Kapitel 3

Sammenligning af to binomialfordelinger

Vi betragter følgende model: X_1 og X_2 er uafhængige, binomialfordelte med (kendte) antalsparametre n_1 og n_2 og ukendte sandsynlighedsparametre p_1 og p_2 i $]0,1[$.

Eksempel 3.1.

Nedenfor er 7006 låntagere i BRF Kredit, som pr. 1/1 1990 ikke havde betalt deres termin for december 1989 rettidigt, klassificeret efter ejendommens kategori (parcelhus eller ejerlejlighed) og efter om de senere (fordi de aldrig fik betalt) gik på tvangsauktion. Spørgsmålet er, om tvangsauktionshyppigheden kan antages at være den samme for parcelhuse og ejerlejligheder.

	Tv.auk.	Ikke tv.auk.	Sum
huse	656	5333	5989
lejl.	148	869	1017
Sum	804	6202	7006

En naturlig model går her ud på at opfatte rækkesummerne som faste, og antage at hver låntager går på tvangsauktion med en sandsynlighed som er p_1 eller p_2, afhængigt af ejendommens kategori. Hvis låntagerne desuden opfører sig uafhængigt, får vi netop den beskrevne binomialfordelingsmodel med $n_1 = 5989$, $n_2 = 1017$ og observationer $x_1 = 656$, $x_2 = 148$.

Likelihoodfunktionen bliver

$$L(p_1, p_2) = \binom{n_1}{x_1} p_1^{x_1} (1-p_1)^{n_1-x_1} \binom{n_2}{x_2} p_2^{x_2} (1-p_2)^{n_2-x_2}$$

og logaritmen til den (uden hensyn til binomialkoefficienterne) er

$$l(p_1, p_2) = x_1 \log p_1 + (n_1 - x_1)\log(1-p_1) + x_2 \log p_2 + (n_2 - x_2)\log(1-p_2).$$

3.1. Estimation.

Da log–likelihooden er sum af en funktion af p_1 og en funktion af p_2, kan maksimering af den foretages ved maksimering af hvert af disse led for sig. Og da hvert af de to led netop er en log–likelihood for en model af den type vi betragtede i kapitel 2, ser man umiddelbart at maksimaliseringsestimatorerne for p_1 og p_2 bliver de tilsvarende relative hyppigheder,

$$\hat{p}_1 = \frac{x_1}{n_1} \text{ og } \hat{p}_2 = \frac{x_2}{n_2}.$$

Usikkerheder på disse estimater kan ligeledes angives præcis som beskrevet i kapitel 2.

Eksempel 3.1 fortsat. I tvangsauktionseksemplet fås (regn selv efter)

$$p_1 = 0.110 \pm 0.008 \text{ og } p_2 = 0.146 \pm 0.022.$$

3.2. Testning.

Estimationsresultatet antyder (da de to intervaller er disjunkte) at $p_2 > p_1$. Men et mere solidt grundlag for at udtale os om dette får vi ved at teste hypotesen om, at de to parametre er ens. Vi opstiller altså hypotesen

$$p_1 = p_2 = p, \quad p \in]0,1[.$$

Likelihood'en i den tilsvarende model med én parameter p får vi ved at erstatte p_1 og p_2 med p i udtrykket for likelihood'en i den oprindelige model:

$$L(p) = p^{x_1}(1-p)^{n_1-x_1} p^{x_2}(1-p)^{n_2-x_2} = p^{x_1+x_2}(1-p)^{n_1+n_2-x_1-x_2}.$$

Bemærk, at denne funktion også kan fortolkes som likelihood'en for en simpel binomialfordelingsmodel, nemlig den der fortolker x_1+x_2 som den observerede værdi af en binomialfordelt variabel med antalsparameter n_1+n_2 og sandsynlighedsparameter p. Det passer fint med, at vi under antagelsen $p_1 = p_2 = p$ ifølge binomialfordelingens foldningsegenskab (Ssr., opgave 3.2.3 (b)) har, at $X_1 + X_2$ er binomialfordelt (n_1+n_2, p). Heraf følger umiddelbart, at ML–estimatoren for p under hypotesen er

