Lab 03A: Week 8 (Solutions)

Contents

The **specific aims** of this lab are:

- understand the issues around multiple testing and how to overcome these issues
- practice performing (one-way) ANOVA and interpreting the output
- check ANOVA assumptions

The unit **learning outcomes** addressed are:

- LO1 Formulate domain/context specific questions and identify appropriate statistical analysis.
- LO2 Extract and combine data from multiple data resources.
- LO3 Construct, interpret and compare numerical and graphical summaries of different data types including large and/or complex data sets.
- LO5 Identify, justify and implement appropriate parametric or non-parametric statistical tests.
- LO6 Formulate, evaluate and interpret appropriate linear models to describe the relationships between multiple factors.
- LO8 Create a reproducible report to communicate outcomes using a programming language.

1 Quick quiz

1.1 Multiple testing

1.1 Multiple testing

I generate 4 samples of size $n = 20$ from a $N(3, 1)$ distribution and for each of these 4 samples, I test the null hypothesis $H_0 : \mu = 3$ against the alternative $H_1 : \mu \neq 3$ using $\alpha = 0.1$ as my level of significance.

What's the probability that I falsely reject **at least one** of the four null hypotheses?

a. 0.1

b. 0.3439

c. 0.4

d. 0.0001

e. 0

b. $1 - (1 - \alpha)^m$ where α is the level of significance and m are the number of tests performed.

1.2 Checking for normality

Which of these is the best (graphical) method to determine whether a set of data come from a normal distribution?

a. histogram

b. boxplot

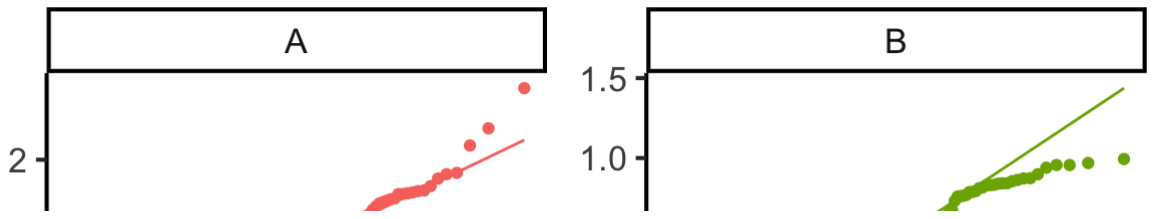
c. normal quantile (QQ) plot

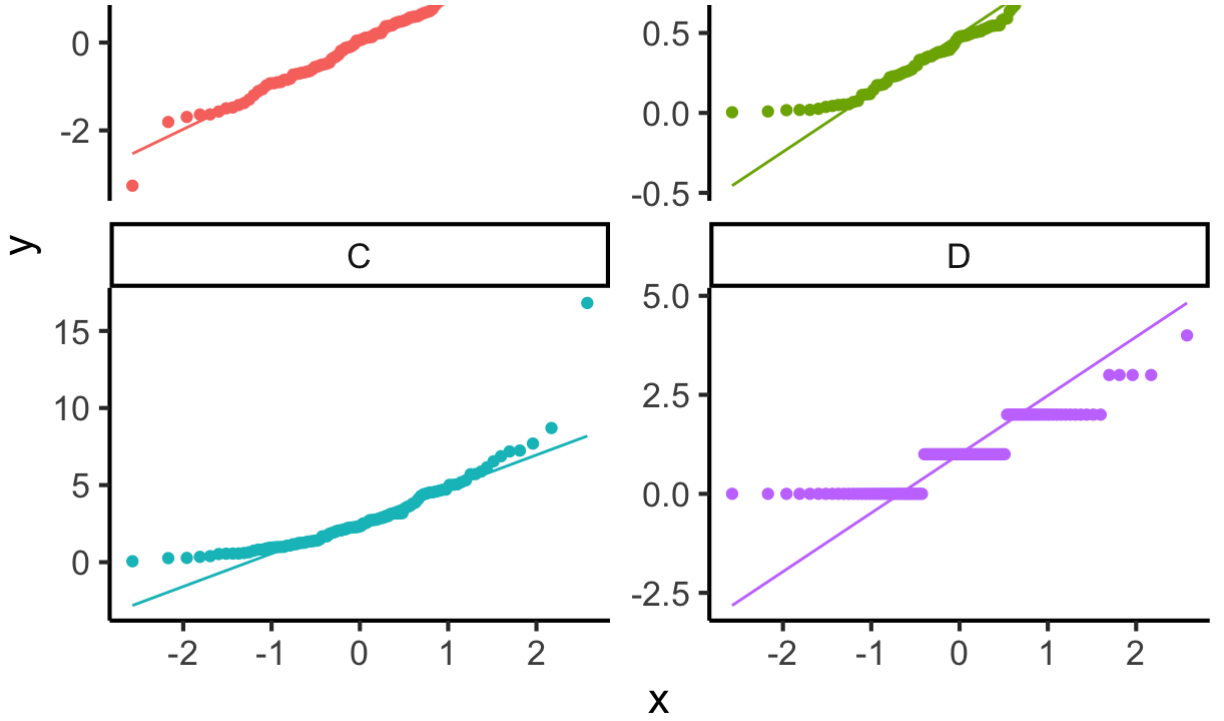
d. all of the above

c.

