FASS512: Significance testing

Professor Patrick Rebuschat, [p.rebuschat@lancaster.ac.uk](mailto:p.rebuschat@lancaster.ac.uk)

Please work through the following handout at your own pace.

This is a loooooong handout.

We will start during class but you probably need to complete this at home.

Please bring questions to our next session.

**References for this handout**

Many of the examples and data files from our class come from these excellent textbooks:

* Andrews, M. (2021). *Doing data science in R*. Sage.
* Brown, D. S. (2021). *Statistics and data visualization using R. The art and practice of data analysis*. Sage.
* Cumming, G. & Calin-Jaegeman, R. J. (2017). *Introduction to the New Statistics: Estimation, Open Science, and Beyond*. Routledge.
* Crawley, M. J. (2013). *The R book*. Wiley.
* Fogarty, B. J. (2019). *Quantitative social science data with R*. Sage.
* Winter, B. (2019). *Statistics for linguists. An introduction using R*. Routledge.

Are you ready? Then let’s start on the next page! ☟

Task 1: *Significance testing: Testing for normality*

In this week's session, we start looking at the more formal hypothesis tests for continuous variables. The focus here will not be on the mathematical theory and equations underlying the tests, but rather on their computation using R. The interpretation of these tests, which is what you really need to know about, follows straightforwardly from what we covered in the last three sessions.

*Loading and inspecting our data set*

Let’s start by loading our data set. Please download and load the language\_exams file from our Moodle site.

language\_exams <- read.table('language\_exams\_data.csv',

header = T,

row.names = 1,

sep = ",")

The table contains three columns:

* a number to identify the student (student\_nr);
* our participants’ native languages, first\_lang, with three levels: Language A, Language B, Language C;
* the level of study of our participants, study\_level, with two levels: undergraduate or postgraduate;
* the scores of our participants on three foreign language exams (lang\_exam\_1, lang\_exam\_2, lang\_exam\_3), which our participants took in term 1, term 2, and term 3 respectively;
* the scores on a maths exam (maths\_exam).

There are 200 observations (N = 200).

Let’s have a quick look at the data set to get a rough idea.

View(language\_exams)

summary(language\_exams)

As mentioned in class, many statistical tests assume random sampling from the population (to avoid sample bias) and that the population is normally distributed. We saw a few weeks ago that the data in our sample might not appear normally distributed, especially when our sample is small, but that this does not necessarily mean the population is not normally distributed.

It’s good practice to check if your data is normally distributed before conductig significance test as this will affect your test choice.

*Checking for normality: Visual inspection*

A quick way to check for normality is to visually inspect your data. In the previous weeks, we have used histograms for this. Just plotting the frequency of scores can give us a rough idea of the distribution in our sample.

Today, we want to look at another widely-used visual method that can be used to check whether a data variable is (approximately) normally distributed, the QQ-plot.

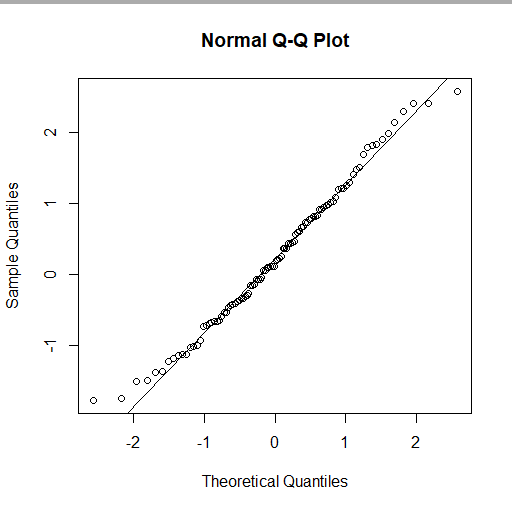
The QQ-plot plots the quantiles of the observed data variable against the quantiles that we would theoretically expect if the data followed a normal distribution exactly.

The plot contains two elements: a set of points (white bubbles) representing the observed data and a diagonal straight line representing the theoretical normal distribution.

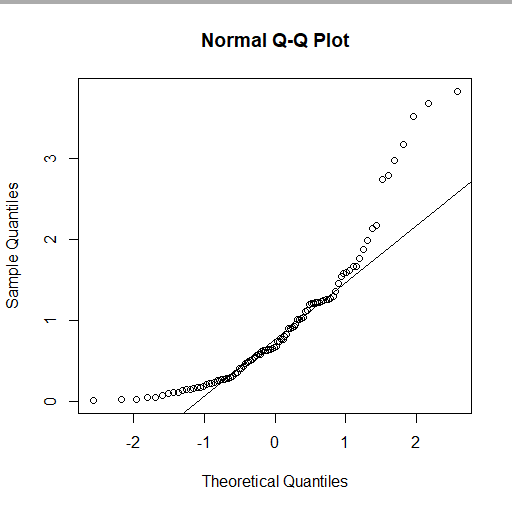
If the data distribution is approximately normal, then we expect the data points to sit **more or less on top of the straight diagonal line**; any deviation from it should be minor and pretty random in its pattern. If we see larger and clearer deviations from the straight diagonal, then we should conclude that our data are not approximately normal.

Bear in mind, though, that random sampling error means that we can expect to see some deviation, even with data drawn from a normal distribution.

Here is a first QQ-plot example. This data is drawn from **a normal distribution**. As you can see, there is some minor deviation bottom left and top right, but otherwise close to the diagonal line.



And here is a QQ-plot example with data drawn from a **non-normal distribution** (in this case, a Gamma distribution). Note the elongated S-shape to the line of data points, with considerable deviation, except in the middle part of the data.



