# Ejercicios Resueltos de Estadística:
# Tema 2: Descripciones bivariantes y regresión

**1. En un estudio de la Seguridad e Higiene en el Trabajo se contrastó la incidencia del tabaquismo en la gravedad de los accidentes laborales. Considerando una gradación de Muy fumador hasta No fumador como media del tabaquismo, y una gradación de Muy grave a Leve en el tipo de accidente. Se extrajo una muestra de 525 individuos que habían sufrido un accidente laboral. Los resultados se presentan en la siguiente tabla de contingencia(tabla de doble entrada):**

| | Muy Grave | Grave | Lesiones Med | Leves |
|---|---|---|---|---|
| Muy Fumador | 20 | 10 | 10 | 30 |
| Fumador | 30 | 40 | 20 | 50 |
| Fumador Esporádico | 10 | 60 | 80 | 60 |
| No Fumador | 5 | 20 | 30 | 50 |

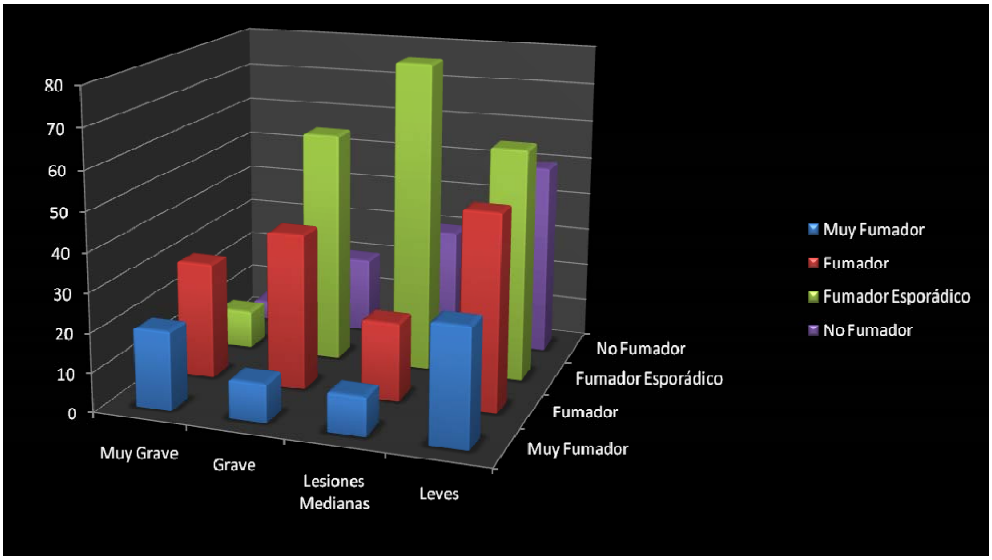
**Se pide:**

1. **Representar los datos anteriores gráficamente**
2. **Calcular las distribuciones marginales para cada una de las variables de estudio.**
3. **Construir una tabla de distribución de frecuencias porcentuales donde aparezcan las distribuciones de la variable de tipo de Lesión condicionada a cada una de las variables del Fumador.**
4. **Estudiar si las variables están asociadas o no por medio de una medida descriptiva. Realizar un análisis gráfico y comentar los resultados.**

---

**SOLUCIÓN:**

a)




b) Se obtiene a partir de la tabla de doble entrada sumando las frecuencias y las filas, o bien por columnas según el caso.

| Marg. Tabaquismo | FREC. | Marg.Accid. Lab. | FREC. |
|---|---|---|---|
| Muy fumador | 70 | Muy grave | 65 |
| Fumador | 140 | Grave | 130 |
| Fumador Esporádico | 210 | Lesión media | 140 |
| No fumador | 105 | Leve | 190 |
| | 525 | | 525 |

c) La distribución de una variable condicionada a que otra variable tome un determinado valor de la distribución de frecuencias de la variable cuando mantenemos fijo el valor condicionante de otra variable.

| | Muy Grave | Grave | Lesión Med. | Leve | |
|---|---|---|---|---|---|
| Muy Fum. | 28.57 | 14.29 | 14.29 | 42.86 | 100% |
| Fumador | 21.43 | 28.57 | 14.29 | 35.71 | 100% |
| Fum.Espor. | 4.76 | 28.57 | 38.10 | 28.57 | 100% |
| No Fum. | 4.76 | 19.05 | 28.57 | 47.62 | 100% |
| Marg.Lesión | 12.38 | 24.76 | 26.67 | 36.19 | 100% |

Como ejemplo del cálculo de la distribución porcentual del Tipo de lesión condicionado al individuo sea Muy Fumador se realizará dividiendo cada una de las frecuencias de la fila Fumador entre el número total de Muy Fumadores y después multiplicariamos como ((20/70)*100=28.57; (10/70)*100=14.29,…).

d) (Este apartado lo vamos a realizar sobre una misma tabla)

La medida descriptiva de la asociación entre las variables viene dada a través de la medida que indica la distancia relativa que existe entre la tabla de frecuencias observadas en la tabla de frecuencias esperadas si las variables fueran independientes. La expresión para las frecuencias esperadas es la siguiente:

$$E_{ij} = \frac{F_i \times C_j}{n}$$

Donde E es la frecuencia esperada en la celda (i,j), F es la suma de las frecuencias de f y C es la suma de las frecuencias de la fila j.

La distancia relativa al cuadrado que existe entre una celda de la tabla de frecuencias observadas es la misma celda de la tabla de esperadas viene dada por:

$$z_{ij}^2 = \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Y la suma de todas ellas recibe el nombre de $x^2$(ji-cuadrado).

Por otra parte podemos estudiar cuáles son los pares de categorías que influyen en mayor medida en la existencia de la asociación. Este lo realizaremos por medio de análisis gráfico atendiendo al siguiente criterio:

$[z_{ij}]<1.645$, le asignaremos el símbolo **.** (influencia muy débil)

$1.645<[z_{ij}]$ 1.960, le asignamos o. (influencia débil)

$1.960<[z_{ij}]$ 2.576 le asignamos O (influencia fuerte)

$[z_{ij}] >2.576$, le asignamos @ (influencia muy fuerte)

La tabla donde se refleja lo expuesto es la siguiente:

| | Muy Grave | Grave | Lesión Med. | Leve | Marg.Tab |
|---|---|---|---|---|---|
| M.F Obs. | 20 | 10 | 10 | 30 | 70 |
| M.F Esp | 8.667 | 17.333 | 18.667 | 25.333 | 70 |
| M.F z | 3.850 | -1.761 | -2.006 | 0.927 | 70 |
| M.F Sim. | @ | O | O | . | 70 |
| F. Obs. | 30 | 40 | 20 | 50 | 140 |
| F Esp. | 17.333 | 34.667 | 37.333 | 50.667 | 140 |
| F: z | 3.043 | 0.906 | -2.837 | -0.094 | 140 |

| | | | | | |
|---|---|---|---|---|---|
| F. Sim. | @ | . | @ | . | 140 |
| F.E Obs. | 10 | 60 | 80 | 60 | 210 |
| F.E Esp. | 26 | 52 | 56 | 76 | 210 |
| F.E z | -3.138 | 1.109 | 3.207 | -1.835 | 210 |
| F.E Sim. | @ | . | @ | O | 210 |
| No F. Obs. | 5 | 20 | 30 | 50 | 105 |
| No F. Esp. | 13 | 26 | 28 | 38 | 105 |
| No F. z | -2.219 | -1.177 | 0.378 | 1.947 | 105 |
| No F. Sim.. | O | O | . | o | 105 |
| Marg. Lesión | 65 | 130 | 140 | 190 | 525 |

$X^2$ = 75.917 este valor depende del tamaño de la muestra y de la forma de la tabla, por tanto utilizaremos el valor V de Cramer como medida descriptiva de la asociación entre variables, esta medida esta comprendida entre 0 y 1, siendo las variables independientes cuande vale 0 y existiendo asociación perfecta cuando vale1. La expresión para V es:

$$V = \sqrt{\frac{X^2}{n \times [\min(filas, columnas) - 1]}}$$

En este caso vale 0.220.

---

**2. En un estudio sobre el sexismo en el trabajo se contrastaron las variables sexo y nivel de ingresos. Los resultados obtenidos sobre una muestra de 528 individuos se presentan en una tabla de doble entrada:**

| | **Alto** | **Medio Bajo** | **Bajo** |
|---|---|---|---|
| **Hombre** | **50** | **135** | **78** |
| **Mujer** | **20** | **147** | **98** |

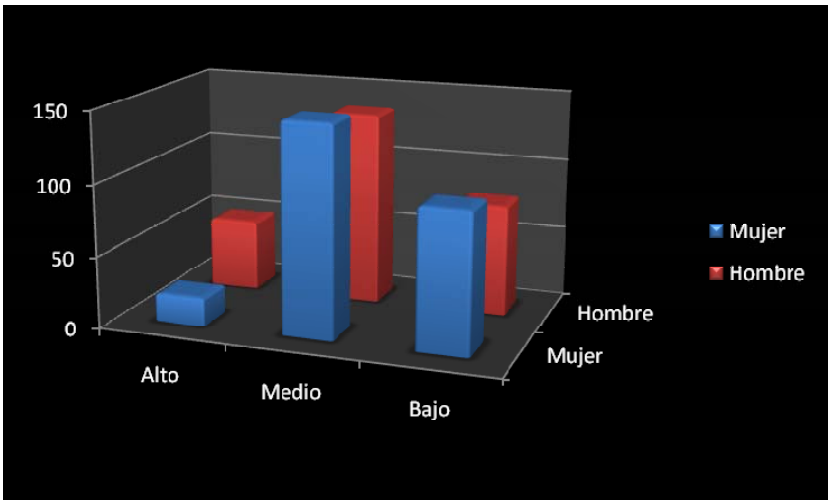
**Se pide:**

**a) Representar gráficamente las variables en estudio.**
**b) Calcular una medida descriptiva del nivel de asociación entre ambas variables. Realizar un análisis gráfico y analizar los resultados.**

---

**SOLUCIÓN:**

a)




b)

| | Alto | Medio | Bajo | Marg.Sexo |
|---|---|---|---|---|
| Hombre Obs. | 50 | 135 | 78 | 263 |
| Hombre Esp | 34.867 | 140.466 | 87.667 | 263 |
| Hombre z. | 2.563 | -0.461 | -1.032 | 263 |
| Hombre Sim. | O | . | . | 263 |
| Mujer Obs. | 20 | 147 | 98 | 265 |
| Mujer Esp. | 35.133 | 141.534 | 88.333 | 265 |
| Mujer z. | -2.553 | 0.459 | 1.029 | 265 |
| Mujer Sim. | O | . | . | 265 |
| Marg. Salario | 70 | 176 | 176 | 528 |

V=0.172

**3. De una determinada empresa se conocen los siguientes datos, referidos al volumen de ventas ( en millones de pesetas) y al gasto en publicidad ( en miles de pesetas) de los últimos 6 años:**

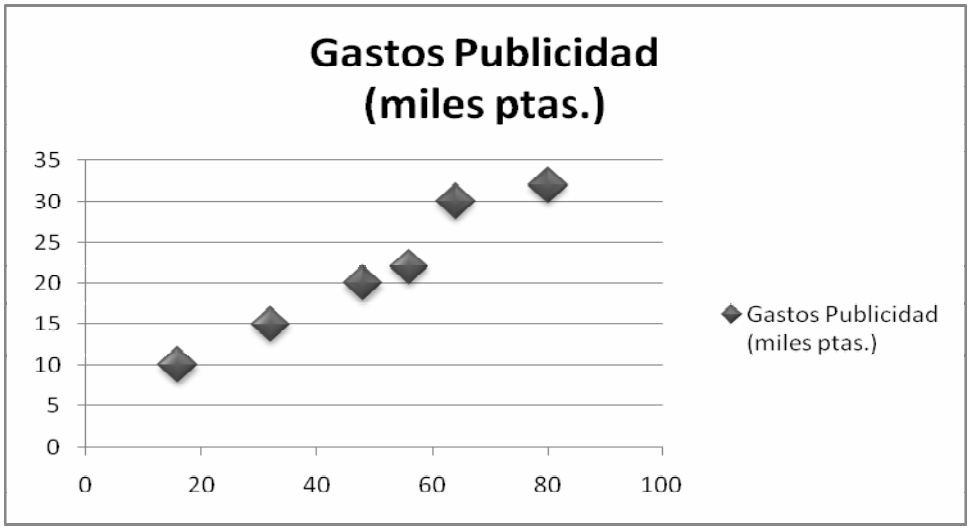
| Volumen de ventas(mill. Ptas) | Gastos Publicidad(miles ptas.) |
|---|---|
| 10 | 16 |
| 15 | 32 |
| 20 | 48 |
| 22 | 56 |
| 30 | 64 |
| 32 | 80 |

**a) ¿ Existe relación lineal entre las ventas de la empresa y sus gastos en publicidad? Razona la respuesta.**
**b) Obtener las rectas de regresión mínimo cuadrático.**
**c)¿ Qué volumen de ventas de la empresa se podría esperar en un año que se gaste de publicidad 60000 pesetas? ¿ Y para un gasto en publicidad de 200000 pesetas?**
**d) Si lo único que interesase fuese la evolución del volumen de ventas en términos de gastos en publicidad, sin tener en cuenta la cantidad concreta de cada uno de ellas, ¿existiría correlación ordinal entre ambas variables?**

---

**SOLUCIÓN:**

a) Dibujamos primero el diagrama de dispersión:



Observándolo podemos decir que existe relación lineal entre ambas variables.

Ahora calculamos el coeficiente de determinación lineal para obtener una medida descriptiva del grado de asociación lineal que existe entre las variables. La expresión del coeficiente de determinación es:

$$r^2 = \frac{S_{xy}^2}{S_x^2 \times S_y^2}$$

Donde $S_{xy}$ representa la covarianza de las variables X e Y. Cuya expresión simplificada

es:

$$S_{xy} = \frac{\sum X_i Y_i}{n} - \bar{x} \times \bar{y}$$

Para clarificar la forma de cálculo construimos la siguiente tabla: ( variable X= Gastos de publicidad y variable Y= Volumen de ventas)

| Y | X | $Y^2$ | $X^2$ | XY |
|---|---|---|---|---|
| 10 | 16 | 100 | 256 | 160 |
| 15 | 32 | 225 | 1024 | 480 |
| 20 | 48 | 400 | 2304 | 960 |
| 22 | 56 | 484 | 3136 | 1232 |
| 30 | 64 | 900 | 4096 | 1920 |
| 32 | 80 | 1024 | 6400 | 2560 |
| **129** | **296** | **3133** | **17216** | **7312** |

$\bar{X}$= 49.333; Y=21.5; $s_x$=20.870; $s_{xy}$=158

Substituyendo obtenemos que $r^2$ vale 0.956 que es lo que cabía esperar después de observar el diagrama de dispersión.

b) Si expresamos las rectas de regresión como $y^*$= a+bx y $x^*$=c+dy los coeficientes de los calculados son como:

$$b = \frac{S_{xy}}{S_x^2} \qquad a = \bar{y} - b \times \bar{x}$$

$$d = \frac{S_{xy}}{S_y^2} \qquad c = \bar{x} - d \times \bar{y}$$

Aplicándolas a este problema obtenemos las rectas de regresión:

$Y^*=3.604+0.363x$ ; $X^*=-7.356+2.637y$

c) Para realizar la predicción del volumen de ventas utilizamos la recta de regresión que tienen las ventas en función de los gastos en publicidad. Para un gasto en publicidad de 60000 pesetas obtendremos un volumen de ventas de $x^*=3.604+0.363*60=25.384$ millones de pesetas.

Si el gasto es de 200 millones de pesetas no podemos utilizar la recta de regresión puesto que el valor 200 esta fuera del recorrido del gasto en publicidad. Si sustituimos nos da un valor de 76204 millones de pesetas, pues las rectas sólo son válidas dentro del rango o para valores próximos a los extremos del recorrido.

d) Para solucionar este apartado calculamos el coeficiente de correlación ordinal de Spearman. El coeficiente de Spearman consiste en calcular el coeficiente de correlación lineal de los datos transformados a través de la función rango.

| **Y** | 10 | 15 | 20 | 22 | 30 | 32 | |
|---|---|---|---|---|---|---|---|
| **X** | 16 | 32 | 48 | 56 | 64 | 80 | |
| **Rang Y** | 1 | 2 | 3 | 4 | 5 | 6 | |
| **Rang X** | 1 | 2 | 3 | 4 | 5 | 6 | |
| $\mathbf{d_i}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $\mathbf{D_i^2}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

El coeficiente de Spearman cuando no existen empates en los rangos, como ocurre en estos datos, tiene la siguiente expresión:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

En este caso $r_s$ es 1 por tanto existe correlación ordinal positiva y perfecta, es decir a mayor gasto en publicidad mayor volumen de ventas.

(Podemos observar que la correlación lineal no es perfecta y sin embargo la correlación ordinaria si lo es).

