Statistical hypothesis testing with paired data

Henry W J Reeve

henry.reeve@bristol.ac.uk

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.

What will we cover today?

- We will introduce the concept of statistical hypothesis testing for differences between paired samples.
- We will introduce the paired t-test for testing for differences in the value of the population mean.
- The paired t-test applies when either differences are approximately Gaussian or the sample size is large.
- We will consider the concept of effect size for assessing the magnitude of differences.

Comparing two samples

Hypothesis testing is a general methodology for drawing conclusions from data.

Consider the following example scenarios:

1. A farmer wants to know if applying different types of soil treatment will modify their crop yield.

2. A veterinarian wants to know if the weight of mice will change following a particular treatment.

3. A pharmaceutical company wants to know if a treatment for a medical condition is effective.

In each of these cases we want to compare and contrast a variable under two conditions.

This leads to the topic of two sample hypothesis testing.

Paired data

Paired data consists occurs when you have access to two samples:

$$X_1^0, \cdots, X_n^0 \qquad \text{and} \qquad X_1^1, \cdots, X_n^1$$

These correspond to a variable measured across two samples with a natural pairing between the samples.

$$\begin{tikzcd}
X_1^0 \arrow[d, leftrightarrow] & X_2^0 \arrow[d, leftrightarrow] & X_3^0 \arrow[d, leftrightarrow] & \cdots & X_n^0 \arrow[d, leftrightarrow] \\
X_1^1 & X_2^1 & X_3^1 & \cdots & X_n^1
\end{tikzcd}$$

Example: The effect of soil treatment on yield

Suppose a farmer wants to know if the yields of his wheat fields will change following soil treatment.

The farmer has a collection of $n = 15$ fields.

An experiment takes place over two years.

Each of the fields are treated in one of the two years and untreated in another.

Let $X_1^0, \cdots, X_n^0$ be the number of bushels for fields when untreated and $X_1^1, \cdots, X_n^1$ when untreated.

Note that the data are **paired**:

For each $i = 1, \cdots, n$, X_i^0 and X_i^1 denote yields with and without treatment for the same field.

Hypothesis testing

Suppose we have a clear research hypothesis and some high-quality data from a well-deigned experiment.

The key stages of statistical hypothesis testing are as follows:

1. Form our statistical hypothesis including a null hypothesis and an alternative hypothesis.

2. Apply model checking to validate any modelling assumptions.

3. Choose our desired significance level.

4. Select an appropriate statistical test.

5. Compute the numerical value of the test statistic from data.

6. Compute a p-value based upon the test statistic.

7. Draw conclusions based upon the relationship between the p-value and the significance level.

Null and alternative hypothesis

The **statistical hypothesis** frames the research question in terms of the parameters of a statistical model.

There are two hypotheses:

H_0 : The **null hypothesis** is our default position typically declaring an absence of an interesting phenomena.

H_1 : The **alternative hypothesis** is the of something interesting difference we'd like to demonstrate.

Example

Research question: Will the amount of wheat produced change following the soil treatment?

Null and alternative hypothesis

Research question: Will the amount of wheat produced change following the soil treatment?

Let $X_1^0, \cdots, X_n^0$ be the yields of untreated fields and let $X_1^1, \cdots, X_n^1$ be the yields of treated fields.

For each $i = 1, \cdots, n$, X_i^0 and X_i^1 denote untreated and treated yield for the same field.

For each $i = 1, \cdots, n$, we let $Y_i := X_i^1 - X_i^0$ be the difference between the two weights.

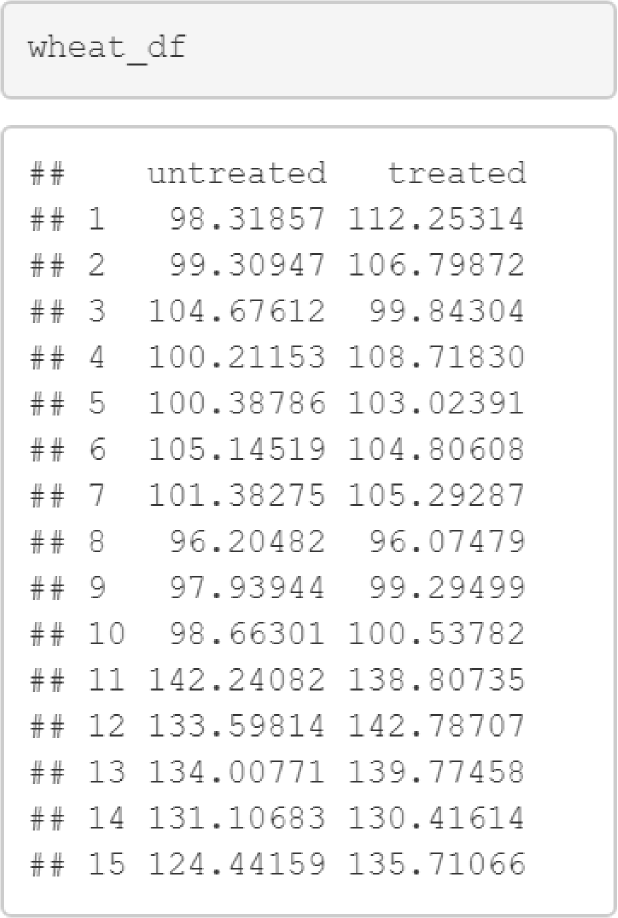
We model the changes as i.i.d. draws from a Gaussian distribution: $Y_1, \cdots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$

Null hypothesis: $\mathrm{H}_0 : \quad \mu = 0$

Alternative hypothesis: $\mathrm{H}_1 : \quad \mu \neq 0$

Checking our assumptions

For each $i = 1, \cdots, n$, we let $Y_i := X_i^1 - X_i^0$ be the difference between the two weights.

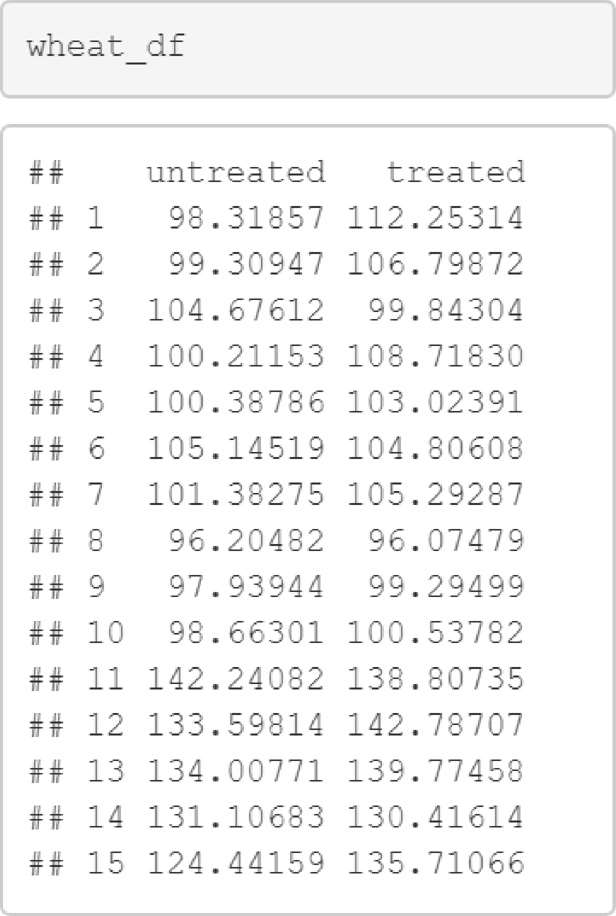


```
wheat_df
```

```
##    untreated   treated
## 1   98.31857 112.25314
## 2   99.30947 106.79872
## 3  104.67612  99.84304
## 4  100.21153 108.71830
## 5  100.38786 103.02391
## 6  105.14519 104.80608
## 7  101.38275 105.29287
## 8   96.20482  96.07479
## 9   97.93944  99.29499
## 10  98.66301 100.53782
## 11 142.24082 138.80735
## 12 133.59814 142.78707
## 13 134.00771 139.77458
## 14 131.10683 130.41614
## 15 124.44159 135.71066
```

