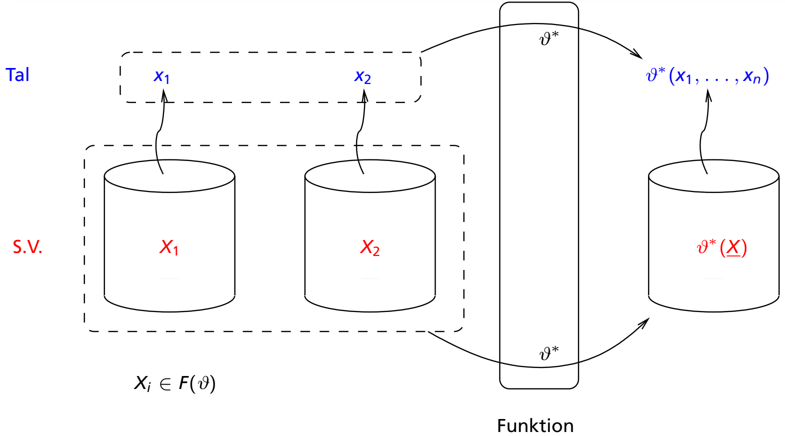
Föreläsning 9, Matematisk statistik 7.5 hp för E
Konfidensintervall

Stas Volkov

Stickprov & Skattning

Ett **stickprov**, $x_1, x_2, \ldots, x_n$, är **observationer** av s.v. $X_1, \ldots, X_n$ från någon fördelning $X_i \in F(\vartheta)$ där ϑ är en okänd **parameter**.

En **skattning** av ϑ, $\vartheta^*(x_1, \ldots, x_n)$ är en observation av den s.v. $\vartheta^*(X_1, \ldots, X_n)$. Båda betecknas oftast bara med ϑ^*.



Minsta kvadrat-metoden, MK

Om $E(X_i) = \mu_i(\vartheta)$ så fås **MK-skattningen** av ϑ genom att minimera förlustfunktionen

$$Q(\vartheta) = \sum_{i=1}^{n} \Big(x_i - \mu_i(\vartheta) \Big)^2$$

m.a.p. ϑ.

Maximum likelihood-metoden, ML

ML-skattningen av ϑ fås genom att maximera likelihood-funktionen $L(\vartheta; x_1, \ldots, x_n)$ m.a.p. ϑ.

$$L(\vartheta) = p_X(x_1) \cdot \ldots \cdot p_X(x_n) \qquad \text{(diskr.)}$$
$$L(\vartheta) = f_X(x_1) \cdot \ldots \cdot f_X(x_n) \qquad \text{(kont.)}$$

Konfidensintervall

Ett **konfidensintervall** för en parameter ϑ täcker rätt värde på ϑ med sannolikheten $1 - \alpha$.

$1 - \alpha$ kallas **konfidensgrad**. Vanliga värden är 0.95, 0.99 och 0.999.

Ett **tvåsidigt** konfidensintervall är alltså **två** skattningar a_1^*, a_2^* så att

$$\mathrm{P}\Big(a_1^*(X_1, \ldots, X_n) < \vartheta < a_2^*(X_1, \ldots, X_n)\Big) = 1 - \alpha$$

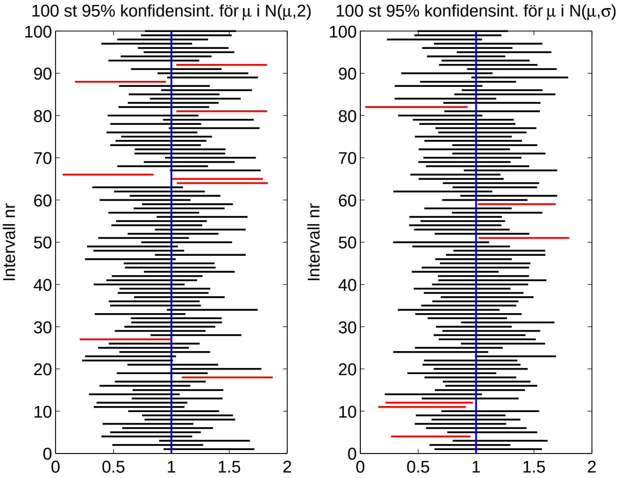
Ett **ensidigt** konfidensintervall är **en** skattning a_1^* eller a_2^* så att

$$\mathrm{P}\Big(a_1^*(X_1, \ldots, X_n) < \vartheta < \infty\Big) = 1 - \alpha$$

eller

$$\mathrm{P}\Big(-\infty < \vartheta < a_2^*(X_1, \ldots, X_n)\Big) = 1 - \alpha$$

Andelen $1-\alpha$ av intervallen täcker rätt värde i långa loppet



α-kvantil, x_α

En **kvantil**, x_α, till en stokastisk variabel X är en gräns som överskrids med sannolikhet α. Den fås som lösning till någon av följande ekvationer.

