BI1363 HT 2020
Analys av kategoridata
Oktober 2020

Adam Flöhr, BT, SLU

Analys av kategoridata

Motsvarar *Biometri*, kap 9

I korthet

Vi undersöker en variabel som ger en indelning i **kategorier**

Observerad data kan sammanfattas i en **frekvenstabell**

Från någon hypotes kan vi beräkna **förväntade värden** för tabellen

Observerade och förväntade värden kan jämföras med ett χ^2**-test**

Testet kan användas för att testa om data följer en viss fördelning (**modellanpassning**)

Och för att testa för samband mellan två kategorivariabler (**homogenitets- och oberoendetest**)

Kategoridata

Undersöker en egenskap där de möjliga utfallen ger en kategori-indelning

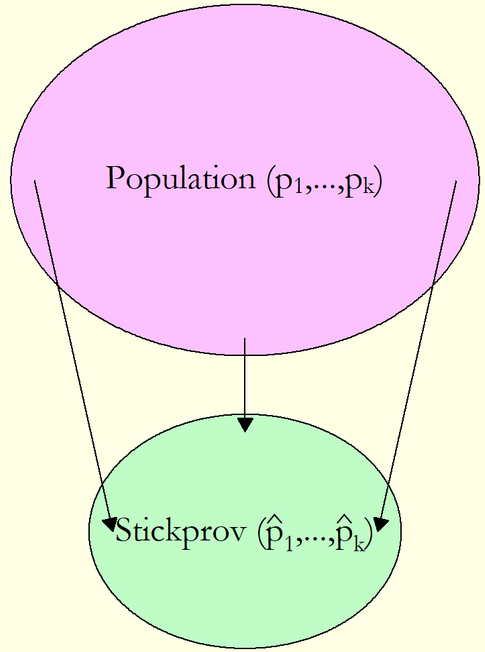
Typiskt en variabel på nominalskala (t.ex art eller nationalitet) eller ordinalskala (t.ex kundnöjdhet)

Vi har $\boldsymbol{k}$ klasser och en klass relativa storlek i populationen ges av $\boldsymbol{p_i}$

Drar ett stickprov av storlek $\boldsymbol{n}$ och tittar på antalet i stickprovet i respektive klass

Kan skatta $\boldsymbol{p}$ för respektive klass med $\hat{\boldsymbol{p}} = \frac{\textbf{antal i klassen}}{\boldsymbol{n}}$

Fallet med binär data är ett särfall där $\boldsymbol{k = 2}$



Frekvenstabeller

En enskild kategorivariabel kan beskrivas med en enkel frekvenstabell

Status	Antal
Frisk	102
Sjuk	198

Två variabel kan beskrivas med en korstabell

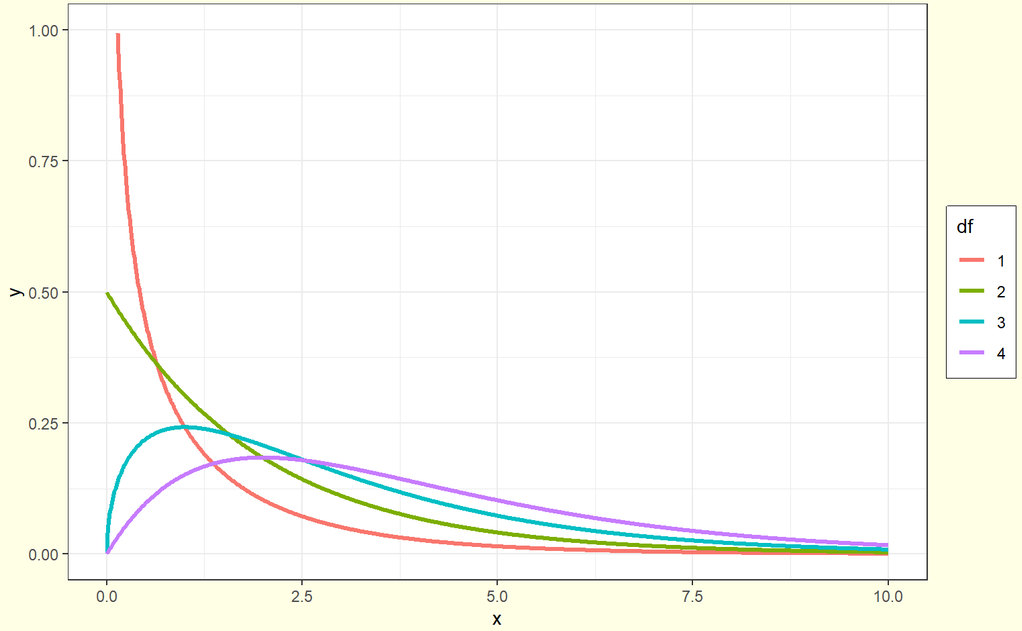
Status \ Läge	Norr	Syd
Frisk	148	50
Sjuk	52	50

