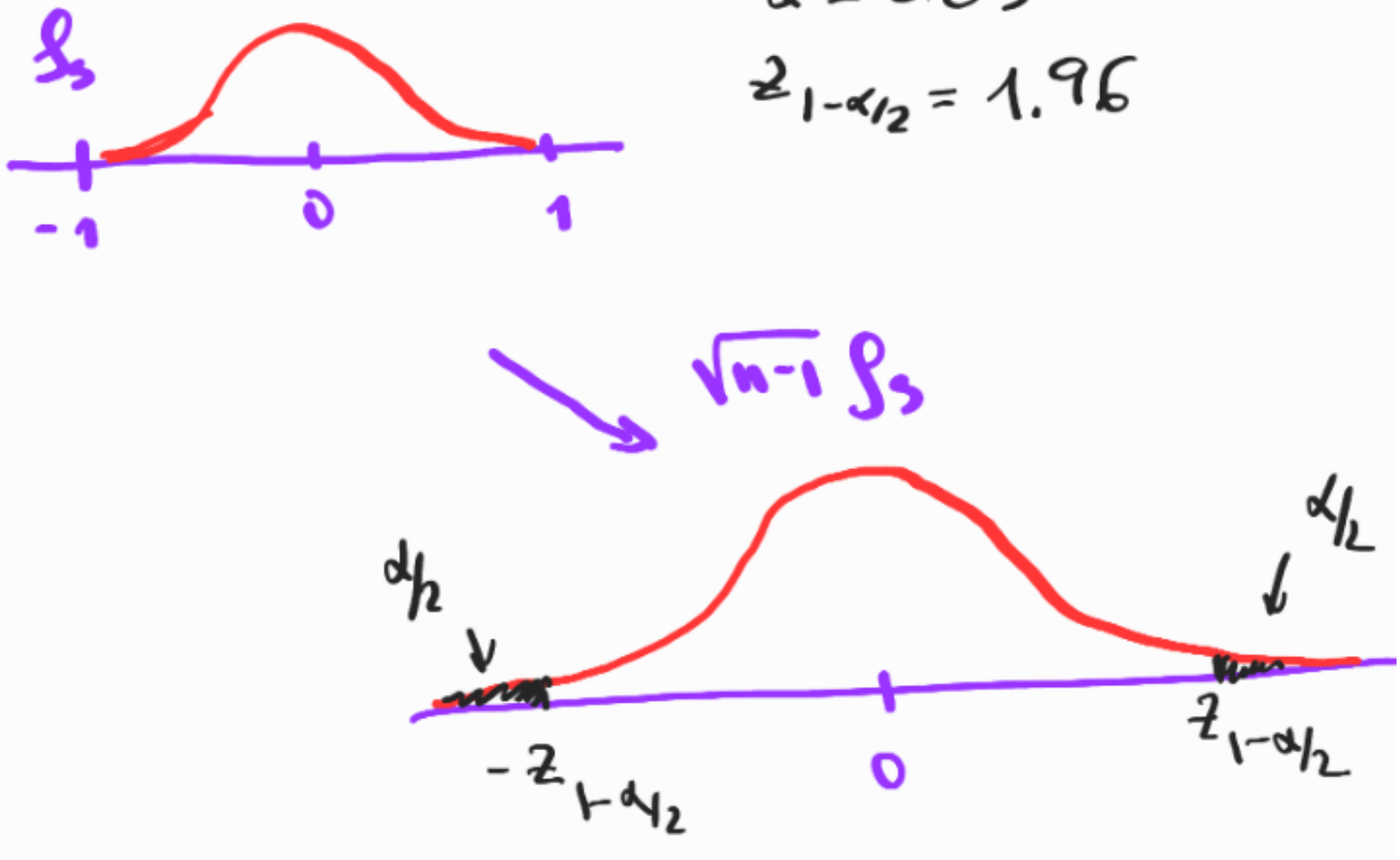
$$E(\rho_s) = 0, \quad var(\rho_s) = \frac{1}{n-1}$$

