Inference for two independent means

Put your name here

Put the date here

Introduction

If we have a numerical variable and a categorical variable with two categories, we can think of the numerical variable as response and the categorical variable as explanatory. The idea is that the two categories sort your numerical data into two groups which can be compared. Assuming the two groups are independent of each other, we can use them as samples of two larger populations. This leads to inference to decide if the difference between the two groups is statistically significant and then estimate the difference between the two populations represented. The relevant hypothesis test is called a two-sample t-test (or Welch's t-test, to be specific).

Instructions

Presumably, you have already created a new project and downloaded this file into it. From the `Run` menu above, select `Run All` to run all existing code chunks.

When prompted to complete an exercise or demonstrate skills, you will see the following lines in the document:

<div align="center">

—————————————————— ANSWER ——————————————————

——

</div>

These lines demarcate the region of the R Markdown document in which you are to show your work.

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
```

Be sure to remove the line `# Add code here` when you have added your own code. You should run each new code chunk you create by clicking on the dark green arrow in the upper-right corner of the code chunk.

Sometimes you will be asked to type up your thoughts. That will appear in the document with the words, "Please write up your answer here." 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. You may need to use inline R code in these sections.

When you are finished with the assignment, knit to PDF and proofread the PDF file **carefully**. Do not download the PDF file from the PDF viewer; rather, you should export the PDF file to your computer by selecting the check box next to the PDF file in the Files plane, clicking the `More` menu, and then clicking `Export`. Submit your assignment according to your professor's instructions.

Load Packages

We load the standard `mosaic` package as well as the `MASS` for the `cabbages` data. The `broom` package gives us tidy output.

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

We set the seed to make our simulations reproducible.

```
set.seed(1729)
```

Research question

We have data on two cultivars of cabbage called "c39" and "c52". Is there a difference in weight of the cabbage heads between these two varieties?

Every day I'm shuffling

Whenever there are two groups, the obvious null hypothesis is that there is no difference between them.

Consider the cultivar types c39 and c52. If there were truly no difference in weight between these cultivars, then it shouldn't matter if we know the cultivar or not. It becomes irrelevant under the assumption of the null.

We can simulate this assumption by shuffling the names of the cultivars. More concretely, we can randomly assign cultivar labels to each head of cabbage and then calculate the average weight in each cultivar groups. Since the cultivar labels are random, there's no reason to expect a difference between the two average weights other than random fluctuations due to sampling variability.

The actual mean weights of the c39 and c52 cultivars can be found using the `mean` command and the tilde notation. The following command should be read aloud as "calculate the mean head weight *by* cultivar type," or "*grouped by* cultivar type."

```
mean(HeadWt ~ Cult, data = cabbages)
```

```
##      c39      c52
## 2.906667 2.280000
```

The difference between the means is calculated with the `diffmean` command. We'll store this result as `obs_diff` for "observed difference".

```
obs_diff <- diffmean(HeadWt ~ Cult, data = cabbages)
obs_diff
```

```
##   diffmean
## -0.6266667
```

(Note that the order of subtraction here is c52 minus c39.)

This is the list of cultivars in the actual data:

```
cabbages$Cult
```

```
##  [1] c39 c39 c39 c39 c39 c39 c39 c39 c39 c39 c39 c39 c39 c39 c39 c39 c39
## [18] c39 c39 c39 c39 c39 c39 c39 c39 c39 c39 c39 c39 c52 c52 c52 c52
## [35] c52 c52 c52 c52 c52 c52 c52 c52 c52 c52 c52 c52 c52 c52 c52 c52
## [52] c52 c52 c52 c52 c52 c52 c52 c52 c52
## Levels: c39 c52
```

This is what happens when we `shuffle` them.

```
shuffle(cabbages$Cult)
```

```
##  [1] c52 c39 c52 c39 c52 c39 c39 c39 c52 c52 c39 c39 c39 c52 c39 c52 c52
## [18] c52 c52 c52 c52 c39 c39 c52 c39 c39 c52 c39 c52 c39 c39 c52 c52 c52
## [35] c39 c39 c52 c39 c52 c52 c52 c52 c39 c39 c39 c52 c52 c52 c39 c52 c39
## [52] c39 c52 c39 c39 c52 c39 c52 c39 c39
## Levels: c39 c52
```

Now we can calculate the group means and their difference for shuffled data. Let's do it few times.

