Lab 03B: Week 9 (Solutions)

Contents

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

- practice performing (one-way) ANOVA and interpreting the output
- check ANOVA assumptions using residuals
- conduct post-hoc testing using contrasts
- perform rank based testing of multiple means (Kruskal-Wallis test)

The unit **learning outcomes** addressed are:

- LO1 Formulate domain/context specific questions and identify appropriate statistical analysis.
- LO3 Construct, interpret and compare numerical and graphical summaries of different data types including large and/or complex data sets.
- 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 Recap: multiple comparisons

1.1 Simultaneous confidence intervals

A 95% confidence interval for a *single* population contrast $\sum_{i=1}^{g} c_i \mu_i$ (where $\sum_{i=1}^{g} c_i = 0$) is of the form

$$\underline{g} \qquad \left(\overline{\underline{g} \quad c^2} \right)$$

$$\sum_{i=1} c_i \bar{y}_{i\cdot} \pm m \left(\hat{\sigma} \sqrt{\sum_{i=1} \frac{c_i}{n_i}} \right)$$

where the **multiplier** m is the upper 2.5% quantile from the t_{N-g} distribution (recall N is the total sample size); the quantity in round brackets is the *standard error*. When the model is correct this procedure "works" 95% of the time in repeated experiments.

However if we are constructing several of these at once, while each one individually may work 95% of the time, having *all of them* "work" (simultaneously) is not guaranteed to the same degree. We can fix this by *increasing the multiplier* m. We have discussed 3 different approaches:

1.1.1 The Bonferroni method

If we are constructing k simultaneous $100(1-\alpha)\%$ confidence intervals, instead of using the upper $\alpha/2$ quantile, we use the upper $\alpha/(2k)$ quantile, i.e. as if we were constructing individual $100(1-\alpha/k)\%$ intervals. This procedure is in general *conservative* i.e. the resultant true confidence level is typically greater than desired (i.e. the multiplier is bigger than it needs to be).

1.1.2 Tukey's method

This provides the *exact* multiplier one needs when

- we are looking at all pairwise comparisons and
- the sample sizes are all the same.

When these two conditions hold, it is the best we can do i.e. it gives the smallest multiplier m that does the job. When the sample sizes are *unequal* it is conservative (although possibly less so than the corresponding Bonferroni multiplier.)

1.1.3 Scheffé's method

This provides the *exact* multiplier one needs when considering *all possible contrasts*, and thus permits "unlimited data snooping". The multiplier is taken from the $\sqrt{(g-1)F}$ distribution, where here F denotes the distribution of the corresponding F-statistic i.e. $F_{g-1,N-g}$. This multiplier is thus conservative when considering only a finite number of contrasts, but again may be smaller than the corresponding Bonferroni multiplier.

1.2 Implementation

The Bonferroni correction is straightforward to implement. But in general, the **emmeans** package provides a convenient way to implement a variety of post hoc tests. The generic code below shows how this might be done for a hypothetical dependent variable `y` and factor variable `group`.

```r
# this code won't actually run, we haven't defined y or group
library(emmeans)
one_way = aov(y ~ group)
one_way_em = emmeans(one_way, ~ group)
one_way_pairs = contrast(one_way_em, method = "pairwise", adjust = "bonferroni")
# alternatively, can use pairs()
# one_way_pairs = pairs(one_way_em, adjust = "bonferroni")
plot(one_way_pairs)
# Tukey's method
contrast(one_way_em, method = "pairwise", adjust = "tukey")
# Scheffe's method:
contrast(one_way_em, method = "pairwise", adjust = "scheffe")
```

2 Questions

2.1 Pain thresholds

Recall the pain/hair colour data. Below we change the factor order from alphabetical to "lightest to darkest":

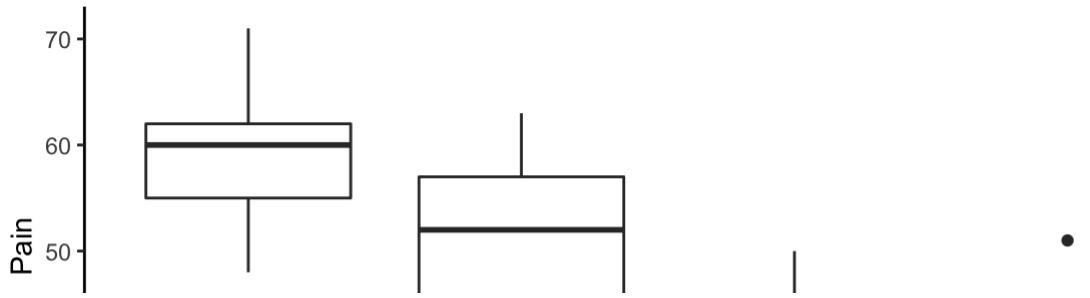
```r
library(tidyverse)
pain = read_tsv("https://raw.githubusercontent.com/DATA2002/data/master/blonds.txt")
glimpse(pain)
```

