Starting Off

In America, our criminal justice system is setup where people are suppose to be presumed innocent and it is the state's job to prove that they are guilty. Why is our system set up this way and not vice versa?

Hypothesis Testing

Data Science Immersive

Introduction to Inferences

There are two types of statistical inferences:

1. Estimation - Use information from the sample to estimate (or predict) the parameter of interest.

Using the result of a poll about the president's current approval rating to estimate (or predict) his or her true current approval rating nationwide.

2. Statistical Test - Use information from the sample to determine whether a certain statement about the parameter of interest is true.

The president's job approval rating is above 50%.

Hypothesis Testing

A hypothesis, in statistics, is a statement about a population where this statement typically is represented by some specific numerical value.

In testing a hypothesis, we use a method where we gather data in an effort to gather evidence about the hypothesis.

Below these are summarized into seven steps to conduct a test of a hypothesis.

1. **Setting up two competing hypotheses and check conditions.**
2. **Set some level of significance called alpha.**
3. **Identify the sampling distribution.**
4. **Calculate a test statistic.**
5. **Calculate probability value (p-value), or find rejection region**.
6. **Make a test decision about the null hypothesis**.
7. **State an overall conclusion**.

The Null and Alternative Hypothesis

The two hypotheses are named the null hypothesis and the alternative hypothesis.

Null hypothesis

The null hypothesis is typically denoted as *H0*. The null hypothesis states the "status quo". This hypothesis is assumed to be true until there is evidence to suggest otherwise.

Alternative hypothesis

The alternative hypothesis is typically denoted as *Ha* or *H1*. This is the statement that one wants to conclude. It is also called the research hypothesis.

We usually set the hypothesis that one wants to conclude as the alternative hypothesis, also called the research hypothesis.

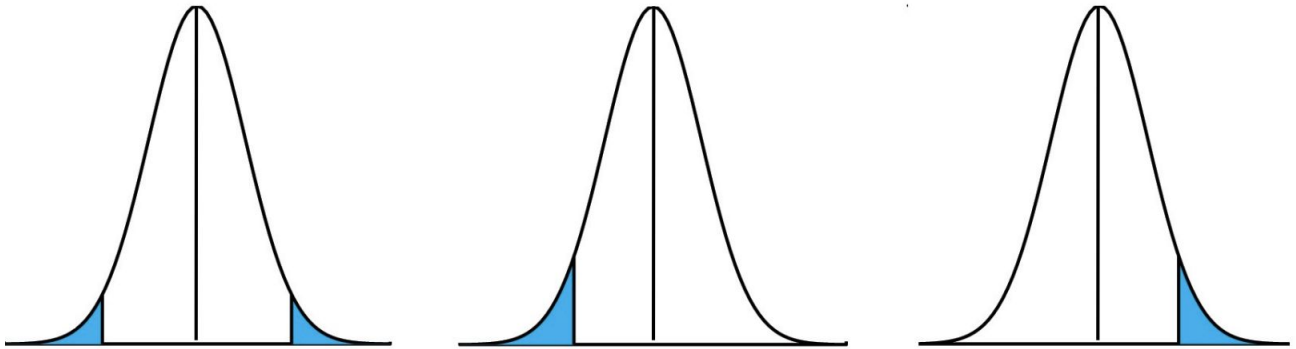
The Null and Alternative Hypothesis

There are three types of alternative hypotheses:

Two-sided test - The population parameter is not equal to a certain value.

Left-tailed test - The population parameter is less than a certain value.

Right-tailed test - The population parameter is greater than a certain value.



The Null and Alternative Hypothesis

The null hypothesis in each case would be:

$$H_0 : p = p_0 \qquad H_0 : \mu = \mu_0$$

Below are the possible alternative hypothesis from which we would select only one of them based on the research question. The symbols p_0 and μ_0 are just used in these general statements. In practice, these get replaced by the parameter value being tested.

Two-sided test - $H_a : p \neq p_0$ $\qquad H_a : \mu \neq \mu_0$

Left-tailed test - $H_a : p < p_0$ $\qquad H_a : \mu < \mu_0$

Right-tailed test - $H_a : p > p_0$ $\qquad H_a : \mu > \mu_0$

Practice: Null and Alternative Hypothesis

When debating the marketing plan for Flatiron School in NYC, the following question is asked: "Over half of the students are from NYC?"