```
diffmean(HeadWt ~ shuffle(Cult), data = cabbages)
```

```
##    diffmean
## 0.07333333
```

```
diffmean(HeadWt ~ shuffle(Cult), data = cabbages)
```
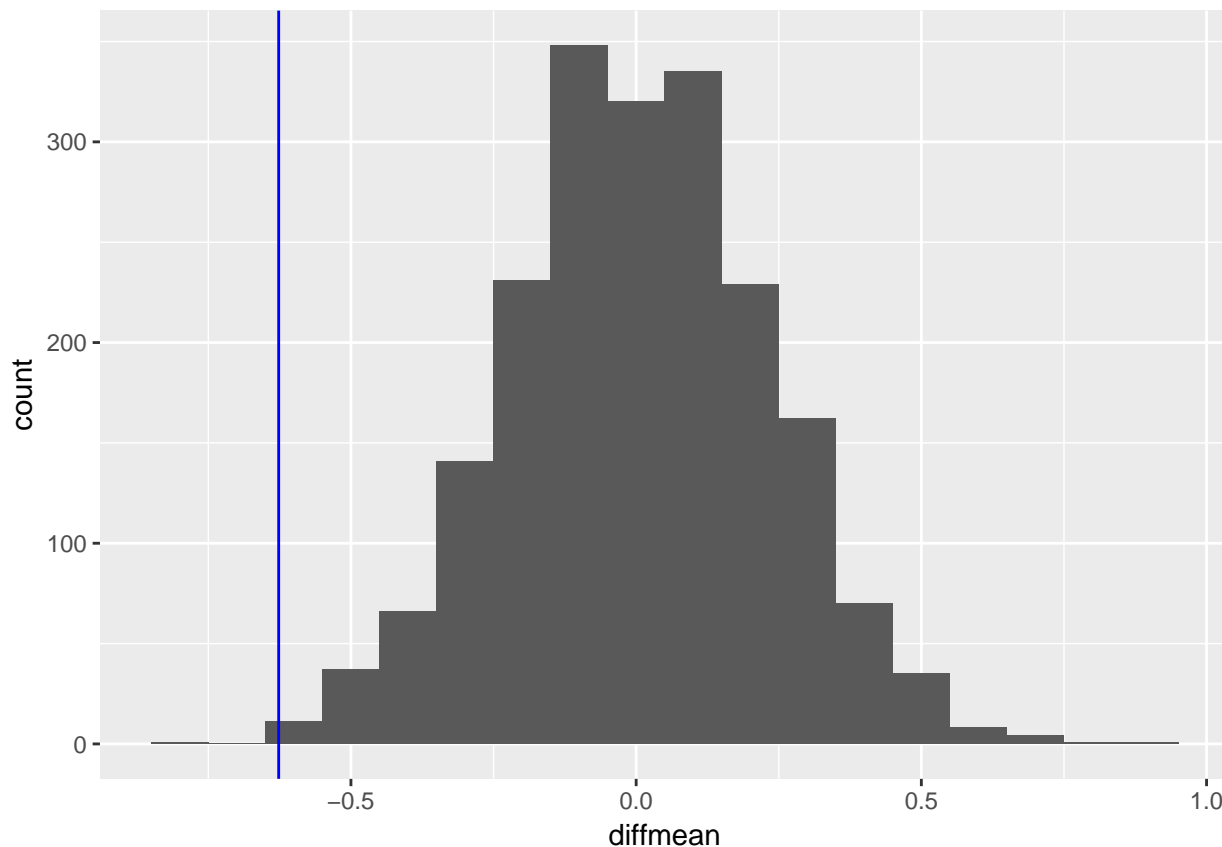
```
##    diffmean
## -0.1866667
```

```
diffmean(HeadWt ~ shuffle(Cult), data = cabbages)
```

```
##    diffmean
## 0.08666667
```

We use the do command to do this a bunch of times and graph the results along with our observed difference.

