



Chapter 11

Introduction to Hypothesis Testing

Statistical Inference – Hypothesis testing

Hypothesis testing is the second form of statistical inference. It also has greater applicability. The following examples give a snapshot of what it is.

Is the body temperature of teenagers higher than the normal human body temperature of 98.4 degrees? A random sample of 25 teenagers show an average temperature of 98.8 degrees, with a standard deviation of 0.5 degrees. Is the original claim true?

Suppose a production line operates with a mean filling weight of 16 ounces per container. Since over- or under-filling can be dangerous, a quality control inspector samples 30 items to determine whether or not the filling weight has to be adjusted. The sample revealed a mean of 16.32 ounces. From past data, the standard deviation is known to be .8 ounces. Is the filling process is in control?

An insurance company is reviewing its current policy rates. When originally setting the rates they believed that the average claim amount was \$1,800. They are concerned that the true mean is actually higher than this, because they could potentially lose a lot of money. They randomly select 40 claims, and calculate a sample mean of \$1,950. Assuming that the standard deviation of claims is \$500, test to see if the insurance company should be concerned.

Hypothesis Testing

Test of a Hypothesis

- A procedure leading to a decision about a particular hypothesis
- Hypothesis-testing procedures rely on using the information in a **random sample from the population of interest**
- If this information is *consistent* with the hypothesis, then we will do not reject the (null) hypothesis;
- if this information is *inconsistent* with the hypothesis, then we reject the (null) hypothesis in favor of the alternate hypothesis.

Statistical Hypotheses

A statistical hypothesis is a statement about the parameters of one or more populations.

Your friend says that his spider travels exactly 50 centimeters in one second which you don't agree to. This problem can be converted into a hypothesis testing problem.

Let $H_0 : \mu = 50$ centimeters per second and

$H_1 : \mu \neq 50$ centimeters per second

The statement $H_0 : \mu = 50$ is called the **null hypothesis**.

The statement $H_1 : \mu \neq 50$ is called the **alternative hypothesis**.

One-sided Alternative Hypotheses

$H_0 : \mu = 50$ centimeters per second

$H_1 : \mu < 50$ centimeters per second

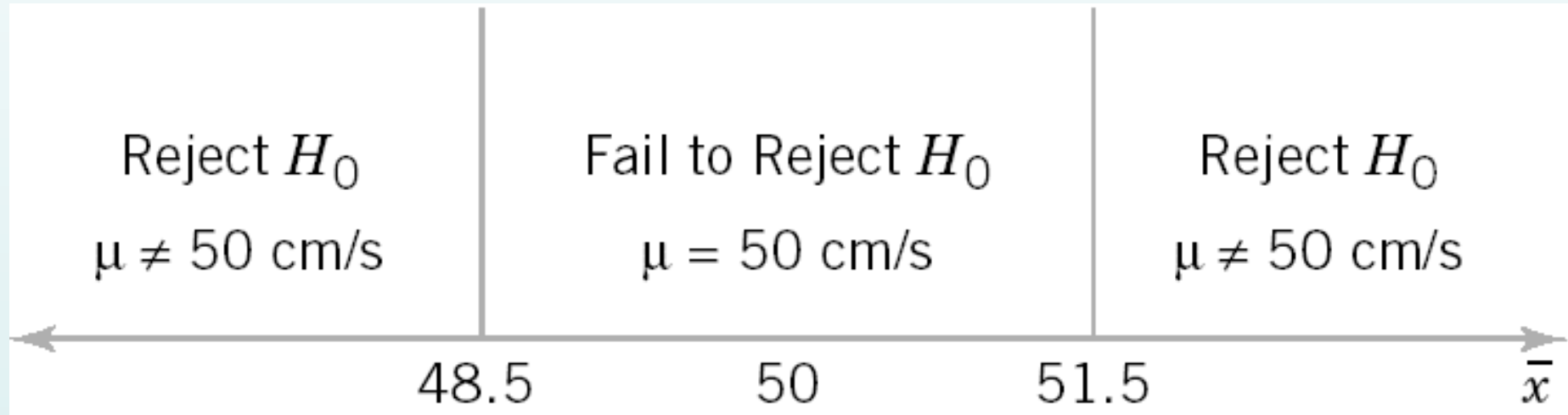
$H_0 : \mu = 50$ centimeters per second

$H_1 : \mu > 50$ centimeters per second

Tests of Statistical Hypotheses

$H_0 : \mu = 50$ centimeters per second

$H_1 : \mu \neq 50$ centimeters per second



Is the Dean right? (from Ch 9)

Salaries of a Business School's Graduates

In the advertisements for a large university, the dean of the School of Business *claims* that the average salary of the school's graduates one year after graduation is \$800 per week with a standard deviation of \$100.

A second-year student in the business school who has just completed her statistics course would like to check whether the claim about the mean is correct, that is, whether the dean is right.

Is the Dean right?

She surveys 25 people who graduated one year ago and finds the sample mean to be \$750 / week.

So, if the dean says \$800 / week but the sample of $n = 25$ shows a mean of \$750 / week, could she conclude that the dean is wrong?

Is the Dean right?

We want to find the probability that the sample mean is less than \$760. Thus, we seek

$$P(\bar{X} < 750)$$

The distribution of X , the weekly income, is likely to be positively skewed, but not sufficiently so to make the distribution of \bar{X} nonnormal. As a result, we may assume that \bar{X} is normal with mean

$$\mu_{\bar{X}} = \mu = 800$$

and standard deviation

$$\sigma_{\bar{X}} = \sigma / \sqrt{n} = 100 / \sqrt{25} = 20$$

Is the Dean right?

Thus,

$$\begin{aligned} & P(\bar{X} < 750) \\ &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{750 - 800}{20}\right) \\ &= P(Z < -2.5) \\ &= .5 - .4938 \\ &= .0062 \end{aligned}$$

The probability of observing a sample mean as low as \$750 when the population mean is \$800 is extremely small. Because this event is quite unlikely, we would have to conclude that the dean's claim is not justified.

