
18. χ^2 METODEN

18.1 Inledning

I detta kapitel diskuterar vi något om analys av variabler som har en ordinal- eller nominalskala. De metoder för hypotestest vi diskuterade i föregående kapitel är inte alltid användbara då variablerna har en ordinal- eller nominalskala. Därför måste vi utveckla några nya metoder för dessa situationer. χ^2 -metoden är en metod som har utvecklats för att analysera data då variabler har ordinal- eller nominalskala.

χ^2 -metoden kan användas för prövning av hypoteser i flera olika slag av problem. Vi skall här först beskriva hur man prövar om en stokastisk variabel följer en viss sannolikhetsfördelning med kända eller okända parametrar. Sedan skall vi beskriva hur man prövar om två stokastiska variabler har samma fördelning och slutligen om två stokastiska variabler är oberoende av varandra.

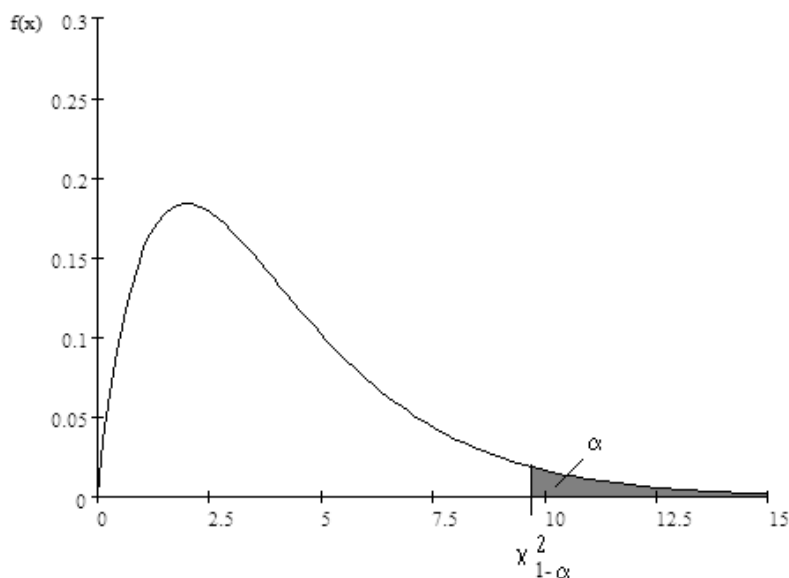
De värden som en variabel med ordinal- eller nominalskala kan anta är ett antal kategorier. T ex kanske variabeln "bedömning av nästa års konjunktur" har kategorierna "högkonjunktur", "varken låg eller hög konjunktur" och "lågkonjunktur". Ett annat exempel är variabeln "färdssätt till arbetet" som kan ha kategorierna "gå", "cykla", "åka buss" och "åka bil". När vi gör observationer noterar vi i vilken kategori observationerna faller och vi räknar hur många observationer vi får i varje kategori. Låt n_i beteckna den observerade frekvensen i den i -te kategorin. Om vi t ex har observerat 18 individer som tror på en högkonjunktur nästa år är den observerade frekvensen 18 för kategorin "högkonjunktur".

χ^2 -metoden bygger alltid på jämförelser mellan observerade frekvenser, n_i , och sk förväntade frekvenser, $E[n_i]$, för variablernas kategorier. De förväntade frekvenserna är de frekvenser vi kan förvänta oss om den hypotes vi vill pröva är sann.

Om hypotesen är sann bör den observerade och den förväntade frekvensen vara ungefär lika, dvs $(n_i - E[n_i])^2$ bör vara liten för varje kategori. Det samma gäller om vi bildar relativtal, $(n_i - E[n_i])^2 / E[n_i]$. Om vi beräknar detta relativtal för varje kategori och summerar får vi

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - E[n_i])^2}{E[n_i]}$$

där k är antalet kategorier. Eftersom n_i är en stokastisk variabel, kommer även χ^2 att vara det. Från avsnitt 14.3 vet vi att den stokastiska variabeln χ^2 är approximativt χ^2 -fördelad. Approximationen är i allmänhet tillräckligt god om alla förväntade frekvenser är stora. Som tumregel brukar man kräva att $E[n_i] \geq 5$ för alla klasser. Om detta inte är uppfyllt kan vi slå ihop intilliggande kategorier med små förväntade frekvenser till kategorier med större förväntade frekvenser.



