Föreläsning 8

den 27 april 2021 10:00

Matematisk Statistik och Diskret Matematik, MVE051/MSG810, VT22

Timo Vilkas

Chalmers - Göteborgs Universitet

Föreläsning 8
26 april 2022

Intervallskattning - Konfidensintervall

När man väljer ut ett stickprov (av storlek n) ur en population, är som bekant stickprovsmedelvärdet $\overline{X}$ en (väntevärdesriktig) punktskattning av medelvärdet μ.
Det är dock mer meningsfullt att skatta genom att ange ett intervall, i vilket den okända parametern μ med stor sannolikhet ligger. Ett sådant intervall kallas **konfidensintervall**.
Det finns intervall av olika konfidensgrader: Ju större sannolikheten att μ faktiskt ligger i intervallet ska vara, desto större intervall måste man välja.

Hur bestämmer man ett konfidensintervall?

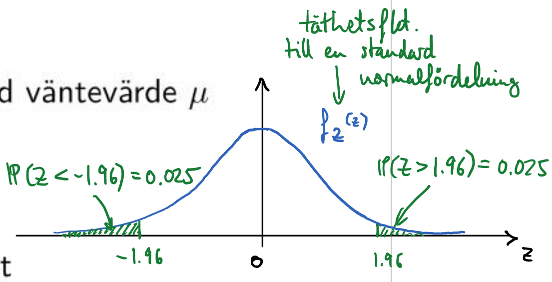
Strategin bestäms i första hand av det som är känt/okänt eller antaganden man gör:

- Stickprovet dras från en normalfördelning med **känd** varians.
- Stickprovet dras från en **icke-normal** eller okänd fördelning.
- Stickprovet dras från en normalfördelning med **okänd** varians.

Stickprov ur en normalfördelning

Antag att värden i populationen är normalfördelade med väntevärde μ och varians σ^2. Låt $X_1, \ldots, X_n$ vara ett slumpmässigt valt stickprov (dvs. oberoende $\mathcal{N}(\mu, \sigma^2)$ variabler).

- Stickprovsmedelvärdet $\overline{X}$ är normalfördelat med väntevärde μ och varians $\frac{\sigma^2}{n}$.
- $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$
- $\mathbb{P}(-1.96 < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < 1.96) = 0.95$ eller ekvivalent $\mathbb{P}(\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}) = 0.95$
- Om man nu, baserat på utfallet $\overline{X} = \overline{x}$, väljer intervallet $(\overline{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}})$, så ligger sannolikheten att det innehåller väntevärdet μ på 0.95. Därför kallas det för ett 95% konfidensintervall (KI).



Exempel

Antag att en forskare, som är intresserad av att skatta den genomsnittliga koncentrationen av någon enzym i en given population, väljer ut ett stickprov av 10 individer, bestämmer koncentrationen hos dem och hittar ett stickprovsmedelvärde på $\overline{x} = 22$. Antag att variabeln (alltså koncentrationen) kan betraktas som normalfördelad med väntevärde μ och varians $\sigma^2 = 40$. Hitta ett 95% konfidensintervall för μ.

Lösning: Enligt idén ovan kan vi ta intervallet

$$(\overline{x} - 1.96 \cdot \tfrac{\sigma}{\sqrt{n}}, \overline{x} + 1.96 \cdot \tfrac{\sigma}{\sqrt{n}}).$$

Då $\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{40}{10}} = 2$, ges intervallet av

$$(22 - 1.96 \cdot 2, 22 + 1.96 \cdot 2)$$

som blir

$$(18.08, 25.92).$$

Konfidensintervall för μ av en normalfördelad population med känd σ

I allmänhet, ges ett konfidensintervall för μ (av en normalfördelad population med känd σ) av **konfidensgrad** $(1 - \alpha)\,100\%$ av

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},$$

där $z_{\alpha/2}$ är $(1 - \frac{\alpha}{2})\,100$-percentilen till standard normalfördelningen, dvs. talet så att $\mathbb{P}(Z > z_{\alpha/2}) = \frac{\alpha}{2}$ för $Z \sim \mathcal{N}(0, 1)$.
$z_{\alpha/2}$ kallas också **reliabilitetskoefficient**.
Det symmetriska konfidensintervallet för μ till grad $(1 - \alpha)\,100\%$ ges alltså av

$$\text{punktskattning} \pm (\text{reliabilitetskoefficient}) \times (\text{standardfel})$$

Interpretation

- **Sannolikhetsteoretisk interpretation:** Om vi upprepar det (bestämma konfidensintervallet $\overline{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ med hjälp av ett stickprov av storlek n ifrån samma normalfördelad population med känd varians), kommer i det långa loppet en andel av $(1-\alpha)$ av dessa innehålla den okända parametern μ (populationens medelvärde).
- **Praktisk interpretation:** Om stickprovet dras av en normalfördelad population med känd varians, är vi $100(1-\alpha)$ procent säkra att intervallet $\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ innehåller populationens medelvärde μ.