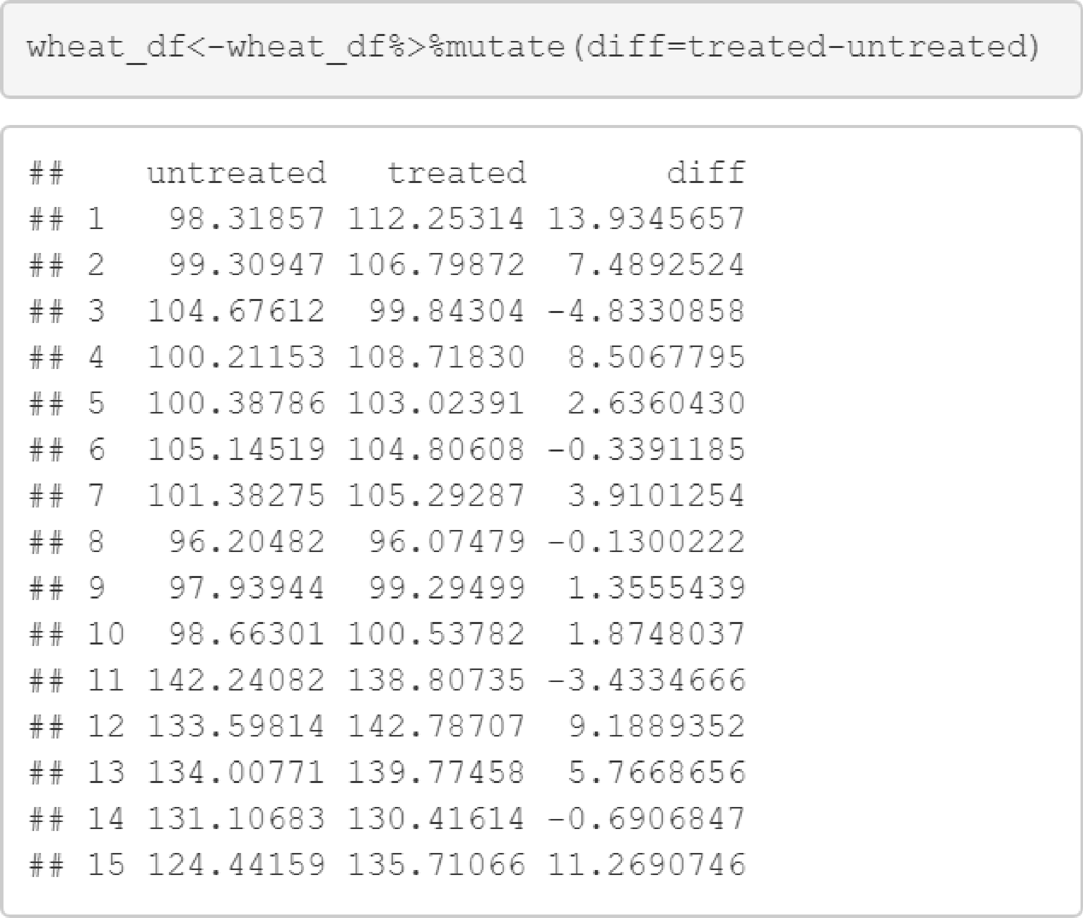
Checking our assumptions

For each $i = 1, \cdots, n$, we let $Y_i := X_i^1 - X_i^0$ be the difference between the two weights.



```
wheat_df
```

```
##    untreated   treated
## 1   98.31857 112.25314
## 2   99.30947 106.79872
## 3  104.67612  99.84304
## 4  100.21153 108.71830
## 5  100.38786 103.02391
## 6  105.14519 104.80608
## 7  101.38275 105.29287
## 8   96.20482  96.07479
## 9   97.93944  99.29499
## 10  98.66301 100.53782
## 11 142.24082 138.80735
## 12 133.59814 142.78707
## 13 134.00771 139.77458
## 14 131.10683 130.41614
## 15 124.44159 135.71066
```



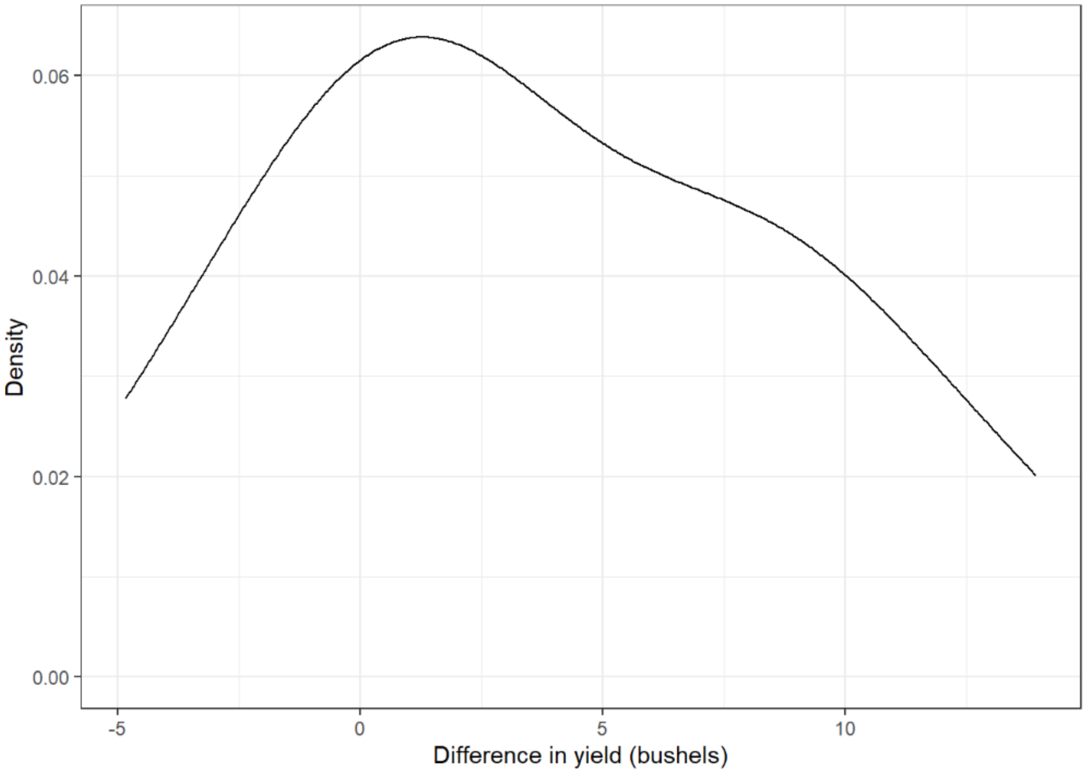
```
wheat_df<-wheat_df%>%mutate(diff=treated-untreated)
```

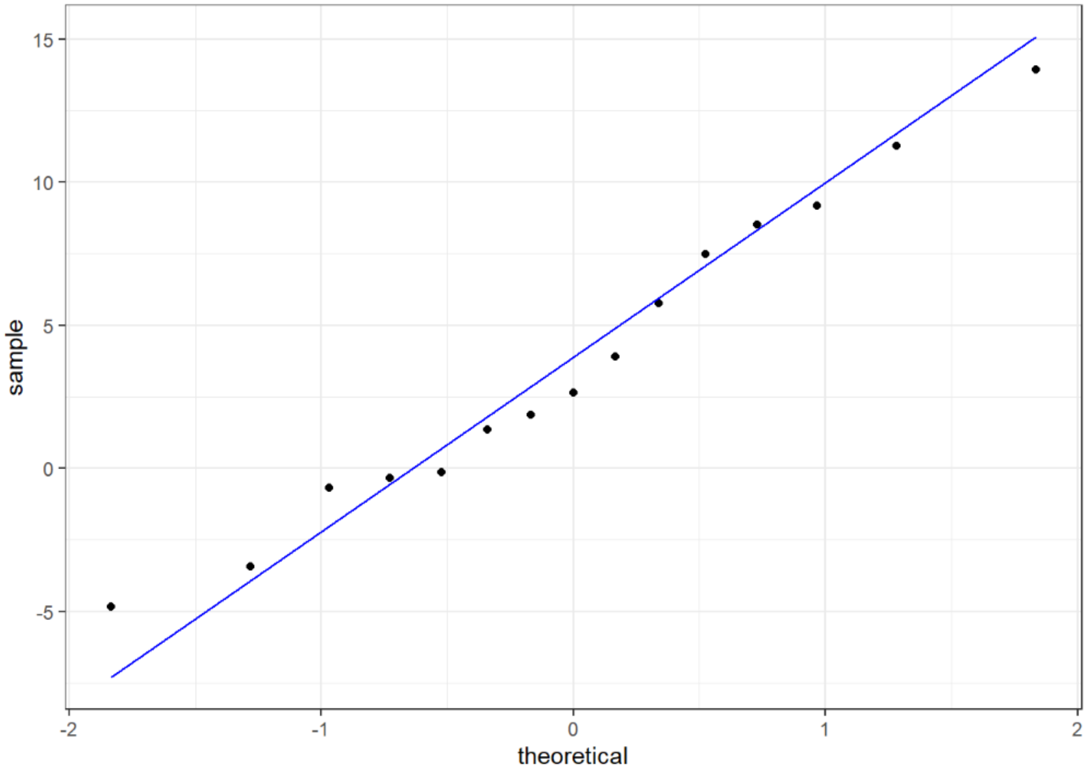
```
##    untreated   treated       diff
## 1   98.31857 112.25314 13.9345657
## 2   99.30947 106.79872  7.4892524
## 3  104.67612  99.84304 -4.8330858
## 4  100.21153 108.71830  8.5067795
## 5  100.38786 103.02391  2.6360430
## 6  105.14519 104.80608 -0.3391185
## 7  101.38275 105.29287  3.9101254
## 8   96.20482  96.07479 -0.1300222
## 9   97.93944  99.29499  1.3555439
## 10  98.66301 100.53782  1.8748037
## 11 142.24082 138.80735 -3.4334666
## 12 133.59814 142.78707  9.1889352
## 13 134.00771 139.77458  5.7668656
## 14 131.10683 130.41614 -0.6906847
## 15 124.44159 135.71066 11.2690746
```

Checking our assumptions

For each $i = 1, \cdots, n$, we let $Y_i := X_i^1 - X_i^0$ be the difference between the two weights.