Figur 18.1 χ^2 -fördelningen och det kritiska värdet $\chi^2_{1-\alpha}$ som bestäms så att $P(\chi^2 > \chi^2_{1-\alpha} \mid H_0 \text{ är sann}) = \alpha$

χ^2 -fördelningen har en parameter, ν . Parametern kallas frihetsgradtal. Hur många frihetsgrader den stokastiska variabeln χ^2 har bestäms av bl a antalet kategorier. Vi återkommer till hur antalet frihetsgrader bestäms vid de olika tillämpningarna.

Om hypotesen är sann bör det observerade värdet på χ^2 -variabeln vara litet. Ett stort värde på χ^2 tyder alltså på att hypotesen är falsk. Antag att vi prövar på signifikansnivån α , dvs sannolikheten att förkasta en sann hypotes är α . Då förkastar vi hypotesen om det observerade χ^2 -värdet är större än det tal som har markerats med $\chi^2_{1-\alpha}$ i figur 18.1. Alternativt kan vi på motsvarande sätt som tidigare beräkna ett p -värde. Om c är det observerade χ^2 -värdet är alltså

$$p = P(\chi^2 > c).$$

18.2 Goodness of fit-test

I detta avsnitt beskrivs hur en hypotes om att en stokastisk variabel är fördelad enligt en viss given fördelning prövas med χ^2 -metoden. Testet bygger på att jämföra hur väl en given teoretisk fördelning och en observerad fördelning överensstämmer med varandra. För att mäta avvikelsen mellan de två fördelningarna använder vi χ^2 -metoden. Test av detta slag brukar benämnas goodness of fit-test.

Exempel 18.1

Antag att vi vill undersöka om en tärning är symmetrisk eller ej. Tärningen kastas 120 ggr. Vår hypotes är att varje sida har sannolikheten $1/6$ att komma upp när tärningen kastas. Följaktligen förväntar vi oss att varje sida kommer upp $\frac{1}{6} \cdot 120 = 20$ gånger då tärningen kastas 120 gånger.

Antalet frihetsgrader är lika med antalet kategorier minus ett. Vi har sex kategorier och får därför $6 - 1 = 5$ frihetsgrader. På signifikansnivån $\alpha = 0,05$ får vi ur tabell att $\chi^2_{0,95}(5) = 11,1$ dvs vi förkastar hypotesen om det observerade χ^2 -värdet är större än 11,1.

Antag att 18 kast resulterade i att sidan med en prick kom upp, 23 kast med sidan med två prickar kom upp osv. Vi får då följande beräkningar och resultat:

Klass	n_i	$E[n_i]$	$(n_i - E[n_i])^2$	$\frac{(n_i - E[n_i])^2}{E[n_i]}$
1	18	20	4	4/20
2	23	20	9	9/20
3	16	20	16	16/20
4	21	20	1	1/20
5	18	20	4	4/20
6	24	20	16	16/20
Summa	120	120		50/20=2,5

Eftersom $\chi^2 = 2,5$ är mindre än 11,1 kan vi inte förkasta hypotesen, dvs hypotesprövningen har inte kunnat påvisa att tärningen skulle vara skev. ■

I exempel 18.1 är parametrarna i sannolikhetsfördelningen kända (likformig fördelning med $p = 1/6$). I vissa fall är parametrarna inte kända utan måste först skattas. Antalet frihetsgrader måste då reduceras med antal skattade parametrar. Om vi behöver skatta m stycken parametrar blir alltså antalet frihetsgrader

$$\nu = k - 1 - m$$

där k är antalet kategorier hos den stokastiska variabeln.

Övning 18.1

En livsmedelskedja genomförde en undersökning av kundernas preferenser för olika märken av bröd. 200 kunder fick prova fyra olika märken av vitt matbröd och ange vilken de tyckte bäst om. Följande frekvenser observerades:

Märke	A	B	C	D
frekvens	28	41	81	50

Pröva hypotesen att kunderna föredrar alla märken lika mycket. Använd χ^2 -metoden och signifikansnivån 0,05.

Övning 18.2

En viss typ av flugor kan dels ha gul eller svart kropp och dels ha normala eller korta vingar. Enligt genetisk teori förekommer dessa varianter i proportionerna 9:3:3:1, dvs av 16 flugor har i genomsnitt 9 gul kropp med normala vingar, 3 har gul kropp med korta vingar, 3 har svart kropp med normala vingar och 1 har svart kropp med korta vingar. 100 flugor har fångats och undersökts, varvid man observerade frekvenserna 57, 20, 12 respektive 11.

- a) Undersök med hjälp av χ^2 –metoden om observationerna stöder den genetiska modellen.
- b) Vad skulle resultatet bli om man i stället hade undersökt 200 flugor med dubbelt så höga frekvenser i varje kategori, dvs 114, 40, 24 respektive 22? Förklara eventuella skillnader i slutsatser från uppgift a.

