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

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

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

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 varibel y som er ikke-negativ). Fordelingsfunktionen $F_n(\chi^2)$ er produktet af frekvensfunktionerne (indbyrdes uafhængige stokastiske variable) integreret over volumenet givet ved $y = \sum u_i^2 \leq \chi^2$:

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

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

$$C_n = \frac{\pi^{\frac{n}{2}}}{\Gamma(1+\frac{n}{2})}, \tag{3}$$

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

$$F_n(\chi^2) = \int_0^{r=\sqrt{\chi^2}} S_n(r)(2\pi)^{-\frac{n}{2}} e^{-r^2/2} dr, \tag{4}$$

og dermed, at

$$\begin{aligned} f_n(\chi^2) = F_n'(\chi^2) &= \frac{dF_n}{dr}\frac{dr}{d\chi^2} = S_n\left(\sqrt{\chi^2}\right)(2\pi)^{-\frac{n}{2}} e^{-\chi^2/2}\,\frac{1}{2\sqrt{\chi^2}} \\ &= n\left(\sqrt{\chi^2}\right)^{n-1}\frac{\pi^{\frac{n}{2}}}{\Gamma(1+\frac{n}{2})}(2\pi)^{-\frac{n}{2}} e^{-\chi^2/2}\,\frac{1}{2\sqrt{\chi^2}} \end{aligned} \tag{5}$$

eller

$$f_n(\chi^2) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\left(\chi^2\right)^{\frac{n}{2}-1} e^{-\chi^2/2}, \tag{6}$$

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

Betragter vi den stokastiske størrelse

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2, \tag{7}$$

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

$$n\frac{S_0^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n} U_i^2 \tag{8}$$

beskrevet ved χ^2-fordelingen med n-frihedsgrader [højresiden er den samme som i ligning (1)]. Indfører vi den stokastiske middelværdi

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{9}$$

kan (7) omskrives til

$$nS_0^2 = \sum_i \left[(X_i - \overline{X}) + (\overline{X} - \mu)\right]^2 = \sum_i (X_i - \overline{X})^2 + n(\overline{X} - \mu)^2, \tag{10}$$

idet ligning (9) medfører, at produktet mellem de to led i den firkantede parentes summerer til 0. Dermed har vi, at

$$n\frac{S_0^2}{\sigma^2} = \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2 + \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2. \tag{11}$$

Ifølge §24.6 teorem 2 er $\overline{X}$ normalfordelt med middelværdien μ og variansen σ^2/n, dvs. at første led på højre side af (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 (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

$$\begin{aligned} &\int_0^\infty \chi^2 f_n(\chi^2) d\chi^2 = n \qquad \text{og variansen} \\ &\int_0^\infty \chi^4 f_n(\chi^2) d\chi^2 - n^2 = 2n \end{aligned} \tag{12}$$

med $f_n(\chi^2)$ givet ved ligning (6).] Stikprøveværdien af venstre side af ligning (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 (11), fås $\sum_i (x_i - \overline{x})^2/\sigma^2 \simeq n - 1$, eller at

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \simeq \sigma^2, \tag{13}$$

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

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

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 i stedet

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = U\frac{\sigma}{S}. \tag{14}$$

Ifølge (11) og (13) er

$$\frac{S^2}{\sigma^2} = \frac{Y_p}{p}, \tag{15}$$

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

$$\frac{T^2}{p} = \frac{Y_1}{Y_p} \qquad \text{med} \qquad Y_1 = U^2, \tag{16}$$

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

$$G(t^2/p) = \int_{y_p=0}^{\infty} \int_{y_1=0}^{y_p t^2/p} f_1(y_1) f_p(y_p) dy_1 dy_p. \tag{17}$$

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

$$2f_t(t) = \pm\frac{d(t^2/p)}{dt}\frac{dG(t^2/p)}{d(t^2/p)} = (2|t|/p)\int_0^{\infty} \left[f_1(y_1)\right]_{y_1=y_pt^2/p} y_p f_p(y_p) dy_p, \tag{18}$$

og benyttes ligning (6) fås

$$\begin{aligned} f_t(t) &= (|t|/p)\int_0^{\infty} \left[\frac{1}{\sqrt{2\pi}}\, y_1^{-\frac{1}{2}} e^{-y_1/2}\right]_{y_1=y_pt^2/p} y_p \frac{1}{2^{\frac{p}{2}}\Gamma(\frac{p}{2})}(y_p)^{\frac{p}{2}-1}\, e^{-y_p/2} dy_p \\ &= \frac{|t|}{p\sqrt{\pi}}\sqrt{\frac{p}{t^2}}\frac{1}{2^{\frac{p+1}{2}}\Gamma(\frac{p}{2})}\int_0^{\infty} (y_p)^{\frac{p-1}{2}}\, e^{-(1+\frac{t^2}{p})y_p/2} dy_p, \end{aligned} \tag{20}$$

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

$$f_t(t) = \frac{1}{\sqrt{p\pi}}\frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2})}\left(1 + \frac{t^2}{p}\right)^{-\frac{p+1}{2}}, \tag{21}$$

som er frekvensfunktionen til T i ligning (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 nærmere sig normalfordelingen, når antallet af frihedsgrader vokser. Som nævnt har χ^2-fordelingen med n frihedsgrader middelværdien $\mu = n$ og variansen $\sigma^2 = 2n$. 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 at usikkerheden knyttet til bestemmelsen af σ ud fra s bliver lille.

Appendix

Enhedskuglens volumen i n dimensioner kan bestemmes ved at udregne integralet

$$I_n = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-(x_1^2 + x_2^2 + \cdots + x_n^2)} dx_1 dx_2 \cdots dx_n \tag{22}$$

dels direkte, som et produkt af n Gauss integraler:

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi} \qquad \Rightarrow \qquad I_n = \pi^{\frac{n}{2}}, \tag{23}$$

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

$$\begin{aligned} I_n &= \int_0^{\infty} S_n(r)\, e^{-r^2} dr = \int_0^{\infty} n r^{n-1} C_n\, e^{-r^2} dr \\ &= n C_n \tfrac{1}{2} \int_0^{\infty} z^{\frac{n}{2}-1} e^{-z} dz = \tfrac{n}{2} C_n \Gamma(\tfrac{n}{2}) = C_n \Gamma(1 + \tfrac{n}{2}). \end{aligned} \tag{24}$$

Ved at sammenholde de to resultater, fås

$$C_n = \frac{\pi^{\frac{n}{2}}}{\Gamma(1 + \frac{n}{2})}. \tag{25}$$