---

**4. Un banco estatal de cierto país está estudiando la posibilidad de bajar los tipos de interés para incentivar la inversión privada, y así abrir la posibilidad de creación de puestos de trabajo. Para ello contrasta los tipos de interés real de diferentes países con la**

**inversión privada en los mismos, todo ello durante el último período. Obteniéndose los resultados que aparecen reflejados en la siguiente tabla:**

**Tipos de Interés(en tantos por uno)**

| **INVERSION(miles mills** | **0.05-0.10** | **0.10-0.15** | **0.15-0.20** | **0.20-0.25** |
|---|---|---|---|---|
| **10-50** | | | **2** | **6** |
| **50-100** | | **1** | **5** | |
| **100-150** | **1** | **4** | | |
| **150-200** | **5** | **1** | | |

**a)¿Existe relación lineal entre ambas variables? Razona la respuesta.**
**b)Construye la recta de regresión que explica la inversión en fluencia de los tipos de interés real.**
**c)¿Cómo variaría la inversión si se produce un incremento de una unidad en los tipos de interés real? Razónalo sin necesidad de hacer ningún cálculo.**
**d)Si el tipo de interés real baja de 0.18 a 0.09, ¿cómo variaría la inversión?**

---

**SOLUCIÓN:**

Para facilitar el seguimiento de los cálculos necesarios para resolver el problema construimos la siguiente tabla resumen: (variable X=tipo de interés real; variable Y=inversión).

| Y X | 0.075 | 0.125 | 0.175 | 0.225 | Marg.Y | $f_ix_i$ | $f_iy_i$ |
|---|---|---|---|---|---|---|---|
| 30 | 0 | 0 | 2 | 6 | 8 | 240 | 7200 |
| 75 | 0 | 1 | 5 | 0 | 6 | 450 | 33750 |
| 125 | 1 | 4 | 0 | 0 | 5 | 625 | 78120 |
| 175 | 5 | 1 | 0 | 0 | 6 | 1050 | 183750 |
| Marg. X | 6 | 6 | 7 | 6 | 25 | 2365 | 302850 |
| $f_ix_i$ | 0.45 | 0.75 | 1.225 | 1.35 | 3.775 | | |
| $f_ix_i^2$ | 0.03375 | 0.09375 | 0.21438 | 0.30375 | 0.64563 | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| $f_{ij}y_ixj$ | 0 | 0 | 10.5 | 40.5 | 285.375 | |
| | 0 | 9.375 | 65.625 | 0 | | |
| | 9.375 | 62.5 | 0 | 0 | | |
| | 65.625 | 21.875 | 0 | 0 | | |

X(media)=0.151; Y(media)=94.6; $s_x$=0.055; $s_{y=}$56.248; $s_{xy}$=-2.870

1. Para estudiar la relación lineal entre las variables tipo de interés e inversión utilizaremos el coeficiente de determinación como medida descriptiva de este hecho.
2. $y^*$=237.863-948.760x
3. El incremento en una unidad de la variable independiente coincide con el valor de la pendiente de la recta; en este caso el incremento será de -948.760( observamos que en este problema el incremento es ficticio pues 1 se sale del recorrido de la variable independiente).
4. El incremento será el producto entre la pendiente y la diferencia entre el tipo de interés en los dos estados, es decir, aumenta en -948.760*(0.09-0.18)=85.388miles de millones.

---

**5. Una compañía discográfica ha recopilado la siguiente información sobre 15 grupos musicales, a saber, el número de conciertos dados este verano y las ventas de discos de estos grupos( en miles de LPs), obteniendo los siguientes datos:**

**CONCIERTOS**

| **LPs** | **10-30** | **30-50** | **50-70** |
|---|---|---|---|
| **1-6** | **3** | **2** | **1** |
| **6-11** | **1** | **4** | **1** |
| **11-16** | **2** | **1** | **5** |

**a)Calcula el número medio de LPs vendidos por estos grupos.**
**b)Obtener la recta de regresión que explica la dependencia lineal**
**c)Si un grupo musical ha vendido 1800 LPs,¿Qué número de conciertos se prevee este verano?**

---

**SOLUCIÓN:**

a) 9000 LPs

c) y*=28.22+1.42x

d) y*=28.22+1.42*1.8=30776 Conciertos.

---

**6. Con objeto de analizar si existe relación lineal entre el consumo de energía eléctrica(kw.hora), variable X y el volumen de producción en millones de pesetas, variable Y, de una empresa se ha obtenido la siguiente información:**

$\overline{x} = 0.151; \overline{y} = 94.6; S_x = 0.055; S_y = 56.248; S_{xy} = -2.870$

**Se pide:**

1. **Ajústese la recta de regresión lineal que explica el consumo de electricidad en $f_i$ del volumen de producción. Razónese la validez de la recta ajustada**

---

**SOLUCIÓN:**

a) y*=-10.746+2.202x

b) r=0.959

---

**7. Una empresa de manufacturas basa las predicciones de sus ventas anuales en los resultados oficiales de la demanda total en la industria. A continuación se dan los datos de demanda total y las ventas efectuadas por la empresa en los últimos 11 años.**

| demanda total (miles de tm) | ventas (miles de tm) |
|---|---|
| 200 | 9 |
| 220 | 6 |
| 400 | 12 |
| 330 | 7 |
| 210 | 5 |
| 390 | 10 |
| 280 | 8 |
| 140 | 4 |
| 280 | 7 |
| 290 | 10 |
| 380 | 14 |

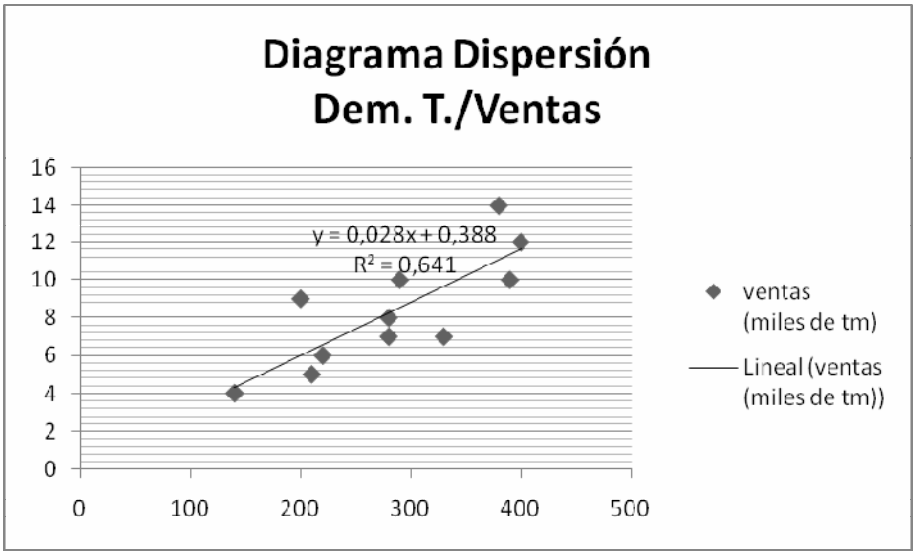
1. **Dibujar los diagramas de dispersión de los datos.**
2. **Trazar la recta que mas se ajuste a los datos.**

3. **Por medio de un ajuste mínimo cuadrático encontrar la recta que más se ajuste a las ventas de la empresa en función de la demanda total. Si la demanda total industrial es de 300000 toneladas, ¿Qué volumen de ventas se predeciría usando la recta de regresión calculada?¿y si utilizamos la recta encontrada en el apartado b)?**
4. **Realiza la validez del ajuste lineal realizado en el apartado anterior.**
   **Utilizando el método robusto de ajuste de una recta basado en la mediana, para obtener una recta de ajuste en los términos del apartado c). Realiza la predicción del apartado c. utilizando esta recta**

---

**SOLUCIÓN:**

1. X=Demanda Total, Y=Ventas




2. y*=0.422+0.028x; y*=0.422+0.028*300=8.822 Miles de Ton.
3. r=0.801; $r^2$=0.642.
4. Para calcular la recta robusta de ajuste basada en la mediana se procede de la siguiente forma:
   1. Se divide la muestra ordenada por la variable X en tres partes aproximadamente iguales, en este caso hemos tomado 4, 3 y 4.
   2. Se calcula la mediana para las variables X e Y en el primer y tercer subconjunto de datos.

      Primer subconjunto: $x_1$=Me(X)=205 ; $y_1$=Me(Y)=5.5

      Tercer subconjunto: $x_2$=Me(X)=385 ; $y_2$=Me(Y)=11

   3. Uniendo los puntos obtenemos la recta robusta de ajuste. La expresión para la pendiente($b_r$) y para el término independiente($a_r$) son:

$$b_r = \frac{y_2 - y_1}{x_2 - x_1}$$

$$a_y = \frac{x_2 y_1 - x_1 y_2}{x_2 - x_1}$$

Sustituyendo obtenemos $y^R$=-0.764+0.031*300=8.536 miles de Ton.( NOTA: ambas rectas están dibujadas sobre el diagrama de dispersión. El signo **.** del gráfico corresponde a los puntos( $x_1,y_1$) y ($x_2,y_2$) y el signo (cuadrado) a los datos del problema).

---

**8. Se está estudiando la relación entre el número de años que una persona está afiliada al sindicato y el nivel de satisfacción con la actuación de dicho sindicato. Para ello se parte de los datos de 7 individuos tomados aleatoriamente de personas adscritas a partidos políticos, obteniéndose:**

| **Años** | **8** | **7** | **10** | **3** | **6** | **13** | **4** |
|---|---|---|---|---|---|---|---|
| **Satisfacción** | **7** | **5** | **8** | **5** | **9** | **9** | **3** |

1. **Calcular el coeficiente de correlación lineal. Comentar el resultado obtenido.**
2. **Predecir el índice de satisfacción de una persona que lleva 11 años militando al sindicato. Conociendo que el índice de satisfacción es de 6 predecir los años que lleva en el sindicato**

---

**SOLUCIÓN:**

1. r=0.711
2. y*=3.118+0.474x ; y*=3.118+0.474*11=8.332 en la escala de satisfacción.
3. x*=0.270+1.068*y; x*=0.270+1.068*6= 6.678 años.

---

**9. En una región vinícola se observó la evolución del precio( en pesetas/litro) y la cantidad de producción( en toneladas) durante algunos años. Mirad la tabla:**

| Producción | 25-35 | 35-45 | 45-55 | 55-65 |
|---|---|---|---|---|
| 100 | 2 | | | 5 |
| 110 | | | 1 | |
| 120 | | | 3 | 1 |
| 140 | | 4 | 2 | |
| 160 | 2 | 3 | 1 | |
| 200 | 5 | 2 | | 1 |

1. **Calcula la recta de regresión lineal que pone el precio en función de la producción.**
2. **Analiza razonadamente la validez de la recta obtenida anteriormente**.

**¿Entre que valores estará el precio cuando la producción está entre 115 y 135 toneladas? Razona la respuesta.**

---

**SOLUCIÓN:**

1. y*=68.291-0.167x
2. r=0.556 ; $r^2$=0.309
3. y*=68.291-0.167*115= 49.086 e y*=68.291-0.167*135=45.746 ; el precio estará entre 45.746 y 68.291 pesetas

---

**10. Dados los siguientes conjuntos de datos:**

| U | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| V | 3 | 5 | 6 | 5 | 7 | 9 | 10 | 9 | 10 | 10 |
| W | 4.543 | 4.543 | 4.543 | 4.543 | 4.543 | 4.543 | 4.543 | 4.543 | 4.543 | 14.117 |
| X | 6.646 | 6.646 | 6 | 6 | 6 | 7 | 7 | 5.684 | 8.838 | 14.186 |

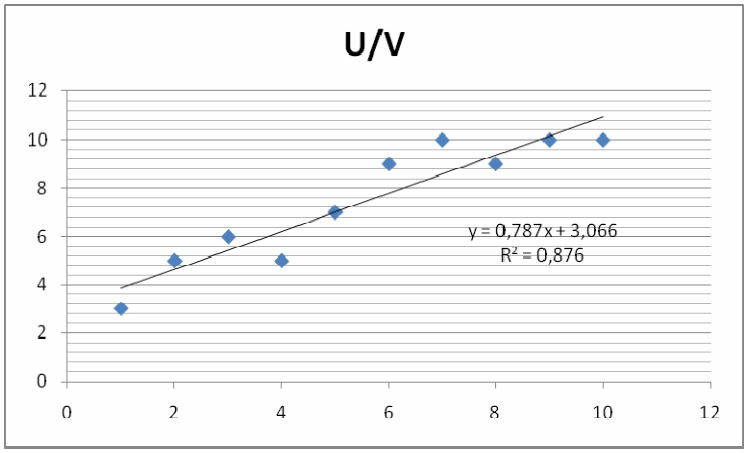
1. **Dibujar el diagrama de dispersión de cada uno de los conjuntos de datos.**
2. **Calcular la recta de regresión de cada uno de los conjuntos de datos y dibujarla en el diagrama de dispersión, considerando como variables independientes las variables U,W,X.**
3. **Calcular el coeficiente de correlación lineal para cada uno de los conjuntos.**
4. **¿Qué podemos observar?**
5. **Eliminando los outliers vuelve a calcular loa apartados b y c.**
6. **¿Qué otras rectas te parecerían mas adecuadas en los conjuntos anteriores? Razona la respuesta.**
7. **Calcula la recta de ajuste robusto**

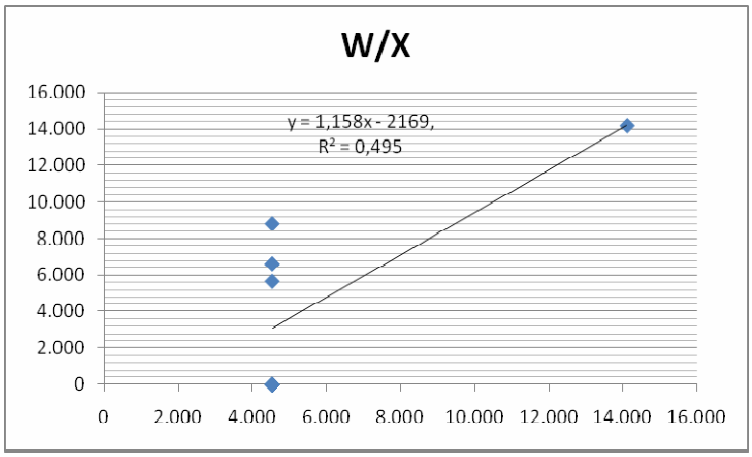
**¿Qué conclusiones podemos extraer de este problema?**

---

**SOLUCIÓN:**

a)




2. v*=3.067+0.788u ; x*=3.067+0.788w
3. $r_{uv}$=0.877 ; $r_{wx}$=0.877
4. Podemos observar que ambas rectas son exactamente iguales y que la relación lineal en ambas es la misma, pero se puede apreciar en las variables U/V la recta es más representativa que en el otro caso. Observamos que la presencia del outlier puede cambiar el resultado esperado.
5. En el diagrama de dispersión U/V no se observan outliers. En el diagrama W/X se advierte la presencia de un outlier, que es el punto( 14.117,14.186) si lo eliminamos obtenemos la x*=6.646 y la asociación lineal entre ellas es nula,
6. Este apartado se realizará para las variablea U/V $v^R$=3.571+0.714u.
7. La principal conclusión es que hay que dibujar siempre el diagrama de dispersión de datos.

---

**11. Un gerente de recursos humanos desea determinar el salario que debe pagar acierta categoría de obreros. Para determinar dicho salario que debe pagar a cierta categoría de obreros. Para determinar dicho salario se realiza un estudio en el que intervienen las variables Salario Mensual( en miles de ptas), Nivel de Producción Anual en la Empresa( en millones de ptas) y Nivel de especialización Media del Trabajador ( de 0 a 10). El gerente obtiene esta serie de resultados:**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Sal. | 123.4 | 135.7 | 115.9 | 100.6 | 98.7 | 150.4 | 124.6 | 110.0 | 138.6 | 123.4 |
| Prod. | 300.5 | 325.9 | 298.6 | 200.9 | 300.4 | 359.8 | 279.6 | 215.6 | 250.0 | 300.0 |
| Esp. | 4.3 | 5.5 | 7.8 | 4.9 | 4.3 | 8.5 | 6.4 | 5.6 | 5.3 | 5.0 |

**Se pide:**

1. **Calcular el plano de regresión lineal mínimo cuadrático que explica el salario en función de la producción y del nivel de especialización.**
2. **Estudia la validez de la función obtenida en el apartado anterior por medio de una medida descriptiva. ¿ Cuánto vale la varianza residual?**
3. **Calcula el coeficiente de correlación parcial para dos variables explicativas.**
4. **Comenta los resultados.**

**¿Qué salario se debería pagar si el nivel de producción fuese de 315 millones de ptas. y el nivel medio de especialización de 6.6?**

---

**SOLUCIÓN:**

1) Variable Y=Salario $X_i$=Producción $X_2$=Nivel de especialización.

La tabla de cálculos es:

| **Y** | **$X_1$** | **$X_2$** | **$Y^2$** | **$X_1^2$** | **$X_2^2$** | **$YX_1$** | **$YX_2$** | **$X_1X_2$** |
|---|---|---|---|---|---|---|---|---|
| 123.4 | 300.5 | 4.3 | 15227.6 | 90300.1 | 18.5 | 37081.7 | 530.6 | 1292.2 |
| 135.7 | 325.9 | 5.5 | 18414.5 | 106210.8 | 30.3 | 44224.6 | 746.5 | 1792.5 |
| 115.9 | 298.6 | 7.8 | 13432.8 | 89162.0 | 60.8 | 34607.7 | 904.0 | 2329.1 |
| 100.6 | 200.9 | 4.9 | 10120.4 | 40360.8 | 24.0 | 20210.5 | 492.9 | 984.4 |
| 98.7 | 300.4 | 4.3 | 9741.7 | 90240.2 | 18.5 | 29649.5 | 424.5 | 1291.7 |
| 150.4 | 359.8 | 8.5 | 22620.2 | 129456.0 | 72.3 | 54114.0 | 1278.4 | 3058.3 |
| 124.6 | 279.6 | 6.4 | 15525.2 | 78176.2 | 41.0 | 34838.32 | 797.4 | 1789.4 |
| 110.0 | 215.6 | 5.6 | 12100.0 | 46483.4 | 31.4 | 23716.0 | 616.0 | 1207.4 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 138.6 | 250.0 | 5.3 | 19210.0 | 62500.0 | 28.1 | 34650.0 | 734.6 | 1325.0 |
| 123.4 | 300.0 | 5.0 | 15227.6 | 90000.0 | 25.0 | 37020.0 | 617.0 | 1500.0 |
| **1221.3** | **2831.3** | **57.6** | **151619.8** | **822889.6** | **349.7** | **350112.2** | **7141.8** | **16570.0** |

La recta a construir tendrá la forma $y^*= a+b_1x_1+b_2x_2$ y para calcular los coeficientes de la recta aparece un Sistema de Ecuaciones Come éste:

$$\sum_{i=1}^{n} y_i = an + b_1\sum_{i=1}^{n} x_{1i} + b_2\sum_{i=1}^{n} x_{2i}$$

$$\sum_{i=1}^{n} y_i x_{1i} = a\sum_{i=1}^{n} x_{1i} + b_1\sum_{i=1}^{n} x_{1i}^2 + b_2\sum_{i=1}^{n} x_{2i}x_{1i}$$

$$\sum_{i=1}^{n} y_i x_{2i} = a\sum_{i=1}^{n} x_{2i} + b_1\sum_{i=1}^{n} x_{1i}x_{2i} + b_2\sum_{i=1}^{n} x_{2i}^2$$

Con estos datos el sistema de ecuaciones a resolver es:

$1221.3 = 10a +2831.3b_1 +57.6b_2$

$350112.2 =2831.3a +822889.6\ b_1 +16570.0b_2$

$7141.8 =57.6a +16570.0b_1 +349.7b_2$

Que tiene por solución $a=56198$ $b_1=0.158$ $b_2=3.664$ . Por tanto el plano de regresión es : $y^*=56.198+0.158x_1+3.664x_2$

$$y^* = 56{,}198 + 0{,}158x^1 + 3{,}664x_2$$

2) Las medidas descriptivas que vamos a utilizar son el coeficiente de determinación y el de correlación. Y definidas serían así:

$$\sum_{i=1}^{n}(y_i - Media_y)^2 = \sum_{i=1}^{n}(y_i^* - Media_y)^2 + \sum_{i=1}^{n}(y_i - y_i^*)^2$$

El coeficiente de determinación múltiple viene expresado como:

$$R^2_{yx_1x_2} = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT} = 1 - \frac{S_e^2}{S_y^2}$$

Donde $s^2$ es la variable residual.

