

Grundläggande matematisk statistik

Anpassningstest och Oberoendetest

Uwe Menzel, 2018
uwe.menzel@matstat.org
www.matstat.org

Goodness-of-fit Test (Anpassningstest)

- testar om en statistisk modell kan stämma, dvs. om modellen överensstämmer med observerade data (stickprov) eller inte
 - följer observationerna en förmodad fördelning?

Modell: rättvis tärning, dvs. $p_i = 1/6$ $i = 1 \dots 6$

Stickprov: $n = 120$ kast



ögontalet	1	2	3	4	5	6	
E_i : förväntade frekvenser	20	20	20	20	20	20	$E_i = p_i \cdot n$
O_i : observerade frekvenser	21	16	28	24	19	12	$\sum O_i = 120$

en restriktion

Med denna observation i handen:

- kan vi fortfarande tro att modellen (rättvis tärning) stämmer?
- ... eller måste denna (noll-)hypotes förkastas?

www.matstat.org

Goodness-of-fit Test

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

O_i : observerade värden (stickprov)
 E_i : förväntade värden (modell)
 k : antalet grupper ($k = 6$ för tärning)

- Om nollhypotesen stämmer och observationerna är ungefär lika med de förväntade modellvärden borde χ^2 vara liten, eftersom $O_i \approx E_i$
- Om observationerna däremot avviker starkt från modellvärdena blir χ^2 stor och vi är benägna att förkasta nollhypotesen.

Fördelning för testvariabeln under H_0 :

- Om H_0 stämmer och om de förväntade värdena inte är för små i någon kategori ($E_i \geq 5$) är testvariabeln χ^2 -fördelad med f frihetsgrader.
- f beror på antalet restriktioner och antalet parametrar som måste skattas från stickprovet för att upprätta modellen (se nere)

under H_0 :

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

både testvariabeln och fördelningen kallas χ^2 !

www.matstat.org

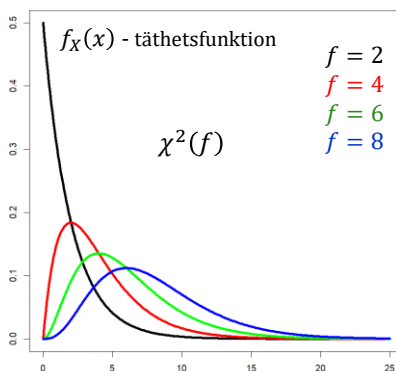
Goodness-of-fit Test

Frihetsgrader

Täthetsfunktionen $\chi^2(f)$ beror på antalet frihetsgrader (f).

Antalet frihetsgrader för anpassningstestet:

$$f = k - r - p$$



k = antalet kategorier ($k = 6$ för tärning)

r = antalet restriktioner ($r = 1$ för tärning: $\sum O_i = 120$)

p = antalet parametrar som måste skattas med hjälp av stickprovet för att upprätta modellen ($p = 0$ för tärning)

ofta är $r = 1$ och $p = 0$



$$f = k - 1$$

t. ex. tärning

www.matstat.org

Goodness-of-fit Test

(Anpassningstest)

Testvariabel χ^2 för två stickprov (tärning):

	1	2	3	4	5	6
E_i	20	20	20	20	20	20
O_i	21	16	28	24	19	12



$$\begin{aligned}\chi^2 &= \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{1^2}{20} + \frac{4^2}{20} + \frac{8^2}{20} + \frac{4^2}{20} + \frac{1^2}{20} + \frac{8^2}{20} \\ &= \underline{\underline{8.1}}\end{aligned}$$

Skillnad mellan modell och observation liten $\rightarrow \chi^2$ liten.

	1	2	3	4	5	6
E_i	20	20	20	20	20	20
O_i	0	3	2	1	1	113



$$\begin{aligned}\chi^2 &= \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{20^2}{20} + \frac{17^2}{20} + \frac{18^2}{20} + \frac{19^2}{20} + \frac{19^2}{20} + \frac{93^2}{20} \\ &= \underline{\underline{519.2}}\end{aligned}$$

Skillnad mellan modell och observation stor $\rightarrow \chi^2$ stor.

www.matstat.org

Goodness-of-fit Test

Testvariabeln χ^2 är $\chi^2(f)$ -fördelad om H_0 stämmer. Under H_0 borde alltså en observation av χ^2 inte ligga för långt ifrån centrumet av täthetsfunktionen för $\chi^2(f)$. Om observationen däremot ligger för långt på höger svansen av denna täthetsfunktion förkastar vi nollhypotesen, eftersom observationen är osannolik under H_0 . Det **kritiska området** är alltså:

$$\Omega_{krit} = \{ \chi^2 > \chi_{\alpha}^2(f) \} \quad \chi_{\alpha}^2(f): \text{kvantil för } \chi^2\text{-fördelning med } f \text{ frihetsgrader, för signifikansnivå } \alpha$$

- H_0 : Modell och observation stämmer överens
- H_a : Modell och observation stämmer **inte** överens

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

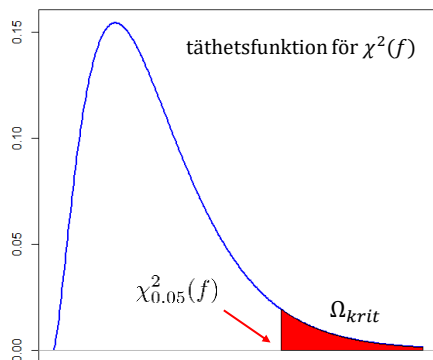
- χ^2 är alltid positiv
- H_0 förkastas om χ^2 blir för stort.

www.matstat.org

Goodness-of-fit Test

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

$\chi^2_{\alpha}(f)$: kvantil för χ^2 -fördelning med f frihetsgrader, för signifikansnivå α



Testvariabel:

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

O_i : observerade värden (stickprov)

E_i : förväntade värden (modell)

$f = k - 1$

www.matstat.org

Goodness-of-fit Test

Anpassningstest

Exempel 1:

- Är risken för hjärtattack jämnfördelad över veckodagarna – eller är en måndag (eller någon annan dag) kanske farligare?
- Med andra ord: är avvikelserna för måndags-värdet vanligt brus eller är den statistiskt signifikant?

söndag	måndag	tisdag	onsdag	torsdag	fredag	lördag
24	36	27	26	32	26	29

$n = 200$ patienter

$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = \frac{1}{7}$ modell: jämn fördelning

$E_i = n \cdot p_i = \frac{200}{7} \approx 28.57$ förväntade antal fall per dag, enligt modellen

Om hjärtattacker var jämnfördelade över veckan och man registrerar 200 fall, så borde man ha ungefär 28 eller 29 fall per veckodag ...

www.matstat.org

Goodness-of-fit Test

söndag	måndag	tisdag	onsdag	torsdag	fredag	lördag
24	36	27	26	32	26	29

$$E_i = n \cdot p_i = \frac{200}{7} \approx 28.57 \quad \text{förväntade antal fall per dag, enligt modellen}$$

