Hypothesis Testing

Hypothesis Testing

- is the procedure used in inferential statistics to estimate population parameters based on sample data.
- Involves the use of statistical tests to determine the likelihood of certain population outcomes.

Hypothesis testing usually begins with a research question.

For Example,

Research Question: Suppose it is known that scores on a standardized test of reading comprehension for 4th graders is normally distributed with a $\mu = 70$ and $\sigma = 10$. A researcher wants to know if a new reading technique has an effect on comprehension. A random sample of n = 25 fourth graders is taught the technique and then tested for reading comprehension. A sample mean of M = 75 is obtained. Does the sample mean (M) differ enough from the population mean (μ) to conclude that the reading technique made a difference in level of comprehension?

Hypothesis Testing Steps

1. Formulate hypotheses.
2. Indicate the alpha level and determine critical values.
3. Calculate relevant statistics.
4. Make a decision and report the results.

Step 1: Formulate Hypotheses

There are two mutually exclusive hypotheses:

- The **null hypothesis** (H_0) attributes any differences between our obtained sample mean and the population mean to chance. It can variously be described as a statement of:

  - *chance* (i.e., any differences found are simply due to chance, or random sampling error).
  - *equality* (i.e., there is no true difference between our obtained statistic and the population parameter being predicted; they are essentially equal).
  - *ineffective treatment* (i.e., the independent variable had no effect).

- The **alternative hypothesis** (H_1) describes:

  - *true differences* rather than just chance differences.
  - *the effectiveness of treatment* of the independent variable.

For our example, the **null hypothesis** states that the new reading technique does not have an effect on comprehension, that there is no true difference between the mean of the population (μ) and the mean of the sample (M), and that any differences found are due simply to chance, or random sampling error.

$$H_0: \ \mu = 70$$

In essence, H_0 states that the new reading technique would not change the mean level of reading comprehension. The population mean (μ) would still be 70.

The **alternative hypothesis**, on the other hand, states that the new reading technique does have an effect on comprehension and that differences between M and μ are more than chance differences. For our example, the alternative hypothesis would be written as:

$$H_1: \mu \neq 70$$

H_1 predicts that the mean for reading comprehension *would* be different for the population of fourth graders who were taught to read using the new technique.

Directionality of the Alternative Hypothesis:

- A **nondirectional alternative hypothesis** merely states that the value of μ is something other than the value predicted in H_0.
- A **directional alternative hypothesis** specifies the "direction" of expected difference.

The alternative hypothesis is always written as the exact opposite of the null hypothesis and could “alternatively” be written as:

- H_1: $\mu \neq 70$ (nondirectional) (for H_0: $\mu = 70$)
- H_1: $\mu > 70$ (directional) (for H_0: $\mu \leq 70$)
- H_1: $\mu < 70$ (directional) (for H_0: $\mu \geq 70$)

In our research problem, we are using a nondirectional alternative hypothesis (H_1: $\mu \neq 70$).

A Brief Lesson in the Backwards Logic of Hypothesis Testing:

- Although it may seem backwards, it is the null hypothesis (H_0) that is tested and assumed to be true, but the hope is that it can be rejected, thereby giving indirect support to the alternative hypothesis (H_1).
- We attempt to show the falsity of the null hypothesis so that we can, in roundabout fashion, confirm what we believed to be true all along in our alternative hypothesis.

Accordingly, how do we know whether or not to reject our H_0 that $\mu = 70$, thereby giving support to our H_1 that $\mu \neq 70$? We must first determine what the sampling distribution of means would look like if the null hypothesis were true, if $\mu = 70$.

If $\mu = 70$, we know that the mean of the sampling distribution of means would also be 70.

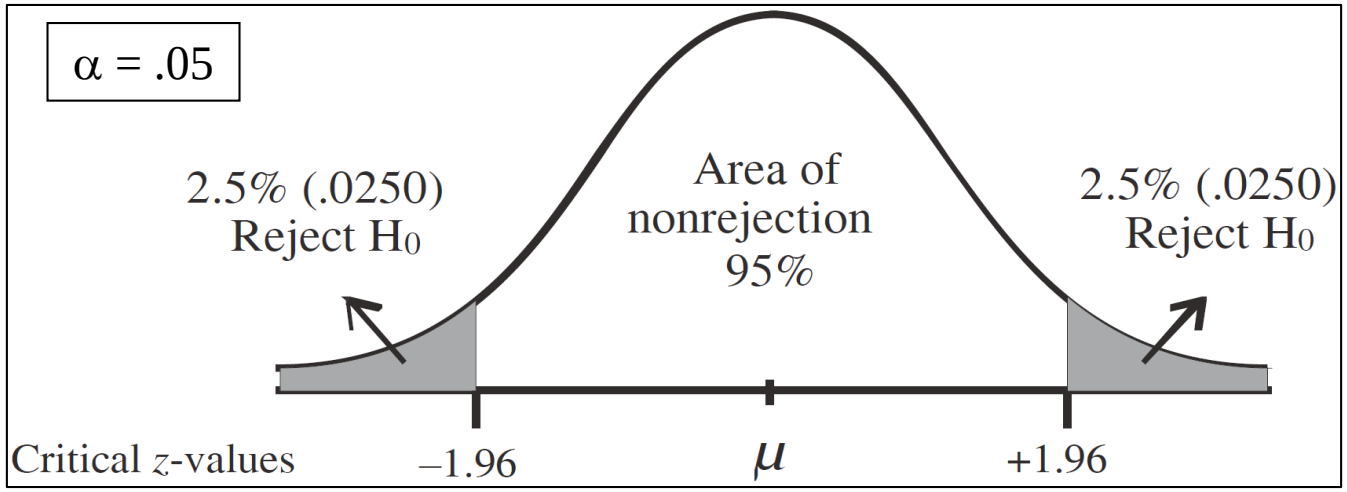
- Now determine the probability of obtaining sample means close to or far away from that value by looking in the z-table.
- If our obtained sample mean has a low probability of occurrence if H_0 is true (if $\mu = 70$), then we reject H_0.
- If our obtained sample mean has a high probability of occurrence, then we fail to reject H_0.

