OXFORD UNIVERSITY STATISTICAL CONSULTING

Introduction to statistical testing

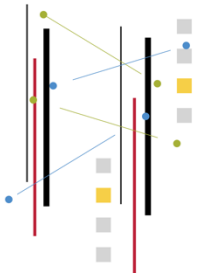
Chi squared test

- Chi squared test is used to test independence of two (or more) variables on a contingency table

$$X^2 = \sum \frac{(X_{ij} - Np_{ij})^2}{Np_{ij}} = \sum \frac{(O-E)^2}{E}$$

- Example: Educational status for males and females
- Question: is the difference in frequencies between males and females statistically significant?

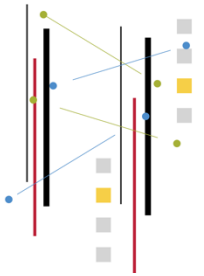
	caseid	sex	degree
1	1	Female	Bachelor
2	2	Male	Lt High School
3	3	Female	High School
4	4	Female	Bachelor
5	5	Female	High School
6	6	Male	High School
7	7	Male	High School
8	8	Male	Bachelor



Chi squared test – in R

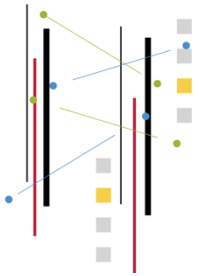
- Making a contingency table

```
> table(gss$degree, gss$sex)

                  Male Female
Lt High School    5153   6669
High School      12340  16947
Junior College    1272   1798
Bachelor          3822   4180
Graduate          2091   1779
```



Chi squared test - Assumptions

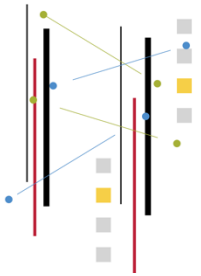
- Independence of observations
- Normality
  - Violated if expected cell frequencies are small: rule of thumb – should be >5
  - Use Fisher's exact test instead



Chi squared test – in R

- Running the test

```
> table(gss$degree, gss$sex)->contTable
> chisq.test(contTable)

        Pearson's Chi-squared test

data:  contTable
X-squared = 254.35, df = 4, p-value < 2.2e-16
```

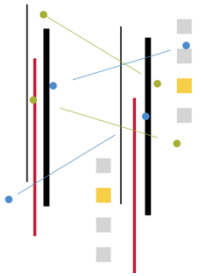


Chi squared test – in R

- Sanity check

```
> table(gss$degree, gss$sex)

                  Male Female
Lt High School    5153   6669
High School      12340  16947
Junior College    1272   1798
Bachelor          3822   4180
Graduate          2091   1779
```

```
> chisq.test(contTable)->chiSquareEducation
> chiSquareEducation$observed

                  Male Female
Lt High School    5153   6669
High School      12340  16947
Junior College    1272   1798
Bachelor          3822   4180
Graduate          2091   1779
```



Chi squared test – in R

- Observed vs Expected values ($X^2 = \sum \frac{(O-E)^2}{E}$)

```
> chiSquareEducation$observed
```

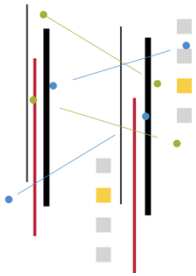
	Male	Female
Lt High School	5153	6669
High School	12340	16947
Junior College	1272	1798
Bachelor	3822	4180
Graduate	2091	1779

```
> chiSquareEducation$expected
```

	Male	Female
Lt High School	5204.962	6617.038
High School	12894.410	16392.590
Junior College	1351.652	1718.348
Bachelor	3523.101	4478.899
Graduate	1703.874	2166.126

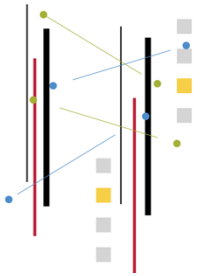
```
> chiSquareEducation$residuals
```

	Male	Female
Lt High School	-0.7202375	0.6387820
High School	-4.8823675	4.3301947
Junior College	-2.1665341	1.9215093
Bachelor	5.0357171	-4.4662012
Graduate	9.3784959	-8.3178320



T-test

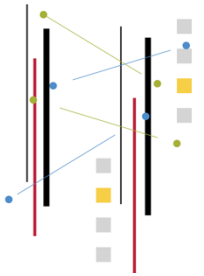
- T-tests are used for continuous data
- One sample t-test is used to compare a sample against a known mean
  - e.g. the height of people in a classroom vs average national
- Two sample t-test to compare two different samples
  - e.g. plant yield in fields under one treatment versus another
  - A special case is paired t-test where the samples are paired
  - e.g. average monthly temperature between 2 years



T-test – One sample (and paired…)

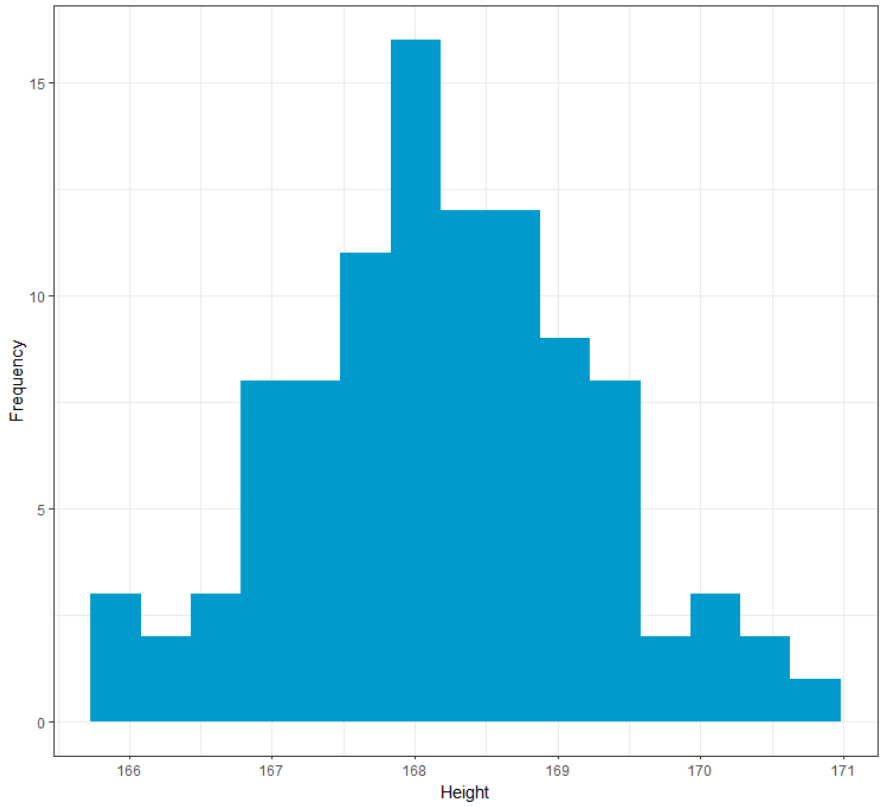
- $\bar{X}$ is the sample mean,
- μ is the true population mean,
- s is the sample standard deviation
- N is the sample size

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$



T-test – One sample

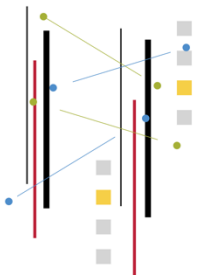
Is the height of the female students in a classroom significantly different to the average for women in the UK?



