Universidad EAFIT
Escuela de ciencias aplicadas e ingeniería
Estadística II
Tareas
Semestre 2023-1

Nombre: Julieth Stefanny Escobar Ramírez y Sara Gallego Villada

1 Pregunta 1

- Mostrar que $-1 \leq \rho \leq 1$
- Relación entre correlación muestral y coseno del ángulo entre 2 vectores

Para saber que $-1 \leq \rho \leq 1$ mostraremos la relación entre el coseno del ángulo entre 2 vectores y la correlación muestral, dado que lo podemos derivar de que el $-1 \leq cos(\theta) \leq 1$, por lo cual es inmmediato que $-1 \leq \rho \leq 1$. El coeficiente de correlación de Pearson se puede ver como una variante del producto interno.
El producto interno se puede definir como sigue:

$$Inner(x, y) = \sum_i x_i y_i = \langle x, y \rangle$$

Como se puede observar, si x tiende a ser grande cuando y también es grande o pequeño cuando y es pequeño, el producto interno es mayor, es decir los vectores son más parecidos. Dado a que el producto interno no esta acotado entre -1 y 1 unadorma es dividirlo por la norma de los dos vectores, lo cual deja como resultado el coseno entre el angulo de los dos vectores, es decir:

$$cos(\theta) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} = \frac{\langle x, y \rangle}{||x||||y||}$$

Ahora bien, el coseno entre dos vectores no es invariante ante cambios de ubicación, mientras la correlación si, esto se debe a que la correralación es es el coseno entre x y y centrados. Aunque el coseno se piensa en términos de ángulos, podriamos compararlo con la correlación pensando los vectores como muestras emparejadas.

Lo anterior lo podemos ver así,
Sea $\mathbf{u} = (u_1, u_2, ...u_n)$ y $\mathbf{v} = (v_1, v_2, ..., v_n)$ donde los elementos son deviaciones de medias, es decir,

$$u_i = x_i - \bar{x} \text{ y } v_i = y_i - \bar{y}$$

Ahora podemos escirbir la covarianza y las varianzas muestrales:

$$r_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n} u_i v_i = \frac{1}{n-1}\mathbf{u} \cdot \mathbf{v}$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n} u_i^2 = \frac{1}{n-1}||\mathbf{u}||^2$$

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n} v_i^2 = \frac{1}{n-1}||\mathbf{v}||^2$$

Ahora como vimos anteriormente, por la ley de cosenos ($\mathbf{u} \cdot \mathbf{v} = ||u||||v||cos(\theta)$), el coeficiente de correlación seria:

$$\rho_{X,Y} = \frac{r_{X,Y}}{s_X s_Y} = \frac{\mathbf{u} \cdot \mathbf{v}}{||\mathbf{u}||\ ||\mathbf{v}||} = \cos\theta.$$

Finalmente, dado que $-1 \leq cos(\theta) \leq 1$ podemos decir que $-1 \leq \rho \leq 1$
Referencia 1 y Referencia 2

2 Test de normalidad:

2.1 Test de Shapiro-Wilk:

La prueba de Shapiro-Wilk se usa para contrastar la normalidad de un conjunto de datos. Se plantea como hipótesis nula que una muestra $x_1, ..., x_n$ proviene de una población normalmente distribuida.

El estadístico de prueba es:

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

La hipotesis nula se rechaza si W es demasiado pequeño. El valor de W puede oscilar entre 0 y 1.

Interpretación:
Siendo la hipótesis nula que la población está distribuida normalmente, si el p-valor es menor al nivel de significancia entonces la hipótesis nula es rechazada, por lo tanto se concluye que los datos no vienen de una distribución normal. Si el p-valor es mayor al nivel de significancia, se concluye que no se puede rechazar dicha hipótesis.
La normalidad se verifica confrontando dos estimadores alternativos de la varianza σ^2:
-Un estimador no paramétrico al numerador, y Un estimador paramétrico (varianza muestral), al denominador.

