

# CFGS DOCUMENTACIÓ I ADMINISTRACIÓ SANITÀRIES

MP 9. Validació i explotació de dades

UF1. Estadística i epidemiologia

NF 1. Estadística

## ESTADÍSTICA BIVARIANT

|                                                                                   |                                              |                                               |
|-----------------------------------------------------------------------------------|----------------------------------------------|-----------------------------------------------|
|  | CFGS DOCUMENTACIÓ I ADMINISTRACIÓ SANITÀRIES | MP9: VALIDACIÓ i EXPLOTACIÓ DE BASES DE DADES |
| Data de realització: 12/02/2018<br>Realitzat per: Margarita Pérez                 |                                              | UF1 NF1. Estadística bivariant                |

## **Estadística descriptiva bivariant**

### **OBJECTIUS**

- Conèixer la finalitat de l'estadística bivariant i treballar-la des d'un punt de vista descriptiu
- Entendre que els resultats numèrics no sempre expressen relacions entre variables, sinó que cal analitzar cada situació amb molta cura
- Fer ús de les taules creuades per estudiar la relació entre variables
- Analitzar les relacions entre variables mitjançant la comparació amb els diagrames de barres combinats
- Analitzar el núvol de punts d'una distribució bivariant (diagrama de dispersió)
- Interpretar la covariància i el coeficient de correlació lineal
- Decidir si el model lineal és adequat per descriure una població estadística i, si escau, representar gràficament la recta de regressió
- Fer prediccions a partir d'un model lineal

### **1. INTRODUCCIÓ A L'ESTADÍSTICA BIVARIANT.**

Fins ara hem considerat experiments en què observàvem un únic caràcter de cada individu. És molt freqüent en estadística que, quan es realitza un experiment, s'observin dos caràcters dels individus de la mostra. Aquests experiments es coneixen amb el nom d'experiments bidimensionals o bivariats i donen lloc a les distribucions bidimensionals.

Les **variables estadístiques bidimensionals** les representem pel parell (X,Y), on X és una variable unidimensional que pren els valors  $x_1, x_2, x_3, \dots, x_n$  i Y és una altra variable unidimensional que pren els valors  $y_1, y_2, y_3, \dots, y_n$ . Per tant, la variable estadística bidimensional (X,Y) pren aquests valors:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

## 2. TAULES BIDIMENSIONALS DE FREQUÈNCIES

La manera més habitual d'expressar els valors obtinguts en una distribució estadística bidimensional consisteix a representar-los mitjançant una **taula de contingència o taula de doble entrada**. Al marge superior s'escriuen els resultats d'una de les variables i al marge esquerre els valors de l'altre variable. En les caselles de la taula s'indiquen les freqüències absolutes, les freqüències relatives o percentatges corresponents a les dues variables simultàniament. S'utilitza per a representar dues variables qualitatives, dues quantitatives o una variable quantitativa i una altre qualitativa.

Per expressar les freqüències absolutes en relatives només hem de dividir les freqüències absolutes de cada casella pel nombre total de dades. Exemple :  $98 / 339 = 0.29$ ,  $120 / 339 = 0.35$

A les taules de doble entrada és freqüent incloure al final de cada fila i de cada columna, una fila i una columna de **totals**; aquestes distribucions de totals reben el nom de **distribucions marginals**.

| Nota de català | Nota de castellà |        |           | Total |
|----------------|------------------|--------|-----------|-------|
|                | Insuf            | Aprov. | Not./Exc. |       |
| Insuf.         | 98               | 22     | 0         | 120   |
| Aprov.         | 59               | 97     | 12        | 168   |
| Not./Exc.      | 1                | 20     | 30        | 51    |
| <b>Total</b>   | 158              | 139    | 42        | 339   |

La freqüència marginal representa la distribució de cadascuna de les variables sense tenir en compte l'altre variable (taula: color rosa = freqüència marginal castellà, groc = freqüència marginal català).

S'ha de tenir en compte que a partir de les dades creuades, es poden construir les distribucions marginals, però no al revés. Amb una taula creuada ja es poden estudiar les relacions entre les dues variables.

