MAT 2377
Probability and Statistics for Engineers

Chapter 6
Hypothesis Testing

P. Boily (uOttawa)

Winter 2021

Based on course notes by Rafał Kulik

Contents

Scenario – Claims and Suspicions

Consider the following scenario: person A claims they have a fair coin, but for some reason, person B is suspicious of the claim, believing the coin to be biased in favour of tails.

Person B flips the coin 10 times, expecting a low number of heads which they intend to use as **evidence** against the claim. Let $X = \#$ of Heads.

Suppose $X = 4$. This is less than expected for a binomial random variable $X \sim \mathcal{B}(10, 0.5)$ since $\mathrm{E}[X] = 5$; the results are more in line with a coin for which $P(\text{Head}) = 0.4$.

Does this data really constitute evidence against the claim $P(\text{Head}) = 0.5$?

If the coin is fair, then $X \sim \mathcal{B}(10, 0.5)$; $X = 4$ is still close to $\mathrm{E}[X]$; in fact, $P(X = 4) = 0.205$ (as opposed to $P(X = 5) = 0.246$) so the event $X = 4$ is still quite likely. It would seem that there is no evidence against the claim that the coin is fair.
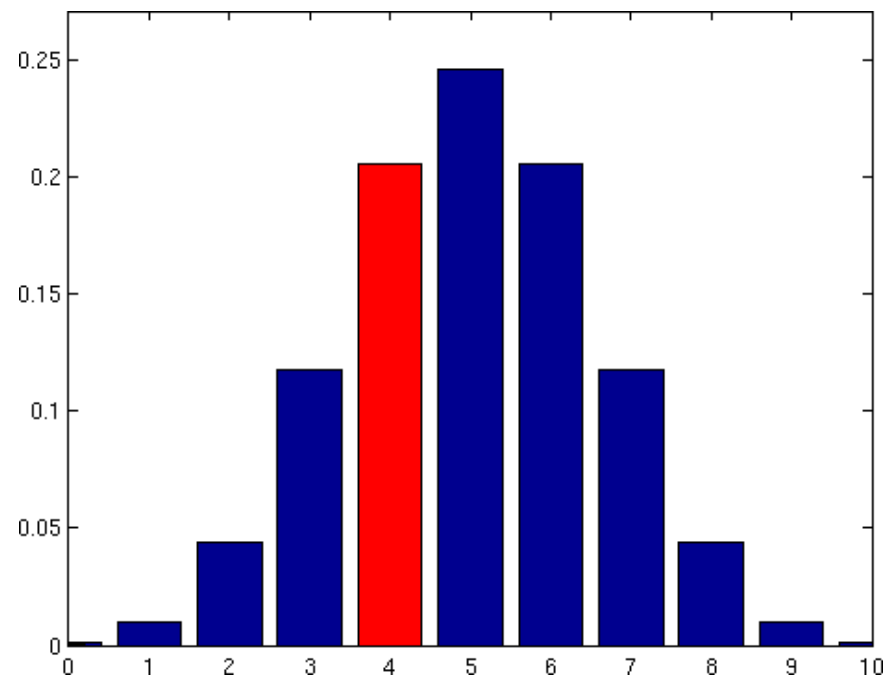
The way the sentence *"It would seem that there is no evidence against the claim that the coin is fair"* is worded is very important.

We did not reject the claim that $P(\text{Head}) = 0.5$ (i.e. that the coin is symmetric), but it **doesn't mean that in fact** $P(\textbf{Head}) = 0.5$.

Accepting, or rather, **not rejecting** a claim is a **very weak statement**.

To see why, let's consider a person C, who claims that the coin from the example above has $P(\text{Head}) = 0.3$. Under $X \sim \mathcal{B}(10, 0.3)$, the event $X = 4$ is still quite likely, with $P(X = 4) = 0.22$; we **do not have enough evidence to reject** either $P(\text{Head}) = 0.5$ or $P(\text{Head}) = 0.3$.

However, **rejecting** a claim is a **very strong statement**!

Let's say that person B convinces person A to flip the coin another 90 times. Another 36 Heads occur, giving a total of 40 out of 100.

What can we say now? Does this constitute any evidence against the claim? If so, how much?

Let $Y \sim \mathcal{B}(100, 0.5)$; $Y = 40$ is smaller than what we would expect as $\mathrm{E}[Y] = 50$ if the claim is true; $Y = 40$ is again more in agreement with $P(\mathsf{Head}) = 0.4$.

This event **does not** lie in the probability mass centre of the distribution; it falls in distribution tail (an area of lower probability). For $Y \sim \mathcal{B}(100, 0.5)$, $P(Y = 40) = 0.011$ (compare this with the previous value 0.205). Thus, if the coin is fair, the event $Y = 40$ is quite **unlikely**.

Values down in the lower tail provide **some evidence** against the claim. The question is: **how do we quantify the evidence?**
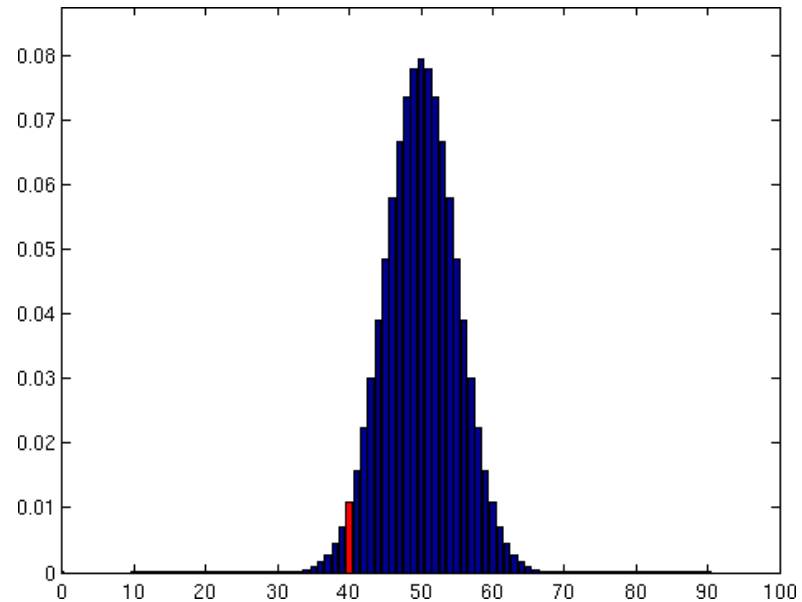
Since values which are "further down the left tail" provide evidence against the claim of a fair coin (in favour of a coin baised against Heads), we will use the actual tail area that goes with the observation: the smaller the tail area, the greater the evidence against the claim.

For 4 Heads out of 10 tosses, the evidence is the $p-$**value** $P(X \leq 4)$ if the claim is true, i.e. $P(X \leq 4)$ when $X \sim \mathcal{B}(10, 0.5)$, i.e. 0.377. Thus, if $P(\text{Head}) = 0.5$, the event $X \leq 4$ is still very likely: we would see evidence that extreme (or more) $\approx 38\%$ of the time (simply by chance).

For 40 Heads out of 100 tosses, the evidence is the $p-$**value** $P(Y \leq 40)$ if the claim is true, i.e. $P(Y \leq 40)$ when $Y \sim \mathcal{B}(100, 0.5)$, i.e. 0.028.

Thus, if $P(\text{Head}) = 0.5$, the event $Y \leq 40$ is very unlikely: we would only see evidence that extreme (or more) $\approx 3\%$ of the time.

Another way to look at the $p-$value: it's the **area of the tail** of the distribution under the assumption that the claim is true.

<div align="center">

smaller $p-$value $\Leftrightarrow$ more evidence against claim

</div>

A traditional language and notation has evolved to describe this approach to "testing hypotheses":

1. The "claim" is called the **null hypothesis** and denoted H_0.

2. The "suspicion" is called the **alternative hypothesis** and denoted H_1.

3. The (random) quantity we use to measure evidence is called a **test statistic**. We need to know its distribution when H_0 is true.

4. The $p-$**value** quantifies "the evidence against H_0".

Examples: consider the coin tossing situation described previously. The null hypothesis is
$$H_0 : P(\text{Head}) = 0.5 \, .$$
The alternative hypothesis is

$$H_1 : P(\text{Head}) < 0.5 \, .$$

The coin is tossed n times: the test statistic is $X = \#$ of Heads in n tosses.

1. If $n = 10$ and $X = 4$, the $p-$value is $P(X \leq 4) = 0.377$, assuming that $X \sim \mathcal{B}(10, 0.5)$.

2. If $n = 100$ and $X = 40$, the $p-$value is $P(X \leq 40) = 0.028$, assuming that $X \sim \mathcal{B}(100, 0.5)$.

How Small Does the $p-$Value Need to Be?

We concluded that 38% was "not that small". How small does a $p-$value need to be before we have "compelling evidence" against H_0?

There is no easy answer to this question. It depends on many factors, including what penalties we might pay for being wrong. Typically, we look at the probability of making a **type I error**, $\alpha = P(\text{reject } H_0 | H_0 \text{ is true})$:

- if $p-$value $\leq \alpha$, then we **reject** H_0 in favour of H_1;

- if $p-$value $> \alpha$, then **there is not enough evidence to reject** H_0 (**which is not the same as accepting** H_0).

By convention, we often use $\alpha = 0.01$ or $\alpha = 0.05$.

🔥**Kareem Carr**🔥
@kareem_carr

Don't know what a P-VALUE is?

Don't know why P-VALUES work?

Don't know why sometimes P-VALUES don't work?

THIS IS THE THREAD FOR YOU. 🧵

11:55 AM · Oct 4, 2020 · Twitter Web App

49 Retweets   **3** Quote Tweets   **212** Likes

🔥**Kareem Carr**🔥 @kareem_carr · 19m
Replying to @kareem_carr
DEFINITION OF A P-VALUE. Assume your theory is false. The P-VALUE is the probability of getting an outcome as extreme or even more extreme than what you got in your experiment.

💬 1   🔁 3   ♡ 13

🔥**Kareem Carr**🔥 @kareem_carr · 19m

THE LOGIC OF THE P-VALUE. Assume my theory is false. The probability of getting extreme results should be very small but I got an extreme result in my experiment. Therefore, I conclude that this is strong evidence that my theory is true. That's the logic of the p-value.

💬 2   🔁 1   ♡ 9   ⬆️

🔥**Kareem Carr**🔥 @kareem_carr · 19m

THE P-VALUE IS REASONABLE IN THEORY BUT TRICKY IN PRACTICE. In my opinion, the p-value is just a mathematical version of the way humans think. If we see something that seems unlikely given our beliefs, we often doubt those beliefs. In practice, the p-value can be tricky to use.

💬 1   🔁 2   ♡ 13   ⬆️

🔥**Kareem Carr**🔥 @kareem_carr · 19m

THE P-VALUE REQUIRES A GOOD DEFINITION OF WHEN YOUR THEORY IS FALSE. There are usually an infinite number of ways to define a world where your theory is false. P-values often fail when people use overly simplistic mathematical models of the processes that created their data.

💬 1   🔁   ♡ 12   ⬆️

🔥**Kareem Carr**🔥 @kareem_carr · 19m

If the mismatch between their mathematical models of the world and the actual world is too large then the probabilities we compute can become completely disconnected from reality.

💬 1   🔁   ♡ 4   ⬆️

🔥**Kareem Carr**🔥 @kareem_carr · 19m   ⌄

THE P-VALUE MAY REQUIRE AN ACCURATE MODEL OF YOU (THE OBSERVER). The probability of getting the result you got depends on many things. If you sometimes do things like throw out data or repeat measurements then you're part of the system.

💬 1   ⟲   ♡ 11   ⬆

🔥**Kareem Carr**🔥 @kareem_carr · 19m   ⌄

Your behavior affects the probability of getting your experimental results. Therefore, to be completely realistic, you need to have an ACCURATE model of your own behavior when you gather and analyze data. This is hard and a big part of why the p-value often fails as a tool.

💬 1   ⟲   ♡ 8   ⬆

🔥**Kareem Carr**🔥 @kareem_carr · 19m   ⌄

BY DEFINITION, P-VALUES MUST SOMETIMES BE WRONG. When using p-values, we're working off of probabilities. By logic of the p-value itself, even with perfect use, some of your decisions will be wrong. You have to embrace this if you're going to use the p-values.

💬 1   ⟲   ♡ 9   ⬆

🔥**Kareem Carr**🔥 @kareem_carr · 19m   ⌄

Badly defining what it means for your model to be false. Inaccurately modeling the chances of getting your data including your own behaviors. Not treating a p-value as a decision rule that can sometimes be wrong. These factors all contribute to misuse of the p-value in practice.

💬 1   ⟲   ♡ 5   ⬆

🔥**Kareem Carr**🔥 @kareem_carr · 19m   ⌄

Hope this cleared some things up for you. Thanks for coming to my p-value TED talk!

💬 3   ⟲   ♡ 12   ⬆

6.1 – Hypothesis Testing

A **hypothesis** is a conjecture concerning the value of a population parameter.

Hypothesis testing require two competing hypotheses:

- a **null hypothesis**, denoted by H_0;

- an **alternative hypothesis**, denoted by H_1 or H_A.

The hypothesis is tested by evaluating experimental evidence:

- we **reject** H_0 if the evidence against H_0 is **strong**;

- we **fail to reject** H_0 if the evidence against H_0 is **insufficient**.

If the evidence against H_0 is strong enough, we reject H_0 **in favour of** H_1. We say that the evidence against H_0 in favour of H_1 is **significant**. If the evidence against H_0 is not strong enough, then we fail to reject H_0. In that case, we say that the evidence against H_0 is **non-significant**.

When we fail to reject H_0, we **do NOT accept** H_0. When that happens, we simply do not have enough evidence to reject H_0.

The hypotheses should be formulated **prior to the experiment** or the study. The experiment or study is then conducted to evaluate the evidence against the null hypothesis.

In order to avoid **data snooping**, it is crucial that we do not formulate H_1 after looking at the data.

Scientific hypotheses can be often expressed in terms of whether an effect is found in the data.

In this case, we use the following null hypothesis:

$$H_0 : \quad \text{there is no effect.}$$

against the alternative hypothesis:

$$H_1 : \quad \text{there is an effect.}$$

Errors in Hypothesis Testing

Two types of errors can be committed when testing H_0 against H_1.

	Decision: reject H_0	Decision: fail to reject H_0
Reality: H_0 is True	Type I Error	No Error
Reality: H_0 is False	No Error	Type II Error

- If we reject H_0 when H_0 is true $\Rightarrow$ we have committed a **type I error**.

- If we fail to reject H_0 when H_0 is false $\Rightarrow$ **type II error**.

Examples:

1. If we conclude that a drug treatment is useful for treating a particular disease, but this is not the case in reality, then we have committed an error of type I.

2. If we cannot conclude that a drug treatment is useful for treating a particular disease, but in the reality the treatment is effective, then we have committed an error of type II.

What type of error is worst?

Probability of Committing Errors and Power

The probability of committing type I error is usually denoted by

$$\alpha = P(\text{reject } H_0 | H_0 \text{ is true}).$$

The probability of committing type II error is

$$\beta = P(\text{fail to reject } H_0 | H_0 \text{ is false}).$$

The **power of a test** is the probability of correctly rejecting H_0:

$$\text{Power} = P(\text{reject } H_0 | H_0 \text{ is false}) = 1 - \beta.$$

Conventional values of α, β, and Power are 0.05, 0.2, and 0.8, respectively.

6.2 – Types of Null and Alternative Hypotheses

Let μ be the population parameter of interest. The hypotheses are expressed in terms of the values of this parameter.

The null hypothesis is a **simple hypothesis**, that is, it is of the form:

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

where μ_0 is some candidate value ("simple" means that it is assumed to be a single value.)

The alternative hypothesis H_1 is a **composite hypothesis**, i.e. it contains more than one candidate value.

Depending on the context, hypothesis testing takes on one of the following three forms:

$$H_0 : \mu = \mu_0, \quad \text{where } \mu_0 \text{ is a number}$$

- against **a two-sided alternative**: $H_1 : \mu \neq \mu_0$;

- against **a left-sided alternative**: $H_1 : \mu < \mu_0$, or

- against **a right-sided alternative**: $H_1 : \mu > \mu_0$.

The formulation of the alternative hypothesis depends on our research hypothesis and is determined prior to experiment or study.

Examples: investigators often want to verify if new experimental conditions lead to a change in a specific population parameter.

An investigator claims that the use of a new type of soil will produce taller plants on average compared to the use of traditional soil. The mean plant height under the use of traditional soil is 20 cm.

1. Formulate the hypotheses that will be tested to verify the claim.

2. If another investigator suspects the opposite, that is, that the mean plant height when using the new soil will be smaller than the mean plant height with old soil. What hypotheses should be formulated?

3. A 3rd investigator believes that there will be an effect, but is not sure if the effect with be to produce shorter or taller plants. What hypotheses should be formulated then?

Solution: let μ represent the mean plant height with the new type of soil. In all three cases, the null hypothesis is $H_0 : \mu = 20$. The alternative hypothesis depends on the situation:

1. $H_1 : \mu > 20$.