Aplicación:
La prueba de normalidad de Shapiro-Wilk es aplicable cuando se analizan muestras compuestas por menos de 50 elementos (muestras pequeñas).

Supuestos:

-Una muestra menor a 50 datos.
-Observaciones independiente.
-Muestreo aleatorio.
-Variables en escala intervalar o razón.

Univariante:

Consideremos los datos estandarizados

$$Z_i = \frac{X_i - \bar{X}}{S}$$

El estadístico de Shapiro-Wilk se construye de la siguiente manera:

$$W = \sum_{i=1}^{n} a_{i,n}(Z_{(n-i+1):n} - Z_{i:n})$$

siendo $Z_{1:n} < ... < Z_{n:n}$ la muestra ordenada de los datos estandarizados y $a_{i,n}$ ciertas constantes. Consiste en calcular las distancias entre los datos de la muestra ordenada, simétricos respecto de la mediana, esto es, la distancia entre el primero y el último, el segundoy el penúltimo, y así sucesivamente; en general el $Z_{i:n}$ y el $Z_{(ni+1):n}$. El propósito es comparar estas distancias con las que habría en una muestra de observaciones normales.

Multivariante:
Si $W_1, ..., W_d$ son los estadísticos de Shapiro-Wilk de cada componente de $Z_1, ..., Z_n$, entonces podemos considerar el estadístico de Shapiro-Wilk multivariante:

$$W = \frac{1}{d}\sum_{j=1}^{d} W_j$$

Uso en R:
Las hipótesis estadísticas son las siguientes:

-H_0: La variable presenta una distribución normal
-H_1: La variable presenta una distribución no normal

Toma de decisión:

Sig(p valor) > alfa: No rechazar H_0 (normal).
Sig(p valor) < alfa: Rechazar H_0 (no normal).

Para hacer los test de normalidad solo tomamos los datos del 2023.
Primero hicimos el test de shapiro-Wilk univariado, para este test tomamos de nuestra base de datos la columna de $rentabilidad - mensual$, esta columna contiene valores que representan la rentabilidad mensual de los Fondos de Inversión Colectiva (FIC).
El resultado del test univariado utilizando la libreria mvnormtest fue:

```
      Shapiro-Wilk normality test

data:  datos_aleatorios[, 14]
W = 0.058168, p-value < 2.2e-16
```

Por el p-value obtenido podemos decir que los datos no tienen un comportamiento de normalidad.

Ahora para test de Shapiro-Wilk multivariado usamos 2 columnas de nuestra base de datos, las cuales fueron: $rentabilidad-mensual$, $rentabilidad-semestral$ y $rentabilidad-anual$, estas columnas contienen valores que representan la rentabilidad mensual, semestral y anual respectivamente.
El resultado del test multivariado fue:

```
      Shapiro-Wilk normality test

data:  Z
W = 0.077127, p-value < 2.2e-16
```

2.2 Test Jarque-Bera

El test de Jarque-Bera es ua prueba que tiene propiedades óptimas de potencia asintótica, como una distribución chi-cuadrado, con dos grados de libertad, además puede usarse para pobar la hipótesis nula que se describirá más adelante. Esta prueda de bondad de ajuste es para comprobar si una muetra de datos tiene asimetria y la curtosis de una distribución normal. Básicamente es la suma de cuadrados de dos normales estandarizadas asintóticamente independientes
Matemáticamente se define de la siguiente forma:

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right)$$

donde n es el número de observaciones en la muestra, S es la asimetría de la muestra y K es la curtosis de la muestra. En un contexto univariado definimos S como:

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^{3/2}}$$

Y K lo definimos como:

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2}$$

Recordando que $\hat{\mu}_3$ es la tercera estimación de momentos central y $\hat{\mu}_4$ son la cuarta, además $\bar{x}$ es la media muestral y $\hat{\sigma}^2$ la estimación del segundo momento, es decir la varianza.

