χ^2-test

χ^2-testet kan användas för att testa hypoteser rörande flera andelar (test av fördelning), test av likafördelning (homogenitetstest) och test av oberoende (kontigenstest). Här är hypoteserna inte bara utsagor om en parameter utan om en hel fördelning: t.ex. testa

$$H_0 : X \text{ är Po}(c), \text{ något } c$$

mot

$$H_1 : X \text{ är inte Poissonfördelad.}$$

χ^2-testet utgår ifrån multinomialfördelningen.

Kategoridata: vi gör n oberoende mätningar av en storhet och grupperar observationerna i r stycken kategorier. Till exempel

”blå” ”grön” ”gul”

eller

$$(-\infty, 25] \qquad (25, 72) \qquad [72, \infty).$$

Antag att varje observation har sannolikhet p_i att hamna i kategori i, $i = 1, \ldots, r$. $\sum_{i=1}^r p_i = 1$. Låt X_i vara antalet observationer som hamnat i kategori i. Modell:

$$X_i \sim \text{Bin}(n, p_i), \quad i = 1, \ldots, r.$$

Det förväntade antalet observationer i kategori i är

$$\mathsf{E}[X_i] = np_i.$$

Notera att $X_1, \ldots, X_r$ inte är oberoende. $\sum_{i=1}^r X_i = n$. Vektorn $(X_1, \ldots, X_r)$ är multinomialfördelad

$$\mathsf{P}(X_1 = x_1, \ldots, X_r = x_r) = \frac{n!}{x_1! \cdots x_r!} p_1^{x_1} \cdots p_r^{x_r}.$$

Med $r = 2$ fås binomialfördelningen.

Sats.

$$Q = \sum_{i=1}^r \frac{(X_i - np_i)^2}{np_i} \overset{\text{approx}}{\sim} \chi^2(r-1)$$

Alla test framgent bygger på detta resultat. Approximationen fungerar bäst då $np_i \geq 5$, dvs. det förväntade antalet observationer i varje kategori får inte vara för litet.

Exempel (Test på andelar/fördelning): Antalet bilar på en parkeringsplats: Under 30 dagar observerades följande kategoridata

Kategori	1	2	3	4	
Antal bilar	0–3	4–5	6–7	8–	
Frekvens, x_i	9	9	9	3	$n = 30$

Låt X beskriva antalet bilar på parkeringsplatsen. Vi vill testa

$$H_0 : X \text{ är Po}(5.5) \quad \text{mot} \quad H_1 : X \text{ är inte Po}(5.5)$$

på nivå $\alpha = 0.10$. Om H_0 är sann så är

$$\begin{aligned} p_1 &= \mathsf{P}(X \leq 3) = 0.2017 \\ p_2 &= \mathsf{P}(4 \leq X \leq 5) = 0.32722 \\ p_3 &= \mathsf{P}(6 \leq X \leq 7) = 0.28057 \\ p_4 &= \mathsf{P}(8 \leq X) = 0.19051 \end{aligned}$$

Detta ger

Antal bilar	0–3	4–5	6–7	8–	
Frekvens, x_i	9	9	9	3	30
Förväntat värde, np_i	6.051	9.8166	8.417	5.7154	30

Vi jämför x_i med np_i genom

$$q = \sum_i \frac{(x_i - np_i)^2}{np_i} = 2.8357$$

som om H_0 är sann är ett utfall på en $\chi^2(3)$-fördelad stokastisk variabel. Om H_0 inte är sann så blir q stor. Vi förkastar H_0 för stora värden på q och ur tabell fås $\chi^2_{0.10} = 6.25$. Då $q < 6.25$ förkastas ej H_0. Det är inte orimligt att X är Poisson(5.5)-fördelad.

För att testa hypoteser som $H_0 : X$ är Po(c), något c. Skatta c med $c^* = 4.87$. Skattningen ger

$$p_1^* = 0.28424 \quad p_2^* = 0.3551 \quad p_3^* = 0.24084 \quad p_4^* = 0.11982$$

Nu jämförs x_i med de skattade förväntade antalen np_i^*

$$q = \sum_{i=1}^{r} \frac{(x_i - np_i^*)^2}{np_i^*} = 0.81709$$

som är ett utfall på en $\chi^2(4-1-1) = \chi^2(2)$-fördelad stokastisk variabel. Generellt,

(Antalet frihetsgrader) = (Antalet kategorier − 1) − (Antalet skattade parameterar).

Ur $\chi^2(2)$-tabeller fås $\chi^2_{0.10} = 4.61$ så vi förkastar inte H_0. Det är inte orimligt att X är Poissonfördelad.

Exempel (test av likafördelning (homogenitetstest)): Används t.ex. för att testa om $Y_1, \ldots, Y_s$ har samma fördelning (oavsett vilken). Bilda r kategorier och gör n_i mätningar på Y_i, $i = 1, \ldots, s$. Låt X_{ij} vara antalet gånger som man observerar ett utfall av Y_i i kategori j och inför $N = \sum_{i=1}^{s} n_i$.

	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
					N

Med $p_{ij} = \mathsf{P}(Y_i \in \{\text{"Kategori } j\text{"}\})$ kan vår nollhypotes kan skrivas som

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

Om $Y_1, \ldots, Y_s$ har samma fördelning skattas sannolikheten för $\mathsf{P}(Y \in \{\text{"Kategori } j\text{"}\})$ med

$$p_j^* = \frac{1}{N} \sum_{i=1}^{s} X_{ij}$$

De skattade förväntade antalet i kategori j i serie i är således $n_i p_j^*$. Dessa jämförs med X_{ij} enligt

$$Q = \sum_{i,j}\sum \frac{(X_{ij} - n_i p_j^*)^2}{n_i p_j^*}$$

som under H_0 är approximativt $\chi^2((r-1)(s-1))$-fördelad. Som vanligt förkastar vi H_0 för stora värden på Q.

Årsinkomst 1996 [tkr]	0–39.9	40.0–79.9	80.0–119.9	120.0–159.9	160.0–199.9	200.0–279.9	280.0–359.9	360.0–	
Män :	92	62	109	163	193	261	77	66	$n_1 = 1023$
Kvinnor:	100	196	198	225	199	114	26	15	$n_2 = 1073$
Totalt, m_j:	192	258	307	388	392	375	103	81	$N = 2096$
$p_j^* = m_j/N$:	.0916	.123	.146	.185	.187	.179	.0491	.0386	
Män, $n_1 p_j^*$:	93.7	125.9	149.8	189.4	191.3	183	50.3	39.5	1023
Kvinnor, $n_2 p_j^*$:	98.3	132.1	157.2	198.6	200.7	192	52.7	41.5	1073

$$q = \sum_{i,j} \frac{(x_{ij} - n_i p_j^*)^2}{n_i p_j^*} = 219.65$$

Ur $\chi^2(7)$-tabeller : $\chi^2_{0.0001} = 29.878$. Förkasta en hypotes om samma lönefördelning på nivå 0.0001.