Inference for paired data

Put your name here

Put the date here

Introduction

In this assignment we will learn how to run inference for two paired numerical variables.

Instructions

Presumably, you have already created a new project and downloaded this file into it. Please knit the document and work back and forth between this R Markdown file and the PDF output as you work through this module.

When you are finished with the assignment, knit to PDF one last time, proofread the PDF file **carefully**, export the PDF file to your computer, and then submit your assignment.

Sometimes you will be asked to add your own R code. That will appear in this document as a code chunk with a request for you to add your own code, like so:

```
## Add code here to [do some task]...
```

Be sure to remove the line `## Add code here to [do some task]...` when you have added your own code.

Sometimes you will be asked to type up your thoughts. That will appear in the document as follows:

Please write up your answer here.

Again, please be sure to remove the line "Please write up your answer here" when you have written up your answer. In these areas of the assignment, please use contextually meaningful full sentences/paragraphs (unless otherwise indicated) and proper spelling, grammar, punctuation, etc. This is not R code, but rather a free response section where you talk about your analysis and conclusions. If you need to use some R code as well, you can use inline R code inside the block between `\begin{answer}` and `\end{answer}`, or if you need an R code chunk, please go outside the `answer` block and start a new code chunk.

Load Packages

We load the standard `mosaic` package as well as the `MASS` package for the `immer` data and the `broom` package for tidy output.

```
library(MASS)
library(broom)
library(mosaic)
```

Paired data

Sometimes data sets have two numerical variables that are related to each other. For example, a diet study might include a pre-weight and a post-weight. The research question is not about either of these variables

directly, but rather the difference between the variables, for example how much weight was lost during the diet.

When this is the case, we run inference for paired data. The procedure involves calculating a new variable `d` that represents the difference of the two paired variables. The null hypothesis is almost always that there is no difference between the paired variables, and that translates into the statement that the average value of `d` is zero.

Research question

The `immer` data frame (from the `MASS` package) has data on five varieties of barley grown in six locations in each of 1931 and 1932. The two variables `Y1` and `Y2` measure the yield in 1931 and 1932, respectively. The question of interest here is whether there is a difference in the yield between those two years.

Inference for paired data

The key idea here is that we don't actually care about the yields themselves. All we care about is if there is a difference between the years. These are not two independent variables because each row represents a single combination of location and variety. Therefore, the two measurements are "paired" and should be treated as a single numerical variable of interest, representing the difference between `Y1` and `Y2`.

Since we're only interesting in analyzing the one numerical variable `d`, this process is nothing more than a one-sample t-test. Therefore, there is really nothing new in this assignment.

Let's go through the rubric.

Exploratory data analysis

Use data documentaton (help files, code books, Google, etc.), the str command, and other summary functions to understand the data.

[Type `library(MASS)` then `?immer` at the Console to read the help file.]

```
str(immer)
```

```
## 'data.frame':    30 obs. of  4 variables:
##  $ Loc: Factor w/ 6 levels "C","D","GR","M",..: 5 5 5 5 5 6 6 6 6 6 ...
##  $ Var: Factor w/ 5 levels "M","P","S","T",..: 1 3 5 4 2 1 3 5 4 2 ...
##  $ Y1 : num  81 105.4 119.7 109.7 98.3 ...
##  $ Y2 : num  80.7 82.3 80.4 87.2 84.2 ...
```

Prepare the data for analysis.

