Simple statistics using R

Julien Martin
University of Ottawa

learning outcomes

- introduce you to some basic statistics in R ✔️
- focus on linear models ✔️
- fit simple linear models in R ✔️
- check linear model assumptions in R ✔️

statistics using R

- many, many statistical tests available in R
- range from the simple to the highly complex
- many are included in standard base installation of R
- you can extend the range of statistics by installing additional packages

statistics using R

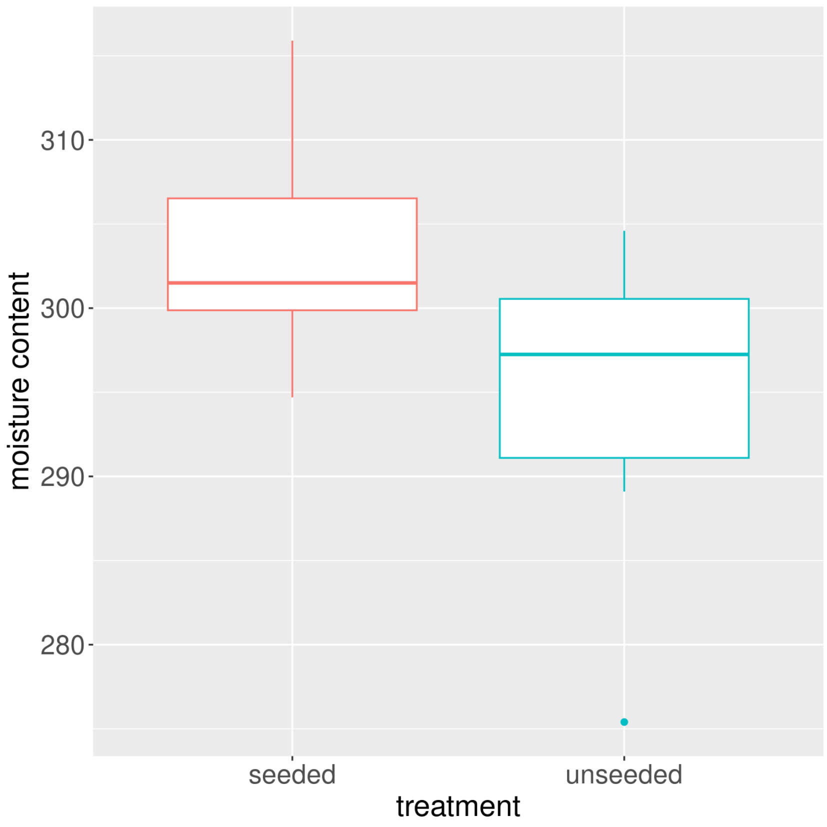
an example

- does seeding clouds with dimethylsulphate alter the moisture content of clouds (can we make it rain!)
- 10 random clouds were seeded and 10 random clouds unseeded
- what’s the null hypothesis?
- no difference in mean moisture content between seeded and unseeded clouds

statistics using R

- plot these data
- interpretation?
- what type of statistical test do you want to use?

```
str(clouds)
## 'data.frame':    20 obs. of  2 variables:
##  $ moisture : num  301 302 299 316 307 ...
##  $ treatment: chr  "seeded" "seeded" "seeded" "s
```



statistics using R

```r
t.test(clouds$moisture~clouds$treatment, var.equal=TRUE)
```

```
## 
##     Two Sample t-test
## 
## data:  clouds$moisture by clouds$treatment
## t = 2.5404, df = 18, p-value = 0.02051
## alternative hypothesis: true difference in means between group seeded and group unseeded is not equal to 0
## 95 percent confidence interval:
##   1.482679 15.657321
## sample estimates:
##   mean in group seeded mean in group unseeded
##                 303.63                 295.06
```

- reject or fail to reject the null hypothesis?

statistics using R

- biological interpretation?
- assumptions?
    - normality within each group?
    - equal variance between groups?
- could test for normality with Shapiro-Wilks test for each group separately (I'll show you a much better ways to do this later)

```
# normality for seeded streatment
shapiro.test(clouds$moisture[clouds$treatment=="seeded"])

# normality for unseeded streatment
shapiro.test(clouds$moisture[clouds$treatment=="unseeded"])
```

statistics using R

- null hypotheses?

```
# normality for seeded streatment
shapiro.test(clouds$moisture[clouds$treatment=="seeded"])
```

```
##
##     Shapiro-Wilk normality test
##
## data:  clouds$moisture[clouds$treatment == "seeded"]
## W = 0.93919, p-value = 0.544
```

```
# normality for unseeded streatment
shapiro.test(clouds$moisture[clouds$treatment=="unseeded"])
```

```
##
##     Shapiro-Wilk normality test
##
## data:  clouds$moisture[clouds$treatment == "unseeded"]
## W = 0.87161, p-value = 0.1044
```

