Inference for Numerical Data

[Put your name here]

In this assignment we will learn how to run hypothesis tests for numerical data in R.

If we have one numerical variable, we will run a one sample t-test. If we have two paired numerical variables, we will run a paired t-test. If we have a categorical grouping variable with two categories along with a numerical response variable, we will run a two-sample t-test (Welch's t-test, to be specific).

If there are three or more categories in the explanatory variable, that will require Analysis of Variance (ANOVA), which will be covered in a separate assignment.

Instructions

Presumably, you have already created a new project and downloaded this file into it. Please knit the document (probably to HTML as you're working) and work back and forth between this R Markdown file and the knit output to answer the following questions.

When you are finished with the assignment, knit to PDF, export the PDF file to your computer, and then submit to the corresponding assignment in Canvas. Be sure to submit before the deadline!

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]...
```

When you see that in a code chunk, you need to type some R code to complete a task.

Sometimes you will be asked to type up your thoughts. Instructions to do that will be labeled as follows. If you are currently reading this in the knit output, please look back at the R Markdown file to see the following text:

(When you knit the document, you can't see the text from the line above. That's because the crazy notation surrounding that text marks it as a "comment", and therefore it doesn't appear in the output.) 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 (although it may include inline R code when necessary), but rather a free response section where you talk about your analysis and conclusions.

Getting started

Make sure you're in a project

If you're looking at this document, you should have already created a project and uploaded this R Markdown file to that project folder.

Save your file!

The first thing we **always** do is save our file. You'll probably want to save this under a new name. Go to the "File" menu and then "Save As". Once you've saved the file with the new name, from them on it's easier to just hit Ctrl-S (or Cmd-S on a Mac) to keep saving it periodically.

Remember that file names should not have any spaces in them. (In fact, you should avoid other kinds of special characters as well, like periods, commas, number signs, etc. Stick to letters and numerals and

you should be just fine.) If you want a multiword file name, I recommend using underscores like this: `this_filename_has_spaces_in_it`.

Load Packages

We load the standard `mosaic` package as well as the `OIdata` package to get the `teacher` data, the `openintro` package for the `hsb2` data, the `MASS` packages for the `immer` data and the `cabbages` data, and the `mosaicData` package for the `Galton` data.

```
library(mosaic)
library(OIdata)
data(teacher)
library(openintro)
library(MASS)
library(mosaicData)
```

As we'll be talking about teacher salary data, these numbers have just enough digits to trigger R's annoying habit of converting things to scientific notation. The following code will fix that problem. At the same time, we'll also tell R just to use two decimal places.

