A hypothesis $H$ is a claim or conjecture about a feature of a population, often expressed in terms of the value(s) of a parameter. For instance

$$H : \mu = 15 \quad , \quad H : \mu \neq 15$$

$$H : \sigma^2 = 2 \quad , \quad H : \sigma^2 < 2$$

Suppose we have two contrasting hypotheses, $H_0$ and $H_1$.

$H_0$ would be called the null hypothesis and represents the “current belief” of “status quo”. $H_1$ would be called the alternative hypothesis and represents a “new belief” or a “research hypothesis”.

And suppose we also have data directly related to the hypotheses $H_0$ and $H_1$.

The test of hypothesis problem consists of deciding whether the data offer support in favor of $H_0$ or in favor of $H_1$.

This problem is analogous to the problem faced by a jury member in a trial involving a person accused of a crime. In this case we have:

$H_0$: the accused is innocent

$H_1$: the accused is guilty

Data: the evidence presented by the defense and prosecutor.
A test of hypothesis is a rule to decide, based on the data, when we should reject $H_0$ (it also follow from the test when not to reject $H_0$). Any test has two “ingredients”:

- The *test statistic*, which is the function of the data that the test uses to decide between $H_0$ and $H_1$;
- The *rejection region*, which is the set of values of the test statistic that lead to the rejection of $H_0$.

For any test of hypothesis there are two possible types of errors:

- *type I error*: reject $H_0$ when $H_0$ is true
- *type II error*: do not reject $H_0$ (or accept $H_1$) when $H_1$ is true

By viewing the testing of hypothesis problem as a decision problem with two possible actions we have, depending on the state of nature and our decision, four possible end results:

<table>
<thead>
<tr>
<th></th>
<th>$H_0$ true</th>
<th>$H_1$ true</th>
</tr>
</thead>
<tbody>
<tr>
<td>reject $H_0$</td>
<td>type I error</td>
<td>correct decision</td>
</tr>
<tr>
<td>do not reject $H_0$</td>
<td>correct decision</td>
<td>type II error</td>
</tr>
</tbody>
</table>

As for any statistical procedure there are good and bad tests, and some tests are better than others. We would assess the goodness of a test
by its probabilities of type I and type II errors:

\[ \alpha = P(\text{type I error}) \quad \text{and} \quad \beta = P(\text{type II error}) \]

**Example:** Let \( X_1, \ldots, X_{25} \) be a random sample from the \( N(\mu, 4) \) distribution, where \( \mu \) is unknown. We want to test the hypotheses:

\[ H_0 : \mu = 1 \quad \text{versus} \quad H_0 : \mu = 3 \]

(for this initial example we assume these two are the only possible values \( \mu \) can take). Consider the following test:

Test 1: reject \( H_0 \) if \( \bar{X} > 2 \). The test statistic of this test is \( \bar{X} \) and the rejection region is \((2, \infty)\). To compute \( \alpha \) and \( \beta \) it is key to know what is the distribution of \( \bar{X} \):

\[ \bar{X} \sim N(1, 4/25) \text{ when } H_0 \text{ is true}; \]
\[ \bar{X} \sim N(3, 4/25) \text{ when } H_1 \text{ is true.} \]

Then

\[ \alpha_1 = P(\text{type I error of test 1}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true}) \]
\[ = P(\bar{X} > 2 \text{ when } \mu = 1) = P(Z > 2.5) = 0.0062 \]

(if we do this test a large number of times, we will make a type I error on average 62 out of every 10000 times). Also

\[ \beta_1 = P(\text{type II error of test 1}) = P(\text{not to reject } H_0 \text{ when } H_1 \text{ is true}) \]
\[ = P(\bar{X} < 2 \text{ when } \mu = 3) = P(Z < -2.5) = 0.0062 \]
(the two probabilities are equal in this particular example, but they are usually different). Note that, except for extreme cases, in general it does not hold that $\alpha + \beta = 1$ (why?).

Consider Test 2: reject $H_0$ if $\bar{X} > 2.2$.

Then by a similar calculation as above we have

$$\alpha_2 = P(\text{type I error of test 2}) = 0.0013$$

$$\beta_2 = P(\text{type II error of test 2}) = 0.0228$$

Which test is better? In terms of type I error test 2 is better than test 1, but the opposite holds in terms of type II error. This example illustrate the following fact: it is not possible to modify a test to reduces the probabilities of both types of error. If one changes the test to reduce the probability of one type of error, the new test will inevitable have a larger probability for the other type of error.