1.3 Identifying normal

Which of these normal QQ plots indicate that the data come from a normal population?





a.

1.4 Which test?

1. Which method of inference is most useful for testing whether average adult body temperature is 37 degrees Celsius?

a. 1-sample t-test

b. 2-sample t-test

c. paired t-test

d. ANOVA

a.

2. Which method of inference is most useful for testing for a difference in the average scores of 2 judges who judge the same set of dives?

a. 1-sample t-test

b. 2-sample t-test

c. paired t-test

d. ANOVA

c.

3. Which method of inference is most useful for testing whether house prices in Sydney are more expensive than in Melbourne?

a. 1-sample t-test

b. 2-sample t-test

c. paired t-test

d. ANOVA

b.

4. To compare the average leaf thickness of 3 different species of conifers, a scientist randomly samples 10 leaves from each species. What hypothesis test is appropriate to determine if there is a difference among the mean leaf thicknesses of the 3 species?

a. 1-sample t-test

b. 2-sample t-test

c. paired t-test

d. ANOVA

d.

5. To compare the average leaf thickness of 3 different species of conifers, a scientist randomly samples 10 leaves from each species. What is the distribution of the ANOVA test statistic T, if H_0 is true?

a. $T \sim F_{3,\,30}$

b. $T \sim F_{2,\,30}$

c. $T \sim F_{3,\,27}$

d. $T \sim F_{2,\,27}$

e. $T \sim F_{3,\,10}$

f. $T \sim F_{2,\,10}$

d.

6. To compare the average leaf thickness of 3 different species of conifers, a scientist randomly samples 10 leaves from each species. An ANOVA yields a p-value of 0.007. What is the appropriate conclusion?

a. There is no evidence of a difference among the mean leaf thicknesses of the 3 conifer species.

b. There is strong evidence that at least 2 of the mean leaf thicknesses are different.

c. There is strong evidence that all 3 mean leaf thicknesses are different.

d. There is some evidence that the mean leaf thicknesses are different, but not enough to reject H_0.

b.

2 Group question

The lifetimes of 10 Brand A batteries, 12 Brand B batteries, and 9 Brand C batteries were recorded. A partially filled in ANOVA table is below. What is the value of the test statistic F?

```
           Df Sum Sq Mean Sq F value   Pr(>F)
Brand      __  238.1  ______   _____   ______
Residuals  __  _____  ______
Total      __ 1524.1
```

a. 1.666

b. 2.348

c. 2.592

d. 2.777

c.

3 Questions

3.1 Critical Flicker Frequency

If a light is flickering but at a very high frequency, it appears to not be flickering at all. Thus there exists a “critical flicker frequency” where the flickering changes from “detectable” to “not detectable” and this varies from person to person.

The critical flicker frequency and iris colour for 19 people were obtained as part of a study into the relationship between critical frequency flicker and eye colour. They are given in the file `flicker.txt`

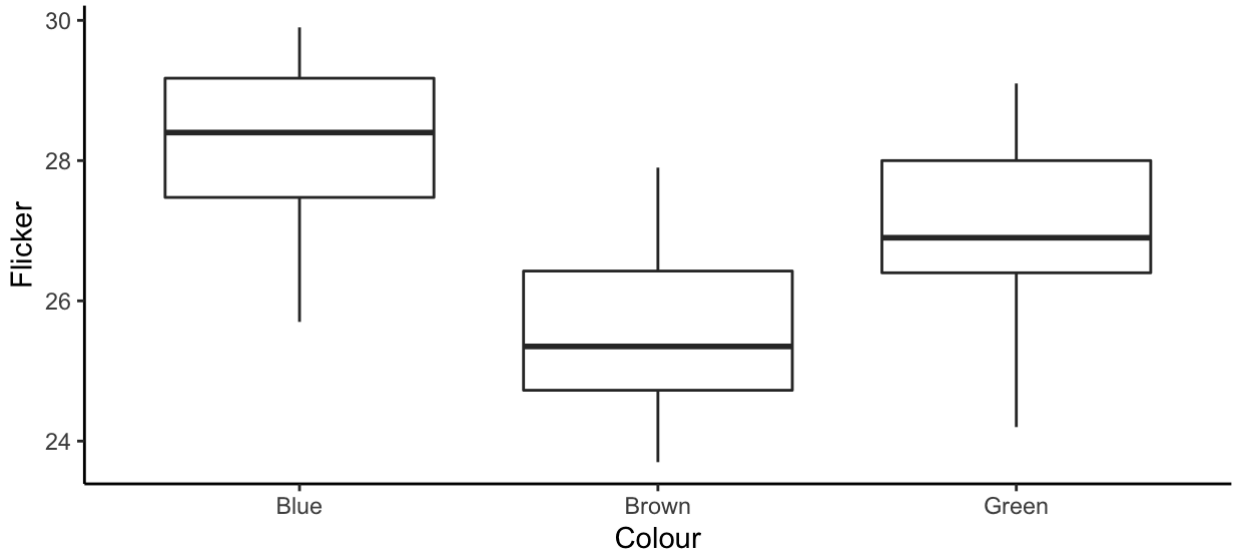
relationship between critical frequency flicker and eye colour. They are given in the file flicker.txt.

```
library(tidyverse)
flicker = read_tsv("https://raw.githubusercontent.com/DATA2002/data/master/flicker.txt")
glimpse(flicker)
```

