Steve Doogue

Common statistical tests are linear models: a work through

Contents

1

Introduction

This is a reworking of the book Common statistical tests are linear models (or: how to teach stats), written by Jonas Lindeløv. The book beautifully demonstrates how many common statistical tests (such as the t-test, ANOVA and chi-squared) are special cases of the linear model. The book also demonstrates that many non-parametric tests, which are needed when certain test assumptions do not hold, can be approximated by linear models using the *rank* of values.

This approach offers a way to greatly simplify the teaching of introductory statistics, using the simple model of the form $y = a + b \cdot x$ which is familiar to most students. This approach brings coherence to a wide-range of statistical tests, which are usually taught to students as independent tools with a potentially-overwhelming array of names. The approach also helps to explain the intuition underlying statistical tests, drawing on the familiar concept of linear regressions, which emphasizes understanding over rote learning.

The purpose of creating this book is to solidify my understanding of this approach. I do this by reproducing the examples provided by Lindeløv; by expanding on areas where there were gaps in my knowledge; and by paraphrasing some of the explanations, using concepts and terms with which I am more familiar. The book may also be helpful to others who want to follow along with Lindeløv's book, but who require more background on some of the concepts being discussed.

Credit for this book should be attributed to Jonas Lindeløv, though I am of course responsible for any errors in my own interpretation.

Note that some of the data used in this book varies from that used by Lindeløv, and so some of the test results differ. However, the concepts being discussed are exactly the same.

Steve Doogue

2

The data

Note that some of the data used in this book varies from those used by Lindeløv. Some of the test results will therefore differ.

2.1 Sample values

Most of the examples in the book are based on three imaginary samples (`x`, `y` and `y2`). Each is normally distributed and made up of 50 observations.

We start by creating a function that will allow us to produce samples of a given size (`N`) with a specified mean (`mu`) and standard deviation (`sd`):

```
rnorm_fixed <- function(n, mu = 0, sd = 1) {
  as.numeric(scale(rnorm(n)) * sd + mu)
}
```

Now we can create our three samples:

```
# Set the seed so our 'random' results are reproducible
set.seed(40)

# Create the samples (use the same order as original book - y, x, then y2)
y <- rnorm_fixed(50, mu = 0.3, sd = 2)
x <- rnorm_fixed(50, mu = 0, sd = 1)
y2 <- rnorm_fixed(50, mu = 0.5, sd = 1.5)
```

Let's also combine these samples into 'wide' and 'long' data frames. This doesn't change any of the values, it just rearranges the data

into different layouts which can sometimes be easier to work with (e.g. when producing plots):

```
# The wide layout is a dataframe with three columns, one for each of x, y a
mydata_wide <- tibble(x = x, y = y, y2 = y2)

# the long layout has two columns: one listing the group to which each
# observation elongs (x, y or y2), and another column with the correspondin
mydata_long <- mydata_wide %>%
  gather(group, value, x:y2)
```

Here's what our three samples look like when plotted. Note the different mean (0.0, 0.3, and 0.5) and the different 'spread' of values for each group (reflecting their different standard deviations).

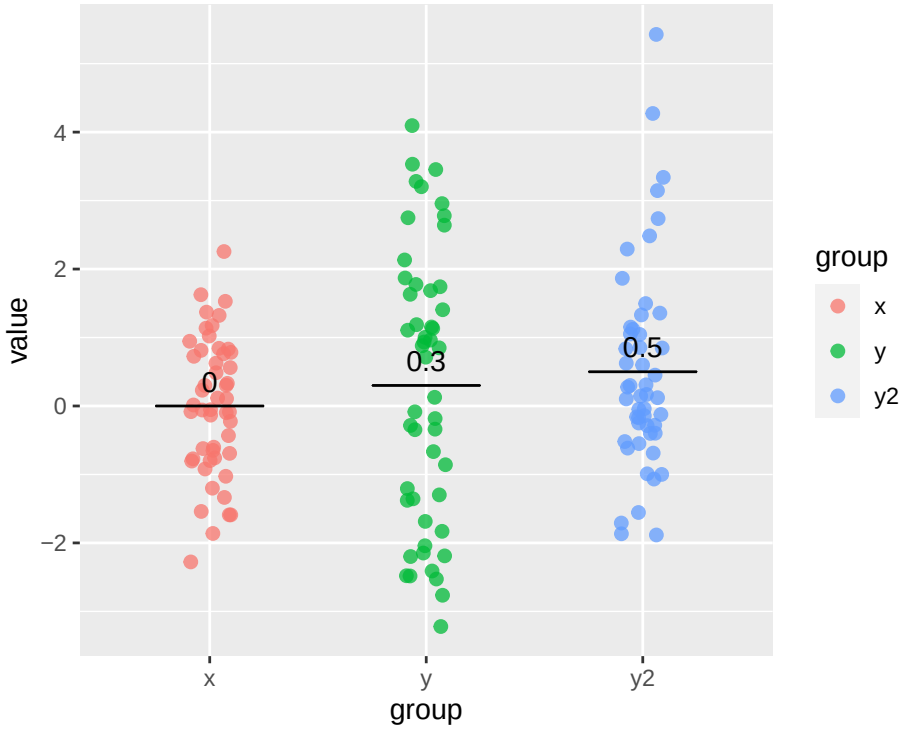

FIGURE 2.1 Our sample data

2.2 'Signed rank' values

Most common tests demonstrated in the book use the *actual* values from the samples above. However, some tests (i.e. non-parametric tests) also use their *rank-transformed* values. In some cases, the *signed rank* of the values is used.

What is meant by signed ranks? The signed rank is found by: 1. taking the *absolute* value of each observation in the original sample; that is, expressing all the values as a positive number; 2. finding the rank of each of these absolute values, where the smallest absolute value has a rank of 1; and 3. giving each of these ranks the same sign (+ or -) as the original value.

For example, the numbers -2, 3, 7, -25, -30, 31 would have the signed ranks of -1, 2, 3, -4, -5, 6.

We create a function, `signed_rank()`, that we can use later to convert our actual values into signed ranks:

```
# Takes any list of values, x, and calculates their signed rank
signed_rank <- function(x) sign(x) * rank(abs(x))
```

3

The linear model

3.1 Overview of the linear model

The premise of this book is that many common statistical tests are really just special cases of the **linear model**.

This section provides an overview of the linear model. The description will be somewhat informal as it aims to provide an intuitive explanation rather than a rigorous technical description.

The linear model, or linear regression model, estimates the relationship between one continuous variable and one or more other variables. It is assumed that the relationship can be described as a straight line (hence the term 'linear').

For example, say we are looking at a variable y and we want to know its relationship with a variable x. We assume that the relationship can be expressed as a mathematical relationship between y (the dependent, or response variable) and x (the explanatory variable):

$$y = \beta_0 + \beta_1 x$$

To illustrate what this equation is showing, imagine we have six observations of variables x and y, which can be plotted as follows:

We assume that this relationship can be represented by a straight line. The line is composed of an intercept (β_0) and a slope (β_1). Each point on this line represents our *predicted* value of y for a given value of x.

So how do we estimate the intercept and slope? In other words, how do we estimate what the values of β_0 and β_1 should be?

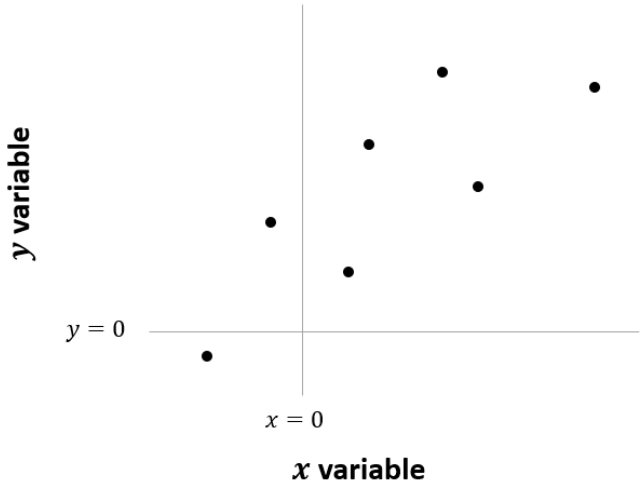

FIGURE 3.1 Six observations of x and y

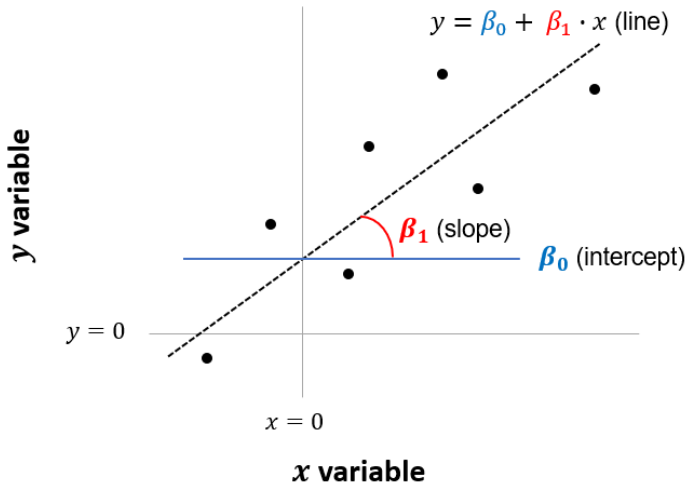

FIGURE 3.2 Intercept and slope of model

Linear regression estimates the line (i.e., slope and intercept) that minimizes the difference between our *predicted* values of y and the *actual* values of y that were observed in the original sample. These difference are referred to as "residuals".