Una expresión de la varianza residual que simplifique el cálculo es:

$$S_e^2 = \frac{\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n}(x_{1i}y_i) - b_2 \sum_{i=1}^{n}(x_{2i}y_i)}{n}$$

En este caso $R^2$=0.438 R=0.661 $s^2$=138.506

3 El coeficiente de correlación parcial entre la variable dependiente y una variable explicativa mide la fuerza de la relación lineal entre ambas cuando eliminamos el efecto lineal de las otras variables explicativas. Su resultado es:

$$r^2_{yx_ix_j} = \frac{SCE(x_j) - SCE(x_i, x_j)}{SCE(x_j)}$$

Donde SCE($x_i$) es la variación no explicada . Observando esta expresión vemos que el coeficiente de determinación parcial nos da el incremento relativo de la variable aplicada. El coeficiente de correlación parcial se consigue de la raiz cuadrada del coeficiente de determinación. En este caso SCE($x_1$)=1582.51 SCE($x_2$)=1643.06 SCE($x_1$,$x_2$)=1385.06 $r^2$=0.157 $r^2$=0.124

y*=56.198+0.158*315+3.664*6.6=130.15

---

**12. El gerente de una determinada empresa desea conocer, de forma aproximada, la demanda anual de producto que se realizará a la empresa en años futuros. Para determinar esta demanda realiza un estudio en el que intervienen las variables Precio Medio del Producto en un Año (en pesetas), Tasa de inflación Anual (IPC)(en tantos por uno) y la Demanda Anual (en miles de millones de pesetas). En una muestra de 20 años obtiene los siguientes resultados: (Demanda=Y, Precio=$x_1$, IPC=$x_2$)**

$$\sum y = 16.945; \sum x_1 = 3230; \sum x_2 = 1.1;$$

$$\sum YX_1 = 2609.452; \sum YX_2 = 0.83631; \sum X_1X_2 = 188.81$$

$$\sum Y^2 = 15.5035; \sum X_1^2 = 538638; \sum X_2^2 = 0.0738$$

**Calcular:**

a) **Determinar a partir del coeficiente de correlación lineal múltiple la validez de la función anterior. ¿Cuánto vale la varianza residual?**
b) **Determinar el coeficiente de correlación parcial para cada una de las variables explicativas y calcula el coeficiente de determinación para $YX_1$ e $YX_2$.**
c) **¿Qué variación se produce en la demanda si el precio se incrementa en 3 unidades permaneciendo fijo el IPC?¿y si se reduce el IPC en 0,03 permaneciendo fijo el precio? Razona la respuesta.**
d) **¿Qué variación porcentual se produciría en la demanda si el precio varia de 156 a 159 pesetas y el IPC permanece constante e igual a 0,04?**
e) **¿Qué volumen de demanda predecirías para un año en que el precio es de 159 pesetas y el IPC anual previsto es del 3,5%?**

---

**SOLUCIÓN:**

a) $S_e^2 = 0{,}008523;\ \ R^2_{yx1x2} = 0{,}851363$

b) SCE($x_1$)=0,195185; SCE($x_2$)=0,458711; SCE($x_1, x_2$)=0,17046; $r^2_{y,x1\cdot x2}$=0,6283934; $r^2_{y,x2\cdot x1}$=0,1266746; $r^2_{y,x1} = 0{,}829803$; $r^2_{y,x2} = 0{,}600013$.

c) La variación en la demanda será tres veces la pendiente correspondiente a la variable Precio, en este caso se produciría una reducción en la demanda de 0,0184 miles de millones de pesetas. La variación en la demanda será un aumento de demanda de 0,061 miles de millones de pesetas.

d) La variación porcentual es el cociente entre la variación total y la situación inicial por 100. En este problema será de:

(-0,0184/0.9115956)x100=-2,018%

e) y*=1.95191-0,00614697·159-2,03495·0,035=0,903 miles de millones de pesetas.

---

**13. Las calificaciones obtenidas por 9 alumnos en los exámenes del primer trimestre y del segundo son:**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **1º** | **5** | **7** | **6** | **9** | **3** | **1** | **2** | **4** | **6** |
| **2º** | **6** | **5** | **8** | **6** | **4** | **2** | **1** | **3** | **7** |

**Calcular:**

1- **Si existe correlación entre los resultados.**
2- **Las rectas de regresión de y sobre x y de x sobre y**

---

**SOLUCIÓN:**
Construimos la siguiente tabla:

| $x_i$ | $y_i$ | $x_i^2$ | $y_i^2$ | $x_i\, y_i$ |
|---|---|---|---|---|
| 5 | 6 | 25 | 36 | 30 |
| 7 | 5 | 49 | 25 | 35 |
| 6 | 8 | 36 | 64 | 48 |
| 9 | 6 | 81 | 36 | 54 |
| 3 | 4 | 9 | 16 | 12 |
| 1 | 2 | 1 | 4 | 2 |
| 2 | 1 | 4 | 1 | 2 |
| 4 | 3 | 16 | 9 | 12 |
| 6 | 7 | 36 | 49 | 42 |
| 43 | 42 | 257 | 240 | 237 |

$$\bar{x} = \frac{43}{9} = 4{,}78$$

$$\bar{y} = \frac{42}{9} = 4{,}67$$

$$S_x^2 = \frac{257}{9} - 4{,}78^2 = 5{,}71$$

$$S_y^2 = \frac{240}{9} - 4{,}67^2 = 4{,}86$$

$$S_{xy} = \frac{237}{9} - 4{,}78 \times 4{,}67 = 4{,}01$$

$$r = \frac{S_{xy}}{S_x S_y} = \frac{4{,}01}{\sqrt{5{,}71}\sqrt{4{,}86}} = 0{,}36$$

$$y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x}) \Rightarrow y - 4{,}67 = \frac{4{,}01}{5{,}71}(x - 4{,}78)$$

$$x - \bar{x} = \frac{S_{xy}}{S_y^2}(y - \bar{y}) \Rightarrow x - 4{,}78 = \frac{4{,}01}{4{,}86}(x - 4{,}67)$$

---

**14.** : **Calcular el coeficiente de correlación y las ecuaciones de las rectas de regresión de la distribución adjunta.**

| $x_i$ / $y_j$ | **1,65-1,70** | **1,70-1,75** | **1,75-1,80** |
|---|---|---|---|
| **70-75** | **1** | | |
| **75-80** | **2** | **2** | |

| | | | |
|---|---|---|---|
| **80-85** | **1** | **1** | **3** |

**$x_i$ = Tallas** **$y_j$ = Pesos**

---

**SOLUCIÓN:**

Efectuamos un cambio de variable mediante

$x_i = x_0 + a\, x_i'$ $x_0 = 1{,}725$ ; $a = 0{,}05$

$y_j = y_0 + b\, y_j'$ $y_0 = 77{,}5$ ; $b = 5$

Llamando: $x_i$ = Tallas e $y_j$ = Peso, construimos la siguiente tabla:

| $x_i'$ / $y_i'$ | | -1 | 0 | 1 | $f_j$ | $f_j\, y_j$ | $f_j\, y_j^2$ |
|---|---|---|---|---|---|---|---|
| | $x_i$ / $y_j$ | 1,675 / 1,65-1,70 | 1,725 / 1,70-1,75 | 1775 / 1,75-1,80 | | | |
| -1 | 72,5 70-75 | 1 | | | 1 | -1 | 1 |
| 0 | 77,5 75-80 | 2 | 2 | | 4 | 0 | 0 |
| 1 | 82,5 80-85 | 1 | 1 | 3 | 5 | 5 | 5 |
| $f_{i\cdot}$ | | 4 | 3 | 3 | 10 | 4 | 6 |
| $f_{i\cdot}\, x_i'$ | | -4 | 0 | 3 | -1 | | |
| $f_{i}\, x_i'^2$ | | 4 | 0 | 3 | 7 | | |

$$\bar{x}' = \frac{\Sigma_i f_i x_i'}{n} = -\frac{1}{10} \Rightarrow \bar{x} = x_o + a\bar{x}' = 1{,}725 + 0{,}005\frac{-1}{10} = 1.72$$

$$\bar{y}' = \frac{\Sigma_j f_j y_j'}{n} = -\frac{4}{10} \Rightarrow \bar{y} = y_o + b\bar{y}' = 77{,}5 + 5\frac{4}{10} = 79{,}5$$

$$S_{x'}^2 = \frac{\Sigma_i f_i x_i'^2}{n} - \bar{x}'^2 = \frac{7}{10} - (-\frac{1}{10})^2 = 0{,}69$$

$$S_{y'}^2 = \frac{\sum_j f_j y_j'^2}{n} - \bar{y}'^2 = \frac{6}{10} - \left(\frac{4}{10}\right)^2 = 0{,}44$$

$$S'_{x'y'} = \frac{\sum f_{ij} x_i' y_j'}{n} - \bar{x}'\bar{y}' = \frac{3}{10} - \left(-\frac{1}{10}\right)\left(\frac{4}{10}\right) = 0{,}26$$

1- Coeficiente de correlación

$$r = r' = \frac{S_{x'y'}}{S_{x'} S_{y'}} = \frac{0{,}26}{\sqrt{0{,}69}\sqrt{0{,}44}} = \frac{0{,}26}{0{,}55} = 0{,}47$$

2- Recta de regresión de y sobre x

$$b_{yx} = \frac{b}{a} b'_{yx} = \frac{5}{0{,}05} \times \frac{S_{x'y'}}{S_{x'}^2} = \frac{5}{0{,}05} \times \frac{0{,}26}{0{,}69} = 37{,}68$$

Luego

$$y - \bar{y} = b_{yx}(x - \bar{x}) \Rightarrow y - 79{,}5 = 37{,}68(x - 1{,}72)$$

3- Recta de regresión de x sobre y

$$S_{xy} = \frac{a}{b} b'_{xy} = \frac{0{,}05}{5} \times \frac{S_{xy}}{S_y^2} = \frac{0{,}05}{5} \times \frac{0{,}26}{0{,}44} = 0{,}01$$

Luego

$$x - \bar{x} = b_{xy}(y - \bar{y}) \Rightarrow x - 1{,}72 = 0{,}01(y - 79{,}5)$$

---

**15. Elegidos 50 matrimonios al azar y preguntada la edad de ambos al contraer matrimonio, se obtuvo la siguiente tabla bidimensional:**

| $x_i$ / $y_i$ | 15-20 | 20-25 | 25-30 | 30-35 | 35-40 |
|---|---|---|---|---|---|
| 15-18 | 3 | 2 | 3 | | |
| 18-21 | | 4 | 2 | 2 | |
| 21-24 | | 7 | 10 | 6 | 1 |
| 24-27 | | | 2 | 5 | 3 |

**x = Edad del marido.** **y = Edad de la esposa.**

**Calcular:**

1- **Recta de regresión de y sobre x.**
2- **Recta de regresión de x sobre y.**

---

**SOLUCIÓN:**
Construimos la siguiente tabla:

| $x_i'$ / $y_j'$ | | -2 | -1 | 0 | 1 | 2 | $f_{.j}$ | $f_{.j}\, y_j'$ | $f_{.j}\, y_j'^2$ |
|---|---|---|---|---|---|---|---|---|---|
| | $x_i$ / $y_j$ | 17,5 15-20 | 22,5 20-25 | 27,5 25-30 | 32,5 30-35 | 37,5 35-40 | | | |
| -1 | 16,5 15-18 | 3 | 2 | 3 | | | 8 | -24 | 72 |
| 0 | 19,5 18-21 | | 4 | 2 | 2 | | 8 | -8 | 8 |
| 1 | 22,5 21-24 | | 7 | 10 | 6 | 1 | 24 | 24 | 24 |
| 3 | 22,5 24-27 | | | 2 | 5 | 3 | 10 | 30 | 90 |
| $f_{i.}$ | | 3 | 13 | 17 | 13 | 4 | 50 | 22 | 194 |
| $f_{i.}\, x_i'$ | | -6 | -13 | 0 | 13 | 8 | 2 | | |
| $f_{i.}\, x_i'^2$ | | 12 | 13 | 0 | 13 | 16 | 54 | | |

Hemos efectuado el cambio de variable

$$x_i = x_o + a x_i'; \quad x_o = 27{,}5 \;\; y \;\; a = 5$$
$$y_j = y_o + a y_j'; \quad y_o = 21 \;\; y \;\; a = 1{,}5$$

$$\bar{x}' = \frac{\sum_i f_i x_i'}{n} = -\frac{2}{50} \Rightarrow \bar{x} = x_o + a\bar{x}' = 127{,}5 + 5\frac{2}{50} = 27{,}7$$

$$\bar{y}' = \frac{\sum_j f_j y_j'}{n} = -\frac{22}{50} \Rightarrow \bar{y} = y_o + b\bar{y}' = 21 + 1{,}5\frac{22}{50} = 21{,}66$$

$$S_{x'}^2 = \frac{\sum_i f_i x_i'^2}{n} - \bar{x}'^2 = \frac{54}{50} - \left(\frac{2}{50}\right)^2 = 1{,}08$$

$$S_{y'}^2 = \frac{\Sigma_j f_j y_j'^2}{n} - \bar{y}'^2 = \frac{194}{50} - \left(\frac{22}{50}\right)^2 = 3{,}69$$

$$S'_{x'y'} = \frac{\Sigma f_{ij} x_i' y_j'}{n} - \bar{x}'\bar{y}' = \frac{60}{50} - \left(-\frac{2}{50}\right)\left(\frac{22}{50}\right) = 1{,}18$$

Resulta:

Coeficiente de regresión de y sobre x

$$b_{yx} = \frac{b}{a} b'_{yx} = \frac{b}{a} \times \frac{S_{x'y'}}{S_{x'}^2} = \frac{1{,}5}{5} \times \frac{1{,}18}{1{,}08} = 0{,}327$$

Recta de regresión de y sobre x

$$y - \bar{y} = b_{yx}(x - \bar{x}) \Rightarrow y - 21{,}66 = 0{,}327(x - 27{,}7)$$

Coeficiente de regresión de x sobre y

$$b_{yx} = \frac{b}{a} b'_{yx} = \frac{a}{b} \times \frac{S_{x'y'}}{S_{x'}^2} = \frac{5}{1{,}5} \times \frac{1{,}18}{3{,}69} = 1{,}06$$

Recta de regresión de x sobre y

$$x - \bar{x} = b_{xy}(y - \bar{y}) \Rightarrow x - 27{,}7 = 1{,}06(y - 21{,}66)$$

---

**16. Se han estudiado los pesos en kg y las tallas en cm de 70 individuos obteniéndose los datos de la tabla siguiente:**

| PESOS/TALLAS | 159-161 | 161-163 | 163-165 | 165-167 | 167-169 | 169-171 |
|---|---|---|---|---|---|---|
| 48 | 3 | 2 | 2 | 1 | 0 | 0 |
| 51 | 2 | 3 | 4 | 2 | 2 | 1 |
| 54 | 1 | 3 | 6 | 8 | 5 | 1 |
| 57 | 0 | 0 | 1 | 2 | 8 | 3 |
| 60 | 0 | 0 | 0 | 2 | 4 | 4 |

**1. Hallar el peso medio y la talla media así como el error cometido al resumir pesos y tallas por sus valores medios ¿ Que media es mejor?**

**2. Hallar la distribución según las tallas de los individuos que pesan 54 kg y la distribución según los pesos de los individuos que miden entre 161 cm y 167 cm. Hallar media y varianza de las dos distribuciones condicionadas.**

---

**SOLUCIÓN:**

Si llamamos X a la variable pesos e Y a la variable tallas, los datos pueden arreglarse en una tabla de doble entrada como sigue para realizar los cálculos:

| X/Y | 160 | 162 | 164 | 166 | 168 | 170 | Ni. | Ni.xi | $Ni.xi^2$ |
|---|---|---|---|---|---|---|---|---|---|
| **48** | 3 | 2 | 2 | 1 | 0 | 0 | **8** | **384** | **18432** |
| **51** | 2 | 3 | 4 | 2 | 2 | 1 | **14** | **714** | **36414** |
| **54** | 1 | 3 | 6 | 8 | 5 | 1 | **24** | **1296** | **69984** |
| **57** | 0 | 0 | 1 | 2 | 8 | 3 | **14** | **798** | **45486** |
| **60** | 0 | 0 | 0 | 2 | 4 | 4 | **10** | **600** | **36000** |
| **n.,j** | **6** | **8** | **13** | **15** | **19** | **9** | **70** | **3792** | **206316** |
| **n.j yj** | **960** | **1296** | **2132** | **2490** | **3192** | **1530** | **11600** | | |
| **$n.j\ yj^2$** | **153600** | **209952** | **349648** | **413340** | **536256** | **260100** | **1922896** | | |

Para hallar el peso medio y la talla media se calcularán las medias de las distribuciones marginales de X e Y respectivamente. Asimismo para cuantificar el error cometido al resumir pesos y tallas por sus valores medios se cuantificarán los coeficientes de variación de pearson para ambas marginales. Las distribuciones marginales de X e Y son las siguientes:

| X | ni. | Y | n.j |
|---|---|---|---|
| 48 | 8 | 160 | 6 |
| 51 | 14 | 162 | 8 |
| 54 | 24 | 164 | 13 |
| 57 | 14 | 166 | 15 |
| 60 | 10 | 168 | 19 |
| | | 170 | 9 |

Tenemos lo siguiente:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{5} ni = \frac{3792}{70} = 54{,}17$$

$$\overline{Y} = \frac{1}{N}\sum_{j=1}^{6} n_j x_j = \frac{11600}{70} = 165{,}71$$

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{5} n_i x_i^2 - \overline{X^2} = \frac{206316}{70} - 54{,}17^2 = 12{,}98$$

$$\sigma_y^2 = \frac{1}{N}\sum_{j=1}^{6} n_j y_j^2 - \overline{Y^2} = \frac{1922896}{70} - 165{,}71^2 = 10{,}13$$

$$V_x = \frac{\sigma_x}{\overline{X}} = \frac{\sqrt{12{,}98}}{54{,}17} = 0{,}0665 \cong 6{,}65\%$$

$$V_y = \frac{\sigma_y}{\overline{Y}} = \frac{\sqrt{10{,}13}}{165{,}71} = 0{,}0192 \cong 1{,}92\%$$

Se observa que el menor coeficiente de variación es el relativo a la talla media, que resulta ser así un promedio más adecuado.

