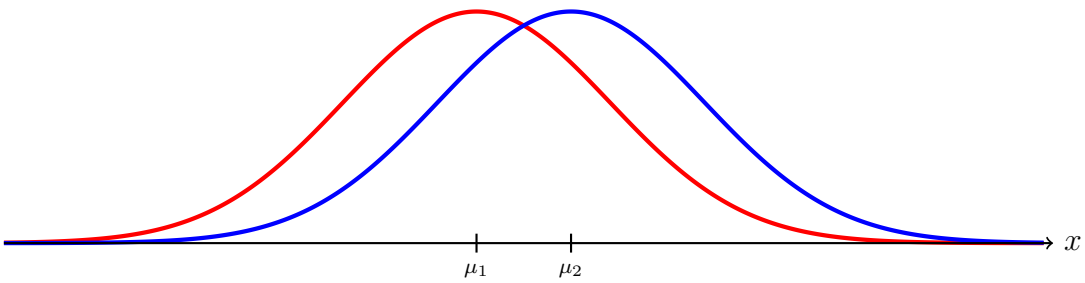
Föreläsning 4: Konfidensintervall (forts.)

Johan Thim (johan.thim@liu.se)

13 september 2018

1 Skillnad mellan parametrar

Vi kommer nu fortsätta med att konstruera konfidensintervall och vi kommer betrakta lite olika situationer där vi börjar med att titta på framförallt skillnader mellan olika mätningar. En rimlig fråga är om det föreligger någon skillnad mellan till exempel väntevärden för två stycken stickprov. Antag att vi har två slumpmässiga stickprov från två normalfördelningar. Vi vet inte direkt om fördelningarna har samma parametrar, så situationen skulle kunna se ut enligt följande.



Hur avgör vi om till exempel $\mu_1 = \mu_2$? Eller snarare om det är så att $\mu_1 \neq \mu_2$? Eller kanske om $\mu_2 > \mu_1$? Går det att avgöra om varianserna skiljer sig åt? Vad gör vi om inte stickprovet är från en normalfördelning?

2 Linjärkombinationer av normalfördelningar

Låt $X_1, \ldots, X_m$ och $Y_1, \ldots, Y_n$ vara oberoende slumpmässiga stickprov från $N(\mu_1, \sigma_1^2)$ respektive $N(\mu_2, \sigma_2^2)$. Om c_1 och c_2 är konstanter, kan vi hitta ett konfidensintervall för linjärkombinationen $c_1\mu_1 + c_2\mu_2$? Svaret beror på vilka antaganden vi gör. Vi börjar med att hitta en lämplig stokastisk storhet. Vi ser att

$$E(c_1\overline{X} + c_2\overline{Y}) = c_1\mu_1 + c_2\mu_2 \quad \text{och} \quad V(c_1\overline{X} + c_2\overline{Y}) = c_1^2\,\frac{\sigma_1^2}{m} + c_2\,\frac{\sigma_2^2}{n},$$

så eftersom vi har oberoende normalfördelade variabler gäller att

$$Z = \frac{c_1\overline{X} + c_2\overline{Y} - (c_1\mu_1 + c_2\mu_2)}{\sqrt{c_1^2\,\frac{\sigma_1^2}{m} + c_2\,\frac{\sigma_2^2}{n}}} \sim N(0,1). \tag{1}$$

Om vi känner σ_1 och σ_2 räcker detta för att ställa upp ett resultat.

2.1 Känd varians

Kända varianser

Antag att följande värden är uppmätta.

x_i	47.7	55.6	51.3	46.1	54.9			
y_i	29.2	47.8	30.9	37.7	27.9	40.1	41.5	40.9

Låt x_i vara observationer av stokastiska variabler $X_i \sim \mathrm{N}(\mu_1, 4^2)$ och y_i observationer av stokastiska variabler $Y_i \sim \mathrm{N}(\mu_2, 9^2)$, där samtliga variabler är oberoende. Ange ett 95% konfidensintervall för $\mu_1 - 2\mu_2$.

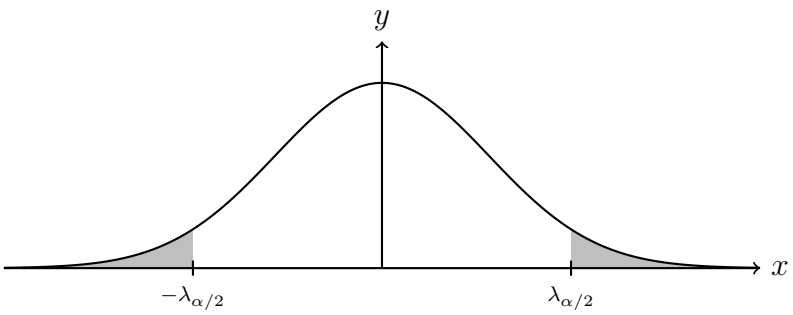
Lösning:
Låt $W = \overline{X} - 2\overline{Y}$. Varför? Denna storhet har egenskapen att $E(W) = E(\overline{X}) - 2E(\overline{Y}) = \mu_1 - 2\mu_2$, vilket är precis vad vi är intresserade av. Vidare är

$$V(W) = V(\overline{X}) + (-2)^2 V(\overline{Y}) = \frac{4^2}{5} + 4\frac{9^2}{8} = 43.7.$$

En lämplig teststorhet ges av

$$Z = \frac{W - (\mu_1 - 2\mu_2)}{\sqrt{V(W)}} \sim \mathrm{N}(0, 1).$$

Obligatorisk principfigur!