```
sims <- do(2000) * diffmean(HeadWt ~ shuffle(Cult), data = cabbages)
ggplot(sims, aes(x = diffmean)) +
    geom_histogram(binwidth = 0.1) +
    geom_vline(xintercept = obs_diff, color = "blue")
```

3

No surprise that this histogram looks nearly normal, centered at zero. Our observed difference is quite far out into the tail of this simulated sampling distribution, so it appears that our actual data would be unlikely due to pure chance alone if the null hypothesis were true.

We can even find a P-value by calculating how many of our sampled values are as extreme or more extreme than the observed data difference. The command below accomplishes this by calculating the percentage of samples in the left tail and then multiplying by 2 to make it a two-sided test. (It's two-sided because we didn't have any preconceptions about which cultivar would be heavier.)

```
2 * prop(sims$diffmean <= obs_diff)
```

```
##   TRUE
## 0.003
```

Indeed, this is a small P-value.

The sampling distribution model

In the previous section, we simulated the sampling distribution under the assumption of a null hypothesis of no difference between the groups. It certainly looked like a normal model, but which normal model? The center is obviously zero, but what about the standard deviation?

Let's assume that both groups come from populations that are normally distributed with normal models $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$. If we take samples of size n_1 from group 1 and n_2 from group 2, some fancy math shows that the distribution of the differences between sample means is

4

$$N \left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right).$$

Under the assumption of the null, the difference of the means is zero ($\mu_1 - \mu_2 = 0$). Unfortunately, though, we make no assumption on the standard deviations.

If we were testing two proportions with categorical data, we could pool our data. If the groups were truly equal, then they would share the same proportion, so we could use a pooled proportion to estimate it. With numerical data, we can calculate a pooled mean, but that doesn't help with the unknown standard deviations. Nothing in the null hypothesis suggests that the standard deviations of the two groups should be the same.[1]

It should be clear that the only solution is to substitute the sample standard deviations s_1 and s_2 for the population standard deviations σ_1 and σ_2.

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

However, s_1 and s_2 are not perfect estimates of σ_1 and σ_2; they are subject to sampling variability too. This extra variability means that a normal model is no longer appropriate as the sampling distribution model.

In the one-sample case, a Student t model with $df = n - 1$ was the right choice. In the two-sample case, we don't know the right answer. Yes, you read that correctly. Statisticians have not found a formula for the correct sampling distribution. It is a famous unsolved problem, called the Behrens-Fisher problem.

Several researchers have proposed solutions that are "close" though. One compelling one is called "Welch's t-test". Welch showed that even though it's not quite right, a Student t model is very close as long as you pick the degrees of freedom carefully. Unfortunately, the way to compute the right degrees of freedom is crazy complicated. Fortunately, R is good at crazy complicated computations. The `t.test` command uses the Welch's t-test by default when there are two groups.

Let's go through the full rubric using the cabbage example.

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 `?cabbages` to read the help file.]

```
str(cabbages)
```

```
## 'data.frame':    60 obs. of  4 variables:
##  $ Cult  : Factor w/ 2 levels "c39","c52": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Date  : Factor w/ 3 levels "d16","d20","d21": 1 1 1 1 1 1 1 1 1 1 ...
##  $ HeadWt: num  2.5 2.2 3.1 4.3 2.5 4.3 3.8 4.3 1.7 3.1 ...
##  $ VitC  : int  51 55 45 42 53 50 50 52 56 49 ...
```

We can also use `favstats` to see summary statistics of head weight for each cultivar.

```
favstats(HeadWt ~ Cult, data = cabbages)
```

---

[1]In the extremely rare situation in which one can assume equal standard deviations in the two groups, then there is a way to run a pooled t test. But this "extra" assumption of equal standard deviations is typically questionable.

