Name: Chandrakant B Thakur

Statistics Assignment 3

1. Write the Gaussian Distribution empirical formula.

68-95-99.7 % Rule or Empirical Rule:
Gaussian distribution is symmetric distribution. This means if we draw the Probability Density Function (Pdf) of normal distribution then the Pdf of both sides of the mean value will be the mirror image of each other. The Pdf of the Gaussian distribution is a bell-shaped curve that is symmetric.

according to the Empirical rule, if a random variable follows Gaussian distribution then it has also three properties, and these properties are also called the Empirical formula or 68-95-99.8 % formula, and the three properties of the Empirical formula are as follows:

1. $P [\mu - \sigma <= X <= \mu + \sigma] \approx 68\ \%$

So the first formula basically says that the probability of a variable that falls within the range of $\mu - \sigma$ and $\mu + \sigma$ is 68 %. which means 68 % of the data points belonging to the random variable X fall within the range of the first standard deviation.

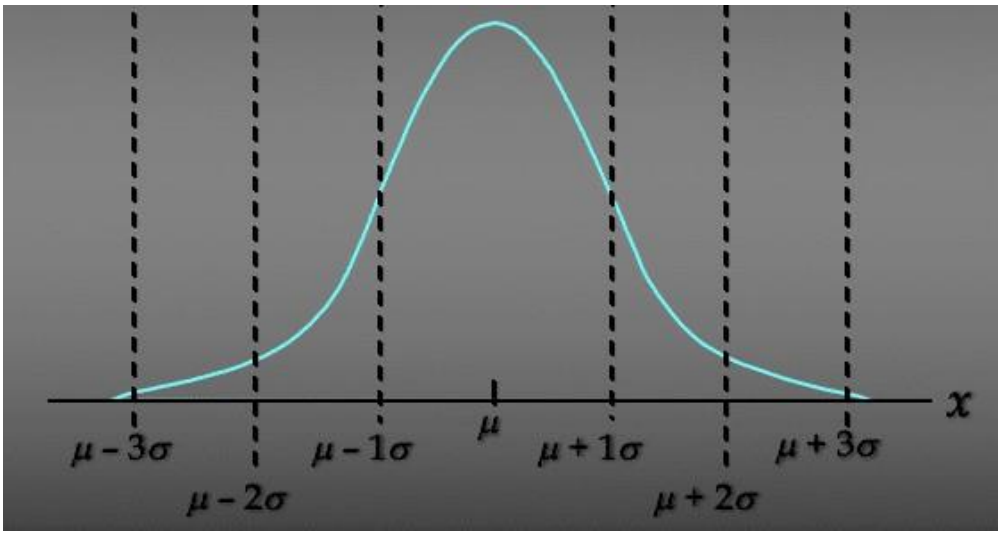
2. $P [\mu - 2\sigma <= X <= \mu + 2\sigma] \approx 95\ \%$

So the second formula basically says that the probability of a variable that falls within the range of $\mu - 2\sigma$ and $\mu + 2\sigma$ is 95 %. which means 95% of the data points belonging to the random variable X fall within the range of the second standard deviation.

3. $P [\mu - 3\sigma <= X <= \mu + 3\sigma] \approx 99.8\ \%$

So the second formula basically says that the probability of a variable that falls within the range of $\mu - 3\sigma$ and $\mu + 3\sigma$ is 99.8 %. which means 99.8% of the data points belonging to the random variable X fall within the range of the third standard deviation.

