2. Oberoende-test

- Oberoendet av två kriterier för klassifikation undersökes
- Exempel: Vi vill veta om rökandet är beroende av kön
- Vi tar ett stickprov ur befolkningen (n=250) och klassificera personerna enligt dessa två kriterier:

	Kv.	M.
R	20	30
R*	80	120

Varje person hamnar i exakt en cell.

uwe.menzel@genpat.uu.se

Kontingenstabell (Korstabell)

	Kv.	M.	Σ
R	20	30	50
R*	80	120	200
Σ	100	150	250

Vi har 100 kvinnor, 20 av de rökar.
Vi har 150 män. 30 av de rökar.
Totalt rökar 50 av 250.

Om rökandet är oberoende av kön: vi förväntar oss att det finns samma andel rökare bland kvinnor som bland män!

Här är det så:
Det finns totalt 20% rökare (50 av 250, sista kolumn).
Det är också 20% av kvinnorna och 20% av män som rökar!

Vad vi förväntar oss om båda kriterier är oberoende:

	Kv.	M.	Σ
R	20	30	50
R*	80	120	200
Σ	100	150	250

	Kv.	M.	Σ
R	a	b	R_1
R*	c	d	R_2
Σ	K_1	K_2	n

Oberoendet kräver:

$$\frac{20}{100}=\frac{50}{250} \quad kvinnor$$

$$\frac{30}{150}=\frac{50}{250} \quad män$$

Allmänt:

$$\frac{a}{K_1}=\frac{R_1}{n} \Rightarrow a=\frac{R_1\cdot K_1}{n}$$

$$\frac{b}{K_2}=\frac{R_1}{n} \Rightarrow b=\frac{R_1\cdot K_2}{n}$$

Vad vi förväntar oss om båda kriterier är oberoende:

$$a=\frac{R_1\cdot K_1}{n}$$

$$b=\frac{R_1\cdot K_2}{n}$$

$$c=\frac{R_2\cdot K_1}{n}$$

$$d=\frac{R_2\cdot K_2}{n}$$

	Kv.	M.	Σ
R	a	b	R_1
R*	c	d	R_2
Σ	K_1	K_2	n

detsamma gäller också för andelen icke-rökare

Förväntade värden i varje cell om båda kriterier är oberoende:

	Kv.	M.	Σ
R	$\frac{R_1\cdot K_1}{n}$	$\frac{R_1\cdot K_2}{n}$	R_1
R*	$\frac{R_2\cdot K_1}{n}$	$\frac{R_2\cdot K_2}{n}$	R_2
Σ	K_1	K_2	n

Stickprov beror av slumpen

Ett mer realistiskt stickprov.
Ett sådant stickprov kan uppstå även vid oberoendet:

	Kv.	M.	Σ
R	24	26	50
R*	76	124	200
Σ	100	150	250

.. stickprovet avviker ju inte mycket från de förväntade värden:

	Kv.	M.	Σ
R	20	30	50
R*	80	120	200
Σ	100	150	250

Ett annat stickprov

Ett stickprov som avviker mycket från de förväntade värden. Vi misstänker att båda kriterier är inte oberoende.

	Kv.	M.	Σ
R	4	46	50
R*	96	104	200
Σ	100	150	250

... ty stickprovet avviker ju mycket från de förväntade värden:

	Kv.	M.	Σ
R	20	30	50
R*	80	120	200
Σ	100	150	250

Test för oberoendet

- Nollhypotes: oberoendet
- Nollhypotesen förkastas om stickprovet avviker för mycket från de förväntade värden, om det är osannolikt att ett sådant stickprov kommer till stånd under H_0
- Igen: Var går gränsen? Vilket värde får skillnaden maximalt ha för att ”bibehålla” nollhypotesen? (Hur mäter jag skillanden överhuvudtaget?) → vi behöver en fördelning för ”skillnaden”, dvs. sannolikheten att en viss skilland uppstår under H_0! Om denna skillnad är ”osannolik”, förkastar vi H_0.

En summa som ”mäter” skillnaden

observed values

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

expected values

Fördelningen av denna testvariabel kan man räkna ut, givet att nollhypotesen gäller (och att n är stor) → Chi-Square fördelningen

Lite förvirrande: testvariabeln kallas av någon anledning χ^2 , och den är också χ^2-fördelad.

Antalet frihetsgrader och kritiskt område för oberoende-testet

r=antalet råder c=antalet kolumner

$$df = (r-1)\cdot(c-1)$$

$$\Omega_{krit} = \{\chi^2 > \chi^2_\alpha\}$$

χ^2 är positiv och vi förkastar nollhypotesen om den är stor → upper tail test.

