F15 χ^2-TEST

Kap. 13.10
Linnéa Gustafsson
linneag2@kth.se

- χ^2-test:
  - Test av fix nollhypotes
  - Homogenitetstest
  - Test av oberoende
  - Nollhypotes med skattade parametrar

χ^2-test

TEST AV GIVEN FÖRDELNING

Används när nollhypotesen är att vi har en viss sannolikhetsfunktion

D.v.s. $H_0: \; P(A_1) = p_1, \; P(A_2) = p_2, \; \dots$

antalet resultat

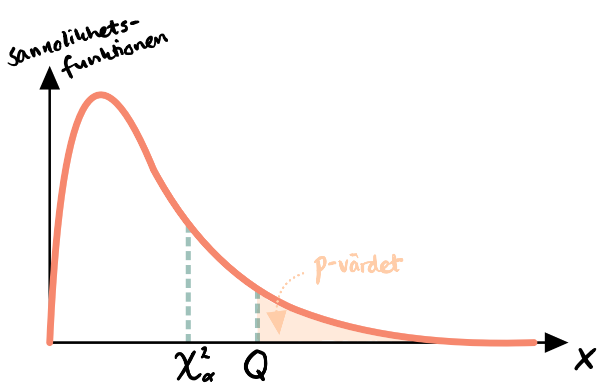
Gå in i § 14.3: $\quad Q = \sum_{i=1}^{r} \frac{(x_i - np_i)^2}{np_i}$

Eg. Q_{obs}

Om Q stort så förkastas H_0 på risknivån α.

När är Q stort?
När $Q > \chi^2_\alpha(f)$ förkastas H_0.

Tab. 4: $P(X > \chi^2_\alpha(f)) = \alpha$



Påminner om binomialfördeln.

Liten bakgrund till Q

$$Q = \sum_{i=1}^{r} \left(\frac{X_i - np_i}{\sqrt{np_i}} \right)^2 = \sum_{i=1}^{r} Y_i^2 = [\, Y_i\text{:na approximativt } N(0,1) \text{ om alla } np_i \geq 5 \,] \Rightarrow$$

$\Rightarrow$ Q approximativt χ^2-fördelad, se § 10 i F.S.

Tenta 2019-01:15

Givet: Ett roulettebord

Vi spelar 8000 ggr.

Korrekt bord: röd : svart : noll = 18 : 18 : 1

Testresultatet: röd: 3751, svart: 4018, noll: 231

$\alpha = 0.01$

Här är nollhypotesen att vi har en viss sannolikhetsfunktion $\Rightarrow$ Test av given fördelning $\Rightarrow$ § 14.3

$n = 8000$

A_1 - rött, A_2 - svart, A_3 - nollor

$r = 3$

$x_1 = 3751$, $x_2 = 4018$, $x_3 = 231$

H_0: $P(A_1) = p_1 = \frac{18}{37}$

$P(A_2) = p_2 = \frac{18}{37}$

$P(A_3) = p_3 = \frac{1}{37}$

Villkor för χ^2-test är att alla $np_i > 5$

$np_3 = 8000 \cdot \frac{1}{37} > 5 \Rightarrow \chi^2$-test ok

$$Q = \sum_{i=1}^{r} \frac{(x_i - np_i)^2}{np_i} \approx 10.20$$

Enklare kanske att använda beräkningsaspekt, se § 14.3

$$Q = \sum_{i=1}^{r} \frac{x_i^2}{np_i} - n$$

$$\chi^2_\alpha(r-1) = \chi^2_{0.01}(2) = [\text{tab. } 4] = 9.21$$

$Q = 10.20 > \chi^2_{0.01}(2) = 9.21 \Rightarrow$ Förkastar H_0 på nivån 1% $\Rightarrow$ Fuskroulette!

Ex 13.18

Y = antal ihjälsparkade soldater i en rapport

Antal döda	Antal rapporter
0	144
1	91
2	32
3	11
4	2
≥5	0

H_0 : $Y \in Po$ - fördelningen

D.v.s. nollhypotesen är att vi har en viss sannolikhetsfunktion.