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

$$= \frac{(24 - 28.57)^2}{28.57} + \frac{(36 - 28.57)^2}{28.57} + \frac{(27 - 28.57)^2}{28.57} + \frac{(26 - 28.57)^2}{28.57}$$

$$+ \frac{(32 - 28.57)^2}{28.57} + \frac{(26 - 28.57)^2}{28.57} + \frac{(29 - 28.57)^2}{28.57} = \frac{103.71}{28.57} = \underline{\underline{3.63}}$$

$$\text{Kritiskt område: } \Omega_{krit} = \{ \chi^2 > \chi_{\alpha}^2(f) \} = \{ \chi^2 > \chi_{0.05}^2(6) \} = \{ \chi^2 > 12.59 \}$$

$$\text{Frihetsgrader: } f = 7 - 1 = 6 \quad (\text{en restriktion: } n = 200)$$

tabell

www.matstat.org

Anpassningstest

$$\text{Kritiskt område: } \Omega_{krit} = \{ \chi^2 > \chi_{\alpha}^2(f) \} = \{ \chi^2 > \chi_{0.05}^2(6) \} = \{ \chi^2 > 12.59 \}$$

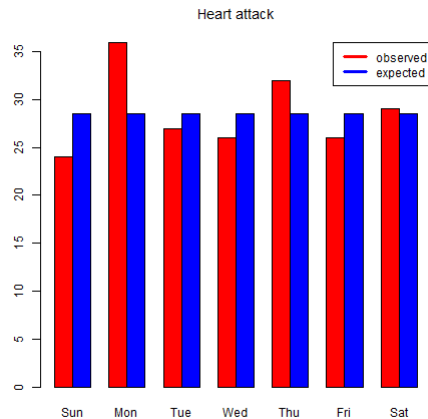
$$\chi^2\text{-fördelningens kvantiler, } \chi_{\alpha}^2(f), \text{ forts.} \quad \alpha = 0.05$$

$f \backslash \alpha$	0.25	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	1.3233	2.7055	3.8415	5.0239	6.6349	7.8794	10.828	12.116
2	2.7726	4.6052	5.9915	7.3778	9.2103	10.597	13.816	15.202
3	4.1083	6.2514	7.8147	9.3484	11.345	12.838	16.266	17.730
4	5.3853	7.7794	9.4877	11.143	13.277	14.860	18.467	19.997
5	6.6257	9.2364	11.070	12.833	15.086	16.750	20.515	22.105
6	7.8408	10.645	12.592	14.449	16.812	18.548	22.458	24.103
7	9.0371	12.017	14.067	16.013	18.475	20.278	24.322	26.018
8	10.219	13.362	15.507	17.535	20.090	21.955	26.124	27.868

Facit: $\chi^2 = 3.62 < 12.59 \rightarrow$ observationen befinner sig **inte** i den kritiska regionen. Nollhypotesen förkastas **inte**. Vi kan **inte** säga att antalet hjärtattacker är signifikant högre på någon veckodag.

www.matstat.org

Anpassningstest



Barplot:
jämförelse "observed"
och "expected"

På måndag skiljer sig det observerade värdet ganska mycket från det förväntade. Ändå visade ett χ^2 -test att måndags-avvikelsen inte är statistiskt signifikant (på signifikansnivå 0.05).

www.matstat.org

Anpassningstest



Exempel 2:

- **Hypotes/Modell:** Fåglarna föredrar inte någon sorts träd, de söker frö i alla träd med jämn sannolikhet – ju mer träd av en viss sort som förekommer, desto mer fåglar finns hos den sorten ...
- **Stickprov:** $n = 156$ fåglar observerades i en skog i Oregon

Mannan, R.W., and E.C. Meslow. 1984. Bird populations and vegetation characteristics in managed and old-growth forests, northeastern Oregon. *J. Wildl. Manage.* 48: 1219-1238.

	ädelgran	furu	kust gran	lärk	summa
kronornas volym	54%	40%	5%	1%	100%
expected	$156 \cdot 0.54 = 84.24$	$156 \cdot 0.4 = 62.4$	$156 \cdot 0.05 = 7.8$	$156 \cdot 0.01 = 1.56$	156
observed	70	79	3	4	156

1.56 för liten! → kustgran och lärk sammanfattas

www.matstat.org

Goodness-of-fit Test (Anpassningstest)



$n = 156$	ädelgran	furu	kustgran & lärk
expected	84,24	62,4	9,36
observed	70	79	7

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

$$= \frac{(70 - 84.24)^2}{84.24} + \frac{(79 - 62.4)^2}{62.4} + \frac{(7 - 9.36)^2}{9.36} = \underline{7.418} \quad \text{observation}$$

tillsammans
är $E_i \geq 5$

$$\Omega_{krit} = \{ \chi^2 > \chi^2_{\alpha}(f) \} = \{ \chi^2 > \chi^2_{0.05}(2) \} = \{ \chi^2 > 5.99 \} \quad \text{kritiskt område}$$

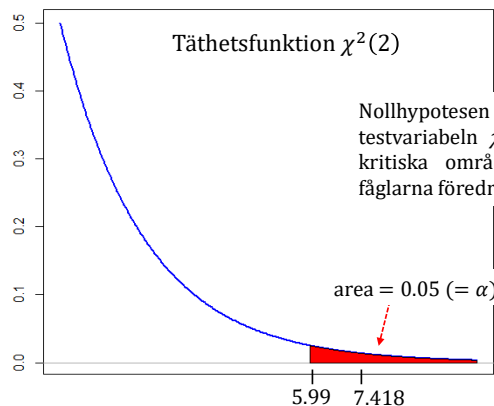
www.matstat.org

Goodness-of-fit Test (Anpassningstest)



$$\Omega_{krit} = \{ \chi^2 > \chi^2_{\alpha}(f) \} = \{ \chi^2 > \chi^2_{0.05}(2) \} = \{ \chi^2 > 5.99 \} \quad \text{kritiskt område}$$

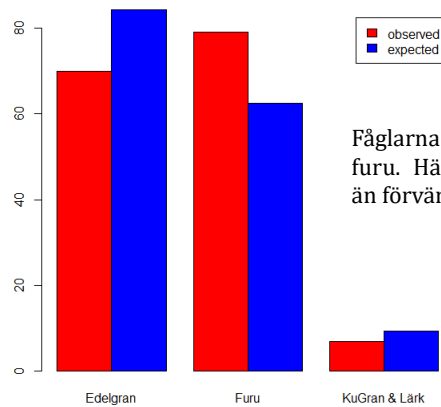
$$\chi^2 = 7.418 \quad \text{observation}$$



Nollhypotesen förkastas eftersom testvariabeln $\chi^2 = 7.42$ ligger i det kritiska området. Det betyder att fåglarna föredrar träd av en viss sort.

www.matstat.org

Goodness-of-fit Test (Anpassningstest)



Fåglarna föredrar kategori 2, alltså furu. Här observerades fler fåglar än förväntat under H_0 .

www.matstat.org

Oberoendetest

- testar om kategoriska slumpvariabler är oberoende
 - är två variabler för klassifikation av en population oberoende?

Exempel:

- två variabler för klassifikation (av en slumpmässigt vald person): kön och rökande (2 nominaler variabler)
- Beror rökandet på kön?
- Antagandet (**Modell**): rökande och kön är oberoende
- **Stickprov**: $n = 250$ slumpmässigt valda personer; klassificeras enligt de två variablerna :

	kvinna	man
R	20	30
R^*	80	120

varje person hamnar i en av tabellens celler

R = rökare

R^* = icke-rökare

www.matstat.org

Oberoendetest

Kontingenstabell (korstabell): r d- och kolumnsummor till gs!

	Kv.	M.	Σ	
R	20	30	50	----- 50 r�kare
R*	80	120	200	----- 200 icke-r�kare
Σ	100	150	250	----- $n = 250$

Nollhypotes: r kandet beror inte p  k n, b da variabler  r oberoende.

