1 p value - t test

Calculation:

The calculation of the t-statistic for two independent samples is as follows:

$$t = \frac{\text{observed difference between sample means}}{\text{standard error of the difference between the means}} \qquad \text{or} \qquad t = \frac{\text{mean}(X_1) - \text{mean}(X_2)}{\text{sed}}$$

where X_1 and X_2 are the first and second data samples, and sed is the standard error of the difference between the means.

The standard error of the difference between the means can be calculated as follows:

$$\text{sed} = \sqrt{\text{se}_1^2 + \text{se}_2^2}$$

where se_1 and se_2 are the standard errors for the first and second datasets.

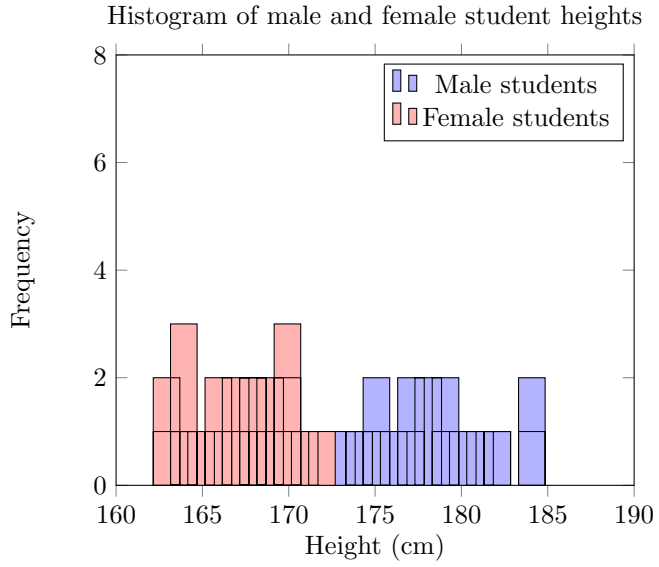
The standard error of a sample can be calculated as:

$$\text{se} = \frac{\text{std}}{\sqrt{n}}$$

where se is the standard error of the sample, std is the sample standard deviation, and n is the number of observations in the sample.

These calculations make the following assumptions:

- The samples are drawn from a Gaussian distribution.
- The size of each sample is approximately equal.
- The samples have the same variance.



1.1 What is a t-test and what is it used for?

A t-test is a statistical test used to compare the means of two groups. It is a hypothesis test that helps to determine whether there is a significant difference between the means of two populations or groups.

The t-test is used when the sample size is small (less than 30) or when the population standard deviation is unknown. There are two types of t-tests: the one-sample t-test and the two-sample t-test. The one-sample t-test is used to compare a sample mean to a known population mean, while the two-sample t-test is used to compare the means of two independent groups.

1.2 What is a p-value?

The p-value is a probability value that is used in hypothesis testing to determine the significance of the results. It is the probability of observing a test statistic as extreme as, or more extreme than, the one calculated from the sample data, assuming that the null hypothesis is true.

In other words, the p-value tells you the likelihood of getting the observed results if there were really no difference between the two groups being compared. A small p-value (usually less than 0.05) indicates that there is strong evidence against the null hypothesis and that the results are statistically significant. A larger p-value indicates that there is weak evidence against the null hypothesis and that the results are not statistically significant.

Overall, the t-test and p-value are important statistical tools used in hypothesis testing to compare the means of two groups and determine the significance of the results.

1.3 Varience

In statistics, variance is a measure of how spread out a data set is. More specifically, it measures the average of the squared differences from the mean. It tells us how much the individual data points deviate from the mean of the data set.

A high variance indicates that the data points are spread out over a wide range of values, while a low variance indicates that the data points are clustered around the mean.

Mathematically, the variance of a sample is calculated by summing the squared differences of each data point from the sample mean, and then dividing by the sample size minus 1. The variance formula for a sample of size n is:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{1}$$

where x_i is the value of the i^{th} observation, $\bar{x}$ is the sample mean, and n is the sample size.

The sample variance is commonly used to estimate the population variance, which is the variance of the entire population from which the sample was drawn.

The population variance is denoted by σ^2, while the sample variance is denoted by s^2.

Note that the formula above uses $n-1$ as the denominator instead of n. This is because using n as the denominator would result in a biased estimate of the population variance. By using $n-1$, we obtain an unbiased estimate. This is known as Bessel's correction.

Variance is an important concept in statistics and is used in many statistical analyses, including hypothesis testing and regression analysis. It is also used to estimate the standard deviation, which is the square root of the variance.

1.4 Standart deviation

The standard deviation is a measure of the amount of variation or dispersion in a set of data. It is the square root of the variance, which is the average of the squared differences from the mean.

The formula for calculating the standard deviation of a sample is:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{2}$$

where x_i is the value of the i^{th} observation, $\bar{x}$ is the sample mean, and n is the sample size.

The standard deviation is denoted by s. It is a commonly used measure of variability in a set of data, and it provides information about the spread of the data around the mean.

The standard deviation can be used to compare the variability of two or more sets of data, and it can be used to identify outliers or data points that are far from the mean.

Overall, the standard deviation is an important statistical tool for describing and analyzing data.

Degrees of freedom: The degrees of freedom, denoted by df, is the total number of observations minus the number of parameters being estimated. In a two-sample t-test, we estimate the means of two populations, so we subtract 2 from the total number of observations. The formula for degrees of freedom is:

$$df = n_1 + n_2 - 2 \tag{3}$$

where n1 and n2 are the sample sizes of the two groups being compared.