Chapter-Opening Example SSA Envelope Plan

Federal Express (FedEx) sends invoices to customers requesting payment within 30 days. Currently the mean and standard deviation of the amount of time taken to pay bills are 24 days and 6 days, respectively.

The chief financial officer (CFO) believes that **including a stamped self-addressed (SSA) envelope** would decrease the amount of time. She calculates that the improved cash flow from a 2-day decrease in the payment period would pay for the costs of the envelopes and stamps. To test her belief she randomly selects 220 customers and includes a stamped self-addressed envelope with their invoices.

The numbers of days until payment is received were recorded. The average was 21.63. Can the CFO conclude that the plan will be profitable? If he does, could he be wrong?

Nonstatistical Hypothesis Testing

A criminal trial is an example of hypothesis testing without the statistics.

In a trial a jury must decide between two hypotheses. The null hypothesis is

H_0 : The defendant is innocent

The alternative hypothesis or research hypothesis is

H_1 : The defendant is guilty

The jury does not know which hypothesis is true. They must make a decision on the basis of evidence presented.

Concepts of Hypothesis Testing

There are **two** hypotheses. One is called the *null hypothesis* and the other the *alternative* or *research hypothesis*. The usual notation is:

pronounced
H "nought"

H_0 : — *the 'null' hypothesis*

H_1 : — *the 'alternative' or 'research' hypothesis*

The null hypothesis (H_0) will always state that the ***parameter equals the value*** specified in the alternative hypothesis (H_1)

Concepts of Hypothesis Testing

Two possible errors can be made in any test:

A Type I error occurs when we reject a true null hypothesis and

A Type II error occurs when we don't reject a false null hypothesis.

There are probabilities associated with each type of error:

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

$$P(\text{Type II error}) = \beta$$

α is called the *significance level*.

Types of Errors

A Type I error occurs when we *reject* a *true* null hypothesis (i.e. Reject H_0 when it is TRUE)

A Type II error occurs when we *don't reject* a *false* null hypothesis (i.e. Do NOT reject H_0 when it is FALSE)

Decision	H_0 is True	H_0 is False
Fail to reject H_0	No error	Type II error
Reject H_0	Type I error	No error

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$$

$$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false})$$

Type I error – how to compensate

DALLAS – A Texas man imprisoned for nearly two decades for the killings of a grandmother and five children that he didn't commit will receive \$1.4 million in compensation after Gov. Rick Perry signed a law with a provision specifically addressing his case.

Anthony Graves had been declared innocent by a special prosecutor last year in the 1992 killings of the six family members outside Houston. However, because of the wording of the order exonerating him, the 45-year-old former inmate has been unable to collect under a 2009 Texas law that gives exonerees \$80,000 for every year they spent in prison.

Graves said Wednesday he's grateful to Perry, but that the money "doesn't even come close" to making up for the time he spent in prison.

"I lost 18 years of my life," he said. "It wasn't like I hit the lottery."

After a federal appeals court overturned his conviction and ordered a new trial, the special prosecutor pronounced Graves innocent. But Graves was denied compensation by Texas Comptroller Susan Combs because the order detailing his exoneration lacks the phrase **"actual innocence."**

Example 11.1

The manager of a department store is thinking about establishing a new billing system for the store's credit customers.

She determines that the new system will be cost-effective only if the mean monthly account is more than \$170. A random sample of 400 monthly accounts is drawn, for which the sample mean is \$178.

The manager knows that the accounts are approximately normally distributed with a standard deviation of \$65. Can the manager conclude from this that the new system will be cost-effective?

Example 11.1

IDENTIFY

The system will be cost effective if the mean account balance for all customers is greater than \$170.

We express this belief as our research hypothesis, that is:

$$H_1: \mu > 170 \quad (\text{this is what we want to determine})$$

Thus, our null hypothesis becomes:

$$H_0: \mu = 170 \quad (\text{this specifies a single value for the parameter of interest})$$

Example 11.1

IDENTIFY

What we want to show:

$$H_0: \mu = 170 \text{ (we'll } *assume* \text{ this is true)}$$

$$H_1: \mu > 170$$

We know:

$$n = 400,$$

$$\bar{x} = 178, \text{ and}$$

$$\sigma = 65$$

What to do next?!

Example 11.1

COMPUTE

To test our hypotheses, we can use two different approaches:

The *rejection region* approach (typically used when computing statistics manually), and

The *p-value* approach (which is generally used with a computer and statistical software).

We will explore both in turn...

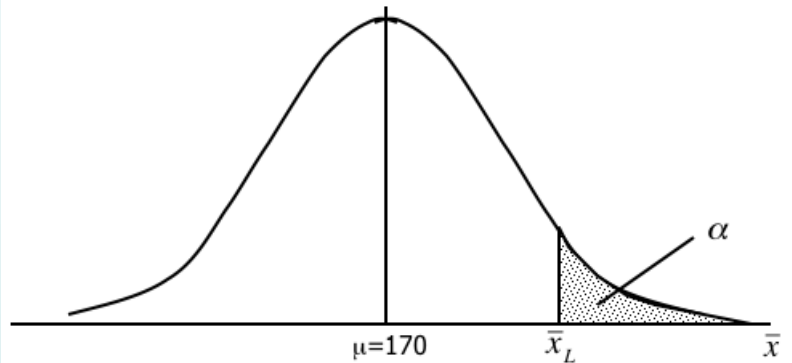
Example 11.1 Rejection region

It seems reasonable to reject the null hypothesis in favor of the alternative if the value of the sample mean is *large* relative to 170, that is if $\bar{x} > \bar{x}_L$.

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

$$= P(\text{reject } H_0 \text{ given that } H_0 \text{ is true})$$

$$\alpha = P(\bar{x} > \bar{x}_L)$$



Example 11.1

COMPUTE

All that's left to do is calculate \bar{x}_L and compare it to 170.

$$P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} > \frac{\bar{x}_L - \mu}{\sigma/\sqrt{n}}\right) = P\left(Z > \frac{\bar{x}_L - \mu}{\sigma/\sqrt{n}}\right) = \alpha$$
$$P(Z > z_\alpha) = \alpha$$
$$\therefore \frac{\bar{x}_L - \mu}{\sigma/\sqrt{n}} = z_\alpha$$

we can calculate this based on any level of significance (α) we want...

