Föreläsning 8: Konfidensintervall (forts)

Johan Thim (johan.thim@liu.se)

January 10, 2022

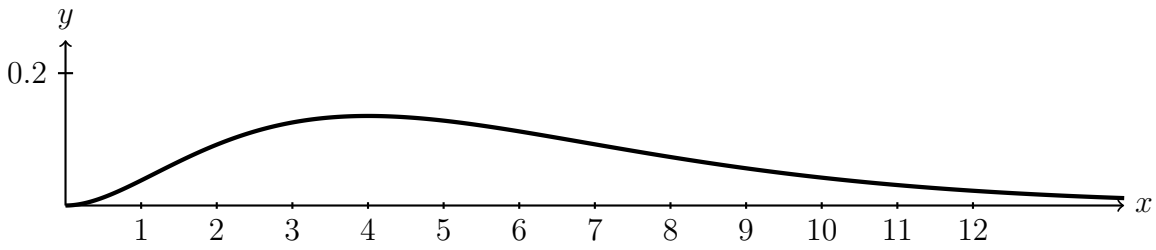
1 χ^2- och t-fördelning

För att hantera situationen när vi inte känner till variansen för den normalfördelning som vi tagit stickprovet ifrån så måste vi introducera ett par nya fördelningar. Först och främst formulerar vi följande sats där χ^2-fördelningen finns i mer detaljer i avsnitt 11. Fördelningen uppstår alltså när vi summerar kvadrater av oberoende normalfördelningar.

Sats. Om $X_1, X_2, \ldots, X_n$ är oberoende och $X_k \sim N(0,1)$ så är

$$\sum_{k=1}^{n} X_k^2 \sim \chi^2(n).$$

Vi kan skissa ett exempel på hur täthetsfunktionen ser ut (här med $n = 6$). Utseendet beror givetvis på parametern n (som brukar kallas för antalet frihetsgrader).



Så varför är vi intresserade av denna kvadratsumma? Tänk på hur vi definierat stickprovsvariansen. Man kan visa (se avsnitt 14) följande sats.

Sats. Låt $X_1, X_2, \ldots, X_n$ vara oberoende likafördelade stokastiska variabler där $X_k \sim N(\mu, \sigma)$ för $k = 1, 2, \ldots, n$. Då gäller att

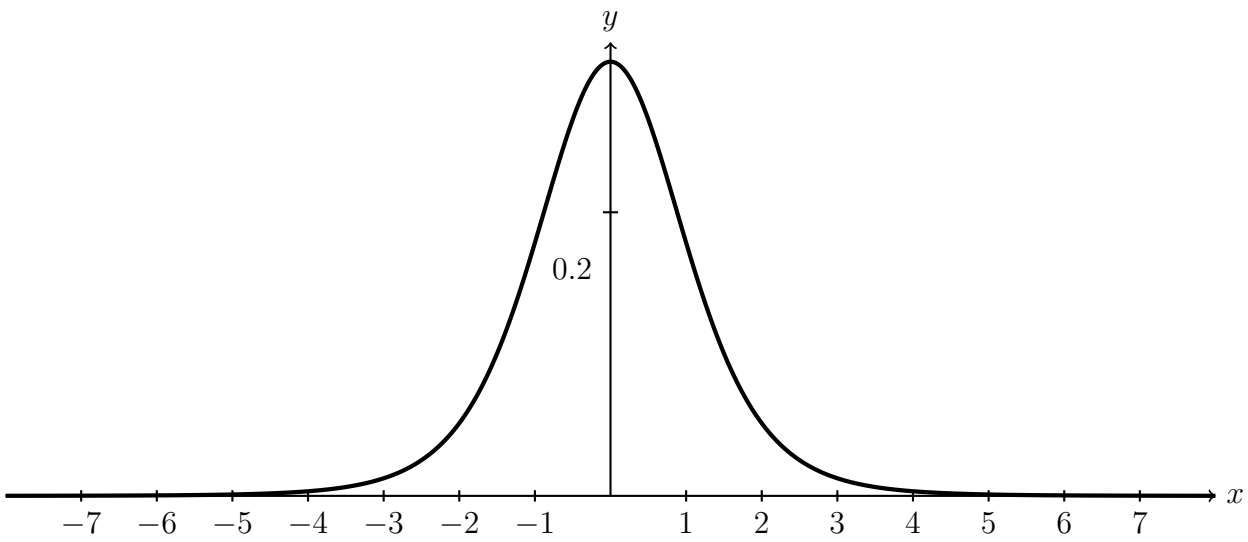
$$\frac{1}{\sigma^2}\sum_{k=1}^{n}(X - \overline{X})^2 \sim \chi^2(n-1).$$

Anledningen till att vi får $n-1$ trots att det är n termer i summan beror på att medelvärdet $\overline{X}$ inte är oberoende av $X_1, X_2, \ldots, X_n$. Det faktum att det blir just $n-1$ är inte självklart (se avsnitt 14).

Med dessa resultat på plats så introducerar vi en till ny fördelning: t-fördelningen. Vi nöjer oss med att karaktärisera den som den fördelning som uppstår om kvoten $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ bildas. Se avsnitt 12 för mer detaljer.

Sats. Om $X_1, X_2, \ldots, X_n$ är oberoende och likafördelade stokastiska variabler med fördelningen $N(\mu, \sigma)$ så är $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$, där S^2 är stickprovsvariansen.

Vi kan även skissa ett exempel på hur täthetsfunktionen ser ut för t-fördelningen ($n = 6$). Likt χ^2-fördelningen så beror utseendet givetvis på parametern n (som även här brukar kallas för antalet frihetsgrader).



Tätheten är symmetrisk kring origo och påminner en hel del om normalfördelning. Faktum är att om $n \to \infty$ så får vi tillbaka normalfördelningen.

2 Konfidensintervall för μ i normalfördelning

2.1 Konfidensintervall för μ när σ är känd

Vi tog fram detta resultat förra föreläsningen så låt oss bara kortfattat beskriva vad vi gjorde. Låt $x_1, x_2, \ldots, x_n$ vara ett stickprov från en $N(\mu, \sigma)$-fördelning där vi känner σ och vill hitta ett konfidensintervall för μ. En punktskattning för väntevärdet ges av

$$\widehat{M} = \bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \sim N(\mu, \sigma/\sqrt{n}).$$

Vi skapar testvariabeln

$$Z = \frac{\widehat{M} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Det följer då att vi kan välja ett tal $\lambda_{\alpha/2}$ så att $P(-\lambda_{\alpha/2} < Z < \lambda_{\alpha/2}) = 1 - \alpha$. Om man löser ut μ ur olikheten i sannolikhetsmåttet finner vi till slut intervallet, med konfidensgrad $1 - \alpha$,

$$I_\mu = \left(\overline{x} - \lambda_{\alpha/2} \frac{\sigma}{\sqrt{n}},\ \overline{x} + \lambda_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right),$$

där vi ersatt $\widehat{M}$ med den observerade punktskattningen $\widehat{\mu} = \overline{x}$ (medelvärdet av observationerna) för att få ett konfidensintervall; se föregående föreläsning för detaljerna.

