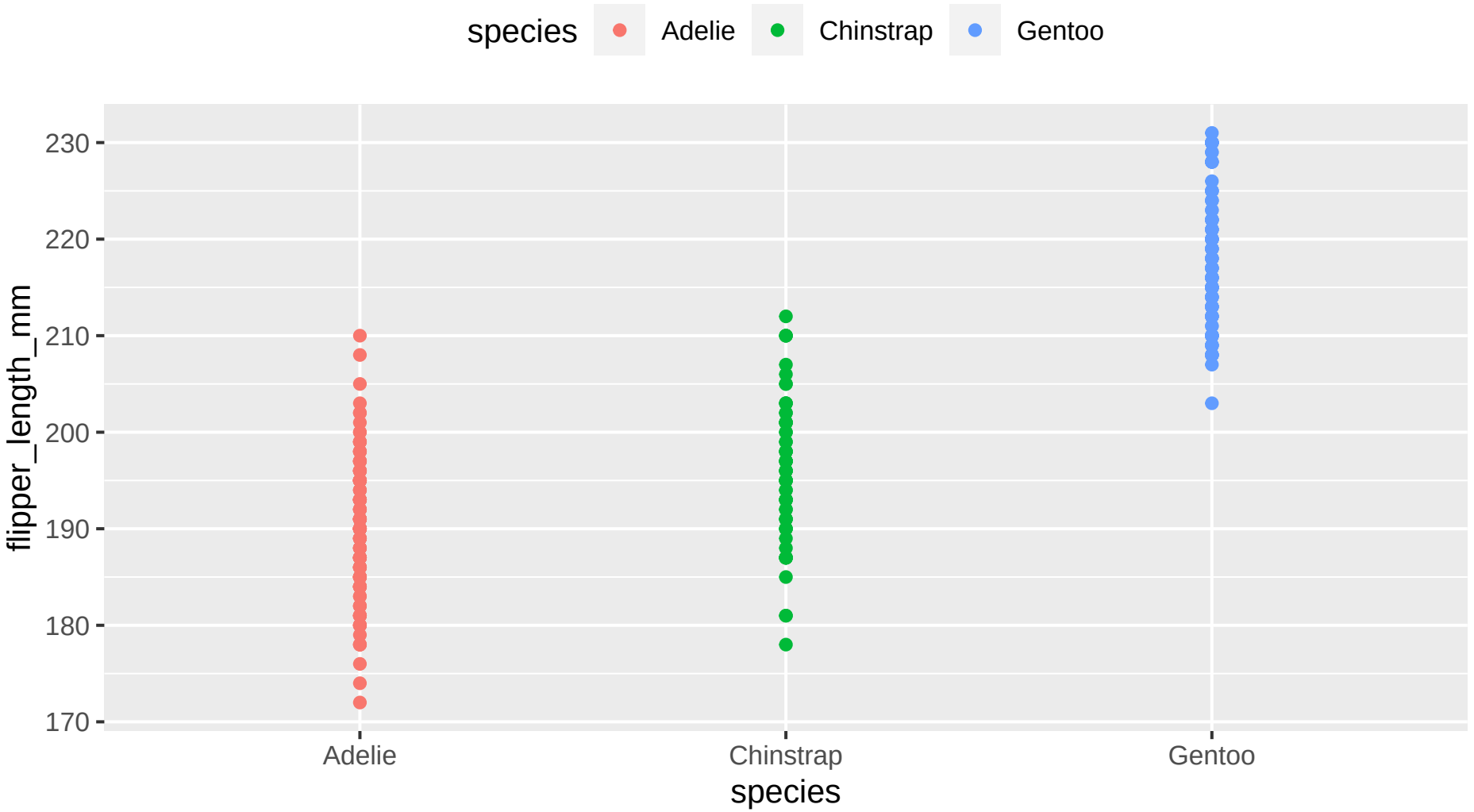
Statistik för Biologer
F8: ANOVA

Shaobo Jin

Matematiska institutionen

Jämförelse av grupper

I många studier är vi intresserade av jämförelse av två eller flera grupper:



Jämförelse av grupper: parvis jämförelse

För att jämföra två grupper kan vi använda t-testet (Föreläsning 4) eller rangmetoder (Föreläsning 7). Nu vill vi jämföra alla tre grupper med varandra.

1. Grupp 1 ska jämföras med 2 andra: grupp 2, grupp 3,
2. Grupp 2 ska slutligen jämföras med 1 annan: grupp 3

Vi skulle kunna göra jämförelsen med $2 + 1 = 3$ test.

