Hypothesis Testing

Dr. Shafayat Abrar

Statistical Hypotheses

> **Statistical Hypothesis**
>
> A **statistical hypothesis** is a statement about the parameters of one or more populations.

As we use probability distributions to represent data or populations, a **statistical hypothesis** may be thought of as a statement about the probability distribution of a random variable. The hypothesis usually involves one or more parameters of this distribution.

Consider a *rocket*. Suppose that we are interested in the *burning rate* of the *solid propellant*. Burning rate is a *random variable* that can be described by a probability distribution. Suppose that our interest focuses on the ***mean*** burning rate (a parameter of this distribution). Specifically, we are interested in deciding whether or not the mean burning rate is 50 cent

$$H_0\colon \mu = 50 \text{ centimeters per second} \qquad H_1\colon \mu \neq 50 \text{ centimeters per second}$$

The statement cm/sec is called the **null hypothesis**. This is a claim that is initially assumed to be true. The statement cm/sec is called the **alternative hypothesis** and it is a statement that condradicts the null hypothesis.

$$H_0: \mu = 50 \text{ centimeters per second} \qquad H_1: \mu \neq 50 \text{ centimeters per second}$$

Because the alternative hypothesis specifies values of that could be either greater or less than 50 cm/sec, it is called a **two-sided alternative hypothesis.**

In some situations, we may wish to formulate a **one-sided alternative hypothesis,** as in

$$H_0: \mu = 50 \text{ centimeters per second} \qquad H_1: \mu < 50 \text{ centimeters per second}$$

or

$$H_0: \mu = 50 \text{ centimeters per second} \qquad H_1: \mu > 50 \text{ centimeters per second}$$

- We will always state the null hypothesis as an **equality** claim.
- However, when the alternative hypothesis is stated with the sign, the implicit claim in the null hypothesis can be taken as
- When the alternative hypothesis is stated with the sign, the implicit claim in the null hypothesis can be taken as .

It is important to remember that hypotheses are always statements about the population or distribution under study, not statements about the sample.

The value of the population parameter specified in the null hypothesis (50 cm/sec in the preceding example) is usually determined in one of three ways:

1. First, it may result from past experience or knowledge of the process or even from previous tests or experiments. The objective of hypothesis testing, then, is usually to determine whether the parameter value has changed.
2. Second, this value may be determined from some theory or model regarding the process under study. Here the objective of hypothesis testing is to verify the theory or model.
3. A third situation arises when the value of the population parameter results from external considerations, such as design or engineering specifications, or from contractual obligations. In this situation, the usual objective of hypothesis testing is conformance testing.

- A **procedure** leading to a **decision** about the **null hypothesis** is called a **test of a 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 null hypothesis, we will not reject it; however, if this information is inconsistent with the null hypothesis, we will conclude that the null hypothesis is false and reject it in favour of the alternative.

- We emphasize that the truth or falsity of a particular hypothesis can never be known with certainty unless we can examine the entire population. This is usually impossible in most practical situations.

- Therefore, a hypothesis-testing procedure should be developed with the probability of reaching a wrong conclusion in mind. Testing the hypothesis involves taking a random sample, computing a **test statistic** from the sample data, and then using the test statistic to make a decision about the null hypothesis.

Tests of Statistical Hypotheses

To illustrate the general concepts, consider the propellant burning rate problem introduced earlier. The null hypothesis is that the mean burning rate is 50 centimeters per second, and the alternate is that it is not equal to 50 centimeters per second. That is, we wish to test