T-test – One sample

- We want to test the difference in height between the classroom and the UK average for women (164.3 cm)
  - H_0: The difference is zero
  - H_1:The difference is not zero
- The sample size is (N) is 100 students
- The sample average height ($\bar{X}$) is 168.2 cm
- The sample standard deviation (s) is 1.04 cm
- The population average is (μ) is 164.3 cm
- Degrees of freedom are $N - 1 = 99$

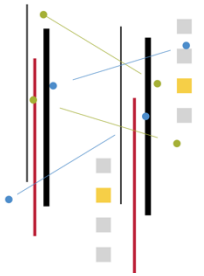
$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$



T-test – Assumptions

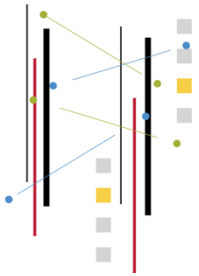
- Independence
- Normality in outcome (Shapiro-Wilk test)

```
> shapiro.test(studentHeights)

        Shapiro-Wilk normality test

data:  studentHeights
W = 0.99323, p-value = 0.9017
```



T-test – One sample - in R

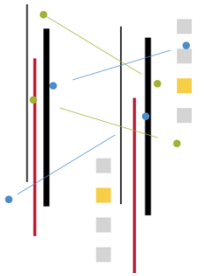
```
> t.test(studentHeights, mu=164.3)

        One Sample t-test

data:  studentHeights
t = 37.306, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 164.3
95 percent confidence interval:
 167.9714 168.3839
sample estimates:
mean of x
 168.1777
```



T-test – Paired

- A paired t-test is when we have a built-in structure in the data
  - For example we're comparing the before and after effect in a sample
  - Or we're comparing the temperatures between two years and we want to pair up Jan with Jan etc
- We can simplify the problem by reframing it:
  - Instead of looking at whether the two years are significantly different, we can see if their difference is significantly different to 0
  - This is essentially a one sample t-test with $\mu = 0$

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

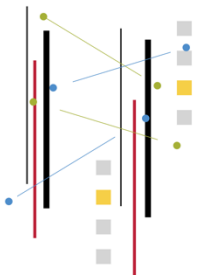


T-test – Paired

So what changes do we need to make?

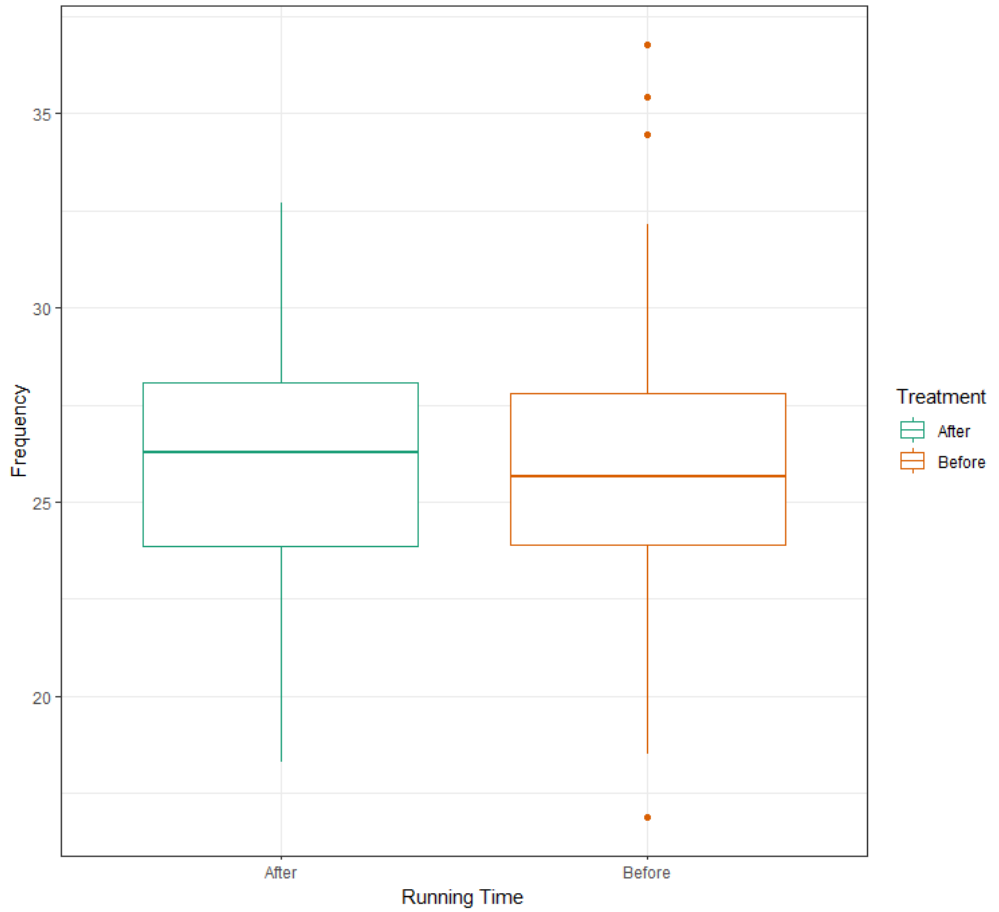
- Our sample is now the difference between the two samples
- Therefore $\bar{X}$ is the **average difference** between the
- μ is 0
- s is the standard deviation of the **differences**
- N is the number of differences we had. This is also equal to the **number of pairs** we had to start with
- Degrees of freedom are $N - 1$

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$



T-test – Paired

Does the running time of students change after a training regime?

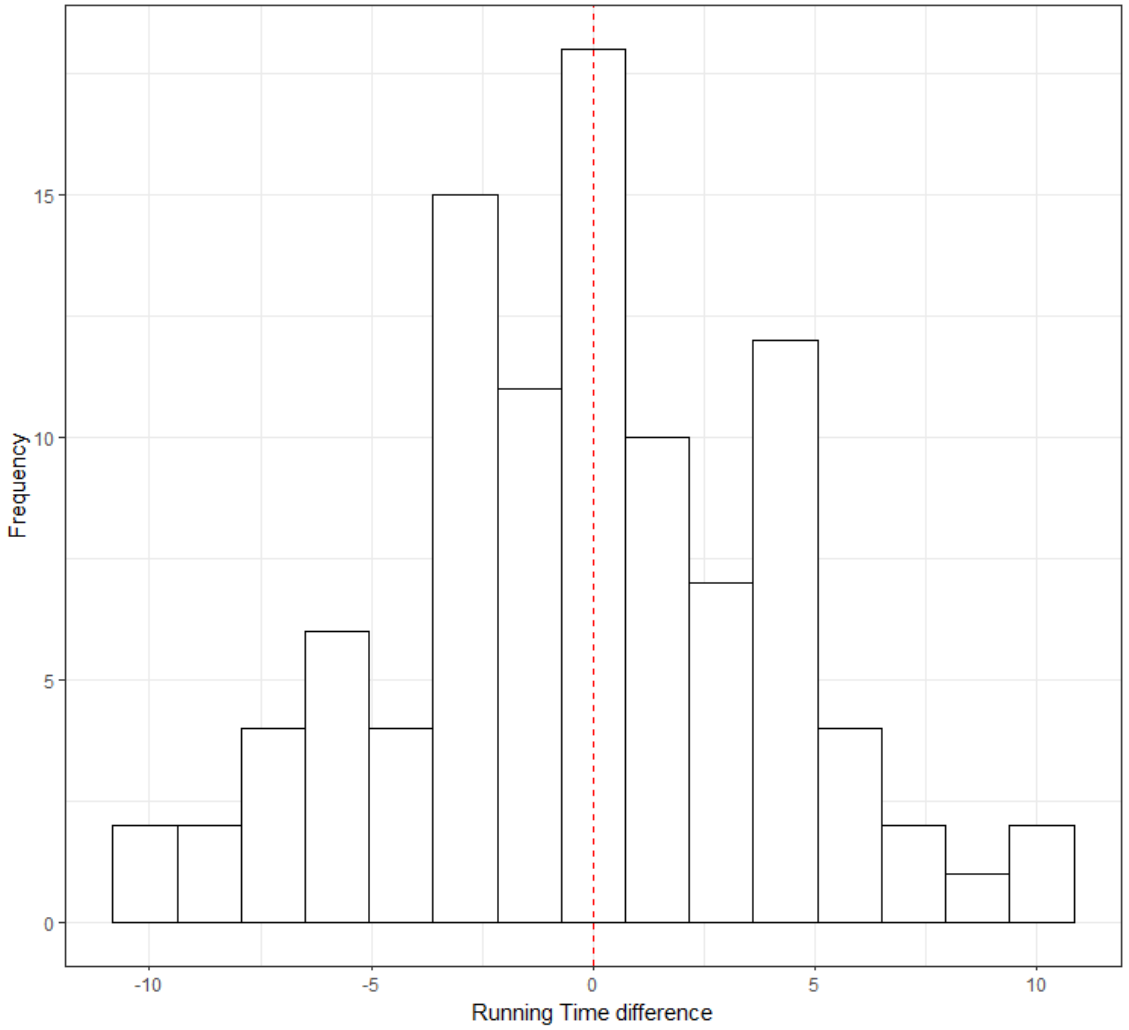


T-test – Paired

- We want to test the difference in running times for students before and after a training regime
  - H_0: The difference is zero
  - H_1:The difference is not zero
- The sample size is (N) is 100 students
- The sample average difference in time is $(\bar{X})$ is -0.196 sec
- The sample difference standard deviation (s) is 4.167 sec
- The difference we're comparing against (μ) is 0 sec
- Degrees of freedom are number of pairs (or students) $N - 1 = 99$