Eftersom

$$P(-\lambda_{\alpha/2} < Z < \lambda_{\alpha/2}) = 1 - \alpha$$

kan vi ur olikheten lösa ut sambandet

$$W - \lambda_{\alpha/2}\sqrt{V(W)} < \mu_1 - 2\mu_2 < W + \lambda_{\alpha/2}\sqrt{V(W)}.$$

Vi skattar W med $w = \overline{x} - 2\overline{y} = 51.12 - 2 \cdot 37 = -20.9$. Ur tabell finner vi att $\Phi(1.96) = 0.95 + 0.025 = 0.975$, så $\lambda_{\alpha/2} = 1.96$. Alltså blir intervallet

$$\begin{aligned} I_{\mu_1 - 2\mu_2} &= \left(-20.9 - 1.96 \cdot \sqrt{43.7},\ -20.9 + 1.96 \cdot \sqrt{43.7}\right) \\ &= (-33.85,\ -7.95). \end{aligned}$$

Vad säger detta oss? Jo, att med 95% säkert så ligger det verkliga värdet för $\mu_1 - 2\mu_2$ i intervallet $(-33.85, -7.95)$. Till exempel ser vi att noll inte finns med i intervallet, så det måste vara så att $2\mu_2 > \mu_1$ med hög säkerhet!

2.2 Okända men likadana varianser ($\sigma_1 = \sigma_2$)

Så om vi inte känner till vad varianserna är behöver vi skatta dessa. Om vi dessutom antar att $\sigma_1 = \sigma_2$ får vi ett enklare resultat, så vi börjar med det. Om vi nyttjar att $\sigma_1 = \sigma_2 = \sigma$ i ekvation (1) erhåller vi att

$$Z = \frac{c_1\overline{X} + c_2\overline{Y} - (c_1\mu_1 + c_2\mu_2)}{\sigma\sqrt{\frac{c_1^2}{m} + \frac{c_2^2}{n}}} \sim N(0,1).$$

Men vi vet fortfarande inte vad σ är, så vi ersätter σ med stickprovsstandardavvikelsen s. Eftersom vi har två stickprov viktar vi ihop dessa på sedvanligt sätt:

$$s^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}.$$

Motsvarande stickprovsvariabel S^2 uppfyller som bekant att $\dfrac{(m+n-2)S^2}{\sigma^2} \sim \chi^2(m+n-2)$ och enligt Gossets sats blir

$$T = \frac{c_1\overline{X} + c_2\overline{Y} - (c_1\mu_1 + c_2\mu_2)}{S\sqrt{\frac{c_1^2}{m} + \frac{c_2^2}{n}}} \sim t(m+n-2).$$

Okänd varians

Samma siffror som i exemplet ovan, men nu vet vi inte vad standardavvikelserna är. Antag att de är lika, dvs att $\sigma_1 = \sigma_2 = \sigma$. Finn ett 95% K.I. för $\mu_1 - \mu_2$ (inte samma uttryck som sist!). Kan du säga något om påståendet att $\mu_1 > \mu_2$?

Lösning:
Vi antar alltså här att $X_i \sim \mathrm{N}(\mu_1, \sigma^2)$ och $Y_i \sim \mathrm{N}(\mu_2, \sigma^2)$.
Vi kan skatta varianserna för varje serie med de vanliga stickprovsvarianserna, så

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad \text{och} \qquad s_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \overline{y})^2$$

är kända storheter. Dessa viktas ihop enligt

$$s^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}.$$

Det följer nu att

$$T = \frac{c_1\overline{X} + c_2\overline{Y} - (c_1\mu_1 + c_2\mu_2)}{S\sqrt{\frac{c_1^2}{n} + \frac{c_2^2}{m}}} \sim t(n+m-2).$$

