18/10/2019

6 Introduction to Probability & Statistics

(Manoj M)

+ Statistics.

Descriptive Statistics	Inferential Statistics
1. Concerned with the describing the target population.	Make inferences from the sample and generalize them to the population
2. Organize, analyze & present the data in a meaningful manner	Compares, test and predicts ~~Final results is the~~ future ~~probability/scores~~ outcomes.
3. Final results are shown in form of charts, tables and Graphs.	Final result is the probability scores.
4. Describes the data which is already known in detail	Tries to make conclusions about the population that is beyond the data available.
5. Tools - Measures of central tendency (mean/median/mode), Spread of data (range, standard deviation etc)	Tools - Hypothesis Tests, Analysis of Variance etc.

→ Inferential statistics is for forecasting Purpose

→ 100 %. Accurate : No

→ Mean : Average Value

→ Inbuild Tool for Descriptive Statistics in Excel

Go To File → Options → Add-ins → Click on "Go" → Select "Analysis Tool Pack" → Click "ok"

It will be under "Data" Ribbon (Data Analysis)

Data → Data Analysis → Descriptive Statistics → select Input Range → Grouped by "Columns" → check "Labels in first row" box → Select output Range → check "Summary Statistics" box → Click "ok"

→ Median : Middle Value of Selected Array

	Mean	Median		
Company A	8	7		Junior
				Mid ✓
Company B	7	8	✓	High

Company A is having higher mean due to presence of Outlier.

Outlier : The value which upsets the mean value

→ Mode : frequently occuring number
Number which is having higher freq.

→ Range : Max Value — Min Value.

→ Standard Deviation : Squore root of Variance

$$SD = \sqrt{Var} = \sqrt{\frac{(X-\mu)^2}{N}} = \sigma$$

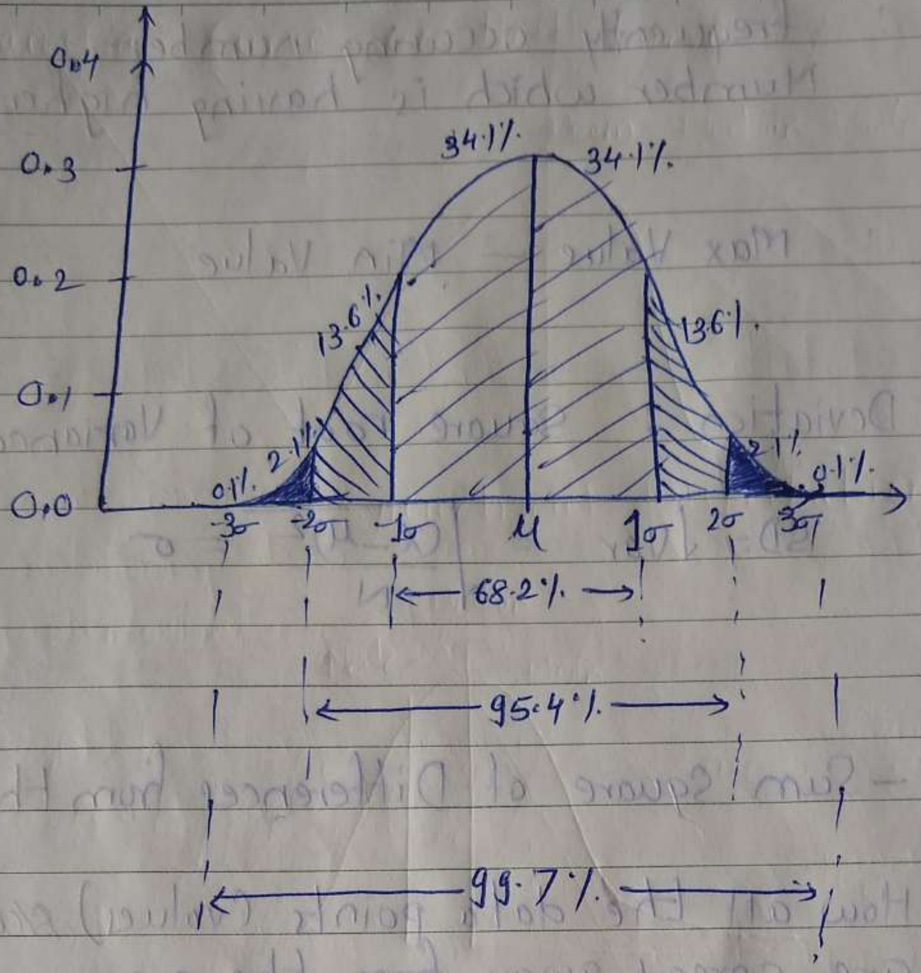
→ Variance - Sum Square of Differences from the mean

