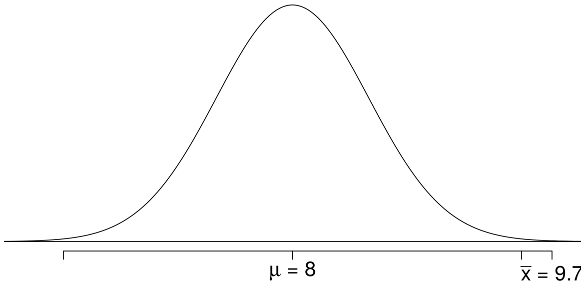
Lecture 17

Formal testing using p-values

Test statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the test statistic.



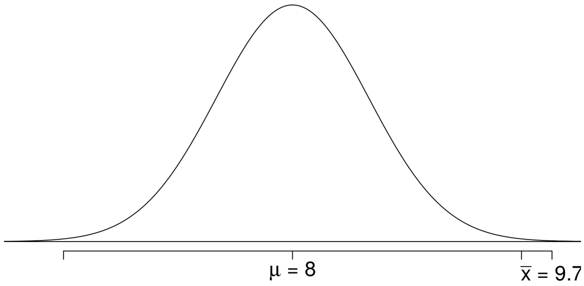
$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} \approx 0.5\right)$$

$$Z = \frac{9.7 - 8}{0.5} = 3.4$$

The sample mean is 3.4 standard errors away from the hypothesized value. Is this considered unusually high? That is, is the result **statistically significant**?

Test statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the test statistic.



$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} \approx 0.5\right)$$

$$Z = \frac{9.7 - 8}{0.5} = 3.4$$

The sample mean is 3.4 standard errors away from the hypothesized value. Is this considered unusually high? That is, is the result **statistically significant**?

Yes, and we can quantify how unusual it is using a p-value.

p-values

- We then use this test statistic to calculate the **p-value**, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.

p-values

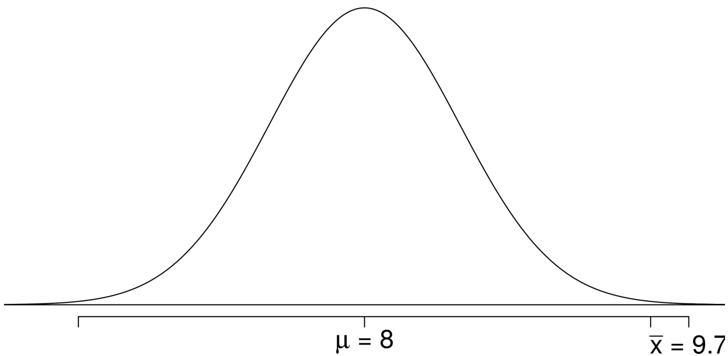
- We then use this test statistic to calculate the **p-value**, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is **low** (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence **reject** H_0.

p-values

- We then use this test statistic to calculate the **p-value**, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is **low** (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence **reject** H_0.
- If the p-value is **high** (higher than α) we say that it is likely to observe the data even if the null hypothesis were true, and hence **do not reject** H_0.

Number of university applications - p-value

p-value: probability of observing data at least as favorable to H_A as our current data set (a sample mean greater than 9.7), if in fact H_0 were true (the true population mean was 8).



$$P(\bar{x} > 9.7 \mid \mu = 8) = P(Z > 3.4) = 0.0003$$

Number of university applications - Making a decision

- p-value = 0.0003

Number of university applications - Making a decision

- p-value = 0.0003
  - If the true average of the number of universities Trent students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Trent students who on average apply to 9.7 or more schools.
  - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.

Number of university applications - Making a decision

- *p*-value = 0.0003
  - If the true average of the number of universities Trent students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Trent students who on average apply to 9.7 or more schools.
  - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.
- Since the *p*-value is **low** (lower than 5%) we **reject** H_0.