The classical approach to deal with this dilemma and select a good test is the following:

The type I error is the one considered more serious, so we would like to use a test that has small $\alpha$. Then we fix a value of $\alpha$ close to zero, either subjectively or by tradition (say 0.05 or 0.01), and then find the test that has this $\alpha$ as its probability of type I error. This $\alpha$ will also be called the \textit{significance level} of the test.
Remarks

- For any testing problem there are several possible tests, but some of them make little or no sense. For the previous example tests of the form reject $H_0$ when $\bar{X} > c$, with $c > 1$ are worth considering, but tests of the form reject $H_0$ when $\bar{X} < c$ make no sense for this example (why?)
- We will consider only problems where the hypotheses refer to the values of a parameter of interest, generically called $\theta$. In addition, the null hypothesis will always be of the form $H_0 : \theta = \theta_0$, where $\theta_0$ is a known value. And the alternative hypothesis will always be one of the three possible forms: $H_1 : \theta < \theta_0$, $H_1 : \theta > \theta_0$ or $H_1 : \theta \neq \theta_0$. The first two are called one-sided alternatives and the last one is called two-sided alternative.
- The classical approach to test hypothesis does not treat $H_0$ and $H_1$ symmetrically, and does so in several ways. Because of that, the issue of what hypothesis to consider as the null and what as the alternative has an important bearing on the conclusion and interpretation of a test of hypotheses.
Example: A mixture of pulverized fuel ash and cement to be used for grouting should have a compressive strength of more than 1300 KN/m$^2$. The mixture cannot be used unless experimental evidence suggests that this requirement is met. Suppose that compressive strengths for specimens of this mixture are normally distributed with mean $\mu$ and standard deviation $\sigma = 60$.

(a) What are the appropriate null and alternative hypotheses?

$$H_0 : \mu = 1300 \quad \text{versus} \quad H_1 : \mu > 1300$$

(b) Let $\bar{x}$ be the sample mean of 20 specimens, and consider the test that rejects $H_0$ if $\bar{X} > 1331.26$. What is the probability of type I error?

Since $\bar{X} \sim N(1300, 60^2/20)$ when $H_0$ is true, we have

$$\alpha = P(\bar{X} > 1331.26 \text{ when } \mu = 1300)$$

$$= P(Z > \frac{1331.26 - 1300}{60/\sqrt{20}}) = 1 - P(Z \leq 2.33) = 0.0099$$

(c) What is the probability of type I error when $\mu = 1350$?

Since $\bar{X} \sim N(1350, 60^2/20)$ when $\mu = 1350$, we have

$$\beta(1350) = P(\bar{X} \leq 1331.26 \text{ when } \mu = 1350)$$

$$= P(Z \leq \frac{1331.26 - 1350}{60/\sqrt{20}}) = P(Z \leq -1.4) = 0.0808$$
(d) How the test in part (b) needs to be changed to obtain a test with a probability of type I error? What would be \( \beta(1350) \) for this new test?

Need to change the cut-off value 1331.21 with a new one, say \( c \), for which it holds

\[
0.05 = P(\bar{X} > c \text{ when } \mu = 1300) = P(Z > \frac{c - 1300}{60/\sqrt{20}})
\]

This implies that \( \frac{c - 1300}{60/\sqrt{20}} = z_{0.05} = 1.645 \), and solving for \( c \), we have \( c = 1322.07 \)

The above show an important point: there is a one-to-one correspondence between the cut-off value of a test and the probability of type I error.

The new test has a larger \( \alpha \) than the test in (b), so it must have a smaller \( \beta \). Indeed

\[
\beta(1350) = P(\bar{X} \leq 1322.07 \text{ when } \mu = 1350) = P(Z \leq \frac{1322.07 - 1350}{60/\sqrt{20}}) = P(Z \leq -2.08) = 0.0188
\]
Test About $\mu$ in Normal Populations:

Case When $\sigma^2$ is Known

Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution, where $\sigma^2$ is assume known. We want to test the null hypothesis $H_0 : \mu = \mu_0$, with $\mu_0$ a known constant, against one of the following alternatives hypothesis:

$$H_1 : \begin{cases} 
\mu < \mu_0 \\
\mu > \mu_0 \\
\mu \neq \mu_0
\end{cases}$$