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

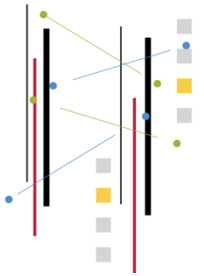
T-test – Paired



T-test – Assumptions

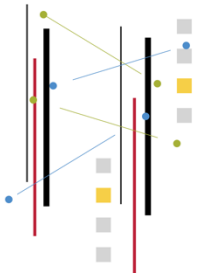
- Independence
- Normality in outcome (Shapiro-Wilk test)

```
> shapiro.test(c(runningTimeBefore-runningTimeAfter))

        Shapiro-Wilk normality test

data:  c(runningTimeBefore - runningTimeAfter)
W = 0.99308, p-value = 0.8928
```



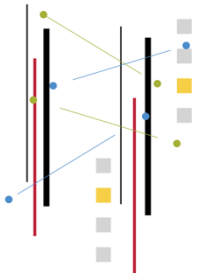
T-test – One sample - in R

```
> t.test(runningTimeBefore,runningTimeAfter, paired = TRUE)

        Paired t-test

data:  runningTimeBefore and runningTimeAfter
t = -0.47151, df = 99, p-value = 0.6383
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -1.0234082  0.6304104
sample estimates:
mean difference
     -0.1964989
```



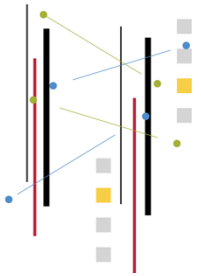
T-test – One sample

- What if the normality assumption does not hold?
  - T-test is relatively robust to normality violations if the sample size is “large enough”
  - If the distribution is relatively symmetrical we can get away with the violation
  - Data can be transformed to make it more “normal” but it’s a trial and error case
- Non-parametric tests exist…
  - They do not make an assumption of normality
  - They are more conservative in their approach - which makes them more reliable when we really doubt the normality assumption



One-sample Wilcoxon

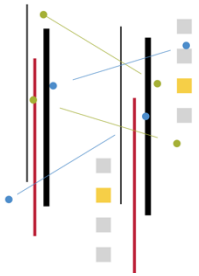
- You can view this as the non parametric equivalent to a one-sample t-test
- In this case we are comparing our sample to a specific median, as opposed to a mean like in the case of the t-test
- The full name of the test is Wilcoxon signed rank test
  - This gives you an indication of its mechanics



One-sample Wilcoxon

A doctor believes that the median number of times she sees her patients in a year is 5. Is she right?

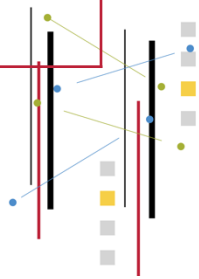
Patient ID	Number of Visits
1	9
2	10
3	8
4	4
5	8
6	3
7	0
8	10
9	15
10	9



One-sample Wilcoxon

The first step is to subtract from every patient the median of the null hypothesis (5 in this case)

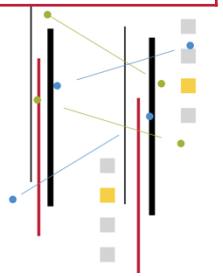
Patient ID	Number of Visits	D=Visits-Median (5)
1	9	4
2	10	5
3	8	3
4	4	-1
5	8	3
6	3	-2
7	0	-5
8	10	5
9	15	10
10	9	4



One-sample Wilcoxon

Next we will rank the absolute value of that difference. In case of ties we will give all of the entries the average rank

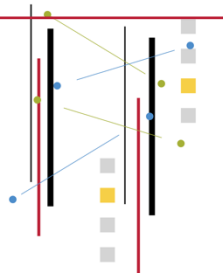
Patient ID	Number of Visits	D=Visits-Median (5)	Rank of|D|
1	9	4	5.5
2	10	5	8
3	8	3	3.5
4	4	-1	1
5	8	3	3.5
6	3	-2	2
7	0	-5	8
8	10	5	8
9	15	10	10
10	9	4	5.5



One-sample Wilcoxon

Finally we will add the sign back to the rank (hence Wilcoxon signed rank test)

Patient ID	Number of Visits	D=Visits-Median (5)	Rank of|D|	Signed rank
1	9	4	5.5	5.5
2	10	5	8	8
3	8	3	3.5	3.5
4	4	-1	1	-1
5	8	3	3.5	3.5
6	3	-2	2	-2
7	0	-5	8	-8
8	10	5	8	8
9	15	10	10	10
10	9	4	5.5	5.5



One-sample Wilcoxon

We can now add together all the positive ranks and all the negative ones:

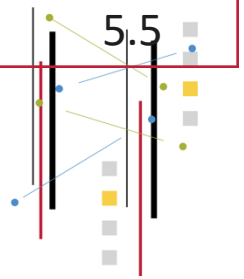
$$+R = 5.5 + 8 + 3.5 + 3.5 + 8 + 10 + 5.5 = 44$$

$$-R = 1 + 2 + 8 = 11$$

$$T_{statistic} = \min(+R, -R) = 11$$

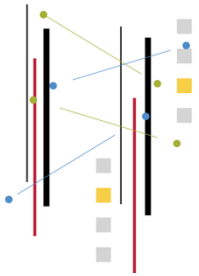
$$p_{value} > 0.05$$

Patient ID	Number of Visits	D=Visits-Median (5)	Rank of|D|	Signed rank
1	9	4	5.5	5.5
2	10	5	8	8
3	8	3	3.5	3.5
4	4	-1	1	-1
5	8	3	3.5	3.5
6	3	-2	2	-2
7	0	-5	8	-8
8	10	5	8	8
9	15	10	10	10
10	9	4	5.5	5.5



One-sample Wilcoxon in R

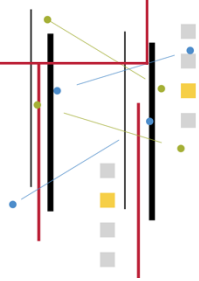
```
> wilcox.test(doctorVisits, mu = 5, exact = FALSE)

        Wilcoxon signed rank test with continuity correction

data:  doctorVisits
V = 44, p-value = 0.1016
alternative hypothesis: true location is not equal to 5
```



Paired Wilcoxon

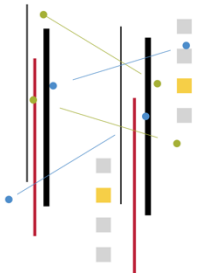
- This would be the non-parametric equivalent to a paired t-test
- It works in the exact same way as the one sample Wilcoxon, but in this case we rank the differences between the samples

Patient ID	Number of Visits in 2016	Number of Visits in 2017
1	9	8
2	10	5
3	8	15
4	4	3
5	8	2
6	3	13
7	0	1
8	10	12
9	15	14
10	9	7



Paired Wilcoxon in R

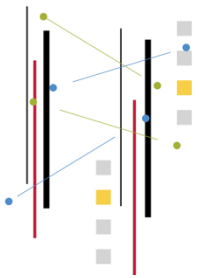
```
> wilcox.test(doctorVisits2016, doctorVisits2017, paired = TRUE, exact=FALSE)

        Wilcoxon signed rank test with continuity correction

data:  doctorVisits2016 and doctorVisits2017
V = 28, p-value = 1
alternative hypothesis: true location shift is not equal to 0
```