How all the data points (values) ~~scatter~~ are spread away from the mean.

X = Value
$\bar{X}$ = Mean. = μ

$$Var = \sum_{i=1}^{n}\frac{(X-\bar{x})^2}{N} \quad or \quad \frac{(X-\mu)^2}{N}$$

Variance does not speak the language of number or dataset.



μ = Mean Value

σ = SD

→ 68.2% of data fall 1 SD from the mean

→ 95.4% of data fall 2 SD from mean

→ 99.7% of data fall 3 SD from mean

→ It is called as 3σ Rule (3 Sigma Rule)

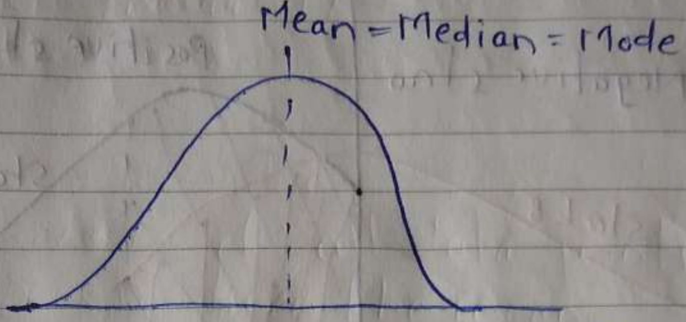
→ Percentages are Pre-defined

Chebyshev's Inequality

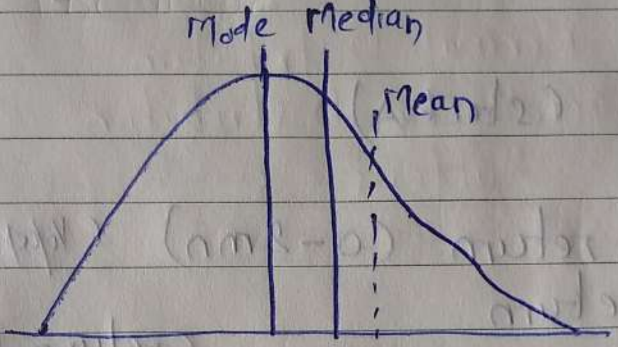
→ Atleast 75% of all data points will lie within 2 SD ~~from~~ of the mean.

→ Atleast 89% of all data points will lie within 3 SD of the mean

Gaussian Curve (Bell Curve)