But first, we have to specify what we mean by a low or high probability of occurrence. We do this in Step 2.

Step 2: Indicate the Alpha Level and Determine Critical Values

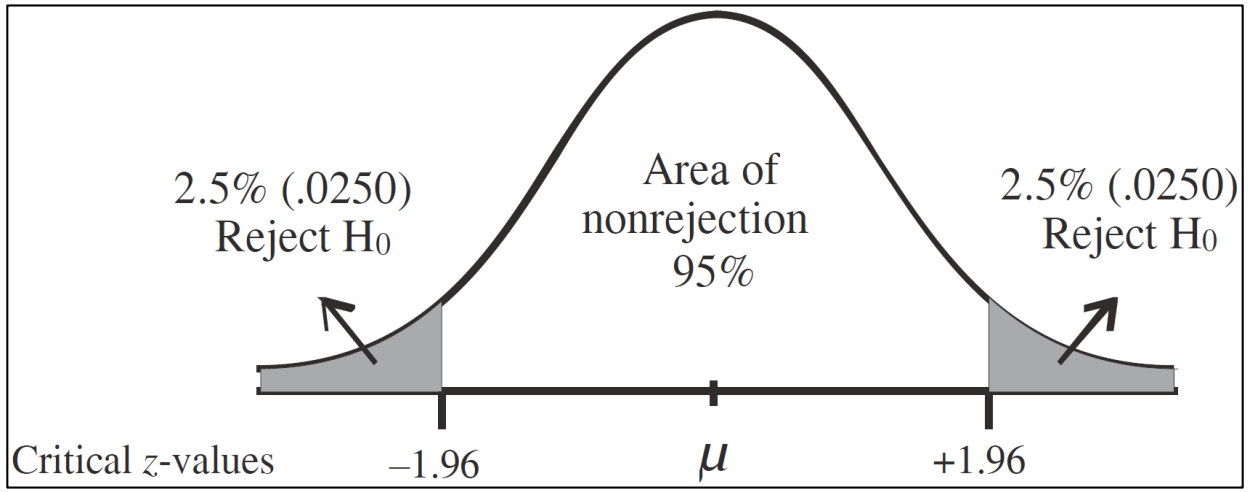
- The **alpha level** (α) is a probability level set by the researcher that defines the point at which H_0 should be rejected.
- Most often is set at .05 or .01, or .001.
- Also defines the point at which the critical region begins.

- **<u>Critical region</u>** – area of a sampling distribution in which score outcomes are highly unlikely.
- Remember, alpha levels are probability levels. Thus, when $\alpha = .05$ and if H_0 were true, the probability of obtaining a sample value that falls in the critical region would be .05 or less.
- Scores in this area are unlikely to have occurred by chance and therefore lead to rejection of H_0.
- This area of rejection is separated from the area of non-rejection by **<u>critical values</u>**, values that designate the point of rejection of H_0.