T-test – Two samples

- A two sample t-test is when we have two samples without a built-in structure. For example:
  - Comparing the running times between students from two different schools
  - Comparing the effect of a soporific treatment between two groups
- We introduce here the concept of the pooled sample variance (s_p^2).
- Degrees of freedom: Two sample averages have to be calculated for the test, therefore df= N_1 + N_2 -2

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_p^2}{N_1} + \frac{s_p^2}{N_2}}}$$



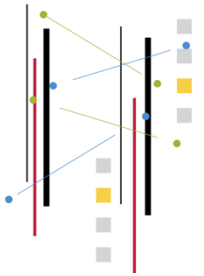
T-test – Two samples

- We want to test the difference in running times for students in two schools
  - H_0: The difference is zero
  - H_1:The difference is not zero
- The sample size for each school (N_1 and N_2) is 100 students
- The sample average for school 1 is ($\overline{X_1}$) is 35.15 sec
- The sample average for school 2 is ($\overline{X_2}$) is 26.05 sec
- The pooled variance is (s_p^2) is 30.38 sec^2
- Degrees of freedom are number of pairs (or students) $100+100-2 = 198$

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_p^2}{N_1} + \frac{s_p^2}{N_2}}}$$

T-test – Two samples - Assumptions

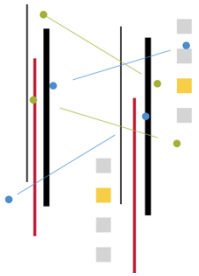
- Independence
- Normality in outcome (Shapiro-Wilk test)
- Homogeneity of variance in both groups (Levene's test or F test)
- T-test robust to departure from homogeneity if sample sizes equal, and to normality if shape of distributions symmetric/roughly equal
  - If variances are not homogeneous, R implements Welch t-test



T-test – Two samples - Assumptions

- Normality in outcome (Shapiro-Wilk test)

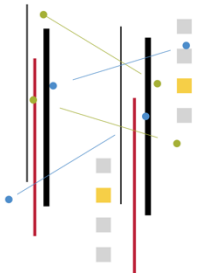
```
> shapiro.test(runningTimeSchool1)

        Shapiro-Wilk normality test

data:  runningTimeSchool1
W = 0.98742, p-value = 0.4664

> shapiro.test(runningTimeSchool2)

        Shapiro-Wilk normality test

data:  runningTimeSchool2
W = 0.98722, p-value = 0.4525
```



T-test – Two samples - Assumptions

- Homogeneity of variance in both groups (Levene's test or F test)

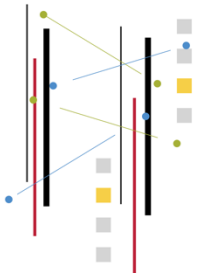
```
> library(car)
> leveneTest(runningTimesSchools$Time, group = runningTimesSchools$School)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   1  6.1877 0.01369 *
      198
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Warning message:
In leveneTest.default(runningTimesSchools$Time, group = runningTimesSchools$School) :
  runningTimesSchools$School coerced to factor.
```



T-test – Two samples - Assumptions

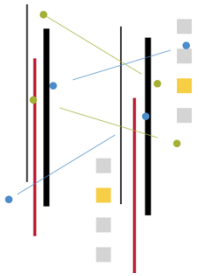
- Normality in outcome (Shapiro-Wilk test)
- Homogeneity of variance (Levene's test)

```
> library(rstatix)
> runningTimesSchools %>% group_by(School) %>% shapiro_test(Time)
# A tibble: 2 × 4
  School   variable statistic     p
  <chr>    <chr>        <dbl> <dbl>
1 School 1 Time         0.987 0.466
2 School 2 Time         0.987 0.452
> runningTimesSchools %>% levene_test(Time ~ School)
# A tibble: 1 × 4
    df1   df2 statistic      p
  <int> <int>     <dbl>  <dbl>
1     1   198      6.19 0.013Z
Warning message:
In leveneTest.default(y = y, group = group, ...) : group coerced to factor.
```



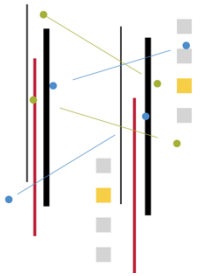
T-test – Two samples in R

```
> t.test(runningTimeSchool1,runningTimeSchool2)

        Welch Two Sample t-test

data:  runningTimeSchool1 and runningTimeSchool2
t = 20.648, df = 187.6, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 8.218340 9.954531
sample estimates:
mean of x mean of y
 35.14192  26.05548
```



Mann-Whitney U test

- The non parametric equivalent to a two sample t-test is the Mann-Whitney test
- It works in a very similar way to the Wilcoxon test, i.e. by ranking samples and then calculating a test statistic from the signed ranks for each group
- Wikipedia has an excellent example:
  - If Aesop decided to increase his sample size for the “Tortoise vs Hare” race, would there be a clear winner?



Tortoise vs Hare…

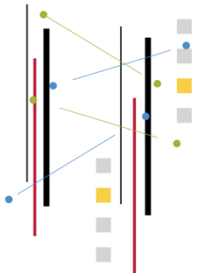
- Suppose we have 6 tortoises and 6 hares
- They cross the finish line in the following order
- We can calculate the statistic for the tortoises by adding up how many hares each tortoise beats
  - The first one beats 6
  - The second one beats 1 and so on
  - This means the statistic for the tortoises is

$$U_T = 6 + 1 + 1 + 1 + 1 + 1 = 11$$

- Similarly:

$$U_H = 5 + 5 + 5 + 5 + 5 + 0 = 25$$

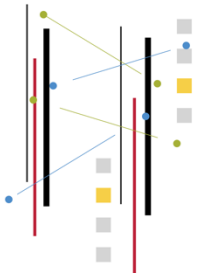
ID	Rank
Tortoise 1	1
Hare 1	2
Hare 2	3
Hare 3	4
Hare 4	5
Hare 5	6
Tortoise 2	7
Tortoise 3	8
Tortoise 4	9
Tortoise 5	10
Tortoise 6	11
Hare 6	12



Tortoise vs Hare…

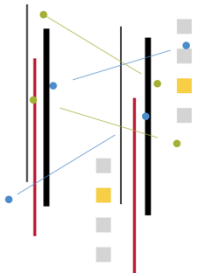
- We use the smallest U score (in this case $U_T = 11$) to find the critical value. Or alternatively:

```
> wilcox.test(Rank ~ Group,data=tortoiseHare)

        Wilcoxon rank sum exact test

data:  Rank by Group
W = 11, p-value = 0.3095
alternative hypothesis: true location shift is not equal to 0
```



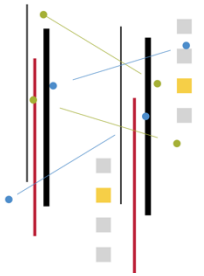
Analysis of Variance (ANOVA)

- Extension of t-test to more than two means
- Allows for the inclusion of categorical independent variables (**factors)**
- **One-way ANOVA**: one continuous dependent variable, one categorical independent variable (factor), with more than two categories/levels
- **Factorial ANOVA**: one continuous dependent variable, two or more categorical independent variables (factors), including all combinations of levels



One-way ANOVA - examples

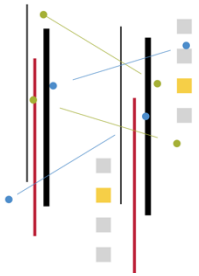
- Does memorising (a list of words) differ between different approaches to memorising (counting, rhyming, imagery…) assigned to 5 different groups of test participants?
- Is there a difference in anxiety levels at exam time between different student groups?
- Is there a difference in plant growth between different fertilisers?