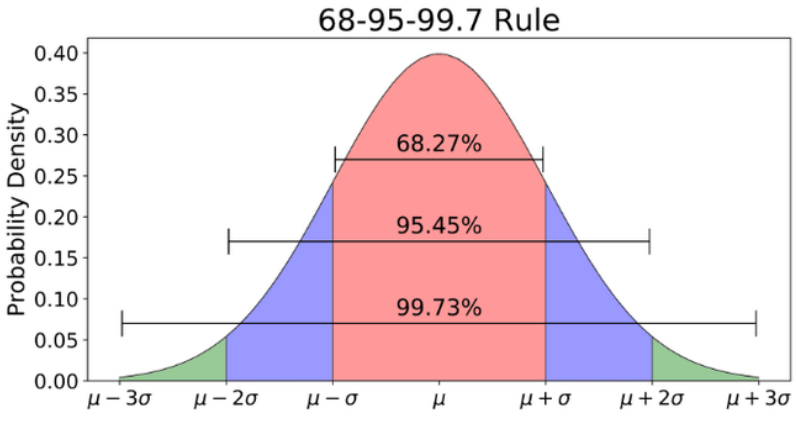
Name: Chandrakant B Thakur



2. What is the Z-score, and why is it important?

zscore is also known as standard score: gives us an idea of how far a data point is from the mean. It indicates how many standard deviations an element is from the mean. Hence, Z-Score is measured in terms of standard deviation from the mean. For example, a standard deviation of 2 indicates the value is 2 standard deviations away from the mean. In order to use a z-score, we need to know the population mean (μ) and also the population standard deviation (σ).
The Formula for Z-Score
A z-score can be calculated using the following formula.

$$z = (X - \mu) / \sigma$$



The 68-95-99.7 Rule for a Normal Distribution

3. What is an outlier, exactly?

A definition of outliers in statistics can be considered a section of data used to represent an extraordinary range from one point to another point. Or we can say that it is the data that remains outside of the other given values with a set of data. If one had Pinocchio within a class of teenagers, his nose's length would be considered an outlier than the other children.

Name: Chandrakant B Thakur

The IQR (Interquartile Range) is not affected by the outliers. One of the most significant reasons is that people mostly prefer to use the IQR while measuring the "spread" of the given data. As the IQR considers the range of the middle that is 50% of the given data value, it does not affect the value of outliers.

How To Find An Outlier In Statistics Using The Interquartile Range (IQR)?

An outlier is described as a data point that ranges above 1.5 IQRs under the first quartile (Q1). Moreover, it lies over the third quartile (Q3) within a set of data.

Low = (Q1) – 1.5 IQR,

High = (Q3) + 1.5 IQR

4. What are our options for dealing with outliers in our dataset?

1. Box Plot

A box plot is a graphical depiction of the distribution of statistics. It makes use of the median as well as the lower and upper quartiles. A Box plot can readily spot an unusual point in the data set since any point above or below the whiskers is an anomaly. Sometimes referred to as the "Univariate method."

Box Plot is a statistical plot to visualize descriptive

statistics(Mean, Median, Q1, Q2, IQR, Minimum, Maximum).

2. Histogram

A histogram in which the majority of the information is on one side while a few observations appear distant from the main group are termed as outliers. Observations outliers.

Histogram also detects outliers.

Inter Quartile Range(IQR)

The interquartile range rule is important for spotting outliers. Inter Quartile Range score or middle 50% or H-spread is a measure of statistical dispersion, being equal to the difference between the 75th percentile and 25th percentile i.e., third quartile(Q3) and first quartile(Q1)

IQR=Q3-Q1

Name: Chandrakant B Thakur

We identify the outliers as values less than Q1 -(1.5 * IQR) or greater than Q3+(1.5 * IQR)

3. Standard Deviation

A measure of how the values in a data set vary or deviate from the mean.

We identify the outliers as values less than (Mean — 3*SD)or greater than (Mean + 3*SD).

4. Scatter Plot

A scatter plot helps in determining the degree of correlation between two numerical variables, such as a simple linear relationship between X and Y. An outlier is any observation that deviates from the ordinary.

5. Write the sample and population variances equations and explain Bessel Correction.

variance is a way to measure the spread of values in a dataset.

The formula to calculate **population variance** is:

$\boldsymbol{\sigma^2} = \Sigma (x_i - \mu)^2 / N$

where:

- **Σ**: A symbol that means "sum"
- **μ**: Population mean
- $\mathbf{x_i}$: The i^{th} element from the population
- **N**: Population size

The formula to calculate **sample variance** is:

$\mathbf{s^2} = \Sigma (x_i - x)^2 / (n-1)$

where:

- **x**: Sample mean
- $\mathbf{x_i}$: The i^{th} element from the sample
- **n**: Sample size

Bessel correction refers to the n-1 part used as the denominator in the formula of sample variance or sample distribution.

Name: Chandrakant B Thakur

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