# Introduction to Hypothesis Testing

**OPRE 6301**

# Motivation . . .

The purpose of hypothesis testing is to determine whether there is enough statistical evidence in favor of a certain belief, or hypothesis, about a parameter.

Examples:

Is there statistical evidence, from a random sample of potential customers, to support the hypothesis that more than 10% of the potential customers will purchase a new product?

Is a new drug effective in curing a certain disease? A sample of patients is randomly selected. Half of them are given the drug while the other half are given a placebo. The conditions of the patients are then measured and compared.

These questions/hypotheses are similar in spirit to the discrimination example studied earlier. Below, we provide a basic introduction to hypothesis testing.

# Criminal Trials . . .

The basic concepts in hypothesis testing are actually quite analogous to those in a criminal trial.

Consider a person on trial for a “criminal” offense in the United States. Under the US system a jury (or sometimes just the judge) must decide if the person is innocent or guilty while in fact the person may be innocent or guilty. These combinations are summarized in the table below.

| | | Person is: | |
|---|---|---|---|
| | | Innocent | Guilty |
| Jury Says: | Innocent | No Error | Error |
| | Guilty | Error | No Error |

Notice that there are two types of errors. Are both of these errors *equally* important? Or, is it *as bad* to decide that a guilty person is innocent and let them go free *as it is* to decide an innocent person is guilty and punish them for the crime? Or, is a jury supposed to be totally *objective*, not assuming that the person is either innocent or guilty and make their decision based on the weight of the evidence one way or another?

In a criminal trial, there actually is a **favored assumption**, an initial bias if you will. The jury is *instructed* to assume the person is innocent, and only decide that the person is guilty if the evidence convinces them of such.

When there is a favored assumption, the *presumed innocence* of the person in this case, and the assumption is true, but the jury decides it is false and declares that the person is guilty, we have a so-called **Type I error**.

Conversely, if the favored assumption is false, i.e., the person is really guilty, but the jury declares that it is true, that is that the person is innocent, then we have a so-called **Type II error**.

Thus,

Favored Assumption: Person is Innocent

| | | Person is: | |
|---|---|---|---|
| | | Innocent | Guilty |
| Jury Says: | Innocent | No Error | Type II Error |
| | Guilty | Type I Error | No Error |

In some countries, the favored assumption is that the person is guilty. In this case the roles of the Type I and Type II errors would *reverse* to yield the following table.

Favored Assumption: Person is Guilty

| | | Person is: | |
|---|---|---|---|
| | | Innocent | Guilty |
| Jury Says: | Innocent | No Error | Type I Error |
| | Guilty | Type II Error | No Error |

Let us assume that the favored assumption is that the person is innocent. Assume further that in order to declare the person guilty, the jury must find that the evidence convinces them *beyond a reasonable doubt.*

Let

$$\begin{aligned} \alpha &= P(\text{Type I Error}) \\ &= P(\text{Jury Decides Guilty} \mid \text{Person is Innocent}) \,. \end{aligned}$$

Then, "beyond a reasonable doubt" means that we should keep $\alpha$ small. For the alternative error, let

$$\begin{aligned} \beta &= P(\text{Type II Error}) \\ &= P(\text{Jury Decides Innocent} \mid \text{Person is Guilty}) \,. \end{aligned}$$

Clearly, we would also like to keep $\beta$ small.

It is important to realize that the conditional probabilities $\alpha$ and $\beta$ depend on our **decision rule**. In other words, we have control over these probabilities.

We can make $\alpha = 0$ by not convicting anyone; however, every guilty person would then be released so that $\beta$ would then equal 1. Alternatively, we can make $\beta = 0$ by convicting everyone; however, every innocent person would then be convicted so that $\alpha$ equals 1. Although the relationship between $\alpha$ and $\beta$ is not simple, it should be clear that they move in opposite directions.

