Topic 10: Hypothesis testing

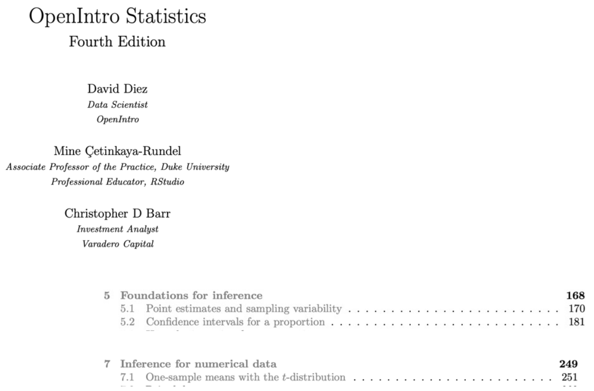
1

Learning Outcomes

1. Carry out hypothesis tests for a single mean.
2. Use the *p*-value approach for making decisions in hypothesis tests.
3. Understand types of testing errors
4. Understand the relationship between hypothesis testing and confidence intervals and the advantages of interval estimation
5. Additional reading material : Open Intro book Chapters 5.1 & 7.1

2

2



OpenIntro Statistics

Fourth Edition

David Diez
Data Scientist
OpenIntro

Mine Çetinkaya-Rundel
Associate Professor of the Practice, Duke University
Professional Educator, RStudio

Christopher D Barr
Investment Analyst
Varadero Capital

3

3

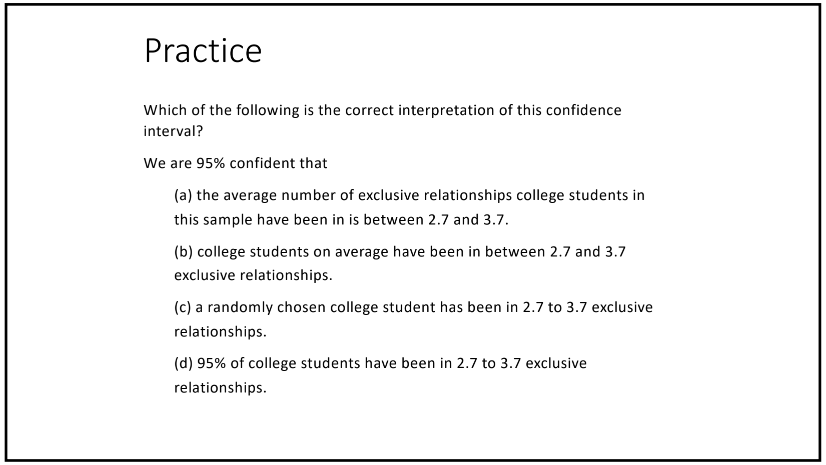
Recap: Inference using Confidence Interval Estimation

A claim has been made that college students have been in, on average, at least 4 exclusive relationships. Data collected on a random sample of 50 college students yielded a mean of 3.2 and a standard deviation of 1.74.

Do these data provide evidence for or against the hypothesis claimed ?

The corresponding 95% CI is [2.7, 3.7].

4



Practice

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

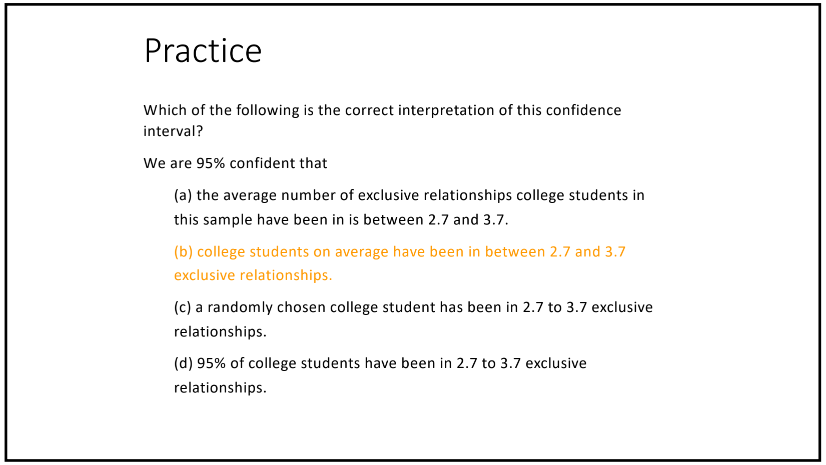
(a) the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

(b) college students on average have been in between 2.7 and 3.7 exclusive relationships.

(c) a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.

(d) 95% of college students have been in 2.7 to 3.7 exclusive relationships.

5



Practice

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

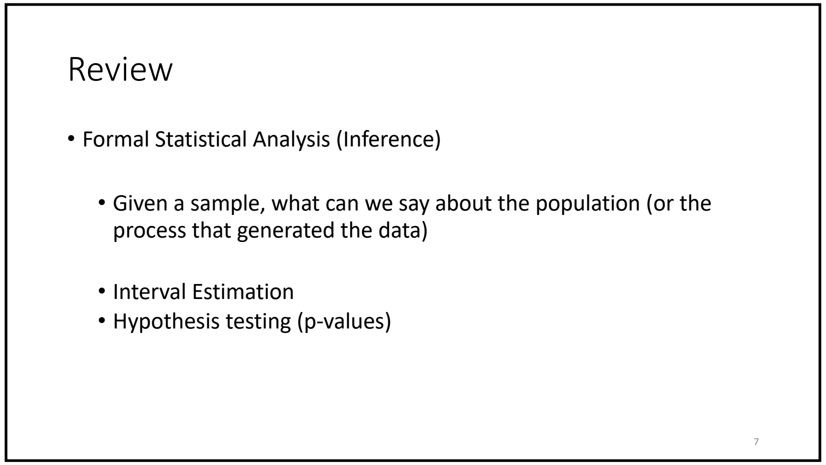
(a) the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

(b) college students on average have been in between 2.7 and 3.7 exclusive relationships.

(c) a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.

(d) 95% of college students have been in 2.7 to 3.7 exclusive relationships.

6

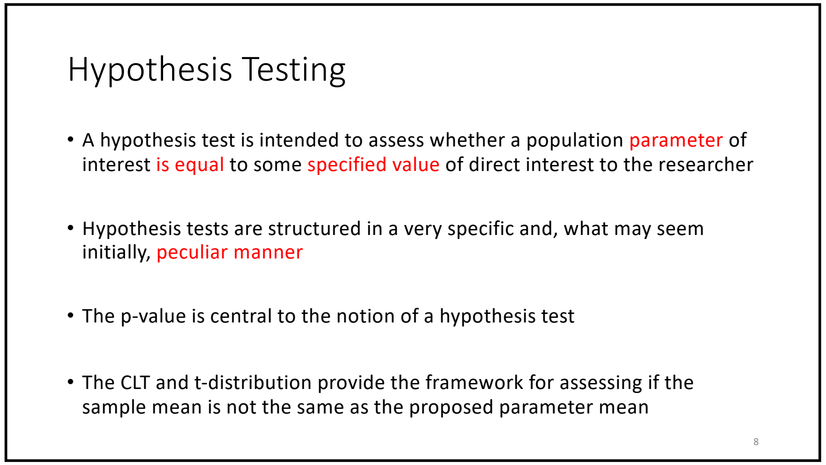


Review

- Formal Statistical Analysis (Inference)
  - Given a sample, what can we say about the population (or the process that generated the data)
  - Interval Estimation
  - Hypothesis testing (p-values)

7

7



Hypothesis Testing

- A hypothesis test is intended to assess whether a population parameter of interest is equal to some specified value of direct interest to the researcher
- Hypothesis tests are structured in a very specific and, what may seem initially, peculiar manner
- The p-value is central to the notion of a hypothesis test
- The CLT and t-distribution provide the framework for assessing if the sample mean is not the same as the proposed parameter mean

