Maestría en Computo Estadístico
Inferencia Estadística
Tarea 9
13 de diciembre de 2020
Enrique Santibáñez Cortés
Repositorio de Git: Tarea 9, IE.

Ejercicio de Bondad de ajuste.

A una muestra de 50 personas se les entrevista sobre su ingreso mensual y se desea encontrar la distribución que siguen estos datos. Use la prueba de Lilliefors, Anderson-Darling y Shapiro-Wilf para analizar si los datos provienen de una distribución normal. Datos:

$$8475, 7784, 8587, 8491, 8086, 9110, 7788, 8819, 8004, 8617, 8581, 9099, 8538, 8656, 8236, 8641, 9120, 7924, 8762, 8708, 9477, 8470, 9269, 8922, 9085, 7020, 9035, 8501, 8111, 7906, 8979, 8012, 7906, 8851, 7989, 8147, 9595, 8581, 8513, 8234, 8467, 8501, 8470, 8028, 7551, 8990, 9765, 9623, 8424, 8605.$$

RESPUESTA

Definición: 1 ***Lilliefors para Normalidad (ambos parámetros desconocidos)***

1. *Dada* $X_1, \cdots, X_n$ *de* $F_X(x)$ *estimar los parámetros* μ *y* σ^2*,*

$$\hat{\mu} = \bar{x}, \quad y \quad \hat{\sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

2. *Transforma la muestra* $X_1, \cdots, X_n$ *por medio de la siguiente relación:*

$$z_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

3. *Con la muestra* $z_1, \cdots z_n$ *llevamos acabo la prueba de Kolmogorov−Smirov para contrastar la hipótesis*

$$H_0 : F_Z(z) = N(0,1) \quad vs \quad H_a : F_Z(z) \neq N(0,1).$$

   Con estadístico de prueba

$$D_n = \sup_{z \in R} |F_n(z) - F_Z^*(z)|, \quad F_Z^*(z) = \phi(z)$$

4. *La desición de rechazo es, rechazar* H_0 *si* D_n *o rechazar* H_0 *con un nivel de significancia* α *si*

$$D_n > \omega_{1-\alpha}.$$

Definición: 2 ***La metodología para la prueba de Anderson−Darling es***

1. *Plantear el juego de hipótesis, en este caso plantearemos que la hipótesis nula es la función acumulada es igual a la función acumulada de una normal*

$$H_0 : F_X(x) = F_X * (x) \quad vs \quad H_a : F_X(x) \neq F_x^*(x)$$

2. *Dada $x_1, \cdots, x_n$ m.a. de $F_X(x)$ se transforma la muestra mediante el cambio $u_i = F_X^*(x_i)$, de tal forma que se obtiene la muestra $u_1, \cdots, u_n$.*

3. *Ordenar la muestra de menor a mayor obteniendo la muestra ordenada*

$$u_{(1)}, \cdots, u_{(n)}.$$

4. *Calcular el estadístico de prueba:*

$$\begin{aligned} A_n^2 &= -n - \sum_{i=1}^{n} \left(\frac{2_i - 1}{n} \right) \log(u_{(i)}) \\ &- \sum_{i=1}^{n} \left(\frac{2_i - i}{n} \right) \log(1 - u_{(n-i+1)}) \end{aligned}$$

5. *Rechazar H_0 si $A_n^2 > \omega_{1-\alpha}$ donde $\omega_{1-\alpha}$ es el cuantil asociado a la distribución de A_n^2 bajo H_0 la cual puede simularse por medio de un programa en R.*

   Lo anterior es considerando que todos los parámetros de la distribución. Si no se conocen ningún párametro entonces el estadístico de prueba se transforma a

$$\left(1 + \frac{4}{n} - \frac{25}{n^2} \right) A_n^2$$

.

Definición: 3 ***Metodología de Shapiro– Wilk*** *(solo para contrastar normalidad). El estadístico de prueba es*

$$W = \frac{\left(\sum_{i=1}^{n} a_i (x_{(n-i+1)} - x_{(i)}) \right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

donde $x_{(i)}$ es el número que ocupa la $i-$ésima posición en la muestra, a_i se obtiene de

$$\begin{pmatrix} a_i & \cdots & a_n \end{pmatrix} = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}},$$

con $m = \begin{pmatrix} m_1 & \cdots & m_n \end{pmatrix}^T$ siendo m_i el valor medio del estadístico ordenado de variables aleatorias iid y V es la matriz de covarianzas de ese estadístico de orden. La decisión de rechazo es, rechazar la hipótesis nula (la datos se distribuyen normal) si W es demasiado pequeño (el valor W puede oscilar entre 0 y 1).

