17C Laboratory & Professional Skills: Data Analysis

Laboratory & Professional skills for Bioscientists
Term 2: Data Analysis in R

One sample tests: one-sample *t*-test, paired-sample *t*-test and one-sample Wilcoxon

Summary of this week and next

- We will consider tests for one-, two- and paired-samples. These are the *t*-tests and their non-parametric equivalents. We will apply what we know about choosing appropriate tests
- Two lectures.

Overview of topics

Week	Topic	
2	Introduction. Logic of hypothesis testing	Foundation
3	Hypothesis testing, variable types	
4	Chi-squared tests	Hypothesis testing
5	The normal distribution, summary statistics and CI	Estimation
6 and 7	**One- and two-sample tests (2 lectures)**	
8	One-way ANOVA and Kruskal-Wallis	Hypothesis testing
9	Two-way ANOVA incl understanding the interaction	
10	Correlation and regression	

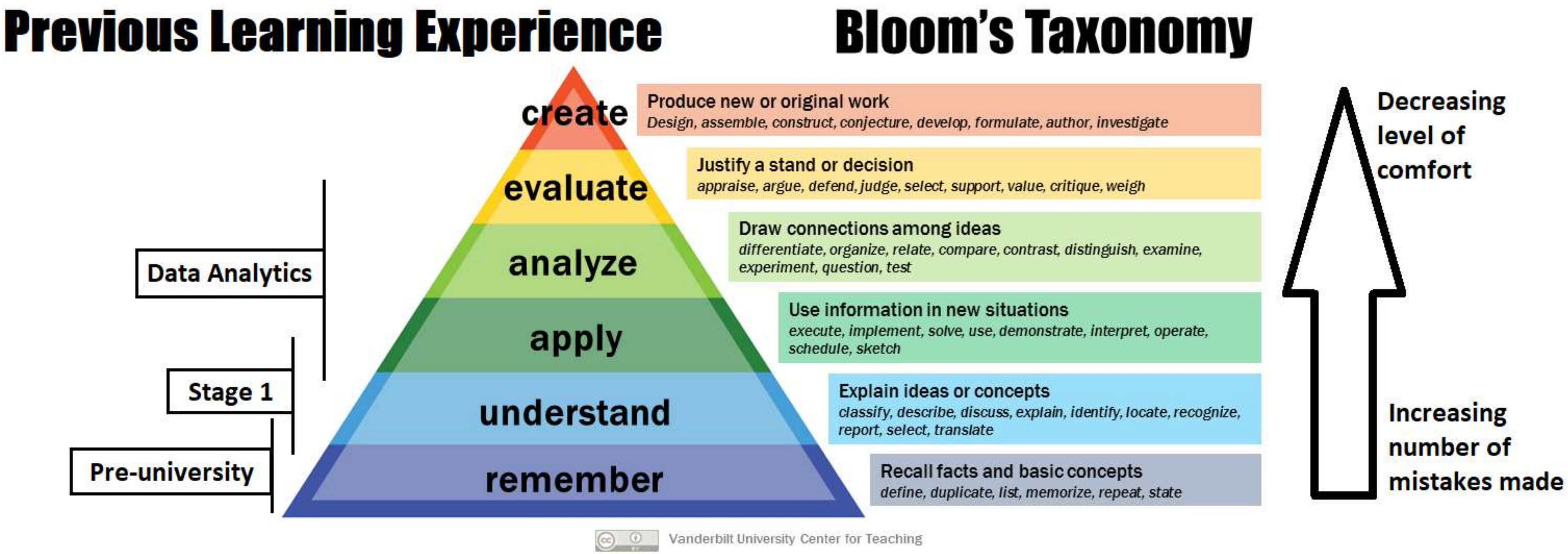
Lecture 2

1. Estimation
   – what is the mean of the population?
2. Hypotheses testing

   e.g., is there a difference between 2 means (*t*-test)

   e.g., is the expected number of observations what we expect (chi-squared test)

"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

Learning objectives for the 2 weeks

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

- Explain dependent and independent samples (MLO 2)
- Select, appropriately, *t*-tests and their non-parametric equivalents (MLO 2)
- Apply, interpret and evaluate the legitimacy of the tests in R (MLO 3 and 4)
- Summarise and illustrate with appropriate R figures test results scientifically (MLO 3 and 4)

Revision Lectures 1 and 2

Choosing tests

Regardless, the choice of statistic depends on ....

1. Type of data

The type of values a variable can take: Discrete or continuous?

2. Their role in the analysis

Which is the response and which is/are explanatory?

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 ☺

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

Student's *t*-test

'Student' was William Sealy Gosset

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

Checking Assumptions

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

When data are not normally distributed

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

One-sample *t*-tests

We often want to know if the mean of a sample differs from some reference value

Comparing a measure of water quality to a reference value

Confidence intervals: <u>small samples</u>

19 lactate dehydrogenase solutions to a recipe that <u>should</u> yield a concentration of 1.5 µmols l^{-1}

How good is the recipe/ability to follow the recipe?