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Eksempel 3.1 fortsat. I BRF–eksemplet betyder hypotesen $p_1 = p_2 = p$, at sandsynligheden for tvangsauktion er den samme for huse og lejligheder. Det forekommer helt rimeligt, at man i så fald skal estimere den fælles sandsynlighed for tvangsauktion ved

$$\hat{p} = \frac{656 + 148}{5989 + 1017} = \frac{804}{7006} = 0.115,$$

altså den relative hyppighed af tvangsauktioner beregnet uden hensyn til ejendomskategori.

Kvotientteststørrelsen bliver

$$q = \frac{\left(\frac{x_1+x_2}{n_1+n_2}\right)^{x_1+x_2}\left(\frac{n_1+n_2-x_1-x_2}{n_1+n_2}\right)^{n_1+n_2-x_1-x_2}}{\left(\frac{x_1}{n_1}\right)^{x_1}\left(\frac{n_1-x_1}{n_1}\right)^{n_1-x_1}\left(\frac{x_2}{n_2}\right)^{x_2}\left(\frac{n_2-x_2}{n_2}\right)^{n_2-x_2}}.$$

Dette besværlige udtryk kan skrives noget simplere, hvis vi ændrer notationen som følger. Vi lader y_1 og y_2 betegne de komplementære til x_1 og x_2, dvs.

$$y_1 = n_1 - x_1,\ y_2 = n_2 - x_2.$$

Desuden indføres betegnelserne

$$x. = x_1 + x_2,\ y. = y_1 + y_2 \text{ og } n. = n_1 + n_2.$$

Det betyder at antalstabellen

	"Succes"	"Fiasko"	Sum
Gruppe 1	x_1	$n_1 - x_1$	n_1
Gruppe 2	x_2	$n_2 - x_2$	n_2
Sum	$x_1 + x_2$	$n_1 + n_2 - x_1 - x_2$	$n_1 + n_2$

herefter kan skrives

	"Succes"	"Fiasko"	Sum
Gruppe 1	x_1	y_1	n_1
Gruppe 2	x_2	y_2	n_2
Sum	$x.$	$y.$	$n.$

Med denne notation kan kvotienttestørrelsen nu skrives

$$q = \frac{x.^{x.}\, y.^{y.}\, n_1^{n_1} n_2^{n_2}}{x_1^{x_1} y_1^{y_1} x_2^{x_2} y_2^{y_2} n.^{n.}}$$

Med betegnelsen $k(x)$ for funktionen

$$k(x) = 2x \log x$$

får vi derfor

$$-2\log q =$$
$$k(x_1) + k(x_2) + k(y_1) + k(y_2) - k(n_1) - k(n_2) - k(x.) - k(y.) + k(n.).$$

Denne formel er nem at huske, og nem at benytte til beregning af $-2\log q$ på en lommeregner. Sagt i ord: Kvotientteststørrelsen beregnes ved at man på hver plads i antalstabellen (inklusive række– og søjlesummer og totalsummen) anvender funktionen $k(x) = 2x\log x$. Kvotientteststørrelsen $-2\log q$ er en sum af disse størrelser, regnet med fortegn. For selve indgangene i tabellen og totalsummen benyttes positivt fortegn, for række– og søjlesummer negativt.

Hypotesen skal forkastes, hvis denne størrelse er ekstremt stor i sin fordeling. Ifølge generel asymptotisk teori — som vi desværre ikke er i stand til at berette om i detaljer — kan denne fordeling approksimeres med en χ^2–fordeling med $2-1=1$ frihedsgrad.

Eksempel 3.1 fortsat. For BRF–tallene får vi

$$
\begin{aligned}
-2\log q = 2(&656\log 656 + \cdots + 869\log 869 + 7006\log 7006 \\
&-5989\log 5989 - \cdots - 6202\log 6202) = 10.46.
\end{aligned}
$$

Da denne størrelse er tæt på 99.9%–fraktilen 10.828 (og i hvert fald større end 99.5%–fraktilen 7.879), må hypotesen helt afgjort forkastes.

