UNIVERSITAT POLITÈCNICA DE VALÈNCIA

INFORME ESTADÍSTICO SOBRE LOS DIAMANTES

(TERCERA ENTREGA)

Ismael Fernández Herreruela

Grupo D

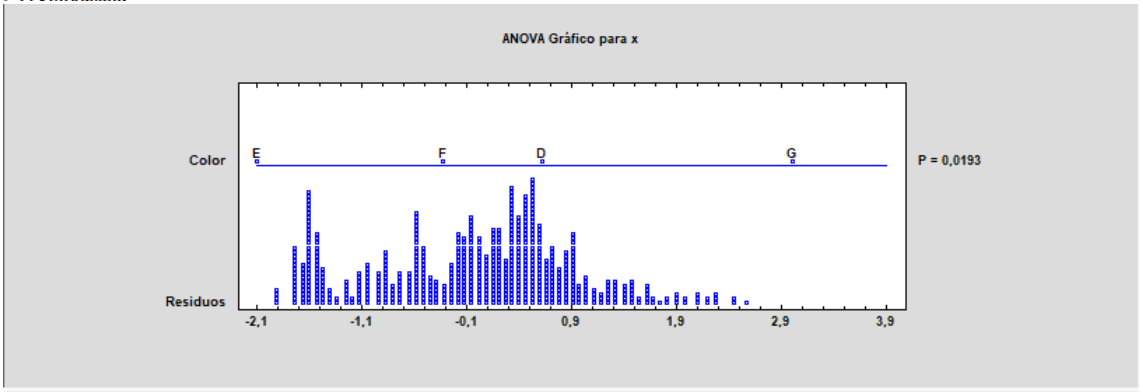
EJERCICIOS

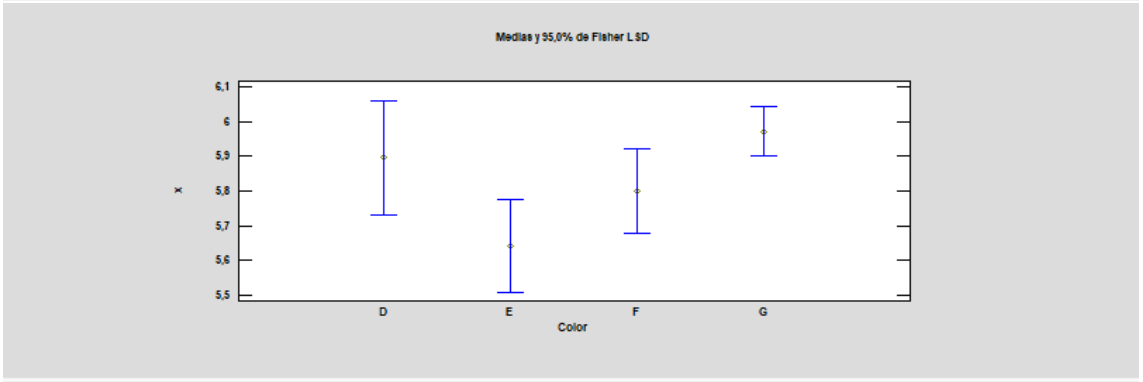
ACTIVIDAD 30

Análisis de Varianza para x - Suma de Cuadrados Tipo III

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
EFECTOS PRINCIPALES					
A:Color	8,97533	3	2,99178	3,33	0,0193
RESIDUOS	540,855	602	0,898431		
TOTAL (CORREGIDO)	549,831	605			

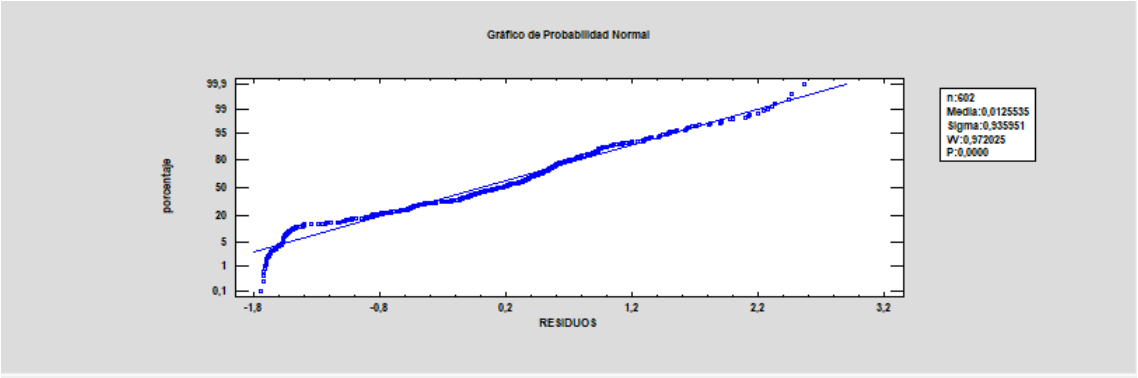
Todas las razones-F se basan en el cuadrado medio del error residual





Como podemos observar en el grafico de medias, prácticamente todos los intervalos se solapan, en el caso del intervalo G no se llega a solapar con E pero si solapa con F y D. No todos tienen la misma amplitud porque cada uno tiene un numero diferente de elementos por lo que unos abarcaran mas que otros.

El efecto principal del color en el estudio fue estadísticamente significativo. Esto se indica mediante el valor p (0,0193), que es más bajo que los niveles de significación comúnmente utilizados, como 0,05. Además, la relación F (3,33) es mayor que 1, lo que refuerza la significación estadística del efecto de color. Esto significa que el color tiene un efecto medible sobre la variable de respuesta en estudio. En conclusión, el análisis indicó que el color tuvo un efecto estadísticamente significativo en las variables de respuesta estudiadas.



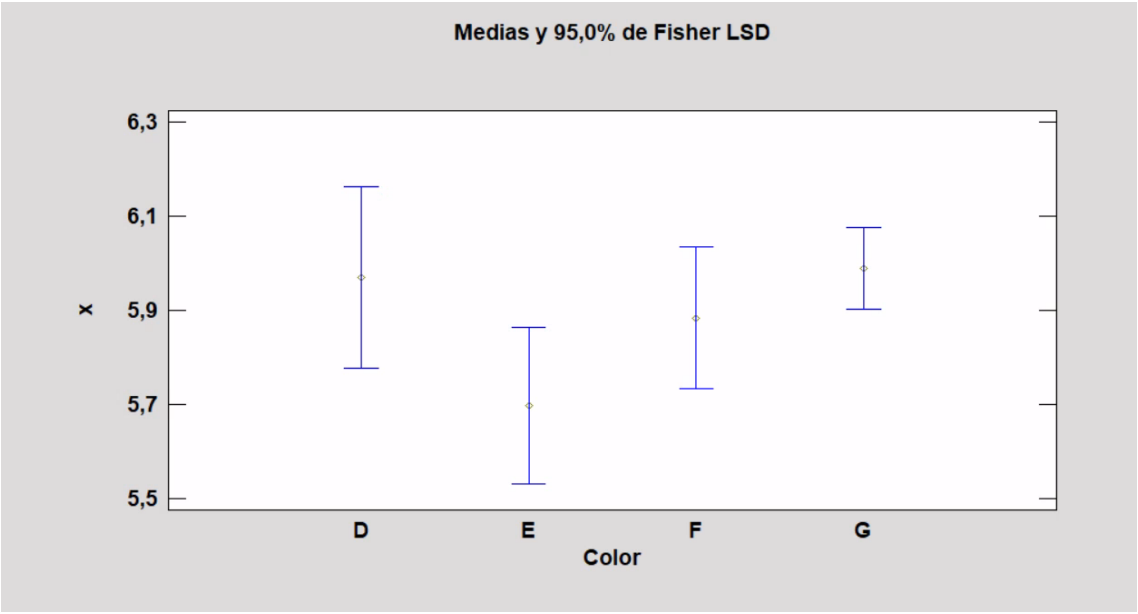
Observamos que se trata de una distribución platicurtica sin datos anómalos.