```
options(scipen = 10, digits = 2)
```

Inference for a single mean

The `teacher` data from the `OIdata` package contains information on 71 teachers employed by the St. Louis Public School in Michigan. Among the teachers in the data set, the mean total salary is $70288.64. According to Google, the average teacher salary in Michigan was $63,024 in 2010. So does this data suggest that the teachers in this region are paid differently than teachers in other parts of Michigan?

Let's walk through the rubric.

Hypotheses

Identify the sample and a reasonable population of interest.

The sample is 71 teachers employed by the St. Louis Public School in Michigan. We are using these 71 teachers as a hopefully representative sample of all teachers in that region of Michigan.

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

H_0 : Teachers in the St. Louis region earn $63,024 on average. (In other words, these teachers are the same as the teachers anywhere else in Michigan.)

H_A : Teachers in the St. Louis region do not earn $63,024 on average. (In other words, these teachers are *not* the same as the teachers anywhere else in Michigan.)

Express the null and alternative hypotheses in symbols.

$H_0 : \mu = 63024$

$H_A : \mu \neq 63024$

Model

Identify the correct sampling distribution model.

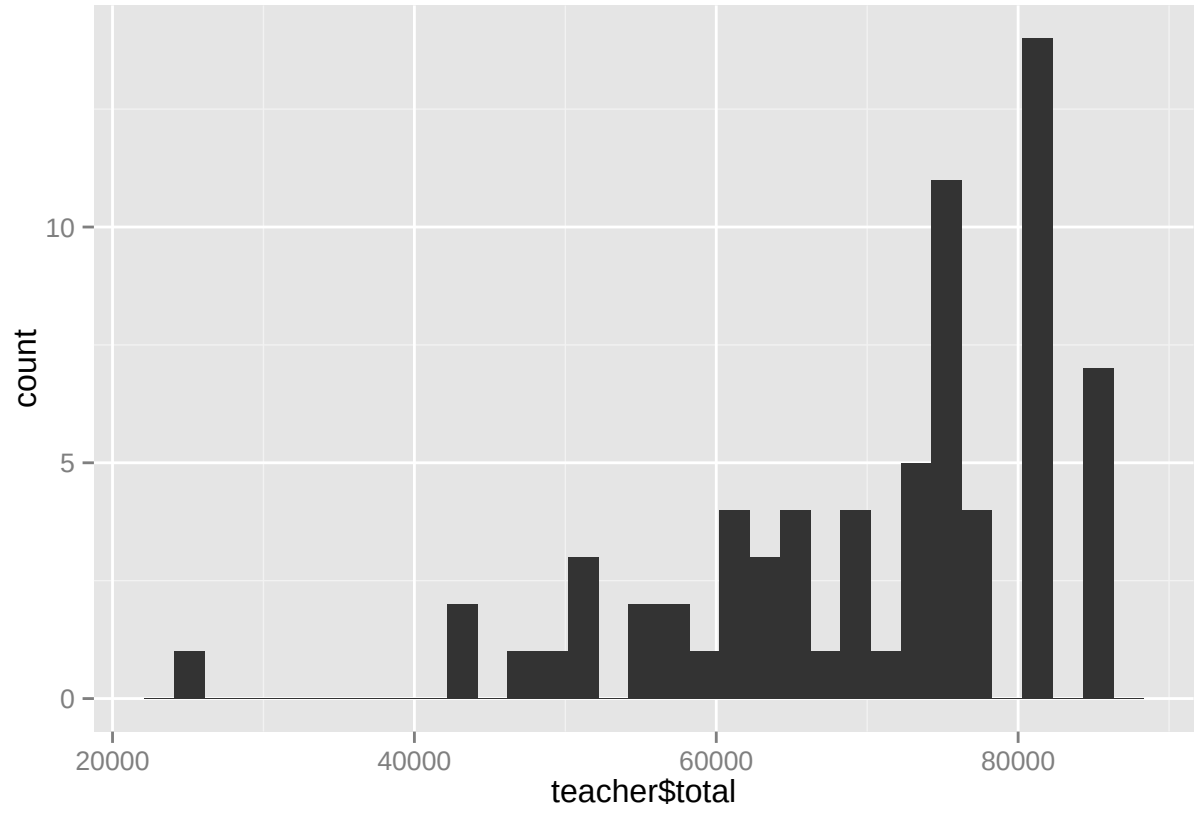
We will use a t model with 70 degrees of freedom.

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

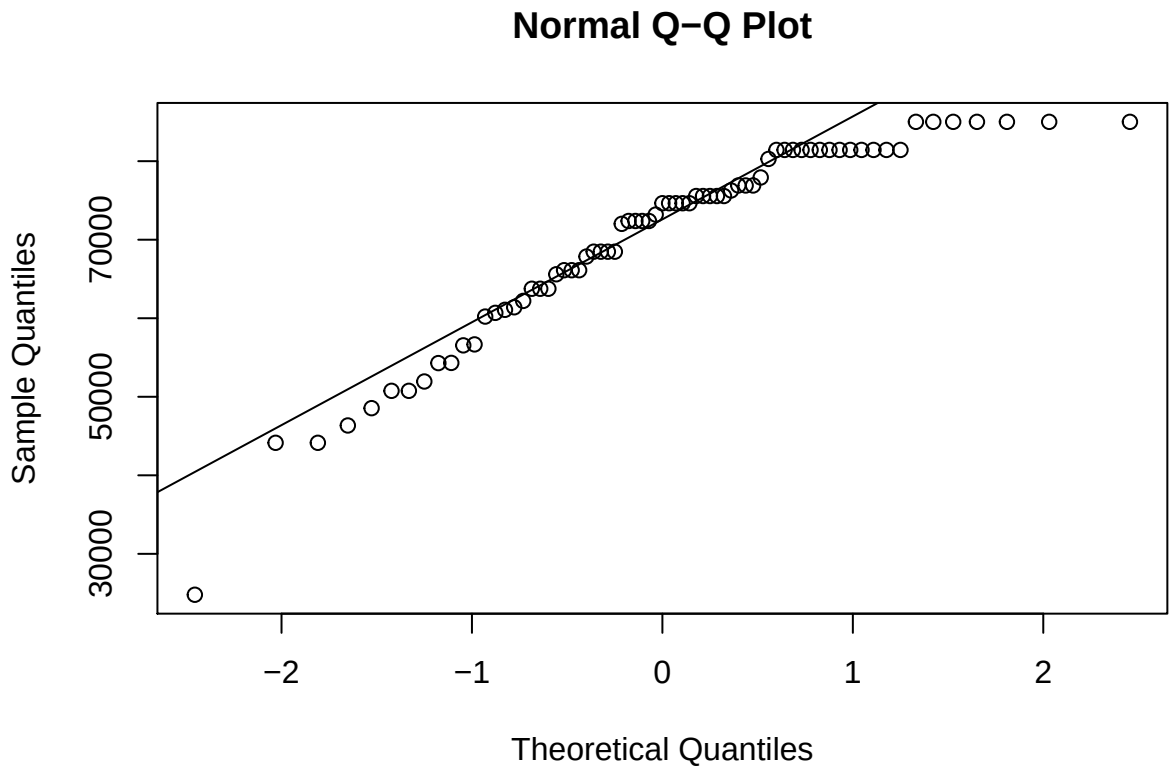
- Random
  - We know this isn't a random sample. We're not sure if this school is representative of other schools in the region, so we'll proceed with caution.
- 10%
  - This is also suspect, as it's not clear that there are 710 teachers in the region. One way to look at it is this: if there are 10 or more schools in the region, and all the school are about the size of the St. Louis Public School under consideration, then we should be okay.
- Nearly Normal
  - For this, we note that the sample size is much larger than 30, so we should be okay. Nevertheless, it's never a bad idea to check a histogram and QQ-plot of the data.

```
qplot(teacher$total)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



```
qqnorm(teacher$total)
qqline(teacher$total)
```



We see that these salaries are seriously skewed. Again, though, because our sample size is large enough, that means the sampling distribution model should be fine.

Mechanics

Compute the test statistic.

The `t.test` command gives us everything we need. Unlike in the tests we used for proportions for which the z-score was not part of the output, here the t-score is part of the output

```
test <- t.test(teacher$total,  mu = 63024)
test
```

```
##
##  One Sample t-test
##
## data:  teacher$total
## t = 5, df = 70, p-value = 0.000006
## alternative hypothesis: true mean is not equal to 63024
## 95 percent confidence interval:
##  67344 73233
```

```
## sample estimates:
## mean of x
##     70289
```

If our results are stored as `test`, then the test statistic is stored as `test$statistic`:

```
test$statistic
```

```
##   t
## 4.9
```

Mechanics

Compute the test statistic.

