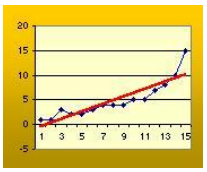


Temas de Estadística Práctica
Antonio Roldán Martínez

Proyecto http://www.hojamat.es/

Tema 4: Distribuciones bidimensionales. Correlación.

Resumen teórico

Resumen teórico de los principales conceptos estadísticos

Distribuciones bidimensionales. Correlación.

Distribuciones bidimensionales

En algunos experimentos las medidas que se obtienen son dobles, pertenecientes a dos variables distintas, a las que llamaremos X e Y respectivamente.

Este tipo de estudios es muy frecuente. Daremos algunos ejemplos:

- Comparación entre mortalidad y natalidad
- Ídem entre extensión y población de diversos países.
- Diferencias de renta entre la población en general y los titulados universitarios.
- Pruebas *pretest* y *postest*.
- Influencia de la latitud en la temperatura media.
- Ídem de las horas de estudio en la calificación en una asignatura.
- Etc.

Tipos de variables

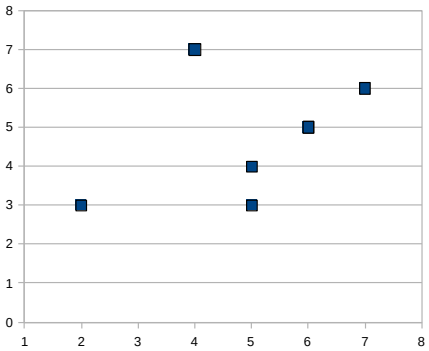
Las dos variables que se comparan pueden ser de igual naturaleza, ambas nominales u ordinales o de intervalo, o de distinta, lo que da lugar a muchos casos posibles, que es imposible estudiarlos todos en este curso.

Incluimos algunos ejemplos:

Tablas simples de comparación de dos datos cuantitativos

Alumnos	X: Examen de Geografía	Y: Examen de Matemáticas
Julia	4	7
Pedro	6	5
Miguel	5	4
Marta	2	3
.....	.....	.....

En estos casos cada par de valores representa a un sujeto o medición. Se representan mediante gráficos de dispersión XY



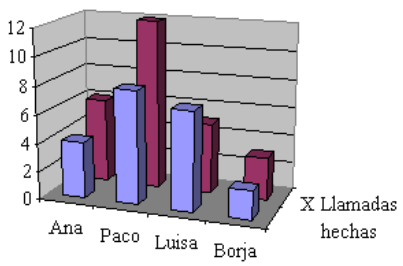
Tablas de doble entrada

En ellas la X y la Y pueden ser de naturaleza muy distinta, por lo que se disponen en tabla de doble entrada. Cuando existen frecuencias, es el mejor método, pues permite tratar una variable por columnas y otra por filas.

La siguiente tabla muestra la distribución de las llamadas telefónicas con origen o destino en los cuatro hijos de una pareja.

	X Llamadas hechas	Y Llamadas recibidas
Ana	4	6
Paco	8	12
Luisa	7	5
Borja	2	3

Estas tablas de doble entrada con frecuencias admiten una representación gráfica muy intuitiva mediante barras (columnas) ordenadas en varios conjuntos mediante tres ejes.



Tipos de frecuencias en una distribución bidimensional

Para aclarar las definiciones de los tipos de frecuencias usaremos la siguiente tabla:

Meses Niveles	Enero	Febrero	Marzo	Abril	Mayo
A	4	6	7	8	8
B	3	3	6	5	9
C	9	7	7	13	14

Frecuencias conjuntas

Se representan por n_{ij}, y son las frecuencias incluidas en la tabla primitiva de entrada. Los subíndices ***i*** y ***j*** representan la fila y columna en la que está situada la frecuencia.Así, en la tabla $n_{13} = 7$ y $n_{34} = 13$

Llamaremos N a la suma total de estas frecuencias. En el ejemplo, N es 109.

Representaremos este hecho mediante un sumatorio doble sin índices, para no complicar las fórmulas:

$$\sum\sum n_{ij} = N$$

Al conjunto de las frecuencias conjuntas lo denominaremos como **Distribución conjunta de las dos variables**.

Frecuencias marginales

Llamaremos ***frecuencia marginal*** de un valor de X, a la que le corresponde a ese valor si no tenemos en cuenta la existencia de Y. En la práctica coincide con la suma de todas las frecuencias contenidas en **la fila correspondiente a ese valor**.

