Test av likafördelning (homogenitetstest) Kan användas för att testa om $Z_1, \ldots, Z_s$ har samma fördelning (oavsett vilken). Dela in de möjliga värdena i r kategorier $A_1, \ldots, A_r$. Observera n_i oberoende utfall av Z_i och låt x_{ij} vara antalet utfall i kategori j.

Modell: x_{ij} är utfall av X_{ij} där X_{ij} är $\text{Bin}(n_i, p_{ij})$. Formellt vill vi testa

$$H_0 : p_{1j} = p_{2j} = \cdots = p_{sj} \text{ alla } j = 1, \ldots, r$$

mot H_1 : inte H_0.

Observationerna sammanfattas av

	Kategori, j				
	1	2	$\cdots$	r	
Serie 1	x_{11}	x_{12}	$\cdots$	x_{1r}	n_1
Serie 2	x_{21}	x_{22}	$\cdots$	x_{2r}	n_2
$\vdots$			$\vdots$		$\vdots$
Serie s	x_{s1}	x_{s2}	$\cdots$	x_{sr}	n_s
Kolonnsumma:	m_1	m_2	$\cdots$	m_r	N

En gemensam fördelning $p_1, \ldots, p_r$ skattas med $(p_1)^*_{\text{obs}}, \ldots, (p_r)^*_{\text{obs}}$ där $(p_j)^*_{\text{obs}} = m_j/N$. En skattning av $n_i p_j$ är således $n_i(p_j)^*_{\text{obs}} = \frac{n_i m_j}{N}$. Vi förkastar en hypotes om en gemensam fördelning för stora värden på

$$q = \sum_{i,j} \sum \frac{\left(x_{ij} - \frac{n_i m_j}{N}\right)^2}{\frac{n_i m_j}{N}}$$

som om H_0 är sann är ett utfall av en (approximativt) $\chi^2((r-1)(s-1))$-fördelad stokastisk variabel.

Exempel: Likafördelning av lön för två observationsserier

	Årsinkomst 2002 [tkr]				
	$0 - 99.9$	$100.0 - 199.9$	$200.0 - 299.9$	$300.0-$	
Män:	7	20	15	8	$n_1 = 50$
Kvinnor:	18	22	7	3	$n_2 = 50$
Totalt, m_j:	25	42	22	11	$N = 100$
$(p_j)^*_{\text{obs}} = m_j/N$:	.25	.42	.22	.11	
Män, $n_1(p_j)^*_{\text{obs}}$:	12.5	21	11	5.5	50
Kvinnor, $n_2(p_j)^*_{\text{obs}}$:	12.5	21	11	5.5	50
	25	42	22	11	100

$$q = \sum_{i,j} \frac{(x_{ij} - \frac{n_i m_j}{N})^2}{\frac{n_i m_j}{N}} = 10.12$$

Ur $\chi^2(3)$-tabeller : $\chi^2_{0.01} = 11.3$ $\chi^2_{0.025} = 9.35$. Testets p-värde ligger mellan 1% och 2.5%.

Test av oberoende (kontingenstabell) Kan användas för att testa om den stokastiska variabeln Y är oberoende av Z. Dela in värdemängden för Y i s kategorier $B_1, \ldots, B_s$, och värdemängden för Z i r kategorier, $A_1, \ldots, A_r$. Gör N oberoende observationer på (Y, Z) och låt x_{ij} vara antalet utfall av (Y, Z) i kategori (i, j).

Modell: x_{ij} är utfall av X_{ij} där X_{ij} är $\text{Bin}(N, p_{ij})$, $p_{ij} = \mathsf{P}((Y, Z) \in (B_i, A_j))$. Om oberoende så är $p_{ij} = \mathsf{P}(Y \in B_i)\,\mathsf{P}(Z \in A_j)$.

Formellt vill vi testa

$$H_0 : p_{ij} = \mathsf{P}(Y \in B_i)\,\mathsf{P}(Z \in A_j) \text{ alla } i = 1, \ldots, s,\ j = 1, \ldots, r$$

mot H_1 : inte H_0.

Observationerna sammanfattas av

	Kategori, j				
	1	2	$\cdots$	r	**Radsumma**
Serie 1	x_{11}	x_{12}	$\cdots$	x_{1r}	n_1
Serie 2	x_{21}	x_{22}	$\cdots$	x_{2r}	n_2
$\vdots$			$\vdots$		$\vdots$
Serie s	x_{s1}	x_{s2}	$\cdots$	x_{sr}	n_s
Kolonnsumma:	m_1	m_2	$\cdots$	m_r	N

$\mathsf{P}(Y \in B_i)$ skattas med n_i/N och $\mathsf{P}(Z \in A_j)$ skattas med m_j/N så

$$Np_{ij} = \{\text{oberoende}\} = N\mathsf{P}(Y \in B_i)\,\mathsf{P}(Z \in A_j) \text{ skattas med } N\frac{n_i}{N}\frac{m_j}{N}.$$

Vi förkastar en hypotes om oberoende för stora värden på

$$q = \sum_{i,j}\sum \frac{\left(x_{ij} - \frac{n_i m_j}{N}\right)^2}{\frac{n_i m_j}{N}}$$

som om H_0 är sann är ett utfall av en (approximativt) $\chi^2((r-1)(s-1))$-fördelad stokastisk variabel.

