Föreläsning 3: Konfidensintervall

Johan Thim (johan.thim@liu.se)

5 september 2018

"[we are] Explorers in the further regions of experience. Demons to some. Angels to others."
–Pinhead

1 Intervallskattningar

Vi har nu studerat hur man mer eller mindre systematiskt kan hitta skattningar för okända parametrerar när vi har stickprov från en fördelning som beror på parametern. Den naturliga följdfrågan är givetvis hur "bra" skattningen är. Vi har vissa mått i form av väntevärdesriktighet, konsistens och effektivitet, men går det att säga något med en given sannolikhet? Kan vi hitta ett intervall som med en viss given sannolikhet måste innehålla den okända parametern?

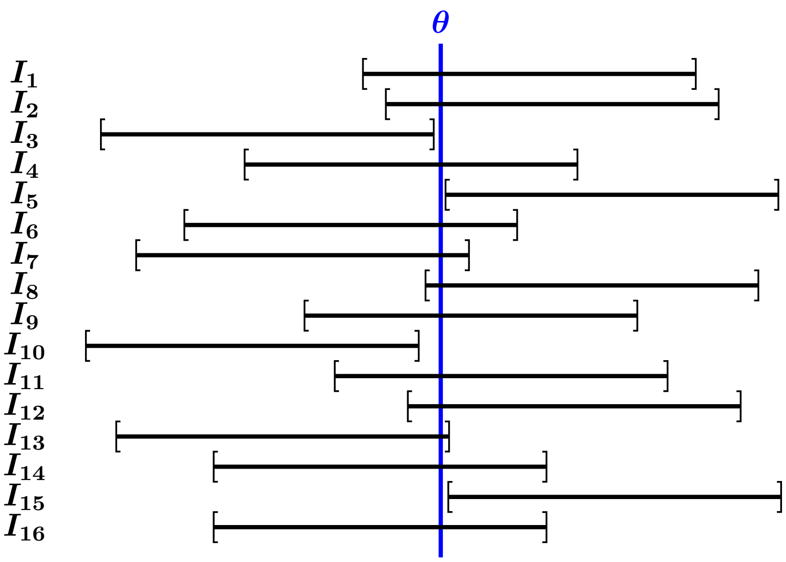
Konfidensintervall

Definition. Låt $x_1, \ldots, x_n$ vara ett stickprov av en fördelning som beror på en okänd parameter θ och låt $\alpha \in [0, 1]$. Ett intervall $I_\theta^{1-\alpha} = (\widehat{\theta}_L, \widehat{\theta}_U)$ kallas för ett **konfidensintervall** för θ med **konfidensgrad** $1 - \alpha$ om

$$P(\widehat{\Theta}_L < \theta < \widehat{\Theta}_U) = 1 - \alpha.$$

Gränserna $\widehat{\theta}_L = a(x_1, \ldots, x_n)$ och $\widehat{\theta}_U = b(x_1, \ldots, x_n)$ är skattningar som beräknas från stickprovet. Dessa ändpunkter kallas **konfidensgränser**.

Så hur fungerar detta i praktiken? Säg att vi har tillgång till 100 olika stickprov $\boldsymbol{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})$ från en och samma fördelning som beror på samma okända parameter θ. Vi hittar konfidensintervall för alla 100 stickproven med konfidensgrad $1-\alpha$. Då kommer $100 \cdot (1-\alpha)$ av dessa intervall att innehålla θ (i snitt).



Av 16 intervall är det 4 som inte innehåller det verkliga värdet på θ. Så med andra ord verkar det som att ungefär $12/16 = 3/4$ av intervallen innehåller det verkliga värdet på θ. Detta innebär att konfidensgraden vid skattningen är ungefär 75%.

Notera att det är gränserna i konfidensintervallet som är stokastiska variabler (eller skattningar därav). Storheten θ är okänd (och behöver inte ens ligga i intervallet).

Inga intervall är mer värda

Vi kan inte säga att till exempel I_9 är ett ”bättre” intervall än I_{13}, utan det är en binär fråga: gäller det att $\theta \in I_k$ eller inte.

Så då kommer vi till nästa rimliga fråga: hur hittar vi systematiskt konfidensintervall med given konfidensgrad?

Konstruktion av konfidensintervall

1. Ställ upp en lämplig skattningsvariabel $\widehat{\Theta}$ för θ. Här kan vi använda de metoder vi tagit fram tidigare (moment-, MK- och ML-skattningar till exempel).

2. Konstruera en hjälpvariabel H (teststorhet) utifrån $\widehat{\Theta}$. Hjälpvariabeln får endast innehålla kända storheter utöver θ (och om θ förekommer flera gånger kan vi behöva skatta bort en del instanser för att få något användbart).

