Chi-Square test

$$\chi^2$$

1. *Anpassningstest (Goodness of Fit)*
2. *Oberoendetest (Independence Test)*

1. Anpassningstest

uwe.menzel@genpat.uu.se

Vad gör ett anpassningstest?

- Hur bra passar en statistisk modell till observerade data? (är modellen bra?)
- Följer observationerna en förmodad fördelning?

1. Är mina data normalfördelade?:
   - Kolmogorov–Smirnov test (Minitab)
   - Shapiro–Wilk test
   - Anderson–Darling test
   - ...
2. Multinomial experiment: Får jag resultat som jag förväntar mig enligt min modell?:
   - Pearson's chi-square test, Goodness of Fit test

Multinomial experiment

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

Experiment: kastar tärningen 120 (=n) gånger

	1	2	3	4	5	6
expected $E_i=p_i \cdot n$	20	20	20	20	20	20
observed	21	16	28	24	19	12

Summan över råden måste vara 120 – en restriktion

Med detta resultat i handen: Kan jag fortfaranda tro att min modell (rättvis tärning) stämmer? ... eller måste jag förkasta denna (noll)hypotes?

Multinomial experiment

Om resultatet hade varit som i tabellen nedan hade jag väl förkastad nollhypotesen (rättvis tärning) ...

	1	2	3	4	5	6
expected $E_i=p_i \cdot n$	20	20	20	20	20	20
observed	0	3	2	1	1	113

$\Sigma\, o_i = 120$

Var går gränsen? Vilket värde får skillnaden mellan ”observed” och ”expected” maximalt ha för att bibehålla H_0? (Hur mäts skillnaden över-huvudtaget?) → vi behöver en *fördelning* för ”skillnaden”, dvs. sannolikheten att en viss skillnad uppstår! Om denna sannolikhet är liten, förkastar vi H_0.

En testvariabel som ”mäter” skillnaden

observed values

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

expected values

Fördelningen av denna testvariabel kan räknas ut, givet att nollhypotesen gäller (och att n är stor)
→ Chi-Square fördelningen (χ^2)

OBS: Testvariabeln är χ^2-fördelad och kallas ofta också χ^2

χ^2 mäter "skillnaden" till nollhypotesen

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

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

Skillnad mellan modell och observation liten → χ^2 liten

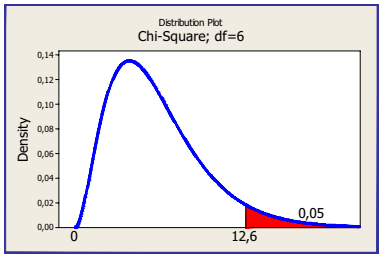
	1	2	3	4	5	6
E_i	20	20	20	20	20	20
O_i	0	3	2	1	1	113

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

Skillnad mellan modell och observation stor → χ^2 stor

Pearson's chi-square test

- Fördelningen för testvariabeln är känd under H_0 → vi kan räkna ut "hur sannolikt" varje värde av testvariabeln är.
- Vi förkastar nollhypotesen (dvs. modellen) om var observation leder till ett χ^2-värde som är mycket osannolikt under H_0 (t.ex < 5%)

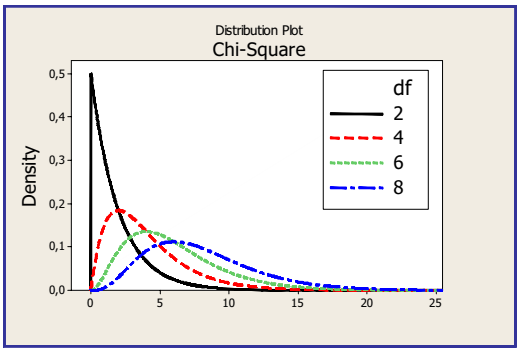


χ^2 är alltid positiv och nollhypotesen förkastas för stora värden → upper tail test

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

$$\Omega_{krit} = \{ \chi^2 > \chi^2_\alpha \} \quad \textit{upper tail}$$

PDF beror på antalet frihetsgrader



Antalet frihetsgrader för χ^2-testet

antalet celler (6 för tärningen)

antalet linjära restriktioner (1 för tärningen: summa för en råd = n)

$$df = k - r - p$$

antalet parametrar som skattas för att få en modell

$$df = k - 1$$

oftast är det bara så här, dvs. r=1 och p=0

Förutsättningar

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

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

← alla $E_i \geq 5$

" I don't like mondays ..."