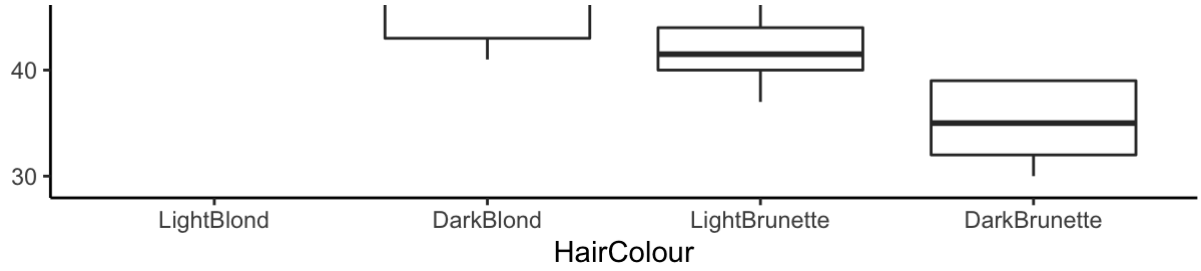
```
Rows: 19
Columns: 2
$ HairColour <chr> "LightBlond", "LightBlond", "LightBlond", "LightB…
$ Pain       <dbl> 62, 60, 71, 55, 48, 63, 57, 52, 41, 43, 42, 50, 4…
```

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

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

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






```
pain_sum = pain %>%
    group_by(HairColour) %>%
    summarise(n = n(), ybar = mean(Pain))
pain_sum
```

```
# A tibble: 4 × 3
  HairColour        n  ybar
  <fct>         <int> <dbl>
1 LightBlond        5  59.2
2 DarkBlond         5  51.2
3 LightBrunette     4  42.5
4 DarkBrunette      5  37.4
```

```
ni = pain_sum %>%
    pull(n)
ybar_i = pain_sum %>%
    pull(ybar)
```

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

```
            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
```

1. Compute the standard error of each pairwise difference (note, there are only two different standard errors over the $\binom{4}{2} = 6$ pairwise differences).

The "Residual Standard Error" is `sig.hat` given below:

```
sig.hat = sqrt(66.8)
sig.hat
```

```
[1] 8.173127
```

Where both sample sizes are 5, the standard error is then

```
se.55 = sig.hat * sqrt(2/5)
se.55
```

```
[1] 5.169139
```

When one sample is 4 and one sample is 5 (i.e. any comparison with LightBrunette), the standard error is

```
se.45 = sig.hat * sqrt((1/5) + (1/4))
se.45
```

```
[1] 5.4827
```

2. Compute t-statistics for all 6 pairwise comparisons. The output below may be useful:

```
diff_mat = outer(ybar_i, ybar_i, "-")
diff_mat
```

```
      [,1]  [,2] [,3] [,4]
[1,]   0.0   8.0 16.7 21.8
[2,]  -8.0   0.0  8.7 13.8
[3,] -16.7  -8.7  0.0  5.1
[4,] -21.8 -13.8 -5.1  0.0
```

We can get a matrix of standard errors using the fancy `outer()` command

```
se.mat = sig.hat * sqrt(outer(1/ni, 1/ni, "+"))
se.mat
```

```
         [,1]     [,2]     [,3]     [,4]
[1,] 5.169139 5.169139 5.482700 5.169139
[2,] 5.169139 5.169139 5.482700 5.169139
[3,] 5.482700 5.482700 5.779273 5.482700
[4,] 5.169139 5.169139 5.482700 5.169139
```

The "ratio" below gives the t-statistics:

```
diff_mat/se.mat
```

```
          [,1]      [,2]       [,3]      [,4]
[1,]  0.000000  1.547646  3.0459445 4.2173366
[2,] -1.547646  0.000000  1.5868094 2.6696902
[3,] -3.045944 -1.586809  0.0000000 0.9301986
[4,] -4.217337 -2.669690 -0.9301986 0.0000000
```