Validating a method to determine Glucose concentration

One-sample *t*-tests

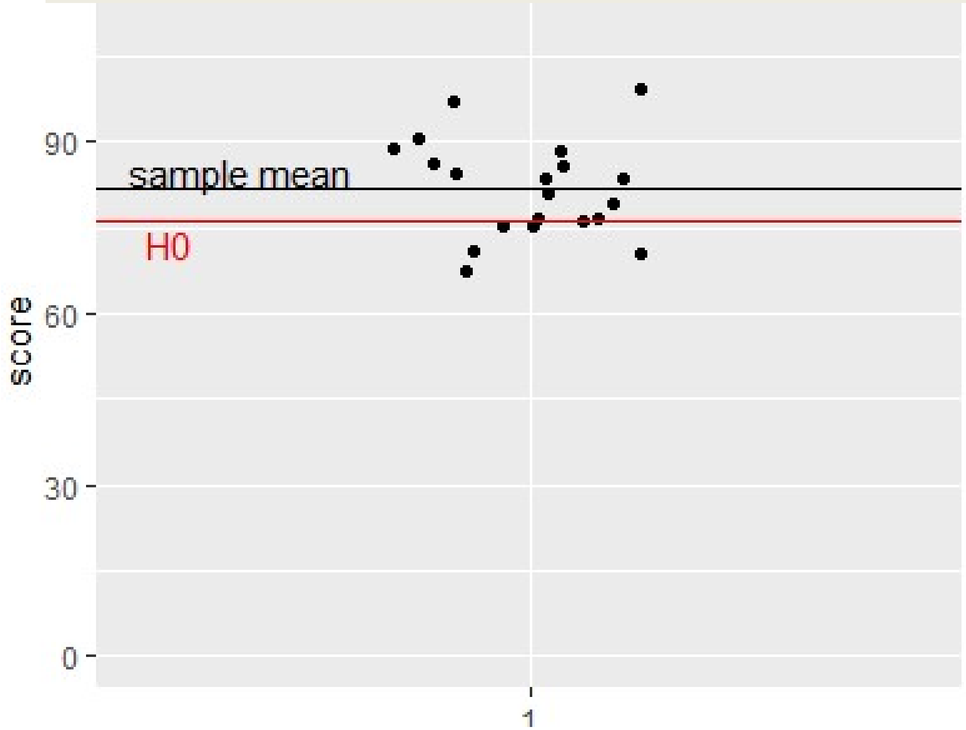
Tests whether the mean of a single sample differs from an expected value (i.e., H_0)

- 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

```
score %>%
  summarise(mean(score),
            sd(score),
            length(score))

mean(score) sd(score) length(score)
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
14	81.12

One-sample t-tests - example

- H_0: mean = 76 vs H_1: 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

- H_0: mean = 76 vs H_1: mean ≠ 76

- Standard formula for all *t*-tests

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

Summary of the observed data

What is the expected value if H_0 true

- d.f.= n - 1

One-sample *t*-tests - example

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

$$\bar{x} = 80.168$$

$$\mu = 76.00$$

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

One-sample *t*-tests - example

Run the *t*-test

Manual:

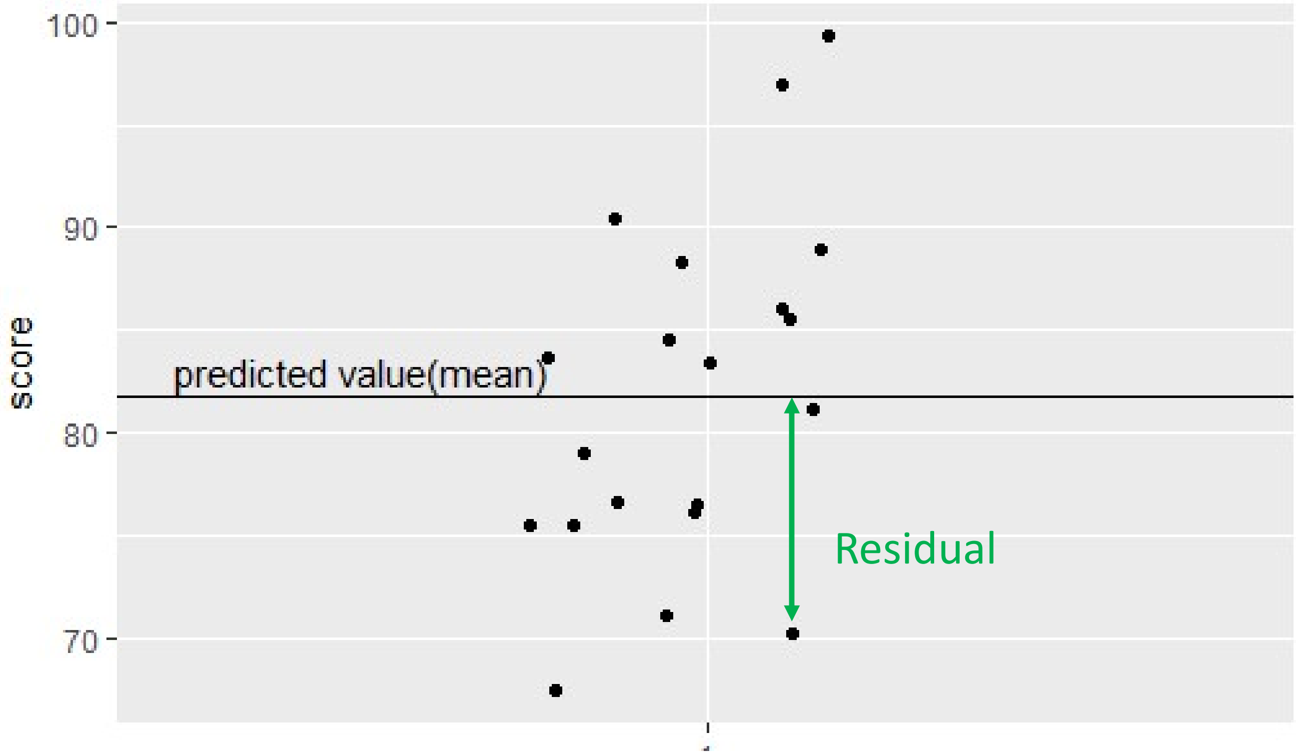
```
t.test(x, y = NULL, alternative =
  c("two.sided", "less", "greater"), mu = 0,
  paired = FALSE, var.equal = FALSE, conf.level
  = 0.95, ...)
```

One-sample *t*-tests - example

