

Coeficiente de correlación (Bootstrap)

María Leyenda Rodríguez

04/03/2010

Coefficiente de correlación (Bootstrap)

Sea $X=(Y,Z)$ una variable aleatoria bidimensional y consideremos $(Y_1,Z_1),\dots,(Y_n, Z_n)$ una m.a.s. de X . Denotemos por F la distribución de X (distribución conjunta) y por F_Y, F_Z las marginales correspondientes, ambas distribuciones Normales. Queremos contrastar:

$$H_0: Y, Z \text{ independientes, vs, } H_a: Y, Z \text{ dependientes}$$

Bajo la hipótesis nula de independencia, el coeficiente de correlación $\rho=0$. Si consideramos como estadístico de contraste el coeficiente de correlación muestral, $\hat{\rho}$, rechazaremos H_0 cuando el coeficiente de correlación muestral sea grande. Por tanto, el p-valor del estadístico de contraste será $\mathbb{P}(\hat{\rho} \geq \hat{\rho}_{obs} | H_0)$.

Bajo H_0 , podemos considerar la muestra observada como:

$$(Y_{(1)}, \dots, Y_{(n)}, Z_{(1)}, \dots, Z_{(n)})$$

por lo que podemos obtener remuestras como:

$$X^* = \{(y_{(1)}, Z_1^*), \dots, (y_{(n)}, Z_n^*)\}, \quad X^* = \{(Y_1^*, z_{(1)}), \dots, (Y_n^*, z_{(n)})\}$$

$$X^* = \{(Y_1^*, Z_1^*), \dots, (Y_n^*, Z_n^*)\}$$

1.- Caso 1: Considera Y, Z variables independientes con distribución $F_Y=N(0,4)$ y $F_Z=N(0,1)$. Para $n=20$ y $B= 500$, obtén una estimación de la distribución de $\hat{\rho}$. Genera una muestra de (Y,Z) y aproxima el p-valor asociado al coeficiente de correlación muestral.

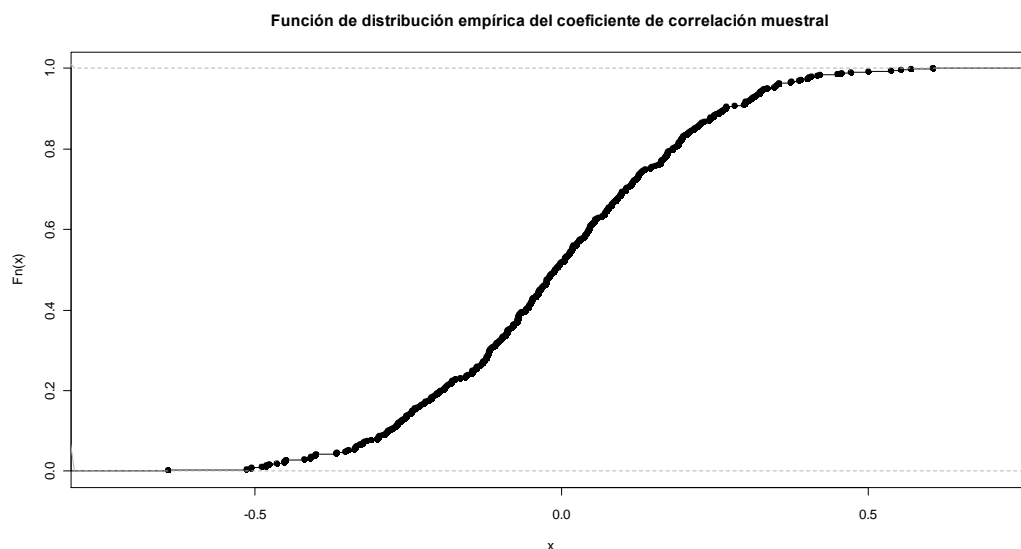
En primer lugar, se obtiene una estimación de la distribución de $\hat{\rho}$ mediante Bootstrap. Para ello se genera una muestra de Y , una muestra de Z ; ambas de tamaño $n=20$.

A continuación, se obtiene $B=500$ remuestras de cada una de las muestras obtenidas. De esta forma se consiguen 500 remuestras de la forma $X^* = \{(Y_1^*, Z_1^*), \dots, (Y_n^*, Z_n^*)\}$, pues las variables son independientes. Luego, es calculado el coeficiente de correlación muestral para cada una de las remuestras.

