Module 2

Data Analytics with Python – Statistics

Section: Advance Statistics

Introduction to Hypothesis testing

There are two main methods used in inferential statistics:

- Estimation
- Hypothesis testing

A Hypothesis is a statement or an assumption about relationships between variables. Alternatively, a hypothesis is a tentative explanation for certain behaviors, phenomenon or events that have occurred or will occur.

Hypothesis Statement

If you are going to propose a hypothesis, it's customary to write a statement. Your statement will look like this:

"If I…(do this to an independent variable)….then (this will happen to the dependent variable)."

For example:

- If I (decrease the amount of water given to herbs) then (the herbs will increase in size).
- If I (give patients counseling in addition to medication) then (their overall depression scale will decrease).
- If I (give exams at noon instead of 7) then (student test scores will improve).
- If I (look in this certain location) then (I am more likely to find new species).

Criteria for Hypothesis Construction

A hypothesis should meet the following criteria:

- It should be empirically testable, whether it is right or wrong.
- It should be specific and precise.
- The statements in the hypothesis should not be contradictory.
- It should specify variables between which the relationship is to be established.
- It should describe one issue only.

Null Hypothesis

If you trace back the history of science, the null hypothesis is always the accepted fact.

Simple examples of null hypotheses that are generally accepted as being true are:

- DNA is shaped like a double helix.
- There are 8 planets in the solar system (excluding Pluto).

Steps In Hypothesis Testing

The following are the steps for constructing a hypothesis test.

- Select the type of hypothesis
    a) Null Hypothesis (H0)
    b) Alternative Hypothesis (Ha or H1)
- Establish Critical or Rejection region
- Select the Suitable Test of significance or Test Statistic
- Check whether the test involves one sample, two samples, or multiple samples?
- Check whether two or more samples used are independent or related?
- Is the measurement scale nominal, ordinal, interval, or ratio?
- The choice of a probability distribution of a sample statistic is guided but the sample size n and the value of population standard deviation.
- Formulate a Decision Rule to Accept Null Hypothesis.
    a) Accept H0 if the test statistic value falls within the area of acceptance.
- Reject otherwise.

Errors in Hypothesis Testing

There are two types of error

- Type I Error
- Type II Error

type I Error

- It is also known as an error of the first kind, occurs when the null hypothesis (H0) is true, but is rejected.
- A type I error may be compared with a so called false positive.
- A Type I error occurs when we believe a falsehood.
- The rate of the type I error is called the size of the test and denoted by the Greek letter α (alpha).It usually equals the significance level of a test. If type I error is fixed at 5%, it means that there are about 5 chances in 100 that we will reject H0 when H0 is true.

Type II Error

- It is also known as an error of the second kind, occurs when the null hypothesis is false, but erroneously fails to be rejected.
- Type II error means accepting the hypothesis which should have been rejected. A type II error may be compared with a so-called False Negative.
- A Type II error is committed when we fail to believe a truth.
- A type II error occurs when one rejects the alternative hypothesis (fails to reject the null hypothesis) when the alternative hypothesis is true.
- The rate of the type II error is denoted by the Greek letter β (beta) and related to the power of a test (which equals 1-β).

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error False positive	Correct outcome True positive
Fail to reject null hypothesis	Correct outcome True negative	Type II error False negative

p-Value

- When you perform a hypothesis test in statistics, a p-value helps you determine the significance of your results. Hypothesis tests are used to test the validity of a claim that is made about a population. This claim that's on trial, in essence, is called the null hypothesis.
- The alternative hypothesis is the one you would believe if the null hypothesis is concluded to be untrue. The evidence in the trial is your data and the statistics that go along with it.
- All hypothesis tests ultimately use a p-value to weigh the strength of the evidence (what the data are telling you about the population)

p-Value Interpretation

The p-value is a number between 0 and 1 and interpreted in the following way:

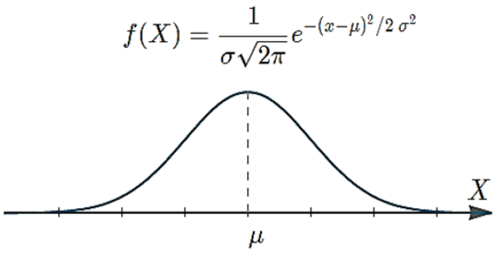
- A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- The p-values very close to the cutoff (0.05) are considered to be marginal (could go either way).
- Always report the p-value so your readers can draw their own conclusions.

Normal Distribution

- The Normal Probability Distribution is very common in the field of statistics.
- The random variable X is said to be normally distributed with mean μ and standard deviation σ if its probability distribution is given by

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\,\sigma^2}$$

(Where −∞ < μ < ∞ and 0 < σ2 < 1 are arbitrary parameters. If X has a normal distribution with parameters μ and σ2, then we write X ~ N (μ, σ2).)


