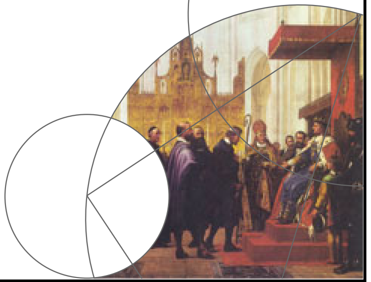
Det Natur- og Biovidenskabelige Fakultet

Analyse af en enkelt stikprøve: estimation og konfidensinterval

Anders Tolver
Institut for Matematiske Fag



Dagens program

Dagens emne: **Analyse af en enkelt stikprøve** (one sample).

Dagens forelæsninger dækkes primært af Kap. 4.2, 4.4 og 5.3.1-5.3.3 i lærebogen.

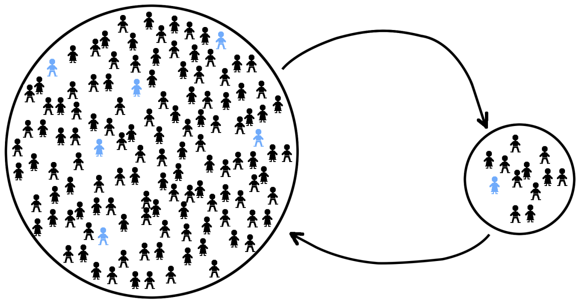
Formiddag:

- Intro/motivation (problemformulering)
- Egenskaber ved gennemsnit, CLT (matematik)
- Statistisk model, estimation og standard error (løsning)
- Konfidensinterval

Eftermiddag (video):

- Besvarelse af Quiz
- Analyse af transformeret stikprøve illustreret ved gæt på punktplot (opfølgning på HS.11 mm).

Dagens statistiske problemstilling



- **Response**: kvantitativ, kontinuert variabel
- **Interesseparameter**: middelværdien (μ) i populationen
- **Data**: (tilfældig) stikprøve fra populationen ($y_1, \ldots, y_n$)

Hvordan bruger vi data $y_1, \ldots, y_n$ til at udtale os om værdien af μ?

Løsningsstrategi og udfordring

- **Estimat**: gennemsnittet

$$\overline{y} = \frac{y_1 + \ldots + y_n}{n}$$

  er vores bedste bud på ukendt middelværdi μ

- **Konfidensinterval**: vil gerne finde interval

$$[q_{\mathrm{low}}; q_{\mathrm{up}}]$$

  omkring $\overline{y}$ som med stor sandsynlighed indeholder μ

Konfidensintervallet afspejler usikkerheden på et gennemsnit af en stikprøve med n observationer.

Hvordan/hvornår kan vi sige noget om usikkerheden på et gennemsnit?

Afstand mellem punkter

Eksempel:

- Alle studerende på StatData1 2022 har forsøgt at afsætte to punkter med afstand 8 cm på en farvet seddel
- En stikprøve består af målinger af afstanden for $n = 20$ tilfældigt udvalgte sedler

Gennemsnit i stikprøve er 7.65 cm, men hvor stor variation skal vi forvente, hvis vi trækker ny stikprøve?

Og kunne man forestille sig at gennemsnit i hele populationen er 8.0 cm?

Ingen forklarende variable i analysen (men kunne faktisk vælge at inddrage farven på sedlen).

Notation og terminologi

Lad os kalde **populationsgennemsnittet** μ. Interesseret i at bruge data (stikprøven) til at sige noget begavet om μ:

- **Estimat** (punktestimat) for populationsgennemsnittet. Naturligt at bruge stikprøvegennemsnittet: $\hat{\mu} = \bar{y}$
- Usikkerhed på estimatet: **Standard error** betegnes $\mathrm{SE}(\hat{\mu})$
- Et interval af μ-værdier der passer med data: **konfidensinterval** (intervalestimat) konstrueres som

$$\hat{\mu} - \text{noget} \cdot \mathrm{SE}(\hat{\mu})$$

Vi har brug for at sige noget om fordelingen af et gennemsnit!

