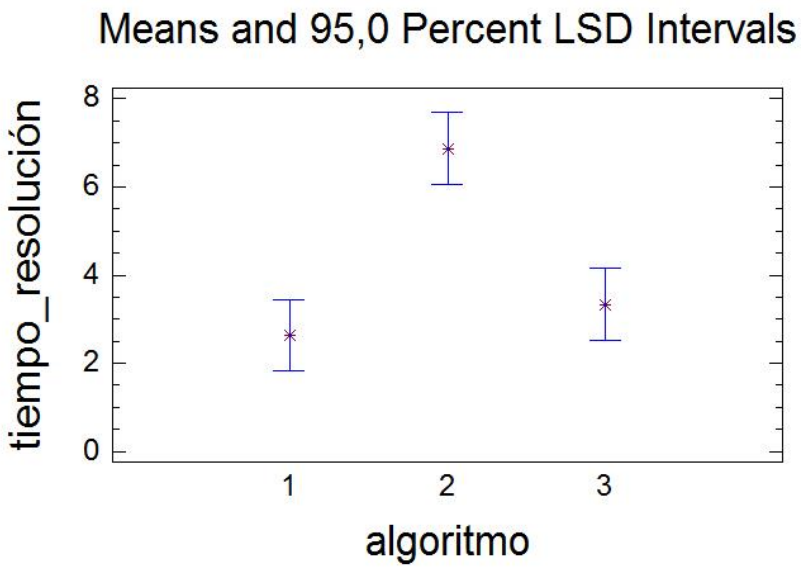
EJERCICIOS U.D.5 INTRODUCCIÓN A LA INFERENCIA ESTADÍSTICA

U.D. 5.3 INTRODUCCIÓN AL ANÁLISIS DE LA VARIANZA

1. Se desea comparar tres algoritmos de inversión de matrices para ver si alguno resulta en promedio más rápido. Para cada algoritmo se han generado 6 matrices y se han constatado los tiempos de resolución que da la tabla siguiente:

Algoritmo 1	3,6 3,9 1,9 3,9 1,7 0,8
Algoritmo 2	6,3 6,5 8,7 7,2 4,6 7,9
Algoritmo 3	3,5 1,8 2,3 5,3 3,6 3,5

a) Estudia con el ANOVA si el factor algoritmo influye significativamente sobre la media del tiempo de resolución.
(SCTotal=88,2511, SCAlgoritmo=61,7911 α=5%).
b) Interpreta el gráfico de intervalos LSD.



SOLUCIÓN:

a) En el enunciado se da la SCTotal=88,2511 que tiene 18-1=17 grados de libertad, y la SCAlgoritmo=61,7911 con 3-1=2 grados de libertad.

La SCResidual=SCTotal-SCAlgoritmo= 88,2511-61,7911=26,46 con 17-2=15 grados de libertad.

La Tabla resumen del ANOVA queda:

O.Variabilidad	Sumas de Cuadrados	Grados de libertad	Cuadrados Medios	F-ratio
Algoritmo	61,7911	2	30,8955	17,51
Residual	26,46	15	1,764	
Total	88,2511	17		

La F de tabla con 2 y 15 grados de libertad para α=5% resulta igual a 3,68. Como F-Ratio=17,51 > F-Tabla, entonces el efecto de algoritmo sobre el tiempo medio de resolución es significativo.

b) En el gráfico de intervalos LSD se aprecia que con el algoritmo 2 el tiempo medio de resolución es significativamente mayor que con los algoritmos 1 y 3. Entre los algoritmos 1 y 3 no hay diferencias significativas (los intervalos se solapan).

2. En un sistema informático con una determinada configuración, se ha medido durante varios días el tiempo medio de respuesta (en segundos) para cuatro niveles de carga (en consultas por minuto), con los siguientes resultados:

	Carga=3	Carga=5	Carga=7	Carga=9
Número de datos	4	4	4	4
Media observada	0,8	1,625	3	6,125

a) Completa la tabla del ANOVA siguiente y analiza si el efecto de la carga sobre el tiempo medio de respuesta es significativo.

Analysis of Variance for Trespuesta

Source	Sum of Squares	Df	Mean Square	F-Ratio
MAIN EFFECTS				
A:carga	65,7825			
RESIDUAL				
TOTAL	66,1575			

b) Interpreta a nivel descriptivo el efecto de la carga.

SOLUCIÓN:

a)La tabla del ANOVA da la suma de cuadrados total y la suma de cuadrados de carga. Faltan los grados de libertad y la suma de cuadrados residual, para poder calcular los cuadrados medios y a partir de ellos la F-Ratio del efecto de carga.

Grados de libertad totales= 16-1=15
Grados de libertad de carga=4-1=3

La SCResidual se obtiene restando a la total la del efecto de carga:
SCResidual= 66,1575-65,7825=0,375 con 15-3=12 grados de libertad

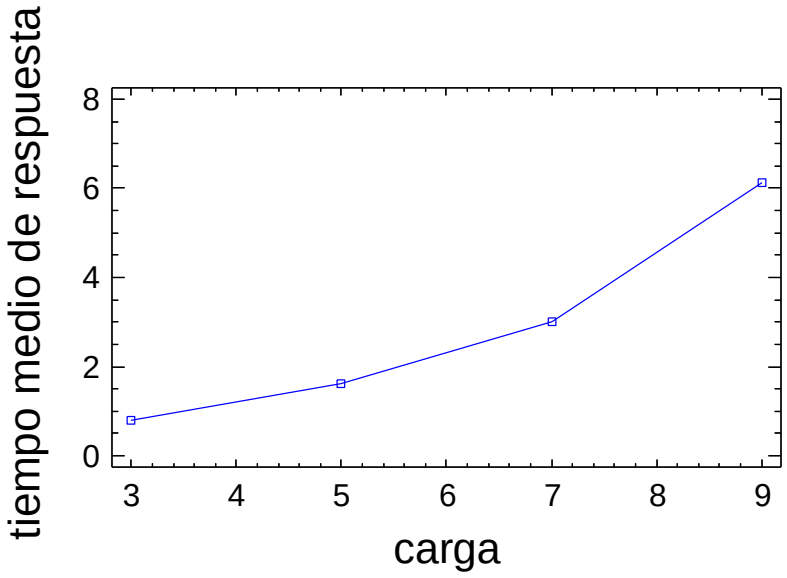
La Tabla del ANOVA con estos cálculos y los CMedios y la F-Ratio resulta:

Analysis of Variance for Trespuesta

Source	Sum of Squares	Df	Mean Square	F-Ratio
MAIN EFFECTS				
A:carga	65,7825	3	21,9275	701,68
RESIDUAL	0,375	12	0,03125	
TOTAL	66,1575	15		

La F-tabla para 3 y 12 grados de libertad con α=5% resulta 3,49, y con α=1% F-tabla=5,95. Como F-Ratio=701,68 >> F(1%) el efecto de la carga cobre el tiempo medio de respuesta es muy significativo estadísticamente.

b) La interpretación descriptiva del efecto de carga se realiza con el gráfico en el que se representan las medias observadas para el tiempo de respuesta con cada nivel de carga. Dicho gráfico es:



Se observa que el efecto de la carga es lineal positivo: a mayor carga mayor tiempo de respuesta medio. También tiene dicho efecto una componente cuadrática positiva: el aumento de tiempo medio de respuesta al aumentar la carga es cada vez mayor.