ACTIVIDAD 31

Análisis de Varianza para x - Suma de Cuadrados Tipo III

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
EFECTOS PRINCIPALES					
A:Color	4,37098	3	1,45699	1,62	0,1845
B:Corte	4,48491	2	2,24246	2,49	0,0840
INTERACCIONES					
AB	1,06407	6	0,177345	0,20	0,9777
RESIDUOS	535,538	594	0,901579		
TOTAL (CORREGIDO)	549,831	605			

Todas las razones-F se basan en el cuadrado medio del error residual

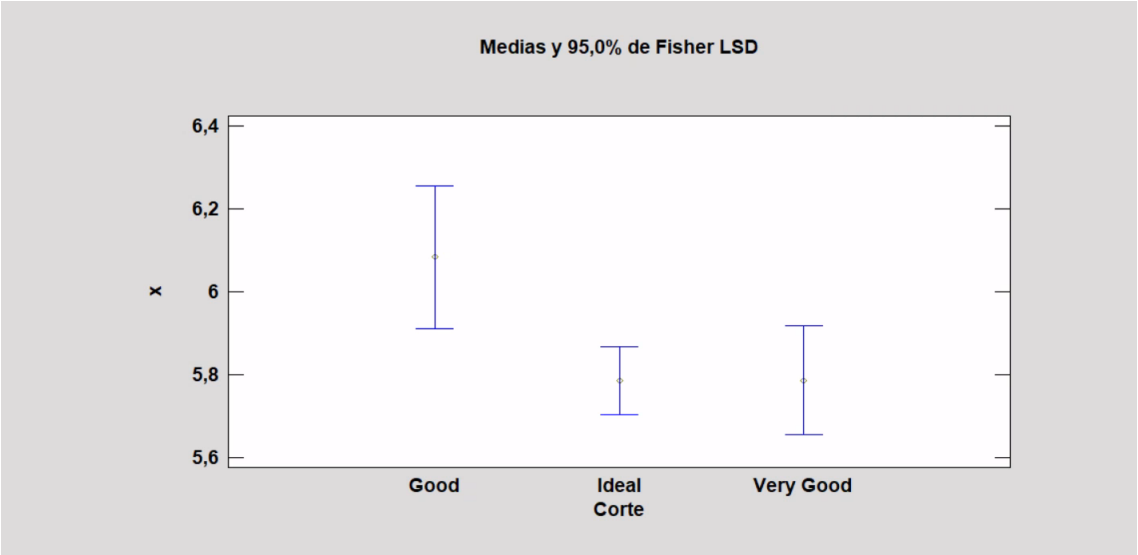


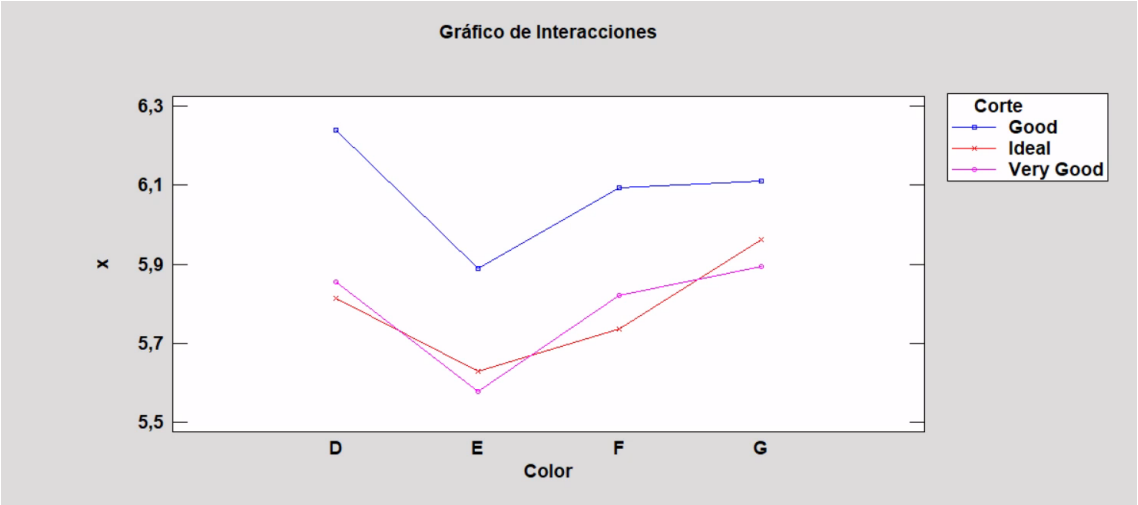
Análisis de Varianza para y - Suma de Cuadrados Tipo III

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
EFECTOS PRINCIPALES					
A:Color	7,61897	3	2,53966	2,88	0,0354
B:Corte	3,78444	2	1,89222	2,15	0,1179
RESIDUOS	529,247	600	0,882078		
TOTAL (CORREGIDO)	541,212	605			

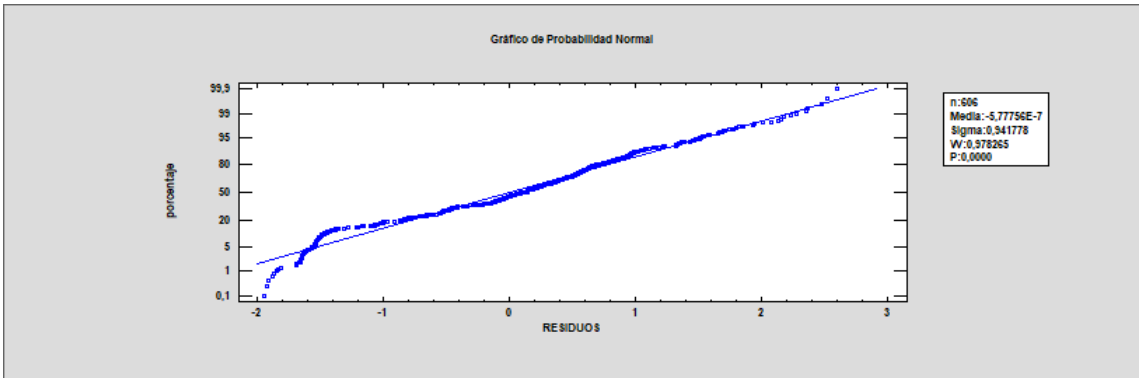
Todas las razones-F se basan en el cuadrado medio del error residual

Podemos ver que en este caso la variable A es significativa ya que tiene un valor-p menor a 0,05.