3. Stäng in hjälpvariabeln i ett intervall $I = (c, d)$ så att $P(c < H < d) = 1 - \alpha$.

4. Lös ut θ ur olikheten $c < H < d$:
$$c < H < d \quad \Leftrightarrow \quad a(X_1, \ldots, X_n) < \theta < b(X_1, \ldots, X_n)$$
vilket ger att $P(a(X_1, \ldots, X_n) < \theta < b(X_1, \ldots, X_n)) = 1 - \alpha$.

5. Ersätt de stokastiska storheterna $X_1, \ldots, X_n$ med observationerna $x_1, \ldots, x_n$ vilket ger intervallet
$$I_\theta^{1-\alpha} = (a(x_1, \ldots, x_n), b(x_1, \ldots, x_n)).$$

Något som kommer bli viktigt är följande definition från sannolikhetsteorin.

Kvantil

Definition. En α-kvantil λ_α för en stokastisk variabel X är ett tal λ_α sådant att
$$P(X > \lambda_\alpha) = \alpha.$$

Vi finner ofta kvantiler i tabell endera genom en explicit kvantiltabell eller genom att söka upp sannolikheten $1 - \alpha$ och identifiera (approximativt) vilket värde på x som gör att vi erhåler $F(x) = 1 - \alpha$, där F är fördelningsfunktionen. Saknar vi tabell får vi istället lösa ekvationen
$$1 - \alpha = \int_{-\infty}^{\lambda_\alpha} f_X(x)\, dx.$$

Observera att svaret inte nödvändigtvis är entydigt.

2 χ^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.

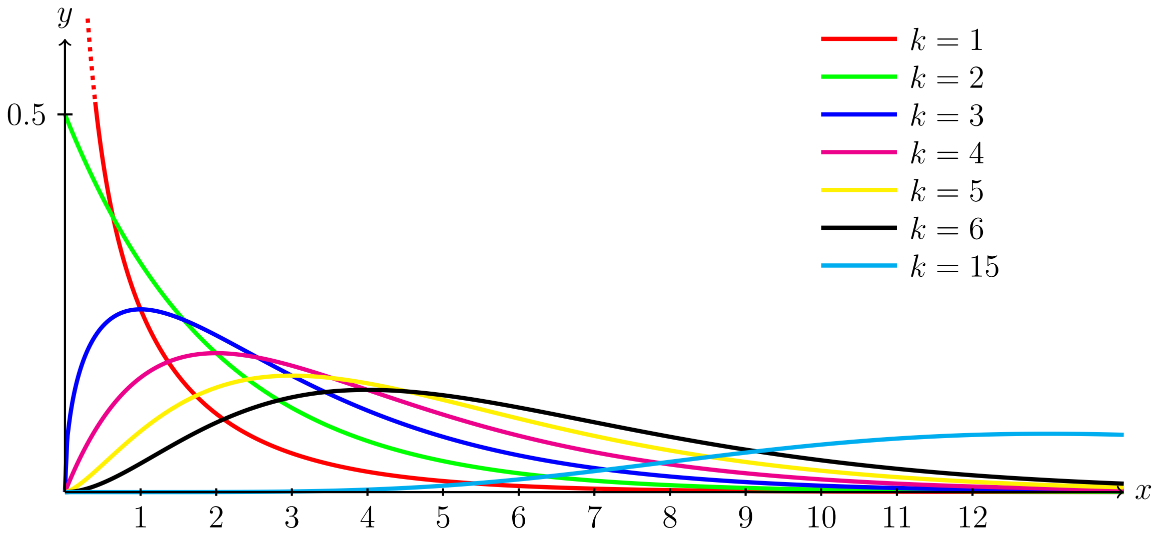
χ^2-fördelning

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 $V(X) = E(X^2) - E(X)^2 = k^2 + 2k - k^2 = 2k$. $\square$

[1] Se avsnitt 10 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 $f_{X+Y} \sim \chi^2(\nu_1 + \nu_2)$. $\square$

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:

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

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?). $\square$

3 t-fördelningen

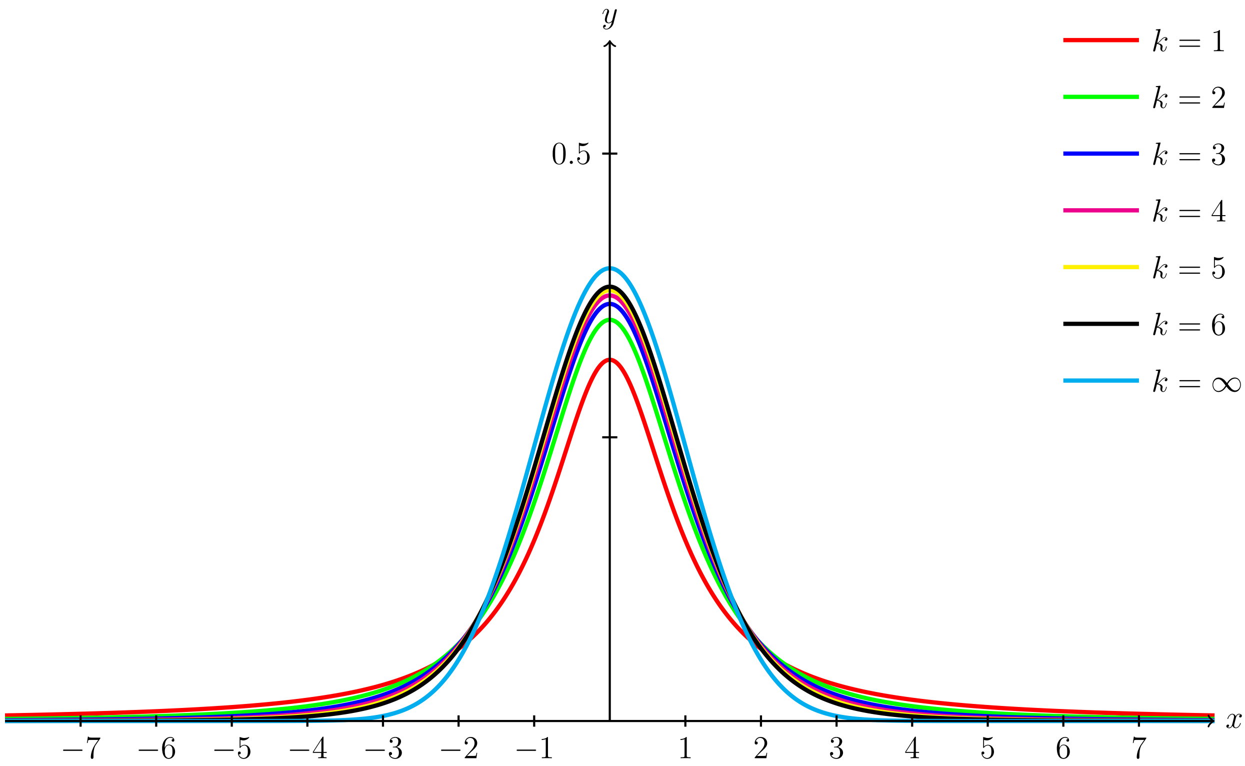
t-fördelning

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

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

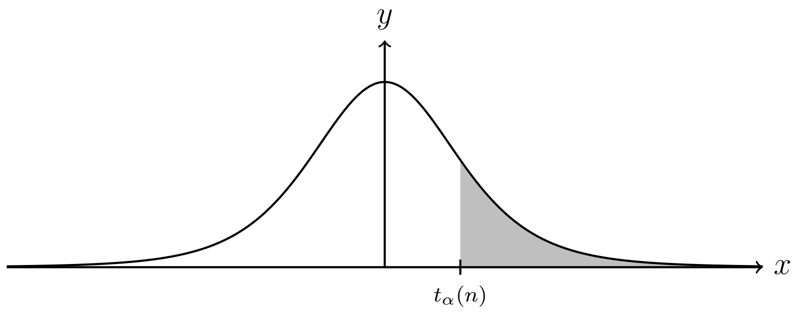
där vi bytte variabel så $x\sqrt{\nu-2} = u\sqrt{\nu}$ och utnyttjade att integralen som dök upp är precis integralen av täthetsfunktionen för en $t(\nu-2)$-fördelad variabel (om $\nu > 2$). Vi förenklar uttrycket och finner att

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

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

3.1 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

4 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 definerar 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

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

5 Cochrans sats

Vi ska nu betrakta ett speciallfall av en ganska generell sats (Cochrans sats) ..

