χ^2 –test i matematikundervisningen

Ole Witt-Hansen 2014

Indhold

1. χ^2 - test i gymnasiets matematikundervisning

I januar nummeret 2008 af LMFK bladet havde jeg en artikel, hvor jeg harcelerede lidt over, at regression og især χ^2. fordeling havde fundet indpas i matematikundervisningen samtidig med, at sandsynlighedsregningen helt forsvandt fra matematikundervisningen efter reformen 2005.

At man anvender regression og χ^2- test i samfundsfag og biologi, mener jeg ikke er nogen relevant begrundelse for at indføre det i matematik. I disse fag anvendes der masser af formler uden forklaring, men jeg synes (stadig), at man i matematik skal kunne forklare det, der står i bøgerne for eleverne, hvilket tidligere jo også havde den forudsætning, at læreren selv forstod det.
I den forbindelse har jeg hørt mange lærerfrustrationer, når χ^2 test skal "forklares".

I fysik behandlede man tidligere måleresultaterne grafisk med millimeterpapir og logaritmiske papirer, hvor eleverne selv skulle afsætte punkterne og selv aflæse de relevante oplysninger af graferne og hvor eleverne selv skulle kunne vurdere måleresultaterne i forhold til punkternes beliggenhed omkring en linie. Dette i stedet for, at skulle henvise til den kryptiske korrelationskoefficient.

Men nu er den naturvidenskabelige pædagogiske fornuft jo for længst nedkæmpet til fordel for den IT-fikserede fliptur i undervisningspolitisk korrekthed, som har hærget undervisningen i matematik og fysik gennem de sidste 10 år.

2. Sandsynlighedstætheden for χ^2

Men tilbage til regression og Chi-i-anden. test. Begge begreber refererer til normalfordelingen med middelværdi μ og spredning σ.

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

Har man n normalfordelte målinger x_1, x_2 , x_3, … x_n , hvor spredningen σ_1, σ_2, σ_3, … σ_n er kendte og $\bar{x}_i$ betegner de teoretiske middelværdier, så er sandsynligheden for at få et resultat i intervallet $dx_1 dx_2 dx_3 ... dx_n$ lig med.

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

Indfører man de normerede variable q_i , ved $q_i = \frac{x_i - \bar{x}}{\sigma_i}$ og

$$Q(q_1, q_2, q_3, ..., q_n) dq_1 dq_2 dq_3 ... dq_n = P(x_1, x_2, x_3, ..., x_n) dx_1 dx_2 dx_3 ... dx_n$$

får man

$$Q(q) = Q(q_1, q_2, q_3, ..., q_n) = \left(\frac{1}{2\pi}\right)^{N/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{N} {q_i}^2\right)$$

Q afhænger kun af variablene q_i gennem summen: $\chi^2 = \sum_{i=1}^{N} {q_i}^2$.

χ^2 anvendes som bekendt til at vurdere hvorvidt resultatet af N målinger ligger inden for den statistiske usikkerhed.

Herefter bliver udtrykket for $Q(q_1,q_2,q_3,...,q_n)=\left(\frac{1}{2\pi}\right)^{N/2}\exp\left(-\frac{1}{2}\chi^2\right)$

χ^2 har en tabuleret sandsynlighedsfordeling, som skrives $F(\chi^2)d\chi^2$.
Udledningen af udtrykket for $F(\chi^2)$ er ret kompliceret i de fleste fremstillinger, men det kan udledes relativt simpelt, hvis man betragter $\chi^2=\sum_{i=1}^{N}{q_i}^2$, som afstanden ud til et punkt i et N-dimensionalt rum.[1]

Volumen af en kugleskal med radius *r* i et *N*-dimensionalt rum er nødvendigvis proportional med r^{N-1}. I planen er rumfangselementet i polære koordinater $dV_2 = rdrd\varphi$. I rummet er rumfangselementet $dV_3 = r^2 sin\theta\, dr\, d\theta\, d\varphi$.
Udregner man Jacobi determinanten for omregning fra kartesiske koordinater til polære koordinater, vil alle koordinater x_i have en faktor *r* gange en funktion af de *N-1* vinkler. Ved den partielle differentiation af x_i'erne, vil *r* forsvinde i netop en søjle, og hvert led i determinanten, vil derfor have faktoren r^{N-1}. Rumfangselementet i et N-dimensionalt rum, må derfor være proportionalt med denne faktor.