- Para valores pequeños ($n < 30$) se calcula la distribución exacta, que se presenta en tablas
- Se puede calcular entonces una forma de TCL que nos aproxima a una variable normal, independientemente de la distribución
- En ningún caso el cálculo depende de la distribución de los datos.
- La variable estandarizada ($n \geq 30$) es

$$\sqrt{n-1}\rho_s \approx \mathcal{N}(0,1)$$

- Para el test de significancia α, la región crítica es

$$\sqrt{n-1}|\rho_s| \geq z_{1-\alpha/2}$$


$\alpha = 0.05$
$z_{1-\alpha/2} = 1.96$
ρ_s
-1
0
1
$\sqrt{n-1}\,\rho_s$
$\alpha/2$
$\alpha/2$
$-z_{1-\alpha/2}$
0
$z_{1-\alpha/2}$

El test en R

- R tiene implementado el test de rangos de una muestra de Spearman
- Está en un paquete que incluye:
  - El test de correlación directa de Pearson (es decir, sin tomar los rangos). Este test para muestras chicas depende de la distribución
  - El test de rangos de una muestra (Test de Spearman) que estudiamos
  - El test de rangos de dos muestras, que estudiaremos a continuación

- El comando es
  ```
  cor.test(datos,sort(datos),method="spearman")
  ```
- `cor.test` es por test de correlación
- `datos` es nuestra muesta de datos (como vector)
- El input son siempre dos muestras
- Si queremos hacer un test de **una** muestra, ponemos en el lugar de la segunda la muestra ordenada `sort(datos)`
- Veamos un ejemplo:

Ejemplo 1:

```
> # Aplicamos el test de rangos de Spearman
> # Primero con una muestra
> n<-30
> datos<-rnorm(30)
> cor.test(datos,sort(datos),method = "spearman")

        Spearman's rank correlation rho

data:  datos and sort(datos)
S = 5370, p-value = 0.3013
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1946607
```

- No se rechaza H_0 porque el p-valor es mayor que 0.05
- La muestra no es “sospechosa”

Ejemplo 1:

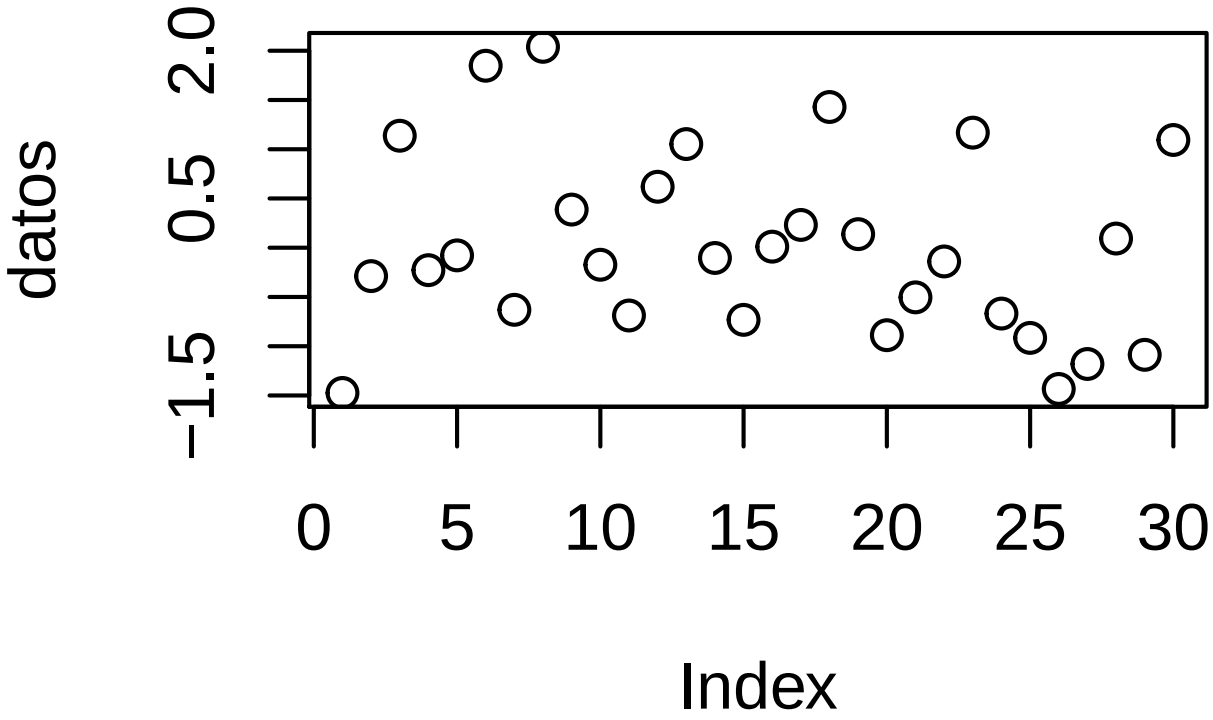


Figura: No se nota tendencia en los datos

Ejemplo 2:

Agregamos una tendencia creciente a los datos

```
> n<-30
> coef<-0.1
> datos2<-rnorm(30)+coef*(1:n)
> cor.test(datos2,sort(datos2),method = "spearman")

        Spearman's rank correlation rho

data:  datos2 and sort(datos2)
S = 1734, p-value = 0.0004029
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.614238
```

Ejemplo 2:

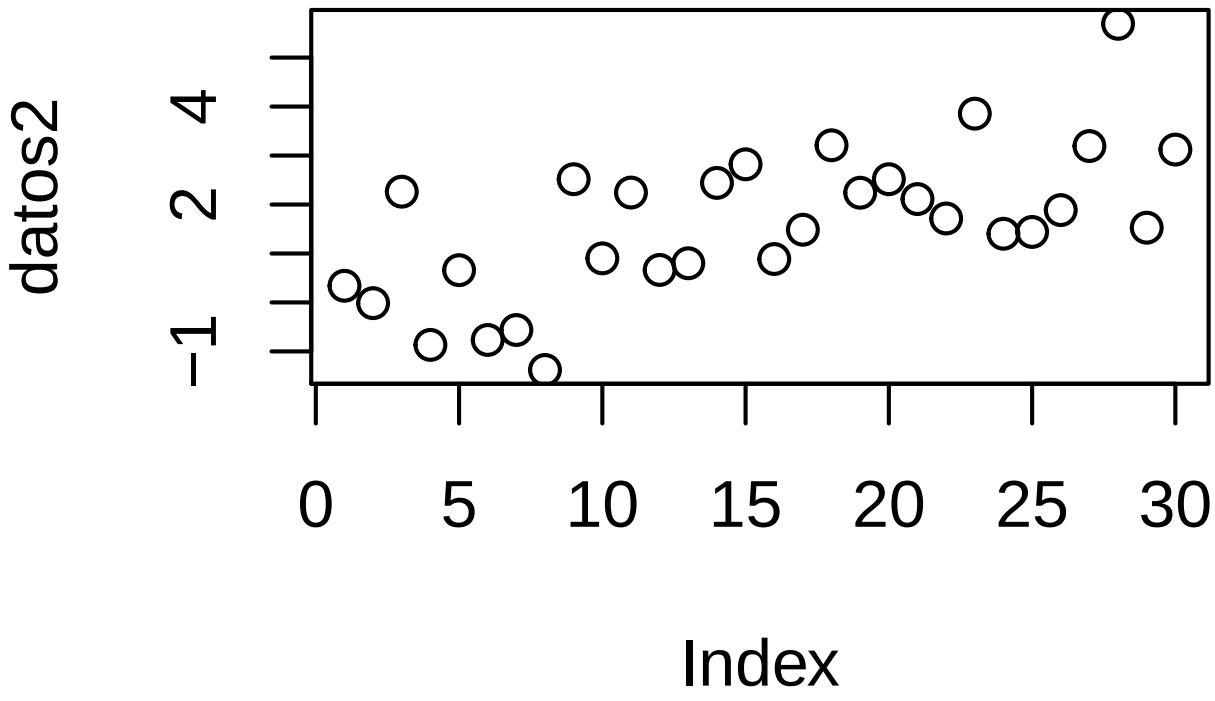


Figura: Se nota tendencia creciente en los datos

Ejemplo 3:

Agregamos una pequeña tendencia creciente a los datos

```
> n<-30
> coef<-(-0.05)
> datos3<-rnorm(30)+coef*(1:n)
> cor.test(datos3,sort(datos3),method = "spearman")

        Spearman's rank correlation rho

data:  datos3 and sort(datos3)
S = 6472, p-value = 0.01579
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.439822
```

- Se rechaza la hipótesis nula. El *p*-valor es menor que 0,05

Ejemplo 3:

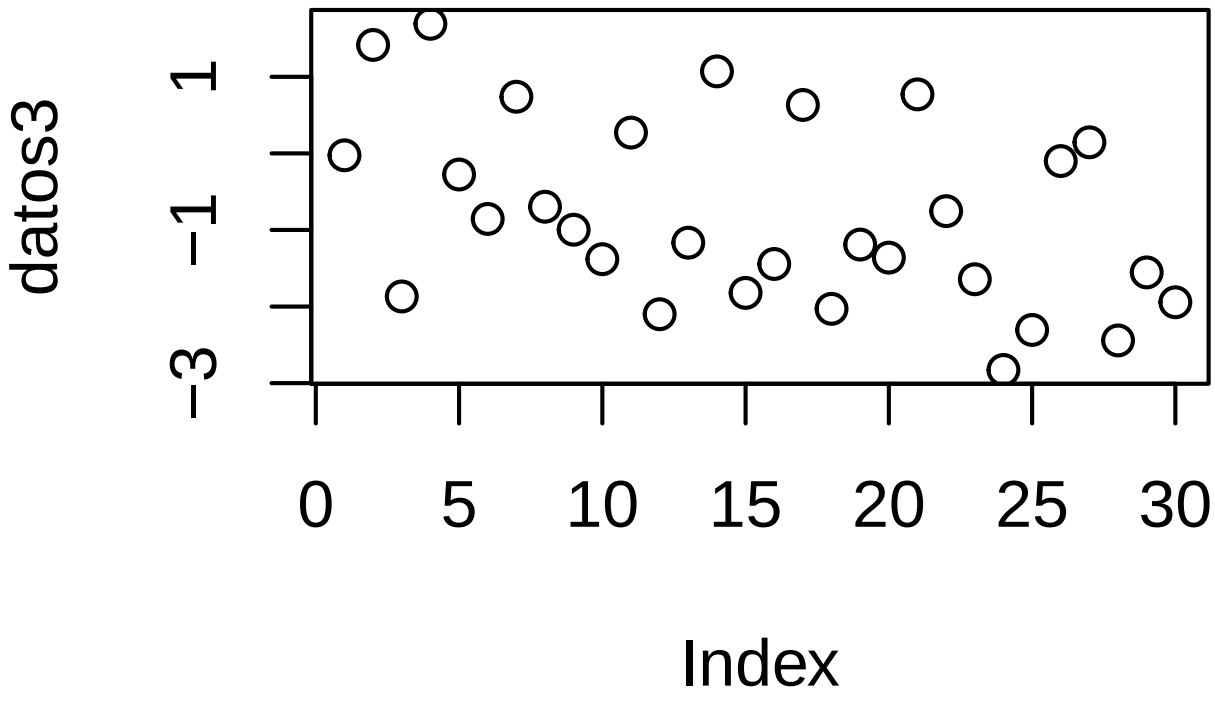


Figura: ¿Se nota una tendencia decreciente en los datos?

