CAL RECORDAR-2 (Bloc 2)

DISTRIBUCIONS BIDIMENSIONALS

INTRODUCCIÓ

Moltes vegades existeixen situacions a la vida quotidiana en les que interessa saber si dues característiques d'una població estan relacionades i, si ho estan, esbrinar en quina mesura.

Com per exemple, quina relació hi ha entre les hores d'estudi i els resultats acadèmics d'un grup d'alumnes de Batxillerat. En un conjunt de famílies, esbrinar si l'estatura mitjana dels pares i l'estatura mitjana dels fills té alguna relació.

Aquestes i altres preguntes únicament es podran respondre amb estudis estadístics adients. Aquesta part de l'estadística s'anomena estadística bivariant (o estudi de les distribucions bidimensionals).

Els estudis corresponents per mirar de donar resposta a aquestes preguntes es van iniciar a Anglaterra a finals del segle XIX. Es van fer nombrosos treballs dedicats majoritàriament a la relació entre diverses variables biològiques. Sir Francis Galton, va ser un dels artífexs i va dedicar gran part de la seva vida a estudiar les relacions entre les característiques hereditàries entre pares i fills. A partir dels resultats de les seves observacions van sorgir els conceptes de c*orrelació* i *regressió* i en honor al seu deixeble Karl Pearson, el coeficient de *correlació lineal* o *coeficient de Pearson*.

DISTRIBUCIÓ ESTADÍSTICA BIDIMENSIONAL

Una distribució estadística bidimensional és el resultat d'una presa de dades en una població de manera que, per a cada individu de la població, s'ha observat i mesurat el valor de dues variables estadístiques.

Quan observem dues variables estadístiques, X i Y (a les variables estadístiques s'acostuma a anomenar-les en lletres majúscules i els valors individuals corresponents a cada variable amb la mateixa lletra en minúscula) corresponents a cadascun dels individus d'una població, tindrà associats un parell de valors (x, y).

Recorda: *cada observació d'un element de la mostra està representada per un parell de valors (x,y).*

No es tracta de fer un estudi aïllat de cadascuna de les variables, sinó d'estudiar la relació o dependència que pugui existir entre elles, en el cas que aquesta relació existeixi.

L'objectiu fonamental d'aquesta unitat és representar, estudiar, quantificar i formalitzar matemàticament la relació entre les variables, per poder descobrir i/o explicar el fenomen i, si fos possible, fer-ne prediccions.

ORGANITZACIÓ DE DADES

Taules de contingència o taules de doble entrada

La manera més habitual d'expressar els valors obtinguts en una distribució estadística bidimensional és mitjançant una *taula de contingència* o *taula de doble entrada*.

Al marge superior de la taula s'escriuen els resultats d'una de les variables i al marge esquerre, els valors de l'altra variable.

En les caselles de la taula s'indiquen simultàniament les freqüències absolutes, les freqüències relatives o els percentatges corresponents a les dues variables.

També s'acostuma a afegir al final de cada fila i de cada columna les anomenades distribucions marginals, que donen els totals de cada resultat.

Les distribucions marginals de les variables X i Y s'obtenen a partir de la taula de contingència, considerant cada variable per separat. Representen les freqüències dels valors d'una variable, independentment dels valors de l'altra.

A partir de les distribucions marginals podem calcular la mitjana i la desviació típica de cadascuna de les variables, estudiant-les com una variable unidimensional.

Diagrama de dispersió o núvol de punts

És un gràfic que permet il·lustrar les dades de dues variables.

Recorda: ***la representació gràfica dels parell (x,y) s'anomena núvol de punts.***

Per visualitzar els valors obtinguts en la presa de dades d'una distribució bidimensional, s'utilitzen uns eixos de coordenades cartesianes: se situen a l'eix d'abscisses els valors d'una variable i a l'eix d'ordenades els valors de l'altra variable.

Així, les dues mesures de cada unitat experimental es consideren com un parell ordenat que queda representat com un únic punt del pla en un sistema de coordenades.

El conjunt de tots els punts representats forma el *diagrama de dispersió* o *núvol de punts*.

La forma del diagrama de dispersió o núvol de punts ens permet intuir si existeix o no relació entre les dues variables estudiades, si aquesta relació és directa o inversa i la intensitat d'aquesta relació.

El punt que té per coordenades les mitjanes aritmètiques d'ambdues variables es diu **centre** o **punt mitjà de la distribució**. És important saber que el punt mitjà de la distribució no ha de formar part necessàriament del núvol de punts o diagrama de dispersió.

ANÀLISI DE DADES

El principal objectiu quan s'estudien distribucions bidimensionals és determinar si hi ha algun tipus de relació estadística entre les dues variables que es consideren, és a dir, si els canvis en una de les variables incideixen en els canvis de l'altra.

Aquestes relacions no s'han d'interpretar com si fossin *dependències funcionals*.

Dependència o relació estadística

Es parla de *dependència* o *relació estadística* quan el diagrama de dispersió tendeix a aproximar-se a la representació d'una funció. En aquest cas, es considera que les variables estadístiques són dependents.

Si quan una variable creix, també creix l'altra es diu que la relació (*correlació*) és directa o positiva i, si quan una variable creix l'altra decreix, es diu que la relació (*correlació*) és inversa o negativa.

Si els valors d'una variable no influeixen en els valors de l'altra, direm que les variables X i Y són independents.

Recorda: ***s'entén per correlació el grau de relació existent entre dues variables.***

Covariància

Definició: és la mitjana dels productes de les desviacions de X i Y respecte de les seves mitjanes.