$$H_0\colon \mu = 50 \text{ centimeters per second}$$

$$H_1\colon \mu \neq 50 \text{ centimeters per second}$$

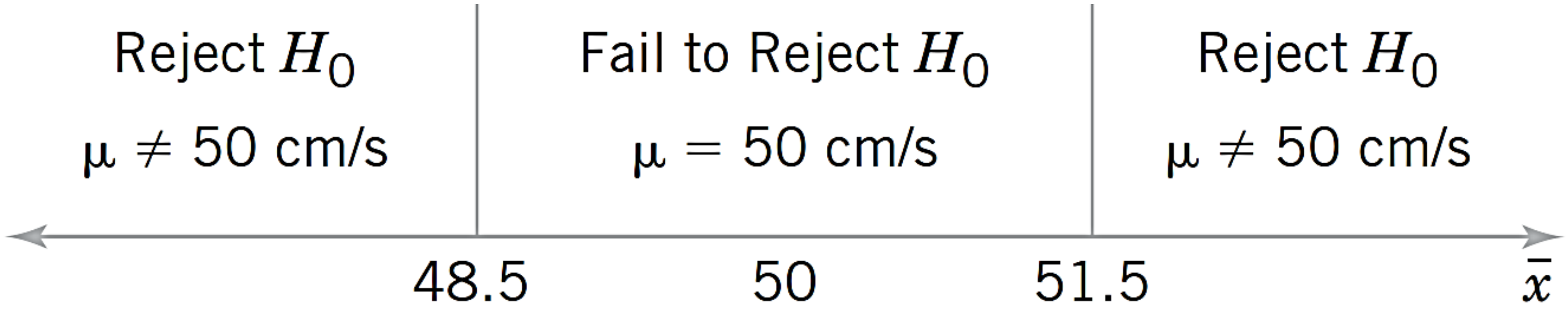
Suppose that a sample of $n = 10$ specimens is tested and that the sample mean burning rate $\bar{x}$ is observed. The sample mean is an estimate of the true population mean μ. A value of the sample mean $\bar{x}$ that falls close to the hypothesized value of $\mu = 50$ centimeters per second does not conflict with the null hypothesis that the true mean μ is really 50 centimeters per second. On the other hand, a sample mean that is considerably different from 50 centimeters per second is evidence in support of the alternative hypothesis H_1. Thus, the sample mean is the test statistic in this case.

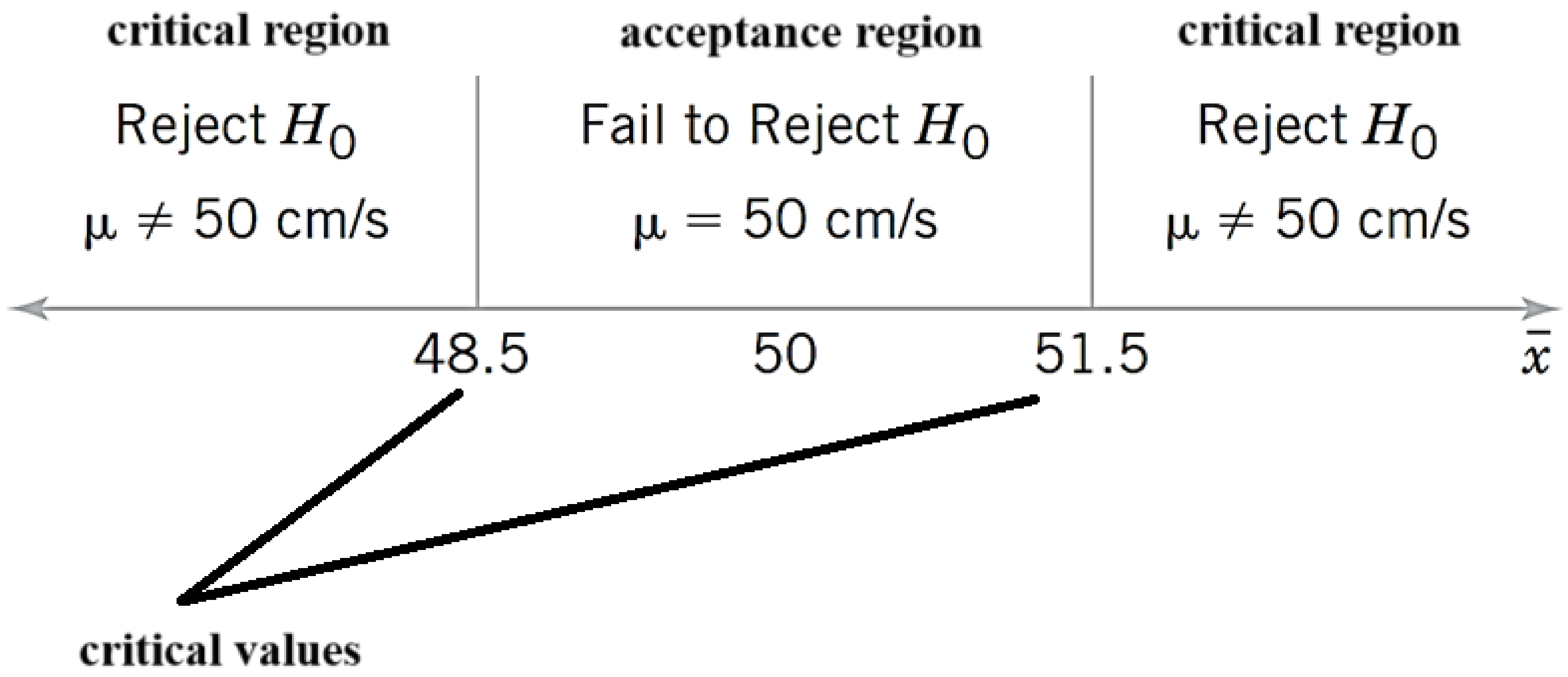
The sample mean can take on many different values. Suppose that if $48.5 \leq \overline{X} \leq 51.5$, we will not reject the null hypothesis $H_0 : \mu = 50$, and if either $\overline{X} < 48.5$ or $\overline{X} > 51.5$, we will reject the null hypothesis in favor of the alternative hypothesis $H_1 : \mu \neq 50$.

