7 One-Sample Hypothesis Tests (Cont'd)

MTH 3240 Environmental Statistics

Spring 2020

Objectives

Objectives:

- Recognize data snooping and explain why it can lead to incorrect conclusions in hypothesis testing.
- Differentiate between Type I and II errors.
- State the relationship between the level of significance and the probability of a Type I error.
- Differentiate between statistical significance and practical importance.

Data Snooping

- Choosing a **direction** for a **one-sided** H_a is intended to be a **prediction** of what you think the data will indicate.

Data Snooping

- Choosing a **direction** for a **one-sided** H_a is intended to be a **prediction** of what you think the data will indicate.
- ***Data snooping*** refers to waiting until **after you've looked at the data** to choose a direction for H_a, **and then** testing H_a in the direction that matches what you **already see in the data**.

Data Snooping

- Choosing a **direction** for a **one-sided** $\boldsymbol{H_a}$ is intended to be a **prediction** of what you think the data will indicate.

- ***Data snooping*** refers to waiting until **after you've looked at the data** to choose a direction for H_a, **and then** testing H_a in the direction that matches what you **already see in the data.**

- Data snooping is "**cheating**" because it results in an **artificially small p-value**, which can lead to mistakenly declaring a spurious result statistically significant.

- A **one-sided** H_a should **only** be used if you have a specific direction in mind **prior** to looking at the data.

- A **one-sided** H_a should **only** be used if you have a specific direction in mind **prior** to looking at the data.

  Otherwise, use a **two-sided** H_a.

- A **one-sided** H_a should **only** be used if you have a specific direction in mind **prior** to looking at the data.

  Otherwise, use a **two-sided** H_a.

- The next example shows that **data snooping** can lead to a **p-value** that's **half as large as it's supposed to be**.

Example

A laboratory quality assurance study was carried out to **look for signs of systematic bias** in a lab's measurements of total organic carbon (**TOC**), an indicator of water quality.

Example

A laboratory quality assurance study was carried out to **look for signs of systematic bias** in a lab's measurements of total organic carbon (**TOC**), an indicator of water quality.

Sixteen certified standard solutions having **50** mg/L TOC were randomly inserted into the lab's work stream. Lab analysts were unaware of the presence of these standard solutions.

If there's **bias**, their measurements will tend to systematically **differ from 50** in the **direction of the bias**.

If there's **bias**, their measurements will tend to systematically **differ from 50** in the **direction of the bias**.

But If there's **no bias**, they should **equal 50** *on average*.

If there's **bias**, their measurements will tend to systematically **differ from 50** in the **direction of the bias.**

But If there's **no bias**, they should **equal 50** *on average*.

Because there **isn't** a particular direction in mind for the bias, the **appropriate** hypotheses to test are

$$\begin{aligned} H_0 : \mu &= 50 \\ H_a : \mu &\neq 50 \end{aligned}$$

where μ is the lab's true (unknown) population **mean** measurement result for **50** mg/L standard solutions.

H_0 says there's **no bias**. H_a says there **is bias**, but doesn't specify a direction.

H_0 says there's **no bias**. H_a says there **is bias**, but doesn't specify a direction.

We'll use **level of significance** $\alpha = 0.01$.

H_0 says there's **no bias**. H_a says there **is bias**, but doesn't specify a direction.

We'll use **level of significance** $\alpha = 0.01$.

Here are the lab's results for $n = 16$ of the standard solutions:

50.3	51.2	50.5	50.2	49.9	50.2	50.3	50.5
49.3	50.0	50.4	50.1	51.0	49.8	50.7	50.6

H_0 says there's **no bias**. H_a says there **is bias**, but doesn't specify a direction.

We'll use **level of significance** $\alpha = 0.01$.