```
##   Cult min    Q1 median   Q3 max     mean         sd  n missing
## 1  c39 1.6 2.500    2.8 3.45 4.3 2.906667 0.8046988 30       0
## 2  c52 1.0 1.625    2.1 2.80 4.2 2.280000 0.8563757 30       0
```

Prepare the data for analysis.

The cultivar variable `Cult` is already a factor variable, as it should be.

Make tables or plots to explore the data visually.
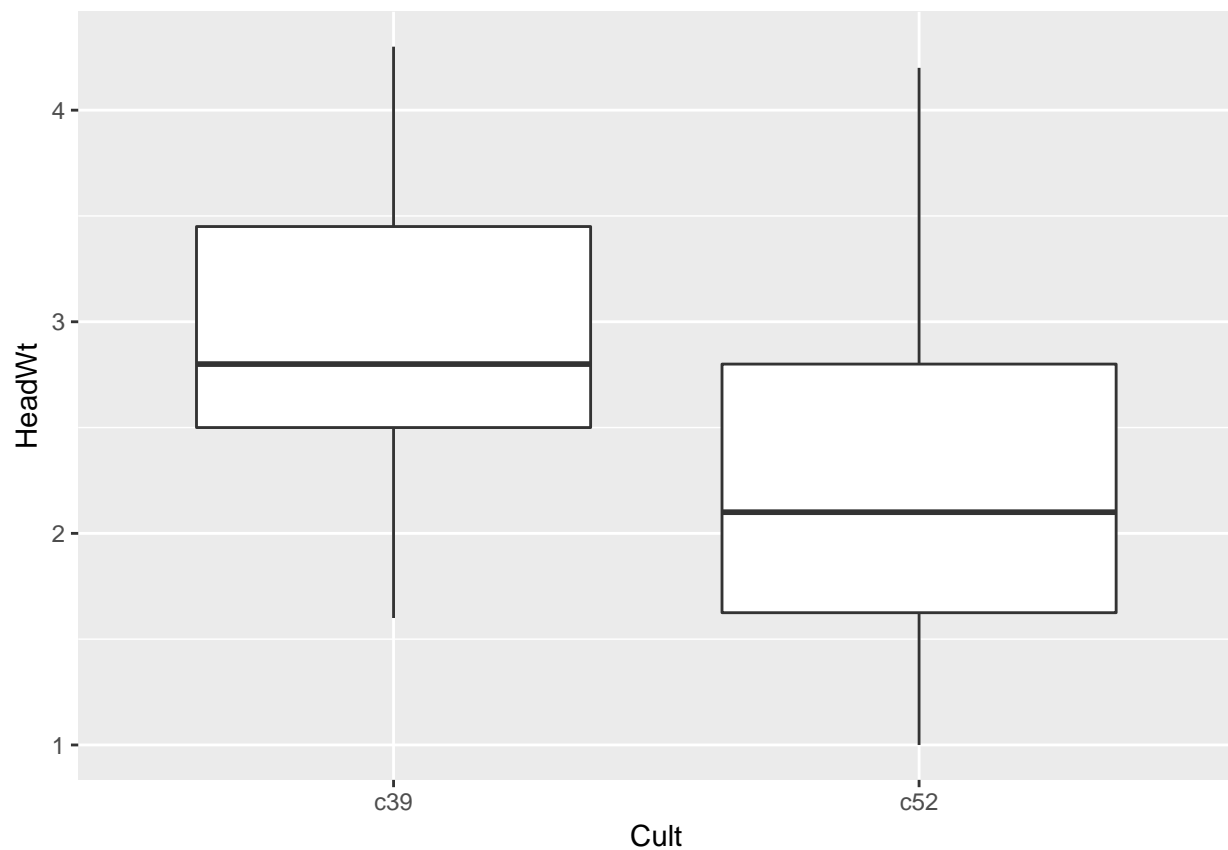
How many cabbages of each cultivar type do we have?
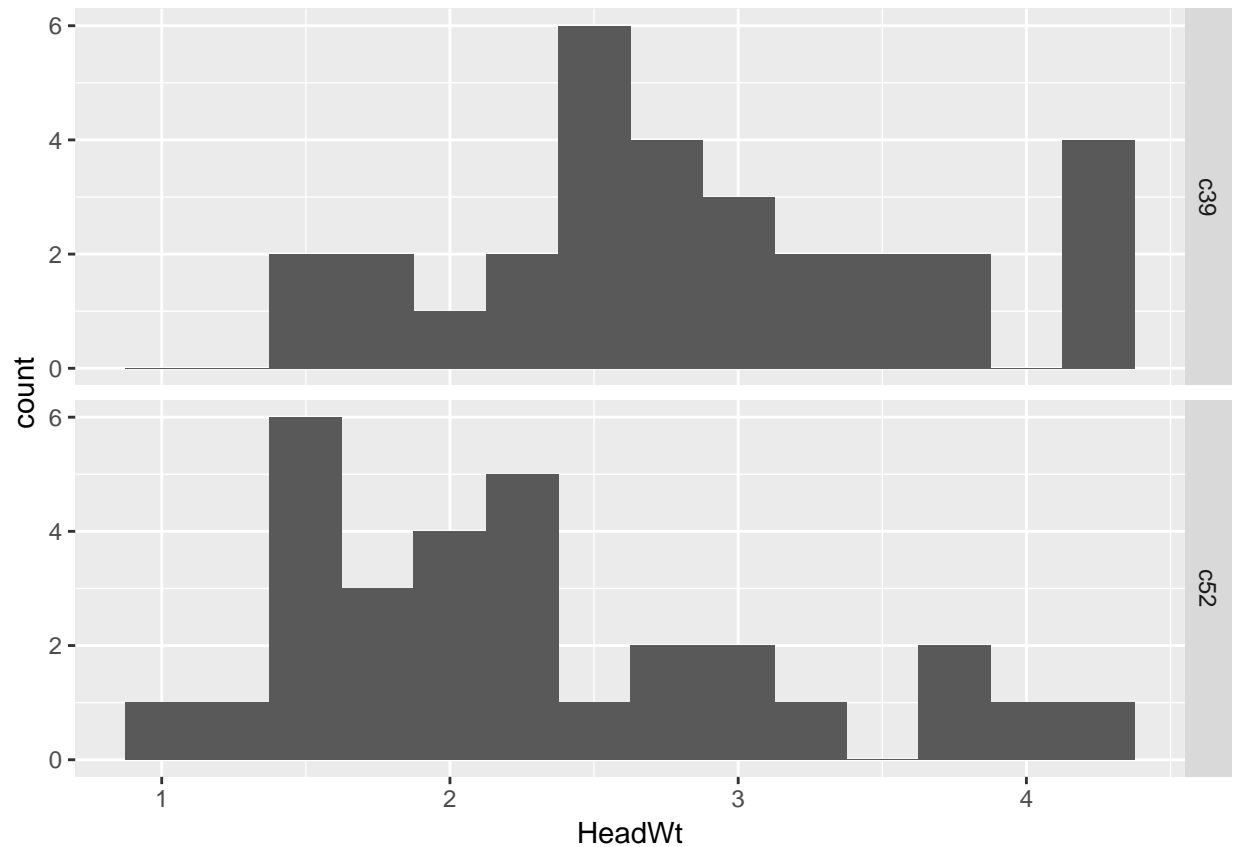
```
table(cabbages$Cult)
```

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

With a numerical response variable and a categorical explanatory variable, there are two useful plots: a side-by-side boxplot and a stacked histogram.

