Statistik och Dataanalys I

Föreläsning 17 - Konfidensintervall för ett väntevärde

Oskar Gustafsson

Statistiska institutionen
Stockholms universitet

Översikt

- Samplingfördelningen för ett medelvärde
- Konfidensintervall för ett väntevärde
- Centrala gränsvärdessatsen och stora talens lag

Samplingfördelning för $\bar{X}$ - normalmodellen

- **Modell för populationen**: $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma)$.
- Vi **skattar** populationens väntevärde $\mu = E(X)$ med

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

> Samplingfördelning $\bar{X}$
>
> Om $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma)$, och σ känd
>
> $$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- Vi måste bevisa tre saker:
  - $\bar{X}$ är normalfördelad.
  - $E(\bar{X}) = \mu$
  - $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Samplingfördelning för $\bar{X}$ - normalmodellen

- **Normalfördelning**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

är en **summa** av normalfördelade variabler, **skalad** med $1/n$.
F13: $\bar{X}$ är normalfördelad.

- **$\bar{X}$ är väntevärdesriktig för μ**

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{\text{skalning}}{=} \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) \overset{\text{summa}}{=} \frac{1}{n}\left(\sum_{i=1}^{n} E(X_i)\right)$$

$$\overset{\text{modell}}{=} \frac{1}{n}\left(\sum_{i=1}^{n} \mu\right) \overset{\text{samma termer}}{=} \frac{1}{n}(n\mu) = \mu$$

- **Varians/Standardavvikelse**

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{\text{skalning}}{=} \left(\frac{1}{n}\right)^2 Var\left(\sum_{i=1}^{n} X_i\right) \overset{\text{summa, iid}}{=} \frac{1}{n^2}\left(\sum_{i=1}^{n} Var(X_i)\right)$$

$$\overset{\text{modell}}{=} \frac{1}{n^2}\left(\sum_{i=1}^{n} \sigma^2\right) \overset{\text{samma termer}}{=} \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

KI väntevärde - normalpopulation med känd varians

- Eftersom $Var(\bar{X}) = \frac{\sigma^2}{n}$ så har vi alltså att

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

- Från samplingfördelningen $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ kan vi skapa ett konfidensintervall:

Antag: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma)$, σ känd

$(1\text{-}\alpha)$**%-igt konfidensintervall för väntevärde** μ

$$\bar{x} \pm z_{\alpha/2} \cdot SD(\bar{x})$$

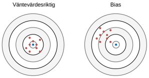
$$SD(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Normalpopulation med okänd varians

- Variansen i populationen, σ^2, är oftast **okänd**.
- Vi kan **skatta σ^2 med stickprovsvariansen**

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- Varför $n-1$? För att s^2 **är väntevärdesriktig för σ^2**



$$E(s^2) = \sigma^2$$

- **OM μ är känd** kan vi skatta σ^2 väntevärdesriktigt med

$$\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

- "Förlorar en frihetsgrad" när vi skattar μ med $\bar{x}$.
- SDM-boken (s. 538): stickprovet kommer ligga närmare $\bar{x}$ än μ i genomsnitt. Avvikelserna $x_i - \bar{x}$ blir för små i genomsnitt.

Okänd varians $\Longrightarrow$ student-t fördelning

- Om σ^2 är **känd**

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

och genom standardisering

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

- Om **okänd** σ^2 **skattas** med s^2 är $\bar{X}$ **student-t** fördelad:

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

- Alltså: när standardavvikelsen i populationen måste skattas får den standardiserade $\bar{X}$ **tyngre svansar**.
- När σ är känd: **standardavvikelse** för $\bar{X}$: $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
- När σ skattas: **standardfel** för $\bar{X}$: $SE(\bar{X}) = \frac{s}{\sqrt{n}}$

