Några exempel på χ^2-test

Oberoendetest

Ett stickprov omfattande 300 personer samlades in för att undersöka om hårfärgen är oberoende av personens kön. Som vanligt formuleras noll- och alternativ hypotes och en lämplig testvariabel väljs.

H_0 : Hårfärgen är oberoende av kön i den population stickprovet drogs från.

H_1 : Hårfärgen är inte oberoende av kön i den population stickprovet drogs från.

De insamlade data sammanfattas i följande korstabell (2×4):

	Hårfärg				
Kön	Svart	Brun	Blond	Röd	Summa
Man	32	43	16	9	100
Kvinna	55	65	64	16	200
Summa	87	108	80	26	300

De förväntade frekvenserna (e_{ij}) beräknas på följande sätt:

$$e_{ij} = \frac{Rad_i}{n} \cdot \frac{Kol_j}{n} \cdot n = \frac{Rad_i \cdot Kol_j}{n}.$$

Vi kompletterar tabellen ovan med de beräknade förväntade frekvenserna i röd text.

	Hårfärg				
Kön	Svart	Brun	Blond	Röd	Summa
Man	32	43	16	9	100
	29,000	36,000	26,667	8,333	
Kvinna	55	65	64	16	200
	58,000	72,000	53,333	16,667	
Summa	87	108	80	26	300

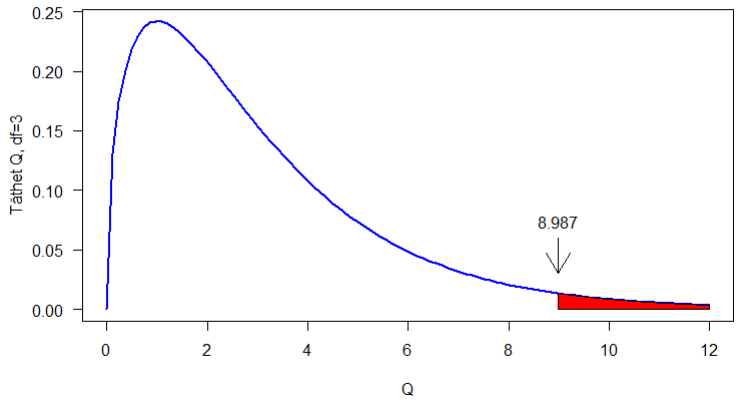
Nu kan vi gå vidare och beräkna det observerade värdet på testvariabeln. Vi beteckna observerade frekvenser med o_{ij} och de förväntade med e_{ij}. Testvariabeln är

$$Q = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

och den är approximativt χ^2-fördelad om nollhypotesen är sann. Med aktuella observerade och förväntade frekvenser blir det

$$\begin{aligned} Q &= \frac{(32-29,000)^2}{29,000} + \frac{(43-36,000)^2}{36,000} + \frac{(16-26,667)^2}{26,667} + \frac{(9-8,333)^2}{8,333} \\ &= \frac{(55-58,000)^2}{58,000} + \frac{(65-72,000)^2}{72,000} + \frac{(64-53,333)^2}{53,333} + \frac{(16-16,667)^2}{16,667} \\ &= 8,987 \end{aligned}$$

och vi behöver antalet frihetsgrader. Det får vi med $df = (r-1)(k-1) = 1 \cdot 3 = 3$. Grafen nedan visar täthetsfunktionen med 3 frihetsgrader.



Nu kan vi räkna ut sannolikheten för att få den observerade värdet på testvariabeln eller något extremare, dvs. det rödmarkerade området i grafen (linjen inkluderad). Med hjälp av R får vi den önskade sannolikheten som

$$1 - \text{pchisq}(8.987, 3) = 0.02946423.$$

Funktionen pchisq() ger sannolikheten för värden från och med 0 till och med det aktuella värdet (8.987). Därför använder vi $1 - \text{pchisq}(8.987, 3)$. Den beräknade sannolikheten är mindre än 0.05 och vi förkastar nollhypotesen.