More specifically, linear regression minimizes the sum of the *squared* value of these residuals. So in the figure below, linear regression is used to estimate the line that would minimize the combined area of the purple squares. For this reason, the method is sometimes referred to as an "ordinary least squares" (OLS) regression.

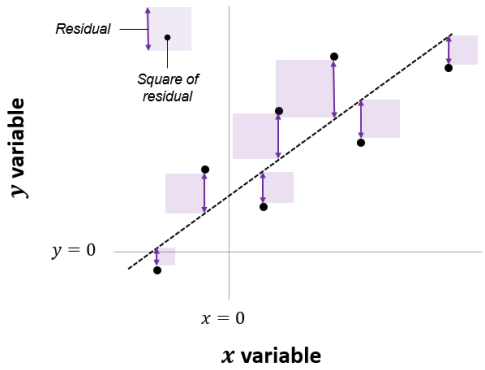

FIGURE 3.3 Line of best fit minimizes the sum of squared residuals

In the example above, we considered a dependent variable (y) that was being "explained" by one other variable, or predictor (x). But this can be expanded to include multiple predictors, for example with the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 ... + \beta_i x_i$$

This is referred to as *multiple regression*. In this case we still have only one intercept (β_0) but many slopes that need to be estimated (β_1, β_2 and so on). The same method can be used to estimate these parameters as was illustrated above, i.e. an OLS regression which minimizes the sum of squared residuals. This cannot be illustrated in a two-dimensional diagram, as was the case when a single predictor was used, but the exact same concept applies.

So that's the linear model. The key premise of Lindeløv's book many statistical tests are just special cases of this:

> Everything below, from one-sample t-test to two-way ANOVA are just special cases of this system. Nothing more, nothing less.

3.2 Estimating linear models in R

To estimate linear models in R we use the `lm()` function.

For example, say we want to estimate the relationship between our sampled data for `x` and `y`. We will apply the following model: $y = \beta_0 + \beta_1 x$

In R, this can be written as follows:

```
# Represents y = beta0 + beta1 * x
lm(y ~ 1 + x)
```

In the original book, the specified model is accompanied by the null hypothesis, $H_0 : \beta_1 = 0$. This is equivalent to saying that there is no relationship between `x` and `y`. The output from our linear model tells us whether there could be grounds for rejecting this null hypothesis of "no relationship", in favor of the alternative hypothesis that there *is* a relationship.

A key output of interest, or test statistic, is the p-value.

- A small p-value (conventionally 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis of no relationship.[1]
- A larger p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis of no relationship.

[1] The p-value is the probability of getting a value as extreme or more extreme than the one observed in your sample, under the assumption that the null hypothesis is true.

Let's see the output of the linear model for our sample data:

```
lm(y ~ 1 + x) %>% summary() #%>% print(digits = 5)
```

```
##
## Call:
## lm(formula = y ~ 1 + x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3393 -1.6593  0.3349  1.3629  3.5214
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3000     0.2780   1.079    0.286
## x            -0.4636     0.2808  -1.651    0.105
##
## Residual standard error: 1.966 on 48 degrees of freedom
## Multiple R-squared:  0.05372,    Adjusted R-squared:  0.03401
## F-statistic: 2.725 on 1 and 48 DF,  p-value: 0.1053
```

As seen in the output above, β_1 (the coefficient on `x`) has a p-value of 0.1053. This means we would fail to reject the null hypothesis that there was no relationship between `x` and `y`.

3.3 Assumptions

The inferences drawn from the linear model, as described above, are only valid if a number of assumptions hold. The following diagram, found in the book Broadening Your Statistical Horizons, can be used to explain these assumptions:

The four main assumptions can be remembered using the acronym 'LINE':

L: there is a **linear** or straightline relationship between the mean response (Y) and the explanatory variable (X). In the figure above,

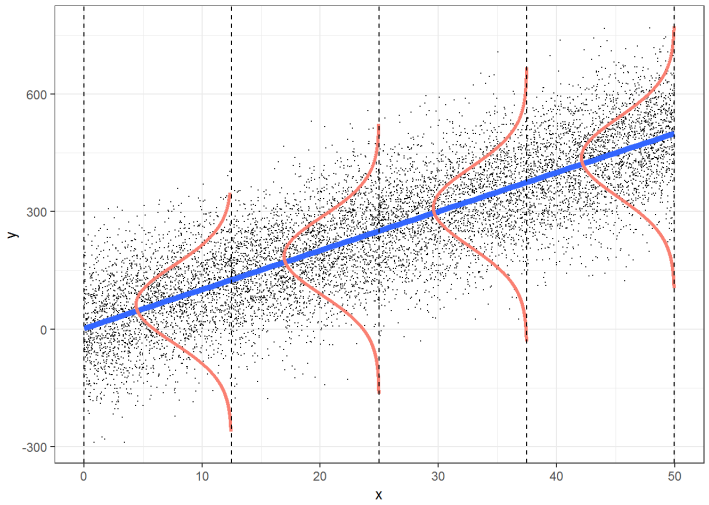

FIGURE 3.4 Ordinary least squares assumptions

the mean value for Y at each level of X falls on the blue regression line.

I: the errors are **independent** — there's no connection between how far any two points lie from the regression line. This is usually more of an issue in time series data (i.e. where data is collected from the same entity over time, e.g. a stock price) than cross-sectional data (where data is collected from many entities a single point in time, e.g. exam marks for a group of students).

N: the dependent variable is **normally** distributed at each level of X. At each level of X, the values for Y follow a normal or 'bell-shaped' distribution as shown in red in the figure above.

E: there is **equal variance** (or 'homoscedasticity') - the variance or 'spread' of the responses is equal for all levels of X. The spread in the Y's for each level of X is the same, as shown above.

A more detailed explanation of these assumptions is provided here by Laerd.

4

Correlations

Correlation is a measure of the strength and direction of association that exists between two variables. Correlation coefficients (r) assume values in the range from -1 to $+1$, where ± 1 indicates the strongest possible positive or negative correlation and 0 indicates no linear association between the variables.

4.1 Pearson correlation

We start by looking at the Pearson correlation coefficient. Here we compare the built-in test (in R) for the Pearson correlation with the equivalent linear model.

The equivalent linear model is the basic regression of y on x, specified as follows:

$$y = \beta_0 + \beta_1 x \qquad H_0 : \beta_1 = 0$$

The two tests are written using the following R code:

```
# Pearson (built in test)
cor.test(y, x, method = "pearson")
# Linear model
lm(y ~ 1 + x) %>% summary() %>% print(digits = 8)
```

If you were to run the code above, you would see the following the key statistics found in the output of each test:

The output shows that the correlation coefficient (r) has a p-value of 0.1053, which is exactly the same as the p-value for the slope of the linear model. In this case, we would not reject the null

TABLE 4.1 Pearson vs linear model

Test	r	slope	t	p-value
cor.test (Pearson)	-0.2318	NA	-1.6507	0.1053
lm	NA	-0.4636	-1.6507	0.1053

hypothesis that there was no correlation between the two variables (at the 0.05 level of significance).

The main difference is that the linear model returns the *slope* of the relationship, β_1 (which in this case is -0.4636), rather than the correlation coefficient, r. The slope is usually much more interpretable and informative than the correlation coefficient.

Additional note (optional):

It may be useful to understand how the Pearson correlation coefficient (r) and the regression coefficient or slope (β_1) are related, which is by the following formula:

$$\beta_1 = r \cdot sd_y / sd_x$$

This shows that:

- When both `x` and `y` have the same standard deviations (sd_x and sd_y) then the slope (β_1) will be equal to the correlation coefficient (r)
- The ratio of the slope to the correlation coefficient (β_1 / r) is equal to the ratio of the standard deviations (sd_y / sd_x). In this example, the standard deviation of `y` is exactly twice as large as `x`, which is why the slope has twice the magnitude of the correlation coefficient.
- The slope from the linear model will always have the same sign (+ or -) as the correlation coefficient (as standard deviations are always positive).

4.2 Spearman correlation

There will be times when it is more appropriate to use the **Spearman rank correlation** than the Pearson correlation. This could be the case when:

1. The relationship between the variables is not linear, i.e. not a straight line;
2. The data is not normally distributed;[1]
3. The data has large outliers; or
4. When you are working with ordinal rather than continuous data.[2]

The Spearman correlation is a *non-parameteric* test as it does not require that the parameters of the linear model hold true. For example, there does not need to be a linear relationship between the two variables, and the data does not need to be normally distributed.

