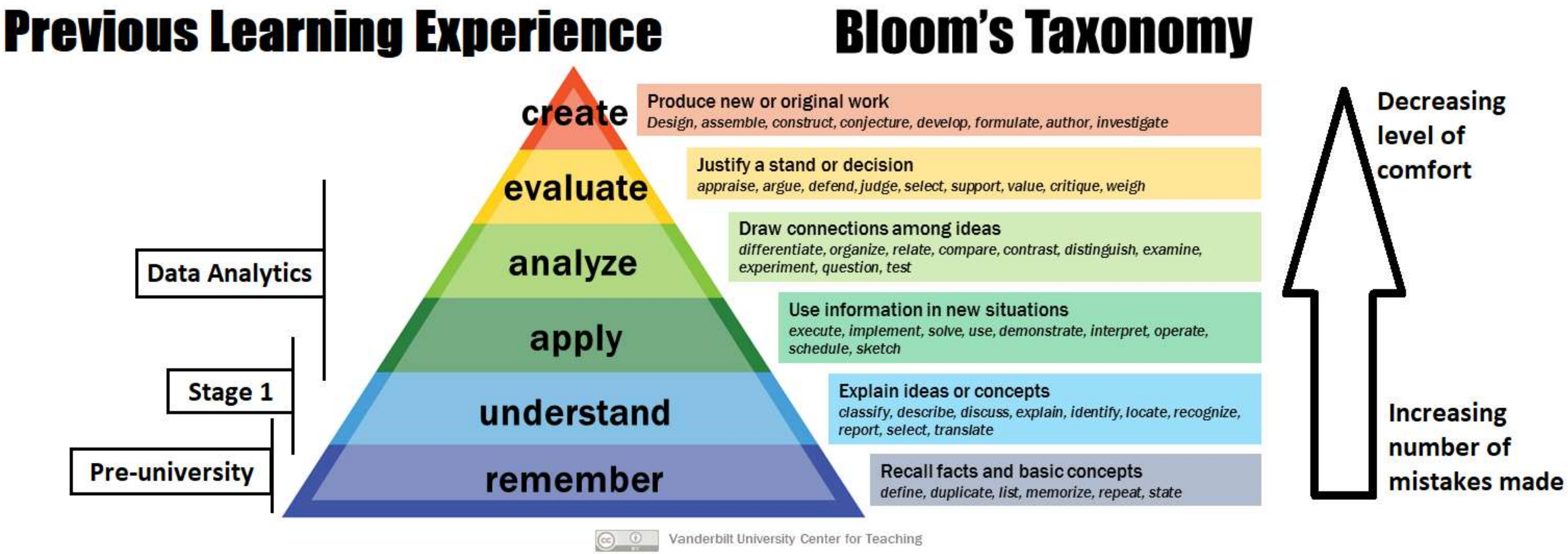
"It's hard........"



Previous Learning Experience
Bloom's Taxonomy
create
Produce new or original work
Design, assemble, construct, conjecture, develop, formulate, author, investigate
evaluate
Justify a stand or decision
appraise, argue, defend, judge, select, support, value, critique, weigh
analyze
Draw connections among ideas
differentiate, organize, relate, compare, contrast, distinguish, examine, experiment, question, test
apply
Use information in new situations
execute, implement, solve, use, demonstrate, interpret, operate, schedule, sketch
understand
Explain ideas or concepts
classify, describe, discuss, explain, identify, locate, recognize, report, select, translate
remember
Recall facts and basic concepts
define, duplicate, list, memorize, repeat, state
Data Analytics
Stage 1
Pre-university
Decreasing level of comfort
Increasing number of mistakes made
Vanderbilt University Center for Teaching

17C

Laboratory & Professional Skills:
Data Analysis

Emma Rand
Data Analysis in R

Week 5 Introduction to one- and two-sample tests. One sample tests

Last week

- The normal distribution - because it is the basis of many tests (parametric tests such as *t*-test, regression and ANOVA)
  – Properties of normal distributions
  – Sampling distribution of the mean and the standard error
  – Confidence intervals
- In RStudio
  – Calculate probabilities and quantiles from normal distributions
  – Calculate confidence intervals

Summary of this week

We will cover

- General intro to one- and two-sample tests
- one sample-tests
  - The one-sample t-test
  - The paired sample t-test
  - The one sample Wilcoxon