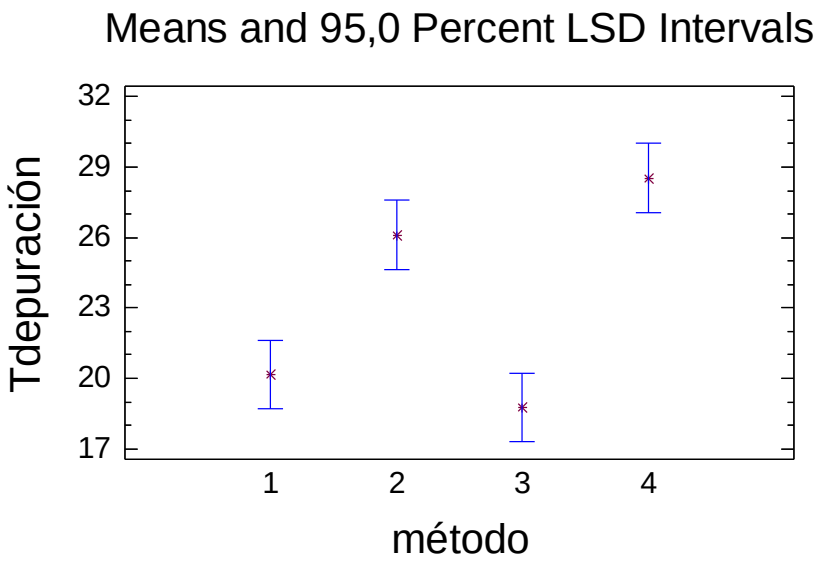
3. Se dispone de cuatro métodos de depuración de programas en un determinado lenguaje. Con el fin de optimizar el tiempo de depuración, se ha probado cada método con 6 programas distintos y se ha calculado la tabla de ANOVA siguiente:

Analysis of Variance for Tdepuración

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:método					
RESIDUAL	119,148				
TOTAL	513,07				

a) Estudia si el factor método influye significativamente sobre el tiempo medio de depuración.

b) Interpreta el gráfico de intervalos LSD siguiente:



SOLUCIÓN:

a)La tabla de ANOVA da las SCResidual y SCTotal. Para calcular la suma de cuadrados del efecto de método de depuración:
SCMétodo=SCTotal-SCResidual=513,07-119,148=393,922
Con 4-1=3 grados de libertad.
Los grados de libertad totales son 24-1=23 y los residuales 23-3=20.

Con estos cálculos y los de las sumas de cuadrados, cuadrados medios y F-Ratio la tabla queda:

Analysis of Variance for Tdepuración

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:método	393,922	3	131,307	22,04	<0,01
RESIDUAL	119,148	20	5,9574		
TOTAL	513,07	23			

La F de tabla para 3 y 20 grados de libertad, con α=5% resulta F-tabla=3,1, y con α=1% F-tabla=4,94. Como F-Ratio=22,04 >>4,94, el efecto de método sobre el tiempo medio de depuración es muy significativo.

b) En la representación gráfica de los cuatro intervalos LSD, se observa que no hay diferencias significativas en el tiempo medio de depuración entre los métodos 1 y 3 (se solapan los intervalos). Con estos dos métodos el tiempo medio de depuración es significativamente menor que con los métodos 2 y 4. Entre los métodos 2 y 4 tampoco hay diferencias significativas.

4. Una empresa está analizando la velocidad de tres tipos de procesadores en función de la potencia del ventilador que incorpora el ordenador. Para ello, ha realizado un experimento considerando estos dos factores (potencia y tipo de procesador), ambos a tres niveles, en el que cada tratamiento se ha ensayado tres veces. La siguiente tabla presenta los resultados medios obtenidos.

		Potencia		
		0,25	0,5	0,75
Tipo de Proces.	A	52,5	56,4	58,0
	B	59,7	52,9	55,0
	C	53,0	57,8	57,4

A partir de este experimento, se ha llevado a cabo un Análisis de la Varianza (ANOVA), parte de cuya tabla se presenta a continuación.

Analysis of Variance for Velocidad

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:Tipo Procesador	0.846667				
B:Potencia	13.8467				
INTERACTIONS					
AB	149.473				
RESIDUAL	8.18				
TOTAL	172.347				

a) Completa la tabla del ANOVA, interpretando de manera detallada los resultados y justificando las conclusiones obtenidas, para un riesgo de primera especie del 5%.

b) Construye e interpreta los gráficos para los efectos que resultan significativos.

SOLUCIÓN:

a) La tabla del ANOVA completa queda de la siguiente forma:

Analysis of Variance for Velocidad

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:Tipo Procesador	0.846667	2	0.42335	0.9325	>0.05
B:Potencia	13.8467	2	6.92335	15.2497	<0.05
INTERACTIONS					
AB	149.473	4	37.36825	82.309	<0.05
RESIDUAL	8.18	18	0.454		
TOTAL	172.347	26			

Con los valores F-ratio de la tabla se puede realizar el contraste F de la significación de cada uno de los tres efectos en estudio.

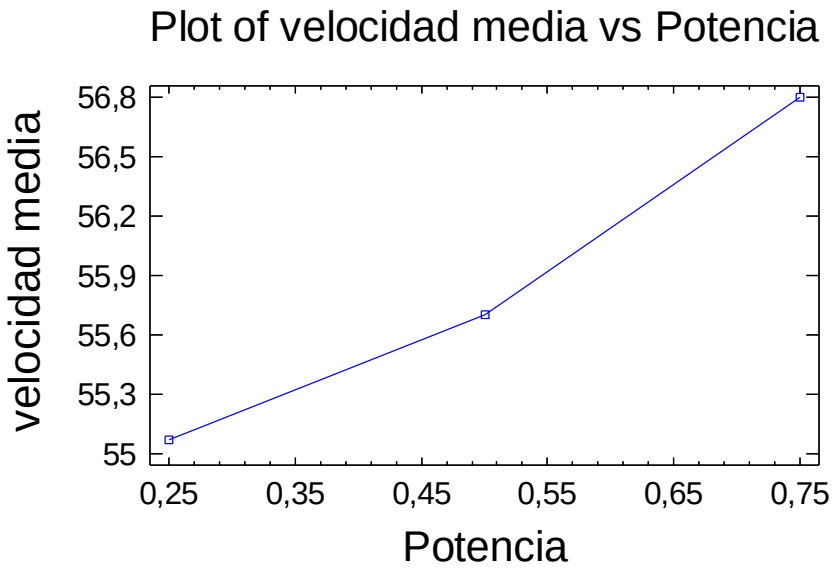
Para el efecto de tipo de procesador, la F-ratio sale menor que la F de tablas con 2 y 18 grados de libertad para un riesgo de primera especie igual al 5%:
F-ratio=0.9325<$F_{2,18}^{\alpha=0.05}$ =3.55⇒ El efecto de tipo de procesador sobre la media de la velocidad no es significativo. Por tanto dicha media no difiere entre ninguno de los tres tipos de procesadores utilizados en el experimento.