Exempel

En sjukgymnast vill bestämma/skatta, med 99 procent säkerhet, medelvärdet på den maximala styrkan av en viss muskel inom en given grupp av individer. Hon går med på antagandet att värden på styrkan är i stort sett normalfördelade med varians 144. Ett stickprov av 15 individer gav ett medelvärde på 93.11.

Lösning: Här är $\alpha = 0.01$. I tabell V hittar vi $z_{\alpha/2} = 2.575$, som talet så att $\mathbb{P}(Z > z_{\alpha/2}) = 0.005$ för en standard normalfördelad s.v. Z. Det symmetriska 99% konfidensintervallet till μ ges då av

$$\left(93.11 - 2.575 \cdot \sqrt{\frac{144}{15}}, 93.11 + 2.575 \cdot \sqrt{\frac{144}{15}}\right),$$

alltså

$$(85.13, 101.09).$$

Stickprov ur en icke-normal fördelning

I vanliga fall är fördelningen i populationen inte en normalfördelning eller inte ens känd. Hur ska man då gå till väga för att bestämma konfidensintervall för medelvärdet μ?

Sats (Centrala gränsvärdessatsen, *central limit theorem*)

Låt $X_1, \ldots, X_n$ vara ett stickprov av storlek n ur en fördelning med väntevärde μ och varians σ^2. För tillräckligt stora n är då $\overline{X}$ approximativt normalfördelat med väntevärde μ och varians σ^2/n.

Bevis: standard med karakteristiska funktioner, utanför kursen.

Anmärkning: Som tumregel är n tillräckligt stort ifall $n \geq 25$.

Exempel

Punktlighet hos patienter undersöks i en studie. Ett stickprov på 35 patienter visade att de kom 17.2 minuter för sent i genomsnitt. Tidigare undersökningar visade att standardavvikelsen på förseningen ligger på ungefär 8 minuter. Datan ger upphov till antagandet att förseningen inte är normalfördelad.

1. Vilken approximation är lämplig för fördelningen av stickprovsmedelvärdet $\overline{X}$? Med vilka parametrar?
2. Bestäm ett 90% konfidensintervall för μ, medelvärdet av förseningen till en läkartid i populationen.

$n = 35 \geq 25$
$\bar{x} = 17.2$

$\sigma = 8$

Exempel (Lösning)

1. Då $n > 25$ är $\overline{X}$ enligt centrala gränsvärdessatsen approximativt normalfördelat med väntevärde μ och varians $\frac{\sigma^2}{n} = \frac{64}{35}$. ⇝ $\frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{35}} \approx 1.3522$
okänt

2. Konfidensgrad 90% motsvarar $\alpha = 0.1$; $z_{\alpha/2} = z_{0.05} = 1.645$.

⇒ Ett (symmetriskt) 90% K.I. för μ ges då av

$$\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) = (17.2 - 1.645 \cdot 1.3522, 17.2 + 1.645 \cdot 1.3522)$$
$$= (15.0, 19.4).$$

Fördelningen av en normalfördelad populationens stickprovsvarians

Låt $X_1, \ldots, X_n$ vara ett stickprov ur populationen. Vi hittade att $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ är en (väntevärdesriktig) estimator för variansen σ^2.

Sats

Antag att $X_1, \ldots, X_n$ är ett stickprov (dvs. oberoende kopior) ur en normalfördelning med väntevärde μ och varians σ^2. Slumpvariabeln

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2}$$

har då en χ^2- fördelning med $n-1$ frihetsgrader.

Konfidensinterval för variansen

- Vill vi hitta ett $100(1-\alpha)\%$ KI för σ^2, letar vi först efter ett $100(1-\alpha)\%$ KI för $\frac{(n-1)S^2}{\sigma^2}$:

  Ett motsvarande KI för $\frac{(n-1)S^2}{\sigma^2}$ ges av $(q_{\alpha/2}, q_{1-\alpha/2})$, där q_p betecknar $100p$–percentilen *(eng. p-quantile)* till χ^2-fördelningen med $n-1$ frihetsgrader, dvs.

  dvs. $\left.\begin{array}{l}\mathbb{P}(X < q_{\alpha/2}) \\ \mathbb{P}(X > q_{1-\frac{\alpha}{2}})\end{array}\right\} = \frac{\alpha}{2}$ för $X \sim \chi^2_{n-1}$

$$\mathbb{P}\left(q_{\alpha/2} \le \frac{(n-1)S^2}{\sigma^2} \le q_{1-\alpha/2}\right) = 1-\alpha.$$

- En enkel algebraisk omformning ger då att följande är ett $100(1-\alpha)\%$ KI för σ^2:

$$\left(\frac{(n-1)S^2}{q_{1-\alpha/2}}, \frac{(n-1)S^2}{q_{\alpha/2}}\right).$$

Exempel

Stickprov ur en normalfördelad population:

$$9.7,\ 12.3,\ 11.2,\ 5.1,\ 24.8,\ 14.8,\ 17.7$$

Bestäm ett 95% konfidensintervall för populationens varians och standardavvikelse.

