

Supplement til sandsynlighedsregning og matematisk statistik

1. Bevis for teorem 4b i §23.3

Teorem 4:

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

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

Det er teorem (4b), hvor de enkelte elementer kan optræde fra 0 til k gange i de enkelte kombinationer, der skal bevises.

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 gyldig 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 disse muligheder skal vi addere alle dem, hvor det n te element optræder 1 gang, svarende til alle mulige 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, dvs. det samlede antal kombinationer er:

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

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

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

Ligning (13), som benyttes i beviset, kan igen bevises ved induktion samt brug af ligning (11) (som vi udledte ved regneøvelserne).

Alternativt bevis for Teorem 4b

Teorem 4b 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, svarer til at dette element er udtrukket.

De n kasser stilles på en række, og de k (sorte) kugler fordeles, fx som vist i figur a). 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 og de resterende til at være skillevægge. Et eksempel er vist i figur c). Her er skillevæggene vist som hvide kugler, dvs. 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

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

a) 

b) 

c) 

2. Moivre's sætning

Sandsynlighedstætheden for binomialfordelingen:

$$f(x) = \binom{n}{x} p^x q^{n-x} \quad (x = 0, 1, \dots, n) \quad (2.1)$$

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

(I)

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

$$f(x) = \frac{n(n-1) \cdots (n-x+1)}{n^x} \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \quad (2.2)$$

hvor

$$\frac{n(n-1) \cdots (n-x+1)}{n^x} = 1 \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{x-1}{n}\right) \rightarrow 1 \quad (2.3)$$

for $n \rightarrow \infty$ og

$$\begin{aligned} \left(1 - \frac{\lambda}{n}\right)^n &= 1 - \frac{n}{1!} \frac{\lambda}{n} + \frac{n(n-1)}{2!} \frac{\lambda^2}{n^2} - \frac{n(n-1)(n-2)}{3!} \frac{\lambda^3}{n^3} + \cdots \\ &= 1 - \lambda + \left(1 - \frac{1}{n}\right) \frac{\lambda^2}{2!} - \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \frac{\lambda^3}{3!} + \cdots \\ &\rightarrow 1 - \lambda + \frac{\lambda^2}{2!} - \frac{\lambda^3}{3!} + \cdots = e^{-\lambda} \end{aligned} \quad (2.4)$$

Sidste faktor i (2.2) går mod 1, dvs.

$$f(x) \rightarrow \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{for } n \rightarrow \infty \quad (2.5)$$

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 $n \rightarrow \infty$. 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

$$Z = \frac{X - \mu}{\sigma} \quad (2.6)$$

I grænsen $n \rightarrow \infty$ bliver Z en kontinuert stokastisk variabel og $F(x = \sigma z + \mu)$ går over i standard normalfordelingen, med middelværdien 0 og spredningen 1, dvs. $F(\sigma z + \mu) \rightarrow \Phi(z)$ (De Moivre og Laplace's grænseværdisætning), svarende til ligning (11) på side 1189 i bogen (hvis ± 0.5 i α og β negligeres).

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, $r\sqrt{npq}/np \rightarrow 0$ for $n \rightarrow \infty$, og dermed at vi dels kan benytte Stirlings formel for alle faktorerne i binomialkoefficienten, n over x , og dels at vi kan udnytte en rækkeudvikling mht. Δ/μ , hvor $\Delta = x - \mu$:
Anvendes Stirlings formel (ligning (7) side 1163) i (2.1) fås:

$$\begin{aligned} f(x) &\approx \frac{\sqrt{2\pi n} n^n e^{-n}}{\sqrt{2\pi x} x^x e^{-x} \sqrt{2\pi(n-x)} (n-x)^{n-x} e^{-n+x}} p^x q^{n-x} \\ &= \sqrt{\frac{n}{2\pi x(n-x)}} \left(\frac{np}{x}\right)^x \left(\frac{nq}{n-x}\right)^{n-x} \end{aligned} \quad (2.7)$$

