Statistics Assignment 6

1. What is a Cumulative Distribution Function, and how does it work?
Answer:

The Cumulative Distribution Function (cdf) calculates the cumulative probability for a given x-value. Cumulative probability describes the probability that a random variable X with a given probability distribution will be found at a value less than or equal to x. i.e. it gives the area under the probability density function from minus infinite to x.

For a random variable y, cdf is generated by the summation of the pdf of all the values less than or equal to y.

2. When should we use a t-test vs a z-test?
Answer:

To compare Sample Mean with Population Mean, we use either t-test or z-test. Whenever Standard Deviation of Population is given and Sample Size is greater than or equal to 30 i.e. $n \geq 30$, we use Z-test. Otherwise we can use T-test.

3. How do we examine two category characteristics?
Answer:

To examine two category characteristics, we will use Chi-Square Test.
Chi-Square test is a non-parametric test which is performed on categorical data. It is usually a comparison between two categorical variables.

<u>Chi-Square Testing:-</u>

i. Define Null Hypothesis (H0)
ii. Define Alternate Hypothesis (H1)
iii. Calculate α value.
iv. Calculate Degree of Freedom
v. State Decision Boundary.
vi. State Chi Square Test i.e. $X^2 = \sum \frac{(f_o - f_e)}{f_e}$ where, f_o = observed value, f_e = expected value.
vii. State Decision – Whether to accept H0 or not

4. Explain the concept of Chebyshev's Inequality?
Answer:

For a random variable X belongs to Gaussian Distribution, we can use empirical formula i.e. 68-95-99.7 rule.
But for a random variable Y which does not belong to Gaussian Distribution, we will use Chebyshev's Inequality.

Chebyshev's Inequality = $\Pr(\mu - k\sigma \ \leq Y \leq \ \mu + k\sigma) \geq 1 - \frac{1}{k^2}$, where k > 1
It means the probability of an observation which is k standard deviation away from mean is greater than or equal to $1 - \frac{1}{k^2}$.

According to this inequality:
At k =2, $\Pr(\mu - k\sigma \ \leq Y \leq \ \mu + k\sigma) \geq 75\%$
At k=3, $\Pr(\mu - k\sigma \ \leq Y \leq \ \mu + k\sigma) \geq 88.9\%$
Thus, for random variable Y, within two standard deviation away from the mean contains 75% of the values, and within three standard deviation away from the mean contains 88.9% of the values.

5. Explain the concept of Pareto Distribution?
Answer:
Pareto Distribution is based on Power Law probability distribution. Power Law states that (in graphical representation) 80% of y-axis is satisfy by 20% of x-axis and 80% of x-axis is satisfy by 20% of y-axis. Basically it is 80-20 rule.
For example, 80% of a team project is done by 20% of team member, 80% of matches are won by 20% of team players.

We can transform the Pareto Distribution into Gaussian Distribution using BoxCox Transformation. And we do this transformation for better performance of ML algorithms.
Using Q-Q Plot, we can ensure that the transformed distribution is Gaussian or not.