En la tabla del ejemplo, la frecuencia marginal de B es 26, suma de las frecuencias de la segunda fila. La frecuencia marginal de la fila ***i*** se representará por n_{i*}

De la misma forma se define la frecuencia marginal en la variable Y, como la que tendría si no se tuviera en cuenta la X, o la suma de la columna correspondiente. En el ejemplo, la frecuencia marginal de Marzo es $n_{*3} = 20$

Frecuencias condicionadas

Son las frecuencias que posee una variable si sólo consideramos **un valor** (o varios) de la otra variable. En la práctica se traduce a considerar sólo una fila o sólo una columna, según el valor elegido.

Las frecuencias condicionadas se representan con este símbolo: $\mathbf{n_{x/y}}$, que se puede leer como *Frecuencia de x condicionada por y*.

En la tabla del ejemplo, la distribución de X condicionada a Marzo es la columna A=7, B=6, C=7. Las frecuencias condicionadas son más representativas si se convierten en proporciones o porcentajes.

Medidas en una distribución bidimensional

Al existir dos variables X e Y, las medidas también son dobles. Así, consideraremos las siguientes:

Media de X

Tiene la misma definición que en el caso unidimensional. Viene dada por la fórmula

$$\bar{x} = \frac{\sum x}{N}$$

si los datos están aislados y por esta otra

$$\bar{x} = \frac{\sum x \cdot n}{N} = \sum x \cdot f$$

si están agrupados.

Media de la Y

Se define de forma similar:

$$\bar{y} = \frac{\sum y}{N}$$

y para agrupados

$$\bar{y} = \frac{\sum y \cdot n}{N} = \sum y \cdot f$$

(Las siguientes definiciones las desarrollaremos sólo para aislados, pues su traducción es fácil)

Varianzas y desviaciones típicas

También serán dobles:

La varianza de X será

$$s_x^2 = \frac{\sum (x_i - \bar{x})^2}{N} = \frac{\sum x_i^2}{N} - \bar{x}^2$$

y su desviación típica $\mathbf{s_x}$ será la raíz cuadrada de esa expresión.

En el caso de Y la definición es similar:

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{N} = \frac{\sum y_i^2}{N} - \bar{y}^2$$

Covarianza

Esta medida es muy interesante. Mide el *paralelismo* existente entre ambas variables (en función **sólo** de los datos presentes en la tabla). Si la covarianza es grande, manifestará la existencia de un cierto paralelismo o dependencia (en sentido estadístico) entre X e Y. Si es pequeña, indicará que ambas variables se comportan de manera más independiente.

Su definición es:

$$s_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{N} = \frac{\sum xy}{N} - \bar{x}\bar{y}$$

y puede ser positiva, cero o negativa.

El significado de la varianza es el siguiente:

Si en el numerador la mayoría de los productos son positivos, será porque las diferencias de X y de Y tienen el mismo signo. Eso significa que para X mayor que la media, la Y también lo es, y al contrario, a valores pequeños de X le corresponden pequeños en Y. Por tanto, los productos serán mayoritariamente positivos y la varianza crecerá.

Una varianza positiva y alejada del valor cero indica un cierto paralelismo entre X e Y, en el que a valores mayores de X le corresponden los mayores en Y.

Si los productos son mayoritariamente negativos, es que las diferencias tienen distintos signos, por lo que

Una varianza negativa y alejada del cero indica un paralelismo inverso, en el que a valores pequeños de X le corresponden valores grandes de Y, y a la inversa.

Por último, si están muy repartidos los productos positivos y negativos, es que apenas existe paralelismo, y la varianza se acercará a cero.

El problema de la varianza es que carece de un valor máximo, por lo que es difícil juzgar si la correspondencia entre las dos variables es la mejor posible.

Coeficiente de correlación

Como en el caso de una variable, la *covarianza* no es adecuada para establecer comparaciones entre medidas muy diferentes, además del inconveniente de no tener un valor máximo, lo que impide valorar el grado de paralelismo existente en los datos.

Para normalizar la covarianza procederemos como en el Coeficiente de Variación: dividiremos dicha covarianza entre las dos desviaciones típicas (de X y de Y respectivamente). Al resultado le daremos el nombre de *Coeficiente de correlación* y lo representaremos por **r**.

$$r = \frac{s_{xy}}{s_x s_y}$$

El coeficiente **r** también recibe el nombre de ***Coeficiente de Pearson*** o también ***Coeficiente de correlación producto-momento***.

