⑤ We want to test

H_0: there's no relationship between response and ethnic origin

we use chi-square test of independency H_0: $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{(I-1)(J-1)}$$ reject H_0 if X^2 too big

where O_{ij}: observed frequencies

E_{ij}: expected frequencies

I: cells = number of rows = 9

J multi dist = number of columns = 2

Origin	Yes	No	Total
Italian	78	47	125
North Eu	56	29	85
Other Eu	43	29	72
English	53	32	85
Irish	43	30	73
Fr. Canadian	36	22	58
French	42	23	65
Portuguese	29	7	36
Total	380	219	599

$$E_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$$

$E_{11} = \frac{380\times125}{599} = 79.3$ $E_{21} = \frac{380\times73}{599} = 46.3$

$E_{12} = \frac{219\times125}{599} = 45.7$ $E_{22} = \frac{219\times73}{599} = 26.69$

$E_{21} = \frac{380\times85}{599} = 53.92$ $E_{61} = \frac{380\times58}{599} = 36.7$

$E_{12} = \frac{85\times219}{599} = 31.08$ $E_{62} = \frac{219\times58}{599} = 21.11$

$E_{31} = \frac{380\times72}{599} = 45.68$ $E_{71} = \frac{380\times65}{599} = 23.7$

$E_{32} = \frac{219\times72}{599} = 26.32$ $E_{72} = \frac{219\times65}{599} = 23.7$

$E_{41} = \frac{380\times85}{599} = 53.92$ $E_{81} = \frac{380\times36}{599} = 22.84$

$E_{42} = \frac{219\times85}{599} = 31.08$ $E_{82} = \frac{219\times36}{599} = 13.16$

$$X^2 = \frac{(78-79.3)^2}{79.3} + \frac{(47-45.7)^2}{45.7} + \cdots + \frac{(29-22.84)^2}{22.84} + \frac{(7-13.6)^2}{13.6}$$

$= 6.03$

$\chi^2_{(18-7)(2-1)}(1-0.05) = \chi^2_{(7)}(0.95) = 14.0671$

P-value $= P(X^2 > 6.03) = 0.5363$

$\Rightarrow$ We don't reject H_0

⑦ We test H_0: there is no relationship between major and grade

⇒ Chi-square test of homogeneity

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{((I-1)(J-1))}$$

Grade	Psychology	Biology	Other	Total
A	8	15	13	36
B	14	19	15	48
C	15	4	7	26
D-F	3	1	4	8
Total	40	39	39	118

$E_{11} = \frac{40 \times 36}{118} = 12.20$ $\quad E_{31} = \frac{40 \times 26}{118} = 8.81$

$E_{12} = \frac{39 \times 36}{118} = 11.90$ $\quad E_{32} = \frac{39 \times 26}{118} = 8.59$

$E_{13} = \frac{39 \times 36}{118} = 11.90$ $\quad E_{33} = \frac{39 \times 26}{118} = 8.54$

$E_{21} = \frac{40 \times 48}{118} = 16.27$ $\quad E_{41} = \frac{40 \times 8}{118} = 2.64$

$E_{22} = \frac{39 \times 48}{118} = 15.86$ $\quad E_{42} = \frac{39 \times 8}{118} = 2.64$

$E_{23} = \frac{39 \times 48}{118} = 15.86$ $\quad E_{43} = \frac{39 \times 8}{118} = 2.64$

$$\chi^2 = \frac{(8-12.2)^2}{12.2} + \frac{(15-11.9)^2}{11.9} + \dots + \frac{(4-2.64)^2}{2.64} = 12.18$$

$$\chi^2_{((4-1)(3-1))}(1-0.05) = \chi^2_{(6)}(0.95) = 12.5916$$

P-value = 0.0581

⇒ we don't reject H_0

(13)

# children younger sister \ # children older sister	0	1	2 ≥	Total
0	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
1	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
2 ≥	n_{31}	n_{32}	n_{33}	$n_{3\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	$n_{\cdot\cdot}$

a) The chi-square test of independence

$H_0: \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ $\quad i = 1,2,3$ $\quad j = 1,2,3$ where π_{ij} are the multinomial cell probabilities of the counts n_{ij}

Which is tested with $\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \sum_i\sum_j \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}} \sim \chi^2_{((I-1)(J-1))}$

b) $H_0: \sum \pi_{ij} = \sum \pi_{ji}$

$\Leftarrow$ $\pi_{ij} = \pi_{ji}$ $\quad i \neq j$

$i \neq j$ $\qquad\qquad$ $i = j$

$\Rightarrow \hat{\pi}_{12} = \hat{\pi}_{21} = \frac{n_{12}+n_{21}}{2n}$ $\qquad$ $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$