To run a normal QQ-plot in R, we need the following lines of code, in which we use the qqnorm() and qqline() functions. The commands must come in the right order.

To illustrate, we first create an object qq\_exam\_1 with the scores from language exam 1.

qq\_exam\_1 <- language\_exams$lang\_exam\_1

Then, we use the following commands to generate the plot.

qqnorm(qq\_exam\_1)

qqline(qq\_exam\_1)



As you can see, most of the scores are on top of our diagonal line, with some deviation at the top right. But this graph suggest that the data is approximately normally distributed. (We will shortly run a statistical test to confirm this!)

A good way of getting used to what degree of variation you can expect to see in normally distributed data is to run several plots using data sampled from a theoretical normal distribution.

You can do this with these lines of code.

Every time we run the code below we are creating a new data set with 80 observations from a normal distribution.

Run the code **four or five times**, and each time, observe the QQ-plot. How well does our sample data fit on the diagonal line? Does it change much from sample to sample?

sample\_80 <- rnorm(80)

qqnorm(sample\_80)

qqline(sample\_80)

Let’s now see what happens when we increase the sample size. This gives you a good idea of how sample size affects the normality of our data.

sample\_20 <- rnorm(20)

qqnorm(sample\_20)

qqline(sample\_20)

sample\_50 <- rnorm(50)

qqnorm(sample\_50)

qqline(sample\_50)

sample\_200 <- rnorm(200)

qqnorm(sample\_200)

qqline(sample\_200)

Again, compare the fit on the diagonal line. When is the fit best? With 20 observations (which is a very, very common sample size in man experiments), with 50 observations, or with 200 observations?

*Statistical test for normality*

Visual inspection is important, it gives us a rough idea of whether our sample data is normally distributed. However, there are several tests that allow us to directly test whether our population data are likely to be normally distributed.

The Shapiro-Wilk’s test is widely recommended for normality test. This test is based on the correlation between the data and the corresponding normal scores. We can use this to test if a given variable is normally distributed.

Note that, normality test is sensitive to sample size. Small samples most often pass normality tests. Therefore, it’s important to combine visual inspection and significance test in order to take the right decision.

The R function shapiro.test() can be used to perform the Shapiro-Wilk test of normality for one variable (univariate).

For example, if we would like to know whether the scores for language exam 2 are normally distributed, we can run the test with the following command.

shapiro.test(language\_exams$lang\_exam\_2)

Shapiro-Wilk normality test

data: language\_exams$lang\_exam\_2

W = 0.98992, p-value = 0.1741

The test statistic is W = 0.99, and p = 0.17, which is not statistically significant. This suggest that the data is normally distributed. (If the p-value were below .05, it would mean the data significantly deviate from a normal distribution.)

If we would like to test the normality of several variables at the same time, rather than doing many tests in sequence, we can use the following code.

First, we create a new object. This will have the continuous data from our sample, i.e. the language scores for exams 1, 2, 3 and the scores of the maths exam.

normality\_data = subset(language\_exams,

select = c(lang\_exam\_1, lang\_exam\_2, lang\_exam\_3, maths\_exam))

We then use the apply() function to apply the Shapiro-Wilk’s test to each variable. In the apply() function, the first argument is the data set (normality\_data). The second argument is the “margin”, and if we use the number “2” for margin we specify that the function should apply to the columns in our data set. The third argument is the function we wish to apply, here, the shapiro.test(), of course.

If you run this, the test will be applied to all the variables one after the other, saving us a bit of time.

apply(normality\_data, 2, shapiro.test)

$lang\_exam\_1

Shapiro-Wilk normality test

data: newX[, i]

W = 0.9833, p-value = 0.01785

$lang\_exam\_2

Shapiro-Wilk normality test

data: newX[, i]

W = 0.98992, p-value = 0.1741

$lang\_exam\_3

Shapiro-Wilk normality test

data: newX[, i]

W = 0.99431, p-value = 0.6467

$maths\_exam

Shapiro-Wilk normality test

data: newX[, i]

W = 0.98376, p-value = 0.02086

From the output, we see that the variables are normally distributed, with the exception of the scores in the maths exam and those in language exam 1. We need to bear this in mind when using these variables in our tests later. For example, we could use non-parametric significance tests, which do now assume a normal distribution in the sample.

Task 2: *Significance testing for continuous data: Testing variances and test selection*

Now that we checked whether our data is normally distributed, let’s do some significance testing. We will focus on the classic case: continuous data, and two groups we wish to compare.

To illustrate, we will compare the performance of the undergraduates and postgraduates on language exam 2. Let’s create two subsets of data, one for the undergraduates, exam\_2.ug, one for the postgraduates, exam\_2.pg.

exam\_2.ug <-

subset(language\_exams, study\_level == 'Undergrad')$lang\_exam\_2

exam\_2.pg <-

subset(language\_exams, study\_level == 'Postgrad')$lang\_exam\_2

*Choosing the right test*

As discussed in class, selecting an appropriate significance test depends on four basic questions:

1. What kind of data is involved, continuous or categorical?
2. How many groups are being compared?
3. Are the group variances approximately equal?
4. Are the groups independent of one another?

*Questions 1 and 2*

Given the present data set, the answers to questions 1 and 2 are obvious. We know that our language scores are a continuous variable (we will cover categorical tests next week), and we know we are comparing two groups (undergraduates and postgraduates). This leaves questions 3 and 4.

*Question 3: Variances*

Question 3 is about group variances, and strictly speaking this refers to the population-level variances of the two groups, not the variances of the actual samples that we have observed. So it is not as easy as just saying that the variance of sample 1 is larger or smaller than the variance of sample 2.