Typ I-Fel Vid Hypotesprövning

	Sanning	
Vår slutsats	H_0 är rätt	H_1 är rätt
H_0 är rätt		**Typ II-fel**
H_1 är rätt	**Typ I-fel**	

- Att felaktigt förkasta H_0 trots att H_0 är sann kallas för ett **typ I-fel** (ett falskt positivt resultat)
- Sannolikhetet för ett typ I-fel när vi gör ett test är lika med signifikansnivån α.
- För vart och ett av de 3 testen är sannolikheten för ett typ I fel alltså lika med α.
- Vad blir sannolikheten för minst ett typ I-fel?

Antal fel vid hypotesprövning: parvis jämförelse

Vi väljer $\alpha = 0.05$. Vi antar förenklat att

1. testen är oberoende.
2. H_0 stämmer för alla 3 test.

Då är X =antal typ I-fel binomialfördelad med $n = 3$ och $p = 0.05$. Sannolikheten för minst ett typ I-fel (minst ett falskt positivt resultat) blir då:

$$P(X \geq 1) \quad = \quad 1 - P(X = 0)$$

```
1 - dbinom(0, 3, 0.05)

## [1] 0.142625
```

Antal fel vid hypotesprövning: parvis jämförelse

Vi vill jämföra k grupper och väljer $\alpha = 0.05$. Vi antar förenklat att

1. testen är oberoende.
2. H_0 stämmer för alla test.

Då är X =antal typ I-fel binomialfördelad med $n = k\left(k-1\right)/2$ och $p = 0.05$. Sannolikheten för minst ett typ I-fel (minst ett falskt positivt resultat) blir då:

$$P\left(X \geq 1\right) \quad = \quad 1 - P\left(X = 0\right)$$

Om vi har $k = 10$ grupper

```
# k = 10
1 - dbinom(0, 10 * (10 - 1) / 2, 0.05)

## [1] 0.9005597
```

Vad kan vi göra?

Vad ska vi då göra om vi vill testa många olika hypoteser?

1. Alternativ 1: **ANOVA** (**An**alysis **o**f **Va**riance, variansanalys). Börja med att testa en "global hypotes": finns det minst en grupp som avviker från övriga?
2. Alternativ 2: Justera p-värdet eller signifikansnivån för att minska risken för falskt positiva resultat
   - Variant 1: ändra signifikansnivån α "lagom mycket" så att risken för falskt positiva resultat minskar
   - Variant 2: skala om p-värdet beroende på antal test, men jämför med ursprungliga α

Övergripande hypoteser

Vi vill jämföra 3 eller flera grupper med varandra. Vi kan då behöva göra väldigt många test, vilket ökar risken för att råka ut för typ I-fel. **ANOVA** börjar med att testa följande hypoteser:

H_0 : alla grupper har samma medelvärde

H_1 : minst en grupp har ett avvikande medelvärde

Om det här första testet inte ger ett signifikant resultat så är vi klara! Annars kan vi gå vidare med parvisa jämförelser av de olika grupperna.

Vad är ANOVA? Perspektiv 1

Variationen i våra data beror på olika saker:

1. Skillnader mellan grupperna.
2. Skillnader som beror på övriga faktorer (slumpavvikelser, mätfel, osv)

Idén bakom ANOVA är att räkna ut ett antal kvadratsummor som beskriver variationen i data, i stil med denna:

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

Kvadratsummor för variation

Vi beräknar tre sådana kvadratsummor (**sum of squares**, **SS**):

1. SSE: variation som beror på slumpavvikelser/mätfel inom grupp
2. SST: variation som beror på skillnaden mellan grupperna (treatments, behandlingar)
3. Total SS: SSE + SST

Vi kan räkna om dessa till hur stora de genomsnittliga avvikelserna är:

1. MSE: genomsnittlig variation mellan två mätningar i samma grupp
2. MST: genomsnittlig variation mellan två mätningar från olika grupper

Hypotesprövning

Om H_0 stämmer och grupperna inte skiljer sig åt så borde MSE (variationen mellan mätningar som hör till samma grupp) vara ungefär lika stor som eller större än MST (variationen mellan mätningar som tillhör olika grupper).

Det vill säga, vi borde få att

$$\frac{MST}{MSE} \text{ är inte så stor.}$$

Om H_0 stämmer följer kvoten $F = \frac{MST}{MSE}$ den så kallade **F-fördelningen**. Vi kan därför räkna ut sannolikheten att få en F-kvot som är minst så stor som den observerade, vilket ger oss p-värdet för testet.