We create a new variable `d` that represents the difference between the yields `Y1` from 1931 and `Y2` from 1932. This uses the `mutate` command that adds an extra column to our data frame. Because we are subtracting `Y2 - Y1`, positive values of `d` mean the yield *increased* from 1931 to 1932 and negative values of `d` mean the yield *decreased* from 1931 to 1932.

```
immer_d <- mutate(immer, d = Y2 - Y1)
```

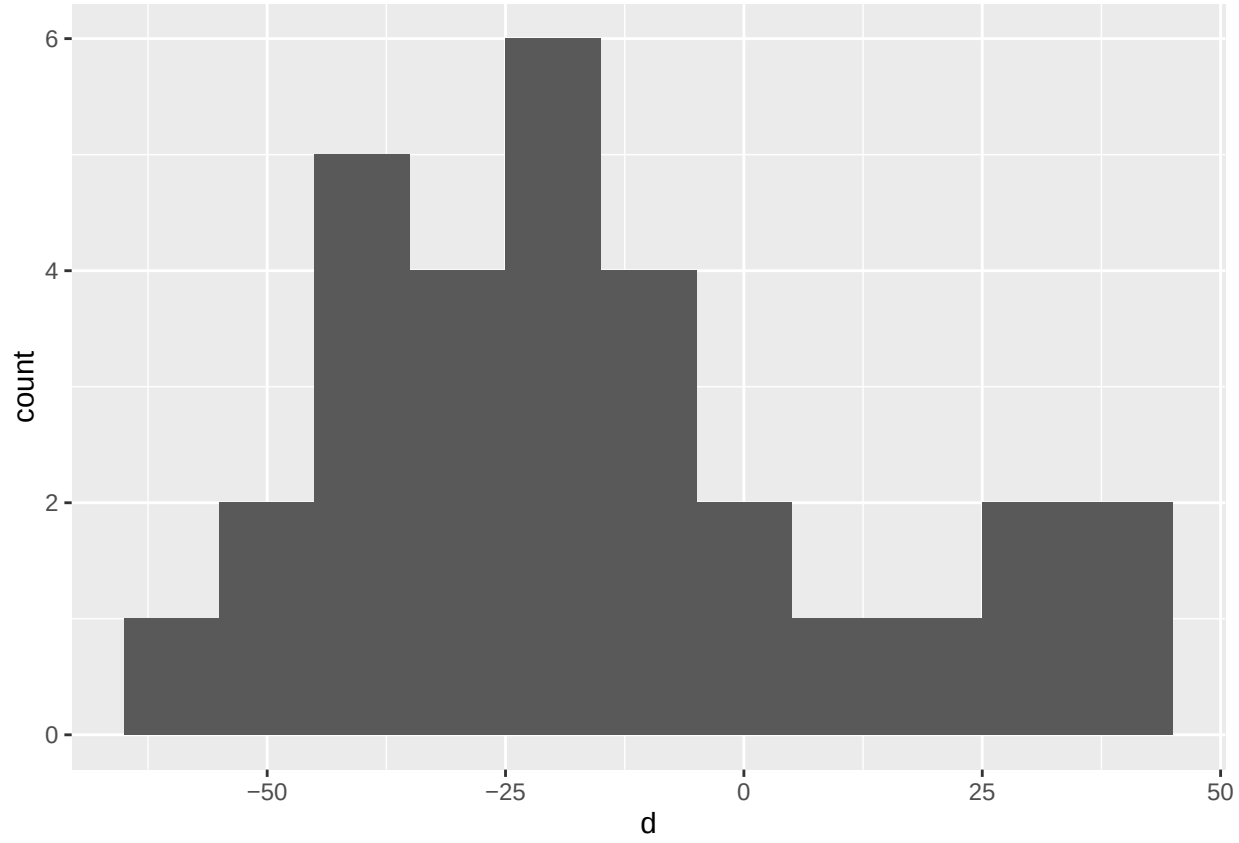
Make tables or plots to explore the data visually.

Here are summary statistics, a histogram, and a QQ plot for `d`.