To answer this question, the most straightforward procedure is to use the F-test. This statistical test assesses whether the variances of two variables are equal. The test assumes that our data is normally distributed, but we already confirmed that this is the case (see our Shapiro-Wilk’s test for language exam 2). We have two options.

We can calculate the F-test statistics by computing the ratio of the two variances with variable A / variable B. The more this ratio deviates from 1, the stronger the evidence for unequal population variances.

We first compute the variances for the two subsamples, exam\_1.ug and exam\_1.pg, using the var() function.

var(exam\_2.ug)

[1] 284.2977

var(exam\_2.pg)

[1] 214.3439

And then we compute their ratio, with the largest of the two results entered first.

var(exam\_2.ug) / var(exam\_2.pg)

[1] 1.326363

Alternatively, we can run the F-test as follows. Our null hypothesis is that the two variances are equal.

var\_test <- var.test(lang\_exam\_2 ~ study\_level, data = language\_exams)

var\_test

F test to compare two variances

data: lang\_exam\_2 by study\_level

F = 0.75394, num df = 99, denom df = 99, p-value = 0.1617

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.5072833 1.1205333

sample estimates:

ratio of variances

0.7539415

As you can see, the p-value indicates that the F statistics is compatible with our null hypothesis, F = 0.75394, p = .162. The result is not significant, so we do not reject the null hypothesis that the two variances are equal.

If we wish to test the homogeneity of variance of more than two groups, there are other tests, e.g. Bartlett’s test (for normally distributed data) or Levene’s test (for data that is not normally distributed).

To use the Bartlett’s test, we use the **bartlett.test**() function as in the following command. The first argument is the formula (exam\_2 ~ first\_lang), the second argument is the data set (language\_exams).

bart\_test <- bartlett.test(lang\_exam\_2 ~ first\_lang, data = language\_exams)

bart\_test

Bartlett test of homogeneity of variances

data: lang\_exam\_2 by first\_lang

Bartlett's K-squared = 4.5247, df = 2, p-value = 0.1041

The test statistic is Bartlett’s K2, and as we can see, the test is not significant, p = .104. This means that that the variance in the language exam 2 scores is not statistically significantly different for undergraduates and postgraduates. We have “equal variances.”

*Question 4*

Finally, question 4 asks whether the two samples are independent.

Sometimes, we sample data from the same participants on more than one occasion in order to look for changes in a particular variable (“before and after” design). For example, we might want to see if a given teaching method is effective in promoting language development. To investigate this, we need to test the all participants in our study at least twice, once before they start the educational intervention (pre-test) and again after they complete the intervention (post-test). In this case, we say that the samples (pre-test scores and post-test scores) are dependent on one another as they come from the same participants. Dependent data is also known as “paired data”.

In the present case, we are clearly looking at two different groups, undergraduate and postgraduate students, and their performance on language exam 2. For this reason, we say they are independent of one another.

The answers to questions 1 to 4 lead us to the following decision: To compare two independent groups with continuous data and approximately equal variances, we use the standard t-test.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Test name | Groups | Variance |
| Parametric | (standard) t-test | Two independent groups | approximately equal variances |
|  | Welch t-test | Two independent groups | unequal variances |
|  | paired t-test | Two dependent groups |  |
|  | (standard) one-way ANOVA | More than two groups | approximately equal variances |
|  | Welch one-way ANOVA | More than two groups | unequal variances |
| Non-parametric | Wilcoxon-Mann-Whitney test | Two independent groups |  |
|  | Wilcoxon signed rank test | Two dependent groups |  |
|  | Kruskal-Wallis test | More than two groups |  |

Five of the tests in the table above are **parametric** tests. This means that they are based on the theoretical assumption that the statistics of interest follow (roughly) a particular probability distribution, such as the normal distribution. As a consequence of the Central Limit Theorem (which we covered in session 7), this is often a fairly safe assumption, though it becomes increasingly less safe as sample sizes get smaller. Thus, if you want to be on the safe side, you can instead use one of a corresponding set of non-parametric tests.

**Non-parametric** tests do not make any assumptions about the distributions that the statistics of interest follow (though they do assume that the distributions for the two groups being compared have roughly the same shape). The relevant non-parametric alternatives can also be found in the table above.

Task 3: *Significance testing for continuous data: Tests for two independent groups*

We are now ready for testing.

To run the statistical tests, it is easiest to pull out the two groups first using the subset command and give them names like we did above (exam\_2.ug and exam\_2.pg). There are sometimes other ways of specifying the groups inside a command, but these are only straightforward if your original grouping variable contains two values exactly. This may not always be the case. For instance, you may have included a third value to account for missing data, or you may just want to perform a test on two groups selected out of a larger number. Also, the alternative methods are not available for some commands. The method we have used above, and which is represented here again, works in all cases and is therefore the simplest "recipe" to learn at first.

exam\_2.ug <-

subset(language\_exams, study\_level == 'Undergrad')$lang\_exam\_2

exam\_2.pg <-

subset(language\_exams, study\_level == 'Postgrad')$lang\_exam\_2

If you inspect the two new variables, exam\_2.ug and exam\_2.pg, you will see that they only contain the exam 2 scores of undergraduate and postgraduate students, respectively. The rest of the data (native languages, language exams 1 and 3, the maths scores) has not been included.

*Tests for two independent groups*

*Standard t-test*

So, to calculate a standard t-test, i.e. assuming equal variances, the command is:

t.test(exam\_2.ug, exam\_2.pg, var.equal = T)

Two Sample t-test

data: exam\_2.ug and exam\_2.pg

t = 1.2916, df = 198, p-value = 0.198

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.519427 7.287715

sample estimates:

mean of x mean of y

62.29971 59.41556

As discussed in class, the most important thing here is the p-value (highlighted in red). To report this test in a paper, you would also need to include the t-value (highlighted in green) and the df (highlighted in blue).