Student-t fördelning

Student-t distribution

$X \sim t_\nu(0,1)$

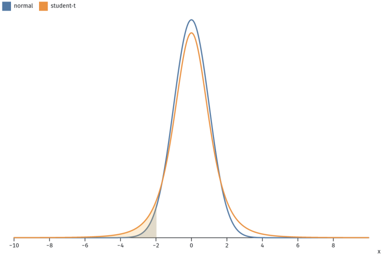
$E(X) = 0$ om $\nu > 1$

$Var(X) = \dfrac{\nu}{\nu - 2}$ om $\nu > 2$

Frihetsgrader ν : 4

Kvantil: -1.96

Normalfördelning: $P(X \leq -1.96) = 0.02500$
Student-t fördelning: $P(X \leq -1.96) = 0.06078$



Mattias Villani Student-t fördelning

Observable

K.I. väntevärde - normalpopulation, okänd varians

Antag: $X_1, X_2, \ldots, X_n \stackrel{iid}{\sim} N(\mu, \sigma)$, med σ okänd.

$(1\text{-}\alpha)$**%-igt konfidensintervall för väntevärde** μ

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot SE(\bar{x})$$

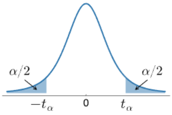
$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

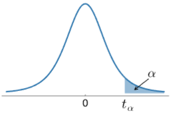
- **Kritiskt värde** $t_{\alpha/2,n-1}$ student-t med $n-1$ frihetsgrader.

Student-t tabell

Tvåsidigt



Ensidigt



Konfidensnivå:	80%	90%	95%	98%	99%
Tvåsidig sannolikhet:	0.200	0.100	0.050	0.020	0.010
Ensidig sannolikhet:	0.100	0.050	0.025	0.010	0.005
df					
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861

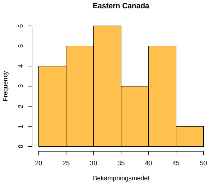
Bekämpningsmedel i odlad lax

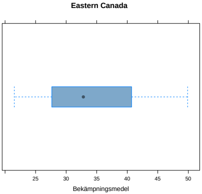
SDMdata Farmed_salmon

- Datamaterial med gifter i 153 laxar vid 8 olika platser.
- Här: Bekämpningsmedel (`Total.pestocide`) i Eastern Canada med $n = 24$ laxar:

```
x=c(25.739, 24.799, 27.563, 21.511, 23.821, 23.311, 49.883, 42.352, 44.598, 31.353,
33.837, 33.915, 41.668, 42.383, 43.638, 39.768, 35.256, 36.270, 29.630, 31.266, 32.577,
33.056, 29.789, 27.737)
```

- Modell: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma)$, och σ okänd.
- **Normalfördelad population?** Kolla stickprovet:





- Svårt se med få observationer. Histogram ok. Inga outliers.

Bekämpningsmedel i odlad lax

- **95%-igt konfidensintervall för μ:** $(30.332, 36.811)$

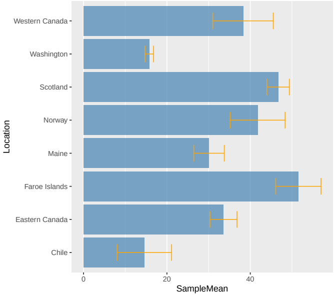
$$\bar{x} \pm t_{0.025,n-1} \frac{s}{\sqrt{n}}$$
$$33.572 \pm t_{0.025,23} \frac{7.671}{\sqrt{24}}$$
$$33.572 \pm 2.069 \frac{7.671}{\sqrt{24}}$$

- $t_{0.025,23} = 2.069$ från tabell, eller R: `qt(0.975, df = 23)`.
- Standardavvikelsen s beräknas i R som `sd(x)`, eller för hand:

$$s^2 = \frac{\sum_{i=1}(x_i - \bar{x})^2}{n-1}$$

Bekämpningsmedel i odlad lax

- **68%-igt konfidensintervall för** μ: $(31.980, 35.163)$
- **95%-igt konfidensintervall för** μ: $(30.332, 36.811)$
- **99%-igt konfidensintervall för** μ: $(29.176, 37.968)$
- Högre konfidens $\Longrightarrow$ bredare intervall.
- 95%-iga konfidensintervall alla orter:



- Se R-koden confidence_intervals_salmon.R på kurswebbsidan.