Finalmente, se obtiene la distribución empírica del coeficiente de correlación muestral mediante aproximación por Monte-Carlo (SLLN).

$$\hat{G}(x) = \mathbb{P}_F(\bar{X} \leq x) = Dist_{T,F}(x) \approx \frac{1}{B} \sum_{b=1}^B I \{T(X^{*(b)}) \leq x\}$$

La distribución empírica es un buen estimador de la función de distribución.



En segundo lugar, se genera una muestra de la variable bidimensional $X_0 = (Y_0, Z_0)$; donde Y_0, Z_0 son variables con distribución $F_{Y_0} = N(0,4)$ y $F_{Z_0} = N(0,1)$ independientes y se aproxima el p-valor asociado al coeficiente de correlación muestral.

Como se dijo anteriormente el estadístico del contraste es $\mathbb{P}(\hat{\rho} \geq \hat{\rho}_{obs} | H_0)$. Por tanto, el p-valor asociado al coeficiente de correlación muestral.

$$\mathbb{P}(\hat{\rho} \geq \hat{\rho}_{obs} | H_0) = 1 - \mathbb{P}(\hat{\rho} < \hat{\rho}_{obs} | H_0) = F_n(\hat{\rho}_{obs}) = 0.692$$

Siendo $\hat{\rho}_{obs}$ el coeficiente de correlación muestral asociado a la muestra de la variable bidimensional $X = (Y, Z)$ y F_n la función de distribución empírica del coeficiente de correlación muestral.

Nótese que el resultado se ha obtenido por simulación, es decir, depende de la muestra y de las remuestras generadas.

En este caso, se obtiene que se acepta la hipótesis nula, es decir, que se acepta que las variables Y_0 e Z_0 sean independientes.

1.- Caso 2: Considera Y, Z variables Normales con media 0, varianzas $\sigma_Y^2 = 4, \sigma_Z^2 = 1$, independientes. Para $n=20$ y $B= 500$, obtén una estimación de la distribución de $\hat{\rho}$ y del p-valor asociado al coeficiente de correlación de una muestra de $(X, Y) \sim (F_Y, F_Z)$ con coeficiente de correlación $\rho=0.25$.

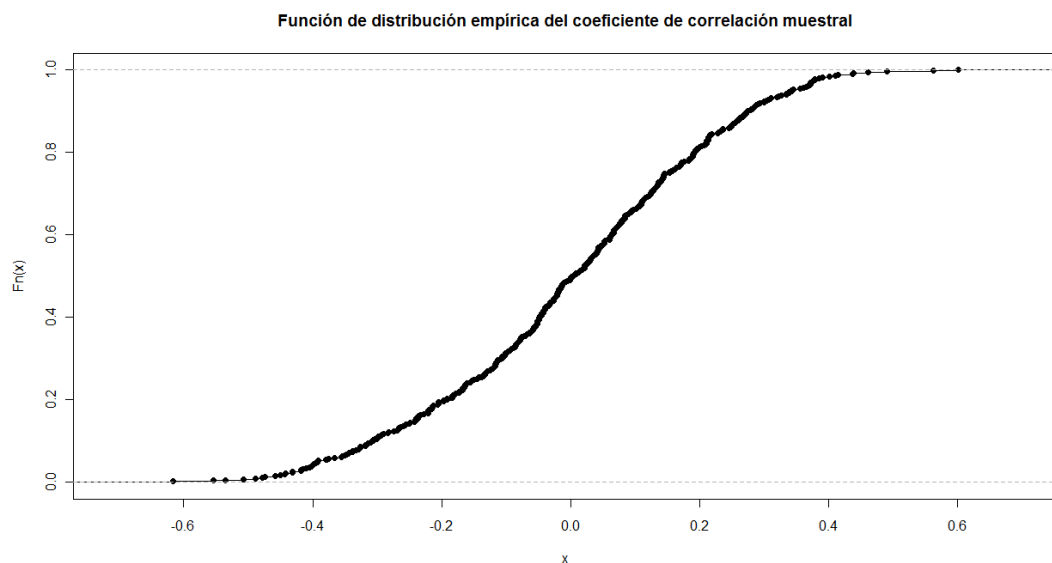
En primer lugar, se obtiene una estimación de la distribución de $\hat{\rho}$ mediante Bootstrap. Para ello se genera una muestra de Y , una muestra de Z ; ambas de tamaño $n=20$.