La distribución según las tallas de los individuos que pesan 54 kg es la distribución de Y condicionada a X=54, y la distribución según los pesos de los individuos que miden entre 161 cm y 167 cm es la distribución de X condicionada a Y=162,164, 166.

| **X/Y=162, 164, 166** | **n i/j=2, 3, 4** | **Y/X=54** | **n j/i=3** |
|---|---|---|---|
| 48 | 5 | 160 | 1 |
| 51 | 9 | 162 | 3 |
| 54 | 17 | 164 | 6 |
| 57 | 3 | 166 | 8 |

| | | | |
|---|---|---|---|
| 60 | 2 | 168 | 5 |
| | | 170 | 1 |

La media y la varianza de cada una de estas dos distribuciones condicionadas se calcula de la misma forma que para cualquier distribución de frecuencias.

$$\overline{X}/_{Y=162,164,166} = \frac{1}{N}\sum_{i=1}^{5} n_{i/j=2,3,4} x_i = \frac{1908}{36} = 53$$

$$\overline{Y}/_{X=54} = \frac{1}{N}\sum_{j=1}^{6} n_{j/i=3} y_j = \frac{3968}{24} = 165{,}33$$

$$\sigma^2_{x/y=162,164,166} = \frac{1}{N}\sum_{i=1}^{5} n_{i/j=2,3,4} x_i^2 - 53^2 = \frac{101448}{36} - 53^2 = 9$$

$$\sigma^2_{y/x=54} = \frac{1}{N}\sum_{j=1}^{6} n_{j/i=3} y_{j^i}^2 - 165{,}33^2 = \frac{656176}{24} - 165{,}33^2 = 5{,}55$$

---

**17. Se considera la variable bidimensional (X, Y) cuya distribución de frecuencias se presenta en la tabla siguiente:**

| X/Y | 15 | 24 | 27 | 30 |
|---|---|---|---|---|
| 12 | 3 | 4 | 2 | 5 |
| 15 | 6 | 8 | 4 | 10 |
| 19 | 9 | 12 | 6 | 15 |

1. **Estudiar si las dos variables son independientes utilizando la distribución conjunta y las marginales.**
2. **Estudiar si las dos variables son independientes utilizando las distribuciones marginales y las condicionadas.**
3. **Hallar la covarianza de X e Y.**

---

**SOLUCIÓN:**

Para estudiar la independencia de las dos variables utilizando la distribución conjunta y las marginales tenemos que comprobar quefij=fi.f.j ∨ i,j.

La primera tarea será construir una tabla con la distribución conjunta (fij=nij/N) y con las marginales (fi.=ni./N y f .j=n.j/N).

| X/Y | 15 | 24 | 27 | 30 | ni. |
|---|---|---|---|---|---|
| **12** | 3 | 4 | 2 | 5 | 14 |
| **15** | 6 | 8 | 4 | 10 | 28 |
| **19** | 9 | 12 | 6 | 15 | 42 |
| **n.j** | 18 | 24 | 12 | 30 | 84 |
| **fij** | | | | | fi. |
| | 0,03571429 | 0,0476191 | 0,02380952 | 0,05952381 | 0,1666667 |
| | 0,07142857 | 0,0952381 | 0,04761905 | 0,11904762 | 0,3333333 |
| | 0,10714286 | 0,1428571 | 0,07142857 | 0,17857143 | 0,5 |
| **f.j** | 0,21428571 | 0,2857143 | 0,14285714 | 0,35714286 | 1 |

Ya estamos en condiciones de probar que $f_{ij} = f_i \cdot f_j \forall i, j$. Para ello ordenaremos los cálculos $f_i \cdot f_j$ como se indica a continuación:

| | | | |
|---|---|---|---|
| 0,21428*0,16666 | 0,28571*0,16666 | 0,14285714*0,16666 | 0,37142*0,16666 |
| 0,21428*0,33333 | 0,28571*0,33333 | 0,14285714*0,33333 | 0,37142*0,33333 |
| 0,21428*0,5 | 0,28571*0,5 | 0,14285714*0,5 | 0,37142*0,5 |

Observamos que, una vez realizados estos cálculos, se obtiene la tabla de la distribución conjunta fij.

| | | | | |
|---|---|---|---|---|
| fij | 0,035714286 | 0,04761905 | 0,02380952 | 0,05952381 |
| | 0,071428571 | 0,0952381 | 0,04761905 | 0,11904762 |
| | 0,107142857 | 0,14285714 | 0,07142857 | 0,17857143 |
| | 0,214285714 | 0,28571429 | 0,14285714 | 0,35714286 |
| | | | | |

Para estudiar la independencia de las dos variables utilizando las distribuciones marginales y las condicionadas tenemos que comprobar que $f_{i/j} = f_j \forall i, j.$

| | **fi/j=1** | **fi/j=2** | **fi/j=3** | **fi/j=4** | **nj.** |
|---|---|---|---|---|---|
| **fj/i=1** | 3/14 | 4/14 | 2/14 | 5/14 | 14 |
| **fj/i=2** | 6/28 | 8/28 | 4/28 | 10/28 | 28 |
| **fj/i=3** | 9/42 | 12/42 | 6/42 | 15/42 | 42 |
| **f.j** | 18/84 | 24/84 | 12/84 | 30/84 | 84 |

Observamos que se cumple fj/i $\forall$ i,j ya que:

$3/14 = 6/28 = 9/42 = 18/84$

$4/14 = 8/28 = 12/42 = 24/84$

$2/14 = 4/28 = 6/42 = 12/84$

$5/14 = 10/28 = 15/42 = 30/84$

Para estudiar la independencia de las dos variables utilizando las distribuciones marginales y la condicionadas también podríamos comprobar que fi/j=fi. $\forall$ i,j.

| | **fi/j=1** | **fi/j=2** | **fi/j=3** | **fi/j=4** | **f.j** |
|---|---|---|---|---|---|
| | 3/18 | 4/24 | 2/12 | 5/30 | 14/84 |
| | 6/18 | 8/24 | 4/12 | 10/30 | 28/84 |
| | 9/18 | 12/24 | 6/12 | 15/30 | 42/84 |
| **nj.** | 18 | 24 | 12 | 30 | 84 |

Observamos que se cumple fi/j=fi. $\forall$ i,j ya que:

$3/18 = 4/24 = 2/12 = 5/30 = 14/84$

$6/18 = 8/24 = 4/12 = 10/30 = 24/84$

$9/18 = 12/24 = 6/12 = 15/30 = 42/84$

La covarianza entre X e Y viene dada por la expresión:

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{h}\sum_{j=1}^{k}(x_i - \overline{x})(y_j - \overline{y})n_{ij}$$

Luego para su cálculo necesitamos las medias de las dos marginales X e Y, que se calcularán con los datos de la tabla:

| X/Y | 15 | 24 | 27 | 30 | nj. |
|---|---|---|---|---|---|
| 12 | 3 | 4 | 2 | 5 | 14 |
| 15 | 6 | 8 | 4 | 10 | 28 |
| 19 | 9 | 12 | 6 | 15 | 42 |
| n.j | 18 | 24 | 12 | 30 | 84 |

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{3} n_i x_i = \frac{1386}{84} = 16{,}5$$

$$\overline{Y} = \frac{1}{N}\sum_{j=1}^{4} n_j x_j = \frac{2070}{84} = 24{,}64$$

La covarianza, que será cero debido a la independencia, puede calcularse como sigue

$$\sigma_{xy} = \frac{1}{N}[(12-16{,}5)(15-24{,}6)+(12-16{,}5)(24-24{,}6)+(12-16{,}5)(27-24{,}6)+(12-16{,}5)(30-24{,}6)+$$
$$(15-16{,}5)(15-24{,}6)+(15-16{,}5)(24-24{,}6)+(15-16{,}5)(27-24{,}6)+(15-16{,}5)(30-24{,}6)+$$
$$(19-16{,}5)(15-24{,}6)+(19-16{,}5)(24-24{,}6)+(19-16{,}5)(27-24{,}6)+(19-16{,}5(30-24{,}6)+=0$$

La covarianza también puede calcularse de la forma siguiente

$$m_{11} = \sigma_{XY} = \frac{1}{N}\sum_{i,j} x_i y_j n_{ij} - \overline{X}\,\overline{Y} = \frac{34155}{84} - 16{,}5 * 24{,}64 = 0$$

$$\sum_{i,j} x_i y_j n_{ij} = 34155$$

| **X/Y** | **15** | **24** | **27** | **30** | **ni.** | **xiyin1j** | **xiz2n2j** | **xiz3n3j** | **xiz4n4j** | **∑xizinij** |
|---|---|---|---|---|---|---|---|---|---|---|

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **12** | 3 | 4 | 2 | 5 | 14 | 540 | 1152 | 648 | 1800 | 4140 |
| **15** | 6 | 8 | 4 | 10 | 28 | 1350 | 2880 | 1620 | 4500 | 10350 |
| **19** | 9 | 12 | 6 | 15 | 42 | 2565 | 5472 | 3078 | 8550 | 19665 |
| **n.j** | 18 | 24 | 12 | 30 | 84 | 4455 | 9504 | 5346 | 14850 | 34155 |

---

**18. En una empresa se toma una muestra de 100 trabajadores con la finalidad de estudiar si hay relación entre su edad X y los días que están de baja en el año Y. se obtienen los siguientes resultados:**

| X/Y | 0-20 | 20-40 | 40-60 | ni. |
|---|---|---|---|---|
| **18-30** | **28** | **2** | **0** | **30** |
| **30-40** | **26** | **15** | **4** | **45** |
| **40-50** | **6** | **14** | **5** | **25** |

1. **¿es simétrica la distribución del número de días de baja de los trabajadores?**
2. **¿Cuál es la edad más frecuente de los trabajadores que piden la baja?**
3. **Ajustar mediante un modelo exponencial los días de baja en función de la edad.**
4. **realizar el mismo ajuste considerando un modelo lineal.**
5. **¿Cuál de los ajustes es el mejor?**

---

**SOLUCIÓN:**

Para realizar los cálculos necesarios elaboramos la tabla siguiente:

| X/Y | 10 | 30 | 50 | ni. | xini. | $^{xi2ni}$. | ci | hi=ni/ci |
|---|---|---|---|---|---|---|---|---|
| 24 | 28 | 2 | 0 | 30 | 720 | 17280 | 12 | 2,5 |
| 35 | 26 | 15 | 4 | 45 | 1575 | 55125 | 10 | 4,5 |
| 45 | 6 | 14 | 5 | 25 | 1125 | 50625 | 10 | 2,5 |
| n.j | 60 | 31 | 9 | 100 | 3420 | 123030 | | |
| yjn.j | 600 | 930 | 450 | 1980 | | | | |
| $^{yj2n}$.j | 6000 | 27900 | 22500 | 56400 | | | | |
| $^{yj3n}$.j | 60000 | 837000 | 1E+06 | 2022000 | | | | |
| N.j | 60 | 91 | 100 | | | | | |

Realizamos los siguientes cálculos:

$$a_{10} = \overline{X} = \frac{1}{N}\sum_{i=1}^{3} n_i x_i = \frac{3420}{100} = 34{,}2$$

$$a_{01} = \overline{Y} = \frac{1}{N}\sum_{j=1}^{3} n_j x_j = \frac{1980}{100} = 19{,}8$$

$$m_{20} = \sigma_x^2 = \frac{1}{N}\sum_{i=1}^{3} n_i x_i^2 - \overline{X^2} = a_{20} - a_{10}^2 = \frac{123030}{100} - 34{,}2^2 = \underbrace{1230{,}3}_{a_{20}} - 1169{,}64 = 60{,}66$$

$$m_{02} = \sigma_y^2 = \frac{1}{N}\sum_{i=1}^{3} n_i y_i^2 - \overline{Y^2} = a_{02} - a_{01}^2 = \frac{56400}{100} - 19{,}8^2 = \underbrace{564}_{a_{02}} - 392{,}04 = 171{,}96$$

$$a_{03} = \overline{Y} = \frac{1}{N}\sum_{i=1}^{3} n_i x_i^3 = \frac{1980}{100} = 19{,}8$$

Para estudiar la asimetría del número de días de baja de los trabajadores calculamos el coeficiente de asimetría de Fisher de la variable marginal Y como sigue:

$$g_{01} = \frac{m_{03}}{\sigma_Y^3} = \frac{2243{,}184}{(\sqrt{171{,}96})^3} = 0{,}99$$

Se observa que hay una ligera asimetría hacia la derecha, pero muy pequeña. Los días de baja se distribuyen casi simétricamente a lo largo del año.

Para calcular la edad más frecuente de los trabajadores que piden la baja hallaremos la moda de la variable marginal X. Observamos que el intervalo modal es [30,40] ya que es el que tiene mayor frecuencia ni. El cálculo de la moda se realiza como sigue:

$$M_0 = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} c_i = 30 + \frac{2{,}5}{2{,}5 + 2{,}5} 10 = 35 años$$

Ahora intentaremos ajustar los días de baja en función de la edad de los trabajadores mediante un modelo de regresión exponencial de ecuación $y=ab^x$

$$y = ab^x \Rightarrow Log(y) = Log(a) + Log(b)x$$

La regresión exponencial es equivalente a la regresión lineal con variable dependiente log(y) y con variable independiente x. Los cálculos para esta regresión se presentan en la tabla siguiente:

| **X/z** | 1 | 1,4771 | 1,699 | ni. | xiz1n1j | xiz2n2j | xiz3n3j | ∑xizinij |
|---|---|---|---|---|---|---|---|---|
| 24 | 28 | 2 | 0 | 30 | 672 | 70,9008 | 0 | 742,9008 |
| 35 | 26 | 15 | 4 | 45 | 910 | 775,478 | 237,86 | 1923,338 |
| 45 | 6 | 14 | 5 | 25 | 270 | 930,573 | 382,275 | 1582,848 |
| n.j | 60 | 31 | 9 | 100 | 1852 | 1776,95 | 620,135 | 4249,086 |
| zjn.j | 60 | 45,7901 | 15,291 | 121,081 | | | | |
| $^{zj2n}$.j | 60 | 67,63656 | 25,979 | 153,616 | | | | |

El parámetro log(b) se estima por mínimos cuadrados mediante:

$$Log(b)\frac{\sigma_{XZ}}{\sigma_X^2} = \frac{\frac{1}{N}\sum_{i,j} x_i y_j n_{ij} - \overline{X}}{\sigma_X^2}$$

Los parámetros finales buscados a y b del modelo exponencial se estimarán mediante:

$$a = 10^{0,6013} = 10,04$$

$$b = 10^{0,01782} = 3,99$$

El modelo estimado tiene la ecuación $y = 3,99(1,04)^3$

Para medir la calidad de este ajuste podemos utilizar el coeficiente de determinación $R^2$ que se calcula como:

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_Y^2} = 1 - \frac{\sum_{i=1}^{3}\sum_{j=1}^{3}(y_i - (3,99(1,04)^{x_i}))^2 n_{ij}}{\sigma_y^2} = 1 - \frac{126,14}{171,96} = 0,26$$

El ajuste no es de calidad porque $R^2$ esta más cerca del cero que de la unidad. El ajuste por regresión lineal de la forma y= a+bx siendo:

$$b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{\frac{1}{N}\sum_{i,j} x_i y_j n_{ij} - \overline{X}\,\overline{Y}}{\sigma_x^2} = \frac{\frac{1}{100}72860 - 34{,}20 * 19{,}80}{60{,}66} = 0{,}854$$

$$a = \overline{Y} - b\overline{X} = 19{,}8 - 0{,}854 * 34{,}20 = -9{,}4$$

El coeficiente de determinación será en este caso el cuadrado del coeficiente de correlación que se calcula como sigue:

$$r^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2} = \frac{(\frac{1}{N}\sum_{i,j} x_i y_j n_{ij} - \overline{X}\overline{Y})^2}{\sigma_X^2 \sigma_Y^2} = \frac{\left(\frac{1}{100}728960 - 34{,}20 * 19{,}80\right)^2}{60{,}66 * 171{,}96} = 0{,}2536$$

Para realizar el calculo de $\sum_{i,j} x_i y_j n_{ij} = 72860$ se utiliza la siguiente tabla:

| X/Y | 10 | 30 | 50 | $n_i$ | $x_i y_j n_{ij}$ | $x_{i2} y_{j2} n_{ij2}$ | $x_{i3} y_{j3} n_{ij3}$ | $\sum x_i y_j n_{ij}$ |
|---|---|---|---|---|---|---|---|---|
| 24 | 28 | 2 | 0 | 30 | 6720 | 1440 | 0 | 8160 |
| 35 | 26 | 14 | 4 | 45 | 9100 | 15750 | 7000 | 31850 |
| 45 | 6 | 14 | 5 | 25 | 2700 | 18900 | 11250 | 32850 |
| $n_j$ | 60 | 31 | 9 | 100 | 18520 | 36090 | 18250 | 72860 |

El ajuste lineal tampoco es de calidad por que R² esta más cerca de cero que de la unidad. Además el ajuste exponencial es mejor que el ajuste lineal por que su coeficiente de determinación es mayor (0,26>0,25536).