One-way ANOVA – Null hypothesis

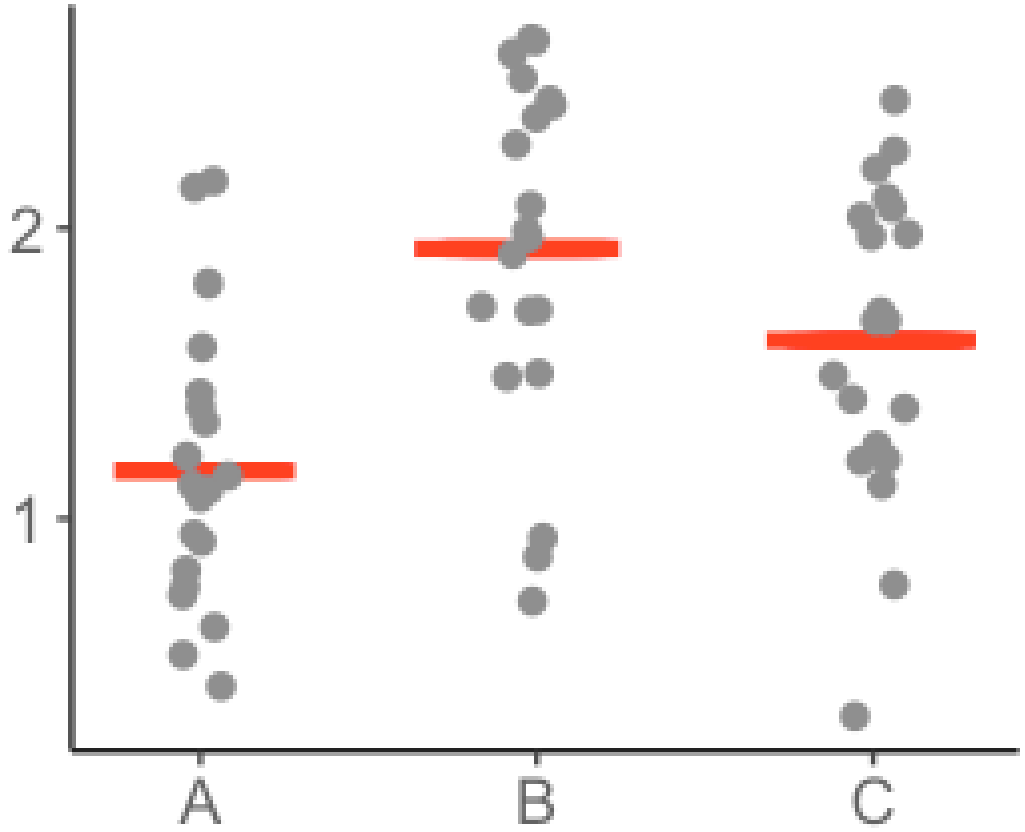
H_0: $\mu_1 = \mu_2 = \mu_3$=… μ_4 against H_1: $\mu_i \neq \mu_j$ for some $i \neq j$

If this is is rejected we don't know which population mean differs, only that at least one is different from at least one other population mean



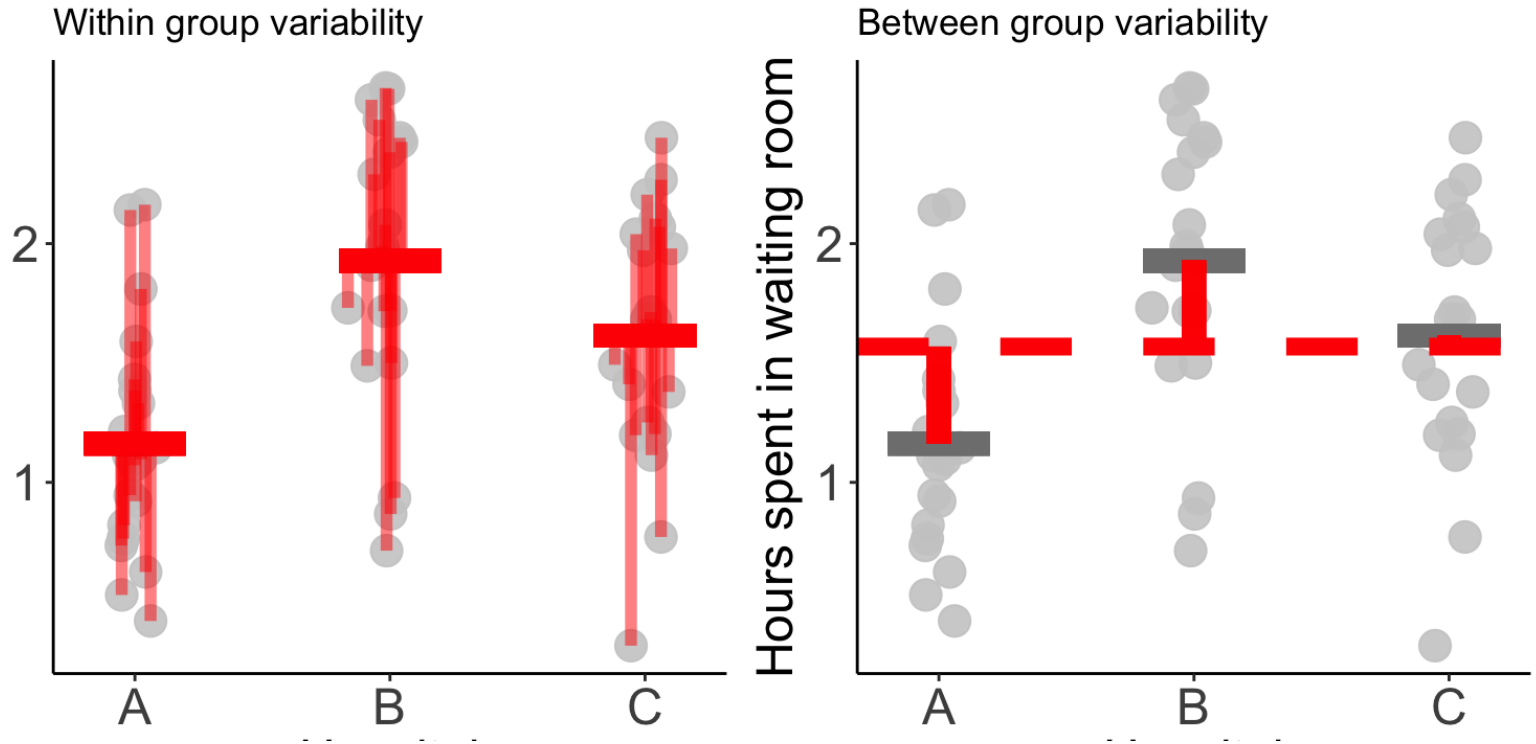
One-way ANOVA

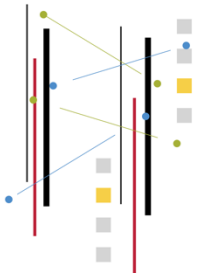
- How are patients responding to three treatments for insomnia?
- Are the differences in mean enough to demonstrate different effect of treatment or is it noise?
- Differences between means, can only be assessed in the context of the overall variability in the data



One-way ANOVA – Between group vs Within group

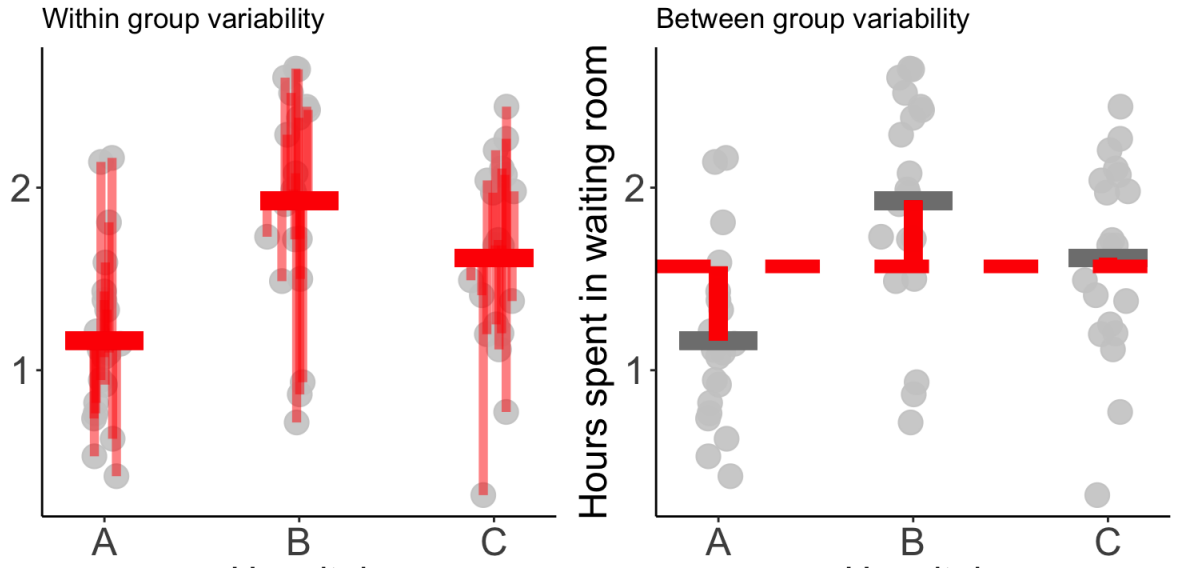
- If variability between groups > variability within groups, there is evidence that population means differ

