IMM, 2002-10-10
Poul Thyregod

Lidt om fordelinger, afledt af normalfordelingen

1 Introduktion

I forbindelse med inferens i normalfordelinger optræder forskellige fordelinger, der er afledt af normalfordelingen, t-fordelingen, F-fordelingen og χ^2-fordelingen.

Denne note giver en kort introduktion til disse fordelinger og deres forbindelse med normalfordelingsmodeller.

2 χ^2-fordelingen

(Se også Petrucelli side 408)

χ^2- fordelingen er en familie af sandsynlighedsfordelinger på den positive reelle akse med tæthedsfunktion

$$p(x) = \frac{1}{2\Gamma(f/2)} \left(\frac{x}{2}\right)^{f/2-1} \exp(-x/2) \qquad \text{for } 0 < x < \infty$$

hvor $\Gamma(\dot{})$ angiver Gammafunktionen (se nedenfor afsnit 5).

Størrelsen f er en heltallig parameter, der fastlægger hvilken af de mange χ^2-fordelinger, vi betragter. Parameteren f betegnes *frihedsgraderne* (degrees of freedom) for fordelingen.
I anvendelser er frihedsgraderne sædvanligvis kendt.

En stokastisk variabel, der følger en χ^2-fordeling med f frihedsgrader symboliseres ofte ved χ^2_f.

Lærebogens appendix A.5 indeholder tabeller over fraktiler (kritiske værdier) for χ^2-fordelingen.

Makroen Nprobsrev giver mulighed for beregning af værdier for den kumulerede fordelingsfunktion for χ^2-fordelingen. Tilsvarende kan makroen Invprobrev bruges til at beregne fraktiler for fordelingen.

(χ^2 fordelingen er et specialtilfælde af en lidt større familie af fordelinger, de såkaldte *gammafordelinger*.

2.1 Fordelingens momenter

Såfremt $Y \sim \chi^2_f$ gælder

$$\mathsf{E}[Y] = f; \qquad \mathsf{V}[Y] = 2f \tag{1}$$

2.2 Relation til normalfordelingen

χ^2-fordelingen optræder naturligt i forbindelse med normalfordelingsmodeller.

Der gælder således:

Sætning:
Lad $Y \sim \mathrm{N}(0,1)$ og lad $Z = Y^2$. Da gælder

$$Z \sim \chi^2_1$$

Sætningen vises ved brug af resultater vedrørende fordelingen af transformationer af stokastiske variable.

Ved brug af dette resultat ser man, at hvis $Y \sim \mathrm{N}(\mu, \sigma^2)$ da gælder

$$\frac{(Y-\mu)^2}{\sigma^2} \sim \chi^2_1 \tag{2}$$

Der gælder endvidere følgende

Sætning:

Lad $Y_1, Y_2, \ldots, Y_n$ være uafhængige standardiserede normalt fordelte variable med $Y_i \sim \mathrm{N}(0,1)$, og lad

$$\begin{aligned}\mathrm{SS} &= \sum_{i=1} Y_i^2 \\ \mathrm{SST} &= \sum_{i=1} (Y_i - \overline{Y})^2\end{aligned}$$

Da gælder

1. $\mathrm{SS} \sim \chi_n^2$
2. $\mathrm{SST} \sim \chi_{n-1}^2$
3. $n\overline{Y}^2 \sim \chi_1^2$
4. $\mathrm{SS} = \mathrm{SST} + n\overline{Y}^2$, og
   SST og $n\overline{Y}^2$ er stokastisk uafhængige

- o -

Bemærk, at opspaltningen

$$\mathrm{SS} = \mathrm{SST} + n\overline{Y}^2$$

netop svarer til opspaltningen

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 + n\overline{Y}^2$$

som blev behandlet i noten om frihedsgrader.

Det er ikke overraskende, at $n\overline{Y}^2 \sim \chi_1^2$, for der gælder jo, at $\overline{Y} \sim \mathrm{N}(0, 1/n)$ og så følger dette resultat af (2).