2.2 Okänd varians

Låt $x_1, x_2, \ldots, x_n$ vara ett stickprov från en $N(\mu, \sigma)$-fördelning där vi *inte* vet vad σ är och vi vill hitta ett konfidensintervall för μ. En punktskattning för väntevärdet ges av

$$\widehat{M} = \bar{X} = \frac{1}{n} \sum_{k=1}^{n} X_k \sim N(\mu, \sigma/\sqrt{n}).$$

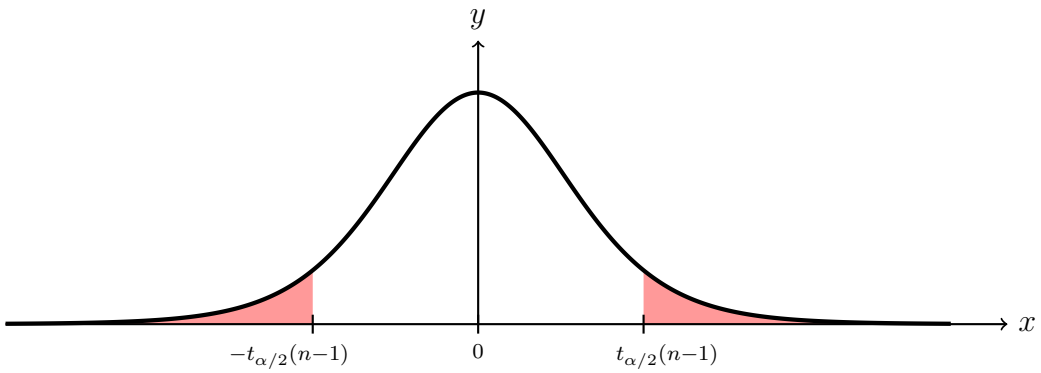
Eftersom σ är okänd behöver vi en skattning och förslagsvis väljer vi stickprovsstandardavvikelsen. Vi skapar sedan testvariabeln

$$T = \frac{\widehat{M} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

där faktumet att T är t-fördelad följer från Gossets sats. Det följer då att vi kan välja ett tal $t_{\alpha/2}$ så att

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

Talet $t_{\alpha/2}(n-1)$ är $\alpha/2$-kvantilen för en $t(n-1)$-fördelning (vi finner denna i tabell).



Vi löser ut μ ur olikheten i sannolikhetsmåttet i ekvation (1) ovan:

$$\begin{aligned} -t_{\alpha/2}(n-1) < T < t_{\alpha/2}(n-1) \quad &\Leftrightarrow \quad -t_{\alpha/2}(n-1)\,\frac{S}{\sqrt{n}} < \widehat{M} - \mu < t_{\alpha/2}(n-1)\,\frac{S}{\sqrt{n}} \\ &\Leftrightarrow \quad \widehat{M} - t_{\alpha/2}(n-1)\,\frac{S}{\sqrt{n}} < \mu < \widehat{M} + t_{\alpha/2}(n-1)\,\frac{S}{\sqrt{n}} \end{aligned}$$

Om vi ersätter $\widehat{M}$ med den observerade punktskattningen $\widehat{\mu} = \overline{x}$ (medelvärdet av observationerna) och S med stickprovsstandardavvikelsen så får vi ett konfidensintervall

$$I_\mu = \left(\overline{x} - t_{\alpha/2}(n-1)\,\frac{s}{\sqrt{n}},\ \overline{x} + t_{\alpha/2}(n-1)\,\frac{s}{\sqrt{n}} \right)$$

med konfidensgrad $1 - \alpha$.

Exempel

Samma exempel som tidigare där man vid en mätning av en process fick man följande mätdata:

6.04 4.96 4.93 3.40 7.04 4.73 3.57 7.70 4.55 3.82

En som arbetar med processen håller inte med om att standardavvikelsen kan antas vara given, utan tycker att man måste skatta den utifrån datan. Hjälp personen i fråga med att ställa upp ett 99% konfidensintervall för väntevärdet μ då mätningarna är ett stickprov på en normalfördelad variabel $X \sim N(\mu, \sigma)$ och σ är okänd.

Lösning. Vi betraktar siffrorna som ett stickprov från oberoende s. v. $X_j \sim N(\mu, \sigma)$. Vi punktskattar med $\widehat{M} = \bar{X} \sim N(\mu, \sigma/\sqrt{10})$ som tidigare och skattar σ med s, där

$$s^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})^2 \approx 2.0842$$

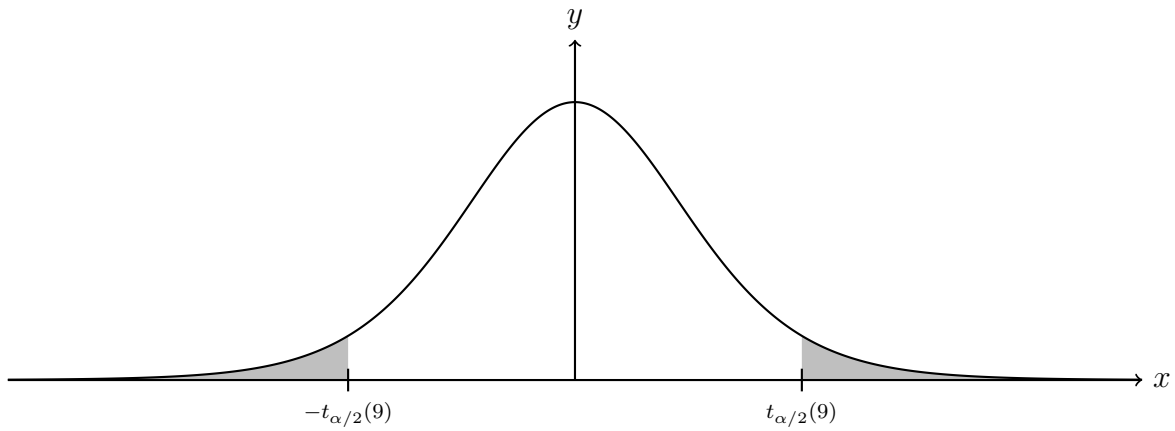
är stickprovsvariansen. Vi skapar testvariabeln

$$T = \frac{\widehat{M} - \mu}{S/\sqrt{10}} \sim t(9).$$

Som i förra deluppgiften följer det att

$$P\big(-t_{\alpha/2}(9) < T < t_{\alpha/2}(9)\big) = 1 - \alpha,$$

där $t_\beta(9)$ är kvantilerna till $t(9)$-fördelningen, $\beta \in [0, 1]$.



Ur tabell finner vi $t_{0.005}(9) = 3.25$. Genom att lösa ut μ ur olikheten i sannolikhetsmåttet får vi

$$\widehat{M} - \frac{3.25 \cdot S}{\sqrt{10}} < \mu < \widehat{M} + \frac{3.25 \cdot S}{\sqrt{10}}.$$

Om vi ersätter $\widehat{M}$ med de observerade punktskattningarna $\widehat{\mu} = 5.074$ (medelvärdet av observationerna) och $s = \sqrt{2.0842} = 1.444$ (stickprovsstandardavvikelsen) så får vi ett konfidensintervall $I_\mu = (3.59, 6.56)$ med konfidensgrad 99%.

