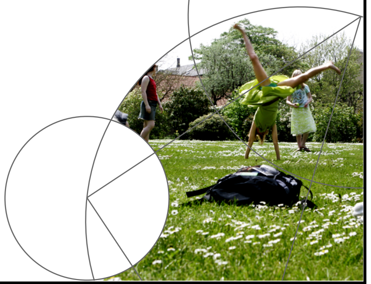
Faculty of Life Sciences

Statistisk inferens

En enkelt stikprøve og lineær regression
Stat. modeller, estimation og konfidensintervaller

Claus Ekstrøm
E-mail: ekstrom@life.ku.dk



Program

- Fordeling af gennemsnit
- Statistisk inferens for en enkelt stikprøve
  - statistisk model
  - estimation og præcision af estimater
  - t-fordelingen
  - konfidensintervaller
- Statistisk inferens for lineær regression

Gennemsnittet

Krabbedata:

- Interesset i den gennemsnitlige vægt i populationen — μ
- Har en stikprøve på 162 krabber: $y_1, \ldots, y_{162}$.
- Stikprøvestørrelser, $\bar{y} = 12.76$ og $s = 2.25$.
- Specielt, $\hat{\mu} = \bar{y} = 12.76$

Men:

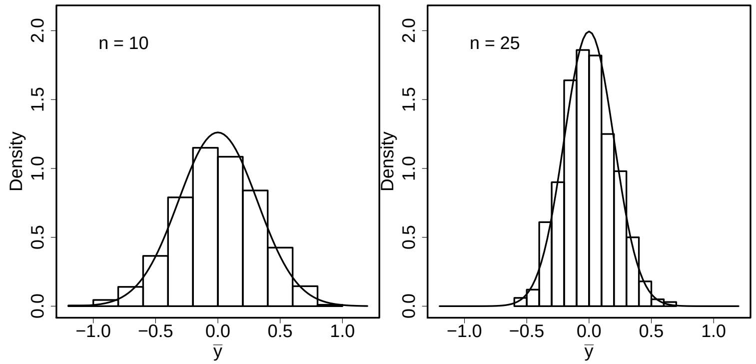
- Hvor meget kan stole på dette estimat? Hvor præcist er det?
- Hvad ville der ske hvis vi indsamlede 162 andre krabber?

Hvis vi bruger normalfordelingen, kan vi faktisk svare meget præcist på disse spørgsmål!

Vil lave konfidensinterval for μ. Dette kræver en statistisk model.

Fordeling af gennemsnit

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



Middelværdi? — Spredning? — fordeling?

Fordeling af gennemsnit

Husk fra sidst at sum af to normalfordelte variable og skalering af normalfordelte variable igen er normalfordelt.

Udvidelse til sum af n uafhængige $N(\mu, \sigma^2)$-variable:

- $y_1 + y_2 + \cdots + y_n \sim N(n\mu, n\sigma^2)$
- $\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n) \sim N(\mu, \sigma^2/n)$

Altså: $\hat{\mu} = \bar{y}$ er normalfordelt med middelværdi μ og spredning $\sigma/\sqrt{n}$.

Det fortæller os om variationen af $\bar{y}$!

Den centrale grænseværdisætning

Et af hovedresultaterne indenfor statistik *og* årsagen til at normalfordelingen er så pokkers vigtig.

Den centrale grænseværdisætning

Lad $Y_1, \ldots, Y_n$ være uafhængige variable med *samme fordeling* med middelværdi μ og spredning σ. Så er

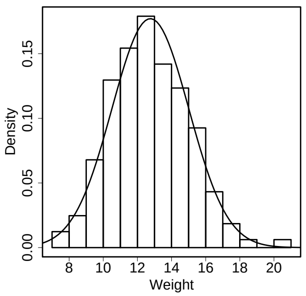
$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N(\mu, \sigma^2/n)$$

Specielt

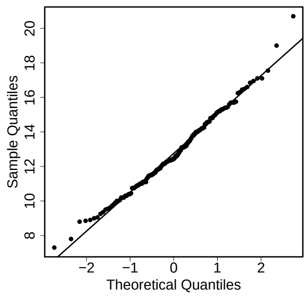
$$P\left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq z\right) \to \Phi(z)$$

Statistisk model

Histogram og N-tæthed



QQ-plot



Statistisk model: $y_1, \ldots, y_{162}$ er uafhængige og $y_i \sim N(\mu, \sigma^2)$

- Normalfordelt
- Alle y_i har samme middelværdi og samme spredning
- Uafhængighed — "deler ikke information"

Estimation

Statistisk model:

$$y_1, \ldots, y_{162} \sim N(\mu, \sigma^2) \text{ uafhængige}$$

Parametre i modellen

- middelværdien μ — gennemsnittet i populationen
- spredningen σ — spredningen i populationen

Estimation: populationsparametrene estimeres ved stikprøvestørrelserne.

