Maestría en Ciencia de Datos

Introducción a la Estadística Usando Software

Curso 2023

Examen

30/05/2023

Diego Velasco

Ejercicio 1

Se considera la muestra de 25 datos presentada a continuación:

26.7964	17.4528	26.1574	38.6935	28.0100
36.2080	20.3004	34.5870	37.5912	23.8766
28.1384	48.5347	8.4779	21.7938	28.2164
37.5639	31.1565	31.3047	17.7121	20.0819
28.7006	1.4869	1.7976	25.8220	26.0392

Tabla 1. Datos de entrada - Ejercicio 1

Parte a)

En primer lugar, se realiza la carga del paquete de estadística y los datos a utilizar. El código implementado en Octave es mostrado a continuación:

```
clc
close all
clear all
pkg load statistics

# Generación de datos
datos = [26.7964 17.4528 26.1574 38.6935 28.0100 36.2080 20.3004 34.5870 37.5912
23.8766 28.1384 48.5347 8.4779 21.7938 28.2164 37.5639 31.1565 31.3047 17.7121
20.0819 28.7006 1.4869 1.7976 25.8220 26.0392];
```

Se presume que la distribución de los datos es $N(\mu, \sigma^2 = 100)$, y se desea realizar el siguiente test unilateral:

$$\begin{cases} H_0: \mu = 23 \\ H_1: \mu > 23 \end{cases}$$

Para poder definir la región crítica, se utiliza el resultado de que, siendo $X_1, X_2, \dots, X_n$ una muestra de variables aleatorias iid con distribución normal $N(\mu, \sigma)$, el promedio de dicha muestra tiene la misma media μ y varianza igual a $\frac{\sigma^2}{n}$, de manera que el estadístico $d = \sqrt{n}\left(\frac{\overline{X_n} - \mu}{\sigma_n}\right)$ tendrá distribución normal estándar.

De esta manera, se considera $d = \sqrt{n}\left(\frac{\overline{X_n} - \mu}{\sigma_n}\right)$ como el estadístico del test de hipótesis a implementar. Considerando H_0 como válida, se tiene que $\mu_0 = 23$, lo que permite calcular el estadístico para la muestra en cuestión. Esto es implementado en Octave mediante el siguiente código:

```
n = length(datos);
mu0 = 23; # Valor de mu bajo Ho
medias = mean(datos,2);
sigma = 10; # Valor de sigma conocido

# Test de hipótesis para mu
```

```
# Ho: mu = 23
# H1: mu > 23

# TEST UNILATERAL
d = sqrt(n)*((medias-mu0)/sigma); # Estadístico del test
```

Para el caso en cuestión, se tiene que $n = 25$, $\overline{X_n} = 25.860$ y $\sigma = 10$, lo que arroja un valor del estadístico de $d = 1.43$.

Luego, dado que este estadístico debería seguir una distribución normal estándar, es posible calcular la probabilidad de que dicho estadístico tome un valor tanto o más extremo que el obtenido para la muestra en cuestión (p-valor) y así contrastar dicha probabilidad con el nivel de significancia deseado, el cual se toma como 0.10 y 0.05. Siendo F_Φ la función de distribución normal estándar, dicha probabilidad es calculada como $P(Z_n > d) = 1 - F_\Phi(d)$. Esto es implementado en Octave mediante el siguiente código:

```
# Right-tailed (H1: mu > mu0)
alfa1 = 0.10;
alfa2 = 0.05;
p_value = (1-normcdf(d))

# Resultado con alfa = 0.10
if p_value < alfa1
    sprintf('TEST UNILATERAL DERECHO: p-valor = %.4f. Ho es rechazada con alfa =
%d', p_value, alfa1)
else
    sprintf('TEST UNILATERAL DERECHO: p-valor = %.4f. Ho no puede ser rechazada
con alfa = %d', p_value, alfa1)
end

# Resultado con alfa = 0.05
if p_value < alfa2
    sprintf('TEST UNILATERAL DERECHO: p-valor = %.4f. Ho es rechazada con alfa =
%d', p_value, alfa2)
else
    sprintf('TEST UNILATERAL DERECHO: p-valor = %.4f. Ho no puede ser rechazada
con alfa = %d', p_value, alfa2)
end
```

El p-valor obtenido para el caso en cuestión es de 0.076359. Dado que dicho valor es menor a 0.10, se deduce que se tiene suficiente evidencia para poder rechazar H_0 trabajando al 10%. Sin embargo, dado que dicho valor es mayor a 0.05, se deduce que no se tiene suficiente evidencia para poder rechazar H_0 trabajando al 5%.