Observed

	Kv.	M.	Σ
R	24	26	50
R*	76	124	200
Σ	100	150	250

Expected (H_0)

	Kv.	M.	Σ
R	20	30	50
R*	80	120	200
Σ	100	150	250

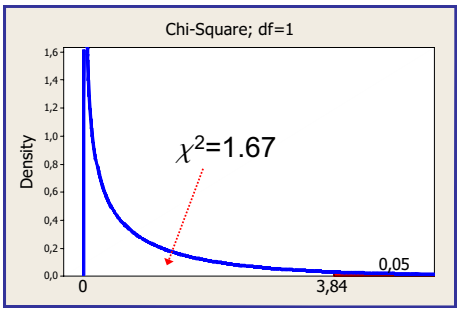
$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad \textit{summa över alla celler}$$

$$= \frac{(24-20)^2}{20} + \frac{(26-30)^2}{30} + \frac{(76-80)^2}{80} + \frac{(124-120)^2}{120}$$

$$= 1.67$$

$$\Omega_{krit} = \{\chi^2 > \chi^2_\alpha\} = \{\chi^2 > \chi^2_{0.05}(1)\} = \{\chi^2 > 3.84\}$$

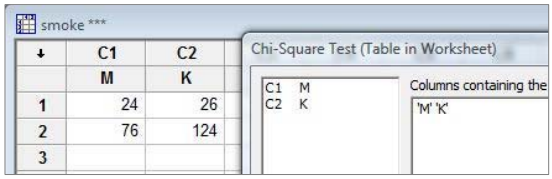
Resultat



H_0 förkastas inte. Rökandet kan vara oberoende av kön.

Minitab

Stat / Tables / Chi-Square Test



Minitab results

Chi-Square Test: M; K

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

	M	K	Total
1	24	26	50
	20,00	30,00	
	0,800	0,533	
2	76	124	200
	80,00	120,00	
	0,200	0,133	
Total	100	150	250

observed
expected

Chi-Sq = 1,667; DF = 1; P-Value = 0,197

Sammanfattning oberoende-test

	A	A*	Σ
B	O_1	O_2	R_1
B*	O_3	O_4	R_2
Σ	K_1	K_2	n

$$E_1 = \frac{R_1 \cdot K_1}{n} \quad E_2 = \frac{R_1 \cdot K_2}{n} \quad E_3 = \frac{R_2 \cdot K_1}{n} \quad E_4 = \frac{R_2 \cdot K_2}{n}$$

	A	A*	Σ
B	E_1	E_2	R_1
B*	E_3	E_4	R_2
Σ	K_1	K_2	n

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad \text{summa över alla celler}$$

$$\Omega_{krit} = \{\chi^2 > \chi^2_\alpha(df)\} \qquad df = (r-1)\cdot(c-1)$$

Inte limiterad till 2 x 2 tabeller!

Förutsättningar

- bara om n är stor har summan en χ^2-fördelning → $E_i \geq 5$ i varje cell
- slumpmässigt stickprov (som vanligt)

	A	A*	Σ
B	E_1	E_2	R_1
B*	E_3	E_4	R_2
Σ	K_1	K_2	n

alla $E_i \geq 5$

Föredrar två muränarter olika miljöer?

Experiment: det räknades hur ofta man såg två muränarter i ett visst område i Belize. Det noterades vilka arter sågs i vilken miljö.

Young, R.F., and H.E. Winn. 2003. Activity patterns, diet, and shelter site use for two species of moray eels, *Gymnothorax moringa* and *Gymnothorax vicinus*, in Belize. Copeia 2003: 44-55.

Föredrar två muränarter olika miljöer? Observationer

	G. moringa	G. vicinus	Σ
gräs	127	116	243
sand	99	67	166
gräns	264	161	425
Σ	490	344	834

Observed and expected

243 · 344 / 834

O	G. moringa	G. vicinus	Σ
gräs	127	116	243
sand	99	67	166
gräns	264	161	425
Σ	490	344	834

E	G. moringa	G. vicinus	Σ
gräs	142.8	100.2	243
sand	97.5	68.5	166
gräns	249.7	175.3	425
Σ	490	344	834

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{(127-142.8)^2}{142.8} + \frac{(116-100.2)^2}{100.2} + \frac{(99-97.5)^2}{97.5}$$
$$+ \frac{(67-68.5)^2}{68.5} + \frac{(264-249.7)^2}{249.7} + \frac{(161-175.3)^2}{175.3}$$
$$= 1.75 + 2.491 + 0.023 + 0.033 + 0.819 + 1.167 = 6.26$$

$$df = (r-1)\cdot(c-1) = 2\cdot 1 = 2$$
$$\Omega_{krit} = \{\chi^2 > \chi^2_\alpha(df)\} = \{\chi^2 > \chi^2_{0.05}(2)\} = \{\chi^2 > 5.99\}$$

