Statistical Methods

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

August 20, 2019

Topics

Objectives

Objectives:

- Explain the meaning of the terms *hypothesis*, *test statistic*, *level of significance*, *rejection region*, *p-value*, and *decision rule*.
- Carry out a one-sample t test for a population mean.
- Distinguish between Type I and Type I errors.
- Know the relationship between the level of significance and the Type I error probability.

Introduction to Hypothesis Testing

- A ***hypothesis*** is a claim about the value(s) of one or more population parameters, (e.g. μ).

Introduction to Hypothesis Testing

- A ***hypothesis*** is a claim about the value(s) of one or more population parameters, (e.g. μ).
- A ***hypothesis test*** is a statistical method for deciding between two hypotheses:

Introduction to Hypothesis Testing

- A ***hypothesis*** is a claim about the value(s) of one or more population parameters, (e.g. μ).
- A ***hypothesis test*** is a statistical method for deciding between two hypotheses:
  - The ***null hypothesis*** ($\boldsymbol{H_0}$) is the hypothesis we seek to **discredit**, but to which we give the **benefit of the doubt**.

Introduction to Hypothesis Testing

- A ***hypothesis*** is a claim about the value(s) of one or more population parameters, (e.g. μ).
- A ***hypothesis test*** is a statistical method for deciding between two hypotheses:
  - The ***null hypothesis*** (H_0) is the hypothesis we seek to **discredit**, but to which we give the **benefit of the doubt**.
  - The ***alternative hypothesis*** (H_a) is the hypothesis we seek to **substantiate**.

- The **conclusion** in any hypothesis test will be to either

Reject H_0 or **Fail to Reject H_0**.

- The **conclusion** in any hypothesis test will be to either

  Reject H_0 or **Fail to Reject H_0**.

- The decision is based on whether a ***test statistic*** provides compelling evidence against H_0, ...

- The **conclusion** in any hypothesis test will be to either

Reject H_0 or **Fail to Reject H_0**.

- The decision is based on whether a ***test statistic*** provides compelling evidence against H_0, ...

  ... as determined by comparing its value to the sampling distribution it *would* follow *if* H_0 was true.

- The **conclusion** in any hypothesis test will be to either

  Reject H_0 or **Fail to Reject H_0**.

- The decision is based on whether a ***test statistic*** provides compelling evidence against H_0, ...

  ... as determined by comparing its value to the sampling distribution it *would* follow *if* H_0 was true.

- A ***decision rule*** specifies when the evidence against H_0 is so compelling that H_0 should be rejected.

- There are **two approaches** to developing a decision rule:

- There are **two approaches** to developing a decision rule:

    1. The ***rejection region approach***.

- There are **two approaches** to developing a decision rule:

  1. The ***rejection region approach***.

  2. The ***p-value approach***.

- There are **two approaches** to developing a decision rule:

  1. The ***rejection region approach***.
  2. The ***p-value approach***.

  In either case, we first choose a ***level of significance*** α, which indicates how strong the evidence against H_0 needs to be before we're willing to reject H_0.

- There are **two approaches** to developing a decision rule:

  1. The ***rejection region approach***.
  2. The ***p-value approach***.

  In either case, we first choose a ***level of significance*** α, which indicates how strong the evidence against H_0 needs to be before we're willing to reject H_0.

  A **smaller** α requires **stronger evidence**.

- There are **two approaches** to developing a decision rule:

  1. The ***rejection region approach***.

  2. The ***p-value approach***.

  In either case, we first choose a ***level of significance*** α, which indicates how strong the evidence against H_0 needs to be before we're willing to reject H_0.

  A **smaller** α requires **stronger evidence**.

  The most commonly used values for α are **0.01**, **0.05**, and **0.10**.

- A ***rejection region*** is the **set** all **test statistic values** for which H_0 should be rejected.

- A ***rejection region*** is the **set** all **test statistic values** for which H_0 should be rejected.

  It's chosen in such a way that when H_0 is true, the test statistic will fall into that region just by chance with probability α.

Decision Rule (RR approach):

Reject H_0 if the test statistic falls in the rejection region.
Fail to reject H_0 otherwise.

- The ***p-value*** is a **probability** that answers the question:

  *"**If H_0 was true**, what's the chance we'd get a test statistic value that's as contradictory to H_0 (and consistent with H_a) as the one we got?"*

Decision Rule (P-value approach):

Reject H_0 if p-value $< \alpha$.
Fail to reject H_0 if p-value $\geq \alpha$.

- We say that a result is ***statistically significant*** when we reject H_0.

- We say that a result is ***statistically significant*** when we reject H_0.

  A statistically significant result is one that isn't likely just due to chance variation.

Steps in Performing a Hypothesis Test:

1. Identify and define the parameter(s) of interest.
2. State the null and alternative hypotheses.
3. Choose a level of significance α.
4. Check any assumptions required for the test.
5. Calculate the test statistic value.
6. Compute the p-value or determine the rejection region.
7. State the conclusion (using the decision rule).

One-Sample t Test for μ (8.3)

- Suppose $X_1, X_2, \ldots, X_n$ are a random sample from a population whose (unknown) mean is μ.

One-Sample t Test for μ (8.3)

- Suppose $X_1, X_2, \ldots, X_n$ are a random sample from a population whose (unknown) mean is μ.
- We'll see how to use the sample to decide if μ is different from some **hypothesized value** $\boldsymbol{\mu_0}$.

- Because we're seeking to "disprove" the claim that μ is equal to μ_0, the **null hypothesis** is that it *is* equal to μ_0 .

Null Hypothesis:

$$H_0 : \ \mu \ = \ \mu_0$$

- The **alternative hypothesis** will depend on what we're trying to "prove":

Alternative Hypothesis: The alternative hypothesis will be one of

1. $H_a: \mu > \mu_0$ (**one-sided, upper-tailed**)
2. $H_a: \mu < \mu_0$ (**one-sided, lower-tailed**)
3. $H_a: \mu \neq \mu_0$ (**two-sided, two-tailed**)

depending on what we're trying to verify using the data.

One-Sample t Test Statistic:

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

One-Sample t Test Statistic:

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

- T measures how many standard errors $\bar{X}$ is away from $\boldsymbol{\mu_0}$.

One-Sample t Test Statistic:

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

- T measures how many standard errors $\bar{X}$ is away from μ_0.
- $\bar{X}$ is an estimator of the unknown population mean μ, so ...

One-Sample t Test Statistic:

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

- T measures how many standard errors $\bar{X}$ is away from μ_0.
- $\bar{X}$ is an estimator of the unknown population mean μ, so ...
  1. T will be approximately **zero** (most likely) if $\mu = \mu_0$.
  2. It will be **positive** (most likely) if $\mu > \mu_0$.
  3. It will be **negative** (most likely) if $\mu < \mu_0$.

1. **Large positive** values of T provide **evidence against H_0 in favor of**
   $H_a : \mu > \mu_0$.
2. **Large negative** values of T provide **evidence against H_0 in favor of**
   $H_a : \mu < \mu_0$.
3. **Large positive *and* large negative** values of T provide **evidence**
   against H_0 in favor of
   $H_a : \mu \neq \mu_0$.

- If either:

  1. The sample is from a $N(\mu, \sigma)$ population, or
  2. The sample size n **is large**,

- If either:

  1. The sample is from a $N(\mu, \sigma)$ population, or
  2. The sample size n **is large**,

  then
$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

- If either:

  1. The sample is from a $N(\mu, \sigma)$ population, or
  2. The sample size n **is large**,

  then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

- It follows that **if H_0 is true** (so $\mu = \mu_0$),

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1).$$

Sampling Distribution of the Test Statistic Under H_0:
If T is the one-sample t test statistic, then when

$$H_0: \ \mu \ = \ \mu_0$$

is true,

$$T \ \sim \ t(n-1).$$

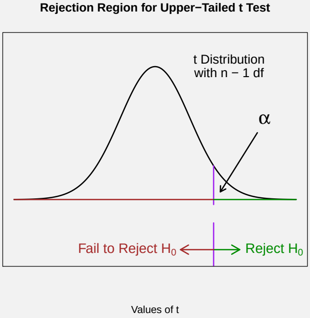
- The $t(n-1)$ curve gives us:

- The $t(n-1)$ curve gives us:
  - The ***rejection region*** as the **extreme 100α% of t values** (in the direction(s) specified by H_a).