The conditional probability $\alpha$ allows us to formalize the concept of "reasonable doubt." If we set a low threshold for $\alpha$, say 0.001, then it will require more evidence to convict an innocent person. On the other hand, if we set a higher threshold for $\alpha$, say 0.1, then less evidence is required. If mathematics could be precisely applied to the court-room (which it can't), then for an acceptable level/threshold for $\alpha$, we could *attempt* to determine how convincing the evidence would have to be to achieve that threshold for $\alpha$. This is precisely what we do in hypothesis testing. The difference between court-room trials and hypothesis tests in statistics is that in the latter we could more easily **quantify** (due in large part to the central limit theorem) the relationship between our decision rule and the resulting $\alpha$.

# Testing Statistical Hypotheses . . .

In the case of the jury trial, the favored assumption is that the person is innocent. In statistical inference, one also works with a favored assumption. This favored assumption is called the **null hypothesis**, which we will denote by $H_0$. The "alternative" (or antithesis) to the null hypothesis is, naturally, called the **alternative hypothesis**, which we will denote by $H_1$ (or $H_A$). The setup of these hypotheses depends on the application context.

As an example, consider a manufacturer of computer devices. The manufacturer has a process which coats a computer part with a material that is supposed to be 100 microns thick (one micron is 1/1000 of a millimeter). If the coating is too thin, then proper insulation of the computer device will not occur and it will not function reliably. Similarly, if the coating is too thick, the device will not fit properly with other computer components.

The manufacturer has calibrated the machine that applies the coating so that it has an average coating depth of 100 microns with a standard deviation of 10 microns. When calibrated this way, the process is said to be “in control.”

Any physical process, however, will have a tendency to drift. Mechanical parts wear out, sprayers clog, etc. Accordingly, the process must be monitored to make sure that it is “in control.” How can statistics be applied to this problem?

In this manufacturing problem, the natural favored assumption, or the null hypothesis, is that the process is in control. The alternative hypothesis is therefore that the process is out of control. The analogy to the jury trial is:

Null Hypothesis: Process is In Control
Alternative Hypothesis: Process is Out of Control

| | | Process is: | |
|---|---|---|---|
| | | In Control | Out of Control |
| We Say: | In Control | No Error | Type II Error |
| | Out of Control | Type I Error | No Error |

Observe however that the above null hypothesis is not sufficiently precise: We need to define what it means to say that the process is in or out of control.

We shall say that the process is in control if the mean is 100 and out of control if the mean is not equal to 100. Thus,

$$\begin{aligned} H_0 : &\quad \mu = 100 \\ H_1 : &\quad \mu \neq 100 \end{aligned}$$

and we now have:

Null Hypothesis: $\mu = 100$

Alternative Hypothesis: $\mu \neq 100$

| | | Mean: | |
|---|---|---|---|
| | | Is 100 | Is Not 100 |
| We Say: | Mean is 100 | No Error | Type II Error |
| | Mean is not 100 | Type I Error | No Error |

The next step is to define reasonable doubt, or equivalently a threshold for $\alpha$. This is an individual choice but, like the construction of a confidence interval, the two most commonly-used values are 0.05 (or about a one in twenty chance) and 0.01 (or about a one in one hundred chance). Let us pick a threshold of 0.05. It is common practice to abbreviate this statement simply as: Let $\alpha = 0.05$.

As in a criminal trial, we must now collect evidence. In a statistical analysis, the evidence comes from data. In order to use data as evidence for or against the null hypothesis, we need to know how to appropriately summarize information contained in the data. Since our hypothesis is a statement about the population mean, it is natural to use the sample mean as our evidence.

Suppose we intend to take a sample of (say) 4 chips and compute the sample mean $\bar{X}$. How do we assess the evidence to be manifested in the value of $\bar{X}$?

From the central limit theorem, we know that *if our null hypothesis is true* (thus we are computing a conditional probability), then

$$\begin{aligned} & P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} \leq \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\ &\quad = P\left(100 - 1.96\,\frac{10}{\sqrt{4}} < \bar{X} \leq 100 + 1.96\,\frac{10}{\sqrt{4}}\right) \\ &\qquad\qquad \text{(note that } \mu = 100\text{, according to } H_0\text{)} \\ &\quad = P(90.2 < \bar{X} \leq 109.8) \\ &\quad = 0.95\,. \end{aligned}$$

We now have a procedure for determining if the observed $\bar{X}$ is beyond a reasonable doubt . . .

