Supplement til Kreyszig

Forelæsningsnoter til Matematik F2

Indholdsfortegnelse side

Jens Jensen

Ørsted Laboratoriet, NBIfAFG, December 2002

1. Numerisk integration. Fejlvurdering af trapez og Simpson algoritmerne

Kreyszig bestemmer fejlen af trapezmetoden til at være; (3*) i §17.5 (side 959 i 7. udgave),

$$\Delta\epsilon = -\frac{h^3}{12}f''(t)$$

for det $i$te delinterval mellem x_i og $x_{i+1} = x_i + h$. Intervallets længde er $\Delta x = h$ og $x_i < t < x_{i+1}$. I grænsen $h \to 0$ haves, at $x_{i+1} \to x_i = x$ og

$$\frac{d\epsilon}{dx} \simeq \frac{\Delta\epsilon}{\Delta x} = \frac{\Delta\epsilon}{h} = -\frac{h^2}{12}f''(x) \quad \Rightarrow \quad \epsilon = \int_a^b \frac{d\epsilon}{dx}dx \simeq -\frac{h^2}{12}\int_a^b f''(x)dx$$

og vi får dermed

$$\boxed{\epsilon(\text{trapez}) = -\frac{h^2}{12}[f'(b) - f'(a)] + \mathcal{O}(h^3)} \tag{1.1}$$

Korrektionerne til denne fejlvurdering er af størrelsesordenen h^3 (hvilket også gælder Kreyszig's resultater).

I Simpsons tilfælde benyttes et 2. grads polynomium som tilnærmelse for $f(x)$ og afvigelsen er, $x_0 \le x \le x_2$,

$$f(x) - p_2(x) = \frac{1}{3!}(x - x_0)(x - x_1)(x - x_2)\Big[f^{(3)}(x_1) + \tfrac{1}{4}(x - x_1)f^{(4)}(x_1)\Big] + \mathcal{O}(h^5)$$

Den tredie afledede af venstre side er $f^{(3)}(x)$ og højre side af ligningen giver $f^{(3)}(x_1) + [x - \frac{1}{4}(x_0 + 2x_1 + x_2)]f^{(4)}(x_1) = f^{(3)}(x_1) + (x - x_1)f^{(4)}(x_1)$.

Indføres $s = (x - x_1)/h$ fås

$$\begin{aligned}\Delta\epsilon &= \int_{x_0}^{x_2}[f(x) - p_2(x)]dx \\ &= \frac{h^4}{6}\int_{-1}^{1}(s+1)s(s-1)\Big[f^{(3)}(x_1) + \tfrac{1}{4}hsf^{(4)}(x_1)\Big]ds = -\frac{h^5}{90}f^{(4)}(x_1)\end{aligned}$$

Leddet proportionalt med h^4 forsvinder ved integrationen, og vi har derfor medtaget korrektionen til næste orden i h. Benyttes $\Delta x = x_2 - x_0 = 2h$ fås ved samme procedure som ovenfor:

$$\boxed{\epsilon(\text{Simpson}) = -\frac{h^4}{180}[f'''(b) - f'''(a)] + \mathcal{O}(h^5)} \tag{1.2}$$

Simpson algoritmen er ikke som umiddelbart forventet en 3. ordens, men en 4. ordens metode.

De udledte udtryk for fejlene i de to tilfælde (1.1) og (1.2) er langt lettere at benytte og mere præcise end Kreyszig's intervalafgrænsninger af fejlene.

2. Dekomponering af brøker (Laplace transformation)

I regninger med Laplace-transformationen sker det tit, at $F(s)$ er en *ægte brudden rational funktion*, dvs.

$$F(s) = \frac{P_n(s)}{Q_m(s)}$$

hvor $P_n(s)$ og $Q_m(s)$ er polynomier i s af $n$te henholdsvis $m$te grad og $n < m$ (ægte). For at finde den inverse må $F(s)$ *dekomponeres* til en sum af *stambrøker*: Betegner a_i de m (inkl. komplekse) rødder i $Q_m(s)$, dvs.
$Q_m(s) = K(s-a_1)^k(s-a_2)^l\cdots$, så er $F(s)$ entydigt bestemt ved udtrykket

$$F(s) = \frac{b_k}{(s-a_1)^k} + \frac{b_{k-1}}{(s-a_1)^{k-1}} + \cdots + \frac{b_1}{s-a_1} + \frac{c_l}{(s-a_2)^l} + \cdots \qquad (2.1)$$

hvor konstanterne i stambrøkernes tællere findes ved at sætte hele udtrykket på fælles brøkstreg med Q_m i nævneren og benytte, at denne brøks tæller skal være lig $P_n(s)$. Er alle konstanter i $P_n(s)$ og $Q_n(s)$ reelle vil eventuelle komplekse rødder optræde i konjugerede par, og

$$\frac{d_h}{(s-\alpha-i\beta)^h} + \frac{d_h^*}{(s-\alpha+i\beta)^h} = \frac{e_h s + f_h}{[(s-\alpha)^2+\beta^2]^h} + \cdots + \frac{e_1 s + f_1}{(s-\alpha)^2+\beta^2} \qquad (2.2)$$

dvs., at eventuelle komplekse led i (2.1) kan erstattes med led svarende til højre siden af ligning (2.2).

Eksempel:

$$F(s) = \frac{P_1(s)}{Q_5(s)} = \frac{s-1}{s^2(s-2)(s^2+s+1)} = \frac{b_2}{s^2} + \frac{b_1}{s} + \frac{c}{s-2} + \frac{ds+e}{s^2+s+1}$$
$$= \frac{(b_2+b_1 s)(s-2)(s^2+s+1) + cs^2(s^2+s+1) + (ds+e)s^2(s-2)}{s^2(s-2)(s^2+s+1)}$$

Samles leddene i tælleren for hver orden af s fås følgende ligninger:

$$\begin{aligned}
s^4: &\quad b_1 + c + d = 0\\
s^3: &\quad b_2 - b_1 + c - 2d + e = 0\\
s^2: &\quad -b_2 - b_1 + c - 2e = 0\\
s^1: &\quad -b_2 - 2b_1 = 1\\
s^0: &\quad -2b_2 = -1\\
&\Rightarrow \quad b_2 = \tfrac{1}{2}; \qquad b_1 = -\tfrac{3}{4}; \qquad c = \tfrac{1}{28}; \qquad d = \tfrac{5}{7}; \qquad e = \tfrac{1}{7}
\end{aligned}$$

Bemærk, at c kan bestemmes langt lettere ved at gange brøkomskrivningerne med $s-2$ og derefter sætte $s = 2$, og tilsvarende for b_2 (gang med s^2 og sæt $s = 0$).

Benyttes omskrivningen $\dfrac{1}{7}\cdot\dfrac{5s+1}{s^2+s+1} = \dfrac{1}{7}\cdot\dfrac{5(s+\frac{1}{2})-\frac{3}{2}}{(s+\frac{1}{2})^2+\frac{3}{4}}$ fås

$$f(t) = \mathcal{L}^{-1}\{F(s)\} = \tfrac{1}{2}t - \tfrac{3}{4} + \tfrac{1}{28}e^{2t} + \tfrac{1}{7}e^{-t/2}\Big(5\cos\sqrt{\tfrac{3}{4}}\,t - \sqrt{3}\sin\sqrt{\tfrac{3}{4}}\,t\Big)$$

NB: Mathematica dekomponerer brøker med kommandoen: Apart[brøk]

3. Permutationer og kombinationer, teorem 3

Teorem 3 i §22.4 i 8. udgave af Kreyszig (teorem 4 i §23.3 i 7. udgave) lyder:

Antallet af forskellige kombinationer med k elementer, der kan dannes ud af n forskellige elementer, er

uden gentagelser: $\binom{n}{k}$ *med gentagelser:* $\binom{n+k-1}{k}$

Her vil vi gerne bevise den sidste del af teoremet, ligning (4b). I dette tilfælde kan det enkelte element optræde fra 0 til k gange i de forskellige kombinationer. Lærebogen foreslår at bevise sætningen ved induktion:

Teoremet er korrekt for $n = 1$ for en vilkårlig værdi af k: Der er kun den ene mulighed, at alle k elementer er ens.

Teoremet antages korrekt for $n - 1$ elementer (for en vilkårlig værdi af k), og skal med denne forudsætning vises at være gyldigt for n elementer:

Vi starter med de $n - 1$ elementer og danner alle mulige kombinationer med k elementer (med gentagelser): $\binom{n-1+k-1}{k}$.

Til dette antal muligheder skal adderes dem, hvor det $n$te element optræder 1 gang. Dette er lig med antallet af kombinationer med $k - 1$ elementer valgt ud blandt de resterende $n - 1$ elementer: $\binom{n-1+k-1-1}{k-1}$.