2. $H_1 : \mu < 20$.

3. $H_1 : \mu \neq 20$.

For each H_1, the corresponding $p-$values would be computed differently when testing H_0 against H_1.

6.3 – Test Statistics and Critical Regions

To test a statistical hypothesis statistics we use a **test statistic**. A test statistic is a function of the random sample and the population parameter of interest.

We reject H_0 if the value of the test statistic is in the **critical region** or **rejection area**. The critical region is a subset of the real numbers.

The critical region is obtained using the definition of errors in hypothesis testing. We select the critical region so that

$$\alpha = P(\text{reject } H_0 | H_0 \text{ is true})$$

is equal to some pre-determined value, like 0.05 or 0.01.

Examples: a new curing process developed for a certain type of cement results in a mean compressive strength of 5000 kg/cm^2, with a standard deviation of 120 kg/cm^2. We test the hypothesis $H_0 : \mu = 5000$ against the alternative $H_1 : \mu < 5000$ with a random sample of 49 pieces of cement. Let's assume that the critical region in this specific instance is $\overline{X} < 4970$, that is, we would reject H_0 if $\overline{X} < 4970$.

1. Find the probability of committing a type I error when H_0 is true.

   Solution: by definition, we have

   $$\alpha = P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ is true}) = P(\overline{X} < 4970 | \mu = 5000).$$

   Thus, according to the CLT,

   $$\alpha \approx P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \frac{4970 - 5000}{120/7}\right) \approx P(Z < -1.75) \approx 0.0401.$$
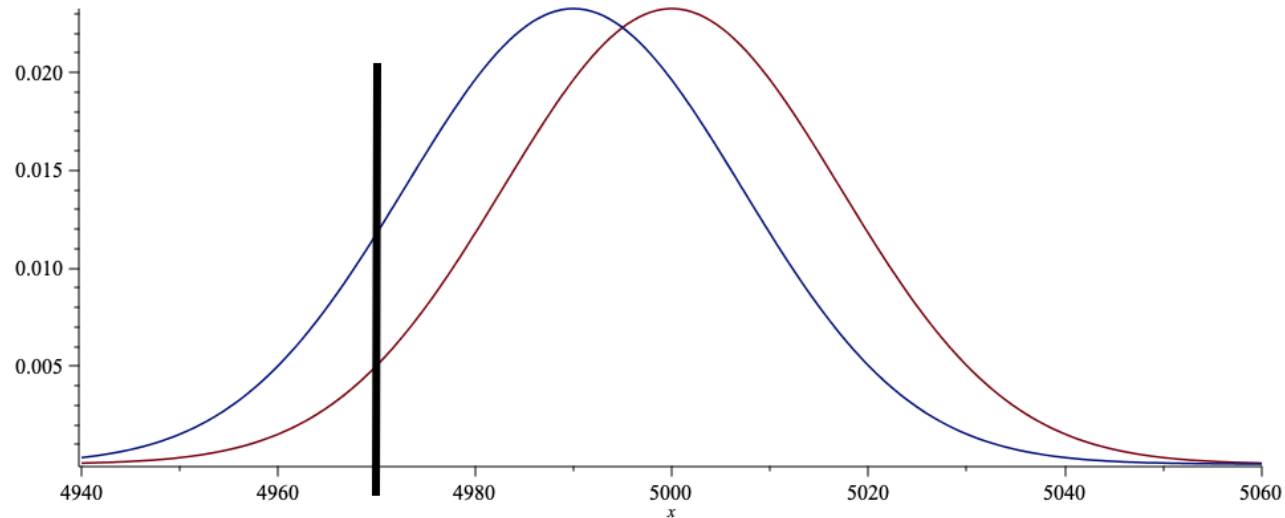
Sampling distribution of $\overline{X}$ under H_0 in **red** (mean $= 5000$, sd $= 120/7$)

Sampling distribution of $\overline{X}$ under H_1 in **blue** (mean $= 4990$, sd $= 120/7$)

Critical region $\overline{X} < 4970$ to the left of vertical **black** line
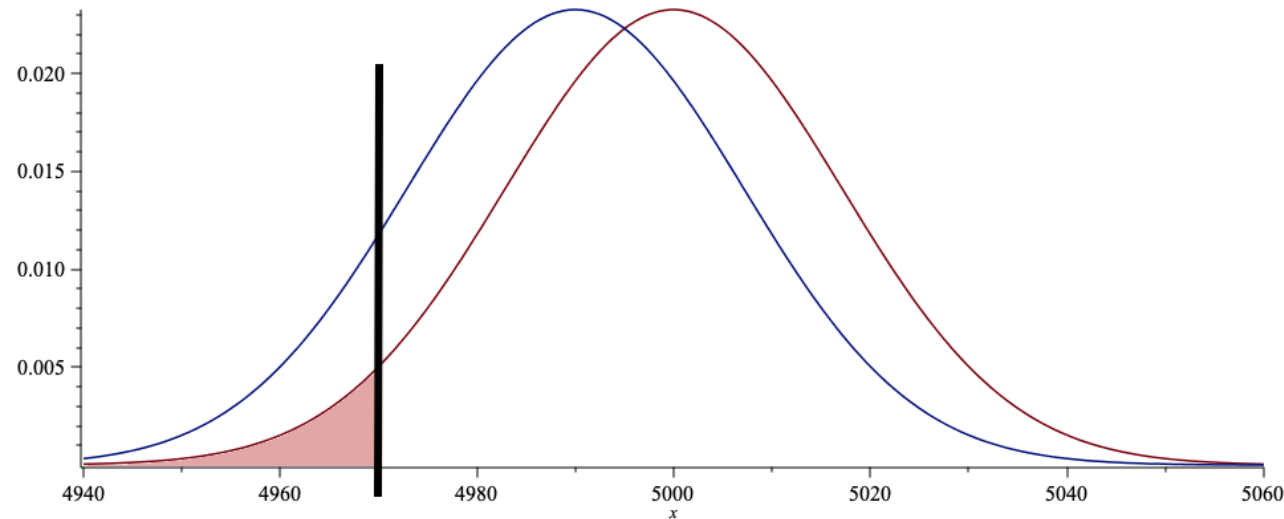
$$\alpha = P(\text{reject } H_0 | H_0 \text{ is true}) = P(\overline{X} < 4970 | \mu = 5000)$$

Reject $H_0 \Rightarrow$ to the left of $\overline{X} = 4970$ (in the critical region)

H_0 is true $\Rightarrow$ under the **red** sampling p.d.f. ($\mu = 5000$)

2. Evaluate the probability of committing a type II error if μ is actually 4990, say (and not 5000, as claimed by H_0).

   Solution: by definition, we have

   $$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 | H_0 \text{ is false})$$
   $$= P(\overline{X} > 4970 | \mu = 4990).$$

   Thus, according to the CLT, we have

   $$\beta = P(\overline{X} > 4970) = P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} > \frac{4970 - 4990}{120/7}\right) \approx P(Z > -1.17)$$
   $$= 1 - \texttt{pnorm}(-1.17, 0, 1) \approx 0.879\,.$$

$$\beta = P(\text{fail to reject } H_0 | H_0 \text{ is false}) = P(\overline{X} > 4970 | \mu = 4990)$$

Fail to reject $H_0 \Rightarrow$ to the right of $\overline{X} = 4970$ (outside the critical region)

H_0 is false $\Rightarrow$ under the **blue** sampling p.d.f. ($\mu = 4990$)
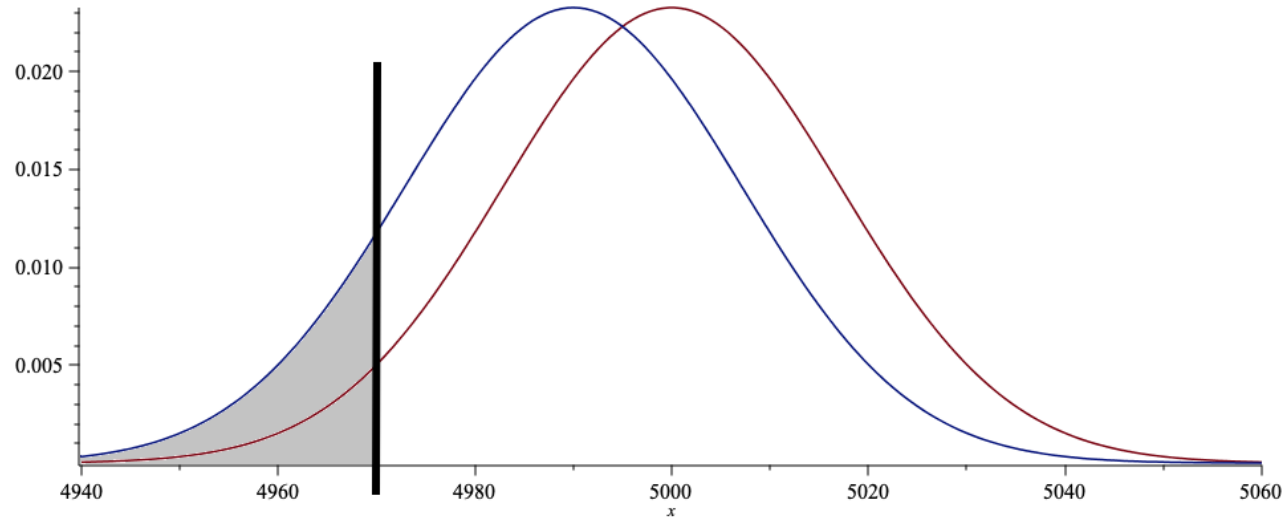
$$\text{Power} = P(\text{reject } H_0 | H_0 \text{ is false}) = P(\overline{X} < 4970) = 1 - \beta$$