Plot the null distribution.

We use the `pdist` command we already know and love, but now we are using a t distribution and not a normal distribution.

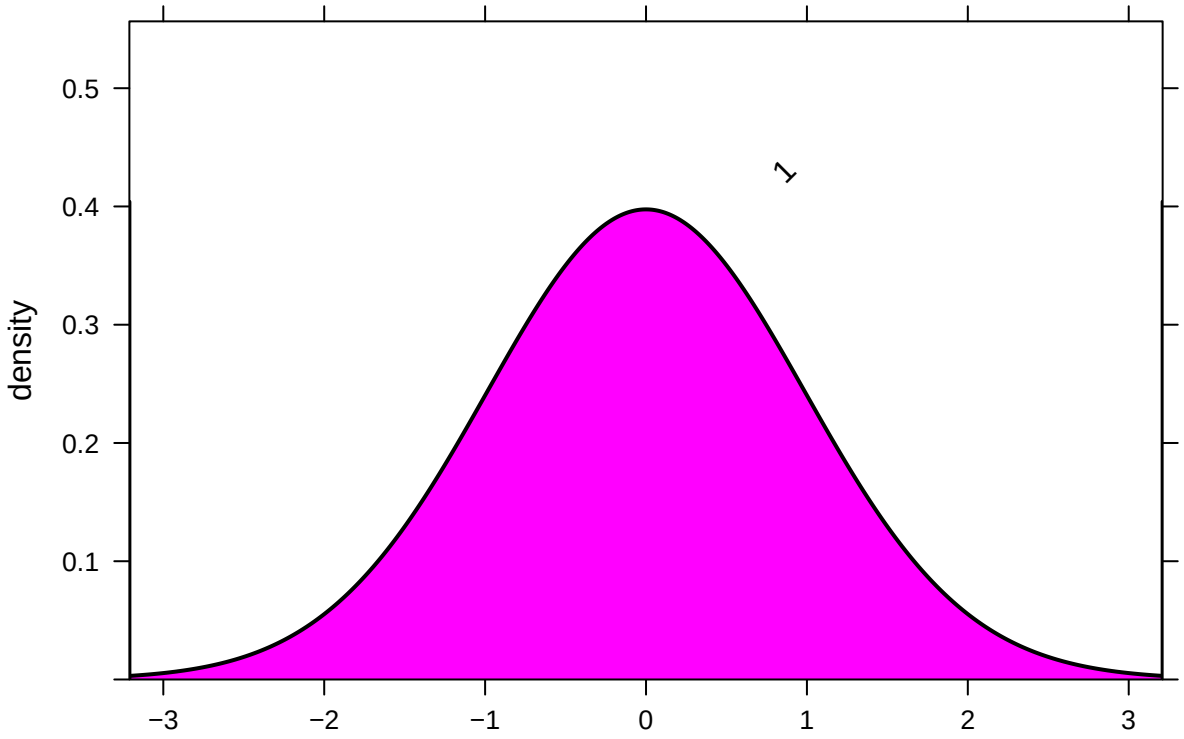
Because the t model has a degrees of freedom parameter, we also have to include this in the `pdist` command. We know there are 70 degrees of freedom; however, this number 70 also appears in the output of the `t.test` command. If we want our commands to be as general as possible (so we can reuse them in the future more easily), we should take advantage of this. The degrees of freedom are stored in `test$parameter`:

```
test$parameter
```

```
## df
## 70
```

So now let's put it all together:

```
pdist(dist = "t", df = test$parameter, q = test$statistic)
```




```
## t
## 1
```

Calculate the P-value.

The P-value is also part of the output.

```
test$p.value
```

```
## [1] 0.0000055
```

Conclusion

State the statistical conclusion.

We reject the null.

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

There is sufficient evidence to suggest that teachers in this region of Michigan do not make the same average salary as teachers in the rest of Michigan.

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.

By rejecting the null, we run the risk of making a Type I error. This could happen if, for example, teachers in this region generally do make $63,024 on average, but the teachers at this school are an unusual sample.

Confidence Interval

Conditions

All necessary conditions have been checked already.

Calculation

This is already stored in our previous output:

```
test$conf.int
```

```
## [1] 67344 73233
## attr(,"conf.level")
## [1] 0.95
```

Conclusion

We are 95% confident that the true mean salary for teachers in the St. Louis region of Michigan is captured in the interval ($67344.3, $73232.98).

Your turn!

In the High School and Beyond survey (the `hsb2` data set), among the many scores that are recorded are standardized math scores. Suppose that these scores are normalized so that a score of 50 represents some kind of international average. (This is not really true. I had to make something up here to give you a baseline number with which to work.) The question is, then, are American students different from this international baseline?

Follow the rubric, copying and pasting thoughtfully from the example above, to answer this question.

Inference for paired data

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.

Here we are looking at two numerical variables, but 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`.

Let's examine this difference. In fact, to make it convenient, we'll use a new command that will add `d` as a column to the existing data set.