Næste led er de kombinationer, hvor det $n$te element optræder to gange, og hvor der skal vælges $k - 2$ elementer ud af de $n - 1$ elementer osv., indtil vi når frem til, at alle k-elementer er det $n$te element. Dermed bliver det samlede antal kombinationer:

$$\binom{n+k-2}{k} + \binom{n+k-3}{k-1} + \binom{n+k-4}{k-2} + \cdots + \binom{n}{2} + \binom{n-1}{1} + 1$$

som kan omskrives til

$$\sum_{s=0}^{k} \binom{n-2+s}{s} = \sum_{s=0}^{k} \binom{n-2+s}{n-2}$$

Benyttes ligning (13) i bogen (med k erstattet med $n-2$ og øverste sum grænse $n-1$ med k) fås, at summen er lig

$$\binom{k+1+n-2}{n-2+1} = \binom{n+k-1}{n-1} = \binom{n+k-1}{k} \qquad Q.E.D.$$

Ligning (13), som benyttes i beviset, kan igen bevises ved induktion samt brug af ligning (11) (regneøvelse-opgave).

Alternativt bevis for Teorem 3

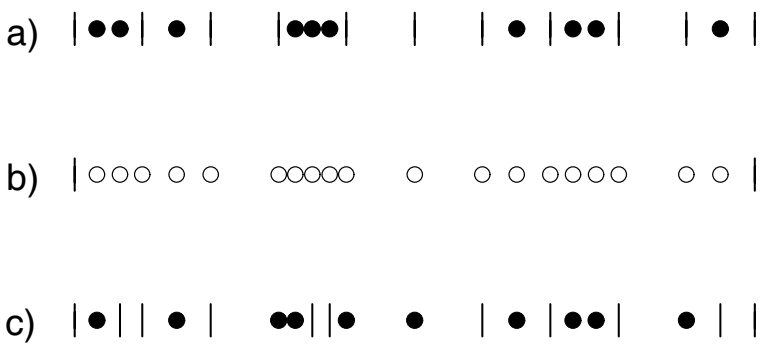
Teorem 3 kan omformuleres til, at antallet af måder k kugler kan fordeles i n kasser er

$$\binom{n+k-1}{k}$$

hvis der ikke er begrænsninger på hvor mange kugler, der kan være i de enkelte kasser (svarende til Bose-statistik). En kugle i en kasse betyder at denne kasse (element) er udtrukket.

De n kasser stilles på en række, og de k (sorte) kugler fordeles, fx som vist i figur a), hvor $n = k = 10$. De k kugler plus de $n - 1$ skillevægge betragtes under et, angivet som hvide kugler i figur b). En vilkårlig kombination kan findes ved at udvælge k af de hvide kugler i b) til at være sorte kugler og de resterende til at være skillevægge. Et eksempel er vist i figur c). Her er der er 1 kugle i 1. kasse, 0 i 2. kasse, 1 i 3. kasse, 2 i 4. kasse, osv. Antallet af kombinationer er derfor det samme som antallet af måder, vi kan udvælge k af de $n - 1 + k$ hvide kugler i b) til at være sorte kugler, hvilket netop er

$$\binom{n+k-1}{k}$$



4. Moivre's sætning

Sandsynlighedstætheden for binomialfordelingen:

$$f(x) = \binom{n}{x} p^x q^{n-x} \qquad (x = 0, 1, \cdots, n) \tag{4.1}$$

$(q = 1 - p)$ kan betragtes i to grænser:

(I)

For $n \to \infty$ og $p \to 0$, således at $\lambda = np$ holdes konstant går binomialfordelingen mod Poisson-fordelingen: Indføres λ kan $f(x)$ skrives

$$f(x) = \frac{n(n-1)\cdots(n-x+1)}{n^x} \frac{\lambda^x}{x!} \Big(1 - \frac{\lambda}{n}\Big)^n \Big(1 - \frac{\lambda}{n}\Big)^{-x} \tag{4.2}$$

hvor

$$\frac{n(n-1)\cdots(n-x+1)}{n^x} = 1\Big(1 - \frac{1}{n}\Big)\cdots\Big(1 - \frac{x-1}{n}\Big) \to 1 \tag{4.3}$$

for $n \to \infty$ og

$$\begin{aligned} \Big(1 - \frac{\lambda}{n}\Big)^n &= 1 - \frac{n}{1!}\frac{\lambda}{n} + \frac{n(n-1)}{2!}\frac{\lambda^2}{n^2} - \frac{n(n-1)(n-2)}{3!}\frac{\lambda^3}{n^3} + \cdots \\ &= 1 - \lambda + \Big(1 - \frac{1}{n}\Big)\frac{\lambda^2}{2!} - \Big(1 - \frac{1}{n}\Big)\Big(1 - \frac{2}{n}\Big)\frac{\lambda^3}{3!} + \cdots \\ &\to 1 - \lambda + \frac{\lambda^2}{2!} - \frac{\lambda^3}{3!} + \cdots = e^{-\lambda} \end{aligned} \tag{4.4}$$

Sidste faktor i (4.2) går mod 1, dvs.

$$f(x) \to \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{for } n \to \infty \tag{4.5}$$

som er Poisson-sandsynlighedstætheden med middelværdi $\mu = \lambda$.

(II)

Ud fra binomial-fordelingen kan vi nå frem til en kontinuert fordeling ved at fastholde p under grænseovergangen $n \to \infty$. Herved går middelværdien $\mu = np$ og spredningen $\sigma = \sqrt{npq}$ begge mod ∞. Dette nødvendiggør en normering af den oprindelige stokastiske variabel X, som erstattes med

$$Z = \frac{X - \mu}{\sigma} \tag{4.6}$$

I grænsen $n \to \infty$ bliver Z en kontinuert stokastisk variabel, og $f(z = (x-\mu)/\sigma)$ går over i standard normalfordelingen med middelværdien 0 og spredningen 1, dvs. $f(z) \to \phi(z)$, eller når x er heltallig, $F(x) \to \Phi(z = (x + 0.5 - \mu)/\sigma)$ svarende til ligning (11) på side 1090 (1189) i 8. (7.) udgave af bogen (De Moivre og Laplace's grænseværdisætning).

I beviset udnyttes det, at det kun er opførslen af $f(x)$ i en tæt omegn af $x = \mu$, der er afgørende, $f(x) \approx 0$ hvis $|x-\mu| > r\sigma$, hvor r er af størrelsesordenen 5. Det betyder at omegnen relativt set, $r\sigma/\mu = r\sqrt{npq}/np \to 0$ for $n \to \infty$, og dermed at vi dels kan benytte Stirlings formel for alle faktorerne i binomialkoefficienten, n over x, og dels udnytte en rækkeudvikling mht. Δ/μ, hvor $\Delta = x - \mu$:

Anvendes Stirlings formel [ligning (7) side 1067 (1163)] i (4.1) fås:

$$\begin{aligned} f(x) &\approx \frac{\sqrt{2\pi n}n^n e^{-n}}{\sqrt{2\pi x}x^x e^{-x}\sqrt{2\pi(n-x)}(n-x)^{n-x}e^{-n+x}}\, p^x q^{n-x} \\ &= \sqrt{\frac{n}{2\pi x(n-x)}}\left(\frac{np}{x}\right)^x \left(\frac{nq}{n-x}\right)^{n-x} \end{aligned} \tag{4.7}$$

Rækkeudvikles kvadratroden til 0te orden i Δ erstattes den af $1/\sqrt{2\pi npq}$. For det næste led fås:

$$\left(\frac{np}{x}\right)^x = \exp\left[x\ln\left(\frac{\mu}{x}\right)\right] = \exp\left[(\mu+\Delta)\ln\left(\frac{\mu}{\mu+\Delta}\right)\right]$$

hvor

$$\ln\left(\frac{\mu}{\mu+\Delta}\right) = -\ln\left(1+\frac{\Delta}{\mu}\right) \approx -\frac{\Delta}{\mu} + \frac{1}{2}\left(\frac{\Delta}{\mu}\right)^2$$

og dermed

$$\left(\frac{np}{x}\right)^x \approx \exp\left[-\Delta - \frac{\Delta^2}{2\mu}\right] = \exp\left[np - x - \frac{(x-np)^2}{2np}\right] \tag{4.8}$$

og helt analogt

$$\left(\frac{nq}{n-x}\right)^{n-x} \approx \exp\left[nq - (n-x) - \frac{(n-x-nq)^2}{2nq}\right] = \exp\left[\Delta - \frac{\Delta^2}{2nq}\right] \tag{4.9}$$

