UNIVERSITAT POLITÈCNICA DE VALÈNCIA

INFORME ESTADÍSTICO SOBRE LOS DIAMANTES

(SEGUNDA ENTREGA)

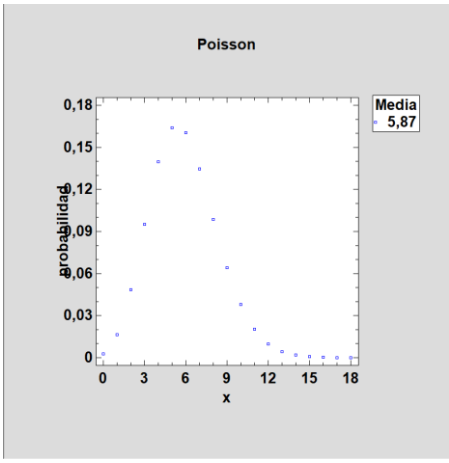
Ismael Fernández Herreruela

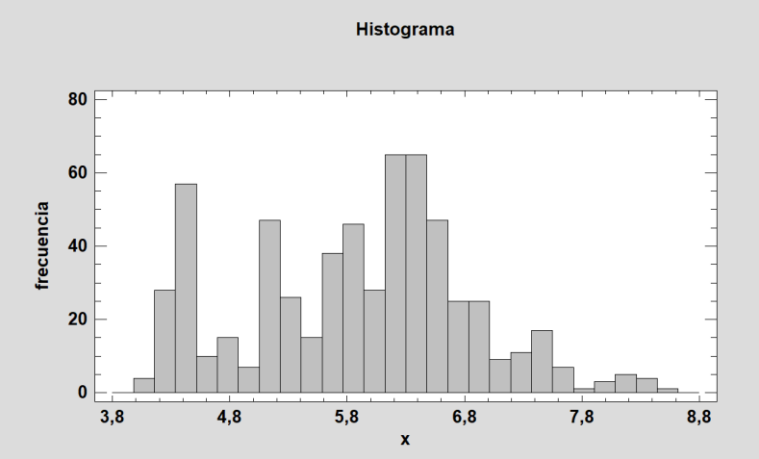
Grupo D

EJERCICIOS

Actividad 17

Todas las variables son continuas por lo que no pueden ser distribuciones normales ni tampoco pueden ser de Poisson. En mi caso tengo que hacer el apartado a).





En ambos graficos podemos ver que una vez se llega a la media, empieza a disminuir significativamente.

Actividad 18

Pruebas de bondad de la variable Quilates:

	Exponencial	*Lognormal*	*Triangular*	*Uniforme*
DMAS	0,120449	0,0957211	0,192898	0,423093
DMENOS	0,296002	0,138061	0,0282185	0,0126301
DN	0,296002	0,138061	0,192898	0,423093
Valor-P	0,0	0,0	0,0	0,0

Pruebas de bondad de la variable x:

	Exponencial	*Lognormal*	*Triangular*	*Uniforme*
DMAS	0,246296	0,0826697	0,101284	0,255415
DMENOS	0,506439	0,100874	0,0254269	0,0325097
DN	0,506439	0,100874	0,101284	0,255415
Valor-P	0,0	0,00000880991	0,00000796764	0,0

Podemos observar que en ambos casos no se tratan ni de distribuciones exponenciales ni uniformes ya que el P-Valor es menor que 0,05.

Actividad 19

Resumen Estadistico para x

Recuento	606
Promedio	5,87848
Desviación Estándar	0,953316
Coeficiente de Variación	16,217%
Mínimo	4,06
Máximo	8,54
Rango	4,48
Sesgo Estandarizado	0,169543
Curtosis Estandarizada	-2,32269

Resumen Estadístico para y

Recuento	606
Promedio	5,87568
Desviación Estándar	0,945815
Coeficiente de Variación	16,0971%
Mínimo	4,05
Máximo	8,51
Rango	4,46
Sesgo Estandarizado	0,0652404
Curtosis Estandarizada	-2,35301

Podemos observar que en ambas variables nos quedamos bastante cerca de que sean distribuciones normales pero no llegan a serlo.

Actividad 20

Ninguna sigue una distribución normal, aunque las variables “x,y,z” si que se aproximan a una distribución normal.