3 Prediktionsintervall

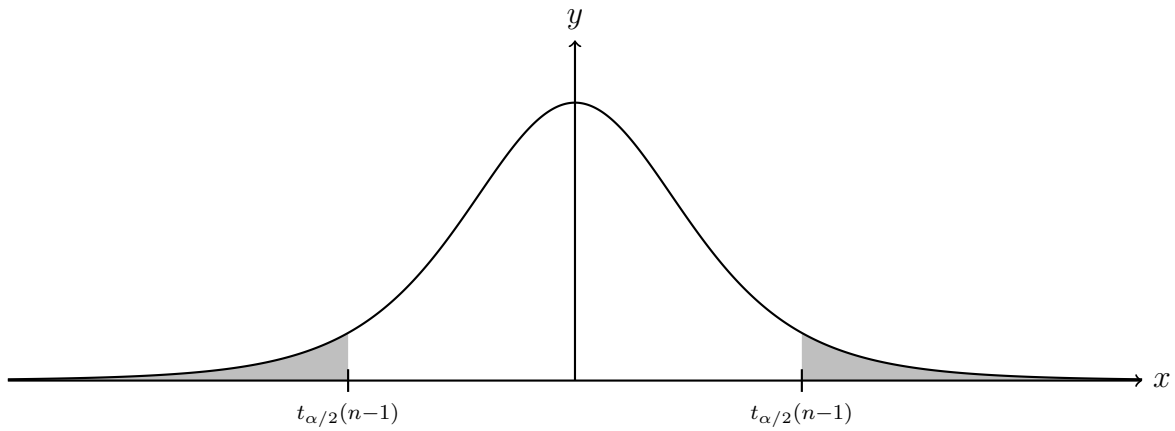
Vi har hittat konfidensintervall för väntevärdet, men kan man säga något om var en enskild observation hamnar? Låt $X_1, X_2, \ldots, X_n$ vara ett slumpmässigt stickprov från $N(\mu, \sigma)$. Vi vill stänga in en enskild observation av en variabel X_0 (som antages vara oberoende) från denna fördelning. Givetvis vill vi utnyttja stickprovet, så vi betraktar variabeln $X_0 - \overline{X}$ som är normalfördelad med

$$E(X_0 - \overline{X}) = 0 \quad \text{och} \quad V(X_0 - \overline{X}) = \sigma^2 + \frac{\sigma^2}{n}.$$

Alltså kommer

$$T = \frac{X_0 - \overline{X}}{S\sqrt{1 + \frac{1}{n}}} \sim t(n-1),$$

eftersom S^2 fortfarande är $\chi^2(n-1)$-fördelad. Vi kan på samma sätt som tidigare stänga in denna variabel med sannolikhet $1 - \alpha$,



och sedan lösa ut X_0:

$$-t_{\alpha/2}(n-1) < \frac{X_0 - \overline{X}}{S\sqrt{1 + \frac{1}{n}}} < t_{\alpha/2}(n-1)$$

$$\Leftrightarrow \quad \overline{X} - t_{\alpha/2}(n-1)\, S\sqrt{1 + \frac{1}{n}} < X_0 < \overline{X} + t_{\alpha/2}(n-1)\, S\sqrt{1 + \frac{1}{n}}.$$

Vi ersätter nu $\overline{X}$ med det observerade medelvärdet $\overline{x}$ och S med stickprovsstandardavvikelsen s och får då intervallet

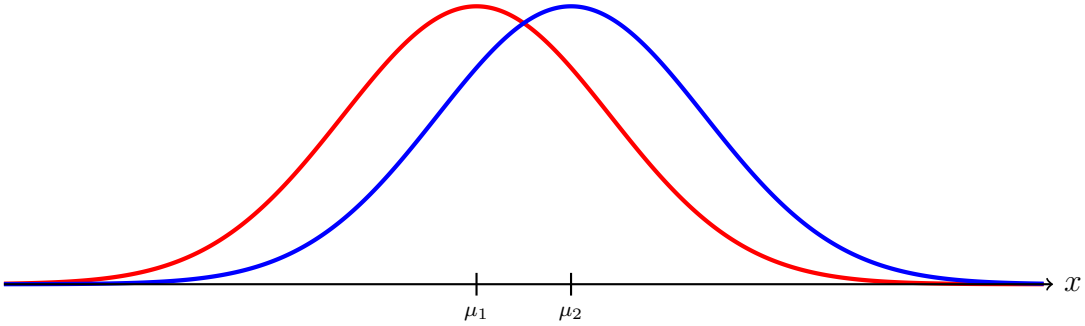
$$I_{X_0} = \left(\overline{x} - t_{\alpha/2}(n-1)\, s\sqrt{1 + \frac{1}{n}},\ \overline{x} + t_{\alpha/2}(n-1)\, s\sqrt{1 + \frac{1}{n}} \right).$$

Om vi jämför intervallen för väntevärde respektive predikterat värde ser vi att $I_{X_o} \subset I_\mu$ alltid gäller med den metod vi använt ovan. Ett prediktionsintervall blir alltså alltid bredare om vi utgår från samma stickprov och samma konfidensgrad.

Se upp med om en fråga ställs angående konfidensintervall för väntevärde eller ett prediktionsintervall. Det är helt olika frågor! Svarar du med ett konfidensintervall för väntevärdet när det efterfrågas ett prediktionsintervall blir det noll poäng.

4 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 framför allt 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?

5 Linjärkombinationer av normalfördelningar

Låt $X_1, \ldots, X_m$ och $Y_1, \ldots, Y_n$ vara oberoende slumpmässiga stickprov från $N(\mu_1, \sigma_1)$ respektive $N(\mu_2, \sigma_2)$. 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

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

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

5.1 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 $X_i \sim \mathrm{N}(\mu_1, 4)$ och y_i observationer av stokastiska variabler $Y_i \sim \mathrm{N}(\mu_2, 9)$, 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

$$E(W) = E(\overline{X}) - 2E(\overline{Y}) = \mu_1 - 2\mu_2,$$

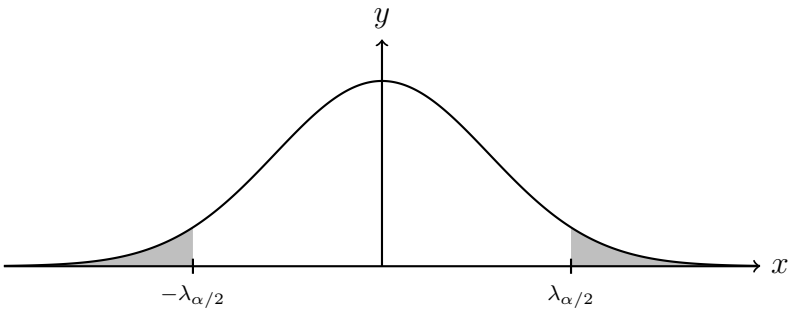
vilket är precis vad vi är intresserade av. Vidare är

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

En lämplig teststorhet ges av

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

Dags för en 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 $w = \overline{x} - 2\overline{y} = 51.12 - 2 \cdot 37 = -22.88$. Ur tabell finner vi att $\Phi(1.96) = 0.95 + 0.025 = 0.975$, så $\lambda_{\alpha/2} = 1.96$. Alltså blir intervallet

$$\begin{aligned} I_{\mu_1 - 2\mu_2} &= \left(-22.88 - 1.96 \cdot \sqrt{43.7},\ -22.88 + 1.96 \cdot \sqrt{43.7}\right) \\ &= (-35.84,\ -9.92). \end{aligned}$$

Vad säger detta oss? Jo, att med 95% säkerhet så ligger det verkliga värdet för $\mu_1 - 2\mu_2$ i intervallet $(-35.84, -9.92)$. 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!