Sats. Låt $X_1, X_2, \ldots, X_n$ vara oberoende likafördelade stokastiska variabler där $X_k \sim N(\mu, \sigma^2)$ 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 $Y_k \sim N(0, \sigma^2)$. 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

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

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 $\dfrac{1}{\sigma^2}\boldsymbol{X}^T Q\boldsymbol{X} \sim \chi^2(n - 1)$. $\square$

6 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

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

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. □

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

7 Konfidensintervall för μ och σ i normalfördelning

7.1 Konfidensintervall för μ när σ är känd

Låt $x_1, x_2, \ldots, x_n$ vara ett stickprov från en $N(\mu, \sigma^2)$-fördelning där vi känner σ och vill hitta ett konfidensintervall för μ. En punktskattning för väntevärdet ges av

$$
\widehat{M} = \bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \sim N(\mu, \sigma^2/n).
$$

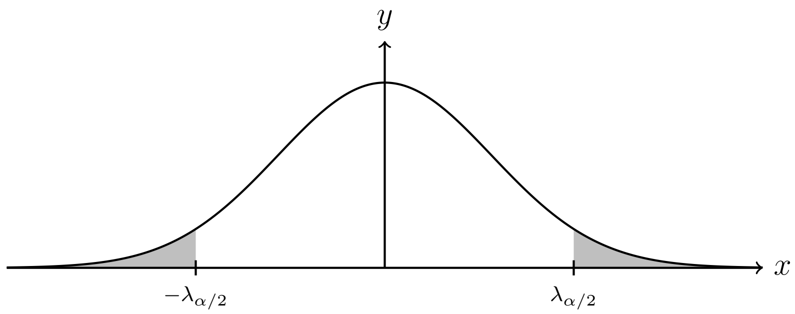
Vi skapar testvariabeln

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

Det följer då att vi kan välja ett tal $\lambda_{\alpha/2}$ så att

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

Talet $\lambda_{\alpha/2}$ är $\alpha/2$-kvantilen för en $N(0,1)$-fördelning och ges av $\lambda_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$. Eftersom vi saknar explicit uttryck för denna invers är det enklast (utan dator åtminstone) att slå i tabell. Standardtabell som finns i formelsamlingen enligt nedan (vi får uttnyttja symmetri för att finna sannolikheter mindre än 0.5).



Det skuggade områdena är sannolikheten att $P(Z < -\lambda_{\alpha/2}) + P(Z > \lambda_{\alpha/2})$.

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

$$\begin{aligned} -\lambda_{\alpha/2} < Z < \lambda_{\alpha/2} \quad &\Leftrightarrow \quad -\lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \widehat{M} - \mu < \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ &\Leftrightarrow \quad \widehat{M} - \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \widehat{M} + \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} \end{aligned}$$

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

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

med konfidensgrad $1 - \alpha$.

Exempel

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

Antag att mätningarna är ett stickprov på en normalfördelad variabel $X \sim N(\mu, 2^2)$ (man tycker sig veta så pass mycket om processen att standardavvikelsen anses vara känd). Beräkna ett 99% konfidensintervall för väntevärdet μ.

Lösning. Vi betraktar siffrorna som ett stickprov från oberoende s. v. $X_j \sim N(\mu, \sigma^2 = 4)$. Vi punktskattar väntevärdet μ med

$$\widehat{M} = \bar{X} = \frac{1}{10}\sum_{j=1}^{10} X_j \sim N(\mu, 4/10).$$

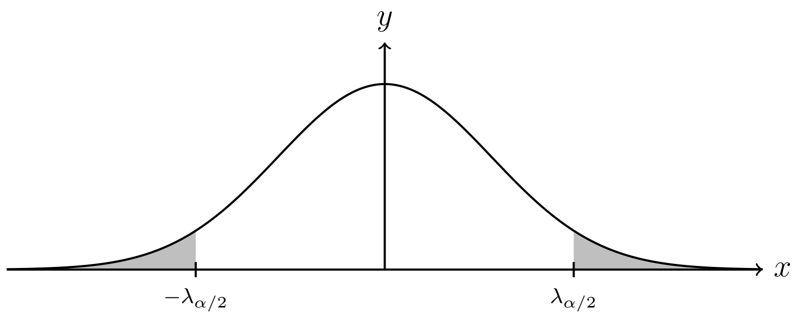
Vi skapar testvariabeln

$$Z = \frac{\widehat{M} - \mu}{\sigma/\sqrt{10}} \sim N(0,1).$$

Det följer då att

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

och då vi söker ett 99% konfidensintervall så är $\alpha = 0.01$ och $\lambda_{0.005} \approx 2.575$ (det sista ur tabell).