## Procedure

(a) Compute $\bar{X}$.

(b) If $\bar{X}$ falls between 90.2 and 109.8, then this is what would happen approximately 95% of the time if the manufacturing process was in control. We would therefore declare that it is in control.

(c) If $\bar{X}$ falls outside the range of (90.2, 109.8), then either one of two things could have happened: (I) the process is in control and it just fell outside by chance (which will happen about 5% of time); or (II) the process is out of control. While we can't be sure which is the case, but since the evidence is beyond a reasonable doubt, we would declare that the manufacturing process is out of control.

It is important to realize that under (b), just because you decide/declare that the process is in control, it does not mean that the process in fact is in control. Similarly under (c), if you decide/declare that the process is out of control, it might just be a random occurrence of an $\bar{X}$ outside the control limits; and this would happen with probability 0.05, assuming that $H_0$ is true.

The procedure described above is an example of *quality control*. To ensure that a manufacturing process is in control, one could collect data (say) daily and check the observed $\bar{X}$s against the control limits to determine if the process is performing as it should be.

## Summary

An often-heard statement is that “Statistics has proved such and such.” There is of course also the even stronger statement that “Statistics can prove anything.” In fact, now that you have examined the structure of statistical logic, it should be clear that one does **not** “prove” anything. Rather, we are only looking at consistency or inconsistency between the observed data and the proposed hypotheses.

If the observed data is “consistent” with the null hypothesis (in our example, this means that the sample mean falls between 90.2 and 109.8), then instead of “proving the hypothesis true,” the proper interpretation is that there is no reason to doubt that the null hypothesis is true.

Similarly, if the observed data is “inconsistent” with the null hypothesis (in our example, this means that the sample mean falls outside the interval (90.2, 109.8)), then either a rare event has occurred (rareness is judged by thresholds 0.05 or 0.01) and the null hypothesis is true, or in fact the null hypothesis is not true.

# One-Tail Tests . . .

In the quality-control problem, we compared $\bar{X}$ against a pair of upper and lower control limits. This is an example of a **two-tail** test. Below, we will discuss an example of a **one-tail** test.

The manager of a department store is interested in the cost effectiveness of establishing a new billing system for the store's credit customers. After a careful analysis, she determines that the new system is justified only if the mean monthly account size is *more than* \$170. The manager wishes to find out if there is sufficient statistical support for this.

The manager takes a random sample of 400 monthly accounts. The sample mean turns out to be \$178. Historical data indicate that the standard deviation of monthly accounts is about \$65.

Observe that what we are trying to find out is whether or not there is sufficient support for the hypothesis that the mean monthly accounts are “more than $170.” The standard procedure is then to let $\mu > 170$ be the *alternative* hypothesis. For this reason, the alternative hypothesis is also often referred to as the **research hypothesis**.

It follows that the null hypothesis should be defined as $\mu = 170$. Note that we do *not* use $\mu \leq 170$ as the null hypothesis; this is because the null hypothesis must be precise enough for us to determine a “unique” sampling distribution. The choice $\mu = 170$ also gives $H_0$, our favored assumption, the *least* probability of being rejected.

Thus,

$$\begin{aligned} H_0: &\quad \mu = 170 \\ H_1: &\quad \mu > 170 \end{aligned}$$

where $H_1$ is *what we want to determine* and $H_0$ specifies a *single* value for the parameter of interest.