A partir de les distribucions marginals podem calcular la mitjana i la desviació tipus de cadascuna de les variables quantitatives, estudiant-les com una variable unidimensional. Per a variables qualitatives, com no se'ls pot sotmetre a operacions de sumes, restes i divisions, en venir expressades en escales nominals o ordinals no té sentit parlar de mitjanes marginals, condicionades, variàncies, etc; sí podríem calcular la moda, i la mediana si utilitzem escales ordinals.

A partir d'una distribució bidimensional, a més de les distribucions marginals, es poden obtenir **distribucions unidimensionals condicionades** de X i de Y.

Seguint l'exemple anterior, podríem estar interessats en conèixer les notes de català, però només dels alumnes que han obtingut un aprovat en castellà. És a dir, quina és la distribució condicionada del caràcter "notes de català" a que la "nota de castellà" sigui "aprovat" ( $X/Y=\text{aprovat}$ ). Podríem escriure també les distribucions condicionades ( $X/Y=\text{insuficient}$ ), ( $X/Y=\text{notable/excel·lent}$ ), ( $Y/X=\text{insuficient}$ ), ( $Y/X=\text{aprovat}$ ), ( $Y/X=\text{notable/excel·lent}$ ).

| Nota de català | Nota de castellà |        |           | Total      |
|----------------|------------------|--------|-----------|------------|
|                | Insuf            | Aprov. | Not./Exc. |            |
| Insuf.         | 98               | 22     | 0         | 120        |
| Aprov.         | 59               | 97     | 12        | 168        |
| Not./Exc.      | 1                | 20     | 30        | 51         |
| <b>Total</b>   | 158              | 139    | 42        | <b>339</b> |

La taula de doble entrada es pot transformar en una **taula simple**, on X i Y designen les variables unidimensionals i  $n_i$  la freqüència absoluta del parell  $(x_i, y_i)$ :

| Variable X<br>$X_i$ | Variable Y<br>$Y_i$ | $n_{ij}$ |
|---------------------|---------------------|----------|
|                     |                     |          |
|                     |                     |          |
|                     |                     |          |
|                     |                     |          |

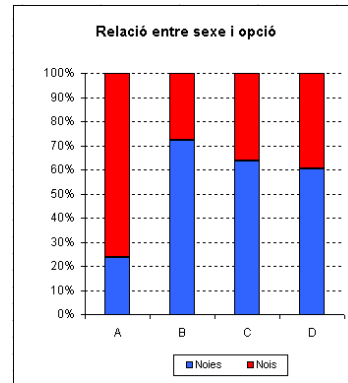
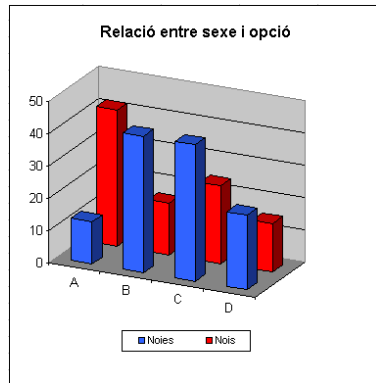
Si les variables X o Y són qualitatives la taula es denomina **taula de contingència**.

### **3. REPRESENTACIÓ GRÀFICA D'UNA VARIABLE BIDIMENSIONAL**

#### **3.1 VARIABLES QUALITATIVES**

En el cas de l'estudi de dues variables qualitatives X i Y, la visualització de les relacions entre les variables es realitza mitjançant la representació de diagrames de barres juxtaposades i/o apilades i diagrames de sectors múltiples.

Exemples:



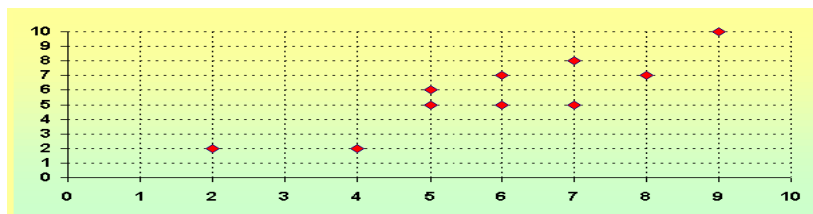
### 3.3 VARIABLE QUALITATIVA I QUANTITATIVA

En el cas d'una variable qualitativa i una quantitativa la relació entre les variables es pot evidenciar a través dels diagrames de caixa múltiples o histograma

### 3.3 VARIABLES QUANTITATIVES

El **diagrama de dispersió** (o bé **núvol de punts**) consisteix en una representació sobre un sistema d'eixos en què es dibuixa un punt (o un altre símbol) de coordenades  $(x_i, y_i)$  per cada individu de la població o mostra estudiada, on  $x_i$  és el valor observat de la primera variable i  $y_i$  el valor de la segona variable corresponent a aquest individu.

