Statistical inference and linear models

February 20, 2018

If you get bored

- Go to the last slide for bonus exercises
- Work on code for your research and ask question during exercise time
- But try and keep an eye out, just in case

1. Statistical inference

2. t-test, ANOVA, regression: all is one, one is all

General approach

1. Scientific question

General approach

1. Scientific question

$\downarrow$

2. Model and Statistical question

General approach

1. Scientific question

$\downarrow$

2. Model and Statistical question $\longleftrightarrow$ 3. Data collection

General approach

$$
\begin{array}{c}
\text{1. Scientific question} \\
\downarrow \\
\text{2. Model and Statistical question} \longleftrightarrow \text{3. Data collection} \\
\downarrow \\
\text{4. Estimation} \\
\text{4.b Uncertainty and statistical significance}
\end{array}
$$

General approach

1. Scientific question

2. Model and Statistical question

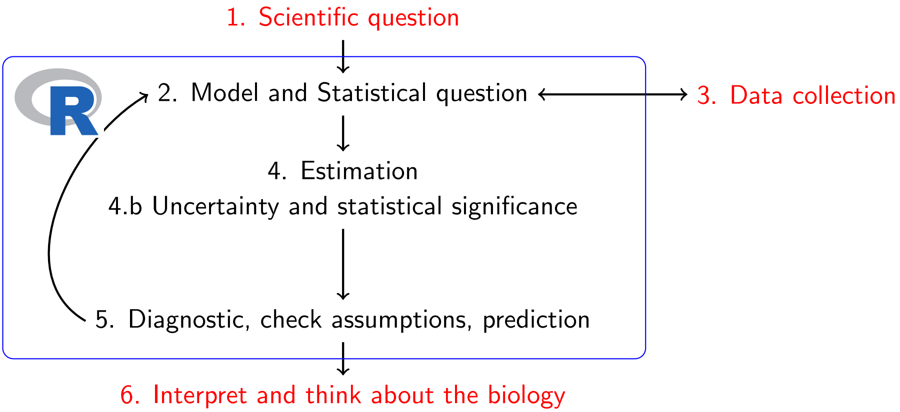
3. Data collection

4. Estimation

4.b Uncertainty and statistical significance

5. Diagnostic, check assumptions, prediction

General approach



Reminder t.test

Reminder t.test

```r
data("iris")
```

Reminder t.test

```
data("iris")
```

One t-test for sepal length between *setosa* and *versicolor*:

```
t.test(x = iris$Sepal.Length[iris$Species == "setosa"],
       y = iris$Sepal.Length[iris$Species == "versicolor"])
```

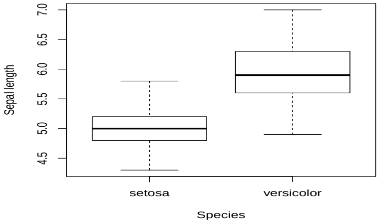
```


Welch Two Sample t-test

data:  iris$Sepal.Length[iris$Species == "setosa"] and iris$Sepal.Le
t = -10.521, df = 86.538, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.1057074 -0.7542926
sample estimates:
mean of x mean of y
    5.006     5.936
```

Reminder t.test: Are they different?

```
boxplot(Sepal.Length ~ Species,
        data = iris[iris$Species %in% c("setosa","versicolor"),],
        drop = TRUE, ylab="Sepal length", xlab="Species")
```



- Means: 5.006 vs. 5.936
- Standard deviation: 0.35 and 0.52
- Standard error (SD/ $\sqrt{n}$): 0.05 and 0.07

When do we know it is different?

Estimation

Estimated difference = $\color{red}{\text{Mean}_1 - \text{Mean}_2}$

When do we know it is different?

Estimation

Estimated difference $= \text{Mean}_1 - \text{Mean}_2$

Quantify uncertainty / Statistical significance

t-statistic unlikely to be *large* when estimated means are equal

$$t = \frac{\text{Mean}_1 - \text{Mean}_2}{\text{Variation}} \frac{\sqrt{\text{Sample Size}}}{\sqrt{2}}$$

1. Larger absolute difference
2. Smaller variability
3. Larger sample size

When do we know it is different?