For Example,

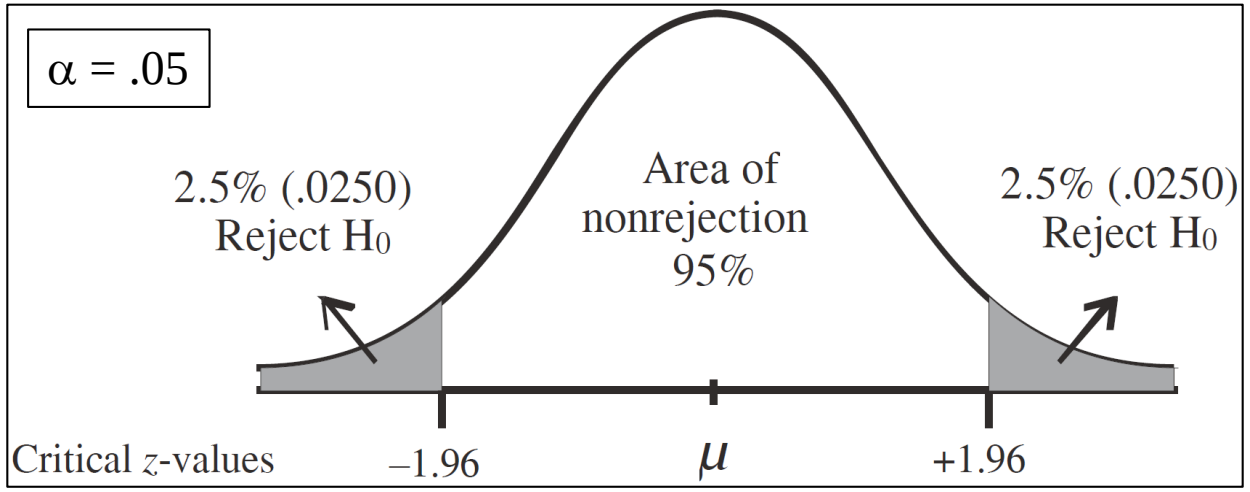
α = .05
Nondirectional

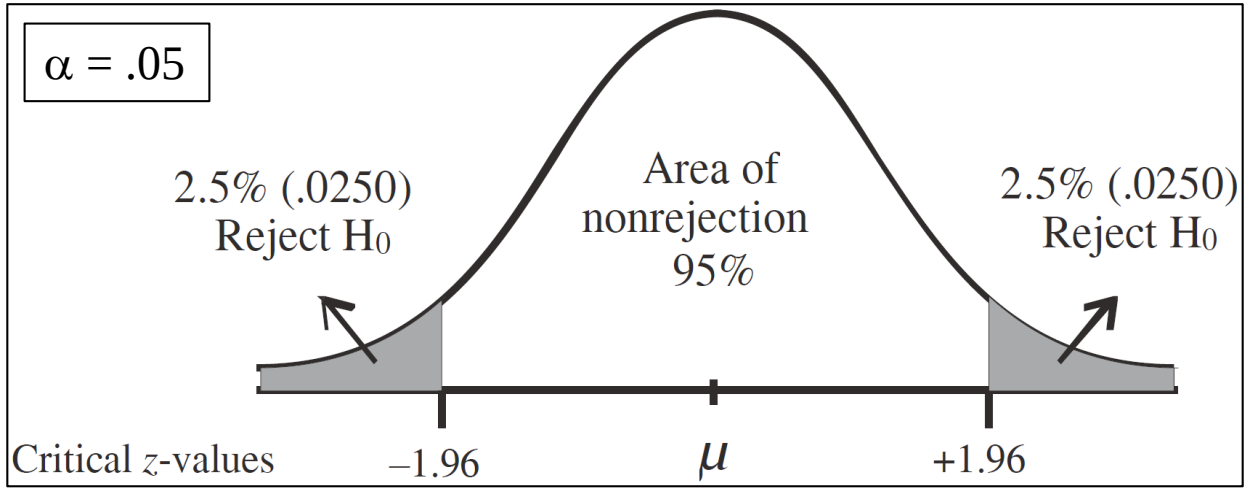


The shaded area represents the critical region which leads to rejection of H_0. Sample means that fall in these areas have a 5% or less likelihood of occurring by chance alone if H_0 is true.

This low probability makes H_0 unreasonable enough to be rejected.

- In this case, the alternative hypothesis (H_1) that scores this extreme (this far away from μ) are probably due to something other than chance makes more sense.
- "Something other than chance" refers to the effect of an independent variable.



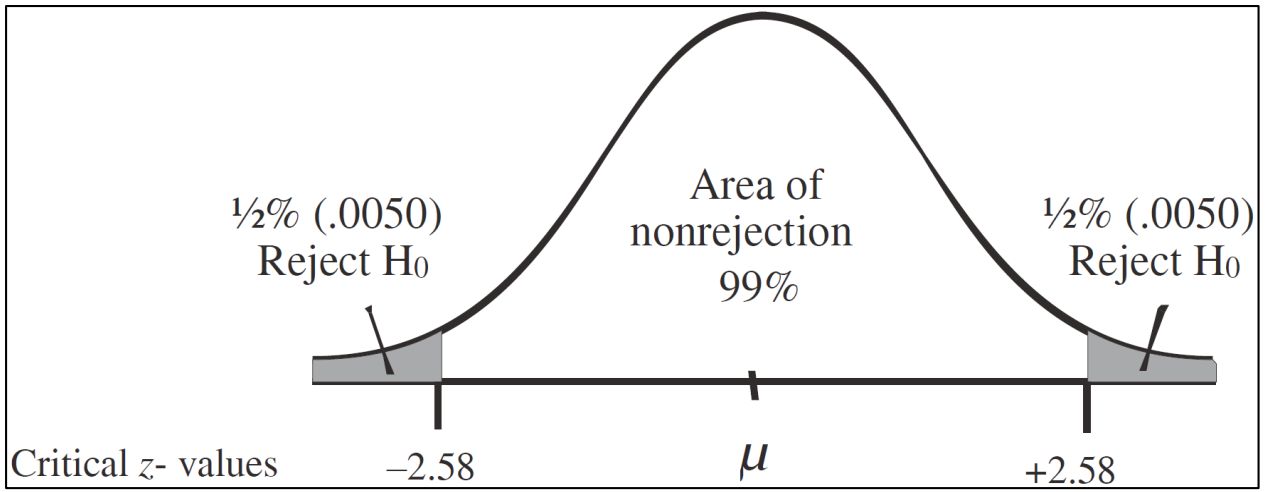


Since our alternative hypothesis is non-directional, the proportion associated with our alpha level of .05 is divided in two, with .0250 at each end. The critical values of ±1.96 separate these regions of rejection from the rest of the distribution.

Where do you find these critical values? Look in the z-table.

- These are the values associated with the extreme 2.5% (.0250) of the normal distribution curve (look in column C).

$\alpha = .01$
Nondirectional



The critical values that separate the areas of rejection for an alpha level of .01 are ±2.58. These are the z-scores associated with the extreme ½% (.0050) of the normal distribution curve.

When the alternative hypothesis is ***nondirectional***, use a two-tailed test.

- If a ***directional alternative hypothesis*** is specified, the entire rejection area is contained in one end of the distribution.

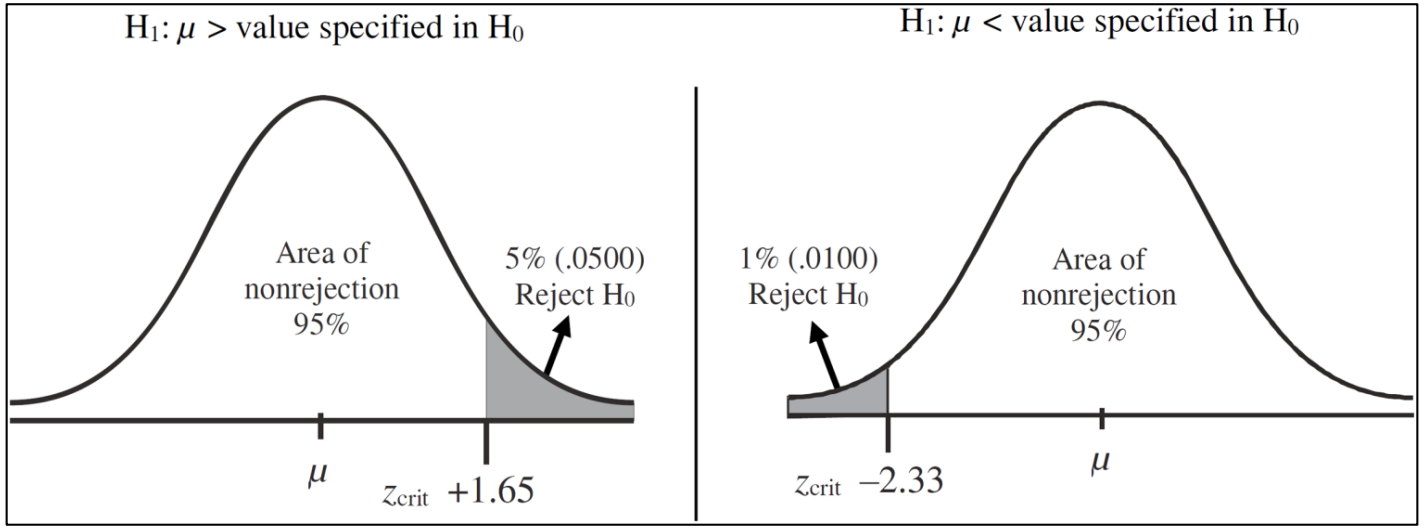
  - If H_1 specifies that μ would be greater than (>) a particular value, then the rejection region would be in the right tail only.

