

TEMA 4

DISTRIBUCIONES BIDIMENSIONALES

Para cada individuo de la población se analizarán dos características de interés. Se tendrán n observaciones del tipo (x_i, y_j) .

La expresión general de una **distribución conjunta** es:

Distribuciones marginales.- Se llama **distribución marginal de cada componente n_i** a la distribución univariante de dicho componente, es decir, a su distribución en la población considerado aisladamente:

Distribuciones condicionadas.- La distribución condicionada de una variable x_i se define cuando el valor de otra variable y_j se supone fijo e igual a un valor concreto. La distribución condicionada es la distribución univariante de x_i en los elementos de la población que tienen como valor de y_j el valor fijado:

REPRESENTACIONES GRÁFICAS

Diagrama de dispersión.- Los datos x_i se representan en el eje X y en el eje Y se representan las y_i . Las coordenadas que nos proporcionan estos pares nos permiten realizar gráficos de este tipo:

Histograma 3D.- Es otra forma de representar los datos de una distribución bidimensional. Su principal inconveniente es que sólo puede utilizarse cuando las observaciones están agrupadas en intervalos de clase:

MOMENTOS

Esto es una ampliación de la teoría de momentos vista en el Tema 2, pero para variables bidimensionales.

Momentos respecto al origen.-

Si $r = 1$ y $s = 0$

Si $r = 0$ y $s = 1$

Momentos centrales de orden (r, s) .- Se denominan así cuando $p_1 =$, y $p_2 =$

Casos particulares:

$$m_{20} = \bar{x}^2$$

$$m_{02} = \bar{y}^2$$

Covarianza.- Si $r = s = 1$, el momento resultante es lo suficientemente importante como para merecer un nombre propio. Se trata de la **covarianza**, que mide la relación lineal existente entre X e Y .

Propiedad de la covarianza.- $S_{xy} = S_{yx}$

Vector de medias.- En el estudio de variables cuantitativas bidimensionales, las dos observaciones asociadas a un individuo pueden considerarse como un vector X , con componentes los valores que en $\bar{X}$ toma cada variable. El conjunto de datos se representa por la secuencia de vectores $x_1 \dots x_n$. Se llamará **vector de medias**

de la variable bidimensional al vector de dimensión 2 cuyos componentes son las medias aritméticas de cada variable:

Matriz de varianzas-covarianzas.- Es la matriz cuadrada simétrica que tiene en la diagonal principal las varianzas de las observaciones, y fuera de ellas las covarianzas entre variables:

REGRESIÓN

Regresión.- Teoría que trata de expresar mediante una expresión matemática la relación que existe entre las variables. Dado un conjunto de puntos, el dibujo de su nube nos puede indicar si existe algún tipo de relación entre las variables.

La idea de **mejor recta o curva que mejor se ajusta** es aquella curva más próxima a la nube de puntos, aquella que posee la menor distancia a los puntos de la nube.

Recta de Regresión.- Dadas dos variables se trata de ver las relaciones de dependencia que puedan darse entre ellas. Esta relación puede ser de tres tipos:

Dependencia exacta.- Es la que se da en las distribuciones que siguen una fórmula matemática.

Dependencia lineal.- El conocimiento de una de las variables da una información incompleta sobre la otra.

III. **Independencia.**- En este caso una variable no da ninguna información sobre la otra.

En las distribuciones bidimensionales nos interesa saber si hay alguna relación lineal entre X e Y , es decir, si podemos expresar Y como función lineal de X , aunque distorsionada por un error aleatorio. La ecuación de una relación lineal será:

donde X es la variable explicativa e Y la variable aleatoria. $y = a + bx$ es la llamada **recta de regresión**, a la que se suma el error.

Método de los mínimos cuadrados.- Se utiliza para conocer los parámetros a y b . El método consiste en escoger los parámetros que hagan mínima la suma de los errores al cuadrado:

siendo los valores de y . Haciendo cuentas nos queda la **ecuación punto-pendiente de la recta de regresión de Y sobre X** :

fig 1.- Recta de regresión lineal.

La pendiente de una recta de regresión lineal como la de la **fig 1.** es .

Coefficiente de correlación lineal.- Para decidir si una recta se ajusta bien a los datos utilizamos esta medida adimensional, que se calcula según esta expresión:

Propiedades del coeficiente de correlación lineal:

$$-1 \leq r_{xy} \leq 1$$

$$r_{xy} = r_{yx}$$

$r_{xy} = 0$ posible sólo en el caso en que la distribución mantiene un valor constante ($Y=5$, por ejemplo). En este caso nos queda una imagen parecida a esta.

• *fig 2.- Función de correlación con $r_{xy} = 0$.*

• En este caso lo único que se ve representado son los errores aleatorios.

• En caso de $r_{xy}=1$ nos quedaría un gráfico como el de la *fig 1*. La recta en este caso es creciente.

• En caso de $r_{xy} = -1$ nos quedaría un gráfico como el de la *fig 1*. pero con la recta decreciente.

•

En el resto de los casos la distribución seguirá una gráfica como la de la *fig 3*.

fig 3.- Mejor curva recta que mejor se ajusta a una nube de puntos.

Coefficiente de determinación.- Se usa mucho en la práctica.

Multiplicado por 100 nos da el **porcentaje de correlación lineal** existente entre las variables. El coeficiente de determinación se interpreta como la proporción de variación de Y explicado por el modelo lineal. Suele aceptarse la independencia lineal si porcentaje de correlación lineal es **superior** al 90%.

Nota importante: La ecuación de la **recta de regresión** de X sobre Y es:

Estadística 1º E.T.I.S. Facultad de Informática da Coruña Curso 1.997-1.998