La covariància es representa per s_{xy} i l'expressió algèbrica que permet calcular-la és:

$$s_{xy} = \frac{\sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y}) n_i}{\sum_{i=1}^{m} n_i}$$

Però, l'expressió que s'utilitza habitualment (degut a que és una expressió molt més senzilla) per calcular la covariància d'una distribució de dues variables estadístiques és:

$$s_{xy} = \frac{\sum_{i=1}^{m} x_i y_i n_i}{\sum_{i=1}^{m} n_i} - \bar{x} \cdot \bar{y}$$

La covariància també es pot expressar així: σ_{xy}

- La covariància és una mesura que permet saber si la relació entre les dues variables és directa o inversa, i si hi ha o no relació entre aquestes variables.

Si la covariància és positiva, la relació entre les dues variables és directa; i si és negativa, la relació és inversa.

Quan el resultat de la covariància és zero, no un valor molt proper a zero, això indica que les dues variables són independents, és a dir, no hi ha cap tipus de relació. És el que es coneix amb el nom de relació nul·la.

Coeficient de correlació lineal o de Pearson

El coeficient de correlació lineal ens indica el grau d'intensitat de la relació lineal, en cas que n'hi hagi.

Es representa per **r** i es defineix com a:

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

Es calcula dividint la covariància pel producte de les desviacions típiques de cadascuna de les variables.

El signe del coeficient de correlació lineal coincideix amb el signe de la covariància.

Propietats del coeficient de correlació lineal

1	Els valors del coeficient **r** sempre estan entre −1 i 1, inclosos els valors extrems: $-1 \leq r \leq 1$
2	Si $r = -1$, els punts del diagrama de dispersió se situen sobre una línia recta decreixent. En cas que $r = 1$, la recta és creixent.
3	Si **r** és un valor proper a -1, hi haurà una relació lineal forta i inversa entre les dues variables; el núvol de punts tendirà a una línia recta decreixent. Si **r** pren un valor proper a 1, la relació serà lineal forta i directa; el diagrama de dispersió tendirà a una línia recta creixent.
4	Si el valor de **r** és proper a zero, tant amb valor positiu com negatiu, hi haurà una relació lineal molt feble entre les dues variables, tret que el diagrama de dispersió ens indiqui l'existència d'alguna relació no lineal.

REGRESSIÓ LINEAL

De vegades, l'estudi de la relació entre dues variables estadístiques té com a objectiu fer prediccions sobre els valors de cadascuna de les variables.

L'objectiu de la regressió lineal és, per tant, determinar una relació funcional entre les dues variables del tipus y = a x + b, que ens permeti fer prediccions sobre els valors d'una de les dues variables.

Es tracta, doncs, de determinar una recta que s'aproximi el màxim possible als punts del diagrama, és a dir, la que *"millor s'ajusta"* al núvol de punts. Això s'aconsegueix amb un mètode anomenat d'ajustament pels mínims quadrats.

Fet això s'obté que la recta de regressió de Y sobre X té per equació:

$$y - \overline{y} = \frac{s_{xy}}{s_x^2}\left(x - \overline{x}\right)$$

Aquesta recta s'utilitza per predir el valor de y_i si es coneix el corresponent valor de x_i.

Hi ha una altra recta de regressió, que és la de X sobre Y, i que es fa servir per predir el valor de $\mathbf{x_i}$ conegut el corresponent valor de $\mathbf{y_i}$. La seva equació és:

$$x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y})$$

Observant l'equació de cadascuna de les dues rectes de regressió es pot deduir fàcilment que el punt mitjà de la distribució pertany a ambdues rectes. Per tant, les dues rectes de regressió es tallen en el punt de coordenades $(\bar{x}, \bar{y})$, llevat que $r = -1$ o $r = 1$, en què les dues rectes de regressió són coincidents.

Important: obtenir les rectes de regressió i les prediccions de les variables només té sentit quan la correlació lineal és forta i es fan estimacions de valors propers a les dades observades.

No té sentit trobar les rectes de regressió quan el coeficient de correlació lineal sigui un valor proper a zero, ja que les possibles estimacions no serien gens fiables.

PER SABER-NE MÉS

UNA MICA D'HISTÒRIA

La teoria de la correlació i la regressió és molt recent, i el seu descobriment és degut al metge britànic Sir Francis Galton (1822-1911). Els seus treballs es van desenvolupar a l'entorn de l'estudi de l'herència i l'expressió matemàtica dels fenòmens vinculats a aquesta. El context històric ho afavoria, ja que va néixer el mateix any que Georges Mendel, amb el qual mantenia una gran amistat; a més, Galton era cosí de Charles Darwin.

El 1869, Galton va publicar el llibre *Hereditary Genius*, en el qual, a través de l'estudi de problemes d'herència, va arribar al concepte de correlació, i va ser el primer en assignar a un conjunt de variables un nombre que permetia obtenir una mesura del grau de relació existent entre aquestes. Va inferir que les persones excepcionalment altes solien tenir fills d'estatura menor que la dels seus progenitors. Aquesta observació va portar Galton a enunciar el seu *principi de la mediocritat*, aplicable a les talles d'una

generació respecte de les següents. Aquest va ser l'origen de l'actual anàlisi de la regressió.

L'observació de Galton és, sens dubte, certa, però el supòsit de la mediocritat és totalment fals i es considera actualment com una de les fal·làcies de la regressió. La justificació que es dóna avui dia a aquest fet és que els valors extrems d'una distribució són deguts en gran part a l'atzar, per això els factors genètics que produeixen una talla excepcional per excés o per defecte no passen als fills.

Els treballs de Galton van ser continuats i millorats pel matemàtic britànic Karl Pearson (1857-1936). A Pearson devem aportacions tan importants com el coeficient de correlació **r** de Pearson, la distribució $\chi 2$ o el test de Pearson per a l'estudi de la bondat de l'ajust d'una distribució empírica a una altra de teòrica.