This is illustrated in Figure below. The values of $\overline{X}$ that are less than 48.5 and greater than 51.5 constitute the critical region for the test; all values that are in the interval $48.5 \leq \overline{X} \leq 51.5$ form a region for which we will fail to reject the null hypothesis.

By convention, this is usually called the acceptance region. The boundaries between the critical regions and the acceptance region are called the critical values. In our example, the critical values are 48.5 and 51.5 . It is customary to state conclusions relative to the null hypothesis H_0.

Therefore, we reject H_0 in favor of H_1 if the test statistic falls in the critical region and fails to reject H_0 otherwise.




critical region
acceptance region
critical region
Reject H_0
Fail to Reject H_0
Reject H_0
$\mu \neq 50$ cm/s
$\mu = 50$ cm/s
$\mu \neq 50$ cm/s
48.5
50
51.5
$\bar{x}$
critical values

This decision procedure can lead to either of two wrong conclusions. For example, the true mean burning rate of the propellant could be equal to 50 centimeters per second. However, for the randomly selected propellant specimens that are tested, we could observe a value of the test statistic $\overline{x}$ that falls into the critical region. We would then reject the null hypothesis H_0 in favor of the alternate H_1 when, in fact, H_0 is really true. This type of wrong conclusion is called a **type I error**.

Type I Error

Rejecting the null hypothesis H_0 when it is true is defined as a **type I error**.

Now suppose that the true mean burning rate is different from 50 centimeters per second, yet the sample mean $\overline{x}$ falls in the acceptance region. In this case, we would fail to reject H_0 when it is false, and this leads to the other type of error.

Type II Error

Failing to reject the null hypothesis when it is false is defined as a **type II error**.