---

**19. Los ahorros S y los ingresos Y mensuales en cientos de euros de una muestra de 10 familias de una determinada región se presentan en la siguiente tabla:**

| **S** | **1,9** | **1,8** | **2,0** | **2,1** | **1,9** | **2,0** | **2,2** | **2,3** | **2,7** | **3,0** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Y** | **20,5** | **20,8** | **21,2** | **21,7** | **22,1** | **22,3** | **22,2** | **22,6** | **23,1** | **23,5** |

1. **Ajustar los datos anteriores a un modelo lineal que explique lo ahorros familiares en función de los ingresos de la región dada.**
2. **Ajustar los datos anteriores a un modelo lineal parabólico que explique los ahorros familiares en función de los ingresos para la región dada.**
3. **¿Qué ajuste es el mejor?**
4. **¿Qué ahorro se puede prever para una familia de la región que ingrese 2500 euros mensuales?**

---

**SOLUCIÓN:**

Comenzaremos elaborando una tabla de datos adecuada para los cálculos a realizar en el problema.

| $S_I$ | $y_i$ | $S_i y_i$ | $y_i^2$ | $S_i y_i^2$ | $y_i^3$ | $y_i^4$ |
|---|---|---|---|---|---|---|
| 1,9 | 20,5 | 38,95 | 420,25 | 798,475 | 8615,125 | 176610,063 |
| 1,8 | 20,8 | 37,44 | 432,64 | 778,752 | 8998,912 | 187177,37 |
| 2 | 21,2 | 42,4 | 449,44 | 898,88 | 9528,128 | 201996,314 |
| 2,1 | 21,7 | 45,57 | 470,89 | 988,869 | 10218,313 | 221737,392 |
| 1,9 | 22,1 | 41,99 | 488,41 | 927,979 | 10793,861 | 238544,328 |
| 2 | 22,3 | 44,6 | 497,29 | 994,58 | 11089,567 | 247297,344 |
| 2,2 | 22,2 | 48,84 | 492,84 | 1084,248 | 10941,048 | 242891,266 |
| 2,3 | 22,6 | 51,98 | 510,76 | 1174,748 | 11543,176 | 260875,778 |
| 2,7 | 23,1 | 62,37 | 533,61 | 1440,747 | 12326,391 | 284739,632 |
| 3 | 23,5 | 70,5 | 552,25 | 1656,75 | 12977,875 | 304980,063 |
| Suma->21,9 | 220 | 484,64 | 4848,38 | 10744,028 | 107032,296 | 2366849,55 |

El modelo lineal $S_1 = a + by$ puede ajustarse mediante el sistema de ecuaciones normales siguientes:

$$\sum_{i=1}^{10} S_i = Na + b\sum_{i=1}^{10} y_i$$

$$\sum_{i=1}^{10} S_i y_i = a\sum_{i=1}^{10} y_i + b\sum_{i=1}^{10} y_i^2$$

21,9=10a+220b

484,64=220a+4848,38b

a=-5,4

b=0,34

Luego el modelo lineal ajustado será:

$$S_i = -5{,}4 + 0{,}34 y_i$$

Para medir la calidad de ajuste lineal utilizamos el coeficiente de correlación o su cuadrado, el coeficiente de determinación, que se calcula como sigue:

$$r^2 = \frac{\sigma_{YS}^2}{\sigma_y^2 \sigma_s^2} = \frac{0{,}284^2}{0{,}838 * 0{,}1329} = 0{,}72421527$$

Se observa que la calidad del ajuste es buena por que el coeficiente de determinación es alto (el coeficiente de correlación vale $\sqrt{0{,}72421527 = 0{,}851}$ que es un valor elevado indicativo de alto grado de relación entre el ahorro y la renta de las familias.

---

**20. La inversión K y el producto interior bruto y se relacionan mediante la expresión $y = ak^c$. Se pide ajustar una función Cobb-Douglas a los datos siguientes:**

| $y_i$ | **2,6** | **2,9** | **3,4** | **4,1** | **5,1** | **6,0** | **7,2** | **9,2** | **11,2** | **13,1** | **15,2** | **17,3** | **19,9** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $K_i$ | **0,6** | **0,6** | **0,8** | **1,0** | **1,3** | **1,4** | **1,6** | **1,9** | **2,2** | **2,5** | **2,9** | **3,5** | **3,9** |

---

**SOLUCIÓN:**

Se trata de un ajuste tipo potencial. Todo este tipo de ajustes se resuelve aplicando logaritmos para linea rizar de la siguiente forma:

$$y = aK^c \Rightarrow \underbrace{Log(y)}_{Z} = \underbrace{Log(a)}_{A} + c\underbrace{Log(K)}_{X} \Rightarrow Z = A + cx$$

11,2137= 13 A + 2,54022 c

3,18349 = 2,54022 A + 1,3963c

A=0,6471

C=1,1

Luego el modelo de Cobb-Douglas ajustado será:

$$y = 4{,}4k^{1{,}1}$$

Para medir la calidad del ajuste potencial medimos la del ajuste lineal al que es equivalente utilizando el coeficiente de correlación o su cuadrado el coeficiente de determinación se calcula:

$$r^2 = \frac{\sigma_{xZ}^2}{\sigma_x^2 \sigma_Z^2} = \frac{0{,}763^2}{0{,}0692 * 0{,}0851} = 0{,}98$$

Se observa que la calidad del ajuste lineal es buena por que el coeficiente de determinación es alto( el coeficiente de correlación vale $\sqrt{0{,}98 = 0{,}994}$ , que es un valor elevado indicativo del alto grado de relación entre Z y Xi.

---

**21. La siguiente tabla muestra el número de gérmenes patógenos por centímetro cúbico de un determinado cultivo según el tiempo transcurrido:**

| **Nº de Horas** | **0** | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|---|
| **Nº de gérmenes** | **20** | **26** | **33** | **41** | **47** | **53** |

a) **Calcúlala recta de regresión para predecir el numero de gérmenes por cm$^3$ en función del tiempo.**
b) **¿Qué cantidad de gérmenes por cm$^3$ es predecible encontrar cuando hayan transcurrido 6 horas? ¿Es buena esa predicción?**

---

**SOLUCIÓN:**

a) $y = 19{,}81 + 6{,}74x$, donde: $x \rightarrow$ número de horas, $y \rightarrow$ número de gérmenes
b) $\hat{y}(6) = 60{,}25 \approx 60$ *gérmenes*

Es una buena predicción, puesto que $r = 0{,}999$ (y 6 está cercano al intervalo de valores considerado

---

**22. En un depósito cilíndrico, la altura del agua que contiene varia conforme pasa el tiempo según esta tabla:**

| **Tiempo (h)** | **8** | **22** | **27** | **33** | **50** |
|---|---|---|---|---|---|
| **Altura (m)** | **17** | **14** | **12** | **11** | **6** |

a) **Halla el coeficiente de correlación lineal entre el tiempo y la altura e interprétalo.**

**r=-0,997. Hay una relación muy fuerte entre las dos variable, y negativa. A medida que pasa el tiempo la altura va bajando (se va consumiendo el agua)**

**b) ¿Cual será la altura del agua cuando hayan transcurrido 40 horas?**
**c) Cuando la altura del agua es de 2m, suena una alarma¿Qué tiempo ha de pasar para que avise la alarma?**

---

**SOLUCIÓN:**

a) $r = -0{,}997$. Hay una relación muy fuerte entre dos variables, y negativa. A medida que pasa el tiempo, la altura va bajando (se va consumiendo el agua).
b) La recta de regresión es $y = 19{,}37 - 0{,}26x$, donde $x \rightarrow tiempo$, $y \rightarrow altura$.
$\hat{y}(40) = 8{,}97m$
c) $2 = 19{,}37 - 0{,}26x \rightarrow x = 66{,}8h$

---

**23. En una cofradía de pescadores las capturas registradas de cierta variedad de pescados, en kilogramos y el precio de subasta en lonja, en euros/kg, fueron los siguientes:**

| X(kg) | 2000 | 2400 | 2500 | 3000 | 2900 | 2800 | 3160 |
|---|---|---|---|---|---|---|---|
| Y(euros/kg) | 1,80 | 1,68 | 1,65 | 1,32 | 1,44 | 1,50 | 1,20 |

**a) ¿Cual es el precio medio registrado?**
**b) Halla el coeficiente de correlación lineal e interprétalo.**
**c) Estima el precio que alcanzaría en la lonja el kilote esa especie si se pescasen 2600kg**

---

**SOLUCIÓN:**

a) $\bar{y} = 1{,}51$
b) $r = -0{,}97$. La relación entre las variables es fuerte y negativa. A mayor cantidad de pescado, menos es el precio por kilo.
c) La recta de regresión es $y = 2{,}89 - 0{,}0005x$
$\hat{y}(2600) = 1{,}59\ euros$

---

**24. Las calificaciones de 40 alumnos obtenidas en el examen parcial (x) y en el examen final (Y) de una asignatura han sido las siguientes:**

| X | Y | X | Y | X | Y | X | Y |
|---|---|---|---|---|---|---|---|
| 4 | 3 | 8 | 9 | 8 | 7 | 2 | 0 |
| 5 | 8 | 0 | 3 | 9 | 6 | 5 | 3 |
| 1 | 3 | 2 | 3 | 9 | 10 | 4 | 6 |
| 6 | 3 | 10 | 10 | 8 | 7 | 7 | 5 |
| 1 | 0 | 4 | 8 | 5 | 3 | 6 | 7 |
| 2 | 1 | 8 | 7 | 3 | 2 | 15 | 7 |
| 2 | 0 | 2 | 0 | 4 | 1 | 6 | 4 |
| 4 | 2 | 6 | 3 | 3 | 2 | 3 | 0 |
| 5 | 6 | 6 | 6 | 0 | 0 | 9 | 8 |
| 6 | 5 | 5 | 3 | 2 | 1 | 0 | 10 |

**Formar la tabla estadística de doble entrada.**

---

**SOLUCIÓN:**

Tomando en filas los valores de x y en columnas los valores de y podremos hacer:

| X \ Y | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | I 1 | I 1 | III 3 | I 1 | | | | | | | |
| 1 | | | II 2 | | I 1 | | | | | | |
| 2 | | | | II 2 | I 1 | | | | | | |
| 3 | I 1 | I 1 | I 1 | | I 1 | III 3 | II 2 | | | | |
| 4 | | | | | | | I 1 | | | | |
| 5 | | | | | | | I 1 | I 1 | | | |
| 6 | | | | | I 1 | I 1 | I 1 | | | I 1 | |
| 7 | | | | | | I 1 | I 1 | | III 3 | | |
| 8 | | | | | I 1 | I 1 | | | | I 1 | |
| 9 | | | | | | | | | I 1 | | |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **10** | | | | | | | | I 1 | | I 1 | I 1 | |

**25. Las alturas (x) y los pesos (y) de 20 hombres son los siguientes:**

| **X** | **Y** | **X** | **Y** |
|---|---|---|---|
| **1.72** | **63** | **1.76** | **71** |
| **1.70** | **75** | **1.70** | **70** |
| **1.70** | **68** | **1.69** | **66** |
| **1.68** | **70** | **1.66** | **60** |
| **1.75** | **74** | **1.78** | **74** |
| **1.69** | **72** | **1.74** | **69** |
| **1.71** | **67** | **1.70** | **65** |
| **1.69** | **69** | **1.69** | **71** |
| **1.67** | **70** | **1.71** | **73** |
| **1.74** | **84** | **1.78** | **69** |

**Establecer la distribución correspondiente y hallar las medias aritméticas y las desviaciones estándar marginales.**

**SOLUCIÓN:**

La distribución de frecuencias será la siguiente:

| **Y / X** | **60-65** | **65-70** | **70-75** | **75-80** | **80-85** | **TOTAL** |
|---|---|---|---|---|---|---|
| **1.65-1.70** | I 1 | II 2 | IIII 4 | | | 7 |
| **1.70-1.75** | I 1 | IIII 4 | I 1 | I 1 | I 1 | 9 |
| **1.75-1.80** | | I 1 | III 3 | | | 4 |
| **TOTAL** | 2 | 7 | 9 | 1 | 1 | 20 |

Con las tablas de cálculo correspondiente a las distribuciones marginales podremos calcular las medias y las desviaciones estándar pedidas:

Distribución marginal de Y:

| $L_{i-1}$-$L_i$ | $n_i$ | $y_i$ | $n_i\, y_i$ | $y_i^2 n_i$ |
|---|---|---|---|---|
| 1,65-1,70 | 7 | 1,675 | 11,725 | 19,639 |
| 1,70-1,75 | 9 | 1,725 | 15,525 | 26,781 |
| 1,75-1,80 | 4 | 1,775 | 7,100 | 12,603 |
| **TOTAL** | **20** | | **34,350** | **59,023** |

**Y** = 34,350/20 = 1,7175

$\mathbf{S_y^2}$ = 59,023/20-$(34,350/20)^2$ = 0,0013

$\mathbf{S_y}$ = 0,037.

Distribución marginal de X:

| $L_{i-1}$-$L_i$ | $n_i$ | $y_i$ | $n_i\, y_i$ | $y_i^2 n_i$ |
|---|---|---|---|---|
| 60-65 | 2 | 62.5 | 125._ | 7812.50 |
| 65-70 | 7 | 67.5 | 472.5 | 31893.75 |
| 70-75 | 9 | 72.5 | 652.5 | 47306.25 |
| 75-80 | 1 | 77.5 | 77.5 | 6006.25 |
| 80-85 | 1 | 82.5 | 82.5 | 6806.25 |
| **TOTAL** | **20** | | **1410._** | **99825._** |

**X** = 1410/20 = 70.50

$\mathbf{S_x^2}$ = 99825/20 – $(1410/20)^2$ = 21._

$\mathbf{S_x}$ = 4,58

---

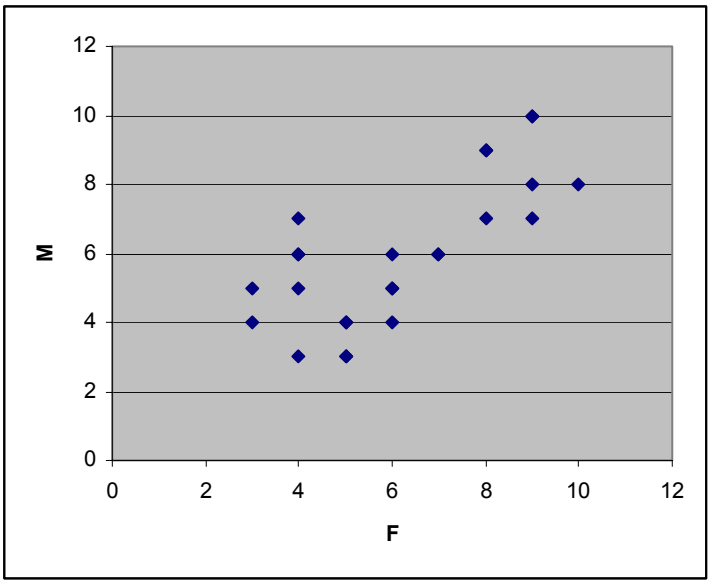
**26. Representar el diagrama de dispersión correspondiente a las notas de 25 alumnos en Física y Matemáticas, siendo éstas:**

| F | M | F | M | F | M |
|---|---|---|---|---|---|
| 7 | 6 | 5 | 4 | 5 | 3 |
| 6 | 6 | 9 | 10 | 4 | 6 |
| 3 | 5 | 10 | 8 | 4 | 7 |
| 4 | 6 | 4 | 5 | 8 | 9 |
| 6 | 5 | 6 | 4 | 9 | 7 |
| 3 | 4 | 9 | 8 | | |
| 5 | 3 | 6 | 5 | | |
| 8 | 9 | 7 | 6 | | |
| 5 | 4 | 4 | 3 | | |
| 9 | 10 | 8 | 7 | | |

---

**SOLUCIÓN:**

Tomando las notas de Física en abscisas y las de Matemáticas en ordenadas, tendremos la siguiente nube de puntos o diagrama de dispersión:



**27. Sea una distribución bidimensional en donde Syx = 4.1, Sy^2 = 9 y el coeficiente de regresión de la recta de Y/X es b = -1,1.**

**Determínese:**

**a) Las dos rectas de regresión de Y/X y X/Y, sabiendo que x = 2, y = 5.**

**b) El coeficiente de correlación lineal.**

---

**SOLUCIÓN:**

a) Los coeficientes de regresión de las serían

Y/X b = -1,1 X/Y b' = Sxy/Sy^2 = 4,1/9 = 0,46

lo cual es imposible, ya que los dos coeficientes de regresión deben ser del mismo signo, puesto que, como

b = Sxy/Sx^2 b' = Sxy/Sy^2

y las varianzas son no negativas, entonces el signo de b y b' debe ser el mismo que la covarianza Sxy.

Como en este caso Sxy = 4,1 > 0, no puede ser b = -1,1, resultado que necesariamente debe estar equivocado.

Aceptando como verdadero valor Sxy = 4,1, lo único que podemos determinar es la recta de regresión de X sobre Y

X/Y x*-x = Sxy/Sy^2(y-y) x*-2 = 4,1/9(y-5) x* = -0,3+0,46y.

b) Por los mismos motivos que antes no se puede determinar r, ya que, como

r = Sxy/SxSy b = -1,1 = 4,1/Sx^2 Sx^2 = 4,1/-1,1<0

lo cual no puede ser.

Este coeficiente r debe tener también el mismo signo que Sxy, b y b'.

---

**28. Estúdiese en cuáles de los casos que a continuación se relacionan los resultados que se ofrecen son compatibles entre si:**

**a) rxy = -0,3 y* = 4x+5.**

**b) Sxy = 100 Sx = 5 Sy^2 = 400 Se^2= 0.**

**c) y = 5x+8 y = 1/5x+9 rxy = 0,2.**

**d) y* = 1/2x+4  x* = y+4  x = 16  y = 12.**

---

**SOLUCIÓN:**

a) Como el coeficiente de regresión de la recta b = 4 es positivo, no puede ser el coeficiente de correlación lineal negativo.

b) El coeficiente de correlación lineal es

r = Sxy/SxSy = 100/5*20 = 1

que también puede expresarse como

r = sqrt(1-Se^2/Sy^2) = sqrt(1-0/400) = 1

en donde Se^2 es la varianza de los residuos o de los errores.

Por tanto, estos resultados sí son consistentes.

c) En este caso existen dos posibilidades:

Caso 1. Que las rectas sean

Y/X  y* = 5x+8  X/Y  x* = 5y-45

con lo que el coeficiente de correlación lineal sería

r = sqrt(b*b') = sqrt(5*5) = 5>1

que es mayor que la unidad, lo cual es imposible.

Caso 2. Que las rectas sean

Y/X  $y^* = 1/5x+9$  X/Y  $x^* = 1/5y-8/5$

y, por tanto

$r = sqrt(b*b') = sqrt(1/5*1/5) = 1/5 = 0,2$

lo que concuerda con lo señalado en el enunciado.

d) Sabemos que el punto de corte entre las dos rectas de regresión debe ser (x,y); para comprobar que en este caso se verifica esta propiedad resolveremos el sistema de ecuaciones formado por estas dos ecuaciones

$y = 1/2x+4$  $2y-x = 8$

$x = y+4$  $-y+x = 4$

de donde

$y = 12 = y$

$x = 4+y = 4+12 = 16 = x$

que son precisamente los dos valores medios que nos ofrecen.