3. Using the output below and the Bonferroni method, determine the appropriate multiplier for constructing 6 simultaneous confidence intervals at both the 95% and 99% confidence levels.

```
upper.tail.area = c(0.05, 0.025, 0.05/6, 0.025/6, 0.01, 0.005, 0.01/6,
    0.005/6)
t.quantile = qt(1 - upper.tail.area, df = 15)
cbind(upper.tail.area, t.quantile)
```

```
cbind(upper.tail.area, t.quantile)
```

```
     upper.tail.area t.quantile
[1,]    0.0500000000   1.753050
[2,]    0.0250000000   2.131450
[3,]    0.0083333333   2.693739
[4,]    0.0041666667   3.036283
[5,]    0.0100000000   2.602480
[6,]    0.0050000000   2.946713
[7,]    0.0016666667   3.483677
[8,]    0.0008333333   3.821973
```

For 95% intervals the multiplier is

```
m = qt(1 - 0.025/6, df = 15)
m
```

```
[1] 3.036283
```

For 99% intervals the multiplier is

```
m = qt(1 - 0.005/6, df = 15)
m
```

```
[1] 3.821973
```

4. Which differences are significant at the

- 5% level
- 1% level

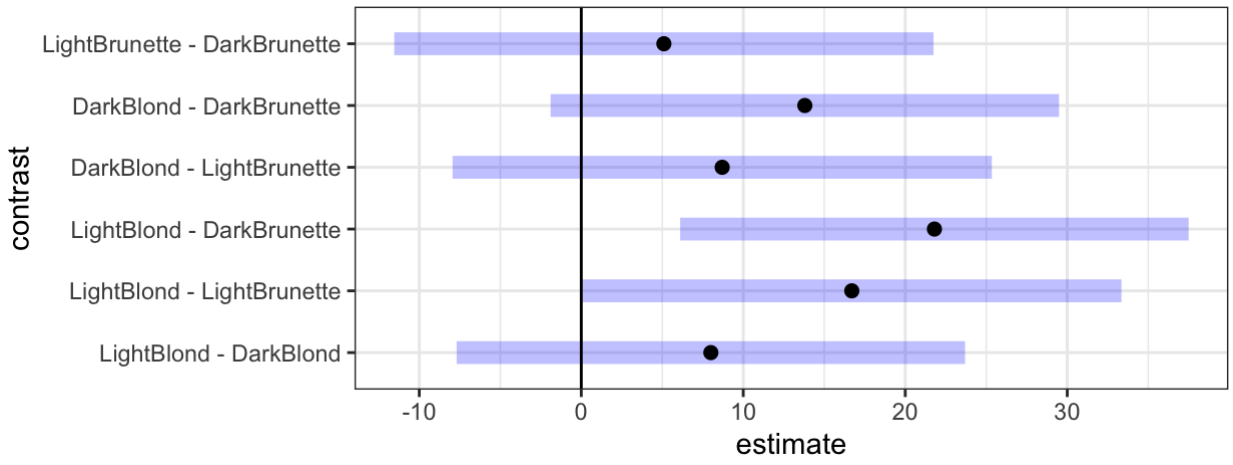
Any t-statistics bigger (in absolute value) than the 95% multiplier are significant at the 5% level. This includes

- `LightBrunette--LightBlond`
- `DarkBrunette--LightBlond`

The latter is also bigger than the 99% multiplier, so it is also significant at the 1% level.

5. Check your answers using the **emmeans** package. Do your conclusions change if you use Tukey's or Scheffe's method?

```
library(emmeans)
pain_em = emmeans(pain_aov, ~HairColour)
# pairs(pain_em, adjust = 'bonferroni')
bonf = contrast(pain_em, method = "pairwise", adjust = "bonferroni")
plot(bonf) + theme_bw() + geom_vline(xintercept = 0)
```




```
contrast(pain_em, method = "pairwise", adjust = "tukey")
```

```
contrast                    estimate   SE df t.ratio p.value
LightBlond - DarkBlond           8.0 5.17 15   1.548  0.4356
LightBlond - LightBrunette      16.7 5.48 15   3.046  0.0366
LightBlond - DarkBrunette       21.8 5.17 15   4.218  0.0037
DarkBlond - LightBrunette        8.7 5.48 15   1.587  0.4147
DarkBlond - DarkBrunette        13.8 5.17 15   2.670  0.0741
LightBrunette - DarkBrunette     5.1 5.48 15   0.930  0.7893

P value adjustment: tukey method for comparing a family of 4 estimates
```

```
contrast(pain_em, method = "pairwise", adjust = "scheffe")
```

```
contrast                    estimate   SE df t.ratio p.value
LightBlond - DarkBlond           8.0 5.17 15   1.548  0.5137
LightBlond - LightBrunette      16.7 5.48 15   3.046  0.0589
LightBlond - DarkBrunette       21.8 5.17 15   4.218  0.0071
DarkBlond - LightBrunette        8.7 5.48 15   1.587  0.4932
DarkBlond - DarkBrunette        13.8 5.17 15   2.670  0.1109
LightBrunette - DarkBrunette     5.1 5.48 15   0.930  0.8330

P value adjustment: scheffe method with rank 3
```