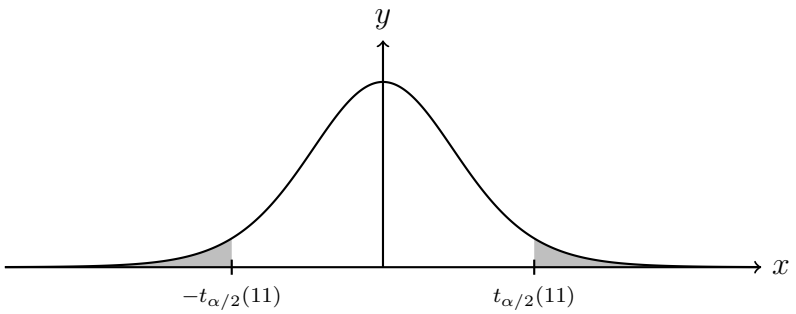
Låt $k := \sqrt{\frac{c_1^2}{n} + \frac{c_2^2}{m}}$. Snarlikt med fallet där vi kände varianserna kan vi stänga in T:

$$P(-t_{\alpha/2}(n+m-2) < T < t_{\alpha/2}(n+m-2)) = 1-\alpha$$

där vi ur olikheten kan lösa ut sambandet

$$T - t_{\alpha/2}(n+m-2)\cdot S\cdot k < c_1\mu_1 + c_2\mu_2 < T + t_{\alpha/2}(n+m-2)\cdot S\cdot k.$$

Vi har $n = 5$ och $m = 8$, så $m+n-2 = 11$ frihetsgrader. Ur tabell finner vi att $t_{0.025}(11) = 2.20$.



Vi kan räkna ut stickprovsvarianserna för x_i och y_i separat (med formel eller miniräknare). Vi erhåller $s_1^2 = 17.822$ och $s_2^2 = 49.009$ (små bokstäver, ej stokastiskt!). Den sammanvägda standardavvikelsen blir då

$$s = \sqrt{\frac{4s_1^2 + 7s_2^2}{11}} = 6.1374.$$

Vidare är $c_1 = 1$ och $c_2 = -1$, så

$$k = \sqrt{\frac{c_1^2}{n} + \frac{c_2^2}{m}} = \sqrt{\frac{1}{5} + \frac{1}{8}} = 0.5701.$$

Alltså blir

$$t_{0.025}(11)s\sqrt{\frac{c_1^2}{n} + \frac{c_2^2}{m}} = 2.20 \cdot 6.1374 \cdot 0.5701 = 7.6976.$$

Vi kan också räkna ut att $\overline{x} - \overline{y} = 14.12$, så det sökta intervallet ges av

$$\begin{aligned} I_{\mu_1-\mu_2} &= (14.12 - 7.70,\ 14.12 + 7.70) \\ &= (6.42,\ 21.82). \end{aligned}$$

Vi ser att noll ej ingår i intervallet, så det förligger troligt att $\mu_1 > \mu_2$.

2.3 Okända varianser ($\sigma_1 \neq \sigma_2$)

Ha ha. Well.. vi har inget användbart exakt samband, men det finns metoder för att hantera även denna situation. Dessa metoder ligger utanför denna kurs, men det kanske kan vara intressant att ha hört talas om dem. Problemet ligger i att uppskatta frihetsgraden ν för $t(\nu)$-fördelningen. Man kan visa (Welch-Satterthwaite-ekvationen) att

$$\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \overset{\text{appr.}}{\sim} \chi^2(\nu), \quad \text{där } \nu = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 \Bigg/ \left(\frac{1}{n_1 - 1}\,\frac{s_1^4}{n_1^2} + \frac{1}{n_2 - 1}\,\frac{s_2^4}{n_2^2}\right).$$

Därifrån kan vi till exempel använda att

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{\text{appr.}}{\sim} t(\nu)$$

för att ställa upp ett konfidensintervall för $\mu_1 - \mu_2$.

3 Stickprov i par

Om stickproven $X_1, \ldots, X_m$ och $Y_1, \ldots, Y_n$ inte är oberoende får vi problem. Åtminstone om inte beroendet är känt. Låt oss betrakta ett vanligt förekommande exempel, nämligen stickprov i par. Av nödvändighet är då $m = n$ så stickproven har samma storlek. Vi tänker oss att x_k är observationer från $X_k \sim \mathrm{N}(\mu_k, \sigma_1^2)$ och $Y_k \sim \mathrm{N}(\mu_k + \Delta, \sigma_2^2)$. Typexemplet är när vi mäter något före och efter en förändring.
Bilda nu ett "nytt" stickprov Z_k av oberoende variabler:

$$Z_k = Y_k - X_k \sim \mathrm{N}(\Delta, \sigma^2),$$

för något σ. Vi är nu tillbaka där vi var föregående föreläsning, så de tekniker vi utvecklade där fungerar även nu.

Exempel

Preparat mot (h)järnbrist. Mätningar (i lämplig enhet) före och efter behandling hos nio patienter.

Person	1	2	3	4	5	6	7	8	9
Före	15.8	12.1	18.2	9.4	11.8	16.6	13.7	13.5	17.5
Efter	14.8	12.4	18.3	9.5	12.2	15.6	13.4	14.4	16.0

Bestäm ett 99% KI av den genomsnittliga effekten hos preparatet. Kan du styrka att det fungerar?

