Examen Parcial 1

Vega-Lagunas Donaldo F.

Departamento: Computación
Curso: Computación aplicada
Profesor: Benito Granados-Rojas, PhD.
Fecha: 15 de septiembre de 2022

Problema 2 Análisis de Datos: Realizar un análisis estadístico y de regresión con los datos obtenidos de la siguiente encuesta:

- Obtener las medidas de tendencia central (μ y σ) para ambos datasets (estatura y talla de calzado).
- Encontrar la función gaussiana que mejor ajuste a las curvas asumiendo una distribución normal.
- Graficar histogramas y funciones gaussianas.
- Encontrar la probabilidad de que una persona tomada al azar se encuentre dentro de la primera desviación estándar para cada variable.
- Devuelva la suma de todos los elementos en y por encima de la diagonal principal.
- Hallar una aproximación lineal a la relación entre estatura y talla de calzado. Tanto para la población general como para subgrupos divididos por sexo y comparar.
- Desarrollar conclusiones a partir de los resultados.

1. Introducción

1.1. Medidas de tendencia central

Las medidas de tendencia central son parámetros estadísticos que notifican sobre el centro de la distribución de la población estadística.

Permiten conocer en qué lugar se ubica el elemento promedio del grupo, comparar resultados obtenidos con relación a los valores observados e interpretar el valor de una misma variable en distintas ocasiones, así como comparar los resultados con otros grupos dentro de las métricas de tendencia central (López 2020).

- Moda: Es el valor que más se repite en un conjunto de datos.

$$m_o$$

- Media: La media es el valor promedio de un conjunto de datos numéricos. Se calcula como la suma total de datos, dividida entre el número total de valores.

$$\bar{x} = \frac{x_1 + x_2 + x_3 \ldots + x_n}{N}$$

- La mediana indica el dato en la posición central del conjunto, que parte la distribución en dos.

$$m_e = \frac{x_n + 1}{2} \qquad m_e = \frac{\frac{x_n}{2} + \frac{x_n}{2} = 1}{2}$$

1.2. Función gaussiana

La función gaussiana es una función continua que se aproxima a la distribución binomial de eventos. La distribución normal esta definida por la expresión:

$$f(x) = ae^{-\frac{(x-b)^2}{2c^2}}$$

La gráfica de la función es simétrica, con forma de campana. El parámetro a es el valor del punto más alto de la campana, b es la posición del centro de la campana y c determina el ancho de la campana (Wikipedia s/f).

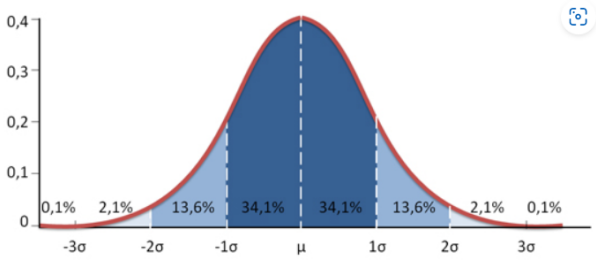

Figura 1: Campana de distribución gaussiana (TyC 2016)

Cuando el valor de la media $\mu = 0$, y desviación $\sigma = 1$, se denomina distribución normal estándar y esta definida por la expresión:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

1.3. Histograma

El histograma es una representación gráfica de un grupo de datos estadísticos agrupados en intervalos numéricos, muestra la distribución de los datos de un conjunto, respecto a una variable numérica.

Se utilizan barras cuya superficie o altura individual es proporcional a la frecuencia de los valores representados. El eje horizontal corresponde a la variable de estudio, el histograma se utiliza cuando esta última es cuantitativa (Westreicher s/f).