5.2 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 (2) 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 $X_i \sim \mathrm{N}(\mu_1, \sigma)$ och $Y_i \sim \mathrm{N}(\mu_2, \sigma)$. 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

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

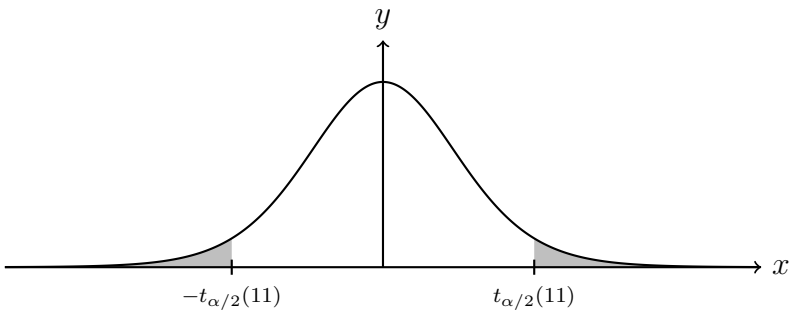
Låt $k := \sqrt{\frac{c_1^2}{n} + \frac{c_2^2}{m}}$. Snarlikt med fallet där vi kände varianserna kan vi stänga in T:

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

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 $\mu_1 > \mu_2$.

5.3 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 $\mu_1 - \mu_2$.

6 Enkelsidiga konfidensintervall

De konfidensintervall vi arbetat med har varit tvåsidiga i den meningen att båda gränserna har varit observationer av stokastiska variabler. Det innebär att vi lagt ut den osäkerhet vi tillåter på båda ”svansarna” i fördelningen. Men det är givetvis inte nödvändigt. Kanske är vi bara intresserade av gränsen åt ena hållet?

Typexemplet är konfidensintervall för variansen (även om vi inte hinner dit i denna kurs). Att variansen är liten brukar inte vara något större bekymmer, så vi lägger allt krut på att hålla koll på gränsen uppåt. Men det kan även handla om väntevärdet (eller ett predikterat värde). Kanske mäter vi något där vi inte får överstiga en viss nivå. Kanske en situation där det inga problem är om koncentrationen av något skadligt ämne är låg, men ett betydligt större problem om koncentrationen är hög?

Så hur åstadkommer vi detta? Vi betraktar ett exempel.

Exempel

Belinda är en hobbykemist som experimenterar med organiska peroxider. Hon försöker syntetisera hexametylentriperoxidiamin (HMTD) med två snarlika metoder. Den första använder citronsyra medan den andra använder isättika. Belinda är intresserad om utbytet blir bättre med citronsyra för att se om det är värt det extra besväret då denna metod producerar mer värme och kräver större försiktighet. Hon har gjort 10 experiment för varje metod och utbytet (beräknat som en kvot med mängden hexametylendiamin som används) i procent avrundat till heltal kan ses nedan. Vi antar att mätningar är normalfördelade och att olika tillverkningsomgångar är oberoende. Vi antar också att variansen är densamma för båda metoderna (rimligt?).

	Utbyte										$\overline{x}$	s
Citronsyra	55	36	55	64	53	58	55	45	51	40	51.2	8.5088
Isättika	50	38	39	40	27	54	47	40	53	35	42.3	8.5641

Utför ett test med åtminstone ett konfidensintervall för att se om metoden med citronsyra ger ett bättre utbyte än isättika. Använd konfidensgraden 90%.

Lösning. Modellen för citronsyran är att $X_i \sim N(\mu_1, \sigma)$ och för isättikan gäller $Y_i \sim N(\mu_2, \sigma)$ (samma varians). Alla variabler antas vara oberoende.

Vi viktar ihop varianserna:

$$s^2 = \frac{9s_1^2 + 9s_2^2}{18} = \frac{1}{2}\left(s_1^2 + s_2^2\right).$$

Det följer nu av Cochrans and Gossets satser att

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{10} + \frac{1}{10}}} \sim t(18),$$

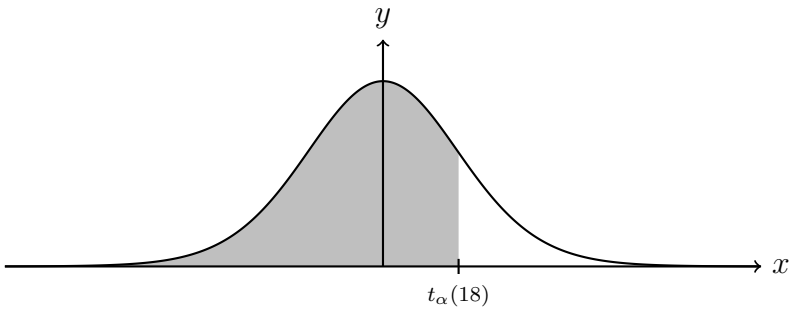
och vi vill att

$$P(T < t_\alpha(18)) = 1 - \alpha,$$

där vi då kan lösa olikheten för att finna att

$$\overline{X} - \overline{Y} - t_\alpha(18) \cdot \frac{S}{\sqrt{5}} < \mu_1 - \mu_2.$$

Vi använder ett enkelsidigt intervall eftersom vi endast vill undersöka om $\mu_1 > \mu_2$. Från en tabell finner vi att $t_{0.10}(18) = 1.3304$.



Som en observation av S använder vi $\sqrt{s^2}$, så

$$t_{0.10}(18)\frac{s}{\sqrt{5}} = 1.3304 \cdot 3.8176 = 5.0790.$$

Eftersom $\overline{x} - \overline{y} = 8.9$ så ges därmed det eftersökta intervallet av

$$\begin{aligned} I_{\mu_1-\mu_2} &= (8.9 - 5.0790,\, \infty) \\ &= (3.821,\, \infty). \end{aligned}$$

Här ser vi att 0 inte är med i intervallet, så vi kan säga att citronsyran verkar ge bättre utbyte (det är rimligt att $\mu_1 > \mu_2$) på den här signifikansnivån.

7 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 $X_k \sim \mathrm{N}(\mu_k, \sigma_1)$ och $Y_k \sim \mathrm{N}(\mu_k + \Delta, \sigma_2)$. Typexemplet är när vi mäter något före och efter en förändring.

Bilda nu ett ”nytt” stickprov Z_k av oberoende variabler:

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

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 (någon enhet) före och efter behandling.

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 funktionen?

Lösning: Låt x_i vara värde före behandling för person i och y_i motsvarande värde efter behandling. Vi antar att olika personer är oberoende, att x_i är observationer av $X_i \sim \mathrm{N}(\mu_i, \sigma_1)$ och att y_i är observationer av $Y_i \sim \mathrm{N}(\mu_i + \Delta, \sigma_2)$. Bilda $Z_i = Y_i - X_i \sim \mathrm{N}(\Delta, \sigma)$. 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 från tabell finner vi att $t_{\alpha/2}(8) = t_{0.005}(8) = 3.36$. 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.

8 Konfidensintervall via CGS

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

9 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 $X \sim \mathrm{Hyp}(N, 400, p)$, där N är antalet vuxna i Sverige (rimligen ca 8 miljoner). Då $400 \ll 8000000$ är det helt rimligt att anta att $X \overset{\text{appr.}}{\sim} \mathrm{Bin}(400, p)$. 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,

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

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

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

10 Jämförelse av två andelar

Exempel

Anonyme Alva har tillgång till två maskiner för att pressa piller. I dessa pressar Alva en perfekt homogeniserad blandning av obsykra bensodiazepiner med en tillsats av fentanyl för lite extra skjuts. Vid uppmätning fann man att Maskin 1 producerade 20 defekta enheter av 400 (där totala mängden fentanyl blir farligt hög för opiatnaiva individer), och att Maskin 2 producerade 60 defekta enheter av 600. Är det någon skillnad på andelen felaktikga piller producerade med de olika maskinerna? Svara med approximativt signifikansnivån 5%.