---

**29. Dada la distribución bidimensional**

| | | | | | |
|---|---|---|---|---|---|
| **xi** | **10** | **20** | **30** | **40** | **50** |
| **yj** | **200** | **180** | **150** | **120** | **100** |

**a) Ajústese una recta por el procedimiento de los mínimos cuadrados.**

**b) Calcúlese el coeficiente de correlación lineal y explíquese su significado**

**SOLUCIÓN:**

a) Formemos la siguiente tabla:

| | | | | | | |
|---|---|---|---|---|---|---|
| xi | 10 | 20 | 30 | 40 | 50 | = 150 |
| yj | 200 | 180 | 150 | 120 | 100 | = 750 |
| xi^2 | 100 | 400 | 900 | 1600 | 2500 | = 5500 |
| yj^2 | 40000 | 32400 | 22500 | 14400 | 10000 | = 119300 |
| xiyj | 2000 | 3600 | 4500 | 4800 | 5000 | = 19900 |

La recta de regresión de Y sobre X, ajustada por mínimos cuadrados, es

$y^* = a+bx$

siendo

$b = Sxy/Sx^2 \qquad a = y-bx$

Determinemos las medias, varianzas y covarianzas

$x = Exi/N = 150/5 = 30 \qquad y = Eyj/N = 750/5 = 150$

$Sx^2 = a20- a10^2 \quad a10 = x = 30 \quad a20 = Exi^2/N = 5500/5 = 1100$

$Sx^2 = 1100-30^2 = 200$

$Sy^2 = a02- a01^2 \quad a01 = y = 150 \quad a02 = Eyj^2/N = 119300/5 = 23860$

$Sy^2 = 23860-150^2 = 1360$

$Sxy = a11- a10*a01 \quad a11 = Eexiyj/N = 19900/5 = 3980$

$Sxy = 3980-30*150 = -520.$

Por tanto,

$b = Sxy/Sx^2 = -520/200 = -2,6 \quad a = y-bx = 150-(-2,6)*30 = 228.$

De donde la recta ajustada es

$y = 228-2,6.$

b) El coeficiente de correlación lineal es

r = Sxy/SxSy = -520/sqrt200*sqrt1360 = -0,99.

Como el coeficiente de correlación es negativo, nos indica que la asociación es de tipo inverso; al estar muy próximo a -1, podemos decir que el grado de asociación lineal es muy fuerte y que, por lo tanto, el poder explicativo de la variable X sobre la variable Y es muy grande.

---

**30. En un determinado sector, la producción y las exportaciones durante los últimos años han sido:**

| **Años** | **1982** | **1983** | **1984** | **1985** | **1986** |
|---|---|---|---|---|---|
| **Producción (10^6 Ptas.)** | **400** | **420** | **440** | **480** | **500** |
| **Exportaciones (10^6 Ptas.)** | **80** | **80** | **90** | **92** | **98** |

**a) Si se estima que la producción en el ejercicio 1988 va a ser de 640 millones de pesetas y que las condiciones del mercado internacional no van a variar, ¿cuál será el volumen de expotación previsible?**

**b) ¿En qué medida esta prevsión puede ser o no aceptable?**

---

**SOLUCIÓN:**

a) No es difícil defender la hipótesis de que el volumen de expotaciones es una variable que depende de la producción. Si las exportaciones las representamos por X y la producción por Y, la especificación lineal de esta hipótesis viene dada por

x = a+by

Para estimar por mínimos cuadrados los parametros a y b, formaremos la tabla

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| xj | 80 | 80 | 90 | 92 | 98 | = | 440 |
| yi | 400 | 420 | 440 | 480 | 500 | = | 2240 |
| xj^2 | 6400 | 6400 | 8100 | 8464 | 9604 | = | 38968 |
| yi^2 | 160000 | 176400 | 193600 | 230400 | 250000 | = | 1010400 |
| xjyi | 32000 | 33600 | 39600 | 44100 | 49000 | = | 198360 |

Como

x = Exj/N = 440/5 = 88  y = Eyi/N = 2240/5 = 448

Sx^2 = a02-a01^2  a01 = x = 88  a02 = Exj^2/N = 38968/5 = 7793,6

Sx^2 = 7793,6-88^2 = 49,6

Sy^2 = a20-a10^2  a10 = y = 448  a20 = Eyi^2/N = 1010400/5 = 202080

Sy^2 = 202080-448^2 = 1376

a11 = EExjyi/N = 198360/5 = 39672

Sxy = 39672-448*88 = 248

Tendremos que

b = Sxy/Sy^2 = 248/1376 = 0,18  a = x-by = 88-0,18*448 = 7,36

El modelo ajustado es

x = 7,36+0,18y

Se estima que la producción en 1988 va ser de 640 millones de pesetas y que las condiciones del mercado internacional no cambian. Esta última hipótesis nos faculta para poder seguir utilizando el modelo lineal ajustado por tanto,

x = 7,36+0,18*640 = 122,56 millones de pesetas.

Las exportaciones se situarán, pues, sobre los 122,56 millones de pesetas.

b) Para estudiar la bondad de la predicción calcularemos previamente el coeficiente de correlación lineal,

r = Sxy/SxSy = 248/sqrt49,6*sqrt1376 = 0,95

Estadísticamente, al ser elevado el grado de asociación lineal entre las variables, debemos aceptar como muy posible el resultado.

---

**31. En un determinado estudio médico se pretende medir la relación existente entre la exposición al ruido y la hipertensión. Los siguientes datos han sido extraídos del Jourrnal of Sound and Vibration:**

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Y** | **1** | **0** | **1** | **2** | **5** | **1** | **4** | **6** | **2** | **3** | **5** | **4** | **6** | **8** | **4** | **5** |
| **X** | **60** | **63** | **65** | **70** | **70** | **70** | **80** | **80** | **80** | **80** | **85** | **89** | **90** | **90** | **90** | **90** |

| | | | | |
|---|---|---|---|---|
| **Y** | **7** | **9** | **7** | **6** |
| **X** | **94** | **100** | **100** | **100** |

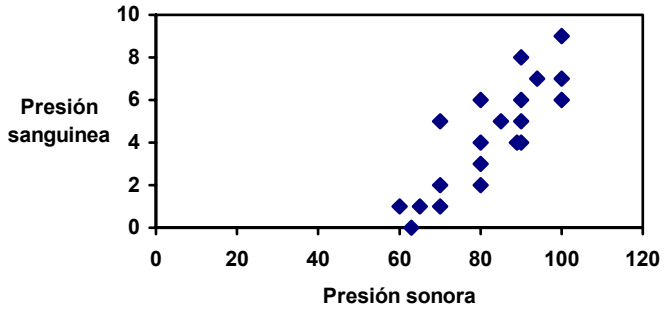
**Donde X representa la presión sonora en dB, e Y el aumento de la presión sanguínea en mmHg.**

1) **Realizar un diagrama de dispersión de Y frente a X.**
2) **Realizar el modelo de regresión lineal simple.**

---

**SOLUCIÓN:**

1) A partir de los datos experimentales que nos proporcionan, obtenemos el siguiente gráfico de dispersión:




2) Obtenemos las medidas muestrales:

$$\bar{x} = 82.3$$
$$\bar{y} = 4.3$$

Y las varianzas y covarianza muestral:

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i = 158.432$$

$$S_y^2 = \frac{1}{n}\sum_{i=1}^{k}(y_i - \bar{y})^2 n_i = 6.537$$

$$S_{xy} = \frac{1}{n}\sum_{i}\sum_{j}(x_i - \bar{x})(y_j - \bar{y})n_{ij} = 27.168$$

Por lo que la ecuación de la recta de regresión es:

$$y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x})$$

$$y = 0.171x - 9.813$$

---

**32. Sea (X,Y) una variable aleatoria bidimensional con función de densidad conjunta**

$f(x,y) = xy$ **si** $0 \le x \le 1$, $0 \le y \le 1$

**Obtenga la recta de regresión de Y sobre X.**

---

**SOLUCIÓN:**

Las correspondientes funciones de densidad marginales son:

$$fX(x) = \int_0^1 f(x,y)\partial y = \int_0^1 xy\partial y = \frac{x}{2}$$

$$fY(y) = \int_0^1 f(x,y)\partial x = \int_0^1 xy\partial x = \frac{y}{2}$$

Se obtiene entonces:

$$\alpha_{10} = E[X] = \int_0^1 xfX(x)\partial x = \frac{1}{6} = E[Y] = \alpha_{01}$$

$$E[X^2] = \int_0^1 x^2 fX(x)\partial x = \frac{1}{8} = E[Y^2]$$

y por lo tanto:

$$\mu_{20} = \sigma_x^2 = E(X^2) - [E(X)]^2 = \frac{7}{72}$$

Además: $\alpha_{11} = E[XY] = \int_0^1 \int_0^1 xyf(x,y)\partial y\partial x = \frac{1}{9}$

$$\mu_{11} = Cov = \alpha_{11} - \alpha_{10} \cdot \alpha_{01} = \frac{1}{9} - \frac{1}{6}\cdot\frac{1}{6} = \frac{1}{9} - \frac{1}{36}$$

La recta de regresión de Y sobre X es, por lo tanto:

$$y - \alpha_{01} = \frac{\mu_{11}}{\mu_{20}}(x - \alpha_{10})$$

$$y - \frac{1}{6} = \frac{\frac{1}{9} - \frac{1}{36}}{\frac{7}{72}}\left(x - \frac{1}{6}\right)$$

es decir: $$y - \frac{1}{6} = \frac{6}{7}\left(x - \frac{1}{6}\right)$$

---

**33. Sea la variable aleatoria bidimensional (X,Y) que asigna probabilidades iguales a los puntos: (1,1); (2,3); (3,2); (4,4); obtener la recta de regresión mínimo cuadrática de Y sobre X.**

---

**SOLUCIÓN:**

Recta de regresión de Y sobre X:

$$y - \alpha_{01} = \frac{\mu_{11}}{\mu_{20}}(x - \alpha_{10})$$

$$\alpha_{10} = E[X] = \sum_{i=1}^{4} x_i P(X = x_i) = 1\frac{1}{4} + 2\frac{1}{4} + 3\frac{1}{4} + 4\frac{1}{4} = \frac{10}{4} = 2.5$$

$$\alpha_{01} = E[Y] = \sum_{j=1}^{4} y_j P(Y = y_j) = 2.5$$

$$E[X^2] = \sum_{i=1}^{4} x_i^2 P(X = x_i) = \frac{1}{4}(1^2 + 2^2 + 3^2 + 4^2) = \frac{30}{4}$$

$$\mu_{20} = \sigma_x^2 = E\left(X^2\right) - \left[E(X)\right]^2 = \frac{30}{4} - (2.5)^2 = 1.25$$

$$\alpha_{11} = E[XY] = \sum_{i=1}^{4}\sum_{j=1}^{4} x_i y_j P\left(X = x_i, Y = y_j\right) = 1{\cdot}1\frac{1}{4} + 2{\cdot}3\frac{1}{4} + 3{\cdot}2\frac{1}{4} + 4{\cdot}4\frac{1}{4} = \frac{29}{4} = 7.25$$

$$\mu_{11} = Cov = \alpha_{11} - \alpha_{10}{\cdot}\alpha_{01} = 7.75 - 2.5{\cdot}2.5 = 1$$

La recta de regresión de Y sobre X es:

$$y - 2.5 = \frac{1}{1.25}(x - 2.5)$$

$$y = 0.8x + 0.5$$

---

**34. Las notas obtenidas por 10 alumnos en matemáticas y música son:**

| **Alum.** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Mat.** | **6** | **4** | **8** | **5** | **3.5** | **7** | **5** | **10** | **5** | **4** |
| **Mús.** | **6.5** | **4.5** | **7** | **5** | **4** | **8** | **7** | **10** | **6** | **5** |

**Calcular la covarianza, correlación y rectas de regresión.**

---

**SOLUCIÓN:**

Indiquemos por X la nota de matemáticas y por Y la nota de música.

Medias: $\bar{x} = \frac{1}{n}\sum x_i = 5.75$

$$\bar{y} = \frac{1}{n}\sum y_i = 6.3$$

Covarianza: $S_{xy} = \frac{1}{n}\sum x_i y_i - \overline{xy} = 3.075$

Varianzas:

$$S_x^2 = \frac{1}{n}\sum x_i^2 - \overline{x}^2 = 3.763$$

$$S_y^2 = \frac{1}{n}\sum y_i^2 - \overline{y}^2 = 2.96$$

Coef. de correlación: $r = \frac{S_{xy}}{S_x S_y} = \frac{3.075}{\sqrt{3.763 \cdot 2.96}} = 0.9214$

Coef. de regresión:

$$b_{21} = \frac{S_{xy}}{S_x^2} = 0.817$$

$$b_{12} = \frac{S_{xy}}{S_y^2} = 1.039$$

Recta de regresión de Y sobre X: $y - \overline{y} = b_{21}(x - \overline{x})$

$$y - 6.3 = 0.817(x - 5.75)$$

Recta de regresión de X sobre Y: $y - \overline{y} = b_{12}(x - \overline{x})$

$$x - 5.75 = 1.039(y - 6.3)$$

---

**35. Para realizar un estudio sobre la utilización de una impresora en un determinado departamento, se midió en un día los minutos transcurridos entre las sucesivas utilizaciones (X) y el número de páginas impresas (Y) obteniéndose los siguientes resultados:**

| | |
|---|---|
| **X** | **9 9 4 6 8 9 7 6 9 9 9 8 8 9 8 9 9 9 10 9 15 10 12 12 10 10 12 10 10 12 12 10** |
| **Y** | **3 8 3 8 3 8 8 8 3 8 12 12 8 8 8 12 12 20 8 20 8 8 20 8 8 12 8 20 20 3 3 20** |

a) **Escribir la distribución de frecuencias conjunta. ¿Cuál es el porcentaje de veces que transcurre más de nueve minutos desde la anterior utilización y se imprimen menos de 12 páginas? ¿Cuántas veces se imprimen menos de 12 páginas y transcurren 9 minutos desde la anterior utilización?**
b) **Frecuencias marginales. ¿Cuantas veces se imprimen como mucho 12 páginas? ¿Cuántas páginas como mucho se imprimen en el 80 % de las ocasiones?**
c) **Dibujar el diagrama de dispersión.**

---

**SOLUCIÓN:**

a) **Escribir la distribución de frecuencias conjunta. ¿Cuál es el porcentaje de veces que transcurre más de nueve minutos desde la anterior utilización y se imprimen menos de 12 páginas? ¿Cuántas veces se imprimen menos de 12 páginas y transcurren 9 minutos desde la anterior utilización?**

| $x_i \backslash y_j$ | 3 | 8 | 12 | 20 | $n_{i.}$ | $f_{i.}$ |
|---|---|---|---|---|---|---|
| 4 | 1/0,03 | - | - | - | 1 | 0,03 |
| 6 | - | 2/0,06 | - | - | 2 | 0,06 |
| 7 | - | 1/0,03 | - | - | 1 | 0,03 |
| 8 | 1/0,03 | 2/0,06 | 1/0,03 | - | 4 | 0,12 |
| 9 | 2/0,06 | 4/0,12 | 3/0,09 | 2/0,06 | 11 | 0,34 |
| 10 | - | 3/0,09 | 1/0,03 | 3/0,09 | 7 | 0,22 |
| 12 | 2/0,06 | 2/0,06 | - | 1/0,03 | 5 | 0,16 |
| 15 | - | 1/0,03 | - | - | 1 | 0,03 |
| $n_{.j}$ | 6 | 15 | 5 | 6 | 32 | - |
| $f_{.j}$ | 0,19 | 0,47 | 0,16 | 0,19 | - | 1 |

Más de 9 min. → 13 → Menos de 12 Pág. → 8 → $\frac{8}{32} = 0.25 \Rightarrow 25\%$

9 min. → 11 → Menos de 12 Pág. → 6 → $\frac{6}{32} = 0.19 \Rightarrow 19\%$

b) **Frecuencias marginales. ¿Cuantas veces se imprimen como mucho 12 páginas? ¿Cuántas páginas como mucho se imprimen en el 80 % de las ocasiones?**

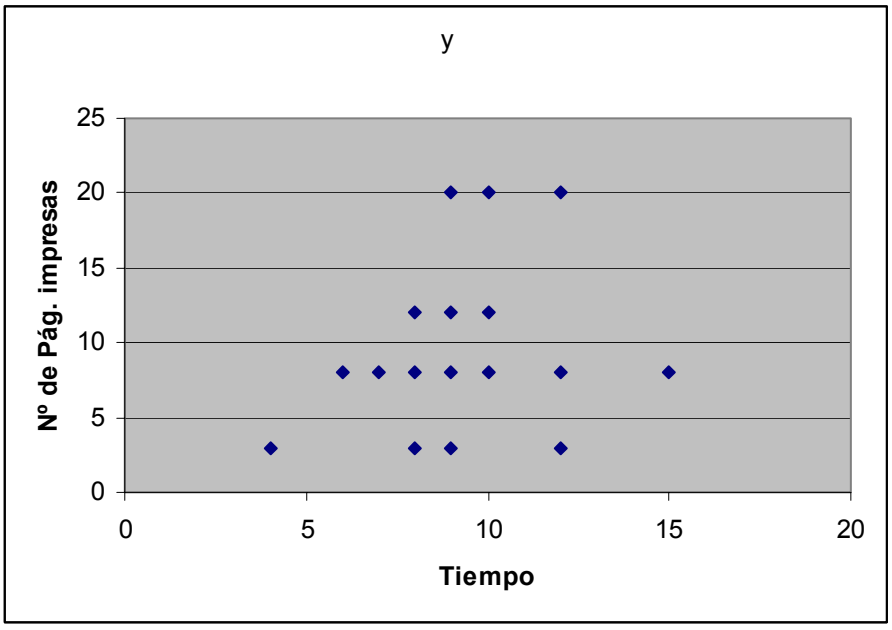
| $Y_j$ | 3 | 8 | 12 | 20 | |
|---|---|---|---|---|---|
| $n_{.j}$ | 6 | 15 | 5 | 6 | 32 |
| $f_{.j}$ | 0,19 | 0,47 | 0,16 | 0,19 | |
| $N_{.j}$ | 6 | 21 | 26 | 32 | |
| $F_{.j}$ | 0.19 | 0.66 | 0.81 | 1 | |