The Spearman rank correlation is the same as a Pearson correlation but using the *rank* of the values in our samples. This is an approximation only, which Lindeløv shows is approximate when the sample size is greater than 10 and almost perfect when the sample is greater than 20.

This is also the same as the linear model using rank-transformed values of x and y:

$$rank(y) = \beta_0 + \beta_1 \cdot rank(x) \qquad H_0 : \beta_1 = 0$$

For a comparison of the Spearman test, the Pearson test (using

[1] Technically the variables should have *bivariate* normality, which means they are normally distributed when added together, but this is complex and so it is common just to assess whether the variables are individually normal (explained here on the Laerd website). If bi-variate normality does not hold then you will still get a fair estimate of r, but the inferential tests (t-statistics and p-values) could be misleading (explained here).

[2] see Chapter 9 for a description of the different types of data.

TABLE 4.2 Spearman, Pearson (ranks) and linear model (ranks)

Test	correlation	slope	p.value
cor.test (Spearman)	-0.2266	NA	0.1135
cor.test (Pearson with ranks)	-0.2266	NA	0.1135
lm (with ranks)	NA	-0.2266	0.1135

ranks) and the linear model (also using ranks) we run the following code:

```
# Spearman
cor.test(y, x, method = "spearman")

# Pearson using ranks
cor.test(rank(y), rank(x), method = "pearson")

# Linear model using rank
lm <- lm(rank(y) ~ 1 + rank(x))
  lm %>% summary() %>% print(digits = 5) # show summary ouput
```

The output of this code is as follows:

This shows that the results are all the same (or at least very close approximations). When ranks are used, the slope of the linear model (β_1) has the same value as the correlation coefficient (r). Note that the slope from the linear model has an intuitive interpretation, which is the number of ranks y changes for each change in rank of x.

Given the similarity of these tests, Lindeløv notes that:

> One interesting implication is that many "non-parametric tests" are about as parametric as their parametric counterparts with means, standard deviations, homogeneity of variance, etc. - just on rank-transformed data.

Finally, the figure below (reproduced from the original book) illustrates how the Pearson and Spearman correlations are equivalent

to linear models, with the latter being based on rank-transformed values.

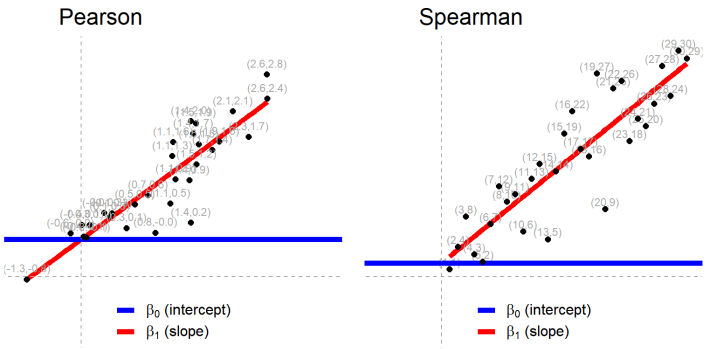

FIGURE 4.1 Pearson and Spearman correlations as linear models

5

One mean

This set of tests deals with a single mean. They tell us whether, based on our sample, we have reason to believe that the mean of the underlying population differs from an some specific level (the 'null hypothesis').

In some cases we might have two samples with *paired* observations (e.g. an athlete's average running speed before and after a new training technique is used). In this case we take the *difference* between the paired observations (the change in each athlete's running speed) and treat this as a single measurement. We then assess whether, based on the mean of these differences in our sample, we have reason to believe that the 'true' mean differs from a specified level (e.g. zero, representing no change in running speed).

The *one sample* and *paired sample* cases are addressed in turn, below.

5.1 One sample

5.1.1 One-sample t-test

A one-sample t-test (or one-sample Student's t-test) is used to determine whether a sample could have come from a population with a specified mean. This population mean is not always known and may only be theoretical / hypothesized.

For example, say we take a sample of pupils who have been taught using a new teaching method. These pupils receive an average score of 70% in an end-of-year test, while the average score for all pupils

nationally is 60%. We want to know how often we are likely to see a sample average of 70% (or more extreme) if the 'true' average for pupils receiving the new teaching method was 60%, i.e. if there was no difference from the national average.

Student's t-test:

In R, we can run a one-sample Student's t-test using the built-in `t.test()`. The documentation for this test can be found here.

Equivalent linear model:

The one-sample t-test is the same as a linear model containing a dependent variable, y, and a constant only. That is, a model with the form:

$$y = \beta_0 \qquad H_0 : \beta_0 = 0$$

This is the same as the equation $y = \beta_0 + \beta_1 \cdot x$ that was shown earlier (Chapter 3), but having dropped the last term. This is shown graphically below. Note that this is the same as a linear regression in which all values of x are treated as zero:

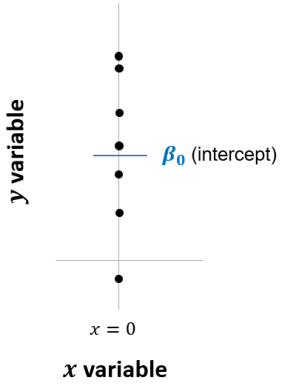

FIGURE 5.1 Linear model equivalent to one-sample t-test

The estimated intercept (β_0) is simply the average of all the values in our sample. The linear regression returns the t-statistic and p-value based on the null hypothesis that the true average is equal to zero (i.e. $\beta_0 = 0$).

TABLE 5.1 One-sample t-test and linear model

Test	mean / intercept	t	p.value	conf.low	conf.high
t.test	0.3	1.0607	0.294	-0.2684	0.8684
lm	0.3	1.0607	0.294	-0.2684	0.8684

Comparison:

Here is how you would run the t-test and equivalent linear model in R:

```
# t-test (built-in test)
t.test(y, mu = 0, alternative = "two.sided")

# Linear model
lm <- lm(y ~ 1)
  lm %>% summary() %>% print(digits = 5) # show summary output
  confint(lm) # show confidence intervals
```

The output of this code, including the 95 percent confidence intervals, is shown in the table below. The outputs are exactly the same!

Based on the p-value above, despite having a sample average of 0.3 we would not reject the null hypothesis that the true or population average was zero (at the 0.05 level of significance).

Testing a null hypothesis other than zero:

By default, the t-test has a null hypothesis is that the sample comes from a population with a mean of *zero*. This was written as $H_0 : \beta_0 = 0$ above. What if we want to use a different null hypothesis? For example, to how likely it is that our sample `y` came from a population with a mean of 0.1?

This can be done by specifying the null hypothesis in the built-in t-test (in this case using the argument `mu = 0.1`) or, in the linear model, by first subtracting this amount from the dependent variable (using `y - 0.1` as the dependent variable).

The code is as follows:

```
# t-test (built-in test)
t.test(y, mu = 0.1, alternative = "two.sided")

# Linear model
lm((y - 0.1) ~ 1) %>%
  summary() %>%
  print(digits = 5) # show summary output
```

The two tests above give identical t statistics and p-values.

5.1.2 Wilcoxen signed-rank

What if we cannot assume that our population is normally distributed? Here we need to use a non-parametric version of the t-test.

Wilcoxen signed-rank test:

In this case we could use the Wilcoxon signed-rank test, which is the non-parametric version of the t-test. Specifically, we use the one-sample Wilcoxon signed-rank test. This is used to determine whether the *median* of the sample is equal to a theoretical value, such as zero, under the assumption that the data is symmetrically distributed.

The Wilcoxon signed-rank test is very similar to the linear model described above, but using the *signed ranks* of y instead of y itself. The concept of signed ranks was explained in section 2.2.

It is implemented in R using the wilcox.test, as shown below.

Equivalent linear model:

In this case the equivalent linear model is:

$$signed\_rank(y) = \beta_0 \qquad H_0 : \beta_0 = 0$$

Lindeløv shows that the linear model will be a good approximation of the Wilcoxon signed-rank test when the sample size is larger than 14 and almost perfect when the sample size is larger than 50.

TABLE 5.2 Wilcox signed rank test and linear model

Test	mean / intercept	p.value
wilcox.test	NA	0.3693
lm (signed ranks)	3.74	0.3721

Comparison:

The code below is for the wilcox.test in R and the equivalent linear model using signed ranks:

```
# Wilcox test (built-in)
wilcox.test(y)

# Equivalent linear model
lm <- lm(signed_rank(y) ~ 1)
  lm %>% summary() %>% print(digits = 8) # show summary output
```

The two tests give similar (though not quite identical) p-values.

5.2 Paired samples

5.2.1 Paired-sample t-test

A **paired-sample t-test** (sometime called a dependent-sample t-test) is used to compare two population means where you have two samples, in which observations in one sample can be paired with observations in the other sample.

Common applications of the paired-sample t-test include controlled studies or repeated-measures designs. Examples of when you might use this test include:

- Before and after observations on the same subjects (e.g. students' test results before and after taking a course).
- A comparison of two different treatments, where where the treat-

ments are applied to the same subjects (e.g. athletes' ability to lift weights following two different warm-up routines).
- A comparison of two different measurements, where where the measurements are applied to the same subjects (e.g. blood pressure measured using two types of machines).