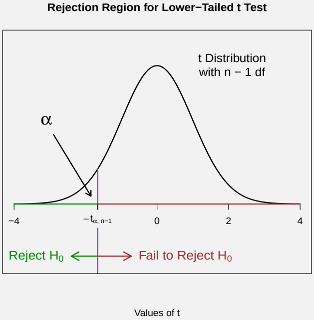
- The $t(n-1)$ curve gives us:
  - The ***rejection region*** as the **extreme 100α% of t values** (in the direction(s) specified by H_a).
  - The ***p-value*** as the **tail area(s) beyond the observed t value** (in the direction(s) specified by H_a).

Rejection Region: The **rejection region** is the **set of** t **values** in the tail of the $t(n-1)$ curve:

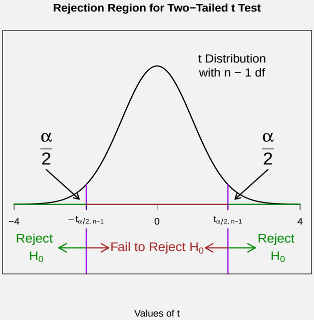
1. To the **right of** $t_{\alpha,\, n\,-\,1}$ if the alternative hypothesis is $H_a : \mu > \mu_0$:



2. To the **left of** $-t_{\alpha,\, n\,-\,1}$ if the alternative hypothesis is $H_a : \mu < \mu_0$:

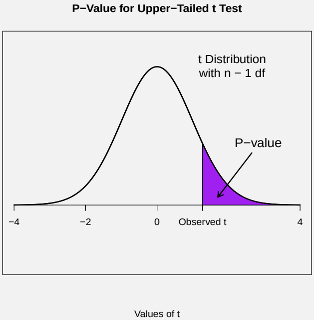


3. To the **left of** $-t_{\alpha/2,\,n\,-\,1}$ **and right of** $t_{\alpha/2,\,n\,-\,1}$ if the alternative hypothesis is $H_a : \mu \neq \mu_0$:

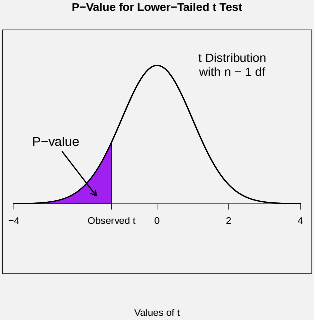


P-Value: The **p-value** is the **tail area** under the $t(n-1)$ curve:

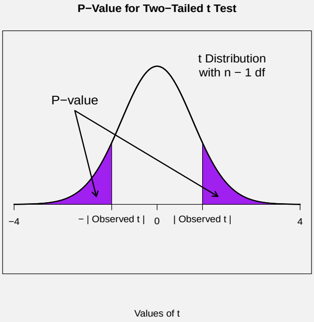
1. To the **right** of the **observed** t if the alternative hypothesis is $H_a : \mu > \mu_0$:



2. To the **left** of the **observed** t if the alternative hypothesis is $H_a : \mu < \mu_0$:



3. To the **left of** $-|t|$ **and right of** $|t|$ if the alternative hypothesis is $H_a : \mu \neq \mu_0$:



- The rejection region and p-value approaches **always reach the same conclusion.**

- The rejection region and p-value approaches **always reach the same conclusion.**

  (The **p-value** will be less than α if and only if t is in the **rejection region**).

Example

A quality control engineer monitors a machine that puts cereal into boxes.

Example

A quality control engineer monitors a machine that puts cereal into boxes.

According to the label, each box is supposed to contain **16** oz of cereal.

Example

A quality control engineer monitors a machine that puts cereal into boxes.

According to the label, each box is supposed to contain **16** oz of cereal.