Lösning. 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,

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

Då följer det att

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

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.

11 $(\star\star)\chi^2$-fördelningen

En situation som dyker upp frekvent i statistik inferens är summor av kvadrater av normalfördelade variabler, så en naturlig fråga är så klart vilken fördelning en sådan summa får (åtminstone då variablerna antas vara oberoende). Svaret fås i form av χ^2-fördelningen.

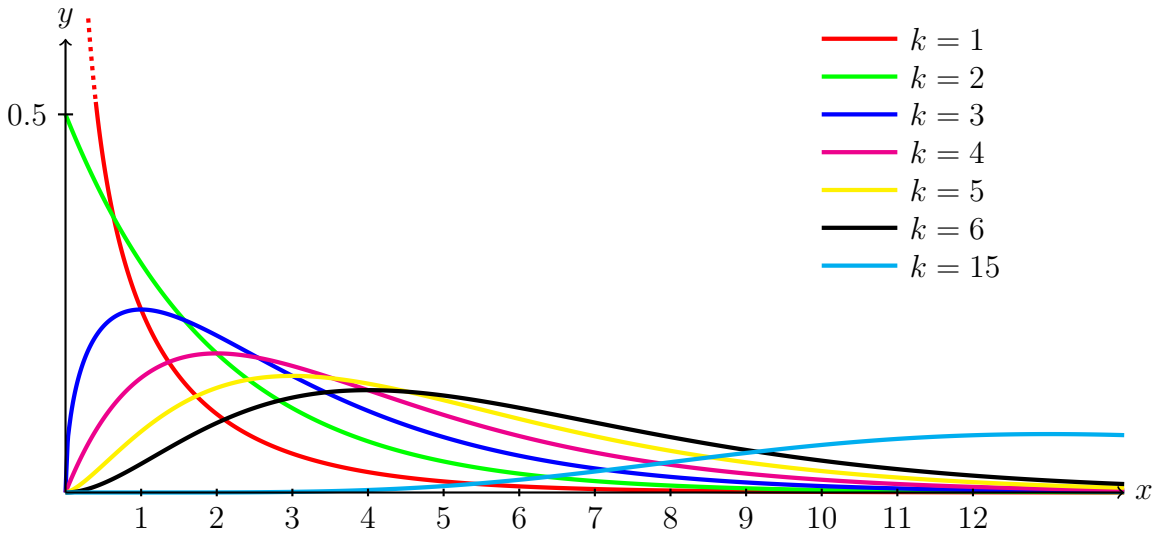
χ^2-fördelning

Definition. Om X är en stokastisk variabel med täthetsfunktionen

$$f_X(x) = \frac{1}{2^{k/2}\Gamma(k/2)}\, x^{k/2-1}e^{-x/2}, \quad x \geq 0 \text{ om } k > 1,$$

kallar vi X för $\chi^2(k)$-fördelad med k frihetsgrader, där $k = 1, 2, \ldots$

Här är Γ gamma-funktionen[1] och $\Gamma(n) = (n-1)!$ och $\Gamma(n+1/2) = \dfrac{(2n)!}{4^n\, n!}\sqrt{\pi}$ om $n \in \mathbf{N}$.



Om $X \sim \chi^2(k)$ är $E(X) = k$ och $V(X) = 2k$.

Bevis. Låt täthetsfunktionen skrivas $f(x) = cx^{k/2-1}e^{-x/2}$. Då gäller att

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

På samma sätt följer att

$$E(X^2) = c\int_0^\infty x^{k/2+1}e^{-x/2}\,dx = 2\left(\frac{k}{2}+1\right)c\int_0^\infty x^{k/2}e^{-x/2}\,dx = (k+2)E(X) = k^2+2k,$$

så $V(X) = E(X^2) - E(X)^2 = k^2 + 2k - k^2 = 2k$. □

[1] Se avsnitt 16 nedan för mer detaljer.

Sats. Om $X \sim \chi^2(\nu_1)$ och $Y \sim \chi^2(\nu_2)$ är oberoende så är $X + Y \sim \chi^2(\nu_1 + \nu_2)$.

Bevis. Enklast är att betrakta Fouriertransformen för täthetsfunktionen (alternativt den närbesläktade **karakteristiska funktionen** definierad enligt $E(e^{itX})$). Det är nämligen så att

$$\mathcal{F}(f_X)(t) = (1+2it)^{-\nu_1/2}, \quad \mathcal{F}(f_Y)(t) = (1+2it)^{-\nu_2/2}$$

och

$$\mathcal{F}(f_X * f_Y) = \mathcal{F}(f_X)\mathcal{F}(f_Y) = (1+2it)^{-(\nu_1+\nu_2)/2},$$

så $f_{X+Y} \sim \chi^2(\nu_1 + \nu_2)$. □

Sats. Om $X_1, X_2, \ldots, X_n$ är oberoende och $X_k \sim N(0,1)$ så är

$$\sum_{k=1}^{n} X_k^2 \sim \chi^2(n).$$

Bevis. Eftersom variablerna är oberoende ges den simultana täthetsfunktionen av

$$f(x_1, \ldots, x_n) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} = \frac{1}{2^{n/2}\pi^{n/2}} \exp\left(-\frac{1}{2}\left(x_1^2 + \cdots + x_n^2\right)\right).$$

Vi söker fördelningen för $Z = X_1^2 + \cdots X_n^2$, så låt oss ställa upp fördelningsfunktionen:

$$\begin{aligned} F_Z(z) = P(Z \le z) &= \int_{x_1^2+\cdots x_n^2 \le z} f(x_1, \ldots, x_n) dx_1\, dx_2 \cdots dx_n \\ &= \frac{1}{2^{n/2}\pi^{n/2}} \int_{S^{n-1}} \int_0^{\sqrt{z}} r^{n-1} e^{-r^2/2}\, dr\, dS \\ &= \frac{1}{2^{n/2}\pi^{n/2}} \frac{2\pi^{n/2}}{\Gamma(n/2)} \int_0^{\sqrt{z}} r^{n-1} e^{-r^2/2}\, dr \\ &= \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^{z} t^{n/2-1} e^{-t/2}\, dt, \end{aligned}$$

där S^{n-1} är enhetssfären i $\mathbf{R}^n$ och dS är ytmåttet på S^{n-1}. Då enhetssfären har ytmåttet $|S^{n-1}| = \dfrac{2\pi^{n/2}}{\Gamma(n/2)}$ följer likheten ovan efter ett variabelbyte i sista integralen (låt $t = r^2$). Analysens huvudsats medför nu att (för $z > 0$) att

