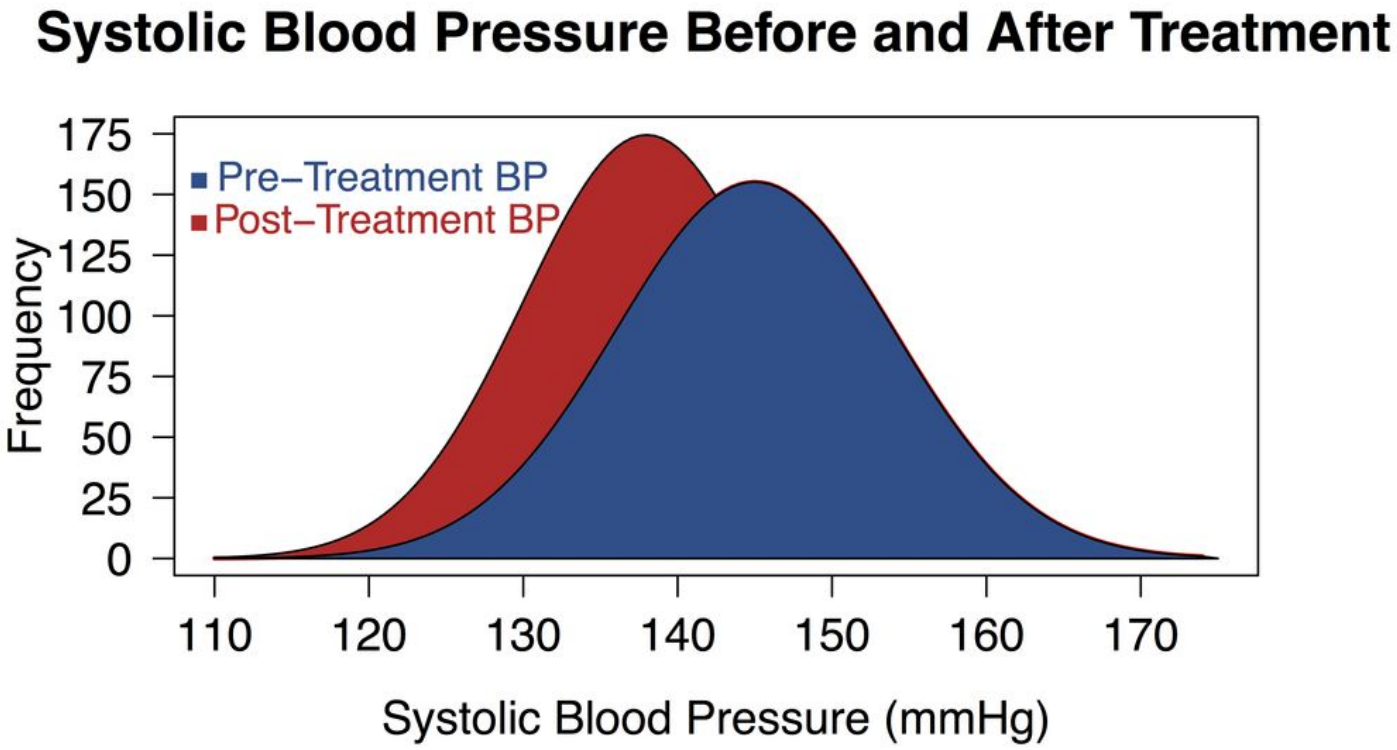
Comparing Two Populations

Data Science Immersive

FLATIRON SCHOOL

Problem

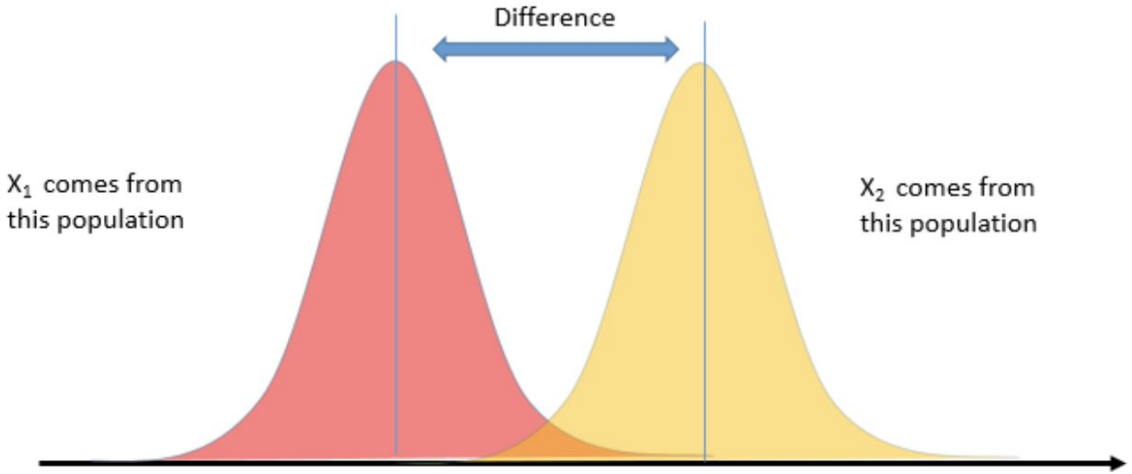

Systolic Blood Pressure Before and After Treatment
Pre−Treatment BP
Post−Treatment BP
Frequency
175
150
125
100
75
50
25
0
110
120
130
140
150
160
170
Systolic Blood Pressure (mmHg)

Bivariate Analysis

Bivariate data is when you are studying two variables.

- If the measurements are **categorical** (Smoker?) and taken from two distinct groups (e.g. Female, Male) the analysis will involve comparing two independent proportions.
- If the measurements are **quantitative** (e.g. GPA) and taken from two distinct groups (e.g. Graduate, Undergraduate) the analysis will involve comparing two independent means.
- If the measurements are quantitative (e.g. Weight) and taken twice from each subject (e.g. subject's weight before and after dieting) the analysis will involve comparing two dependent means.

Problem


Difference
X_1 comes from this population
X_2 comes from this population

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 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 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 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 95% 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.

Another way to consider this is how many measurements are taken off each subject. If only one measurement, then independent; if two measurements, then paired.

Comparing Two 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 view of this, there are two options for estimating the variances for the 2-sample t-test with independent samples:

1. 2-sample t-test using pooled variances
2. 2-sample t-test using separate 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.

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

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 (machine.txt), in seconds, are shown in the following table.

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?

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.

Pooled Variance

We can perform the separate variances test using the following test statistic:

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

This calculation for the exact degrees of freedom is cumbersome and is typically done by software. An alternate, conservative option to using the exact degrees of freedom calculation can be made by choosing the smaller of n-1 for the two samples

Comparing Two Population Variances

F-test to Compare Two Population Variances

To compare the variances of two quantitative variables, the hypotheses of interest are:

$$H_o : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

and then

$$H_a : \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \quad \text{or} \quad H_a : \frac{\sigma_1^2}{\sigma_2^2} < 1 \quad \text{or} \quad H_a : \frac{\sigma_1^2}{\sigma_2^2} > 1$$

Code Examples and Resources

https://stats.stackexchange.com/questions/210830/why-is-f-test-so-sensitive-for-the-assumption-of-normality

https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/

https://newonlinecourses.science.psu.edu/stat500/node/48/