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

$$\left(\frac{np}{x}\right)^x \approx \exp \left[-\Delta - \frac{\Delta^2}{2\mu} \right] = \exp \left[np - x - \frac{(x - np)^2}{2np} \right] \quad (2.8)$$

og helt analogt

$$\left(\frac{nq}{n-x}\right)^{n-x} \approx \exp \left[nq - (n-x) - \frac{(n-x - nq)^2}{2nq} \right] = \exp \left[\Delta - \frac{\Delta^2}{2nq} \right] \quad (2.9)$$

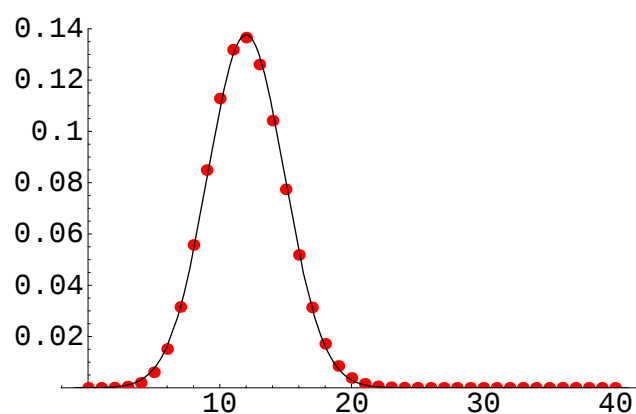
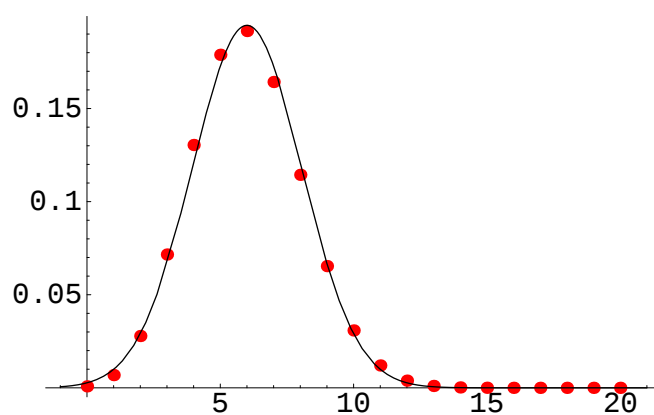
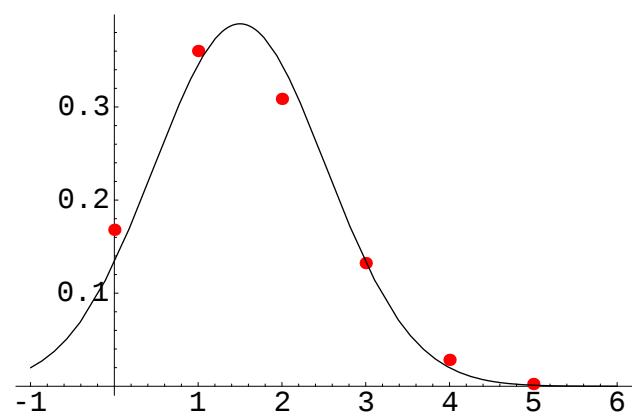
Ganger vi de to faktorer sammen fås

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

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

$$f(x) \approx \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{(x - \mu)^2}{2\sigma^2} \right] \quad (2.10)$$

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 (2.10) 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 $\Phi(z)$.