```
t.test(data = score, 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 77
95 percent confidence interval:
 77.80908 85.79692
sample estimates:
mean of x
   81.803
```

One-sample *t*-tests - example

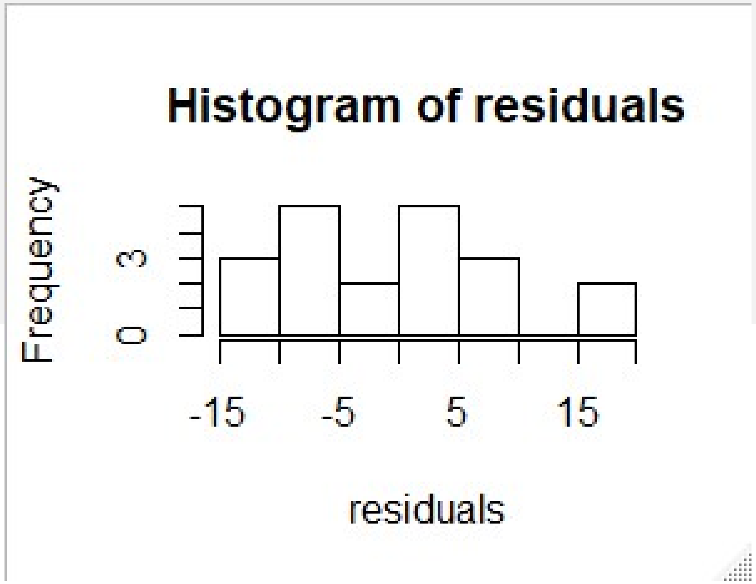
Checking the assumptions: normally and homogenously distributed residuals



One-sample *t*-tests - example

Checking the assumptions: normally and homogenously distributed residuals

```
residuals <- score$score - mean(score$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 ($\bar{x} = 81.2$) is significantly higher than 76 (*t* = 2.52; *d.f.* = 19; *p* = 0.021).

Paired-sample *t*-tests

- Really a one-sample test
- Two samples but values are not independent (could not reorder)

Patient	Drug	Placebo
1	14	18
2	26	29
3	21	24
etc		

- N.b. not ‘tidy’ data

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 vs H_1: 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

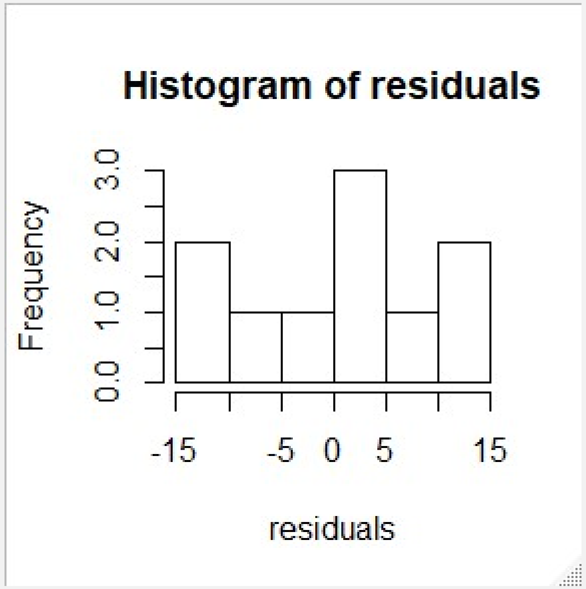
Run paired sample *t*-test

```
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)
```