Reject $H_0 \Rightarrow$ to the left of $\overline{X} = 4970$ (in the critical region)

H_0 is false $\Rightarrow$ under the **blue** sampling p.d.f. ($\mu = 4990$)
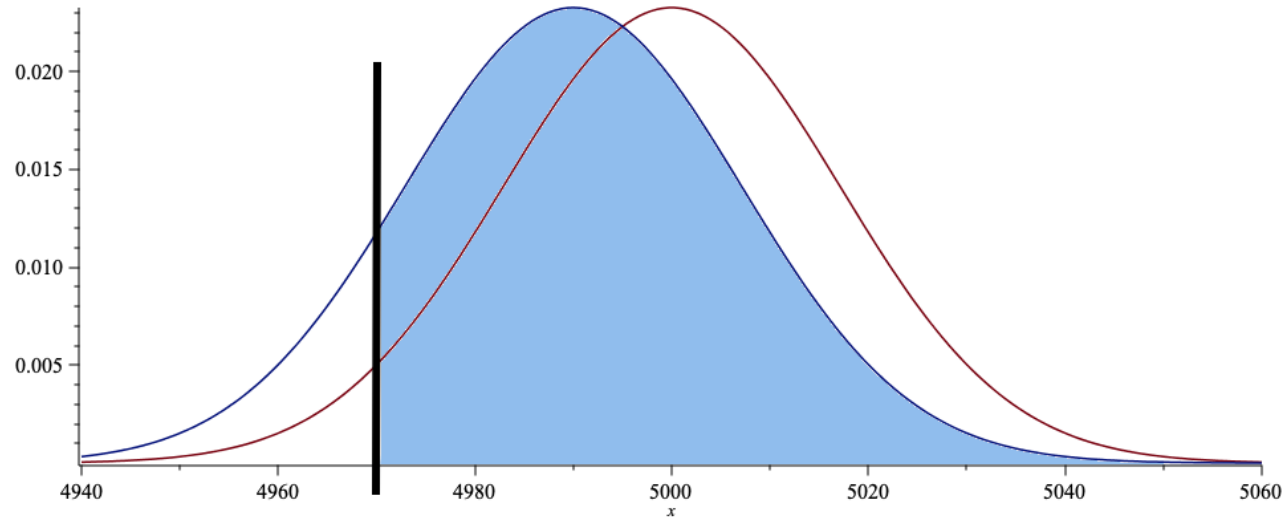
3. Evaluate the probability of committing a type II error if μ is actually 4950, say (and not 5000, as claimed by H_0).

   Solution: by definition, we have

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

   $$= P(\overline{X} > 4970 | \mu = 4950).$$

   Thus, according to the CLT, we have

   $$\beta = P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} > \frac{4970 - 4950}{120/7}\right) \approx P(Z > 1.17)$$

   $$= 1 - \texttt{pnorm}(1.17, 0, 1) \approx 0.121 \,.$$
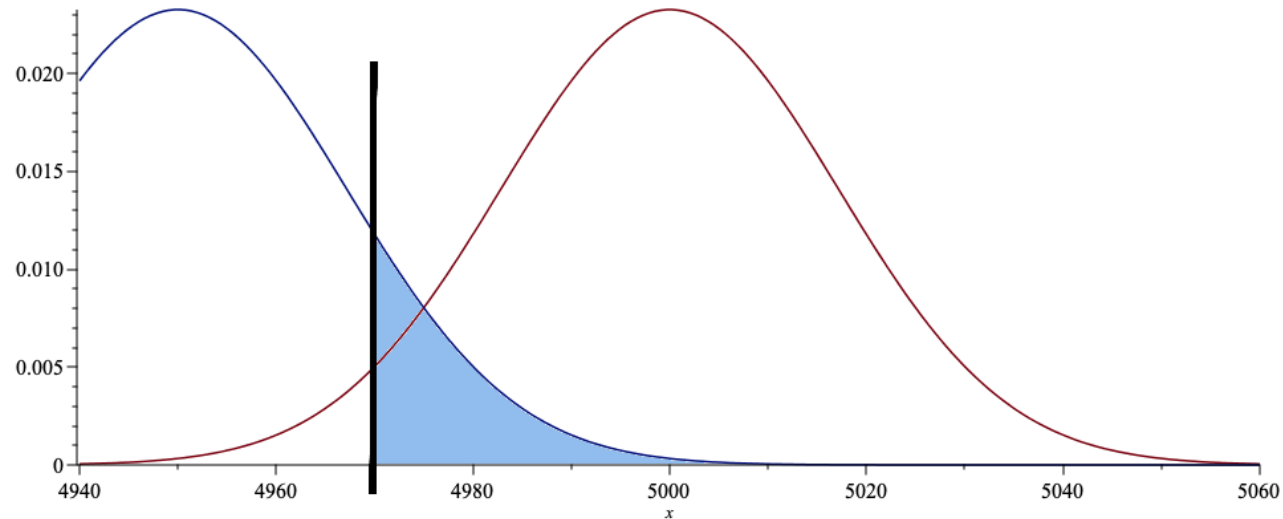
$$\beta = P(\text{fail to reject } H_0 | H_0 \text{ is false}) = P(\overline{X} > 4970)$$

Fail to reject $H_0 \Rightarrow$ to the right of $\overline{X} = 4970$ (outside the critical region)

H_0 is false $\Rightarrow$ under the **blue** sampling p.d.f. ($\mu = 4950$)

6.4 – Test for a Mean with Known Variance

Suppose $X_1, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2, and let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and denote the sample mean:

- if the population is normal, then $\overline{X} \overset{\text{exact}}{\sim} \mathcal{N}(\mu, \sigma^2/n)$;

- if the population is **not** normal, then as long as n is **large enough**, we have $\overline{X} \overset{\text{approx}}{\sim} \mathcal{N}(\mu, \sigma^2/n)$, according to the CLT.

In this section, we assume that the population variance σ^2 is known, and that the hypothesis concerns the **unknown** population mean μ.

Explanation: Left-Sided Alternative

Consider the unknown population mean μ. Suppose that we would like to test

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_1 : \mu < \mu_0,$$

where μ_0 is some candidate value for μ.

To evaluate the evidence against H_0, we compare $\overline{X}$ to μ_0: under H_0,

$$Z_0 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \overset{\text{approx}}{\sim} \mathcal{N}(0, 1).$$

We say that $z_0 = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$ is the observed value of the $Z-$**test statistic** Z_0. If $z_0 < 0$, we have evidence that $\mu < \mu_0$. However, we only reject H_0 in favour of H_1 if the evidence is **significant**.

Critical Region: Let α be the level of significance. We reject H_0 in favour of H_1 only if

$$z_0 \leq -z_\alpha .$$

The corresponding $p-$**value** for this test is the probability of observing evidence as or more extreme than our current evidence in favour of H_1, assuming that H_0 is true (that is, simply by chance). Even more extreme in this case means further to the left; so

$$p\text{-value} = P(Z \leq z_0) = \Phi(z_0),$$

where z_0 is the observed value for the Z-test statistic.