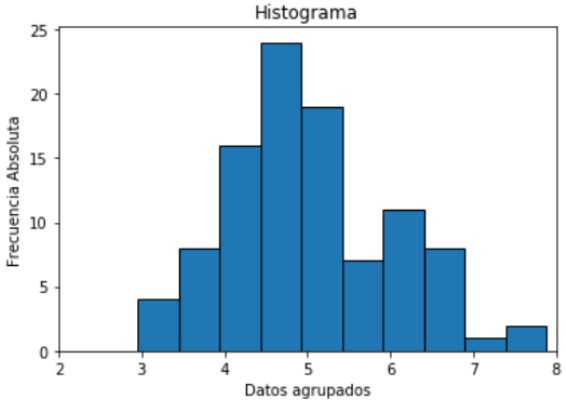

Figura 2: Histograma (Superprof s/f)

1.4. Aproximación Lineal

Una aproximación lineal es una aproximación de una función general utilizando una función lineal. Es utilizada para resolver o aproximar soluciones a ecuaciones. También se denomina linealización o desarrollo de Primer Orden de la función (HMN s/f).

$$f(x) \approx f(a) + f\,'(a)(x - a)$$

2. Procedimiento y Resultados

2.1. Análisis y extracción de datos

Se recabaron datos reales a través de un formulario en linea referente a la estatura, talla de calzado y sexo de cada participante. Los datos se exportaron a una hoja de Excel, donde se realizo la limpieza de los mismos.

Debido a que varios participantes ingresaron su estatura en mts, fue necesario realizar la conversión a cm directamente en la hoja de calculo. De igual forma, se convirtieron manualmente los datos ingresados como texto. De esta forma se consiguió un listado de datos adecuado para su procesamiento. Utilizando un Live Script de MATLAB y la instrucción $readtable$, se generó una tabla leyendo datos orientados a columnas de la hoja de cálculo de Excel. La tabla se almacenó en una variable y por medio de la instrucción $table2array$, se convirtió a un arreglo para poder utilizar los datos.

```
T=readtable('Data.xlsx')
```

```
T = 200×3 table
```

	Estatura	Talla	Sexo
1	169	27.5000	'Masculino'
2	175	26	'Masculino'
3	173	27	'Masculino'
4	173	27	'Femenino'
5	175	26	'Masculino'

Figura 3: Lectura de datos previa a importación en arreglo de MATLAB

2.2. Medidas de tendencia central para ambos datasets

Para obtener las medidas de tendencia central, se utilizó la función $mean$ que devuelve el valor medio de un arreglo y std que devuelve la desviación estándar de los elementos del arreglo. Se genera un valor de μ y σ para estatura y talla de calzado.

```
% 1. Medidas de tendencia central

mu=mean(a)
sigma=std(a)

fprintf('Valor Promedio de Estatura: %f',mu(:,1))
fprintf('Valor Promedio de Talla de Calzado: %f',mu(:,2))

fprintf('Desviacion Estandar de Estatura: %f',sigma(:,1))
fprintf('Desviacion Estandar de Talla de Calzado: %f',sigma(:,2))
```

```
mu = 1×2
    167.1150   25.7425

sigma = 1×2
     9.4980    1.9159

Valor Promedio de Estatura: 167.115000
Valor Promedio de Talla de Calzado: 25.742500

Desviacion Estandar de Estatura: 9.498044
Desviacion Estandar de Talla de Calzado: 1.915933
```

Figura 4: Calculo de medidas de tendencia central para Estatura y Talla de Calzado

2.3. Función gaussiana de ajuste a curvas asumiendo una distribución normal

Para encontrar la función gaussiana que mejor ajuste a las curvas asumiendo una distribución normal, primero se define un vector fila utilizando la función $linspace$ con el rango que se utilizará para graficar la función.