Example 11.1

COMPUTE

At a 5% significance level (i.e. $\alpha = 0.05$), we get

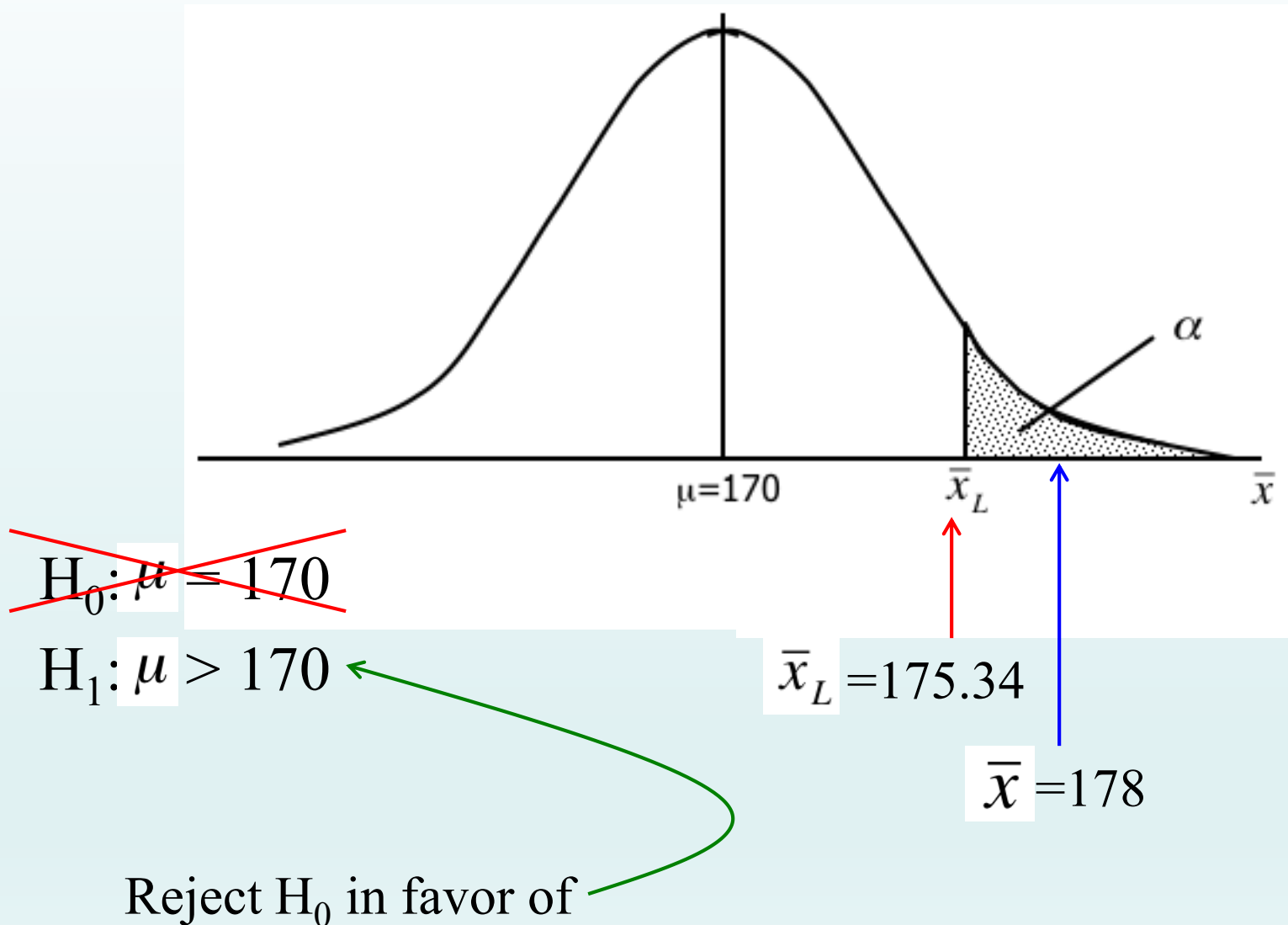
$$\frac{\bar{x}_L - \mu}{\sigma / \sqrt{n}} = z_{\alpha} \quad \& \quad z_{\alpha} = z_{.05} = 1.645$$

$$\text{gives: } \frac{\bar{x}_L - 170}{65 / \sqrt{400}} = 1.645$$

Solving we compute $\bar{x}_L = 175.34$

Since our sample mean (178) is ***greater than*** the critical value we calculated (175.34), we reject the null hypothesis in favor of H_1 , i.e. that: $\mu > 170$ and that it is cost effective to install the new billing system

Example 11.1 The Big Picture



Standardized Test Statistic

An easier method is to use the standardized test statistic:

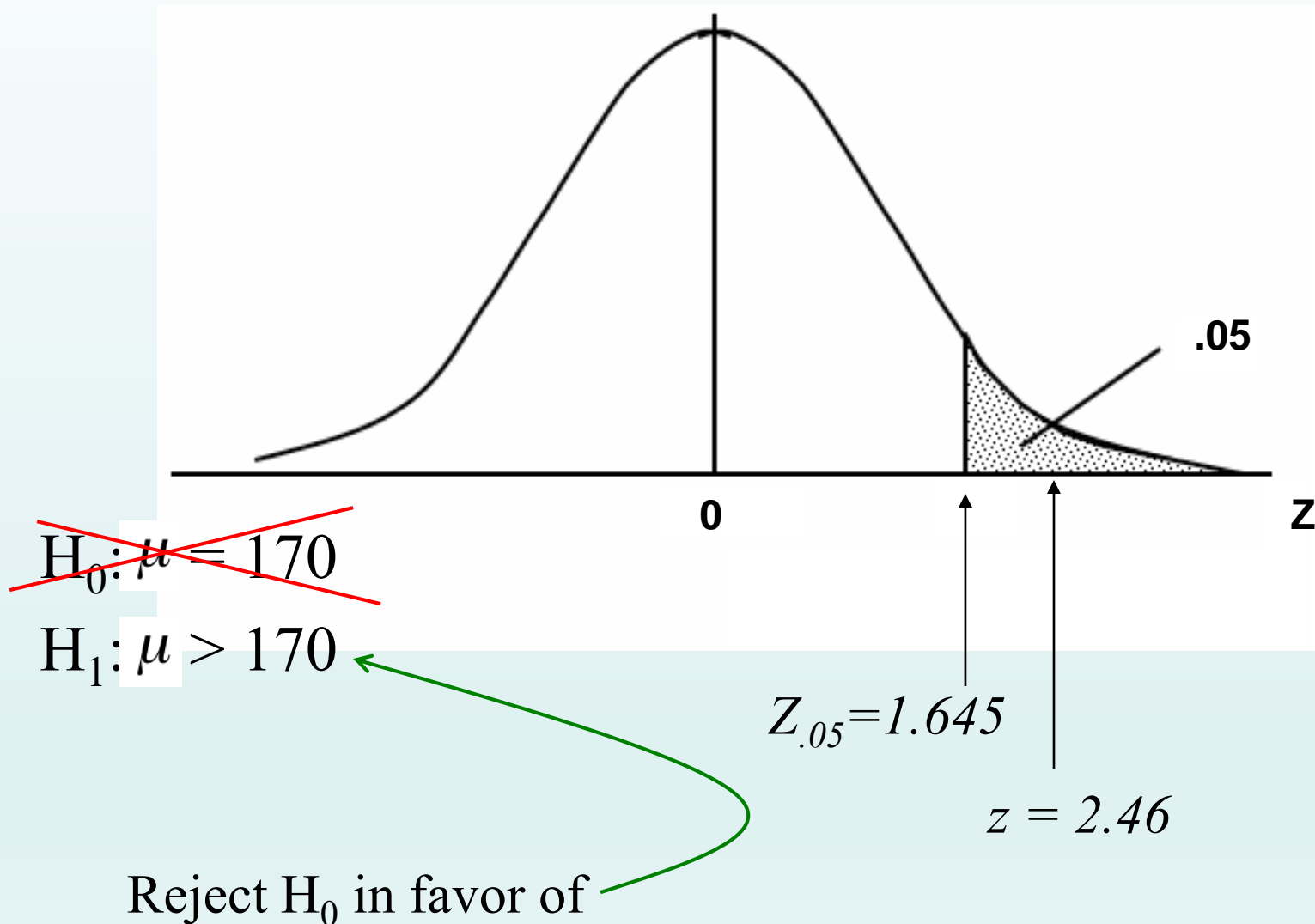
$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

and compare its result to z_{α} : (rejection region: $z > z_{\alpha}$)

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{178 - 170}{65 / \sqrt{400}} = 2.46$$

Since $z = 2.46 > 1.645 (z_{.05})$, we reject H_0 in favor of $H_1 \dots$

Example 11.1... The Big Picture Again



p-Value of a Test

The *p-value* of a test is the probability of observing a test statistic at least as extreme as the one computed given that the null hypothesis is true.

In the case of our department store example, what is the *probability* of observing a sample mean *at least as extreme* as the one already observed (i.e. $\bar{x} = 178$), given that the null hypothesis ($H_0: \mu = 170$) is true?

$$P(\bar{x} > 178) = P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} > \frac{178 - 170}{65/\sqrt{400}}\right) = P(Z > 2.46) = .0069$$

p-value

Interpreting the p-value

Compare the p-value with the selected value of the significance level:

If the p-value is less than α , we judge the p-value to be small enough to reject the null hypothesis.

If the p-value is greater than α , we do not reject the null hypothesis.

Since $p\text{-value} = .0069 < \alpha = .05$, we reject H_0 in favor of H_1

Interpreting the p-value

Overwhelming Evidence
(Highly Significant)

Strong Evidence
(Significant)

Weak Evidence
(Not Significant)

No Evidence
(Not Significant)

The **P-value** is the smallest level of significance that would lead to rejection of the null hypothesis H_0 with the given data.

P-value is the **observed significance level**.

0

.01

.05

.10

p=.0069

Propellant Burning Rate

Air crew escape systems are powered by a solid propellant. The burning rate of this propellant is an important product characteristic. Specifications require that the mean burning rate must be 50 centimeters per second and the standard deviation is $\sigma = 2$ centimeters per second. The significance level of $\alpha = 0.05$ and a random sample of $n = 25$ has a sample average burning rate of $\bar{x} = 51.3$ centimeters per second. Draw conclusions.

The seven-step procedure is

- 1. Parameter of interest:** The parameter of interest is μ , the mean burning rate.
- 2. Null hypothesis:** $H_0: \mu = 50$ centimeters per second
- 3. Alternative hypothesis:** $H_1: \mu \neq 50$ centimeters per second

Conclusions of a Test of Hypothesis

If we reject the null hypothesis, we conclude that there is enough evidence to infer that the alternative hypothesis is true.

If we do not reject the null hypothesis, we conclude that there is not enough statistical evidence to infer that the alternative hypothesis is true.

Remember: The alternative hypothesis is the more important one. It represents what we are investigating.

Chapter-Opening Example SSA Envelope Plan

Federal Express (FedEx) sends invoices to customers requesting payment within 30 days.

The bill lists an address and customers are expected to use their own envelopes to return their payments.

Currently the mean and standard deviation of the amount of time taken to pay bills are 24 days and 6 days, respectively.

The chief financial officer (CFO) believes that including a stamped self-addressed (SSA) envelope would decrease the amount of time.

Chapter-Opening Example SSA Envelope Plan

She calculates that the improved cash flow from a 2-day decrease in the payment period would pay for the costs of the envelopes and stamps.

Any further decrease in the payment period would generate a profit.

To test her belief she randomly selects 220 customers and includes a stamped self-addressed envelope with their invoices.

The numbers of days until payment is received were recorded. Can the CFO conclude that the plan will be profitable?