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

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

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

$$Y = \sum_{i=1}^n U_i^2, \quad (3.1)$$

hvor U_i , for hvert i , er en uafhængig stokastisk variabel med en frekvensfunktion, som er den normaliserede normalfordeling:

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

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 $y = \sum u_i^2 \leq \chi^2$:

$$F_n(\chi^2) = \int_{y \leq \chi^2} du_1 \int du_2 \cdots \int du_n (2\pi)^{-\frac{n}{2}} e^{-y/2}. \quad (3.2)$$

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

$$C_n = \frac{\pi^{\frac{n}{2}}}{\Gamma(1 + \frac{n}{2})}, \quad (3.3)$$

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 (3.2) kan vi uden videre integrere over alle de $n - 1$ frihedsgrader, der er uafhængige af $y = r^2$, hvilket giver:

$$F_n(\chi^2) = \int_0^{r=\sqrt{\chi^2}} S_n(r) (2\pi)^{-\frac{n}{2}} e^{-r^2/2} dr, \quad (3.4)$$

og dermed, at

$$\begin{aligned} f_n(\chi^2) &= F'_n(\chi^2) = \frac{dF_n}{dr} \frac{dr}{d\chi^2} = S_n(\sqrt{\chi^2}) (2\pi)^{-\frac{n}{2}} e^{-\chi^2/2} \frac{1}{2\sqrt{\chi^2}} \\ &= n(\sqrt{\chi^2})^{n-1} \frac{\pi^{\frac{n}{2}}}{\Gamma(1 + \frac{n}{2})} (2\pi)^{-\frac{n}{2}} e^{-\chi^2/2} \frac{1}{2\sqrt{\chi^2}} \end{aligned} \quad (3.5)$$

eller

$$f_n(\chi^2) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} (\chi^2)^{\frac{n}{2}-1} e^{-\chi^2/2}, \quad (3.6)$$

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

Betragter vi den stokastiske størrelse

$$S_0^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2, \quad (3.7)$$

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

$$n \frac{S_0^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^n U_i^2 \quad (3.8)$$

beskrevet ved χ^2 -fordelingen med n -frihedsgrader [højresiden er den samme som i ligning (3.1)]. Indfører vi den stokastiske middelværdi

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (3.9)$$

kan (3.7) omskrives til

$$nS_0^2 = \sum_i [(X_i - \bar{X}) + (\bar{X} - \mu)]^2 = \sum_i (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2, \quad (3.10)$$

idet ligning (3.9) medfører, at produktet mellem de to led i den firkantede parentes summerer til 0. Dermed har vi, at

$$n \frac{S_0^2}{\sigma^2} = \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2 + \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2. \quad (3.11)$$

Ifølge §24.6 teorem 2 er $\bar{X}$ normalfordelt med middelværdien μ og variansen σ^2/n , dvs. at første led på højre side af (3.11) netop er kvadratet på en stokastisk variabel, som har en normaliserede normalfordeling, og dermed, at dette led har en χ^2 -fordeling med 1 frihedsgrad. Nu er det let at indse, at summen af to (uafhængige) χ^2 -fordelinger med frihedsgraderne n_1 og n_2 giver χ^2 -fordelingen med $n_1 + n_2$ frihedsgrader. Det kan vises, at de to led på højre side af (3.11) er statistisk uafhængige (kovariansen er 0), hvoraf følger, at sidste led har en χ^2 -fordeling med $n - 1$ frihedsgrader. Middelværdien af χ^2 -fordelingen med n frihedsgrader er n . [Det kan vises, at

$$\begin{aligned} \int_0^\infty \chi^2 f_n(\chi^2) d\chi^2 &= n && \text{og variansen} \\ \int_0^\infty \chi^4 f_n(\chi^2) d\chi^2 - n^2 &= 2n \end{aligned} \quad (3.12)$$

med $f_n(\chi^2)$ givet ved ligning (3.6).] Stikprøveværdien af venstre side af ligning (3.11) (store bogstaver erstattes med stikprøveværdier givet ved de tilsvarende små bogstaver) er tilnærmet lig med middelværdien af højre side. Dvs. at $ns_0^2/\sigma^2 \simeq 1 + n - 1$ eller som forventet, at $s_0 \simeq \sigma$. Betragter vi alene det sidste led i (3.11), fås $\sum_i (x_i - \bar{x})^2/\sigma^2 \simeq n - 1$, eller at

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \simeq \sigma^2, \quad (3.13)$$

hvor s^2 er størrelsen, der bestemmes direkte af stikprøven, hvorimod udregningen af s_0^2 kræver kendskab til μ .