Here are the lab's results for $n = 16$ of the standard solutions:

50.3	51.2	50.5	50.2	49.9	50.2	50.3	50.5
49.3	50.0	50.4	50.1	51.0	49.8	50.7	50.6

The **sample mean** and **standard deviation** are

$$\bar{X} = 50.31 \qquad \text{and} \qquad S = 0.46.$$

The **standard error** of $\bar{X}$ is

$$\boldsymbol{S_{\bar{X}}} = \frac{0.46}{\sqrt{16}} = \mathbf{0.115}.$$

so the **test statistic** is

$$\boldsymbol{t} = \frac{50.31 - 50}{0.115} = \mathbf{2.70}.$$

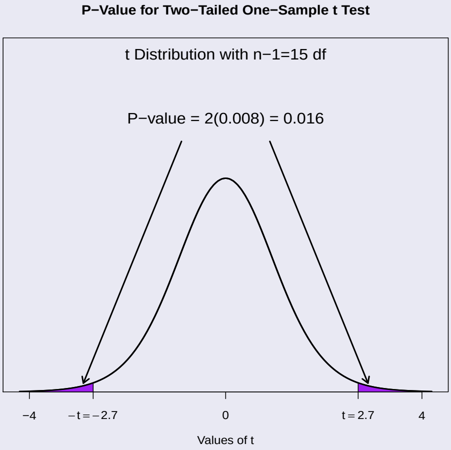
The **standard error** of $\bar{X}$ is

$$\boldsymbol{S_{\bar{X}}} = \frac{0.46}{\sqrt{16}} = \mathbf{0.115}.$$

so the **test statistic** is

$$\boldsymbol{t} = \frac{50.31 - 50}{0.115} = \mathbf{2.70}.$$

For the **two-sided** test, the **p-value** is the sum of the **two tail areas** shown below.


P−Value for Two−Tailed One−Sample t Test
t Distribution with n−1=15 df
P−value = 2(0.008) = 0.016
−4
−t = −2.7
0
t = 2.7
4
Values of t

From a t table, using $n - 1 = 15$ df, the **p-value** is $2(0.0082) = \mathbf{0.0164}$.

Using $\boldsymbol{\alpha = 0.01}$, we'd **fail to reject $\boldsymbol{H_0}$**.

Now suppose that we had **data snooped**, and decided, **after** noticing that $\bar{X} = 50.31$ is **greater than 50**, to do a one-sided, **upper-tailed** test of

$$\begin{aligned} H_0 : \mu &= 50 \\ H_a : \mu &> 50 \end{aligned}$$

using $\alpha = 0.01$ again.

The test statistic would still be $t = 2.70$, but now the p-value would be **just the upper tail** area, which is **0.0082**.

The test statistic would still be $t = 2.70$, but now the p-value would be **just the upper tail** area, which is **0.0082**.

This p-value is **half** of what it was for the two-tailed test, and using $\alpha = 0.01$, now we'd **reject** H_0.

The test statistic would still be $t = 2.70$, but now the p-value would be **just the upper tail** area, which is **0.0082.**

This p-value is **half** of what it was for the two-tailed test, and using $\alpha = 0.01$, now we'd **reject H_0.**

Here, we'd **mistakenly** conclude there's bias in the positive direction, and might **unnecessarily** recommend corrective actions.

Type I and II Errors

- Any time we carry out a hypothesis test, there's always a possibility that we might reach the **wrong conclusion.**

Type I and II Errors

- Any time we carry out a hypothesis test, there's always a possibility that we might reach the **wrong conclusion.**

  A ***Type I error*** ("false positive") occurs when we mistakenly **reject** H_0 even though in fact H_0 **is true.**

Type I and II Errors

- Any time we carry out a hypothesis test, there's always a possibility that we might reach the **wrong conclusion.**

  A ***Type I error*** ("false positive") occurs when we mistakenly **reject** H_0 even though in fact H_0 **is true.**

  A ***Type II error*** ("false negative") occurs when we mistakenly ***fail*** **to reject** H_0 even though H_a **is true.**

Type I and II Errors

		True State of Nature: H_0	True State of Nature: H_a
Your Decision	**Reject H_0**	Type I Error	Correct Decision
	Fail to Reject H_0	Correct Decision	Type II Error

Exercise

Let μ denote the true (unknown) population **mean radioactivity level** in a certain lake.

