**COEFICIENTE DE CORRELACIÓN DE PEARSON PARA DATOS AGRUPADOS EN INTERVALOS**

**Autor:** Mario Orlando Suárez Ibujes

mgsmariosuarez@gmail.com

mosuarez@utn.edu.ec

**Telf:** 06 2632 166

085619601

**FÓRMULA**

$$r=\frac{n∙\sum\_{}^{}f∙dx∙dy-\left(\sum\_{}^{}fx∙dx\right)\left(\sum\_{}^{}fy∙dy\right)}{\sqrt{\left[n∙\sum\_{}^{}fx∙dx^{2}-\left(\sum\_{}^{}fx∙dx\right)^{2}\right]\left[n∙\sum\_{}^{}fy∙dy^{2}-\left(\sum\_{}^{}fy∙dy\right)^{2}\right]}}$$

Donde

n = número de datos.

f = frecuencia de celda.

fx = frecuencia de la variable X.

fy = frecuencia de la variable Y.

dx = valores codificados o cambiados para los intervalos de la variable X, procurando que al intervalo central le corresponda dx = 0, para que se hagan más fáciles los cálculos.

dy = valores codificados o cambiados para los intervalos de la variable X, procurando que al intervalo central le corresponda dy = 0, para que se hagan más fáciles los cálculos.

**EJEMPLOS ILUSTRATIVOS**

**N° 1**

Con los siguientes datos sobre los Coeficientes Intelectuales (X) y de las calificaciones en una prueba de conocimiento (Y) de 50 estudiantes:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| N° de estudiante | X | Y | N° de estudiante | X | Y |
|  1 | 76 | 28 | 26 | 88 | 40 |
| 2 | 77 | 24 | 27 | 88 | 31 |
| 3 | 78 | 18 | 28 | 88 | 35 |
| 4 | 79 | 41 | 29 | 88 | 26 |
| 5 | 79 | 43 | 30 | 89 | 30 |
| 6 | 80 | 45 | 31 | 89 | 24 |
| 7 | 80 | 34 | 32 | 90 | 18 |
| 8 | 80 | 18 | 33 | 90 | 11 |
| 9 | 82 | 40 | 34 | 90 | 15 |
| 10 | 82 | 35 | 35 | 91 | 38 |
| 11 | 83 | 30 | 36 | 92 | 34 |
| 12 | 83 | 21 | 37 | 92 | 31 |
| 13 | 83 | 22 | 38 | 93 | 33 |
| 14 | 83 | 23 | 39 | 93 | 35 |
| 15 | 84 | 25 | 40 | 93 | 24 |
| 16 | 84 | 11 | 41 | 94 | 40 |
| 17 | 84 | 15 | 42 | 96 | 35 |
| 18 | 85 | 31 | 43 | 97 | 36 |
| 19 | 85 | 35 | 44 | 98 | 40 |
| 20 | 86 | 26 | 45 | 99 | 33 |
| 21 | 86 | 30 | 46 | 100 | 51 |
| 22 | 86 | 24 | 47 | 101 | 54 |
| 23 | 86 | 16 | 48 | 101 | 55 |
| 24 | 87 | 20 | 49 | 102 | 41 |
| 25 | 88 | 36 | 50 | 102 | 45 |

1) Elaborar una tabla de dos variables

2) Calcular el coeficiente de correlación

**Solución:**

1) En la *tabla de frecuencias de dos variables*, cada recuadro de esta tabla se llama una *celda* y corresponde a un par de intervalos, y el número indicado en cada celda se llama *frecuencia de celda*. Todos los totales indicados en la última fila y en la última columna se llaman *totales marginales o frecuencias marginales*, y corresponden, respectivamente, a las frecuencias de intervalo de las distribuciones de frecuencia separadas de la variable X y Y.

Para elaborar la tabla se recomienda:

- Agrupar las variables X y Y en un igual número de intervalos.

- Los intervalos de la variable X se ubican en la parte superior de manera horizontal (fila) y en orden ascendente.

- Los intervalos de la variable Y se ubican en la parte izquierda de manera vertical (columna) y en orden descendente.