Boomtown Rats

Hjärtattack (n=200 patienter):

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

Risken jämnfördelad – eller är måndag farligare?

$$H_0: \quad p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = 1/7$$

$$n = 200 \qquad E_i = n \cdot p_i = 200/7 \approx 28.57 \quad \textit{expected}$$

$$n_i \geq 5 \quad \textit{okay}$$

Om hjärtattack är jämnfördelade och man registrerar 200 fall, så skulle det bli omkring 28/29 per dag ...

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

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

$$+ \frac{(32-28.57)^2}{28.57} + \frac{(26-28.57)^2}{28.57} + \frac{(29-28.57)^2}{28.57}$$

$$= \frac{103.71}{28.57} = 3.63$$

H_0 förkastas inte!

$$\Omega_{krit} = \{\chi^2 > \chi^2_\alpha\} = \{\chi^2 > \chi^2_{0.05}(6)\} = \{\chi^2 > 12.59\} \qquad \textit{tabell}$$

Tabell χ^2

Table 6. Percentage points of the χ^2 distributions

icke symmetrisk!

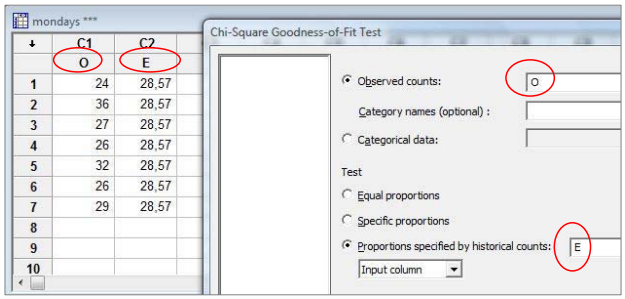
α=0.05

d.f.	$\chi^2_{0.995}$	$\chi^2_{0.990}$	$\chi^2_{0.975}$	$\chi^2_{0.950}$	$\chi^2_{0.050}$	$\chi^2_{0.025}$	$\chi^2_{0.010}$
1	0.0000393	0.0001571	0.0009821	0.0039321	3.84146	5.02389	6.63490
2	0.0100251	0.0201007	0.0506356	0.102587	5.99147	7.37776	9.21034
3	0.0717212	0.114832	0.215795	0.351846	7.81473	9.34840	11.3449
4	0.206990	0.297110	0.484419	0.710721	9.48773	11.1433	13.2767
5	0.411740	0.554300	0.831211	1.145476	11.0705	12.8325	15.0863
6	0.675727	0.872085	1.237347	1.63539	12.5916	14.4494	16.8119
7	0.989265	1.239043	1.68987	2.16735	14.0671	16.0128	18.4753
8	1.344419	1.646482	2.17973	2.73264	15.5073	17.5346	20.0902
9	1.734926	2.087912	2.70039	3.32511	16.9190	19.0228	21.6660

chisquare.MPJ

Minitab

Stat / Tables / Chi-Square Goodness-of-Fit Test



Minitab

$$\frac{(O_1 - E_1)^2}{E_1} = \frac{(24-28.57)^2}{28.57}$$

***Chi-Square Goodness-of-Fit Test* for Observed Counts in Variable: O**

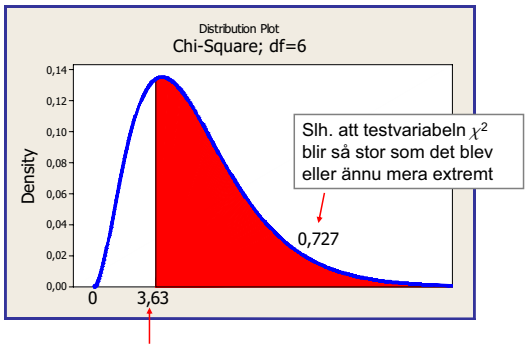
Category	Observed	Historical Counts	Test Proportion	Expected	Contribution to Chi-Sq
1	24	28,57	0,142857	28,5714	0,73143
2	36	28,57	0,142857	28,5714	1,93143
3	27	28,57	0,142857	28,5714	0,08643
4	26	28,57	0,142857	28,5714	0,23143
5	32	28,57	0,142857	28,5714	0,41143
6	26	28,57	0,142857	28,5714	0,23143
7	29	28,57	0,142857	28,5714	0,00643

N	DF	Chi-Sq	P-Value
200	6	3,63	0,727

Nollhypotesen förkastas inte, högt p-värde.