La función de densidad de probabilidad para un variable aleatoria continua se define como:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Sustituyendo los valores de μ y σ, se obtienen las funciones gaussianas para ambos datasets.

$$f_{Estatura}(x) = \frac{1}{\sqrt{2\pi}9{,}49} e^{\frac{-(x-167{,}11)^2}{2(9{,}49)^2}} \qquad f_{Talla}(x) = \frac{1}{\sqrt{2\pi}1{,}91} e^{\frac{-(x-25{,}74)^2}{2(1{,}91)^2}}$$

Posteriormente se definen ambas funciones gaussianas en MATLAB, sustituyendo los valores obtenidos de μ y σ para cada dataset, y se les asigna una variable a cada una para poder ser graficadas contra el rango definido. Cada variable devuelve un arreglo de datos que se utilizara al graficar cada función.

```
% 2. Función Gaussiana de la distribucion normal.

% Gaussiana de estatura
xe=linspace(130,200,200);
fge=1./(sqrt(2*pi)*sigma(1))*exp(-((xe-mu(1)).^2)/(2*sigma(1).^2));

% Gaussiana de talla
xt=linspace(19,32,200);
fgt=1./(sqrt(2*pi)*sigma(2))*exp(-((xt-mu(2)).^2)/(2*sigma(2).^2));
```

Figura 5: Definición de funciones gaussianas para ambos datasets con valores de μ y σ

2.4. Gráfica de histogramas y funciones gaussianas

Con ambas funciones gaussianas definidas, se grafica el vector fila contra la columna de datos de estatura y talla utilizando la función $plot$. La instrucción $hold$ mantiene graficada la función gaussiana en ambos juegos de datos para poder superponer la gráfica con el histograma y realizar un comparativo.

Los histogramas para cada conjunto de datos se generan a través de la función $histogram$, tomando como argumento la columna de datos, número de bins y un argumento de normalización para mostrar el histograma escalado.

En la Fig. 6 se muestra el código para realizar la gráfica de ambos elementos, únicamente se sustituye el argumento de la columna de datos para graficar estatura y la talla de calzado.

```
% 3. Graficar histogramas y funciones gaussiana
plot(xe,fge)
hold on
histogram(a(:,1),10,'Normalization','pdf')
hold off
title('Estatura')
grid on
```

Figura 6: Segmento para graficar las función gaussiana e histograma normalizados

Utilizando los datos correspondientes a la estatura, se observa una estructura similar entre la función gaussiana y el histograma, con la mayoría de datos agrupados dentro de la campana.

En el caso de la talla, se observa una diferencia significativa en la presentación de la función gaussiana y el histograma. La la mayor cantidad de datos se encuentra agrupada bajo la campana, pero con una estructura más conservativa.

2.5. Probabilidad dentro de la primera desviación estándar

Para el calculo de la probabilidad dentro de la primera desviación estándar, se utiliza la función de distribución acumulativa cdf evaluada en los valores de x, con el parámetro de distribución

