

STAT111: Statistics and Probability I

Lecture 3: Descriptive Statistics for Univariate Data

Dr. R. Minkah

University of Ghana
Department of Statistics and Actuarial Science

January 30, 2021

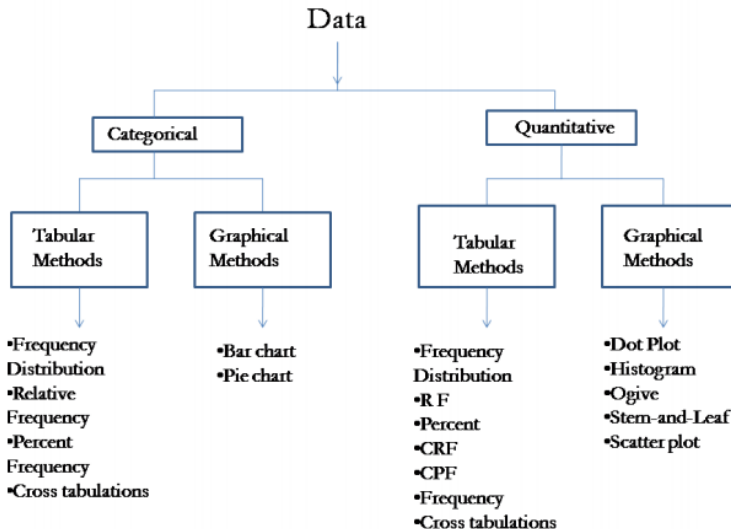
Outline

- Definitions
- Why do we use descriptive statistics
- Measures of Location
 - ① Arithmetic mean
 - ② Geometric mean
 - ③ Harmonic mean
 - ④ Relations between arithmetic, geometric and harmonic means
 - ⑤ Median, Mode and their relationships
 - ⑥ Other measures of location (Quartiles, Deciles and Percentiles)
- Measures of Variability
 - ① Variance, Standard deviation, Coefficient of Variation
 - ② Other Measures of Variation (Range, Inter-Quartile Range, Semi-Interquartile Range)
- Measures of Symmetry and Distribution
 - ① Skewness and Kurtosis
 - ② Characterization of Symmetry

Definitions

- **STATISTICS** is the science pertaining to the collection, organisation, analysis, interpretation or explanation and presentation of data
- **DESCRIPTIVE STATISTICS** is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way. It is the first step of analysis of collected data.
- The purpose of descriptive statistics is to make data sensible or meaningful. it gives a fair idea of how the data behaves.

Recap



Measures of Location

- Implies measures of central and other locations
- The most common measures of location are the mean, mode and median (central location)
- Others include Quartiles, Deciles and Percentiles (Both central and non-central locations)

Mean

- **Mean:** Mean is the most commonly used measure of central tendency. The types of mean include:
 - ① Arithmetic mean
 - ② Geometric mean and
 - ③ Harmonic mean
- The arithmetic mean or simply the mean (average) measures the central location for the data.
- Computed by adding up all values divided by the number of values.
- The population and the sample means are denoted by μ and \bar{x} respectively.

Arithmetic Mean

Population Mean

$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$



Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$



- Here N and n are the population and sample sizes respectively.

Arithmetic Mean for Group Data I

- If the numbers x_1, x_2, \dots, x_n occur with frequencies f_1, f_2, \dots, f_n respectively,
- The mean is given by $\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{\sum f_i}$
- If data is grouped with classes, then we use the class mark, x_i 's for the computation of the mean.

Arithmetic Mean for Group Data II

Example: The following table gives the frequency distribution of the number of orders received each day during the past 50 days at the office of a mail-order company. Calculate the mean.

Number of order	f
10 – 12	4
13 – 15	12
16 – 18	20
19 – 21	14
	$n = 50$

Solution:

Number of order	f	x	fx
10 – 12	4	11	44
13 – 15	12	14	168
16 – 18	20	17	340
19 – 21	14	20	280
	$n = 50$		$= 832$

X is the midpoint of the class. It is adding the class limits and divide by 2.

$$\bar{x} = \frac{\sum fx}{n} = \frac{832}{50} = 16.64$$

Harmonic Mean I

- **Harmonic mean:** A type of mean calculated by dividing the number of values in the data series by the sum of reciprocals of each of the value in the series.
- It is the lowest among the three means
- Harmonic Mean is often used to calculate the average of ratios or rates.
- It is the most appropriate measure since it equalizes the weights of each data points.