- fail to reject null hypotheses for both groups, therefore not different from normal

statistics using R

- test equal variance using an *F* test
- null hypothesis?

```
var.test(clouds$moisture~clouds$treatment)
```

```
## 
##  F test to compare two variances
## 
## data:  clouds$moisture by clouds$treatment
## F = 0.57919, num df = 9, denom df = 9, p-value = 0.4283
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1438623 2.3318107
## sample estimates:
## ratio of variances 
##          0.5791888
```

- fail to reject null hypotheses and therefore variances are equal

linear models in R

- an alternative, but equivalent approach is to use a linear model to compare the means in each group
- general linear models are generally thought of as simple models, but can be used to model a wide variety of data and exp. designs
- traditionally statistics is performed (and taught) like using a recipe book (ANOVA, *t*-test, ANCOVA etc)
- general linear models provide a coherent and theoretically satisfying framework on which to conduct your analyses

what are linear models?

- *t*-test
- ANOVA
- factorial ANOVA
- ANCOVA
- linear regression
- multiple regression
- etc, etc

model formulae

- general linear modelling is based around the concept of model formulae

```
response variable ~ explanatory variable(s) + error
```

- literally read as ‘*variation in response variable modelled as a function of the explanatory variable(s) plus variation not explained by the explanatory variables*’
- it's the attributes of the response and explanatory variables that determines the type of linear model fitted

`response ~ continous variable`	equivalent to simple linear regression
`response ~ categorical variable`	equivalent to one-way ANOVA

linear modelling in R

- the function for carrying out linear regression in R is `lm()`

-the response variable comes first, then the tilde ~ then the name of the explanatory variable

```
clouds.lm <- lm(moisture ~ treatment, data=clouds)
```

- how does R know that you want to perform a *t*-test (ANOVA)?

```
class(clouds$treatment)
## [1] "character"
```