Vi vil nu først bestemme (på nær en konstant) bidraget til $F(\chi^2)d\chi^2$ fra en kugleskal mellem χ og χ + dχ, hvor altså χ er konstant.

$$F(\chi^2)d\chi^2 = F(\chi^2)2\chi d\chi = \int_{\substack{kugleskal\\ \chi,d\chi}} Q(q_1,q_2..q_n)dq_1dq_2...dq_n = \left(\frac{1}{2\pi}\right)^{N/2}\exp\left(-\frac{1}{2}\chi^2\right)\int_{\substack{kugleskal\\ \chi,d\chi}} dq_1dq_2...dq_n$$

Det sidste integral, (når man integrerer over de *N-1* vinkler) er ifølge det foregående proportionalt med radius i "kuglen i N-1 potens", som er χ^{N-1}. Samler vi integralet over vinklerne og de øvrige konstanter i en faktor 2*C*, finder man derfor:

$$F(\chi^2)d\chi^2 = F(\chi^2)2\chi d\chi = (2C)\chi^{N-1}\exp\left(-\frac{1}{2}\chi^2\right)d\chi = C\chi^{N-2}\exp\left(-\frac{1}{2}\chi^2\right)2\chi d\chi$$

Konstanten C, kan derefter bestemmes ved normaliseringsbetingelsen: $\int_0^\infty F(\chi^2)d\chi^2 = 1$

Gammafunktionen *Γ* er defineret ved integralet: $\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt$.

Der gælder som bekendt for heltallige og positiv *n, at, Γ(n+1)=n!*

Ovenstående normaliseringsintegral, kan derfor udtrykkes ved Gamma funktionen ved substitutionen: $t=\frac{1}{2}\chi^2 \Rightarrow \chi=\sqrt{2t}$ *og* $dt=\chi d\chi$. Herved får man:

$$\int_0^\infty F(\chi^2)d\chi^2 = 2C\int_0^\infty \chi^{N-1}\exp\left(-\frac{1}{2}\chi^2\right)d\chi = 2C\int_0^\infty (2t)^{\frac{N-1}{2}}\exp(-t)\frac{1}{\sqrt{2t}}dt = C\cdot 2^{\frac{N}{2}}\int_0^\infty t^{\frac{N}{2}-1}\exp(-t)dt$$

Det sidste integral er $C\cdot 2^{\frac{N}{2}}\Gamma(\frac{N}{2})$ så normaliseringsbetingelsen giver: $C=\left(2^{\frac{N}{2}}\Gamma(\frac{N}{2})=1\right)^{-1}$.

Herefter følger udtrykket for fordelingsfunktionen for χ^2.

$$F(\chi^2)d\chi^2 = \frac{1}{2^{\frac{N}{2}}\Gamma(\frac{N}{2})}e^{-\frac{1}{2}\chi^2}(\chi^2)^{\frac{N}{2}-1}d\chi^2$$

3. Anvendelse af χ^2 til statistisk behandling af måledata. Frihedsgrader.

Ovenstående er sandsynlighedstætheden for χ^2 er kendt som χ^2 fordelingsfunktionen. Sandsynligheden for at få en værdi af χ^2 , som overstiger χ_0^2 er givet ved:

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

N er i denne formel det samme som antallet af uafhængige variable. *N* betegnes imidlertid som antallet af frihedsgrader. Mere generelt er antallet af frihedsgrader lig med antallet af uafhængige variable minus antallet af lineære relation, der findes mellem disse variable.

Hvis man f.eks. i formelen for $\chi^2 = \frac{(x_1-\mu)^2}{\sigma^2}+\frac{(x_2-\mu)^2}{\sigma^2}+...+\frac{(x_n-\mu)^2}{\sigma^2}$ erstatter den teoretiske middelværdi μ med det beregnede gennemsnit $\bar{x} = \frac{1}{n}(x_1+x_2+x_3+...+x_n)$, så er der netop en lineær relation mellem de uafhængige variable, og antallet af frihedsgrader er N – 1.

Formlen $F(\chi^2)d\chi^2 = \frac{1}{2^{\frac{N}{2}}\Gamma(\frac{N}{2})}e^{-\frac{1}{2}\chi^2}(\chi^2)^{\frac{N}{2}-1}d\chi^2$ er imidlertid den samme, når blot N erstattes af N-1. Beviset for dette er imidlertid ret utilgængeligt.