Övning 18.3

Enligt en hypotes åker hälften av alla anställda vid ett företag bil till och från arbetet. Av de övriga åker $2/3$ kollektivtrafik och $1/3$ cyklar eller går till och från arbetet. För att pröva denna hypotes undersöktes ett slumpmässigt urval om $n = 50$ anställda vid företaget. Det visade sig då att 22 anställda åkte bil och 10 åkte kollektivtrafik. Genomför provningen av hypotesen. Använd signifikansnivån 0,05.

Övning 18.4

Enligt en teori är könet hos ett barn oberoende av könet hos eventuella äldre syskon. Om oberoende råder kommer, som bekant, antal flickor i en slumpmässigt vald familj att vara en binomialfördelad stokastisk variabel. För att testa teorin om oberoende undersöktes antal pojkar och antal flickor hos 150 fyrabarnsfamiljer. Resultatet blev

1.	Antal flickor	0	1	2	3	4
	Antal familjer	19	37	41	38	15

Pröva teorin om oberoende med hjälp av ett χ^2 -test. För att förenkla beräkningarna kan du utgå från att andelen pojkar bland alla födselar är 0,5. Använd signifikansnivån 5%.

Övning 18.5

Vid en industri mättes förslitningen av en viss maskindel på 100 maskiner av samma slag. En frekvenstabell med förslitningen mätt i mm fick nedanstående utseende:

Förslitning (mm)	Antal maskiner
- 1,275	0
1,275 - 1,325	1
1,325 - 1,375	6
1,375 - 1,425	9
1,425 - 1,475	15
1,475 - 1,525	13
1,525 - 1,575	22
1,575 - 1,625	19
1,625 - 1,675	9
1,675 - 1,725	2
1,725 - 1,775	4
1,775 -	0

Pröva hypotesen om att förslitningen kan beskrivas med en normalfördelad stokastisk variabel enligt följande arbetsgång:

- Bestäm medelvärde och standardavvikelse (använd två decimaler) i det klassindelade materialet.
- Transformera observationsmaterialet så att det får medelvärdet 0 och standardavvikelsen 1.
- Beräkna förväntade frekvenser.
- Slå ihop klasser om det förväntade antalet observationer i någon klass är för litet (mindre än 5).
- Bestäm antalet frihetsgrader ν och vid vilket värde på χ^2 som hypotesen förkastas.
- Beräkna det observerade χ^2 -värdet.
- Drag slutsats.

18.3 Test av homogenitet

I föregående avsnitt beskrevs ett test för att avgöra om en stokastisk variabel har en viss specificerad sannolikhetsfördelning. I detta avsnitt beskriver vi ett test för att avgöra om två eller flera stokastiska variabler har samma sannolikhetsfördelning. Här specificerar vi dock inte vilken sannolikhetsfördelning de har. Liksom i föregående avsnitt använder vi χ^2 -måttet för att mäta hur stor skillnaden mellan fördelningarna är. Test av det här slaget brukar kallas homogenitetstest, dvs vi testar om två eller flera populationer är homogena vad avser sannolikhetsfördelning.

Exempel 18.2

En undersökning av folks ätvanor när de ser på TV genomfördes i fem länder, Sverige, Italien, England, USA och Kina. Tabellen nedan visar resultaten på frågan om man brukar äta chips, nötter eller godis när man tittar på TV.

	Sverige	Italien	England	USA	Kina	Totalt
Ofta	512	395	685	743	532	2867
Ibland	125	128	130	131	184	698
Sällan	212	404	32	71	181	900
Totalt	849	927	847	945	897	4465

Vi vill nu undersöka om fördelningen över svarskategorierna kan vara densamma i de olika länderna, dvs om fördelningen är homogen. Totalt för de fem länderna var det 2867 som svarade "ofta". Detta är $2867/4465 = 0,6521$ av alla svar. Om det inte är någon skillnad mellan länderna bör alltså andelen som svarat "ofta" vara ungefär 64,21 %. För Sverige innebär det att vi förväntar oss att $0,6421 \cdot 849 = 545,14$ svarar "ofta". För Italien blir motsvarande förväntade frekvens $0,6421 \cdot 927 = 595,23$, osv. På samma sätt är andelen som svarat "ibland" lika med $698/4465 = 0,1563$ oavsett land. Om det är samma fördelning i alla länder förväntar vi oss då att $0,1563 \cdot 849 = 132,7$ svarar "ibland" i Sverige, $0,1563 \cdot 927 = 144,9$ svarar "ibland" i Italien osv. Under hypotesen att fördelningen mellan svarsalternativen är densamma i de fem länderna kan vi beräkna förväntade frekvenser på detta sätt. De resulterade förväntade frekvenserna sammanfattas i följande tabell:

	Sverige	Italien	England	USA	Kina	Totalt
Ofta	545,1	595,2	543,9	606,8	576,0	2867
Ibland	132,7	144,9	132,4	147,7	140,2	698
Sällan	171,1	186,9	170,7	190,5	180,8	900
Totalt	849	927	847	945	897	4465

Nästa steg i analysen är att jämföra de observerade frekvenserna med de förväntade frekvenserna med hjälp av χ^2 -måttet. Vi får då att

$$\chi^2 = \frac{(512 - 545,1)^2}{545,1} + \frac{(395 - 595,2)^2}{595,2} + \dots + \frac{(181 - 180,8)^2}{180,8} = 607,742$$

För att avgöra om detta är ett så stort värde så att hypotesen om lika fördelning kan förkastas, behöver vi först bestämma antalet frihetsgrader. Vi ser då att varje land har tre svarskategorier men har också en restriktion, nämligen att summan av frekvenserna är given. Vet vi t ex frekvenserna för svarskategorierna "ofta" och "ibland" kan vi alltid beräkna frekvensen för svarsalternativet "sällan". På motsvarande sätt har varje svarsalternativ restriktionen att summan av frekvenserna är given. Vet vi frekvenserna för svarsalternativet "ofta" för fyra av länderna kan vi alltid räkna ut frekvensen för det femte landet. Detta gör att antalet frihetsgrader är

$$\nu = (\text{antal kategorier} - 1) \cdot (\text{antal populationer} - 1).$$

I detta exempel har vi alltså att $\nu = (3 - 1) \cdot (5 - 1) = 2 \cdot 4 = 8$. Kritiskt värde i χ^2 -fördelningen med 8 frihetsgrader och signifikansnivån $\alpha = 0,05$ är 15,507. Vårt observerade χ^2 -värde, 607,742, överstiger det kritiska värdet avsevärt och vi kan förkasta hypotesen om lika fördelning mellan länderna.

När en hypotes förkastas är det ofta intressant att se var det är stora skillnader mellan observerade frekvenser och förväntade frekvenser. För att studera detta kan vi beräkna residualer, dvs skillnader mellan observerade och förväntade frekvenser. För att få skillnaderna på en gemensam skala dividerar vi dem med sina respektive standardavvikelser. I detta fall vet vi att standardavvikelsen är kvadratroten ur den förväntade frekvensen. För svarsalternativet "ofta" blir alltså residualen för Sverige

$$\frac{512 - 545,1}{\sqrt{545,1}} = -1,4.$$

Dessa tal, som kallas standardiserade residualer eller Pearson residualer, sammanfattas i nedanstående tabell. Observera att kvadraten på de standardiserade residualerna är precis termerna som ingår i χ^2 -värdet.

	Sverige	Italien	England	USA	Kina
Ofta	-1,4	- 8,2	6,1	5,5	-1,8
Ibland	- 0,7	- 1,4	- 0,2	-1,4	3,7
Sällan	3,1	15,9	-10,6	-8,7	0,0

En analys av de standardiserade residualerna visar på att svarsalternativet "sällan" förekommer i långt högre frekvens Italien medan det förekommer med betydligt lägre frekvenser i England och USA jämfört med genomsnittet i undersökningen. ■

Övning 18.6

En butikskedja önskar undersöka i vilken utsträckning kunderna använder butikskedjans rabattkuponger. Man har därför valt ut tre butiker med ungefär samma försäljningsvolym och tagit ett slumpmässigt urval om 200 kunder från vardera butiken. Följande resultat erhöles:

Använder rabattkuponger	Butik A	Butik B	Butik C
Ofta	65	97	55
Ibland	98	89	77
Sällan	37	14	68
Totalt	200	200	200

Pröva hypotesen att fördelningen av hur ofta rabattkupongen används är lika för de tre butikerna.

Övning 18.7

Inför ett riksdagsval håller en politiker på att formulera en valstrategi. Som ett led i detta arbete behöver hon veta om hon har lika stort stöd i alla delar av valdistriktet. Ett slumpmässigt utval gav följande data:

	Östra	Norra	Västra	Södra
1. Antal som sympatiserar	21	29	23	16
Antal som inte sympatiserar	39	41	27	24

Kan politiern dra slutsatsen att hon är lika populär i alla delar av valdistriktet? Använd signifikansnivån 0,05.