Student's t-fordeling:

Vi har netop betragtet den stokastiske variabel

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

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

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = U \frac{\sigma}{S}. \quad (3.14)$$

Ifølge (3.11) og (3.13) er

$$\frac{S^2}{\sigma^2} = \frac{Y_p}{p}, \quad (3.15)$$

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

$$\frac{T^2}{p} = \frac{Y_1}{Y_p} \quad \text{med} \quad Y_1 = U^2, \quad (3.16)$$

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 $y_1/y_p \leq t^2/p$

$$G(t^2/p) = \int_{y_p=0}^{\infty} \int_{y_1=0}^{y_p t^2/p} f_1(y_1) f_p(y_p) dy_1 dy_p. \quad (3.17)$$

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

$$2f_t(t) = \pm \frac{d(t^2/p)}{dt} \frac{dG(t^2/p)}{d(t^2/p)} = (2|t|/p) \int_0^{\infty} [f_1(y_1)]_{y_1=y_p t^2/p} y_p f_p(y_p) dy_p, \quad (3.18)$$

og benyttes ligning (3.6) fås

$$\begin{aligned} f_t(t) &= (|t|/p) \int_0^{\infty} \left[\frac{1}{\sqrt{2\pi}} y_1^{-\frac{1}{2}} e^{-y_1/2} \right]_{y_1=y_p t^2/p} y_p \frac{1}{2^{\frac{p}{2}} \Gamma(\frac{p}{2})} (y_p)^{\frac{p}{2}-1} e^{-y_p/2} dy_p \\ &= \frac{|t|}{p\sqrt{\pi}} \sqrt{\frac{p}{t^2}} \frac{1}{2^{\frac{p+1}{2}} \Gamma(\frac{p}{2})} \int_0^{\infty} (y_p)^{\frac{p-1}{2}} e^{-(1+\frac{t^2}{p})y_p/2} dy_p, \end{aligned} \quad (3.20)$$

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

$$f_t(t) = \frac{1}{\sqrt{p\pi}} \frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2})} \left(1 + \frac{t^2}{p} \right)^{-\frac{p+1}{2}}, \quad (3.21)$$

som er frekvensfunktionen til T i ligning (3.14), når $p = n - 1$. Dermed har vi udledt teorem (3) i §24.6.

En nærmere analyse af de to fordelinger kan findes som Mathematica programmer. Teorem 5 i §24.6 ("Central limit theorem") fortæller, at begge fordelinger nærmer sig normalfordelingen, når antallet af frihedsgrader vokser. Som nævnt har χ^2 -fordelingen med n frihedsgrader middelværdien $\mu = n$ og variansen $\sigma^2 = 2n$. t -fordelingen med p frihedsgrader har middelværdien $\mu = 0$ (frekvensfunktionen er lige i t). Variansen divergerer for $p = 1$ og $p = 2$. For $p \geq 3$ er variansen $\sigma^2 = p/(p-2)$. Dvs. at for store p bliver variansen 1, svarende til at usikkerheden knyttet til bestemmelsen af σ ud fra s bliver lille.

Appendix

Enhedskuglens volumen i n dimensioner kan bestemmes ved at udregne integralet

$$I_n = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-(x_1^2 + x_2^2 + \cdots + x_n^2)} dx_1 dx_2 \cdots dx_n \quad (3.22)$$

dels direkte, som et produkt af n Gauss integraler:

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi} \quad \Rightarrow \quad I_n = \pi^{\frac{n}{2}}, \quad (3.23)$$

dels ved at integrere over alle andre frihedsgrader end $r = (x_1^2 + x_2^2 + \cdots + x_n^2)^{1/2}$, hvorved

$$\begin{aligned} I_n &= \int_0^{\infty} S_n(r) e^{-r^2} dr = \int_0^{\infty} nr^{n-1} C_n e^{-r^2} dr \\ &= nC_n \frac{1}{2} \int_0^{\infty} z^{\frac{n}{2}-1} e^{-z} dz = \frac{n}{2} C_n \Gamma\left(\frac{n}{2}\right) = C_n \Gamma\left(1 + \frac{n}{2}\right). \end{aligned} \quad (3.24)$$