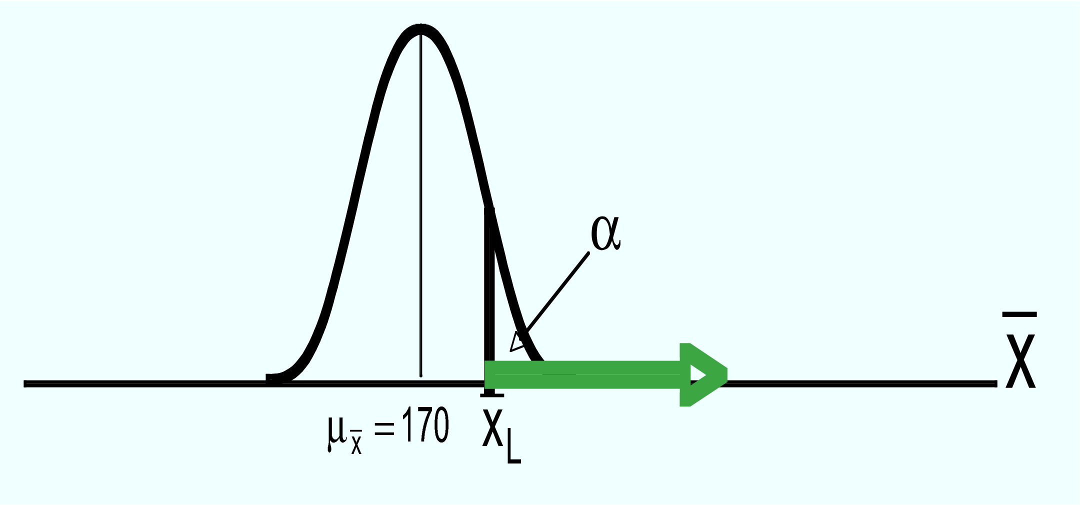
How do we test such a pair of hypotheses? There are two *equivalent* approaches ...

## Rejection-Region Approach

This approach is similar to what we did in the quality-control problem. However, we will just have one “upper” control limit, and hence the name one-tail test.

Clearly, if the sample mean is “large” relative to 170, i.e., if $\bar{X} > \bar{X}_L$ for a suitably-chosen control limit $\bar{X}_L$, then we should reject the null hypothesis in favor of the alternative.

Pictorially, this means that for a given $\alpha$, we wish to find $\bar{X}_L$ such that:



Formally, from the central limit theorem, we know that *if our null hypothesis is true*, i.e., if $\mu = 170$, then

$$P\left(\bar{X} \le 170 + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha\,. \qquad (1)$$

Hence, if we let

$$\bar{X}_L = 170 + z_\alpha \frac{\sigma}{\sqrt{n}}\,,$$

then $P(\bar{X} > \bar{X}_L) = \alpha$ and the **rejection region** is the interval $(\bar{X}_L, \infty)$.

For $\alpha = 0.05$, we have

$$\bar{X}_L = 170 + 1.645 \frac{65}{\sqrt{400}} = 175.34\,.$$

Since the observed sample mean $\bar{X} = 178$ is greater than 175.34, we reject the null hypothesis in favor of the research hypothesis (which is what we are investigating). In other words, statistical evidence suggests that the installation of the new billing system will be cost effective.

The critical-region approach can also be implemented via the standardized variable $Z$, as follows.

Observe that (1) is equivalent to:

$$P\left(Z = \frac{\bar{X} - 170}{\sigma/\sqrt{n}} \le z_\alpha\right) = 1 - \alpha\,. \tag{2}$$

Hence, $P(Z > z_\alpha) = \alpha$ and, in terms of $Z$, the rejection region is the interval $(z_\alpha,\, \infty)$.

For $\alpha = 0.05$, we have

$$Z = \frac{178 - 170}{65/\sqrt{400}} = 2.46\,.$$

Since 2.46 is greater than $z_\alpha = 1.645$, we reject $H_0$ in favor of $H_1$. This conclusion is of course the same as the previous one.