Decision Rule: if the $p-$value $\leq \alpha$, then we **reject H_0 in favour of H_1**. If the $p-$value $> \alpha$, we **fail to reject H_0.**
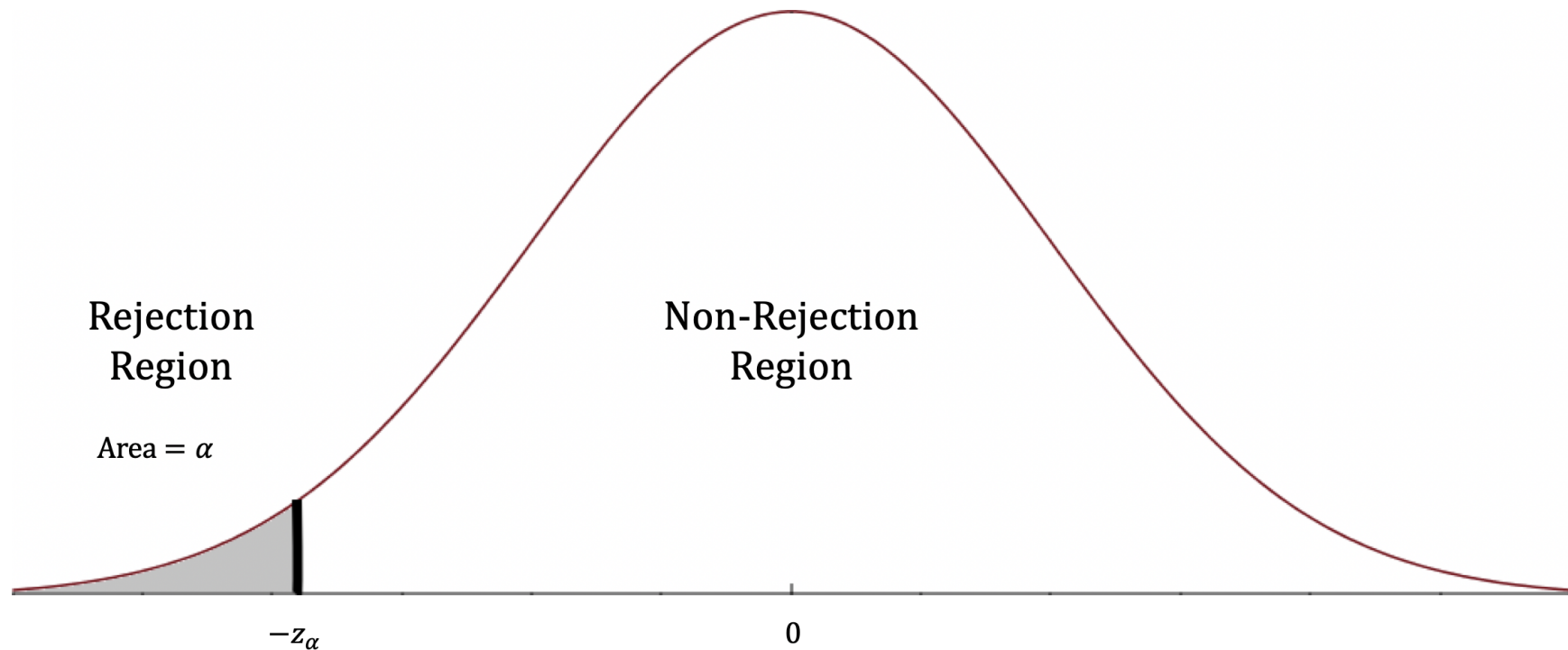
Left-Sided Test: $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$.

At significance α, if $z_0 = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha$, we reject H_0 in favour of H_1.
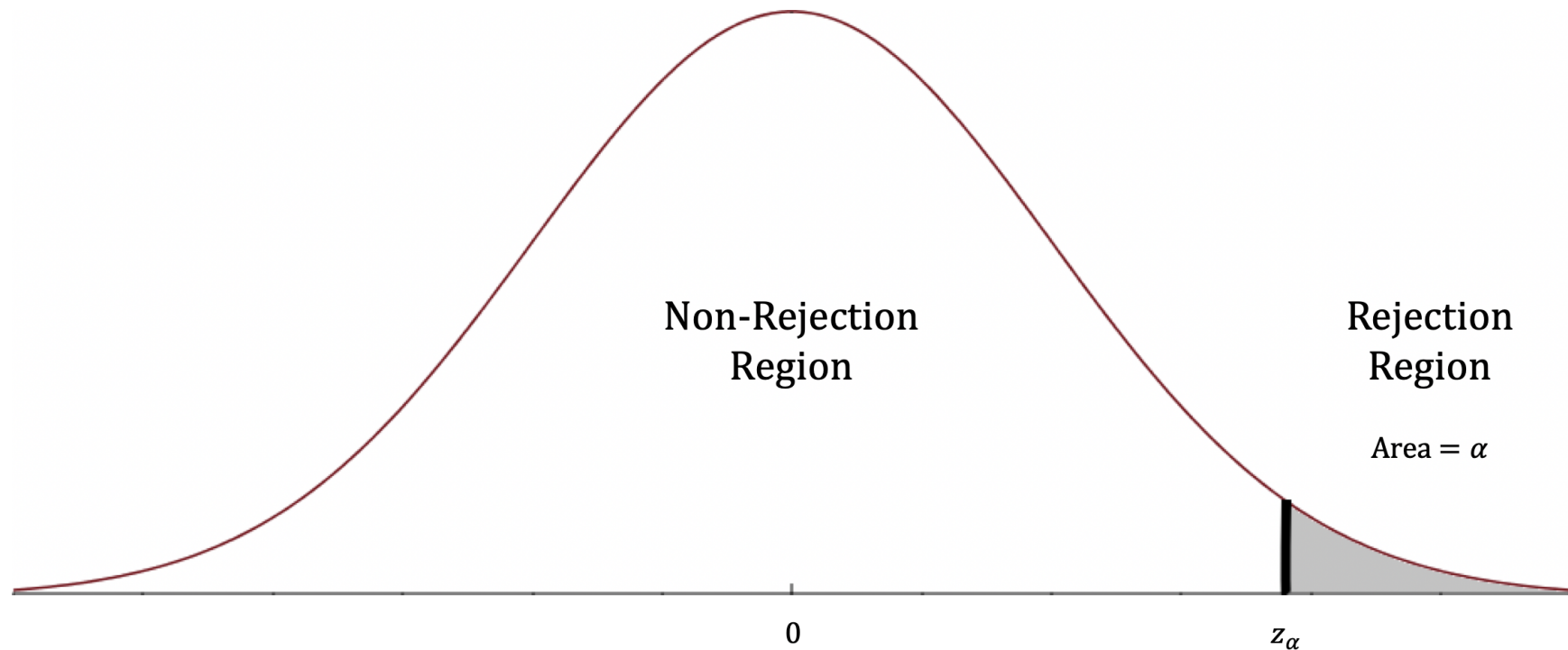
Right-Sided Test: $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$.

At significance α, if $z_0 = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha$, we reject H_0 in favour of H_1.
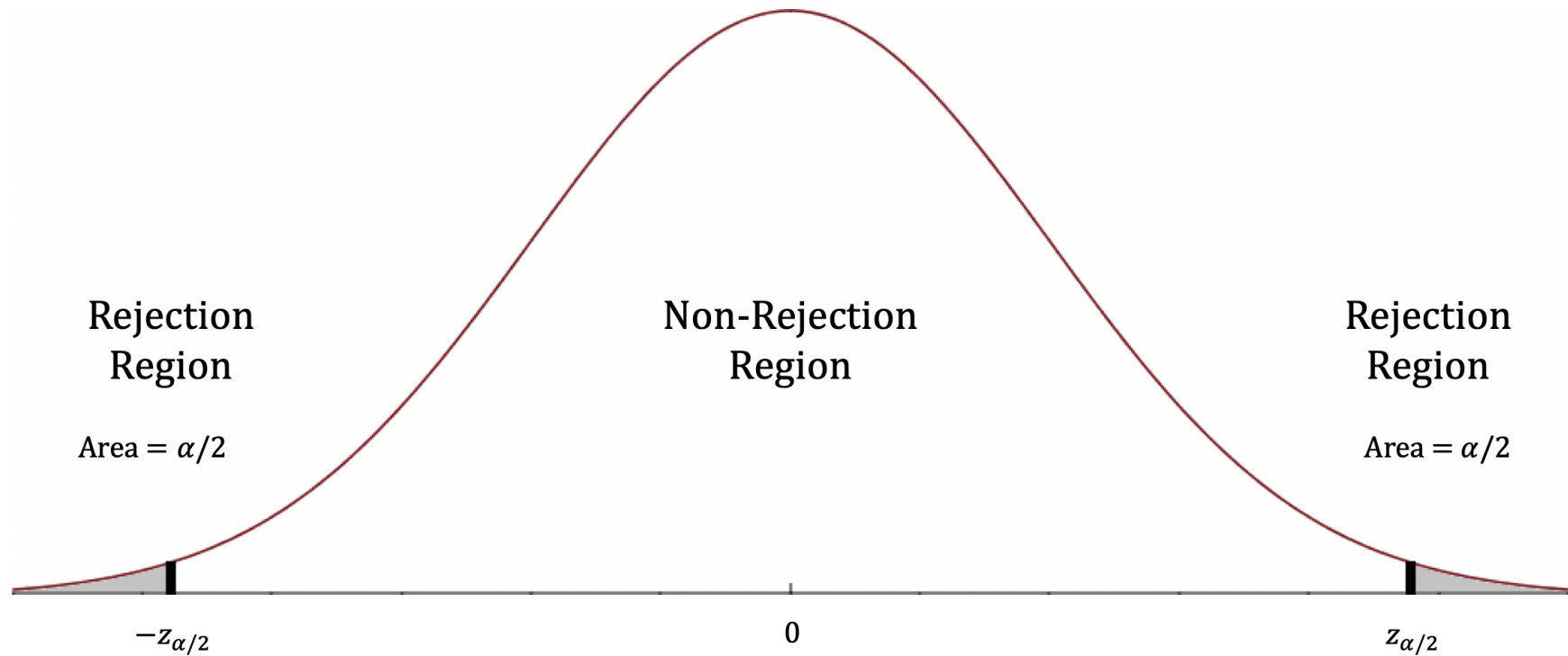
Two-Sided Test: $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.

