Hypothesis testing

Research hypotheses are conjectures or suppositions that motivate the research.

Statistical hypotheses involve restating the research hypotheses in such a way that they may be addressed by statistical techniques.

Formally, a statistical hypothesis testing problem includes two hypotheses. These hypotheses are referred to as the *null hypothesis* (H_0), and the *alternative hypothesis* (H_A).

In statistical hypothesis testing, we start off believing the null hypothesis, and see if the data provide enough evidence to abandon our belief in H_0 in favor of the alternative hypothesis H_A.

In the previous example, $H_0 : \mu = \mu_0 = 211$ and we considered the alternative hypothesis ($H_A : \mu \neq \mu_0 = 211$).

We made plans to collect a sample of 12 US men (with a particular disease) and to record the cholesterol level of each. We want to know if the underlying population mean for this population is different than 211. The null and alternative hypotheses are as follows:

$$H_0 : \mu = \mu_0 = 211 \text{ and } H_A : \mu \neq \mu_0 = 211$$

We collect the sample of 12 US men (with a particular disease), record the cholesterol level of each, and figure out that the sample mean, $\bar{x}$, is 217.

We want to know if the difference between 211 and 217 is too big to be attributed to chance alone. If there is evidence that the sample could NOT have come from a population with mean 211, we reject the null hypothesis.

To see if there is evidence that the sample could NOT have come from a population with mean $\mu = 211$, we assume that it did. We then calculate the probability that we see a sample mean as far away (or farther) from 211 than the one we observed.

If this probability is very small, we conclude that our assumption was wrong and that the sample could NOT have come from a population with mean 211 (we reject the null hypothesis).

What do we know about the distribution of the sample mean ?
We know by the CLT that

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

is approximately standard normal.

Assume that the null hypothesis, H_0, is true. If $\mu = 211$, then

$$Z = \frac{\bar{X} - 211}{46/\sqrt{12}}$$

is approximately standard normal.

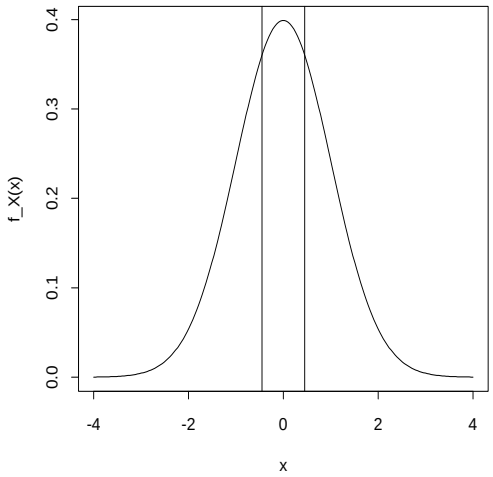
We evaluate

$$z = \frac{217 - 211}{46/\sqrt{12}}$$

and see if this looks like it might be a sample from a standard normal. In particular, we calculate the probability that we would observe a z this far away (or farther) from 0.

$$z = \frac{217 - 211}{46/\sqrt{12}} = 0.45$$

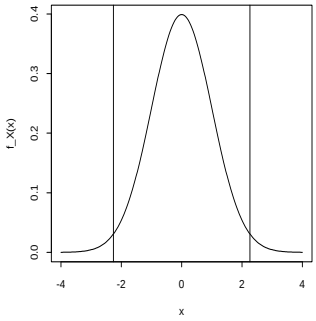
$$z = \frac{217 - 211}{46/\sqrt{12}} = 0.45$$



$P(Z < -0.45) + P(Z > 0.45) = 0.326 + 0.326 = 0.652$. This probability is relatively large, so it doesn't seem unusual that z is a sample from a standard normal. Thus, our assumption that true population mean was 211 seems reasonable. As a result, we do not reject the null.

If instead the sample mean had been $\bar{x} = 241$ We evaluate $z = \frac{241-211}{46/\sqrt{12}}$ and see if this looks like it might be a sample from a standard normal. In particular, we calculate the probability that we would observe a z this far away (or farther) from 0.

$$z = \frac{241 - 211}{46/\sqrt{12}} = 2.26$$



$P(Z < -2.26) + P(Z > 2.26) = 0.012 + 0.012 = 0.024$. This probability is very small, so it does seem unusual that z is a sample from a standard normal. Thus, our assumption that the true population mean was 211 seems unreasonable. As a result, we reject the null.