Paired-sample t-test / dependent-sample t-test:

The paired-sample t-test determines whether the average difference between two sets of observations is zero. To run this in R, we use the same Student's t-test as above, but now include both variables as arguments and specify that we have paired observations (using the argument `paired = TRUE`).

Equivalent linear model:

The equivalent linear model is the *difference* between the two observations regressed against a constant:

$$y_2 - y_1 = \beta_0 \qquad H_0 : \beta_0 = 0$$

Comparison:

We can compare the t-test and linear model using our sample data (section 2.1) as an example. Recall that `y` had a mean of 0.3, and `y2` had a mean of 0.5.

The code and the outputs are shown below. Again, the linear model gives exactly the same results as the t-test!

```
# t-test (built-in test)
t.test(y2, y, paired = TRUE, mu = 0, alternative = "two.sided")

# Equivalent linear model
lm <- lm(y2 - y ~ 1)
  lm %>% summary() %>% print(digits = 8)
  confint(lm)
```

This shows us that the difference in averages between `y` and `y2` is 0.2, as would be expected, but that this difference is not statistically significantly difference from zero ($p = 0.601$) at the 0.05 level of significance.

TABLE 5.3 Paired-sample t-test and linear model

Test	mean of diff / intercept	t	p.value	conf.low	conf.high
t.test	0.2	0.5264	0.601	-0.5635	0.9635
lm	0.2	0.5264	0.601	-0.5635	0.9635

5.2.2 Wilcoxen matched pairs

If the necessary assumptions do not hold for a paired-sample t-test - such as normally distributed data - we can use the non-parametric counterpart. This is the **Wilcoxen matched pairs** test. The only difference from the Wilcoxon signed-rank test is that it's testing the signed ranks of the pairwise $y - x$ differences.

This is used to test the null hypothesis that the *median* of the differences in our paired observations are zero. An assumption is that the values of the pairwise differences are symmetrically distributed, as explained by Laerd.

For this we use the built-in wilcox.test in R, including both variables and specifying that we have paired observations (the argument `paired = TRUE`).

The equivalent linear model is exactly the same above but using *signed rank* of the difference between observations, i.e.:

$$signed\_rank(y_2 - y_1) = \beta_0 \qquad H_0 : \beta_0 = 0$$

A comparison of the outputs is shown below. The p-values are almost identical. The t-test has also been included, this time using signed ranks, to show that this is the same as the linear model.

```
#  Wilcox test (built-in test)
wilcox.test(y2, y, paired = TRUE, mu = 0, alternative = "two.sided")

# Equivalent linear model
lm <- lm(signed_rank(y2 - y) ~ 1)
  lm %>% summary() %>% print(digits = 8)
```

TABLE 5.4 Wilcox test and linear model, with paired samples

Test	mean / intercept	p.value
wilcox.test	NA	0.8243
lm (signed ranks)	0.94	0.8232
t.test (signed ranks)	0.94	0.8232

```
# t-test using signed ranks (built-in test)
t.test(signed_rank(y2 - y), mu = 0, alternative = "two.sided")
```

Based on the results above, we would not reject the null hypothesis that the *median* change between `y` and `y2` was zero (p = 0.8232), at the 0.05 level of significance.

6

Two means (independent samples)

These tests compare the means of two independent or unrelated groups, in order to determine whether there is statistical evidence that the population means are significantly different. Examples include:

- Do first year graduate salaries differ based on gender? and
- Is there is a difference in test anxiety based on educational level (undergraduate vs postgraduate)?

A key question is whether or not the **variances** of the two samples being tested are equal.[1] In general terms:

- If the samples have *equal* variance then an **independent-sample t-test** (Student's t-test) could be used.
- If the samples have *unequal* variance then the **Welch's t-test** can be used, as this does not assume identical variances.
- If the samples are *not normally distibuted*, or if the other requirements of the above tests are not met, then the non-parametric **Mann-Whitney U** test could be used.

This is somewhat simplistic, and the choice of tests is not always clear (see the discussion here on StackExchange for example). The following sections goes through each of these tests in turn, and describes how they can be approximated by an equivalent linear model.

[1] This can be tested with the Levene's Test for Equality of Variances.

6.1 Independent t-test

The independent t-test can be used to compare the means of two unrelated groups. Assumptions include:

- Independence of observations (independent samples / groups);
- Normal distribution (approximately) of the dependent variable for each group, though this might not be an issue for large samples;
- No outliers; and
- Homogeneity of variances; i.e. variances are approximately equal across the two groups.

Student's t-test:

In R, we can assesses the difference between two groups using the t.test. This has the default assumptions that the two samples are *not* paired, and that their variances are *not* equal.

Equivalent linear model:

The equivalent linear model uses a dummy variable to represent the two groups (see the explanation of dummy variables explained by Lindeløv in section 5.1.3 of the original book).

The linear model takes the form:

$$y = \beta_0 + \beta_1 \cdot x_i \qquad H_0 : \beta_0 = 0$$

where x_i is a dummy variable taking the value of 0 or 1, indicating whether observation i is from the reference group (0) or the other group (1). For example, y could be a measure of income and x_i could be a dummy variable taking the value of 0 for women and 1 for men.

The use of dummy variables is illustrated below. The intercept (β_0) is the sample average for observations in our reference group (in this example, women) when $x = 0$. The slope (β_1) is the difference in the averages of the two groups. In this example, when $x = 1$ (indicating men), the average salary is equal to $\beta_0 + \beta_1$. Therefore,

β_1 is the difference between the two groups, and we want to assess whether this difference is significantly different from zero.

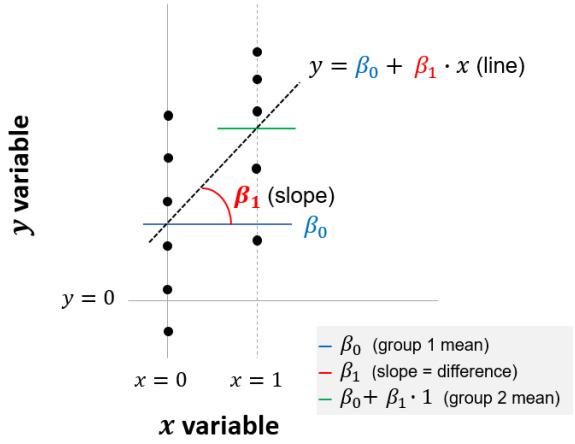

FIGURE 6.1 Linear model equivalent to independent-sample t-test

Comparison:

Let's say we want to compare the means of two groups in our example dataset, `y` and `y2`.

We start by creating a dummy variable to represent the two groups. The new variable is called `group_y2` and takes a value of 0 for observations in group `y` and a value of 1 for observations in group `y2`. This is done using the following code:

Here's a sample of six rows from our new data set, which now includes the dummy variable:

Now we've organized our data, we can compare the results of the built-in independent t-test with the linear model:

```
# independent t-test (built-in test)
t.test(y2, y, var.equal = TRUE)
# default is mu = 0, i.e. the null hypothesis that the difference in means

# Linear model
lm <- lm(value ~ 1 + group_y2, data = mydata_dummy)
```

TABLE 6.1 Some randomly selected rows from our data set

group	value	group_y2
y	1.1535602	0
y	1.1881680	0
y	-1.3557817	0
y2	0.1030347	1
y2	-0.2808907	1
y2	0.6238886	1

TABLE 6.2 Independent-sample t-test (equal variance) and linear model

Test	mean.y	mean.y2	difference	t	p.value	conf.low	conf.high
t-test	0.3	0.5	NA	0.5657	0.5729	-0.5016	0.9016
lm (with dummy)	0.3	NA	0.2	0.5657	0.5729	-0.5016	0.9016

```
lm %>% summary() %>% print(digits = 8) # show summary output
confint(lm) # show confidence intervals
```

The outputs from the independent t-test and the equivalent linear model are shown below. The t statistic, p-value and confidence intervals are identical. While the t-test reports the averages of the two samples, the linear model reports the average of the reference group (0.3 for `y`) and the *difference* between the average of this reference group and the other group indicated by the dummy variable (a difference of +0.2).

6.2 Welch's t-test

If the two samples have unequal variance then Welch's t-test could be used.

In fact, there is an argument that we should use Welch's t-test by default, rather than the independent Student's t-test, because Welch's t-test performs better than the t-test whenever sample sizes and variances are unequal between groups, and gives the same result when sample sizes and variances are equal.

Welch's t-test:

The Welch t-test is identical to the independent-sample Student's t-test described above, except that it does not assume equal variance of the two samples. In R, we can use the built-in t.test. When entering the code, we just need to specify that the sample variances are **not equal** (though this is the default assumption anyway).

Equivalent linear model:

The equivalent linear model is based on generalized least squares (GLS). The GLS approach makes the assumption that there is a relationship between the variance of observations and one of the independent variables used in our regression (in this example, the gender dummy variable). In a GLS regression, less weight is given to the observations with a higher variance.

