

SF1901: SANNOLIKHETSLÄRA OCH STATISTIK

FÖRELÄSNING 13.

MER OM χ^2 -TEST OCH LIKNANDE

Jan Grandell & Timo Koski

25.02.2015



- χ^2 -test
- χ^2 -test med skattade parametrar
- små talens lag (Bortkiewicz)
- homogenitetstest
- oberoendetest

χ^2 -testet är ett så kallat "**goodness of fit**" -test. "**goodness of fit**" översätts ibland som **anpassningsgrad**

Den enklaste situationen:

Ett försök kan utfalla på r olika sätt: $A_1, A_2, \dots, A_r$. Låt $x_1, x_2, \dots, x_r$ vara antalet gånger som alternativen $A_1, A_2, \dots, A_r$ förekommer i n försök.

Låt $p_1, p_2, \dots, p_r$ vara givna sannolikheter, dvs $\sum_{i=1}^r p_i = 1$. Vi vill testa

$$H_0 : P(A_i) = p_i \text{ för } i = 1, \dots, r$$

mot

$$H_1 : \text{ej alla } P(A_i) = p_i.$$

För att göra detta bildar vi

$$Q_{\text{obs}} = \sum_{i=1}^r \frac{(x_i - np_i)^2}{np_i}.$$

Man kan visa att Q är approximativt $\chi^2(r-1)$ -fördelad under H_0 .

För att göra resultatet troligt, betraktar vi $r = 2$. Då gäller, med $X = X_1$ och $p = p_1$ att

$$\begin{aligned} Q &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} = \frac{(X - np)^2}{np} + \frac{(n - X - n(1 - p))^2}{n(1 - p)} \\ &= \frac{(X - np)^2}{np} + \frac{(X - np)^2}{n(1 - p)} = \frac{(X - np)^2}{np(1 - p)}. \end{aligned}$$

Eftersom X är $\text{Bin}(n, p)$ så gäller att $\frac{X - np}{\sqrt{np(1 - p)}}$ är appr. $N(0, 1)$. Således följer att $\frac{(X - np)^2}{np(1 - p)}$ är appr. $\chi^2(1)$.

Vi gör nu följande test:

Förkasta H_0 om

$$Q_{\text{obs}} > \chi^2_{\alpha}(r-1).$$

Om n är stort, har detta test approximativt signifikansnivån α .

Det går inte att generellt ange hur stort n skall vara för att approximationen skall vara tillfredsställande. Som tumregel brukar man använda: Tillse att insatta värden på np_j är minst lika med 5.

χ^2 -TEST: KAST MED EN TÄRNING

Vid 96 kast med en tärning erhöles följande antal ettor, tvåor, etc: 15, 7, 9, 20, 26, 19. Man önskar pröva om tärningen kan antas vara symmetrisk. Hypotesen H_0 blir alltså

$$H_0 : P(A_1) = 1/6, \dots, P(A_6) = 1/6.$$

χ^2 -metoden kan användas, ty alla $np_j = 96 \cdot 1/6 = 16$ och är alltså tillräckligt stora.

χ^2 -TEST: KAST MED EN TÄRNING

Man får

$$Q_{\text{obs}} = \frac{(15 - 16)^2}{16} + \frac{(7 - 16)^2}{16} + \frac{(9 - 16)^2}{16} + \frac{(20 - 16)^2}{16} + \frac{(26 - 16)^2}{16} + \frac{(19 - 16)^2}{16} = 16.0.$$

Antalet frihetsgrader är $f = r - 1 = 6 - 1 = 5$. Eftersom Q_{obs} något överstiger $\chi^2_{0.01}(5) = 15.1$ kan man påstå att tärningen inte är symmetrisk: Resultatet är signifikant**. Med dator erhålls $P = P(Q > 16.0) \approx 0.0068$ då Q är $\chi^2(5)$.

χ^2 -TEST MED SKATTNING AV PARAMETRAR

Ofta vill vi låta sannolikheterna $p_1, p_2, \dots, p_r$ bero av en okänd parameter $\theta = (\theta_1, \dots, \theta_s)$, och testa hypotesen

$$H_0 : P(A_i) = p_i(\theta), \text{ för } i = 1, \dots, r,$$

och för något värde på θ .



KTH Matematik

Skattar vi θ med ML-metoden, och bildar

$$Q_{\text{obs}} = \sum_{i=1}^r \frac{(x_i - np_i(\theta_{\text{obs}}^*))^2}{np_i(\theta_{\text{obs}}^*)},$$

så är Q approximativt $\chi^2(r - s - 1)$ -fördelad under H_0 .
Detta resultat kallas ibland för *stora χ^2 -satsen*.

