

SF1901 Sannolikhetsteori och statistik I

Jimmy Olsson

Föreläsning 14
13 december 2016



Idag

χ^2 -metoden

Test av given fördelning

Homogenitetstest



Idag

χ^2 -metoden

Test av given fördelning

Homogenitetstest



Exempel: Väljarundersökning*

- ▶ I SKOP:s väljarbarometer i december 2012 sade sig 473 personer från hela riket stödjade det röd-gröna blocket, 440 allianspartierna och 90 gruppen partier övriga. I januari 2012 var de motsvarande siffrorna i SKOP:s väljarbarometer 479, 493 och 73.
- ▶ undersök om det överhuvudtaget skett någon förändring i väljaropinionen mellan de två undersökningstillfällena, där alla 3 grupperna (rödgröna blocket, allianspartierna samt övriga) analyseras samtidigt. Nivå 5%.

*Tentamen mars 2013, Uppg. 5(b)

Preludium: χ^2 -fördelningen igen

- ▶ Följande resultat kommer att vara till användning.

Sats

Om $X \in N(0, 1)$ så är $X^2 \in \chi^2(1)$.

Sats

Om $X \in \chi^2(k_1)$ och $Y \in \chi^2(k_2)$ är oberoende så gäller att

$$X + Y \in \chi^2(k_1 + k_2).$$



Idag

χ^2 -metoden

Test av given fördelning

Homogenitetstest



Test av given fördelning

- ▶ Vi utför en serie n oberoende försök, av vilka vart och ett kan utfalla på r olika sätt A_1, A_2, \dots, A_r med sannolikheterna $\mathbb{P}(A_1), \mathbb{P}(A_2), \dots, \mathbb{P}(A_r)$.
- ▶ Låt x_1, x_2, \dots, x_r vara de absoluta frekvenser som erhålls i en sådan serie.
- ▶ Det gäller alltså att $\sum_{j=1}^r \mathbb{P}(A_j) = 1$ och $\sum_{j=1}^r x_j = n$.
- ▶ Vi vill testa

$$H_0 : \mathbb{P}(A_1) = p_1, \mathbb{P}(A_2) = p_2, \dots, \mathbb{P}(A_r) = p_r$$

för några specificerade sannolikheter p_1, p_2, \dots, p_r (där $\sum_{j=1}^r p_j = 1$).

- ▶ Mothypotesen är helt enkelt att H_0 inte gäller.



Test av given fördelning (forts.)

- ▶ Notera att antalet utfall i A_j har $\text{Bin}(n, p_j)$ -fördelning.
- ▶ Då förväntat antal utfall i A_j är np_j använder vi testvariabeln

$$Q = \sum_{j=1}^r \frac{(x_j - np_j)^2}{np_j}.$$

- ▶ Under H_0 kan testvariabeln Q visas vara *approximativt* $\chi^2(r-1)$ -fördelad.
- ▶ Testet (på nivån α) utförs följaktligen genom att

$$\begin{cases} \text{förkasta } H_0 & \text{om } Q > \chi_{\alpha}^2(r-1), \\ \text{ej förkasta } H_0 & \text{om } Q \leq \chi_{\alpha}^2(r-1). \end{cases}$$



Test av given fördelning (forts.)

- ▶ Tumregel: approximationen fungerar väl om $np_j \geq 5$ för alla j .
- ▶ Man bör alltså inte välja r för stort så att p_j :na blir för små.
- ▶ Dock vill man ha r stort för att få en detaljerad indelning av resultaten och test med stor styrka. Kräver att n är stort.



Exempel: manipulerad roulett?



Exempel: manipulerad roulett?[†]

- ▶ Man misstänker att ett roulettebord på ett kasino är manipulerat och genomför ett test med 8000 försök. Om rouletten är korrekt skall röd, svart och grön (nollan) komma upp i proportionerna 18:18:1. Testresultatet gav röd: 3751, svart: 4018, grön: 231.
- ▶ Avgör med felrisken 1% om rouletten är korrekt. Det måste klart framgå av svaret vad slutsatsen är.

[†]Tentamen maj 2014, Uppg. 4

Test av fördelning med skattade parametrar

- ▶ χ^2 -testet kan utvidgas till följande situation.
- ▶ Antag att vi har ett stickprov av storlek n .
- ▶ Vi vill pröva hypotesen H_0 att data härrör från någon fördelning med parameter θ .
- ▶ Vi betraktar då åter en indelning A_1, A_2, \dots, A_r av data och låter

$$H_0 : \mathbb{P}(A_1) = p_1(\theta), \mathbb{P}(A_2) = p_2(\theta), \dots, \mathbb{P}(A_r) = p_r(\theta),$$

där $p_j(\theta)$ är sannolikheten att erhålla A_j under H_0 .