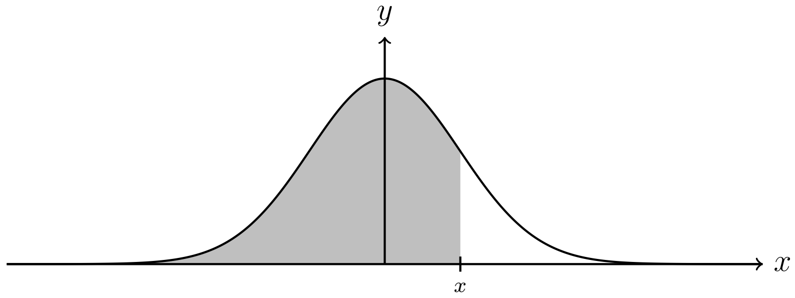


Det skuggade området är $\alpha \cdot 100\%$ av sannolikhetsmassan jämt fördelad på svansarna.
Vi löser ut μ ur olikheten i sannolikhetsmåttet i ekvation (2) ovan och erhåller att

$$\widehat{M} - \frac{2.575 \cdot 2}{\sqrt{10}} < \mu < \widehat{M} + \frac{2.575 \cdot 2}{\sqrt{10}}.$$

Om vi ersätter $\widehat{M}$ med den observerade punktskattningen $\widehat{\mu} = 5.074$ (medelvärdet av observationerna) så får vi ett konfidensintervall $I_\mu = (3.45, 6.70)$ med konfidensgrad 99%.

7.1.1 Normalfördelningstabell



Det skuggade området är sannolikheten att $P(Z \leq x)$, där $Z \sim N(0,1)$.

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

7.2 Okänd varians

Låt $x_1, x_2, \ldots, x_n$ vara ett stickprov från en $N(\mu, \sigma^2)$-fördelning där vi *inte* vet vad σ är och vi vill hitta ett konfidensintervall för μ. En punktskattning för väntevärdet ges av

$$\widehat{M} = \bar{X} = \frac{1}{n} \sum_{k=1}^{n} X_k \sim N(\mu, \sigma^2/n).$$

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

$$P(-t_{\alpha/2}(n-1) < T < t_{\alpha/2}(n-1)) = 1 - \alpha. \tag{3}$$

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 (3) 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 $X \sim N(\mu, \sigma^2)$ och σ är okänd.

Lösning. Vi betraktar siffrorna som ett stickprov från oberoende s. v. $X_j \sim N(\mu, \sigma^2)$. Vi punktskattar med $\widehat{M} = \bar{X} \sim N(\mu, \sigma^2/10)$ som tidigare och skattar σ med s, där

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

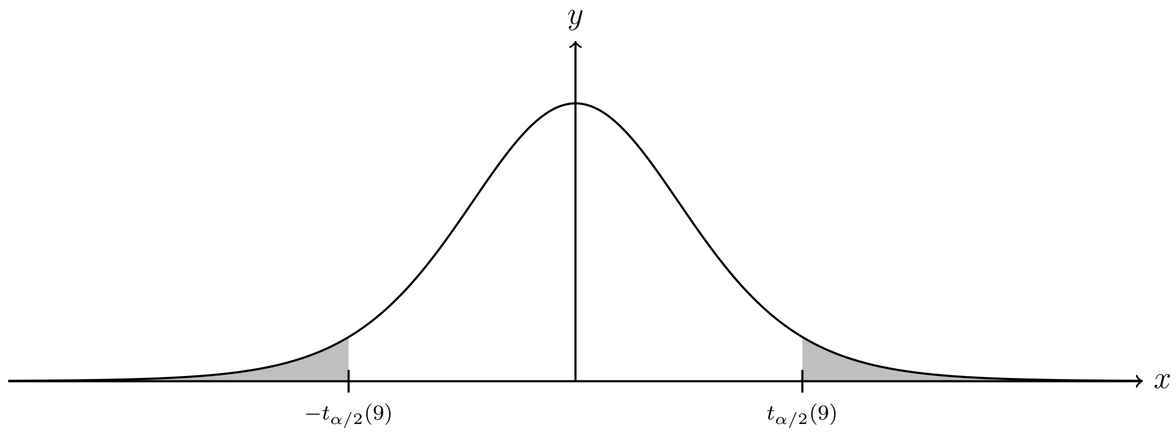
är stickprovsvariansen. Vi skapar testvariabeln

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

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%.