A continuación, se obtiene $B=500$ remuestras de cada una de las muestras obtenidas. De esta forma se consiguen 500 remuestras de la forma $X^* = \{(Y_1^*, Z_1^*), \dots, (Y_n^*, Z_n^*)\}$, pues las variables son independientes. Luego, es calculado el coeficiente de correlación muestral para cada una de las remuestras.

Finalmente, obtenemos la distribución empírica del coeficiente de correlación muestral mediante aproximación por Monte-Carlo (SLLN).

$$\hat{G}(x) = \mathbb{P}_F(\bar{X} \leq x) = \text{Dist}_{T,F}(x) \approx \frac{1}{B} \sum_{b=1}^B I\{T(X^{*(b)}) \leq x\}$$

La distribución empírica es un buen estimador de la función de distribución.



En segundo lugar, se estima el p-valor asociado al coeficiente de correlación de una muestra de $X=(Y, Z) \sim (F_Y, F_Z)$ con coeficiente de correlación $\rho=0.25$. Para ello se genera la muestra bidimensional con vector de medias $(0,0)$ y matriz de varianzas covarianzas $\begin{pmatrix} \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{ZY} & \sigma_Z^2 \end{pmatrix} = \begin{pmatrix} 4 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ donde $0.5 = \sigma_{YZ} = \sigma_{ZY} = \rho \sigma_Y \sigma_Z$. Esta muestra bidimensional es generada en R mediante el comando *rmvnorm* que se encuentra en el paquete *mvtnorm* (*Multivariate normal density and random deviates*). La generación de la muestra se lleva a cabo mediante la siguiente sentencia:

```
rmvnorm(2*n, mean=c(0,0), sigma=matrix(c(4,cov,cov,1),ncol=2))
```

Como el estadístico del contraste es $\mathbb{P}(\hat{\rho} \geq \hat{\rho}_{obs} | H_o)$. Se tiene que, el p-valor asociado al coeficiente de correlación muestral.

$$\mathbb{P}(\hat{\rho} \geq \hat{\rho}_{obs} | H_o) = 1 - \mathbb{P}(\hat{\rho} < \hat{\rho}_{obs} | H_o) = F_n(\hat{\rho}_{obs}) = 0.008$$

Siendo $\hat{\rho}_{obs}$ el coeficiente de correlación muestral asociado a la muestra de la variable bidimensional $X= (Y, Z)$ y F_n la función de distribución empírica del coeficiente de correlación muestral.

Nótese que el resultado se ha obtenido por simulación, es decir, depende de la muestra y de las remuestras generadas.

En este caso, se obtiene que se rechaza la hipótesis nula, es decir, la variables no son independientes.

3.- En el primer caso, obtén un intervalo de confianza Bootstrap para ρ , considerando $\alpha=0.05$ (intervalo de nivel $(1-\alpha)$ y evalúa su cobertura real del siguiente modo.

a) Calcula el IC por Bootstrap para ρ , con $B=500$ y $\alpha=0.05$.

Se obtiene el coeficiente de correlación muestral para cada una de las $B=500$ remuestras de tamaño $n=20$, siguiendo el mismo procedimiento que en el **Caso 1**.

A continuación, se calcula el IC por Bootstrap para ρ .

Este intervalo se obtiene mediante la expresión:

$$\left(T(X) - v^* \left(1 - \frac{\alpha}{2} \right), T(X) - v^* \left(\frac{\alpha}{2} \right) \right)$$

En esta expresión, $T(X)$ es el coeficiente de correlación muestral observado, $\hat{\rho}_{obs}$, y $v^* \left(1 - \frac{\alpha}{2} \right)$, $v^* \left(\frac{\alpha}{2} \right)$ son los cuantiles de $\hat{G}(x) = Dist_T^*(x)$ que son obtenidos mediante el siguiente algoritmo.