  - If H_1 specifies that μ would be less than (<) a particular value, then the rejection region would be in the left tail only.

  - Since interest is only in one tail of the distribution, the proportion is not divided in half.

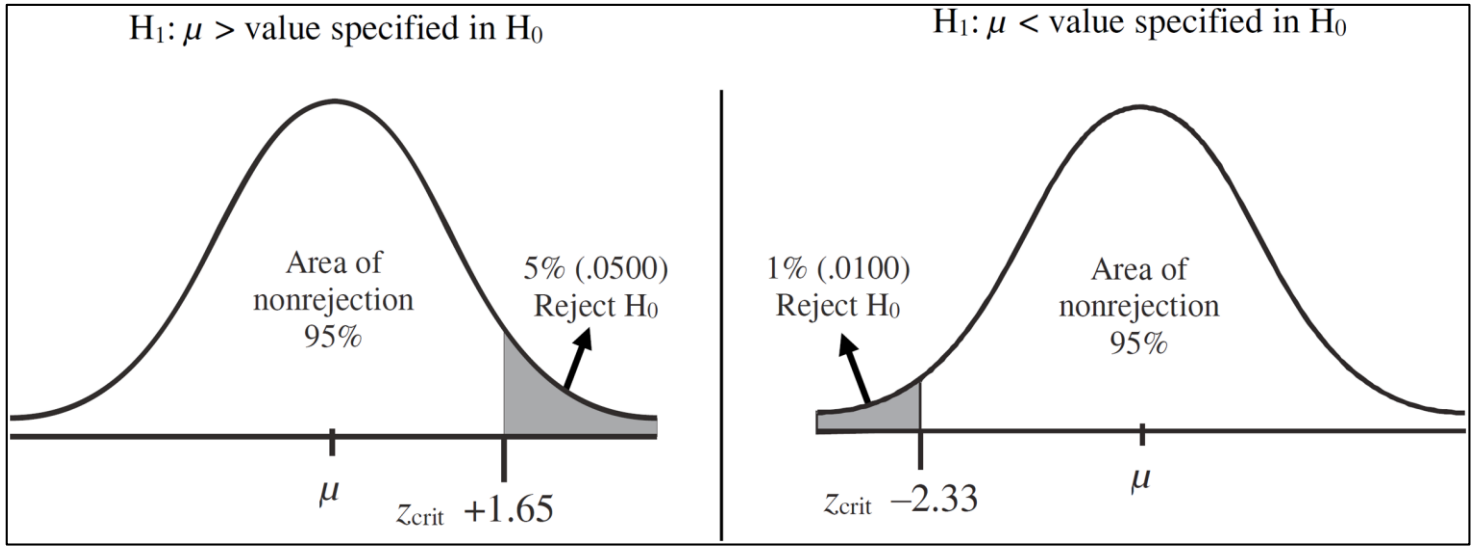
α = .05
Directional

α = .01
Directional

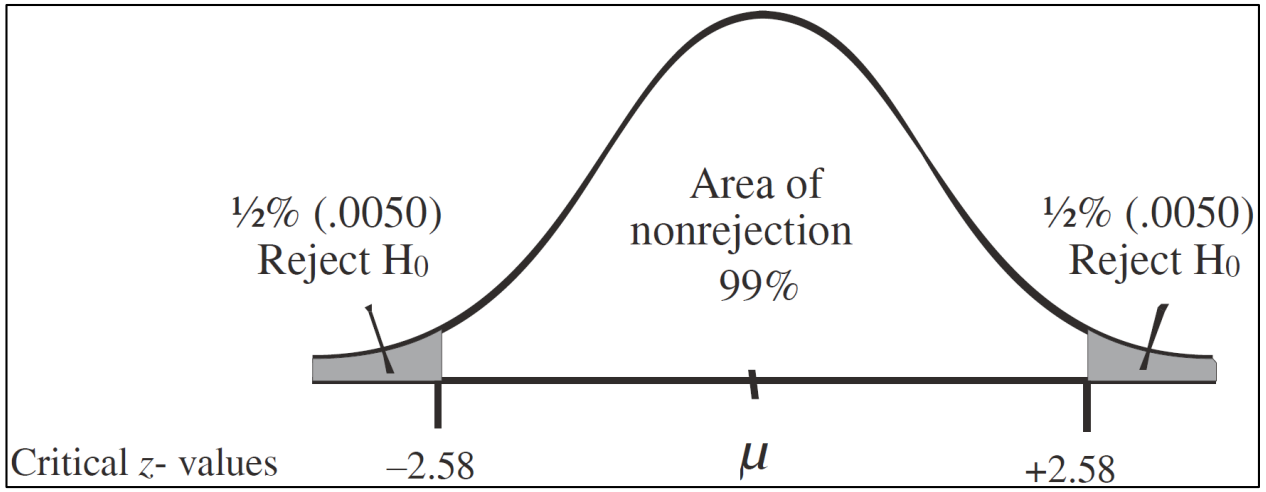


If you are using a one-tailed test, the critical value for rejecting H_0 (z_{crit}) will be only positive (+) or negative (−), depending on the direction specified in H_1.