Para el efecto de la potencia la F-ratio sale mayor que la F de tablas con 2 y 18 grados de libertad para un riesgo de primera especie igual al 5%:
F-ratio=15.2497 > $F_{2,18}^{\alpha=0.05}$ =3.55⇒ Efecto significativo. Por tanto cuando cambia la potencia cambia significativamente la velocidad media, y por tratarse de un factor cuantitativo dicho cambio podrá ser lineal y/o cuadrático.

Con el efecto de la interacción entre tipo de procesador y potencia la F-ratio también sale mayor que la F de tablas con 4 y 18 grados de libertad para un riesgo de primera especie igual al 5%:
F-ratio=82.309 > $F_{4,18}^{\alpha=0.05}$ =2.93⇒ Efecto significativo. Por tanto el efecto de la potencia sobre la velocidad media cambia significativamente según el tipo de procesador. También se puede interpretar de forma equivalente de la siguiente manera: el efecto del tipo de procesador sobre la velocidad media difiere significativamente según la potencia.

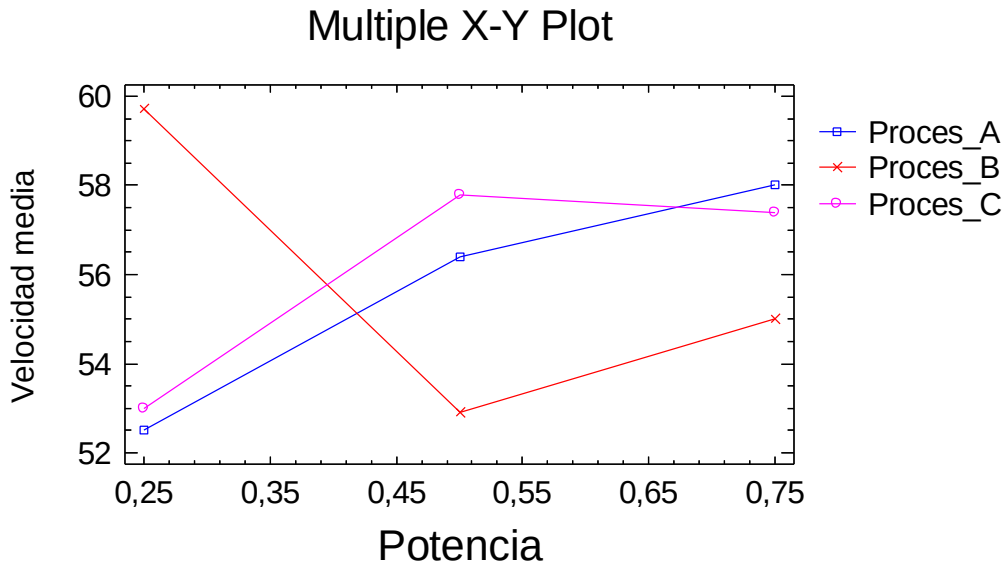
b) Los efectos que hay que interpretar son el de potencia y el de la interacción procesador*potencia.

El gráfico de las medias de la velocidad observada con cada potencia es:



Se observa que cuando aumenta la potencia aumenta la velocidad media: hay un efecto de potencia lineal positivo. Hay también una ligera curvatura positiva ya que la velocidad media aumenta más al cambiar de potencia 0,5 a 0,75, que entre potencia 0,25 y 0,5.

Para interpretar la interacción se representa el gráfico con las velocidades medias observadas para cada combinación de los dos factores procesador y potencia:



Se observa que con procesadores A y B el efecto de la potencia sobre la velocidad media, es lineal positivo y cuadrático negativo. Con procesador A el efecto cuadrático es menor que con procesador C. Sin embargo con el procesador C el efecto de la potencia sobre la velocidad media es lineal negativo y cuadrático positivo.

5. La velocidad de ejecución de un programa varía en función del número de núcleos utilizados (2, 4 o 6) y del tipo de procesador utilizado (A, B o C). La tabla adjunta muestra las medias de los resultados de un Diseño de Experimentos con tres pruebas para cada tratamiento, cuyo fin es optimizar dicha velocidad.

		Núcleos		
		2	**4**	**6**
Tipo procesador	**A**	43,33	60,67	64,67
	B	53,67	52,67	55,33
	C	51,83	57,5	56,67

La tabla siguiente recoge algunas sumas de cuadrados obtenidas para realizar un ANOVA:

```
Analysis of Variance for Velocidad
-------------------------------------------------------------------------------
Source               Sum of Squares     Df    Mean Square    F-Ratio   P-Value
-------------------------------------------------------------------------------
MAIN EFFECTS
 A:Procesador              24,963
 B:Núcleos                430,907

INTERACTIONS
 AB

RESIDUAL                  149,853
-------------------------------------------------------------------------------
TOTAL                    1013,43
-------------------------------------------------------------------------------
```

a) Completa la tabla del ANOVA y explica detalladamente qué efectos son estadísticamente significativos (α=5%).

b) Construye e interpreta los gráficos de medias para los efectos que en el apartado anterior han resultado significativos.

SOLUCIÓN:

a) En la Tabla del ANOVA del enunciado, faltan la suma de la interacción procesadorxnúcleos, los grados de libertad totales, de los efectos y residuales, los cuadrados medios y las F para los contrastes de significación. La suma de cuadrados de la interacción se puede obtener restando a la total las de los efectos de procesador, núcleos y la residual:

SCprocesadorxnúcleos=SCTotal-SCprocesador-SCnúcleos-SCresidual=
= 1013,43-24,963-430,907-149,853=407,707

Como hay 27 datos (3 repeticiones por 9 tratamientos), los grados de libertad totales son 27-1=26
Como hay tres tipos de procesador, los grados de libertad de este factor serán 3-1=2. El factor número de núcleos tiene tres niveles, por lo que sus grados de libertad son 3-1=2.

La interacción procesadorxnúcleos tendrá por tanto, 2x2=4 grados de libertad. Los grados de libertad residuales se obtendrán restando a los totales los de los efectos en estudio:
Grados de libertad residuales= 26-2-2-4=18.

