MA4605 Lab: Comparing Two Independent Groups

This lab will examine a completely randomised design, the simplest experimental designs, where two independent groups are compared.

## Example 1: Bacterial Growth on Meat

The shelf life of stored meats is the time a prepackaged cut remains safe and nutritious. Standard packaging in ambient air atmosphere has a shelf life of approximately 48 hours, after which the meat quality tends to deteriorate. Recent studies have suggested that using controlled gas atmospheres (such as CO) are possible alternatives to existing packagings.

### Research hypothesis:

The investigator believes that a controlled gas atmosphere would provide a more effective packaging environment for meat storage. They are interested in answering the following question “Is bacterial growth different with standard packaging versus packaging using a controlled gas atmosphere?”.

### Treatments:

The treatments developed by the investigator to answer this questions were

1. standard packaging using ambient air and
2. packaging using 100% CO.

### Experimental design:

The researcher decided to use steaks as his experimental units. Twenty beef steaks of approximately the same size (25g) were randomly assigned to each of the packaging conditions. Each steak was packaged separately in its assigned condition and the number of bacteria (measured in count/cm) on the meat was measured after nine days of storage at 4C.

## Exploratory analysis

### Exploratory plots

The data for this example are stored in the “Bacteria.csv” file on RStudio cloud. First, load the data into R using:

steak = read.csv(file="../data/Bacteria.csv")

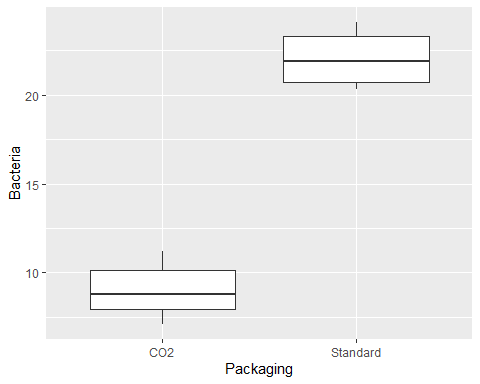
The dataset is now stored in a **dataframe** called **steak**. The dataframe has two columns:

1. **Packaging** - contains what packaging was used (CO or Standard).
2. **Bacteria** - contains the corresponding bacterial levels.

The first step in any statistical analysis is to perform an exploratory analysis of the data. This involves generating appropriate plots and summary statistics to describe the data, identify any outlying values, etc.

Let’s initially create a boxplot:

ggplot(steak, aes(x=Packaging, y=Bacteria)) + geom\_boxplot()



The boxplots are a very useful summary of our data. They show us that:

* The bacterial levels for the standard packaging are much higher than those for the CO packaging. This indicates a clear difference between the bacterial levels in the two packaging types.
* The spread of the boxplots also looks to be reasonably similar, indicating that the variability in the two groups is approximately the same.

Boxplots can also be used to check the assumptions required for the -test to compare the group means. The assumptions we need to check are:

1. Is the sample size large? If yes, check for any very extreme outliers. Otherwise, ok.
2. Is the sample size small? If yes, check that the data are approximately Normally distributed in each group. No extreme skewness or outliers.

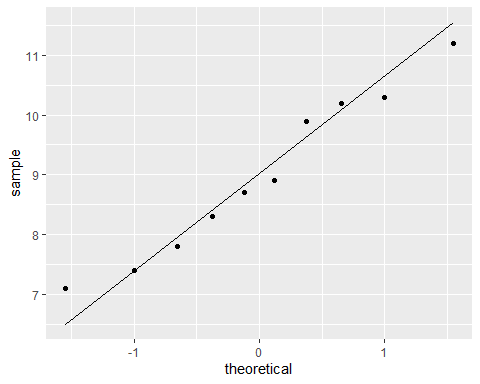
Since here, , the sample size in each group is small, and it is important to examine the boxplots to assess the distribution of the data in each group and identify any outliers.

* The shapes of the boxplots suggest that the data for both types of packaging arose from symmetric (possibly Normal) distributions. (The median is in the middle of the box, the tails are approximately symmetric. There are no obvious outliers.)