Ejemplo 4:

Ahora agregamos una tendencia quebrada, que sube primero y baja después:

```
> coef<-0.1
> sesgo<-c(c(1:15),c(15:1))
> datos4<-rnorm(30)+coef*sesgo
> cor.test(datos4,sort(datos4),method = "spearman")

        Spearman's rank correlation rho

data:  datos4 and sort(datos4)
S = 5424, p-value = 0.2719
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.2066741
```

¿Cuál es la conclusión?

- El test no rechaza
- El p-valor es mayor que 0,05
- No se da cuenta lo que ocurre

Ejemplo 4:

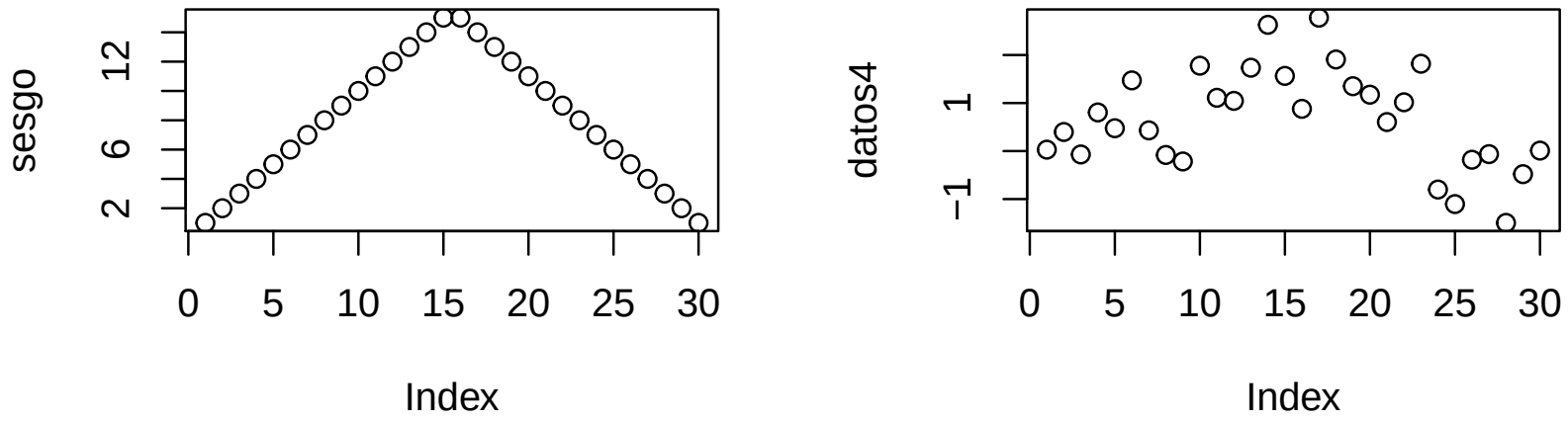


Figura: Los datos no son i.i.d. Se agregaron a los datos i.i.d. una pequña perturbación con la forma del sesgo

Test de dos muestras

- Queremos testear la independencia de dos muestras aleatorias simples:

$$H_0 : (X_k)_{1 \leq k \leq n} \text{ y } (Y_k)_{1 \leq k \leq n} \text{ son idependientes}$$

- La hipótesis alternativa es muy amplia, y el estadístico que propuso Spearman detecta crecimiento o decrecimiento simultáneo, lo que indica dependencia
- A datos mayores de X corresponden datos mayores de Y si la asociación es positiva
- Podrían ser menores las Y y la asociación es negativa

- Para eso comparamos los rangos R^X de la muestra X con los rangos R^Y de la muestra Y

- Calculamos

$$S = \sum_{k=1}^{n} (R_k^X - R_k^Y)^2$$

- Si tenemos asociación positiva, los rangos menores de X corresponden a rangos menores de Y y tenemos un S relativamente pequeño