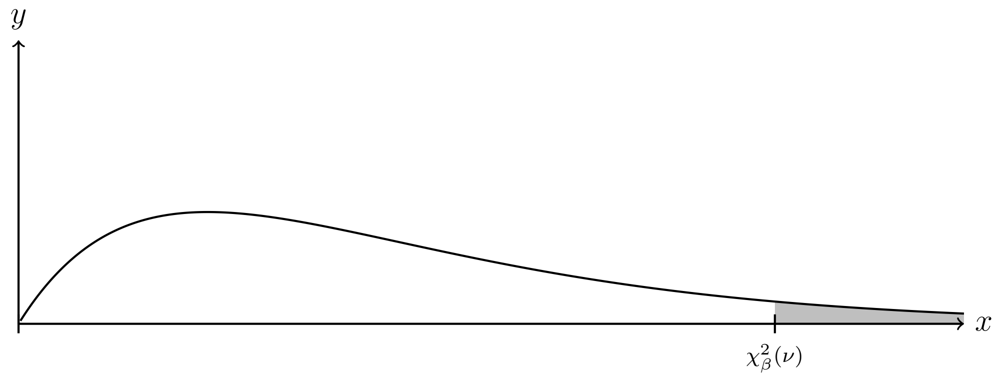
7.3 Konfidensintervall för varians

Kan man avgöra om en gissning på variansen är rimlig? Vi behöver en testvariabel där vi känner fördelningen. Enligt Cochrans sats är $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, så vi kan stänga in denna variabel. Således är

$$\begin{aligned}
\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}(n-1) \quad &\Leftrightarrow \quad \frac{1}{\chi^2_{1-\alpha/2}(n-1)} > \frac{\sigma^2}{(n-1)S^2} > \frac{1}{\chi^2_{\alpha/2}(n-1)} \\
&\Leftrightarrow \quad \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)},
\end{aligned}$$

där $\chi^2_\beta(\nu)$ är β-kvantilen för $\chi^2(\nu)$-fördelningen:

$$P(V > \chi^2_\beta(\nu)) = \beta, \quad \text{då } V \sim \chi^2(\nu).$$



k \ α	0.0500	0.0250	0.0100	0.0010	0.9500	0.9750	0.9900	0.9990
1	0.0039	0.0010	0.0002	0.0000	3.8415	5.0239	6.6349	10.8276
2	0.1026	0.0506	0.0201	0.0020	5.9915	7.3778	9.2103	13.8155
3	0.3518	0.2158	0.1148	0.0243	7.8147	9.3484	11.3449	16.2662
4	0.7107	0.4844	0.2971	0.0908	9.4877	11.1433	13.2767	18.4668
5	1.1455	0.8312	0.5543	0.2102	11.0705	12.8325	15.0863	20.5150
6	1.6354	1.2373	0.8721	0.3811	12.5916	14.4494	16.8119	22.4577
7	2.1673	1.6899	1.2390	0.5985	14.0671	16.0128	18.4753	24.3219
8	2.7326	2.1797	1.6465	0.8571	15.5073	17.5345	20.0902	26.1245
9	3.3251	2.7004	2.0879	1.1519	16.9190	19.0228	21.6660	27.8772
10	3.9403	3.2470	2.5582	1.4787	18.3070	20.4832	23.2093	29.5883
11	4.5748	3.8157	3.0535	1.8339	19.6751	21.9200	24.7250	31.2641
12	5.2260	4.4038	3.5706	2.2142	21.0261	23.3367	26.2170	32.9095
13	5.8919	5.0088	4.1069	2.6172	22.3620	24.7356	27.6882	34.5282
14	6.5706	5.6287	4.6604	3.0407	23.6848	26.1189	29.1412	36.1233
15	7.2609	6.2621	5.2293	3.4827	24.9958	27.4884	30.5779	37.6973
16	7.9616	6.9077	5.8122	3.9416	26.2962	28.8454	31.9999	39.2524
17	8.6718	7.5642	6.4078	4.4161	27.5871	30.1910	33.4087	40.7902
18	9.3905	8.2307	7.0149	4.9048	28.8693	31.5264	34.8053	42.3124
19	10.1170	8.9065	7.6327	5.4068	30.1435	32.8523	36.1909	43.8202
20	10.8508	9.5908	8.2604	5.9210	31.4104	34.1696	37.5662	45.3147
21	11.5913	10.2829	8.8972	6.4467	32.6706	35.4789	38.9322	46.7970
22	12.3380	10.9823	9.5425	6.9830	33.9244	36.7807	40.2894	48.2679
23	13.0905	11.6886	10.1957	7.5292	35.1725	38.0756	41.6384	49.7282
24	13.8484	12.4012	10.8564	8.0849	36.4150	39.3641	42.9798	51.1786
25	14.6114	13.1197	11.5240	8.6493	37.6525	40.6465	44.3141	52.6197
26	15.3792	13.8439	12.1981	9.2221	38.8851	41.9232	45.6417	54.0520
27	16.1514	14.5734	12.8785	9.8028	40.1133	43.1945	46.9629	55.4760
28	16.9279	15.3079	13.5647	10.3909	41.3371	44.4608	48.2782	56.8923
29	17.7084	16.0471	14.2565	10.9861	42.5570	45.7223	49.5879	58.3012
30	18.4927	16.7908	14.9535	11.5880	43.7730	46.9792	50.8922	59.7031
40	26.5093	24.4330	22.1643	17.9164	55.7585	59.3417	63.6907	73.4020
50	34.7643	32.3574	29.7067	24.6739	67.5048	71.4202	76.1539	86.6608
60	43.1880	40.4817	37.4849	31.7383	79.0819	83.2977	88.3794	99.6072
70	51.7393	48.7576	45.4417	39.0364	90.5312	95.0232	100.4252	112.3169
80	60.3915	57.1532	53.5401	46.5199	101.8795	106.6286	112.3288	124.8392
90	69.1260	65.6466	61.7541	54.1552	113.1453	118.1359	124.1163	137.2084
100	77.9295	74.2219	70.0649	61.9179	124.3421	129.5612	135.8067	149.4493