Om nollhypotesen st mmer borde andelen r kare hos kvinnor och andelen r kare hos m n vara ungef r lika stora (och d rmed ocks  ungef r lika stor som den totala andelen r kare r knat f r b da k n).

F r tabellen ovan:

-   Den totala andelen r kare  r 20% (50/250, sista kolumn).
-   Andelen r kare hos kvinnor  r 20%.
-   Andelen r kare hos m n  r 20%.

→ Variablerna "k n" och "r kandet"  r oberoende.

www.matstat.org

Oberoendetest

Under H_0 (oberoendet):

	kv.	m.	Σ
R	20	30	50
R*	80	120	200
Σ	100	150	250

allm nt:

	kv.	m.	Σ
R	a_{11}	a_{12}	R_1
R*	a_{21}	a_{22}	R_2
Σ	K_1	K_2	n

Oberoendet kr ver (f r andelen r kare):

$$\frac{20}{100} \cong \frac{50}{250} \quad \text{kvinnor: andel r kare} \cong \text{totala andelen r kare}$$

$$\frac{30}{150} \cong \frac{50}{250} \quad \text{m n: andel r kare} \cong \text{totala andelen r kare}$$



Kriterier f r oberoendet:

$$\frac{a_{11}}{K_1} \cong \frac{R_1}{n} \rightarrow a_{11} \cong \frac{R_1 \cdot K_1}{n}$$

$$\frac{a_{12}}{K_2} \cong \frac{R_1}{n} \rightarrow a_{12} \cong \frac{R_1 \cdot K_2}{n}$$

www.matstat.org

Oberoendetest

Under H_0 (oberoendet):

	kv.	m.	Σ
R	a_{11}	a_{12}	R_1
R^*	a_{21}	a_{22}	R_2
Σ	K_1	K_2	n

$$a_{11} \cong \frac{R_1 \cdot K_1}{n} \quad a_{12} \cong \frac{R_1 \cdot K_2}{n} \quad a_{21} \cong \frac{R_2 \cdot K_1}{n} \quad a_{22} \cong \frac{R_2 \cdot K_2}{n}$$

formler för andelen
icke-rökare.

analoge formler gäller också för
andelen icke-rökare.

Oberoendet:

en cell i tabellen $\cong$ rādsumma $\cdot$ kolumnsumma / n

www.matstat.org

Oberoendetest

Även om nollhypotesen stämmer kan vi inte förvänta oss att andelarna är **precis** lika stora, eftersom stickprovet styrs av slumpen:

Realistisk tabell:

	kv.	m.	Σ
R	24	26	50
R^*	76	124	200
Σ	100	150	250

Realistiskt stickprov: en sådan korstabell kan uppstå även om båda variabler är oberoende ..

Idealisk tabell :

	kv.	m.	Σ
R	20	30	50
R^*	80	120	200
Σ	100	150	250

... för att korstabellen avviker ju inte mycket från den **idealiska** (under H_0 förväntade) **tabellen**.

www.matstat.org

Oberoendetest

Först när stickprovsvärdena avviker **mycket** från värdena vi förväntar oss under H_0 är vi benägna att förkasta nollhypotesen.

Stickprov:

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

Får vi en sådan korstabell kan vi förmoda att variablerna **rökandet och kön är icke oberoende**.

Idealisk tabell :

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

... för att stickprovet avviker mycket från den under H_0 förväntade tabellen.

www.matstat.org

Oberoendetest

- **Nollhypotes:** variablerna är oberoende
- **Alternativ hypotes:** variablerna är icke oberoende

Nollhypotesen förkastas när stickprovs-korstabellen avviker för mycket från korstabellen som förväntas under H_0 (idealisk tabell).

Testvariabel: $\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ O_{ij} : observerade värden i cell i, j
 E_{ij} : förväntade värden i cell i, j

Om de observerade tabellvärdena avviker för mycket från korstabellen som förväntas under H_0 blir χ^2 stor. Vi förkastar alltså nollhypotesen för stora värden av χ^2 .

Fördelning för testvariabeln under H_0 :

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(f)$$

χ^2 -fördelning med f frihetsgrader

www.matstat.org

Oberoendetest

Fördelning för testvariabeln under H_0 : $\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(f)$ χ^2 -fördelning med f frihetsgrader

Förutsättningar för χ^2 -fördelning: $E_i \geq 5$ i varje cell (n stor)

Frihetsgrader: $f = (r - 1) \cdot (c - 1)$ $r =$ antalet rader i korstabellen
 $c =$ antalet kolumner i korstabellen
 (utan marginalsummer!)

Testvariabeln χ^2 är $\chi^2(f)$ -fördelad om H_0 stämmer. Under H_0 borde alltså en observation av χ^2 inte ligga för långt ifrån centrumet av täthetsfunktionen för $\chi^2(f)$. Om observationen däremot ligger för långt på höger svansen av denna täthetsfunktion förkastar vi nollhypotesen. Det **kritiska området** är alltså:

$$\Omega_{krit} = \{ \chi^2 > \chi^2_{\alpha}(f) \} \quad \chi^2 \text{ är alltid positiv; } H_0 \text{ förkastas när } \chi^2 \text{ blir stor.}$$

www.matstat.org

Oberoendetest

Observed:

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

Expected (under H_0):

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

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

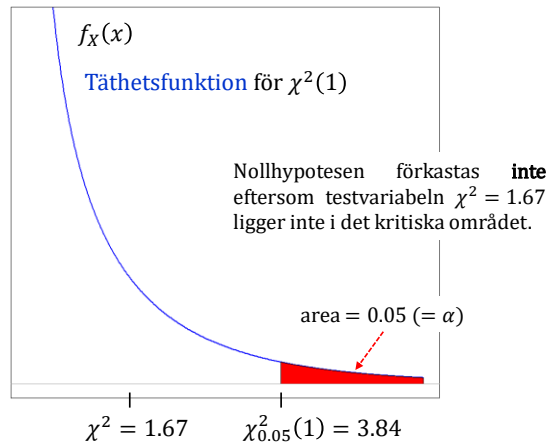
$$= \frac{(24 - 20)^2}{20} + \frac{(26 - 30)^2}{30} + \frac{(76 - 80)^2}{80} + \frac{(124 - 120)^2}{120} = \frac{103.71}{28.57} = \underline{\underline{1.67}}$$

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

H_0 förkastas **inte**. Vi kan inte visa på signifikansnivå $\alpha = 0.05$ att rökandet och kön är beroende variabler.

www.matstat.org

Oberoendetest



H_0 förkastas **inte**. Vi kan inte visa på signifikansnivå $\alpha = 0.05$ att rökandet och kön är beroende variabler.

www.matstat.org

Oberoendetest

Exempel: Föredrar två muränsarter olika miljöer?

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.



Stickprov: Det räknades hur ofta man såg två muränsarter i ett visst område i Belize. Det noterades vilka arter som sågs i vilken miljö.:

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

Två nominala variabler:

- art
- miljö

www.matstat.org

Oberoendetest

Exempel: Föredrar två muränsarter olika miljöer ?

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

H_0 : Arterna föredrar samma miljö.

H_a : Arterna föredrar olika miljöer.

Under H_0 förväntar vi oss att de procentuala andelarna för gräs, sand och gräns är ungefär lika för båda arterna, alltså t. ex. 30% gräs, 20% sand och 50% gräns för **båda** arterna.