Fordelingsfunktionen for $F(\chi^2)d\chi^2$ skrives traditionelt:

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

Denne funktion er tabuleret og kan i øvrigt findes på en CAS.

Hvis χ^2 er lig med 0, fordi alle observationer er lig med middelværdien, så er sandsynligheden P = 1. Jo større $P(\chi^2 > {\chi_0}^2)$ er jo bedre er observationerne statistisk set. Skal man foretage en test med et signifikansniveau på 5%, skal sandsynligheden for et resultat, som er større end det udregnede, altså være mindre end 0,05. (Da dette kun er lidt sandsynligt, hvis afvigelsen er statistisk) Acceptbetingelsen altså, at

$$P(\chi^2 > {\chi_0}^2) > 0{,}95$$

Og hypotesen forkastes, hvis $P(\chi^2 > {\chi_0}^2) < 0{,}05$

.

Især det sidste, har jeg erfaret giver vanskeligheder, når det skal forklares til eleverne – afsluttende med "Sådan er det bare", men sådanne "forklaringer" synes jeg man bør overlade til andre fag end matematik.

Hvis de teoretiske middelværdier og spredninger ikke er kendte, kan man anvende de to estimater:

$$\bar{x} = \tfrac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{og} \qquad s^2 = \tfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Begrundelsen for disse formler er naturligvis, at $E(\bar{x}) = \mu$ og $E(s^2) = \sigma^2$.

Man kan imidlertid ikke anvende disse udtryk, hvis man kun har én observation.

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

4. Pearsons χ^2 test

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 vanskelig, at gennemskue, hvorledes de to udtryk, står for den samme statistiske deskriptor. Forklaringen er ikke helt triviel. Bemærk især, at i modsætning til den traditionelle χ^2 –test, så kan formlen anvendes, selv om man kun har en observation. Forklaringen følger i noten: Chi-i-anden test og Pearsons formel.

Jeg skal ikke bebrejde nogen, at den teoretiske forklaring på χ^2 –test og Pearsons formel ikke står i en lærebog for gymnasiet, da det jo ligger langt over den elementære sandsynlighedsregning – som man i øvrigt ikke lærer længere i gymnasiet efter 2005, men jeg synes ufortrødent, at det som der står i en matematiklærebog – i hvert fald på A-niveau – bør kunne forklares for eleverne.

5. Grænserne for usikkerhed i en stikprøve

En anden formel, som står refereret i matematikbøgerne er grænserne for den usikkerhed, der er på en stikprøve, hvor n er størrelsen af stikprøven og p er sandsynligheden for udfaldet Formlen er:

$$f = 1{,}96\sqrt{\frac{p(p-1)}{n}}$$

.

Kvadratrodsfaktoren er jo spredningen på frekvensen fra binomialfordelingen, men hvorfor 1,96?

Svaret er det simple, at -1,96 er 2,5% fraktilen for normalfordelingen, altså, at

$$\Phi(\frac{x-\mu}{\sigma}) = \Phi(\frac{\mu-\alpha\sigma-\mu}{\sigma}) = \Phi(-\alpha) = 0{,}025 \quad \Leftrightarrow \ \alpha = 1{,}96$$

Hvis observationssættet er normalfordelt vil 2,5% af observationerne – statistisk set – ligge under $\mu - 1{,}96\sigma$. Da normalfordelingen er symmetrisk omkring middelværdien, ligger 2,5% af observationerne over $\mu + 1{,}96\sigma$ og 95% af observationerne vil derfor ligge i intervallet $[\mu - 1{,}96\sigma , \mu - 1{,}96\sigma]$. Tager man spredningen fra binomialfordelingen, kan man ved gange den med 1,96, med et signifikansniveau på 95% vide, at ”den rigtige værdi” ligger i dette interval. Dette er indholdet af formlen ovenfor.

Efter reformen, står der en del i lærebøgerne i fysik som matematik, som ikke længere bliver forklaret for eleverne, blandt andet på grund af manglende forudsætninger. Og det kan man jo have en mening om. At de bliver undervist i formler, som læreren heller ikke forstår, kan man vist kun have én mening om.

[1] Mathematical methods of physics. Jon Mathews, Robert L. Walker

Ole Witt-Hansen