α	p-valor	Conclusión
0.05	0.076359	No se rechaza H_0
0.10	0.076359	Se rechaza H_0

Tabla 2. Resultados de test de Hipótesis - Ejercicio 1

Parte b)

En esta parte se realiza el mismo test que en la parte anterior, pero por simulación. Para ello, se simulan 1.000.000 muestras de 25 datos cada una, distribuidas de manera normal con parámetros μ = 23 y σ = 10. Luego, para cada muestra se calcula el promedio de los datos, y se determina la cantidad de muestras para las cuales el promedio es mayor o igual al obtenido de la muestra original. El cociente entre esta cantidad y la cantidad total de muestras dará la proporción de muestras cuyo estimador toma un valor tanto o más extremo que el estimado a partir de la muestra original, por lo que se espera que se aproxime al p-valor calculado en la parte anterior. Esto es implementado mediante el siguiente código:

```
# Por simulacion
cant_muestras = 1000000;
muestras = normrnd(mu0, sigma, cant_muestras, n);

media_muestras = mean(muestras, 2);

pseudo_p_val = size(find(media_muestras >= medias ))(1)/cant_muestras

# Resultado con alfa = 0.10
if pseudo_p_val < alfa1
    sprintf('TEST SIMULADO: pseudo p-valor = %.4f. Ho es rechazada con alfa =
%d', pseudo_p_val, alfa1)
else
    sprintf('TEST SIMULADO: pseudo p-valor = %.4f. Ho no puede ser rechazada con
alfa = %d', pseudo_p_val, alfa1)
end

# Resultado con alfa = 0.05
if pseudo_p_val < alfa2
    sprintf('TEST SIMULADO: pseudo p-valor = %.4f. Ho es rechazada con alfa =
%d', pseudo_p_val, alfa2)
else
    sprintf('TEST SIMULADO: pseudo p-valor = %.4f. Ho no puede ser rechazada con
alfa = %d', pseudo_p_val, alfa2)
end
```

Para el caso en cuestión, la cantidad de muestras cuyo estimador cae por fuera del intervalo es de 76118, lo que da una proporción de 0.076118, valor muy cercano al p-valor de 0.076359 obtenido en la parte anterior. Dado que los p-valores hallados son muy similares y que se cumplen las mismas relaciones con los valores de α considerados, las conclusiones son las mismas que en la parte anterior.

Ejercicio 2

Se considera la muestra de 100 datos contenida en el archivo muestra.mat.

Parte a)

Se plantea la realización de las pruebas de aleatoriedad de Rachas y de Spearman de manera de ver si es razonables afirmar que los datos son independientes e idénticamente distribuidos (i.i.d).

Test de rachas

Se considera una racha como una secuencia de observaciones en orden creciente o decreciente. Una racha creciente se termina cuando el dato en posición *i+1* es menor al dato en posición *i*, mientras que una racha decreciente se termina cuando el dato en posición *i+1* es mayor al dato en posición *i.* Tomando dos datos cualesquiera de un set de datos i.i.d, la probabilidad de que el segundo dato sea mayor o menor que el primero es de 0.5, lo que implica que es esperable que la cantidad de rachas en un set de datos iid esté razonablemente alejado tanto de la unidad (datos ordenados) como del tamaño de la muestra *n*.

Teniendo *n* realizaciones de una variable aleatoria, se tienen *n!* posibles configuraciones en las cuales los datos pueden estar ordenados. Cada posible configuración está asociada con una probabilidad de ocurrencia, la cual es igual al producto de la probabilidad de ocurrencia de cada observación si la muestra es iid. Dado que varias configuraciones distintas tendrán la misma cantidad de rachas, es posible obtener la probabilidad de ocurrencia de cada número de rachas para cada valor de *n,* pudiendo así contrastar la cantidad de rachas de una muestra cualquiera con su probabilidad de ocurrencia y así poder decidir si es razonable afirmar que los datos de dicha muestra son efectivamente iid. La tabla presentada en el curso presenta estas probabilidades en formato *P(R≤R')* y *P(R≥R')* para cada valor de *n*, siendo *R'* los posibles valores de rachas para el valor de *n.* Dichas probabilidades corresponden a los p-valores del test de rachas.