The **t-value** is simply the test statistic, i.e. the number that comes out of the mathematical equation underlying this test.

The df stands for "**degrees of freedom**" and is defined as the number ‬of independent values in the calculation that are free to vary. The degrees of freedom are of far less practical importance nowadays, because we no longer have to look up t-values in books of printed tables to find out the associated p-value; but they are still theoretically important and must therefore be reported in your results. For a standard t-test, the df is equal to N – 2 (i.e. N minus two, where N is the combined sample size), but, as we’ll see in a moment, the df is adjusted as part of the Welch procedure for unequal variances. (Here, we were testing two groups of 100, so my combined sample size was 200 and the df is therefore 200 minus 2 = 198.) Note that the t-test output in R also gives you a 95% confidence interval (highlighted in purple) for the difference between the means of the two groups. Finally, on the last line of the output, you have (for reference) the observed means of the two samples.

The outputs of most significance tests in R have a very similar format, so, in what follows, I will only comment on any additional features of importance.

We could report the result of the significance test as follows, following APA style. “An independent samples t-test showed that the difference betweeen groups was not significant, t(198) = -1.2916, p-value = .198.” (The degrees of freedom are in the parentheses.)

*Welch t-test*

To calculate a Welch t-test, i.e. assuming unequal variances, the command is as follows. Note, all we did was to change the var.equal setting from T to F.)

t.test(exam\_2.ug, exam\_2.pg, var.equal = F)

Welch Two Sample t-test

data: exam\_2.ug and exam\_2.pg

t = 1.2916, df = 194.18, p-value = 0.198

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.519960 7.288248

sample estimates:

mean of x mean of y

62.29971 59.41556

Note here the difference in the degrees of freedom (df), which is now smaller than the one we obtained with the standard t-test, and is also no longer a whole number. Adjusting the df (using a formula derived by B. L. Welch and F. E. Satterthwaite) provides a better approximation to the true distribution of the statistic for the “unequal variances” case. In fact, many researchers prefer always to use the Welch t-test without bothering to check the ratio of variances, because it is more robust than the standard t-test. It is also the default case in R, which means that you don’t actually need to include the flag var.equal = F in the command.

The report would be something like this. “A Welch t-test showed that the difference betweeen groups was not significant, t(194.18) = -1.2916, p-value = .198.” (The degrees of freedom are in the parentheses.)

*Wilcoxon-Mann-Whitney test*

To calculate a non-parametric Wilcoxon-Mann-Whitney test , the command is as follows. Note that here you have to specify that you want to see a 95% confidence interval, conf.int = 0.95, otherwise this command will only output a p-value.

wilcox.test(exam\_2.ug, exam\_2.pg, conf.int = 0.95)

Wilcoxon rank sum test with continuity correction

data: exam\_2.ug and exam\_2.pg

W = 5572, p-value = 0.1626

alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:

-1.326931 7.629876

sample estimates:

difference in location

3.236362

Note the different terminology from what we had with the t-tests, i.e., “location shift” and “difference in location”. This reminds us that the Wilcoxon-Mann-Whitney test does not test the null hypothesis that the means (or even the medians) of the two populations are equal. It tests, instead, the null hypothesis that the distributions of the two populations coincide exactly, against the alternative that one of the distributions is shifted by a certain amount to the right or left of the other, as illustrated in the diagram below:

Our report could be as follows. “A Wilcoxon-Mann-Whitney testshowed that the difference betweeen groups was not significant, W = 5572, p-value = .163.” There are no degrees of freedom (df) to report with the Wilcoxon-Mann-Whitney test.

Note: APA guidelines say we can round p-values to two or three decimals, so p = .1626 became p = .163.

Task 4: *Significance testing for continuous data: Two paired groups*

Sometimes, we wish to compare dependent groups. For example, we might want to compare the performance of the same participants on different tests, say language exams 1 and 2.

Let’s try this out to illustrate how to do significance testing with paired data.

*Tests for paired data*

In our data set, we have scores for several language exams from the same participants. We can use scores on language\_exam\_1 and language\_exam\_2 to try out our paired-samples t-tests.

First, let’s create two new objects, one for all scores in exam 1, the other for all scores in exam 2. The scores of undergraduates and postgraduates are now combined.

exam\_1 <- subset(language\_exams)$lang\_exam\_1

exam\_2 <- subset(language\_exams)$lang\_exam\_2

*Paired t-test*

To calculate a paired t-test, then, the command is:

t.test(exam\_2, exam\_1, paired = T)

Paired t-test

data: exam\_2 and exam\_1

t = 12.553, df = 199, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

14.58768 20.02503

sample estimates:

mean of the differences

17.30636

Note that the statistic at the bottom (and, by extension, the 95% CI) relate to the mean of the differences between the pairs of observations, and not the difference between the means of the two sets of figures taken collectively. These are not the same thing.

Note also that I entered exam 2 as the first argument and exam 1 as the second. This is because the mean score for exam 2 is larger than the mean score for exam 1. Putting the larger value first avoids getting a negative score, but doesn’t other affect the calculations.

Our report for the test statistic: “A paired-samples t-testshowed that the difference betweeen groups was significant, t(199) = 12.553, p < .001.”

Note: Remember what we said about reporting exact p-values only when they are above p = .01 or .001. For lower values, we can use report inequalities, as in p < .01 or p < .001, as is the case here. The actual p-value is p = .00000000000003731, so we use p < .001.