Exercise

Let μ denote the true (unknown) population **mean radioactivity level** in a certain lake.

The value **5** pCi/L is considered the dividing line between **safe** and **hazardous** water.

Exercise

Let μ denote the true (unknown) population **mean radioactivity level** in a certain lake.

The value **5** pCi/L is considered the dividing line between **safe** and **hazardous** water.

To decide whether the lake's water is safe, a random sample of **50** water specimens is selected, and the radioactivity measured in each specimen.

a) Suppose we decide to test the hypotheses

$$\begin{aligned} H_0 : \mu &\leq 5 \\ H_a : \mu &> 5 \end{aligned}$$

Which of following is a **Type I error** which is a **Type II error**?

a) Suppose we decide to test the hypotheses

$$\begin{aligned} H_0 : \mu &\leq 5 \\ H_a : \mu &> 5 \end{aligned}$$

Which of following is a **Type I error** which is a **Type II error**?

A. In reality the **water is safe**, but we **conclude it's hazardous.**

a) Suppose we decide to test the hypotheses

$$\begin{aligned} H_0 : \mu &\leq 5 \\ H_a : \mu &> 5 \end{aligned}$$

Which of following is a **Type I error** which is a **Type II error**?

A. In reality the **water is safe**, but we **conclude it's hazardous.**

B. In reality the **water is hazardous**, but we **conclude it's safe.**

b) In Part *a*, which type of error has **more serious consequences**?

c) Now suppose instead that the hypotheses are

$$
\begin{aligned}
H_0 : \mu &\geq 5 \\
H_a : \mu &< 5
\end{aligned}
$$

Now which of following is a **Type I error** which is a **Type II error**?

c) Now suppose instead that the hypotheses are

$$\begin{aligned} H_0 : \mu &\geq 5 \\ H_a : \mu &< 5 \end{aligned}$$

Now which of following is a **Type I error** which is a **Type II error**?

A. In reality the **water is safe**, but we **conclude it's hazardous.**

c) Now suppose instead that the hypotheses are

$$H_0 : \mu \geq 5$$
$$H_a : \mu < 5$$

Now which of following is a **Type I error** which is a **Type II error**?

A. In reality the **water is safe**, but we **conclude it's hazardous.**

B. In reality the **water is hazardous**, but we **conclude it's safe.**

d) In part *c*, which type of error has **more serious consequences**?

Level of Significance as the Type I Error Probability

- The **level of significance** α turns out to be the **probability** of making a **Type I error** (when H_0 is true).

Level of Significance as the Type I Error Probability

- The **level of significance** α turns out to be the **probability** of making a **Type I error** (when H_0 is true).

For a hypothesis test using level of significance α, when H_0 is true,

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

Level of Significance as the Type I Error Probability

- The **level of significance** α turns out to be the **probability** of making a **Type I error** (when H_0 is true).

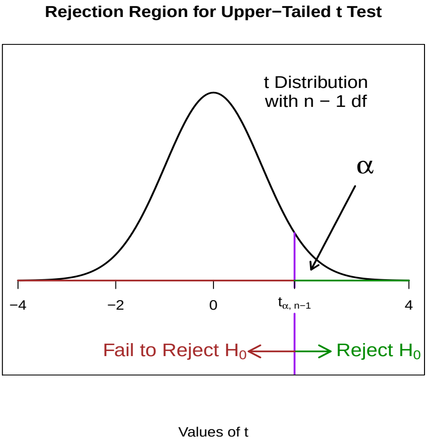
> For a hypothesis test using level of significance α, when H_0 is true,
> $$P(\text{Type I error}) = \alpha.$$

- Thus using $\alpha = \mathbf{0.05}$, if H_0 was true, there'd be a **5% chance** we'd **mistakenly** reject H_0.