TABLE **Decisions in Hypothesis Testing**

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

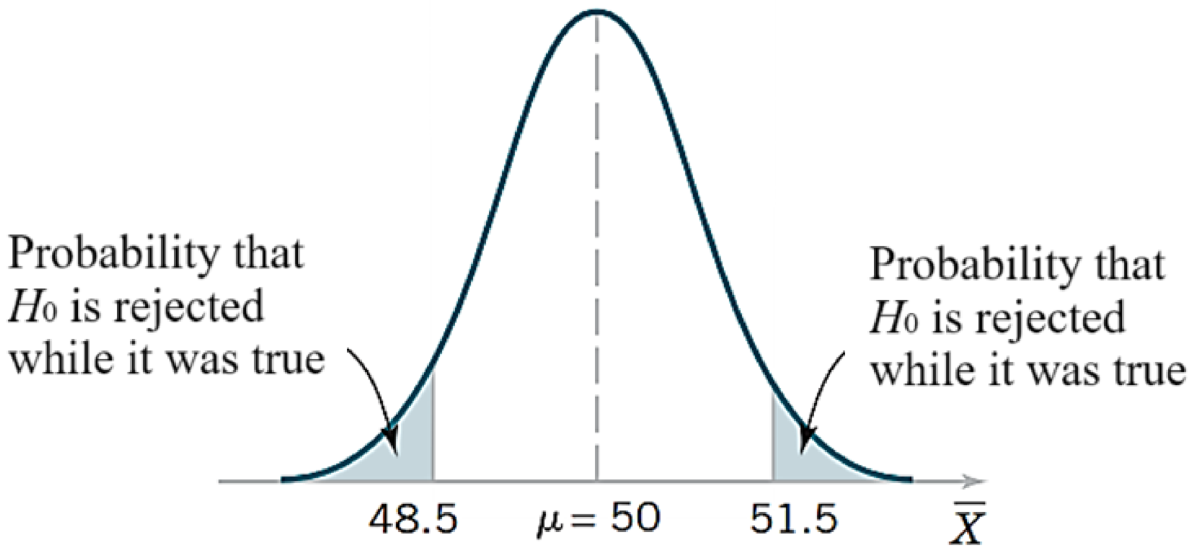
Probability of Type I Error

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

Sometimes the type I error probability is called the significance level, the α-error, or the size of the test. In the propellant burning rate example, a type I error will occur when either $\overline{X} > 51.5$ or $\overline{X} < 48.5$ when the true mean burning rate really is $\mu = 50$ cm/sec.

Suppose that the standard deviation of the burning rate is $\sigma = 2.5$ cm/sec and that the burning rate has a distribution for which the conditions of the central limit theorem apply, so the distribution of the sample mean is approximately normal with mean $\mu = 50$ and standard deviation $\sigma/\sqrt{n} = 2.5/\sqrt{10} = 0.79$.

The probability of making a type I error (or the significance level of our test) is equal to the sum of the areas that have been shaded in the tails of the normal distribution in Figure below. We may find this probability as



This implies that **5.74%** of all random samples would lead to **rejection** of the hypothesis H_0 when the true mean burning rate is **really** 50 cm/sec.

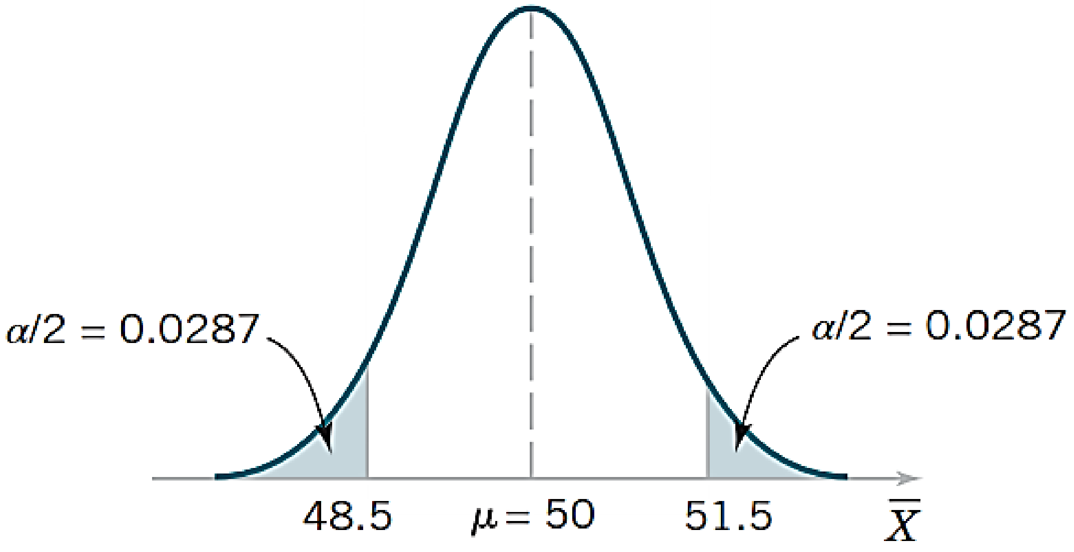
$$\alpha = P(\overline{X} < 48.5 \text{ when } \mu = 50) + P(\overline{X} > 51.5 \text{ when } \mu = 50)$$