To answer this question, we can set it up as a hypothesis testing problem and use data collected to answer it. This example is about a population proportion and thus we set up the hypotheses in terms of *p*. Here the value p_o is 0.5 since more than 0.5 constitute a majority.

The hypothesis set up would be a right-tailed test:

$$H_0 : p = 0.5 \qquad H_a : p > 0.5$$

Practice: Null and Alternative Hypothesis

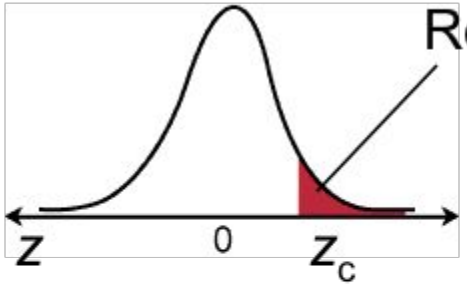
1.A consumer test agency wants to see the whether the mean lifetime of a brand of tires is greater than 42,000 miles as the tire manufacturer advertises.

2. The length of a cut of lumber from a store is supposed to be 8.5 feet. A builder wants to check whether the shipment of lumber she receives has a mean length different from 8.5 feet.

3. A political news company believes the national approval rating for the current president has fallen below 40%.

Rejection Regions

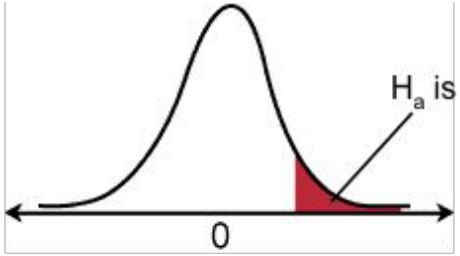
Sampling distribution for $\bar{x}$



The **rejection region** is the range of values for which the *null hypothesis is not probable*. It is always in the direction of the alternative hypothesis. Its area is equal to α.

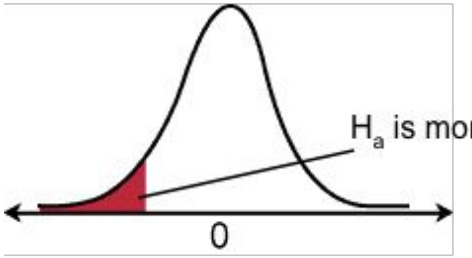
A **critical value** separates the rejection region from the non-rejection region.

The Null and Alternative Hypothesis



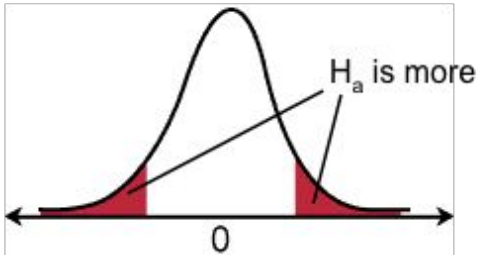
Right-tail test

$H_a: \mu > \text{value}$



Left-tail test

$H_a: \mu < \text{value}$



Two-tail test

$H_a: \mu \neq \text{value}$

Type I and Type II Errors

How do we determine whether to reject the null hypothesis? It begins the level of significance α, which is the probability of the Type I error.

What is Type I error and what is Type II error?

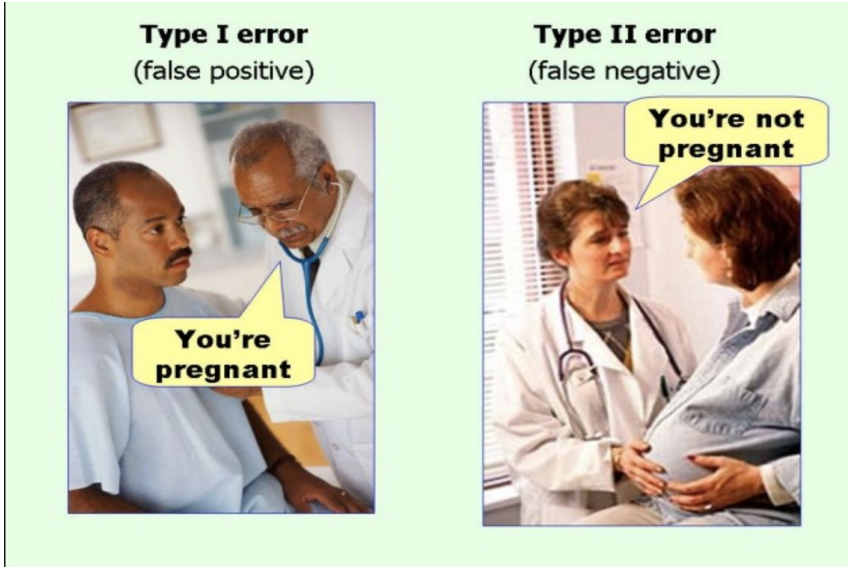
Decision	*Reality*	
	H_0 is true	H_0 is false
Reject H_0	Type I error	Correct
Fail to Reject H_0	Correct	Type II error