Ahora, definamos el test para p variables:
Sea $x = (x_1, x_2, ..., x_p)$ y $y = (y_1, y_2, ..., y_p)$ vectores aleatorios independientes e identicamente distribuidos con vector de medias $\mu = (\mu_1, \mu_2, ..., \mu_3)$ y matriz de covarianza Σ con $\Sigma > 0$.
Sea $\beta_{M,1}$ la asimetria, y $\beta_{M,2}$ la curtosis. Las cuales definiremos :

$$\beta_{M,1} = E[\{(x-\mu)'\Sigma^{-1}(y-\mu)\}^3]$$

$$\beta_{M,2} = E[\{(x-\mu)'\Sigma^{-1}(x-\mu)\}^2]$$

donde por la forma de en que se distribuyen:

$$\beta_{M,1} = 0 \text{ y } \beta_{M,2} = p(p+2)$$

Ahora de forma muestral: Sea $x_1.x_2, ..., x_N$ muestras de tamaño N de una población p-dimensional, por lo que el vector de medias y covarianza muestral son:

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j$$

$$S = \frac{1}{N}\sum_{j=1}^{N}(x_j - \bar{x})((x_j - \bar{x})'$$

respectivamente. con lo anterior podemos definir la asimetria y curtosis muestral:

$$b_{M,1} = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\{(x_i - \bar{x})'S^{-1}(x_j - \bar{x})\}^3$$

$$b_{M,2} = \frac{1}{N}\sum_{i=1}^{N}\{(x_i - \bar{x})'S^{-1}(x_i - \bar{x})\}^2$$

Ahora, considerando una población $N_p(\mu, \Sigma)$ podemos pantear el test multivariado que va a ser:

$$MJB_M = N\left\{\frac{b_{M,1}}{6} + \frac{b_{M,2} - p(p+2)^2}{8p(p+2)}\right\} \tag{1}$$

Donde N es la muestra aleatorua de $N_p(\mu, \Sigma)$, $\Sigma > 0$ y p las variables
MJB_M es asíntoticamente distribuida como χ^2_{f+1}.
Con $f = p(p+1)(p+2)/6$ La información fue obtenida de la siguiente Referencia

La hipótesis nula para es te test es de que los datos pertenecen a una distribución normal.
Esta es una hipótesis conjunta de que la asimetría y el exceso de curtosis son nulos (asimetría = 0 y curtosis = 3). Referencia

Ahora, las hipótesis nulas son las siguientes:

-H_0: La variable presenta una distribución normal multivariada/univariada -H_1: La variable presenta una distribución no normal multivariada/univariada

Es estadístico de prueba fue el prentado en (1) Toma de decisión:

$p - valor > \alpha$: No rechazar H_0
(aceptamos noralidad).

$p - valor < \alpha$: Rechazar H_0
(rechazamos que sea normal).

En R

De nuestra base de datos tomamos la columna de $rentabilidad - mensual$ para hacerle el test, esta columna se compone de valores que representan la rentabilidad mensual de los Fondos de Inversión Colectiva (FIC), en este caso hay diversos fondos y por fechas. Los resultados de este test univariado, utlizando la libreria t-series fue:

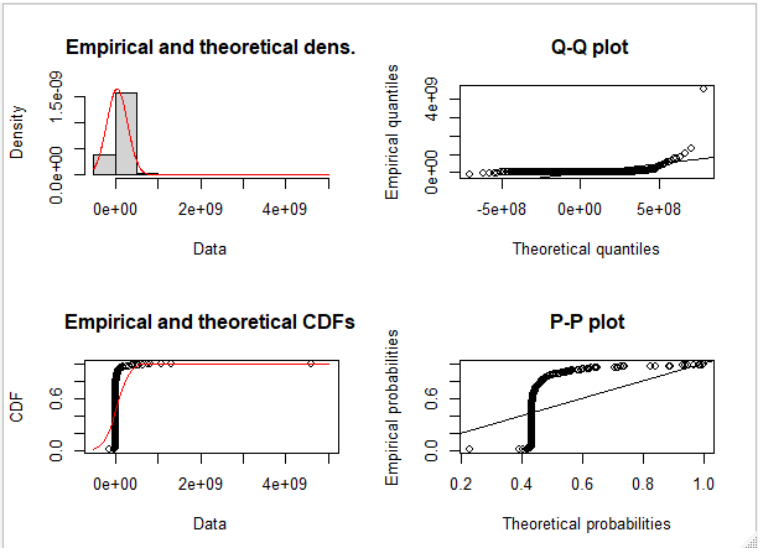
```
      Jarque Bera Test

data:  datos$rentabilidad_mensual
X-squared = 134681, df = 2, p-value < 2.2e-16
```