Como mucho 12 Pág. → 6+15+5 = 26

$$N_{i-1} \angle \frac{n*k}{100} \le N_i \Rightarrow 21 \angle \frac{32*80}{100} = 25.6 \le 26$$

$$P_{80} = 12 Pág.$$

**c) Dibujar el diagrama de dispersión.**




---

**36. Se midió el tiempo en segundos que tardaron en grabarse los mismos 24 ficheros en cada uno de los dos tipos de discos ($3^{1/3}$ y $5^{1/4}$). Los tiempos observados fueron:**

| | |
|---|---|
| $3^{1/3}$ | **1.2 1 1.1 0.5 1.1 1.5 1 1.4 1.4 1.3 0.4 1.2 0.4 0.3 0.3 1.5 1.4 1.1 1.2 1.2 0.4 0.5 1.3 1.5** |
| $5^{1/4}$ | **1.3 1.1 1.2 0.4 1.2 1.4 1.1 1.6 1.6 1.5 0.4 1.5 0.4 0.3 0.3 1.6 1.3 1.1 1.3 1.1 0.4 0.4 1.4 1.6** |

**a) Construye la tabla de frecuencias conjuntas. ¿Cuál es el porcentaje de ficheros que tardan menos de 1.5 segundos en el primer tipo de disco y más de 1.4 en el segundo? ¿Cuántos ficheros tardan en grabarse entre 0.6 y 1.2 segundos en el primer tipo de disco? ¿Cuánto tiempo tarda como mucho en grabarse al menos el 90,5 de los ficheros en el segundo tipo de disco?**

**b) Hallar la tabla de frecuencias condicionales de los tiempos en el disco de $5^{1/4}$ de aquellos programas que tardaron 1.2 en el disco de $3^{1/3}$ ¿Cuál es la proporción de estos programas que tardan en grabarse más de 1.5 segundos en el de $5^{1/4}$?**

**c) Representar gráficamente los datos y comentar el gráfico obtenido.**

**d) Si un fichero tarda 0.8 segundos en grabarse en el primer tipo de disco, ¿cuantos segundos tardará en grabarse en el segundo tipo? Da una medida de fiabilidad. ¿Confirma esta medida lo comentado en el apartado c?**

---

**SOLUCIÓN:**

**a) Construye la tabla de frecuencias conjuntas. ¿Cuál es el porcentaje de ficheros que tardan menos de 1.5 segundos en el primer tipo de disco y más de 1.4 en el segundo? ¿Cuántos ficheros tardan en grabarse entre 0.6 y 1.2 segundos en el primer tipo de disco? ¿Cuánto tiempo tarda como mucho en grabarse al menos el 90,5 de los ficheros en el segundo tipo de disco?**

| $x_i \backslash y_j$ | 0,3 | 0,4 | 1,1 | 1,2 | 1,3 | 1,4 | 1,5 | 1,6 | $n_{i.}$ | $f_{i.}$ | $N_{i.}$ | $F_{i.}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0,3 | 2/0,08 | - | - | - | - | - | - | - | 2 | 0,083 | 2 | 0,083 |
| 0,4 | - | 3/0,125 | - | - | - | - | - | - | 3 | 0,125 | 5 | 0,208 |
| 0,5 | - | 2/0,08 | - | - | - | - | - | - | 2 | 0,083 | 7 | 0,292 |
| 1 | - | - | 2/0,08 | - | - | - | - | - | 2 | 0,083 | 9 | 0,375 |
| 1,1 | - | - | 1/0,04 | 2/0,08 | - | - | - | - | 3 | 0,125 | 12 | 0,500 |
| 1,2 | - | - | 1/0,04 | 2/0,08 | - | - | 1/0,04 | - | 4 | 0,167 | 16 | 0,667 |
| 1,3 | - | - | - | - | - | 1/0,04 | 1/0,04 | - | 2 | 0,083 | 18 | 0,750 |
| 1,4 | - | - | - | - | 1/0,04 | - | - | 2/0,08 | 3 | 0,125 | 21 | 0,875 |
| 1,5 | - | - | - | - | - | 1/0,04 | - | 2/0,08 | 3 | 0,125 | 24 | 1,000 |
| $n_{.j}$ | 2 | 5 | 4 | 4 | 1 | 2 | 2 | 4 | 24 | - | | |
| $f_{.j}$ | 0,083 | 0,208 | 0,167 | 0,167 | 0,042 | 0,083 | 0,083 | 0,167 | - | 1 | | |
| $N_{.j}$ | 2 | 7 | 11 | 15 | 16 | 18 | 20 | 24 | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $F_{.j}$ | 0,083 | 0,292 | 0,458 | 0,625 | 0,667 | 0,750 | 0,833 | 1,000 |

Menos de 1.5 seg. En x → 21 → Mas de 1.4 seg. En Y → 4

$$\frac{4}{24} = 0.17 \Rightarrow 17\%$$

Entre 0.6 y 1.2 seg. En X → $\frac{9}{24} = 0.375 \Rightarrow 37.5\%$

$$N_{i-1} \angle \frac{n * k}{100} \leq N_i \Rightarrow 20 \angle \frac{24 * 90.5}{100} = 21.72 \leq 24$$

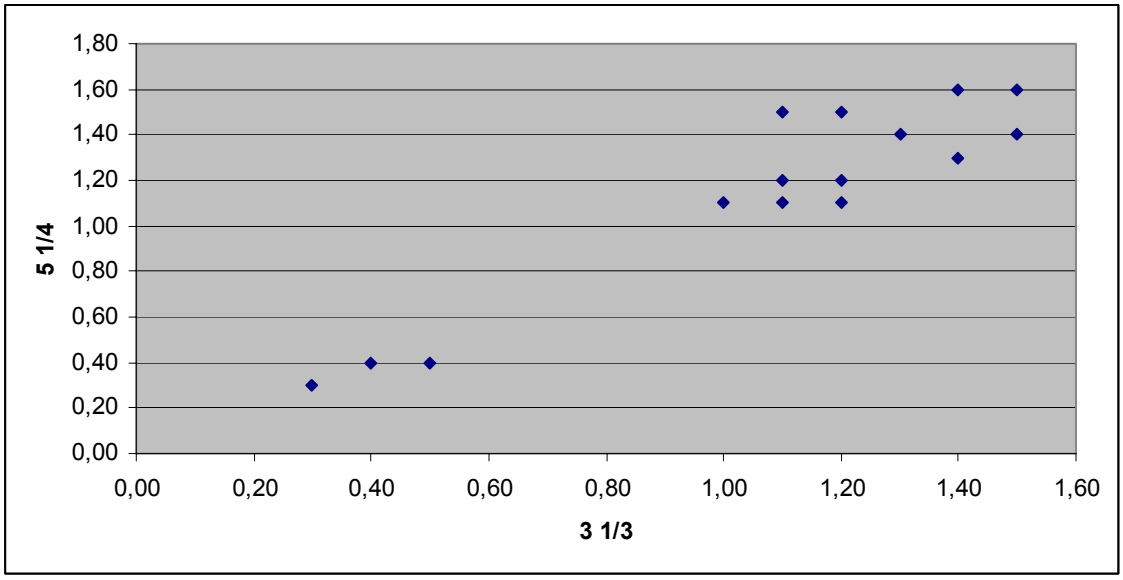
$$P_{90.5} = 1.6 seg.$$

**b) Hallar la tabla de frecuencias condicionales de los tiempos en el disco de $5^{1/4}$ de aquellos programas que tardaron 1.2 en el disco de $3^{1/3}$ ¿Cuál es la proporción de estos programas que tardan en grabarse más de 1.5 segundos en el de $5^{1/4}$?**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $Y_j$ | 0,3 | 0,4 | 1,1 | 1,2 | 1,3 | 1,4 | 1,5 | 1,6 | |
| $n_{.j}$ | 0 | 0 | 1 | 2 | 0 | 0 | 1 | 0 | 4 |
| $f_j$ | 0 | 0 | 0,25 | 0,5 | 0 | 0 | 0,25 | 0 | |
| $N_{.j}$ | 0 | 0 | 1 | 3 | 3 | 3 | 4 | 4 | |
| $F_{.j}$ | 0 | 0 | 0,25 | 0,75 | 0,75 | 0,75 | 1 | 1 | |

Mas de 1.5 seg → 0 → 0%

**c) Representar gráficamente los datos y comentar el gráfico obtenido.**




Se puede observar como los puntos describen una línea recta difusa.

**d) Si un fichero tarda 0.8 segundos en grabarse en el primer tipo de disco, ¿cuantos segundos tardará en grabarse en el segundo tipo? Da una medida de fiabilidad. ¿Confirma esta medida lo comentado en el apartado c?**

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} * (x - \bar{x})$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i * n_i}{n} \Rightarrow \bar{x} = 1.008$$

$$\bar{y} = \frac{\sum_{j=1}^{n} y_j * n_j}{n} \Rightarrow \bar{y} = 0.97$$

$$Sn_x^2 = \frac{\sum_{i=1}^{n} x_i^2 * n_i}{n} - \bar{x}^2 \Rightarrow Sn_x^2 = 0.1739$$

$$Sn_x = \sqrt{Sn_x^2} = 0.417$$

$$Sn_y^2 = \frac{\sum_{j=1}^{n} y_j^2 * n_j}{n} - \bar{y}^2 \Rightarrow Sn_y^2 = 0.3895$$

$$Sn_y = \sqrt{Sn_y^2} = 0.624$$

$$S_{xy} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} n_{ij} * x_i * y_j}{n} - \bar{x} * \bar{y} = 0.1756$$

$$y - 0.97 = \frac{0.1756}{0.1739} * (x - 1.008) \rightarrow y = 1.00977 * x - 0.0478$$

Si x = 0.8 → y = 0.75996

Medida de fiabilidad

$$r_{xy} = \frac{S_{xy}}{S_x * S_y} = 0.6748$$

Es una medida de fiabilidad mala puesto que no llega al 70 %, aunque este cerca.

---

**37. Las siguientes son las calificaciones obtenidas por los 25 alumnos de un grupo de Bachillerato en las asignaturas de Biología y Química:**

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **B** | **4** | **5** | **5** | **5** | **6** | **6** | **7** | **7** | **7** | **7** | **7** | **7** | **7** | **8** | **8** | **8** | **8** | **8** | **8** | **9** | **9** | **9** | **9** | **10** |
| **Q** | **3** | **5** | **5** | **6** | **7** | **7** | **7** | **7** | **7** | **7** | **8** | **8** | **8** | **7** | **7** | **8** | **8** | **8** | **8** | **8** | **8** | **8** | **10** | **10** | **10** |

a) **Obtener la tabla de frecuencias conjunta.**
b) **¿Qué proporción de alumnos obtienen más de un cinco en ambas asignaturas? ¿Qué proporción de alumnos obtienen más de un cinco en Biología? ¿Qué proporción de alumnos obtienen más de un cinco en Químicas?**
c) **Hallar la distribución de frecuencias condicionales de la calificación en Biología de los estudiantes que obtuvieron un 7 en Químicas. ¿Qué proporción de estos estudiantes obtuvieron notable en Biología?**
d) **Representar gráficamente. Comentar el resultado.**
e) **Hallar el coeficiente de correlación. Comentar el resultado.**

---

**SOLUCIÓN:**

a) **Obtener la tabla de frecuencias conjunta.**

| $x_i \backslash y_j$ | 3 | 5 | 6 | 7 | 8 | 10 | $n_{i.}$ | $f_{i.}$ | $N_{i.}$ | $F_{i.}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 1/0,04 | - | - | - | - | - | 1 | 0,040 | 1 | 0,040 |
| 5 | - | 2/0,08 | 1/0,04 | - | - | - | 3 | 0,120 | 4 | 0,160 |
| 6 | - | - | - | 2/0,08 | - | - | 2 | 0,080 | 6 | 0,240 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 7 | - | - | - | 4/0,16 | 3/0,12 | - | 7 | 0,280 | 13 | 0,520 |
| 8 | - | - | - | 2/0,08 | 4/0,16 | - | 6 | 0,240 | 19 | 0,760 |
| 9 | - | - | - | - | 3/0,12 | 2/0,12 | 5 | 0,200 | 24 | 0,960 |
| 10 | - | - | - | - | - | 1/0,04 | 1 | 0,040 | 25 | 1,000 |
| $n_{.j}$ | 1 | 2 | 1 | 8 | 10 | 3 | 25 | - | | |
| $f_{.j}$ | 0,040 | 0,080 | 0,040 | 0,320 | 0,400 | 0,120 | - | 1 | | |
| $N_{.j}$ | 1 | 3 | 4 | 12 | 22 | 25 | | | | |
| $F_{.j}$ | 0,040 | 0,120 | 0,160 | 0,480 | 0,880 | 1,000 | | | | |

**b) ¿Qué proporción de alumnos obtienen más de un cinco en ambas asignaturas? ¿Qué proporción de alumnos obtienen más de un cinco en Biología? ¿Qué proporción de alumnos obtienen más de un cinco en Químicas?**

Más de un 5 en B y en Q $\rightarrow \frac{25-4}{25} * 100 = 84\%$

Más de un 5 en B $\rightarrow \frac{25-4}{25} * 100 = 84\%$

Más de un 5 en Q $\rightarrow \frac{25-3}{25} * 100 = 88\%$

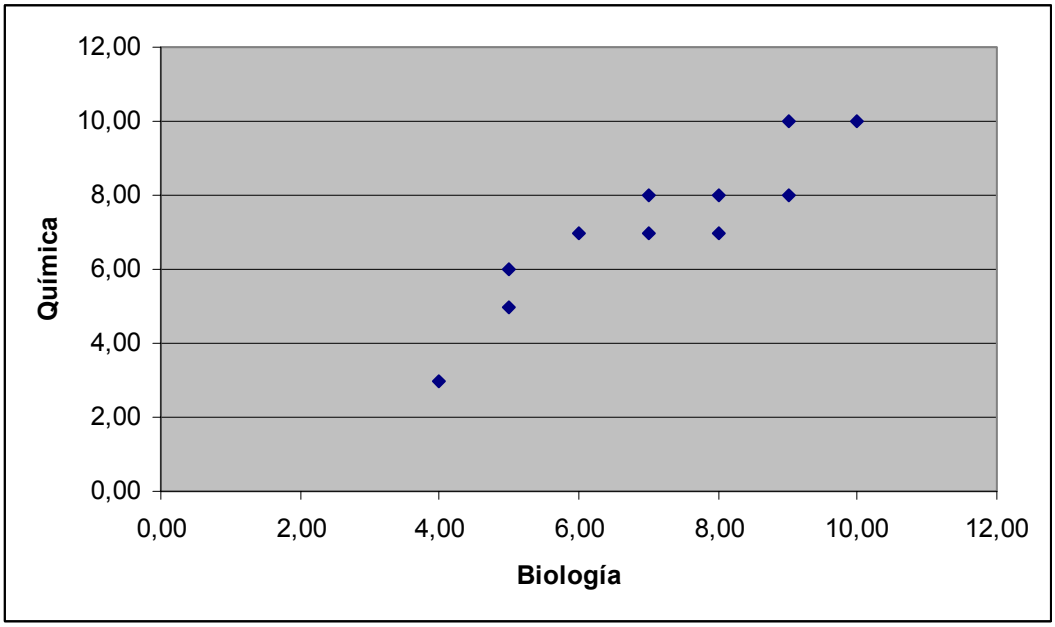
**c) Hallar la distribución de frecuencias condicionales de la calificación en Biología de los estudiantes que obtuvieron un 7 en Químicas. ¿Qué proporción de estos estudiantes obtuvieron notable en Biología?**

| $x_i$ | Y = 7 | $n_{i.}$ | $f_{i.}$ | $N_{i.}$ | $F_{i.}$ |
|---|---|---|---|---|---|
| 4 | - | 0 | 0,000 | 0 | 0,000 |
| 5 | - | 0 | 0,000 | 0 | 0,000 |
| 6 | 2/0,08 | 2 | 0,250 | 2 | 0,250 |
| 7 | 4/0,16 | 4 | 0,500 | 6 | 0,750 |
| 8 | 2/0,08 | 2 | 0,250 | 8 | 1,000 |
| 9 | - | 0 | 0,000 | 8 | 1,000 |

| | | | | | |
|---|---|---|---|---|---|
| 10 | - | 0 | 0,000 | 8 | 1,000 |
| | | 8 | | | |

Notable = 7-8 $\rightarrow \frac{6}{8} * 100 = 75\%$

**d) Representar gráficamente. Comentar el resultado.**




Se puede observar como a mas nota en biología se tiende a sacar mas nota en química, es una relación lineal ascendente.

**e) Hallar el coeficiente de correlación. Comentar el resultado.**

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i * n_i}{n} \Rightarrow \bar{x} = 7.32$$

$$\bar{y} = \frac{\sum_{j=1}^{n} y_j * n_j}{n} \Rightarrow \bar{y} = 7.4$$

$$Sn_x^2 = \frac{\sum_{i=1}^{n} x_i^2 * n_i}{n} - \bar{x}^2 \Rightarrow Sn_x^2 = 2.2176$$

$$Sn_x = \sqrt{Sn_x^2} = 1.489$$

$$Sn_y^2 = \frac{\sum_{j=1}^{n} y_j^2 * n_j}{n} - \bar{y}^2 \Rightarrow Sn_y^2 = 2.32$$

$$Sn_y = \sqrt{Sn_y^2} = 1.523$$

$$S_{xy} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} n_{ij} * x_i * y_j}{n} - \bar{x} * \bar{y} = 1.992$$

$$r_{xy} = \frac{S_{xy}}{S_x * S_y} = 0.878$$

Aquí observamos lo que ya se había comentado al ver el gráfico, y es que hay una tendencia lineal ascendente, por lo que a mayores notas en Biología, ese mismo alumno, tendrá mayores notas de Química.