$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\,\sigma^2}$
X
μ

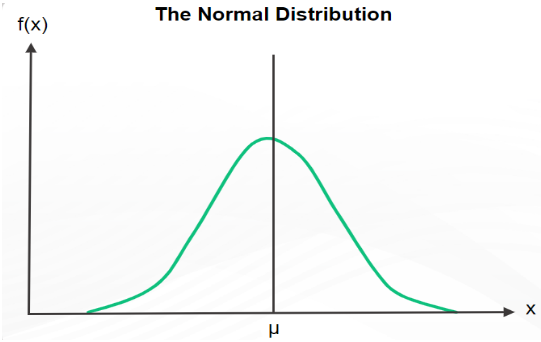
A normal curve

Properties of Normal Distribution

The properties of normal distribution are as follows:

- The normal curve is symmetrical about the mean μ.
- The mean is at the middle and divides the area into halves.
- The total area under the curve is equal to 1.
- It is completely determined by its mean and standard deviation σ (or variance σ^2)

In a normal distribution, only 2 parameters are needed, namely μ and σ^2.



Z-test

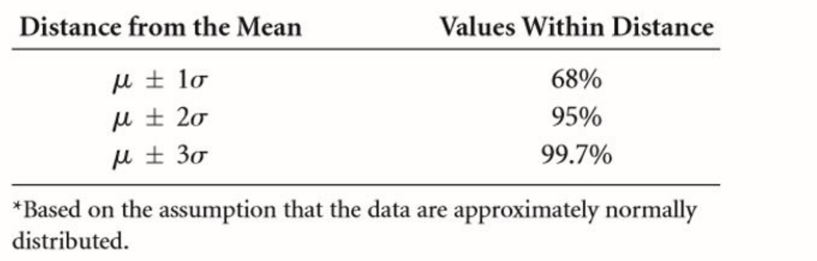
Z-test is a statistical method to determine whether the distribution of the test statistics can be approximated by a normal distribution. It is the method to determine whether two sample means are approximately the same or different when their variance is known and the sample size is large (should be >= 30).

Z score - Area Under the Normal Curve

- The probability that a variable is within range in a normal distribution is calculated by finding the area under the normal curve.
- The area depends on the values of μ (mean) and σ (standard deviation).
- The z-score table is used to find the area under the normal curve.
- Z-score is the standardized value of observation x from a distribution that has mean μ and standard deviation σ.
- In a z-score table, the left most column means the number of standard deviations above the mean to 1 decimal place, the top row gives the second decimal place, and the intersection of a row and column gives the probability.

Empirical Rule

The empirical rule is an important rule of thumb that is used to state the approximate percentage of values that lie within a given number of standard deviations from the mean of a set of data if the data are normally distributed.



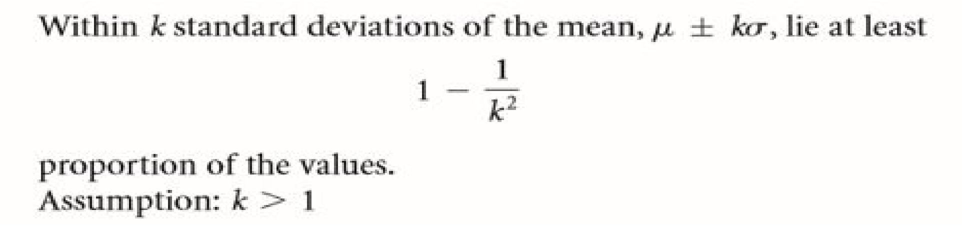
Distance from the Mean	Values Within Distance
$\mu \pm 1\sigma$	68%
$\mu \pm 2\sigma$	95%
$\mu \pm 3\sigma$	99.7%

*Based on the assumption that the data are approximately normally distributed.

The empirical rule applies only when data are known to be approximately normally distributed. when data are not normally distributed or when the shape of the distribution is unknown, we have to use Chebyshev’s theorem

Chebyshev’s Theorem

It applies to all distributions regardless of their shape and thus can be used whenever the data distribution shape is unknown or is non normal.



