Statistik för bioteknik
Föreläsning 12.
χ^2-test och liknande

Jan Grandell & Timo Koski

04.12.2017

KTH Matematik

Innehåll

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

KTH Matematik

χ^2-TEST

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

Den enklaste situationen:

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

KTH VETENSKAP OCH KONST
KTH Matematik

χ^2-TEST

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

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

mot

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

KTH VETENSKAP OCH KONST
KTH Matematik

För att göra detta bildar vi teststatistikan

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

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

χ^2-FÖRDELNING: HUR UPPSTÅR DEN?

$X_1, X_2, \ldots, X_n$, $X_i \sim \mathcal{N}(0,1)$, oberoende.

$$X_1^2 + X_2^2 + \ldots + X_n^2 \sim \chi^2(n).$$

Observera att $X_1^2 + \ldots + X_n^2 > 0$.

KTH VETENSKAP OCH KONST

KTH Matematik

Skewness to the right

$$X \sim \chi^2(k)$$

- $E(X) = k$
- $Median(X) \approx k\left(1 - \frac{2}{9k}\right)^3$
- $Mode = \max(k-2, 0)$.

I.e., $Mode < Median < Mean$

KTH Matematik

För att göra resultatet (d.vs. Q är approximativt $\chi^2(r-1)$-fördelad) troligt, betraktar vi $r = 2$. Då gäller, med $X = X_1$ och $p = p_1$ att

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

$$= \frac{(X - np)^2}{np} + \frac{(X - np))^2}{n(1-p)} = \frac{(X - np)^2}{np(1-p)}\,.$$

Eftersom X är $\mathrm{Bin}(n, p)$ 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. $\chi^2(1)$.

χ^2-TEST

Vi gör nu följande test:

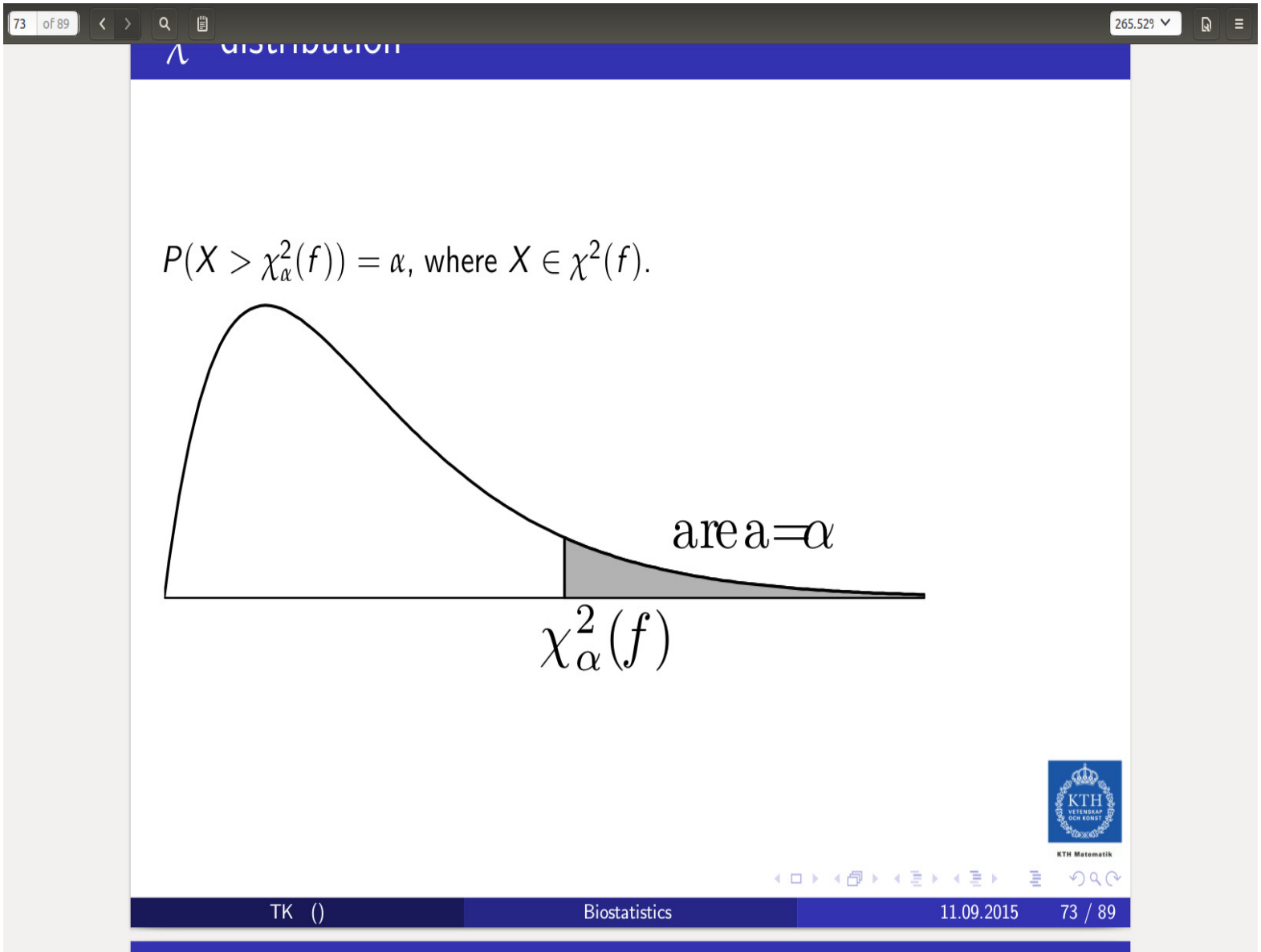
Förkasta H_0 om

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

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

χ^2-FÖRDELNING

$P(X > \chi^2_\alpha(f)) = \alpha$, där $X \in \chi^2(f)$.



KTH Matematik

χ^2-TEST

The critical region and decision are one-tailed:

- Reject H_o if
$$Q > \chi^2_\alpha(r-1).$$
- Otherwise you fail to reject H_o

If n is large, the approximative level of significance is α. Practical requirement is that all $np_j \geq 5$.

KTH VETENSKAP OCH KONST
KTH Matematik

χ^2-TEST

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.

KTH VETENSKAP OCH KONST
KTH Matematik

χ^2-test: Kast med en tärning

Vid 96 kast med en tärning erhölls 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

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

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 $p_1, p_2, \ldots, p_r$ bero av en okänd parameter $\theta = (\theta_1, \ldots, \theta_s)$, och testa hypotesen

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

och för något värde på θ.

χ^2-TEST MED SKATTNING AV PARAMETRAR

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*.

KTH Matematik

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

Grundregeln är att antalet frihetsgrader fås av

antalet fria kvadratsummor − antalet skattade parametrar.

KTH Matematik

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

KTH Matematik

Die Gesamtstärke eines Armeekorps (=armekår) im deutschen Heer betrug 1554 Offiziere, 43.317 Mann, 16.934 Pferde und 2933 Fahrzeuge.

Totalt omkom $1 \cdot 91 + 2 \cdot 32 + 3 \cdot 11 + 4 \cdot 2 = 196$ *soldater under tjugo år i 14 armekårer. Således var det en mycket sällsynt händelse att bli ihjälsparkad av en häst.*

De små talens lag En händelse av låg frekvens i en stor population följer en Poissonfördelning.

KTH VETENSKAP OCH KONST
KTH Matematik

De små talens lag

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

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

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

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

och

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

KTH Matematik

De små talens lag

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

$$H_0: \quad \text{"data kommer från Poisson-fördelning"}$$

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 på ståendet att antalet ihjälsparkade soldater per år och armékår är Poisson-fördelat.

Om vi på *förhand* hade velat testa om data kom från just $\text{Po}(0.7)$-fördelningen hade vi jämfört med $\chi^2_{0.05}(3) = 7.81$.

De små talens lag

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.

KTH Matematik

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`

KTH Matematik

Homogenitetstest

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

Homogenitetstest

Antag att s serier utförts med följande resultat:

Serie	Absoluta frekvenser för A_1	A_2	$\dots$	A_r	Antal försök
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.

KTH Matematik

Homogenitetstest

Vi anser att serierna är *homogena* om hypotesen

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

KTH Matematik

Homogenitetstest

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.

KTH VETENSKAP OCH KONST
KTH Matematik

Homogenitetstest

Frihetsgraderna fås på följande sätt:

$$\text{antalet fria kvadratsummor} - \text{antalet skattade parametrar}$$

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

Oberoendetest

Vi tar nu ett stickprov om n enheter, där varje enhet klassifiseras 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 A_1	A_2	$\dots$	A_r	Total	Antal försök
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	

KTH Matematik

Oberoendetest

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.