Lösning:
Låt x_i vara värde före behandling för person i och y_i motsvarande efter. Vi antar att olika personer är oberoende och att x_i är observationer från $X_i \sim \mathrm{N}(\mu_i, \sigma_1^2)$ och $Y_i \sim \mathrm{N}(\mu_i + \Delta, \sigma_2^2)$. Bilda $Z_i = Y_i - X_i \sim \mathrm{N}(\Delta, \sigma^2)$. Vi har nu en enda serie $z_i = y_i - x_i$ som ges enligt

z_i	-1.0	0.3	0.1	0.1	0.4	-1.0	-0.3	0.9	0.5

Vi räknar ut $s = 0.7886$ och $\overline{z} = 0.2222$. Vidare är $n - 1 = 8$ och $\alpha = 0.01$, så $t_{\alpha/2}(8) = t_{0.005}(8) = 3.36$. Alltså,

$$I_\Delta = (0.222 - 3.36 \cdot 0.7886/\sqrt{9},\ 0.222 + 3.36 \cdot 0.7886/\sqrt{9}) = (-0.66,\ 1.11).$$

Eftersom nollan finns med kan vi inte förkasta att $\Delta = 0$ (med 99% säkerhet). Preparatet kan alltså vara verkningslöst.

4 Jämförelse av varianser

"The box. You opened it. We came."
–Pinhead

Vi antog tidigare att stickproven hade samma varians (för att kunna ställa upp en lämplig teststorhet). Hur vet vi det? Kan vi på något sätt avgöra om det antagandet är rimligt? Vi vill alltså jämföra varianserna för två stickprov och för att göra det behöver vi introducera en ny fördelning (ljuva lycka!).

4.1 F-fördelningen

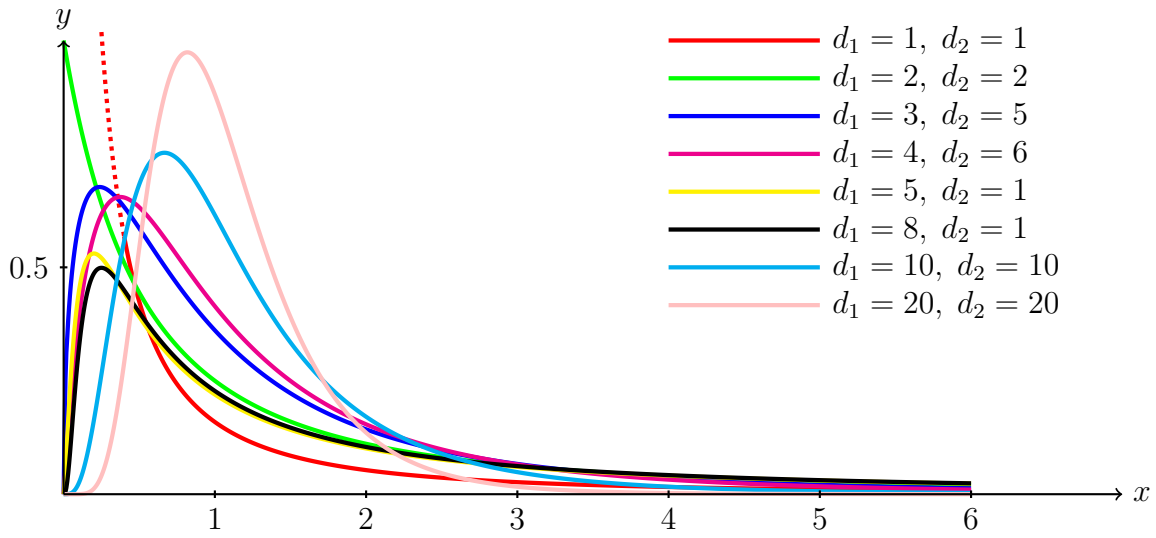
F-fördelning

Definition. Vi kallar $X \sim \mathrm{F}(d_1, d_2)$ **F-fördelad** med frihetsgraderna $d_1 > 0$ och $d_2 > 0$ om

$$f_X(x) = \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}\, x\right)^{-\frac{d_1+d_2}{2}}, \quad x \geq 0,$$

där $B(a,b) = \dfrac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}$ är beta-funktionen.

Notera att $X \sim \mathrm{F}(d_1, 1) \;\Leftrightarrow\; X \sim \chi^2(d_1)$.



Sats. Om $V_1 \sim \chi^2(d_1)$ och $V_2 \sim \chi^2(d_2)$ är oberoende så gäller att

$$\frac{V_1/d_1}{V_2/d_2} \sim F(d_1, d_2).$$

Bevis. Vi börjar med att betrakta hur man kan hitta täthetsfunktionen för kvoten $Z = X/Y$ av två oberoende stokastiska variabler X och Y. Vi antar att respektive täthetsfunktion är kontinuerlig. Det gäller att $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ och

$$\begin{aligned} F_Z(z) &= P\left(\frac{X}{Y} \leq z\right) = P(X \leq Yz,\, Y > 0) + P(X \geq Yz,\, Y < 0) \\ &= \int_0^\infty \int_{-\infty}^{yz} f_{X,Y}(x,y)dx\,dy + \int_{-\infty}^0 \int_{yz}^\infty f_{X,Y}(x,y)dx\,dy \\ &= \int_0^\infty f_Y(y) F_X(yz)\,dy + \int_{-\infty}^0 f_Y(y)(1 - F_X(yz))\,dy, \end{aligned}$$

från vilket det följer att

$$f_Z(z) = \frac{d}{dz}F_Z(z) = \int_0^\infty y f_Y(y) f_X(yz)\,dy + \int_{-\infty}^0 -y f_Y(y) f_X(yz)\,dy$$
$$= \int_{-\infty}^\infty |y| f_Y(y) f_X(yz)\,dy.$$