8

8

Null and alternative hypotheses

- The null hypothesis is a claim to be tested – often the skeptical claim of “no effect”.. eg

$$H_0\colon\ \mu = \mu_0$$

- The alternative hypothesis is an alternative claim under consideration, often represented by a range of parameter values – eg

$$H_1\colon\ \mu \neq \mu_0$$

- We only reject the null in favour of the alternative if there is strong supporting evidence.
- We decide a priori how much evidence is “strong” enough to reject the null

9

9

Stages in Hypothesis Testing

1. Null Hypothesis: The hypothesis that the population parameter is equal to some claimed value (H_0)
2. Study or Alternative Hypothesis: The hypothesis that must be true if the null hypothesis is false (H_1)
3. **Collect appropriate data**
4. Assess, through a test statistic, how probable (the p-value) it would be to observe data as or more extreme than the data actually collected if, in fact, the Null Hypothesis was true
5. Come to a conclusion whether or not to reject the Null Hypothesis

10

10

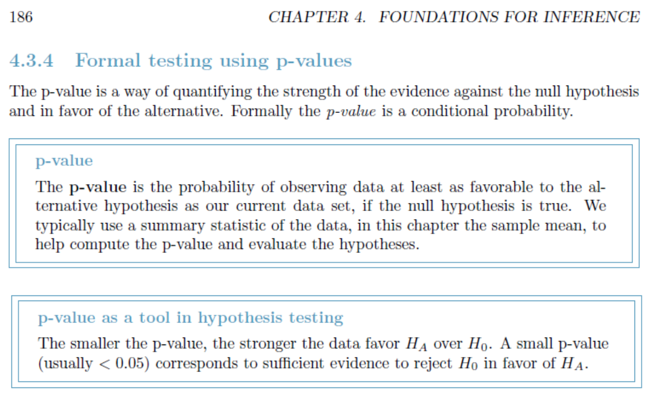
Rejecting/not rejecting the null

- If we do not reject the null hypothesis in favour of the alternative, we are saying that the effect indicated by the sample is due only to sampling variation.

- If we do reject the null hypothesis in favour of the alternative, we are saying that the effect indicated by the sample is real, in that it is more than can be attributed to sampling variation.

11

11



186 *CHAPTER 4. FOUNDATIONS FOR INFERENCE*

4.3.4 Formal testing using p-values

The p-value is a way of quantifying the strength of the evidence against the null hypothesis and in favor of the alternative. Formally the *p-value* is a conditional probability.

p-value

The **p-value** is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis is true. We typically use a summary statistic of the data, in this chapter the sample mean, to help compute the p-value and evaluate the hypotheses.

p-value as a tool in hypothesis testing

The smaller the p-value, the stronger the data favor H_A over H_0. A small p-value (usually < 0.05) corresponds to sufficient evidence to reject H_0 in favor of H_A.

12

12

One-Sample Tests for the population mean

1. Specify the hypotheses about μ
2. Calculate a test statistic – based on the sampling distribution of the sample mean
3. See how extreme the test statistic is if the null hypothesis was true – compare the test statistic with the t or Normal distribution
4. Make a decision: reject the null or don't reject it.

Strategy

- If the sample came from the population in question the sample mean should be 'close' to the population mean in question
- 'Close' needs to take into account the sample size used and the variability in the measure (i.e. the standard error)
- For testing means, the Central Limit Theorem or t distribution (or the bootstrap) is key

Tests on the Mean of a Normal Distribution, Variance Unknown

One-Sample *t*-Test

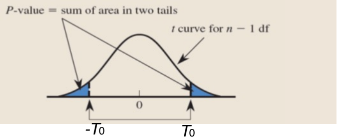
Null Hypothesis

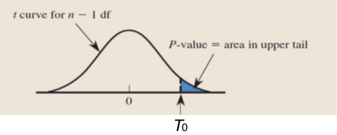
H_0: $\mu = \mu_0$

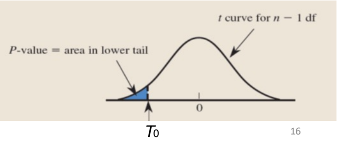
Test statistic: $$T_0 = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

	Alternative hypothesis	Rejection criteria
Two sided hypotheses test →	H_1: $\mu \neq \mu_0$	$T_0 > t_{\alpha/2,n\text{-}1}$ or $T_0 < -t_{\alpha/2,n-1}$
One sided hypotheses tests	H_1: $\mu > \mu_0$	$T_0 > t_{\alpha,n-1}$
	H_1: $\mu < \mu_0$	$T_0 < -t_{\alpha,n-1}$

Alternative hypothesis	Rejection criteria
H_1: $\mu \neq \mu_0$	$T_0 > t_{\alpha/2,n\text{-}1}$ or $T_0 < -t_{\alpha/2,n-1}$
H_1: $\mu > \mu_0$	$T_0 > t_{\alpha,n-1}$
H_1: $\mu < \mu_0$	$T_0 < -t_{\alpha,n-1}$







Typically α is set at 0.05

Terms and conditions:

- Independence: random sample/assignment
- Normality: for small samples where we use the t distribution, we require the observations to be approximately normally distributed. For larger (*n* ≥ 30) samples, no extreme skew we can use the CLT and do not require the observations to be normally distributed.

17

p-values and (α) significance levels …

- A p-value ≤ 0.05 is (typically) considered as sufficient evidence against a null hypothesis (ie sufficient evidence to reject the null).
- If the p-value for the test of a parameter with 2-sided alternative is <0.05, the 95% Confidence Interval will not include the parameter.

18

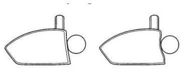
Statistical Significance

- Whenever the p-value is less than a particular threshold, the result is said to be "statistically significant" at that level.
- The threshold should be decided a priori, before you calculate the test statistic
- For example, if the threshold is p ≤ 0.05, the result is statistically significant at the 5% level; if p ≤ 0.01, the result is statistically significant at the 1% level, and so on.
- If a result is statistically significant at the 100α% level, we can also say that the null hypothesis is "rejected at level 100α%."

19

Example: Golf Club Design

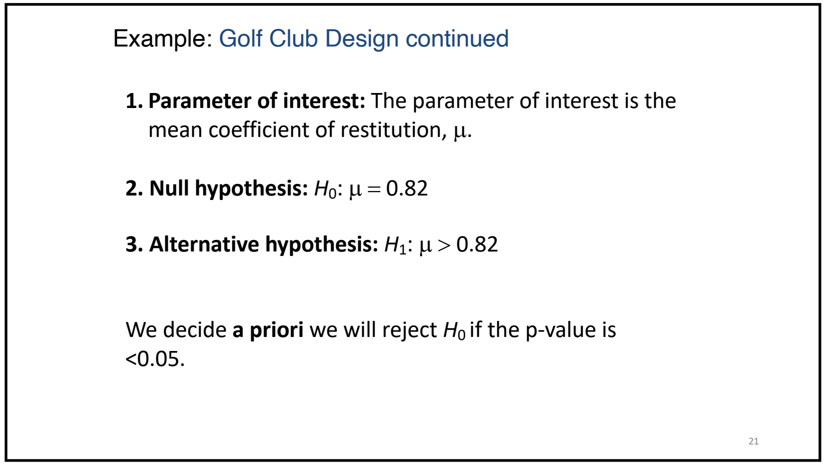
An experiment was performed in which 15 drivers produced by a particular club maker were selected at random and their coefficients of restitution measured. It is of interest to determine if there is evidence (with $\alpha = 0.05$ significance level) to support a claim that the mean coefficient of restitution *exceeds* 0.82.



The sample mean and sample standard deviation are $\bar{x} = 0.83725$ and $s = 0.02456$.

The objective of the experimenter is to demonstrate that the mean coefficient of restitution exceeds 0.82, hence a one-sided alternative hypothesis is appropriate.