- $\hat{\mu} = \bar{y}$ — det er faktisk LS estimatet
- $\hat{\sigma} = s$

Præcision af $\hat{\mu}$

Estimatet $\hat{\mu}$ siger ikke noget om præcisionen. Men vi ved jo at

- $\mathrm{sd}(\bar{y}) = \sigma/\sqrt{n}$
- $\bar{y}$ ligger i $\mu \pm 1.96 \cdot \sigma/\sqrt{n}$ med 95% sandsynlighed.

så $\bar{y}$ rammer rigtigt "i gennemsnit" og bliver mere og mere præcist jo større n bliver.

Åh-åh: kender ikke σ — kun estimatet s!

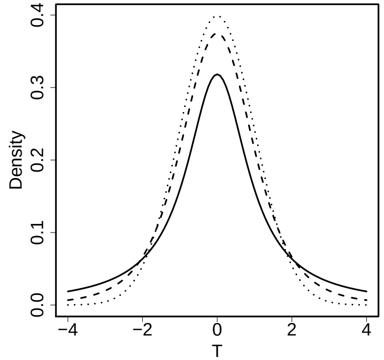
- Standard error af $\bar{y}$ — estimeret spredning:

$$\mathrm{SE}(\bar{y}) = s/\sqrt{n}$$

- $\bar{y}$ ligger i $\mu \pm ??? \cdot s/\sqrt{n}$ med 95% sandsynlighed. Fraktilen skal ændres for at tage højde i usikkerheden i estimatet for σ.

t-fordelingen

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:

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

t-fordelingen med $n-1$ frihedsgrader.

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

Skål

Øl

Gosset = Student

Konfidensinterval for μ

Hvis $t_{0.975,n-1}$ er 97.5%-fraktilen i t_{n-1}-fordelingen:

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

Hvis vi flytter rundt og isolerer μ:

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

Altså: intervallet

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

indeholder populationsmiddelværdien μ med ssh. 95%.

Intervallet kaldes et 95% konfidensinterval for μ.

Konfidensintervaller: krabbedata

Husk: $n = 162$, $\bar{y} = 12.75$ og $s = 2.25$.

Fraktiler:

```
> qt(0.975,161)
[1] 1.974808
> qt(0.95,161)
[1] 1.654373
```

Beregn:

- Standard error, $\mathrm{SE}(\hat{\mu})$?
- 95% konfidensinterval?
- 90% konfidensinterval?

Konfidensintervaller: fortolkning

95%-konfidensinterval for μ

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

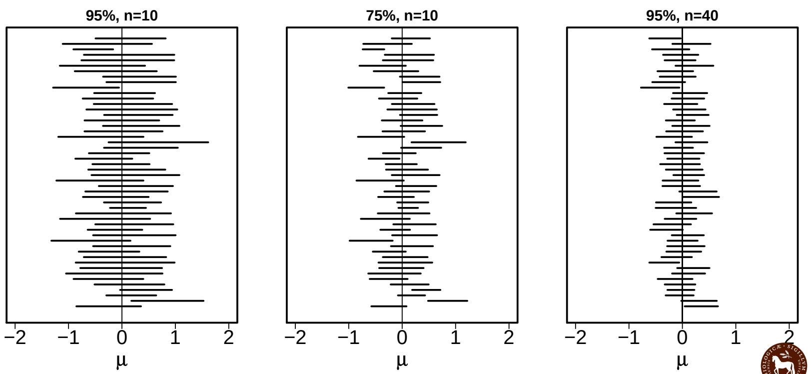
Fortolkning: intervallet indeholder med 95% sandsynlighed populationsgennemsnittet μ.

- Hvordan beregnes et 90%-konfidensinterval? Bliver det bredere eller smallere?
- Hvad sker der hvis stikprøvestørrelsen n vokser? Bliver det tilsvarende konfidensinterval bredere eller smallere?

Konfidensintervaller: fortolkning

Hvis vi gentog eksperimentet mange gange, så ville 95% af CI'erne indeholde populationsgennemsnittet.

Konfidensintervaller for 50 datasæt fra $N(0,1)$.