2.2 Tablet

This data contains the level of chlorpheniramine maleate in tablets from seven labs (Rice 1995, 443–44). The purpose of the experiment was to study the consistency between labs. For each of four manufacturers, composites were prepared by grinding and mixing together tablets in order to measure the amount of chlorpheniramine maleate. Seven labs were asked to make 10 determinations on each composite (Kirchhoefer 1979). The data for the 7 labs are provided in the file `tablet1.txt`.

We start by reading in the data and use the `pivot_longer()` function from the **tidyr** package to reshape the data from *wide* to *long* format

Reshape the data from wide to long format.

```
library(tidyverse)
tablet = read_tsv("https://raw.githubusercontent.com/DATA2002/data/master/tablet1.txt")
glimpse(tablet)
```

```
Rows: 10
Columns: 7
$ Lab1 <dbl> 4.13, 4.07, 4.04, 4.07, 4.05, 4.04, 4.02, 4.06, 4.10, 4…
$ Lab2 <dbl> 3.86, 3.85, 4.08, 4.11, 4.08, 4.01, 4.02, 4.04, 3.97, 3…
$ Lab3 <dbl> 4.00, 4.02, 4.01, 4.01, 4.04, 3.99, 4.03, 3.97, 3.98, 3…
$ Lab4 <dbl> 3.88, 3.88, 3.91, 3.95, 3.92, 3.97, 3.92, 3.90, 3.97, 3…
$ Lab5 <dbl> 4.02, 3.95, 4.02, 3.89, 3.91, 4.01, 3.89, 3.89, 3.99, 4…
$ Lab6 <dbl> 4.02, 3.86, 3.96, 3.97, 4.00, 3.82, 3.98, 3.99, 4.02, 3…
$ Lab7 <dbl> 4.00, 4.02, 4.03, 4.04, 4.10, 3.81, 3.91, 3.96, 4.05, 4…
```

```
tabdat = tablet %>%
    pivot_longer(cols = everything(), names_to = "lab", values_to = "measurement")
glimpse(tabdat)
```

```
Rows: 70
Columns: 2
$ lab         <chr> "Lab1", "Lab2", "Lab3", "Lab4", "Lab5", "Lab6", …
$ measurement <dbl> 4.13, 3.86, 4.00, 3.88, 4.02, 4.02, 4.00, 4.07, …
```

Produce some basic summary statistics and generate side by side box plots.

```
tabdat %>%
  ggplot() +
  aes(x = lab, y = measurement, fill = lab) +
  geom_boxplot() +
  theme_classic() +
  labs(y = "Chlorpheniramine maleate (mg)",
       x = "Lab", fill = "")
```