- To see why, consider an upper-tailed test using the rejection region approach. In this case:

- To see why, consider an upper-tailed test using the rejection region approach. In this case:
  - The **probability** that the test statistic t will fall to the right of the critical value $t_{\alpha,n-1}$ (when H_0 is true) is α. See the next slide.

- To see why, consider an upper-tailed test using the rejection region approach. In this case:
  - The **probability** that the test statistic t will fall to the right of the critical value $t_{\alpha,n-1}$ (when H_0 is true) is α. See the next slide.
  - If this happens, a **Type I error** is committed.



The $t(n-1)$ distribution is the sampling distribution of the test statistic t *when* H_0 *is true*.

Choosing a Level of Significance

- Because the level of significance is the *probability of making a Type I error*, **if a Type I error** has very **serious consequences**, we should use a **small value for** α (e.g. 0.01).

Exercise

Suppose again μ is the (unknown) population **mean radioactivity level** in a lake.

Exercise

Suppose again μ is the (unknown) population **mean radioactivity level** in a lake.

Suppose we want to test the hypotheses

$$\begin{aligned} H_0 : \mu &\geq 5 \\ H_a : \mu &< 5 \end{aligned}$$

Exercise

Suppose again μ is the (unknown) population **mean radioactivity level** in a lake.

Suppose we want to test the hypotheses

$$\begin{aligned} H_0 : \mu &\geq 5 \\ H_a : \mu &< 5 \end{aligned}$$

Recall (previous exercise) that a **Type I error** would result if we **conclude** the **water's safe** even though in fact **it's hazardous.**

Exercise

Suppose again μ is the (unknown) population **mean radioactivity level** in a lake.

Suppose we want to test the hypotheses

$$\begin{aligned} H_0 : \mu &\geq 5 \\ H_a : \mu &< 5 \end{aligned}$$

Recall (previous exercise) that a **Type I error** would result if we **conclude** the **water's safe** even though in fact **it's hazardous.**

Which level of significance, **0.10**, **0.05**, or **0.01**, should we use?

- But there's a trade-off in using a small α to guard against a Type I error:

- But there's a trade-off in using a small α to guard against a Type I error:

  Using a **smaller $\boldsymbol{\alpha}$ requires stronger evidence against $\boldsymbol{H_0}$** (i.e. t farther away from zero) before we're willing to reject H_0 ...

- But there's a trade-off in using a small α to guard against a Type I error:

  Using a **smaller $\boldsymbol{\alpha}$ requires stronger evidence against $\boldsymbol{H_0}$** (i.e. t farther away from zero) before we're willing to reject H_0 ...

  ... but requiring stronger evidence against H_0 means we're less likely to reject H_0 **even when $\boldsymbol{H_a}$ is true.**

- But there's a trade-off in using a small α to guard against a Type I error:

  Using a **smaller α requires stronger evidence against H_0** (i.e. t farther away from zero) before we're willing to reject H_0 ...

  ... but requiring stronger evidence against H_0 means we're less likely to reject H_0 **even when H_a is true.**

  In other words, using a **smaller α** makes it **more likely that we'll make a Type II error** (when H_a is true).

Tradeoff Between Type I and II Error Probabilities

	Value of α		
	Small (e.g. 0.01)	**Medium (e.g. 0.05)**	**Large (e.g. 0.10)**
Probability of Type I Error	Small	Medium	Large
Probability of Type II Error	Large	Medium	Small

Statistical Significance Versus Practical Importance

- The **p-value** of a hypothesis test indicates **how strong** the **evidence** against H_0 is:

P-value	**Strength of Evidence**
P-value > 0.10	Weak
$0.05 <$ P-value < 0.10	Moderate
$0.01 <$ P-value < 0.05	Strong
P-value < 0.01	Very Strong

Statistical Significance Versus Practical Importance

- The **p-value** of a hypothesis test indicates **how strong** the **evidence** against H_0 is:

P-value	**Strength of Evidence**
P-value > 0.10	Weak
$0.05 <$ P-value < 0.10	Moderate
$0.01 <$ P-value < 0.05	Strong
P-value < 0.01	Very Strong

A **small p-value** indicates that a difference or effect was **detected**, but **not necessarily** that it's *large*.