α = .05, one-tailed **α = .01, one-tailed**

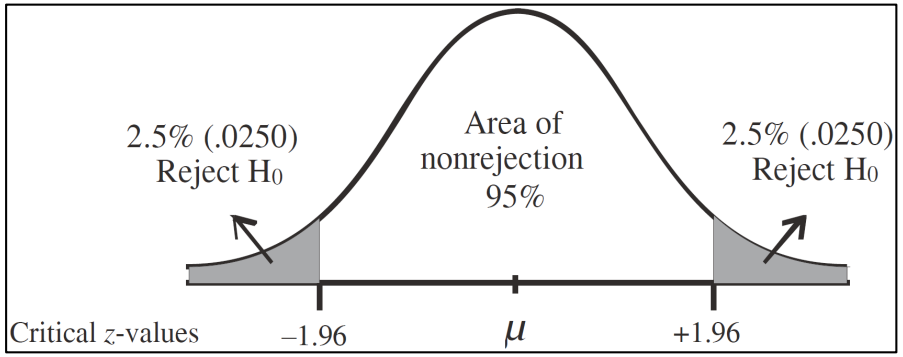


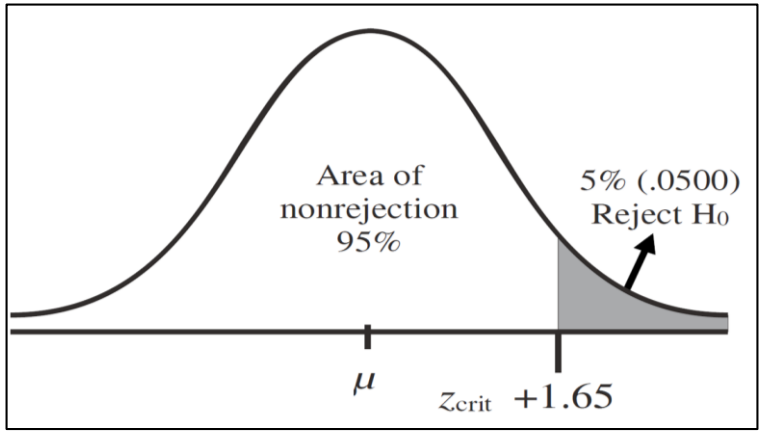
For a two-tailed test, the critical values to be considered will be both positive and negative (±)



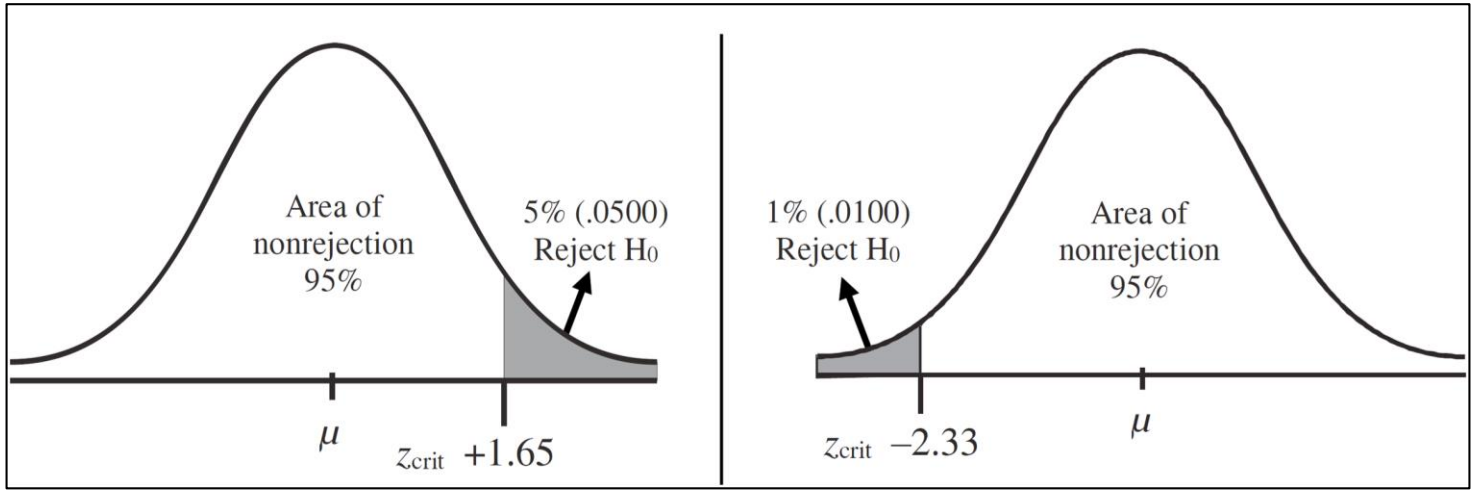
It is harder to reject H_0 using a two-tailed test because the rejection region is divided in two and distributed in both tails.

- The further into the tails we have to go, the more extreme our obtained mean has to be to fall in the rejection region.



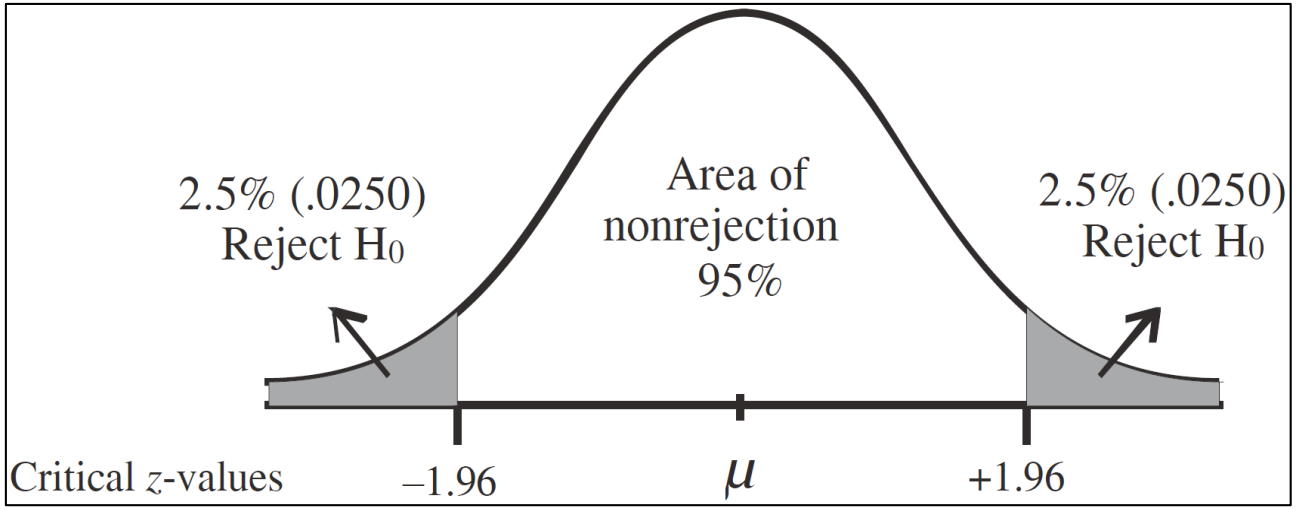


Likewise, an alpha level of .01 is more rigorous than an alpha level of .05 because larger differences would be necessary to reject H_0, showing greater effects.