R practice

- No workshop
- Activities to consolidate previous skills: Workflow basics and projects

Summary of this week

- We will cover one sample-tests
  - **The one-sample t-test**
  - **The paired sample t-test**
  - The one sample Wilcoxon
- R practice
  - No workshop
  - Activities to consolidate previous skills : Workflow basics and projects

‘Parametric’ tests –

Based on the normal distribution

Summary of this week

- We will cover one sample-tests
  - The one-sample t-test
  - The paired sample t-test
  - **The one sample Wilcoxon** Non-parametric
- R practice
  - No workshop
  - Activities to consolidate previous skills : Workflow basics and projects

Learning objectives

By actively following the material and carrying out the independent study the successful student will be able to :

- Appreciate that *t*-tests are based on the normal distribution and have assumptions relating to it
- Understand the principles of *t*-tests
- Select, appropriately, one-sample *t*-tests and their non-parametric equivalent (MLO 2)
- Recognise when two samples are not independent (MLO 2)
- Know what functions are used in R to run these tests and how to interpret them(MLO 3 and 4)
- Know how to state the results of these tests scientifically (MLO 3 and 4)

Introduction to one- and two-sample tests

T-tests and their non-parametric equivalents

Reminder: The choice of test depends on ….

1. Type of data

The type of values a variable can take: <u>Discrete</u> or <u>continuous</u>?

2. Their role in the analysis

Which is the response and which is/are explanatory?

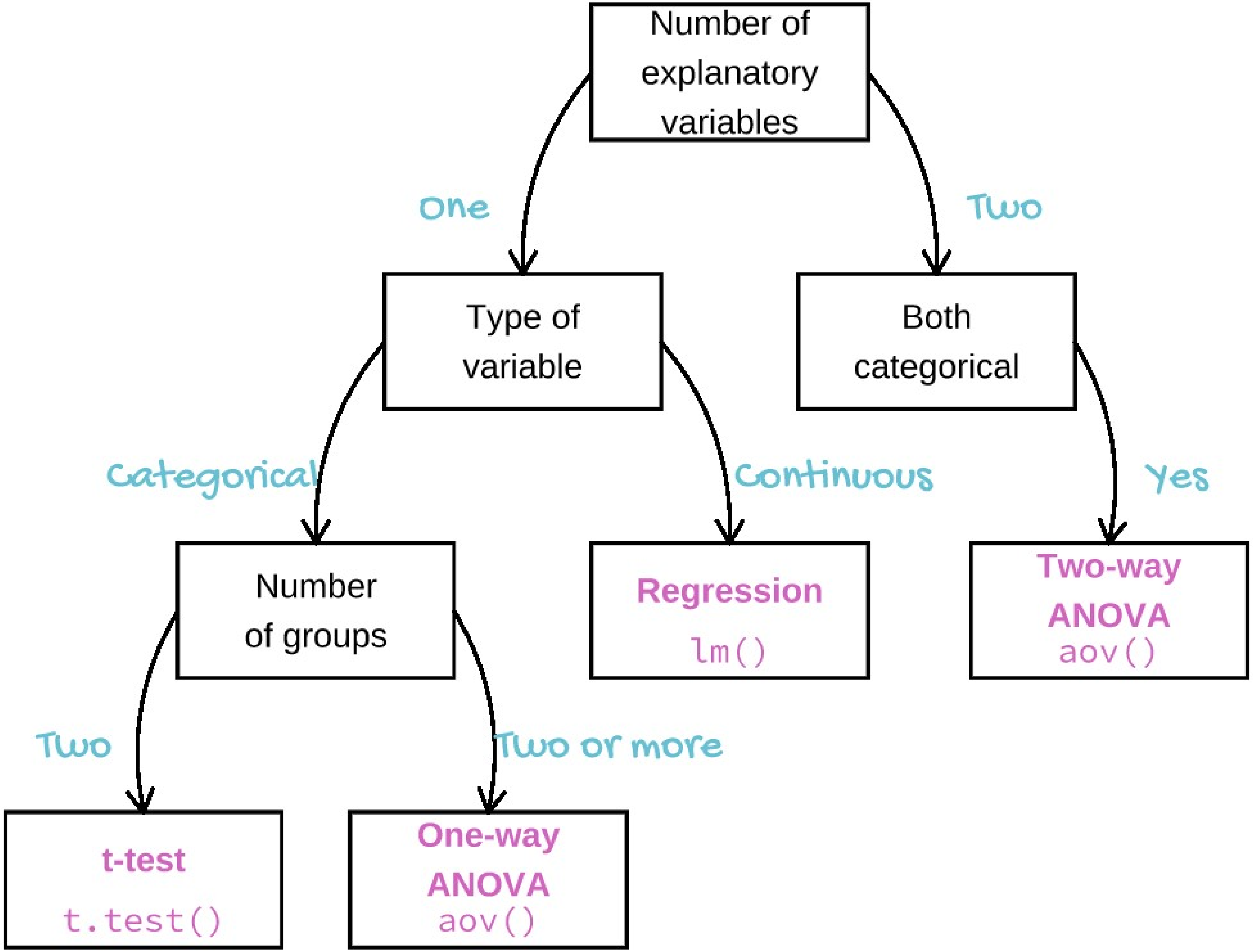
(week 3 Hypothesis testing, data types, reading data in to R and saving figures