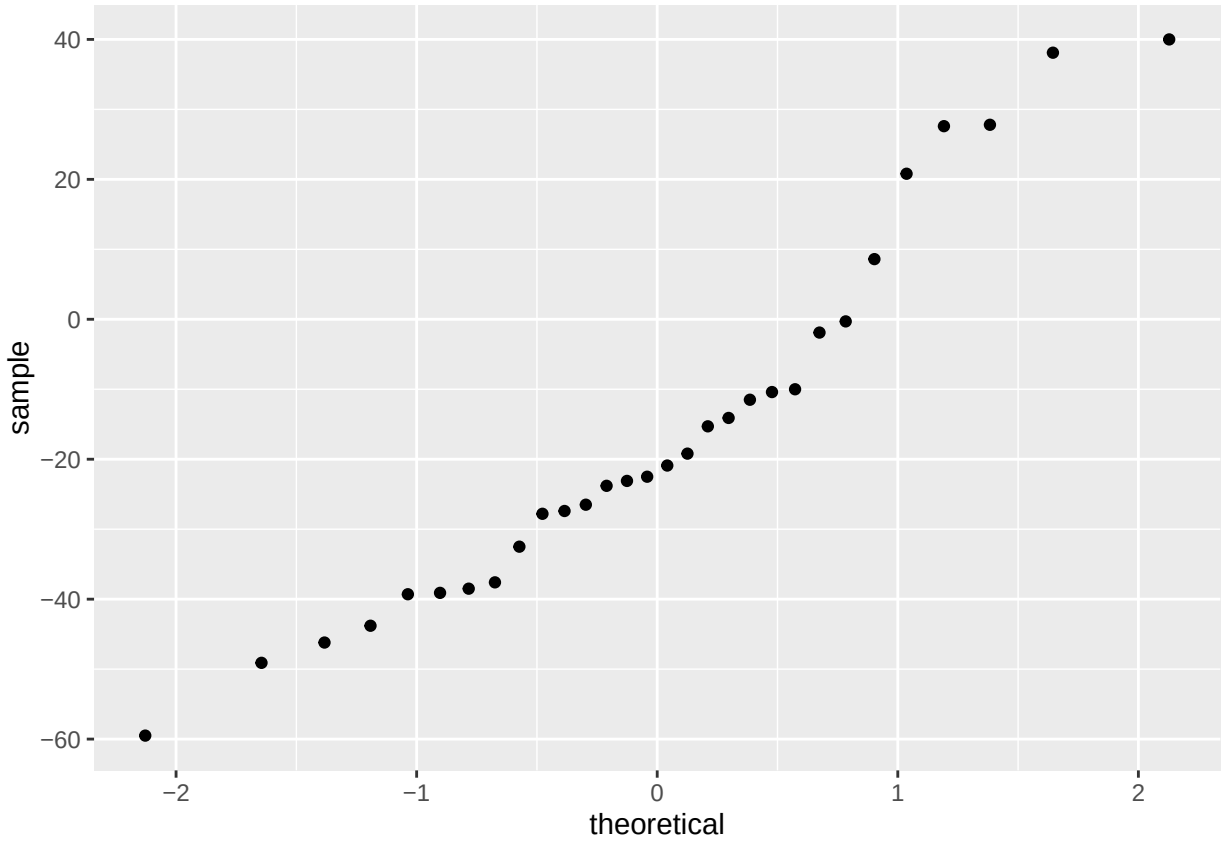
```
favstats(immer_d$d)
```

```
##    min      Q1 median     Q3 max      mean      sd  n missing
##  -59.5 -36.325  -21.7 -3.925  40 -15.91333 26.2218 30       0
```

```
ggplot(immer_d, aes(x = d)) +
    geom_histogram(binwidth = 10)
```




```
ggplot(immer_d, aes(sample = d)) +
    geom_qq()
```