The z-values that correspond to the critical values 48.5 and 51.5 are

$$z_1 = \frac{48.5 - 50}{0.79} = -1.90 \quad \text{and} \quad z_2 = \frac{51.5 - 50}{0.79} = 1.90$$

Therefore,

$$\alpha = P(z < -1.90) + P(z > 1.90) = 0.0287 + 0.0287 = 0.0574$$



From an inspection, notice that we can reduce α by widening the acceptance region. For example, if we make the critical values 48 and 52, the value of α is

$$\begin{aligned}\alpha &= P\left(z < -\frac{48-50}{0.79}\right) + P\left(z > \frac{52-50}{0.79}\right)\\ &= P(z < -2.53) + P(z > 2.53)\\ &= 0.0057 + 0.0057 = 0.0114\end{aligned}$$

This implies that 1.144% of all random samples would lead to rejection of the hypothesis $H_0 : \mu = 50$ cm/sec when the true mean burning rate is really 50 cm/sec.

Effect of changing critical values

We could also reduce α by increasing the sample size. If $n = 16, \sigma/\sqrt{n} = 2.5/\sqrt{16} = 0.625$ and using the original critical region from Figure 9.1, we find

$$z_1 = \frac{48.5-50}{0.625} = -2.40 \quad \text{and} \quad z_2 = \frac{51.5-50}{0.625} = 2.40$$

Therefore,

$$\alpha = P(Z < -2.40) + P(Z > 2.40) = 0.0082 + 0.0082 = 0.0164$$

This implies that 1.64% of all random samples, taken 16 at a time, would lead to rejection of the hypothesis $H_0 : \mu = 50$ cm/sec when the true mean burning rate is really 50 cm/sec.

Effect of changing sample size

Probability of Type II Error

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

To calculate β (sometimes called the β-error), we must have a specific alternative hypothesis; that is, we must have a particular value of μ under the alternative hypothesis.