$$f_Z(z) = F_Z'(z) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}.$$

För $z < 0$ är givetvis $f_Z(z) = 0$ (varför?). □

12 $(\star)$*t*-fördelningen

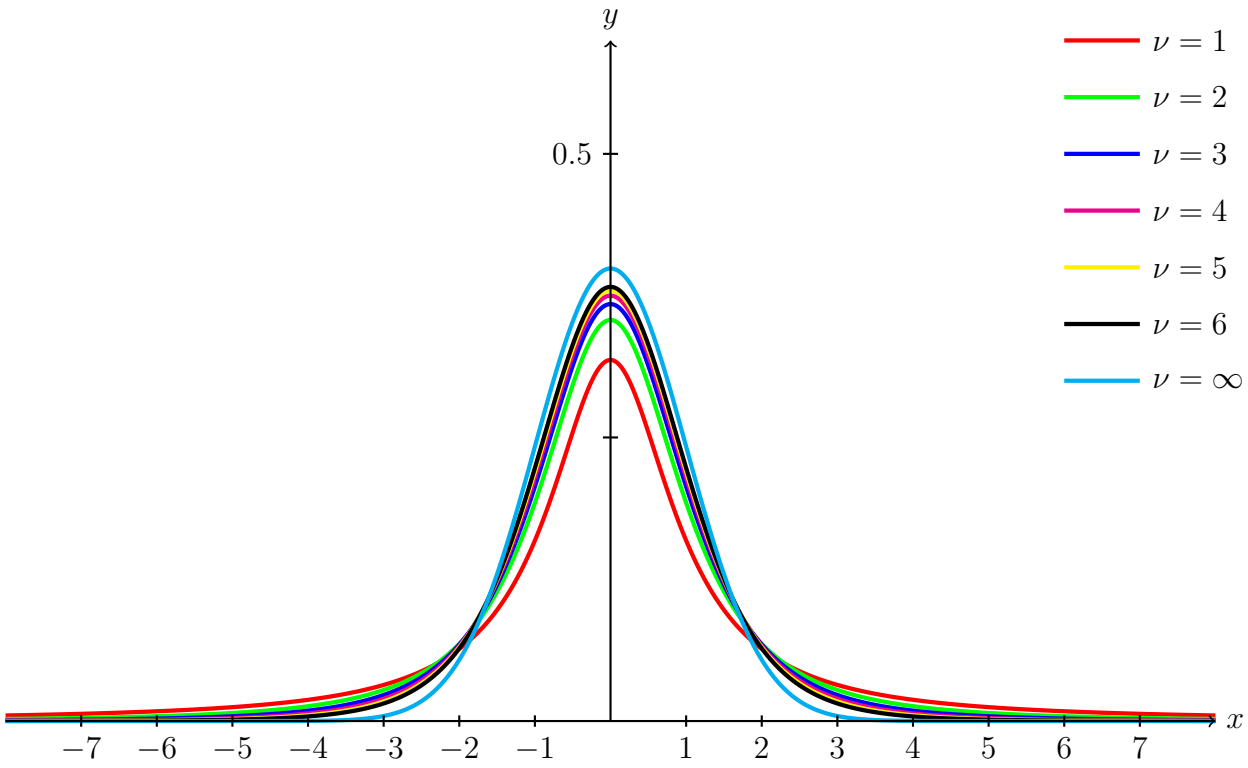
t-fördelning

Definition. Om X är en stokastisk variabel med täthetsfunktionen

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad x \in \mathbf{R} \text{ och } \nu > 0,$$

kallar vi X för $t(\nu)$-fördelad med ν frihetsgrader.

Denna fördelning är symmetrisk och om antalet frihetsgrader går mot oändligheten konvergerar täthetsfunktionen mot täthetsfunktionen för normalfördelning.



Sats. Om $X \sim t(\nu)$ är $E(X) = 0$ (om $\nu > 1$) och $V(X) = \nu/(\nu-2)$ (om $\nu > 2$).

Bevis. Om $\nu > 1$ är integralen $E(X)$ absolutkonvergent (visa det) och då integranden är udda blir således $E(X) = 0$. För att beräkna $E(X^2)$ låter vi $c_\nu = \dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}$.

Om $\nu > 2$ ser vi genom partialintegration att

$$\begin{aligned}
E(X^2) &= c_\nu \int_{-\infty}^{\infty} x \cdot x \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} dx = c_\nu \int_{-\infty}^{\infty} \frac{\nu}{\nu-1} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu-1}{2}} dx \\
&= c_\nu \frac{\nu}{\nu-1} \int_{-\infty}^{\infty} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu-1}{2}} dx \\
&= \frac{c_\nu}{c_{\nu-2}} \frac{\nu^{3/2}}{(\nu-1)\sqrt{\nu-2}} c_{\nu-2} \int_{-\infty}^{\infty} \left(1 + \frac{u^2}{\nu-2}\right)^{-\frac{\nu-1}{2}} du = \frac{c_\nu}{c_{\nu-2}} \frac{\nu^{3/2}}{(\nu-1)\sqrt{\nu-2}}
\end{aligned}$$

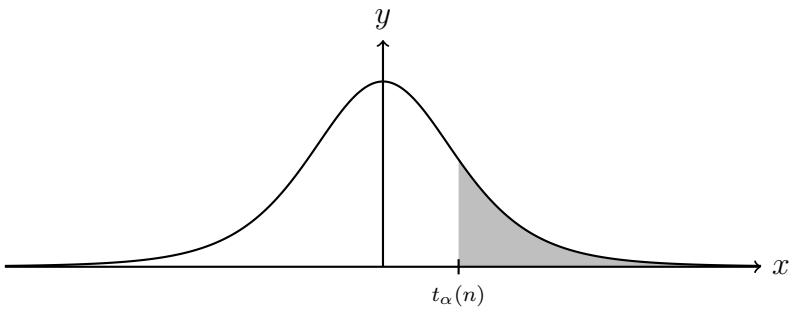
där vi bytte variabel så $x\sqrt{\nu-2} = u\sqrt{\nu}$ och utnyttjade att integralen som dök upp är precis integralen av täthetsfunktionen för en $t(\nu-2)$-fördelad variabel (om $\nu > 2$). Vi förenklar uttrycket och finner att

$$\begin{aligned}
\frac{c_\nu}{c_{\nu-2}} \frac{\nu^{3/2}}{(\nu-1)\sqrt{\nu-2}} &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)\Gamma\left(\frac{\nu-2}{2}\right)\sqrt{(\nu-2)\pi}}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{\nu-1}{2}\right)} \frac{\nu^{3/2}}{(\nu-1)\sqrt{\nu-2}} \\
&= \frac{\frac{\nu-1}{2}\Gamma\left(\frac{\nu-1}{2}\right)\Gamma\left(\frac{\nu-2}{2}\right)}{\frac{\nu-2}{2}\Gamma\left(\frac{\nu-2}{2}\right)\Gamma\left(\frac{\nu-1}{2}\right)} \frac{\nu}{\nu-1} = \frac{\nu}{\nu-2},
\end{aligned}$$

där vi nyttjat att $\Gamma(z+1) = z\Gamma(z)$. Eftersom $E(X) = 0$ följer det nu att $V(X) = E(X^2)$. □

12.1 *t*-fördelningens kvantiler

Kvantilerna för t-fördelningen är de tal $t_\alpha(n)$ sådana att $P(T > t_\alpha(n)) = 1 - \alpha$. Det vill säga gränser $t_\alpha(n)$ sådana att för $T \sim t(n)$ gäller att andelen α av sannolikhetsmassan ligger till höger om $t_\alpha(n)$. Eftersom gränserna är jobbiga att räkna fram för hand brukar vi använda tabellverk enligt nedan (studera även formelsamlingen).



n \ α	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
70	1.294	1.667	1.994	2.381	2.648	3.211	3.435
80	1.292	1.664	1.990	2.374	2.639	3.195	3.416
90	1.291	1.662	1.987	2.368	2.632	3.183	3.402
100	1.290	1.660	1.984	2.364	2.626	3.174	3.390
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