$\Rightarrow$ Test av given fördelning $\Rightarrow$ § 14.3

n = antal rapporter = 280

A_1 – 0 döda, A_2 – 1 död, A_3 – 2 döda, A_4 – 3 döda, A_5 – 4 döda, A_6 – fler än 4 döda

$r = 6$

$x_1 = 144, \quad x_2 = 91, \quad x_3 = 32, \quad x_4 = 11, \quad x_5 = 2, \quad x_6 = 0$

$$p_1 = P(A_1) = P(Y=0) = [Po(\mu)] = \frac{\mu^0}{0!} e^{-\mu} = e^{-\mu}$$

$$p_2 = P(A_2) = P(Y=1) = \frac{\mu^1}{1!} e^{-\mu} = \mu e^{-\mu}$$

$$p_3 = \frac{\mu^2}{2!} e^{-\mu}$$

$$p_4 = \frac{\mu^3}{3!} e^{-\mu}$$

$$p_5 = \frac{\mu^4}{4!} e^{-\mu}$$

$$p_6 = 1 - \sum_{i=1}^{5} p_i$$

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

μ ej given i H_0 $\Rightarrow$ Vi måste skatta μ ur data

Enl. § 3 gäller att om $Y \in Po(\mu)$ så är $E(Y) = \mu$

Vi skattar μ med $\bar{y}$

$$\mu^*_{obs} = \bar{y} = \frac{144 \cdot 0 + 91 \cdot 1 + 32 \cdot 2 + 11 \cdot 3 + 2 \cdot 4}{280} = 0.7$$

$$\Rightarrow \; np^*_{1obs} = 280 \cdot e^{-\mu^*_{obs}} = 280 \cdot e^{-0.7} = 139.0$$

$$np^*_{2obs} = 280 \cdot 0.7 \cdot e^{-0.7} = 97.3$$

$$np^*_{3obs} = 280 \cdot \frac{0.7^2}{2} e^{-0.7} = 34.1$$

$$np^*_{4obs} = 280 \cdot \frac{0.7^3}{3!} e^{-0.7} = 7.95$$

$$np^*_{5obs} = 280 \cdot \frac{0.7^4}{4!} e^{-0.7} = \text{något} < 5$$

$$np^*_{6obs} = 280 \left(1 - \sum_{i=1}^{r} p^*_{iobs}\right) = \text{något} < 5$$

Nu är inte alla $np^*_{iobs} \geq 5$. Vi slår då ihop resultaten 5 och 6 vilket ger

$$np^*_{5obs} + n^*_{6obs} = 1.65 < 5$$

Då slår vi ihop resultat 4, 5 och 6 $\Rightarrow$

$$np^*_{4obs} + np^*_{5obs} + np^*_{6obs} = 9.6 > 5$$

Nu har vi 4 resultat, d.v.s.

$r = 4$

A_1 – 0 döda, A_2 – 1 död, A_3 – 2 döda, A_4 – ≥ 3 döda

$x_1 = 144$, $x_2 = 91$, $x_3 = 32$, $x_4 = 11 + 2 + 0 = 13$

$np^*_{1obs} = 139.0$, $np^*_{2obs} = 97.3$, $np^*_{3obs} = 34.1$, $np^*_{4obs} = 9.6$

$$\Rightarrow \; Q' = \sum_{i=1}^{4} \frac{(x_i - np^*_{iobs})^2}{np^*_{iobs}} = 1.95$$

Vi ska jämföra med $\chi^2_\alpha(r-k-1)$

$\alpha = 0.05$

$r = 4$

Vi har skattat $k = 1$ parametrar ur data (μ) för att skatta $p_1, p_2, \ldots$

$$\Rightarrow \; \chi^2_\alpha(r-k-1) = \chi^2_{0.05}(4-1-1) = [\text{tab. } 4] = 5.99$$

$Q' < \chi^2_{0.05}(2) \Rightarrow$ Vi kan ej förkasta H_0 på risknivån 5 %

HOMOGENITETSTEST

Används när nollhypotesen är att vi har samma sannolikhetsfunktion i alla grupperna. ⇒ § 14.3

$$Q = \sum_{i=1}^{s}\sum_{j=1}^{r} \frac{\left(x_{ij} - \frac{n_i m_j}{N}\right)^2}{\frac{n_i m_j}{N}}$$

Villkor för homogenitetstest: alla $\frac{n_i m_j}{N} \geq 5$

Om $Q > \chi^2_{\alpha}[(r-1)(s-1)]^2$ så förkastas H_0

Tenta 2018-08: 5

Montering	Fungerar	F1	F2
A-läge	27	15	8
B-läge	22	7	21

Undersök om monteringsläget har någon betydelse.