La variabilidad explicada por la interacción entre las variables A y B en esta situación se muestra mediante la suma de cuadrados de la interacción AB, que en este caso es 1.06407. No se encuentran pruebas convincentes para apoyar la hipótesis, que tiene un valor P alto de 0,9777, una hipótesis nula de que la interacción es cero. Esto sugiere que la interacción AB no afecta la variable de respuesta de manera estadísticamente significativa.



Resumen Estadístico para RESIDUOS

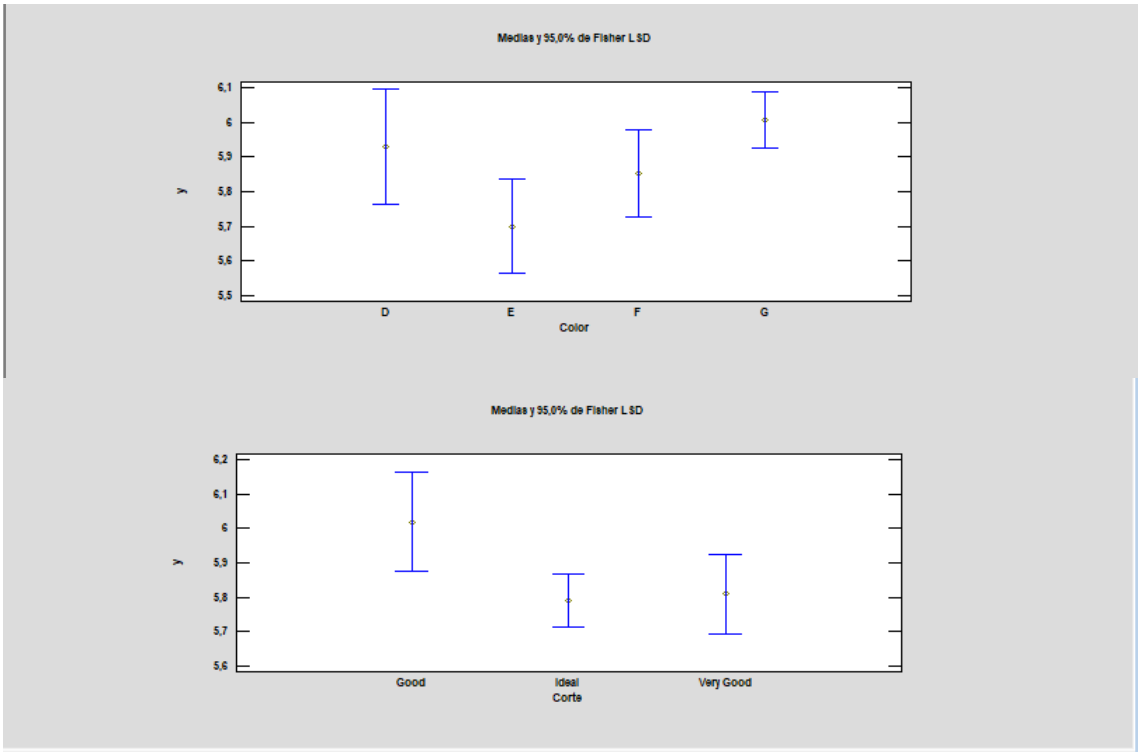
Recuento	606
Promedio	-5,77756E-7
Desviación Estándar	0,941778
Coeficiente de Variación	-1,63006E8%
Mínimo	-1,94092
Máximo	2,60014
Rango	4,54106
Sesgo Estandarizado	-0,431921
Curtosis Estandarizada	-1,99373

El análisis estadístico de los residuos sugiere que en promedio están equilibrados en torno a cero, pero tienen una dispersión moderada. Además, la distribución de los residuos parece estar sesgada hacia la izquierda y tener colas más ligeras y achatadas en comparación con una distribución normal.

ACTIVIDAD 32

Análisis de Varianza para y - Suma de Cuadrados Tipo III

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
EFECTOS PRINCIPALES					
A:Color	7,61897	3	2,53966	2,88	0,0354
B:Corte	3,78444	2	1,89222	2,15	0,1179
RESIDUOS	529,247	600	0,882078		
TOTAL (CORREGIDO)	541,212	605			



Color:

Los diferentes colores entre los factores analizados son estadísticamente significativamente diferentes. Dado que el valor P es menor que un umbral de significación comúnmente

utilizado, como 0,05, podemos concluir que el color tiene un efecto significativo en la variable medida.

Corte:

No se encontraron diferencias estadísticamente significativas entre los distintos cortes de los factores analizados. El valor P es mayor que el umbral de significación de 0,05, lo que indica que el valor de corte no tuvo un efecto significativo en la variable medida en el análisis.

En general, estas conclusiones prácticas indican que el factor de color tiene un efecto significativo en la variable medida, mientras que el factor de corte no muestra una influencia estadísticamente significativa.

Para el efecto principal A: Color, el valor P es 0,0354. Esto significa que hay un 3,54 % de posibilidades de obtener un resultado tan extremo como el observado, suponiendo que no haya una diferencia real entre los grupos de colores. Si se establece un nivel de significancia de alfa=0,05 (5%), entonces el valor P obtenido es menor que alfa. Por tanto, podemos concluir que el efecto principal A: Color es estadísticamente significativo al 95% de nivel de confianza.

Por otro lado, para el efecto principal B: Corte, el P-valor es 0.1179. Esto significa que la probabilidad de obtener un resultado tan extremo como el observado es del 11,79 %, suponiendo que no exista una diferencia real entre los grupos de corte. Si establece un nivel de significancia de alfa=0,05 (5%), obtiene un valor de p mayor que alfa. Por lo tanto, no podemos concluir que el efecto principal B: Corte sea estadísticamente significativo al 95% de nivel de confianza.

Si el valor alfa inicial no alcanza un valor razonable (como alfa=0,05), puede deberse a las siguientes razones

Tamaño de la muestra: si las muestras utilizadas en el estudio son pequeñas, la capacidad del análisis para detectar diferencias estadísticamente significativas puede verse limitada. Es más probable que se detecten efectos significativos a medida que aumenta el tamaño de la muestra.

Variabilidad de los datos: si los datos tienen una gran variabilidad dentro del grupo y una pequeña variabilidad entre grupos, puede ser difícil detectar diferencias significativas. En estos casos, se requieren tamaños de muestra más grandes o una variabilidad reducida en los datos para lograr significación estadística.