Vi noterar även att för $r > 0$ gäller att

$$P\left(\frac{X}{r} \le x\right) = P(X \le rx) \quad \Rightarrow \quad f_{X/r}(x) = r f_X(rx).$$

Således ges täthetsfunktionerna för V_1/d_1 och V_2/d_2 av

$$f_{V_1/d_1}(x) = \frac{d_1^{d_1/2}}{2^{d_1/2}\Gamma\left(\frac{d_1}{2}\right)} x^{d_1/2-1} e^{-d_1 x/2}, \quad x \ge 0$$

och

$$f_{V_2/d_2}(y) = \frac{d_2^{d_2/2}}{2^{d_2/2}\Gamma\left(\frac{d_2}{2}\right)} y^{d_2/2-1} e^{-d_2 y/2}, \quad y \ge 0,$$

så enligt resultatet ovan för kvoten $\dfrac{V_1/d_1}{V_2/d_2}$ erhåller vi att

$$\begin{aligned}
f_Z(z) &= \int_0^\infty y f_{V_2/d_2}(y) f_{V_1/d_1}(yz)\,dy \\
&= \frac{d_2^{d_2/2} d_1^{d_1/2}\, z^{d_1/2-1}}{2^{(d_1+d_2)/2}\,\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)} \int_0^\infty y^{d_1/2+d_2/2-1} e^{-(y(d_2+d_1 z))/2}\,dy \\
&= \Bigg/ \text{variabelbyte: } \begin{matrix} u = y(d_2 + d_1 z) \\ dy = (d_2 + d_1 z)^{-1} du \end{matrix} \Bigg/ \\
&= \frac{d_2^{d_2/2} d_1^{d_1/2}\, z^{d_1/2-1} (d_2 + d_1 z)^{-(d_1+d_2)/2}}{2^{(d_1+d_2)/2}\,\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)} \int_0^\infty u^{(d_1+d_2)/2-1} e^{-u/2}\,du \\
&= \frac{d_2^{d_2/2-1} d_1^{d_1/2-1}\Gamma\left(\frac{d_1+d_2}{2}\right)\, z^{d_1/2-1} (d_2 + d_1 z)^{-(d_1+d_2)/2}}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)}
\end{aligned}$$

eftersom

$$\frac{1}{2^{(d_1+d_2)/2}\,\Gamma\left(\frac{d_1+d_2}{2}\right)} \int_0^\infty u^{(d_1+d_2)/2-1} e^{-u/2}\,du = 1$$

då detta är integralen av täthetsfunktionen för en stokastisk variabel $U \sim \chi^2(d_1 + d_2)$. Vi kan hyffsa till slutresultatet för $f_Z(z)$ genom att bryta ut d_2 ur $(d_2 + d_1 z)^{-(d_1+d_2)/2}$ och använda beta-funktionen:

$$\begin{aligned}
f_Z(z) &= \frac{d_2^{d_2/2}\, z^{d_1/2-1} d_1^{-d_2/2} \left(1 + \frac{d_1}{d_2} z\right)^{-(d_1+d_2)/2}}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \\
&= \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_2}{d_1}\right)^{-d_2/2} z^{d_1/2-1} \left(1 + \frac{d_1}{d_2} z\right)^{-(d_1+d_2)/2},
\end{aligned}$$

vilket är precis vad vi ville visa. □

Sats. Om $X \sim \mathrm{F}(d_1, d_2)$ så är

$$E(X) = \frac{d_2}{d_2 - 2},\ d_2 > 2, \quad \text{och} \quad V(X) = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2 - 2)^2(d_2 - 4)},\ d_2 > 4.$$