20



Example: Golf Club Design continued

1. Parameter of interest: The parameter of interest is the mean coefficient of restitution, μ.

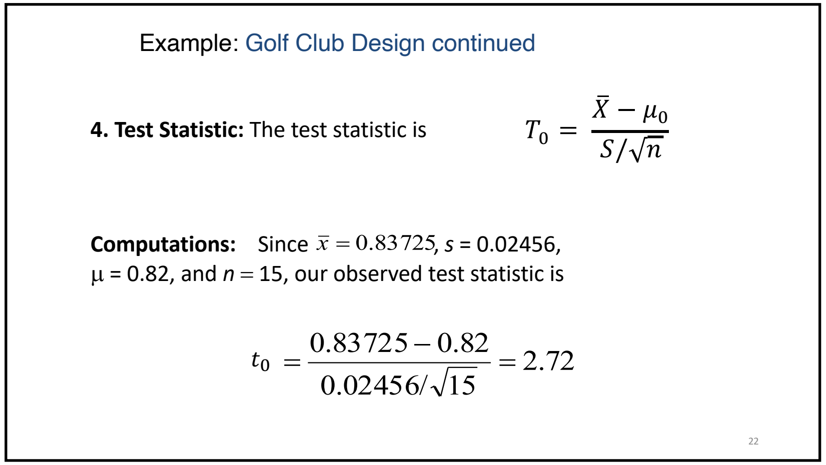
2. Null hypothesis: H_0: $\mu = 0.82$

3. Alternative hypothesis: H_1: $\mu > 0.82$

We decide **a priori** we will reject H_0 if the p-value is <0.05.

21

21



Example: Golf Club Design continued

4. Test Statistic: The test statistic is

$$T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

Computations: Since $\bar{x} = 0.83725$, $s = 0.02456$, $\mu = 0.82$, and $n = 15$, our observed test statistic is

$$t_0 = \frac{0.83725 - 0.82}{0.02456/\sqrt{15}} = 2.72$$

22

22

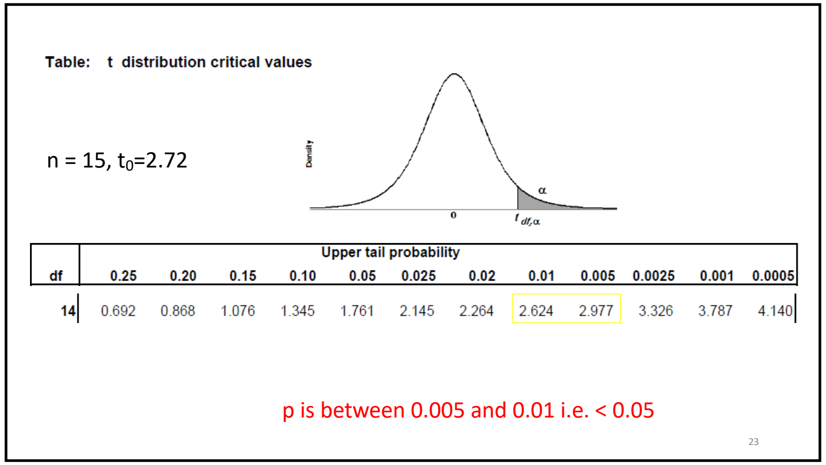


Table: t distribution critical values

n = 15, t_0=2.72

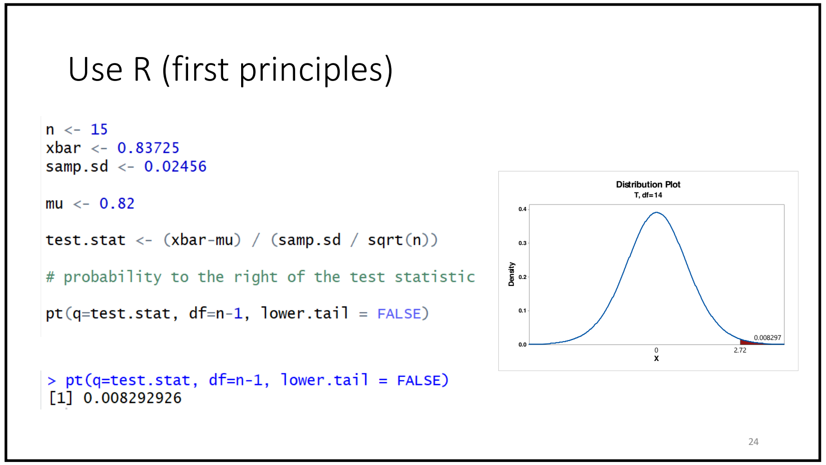
Density
α
0
$t_{df,\alpha}$

df	Upper tail probability 0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140

p is between 0.005 and 0.01 i.e. < 0.05

23

23



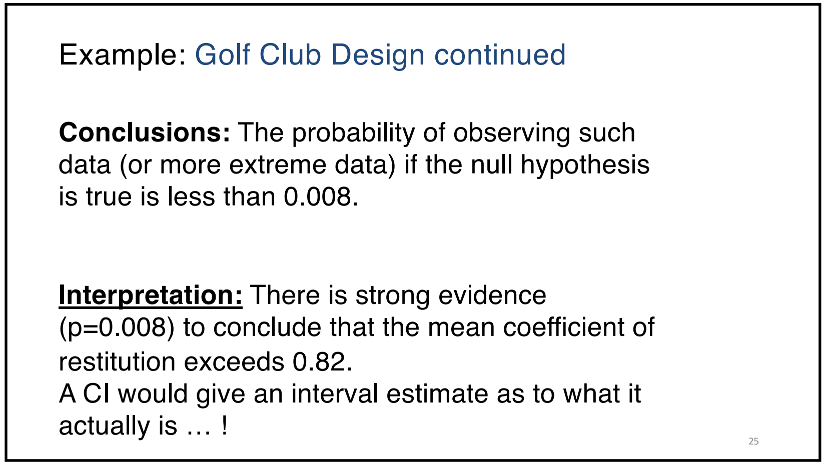
Use R (first principles)

```
n <- 15
xbar <- 0.83725
samp.sd <- 0.02456

mu <- 0.82

test.stat <- (xbar-mu) / (samp.sd / sqrt(n))

# probability to the right of the test statistic

pt(q=test.stat, df=n-1, lower.tail = FALSE)
```

```
> pt(q=test.stat, df=n-1, lower.tail = FALSE)
[1] 0.008292926
```

24

24



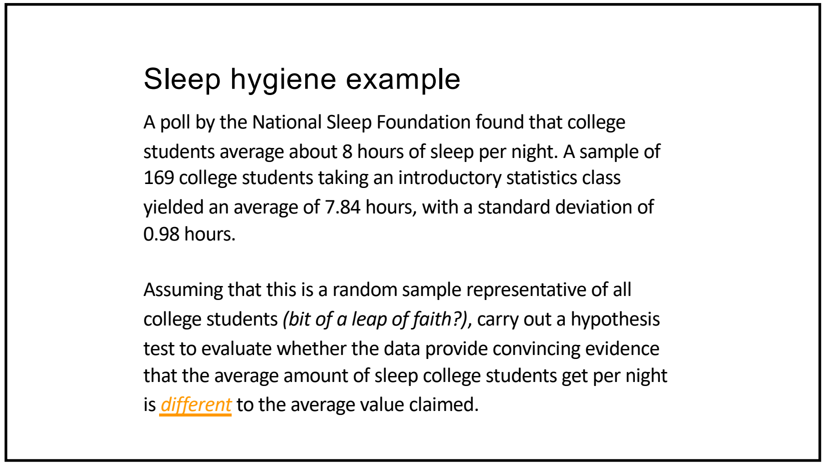
Example: Golf Club Design continued

Conclusions: The probability of observing such data (or more extreme data) if the null hypothesis is true is less than 0.008.

Interpretation: There is strong evidence (p=0.008) to conclude that the mean coefficient of restitution exceeds 0.82.
A CI would give an interval estimate as to what it actually is … !