Konfidensintervall för väntevärde i R

```
> t.test(pesticides_EC, conf.level = 0.95)

        One Sample t-test

data:  pesticides_EC
t = 21.439, df = 23, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 30.33231 36.81103
sample estimates:
mean of x
 33.57167
```

KI väntevärde - normalpop, okänd varians, $n \geq 30$

Antag: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(0, \sigma)$, med σ okänd och $n \geq 30$.

Approximativt $(1\text{-}\alpha)$%-igt K.I. för väntevärde μ

$$\bar{x} \pm z_{\alpha/2} \cdot SE(\bar{x})$$

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Centrala gränsvärdessatsen

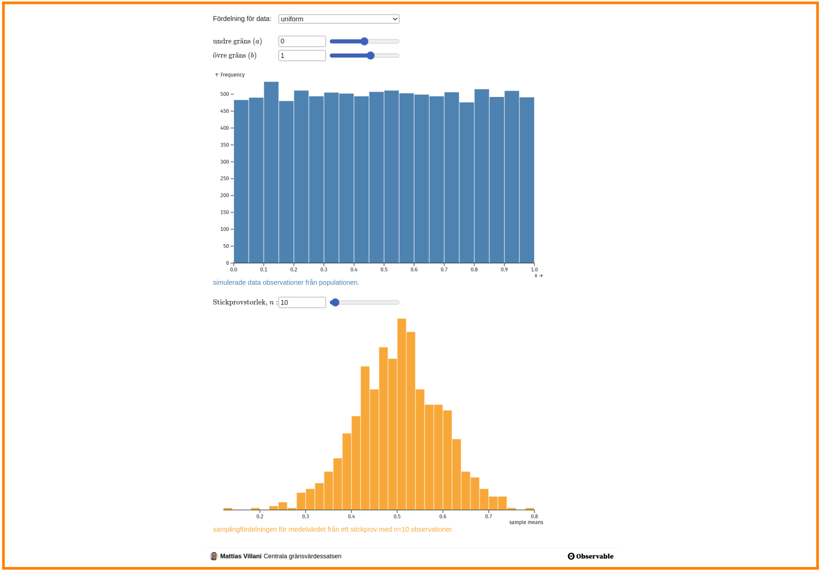
Om $X_1, X_2, \ldots, X_n$ är oberoende från en population med godtycklig fördelning (med ändlig varians σ^2) så är

samplingfördelningen för medelvärdet approximativt normalfördelat i stora stickprov:

$$\bar{X} \overset{\text{approx}}{\sim} \mathrm{N}\Big(\mu, \frac{\sigma}{\sqrt{n}}\Big) \text{ för tillräckligt stort } n$$

Tumregel: $n \geq 30$ är tillräckligt.

Centrala gränsvärdessatsen (CGS) - interaktivt



Stora talens lag

Om $X_1, X_2, \ldots, X_n$ är oberoende från en population med godtycklig fördelning (med ändligt väntevärde μ) så blir

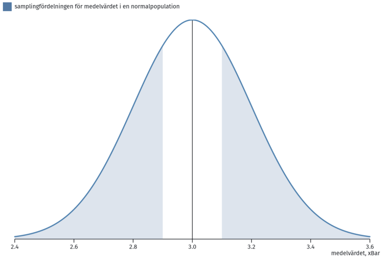
samplingfördelningen för medelvärdet alltmer koncentrerad kring μ när stickprovsstorleken n ökar.

Stora talens lag - interaktivt

Samplingfördelningen koncentreras kring μ

n	1	
ϵ	0.1	
μ	3	
σ	0.2	

$P(|\bar{X}_n - \mu| > 0.1) = 0.6171$



Mattias Villani Stora talens lag

Observable

Konfidensintervall för μ - tre olika situationer

- **Normalpopulation** med **känd varians** σ^2

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

- **Normalpopulation** med **okänd varians** skattad med s^2

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

- **Godtycklig populationsmodell** och $n \geq 30$ (CGS)

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Credits

Dessa slides skapades för kursen statistik och dataanalys 1 av Mattias Villani HT 2023, och har modifierats av Oscar Oelrich VT 2024, och Oskar Gustafsson för VT 2025.