We don't know the true relationship between variance and our independent variable(s) in the underlying population, so this relationship is estimated from our sample. The estimate is based on the relationship between residuals (i.e. observed values minus predicted values, based on an OLS regression) and an independent variable. Once we have estimated this relationship, we then re-weight each observation in our sample by dividing each observation by the predicted variance. The final GLS regression is then run on these re-weighted observations. A more detailed explanation of the GLS approach (without using matrix algebra!) can be found here.

In this example, when applying the GLS model in R, we specify that there is a different variance for group `y` and group `y2`.

Comparison:

TABLE 6.3 Independent-sample t-test (unequal variance) and GLS model

Test	mean.y	mean.y2	difference	t	p.value	conf.low	conf.high
t-test	0.3	0.5	NA	0.5657	0.5730	-0.5023	0.9023
lm (GLS)	0.3	NA	0.2	0.5657	0.5729	-0.4930	0.8930

The Welch's t-test and the equivalent linear model are carried out as follows:

```
options(digits = 10)
# t-test (built-in test)
t.test(y2, y, mu = 0, var.equal = FALSE, alternative = "two.sided")
# Note the assumption that variances are false

# Linear model (GLS)
lm <- nlme::gls(value ~ 1 + group_y2,
                weights = nlme::varIdent(form = ~1|group),
                method = "ML",
                data = mydata_dummy)
  lm %>% summary() %>% print(digits = 8) # show summary output
  confint(lm) # show confidence intervals
```

Here are the results, which are almost identical:

Note that we get the same estimates for the means of y and y_2 (of 0.3 and 0.5, respectively) from the t-test both with and without the assumption of equal variance. The unequal variance (heteroscedasticity) does not affect our estimates of the population means, but rather our assessment of whether or differences are statistically significant, in the form of t statistics, p-values and confidence intervals.

6.3 Mann-Whitney U test

If our usual assumptions don't hold (e.g. normal distributions, or if we're working with ordinal data) we can use a non-parametric version of these tests instead. When comparing two independent samples, this would be the Mann-Whitney U test.

Mann-Whitney U test:

This tests the null hypothesis that it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample. In other words, if the null hypothesis was rejected, you would infer that values from one population are more likely to be higher or lower than values from another population.

The Man-Whitney U test is a Wilcoxon test but with two samples. It can be run using R's built-in wilcox.test, this time with the (default) assumption that the observations are not paired.

Equivalent linear model:

Another way to describe the Mann-Whitney U test is as a test of mean ranks. It first ranks all your values for the dependent variable (e.g. income) from high to low, with the smallest rank assigned the smallest value. It then computes the mean rank in each group (e.g. male or female), and then computes the probability than random shuffling of those values between two groups would end up with the mean ranks as far apart as, or further apart, than you observed (see for example StackExchange and Laerd). If the resulting p-value was sufficiently small, you would infer that the difference in mean ranks between the samples was probably not due to chance, and that the values from one population were higher than those from another.

The equivalent linear model, with a close approximation, is a regression using the rank of the dependent variable:

$$rank(y_i) = \beta_0 + \beta_1 \cdot x_i \qquad H_0 : \beta_0 = 0$$

TABLE 6.4 Mann-Whitney U and equivalent linear model

Test	p-value	rank diff
wilcox.test	0.7907	NA
lm (ranks)	0.7896	1.56

where again x_i is a dummy variable indicating the group to which an observation belongs (in this example, 0 for women and 1 for men). Note that this uses the rank of the dependent variable, and not the signed rank as was the case with matched pairs.

Comparison:

Here is how you would run the Man-Whitney U test and equivalent linear model in R:

```
# Wilcoxon / Mann-Whitney U test (built-in)
wilcox.test(y, y2, paired = FALSE)

# Equivalent linear model
lm <- lm(rank(value) ~ 1 + group_y2, data = mydata_dummy)
  lm %>% summary() %>% print(digits = 8) # show summary output
```

The results are shown below. Both tests have very similar p-values. On this basis, we would not reject the null hypothesis that a value randomly selected from group `y` was not more likely to be higher or lower than one sampled from `y2`.

Digression - Mann-Whitney U test and medians:

It's worth noting that, with additional assumptions, the Man-Whitney U test is also a test for different *medians*.

This requires the assumption that the two samples have an equal shape, and therefore an equal variance. Laerd illustrates this concept using the diagram below: on the left, the two samples (men and women) have the same shape, so the Mann-Whitney U test tests for differences in the median score for men and women. On the right, the samples have different shapes, so it is a more gen-

eral test for whether a randomly-selected woman's score is higher than a randomly-selected man's score (without telling us anything about medians).

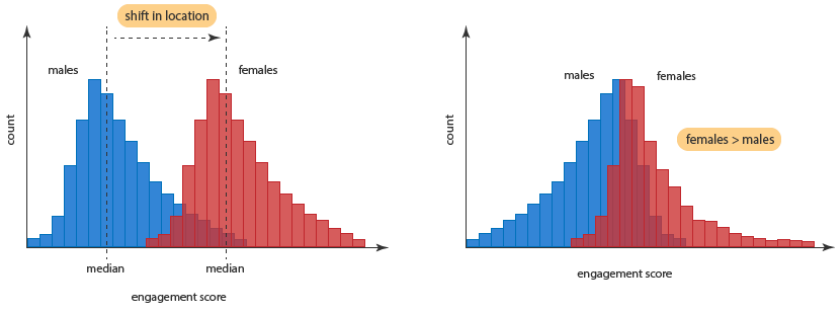

FIGURE 6.2 Man-Whitney U test and sample shapes

7

Three or more means

7.1 One-way ANOVA

The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent groups.

Examples include:

- Is there a difference in academic outcomes for pupils from ten different schools?
- Is there a difference daily coffee consumption between people in three different countries?

Intuitively, ANOVA is based on comparing the variance (or variation) between the groups, to variation within each particular group. If the 'between' variation is much larger than the 'within' variation, we are more likely to conclude that the means of the different groups are not equal. If the 'between' and 'within' variations are more similar in size, then we are less likely to conclude that there is a significant difference between sample means.

Why can't we just compare the means of every possible pair of groups, and see if any differences are statistically significant? The reason is that as the number of groups increases, the more likely we are to see differences that are due to chance alone. This means we are more likely to commit a Type I error, rejecting the null hypothesis (that there is no difference between the means) when the null hypothesis is in fact true.[1] An ANOVA controls for this additional

[1]For example, if there were only two groups, we would be carrying out one comparison: the mean of Group A vs the mean of Group B. At a 0.05 level

risk of Type I errors, maintaining the overall or experimentwise error rate, which is typically $\alpha = 0.05$.

It is important to note that the one-way ANOVA is an omnibus test statistic and cannot tell you *which* specific groups were statistically significantly different from each other, only that at least two groups were different. To determine which specific groups differed from each other, you would need to use a post hoc test.[2]

Dataset:

To illustrate, we will create a new dataset (`mydata_anova1`). We assume three groups (A, B and C) of normally-distributed variables, with means of 0, 1 and 0.5 respectively. We also create dummy variables for groups B and C (group A is our reference group, and so does not require an indicator):

```
# Create dataset 'mydata_anova1' which is three groups:
set.seed(40)                          # Makes the randomised figures reproducib
n <- 20                               # Sample size of 20 for each group
mydata_anova1 <- data.frame(
                    value = c(rnorm_fixed(n, mu = 0, sd = 1),      # Group A
                              rnorm_fixed(n, mu = 1, sd = 1),     # Group B
                              rnorm_fixed(n, mu = 0.5, sd = 1)), # Group C
                    group = rep(c("a", "b", "c"), each = N)
                    ) %>%
  # Explicitly add indicator/dummy variables
  mutate(group_b = if_else(group == "b", 1, 0)) %>%              # Group B du
  mutate(group_c = if_else(group == "c", 1, 0))                  # Group C du
```

of significance, there would be a 5% chance of a Type I error. If we had three groups, there would be three comparisons (Group A vs Group B, Group A vs Group C, and Group B vs Group C), and we would have a 14.3% (1-0.95^3) chance of a Type I error.

[2]Post hoc tests attempt to control the experimentwise error rate (usually $\alpha = 0.05$) in the same manner that the one-way ANOVA is used instead of multiple t-tests. Laerd suggests using the Tukey test (where there is homogeneity of variances in your samples) or the Games Howell test (where there is not).

TABLE 7.1 Some randomly selected rows from our dataset

value	group	group_b	group_c
0.5685977	a	0	0
0.5873537	a	0	0
2.2828164	b	1	0
0.5557109	b	1	0
1.2658245	c	0	1
0.1858697	c	0	1

Here's a sample of six rows from our new dataset, which includes the dummy variables:

Note that the data used in the remainder of this book varies from that used by Lindeløv in the original version, but the principles being discussed are exactly the same.

ANOVA function:

R has a package for ANOVA, in this case `car::Anova(aov())`. However, this is simply a 'wrapper' around the equivalent linear model, described below, and yields identical results.