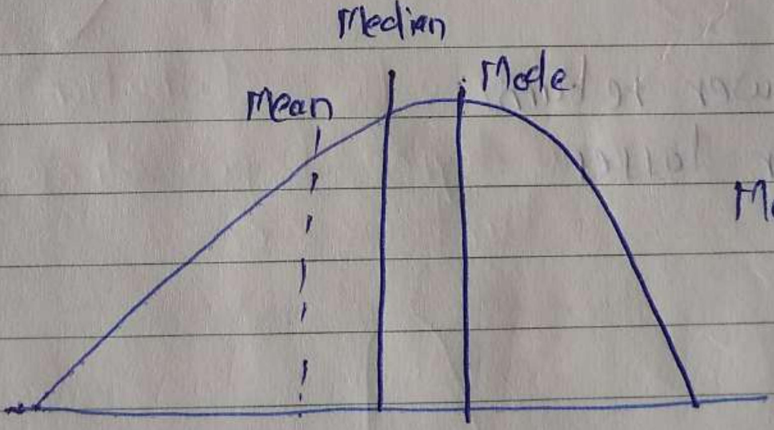


Symmetrical Distribution



Mean > Median

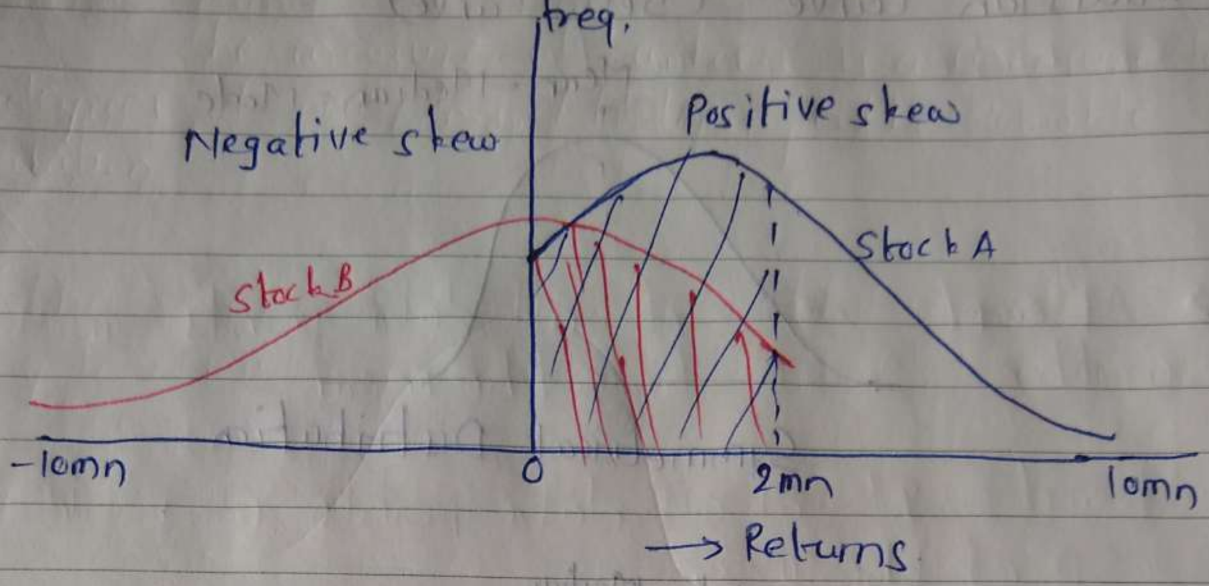
Positive skew



Mean < Median

Negative skew

Stock Returns Example for skewness



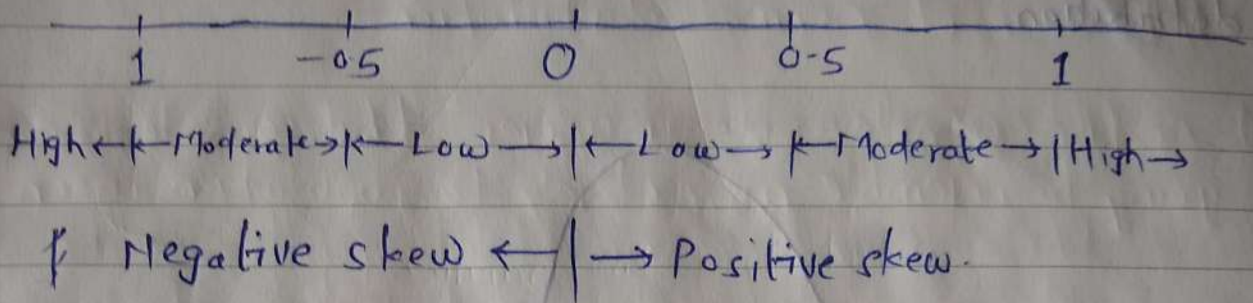
→ Positive skew. (stock A)

* frequent lower return (0-2mn) ([illegible])
* Rare higher return (return/gains)

→ Negative skew (stock B)

* frequent lower return
* Rare higher losses

Skew


1
-0.5
0
0.5
1
High ← |← Moderate → |← Low → | ← Low → |← Moderate → | High →
Negative skew ← | → Positive skew.