- More precisely, a **small p-value** can arise in either of **two ways**:

- More precisely, a **small p-value** can arise in either of **two ways**:
  - The sample size n is **small**, but the difference or effect being tested for is **large**.
  - The difference or effect being tested for is **small**, but the sample size n is **large**.

- Thus a study result that's **statistically significant** (p-value $< \alpha$) isn't necessarily one that has **practical importance**.

- Thus a study result that's **statistically significant** (p-value $< \alpha$) isn't necessarily one that has **practical importance.**

  The former only means that a difference or effect was **detected**, but doesn't say anything about its **size.**

- Thus a study result that's **statistically significant** (p-value $< \alpha$) isn't necessarily one that has **practical importance.**

  The former only means that a difference or effect was **detected**, but doesn't say anything about its **size.**

  Differences or effects that are so **small** as to **not** to have any impactful consequences can nonetheless be found to be statistically significant when the **sample size** n is **large.**

- **Example**: A study may find statistically significant evidence that clearcutting a forest caused an increase in a nearby stream's temperature (via increased solar radiation).

- **Example**: A study may find statistically significant evidence that clearcutting a forest caused an increase in a nearby stream's temperature (via increased solar radiation).

  But the increase may be so small that the stream's biology is unaffected.

- In statistical parlance, studies that use **very large sample sizes** are said to have very high ***power*** for detecting (even small) differences or effects.

- In statistical parlance, studies that use **very large sample sizes** are said to have very high ***power*** for detecting (even small) differences or effects**.**

  The next example illustrates how a **large** n can lead to a **small p-value**.

Example

Consider a **one-sample** t test of

$$\begin{aligned} H_0 : \mu &= 5 \\ H_a : \mu &> 5 \end{aligned}$$

and suppose

$$\bar{X} = \mathbf{5.1} \qquad \text{and} \qquad S = \mathbf{1.3}$$

The **test statistic** is

$$t = \frac{\bar{X} - \mu_0}{S_{\bar{X}}}, \quad \text{where} \quad S_{\bar{X}} = \frac{S}{\sqrt{n}},$$

The **test statistic** is

$$t \;=\; \frac{\bar{X} - \mu_0}{S_{\bar{X}}}, \quad \text{where} \quad S_{\bar{X}} \;=\; \frac{S}{\sqrt{n}},$$

i.e.

$$t \;=\; \frac{\mathbf{5.1 - 5}}{\boldsymbol{S_{\bar{X}}}}, \quad \text{where} \quad S_{\bar{X}} \;=\; \frac{\mathbf{1.3}}{\boldsymbol{\sqrt{n}}}$$

The **test statistic** is

$$t \;=\; \frac{\bar{X} - \mu_0}{S_{\bar{X}}}, \quad \text{where} \quad S_{\bar{X}} \;=\; \frac{S}{\sqrt{n}},$$

i.e.

$$t \;=\; \frac{\mathbf{5.1 - 5}}{\boldsymbol{S_{\bar{X}}}}, \quad \text{where} \quad S_{\bar{X}} \;=\; \frac{\mathbf{1.3}}{\boldsymbol{\sqrt{n}}}$$

The **p-value** can differ depending on how big the sample size $\boldsymbol{n}$ was:

If $\boldsymbol{n = 10}$:
$t \;=\; 0.24$
df $=\; 9$
p-value $=\; 0.4079$

If $\boldsymbol{n = 1{,}000}$:
$t = 2.43$
df $=\; 999$
p-value $=\; 0.0076$