The probability of Type II error is denoted by: β

Errors

Let's set up the null and alternative hypotheses so that a Type I error is more serious.

- Type I error: false positive
- Type II error: false negative



Important Terms

Test statistic: The sample statistic one uses to either reject H_o (and conclude H_a) or not to reject H_o.

Critical values: The values of the test statistic that separate the rejection and non-rejection regions.

Rejection region: the set of values for the test statistic that leads to rejection of H_o.

Non-rejection region: the set of values not in the rejection region that leads to non-rejection of H_o.

***P*-value:** The *p*-value (or probability value) is the probability that the test statistic equals the observed value or a more extreme value under the assumption that the null hypothesis is true.

Hypothesis Testing

Below these are summarized into seven such steps to conducting a test of a hypothesis.

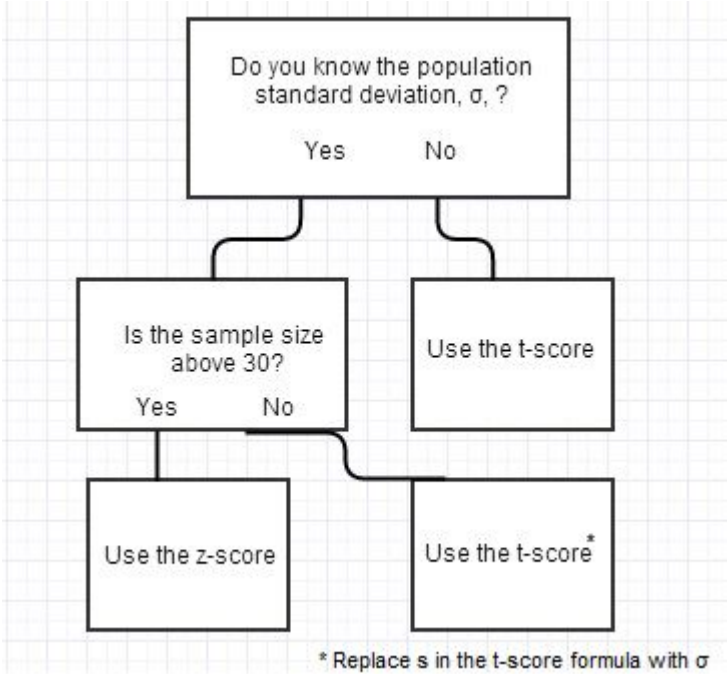
1. **Setting up two competing hypotheses and check conditions**
2. **Set some level of significance called alpha.**
3. **Identify the sampling distribution.**
4. **Calculate a test statistic.**
5. **Calculate probability value (p-value), or find rejection region**.
6. **Make a test decision about the null hypothesis**.
7. **State an overall conclusion**.

Hypothesis testing

A **z-score** and a **t-score** are both used in **hypothesis testing**.

The general rule of thumb for *when* to use a t-score is when your sample:

- Has a sample size below 30,
- Has an unknown population standard deviation.



*

One-Sided *T*-Test for a Mean

A health group claims the mean sodium content in one serving of a cereal is greater than 230 mg. You work for a national health service and are asked to test this claim. You find that a random sample of 52 servings has a mean sodium content of 232 mg and a standard deviation of 10 mg. At $\alpha = 0.05$, do you have enough evidence to accept the group's claim?

1. Write the null and alternative hypothesis.

2. State the level of significance. $\alpha = 0.05$

3. Determine the sampling distribution.

Since the sample size is at least 30, the sampling distribution is normal, but the population standard deviation is unknown.