Choosing tests: 3 steps

1. What is a one sentence description of what you want to know?
2. What are your explanatory variables?
   - Categories: *t*-tests, ANOVA, Wilcoxon, Mann-Whitney
   - Continuous: Regression, correlation
3. What is your response variable?
   - Normally distributed: *t*-tests, ANOVA, regression
   - Counts: Chi-squared or stage 2 ☺

Choosing tests: 3 steps

1. What is a one sentence description of what you want to know?
2. What are your explanatory variables?
   - Categories: *t*-tests, ANOVA, Wilcoxon, Mann-Whitney
   - Continuous: Regression, correlation
3. What is your response variable?
   - Normally distributed: *t*-tests, ANOVA, regression
   - Counts: Chi-squared or stage 2 ☺


Number of explanatory variables
One
Two
Type of variable
Both categorical
Categorical
Continuous
Yes
Number of groups
Regression
lm()
Two-way ANOVA
aov()
Two
Two or more
t-test
t.test()
One-way ANOVA
aov()

Types of *t*-test

1. One-sample
   - Compares the mean of sample to a particular value (compares the response to a reference)
     - Includes paired-sample test – compares the mean difference to zero (i.e., compares dependent means)

2. Two-sample
   - Compares two (independent) means to each other

t-tests in general

Assumptions

All *t*-tests assume the “residuals” are normally distributed and have homogeneity of variance

A residual is the difference between the predicted and observed value

Predicted value is the mean / group mean *t*-tests in general: assumptions

Checking Assumptions

- Common sense
  - Data should be continuous
  - No/few repeats
- Plot the residuals
- Using a test in R

When residuals are not normally distributed

- Transform (not really covered)
  - E.g. Log to remove skew
- Use a non-parametric test (covered)
  - Fewer assumptions
  - Generally less powerful

The one-sample *t*-test

A parametric test

One-sample *t*-tests

Tests whether the mean of a single sample differs from an expected value (i.e., H_0)

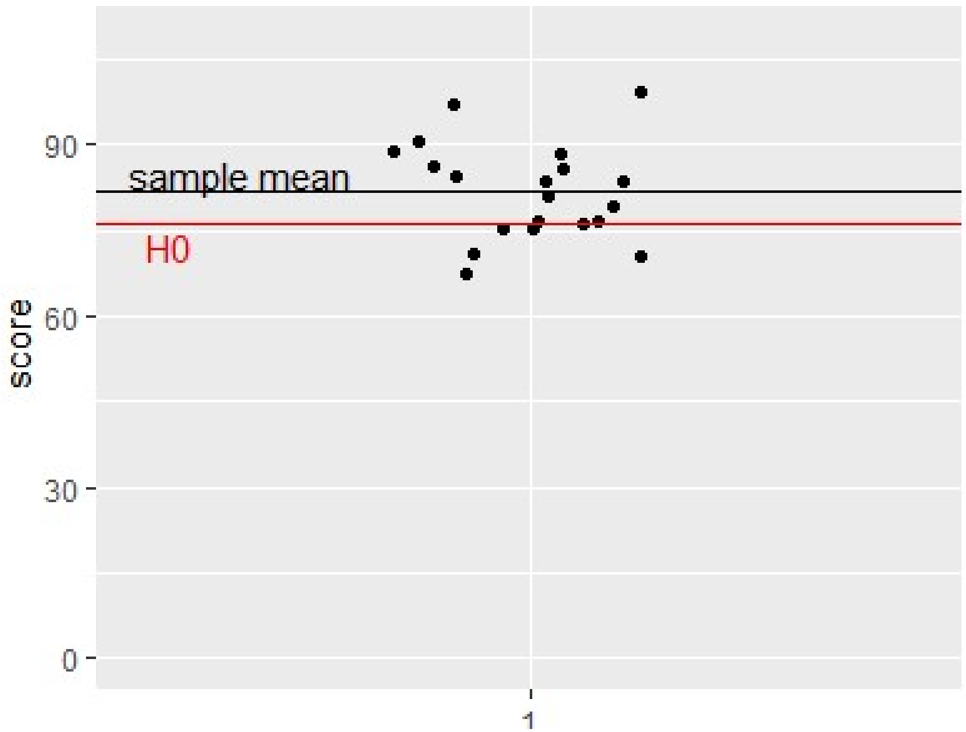
- Example: Fields are sprayed if crop plants have a disease score* of 76.
- 20 plants in a field are measured
- Is their mean significantly different from the reference of 76?