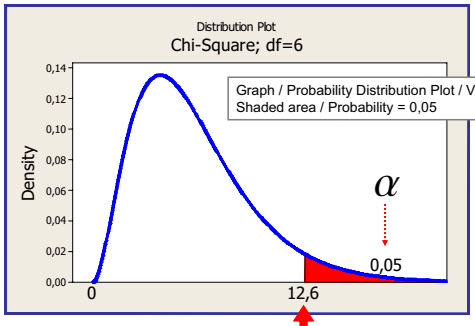
k-1 *hade vi förut*

P-värdet



det var värdet för testvariabeln χ^2

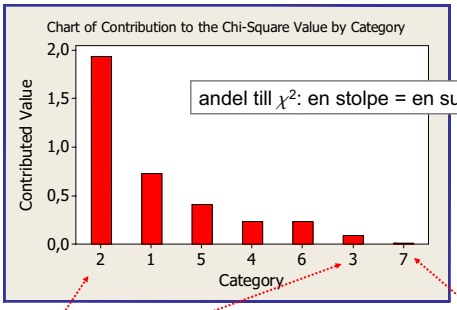
Kvantil



det var vad vi också hittade i tabellen

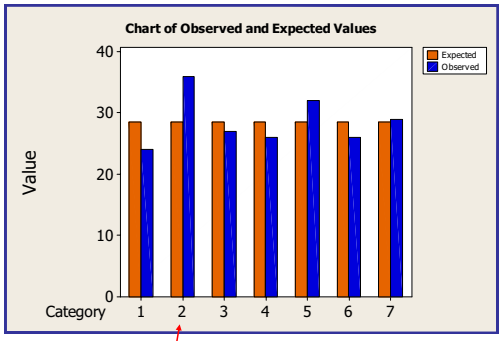
$$\chi^2_\alpha(f) = \chi^2_{0.05}(6)$$

Vilket värde avviker hur mycket från vara förväntningar?



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

Jämförelse ”observed” och ”expected”



på måndag skiljer sig det observerade värdet mest från förväntningen

Var fåglarna söker frö

Antagandet (modell): Fåglarna föredrar *inte* någon sorts träd, de söker frö i alla träd med jämn sannolikhet – ju mer träd av en viss sort förekommer, desto mer fåglar finns där ...

Experiment: n=156 fåglar observerades i en skog i Oregon

	ädelgran	furu	kust gran	lärk	summa
kronornas volym	54%	40%	5%	1%	100%
expected	156·0,54=84,24	156·0,4=62,4	156·0,05=7,8	156·0,01=1,56	156
observed	70	79	3	4	156

för liten!!

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**.

sammanfattat → $E_i \geq 5$

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

$n_i \geq 5 \quad okay$

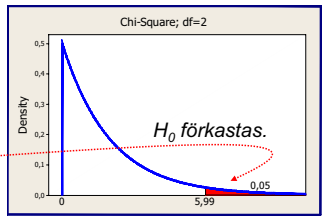
$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(70-84.24)^2}{84.24} + \frac{(79-62.4)^2}{62.4} + \frac{(7-9.36)^2}{9.36}$$

$$= 2.407 + 4.416 + 0.595$$

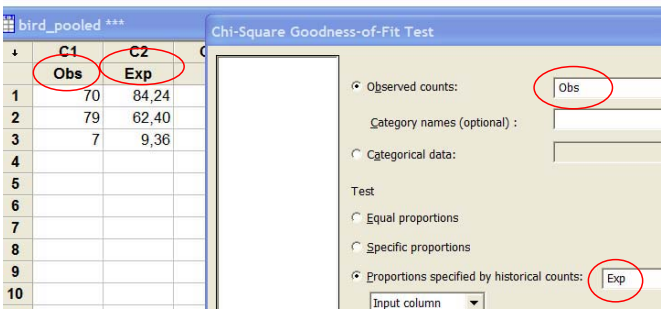
$$= 7.418$$

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



Minitab

Stat / Tables / Chi-Square Goodness of Fit



Minitab

Results for: bird_pooled

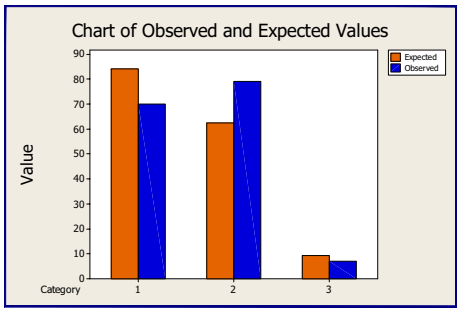
Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: Obs