We model the changes as i.i.d. draws from a Gaussian distribution: $Y_1, \cdots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$

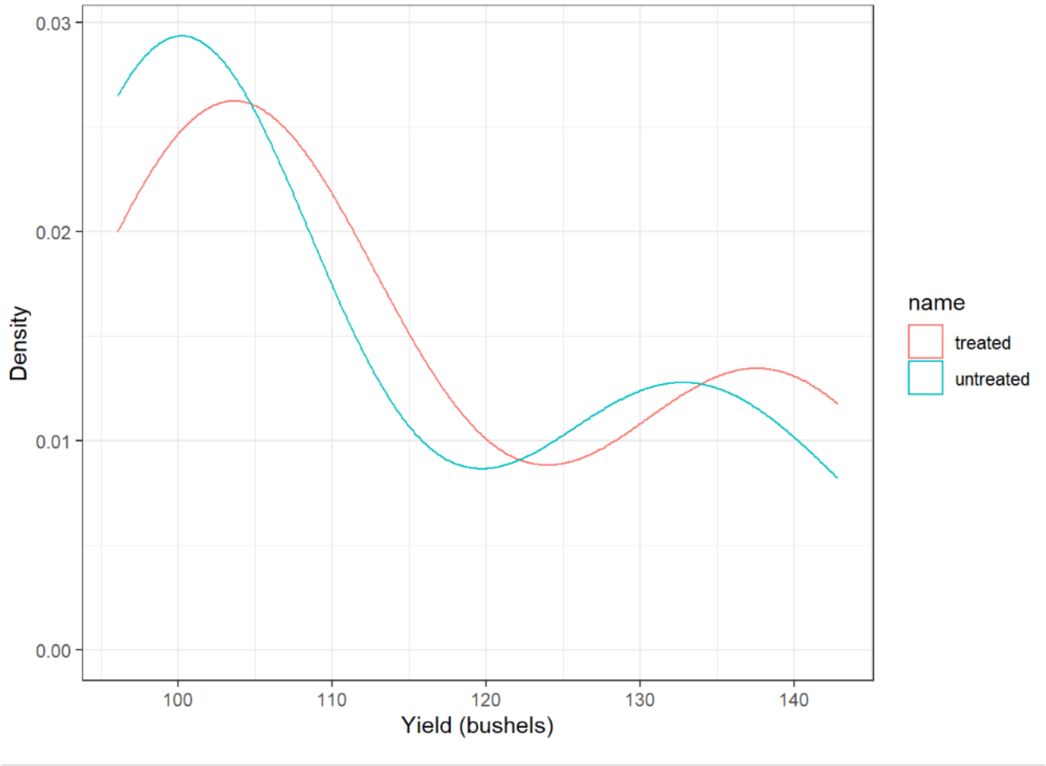


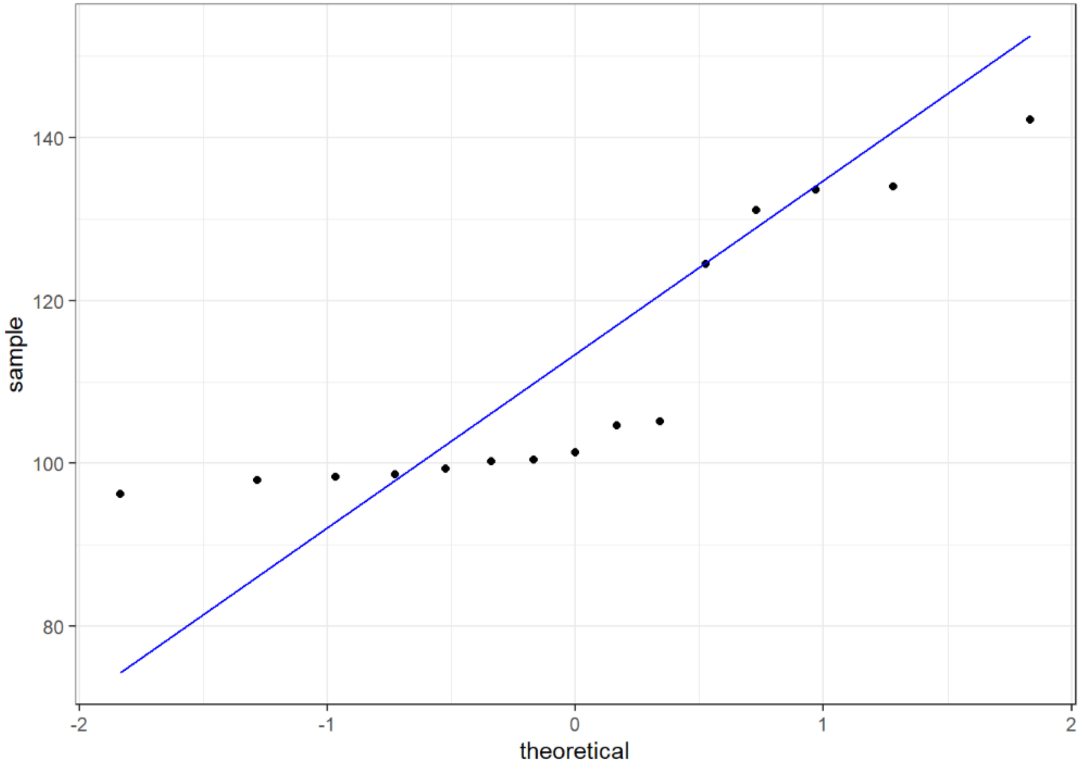


Checking our assumptions

We model the changes as i.i.d. draws from a Gaussian distribution: $Y_1, \cdots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$

Note that this does not require that $X_1^0, \cdots, X_n^0$ and $X_1^1, \cdots, X_n^1$ are Gaussian.