Ved at sammenholde de to resultater, fås

$$C_n = \frac{\pi^{\frac{n}{2}}}{\Gamma\left(1 + \frac{n}{2}\right)}. \quad (3.25)$$

4. “Central limit” teorem

Vi skal supplere bogen med nogle vigtige resultater. Vi vil starte med at betragte den moment-genererende funktion (§23.5 opgave 20 og §23.6 opgave 19):

$$G(t) = \langle e^{tX} \rangle = \begin{cases} \int_{-\infty}^{\infty} e^{tx} f(x) dx \\ \sum_i e^{tx_i} f(x_i) \end{cases} \quad (4.1)$$

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 udledt, at

$$G_{\text{bin}}(t) = (pe^t + q)^n \quad \text{og} \quad G_P(t) = e^{-\mu} e^{\mu e^t} \quad (4.2)$$

For Gauss-fordelingen er:

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

hvor eksponenten kan omskrives til

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

og dermed, at

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

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

$$G(t) = e^{(\mu t + \frac{1}{2}\sigma^2 t^2)}. \quad (4.3)$$

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:

$$G_y(t) = \langle e^{tY} \rangle = \langle e^{ta_1 X_1} e^{ta_2 X_2} \rangle = \langle e^{ta_1 X_1} \rangle \langle e^{ta_2 X_2} \rangle = G_1(a_1 t) G_2(a_2 t) \quad (4.4)$$

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 (4.3) og (4.4) kan nu benyttes til at bevise den generelle sætning **"the central limit theorem"**:

Antag X_i , $i = 1, \dots, n$ 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

$$Y_n = \frac{1}{n} \sum_{i=1}^n X_i \quad (4.5)$$

har følgende egenskaber:

- (i) Middelværdien er $\mu = \langle Y_n \rangle = \frac{1}{n} \sum_i \mu_i$.
- (ii) Variansen er $\sigma^2 = \frac{1}{n^2} \sum_i \sigma_i^2$.
- (iii) Fordelingsfunktionen for $Z_n = \frac{Y_n - \mu}{\sigma}$ vil (asymptotisk) $\rightarrow \Phi(z)$ (den normerede normalfordeling) for $n \rightarrow \infty$.

Bemærk, at teoremet i denne udgave er mere generelt end lærebogens version, teorem 5 i §24.6. Punkterne (i) og (ii) følger direkte af teorem 1 og 2 i §23.8. Punkt (iii) kan bevises ved at udnytte (4.4), 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 $G_i(t/n) = G_i(0) + G'_i(0)(t/n) + \frac{1}{2}G''_i(0)(t/n)^2 \dots$ giver nærmest definitions-mæssigt:

$$G_i\left(\frac{t}{n}\right) = 1 + \mu_i \frac{t}{n} + \frac{1}{2}(\sigma_i^2 + \mu_i^2) \frac{t^2}{n^2} + \dots = \exp \left[\mu_i \frac{t}{n} + \frac{1}{2} \sigma_i^2 \frac{t^2}{n^2} \right] + \mathcal{O}\left(\frac{1}{n^3}\right)$$

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

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

som netop, ifl. (4.3), er den moment-genererende funktion for normalfordelingen med middelværdi μ (i) og varians σ^2 (ii).

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 $\sigma^2 = \sigma_1^2 + \sigma_2^2$.

Dette resultat udledes let ved at benytte (4.3) og (4.4), idet

$$G_y(t) = G_1(t)G_2(t) = e^{(\mu_1 t + \frac{1}{2}\sigma_1^2 t^2)} e^{(\mu_2 t + \frac{1}{2}\sigma_2^2 t^2)} = e^{[(\mu_1 + \mu_2)t + \frac{1}{2}(\sigma_1^2 + \sigma_2^2)t^2]} \quad (4.6)$$

[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.:

$$F(y) = \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{y-x_1} f_1(x_1)f_2(x_2)dx_1dx_2 = \int_{-\infty}^{\infty} f_1(x_1)F_2(y-x_1)dx_1$$

eller

$$f(y) = F'(y) = \int_{-\infty}^{\infty} f_1(x)f_2(y-x)dx \quad (4.7)$$

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 $\sigma_1^2 + \sigma_2^2$.]