4. Find the test statistic and standardize it.

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{52}} = 1.387$$

$n = 52 \quad s = 10$

$\bar{x} = 232$

$$z = \frac{232 - 230}{1.387} = 1.44$$

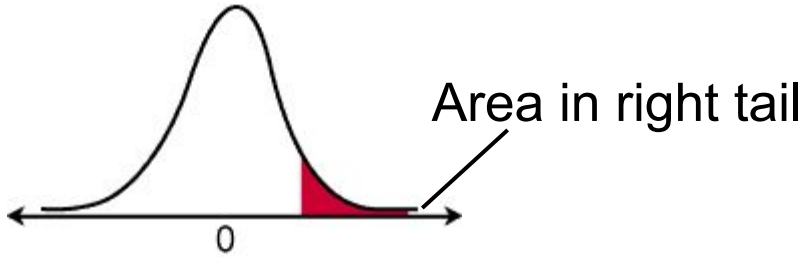
Test statistic

5a. Calculate the P-value for the test statistic.

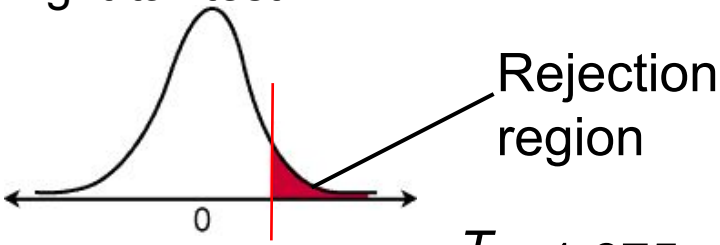
Since this is a right-tail test, the P-value is the area found to the right of $t = 1.44$ in the normal distribution.

From the table P = 1 – 0.9220

P = 0.078



5b. Find the critical values. Since H_a contains the > symbol, this is a right-tail test.



$T_c = 1.675$

$\alpha = 0.05$ so find 0.0500 in the T table which corresponds to a T_c of 1.675

6. Make your decision.

$t = 1.44 < T_c = 1.675$

The test statistic does not fall in the rejection region, so fail to reject H_0

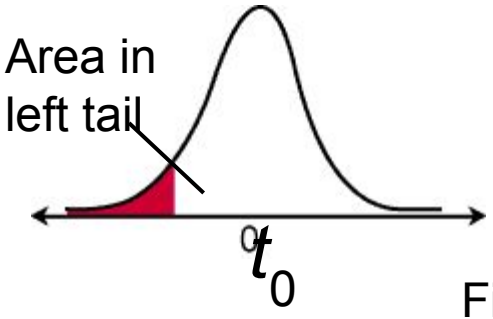
7. Interpret your decision.

There is not enough evidence to accept the group's claim that there is greater than 230 mg of sodium in one serving of its cereal.

*

The *t* Sampling Distribution

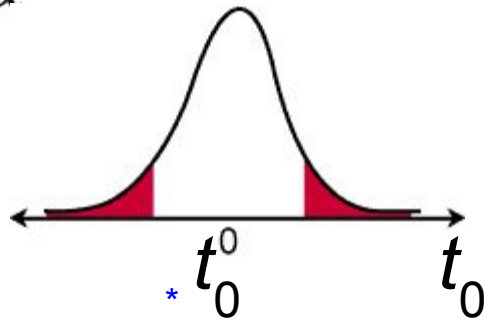
Find the critical value t_0 for a left-tailed test given α = 0.01 and n = 18.



d.f. = 18 – 1 = 17

t_0 = –2.567

Find the critical values $-t_0$ and t_0 for a two-tailed test given α = 0.05 and n = 11.



$-t_0$ = –2.228 and t_0 = 2.228

d.f. = 11 – 1 = 10

Testing μ –Small Sample

A university says the mean number of classroom hours per week for full-time faculty is 11.0. A random sample of the number of classroom hours for full-time faculty for one week is listed below. You work for a student organization and are asked to test this claim. At α = 0.01, do you have enough evidence to reject the university's claim?