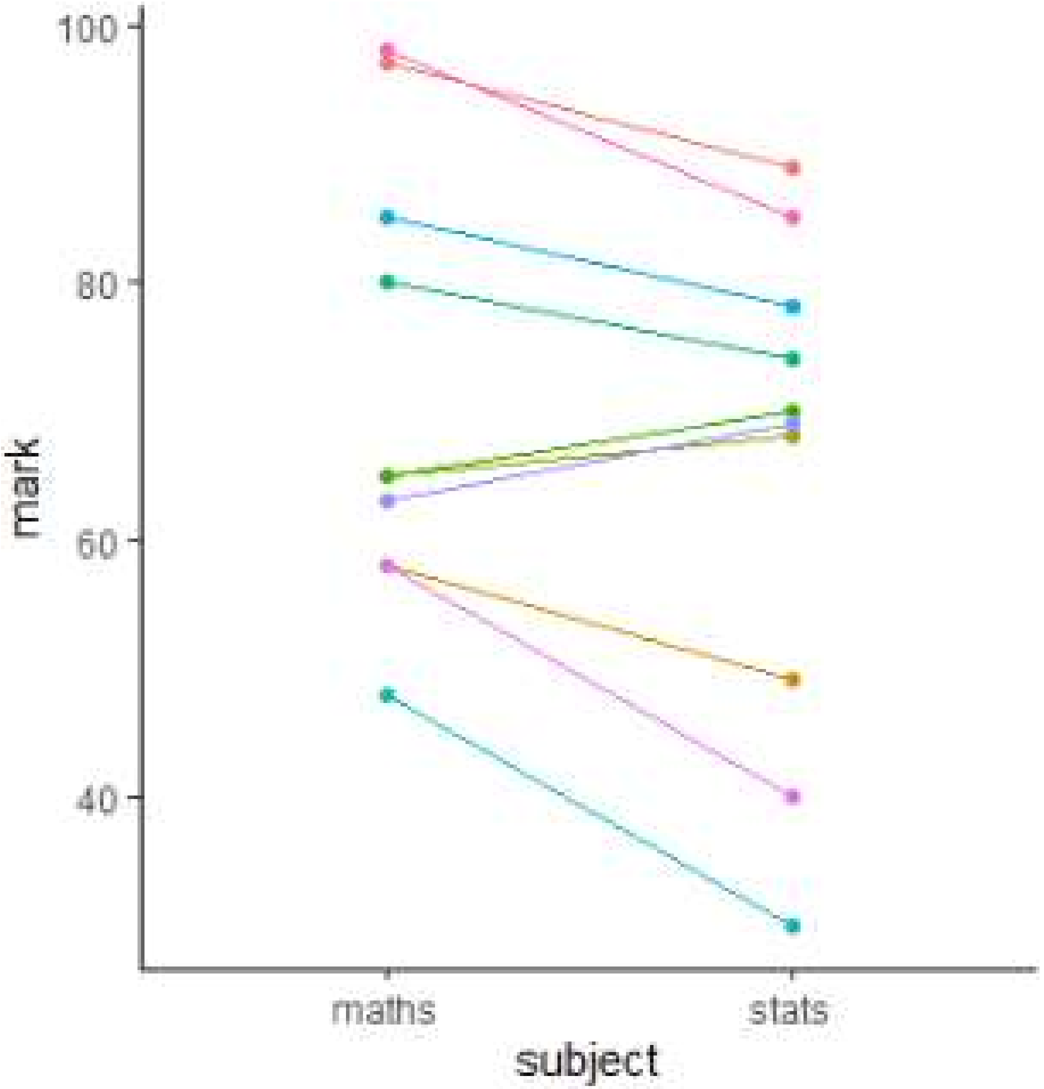

```
Shapiro-Wilk normality test

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



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 lecture): 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%.

Learning objectives for the week

By attending the lectures and practical the successful student will be able to

- Explain dependent and independent samples (MLO 2)
- Select, appropriately, *t*-tests and their non-parametric equivalents (MLO 2)
- Apply, interpret and evaluate the legitimacy of the tests in R (MLO 3 and 4)
- Summarise and illustrate with appropriate R figures test results scientifically (MLO 3 and 4)