- ▶ Vi låter x_j beteckna antalet observationer som ligger i A_j .
- ▶ Om θ är okänd ersätts denna med en under H_0 konstruerad skattning θ^* baserad på stickprovet.



Test av fördelning med skattade parametrar (forts.)

- ▶ Testvariabeln blir

$$Q = \sum_{j=1}^r \frac{(x_j - np_j(\theta^*))^2}{np_j(\theta^*)}.$$

- ▶ Under H_0 kan testvariabeln Q visas vara *approximativt* $\chi^2(r - k - 1)$ -fördelad, där k är *antalet skattade parametrar*.



Exempel: ihjälsparkade kavallerister

► Saxat ur Blom:

Exempel 13.18 *Test av Poisson-fördelning*

I en klassisk datamängd undersöktes antalet ihjälsparkade soldater vid 14 tyska armékårar 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

- Vi vill testa om data verkar komma från en $Po(\theta)$ -fördelning. Gör ett test på nivån 5%!



Idag

χ^2 -metoden

Test av given fördelning

Homogenitetstest



Homogenitetstest

- ▶ Antag att man utfört s serier av oberoende försök (av ev. olika längd) och erhållit data på formen

serie	A_1	A_2	\cdots	A_r	antal försök
1	x_{11}	x_{12}	\cdots	x_{1r}	n_1
2	x_{21}	x_{22}	\cdots	x_{2r}	n_2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
s	x_{s1}	x_{s2}	\cdots	x_{sr}	n_s
summa	$x_{.1}$	$x_{.2}$	\cdots	$x_{.r}$	n

- ▶ Vi vill undersöka om serierna kan anses *homogena*, dvs. om sannolikheterna $p_1 = \mathbb{P}(A_1)$, $p_2 = \mathbb{P}(A_2)$, \dots , $p_r = \mathbb{P}(A_r)$ är samma för alla serier.
- ▶ I regel är fördelningen (sannolikheterna p_1, \dots, p_r) helt ospecificerad.



Homogenitetstest (forts.)

- ▶ Genom att tillämpa χ^2 -metoden får vi testvariabeln

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

där skattningarna av sannolikheterna ges av

$$p_j^* = \frac{1}{n} \sum_{i=1}^s x_{ij} = \frac{x_{.j}}{n}.$$

- ▶ Under H_0 kan testvariabeln Q visas vara *approximativt* χ^2 -fördelad med

$$s(r - 1) - (r - 1) = (r - 1)(s - 1)$$

frihetsgrader. Tumregel: åter $n_i p_j^* \geq 5$ för alla i, j .



Exempel: Väljarundersökning*

- ▶ I SKOP:s väljarbarometer i december 2012 sade sig 473 personer från hela riket stödjade det röd-gröna blocket, 440 allianspartierna och 90 gruppen partier övriga. I januari 2012 var de motsvarande siffrorna i SKOP:s väljarbarometer 479, 493 och 73.
- ▶ undersök om det överhuvudtaget skett någon förändring i väljaropinionen mellan de två undersökningstillfällena, där alla 3 grupperna (rödgröna blocket, allianspartierna samt övriga) analyseras samtidigt. Nivå 5%.

*Tentamen mars 2013, Uppg. 5(b)

Exempel: Väljarundersökning (forts.)

- ▶ Lösning: homogenitetstest med $r = 3$ och $s = 2$ och data enligt tabellen

Barometer	Rödgröna	Alliansen	Övriga	Totalt
December 2012	473	440	90	1003
Januari 2012	479	493	73	1045
Totalt	952	933	163	2048

- ▶ Vi erhåller den observerade testvariabeln

$$\begin{aligned} Q_{\text{obs}} &= \frac{(473 - 1003 \frac{952}{2048})^2}{1003 \frac{952}{2048}} + \frac{(440 - 1003 \frac{933}{2048})^2}{1003 \frac{933}{2048}} + \frac{(90 - 1003 \frac{163}{2048})^2}{1003 \frac{163}{2048}} \\ &+ \frac{(479 - 1045 \frac{952}{2048})^2}{1045 \frac{952}{2048}} + \frac{(493 - 1045 \frac{933}{2048})^2}{1045 \frac{933}{2048}} + \frac{(73 - 1045 \frac{163}{2048})^2}{1045 \frac{163}{2048}} \\ &= 3.96. \end{aligned}$$

- ▶ Slutsats?



Nästa föreläsning

- ▶ Oberoendetest,
- ▶ summering,
- ▶ genomgång av extenta (11 januari 2016).