25

25

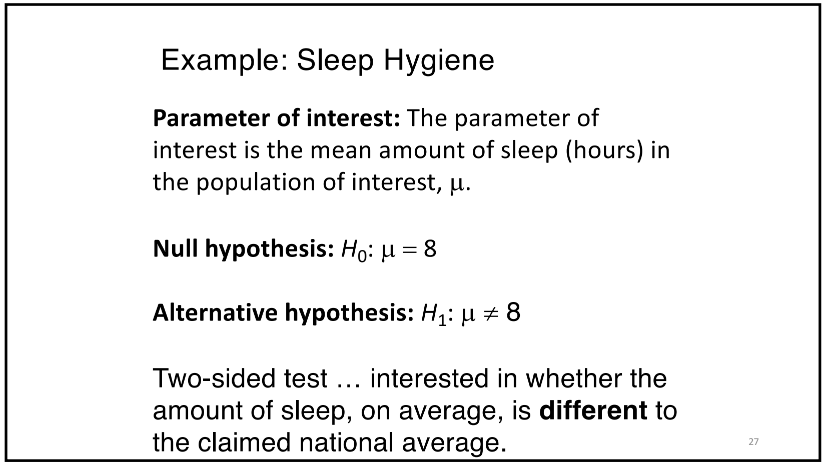


Sleep hygiene example

A poll by the National Sleep Foundation found that college students average about 8 hours of sleep per night. A sample of 169 college students taking an introductory statistics class yielded an average of 7.84 hours, with a standard deviation of 0.98 hours.

Assuming that this is a random sample representative of all college students *(bit of a leap of faith?)*, carry out a hypothesis test to evaluate whether the data provide convincing evidence that the average amount of sleep college students get per night is *different* to the average value claimed.

26



Example: Sleep Hygiene

Parameter of interest: The parameter of interest is the mean amount of sleep (hours) in the population of interest, μ.

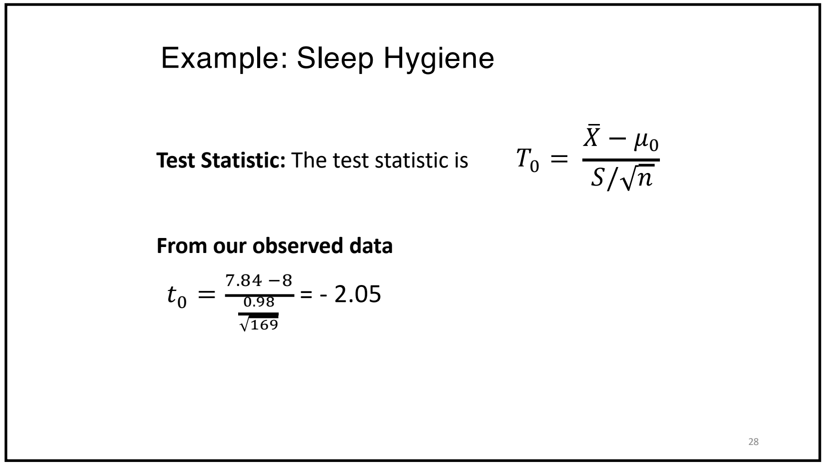
Null hypothesis: H_0: $\mu = 8$

Alternative hypothesis: H_1: $\mu \neq 8$

Two-sided test … interested in whether the amount of sleep, on average, is **different** to the claimed national average.

27

27



Example: Sleep Hygiene

Test Statistic: The test statistic is $T_0 = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$

From our observed data

$t_0 = \dfrac{7.84 - 8}{\frac{0.98}{\sqrt{169}}}$ = - 2.05

28

28

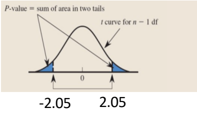
Example: Sleep Hygiene

Test Statistic: The test statistic is $T_0 = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$

From our observed data

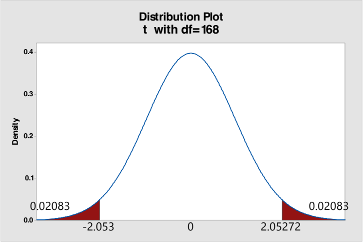
$$t_0 = \frac{7.84 - 8}{\frac{0.98}{\sqrt{169}}} = -\,2.05$$

p-value : calculate area to the right of 2.05 and to the left of -2.05 in a t distribution with 169-1 degrees of freedom.



Example: Sleep Hygiene

P-value : calculate area to the right of 2.05 and to the left of -2.05 in a t distribution with 169-1 degrees of freedom.



Example: Sleep Hygiene

p-value : use the symmetry in the distribution i.e. calculate area to the right of 2.05 in a t distribution with 169-1 degrees of freedom and double it.

```
> 2 * pt(q= 2.052717, df=168, lower.tail = FALSE)
[1] 0.04165098
```

p-value : 0.0416

Example: Sleep Hygiene

Conclusions: The probability of observing such data (or more extreme) if the null hypothesis is true is = 0.04.

Interpretation: As the p-value is less than 0.05, there is evidence (at the 5% significance level) that the mean hours sleeping is different from the national average of 8.

Example: Sleep Hygiene using R

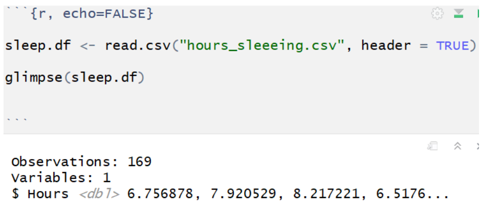
Parameter of interest: The parameter of interest is the mean amount of sleep (hours) in the population of interest, μ.

Null hypothesis: H_0: $\mu = 8$

Alternative hypothesis: H_1: $\mu \neq 8$

Two-sided test … interested in whether the amount of sleep, on average, is different to claimed national average.

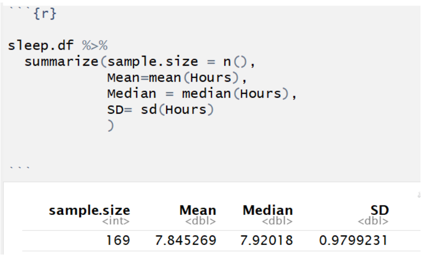
33

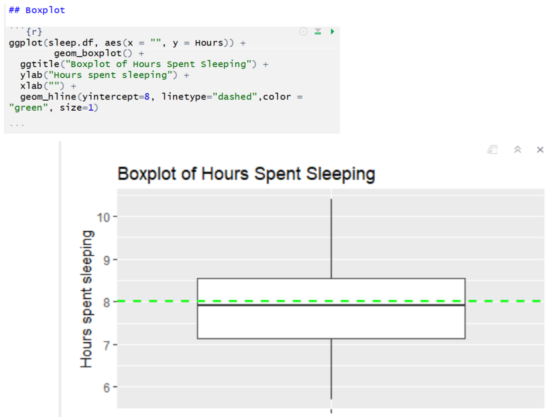


~~~
```{r, echo=FALSE}
sleep.df <- read.csv("hours_sleeeing.csv", header = TRUE)
glimpse(sleep.df)

```
~~~

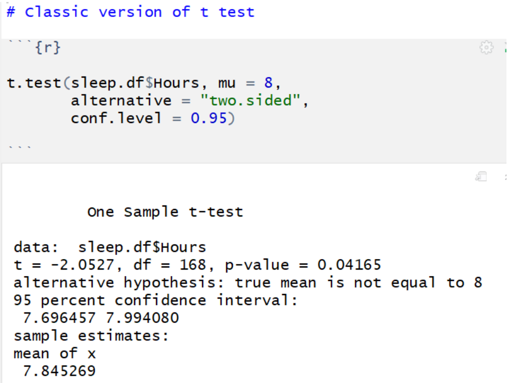
```
Observations: 169
Variables: 1
$ Hours <dbl> 6.756878, 7.920529, 8.217221, 6.5176...
```

34



~~~
```{r}
sleep.df %>%
  summarize(sample.size = n(),
            Mean=mean(Hours),
            Median = median(Hours),
            SD= sd(Hours)
            )

~~~
```

sample.size <int>	Mean <dbl>	Median <dbl>	SD <dbl>
169	7.845269	7.92018	0.9799231