- En primer lugar, consideremos los estadísticos ordenados, $T(X^{*(b)})$, $b=1, \dots, B$

$$T_{(1)}^* \leq \dots \leq T_{(B)}^*$$

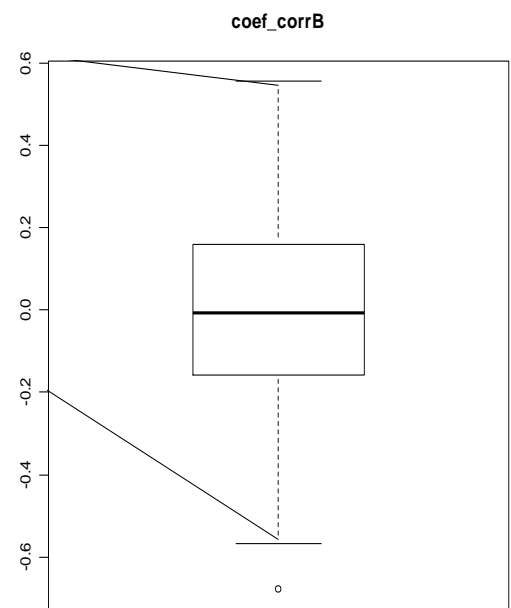
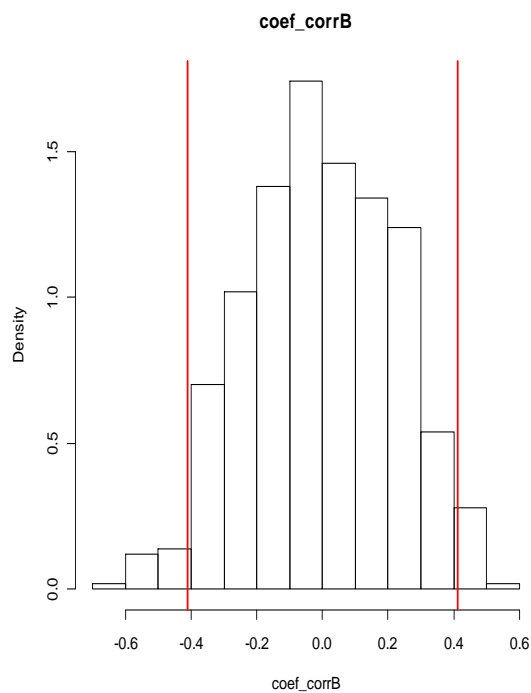
- Sean $k1 = [B \frac{\alpha}{2}]$ y $k2 = [B(1 - \frac{\alpha}{2})]$, donde $[\cdot]$ denota la parte entera.

- Finalmente,

$$v^* \left(\frac{\alpha}{2} \right) = T_{(k1)}^*, \quad v^* \left(1 - \frac{\alpha}{2} \right) = T_{(k2)}^*$$

El intervalo de confianza determinado por Bootstrap para ρ es:
-0.5099937 0.3130879

A continuación, se representa el histograma relativo a la estimación del coeficiente de correlación mediante Bootstrap junto con los cuantiles $v^*\left(\frac{\alpha}{2}\right)=T_{(k1)}^*$, $v^*\left(1-\frac{\alpha}{2}\right)=T_{(k2)}^*$ de $\hat{G}(x) = Dist_T^*(x)$. Además, se representa mediante un box-plot (diagrama de cajas) donde se concentran los valores de la estimación del coeficiente de correlación mediante Bootstrap, que como podemos observar están entorno a cero.



- b) Repita el procedimiento anterior $M=500$ veces y calcule el porcentaje de veces que ρ_0 (valor verdadero de la correlación) cae en el IC.

Con el fin de obtener M intervalos de confianza Bootstrap, se repite M veces con el procedimiento anterior y se guarda cada uno de los intervalos de confianza construidos en cada una de las filas de una matriz; por tanto, se obtiene una matriz de M filas y 2 columnas.

Para obtener el porcentaje de veces que ρ_0 cae dentro del intervalo de confianza seguimos el siguiente algoritmo:

- En primer lugar se cuentan cuantos de los intervalos de confianza tienen el extremo inferior menor que cero y el extremo superior mayor que cero; es decir, se cuentan cuantos intervalos contienen al cero que es el verdadero valor del coeficiente de correlación (ρ_0). En R esto se realiza mediante la siguiente sentencia:

```
coverage<-sum((ICM[,1]<=0)*(ICM[,2]>=0))
```

- A continuación, se calcula el porcentaje de intervalos en los que ρ_0 cae dentro del intervalo de confianza de la siguiente forma:

```
porcentaje<-coverage/M
```

De esta manera, se obtiene que ρ_0 cae dentro del 94.4% de los intervalos de confianza.