Bevis. Välj två oberoende stokastiska variabler $V_1 \sim \chi^2(d_1)$ och $V_2 \sim \chi^2(d_2)$. Eftersom vi visade ovan att $\dfrac{V_1/d_1}{V_2/d_2} \sim F(d_1, d_2)$ följer det att

$$E(X) = E\left(\frac{V_1/d_1}{V_2/d_2}\right) = \Big/ \frac{V_1}{d_1} \text{ och } \frac{V_2}{d_2} \text{ oberoende } \Big/ = E\left(\frac{V_1}{d_1}\right) E\left(\frac{d_2}{V_2}\right) = \frac{d_1\, d_2}{d_1} E\left(\frac{1}{V_2}\right),$$

där vi nyttjat att $E(V_1) = d_1$. Vi beräknar $E(1/V_2)$:

$$\begin{aligned} E\left(\frac{1}{V_2}\right) &= c\int_0^\infty x^{d_2/2-2} e^{-x/2} dx = c\left(\left[\frac{1}{d_2/2-1}\, x^{d_2/2-1} e^{-x/2}\right]_0^\infty + \frac{1}{d_2-2}\int_0^\infty x^{d_2/2-1} e^{-x/2}\, dx\right) \\ &= \frac{1}{d_2-2}\int_0^\infty f_{V_2}(x)\, dx = \frac{1}{d_2-2}, \end{aligned}$$

under förutsättning att $d_2 > 2$. Således blir

$$E(X) = \frac{d_2}{d_2 - 2} \text{ om } d_2 > 2.$$

När det gäller variansen använder vi ett analogt resonemang:

$$\begin{aligned} V(X) &= V\left(\frac{V_1/d_1}{V_2/d_2}\right) = \frac{d_2^2}{d_1^2}\left(E\left(\frac{V_1^2}{V_2^2}\right) - E\left(\frac{V_1}{V_2}\right)^2\right) \\ &= \Big/ V_1 \text{ och } V_2 \text{ oberoende } \Big/ = \frac{d_2^2}{d_1^2}\left(E\left(V_1^2\right) E\left(\frac{1}{V_2^2}\right) - E\left(V_1\right)^2 E\left(\frac{1}{V_2}\right)^2\right) \\ &= \frac{d_2^2}{d_1^2}\left(\left(V\left(V_1\right) + E(V_1)^2\right) E\left(\frac{1}{V_2^2}\right) - \frac{d_1^2}{(d_2-2)^2}\right) \\ &= \frac{d_2^2}{d_1^2}\left(\left(2d_1 + d_1^2\right) E\left(\frac{1}{V_2^2}\right) - \frac{d_1^2}{(d_2-2)^2}\right) \end{aligned}$$

eftersom $V(V_1) = 2d_1$ och vi använt resultatet för $E(1/V_2)$ ovan. Vi partialintegrerar nu för att beräkna $E(1/V_2^2)$:

$$\begin{aligned} E\left(\frac{1}{V_2^2}\right) &= c\int_0^\infty x^{k/2-3} e^{-x/2}\, dx \\ &= c\left(\left[\frac{1}{d_2/2-2}\, x^{d_2/2-2} e^{-x/2}\right]_0^\infty + \frac{1}{d_2-4}\int_0^\infty x^{d_2/2-2} e^{-x/2}\, dx\right) \\ &= \frac{1}{d_2-4} E\left(\frac{1}{V_2}\right) = \frac{1}{(d_2-4)(d_2-2)}, \end{aligned}$$

om $d_2 > 4$ och vi nyttjat kalkylen för $E(1/V_2)$ ovan.
Alltså blir

$$V(X) = \frac{d_2^2}{d_1^2}\left(\frac{2d_1 + d_1^2}{(d_2-4)(d_2-2)} - \frac{d_1^2}{(d_2-2)^2}\right) = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2-2)^2(d_2-4)},$$

vilket var precis vad vi ville visa. □

Sats. Om $X \sim F(d_1, d_2)$ så är $1/X \sim F(d_2, d_1)$.

Bevis. Låt $V = 1/X$ och antag att $v > 0$. Då gäller att

$$F_V(v) = P(1/X \le v) = P(X \ge 1/v) = 1 - F_X(1/v) \quad \Rightarrow \quad f_V(v) = \frac{1}{v^2} f_X(1/v),$$

så

$$\begin{aligned} f_V(v) &= \frac{1}{v^2} \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_2}{d_1}\right)^{-d_2/2} \left(\frac{1}{v}\right)^{d_1/2-1} \left(1 + \frac{d_1}{d_2}\frac{1}{v}\right)^{-(d_1+d_2)/2} \\ &= \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_2}{d_1}\right)^{-d_2/2} \left(\frac{1}{v}\right)^{d_1/2+1} \left(\frac{d_1}{d_2 v}\right)^{-(d_1+d_2)/2} \left(\frac{d_2}{d_1}v + 1\right)^{-(d_1+d_2)/2} \\ &= \frac{1}{B\left(\frac{d_2}{2}, \frac{d_1}{2}\right)} \left(\frac{d_1}{d_2}\right)^{-d_1/2} v^{d_2/2-1} \left(1 + \frac{d_2}{d_1}v\right)^{-(d_2+d_1)/2} \end{aligned}$$

eftersom $B(a, b) = B(b, a)$. Således är $V \sim F(d_2, d_1)$. □

Sats. Om $T \sim t(n)$ så är $T^2 \sim F(1, n)$.