Back to our example,

We will use an $\alpha = .05$. In step 1, we used the nondirectional alternative hypothesis $H_1: \mu \neq 70$. Thus, we will be conducting a two-tailed test and our critical values are ± 1.96.



Step 3: Calculate Relevant Statistics

Let's examine our data and calculate the "relevant statistics."

$\mu = 70$
$\sigma = 10$
$n = 25$
$M = 75$

The relevant statistics in this case are the standard error of the mean and a z-score.

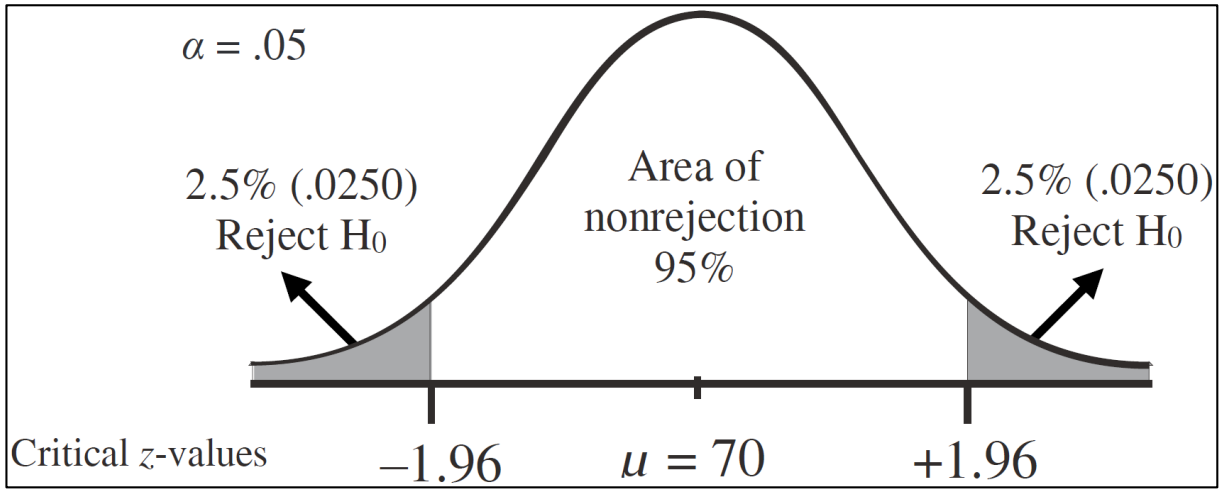
$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$

$$z_{obt} = \frac{M - \mu}{\sigma_M} = \frac{75 - 70}{2} = +2.50$$

Step 4: Make A Decision and Report The Results

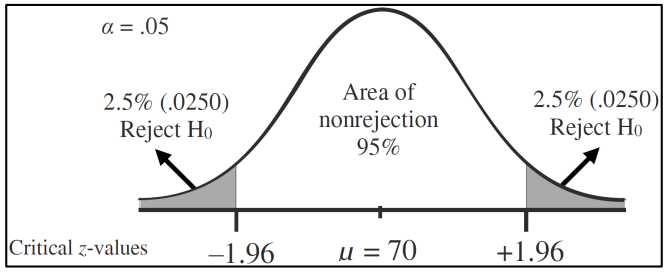
Finally, we must decide whether or not to reject the null hypothesis that H_0: $\mu = 70$. We are using an $\alpha = .05$ and a non-directional alternative hypothesis.

Sampling distribution for our example:



For our example, the z_{crit} values are ±1.96, meaning that in order to reject H_0, we would need a sample mean with an obtained z-score value beyond 1.96 (in either direction). Our obtained z-score was +2.50, which falls in the rejection region. Thus, we will reject H_0: $\mu = 70$.

We will reject the null hypothesis because if it were true that $\mu = 70$, there would only be a 5% or less chance that we would have obtained a sample mean as far from that value as the one we obtained (i.e., M = 75). The alternative hypothesis that H_1: $\mu \neq 70$ for the population of fourth graders using the new reading technique makes more sense. We conclude therefore that the new reading technique did have an effect on reading comprehension.



Results Format.

The new reading program had a significant effect on reading comprehension. Reject H_0, $z_{obt} = +2.50$, $p < .05$.

Notice that:

- "Significant" means greater than chance.
- The z-score that is reported is the *obtained* value of z, rather than the z_{crit} value.
- The "p" stands for probability. Here, if H_0 were true, the probability would be less than .05 of obtaining a sample mean that falls in the critical region. Use the alpha level (α) that has been pre-set by the researcher as the "p" value.
  - If z_{obt} is less extreme than z_{crit}, *p* will be greater than (>) the alpha level and we will "fail to reject" H_0."
  - If z_{obt} is more extreme than z_{crit}, *p* will be less than (<) the alpha level and we will "reject H_0."

Summary of Problem

Research question: Given: $\mu = 70$ and $\sigma = 10$ for reading comprehension, does the new reading technique have an effect on comprehension?

Step 1: Formulate hypotheses:

H_0: $\mu = 70$
H_1: $\mu \neq 70$

Step 2: Indicate the alpha level and determine critical values:

$\alpha = .05$
$Z_{crit} = \pm 1.96$

Step 3: Calculate relevant statistics:

Given: $n = 25$ $\mu = 70$ $\sigma = 10$ $M = 75$

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2 \qquad z_{obt} = \frac{M - \mu}{\sigma_M} = \frac{75 - 70}{2} = +2.50$$

Step 4: Make a decision and report the results:

The new reading program had a significant effect on reading comprehension. Reject H_0, $z_{obt} = +2.50$, $p < .05$.