Linjär regression

Betrakta följande modell:

$$Y(x) \text{ är } \mathrm{N}(\mu(x), \sigma)$$

det vill säga Y är en normalfördelad stokastisk variabel vars väntevärde beror på en underliggande parameter x. Vi kommer enbart att betrakta funktioner

$$\mu(x) = \alpha + \beta x,$$

det vill säga att väntevärdet för $Y(x)$ är linjär i x. Problemet nu är hur vi med observationer

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

skall skatta α och β.

Med minsta kvadratmetoden får vi att skattningarna av α och β är de värden som minimerar

$$Q(\alpha,\beta) = \sum_{i=1}^{n}(y_i - \mu(x_i))^2 = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2.$$

Derivering ger ekvationen

$$\begin{aligned}\frac{\partial}{\partial\alpha}Q(\alpha,\beta) &= -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = -2\left(\left(\sum_{i=1}^{n} y_i\right) - n\alpha - \beta\left(\sum_{i=1}^{n} x_i\right)\right)\\ &= -2n\left(\bar{y} - \alpha - \beta\bar{x}\right) = 0,\end{aligned}$$

eller $\alpha = \bar{y} - \beta\bar{x}$. Vidare,

$$\begin{aligned}\frac{\partial}{\partial\beta}Q(\alpha,\beta) &= \frac{\partial}{\partial\beta}\sum_{i=1}^{n}(y_i - (\bar{y} - \beta\bar{x} + \beta x_i))^2 = -2\sum_{i=1}^{n}(y_i - \bar{y} - \beta(x_i - \bar{x}))(x_i - \bar{x})\\ &= -2\left(\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - \beta\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = 0\end{aligned}$$

eller $\beta = (\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}))/\sum_{i=1}^n (x_i - \bar{x})^2$. Alltså, MK-skattningarna av α och β är

$$\begin{aligned} \alpha^*_{\text{obs}} &= \bar{y} - \beta^*_{\text{obs}}\bar{x} \\ \beta^*_{\text{obs}} &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})} = \frac{S_{xy}}{S_{xx}}. \end{aligned}$$

Skattningarna beskrivs av motsvarande stokastiska variabler:

$$\alpha^* = \overline{Y} - \beta^*\bar{x} \qquad \beta^* = \frac{S_{xY}}{S_{xx}}.$$

Man kan visa att de stokastiska variablerna α^* och β^* är beroende, men att β^* och $\overline{Y}$ är oberoende.

Notera att β^* är en linjärkombination av oberoende normalfördelade stokastiska variabler och alltså är normalfördelad. Detsamma gäller α^*.

Slutligen, variansen σ^2 skattas med

$$s^2 = \frac{1}{n-2}\sum_{i=1}^n (y_i - (\alpha^*_{\text{obs}} + \beta^*_{\text{obs}}(x_i - \overline{x})))^2 = \frac{1}{n-2}\left(S_{yy} - \beta^*_{\text{obs}} S_{xy}\right)$$

Variabeln S^2 är oberoende av β^* och $\overline{Y}$.

Sammanfattning:

$$\begin{aligned} \alpha^* &= \overline{Y} - \beta^*\bar{x} \text{ är N}\left(\alpha, \sigma\sqrt{\frac{1}{n} + \frac{1}{S_{xx}}}\right) \\ \beta^* &= \frac{S_{xY}}{S_{xx}} \text{ är N}\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right) \\ \frac{(n-2)S^2}{\sigma^2} &= \frac{Q(\alpha^*, \beta^*)}{\sigma^2} \text{ är } \chi^2(n-2) \end{aligned}$$

Fixera en punkt x_0. Ett konfidensintervall för $\mathsf{E}[Y(x_0)] = \alpha + \beta x_0$ erhålls som följer. En skattning av $\mu_0 = \alpha + \beta x_0$ är $\mu^*_{\text{obs}} = \alpha^*_{\text{obs}} + \beta^*_{\text{obs}} x_0$ som är ett utfall av den normalfördelade variabeln

$$\mu^*_0 = \alpha^* + \beta^* x_0$$

som har väntevärde

$$\mathsf{E}[\mu^*_0] = \mathsf{E}[\alpha^* + \beta^* x_0] = \alpha + \beta x_0 = \mu_0$$

och varians

$$\mathrm{V}(\mu^*_0) = \mathrm{V}(\alpha^* + \beta^* x_0) = \mathrm{V}\left(\overline{Y} + \beta^*(x_0 - \bar{x})\right) = \mathrm{V}\left(\overline{Y}\right) + (x_0 - \bar{x})^2 \mathrm{V}(\beta^*) = \frac{\sigma^2}{n} + \sigma^2\frac{(x_0 - \overline{x})^2}{S_{xx}}.$$

(Här utnyttjade vi att $\overline{Y}$ och β^* är oberoende!) Alltså är

$$\mu^*_0 \quad \text{N}\left(\mu_0, \sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right)$$

och ett konfidensintervall för μ_0 ges av

$$\mu_0 \in \mu^*_{\text{obs}} \pm t_{\gamma/2}\, s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

där t-kvantilen bestäms ur t-fördelningen med $n-2$ frihetsgrader.