35



```
## Boxplot

```{r}
ggplot(sleep.df, aes(x = "", y = Hours)) +
        geom_boxplot() +
  ggtitle("Boxplot of Hours Spent Sleeping") +
  ylab("Hours spent sleeping") +
  xlab("") +
  geom_hline(yintercept=8, linetype="dashed",color =
"green", size=1)
```
```

36
```



~~~
# Classic version of t test
```{r}

t.test(sleep.df$Hours, mu = 8,
       alternative = "two.sided",
       conf.level = 0.95)

```

        One Sample t-test

data: sleep.df$Hours
t = -2.0527, df = 168, p-value = 0.04165
alternative hypothesis: true mean is not equal to 8
95 percent confidence interval:
 7.696457 7.994080
sample estimates:
mean of x
 7.845269
~~~

Statistical Significance Is Not the Same as Practical Significance

- When a result has a small p-value, we say that it is “statistically significant.” In common usage, the word significant means “important.” It is therefore tempting to think that statistically significant results must always be important.
- This is not the case. Sometimes statistically significant results do not have any scientific or practical importance.
- A difference is only a difference if it makes a difference.

Statistical Significance Is Not the Same as Practical Significance continued …

- The p-value does not measure practical significance. What it does measure is the degree of confidence we can have that the true value is really different from the value specified by the null hypothesis.
- When the p-value is small, then we can be confident that the true value is really different. This does not necessarily imply that the difference is large enough to be of practical importance.

Connection between Hypothesis Tests and Confidence Intervals

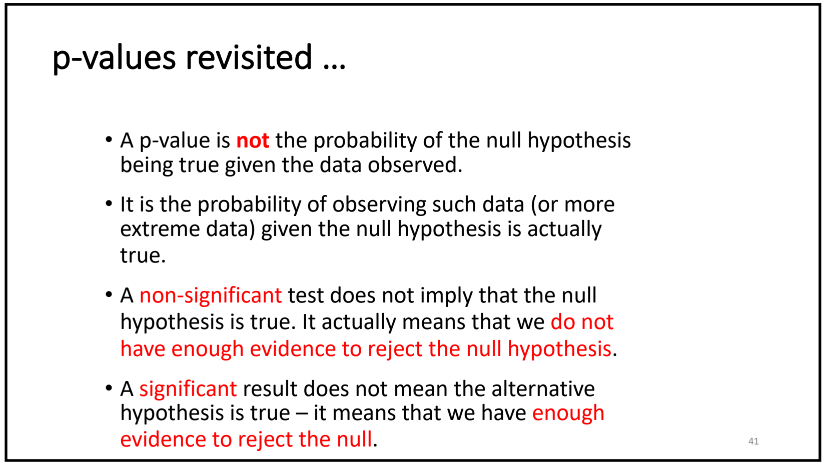
A close relationship exists between the test of a hypothesis for θ, and the confidence interval for θ.

If $[l, u]$ is a 95% confidence interval for the parameter θ, the test of the null hypothesis against a 2-sided alternative at the 0.05 significance level

$$H_0: \theta = \theta_0$$