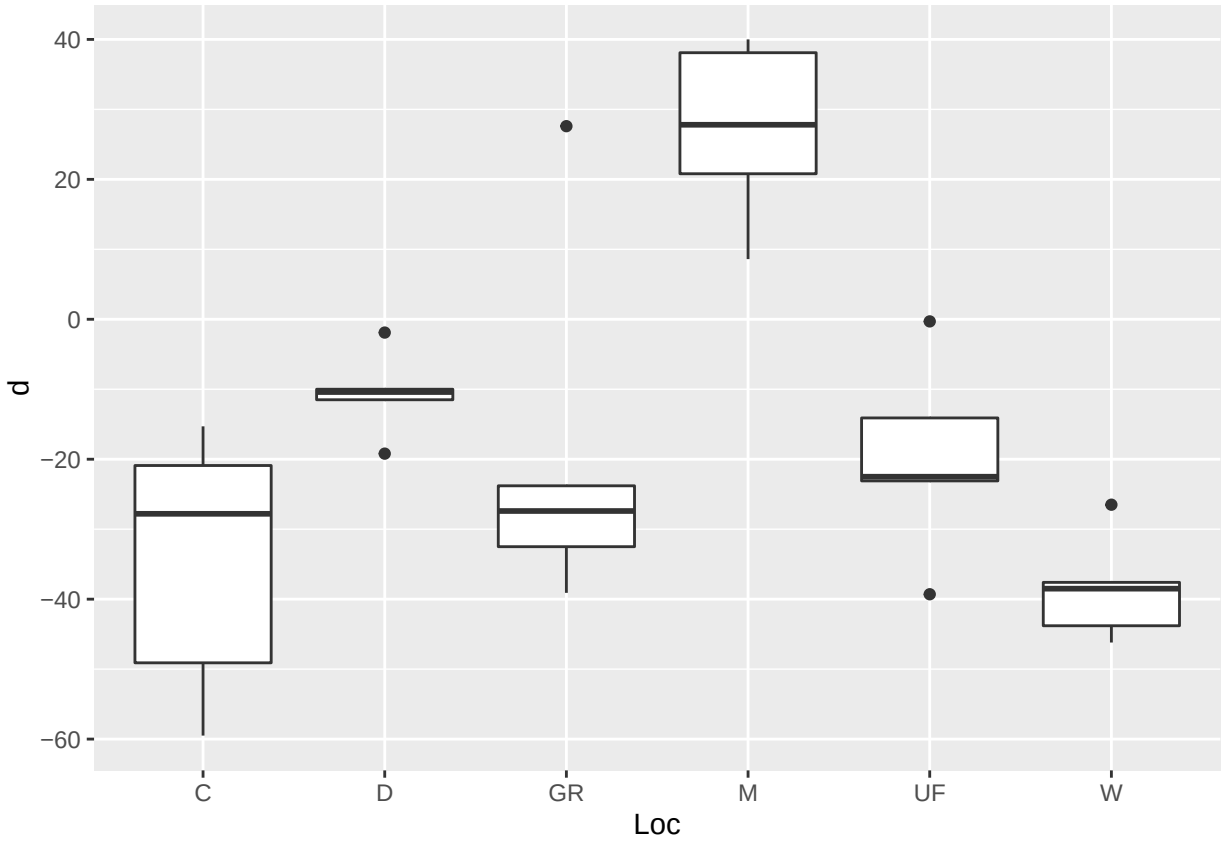


This distribution appears to be somewhat bimodal. This is a problem for inference because the mean of a bimodal distribution is not a good measure of center for the distribution. The typical method for dealing with bimodal data is to see if you can determine the reason for the two peaks. If there is a way of separating the data into the two groups, we should do that. Assuming there is enough data in both groups, they should be analyzed separately.

Type `View(immer_d)` at the Console and sort the `d` column in descending order by clicking on the arrow in the column header. (You may have to click twice to get descending order.)