Para el caso en cuestión, la cantidad de rachas es determinada mediante la función *rachas* implementada en Octave. El código es presentado a continuación:

```
function r = rachas(datos)
    n = length(datos);
    Y = datos(2:n) > datos(1:n-1);
    rachas = (Y(2:n-1) != Y (1:n-2));
    r = sum(rachas) + 1;
```

Esto arroja un total de 72 rachas. Dado que la tabla mencionada cuenta con p-valores para *n≤25*, se utiliza el ajuste de Levene, calculando entonces el p-valor del test a partir del número de rachas *R* como se indica a continuación:

$$Si\ R < \frac{2n-1}{3} \rightarrow p = P(Z \leq z_L), con\ Z \sim N(0,1)\ y\ z_L = \frac{R + 0.5 - \frac{2n-1}{3}}{\sqrt{\frac{(16n-29)}{90}}}$$

$$Si\ R > \frac{2n-1}{3} \rightarrow p = P(Z \geq z_R), con\ Z \sim N(0,1)\ y\ z_R = \frac{R - 0.5 - \frac{2n-1}{3}}{\sqrt{\frac{(16n-29)}{90}}}$$

Esto es implementado en Octave en el archivo .m entregado. La fracción de código en la cual se realiza este ajuste es mostrada a continuación:

```
# Ajuste de Levene
zL_r = (R + 0.5 - (2*n-1)/3)/(sqrt((16*n-29)/90));
zR_r = (R - 0.5 - (2*n-1)/3)/(sqrt((16*n-29)/90));

if R < (2*n-1)/3
    p_value_rachas = normcdf(zL_r);
else
    p_value_rachas = 1 - normcdf(zR_r);
end
```

Dado que $R = 72 > \frac{2n-1}{3} = 66.33$, se utiliza el segundo renglón del ajuste planteado, lo que arroja un p-valor de 0.1081. Este valor representa la probabilidad de obtener un número de rachas igual a 72 o más extremo (entendiendo "más extremo" como "mayor" en este caso particular). Dado que dicho valor es mayor a 0.05 y 0.10, se considera que la muestra supera el test de rachas trabajando al 5% y 10% respectivamente.

Test de correlación de rangos de Spearman

Se consideran las siguientes definiciones:

- Rango: posición que ocuparía el dato X_i si se ordenan los datos de menor a mayor
- Índice: posición que ocupa el dato X_i en la configuración inicial.

El test de correlación de Spearman consiste en hallar el coeficiente de correlación entre ambos vectores, el cual corresponde a la pendiente de la recta que mejor ajusta los puntos de ambos vectores. Este coeficiente es determinado mediante la función *spearman* implementada en Octave. El código es presentado a continuación:

```
function RS = spearman(datos)
    n = length(datos);
    [a b] = sort(datos);
    [indices rangos] = sort(b);

    RS = 1 - 6*(sum((rangos-indices).^2))/(n*(n^2-1));
```

Una vez hallado dicho coeficiente, se determina el p-valor del test en base a valores tabulados y se lo contrasta con el nivel de significancia elegido, pudiendo así decidir si se tiene suficiente evidencia como para rechazar la hipótesis de aleatoriedad. En este caso, dado que el tamaño de la muestra es mayor a 30, se utiliza nuevamente un ajuste, calculando el p-valor de la siguiente manera:

$$Si\ RS < 0 \rightarrow p = P(Z \leq z_L), con\ Z \sim N(0,1)\ y\ z_L = RS\sqrt{n-1}$$

$$Si\ RS > 0 \rightarrow p = P(Z \geq z_R), con\ Z \sim N(0,1)\ y\ z_R = RS\sqrt{n-1}$$

Esto es implementado en Octave en el archivo .m entregado. La fracción de código en la cual se realiza este ajuste es mostrada a continuación:

```
# TEST DE SPEARMAN
RS = spearman(datos);

# Ajuste normal
z_sp = RS*sqrt(n-1);

if RS < 0
    p_value_sp = normcdf(z_sp)
else
    p_value_sp = 1 - normcdf(z_sp)
end
```

Dado que $RS = -0.1215 < 0$, se utiliza el primer renglón del ajuste planteado, lo que arroja un p-valor de 0.1133. Dado que dicho valor es mayor a 0.05 y 0.10, se considera que la muestra supera el test de correlación de rangos de Spearman trabajando al 5% y 10% respectivamente.