Equivalent linear model:

The linear model assumes that the dependent variable can be predicted with a single mean for each group:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... \qquad H_0 : y = \beta_0$$

This assumption is illustrated below, in which there are assumed to be three groups (A, B and C). This extends the case with two groups, as was illustrated in the previous chapter, to cases with three or more groups. In this case, members of group B are identified by the dummy variable x_1, with the coefficient (or slope) of β_1, and members of group C are identified by the variable x_2, with the slope of β_2.

The null hypothesis of the linear model is that β_1 and β_2 are

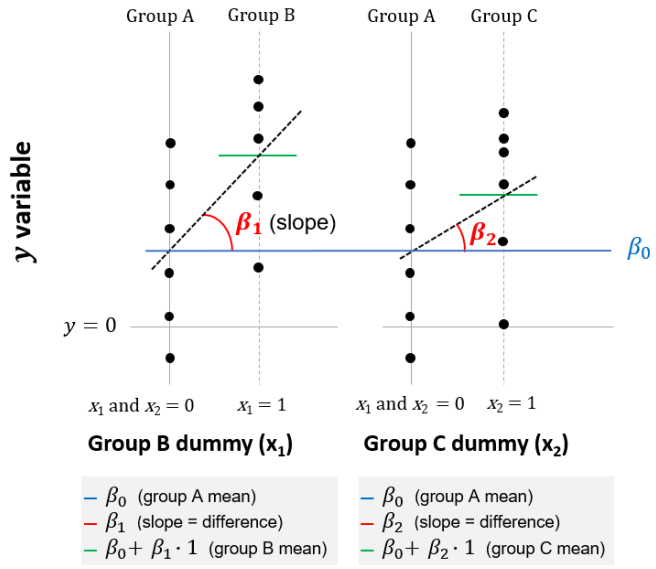

FIGURE 7.1 Linear model equivalent to ANOVA with three groups

both zero; or equivalently, that all groups have the same mean of β_0. To test this hypothesis, an F-test is used. The F statistic in a regression is the result of a test where the null hypothesis is that all of the regression coefficients are equal to zero. The F-test compares your full model to one with no predictor variables (the intercept only model), and decides whether your added variables improved the model. If you get a significant result, then whatever coefficients you included in your full model improved the model's fit (beyond what could be expected by chance alone).

In R, an F-test is carried out every time you run a linear regression, i.e. you do not have to specify it as an additional test.

Comparison:

The following code compares the ANOVA test in R with the identical linear model using dummy variables:

```
# Anova
car::Anova(aov(value ~ group, data = mydata_anova1))
```

TABLE 7.2 One-way ANOVA and equivalent linear model

Test	df	df.residual	F.statistic	p.value
Anova	2	57	5	0.00998
lm	2	57	5	0.00998

```
# Linear model
lm <- lm(value ~ 1 + group_b + group_c, data = mydata_anova1)
  lm %>% summary() %>% print(digits = 8) # show summary output
```

Here are the results, which are identical:

Here we would reject the null hypothesis that there was no differences between the means of any of our groups at the 0.05 level of significance (because p = 0.00998).

It should be emphasised that the results of the ANOVA and the linear model are identical *by construction*, as they are both an F-test that compares the full model (with group dummies) to a model with an intercept only.

7.1.1 Kruskal-Wallis

The non-parametric version of the ANOVA is the Kruskal-Wallis test. We would need to use this test if our dependent variable was ordinal rather than continuous. We would also use the non-parametric version if other assumptions of the one-way ANOVA did not hold, including (1) that the dependent variable was approximately normally distributed for each category of the independent variable, and (2) homogeneity of variances.

Equivalent linear model:

The Kruskal-Wallis is essentially a one-way ANOVA test on **ranks**. It can be expressed as the following linear model:

$rank(y) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... \qquad H_0 : y = \beta_0$

Comparison:

TABLE 7.3 Kruskal-Wallis test and equivalent linear model

Test	df	p.value
Kruskal	2	0.0203
lm	2	0.0177

Here is a comparison of the Kruskal-Wallis test and the equivalent linear model (the equivalent ANOVA test is also included for completeness, which we've seen is just a 'wrap' around the linear model).

```
# Kruskal-Wallis
kruskal.test(value ~ group, data = mydata_anova1)

# Linear model on ranks
lm <- lm(rank(value) ~ 1 + group_b + group_c, data = mydata_anova1)
  lm %>% summary() %>% print(digits = 8) # show summary output

# Anova on ranks (which is a wrapper around the linear model above)
car::Anova(aov(rank(value) ~ group, data = mydata_anova1))
```

The p-value of the two tests are similar, though not identical. In this example, we would reject the null hypothesis that all groups had equal means (at the 0.05 level of significance).

7.2 Two-way ANOVA

The two-way ANOVA compares the means of groups that have been split on two independent variables, or 'factors'.

For example: is there an interaction between gender and educational level on test anxiety among university students? Here gender (males / females) and education level (high school / undergraduate / postgraduate) are your independent variables or factors.

A two-way ANOVA tests three hypotheses:

- That the population means of the first factor (e.g. each gender) are equal;
- That the population means of the second factor (e.g. each education level) are equal; and
- That there is no interaction between the two factors - i.e. that the relationship between anxiety and gender does not depend on education level, or that the relationship between anxiety and education does not depend on gender.

The first two hypotheses relate to the relationship between each factor and the dependent variable, referred to as 'main effects'. Each of these is like a one-way ANOVA, but in the context of a larger model. The third hypothesis relates to the 'interaction effect'. Here we will focus on the interaction effect.

Updated dataset:

To show the modelling in R, we'll add another factor to our example dataset, `mood`, which reports whether a person is happy or sad. We also use this to create a dummy variable, `mood_happy`, which takes the value of 1 if the person is happy or 0 if they are sad.

```
mydata_anova2 <- mydata_anova1 %>%
  # 60 observations in total
  mutate(mood = rep(c("happy", "sad"), 30)) %>%
  # The dummy variable
  mutate(mood_happy = if_else(mood == "happy", 1, 0))
```

Here's a selection of six rows from the updated dataset:

Equivalent linear model:

The two-way ANOVA can test for the interaction between two factors (let's ignore the main effects for now). It is equivalent to the following linear model, which is now expressed using matrix notation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 X_3 \qquad H_0 : \beta_3 = 0$$

Here X_1 and X_2 represent the two factors in our model (in this

TABLE 7.4 Some randomly selected rows from our dataset

value	group	group_b	group_c	mood	mood_happy
0.5685977	a	0	0	happy	1
0.5873537	a	0	0	sad	0
2.2828164	b	1	0	happy	1
0.5557109	b	1	0	sad	0
1.2658245	c	0	1	happy	1
0.1858697	c	0	1	sad	0

example, the 'group' and 'mood' of each observation, respectively). Each β_i is a vector of values that relates to the levels within each factor. In our example, β_1 will have two values that correspond to the group dummy variables (`group_b` and `group_c`) and β_2 will be single value corresponding to the dummy variable for mood (`mood_happy`). The intercept β_0, to which all other βs are relative, is now the mean for the first level of all factors (people in group A who are sad).

β_3 is a vector that relates to the interactions of the factors. Here, it will be a vector comprising two values, corresponding to the combinations of the group and mood dummy variables (`group_b` and `mood_happy`, and `group_c` and `mood_happy`). The null hypothesis (of $\beta_3 = 0$) is that all values in this vector are zero, and that there is no interaction between group and mood when explaining the dependent variable, `value`.

To test this null hypothesis, we are going to carry out an F-test of two nested models:

- a full model, which includes both factors and the interaction terms; and
- a restricted or null model, which includes both factors but no interaction term.

This is similar to the F-test used in the one-way ANOVA above,

but in this case our null model includes the two factors and not just an intercept term.

Comparison:

The code for running the ANOVA, using the package in R, is as follows:

```
# Two-way ANOVA, built-in function
car::Anova(aov(value ~ mood + group + mood:group, data = mydata_anova2))
```

The equivalent linear model, which includes an interaction between the two group dummy variables and the mood dummy variable, is specified as follows:

```
lm(value ~ 1 + group_b + group_c + mood_happy +
           group_b:mood_happy + group_c:mood_happy,
         data = mydata_anova2)
```

The results of the above model will give us the p-values of *all* the interaction terms (in this case, two) and tell us if any of these are statistically significant. But recall that our null hypothesis is that *none* of the interaction terms are significant, and we can't rely on individual tests for this because of the increased risk of errors (as explained in the previous section on one-way ANOVAs). This is why we use the two-way ANOVA, which is an F-test that compares the full and null models:

```
# null model, without interactions
null <- lm(value ~ 1 + group_b + group_c + mood_happy, data = mydata_anova2)

# full model, with interactions
full <- lm(value ~ 1 + group_b + group_c + mood_happy + group_b:mood_happy +
             group_c:mood_happy, data = mydata_anova2)

# ANOVA using the two models above.
anova(null, full, test = "F")    # anova() uses an F test by default,
                                 # but here it's made explicit
```

The results of the two approaches are presented in the table below.

TABLE 7.5 Two-way ANOVA and equivalent linear model

Test	df	df.res	F.value	p.value
ANOVA	2	54	0.2977	0.7437
lm	2	54	0.2977	0.7437

This shows that the two approaches are identical F-tests with the same resulting p-value.