*Wilcoxon signed rank test*

If our data is not normally distributed, we can use the Wilcoxon signed rank test as the non-parametric option.

wilcox.test(

exam\_2,

exam\_1,

paired = T,

correct = F,

conf.int = 0.95

)

Wilcoxon signed rank test

data: exam\_2 and exam\_1

V = 17919, p-value < 2.2e-16

alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:

14.85871 20.44333

sample estimates:

(pseudo)median

17.66846

The statistic used here is called the “(pseudo-)median”. The pseudo-median and its accompanying 95% CI are somewhat analogous to the “mean of the differences” in the case of the paired t-test. It is the median value of the pairwise differences, not the difference between the medians of the two sets of figures.

Our report for the test statistic: “A Wilcoxon signed rank test indicated that the difference between groups was significant, V = 17919, p < .001.”

*Testing directional versus non-directional hypotheses*

All of the commands presented so far are tests of non-directional hypotheses (commonly also known as “two-tailed” tests). This is by far the most common procedure used in everyday research practice.

To perform a test of a directional (or “one-tailed”) hypothesis, you would need to include one of the following additional statements inside the brackets of the relevant test command: alternative = "greater" and alternative = "less".

Consider the following example.

t.test(exam\_2.ug,

exam\_2.pg,

var.equal = T,

alternative = "less")

Two Sample t-test

data: exam\_2.ug and exam\_2.pg

t = 1.2916, df = 198, p-value = 0.901

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf 6.574415

sample estimates:

mean of x mean of y

62.29971 59.41556

The statement alternative = "less" tests a working hypothesis that “the difference in means is less than zero” against a null hypothesis that “the difference is equal to zero”. Similarly, alternative = "greater" tests a working hypothesis that “the difference in means is greater than zero” against a null hypothesis that “the difference is equal to zero”.

Our report needs to include the fact that it was directional test. We can do this by adding that the test was one-tailed, i.e. it checked either one tail of the distribution or the other. “A paired-samples t-test (one-tailed) showed that the difference was not significantly different, t(198) = 1.2916, p = .901.”

*A note on scientific notation and funny-looking p-values*

When working with R, as we saw with an example earlier, you may sometimes get a very funny-looking p-value like the one below.

p-value = 2.197e-05

This notation is used by R to report very small numbers with many decimal places. What the above example means is that the number to the left of the letter "e", i.e., 2.197, really begins in the fifth place after a decimal point; in other words, it is preceded by a decimal point and four zeroes. The actual number is therefore 0.00002197, a p-value very much smaller than the usual significance threshold of 0.05.

Task 5: *The perils of multiple testing*

There is always a chance that the result of any particular significance test will be incorrect as a consequence of random sampling error.

* It may show a significant difference when one is actually not present at the population level; this is known as a Type I error or false positive.
* It may fail to detect a true difference that exists; this is known as a Type II error or false negative.

Performing multiple tests on the same body of data can increase the chance of encountering Type I (false positive) errors among the results, which is why we should not run multiple t-tests if we have lots of different groups.

Several statistical techniques have been developed to address that problem, typically by requiring a stricter significance threshold for individual comparisons, so as to compensate for the number of inferences being made. We will get back to this when discussing ANOVA below.

Here is a quick task from Winters (2019) that demonstrates how easy it is to encounter Type I and Type II errors.

*A Type I error*

In the task, you will run several t-tests. Each t-test compare two sets of random data, each with ten observations, both drawn from a normal distribution. The set.seed() command ensures that we get the same ‘random’ numbers.

*A Type I error*

set.seed(42)# set random number seed

Run the following t-tests, each time observing the p-value. In each t-test, you are comparing two groups. Their respective observations (each n = 10) are randomly sampled from the same normal distribution, with μ = 0. This means that there should be no significant difference between our groups. Let’t see what happens.

t.test(rnorm(10), rnorm(10)) # p = not sig

t.test(rnorm(10), rnorm(10)) # p = not sig

t.test(rnorm(10), rnorm(10)) # p = not sig

t.test(rnorm(10), rnorm(10)) # p = sig

As you can see, the fourth t-test suggests that there is a significant difference. This is a Type I error since there actually isn’t a difference in the population.

*A Type II error*

We can now use R to simulate Type II errors, too. Again, run the following tests. Here, we have again two groups, but this time one with M = 1 and another with M = 0. That is, there is a difference in the population. This means any result that is p > 0.05 is a false negative.

Run the following commands and observe the p-values. As you will notice, the third test is not statistically significant, p > .05, a Type II error.

set.seed(42)

t.test(rnorm(10, mean = 1), rnorm(10, mean = 0))

t.test(rnorm(10, mean = 1), rnorm(10, mean = 0))

t.test(rnorm(10, mean = 1), rnorm(10, mean = 0))

*How do we deal with multiple testing?*

One influential school of thought says that we should aim to correct for multiple testing in our testing procedures. This practice is perhaps more commonly encountered in confirmatory research than in exploratory research; in exploratory research, we are often more interested in identifying the more likely hypotheses that deserve further replication studies, and thus a bit less worried about false positives.

The simplest (and probably the commonest) form of correction is known as the Bonferroni correction. This involves dividing the p-value that we intend to use as our significance level (i.e. usually 0.05) by the number of pairwise tests to be performed, in order to arrive at a much stricter p-value which we will then use in place of the original 0.05. So, if we were comparing four groups of data, we would divide 0.05 by 6 (the necessary number of pairwise tests for four groups) and arrive at a new significance criterion of p < 0.0083.