Dado que el p valor obtenido para ambos tests es mayor a los niveles de significancia considerados, se deduce que no se tiene suficiente evidencia para rechazar la hipótesis de aleatoriedad, de manera que es razonable afirmar que los datos son iid.

Parte b)

El test de Kolmogorov y Smirnov se basa en el teorema de Glivenko-Cantelli, el cual establece que, siendo $X_1, X_2..,X_n$ una sucesión de variables aleatorias iid con distribución *F* y siendo *Fn* la dunción de distribución empírica para la muestra de tamaño *n,* entonces se cumple que:

$$sup_{x \in R} |F_n(x) - F(x)| \rightarrow 0$$

con probabilidad 1.

El test consiste en suponer una distribución *F(x)* definida (en este caso, una distribución normal con μ=3 y σ=2), calcular la función de distribución empírica para la muestra en cuestión, y calcular el supremo de la diferencia entre ambas funciones. Luego, se compara dicho valor con los valores límites tabulados, calculados para el nivel de significancia deseado, y se decide si la hipótesis nula es rechazada o no. Cabe destacar que, para cada nivel de significancia α, los valores tabulados corresponden a supremos *KS'* tal que *P(KS>KS') = α.*

Para el caso en cuestión, el supremo mencionado se calcula mediante la función *kolsmi_normal,* implementada en Octave, la cual calcula el valor límite del supremo y grafica la función de distribución teórica y empírica. El código se presenta a continuación:

```
function KS = kolsmi_normal(datos, mu, sigma)
    n = length(datos);
    x = sort([datos datos]);
    Fn = sort([0 (1:n-1)/n (1:n-1)/n 1]);

    F = normcdf((x-mu)/sigma);

    KS = max(abs(Fn - F));

    #PLOTEO
    figure
    plot (x,Fn, 'b')
    hold on
    plot (x,F, 'r')
    tit = sprintf('Fn(x) y F(x) - Distribución normal');
    title(tit, 'fontsize', 35) # plot title # set limit for y axis
    set(gca, 'fontsize', 35) # set axis fontsize
    ylabel('Fn(x) y F(x)', 'fontsize', 35, 'fontweight', 'bold') # set ylabel
    xlabel('x', 'fontsize', 35, 'fontweight', 'bold') # set xlabel
    hold off
```

Luego, dado que el tamaño de la muestra es mayor a 35, se utiliza el resultado asintótico en función de *n*. La implementación en Octave de la búsqueda del supremo y la comparación contra valores tabulados es presentada a continuación:

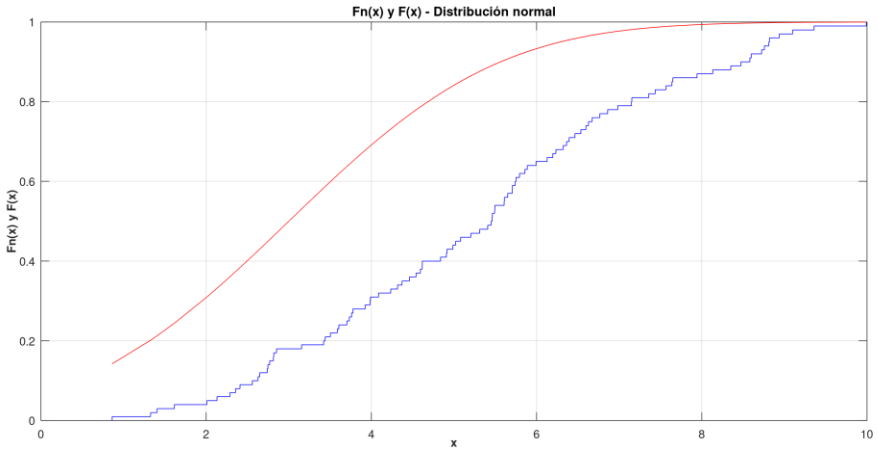
```
mu = 3;
sigma = 2;

KS = kolsmi_normal(datos, mu, sigma)

# Ajuste para n>35
if alfa == 0.20
    KS_lim = 1.07/sqrt(n)
elseif alfa == 0.15
    KS_lim = 1.14/sqrt(n)
elseif alfa == 0.10
    KS_lim = 1.22/sqrt(n)
elseif alfa == 0.05
    KS_lim = 1.36/sqrt(n)
elseif alfa == 0.01
    KS_lim = 1.63/sqrt(n)
end
```