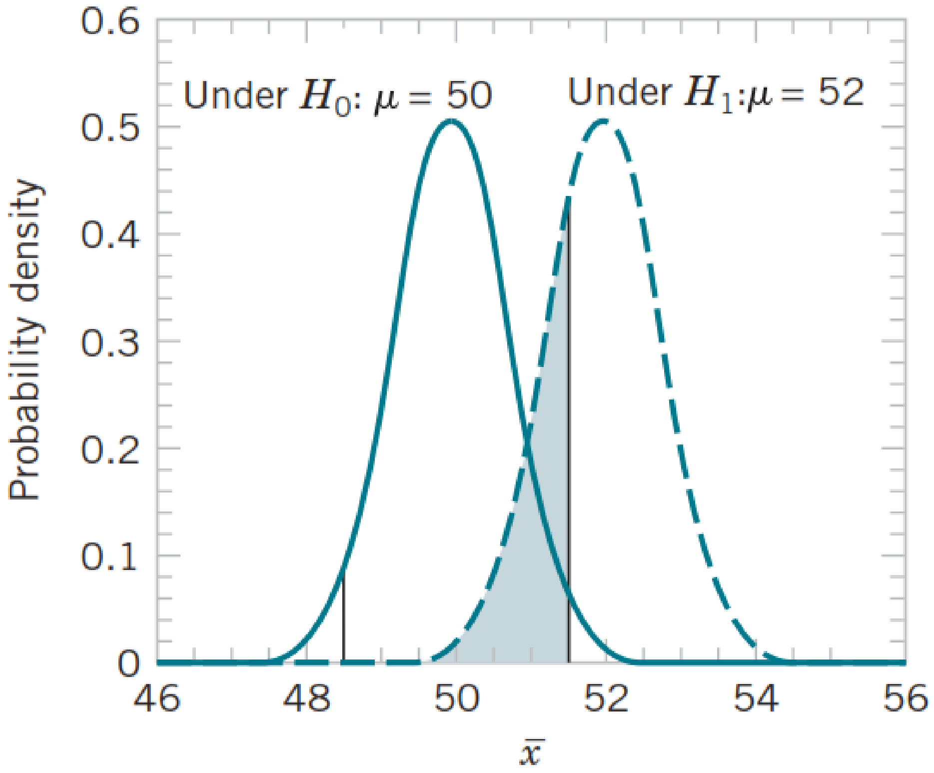
For example, suppose that it is important to reject the null hypothesis $H_0 : \mu = 50$ whenever the mean burning rate μ is greater than 52 cm/sec or less than 48 cm/sec.

We could calculate the probability of a type II error β for the values $\mu = 52$ and $\mu = 48$ and use this result to tell us something about how the test procedure would perform.

Specifically, how will the test procedure work if we wish to reject H_0, for a mean value of $\mu = 52$ or $\mu = 48$?

Because of symmetry, it is necessary to evaluate only one of the two cases - say, find the probability of failing to reject the null hypothesis $H_0 : \mu = 50$ cm/sec when the true mean is $\mu = 52$ centimeters per second.

$$\beta = P(48.5 \le \overline{X} \le 51.5 \text{ when } \mu = 52)$$



The probability of type II error when $\mu = 52$ and $n = 10$.

The z-values corresponding to 48.5 and 51.5 when $\mu = 52$ are

$$z_1 = \frac{48.5 - 52}{0.79} = -4.43 \quad \text{and} \quad z_2 = \frac{51.5 - 52}{0.79} = -0.63$$