Pequeños efectos: si el verdadero efecto del factor que se está probando es pequeño en relación con la variabilidad general de los datos, es posible que se requiera un tamaño de muestra más grande para detectar una diferencia significativa con un nivel de confianza determinado.

ACTIVIDAD 33

Análisis de Varianza para y - Suma de Cuadrados Tipo III

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
EFECTOS PRINCIPALES					
A:Color	4,13686	3	1,37895	1,55	0,2003
B:Corte	4,21104	2	2,10552	2,37	0,0946
INTERACCIONES					
AB	1,07878	6	0,179797	0,20	0,9761
RESIDUOS	528,168	594	0,889172		
TOTAL (CORREGIDO)	541,212	605			

Todas las razones-F se basan en el cuadrado medio del error residual

Como resultado obtenemos esta tabla en la que podemos observar que con un índice de 0,05 no obtenemos ningún valor significativo por lo que la tabla de valores definitivos no tendrá ningún efecto.

La interpretación no difiere significativamente según los resultados de la tabla. La interacción AB tampoco es estadísticamente significativa, como tampoco lo son los efectos principales A (Color) y B (Corte). Debido a que el valor de p está tan cerca del nivel de significancia de 0.05, existe una sugerencia de significación potencial en el efecto principal B (corte).

No sería necesario analizar la gráfica de la doble interacción mediante pruebas estadísticas formales si la interacción AB no se considera significativa en el análisis de varianza. No obstante, para comprender cualquier patrón o tendencia que pueda estar presente en los datos, podría ser útil en algunas circunstancias mirar el gráfico.

Dado que la doble interacción no es significativa, en este caso no aporta ninguna información adicional relevante en cuanto a la respuesta.

Resumen Estadístico para RESIDUOS

Recuento	606
Promedio	-7,11997E-7
Desviación Estándar	0,934348
Coeficiente de Variación	-1,31229E8%
Mínimo	-1,85708
Máximo	2,58969
Rango	4,44677
Sesgo Estandarizado	-0,534888
Curtosis Estandarizada	-2,04555

FALTA PAPEL PROBABILISTICO NORMAL

Con el sesgo que hemos obtenido podemos destacar que la lista de residuos tendrá una ligera asimetría negativa. También hay que resaltar que se trata de una distribución platicúrtica ya que hemos obtenido el valor de -2,04 en la curtosis estandarizada.

La ausencia de una asimetría positiva en los datos se puede inferir del hecho de que la asimetría en este caso es negativa, lo cual es relevante cuando se habla de la asimetría positiva. Pero si se observa un sesgo positivo, se recomienda analizar la posibilidad de que haya valores atípicos u otras anomalías en los datos. También puede pensar en transformar los datos para producir una distribución más simétrica, como aplicar una transformación logarítmica o de raíz cuadrada. Esto puede disminuir la asimetría positiva y mejorar la idoneidad de los datos para análisis y modelos estadísticos particulares.

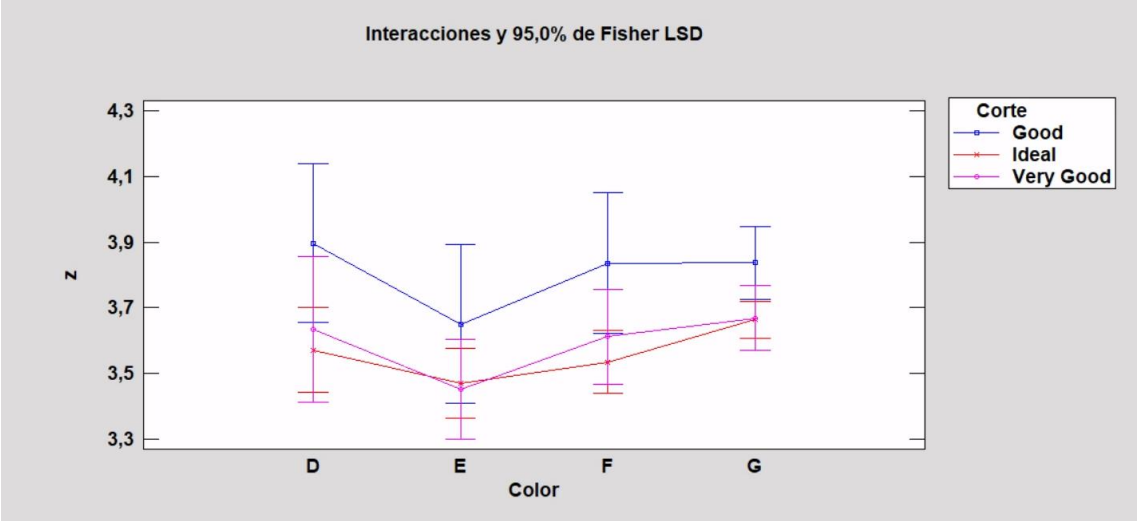
ACTIVIDAD 34

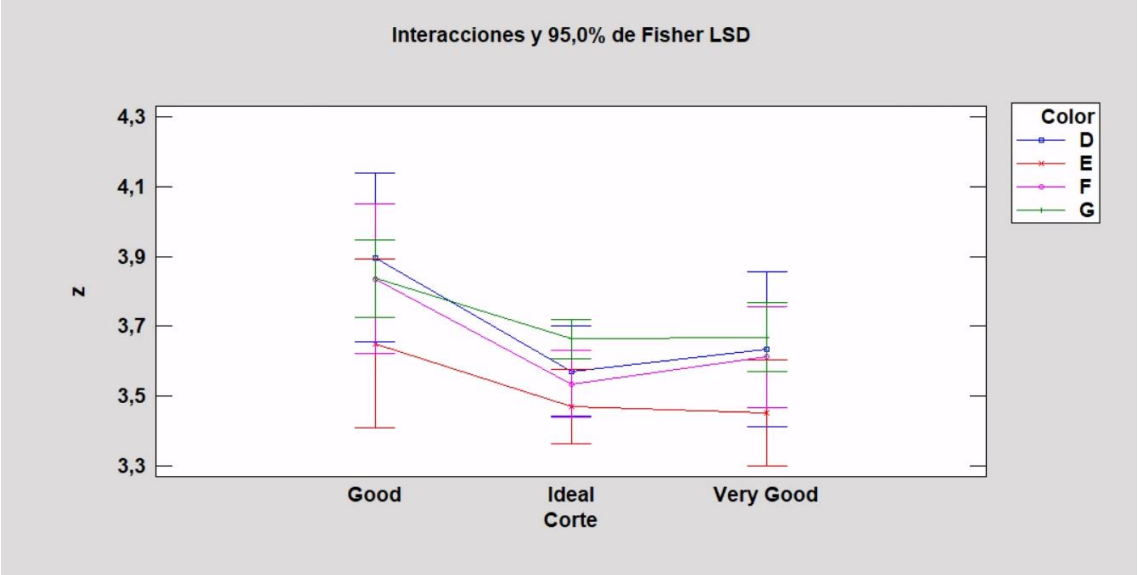
Análisis de Varianza para z - Suma de Cuadrados Tipo III