Para elaborar los intervalos se procede a realizar los cálculos respectivos:

En la variable X:

Calculando el Rango se obtiene:

$$R=x\_{máx}-x\_{mín}=102-76=26$$

Calculando el número de intervalos se obtiene:

$$n\_{i}=1+3,32∙log\left(n\right)=1+3,32∙log50=6$$

Calculando el ancho se obtiene:

$$i=\frac{R}{ni}=\frac{26}{6}=4,33$$

En la variable Y:

Calculando el Rango se obtiene:

$$R=y\_{máx}-y\_{mín}=55-11=44$$

Calculando el número de intervalos se obtiene:

$$n\_{i}=1+3,32∙log\left(n\right)=1+3,32∙log50=6$$

Calculando el ancho se obtiene:

$$i=\frac{R}{ni}=\frac{44}{6}=7,33$$

**Nota:** Para la variable X se tomará un ancho de intervalo igual a 5 y para la variable Y un ancho de intervalo igual a 8 para obtener un número de intervalos igual a 6 para cada variable.

Contando las frecuencias de celda para cada par de intervalos de las variables X y Y se obtiene la siguiente tabla de frecuencias de dos variables:

|  |  |  |
| --- | --- | --- |
|  | Coeficientes Intelectuales (X) |  |
| 76-80 | 81-85 | 86-90 | 91-95 | 96-100 | 101-105 | fy |
|  | 51-58 |  |  |  |  | 1 | 2 | 3 |
| 43-50 | 2 |  |  |  |  | 1 | 3 |
| 35-42 | 1 | 3 | 3 | 3 | 3 | 1 | 14 |
| 27-34 | 2 | 2 | 3 | 3 | 1 |  | 11 |
| 19-26 | 1 | 4 | 5 | 1 |  |  | 11 |
| 11-18 | 2 | 2 | 4 |  |  |  | 8 |
| fx | 8 | 11 | 15 | 7 | 5 | 4 | 50 |

**Interpretación:**

- El número 5 es la frecuencia de la celda correspondiente al par de intervalos 86-90 en Coeficiente Intelectual y 19-26 en Calificación obtenida en la prueba de conocimiento.

- El número 8 en la fila de fx es el total marginal o frecuencia marginal del intervalo 76-80 en Coeficiente Intelectual.

- El número 14 en la columna de fy es el total marginal o frecuencia marginal del intervalo 35-42 en Calificación obtenida en la prueba de conocimiento.

- El número 50 es total de frecuencias marginales y representa al número total de estudiantes.

2) Realizando los cálculos respectivos se obtiene la siguiente tabla:

|  |  |
| --- | --- |
|  |  |
| X | 76-80 | 81-85 | 86-90 | 91-95 | 96-100 | 101-105 |
|  | Y |  dx dy | -2 | -1 | 0 | 1 | 2 | 3 | fy | fy·dy | fy·dy2 | f·dx·dy |
| 51-58 | 3 |  |  |  |  | 1 | 2 | 3 | 9 | 27 | 24 |
|  | 6 |  | 18 |
| 43-50 | 2 | 2 |  |  |  |  | 1 | 3 | 6 | 12 | -2 |
|  | -8 |  | 6 |
| 35-42 | 1 | 1 | 3 | 3 | 3 | 3 | 1 | 14 | 14 | 14 | 7 |
|  | -2 |  | -3 |  | 0 |  | 3 |  | 6 |  | 3 |
| 27-34 | 0 | 2 | 2 | 3 | 3 | 1 |  | 11 | 0 | 0 | 0 |
|  | 0 |  | 0 |  | 0 |  | 0 |  | 0 |
| 19-26 | -1 | 1 | 4 | 5 | 1 |  |  | 11 | -11 | 11 | 5 |
|  | 2 |  | 4 |  | 0 |  | -1 |
| 11-18 | -2 | 2 | 2 | 4 |  |  |  | 8 | -16 | 32 | 12 |
|  | 8 |  | 4 |  | 0 |
| fx | 8 | 11 | 15 | 7 | 5 | 4 | 50 | 2 | 96 | 46 |
| fx·dx | -16 | -11 | 0 | 7 | 10 | 12 | 2 |  |
| fx·dx2 | 32 | 11 | 0 | 7 | 20 | 36 | 106 |
| f·dx·dy | 0 | 5 | 0 | 2 | 12 | 27 | 46 |

**Nota:**

Los números de las esquinas de cada celda en la anterior tabla representan el producto f·dx·dy, así por ejemplo, para obtener el número el número -8 de los intervalos 76-80 en X y 43-50 en Y se obtiene multiplicando 2·(-2)·(2) = -8. Para obtener el número 6 de los intervalos 96-100 en X y 51-58 en Y se obtiene multiplicando 1·2·3 = 6.

Los números de la última columna (24, -2, 7, 0, 5 y 12) se obtienen sumando los números de las esquinas en cada fila, así por ejemplo, para obtener el número 24 se suma 6 + 18 = 24.

Los números de la última fila (0, 5, 0, 2, 12 y 27) se obtienen sumando los números de las esquinas en cada columna, así por ejemplo, para obtener el número 27 se suma 18 + 6 + 3 = 27.