Effect Size

If we reject H_0, we conclude that treatment was effective.

- But, how effective was it?
- In other words, what was the magnitude of the effect?

Cohen's *d* is a measure of effect size that involves comparing mean differences and dividing by the standard deviation.

For our research problem,

$$d = \frac{|M - \mu|}{\sigma} \qquad d = \frac{|75 - 70|}{10} = .5$$

The following guidelines are used for judging the size of d:

d = .20 to .49 – Small effect
d = .50 to .79 – Moderate effect
d = .80 and above – Large effect

For our problem, what this value tells us is that our effect size is one-half of a standard deviation, a moderate effect.

Assumptions

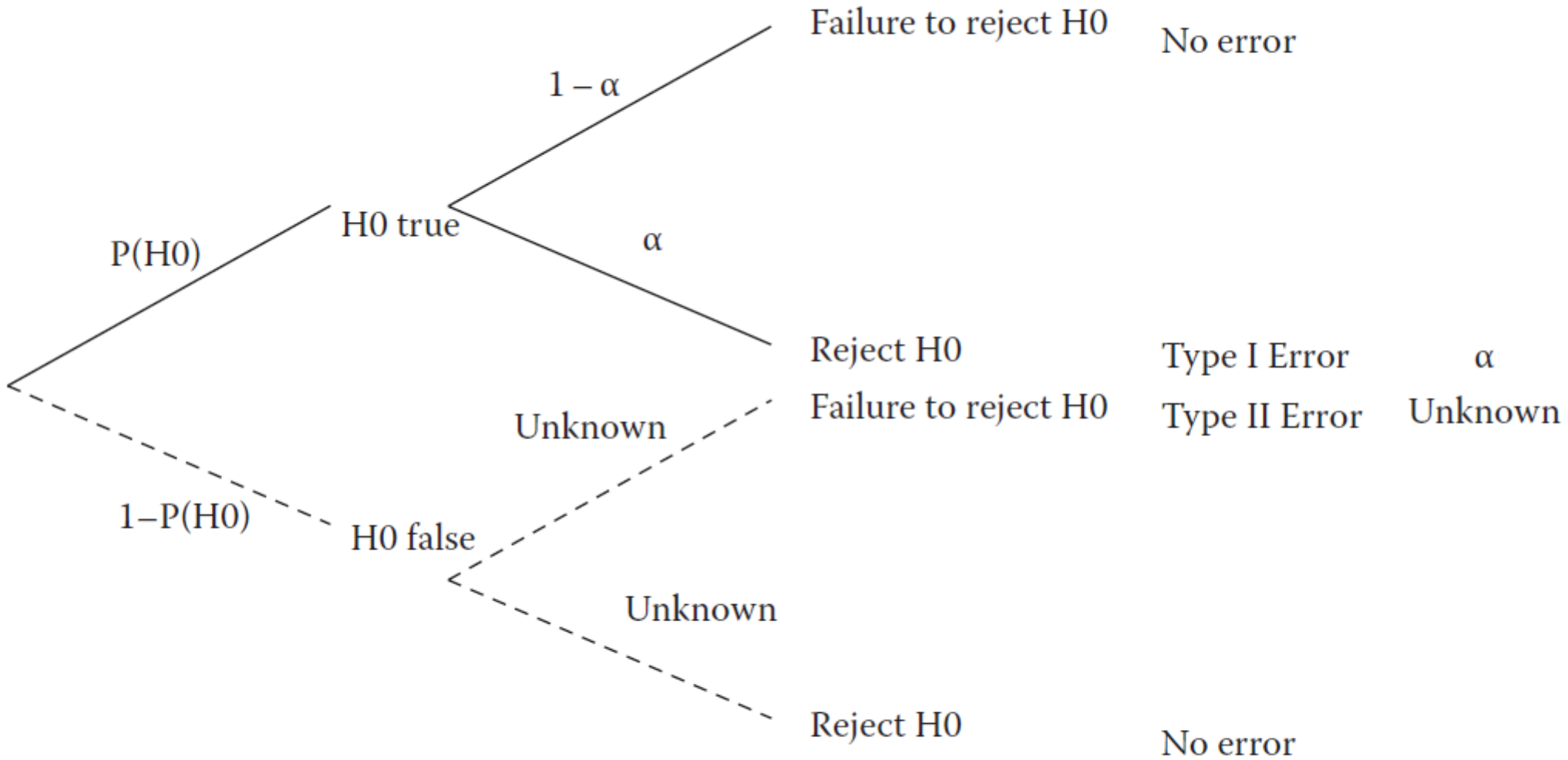
Statistical procedures that test hypotheses about populations are called **parametric tests**. Such tests should live up to certain assumptions in order for the results to be well-founded. Assumptions for the z-test are as follows:

- Independent and random selection of subjects.
- Normal distribution of dependent variable in the population.
- σ is known.