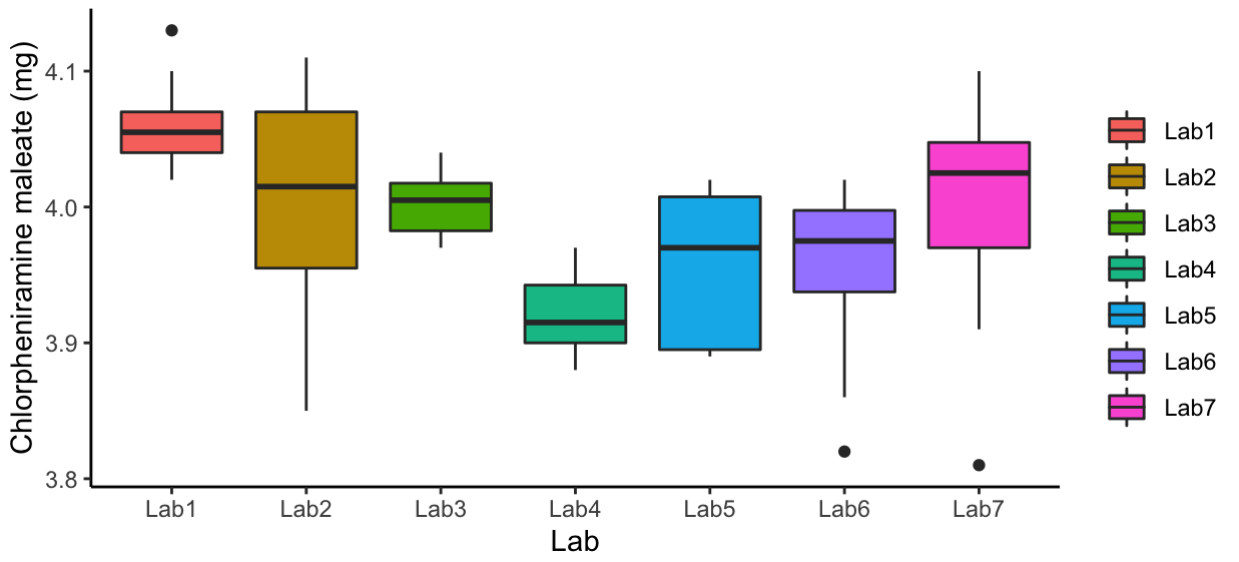


Figure 1: Figure 1: Boxplots of determinations of amounts of chlorpheniramine maleate in tablets by seven laboratories.

The boxplots in Figure 1 show some differences in the medians. Are these differences significant? We can address this by asking a variety of questions.

1. Is the mean level of chlorpheniramine maleate in tablets from Lab 1 different from 4.0? State the null hypothesis.

Let μ_i be the mean level of chlorpheniramine in tablets from Lab i.

This is a one sample t-test

$H_0\colon\ \mu_1 = 4$ vs $H_1\colon\ \mu_1 \neq 4$

```
t.test(tablet$Lab1, mu = 4)
```

```
    One Sample t-test

data:  tablet$Lab1
t = 6.0157, df = 9, p-value = 0.0001986
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 4.038685 4.085315
sample estimates:
mean of x
    4.062
```

2. Is the mean level of chlorpheniramine maleate in tablets from Lab 1 different from that from Lab 3?

This is a two sample t-test

$H_0\colon\ \mu_1 = \mu_3$ vs $H_1\colon\ \mu_1 \neq \mu_3$

```
t.test(tablet$Lab1, tablet$Lab3)
```

```
    Welch Two Sample t-test

data:  tablet$Lab1 and tablet$Lab3
t = 4.6692, df = 16.227, p-value = 0.0002475
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.03224345 0.08575655
sample estimates:
mean of x mean of y
    4.062     4.003
```

```
# or equivalently tabdat %>% filter(lab %in% c('Lab1','Lab3')) %>%
# t.test(measurement ~ lab, data = .)
```