Within k standard deviations of the mean, $\mu \pm k\sigma$, lie at least

$$1 - \frac{1}{k^2}$$

proportion of the values.
Assumption: $k > 1$

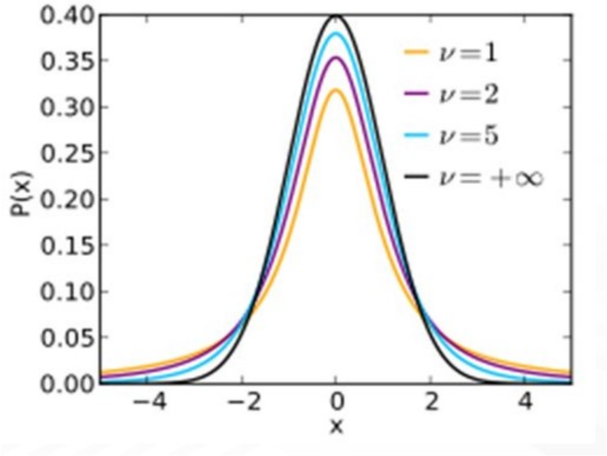
Chebyshev’s Theorem compared to The Empirical Rule

- The Empirical Rule also describes the proportion of data that fall within a specified number of standard deviations from the mean. However, there are several differences between Chebyshev’s Theorem and the Empirical Rule.
- Chebyshev’s Theorem applies to all probability distributions where you can calculate the mean and standard deviation. On the other hand, the Empirical Rule applies only to the normal distribution.
- Chebyshev’s Theorem provides approximations. Conversely, the Empirical Rule provides exact answers for the proportions because the data are known to follow the normal distribution.
- If data follow the normal distribution, use the Empirical Rule. Otherwise, Chebyshev’s Theorem is best choice.

t-Test

- A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features.
- The t-test is one of many tests used for the purpose of hypothesis testing in statistics.
- Calculating a t-test requires three key data values. They include the difference between the mean values from each data set, the standard deviation of each group, and the number of data values of each group.
- There are several different types of t-test that can be performed depending on the data and type of analysis required.
- It is type of test statistic in which hypothesis tests use the test statistic that is calculated from your sample to compare your sample to the null hypothesis.

t-Distribution Curve



The t-distribution (also called Student's t-distribution) is a family of distributions that look almost identical to the normal distribution curve, only a bit shorter and fatter.
The t-distribution is used instead of the normal distribution when you have small samples.

t-Distribution Formula

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

Where

$\bar{x}$ is the mean of the first sample;

μ is the mean of the second sample;

$\frac{s}{\sqrt{N}}$ is the estimate of the standard error of the difference between the means.

Assumptions of t-Distribution

- The sample is drawn from the Normal population
- The sample observations are independent
- The population standard deviation σ is unknown

Paired T-Test

The Paired t-test is performed when the samples typically consist of matched pairs of similar units, or when there are cases of repeated measures. For example, there may be instances of the same patients being tested repeatedly—before and after receiving a particular treatment. In such cases, each patient is being used as a control sample against themselves.
Correlated or paired t-tests are of a dependent type, as these involve cases where the two sets of samples are related.

The formula for computing the t-value and degrees of freedom for a paired t-test is:

$$T = \frac{mean1 - mean2}{\frac{s(\mathrm{diff})}{\sqrt{(n)}}}$$

where:

mean1 and mean2=The average values of each of the sample sets
s(diff)=The standard deviation of the differences of the paired data values
n=The sample size (the number of paired differences)
n−1=The degrees of freedom

Degrees of freedom refers to the values in a study that has the freedom to vary and are essential for assessing the importance and the validity of the null hypothesis. Computation of these values usually depends upon the number of data records available in the sample set.

Other Type of t- Test is called independent t- Test which is further divided as below

- Equal Variance (or Pooled) T-Test
- Unequal Variance T-Test

Equal Variance (or Pooled) T-Test

The equal variance t-test is used when the number of samples in each group is the same, or the variance of the two data sets is similar. The following formula is used for calculating t-value and degrees of freedom for equal variance t-test:

$$\text{T-value} = \frac{mean1 - mean2}{\frac{(n1-1)\times var1^2+(n2-1)\times var2^2}{n1+n2-2} \times \sqrt{\frac{1}{n1} + \frac{1}{n2}}}$$

where:
mean1 and mean2=Average values of each of the sample sets
var1 and var2=Variance of each of the sample sets
n1 and n2=Number of records in each sample set

Unequal Variance T-Test

The unequal variance t-test is used when the number of samples in each group is different, and the variance of the two data sets is also different. This test is also called the Welch's t-test. The following formula is used for calculating t-value and degrees of freedom for an unequal variance t-test:

$$\text{T-value} = \frac{mean1 - mean2}{\sqrt{\left(\frac{var1}{n1} + \frac{var2}{n2}\right)}}$$

where:
mean1 and mean2=Average values of each of the sample sets
var1 and var2=Variance of each of the sample sets
n1 and n2=Number of records in each sample set

ANOVA

- **ANOVA** refers to **analysis of variance** and is a statistical procedure used to test the degree to which two or more groups vary or differ in an experiment.
- Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.
- ANOVA checks the impact of one or more factors by comparing the means of different samples.