Some Definitions

Based on our analysis of the data and resulting conclusions, there are two types of errors that we could have made:

• A **type I error** occurs when the null hypothesis is rejected when it is, in fact, true. The probability of a type I error is denoted by α.

• A **type II error** occurs when the null hypothesis is not rejected when it is, in fact, false. The probability of a type II error is denoted by β.

A *test statistic* is a statistic (function of random variables which does not contain any unknown parameters) used to evaluate the validity of a null hypothesis. Oftentimes, it is a function of observable random variables (e.g., $\bar{X}$) and the parameter of interest in the hypothesis test (e.g., $\mu = \mu_0$). It is constructed so that we know its distribution.

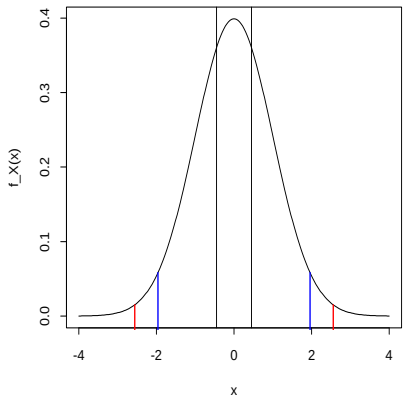
A **p-value** for a hypothesis test is the probability, computed under the null hypothesis, that the value of the test statistic would be as extreme or more extreme than the one observed.

If the p-value is small, this denotes an unusual observation if H_0 is true. In this case, we might doubt the fact that H_0 is true and reject the null hypothesis.

How "small" is small ?

When we define the type I error, α, we are defining what we mean by "unusual" observation. H_0 is rejected when the p-value is less than α.

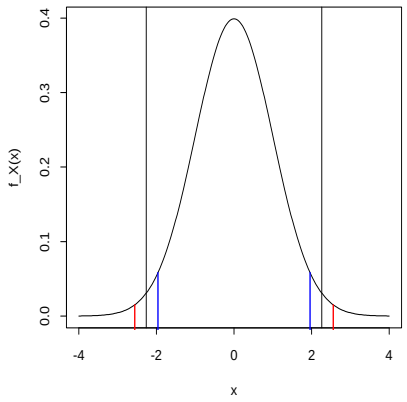
$$z = \frac{217 - 211}{46/\sqrt{12}} = 0.45$$



$$P(Z < -0.45) + P(Z > 0.45) = 0.326 + 0.326 = 0.652$$

Here, the p-value is 0.652.
Consider $\alpha = 0.05$ ($z_{\alpha/2}$ shown in blue), H_0 is not rejected since p-value is not less than α.
Consider $\alpha = 0.01$ ($z_{\alpha/2}$ shown in red), H_0 is not rejected since p-value is not less than α.

$$z = \frac{241 - 211}{46/\sqrt{12}} = 2.26$$



$$P(Z < -2.26) + P(Z > 2.26) = 0.012 + 0.012 = 0.024$$

Here, the p-value is 0.024.
Consider $\alpha = 0.05$ ($z_{\alpha/2}$ shown in blue), H_0 IS rejected since p-value is less than α.
Consider $\alpha = 0.01$ ($z_{\alpha/2}$ shown in red), H_0 is not rejected since p-value is not less than α.

General Approach to Hypothesis Testing

1. Define the research question and formulate the appropriate null and alternative hypotheses.
2. Decide on your type I error rate (α).
3. Assume that the null is true, construct a test statistic, and identify what is known about the distribution of the test statistic under this assumption.
4. Collect a sample and evaluate the test statistic.
5. Calculate the p-value. Note that calculation of this value depends on the form of the alternative hypothesis.
6. If the p-value is less than α, reject the null. Otherwise, do not reject the null.

Note: We have not yet considered type II error rates or size of the sample. More on this later.

From the Physician Worklife Study, consider the random variable X which is a measure of time pressure. Assume that we know that the standard deviation of X, σ, is 16 minutes.

X= time needed for a new patient visit - time allotted

We would like to know whether or not physicians have adequate time for office visits with new patients. Let's turn this into a statistical hypothesis testing problem.

Let μ be the mean of X in the population of practicing US physcians. The null hypothesis would simply be that physicians on average are allotted just the right amount of time.

$$H_0 : \mu = 0$$

Consider a *two-sided* alternative hypothesis. A two-sided alternative hypothesis is just a negation of the null hypothesis, allowing for the population parameter to be either larger or smaller than the value expressed by the null hypothesis.