The right test to use would depend on the alternatives hypothesis that is used. To start, consider the case when we want to test the hypotheses

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

The right kind of test in this case is to reject $H_0$ when $\bar{X} < c$, for some constant $c$ (why ?). The question then is how to choose the cut-off value $c$ ? We use the classical approach described before:

we fix a small probability $\alpha$ and choose the value of $c$ so the resulting test has this $\alpha$ as its probability of type I error (significance level)

$$\alpha = P(\bar{X} < c \text{ when } \mu = \mu_0) = P(Z < \frac{c - \mu_0}{\sigma / \sqrt{n}})$$

From this follows that $\frac{c - \mu_0}{\sigma / \sqrt{n}} = z_{1-\alpha} = -z_{\alpha}$, and $c = \mu_0 - z_{\alpha} \sigma / \sqrt{n}$.
Then for this case the test of hypotheses having significance level $\alpha$ is:

reject $H_0$ if $\bar{X} < \mu_0 - z_\alpha \sigma / \sqrt{n}$ or equivalently if $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} < -z_\alpha$.

Following a similar reasoning and computation we have that for the case when we want to test the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0$$

the test with significance level $\alpha$ is:

reject $H_0$ if $\bar{X} > \mu_0 + z_\alpha \sigma / \sqrt{n}$ or equivalently if $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} > z_\alpha$.

Note that in the previous two cases the rejection region is one-sided and of the same “form” as the alternative hypothesis.

Finally, to test the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0$$

with significance level $\alpha$ we use the test:

reject $H_0$ if $|Z| = \left| \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \right| > z_{\alpha/2}$

Note that in this case the rejection region is two-sided, as is the alternative hypothesis. Values of $Z$ that are significantly different from zero, either positive or negative, provide evidence in favor of $H_1$, so they call for the rejection of $H_0$.

Note also that in this case we use the $z$-value corresponding to $\alpha/2$ instead of $\alpha$. 
Sample Size Calculation

Suppose we want to test the hypotheses

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

and for that we can take observations $X_1, \ldots, X_n$, assumed to be a random sample from a $N(\mu, \sigma^2)$ distribution with $\sigma$ known, but we have not decided yet how many we would take. After we collect the data, we would like to use the test with significance level $\alpha$ (as usual), and in addition we want the test to have probability of type II error equal to $\beta$ when $\mu = \mu_1$ ($\alpha, \beta$ and $\mu_1$ are given). The question is: what the sample size $n$ needs to be for these to hold?

As we saw before, the test with significance level $\alpha$ is:

reject $H_0$ if $\bar{X} < \mu_0 - z_\alpha \sigma / \sqrt{n}$ or equivalently if $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} < -z_\alpha$.

Also, if we want that this test has probability of type II error when $\mu = \mu_1$ equal to $\beta$, it must hold that

$$\beta = \beta(\mu_1) = P(\bar{X} > \mu_0 - z_\alpha \sigma / \sqrt{n} \text{ when } \mu = \mu_1)$$

$$= P(\tilde{Z} > \frac{\mu_0 - \mu_1}{\sigma / \sqrt{n}} - z_\alpha) \quad \text{where } \tilde{Z} \sim N(0, 1)$$

which is equivalent to

$$P(\tilde{Z} \leq \frac{\mu_0 - \mu_1}{\sigma / \sqrt{n}} - z_\alpha) = 1 - \beta$$
The latter equation implies that \( \frac{\mu_0 - \mu_1}{\sigma / \sqrt{n}} - z_\alpha = z_\beta \), and solving for \( n \) in this equation we get

\[
n = \left( \frac{\sigma (z_\alpha + z_\beta)}{\mu_0 - \mu_1} \right)^2
\]

The required sample size is the above value rounded up.

The same sample size as above would be used for the case we are testing the hypotheses

\[
H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0
\]

But for the case we are testing the hypotheses

\[
H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0
\]

the required same sample size is

\[
n = \left( \frac{\sigma (z_{\alpha/2} + z_\beta)}{\mu_0 - \mu_1} \right)^2
\]
Example: The drying time of a certain type of paint is known to be normally distributed with mean 75 min and standard deviation 9 min. A chemical lab has proposed a new additive that claim to decrease the mean drying time. It is believed that drying times with the additive will remain normally distributed with the same standard deviation. To check the lab’s claim we want to test the hypothesis

\[ H_0 : \mu = 75 \quad \text{versus} \quad H_1 : \mu < 75 \]

(a) Test this hypothesis based on a random sample of 25 observations for which \( \bar{x} = 72.3 \), using \( \alpha = 0.01 \).