Ovenstående resultat kan benyttes til at vise, at såfremt $Y_1, Y_2, \ldots, Y_n$ er uafhængige normalt fordelte variable med $Y_i \sim \mathrm{N}(\mu, \sigma^2)$, og man sætter

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \tag{3}$$

Da gælder, at

$$\mathsf{E}[S^2] = \sigma^2$$

jvf (1), hvilket begrunder at vi dividerer kvadratafvigelsessummen med dens frihedsgrader, $n-1$, og ikke med antallet, n, af observationer.

Man finder endvidere af (1) at

$$\mathsf{V}[S^2] = \frac{2(\sigma^2)^2}{n-1}$$

3 t-fordelingen

(Se også Petrucelli side 248)

t-fordelingen (undertiden betegnet Students t-fordeling) er en familie af sandsynlighedsfordelinger på den reelle akse med tæthedsfunktion

$$p(t) = \frac{\Gamma((f+1)/2)}{\sqrt{f\pi}\Gamma(f/2)} \left(1 + \frac{t^2}{f}\right)^{-(f+1)/2}$$

hvor $\Gamma(\dot{})$ angiver Gammafunktionen (se nedenfor afsnit 5).

Størrelsen f er en heltallig parameter, der fastlægger hvilken af de mange t-fordelinger, vi betragter. Parameteren f betegnes *frihedsgraderne* (degrees of freedom) for fordelingen.

I anvendelser er frihedsgraderne sædvanligvis kendt.

En stokastisk variabel, der følger en t-fordeling med f frihedsgrader symboliseres ofte ved t_f,

Lærebogens appendix A.4 indeholder tabeller over fraktiler (kritiske værdier) for t-fordelingen.

Makroen Nprobs (og Nprobsrev) giver mulighed for beregning af værdier for den kumulerede fordelingsfunktion for t-fordelingen. Tilsvarende kan makroen Invprobs (og Invprobrev) bruges til at beregne fraktiler for fordelingen.

3.1 Fordelingens momenter

Lad $Y \sim \mathrm{t}_f$.

For $f > 1$ har fordelingen en middelværdi, og der gælder

$$\mathsf{E}[Y] = 0 \quad \text{for } f > 1$$

For $f > 2$ har fordelingen en varians, og der gælder

$$\mathsf{V}[Y] = \frac{f}{f-2} \quad \text{for } f > 2$$

3.2 Relation til normalfordelingen

Sætning:

Lad $U \sim \mathrm{N}(0,1)$ og lad $Z \sim \chi^2_f$ være **uafhængige** stokastiske variable.

Da gælder, at den stokastiske variable

$$T = \frac{U}{\sqrt{Z/f}}$$

følger en t_f-fordeling, dvs en t-fordeling med f frihedsgrader.

Det er netop dette resultat, vi bruger i forbindelse med inferens i normalfordelingsmodeller.

Lad nemlig $Y_1, Y_2, \ldots, Y_n$ være uafhængige normalt fordelte variable med $Y_i \sim \mathrm{N}(\mu, \sigma^2)$, og lad $\overline{Y}$ betegne gennemsnittet, og S^2 betegne den empiriske varians (3). Da gælder jo

$$\begin{aligned} U &= \frac{\overline{Y} - \mu}{\sigma / \sqrt{n}} \sim \mathrm{N}(0{,}1) \\ Z &= f \cdot S^2 / \sigma^2 \sim \chi^2_f \end{aligned}$$

med $f = n - 1$ og U og Z er netop stokastisk uafhængige jvf resultatet i afsnit 2.2.

Den sædvanlige t-størrelse kan jo netop udtrykkes som

$$T = \frac{U}{\sqrt{Z/f}}$$

idet σ forkortes ud.

Resultatet viser også, at udtrykket for den sædvanlige t-størrelse for en-stikprøvesituationen

$$Z = \frac{\overline{Y} - \mu_0}{\widehat{\sigma}(\overline{Y})} = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}} \tag{4}$$

referer **frihedsgraderne**, f til estimatet for σ^2, mens størrelsen n refererer til antallet af observationer, der indgår i gennemsnittet $\overline{Y}$.