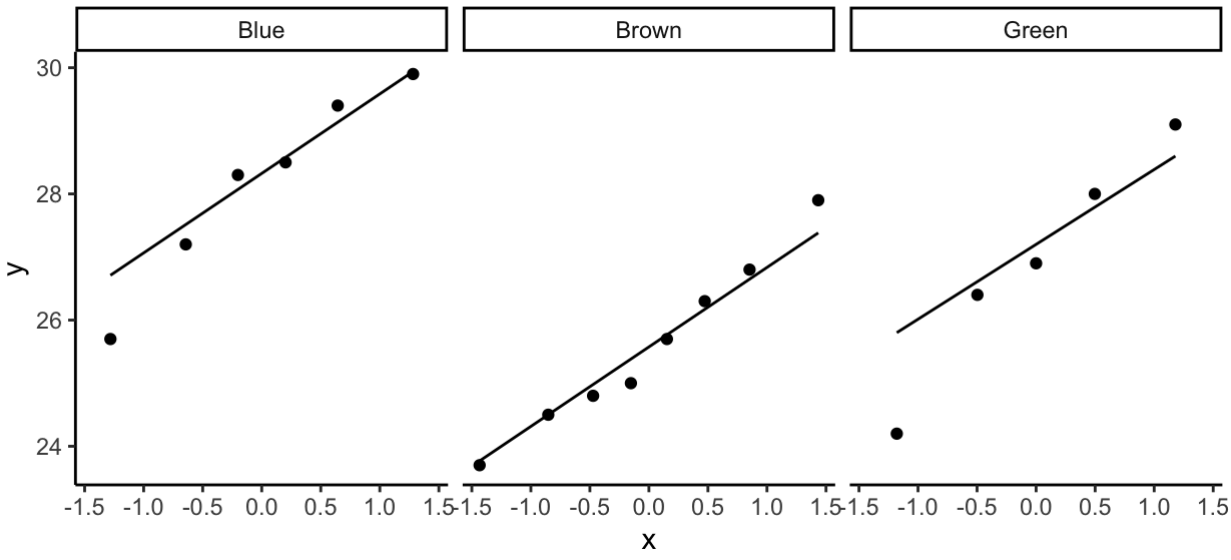
```
Rows: 19
Columns: 2
$ Colour  <chr> "Brown", "Brown", "Brown", "Brown", "Brown", "Brown"…
$ Flicker <dbl> 26.8, 27.9, 23.7, 25.0, 26.3, 24.8, 25.7, 24.5, 26.4…
```

a. Generate side by side boxplots as well as normal QQ plots for the flicker data. Do your plots support the assumptions required for an ANOVA test to be valid? Explain.

```
ggplot(flicker, aes(x = Colour, y = Flicker)) + geom_boxplot() + theme_classic()
```




```
ggplot(flicker, aes(sample = Flicker)) + geom_qq() + geom_qq_line() + facet_wrap(~Colour)
    +
  theme_classic()
```



The QQ-plots look OK, in that the points are reasonably close to the line. The boxplots look symmetric, there are no outliers and they have similar spread. We can conclude that each population looks approximately normal and the equal variance assumption is reasonable.