*Arbitrary scale

One-sample *t*-tests - example

```
plants %>%
  summarise(mean = mean(score),
            sd = sd(score),
            n = length(score))

       mean     sd          length
1      81.803   8.533749    20
```



	score
1	76.11
2	76.52
3	83.37
4	88.28
5	83.67
6	67.40
7	75.43
8	97.03
9	75.46
10	90.42
11	99.30
12	79.00
13	85.55

One-sample t-tests - example

- H_0: mean = 76

- Standard formula for all t-tests

$$t = \frac{statistic - hypothesised\ value}{s.e.of\ \ statistic}$$

- d.f.= n - 1

One-sample *t*-tests - example

$$t = \frac{statistic - hypothesised\ value}{s.e. of\ statistic}$$

$$\bar{x} = 81.8 \qquad \mu = 76.00$$

Is the difference between the obtained value and the expected value big relative to the variability?

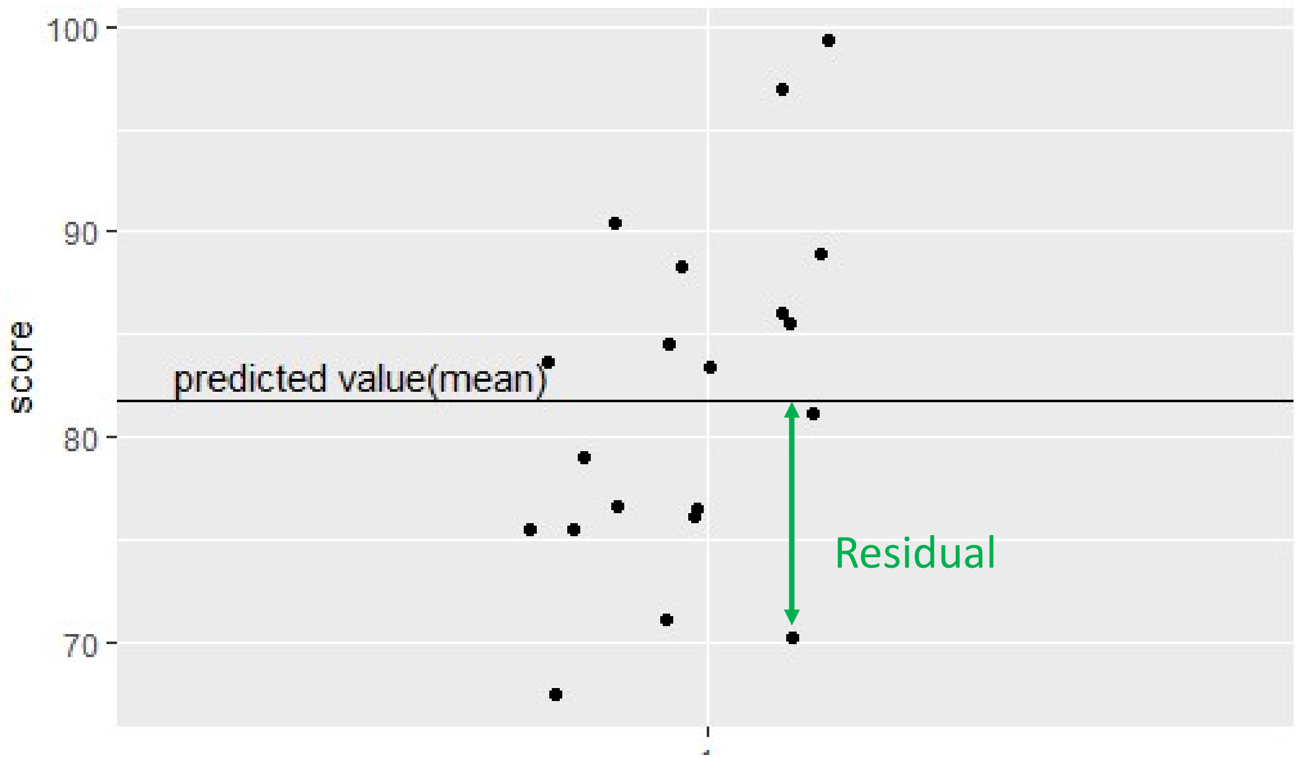
One-sample *t*-tests - example

```
t.test(data = plants, score, mu = 76)
             One Sample t-test

data:  score
t = 2.517, df = 19, p-value = 0.02097
alternative hypothesis: true mean is not equal to 76
95 percent confidence interval:
 77.80908 85.79692
sample estimates:
mean of x
   81.803
```