13 ($\star$)Vektorer med stokastiska variabler

Låt $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ vara en vektor vars komponenter är stokastiska variabler. Vi strävar efter att skriva vektorer som kolonnvektorer. Det faller sig naturligt att definera väntevärdet av $\boldsymbol{X}$ genom

$$E(\boldsymbol{X}) = (E(X_1), E(X_2), \ldots, E(X_n)).$$

På samma sätt definierar vi väntevärdet av en matris av stokastiska variabler. Variansen blir lite konstigare så vi introducerar den så kallade kovariansmatrisen mellan två vektorer (av samma dimension). Låt $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ och definiera $C(\boldsymbol{X}, \boldsymbol{Y})$ enligt

$$C(\boldsymbol{X}, \boldsymbol{Y}) = \begin{pmatrix} C(X_1, Y_1) & C(X_1, Y_2) & \cdots & C(X_1, Y_n) \\ C(X_2, Y_1) & C(X_2, Y_2) & \cdots & C(X_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(X_n, Y_1) & C(X_n, Y_2) & \cdots & C(X_n, Y_n) \end{pmatrix}$$

där $C(X_i, Y_j) = E(X_i Y_j) - E(X_i)E(Y_j)$ är kovariansen mellan X_i och Y_j.

En stor anledning att blanda in vektorer och matriser är givetvis att få tillgång till maskineriet från linjär algebra. Kovariansen (en matris) mellan två vektorer $\boldsymbol{X}$ och $\boldsymbol{Y}$ kan då lite mer kompakt skrivas

$$C(\boldsymbol{X}, \boldsymbol{Y}) = E(\boldsymbol{X}\boldsymbol{Y}^T) - E(\boldsymbol{X})E(\boldsymbol{Y})^T,$$

där $(\,\cdot\,)^T$ innebär transponering. En produkt $A = \boldsymbol{x}\boldsymbol{y}^T$ brukar kallas för den yttre produkten och består av element $(a)_{ij} = x_i y_j$, $i, j = 1, 2, \ldots, n$. Detta är alltså *inte* skalärprodukten $(\boldsymbol{X}^T\boldsymbol{Y})$.

Låt $A, B \in \mathbf{R}^{n\times n}$ vara matriser. Då är $A\boldsymbol{X}$ en linjärkombination av $X_1, X_2, \ldots, X_n$ och $B\boldsymbol{Y}$ en linjärkombination av $Y_1, Y_2, \ldots, Y_n$. Dessutom kan *alla* linjärkombinationer skrivas på detta sätt. Vidare gäller nu tack varje linjäriteten att

$$E(A\boldsymbol{X}) = AE(\boldsymbol{X}) \quad \text{och} \quad C(A\boldsymbol{X}, B\boldsymbol{Y}) = A\boldsymbol{X}(B\boldsymbol{Y})^T = A\boldsymbol{X}\boldsymbol{Y}^T B^T.$$

14 ($\star$)Cochrans sats

Sats. Låt $X_1, X_2, \ldots, X_n$ vara oberoende likafördelade stokastiska variabler där $X_k \sim N(\mu, \sigma)$ för $k = 1, 2, \ldots, n$. Då gäller att

$$\frac{1}{\sigma^2}\sum_{k=1}^{n}(X - \overline{X})^2 \sim \chi^2(n-1).$$

Bevis. Låt $Y_k = X_k - \mu$ så att $Y_k \sim N(0, \sigma)$. Vi ser att

$$\sum_{k=1}^{n}(X_k - \overline{X})^2 = \sum_{k=1}^{n}(Y_k - \overline{Y})^2.$$

Låt J vara $n \times n$-matrisen vars samtliga element är 1 och låt $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$. Då kan vi skriva

$$\begin{pmatrix} Y_1 - \overline{Y} \\ Y_2 - \overline{Y} \\ \vdots \\ Y_n - \overline{Y} \end{pmatrix} = I\boldsymbol{Y} - \frac{1}{n}J\boldsymbol{Y} = Q\boldsymbol{Y},$$

där $Q = I - (1/n)J$. Låt $P = I - Q = (1/n)J$. Då är

$$P + Q = I, \qquad P^2 = P^T = P, \qquad Q^2 = Q^T = Q \qquad \text{samt } PQ = QP = 0.$$

Matriserna P och Q representerar alltså ortogonala projektioner på $\mathbf{R}^n$ och av naturliga skäl är $\operatorname{rank}(P) = 1$ så $\operatorname{rank}(Q) = n - 1$ (eftersom $P + Q = I$).

Vidare gäller att

$$C(P\boldsymbol{Y}, Q\boldsymbol{Y}) = PC(\boldsymbol{Y}, \boldsymbol{Y})Q^T = P\sigma^2 I Q^T = PQ^T = PQ = 0$$

då $E(\boldsymbol{Y}) = 0$ och $C(Y_i, Y_j) = \sigma^2$ om $i = j$ och $C(Y_i, Y_j) = 0$ då $i \neq j$ eftersom olika Y_k är oberoende. Således är $Y_i - \overline{Y}$ och $\overline{Y}$ oberoende stokastiska variabler (eftersom kovariansen noll mellan normalfördelade variabler är ekvivalent med oberoende). Eftersom $\operatorname{rank}(Q) = n - 1$ så kan vi representera $Q\boldsymbol{Y}$ i en ortogonal bas så att

$$\frac{1}{\sigma^2}\sum_{k=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{\sigma^2}(Q\boldsymbol{Y})^T Q\boldsymbol{Y} = \frac{1}{\sigma^2}\boldsymbol{Y}^T Q\boldsymbol{Y} = Z_1^2 + Z_2^2 + \cdots + Z_{n-1}^2,$$

där $Z_k \sim N(0,1)$ och dessa variabler är oberoende. Vi kan nu nyttja den tidigare satsen om att summan av n stycken kvadrater av $N(0,1)$-fördelade variabler är $\chi^2(n)$-fördelad för att dra slutsatsen att $\dfrac{1}{\sigma^2}\boldsymbol{X}^T Q\boldsymbol{X} \sim \chi^2(n-1)$. □

15 $(\star\star)$ t- och χ^2-fördelning; Gossets sats

Det finns givetvis en anledning till att vi studerar just dessa två fördelningar. William Gosset bevisade nämligen följande sats.

Sats. Låt $Z \sim N(0,1)$ och $V \sim \chi^2(\nu)$ vara oberoende. Då är $\dfrac{Z}{\sqrt{V/\nu}} \sim t(\nu)$.