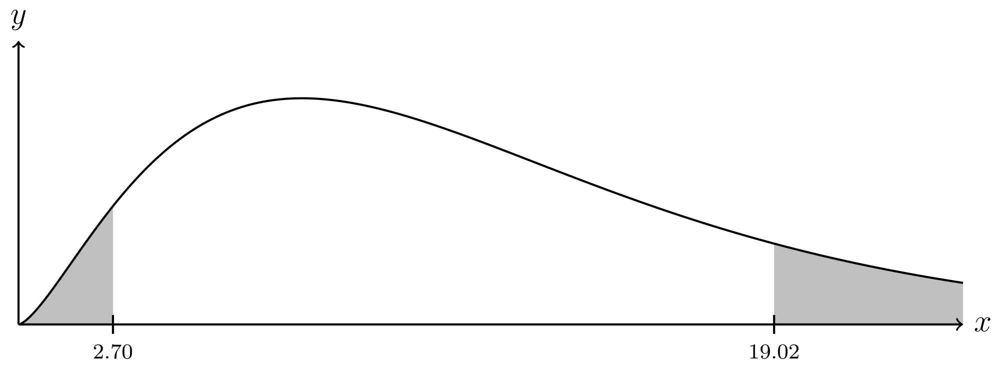
Exempel

Samma exempel (igen!) 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

Ställ upp ett 95%-igt konfidensintervall för variansen. Var antagandet att $\sigma = 2$ rimligt?

Lösning. Stickprovsstorleken är $n = 10$ och vi låter $V = \dfrac{9\,S^2}{\sigma^2}$. Under antagande att det är ett stickprov från en $N(\mu, \sigma^2)$-fördelning kommer $V \sim \chi^2(9)$. Ur tabell hittar vi gränser a och b så att $P(a < V < b) = 0.95$ genom att välja $a = \chi^2_{0.975}(9) = 2.70$ och $b = \chi^2_{0.025}(9) = 19.02$.



Vi löser nu ut σ^2 ur olikheten:

$$a < \frac{9\,S^2}{\sigma^2} < b \quad \Leftrightarrow \quad \frac{9\,S^2}{b} < \sigma^2 < \frac{9\,S^2}{a}$$

och skattar S^2 med stickprovsvariansen $s^2 = 2.08$ och erhåller då intervallet

$$I_{\sigma^2} = (0.98,\ 6.93)\,.$$

Vi kan utifrån detta även skatta ett konfidensintervall för standardavvikelsen enligt

$$I_{\sigma} = (0.99,\ 2.63)\,.$$

Eftersom $2 \in I_\sigma$ kan vi inte säga att $\sigma = 2$ är orimligt.

8 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. 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; se nästa avsnitt). 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 par exempel.

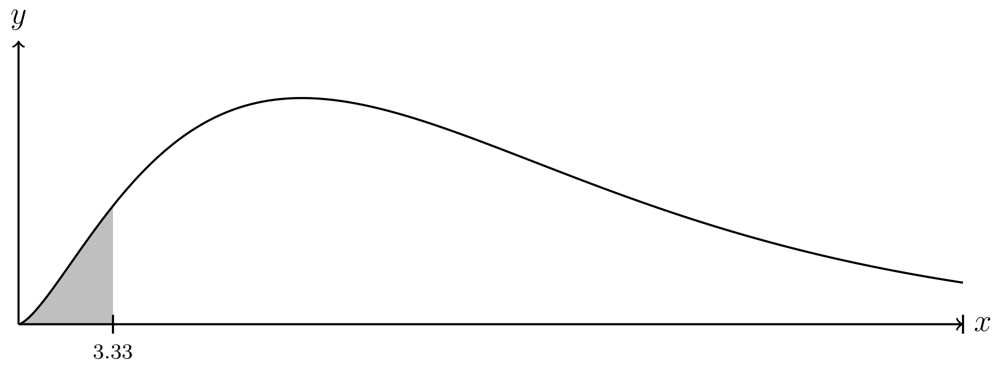
Exempel

Samma exempel (igen igen!) som tidigare där man vid en mätning av en process fick man följande mätdata:

$$6.04 \quad 4.96 \quad 4.93 \quad 3.40 \quad 7.04 \quad 4.73 \quad 3.57 \quad 7.70 \quad 4.55 \quad 3.82$$