$$H_1: \theta \neq \theta_0$$

will lead to rejection of H_0 if and only if θ_0 is **not** in the 95% CI $[l, u]$.

And similarly for your alpha of choice e.g. 90% CI and p < 0.10 ….

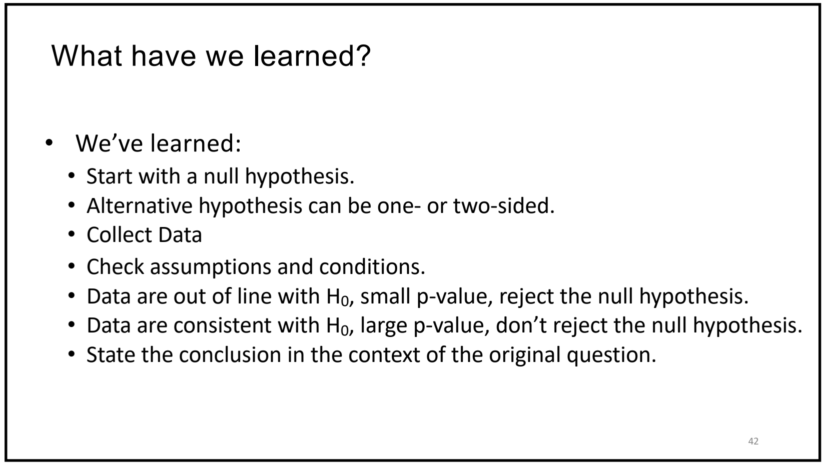


p-values revisited …

- A p-value is **not** the probability of the null hypothesis being true given the data observed.
- It is the probability of observing such data (or more extreme data) given the null hypothesis is actually true.
- A non-significant test does not imply that the null hypothesis is true. It actually means that we do not have enough evidence to reject the null hypothesis.
- A significant result does not mean the alternative hypothesis is true – it means that we have enough evidence to reject the null.

41

41

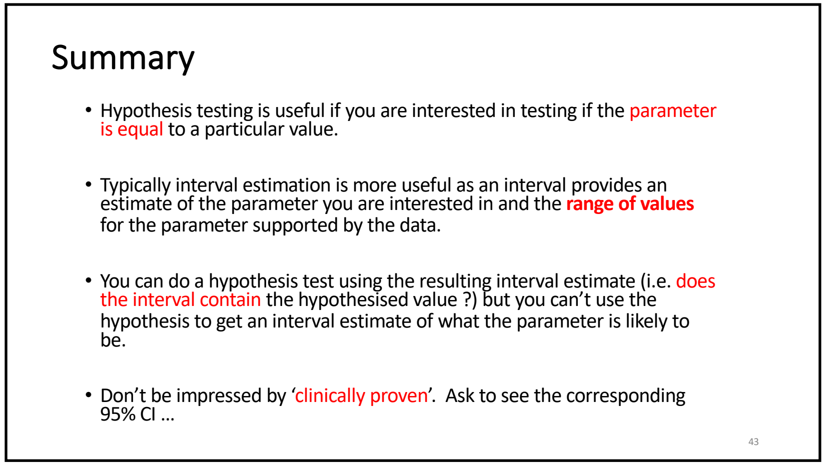


What have we learned?

- We've learned:
  - Start with a null hypothesis.
  - Alternative hypothesis can be one- or two-sided.
  - Collect Data
  - Check assumptions and conditions.
  - Data are out of line with H_0, small p-value, reject the null hypothesis.
  - Data are consistent with H_0, large p-value, don't reject the null hypothesis.
  - State the conclusion in the context of the original question.

42

42

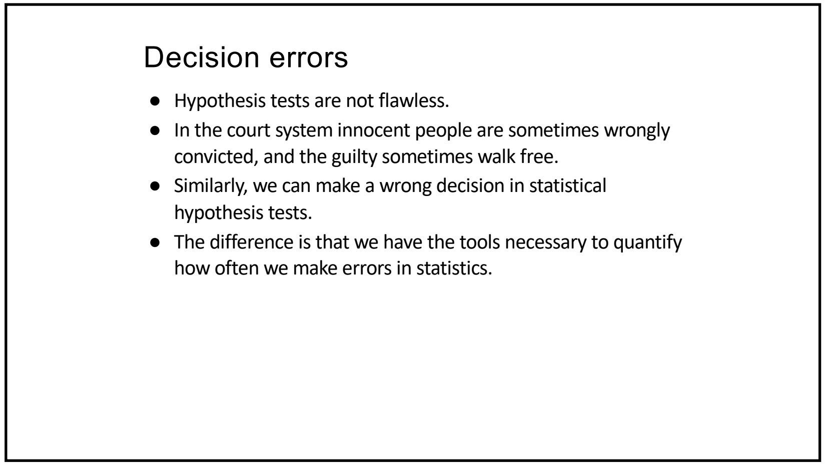


Summary

- Hypothesis testing is useful if you are interested in testing if the parameter is equal to a particular value.
- Typically interval estimation is more useful as an interval provides an estimate of the parameter you are interested in and the **range of values** for the parameter supported by the data.
- You can do a hypothesis test using the resulting interval estimate (i.e. does the interval contain the hypothesised value ?) but you can't use the hypothesis to get an interval estimate of what the parameter is likely to be.
- Don't be impressed by 'clinically proven'. Ask to see the corresponding 95% CI …

43

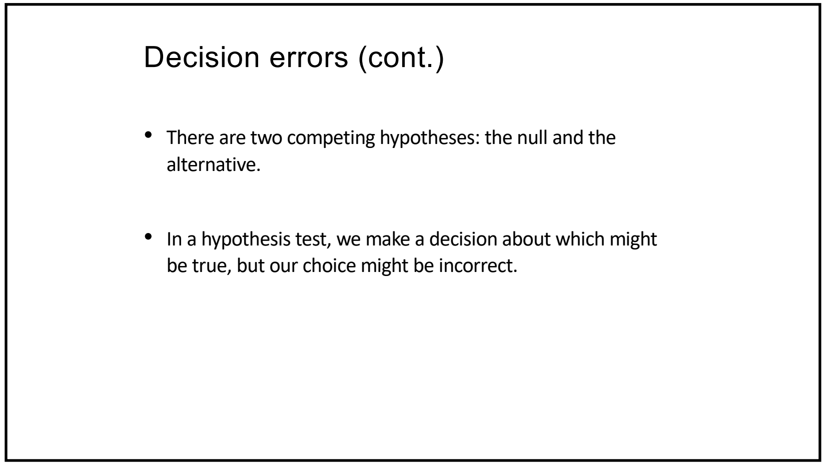
43



Decision errors

- Hypothesis tests are not flawless.