It appears that location "M" is almost entirely responsible for all positive values of yield (with the exception of one plot). All other values are negative. This suggests that something different is happening at location "M" versus all other locations. We can also verify this by looking at a side-by-side boxplot, grouped by location.

```
ggplot(immer_d, aes(x = Loc, y = d)) +
    geom_boxplot()
```



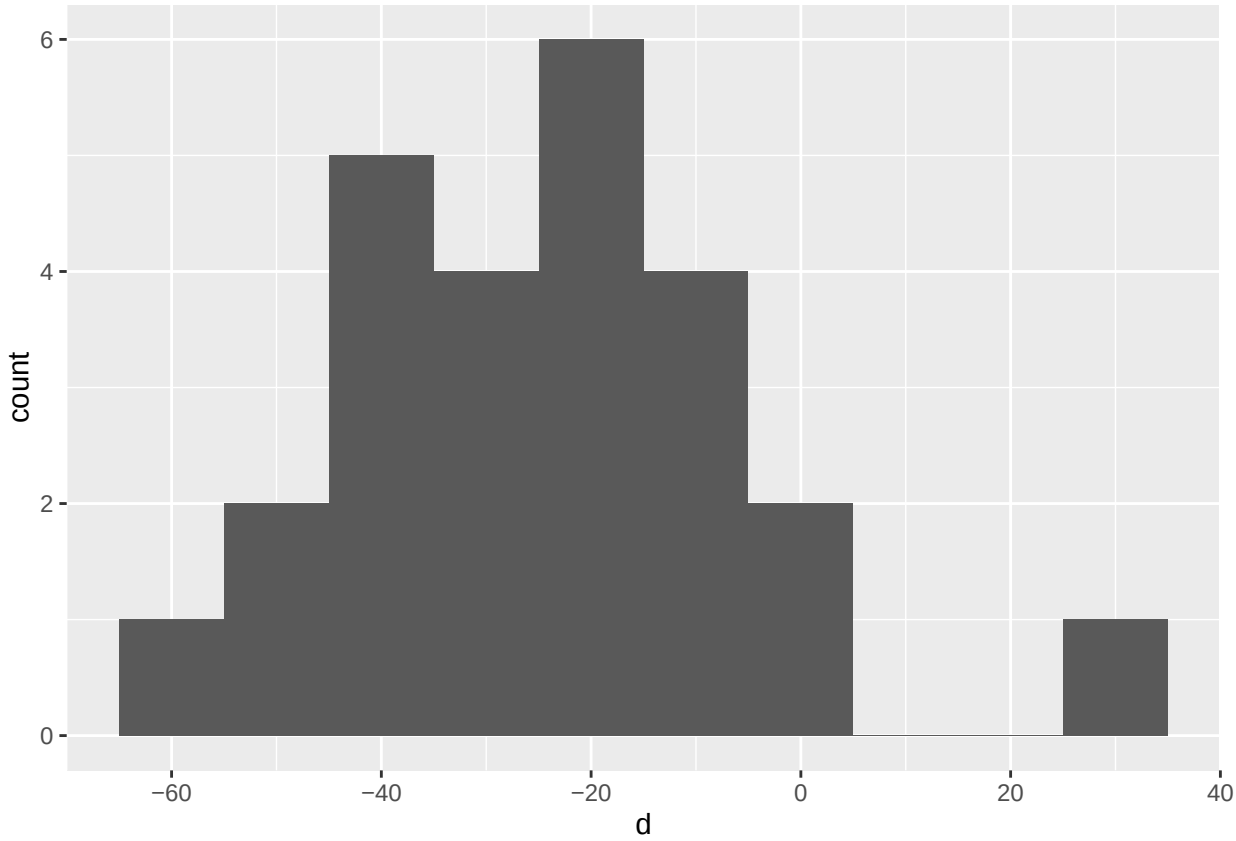
If we separate the data into two groups ("M" and all other locations), there is not enough data from location "M" to analyze this separately. Therefore, we will just remove group "M" from the analysis and note this in our conclusion.

We accomplish this in R by filtering the rows that are not at location "M". The symbol `!=` means "not equal to."