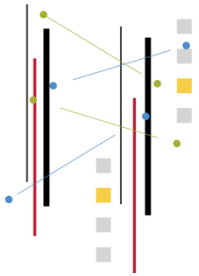




One-way ANOVA – Between group vs Within group

- Compare “between sum of squares” to “within sum of squares”:
  - SS_B: squared distances from the group means to the global mean (average across all subjects)
  - SS_W: squared distances from each observation to group means; also called **residual variability or SS error**
  - SS_T: squared distances from each observation to the overall mean

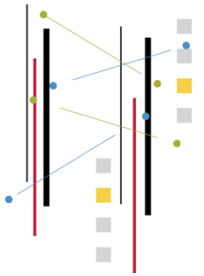




One-way ANOVA – Test statistic

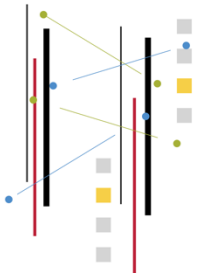
- Test statistic follows F distribution
  - k=number of groups
  - N=number of individual observations
- Assumptions
  - Samples are independent
  - Data are normally distributed
  - Each group has the same population variance
  - No significant outliers

$$F = \frac{\frac{SS_B}{k-1}}{\frac{SS_W}{N-k}}$$



One-way ANOVA – Testing assumptions

- Normality:
  - Boxplot of residuals by group, Q-Q plot
  - Shapiro-Wilk test
- Homogeneity of variances:
  - Boxplot of residuals by group, residuals against fitted
  - Most commonly used hypothesis tests: Levene's test



One-way ANOVA

Is dose level of vitamin C significantly associated with odontoblast length in guinea pigs?

ToothGrowth {datasets} R Documentation

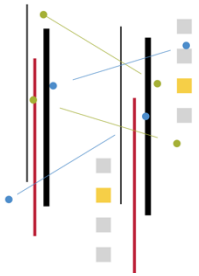
The Effect of Vitamin C on Tooth Growth in Guinea Pigs

Description

The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC).

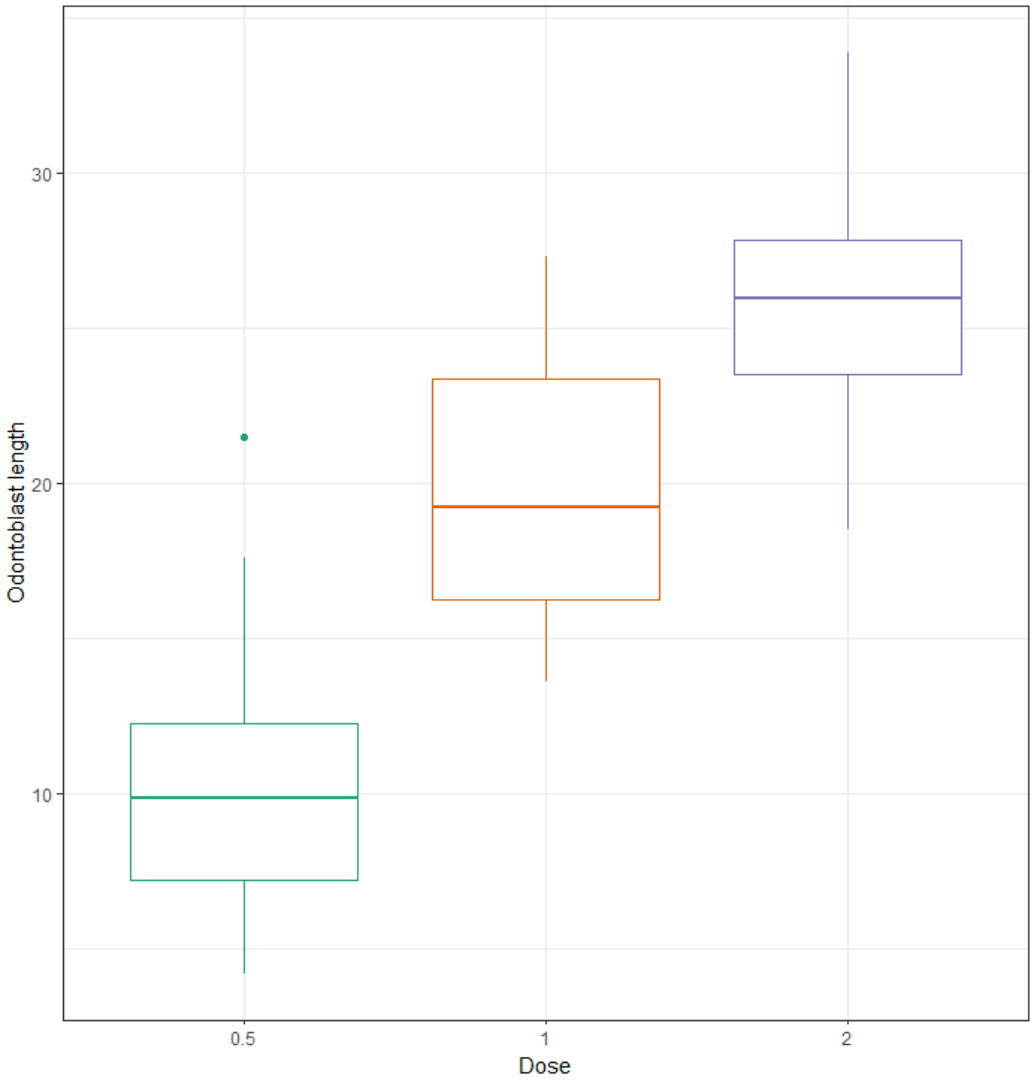
Usage

```
ToothGrowth
```



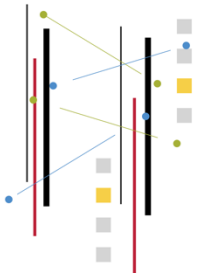
One-way ANOVA

Is dose level of vitamin C significantly associated with odontoblast length in guinea pigs?