```
ggplot(cabbages, aes(x = Cult, y = HeadWt)) +
    geom_boxplot()
```
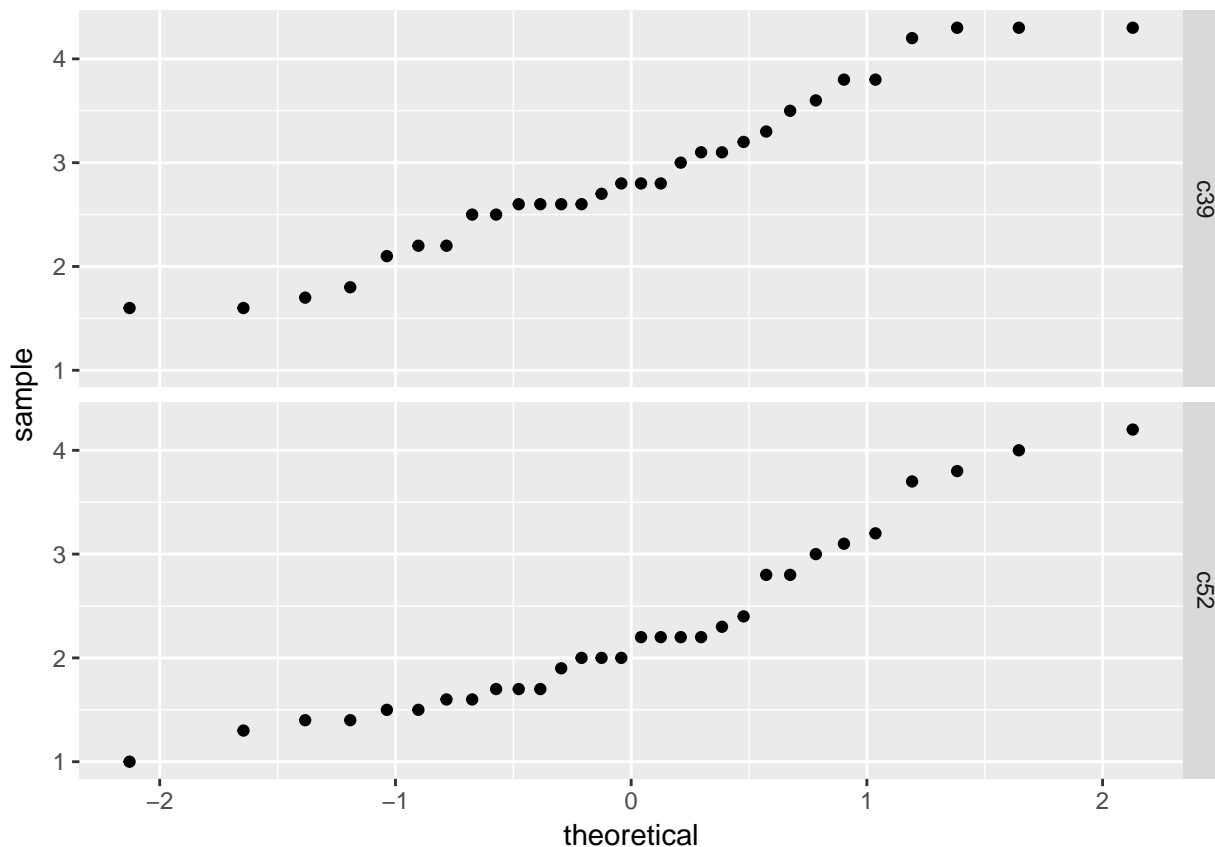