$\hat{\pi}_{13} = \hat{\pi}_{31} = \frac{n_{13}+n_{31}}{2n}$

$\hat{\pi}_{23} = \hat{\pi}_{32} = \frac{n_{23}+n_{32}}{2n}$

$$\chi^2 = \sum_i\sum_j \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \sum_i\sum_j \frac{(n_{ij}-n\hat{\pi}_{ij})^2}{n\hat{\pi}_{ij}} = \sum_{i=j}\frac{\left(n_{ij}-n\frac{n_{ij}}{n}\right)^2}{n\frac{n_{ij}}{n}} + \sum_{i\neq j}\frac{\left(n_{ij}-n\left(\frac{n_{ij}+n_{ji}}{2n}\right)\right)^2}{n\left(\frac{n_{ij}+n_{ji}}{2n}\right)}$$

$$= \sum_{i\neq j}\frac{\left(n_{ij}-\frac{n_{ij}+n_{ji}}{2}\right)^2}{\frac{n_{ij}+n_{ji}}{2}}$$

$\chi^2 \sim \chi^2_{(2)}$

df = number of free parameters under H_1 − number of estimated parameter under H_0

= (9−1) − 6

= 2

The null hypothesis are not equivalent, for example, if the younger sister had exactly the same number of children as the older a) would be false and b) true

(21)

	$\bar{D}$	D
$\bar{X}$	π_{00}	π_{01}
X	π_{10}	π_{11}

$$\text{odds}(D|X) = \frac{\pi_{11}}{\pi_{10}}$$

$$\text{odds}(D|\bar{X}) = \frac{\pi_{01}}{\pi_{00}}$$

$$\hat{\Delta} = \frac{n_{00} n_{11}}{n_{01} n_{10}}$$

$$\Delta = \frac{\text{odds}(D|X)}{\text{odds}(D|\bar{X})} = \frac{\pi_{11}\pi_{00}}{\pi_{01}\pi_{10}}$$

	Normal	Diabetic
BB	49	39
Bb or bb	4	12

$$\hat{\Delta} = \frac{49 \cdot 12}{39 \cdot 4} = 3.77$$

The odds of contracting diabetis are increased by a factor of 3.77 if patient has Bb or bb allele

Chapter 14

②										
x	0.34	1.38	-0.65	0.68	1.40	-0.88	-0.30	-1.18	0.50	-1.75
y	0.27	1.34	-0.53	0.35	1.28	-0.98	-0.72	-0.81	0.64	-1.59

a) Find $y = \alpha + \beta x$ $\quad y = \beta_0 + \beta_1 x$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

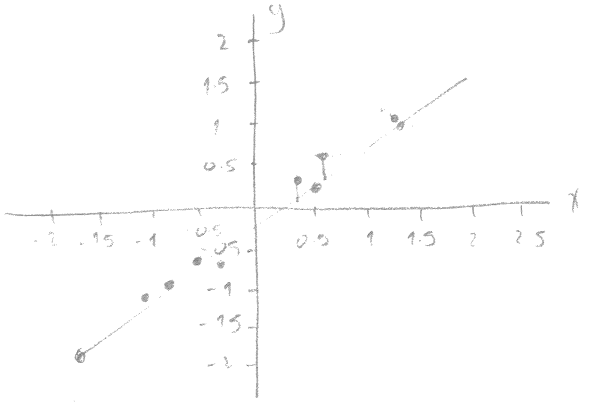
$\bar{x} = -0.046$

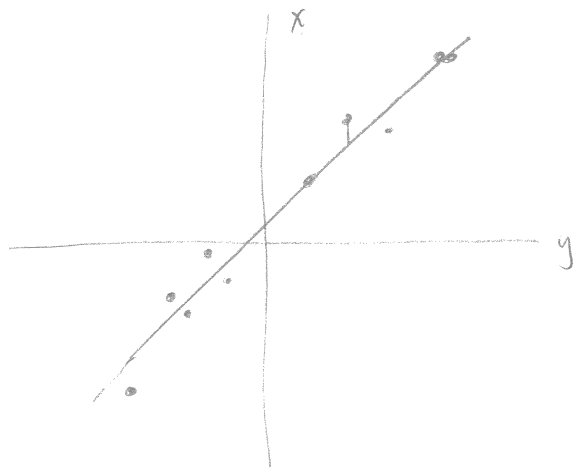
$\bar{y} = -0.075$

$$\hat{\beta}_1 = \frac{(0.34+0.046)(0.27+0.075) + \dots + (-1.75+0.046)(1.59+0.075)}{(0.34+0.046)^2 + \dots + (-1.75+0.046)^2} = \frac{9.4177}{10.4130} = 0.9044$$

$\hat{\beta}_0 = -0.075 - (0.9044)(-0.046) = -0.0334$

$\Rightarrow y = -0.0334 + 0.9044x$





b) Find $x = \alpha + \beta y$

$$\beta_0 = \bar{x} - \hat{\beta}_1 \bar{y} \qquad \hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (y_i - \bar{y})^2}$$

$$\hat{\beta}_1 = \frac{(0.27+0.075)(0.34+0.046) + \dots + (1.59+0.075)(-1.75+0.046)}{(0.27+0.075)^2 + \dots + (1.59+0.075)^2} = \frac{9.4177}{8.9267} = 1.0550$$

$\hat{\beta}_0 = -0.046 - (1.0550)(-0.075) = 0.0331$

$\Rightarrow x = 0.0331 + 1.0550y \Rightarrow$ c) Invert result from b) to obtain $y = -0.0314 + 0.9479x$ so, not the same line as in a)

c) because the distances (projections) to minimize in a) are not the same as in b which one to use is not a problem, its defined by which one is the independent (predictor) variable (in whose terms the properties, such as variance will be defined) and which one is the dependent (response) variable.