Bevis. Låt $V = T^2$ och antag att $v \ge 0$. Då gäller att

$$F_V(v) = P(T \le v) = P(-\sqrt{v} \le T \le \sqrt{v}) = F_T(\sqrt{v}) - F_T(-\sqrt{v}),$$

så

$$\begin{aligned} f_V(v) = F_V'(v) &= \frac{1}{2\sqrt{v}} f_T(\sqrt{v}) - \frac{-1}{2\sqrt{v}} f_T(-\sqrt{v}) \\ &= \frac{1}{\sqrt{v}} f_T(\sqrt{v}) = \frac{1}{\sqrt{v}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} v^{-1/2} \left(1 + \frac{v}{n}\right)^{-(n+1)/2} \\ &= \frac{\Gamma\left(\frac{n}{2} + \frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{1}{n}\right)^{1/2} v^{-1/2} \left(1 + \frac{v}{n}\right)^{-(n+1)/2} \\ &= \frac{1}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \left(\frac{1}{n}\right)^{1/2} v^{-1/2} \left(1 + \frac{1}{n}v\right)^{-(n+1)/2}, \end{aligned}$$

eftersom $f_T(-t) = f_T(t)$ och $\Gamma(1/2) = \sqrt{\pi}$. Således är $V \sim F(1, n)$. □

4.2 Jämförelse av två varianser

Låt $X_1, \ldots, X_{n_1}$ och $Y_1, \ldots, Y_{n_2}$ vara oberoende slumpmässiga stickprov från $N(\mu_1, \sigma_1^2)$ respektive $N(\mu_2, \sigma_2^2)$. Då vet vi att

$$\frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1) \quad \text{och} \quad \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1).$$

Det följer då enligt ovan att

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1).$$

Exempel

Betrakta det tidigare exempel igen, där vi hade

x_i	47.7	55.6	51.3	46.1	54.9			
y_i	29.2	47.8	30.9	37.7	27.9	40.1	41.5	40.9

Antag att x_i är oberoende observationer av $N(\mu_1, \sigma_1^2)$ och att y_i är oberoende observationer av $N(\mu_2, \sigma_2^2)$. Ange ett 95% konfidensintervall för σ_1/σ_2.

Lösning. Låt $F = \dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$. På grund av antagandet följer det att $F \sim F(4,7)$. Vi söker ett konfidensintervall med konfidensgrad 95% så vi behöver gränser a och b så att

$$P(F < a) = 0.025 \quad \text{och} \quad P(F > b) = 0.025.$$

Ur tabell finner vi att $a = 0.1102$ och $b = 5.5226$ (i MATLAB `finv([0.025 0.975], 4, 7)`). Notera att tabeller oftast endast innehåller värden för sannolikheter ≥ 0.5. Anledning till det är att vi kan använda att

$$F \sim F(m,n) \quad \Rightarrow \quad \frac{1}{F} \sim F(n,m).$$

Konkret för oss just nu blir det således

$$0.025 = P(F < a) = P\left(\frac{1}{a} < \frac{1}{F}\right) = 1 - P\left(\frac{1}{F} \leq \frac{1}{a}\right) \quad \Leftrightarrow \quad P\left(\frac{1}{F} \leq \frac{1}{a}\right) = 0.975.$$

Vi försöker nu lösa ut σ_1/σ_2:

$$a < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} < b \quad \Leftrightarrow \quad \frac{1}{b}\frac{S_1^2}{S_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{a}\frac{S_1^2}{S_2^2}.$$

Vi skattar nu S_1^2 och S_2^2 med respektive stickprovsvarians:

$$s_1^2 = 17.822 \quad \text{och} \quad s_2^2 = 49.0086.$$

Ett konfidensintervall för σ_1^2/σ_2^2 ges alltså av

$$I = \left(\frac{1}{5.5226}\frac{17.822}{49.0086}, \ \frac{1}{0.1102}\frac{17.822}{49.0086}\right) = (0.0658, \ 3.2999).$$

Vill vi ha ett konfidensintervall för σ_1/σ_2 tar vi helt enkelt roten ur gränserna:

$$I_{\sigma_1/\sigma_2} = (\sqrt{0.0658}, \ \sqrt{3.2999}) = (0.2566, \ 1.8165).$$

I MATLAB kan man använda funktionen `vartest2` för att skapa konfidensintervallet.

```
>> x = [47.7  55.6  51.3  46.1  54.9 ];
>> y = [29.2  47.8  30.9  37.7  27.9  40.1  41.5  40.9];
>> [H P CI] = vartest2(x,y,0.05,'both')
H =
     0
P =
    0.3452
CI =
    0.0658    3.2998
```

Vad H och P representerar kommer vi till på nästa föreläsning.

5 Konfidensintervall via CGS

Så vad gör vi om stickprovet inte är från en normalfördelning?

6 Stickprov för andel

Exempel

Ett företag som sysslar med opinionsanalys väljer slumpmässigt ut 400 vuxna i Sverige och frågar om de har åsikt A. Av dessa svarar 80 ja (alla svarar). Bestäm ett approximativt 95% konfidensintervall för andelen av den stora populationen som håller åsikt A.