b. Use the `aov()` function to perform an ANOVA test for the equality of means flicker level across each of the three eye colours.

```
flicker_anova = aov(Flicker ~ Colour, data = flicker)
flicker_anova
```

```
Call:
   aov(formula = Flicker ~ Colour, data = flicker)

Terms:
                   Colour Residuals
Sum of Squares  22.99729  38.31008
Deg. of Freedom        2        16

Residual standard error: 1.547378
Estimated effects may be unbalanced
```

```
summary(flicker_anova)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Colour       2  23.00  11.499   4.802 0.0232 *
Residuals   16  38.31   2.394
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. Using the output, write out the hypothesis test in full. Be sure to state the null and alternative hypothesis, assumptions, test statistic (with distribution), observed test statistic, p-value and an appropriate conclusion.

1. **Hypotheses:** H_0: $\mu_1 = \mu_2 = \mu_3$ vs H_1: at least one $\mu_i \neq \mu_j$.
2. **Assumptions:** Observations are independent within each of the 3 samples. Each of the 3 populations have the same variance, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma$. Each of the 3 populations are normally distributed (or the sample sizes are large enough such that you can rely on the central limit theorem).
3. **Test statistic:** $T = \frac{\text{Treatment Mean Sq}}{\text{Residual Mean Sq}}$. Under H_0, $T \sim F_{g-1,\, N-g}$ where $g = 3$ is the number of groups and $N = 19$ is the total sample size.
4. **Observed test statistic:** $t_0 = \frac{11.499}{2.394} = 4.8$.
5. **p-value:** $P(T \geq t_0) = P(F_{2,16} \geq t_0) = 0.023$.

6. **Decision:** As the p-value is less than 0.05 we reject the null hypothesis and conclude that the population mean flicker sensitivity of at least one eye colour is significantly different to the others.

3.2 Blonds

In an investigation into the relationship between tolerance to pain and hair colour, men and women of various ages were divided into 4 groups based on hair colour and given a pain sensitivity test. Each person's "pain threshold score" (higher score means higher pain tolerance) is recorded in the file `blonds.txt`.

```
pain = read_tsv("https://raw.githubusercontent.com/DATA2002/data/master/blonds.txt")
glimpse(pain)
```

```
Rows: 19
Columns: 2
$ HairColour <chr> "LightBlond", "LightBlond", "LightBlond", "LightB…
$ Pain       <dbl> 62, 60, 71, 55, 48, 63, 57, 52, 41, 43, 42, 50, 4…
```

1. Change `HairColour` so that the ordering is preserved from lightest to darkest. Hint use: `factor()`

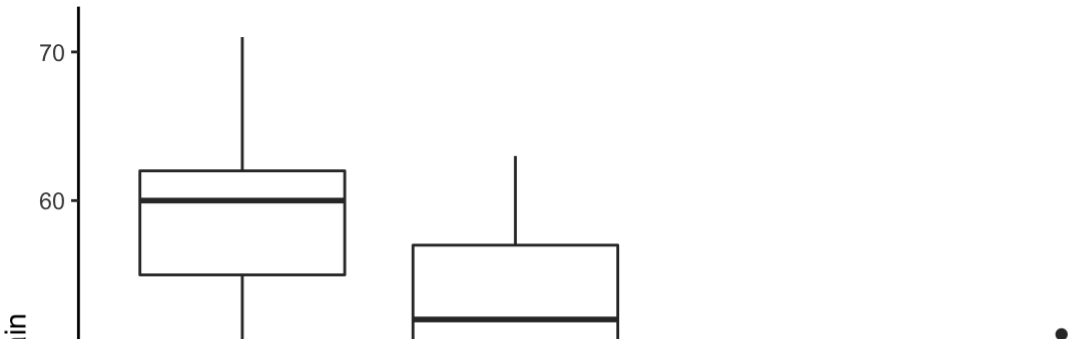
The order of these should be changed from alphabetical, there is a natural ordering, from lighter to darker. This is done as follows:

```
pain = pain %>%
    mutate(HairColour = factor(HairColour, levels = c("LightBlond", "DarkBlond",
        "LightBrunette", "DarkBrunette")))
levels(pain$HairColour)
```

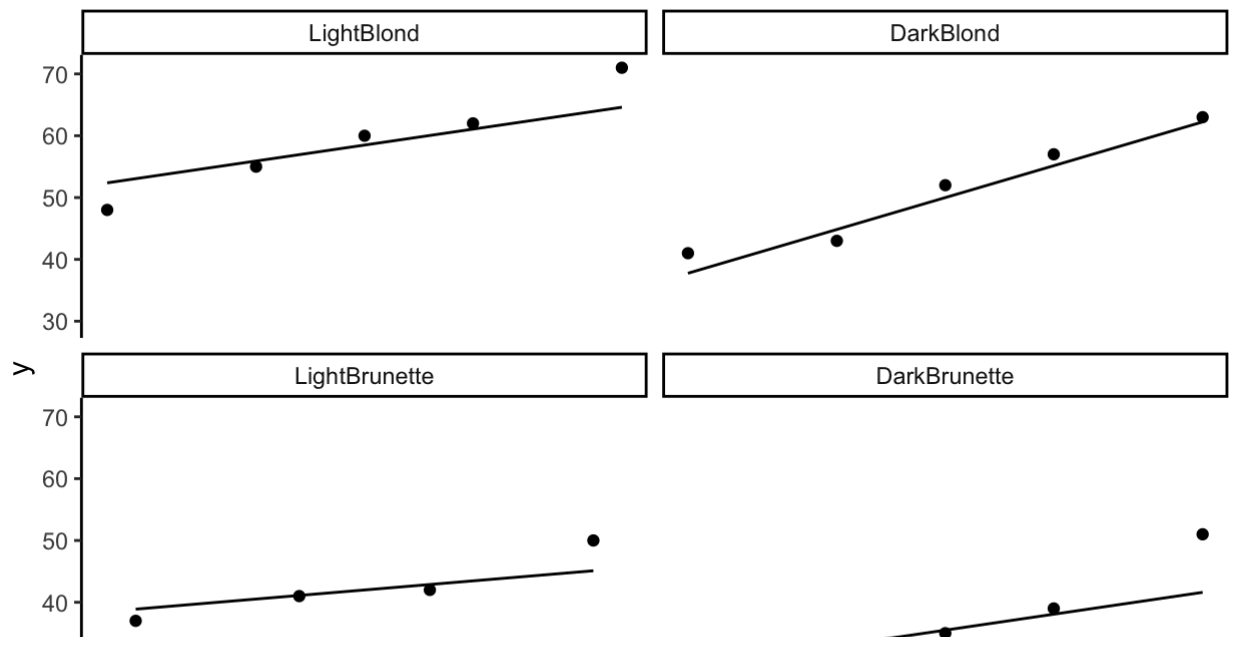
```
[1] "LightBlond"    "DarkBlond"     "LightBrunette" "DarkBrunette"
```

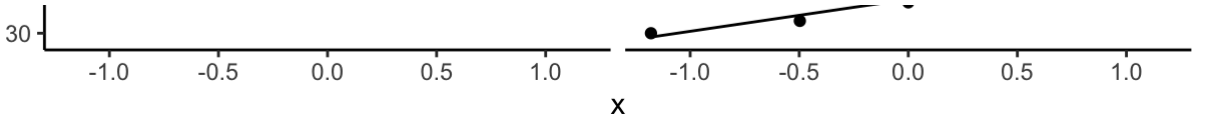
2. Generate boxplots and QQ plots to check the ANOVA assumptions.

```
ggplot(pain, aes(x = HairColour, y = Pain)) + geom_boxplot() + theme_classic()
```




```
ggplot(pain, aes(sample = Pain)) + geom_qq() + geom_qq_line() + facet_wrap(~HairColour) +
    theme_classic()
```





It is hard to say anything conclusive about the ANOVA assumptions with so few observations in the different groups. Should be careful not to read too much into boxplots with so few observations, but the spreads look roughly similar. Also with the QQ-plots, can't be too conclusive because of the low sample size, but the points are all reasonably close to the lines so the normality assumption doesn't appear to be violated.

3. What do the boxplots suggest about the null hypothesis that pain thresholds are the same regardless of hair colour?

A shocking apparent effect! Looks like as hair colour darkens, pain thresholds decrease.

4. Test this hypothesis formally using ANOVA. Does there seem to be a relationship between hair colour and pain threshold?!

```
pain_anova = aov(Pain ~ HairColour, data = pain)
summary(pain_anova)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
HairColour   3   1361   453.6   6.791 0.00411 **
Residuals   15   1002    66.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

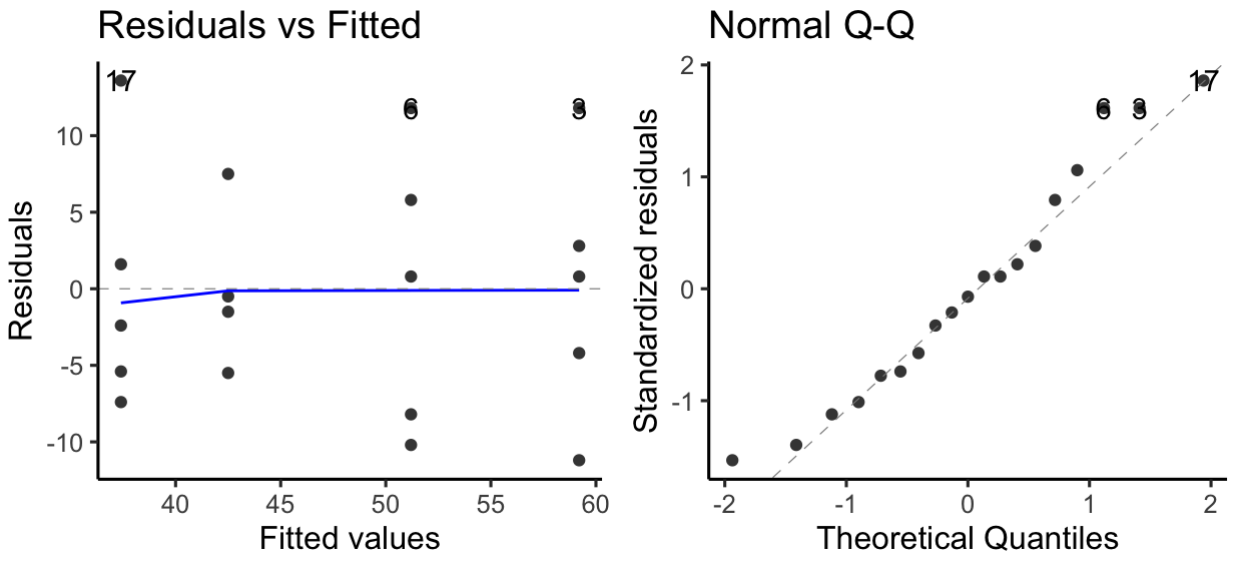
Let μ_1, μ_2, μ_3 and μ_4 be the population mean pain thresholds of light blond, dark blond, light brunette and dark brunette, respectively.