El valor del supremo KS vale 0.4209 para el caso en cuestión. Asimismo, dado que se trabaja al 5% y al 10%, el valor límite de KS vale $\frac{1.36}{\sqrt{n}} = 0.1360$ y $\frac{1.22}{\sqrt{n}} = 0.1220$ respectivamente. Dado que el valor del supremo para la muestra en cuestión es mayor a los valores límite, se rechaza la hipótesis nula, de manera que no es razonable afirmar que los datos tienen distribución normal de parámetros μ=3 y σ=2.

α	KS	KS_{lim}	Conclusión
0.05	0.4209	0.1360	Se rechaza H_0
0.10	0.4209	0.1220	Se rechaza H_0

Tabla 3. Resumen de resultados para test de Kolmogorov y Smirnov - Ejercicio 2

Las funciones de distribución teórica y empírica calculada con los parámetros ya mencionados son mostradas a continuación.



Parte c)

La prueba de normalidad de Lilliefors utiliza el mismo estadístico que la prueba de Kolmogorov y Smirnov, en el caso en que la media y el desvío de la distribución no son conocidas y por tanto son estimados utilizando toda la muestra. Dado que estimar estos parámetros en base a la muestra tiende a sesgar los resultados hacia no rechazar la hipótesis nula, se utilizan valores límites del estadístico más exigentes para esta prueba. La implementación en Octave de esta prueba es mostrada a continuación. En ella, se estiman los parámetros μ y σ en base a la muestra, se aplica el test de Kolmogorov y Smirnov con dichos parámetros, y se compara el estadístico obtenido con los valores más exigentes obtenidos de la tabla de Lilliefors para el ajuste normal.

```
# PARTE c)
mu = mean(datos);
sigma = std(datos);

KSL = kolsmi_normal(datos, mu, sigma)

# Ajuste para n>35
if alfa == 0.20
    KSL_lim = 0.736/sqrt(n)
elseif alfa == 0.15
    KSL_lim = 0.768/sqrt(n)
```

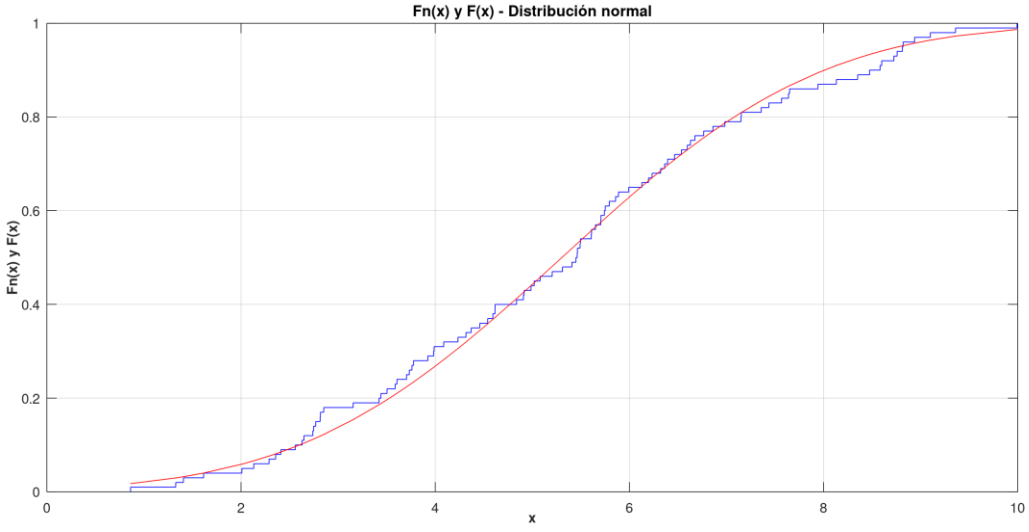
```
elseif alfa == 0.10
    KSL_lim = 0.805/sqrt(n)
elseif alfa == 0.05
    KSL_lim = 0.886/sqrt(n)
elseif alfa == 0.01
    KSL_lim = 1.031/sqrt(n)
end
```