The test with significance level \( \alpha = 0.01 \) is:

reject \( H_0 \) if \( Z < -z_{0.01} = -2.33 \). For the observed data we have

\[ z_{\text{obs}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{72.3 - 75}{9/5} = -1.5 \]

Since \( z_{\text{obs}} \) does not fall in the rejection region \((-1.5 \neq -2.33)\), the conclusion is not to reject \( H_0 \).

(b) What is \( \alpha \) for the test that rejects \( H_0 \) if \( Z < -2.88 \) ?

\[ \alpha = P(Z < -2.88) = 0.002 \]
(c) For the test in (b), compute the probability of type II error when 
\( \mu = 70 \)

First note that \( Z < -2.88 \) is equivalent to \( \bar{X} < 75 - 2.88 \left( \frac{9}{5} \right) = 69.82 \) (why?). Then

\[ \beta(70) = P(\bar{X} > 69.82 \text{ when } \mu = 70) = P(\tilde{Z} > -0.1) = 0.5398 \]

(d) If the test in part (b) is used, what is the sample size required to ensure that \( \beta(70) = 0.01 \) ?

\[ n = \left( \frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu_1} \right)^2 = \left( \frac{9(2.88 + 2.33)}{75 - 70} \right)^2 = 87.95 \]

so we need to collect \( n = 88 \) observations.
Test About $\mu$ When $n$ is Large

Suppose $X_1, \ldots, X_n$ is a random sample from any distribution, not necessarily normal, with mean $\mu$ and variance $\sigma^2$, both unknown.

And suppose we want to test the null hypothesis $H_0 : \mu = \mu_0$.

The test statistic in this case would be

$$ Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} $$

The key point to note is that, if $n$ is moderate or large, we have by the CLT that $Z \overset{\text{approx}}{\sim} N(0, 1)$ when $H_0$ is true.

Depending on the alternative hypothesis that is being tested, the test with significance level $\alpha$ is

- $H_1$ reject $H_0$ if
  - $\mu < \mu_0$ : $Z < -z_\alpha$
  - $\mu > \mu_0$ : $Z > z_\alpha$
  - $\mu \neq \mu_0$ : $|Z| > z_{\alpha/2}$
Test About $\mu$ in Normal Populations:

Case When $\sigma^2$ is Unknown

Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution, where now $\sigma^2$ is unknown. We want to test the null hypothesis $H_0 : \mu = \mu_0$ against one of the usual alternatives hypothesis $H_1$.

The test statistic in this case would be

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

and the key point to note is that $T \sim t_{(n-1)}$ when $H_0$ is true.

Depending on the alternative hypothesis that is being tested, the test with significance level $\alpha$ is

- $H_1$ reject $H_0$ if
  - $\mu < \mu_0$ $T < -t_{\alpha,n-1}$
  - $\mu > \mu_0$ $T > t_{\alpha,n-1}$
  - $\mu \neq \mu_0$ $|T| > t_{\frac{\alpha}{2},n-1}$
**Example:** A sample of twelve radon detectors of a certain type was selected, and each was exposed to 100pCi/L of radon. The resulting readings were:

105.6 90.9 91.2 96.9 96.5 91.3 100.1 105.0 99.6 107.7 103.3 92.4

(a) Do these data suggest that the population mean reading under these conditions differs from 100? Use $\alpha = 0.05$.

The hypotheses we need to test are

$$H_0 : \mu = 100 \quad \text{versus} \quad H_1 : \mu \neq 100$$

The test with significance level $\alpha = 0.05$ is:

reject $H_0$ if $|T| > t_{0.025,11} = 2.201$, where $T = \frac{\bar{X} - 100}{S/\sqrt{n}}$.

For these data we have $\bar{x} = 98.375$ and $s = 6.1095$, so

$$T_{\text{obs}} = \frac{98.375 - 100}{6.1095/\sqrt{12}} = -0.9214.$$  

Since $T_{\text{obs}}$ does not fall in the rejection region, we do not reject $H_0$ and conclude that the reading population mean of these radon detectors is not significantly different from 100.