Oberoendetest

Frihetsgraderna fås på följande sätt:

$$\text{antalet fria kvadratsummor} - \text{antalet skattade parametrar}$$

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

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

KTH Matematik

Oberoendetest: ett exempel

A_1 = case, en individ med en viss sjukdom, A_2 = control, en individ utan denna sjukdom. B_1 = har aldrig använt en viss terapi. B_2 = har tidigare använt en viss terapi, B_3 = använder för tillfället terapin med metod 1, B_3 = använder för tillfället terapin med metod 2.

Serie Egenskap	A_1	A_2	Total
B_1	71	208	$x_{1\cdot} = 279$
B_2	22	53	$x_{2\cdot} = 75$
B_3	32	27	$x_{3\cdot} = 59$
B_4	30	93	$x_{4\cdot} = 123$
Total	$x_{\cdot 1} = 155$	$x_{\cdot 2} = 381$	536

Oberoendetest: ett exempel

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

Vi får:

$$(p_{1\cdot}^*)_{\text{obs}} = \frac{x_{1\cdot}}{n} = \frac{279}{536}$$

$$(p_{2\cdot}^*)_{\text{obs}} = \frac{75}{536}, (p_{3\cdot}^*)_{\text{obs}} = \frac{59}{536}, (p_{4\cdot}^*)_{\text{obs}} = \frac{123}{536}$$

$$p_{\cdot 1}^* = (p_{\cdot 1}^*)_{\text{obs}} = \frac{x_{\cdot 1}}{n} = \frac{155}{536}, (p_{\cdot 2}^*)_{\text{obs}} = \frac{x_{\cdot 2}}{n} = \frac{381}{536},$$

KTH Matematik

Oberoendetest: ett exempel

De förväntade frekvenserna:

$$np_{i\cdot}^* \, p_{\cdot j}^*$$

blir

$$np_{1\cdot}^* \, p_{\cdot 1}^* = 536 \cdot \frac{279}{536} \cdot \frac{155}{536} = 80.68$$

$$np_{1\cdot}^* \, p_{\cdot 2}^* = 536 \cdot \frac{279}{536} \cdot \frac{381}{536} = 198.32$$

$$np_{2\cdot}^* \, p_{\cdot 1}^* = 536 \cdot \frac{75}{536} \cdot \frac{155}{536} = 21.69$$

$$np_{2\cdot}^* \, p_{\cdot 2}^* = 536 \cdot \frac{75}{536} \cdot \frac{381}{536} = 53.31$$

KTH VETENSKAP OCH KONST
KTH Matematik

Oberoendetest: ett exempel

De förväntade frekvenserna (forts.)

$$np_{i\cdot}^* \, p_{\cdot j}^*$$

blir

$$np_{3\cdot}^* \, p_{\cdot 1}^* = 536 \cdot \frac{59}{536} \cdot \frac{155}{536} = 17.06$$

$$np_{3\cdot}^* \, p_{\cdot 2}^* = 536 \cdot \frac{59}{536} \cdot \frac{381}{536} = 41.94$$

$$np_{4\cdot}^* \, p_{\cdot 1}^* = 536 \cdot \frac{123}{536} \cdot \frac{155}{536} = 35.57$$

$$np_{4\cdot}^* \, p_{\cdot 2}^* = 536 \cdot \frac{123}{536} \cdot \frac{381}{536} = 87.43$$

KTH VETENSKAP OCH KONST
KTH Matematik

Oberoendetest: ett exempel

$$Q = \frac{(71-80.68)^2}{80.68} + \frac{(208-198.32)^2}{198.32} + \frac{(22-21.69)^2}{21.69} + \frac{(53-53.32)^2}{53.32}$$
$$+ \frac{(32-17.06)^2}{17.06} + \frac{(27-41.94)^2}{41.94}$$
$$+ \frac{(30-35.57)^2}{35.37} + \frac{(93-87.43)^2}{87.43}$$
$$= 21.268$$

Antalet frihetsgrader är $(r-1)(s-1) = (2-1)(4-1) = 3$

KTH Matematik

Oberoendetest: P-value

P-värdet

```
>> 1-chi2cdf(21.268,3)
ans =
9.2610e-05
```

Kritiska värdet, signifikansnivå $= 0.05$

```
>> chi2inv(0.95,3)
ans =
7.8147
```

Kursens tabellsamling ger $\chi_{0.05}(3) = 7.81$. Det gäller alltså $7.81 < Q = 21.268$, d.v.s. teststatistikan ligger i det kritiska området. Alltså förkastas nollhypotesen om oberoendet på signifikansnivån 5%. Det finns en kontingens mellan A_j:na och B_i:na.

KTH Matematik