Descriptive Statistics

1. Central Tendency
   1. Mean – average
   2. Mode – max. frequency occurrence. Eg. 1, 2, 3, 4, 4, 4, 5, 6….mode = 4
   3. Median – central value. Eg. 1, 2, 3, 4, 5, 6, 7….median = 4
2. Measures of Dispersion
   1. Range = Xmax – Xmin
   2. Mean Deviation = Xi – Xmean
   3. Squared Mean Deviation = (Xi – Xmean)2
   4. Variance = sum of Squared Mean Deviation / number of data points

Variance = ∑(Xi – Xmean)2 / n

Common convention is to tweak variance. This is Bassel’s correction.

Variance = ∑(Xi – Xmean)2 / **(n – 1)**

* 1. Standard Deviation = How far your data points lie from the mean. It is a measure of dispersion

Standard Deviation = Square root of Variance

Standard Deviation = **sq. root[**(∑(Xi – Xmean)2 / **(n – 1)]**

* 1. Outlier = May represent data errors or rare points which lie at the extreme ends of data sets. Lies many standard deviations away from the mean.
  2. Inter Quartile Rane
     1. Q1 = 25th percentile, i.e. 25% data is below Q1
     2. Q2 = mean = 50th percentile, i.e. 50% data is below Q2
     3. Q3 = 75th percentile, i.e. 75% data is below Q3

**Inter Quartile Range = Q3 – Q1**

* 1. As staked grow, variance grows faster than mean as Variance = (Xi – Xmean)2 / (n-1)
  2. **Gaussian Normal Distribution (Normal Distribution):**
     1. Consists of: **mean = ****Standard Deviation =** ****
     2. 68% points will lie between ( - ) and ( + )i.e. between 1st standard deviation of the mean.
     3. 95% points will lie between ( -2 ) and ( + 2)i.e. between 2nd standard deviation of the mean.
     4. 99% points will lie between ( - 3) and ( + 3)i.e. between 3rd standard deviation of the mean.
     5. If Std. Deviation () is small, it gives a tall Normal Distribution

If Std. Deviation () is large, it gives a wider Normal Distribution i.e. many points away from mean

1. Sampling Distribution: **Probability Distribution** of a population statistic (mean, variance) for a particular sample.

is for Population mean.

1. Confidence Level: Depends on **size of the sample** data from population and the **variance** of the data.

For a sampling distribution:

Mean = sample mean

Variance = sample variance / n (n is data points in sample)

Std. Dev. = sample

1. Confidence Intervals: ‘We can be 99% confident that the average is between **<<lower range>> and <<upper range>>**’.
2. Low variability = high Confidence level
3. High variability = low Confidence level
4. Higher the sample size, higher the confidence level

**Generalisation:**

* 68% Confidence that  is within 1 of Xmean
* 95% Confidence that  is within 2 of Xmean
* **(100-p)% Confidence that  is within Z of Xmean**
* **p = level of significance, (100 – p) gives confidence in our estimate**
* **Z = number of standard deviations corresponding to value of p**
* ’s Range is centered around sample mean.
* Greater the range, greater the confidence that population mean lies within that range.

1. **Skewness and Kurtosis:**

Skewness = measure of asymmetry around the mean (how the population data is distributed around the sample mean)

For a perfect normal distribution, skewness is zero.

Positive skewness: When outlier falls in higher side. Medial is on left of mean (median < mean)

Negative skewness: When outlier falls in lower side. Medial is on right of mean (median > mean)

**Kurtosis:** frequency of occurrence of extremes (outliers).

For normal distribution, Kurtosis = 3

If Kurtosis > 3, then chances of irregularities is high

If Kurtosis < 3, then chances of irregularities is low

1. **Correlation and Covariance:**

When a variance is obtained from two or more factors, it is called covariance.

Covariance is positive if increase in one variable corresponds to increase in another variable & vice versa.

Covariance is negative if increase in one variable corresponds to decrease in another variable & vice versa

Covariance matrix summarises the relationships between columns

Covariance values are not scalable ranges, hence correlation is used instead. Here, range is between -1 to 1. -1 and 1 shoes strong correlation while 0 shows weak correlation.

Correlation(x,y) = Covariance(x,y) / sq. rt. [(variance(x)] \* sq. rt. [variance(y)])

.