The machine will need to be adjusted if the boxes are systematically being **under-filled** or **over-filled**.

Example

A quality control engineer monitors a machine that puts cereal into boxes.

According to the label, each box is supposed to contain **16** oz of cereal.

The machine will need to be adjusted if the boxes are systematically being **under-filled** or **over-filled**.

From past experience, the engineer knows that the weight (ounces) of the cereal in a box follows a **normal** distribution.

To decide if the boxes are being **under-filled or overfilled**, the engineer will test the **hypotheses**

$$\begin{aligned} H_0: \mu &= 16 \\ H_a: \mu &\neq 16 \end{aligned}$$

where μ is the true (unknown) population mean weight.

To decide if the boxes are being **under-filled or overfilled**, the engineer will test the **hypotheses**

$$\begin{aligned} H_0: \ \mu &= 16 \\ H_a: \ \mu &\neq 16 \end{aligned}$$

where μ is the true (unknown) population mean weight.

A random sample of **ten** boxes gives

$$\bar{x} = \mathbf{16.6} \qquad \text{and} \qquad s = \mathbf{0.9}.$$

The observed **test statistic** is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

The observed **test statistic** is

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{16.6 - 16}{0.9/\sqrt{10}} \end{aligned}$$

The observed **test statistic** is

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{16.6-16}{0.9/\sqrt{10}} \\ &= \mathbf{2.11}. \end{aligned}$$

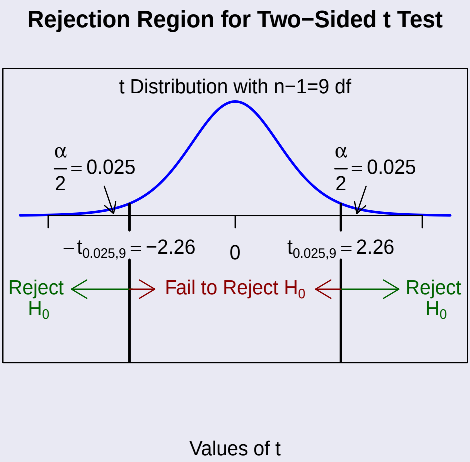
Thus the **sample mean** weight, $\bar{x} = \mathbf{16.6}$, is about **2.11 standard errors above 16** ounces.

For the **rejection region**, using a **level of significance** $\alpha = 0.05$, the t **critical value** is

$$t_{0.025,\,9} \;=\; 2.262,$$

and so the decision rule is

Reject H_0 if $t < -2.262$ or $t > 2.262$.
Fail to reject H_0 otherwise.


Rejection Region for Two−Sided t Test
t Distribution with n−1=9 df
$\frac{\alpha}{2} = 0.025$
$\frac{\alpha}{2} = 0.025$
$-t_{0.025,9} = -2.26$
0
$t_{0.025,9} = 2.26$
Reject H_0
Fail to Reject H_0
Reject H_0
Values of t

Because the test statistic, $t = 2.11$, is **not** in the rejection region, we **fail to reject H_0.**

Because the test statistic, $t = 2.11$, is **not** in the rejection region, we **fail to reject H_0.**

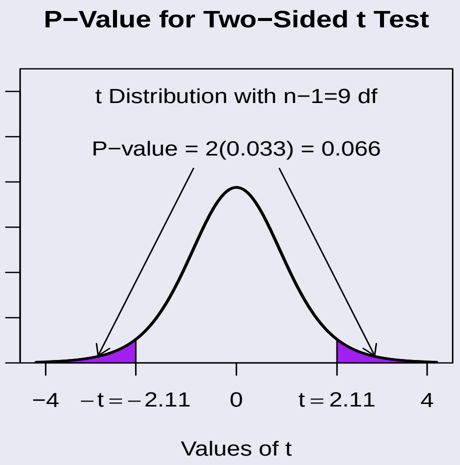
Thus the t value we got is **not** among the most extreme **5%** of values we'd get **if** the **population mean** μ was **16** ounces.

There's **no statistically significant evidence** that the population mean cereal box weight μ is different from 16 ounces.