$$F_X(x_\alpha) = 1 - \alpha \iff \int_{-\infty}^{x_\alpha} f_X(x)\,dx = 1 - \alpha \iff \int_{x_\alpha}^{\infty} f_X(x)\,dx = \alpha$$

Sats 6.1 — Standardiserad normalfördelning

Om $X \in N(\mu, \sigma)$, $E(X) = \mu$, $V(X) = \sigma^2$ så är

$$\frac{X - \mu}{\sigma} \in N(0, 1) \text{ med kvantiler } \lambda_\alpha$$

Konfidensintervall för μ då $X_i \in \mathsf{N}(\mu, \sigma)$, i fallet σ **känd**

- En skattning av μ är:

$$\mu^* = \frac{1}{n} \sum_{i=1}^{n} X_i$$

  med $\mathsf{E}(\mu^*) = \mu$ och $\mathsf{D}(\mu^*) = \frac{\sigma}{\sqrt{n}}$.

- Enligt Sats 6.1 är

$$\frac{\mu^* - \mu}{\mathsf{D}(\mu^*)} \in \mathsf{N}(0, 1).$$

- Vi söker nu två tal så att:

$$\mathsf{P}\left(? < \frac{\mu^* - \mu}{\mathsf{D}(\mu^*)} < ? \right) = 1 - \alpha$$

- Konfidensintervallet för μ blir då: $I_\mu = \mu^* \pm \lambda_{\alpha/2} \cdot \mathsf{D}(\mu^*)$.

Konfidensintervall för μ då $X_i \in \mathsf{N}(\mu, \sigma)$, i fallet σ **okänd**

Om σ är okänd ersätts $\mathsf{D}(\mu^*)$ med medelfelet:

$$d(\mu^*) = \frac{s}{\sqrt{n}} \qquad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Men, nu är $\dfrac{\mu^* - \mu}{d(\mu^*)}$ inte $N(0, 1)$ ☹

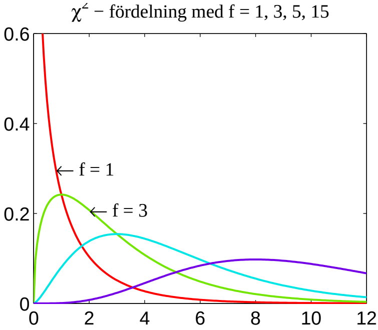
χ^2-fördelning (*uttalas "chi-två"*)

- $Y \in \chi^2(f)$. f kallas antal frihetsgrader.
- α-kvantil: $\chi^2_\alpha(f)$. Se Tabell 4.

Om $X_1, \ldots, X_n \in N(\mu, \sigma)$ och oberoende så gäller

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \in \chi^2(n)$$

...men $$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \in \chi^2(n-1)$$

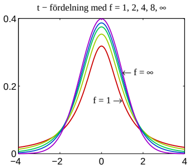


Student's t-fördelning

- $X \in t(f)$. f kallas antal frihetsgrader.
- α-kvantil: $t_\alpha(f)$. Se Tabell 3.

Om $X \in N(0,1)$ och $Y \in \chi^2(f)$ är oberoende gäller

$$\frac{X}{\sqrt{Y/f}} \in t(f)$$



och speciellt för $X_i \in N(\mu, \sigma)$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \in t(n-1)$$

där $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S^2 = Q/f$, $Q = \sum_{i=1}^{n}(X_i - \bar{X})^2$, och $f = n - 1$.

$\bar{X}$ och S^2 råkar vara oberoende (förvånande nog!)

Student — William Sealy Gosset

Konfidensintervall för μ i $N(\mu, \sigma)$

$x_1, \ldots, x_n$ observationer av $X_i \in N(\mu, \sigma)$

σ känd:

$$I_\mu = \mu^* \pm \lambda_{\alpha/2} \cdot D(\mu^*) = \bar{x} \pm \lambda_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

σ okänd:

$$I_\mu = \mu^* \pm t_{\alpha/2}(f) \cdot d(\mu^*) = \bar{x} \pm t_{\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}}$$

Kvantilerna ges då av:

- $\lambda_{\alpha/2}$ är $N(0, 1)$-fördelningens $\alpha/2$-kvantil (Tabell 2)
- $t_{\alpha/2}(n-1)$ är $t(n-1)$ fördelningens $\alpha/2$-kvantil (Tabell 3)

Exempel: Sockerinnehåll i sockerbetor

Sockerbetor har i regel ett sockerinnehåll på $16 - 19\,\%$ (enligt Wikipedia). Anta att sockerinnehållet i en godtycklig beta beskrivas av $X_i \in \mathrm{N}(\mu, \sigma)$ med σ okänd. I ett visst betlass undersökte man sockerhalten hos 25 slumpmässigt utvalda betor.

$$\frac{1}{25}\sum_{i=1}^{25} x_i = 16.8 \qquad \sum_{i=1}^{25}(x_i - \bar{x})^2 = 4.8$$

Gör ett $95\,\%$-konfidensintervall för den förväntade sockerhalten i betlasset.