Para obtener el número 2 de la antepenúlmina columna se obtiene sumando los resultados de fy·dy, es decir, representa la ∑ fy·dy

Para obtener el número 2 de la antepenúlmina fila se obtiene sumando los resultados de fx·dx, es decir, representa la ∑ fy·dy

Para obtener el número 96 de la penúltima columna se obtiene sumando los resultados de fy·dy2, es decir, representa ∑ fy·dy2

Para obtener el número 106 de la penúltima fila se obtiene sumando los resultados de fx·dx2, es decir, representa ∑ fx·dx2

Para obtener último número de la última columna se obtiene sumando los resultados de la última columna (46=24-2+7+0+5+12), es decir, representa ∑f·dx·dy.

Para obtener último número de la última fila se obtiene sumando los resultados de la última fila (46=0+5+0+2+12+27), y tiene que ser igual al último número de la última columna como comprobación que los cálculos de la tabla han sido correctos.

Observando los datos en la tabla anterior se reemplaza los valores en la ecuación del Coeficiente de Correlación de Pearson para datos agrupados se obtiene:

$$r=\frac{n∙\sum\_{}^{}f∙dx∙dy-\left(\sum\_{}^{}fx∙dx\right)\left(\sum\_{}^{}fy∙dy\right)}{\sqrt{\left[n∙\sum\_{}^{}fx∙dx^{2}-\left(\sum\_{}^{}fx∙dx\right)^{2}\right]\left[n∙\sum\_{}^{}fy∙dy^{2}-\left(\sum\_{}^{}fy∙dy\right)^{2}\right]}}$$

$$r=\frac{50∙46-\left(2\right)\left(2\right)}{\sqrt{\left[50∙106-(2)^{2}\right]\left[50∙96-(2)^{2}\right]}}=\frac{2300-4}{\sqrt{\left[5300-4\right]\left[4800-4\right]}}=\frac{2296}{\sqrt{\left[5296\right]\left[4796\right]}}$$

$$r=\frac{2296}{\sqrt{25399616}}=\frac{2296}{5039,803}=0,456$$

Existe una correlación positiva moderada

**N° 2**

Dada la siguiente tabla de frecuencias de dos variables, con los datos sobre los pesos en kilogramos de habitantes de dos barrios diferentes en una ciudad, determinar el tipo de correlación que existe entre ellas mediante el coeficiente de Pearson.

|  |  |  |
| --- | --- | --- |
|  | X |  |
| 40-49 | 50-59 | 60-69 | 70-79 | 80-89 | 90-99 | fy |
| Y | 90-99 |  |  |  | 3 | 3 | 4 | 10 |
| 80-89 |  |  | 8 | 2 | 2 | 4 | 16 |
| 70-79 |  |  | 5 | 10 | 8 | 1 | 24 |
| 60-69 | 8 | 1 | 2 | 5 | 2 |  | 18 |
| 50-59 | 3 | 10 | 6 | 2 |  |  | 21 |
| 40-49 | 4 | 6 | 1 |  |  |  | 11 |
| fx | 15 | 17 | 22 | 22 | 15 | 9 | 100 |