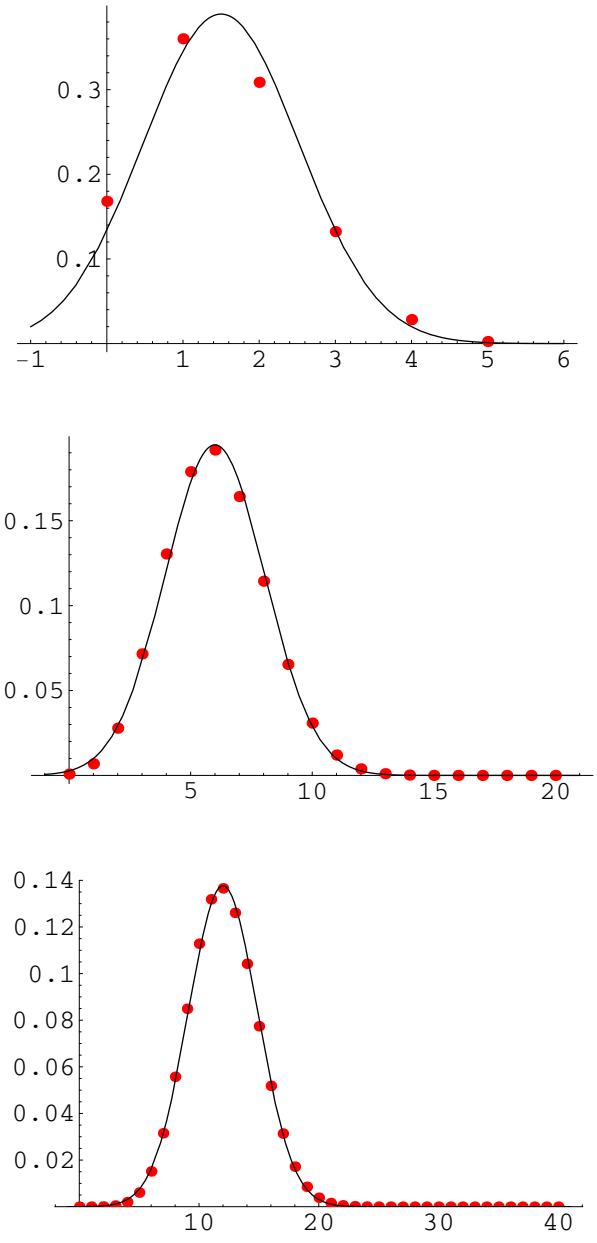
Ganger vi de to faktorer sammen fås

$$\left(\frac{np}{x}\right)^x \left(\frac{nq}{n-x}\right)^{n-x} \approx \exp\left[-\Delta - \frac{\Delta^2}{2np} + \Delta - \frac{\Delta^2}{2n(1-p)}\right] = \exp\left[-\frac{\Delta^2}{2np(1-p)}\right]$$

og når $npq = np(1-p)$ erstattes med σ^2 bliver slutresultatet

$$f(x) \approx \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{4.10}$$

som er gyldigt i grænsen $\mu = np \gg 1$. Det kan tilføjes, at benyttes samme fremgangsmåde kan det vises at sandsynlighedstætheden for Poisson fordelingen for store værdier af μ kan tilnærmes med den normale fordeling (4.10) med $\sigma = \sqrt{\mu}$. Bemærk, at bogens tabel A6 over Poisson-fordelingen kun medtager værdier af $\mu \leq 5$. For større værdier af μ kan man benytte $F(x) \simeq \Phi((x+0.5-\mu)/\sqrt{\mu})$, når x er heltallig.



Figuren viser sandsynlighedstætheden for en binomialfordeling med $p = 0.3$ og $n = 5$, 20 og 40 (punkterne) og den tilsvarende normale fordelingstæthed (de kontinuerte kurver).

5. χ^2-fordeling og Student's t-fordeling

χ^2-fordelingen afhænger kun af antallet af frihedsgrader n, og er defineret ved:

$$Y = \sum_{i=1}^{n} U_i^2 \tag{5.1}$$

hvor U_i, for hvert i, er en uafhængig stokastisk variabel med en frekvensfunktion, som er den standardiserede normale fordelingstæthed:

$$\phi(u_i) = \frac{1}{\sqrt{2\pi}}\, e^{-u_i^2/2}$$

Y er den stokastiske variabel svarende til χ^2 (bemærk, at χ^2 opfattes som en betegnelse for en variabel y som er ikke-negativ). Fordelingsfunktionen $F_n(\chi^2)$ er produktet af frekvensfunktionerne (indbyrdes uafhængige stokastiske variable) integreret over volumenet givet ved $y = \sum u_i^2 \le \chi^2$:

$$F_n(\chi^2) = \int_{y\le\chi^2} du_1 \int du_2 \cdots \int du_n (2\pi)^{-\frac{n}{2}} e^{-y/2} \tag{5.2}$$

Funktionen, der skal integreres, afhænger kun af $r = \sqrt{y}$, som kan opfattes som radien af en n-dimensional kugle. Volumenet af den n-dimensional kugle er $V_n = r^n C_n$, hvor enhedskuglens volumen (se appendix) er

$$C_n = \frac{\pi^{\frac{n}{2}}}{\Gamma(1+\frac{n}{2})} \tag{5.3}$$

og $\Gamma(1+a) = a\Gamma(a)$ er gammafunktionen. Volumenet V_n kan findes ved integration af "overfladearealet" $S_n(r) = dV_n/dr = nr^{n-1}C_n$ mht. r. I ligning (5.2) kan vi uden videre integrere over alle de $n-1$ frihedsgrader, der er uafhængige af $y = r^2$, hvilket giver:

$$F_n(\chi^2) = \int_0^{r=\sqrt{\chi^2}} S_n(r)(2\pi)^{-\frac{n}{2}} e^{-r^2/2} dr \tag{5.4}$$

og dermed, at

$$\begin{aligned} f_n(\chi^2) = F_n'(\chi^2) &= \frac{dF_n}{dr}\,\frac{dr}{d\chi^2} = S_n\Big(\sqrt{\chi^2}\Big)\,(2\pi)^{-\frac{n}{2}} e^{-\chi^2/2}\,\frac{1}{2\sqrt{\chi^2}} \\ &= n\Big(\sqrt{\chi^2}\Big)^{n-1} \frac{\pi^{\frac{n}{2}}}{\Gamma(1+\frac{n}{2})} (2\pi)^{-\frac{n}{2}} e^{-\chi^2/2}\,\frac{1}{2\sqrt{\chi^2}} \end{aligned} \tag{5.5}$$

eller

$$f_n(\chi^2) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\Big(\chi^2\Big)^{\frac{n}{2}-1} e^{-\chi^2/2} \tag{5.6}$$

som er sandsynlighedstætheden for χ^2-fordelingen med n frihedsgrader.

Det kan vises, at middelværdien og variansen af χ^2-fordelingen er henholdsvis n og $2n$:

$$\begin{aligned}\langle\chi^2\rangle &= \int_0^\infty \chi^2 f_n(\chi^2)d\chi^2 = n \\ \sigma^2(\chi^2) = \langle\chi^4\rangle - \langle\chi^2\rangle^2 &= \int_0^\infty \chi^4 f_n(\chi^2)d\chi^2 - n^2 = 2n\end{aligned} \tag{5.7}$$

For at udlede teorem 3 i §23.3 (teorem 4 i §24.6) betragter vi den stokastiske størrelse

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{5.8}$$

hvor X_i er normalt fordelt med middelværdien μ og variansen σ^2, og dermed er

$$n\frac{S_0^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n} U_i^2 \tag{5.9}$$

beskrevet ved χ^2-fordelingen med n-frihedsgrader [højresiden er den samme som i ligning (5.1)]. Indfører vi den stokastiske middelværdi

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{5.10}$$

kan (5.8) omskrives til

$$nS_0^2 = \sum_i \Big[(X_i - \overline{X}) + (\overline{X} - \mu)\Big]^2 = \sum_i (X_i - \overline{X})^2 + n(\overline{X} - \mu)^2 \tag{5.11}$$

idet ligning (5.10) medfører, at produktet mellem de to led i den firkantede parentes summerer til 0. Dermed har vi, at

$$n\frac{S_0^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2 + \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2 \tag{5.12}$$

Ifølge §23.3 teorem 1 (§24.6 teorem 2) er $\overline{X}$ normalfordelt med middelværdien μ og variansen σ^2/n, dvs. at sidste led på højre side af (5.12) netop er kvadratet på en stokastisk variabel, som har en standardiserede normalfordeling, og dermed, at dette led har en χ^2-fordeling med 1 frihedsgrad. Af definitionen (5.1) fremgår det umiddelbart, at summen af to (uafhængige) χ^2-fordelinger med frihedsgraderne n_1 og n_2 er givet ved χ^2-fordelingen med $n_1 + n_2$ frihedsgrader. Det kan vises, at de to led på højre side af (5.12) er statistisk uafhængige (kovariansen er 0), hvoraf følger, at første led har en χ^2-fordeling med $n - 1$ frihedsgrader.