Number of university applications - Making a decision

- p-value = 0.0003
  - If the true average of the number of universities Trent students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Trent students who on average apply to 9.7 or more schools.
  - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.
- Since the p-value is **low** (lower than 5%) we **reject** H_0.
- The data provide convincing evidence that Trent students apply to more than 8 schools on average.

Number of university applications - Making a decision

- p-value = 0.0003
    - If the true average of the number of universities Trent students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Trent students who on average apply to 9.7 or more schools.
    - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.
- Since the p-value is **low** (lower than 5%) we **reject** H_0.
- The data provide convincing evidence that Trent students apply to more than 8 schools on average.
- The difference between the null value of 8 schools and observed sample mean of 9.7 schools is **not due to chance** or sampling variability.

A poll by the National Sleep Foundation (USA) found that college students average about 7 hours of sleep per night. A sample of 169 college students taking an introductory statistics class yielded an average of 6.88 hours, with a standard deviation of 0.94 hours. Assuming that this is a random sample representative of all college students *(bit of a leap of faith?)*, a hypothesis test was conducted to evaluate if college students on average sleep **less than** 7 hours per night. The p-value for this hypothesis test is 0.0485. Which of the following is correct?

- Fail to reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.
- Reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.
- Reject H_0, the data prove that college students sleep more than 7 hours on average.
- Fail to reject H_0, the data do not provide convincing evidence that college students sleep less than 7 hours on average.
- Reject H_0, the data provide convincing evidence that college students in this sample sleep less than 7 hours on average.

Two-sided hypothesis testing with p-values

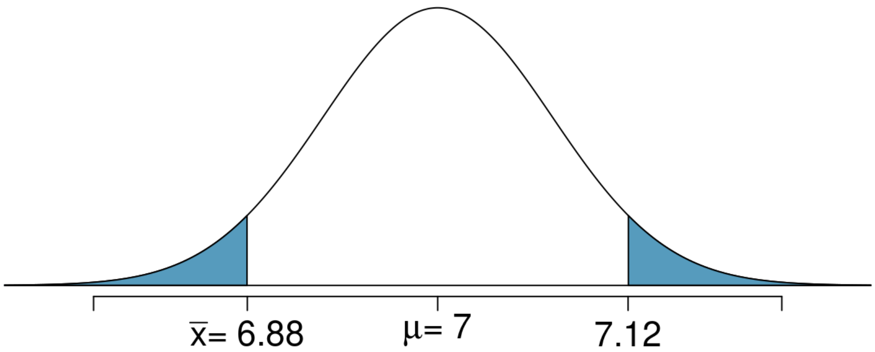
- If the research question was “Do the data provide convincing evidence that the average amount of sleep college students get per night is **different** than the national average?”, the alternative hypothesis would be different.

$$H_0 : \mu = 7$$
$$H_A : \mu \neq 7$$

Two-sided hypothesis testing with p-values

- If the research question was “Do the data provide convincing evidence that the average amount of sleep college students get per night is **different** than the national average?”, the alternative hypothesis would be different.
- Then the p-value **would change as well**:

$$\text{p-value} = 0.0485 \times 2 = 0.097$$



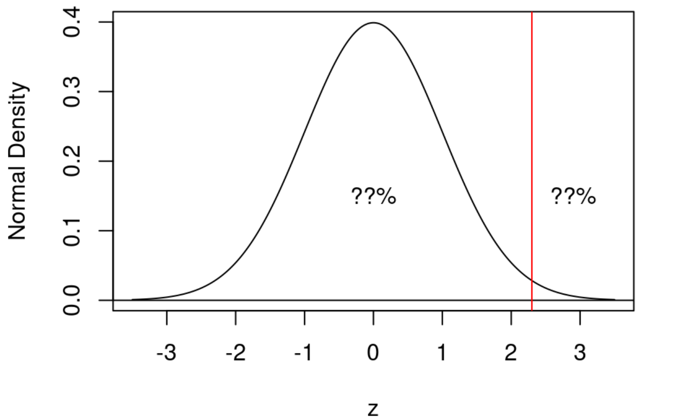
Computing the -value

How do we actually compute the *p*-value? We use **pnorm()**! There's a reason we made you learn about it!