- En caso de asociación negativa, los valores pequeños de R^X corresponden a valores grandes de R^Y y reciprocamente, eso produce valores grandes de S

- La correlación de Spearman es

$$\rho_s = 1 - \frac{6S}{n(n^2 - 1)}$$

Región crítica

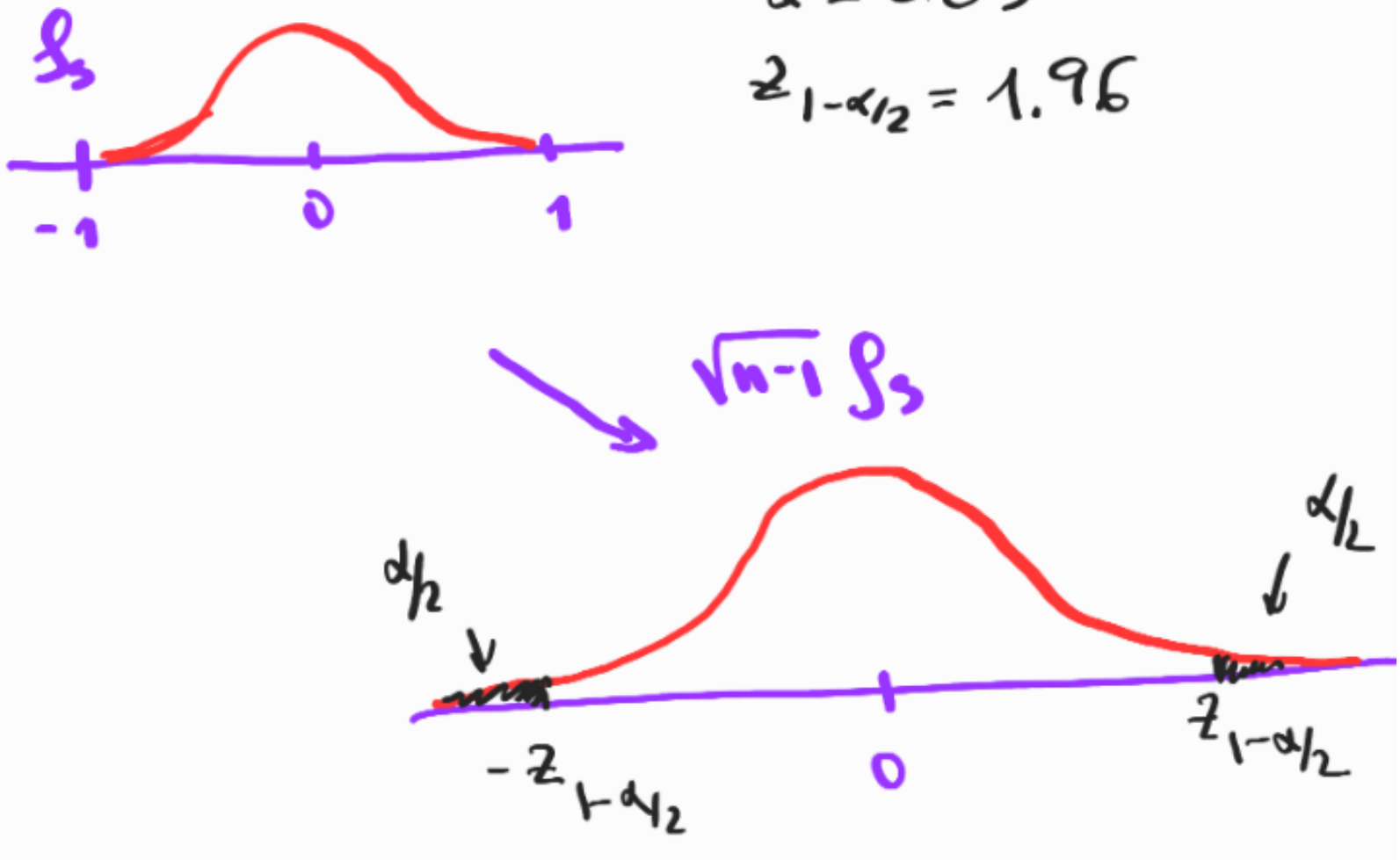
- La determinación de la región crítica es similar
- Para valores grandes de n tenemos