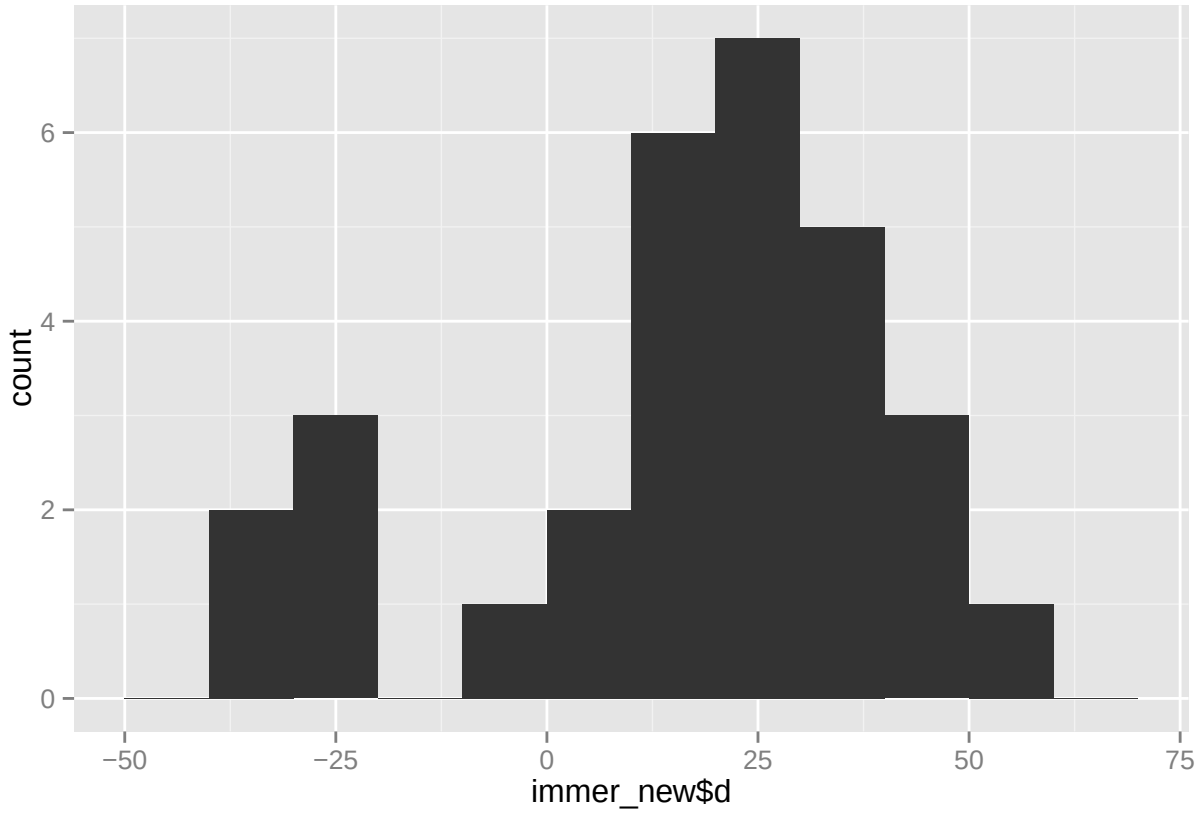
```
immer_new <- mutate(immer, d = immer$Y1 - immer$Y2)
favstats(immer_new$d)
```

```
##  min  Q1 median Q3 max mean sd  n missing
##  -40 3.9     22 36  60   16 26 30       0
```

Because of the order of subtraction, positive numbers mean more yield in 1931 than in 1932, whereas negative numbers represent the opposite.

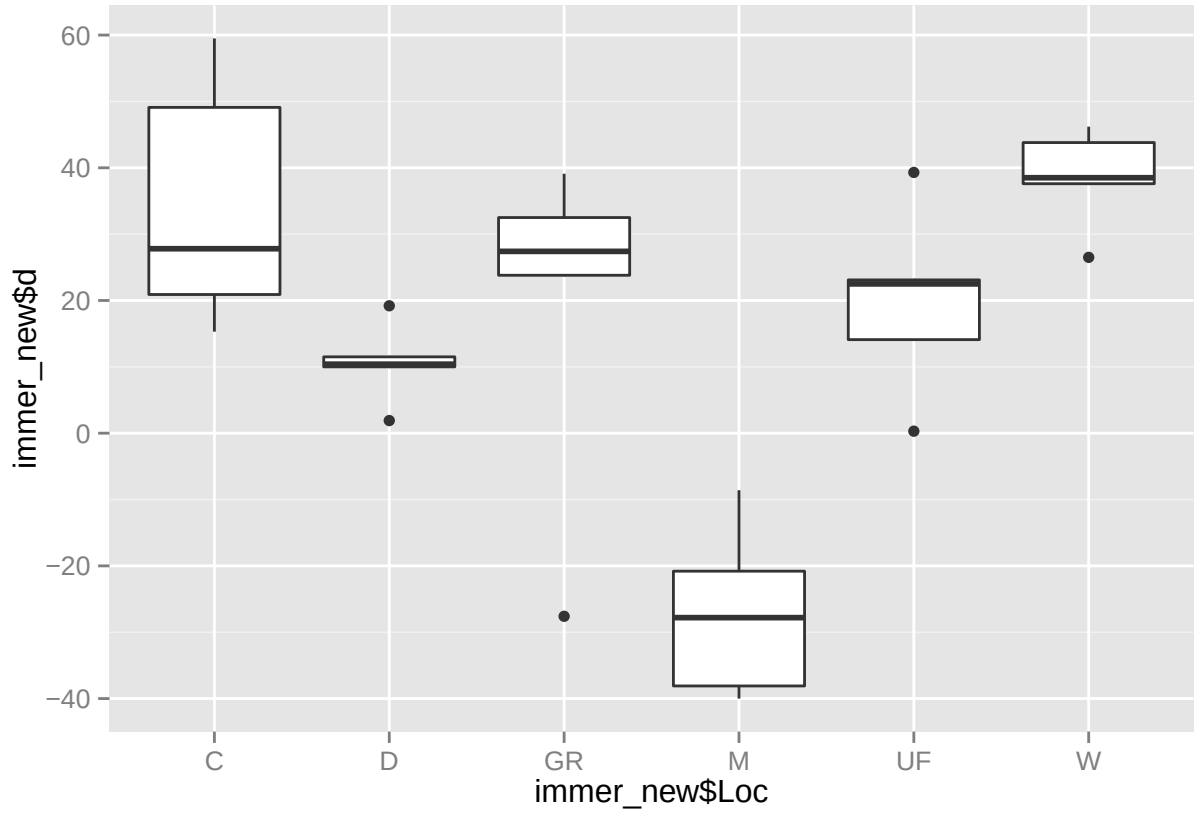
Here is the histogram.

```
qplot(immer_new$d, binwidth = 10)
```



