Mer Variansanalyse og Bootstrapping

Plan for forelesningen

- Litt mer om interaksjoner
- Variansanalyse med repetert målinger
- Bootstrapping
    - Ikke direkte relatert til variansanalyse, men veldig nyttig å kunne

PSY2014, 5. april 2022

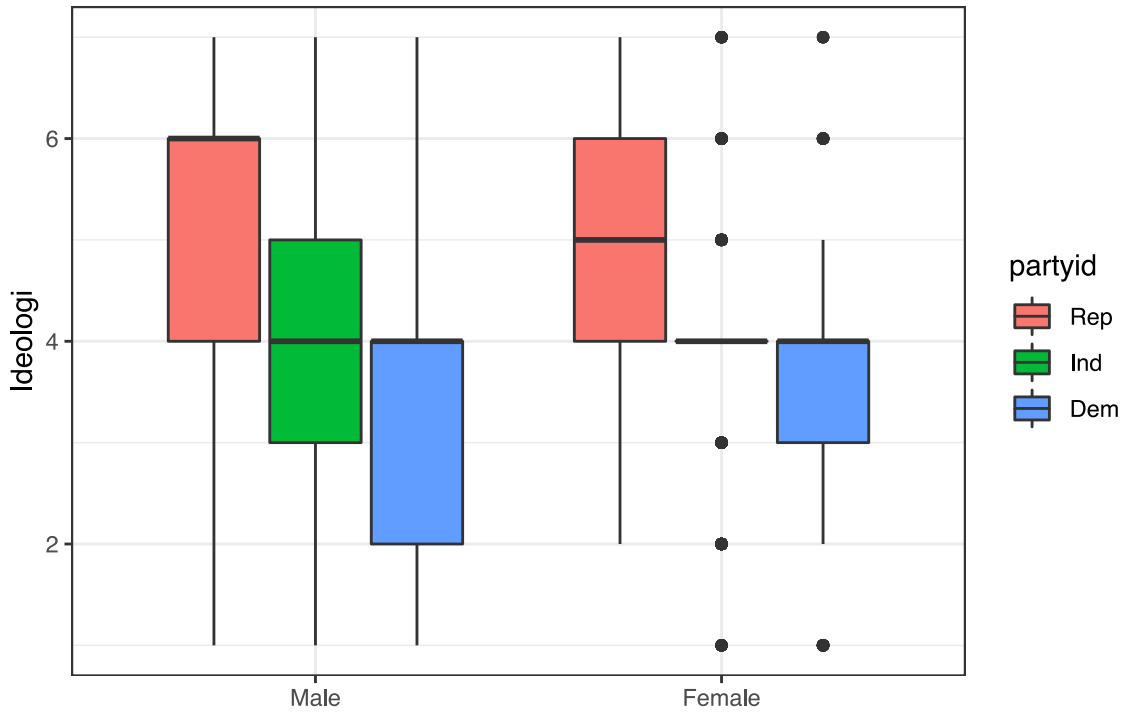
Interaksjoner

- Når vi har to eller flere kategoriske variabler i en variansanalyse
- Eksempeldata også brukt i Kapittel 12 i boka:

```
head(party_data)
```

```
## # A tibble: 6 × 3
##   ideologi partyid sex
##      <dbl> <fct>   <fct>
## 1        1 Dem     Male
## 2        1 Dem     Male
## 3        1 Dem     Male
## 4        1 Dem     Male
## 5        1 Dem     Male
## 6        1 Dem     Male
```

- Avhengig variabel **Ideologi**, på skala: 1-7, veldig liberal til veldig konservative.
- Ideologi avhenger av partitilhørighet, men avhenger det også av kjønn?



Steg 1, toveis anova med interaksjoner.

- Dummyvariabler for parti, $p_1 = 1$ for demokrater og $p_2 = 1$ for uavhengige.
- Dummaryvariabel for kjønn: $s = 1$ for kvinner.

Regresjonsmodell:

$$E(y) = \alpha + \beta_1 p_1 + \beta_2 p_2 + \beta_3 s + \beta_4 p_1 s + \beta_5 p_2 s$$

Impliserer, ved å plugge inn riktige verdier for dummyvariabler:

- Republikanere: $E(y) = \alpha + \beta_3$ for kvinner og $E(y) = \alpha$ for menn.
- Uavhengige: $E(y) = \alpha + \beta_2 + \beta_3 + \beta_5$ for kvinner og $E(y) = \alpha + \beta_2$ for menn.
- Demokrater: $E(y) = \alpha + \beta_1 + \beta_3 + \beta_4$ for kvinner og $E(y) = \alpha + \beta_1$ for menn.