Egenskaber ved gennemsnittet

Fordeling af gennemsnit

Vi forestiller os at vi ser **mange datasæt** der hver især består af n observationer. For hvert datasæt beregner vi gennemsnittet.

Stikprøve 1 (n observationer)	$\rightarrow$	$\bar{y}_1$
Stikprøve 2 (n observationer)	$\rightarrow$	$\bar{y}_2$
$\vdots$	$\vdots$	$\vdots$
Stikprøve 1000 (n observationer)	$\rightarrow$	$\bar{y}_{1000}$

Hvordan ser histogrammet for $\bar{y}_1, \dots, \bar{y}_{1000}$ ud?

Gennemsnit af normalfordelte variable

Infobox 4.3 Hvis $Y_1, \ldots, Y_n$ er uafhængige og alle $Y_i \sim N(\mu, \sigma^2)$, så er gennemsnittet $\bar{Y}$ også normalfordelt:

$$\bar{Y} = \frac{1}{n}(Y_1 + \cdots + Y_n) \sim N(\mu, \sigma^2/n)$$

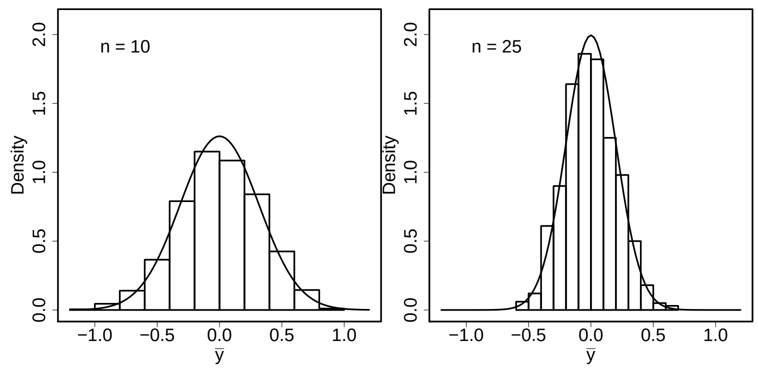
Specielt gælder:

$$\mathrm{sd}(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$$

Lad os prøve at illustrere det...

Fordeling af gennemsnit

Histogrammer over 1000 gennemsnit af n stk. $N(0,1)$ variable.



Ser faktisk ud til at være **normalfordelt** som Infobox 4.3 forudsagde. Passer middelværdi og spredning?

Repetition: er data normalfordelte?

I mandags så vi at jeres **gæt på afstande** kan beskrives ved normalfordeling.

Hvis data $y_1, \ldots, y_n$ er normalfordelt, **så** vil...

- tæthed for $N(\bar{y}, s^2)$ være en god approks. til histogrammet
- punkterne i QQ-plottet ligge omkring den rette linie med skæring $\bar{y}$ og hældning s

Her er: $\bar{y}$ gennemsnit og s stikprøvespredning.

Systematiske afvigelser er tegn på at data **ikke** er normalfordelte.

- Jo mindre n, jo større afvigelser kan vi acceptere
- Histogrammet dur kun for n nogenlunde stor

Model, estimation, standard error

Statistisk model

Data: $y_1, \ldots, y_n$. Målinger på repræsentativ stikprøve.

Statistisk model: $y_1, \ldots, y_n$ er uafhængige og alle normalfordelte med samme middelværdi μ og samme spredning σ.

En statistisk model angiver de antagelser vi gør os om hvordan "'de mekanismer"' der har genereret data.

Hvad betyder **uafhængighed?**

- Løst: Ingen information i én observation om nogle af de andre
- Eksempler på ikke-uafhængige data?

To ukendte **parametre** i modellen: Populationsgennemsnittet μ og populationsspredningen σ.