Beräkning av korstabellen som förväntas under H_0 ("idealisk tabell") :

$$E_{ij} = \frac{R_i \cdot K_j}{n}$$

E_{ij} : förväntade värden i cell i, j
 R_i : radsumma för rad i
 K_j : kolumnsumma för kolumn j

Oberoendetest

observerat

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

förväntat

E_{ij}	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

$$\frac{243 \cdot 344}{834}$$

$$\chi^2 = \sum_{i,j}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \text{testvariabel}$$

$$= \frac{(127 - 142.8)^2}{142.8} + \frac{(116 - 100.2)^2}{100.2} + \frac{(99 - 97.5)^2}{97.5} + \frac{(67 - 68.5)^2}{68.5} + \frac{(264 - 249.7)^2}{249.7} + \frac{(161 - 175.3)^2}{175.3} = \underline{\underline{6.26}}$$

Det jämförs respektive celler i båda tabeller.

$$f = (r - 1) \cdot (c - 1) = 2 \cdot 1 = 2 \quad \text{frihetsgrader}$$

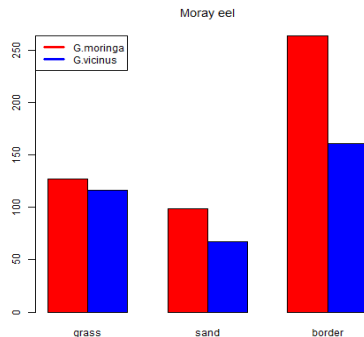
$$\Omega_{krit} = \{\chi^2 > \chi_{\alpha}^2\} = \{\chi^2 > \chi_{0.05}^2(2)\} = \{\chi^2 > \underline{\underline{5.99}}\}$$

www.matstat.org

Oberoendetest

Testvariabel: $\chi^2 = 6.26$ Kritiskt område: $\Omega_{krit} = \{ \chi^2 > 5.99 \}$

Testvariabeln ligger i det kritiska området. H_0 förkastas. Variablerna art och miljö är beroende, olika arter föredrar olika miljöer.



Proportionerna för de tre miljöer är ganska olika så att nollhypotesen kan förkastas.



?chisq.test

www.matstat.org

Fishers Exakt Test

- testar om kategoriska slumpvariabler är oberoende (som χ^2 -Test ovan)
- Skillnad till χ^2 -testet:
 - ingen normalapproximation (därför "exakt" test)
 - därför också användbart för $E_i \leq 5$ (eller just då!)
- H_0 : ingen association mellan variablerna (oberoendet)
- H_a : variablerna är beroende



I följande ska **p-värdet** för Fisher-testet beräknas.

Kom ihåg(föreläsning F11):

p-värdet är sannolikheten av en erhållen observation, eller ännu mer extrema observationer, givet att nollhypotesen är sann.

I motsats till alla andra fall som hittills behandlades är testresultatet för Fisher-testet inte något numeriskt värde på en testvariabel, utan själva korstabellen som härledas från stickprovet (= den "observerade tabellen").

www.matstat.org

Fisher-Test

För att beräkna p -värdet

- måste vi alltså först beräkna sannolikheten (under H_0) att den observerade tabellen kommer till stånd,
- ... och sedan addera sannolikheterna (under H_0) av alla tabeller som är ännu mer extrema än den observerade.

Exempel: Succé för 2 terapiformer A och B (5 patienter i grupp A, 7 patienter i grupp B, 8 blir friska, 4 blir inte friska)

	frisk	sjuk	Σ
A	2	3	5
B	6	1	7
Σ	8	4	12

"observerad tabell" =
testresultat

- 40% har blivit friska med terapi A
- 85% har blivit friska med terapi B
- 67% har blivit friska totalt

Är den observerade skillnaden
för andelen tillfrisknade patienter
statistiskt signifikant ?

www.matstat.org

Fisher-Test

	frisk	sjuk	Σ
A	2	3	5
B	6	1	7
Σ	8	4	12

Sannolikheten att denna
tabell kommer till stånd
"under H_0 " måste beräknas.

Sannolikheten för den observerade tabellen under H_0 kan beräknas genom att betrakta **marginalsummorna** (Σ) som **givna**. Vi förutsätter alltså att

- 5 patienter kommer att behandlas med terapi A
- 7 patienter kommer att behandlas med terapi B
- 8 patienter kommer att bli friska
- 4 patienter kommer inte att bli friska

Vi förutsätter alltså från början att 8 patienter blir friska. Detta motsvarar i viss mån nollhypotesen: terapin har inte något inflytande på det framtida hälsotillståndet.

www.matstat.org

Fisher-Test

Sannolikheten för den observerade tabellen under H_0 är lika med sannolikheten att precis 2 av de 8 som kommer att tillfriskna plockas ut slumpmässigt för terapi A. Sannolikheten att en cell i en tabell med givna marginalsummor antar ett visst värde är **hypergeometriskt** fördelat (se föreläsning **F4**):

	frisk	sjuk	Σ
A	$a_{11} = 2$	$a_{12} = 3$	5
B	$a_{21} = 6$	$a_{22} = 1$	7
Σ	8	4	12

Celler:

$a_{11} = \{A, \text{frisk}\}$

$a_{12} = \{A, \text{sjuk}\}$

$a_{21} = \{B, \text{frisk}\}$

$a_{22} = \{B, \text{sjuk}\}$

$$P(a_{11} = 2) = \frac{\binom{8}{2} \cdot \binom{4}{3}}{\binom{12}{5}}$$

- **Nämnare:** Vi väljer 5 av totalt 12 patienter för terapi A.
- **Täljare, faktor 1:** Bland dessa 5 är 2 av dem 8 som kommer att tillfriskna
- **Täljare, faktor 2:** Bland dessa 5 är 3 av dem 4 som inte kommer att tillfriskna

www.matstat.org

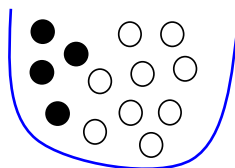
Fisher-Test

Analogi till urnmodellen (föreläsning **F4**):

Man kan föreställa sig att namnen på alla patienter skrivs på kulor som kommer att läggas i urnan. 5 patienter ska plockas ut för terapi A genom att dra 5 kulor slumpmässigt utan återläggning. Bland kulorna i urnan finns 8 som bär namn på patienter som blir friska, och 4 bär namn på patienter som inte blir friska. Vi söker sannolikheten att vi får 2 namn på patienter som tillfriskna bland de 5 som dras.

$$P(a_{11} = 2) = \frac{\binom{8}{2} \cdot \binom{4}{3}}{\binom{12}{5}}$$

↑
P("dra två vita")



$w = 8$ kommer att bli friska
 $s = 4$ kommer att förbli sjuka
 $N = 12$ totala antalet kulor
 $n = 5$ antalet dragna kulor

www.matstat.org

Fisher-Test

Det verkar vara godtyckligt att just beräkna sannolikheten för cell $a_{11} = 2$.
I själva verket spelar det dock ingen roll för vilken cell sannolikheten beräknas:

	frisk	sjuk	Σ
A	$a_{11} = 2$	$a_{12} = 3$	5
B	$a_{21} = 6$	$a_{22} = 1$	7
Σ	8	4	12

Celler:

$a_{11} = \{A, \text{frisk}\}$

$a_{12} = \{A, \text{sjuk}\}$

$a_{21} = \{B, \text{frisk}\}$

$a_{22} = \{B, \text{sjuk}\}$

$$P(a_{12} = 3) = \frac{\binom{4}{3} \cdot \binom{8}{2}}{\binom{12}{5}} \quad P(a_{11} = 2) = \frac{\binom{8}{2} \cdot \binom{4}{3}}{\binom{12}{5}} \quad \text{plocka 5 av 12 för terapi A}$$

$$P(a_{21} = 6) = \frac{\binom{8}{6} \cdot \binom{4}{1}}{\binom{12}{7}} \quad P(a_{22} = 1) = \frac{\binom{4}{1} \cdot \binom{8}{6}}{\binom{12}{7}} \quad \text{plocka 7 av 12 för terapi B}$$