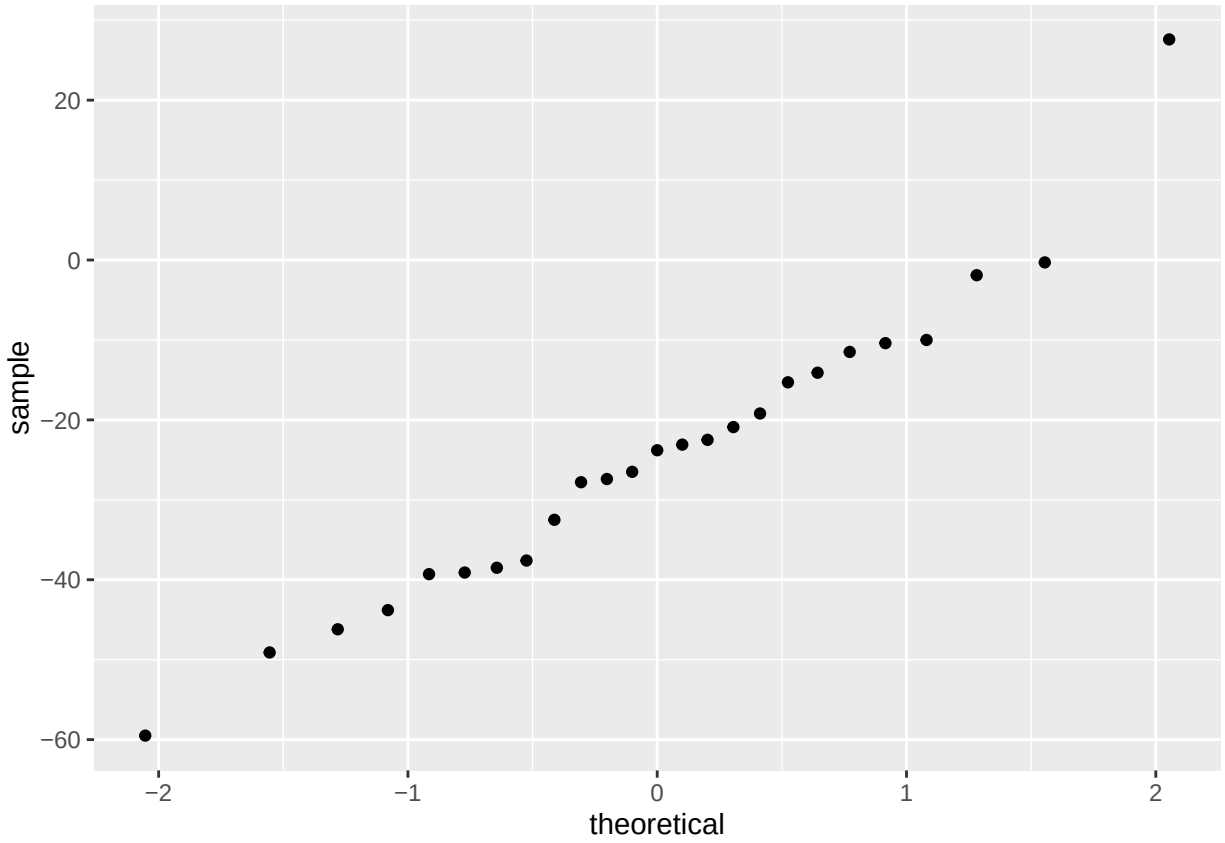
```
immer_d <- filter(immer_d, Loc != "M")
```

Here is the new histogram and QQ plot:

```
ggplot(immer_d, aes(x = d)) +
    geom_histogram(binwidth = 10)
```




```
ggplot(immer_d, aes(sample = d)) +
    geom_qq()
```



There is still a significant outlier. There is no real justification for removing it, so we need to run the analysis once with the outlier included and once with it excluded.

Hypotheses

Identify the sample (or samples) and a reasonable population (or populations) of interest.

The sample consists of 25 plots representing five varieties of barley grown in five different locations. (There were 30 plots across six locations, but remember that we removed the observations from one of the locations to get rid of the extra mode.) The population is all possible locations at which we might try growing these varieties of barley.