11.8 8.6 12.6 7.9 6.4 10.4 13.6 9.1

1. Write the null and alternative hypothesis

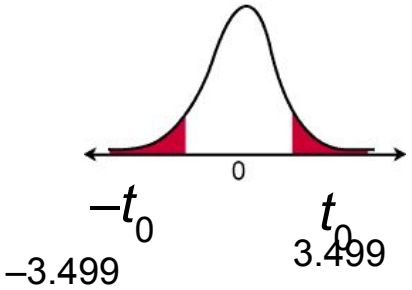
2. State the level of significance α = 0.01

3. Determine the sampling distribution

Since the sample size is 8, the sampling distribution is a t-distribution with 8 – 1 = 7 d.f.

*

Since H_a contains the ≠ symbol, this is a two-tail test.



4. Find the critical values.

5a. Find the rejection region.

5 b. Find the test statistic and standardize it

$n = 8$ $\bar{x}$ 10.050 $s = 2.485$

$$t = \frac{10.050 - 11.0}{\frac{2.485}{\sqrt{8}}} = \frac{-0.95}{0.878} = -1.08$$

6. Make your decision.

t = –1.08 does not fall in the rejection region, so fail to reject H_0 at = 0.01

7. Interpret your decision.

There is not enough evidence to reject the university's claim that faculty spend a mean of 11 classroom hours.

Practice

The mean length of the lumber is supposed to be 8.5 feet. A builder wants to check whether the shipment of lumber she receives has a mean length different from 8.5 feet. If the builder observes that the sample mean of 61 pieces of lumber is 8.3 feet with a sample standard deviation of 1.2 feet. What will she conclude?

1. Setting up two competing hypotheses and check conditions
2. Set some level of significance called alpha.
3. Identify the sampling distribution.
4. Calculate a test statistic.
5. Calculate probability value (p-value), or find rejection region.
6. Make a test decision about the null hypothesis.
7. State an overall conclusion.

*

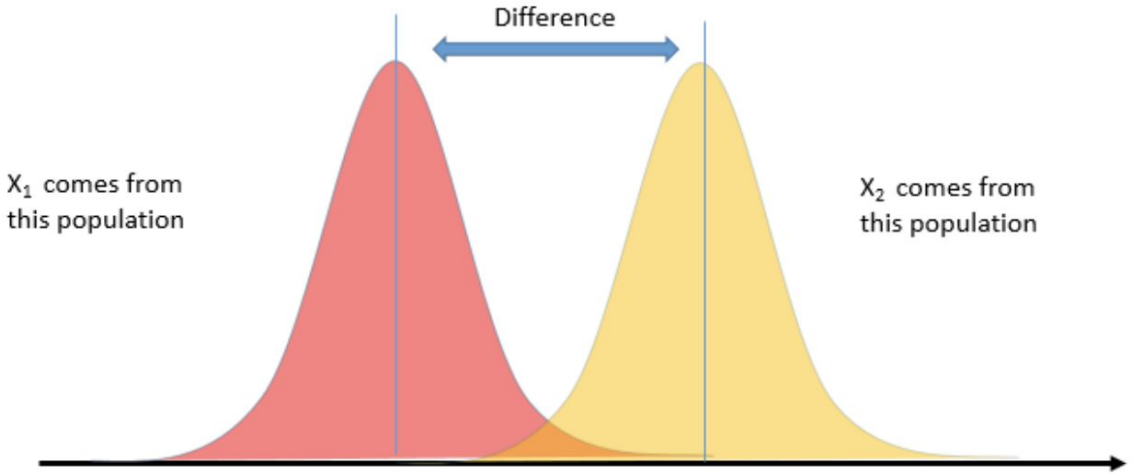
*

Comparing Two Populations

Males and females were asked about what they would do if they received a $100 bill by mail, addressed to their neighbor, but wrongly delivered to them. Would they return it to their neighbor? Of the 69 males sampled, 52 said "yes" and of the 131 females sampled, 120 said "yes."

Does the data indicate that the proportions that said "yes" are different for male and female? How do we begin to answer this question?

Comparing Two Populations



Examples

- Chemistry - do inputs from two different barley fields produce different yields?
- Astrophysics - do star systems with near-orbiting gas giants have hotter stars?
- Economics - demography, surveys, etc.
- Medicine - BMI vs. Hypertension, etc.
- Business - which ad is more effective given engagement?

Comparing Two Population Proportions

When we have a categorical variable of interest measured in two populations, it is quite often that we are interested in comparing the proportions of a certain category for the two populations.