SSA Envelope Plan

IDENTIFY

The objective of the study is to draw a conclusion about the mean payment period. Thus, the parameter to be tested is the population mean.

We want to know whether there is enough statistical evidence to show that the population mean is less than 22 days. Thus, the alternative hypothesis is

$$H_1: \mu < 22$$

The null hypothesis is

$$H_0: \mu = 22$$

The test statistic is

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

We wish to reject the null hypothesis in favor of the alternative only if the sample mean and hence the value of the test statistic is small enough.

As a result we locate the rejection region in the left tail of the sampling distribution.

We set the significance level at 10%.

SSA Envelope Plan

COMPUTE

Rejection region: $z < -z_{\alpha} = -z_{.10} = -1.28$

From the data in [Xm11-00](#) we compute

and
$$\bar{x} = \frac{\sum x_i}{220} = \frac{4,759}{220} = 21.63$$

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{21.63 - 22}{6 / \sqrt{220}} = -.91$$

$$\text{p-value} = P(Z < -.91) = .5 - .3186 = .1814$$

SSA Envelope Plan

INTERPRET

Conclusion: There is not enough evidence to infer that the mean is less than 22.

There is not enough evidence to infer that the plan will be profitable.

Example 11.2

In recent years, a number of companies have been formed that offer competition to AT&T in long-distance calls.

All advertise that their rates are lower than AT&T's, and as a result their bills will be lower.

AT&T has responded by arguing that for the average consumer there will be no difference in billing.

Suppose that a statistics practitioner working for AT&T determines that the mean and standard deviation of monthly long-distance bills for all its residential customers are \$17.09 and \$3.87, respectively.

Example 11.2

He then takes a random sample of 100 customers and recalculates their last month's bill using the rates quoted by a leading competitor.

Assuming that the standard deviation of this population is the same as for AT&T, can we conclude at the 5% significance level that there is a difference between AT&T's bills and those of the leading competitor?

Example 11.2

IDENTIFY

The parameter to be tested is the mean of the population of AT&T's customers' bills based on competitor's rates.

What we want to determine whether this mean differs from \$17.09. Thus, the alternative hypothesis is

$$H_1: \mu \neq 17.09$$

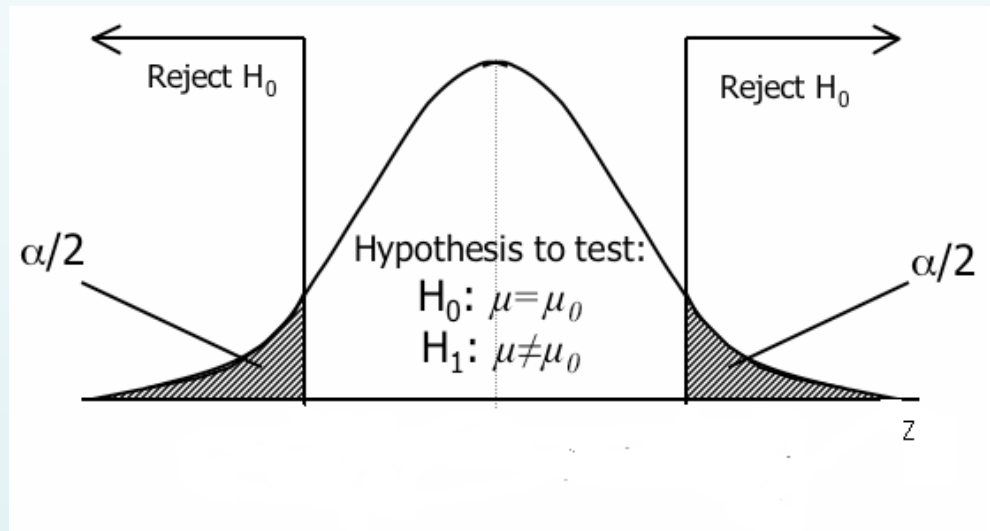
The null hypothesis automatically follows.

$$H_0: \mu = 17.09$$

Example 11.2

IDENTIFY

The rejection region is set up so we can reject the null hypothesis when the test statistic is large **or** when it is small.



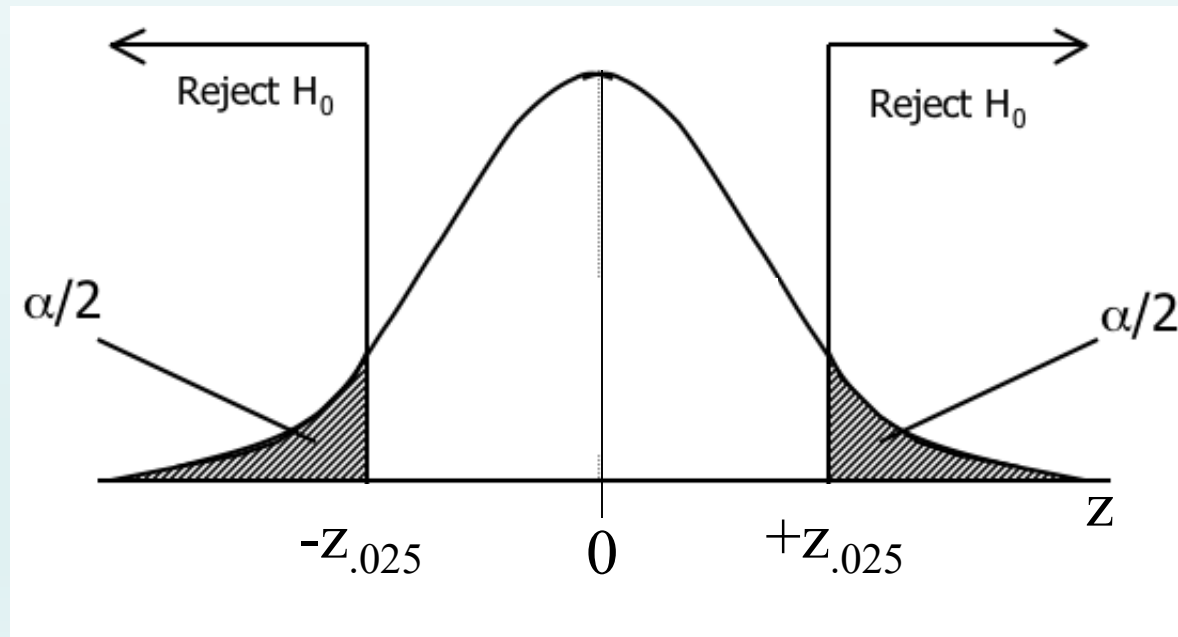
That is, we set up a two-tail rejection region. The total area in the rejection region must sum to α , so we divide this probability by 2.