Los cuadrados medios se obtienen dividiendo las sumas de cuadrados por los grados de libertad. Las F para los contrastes de significación de los efectos se calculan dividiendo los cuadrados medios de los efectos por el cuadrado medio residual. Con todos estos cálculos la tabla del ANOVA queda:

Analysis of Variance for Velocidad

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:Procesador	24,963	2	12,4815	1,499	>0,05
B:Núcleos	430,907	2	215,4535	25,8796	<0,05
INTERACTIONS					
AB	407,707	4	101,926	12,24	<0,05
RESIDUAL	149,853	18	8,3252		
TOTAL	1013,43	26			

Para estudiar la significación de los efectos se comparan las F calculadas con las de las tablas con los grados de libertad del efecto y residuales. Así, la F calculada para el efecto de procesador es F=1,499 < F tablas con 2 y 18 grados de libertad para un riesgo de primera especie del 5% (F tablas=3,55). Por tanto el efecto de procesador no es significativo. Esto implica que la velocidad media de ejecución del programa no difiere significativamente entre los tres tipos de procesadores. El P-value de la F calculada para este efecto será mayor que el riesgo de primera especie 0,05.

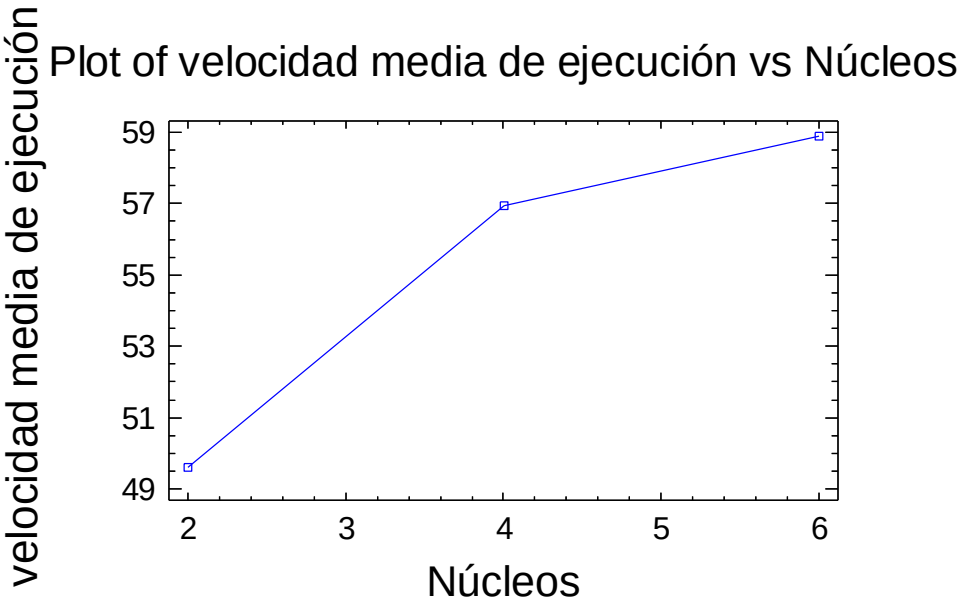
La F para número de núcleos es 25,8796, mayor que la F tablas (la misma que antes 3,55). Por tanto la velocidad media de ejecución difiere significativamente según el número de núcleos. Como número de núcleos es un factor cuantitativo, que su efecto sea significativo indica que tiene un efecto o lineal y/o cuadrático significativo sobre la velocidad media de ejecución. El P-value para la F de este efecto será menor que el riesgo de primera especie 0,05. La interacción procesadorxnúcleos también es significativa, ya que su F calculada=12,24 es mayor que la F de tablas con 4 y 18 grados de libertad y riesgo de primera especie 5% (F-tablas=2,93). Por tanto el efecto de número de núcleos sobre la velocidad media de ejecución difiere significativamente según el tipo de procesador que se utilice. El P-value de esta interacción será menor que el riesgo de primera especie 0,05.

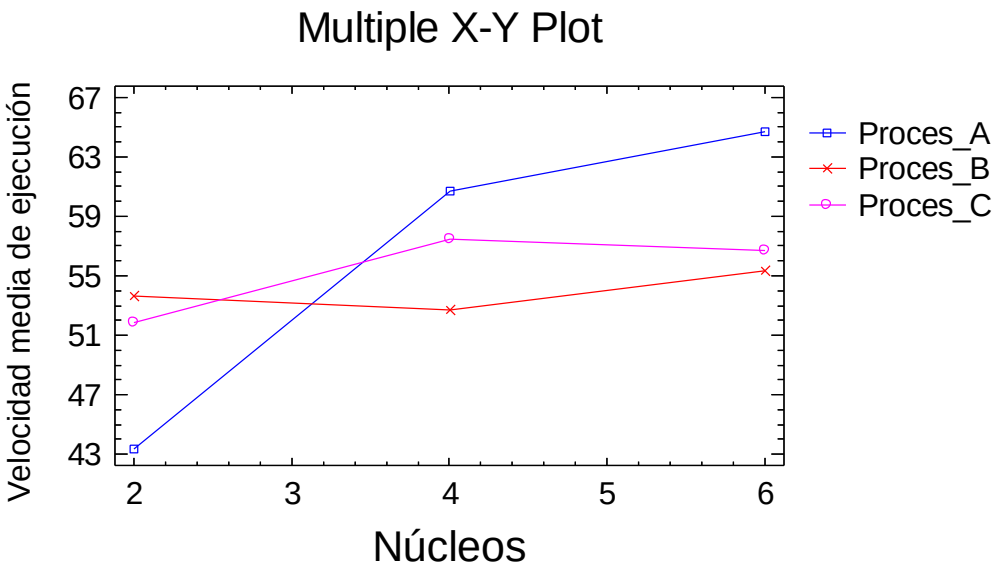
b) Han resultados significativos los efectos del factor cuantitativo núcleos y la interacción entre procesador y núcleos.

Para interpretar el efecto de núcleos se representan las velocidades medias observadas para cada nivel de dicho factor. El gráfico resultante se muestra en la página siguiente. Se observa que cuando aumenta núcleos la velocidad media de ejecución aumenta de

forma lineal y cuadrática. El efecto cuadrático es negativo puesto que entre núcleos 4 y 6 hay un menor aumento que entre núcleos 2 y 4.

Para interpretar el efecto de la interacción procesador*núcleos se representan las 9 medias observadas en las 9 combinaciones entre estos dos factores. El gráfico se recoge también en la página siguiente. Se observa que el efecto de núcleos sobre la velocidad media de ejecución, es mayor con procesador A, teniendo una componente lineal positiva y cuadrática negativa. Con procesador B el efecto lineal y cuadrático es muy pequeño, situación que se repite con procesador C. Con este último prácticamente no cambia la velocidad media de ejecución cuando aumenta núcleos.