---

**38. Los siguientes datos corresponden a los tiempos en segundos que tardaron en ejecutarse seis programas elegidos al azar en el entorno Windows y en DOS:**

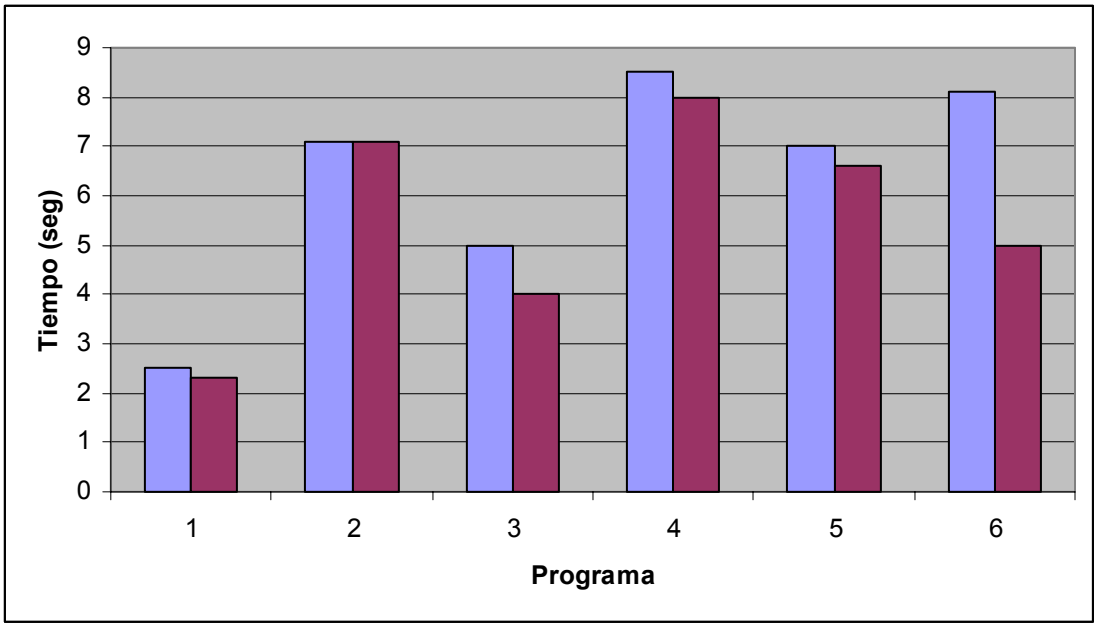
| | **Programa** |
|---|---|
| **Windows** | **2.5 7.1 5 8.5 7 8.1** |
| **DOS** | **2.3 7.1 4 8 6.6 5** |

a) **Representar gráficamente los datos.**
b) **Si un programa tarda 3 segundos en ejecutarse en Windows, ¿cuanto tardará en ejecutarse en DOS?**
c) **Si un programa tarda 6 segundos en ejecutarse en DOS, ¿cuanto tardará en ejecutarse en Windows?**
d) **Dar una medida de fiabilidad de los anteriores resultados.**

---

**SOLUCIÓN:**

a) **Representar gráficamente los datos.**




**b) Si un programa tarda 3 segundos en ejecutarse en Windows, ¿cuanto tardará en ejecutarse en DOS?**

Datos de Windows:

$$\mathrm{Re} = 8.5 - 2.5 = 6$$

$$n = 6 \leq 50 \Rightarrow m = \sqrt{n} = \sqrt{6} = 2.5 \Rightarrow C_i = 3$$

$$a_i = \frac{\mathrm{Re}}{C_i} = \frac{6}{3} = 2 \Rightarrow a_i = 2$$

Datos de Dos:

$$\mathrm{Re} = 8 - 2.3 = 5.7$$

$$n = 6 \leq 50 \Rightarrow m = \sqrt{n} = \sqrt{6} = 2.5 \Rightarrow C_i = 3$$

$$a_i = \frac{\mathrm{Re}}{C_i} = \frac{5.7}{3} = 1.9 \Rightarrow a_i = 2$$

| $x_i \backslash y_j$ | [2,5-4,5) | [4,5-6,5) | [6,5-8,5) | $c_i$ | $n_{i.}$ | $f_{i.}$ | $N_{i.}$ | $F_{i.}$ |
|---|---|---|---|---|---|---|---|---|
| [2,5-4,5) | 1/0,17 | - | - | 3,5 | 1 | 0,167 | 1 | 0,167 |
| [4,5-6,5) | 1/0,17 | - | - | 5,5 | 1 | 0,167 | 2 | 0,333 |
| [6,5-8,5) | - | 1/0,17 | 3/0,5 | 7,5 | 4 | 0,667 | 6 | 1,000 |

| $c_j$ | 3,5 | 5,5 | 7,5 |
|---|---|---|---|
| $n_{.j}$ | 2 | 1 | 3 |
| $f_{.j}$ | 0,333 | 0,167 | 0,500 |
| $N_{.j}$ | 2 | 3 | 6 |
| $F_{.j}$ | 0,333 | 0,500 | 1,000 |

| 6 | - |
|---|---|
| - | 1 |

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} * (x - \bar{x})$$

$$\bar{x} = \frac{\sum_{i=1}^{n} c_i * n_i}{n} \Rightarrow \bar{x} = 6.5$$

$$\bar{y} = \frac{\sum_{j=1}^{n} c_j * n_j}{n} \Rightarrow \bar{y} = 5.83$$

$$Sn_x^2 = \frac{\sum_{i=1}^{n} c_i^2 * n_i}{n} - \bar{x}^2 \Rightarrow Sn_x^2 = 2.33$$

$$Sn_x = \sqrt{Sn_x^2} = 1.527$$

$$Sn_y^2 = \frac{\sum_{j=1}^{n} c_j^2 * n_j}{n} - \bar{y}^2 \Rightarrow Sn_y^2 = 3.26$$

$$Sn_y = \sqrt{Sn_y^2} = 1.8$$

$$S_{xy} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} n_{ij} * x_i * y_j}{n} - \bar{x} * \bar{y} = 2.355$$

$$y - 5.83 = \frac{2.355}{2.33} * (x - 6.5) \rightarrow y = 1.011 * x - 0.7397$$

Si x = 3 → <u>y = 2.29 seg</u>

**c) Si un programa tarda 6 segundos en ejecutarse en DOS, ¿cuanto tardará en ejecutarse en Windows?**

Si y = 6 → x = 6.666 seg

**d) Dar una medida de fiabilidad de los anteriores resultados.**

$$r_{xy} = \frac{S_{xy}}{S_x * S_y} = 0.8568$$

Es una medida de fiabilidad buena puesto que llega al 70 % y lo sobrepasa hasta llegar a un 85.68 %.

---

**39. Un determinado partido político, se plantea el problema de hasta que punto le pueden compensar los gastos de la campaña de propaganda para las futuras elecciones. En las últimas elecciones, los gastos de publicidad y el número de diputados elegidos han sido:**

| Gastos publicidad en miles de pesetas | Diputados elegidos |
|---|---|
| **1500** | **3** |
| **1750** | **4** |
| **3250** | **4** |
| **4000** | **6** |
| **5000** | **8** |

**La comisión electoral está estudiando la posibilidad de un presupuesto de propaganda de diez millones de pesetas.**

**a) ¿Cuál será el número de diputados que serían elegidos de ese partido de acuerdo con este presupuesto, si la imagen del partido no varía respecto a las elecciones anteriores?**
**b) ¿Con qué confianza se puede esperar ese resultado?**
**c) ¿Cuál sería el porcentaje de causas diferentes a la publicidad que influirían en las elecciones?**

---

**SOLUCIÓN:**

**a) ¿Cuál será el número de diputados que serían elegidos de ese partido de acuerdo con este presupuesto, si la imagen del partido no varía respecto a las elecciones anteriores?**

| $x_i \backslash y_j$ | 3 | 4 | 6 | 8 | $n_{i.}$ | $f_{i.}$ | $N_{i.}$ | $F_{i.}$ |
|---|---|---|---|---|---|---|---|---|
| 1500 | 1/0,2 | - | - | - | 1 | 0,200 | 1 | 0,200 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1750 | - | 1/0,2 | - | - | 1 | 0,200 | 2 | 0,400 |
| 3250 | - | 1/0,2 | - | - | 1 | 0,200 | 3 | 0,600 |
| 4000 | - | - | 1/0,2 | - | 1 | 0,200 | 4 | 0,800 |
| 5000 | - | - | - | 1/0,2 | 1 | 0,200 | 5 | 1,000 |
| $n_{.j}$ | 1 | 2 | 1 | 1 | 5 | - | | |
| $f_{.j}$ | 0,200 | 0,400 | 0,200 | 0,200 | - | 1 | | |
| $N_{.j}$ | 1 | 3 | 4 | 5 | | | | |
| $F_{.j}$ | 0,200 | 0,600 | 0,800 | 1,000 | | | | |

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} * (x - \bar{x})$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i * n_i}{n} \Rightarrow \bar{x} = 3100$$

$$\bar{y} = \frac{\sum_{j=1}^{n} y_j * n_j}{n} \Rightarrow \bar{y} = 5$$

$$Sn_x^2 = \frac{\sum_{i=1}^{n} x_i^2 * n_i}{n} - \bar{x}^2 \Rightarrow Sn_x^2 = 1765000$$

$$Sn_x = \sqrt{Sn_x^2} = 1328.53$$

$$Sn_y^2 = \frac{\sum_{j=1}^{n} y_j^2 * n_j}{n} - \bar{y}^2 \Rightarrow Sn_y^2 = 3.2$$

$$Sn_y = \sqrt{Sn_y^2} = 1.789$$

$$S_{xy} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} n_{ij} * x_i * y_j}{n} - \bar{x} * \bar{y} = 2200$$

$$y - 5 = \frac{2200}{1765000} * (x - 3100) \rightarrow y = 1.246 * 10^{-3} * x + 1.136$$

Si x = 10 000 → y = 13.596 Diputados

**b) ¿Con qué confianza se puede esperar ese resultado?**

$$r_{xy} = \frac{S_{xy}}{S_x * S_y} = 0.9256$$

92.56 % de confianza

**c) ¿Cuál sería el porcentaje de causas diferentes a la publicidad que influirían en las elecciones?**

- La cantidad de recursos publicitarios utilizados por otros partidos.
- El número máximo de diputados.

---

**40. La resistencia del papel utilizado en la fabricación de cajas de cartulina (Y) está relacionado con la concentración de madera dura en la pulpa original (X). Bajo condiciones controladas, una planta piloto fabrica 16 muestras con un lote diferente de pulpa y mide la resistencia a la tensión. Los datos obtenidos son los siguientes:**

| **X** | **1** | **1.5** | **1.5** | **1.5** | **2** | **2** | **2.2** | **2.4** | **2.5** | **2.5** | **2.8** | **2.8** | **3** | **3** | **3.2** | **3.3** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Y** | **101.4** | **117.4** | **117.1** | **106.2** | **131.9** | **146.9** | **146.8** | **133.9** | **111** | **123** | **125.1** | **145.1** | **134.3** | **144.5** | **143.7** | **146.9** |

**a) Representar gráficamente los datos y comentar los resultados.**
**b) Hallar el coeficiente de correlación.**
**c) Ajustar un modelo de regresión lineal. Predecir la resistencia de una caja fabricada con pulpa cuya concentración es 2.3.**

---

**SOLUCIÓN:**

**a) Representar gráficamente los datos y comentar los resultados.**

Datos de X:

$$\text{Re} = 3.3 - 1 = 2.3$$
$$n = 16 \leq 50 \Rightarrow m = \sqrt{n} = \sqrt{16} = 4 \Rightarrow C_i = 4$$
$$a_i = \frac{\text{Re}}{C_i} = \frac{2.3}{4} = 0.575 \Rightarrow a_i = 0.6$$

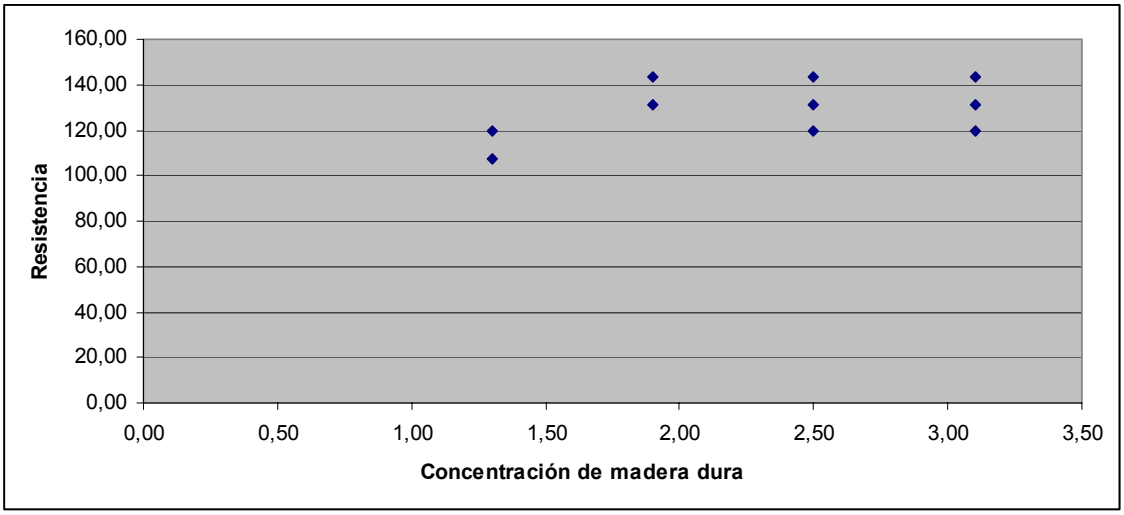
Datos de Y:

$$\text{Re} = 146.9 - 101.4 = 45.5$$

$$n = 16 \leq 50 \Rightarrow m = \sqrt{n} = \sqrt{16} = 4 \Rightarrow C_i = 4$$

$$a_i = \frac{\text{Re}}{C_i} = \frac{45.5}{4} = 11.375 \Rightarrow a_i = 12$$

| $x_i \backslash y_j$ | [101,4-113,4) | [113,4-125,4) | [125,4-137,4) | [137,4-149,4) | $c_j$ | $n_{i.}$ | $f_{i.}$ | $N_{i.}$ | $F_{i.}$ |
|---|---|---|---|---|---|---|---|---|---|
| [1-1,6) | 2/0,125 | 2/0,125 | - | - | 1,3 | 4 | 0,250 | 4 | 0,250 |
| [1,6-2,2) | - | - | 1/0,06 | 1/0,06 | 1,9 | 2 | 0,125 | 6 | 0,375 |
| [2,2-2,8) | - | 2/0,125 | 1/0,06 | 1/0,06 | 2,5 | 4 | 0,250 | 10 | 0,625 |
| [2,8-3,4) | - | 1/0,06 | 1/0,06 | 4/0,25 | 3,1 | 6 | 0,375 | 16 | 1,000 |
| $c_j$ | 107,4 | 119,4 | 131,4 | 143,4 | | | | | |
| $n_{.j}$ | 2 | 5 | 3 | 6 | | 16 | - | | |
| $f_{.j}$ | 0,125 | 0,313 | 0,188 | 0,375 | | - | 1 | | |
| $N_{.j}$ | 2 | 7 | 10 | 16 | | | | | |
| $F_{.j}$ | 0,125 | 0,438 | 0,625 | 1,000 | | | | | |




Lo que se observa es una variación muy leve de la resistencia a medida que aumentamos la concentración.

**b) Hallar el coeficiente de correlación.**

$$\bar{x} = \frac{\sum_{i=1}^{n} c_i * n_i}{n} \Rightarrow \bar{x} = 2.35$$

$$\bar{y} = \frac{\sum_{j=1}^{n} c_j * n_j}{n} \Rightarrow \bar{y} = 129.15$$

$$Sn_x^2 = \frac{\sum_{i=1}^{n} c_i^2 * n_i}{n} - \bar{x}^2 \Rightarrow Sn_x^2 = 0.5175$$

$$Sn_x = \sqrt{Sn_x^2} = 0.719$$

$$Sn_y^2 = \frac{\sum_{j=1}^{n} c_j^2 * n_j}{n} - \bar{y}^2 \Rightarrow Sn_y^2 = 165.9375$$

$$Sn_y = \sqrt{Sn_y^2} = 12.88$$

$$S_{xy} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} n_{ij} * x_i * y_j}{n} - \bar{x} * \bar{y} = 5.9625$$

$$r_{xy} = \frac{S_{xy}}{S_x * S_y} = 0.644$$

No tiene mucha fiabilidad.

**c) Ajustar un modelo de regresión lineal. Predecir la resistencia de una caja fabricada con pulpa cuya concentración es 2.3.**

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} * (x - \bar{x})$$

$$y - 129.15 = \frac{5.9625}{0.5175} * (x - 2.35) \rightarrow y = 11.52 * x + 102.07$$

Si x = 2.3 → <u>y = 128.566</u>

**41. Sabiendo que $\bar{x}$ = 3, $s^2_x$ = 6, $s^2_y$ = 8 y que la recta de regresión de Y sobre X es:**

**y=4 – 0.667x, obtener la recta de regresión de X sobre Y.**

**SOLUCIÓN:**

$$y-\bar{y}=\frac{S_{xy}}{S_x^2}*(x-\bar{x})$$

$$y=\left(\frac{S_{xy}}{S_x^2}\right)*x+\left(\bar{y}-\frac{S_{xy}}{S_x^2}*\bar{x}\right)=4-0.667*x \qquad \Rightarrow \frac{S_{xy}}{S_x^2}=-0.667 \qquad \Rightarrow \bar{y}-\frac{S_{xy}}{S_x^2}*\bar{x}=4$$

$$S_{xy}=-0.667*S_x^2=-0.667*6=-4.002\approx -4$$

$$\bar{y}=4+\frac{S_{xy}}{S_x^2}*\bar{x}=4-0.667*3=1.999\approx 2$$

$$x-\bar{x}=\frac{S_{xy}}{S_y^2}*(y-\bar{y})$$

$$x-3=\frac{-4}{8}*(y-2)\rightarrow \underline{x=4-0.5*y}$$

**42. Hallar la recta de regresión de Y sobre X sabiendo que $\bar{x}$ = 4.1, $\bar{y}$ = 2.3 y la recta pasa por el punto (5.9, 3.5).**

**SOLUCIÓN:**

$$y=\left(\frac{S_{xy}}{S_x^2}\right)*x+\left(\bar{y}-\frac{S_{xy}}{S_x^2}*\bar{x}\right)=a+b*x \qquad \Rightarrow \frac{S_{xy}}{S_x^2}=b \qquad \Rightarrow a=\bar{y}-\frac{S_{xy}}{S_x^2}*\bar{x}=\bar{y}-b*\bar{x}$$

Si pasa por el punto (5.9, 3.5):

$$y = a + b * x = \left(\bar{y} - b * \bar{x}\right) + b * x$$

$$3.5 = \left(2.3 - b * 4.1\right) + b * 5.9 \Rightarrow b = 0.6667$$

$$a = \bar{y} - b * \bar{x} = 2.3 - 0.667 * 4.1 = -0.433$$