Lösning. Vi låter X vara antalet som svarar ja. Då är egentligen $X \sim \text{Hyp}(N, 400, p)$, där N är antalet vuxna i Sverige (rimligen ca 8 miljoner). Då $400 \ll 8000000$ är det helt rimligt att anta att $X \overset{\text{appr.}}{\sim} \text{Bin}(400, p)$. Vi vill skatta den okända andelen p och väljer som skattningsvariabel

$$\widehat{P} = \frac{X}{400}$$

Vi har observerat att $\widehat{p} = 80/400 = 0.2$.
Binomialfördelningen är lite jobbig eftersom den är diskret, så vi försöker oss på en approximation. Eftersom

$$400 \cdot \widehat{p} \cdot (1 - \widehat{p}) = 400 \cdot 0.2 \cdot 0.8 = 64$$

är ordentligt större än 10 är det rimligt att approximera binomialfördelningen med normalfördelning. Alltså,

$$\widehat{P} \overset{\text{appr.}}{\sim} \text{N}(p, p(1-p)/400).$$

Låt oss bilda

$$Z = \frac{\widehat{P} - p}{\sqrt{\widehat{p}(1-\widehat{p})/400}} \overset{\text{appr.}}{\sim} N(0, 1).$$

Observera att vi ersatt med det skattade värdet på p i kvadratroten (men **inte** i täljaren). Vi nyttjar här alltså medelfelet d, dvs

$$d(\widehat{P}) = \sqrt{\widehat{p}(1-\widehat{p})/400} = 0.02.$$

Vi kan nu räkna precis som om vi känner standardavvikelsen exakt, så om vi söker ett approximativt 95% K.I. erhåller vi

$$I_p = (0.2 - 1.96 \cdot 0.02,\ 0.2 + 1.96 \cdot 0.02) = (0.16, 0.24).$$

7 Jämförelse av två andelar

Antag att vi har två maskiner. Vid uppmätning fann man att Maskin 1 producerade 20 defekta enheter av 400, och att Maskin 2 producerade 60 defekta enheter av 600.
Modell: Låt X vara antal defekta enheter från Maskin 1 och Y antal defekta enheter från Maskin 2. Under lämpligt oberoendeantagande vet vi att $X \sim \text{Bin}(400, p_1)$ och $Y \sim \text{Bin}(600, p_2)$ där p_1 och p_2 är de verkliga felsannolikheterna. Vi skattar lämpligen med

$$\widehat{P_1} = \frac{X}{400} \qquad \text{och} \qquad \widehat{P_2} = \frac{Y}{600}.$$

Vi har observerat att $\widehat{p_1} = 20/400 = 0.05$ och $\widehat{p_2} = 60/600 = 0.10$. Alltså är $\widehat{p_1} - \widehat{p_2} = -0.05$. Är detta signifikant? För att svara på frågan behöver vi räkna lite sannolikheter. Eftersom både $n_1\widehat{p_1}(1-\widehat{p_1})$ och $n_2\widehat{p_2}(1-\widehat{p_2})$ är mycket större än 10 är det rimligt att approximera binomialfördelningen med normalfördelning. Alltså,

$$\widehat{P_1} \overset{\text{appr.}}{\sim} \mathrm{N}\left(p_1, \frac{p_1(1-p_1)}{400}\right) \qquad \text{och} \qquad \widehat{P_2} \overset{\text{appr.}}{\sim} \mathrm{N}\left(p_2, \frac{p_2(1-p_2)}{600}\right).$$

Då följer det att

$$\widehat{P_1} - \widehat{P_2} \overset{\text{appr.}}{\sim} \mathrm{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{400} + \frac{p_2(1-p_2)}{600}\right).$$

Vi bildar nu

$$Z = \frac{\widehat{P_1} - \widehat{P_2} - (p_1 - p_2)}{\sqrt{\widehat{p_1}(1-\widehat{p_1})/400 + \widehat{p_2}(1-\widehat{p_2})/600}} \overset{\text{appr.}}{\sim} N(0,1).$$

Observera att vi ersatt med skattade värden på p_1 och p_2 i kvadratroten (men **inte** i täljaren). Det blir fortfarande approximativt (men lite sämre så klart) normalfördelat, men underlättar mycket för beräkningar. Vi har

$$\sqrt{\widehat{p_1}(1-\widehat{p_1})/400 + \widehat{p_2}(1-\widehat{p_2})/600} = 0.0164.$$

Vi kan nu räkna precis som om vi känner standardavvikelsen exakt, så om vi söker ett approximativt 95% K.I. erhåller vi

$$I_{p_1-p_2} = (-0.05 - 1.96 \cdot 0.0164,\ -0.05 + 1.96 \cdot 0.0164) = (-0.08, -0.02).$$

Endast negativa värden, så $p_1 < p_2$ med hög sannolikhet! Maskin 2 är antagligen sämre.