3. Perform a one-way ANOVA to test if the mean levels of chlorpheniramine maleate differ across the seven labs.

```
lab_anova = aov(measurement ~ lab, data = tabdat)
summary(lab_anova)
```

```
            Df Sum Sq  Mean Sq F value   Pr(>F)
lab          6 0.1247 0.020790    5.66 9.45e-05 ***
Residuals   63 0.2314 0.003673
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Short form summary:

The one-way ANOVA shows a significant difference between the means of the 7 labs ($F = 5.66, \text{df} = 6, 36,\ p < 0.001$).

Write out the steps in full:

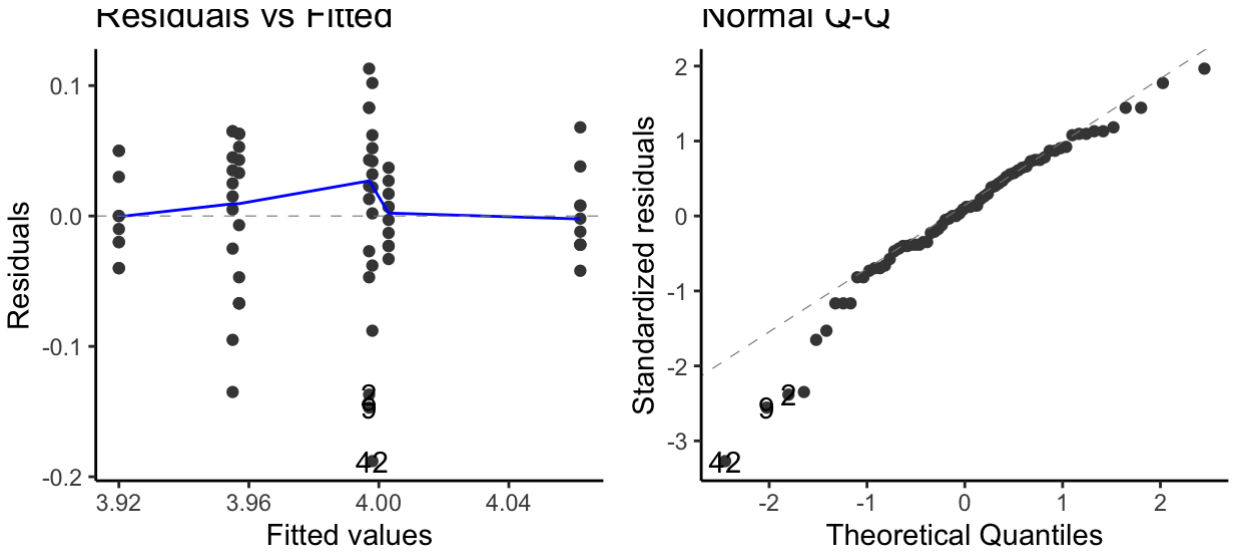
Let μ_1, μ_2, μ_3, μ_4, μ_5, μ_6 and μ_7 be the population mean level of chlorpheniramine maleate from each of the seven labs.

1. **Hypotheses:** H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$ vs H_1: at least one $\mu_i \neq \mu_j$.
2. **Assumptions:** Observations are independent within each of the 7 samples. Each of the 7 populations have the same variance. Each of the 4 populations are normally distributed. [These are checked in the next question.]
3. **Test statistic:** $T = \frac{\text{Treatment Mean Sq}}{\text{Residual Mean Sq}}$. Under H_0, $T \sim F_{g-1,\, N-g}$ where $g = 7$ and $N = 70$.
4. **Observed test statistic:** $t_0 = \frac{0.020790}{0.003673} = 5.66$.
5. **p-value:** $P(T \geq 5.66) = P(F_{6,\,63} \geq 5.66) < 0.001$.
6. **Decision:** As the p-value is very small we reject the null hypothesis and conclude that the population mean level of chlorpheniramine maleate of at least one lab is significantly different to the others.

4. Obtain a QQ plot of the residuals and comment on the validity of the ANOVA assumptions.

We can use the **ggfortify** function to speed up the process of generating residual plots.

```
library(ggfortify)
autoplot(lab_anova, which = c(1, 2)) + theme_classic()
```



In the left hand plot, we're looking for changes in the spread of the residuals across the range of fitted values. It looks like there might be a bit more variation in the center than at the extremes, but the side by side boxplots earlier showed that the constant variance assumption was more or less OK in that the spreads of data was not wildly different between the labs.

In the right hand plot, there are a few observations at the lower end that deviate from the diagonal line, so the residuals may not be normally distributed. However, the discrepancy is not large and the total sample size is large enough that the central limit theorem will ensure our inferences are at least approximately valid.

We could generate these plots "manually" by extracting the fitted values and residuals from the ANOVA object.