χ^2-fördelning

Tester av frekvenser baseras på en testfördelning som kallas en χ^2-fördelning

En χ^2-fördelning uppstår som summan av kvadrerade standardiserade normalfördelningar

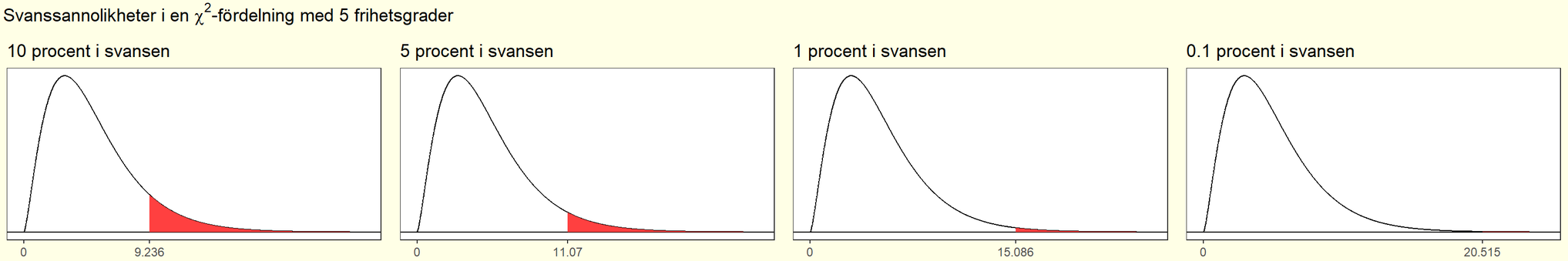
Den defineras av en parameter, antalet *frihetsgrader*, som ges av antalet termer i summan



Biometri, tabell 6

Som tidigare (med normalfördelningen och t-fördelningen) kommer vi vilja uppskatta svanssannolikheten

Eftersom χ^2-fördelningen bygger på kvadrerade värden är vi bara intresserade av den högra svansen



Tabellvärden för x-axeln betecknas $\chi^2_{(1-\alpha,df)}$ och kan hämtas från en tabell över χ^2-fördelningen, t.ex tabell 6 i *Biometri*

	$F(\chi^2)$								
d.f.	0.75	0.9	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995
5	6.626	9.236	11.070	12.833	15.086	16.750	18.386	20.515	22.105

För ett tabellvärde motsvarande fem procent i svansen och $df = 5$ tittar vi på $\chi^2_{(0.95,5)} = 11.070$

Test av modellanpassning

χ^2-testet för modelanpassning används för att testa om observerade frekvenser kommer från en given fördelning

Exempel kan vara

- om alla färger är lika vanliga i en M&M-förpackning
- om en art förekommer lika ofta i flera habitat
- om delar av befolkning förekommer proportionellt i företagsstyrelser

Testet genomförs genom att beräkna *förväntade värden* för varje klass

Förväntade och observerade värden vägs samman med en testfunktion och p-värdet beräknas från en χ^2-fördelning