Konfidensintervall baserade på normal(approximation)

Om $\vartheta^* \in N(\vartheta, D(\vartheta^*))$ eller $\vartheta^* \underset{\sim}{\in} N(\vartheta, D(\vartheta^*))$:

$D(\vartheta^*)$ **känd** känd

$$I_\vartheta = \vartheta^* \pm \lambda_{\alpha/2} \cdot D(\vartheta^*)$$

$D(\vartheta^*)$ **okänd** okänd så skattas med $d(\vartheta^*)$

$$I_\vartheta = \vartheta^* \pm t_{\alpha/2}(f) \cdot d(\vartheta^*) \quad \text{om } d(\vartheta^*) \text{ innehåller } \sigma^* = s = \sqrt{\frac{Q}{f}}$$

$$I_\vartheta = \vartheta^* \pm \lambda_{\alpha/2} \cdot d(\vartheta^*) \quad \text{annars}$$

Ex: Konfidensintervall för p då $X \in \mathrm{Bin}(n, p)$

Vi vill uppskatta hur vanligt det är att det snöar i april i Målilla och konstaterar att under de 300 aprildagarna under perioden 1988–1997 så snöade det under 71 dagar. Antag att olika dagar är oberoende av varandra.

Beräkna ett approximativt 95 % konfidensintervall för sannolikheten att det snöar en slumpmässigt vald aprildag i Målilla.

Sammanvägd variansskattning

Om vi här
$$\begin{cases} x_1, \ldots, x_{n_x} & \text{obs. av } X_i \in \mathsf{N}(\mu_x, \sigma) \\ y_1, \ldots, y_{n_y} & \text{obs. av } Y_i \in \mathsf{N}(\mu_y, \sigma) \end{cases}$$

kan den gemensamma variansen σ^2 skattas med

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x - 1 + n_y - 1} = \frac{Q}{f}, \qquad \left(\text{där } \frac{Q}{\sigma^2} \in \chi^2(f)\right)$$

Ett konfidensintervall för $\mu_x - \mu_y$ blir t.ex.

$$I_{\mu_x - \mu_y} = \bar{x} - \bar{y} \pm t_{\alpha/2}(f) \cdot s_p \cdot \sqrt{1/n_x + 1/n_y}$$

eftersom $\mu_x^* - \mu_y^* = \bar{X} - \bar{Y} \in N\left(\mu_x - \mu_y, \sigma\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}\right)$

Stickprov i par

Vid många mätsituationer är det vanligt att man mäter före och efter en behandling på n inbördes olika föremål.
Modell (observera att paret X_i och Y_i **inte** är oberoende!)
(Före, efter, och skillnad):

$$\begin{array}{cccl} x_1 & \dots & x_n & \text{obs. av } X_i \in N(\mu_i, \sigma_x) \\ y_1 & \dots & y_n & \text{obs. av } Y_i \in N(\mu_i + \Delta, \sigma_y) \\ z_1 = y_1 - x_1 & \dots & z_n = y_n - x_n & \text{obs. av } Z_i = Y_i - X_i \in N(\Delta, \sigma) \end{array}$$

Vi vill nu skatta effekten av behandlingen (Δ).
Skatta Δ med $\bar{z}$ och gör konfidensintervall som vanligt för ett stickprov, dvs

$$I_\Delta = \bar{z} \pm t_{\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}},$$

$$\text{där} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})^2.$$

Stickprov i par?

- Blodtrycket hos ett antal patienter mäts förre och efter behandling med blodtryckssänkande medicin; konfidensintervall för sänkningen?
- Luftkvaliteten mäts längs Hornsgatan i Stockholm vintern 2009 (dubbdäck fortfarande tillåtna) och 2010 (efter dubbdäcksförbud); konfidensintervall för skillnaden i luftkvalitet?
- pH-värdet möts varje dag i Höjeå förre och efter Lunds reningsverk; konfidensintervall för skillnaden?

Ensidiga konfidensintervall

Konfidensintervall kan även vara uppåt eller nedåt begränsade. De konstrueras allmänt genom att

- Ta ena gränsen i ett tvåsidigt konfidensintervall
- Byt ut $\alpha/2 \longrightarrow \alpha$ för att få rätt konfidensgrad
- Låt den andra gränsen bli så stor/liten som möjligt

Ex. Om det tvåsidiga intervallet ges av $\bar{x} \pm \lambda_{\alpha/2} \cdot \dfrac{\sigma}{\sqrt{n}}$ fås följande ensidiga konfidensintervall:

Nedåt begränsat intervall: $\left(\bar{x} - \lambda_\alpha \cdot \frac{\sigma}{\sqrt{n}} \quad , \quad +\infty\right)$

Uppåt begränsat intervall: $\left(-\infty \quad , \quad \bar{x} + \lambda_\alpha \cdot \frac{\sigma}{\sqrt{n}}\right)$

Ensidiga KI är framförallt användbara vid ensidiga hypotestest.