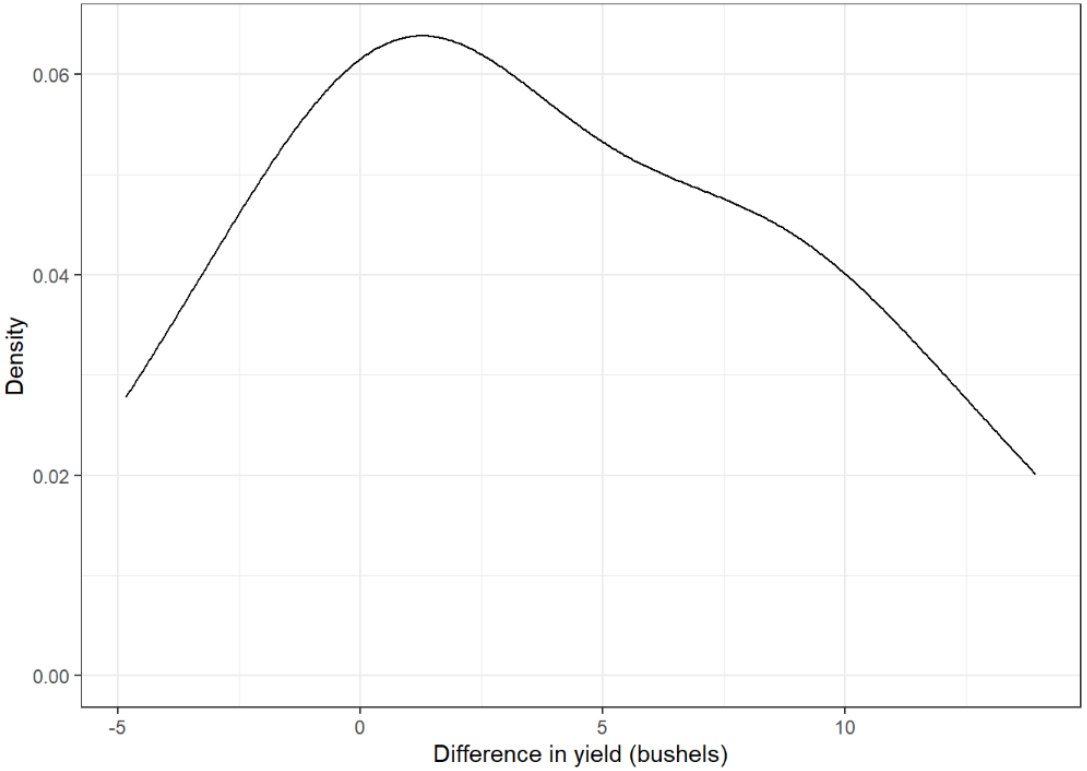


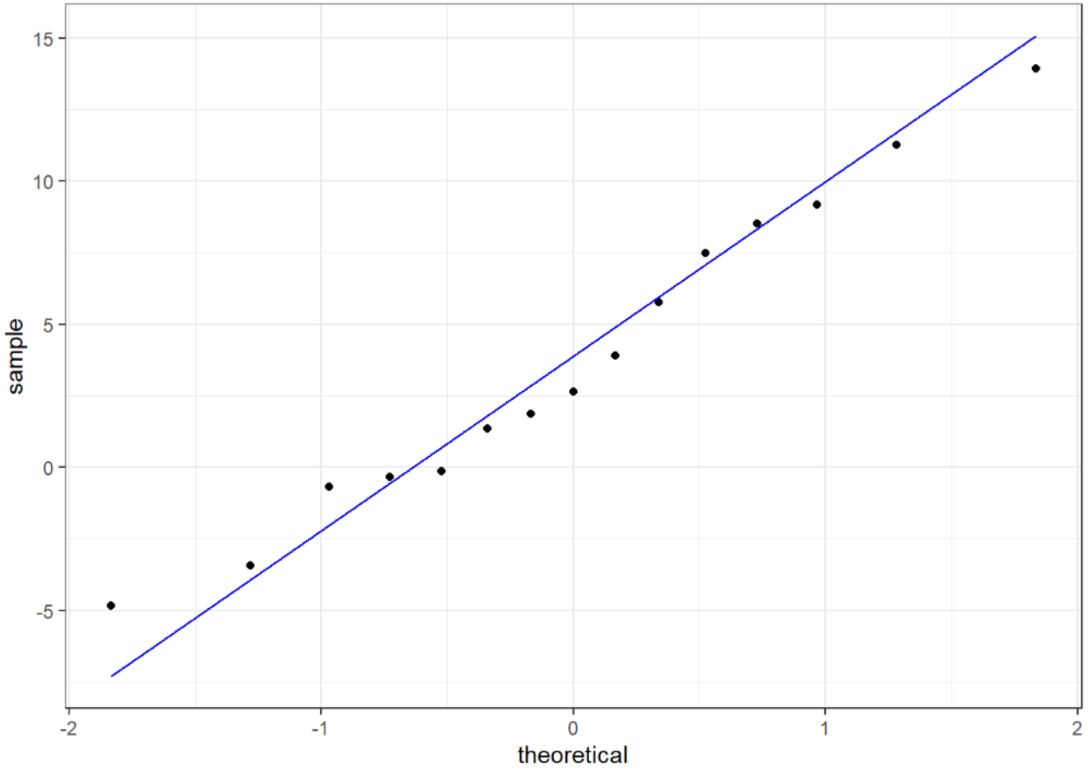


Checking our assumptions

For each $i = 1, \cdots, n$, we let $Y_i := X_i^1 - X_i^0$ be the difference between the two weights.

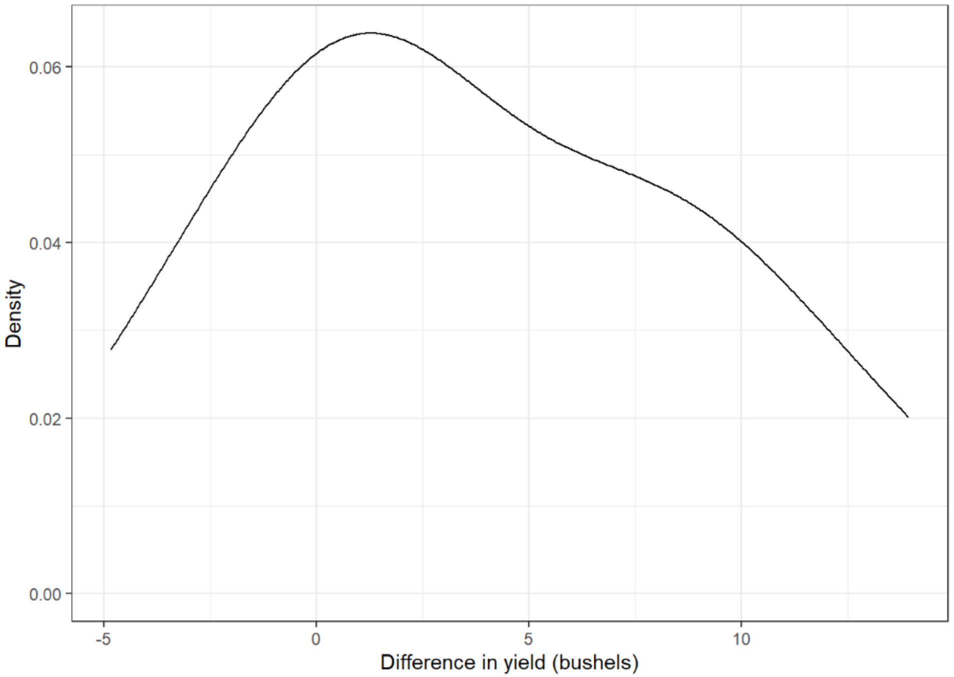
We model the changes as i.i.d. draws from a Gaussian distribution: $Y_1, \cdots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$

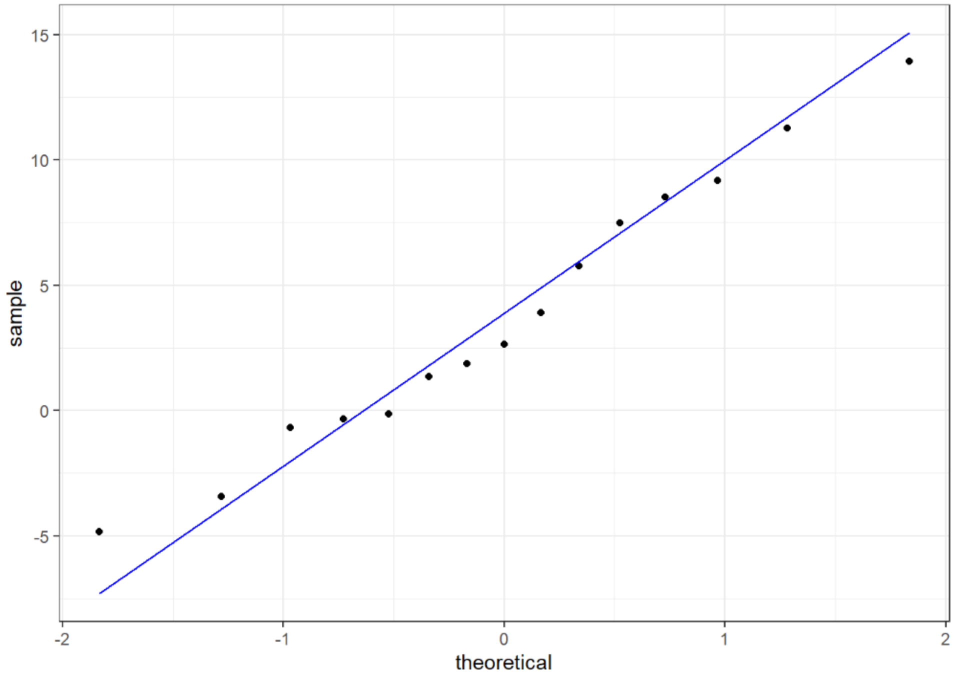




Choosing a significance level

We model the changes as i.i.d. draws from a Gaussian distribution: $Y_1, \cdots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$





Next we choose a significance level of $\alpha = 0.05$.

This must be done before carrying out the statistical hypothesis test.

The paired t-test

For each $i = 1, \cdots, n$, we let $Y_i := X_i^1 - X_i^0$ be the difference between the two weights.

We model the changes as i.i.d. draws from a Gaussian distribution: $Y_1, \cdots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$

Our null hypothesis is $\mathrm{H}_0 : \quad \mu = 0$ and our alternative hypothesis is $\mathrm{H}_1 : \quad \mu \neq 0$.

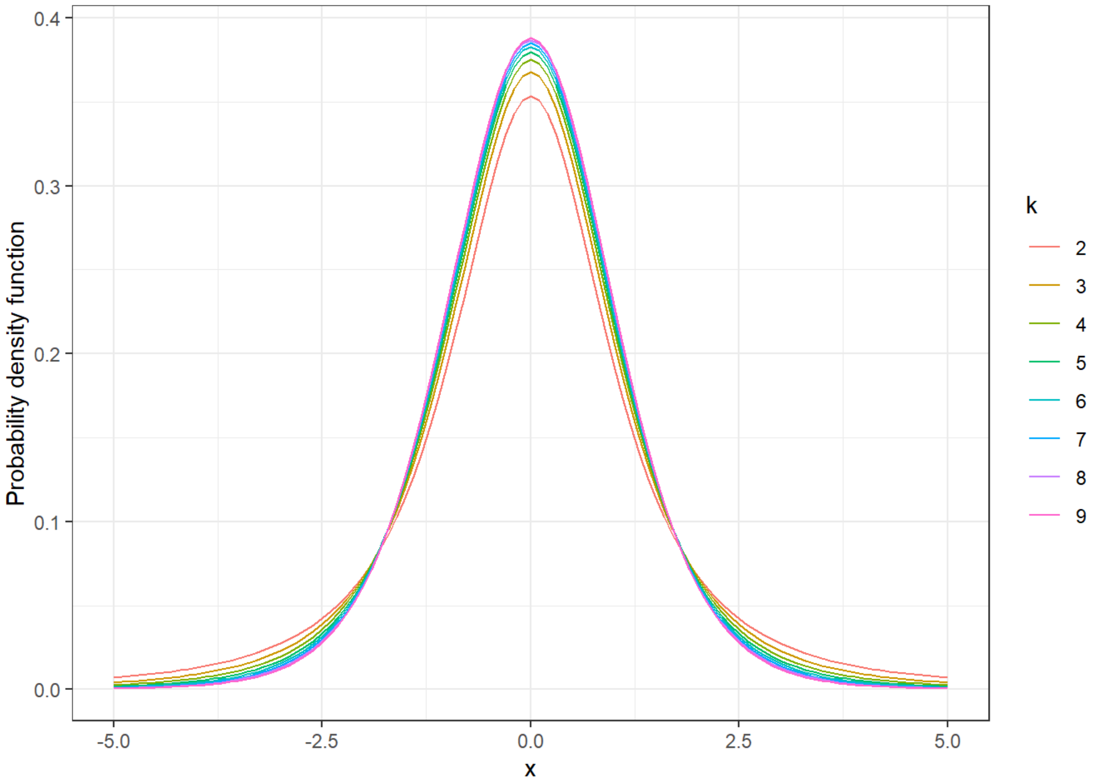
The canonical procedure for this setting is the **paired t-test** with test statistic,

$$T := \frac{\overline{Y}}{S_Y / \sqrt{n}} \quad \text{where} \quad \overline{Y} := \frac{1}{n} \sum_{i=1}^{n} Y_i \quad \text{and} \quad (S_Y)^2 := \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})$$

Under the null-hypothesis the paired t-test statistic is t-distributed with n-1 degrees of freedom.