Bevis. Eftersom Z och V är oberoende ges den simultana täthetsfunktionen av

$$f(z, v) = \frac{1}{\sqrt{2\pi}\, 2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} e^{-z^2/2} v^{\nu/2-1} e^{-v/2}, \quad z \in \mathbf{R},\ v \geq 0.$$

Låt $T = \dfrac{Z}{\sqrt{V/\nu}}$ och $c = \dfrac{1}{\sqrt{2\pi}\, 2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)}$. Vi söker täthetsfunktionen f_T för T. Betrakta

$$P(T \leq t) = \iint_{z/\sqrt{v/\nu} \leq t} f(z, v)\, dz\, dv \quad = c \iint_{z/\sqrt{v/\nu} \leq t} e^{-z^2/2} v^{\nu/2-1} e^{-v/2}\, dz\, dv.$$

Vi gör ett variabelbyte,

$$\begin{cases} u\sqrt{\dfrac{v}{\nu}}, \\ w = v \end{cases} \qquad \Rightarrow \qquad \frac{d(z,v)}{d(u,w)} = \begin{vmatrix} \sqrt{\frac{w}{\nu}} & \frac{u}{2\nu\sqrt{\frac{w}{\nu}}} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{w}{\nu}},$$

så integralen blir

$$\begin{aligned} c\iint_{u\le t} \sqrt{\frac{w}{\nu}} e^{-u^2 w/(2\nu)} w^{\nu/2-1} e^{-v/2}\, dz\, dv &= \frac{c}{\sqrt{\nu}} \int_{-\infty}^{t}\int_{0}^{\infty} w^{\frac{\nu+1}{2}-1} \exp\left(-\frac{w}{2}\left(1+\frac{u^2}{\nu}\right)\right) dw\, du \\ &= \frac{c}{\sqrt{\nu}} \int_{-\infty}^{t}\int_{0}^{\infty} \frac{r^{\frac{\nu+1}{2}-1}}{\left(1+\frac{u^2}{\nu}\right)^{\frac{\nu+1}{2}}} e^{-\frac{r}{2}} dr\, du, \\ &= \frac{c}{\sqrt{\nu}} \int_{-\infty}^{t} \left(1+\frac{u^2}{\nu}\right)^{-\frac{\nu+1}{2}} \int_{0}^{\infty} r^{\frac{\nu+1}{2}-1} e^{-\frac{r}{2}} dr\, du, \end{aligned}$$

där vi gjorde ett variabelbyte $r = w\left(1+\dfrac{u^2}{\nu}\right)$ i den innersta integralen och bröt ut den faktor som inte beror på u. Den innersta integralen är nu nästan (upp till normeringskonstanten) integralen av täthetsfunktionen för en $\chi^2(\nu+1)$-variabel, så

$$\int_0^\infty r^{\frac{\nu+3}{2}-1} e^{-\frac{r}{2}} dr = 2^{(\nu+1)/2}\Gamma\left(\frac{\nu+1}{2}\right).$$

Således ges fördelningsfunktionen

$$F_T(t) = \frac{2^{(\nu+1)/2}\,\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu}\sqrt{2\pi}\,2^{\nu/2}\,\Gamma\left(\frac{\nu}{2}\right)} \int_{-\infty}^{t} \left(1+\frac{u^2}{\nu}\right)^{-\frac{\nu+1}{2}} du = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \int_{-\infty}^{t} \left(1+\frac{u^2}{\nu}\right)^{-\frac{\nu+1}{2}} du$$

vilket efter derivering ger täthetsfunktionen

$$f_T(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

vilket är precis täthetsfunktionen för en $t(\nu)$-fördelad variabel. □

Sats. Om $X_1, X_2, \ldots, X_n$ är oberoende och likafördelade med fördelningen $N(\mu,\sigma)$ så är kvoten $T = \dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$, där S^2 är stickprovsvariansen.

Bevis. Detta följer direkt från föregående resultat och Cochrans sats. Vi kan formulera T enligt

$$T = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \cdot \frac{1}{\frac{1}{\sigma}S} = \frac{Z}{\sqrt{\frac{V}{n-1}}},$$

där $Z \sim N(0,1)$ och $V = \dfrac{1}{\sigma^2}\displaystyle\sum_{k=1}^{n}(X_k-\overline{X})^2 \sim \chi^2(n-1)$ (med $S^2 = \dfrac{1}{n-1}V$). □

16 ($\star$)Bonus: Gammafunktionen

För $z \in \mathbf{C}$ med $\operatorname{Re} z > 0$ är integralen

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$$

absolutkonvergent. Detta är ett sätt att definiera gammafunktionen på. Funktionen ovan går även att analytiskt utvidga till $\operatorname{Re} z \leq 0$ förutom för negativa heltal. Ifrån definitionen ovan kan vi medelst partialintegration erhålla att

$$\Gamma(z+1) = \int_0^\infty x^z e^{-x}\, dx = \left[-x^z e^{-x}\right]_0^\infty + z\int_0^\infty x^{z-1} e^{-x}\, dx = z\Gamma(z).$$

Eftersom $\Gamma(1) = 1$ visar denna likhet att

$$\Gamma(n) = (n-1)!$$

för alla positiva heltal. Gamma-funktionen utvidgar således fakultetet till alla komplexa z förutom negativa heltal. Kopplingen till normaliseringen av χ^2-fördelningen är ganska naturlig. Vi ser att om $X \sim \chi^2(k)$ så är

$$\begin{aligned}
\int_0^\infty f_X(x)\, dx &= \frac{1}{2^{k/2}\Gamma(k/2)} \int_0^\infty x^{k/2-1} e^{-x/2}\, dx \\
&= \Bigg/ \text{variabelbyte: } \begin{matrix} u = x/2 \\ dx = 2du \end{matrix} \Bigg/ = \frac{1}{2^{k/2}\Gamma(k/2)} \int_0^\infty 2^{k/2} u^{k/2-1} e^{-u}\, du \\
&= \frac{1}{\Gamma(k/2)}\Gamma(k/2) = 1.
\end{aligned}$$

Även identiteten $\Gamma(z+1) = z\,\Gamma(z)$ är det vi använde när vi beräknade $E(X)$ och $V(X)$ (gå tillbaka och studera partialintegrationen!).

För att identifiera $\Gamma(n+1/2)$ kan vi till exempel göra variabelbytet $u = \sqrt{x}$ och partialintegrera n gånger:

$$\begin{aligned}
\Gamma(n+1/2) &= \int_0^\infty x^{n+1/2} e^{-x}\, dx = 2\int_0^\infty u^{2n} e^{-u^2}\, du = 2\int_0^\infty u^{2n-1} \cdot (ue^{-u^2})\, du \\
&= -\left[u^{2n-1} e^{-u^2}\right]_0^\infty + (2n-1)\int_0^\infty u^{2n-3} \cdot (ue^{-u^2})\, du \\
&= (2n-1)\left(-\frac{1}{2}\left[u^{2n-3} e^{-u^2}\right]_0^\infty + \frac{2n-3}{2}\int_0^\infty u^{2n-5} \cdot (ue^{-u^2})\, du\right) \\
&= \frac{(2n-1)(2n-3)}{2}\left(-\frac{1}{2}\left[u^{2n-5} e^{-u^2}\right]_0^\infty + \frac{2n-5}{2}\int_0^\infty u^{2n-7} \cdot (ue^{-u^2})\, du\right) \\
&= \cdots = \frac{(2n-1)(2n-3)\cdots(2n-(2n-1))}{2^{n-1}}\int_0^\infty u^{2n-2n} e^{-u^2}\, du \\
&= \frac{(2n-1)(2n-3)\cdots(2n-(2n-1))}{2^{n-1}}\,\frac{\sqrt{\pi}}{2} = \frac{(2n)!}{4^n\, n!}\sqrt{\pi},
\end{aligned}$$

där vi i sista steget använde den välkända identiteten $\displaystyle\int_0^\infty e^{-u^2}\, du = \frac{\sqrt{\pi}}{2}$.

Det finns mer eleganta sätt att ta fram identiteten för $\Gamma(n+1/2)$ genom exempelvis Eulers reflektionsformel:

$$\Gamma(1-z)\Gamma(z) = \frac{\pi}{\sin(\pi z)}, \quad z \notin \mathbf{Z},$$

men den likheten är lite mer komplicerad att bevisa, så vi nöjer oss med ovanstående.