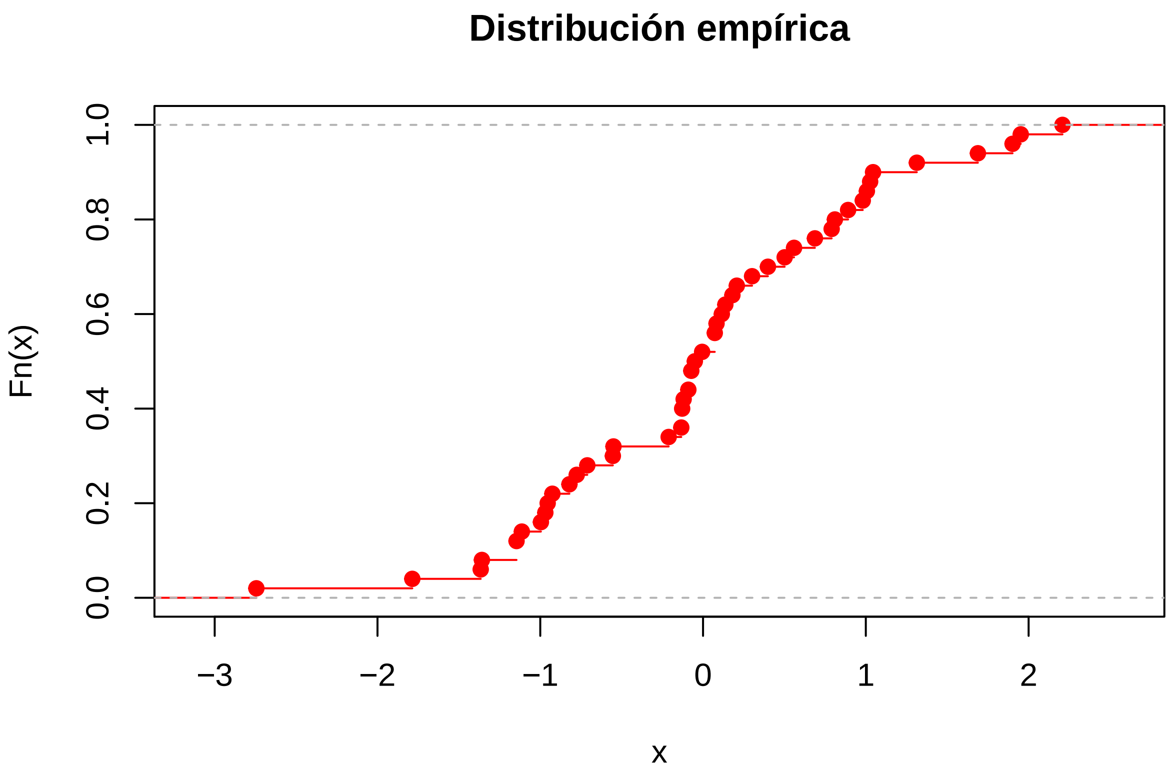
Tenemos la muestra de 50 personas, la cuál desconocemos la media real y la desviación real ocupamos la prueba de Lilliefors para normalidad, procedemos a programar con ayudad de R la metodología de la definición (1).

```r
datos <- c(8475, 7784, 8587, 8491, 8086, 9110, 7788, 8819, 8004, 8617, 8581, 9099, 8538,
           8656, 8236, 8641, 9120, 7924, 8762, 8708, 9477, 8470, 9269, 8922, 9085, 7020,
           9035, 8501, 8111, 7906, 8979, 8012, 7906, 8851, 7989, 8147, 9595, 8581, 8513,
           8234, 8467, 8501, 8470, 8028, 7551, 8990, 9765, 9623, 8424, 8605) # datos del
            # problema

mu_hat <- mean(datos) # estimación de parámetros
var_hat <- var(datos)

z <- (datos-mu_hat)/sqrt(var_hat) # normalizamos los datos

F_z_hat <- ecdf(z) # distribución empírica
plot(F_z_hat, main="Distribución empírica", col="red")
```




```r
F_z <- pnorm(z,0,1) # distribución teórica

D <- abs(F_z_hat(z)-F_z) # estadístico
D_n <- max(D)

D_n
```

```
## [1] 0.0868547
```