Harmonic Mean II

- Harmonic mean is used to determine the average for financial multiples such as Price-to-earnings(P/E) ratio
- one of the most common problems in finance that uses harmonic mean is the ratio of portfolio that consist of several securities
- The formula is, Harmonic Mean, $H = \frac{N}{\sum_{i=1}^N 1/x_i}$, where N is number of data points
- Example, obtain the harmonic mean of 10,13,18,24 and 30

Harmonic Mean III

Example: harmonic mean

$$H = \frac{n}{\sum_{i=1}^n 1/x_i} = \frac{5}{\left[\frac{1}{10} + \frac{1}{13} + \frac{1}{18} + \frac{1}{24} + \frac{1}{30}\right]} = 16.26$$



Geometric Mean

- **Geometric mean** is derived as relevant set of quantities multiplied together and taken n^{th} root
- Mathematically, the geometric mean is given by

$$G = \sqrt[n]{x_1 x_2 x_3 \dots x_n}$$

$$= \sqrt[n]{\prod_{i=1}^n (x_i)}$$

- Example; Find the geometric mean of the data set, 5,6,9,10,15.

Relationship between Arithmetic Geometric and Harmonic Mean I

- Let A, G and H be the Arithmetic, Geometric and Harmonic means respectively of the two real numbers x and y.

$$\Rightarrow A = \frac{x + y}{2}$$

$$G = \sqrt{xy}$$

$$H = \frac{2xy}{x + y}$$

Relationship between Arithmetic Geometric and Harmonic Mean II

$$\begin{aligned}A \times H &= \frac{x+y}{2} \times \frac{2xy}{x+y} \\&= xy \\&= G^2\end{aligned}$$

- Arithmetic Mean (AM) places a high weight on large values while Geometric Mean (GM) places a lower weight to smaller data points
- **Exercise:** If the arithmetic mean and geometric mean of two numbers are 16 and 8 respectively. Find the value of the Harmonic mean.

Median I

- The **median** of a data set is the measure of center that is the *middle value* when the original data values are arranged in order of increasing or decreasing magnitude.
- The median is often denoted by \tilde{x} (pronounced as "x-tilde").
- For a sample of size n , the position of the median is $\frac{1}{2}(n + 1)$.

Median II

- Find the median for this sample of data values: 8,4,9,2,5,12

Example;

- Sort: 2, 4, 5, 8, 9, 12
- Sample size, $n = 6$,
- Position of Median: $\frac{1}{2}(6 + 1) = 3.5$
- \therefore the median is in the 3.5th position in the ordered array and thus obtained as the average of the numbers in the 3rd and 4th positions:

$$\tilde{x} = \frac{5 + 8}{2} = 6.5$$

Mode

- The **Mode** of data set is the value that occurs with the greatest frequency.
- A data set can have one mode, more than one mode or no mode.
- When two data values occur with the same greatest frequency, each one is a mode and the data set is **bimodal**
- When more than two data values occur with the same greatest frequency, each is a mode and the data set is said to be **multimodal**

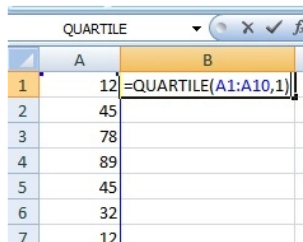
Other measures of locations-Quartiles

Quartile Function Excel

You can also find a quartile in [Microsoft Excel](#) using the Excel quartile function.

1. Type your data into a single column. For example, type your data into cells A1 to A10.
2. Click an empty cell somewhere on the sheet. For example, click cell B1.
3. Type `=QUARTILE(A1:A10,1)` and then press "Enter". This finds the first quartile. To find the third quartile, type `=QUARTILE(A1:A10,3)`.

Note: If your data is in a different cell range other than A1:A10, make sure you change the function to reflect that.