6. Una empresa informática debe decidir la configuración para los discos duros que instala en los equipos informáticos que vende. Los discos pueden ser configurados según tres tipos de distribución de ficheros (A, B y C) y reservando inicialmente un espacio para los archivos del sistema de 20, 40 ó 60 Mb. Para decidir la distribución y el espacio inicial reservado, realiza un experimento con dos repeticiones recogiendo el tiempo medio de acceso (en segundos), para cada una de las nueve combinaciones posibles. Los resultados medios correspondientes a las dos repeticiones realizadas para cada tratamiento son los siguientes:

	20	**40**	**60**
A	20	22	24,5
B	25	28	32
C	21	25	24

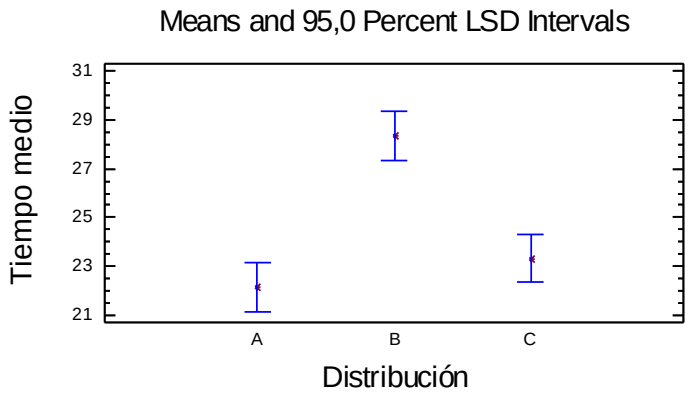
Los resultados del ANOVA aparecen en la siguiente tabla de manera parcial

```
Analysis of Variance for Tiempo medio
-------------------------------------------------------------------------------
Source                 Sum of Squares     Df    Mean Square    F-Ratio    F-tabla
-------------------------------------------------------------------------------
MAIN EFFECTS
 A:Espacio reservad          71,4444
 B:Distribución             128,78
INTERACTIONS
 AB                          15,55

RESIDUAL
-------------------------------------------------------------------------------
TOTAL                        236,278
-------------------------------------------------------------------------------
```

a) Completa la tabla ANOVA anterior, estudiando la significación de los efectos simples y de la interacción doble y explicando los resultados para un error de 1ª especie α=5%.

b) Interpreta el siguiente gráfico de intervalos LSD para el factor distribución:



c) Interpreta a nivel descriptivo con el gráfico de medias el efecto del factor espacio reservado.

d) En un ANOVA con un factor: ¿Qué mide la Suma de Cuadrados Total? ¿Y la del factor? ¿Y la residual?

SOLUCIÓN:

La $SC_{espacio}$ tiene 3-1=2 grados de libertad. La $SC_{distribución}$ tiene 3-1= 2 grados de libertad. La $SC_{expaico\ x\ distrib}$ tiene (3-1)x(3-1)=4 grados de libertad. La SC_{total} tiene 18-1=17 grados de libertad.

$SC_{residual}$ = SCtotal- $SC_{espacio\ reservado}$ - $SC_{distribución}$ - $SC_{espacio\ reservado\ x\ distribución}$ = 20,5
Con 17-2-2-4=9 grados de libertad

Analysis of Variance for Tiempo medio

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:Espacio reservad	71,4444	2	35,7222	15,68	<0,05
B:Distribución	128,78	2	64,39	28,27	<0,05
INTERACTIONS					
AB	15,55	4	3,8875	1,71	>0,05
RESIDUAL	20,5	9	2,27778		
TOTAL	236,278	17			

$F_{2,9}^{\alpha=0,05}$= 4, 26 $F_{4,9}^{\alpha=0,05}$= 3,63

F-ratio de espacio reservado= 15,68 > 4,26 Efecto significativo

F-ratio de distribución = 28,27 > 4,26 Efecto significativo

F-ratio interacción = 1,71 < 3,63 Efecto no significativo

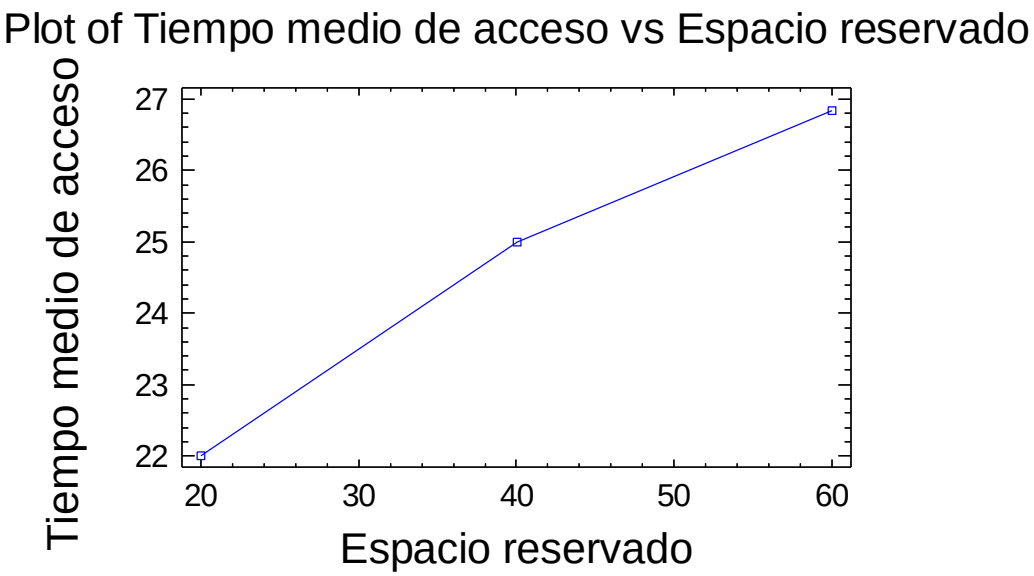
Que el espacio reservado aparezca como significativo en el estudio indica que el tiempo medio de acceso varía con dicho espacio reservado para memoria (no sabemos todavía si lo hace de manera lineal y/o cuadrática).

Que la distribución de ficheros aparezca como significativa indica que existe al menos un tipo de distribución de ficheros cuyo tiempo medio de acceso es diferente a los otros dos.

Que la interacción del espacio reservado con la distribución de ficheros aparezca como no significativa, indica que el tiempo medio de acceso evoluciona según el espacio reservado para memoria de la misma manera con los tres tipos de distribución de ficheros.

b) Se observa en el gráfico de intervalos LSD de distribución que el tiempo medio de acceso difiere significativamente entre las distribuciones tipo A y B, siendo menor con tipo A. También difieren B y C, pero no A y C (en este tercer caso aparecen solapados los intervalos). La distribución con un tiempo de acceso medio significativamente superior es la B.

c) El gráfico con las medias observadas para cada nivel de espacio reservado es:



Se observa que cuando aumenta el espacio reservado aumenta el tiempo medio de acceso de forma prácticamente sólo lineal.

d) La suma de Cuadrados Total está indicando la variabilidad presente en el conjunto global de los datos. Para ello, calcula la suma de los cuadrados de las desviaciones de cada uno de los datos respecto de la media general.