Att testa skillnader i R

Att utföra ett ANOVA-test går till på samma sätt som att anpassa en linjär regressionsmodell, fast med funktionen aov() istället för lm():

```
ANOVA <- aov(flipper_length_mm ~ species, data = penguins)
summary(ANOVA)

##              Df Sum Sq Mean Sq F value Pr(>F)
## species       2  52473   26237   594.8 <2e-16 ***
## Residuals   339  14953      44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
## 2 observations deleted due to missingness
```

Vad är ANOVA? Perspektiv 2: ANOVA och regression

ANOVA och **linjär regression** är i grund och botten samma problem!

- Regression: påverkas y av den numeriska variabeln x?
- ANOVA: påverkas y av den kategoriska variabeln x?

Variationen i våra data beror på olika saker:

- Skillnader mellan grupperna motsvarar effekten av x i en regressionsmodell
- Skillnader som beror på övriga faktorer (slumpavvikelser, mätfel, osv) motsvarar ϵ i en regressionsmodell

Vad är ANOVA? Perspektiv 2: ANOVA och regression

Vi kan skriva vår ANOVA-modell med en formel som liknar formeln vi haft vid linjär regression:

$$y_i = \mu_j + \epsilon_i$$

- y_i är observationen i prov i
- μ_j är medelvärdet för grupp j som provet hör till
- ϵ_i är hur mycket observationen i provet avviker från genomsnittet för stammen

ANOVA som Regression

```
LM <- lm(flipper_length_mm ~ species, data = penguins); summary(LM)

##
## Call:
## lm(formula = flipper_length_mm ~ species, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.9536  -4.8235   0.0464   4.8130  20.0464
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      189.9536     0.5405 351.454  < 2e-16 ***
## speciesChinstrap   5.8699     0.9699   6.052 3.79e-09 ***
## speciesGentoo     27.2333     0.8067  33.760  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.642 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.7782,Adjusted R-squared:  0.7769
## F-statistic: 594.8 on 2 and 339 DF,  p-value: < 2.2e-16
```

ANOVA Eller T-Test vid Jämförelse av Två Grupper

```
ANOVA <- aov(flipper_length_mm ~ sex, data = penguins)
summary(ANOVA)

##              Df Sum Sq Mean Sq F value   Pr(>F)
## sex           1   4246    4246   23.05 2.39e-06 ***
## Residuals   331  60972     184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 11 observations deleted due to missingness

summary(ANOVA)[[1]][["F value"]][1]

## [1] 23.05279
```

ANOVA Eller T-Test vid Jämförelse av Två Grupper

```
Ttest <- t.test(flipper_length_mm ~ sex, data = penguins, var.equal = TRUE)
Ttest

##
##  Two Sample t-test
##
## data:  flipper_length_mm by sex
## t = -4.8013, df = 331, p-value = 2.391e-06
## alternative hypothesis: true difference in means between group female an
## 95 percent confidence interval:
##  -10.068599  -4.216033
## sample estimates:
## mean in group female   mean in group male
##             197.3636             204.5060

Ttest$statistic ^ 2

##        t
## 23.05279
```

Förutsättningar

ANOVA-test bygger på några villkor:

1. Observationerna ska vara oberoende av varandra
2. Observationerna i varje grupp ska vara normalfördelade (kan kringås via bootstrap)
3. Alla grupper ska ha samma varians

Bootstrapvariant

ANOVA fungerar sämre om vi inte har normalfördelade data, särskilt om gruppstorlekarna är obalanserade. I sådana situationer kan vi istället använda bootstrap för att beräkna p-värdet:

```
library(lmboot)
ANOVA.boot(flipper_length_mm ~ species, data = penguins)

## Error in ANOVA.boot(flipper_length_mm ~ species, data =
penguins):  Response must not have any missing values.
```

```
ANOVA <- ANOVA.boot(flipper_length_mm ~ species,
          data = penguins[-c(4, 272), ])
ANOVA[["p-values"]]

## [1] 0
```

Men var finns skillnaden?

Vi har nu sett att det finns minst en grupp som avviker från de övriga. Men vilken är det som avviker?

Eftersom vi ofta vill utföra flera ("multipla") hypotestester så finns ett stort antal metoder för att justera för multiplicitet:

1. Benjamini-Hochberg
2. Tukeys HSD
3. Bonferroni
4. Holm

Scand J Statist 6: 65–70, 1979

A Simple Sequentially Rejective Multiple Test Procedure

STURE HOLM

Chalmers University of Technology, Göteborg

Vad ska vi justera? FWER

De olika metoderna justerar p-värden/signifikansnivåer utifrån olika kriterium.