Resumé: en stikprøve

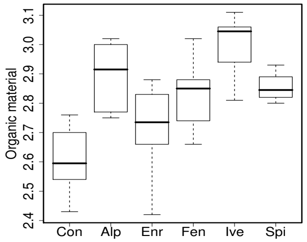
- Statistisk model: $y_1, \ldots, y_{162}$ er uafhængige og $y_i \sim N(\mu, \sigma^2)$
- Parametre, μ og σ: gennemsnit og spredning i populationen
- Estimater: $\hat{\mu} = \bar{y}$ og $\hat{\sigma} = s$
- Fordeling af estimat: $\hat{\mu}$ normalfordelt med middelværdi μ og spredning $\sigma/\sqrt{n}$
- Standard error, dvs. estimeret spredning for estimat: $\mathrm{SE}(\hat{\mu}) = s/\sqrt{n}$
- 95%-konfidensinterval: $\bar{y} \pm t_{n-1,0.975} \cdot \frac{s}{\sqrt{n}} = \hat{\mu} \pm t_{n-1,0.975} \cdot \mathrm{SE}(\hat{\mu})$

Vi kan køre præcis de samme punkter igennem for lineær regression og ensidet variansanalyse (og mange andre modeller).

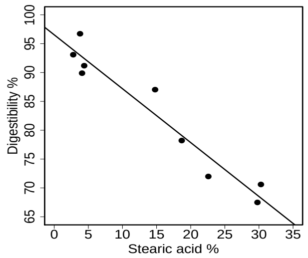
Hvorfor skal vi lære om normalfordelingen (nu)?

Har set tre typer af data/eksperimenter med kontinuerte data:

Ensidet ANOVA



Lineær regression



En stikprøve:

Blood pressure								
96	119	119	108	126	128	110	105	94

Vi skal bruge normalfordelingen for alle tre forsøgstyper/datatyper!

Statistisk model og parametre

Statistisk model: afvigelserne fra den rette linie er normalfordelt

$$y_i = \alpha + \beta \cdot x_i + e_i, \quad e_1, \ldots, e_n \sim N(0, \sigma^2) \text{ uafhængige}$$

Antagelserne er:

- Alle y_i er normalfordelte
- Middelværdien af y_i er $\alpha + \beta \cdot x_i$
- Alle y_i har samme spredning
- Uafhængighed

Parametre (populationsstørrelser)

- Skæring α og hældning β
- Spredning σ om den rette linie

Estimation og fordeling af estimater

Estimaterne $\hat\beta$ og $\hat\alpha$ så I allerede i uge 1...

Estimat for spredning:

$$s = \sqrt{\frac{1}{n-2}\sum_{i=1}^n (y_i - \hat\alpha - \hat\beta \cdot x_i)^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^n r_i^2}$$

$\hat\beta$ og $\hat\alpha$ er normalfordelte:

$$\hat\beta \sim N\left(\beta, \frac{\sigma^2}{\mathrm{SS}_x}\right), \quad \hat\alpha \sim N\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SS}_x}\right)\right), \quad \mathrm{SS}_x = \sum_{i=1}^n (x_i - \bar{x})^2.$$

Igen: Estimaterne rammer rigtigt i gennemsnit, med en præcision der vokser når n vokser.

Standard errors og konfidensintervaller

Fordelinger:

$$\hat\beta \sim N\left(\beta, \frac{\sigma^2}{\mathrm{SS}_x}\right), \quad \hat\alpha \sim N\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SS}_x}\right)\right)$$

Standard errors — estimerede spredninger

$$\mathrm{SE}(\hat\beta) = \frac{s}{\sqrt{\mathrm{SS}_x}}, \quad \mathrm{SE}(\hat\alpha) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SS}_x}}$$

95% konfidensintervaller:

$$\hat\beta \pm t_{0.975,n-2} \cdot \mathrm{SE}(\hat\beta), \quad \hat\alpha \pm t_{0.975,n-2} \cdot \mathrm{SE}(\hat\alpha)$$

Bemærk: t-fordelingen med $n-2$ frihedsgrader — fordi der er 2 middelværdiparametre. Samme som nævner i formel for s, df_e!

Stearinsyredata

```
> model1 = lm(ford~ssyre)
> summary(model1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 96.53336    1.67518   57.63 1.24e-10 ***
ssyre       -0.93374    0.09262  -10.08 2.03e-05 ***

Residual standard error: 2.97 on 7 degrees of freedom
```

Analyse:

- Statistisk model? Fortolkning af parametre?
- Estimater? Konfidensintervaller?

Dagens hovedpunkter

- Centrale grænseværdisætning — hvorfor er den central?
- Statistisk model og parametre
- Estimater, fordeling af estimater, standard error
- Konfidensintervaller: $\text{estimat} \pm t\text{-fraktil} \cdot \text{SE}(\text{estimat})$ og **fortolkning**

Ordliste

Engelsk	Dansk
average/mean	gennemsnit
confidence interval	konfidensinterval
degrees of freedom (df)	frihedsgrader
least squares method	mindste kvadraters metode
sample	stikprøve
standard deviation (sd)	spredning
standard error (SE)	estimeret spredning for estimat"