La Suma de Cuadrados del Factor permite estimar la variabilidad que hay entre los distintos niveles del factor. Dicha Suma de Cuadrados se calcula como la suma de los cuadrados de las desviaciones de la media que hay con cada nivel del factor respecto de la media general, ponderada cada desviación por el número de datos correspondientes a cada nivel.

La Suma de Cuadrados Residual se utiliza para estimar la variabilidad que hay en los datos con cada nivel del factor. Se supone que dicha variabilidad no depende del nivel del factor. Se calcula como la suma de los cuadrados de las desviaciones de cada dato respecto de la media que hay en cada nivel.

7. La siguiente tabla recoge los resultados de un estudio para investigar la influencia del proceso de pegado en la fabricación (procedimiento A, B ó C) y de la temperatura (10, 30 y 50 ºC) sobre la fiabilidad de unos módulos electrónicos. La variable resultado es una intensidad de corriente medida en un determinado punto del módulo (valores elevados de la misma indican altas cotas de fiabilidad).
Los resultados medios del experimento, repetido dos veces en cada tratamiento, son los que se muestran en la siguiente tabla:

PROCEDIMIENTO DE PEGADO	T 10ºC	T 30ºC	T 50ºC
Proced. A	17	21	25,5
Proced. B	20,5	26	31,5
Proced. C	11	15	10,5

Analysis of Variance for Intensidad

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:Pegado	591.444				
B:Temperatura	127.444				
INTERACTIONS					
AB	90.2222				
RESIDUAL	26.0				
TOTAL	835.111				

a) Completa el cuadro resumen del ANOVA, indicando los efectos que resultan o no significativos e interpretando detalladamente las conclusiones obtenidas (α=5%).

b) ¿La evolución de la intensidad con la temperatura es la misma en todos los procedimientos de pegado? ¿Por qué? Interpreta gráficamente a nivel descriptivo este efecto.

c) Si el efecto de un factor no existe en una población, ¿el cuadrado medio correspondiente será en promedio mayor, menor o igual que el cuadrado medio residual? ¿Cuál será la probabilidad de que aparezca significativo en el ANOVA?

SOLUCIÓN:

a)

Analysis of Variance for Intensidad

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:Pegado	591.444	2	295.722	102.37	<0.05
B:Temperatura	127.444	2	63.7222	22.06	<0.05
INTERACTIONS					
AB	90.2222	4	22.5556	7.81	<0.05
RESIDUAL	26.0	9	2.88889		
TOTAL	835.111	17			

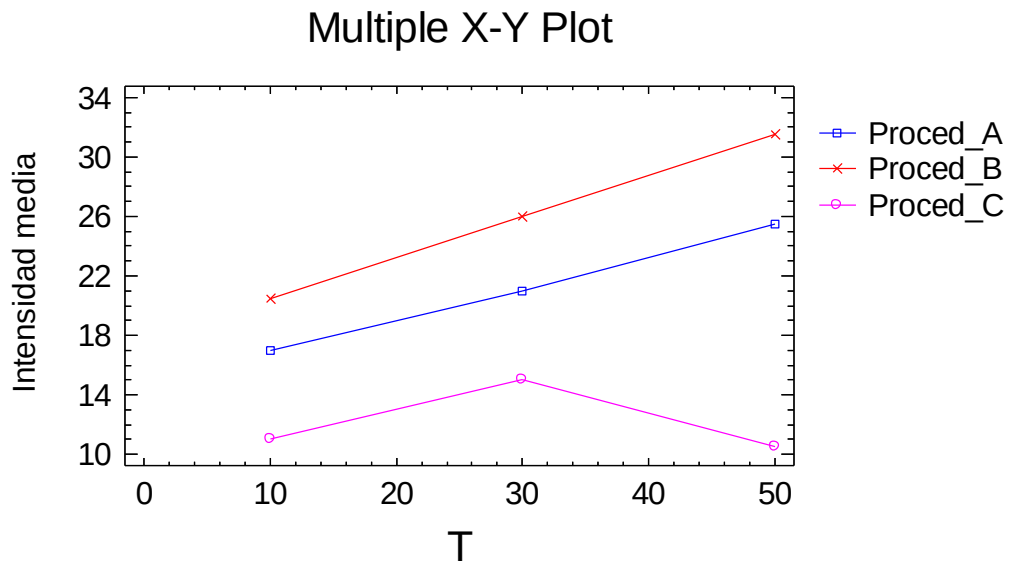
El efecto Pegado aparece significativo (F-Ratio=102,37 > F-Tabla=$F_{2,9}^{\alpha=0,05}$=4,26), lo cual implica que al menos uno de los tres procesos de pegado presenta intensidades medias diferentes a los otros dos.

Como la Temperatura también aparece significativa (F-Ratio = 22,06 > F-Tabla = $F_{2,9}^{\alpha=0,05}$ = 4,26), la intensidad media evoluciona de forma significativa con la Temperatura. Como temperatura es un factor cuantitativo dicha evolución podrá ser lineal y/o cuadrática.

Por último, que la interacción aparezca significativa (F-Ratio=7,81 > F-Tabla = $F_{4,9}^{\alpha=0,05}$= 3,63) quiere decir que la evolución de la intensidad media con la temperatura es diferente en al menos uno de los tres procesos de pegado.

b) La evolución de la intensidad con la temperatura es diferente en al menos uno de los tres procesos de pegado porque la interacción ha resultado significativa. El gráfico que permite interpretar este efecto es el de las medias de cada combinación de los dos factores. Se recoge en la figura siguiente.

Se observa que con los procedimiento A y B el efecto de la temperatura sobre la intensidad media es sólo lineal positivo. Sin embargo con el procedimiento C el efecto de la temperatura sobre la intensidad media es ligeramente cuadrático negativo, y el cambio que experimenta con dicho procedimiento la intensidad media es pequeño.



c) Si el efecto de un factor no existe en una población, el Cuadrado Medio del mismo debería ser en promedio igual al residual.

La probabilidad de que salga significativo en el ANOVA es la probabilidad de decir que es significativo cuando no lo es, es decir, es el riesgo de primera especie α elegido, generalmente el 5% en nuestro caso.

U.D. 5.4 INTRODUCCIÓN A LA REGRESIÓN LINEAL

1. Con el fin de estudiar el rendimiento de una base de datos se han medido a lo largo de un mes la *carga media diaria* del sistema como el nº de consultas por minuto (**X**) y *el tiempo medio de respuesta* en segundos (**Y**).
A partir de los 100 datos reunidos durante el mes del estudio se ha calculado el coeficiente de correlación lineal entre X e Y, resultando su valor **0,85**

a) A la vista del valor de $\mathbf{r_{XY}}$ indicar, justificando la respuesta, cuáles de las siguientes afirmaciones son ciertas:

a.1) La relación lineal entre X e Y es positiva y débil.
a.2) La relación lineal entre X e Y es negativa y fuerte.
a.3) No existe porque $\mathbf{r_{XY}}$ es muy próximo a 0.
a.4) La relación lineal entre X e Y es positiva y fuerte.

b) Indicar y justificar qué forma tendrá, aproximadamente, el Diagrama de Dispersión para X e Y, dado el valor obtenido para $\mathbf{r_{XY}}$ de entre los que se muestran a continuación (ver página siguiente).