- here the explanatory variable is a factor

linear modelling in R

- to display the ANOVA table use the `anova()` function

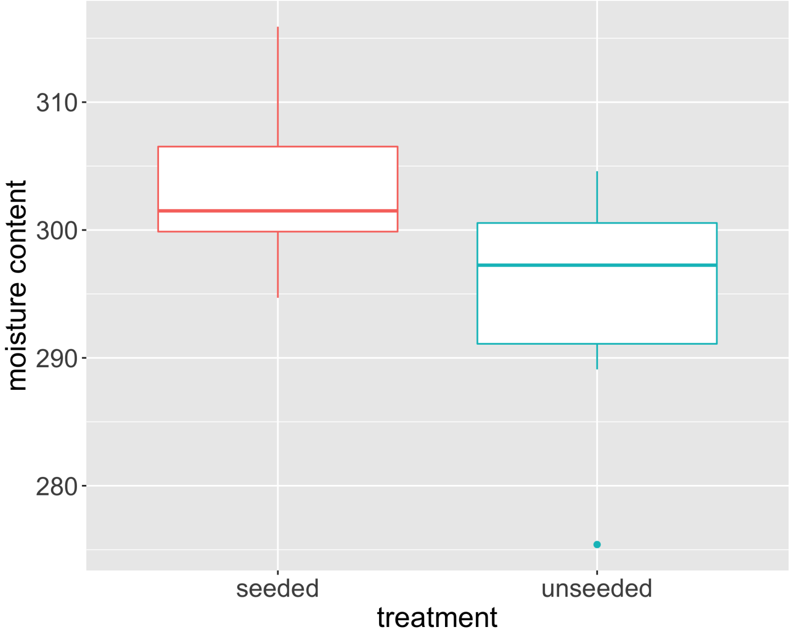
```
anova(clouds.lm)
```

```
## Analysis of Variance Table
##
## Response: moisture
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## treatment  1  367.22  367.22  6.4538 0.02051 *
## Residuals 18 1024.20   56.90
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- do you notice anything familiar about the p value?
- (hint: see the output from the *t*-test we did earlier)

linear modelling in R

- we have sufficient evidence to reject the null hypothesis (as before)
- therefore, there is a significant difference in the mean moisture content between clouds that were seeded and unseeded clouds
- do we accept this inference?
- what about assumptions?
- we could use Shapiro-Wilks and F tests as before
- much better to assess visually by plotting the residuals

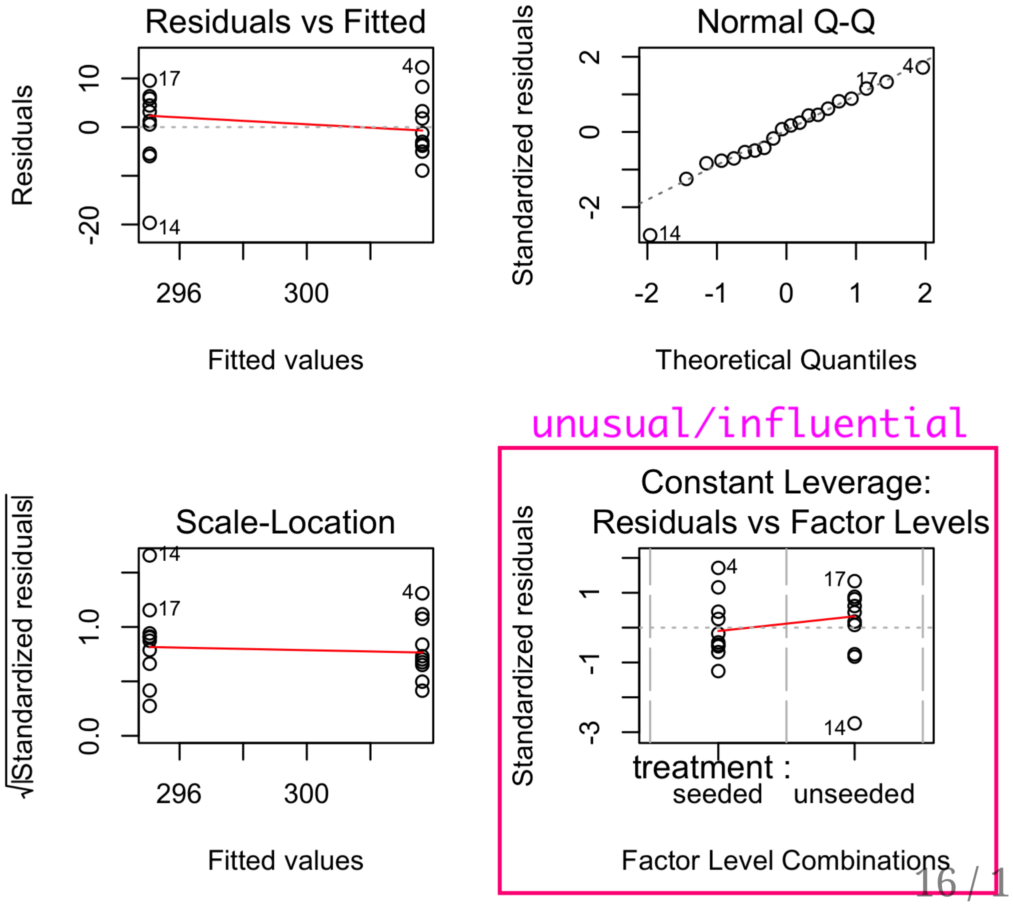


linear modelling in R

- because `clouds.lm` is a linear model object we can do stuff with it
- we can use the `plot()` function directly to display residual plots

```
par(mfrow = c(2, 2))
plot(clouds.lm)
```

- normality assumption
- equal variance assumption
- unusual or influential observations



other linear models

traditional name	model formula	R code
simple linear regression	Y ~ X1 (continuous)	`lm(Y ~ X)`
one-way ANOVA	Y ~ X1 (categorical)	`lm(Y ~ X)`
two-way ANOVA	Y ~ X1 (cat) + X2 (cat)	`lm(Y ~ X1 + X2)`
ANCOVA	Y ~ X1 (cat) + X2 (cont)	`lm(Y ~ X1 * X2)`
multiple regression	Y ~ X1 (cont) + X2 (cont)	`lm(Y ~ X1 + X2)`
factorial ANOVA	Y ~ X1 (cat) * X2 (cat)	`lm(Y ~ X1 * X2)`

]

Thanks!

I created these slides with xaringan and R Markdown using the rutgers css that I slightly modified.

Credit: I borrowed slides from Alex Douglas.