Exemple:

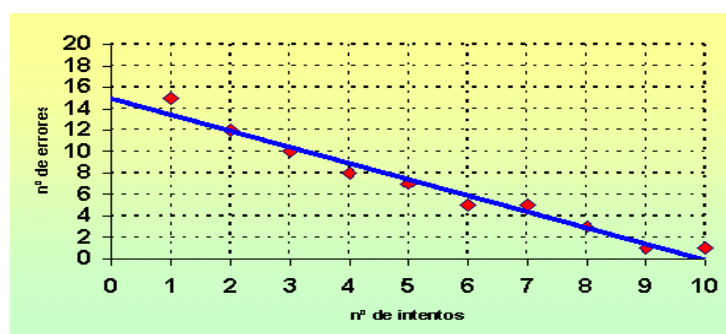


Aquesta representació ens permetrà detectar visualment el grau i el tipus de relació entre les variables. Segons la forma del núvol de punts podem establir:

- Relació lineal o curvilínia: si el núvol de punts es troba al voltant d'una recta o d'una corba, respectivament.
- Relació positiva o directa quan, a mesura que creix una variable, l'altra també creix.
- Relació negativa o inversa quan, a mesura que creix una variable, l'altra decreix.
- Relació nul·la quan no hi ha cap relació entre les dues variables. En aquest cas, els punts del diagrama estan escampats a l'atzar, sense formar cap línia, i es diu que les variables estan incorrelades.

Dit d'una altra manera, es pot observar si hi ha una tendència a l'associació dels *valors grans* d'una variable amb els *valors grans* de l'altra (**associació positiva o directa**) o bé dels *valors grans* d'una variable amb els *valors petits* de l'altra (**associació negativa o inversa**) o bé no es constata cap associació significativa de valors.

Hi ha associació positiva quan la major part dels punts del núvol queden al primer i al tercer quadrant; hi ha associació negativa quan la major part de punts queden al segon i al quart quadrant; si els punts es dispersen per tots els quadrants, no s'observa relació.



#### **4. CÀLCUL DE PARÀMETRES PER A VARIABLES QUANTITATIVES**

Després de comprovar que els valors d'una variable s'associen amb determinats valors de l'altra, ens podem preguntar de seguida si es podria mesurar numèricament el grau de relació o interdependència que se sospita que hi ha entre dues variables. Per això es defineixen els **coeficients de correlació**.

Per calcular el coeficient de correlació lineal corresponent a un conjunt bivariant de  $n$  dades emparellades,  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , es comença per calcular les mitjanes i desviacions tipus de les dues variables.

Tot seguit es calcula la **covariància**, nombre que mesura la variació conjunta global de les dues variables. La fórmula que defineix la covariància ( $\sigma_{xy}$ ) entre dues variables estadístiques és la següent:

$$\sigma_{xy} = \frac{\sum n_i \cdot x_i \cdot y_i}{N} - \bar{x} \cdot \bar{y}$$

Es pot observar la semblança formal de la definició de covariància i la de la variància que estudia la dispersió d'una sola variable.

La covariància és positiva si hi ha una tendència a l'associació positiva, i negativa en cas contrari. Si la covariància és zero ens indica que les 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.

Mesura la variabilitat absoluta i és fortament dependent de les unitats amb què s'expressen les dues variables (amb les quals també cal expressar la covariància). Per tant, convé passar a un altre valor que mesuri la relació intrínseca entre les variables, independentment de les unitats.