Two-sided alternative hypothesis

$$H_A : \mu \neq 0$$

Decide on the type I error rate: $\alpha = 0.05$

Recall that X is the time pressure variable and μ is the mean of X in the general population of US physicians. For large samples, we know that the distribution of

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is approximately standard normal.

Thus, if we assume that the null hypothesis is true, the test statistic

$$\frac{\bar{X}}{\sigma/\sqrt{n}}$$

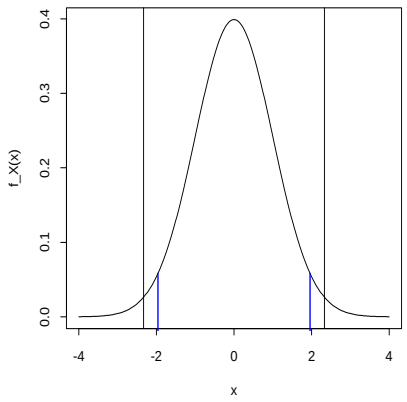
will have a standard normal distribution, approximately.

Collect the sample and evaluate the test statistic In a sample of 40 physicians, $\bar{x} = 5.9$ minutes. Therefore,

$$z = \frac{5.9}{16/\sqrt{40}} = 2.33$$

$$z = \frac{5.9}{16/\sqrt{40}} = 2.33$$

Calculate the p-value:



$$P(Z < -2.33) + P(Z > 2.33) = 0.0099 + 0.0099 = 0.0198$$

Here, the p-value is 0.0198.

Consider $\alpha = 0.05$ ($z_{\alpha/2}$ shown in blue), H_0 is rejected since p-value is less than α. We conclude that physicians, on average, are under time pressure.

Example: one-sided alternative

Consider the example EXACTLY as we have it above EXCEPT specify a one-sided alternative hypothesis. The null and alternative hypotheses are now:

$$H_0 : \mu \leq 0 \text{ versus } H_A : \mu > 0,$$

We'll keep the type I error rate at 0.05 ($\alpha = 0.05$)
The form of the test statistic has not changed. If we assume that the null hypothesis is true,

$$\frac{\bar{X}}{\sigma/\sqrt{n}}$$

will have a standard normal distribution, approximately.

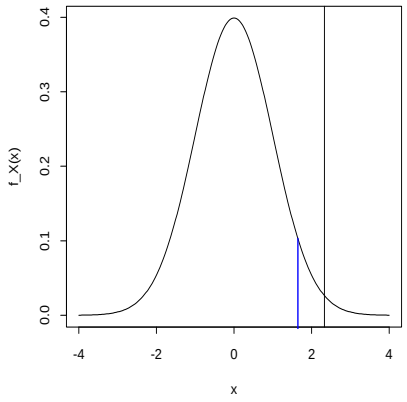
Again, we evaluate the test statistic. In our sample of 40 physicians, $\bar{x} = 5.9$ minutes. Therefore,

$$z = \frac{5.9}{16/\sqrt{40}} = 2.33$$

Find the p-value. Here, "as extreme or more extreme" is just as extreme or greater !!

Calculate the p-value:

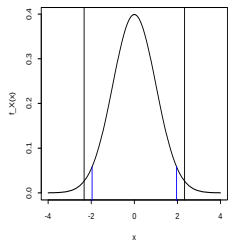
$$z = \frac{5.9}{16/\sqrt{40}} = 2.33$$



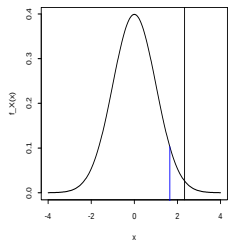
$$P(Z \geq 2.33) = 0.0099$$

Here, the p-value is 0.0099.
Consider $\alpha = 0.05$ (z_α shown in blue), H_0 is rejected since p-value is less than α. We conclude that physicians, on average, are under positive time pressure.

Summary:



p-value $= P(Z \leq -2.33) + P(Z \geq 2.33) = 0.0099 + 0.0099 = 0.0198$
Consider $\alpha = 0.05$ ($z_{\alpha/2}$ shown in blue), H_0 is rejected since p-value is less than α.



p-value $= P(Z \geq 2.33) = 0.0099$
Consider $\alpha = 0.05$ (z_α shown in blue), H_0 is rejected since p-value is less than α.

Note:

We would have rejected the null hypothesis in the first case if the value of the test statistic had been smaller than or equal to -1.96 ($-z_{\alpha/2}$) or if it had been greater than or equal to 1.96 ($z_{\alpha/2}$)
(for $\alpha = 0.05$).