```
ggplot(cabbages, aes(x = HeadWt)) +
    geom_histogram(binwidth = 0.25) +
    facet_grid(Cult ~ .)
```

The individual group histograms looks reasonably normal, although it's a bit hard to tell with a sample size of 30 in each group. Here are the QQ plots to give us another way to ascertain normality of the data.

```
ggplot(cabbages, aes(sample = HeadWt)) +
    geom_qq() +
    facet_grid(Cult ~ .)
```

Commentary: The boxplots and histograms show why statistical inference is so important. It's clear that there is some difference between the two groups, but it's not obvious if that difference will turn out to be statistically significant. There appears to be a lot of variability in both groups, and both groups have a fair number of lighter and heavier cabbage heads.

Hypotheses

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

The samples consist of 30 cabbages from the c39 group and 30 cabbages from the c52 group. 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 (when possible).

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

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

Commentary: Pay close attention to the order of subtraction. It's easiest to make your hypotheses match the order of the `t.test` command we use later in the rubric. **It is the oppposite of the `diffmean` command!** How do we know? Let's run the `t.test` command a little early and look at the output.

```
cabbage_test <- tidy(t.test(HeadWt ~ Cult, data = cabbages))
cabbage_test
```

```
##    estimate estimate1 estimate2 statistic    p.value parameter  conf.low
## 1 0.6266667  2.906667      2.28  2.920875 0.004971691  57.77675 0.1971677
##   conf.high                 method alternative
## 1  1.056166 Welch Two Sample t-test    two.sided
```

The `estimate` is positive, obtained by subtracting `estimate1` minus `estimate2`. Looking back to the mean of both groups that we calculated with the `mean` command, we can see that `estimate1` corresponds to the c39 group and `estimate2` corresponds to the c52 group.

Model

Identify the sampling distribution model.

We use a t model. Since we ran the `t.test` command already, we can see the degrees of freedom:

```
cabbage_test$parameter
```

```
## [1] 57.77675
```

Commentary: For Welch's t test, the degrees of freedom won't usually be a whole number. Be sure you understand that the formula is no longer $df = n - 1$. That doesn't even make any sense as there isn't a single n in a *two*-sample test.

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

- Random
  - 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%
  - 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, though, we meet the condition automatically.

Mechanics
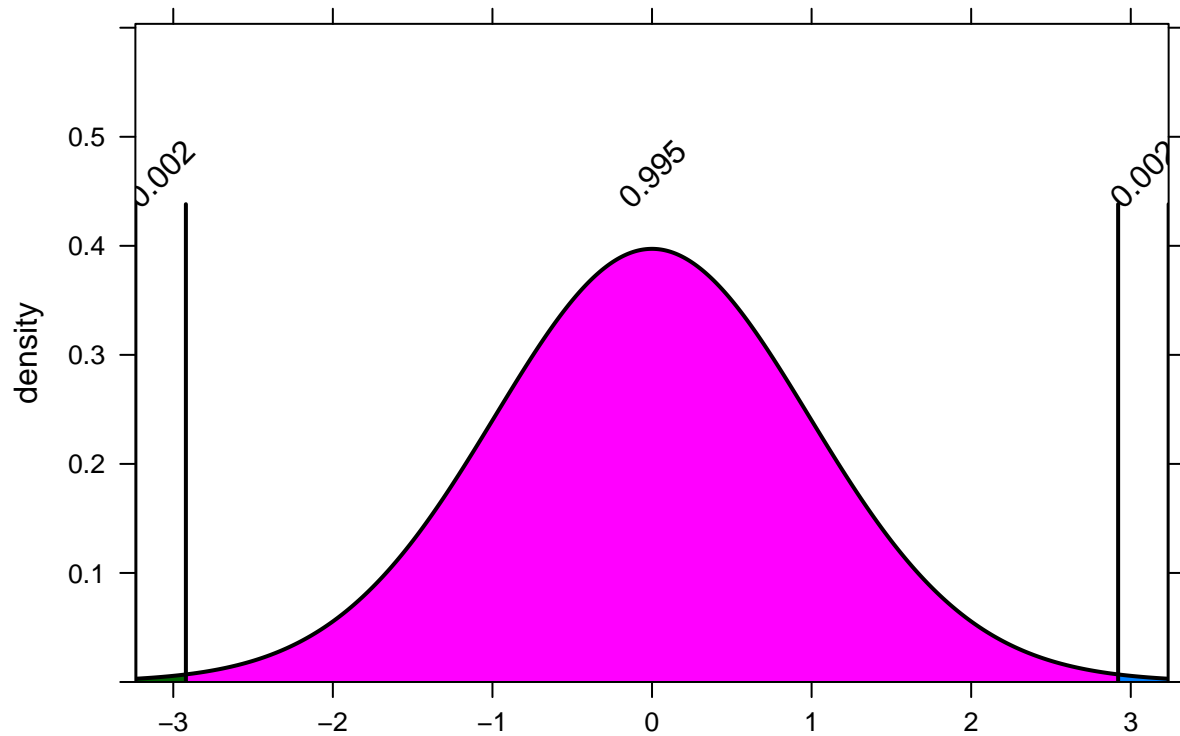
Compute the test statistic.

```
t <- cabbage_test$statistic
t
```

```
## [1] 2.920875
```

The t-score is 2.9208754.

Plot the null distribution.

```r
pdist("t", df = cabbage_test$parameter, q = c(-t, t), invisible = TRUE)
```


Calculate the P-value.

```r
cabbage_test$p.value
```

```
## [1] 0.004971691
```

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 the 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.

If we've made a Type I error, then that means that there might be no difference in the head weight of c39 cabbages and c52 cabbages, but we got some unusual samples that showed a difference.

Confidence interval

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

There are no additional conditions to check.

Calculate the confidence interval.