However, as you will already infer from this example, the Bonferroni correction rapidly becomes extremely conservative and consequently raises the risk of Type II (false negative) errors instead. This trade-off is always present in such corrections, and Type II errors are no less wrong than Type I errors. (In the context of a Covid-19 screening test, for instance, a Type I error might have you stuck at home for a week for no good reason, whereas a Type II error might allow you to unwittingly infect lots of other people…)

Some other scholars (e.g., Perneger, 1998, 1999) have therefore argued that one should not necessarily attempt to apply corrections for multiple testing, not only because they may be too conservative but also because they do not always make sense in practice. This is a very contentious area of statistics, and several alternative approaches to controlling for false positives have also been proposed. However, you may still see the Bonferroni correction used in research papers that you read, and reviewers of your own work might expect to see some comment about (not) correcting for Type I errors, even if you have decided to follow Perneger’s advice. I would therefore strongly encourage you to read Perneger’s very short and accessible papers, in order to find out a bit more about the issues involved in multiple testing.

Perneger, T.V. (1998). What’s wrong with Bonferroni adjustments. British Medical Journal, 316, 1236-1238.

Perneger, T.V. (1999). Adjusting for multiple testing in studies is less important than other concerns. British Medical Journal, 318, 1288.

Task 6: *Significance testing for continuous data: Two or more groups*

As discussed, the problem with conducting multiple t-tests for comparing sample means is that the Type I error rate increases as the number of t-tests increases. Bearing this in mind, R. A. Fisher developed the analysis of variance (ANOVA) test. This allows you to test the hypothesis of equality of population means while maintaing the Type I error rate at the established significance level (usually .05) for the entire set of comparisons.

There are different types of ANOVA, and the textbooks recommended on page 1 provide a helpful introduction. Here, we just want to focus on the **one-way ANOVA** and its non-parametric counterpart, the **Kruskal-Wallis test**.

*One-way ANOVA*

The one-way ANOVA involves the analysis of **one independent variable with two or more levels**. In our data set, this could be variable first\_lang, which records the native languages of our participants, and which has three levels, depending on the language with which our participants grew up, Language A, Language B or Language C. The variable first\_lang is also called our grouping variable or factor.

Our dependent variable will be performance on the maths exam (maths\_exam).

When it comes to performance on the maths exam, does it make a difference what first language you speak?

(We can pretend previous research has shown that some languages facilitate mathematical calculations because of the way the languages function. Or maybe pretend that the participants’ native languages could serve as rough indicators of where our participants grew up. A maths advantage could be due to the fact that speakers of language A are likely to reside in country A, which invests significantly more in maths education than other countries.)

Either way, changes in our dependent variable (math scores) in the ANOVA are presumed to be the result of changes in the independent variable.

We will not go into details, but broadly speaking the ANOVA looks at two types of variance.

* The variance *within* each group, i.e. the variation of scores within Language A group, within Language B group, and within Language C group. This is called the **within-groups variation**
* The variance *across* the groups, i.e. comparing the variances of Language A group, Language B group and Language C group. This is called **between-groups variation** .

The F-ratio is they key statistic for the ANOVA. You get this if you divide the between-groups variation by the within-groups variation. A lower F-ratio (< 1) suggests that there are no significant differences between the means of the samples. A higher ratio suggests that there is more variation.

We always expect some within-groups variation as a result of random sampling. However, if there is also between-groups variation that is not accounted for by the within-groups variation, then we assume that this is due to the independent variable (our factor, the native languages of our participants).

*Assumptions underlying ANOVA*

The ANOVA test is appropriate when:

* The data consists of random, independent samples from our population.
* The data is normally distributed.
* The variables being compared have common variance (equal variances).

*Visualize our data*

Let’s start by visualizing our data. We wish to inspect the performance of our three subgroups (Language A speakers, Language B speakers, Language C speakers) on the maths exam.

A Tukey boxplot gives us an idea of how the scores are distributed. The error bars represent the interquartile range, as discussed a few weeks ago.

boxplot(

maths\_exam ~ first\_lang,

data = language\_exams,

xlab = "Native languages",

ylab = "Maths scores",

frame = FALSE,

col = c("mistyrose", "turquoise", "darkseagreen")

)



The following mean plot is also helpful. This time, the error bars represent 95% confidence intervals.

install.packages('gplots')

library(gplots)

plotmeans(

maths\_exam ~ first\_lang,

data = language\_exams,

frame = FALSE,

xlab = "Native languages",

ylab = "Maths scores",

main = "Mean Plot with 95% CI"

)



*Compute one-way ANOVA test*

We can run a one-way ANOVA to test whether our factor of interest (the participants’ native languages) affects performance on our dependent variable (maths scores). (We should not use a t-test as we have three, not two, groups.)

The aov() function will run the ANOVA on our data, and the summary.aov() function can then be used to summarize the ANOVA model.

In the formula, we specify maths\_exam as our dependent variable and first\_lang as our independent variable (with three levels) or factor. We also tell R where to find the data, see language\_exams.

result\_aov <- aov(maths\_exam ~ first\_lang, data = language\_exams)

We get the summary of our results as follows.

summary(result\_aov)

Df Sum Sq Mean Sq F value Pr(>F)

first\_lang 2 26164 13082 40.23 2.24e-15 \*\*\*

Residuals 197 64061 325

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

To report the results of an ANOVA, we need to know the F statistic, the degrees of freedom (for between and within groups) and the p-value (see “Pr(>F)” above). The model summary indicates significant p-values with asterisks.

We can write this part of the results as follows: “A one-way ANOVA demonstrated that the effect of native language on maths scores was significant, F(2, 197) = 40.23, p < .001.” Or:

“We observed a statistically significant main effect of native language on maths scores, F(2, 197) = 40.23, p < .001.”

*Multiple pairwise-comparison between the means of groups: Tukey’s HSD test*

You will have noticed that there is a significant effect, but we didn’t say where! Are all three groups different from each other, or are just two different?