Estimation

To ukendte **parametre** i modellen: Populationsgennemsnittet μ og populationsspredningen σ.

Vores bedste gæt på parametrene er de tilhørende stikprøvestørrelser.

Estimation:

$$\hat{\mu} = \bar{y}, \quad \hat{\sigma} = s$$

Husk at $\bar{y}$ er normalford. med middelværdi μ og spredning $\sigma/\sqrt{n}$.

Standard error

$\bar{y}$ normalfordelt med middelværdi μ og spredning $\sigma/\sqrt{n}$

Standard error for $\hat{\mu} = \bar{y}$ er den estimerede spredning:

$$\mathrm{SE}(\hat{\mu}) = \mathrm{SE}(\bar{y}) = \frac{s}{\sqrt{n}}$$

Vores gæt på spredningen af $\bar{y}$.

For data vedr. afstande ml. punkter:

$$\hat{\mu} = \bar{\mu} = 7.65, \quad \mathrm{SE}(\hat{\mu}) = \mathrm{SE}(\bar{y}) = \frac{1.484}{\sqrt{20}} = 0.33$$

Evt.: Mindste kvadraters metode

Husk at vi fandt "'den bedste rette linie"' i lineær regression med mindste kvadraters metode.

Vi kan også bruge **mindste kvadraters metode** for en enkelt stikprøve: Vælg μ så residualkvadratsummen er så lille som mulig:

$$\text{Minimér} \quad \sum_{i=1}^{n} (y_i - \mu)^2$$

Residualkvadratsummen viser sig at være mindst for $\mu = \bar{y}$.

Skæg pointe: Minimeres istedet

$$\sum_{i=1}^{N} |y_i - m|,$$

så fås at m skal være *medianen* af y_i'erne.

Konfidensinterval

Konfidensinterval

Har estimat $\bar{y}$ — den værdi der "passer bedst" med vores data. Kaldes sommetider et **punktestimat**.

Ønsker et **intervalestimat** — et interval af μ-værdier der er "i overensstemmelse" med vores data. **Konfidensinterval.**

"Løsningen" viser sig at være

$$\hat{\mu} \pm \textit{noget} \cdot \mathrm{SE}(\hat{\mu})$$

Hvad er dette *noget*?

Konfidensinterval for μ

$\bar{y} \sim N(\mu, \sigma^2/n)$, så

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{y} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Eller — hvis vi omorganiserer så μ står i midten:

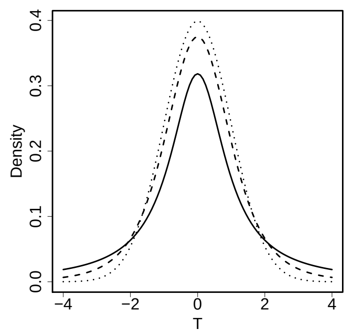
$$P\left(\bar{y} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Hvis vi kendte populationsspredningen σ, så ville vi kunne beregne endepunkterne $\bar{y} \pm 1.96\frac{\sigma}{\sqrt{n}}$.

Men: Vi kender ikke populationsspredningen σ. Oplagt at erstatte σ med s, men så skal 1.96 erstattes med et lidt større tal.

t-fordelingen

$\mathrm{df} = 1, 4$ og $N(0,1)$



Standardisering

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

Fordelingen ændres hvis σ erstattes med s:

$$T = \frac{\sqrt{n}(\bar{y} - \mu)}{\mathbf{s}} \sim t_{n-1}$$

***t*-fordelingen** med $n-1$ frihedsgrader ($\mathrm{df} = n-1$)

- Bredere haler end $N(0,1)$.
- Ligner $N(0,1)$ mere og mere når df vokser.

Konfidensinterval for μ

For kendt σ:

$$P\left(-1.96 < \frac{\sqrt{n}(\bar{y}-\mu)}{\sigma} < 1.96\right) = 0.95$$

Husk at 1.96 er 97.5% fraktilen i $N(0,1)$.