χ^2 -TEST MED SKATTNING AV PARAMETRAR: ANTALET FRIHETSGRADER

Grundregeln är att antalet frihetsgrader fås av

antalet fria kvadratsummor — antalet skattade parametrar.

EXEMPEL: DE SMÅ TALENS LAG

En händelse av låg frekvens i en stor population följer en poissonfördelning.



EXEMPEL: DE SMÅ TALENS LAG

I en klassisk datamängd undersöktes antalet ihjälsparkade soldater vid 14 tyska armékårer från 1875 till 1894 (20 år). De $14 \cdot 20 = 280$ rapporterna fördelade sig som i tabellen.

Antal döda	Antal rapporter	Andel
0	144	0.5143
1	91	0.3250
2	32	0.1143
3	11	0.0393
4	2	0.0071
≥ 5	0	0
Summa	280	1

Totalt omkom $1 \cdot 91 + 2 \cdot 32 + 3 \cdot 11 + 4 \cdot 2 = 196$ soldater. Test om data verkar komma från en $\text{Po}(\theta)$ -fördelning med

$$\theta = \theta_{\text{obs}}^* = 196/280 = 0.7.$$

Vi gör nu indelningen $A_1 = "0 \text{ döda}"$, $A_2 = "1 \text{ död}"$, $A_3 = "2 \text{ döda}"$ samt $A_4 = "3 \text{ eller fler döda}"$. Vi klumpar ihop eftersom vi vill få minst 5 i alla kategorier. $\text{Po}(0.7)$ -fördelningen ger

$$p_1^* = e^{-0.7} = 0.4966, \quad p_2^* = 0.7e^{-0.7} = 0.3476, \quad p_3^* = \frac{0.7^2}{2} e^{-0.7} = 0.1217, \\ p_4^* = 1 - p_1^* - p_2^* - p_3^* = 1 - 1.945e^{-0.7} = 0.0341.$$

Detta ger de förväntade frekvenserna

$$280 p_1^* = 139.0, \quad 280 p_2^* = 97.3, \quad 280 p_3^* = 34.1, \quad 280 p_4^* = 9.6$$

och

$$Q_{\text{obs}} = \frac{(144 - 139.0)^2}{139.0} + \frac{(91 - 97.3)^2}{97.3} + \frac{(32 - 34.1)^2}{34.1} + \frac{(13 - 9.6)^2}{9.6} \approx 1.95$$

Q är approximativt $\chi^2(4 - 1 - 1) = \chi^2(2)$ om

H_0 : "data kommer från Poisson-fördelning"

är sann. Vi har $\chi^2_{0.05}(2) = 5.99$ så H_0 förkastas inte på nivån 5 %.
Slutsatsen är alltså att data på intet sätt strider mot påståendet att antalet ihjälsparkade soldater per år och armékår är Poisson-fördelat.
Om vi på *förhand* hade velat testa om data kom från just $Po(0.7)$ -fördelningen hade vi jämfört med $\chi^2_{0.05}(3) = 7.81$.

L. Bortkiewicz published a book (in German) The Law of Small Numbers in 1898. In this he was the first to note that events with low frequency in a large population followed a Poisson distribution even when the probabilities of the events varied. A striking example was the number of soldiers killed by horse kicks per year per Prussian army corps. Fourteen corps were examined, each for twenty years. For over half the corps-year combinations there were no deaths from horse kicks; for the other combinations the number of deaths ranged up to four. Presumably the risk of lethal horse kicks varied over years and corps, yet the over-all distribution was remarkably well fitted by a Poisson distribution.

DE SMÅ TALENS LAG: A DIFFERENT STORY

The Poisson distribution was discovered through observation of the process of fermentation in beer at the Guinness company. The t-statistic is a measure of how much confidence can be placed in judgments made from small samples. W.S. Gossett invented that measure to enable the quality of brews to be monitored in a cost- effective manner.

The scholars behind the stout; Financial Times 27 December 2005

http://www.johnkay.com/in_action/422



Vi återgår nu till exemplet i början, med ett försök som kan utfalla på r olika sätt: $A_1, A_2, \dots, A_r$. Antag nu att vi har s försöksserier om $n_1, \dots, n_s$ försök vardera. Låt x_{ij} vara antalet gånger som alternativet A_j förekommer i i te försöksserien.

HOMOGENITETSTEST

Antag att s serier utförts med följande resultat:

Serie	Absoluta frekvenser för				Antal försök
	A_1	A_2	$\dots$	A_r	
1	x_{11}	x_{12}	$\dots$	x_{1r}	n_1
2	x_{21}	x_{22}	$\dots$	x_{2r}	n_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
s	x_{s1}	x_{s2}	$\dots$	x_{sr}	n_s
Summa	$x_{\cdot 1}$	$x_{\cdot 2}$	$\dots$	$x_{\cdot r}$	n

Som framgår av tabellen indikerar en punkt i index en summation.