To find out, we need to run so-called “post-hoc” tests. There are many options, but the most-widely used is perhaps Tukey’s HSD test. (If you’re wondering, HSD stands for “honestly significant difference”.)

We can use Tukey’s HSD to conduct multiple pairwise-comparisons between the means of the three groups by means of the TukeyHSD() function.

The function takes the fitted ANOVA as its argument.

TukeyHSD(result\_aov)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = maths\_exam ~ first\_lang, data = language\_exams)

$first\_lang

diff lwr upr p adj

Lang\_B-Lang\_A 4.271679 -3.245571 11.78893 0.3737241

Lang\_C-Lang\_A 25.918238 18.450396 33.38608 0.0000000

Lang\_C-Lang\_B 21.646559 14.447520 28.84560 0.0000000

The diff colum gives you the difference between means, and the lwr and upr columns helpfully provide the lower and upper limits of the 95% confidence intervals. Finally, as you can see in the p adj column, our p-value was adjusted for multiple comparisons. We find that there is a significant difference between Language A group and Language C group, with the latter outperforming the latter on the maths exam by Mdiff = 25.92, p < .001. And there is also a significant difference between Language A group and Language C group, again with the former outperforming the latter, Mdiff = 21.65, p < .001. But there was not significant difference between Language A group and Language B group, Mdiff = 4.27, p = .374.

*Checking the ANOVA assumptions: Variance*

As mentioned, the ANOVA test assumes that, the data are normally distributed and that the variance across groups are homogeneous.

We can run Bartlett’s test to check for the homogeneity of variances using the bartlett.test() function.

bart\_all <- bartlett.test(maths\_exam ~ first\_lang, data = language\_exams)

bart\_all

Bartlett test of homogeneity of variances

data: maths\_exam by first\_lang

Bartlett's K-squared = 23.808, df = 2, p-value = 6.762e-06

The output suggests that there is a significant difference, K2 = 23.808, p < .001, so we have “unequal variances.” This means, of course, we should use a different ANOVA!

*One-way ANOVA with unequal variances*

As you can see in the table on page 11, a good alternative procedure is the Welch one-way ANOVA, which does not assume equal variances.

To run this test, we use the oneway.test() function.

oneway.test(maths\_exam ~ first\_lang, data = language\_exams)

One-way analysis of means (not assuming equal variances)

data: maths\_exam and first\_lang

F = 42.938, num df = 2.00, denom df = 126.28, p-value = 5.932e-15

The results are similar to previous ANOVA, the difference is again significant. We can write this up as follows: “A Welch one-way ANOVA demonstrated that the effect of native language on maths scores was significant, F(2, 126.28) = 42.94, p < .001.”

*Pairwise t-tests with no assumption of equal variances*

Since there is a significant difference somewhere among the groups, let’s conduct our post-hoc tests as before.

Given the unequal variances, we can do this with the following t-test.

pairwise.t.test(language\_exams$maths\_exam, language\_exams$first\_lang,

p.adjust.method = "BH", pool.sd = FALSE)

Pairwise comparisons using t tests with non-pooled SD

data: language\_exams$maths\_exam and language\_exams$first\_lang

Lang\_A Lang\_B

Lang\_B 0.14 -

Lang\_C 5.4e-15 5.1e-09

P value adjustment method: BH

As you can see, the difference between Language C and the other two groups are significant, see the highlighted p-values, both below p < .01. This test used the Benjamini-Hochberg correction (BH) rather than the Bonferroni method used above.

*Non-parametric alternative to the one-way ANOVA test*

Last but not least. If our data is not normally distributed, we can use the **Kruskal-Wallis** **rank sum test** as the non-parametric alternative to the one-way ANOVA.

If you the run the following command with the kruskal.test() function, you should find out soon if there is a significant difference.

kruskal.test(maths\_exam ~ first\_lang, data = language\_exams)

Kruskal-Wallis rank sum test

data: maths\_exam by first\_lang

Kruskal-Wallis chi-squared = 60.378, df = 2, p-value = 7.746e-14

As before, there is a significant effect. “A Kruskal-Wallis rank sum test revealed a significant difference, H(2) = 60.38, p < .001.” The H-statistics is the chi-squared value.

And as before, we need to run post-doc tests to discover where the statistical difference is.

Since the data are not normally distributed, the Tukey HSD test needs to be replaced by a nonparametric alternative. The Dunn test is a good nonparametric post-hoc alternative.

To run the Dunn test, we proceed as follows.

install.packages('FSA')

library(FSA)

dunnTest(maths\_exam ~ as.factor(first\_lang),

data = language\_exams,

method = "bonferroni")

Comparison Z P.unadj P.adj

1 Lang\_A - Lang\_B -1.621220 1.049704e-01 3.149112e-01

2 Lang\_A - Lang\_C -7.276655 3.422000e-13 1.026600e-12

3 Lang\_B - Lang\_C -5.855474 4.756520e-09 1.426956e-08

dunn <- dunnTest(maths\_exam ~ as.factor(first\_lang),

data = language\_exams,

method = "bonferroni")

The head() function is a good way to display the Dunn test results. This is what our results look like.

head(dunn)

$method

[1] "Bonferroni"

$res

Comparison Z P.unadj P.adj

1 Lang\_A - Lang\_B -1.621220 1.049704e-01 3.149112e-01

2 Lang\_A - Lang\_C -7.276655 3.422000e-13 1.026600e-12

3 Lang\_B - Lang\_C -5.855474 4.756520e-09 1.426956e-08

$dtres

[1] " Kruskal-Wallis rank sum test"

[2] ""

[3] "data: x and g"

[4] "Kruskal-Wallis chi-squared = 60.3779, df = 2, p-value = 0"

[5] ""

[6] ""

[7] " Comparison of x by g "

[8] " (Bonferroni) "

[9] "Col Mean-|"

[10] "Row Mean | Lang\_A Lang\_B"

[11] "---------+----------------------"

[12] " Lang\_B | -1.621220"

[13] " | 0.3149"

[14] " |"

[15] " Lang\_C | -7.276654 -5.855473"

[16] " | 0.0000\* 0.0000\*"

[17] ""

[18] "alpha = 0.05"

[19] "Reject Ho if p <= alpha"

The table at the bottom quickly summarizes the results. Again, we find a difference between Language C group and Language A group, and Language C group and Language B group. The asterisks allow us to spot the significant test results quickly.