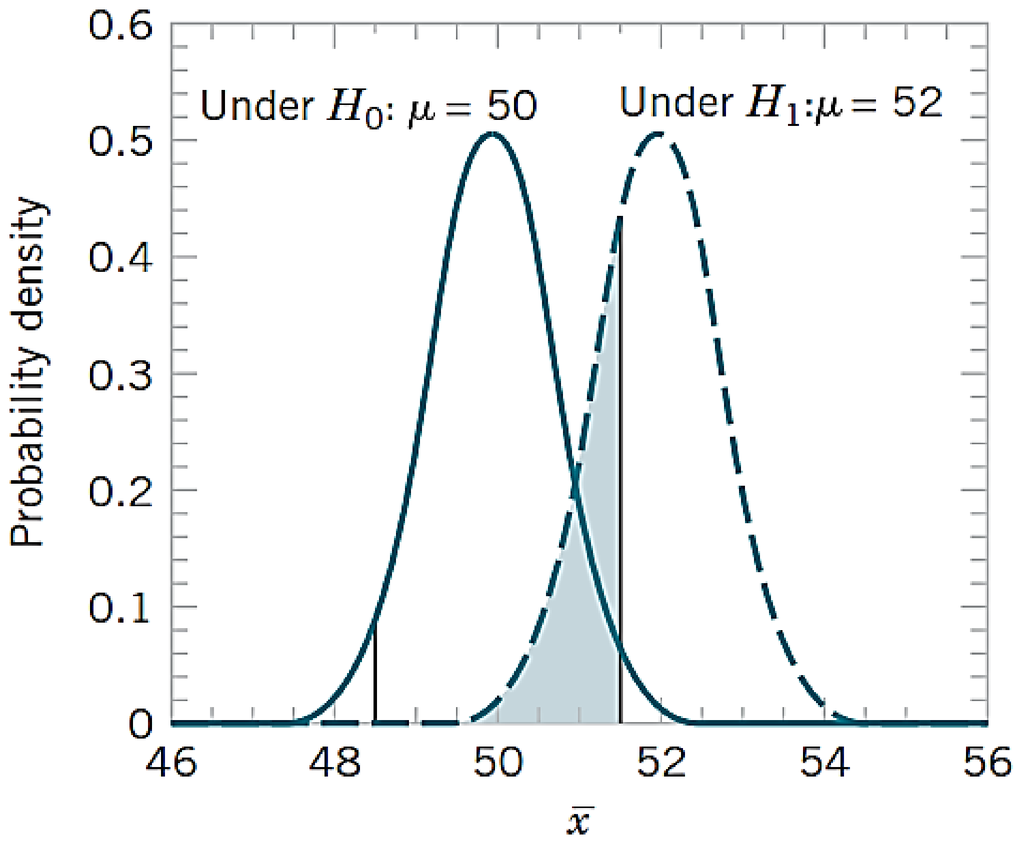
Therefore,

$$\beta = P(-4.43 \le Z \le -0.63) = P(Z \le -0.63) - P(Z \le -4.43)$$
$$= 0.2643 - 0.0000 = 0.2643$$

Thus, if we are testing $H_0 : \mu = 50$ against $H_1 : \mu \ne 50$ with $n = 10$ and the true value of the mean is $\mu = 52$, the probability that we will fail to reject the false null hypothesis is 0.2643.

By symmetry, if the true value of the mean is $\mu = 48$, the value of β will also be 0.2643.

$\beta = P(48.5 \le \overline{X} \le 51.5 \text{ when } \mu = 52)$



The probability of type II error when $\mu = 52$ and $n = 10$.

The z-values corresponding to 48.5 and 51.5 when $\mu = 50.5$ are

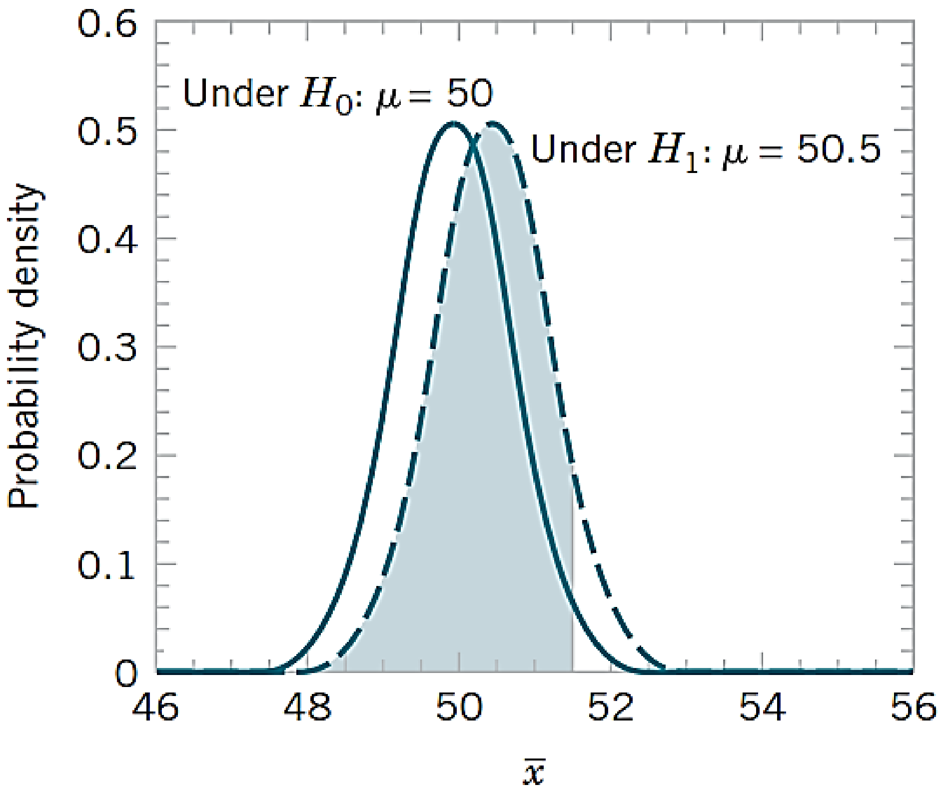
$$z_1 = \frac{48.5 - 50.5}{0.79} = -2.53 \quad \text{and} \quad z_2 = \frac{51.5 - 50.5}{0.79} = 1.27$$