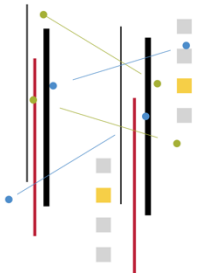
One-way ANOVA – Assumptions

```
> ToothGrowth %>% group_by(dose) %>% shapiro_test(len)
# A tibble: 3 × 4
   dose variable statistic     p
  <dbl> <chr>        <dbl> <dbl>
1   0.5 len          0.941 0.247
2   1   len          0.931 0.164
3   2   len          0.978 0.902
> ToothGrowth %>% levene_test(len~as_factor(dose))
# A tibble: 1 × 4
    df1   df2 statistic     p
  <int> <int>     <dbl> <dbl>
1     2    57     0.646 0.528
```



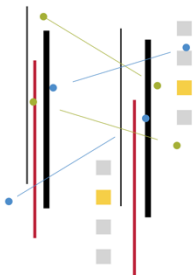
One-way ANOVA in R

```
> library(rstatix)
> ToothGrowth %>% anova_test(len~as_factor(dose))
Coefficient covariances computed by hccm()
ANOVA Table (type II tests)

            Effect DFn DFd      F         p p<.05   ges
1 as_factor(dose)   2  57 67.416 9.53e-16     * 0.703
```



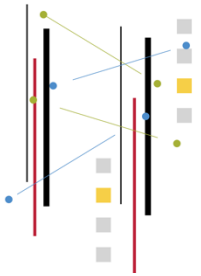
Test	What does it test?	Advantages	Disadvantages
t-tests on all pairs	All pairwise contrasts both simple and complex.	• Simple to run on a computer or hand calculate; • Widely available; • Powerful; • May be used with unequal sized groups.	• Alpha inflation; • Multiple Type I errors; • Unreliable results due to overestimation of differences among pairs.
Tukey	All possible simple contrasts.	• Useful in confirmatory research when combinations of groups is not meaningful; • Available in many statistical packages; • Reduces risk of Type I errors; • May be used when group sizes are unequal.	• Does not test complex contrasts; • Subject to Type II errors and not as powerful as other tests; • Not ideal for exploratory studies; • Not as available as Scheffee or Bonferroni.
Newman-Keuls	All possible simple contrasts.	• More powerful than the Tukey method; • Available in some statistical packages; • Reduces risk of Type II errors; • More likely to find small but significant differences.	• Does not test complex contrasts; • Requires equal group sizes; • Subject to Type 1 errors; • Availability is variable.
Scheffee	Tests of all possible contrasts, both simple and complex.	• Good for both exploratory data analysis and for testing well developed theories; • Can test pairs consisting of combinations of original study groups; • Relatively powerful test; • No need to define contrasts in advance; • Available in many statistical packages; • Reduced risk of Type II errors.	• Alpha inflation higher than for other Multiple Comparison Analysis (MCA) statistics; • Requires equal group sizes; • Tests contrasts not of interest; • More subject to Type 1 errors than other MCA statistics.
Bonferroni	Tests selected contrasts, both simple and complex.	• Preserves alpha; • Can test differences among experimental groups as well as between experimental and control groups; • Available in many statistical packages.	• Groups must be equal in size; • All contrasts must be defined by researcher; • Not used in exploratory studies.
Dunnett	Contrast of control group with each experimental group or combinations of experimental groups. Used when ANOVA has rejected the hypothesis of equality of means.	• Powerful. Good for finding small differences between experimental and control groups; • Specifically tests the experimental groups directly against the control group and thus those differences are more clearly specified.	• Not widely available; • Does not test differences among experimental groups; • Not ideal for exploratory statistical studies.

McHugh, M. Multiple comparison analysis testing in ANOVA; Biochemia Medica 2011;21(3):203–9



One-way ANOVA in R

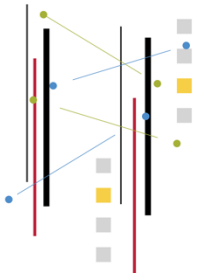
```
> library(rstatix)
> ToothGrowth %>% tukey_hsd(len~as_factor(dose))
# A tibble: 3 × 9
  term             group1 group2 null.value estimate conf.low conf.high    p.adj p.adj.signif
* <chr>            <chr>  <chr>       <dbl>    <dbl>    <dbl>     <dbl>    <dbl> <chr>
1 as_factor(dose) 0.5    1               0     9.13     5.90      12.4   2    e- 8 ****
2 as_factor(dose) 0.5    2               0    15.5     12.3       18.7   1.12e-11 ****
3 as_factor(dose) 1      2               0     6.36     3.14       9.59  4.25e- 5 ****
```



Non-parametric equivalent

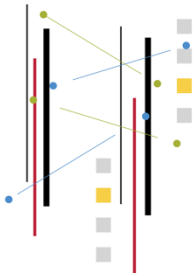
- When the assumptions fail we can use a non-parametric equivalent for a One-way ANOVA such as the Kruskal-Wallis test
- It is essentially an extension of the Wilcoxon rank test to work on multiple groups

```
> library(rstatix)
> ToothGrowth %>% kruskal_test(len~dose)
# A tibble: 1 × 6
  .y.       n statistic    df             p method
* <chr> <int>     <dbl> <int>         <dbl> <chr>
1 len      60      40.7     2 0.000000001148 Kruskal-Wallis
```



Post-hoc test

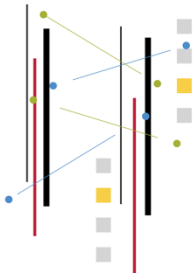
```
> library(rstatix)
> ToothGrowth %>% dunn_test(len~dose)
# A tibble: 3 x 9
  .y.   group1 group2    n1    n2 statistic        p    p.adj p.adj.signif
* <chr> <chr>  <chr>  <int> <int>     <dbl>    <dbl>    <dbl> <chr>
1 len   0.5    1         20    20      3.55 3.78e- 4 7.56e- 4 ***
2 len   0.5    2         20    20      6.36 1.98e-10 5.95e-10 ****
3 len   1      2         20    20      2.81 4.99e- 3 4.99e- 3 **
```



Comparing tests

```
> library(rstatix)
> ToothGrowth %>% anova_test(len~as_factor(dose))
Coefficient covariances computed by hccm()
ANOVA Table (type II tests)

           Effect DFn DFd      F        p p<.05   ges
1 as_factor(dose)   2  57 67.416 9.53e-16     * 0.703
```

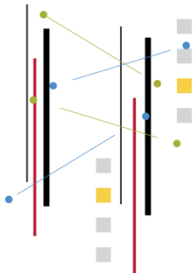
```
> library(rstatix)
> ToothGrowth %>% kruskal_test(len~dose)
# A tibble: 1 × 6
  .y.       n statistic    df             p method
* <chr> <int>     <dbl> <int>         <dbl> <chr>
1 len      60      40.7     2 0.000000000148 Kruskal-Wallis
```



Comparing tests

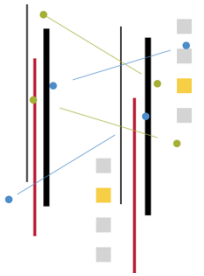
```
> library(rstatix)
> ToothGrowth %>% tukey_hsd(len~as_factor(dose))
# A tibble: 3 × 9
  term           group1 group2 null.value estimate conf.low conf.high    p.adj p.adj.signif
* <chr>          <chr>  <chr>       <dbl>    <dbl>    <dbl>     <dbl>    <dbl> <chr>
1 as_factor(dose) 0.5    1               0     9.13     5.90      12.4  2   e- 8 ****
2 as_factor(dose) 0.5    2               0    15.5     12.3       18.7  1.12e-11 ****
3 as_factor(dose) 1      2               0     6.36     3.14       9.59 4.25e- 5 ****
```

```
> library(rstatix)
> ToothGrowth %>% dunn_test(len~dose)
# A tibble: 3 × 9
  .y.   group1 group2    n1    n2 statistic        p    p.adj p.adj.signif
* <chr> <chr>  <chr>  <int> <int>     <dbl>    <dbl>    <dbl> <chr>
1 len   0.5    1         20    20      3.55 3.78e- 4 7.56e- 4 ***
2 len   0.5    2         20    20      6.36 1.98e-10 5.95e-10 ****
3 len   1      2         20    20      2.81 4.99e- 3 4.99e- 3 **
```