## *p*-Value Approach

The concept of *p*-values introduced earlier can also be used to conduct hypothesis tests. In this context, the *p*-value is defined as the probability of observing *any* test statistic that is at least as extreme as the one computed from a sample, *given that the null hypothesis is true.*

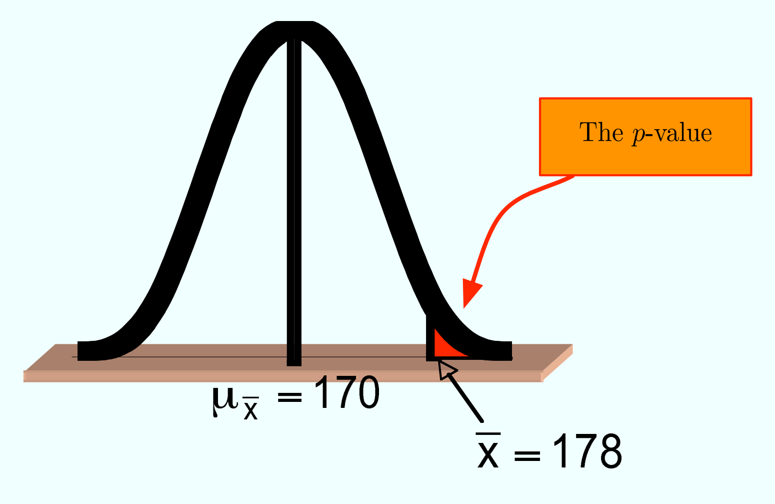
In the monthly account size example, the *p*-value associated with the given sample mean 178 is:

$$\begin{aligned} P\left(\bar{X} \geq 178 \mid \mu = 170\right) &= P\left(Z \geq \frac{178 - 170}{65/\sqrt{400}}\right) \\ &= P\left(Z \geq 2.4615\right) \\ &= 1 - P\left(Z \leq 2.4615\right) \\ &= 0.0069\,, \end{aligned}$$

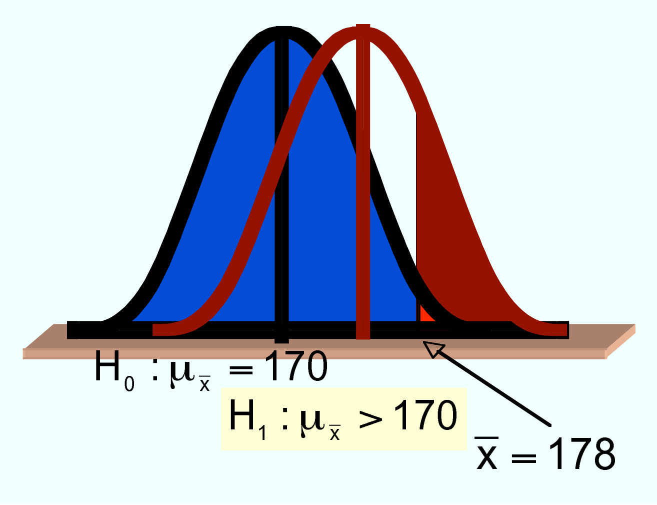
where the last equality comes from

$$\begin{aligned} P\left(Z \leq 2.4615\right) &= \text{NORMSDIST}(2.4615) \\ &= 0.9931\,. \end{aligned}$$

Pictorially, we have:



It is illuminating to observe that under $H_1$ with a $\mu > 170$ but still with the same sample mean 178 (the brown curve below), we would have a *higher* $p$-value:



The $p$-value therefore provides explicit information about the amount of statistical evidence that supports the alternative hypothesis. More explicitly, the smaller the $p$-value, the stronger the statistical evidence against the null hypothesis is.

For our problem, since the $p$-value 0.0069 is (substantially) below the specified $\alpha = 0.05$, the null hypothesis should be rejected in favor of the alternative hypothesis. This conclusion again is the same as before.

In general, we have the following guidelines:

— If the $p$-value is less than 1%, there is *overwhelming* evidence that supports the alternative hypothesis.

— If the $p$-value is between 1% and 5%, there is a *strong* evidence that supports the alternative hypothesis.

— If the $p$-value is between 5% and 10% there is a *weak* evidence that supports the alternative hypothesis.