Alla dessa sannolikheter är lika stora, eftersom $\binom{n}{k} = \binom{n}{n-k}$

www.matstat.org

Fisher-Test

- Beräknas sannolikheten för bara en enda cell av ovanstående tabell, så har man sannolikheten för hela tabellen, eftersom de andra cellerna följer automatiskt om marginalsummorna är givna.
- Detta motsvarar faktumet att ovanstående tabell bara har en **frihetsgrad**.
- Allmänt är **antalet frihetsgrader** för tabeller med c kolumner och r rader $f = (c - 1) \cdot (r - 1)$

	frisk	sjuk	Σ
A	$a_{11} = 2$	$a_{12} = ??$	5
B	$a_{21} = ??$	$a_{22} = ??$	7
Σ	8	4	12

Celler:

$a_{11} = \{A, \text{frisk}\}$

$a_{12} = \{A, \text{sjuk}\}$

$a_{21} = \{B, \text{frisk}\}$

$a_{22} = \{B, \text{sjuk}\}$

Med $a_{11} = 2$ följer automatiskt (för att marginalsummorna är givna):

$$\rightarrow a_{21} = 8 - 2 = 6 \quad \rightarrow a_{12} = 5 - 2 = 3 \quad \rightarrow a_{22} = 4 - a_{12} = 1$$

www.matstat.org

Fisher-Test

	frisk	sjuk	Σ
A	$a_{11} = 2$	$a_{12} = 3$	$Z_1 = 5$
B	$a_{21} = 6$	$a_{22} = 1$	$Z_2 = 7$
Σ	$S_1 = 8$	$S_2 = 4$	$N = 12$

Celler:

$a_{11} = \{A, \text{frisk}\}$

$a_{12} = \{A, \text{sjuk}\}$

$a_{21} = \{B, \text{frisk}\}$

$a_{22} = \{B, \text{sjuk}\}$

$$P(a_{11} = 2) = \frac{\binom{8}{2} \cdot \binom{4}{3}}{\binom{12}{5}} = P(\text{"tabell"})$$

$f = 1$ (en frihetsgrad,
en enda cell bestämmer alla
andra celler)

Allmänt:

gäller även för tabeller
med fler dimensioner!

$$P(\text{"tabell"}) = \frac{\binom{S_1}{a_{11}} \cdot \binom{S_2}{a_{12}}}{\binom{N}{Z_1}} = \frac{S_1! \cdot S_2! \cdot Z_1! \cdot Z_2!}{N! \cdot a_{11}! \cdot a_{12}! \cdot a_{21}! \cdot a_{22}!}$$

www.matstat.org

Fisher-Test

	frisk	sjuk	Σ
A	$a_{11} = 2$	$a_{12} = 3$	$Z_1 = 5$
B	$a_{21} = 6$	$a_{22} = 1$	$Z_2 = 7$
Σ	$S_1 = 8$	$S_2 = 4$	$N = 12$

Celler:

$a_{11} = \{A, \text{frisk}\}$

$a_{12} = \{A, \text{sjuk}\}$

$a_{21} = \{B, \text{frisk}\}$

$a_{22} = \{B, \text{sjuk}\}$

$$P(\text{"tabell"}) = \frac{\binom{S_1}{a_{11}} \cdot \binom{S_2}{a_{12}}}{\binom{N}{Z_1}} = \frac{S_1! \cdot S_2! \cdot Z_1! \cdot Z_2!}{N! \cdot a_{11}! \cdot a_{12}! \cdot a_{21}! \cdot a_{22}!}$$

hypergeometrisk
fördelning

$$= \frac{8! \cdot 4! \cdot 5! \cdot 7!}{12! \cdot 2! \cdot 3! \cdot 6! \cdot 1!} = \frac{14}{99} = \underline{\underline{0.141414}}$$

Sannolikheten för den observerade tabellen under antagandet av nollhypotesen är $P_1 = 0.141414$.

www.matstat.org

Fisher-Test

	frisk	sjuk	Σ
A	$a_{11} = 2$	$a_{12} = 3$	$Z_1 = 5$
B	$a_{21} = 6$	$a_{22} = 1$	$Z_2 = 7$
Σ	$S_1 = 8$	$S_2 = 4$	$N = 12$

Celler:

$a_{11} = \{A, \text{frisk}\}$

$a_{12} = \{A, \text{sjuk}\}$

$a_{21} = \{B, \text{frisk}\}$

$a_{22} = \{B, \text{sjuk}\}$

```
## Hypergeometric Distribution

?dhyper

dhyper(x = 2, m = 8, n = 4, k = 5)
# 0.1414141

# explicit calculation:
p = choose(8,2) * choose(4,3) / choose(12,5)
p # 0.1414141
```



www.matstat.org

Fisher-Test



Minitab

Hypergeometric Distribution

☒ Probability
☐ Cumulative probability
☐ Inverse cumulative probability

Population size (N): 12
 Event count in population (M): 8
 Sample size (n): 5

☐ Input column:
 Optional storage:

☒ Input constant: 2
 Optional storage:

Select OK Cancel

Calc / Probability Distributions

Hypergeometric with $N = 12$, $M = 8$, and $n = 5$
 $x \quad P(X = x)$
 2 0,141414

www.matstat.org

Fisher-Test

För att beräkna p -värdet

- måste vi först beräkna sannolikheten (under H_0) att den observerade tabellen kommer till stånd ... *ok, done!*
- ... och sedan addera sannolikheterna (under H_0) av alla tabeller som är ännu mer extrema än den observerade.

För att bearbeta punkt 2 kan man gå två olika vägar:

- 1) Fundera på vilka tabeller som är ännu mer extrema under nollhypotesen än den observerade tabellen och addera deras sannolikheter (se appendix).
- 2) Beräkna sannolikheterna av **alla** tabeller under H_0 och sedan addera alla sannolikheter som är mindre än eller lika med sannolikheten av den observerade tabellen.

Här ska metod 2) visas → beräkna sannolikheterna av alla möjliga tabeller:

	frisk	sjuk	Σ
A	a_{11}	a_{12}	5
B	a_{21}	a_{22}	7
Σ	8	4	12

Cellen a_{12} är bäst att använda som pivot-element, eftersom a_{12} kan bara anta 5 olika värden: 0, 1, 2, 3, 4 och 5. (de andra cellerna följer återigen automatiskt om a_{12} är fastlagd)

www.matstat.org

Fisher-Test

a_{11}	a_{12}	Z_1
a_{21}	a_{22}	Z_2
S_1	S_2	N

$$P(\text{"tabell"}) = \frac{S_1! \cdot S_2! \cdot Z_1! \cdot Z_2!}{N! \cdot a_{11}! \cdot a_{12}! \cdot a_{21}! \cdot a_{22}!}$$

5	$a_{12} = 0$	5
3	4	7
8	4	12

$$P = 0.070707$$

4	$a_{12} = 1$	5
4	3	7
8	4	12

$$P = 0.353535$$

3	$a_{12} = 2$	5
5	2	7
8	4	12

$$P = 0.424242$$

2	$a_{12} = 3$	5
6	1	7
8	4	12

$$P = 0.141414$$

(observation)

