Pearsons formel for χ^2 – test

Den teoretiske forklaring

Ole Witt-Hansen 2014

Indhold

1. Normalfordelingen og χ^2

Begrebet $\chi 2$,Chi-i-anden eller Chi-square. er en del af teorien for en stokastisk variabel, der er normalfordelt. En variabel med den teoretiske middelværdi μ og spredning σ er normalfordelt, hvis sandsynlighedstætheden er givet ved:

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Sandsynligheden for at man får et resultat (et udfald), som er mindre end *x*, er givet ved fordelingsfunktionen *F(x)*.

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

Har man *n* normalfordelte målinger x_1, x_2 , x_3, … x_n , hvor middelværdierne μ_1, $\mu_2, \mu_3, \ldots \mu_n$ og spredningen σ_1, $\sigma_2, \sigma_3, \ldots \sigma_n$ er kendte er sandsynligheden for at få et resultat i intervallet $dx_1 dx_2 dx_3 \ldots dx_n$ lig med produktet af de enkelte sandsynligheder.

$$P(x_1, x_2, x_3, \ldots, x_n) dx_1 dx_2 dx_3 \ldots dx_n = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}} dx_i$$

I mange tilfælde er middelværdi og spredning den samme for alle målingerne.

Hvis middelværdien er ukendt eller afhænger af en parameter α, kan man bestemme den mest sandsynlige værdi af α, ved at bestemme maximum for funktionen ovenfor, som funktion af α.

Dette kaldes ”principle of maximum likelihood”. I praksis tager man logaritmen til begge sider og dropper konstanterne, samt totallet i nævneren i eksponenten. Herved fremkommer det velkendte udtryk for χ^2.

$$\chi^2 = \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} + \ldots + \frac{(x_n - \mu_n)^2}{\sigma_n^2} \qquad \chi^2 = \sum_{k=1}^{n} \frac{(x_k - \mu_k)^2}{\sigma_k^2}$$

Denne formel danner grundlaget for hypotesetest, funktions fits til måledata og lineær regression.

Udtrykket for sandsynlighedstætheden for χ^2 er udledt i artiklen: Det teoretiske fundament for Chi-i-anden sandsynligheder. Det er:

$$P(\chi^2 > \chi_0^2) = \int_{\chi_0^2}^{\infty} F(\chi^2) d\chi^2$$

Hvis man skal vurdere en sandsynlighed f.eks. $P(\chi^2 > \chi_0^2)$, så gøres det ved hjælp af fordelingsfunktionen:

$$P(\chi^2 > {\chi_0}^2) = \int_{{\chi_0}^2}^{\infty} F(\chi^2) d\chi^2$$

Som altså angiver sandsynligheden for at χ^2 overstiger den observerede værdi ${\chi_0}^2$. Hvis ${\chi_0}^2$ er lig med nul, fordi alle observationer er lig med middelværdien, så er P =1, og jo større P er, jo større er sandsynligheden for at måleværdierne repræsenterer teorien.
Hvis P er stor, f.eks. 0,95, så er sandsynligheden altså kun 0,05 for at man rent statistisk kunne have fået en større værdi.
Accepterer man en hypotese eller verifikation af et observationssæt på dette grundlag, så taler man om hypotesetest med signifikansniveau 0,95. Man forkaster hvis den observerede værdi af χ^2 ikke befinder sig et det område, hvor 95% af observationerne befinder sig i.

Man kan vælge signifikansniveau, 0,90 , 0.95 , 0,99 afhængig af omstændighederne.

$P(\chi^2 > {\chi_0}^2)$ er en tabuleret funktion, som også kan aflæses på de fleste matematik-computere.

2. Pearsons formel for χ^2 test

I lærebøgerne i matematik for gymnasiet, og givetvis i bøger fra samfundsfag og biologi, findes der, imidlertid en formel, som betegnes som χ^2 test, og hyppigt anvendes, men som faktisk ikke ligner de udtryk for χ^2, som udledt ud fra normalfordelingen. Den kaldes for Pearsons formel.

Hvis vi lader O_i betegne de observerede værdier, og E_i de forventede værdier (expectation value), så definerede Karl Pearson en "test value", som fik den samme betegnelse χ^2.

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

Ser man på χ^2, som det er defineret ud fra normalfordelingen, så er det ikke helt ligetil – heller ikke for matematiklærere - at gennemskue, hvorledes de to udtryk, står for den samme statistiske deskriptor. Forklaringen er heller ikke triviel.
Bemærk især, at i modsætning til den traditionelle χ^2 –test, så kan formlen anvendes, selv om man kun har en observation!

Formlen har mange anvendelser, men anvendes ofte til at afgøre om to størrelser er afhængige eller uafhængige.
Hvis to størrelser er statistisk uafhængige, kan man som bekendt finde sandsynligheden for, at de indtræffer samtidig som produktet af de to sandsynligheder.

Hvis ikke, skal man anvende de betingede sandsynligheder. Se f.eks. Ole Witt-Hansen Sandsynlighedsregning, hvor betinget sandsynlighed er ret indgående beskrevet. Bogen findes på hjemmesiden.

P(H) betegner som sædvanlig sandsynligheden for at hændelsen H indtræffer.

I et endeligt sandsynlighedsfelt betegner *n* eller *n(U)* antallet af elementer i udfaldsrummet *U*.
n(H) betegner antallet af elementer i hændelsen *H*.

Sandsynligheden *P(H)* svarer i statistikken til frekvensen af en observation: *f(H)= n(H)/n*

Hændelserne A og B er uafhængige, hvis $P(A \cap B) = P(A)P(B)$

Sandsynligheden for at A indtræffer, givet B skrives: $P(A \mid B)$

Hvis hændelserne A og B er ikke uafhængige er

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \Leftrightarrow P(A \cap B) = P(A \mid B)P(B).$$

Vi vil nu give et teoretisk (dvs. uden at inddrage tal) eksempel på, hvordan man undersøge om to observationer er uafhængige af hinanden.