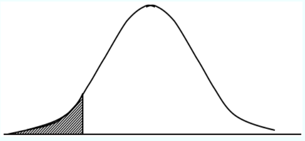
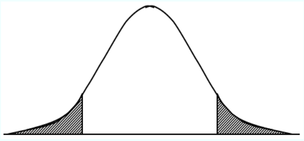
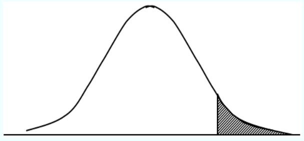
— If the $p$-value exceeds 10%, there is *no* evidence that supports the alternative hypothesis.

Our account-size problem is an example of a **right**-tail test. Depending on what we are trying to find out, we could also conduct a **left**-tail test, i.e., consider a hypotheses pair of the form:

$$\begin{aligned} H_0: &\quad \mu = \mu_0 \\ H_1: &\quad \mu < \mu_0 \end{aligned}$$

Clearly, the method is similar.

In general, we have:

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

# Calculating and Controlling $\beta$ . . .

It is important to have a good understanding of the relationship between Type I and Type II errors; that is, how the probability of a Type II error is calculated and its interpretation.

Consider again the account size problem. Recall that a Type II error occurs when a false null hypothesis is *not* rejected. In that example, we would not reject the null hypothesis if the sample mean $\bar{X}$ is less than or equal to the critical value $\bar{X}_L = 175.34$. It follows that

$$\beta = P(\bar{X} \leq 175.34 \mid H_0 \text{ is false})\,.$$

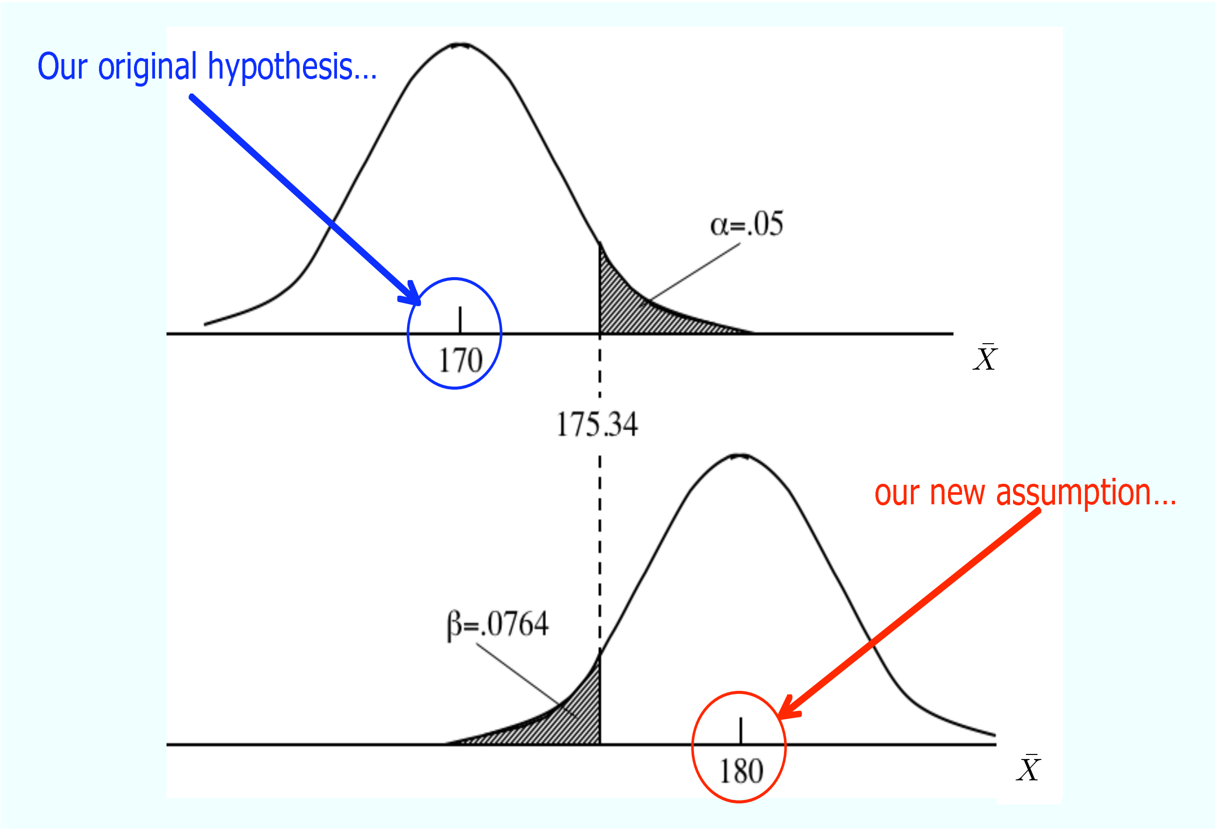
Observe that to compute $\beta$, one has to work with a specific value of $\mu$ that is *greater* than what $H_0$ states, i.e., 170. For the sake of discussion, let us pick 180. For this choice, we have