```
ass_df = data.frame(fitted = lab_anova$fitted.values, resids = lab_anova$residuals)
p1 = ass_df %>%
    ggplot() + aes(sample = resids) + stat_qq() + stat_qq_line() + theme_classic() +
    labs(x = "Theoretical quantiles", y = "Residuals")
p2 = ass_df %>%
    ggplot() + aes(y = resids, x = fitted) + geom_point() + theme_classic() +
    labs(x = "Fitted values", y = "Residuals")
gridExtra::grid.arrange(p2, p1, ncol = 2)
```

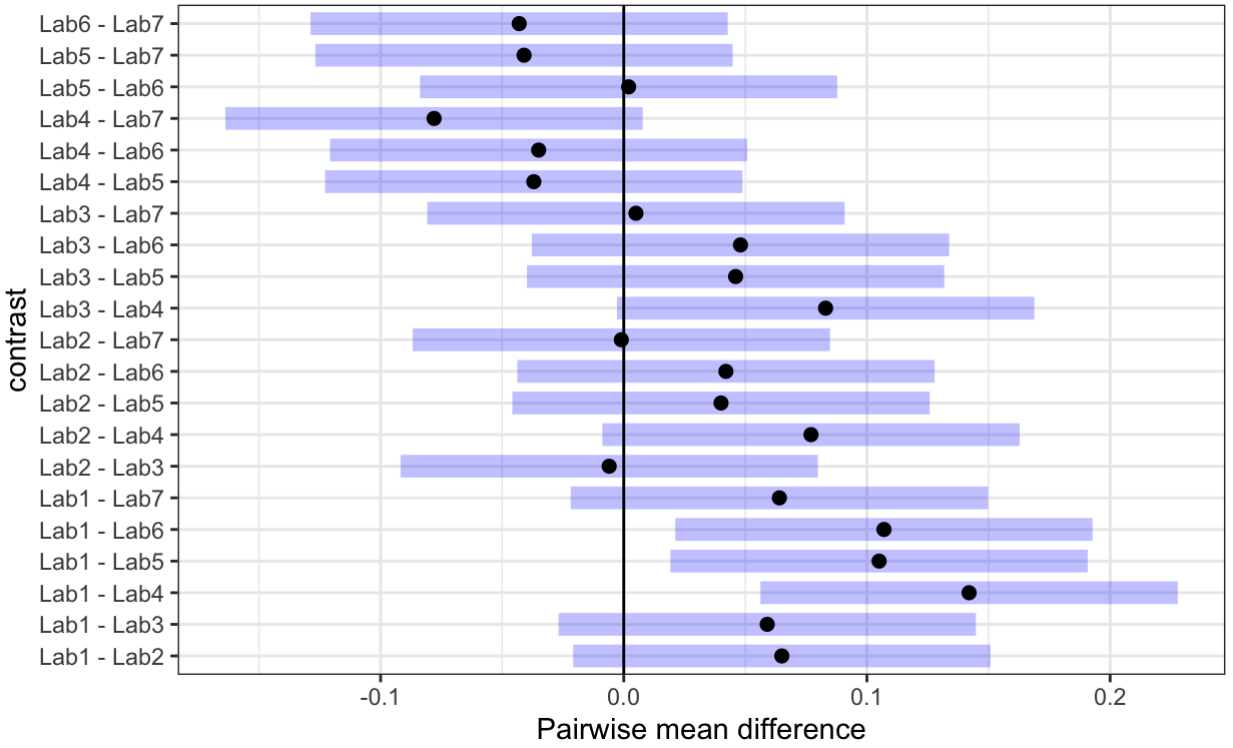
5. Perform pairwise post hoc tests to determine which labs are significantly different.

```
library(emmeans)
lab_em = emmeans(lab_anova, ~ lab)
lab_pair = contrast(lab_em, method = "pairwise", adjust = "bonferroni")
lab_pair %>%
  data.frame() %>%
  filter(p.value < 0.1) %>%
  knitr::kable(digits = 4)
```

contrast	estimate	SE	df	t.ratio	p.value

contrast	estimate	SE	df	t.ratio	p.value
Lab1 - Lab4	0.142	0.0271	63	5.2392	0.0000
Lab1 - Lab5	0.105	0.0271	63	3.8740	0.0054
Lab1 - Lab6	0.107	0.0271	63	3.9478	0.0042
Lab3 - Lab4	0.083	0.0271	63	3.0623	0.0678

```
plot(lab_pair) + theme_bw() +
  labs(x = "Pairwise mean difference") +
  geom_vline(xintercept = 0)
```



We conclude that the population mean level of chlorpheniramine maleate in tablets from Lab 1 is significantly different to Labs 4, 5 and 6.

6. Use a **rank** based approach to testing for a difference between the means of the 7 labs.

```
kruskal.test(measurement ~ factor(lab), data = tabdat)
```

```
    Kruskal-Wallis rank sum test

data:  measurement by factor(lab)
Kruskal-Wallis chi-squared = 29.606, df = 6, p-value =
4.67e-05
```

Short form summary:

The Kruskal-Wallis test shows a significant difference between the means of the 7 labs ($\chi^2 = 29.6, \mathrm{df} = 6, p < 0.001$).

Write out the steps in full:

1. **Hypotheses:** H_0: the level of chlorpheniramine maleate is distributed identically for across all labs (and therefore the mean level is the same across all labs) vs H_1: the level of chlorpheniramine maleate is systematically higher for at least one lab
2. **Assumptions:** Observations are independent within each group and groups are independent of each other. The different groups follow the same distribution (differing only by the location parameter). This looks reasonable from the side-by-side boxplots.
3. **Test statistic:** Under the null hypothesis the Kruskal-Wallis test statistic approximately follows a χ^2 distribution with $g-1$ degrees of freedom $t_0 = 29.6$
4. **p-value:** $P(T \geq t_0) = P(\chi^2_6 \geq 29.606) < 0.001$.
5. **Decision:** As the p-value is very small we reject the null hypothesis and conclude that the population mean of at least one group is significantly different to the others.

Post hoc:

If we wanted to go on and perform nonparametric post hoc tests, we could apply the Bonferroni method to all pairwise comparisons tested by Wilcoxon rank-sum tests.