R setter opp dummyvariablene for oss automatisk:

```
mod ← lm(ideologi ~ partyid * sex, data = party_data)
knitr::kable(coefficients(summary(mod)), digits = 3)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.290	0.078	67.692	0.000
partyidInd	-1.282	0.097	-13.151	0.000
partyidDem	-1.899	0.105	-18.102	0.000
sexFemale	-0.128	0.110	-1.165	0.244
partyidInd:sexFemale	0.090	0.135	0.667	0.505
partyidDem:sexFemale	0.241	0.142	1.699	0.090

I modellen tillater vi at effekten av partitilhørighet på ideologi varierer mellom kjønn. Eller tilsvarende, at effekten av kjønn på ideologi varierer mellom partitilhørigheter. En modell uten interaksjon antar at kjønn og partitilhørighet virker uavhengig av hverandre.

$$E(y) = \alpha + \beta_1 p_1 + \beta_2 p_2 + \beta_3 s$$

- Forskjellen mellom kvinner og menn i et gitt parti er nå β_3, uavhengig av parti.
- I interaksjonsmodellen var forskjellene $\beta_3 + \beta_4$ for demokrater, $\beta_3 + \beta_5$ for uavhengige og β_3 for republikanere.

Skal vi ha med interaksjonen?

Modell uten interaksjon:

```
mod0 ← lm(ideologi ~ partyid + sex, data = party_data)
```

Sammenligning:

```
anova(mod, mod0)
```

```
## Analysis of Variance Table
##
## Model 1: ideologi ~ partyid * sex
## Model 2: ideologi ~ partyid + sex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   2362 3779.4
## 2   2364 3784.5 -2    -5.024 1.5699 0.2083
```

Noen huskeregler for interaksjoner

- Dersom interaksjonen er signifikant, men hovedeffektene ikke er det, skal hovedeffektene likevel med i modellen. Interaksjoner gir ikke mening uten at hovedeffektene også er der.

- Interaksjoner betyr at effekten av én variabel varierer basert på verdien av en annen variabel. Begge må derfor tolkes samtidig. For eksempel, dersom interaksjonen hadde vært signifikant i vår modell for ideologi:
  - Forskjell mellom demokrater og republikanere lik β_1 for menn og $\beta_1 + \beta_4$ for kvinner.
  - Forskjell mellom kvinner og menn lik $\beta_3 + \beta_4$ for demokrater og β_3 for republikanere.

$$E(y) = \alpha + \beta_1 p_1 + \beta_2 p_2 + \beta_3 s + \beta_4 p_1 s + \beta_5 p_2 s$$

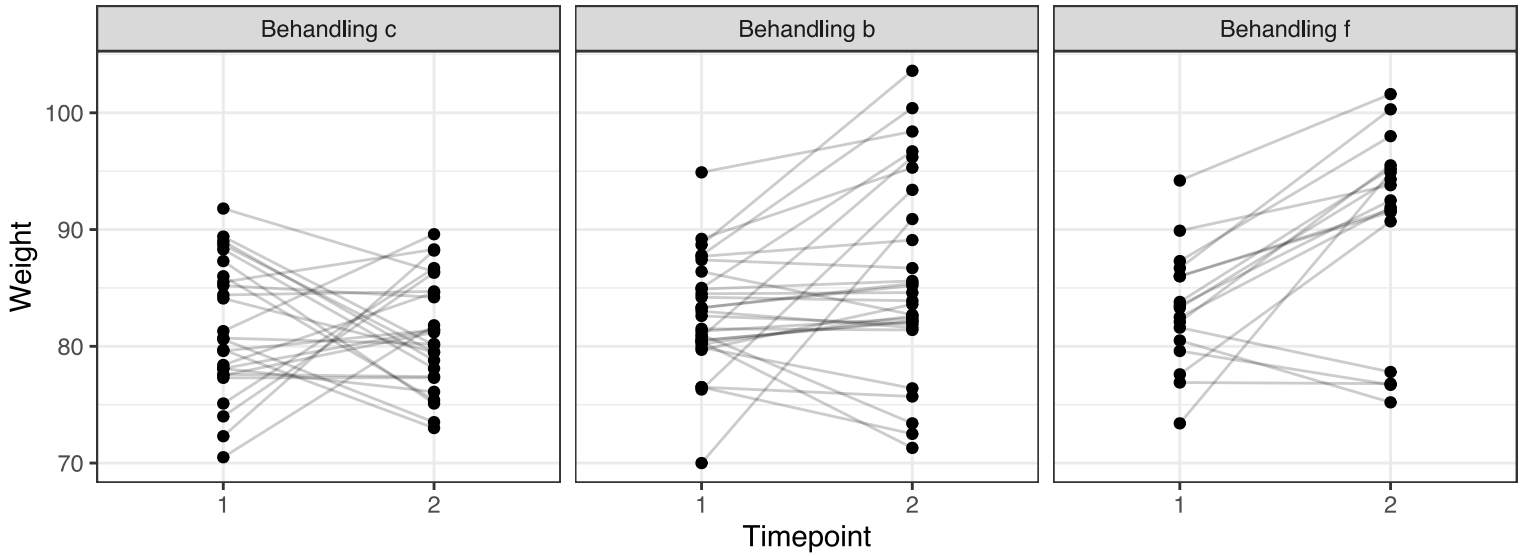
- Data fra boka, tilgjengelig fra https://tinyurl.com/2p99y9d7. Analysen her er litt annerledes.