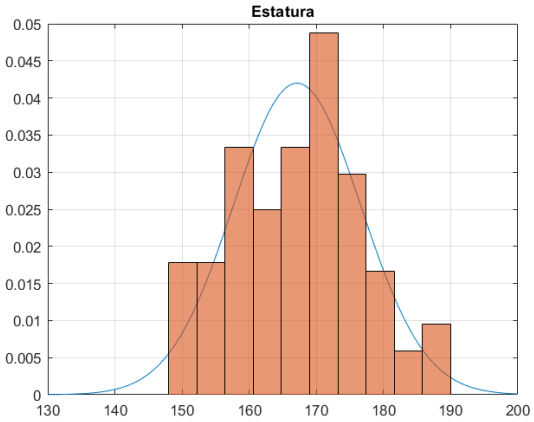

Figura 7: Función gaussiana e histograma del conjunto de datos Estatura

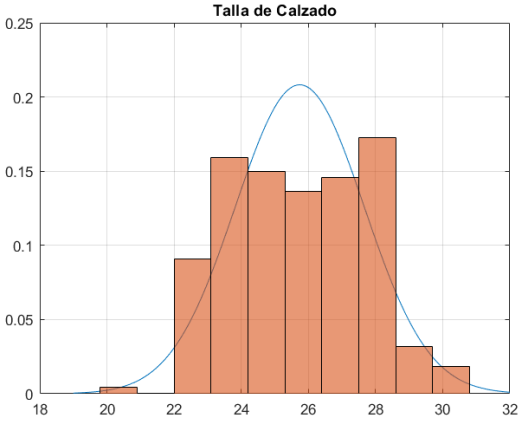

Figura 8: Función gaussiana e histograma del conjunto de datos Talla de Calzado

normal definido y los datos de media μ y desviación estándar σ como parámetros de entrada.

```
% 4. Encontrar probabilidad dentro de primera desviación estándar

fprintf('Probabilidad Estatura')
pe=200*cdf("Normal",mu(1)+sigma(1),mu(1),sigma(1))-100
fprintf('Probabilidad Talla')
pt=200*cdf("Normal",mu(2)+sigma(2),mu(2),sigma(2))-100
```

```
Probabilidad Estatura
pe = 68.2689
Probabilidad Talla
pt = 68.2689
p1 = 1×2
    0.1600   -0.9917
```

Figura 9: Probabilidad dentro de la primera desviación estándar

Se observa que el valor de la probabilidad en ambos casos es igual a 68.27 %, lo que concuerda con el valor del intervalo de la desviación estándar a ambos lados de la media, como se muestra en la Fig. 1 equivalente al 68.27 % de un conjunto.

2.6. Aproximación lineal

Posteriormente se realiza una aproximación lineal a la relación entre estatura y talla de calzado, de la población general, así como la población masculina y femenina exclusivamente.

Se utilizó la función de $polyfit$ para realizar el ajuste polinomial de curvas, que devuelve los coeficientes de un polinomio p(x) de grado n que mejor se aproximen con los datos de y. Los coeficientes de p son dados en potencias descendentes y la longitud de p es $n+1$. A esta función se añade un argumento de salida s, que es utilizado para obtener el cálculo del coeficiente de determinación R^2. Este bloque se aplica a los 3 conjuntos de datos.

```
% 5. Aproximación lineal entre estatura y talla de calzado

% Talla contra estatura [Población General]
[p1,s1]=polyfit(a(:,1),a(:,2),1)
R1=1 - (s1.normr/norm(a(:,2) - mean(a(:,2))))^2
```

```
p1 = 1×2
    0.1600   -0.9917

s1 = struct with fields:
       R: [2×2 double]
      df: 198
   normr: 16.4637
R1 = 0.6289
```

Figura 10: Aproximación lineal Talla vs Estatura [Población General]

Para evaluar exclusivamente a la población masculina o femenina, se genera un arreglo con la tercera columna de los datos importados desde Excel, correspondientes al sexo de cada usuario, y se utiliza nuevamente la función $polyfit$ para para obtener los coeficientes del polinomio. Para realizar el filtrado de datos de acuerdo con el sexo, se utiliza la función $strcmp$ que compara cadenas y buscar coincidencias con el valor de Masculino o Femenino en la columna de datos correspondiente.

```
% Talla contra estatura [Población Masculina]
d=table2array(T(:,3));
[p2,s2]=polyfit(a(strcmp(d,'Masculino'),1),...
    a((strcmp(d,'Masculino')),2),1)
R2=1 - (s2.normr/norm(a((strcmp(d,'Masculino')),2) -...
    mean(a((strcmp(d,'Masculino')),2))))^2

% Talla contra estatura [Población Femenina]
[p3,s3]=polyfit(a((strcmp(d,'Femenino')),1),...
    a((strcmp(d,'Femenino')),2),1)
R3=1 - (s3.normr/norm(a((strcmp(d,'Femenino')),2) -...
    mean(a((strcmp(d,'Femenino')),2))))^2
```

```
p2 = 1×2
    0.1196    6.4418

s2 = struct with fields:
       R: [2×2 double]
      df: 98
   normr: 11.0354
R2 = 0.3602
p3 = 1×2
    0.0918    9.5443

s3 = struct with fields:
       R: [2×2 double]
      df: 98
   normr: 8.9869
R3 = 0.3487
```

Figura 11: Aproximación lineal Talla vs Estatura [Población Masculina y Femenina] (Individual)