This appears to be bimodal, which is a problem. Further investigation shows that one of the locations ("M") is mostly responsible for this mode.

```
qplot(x = immer_new$Loc, y = immer_new$d, geom = "boxplot")
```



We'll have to remove that location:

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

Now let's go through the rubric. Other than a slight change of notation and interpretation, this is really just a one-sample t-test applied to the new variable `d`.

Hypotheses

Identify the sample and a reasonable population 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.

$H_0 : \mu_d = 0$

$H_A : \mu_d \neq 0$

(The subscript d here is just a reminder that we are calculating the mean of the difference between the two years, rather than the mean yield itself in either of the years.)

Model

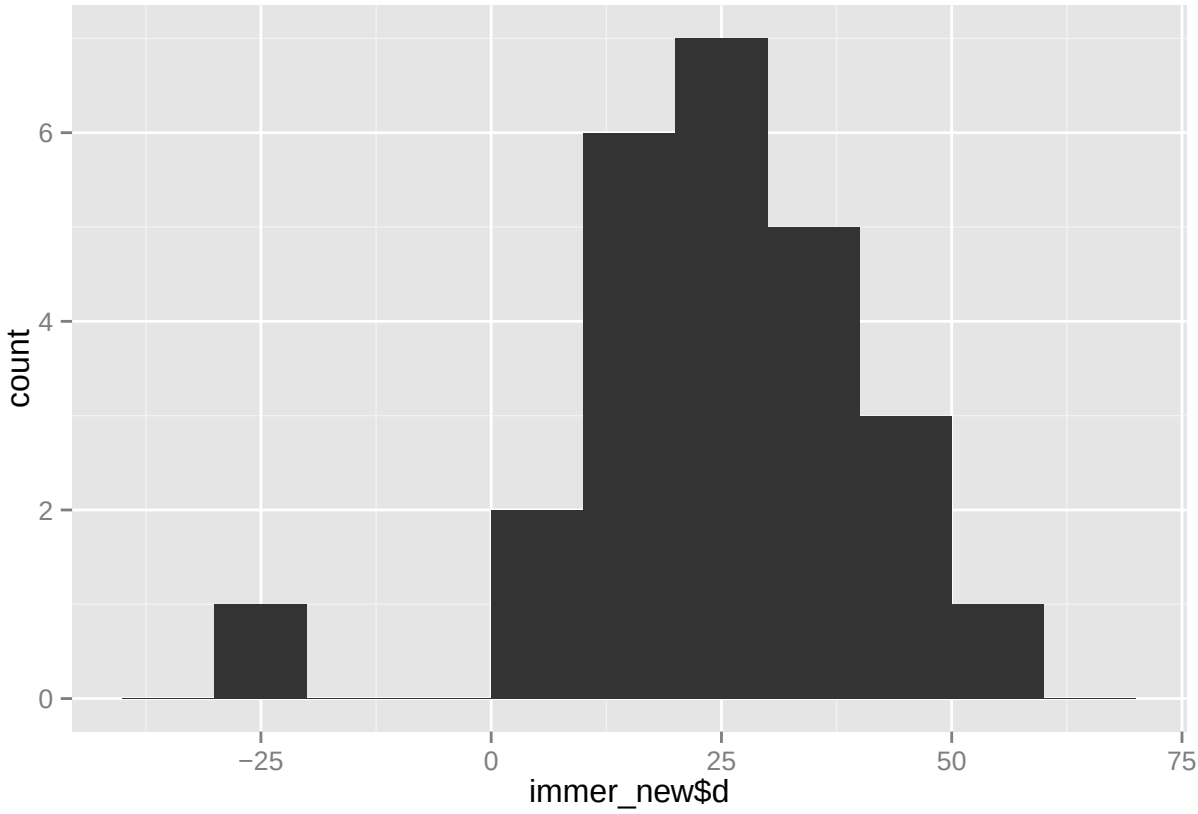
Identify the correct sampling distribution model.

We will use a t model with 24 degrees of freedom.

