**Inferential Statistics**

**Hypothesis Testing:**

Hypothesis testing means analysis of sample data using some techniques like z-test, t-test, f-test, chi-square test and then apply this analysis on population data.

**Steps:**

1. Define null hypothesis.
2. Define alternate hypothesis
3. Statistical Experiment on data (Experiment means z-test, t-test, f-test)
4. Accept or reject the null hypothesis.

Let’s take one general example.

A person is arrest in charge of murder. Here judge have to make decision whether person is innocent or guilty. Here,

Steps:

Define null hypothesis. – Person is innocent

Define alternate hypothesis -- Person is guilty

Statistical Experiment on data -- Police will analyze the data like (DNA, Finger prints)

Accept or reject the null hypothesis. – After doing analysis judge will make decision whether person is innocent (Accept the null hypothesis) or guilty (reject the null hypothesis).

**P Value, Significance value, CI (Confidence Interval):**

P value is shows frequency of the random variable in dataset. The P-value is always decide the domain expert.

If significance value = 0.05 then

CI % = 1-0.05 = 95%

95%

-0.025 0.025

Mean

If p value fall under 95% then accept the null hypothesis. If p value fall outside the 95% then reject the null hypothesis.

e.g. – if p value is 0.07 then accept the null hypothesis.

If p value is 0.01 then reject the null hypothesis

**Different statistical tests which basically used in hypothesis testing:**

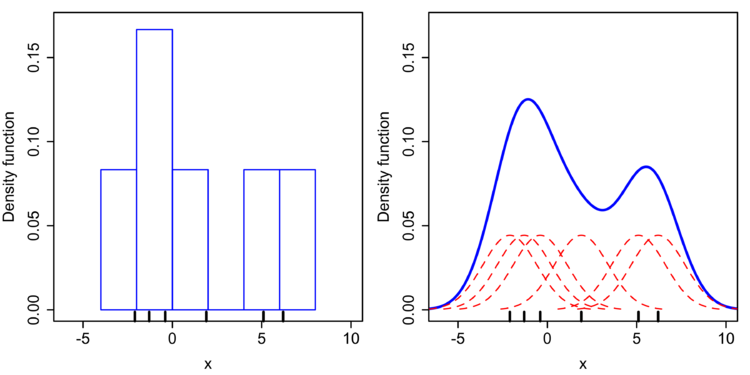
1. Z-test (Comparison of mean)
2. T-test (Comparison of mean)
3. F-test (Comparison of Variance)
4. Anova (Analysis of variance)
5. Chi Square (Comparison between two categorical variable)

**Central limit Theorem:**

Whenever random variable may or may not follow Gaussian distribution, after applying central limit theorem data will follow Gaussian distribution if you knows the population mean and S.D.

**X != G.D (u,S.D)**

**PDF:** It is smoothen version of Histogram. KDE helps to construct the pdf of histogram. If user knows mean and Standard deviation of data then user can construct it. In the below diagram red curves which is formed by data elements is nothing but the kernel.



**CDF:** In simple words CDF means adding data ponits at every point.

E.g.

Score CDF

100 100

80 180

50 230

**Bernoulli distribution:**

Whenever values are continuous then use **pdf** if values are discrete then use **pmf**.

Whenever any event has only two outcomes then that is Bernoulli distribution. E.g. – pass/fail, out/not out, True/False, 0/1

**P (k) = p^k (1-p)^1-k**

**Q = 1-p … if k=0**

**P …if k=1**

**p^k (1-p)^1-k**

This is called as Probability Mass Function.

If p(x = head) = 0.6 then p(x=tail) = 0.4 …. Head/tail this is called as Bernoulli distribution.

**Binomial distribution:**

An experiment is a binomial if it is repeated for fixed no. of time, the trails are independent, if trails have 2 mutually exclusive outcomes, either success or failure and if probability of success is same for all trails (not biased).

**Chebyshev’s Inequality:**

Gaussian distribution: empirical formula

X == G.D.

Probability (mean – S.D <= X <= mean+ S.D. = 68%)

Probability (mean – 2 S.D <= X <= mean+ 2 S.D. = 68%)

Probability (mean –3 S.D <= X <= mean+ 3 S.D. = 68%)

But if X != G.D. and if it has mean and S.D. then

mean – k \* S.D <= X <= mean+ k \* S.D. >= (1 – 1/k^2)

now k=2 **Note – value of k must have >= 2**

Probability (mean – 2 S.D. <= X <= mean +2 S.D.) > ¾

It mean if X != G.D. then probability of any value of X should be fall under 2 S.D. is 75%.

**IMP – If X is follows Gaussian distribution then it is use empirical formula. But if not then it follows chebyshev’s inequality.**

**Power Low Distribution:**

**y**

**X**

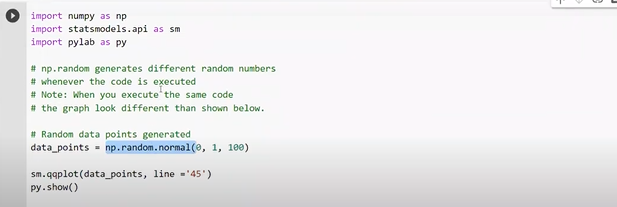
Power low distribution is follow 80-20 rule. 80% of incident happened due to 20% causes.

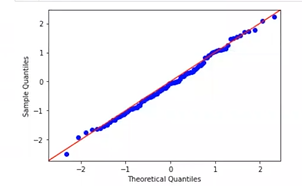
E.g. 80% runs of the Rcb is hit by 20% of batsman. This type of data is called as **pareto distribution**

E.g. 2- 80% of the project completed by 20% of people and 20% of project work completed by 80% of people.

To convert pareto distribution data into log normal distribution we use Boxcox transformation.

**Q-Q plot –** Q-Q plot is used to verify whether distribution follows Gaussian distribution or pareto distribution. (Quantile - Quantile Plot)





For Practical info please check stats.ipynb file.