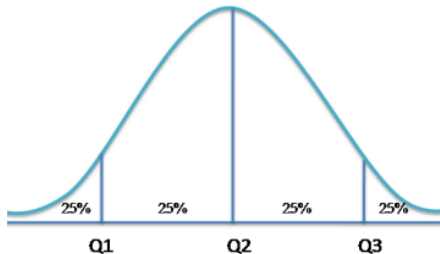
The screenshot shows an Excel spreadsheet with a data table in column A and a formula in cell B1. The data in column A is as follows:

	A	B
1	12	<code>=QUARTILE(A1:A10,1)</code>
2	45	
3	78	
4	89	
5	45	
6	32	
7	12	

Quartiles

- **Quartiles** are measures of location, denoted Q_1 , Q_2 , and Q_3 , which divide a set of data into four groups with about 25% of the values in each group.
- $Q_1 = \frac{(1+n)}{4}$, $Q_2 = \frac{2(n+1)}{4}$, $Q_3 = \frac{3(n+1)}{4}$ th position

Location of Quartiles



First Quartile

2nd Quartile

3rd Quartile

Navigation icons

Quartiles Example

- Example; What is the 3rd quartile (Q_3) of this set?
 $\{1, 2, 3, 3, 4, 5, 5, 5, 6, 6, 7\}$ (Ans = 6)
- Step by step solution
 - First step is to find the median of the data set (5)
 - To find the 1st quartile, you find the middle number above the median (5)
 - To find the 3rd quartile, you find the middle number below the median
 - The set of data below the median is $\{1, 2, 3, 3, 4\}$, hence the 1st quartile is 3
 - The set of data above the median is $\{5, 5, 6, 6, 7\}$, hence the 3rd quartile is 6
 - Alternatively, you can use the formulas above to find the Q_1 and Q_3 .

Deciles

- Deciles are similar to quartiles. But while quartiles sort data into four quarters, deciles sort data into ten equal parts: The 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, 90th and 100th percentiles.
- A decile rank assigns a number to a percentile:

Decile Rank	Percentile
1	10th
2	20th
3	30th
4	40th
5	50th
6	60th
7	70th
8	80th
9	90th

Percentiles I

- A Percentile provides information about how the data are spread over the interval from the smallest value to the largest.
- The p th percentile is a value such that at least p percent of the observations are less than or equal to this value and at least $(100 - p)$ percent of the observations are greater than or equal to this value.
- Calculating the p th percentile;
 - Arrange the data in ascending order (smallest value to the largest value).
 - compute an index i ; $(\frac{p}{100}) \times n$, where p is the percentile of interest and n is the number of observations.

Percentiles II

- If i is not an integer, round it up to the next integer greater than i .
- If i is an integer, the p th percentile is the average of the values in positions i and $i + 1$.

Percentiles and deciles example I

- ABC Investment Company operates a mutual fund targeted at the informal sector workers. The following are the growth rates (in %) of the fund for the last 11-month period:
3, 2, 7, 8, 2, 4, 3, 7.5, 7.2, 2.7, 2.9.
- Determine the 5th decile.

Percentiles and deciles example II

Step by step solution

- First re-arrange the data in ascending order (from smallest to highest):
 $\{2, 2, 2.7, 2.9, 3, 3, 4, 7, 7.2, 7.5, 8\}$
- Establish the 5th decile which is the 50th percentile (median) i.e.

$$\begin{aligned}P_{50} &= \frac{(1 + 11)50}{100} \\&= 12 \times 0.5 \\&= 6 \text{ \{i.e. the 6th data point.\}}\end{aligned}$$

- Therefore, the 5th decile is 3%

Percentiles and deciles example III

- the 5th decile = 50th percentile = median

Measures of variability I

- A measure of variability is a summary statistic that represents the amount of dispersion or spread in a dataset.
- There are four frequently used measures of variability: the range, interquartile range, variance, and standard deviation.
- The simplest measure of variability is the range.

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

- Interquartile range (IQR) is a measure of variability that overcomes the dependency on extreme values.

Measures of variability II

- This measure of variability is the difference between the third quartile Q_3 and the first quartile Q_1 .
- The interquartile range is the range for the middle 50 percent of the data. i.e.

$$IQR = Q_3 - Q_1$$

Variance I

- Variance is a measure of variability that utilizes all the data. The variance is based on the difference between the value of each observation and the mean. The difference between the observed value and its mean $(x_i - \bar{x})$ or $(x_i - \mu)$ is called a *deviation* about the mean
- Mathematically Population Variance is given by,
$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$
 and Sample variance, $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$
- An alternative formula for the computation of sample variance is $s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$

Variance II

- Standard Deviation is defined as the positive square root of the variance.