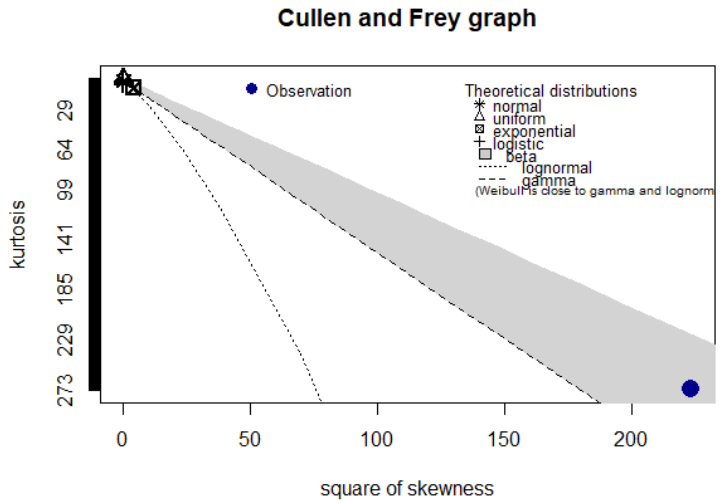
Ahora implementamos del test en unos datos que si son normales para verificar que :

```
     Jarque Bera Test

data:  data
X-squared = 3.409, df = 2, p-value = 0.1819
```



Resultados por fitdist

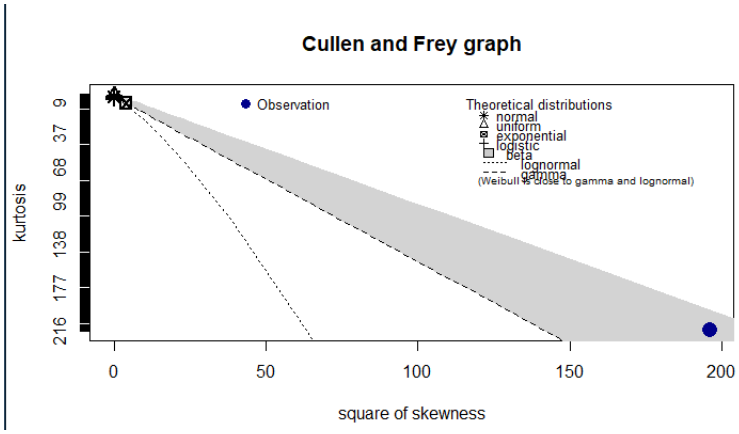


Resultado cullen and Frey

En el resultado de Cullen and Frey se compara la Curtosis y la asimetría nos demuestra que los datos no son normales.
Ahora en multivariado:
Tomamos 2 variables de nuestra base de datos(número inversionistas y rentabilidad diaria), finalmente concluímos que no es normal. Pues:

```
JB = 10648343
vp = 0
```

Nos damos Cuenta que el p-valor es tan pequeño que R lo aproxima a 0, en consiguiente y con nuestro estadístico JB nos damos cuenta que nuestros datos no siguen una distribución normal.



Gráfica Número de inversionistas

Para comprobar gráficamente el resultado de nuestra hipótesis gráficamos uno de los valores y vemos nuevamente que se aleja del supuesto de normalidad.

Para ver los resultados se ajunta el acceso directo a carpeta con códigos y datos

Los datos para esta entrega fueron tomados de: datos abiertos