```
cabbage_test$conf.low
```

```
## [1] 0.1971677
```

```
cabbage_test$conf.high
```

```
## [1] 1.056166
```

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

We are 95% confident that the true difference in mean head weight between c52 and c39 cabbages is captured in the interval (0.1971677 kg, 1.0561656 kg). We obtained this by subtracting c39 minus c52.

Inference using summary statistics

In the previous example, we had access to the actual data frame. In some situations, you are not given the data; rather, all you have are summary statistics about the data. This certainly happens with homework problems from a textbook, but it can happen in "real life" too. If you're reading a research article, you will rarely have access to the original data used in the analysis. All you can see is what the researchers report in their paper.

For a two-sample t-test, often you have nothing but the sample sizes n_1 and n_2, the sample means $\bar{y}_1$ and $\bar{y}_2$, and the sample standard deviations s_1 and s_2.

Unlike the `binom.test` or `prop.test` commands that allow you to use either the raw data or summary statistics, the `t.test` command does not allow this. Instead, you have to calculate the t-score directly and use `pdist` to get the P-value.

For example, suppose you are told only that for the first group $n_1 = 48$, $\bar{y}_1 = 8.8$, and $s_1 = 2.1$; and for the second group $n_2 = 54$, $\bar{y}_2 = 10.4$, and $s_2 = 3.1$.

Using the formula for the standard error that appeared earlier in the module, here's the t score:

```
t <- (8.8 - 10.4)/sqrt(2.1^2/48 + 3.1^2/54)
t
```

```
## [1] -3.080126
```

The trouble we face is that the degrees of freedom will be complicated. In the event that the number of degrees of freedom are reported, we're good. For example, suppose we're told that there are 93.7 degrees of freedom. If that's the case, the P-value can be calculated:

```
2 * pdist("t", df = 93.7, q = t, plot = FALSE)
```

[1] 0.002715621

If you don't have the degrees of freedom, you could google the Welch-Satterthwaite formula and plug in all the necessary values to calculate degrees of freedom for yourself.[2] Or, another common recommendation is to use a conservative estimate for the degrees of freedom by choosing the smaller of $n_1 - 1$ or $n_2 - 1$. If we didn't know the degrees of freedom for the above example, we could use 47 degrees of freedom (48 - 1) and get the following P-value:

```
2 * pdist("t", df = 47, q = t,  plot = FALSE)
```

[1] 0.003452051

While a bit larger than the "correct" P-value, it's still plenty small. Either way we would reject the null.

Be careful: with only summary statistics, we can't do any exploratory data analysis, so it may be impossible to check conditions. The only condition we have to check with the raw data is the nearly normal condition. In this example, though, since both samples are larger than 30, we're not too worried.

Your turn

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

The rubric outline is reproduced below. You may refer to the worked example above and modify it accordingly. Remember to strip out all the commentary. That is just exposition for your benefit in understanding the steps, but is not meant to form part of the formal inference process.

Another word of warning: the copy/paste process is not a substitute for your brain. You will often need to modify more than just the names of the data frames and variables to adapt the worked examples to your own work. Do not blindly copy and paste code without understanding what it does. And you should **never** copy and paste text. All the sentences and paragraphs you write are expressions of your own analysis. They must reflect your own understanding of the inferential process.

Exploratory data analysis

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

<div align="center">━━━━━━━━━━ ANSWER ━━━━━━━━━━</div>

```
# Add code here to understand the data.
```

Prepare the data for analysis. [Not always necessary.]

<div align="center">━━━━━━━━━━ ANSWER ━━━━━━━━━━</div>

---

[2]This is painful.

```
# Add code here to prepare the data for analysis.
```

Make tables or plots to explore the data visually.

──────────────── ANSWER ────────────────

```
# Add code here to make tables or plots.
```

Hypotheses

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

──────────────── ANSWER ────────────────

Please write up your answer here.

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

──────────────── ANSWER ────────────────

H_0: Null hypothesis goes here.

H_A: Alternative hypothesis goes here.

Express the null and alternative hypotheses in symbols (when possible).

──────────────── ANSWER ────────────────

$H_0 : math$

$H_A : math$

Model

Identify the sampling distribution model.

──────────────── ANSWER ────────────────

Please write up your answer here.

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

ANSWER

Please write up your answer here. (Some conditions may require R code as well.)

Mechanics

Compute the test statistic.

ANSWER

```r
# Add code here to compute the test statistic.
```

Plot the null distribution.

ANSWER

```r
# Add code here to plot the null distribution.
```

Calculate the P-value.

ANSWER

```r
# Add code here to calculate the P-value.
```

Conclusion

State the statistical conclusion.

ANSWER

Please write up your answer here.

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

Please write up your answer here.

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.

Please write up your answer here.

Confidence interval

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

Please write up your answer here. (Some conditions may require R code as well.)

Calculate the confidence interval.

```
# Add code here to calculate the confidence interval.
```

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

Please write up your answer here.