Ställ upp ett 95%-igt konfidensintervall för variansen där vi endast är intresserade av hur stor variansen är. Skulle antagandet att $\sigma = 2.5$ vara rimligt? Jämför med föregående exempel.

Lösning. Stickprovsstorleken är $n = 10$ och vi låter även nu $V = \dfrac{9\,S^2}{\sigma^2} \sim \chi^2(9)$. Ur tabell hittar vi en gräns c så att $P(c < V) = 0.95$ genom att välja $c = \chi^2_{0.95}(9) = 3.33$.



Vi löser nu ut σ^2 ur olikheten:

$$c < \frac{9\,S^2}{\sigma^2} \quad \Leftrightarrow \quad \sigma^2 < \frac{9\,S^2}{c}$$

och skattar S^2 med stickprovsvariansen $s^2 = 2.08$ och erhåller då intervallet

$$I_{\sigma^2} = (0,\, 5.62)\,.$$

Vi kan utifrån detta även skatta ett konfidensintervall för standardavvikelsen enligt

$$I_{\sigma} = (0,\, 2.37)\,.$$

Eftersom $2 \in I_\sigma$ kan vi inte säga att $\sigma = 2$ är orimligt. Däremot kan vi säga att $\sigma = 2.5$ är orimligt (vilket vi *inte* kunde göra i föregående exempel!).

9 Prediktionsintervall

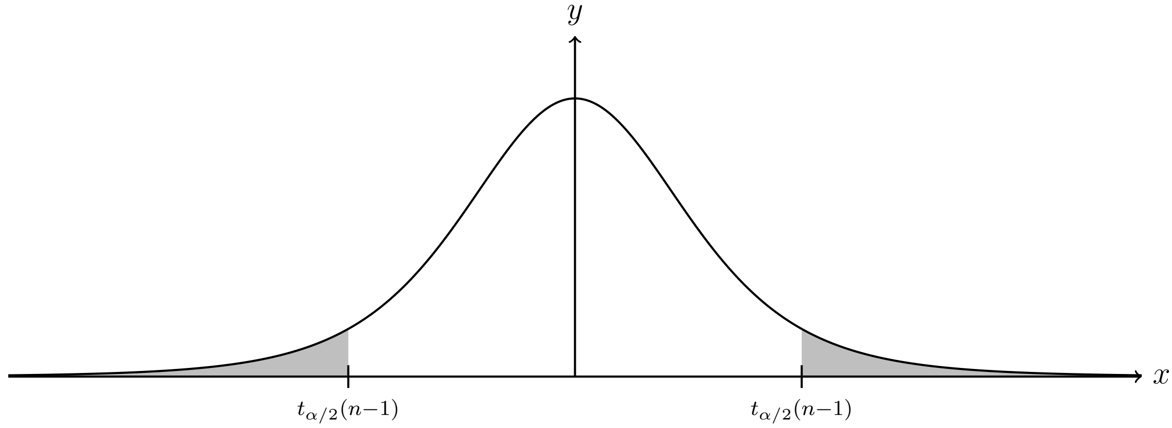
Vi har hittat konfidensintervall för både väntevärde och varians (och därigenom skattat intervall för standardavvikelsen), men kan man säga något om vart en enskild observation hamnar? Det är ju en stokastisk variabel, så det måste gå. Det vanliga är följande manöver.
Låt $X_1, X_2, \ldots, X_n$ vara ett slumpmässigt stickprov från $N(\mu, \sigma^2)$. Vi vill stänga in en enskild observation av en variabel X_0 (som antags 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

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

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. Om vi jämför intervallen för väntevärde respektive predikterat värde ser vi att vi alltid har $I_{X_o} \subset I_\mu$ med den metod vi använt ovan (och aldrig likhet).

10 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 Re $z \leq 0$ förutom för negativa heltal. Ifrån definitionen ovan kan vi medelst partialintegration erhålla att

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

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 identifera $\Gamma(n+1/2)$ 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.