Sample Deviation $S = \sqrt{S^2}$

Population Deviation $\sigma = \sqrt{\sigma^2}$

Coefficient of Variation I

- Coefficient of Variation (CV): This descriptive statistics indicates how large the standard deviation is relative to the mean. It is usually expressed in percentage.
- Coefficient of Variation is computed as:

Population CV

$$CV = \frac{\sigma}{\mu} \times 100$$

•

Coefficient of Variation II

Sample CV

$$CV = \frac{s}{\bar{x}} \times 100$$

-
- Coefficient of Variation is used to compare the variability of sets of data measured in different units. For example, we may wish to know, for a certain population whether body masses measured in kilograms, are more variable than heights measured in centimetres.

Coefficient of Variation III

- Also CV is used to compare the variability of sets of data measured in the same units, but whose means are quite different.
- The following data, 11,16,10,30,24,5,6,12,11,45,9,8,3,4,35,31, represents the number of days spent by COVID 19 patients admitted at the Intensive Care Unit of the University of Ghana Medical Centre. Find
 - 1 the mean,
 - 2 range,
 - 3 interquartile range,
 - 4 variance and standard deviation
 - 5 the coefficient of variation.
 - 6 Comment on your results

Introduction I

- We have learned numerical measures of location and spread .
- What about measures of shape?
- The histogram can give a general idea the shape but the numerical measure of shape gives a precise value.
- We discuss Relative Location, Skewness and Kurtosis in this section.
- Relative Location: The distance of the observation from the mean.
- Skewness is a measure of symmetry: A distribution, or data set, is symmetric if the left and right of the center point are mirror images of each other.

Introduction II

- Kurtosis is a measure of whether the data are longer-tailed or light-tailed relative to a normal distribution (Symmetric distribution). In other words, how tall and sharp the central peak is, relative to a standard bell-shaped curve.

Relative Location I

- Measure of relative location helps to determine how far a particular value is from the mean.
- By using the mean and standard deviation, it is very easy to determine the relative location of any observation.
- A value called z-score (often called the standardized value) is used for such characterisation

$$z_i = \frac{x_i - \bar{x}}{s},$$

where

- z_i : the z- score for x_i ,
 - \bar{x} :the sample mean
 - s :the sample standard deviation.
- The z-score z_i , can be interpreted as the number of standard deviations x_i is from the mean \bar{x} .

Detecting Outliers I

- An **outlier** is a value that “lies outside” (is much smaller or larger than) most of the other values in a set of data.
- For example in the scores 250,290,3,320,850,330,270,280 both 3 and 850 are “**outliers**”.
- **Chebyshev's Theorem** enables us to make statement about the proportion of the data values that must be within a specified number of standard deviations of the mean.
- At least $(1 - \frac{1}{z^2})$ of the data values must be within z standard deviations of the mean, where z is any value greater than 1.

Detecting Outliers II

- At least 0.75 or 75%, of the data values must be within $z = 2$ standard deviations of the mean.
- At least 0.89 or 89%, of the data values must be within $z = 3$ standard deviations of the mean.
- At least 0.94 or 94%, of the data values must be within $z = 4$ standard deviations of the mean.

Detecting Outliers III

- **Empirical Rule**

For data having bell - shaped distribution : Approximately 68% of the data values will be within one standard deviation of the mean. Approximately 95% of the data values will be within two standard deviations of the mean. Almost all of the data values will be within three standard deviations of the mean.

- Experience Statisticians take steps to identify outliers and then review each one carefully.

Detecting Outliers IV

- An outlier may be a data value that has been incorrectly recorded. If so, it can be corrected before further analysis. An outlier may also be from an observation that was incorrectly included in the data set; if so, it can be removed. Finally, an outlier may be an unusual data value that has been correctly and belongs in the data set. In such cases it should remain.
- Standardized values (z-score) can be used to identify outliers. We recommend treating any data with a z-score less than -3 or greater than $+3$ as an outlier.
- Such data values can then be reviewed for accuracy and to determine whether they belong to the data set.