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
EFECTOS PRINCIPALES					
A:Color	1,99254	3	0,66418	1,98	0,1163
B:Corte	2,89236	2	1,44618	4,30	0,0139
INTERACCIONES					
AB	0,339524	6	0,0565873	0,17	0,9851
RESIDUOS	199,612	594	0,336046		
TOTAL (CORREGIDO)	206,709	605			

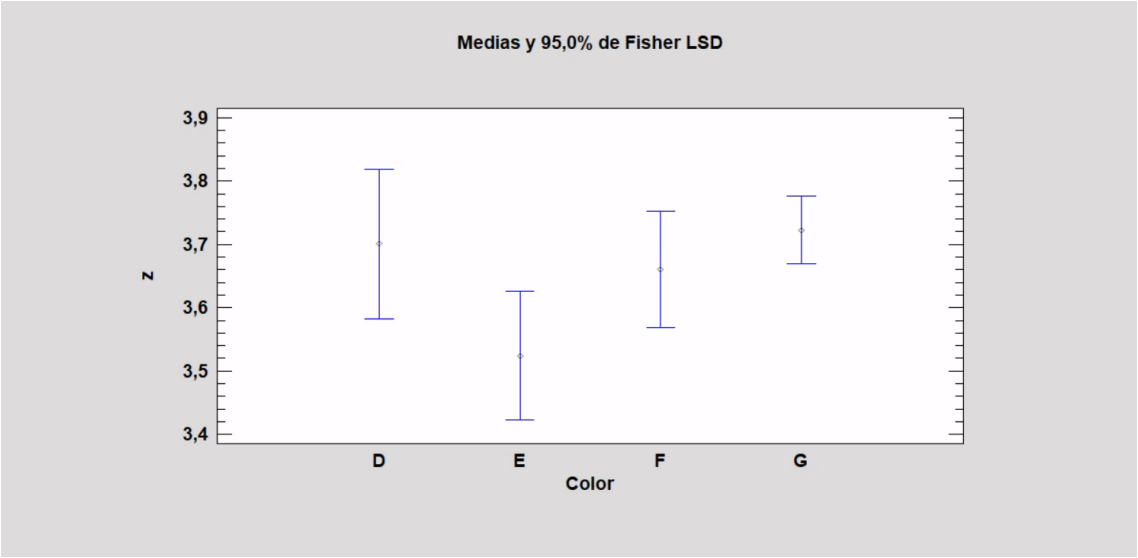
Todas las razones-F se basan en el cuadrado medio del error residual

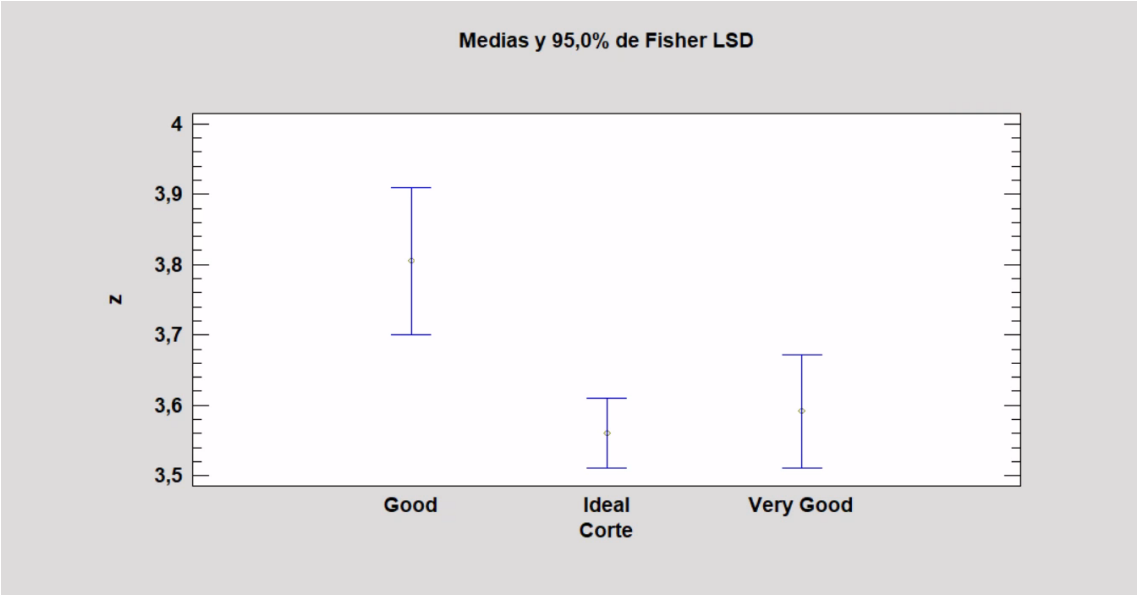
Obtenemos la siguiente tabla y podemos ver que si que tenemos un dato significativo (Corte) para el índice 0,05.





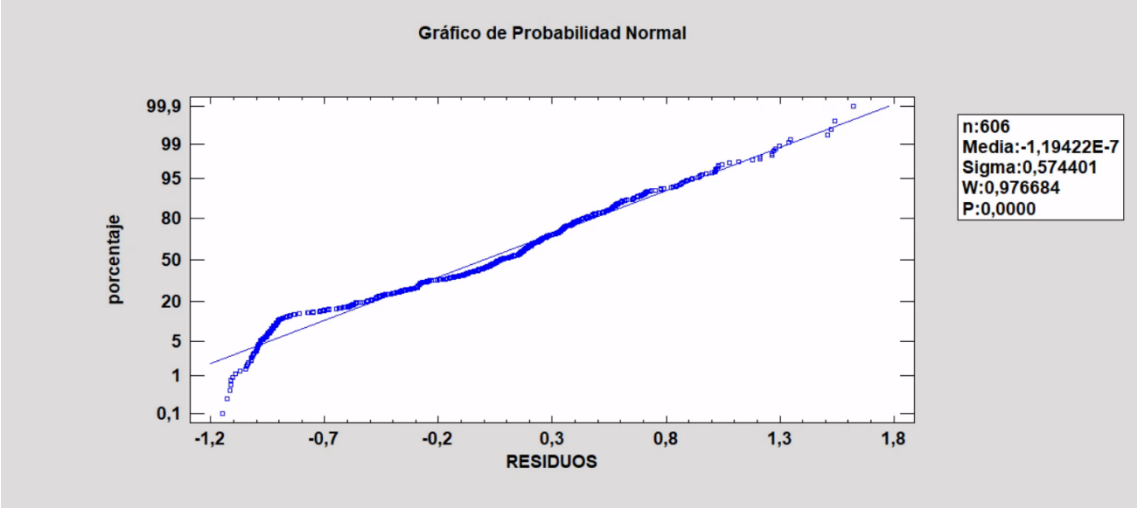
A partir de la primera grafica de doble interacción podemos sacar la información de una manera mas simple que en la segunda, ya que al no haber tantas líneas es mas sencillo comprender el grafico y obtener los datos.