Hvis vi i stedet indsætter estimatet s, så skal vi bruge 97.5% fraktilen i t fordelingen med $n-1$ frihedsgrader:

$$P\left(-t_{0.975,n-1} < \frac{\sqrt{n}(\bar{y}-\mu)}{s} < t_{0.975,n-1}\right) = 0.95$$

Vi flytter rundt så μ står i midten:

$$P\left(\bar{y} - t_{0.975,n-1}\cdot\frac{s}{\sqrt{n}} < \mu < \bar{y} + t_{0.975,n-1}\cdot\frac{s}{\sqrt{n}}\right) = 0.95$$

Konfidensinterval for μ

Foregående slide:

$$P\left(\bar{y} - t_{0.975,n-1}\cdot\frac{s}{\sqrt{n}} < \mu < \bar{y} + t_{0.975,n-1}\cdot\frac{s}{\sqrt{n}}\right) = 0.95$$

Altså: Intervallet

$$\bar{y} \pm t_{0.975,n-1}\cdot\frac{s}{\sqrt{n}} \quad \text{eller} \quad \hat{\mu} \pm t_{0.975,n-1}\cdot \mathrm{SE}(\hat{\mu})$$

indeholder populationsmiddelværdien med 95% sandsynlighed.

Intervallet kaldes et **95% konfidensinterval for μ.**

Konfidensinterval for gennemsnitshøjde

Indhold på de næste 5 slides er også grundigt gennemgået i R programmet!

95% KI for den populationsgennemsnittet for **kvinder:**

$$168.52 \pm 1.983 \cdot \frac{6.64}{\sqrt{104}} = 168.52 \pm 1.29 = (167.23\,,\,169.82)$$

Værdier mellem 167.2 og 169.8 for populationsgennemsnittet er i overensstemmelse med data på 95% konfidensniveau.

95% KI for den populationsgennemsnittet for **mænd:**

$$182.70 \pm 2.010 \cdot \frac{5.54}{\sqrt{50}} = 182.70 \pm 1.57 = (181.13\,,\,184.27)$$

R: Kommentarer

Flere metoder til bestemmelse af konfidensintervallet i situationen med en stikprøve:

- "Manuelt". Brug `qt` til at finde t-fraktilen
- Funktionen `t.test`
- Med `lm` og `confint`

Bemærk: `lm` og `summary` giver flere ting: $\bar{y}$, $\mathrm{SE}(\bar{y})$, s mm.

R: "Manuelt"

```
### Gennemsnit og stikprøvespredning
> mean(kData$hojde, na.rm=TRUE)
[1] 168.524
> sd(kData$hojde, na.rm=TRUE)
[1] 6.639972

### Den relevante t-fraktil
> qt(0.975, df=103)
[1] 1.983264

### Nedre grænse
> 168.524 - 1.9833 * 6.639972/sqrt(104)
[1] 167.2327

### Øvre grænse
> 168.524 + 1.9833 * 6.639972/sqrt(104)
[1] 169.8153
```


R: t.test

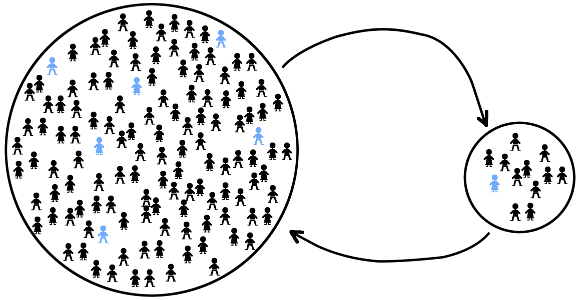
```
> t.test(kData$hojde)

      One Sample t-test

data:  kData$hojde
t = 258.83, df = 103, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  167.2327 169.8153
sample estimates:
mean of x
  168.524
```


R: lm

```
> model <- lm(hojde ~ 1, data=kData)
> summary(model)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 168.5240     0.6511   258.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.64 on 103 degrees of freedom
  (1 observation deleted due to missingness)

> confint(model)
               2.5 %   97.5 %
(Intercept) 167.2327 169.8153
```


Analyse af en enkelt stikprøve: når data ikke er normalfordelte

Population vs stikprøve



- Vi er interesserede i populationen
- Vi har kun målinger på en repræsentativ stikprøve (n)
- Særligt interesseret i populationegennemsnittet μ (ukendt).
- Men vi tror ikke på, at data er normalfordelte!