Background Problem I

In January 2009, 18 men and 18 women entered a half-marathon. Finish times in minutes are as follows (Naples Daily News, January 19, 2009). Times are shown in order of finish:

Men	Women	Men	Women
65.30	122.62	131.80	136.75
66.52	109.03	109.05	138.00
66.85	111.22	110.23	139.00
70.87	111.65	112.90	147.18
87.18	111.93	113.52	147.35
96.45	114.38	120.95	147.50
98.52	118.33	127.98	153.88
100.52	121.25	128.40	154.83

Background Problem II

- 1 What is the median time for men and women runners? compare men and women runners based on their median times.
- 2 Provide a five-number summary (maximum, minimum, 1st, 3rd and median values) for both the men and women.
- 3 Construct box and whisker plots for the groups. Did men or women have the most variation in finish times? Explain.
- 4 Are there any outliers in the data? Explain your answer

Background Problem I I

- Annual sales, in 1000 Ghana Cedis, for 21 pharmaceutical companies listed with the Pharmacy Council of Ghana are as follows:

8408	1374	1872	8879	2459	11413
608	14138	6452	1850	2818	1356
10498	7478	4019	4341	739	2127
3653	5794	8305			

- Provide a five-number summary i.e. minimum, first quartile, median, third quartile, and maximum.
- Do the data contain any outliers?

Background Problem I II

- Suppose a mistake has been done in the entry process and the sales has been entered as 41138 (transposition) million.
Would the method of detecting outliers above identify this problem and allow for correction of the data

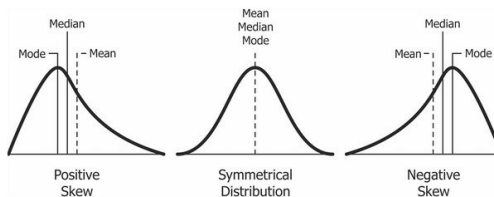
Background Problem II I

The result of a national survey showed that on the average, adults sleep 6.9 hours per night. Suppose that the standard deviation is 1.2 hours.

- 1 Use chebyshev's theorem to calculate the percentage of individuals who sleep between 4.5 and 9.5 hours.
- 2 Use chebyshev's theorem to calculate the percentage of individuals who sleep between 3.9 and 9.9 hours.
- 3 Assume that the number of hours of sleep follows a bell-shaped distribution. Use the empirical rule to calculate the percentage of individuals who sleep between 4.5 and 9.3 hours per day. How does this result compare to the value that you obtained using the Chebychev's theorem in part (1).

Measures of Skewness I

- An important numerical measure of shape of a distribution is called Skewness.
- There are two types of skewness: Positive and Negative



Measures of Skewness II

- **Positive Skewness** means that the tail on the right side of the distribution is longer or fatter. The mean and median will be greater than the mode
- **Negative Skewness** is when the tail of the left side of the distribution is longer or fatter than the tail on the right side. The mean and median will be less than the mode
- If the data are **symmetric**, the skewness is zero: the mean, median and mode are equal.
- The mean is usually greater than the median for positively skewed data.

Measures of Skewness III

- The mean will usually be less than the median for negatively skewed data.
- The median provides the preferred measure of location when the data are highly skewed.
- **So when is the skewness too much?**
A simple rule of thumb:
 - If skewness is between -0.5 and 0.5, the data are fairly symmetrical.
 - If the skewness is between -1 and -0.5 (negatively skewed) or between 0.5 and 1 (positively skewed), the data are moderately skewed.
 - If the skewness is less than -1 (negatively skewed) or greater than 1 (positively skewed), the data are highly skewed.

Measures of Skewness IV

Skewness: Example

Suppose we have house price values ranging from GHS100,000 to GHS1,000,000 with the average being GHS500,000.

- If the peak of the distribution was left of the average value, portraying a *positive skewness* in the distribution. It would mean that many houses were being sold for less than average value, i.e. GHS500,000. This could be for many reasons, but we are not going to interpret those reasons in this chapter.
- If the peak of the distribution of the data was right of the average value, that would mean a *negative skew*. This would mean that houses were being sold for more than the average value.

Computing Measure of Skewness I

- Coefficient of skewness of a data set is:

$$g_1 = \frac{m_3}{m_2^{3/2}}$$

where

-

$$m_3 = \sum_{i=1}^n \frac{(x_i - \bar{x})^3}{n}$$

and

$$m_2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n}.$$

- m_3 is third moment and m_2 is the variance.

Computing Measure of Skewness II

- Many softwares compute an adjusted sample skewness given by

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$

Background Problem III I

- Below are some appliance brand and their rating scaled from 1 to 5, with 5 being best.

