χ2–test i matematikundervisningen

Ole Witt-Hansen, Køge Gymnasium

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.

Men tilbage til regression og χ^2 –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 $x_1, x_2, x_3, \ldots x_n$, hvor spredningen $\sigma_1, \sigma_2, \sigma_3, \ldots \sigma_n$ er kendte og udregner middelværdien som

$$\bar{x}=\tfrac{1}{n}(x_1+x_2+x_3+\ldots+x_n)$$

er sandsynligheden for at få et resultat i intervallet

$$dx_1\cdot dx_2\cdot dx_3\ldots dx_n$$

lig med

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

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

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

får man

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

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 $Q(q)$

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

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 $dV_2 = r\cdot dr\cdot d\varphi$. I rummet er rumfangselementet $dV_3 = r^2\cdot\sin\theta\cdot dr\cdot d\theta\cdot d\varphi$.

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 $\chi + d\chi$, hvor altså χ er konstant.

$$F(\chi^2)d\chi^2=F(\chi^2)2\chi d\chi=\int_{\substack{\text{kugleskal}\\ \chi,d\chi}}Q(q_1,q_2..q_n)\cdot dq_1\cdot dq_2\ldots dq_n$$

$$=\left(\frac{1}{2\pi}\right)^{N/2}\exp\left(-\frac{1}{2}\chi^2\right)\int_{\substack{\text{kugleskal}\\ \chi,d\chi}}dq_1\cdot dq_2\ldots dq_n$$

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 $2C$, 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 positive n, at, $\Gamma(n+1) = n!$

Ovenstående normaliseringsintegral, kan derfor udtrykkes ved Gammafunktionen ved substitutionen:

$$t = \tfrac{1}{2}\chi^2 \Rightarrow \chi = \sqrt{2t} \;\; og \;\; dt = \chi d\chi$$

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.

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

Fordelingsfunktionen for χ^2 er kendt som chi–square fordelingsfunktionen. Sandsynligheden for at få en værdi af χ^2, som ikke overstiger ${\chi_0}^2$ er givet ved:

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

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%, er acceptbetingelsen altså at $P(\chi^2 > {\chi_0}^2) > 0{,}95 \;\Leftrightarrow\; P(\chi^2 < {\chi_0}^2) < 0{,}05$.

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.

Skal man foretage en test, er der principielt ikke noget i vejen for, at lade x_i'erne høre til forskellige observationer og lade μ_i'erne være teoretiske eller påståede værdier y_i.

Problemet er imidlertid, hvordan man skal vurdere spredningen σ_i. Man kan ikke anvende udtrykkene

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

hvis man kun har én observation.

Som beskrevet men ikke redegjort for i nogle matematikbøger, anvender man imidlertid spredningen for Poissonfordelingen, som angiver sandsynligheden for at få netop n observationer i tidsrummet t, når sandsynligheden for at få netop 1 observation pr. tidsenhed er λ. Poissonfordelingen er:

$$P_n(t) = \frac{(\lambda t)^n}{n!}e^{-\lambda t}$$

med middelværdi $< n > = \lambda\cdot t$ og spredning $\sigma = \sqrt{< n >}$.

Dette er den eneste forklaring jeg kan give på de χ^2–test, man anvender i samfundsfag og biologi, og som er refereret i matematikbøger (uden forklaring – heller ikke for læreren), hvor leddene i χ^2 har formen: $\frac{(y_i - x_i)^2}{x_i}$. I nævneren står det observerede antal. Der burde stå kvadratet på spredningen. Men ifølge Poissonfordelingen er spredningen lig med kvadratroden af det forventede antal. Når man i stedet for det forventede antal anvender det observerede antal, giver formlen mening.

Jeg skal ikke bebrejde nogen, at forklaringer på sådanne formler 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 der står i en matematiklærebog – i hvert fald på A–niveau – bør kunne forklares for eleverne.

En anden formel, som står refereret i matematikbøgerne er grænserne for den usikkerhed, der er på en stikprøve: Formlen er:

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

hvor n er størrelsen af stikprøven, og p er sandsynligheden for udfaldet.

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

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

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 $\mu + 1{,}96\sigma$. 95% af observationerne vil derfor ligge i intervallet $[\mu - 1{,}96\sigma\ ,\ \mu + 1{,}96\sigma]$.

Tager man spredningen fra binomialfordelingen, kan man ved at 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 og 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)] Jon Mathews, Robert L. Walker, *Mathematical methods of physics.*