Per determinar els paràmetres estadístics és convenient disposar les dades en una taula d'aquest tipus:

| $x_i$ | $y_i$ | $n_i$      | $x_i \cdot n_i$ | $x_i^2 \cdot n_i$ | $y_i \cdot n_i$ | $y_i^2 \cdot n_i$ | $x_i \cdot y_i \cdot n_i$ |
|-------|-------|------------|-----------------|-------------------|-----------------|-------------------|---------------------------|
|       |       |            |                 |                   |                 |                   |                           |
|       |       |            |                 |                   |                 |                   |                           |
|       |       |            |                 |                   |                 |                   |                           |
|       |       |            |                 |                   |                 |                   |                           |
|       |       | $\sum n_i$ | $\sum x_i$      | $\sum x_i^2$      | $\sum y_i$      | $\sum y_i^2$      | $\sum x_i \cdot y_i$      |

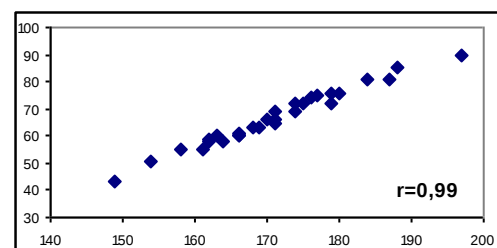
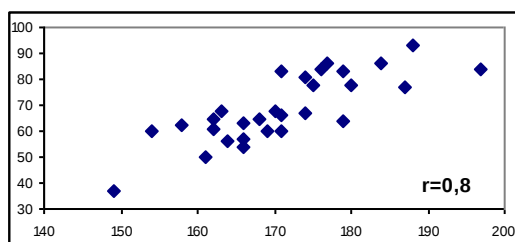
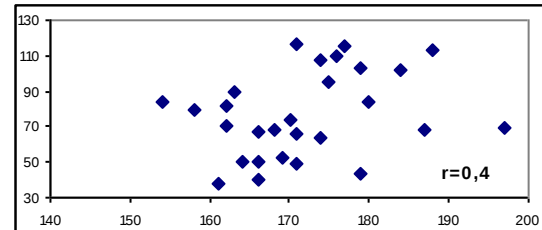
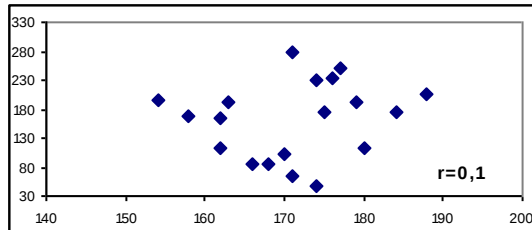
Aquest nou paràmetre és el **coeficient de correlació lineal de Pearson (r)** ens indica el grau d'intensitat de la relació lineal, en cas que n'hagi. Això es constata sobre el núvol de punts per l'agrupament dels punts molt a prop d'una línia recta. Els signe del coeficient de correlació lineal coincideix amb el signe de la covariància.

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

Propietats del coeficient de correlació lineal:

- Els valors del coeficient r sempre estan entre -1 i 1, inclosos els valors extrems.
- 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.
- 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.

- Si el valor  $r$  és proper a zero, tant en 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



Observacions sobre el coeficient de correlació lineal:

- Tal com passa també amb la mitjana i la desviació tipus, el coeficient de correlació de Pearson és fortament influït per l'existència de valors atípics, sobretot si la distribució no és molt nombrosa.
- El coeficient de correlació de Pearson és un *coeficient de correlació lineal*. Quan se'n vulgui donar la interpretació, doncs, cal tenir present que mesura *únicament* la tendència a un determinat tipus de relació entre dues variables: la lineal.
- Un coeficient de correlació lineal petit o nul pot indicar una total dispersió de les dues variables.

## **5. REGRESSIÓ LINEAL PER A VARIABLES QUANTITATIVES**

Una de les fites que es marca el treball científic, i també, doncs, l'estadística, és la possibilitat de predir resultats d'allò que pot passar quan es donen circumstàncies semblants a les ja estudiades.

Amb aquesta idea sorgeix el concepte de *regressió*, que rep el nom a partir d'una experiència de relacions entre variables duta a terme per Sir Francis Galton (1822-1911). Es tracta d'establir, en el marc de l'estadística bivariant, una funció que permeti descriure de la millor manera possible la relació entre les variables i estimar el valor que pot tenir una de les variables (dita *variable dependent* o

resposta) per a un element del que s'ha pogut mesurar el valor de l'altra variable (dita *variable independent, d'entrada o predictora*).