Therefore,

$$\begin{aligned}\beta &= P(-2.53 \leq Z \leq 1.27) = P(Z \leq 1.27) - P(Z \leq -2.53) \\ &= 0.8980 - 0.0057 = 0.8923\end{aligned}$$

Thus, the type II error probability is much higher for the case in which the true mean is 50.5 centimeters per second than for the case in which the mean is 52 centimeters per second. Of course, in many practical situations, we would not be as concerned with making a type II error if the mean were "close" to the hypothesized value. We would be much more interested in detecting large differences between the true mean and the value specified in the null hypothesis.

$\beta = P(48.5 \leq \overline{X} \leq 51.5 \text{ when } \mu = 50.5)$



The probability of type II error when $\mu = 50.5$ and $n = 10$.

Acceptance Region	Sample Size	α	β at $\mu = 52$	β at $\mu = 50.5$
$48.5 < \bar{x} < 51.5$	10	0.0576	0.2643	0.8923
$48 < \bar{x} < 52$	10	0.0114	$\boxed{0.5000}$	$\boxed{0.9705}$
$48.81 < \bar{x} < 51.19$	16	$\boxed{0.0576}$	$\boxed{0.0966}$	$\boxed{0.8606}$
$48.42 < \bar{x} < 51.58$	16	$\boxed{0.0114}$	$\boxed{0.2515}$	$\boxed{0.9578}$

1. The size of the critical region, and consequently the probability of a type I error α, can always be reduced by appropriate selection of the critical values.
2. Type I and type II errors are related. A decrease in the probability of one type of error always results in an increase in the probability of the other provided that the sample size n does not change.
3. An increase in sample size reduces β provided that α is held constant.
4. When the null hypothesis is false, β increases as the true value of the parameter approaches the value hypothesized in the null hypothesis. The value of β decreases as the difference between the true mean and the hypothesized value increases.

Generally, the analyst controls the type I error probability α when he or she selects the critical values. Thus, it is usually easy for the analyst to set the type I error probability at (or near) any desired value. Because the analyst can directly control the probability of wrongly rejecting H_0, we always think of rejection of the null hypothesis H_0 as a *strong conclusion*.

Because we can control the probability of making a type I error (or significance level), a logical question is what value should be used. The type I error probability is a measure of risk, specifically, the risk of concluding that the null hypothesis is false when it really is not. So, the value of α should be chosen to reflect the consequences (economic, social, etc.) of incorrectly rejecting the null hypothesis. Smaller values of α would reflect more serious consequences and larger values of α would be consistent with less severe consequences. This is often hard to do, so what has evolved in much of scientific and engineering practice is to use the value $\alpha = 0.05$ in most situations unless information is available that this is an inappropriate choice. In the rocket propellant problem with $n = 10$, this would correspond to critical values of 48.45 and 51.55.

A widely used procedure in hypothesis testing is to use a type I error or significance level of $\alpha = 0.05$. This value has evolved through experience and may not be appropriate for all situations.