Student’s t distribution

We model the changes as i.i.d. draws from a Gaussian distribution: $Y_1, \cdots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$



Under the null hypothesis, $\mathrm{H}_0 : \ \mu = 0$

Let $\overline{Y} := \frac{1}{n} \sum_{i=1}^{n} Y_i$

& $(S_Y)^2 := \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})$

Then $T := \frac{\overline{Y}}{S_Y / \sqrt{n}}$

is t-distributed with n-1 degrees of freedom.

The paired t-test

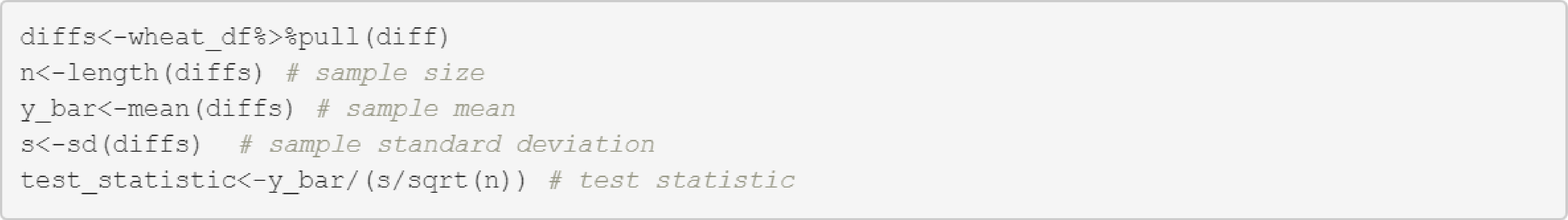
We assume $Y_1, \cdots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$ (i.i.d.). Our null is $\mu = 0$ and our alternative is $\mu \neq 0$.

The canonical procedure for this setting is the **paired t-test** with test statistic,

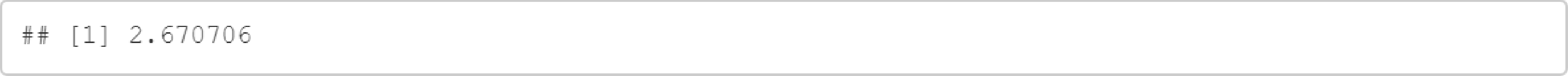
$$T := \frac{\overline{Y}}{S_Y/\sqrt{n}} \qquad \text{where} \quad \overline{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \text{and} \quad (S_Y)^2 := \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})$$

Under the null-hypothesis the paired t-test statistic is t-distributed with n-1 degrees of freedom.

Next, we compute the numerical value of the test statistic based on our two samples $\mathcal{T}$.



```r
diffs<-wheat_df%>%pull(diff)
n<-length(diffs) # sample size
y_bar<-mean(diffs) # sample mean
s<-sd(diffs)  # sample standard deviation
test_statistic<-y_bar/(s/sqrt(n)) # test statistic
```



```
## [1] 2.670706
```

The paired t-test

We assume $Y_1, \cdots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$ (i.i.d.). Our null is $\mu = 0$ and our alternative is $\mu \neq 0$.

The canonical procedure for this setting is the **paired t-test** with test statistic T

Under the null-hypothesis the paired t-test statistic is t-distributed with n-1 degrees of freedom.

Suppose we have computed the numerical value of the test statistic based on our two samples $\mathcal{T}$.

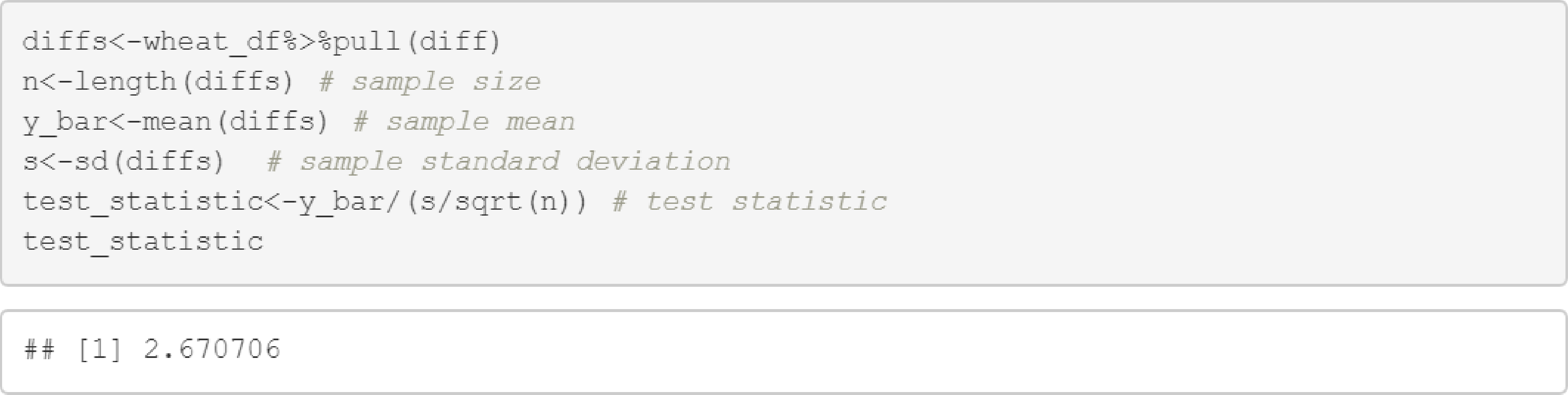
The **p-value** is the probability that T takes on a value at least as extreme as $\mathcal{T}$ under the null H_0 ,

We can compute the **p-value** using properties of Student's t-distribution as follows:

$$p = \mathbb{P}\left(|T| \geq |\tau| | H_0\right) = 2 \cdot \left(1 - F_{n-1}(|\tau|)\right)$$

The paired t-test

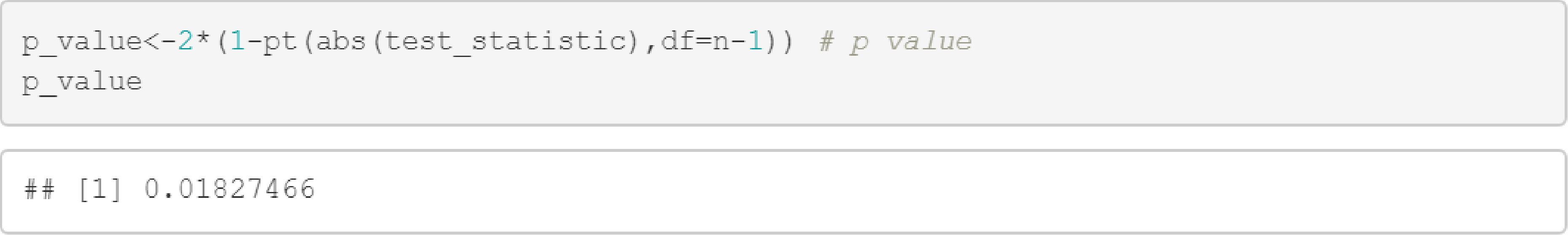
We compute the numerical value of the test-statistic.



```
diffs<-wheat_df%>%pull(diff)
n<-length(diffs) # sample size
y_bar<-mean(diffs) # sample mean
s<-sd(diffs)  # sample standard deviation
test_statistic<-y_bar/(s/sqrt(n)) # test statistic
test_statistic
```

```
## [1] 2.670706
```

Then we compute the p-value

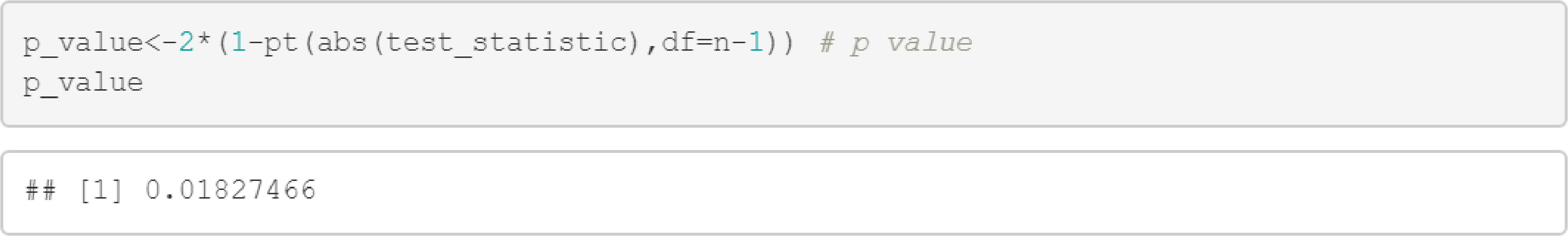


```
p_value<-2*(1-pt(abs(test_statistic),df=n-1)) # p value
p_value
```

```
## [1] 0.01827466
```

The paired t-test

Then we compute the p-value:



```
p_value<-2*(1-pt(abs(test_statistic),df=n-1)) # p value
p_value
```

```
## [1] 0.01827466
```