1. **Hypotheses:** H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ vs H_1: at least one $\mu_i \neq \mu_j$.
2. **Assumptions:** Observations are independent within each of the 4 samples. Each of the 4 populations have the same variance, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma$. Each of the 4 populations are normally distributed.
3. **Test statistic:** $T = \frac{\text{Treatment Mean Sq}}{\text{Residual Mean Sq}}$. Under H_0, $T \sim F_{g-1,\, N-g}$ where $g = 4$ and $N = 19$.
4. **Observed test statistic:** $t_0 = \frac{453.6}{66.8} = 6.791$.
5. **p-value:** $P(T \geq 6.791) = P(F_{3,\,15} \geq 6.791) = 0.004$.
6. **Decision:** As the p-value is less than 0.05 we reject the null hypothesis and conclude that the population mean pain threshold of at least one hair colour group is significantly different to the others.

The code and results below will be introduced in more detail in future weeks, but it provides a more

The code and results below will be introduced in more detail in future weeks, but it provides a more overall way of assessing these assumptions. It's a similar idea, looking for roughly constant spread in the "residuals" across the range of "fitted values" and looking to check if the (standardised) residuals lie close to the dashed line in the normal QQ plot. In this case the spread of the residuals looks roughly similar across the range of fitted values (indicating the equal variance assumption is OK) and the points all lie reasonably close to the dashed line in the QQ plot indicating that the normality assumption is well satisfied.

```
library(ggfortify)
autoplot(pain_anova, which = c(1, 2))
```



3.3 Hedenfalk data

The package **sgof** has a data set `Hedenfalk` (Conde and Una Alvarez 2020). You may need to `install.packages("sgof")`. Use `?Hedenfalk` to find out more about this data set. (If you can't install the package, the data can be found here and the help page can be found here.)

a. How many p-values are in the data set?

b. Generate a histogram of the (unadjusted) Hedenfalk p-values.

c. How many (unadjusted) p-values are significant at the 5% level of significance? What proportion of all p-values in the data set is this?

d. Why is it a good idea to consider adjusted p-values?

e. Using `p.adjust()` find the Bonferroni and BH p-values. Plot histograms of each and find the

e. Using `p.adjust()` find the Bonferroni and BH p-values. Plot histograms of each and find the number of significant results after adjustment for both.

f. Comment on the difference between the Bonferroni method and the BH method.

a.

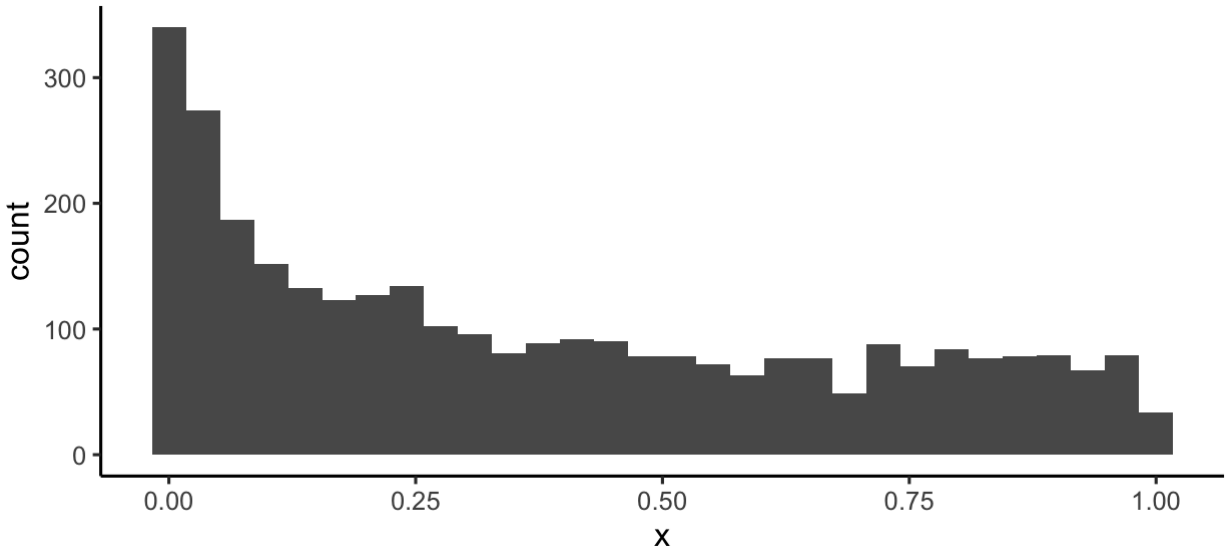
```
# install.packages('sgof')
library(sgof)
# ?Hedenfalk
glimpse(Hedenfalk)
```

```
Rows: 3,170
Columns: 3
$ x      <dbl> 0.0121261800, 0.0750252400, 0.9949211000, 0.041785490…
$ bonf_p <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
$ BH_p   <dbl> 0.13164380, 0.31252301, 0.99712295, 0.24127505, 0.944…
```

```
length(Hedenfalk$x)
```