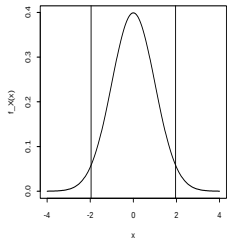
We would have rejected the null hypothesis in the second case if the value of the test statistic had been greater than or equal to 1.645 (z_α) (again $\alpha = 0.05$).

Critical values of a test statistic are the values of the test statistic that determine whether or not the null hypothesis is rejected.

In the test that we just discussed with a two-sided alternative, the critical values are -1.96 and 1.96. In the test with a one-sided alternative, the critical value is 1.645.

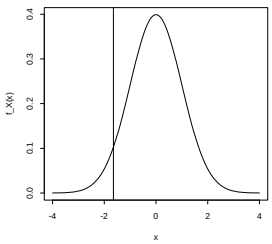
Suppose we are in the standard situation where we are interested in testing $H_0 : \mu = \mu_0$ against some specified alternative. We have a test statistic Z that is Normally distributed. Consider $\alpha = 0.05$.

I. $H_0 : \mu = \mu_0$ and $H_A : \mu \neq \mu_0$



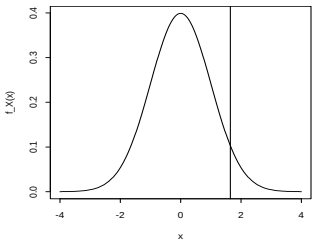
We reject H_0 if $|z| \geq 1.96$. For general α, we reject H_0 if $|z| \geq z_{\alpha/2}$.

II. $H_0 : \mu \geq \mu_0$ and $H_A : \mu < \mu_0$



We reject H_0 if $z \leq -1.645$. For general α, we reject H_0 if $z \leq -z_\alpha$.

III. $H_0 : \mu \leq \mu_0$ and $H_A : \mu > \mu_0$



We reject H_0 if $z \geq 1.645$. For general α, we reject H_0 if $z \geq z_\alpha$.

Suppose we collect 8 pairs of twins. The first twin in the pair is healthy; the second is not. For each twin, we measure grey matter density. Processed data from the 8 pairs is shown below (units not given).

Pair	Twin 1	Twin 2
1	134	128
2	139	136
3	176	163
4	152	153
5	166	157
6	161	154
7	129	125
8	122	124

Is grey matter density in the populations significantly different ?

Consider the population differences, d,

where $\bar{d}$ is the sample mean of the differences, δ is the population mean of the differences, and s_d is the sample standard deviation of the differences. The quantity

$$t = \frac{\bar{d} - \delta}{s_d/\sqrt{n}},$$

is approximately t distributed with $n - 1$ degrees of freedom.

Pair	Twin 1	Twin 2	difference
1	134	128	6
2	139	136	3
3	176	163	13
4	152	153	-1
5	166	157	9
6	161	154	7
7	129	125	4
8	122	124	-2

In this sample, $\bar{d} = 4.875$ and $s_d = 4.998$.

Test $H_0 : \delta = 0$ versus $H_A : \delta \neq 0$, at significance level $\alpha = 0.05$.

Compute the test statistic

$$\frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{4.875}{4.998/\sqrt{8}} = 2.76$$

For 7 degrees of freedom $t_{0.025} = 2.365$, so the null hypothesis is rejected and we conclude that there is a non-zero difference in grey matter density.

Paired Data

Paired data is where two observations are taken from the same individual or from two individuals who are very similar. For example,

- one observation from each eye of an individual,
- a "before" and "after" observation from an individual,
- the response to two different treatments on the same individual,
- a response for both a "case" and a "control" from similar individuals, and
- responses from twins.

In many cases, the difference in response between the two treatments or states can be computed for each pair (such as "twin 1" − "twin 2").

Suppose that we will sample n pairs (X,Y). From each pair we compute the difference $D = X - Y$. Let δ denote the true difference in population means ($\delta = \mu_1 - \mu_2$), where μ_1 is population mean of X and μ_2 is population mean of Y.

With paired data, we are usually interested in testing the hypotheses:

Two-sided

$$H_0 : \delta = 0 \text{ versus } H_A : \delta \neq 0$$

One-sided

$$H_0 : \delta \leq 0 \text{ versus } H_A : \delta > 0$$

or

$$H_0 : \delta \geq 0 \text{ versus } H_A : \delta < 0$$