At significance α, if $|z_0| = \left| \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \right| \geq z_{\alpha/2}$, we reject H_0 in favour of H_1.

Procedure: to test for $H_0 : \mu = \mu_0$, where μ_0 is a constant.

Step 1: set $H_0 : \mu = \mu_0$

Step 2: select an alternative hypothesis H_1 (what we are trying to show using the data). Depending on context, we choose one of these alternatives:

- $H_1 : \mu < \mu_0$ (one-sided test)

- $H_1 : \mu > \mu_0$ (one-sided test)

- $H_1 : \mu \neq \mu_0$ (two-sided test)

Step 3: choose $\alpha = P(\text{type I error})$: typically $\alpha = 0.01$ or 0.05.

Step 4: for the observed sample $\{x_1, \ldots, x_n\}$, compute the observed value of the test statistics $z_0 = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$.

Step 5: determine the critical region as follows:

Alternative Hypothesis	Critical Region
$H_1 : \mu > \mu_0$	$z_0 > z_\alpha$
$H_1 : \mu < \mu_0$	$z_0 < -z_\alpha$
$H_1 : \mu \neq \mu_0$	$\lvert z_0 \rvert > z_{\alpha/2}$

where z_α is the critical value satisfying $P(Z > z_\alpha) = \alpha$, and $Z \sim \mathcal{N}(0,1)$:

α	z_α	$z_{\alpha/2}$
0.05	1.645	1.960
0.01	2.327	2.576

Step 6: compute the associated $p-$value as follows:

Alternative Hypothesis	$p-$**Value**
$H_1 : \mu > \mu_0$	$P(Z > z_0)$
$H_1 : \mu < \mu_0$	$P(Z < z_0)$
$H_1 : \mu \neq \mu_0$	$2 \cdot \min\{P(Z > z_0), P(Z < z_0)\}$

where $Z \sim \mathcal{N}(0,1)$.

Decision Rule: if the $p-$value $\leq \alpha$, then we **reject** H_0 **in favour of** H_1. If the $p-$value $> \alpha$, we **fail to reject** H_0**.**

Let's take a look at some examples!

Examples:

1. Components are manufactured to have strength normally distributed with mean $\mu = 40$ units and standard deviation $\sigma = 1.2$ units. The manufacturing process has been modified, and an increase in mean strength is claimed (the standard deviation remains the same). A random sample of $n = 12$ components produced using the modified process had the following strengths:

   ```
   42.5, 39.8, 40.3, 43.1, 39.6, 41.0,
   39.9, 42.1, 40.7, 41.6, 42.1, 40.8
   ```

   Does the data provide strong evidence that the mean strength now exceeds 40 units? Use $\alpha = 0.05$.

Solution: we follow the previously-outlined procedure. We test for $H_0 : \mu = 40$ against $H_1 : \mu > 40$.

The observed value of the sample mean is $\overline{x} = 41.125$. Hence,

$$p-\text{value} = P(\overline{X} \geq \overline{x}) = P(\overline{X} \geq 41.125)$$

$$= P \left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{41.125 - \mu_0}{\sigma/\sqrt{n}} \right) = P(Z \geq 3.25) \approx 0.006.$$

As the $p-$value is smaller than α, we reject H_0 in favour of H_1.

Another way to see this is that if the model '$\mu = 40$' is true, then it is very unlikely that we would observe the event $\{\overline{X} \geq 41.125\}$ entirely by chance, and so the manufacturing process likely has an effect in the claimed direction.

2. A set of scales works properly if the measurements differ from the true weight by a normally distributed random error term with standard deviation $\sigma = 0.007$ grams. Investigators suspect that the scale is systematically adding to the weights. To test this hypothesis, $n = 10$ measurements are made on a 1.0g "gold-standard" weight, giving a set of measurements which average out to 1.0038g. Does this provide evidence that the scale adds to the measurement weights? Use $\alpha = 0.05$, 0.01.

Solution: let μ be the weight that the scale would record in the absence of random error terms. We test for $H_0 : \mu = 1.0$ against $H_1 : \mu > 1.0$.

The observed sample test statistic is $z_0 = \frac{1.0038 - 1.0}{0.007/\sqrt{10}} \approx 1.7167$. Since

$$z_{0.05} = 1.645 < z_0 = 1.7167 \leq z_{0.01} = 2.327,$$

we reject H_0 for $\alpha = 0.05$, but we fail to reject H_0 for $\alpha = 0.01$. Right?

3. In the previous example, let's assume that we are interested in whether the scale works properly, which means that the investigators think there might be some systematic mis-reading, but they are not sure in which direction the misreading would occur. Does the sample data provide evidence that the scale is systematically biased? Use $\alpha = 0.05$, 0.01.

Solution: let μ be as in the previous example. We test for $H_0 : \mu = 1.0$ against $H_1 : \mu \neq 1.0$.

The test statistic is still $z_0 = 1.7167$.

Since $|z_0| \leq z_{\alpha/2}$ for both $\alpha = 0.05$ and $\alpha = 0.01$, we fail to reject H_0 at either $\alpha = 0.05$ or $\alpha = 0.01$.

Our reading of the test statistic depends on what type of alternative hypothesis we have selected (and so, on the context).

4. The marks for an 'average' class are normally distributed with mean 60 and variance 100. Nine students are selected from the class; their average mark is 55. Is this subgroup 'below average'?

   Solution: let μ be the true mean of the subgroup. We are testing for $H_0 : \mu = 60$ against $H_1 : \mu < 60$. The observed sample test statistic is

   $$z_0 = \frac{55 - 60}{10/\sqrt{9}} = -1.5.$$

   The corresponding $p-$value is

   $$P(\overline{X} \le 55) = P(Z \le -1.5) = 0.07.$$

   Thus there is not enough evidence to reject the claim that the subgroup is 'average', regardless of whether we use $\alpha = 0.05$ or $\alpha = 0.01$.

5. We consider the same set-up as in the previous example, but this time the sample size is $n = 100$, not 9. Is there some evidence to suggest that this subgroup of students is 'below average'?

   Solution: let μ be as before. We are still testing for $H_0 : \mu = 60$ against $H_1 : \mu < 60$, but this time the observed sample test statistic is

   $$z_0 = \frac{55 - 60}{10/\sqrt{100}} = -5.$$

   The corresponding $p-$value is

   $$P(\overline{X} \leq 55) = P(Z \leq -5) \approx 0.00.$$

   Thus we reject the claim that the subgroup is 'average', regardless of whether we use $\alpha = 0.05$ or $\alpha = 0.01$. The sample size plays a role!

Tests and Confidence Intervals

It is becoming more and more common for analysts to bypass the computation of the $p-$value altogether, in favour of a confidence interval based approach.

If α is given, then we reject $H_0 : \mu = \mu_0$ in favour of $H_1 : \mu \neq \mu_0$ if, and only if, μ_0 is **not** in the $100(1 - \alpha)\%$ C.I. for μ.

Example: A manufacturer claims that a particular type of engine uses 20 gallons of fuel to operate for one hour. From previous studies it is known that the amount of fuel used per hour is normally distributed with variance $\sigma^2 = 25$. A sample of size $n = 9$ has been taken and the following value has been obtained: $\overline{X} = 23$. Should we accept the manufacturer's claim? Use $\alpha = 0.05$.