Resumen Estadístico

	x	*y*	*z*
Recuento	606	606	606
Promedio	5,87848	5,87568	3,63398
Desviación Estándar	0,953316	0,945815	0,584523
Sesgo Estandarizado	0,169543	0,0652404	-0,224021
Curtosis Estandarizada	-2,32269	-2,35301	-2,5946

Actividad 21

Para crear las 2 columnas nuevas utilizo el método especificado en el Word. A partir de ahí, he pasado los datos a una hoja Excel para hacer la suma de ambas columnas de una forma mas sencilla y rápida. Una vez realizada la suma he vuelto a pasar los datos al Statgraphics en una nueva columna y al realizar el análisis de esta columna obtenemos los siguientes datos:

Resumen Estadístico para Col_9

Recuento	100
Promedio	18,3772
Desviación Estándar	4,00423
Coeficiente de Variación	21,7891%
Mínimo	10,2055
Máximo	27,1202
Rango	16,9147
Sesgo Estandarizado	0,570746
Curtosis Estandarizada	-1,52206

Sabemos que una suma de variables normales da lugar a una nueva variable normal. Por tanto, podemos calcular su media y su desviación de manera teórica de la siguiente manera:

$E(Y)=E(Y1+Y2)=E(Y1)+E(Y2)=15+4=19$

$\sigma2(Y)= \sigma2(Y1+Y2)=\sigma2(Y1)+ \sigma2(Y2)=16+9=27$

$\sigma=5,19$

Como podemos observar, los resultados teóricos no son exactamente iguales pero son cercanos. Esto se debe a que tenemos una muestra demasiado pequeña como para obtener resultados fiables.

Actividad 22

Utilizare la variable "y", la cual tiene m=5,87 y σ= 0,945815. Cogemos 10 datos aleatorios con un intervalo de confianza de 95%.Dado que es conocida la desviación típica poblacional, podemos usar el valor crítico de la tabla de la normal tipificada que deja un 2,5% de los datos a la derecha cuando es positivo y un 2,5% a la izquierda cuando es negativo. Por lo tanto se nos quedaría 5,87 ±1,96 = (3,91|7,83).

Actividad 23

Percentiles para y

	Percentiles
1,0%	4,22
52,0%	6,08

Como podemos observar, tenemos que el percentil 52 es 6,08 tras analizarlo con el Statgraphics. Para comprobar la hipótesis realizamos una prueba de hipótesis con el programa y observamos los siguientes resultados:

Prueba de Hipótesis para x
Media Muestral = 5,87848
Mediana Muestral = 6,0
Desviación Estándar de la Muestra = 0,953316

Prueba t
Hipótesis Nula: media = 6,08
Alternativa: no igual

Estadístico t = -5,20372
Valor-P = 1,95715E-7
Se rechaza la hipótesis nula para alfa = 0,05.

Se rechaza la hipótesis ya que el p-valor es menor de 0,05 y 0,01que son los valores alfa.

Actividad 24

Utilizare la variable “y” que tiene una media de 5,87 y una desviación típica de 0,945.

Tomamos una muestra de 10 datos con un nivel de confianza de 95%.

Usaremos la siguiente formula: $\frac{(N-1)s^2}{\sigma^2} \sim \chi^2_{N-1}$ donde N es el tamaño de la muestra, en este caso 10.

Como hemos hecho previamente, se dejan 2,5% por la derecha y 2,5% por la izquierda, por lo que sabiendo por esto podemos mirar la tabla de chi-cuadrado y obtener los siguientes resultados: g1=2,7 y g2=19,023. Obtenemos estos resultados ya que buscamos en la tabla mirando por n=9 ya que es el resultado de la muestra menos 1, y buscando por 0,975 y 0,025.

Con estos datos ahora podemos decir que la probabilidad de que $\frac{(N-1)s^2}{\sigma^2}$ se encuentre entre 2,7 y 19,023 es de = P2,7 < $(N-1)s2\sigma2$<19,023)=0.95.

Esto se nos quedaría en el siguiente intervalo [2,7σ2$(N-1)$,19,023σ2$(N-1)$], es decir [2,8×0,945^2(5−1), 11.143×0,945^2(5−1)] = [0,26,|1,887]

Actividad 25

Utilizare la variable “y” la cual tiene una media de 5,87 y una desviación típica de 0,945. Tomamos 12 datos y queremos comprobar cual es la probabilidad de que la varianza sea superior a 3* σ. $P(\chi$>(11×3×0,945)/0,945^2)=34,92. Introduciendo estos datos en el programa obtenemos los siguientes resultados:

Área Cola Superior (>)

Variable	*Dist. 1*	*Dist. 2*	*Dist. 3*	*Dist. 4*	*Dist. 5*
34,92	0,0				

Actividad 26

Para empezar, sabemos que la media de la variable es 5.7 y su desviación estándar es 0.945. La varianza de la población es como resultado 0,893025.

Como se indica en el problema, ahora tomamos dos muestras, cada una de 14 valores. Sea S22 la varianza de la segunda muestra y S12 la varianza de la primera muestra.

La distribución F de Snedecor, que conecta las varianzas de dos muestras independientes de una población normal, se puede utilizar para calcular la probabilidad solicitada.
F = (S2² / σ²) / (S1² / σ²).
En este caso, nos interesa calcular la probabilidad de que la varianza de la segunda muestra sea mayor que el triple de la de la primera, lo que equivale a: P(S2² > 3 S1²).
Esto podría escribirse como: P(S2² / S1² > 3).
Por tanto, debe determinarse la probabilidad de que el estadístico F sea mayor que

3. La fórmula de cálculo de la distribución F es: $F = [(n2 - 1) S2^2] / [(n1 - 1) S1^2]$.
Los dos tamaños de muestra son N1 = N2 = 14. A la luz de esto, la expresión de F se convierte en: $F = S2^2 / S1^2$.
Podemos sacar S22 de la expresión original de F e insertarlo en la expresión de probabilidad de la siguiente manera:.