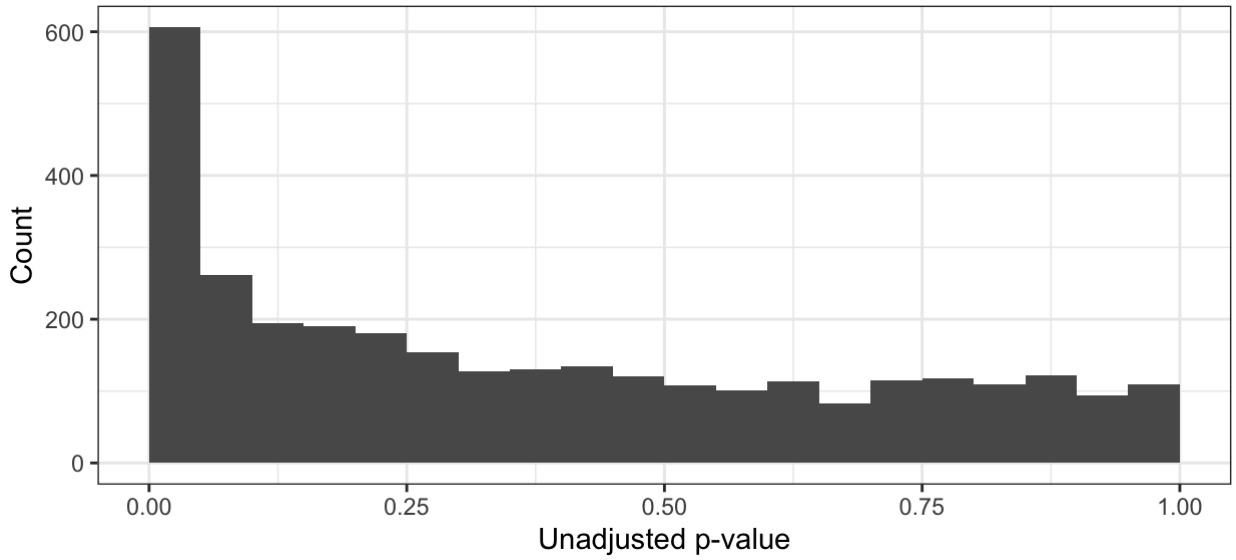
```
[1] 3170
```

```
ggplot(Hedenfalk, aes(x = x)) + geom_histogram()
```



A slightly improved histogram, where the bins don't go below 0 or above 1:

```
ggplot(Hedenfalk, aes(x = x)) + geom_histogram(boundary = 0, binwidth = 0.05) +
    labs(x = "Unadjusted p-value", y = "Count") + theme_bw()
```




```
sum(Hedenfalk$x < 0.05)
```