Vi vil under søge hypotesen: Er en stemme på Socialdemokratiet (S), uafhængigt af om man er ufaglært (U) eller ej.

Antallet af stemmeberettigede betegnes *n*.

Antallet der stemmer på S betegnes: *n(S)*, og antallet, der ikke stemmer på S betegnes $n(\overline{S})$ og sådan fremdeles.

$n(U \cap \overline{S})$ er antallet af ufaglærte, der ikke stemmer på *S*.

Vi har derfor 4 muligheder. $n(U \cap S)$, $n(U \cap \overline{S})$, $n(\overline{U} \cap S)$, $n(\overline{U} \cap \overline{S})$.

Disse 4 muligheder er imidlertid ikke uafhængige, idet der gælder de to relationer

$n(U \cap S) + n(U \cap \overline{S}) = n(U)$ og $n(U \cap S) + n(\overline{U} \cap S) = n(S)$

For så vel for *S*, som for *U* har vi 2-1 frihedsgrader, som er (2-1)(2-1) = 1 frihedsgrad.

Vores hypotese var, at når man stemmer S, så er det uafhængigt af, om man er ufaglært eller ej.

Vi antager altså at: $P(U \cap S) = P(U)P(S)$.

Vi skriver dernæst Pearsons formel op, som har 4 led.

$$\chi^2 = \frac{(n(U \cap S) - n(U)n(S))^2}{n(U)n(S))} + \frac{(n(\overline{U} \cap S) - n(\overline{U})n(S))^2}{n(\overline{U})n(S))} + \frac{(n(U \cap \overline{S}) - n(U)n(\overline{S}))^2}{n(U)n(\overline{S}))} + \frac{(n(\overline{U} \cap \overline{S}) - n(\overline{U})n(\overline{S}))}{n(\overline{U})n(\overline{S}))}$$

Hvis man indsætter virkelige tal, f.eks. fra sidste folketingsvalg, vil vi få en værdi for χ^2.

Hvis vi får en værdi på 3,84, så er sandsynligheden for at få en større værdi en den beregnede omkring 0.95. Værdien ligger altså i et 95% acceptinterval.

Hypotesen kan accepteres, men hvis vi f.eks. får en værdi på 10 χ^2, er sandsynligheden for at få en større værdi 0,00157.
Denne værdi er så lille, at det er så usandsynligt at det udelukkende kan skyldes statistiske tilfældigheder, at vi vil forkaste hypotesen.
Med virkelige tal vil testen sandsynligvis være endnu mere signifikant i dette konstruerede tilfælde.

3. Forklaring på Pearsons formel.

Formålet er nu at give en begrundelse for, at Pearsons formel godt kan opfattes son en χ^2 - test.

Persons test, tager imidlertid ikke udgangspunkt i normalfordelingen, men i binomialfordelingen.

For en stokastisk variabel *X*, hvor sandsynligheden for at den indtræffer er *p*, dvs. *P(X) = p,* så er den binomialfordelt, hvis sandsynligheden for at hændelsen indtræffer netop *q* gange i *n* forsøg er givet ved.

$$P(x=q)=\binom{n}{q}p^q(1-p)^{n-q}$$

middelværdien $\mu = np$ og spredningen $\sigma=\sqrt{np(1-p)}$

Vi opskriver nu Pearsons formel

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

Om to stokastiske variabel *X* og $\overline{X}$, som vi antager, er binomialfordelte, er *X* og $\overline{X}$ ikke uafhængige, idet de udelukker hinanden og tilsammen er hele udfaldsrummet, er antallet af frihedsgrader er 2-1 = 1.
Hvis forsøget gentages n gange og X indtræffer *q* gange, så indtræffer $\overline{X}$ *n-q* gange.

Endvidere er

$$E(X) = np \quad \sigma(X)=\sqrt{np(1-p)} \quad \text{og} \quad E(\overline{X})=n(1-p) \;\; og \;\; \sigma(\overline{X})=\sqrt{n(1-p)}p$$

Ifølge Pearsons formel

$$\chi^2=\sum_{i=1}^{n}\left(\frac{(X_i-np)^2}{np}+\frac{((n-X_i)-n(1-p))^2}{n(1-p)}\right)$$

Vi nøjes med at se på ét af leddene, dropper indeks i, og som vi underkaster en række omskrivninger:

$$\chi_i^{\,2}=\frac{(X-np)^2}{np}+\frac{((n-X)-n(1-p))^2}{n(1-p)}=\frac{(X-np)^2}{np}+\frac{(np-X)^2}{n(1-p)}=$$

$$\frac{(1-p)(X-np)^2+p(np-X)^2}{np(1-p)}=$$

$$\frac{(X-np)^2}{np(1-p)} = \frac{(X-\mu)^2}{\sigma^2}$$

Det sidste udtryk er netop udtrykket for det led som indgår i $\chi^2 = \sum_{k=1}^{n} \frac{(x_k - \mu_k)^2}{\sigma_k^2}$,

Hvilket viser, at Pearsons test faktisk er en χ^2 –test.

Man kan indvende, at χ^2 refererer til normalfordelingen, mens Pearsons test tager udgangspunkt i binomialfordelingen, men man kan vise, (det er langtfra simpelt), at binomialfordelingen nærmer sig asymptotisk til normalfordelingen, når *n* går imod uendelig, samtidig med at *np* holdes konstant.

Det følger heraf, at man kan anvende Persons formel som en χ^2 test med én frihedsgrad.

Ole Witt-Hansen

2014-11-30