SOLUCIÓN:

a) El valor del coeficiente de correlación lineal entre X e Y es $r_{XY} = 0{,}85$, lo cual indica una relación lineal positiva y fuerte (porque está bastante próximo a 1). Por tanto, la única afirmación correcta es la **a.4)**.

b) De los cuatro diagramas de dispersión mostrados en la página siguiente, el único que corresponde a una relación lineal creciente/positiva y fuerte es el diagrama **b.4)**.

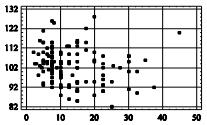
   El diagrama **b.1)** corresponde a un par de variables sin apenas relación lineal (ni de ningún otro tipo), ya que la nube de puntos presenta una forma indefinida.

   El diagrama **b.2)** muestra una relación lineal decreciente/negativa débil o intermedia (nube de puntos descendente y algo dispersa).

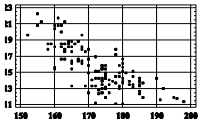
   El diagrama **b.3)** corresponde a una relación no lineal, ya que la nube de puntos es ligeramente curva; el coeficiente de correlación lineal para dicha muestra sería positivo, pero no cercano a 1.

   Por el contrario, el diagrama **b.4)** presenta una nube puntos con forma de recta, creciente y muy concentrada, cosa que indica una relación lineal creciente/positiva y fuerte, lo cual coincide con el coeficiente de correlación $r_{XY} = 0{,}85$.

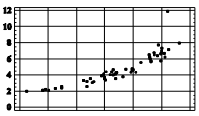
b.1)



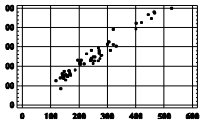
b.2)



b.3)



b.4)



2. Cierta empresa de telemarketing quiere saber el efecto que tienen las llamadas directas en sus ventas. La media de facturación de los 10 últimos meses ha sido de 12,2 miles de euros para una media de 5,5 centenares de llamadas. Estadísticamente saben que las desviaciones típicas son de 1,2 para la facturación y de 3 para las llamadas, con una covarianza de 3,4.

a) Calcular la recta de regresión que nos permita saber la facturación de un mes en función de las llamadas efectuadas.

b) Hallar la facturación prevista para un mes en el que se efectúen 7,2 centenares de llamadas.

c) Averiguar en qué grado, en tanto por cien, la variabilidad de la facturación está explicada por el número de llamadas.

SOLUCIÓN:

a) E(facturación/llamadas) = a + b * Llamadas.

$$b = r * \frac{S_{facturación}}{S_{llamadas}} = \frac{Cov}{S_{facturación} * S_{llamadas}} * \frac{S_{facturación}}{S_{llamadas}} = \frac{3,4}{(3)^2} \approx 0,38$$

$a = \bar{y} - b\bar{x} = 12{,}2 - 0{,}38 * 5{,}5 = 10{,}11$

Facturación media/llamadas = 10,11 + 0,38 * Llamadas

b) E(Facturación/llamadas=7,2) = 10,11 + 0,38 * 7,2 = 12,846

c) $R^2 = r^2 = \left(\frac{Cov}{S_{facturación} * S_{llamadas}} \right)^2 = \left(\frac{3{,}4}{1{,}2 * 3} \right)^2 = 0{,}892 \Rightarrow 89{,}2\%$

Por lo que el 89,2 % de la variabilidad observada en la facturación es explicado por el efecto lineal de las llamadas.

3. Se tiene la siguiente matriz de varianzas-covarianzas relativa a la cantidad invertida y el beneficio obtenido. Sabiendo que si no se invierte nada el beneficio medio es de 10 millones de euros, hallar cual sería el beneficio esperado para una inversión de 4,5 millones de euros. ¿Hasta que punto las variaciones de beneficio están asociadas a las variaciones de inversión?

Covariances

	Beneficio	Inversión
Beneficio	1,40804 (10)	3,38127 (10)
Inversión	3,38127 (10)	9,16667 (10)

SOLUCIÓN:

E(Beneficio/inversión) = a + b * Inversión

$b = r * \frac{S_{beneficio}}{S_{inversión}} = \frac{Cov}{S_{beneficio} * S_{inversión}} * \frac{S_{beneficio}}{S_{inversión}} = \frac{3{,}38}{9{,}17} \approx 0{,}37$

Sabemos que para Inversión = 0 el Beneficio medio = 10, luego:
$10 = a + 0{,}37 * 0 \Rightarrow a = 10$

E(Beneficio/inversión) = 10 + 0,37 * Inversión

Para una inversión de 4,5 millones, el beneficio medio sería de:
E(Beneficio/inversión=4,5) = 10 + 0,37 * 4,5 = 11,665 millones.

Las variaciones del beneficio están asociadas a las variaciones de la inversión según el porcentaje:

$R^2 = r^2 = \left(\frac{Cov}{S_{facturación} * S_{llamadas}} \right)^2 = \left(\frac{3{,}38}{\sqrt{1{,}41 * 9{,}17}} \right)^2 \approx 0{,}884 \Rightarrow 88{,}4\%$

4. En una red de ordenadores se ha realizado un estudio de regresión, comprobándose que la relación entre la carga del sistema y el tiempo de respuesta de cualquier consulta se ajusta mediante una recta de TIEMPO_RESPUESTA en función de la CARGA_SISTEMA con un coeficiente de correlación de 0.9, y se sabe que cuando la carga del sistema es de 6 trabajos, el tiempo medio de respuesta de cualquier consulta fluctúa en el 95% de los casos entre 15 y 35 segundos. ¿Cuánto vale la desviación típica de la variable TIEMPO_RESPUESTA?

SOLUCIÓN:

La recta será
E(TIEMPO_RESPUESTA/CARGA_SISTEMA)= a + b CARGA_SISTEMA.

Se indica que cuando la CARGA_SISTEMA=6, el TIEMPO_RESPUESTA fluctúa en el 95% de los casos entre 15 y 35 segundos. Dicho intervalo se calcula, teniendo en cuenta la hipótesis de normalidad de los datos, con la expresión:
$E(TIEMPO\_RESPUESTA/CARGA\_SISTEMA=6)\pm 2\, s_{residual}$.