```
head(Anorexia)
```

```
## # A tibble: 6 × 4
##   Subject Treat Timepoint Weight
##     <dbl> <fct> <chr>      <dbl>
## 1       1 b     1           80.5
## 2       1 b     2           82.2
## 3       2 b     1           84.9
## 4       2 b     2           85.6
## 5       3 b     1           81.5
## 6       3 b     2           81.4
```

Repeterte målinger

To målinger for tre behandlingsgrupper: Kontroll `c`, kognitiv terapi `b` og familieterapi `f`.



- La $s = 1$ i tidspunkt 2 og 0 i tidspunkt 1.
- La $t_1 = 1$ for `Treat = "b"` og $t_2 = 1$ for `Treat = "f"`.

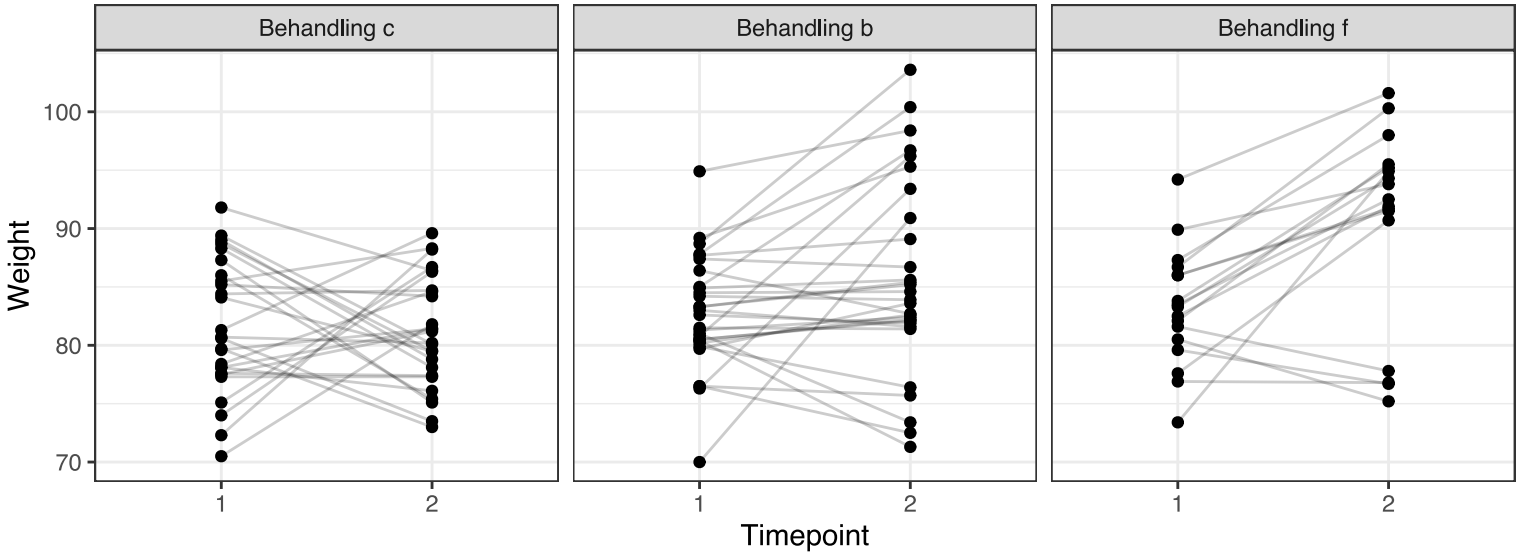
Følgende modell virker naturlig:

$$E(y) = \alpha + \beta_1 s + \beta_2 t_1 + \beta_3 t_2 + \beta_4 s t_1 + \beta_5 s t_2$$

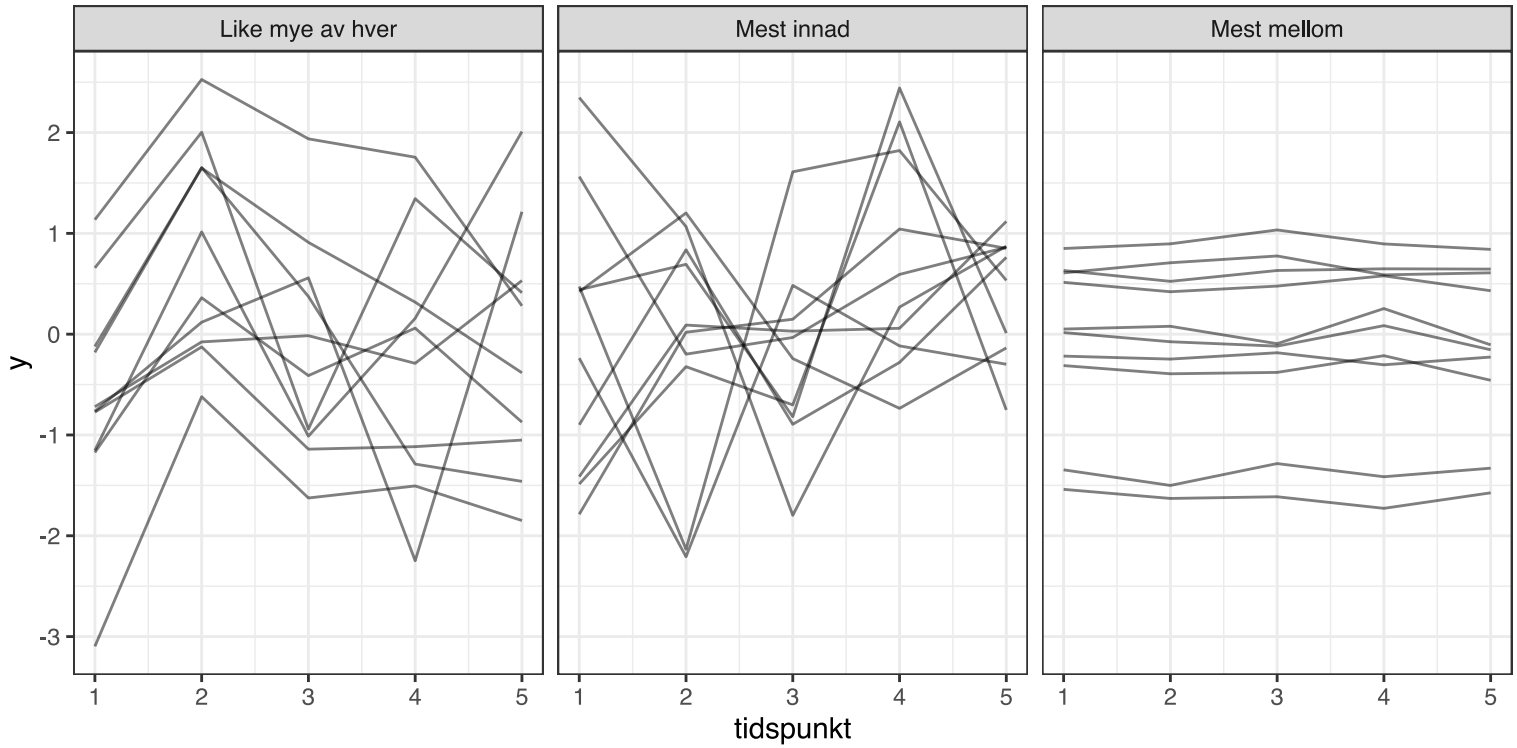
- α er intercept.
- β_1 er endring kontrollgruppa (vekt per tidsenhet).
- β_2 er forskjellen mellom kognitiv terapi og kontroll i tidspunkt 1.
- β_3 er forskjellen mellom familieterapi og kontroll i tidspunkt 1.
- β_4 er forskjell mellom kognitiv terapigruppe og kontrollgruppe i tidspunkt 2.
- β_5 er forskjell mellom familieterapi og kontrollgruppe i tidspunkt 2.

Repeterte målinger

- Problem: Observasjonene er ikke uavhengige. Vi har to stykk fra hver person.



Kilder til varians: Variasjon **mellom** individer og **innad** i individer. Figurene viser eksempler for en studie med 5 tidspunkter.



Random effects

- Vi legger til en dummy-variabel for hver eneste deltager, $z_1, z_2, \ldots, z_{72}$:

$$E(y) = \alpha + \beta_1 s + \beta_2 t_1 + \beta_3 t_2 + \beta_4 s t_1 + \beta_5 s t_2 + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \gamma_{72} z_{72}$$

- Koeffisientene $\gamma_1, \gamma_2, \ldots, \gamma_{72}$ kalles *random intercepts*. De fanger opp den systematiske variasjonen mellom individer. Det vanlige residualleddet fanger opp variasjonen innad i individer.

- Random effects er ikke parametre på samme måte som β-ene. De er i stedet ekstra residualledd som lar oss dele inn i variasjon innad og mellom.

La oss se på deltager 23. Behandlingen er `b`, så $t_1 = 1$ og $t_2 = 0$.