Express the null and alternative hypotheses as contextually meaningful full sentences.

H_0: There is no difference in mean barley yield from 1931 to 1932.

H_A: There is a difference in mean barley yield from 1931 to 1932.

Express the null and alternative hypotheses in symbols.

$H_0 : \mu_d = 0$

$H_A : \mu_d \neq 0$

Commentary: Since we're really just doing a one-sample t-test, we could just call this parameter μ, but the subscript d is a good reminder that it's the mean of the difference variable we care about (as opposed to the mean yield in either 1931 or 1932).

Model

Identify the sampling distribution model.

We use a t model with 24 degrees of freedom.

Check the relevant conditions to ensure that model assumptions are met.

- Random
  - As this is an experiment, the locations and varieties are not chosen at random. The idea here is that all five varieties are tested at five different locations with the hope that these measurements are representative of barley grown in a range of conditions. One concern about this is that we removed one of the locations to achieve unimodality, but this means that our samples are clearly not representative of all possible locations. We would need more information about location "M" to understand what was different about that location and, therefore, what our remaining data represents. (Another way of saying this is that our sample plots are hopefully representative of other plots that are not like location "M".)
- 10%
  - These 25 plots are way less than 10% of all possible locations in which barley could be grown.
- Nearly normal
  - We are now below the minimum sample size we use to consider this condition met automatically. However, the histogram and QQ plot show a reasonably normal shape, with the exception of the outlier.

Mechanics

Compute the test statistic.

```
barley_test <- tidy(t.test(immer_d$Y2, immer_d$Y1, paired = TRUE))
t <- barley_test$statistic
```

The value of t is -6.6335554.

Commentary: The `t.test` command offers an argument `paired = TRUE` that allows you to run a paired t-test from the original data. The order of the variables needs to be consistent with the order of subtraction used to produce `d` (in this case, `Y2 - Y1`).

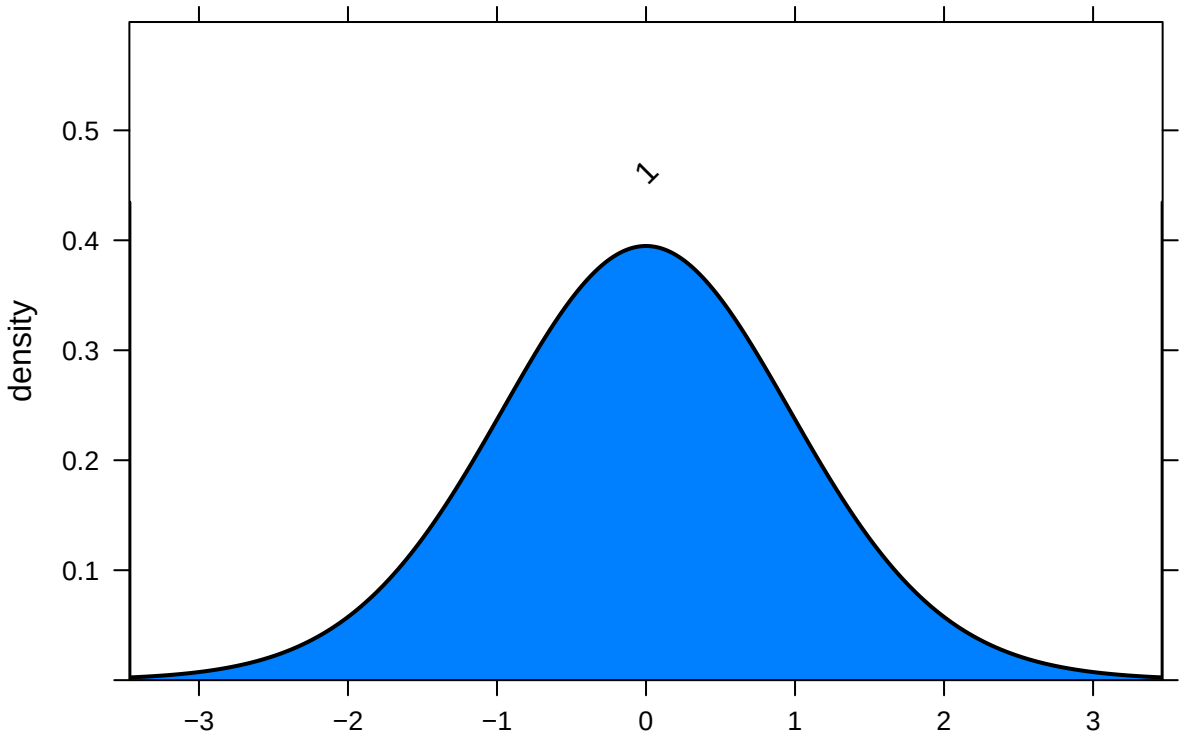
Note that we get the exact same answer if we run it as a one-sample t-test with the difference variable `d`:

```
barley_test_2 <- tidy(t.test(immer_d$d,  mu = 0))
barley_test_2$statistic
```

```
## [1] -6.633555
```