A QQ-plot for each group assesses how well the data are described by the Normal distribution. This plot is more useful than the boxplot for assessing the Normality assumption required when the sample size is small.

For the CO group the QQ-plot is created as follows:

CO2 = steak %>%  
 filter(Packaging == "CO2")  
ggplot(CO2, aes(sample=Bacteria)) + geom\_qq() + geom\_qq\_line()



The points lie approximately along a straight line. The data in this groups are approximately Normally distributed as required.

**Exercise 1:** Create a QQ-plot for the Standard data and comment. Write your answers in your R Markdown file.

### Summary statistics

Next calculate some summary statistics. We use the group\_by() command to calculate the statistics by packaging type. The summarise() command allows us to specify what summary statistics we want to calculate.

steak %>%  
 group\_by(Packaging) %>%  
 summarise(min = min(Bacteria), max = max(Bacteria),  
 mean = mean(Bacteria), med = median(Bacteria),  
 sd = sd(Bacteria),   
 q1 = quantile(Bacteria, probs = 0.25),  
 q3 = quantile(Bacteria, probs = 0.75))

`summarise()` ungrouping output (override with `.groups` argument)

# A tibble: 2 x 8  
 Packaging min max mean med sd q1 q3  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 CO2 7.1 11.2 8.98 8.8 1.37 7.92 10.1  
2 Standard 20.3 24.1 22.0 21.9 1.50 20.7 23.3

There appears to be a clear difference between the mean number of bacteria for the two packaging types. Standard packaging has much higher bacteria counts on average, 22.01 per cm versus 8.98 per cm for CO packaging. The standard deviations in both groups are also reasonably similar indicating that it may be appropriate to assume the variances are the same.

## Formal analysis

To provide a formal answer to the researcher’s question, it is necessary to carry out an appropriate hypothesis test and construct a confidence interval to test if there is a difference in the population mean number of bacteria/cm between the two packaging types.

Before carrying out the -test we should check if the population variances are equal. While the sample standard deviations of the two groups look roughly similar, an -test is performed to formally test this.

var.test(data=steak, Bacteria ~ Packaging, conf.level=0.95)

F test to compare two variances  
  
data: Bacteria by Packaging  
F = 0.83589, num df = 9, denom df = 9, p-value = 0.7938  
alternative hypothesis: true ratio of variances is not equal to 1  
95 percent confidence interval:  
 0.2076222 3.3652704  
sample estimates:  
ratio of variances   
 0.8358856

Notice that the p-value is 0.79 which implies we cannot reject the null hypothesis that the variances are equal. The CI is quite wide as we have small samples. For our purposes we assume the variances are approximately equal.

Next carry out the -test to compare the group means. Remember we are only interested in if there is a difference between the two groups. Therefore this is a two-tailed test. Remember by default, R orders the factor levels in alphabetical order. Therefore and . The null and alterantive hypotheses are:

* :
* :

To carry out this in R:

t.test(Bacteria ~ Packaging, data=steak, alternative='two.sided',  
 var.equal=TRUE, paired=FALSE, mu=0, conf.level=0.95)

Two Sample t-test  
  
data: Bacteria by Packaging  
t = -20.244, df = 18, p-value = 7.811e-14  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -14.38225 -11.67775  
sample estimates:  
 mean in group CO2 mean in group Standard   
 8.98 22.01

NOTE:

* alternative='two.sided' tells R that a two-tailed test required.
* var.equal=TRUE implies assume equal variances.
* paired=FALSE implies these are independent groups.
* mu=0 tells R to assume the difference between the pop. means is 0.
* conf.level=0.95 sets the significance level .

## Conclusion

**Exercise 2:** Using the output above, what is the conclusion? Write your answer in your R Markdown file. You should comment on the p-value AND the confidence interval.

### Tutorial question

**Exercise 3:** Carry out the formal analysis (i.e. the F-test and t-test) by hand showing all calculations. Verify that your calculations and conclusions are the same as those outputted by R. You do not have to put this work in your R Markdown file.