**Department of Computer Engineering**

**Academic Year:** 2022-2023 **Semester:** VIII

**Subject:** Applied Data Science **Class / Division:** BE/CMPN/B

**Name:** Domingos J Pinto **Roll Number:** 007

**Date: Seat Number:** BE/CMPN/B

**Experiment No.: 2**

**Explore the inferential statistics on the given dataset.**

**Aim :** Explore the inferential statistics on the given dataset.

**I-THEORY**

Inferential statistics generally involves generating deductions and/or predictions about a population. In several cases, inferences are made about a population using a sample. Unlike descriptive statistics where a known sample/population data is described, inferential statistics uses sample data to make conclusions about the population.

# Hypothesis Testing

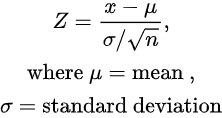
Hypothesis testing is a statistical inference technique used to confirm or refute statements made about a population using the sample data provided. We can think of hypothesis testing as an experiment, an hypothesis is made before the experiment starts. After experimentation, we would confirm if the results agree with the statement or not.

Hypothesis testing is one of the most significant aspects of inferential statistics. There are several tests applied in hypothesis testing and the specific test to use depends on the data and purpose of the test. There are several hypothesis tests you would need to be familiar with in your journey as a data analyst. This article covers the following tests:

* Z Test & T-Test
* Correlation Test
* Chi Square Tests

## Z-Test & T-Test

The Z-Test is a hypothesis test typically used to determine if the means of two populations are significantly different or if the mean of a population is greater than, less than or equivalent to a specific value. This test is used when the variance(s) of the population(s) is/are known. It is also applied when the data follows a [normal distribution](https://www.investopedia.com/terms/n/normaldistribution.asp). When the sample size is large, it is also assumed that the data follows a normal distribution.



Using a case study of the performance of students in 2 classes, the Z Test can be used to ascertain if there is a significant difference in score. In this scenario, the null hypothesis is that the mean scores from the two classes are equal. The hypothesis test would enable us to support or refute this claim. Usually, for hypothesis tests, a 5% **[level of significance](https://statisticsbyjim.com/glossary/significance-level)** is applied and the claim is rejected if the **[p-value](https://www.scribbr.com/statistics/p-value/)** produced is less than the level of significance.

The T-Test has a similar purpose as the Z-Test. However, it is applied when the population standard deviation is not known, or for samples with small sample sizes (n < 30).

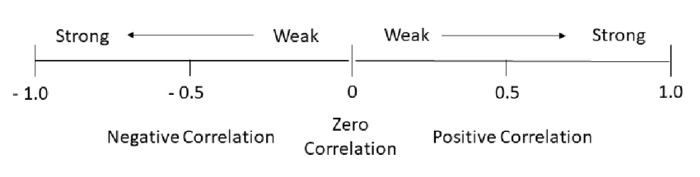
https://miro.medium.com/max/110/1*Kv9whbX5rYIU2ilTg8C0dQ.png

Let us paint another scenario of a coach who trains junior athletes to run a 100meters race. The coach believes that the average speed of her student is 10 seconds. To confirm this, she selects 10 athletes.

## Correlation Test

Correlation describes the degree of relationship between two (or more) variables. For example, there might be a positive relationship between hours of practice and overall performance: “*The more you practice, the better your results in an examination will be*”.

The correlation test tests if the relationship between these variables is statistically significant. The Pearson Correlation Coefficient is a popular correlation coefficient that measures the linear relationship between 2 variables.



For instance, the relationship between test scores and exam scores can be tested using Pearson correlation. The *pearson r*function on Scipy returns the correlation coefficient and tests if the correlation is significant. The null hypothesis for correlation test is that there is no correlation between the variables.

## Chi-Square Tests

There are 3 types of Chi-Square Tests:

* Chi-Square Test of Independence
* Chi-Square Goodness of Fit Test
* Chi-Square Test of Homogeneity

The most popularly of these tests are the Chi Square Test of Independence and the Goodness of Fit Test.

The Chi-Square Goodness of Fit test is used, mostly, to ascertain if the sample data is a true representation of the population. On the other hand, the Chi-Square test of independence is used to determine if the relationship between two categorical variables is significant. It is different from the Correlation test because, unlike the correlation test that focuses on quantitative variables, this chi-square test deals with categorical variables.

The *scipy*’s [stats.chisquare](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html) function is used to compute the goodness of fit test while the [chi2\_contigency](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html) function is used to compute the chi-square test of independence.

Inferential Statistics is an extremely valuable tool for every potential data analyst. From applying sampling techniques in your data collection process to applying hypothesis tests to deduce from your data, it is too valuable to dismiss.

**II-IMPLEMENTATION - Explore the inferential statistics on the dataset.**

**Dataset :** Crop Recommendation

**Description :** It is a dataset that contains which crop can grow based on the parameters mentioned in it. The dataset consist of 7 different parameters and one label section that contains the name of the crop. There are 2200 rows in this dataset.

import numpy as np

import pandas as pd

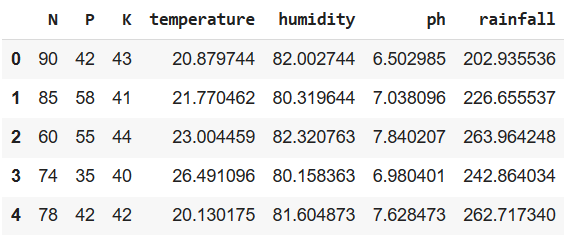
import scipy.stats as stats

from statsmodels.stats.weightstats import ztest

crops = pd.read\_csv("Crop\_recommendation.csv")

del crops["label"]

crops.head()



ztest(crops["N"],crops["P"],value=0)



stats.ttest\_1samp(crops["N"], popmean=10)



def f\_test(group1, group2):

f = np.var(group1, ddof=1)/np.var(group2, ddof=1)

nun = group1.size-1

dun = group2.size-1

p\_value = 1-stats.f.cdf(f, nun, dun)

return f, p\_value

# perform F-test

x = crops["temperature"][:1100]

y = crops["temperature"][1100:]

f\_test(x, y)