Example 11.2

IDENTIFY

At a 5% significance level (i.e. $\alpha = .05$), we have $\alpha/2 = .025$. Thus, $z_{.025} = 1.96$ and our rejection region is:

$$z < -1.96 \quad \text{-or-} \quad z > 1.96$$



Example 11.2

COMPUTE

From the data ([Xm11-02](#)), we calculate $\bar{x} = 17.55$

Using our standardized test statistic: $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

We find that: $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{17.55 - 17.09}{3.87 / \sqrt{100}} = 1.19$

Since $z = 1.19$ is not greater than 1.96, nor less than -1.96 we cannot reject the null hypothesis in favor of H_1 . That is ***“there is insufficient evidence to infer that there is a difference between the bills of AT&T and the competitor.”***

Two-Tail Test p-value

COMPUTE

In general, the p-value in a two-tail test is determined by

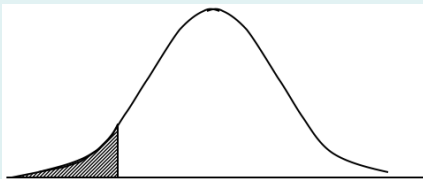
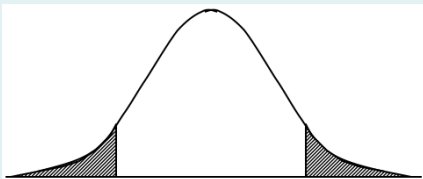
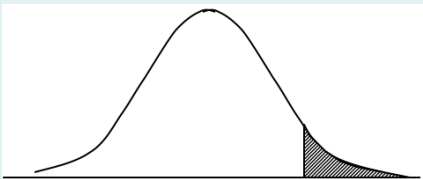
$$\text{p-value} = 2P(Z > |z|)$$

where z is the actual value of the test statistic and $|z|$ is its absolute value.

For Example 11.2 we find

$$\begin{aligned}\text{p-value} &= 2P(Z > 1.19) \\ &= 2(.1170) \\ &= .2340\end{aligned}$$

Summary of One- and Two-Tail Tests...

One-Tail Test (left tail)	Two-Tail Test	One-Tail Test (right tail)
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$
		

Probability of a Type II Error β

It is important that that we understand the relationship between Type I and Type II errors; that is, how the probability of a Type II error is calculated and its interpretation.

Recall Example 11.1...

$$H_0: \mu = 170$$

$$H_1: \mu > 170$$

At a significance level of 5% we rejected H_0 in favor of H_1 since our sample mean (178) was greater than the critical value of \bar{x} (175.34).

Probability of a Type II Error β

A Type II error occurs when a false null hypothesis is not rejected.

In example 11.1, this means that if \bar{x} is less than 175.34 (our critical value) we will **not reject** our null hypothesis, which means that we will not install the new billing system.

Thus, we can see that:

$$\beta = P(\bar{x} < 175.34 \text{ given that the null hypothesis is false})$$

Example 11.1 (revisited)

$\beta = P(\bar{x} < 175.34 \text{ given that the null hypothesis is false})$

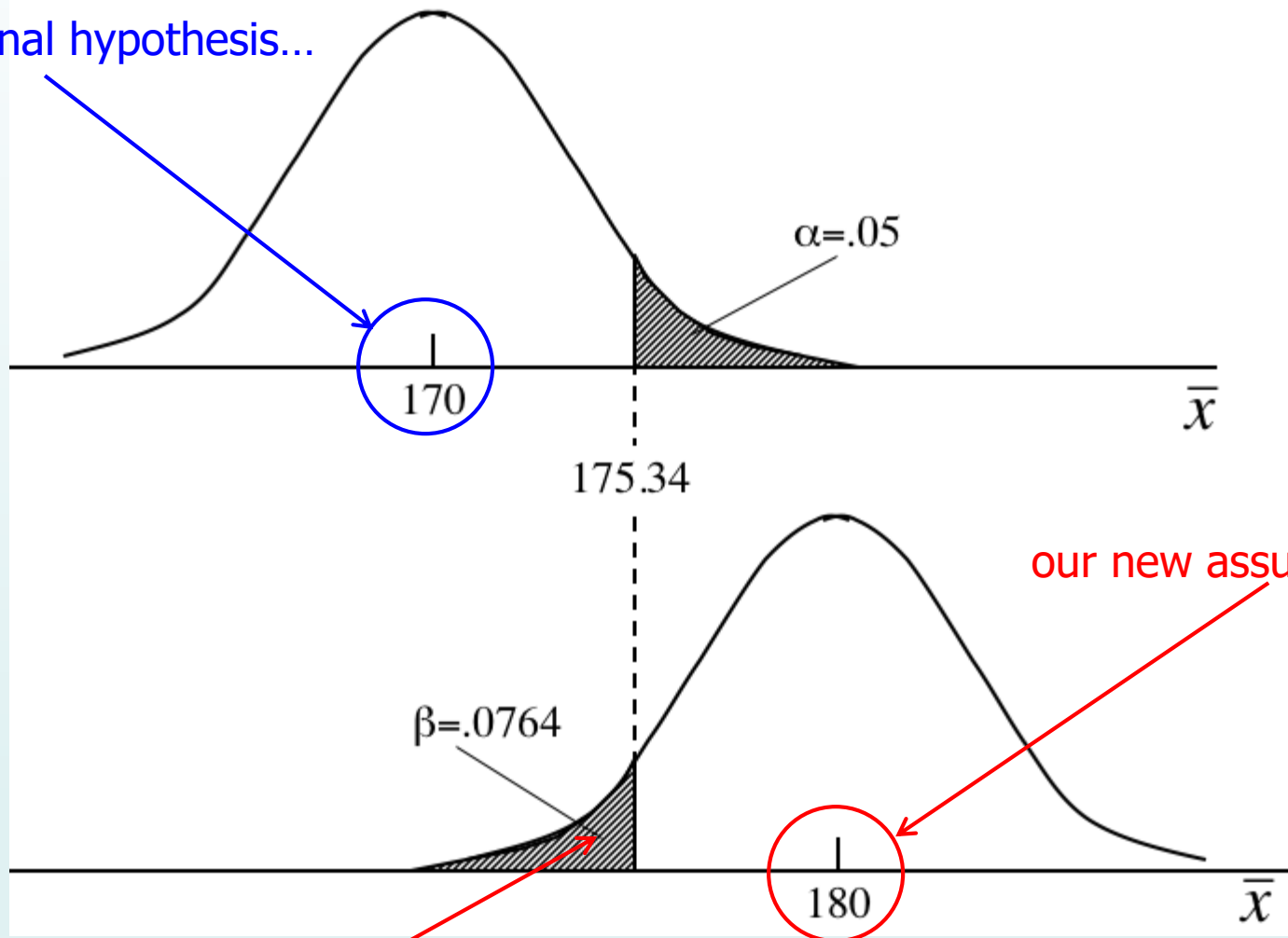
The condition only tells us that the mean $\neq 170$. We need to compute β for some new value of μ . For example, suppose that if the mean account balance is \$180 the new billing system will be so profitable that we would hate to lose the opportunity to install it.