1	$a_{12} = 4$	5
7	0	7
8	4	12

$$P = 0.010101$$

p -värdet blir alltså:

$$p = 0.1414 + 0.0707 + 0.0101$$

$$p \approx 0.222$$

www.matstat.org

Beräkning av sannolikheterna i R:

```
C:\Users\Uwe\Desktop\TALKS_POSTERS\Magdeburg_Talk\R_scripts_for_Magdeburg_Talk.R - R Editor

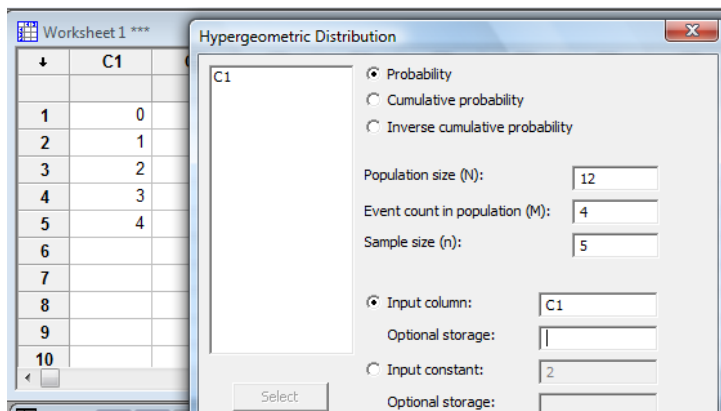
## Hypergeometric Distribution|

white_urn = 8      # the number of white balls in the urn
black_urn = 4      # the number of black balls in the urn
total_drawn = 5    # the number of balls drawn from the urn

for (a12 in 0:4) {
  white_drawn = total_drawn - a12
  p = dhyper(x = white_drawn, m = white_urn, n = black_urn, k = total_drawn)
  p = round(p, 6)
  cat(paste("a12 =", a12, " p =", p, "\n"))
}

# a12 = 0    p = 0.070707
# a12 = 1    p = 0.353535
# a12 = 2    p = 0.424242
# a12 = 3    p = 0.141414
# a12 = 4    p = 0.010101
```

www.matstat.org



Hypergeometric with $N = 12$, $M = 4$, and $n = 5$

x	$P(X = x)$
0	0,070707
1	0,353535
2	0,424242
3	0,141414
4	0,010101

www.matstat.org



Tillämpning av testet i R:

```
## Fischers Exakter Test

A = matrix(c(2,3,6,1), nrow = 2, byrow = TRUE)
A
#      [,1] [,2]
# [1,]    2    3
# [2,]    6    1

fisher.test(A)

#      Fisher's Exact Test for Count Data
#
# data:  A
# p-value = 0.2222
# alternative hypothesis: true odds ratio is not equal to 1
# 95 percent confidence interval:
# | 0.001830735 2.834757137
```

2	3	5
6	1	7
8	4	12

(observation)

www.matstat.org

Tillämpning av testet i Minitab:

2 Proportions (Test and Confidence Interval)

Stat / Basic Statistics / 2 Proportions

Subscripts:

☐ Samples in different columns

First:

Second:

☒ Summarized data

Events: Trials:

First:

Second:

Select

Options...

Help

OK

Cancel

Minitab



Normalapproximation
kan inte tillämpas

Test for difference = 0 (vs not = 0): Z = -1,79 P-Value = 0,074

Fisher's exact test: P-Value = 0,222

* NOTE * The normal approximation may be inaccurate for small samples.

www.matstat.org

Fisher-Test

	frisk	sjuk	Σ
A	$a_{11} = 2$	$a_{12} = 3$	5
B	$a_{21} = 6$	$a_{22} = 1$	7
Σ	8	4	12

$p. value = 0.2222$

- 40% har blivit friska i grupp A
- 85% har blivit friska i grupp B

Nollhypotes (H_0): $p_1 = p_2$ (samma proportioner, samma andel friska)

p -värdet är stort ($p > 0.05$). Nollhypotesen kan därför **inte** förkastas. Det kunde **inte** visas att skillnaden mellan de observerade proportionerna är statistiskt signifikant, dvs. vi måste fortfarande utgå ifrån att båda terapiformer är lika framgångsrika (terapiform och terapiframgång är **oberoende**).

Detta är förvånansvärt om man tittar på själva proportionerna. Orsaken är den dåliga försöksplaneringen – **stickprovet är helt enkelt för litet**.

www.matstat.org

p -värdet för ett stort stickprov med samma proportioner

```
## Fisher's Exakt Test - big sample

A = matrix(c(200,300,600,100), nrow = 2, byrow = TRUE)
A
#           [,1] [,2]
# [1,]    200  300
# [2,]    600  100

fisher.test(A)

#           Fisher's Exact Test for Count Data

# data:  A
# p-value < 2.2e-16
# alternative hypothesis: true odds ratio is not equal to 1
# 95 percent confidence interval:
#  0.08335038 0.14790301
```

www.matstat.org

p-värden för ett litet och ett stort stickprov med samma andel tillfriskade

	frisk	sjuk	Σ
A	2	3	5
B	6	1	7
Σ	8	4	12

p. value = 0.222

- 40% friska i grupp A
- 85% friska i grupp B

	frisk	sjuk	Σ
A	200	300	500
B	600	100	700
Σ	800	400	1200

p. value < 10^{-16}

På grund av att stickprovet är för litet räcker inte den **statistiska noggrannheten** ("power") för att förkasta nollhypotesen. Detta hade man kunnat se i förväg. För detta exempel hade bara det mest extrema resultatet kunnat leda till ett *p*-värde under 0.05 ... för $a_{12} = 4$ (se nästa sida).

www.matstat.org

$$a_{12} = 4$$

Bara denna observation hade kunnat leda till ett signifikant resultat (med $\alpha = 0.05$):

- 20% friska i grupp A
- 100% friska i grupp B

	frisk	sjuk	Σ
A	1	4	5
B	7	0	7
Σ	8	4	12

```

C:\Users\Uwe\Desktop\TALKS_POSTERS\Magdeburg_Talk\R_scripts_for_Magdeburg_Talk.R - R Editor

## Fisher's Exakt Test - the "most extreme" case

A = matrix(c(1,4,7,0), nrow = 2, byrow = TRUE)
A
#      [,1] [,2]
# [1,]    1    4
# [2,]    7    0

fisher.test(A)
#
#      Fisher's Exact Test for Count Data
#
# data:  A
# p-value = 0.0101
  
```

www.matstat.org

Appendix

Anpassningstest och Oberoendetest

Uwe Menzel, 2018
uwe.menzel@matstat.org
www.matstat.org

Goodness-of-fit Test (frihetsgrader)

Täthetsfunktionen $\chi^2(f)$ beror av antalet frihetsgrader (f).

Antalet frihetsgrader för
anpassningstestet:

$$f = k - r - p$$

- k = antalet kategorier (frisk, sjuk = två kategorier)
- r = antalet restriktioner (t. ex.: summan av alla observationer är fastlagd, se exempel tärning ovan)
- p = antalet parametrar som måste skattas ur stickprovet för att upprätta modellen (se följande sidor)

I följande ska två fall behandlas:

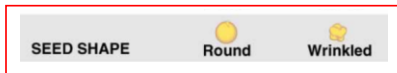
- 1) Ingen parameter måste skattas ur stickprovet ($p = 0$)
- 2) En parameter måste skattas ur stickprovet ($p = 1$)

www.matstat.org

Anpassningstest

1) Ingen parameter måste skattas ur stickprovet ($p = 0$)

Mendel's Experiment:



Två alleler: A, a **allel A är dominant**

Mendel korsade bara heterozygota (Aa) bönor (allel-frekvenserna är båda 0.5):

	A	a
A	AA	Aa
a	Aa	aa

Genotyper AA och Aa blir runda,
bara genotyp aa blir ynklig.