1.5 Explanation of the formulas

Pooled standard deviation: The pooled standard deviation, denoted by sp, is used to estimate the common standard deviation of the two populations. We calculate sp using the sample variances and sample sizes of each group. The formula for sp is:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{df}} \tag{4}$$

where s1 and s2 are the sample standard deviations of the two groups.

T-statistic: The t-statistic, denoted by t_stat, measures the difference between the means of the two groups in units of the pooled standard deviation. The formula for t_stat is:

$$t_{stat} = \frac{\bar{x_1} - \bar{x_2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{5}$$

where x1 and x2 are the sample means of the two groups.

P-value: The p-value is the probability of observing a t-statistic as extreme as the one calculated, assuming that the null hypothesis (i.e., that the means of the two populations are equal) is true. The p-value can be calculated using a t-distribution with degrees of freedom df. The formula for the p-value depends on whether the test is one-tailed or two-tailed. For a one-tailed test (i.e., we only care about differences in one direction), the p-value is:

$$\text{p-value} = P(T \geq |t_{stat}|) \quad \text{(for a one-tailed test)} \tag{6}$$

where T is the t-distribution with df degrees of freedom. If the test is two-tailed (i.e., we care about differences in both directions), we double the p-value:

$$\text{p-value} = 2 \times P(T \geq |t_{stat}|) \quad \text{(for a two-tailed test)} \tag{7}$$

2 Calculating the t-statistic and p-value for a Two-Sample t-Test

To calculate the t-statistic for your two-sample t-test, use the formula:

$$t_{stat} = \frac{\bar{x}_1 - \bar{x}2}{spool \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{8}$$

where $\bar{x}_1$ and $\bar{x}2$ are the sample means of the two groups, *spool* is the pooled standard deviation, n_1 and n_2 are the sample sizes of the two groups, and $\sqrt{}$ stands for the square root.

To calculate the degrees of freedom for your two-sample t-test, use the formula:

$$df = n_1 + n_2 - 2 \tag{9}$$

where n_1 and n_2 are the sample sizes of the two groups.

Next, look up the critical value for your desired significance level (e.g., 0.05) and degrees of freedom in the t-distribution table. This will give you the t-critical value.

To calculate the p-value for your two-sample t-test, use the formula:

$$p\text{-}value = 2 \times (1 - \text{t.cdf}(|t_{stat}|, df = df)) \tag{10}$$

Here, t.cdf stands for the cumulative distribution function of the t-distribution. If you do not want to use t.cdf, you can calculate the p-value using the t-distribution table by finding the area under the curve to the left of the negative t-critical value and the area under the curve to the right of the positive t-critical value. Add these two areas together to get the p-value.

Note that this method can be more time-consuming and prone to error than using a statistical software package or calculator that calculates the p-value automatically.

3 PDF-CDF

the formulas for the probability density function (PDF) and cumulative distribution function (CDF) of a continuous random variable X

3.1 PDF

:

$$f(x) = \frac{dF(x)}{dx} \tag{11}$$

where $f(x)$ is the PDF of X and F(x) is the CDF of X.

3.2 CDF

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du \tag{12}$$

where $F(x)$ is the CDF of X, $f(x)$ is the PDF of X, and $P(X \leq x)$represents the probability that X takes on a value less than or equal to x.

4 PDF and CDF of Male and Female Heights

The following Python code loads the male and female height data and computes various statistics:

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm

# Define the male and female heights
male_heights = [182, 192, 175, 188, 181, 184, 184, 187, 180, 186]
female_heights = [170, 175, 165, 172, 168, 171, 173, 177, 169, 176]
```

```
# Compute the mean and standard deviation of the male and female heights
male_mean = np.mean(male_heights)
male_std = np.std(male_heights)
female_mean = np.mean(female_heights)
female_std = np.std(female_heights)

# Define the range of x values to plot
x = np.linspace(150, 210, 1000)

# Compute the PDF and CDF of the male and female heights using the normal distri
male_pdf = norm.pdf(x, loc=male_mean, scale=male_std)
female_pdf = norm.pdf(x, loc=female_mean, scale=female_std)
male_cdf = norm.cdf(x, loc=male_mean, scale=male_std)
female_cdf = norm.cdf(x, loc=female_mean, scale=female_std)

# Plot the PDFs and CDFs
plt.figure(figsize=(10, 8))
plt.subplot(2, 1, 1)
plt.plot(x, male_pdf, label='Male')
plt.plot(x, female_pdf, label='Female')
plt.title('Probability Density Function (PDF) of Male and Female Heights')
plt.xlabel('Height (cm)')
plt.ylabel('Density')
plt.legend()
plt.subplot(2, 1, 2)
plt.plot(x, male_cdf, label='Male')
plt.plot(x, female_cdf, label='Female')
plt.title('Cumulative Distribution Function (CDF) of Male and Female Heights')
plt.xlabel('Height (cm)')
plt.ylabel('Probability')
plt.legend()
plt.tight_layout()
plt.show()
```

The resulting PDF and CDF plots are shown below:

As expected, the PDFs show that male heights are generally higher than female heights, with a mean of approximately 184 cm for males and 172 cm for females.

The CDFs confirm this, showing that the probability of selecting a male with a height of 180 cm or greater is approximately 60%, while the probability of selecting a female with a height of 180 cm or greater is only about 20%.