Example: the **test statistic** is 2.3, with hypotheses

$$H_0 : \mu = 5 \qquad \text{versus} \qquad H_A : \mu > 5$$

What is the *p*-value?



Example, continued

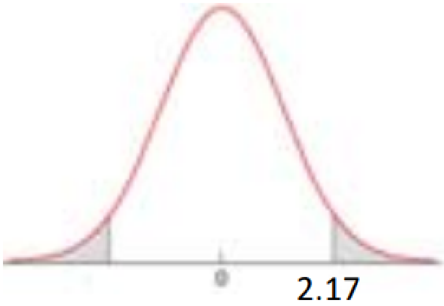
```r
pnorm(2.3, mean = 0, sd = 1, lower.tail = FALSE)
```

```
## [1] 0.01072411
```

So the *p*-value for this **one-tailed hypothesis test** is 0.011. What does this imply?

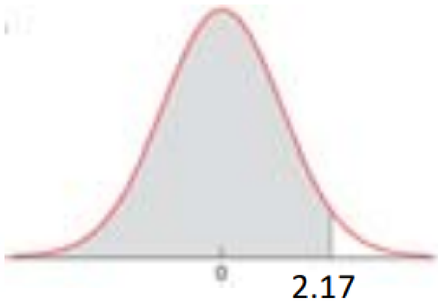
Since $0.011 < 0.05$, we do have evidence at the 95% level to reject the null hypothesis (whatever it is in context), and conclude that $\mu > 5$.

The Alternative Hypothesis …



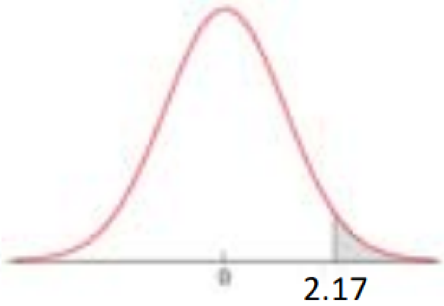
Two-tailed H_A
(≠)

Find the area to the right of z and multiply by 2, (or to the left of z if *z* were negative and multiply by 2).



Left-tailed H_A
(<)

Find the area to the left of z.



Right-tailed H_A
(>)

Find the area to the right of z.

Decision Errors

Decision errors

- Hypothesis tests are not flawless.

Decision errors

- Hypothesis tests are not flawless.
- In the court system innocent people are sometimes wrongly convicted and the guilty sometimes walk free.

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 as well.

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 as well.
- The difference is that we have the tools necessary to quantify how often we make errors in statistics.

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		✓

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*	✓

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.

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.

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.

Hypothesis Test as a trial

If we again 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:

$$\begin{aligned} H_0 &: \text{ Defendant is innocent} \\ H_A &: \text{ Defendant is guilty} \end{aligned}$$

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

Hypothesis Test as a trial

If we again 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:

$$
\begin{aligned}
H_0 &: \text{ Defendant is innocent} \\
H_A &: \text{ Defendant is guilty}
\end{aligned}
$$

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

Hypothesis Test as a trial

If we again 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:

$$\begin{aligned} H_0 &: \text{ Defendant is innocent} \\ H_A &: \text{ Defendant is guilty} \end{aligned}$$

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**

Hypothesis Test as a trial

If we again 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:

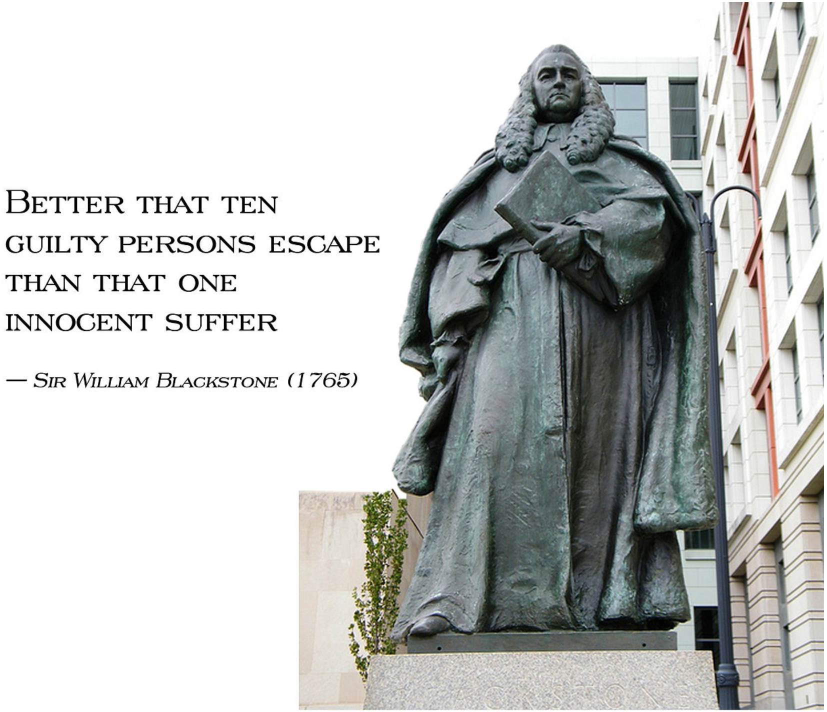
$$\begin{aligned} H_0 &: \text{ Defendant is innocent} \\ H_A &: \text{ Defendant is guilty} \end{aligned}$$

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?

Hypothesis Test as a trial

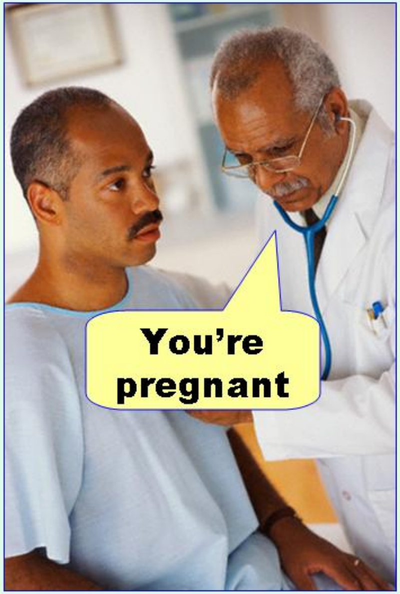


https://en.wikipedia.org/wiki/Blackstone%27s_ratio

Another way to remember

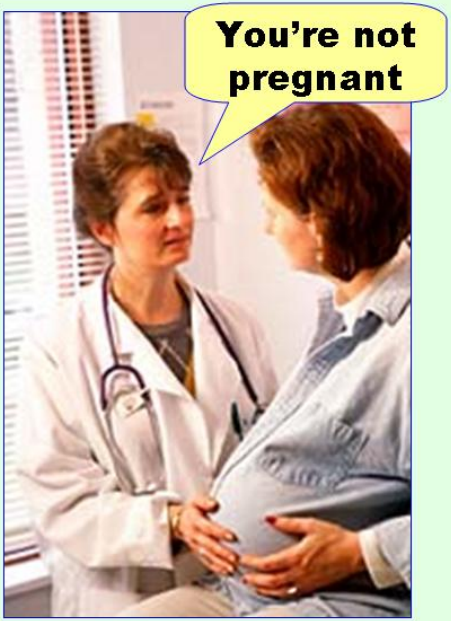
Type I error

(false positive)



Type II error

(false negative)



Another way to remember

For these medical diagnoses, what is happening?

- Null hypothesis is always “nothing going on”: so a **medical test** for pregnancy should have its null as “Not Pregnant”

Another way to remember

For these medical diagnoses, what is happening?

- Null hypothesis is always “nothing going on”: so a **medical test** for pregnancy should have its null as “Not Pregnant”
- So for the man in the left panel, being told “you are pregnant” means **reject the null** - select the alternative.