$$\text{skewness} = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^3}{(N-1)\, s^3}.$$

$$\text{kurtosis} = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^4}{(N-1)\, s^4}$$

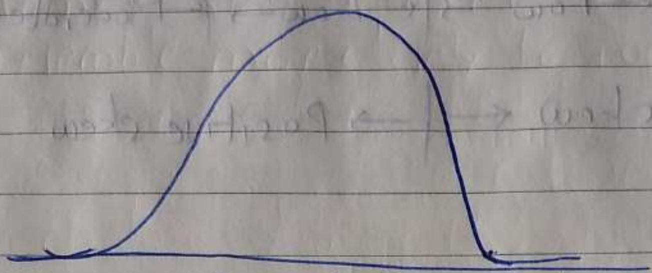
where Y_i = value of ith variable

$\bar{Y}$ = mean of sample

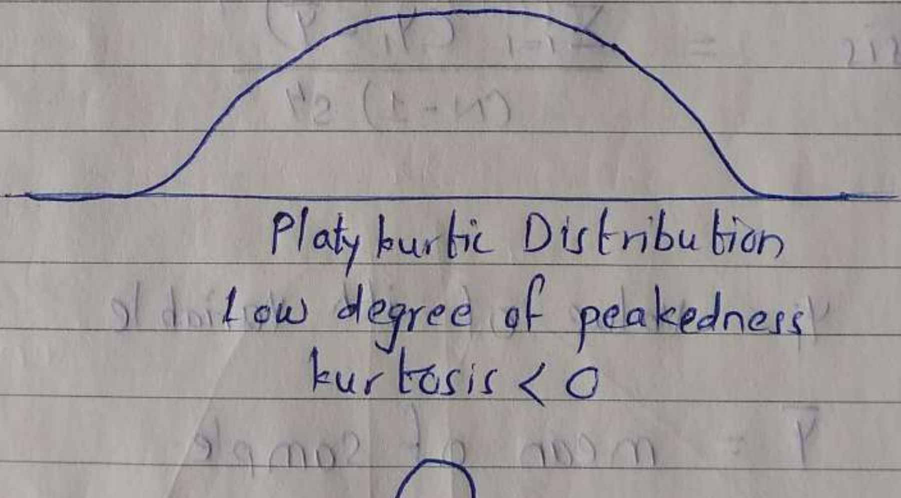
s = SD of sample

Kurtosis :

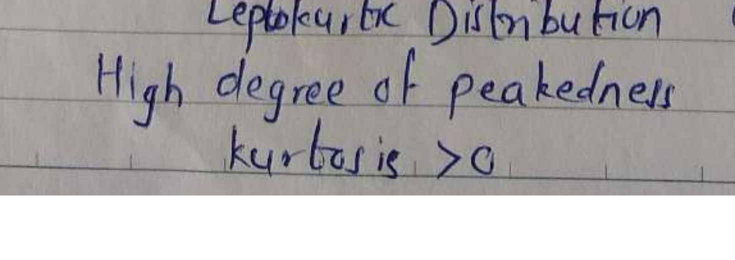
The measure of kurtosis is a measure for the degree of peakedness / flatness in the Variable distribution.



Normal Distribution
Mesokurtic Distribution
kurtosis = 0



Platykurtic Distribution
Low degree of peakedness
kurtosis < 0



Leptokurtic Distribution (Leptokurtic)
High degree of peakedness
kurtosis >0

* Outliers heavily impact skewness & kurtosis but not the standard deviation.

* Hence, SD ~~is~~ can be used widely. (preferred)

* SD gives ~~proper~~ the distribution precisely

* Skewness & kurtosis are used rarely. (only in some applications)

* skewness & kurtosis tells about the extream ends of the data, they don't ~~giv~~ tell about mean. Hence, not widely applicable.

* They are used in financial sector / stock Market.

Standard Error:

Difference between Sample mean and population mean. (Accuracy)

$$SE = \frac{SD}{\sqrt{N}}$$

SE = Standard Error
N = no of observations
SD = Standard Deviation

As the sample size increases, standard error reduces (decreases).