$\beta = P(\bar{x} < 175.34, \text{ given that } \mu = 180), \text{ thus...}$

$$\beta = P\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < \frac{175.34 - 180}{65 / \sqrt{400}}\right) = P(Z < -1.43) = .0764$$

Example 11.1 (revisited)

Our original hypothesis...



$$\beta = P(\bar{x} < 175.34, \text{ given that } \mu = 180)$$

Effects on β of Changing α

Decreasing the significance level α , increases the value of β and vice versa. Change α to .01 in Example 11.1.

Stage 1: Rejection region

$$Z > Z_{\alpha} = Z_{.01} = 2.33$$

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{\bar{X} - 170}{65 / \sqrt{400}} > 2.33$$

$$\bar{X} > 177.57$$

Effects on β of Changing α

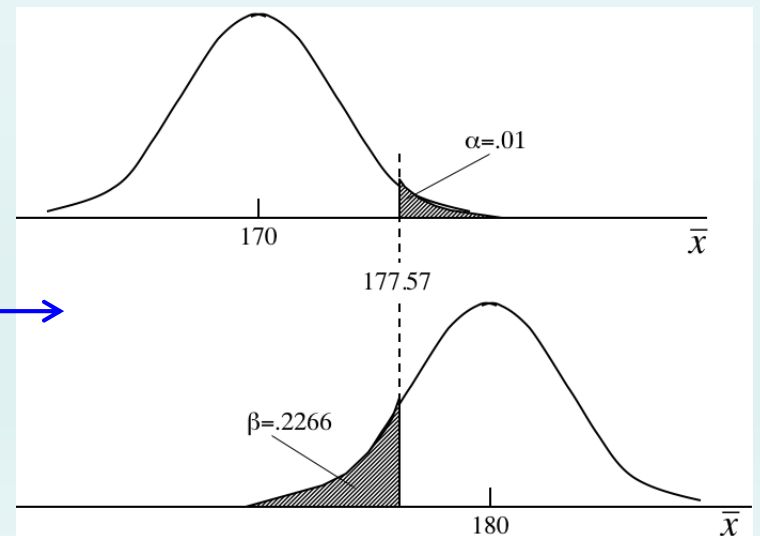
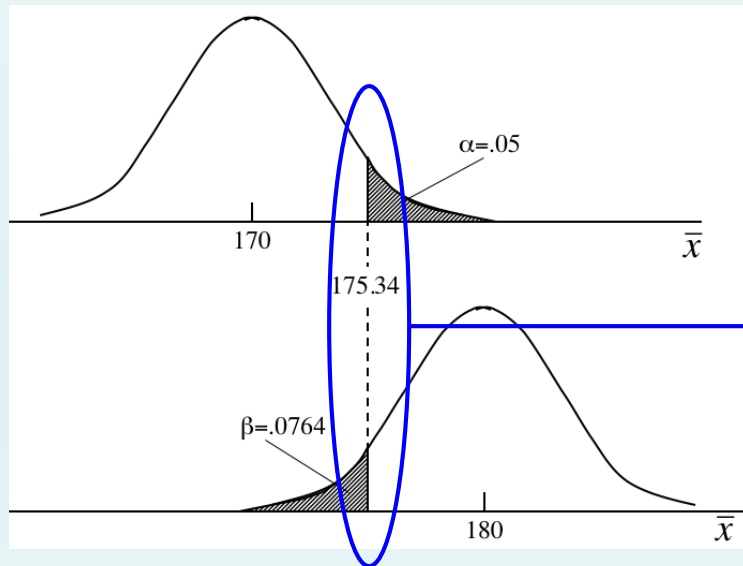
Stage 2 Probability of a Type II error

$$\begin{aligned}\beta &= P(\bar{x} < 177.57 \mid \mu = 180) \\ &= P\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < \frac{177.57 - 180}{65 / \sqrt{400}}\right) \\ &= P(z < -.75) \\ &= .2266\end{aligned}$$

Effects on β of Changing α

Decreasing the significance level α , increases the value of β and vice versa.

Consider this diagram again. Shifting the critical value line to the right (to decrease α) will mean a larger area under the lower curve for β ... (and vice versa)



Judging the Test

A statistical test of hypothesis is effectively defined by the significance level (α) and the sample size (n), *both of which are selected* by the statistics practitioner.

Therefore, if the probability of a Type II error (β) is judged to be too large, we can reduce it by

Increasing α ,
and/or
increasing the sample size, n .