*Task 7: Reporting the results of our tests*

Above, I have provided one-sentence reports for each significance test (e.g., “A paired-samples t-test has shown…”). When reporting the significance test, you generally need to report the two or three items of information printed on the line immediately below “data” in the R output – i.e., the test statistic; the degrees of freedom (where they exist); and the p-value.

Importantly, you should also report the “statistic of interest” and its 95% confidence interval.

The statistic of interest is not the same thing as the test statistic! Compare the two in the table below.

|  |  |  |
| --- | --- | --- |
| Significance test | Test statistic | Statistic of interest |
| t-test (standard and Welch) | t-value | difference in means |
| Wilcoxon-Mann-Whitney | W-value | difference in location |
| Paired t-test | t-value | mean of paired differences |
| Wilcoxon rank sum test | V-value | (pseudo) median of paired differences |

Mostly, these are printed as the very last line of the R output. With the two-sample t-tests, you can either present the two sample means or you can calculate the difference between those two means. If you do the latter, make sure that you perform the subtraction in the same order that the figures appear in the R output, otherwise your result – which may be a negative number – will not match the signs on the 95% CI that “R” has produced.

These “statistics of interest” are your measures of effect size. You should always pay attention to them, and comment on them when appropriate. As we said last time, they tell you how big a difference is, whereas the test of significance merely suggests whether the true difference at the population level is likely to be different from (exactly) zero. (And the 95% CI tells you how precise your effect-size estimate is likely to be.)

Style sheets will generally tell you to round all figures to 2 or 3 decimal places. Within this constraint, you should always quote the exact p-value, down to the smallest number that you can have with the prescribed number of decimal places – i.e., 0.01 for 2 decimal places and 0.001 for 3 decimal places. P-values smaller than this can be written as the inequalities p < 0.01 or p < 0.001, respectively. Otherwise, do not write p-values as inequalities (e.g. p < 0.05).

Following the APA style sheet, the degrees of freedom for the t-test are enclosed in round brackets after the letter “t”, whilst the 95% CI is enclosed in square brackets; and all numbers in APA style are rounded to 2 decimal places.

By way of illustration, have a look at the following reports.

*Example of how you might write-up a t-test result*

“The mean score on language exam 1 was 65.45 (SD = 14.27) for undergraduates and 33.20 (SD = 18.38) for postgraduates. Cohen’s d is 1.56, 95% CI [1.62, 2.30], which represents a large effect size (Cohen, 1988). An independent-samples t-test further showed that this difference between groups is significant, t(198) = 13.863, p-value < .001. That is, undergraduate students significantly outperformed their postgraduate counterparts on the first language exam.”

*Example of how you might write-up an ANOVA test result*

“The mean score on the maths exam was 55.97 (SD = 11.21) for language group A, M = 60.24 (SD = 20.28) for language group B, and M = 81.89 (SD = 20.24) for language group C. Our calculation of Cohen’s d to measure the difference between the three means revealed a large effect, d = 1.553.[[1]](#footnote-2) A one-way ANOVA demonstrated that the effect of native language on maths scores was significant, F(2, 197) = 40.23, p < .001. Post-doc comparisons using Tukey’s HSD test further revealed a significant difference between language group A and C, Mdiff = 25.92, p < .001, and between language group B and C, Mdiff = 25.92, p < .001. The difference between language groups A and B was not significant, Mdiff = 4.27, p = .374. In other words, speakers of language C clearly outperformed the speakers of languages A and B in the maths exam.”

Note: Since the maths scores are not normally distributed, the appropriate test is the Kruskal-Wallis non-parametric test. However, above I report the results of the regular one-way ANOVA with Tukey’s HSD post-hoc comparisons as these are far more frequently used, so showing how to report the ANOVA results with these tests made more sense.

*Final reminders*

The p-value does not tell you whether your results are large or important. That is an inference you can only make on the basis of the effect size, and in the broader context of what you already know about your research area.

The p-value tells you the probability (in the long run) of seeing these or more extreme results, if the null hypothesis is true each time. So, a sufficiently small p-value is generally taken to indicate that the null hypothesis in unlikely to be true.

BUT:

We will always expect to see a small number of statistically significant results, even when the null hypothesis is true. This is a consequence of random sampling. We can only really decide whether the null hypothesis is false by repeating the study a number of times.

Likewise, if you fail to reject the null hypothesis in a study, it could be that your sample was simply too small to reject it on this occasion. Failing to reject the null hypothesis on one occasion does not necessarily mean that it is true, any more than rejecting it means that it is false. Again, studies need to be replicated.

In real life, what we call “point null” hypotheses (like “no difference” or “zero”) are very rarely true. There is nearly always some difference, however small it may be, hence the importance of looking at effect sizes as well.

1. There are ways of calculating Cohen’s d and other standardized effect size measures for more than two groups. [↑](#footnote-ref-2)