Den stokastiske variabel S^2 defineres ved ligningen:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{\sigma^2}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2 \tag{5.13}$$

Sammenlignes med (5.12) ses, at $(n-1)\dfrac{S^2}{\sigma^2}$ er en stokastisk variabel, som har en χ^2-fordeling med $n - 1$ frihedsgrader (teorem 3).

Konfidensintervallet udledes ud fra

$$P\Big(c_1 \le (n-1)\frac{S^2}{\sigma^2} \le c_2\Big) = F(c_2) - F(c_1) = \gamma \tag{5.14}$$

som er opfyldt med $F(c_1) = (1-\gamma)/2$ og $F(c_2) = (1+\gamma)/2$, hvor fordelingsfunktionen F er χ^2-fordelingen med $n-1$ frihedsgrader (tabel A10). For et enkelt udfald (stikprøve) antager S^2 værdien s^2 og

$$\mathrm{CONF}_\gamma\Big\{c_1 \le (n-1)\frac{s^2}{\sigma^2} \le c_2\Big\}$$

eller

$$\mathrm{CONF}_\gamma\Big\{\frac{n-1}{c_2}\,s^2 \le \sigma^2 \le \frac{n-1}{c_1}\,s^2\Big\} \tag{5.15a}$$

hvor

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad ; \qquad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5.15b}$$

Dette resultat kan formuleres lidt anderledes. Vi kan direkte udnytte at fordelingsfunktionen for S^2 er faktoren $\sigma^2/(n-1)$ gange χ^2-fordelingen med $n-1$ frihedsgrader, dvs.:

$$\langle S^2\rangle = \frac{\sigma^2}{n-1}\,\langle\chi^2\rangle\Big|_{n-1} = \frac{\sigma^2}{n-1}\,(n-1) = \sigma^2 \tag{5.16a}$$

og

$$\begin{aligned}\sigma^2(S^2) &= \langle S^4\rangle - \langle S^2\rangle^2 \\ &= \Big(\frac{\sigma^2}{n-1}\Big)^2\sigma^2(\chi^2)\Big|_{n-1} = \frac{\sigma^4}{(n-1)^2}\,2(n-1) = \frac{2\sigma^4}{n-1}\end{aligned} \tag{5.16b}$$

σ^2, uden argument, er variansen af den oprindelige fordeling for X, og er den størrelse vi forsøger at bestemme ud fra stikprøven sammen med en vurdering af resultatets pålidelighed. Vi skal senere se ("the central limit theorem"), at når $n \gg 1$ kan χ^2-fordelingen tilnærmes med normalfordelingen med samme middelværdi og varians. Dette kan udnyttes til at sige, at hvis $n \gg 1$ vil den stokastiske variabel S^2 være beskrevet ved normalfordelingen med middelværdien σ^2 og variansen $2\sigma^4/(n-1)$, eller efter lidt regning

$$\mathrm{CONF}_\gamma\left\{\frac{s^2}{1+c_0\sqrt{\frac{2}{n-1}}} \le \sigma^2 \le \frac{s^2}{1-c_0\sqrt{\frac{2}{n-1}}}\right\} \quad ; \quad \Phi(c_0) = \frac{\gamma+1}{2} \quad ; \quad n \gg 1 \tag{5.17a}$$

eller udtrykt på en mere forenklet, og tilnærmet ($n \ge 100$), måde:

$$\sigma^2 \simeq s^2 \pm c_0\sqrt{\frac{2}{n-1}}\,s^2 \tag{5.17b}$$

hvor $c_0 = 1.96 \simeq 2$, hvis vi benytter et 95% konfidensniveau.

Student's t-fordeling:

Vi har netop betragtet den stokastiske variabel

$$U = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

hvor U^2 er χ^2-fordelt med 1 frihedsgrad. For at vurdere pålidelighedsintervallet ved en bestemmelsen af μ ud fra stikprøveværdien af variansen betragter vi nu i stedet

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = U\frac{\sigma}{S} \tag{5.18}$$

Ifølge (5.13) er

$$\frac{S^2}{\sigma^2} = \frac{Y_p}{p} \tag{5.19}$$

hvor Y_p har en χ^2-fordeling med $p = n - 1$ frihedsgrader. Dvs. vi har

$$\frac{T^2}{p} = \frac{Y_1}{Y_p} \qquad \text{med} \qquad Y_1 = U^2 \tag{5.20}$$

og dermed er fordelingsfunktionen $G(t^2/p)$ for den stokastiske variabel T^2/p bestemt, som arealet af $f_1(y_1)f_p(y_p)$ over området, hvor $y_1/y_p \leq t^2/p$

$$G(t^2/p) = \int_{y_p=0}^{\infty}\int_{y_1=0}^{y_p t^2/p} f_1(y_1)f_p(y_p)dy_1dy_p \tag{5.21}$$

Definitionen af fordelingsfunktioner medfører, at $G(t^2/p) = \pm[F_t(t) - F_t(-t)]$, hvor $+$ $(-)$ svarer til t positiv (negativ). Dermed kan t-frekvensfunktionen, $f_t(t) = F_t'(t)$, bestemmes som

$$2f_t(t) = \pm\frac{d(t^2/p)}{dt}\frac{dG(t^2/p)}{d(t^2/p)} = (2|t|/p)\int_0^{\infty}\Big[f_1(y_1)\Big]_{y_1=y_pt^2/p} y_p f_p(y_p)dy_p \tag{5.22}$$

og benyttes ligning (5.6) fås

$$\begin{aligned} f_t(t) &= (|t|/p)\int_0^{\infty}\Big[\frac{1}{\sqrt{2\pi}}\,y_1^{-\frac{1}{2}}e^{-y_1/2}\Big]_{y_1=y_pt^2/p} y_p\frac{1}{2^{\frac{p}{2}}\Gamma(\frac{p}{2})}(y_p)^{\frac{p}{2}-1}\,e^{-y_p/2}dy_p \\ &= \frac{|t|}{p\sqrt{\pi}}\sqrt{\frac{p}{t^2}}\frac{1}{2^{\frac{p+1}{2}}\Gamma(\frac{p}{2})}\int_0^{\infty}(y_p)^{\frac{p-1}{2}}\,e^{-(1+\frac{t^2}{p})y_p/2}dy_p \end{aligned} \tag{5.23}$$

hvor integralet udregnes til $\Big[\frac{2}{1+t^2/p}\Big]^{\frac{p+1}{2}}\Gamma(\frac{p+1}{2})$, og dermed slutresultatet

$$f_t(t) = \frac{1}{\sqrt{p\pi}}\frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2})}\Big(1+\frac{t^2}{p}\Big)^{-\frac{p+1}{2}} \tag{5.24}$$

som er frekvensfunktionen for T i ligning (5.18) med $p = n - 1$. Hermed har vi udledt teorem 2 i §23.3 (teorem 3 i §24.6).

Idet frekvensfunktionen er lige, $f_t(t) = f_t(-t)$, har vi at

$$P\Big(-c \le T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \le c\Big) = F_t(c) - F_t(-c) = 2F_t(c) - 1 = \gamma \qquad (5.25)$$

således at $F_t(c) = (1+\gamma)/2$, hvor F_t er Student's t-fordeling med $n-1$ frihedsgrader (tabel A9). Konfidensintervallet er derefter bestemt ved

$$\mathrm{CONF}_\gamma\Big\{-c \le \frac{\overline{x} - \mu}{s/\sqrt{n}} \le c\Big\} \quad \text{eller} \quad \mathrm{CONF}_\gamma\Big\{\overline{x} - \frac{cs}{\sqrt{n}} \le \mu \le \overline{x} + \frac{cs}{\sqrt{n}}\Big\} \qquad (5.26)$$

En nærmere analyse af χ^2 og Student's t fordelingerne kan findes som Mathematica programmer på hjemmesiden. Teorem 4 i §23.3 (teorem 5 i §24.6), det såkaldte "Central limit theorem", fortæller at begge fordelinger nærmere sig normalfordelingen, når antallet af frihedsgrader vokser. Som nævnt har χ^2-fordelingen med n frihedsgrader middelværdien $\mu = n$ og variansen $\sigma^2 = 2n$. For Student's t-fordeling med p frihedsgrader fås:

$$\langle T\rangle = 0 \qquad ; \qquad \sigma^2(T) = \langle T^2\rangle - \langle T\rangle^2 = \frac{p}{p-2} \quad (p \ge 3) \qquad (5.27)$$