Spørgsmål

Brug stikprøven til at sige noget om **pop.-gennemsnittet** μ:

- **Estimat** (punktestimat) for populationsgennemsnittet. Naturligt at bruge stikprøvegennemsnittet: $\hat{\mu} = \bar{y}$
- Usikkerhed på estimatet: **Standard error**
- Et interval af μ-værdier der passer med data: **konfidensinterval** (intervalestimat)

Problemer forhold til tidligere analyse:

- Desværre ser data **ikke normalfordelte** ud
- Estimat $\hat{\mu} = \bar{y} \rightarrow$ egenskaberne for **gennemsnittet** er vigtige, men vi kan ikke bruge Infobox 4.3

To løsninger:

- Find transformation så data bliver normalfordelte (R program, øvelser, video)
- Træk på **Den centrale Grænseværdisætning** (CLT)

Den centrale grænseværdisætning (CLT)

I dagens R-program simuleres og beregnes gennemsnit fra population, som ikke er normalfordelt (transporttid til studie).

Overraskende: Gennemsnittet så ud til være normalfordelt uanset om "‘basisfordelingen"’ var en normalfordeling eller ej.

Det er præcis det **den centrale grænseværdisætning** (CLT) siger:

- **Hvis:** $y_1, \ldots, y_n$ er uafhængige og har den **samme fordeling**, med middelværdi μ og spredning σ
- **Så:** $\bar{y}$ approksimativt normalfordelt med middelværdi μ og spredning $\sigma/\sqrt{n}$

Gælder (næsten) uanset hvordan den bagvedliggende fordeling ser ud.

Konsekvenser af CLT

- For **store stikprøver** vil statistiske metoder baseret på normalfordelingsmodeller give fornuftige resultater
- Gælder også selvom data ikke er normalfordelte
- Gælder ikke kun for analyser af en enkelt stikprøve men også for fx. ensidet ANOVA og lineær regression
- **Udfordring:** Svært at afgøre, hvornår stikprøven er stor nok til at retfærdiggøre brug af normalfordelingsmodeller.

Opsummering - eget brug (en enkelt stikprøve)

Modelfiguren: Kontinuert respons, ingen forklarende variable.

Data: $y_1, \ldots, y_n$

Statistisk model: $y_1, \ldots, y_n$ er uafhængige og alle normalfordelte med samme middelværdi μ og samme spredning σ.

Estimation: $\hat{\mu} = \bar{y}$ og $\hat{\sigma} = s$

Standard error for $\hat{\mu}$: $\mathrm{SE}(\hat{\mu}) = \frac{s}{\sqrt{n}}$

95% konfidensinterval for μ: $\bar{y} \pm t_{0.975,n-1} \cdot \frac{s}{\sqrt{n}}$. De værdier af μ der er i overensstemmelse med data.

Bemærk struktur af KI:

$$\text{estimat} \pm t\text{-fraktil} \cdot \text{SE(estimat)}.$$

Opsummering — til eget brug

- Hvad er antagelserne i den statistiske model for en enkelt stikprøve?
- Hvordan estimeres populationsparametrene?
- Hvad er formlen for $\mathrm{SE}(\bar{y})$?
- Hvad er formlen for 95% konfidensintervallet for μ?
- Hvad er fortolkningen af konfidensintervallet?
- Kan du indlæse data fra en Excel og/eller tekstfil?