Considerando la pendiente como $y = mx + b$, se presentan las ecuaciones propias de cada conjunto de datos, junto con el valor de su coeficiente de determinación R^2

- Población General

$$y = 0{,}16x - 0{,}9917 \qquad R^2 = 0{,}6289$$

- Población Masculina

$$y = 0{,}1196x + 6{,}4418 \qquad R^2 = 0{,}3602$$

- Población Femenina

$$y = 0{,}0918x + 9{,}5443 \qquad R^2 = 0{,}3487$$

Por ultimo se grafican la pendientes obtenidas con su correspondiente conjunto de datos. Se utiliza la función $polyval$ para evaluar el polinomio p en cada punto de x. De esta forma se puede mostrar la ecuación correspondiente a cada linea de tendencia obtenida para las 3 poblaciones como leyenda en cada una de las gráficas.

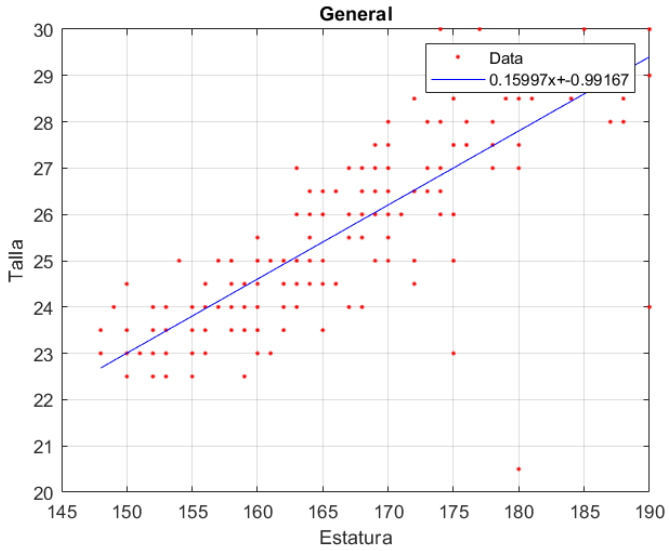

Figura 12: Gráfica de Talla vs Estatura [Población General] con linea de tendencia y ecuación de primer grado

En la gráfica correspondiente a la población general, se observa una menor dispersión de datos, esta afirmación es soportada por el valor de R^2 obtenido, que es mayor a los valores correspondientes para las otras poblaciones. Se entiende que al ser una mayor cantidad de datos evaluados, se generan menos espacios entre entidades.

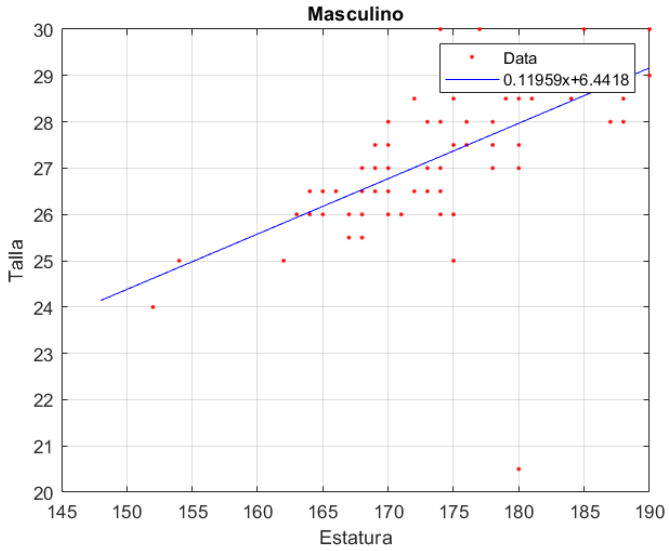

Figura 13: Gráfica de Talla vs Estatura [Población General] con linea de tendencia y ecuación de primer grado

En la gráfica correspondiente a la población masculina, los datos poseen una mayor dispersión entre si, así como un valor más pequeño de R^2. Se observa claramente la relación entre una mayor estatura y un numero de calzado mas grande, sin embargo se pueden apreciar un par de datos que se alejan de la zona de dispersión, lo que podría significar un error generado por el usuario al contestar el formulario con su información personal.