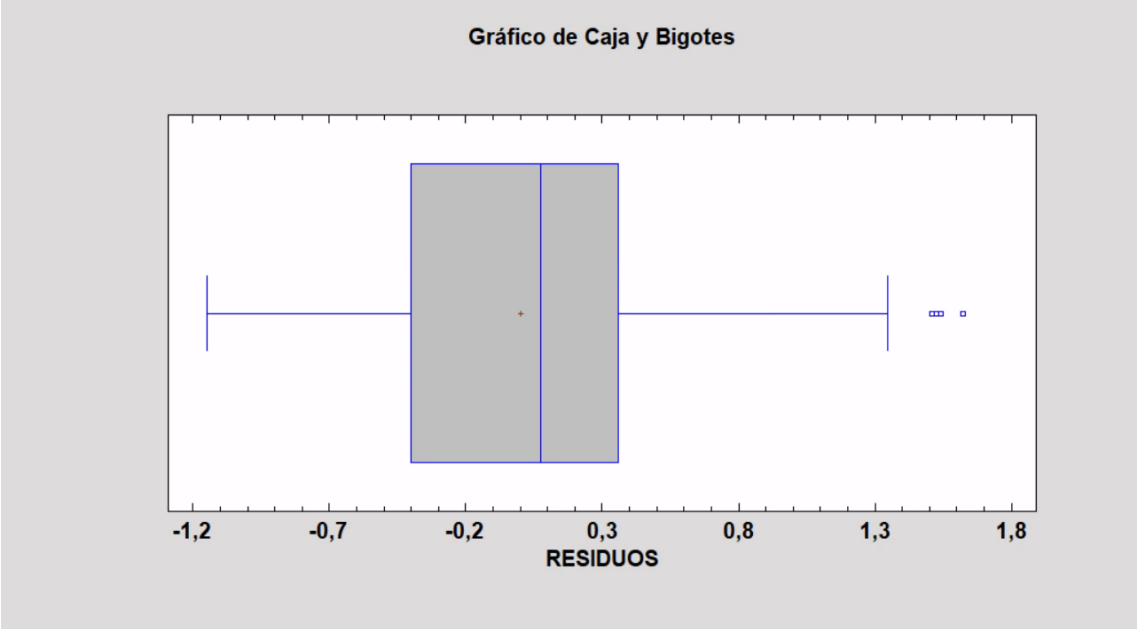




De estos gráficos hay que resaltar que en la variable corte, “Good” no se solapa con ningún otro dato, ya que como habíamos mencionado previamente esta era la variable significativa con un p-valor de 0,0139 el cual es menor que 0,05.

ACTIVIDAD 35





Podemos observar claramente con el grafico que no hay ninguna mezcla de poblaciones. También hay que resaltar que se trata de una distribución platicúrtica. Respecto a los datos anómalos no se aprecia ninguno en la grafica, aunque observando el diagrama de caja y bigotes podemos ver que hay 3-4 datos anómalos pero no son significativos para el resultado.

La afirmación correcta es la número 2.

En el contexto del análisis de varianza (ANOVA), cuando la variable dependiente tiene un sesgo positivo, es mejor ajustar el modelo ANOVA y examinar la distribución de los residuos. Los residuos representan la diferencia entre el valor observado y el valor predicho por el modelo. Si los residuos siguen una distribución asimétrica, esto es una indicación de que el modelo no se ajusta bien a los datos y puede violar los supuestos de ANOVA.

En este caso, la estrategia recomendada es probar diferentes transformaciones de la variable dependiente en un intento de hacer que los residuos se distribuyan de forma razonablemente normal. Esto se puede lograr a través de transformaciones matemáticas como transformaciones logarítmicas, raíces cuadradas, etc. Al aplicar estas transformaciones, buscamos obtener una distribución residual más simétrica y acercarnos a una distribución normal.

Como el sesgo estandarizado es -0,676 se trata de una asimetría negativa.

ACTIVIDAD 36

Covarianzas

	x	y	z	Quilates
x	0,908811	0,899403	0,550054	0,363055
	(606)	(606)	(606)	(606)
y	0,899403	0,894566	0,545064	0,359518
	(606)	(606)	(606)	(606)
z	0,550054	0,545064	0,341667	0,221937
	(606)	(606)	(606)	(606)
Quilates	0,363055	0,359518	0,221937	0,151347
	(606)	(606)	(606)	(606)

La simetría matricial significa que la covarianza entre dos variables es la misma independientemente del orden en que se coloquen. Por ejemplo, la covarianza de x e y es 0,899403 y la covarianza de y y x también es 0,899403. Esto funciona para todas las combinaciones de variables en la matriz. Los valores de la diagonal principal de la matriz representan la covarianza de una variable consigo misma, la varianza de esa variable. Por ejemplo, en la posición (x, x) de la matriz, el valor es 0,908811 y representa la varianza de la variable x. De manera similar, los valores en las ubicaciones (y, y), (z, z) y (Carats, Carats) representan las respectivas varianzas de estas variables. Esta matriz proporciona información útil sobre la relación entre las variables y la variabilidad conjunta. Una covarianza positiva indica una relación directa, es decir, las variables tienden a aumentar o disminuir juntas. Por otro lado, la covarianza negativa indica una relación inversa, donde una variable tiende a aumentar mientras que la otra variable tiende a disminuir. Valores cercanos a cero indican vínculos débiles o nulos.

En resumen, la matriz de covarianza describe la relación y la variabilidad conjunta entre variables, y su simetría asegura que la covarianza entre variables sea independiente del orden. Los valores de la diagonal principal representan las varianzas individuales de las variables.

ACTIVIDAD 37

Correlaciones

	x	y	z	Quilates
x		0,9975	0,9871	0,9789
		(606)	(606)	(606)
		0,0000	0,0000	0,0000
y	0,9975		0,9859	0,9771
	(606)		(606)	(606)
	0,0000		0,0000	0,0000
z	0,9871	0,9859		0,9760
	(606)	(606)		(606)
	0,0000	0,0000		0,0000
Quilates	0,9789	0,9771	0,9760	
	(606)	(606)	(606)	
	0,0000	0,0000	0,0000	

Los valores fuera de la diagonal principal representan correlaciones entre distintas variables. En este caso, el valor de la correlación es muy alto, cercano a 1,0000, lo que indica una fuerte correlación positiva entre las variables. En concreto, los coeficientes de correlación son los siguientes:

La correlación entre x e y es 0,9975.

La correlación entre x y z es 0,9871.

La correlación entre x y quilate es 0,9789.

La correlación entre y y z es 0,9859.

La correlación entre y y el quilate es 0,9771.