- Om Mendel's lag gäller, så förväntar man sig en proportion 3:1 för kvoten runda : ynkliga.
- **Observation:** 423 runda och 133 ynkliga i ett sådant experiment.

www.matstat.org

Anpassningstest

1) Ingen parameter måste skattas ur stickprovet ($p = 0$)

	rund	ynklig	Σ
O_i	423	133	556
E_i	$556 \cdot 3/4 = 417$	$556 \cdot 1/4 = 139$	556

$n = 556$ stickprov

$r = k - 1 = 1$

1 frihetsgrad

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \frac{(423 - 417)^2}{417} + \frac{(133 - 139)^2}{139} = \underline{\underline{0.345}} \quad \text{test-variabel}$$

$$\Omega_{krit} = \{ \chi^2 > \chi^2_{\alpha}(f) \} = \{ \chi^2 > \chi^2_{0.05}(1) \} = \{ \chi^2 > 3.84 \} \quad \text{kritiskt område}$$

Testvariabeln ligger **inte** i det kritiska området. Nollhypotesen förkastas **inte**. En **signifikant skillnad** mellan observation och modell kunde inte påvisas. Stickprovet står inte i motsats till modellen (till Mendels lag).

www.matstat.org

Anpassningstest

2) En parameter måste skattas ur stickprovet ($p = 1$)

- Ofta känner man dock inte till allel- frekvenserna (p, q) från början som i Mendels experiment (där $p = q = 0.5$ var bekant, eftersom bara heterozygota (genotyp Aa) bönor korsades)
- Hardy-Weinberg-Modell:** allel-frekvenserna förblir konstant över generationerna.

Låt allel-frekvenserna i första generationen vara $f(A) = p$ och $f(a) = q$, **båda okända**. För den efterföljande generationen gäller:

$$f(AA) = p^2$$

$$f(Aa) = 2 \cdot p \cdot q$$

$$f(aa) = q^2$$

	A	a
A	$AA (p^2)$	$Aa (pq)$
a	$Aa (pq)$	$aa (q^2)$

www.matstat.org

Anpassningstest

För att beräkna de enligt modellen förväntade frekvenserna AA , Aa och aa måste vi först beräkna allel-frekvenserna p och q ur stickprovet.

Exempel: http://en.wikipedia.org/wiki/Hardy-Weinberg_principle, lite ändrat.



Callimorpha dominula
Linnaeus, 1758

Stickprov:

genotyp	White-spotted (AA)	Intermediate (Aa)	Little spotting (aa)	$\Sigma (=n)$
observerad	1329	268	15	1612

$$p = \frac{2 \cdot \text{antalet}(AA) + \text{antalet}(aA)}{2 \cdot n} = \frac{2 \cdot 1329 + 268}{2 \cdot 1612} = 0.907 \quad \text{frekvens för allel } A$$

$$q = \frac{2 \cdot \text{antalet}(aa) + \text{antalet}(aA)}{2 \cdot n} = 1 - p = 0.093 \quad \text{frekvens för allel } a$$

(n individer har totalt $2 \cdot n$ alleler)

Anpassningstest

Beräkning av de under HW-modellen **förväntade** frekvenserna ("E = expected"):

$$\begin{aligned} E(AA) &= n \cdot p^2 = 1612 \cdot 0.907^2 = 1326.11 \\ E(Aa) &= n \cdot 2 \cdot pq = 271.95 \\ E(aa) &= n \cdot q^2 = 13.94 \end{aligned} \quad \left. \vphantom{\begin{aligned} E(AA) \\ E(Aa) \\ E(aa) \end{aligned}} \right\} \begin{array}{l} \text{dessa värden måste jämföras} \\ \text{med de observerade värdena} \\ \text{(tabell ovan)} \end{array}$$

genotyp:	White-spotted (AA)	Intermediate (Aa)	Little spotting (aa)	$\Sigma (= n)$
O_i	1329	268	15	1612
E_i	1326.11	271.95	13.94	1612

Testvariabel:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \frac{(1329 - 1326.11)^2}{1326.11} + \frac{(268 - 271.95)^2}{271.95} + \frac{(15 - 13.94)^2}{13.94} = \underline{0.144}$$

www.matstat.org

Anpassningstest

Antalet frihetsgrader för anpassningstestet:

$$f = k - r - p$$

- Antalet kategorier: $k = 3$ (genotyp AA, Aa, aa)
- Antalet restriktioner: $r = 1$ (totala antalet fastlagd: $n = 1612$)
- Antalet parametrar som måste skattas ur stickprovet för att etablera modellen: $p = 1$ (när en av de båda allel-frekvenserna beräknades följer den andra automatiskt: $q = p - 1$)

$$\Rightarrow f = 3 - 1 - 1 = 1 \quad \text{allmänt för HW-modell: } f = \text{antalet genotyper minus antalet alleler}$$

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

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = 0.144 \quad \text{testvariabel}$$

Testvariabeln ligger **inte** i det kritiska området. Observationen står inte i motsats till Hardy-Weinberg-modellen.

Fisher-test

Exempel s. 31: Succé för 2 terapiformer A och B

Identifiering av de mera extrema observationerna med hjälp av χ^2 :

- Ovan ansågs sådana resultat som mera extrema (än observationen), som hade en mindre sannolikhet under H_0 än observationen.
- En **alternativ metod** anser sådana tabeller som mera extrema som har ett större "avstånd" till " H_0 -tabellen" än den observerade tabellen. Detta avstånd "mäts" med hjälp av χ^2 :

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \chi^2 = \text{avstånd av en tabell till tabellen som förväntas under } H_0 \text{ ("} H_0\text{-tabellen")}$$

O_{ij} : värde i cellen i, j av en given tabell

$E_{i,j}$: under nollhypotesen förväntat värde i cellen i, j

Viktigt: χ^2 används här **inte** som testvariabel, utan bara för att "mäta" avståndet av en given tabell till " H_0 -tabellen"

www.matstat.org

Fisher-test

För att beräkna $E_{i,j}$ måste vi först hitta tabellen som förväntas under H_0 :

Nollhypotes (H_0): $p_1 = p_2$ (samma andel tillfrisknade för A och B)

	frisk	sjuk	Σ
A	E_{11}	...	5
B	E_{21}	...	7
Σ	8	4	12

Om andelen tillfrisknade är lika stor i båda grupperna måste detta gälla:

$$\frac{E_{11}}{5} = \frac{8}{12} \quad \text{och} \quad \frac{E_{21}}{7} = \frac{8}{12} \quad H_0: \text{andelen tillfrisknade i båda grupperna är } 8/12 = 2/3$$

↑
andel tillfrisknade i
grupp A
↑
andel tillfrisknade i
grupp B

www.matstat.org

Fisher-test

För att beräkna $E_{i,j}$ måste vi först hitta tabellen som förväntas under H_0 :

Nollhypotes (H_0): $p_1 = p_2$ (samma andel sjuka för A och B)

	frisk	sjuk	Σ
A	...	E_{12}	5
B	...	E_{22}	7
Σ	8	4	12

Analogt måste andelen sjuka i båda grupperna vara $1/3$:

$$\frac{E_{12}}{5} = \frac{4}{12} \quad \text{och} \quad \frac{E_{22}}{7} = \frac{4}{12}$$