Anomenarem **funció d'ajust** o **de regressió** a la funció  $y = f(x)$ , que es vol emprar per descriure la relació entre dues variables:

- Per un cert valor  $x_i$  que pot prendre la variable d'entrada, s'anomena **ajust** el nombre que resulta en calcular  $f(x_i)$ .
- Per a una observació emparellada  $(x_i, y_i)$  del valor de les dues variables en un individu de la població, s'anomena **residu** respecte d'una funció d'ajust, la diferència entre el valor de la variable de resposta mesurat realment i el valor que ens donaria la funció d'ajust per a un determinat valor de la variable predictora;  $r_i = y_i - f(x_i)$ .
- Els residus mesuren, doncs, la discrepància entre les observacions reals i el model d'ajust que es vol estudiar.

La situació més habitual és aquella en què la funció  $y = f(x)$  és una funció polinòmica de primer grau (una recta). Quan hom s'ha convençut de la bondat d'un estudi del núvol de punts d'una distribució bivariant mitjançant una línia recta, de primer intuïtivament sobre el gràfic, i posteriorment amb el càlcul del coeficient de correlació, convé saber la forma adequada de dibuixar la recta que millor descriu la relació entre les variables. Aquesta recta rep el nom de **recta de regressió**.

La primera possibilitat intuïtiva per dibuixar la recta, com a primer supòsit, pot ser fer-ho a ull, mirant d'entendre quina és la recta que millor s'adapta al núvol de punts. Quan es vol obtenir la fórmula de la recta de regressió, es fa servir el mètode dels mínims quadrats. Es tracta de raonar quins han de ser els valors dels coeficients  $a$  i  $b$  que fan que la suma dels quadrats dels residus de les observacions reals respecte del model d'ajust donat per la funció  $y = a + b \cdot x$  tingui el valor mínim possible. Aquesta és la idea de millor ajust: residus com més petits millor!

Es pot demostrar que el pendent de la recta de regressió  $y = a + bx$  és:

$$b = \frac{\sigma_{xy}}{\sigma_x^2}$$

- La recta de regressió,  $y = a + bx$ , passa pel punt mitjà de la distribució i, per tant, es pot escriure així:

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2} \cdot (x - \bar{x})$$

Observeu que el signe del pendent de la recta de regressió coincideix amb el signe del coeficient de correlació. De fet, aquests dos nombres tenen el signe de la covariància.

La recta calculada s'anomena **recta de regressió de  $y$  sobre  $x$** , perquè parteix d'una variable d'entrada o predictora, que se suposa que està més sota el control de l'experimentador. Aquesta recta passa pel punt de coordenades  $(\bar{x}, \bar{y})$ , que és el centre de gravetat del núvol de punts..

Una vegada establert el model, en aquest cas el que ens dóna la recta de regressió, podem **fer prediccions del valor que pot tenir la variable de resposta**. El valor que ens dóna la funció model, per a un cert valor  $x_0$ , és a dir, el que hem definit com a **ajust**, és el valor que estimarem. Per exemple, les màquines de pesar que podem trobar al metro o en algunes farmàcies, que comparen el pes que obtenen amb un anomenat pes ideal (després de preguntar dades d'alçada, edat, etcètera), fan prediccions a partir d'un estudi de regressió.

Atès que el coeficient de correlació de Pearson mesura com d'estricta és l'acostament del núvol de punts a una línia recta, és ben intuïtiu que, com més gran és el valor absolut del coeficient més fiabilitat tenen les prediccions que es poden fer i més petit és el marge d'error amb què cal enunciar-les. Així doncs:

- Si  **$r$  és molt petit**, no té sentit fer cap mena d'estimacions o previsions.
- Si  **$r$  s'acosta a  $-1$  o a  $1$** , probablement els valors reals s'acostaran a les estimacions que fem.
- Si  **$r = -1$  o  $r = 1$** , les estimacions coincidiran amb els valors reals; la relació és **funcional**

Exemple: estudi de la alçada en grups familiars:

- alçada del fill = 85 cm + 0,5 alçada del pare ( $Y = 85 + 0,5 X$ )

Si el pare mesura 200 cm ¿quant mesura el fill?

S'espera (prediu)  $85 + 0,5 \times 200 = 185$  cm.

La recta de regressió serveix per fer estimacions, tenint en compte que:

- Els valors obtinguts són aproximacions en termes de probabilitat: és probable que el valor corresponent a  $x$  sigui 0.
- La fiabilitat és major com més fort sigui la correlació.
- La fiabilitat augmenta en augmentar el nombre de dades.
- L'estimació és més fiable per als valors de  $x$  propers a la mitjana.