One-sample *t*-tests - example

Checking the assumptions: normally and homogenously distributed residuals



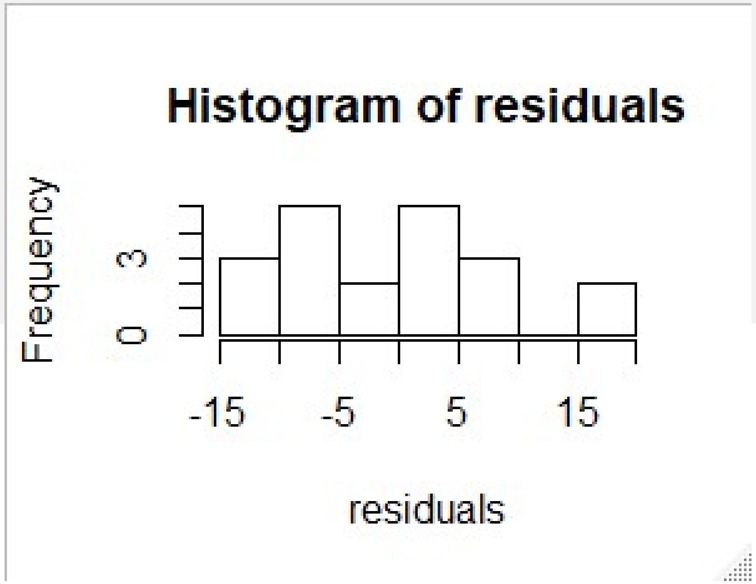
One-sample *t*-tests - example

Checking the assumptions: normally and homogenously distributed residuals

```
residuals <- plants$score - mean(plants$score)
hist(residuals)
shapiro.test(residuals)

Shapiro-Wilk normality test

data:  residuals
W = 0.9725, p-value = 0.8065
```



One-sample *t*-tests - example

Reporting the result: "significance of effect, direction of effect, magnitude of effect"

The disease score for plants in this field ($\bar{x} = 81.8$) is significantly higher than 76 (t = 2.52; $d.f.$ = 19; p = 0.021).

One-sample *t*-tests - summary

- Parametric
- To test whether the mean of a single sample differs from an expected value
- *t* is size of difference relative to the s.e.
- Function in R:
  `t.test(data = df, response, mu = Exp)`
- If $p < 0.05$ the test is significant
- assumptions: normally and homogenously distributed residuals
- Significance, direction, magnitude
- Figure: probably not needed

The paired sample *t*-test

A parametric test

Paired-sample *t*-tests

- Two samples but values are not independent (could not reorder)

Actually a one-sample test

Paired-sample *t*-tests example

Is there a difference between the maths and stats marks of 10 students?