Resultat

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 6.26$$

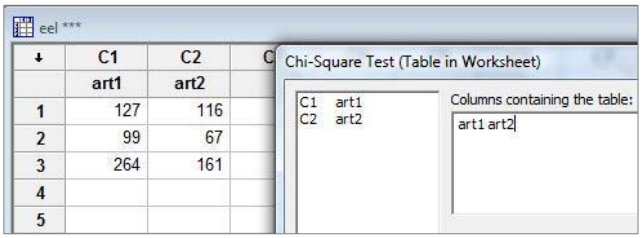
$$\Omega_{krit} = \{\chi^2 > \chi^2_\alpha(df)\} = \{\chi^2 > \chi^2_{0.05}(2)\} = \{\chi^2 > 5.99\}$$

H_0 förkastas. Olika arter föredrar olika mijöer.

Antalet frihetsgrader var 2: Om man fyller i 2 (godtyckliga) celler i 2x3-tabellen, följer alla andra automatiskt (om marginalfördelningar är givna).
Testa gärna det!! Det är lättare än sudoku ...

Minitab

Stat / Tables / Chi-Square Test



Chisquare.MPJ

Results for: eel

Chi-Square Test: art1; art2

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

	art1	art2	Total
1	127	116	243
	142,77	100,23	
	1,742	2,481	
2	99	67	166
	97,53	68,47	
	0,022	0,032	
3	264	161	425
	249,70	175,30	
	0,819	1,166	
Total	490	344	834

observed
expected

Chi-Sq = 6,262; DF = 2; P-Value = 0,044

samma resultat som förut ...
H_0 förkastas

Om E_i < 5 (small sample size)

- ***Fisher's Exact test*** (för korstabeller)
- beror inte på normalapproximation
- H_0 som förut: ingen assoziation mellan båda egenskaper
- Cell-värdena (under H_0) i korstabellen följer en hypergeometrisk fördelning

Exempel: Jag antar att rökandet är oberoende av kön (H_0). Jag tar ett stickprov av 25 personer, därav 5 rökare och 10 kvinnor (det måste vara känt). Hur stor är sannolikheten att ha 2 kvinnliga rökare i detta stickprov?

	Kv.	M.	Σ
R	2		5
R*			20
Σ	10	15	25

Fisher's Exact test

- Fisher's Exact test räknar ut hur sannolikt en viss konfiguration är (marginalsummorna givna):

	Kv.	M.	
R	a	b	a+b
R*	c	d	c+d
	a+c	b+d	n

$$p = \frac{\binom{a+b}{a}\cdot\binom{c+d}{c}}{\binom{n}{a+c}}$$

en viss konfiguration av a,b,c,d

... och dess sannolikhet under H_0

Fisher's Exact test: p-värdet

- p-värdet: sannolikhet för den observerade tabellen plus slh:a för alla tänkbara ännu mera extrema (mera ojämn-fördelade, mindre sannolika) tabeller
  - mera extrema = mindre sannolikheter
  - "tänkbar" heter att r- och k-summorna måste bibehållas!
- om p-värdet är litet → förkasta H_0, ty det är osannolikt att få ett sådant tabell under H_0

Att räkna Fisher's Exact test

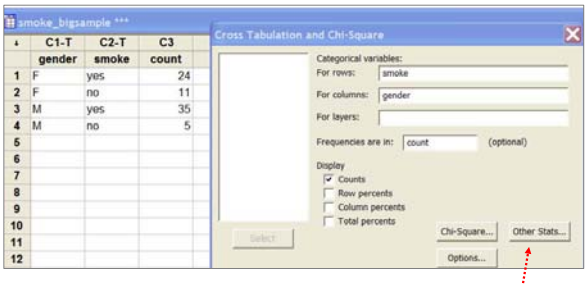
- räkne-intensivt, antalet "tänkbara" tabeller växer snabbt med tabellens storlek (dvs. med antalet rader och kolumner)
- Minitab: bara 2x2 tabeller
- web-site för större tabeller (max. 6x6):

http://www.physics.csbsju.edu/stats/exact_NROW_NCOLUMN_form.html

Minitab- stort stickprov

Fisher är "gömd" här

Stat / Tables / Cross Tabulation and Chi-Square



välj Fisher's Exact Test här, man kan dock inte bestämma vilken alternativ hypotes man vill ha (one-tailed / two-tailed)