```
subset(Anorexia, Subject == 23)
```

```
## # A tibble: 2 × 4
##   Subject Treat Timepoint Weight
##     <dbl> <fct> <chr>      <dbl>
## 1      23 b     1           80.2
## 2      23 b     2           82.6
```

For deltager 23:

- I tidspunkt 1 har vi

$$y = \alpha + \beta_2 + \gamma_{23} + \epsilon_{23,1}$$

- I tidspunkt 2 har vi

$$y = \alpha + \beta_1 + \beta_2 + \beta_4 + \gamma_{23} + \epsilon_{23,2}$$

γ_{23} er lik i begge tidspunktene, og representerer hvordan nivået til deltager 23 skiller seg fra gjennomsnittet. $\epsilon_{23,1}$ og $\epsilon_{23,2}$ er variasjonen innad, og varierer fra gang til gang.

- Forventede verdi for deltager 23 er $E(y) = \alpha + \beta_2 + \gamma_{23}$ og $\alpha + \beta_1 + \beta_2 + \beta_4 + \gamma_{23}$.

- Forventede verdier for en tilfeldig person fra populasjonen, som også er i behandlingsgruppe `b`, er $E(y) = \alpha + \beta_2$ og $\alpha + \beta_1 + \beta_2 + \beta_4$.

Det er mulig å bruke `aov()`-funksjonen i R. Se Appendix A i boka. Her illustrerer jeg funksjonen `lmer()` fra pakka `lme4`.

```
library(lme4)
mod ← lmer(Weight ~ Timepoint * Treat + (1|Subject),
           data = Anorexia)
```

Random effects

```
summary(mod, correlation = FALSE)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Weight ~ Timepoint * Treat + (1 | Subject)
##    Data: Anorexia
##
## REML criterion at convergence: 913.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.10372 -0.43538  0.04229  0.46206  2.34248
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Subject  (Intercept) 11.80    3.435
##  Residual             28.34    5.323
## Number of obs: 144, groups:  Subject, 72
##
## Fixed effects:
##                   Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)         81.558      1.242 127.020  65.639   <2e-16 ***
## Timepoint2          -0.450      1.476  69.000  -0.305   0.7614
## Treatb               1.132      1.711 127.020   0.662   0.5095
## Treatf               1.672      1.976 127.020   0.846   0.3992
## Timepoint2:Treatb    3.457      2.033  69.000   1.700   0.0936 .
## Timepoint2:Treatf    7.715      2.348  69.000   3.285   0.0016 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Varians innad og mellom

Fra utskriften:

```
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Subject  (Intercept) 11.80    3.435
##  Residual             28.34    5.323
## Number of obs: 144, groups:  Subject, 72
```

- Etter at vi har tatt hensyn til de uavhengige variablene, er det større variasjon innad enn mellom.

Fixed effects

```
## Fixed effects:
##                    Estimate Std. Error t value
## (Intercept)          81.558      1.242  65.639
## Timepoint2           -0.450      1.476  -0.305
## Treatb                1.132      1.711   0.662
## Treatf                1.672      1.976   0.846
## Timepoint2:Treatb     3.457      2.033   1.700
## Timepoint2:Treatf     7.715      2.348   3.285
```

Husk modellen:

$$E(y) = \alpha + \beta_1 s + \beta_2 t_1 + \beta_3 t_2 + \beta_4 s t_1 + \beta_5 s t_2 + \textbf{random effects}$$

Det vil si:

- Forventede verdier for gruppe `b` er $E(y) = \alpha + \beta_2$ i tidspunkt 1 og $E(y) = \alpha + \beta_1 + \beta_2 + \beta_4$ i tidspunkt 2.
- Forventede verdier for gruppe `f` $E(y) = \alpha + \beta_3$ i tidspunkt 1 og $E(y) = \alpha + \beta_1 + \beta_3 + \beta_5$ i tidspunkt 2.

Fixed effects

```
## Fixed effects:
##                    Estimate Std. Error t value
## (Intercept)          81.558      1.242  65.639
## Timepoint2           -0.450      1.476  -0.305
## Treatb                1.132      1.711   0.662
## Treatf                1.672      1.976   0.846
## Timepoint2:Treatb     3.457      2.033   1.700
## Timepoint2:Treatf     7.715      2.348   3.285
```

For en ny person fra populasjonen:

$$E(y) = \alpha + \beta_1 s + \beta_2 t_1 + \beta_3 t_2 + \beta_4 s t_1 + \beta_5 s t_2$$

- Effekten av kognitiv terapi er

$$(\alpha + \beta_1 + \beta_2 + \beta_4) - (\alpha + \beta_2) = \beta_1 + \beta_4$$

Forventet verdi $\mathbf{3.457 - 0.450 = 3.007}$.

Hva er p-verdi og konfidensintervall?

Når modellene blir kompliserte, blir det vanskelig å finne nøyaktige formler. Vi trenger da å bruke approksimasjoner. Vi kan få p-verdier for β-ene, men ikke så lett for differansen mellom β-er.

```
library(lmerTest)
mod ← lmer(Weight ~ Timepoint * Treat + (1|Subject),
           data = Anorexia, REML = FALSE)
knitr::kable(coefficients(summary(mod)), digits = 3)
```

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	81.558	1.216	132.542	67.050	0.000
Timepoint2	-0.450	1.445	72.000	-0.311	0.756
Treatb	1.132	1.675	132.542	0.676	0.500
Treatf	1.672	1.935	132.542	0.864	0.389
Timepoint2:Treatb	3.457	1.990	72.000	1.737	0.087
Timepoint2:Treatf	7.715	2.299	72.000	3.356	0.001

Og vi kan få konfidensintervaller for β-ene:

```
confint(mod)
```

```
## Computing profile confidence intervals ...
```

```
##                             2.5 %    97.5 %
## .sig01                  1.6309658  4.716292
## .sigma                  4.4635018  6.193945
## (Intercept)            79.1562387 83.959146
## Timepoint2             -3.3210641  2.421064
## Treatb                 -2.1752071  4.439133
## Treatf                 -2.1475822  5.491021
## Timepoint2:Treatb      -0.4969996  7.410792
## Timepoint2:Treatf       3.1485302 12.280882
```

Hvordan får vi konfidensintervaller for effekten av kognitive terapi $\beta_4 + \beta_1$ eller effekten av familieterapi $\beta_5 + \beta_1$?

Her finnes det en metode som nesten alltid virker.

Bootstrapping

En simuleringsmetode, som fungerer som følger.

1. Bestem et stort tall T, f.eks. 10000
2. Trekk rader fra det opprinnelige datasettet, **med tilbakelegging**, slik at du får T nye like store datasett.
3. Tilpass T modeller, én til hvert av de nye datasettene, og for hver modell lagre verdien for statistikken av interesse, f.eks. $\beta_4 + \beta_1$.

Vi har nå en fordeling med T verdier av $\beta_4 + \beta_1$. Vi kan bruke prosentilene av denne fordelingen for å finne konfidensintervaller og p-verdier.

Bootstrapping - steg for steg

- Finn ut hvor mange observasjoner vi har:

```
(n ← nrow(Anorexia))
```

```
## [1] 144
```

- Trekk `n` nye rader, og ta en titt på dem:

```
nye_rader ← sort(sample(n, n, replace = TRUE))
nye_rader
```

```
##   [1]   2   4   4   5   8  10  11  11  14  15  16  19  19  19  20  22  23
##  [18]  23  27  27  27  29  30  30  30  33  40  41  42  42  44  44  46  48
##  [35]  51  52  52  53  53  54  57  60  60  61  61  61  61  63  64  65  65
##  [52]  67  68  71  71  71  74  74  75  76  76  77  79  79  79  79  80  80
##  [69]  80  81  81  81  82  82  82  83  83  84  86  88  88  88  89  89  91
##  [86]  91  92  92  95  97  97 101 102 103 104 107 108 108 109 110 112 112
## [103] 112 113 113 114 117 117 118 118 120 120 120 120 121 121 122 123 124
## [120] 125 125 125 125 125 125 125 126 127 130 131 131 132 134 134 135 135
## [137] 136 136 137 141 141 142 143 144
```

Lag et nytt datasett basert på `nye_rader`:

```
bootdata ← Anorexia[nye_rader, ]
```

```
head(bootdata)
```

```
## # A tibble: 6 × 4
##   Subject Treat Timepoint Weight
##     <dbl> <fct> <chr>      <dbl>
## 1       1 b     2           82.2
## 2       2 b     2           85.6
## 3       2 b     2           85.6
## 4       3 b     1           81.5
## 5       4 b     2           81.9
## 6       5 b     2           76.4
```

Bootstrapping - steg for steg

Vi tilpasser modell til dette datasettet

```
bootmod ← lmer(Weight ~ Timepoint * Treat + (1|Subject),
               data = bootdata, REML = FALSE)
```

Og lagrer $\beta_4 + \beta_1$

```
fixef(bootmod)
```

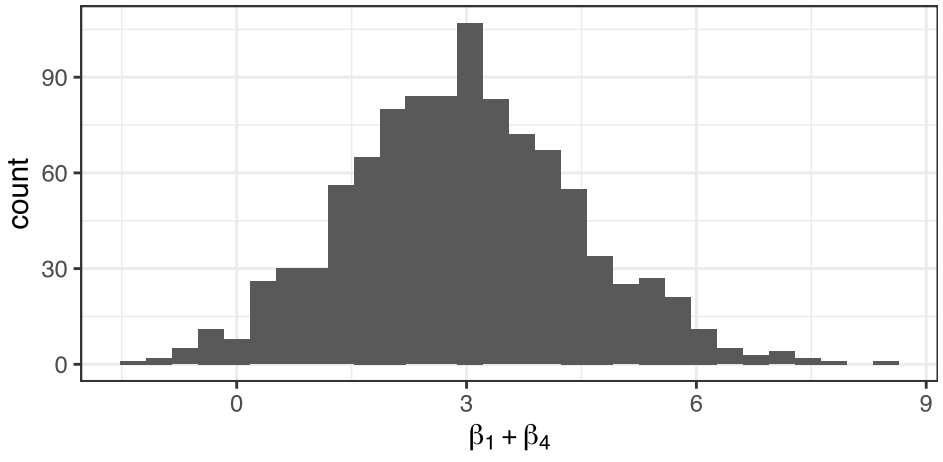
```
##       (Intercept)        Timepoint2            Treatb            Treatf
##         81.4227007          1.2464499          0.5881533          1.7852333
## Timepoint2:Treatb Timepoint2:Treatf
##          3.5567818          8.5061856
```

```
fixef(bootmod)[["Timepoint2"]] + fixef(bootmod)[["Timepoint2:Treatb"]]
```

```
## [1] 4.803232
```