Plot the null distribution.

```
pdist("t", df = barley_test$parameter, q = c(-t, t))
```




```
## [1] 9.999996e-01 3.666165e-07
```

Commentary: The `parameter` stored in the output of `t.test` is the degrees of freedom. We set `q = c(-t, t)` as this is a two-sided test.

Calculate the P-value.

```
barley_test$p.value
```

```
## [1] 7.33233e-07
```

$P < 0.001$.

This is also the place to check the effect of excluding the outlier.

```
immer_d_no_outlier <- filter(immer_d,  d < 20)
immer_d_no_outlier_test <- tidy(t.test(immer_d_no_outlier$Y2,
                                       immer_d_no_outlier$Y1,
                                       paired = TRUE))
immer_d_no_outlier_test$statistic
```

```
## [1] -8.560539
```

```
immer_d_no_outlier_test$p.value
```

```
## [1] 1.316392e-08
```

The P-value is even smaller when excluding the outlier, so this won't affect the conclusion.

Conclusion

State the statistical conclusion.

We reject the null hypothesis.

State (but do not overstate) a contextually meaningful conclusion.

We have sufficient evidence that there is a difference in barley yield from 1931 to 1932. Keep in mind that we have excluded location "M" from consideration. If we had more locations like "M", that might have changed the conclusion.

Identify the possibility of either a Type I or Type II error and state what making such an error means in the context of the hypotheses.

If we made a Type I error, that would mean there was actually no difference in barley yield from 1931 to 1932, but that we got an unusual sample that detected a difference.

Confidence interval

Conditions

All necessary conditions have already been checked.

Calculation

```
barley_test$conf.low
```

```
## [1] -32.13318
```

```
barley_test$conf.high
```

```
## [1] -16.88282
```

Conclusion

We are 95% confident that the true change in barley yield from 1931 to 1932 is captured in the interval(-32.1331758, -16.8828242). This was obtained by subtracting 1931 yield from 1932 yield.

Commentary: it would normally be required that we report these numbers with units of measurement. Unfortunately, the help file does not tell us how the barley yield was measured. Also, don't forget that any time we find a number that represents a difference, we have to be clear in the conclusion about the direction of subtraction. Otherwise, we have no idea how to interpret positive and negative values. (Does this interval mean that yield was larger or smaller in 1932? Since we subtracted `Y2 - Y1` and these numbers are negative, that means yield *decreased* from 1931 to 1932.)

Your turn

A famous early pioneer of statistics, Francis Galton, collected data on the heights of adult children and their parents. We want to know if there is a difference between the heights of mothers and their daughters.

Since this data set includes data for both sons and daughters, let's filter it so we are only looking at daughters.

```
Galton_F <- filter(Galton, sex == "F")
```

Run inference to determine if there is a difference between the heights of mothers and their daughters.