Results Large Sample

Tabulated statistics: smoke; gender

Using frequencies in count

Rows: smoke Columns: gender

	F	M	All
no	11	5	16
yes	24	35	59
All	35	40	75

Cell Contents: Count

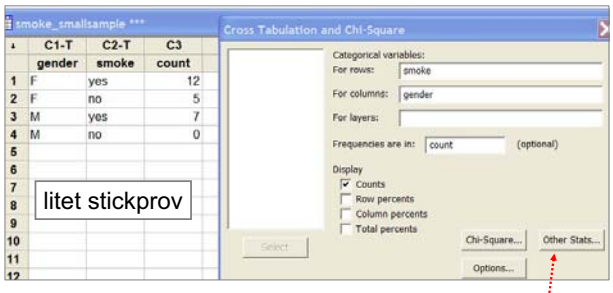
Pearson Chi-Square = 3,985; DF = 1; P-Value = 0,046
Likelihood Ratio Chi-Square = 4,035; DF = 1; P-Value = 0,045

Fisher's exact test: P-Value = 0,054

Minitab- litet stickprov

Fisher är "gömd" här

Stat / Tables / Cross Tabulation and Chi-Square



välj Fisher's Exact Test här

Results Small Sample

Results for: smoke_smallsample

Rows: smoke Columns: gender

	F	M	All
no	5	0	5
yes	12	7	19
All	17	7	24

Cell Contents: Count

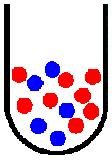
Pearson Chi-Square = 2,601; DF = 1; P-Value = 0,107
Likelihood Ratio Chi-Square = 3,966; DF = 1; P-Value = 0,046

**** NOTE * 2 cells with expected counts less than 5***

Fisher's exact test: P-Value = 0,272

Appendix

Hypergeometriska fördelningen



$N = 13$ *antalet kulor i urnan*

$b = 5$ *antalet blå kulor i urnan*

$n = 3$ *stickprovets storlek*

Man drar n kulor. Hur stor är slh. att få k blåa?

$$X \in Hyp(N, b, n)$$

$$P_X(k) = \frac{\binom{b}{k}\cdot\binom{N-b}{n-k}}{\binom{N}{n}} = \frac{\binom{5}{k}\cdot\binom{8}{3-k}}{\binom{13}{3}}$$

Hypergeometriska fördelningen

$$X \in Hyp(N, b, n) \quad \rightarrow \quad X \in Hyp(13, 5, 3)$$

$$P_X(k) = \frac{\binom{b}{k}\cdot\binom{N-b}{n-k}}{\binom{N}{n}} = \frac{\binom{5}{k}\cdot\binom{8}{3-k}}{\binom{13}{3}}$$

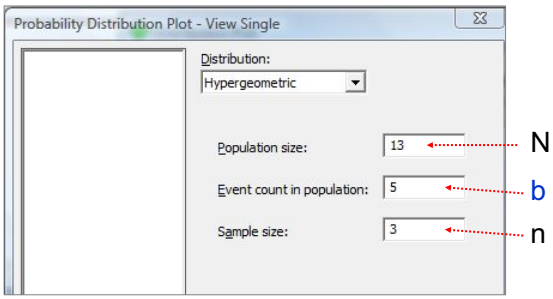
$$P_X(0) = \frac{\binom{5}{0}\cdot\binom{8}{3}}{\binom{13}{3}} = \frac{28}{13\cdot 11} = 0.1958$$

$$P_X(1) = \frac{\binom{5}{1}\cdot\binom{8}{2}}{\binom{13}{3}} = \frac{70}{143} = 0.4895$$

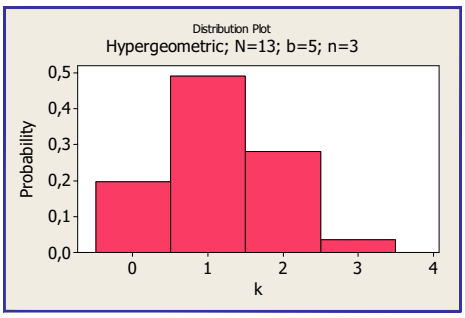
$$P_X(2) = \frac{\binom{5}{2}\cdot\binom{8}{1}}{\binom{13}{3}} = 0.2797$$

$$P_X(3) = \frac{\binom{5}{3}\cdot\binom{8}{0}}{\binom{13}{3}} = 0.0396$$

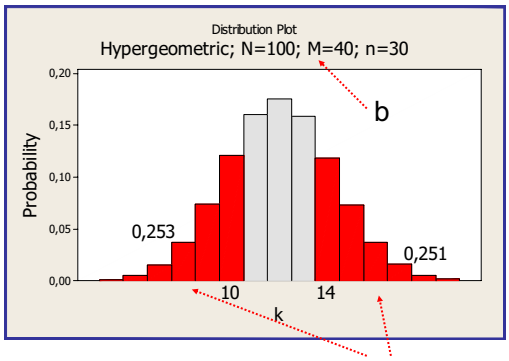
Minitab



Minitab



Graph / View Probability



Shaded area, x-value=10: Minitab gör det så att sannolikheterna på båda sidor är ungefär lika stora