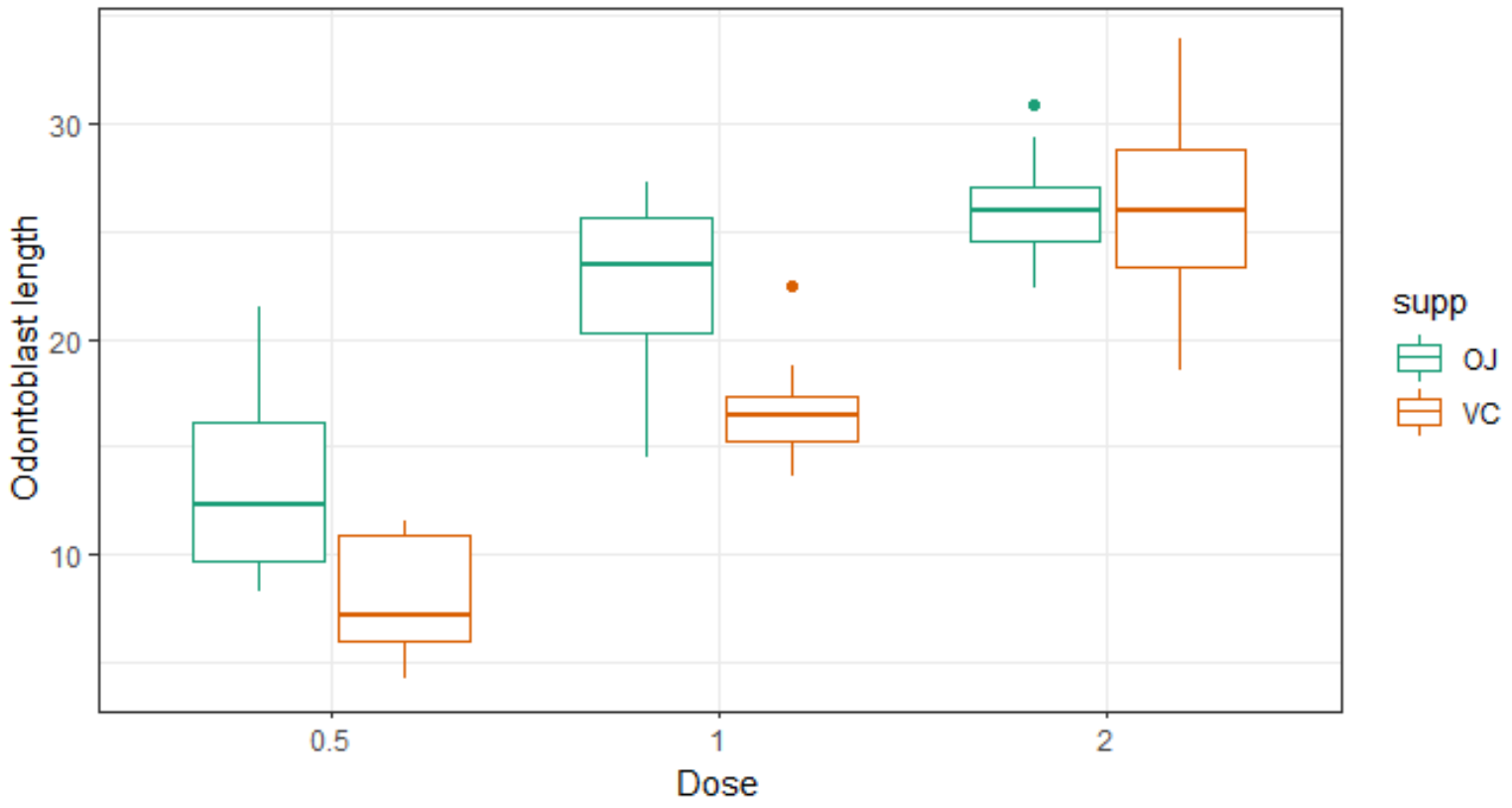


Factorial ANOVA

- More than one independent variable that may influence outcome
- We can “control” for the effect of the second variable
- We can run a factorial design, where we include all combinations of the levels of the independent variables.
- Null hypotheses
  - There is no difference in the means of factor A
  - There is no difference in means of factor B
  - There is no interaction between factors A and B

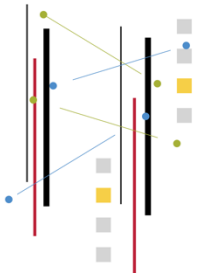


Factorial ANOVA – Guinea pigs



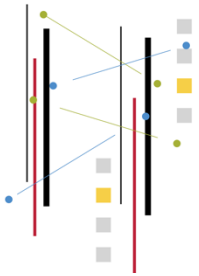
Factorial ANOVA – Assumptions

```
> library(rstatix)
> ToothGrowth %>% group_by(dose, supp) %>% shapiro_test(len)
# A tibble: 6 × 5
  supp   dose variable statistic     p
  <fct> <dbl> <chr>        <dbl> <dbl>
1 OJ      0.5 len          0.893 0.182
2 VC      0.5 len          0.890 0.170
3 OJ      1   len          0.927 0.415
4 VC      1   len          0.908 0.270
5 OJ      2   len          0.963 0.815
6 VC      2   len          0.973 0.919
> ToothGrowth %>% levene_test(len~as_factor(dose)*supp)
# A tibble: 1 × 4
    df1   df2 statistic     p
  <int> <int>     <dbl> <dbl>
1     5    54      1.71 0.148
```



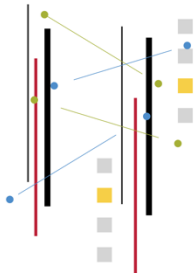
Factorial ANOVA in R

```
> ToothGrowth %>% anova_test(len~as_factor(dose)*supp)
Coefficient covariances computed by hccm()
ANOVA Table (type II tests)

                 Effect DFn DFd      F        p p<.05   ges
1      as_factor(dose)   2  54 92.000 4.05e-18     * 0.773
2                 supp   1  54 15.572 2.31e-04     * 0.224
3 as_factor(dose):supp   2  54  4.107 2.20e-02     * 0.132
```



Factorial ANOVA in R

```
> ToothGrowth %>% tukey_hsd(len~as_factor(dose)*supp)
# A tibble: 19 × 9
   term                  group1 group2 null.value estimate conf.low conf.high    p.adj p.adj.…¹
 * <chr>                 <chr>  <chr>       <dbl>    <dbl>    <dbl>     <dbl>    <dbl> <chr>
 1 as_factor(dose)       0.5    1               0   9.13       6.36     11.9    3.55e-10 ****
 2 as_factor(dose)       0.5    2               0  15.5       12.7      18.3    4.38e-13 ****
 3 as_factor(dose)       1      2               0   6.36       3.60      9.13   2.71e- 6 ****
 4 supp                  OJ     VC              0  -3.70      -5.58     -1.82   2.31e- 4 ***
 5 as_factor(dose):supp  0.5:OJ 1:OJ            0   9.47       4.67     14.3    4.61e- 6 ****
 6 as_factor(dose):supp  0.5:OJ 2:OJ            0  12.8        8.03     17.6    2.13e- 9 ****
 7 as_factor(dose):supp  0.5:OJ 0.5:VC          0  -5.25     -10.0      -0.452  2.43e- 2 *
 8 as_factor(dose):supp  0.5:OJ 1:VC            0   3.54      -1.26      8.34   2.64e- 1 ns
 9 as_factor(dose):supp  0.5:OJ 2:VC            0  12.9        8.11     17.7    1.77e- 9 ****
10 as_factor(dose):supp  1:OJ   2:OJ            0   3.36      -1.44      8.16   3.19e- 1 ns
11 as_factor(dose):supp  1:OJ   0.5:VC          0 -14.7      -19.5      -9.92   2.99e-11 ****
12 as_factor(dose):supp  1:OJ   1:VC            0  -5.93     -10.7      -1.13   7.39e- 3 **
13 as_factor(dose):supp  1:OJ   2:VC            0   3.44      -1.36      8.24   2.94e- 1 ns
14 as_factor(dose):supp  2:OJ   0.5:VC          0 -18.1      -22.9     -13.3    4.86e-13 ****
15 as_factor(dose):supp  2:OJ   1:VC            0  -9.29     -14.1      -4.49   6.91e- 6 ****
16 as_factor(dose):supp  2:OJ   2:VC            0   0.0800    -4.72      4.88   1    e+ 0 ns
17 as_factor(dose):supp  0.5:VC 1:VC            0   8.79       3.99     13.6    2.1 e- 5 ****
18 as_factor(dose):supp  0.5:VC 2:VC            0  18.2       13.4      23.0    4.82e-13 ****
19 as_factor(dose):supp  1:VC   2:VC            0   9.37       4.57     14.2    5.77e- 6 ****
# … with abbreviated variable name ¹p.adj.signif
```



Non-parametric equivalent

- There is no obvious non-parametric equivalent to the two way ANOVA
- Some argue that the Friedman test would be a good option but this is controversial
- In these situations it may be worth considering more complex techniques such as generalised models