$S2^2 = 3\ S1^2$.
$S1^2 / S2^2 = 1/3$.
$F = 1 / (1/3) = 3$.

Como resultado, debemos determinar la probabilidad de que la distribución F tenga más de tres grados de libertad (n1-1 = n2-1 = 13). Usando Statgraphics descubrimos que:.

$P(F > 3) = 0,027$

Por lo tanto, la probabilidad de que la varianza de la segunda muestra sea mayor que tres veces la de la primera muestra es aproximadamente 0,027 o 2,7%.

Actividad 27

Con un 99 por ciento de confianza y usando x como muestra, Statgraphics nos proporciona el siguiente intervalo de confianza para la media de la población: [5.77873; 5.97823].

La interpretación práctica es que, si recolectamos un gran número de muestras de la población y calculamos los intervalos de confianza para cada uno a un nivel del 99 por ciento, el verdadero valor de la media de la población estará contenido dentro de estos intervalos con una probabilidad del 99% y saldrá con un 1% de probabilidad. En otras palabras, si calculamos los intervalos de confianza para cada una de las 100 muestras, 99 de ellas contendrán el verdadero valor medio de la población y solo una no.

Actividad 28

Para la variable “y”, con un 95% de confianza obtenemos el siguiente intervalo: [0,895395; 1,0023], y con un 99% obtenemos el siguiente: [0,880326; 1,02103].
Si queremos más seguridad, deberíamos usar el 99%, pero si solo queremos un intervalo con una alta probabilidad y un intervalo más amplio, el 95% servirá. Cuanto mayor sea el intervalo de confianza, más probable será encontrar el valor real. Es decir, usaremos un intervalo u otro dependiendo de lo que estemos tratando de encontrar con el intervalo. Sin conocer el contexto, es imposible decir qué intervalo es mejor porque, al reducir el error asociado con rechazar la hipótesis nula como verdadera, aumentamos el riesgo asociado con aceptarla como falsa. Por esta razón, debemos lograr un equilibrio y no solo utilizar el nivel de confianza más alto.

Actividad 29

Estadisticas de Resumen
Datos/Variable: x

Color	*Recuento*	*Promedio*	*Varianza*
F	117	5,80137	0,529758
G	329	5,97198	1,17272
D	63	5,89619	0,58715
E	97	5,64289	0,607789
Total	606	5,87848	0,908811

Como podemos observar, F tiene la varianza menor y G tiene la mayor, por lo tanto se nos quedaría una tabla así:

	Muestras	Varianza
F	117	0,529
G	329	1,172

Para comprobar si las diferencias son significativas, primero se crean las hipótesis nula y alternativa: H0: $\mu 1 = \mu 2$ (las medias de las dos poblaciones no son significativamente diferentes) Ha: $\mu 1 \neq \mu 2$ (las medias de las dos poblaciones son significativamente diferentes) donde $\mu 1$ y $\mu 2$ son las medias poblacionales de las dos muestras. Se elige un nivel de significación α (por ejemplo, 0,05), que representa la probabilidad de rechazar H0 cuando H0 es verdadera. El estadístico t de Student se calcula mediante la siguiente fórmula: $t = (\bar{x}1 - \bar{x}2) / [s^2p * (1/n1 + 1/n2)]^{(1/2)}$ donde $\bar{x}1$ y $\bar{x}2$ son las medias muestrales de las dos muestras, n1 y n2 son los tamaños de muestra y s^2p es la varianza combinada de las dos muestras: $s^2p = [(n1-1) * s_1^2 + (n2-1) * s^2] / (n1 + n2 - 2)$ En este caso, hay: $n1 = 117$, $s_1^2 = 0,529$, $\bar{x}1 = 5,8$ $n2 = 329$, $s2^2 = 1,172$, $\bar{x}2 = 5,97$ La varianza es: $s^2p = [(117-1) * 0,529 + (329-1) * 1,172] / (117 + 329 - 2) = 1,019$. Por lo tanto, el estadístico t es: $t = (5,8 - 5,97) / [1,019 * (1/117 + 1/329)]^{(1/2)} = -3,535$ Con 444 grados de libertad (n1 + n2 - 2), el valor t-crítico para un nivel de significancia de $\alpha = 0,05$ es ±1,965. El valor absoluto del estadístico t calculado es mayor que el valor crítico de t, por lo que se rechaza la hipótesis nula. Esto significa que existe suficiente evidencia estadística para confirmar que las medias de la población son diferentes y que la diferencia observada es estadísticamente significativa. En conclusión, con base en los datos presentados, se puede decir que las diferencias observadas son estadísticamente significativas.