"kaggle.com" : For datasets

Central Limit Theorem

(i) As sample size increases, its going to exhibit behaviour of normal distribution

(ii) As the sample size increases, the sample mean going to resemble more closely to population mean

**Quartile :

A Quartile is a type of Quantile.

First Quartile (Q_1) : [Lower Quartile | 25th percentile)

Middle number between smallest value and median of the dataset.

* splits off the lowest 25% of data from highest 75%

Second Quartile (Q_2) : (Median/ 50th Percentile)

Median of the data.

* cuts data in the half.

Third Quartile (Q_3) : (Upper Quartile/ 75th Percentile)

Middle value between median and highest value of the data set

* splits off the highest 25% of the data from the lowest 75%.

Computing Methods :

Method 1 :

1. Use the median to divide dataset into 2 halves

(i) If there is an odd no of data points in the original ordered data set, do not include the median (central value) in either half.

(ii) If there is an even no of data points in the original ordered data set, split this data exactly in half

2. The lower quartile value is the median of the lower half of data
The Upper quartile value is the median of the upper half of data

* This rule is employed by the "TI-83" Calculator "boxplot" and "1-Var Stats" fn.

Method 2 :

1. Use the median to divide the ordered data set into 2 halves

(i) If there are odd no of data points in the original ordered data set, include the median (central value) in both halves.

(ii) If there are even no of data points in the original ordered data set, split the data set exactly in half.

2. The lower quartile value is the median of the lower half of the data. The upper quartile is the median of the upper half of the data.

* The value found by this method are also known as "Tukey's" hinges.

Method 3 :

1. If there are even no of data points , then method 3 is same as 1 & 2.

2. If there are (4n+1) data points, (i) then the lower quartile is 25% of the nth data value plus 75% of the (n+1)th data value ; ~~th~~.

(ii) The upper quartile is 75% of the (3n+1)th data point plus 25% of (3n+2)th data point.

3. If there are (4n+3) data points then

(i) The lower quartile is 75% of the (n+1) the data value plus 25% of the (n+2)th data value

(ii) The Upper Quartile is 25% of the (3n+2)th data point plus 75% of the (3n+3)th data point

$\boxed{\text{e.g. 1}}$

Ordered data set : 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

Count of data points = 11

Method 1 (Ignore the medaan)

Lower half : 6, 7, $\underline{\underline{15}}$, 36, 39 $\quad Q_1 = 15.$

$Q_2 = 40$

Upper half : 41, 42, $\underline{\underline{43}}$, 47. 49 $\quad Q_3 = 43$

Method 2 : (include median in both halves)

Lower half : 6, 7, $\underline{\underline{15, 36}}$, 39, 40 $\quad Q_1 = \frac{15+36}{2}/_{2} = 25.5$

$Q_2 = 40$

Upper half : 40, 41, $\underline{\underline{42, 43}}$, 47, 49 $\quad Q_3 = \frac{42+43}{2} = 42.5$

Method 3 : (4n+3) datapoints [n=2]

Q_1 = 75% of (n+1) + 25% of (n+2)

= 75% of 15 + 25% 36

= 11.25 + 9

Q_2 = 20.25

Q_2 = 40

$Q_3 = 25\%$. of $(3n+2) + 75\%. (3n+3)$

$= 25\%$ of $42 + 75\%$ of 43.

$= 10.5 + 32.25$

$Q_3 = 43.75$

	Method 1	method 2	Method 3.
Q_1	15	25.5	20.25
Q_2	40	40	40
Q_3	43	42.5	43.75.

$\boxed{\text{e.g.: 2}}$

Ordered data set : 7, 15, 36, 39. 40, 41.