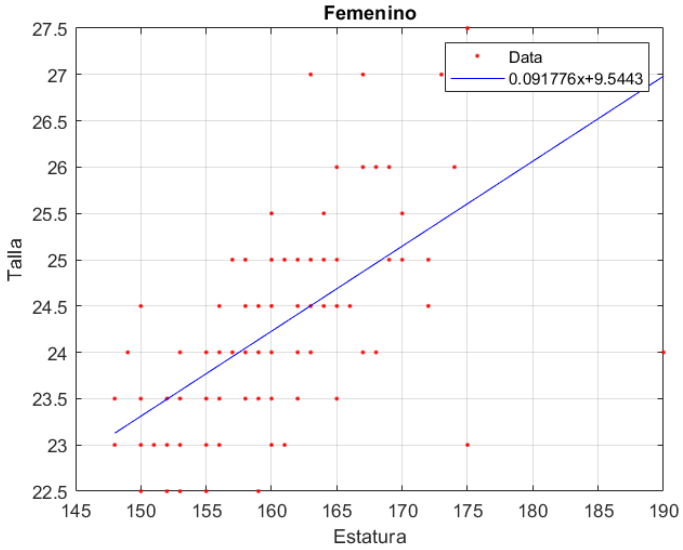

Figura 14: Gráfica de Talla vs Estatura [Población General] con linea de tendencia y ecuación de primer grado

La gráfica correspondiente a la población femenina es similar a la masculina, con el menor valor de R^2 obtenido entre las 3 gráficas. En contraste con la población masculina, se observa una relación entre estaturas menores a 170 cm con un numero de calzado más pequeño. De igual manera se muestran excepciones con bastante presencia, superando la zona de distribución de la mayoría de los datos referente a la talla de calzado.

3. Conclusiones

Se realizó la importación y procesamiento de un set de datos, para visualizar el comportamiento y distribución de información en condiciones reales. Se cumplieron los objetivos de de poner a prueba herramientas de análisis estadístico en MATLAB y desarrollar competencias en la evaluación e interrelación con fines estadísticos de la información recopilada.

En la elaboración de las gráficas correspondientes a la función gaussiana e histograma se observó una distribución de datos similar, lo que corrobora la correlación entre estos elementos. También se comprobó la probabilidad de que una persona tomada al azar se encuentre dentro de la primera desviación estándar, al realizar el cálculo correspondiente en Fig. 9, y determinar el mismo valor observado en la Fig. 1.

Para el desarrollo de este ejercicio, fue necesaria una limpieza y filtrado de datos, ya que muchos usuarios ingresaron su información de manera distinta a la solicitada, además de existir la probabilidad contabilizar datos erróneos, lo que significa que para este tipo de análisis es necesario contar con medidas que minimicen el numero de datos no aptos y así poder optimizar el trabajo.

4. Repositorio

https://github.com/a01745982/Examen1

Referencias

[1] HMN. *Aproximación lineal*. s/f. URL: https://hmn.wiki/es/Linear_approximation.

[2] José López. *Medidas de tendencia central Economipedia*. 2020. URL: https://economipedia.com/definiciones/medidas-de-tendencia-central.html.

[3] Superprof. *El histograma*. s/f. URL: https://www.superprof.es/apuntes/escolar/matematicas/estadistica/descriptiva/histograma.html.

[4] TyC. *Distribución normal o campana de Gauss en Excel*. 2016. URL: http://trucosycursos.es/la-distribucion-normal-o-campana-de-gauss-en-excel/.

[5] Guillermo Westreicher. *Histrograma Economipedia*. s/f. URL: https://economipedia.com/definiciones/histograma.html.

[6] Wikipedia. *Función gaussiana*. s/f. URL: https://es.wikipedia.org/wiki/Funci%5C%C3%5C%B3n_gaussiana.