Variansen divergerer og er dermed ikke defineret for $p = 1$ og $p = 2$. Antager vi igen at $n \gg 1$, så betyder "central limit" teoremet at $F_t(c) \simeq \Phi(c/\sigma(T))$. Indføres $c_0 = c/\sigma(T)$ fås at $\sigma(T) \simeq 1$, og dermed at $c_0 \simeq c$ i grænsen $p = n - 1 \gg 1$, og (5.26) kan tilnærmes med

$$\mathrm{CONF}_\gamma\Big\{\overline{x} - \frac{c_0 s}{\sqrt{n}} \le \mu \le \overline{x} + \frac{c_0 s}{\sqrt{n}}\Big\} \quad ; \quad \Phi(c_0) = \frac{\gamma+1}{2} \quad ; \quad n \gg 1 \qquad (5.28a)$$

eller

$$\mu \simeq \overline{x} \pm c_0 \frac{s}{\sqrt{n}} \qquad (5.28b)$$

Reultatet kan forbedres (ubetydeligt) ved at erstatte c_0 med $c = c_0\sqrt{\frac{n-1}{n-3}}$. Bemærk, at når $n \gg 1$ er resultatet det samme, som hvis s kan erstattes med σ, svarende til teorem 1. Dvs., at i denne grænse er usikkerheden på bestemmelsen af σ ud fra s uden betydning for vurderingen af hvor pålideligt μ bestemmes af $\overline{x}$.

Appendix

Enhedskuglens volumen i n dimensioner, C_n, kan bestemmes ved at udregne integralet

$$I_n = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-(x_1^2+x_2^2+\cdots+x_n^2)} dx_1 dx_2 \cdots dx_n \tag{5.29}$$

dels direkte, som et produkt af n Gauss integraler:

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi} \qquad \Rightarrow \qquad I_n = \pi^{\frac{n}{2}} \tag{5.30}$$

dels ved at udnytte at funktionen under integraltegnene kun afhænger af den radiære variabel $r \equiv (x_1^2 + x_2^2 + \cdots + x_n^2)^{1/2}$.

Volumenet af en n-dimensional kugle med radius r er

$$V_n(r) = \int \cdots \int dx_1 dx_2 \cdots dx_n = \int_0^r S_n(r')dr' = r^n C_n \tag{5.31}$$

og dermed $dV_n = S_n(r)dr = n\, r^{n-1} C_n dr$. Det betyder, at udføres alle de $n-1$ integrationer i (5.29), der er uafhængige af r, fås

$$\begin{aligned} I_n &= \int_0^{\infty} S_n(r)\, e^{-r^2} dr = \int_0^{\infty} n r^{n-1} C_n\, e^{-r^2} dr \\ &= n C_n \tfrac{1}{2} \int_0^{\infty} z^{\frac{n}{2}-1} e^{-z} dz = \tfrac{n}{2} C_n \Gamma(\tfrac{n}{2}) = C_n \Gamma(1+\tfrac{n}{2}) \end{aligned} \tag{5.32}$$

Ved at sammenholde de to resultater for I_n, fås

$$C_n = \frac{\pi^{\frac{n}{2}}}{\Gamma(1+\frac{n}{2})} \tag{5.33}$$

6. "Central limit" teoremet og ophobningsloven

Vi skal supplere bogen med nogle vigtige resultater. Vi vil starte med at betragte den moment-genererende funktion, som indføres i §22.7 opgave 14 (§23.5 opgave 20 og §23.6 opgave 16–19):

$$G(t) = \langle e^{tX} \rangle = \begin{cases} \int_{-\infty}^{\infty} e^{tx} f(x) dx \\ \sum_i e^{tx_i} f(x_i) \end{cases} \tag{6.1}$$

som har egenskaben:

$$\langle X^n \rangle = \left(\frac{d^n G(t)}{dt^n} \right)_{t=0}$$

For de to diskrete fordelinger, henholdsvis binomialfordelingen ($\mu = np$, $\sigma^2 = npq$) og Poisson-fordelingen, har vi fra bogens opgaver, at

$$G_{\text{bin}}(t) = (pe^t + q)^n \qquad \text{og} \qquad G_P(t) = e^{-\mu}\, e^{\mu e^t} \tag{6.2}$$

For Gauss-fordelingen er:

$$G(t) = \langle e^{tX} \rangle = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\left[tx - \frac{(x-\mu)^2}{2\sigma^2}\right]} dx$$

hvor eksponenten kan omskrives til

$$tx - \frac{(x-\mu)^2}{2\sigma^2} = \mu t + \tfrac{1}{2}\sigma^2 t^2 - \frac{(x - \mu - \sigma^2 t)^2}{2\sigma^2}$$

og dermed, at

$$G(t) = e^{\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)} \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu-\sigma^2 t)^2}{2\sigma^2}} dx$$

Gauss-integralet er uafhængigt af at μ er erstattet af $\mu + \sigma^2 t$, og resultatet for den moment-genererende funktion for normalfordelingen er:

$$G(t) = e^{\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)} \tag{6.3}$$

Generelt har vi, at hvis Y er en linearkombination af to *uafhængige* stokastiske variable X_1 og X_2,

$$Y = a_1 X_1 + a_2 X_2$$

så er den moment-genererende funktion for y-fordelingen produktet af de to moment-genererende funktioner $G_1(a_1 t)$ og $G_2(a_2 t)$ for x_1- og x_2-fordelingerne:

$$G_y(t) = \langle e^{tY} \rangle = \langle e^{ta_1 X_1} e^{ta_2 X_2} \rangle = \langle e^{ta_1 X_1} \rangle \langle e^{ta_2 X_2} \rangle = G_1(a_1 t) G_2(a_2 t) \tag{6.4}$$

Disse to resultater, (6.3 og (6.4), kan benyttes til at eftervise teorem 1 i §23.3 (teorem 1–2 i §24.6):

Summen af to uafhængige stokastiske variable, $Y = X_1 + X_2$, som begge har normalfordelinger med middelværdi μ_1, μ_2 og varians σ_1^2, σ_2^2 har normalfordelingen med $\mu = \mu_1 + \mu_2$ og $\sigma^2 = \sigma_1^2 + \sigma_2^2$.

Ifølge (6.4) er

$$G_y(t) = G_1(t)G_2(t) = e^{(\mu_1 t + \frac{1}{2}\sigma_1^2 t^2)} e^{(\mu_2 t + \frac{1}{2}\sigma_2^2 t^2)} = e^{\left[(\mu_1+\mu_2)t + \frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2\right]} \tag{6.5}$$

som er den moment-genererende funktion for den angivne y-fordeling.

[Indskud:

Frekvensfunktionen for $Y = X_1 + X_2$, bestemt ved *vilkårlige* (kontinuerte) frekvensfunktioner, $f_1(x)$ og $f_2(x)$, kan skrives som et *foldningsintegral*: $F(y)$ er produktet af de to frekvensfunktioner integreret over arealet $y' = x_1 + x_2 \leq y$ eller $x_2 \leq y - x_1$, dvs.:

$$F(y) = \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{y-x_1} f_1(x_1) f_2(x_2) dx_1 dx_2 = \int_{-\infty}^{\infty} f_1(x_1) F_2(y - x_1) dx_1$$

eller

$$f(y) = F'(y) = \int_{-\infty}^{\infty} f_1(x) f_2(y - x) dx \tag{6.6}$$

Foldningsintegralet optræder fx når en resonanskurve, $f_1(x)$, udmåles med en eksperimentel usikkerhed beskrevet ved en *opløsningsfunktion*, $R(x) = f_2(x)$ med $\mu_2 = 0$. Antages begge funktioner at være (eller kunne tilnærmes med) Gauss-funktioner, bliver resultatet en Gauss-funktion centreret omkring μ_1 med variansen $\sigma_1^2 + \sigma_2^2$.]

Vi har tidligere vist, at binomialfordelingen for store værdier af n kan tilnærmes med normalfordelingen (med $\mu = np$ og $\sigma^2 = npq$), og tilsvarende for Poissonfordelingen (for $\mu \geq 5$). Resultaterne (6.3) og (6.4) kan nu benyttes til at bevise den generelle sætning **"the central limit theorem"**:

Antag X_i, $i = 1, \ldots, n$ er uafhængige stokastiske variable, som beskrives ved frekvensfunktionerne $f_i(x_i)$ (der alle kan være forskellige) med middelværdi μ_i og varians σ_i^2. Den stokastiske variabel

$$Y_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{6.7}$$

har følgende egenskaber:

(i) Middelværdien er $\mu = \langle Y_n \rangle = \dfrac{1}{n} \sum_i \mu_i$

(ii) Variansen er $\sigma^2 = \dfrac{1}{n^2} \sum_i \sigma_i^2$

(iii) Fordelingsfunktionen for $Z_n = \dfrac{Y_n - \mu}{\sigma}$ *vil (asymptotisk)* $\to \Phi(z)$ *(den standardiserede normalfordeling) for* $n \to \infty$.