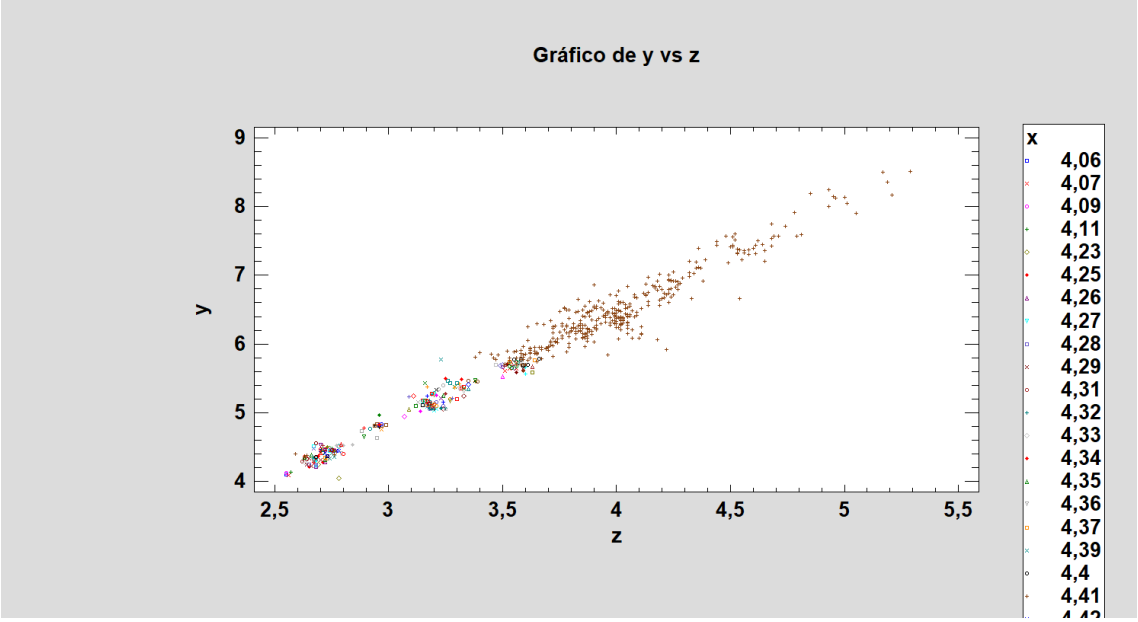
La correlación entre z y quilates es 0.9760.

Estos coeficientes de correlación cercanos a 1,0000 indican que las variables están altamente correlacionadas. En otras palabras, existe una fuerte relación lineal positiva entre las variables x, y, z y quilate. Cuando una variable aumenta, las otras variables tienden a aumentar juntas y viceversa.

En resumen, la matriz de correlación muestra que existe una alta correlación positiva entre las variables, lo que indica que están estrechamente relacionadas. Los valores de la diagonal principal son todos 1,0000, lo que significa que cada variable tiene una correlación perfecta consigo misma.

ACTIVIDAD 38

La pareja de variables con el segundo mayor grado de correlación es la pareja formada por las variables y y z, con un coeficiente de correlación de 0.9859.

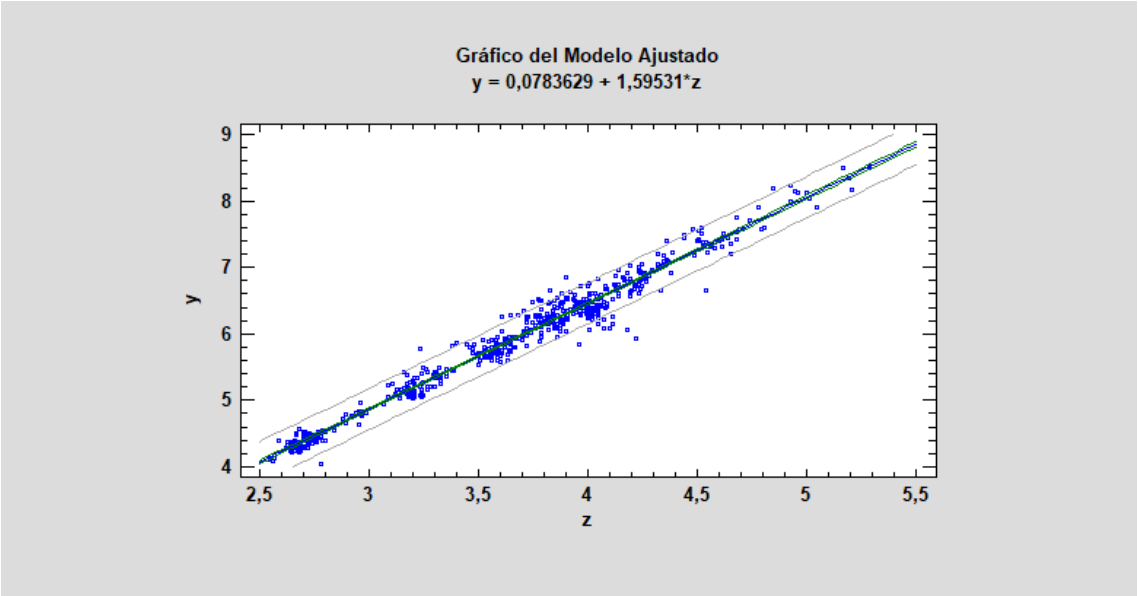


Podemos ver como se traza una línea con pendiente positiva. También podemos observar como los puntos están agrupados y bastante cerca los unos de los otros. Como conclusión podemos decir que la relación es líneal positiva fuerte.

Como hemos mencionado, se trata de una línea con pendiente positiva y que su varianza aumenta de manera constante, también simplemente viendo la forma de la distribución podemos afirmar que se cumple la hipótesis de homocedasticidad.

ACTIVIDAD 39

Utilizare las variables del anterior ejercicio ya que tienen una relación positiva lineal fuerte.



Valores Predichos

	Pronosticado	*Inferior 95%*	*Superior 95%*	*Inferior 95%*	*Superior 95%*
X	*Y*	*Límite Pred.*	*Límite Pred.*	*Límite Conf.*	*Límite Conf.*
2,55	4,1464	3,83498	4,45782	4,11983	4,17297
5,29	8,51754	8,20496	8,83013	8,47965	8,55544

Un intervalo de predicción es una herramienta útil en estadística para estimar el rango de valores dentro del cual se espera que caigan nuevas observaciones futuras. En este caso, los datos presentados parecen ser una serie de observaciones (X) y sus respectivos valores pronosticados (Y) junto con residuos y residuos estandarizados.

Coeficientes

	Mínimos Cuadrados	*Estándar*	*Estadístico*	
Parámetro	*Estimado*	*Error*	*T*	*Valor-P*
Intercepto	0,0783629	0,0405274	1,93358	0,0532
Pendiente	1,59531	0,0110111	144,882	0,0000

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	526,075	1	526,075	20990,87	0,0000
Residuo	15,1375	604	0,0250621		
Total (Corr.)	541,212	605			

Como podemos observar en la tabla el p-valor es menor que 0,05 por lo que si que tenemos una relación significativa. Por otra parte el p-valor de la intercepción es ligeramente mayor que 0,05 pero al ser por tan poco lo podemos considerar significativo.