La condició que la suma dels quadrats de les diferències  $(y_i - \bar{y})^2$  podria substituir-se per la de la suma dels quadrats de les diferències  $(x_i - \bar{x})^2$  no és correcta. En aquest cas obtindríem la recta de regressió de x sobre y. Si la correlació és forta les dues rectes difereixen.

## **6. REGRESSIÓ LINEAL PER A VARIABLES QUALITATIVES**

### **El llenguatge associat als contrastos d'hipòtesis**

De les accepcions que ens dona el diccionari per al mot hipòtesi, la que més d'acord està amb el sentit que se li dona en aquest capítol és aquesta: teoria provisionalment adoptada per explicar certs fets.

En els contrastos d'hipòtesis rep el nom d'**hipòtesi nul·la**,  $H_0$ , l'afirmació de la qual es vol confrontar la veracitat; en aquest cas, que el model estadístic que hem triat per al nostre estudi (i que es caracteritza mitjançant uns valors esperats) és l'adequat. Per contra,  $H_1$  és la hipòtesi alternativa

La finalitat del test consisteix a decidir, a la vista de les dades obtingudes i els càlculs necessaris, quina de les dues situacions següents es dona:

- Es pot entendre que les diferències entre els valors observats i els valors esperats són fruit de l'atzar i no ens han de fer dubtar del model, i en aquest cas podem seguir admetent com a vàlida la hipòtesi nul·la.
- La discrepància entre els valors obtinguts empíricament i el model teòric és significativament gran i podem rebutjar el model teòric.

En un experiment aleatori encaminat a la comprovació d'una hipòtesi estadística, s'anomena **risc de primera espècie** la probabilitat de rebuig de la hipòtesi nul·la en cas que sigui correcta. El màxim risc de primera espècie que es vol admetre en la realització d'una prova estadística de contrast d'unes dades amb un model rep el nom de **nivell de significació de la prova** i és representat habitualment per la lletra grega alfa (l'experiència dels que treballen en estadística demostra que la fixació d'un **nivell de significació del 5 %** és prou correcta).

El **nivell de confiança** se indica per  $1-\alpha$  y habitualmente se da en porcentaje  $(1-\alpha)\%$ . Els valors que s'utilitzen normalment per al nivell de confiança són 95%, 99% y 99,9%.

S'anomena **regió crítica** el subconjunt dels valors que, en cas de ser observats, ens fan rebutjar la hipòtesi nul·la. La regió crítica s'estableix moltes vegades no tant segons les freqüències observades,

sinó segons algun **estadístic** que es calcula a partir d'aquestes, en el benentès que un *estadístic* és una *característica numèrica calculada a partir de les dades obtingudes en una realització de l'experiència* i pot variar d'una realització a una altra.

### Prova d'independència de la $\chi^2$

Una anàlisi típic d'una taula de contingència és el estudi de la possible associació o relació entre les variables  $X$  i  $Y$ . Aquesta anàlisi es fa també amb un **test de khi quadrat**. En aquest cas, s'adopta com a **hipòtesi nul·la  $H_0$** : la independència de les variables en la població considerada i per calcular les freqüències esperades amb aquest enfocament s'adopta un punt de vista probabilístic. Recordeu que:

- Es diu que dos esdeveniments  $A$  i  $B$  són **independents** si i només si:

$$p(A \cap B) = p(A) \cdot p(B)$$

En aquest cas, si suposem que hi ha independència entre les variables (com correspon a la hipòtesi nul·la d'ajust al model que ara interessa) resulta que la probabilitat que un individu de la mostra pertanyi a una casella (fila  $f$ , columna  $c$ ) ha de ser igual a la probabilitat de pertànyer a la fila  $f$  multiplicat per la probabilitat de pertànyer a la columna  $c$ .

Pel que fa a la probabilitat, cal estimar els valors que prendrem com a tals a partir de les freqüències relatives que es dedueixen dels totals marginals obtinguts a partir de les dades.