Por tanto $35-15=20=4\, s_{residual} \Rightarrow s_{residual}=20/4=5$

Como $s^2_{residual}= s^2_{TIEMPO\_RESPUESTA}(1-r^2) \Rightarrow s_{TIEMPO\_RESPUESTA} = \sqrt{\frac{5^2}{(1-0{,}9^2)}}=11{,}47$

5. En un estudio sobre redes de interconexión, se realizaron 20 simulaciones de cierto algoritmo de encaminamiento de mensajes. En cada simulación se generaron 5000 mensajes, y se trabajó con una tasa de generación de mensajes (variable TGM), midiéndose la latencia de éstos en la red (variable LATENCIA). La tasa de generación de mensajes se varió de una simulación a otra. Seguidamente se definen las dos variables manejadas:

LATENCIA es el tiempo medio transcurrido en cada simulación desde que los mensajes son generados en el nodo origen hasta que llegan al nodo destino, expresada en ciclos de reloj.

TGM (Tasa de generación de mensajes): mensajes generados por unidad de tiempo en cada nodo de la red, expresada en mensajes/ciclo/nodo.

Se realizó un análisis de regresión con los datos, obteniendo la siguiente información:

Model fitting results for LATENCIA

Independent variable	coefficient	std. error	t-value	sig. Level
CONSTANT	71,77	8,7547	8,1978	0,0000
TGM	1242,02	131,4050	9,4519	0,0000

Analysis of Variance for the Full Regression

Source	Sum of Squares	DF	Mean Square	F-Ratio	P-value
Model	11361,70	1	11361,70	89,338	0,0000
Error	2289,18	18	127,18		
Total (Corr.)	13650,88	19			

R-squared = 0,8323
Stnd. Error of est. = 11,28

¿Entre qué límites fluctuará en el 95% de los casos la latencia media de 5000 mensajes inyectados en la red con una tasa de generación de 0,07 mensajes/ciclo/nodo?

SOLUCIÓN:

Media de (LATENCIA/TGM=0,07)= 71,77 +1242,02 x 0,07 = 158,71
Desviación típica de (LATENCIA/TGM=0,07)= $\sqrt{CM_{residual}} = \sqrt{127{,}18} = 11{,}28$

Límites del 95% de probabilidad 158,71 ± 2 x 11,28 =[136,15, 181,27]

6. Un multicomputador concurrente está constituido por 512 procesadores interconectados con una topología de hipercubo. Los procesadores generan mensajes cuya longitud (expresada en bytes) fluctúa aleatoriamente, y para que los que se puede medir el retardo, tiempo que tardan los mensajes en llegar desde el procesador que los genera hasta su destino (expresado en ciclos de reloj).

a) ¿Qué variable bidimensional se puede definir en el contexto anterior? ¿Sobre qué población estaría definida?

A partir de un conjunto de valores observados se ha calculado la siguiente recta de regresión del retardo respecto a la longitud del mensaje. (Se adjunta también el ANOVA obtenido).

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: RETARDO
Independent variable: LONGITUD
------------------------------------------------------------------------------
                                Standard          T
Parameter          Estimate       Error       Statistic     P-Value
------------------------------------------------------------------------------
Intercept          31.35861      68.1381       0.4602        0,6424
Slope               2.925091      0.378888     7.7204        0,0000
------------------------------------------------------------------------------
Analysis of Variance
------------------------------------------------------------------------------
Source             Sum of Squares      Df      Mean Square  F-Ratio   P-Value
------------------------------------------------------------------------------
Model                145344.43          1       145344.43    59.6040   0,0000
Residual             282866.23        116         2438.50
------------------------------------------------------------------------------
Total (Corr.) 1698,98 117
R-squared = 33.9423 percent
R-squared (adjusted for d.f.) = 33.3728 percent
Standard Error of Est. = 49.3812
```

b) ¿Qué interpretación tiene el valor R^2=33.9%? ¿Por qué supones que se ha obtenido este valor relativamente bajo?

c) ¿Qué porcentaje de los mensajes de 64 bytes de longitud tarda más de 150 ciclos de reloj en llegar a su destino?

SOLUCIÓN:

a) Variable bidimensional : (longitud, retardo).
Población: mensajes que se puedan generar.

b) R^2=33.9 %. Es el coeficiente de correlación al cuadrado, indica que el 33,9% de la variabilidad observada en el retardo es explicada por el efecto lineal de la longitud de los mensajes. Un valor tan bajo indica que además del efecto lineal de la longitud de mensajes hay otros efectos que también influyen de forma significativa sobre el retardo medio.

c) Para longitud=64 bytes, el retardo medio es = 31,36 + 2,925 x 64 = 218,56. La desviación típica del retardo cuando la longitud es 64 se calcula con la raíz cuadrada del $CM_{residual}$=49,38 Este valor también aparece en la salida de Statgraphics en el campo `Standard Error of Est.`).
El porcentaje de mensajes de 64 bytes de longitud que tarda más de 150 ciclos de reloj en llegar a su destino, se calcula suponiendo distribución normal:
P((retardo/longitud=64) > 150)=P(N(218,56, 49,38)>150)=P(N(0,1)>$\frac{150-218,56}{49,38}$)=
= P(N(0,1) > -1,39) = 1- P(N(0,1) < -1,39) = 1-0,0823 = 0,9177.

Por tanto el porcentaje es 91,77%.

7. Se ha realizado una encuesta entre los alumnos en la que se les preguntaba por su peso (en Kg) y por su estatura (cm). Una vez introducidos los datos en el STATGRAPHICS, se calculó la recta de regresión para relacionar el peso de las chicas con su estatura obteniéndose los siguientes resultados:

Regression Analysis - Linear model: Y = a + b*X

Dependent variable: PESO
Independent variable: ESTATURA-150
Selection variable: SEXO="chicas"

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	45,002	1,92181	23,4165	0,0000
Slope	0,767583	0,132065	5,81218	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	777,89	1	777,89	33,78	0,0000
Residual	921,086	40	23,0271		
Total (Corr.)	1698,98	41			

Correlation Coefficient = 0,676652
R-squared = 45,7858 percent
R-squared (adjusted for d.f.) = 44,4305 percent
Standard Error of Est. = 4,79866

A partir de estos datos, ¿qué tanto por cien de las chicas que miden 167 cm pesarán entre 60 y 70 Kg.?

SOLUCIÓN:

El peso de las chicas que miden 167 cm se distribuye normalmente con media:
E(Peso/estatura=167) = 45 + 0,77 (167-150) = 58,09 Kg
y desviación típica 4,799 (en la salida Standard Error of Est. o raíz cuadrada del cuadrado medio residual de la tabla de ANOVA).

Por tanto:
P(60<(peso/estatura=167)<70)=P(60<N(58,09,4,799)<70) =
$=P(\frac{60-58,09}{4,799} < N(0,1) < \frac{70-58,09}{4,799}) = P(0,39< N(0,1) < 2,48)=$
= P(N(0,1) < 2,48) - P(N(0,1) < 0,39) = 1 - P(N(0,1)>2,48) – 1 + P(N(0,1) > 0,39) =
= -0,0066 + 0,3483 = 0,3417

El porcentaje es 34,17%