Solution: we test for $H_0 : \mu = 20$ against $H_1 : \mu \neq 20$ (what is μ, here?). The observed sample test statistic is

$$z_0 = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{23 - 20}{5/\sqrt{9}} = 1.8.$$

For a $2-$sided test with $\alpha = 0.05$, the critical value is $z_{0.025} = 1.96$. Since $|z_0| \leq z_{0.025}$, z_0 is not in the critical region, and we do not reject H_0.

The advantage of the **confidence interval** approach is that allows to test for various claims simultaneously. Since we know the variance of the underlying population, an approximate $100(1 - \alpha)\%$ C.I. for μ is given by

$$\overline{X} \pm z_{\alpha/2}\sigma/\sqrt{n} = 23 \pm 1.96 \cdot 5/\sqrt{9} = (19.73; 26.26).$$

We would not reject the claim that $\mu = 20$, $\mu = 19.74$, $\mu = 26.20$, etc.

6.5 – Test for a Mean with Unknown Variance

If the data is normal and σ is unknown, we can estimate it by the sample variance

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

As we have seen for confidence intervals, the test statistic

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

follows a **Student's $T-$distribution with $n-1$ degrees of freedom**.

We can follow the same steps as for the test with known variance, with the modified critical regions and $p-$values:

Alternative Hypothesis	Critical Region
$H_1 : \mu > \mu_0$	$t_0 > t_\alpha(n-1)$
$H_1 : \mu < \mu_0$	$t_0 < -t_\alpha(n-1)$
$H_1 : \mu \neq \mu_0$	$|t_0| > t_{\alpha/2}(n-1)$

where $t_0 = \frac{\overline{x}-\mu_0}{S/\sqrt{n}}$, $t_\alpha(n-1)$ is the $t-$value satisfying $P(T > t_\alpha(n-1)) = \alpha$, and T follows Student's $t-$distribution with $n-1$ degrees of freedom.

Alternative Hypothesis	$p-$Value
$H_1 : \mu > \mu_0$	$P(T > t_0)$
$H_1 : \mu < \mu_0$	$P(T < t_0)$
$H_1 : \mu \neq \mu_0$	$2 \cdot \min\{P(T > t_0), P(T < t_0)\}$

Example: consider the following observations, taken from a normal population with unknown mean μ and variance:

18.0, 17.4, 15.5, 16.8, 19.0, 17.8, 17.4, 15.8,
17.9, 16.3, 16.9, 18.6, 17.7, 16.4, 18.2, 18.7.

Test for $H_0 : \mu = 16.6$ against $H_1 : \mu > 16.6$, using $\alpha = 0.05$.

Solution: we test for $H_0 : \mu = 16.6$ against $H_1 : \mu > 16.6$. The sample size is $n = 16$, and the sample mean and sample variance are $\overline{X} = 17.4$ and $S = 1.078$, respectively. The observed sample test statistic

$$t_0 = \frac{\overline{x} - \mu_0}{S/\sqrt{n}} = \frac{17.4 - 16.6}{1.078/4} \approx 2.968,$$

so that the $p-$value is

$$p-\text{value } = P(\overline{X} \geq 17.4) = P(T > 2.968).$$

where T follows Student's $t-$distribution with $\nu = n - 1 = 15$ degrees of freedom. From the $t-$tables we see that, for $\nu = 15$,

$$P\left(T(15) \geq 2.947\right) \approx 0.005 \quad \text{and} \quad P\left(T(15) \geq 3.286\right) \approx 0.0025\,.$$

Hence the corresponding $p-$value is somewhere between these endpoints, i.e. in the interval $(0.0025, 0.005)$, in particular, $p-\text{value} \leq 0.05$, which is strong evidence against $H_0 \colon \mu = 16.6$.

6.6 – Test for a Proportion

The principle is pretty much the same; as we can see in the next example.

Example: a group of 100 adult American Catholics have been asked: "Do you favour allowing women to be priests?" 60 of them answered 'Yes'; is the evidence strong enough to conclude that more than half of American Catholics favour allowing women to be priests?

Solution: let X be the number of people who answered 'Yes'. We assume that $X \sim \mathcal{B}(100, p)$, where p is the true proportion of American Catholics who favour allowing women to be priests.

We test for $H_0 : p = 0.5$ against $H_1 : p > 0.5$. Under H_0, $X \sim \mathcal{B}(100, 0.5)$.

The $p-$value that corresponds to the observed sample is

$$P(X \geq 60) = 1 - P(X < 60) = 1 - P(X \leq 59)$$

$$\approx 1 - P\left(\frac{X+0.5-np}{\sqrt{np(1-p)}} \leq \frac{59+0.5-50}{\sqrt{25}}\right)$$

$$\approx 1 - P(Z \leq 1.9) = 0.0287,$$

where the $+0.5$ comes from the correction to the normal approximation of the binomial distribution.

Thus, we reject H_0 for $\alpha = 0.05$, but we do not reject H_0 for $\alpha = 0.01$.

6.7 – Paired Two-Sample Test

Let $X_{1,1}, \ldots, X_{1,n}$ be a random sample from a normal population with unknown mean μ_1 and unknown variance σ^2; let $X_{2,1}, \ldots, X_{2,n}$ be a random sample from a normal population with unknown mean μ_2 and unknown variance σ^2, with both populations **not independent** of one another (i.e., it's possible that the 2 samples come from the same population, or are measurements on the same units).

We would like to test for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$. In order to do so, we compute the differences $D_i = X_{1,i} - X_{2,i}$ and consider the $t-$test (as we do not know the variance). The test statistic is

$$T_0 = \frac{\overline{D}}{S_D/\sqrt{n}} \sim t(n-1), \text{ where } \overline{D} = \frac{1}{n}\sum_{i=1}^{n}D_i \text{ and } S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2.$$

Example: $n = 10$ engineers' knowledge of basic statistical concepts was measured on a scale from $0-100$ before and after a short course in statistical quality control. The result are as follows:

Engineer	1	2	3	4	5	6	7	8	9	10
Before $X_{1,i}$	43	82	77	39	51	66	55	61	79	43
After $X_{2,i}$	51	84	74	48	53	61	59	75	82	48

Let μ_1 and μ_2 be the mean score before and after the course, respectively, with normally distributed scores. Test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$.

Solution: The differences $D_i = X_{1,i} - X_{2,i}$ are:

Engineer	1	2	3	4	5	6	7	8	9	10
Before X_{1i}	43	82	77	39	51	66	55	61	79	43
After X_{2i}	51	84	74	48	53	61	59	75	82	48
Difference D_i	-8	-2	3	-9	-2	5	-4	-14	-3	-5

The observed sample mean is $\overline{d} = -3.9$, and the observed sample variance is $s_D^2 = 31.21$. The test statistic is

$$T_0 = \frac{\overline{D} - 0}{S_D/\sqrt{n}} \sim t(n-1), \text{ with observed value } t_0 = \frac{-3.9}{\sqrt{31.21/10}} \approx -2.21.$$

We compute

$$P(\overline{D} \leq -3.9) = P(T(9) \leq -2.21) = P(T(9) > 2.21).$$

But $t_{0.05}(9) = 1.833 < t_0 = 2.21 < t_{0.01}(9) = 2.821$, so we reject H_0 when $\alpha = 0.05$, but we do not reject H_0 when $\alpha = 0.01$.

r	$t_{0.40}(r)$	$t_{0.25}(r)$	$t_{0.10}(r)$	$t_{0.05}(r)$	$t_{0.025}(r)$	$t_{0.01}(r)$	$t_{0.005}(r)$
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169

6.8 – Unpaired Two-Sample Test

Let $X_{1,1}, \ldots, X_{1,n}$ be a random sample from a normal population with unknown mean μ_1 and variance σ_1^2; let $Y_{2,1}, \ldots, Y_{2,m}$ be a random sample from a normal population with unknown mean μ_2 and variance σ_2^2, with both populations **independent** of one another.

We want to test

$$H_0 : \mu_1 = \mu_2 \ \text{ against } \ H_1 : \mu_1 \neq \mu_2.$$

Let $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \ , \quad \overline{Y} = \frac{1}{m} \sum_{i=1}^{m} Y_i \ .$ The observed values are again denoted by lower case letters: $\overline{x}, \overline{y}$.

Case 1: σ_1^2 and σ_2^2 are Known

Alternative Hypothesis	Critical Region
$H_1 : \mu_1 > \mu_2$	$z_0 > z_\alpha$
$H_1 : \mu_1 < \mu_2$	$z_0 < -z_\alpha$
$H_1 : \mu_1 \neq \mu_2$	$|z_0| > z_{\alpha/2}$

where $z_0 = \dfrac{\overline{x} - \overline{y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}$, z_α satisfies $P(Z > z_\alpha) = \alpha$, and $Z \sim \mathcal{N}(0, 1)$.