Another way to remember

For these medical diagnoses, what is happening?

- Null hypothesis is always "nothing going on": so a **medical test** for pregnancy should have its null as "Not Pregnant"
- So for the man in the left panel, being told "you are pregnant" means **reject the null** - select the alternative.
  - this is obviously incorrect

Another way to remember

For these medical diagnoses, what is happening?

- Null hypothesis is always “nothing going on”: so a **medical test** for pregnancy should have its null as “Not Pregnant”
- So for the man in the left panel, being told “you are pregnant” means **reject the null** - select the alternative.
    - this is obviously incorrect
    - therefore it is **false**

Another way to remember

For these medical diagnoses, what is happening?

- Null hypothesis is always “nothing going on”: so a **medical test** for pregnancy should have its null as “Not Pregnant”
- So for the man in the left panel, being told “you are pregnant” means **reject the null** - select the alternative.
    - this is obviously incorrect
    - therefore it is **false**
    - but the diagnosis was “positive” (the alternative)
    - this is equivalent to **declaring the defendent guilty, when they are actually innocent**

Another way to remember

For these medical diagnoses, what is happening?

- Null hypothesis is always “nothing going on”: so a **medical test** for pregnancy should have its null as “Not Pregnant”
- So for the woman in the right panel, being told “you are not pregnant” means **fail to reject the null** - there is no evidence against the null state

Another way to remember

For these medical diagnoses, what is happening?

- Null hypothesis is always “nothing going on”: so a **medical test** for pregnancy should have its null as “Not Pregnant”
- So for the woman in the right panel, being told “you are not pregnant” means **fail to reject the null** - there is no evidence against the null state
    - this is obviously incorrect (poor woman!)

Another way to remember

For these medical diagnoses, what is happening?

- Null hypothesis is always “nothing going on”: so a **medical test** for pregnancy should have its null as “Not Pregnant”
- So for the woman in the right panel, being told “you are not pregnant” means **fail to reject the null** - there is no evidence against the null state
  - this is obviously incorrect (poor woman!)
  - therefore it is **false**

Another way to remember

For these medical diagnoses, what is happening?

- Null hypothesis is always “nothing going on”: so a **medical test** for pregnancy should have its null as “Not Pregnant”
- So for the woman in the right panel, being told “you are not pregnant” means **fail to reject the null** - there is no evidence against the null state
  - this is obviously incorrect (poor woman!)
  - therefore it is **false**
  - the diagnosis was “negative” (against the alternative)
  - this is equivalent to **declaring the defendent innocent, when they are actually guilty**

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, $\alpha = 0.05$.

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.

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.

$$P\left(\text{Type 1 error } | H_0 \text{ true}\right) = \alpha$$

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.

$$P\left(\text{Type 1 error } | H_0 \text{ true}\right) = \alpha$$

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

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.

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.

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.

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.

Recap: Hypothesis testing framework

- Set the hypotheses.
- Check assumptions and conditions.
- Calculate a **test statistic** and a p-value.
- Make a decision, and interpret it in context of the research question.

Recap: Hypothesis testing for a population mean

- Set the hypotheses
  - $H_0 : \mu = \text{null value}$
  - $H_A : \mu <$ or $>$ or $\neq$ null value
- Calculate the point estimate
- Check assumptions and conditions
  - Independence: random sample/assignment, 10% condition when sampling without replacement
  - Normality: nearly normal population or $n \geq 30$, no extreme skew – **or use the t distribution** (next chapter)

Recap: Hypothesis testing for a population mean

- Calculate a **test statistic** and a p-value (draw a picture!)

$$Z = \frac{\bar{x} - \mu}{SE}, \text{ where } SE = \frac{s}{\sqrt{n}}$$

- Make a decision, and interpret it in context
  - If p-value $< \alpha$, reject H_0, data provide evidence for H_A
  - If p-value $> \alpha$, do not reject H_0, data do not provide evidence for H_A