2.4 Dependència entre dos fets. Prova χ^2

Suposem que disposem de les històries clíniques de 5000 persones (ja mortes). Per a cada persona se sap si la mort va ser causada o no per càncer de pulmó i si l'individu era fumador o no. Es pot formar la taula següent amb aquests historials.

	Morts per càncer de pulmó	Morts per altres causes	Total
Fumadors	348	3152	3500
No fumadors	82	1418	1500
Total	430	4570	5000

(Encara que aquestes dades són hipotètiques, els registres reals dels hospitals presenten percentatges molt semblants.)

Es tracta de veure si hi ha relació entre el fet de ser fumador o no i el de morir per càncer de pulmó o altres causes. Per determinar si hi ha relació o no, s'efectua el que s'anomena "un test d'hipòtesis". Concretament, en aquest cas, la prova χ^2 (khi quadrat) d'independència.

Aquesta prova es basa en una idea bastant simple. Es comença suposant que no hi ha cap relació entre els dos factors que estem considerant, és a dir, són factors independents l'un de l'altre. Si aquest fos el cas, com que els fumadors representen la fracció $\frac{3500}{5000} = \frac{7}{10}$ sobre la població de morts i els no fumadors el $\frac{1500}{5000} = \frac{3}{10}$, esperaríem que del total de 430 morts per càncer de pulmó els $\frac{7}{10}$ fossin fumadors, és a dir $\frac{7}{10} \cdot 430 = 301$.

Fixem-nos que a partir d'aquesta dada esperada a la primera casella, 301, podem construir una nova taula amb tots els valors esperats si realment no hi hagués cap relació, ja que tenim fixats els totals.

	Morts per càncer de pulmó	Morts per altres causes	Total
Fumadors	301	3199	3500
No fumadors	129	1371	1500
Total	430	4570	5000

Observem que hi ha unes diferències entre el que s'ha observat a la mostra (la qual suposem representativa d'alguna població) i les que s'esperarien si fos independent el fet de fumar o no amb morir de càncer de pulmó. El que s'ha de decidir és si aquestes diferències són degudes a l'atzar (quan hom recull dades mai no s'ajusten exactament als models teòrics) o bé al fet que realment hi ha relació entre els dos factors.

El pas següent consisteix a calcular l'anomenat estadístic χ^2. Si anomenem O_i la freqüència observada a la mostra la casella i-èsima i E_i la freqüència esperada (sota la hipòtesi d'independència) a la mateixa casella, calculem l'expressió següent:

$$V = \sum_i \frac{(E_i - O_i)^2}{E_i}.$$

El resultat matemàtic important (i difícil de demostrar!) en què està basada la prova, és que la distribució de la variable aleatòria V, si no hi ha realment relació entre les variables, és aproximadament (si la mostra és gran) una llei coneguda anomenada χ^2 amb un grau de llibertat. Això vol dir que es poden calcular les probabilitats que aquesta variable prengui determinats valors.

Fixem-nos que si no hi hagués relació entre les variables, tindríem que els $\frac{(E_i-O_i)^2}{E_i}$ haurien de donar números petits, i per tant V donarà un valor més aviat petit. En canvi, si hi ha relació entre les dues variables, V hauria de prendre valors més aviat grans.

Calculem el valor de V en el nostre cas particular

$$V = \frac{(301-348)^2}{301} + \frac{(129-82)^2}{129} + \frac{(3199-3152)^2}{3199} + \frac{(1371-1418)^2}{1371} = 26.765.$$

Si realment no hi hagués relació, podríem calcular la probabilitat d'haver obtingut un resultat tan gran com aquest, és a dir $P\{V \geq 26.765\}$, ja que la distribució χ^2 amb un grau de llibertat és perfectament coneguda (de fet, quan el número de graus de llibertat és 1, es tracta d'una variable normal estàndard elevada al quadrat). Aquesta probabilitat és menor que 0.0001.

Observem que, suposant que no hi havia relació, hem obtingut uns resultats a la nostra mostra amb una probabilitat pràcticament nul·la. Quan passa això, és perquè és "gairebé"segur que la hipòtesi d'independència és falsa. Per tant, afirmaríem que aquestes dues variables estan relacionades.

Cal fer atenció al fet que només hem inferit una possible relació entre fumar i tenir càncer i no pas que el fet de fumar causi càncer. Podria ser que ambdues característiques estiguessin relacionades a causa d'una tercera variable no controlada que les influís alhora.

Si la taula té més de dues files o dues columnes aleshores s'han d'usar altres distribucions χ^2.

Distribució χ^2 amb un grau de llibertat

A la taula següent es donen les probabilitats que el valor de V obtingut seguint el procediment descrit anteriorment sigui més gran que uns certs números.

Valor de $V \geq$	1	2	3	4	5	10
Probabilitat $\leq$	0.318	0.159	0.084	0.046	0.025	0.002

La seva construcció es basa en el fet que per a una mostra prou gran es pot veure que

$$P\{V \geq v\} \simeq \frac{2}{\sqrt{2\pi}} \int_{\sqrt{v}}^{\infty} e^{-x^2/2}\, dx.$$

Aquesta última integral és difícil de calcular ja que es pot demostrar que la funció $e^{-x^2/2}$ no té cap primitiva (funció $g(x)$ tal que $g'(x) = e^{-x^2/2}$) expressable com una combinació finita d'operacions elementals, funcions trigonomètriques, logaritmes i exponencials, i per tant la regla de Barrow no és aplicable a la pràctica. En aquestes situacions el càlcul de l'àrea es fa per mètodes aproximats com el dels trapezis, basat en la idea següent: aproximar l'àrea real, $\int_a^b f(x)\,dx$, per la suma d'àrees de trapezis, $\sum_i \text{àrea}(T_i)$, amb els quals s'ha aproximat la figura real. Vegeu la figura: