Statistik för Biologer
F9: Mer om ANOVA

Shaobo Jin

Matematiska institutionen

Pingviner

Vår förskningsfråga är

Finns det någon skillnad i vikt mellan hanar och honor hos Adeliépingviner?

- Forskarna undersökte pingviner på tre olika platser.
- Om boplatsen påverkar vikten så kan vi inte bara köra ett t-test för att jämföra könen! Vi behöver veta att skillnader i boplats inte påverkar resultatet

Frekvenstabell

Antalet observationer insamlade under olika förutsättningar:

```
library(palmerpenguins)
adelie <- subset(penguins, species == "Adelie")
table(adelie$sex, adelie$island)

##
##           Biscoe Dream Torgersen
##   female      22    27         24
##   male        22    28         23
```

Modell

Vi tänker oss att vikten påverkas dels av vilken plats individen lever på och dels av vilket kön den har:

$$y_i = \mu_j + \tau_k + \epsilon_i$$

- y_i är observationen i prov i (t.ex. vikten för individ i)
- μ_j är medelvärdet för grupp j av en faktor som provet hör till (t.ex. medelvikten för de två könen, $j = 1, 2$)
- τ_k , är medelvärdet för grupp k av en annat faktor som provet hör till (t.ex. medelvikten på de tre platserna, $k = 1, 2, 3$)
- ϵ_i är hur mycket observationen i provet avviker från genomsnittet för stammen

Tvåvägs-ANOVA

Om:

1. Vi vill veta hur två olika variabler påverkar en responsvariabel y, eller
2. Vi vill veta hur en variabel påverkar en responsvariabel y, och ta hänsyn till en störande faktor .

kan vi utföra en **tvåvägs-ANOVA** - en ANOVA med två förklarande variabler.

ANOVA som vi gjorde tidigare är en envägs-ANOVA.

Medelvärden

```
aggregate(body_mass_g ~ island + sex, data = adelie,
          FUN = mean)
```

```
##       island    sex body_mass_g
## 1     Biscoe female    3369.318
## 2      Dream female    3344.444
## 3 Torgersen female    3395.833
## 4     Biscoe   male    4050.000
## 5      Dream   male    4045.536
## 6 Torgersen   male    4034.783
```

Tvåvägs-ANOVA i R

För att göra en mer formell analys och köra en tvåvägs-ANOVA lägger vi in de två förklarande variablerna på den högra sidan i formeln i aov

```
m <- aov(body_mass_g ~ island + sex, data = adelie)
summary(m)
```

```
##              Df   Sum Sq  Mean Sq F value Pr(>F)
## island        2     2064     1032   0.011  0.989
## sex           1 16622756 16622756 170.141 <2e-16 ***
## Residuals   142 13873382    97700
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
## 6 observations deleted due to missingness
```

Tolkning

1. p-värdet för skillnader mellan öarna är $0.989 > 0.05$. Skillnaden är inte signifikant.
2. p-värdet för skillnader mellan könen är mindre än 0.05. Skillnaden är signifikant.

Vi kan dessutom utesluta att skillnaderna vi ser beror på skillnader orsakade av skillnader boplatserna emellan. Detta skrivs ofta på följande vis: "Skillnaden mellan könen, justerad för boplats, är signifikant."

Ett Till Exempel: Kornskörd

I en studie jämfördes 5 olika kornsorter odlade på 6 olika platser. Man ville veta om kornsorterna gav olika stor skörd. För att ta hänsyn till odlingsplatsen tar vi med odlingsplats som en variabel i vår ANOVA.

```
library(MASS)
head(immer, n = 6)
```

```
##   Loc Var    Y1    Y2
## 1  UF   M  81.0  80.7
## 2  UF   S 105.4  82.3
## 3  UF   V 119.7  80.4
## 4  UF   T 109.7  87.2
## 5  UF   P  98.3  84.2
## 6   W   M 146.6 100.4
```

Loc är plats och Var är kornsort. Y1 är skördens storlek i 1931.

Kornskörd: ANOVA

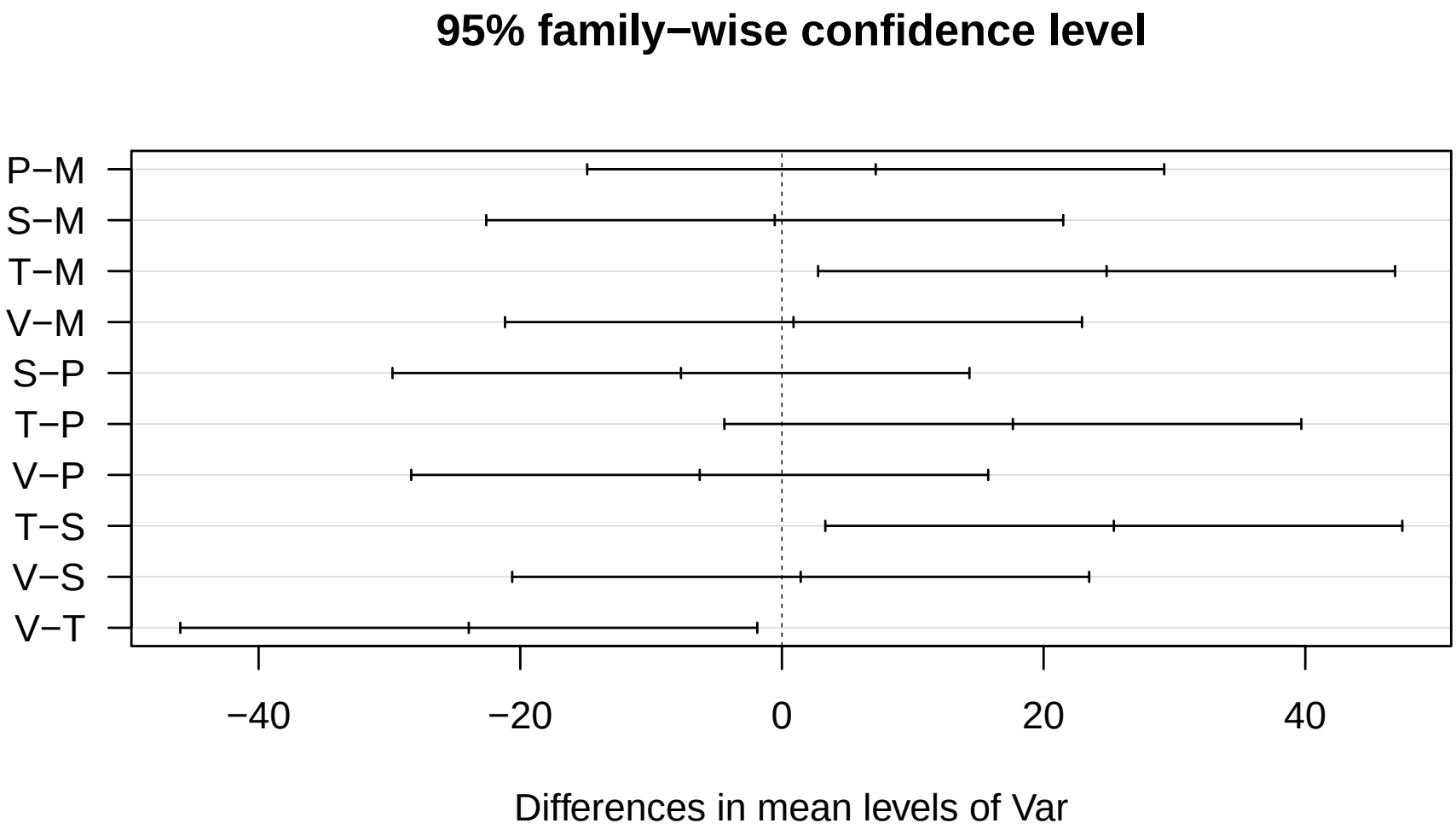
```
m <- aov(Y1 ~ Loc + Var, data = immer)
summary(m)

##             Df Sum Sq Mean Sq F value   Pr(>F)
## Loc          5  17830    3566  21.892 1.75e-07 ***
## Var          4   2757     689   4.231   0.0121 *
## Residuals   20   3258     163
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Vi kan fortsätta med post hoc-tester. Vi behöver då specificera vilken variabel vi vill göra testerna för.

Kornskörd: vilka sorter skiljer sig åt?

```
plot(TukeyHSD(m, which = "Var"))   # which = Namn av variabel
```



Art och kön

Antag att vi istället är intresserade av att jämföra medelvikten för adeliépingviner med medelvikten hos hakremspingviner. Vi tar ut datamaterialet med dessa:

```
penguins2 <- subset(penguins, species != "Gentoo")
```

Vi beräknar medelvärden grupperat på art och kön:

```
aggregate(body_mass_g ~ species, data = penguins2, FUN = mean)
```

```
##     species body_mass_g
## 1    Adelie    3700.662
## 2 Chinstrap    3733.088
```

Ett t-test ger p-värdet 0.59. Det verkar inte finnas någon skillnad i vikt mellan arterna.

Art och kön

Vi beräknar också medelvärden grupperat på art och kön:

```
aggregate(body_mass_g ~ species + sex, data = penguins2,
          FUN = mean)

##     species    sex body_mass_g
## 1    Adelie female    3368.836
## 2 Chinstrap female    3527.206
## 3    Adelie   male    4043.493
## 4 Chinstrap   male    3938.971
```

Hakremshonan väger mer än adeliéhonan. Men hakremshanen väger mindre än adeliéhanen. Det finns ett samspel!

t-Test uppdelat per kön

```
t.test(body_mass_g ~ species,
       data = penguins2[penguins2$sex == "male", ])
```

```
t.test(body_mass_g ~ species,
       data = penguins2[penguins2$sex == "female", ])
```

- Ett t-test för skillnad i vikt mellan adeliéhanar och hakremshanar ger p-värdet 0.1639.
- Ett t-test för skillnad i vikt mellan adeliéhonor och hakremshonor ger p-värdet 0.0085.

Samspel/interaktioner

Vad ska vi tro egentligen? Finns det en skillnad mellan arterna eller inte?

- Det vi ser här är en **samspelseffekt**, även kallat **interaktion**, mellan två variabler.
- Pingvinens vikt påverkas dels av vilken art den tillhör och dels av vilket kön den har.
- Dessutom spelar kombinationen av dessa roll - effekten av arttillhörigheten skiljer sig åt mellan könen!

Modell med samspel

$$y_i = \mu_j + \tau_k + v_{jk} + \epsilon_i$$

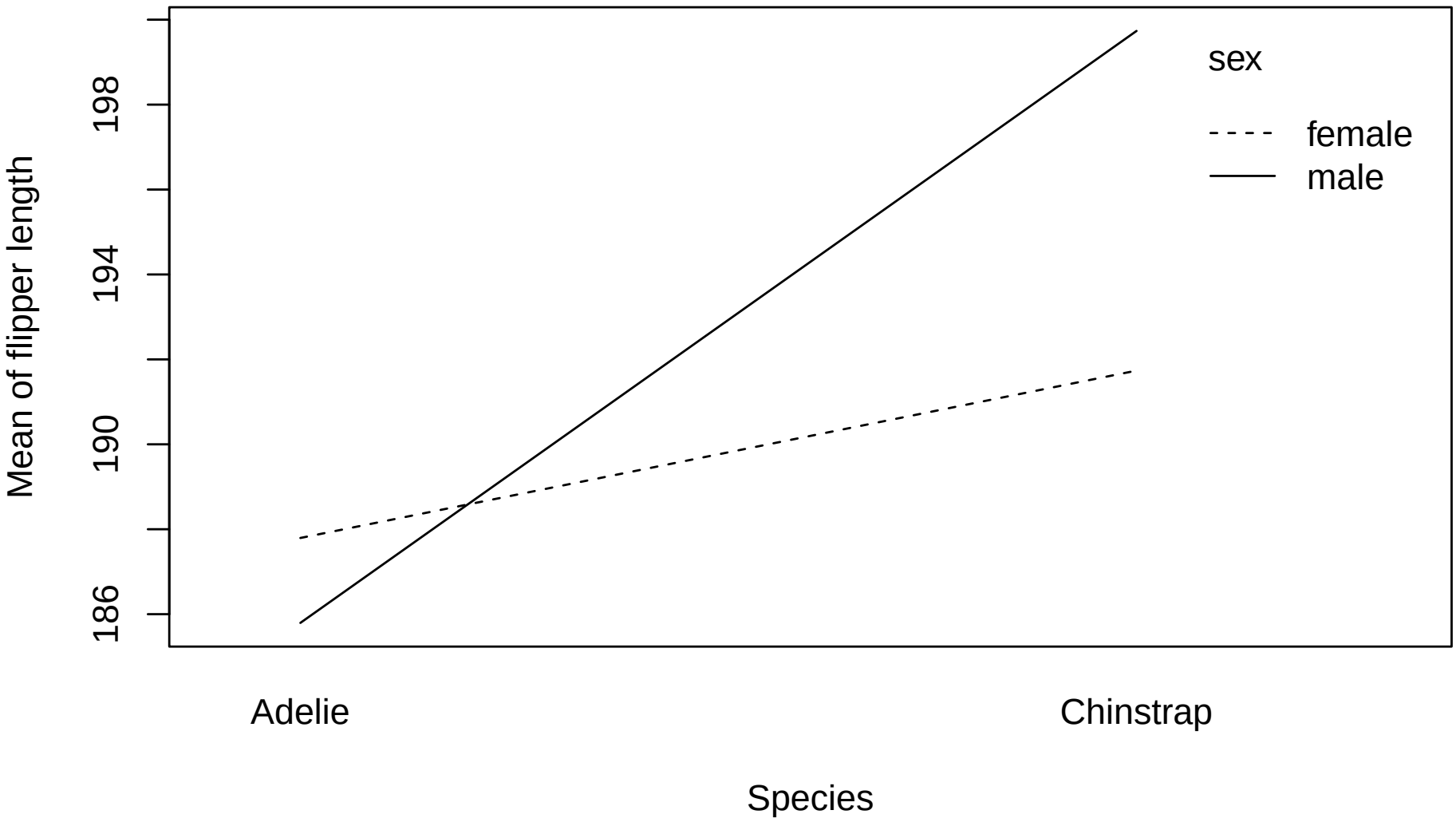
- y_i är observationen i prov i (t.ex. vikten för individ i)
- μ_j är medelvärdet för grupp j av en faktor som provet hör till (t.ex. medelvikten för de två könen, $j = 1, 2$)
- τ_k , är medelvärdet för grupp k av en annat faktor som provet hör till (t.ex. medelvikten på de tre platserna, $k = 1, 2, 3$)
- v_{jk}, är samspelseffekten för faktorerna (t.ex. kön och art)
- ϵ_i är hur mycket observationen i provet avviker från genomsnittet för stammen

Visualisering av samspelseffekt

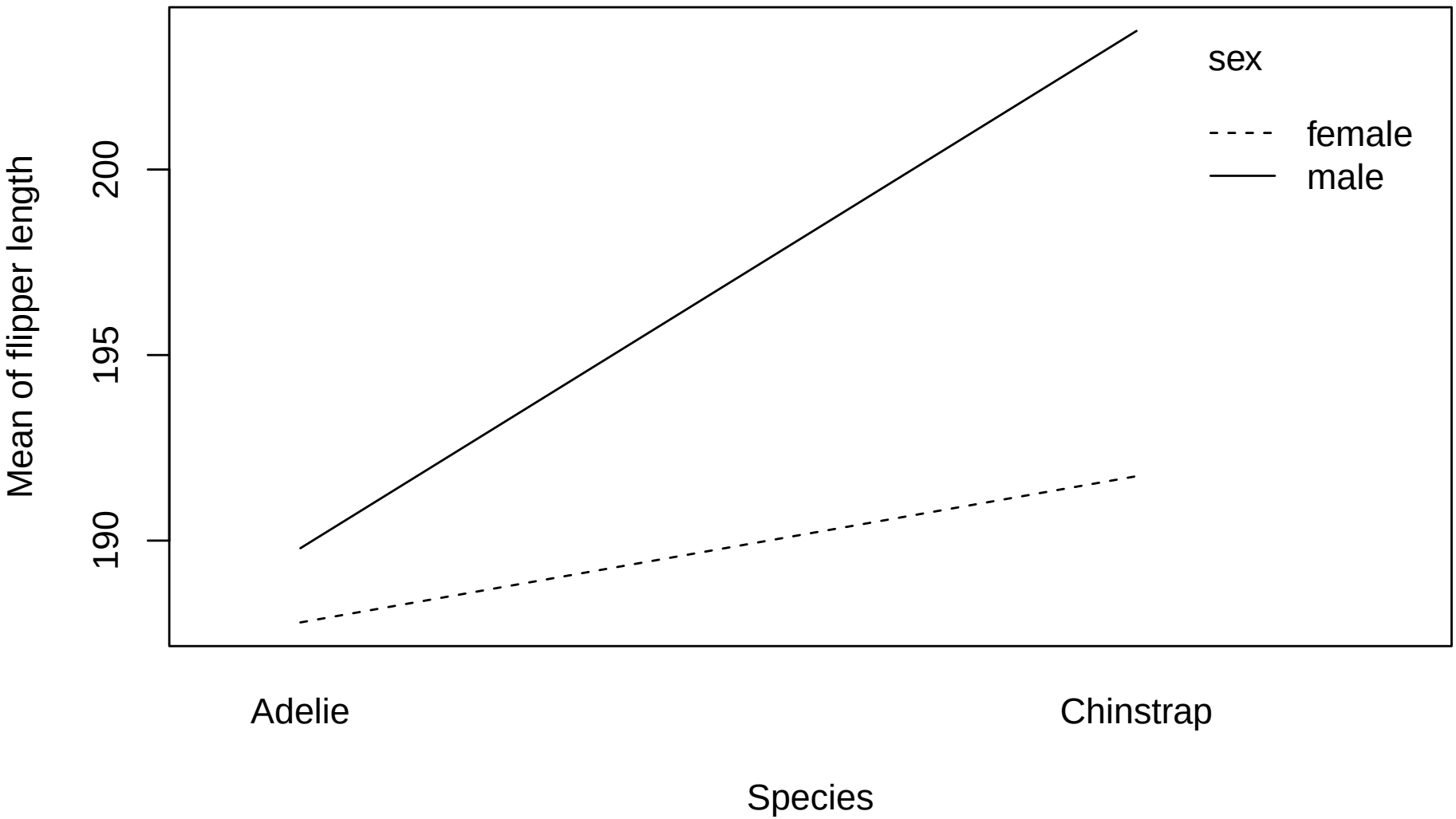
Samspelseffekter brukar visualiseras med en samspelsplot med funktionen interaction.plot().

- Parallella linjer (eller nästan parallella linjer) tyder på att det inte finns en samspelseffekt.
- Icke-parallella linjer tyder på att det finns en samspelseffekt.

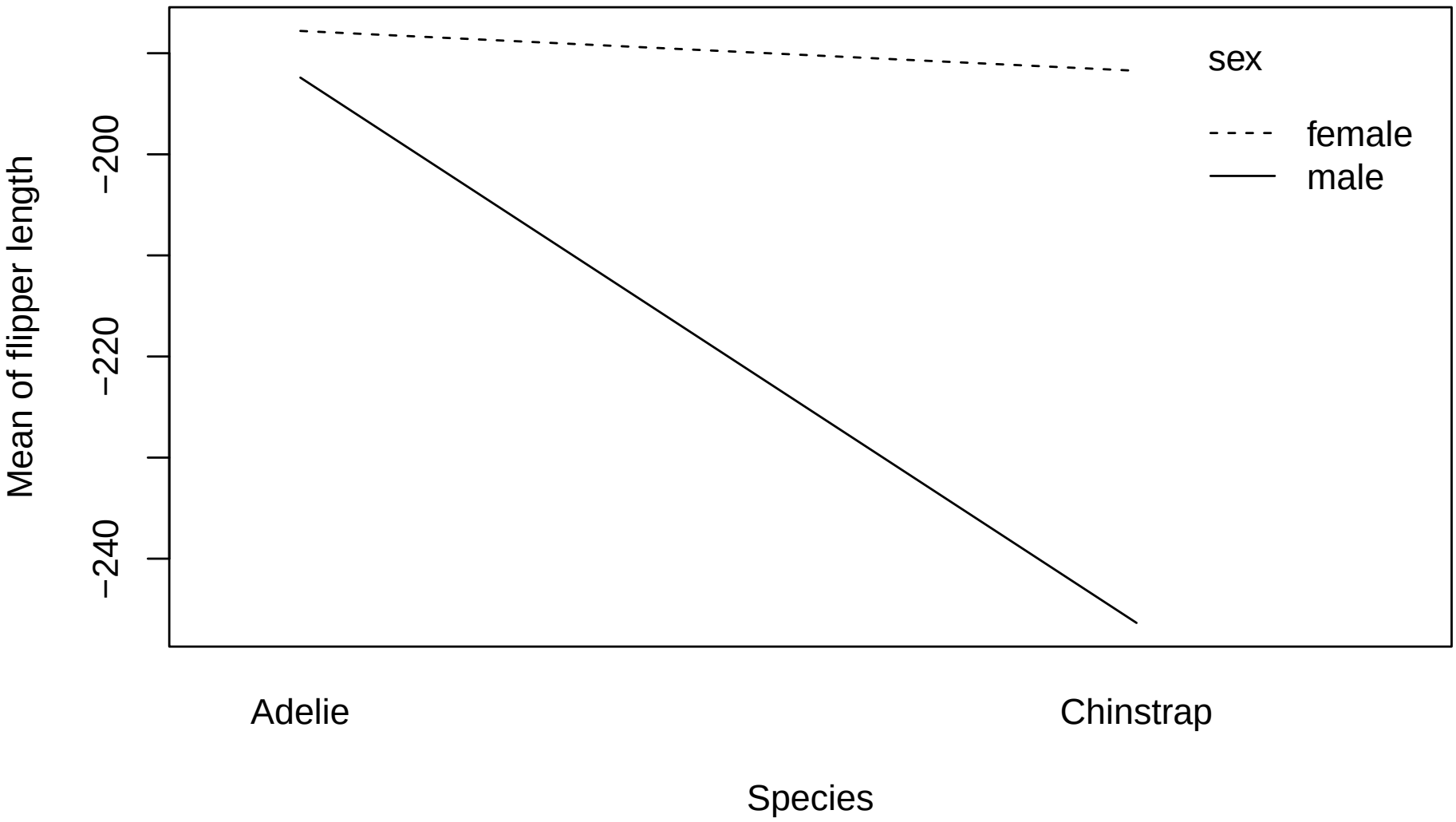
Här finns ett samspel



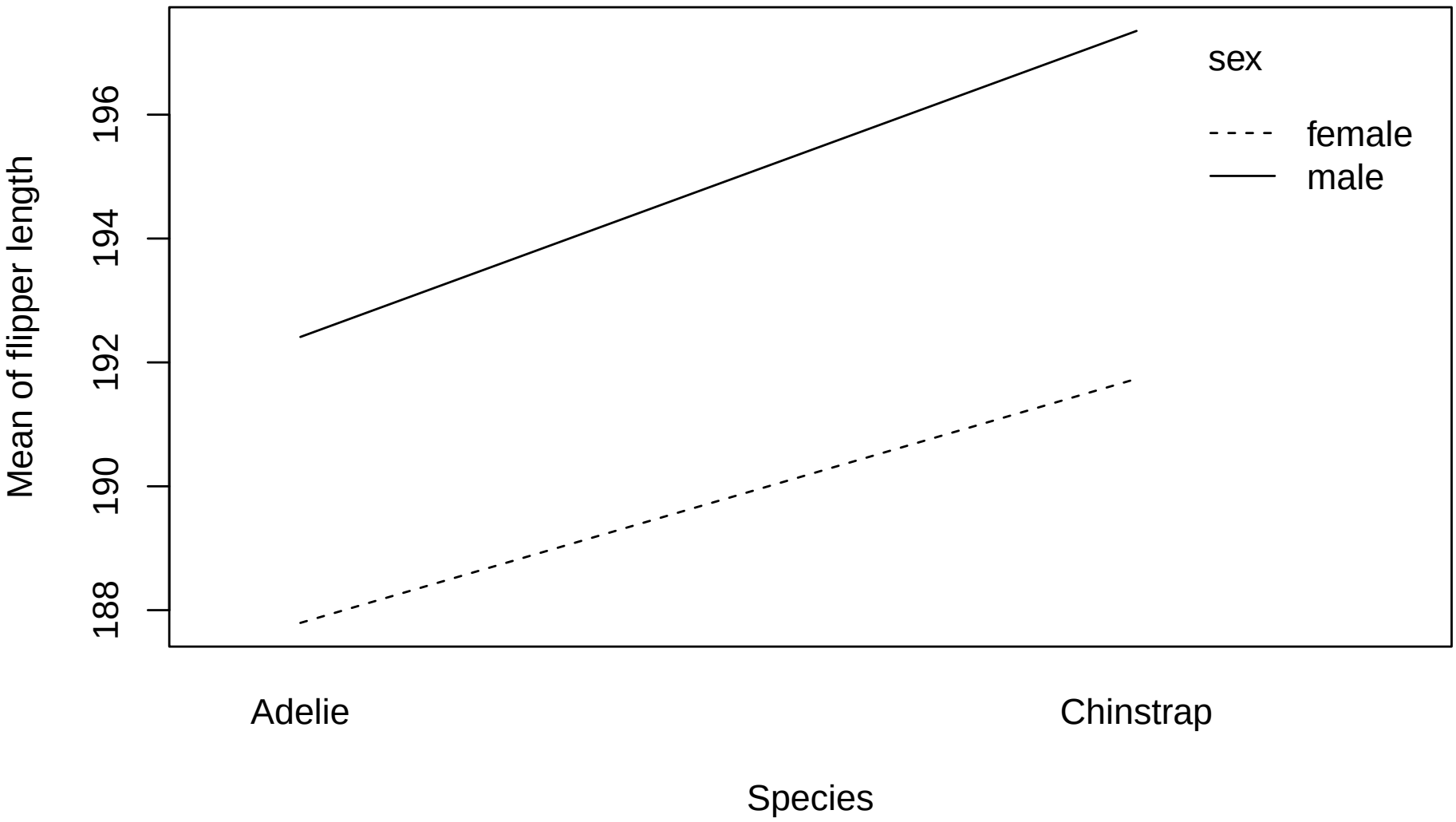
Här finns ett samspel



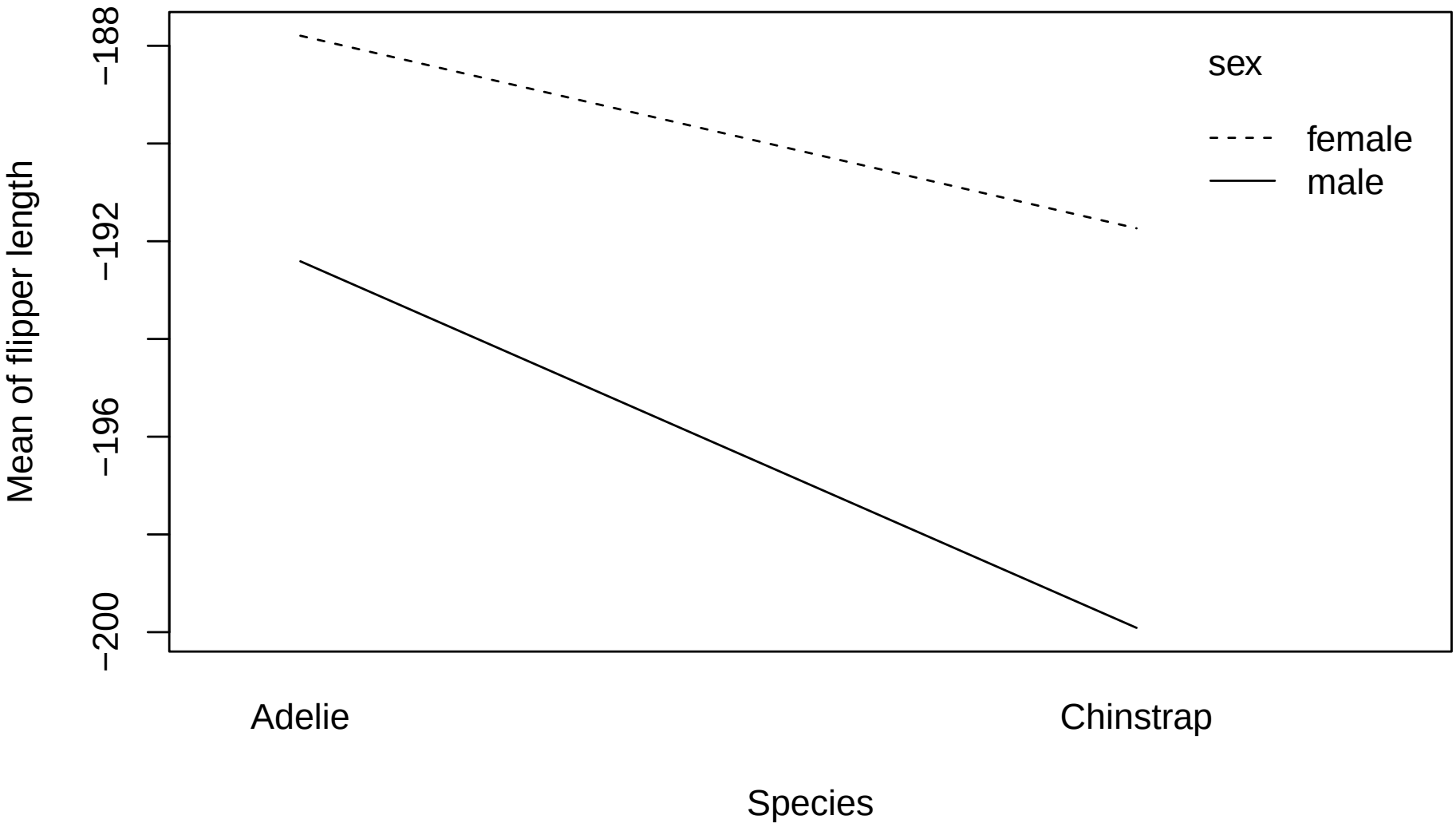
Här finns ett samspel



Här finns inget tydligt samspel

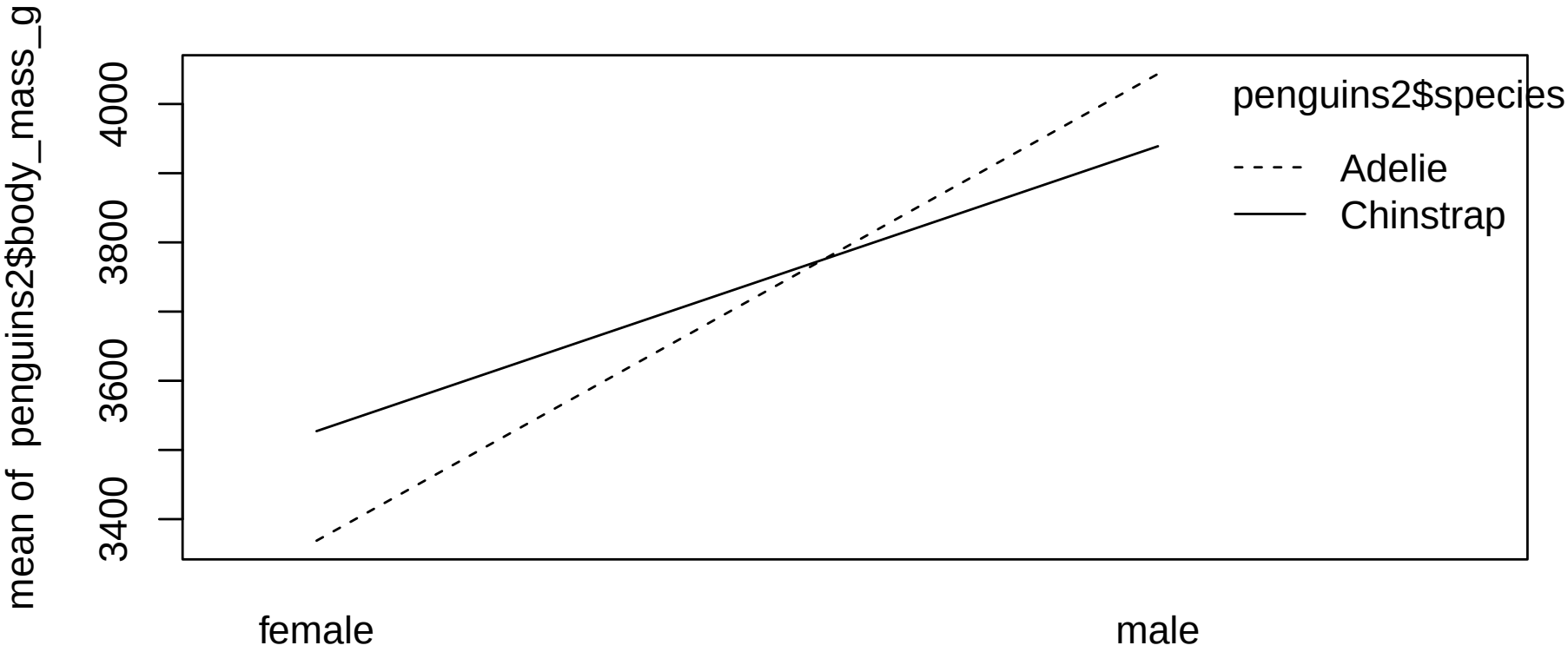


Här finns inget tydligt samspel



Samspelsplot för pingvinvikter

```
interaction.plot(penguins2$sex, penguins2$species,
                 response = penguins2$body_mass_g)
```



ANOVA med samspelseffekt

Om vi vill ha en samspelseffekt i vår ANOVA-modell ersätter vi bara tecknet + med * i formeln:

```
m <- aov(body_mass_g ~ species * sex, data = penguins2)
summary(m)

##                Df   Sum Sq  Mean Sq F value  Pr(>F)
## species         1    33630    33630   0.338 0.56166
## sex             1 18694217 18694217 187.844 < 2e-16 ***
## species:sex     1   801577   801577   8.054 0.00498 **
## Residuals     210 20899209    99520
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
## 6 observations deleted due to missingness
```

Tolkning av resultaten

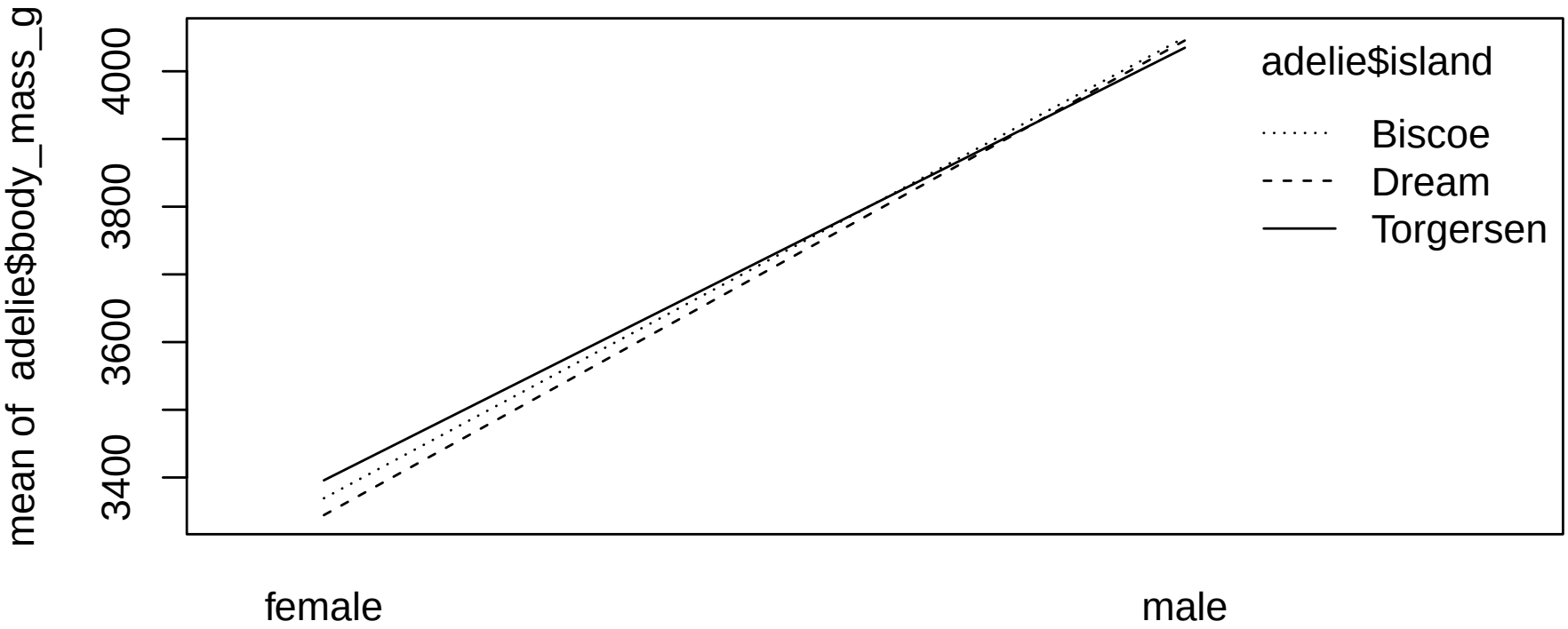
1. Arternas medelvikt skiljer sig inte åt.
2. Medelvikten hos hanar och honor skiljer sig åt.
3. Det finns en signifikant samspelseffekt, vilket visar att skillnaden mellan arterna ser olika ut för honor och hanar.

I många studier är samspelseffekter det intressantaste!

1. Svarar män och kvinnor likadant på ett läkemedel?
2. Påverkar kombinationen av genotyper för gen A och gen B köldtålighet hos gräs?
   - Genotyperna kanske inte har någon inverkan var och en för sig, men tillsammans ger de ökad köldtålighet.

Finns något samspel mellan kön och boplats?

```
interaction.plot(adelie$sex, adelie$island,
                 response = adelie$body_mass_g)
```



ANOVA med samspelseffekt: kön och boplats

```
m <- aov(body_mass_g ~ island * sex, data = adelie)
summary(m)

##                Df   Sum Sq  Mean Sq F value Pr(>F)
## island          2     2064     1032   0.010  0.990
## sex             1 16622756 16622756 168.047 <2e-16 ***
## island:sex      2    24976    12488   0.126  0.881
## Residuals     140 13848406    98917
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
## 6 observations deleted due to missingness
```

Varför försöksplanering?

Det finns tre sorters störande faktorer:

1. Känd och kontrollerbar
   - Exempel: Vilken åker en gröda odlas på. Vilken art en pingvin tillhör. Vilket behanding en patient får.
   - Vi kan välja var vi odlar grödorna. Vi kan välja hur många pingviner från varje art vi undersöker. Vi kan välja behandlingen.
   - Vi planerar vårt försök för att minimera effekten av den störande faktorn, och tar med den störande faktorn som en variabel i vår ANOVA.
2. Känd men inte kontrollerbar
   - Exempel: Hur mycket regn en gröda får.
   - Vid odling utomhus kan vi inte styra hur mycket regn som kommer.
   - Vi kan studera effekten av mängden regn med regression!
3. Okänd och inte kontrollerbar

Randomisering

För att minimera effekten av okända och icke-kontrollerbara störande faktorer använder vi oss av randomisering i vår försöksplanering.

- Slumpa i vilka ordning vi samlar in olika mätningar
  - Pingvinvägning: vågens tillförlitlighet kanske förändras över tid. Det vore dumt att först väga alla adeliépinginver och sedan alla hakremspingviner.
- Slumpa vilken experimentenhet som får vilken behandling
  - Läkemedelsstudie: genom att slumpa ut vilken behandling en patient får kan vi förhoppningsvis undvika systematiska skillnader mellan de patienter som får olika behandlingar.

Blockförsök

Vi ska undersöka hur stor skörden blir för 5 olika kornsorter. Vi har 6 olika åkrar att odla dem på.

- Dålig idé: odla sort 1 på åker 1, sort 2 på åker 2, . . . , sort 5 på åker 5.
- Då kan vi inte veta vad som är skillnader mellan sorterna och vad som är skillnader mellan åkrarna!
- Bättre idé: dela in varje åker i 5 rutor. Odla varje sort på varje åker.
- Då kan vi veta vad som är skillnader mellan sorterna och vad som är skillnader mellan åkrarna!

Blockförsök

Blockning eller **blockförsök** används för att hantera kända och kontrollerbara störande faktorer. Vi ser till att få en fördelning mellan blocken som är "bättre" än den vi hade fått om vi slumpat ut blocken.

- Pingvinvägning: gör lika många mätningar vid varje boplats.
- Läkemedelsstudie: låt lika många män och kvinnor hamna i behandlingsgruppen (men slumpa ut vilka män och kvinnor som hamnar där).

Faktorförsök

I många studier vill vi undersöka effekten av flera variabler. Låt oss säga att vi har k variabler, som alla har två kategorier vardera.

Läkemedelsstudie:

1. Variabel 1: man/kvinna
2. Variabel 2: rökare/icke-rökare
3. Variabel 3: behandling/placebo

I ett **faktorförsök** utför vi vårt experiment för samtliga kombinationer av dessa. I det här fallet får vi 2^k kombinationer (2^k factorial design). Faktorförsök låter oss upptäcka samspelseffekter mellan variablerna. Att titta på kombinationer är mer effektivt än att studera variablerna en och en.

Flervägs-ANOVA

På samma sätt som vi kunde utvidga ANOVA till att hantera två variabler kan vi utvidga det till att hantera ännu fler variabler. En **flervägs-ANOVA** (multiway ANOVA) fungerar på samma sätt som tidigare i R:

```
aov(y ~ x1 + x2 + x3 + x4, data = data)
```

- Vi riskerar att få många och komplicera samspelseffekter när vi har många variabler.
- Vi behöver ofta välja om vi vill ha med samspelseffekter av högre ordning (t.ex. samspel mellan fyra variabler) i modellen eftersom dessa kräver mycket data för att kunna undersökas.

Slumpeffekter

I en del studier representerar en eller flera av de variabler vi behöver ta med i vår ANOVA slumpeffekter.

- Exempel: vi följer en grupp patienter över tid, med upprepade mätningar av blodtryck.
- Vi bör ta med patient-ID i vår ANOVA, eftersom det är rimligt att tro att det finns skillnader mellan patienterna.
- Men vi är inte intresserade av just de här patienterna! De är bara några personer som vi "råkat" ta med i studien.
- Skillnaderna mellan individerna är därmed "slumpmässiga" och patient-ID beskriver en slumpeffekt.

Jämför detta med variabeln rökare/icke-rökare. Där är vi intresserade av just de två värdena. Variabeln beskriver därför en fix effekt. Studier där slumpeffekt ingås analyseras bäst med mixade modeller (finns för både regression och ANOVA).

Sammanfattning

1. ANOVA kan användas för att hantera störande faktorer
2. I planerade försök ser vi till att designa försöket så att vi minimerar effekten av de störande faktorerna
3. ANOVA låter oss undersöka samspeleffekter mellan variabler
4. Ofta är dessa det mest intressanta i hela studien!
5. Kan visualiseras med samspelsplottar