Vi kan sette det hele sammen

```
boot_beta1beta4 ← rerun(1000, {
  nye_rader ← sample(n, n, replace = TRUE)
  bootdata ← Anorexia[nye_rader, ]
  bootmod ← lmer(Weight ~ Timepoint * Treat + (1|Subject),
                 data = bootdata)
  fixef(bootmod)[["Timepoint2"]] + fixef(bootmod)[["Timepoint2:Treatb"]]
})
boot_beta1beta4 ← as.numeric(boot_beta1beta4)
```



95 % konfidensintervall for $\beta_1 + \beta_4$ er

```
quantile(boot_beta1beta4, probs = c(.025, .975))
```

```
##      2.5%     97.5%
## 0.1143038 5.9969556
```

P-verdi for $H_0 : \beta_1 + \beta_4 \leq 0$ mot alternativet $\beta_1 + \beta_4 > 0$ er

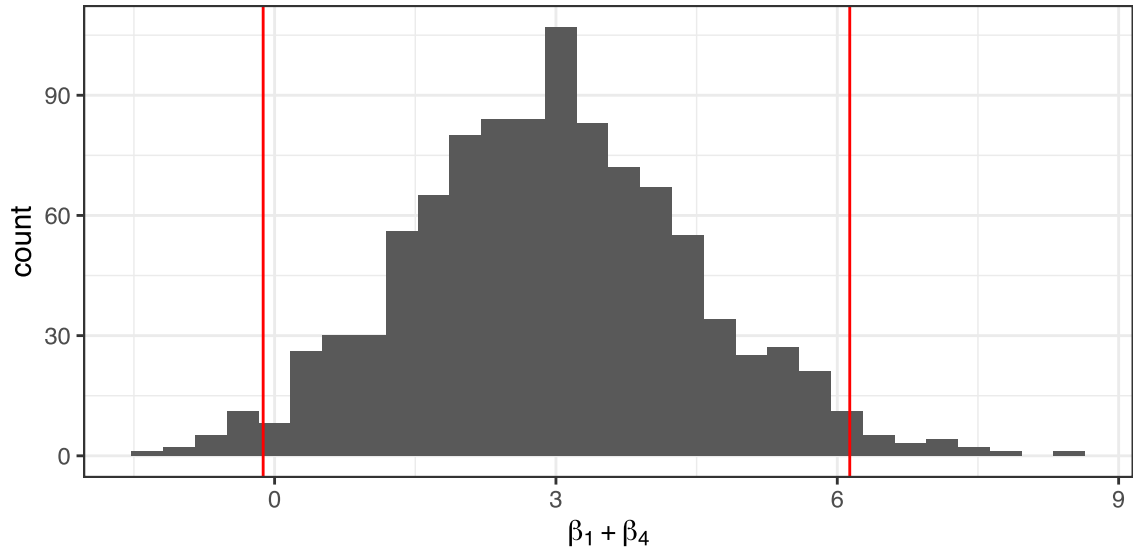
```
mean(boot_beta1beta4 < 0)
```

```
## [1] 0.024
```

P-verdien er ensidig mens konfidensintervallet er tosidig.

Bootstrapping

- Grensene for konfidensintervallet er markert i histogrammet.



Bootstrapping oppsummert

- Veldig nyttig teknikk når du trenger konfidensintervaller som ikke kommer rett ut av programmet.
- R-pakken `boot` kan gjøre jobben i mange tilfeller.
- For mixed models som vi brukte her, kunne vi også brukt `confint(mod, method = "boot")`.

Interaksjoner

- Når vi har to eller flere faktorer i en variansanalyse må vi sjekke om det er interaksjoner mellom dem.

Repeterte målinger

- Når vi har to eller flere repeterte målinger, må vi ta hensyn til at målingene av samme individ er korrelert.
- Random effects lar oss skille mellom variasjonen mellom individer og variasjonen innad i individer.

Bootstrapping

- Simuleringsmetode som gir konfidensintervaller. Nyttig når programmet ikke gir deg akkurat det du er ute etter.

Slides created via the R package **xaringan**.

The chakra comes from remark.js, **knitr**, and R Markdown.