Antalet frihetsgrader ges av $k - 1$ där k är antalet klasser

Test av modellanpassning, schema

Hypoteser

H_0 : data kommer från den antagna fördelning

H_1 : data kommer inte från den antagna fördelningen

Testfunktion

$$\chi^2 = \sum_{\text{alla klasser}} \frac{(O_i - E_i)^2}{E_i}$$

där O_i och E_i är observerat respektive förväntat antal i klass i

Testfördelning

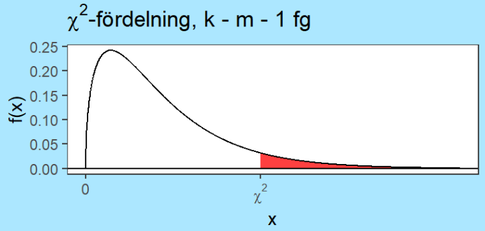
Under nollhypotesen följer χ^2 en χ^2-fördelning med $k - 1$ frihetsgrader, där k är antalet klasser

Förväntade värden E ska vara större än 5

P-värde

P-värdet ges av arean bortom χ^2 i testfördelningen

Vid handräkning uppskattas p-värdet genom att ställa χ^2 mot ett tabellvärde



Svar

P-värdet ställs mot en förbestämd *signifikansnivå* (ofta 5 procent)

Vid ett lågt p-värde förkastas nollhypotesen

Vid ett högt p-värde förkastas ej nollhypotesen

Test av modelanpassning, exempel

I en studie av dagfjärilar fångas hundra fjärilar och sorteras efter familj

Antalen ges av följande

Familj	Tjockhuvuden	Riddarfjärilar	Vitfjärilar	Juvelvingar
O (observerade)	40	23	27	10

Från tidigare studier tror man att fördelningen i området är 40 procent tjockhuvuden och 20 procent var för övriga

Familj	Tjockhuvuden	Riddarfjärilar	Vitfjärilar	Juvelvingar
O (observerade)	40	23	27	10
p (sannolikheter)	0.4	0.2	0.2	0.2

Vi genomför ett test för att se om våra observationer kommer från den etablerade fördelningen

Hypoteser

$\boldsymbol{H_0}$ **:** observerade antal kommer från den tidigare fördelningen

$\boldsymbol{H_1}$ **:** observerade antal kommer inte från den tidigare fördelning

Testfunktion

Förväntade värden

Vi beräknar förväntade värden genom att multiplicera vårt totala antal med sannolikheterna från fördelningen

Familj	**Tjockhuvuden**	**Riddarfjärilar**	**Vitfjärilar**	**Juvelvingar**
O (observerade)	40	23	27	10
p (sannolikheter)	0.4	0.2	0.2	0.2
E (förväntade)	40	20	20	20

Notera särskilt att χ^2-testet alltid beräknas på *antalen*

Kontrollera så att kravet att $E > 5$ är uppfyllt för samtliga klasser

Testfunktionen beräknas med

$$\chi^2 = \sum_{\text{alla klasser}} \frac{(O_i - E_i)^2}{E_i} = \frac{(40-40)^2}{40} + \frac{(23-20)^2}{20} + \frac{(27-20)^2}{20} + \frac{(10-20)^2}{20} = 0 + \frac{9}{20} + \frac{49}{20} + \frac{100}{20} = 7.9$$