The one sample is the difference between the pairs of values

n.b. tidy data

Same student

	subject	mark
1	maths	97
2	maths	58
3	maths	65
4	maths	65
5	maths	80
6	maths	48
7	maths	85
8	maths	63
9	maths	58
10	maths	98
11	stats	89
12	stats	49
13	stats	68
14	stats	70
15	stats	74
16	stats	30
17	stats	78
18	stats	69
19	stats	40
20	stats	85

Paired-sample t-tests - example

- H_0: mean difference = 0
- Standard formula for all t-tests

$$t = \frac{statistic - hypothesised\ value}{s.e.of\ statistic}$$

- $t_{[d.f]} = \frac{\bar{d} - 0}{s.e.\ of\ \bar{d}}$
- d.f.= n – 1 (where n is the number of pairs)

Paired-sample *t*-tests

Run paired sample *t*-test

```
t.test(data = marks, mark ~ subject, paired = TRUE)

        Paired t-test

data:  mark by subject
t = 2.3399, df = 9, p-value = 0.04403
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  0.2159788 12.7840212
sample estimates:
mean of the differences
                    6.5
```

Paired-sample *t*-tests - example

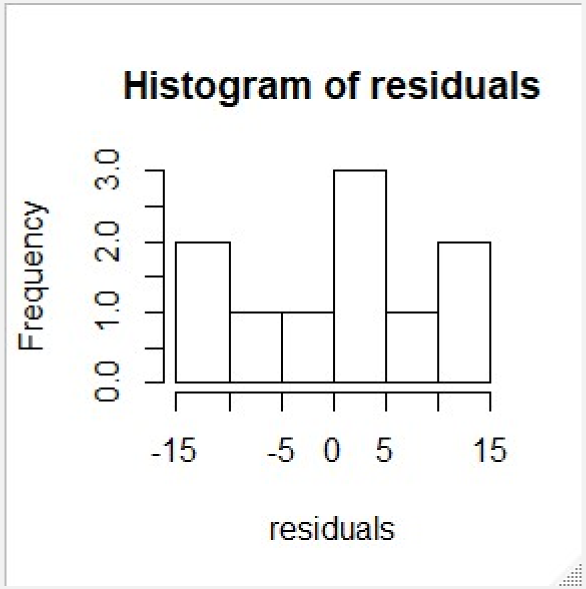
Checking the assumptions: normally and homogenously distributed residuals

```
diffs <- marks$mark[marks$subject == "maths"] -
marks$mark[marks$subject == "stats"]

residuals <- diffs - mean(diffs)
hist(residuals)
shapiro.test(residuals)


Shapiro-Wilk normality test

data:  residuals
W = 0.91246, p-value = 0.2983
```