**Solución:**



Correlación positiva moderada de 0,688

**N° 3**

Dada la siguiente tabla de frecuencias de dos variables, con los datos sobre las calificaciones obtenidos en un curso de 100 estudiantes en la asignatura de Matemática (X) y en la asignatura de Estadística (Y), determinar el tipo de correlación que existe entre ellas mediante el coeficiente de Pearson para datos agrupados.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| N° de estudiante | X | Y | N° de estudiante | X | Y | N° de estudiante | X | Y | N° de estudiante | X | Y |
| 1 | 40 | 60 | 26 | 57 | 73 | 51 | 71 | 86 | 76 | 84 | 83 |
| 2 | 41 | 50 | 27 | 58 | 78 | 52 | 72 | 88 | 77 | 84 | 84 |
| 3 | 42 | 55 | 28 | 60 | 79 | 53 | 72 | 89 | 78 | 85 | 86 |
| 4 | 43 | 59 | 29 | 61 | 60 | 54 | 72 | 70 | 79 | 86 | 88 |
| 5 | 44 | 40 | 30 | 62 | 61 | 55 | 73 | 71 | 80 | 86 | 89 |
| 6 | 45 | 42 | 31 | 63 | 62 | 56 | 74 | 72 | 81 | 86 | 70 |
| 7 | 45 | 49 | 32 | 64 | 63 | 57 | 74 | 73 | 82 | 87 | 78 |
| 8 | 45 | 60 | 33 | 64 | 64 | 58 | 74 | 74 | 83 | 87 | 79 |
| 9 | 45 | 62 | 34 | 65 | 65 | 59 | 75 | 75 | 84 | 88 | 78 |
| 10 | 48 | 66 | 35 | 65 | 66 | 60 | 76 | 76 | 85 | 88 | 77 |
| 11 | 49 | 69 | 36 | 66 | 67 | 61 | 76 | 77 | 86 | 88 | 79 |
| 12 | 50 | 50 | 37 | 66 | 69 | 62 | 77 | 78 | 87 | 88 | 78 |
| 13 | 50 | 52 | 38 | 66 | 50 | 63 | 77 | 79 | 88 | 89 | 78 |
| 14 | 56 | 54 | 39 | 66 | 52 | 64 | 78 | 60 | 89 | 89 | 60 |
| 15 | 56 | 56 | 40 | 67 | 55 | 65 | 78 | 67 | 90 | 89 | 69 |
| 16 | 56 | 59 | 41 | 68 | 56 | 66 | 78 | 65 | 91 | 90 | 90 |
| 17 | 56 | 59 | 42 | 68 | 57 | 67 | 78 | 68 | 92 | 91 | 96 |
| 18 | 56 | 40 | 43 | 68 | 59 | 68 | 79 | 69 | 93 | 92 | 97 |
| 19 | 57 | 45 | 44 | 69 | 40 | 69 | 79 | 50 | 94 | 93 | 99 |
| 20 | 57 | 47 | 45 | 69 | 45 | 70 | 79 | 59 | 95 | 94 | 80 |
| 21 | 57 | 48 | 46 | 69 | 47 | 71 | 80 | 90 | 96 | 95 | 81 |
| 22 | 57 | 49 | 47 | 69 | 49 | 72 | 81 | 94 | 97 | 96 | 82 |
| 23 | 57 | 80 | 48 | 70 | 90 | 73 | 82 | 96 | 98 | 97 | 83 |
| 24 | 57 | 70 | 49 | 70 | 99 | 74 | 82 | 99 | 99 | 98 | 89 |
| 25 | 57 | 72 | 50 | 70 | 80 | 75 | 83 | 80 | 100 | 99 | 70 |

**Solución:**



Correlación positiva moderada de 0,62

**REFERENCIAS BIBLIOGRÁFICAS**

BENALCÁZAR, Marco, (2002), Unidades para Producir Medios Instruccionales en Educación, SUÁREZ, Mario Ed. Graficolor, Ibarra, Ecuador.

DAZA, Jorge, (2006), Estadística Aplicada con Microsoft Excel, Grupo Editorial Megabyte, Lima,

 Perú.

GOVINDEN, Lincoyán, (1985), Introducción a la Estadística, Ed. McGraw Hill. Interamericana

 Editores. S.A., Bogotá, Colombia.

JOHNSON, Robert, (2003), Estadística Elemental, Ed. Math Learning, Ed. Tercera, México DF.

KUBY, Patricia.

KAZMIER, J. Leonard, (1990). Estadística Aplicada a la Administración y la Economía,

 Ed. McGrawHill, Ed. Segunda, Bogotá, Colombia.

LIND, Marchal, (2005), Estadística Aplicada a los Negocios y a la Economía, Ed. McGraw- Hill,

MASON Ed. Décima., Mexico DF.

MARTINEZ, Bencardino, (1981), Estadística Comercial, Ed. Norma, Bogotá, Colombia.

MORENO, Francis, (1993), Estadística Inferencial, Universidad Particular de Loja, Loja, Ecuador.

SÁNCHEZ, Jesús, (2007), Introducción a la Estadística Empresarial, Madrid, España.

SALTOS, Héctor, (1986), Estadística de Inferencia, Ed. Pío XII, Ambato, Ecuador.

SHAO, Stephen, (1980), Estadística para Economistas y Administradores de Empresas, Ed. Herrero

 Hnos, México DF.

SPIEGEL, Murray, (2000), Estadística,Serie de Compendios Schaum, Ed. McGraw-Hill, México.

SUÁREZ, Mario, (2004), Interaprendizaje Holístico de Matemática, Ed. Gráficas Planeta, Ibarra,

 Ecuador.

STEVENSON, William, (1981), Estadística para Administración y Economía, Ed. Harla S.A de C.V.

 México D.F.

WEBSTER, Allen, (2000), Estadística Aplicada a los Negocios y a la Economía, Ed. McGraw Hill.

 Interamericana Editores S.A. Bogotá, Colombia