Vi anser att serierna är *homogena* om hypotesen

$$H_0 : P(A_i) = p_i, \text{ för } i = 1, \dots, r \text{ i alla serierna.}$$

För att testa H_0 bildar vi

$$Q_{\text{obs}} = \sum_{i=1}^s \sum_{j=1}^r \frac{(x_{ij} - n_i p_j^*)^2}{n_i p_j^*},$$

där

$$p_j^* = (p_j^*)_{\text{obs}} = \frac{\sum_{i=1}^s x_{ij}}{\sum_{i=1}^s n_i}.$$

Man kan visa att Q är approximativt $\chi^2((r-1)(s-1))$ -fördelad under H_0 .

Frihetsgraderna fås på följande sätt:

antalet fria kvadratsummor – antalet skattade parametrar

$$= s \cdot (r - 1) - (r - 1) = (r - 1)(s - 1).$$

EXEMPEL: HOMOGENITETSTEST

En opinionsundersökning (i USA), där vi ställer en politiskt känslig fråga med svar: instämmer vs. instämmer ej. Man undrar om svaren är beroende intervjuarens kön. Vi studerar endast mäns svar.

EXEMPEL

Svar	Absoluta frekvenser för		
	Män instämmer	Män instämmer ej	
Man intervjuar	560	240	800
Kvinna intervjuar	308	92	400
Summa	868	332	1200

$$n_1 p_1^* = 800 \cdot \frac{868}{1200} = 578.67, n_1 p_2^* = 800 \cdot \frac{332}{1200} = 221.33$$

$$n_2 p_1^* = 400 \cdot \frac{868}{1200} = 289.33, n_2 p_2^* = 400 \cdot \frac{332}{1200} = 110.67$$

Teststorheten är

$$\begin{aligned} Q_{\text{obs}} &= \frac{(560 - 578.67)^2}{578.67} + \frac{(240 - 221.33)^2}{221.33} \\ &\quad + \frac{(308 - 289.33)^2}{289.33} + \frac{(92 - 110.67)^2}{110.67} \\ &= 0.6062 + 1.574 + 1.204 + 3.149 = 6.529. \end{aligned}$$

Vi väljer $\alpha = 0.05$. Q approximativt $\chi^2(1)$ -fördelad,

$$Q_{\text{obs}} = 6.529$$

Kvantilen $\chi^2_{0.05}(1) = 3.84$.

p-värdet är

$$P(Q \geq 6.529) = 0.011.$$

Nollhypotesen om homogena proportioner förkastas på signifikansnivån 0.05.

Vi tar nu ett stickprov om n enheter, där varje enhet klassificeras efter två egenskaper, A och B . Vi kan skriva detta i en *kontingenstabell*, lik den tabell vi hade i hogenitetstestet.

OBEROENDETEST & KONTINGENSTABELL

Antag att s serier utförts med följande resultat:

Serie Egenskap	Absoluta frekvenser för					Antal försök
	A_1	A_2	$\dots$	A_r	Total	
B_1	x_{11}	x_{12}	$\dots$	x_{1r}	$x_{1\cdot}$	
B_2	x_{21}	x_{22}	$\dots$	x_{2r}	$x_{2\cdot}$	
$\vdots$			$\vdots$		$\vdots$	
B_s	x_{s1}	x_{s2}	$\dots$	x_{sr}	$x_{s\cdot}$	
Total	$x_{\cdot 1}$	$x_{\cdot 2}$	$\dots$	$x_{\cdot r}$	n	

Vi vill nu testa hypotesen

$$H_0 : P(A_j \cap B_i) = P(A_j)P(B_i), \text{ för alla } i \text{ och } j.$$

För att testa H_0 bildar vi

$$Q = \sum_{i=1}^s \sum_{j=1}^r \frac{(x_{ij} - np_{i\cdot}^* p_{\cdot j}^*)^2}{np_{i\cdot}^* p_{\cdot j}^*},$$

där

$$p_{i\cdot}^* = (p_{i\cdot}^*)_{\text{obs}} = \frac{x_{i\cdot}}{n} \quad \text{och} \quad p_{\cdot j}^* = (p_{\cdot j}^*)_{\text{obs}} = \frac{x_{\cdot j}}{n}.$$

Man kan även här visa att Q är approximativt $\chi^2((r-1)(s-1))$ -fördelad under H_0 .



Frihetsgraderna fås på följande sätt:

antalet fria kvadratsummor – antalet skattade parametrar

$$= (sr - 1) - [(r - 1) + (s - 1)] = sr - r - s + 1 = (r - 1)(s - 1).$$

OBSERVERA! Även om homogenitetstestet och kontingenstabellen numeriskt och statistiskt är lika, så är det olika test.