Testfördelning

Under nollhypotesen följer χ^2 en χ^2-fördelning

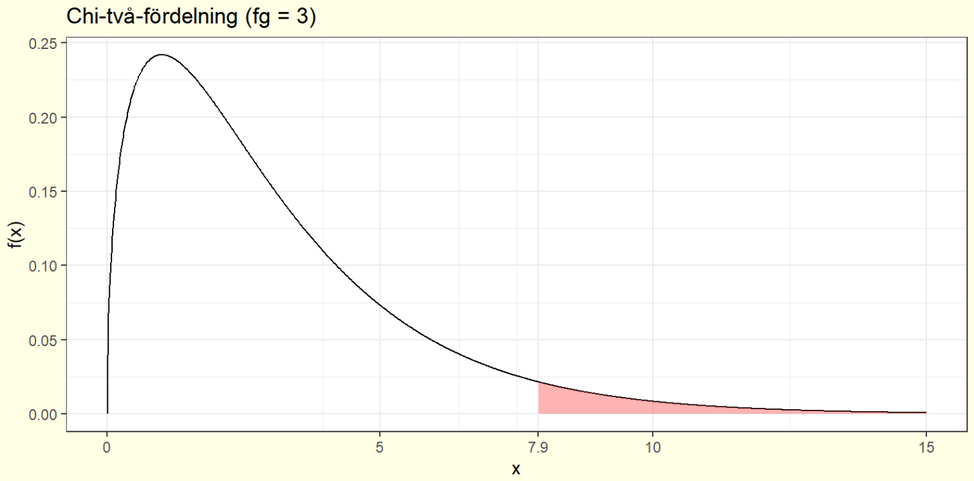
Antalet frihetsgrader ges av $k - 1 = 4 - 1 = 3$

P-värde

P-värdet ges av ytan till höger om vårt observerade χ^2

Vi kan uppskatta p-värdet från *Biometri* tabell 6

	$F(\chi^2)$								
d.f.	0.75	0.9	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995
1	1.323	2.706	3.841	5.024	6.635	7.879	9.141	10.828	12.116
2	2.773	4.605	5.991	7.378	9.210	10.597	11.983	13.816	15.202
3	4.108	6.251	7.815	9.348	11.345	12.838	14.320	16.266	17.730



Vårt observerade värde ligger precis över $\chi^2_{(0.95,3)} = 7.815$. P-värdet måste alltså vara strax under fem procent

En datorberäkning ger det exakta värdet 0.04812

Slutsats

Det finns en statistiskt signifikant skillnad mellan våra observerade värden och den tidigare fördelningen

Test av samband i en korstabell

χ^2-test kan också användas för att testa om det finns något samband mellan två kategorivariabler

Observerade värden presenteras i en korstabell

Marginalsummor hålls konstanta och förväntade värden beräknas som om variablerna vore oberoende

Testfunktionen är densamma som tidigare

Antalet frihetsgrader ges av $(r-1)(k-1)$ där r är antalet rader och k antalet kolumner i korstabellen

Test av korstabell, exempel

I en fortsättning på vår fjärilstudie besöker vi tre olika områden och samlar in populationsdata

Vi vill undersöka om andelen vitfjärilar är densamma oberoende av område

Art \ Område	Område A	Område B	Område C	Summa
Vitfjäril	9	18	19	46
Annan art	41	32	81	154
Summa	50	50	100	200

Hypoteser

H_0 : andelen vitfjärilar är densamma

H_1 : andelen vitfjärilar skiljer sig mellan områden

Testfunktion

Förväntade antal

Förväntade värden beräknas genom att hålla marginalsummor konstanta och beräkna *inre celler* enligt formeln

$$E_{ij} = \frac{\textbf{Radsumma i} \cdot \textbf{Kolumnsumma j}}{\textbf{Totalsumma}}$$

Det första förväntade värdet ges av

$$E_{11} = \frac{50 \cdot 46}{200} = \frac{2300}{200} = 11.5$$

Observerade värden och förväntade värden

Om vi gör beräkningen för förväntade värden för varje cell får vi följande tabeller

Observerade värden (O)

Art \ Område	Område A	Område B	Område C	Summa
Vitfjäril	9	18	19	46
Annan art	41	32	81	154
Summa	50	50	100	200

Förväntade värden (E)

Art \ Område	Område A	Område B	Område C	Summa
Vitfjäril	11.5	11.5	23	46
Annan art	38.5	38.5	77	154
Summa	50	50	100	200