Entonces tenemos que el estadístico de prueba es $D = 0{,}0868547$, con un nivel de significancia $\alpha = 0{,}05$ tenemos que $D_{50,0,05} = 0{,}1246$ (consultado en tablas). Entonces como $D < 0{,}1246$ no hay evidencia significativa para rechazar la hipótesis nula, **por lo que podemos concluir (con 95 % de confianza) que los datos si se distribuyen como una normal.**

Ahora ocupemos la prueba de Anderson−Darling para determinar si los datos se distribuyen como una normal con $\mu = \hat{\mu}$ y $\sigma = \hat{\sigma}$. Para ello ocupemos R para programar la metodología descrita en la definición (2).

```
n <- length(datos) # tamaño de la muestra
u <- pnorm(datos, mu_hat, sqrt(var_hat)) # transformamos la muestra

u_ordenado <- sort(u) # ordenamos u.

s<-0 # calculamos las sumas parciales
for (j in 1:n) {
s<-s+(2*j-1)/n*(log(u_ordenado[j])+log(1-u_ordenado[n-j+1]))
}
A.t<- -n-s # calculamos A
A.t
```

```
## [1] 0.3496661
```

```
A.t_abj = (1+4/n-25/n**2)*A.t
```

Entonces tenemos que el estadístico de prueba cuando se conoces los parametros es $A_n^2 = 0{,}3496661$, pero como en este caso estamos suponiendo que $\mu = \hat{\mu}$ y $\sigma = \hat{\sigma}$, entonces ocupamos el estadístico

$$\left(1 + \frac{4}{n} - \frac{25}{n^2}\right) A_n^2 = 0{,}3741428$$

Entonces, considerando un nivel de significancia de $\alpha = 0{,}05$ tenemos que el cualtin asociado a la distribución de A_n^2 es $\omega_{1-\alpha} = 0{,}751$. Y por lo tanto, **como** $\left(1 + \frac{4}{n} - \frac{25}{n^2}\right) A_n^2 < 0{,}751$ **no existe evidencia significativa para rechazar la hipótesis nula, por lo que podemos concluir (con 95 % de confianza) que los datos si se distribuyen como una normal.**

Por último, considermalos la prueba de Sharpiro−Wilk para probar normalidad. Programamos la metodología descrita en la definición 3, los valores de a se obtuvieron de una tabla para esta prueba.

```
datos_ordenados <- sort(datos) # ordenamos los datos

a <- c(0.3751,0.2574,0.2260,0.2032,0.1847,0.1691,0.1554,0.1430,0.1317,0.1212,0.1113,
       0.1020,0.0932,0.0846,0.0764,0.0685,0.0608,0.0532,0.0459,0.0386,0.0314,0.0244,
       0.0174,0.0104,0.0035) # obtenido de tablas

p_p<-numeric() # inicializamos variables auxiliares.
a_n<-0
D<-0
for(i in 1:(n/2)) { # calculamos el numerado del estadístico de prueba
  p_p[i]<-(datos_ordenados[n+1-i]-datos_ordenados[i])
  a_n<-(a[i]*p_p[i])+a_n
  D<-(datos_ordenados[i]-mu_hat)^2
}

for(i in 1: n) {# calculamos el denominador del estadístico de prueba
  D<-((datos_ordenados[i]-mu_hat)^2)+D
}
```

```r
w_x<-((a_n)^2)/D # calculamos el estadístico de prueba
w_x
```

```
## [1] 0.9822247
```

Entonces, tenemos que el estadístico de prueba es $W = 0{,}983$, y el valor crítico con un $\alpha = 0{,}05$ es 0.947 (obtenido de tablas). Por lo tanto, **podemos concluir que como $W > 0{,}947$ entonces no existe evidencia significativa para rechazar la hipótesis nula, es decir con 95 % de confianza podemos decir que la muestra si provienen de una distribución normal.**

Observamos que las tres pruebas fueron consistentes en las concluciones, por lo que podemos concluir que efectivamente (con 95 % de confianza) que los datos se distribuyen como una normal. $\blacksquare$.