Alternative Hypothesis	$p-$Value
$H_1 : \mu_1 > \mu_2$	$P(Z > z_0)$
$H_1 : \mu_1 < \mu_2$	$P(Z < z_0)$
$H_1 : \mu_1 \neq \mu_2$	$2 \cdot \min\{P(Z > z_0), P(Z < z_0)\}$

Example: a sample of $n = 100$ Albertans yields a sample mean income of $\overline{X} = 33,000\$$. A sample of $m = 80$ Ontarians yields a sample mean income of $\overline{Y} = 32,000\$$. From previous studies, it is known that the population income standard deviations are, respectively, $\sigma_1 = 5000\$$ in Alberta and $\sigma_2 = 2000\$$ in Ontario. Do Albertans earn more than Ontarians, on average?

Solution: we test for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$. The observed difference is $\overline{X} - \overline{Y} = 1000$; the observed test statistic is

$$z_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} = \frac{1000}{\sqrt{5000^2/100 + 2000^2/80}} = 1.82;$$

the corresponding $p-$value is $P\left(\overline{X} - \overline{Y} > 1000\right) = P(Z > 1.82) = 0.035$, and we reject H_0 when $\alpha = 0.05$, but not when $\alpha = 0.01$.

Case 2: σ_1^2 and σ_2^2 are Unknown (Small Samples)

Alternative Hypothesis	Critical Region
$H_1 : \mu_1 > \mu_2$	$t_0 > t_\alpha(n + m - 2)$
$H_1 : \mu_1 < \mu_2$	$t_0 < -t_\alpha(n + m - 2)$
$H_1 : \mu_1 \neq \mu_2$	$|t_0| > t_{\alpha/2}(n + m - 2)$

where $t_0 = \dfrac{\overline{X} - \overline{Y}}{\sqrt{S_p^2/n + S_p^2/m}}$, $S_p^2 = \dfrac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$, $t_\alpha(n + m - 2)$ satisfies $P(T > t_\alpha(n + m - 2)) = \alpha$, and $T \sim t(n + m - 2)$.

Alternative Hypothesis	$p-$Value
$H_1 : \mu_1 > \mu_2$	$P(T > t_0)$
$H_1 : \mu_1 < \mu_2$	$P(T < t_0)$
$H_1 : \mu_1 \neq \mu_2$	$2 \cdot \min\{P(T > t_0), P(T < t_0)\}$

Example: a researcher wants to test whether, on average, a new fertilizer yields taller plants. Plants were divided into two groups: a control group treated with an old fertilizer and a study group treated with the new fertilizer. The following data are obtained:

Sample Size	Sample Mean	Sample Variance
$n = 8$	$\overline{X} = 43.14$	$S_1^2 = 71.65$
$m = 8$	$\overline{Y} = 47.79$	$S_2^2 = 52.66$

Test $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 < \mu_2$.

Solution: We test for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$. The observed difference is -4.65; and the pooled sampled variance is

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2} = \frac{7(71.65) + 7(52.66)}{8+8-2} = 62.155 = 7.88^2.$$

The observed test statistic is

$$t_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2/n + S_p^2/m}} = \frac{-4.65}{7.88\sqrt{1/8 + 1/8}} = -1.18;$$

the corresponding $p-$value is

$$P\left(\overline{X} - \overline{Y} < -4.65\right) = P(T(14) < -1.18) = P(T(14) > 1.18) \in (0.1, 0.25)$$

(from the table), and we do not reject H_0 when $\alpha = 0.05$, or when $\alpha = 0.01$.

r	$t_{0.40}(r)$	$t_{0.25}(r)$	$t_{0.10}(r)$	$t_{0.05}(r)$	$t_{0.025}(r)$	$t_{0.01}(r)$	$t_{0.005}(r)$
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012
14	0.258	0.692	1.345	1.761	2.145	2.624	2.997
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845

Case 3: σ_1^2 and σ_2^2 are Unknown (Large Samples)

Alternative Hypothesis	Critical Region
$H_1 : \mu_1 > \mu_2$	$z_0 > z_\alpha$
$H_1 : \mu_1 < \mu_2$	$z_0 < -z_\alpha$
$H_1 : \mu_1 \neq \mu_2$	$|z_0| > z_{\alpha/2}$

where $z_0 = \dfrac{\overline{X}-\overline{Y}}{\sqrt{S_1^2/n+S_2^2/m}}$, z_α satisfies $P(Z > z_\alpha) = \alpha$, and $Z \sim \mathcal{N}(0,1)$.

Alternative Hypothesis	$p-$Value
$H_1 : \mu_1 > \mu_2$	$P(Z > z_0)$
$H_1 : \mu_1 < \mu_2$	$P(Z < z_0)$
$H_1 : \mu_1 \neq \mu_2$	$2 \cdot \min\{P(Z > z_0), P(Z < z_0)\}$

Example: same as the previous example, but with larger sample sizes: $n = m = 100$. Test $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 < \mu_2$.

Solution: We test for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$. The observed difference is (still) -4.65. The observed test statistic is

$$z_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{S_1^2/n + S_2^2/m}} = \frac{-4.65}{\sqrt{71.65^2/100 + 52.66^2/100}} = -4.17;$$

the corresponding $p-$value is

$$P\left(\overline{X} - \overline{Y} < -4.65\right) = P(Z < -4.17) \approx 0.0000;$$

we reject H_0 when either $\alpha = 0.05$ or $\alpha = 0.01$.

6.9 – Difference of Two Proportions

As always, we can transfer these tests to proportions, using the normal approximation to the binomial distribution.

Example: consider a proportion of recaptured moths in the light-coloured (p_1) and the dark-coloured (p_2) populations. Among the $n_1 = 137$ light-coloured moths, $y_1 = 18$ were recaptured; among the $n_2 = 493$ dark-coloured moths, $y_2 = 131$ were recaptured. Is there a significant difference between the proportion of recaptured moths in both populations?

Solution: We test for $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$. The observed proportions are

$$\hat{p}_1 = \frac{y_1}{n_1} = 0.131; \ \hat{p}_2 = \frac{y_2}{n_2} = 0.266; \ \hat{p}_1 - \hat{p}_2 = -0.135.$$

The corresponding $p-$value is given by $2 \cdot P(\hat{p}_1 - \hat{p}_2 \leq -0.135)$:

$$2 \cdot P\left(\frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{1/n_1 + 1/n_2}} \leq \frac{-0.135 - 0}{\sqrt{0.2365(1 - 0.2365)}\sqrt{1/137 + 1/493}} \right) \cdot$$

We get

$$2 \cdot P(\hat{p}_1 - \hat{p}_2 \leq -0.135) \approx 2P(Z < -3.29) \approx 0.0000,$$

and we reject H_0 when either $\alpha = 0.05$ or $\alpha = 0.01$.

Note: $\hat{p}$ is the **pooled proportion**:

$$\hat{p} = \frac{n_1}{n_1 + n_2}\hat{p}_1 + \frac{n_2}{n_1 + n_2}\hat{p}_2.$$

6.10 – Hypothesis Testing with R

- `t.test(x,mu=5)` tests for $H_0 : \mu = 5$ against $H_1 : \mu \neq 5$ when σ is unknown ($t-$test)

- `t.test(x,mu=5,alternative="greater")` tests for $H_0 : \mu = 5$ against $H_1 : \mu > 5$ when σ is unknown ($t-$test)

- `t.test(x,mu=5,alternative="less")` tests for $H_0 : \mu = 5$ against $H_1 : \mu < 5$ when σ is unknown ($t-$test)

- `t.test(x,y,var.equal=TRUE)` tests for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ in case of two independent samples, when variances are unknown but equal.

- `t.test(x,y,var.equal=TRUE,alternative="greater")` tests for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ in case of two independent samples, when variances are unknown but equal.

- `t.test(x,y,var.equal=TRUE,alternative="less")` tests for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$ in case of two independent samples, when variances are unknown but equal.

Example:

```
> x=c(4,5,4,6,4,4,5)
> t.test(x,mu=5)

One Sample t-test
data:  x
t = -1.4412, df = 6, p-value = 0.1996
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
 3.843764 5.299093
sample estimates:
mean of x
 4.571429
```

Here, we would fail to reject the null hypothesis that the true mean is 5.

Example:

```
> x=c(1,2,1,4,3,2,4,3,2)
> t.test(x,mu=5)


One Sample t-test

data:  x
t = -6.7823, df = 8, p-value = 0.0001403
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
 1.575551 3.313338
sample estimates:
mean of x
 2.444444
```

Here, we reject the null hypothesis that the true mean is 5.