- Si representem per  $T(f)$  el total d'observacions a la fila  $f$ , per  $T(c)$  el total d'observacions a la columna  $c$ , i per  $N$  el nombre total d'observacions, podem escriure:

$$p(\text{fila } f) = \frac{T(f)}{N}; \quad p(\text{columna } c) = \frac{T(c)}{N}$$

- Llavors, sota la hipòtesi nul·la d'independència, la probabilitat  $p(f,c)$  de pertànyer a la casella corresponent a la fila  $f$  i la columna  $c$  s'obté multiplicant les dues anteriors i és:

$$p(f,c) = \frac{T(f) \cdot T(c)}{N^2}$$

- Per tant, la freqüència esperada en cada casella,  $F_{\text{esp}}(f,c)$ , que s'obté multiplicant la probabilitat de pertànyer a la casella pel nombre total de dades observades, és, com en el cas del test d'homogeneïtat:

$$F_{\text{esp}}(f,c) = \frac{T(f) \cdot T(c)}{N}$$

Observem, doncs, que tot i que el plantejament inicial teòric ha estat diferent del que hem fet a l'apartat anterior, les freqüències esperades són les mateixes. Com que per calcular les freqüències

esperades fem servir els mateixos totals (de dades, per files i per columnes) que en el cas anterior, també coincideix el nombre de graus de llibertat amb què cal aplicar la prova de khi quadrat.

### Exemple

La taula següent dóna el color dels ulls (foscos, verds o blaus) i l'alçada classificada en dues categories (alts i baixos) d'una mostra de 100 individus extreta d'una determinada població. Volem estudiar si, a partir de la mostra recollida, es pot pensar que hi ha relació entre aquests caràcters en la població estudiada.

|        | Fosc | Verd | Blau | Total |
|--------|------|------|------|-------|
| Alts   | 23   | 9    | 8    | 40    |
| Baixos | 35   | 16   | 9    | 60    |
| Total  | 58   | 25   | 17   | 100   |

En aquest cas, cal fer un tractament simètric de les dues variables estudiades. Per calcular les **freqüències esperades**, es parteix de la taula observades, des de la qual, es poden calcular les freqüències esperades sota la hipòtesi d'independència que el test ens ha d'ajudar a contrastar. La taula de freqüències esperades és:

|        | Fosc | Verd | Blau | Total |
|--------|------|------|------|-------|
| Alts   | 23,2 | 10,0 | 6,8  | 40    |
| Baixos | 34,8 | 15,0 | 10,2 | 60    |
| Total  | 58   | 25   | 17   | 100   |

D'un cop d'ull ja veiem que hi ha molta semblança entre les freqüències observades i les freqüències esperades en cas d'independència, cosa que ens porta a pensar que podem mantenir la hipòtesi nul·la. No hi ha criteris amb aquesta mostra per rebutjar-la.

Com ho hem de fer amb el test de khi quadrat?

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$O_i$  = freqüència conjunta observada de la fila i i columna j

$E_i$  = freqüència conjunta esperada de la fila i i columna j

Graus de llibertat: (nombre de columnes  $m - 1$ ) \* (nombre de files  $n - 1$ )

Nivell de confiança = 0.95

Para N gran l'estadístic té una distribució aproximada ji-quadrat amb (Columnes-1) \* (Files-1) graus de llibertat si la hipòtesi nul·la és vertadera, és a dir, no existeix associació entre les variables. Per tant, la hipòtesi d'independència ha de rebutjar-se si el valor de l'estadístic de prova  $X^2$  calculat és major que  $X^2$  crític de taula:

Regla de decisió:

**Si  $X^2 \leq$  valor crític s'accepta  $H_0$**

**Si  $X^2 \geq$  valor crític es rebutja  $H_0$**

- Calculem l'estadístic  $X^2$  i resulta ser, en aquest exemple, de 0,52.
- El test s'aplica amb 2 graus de llibertat ( $m = 2$ ,  $n = 3$ ,  $(2 - 1) * (3 - 1) = 2$ ). En aquest cas (vegeu [taula](#)), la regió crítica amb un nivell de significació del 5 % està definida pel valor 5,99.
- Com que el valor observat és molt més petit (molt i molt!) que el que determina la regió crítica, no hi ha criteris amb les dades d'aquesta mostra per rebutjar la hipòtesi nul·la.
- Conclusió: en la població estudiada no hi ha cap relació entre l'alçada i el color dels ulls, es tracta de variables independents.

És clar que, a més de saber fer els càlculs per aplicar la prova de khi quadrat hem d'interpretar els resultats.