(11) If $\bar{x}=0$ $\Rightarrow$ $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$ under the assumptions of the standard statistical model

Standard statistical model $y_i = \beta_0 + \beta_1 x_i + e_i \quad i = 1, \dots, n \quad E(e_i) = 0$

$\text{Var}(e_i) = \sigma^2$

under those assumptions (Th B, p 549) $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \dfrac{-\sigma^2 \sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \dfrac{-\sigma^2 n \bar{x}}{n\sum x_i^2 - (\sum x_i)^2} = 0$

$\text{Corr}(\hat{\beta}_0, \hat{\beta}_1) = \dfrac{\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)}{\sqrt{\text{Var}(\hat{\beta}_0)\,\text{Var}(\hat{\beta}_1)}} = 0$

(15) Find the least squares estimate of β for $y = \beta x$ for (x_i, y_i) $i = 1, \dots, n$

The estimator will be found through $\min S(\beta) = \sum_{i=1}^n (y_i - \beta x_i)^2$

$\dfrac{\partial S}{\partial \beta} = -2\sum_{i=1}^n x_i(y_i - \beta x_i) \qquad \dfrac{\partial S}{\partial \beta} = 0 \iff \sum_{i=1}^n x_i (y_i - \beta x_i) = 0$

$\iff \sum x_i y_i = \beta \sum x_i^2 \iff \hat{\beta} = \dfrac{\sum x_i y_i}{\sum x_i^2}$

Another possibility is to find the mle (which coincides with $\hat{\beta}$). *solution in next page

(19) $\hat{Y} = X\hat{\beta} \qquad \hat{e} = Y - X\hat{\beta} \qquad \hat{\beta} = (X^T X)^{-1} X^T Y$

a) To show $X^T \hat{e} = 0$

$X^T(Y - X\hat{\beta}) = X^T Y - X^T X \hat{\beta} = X^T Y - X^T X (X^T X)^{-1} X^T Y = X^T Y - X^T Y = 0$

$\Rightarrow$ each column of X, x_i satisfy $x_i^T \hat{e} = 0 \Rightarrow$ they are orthogonal

b) If the model contains an intercept term then $X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \dots & x_{2p-1} \\ \vdots & & & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np-1} \end{bmatrix}$

we know that $X^T \hat{e} = 0$, in particular $0 = X_1^T \hat{e} = (1, 1, \dots, 1)\begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \vdots \\ \hat{e}_n \end{pmatrix} = \sum_{i=1}^n \hat{e}_i$

$\Rightarrow E\left(\sum_{i=1}^n e_i\right) = E(0) = 0$

$\star$ $y_i - \beta x_i = \varepsilon_i \sim N(0, \sigma^2)$

$\Rightarrow$ $y_i \sim (\beta x_i, \sigma^2)$

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{1}{2\sigma^2} (y_i - \beta x_i)^2 \right\}$$

$$l(\beta) = -\frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

$$\frac{\partial l}{\partial \beta} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} x_i (y_i - \beta x_i) = 0 \iff \sum x_i (y_i - \beta x_i) = 0$$

(35) $X = [X_1, X_2, X_1 + X_2]$

$$\hat{\beta} = (X^T X)^{-1} X^T Y = \left([X_1\ X_2\ X_1+X_2]^T [X_1\ X_2\ X_1+X_2]\right)^{-1} [X_1\ X_2\ X_1+X_2]^T Y$$

$$= \left[\begin{matrix} X_1^T \\ X_2^T \\ X_1^T + X_2^T \end{matrix}\right] [X_1\ X_2\ X_1+X_2] \Bigg]^{-1} \left[\begin{matrix} X_1^T \\ X_2^T \\ X_1^T + X_2^T \end{matrix}\right] Y$$

$$= \left[\begin{matrix} X_1^T X_1 & X_1^T X_2 & X_1^T X_1 + X_2^T X_2 \\ X_2^T X_1 & X_2^T X_2 & X_2^T X_1 + X_2^T X_2 \\ X_1^T X_1 + X_2^T X_1 & X_1^T X_2 + X_2^T X_2 & X_1^T X_1 + X_1^T X_2 + X_2^T X_1 + X_2^T X_2 \end{matrix}\right]^{-1} \left[\begin{matrix} X_1^T \\ X_2^T \\ X_1^T + X_2^T \end{matrix}\right] Y$$

$\Rightarrow$ (3) = (1)+(2) $\Rightarrow$ $X^T X$ doesn't have full rank $\Rightarrow$ not possible to invert