$\uparrow$ andel sjuka i grupp A $\uparrow$ andel sjuka i grupp B

H_0 : andelen sjuka i båda grupperna är $4/12 = 1/3$

www.matstat.org

Fisher-test

$$\frac{E_{11}}{5} = \frac{8}{12} \quad \frac{E_{12}}{5} = \frac{4}{12} \quad \frac{E_{21}}{7} = \frac{8}{12} \quad \frac{E_{22}}{7} = \frac{4}{12}$$

Allmänt gäller för den under H_0 "förväntade" tabellen:

$$E_{11} = \frac{Z_1 \cdot S_1}{N} \quad E_{21} = \frac{Z_2 \cdot S_1}{N} \quad E_{12} = \frac{Z_1 \cdot S_2}{N} \quad E_{22} = \frac{Z_2 \cdot S_2}{N}$$

Varje under H_0 förväntade cell beräknas alltså som: **radsumma gånger kolumnsumma delad med det totala antalet**:

	frisk	sjuk	Σ
A	$10/3$	$5/3$	5
B	$14/3$	$7/3$	7
Σ	8	4	12

$$E_{ij} = \frac{Z_i \cdot S_j}{N}$$

Under H_0 förväntade (idealisk) tabell

www.matstat.org

Fisher-test

	friska	sjuka	Σ
A	$10/3$	$5/3$	5
B	$14/3$	$7/3$	7
Σ	8	4	12

$$E_{ij} = \frac{Z_i \cdot S_j}{N}$$

Under H_0 förväntade
(idealisk) tabell

- Med hjälp av χ^2 beräknas "avståndet" av alla möjliga tabeller till denna tabell (OBS: även avståndet av den observerade tabellen till denna tabell beräknas).
- För att hitta alla "möjliga" tabeller använder vi igen elementet a_{12} som pivot.
- Elementet a_{12} kan på grund av de givna marginalsummorna bara anta värdena 0,1,2,3 eller 4.
- Är a_{12} fastlagt, så följer alla andra celler automatiskt.

www.matstat.org

Fisher-test

Exempel: avstånd av den ideala (H_0)-tabellen till den observerade tabellen:

observerade tabell

	friska	sjuka	Σ
A	2	3	5
B	6	1	7
Σ	8	4	12

H_0 - tabell

	friska	sjuka	Σ
A	$10/3$	$5/3$	5
B	$14/3$	$7/3$	7
Σ	8	4	12

$$\chi_0^2 = \sum_{i,k=1}^K \frac{(O_{ik} - E_{ik})^2}{E_{ik}} = \frac{(2 - 10/3)^2}{10/3} + \frac{(3 - 5/3)^2}{5/3} + \frac{(6 - 14/3)^2}{14/3} + \frac{(1 - 7/3)^2}{7/3}$$

$$\chi_0^2 = \underline{2.743} \quad \text{"avstånd" observerad tabell} \rightleftharpoons \text{under } H_0 \text{ förväntad tabell}$$

Det beräknas nu χ^2 för alla möjliga tabeller. Alla dessa χ^2 jämförs med χ_0^2 . Om $\chi^2 \geq \chi_0^2$ anses motsvarande tabell som mer extrem än tabellen som motsvarar observationen, och dess sannolikhet under H_0 ingår i p -värdet.

www.matstat.org

Beräkning av avståndet för alla tabeller

```

C:\Users\Uwe\Desktop\TALKS_POSTERS\Magdeburg_Talk\R_scripts_for_Magdeburg_Talk.R - R Editor

## Chi-squared for test of independence
Uwe Menzel, 2010

S1 = 8 # column 1
S2 = 4 # column 2
Z1 = 5 # row 1
Z2 = 7 # row 2
N = 12 # total

E11 = Z1 * S1 / N
E12 = Z1 * S2 / N
E21 = Z2 * S1 / N
E22 = Z2 * S2 / N

E = matrix(c(E11, E12, E21, E22), nrow = 2, byrow = TRUE)

for (a12 in 0:4) {
  a11 = Z1 - a12
  a22 = S2 - a12
  a21 = S1 - a11
  A = matrix(c(a11, a12, a21, a22), nrow = 2, byrow = TRUE)
  chi2 = sum((E-A)^2/E)
  chi2 = round(chi2, 3)
  cat(paste("a12 =", a12, " chi2 =", chi2, "\n"))
}

```

Beräkning χ^2 -"avstånd" för alla tabeller



	C1-T	C2-T	C3
	treatment	state	count
1	control	sick	3
2	therapy	sick	1
3	control	healthy	2
4	therapy	healthy	6
5			
6			
7			
8			
9			
10			

för alla 5 tabeller!

Cross Tabulation and Chi-Square

Categorical variables:

For rows:

For columns:

For layers:

Frequencies are in: (optional)

Display

☒ Counts

☐ Row percents

☐ Column percents

☐ Total percents

Stat / Tables / Cross Tabulation and Chisquare

www.matstat.org

χ^2 och p för alla möjliga tabeller

tabell	χ^2	p									
<table border="1"> <tr><td>5</td><td>0</td><td>5</td></tr> <tr><td>3</td><td>4</td><td>7</td></tr> <tr><td>8</td><td>4</td><td>12</td></tr> </table>	5	0	5	3	4	7	8	4	12	4.286	0.070707
5	0	5									
3	4	7									
8	4	12									
<table border="1"> <tr><td>4</td><td>1</td><td>5</td></tr> <tr><td>4</td><td>3</td><td>7</td></tr> <tr><td>8</td><td>4</td><td>12</td></tr> </table>	4	1	5	4	3	7	8	4	12	0.686	0.353535
4	1	5									
4	3	7									
8	4	12									
<table border="1"> <tr><td>3</td><td>2</td><td>5</td></tr> <tr><td>5</td><td>2</td><td>7</td></tr> <tr><td>8</td><td>4</td><td>12</td></tr> </table>	3	2	5	5	2	7	8	4	12	0.171	0.424242
3	2	5									
5	2	7									
8	4	12									
<table border="1"> <tr><td>2</td><td>3</td><td>5</td></tr> <tr><td>6</td><td>1</td><td>7</td></tr> <tr><td>8</td><td>4</td><td>12</td></tr> </table>	2	3	5	6	1	7	8	4	12	2.743	0.141414
2	3	5									
6	1	7									
8	4	12									
<table border="1"> <tr><td>1</td><td>4</td><td>5</td></tr> <tr><td>7</td><td>0</td><td>7</td></tr> <tr><td>8</td><td>4</td><td>12</td></tr> </table>	1	4	5	7	0	7	8	4	12	8.4	0.010101
1	4	5									
7	0	7									
8	4	12									

Möjliga är alla tabeller som tillåts med hänsyn till de givna rad- och kolumnsummorna

$$\chi^2 = \sum_{i,j=1} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad E_{ij} = \frac{Z_i \cdot S_j}{N}$$

$$P(\text{"table"}) = \frac{S_1! \cdot S_2! \cdot Z_1! \cdot Z_2!}{N! \cdot a_{11}! \cdot a_{12}! \cdot a_{21}! \cdot a_{22}!}$$

Röd markerade: mindre sannolikhet under H_0 jämfört med observationen och även större χ^2 jämfört med observationen (här överensstämmer båda kriterierna). Motsvarande sannolikheter bidrar till p -värdet:

$$p = 0.0707 + 0.1414 + 0.0101 \approx \underline{\underline{0.222}}$$