Estimation

Estimated difference $= \text{Mean}_1 - \text{Mean}_2$

Quantify uncertainty / Statistical significance

t-statistic unlikely to be *large* when estimated means are equal

$$t = \frac{\text{Mean}_1 - \text{Mean}_2}{\text{Variation}} \frac{\sqrt{\text{Sample Size}}}{\sqrt{2}}$$

1. Larger absolute difference
2. Smaller variability
3. Larger sample size

Same for every statistical model

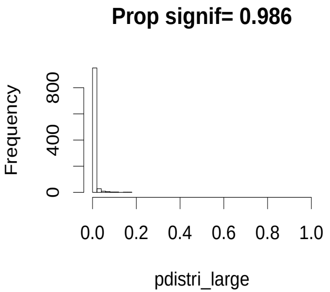
When do we know it is different?

1. Larger absolute difference

```
nbsim <- 1000
pdistri_large <- vector(length = nbsim)
pdistri_small <- vector(length = nbsim)
for (i in 1:nbsim)
  {
  x1 <- rnorm(n = 10, mean = 2, sd = 1)
  x2 <- rnorm(n = 10, mean = 4, sd = 1) #large diff
  x3 <- rnorm(n = 10, mean = 2.5, sd = 1) #small diff
  out_large <- t.test(x1, x2)
  out_small <- t.test(x1, x3)
  pdistri_large[i]<-out_large$p.value
  pdistri_small[i]<-out_small$p.value
}
```

When do we know it is different?

```
par(mfrow=c(1,2), cex=2)
hist(pdistri_large, xlim=c(0,1),
     main=paste("Prop signif=",mean(pdistri_large<0.05)))
hist(pdistri_small, xlim=c(0,1),
     main=paste("Prop signi=",mean(pdistri_small<0.05)))
```




```
par(mfrow=c(1,1))
```

When do we know it is different? Try it!

Exercise

Check the effect of smaller variability and/or larger sample size.

By the way, what are these p-values?

Blabla

Properties

- Reference to a null-model (H_0) with assumptions
- Uniform distribution under H_0 ...
- ... hence proportion(significance under H_0) $=$ significance threshold

T-test exercise

```
t.test(x = ..., y=...., var.equal = TRUE)
t.test(x = ..., y=...., var.equal = FALSE)
```

What if variance are different by chance only?

```
set.seed(1234)
var(rnorm(20, mean = 0, sd = 1))

[1] 1.027806

var(rnorm(20, mean = 0, sd = 1))

[1] 0.6265501
```

Exercise

What option is more correct for var.equal?

A small example

Animal behavior in response to weather
Load data:

```
getwd()
setwd()
```

```
dat.behav <- read.csv(file = "datbehav.csv") # path to file
```

A small example

Animal behavior in response to weather
Load data:

```
getwd()
setwd()
```

```
dat.behav <- read.csv(file = "datbehav.csv") # path to file
```

STEP 1: have a look at your data

```
str(dat.behav)
summary(dat.behav)
plot(dat.behav)
```

t-test

```
fitstudent <- t.test(x = dat.behav$activity[dat.behav$weather==
                                                "rainy"],
                     y = dat.behav$activity[dat.behav$weather==
                                                "sunny"],
                     var.equal = TRUE)
print(fitstudent)


Two Sample t-test

data:  dat.behav$activity[dat.behav$weather == "rainy"] and dat.beha
t = 3.2752, df = 33, p-value = 0.002485
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6138373 2.6270325
sample estimates:
mean of x mean of y
 6.781476  5.161041
```

ANOVA

```
fitanova <- aov(data = dat.behav, formula = activity ~ weather)
summary(fitanova)

            Df Sum Sq Mean Sq F value  Pr(>F)
weather      1  11.25  11.253   10.73 0.00248 **
Residuals   33  34.62   1.049
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear regression

```
fitlm <- lm(data = dat.behav, formula = activity ~ weather)
summary(fitlm)


Call:
lm(formula = activity ~ weather, data = dat.behav)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3547 -0.6028  0.2346  0.6419  1.6534

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.7815     0.4581  14.805 3.94e-16 ***
weathersunny  -1.6204     0.4948  -3.275  0.00248 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.024 on 33 degrees of freedom
```

NB: aov() vs. anova()

```
aov(data = dat.behav, formula = activity ~ weather)
anova(fitlm)
```