$$\sqrt{n-1}\rho_s \approx \mathcal{N}(0,1)$$

- La región crítica es entonces de la misma forma

$$\sqrt{n-1}|\rho_s| \geq z_{1-\alpha/2}$$

para significancia α


$\alpha = 0.05$
$z_{1-\alpha/2} = 1.96$
ρ_s
-1
0
1
$\sqrt{n-1}\,\rho_s$
$\alpha/2$
$\alpha/2$
$-z_{1-\alpha/2}$
0
$z_{1-\alpha/2}$

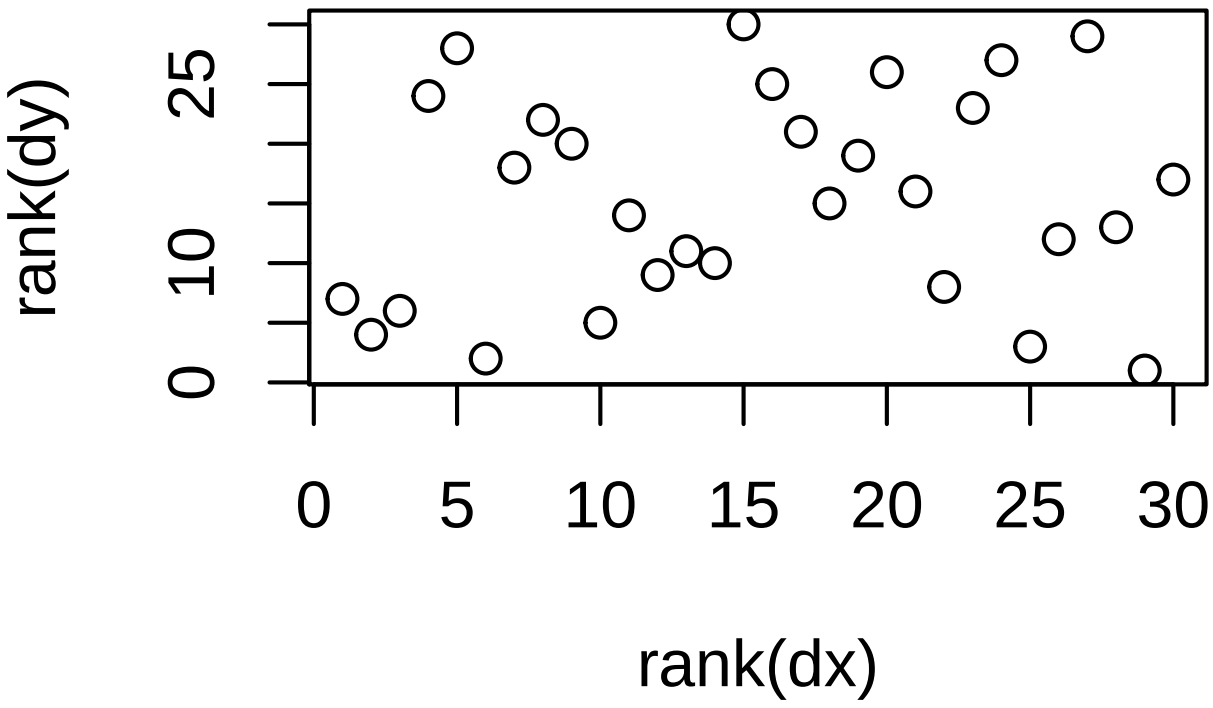
Ejemplo 1: test de dos muestras

```
> n<-30
> dx<-rnorm(n)
> dy<-runif(n)
> cor.test(dx,dy,method="spearman")

        Spearman's rank correlation rho

data:  dx and dy
S = 3954, p-value = 0.5249
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.120356
```