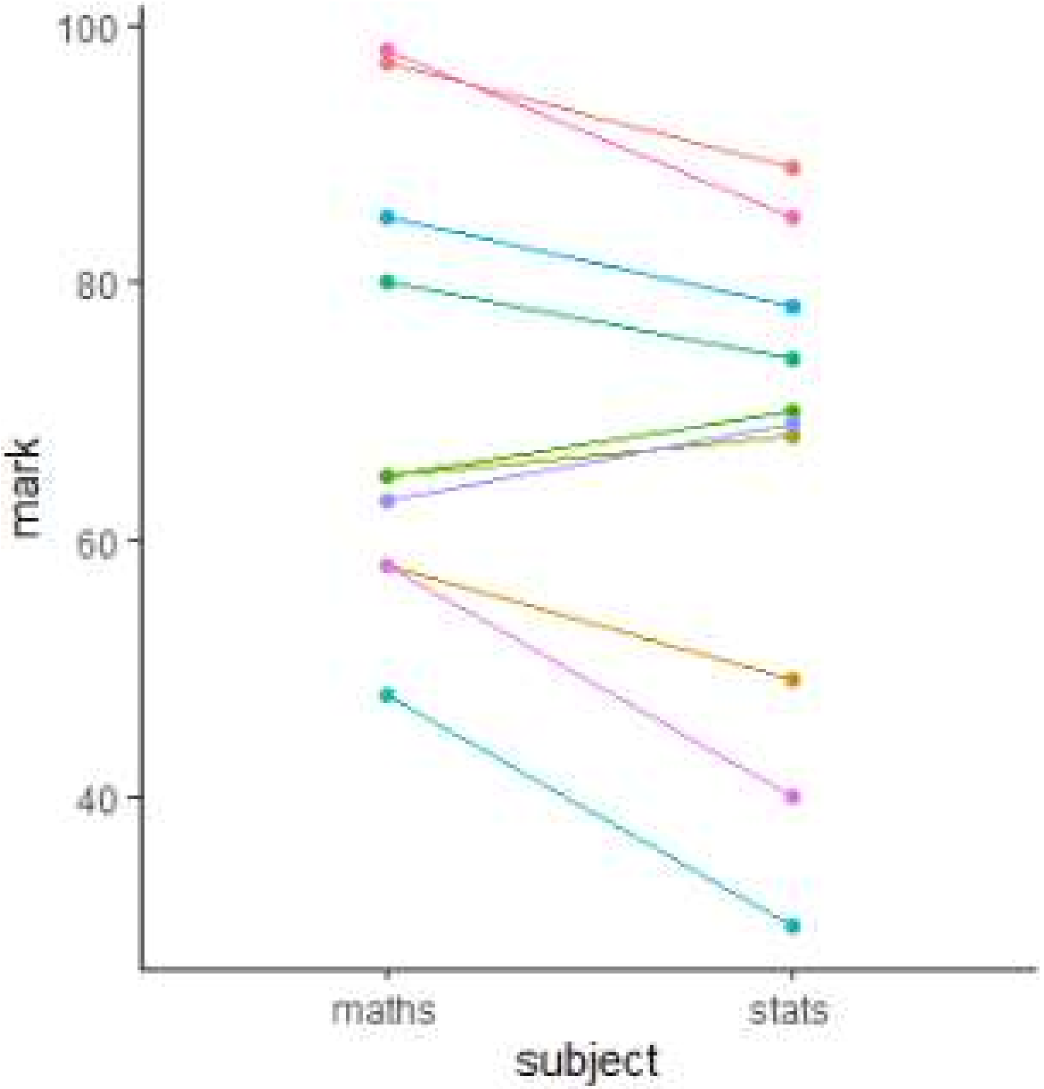


Paired-sample *t*-tests

Reporting the result: "significance of effect, direction of effect, magnitude of effect"

Individual students score significantly higher in maths than in statistics (t = 2.34; $d.f.$ = 9; p = 0.044) with an average difference of 6.5%.

Paired-sample *t*-tests: figure



Paired-sample *t*-tests - summary

- Parametric
- To test whether the mean difference between pairs of values is zero
- *t* is size of mean difference relative to the s.e.
- Function in R:
  `t.test(data = df, response ~ explanatory, paired = TRUE)`
- If $p < 0.05$ the test is significant
- assumptions: normally and homogenously distributed residuals
- Significance, direction, magnitude
- Figure: none or spaghetti plot

The one sample Wilcoxon

Non-parametric equivalent of the paired-sample *t*-test

When the *t*-test assumptions are not met: non- parametric tests

- Non-parametric tests make fewer assumptions
- Based on the ranks rather than the actual data
- Null hypotheses are about the *mean rank* (not the mean)

Non-parametric tests

t-test equivalents

i,.e., the type of question is the same but the response variable is not normally distributed or it is impossible to tell (small samples)

- one – sample *t*-test and paired-sample *t*-test: the one-sample Wilcoxon
- Two-sample *t*-test (next week): two-sample Wilcoxon aka Mann-Whitney

Non-parametric tests

one/paired-sample Wilcoxon

Marks – small sample.

Wilcoxon might be more appropriate

```
wilcox.test(data = marks, mark ~ subject, paired = TRUE)

   Wilcoxon signed rank test with continuity correction

data:  mark by subject
V = 48.5, p-value = 0.03641
alternative hypothesis: true location shift is not equal to 0


Warning message:
In wilcox.test.default(x = c(97L, 58L, 65L, 65L, 80L, 48L, 85L,  :
  cannot compute exact p-value with ties
```

Non-parametric tests

one/paired-sample Wilcoxon

Reporting the result: "significance of effect, direction of effect, magnitude of effect"

Individual students score significantly higher in maths than in statistics (Wilcoxon: $V = 48.5$; $n = 10$; $p = 0.036$) with a median difference of 7.5%.

Wilcoxon- summary

- Non-parametric
- when assumptions for *t*-test not met
- To test whether the mean rank difference between pairs of values is zero
- Function in R:
  `wilcox.test(data = df, response ~ explanatory, paired = TRUE)`
- If $p < 0.05$ the test is significant
- Few assumptions
- Figure: none or spaghetti plot