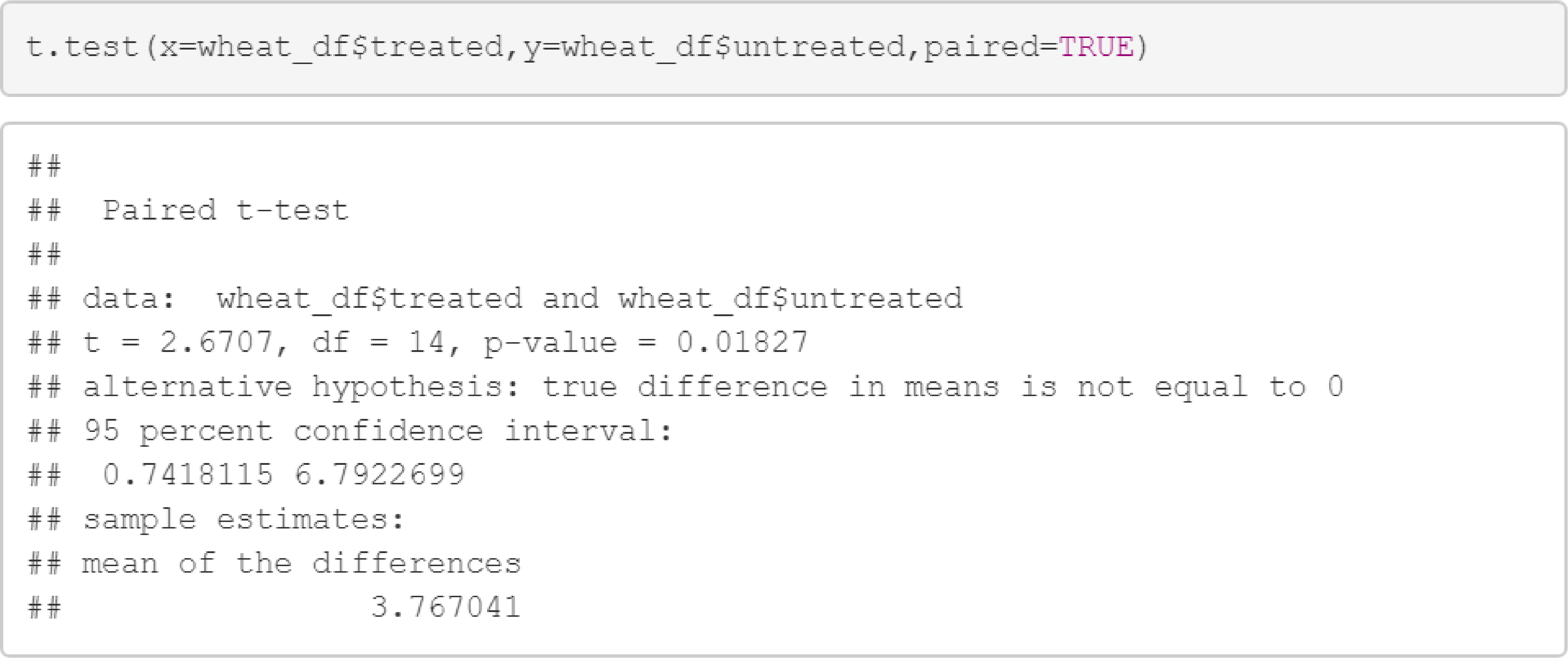
The p-value is 0.0183 which is below the significance level of $\alpha = 0.05$.

Hence, we reject the null hypothesis of $\mathrm{H}_0 : \quad \mu = 0$

We conclude that the alternative hypothesis holds $\mathrm{H}_1 : \quad \mu \neq 0$

The paired t-test

We can efficiently carry out the paired t-test in R as follows.



```
t.test(x=wheat_df$treated,y=wheat_df$untreated,paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  wheat_df$treated and wheat_df$untreated
## t = 2.6707, df = 14, p-value = 0.01827
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7418115 6.7922699
## sample estimates:
## mean of the differences
##                3.767041
```

Experimental design

For our yield experiment we computed a p-value is 0.0183 which is below the significance level.

Hence, we are justified in rejecting the null $\mu = 0$ and concluding that $\mu > 0$....

... provided our assumption that $Y_1, \cdots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$ (independent) holds.

The independence assumption is often violated in practice e.g. Imagine a crop disease across the fields.

Even if we have shown that $\mu > 0$ does it follow that the change was ***caused*** by the treatment?

If all fields are treated in one year and untreated in another there could be another causal factor e.g. weather.

To establish causal relationships we can use **randomized allocation** as discussed in future lectures...

The paired t-test and the one sample t-test

The one-sample t-test

We assume $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. We have $\mathrm{H}_0 : \ \mu = \mu_0$ & $\mathrm{H}_1 : \ \mu \neq \mu_0$

Our test statistic is $\hat{T}_{\mu_0} = \left(\overline{X} - \mu_0\right) / \left(S_X / \sqrt{n}\right)$

The paired t-test

We have paired data $X_1^0, \cdots, X_n^0$ and $X_1^1, \cdots, X_n^1$. We let $Y_i := X_i^1 - X_i^0$

We assume $Y_1, \cdots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$. We have $\mathrm{H}_0 : \ \mu = 0$ & $\mathrm{H}_1 : \ \mu \neq 0$

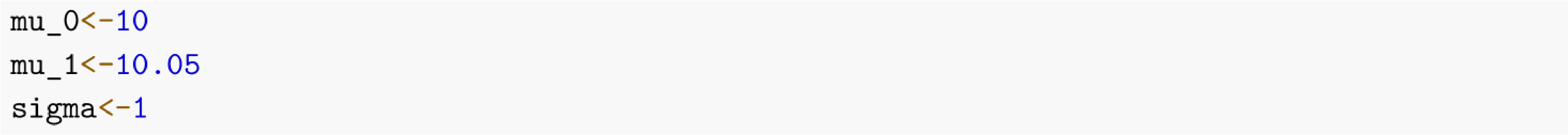
Our test statistic is $\hat{T}_0 = \overline{Y} / \left(S_Y / \sqrt{n}\right)$.

We can view the paired t-test as a special case of the one sample t-test on the differences!

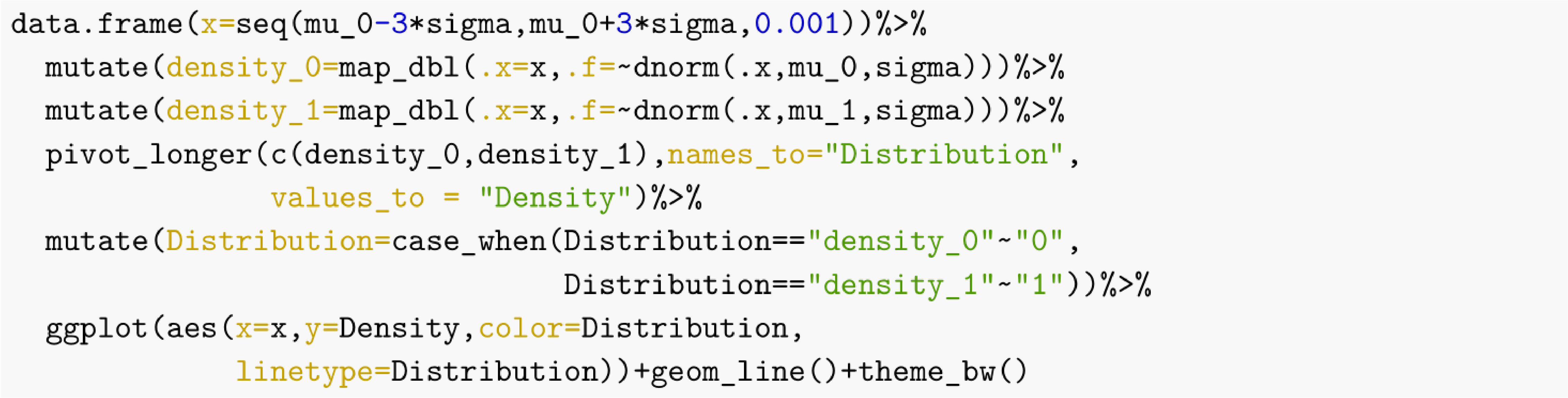
Now take a break!