Marginalsummorna ska vara desamma

Förväntade värden $\boldsymbol{E}$ behöver inte vara heltal

Som tidigare gäller tumregeln $\boldsymbol{E > 5}$ för att $\boldsymbol{\chi^2}$-testet ska vara lämpligt

Testfunktion, beräkning

Testfunktionen är densamma som tidigare. Summan går nu över samtliga celler i korstabellen

$$\chi^2 = \sum_{\text{alla celler}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(9 - 11.5)^2}{11.5} + \cdots + \frac{(77 - 81)^2}{77} = 6.381$$

Testfördelning

Under nollhypotesen följer χ^2 en χ^2-fördelning med $(r-1)(k-1)$ frihetsgrader, där r är antalet rader och k antalet kolumner i korstabellen

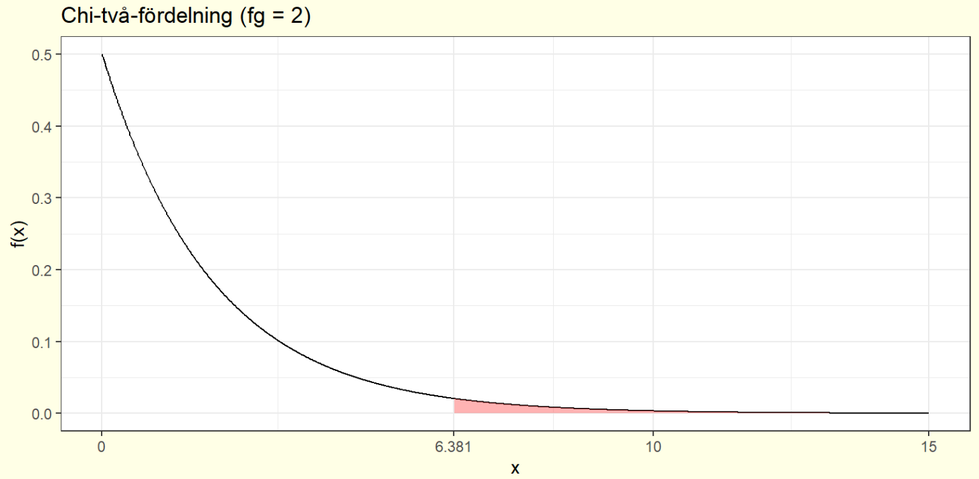
I vårt exempel har vi $df = (3-1)(2-1) = 2$

P-värde

p-värdet ges av ytan till höger om vårt observerade χ^2

Vi kan uppskatta p-värdet från *Biometri* tabell 6

	$F(\chi^2)$								
d.f.	0.75	0.9	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995
1	1.323	2.706	3.841	5.024	6.635	7.879	9.141	10.828	12.116
2	2.773	4.605	5.991	7.378	9.210	10.597	11.983	13.816	15.202



Vårt observerade värde ligger över $\chi^2_{(0.95,2)} = \mathbf{5.991}$. P-värdet måste alltså vara strax under fem procent

En datorberäkning ger det exakta värdet $\mathbf{0.04116}$

Slutsats

Vi förkastar nollhypotesen

Det finns en signifikant skillnad i andelen vitfjärilar mellan områden

Homogenitetstest eller oberoendetest

Boken *Biometri* delar upp tester på korstabeller i två typer

Indelningen har ingen påverkan på hur testet genomförs

Homogenitetstest

Vid ett *homogenitetstest* är antalen i en av variablerna fixerad och man är intresserad av skillnader inom den andra variabeln

- Exemplet med fjärilar i tre områden är ett homogenitetstest, eftersom vi valt att samla in femtio, femtio och hundra fjärilar per område
- Man kan redan innan man samlar in sin data säga vad kolumnsummorna blev

Oberoendetest

Vid ett *oberoendetest* är bägge variablerna slumpmässiga

Vi testar om det finns något samband mellan variablerna

- Till exempel att man samlar in 500 ekorrar och noterar kön och ålder
- Marginalsummor kan ej uppskattas innan data samlats in

Slut