There's **no statistically significant evidence** that the population mean cereal box weight μ is different from 16 ounces.

The result that the engineer got (by taking a random sample) can be explained by chance variation (sampling error).

The **p-value** is the **probability** that by chance we'd get a t value as far away from zero (in either direction) as $\boldsymbol{t = 2.11}$ **if** the **population mean** μ was **16** oz.


P−Value for Two−Sided t Test
t Distribution with n−1=9 df
P−value = 2(0.033) = 0.066
−4
−t = −2.11
0
t = 2.11
4
Values of t

From the **two tail** areas of the **sampling distribution** that the test statistic would follow under H_0 (the t(9) distribution), to the **right** of **2.11** and **left** of **-2.11**,

$$\textbf{p-value} \;=\; 2(0.033) \;=\; \mathbf{0.066}.$$

From the **two tail** areas of the **sampling distribution** that the test statistic would follow under H_0 (the t(9) distribution), to the **right** of **2.11** and **left** of **-2.11**,

$$\textbf{p-value} \;=\; 2(0.033) \;=\; \mathbf{0.066}.$$

Thus we'd get a result like the one we got **6.6%** of the time **even if** the **population mean** μ was **16** ounces.

Using $\alpha = \mathbf{0.05}$, the **decision rule** is

Reject H_0 if p-value < 0.05.
Fail to reject H_0 if p-value ≥ 0.05.

Using $\alpha = 0.05$, the **decision rule** is

Reject H_0 if p-value < 0.05.
Fail to reject H_0 if p-value ≥ 0.05.

Because $0.066 \geq 0.05$, we **fail to reject H_0**.

- The next exercise illustrates the fact that using a **smaller** α means we require **stronger evidence** against H_0 before we're willing to reject H_0.

- The next exercise illustrates the fact that using a **smaller** α means we require **stronger evidence** against H_0 before we're willing to reject H_0.

Exercise

In the last example, if the engineer had used a level of significance $\alpha = 0.10$ instead, would his **conclusion** be any **different**?

- The next exercise illustrates the fact that using a **smaller** α means we require **stronger evidence** against H_0 before we're willing to reject H_0.

Exercise

In the last example, if the engineer had used a level of significance $\alpha = 0.10$ instead, would his **conclusion** be any **different**?

What if he used $\alpha = 0.01$?

Data Snooping: Don't Do It

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

Data Snooping: Don't Do It

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

Data Snooping: Don't Do It

- Choosing a **direction** for a **one-sided** H_a is intended to be a **prediction** of what the data will indicate.
- ***Data snooping*** refers to waiting until **after you've looked at the data** to decide on a direction for H_a, **and then** choosing the direction for H_a that best fits 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 to be real.

- 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 should be**.

Exercise

Suppose the engineer who monitors cereal box weights was to "cheat" by **data snooping**, and deciding, ***after*** and noticing that the sample mean, $\bar{x} = \mathbf{16.6}$, is above the target value **16** oz, to do a **one-sided**, **upper-tailed** test of

$$\begin{aligned} H_0 : \ \mu &= 16 \\ H_a : \ \mu &> 16 \end{aligned}$$

Exercise

Suppose the engineer who monitors cereal box weights was to "cheat" by **data snooping**, and deciding, ***after*** and noticing that the sample mean, $\bar{x} = \mathbf{16.6}$, is above the target value **16** oz, to do a **one-sided**, **upper-tailed** test of

$$\begin{aligned} H_0 : \ \mu &= 16 \\ H_a : \ \mu &> 16 \end{aligned}$$

a) What would the (artificially small) **p-value** be?

Exercise

Suppose the engineer who monitors cereal box weights was to "cheat" by **data snooping**, and deciding, ***after*** and noticing that the sample mean, $\bar{x} = \mathbf{16.6}$, is above the target value **16** oz, to do a **one-sided**, **upper-tailed** test of

$$\begin{aligned} H_0 : \mu &= 16 \\ H_a : \mu &> 16 \end{aligned}$$

a) What would the (artificially small) **p-value** be?

b) Using $\alpha = \mathbf{0.05}$, as before, would the **conclusion** be **different**?