On this basis of the test above, we would fail to reject the null hypothesis that there was no interaction between the two factors in our model (group and mood), at the 0.05 level of significance.

7.3 ANCOVA

This adds a *continuous* independent variable, or covariate, to the model (e.g. age), in addition to one or more categorical independent variables (e.g. gender or education level).

An analysis of covariance (ANCOVA) evaluates whether the mean of the dependent variable is equal across levels of a categorical independent variable, while statistically controlling for the effects of other continuous variables (e.g. age) that are not of primary interest, known as covariates.

Updated dataset:

Here will will add a covariate to our one-way ANOVA above. In addition to the `group` dummy variables, we update our data set to include each subject's `age`, which we assume is correlated with the dependent variable, `value`:

```
# create a new column with the continuous variable 'age'
mydata_anova3 <- mydata_anova1 %>%
  mutate(age = value + rnorm_fixed(nrow(.), sd = 3))
```

TABLE 7.6 Some randomly selected rows from our dataset

value	group	group_b	group_c	age
0.5685977	a	0	0	-3.6341563
0.5873537	a	0	0	3.1155833
2.2828164	b	1	0	2.4620050
0.5557109	b	1	0	-3.5019533
1.2658245	c	0	1	0.8547244
0.1858697	c	0	1	-1.0090546

Here's a selection of six rows from the updated dataset:

ANOVA function:

An ANCOVA can be carried out using the `Anova()` function and including the covariate (in this case `age`) as an independent variable.

```
car::Anova(aov(value ~ group + age, mydata_anova3))
```

Equivalent linear model:

The same results can be achieved by using F-tests to compare two sets of linear models: (i) the full model and the nested model which excludes `age`, and (ii) the full model and the nested model that excludes the `group` dummy variables. Again, the F-tests are carried out using the `anova()` function, which uses an F-test by default.

The full model can be formulated as follows:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_3 \cdot age$$

where the value of y varies by group, as represented here by the dummy variables x_1 and x_2, and also by age.

This can be illustrated below. The ANCOVA tests whether there is difference in the mean `y` for the three groups, after controlling for

age (the vertical shift shown by β_1 and β_2). It also tests whether the slope (β_3) is statistically significant.

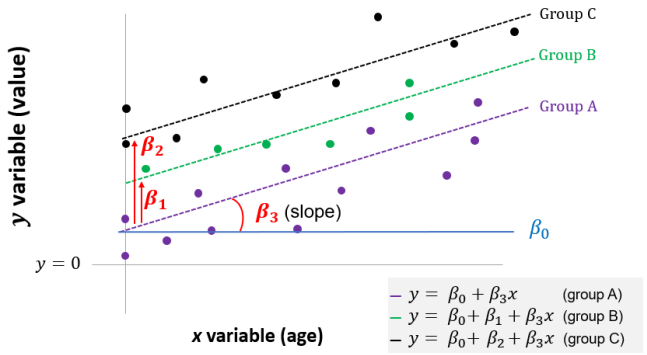

FIGURE 7.2 Linear model equivalent to ANCOVA with three groups

Here we run the linear model in R. The resulting p-values relate to the null hypotheses that `age` and `group` (respectively) have no effect on the dependent variable, in this case `value`.

```
# full model, with group and age variables
full <- lm(value ~ 1 + group_b + group_c + age, mydata_anova3)
# model without age
null_age <- lm(value ~ 1 + group_b + group_c, mydata_anova3)
# model without groups
null_group <- lm(value ~ 1 + age, mydata_anova3)
# result for age
anova(null_age, full)
# results of group
anova(null_group, full)
```

Comparison:

The results of the two approaches are presented in the table below:

Based on these results, we would reject the null hypothesis that there was no relationship between `value` and `group`, even after

TABLE 7.7 ANCOVA and linear model

Term	Model	df	res.df	F.value	p.value
age	Anova	1	NA	5.2002	0.02641
age	lm	1	56	5.2002	0.02641
group	Anova	2	NA	4.6929	0.01305
group	lm	2	56	4.6929	0.01305

controlling for differences in `age`. We would also reject the null hypothesis that `value` was not related to `age`.

8

Proportions and chi squared

This chapter looks at methods used for analyzing relationships in *categorical* data. The variable of interest is not a single continuous variable (e.g. how income or weight varies between groups) but the relative *count* or *proportion* of observations that fall into each category.

A key point is that the chi-squared test used in these cases is equivanet to a test of nested Poisson regression models.

8.1 Goodness of fit

A test of **goodness-of-fit** establishes whether an observed frequency distribution differs from a theoretical distribution. For example, we could test the hypothesis that a random sample with 44 men and 56 women has been drawn from a population in which men and women are equal in frequency, i.e. with the theoretical distribution of 50 men and 50 women.

Sample dataset:

Assume we have one category (mood) with three possible levels ('happy', 'sad' or 'meh'). We have a sample of 120 observations in total, as generated by the code below. We also add dummy variables to represent 'happy' and 'sad' for use in the linear model.

```
mydata_chi1 <- data.frame(mood = c("happy", "sad", "meh"),
                          counts = c(60, 90, 70)) %>%

  # Dummy variables to be used in linear model
```

TABLE 8.1 Categorical data with one variable

mood	counts	mood_happy	mood_sad
happy	60	1	0
sad	90	0	1
meh	70	0	0

```
  mutate(mood_happy = if_else(mood == "happy", 1, 0)) %>%
  mutate(mood_sad = if_else(mood == "sad", 1, 0))
```

In this example, we may want to test whether the proportions of people with each mood in the underlying population are equal, based on our observation of 120 people.

Chi-squared test:

The goodness-of-fit test is based on the following test statistic:

$$\chi^2 = \sum \frac{(O_i - Ei)^2}{E_i}$$

where O_i is the observed count in each category and E_i is the expected count in each category. The test statistic follows a chi-squared (χ^2) distribution where the degrees of freedom are equal to the number of categories minus one, i.e. $d.f. = rows - 1$ (in this example, df = 2).

A worked example of a goodness-of-fit test is provided in this video by Khan Academy.

R's built-in chi-squared test, `chisq.test`, compares the proportion of counts in each category with the expected proportions. By default, the expected proportions in each category are assumed to be equal.

```
#Built-in test
chisq.test(mydata_chi1$counts)
```

```
##
##  Chi-squared test for given probabilities
```

```
##
## data:  mydata_chi1$counts
## X-squared = 6.3636, df = 2, p-value = 0.04151
```

In this example, we would reject the null hypothesis that the proportion of people with each mood are all equal (p = 0.04151).

Equivalent linear model:

The equivalent linear model is a Poisson regression. This is a type of Generalized Linear Model (GLM). Poisson regressions use the natural log of the dependent variable, and so are sometimes referred to as a *log-linear models*. The follow model specification can be used:

$$log(y) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ...$$

where in this example x_1 and x_2 would be dummy variables representing 'happy' and 'sad' moods.

There are two reasons why we need to use a log-linear model rather than the OLS linear regression model, as explained in the book 'Broadening Your Statistical Horizons' by Julie Legler and Paul Roback:

1. Since a Poisson random variable is a count, its minimum value is zero and, in theory, the maximum is unbounded. A line from an OLS is model certain to yield negative values for certain values of the dependent variable, x; however, y can only take values from 0 to ∞.
2. The equal variance assumption in linear regression inference is violated, because as the expected value of a Poisson variable increases so too does the variance. With a Poisson distribution, the variance is equal to the expected value of the dependent variable, that is $E(Y) = VAR(Y)$.

These issues are addressed by taking the natural log of the dependent variable. It is assumed that the resulting logarithm of the dependent variable can be modeled by a linear combination of the independent variable(s) used in the model.

Because we are using the natural log of the dependent variable, the coefficients from our regression, β_1 and β_1, are interpreted as the *percentage* differences in the number of people whose moods are happy or sad, relative to the reference level of 'meh'.[1] Because of this, it can be used as a test for differences of proportions; that is, whether there is a significant difference in the percentage of people in each category.

To assess whether the differences between categories (moods) are statistically significant we compare two nested models, just as we did with the ANOVA and ANCOVA tests in the previous chapter. In this case we compare a 'full' Poisson regression model, which includes our dummy variables, and a 'null' model which does not. Instead of using the F-test to compare nested models (as was the case with ANOVA/ANCOVA), here we use a 'score test', specifically the Rao test. This gives us a test statistic that follows a χ^2 distribution, with the degrees of freedom being equal to the additional parameters in the full model compared to the null model.

The code is as follows:

```
full <- glm(counts ~ 1 + mood_happy + mood_sad,
            data = mydata_chi1,
            family = poisson())
null <- glm(counts ~ 1,  data = mydata_chi1, family = poisson())
anova(null, full, test = "Rao")
```

```
## Analysis of Deviance Table
##
## Model 1: counts ~ 1
## Model 2: counts ~ 1 + mood_happy + mood_sad
##   Resid. Df Resid. Dev Df Deviance    Rao Pr(>Chi)
```

[1] Technically, the percentage difference is equal to e^{β_1} and e^{β_2}, where β_1 and β_2 are continuously compounded rates of growth between 0 and 1. So for example, in the example below, the coefficient on 'sad' is $\beta_2 = 0.2513$. This mean the proportion of people in the 'sad' category are higher by $e^{0.2513} - 1 = 0.286 = 28.6\%$ than the proportion of people in the 'meh' category. This is equal to difference of the count of people in these two categories i.e. 90 'sad' people and 70 'meh' people.