Para el caso en cuestión, se tiene que μ = 5.3042, σ = 2.1095 y KSL = 0.057488. Luego, dado que se trabaja al 5% y 10%, los valores límite de KSL vale $\frac{0.886}{\sqrt{n}} = 0.0886$ y $\frac{0.805}{\sqrt{n}} = 0.0805$. Dado que el valor del supremo es menor al valor límite, no se tiene suficiente evidencia para rechazar la hipótesis nula, de manera que se deduce que es razonable afirmar que los datos tienen distribución normal, trabajando tanto al 5% como al 10%.

α	KSL	KSL_{lim}	Conclusión
0.05	0.057488	0.0886	No hay suficiente evidencia para rechazar H_0
0.10	0.057488	0.0805	No hay suficiente evidencia para rechazar H_0

Tabla 4. Resumen de resultados para test de normalidad de Kolmogorov y Smirnov y Lilliefors - Ejercicio 2

Las funciones de distribución teórica y empírica (calculada con los estimadores de los parámetros a partir de la muestra) son mostradas a continuación.



Parte d)

Dado que el tamaño de muestra planteado en el problema es mayor que 30, nuevamente se debe utilizar el resultado asintótico planteado en la parte anterior. Para un tamaño de muestra de 40 y un nivel de significancia del 5%, el valor límite de KSL es calculado como $\frac{0.886}{\sqrt{n}} = 0.1401$. Esto significa que, para una

muestra de 40 datos distribuidos de forma normal, la probabilidad de obtener un valor de KS mayor o igual a 0.1401 es de 0.05.

A continuación, se generan 10000 muestras con distribución normal de parámetros $\mu = 3$ y $\sigma = 2$, se calcula el valor de KS para cada una de ellas a partir de los estimadores de μ y σ utilizando la función *kolsmi_normal.m* presentada anteriormente, y se calcula la cantidad de muestras para la cual el valor de KS es mayor al valor límite de 0.1401. El código en Octave es presentado a continuación.

```
# PARTE d)
n_sim = 40
alfa_sim = 0.05;

# Ajuste para n>35
if alfa_sim == 0.20
    KSL_lim_sim = 0.736/sqrt(n_sim)
elseif alfa_sim == 0.15
    KSL_lim_sim = 0.768/sqrt(n_sim)
elseif alfa_sim == 0.10
    KSL_lim_sim = 0.805/sqrt(n_sim)
elseif alfa_sim == 0.05
    KSL_lim_sim = 0.886/sqrt(n_sim)
elseif alfa_sim == 0.01
    KSL_lim_sim = 1.031/sqrt(n_sim)
end

cant_muestras_sim = 10000;
mu_sim = 3;
sigma_sim = 2;

x = normrnd(mu_sim, sigma_sim, cant_muestras_sim, n_sim);
KSL_sim = [];

for i = 1:cant_muestras_sim
    KSL_sim(i) = kolsmi_normal(x(i,:),mean(x(i,:)), std(x(i,:)));
end

size(find(KSL_sim>=KSL_lim_sim))(2) / cant_muestras_sim
```

La cantidad de muestras para la cual el valor límite de KS es mayor o igual al valor límite de 0.1401 es de 433, lo que da una proporción de 0.043. Es decir, hay 433 muestras de las 10000 muestras simuladas para las cuales el supremo del estadístico de la prueba de Kolmogorov Smirnov Lilliefors es igual o más extremo que el valor límite, aún cuando los datos tienen distribución normal. Este valor es muy cercano a 0.05, por lo que la simulación es coherente con el resultado teórico.