When we want to check whether two proportions are different or the same, the two-tailed test is appropriate.

$$H_0 : p_1 - p_2 = 0 \qquad H_a : p_1 - p_2 \neq 0$$

OR

$$H_0 : p_1 = p_2 \qquad H_a : p_1 \neq p_2$$

Comparing Two Population Proportions

When the observed number of successes and the observed number of failures are greater than or equal to 5 **for both populations**, then the sampling distribution of $\hat{p_1} - \hat{p_2}$ is approximately normal and we can use z-methods.

The formula for the test statistic is:

$$Z^* = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p^*(1-p^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where,

$$p^* = \frac{x_1 + x_2}{n_1 + n_2}$$

Comparing Two Populations

Males and females were asked about what they would do if they received a $100 bill by mail, addressed to their neighbor, but wrongly delivered to them. Would they return it to their neighbor? Of the 69 males sampled, 52 said "yes" and of the 131 females sampled, 120 said "yes."

Does the data indicate that the proportions that said "yes" are different for male and female? How do we begin to answer this question?

$$Z^* = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p^*(1-p^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad p^* = \frac{x_1 + x_2}{n_1 + n_2}$$

Comparing Two Population Proportions

In 1980, of 750 men 20-34 years old, 130 were found to be overweight. Whereas, in 1990, of 700 men, 20-34 years old, 160 were found to be overweight. At the 5% significance level, do the data provide sufficient evidence to conclude that for men 20-34 years old, a higher percentage were overweight in 1990 than 10 years earlier?

$$p^* = \frac{x_1 + x_2}{n_1 + n_2}$$

$$Z^* = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p^*(1-p^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Comparing Two Population Means

Are the Data Independent Samples or Dependent Samples?

The samples from two populations are independent if the samples selected from one of the populations has no relationship with the samples selected from the other population.

The samples are dependent (also called paired data) if each measurement in one sample is matched or paired with a particular measurement in the other sample.

https://newonlinecourses.science.psu.edu/stat500/lesson/7/7.3/7.3.2

Comparing Two Independent Population Means:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

In order to find a confidence interval for $\mu_1 - \mu_2$ and perform a hypothesis test, we need to find the sampling distribution of $\bar{x}_1 - \bar{x}_2$.

We can show that when the sample sizes are large or the samples from each population are normal and the samples are taken independently, then $\bar{y}_1 - \bar{y}_2$ is normal with mean $\mu_1 - \mu_2$ and standard deviation is

$$\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

Pooled Variances VS. Non-Pooled Variances

In most cases, σ1 and σ2 are unknown, and they have to be estimated using the sampe statistics. When the sample sizes are small, the estimates may not be that accurate and one may get a better estimate for the common standard deviation by pooling the data from both populations if the standard deviations for the two populations are not that different.

Given this, there are two options for estimating the variances for the independent samples:

1. Using pooled variances
2. Using unpooled (or unequal) variances

When we are reasonably sure that the two populations have nearly equal variances, then we use the pooled variances test. Otherwise, we use the separate variances test.

Pooled Variances

An informal check for this is to compare the ratio of the two sample standard deviations. When the sample sizes are nearly equal (admittedly "nearly equal" is somewhat ambiguous so often if sample sizes are small one requires they be equal), then a good Rule of Thumb to use is to see if this ratio falls from 0.5 to 2.

The assumptions/conditions are:

- The populations are independent
- The population variances are equal
- Each population is either normal or the sample size is large.

$$t = \frac{\bar{x}_1 - \bar{x_2} - 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Comparing Two Population Means

Then the common standard deviation can be estimated by the pooled standard deviation:

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

The test statistic is:

$$t^* = \frac{\bar{x}_1-\bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

with degrees of freedom equal to $df = n_1 + n_2 - 2$

Example: Comparing Packing Machines

In a packing plant, a machine packs cartons with jars. It is supposed that a new machine will pack faster on the average than the machine currently used. To test that hypothesis, the times it takes each machine to pack ten cartons are recorded. The results, in seconds, are shown in the following table. Assume that these samples come from normal distributions

New machine					Old machine				
42.1	41.3	42.4	43.2	41.8	42.7	43.8	42.5	43.1	44.0
41.0	41.8	42.8	42.3	42.7	43.6	43.3	43.5	41.7	44.1
$\bar{y}_1 = 42.14, s_1 = 0.683$					$\bar{y}_2 = 43.23, s_2 = 0.750$				

Example: Comparing Packing Machines

Assumption 1: Are these independent samples?