Statistical significance and effect size

Statistical significance is not the same as a meaningful difference between distributions



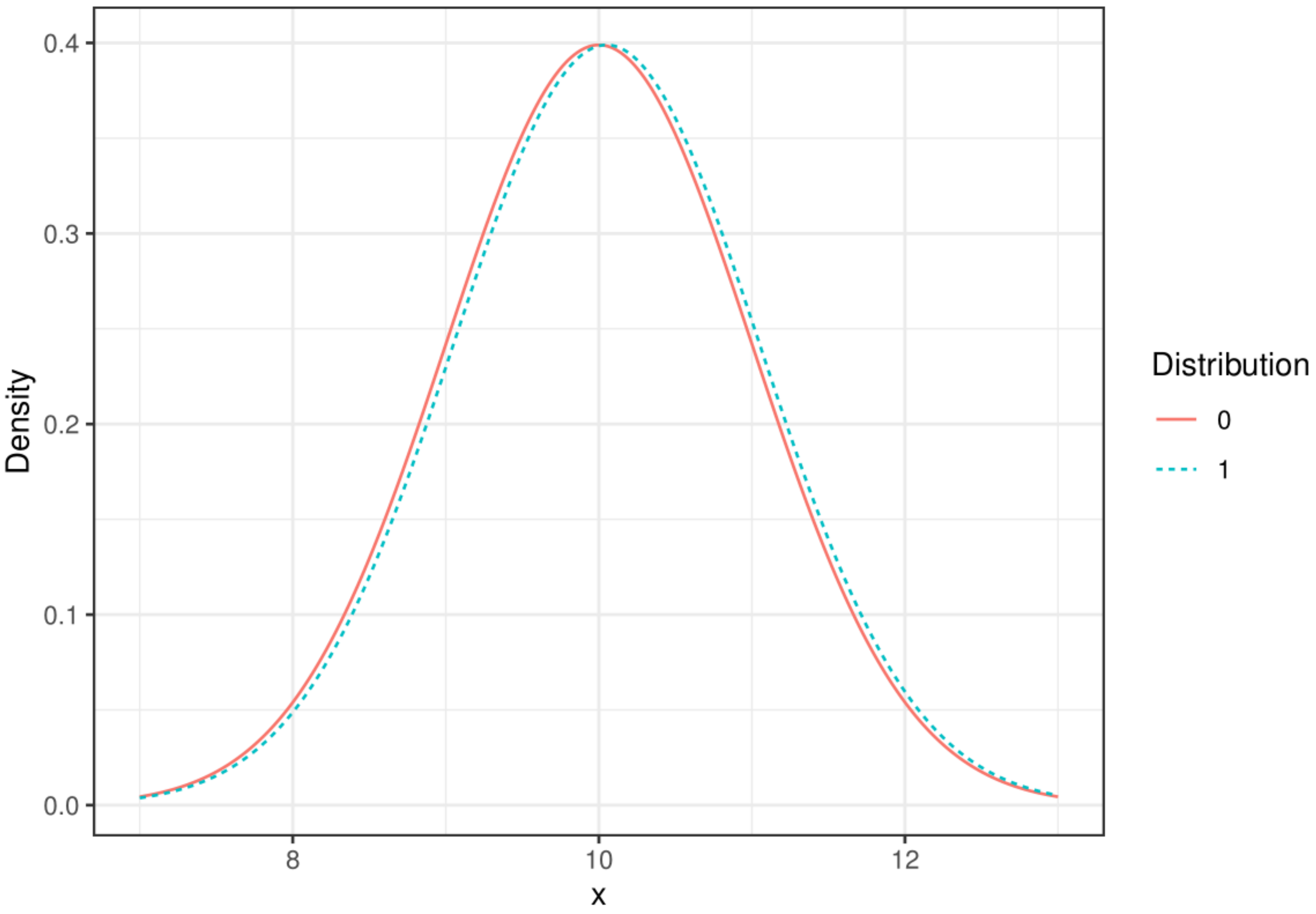
```r
mu_0<-10
mu_1<-10.05
sigma<-1
```



```r
data.frame(x=seq(mu_0-3*sigma,mu_0+3*sigma,0.001))%>%
  mutate(density_0=map_dbl(.x=x,.f=~dnorm(.x,mu_0,sigma)))%>%
  mutate(density_1=map_dbl(.x=x,.f=~dnorm(.x,mu_1,sigma)))%>%
  pivot_longer(c(density_0,density_1),names_to="Distribution",
               values_to = "Density")%>%
  mutate(Distribution=case_when(Distribution=="density_0"~"0",
                                Distribution=="density_1"~"1"))%>%
  ggplot(aes(x=x,y=Density,color=Distribution,
             linetype=Distribution))+geom_line()+theme_bw()
```

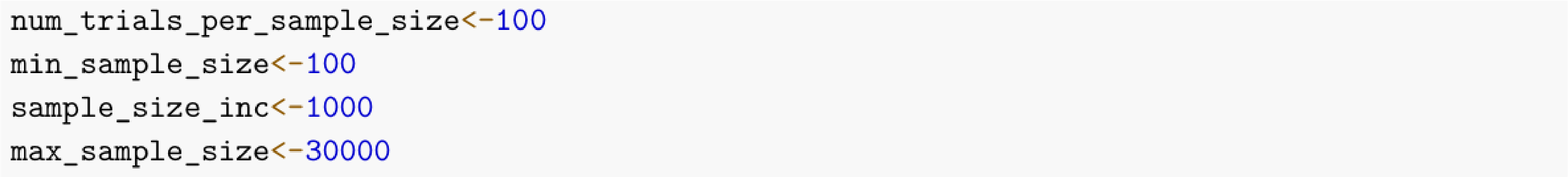
Statistical significance and effect size

Statistical significance is not the same as a meaningful difference between distributions

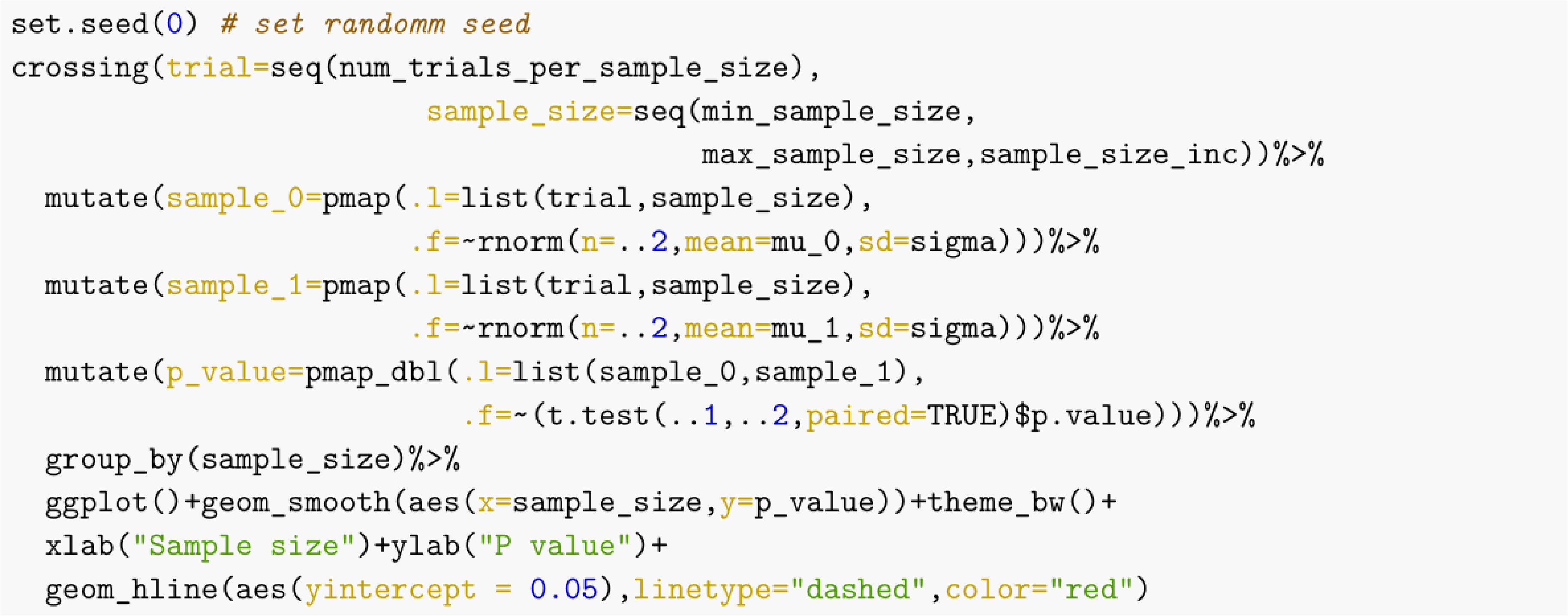


Statistical significance and effect size

Statistical significance is not the same as a meaningful difference between distributions