Bonferroni, Holm och Tukeys HSD: bygger på sannolikheten att få minst ett typ I-fel (family-wise error rate, FWER)

- Används när vi är måna om att undvika falskt positiva resultat
- Blir väldigt konservativt om vi har många tester - styrkan blir låg och vi får svårt att upptäcka de skillnader som finns på riktigt
- Vanligt vid ANOVA

Vad ska vi justera? FDR

De olika metoderna justerar p-värden/signifikansnivåer utifrån olika kriterium.

Benjamini-Hochberg: bygger på den förväntade andelen falska positiva resultat bland alla positiva resultat (false discovery rate, FDR)

- Ger högre styrka än Bonferroni/Holm - vi får större chans att upptäcka de skillnader som faktiskt finns
- Används när vi är beredda att acceptera en del falskt positiva resultat i utbyte mot att vi upptäcker fler av de skillnader som finns på riktigt

Post hoc-analys

Om ANOVA-testet är signifikant kan vi gå vidare och göra ett **post hoc-test** för att se vilka grupper som skiljer sig åt.

- Vanligtvis används Tukeys HSD för detta.
- I princip gör vi t-test där p-värdet justeras beroende på hur många grupper vi vill jämföra.
- De p-värden vi får fram kan sedan jämföras med signifikansnivån α.

Dels kan vi få p-värden för varje parvis jämförelse, och dels kan vi visualisera vilka grupper som skiljer sig åt med hjälp av de tillhörande konfidensintervallen.

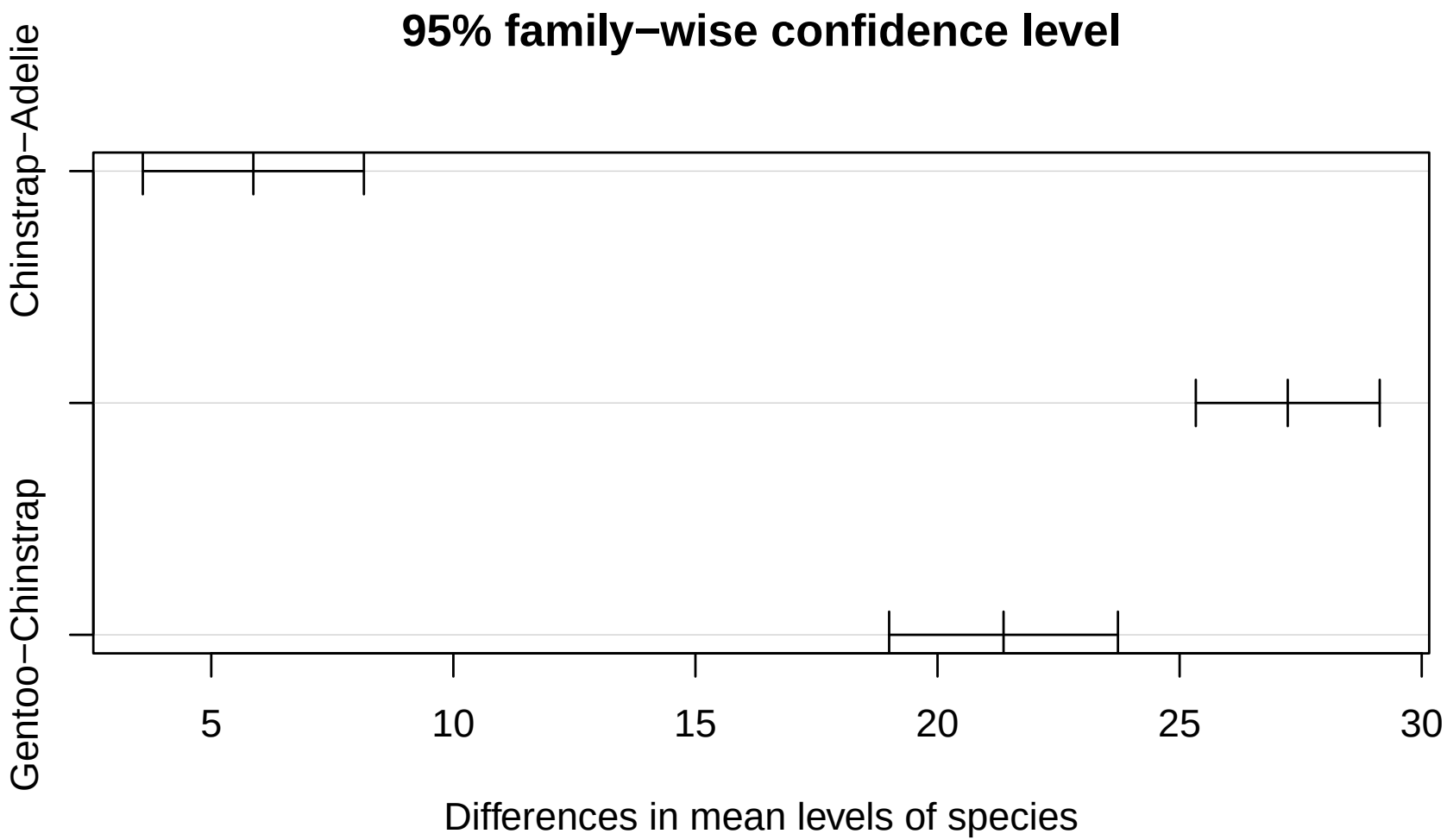
Tukeys HSD i R

```
ANOVA <- aov(flipper_length_mm ~ species, data = penguins)
TukeyHSD(ANOVA)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = flipper_length_mm ~ species, data = pengu
##
## $species
##                         diff       lwr       upr p adj
## Chinstrap-Adelie  5.869887  3.586583  8.153191     0
## Gentoo-Adelie    27.233349 25.334376 29.132323     0
## Gentoo-Chinstrap 21.363462 19.000841 23.726084     0
```