Övning 18.8

För att studera könsskillnader i konsumentbeteende fick 100 slumpmässigt utvalda kvinnor och 100 slumpmässigt utvalda män svara på frågan om vilken av fyra nämnda butiker de helst besöker. Följande resultat erhöles:

	Clas Ohlson	H & M	Teknikmagasinet	MQ
Kvinnor	11	57	6	26
Män	46	4	32	18

Kan man dra slutsatsen att butikerna är lika attraktiva för kvinnor som för män. Använd signifikansnivån 0,05.

18.4 Prövning av hypoteser om oberoende

Vi skall nu ge ett exempel på hur en hypotes om att två stokastiska variabler är oberoende prövas med χ^2 -metoden.

Exempel 18.3

Man vill undersöka ett eventuellt samband mellan variablerna "civilstånd" och "typ av bostad". Ett urval om $n = 100$ individer togs ut och sammanställdes i följande tabell

Typ av bostad	Gift	Ogift	Frånskild	Totalt
Villa	25	5	10	40
Hyseslägenhet	15	25	20	60
Totalt	40	30	30	100

Från sannolikhetsteorin vet vi att dessa två diskreta stokastiska variabler är oberoende om och endast om sannolikheten

$$\begin{aligned}
 & P(\text{"civilstånd"} = x \cap \text{"typ av bostad"} = y) \\
 &= P(\text{"civilstånd"} = x) \cdot P(\text{"typ av bostad"} = y)
 \end{aligned}$$

där x = gift, ogift eller fränskild och y = villa eller hyreslägenhet.

Det betyder att om variablerna är oberoende förväntar vi oss att antalet gifta i villa är

$$\frac{40}{100} \cdot \frac{40}{100} \cdot 100 = \frac{40 \cdot 40}{100} = 16$$

På motsvarande sätt kan vi beräkna de övriga förväntade frekvenserna.

Typ av bostad	Gift	Ogift	Fränskild	Totalt
Villa	16	12	12	40
Hyreslägenhet	24	18	18	60
Totalt	40	30	30	100

Observera att förväntat antal är större än 5 i samtliga klasser. Antalet frihetsgrader beräknas genom

$$\nu = (\text{antal rader} - 1) \cdot (\text{antal kolumner} - 1)$$

Eftersom det är två rader och tre kolumner får vi $(2 - 1) \cdot (3 - 1) = 1 \cdot 2 = 2$ frihetsgrader. Ur tabell får vi att $\chi^2_{0,95}(2) = 5,99$, dvs om vi testar på signifikansnivån 0,05, måste vi förkasta hypotesen om det observerade χ^2 -värdet är större än 5,99.

Vi beräknar nu χ^2 -värdet:

Klass	n_i	$E[n_i]$	$(n_i - E[n_i])^2$	$\frac{(n_i - E[n_i])^2}{E[n_i]}$
v.g.	25	16	81	5,06
v.o.	5	12	49	4,08
v.f.	10	12	4	0,33
h.g.	15	24	81	3,38
h.o.	25	18	49	2,72
h.f.	20	18	4	0,22
	100	100		15,79

Eftersom det observerade χ^2 -värdet, 15,79, överstiger det kritiska värdet 5,99 kan vi förkasta hypotesen om att det föreligger oberoende mellan variablerna "civilstånd" och "typ av bostad". ■

Övning 18.9

För att pröva om det finns ett samband mellan en individs skolutbildning och inställning till äktenskapet utfrågades 400 personer. Man erhöll följande resultat:

Utbildning	Inställning till äktenskap				Totalt
	Mycket dålig	Dålig	God	Mycket god	
Grundskola	18	29	70	115	232
Gymnasium	17	28	30	41	116
Universitet	11	10	11	20	52
Totalt	46	67	111	176	400

Pröva hypotesen att det inte föreligger något samband mellan en individs utbildning och inställning till äktenskapet. Välj signifikansnivån till 0,05.

Övning 18.10

För att undersöka inställningen till rökning gjordes ett urval om 200 personer, vilka vardera fick besvara två frågor:

1. Tror ni att rökning är skadligt för hälsan?
2. Röker ni själv?

Undersökningen gav till resultat att

80 st svarade "ja" på båda frågorna

15 st svarade "nej" på båda frågorna

80 st svarade "nej" på första frågan och "ja" på andra frågan

25 st svarade "ja" på första frågan och "nej" på andra frågan

Man vill undersöka om uppfattningen om den första frågan är oberoende av om man är rökare eller ej. Pröva hypotesen om oberoende på signifikansnivån 0,05.