$$\beta = P(\bar{X} \leq 175.34 \mid \mu = 180)\,.$$

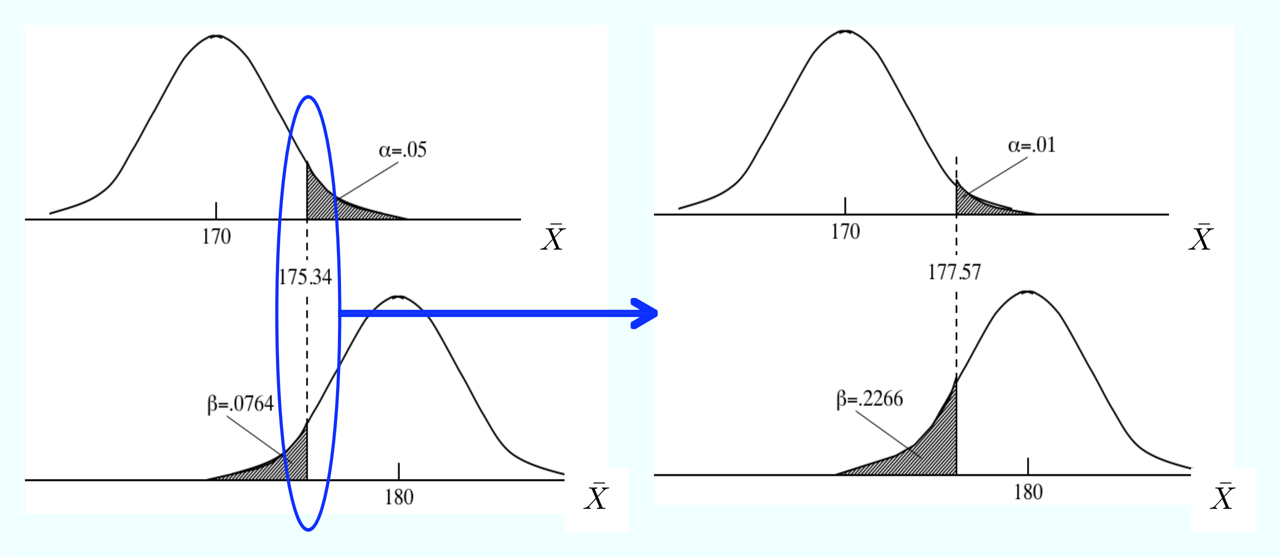
The central limit theorem then tells us that

$$\begin{aligned}\beta &= P\left(Z \leq \frac{175.34 - 180}{65/\sqrt{400}}\right) \\ &= P(Z \leq -1.43) \\ &= 0.0764\,.\end{aligned}$$

Pictorially, this means that:

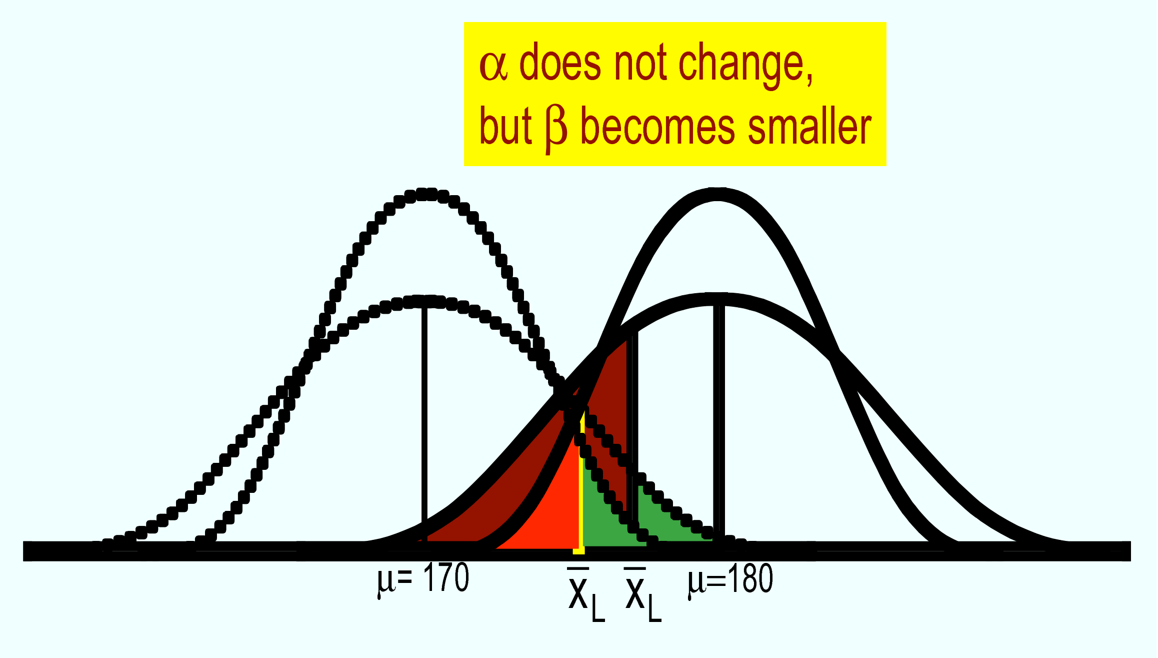


Observe that decreasing the significance level $\alpha$ will result in an increase in $\beta$, and vice versa. Alternatively, this is equivalent to saying that shifting $\bar{X}_L$ to the right (to decrease $\alpha$) will result in a larger $\beta$, and vice versa. Pictorially, we have



It is clear from the above figure that $\alpha$ and $\beta$ move in opposite directions. An important question then is: Is it possible to decrease $\beta$, for a given level of $\alpha$?

The answer is that the *shape* of the sampling distribution must change, and for that to happen, we need to *increase* the sample size $n$:



Formally, . . .

Suppose the sample size $n$ is increased to 1,000; then,

$$\begin{aligned}\bar{X}_L &= \mu + z_\alpha \frac{\sigma}{\sqrt{n}} \\ &= 170 + 1.645\,\frac{65}{\sqrt{1000}} \\ &= 173.38\end{aligned}$$

and hence, for the same alternative $\mu = 180$,

$$\begin{aligned}\beta &= P\left(Z \leq \frac{173.38 - 180}{65/\sqrt{1000}}\right) \\ &= P(Z \leq -3.22) \approx 0\,.\end{aligned}$$

That is, while maintaining the same $\alpha = 0.05$ (and assuming that $\mu = 180$), the probability of committing a Type II error has been reduced to essentially zero!

The conditional probability $1 - \beta$ is called the **power of a test**; it is the probability of taking the correct action of rejecting the null hypothesis when it is false. By increasing $n$, we can improve the power of a test. For the same $\alpha$ *and* the same $n$, the power of test is also used to choose between different tests; a "more powerful" test is one that yields the correct action with greater frequency.