```
pairwise.wilcox.test(x = tabdat$measurement, g = factor(tabdat$lab), p.adjust.method =
        "bonferroni")
```

```
    Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  tabdat$measurement and factor(tabdat$lab)

     Lab1   Lab2   Lab3   Lab4   Lab5   Lab6
Lab2 1.0000 -      -      -      -      -
Lab3 0.0114 1.0000 -      -      -      -
Lab4 0.0036 1.0000 0.0049 -      -      -
Lab5 0.0048 1.0000 1.0000 1.0000 -      -
Lab6 0.0049 1.0000 1.0000 1.0000 1.0000 -
Lab7 0.8423 1.0000 1.0000 0.3228 1.0000 1.0000

P value adjustment method: bonferroni
```

Again, we see that Lab 1 is significantly different to Labs 4, 5 and 6. However, now Lab 3 is also significantly different to Lab 4.

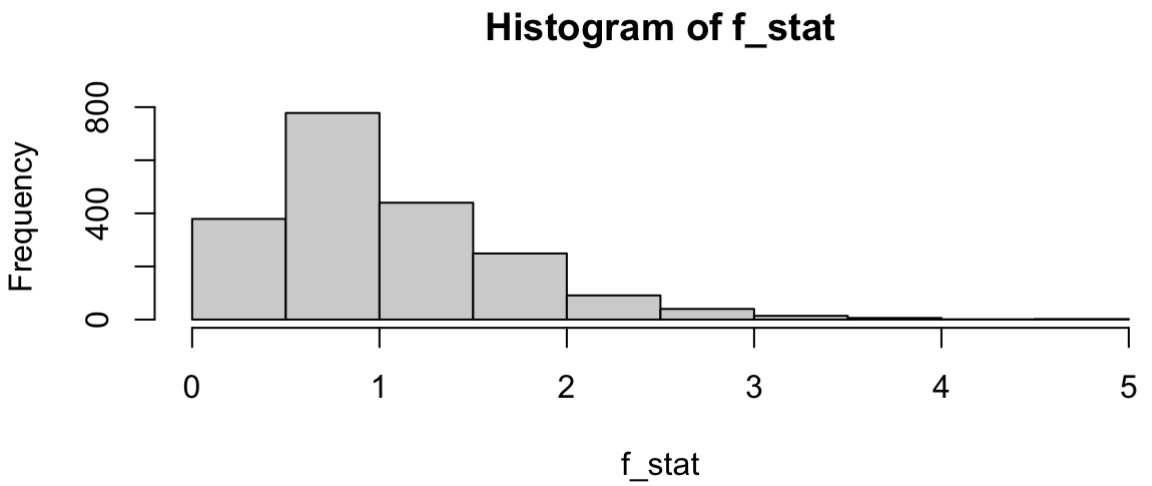
Final comments:

The Kruskal-Wallis test makes no assumption of normality and thus has a wider range of applicability than a standard one-way ANOVA. It is especially useful in small-sample situations. Because data are replaced by their ranks, outliers will have less influence on this nonparametric test than on the ANOVA F test. In some applications, the data might be considered more like a ranking than a measurement -

for example, in wine tasting, judges usually rank the wines - which makes the use of the Kruskal-Wallis test very natural.

7. Use a **permutation** based approach to testing for a difference between the means of the 7 labs.

```
B = 2000
f_stat = vector(mode = "numeric", length = B)
for (i in 1:B) {
    permuted_anova = aov(sample(tabdat$measurement) ~ factor(tabdat$lab))
    f_stat[i] = broom::tidy(permuted_anova)$statistic[1]
}
t_0 = broom::tidy(lab_anova)$statistic[1]
hist(f_stat)
```




```
mean(f_stat >= t_0)
```

```
[1] 0
```

3 For practice after the computer lab

- Use a rank based approach to testing for a difference between the means for the pain threshold data.
- In Larsen and Marx (2012) work through Case Study 12.3.1 and Case Study 12.4.1 and then consider questions 12.3.3 and 12.4.4.
- You can also attempt the DataCamp chapter on comparing many means.

References

Kirchhoefer, R D. 1979. “Semautomated Method for the Analysis of Chlorpheniramine Maleate Tablets: Collaborative Study.” *Journal - Association of Official Analytical Chemists* 62 (6): 1197–1201.

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

Rice, John A. 1995. *Mathematical Statistics and Data Analysis*. Belmont, CA: Duxbury Press.