ANOVA is measured using a statistic known as F-Ratio. It is defined as the ratio of Mean Square (between groups) to the Mean Square (within group).

Mean Square (between groups) = Sum of Squares (between groups) / degree of freedom (between groups)
Mean Square (within group) = Sum of Squares (within group) / degree of freedom (within group)

Between groups: If there are k groups in ANOVA model, then k-1 will be independent. Hence, k-1 degree of freedom.
Within groups: If N represents the total observations in ANOVA (∑n over all groups) and k are the number of groups then, there will be k fixed points. Hence, N-k degree of freedom.

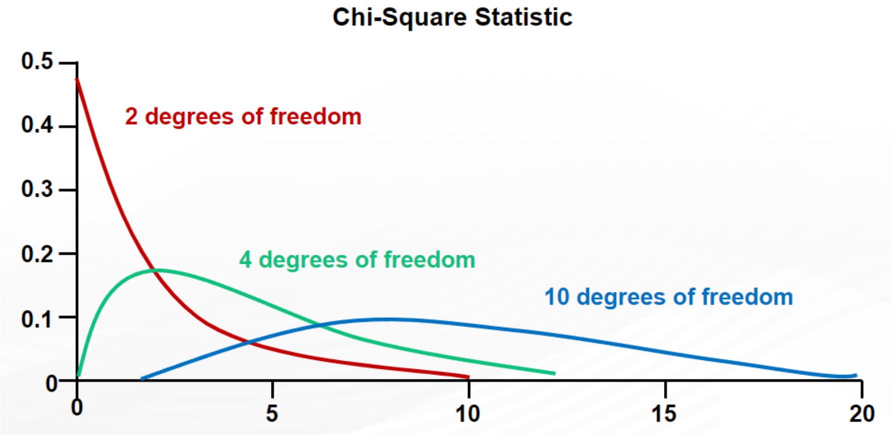
Chi-square Distribution

A chi-square statistic is one way to show a relationship between two categorical variables. The distribution of the chi-square statistic is called the chi-square distribution. The **chi-square distribution** is defined by the following probability density function:

$$Y = Y_0 * (X^2)^{(v/2 - 1)} * e^{-X2 / 2}$$

- Y_0 is a constant that depends on the number of degrees of freedom,
- X^2 is the chi-square statistic,
- *v* = *n* - 1 is the number of degrees of freedom,
- *e* is a constant equal to the base of the natural logarithm system (approximately 2.71828).
- Y_0 is defined, so that the area under the chi-square curve is equal to one.

The graph illustrates chi-square distribution of different sample sizes.



Properties of Chi-square Distribution

The chi-square distribution has the following properties:

- The mean of the distribution is equal to the number of degrees of freedom: μ = v.
- The variance is equal to two times the number of degrees of freedom: $\sigma^2 = 2*v$.
- When the degree of freedom is greater than or equal to 2, the maximum value for Y occurs when $X^2 = v - 2$.
- As the degrees of freedom increase, the chi-square curve approaches a normal distribution.

Example:

The Acme Battery Company has developed a new cell phone battery. On average, the battery lasts 60 minutes on a single charge. The standard deviation is 4 minutes.

Suppose, the manufacturing department runs a quality control test. They randomly select 7 batteries. The standard deviation of the selected batteries is 6 minutes. What would be the chi-square statistic represented by this test?

Solution:

The standard deviation of the population is 4 minutes.
The standard deviation of the sample is 6 minutes.
The number of sample observations is 7.
To compute the chi-square statistic, we plug these data in the chi-square equation, as shown below.

X2 = [(n - 1) * s2] / σ2
X2 = [(7 - 1) * 62] / 42 = **13.5**
(Where X2 is the chi-square statistic, n is the sample size, s is the standard deviation of the sample, and σ is the standard deviation of the population.)

Summary

- A Hypothesis is a statement or an assumption about relationships between variables.
- When you perform a hypothesis test in statistics, a p-value helps you determine the significance of your results
- A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- The normal curve is symmetrical about the mean μ.
- Z-test is a statistical method to determine whether the distribution of the test statistics can be approximated by a normal distribution.
- The empirical rule is an important rule of thumb that is used to state the approximate percentage of values that lie within a given number of standard deviations from the mean of a set of data if the data are normally distributed
- A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features.
- ANOVA refers to analysis of variance and is a statistical procedure used to test the degree to which two or more groups vary or differ in an experiment.
- A chi-square statistic is one way to show a relationship between two categorical variables.