- Utilizamos dos muestras independientes
- No se rechaza la hipótesis nula
- El *p*-valor es mayor que $\alpha = 0{,}05$



No se ve asociación entre rangos pequeños y grandes

Ejemplo 2: test de dos muestras

```
> plot(dx,dy)
> dx2<-dx
> dy2<-dy+0.1*dx
> cor.test(dx2,dy2,method="spearman")

        Spearman's rank correlation rho

data:  dx2 and dy2
S = 2580, p-value = 0.01972
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4260289
```

- Partimos dos las mismas muestras independientes
- Contaminamos la segunda muestra con un 10 % de los valores de la primera
- El estadístico lo nota y se rechaza la hipótesis nula

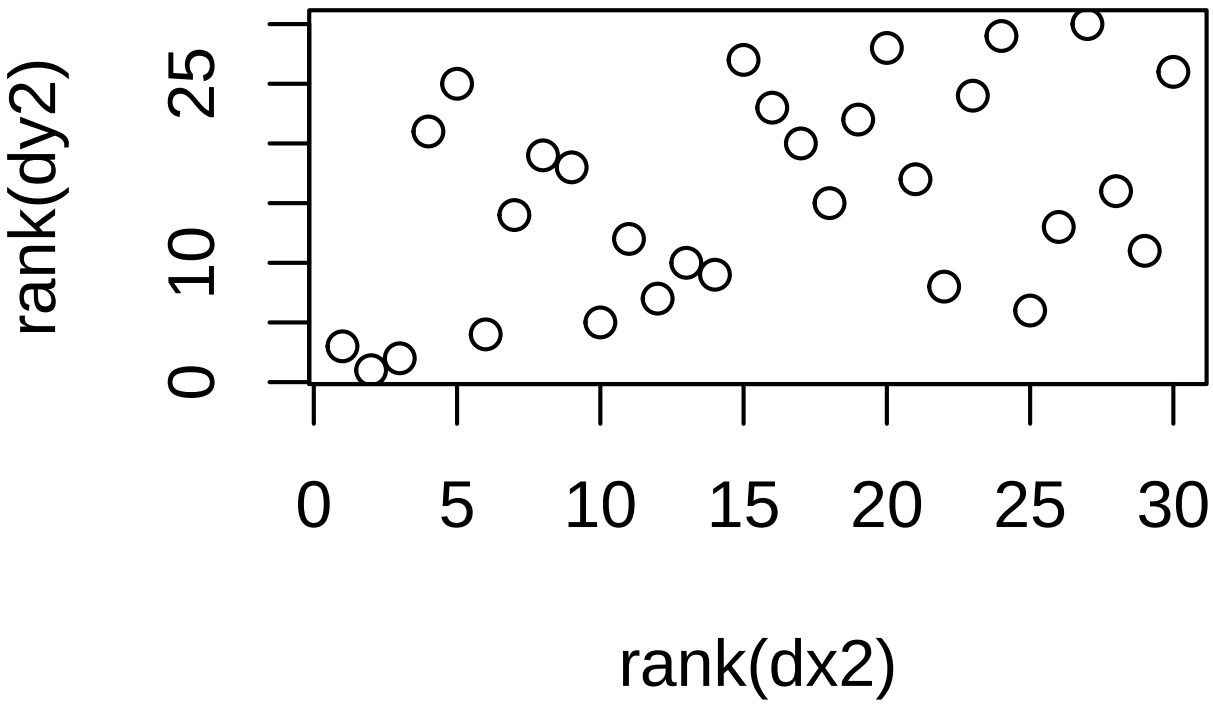


Figura: ¿Se nota la contaminación en los rangos?