- In the court system innocent people are sometimes wrongly convicted, and the guilty sometimes walk free.
- Similarly, we can make a wrong decision in statistical hypothesis tests.
- The difference is that we have the tools necessary to quantify how often we make errors in statistics.

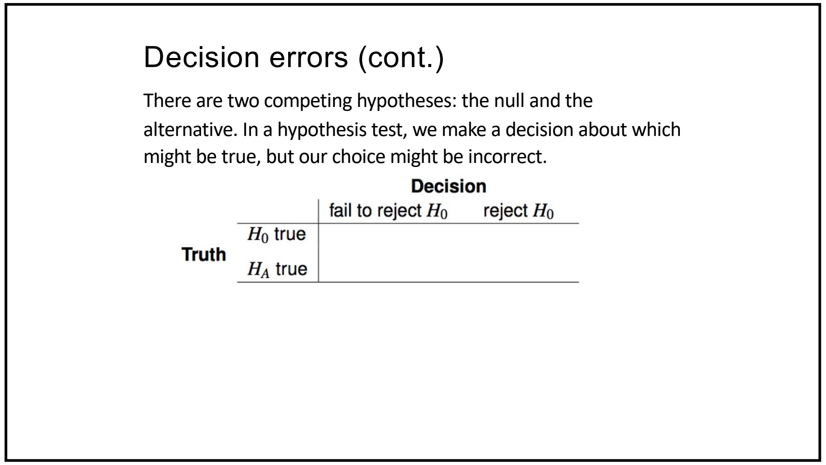
44



Decision errors (cont.)

- There are two competing hypotheses: the null and the alternative.
- In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

45

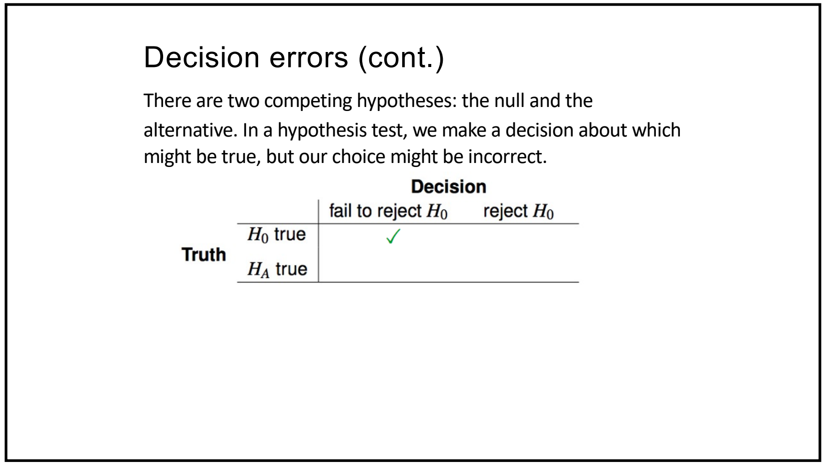


Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		
	H_A true		

46

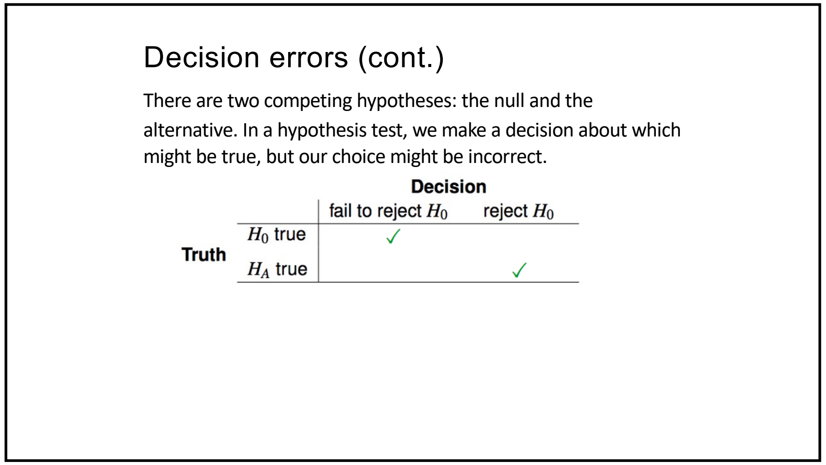


Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	
	H_A true		

47

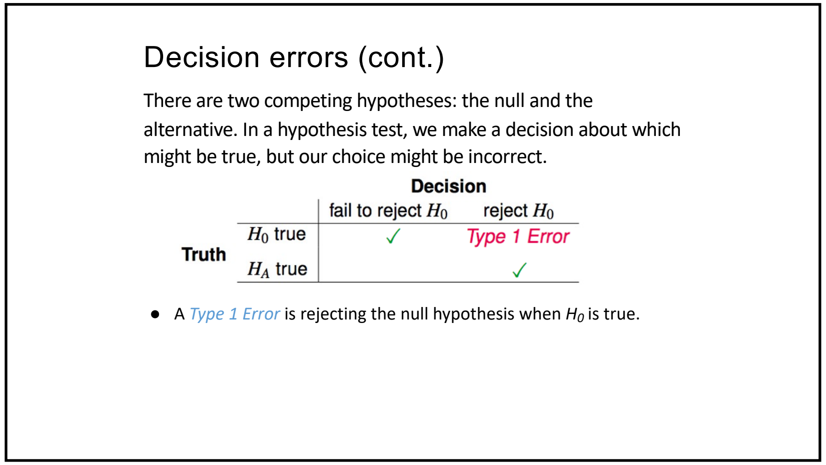


Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	
	H_A true		✓

48



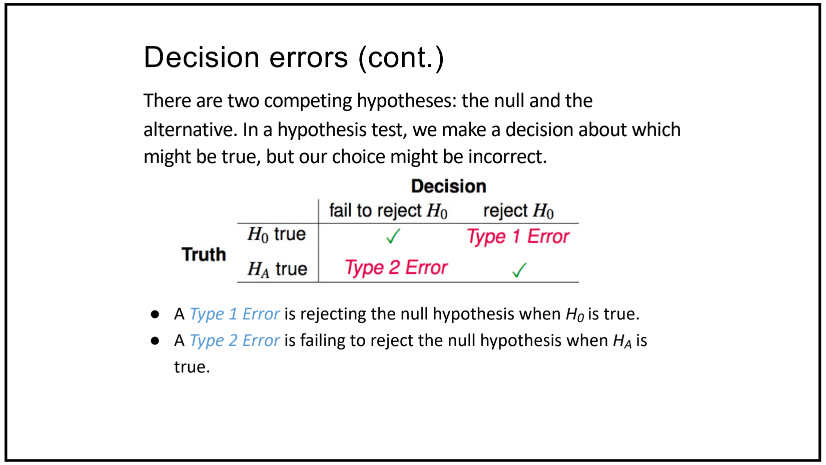
Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true		✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.

49



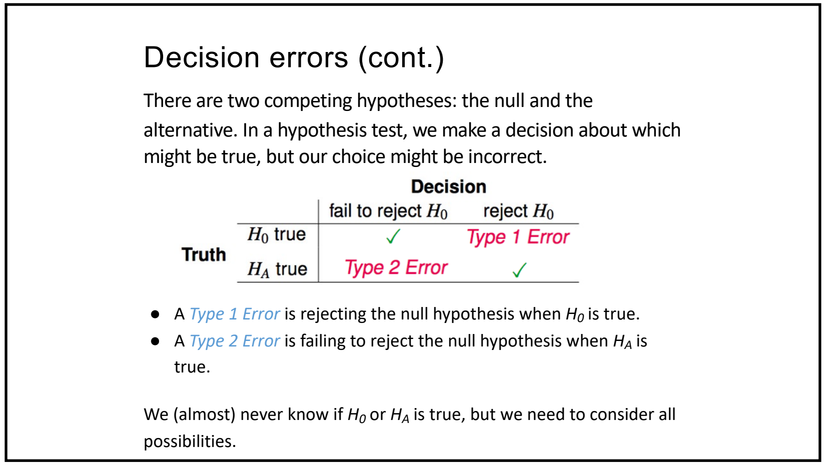
Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.
- A *Type 2 Error* is failing to reject the null hypothesis when H_A is true.

50



Decision errors (cont.)

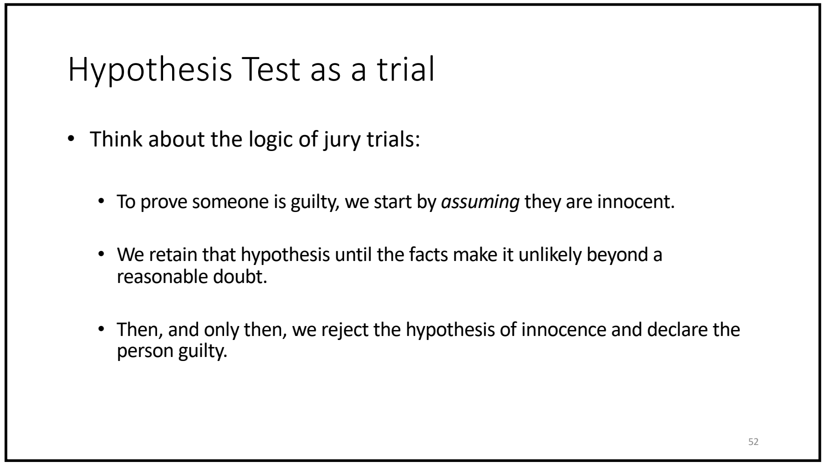
There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.
- A *Type 2 Error* is failing to reject the null hypothesis when H_A is true.

We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

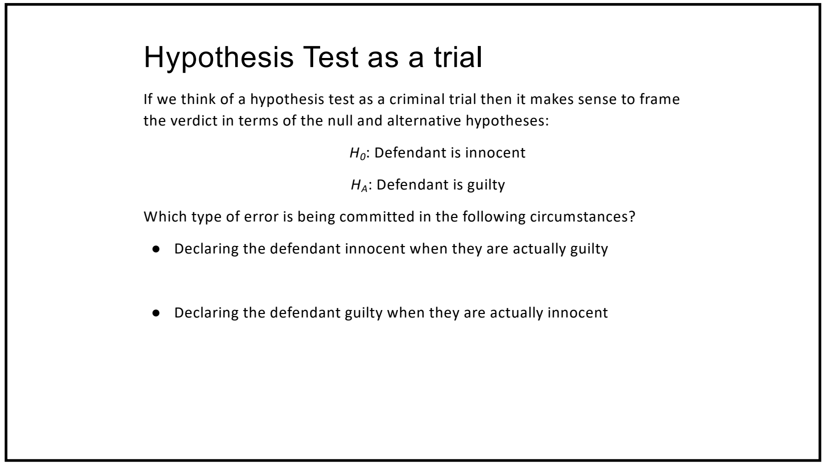
51



Hypothesis Test as a trial

- Think about the logic of jury trials:
  - To prove someone is guilty, we start by *assuming* they are innocent.
  - We retain that hypothesis until the facts make it unlikely beyond a reasonable doubt.
  - Then, and only then, we reject the hypothesis of innocence and declare the person guilty.

52



Hypothesis Test as a trial

If we think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

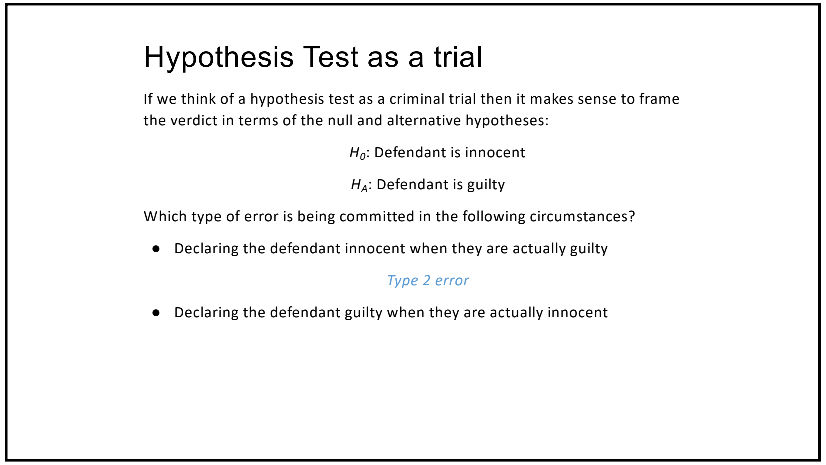
H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty
- Declaring the defendant guilty when they are actually innocent