TABLE 8.2 Goodness-of-fit: chi-squared test and equivalent linear model

model	Chi-squared	df	p.value
chisq.test	6.3636	2	0.04151
glm	6.3636	2	0.04151

```
## 1          2     6.2697
## 2          0     0.0000  2   6.2697 6.3636  0.04151 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparison:

As seen above, the p-value is identical under both the built-in chi-squared test and the equivalent linear model:

8.2 Contingency tables

Contingency tables are used in statistics to summarize the relationship between two categorical variables. They are used to test whether the frequency distribution differs between two or more groups. This can be tested using R"s built-in chi-squared test (`chisq.test`), similar to the goodness-of-fit test above, or again as two nested linear models (Poisson regressions), one of which includes dummy variables representing the interaction between the two categories in our table.

Sample dataset:

As an example, we might like to test whether a subjects mood (happy, sad or meh) is related to sex (male or female). We can create the following data set:

```
mydata_chi2 <- data.frame(
  sex = c("male", "female"),
```

TABLE 8.3 A contingency table

sex	happy	meh	sad
male	70	32	120
female	100	30	110

```
  happy = c(70, 100),
  meh = c(32, 30),
  sad = c(120, 110)
)
```

Chi-squared test:

As shown above, a contingency table is a table that lists the frequencies of occurrence for categories of **two** variables. The first variable is shown in rows, and the second variable is shown in columns.

Contingency tables can be used to assess whether the proportion of observations in one category depends on, or is *contingent* upon, the other category in the table. There are actually two types of tests:

- A test of **homogeneity**. This tests the null hypothesis that *different populations* have the same proportions of some characteristics. The key difference from the test of independence is that there are multiple populations that the data is drawn from. The null hypothesis is that the proportion of X is the same in all populations studied.
- A test of **independence**. A test of independence tests the null hypothesis that there is no association between the two variables in a contingency table where the data is all drawn from *one population*. The null hypothesis is that X and Y are independent.

Both these tests involve the exactly the same mathematical procedures and only differ only in terms of the hypothesis being tested. Some further reading can be found here.

These tests are based on a similar test statistic to the goodness-of-fit test:

$$\chi^2 = \sum \frac{(O_i - Ei)^2}{E_i}$$

where O_i is the observed count in each category and E_i is the expected count in each category. The test statistic follows a chi-squared (χ^2) distribution where the degrees of freedom are found by $d.f. = (rows - 1)(columns - 1)$.

A worked example of a chi-squared test is provided in this video by Khan Academy.

R's built-in chi squared test, `chisq.test`, can be used to assess whether the distribution of observations in one category is contingent on the distribution of observations under the other category. Note that first we must convert our data to a matrix and drop the first column (in this case `sex`):

```
# Convert data to matrix format, need for the built-in chi-squared
mydata_chi2_matrix <- mydata_chi2 %>%
  select(-sex) %>%
  as.matrix()
```

Now carry out the chi-squared test:

```
# Built-in chi-squared
chisq.test(mydata_chi2_matrix)
```

```
##
##  Pearson's Chi-squared test
##
## data:  mydata_chi2_matrix
## X-squared = 5.0999, df = 2, p-value = 0.07809
```

Based on this p-value we would not reject the null hypothesis that sex and mood were independent at the 0.05 level of significance.

Equivalent linear model:

The equivalent linear model is a Poisson regression with interaction

TABLE 8.4 Contingency table data in long format with dummy variables

sex	mood	Freq	mood_happy	mood_meh	sex_male
male	happy	70	1	0	1
female	happy	100	1	0	0
male	meh	32	0	1	1
female	meh	30	0	1	0
male	sad	120	0	0	1
female	sad	110	0	0	0

terms between the two sets of dummy variables (in this case one set of dummy variables is for mood, the other is for sex).

Here we are testing for the interaction between two the two categories, just as we did with the two-way ANOVA (though here our dependent variable is the natural log of the count of observations, rather than the value of a single continuous variable). The test is equivalent to the following linear model, which is expressed using matrix notation:

$$log(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 X_3 \qquad H_0 : \beta_3 = 0$$

Here X_1 and X_2 represent the two categories in our model (in this example, mood and sex). β_3 is a vector that relates to the interactions of the categories. Here, it will be a vector comprising two values, corresponding to the combinations of the sex and mood dummy variables (`male` and `mood_happy`, and `male` and `mood__meh`). The null hypothesis (of $\beta_3 = 0$) is that all values in this vector are zero, and that there is no interaction between sex and mood when explaining the distribution of observations in each category.

We begin by expressing our data in 'long' format, including dummy variables to identify the groups of people who are happy and 'meh'. This following data will be used in our linear model:

The test involves a comparison of two nested linear models: a full

model which includes the interaction terms between the two sets of dummy variables, and the null model which excludes the interaction terms. Again we use the (Rao) score test.

```
full <- glm(Freq ~ 1 + mood_happy + mood_meh + sex_male +
                   mood_happy*sex_male + mood_meh*sex_male,
            data = mydata_chi2_long, family = poisson())
null <- glm(Freq ~ 1 + mood_happy + mood_meh + sex_male,
            data = mydata_chi2_long, family = poisson())
anova(null, full, test = "Rao")
```

```
## Analysis of Deviance Table
##
## Model 1: Freq ~ 1 + mood_happy + mood_meh + sex_male
## Model 2: Freq ~ 1 + mood_happy + mood_meh + sex_male + mood_happy * sex_ma
##     mood_meh * sex_male
##   Resid. Df Resid. Dev Df Deviance    Rao Pr(>Chi)
## 1         2     5.1199
## 2         0     0.0000  2   5.1199 5.0999  0.07809 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the only difference between the nested models is the two interaction terms. Intuitively, we are testing whether the *interaction* of the two categories is statistically significant, i.e. whether the interaction between mood and sex can explain the distribution of observations over and above the individual effects of mood and sex alone.

Comparison:

As summarized in the table below, the linear model gives the same test statistic as the built-in chi-squared test, has the same degrees of freedom (which is equal to the number of interaction terms), and the same p-value.

TABLE 8.5 Contingency table: chi-squared test and linear model

model	Chi-squared	df	p.value
chisq.test	5.0999	2	0.07809
glm	5.0999	2	0.07809

9

Appendix - Types of variables

Variables (or measurements) fall into two categories: **discrete** or **continuous**.

9.1 Discrete variables

Discrete variables are also known as categorical variables. They are descriptions of categories into which observations can fall. Discrete variables can be further categorized as either *nominal* or *ordinal*.

Nominals variables have two or more categories which do not have an intrinsic order. In other words, there is no basis for ranking the categories. Examples of nominal variables include:

- Whether a person has a landline telephone could be categorized as "yes" or "no" (two categories)
- The US state in which a person lives (50 categories)

Ordinal variables have two or more catogories which do have an intrinsic order - that is, they *can* be ordered or ranked. Examples include:

- Levels of agreement, e.g. asking a survey respondent if they (i) strongly agree, (ii) agree, (iii) neither agree nor disagree, (iv) disagree or (v) strongly disagree with a question.
- Educational attainment could be recorded in a survey using four categories: (i) no high school degree, (ii) high school degree, (iii) college degree, or (iv) postgraduate degree. Here the categories can be ranked based on the level of educational attainment.

For ordinal variables, the interval or distance between the cate-

gories does not have a meaningful interpretation. For example, we cannot say that the distance between (i) no highschool degree and (ii) high school degree is the same as the distance bewteen (iii) college degree and (iv) postgraduate degree.

9.2 Continuous variables

Continuous variables are numbers rather than categories. Continuous variables can be further categorized as either *interval* or *ratio* variables.

Interval variables have a numeric value and can be measured along a continuum. The difference between values is interpretable. An example is temperature measure in Fahrenheit: the difference between 20F and 30F is the same as the difference bewtween 30F to 40F. However Fahrenheit is not a ratio variables. For example, 40 degrees is not "twice as hot" as 20 degrees.

Ratio variables are interval variables for which you can construct a meaningful fraction. Examples include height, weight, distance and income. For example, you could say that an income of $40,000 was twice as much as an income of $20,000. "Count" variables are also ratio variables; for example, the number of survey respondents who would vote for a presidential candidate. A condition of ratio variables is that 0 (zero) of the measurement indicates that there is none of that variable (e.g. $0 indicates zero income). This was not the case of temperature measured in Farenheit, as 0F does not mean there is "no temperature".

The four types of variables above form a hierarchy, where ratio variables are the highest:

Nominal < Ordinal < Interval < Ratio

At each level up the hierarchy, the current level includes all of the qualities of the one below it and adds something new.