I R kan vi bygga en matris av observerade data för de båda variablerna hårfärg och kön. Exempelvis på följande sätt

```
r1 <- c(32,43,16,9)
r2 <- c(55,65,64,16)
m <- rbind(r1,r2)
```

Testet genomförs med matrisen

```
tst <- chisq.test(m)
```

Genom att lagra testresultatet i variabeln tst blir det lättare att extrahera information. Vi får den viktigaste informationen med

```
tst
Pearson's Chi-squared test
data: m X-squared = 8.9872, df = 3, p-value = 0.02946
```

Här motsvarar X-squared vår testvariabel Q. Vi kan se att värdena överensstämmer med dem vi räknade fram ovan och vi drar samma slutsats. Om vi vill kontrollera att de observerade värdena stämmer gör vi

```
tst$observed
> tst$observed
     [,1] [,2] [,3] [,4]
r1     32   43   16    9
r2     55   65   64   16
```

De förväntade frekvserna får vi analogt

```
tst$expected
> tst$expected
```

	[,1]	[,2]	[,3]	[,4]
r1	29	36	26.66667	8,333333
r2	58	72	53.33333	16.666667

Homogenitetstest

I en undersökning av biverkningar av medicin mot säsongsallergi rekryterades 3036 personer i fyra grupper som använt Claritin-D, Loratadine, Pseudoephedrine eller placebo. Observerade frekvenser framgår av tabellen nedan. De båda hypoteserna blir

H_0 : Fördelningen av insomnia är lika för varje behandlingsgrupp.

H_1 : Fördelningen av insomnia är inte lika för varje behandlingsgrupp.

	Allergimedicin				
Insomnia	Claritin-D	Loratadine	Pseudoephedrine	Placebo	Summa
Ja	164	22	104	28	318
Nej	859	521	444	894	2718
Summa	1023	543	548	922	3036

Med förväntade frekvenser, beräknade som ovan, får vi följande tabell. De förväntade frekvenserna anges med röda siffror.

	Allergimedicin				
Insomnia	Claritin-D	Loratadine	Pseudoephedrine	Placebo	Summa
Ja	164	22	104	28	318
	107,152	56,875	57,399	96,573	
Nej	859	521	444	894	2718
	915,848	486,125	490,601	825,427	
Summa	1023	543	548	922	3036

Vi kan nu beräkna testvariabelns observerade värde på samma sätt som ovan och vi får $Q = 154,224$. Med hjälp av R räknar vi ut sannolikheten för att få det observerade värdet eller något extremare

$$1 - \text{pchisq}(154,224,3) = 0.$$

Nu undrar vi är det beräknade p-värdet exakt lika med 0 eller är det något mycket litet? Vi kan ta ytterligare hjälp från R och extrahera p-värdet ur tst, jämför ovan.

> tst1$p.value

[1] 3.232003e-33,

vilket är detsamma som $3,232003 \cdot 10^{-33}$ och detta är ett mycket litet tal men inte exakt lika med 0. Vi kan utan vidare förkasta nollhypotesen.

Hela hypotesprövningen kan förstås genomföras med R och det ser då ut som följer.

```
ra1 <- c(164,22,104,28)
ra2 <- c(859,521,444,894)
m2 <- rbind(ra1,ra2)
m2
tst1 <- chisq.test(m2)
tst1
```

Pearson's Chi-squared test

data: m2 X-squared = 154.22, df = 3, p-value < 2.2e-16.

Vi får inte ett exakt p-värde utan utsagan att p-värdet är mindre än $2,2 \cdot 10^{-16}$ och det räcker förstås för att förkasta nollhypotesen. Extraktionen av p-värdet innan förutsätter att vi genomfört hypotesprövningen i R.

Anpassningstest

I en berömd undersökning studerade den ryske ekonomen Ladislaus Bortkiewicz hur många soldater som hade dödats genom spark av häst vid fjorton preussiska armékårer under 20 år (1875-1894). För varje år och armékår noterades antalet omkomna. De 281 obsservationerna blev:

Antal döda	0	1	2	3	4
Frevens	144	91	33	11	2

Sannolikheten för en viss soldat att bli ihjälsparkad av häst var förhoppningsvis liten även vid en preussisk armékår, men det var många som löpte denna risk. Antalet dödade kan således på goda grunder antas kunna beskrivas av en Poissonfördelning. Våra hypoteser blir:

H_0 : variabeln är Poissonfördelad

H_1 : variabeln är inte Poissonfördelad

Poissonfördelningen har en parameter μ som också är fördelningens väntevärde. Lägg märke till att vi inte angett något numeriskt värde på denna parameter i nollhypotesen. För att fortsätta måste vi skatta μ från stickprovet. Vi vet att den bästa skattningen är stickprovets medelvärde $\overline{x}$, som beräknas

$$\overline{x} = \frac{144 \cdot 0 + 91 \cdot 1 + 33 \cdot 2 + 11 \cdot 3 + 2 \cdot 4}{281} = \frac{198}{281} = 0,70.$$

Sannolikheterna för olika utfall beräknas med hjälp av

$$Pr\left(X = x\right) = \frac{\mu^x}{x!} \cdot \mathrm{e}^{-\mu},\ x = 0, 1, 2, \ldots$$

I detta uttryck sätter vi in $x = 0, 1, 2, 3$ och får de skattade sannolikheterna för motsvarande värden. $Pr\left(X \geq 4\right)$ får vi som

$$Pr\left(X \geq 4\right) = 1 - \left(Pr\left(X = 0\right) + Pr\left(X = 1\right) + Pr\left(X = 2\right) + Pr\left(X = 3\right)\right).$$

De skattade sannolikheterna multipliceras med 281. Då får vi de förväntade frekvenserna.