53



Hypothesis Test as a trial

If we think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

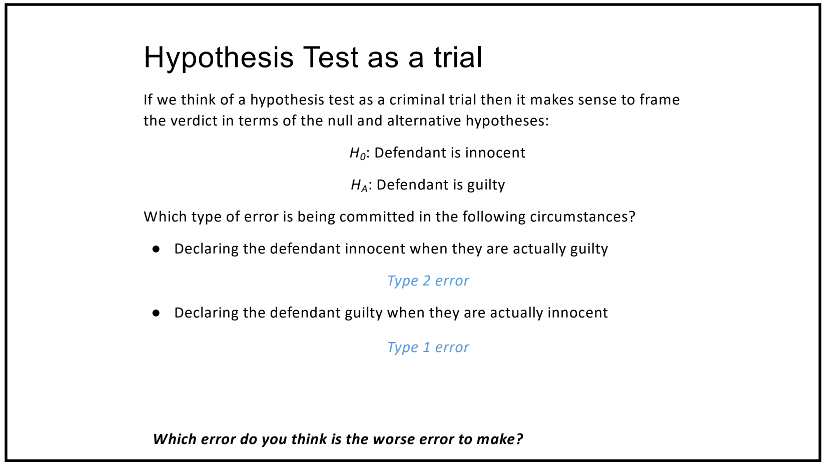
Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

Type 2 error

- Declaring the defendant guilty when they are actually innocent

54



Hypothesis Test as a trial

If we think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

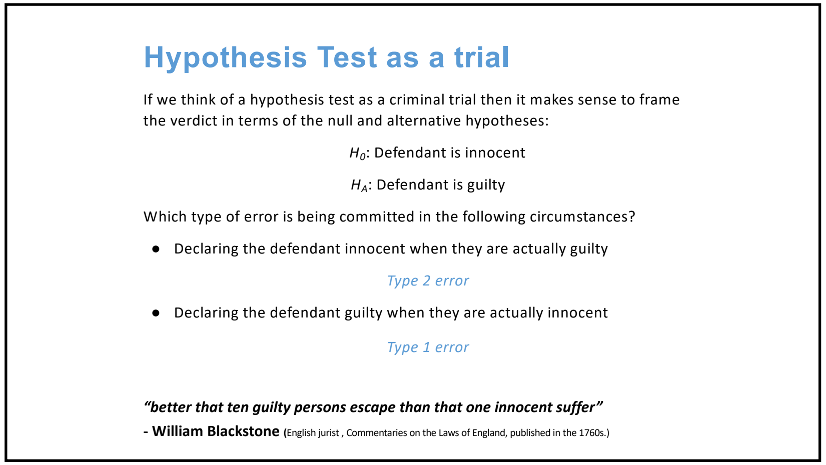
Type 2 error

- Declaring the defendant guilty when they are actually innocent

Type 1 error

Which error do you think is the worse error to make?

55



Hypothesis Test as a trial

If we think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

Type 2 error

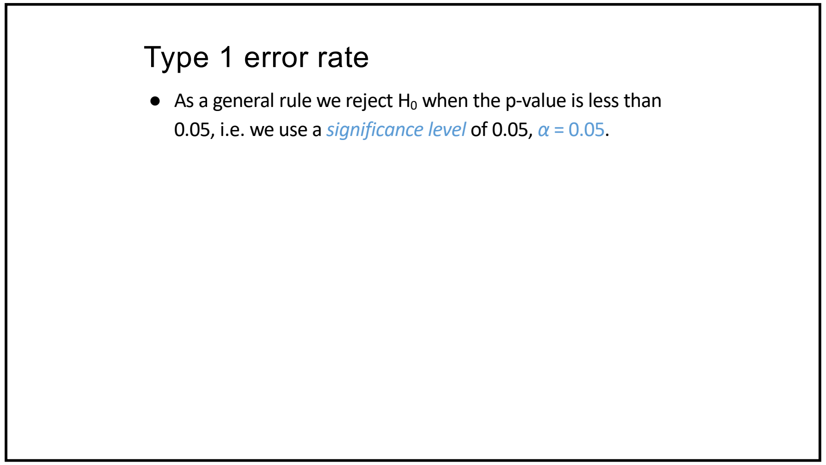
- Declaring the defendant guilty when they are actually innocent

Type 1 error

"better that ten guilty persons escape than that one innocent suffer"

- William Blackstone (English jurist , Commentaries on the Laws of England, published in the 1760s.)

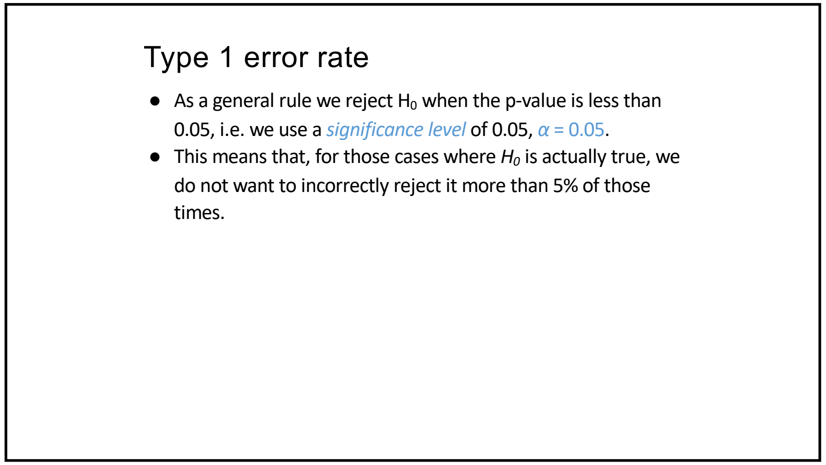
56



Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.

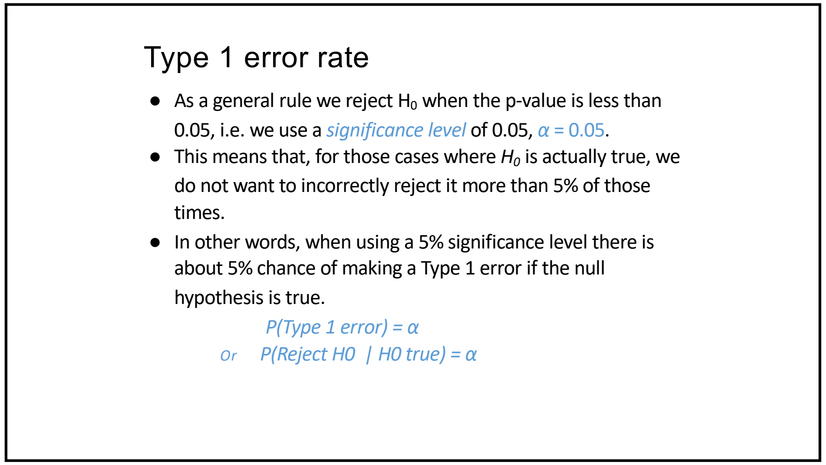
57



Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.

58



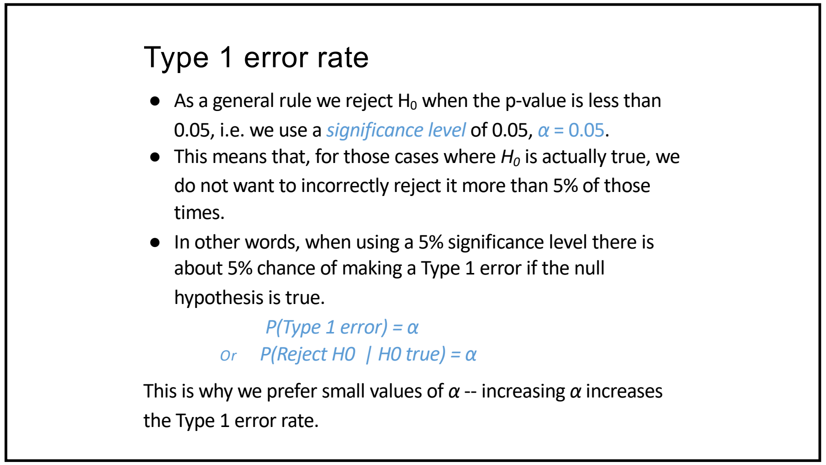
Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error}) = \alpha$$

Or $$P(\text{Reject } H0 \mid H0 \text{ true}) = \alpha$$

59



Type 1 error rate

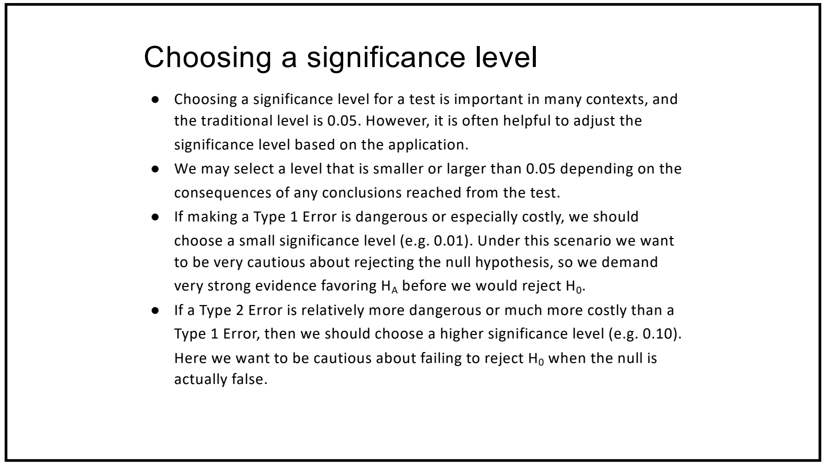
- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error}) = \alpha$$

Or $$P(\text{Reject } H0 \mid H0 \text{ true}) = \alpha$$

This is why we prefer small values of α -- increasing α increases the Type 1 error rate.

60



Choosing a significance level

- Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05. However, it is often helpful to adjust the significance level based on the application.
- We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.
- If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0.
- If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false.

61