Ophobningsloven kan betragtes som et korrolar til “the central limit theorem”. Er Y en generel funktion af X_i kan vi tilnærmelsesvis erstatte funktionen med et lineært udtryk i et område omkring Y 's middelværdi:

$$Y = \psi(X_1, \dots, X_n) \approx \psi(\langle X_1 \rangle, \dots, \langle X_n \rangle) + \sum_i \left(\frac{\partial \psi}{\partial X_i} \right)_{X_j=\langle X_j \rangle} (X_i - \langle X_i \rangle)$$

Hvis de forskellige variable har en Gauss-fordeling, eller hvis der er mange forskellige bidrag (usikkerhedsberegninger!), vil Y , ifl. henholdsvis (4.6) og (4.5), (tilnærmelsesvis) have en normalfordeling med middelværdien

$$\langle Y \rangle = \psi(\langle X_1 \rangle, \dots, \langle X_n \rangle) \quad (4.8a)$$

og variansen

$$\sigma^2 = \sum_i \left(\frac{\partial \psi}{\partial X_i} \right)_{X_j=\langle X_j \rangle}^2 \sigma_i^2 \quad (4.8b)$$

Vi skal benytte ophobningsloven til at udregne variansen af parametrene k_0 og k_1 bestemt ved en lineær regression:

$$y = k_0 + k_1 x. \quad (4.9)$$

En "stikprøve" på n talsæt (x_i, y_i) , antages givet, hvor x_i antages bestemt uden usikkerheder (eller snarere at $k_1 \sigma_x \ll \sigma_y$). "Maksimum likelihood" metoden fører til resultatet, at de mest sandsynlige værdier af de to parametre bestemmes ved at minimere summen af de kvadratiske afvigelse $\sum (y_i - \bar{y}_i)^2$, hvor vi har indført $\bar{y}_i = k_0 + k_1 x_i$:

Resultatet er (§24.12 eller §19.5)

$$\begin{aligned} D &= n \sum x_i^2 - \left(\sum x_i \right)^2 \\ k_0 &= \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{D} \\ k_1 &= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{D}, \end{aligned} \quad (4.10)$$

hvor alle summer går fra $i = 1$ til $i = n$. 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) vil $\sum (y_i - \bar{y}_i)^2 / \sigma_y^2$ have χ^2 -fordelingen med $n - 2$ frihedsgrader, og dermed er $\sigma_y \simeq s_y$, hvor

$$s_y^2 = \frac{1}{n - 2} \sum_i (y_i - \bar{y}_i)^2. \quad (4.11)$$

Benytter vi ophobningsloven er stikprøve-variansen af k_0 :

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

eller

$$s^2(k_0) = \frac{\sum x_j^2}{D} s_y^2. \quad (4.12)$$

Helt analogt kan variansen af k_1 udregnes til:

$$s^2(k_1) = \frac{n}{D} s_y^2. \quad (4.13)$$

Dette sidste resultat svarer til bogens opskrift i Tabel 24.13: $q_0 = (n - 2)s_y^2$ og $(n - 1)s_1^2 = D/n$ og dermed $K = cs(k_1)$. Konstanten c der angiver den relative længde af konfidensintervallet i forhold til standardafvigelsen, bestemmes af Student's t -fordeling med $n - 2$ frihedsgrader (frihedsgraderne reduceres med 2 fordi stikprøven benyttes til at fastlægge 2 parametre k_0 og k_1). Dette er helt analogt med bestemmelsen af konfidensintervallet for μ , når stikprøveværdien af standardafvigelsen benyttes.