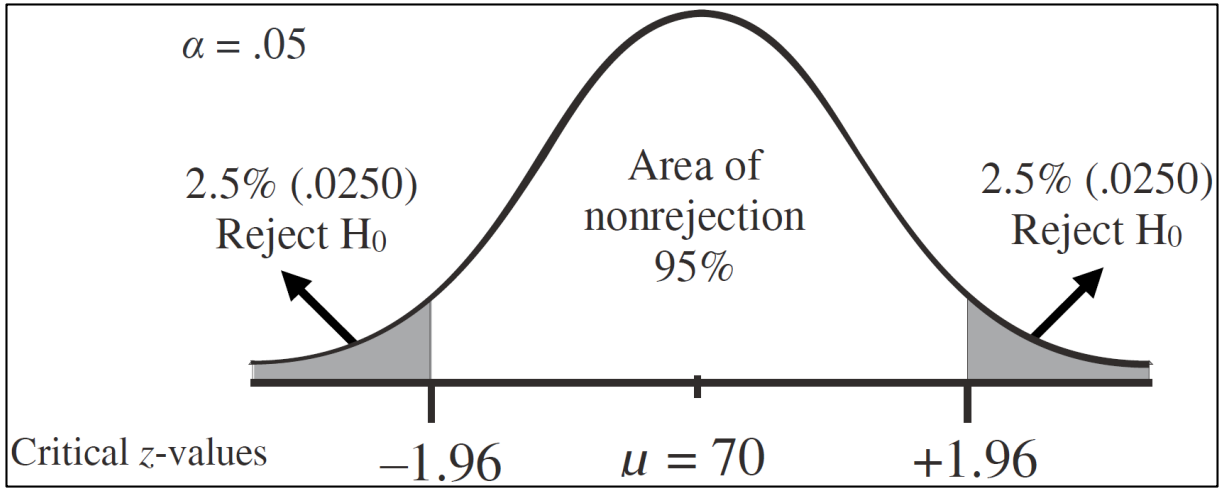
Errors

Two types of error are possible whenever we make a decision to either reject or not to reject H_0.

- A **Type I error** is rejecting a null hypothesis that is in reality true (i.e., saying that treatment had an effect when it actually didn’t).
- A **Type II error** is the failure to reject a null hypothesis that is in reality false (i.e., saying that treatment did not have an effect when it actually did).



Decision tree for hypothesis testing.



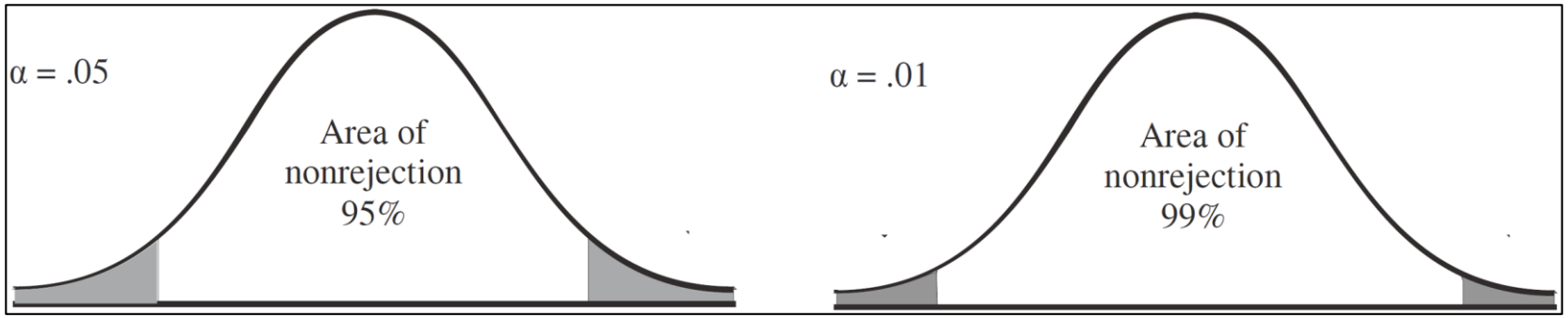
When we reject H_0 using an $\alpha = .05$, we are saying that our obtained statistic is unlikely to have occurred by chance alone (i.e., it would have occurred less than 5% of the time) if H_0 were true.

Such an unlikely event leads us to reject H_0 and to conclude instead that our independent variable had an effect.

However, we could be wrong (Type I error). In fact, we know that using a .05 alpha level will result in scores in the rejection region 5% of the time just by chance alone (without the influence of an independent variable).

Thus, the probability of obtaining a Type I error is defined by the alpha level.

- We could lower the probability of making a Type I error by lowering the alpha level, say to .01.
- Using $\alpha = .01$ reduces the rejection region and we would therefore be rejecting H_0 less often.
- However, using an $\alpha = .01$ rather than an $\alpha = .05$ simultaneously increases the probability of a Type II error – failing to reject a false null hypothesis.



- Because we don’t test actual populations, we have to rely on samples which introduces the possibility of error.
- We never know whether an error has been committed but we do want to keep the *risk* of error at a minimum. If we reject a null hypothesis, we want to be relatively certain that our results were real and not due to sampling error.
- The possibility of error associated with an alpha level of .05 is usually the greatest amount of risk that a researcher is willing to accept.
- If the consequences of making a Type I error are serious, then a lower alpha level may be chosen.
- But this will also have to be weighed against the seriousness of making a Type II error.

For Example,

Suppose a cardiologist had reason to believe that a tablespoon of sugar taken within ten minutes following a stroke would decrease the possibility of paralysis. Since a tablespoon of sugar is unlikely to cause serious complications (even if the doctor is wrong) and since the potential benefit would be great if it did work, the doctor would probably want to minimize the possibility of not discovering the effectiveness of this intervention. Thus, she would want to minimize the possibility of a Type II error. In this case, an alpha level of .05 might be chosen, making it easier to reject H_0.

Another Example,

Assume another researcher believes that moderation drinking helps prevent relapse in people who struggle with alcoholism. In this case, the researcher would want to take care to minimize the possibility of a Type I error – rejecting a true H_0. It would be a sobering mistake to say that moderation drinking is effective if in fact it leads to more serious alcohol problems. Thus, a more stringent alpha level (.01, or even .001) would likely be chosen, making it harder to reject H_0.

Researchers attempt to weigh the consequences and reach a balance between the two types of errors. Alpha levels of .05, .01, and .001 are commonly accepted as achieving this aim.

Power

- The ability of a statistical test to reject a null hypothesis that is in fact false.
- Makes it more likely that a treatment that actually exists will be detected.

- **Alpha (α) level**. Choosing a less rigorous alpha level, such as .05 instead of .01, will increase power.
- **Directionality of the alternative hypothesis**. A one-tailed test is more likely to detect statistical significant because the rejection region is larger. If a two-tailed test is used, the rejection region is split in half with each half moving further into the tails of the distribution, making it more difficult to reject H_0.
- **Sample size**. Larger samples generally produce less sampling error, which will result in larger obtained test statistics.
- **Variability of the data**. Distributions that have less variability will produce a smaller amount of standard error, resulting in larger obtained test statistics.