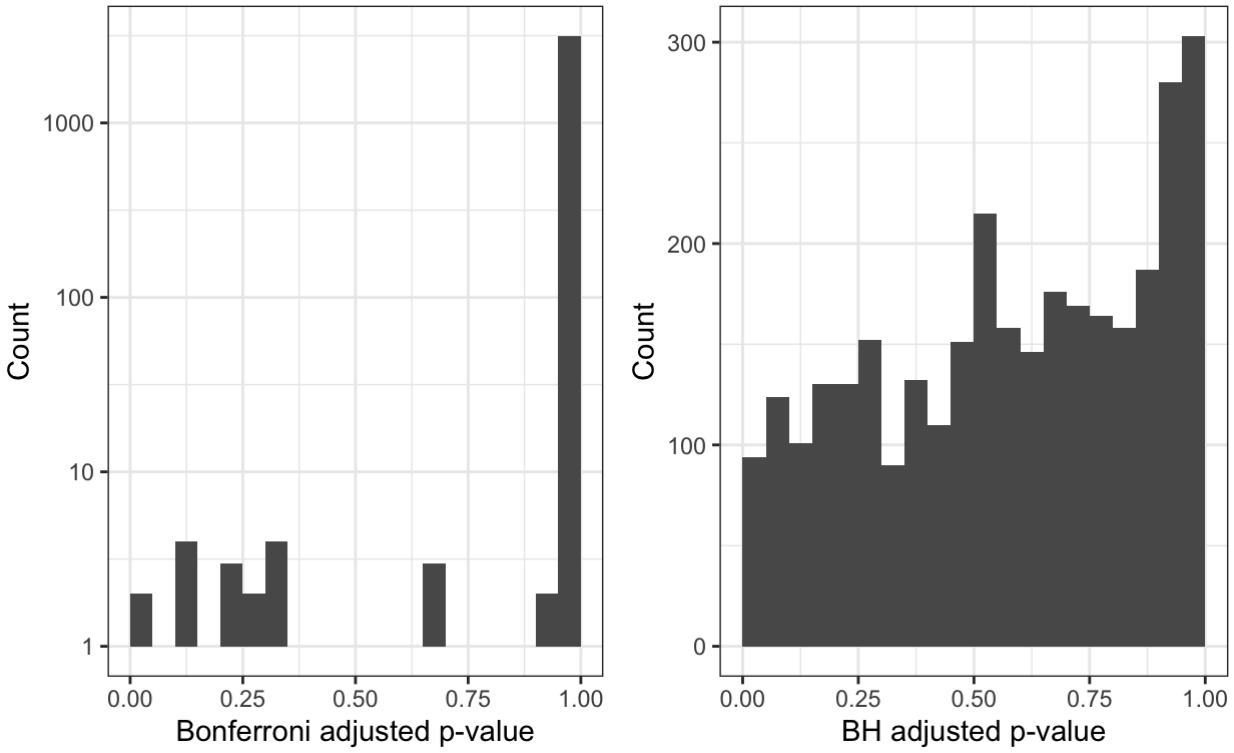
```
[1] 605
```

```
mean(Hedenfalk$x < 0.05)
```

```
[1] 0.1908517
```

With so many tests, it's likely we're seeing a substantial number of false positives. With 3170 p-values, even if **none of them should be rejected** (i.e. even if in reality the null hypothesis is true for all tests), we'd expect to see $3170 \times 0.05 = 158.5$ significant p-values by chance alone.

```
Hedenfalk = Hedenfalk %>%
  mutate(
    bonf_p = p.adjust(x, method = "bonferroni"),
    BH_p = p.adjust(x, method = "BH")
  )
p1 = Hedenfalk %>%
  ggplot() + aes(x = bonf_p) +
  geom_histogram(boundary = 0, binwidth = 0.05) +
  theme_bw() +
  labs(x = "Bonferroni adjusted p-value",
       y = "Count") +
  scale_y_log10()
p2 = Hedenfalk %>%
  ggplot() + aes(x = BH_p) +
  geom_histogram(boundary = 0, binwidth = 0.05) +
  labs(x = "BH adjusted p-value",
       y = "Count") +
  theme_bw()
gridExtra::grid.arrange(p1,p2,ncol=2)
```




```
# sum(Hedenfalk$bonf_p < 0.05) mean(Hedenfalk$bonf_p < 0.05)
# sum(Hedenfalk$BH_p < 0.05) mean(Hedenfalk$BH_p < 0.05)
Hedenfalk %>%
    summarise_at(.vars = vars(bonf_p, BH_p), .funs = list(n_sig = function(x) sum(x <
        0.05), prop_sig = function(x) mean(x < 0.05))) %>%
    kable()
```

bonf_p_n_sig	BH_p_n_sig	bonf_p_prop_sig	BH_p_prop_sig
2	94	0.0006309	0.029653

The Bonferroni method seeks to control the family wise error rate, and can be very conservative. The Benjamini–Hochberg (BH) method looks to control the false discovery rate and tends to allow more false positives. We can see this in the results, where the Bonferroni method finds only two significantly differentially expressed genes whereas the BH procedure identified 94.

4 For practice after the computer lab

From Larsen and Marx (2012) you could work through Case Study 12.2.1 and then consider these questions: 12.2.1, 12.2.2, 12.2.3, 12.2.4, 12.2.5, 12.2.6, 12.2.7, 12.2.8, 12.2.11, 12.2.12 and 12.2.13.

You can also attempt the DataCamp chapter on comparing many means.

References

Conde, Irene Castro, and Jacobo de Una Alvarez. 2020. *Sgof: Multiple Hypothesis Testing*. https://CRAN.R-project.org/package=sgof.

Larsen, Richard J., and Morris L. Marx. 2012. *An Introduction to Mathematical Statistics and Its Applications*. 5th ed. Boston, MA: Prentice Hall.