Bemærk, at vi her er i en helt anden situtation end i de eksempler vi har set på vedrørende kast med en mønt, hvor der skulle meget ekstreme observationer til for at få os til at tvivle på den fundamentale hypotese $p = 0.5$. Her er det nærmest omvendt. Der er ingen som helst grund til at tro på hypotesen $p_1 = p_2$. En eller anden forskel af ikke–mikroskopisk størrelsesorden skal der nok være mellem ejerlejligheder og parcelhuse hvad angår risikoen for tvangsauktion. Spørgsmålet er kun hvor stor den er, og hvilken vej den går. Den oplysning, som testet giver os, er først og fremmest nyttig derved, at den berettiger os til med overbevisning at hævde, at tvangsauktionsrisikoen er størst for lejligheder. Hvis testet ikke havde ført til forkastelse ville det være en anden sag. Så ville vi ikke med rimelighed kunne udtale os om, hvilken ejendomskategori der er mest udsat.

3.3. Pearson's teststørrelse.

Ved at Taylorudvikle $-2\log q$ som funktion af $\hat{p}_1$ og $\hat{p}_2$ omkring $\hat{p}$ til og med anden orden kan man indse, at der for $\hat{p}_1$ og $\hat{p}_2$ nær på hinanden gælder

$$
-2\log q \approx \frac{(\hat{p}_1 - \hat{p}_2)^2}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})} = \left(\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}}\right)^2.
$$

Regningerne er for så vidt elementære, de er bare så besværlige og kedsommelige, at vi ikke vil gengive dem her. Højre side er Pearson's teststørrelse. Det sidste udtryk viser, at denne størrelse kan fortolkes som kvadratet på en størrelse, der har en simpel intuitiv fortolkning. Det virker jo umiddelbart rimeligt, at hypotesen $p_1 = p_2$ skal forkastes når differensen mellem $\hat{p}_1$ og $\hat{p}_2$ er stor. Størrelsen $\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})$ under kvadratrodstegnet er netop variansen for tælleren $\hat{p}_1 - \hat{p}_2$ under hypotesen, udregnet under antagelse af at $p = \hat{p}$. Af den centrale grænseværdisætning og normalfordelingens foldningsegenskab følger, at tælleren $\hat{p}_1 - \hat{p}_2$ er approksimativt normalfordelt, og dermed at Pearson's teststørrelse er approksimativt χ^2–fordelt med 1 frihedsgrad.

Eksempel 3.1 fortsat. For BRF–tallene bliver Pearson's teststørrelse

$$\frac{\left(\frac{656}{5989} - \frac{148}{1017}\right)^2}{\left(\frac{1}{5989} + \frac{1}{1017}\right)\frac{804}{7006}\frac{6202}{7006}} = 11.0858,$$

altså lidt større end $-2\log q$, faktisk større end 99.9%–fraktilen i χ^2–fordelingen med 1 frihedsgrad.

Opgave 3.3.1. Blandt de låntagere i BRF, som pr. 1/1–90 ikke havde betalt december 89 terminen rettidigt, havde de 6176 deres lån i BRF i form af et såkaldt mixlån. Af disse gik indenfor seks måneder 9.42% på tvangsauktion. Den tilsvarende tvangsauktionsprocent for de 2466 låntagere, hvis lån var af en anden type, var 10.50%.

(a) Angiv passende usikkerheder på de oplyste procenter.

(b) Er der statistisk hold i den påstand at låntagere med mixlån har mindre tvangsauktionsrisiko end de andre?

(Bemærk: Det samlede antal 6176+2466 er noget større end totalen 7006 i tabellen på side 14. Det skal man ikke lade sig forvirre af. I dette tilfælde skyldes den store forskel, at oplysningen om boligform ikke var umiddelbart tilgængelig for lån optaget før 1980. Det kan naturligvis give anledning til fortolkningsproblemer, men dem ignorerer vi her).