Dette benyttes blandt andet når man **pooler** variansestimater.

4 F-fordelingen

(Se også Petrucelli side 477 og 529)

F-fordelingen er en familie af sandsynlighedsfordelinger på den positive reelle akse med tæthedsfunktion

$$p(x) = \frac{\Gamma((n+m)/2)}{\Gamma(n/2)\Gamma(m/2)} \left(\frac{n}{m}\right)^{n/2} \frac{x^{n/2-1}}{(1+nx/m)^{(n+m)/2}} \qquad \text{for } 0 < x < \infty$$

hvor $\Gamma(\dot{})$ angiver Gammafunktionen (se nedenfor afsnit 5).

Størrelserne n og m er heltallige parametre, der fastlægger hvilken af de mange F-fordelinger, vi betragter. Parametrene betegnes *frihedsgraderne* (degrees of freedom) for fordelingen.

En stokastisk variabel, der følger en F-fordeling med (n, m) frihedsgrader symboliseres ofte ved $\mathrm{F}_{n,m}$.

Lærebogens appendix A.6 indeholder tabeller over fraktiler (kritiske værdier) for F-fordelingen.

Makroen Nprobsrev giver mulighed for beregning af værdier for den kumulerede fordelingsfunktion for F-fordelingen. Tilsvarende kan makroen Invprobrev bruges til at beregne fraktiler for fordelingen.

ved bestemmelse af fraktiler kan man undertiden have fordel af relationen

$$\mathrm{F}_{n,m;q} = \frac{1}{\mathrm{F}_{m,n;1-q}} \tag{5}$$

4.1 Fordelingens momenter

Lad t $Y \sim \mathrm{F}_{n,m}$.

For $m > 2$ har fordelingen en middelværdi, og der gælder

$$\mathsf{E}[Y] = \frac{m}{m-2} \quad \text{for } m > 2$$

For $m > 4$ har fordelingen en varians, og der gælder

$$\mathsf{V}[Y] = \frac{m^2(2n+2m-4)}{(m-2)^2(m-4)n} \quad \text{for } m > 4$$

4.2 Relation til normalfordelingen

F-fordelingen optræder naturligt i forbindelse med normalfordelingsmodeller.

Der gælder således:

Sætning:
Lad $Y \sim \mathrm{t}_f)$ og lad $Z = Y^2$. Da gælder

$$Z \sim \mathrm{F}_{1,f}$$

altså en F-fordeling med $(1, f)$ frihedsgrader.

I forbindelse med lineære normalfordelingsmodeller benyttes ofte

Sætning:
Lad $Z_1 \sim \chi^2_n$ og $Z_2 \sim \chi^2_m$ være stokastisk uafhængige, og lad

$$Y = \frac{Z_1/n}{Z_2/m}$$

da gælder, at $Y \sim \mathrm{F}_{n,m}$.

F-fordelingen beskriver således fordelingen af forholdet mellem to **uafhængige** variansestimater under antagelse af at disse estimater begge beskriver den **samme** varians, σ^2.

Ofte betegner man derfor parameteren n som frihedsgraderne for **tælleren**, og m som frihedsgraderne for **nævneren**.

5 Gammafunktionen

Gammafunktionen, $\Gamma(\cdot)$ er en funktion af et reelt positivt argument givet ved

$$\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) dx$$

Der gælder relationen

$$x\Gamma(x) = \Gamma(x+1)$$

Specielt har man $\Gamma(1) = 1$ og $\Gamma(1/2) = \sqrt{\pi}$.

For heltallige værdier, n har man derfor

$$\begin{aligned} \Gamma(n) &= 1 \cdot 2 \cdot 3 \cdots (n-1) = (n-1)! \\ \Gamma(n+1/2) &= \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{2^n} \sqrt{\pi} \end{aligned}$$