Category	Observed	Historical Counts	Test Proportion	Expected	Contribution to Chi-Sq
1	70	84,24	0,54	84,24	2,40714
2	79	62,40	0,40	62,40	4,41603
3	7	9,36	0,06	9,36	0,59504

N	DF	Chi-Sq	P-Value
156	2	7,41821	0,024

P<0.05: Nollhypotesen förkastas.
Uppenbarligen föredrar fåglarna vissa träd.

Jämförelse observed/expected



Som det ser ut, föredrar fåglarna kategori 2, alltså furu. Här finns fler fåglar än förväntad under H_0.

Måste man sammanfatta cellerna om n<5?

Obs	Exp
70	84,24
79	62,40
3	7,80
4	1,56

Man får i alla fall en varning.

Category	Historical Observed	Counts	Test Proportion	Expected	Contribution to Chi-Sq
1	70	84,24	0,54	84,24	2,40714
2	79	62,40	0,40	62,40	4,41603
3	3	7,80	0,05	7,80	2,95385
4	4	1,56	0,01	1,56	3,81641

N	DF	Chi-Sq	P-Value
156	3	13,5934	0,004

1 cell(s) (25,00%) with expected value(s) less than 5.

Mendel's experiment

SEED SHAPE	Round	Wrinkled
	A	a

Allel A är dominant

Mendel korsade bara heterozygota 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 runda:ynkliga.

Mendel's Law

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

	rund	ynklig	Σ
O	423	133	556
E	556·¾=417	556·¼=139	556

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

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

Ingen signifikant skillnad mellan observed/expected. Ingen motsats till modellen, dvs. ingen förkastning av Mendels lag genom detta försök.

Hardy-Weinberg-Equilibrium

- Ofta vet man dock inte allel-frekvenserna (p,q) från början som i Mendels experiment (där man hade p=q=0.5 eftersom bara heterozygota bönor används)
- HWE säger dock att allel-frekvenserna förblir konstant

HWE		female	
		A(p)	a(q)
male	A(p)	AA(p^2)	Aa(pq)
	a(q)	Aa(pq)	aa(q^2)

$$f(AA) = p^2$$
$$f(Aa) = 2 \cdot pq$$
$$f(aa) = q^2$$

Exempel från http://en.wikipedia.org/wiki/Hardy-Weinberg_principle, changed!

Skattning av allel-frekvenserna

Callimorpha dominula Linnaeus, 1758

Genotype	White-spotted(AA)	Intermediate (Aa)	Little spotting (aa)	Σ (=n)
Number	1329	268	15	1612

$$p \sim obs(AA) + 0.5 \cdot obs(aA) \quad och \quad q \sim obs(aa) + 0.5 \cdot obs(aA)$$

$$p = \frac{obs(AA) + 0.5 \cdot obs(aA)}{n} = \frac{1329 + 134}{1612} = 0.907$$

$$q = 1 - p = 0.093$$

a

A

Hardy-Weinberg-Equilibrium

Genotype	White-spotted(AA)	Intermediate (Aa)	Little spotting (aa)	Σ (=n)
Number	1329	268	15	1612

Expected:

$p = 0.907 \qquad q = 0.093$

$E(AA) = n \cdot p^2 = 1612 \cdot 0.907^2 = 1326.11$

$E(Aa) = n \cdot 2 \cdot pq = 271.95$

$E(aa) = n \cdot q^2 = 13.94$

Testvariabel:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$
$$= \frac{(1329 - 1326.11)^2}{1326.11} + \frac{(268 - 271.95)^2}{271.95} + \frac{(15 - 13.94)^2}{13.94}$$
$$= 0.0063 + 0.0574 + 0.081$$
$$= 0.144$$

Antalet frihetsgrader och det kritiska värdet

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

Vi har 3 celler, men vi förlorar 1 df för en linjär restriktion och 1 df för en skattning (allmänt för HWE-test: df = antalet genotyper minus antalet alleler)

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

Vi förkastar *inte* nollhypotesen att populationen är i HWE.