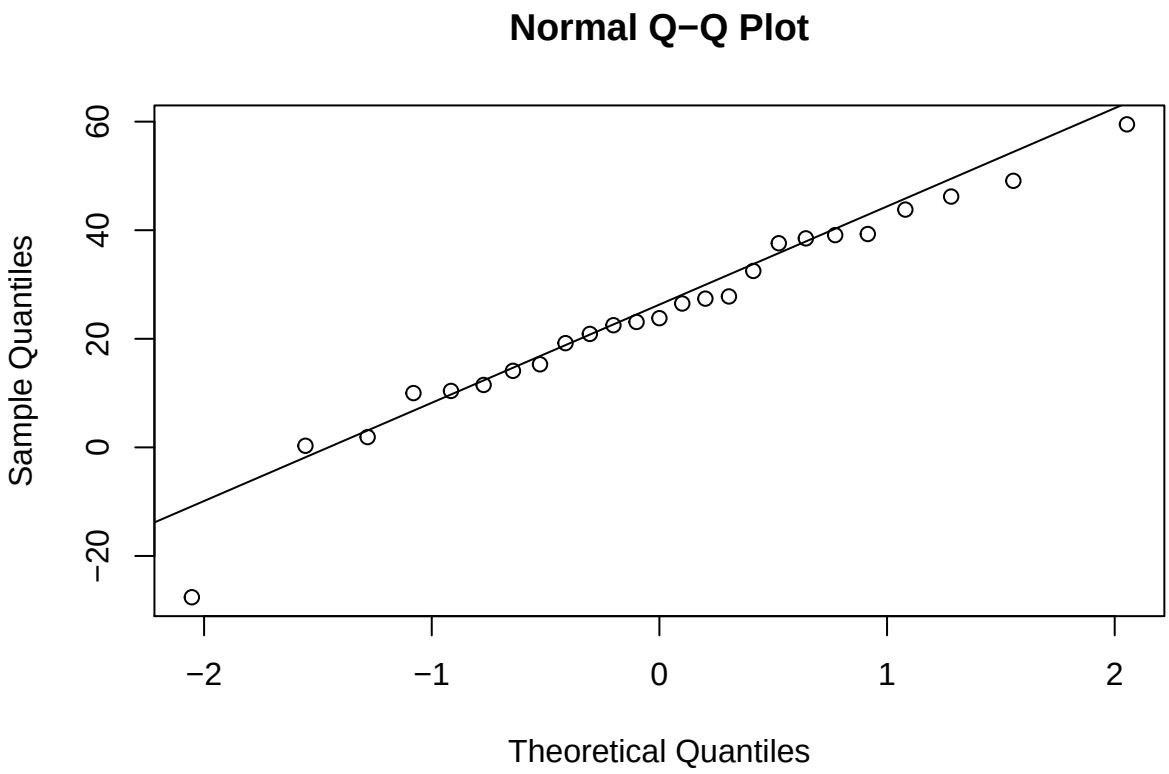
Check the relevant conditions to ensure that the model assumptions are met.

- 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. We need to look at a histogram and a QQ-plot.

```
qplot(immer_new$d, binwidth = 10)
```




```
qqnorm(immer_new$d)
qqline(immer_new$d)
```



There is one serious outlier. Ideally, we should do the analysis with and without this outlier to make sure are conclusions aren't dependent on its presence or absence.

Mechanics

Compute the test statistic.

R can run a paired t-test using the original variables (in others words, without having to use `d`).

```
test_paired <- t.test(immer_new$Y1, immer_new$Y2, paired = TRUE)
test_paired
```

```
##
##  Paired t-test
##
## data:  immer_new$Y1 and immer_new$Y2
## t = 7, df = 20, p-value = 0.0000007
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  17 32
## sample estimates:
## mean of the differences
##                      25
```

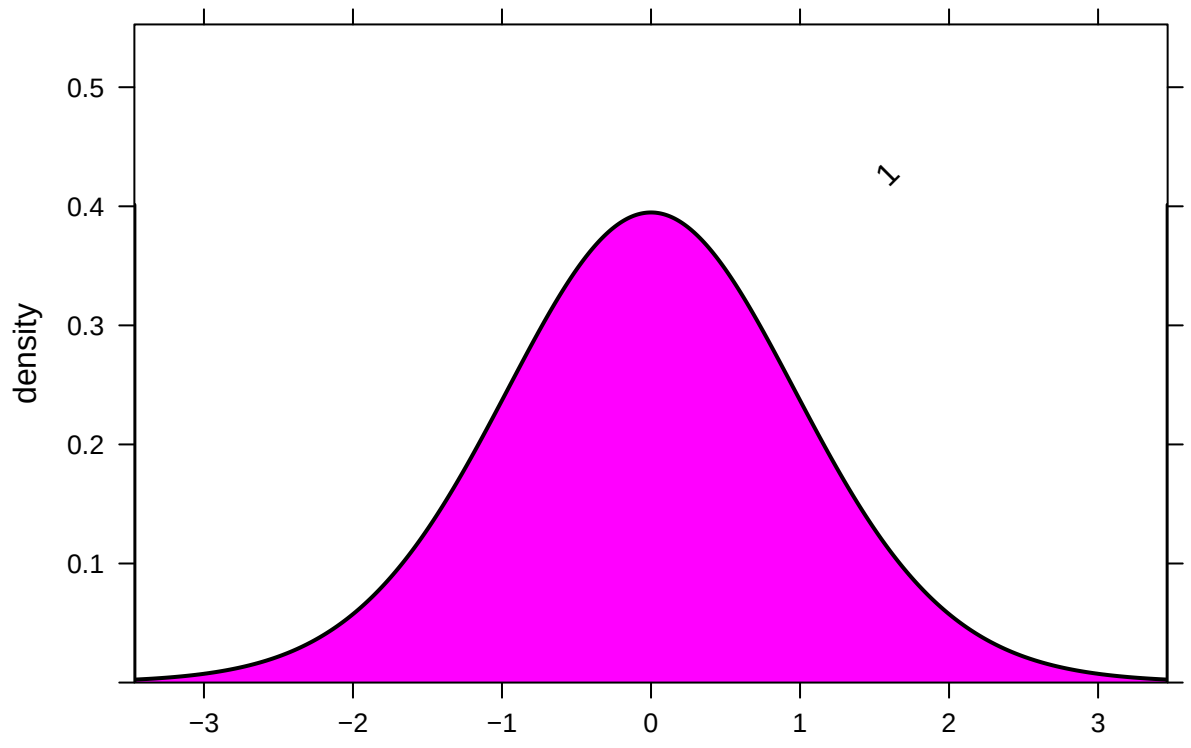
Note that you get the same result if you run a one-sample t-test on `immer_new$d`.

```r
t.test(immer_new$d, mu = 0)
```

```
##
##  One Sample t-test
##
## data:  immer_new$d
## t = 7, df = 20, p-value = 0.0000007
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  17 32
## sample estimates:
## mean of x
##        25
```

Plot the null distribution.

```r
pdist(dist = "t", df = test_paired$parameter, q = test_paired$statistic)
```




```
## t
## 1
```

Calculate the P-value.

```
test_paired$p.value
```

```
## [1] 0.00000073
```

Conclusion

State the statistical conclusion.