Bemærk, at teoremet i denne udgave er mere generelt end lærebogens version, teorem 4 i §23.3 (teorem 5 i §24.6). Punkterne (i) og (ii) følger direkte af teorem 1 og 2 i §22.9 (§23.8). Punkt (iii) kan bevises ved at udnytte (6.4), generaliseret til n led med $a_i = 1/n$, dvs.

$$G_y(t) = \prod_{i=1}^{n} G_i(t/n)$$

En Taylor-rækkeudvikling af $G_i(t/n) = G_i(0) + G_i'(0)(t/n) + \frac{1}{2}G_i''(0)(t/n)^2 \cdots$ giver nærmest definitionsmæssigt:

$$G_i\Big(\frac{t}{n}\Big) = 1 + \mu_i\frac{t}{n} + \tfrac{1}{2}(\sigma_i^2 + \mu_i^2)\frac{t^2}{n^2} + \cdots = \exp\Big[\mu_i\frac{t}{n} + \tfrac{1}{2}\sigma_i^2\frac{t^2}{n^2}\Big] + \mathcal{O}\Big(\frac{1}{n^3}\Big)$$

og dermed, at for $n \gg 1$:

$$G_y(t) \approx \prod_{i=1}^{n} \exp\Big[\mu_i\frac{t}{n} + \tfrac{1}{2}\sigma_i^2\frac{t^2}{n^2}\Big] = \exp\Big[\frac{1}{n}\sum_i \mu_i\, t + \tfrac{1}{2}\frac{1}{n^2}\sum_i \sigma_i^2 t^2\Big]$$

som netop, ifl. (6.3), er den moment-genererende funktion for normalfordelingen med middelværdi μ (i) og varians σ^2 (ii).

Ophobningsloven kan betragtes som et korrolar til "the central limit theorem". Er Y en generel funktion af X_i $(i = 1, \ldots, n)$ kan vi tilnærmelsesvis erstatte funktionen med et lineært udtryk i et område omkring X_i'ernes middelværdier:

$$Y = \psi(X_1, \ldots, X_n) \approx \psi\Big(\langle X_1\rangle, \ldots, \langle X_n\rangle\Big) + \sum_{i=1}^{n}\Big(\frac{\partial\psi}{\partial X_i}\Big)_{X_j=\langle X_j\rangle}\Big(X_i - \langle X_i\rangle\Big)$$

Hvis de forskellige variable har en Gauss-fordeling, eller hvis der er mange forskellige bidrag (usikkerhedsberegninger!), vil Y, ifl. henholdsvis (6.5) og (6.7), (tilnærmelsesvis) have en normalfordeling med middelværdien

$$\langle Y\rangle = \psi\Big(\langle X_1\rangle, \ldots, \langle X_n\rangle\Big) \tag{6.8a}$$

og variansen

$$\sigma^2 = \sum_i \Big(\frac{\partial\psi}{\partial X_i}\Big)^2_{X_j=\langle X_j\rangle}\sigma_i^2 \tag{6.8b}$$

---

7. Regression ("Mindste kvadraters metode")

En stikprøvemåling af Y som funktion af x giver n talpar $(x_1, y_1), \ldots, (x_n, y_n)$. Det antages, at der ikke er nogen usikkerhed forbunden med målingen af x (almindelig variabel), mens Y er en stokastisk variabel der for en bestemt, fastholdt x-værdi har en normalfordeling med middelværdien $\langle Y \rangle = \mu(x) = \kappa_0 + \kappa_1 x$ og variansen σ_y^2. Funktionen $\mu(x)$ kaldes for regressionskurven til Y som funktion af x, heraf navnet på den analyse vi skal foretage.

Variansen σ_y^2 antages at være uafhængig af x, og vi ønsker at bestemme de to konstanter, κ_0 og κ_1, ud fra de n talpar i stikprøven og at vurdere pålideligheden af resultatet, når de to parametre erstattes med de tilsvarende stikprøveværdier k_0 og k_1. Indføres betegnelsen

$$\overline{y}_i = k_0 + k_1 x_i \tag{7.1}$$

er opgaven at bestemme k_0 og k_1, således at værdierne $\overline{y}_i \approx y_i$ for alle i. "Maximum likelihood" metoden fører til resultatet, at de mest sandsynlige værdier af de to parametre bestemmes ved at minimere summen af de kvadratiske afvigelser $\sum(y_i - \overline{y}_i)^2$, dvs. ved at minimere funktionen

$$q = \sum_{i=1}^{n} \left(y_i - k_0 - k_1 x_i\right)^2 \tag{7.2}$$

mht. k_0 og k_1. Minimumet er bestemt af ligningerne:

$$\frac{\partial q}{\partial k_0} = -2\sum_{i=1}^{n}(y_i - k_0 - k_1 x_i) = 0$$
$$\frac{\partial q}{\partial k_1} = -2\sum_{i=1}^{n}(y_i - k_0 - k_1 x_i)x_i = 0$$

eller

$$\begin{aligned} k_0 n + k_1 \sum x_i &= \sum y_i \\ k_0 \sum x_i + k_1 \sum x_i^2 &= \sum x_i y_i \end{aligned} \tag{7.3}$$

hvor alle summer går fra $i = 1$ til $i = n$. Resultatet af denne "mindste kvadraters metode" er, i overensstemmelse med §23.9 eller §18.5 (§24.12 eller §19.5),

$$\begin{aligned} D &= n\sum x_i^2 - \left(\sum x_i\right)^2 \\ k_0 &= \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{D} \\ k_1 &= \frac{n\sum x_i y_i - \sum x_i \sum y_i}{D} \end{aligned} \tag{7.4}$$

Den rette linie bestemt af disse ligninger går igennem måledataernes tyngdepunkt $(x_T, y_T) = \left(\frac{1}{n}\sum x_i, \frac{1}{n}\sum y_i\right)$, dvs.:

$$y_T = k_0 + k_1 x_T \qquad \text{og} \qquad \langle \overline{y}_i \rangle = \frac{1}{n}\sum \overline{y}_i = y_T \tag{7.5}$$

Hvis vi erstatter y_i i ligning (7.4) med den tilsvarende stokastiske variabel $Y = Y_i$ for $x = x_i$ har vi at de stokastiske variable (K_0 og K_1) svarende til k_0 og k_1 er funktioner af de forskellige Y_i, og vi kan benytte ophobningsloven (6.8) til at bestemme stikprøveværdierne af varianserne for κ_0 og κ_1, som vi skal betegne $s^2(k_0)$ og $s^2(k_1)$.

I en eksperimentel situation vil standardafvigelsen af y_i (σ_y) ofte være ukendt, men antager vi, at den er den samme for alle målepunkterne (alle i) kan det vises, at $\sum(Y_i - \overline{Y}_i)^2/\sigma_y^2$ har en χ^2-fordelingen med $n-2$ frihedsgrader. Beviset benytter en omskrivning af $\sum[Y_i - \mu(x_i)]^2/\sigma_y^2$ helt parallelt med ligning (5.12). Frihedsgraderne reduceres her med 2, fordi vi skal benytte to parametre, k_0 og k_1, til at bestemme $\mu(x_i)$ ud fra stikprøven. Analogt med (5.15), eller (5.17), har vi derfor at $\sigma_y \simeq s(y)$, hvor

$$\boxed{s^2(y) = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \overline{y}_i)^2 \qquad ; \quad \overline{y}_i = k_0 + k_1 x_i} \tag{7.6}$$

Hvis vi ønsker at vurdere, hvor stor en tiltro vi kan have til resultatet, er fremgangsmåden den samme som i ligningerne (5.14)–(5.17) med den eneste forskel at $n-1$ skal erstattes med $n-2$. Dvs. groft set (95% konfidens) er $\sigma_y^2 \simeq s^2(y)[1 \pm \sqrt{8/(n-2)}]$, når n er noget større end 10.