Type I and II Errors and Their Probabilities

Type I and II Errors

Type I and II Errors and Their Probabilities

Type I and II Errors

- A ***Type I error*** occurs when H_0 is **mistakenly rejected** (even though H_0 is true).

Type I and II Errors and Their Probabilities

Type I and II Errors

- A ***Type I error*** occurs when H_0 is **mistakenly rejected** (even though H_0 is true).
- A ***Type II error*** occurs when H_0 is **mistakenly *not* rejected** (even though H_a true).

Type I and II Errors and Their Probabilities

Type I and II Errors

- A ***Type I error*** occurs when H_0 is **mistakenly rejected** (even though H_0 is true).
- A ***Type II error*** occurs when H_0 is **mistakenly *not* rejected** (even though H_a true).
- These are analogous to **false positives** and **false negatives** in medical tests.

		True State of Nature	
		$\mathbf{H_0}$	$\mathbf{H_a}$
Your Decision	**Reject $\mathbf{H_0}$**	**Type I Error**	Correct Decision
	Fail to Reject $\mathbf{H_0}$	Correct Decision	**Type II Error**

Type I Error Probabilities and the Level of Significance

- It turns out that the **chance** of making a **Type I error** (when H_0 is true) is α, the **level of significance.**

Type I Error Probabilities and the Level of Significance

- It turns out that the **chance** of making a **Type I error** (when H_0 is true) is α, the **level of significance.**
- To see why, consider the **rejection region** approach.

Type I Error Probabilities and the Level of Significance

- It turns out that the **chance** of making a **Type I error** (when H_0 is true) is α, the **level of significance.**
- To see why, consider the **rejection region** approach.
  - The **rejection region** is the most extreme **100α%** of the sampling distribution that the test statistic would follow **if H_0 was true**.

Type I Error Probabilities and the Level of Significance

- It turns out that the **chance** of making a **Type I error** (when H_0 is true) is α, the **level of significance.**
- To see why, consider the **rejection region** approach.
  - The **rejection region** is the most extreme **100α%** of the sampling distribution that the test statistic would follow **if H_0 was true**.
  - A **Type I error** occurs when the test statistic falls into the **rejection region** even though H_0 is true.

- Takeaway:
  - In order to reject H_0 when $\alpha = \mathbf{0.05}$, we require that the evidence against H_0 be so strong that it would occur by chance only 5% of the time if H_0 was true.

- Takeaway:
  - In order to reject H_0 when $\alpha = \mathbf{0.05}$, we require that the evidence against H_0 be so strong that it would occur by chance only 5% of the time if H_0 was true.
  - In order to reject H_0 when $\alpha = \mathbf{0.01}$, we require even stronger evidence. We require evidence that would occur by chance only 1% of the time.

- The choice of what value to use for α will depend on the consequences of making a Type I error: if they're serious, choose α to be very small.

Exercise

Let μ denote the true mean radioactivity level (pCi/L) in a certain lake.

Exercise

Let μ denote the true mean radioactivity level (pCi/L) in a certain lake.

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

Exercise

Let μ denote the true mean radioactivity level (pCi/L) in a certain lake.

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

To decide whether the water is safe, 50 water specimens are sampled from the lake, and the radioactivity level measured in each specimen.

a) Describe what the **Type I** and **Type II errors** would be (in the context of this problem) for each of the following sets of hypotheses.

$$H_0 : \mu = 5 \qquad H_0 : \mu = 5$$
$$H_a : \mu > 5 \qquad H_a : \mu < 5$$

a) Describe what the **Type I** and **Type II errors** would be (in the context of this problem) for each of the following sets of hypotheses.

$$H_0 : \mu = 5 \qquad\qquad H_0 : \mu = 5$$
$$H_a : \mu > 5 \qquad\qquad H_a : \mu < 5$$

b) If we were to test the second set of hypotheses, which **level of significance** would you recommend, $\alpha = 0.10$, $\alpha = 0.05$, or $\alpha = 0.01$?