We reject the null.

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

We have sufficient evidence to suggest that there is a difference in mean barley yield from 1931 to 1932.

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.

By rejecting the null, we run the risk of making a Type I error. This could happen if the 1931 yields and 1932 yields were the same, but our sample showed a difference.

Confidence Interval

Conditions

All necessary conditions have been checked.

Calculation

This is already stored in our previous output:

```
test_paired$conf.int
```

```
## [1] 17 32
## attr(,"conf.level")
## [1] 0.95
```

Conclusion

We are 95% confident that the true mean difference in barley yields between 1931 and 1932 is captured in the interval (16.88, 32.13). (This difference was obtained by subtracting the 1932 yield from the 1931 yield.)

(Normally, our conclusion should come with units of measurement. Unfortunately, we do not have that information in this problem.)

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 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.

Inference for two independent means

In cases where we have one categorical explanatory variable with two groups and a numerical response variable, we can run a two-sample t-test. (The version we will run is often called Welch's t-test.)

Here we will consider cabbages. We have a variable `cabbages$Cult` containing two cultivars, called "c39" and "c52". We are interested to know if there is a difference in weight of the cabbage heads between these two varieties.

This is not paired data. We have not made two measurements on each cabbage head. Each cabbage plant is independent of all other plants, so we genuinely have two independent groups of cabbages.

Hypotheses

Identify the sample and a reasonable population of interest.

As with a two-sample test of proportions, a two-sample t-test will also deal with two samples drawn from their two respective populations.

The samples consist of 30 cabbages from the c39 group and 30 cabbages from the c52 group.

```
table(cabbages$Cult)
```

```
##
## c39 c52
##  30  30
```

The populations are all cabbages of variety c39 and all cabbages of variety c52.

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

H_0 : There is no difference in the head weight of c39 cabbages and c52 cabbages.

H_A : There is a difference in the head weight of c39 cabbages and c52 cabbages.

Express the null and alternative hypotheses in symbols.

$H_0 : \mu_{c39} - \mu_{c52} = 0$

$H_A : \mu_{c39} - \mu_{c52} \neq 0$

Model

Identify the correct sampling distribution model.

We will use a t model. The degrees of freedom are complicated. Because there are two samples, you can't calculate it as the sample size minus one anymore.

Fortunately, R will compute the degrees of freedom for us. Let's run the t-test a little early so we can read off the answer.

```
test2 <- t.test(cabbages$HeadWt ~ cabbages$Cult)
test2$parameter
```