Antal döda	Observerad frekvens (o)	Skattad sannolikhet	Förväntad frekvens (e)
0	144	0,497	139,66
1	91	0,348	97,79
2	33	0,122	34,28
3	11	0,028	7,87
4	2	0,005	1,40
Summa	281	1,000	281,00

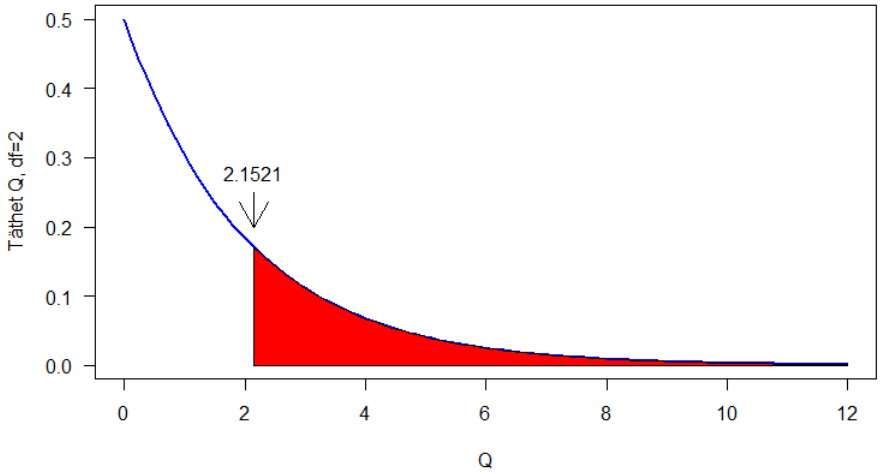
Nu kan vi räkna ut testvariabelns observerade värde och beräkna sannolikheten för att få det observerade värdet eller något extremare, men vi går direkt på hypotesprövningen i R.

```
ob <- c(144,91,33,11,2) # Observerade frekvenser
sa <- c(0.497,0.348,0.122,0.028,0.005) # Sannolikheter för utfallen (antal döda)
tst <- chisq.test(ob, p = sa)
tst
tst$expected
Chi-squared test for given probabilities
data: ob
X-squared = 2.1529, df = 4, p-value = 0.7077
> tst$expected
[1] 139.657 97.788 34.282 7.868 1.405
```

Vi ser här att den sista klassen har en förväntad frekvens av 1,405 som är mindre än 5. Då är inte resultaten helt pålitliga och vi åtgärdar genom att slå ihop de två sista klasserna till en och gör om analysen.

```
ob1 <- c(144,91,33,13)
sa1 <- c(0.497,0.348,0.122,0.033)
tst <- chisq.test(ob1, p = sa1)
tst
tst$expected
Chi-squared test for given probabilities
data: ob1
X-squared = 2.1521, df = 3, p-value = 0.5414
> tst$expected
[1] 139.657 97.788 34.282 9.273
```

Båda erhållna p-värden är mycket större än 0,05 så vi förkastar inte nollhypotesen. Vi kan med gott mod acceptera nollhypotesen att data inte ger anledning av tvivla på att de kan beskrivas med en Poissonfördelning. Vi ser också att i de båda beräkningarna är antalet frihetsgrader olika. Det beror på att vi slog samman två klasser och $(r-1)(k-1) = (2-1)(4-1) = 3$. Nu är det ytterligare en sak vi måste ta hänsyn till och det är att vi skattat parametern μ från stickprovet. Det medför att antalet frihetsgrader minskas med en enhet. Slutligen får vi två frihetsgrader. Har då modet sjunkit för att acceptera nollhypotesen? Vi får räkna om p-värdet med $df = 2$. Det gör vi med $1 - \text{pchisq}(2.1521, 2) = 0.3409$. Slutsatsen blir densamma. Grafen nedan illustrerar situationen.


0.5
0.4
0.3
0.2
0.1
0.0
Täthet Q, df=2
2.1521
0
2
4
6
8
10
12
Q