Judging the Test

For example, suppose we increased n from a sample size of 400 account balances to 1,000 in Example 11.1.

Stage 1: Rejection region

$$z > z_{\alpha} = z_{.05} = 1.645$$

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{\bar{x} - 170}{65 / \sqrt{1,000}} > 1.645$$

$$\bar{x} > 173.38$$

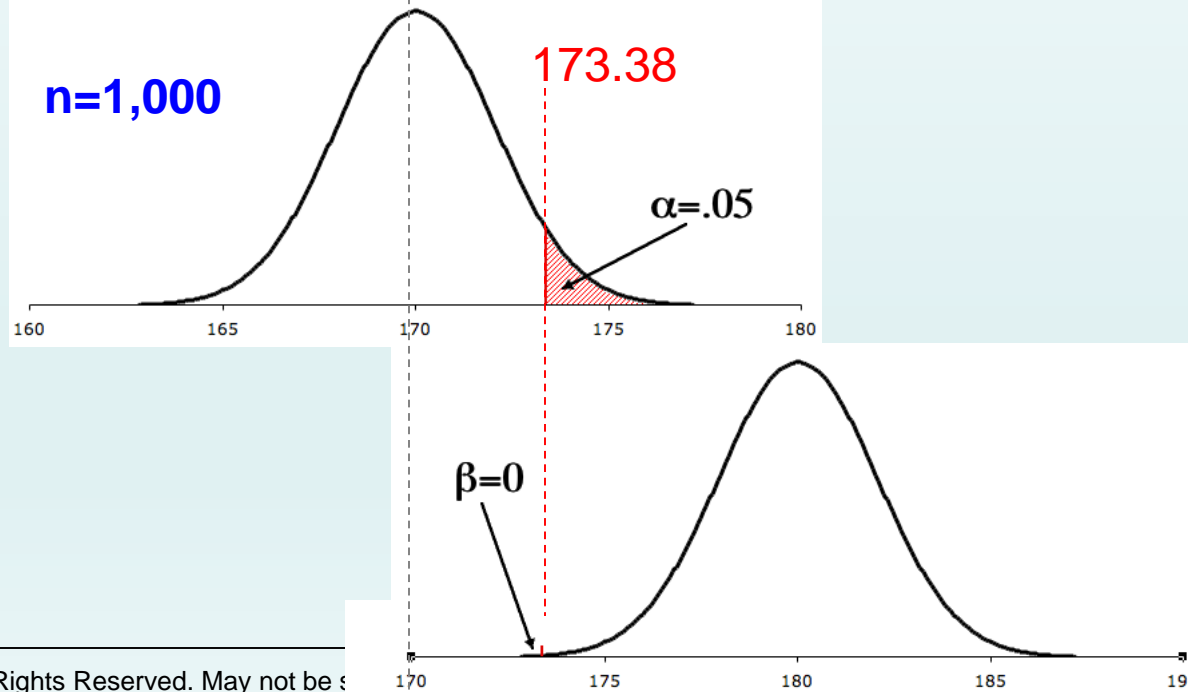
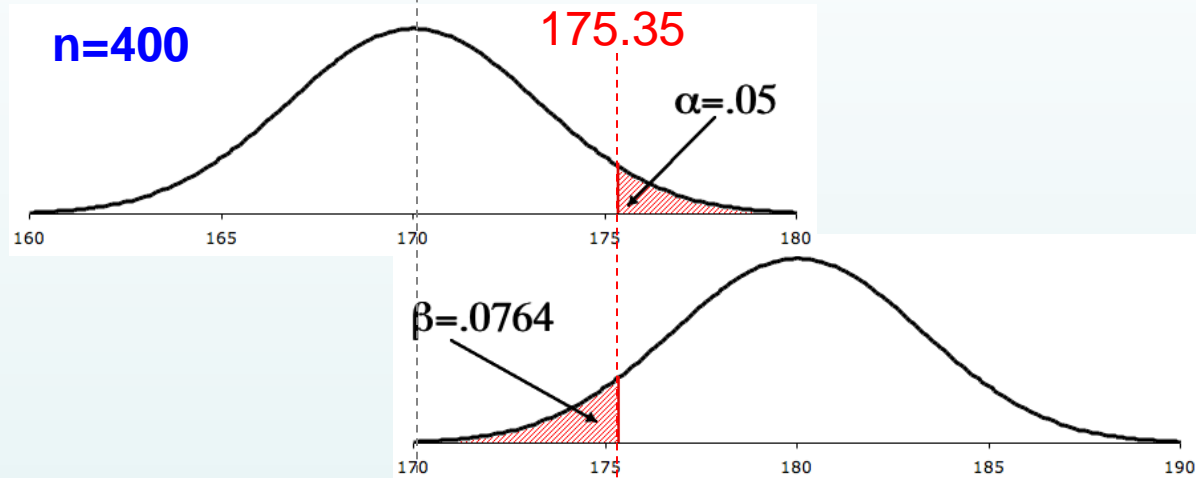
Judging the Test

Stage 2: Probability of a Type II error

$$\begin{aligned}\beta &= P(\bar{x} < 173.38 \mid \mu = 180) \\&= P\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < \frac{173.38 - 180}{65 / \sqrt{1,000}}\right) \\&= P(z < -3.22) \\&= 0 \text{ (approximately)}\end{aligned}$$

Compare β at $n=400$ and $n=1,000$...

By increasing the sample size we reduce the probability of a Type II error:



Developing an Understanding of Statistical Concepts

The calculation of the probability of a Type II error for $n = 400$ and for $n = 1,000$ illustrates a concept whose importance cannot be overstated.

By increasing the sample size we reduce the probability of a Type II error. By reducing the probability of a Type II error we make this type of error less frequently.

Judging the Test

The *power of a test* is defined as $1 - \beta$.

It represents the probability of rejecting the null hypothesis when it is false.

I.e. when more than one test can be performed in a given situation, it is preferable to use the test that is correct more often. If one test has a higher power than a second test, the first test is said to be more powerful and the preferred test.