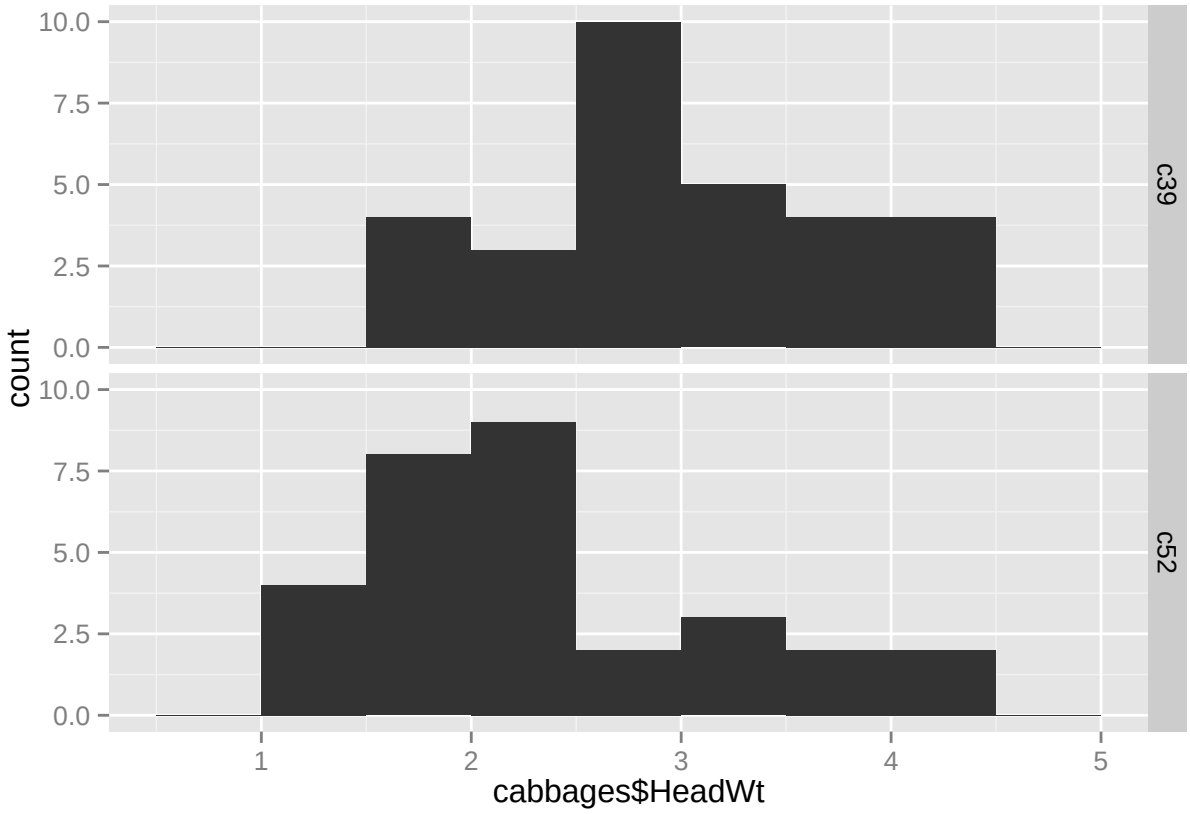
```
## df
## 58
```

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

- Random (for both groups)
  - We have no information at all about these cabbages. We hope that the 30 we have of each kind are representative of all cabbages from the two cultivars.
- 10% (for both groups)
  - 30 is less than 10% of all c39 cabbages and 30 is less than 10% of all c52 cabbages.
- Nearly normal
  - We have to check this for both groups. Since the sample sizes are 30 in each group, this meets the condition.

  As always, though, it's a good idea to look at our data graphically. Making QQ-plots for two groups is kind of a complicated hassle in R, so we'll just consider the histograms.

```
Cult <- factor(cabbages$Cult)
qplot(cabbages$HeadWt,  facets = Cult ~ ., binwidth = 0.5)
```



Both groups seem reasonably unimodal and symmetric with no outliers.

Mechanics

Compute the test statistic.

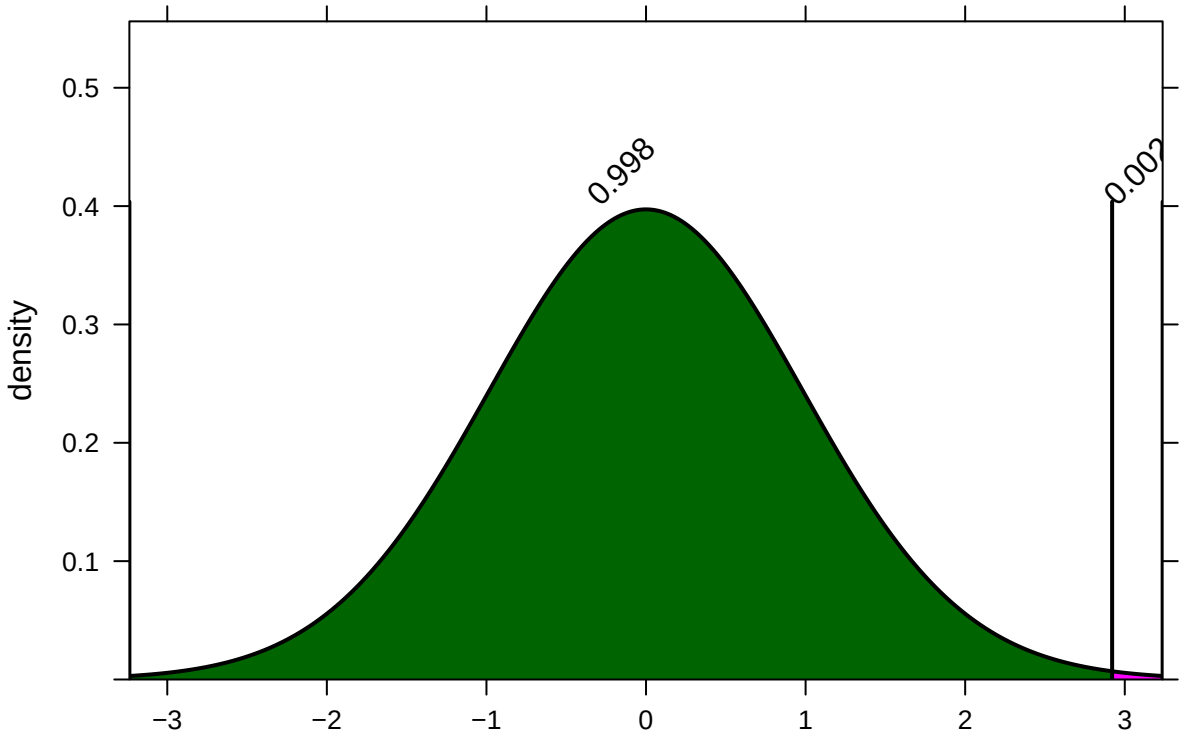
Since we ran the t-test above and stored the result in `test2`, we can just look at the results here.

```
test2
```

```
##
##  Welch Two Sample t-test
##
## data:  cabbages$HeadWt by cabbages$Cult
## t = 3, df = 60, p-value = 0.005
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2 1.1
## sample estimates:
## mean in group c39 mean in group c52
##               2.9               2.3
```

Plot the null distribution.

```
pdist(dist = "t", df = test2$parameter, q = test2$statistic)
```




```
## t
## 1
```

Calculate the P-value.

```
test2$p.value
```

```
## [1] 0.005
```

Conclusion

State the statistical conclusion.

We reject the null.

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

We have sufficient evidence to suggest that there is a difference in the mean head weight of c39 cabbages and c52 cabbages.

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.

By rejecting the null, we run the risk of making a Type I error. This could happen if the two cultivars were on average the same weight, but our samples showed a difference.

Confidence Interval

Conditions

All necessary conditions have been checked.

Calculation

This is already stored in our previous output:

```
test2$conf.int
```

```
## [1] 0.2 1.1
## attr(,"conf.level")
## [1] 0.95
```

Conclusion

We are 95% confident that the true mean difference in c39 and c52 cabbage head weights is captured in the interval (0.2 kg, 1.06 kg). (This difference was obtained by subtracting the c52 weights from the c39 weights.)

Your turn!

Continue to use the `cabbage` data set. This time, explore the ascorbic acid (vitamin C) content of each of the two cultivars.