Efter at $s(y)$ er bestemt kan ophobningsloven (6.8) udnyttes til at beregne stikprøve-variansen af k_0:

$$\begin{aligned} s^2(k_0) &= \sum\left(\frac{\partial k_0}{\partial y_i}\right)^2 s^2(y) = \sum_i \left(\frac{\sum x_j^2 - x_i\sum x_j}{D}\right)^2 s^2(y) \\ &= \left[n\Big(\sum x_j^2\Big)^2 - 2\Big(\sum x_j\Big)^2\sum x_j^2 + \Big(\sum x_j\Big)^2\sum x_j^2\right]\frac{s^2(y)}{D^2} \end{aligned}$$

eller

$$\boxed{s^2(k_0) = \frac{\sum x_j^2}{D}s^2(y)} \tag{7.7}$$

Helt analogt kan variansen af k_1 udregnes til:

$$\boxed{s^2(k_1) = \frac{n}{D}s^2(y)} \tag{7.8}$$

Konfidensintervallerne for κ_0 og κ_1 kan herefter bestemmes på samme måde, som benyttet ved udregningen af konfidensintervallet for μ, (5.25)–(5.28). Parameteren c, som angiver forholdet mellem den halve længde af konfidensintervallet og standardafvigelsen, er givet ved

$$\text{Student's } t\text{-fordeling med } n-2 \text{ frihedsgrader:} \qquad F_t(c) = \frac{1+\gamma}{2} \tag{7.9}$$

og derefter har vi, at

$$\mathrm{CONF}_\gamma\Big\{k_1 - c\,s(k_1) \le \kappa_1 \le k_1 + c\,s(k_1)\Big\} \tag{7.10}$$

og det helt analoge udtryk for κ_0. K_0 og K_1 er ikke stokastisk uafhængige. Det er derimod tilfældet for K_1 og "tyngdepunktsværdien" $Y_T = (1/n)\sum Y(x_i)$ med den teoretiske middelværdi $\mu_T = (1/n)\sum \mu(x_i)$. Der kan derfor opnås en bedre karakteristik af regressionskurven ved at benytte stikprøveværdierne k_1 og y_T i stedet for k_1 og k_0. Konfidensintervallet for μ_T er bestemt af samme Student's t-fordeling (samme c) som ovenfor, og idet $y_T = (1/n)\sum y_i$ fås

$$\mathrm{CONF}_\gamma\Big\{y_T - c\,s(y_T) \le \mu_T \le y_T + c\,s(y_T)\Big\}, \qquad s(y_T) = \frac{s(y)}{\sqrt{n}} \tag{7.11}$$

Man kan sige, at det er tilstrækkeligt at angive stikprøveværdierne af standardafvigelserne, $s(y)$, $s(k_0)$, $s(k_1)$, og $s(y_T)$, idet vurderingerne af hvor godt σ_y, κ_0, κ_1 og μ_T er bestemt, med god tilnærmelse, kan afledes af normalfordelingen, hvis $n \ge 10$. Dermed ved vi for eksempel, at c i (7.10)–(7.11) er ca. 2, hvis vi forlanger 95% konfidens, eller 1, hvis vi nøjes med 68,26% konfidens.

Fremstillingen ovenfor kan generaliseres på forskellige måder:

I) Lineær regression med polynomier af højere grad end 1: Anvendes mindste kvadraters metode på en stikprøve, hvor regressionskurven er et andengrads polynomium

$$\overline{y}_i = k_0 + k_1 x_i + k_2 x_i^2$$

fås følgende ligningssystem:

$$\begin{aligned} k_0 n + k_1 \sum x_i + k_2 \sum x_i^2 &= \sum y_i \\ k_0 \sum x_i + k_1 \sum x_i^2 + k_2 \sum x_i^3 &= \sum x_i y_i \\ k_0 \sum x_i^2 + k_1 \sum x_i^3 + k_2 \sum x_i^4 &= \sum x_i^2 y_i \end{aligned} \tag{7.12}$$

se bogens §18.5 (§19.5 i 7. udgave). Bemærk, at udtrykket er lineært i regressionsparametrene, k_0, k_1 og k_2.

II) Hvis $\sigma_y = \sigma_y(x)$ afhænger af x betyder det, at den funktion der skal minimeres for at bestemme de mest sandsynlige værdier af k_0 og k_1, er:

$$q = \sum_{i=1}^{n} \left(\frac{y_i - \overline{y}_i}{\sigma_y(x_i)}\right)^2 = \sum_{i=1}^{n} \left(\frac{y_i - k_0 - k_1 x_i}{\sigma_y(x_i)}\right)^2 \tag{7.13}$$

Den generelle, ikke-lineære regressionsmetode som præsenteres nedenfor i punkt IV er anvendelig i dette tilfælde. Her foretages minimeringen ved iteration.

III) Lineær regression i det tilfælde hvor både X og Y er stokastiske variable. Det antages at for en fastholdt værdi af parameteren i vil X og Y være stokastisk uafhængige og begge normalfordelte med middelværdierne $\mu_x(i)$ og $\mu_y(i)$ og

standardafvigelserne $\sigma_x(i)$ og $\sigma_y(i)$. Standardafvigelserne antages uafhængige af i og vi indfører forholdet $\xi = \sigma_x/\sigma_y$. Regressionskurven antages lineær $\mu_y = \kappa_0 + \kappa_1\mu_x$, og udtrykt vha. stikprøveværdierne er opgaven at finde de bedst mulige værdier af k_0 og k_1 i ligningen

$$\overline{y}_i = k_0 + k_1\overline{x}_i \tag{7.14}$$

således at $(\overline{x}_i, \overline{y}_i) \approx (x_i, y_i)$ for alle i. I dette tilfælde er funktionen der skal minimeres:

$$q = \sum_{i=1}^{n}\left[\left(\frac{y_i-\overline{y}_i}{\sigma_y}\right)^2 + \left(\frac{x_i-\overline{x}_i}{\sigma_x}\right)^2\right] = \frac{1}{\sigma_y^2}\sum_{i=1}^{n}\left[(y_i-k_0-k_1\overline{x}_i)^2 + \left(\frac{x_i-\overline{x}_i}{\xi}\right)^2\right] \tag{7.15}$$

Her er der $2+n$ ubekendte størrelser k_0, k_1, og $\overline{x}_i$ (alle i). q kan minimeres mht. $\overline{x}_i$ ved en geometrisk betragtning: transformér til et koordinatsystem, hvor x erstattes af x/ξ, og bestem den korteste afstand mellem punktet $(x_i/\xi, y_i)$ og et punkt $(\overline{x}_i/\xi, \overline{y}_i)$ på den rette linie $\overline{y}_i = k_0 + \xi k_1\overline{x}_i/\xi$. Resultatet er

$$q(k_0,k_1) = \mathrm{Min}\left[q(k_0,k_1,\overline{x}_i)\sigma_y^2\right] = \frac{1}{1+(\xi k_1)^2}\sum_{i=1}^{n}(y_i-k_0-k_1x_i)^2 \tag{7.16}$$

Indføres følgende parametre:

$$D_{\alpha\beta} = n\sum\alpha_i\beta_i - \sum\alpha_i\sum\beta_i \quad ; \qquad \alpha_T = \frac{1}{n}\sum\alpha_i \tag{7.17}$$

hvor α og β betyder enten x eller y, så er $q(k_0,k_1)$ minimal når

$$\begin{aligned} k_1 &= \frac{1}{\xi}\left[-F \pm \sqrt{1+F^2}\,\right]; \qquad F = \frac{D_{xx}-\xi^2D_{yy}}{2\xi D_{xy}} \\ k_0 &= y_T - k_1x_T \end{aligned} \tag{7.18}$$

Fortegnet foran kvadratroden er bestemt ved, at k_1 har samme fortegn som D_{xy} (og r). Bemærk at linien går gennem tyngdepunktet, se (7.5). Varianserne udregnes på samme måde som i (7.6)–(7.8):

$$\begin{aligned} s^2(y) &= \frac{1}{n-2}\sum_{i=1}^{n}\frac{(y_i-k_0-k_1x_i)^2}{1+(\xi k_1)^2} = \frac{1+a^2-2ar}{n(n-2)[1+(\xi k_1)^2]}\,D_{yy} \\ s^2(k_1) &= \frac{D_{xx}+\xi^2D_{yy}}{D_{xy}^2}\,nk_1^2s^2(y) = \frac{k_1^2}{n-2}\left[\frac{2}{r^2} - \frac{1}{r}\left(a+\frac{1}{a}\right)\right] \\ s^2(k_0) &= x_T^2s^2(k_1) + \frac{1}{n}[1+(\xi k_1)^2]s^2(y)\,; \qquad s^2(x) = \xi^2s^2(y) \end{aligned} \tag{7.19}$$

Korrelationskoefficienten (§23.10 i 8. udgave af Kreyszig) er defineret

$$r = \frac{s_{xy}}{s_xs_y} = \frac{D_{xy}}{\sqrt{D_{xx}D_{yy}}} \qquad \text{og} \qquad a = k_1\sqrt{\frac{D_{xx}}{D_{yy}}} \tag{7.20}$$