Lösning: Utfallet för stickprovsvariansen blir $s^2 = 39.763$. $\frac{(n-1)S^2}{\sigma^2}$ har $n-1=6$ frihetsgrader och $\alpha = 0.05$. Tabell IV i boken (s.695) levererar percentilerna till χ^2_6, nämligen $q_{0.025} = 1.24$ och $q_{0.975} = 14.4$. Därmed fås ett 95% KI för σ^2:

$$\left(\frac{6\cdot 39.763}{14.4}, \frac{6\cdot 39.763}{1.24}\right) = (16.57, 192.40)$$

och motsvarande 95% KI för σ ges av $(4.07, 13.87)$.

(Students) t-fördelning

Låt Z och Y vara oberoende s.v., där $Z \sim \mathcal{N}(0,1)$ och $Y \sim \chi^2_\gamma$.
Då har slumpvariabeln

$$T := \frac{Z}{\sqrt{Y/\gamma}}$$

den så kallade **(students) t-fördelning** med $\gamma(> 0)$ frihetsgrader.
Egenskaper hos t-fördelningen:

- T är kontinuerlig och motsvarande täthetsfunktion ges av

  $$f_T(t) = \frac{\Gamma(\frac{\gamma+1}{2})}{\Gamma(\gamma/2)\sqrt{\pi\gamma}}\Big(1 + \frac{t^2}{\gamma}\Big)^{-(\gamma+1)/2}, \quad t \in \mathbb{R}.$$

  Dess graf är klockformad och symmetrisk kring $t = 0$ (som motsvarar medelvärdet för $\gamma > 1$, annars är det odefinerat).
- Variansen av t-fördelningen är lika med $\frac{\gamma}{\gamma-2}$ för $\gamma > 2$ ($+\infty$ annars).
- För $\gamma \to \infty$ konvergerar täthetsfunktionen mot den av en standard normalfördelning.

Konfidensintervall för $\overline{X}$ ur en normalfördelad population med okänd varians σ^2

Låt $X_1, \dots, X_n$ vara ett stickprov ur en normalfördelad population med okända parametrar μ och σ^2. Då

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} =: Z \sim \mathcal{N}(0,1) \quad \text{och} \quad \frac{(n-1)S^2}{\sigma^2} =: Y \sim \chi^2_{(n-1)}$$

följer det att

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \cdot \Big(\frac{Y}{n-1}\Big)^{-1/2}$$

är t-fördelad med $n - 1$ frihetsgrader.

Konfidensintervall för $\overline{X}$ ur en normalfördelad population med okänd varians σ^2

- För att hitta ett $100(1-\alpha)\%$ KI för $\overline{X}$, följer vi samma strategi som för en normalfördelning, dvs. vi väljer
$$\overline{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$
där $\overline{x}$ är utfallet på stickprovets medelvärde, s på dess varians och $t_{\alpha/2}$ är talet så att $\mathbb{P}(T > t_{\alpha/2}) = \frac{\alpha}{2}$.
- Percentilerna till t-fördelningen hittas i tabell VI i boken (s.699-700).

Exempel (8.2.2. i boken)

Data om koncentrationen av svaveldioxid (i $\frac{\mu g}{m^3}$) från en skog i Bayern har tagits fram. Ett stickprov innehåller 24 värden med medelvärde $\overline{x} = 53.92\frac{\mu g}{m^3}$ och varians $s^2 = 101.48$, alltså standard avvikelse på $s = 10.07\frac{\mu g}{m^3}$. För att kunna ange ett ett 95% KI för μ behöver vi percentilen $t_{\alpha/2}$ ur tabellen, där $\gamma = 23$ och $\alpha = 0.05$: $t_{0.025} = 2.069$. Motsvarande 95% KI ges därmed av
$$\left(53.92 - \frac{2.069 \cdot 10.07}{\sqrt{24}}, 53.92 + \frac{2.069 \cdot 10.07}{\sqrt{24}}\right),$$
alltså
$$(49.67, 58.17).$$

Strategisk översikt – Konfidensintervall för populationens medelvärde