Y=a+b*X donde a es el intercepto y b es la pendiente -> y=0,078 + 1,595 * z

Esta ecuación nos sirve para calcular el valor Medio en Y esperado para cada valor del eje X. De esta forma se pueden aproximar los resultados de los residuos que se obtendrían dependiendo del valor z. Si el valor z fuese por ejemplo 3 entonces nos quedaría de la siguiente manera:

Y= 0,078+1,595*3

ACTIVIDAD 40

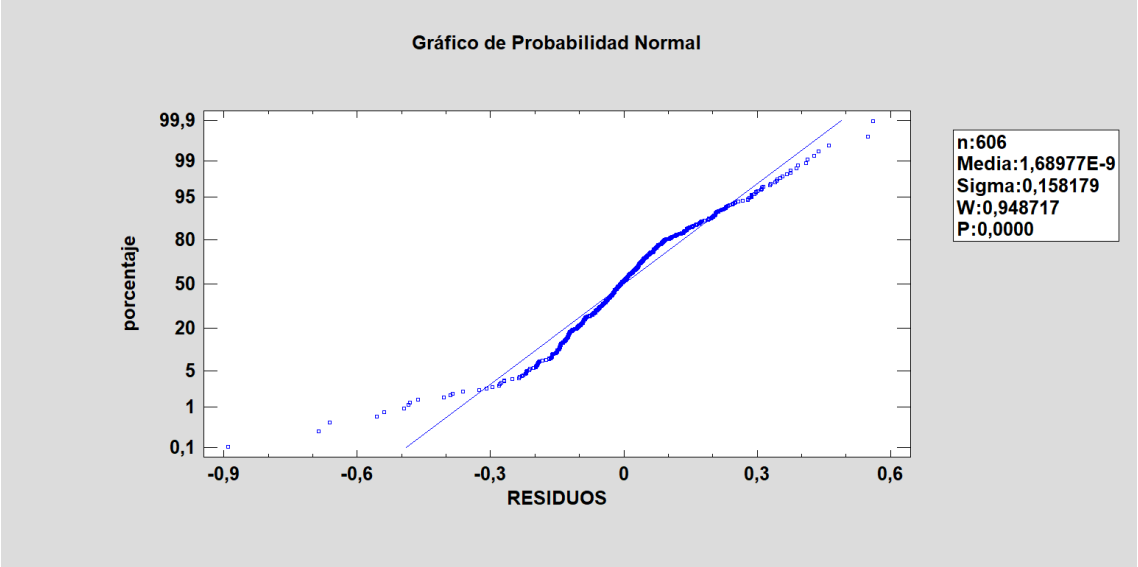
La intersección (a) en la ecuación representa el valor esperado de la variable dependiente cuando la variable independiente es cero. En este caso, el intercepto es 0.078. Esto significa que cuando z es igual a 0, el valor estimado de y es 0,078.

La pendiente (b) en la ecuación representa el cambio medio en la variable dependiente (y) por unidad de cambio en la variable independiente (z). En este ejemplo, la pendiente es 1.595. Esto muestra que por cada unidad de cambio en z, el cambio promedio en y es 1.595.

Es posible que la relación este relacionada con como de grande es el diamante, ya que independientemente de que z sea mas grande que y o al revés, cuando aumenta cualquiera de las 2 variables el diamante es mas grande.

En conclusión, si que hay una relación de causalidad, ya que el tamaño del diamante depende de estas variables.

ACTIVIDAD 41-DESVIACION TIPICA



En resumen, se puede deducir que los residuos tienen una variabilidad relativamente alta en relación con su promedio, una distribución asimétrica con una cola izquierda más pesada y una forma de distribución con colas más pesadas y un pico más pronunciado que una distribución normal. Podemos ver ciertos datos un poco mas separados de los demás pero considero que no son significativos a la hora de obtener el resultado.

La tabla de coeficientes muestra que no hay términos cuadráticos en el modelo. Se utilizan coeficientes estimados para la intersección y la pendiente lineal, pero no hay coeficientes asociados con el término cuadrático. Los valores de los coeficientes tampoco indican una relación significativa entre las variables independiente y dependiente.

En conclusión, no se sospecharon efectos cuadráticos en el modelo, ya que los términos cuadráticos estaban ausentes y ANOVA mostró una variabilidad insuficiente explicada por el modelo.

Coeficientes

	Mínimos Cuadrados	*Estándar*	*Estadístico*	
Parámetro	*Estimado*	*Error*	*T*	*Valor-P*
Intercepto	-5,80656E-8	0,0405274	-0,00000143275	1,0000
Pendiente	1,64435E-8	0,0110111	0,00000149336	1,0000

Como podemos observar en la tabla, el p-valor esta por encima de 0,05 y por lo tanto no es significativo.

Resumen Estadístico para RESIDUOS

Recuento	606
Promedio	1,68977E-9
Desviación Estándar	0,158179
Coeficiente de Variación	9,361E9%
Mínimo	-0,890564
Máximo	0,559935
Rango	1,4505
Cuartil Superior	0,0704332
Sesgo Estandarizado	-3,22155
Curtosis Estandarizada	17,6701

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	526,075	1	526,075	20990,87	0,0000
Residuo	15,1375	604	0,0250621		
Total (Corr.)	541,212	605			

Para obtener la desviación típica cogeremos el cuadrado medio residual y haremos su raíz cuadrada. Nos dará como resultado 0,159.

Si utilizamos la formula previamente definida en el ejercicio 39 obtendriamos lo siguiente:

Y= 0,078+1,595*0,0704332 = 0,1903

Con el 95% fluctuara entre 0,1903 +2* σl y 0,1903 - 2* σ y se nos quedaría el siguiente intervalo:

[-0,1233|0,509]

Dado que el extremo izquierdo es negativo, el intervalo se quedara en [0|0,509] ya que no se pueden tener valores físicos reales menores que 0.

ACTIVIDAD 42

Tras haber realizado todo el trabajo, utilizando diversos métodos de análisis estadístico como pueden ser tablas de frecuencias normales y cruzadas, diagramas de barras, tablas resumen (con datos como máximo, mínimo, rango intercuartílico, media, mediana, desviación típica,

coeficiente de asimetría, coeficiente de curtosis) de las 4 variables continuas, histogramas, papeles probabilísticos, estudios de pauta de variabilidad y en esta ultima entrega los gráficos LSD, anova simple y multifactorial, estudios de residuos y matrices de varianza-covarianza, podemos sacar como conclusión que existe una correlación entre las variables para determinar que diamantes son mejores y cuales peores. Por ejemplo, el tamaño de un diamante puede influir en como se ve su corte, si es un diamante muy pequeño se apreciaran menos sus cortes, en cambio si es un diamante grande, al ser de mas tamaño se pueden visualizar mejor esos cortes. También existe una correlación entre el color y el corte, ya que dependiendo de la calidad del corte puede influir en como se percibe su color