Assumption 2: Are these large samples or a normal population?

Assumption 3: Do the populations have equal variance?

Example: Comparing Packing Machines

The significance level is 5%. Since we may assume the population variances are equal, we first have to calculate the pooled standard deviation:

$$
\begin{aligned}
s_p &= \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \\
&= \sqrt{\frac{(10 - 1)(0.683)^2 + (10 - 1)(0.750)^2}{10 + 10 - 2}} \\
&= \sqrt{\frac{9.261}{18}} \\
&= 0.7173
\end{aligned}
$$

Example: Comparing Packing Machines

The alternative is left-tailed so the critical value is the value a such that $P(T < a) = 0.05$, with $10 + 10 - 2 = 18$ degrees of freedom. The critical value is -1.7341. The rejection region is $t^* < -1.7341$.

Our test statistic, -3.3978, is in our rejection region, therefore, we reject the null hypothesis. With a significance level of 5%, we reject the null hypothesis and conclude there is enough evidence to suggest that the new machine is faster than the old machine.

Example: Comparing Packing Machines

The test statistic is:

$$t^* = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{42.14 - 43.23}{0.7173\sqrt{\frac{1}{10} + \frac{1}{10}}} = -3.398$$

The alternative is left-tailed so the critical value is the value a such that $P(T < a) = 0.05$, with $10 + 10 - 2 = 18$ degrees of freedom. The critical value is -1.7341. The rejection region is $t^* < -1.7341$.

Our test statistic, -3.3978, is in our rejection region, therefore, we reject the null hypothesis. With a significance level of 5%, we reject the null hypothesis and conclude there is enough evidence to suggest that the new machine is faster than the old machine.

What if some of the assumptions are not satisfied:

Assumption 1. What should we do if the assumption of independent samples is violated?

If the samples are not independent but paired, we can use the paired t-test.

Assumption 2. What should we do if the sample sizes are not large and the populations are not normal?

We can use a nonparametric method to compare two samples such as the Mann-Whitney procedure.

Assumption 3. What should we do if the assumption of equal variances is violated?

We can use the separate variances 2-sample t-test.

Unpooled Variances

When the assumption of equal variances is not valid, we need to use separate, or unpooled, variances. The mathematics and theory are complicated for this case and we intentionally leave out the details.

We still have the following assumptions:

- The populations are independent
- Each population is either normal or the sample size is large.

If the assumptions are satisfied, then

$$t^* = \frac{\bar{x}_1 - \bar{x_2} - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)C^2 + (1 - C)^2(n_1 - 1)}$$

Resources

https://newonlinecourses.science.psu.edu/stat500/node/38/
https://machinelearningmastery.com/statistical-hypothesis-tests/

Let's practice: hypothesis testing with t test

- Calculating t tests using scipy

ttest_ind underestimates p for unequal variances:

```
>>> rvs3 = stats.norm.rvs(loc=5, scale=20, size=500)
>>> stats.ttest_ind(rvs1, rvs3)
(-0.46580283298287162, 0.64145827413436174)
>>> stats.ttest_ind(rvs1, rvs3, equal_var = False)
(-0.46580283298287162, 0.64149646246569292)
```

When n1 != n2, the equal variance t-statistic is no longer equal to the unequal variance t-statistic:

```
>>> rvs4 = stats.norm.rvs(loc=5, scale=20, size=100)
>>> stats.ttest_ind(rvs1, rvs4)
(-0.99882539442782481, 0.3182832709103896)
>>> stats.ttest_ind(rvs1, rvs4, equal_var = False)
(-0.69712570584654099, 0.48716927725402048)
```

T-test with different means, variance, and n:

```
>>> rvs5 = stats.norm.rvs(loc=8, scale=20, size=100)
>>> stats.ttest_ind(rvs1, rvs5)
(-1.4679669854490653, 0.14263895620529152)
>>> stats.ttest_ind(rvs1, rvs5, equal_var = False)
(-0.94365973617132992, 0.34744170334794122)
```