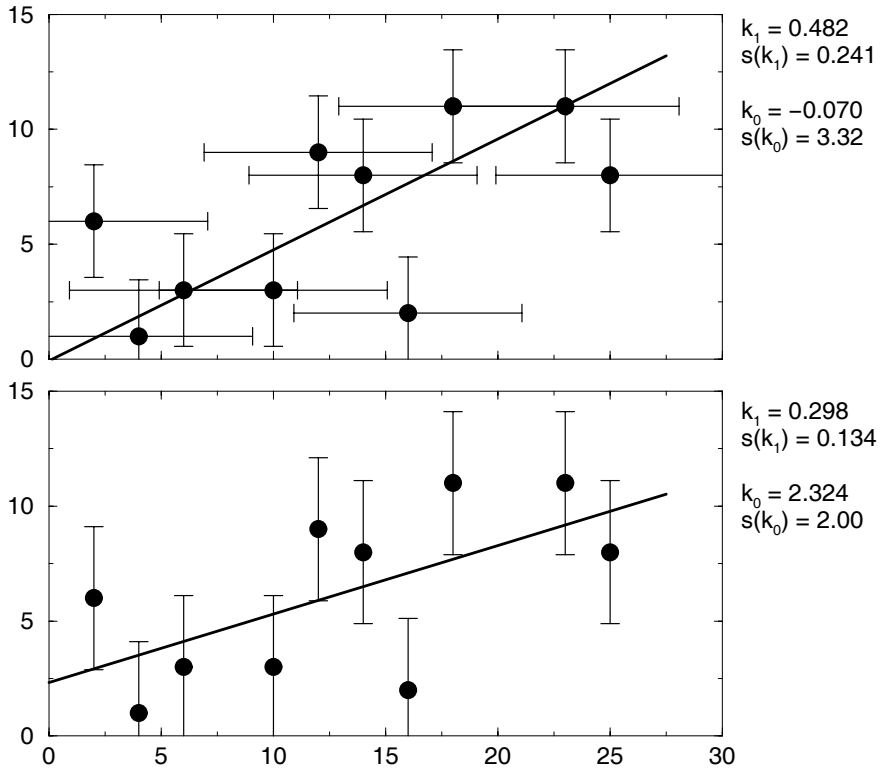
Vi kan skelne mellem tre hovedtilfælde:

i) $(\xi k_1)^2 \ll 1 \quad \Rightarrow \quad k_1 = \dfrac{D_{xy}}{D_{xx}}$; ii) $(\xi k_1)^2 \gg 1 \quad \Rightarrow \quad k_1 = \dfrac{D_{yy}}{D_{xy}}$; eller

iii) $(\xi k_1)^2 \approx 1 \quad \Rightarrow \quad k_1^2 = \dfrac{D_{yy}}{D_{xx}} = k_1(\text{i})k_1(\text{ii})$ ("geometrisk middelværdi").

Hvis korrelationskoefficienten $r \approx \pm 1$ bliver resultatet for regressionslinien uafhængigt af ξ og er dermed uafhængigt af om spredningen skyldes usikkerhed på målingen af x eller y eller af begge størrelser. Med den forudsætning, at der er en lineær sammenhæng mellem x og y, kan vi omvendt slutte, at det er kun hvis $|r|$ ikke er tæt på 1, at det er afgørende at vurdere om usikkerheden på x er betydende ($\xi|k_1| \approx 1$) eller ej.

I figurerne nedenfor analyseres måledatasættet $(x, y) = (2, 6),\ (4, 1),\ (6, 3),\ (10, 3),\ (12, 9),\ (14, 8),\ (16, 2),\ (18, 11),\ (23, 11)$, og $(25, 8)$. Korrelationskoefficienten for disse 10 talsæt er $r = 0.618$ og den nederste figur viser resultatet når spredningen alene skyldes usikkerhed på y, hvorimod den øverste viser resultatet når det antages, at usikkerhederne på x og y er lige betydende [tilfælde iii) $\xi k_1 = 1$]. Begge regressionslinier går gennem tyngdepunktet $(x_T, y_T) = (13, 6.2)$, men har forskellige hældninger. Liniestykkerne på hver side af målepunkterne har længderne: $s(x) = 5.08,\ s(y) = 2.45$ i øverste figur og $s(y) = 3.11$ i nederste. Når standardafvigelserne som her er bestemt af måleresultaterne, vil regressionskurven skære liniestykkerne omkring målepunkterne i ca. 2/3 af tilfældene.



IV) Generel, ikke-lineær regression. Vi starter med forudsætningerne, at vi har en stikprøve med n talpar $(x_1, y_1), \ldots, (x_n, y_n)$, hvor der ikke er nogen usikkerhed forbunden med x. Regressionskurven $\langle Y \rangle = \mu(x)$ antages at være er en generel funktion af x, specificeret ved hjælp af p parametre $\beta_1, \ldots, \beta_p$,

$$\langle Y \rangle = g(x; \beta_1, \ldots, \beta_p) = g(x; \underline{\beta}) \tag{7.14}$$

hvor $\underline{\beta}$ betegner den p dimensionale vektor $(\beta_1, \ldots, \beta_p)$. Funktionen, der skal minimeres er:

$$q = \sum_{i=1}^{n} \left(\frac{y_i - g(x_i; \underline{\beta})}{s_i} \right)^2 \tag{7.15}$$

hvor s_i er standardafvigelsen på målingen af Y for $x = x_i$ (bestemt eksperimentelt eller vurderet på anden måde). Hvis der er mere end et par parametre vil dette udtryk kunne udvikle mange lokale minima, og det vil være svært at finde det "rigtige", som ikke nødvendigvis er det absolutte minimumspunkt. Metoden anvender en iteration, dvs. vi starter ud med at gætte en løsning $\underline{\beta}_0$ og derefter at undersøge opførslen af q i nærheden af dette punkt. For at gøre dette udnyttes en rækkeudvikling af $g(x_i; \underline{\beta})$ til første orden i $\underline{\beta} - \underline{\beta}_0$, og følgende $n \times p$ matrix opstilles:

$$\{\boldsymbol{A}\}_{ij} = \left. \frac{\partial g(x_i; \underline{\beta})}{\partial \beta_j} \right|_{\underline{\beta} = \underline{\beta}_0} \tag{7.16}$$

og $\boldsymbol{A}^T$ betegner den transponerede matrix. For at opskrive resultatet skal vi indføre

$$\{\boldsymbol{L}\}_{ij} = \frac{\delta_{ij}}{s_i^2} \quad ; \quad \underline{\Delta y} = \Big(y_1 - g(x_1; \underline{\beta}_0), y_2 - g(x_2; \underline{\beta}_0), \ldots, y_n - g(x_n; \underline{\beta}_0) \Big) \tag{7.17}$$

$\boldsymbol{L}$ er en (diagonal) $n \times n$ matrix og $\underline{\Delta y}$ en n dimensional vektor. En bedre bestemmelse af minimumspunktet (i nærheden af $\underline{\beta}_0$) kan herefter findes ved at udregne:

$$\underline{\Delta \beta} = \Big(\boldsymbol{A}^T \cdot \boldsymbol{L} \cdot \boldsymbol{A} \Big)^{-1} \cdot \boldsymbol{A}^T \cdot \boldsymbol{L} \cdot \underline{\Delta y} \tag{7.18}$$

Er $\underline{\beta} = \underline{\beta}_0$ et minimumspunkt vil $\underline{\Delta \beta}$ være nulvektoren. Hvis ikke, erstattes $\underline{\beta}_0$ med $\underline{\beta}_0 + \underline{\Delta \beta}$ og regningen gentages. Dette skema gentages derefter indtil man er nået frem til en tilfredsstillende iterativ løsning. I minimumspunktet er variansen af de forskellige parametre bestemt ved:

$$s^2(\beta_i) = \left\{ \Big(\boldsymbol{A}^T \cdot \boldsymbol{L} \cdot \boldsymbol{A} \Big)^{-1} \right\}_{ii} \tag{7.19}$$

Hvis s_i er konstant, $s_i = s(y)$, forenkles regningerne, idet $\boldsymbol{L}$ kan erstattes med enhedsmatricen, når blot det sidste resultat (7.19) multipliceres med $s^2(y)$. Er s_i ikke kendt på anden måde skal der gode argumenter til for ikke at antage en konstant standardafvigelse. Accepteres denne antagelse kan $s(y)$ bestemmes af stikprøven på helt analog måde som i det lineære tilfælde:

$$s^2(y) = \frac{|\underline{\Delta y}|^2}{n - p} \tag{7.20}$$

Bemærk, at vi må forlange at n er større end (eller til nød lig med) p, for at antallet af ubekendte ikke skal blive større end det antal bindinger der er.