También se puede demostrar que este coeficiente es en realidad la covarianza del conjunto si expresamos los datos en medidas típicas ***z*** (ver sesión 3).

El valor de **r** oscila entre -1 y +1, y mide el paralelismo o *correlación* entre X e Y. Si sus valores se acercan a 1 o a -1, diremos que existe correlación **fuerte**, y está cerca del cero, **débil**.

Podemos desarrollar más estos comentarios mediante una tabla:

Valor de r	Comentario
+1	Dependencia funcional positiva (función creciente entre ambas)
Cercana al 1	Correlación fuerte positiva
Cercana al 0	Correlación débil o independencia
Cercana al -1	Correlación fuerte negativa
-1	Dependencia funcional negativa (función decreciente)

Se deben evitar interpretaciones erróneas del coeficiente **r**. Seleccionamos las más frecuentes:

La dependencia es sólo matemática: no supone relación causa-efecto. Las causas nunca son tan simples y pueden existir, pero respecto a una tercera variable.

Se deben evitar demasiados adjetivos como *correlación regular, media, ...* pues el significado exacto de **r** depende de cada experimento en concreto.

Si la relación entre datos es de tipo curvilíneo, el coeficiente **r** pierde representatividad.

A veces, si existe asimetría, **r** no puede acercarse al 1.

Otras medidas de correlación

El coeficiente de correlación de Pearson exige que la escala de medida sea de intervalo o razón. Cuando este supuesto no se cumple, deberemos usar otros coeficientes, aunque muchos de ellos equivalen, en sus cálculos, al coeficiente de Pearson.

Coeficiente de Spearman o de rangos

Si la variable es de tipo ordinal, podemos usar los rangos (número de orden de cada dato) para evaluar la correlación.

Representaremos por **d** a la diferencia entre rangos que presenta un dato en dos ordenaciones distintas. Por ejemplo, supongamos que diez individuos han sido ordenados de forma diferente por dos evaluadores A y B:

Individuos	1	2	3	4	5	6	7	8	9	10
A	2	3	4	1	5	9	10	8	7	6
B	3	5	1	4	2	6	8	10	9	7
D	+1	+2	-3	+3	-3	-3	-2	+2	+2	+1

La suma de todas las diferencias será cero.

La fórmula del coeficiente de Spearman es

$$\rho = 1 - \frac{6 \cdot \sum d^2}{N \cdot (N^2 - 1)}$$

Si existen empates entre ordenaciones se resuelven asignando el rango promedio.

Equivale al coeficiente de Pearson, aunque se calcule mediante otras técnicas. Un coeficiente positivo significará que rangos altos en una de las variables se corresponderán con rangos también altos en la otra, y negativo cuando a los altos en una correspondan bajos en otra.

Coeficiente biserial puntual

Se utiliza cuando X es cuantitativa y la Y dicotómica (variable con sólo dos valores). Por ejemplo, X puede ser la calificación en un examen de Ciencias Sociales, y la Y el hecho de que los alumnos examinados tengan o no una habitación para estudiar ellos solos, sin compartirla con los hermanos.

Los valores de la variable Y se suelen representar por 1 y 0. Puede ser dicotómica en su definición (tener o no tener, aprobar o suspender, ...), o bien haber sido *dicotomizada*, si, por ejemplo, asignamos un 1 a los individuos que superen un valor y 0 a los que no lo superen.

La fórmula de este coeficiente es:

$$r_{bp} = \frac{\bar{X}_1 - \bar{X}_0}{s_x} \cdot \sqrt{pq}$$

donde las medias del numerador corresponden a los valores correspondientes a Y=1 e Y=0 respectivamente, la desviación típica del denominador a la de **todas las X**, y los valores **p** y **q** a las proporciones de sujetos con Y=1 e Y=0 respectivamente.

En la siguiente tabla presentamos un ejemplo de situación en la que es aplicable este coeficiente:

X: Notas en el examen de Ciencias Sociales.

Y: Disposición de habitación de estudio individual, representada por 0 y 1.

X	9	5	4	8	6	9	8	6	6	7
Y	1	1	0	0	0	1	1	0	1	0

Coeficiente de contingencia

Se utiliza para tablas de doble entrada que contengan frecuencias correspondientes a dos variables de cualquier tipo de escala, desde nominal hasta cuantitativa de razón.

Usa la distribución chi-cuadrado χ^2, que se estudia en otra sesión del curso.

Su fórmula es

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$