$\alpha = 0.05$

Vårt H_0 är att sannolikhetsfunktionen är samma för A-läge som för B-läge

⇒ Homogenitetstest ⇒ §14.3 ⇒ $Q = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{\left(x_{ij} - \frac{n_i m_j}{N}\right)^2}{\frac{n_i m_j}{N}}$

$x_{11} = 27,\quad x_{12} = 15,\quad x_{13} = 8$
$x_{21} = 22,\quad x_{22} = 7,\quad x_{23} = 21$

$n_1 = x_{11} + x_{12} + x_{13} = 50$
$n_2 = x_{21} + x_{22} + x_{23} = 50$

$m_1 = x_{11} + x_{21} = 49$
$m_2 = x_{12} + x_{22} = 22$
$m_3 = x_{13} + x_{23} = 29$

$N = \sum n_i = \sum m_j = 100$

Är alla $\frac{n_i m_j}{N} \geq 5$? : $\frac{n_1 m_2}{N} = \frac{50 \cdot 22}{100} > 5$ ⇒ Homogenitetstest ok

Montering	Fungerar	F1	F2	
A-läge	27	15	8	50
B-läge	22	7	21	50
	49	22	29	100

$$Q = \frac{\left(27 - \frac{50 \cdot 49}{100}\right)^2}{\frac{50 \cdot 49}{100}} + \frac{\left(15 - \frac{50 \cdot 22}{100}\right)^2}{\frac{50 \cdot 22}{100}} + \frac{\left(21 - \frac{50 \cdot 29}{100}\right)^2}{\frac{50 \cdot 29}{100}} = 9.2469$$

$$\chi^2_{\alpha}\left[(r-1)(s-1)\right] = \chi^2_{0.05}\left[(3-1)(2-1)\right] = \chi^2_{0.05}(2) = [\text{tab. } 4] = 5.99$$

Homogenitetstest på räknaren: "χ^2-Test"

$Q > \chi^2_{0.05}(2) \Rightarrow$ Förkasta H_0

D.v.s. det har betydelse om vi använder A-läge eller B-läge

OBEROENDETEST

Om vi vill undersöka om två egenskaper A och B är oberoende, där A kan ha utfallen $A_1, A_2, \ldots, A_r$ och B kan ha utfallen $B_1, B_2, \ldots, B_s$ och H_0 är att A och B är oberoende gör man oberoendetest.

Det går numeriskt till på samma sätt som homogenitetstest.

samma Q

Tenta 2017-01: 5

	Sträng vinter Nuuk	Normal vinter Nuuk	Mild vinter Nuuk	
Sträng vinter Köpenhamn	2	19	9	50
Normal vinter Köpenhamn	13	51	22	86
Mild vinter Köpenhamn	12	18	2	32
	27	88	33	148

H_0: danska och grönländska vintertemperaturer är oberoende.

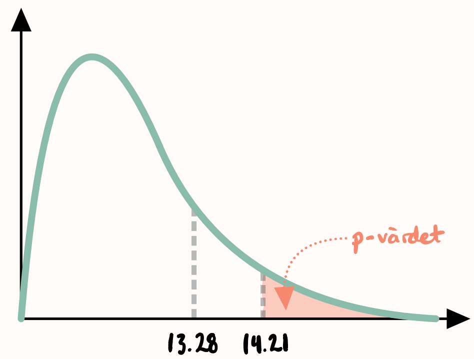
Här har vi ett oberoendetest som utförs rent praktiskt som homogenitetstest.

Kolla att alla $\frac{n_i m_j}{N} \geq 5$: $\frac{30 \cdot 27}{148} = 5.47 > 5 \Rightarrow$ ober.test är ok

$$Q = \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{\left(x_{ij} - \frac{n_i m_j}{N}\right)^2}{\frac{n_i m_j}{N}} = \ldots = 14.21$$

$$\chi^2_{0.01}\left[(3-1)(3-1)\right] = \chi^2_{0.01}(4) = \left[\text{tab. } 4\right] = 13.28$$

$Q > \chi^2_{0.01}(4)$ $\Rightarrow$ Vi förkastar att vintertemperaturerna på Grönland och i Danmark är oberoende på 1%-nivån.



p-värdet = [räknare: 2nd distribution → $\chi^2 cdf$] = $P(14.21,\ 10^{99},\ 4) = 0.006654$