Count = 6

	method 1	Method 2	Method 3.	
Q_1	15	15	15.	
Q_2	37.5	37.5	37.5.	$\left(\frac{36+39}{2}\right)$
Q_3	40	40	40	

Lower half : 7, 15, 36

Upper half : 39, 40, 41.

e.g 3

Ordered data set: 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49, 52, 53

Count of data points: 13

Method 1 (Ignore the median)

Lower half : 6, 7, 15, 36, 39, 40

Upper half : 42, 43, 47, 49, 52, 53

$Q_1 = 25.5$

$Q_2 = 41$

$Q_3 = 48$

Method 2 : (Include median in both halves)

Lower half : 6, 7, 15, 36, 39 40, 41

Upper half : 41, 42, 43, 47, 49, 52, 53

$Q_1 = 36$

$Q_2 = 41$

$Q_3 = 47$.

Method 3 : (4n+1) data points [n = 3]

Q_1 = 25% of nth + 75% of (n+1)

= 25% of 15 + 75% of 36

= 3.75 + 27

Q_1 = 30.75

Q_2 = 41.

$Q_3 = 75\% \text{ of } (3n+1) + 25\% \,(3n+2)$

$= 75\% \text{ of } 47 + 25\% \text{ of } 49$

$= 35.25 + 12.25$

$Q_3 = 47.5.$

	Method 1	Method 2	Method 3
Q_1	25.5	36	30.75
Q_2	41	41	41
Q_3	48	47	47.5

Outliers :

There are methods by which to check for Outliers in the discipline of statistics and statistical Analysis. As is the basic idea of Descriptive statistics, when encouraging an outlier, we have to explain this value by further analysis of the cause or origin of the Outlier. In cases of extreme observations, which are not an infrequent occurrence, the typical values must be analyzed. In the case of quartiles, the Interquartile Range (IQR) may be used to characterize the data when there may be extremities that skew the data ; the interquantile range is a relatively robust statistics (also called as "resistance") compared to the range and standard deviation. There is also a mathematical method to check for outliers and determining "fences", upper and lower limits from which to check for outliers.

After Determining the first & third quartiles and the interquartile range as outlined above, the fences are calculated using following formulas :

Lower fence $= Q_1 - 1.5\,(IQR)$

Upper fence $= Q_3 + 1.5\,(IQR)$

where Q_1 & Q_3 are 1st & 3rd Quratiles. The Lower fence is "Lower limit" & upper fence is "Upper limit" of data and any data lying outside these defined bounds can be considered an "outlier".

← Continued

Anything below the Lower fence or above the Upper fence can be considered such a case. The fences provide a guideline by which to define an outlier, which may be defined in other ways. The fences define a "range" outside which an outlier exists; a way to picture this is a boundry of a fence, outside of which are "outsiders" as opposed to outliers.

$$\boxed{\text{Interquartile Range (IQR)} = Q_3 - Q_1}$$

*** Percentile :

A percentile (or a Centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observation falls.
For example, the 20th percentile is the value below which 20% of the observations may be found.

The term percentile and the related term percentile Rank are often used in the reporting of scores from non-referenced tests.
Foro example, if a score is at the 86th percentile, where 86 is the percentile Rank, it is equal to the value below which 86% of the observations may be found (carefully contrast within the 86th percentile, which means the score is at or below the value which 86% of the observations may be found - every score is in the 100th percentile).

The 25th percentile is also known as first quartile (Q_1), the 50th percentile as the second quartle (Q_2) or median, and the 75th percentile as the third quartile (Q_3).

In general, percentiles and quartiles are specific types of quantities.

Applications :

(i) When ISP's bill "burstable" internal bandwidth, the 95th or 98th percentile usually cuts off the top 5% or 2% of bandwidth peaks in each month and then bills at the nearest rate. In this way infrequent peaks are ignored, and the customer is charged in a fairer way. The reason this statistics is so useful in measuring data throughput is that it gives a very accurate picture of the cost of the bandwidth. The 95th percentile says that 95% of the time, the usage is below this amount. So, the remaining 5% of the time, the usage is above that amount.

(ii) Physicians often use infant and children's weight and height to access their growth in comparison to national averages and percentiles which are found in growth charts.

(iii) The 85th percentile speed of traffic on road is often used as a guideline in setting speed limits & assessing whether such a limit is good too high or low.

(iv) In finance, Value at Risk is a standard measure to assess (in a model dependent way) the quantit under which the portfolio is not expected to sin within a given period of time & given a confidenc value.