3.4. Fishers eksakte test.

Fordelingen af $-2\log Q$ under hypotesen $p_1 = p_2 = p$ afhænger naturligvis af p, selvom denne afhængighed åbenbart er approksimativt uden betydning. Den eksakte P–værdi

$$P = P(-2\log Q \geq -2\log q)$$

er derfor en funktion af p, og da p er ukendt er det principielt umuligt at angive P–værdien eksakt. Man kan imidlertid løse dette problem ved følgende trick. Antag, at vi har observeret $X_1 + X_2$, men ikke X_1 og X_2. I eksempel 3.1 ville det betyde, at vi har observeret det totale antal tvangsauktioner $x. = 804$. Da vi kender $n_1 = 5989$ og $n_2 = 1017$ kan vi også beregne antal ikke–tvangsauktioner $y. = 5989 + 1017 - 804 = 6202$. Kort sagt, vi kender tabellens marginaler, men ikke indmaden. Eftersom marginalerne ikke indeholder nogen som helst information om, hvorvidt hypotesen $p_1 = p_2$ er acceptabel eller ej, vælger vi at *betinge* med den observerede værdi af $x.$, dvs. vi ser nu på den betingede fordeling af (X_1, X_2), givet $X_1 + X_2 = x.$. Ifølge opgave 3.3.6 i Ssr. er denne fordeling under hypotesen beskrevet ved, at X_1 følger en hypergeometrisk fordeling, med parametre der på indlysende måde er givet ved de kendte marginaler (fortolk n_1 som antallet af røde kugler, n_2 som antallet af hvide kugler og $x.$ som stikprøvens størrelse). Da den ukendte parameter p ikke indgår i udtrykkene for denne fordelings punktsandsynligheder (groft sagt fordi vi netop har betinget med størrelsen $x.$, som indeholder al den information vi har om p) kan den *betingede* sandsynlighed for hændelsen $\{-2\log Q \geq -2\log q\}$, givet $X_1 + X_2 = x.$, udregnes ved summation af punktsandsynligheder i denne fordeling, og denne sandsynlighed vælger man så at fortolke som den relevante P–værdi. I praksis gør man det ved direkte at vurdere, hvor ekstremt x_1 ligger i den hypergeometriske fordeling, idet man udregner den tilsvarende halesandsynlighed og ganger den med 2.

Eksempel 3.1 fortsat. For BRF–tallene kan argumentet gengives således. Antallene 5989 og 1017 af henholdsvis huse og ejerlejligheder er givet. Ligeledes er det givet, at 804 af de i alt 7006 husejere går på tvangsauktion. Hvis tvangsauktionsrisikoen var ens for de to ejendomskategorier, ville vi være i samme situation som når man skal trække 804 kugler fra en kasse med 5989 røde og 1017 hvide kugler. Det vil sige at vi under hypotesen forventer, at antal tvangsauktioner blandt husejerne vil være hypergeometrisk fordelt med disse parametre. Hvis derimod tvangsauktionsrisikoen er forskellig for de to ejendomskategorier, kan vi naturligvis få et mere skævt resultat.

Vi får nu at vide at der var $x_1 = 656$ tvangsauktioner blandt husejerne. Spørgsmålet er, om dette med rimelighed kan være et udfald fra denne hypergeometriske fordeling. Den forventede værdi af en stokastisk variabel med denne fordeling er $\frac{5989 \times 804}{7006} = 687.3$, så hvis 656 skal være et ekstremt udfald i denne fordeling må det være fordi den *venstre* halesandsynlighed (summen af punktsandsynlighederne til og med nr. 656) er lille. Den kan udregnes (f.eks. v.h.a. WinT) til 0.0007.

Opgave 3.4.1. (Kilde: Opgaver i Statistik, Inge Henningsen, Københavns Universitet, Institut for Matematisk Statistik 1996). Ved kurset “Statistik 0” på Københavns Universitet gik 39 kvindelige og 60 mandlige studerende til ordinær eksamen sommeren 1990. Af disse bestod 21 kvindelige og 39 mandlige studerende. Tyder dette på en generel kønsforskel hvad angår beståelsesprocenten i dette fag?