Visualisering av skillnader

```
plot(TukeyHSD(ANOVA))
```



Andra justeringarr

Om vi vill använda någon annan metod än Tukeys HSD så kan vi utföra de parvisa testen manuellt (t.ex. med bootstrap-t-test eller Wilcoxon-Mann-Whitney) och spara p-värdena i en vektor. Funktionen p.adjust kan sedan användas för att få justerade p-värden:

```
pvalue <- c(0.812, 0.012, 0.008, 0.122, 0.134, 0.489, 0.001)
p.adjust(pvalue, method = "holm")

## [1] 0.978 0.060 0.048 0.488 0.488 0.978 0.007
```

Detta går även att göra i situationer där vi inte vill använda just ANOVA, exempelvis om vi anpassar ett stort antal regressionsmodeller.

Art och kön

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

```
penguins2 <- subset(penguins, species != "Gentoo")
t.test(body_mass_g ~ species, data = penguins2)

##
##  Welch Two Sample t-test
##
## data:  body_mass_g by species
## t = -0.54309, df = 152.45, p-value = 0.5879
## alternative hypothesis: true difference in means between group Ad
## 95 percent confidence interval:
##  -150.38481   85.53284
## sample estimates:
##    mean in group Adelie mean in group Chinstrap
##                3700.662                3733.088
```

t-Test uppdelat per kön

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

##
##  Welch Two Sample t-test
##
## data:  body_mass_g by species
## t = 1.4088, df = 62.027, p-value = 0.1639
## alternative hypothesis: true difference in means between gro
## 95 percent confidence interval:
##  -43.78869 252.83381
## sample estimates:
##    mean in group Adelie mean in group Chinstrap
##                 4043.493                 3938.971
```

t-Test uppdelat per kön

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

##
##  Welch Two Sample t-test
##
## data:  body_mass_g by species
## t = -2.7206, df = 61.248, p-value = 0.008471
## alternative hypothesis: true difference in means between gro
## 95 percent confidence interval:
##  -274.76257  -41.97796
## sample estimates:
##    mean in group Adelie mean in group Chinstrap
##                3368.836                3527.206
```

Simpsons Paradox

- **Simpsons paradox** innebär att grupper kan ha en viss tendens, men tendensen är omvänd eller osynlig när grupperna sätts ihop.
- En dold variabel (eller en lurande variabel) kan påverka resultatet.
- Man måste ta hänsyn till den dold variabel.
  - Tvåvägs-ANOVA och flervägs-ANOVA

Sammanfattning

1. Typ I-fel: falska upptäckter, falskt positiva resultat
2. Ju fler test vi utför, desto högre blir risken att vi gör minst ett typ I-fel
3. ANOVA: jämförelse av 3+ grupper
   - Börja med att testa om det finns någon grupp som avviker
   - Om ja, gå vidare och undersök vilka grupper som skiljer sig åt
   - Justering för multiplicitet
   - Många olika metoder, som passar för olika tillämpningar