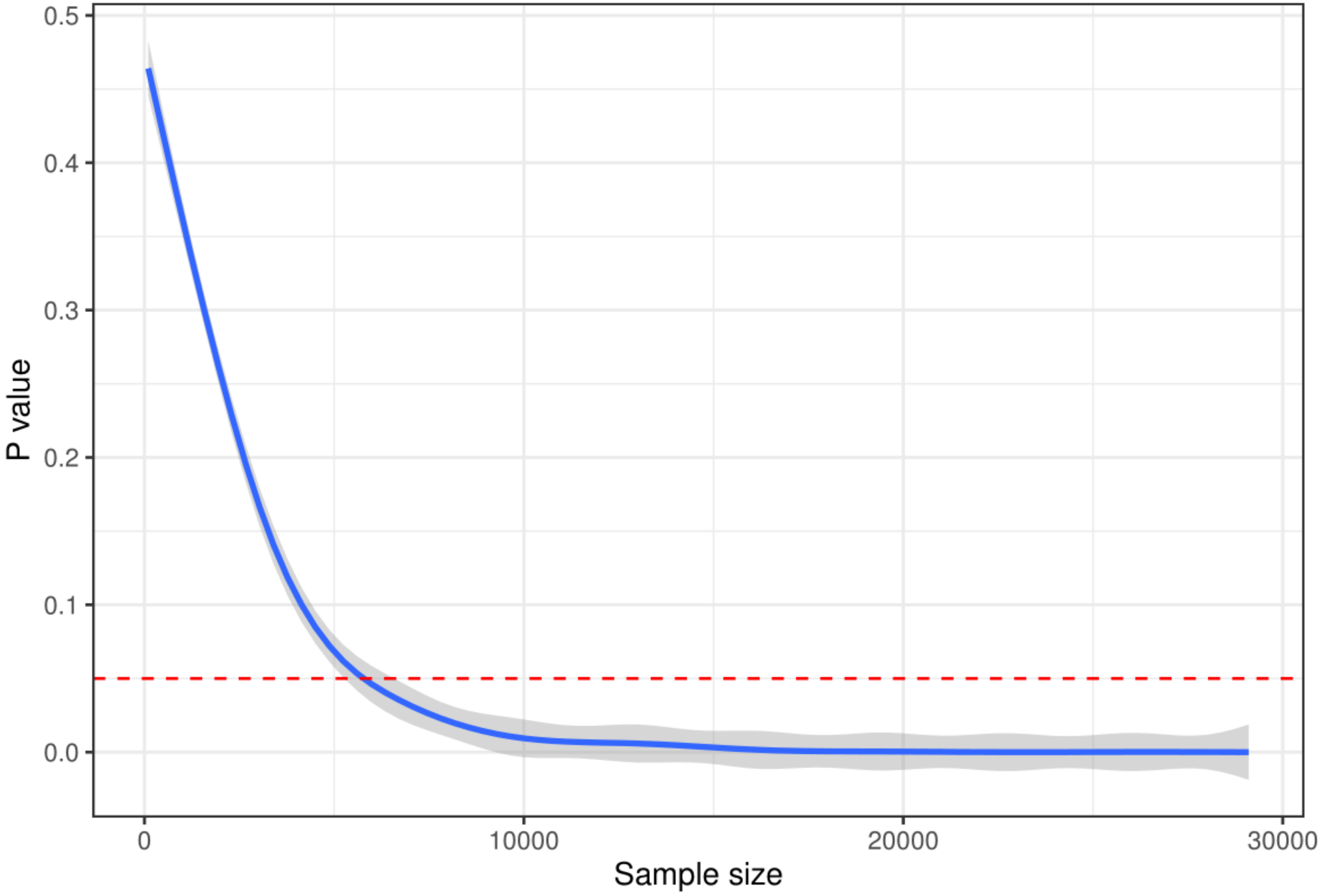
```r
num_trials_per_sample_size<-100
min_sample_size<-100
sample_size_inc<-1000
max_sample_size<-30000
```



```r
set.seed(0) # set randomm seed
crossing(trial=seq(num_trials_per_sample_size),
                        sample_size=seq(min_sample_size,
                                        max_sample_size,sample_size_inc))%>%
  mutate(sample_0=pmap(.l=list(trial,sample_size),
                       .f=~rnorm(n=..2,mean=mu_0,sd=sigma)))%>%
  mutate(sample_1=pmap(.l=list(trial,sample_size),
                       .f=~rnorm(n=..2,mean=mu_1,sd=sigma)))%>%
  mutate(p_value=pmap_dbl(.l=list(sample_0,sample_1),
                          .f=~(t.test(..1,..2,paired=TRUE)$p.value)))%>%
  group_by(sample_size)%>%
  ggplot()+geom_smooth(aes(x=sample_size,y=p_value))+theme_bw()+
  xlab("Sample size")+ylab("P value")+
  geom_hline(aes(yintercept = 0.05),linetype="dashed",color="red")
```

Statistical significance and effect size

Statistical significance is not the same as a meaningful difference between distributions



The necessity of the effect size

Minute differences between populations can yield large test statistics and small p-values, given enough data.

A treatment may cause a statistically significant change, but is this change important?

The **effect size** is a measure for quantifying the magnitude of the observed phenomena.

For paired t-tests we use Cohen’s D statistic to quantify effect size.

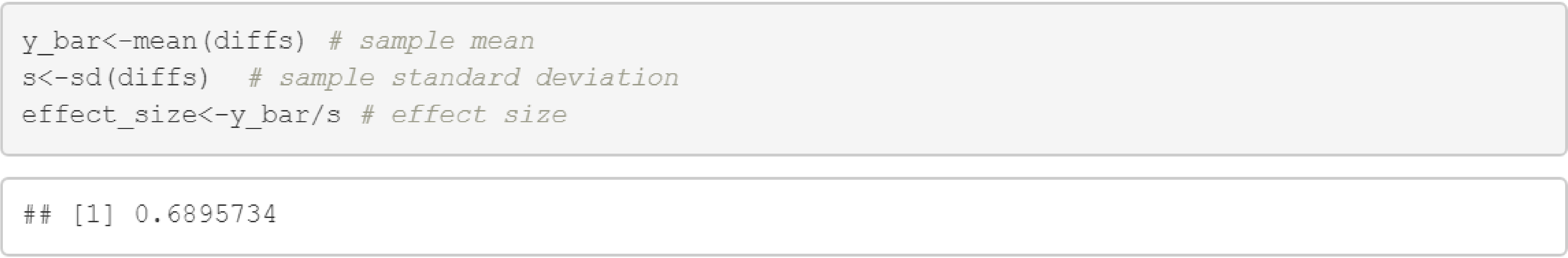
Cohen’s d for the paired t-test

Suppose we carry out a paired t-test on some paired data $X_1^0, \cdots, X_n^0$ and $X_1^1, \cdots, X_n^1$.

Letting $Y_i := X_i^1 - X_i^0$ we want to determine if $Y_1, \cdots, Y_n$ has population mean $\mu \neq 0$.

We can quantify the effect size via Cohen’s d for paired data:

$$d_{\text{paired}} = \frac{\overline{Y}}{S_Y} \quad \text{where} \quad \overline{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{and} \quad (S_Y)^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)$$



```
y_bar<-mean(diffs) # sample mean
s<-sd(diffs)  # sample standard deviation
effect_size<-y_bar/s # effect size
```

```
## [1] 0.6895734
```

Interpreting effect size

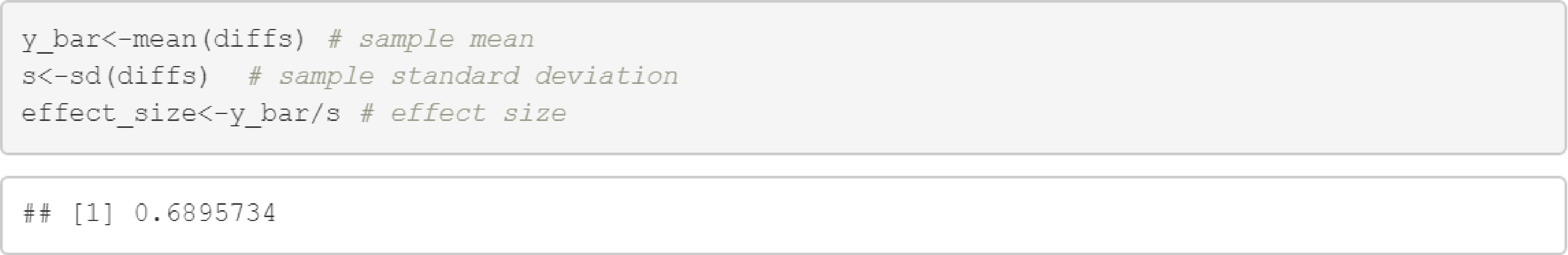
The **effect size** is a measure for quantifying the magnitude of the observed phenomena.

This is interesting when the null hypothesis has been rejected.

Cohen suggested the following guidelines for interpreting effect size with Cohen’s d statistic:

Below 0.2 - small effect,   Around 0.5 – medium effect,   Larger than 0.8 – A large effect.

For the crop yield experiment we have a moderate effect.



```
y_bar<-mean(diffs) # sample mean
s<-sd(diffs)  # sample standard deviation
effect_size<-y_bar/s # effect size
```

```
## [1] 0.6895734
```

Interpreting effect size

Note that effect size is very different to the size of the test.

The **test size** is the probability of Type I error under the null hypothesis:

$$\alpha_{\text{test}} = \mathbb{P}\left(\text{Type I error} \mid \text{H}_0 \text{ is true}\right).$$

The **effect size** is a measure for quantifying the magnitude of the observed phenomena.

<u>Example</u>

For the paired t-test we can use Cohen's d as a measure of effect size.

$$d_{\text{paired}} = \frac{\overline{Y}}{S_Y}$$

Reporting experimental results and effect size

When reporting our results we should include:

1) Our motivating research question;

2) A corresponding statistical question framed in terms of a statistical model;

3) The hypothesis test and its motivation.

4) The numerical value of the test-statistic.

5) The p-value (computed based on the value of the test-statistic).

6) The effect size (this is interesting if we rejected the null and established an effect).

A common misunderstanding about p-values

People often make the following mistake:

“The p-value is the probability that the null hypothesis is true.”

A common misunderstanding about p-values

People often make the following mistake:

~~“The p-value is the probability that the null hypothesis is true.”~~

This is incorrect!

The null hypothesis is a statement about population parameters – not random variables.

Within the classical interpretation of probability such hypotheses either hold or don’t.

What’s random is the data and the test statistics computed based on the data.

The **p-value** is the probability under the null hypothesis that the test statistic takes a value as extreme or more extreme than the observed value.

What have we covered?

- We introduced the concept of statistical hypothesis testing for differences between paired samples.
- We introduced the paired t-test which applies when the differences are approximately Gaussian.
- We don’t need to be so concerned about departures from Gaussian behavior for large sample sizes.
- We also considered the concept of effect size for assessing the magnitude of differences.
- We also discussed a common misunderstanding regarding p-values.

Thanks for listening!

Henry W J Reeve

henry.reeve@bristol.ac.uk

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.