Brand	Rating	Brand	Rating
Sony	4.00	Acatel	4.67
Ericson	4.12	Hp	2.14
Nokia	3.82	Acer	4.09
Hisonic	4.00	Toshiba	4.17
Hisense	4.56	apple	4.88
Panasonic	4.32	Dell	4.26
Blackberry	4.33	Binatone	2.32
Philips	4.50	PSB	4.50
Sharp	4.64	Infinity	4.17
HTC	4.20	Bose	2.17

- Compute the mean and the median.
- Compute the first and the third Quartile.
- Compute the standard deviation.

Background Problem III II

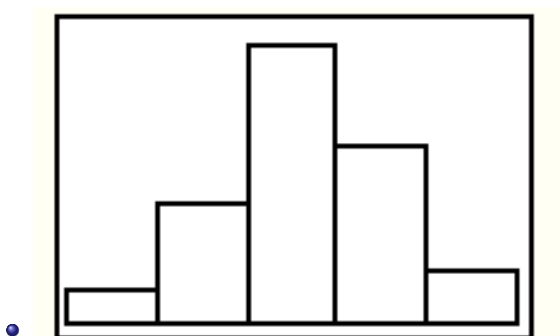
- The skewness of the data is -1.67. Comment on the shape of the distribution.
- What are the z-score associated with Hisense and Sony? Do the data contain any outliers? Explain.

Illustration I

The table below shows data for heights of 100 randomly selected male students, adapted from Spiegel and Stephens (1999,68).

Height (inches)	Class Mark, x	Frequency, f
59.5 to 62.5	61	5
62.5 to 65.5	64	18
65.5 to 68.5	67	42
68.5 to 71.5	70	27
71.5 to 74.5	73	8

Illustration II



- The Histogram shows that the data are slightly skewed left, not symmetric.
But **how highly skewed** are they, compared to other data sets?

Illustration III

- To answer this question, you have to compute the skewness.
- Begin with the sample size and sample mean. (The sample size was given, but it never hurts to check.)

$$\begin{aligned}
 n &= 5 + 18 + 42 + 27 + 8 = 100 \\
 \bar{x} &= \frac{(61 \times 5) + (64 \times 18) + (67 \times 42) + (70 \times 27) + (73 \times 8)}{100} \\
 &= \frac{305 + 1152 + 2814 + 1890 + 584}{100} \\
 &= 67.45
 \end{aligned}$$

- Now with the mean calculated, you can compute the skewness.

Illustration I

Class Mark, x	Frequency, f	xf	(x- \bar{x})	(x- \bar{x}) ² f	(x- \bar{x}) ³ f
61	5	305	-6.45	208.01	-1341.68
64	18	1152	-3.45	214.25	-739.15
67	42	2814	-0.45	8.51	-3.83
70	27	1890	2.55	175.57	447.70
73	8	584	5.55	246.42	1367.63
Σ		6745	n/a	852.75	-269.33
\bar{x}, m_2, m_3		67.45	n/a	8.5275	-2.6933

- Finally, the skewness is

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{-2.6933}{8.5275^{3/2}} = -0.1082.$$

Illustration II

- The adjusted sample coefficient of skewness is computed as follows:

Illustration

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1 = \frac{\sqrt{(100 \times 99)}}{98} \times \frac{-2.6933}{8.5275^{3/2}} = -0.1098$$

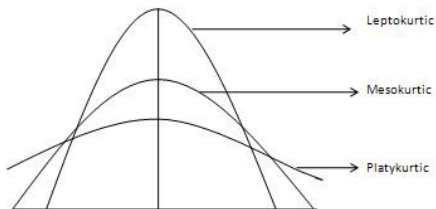


Measures of Kurtosis I

- Kurtosis is all about the tails of the distribution not the peakedness or flatness.
- It is used to describe the extreme values in one versus the other tail.
- It is actually the measure of outliers present in the distribution
- **High kurtosis** in a data set is an indication that data has heavy tails or outliers
- It also calls for investigation wrong data entry or something else.

Measures of Kurtosis II

- **Low kurtosis** in a data indicate that the data has light tails or lack of outliers. (also investigate)



Measures of Kurtosis III

- **Mesokurtic:** This distribution has kurtosis statistic similar to that of the normal distribution. It means that extreme values of the distribution are similar to that of a normal distribution characteristics. This definition is used so that the standard normal distribution has a *kurtosis of three*.
- **Leptokurtic**($Kurtosis > 3$); Distribution is longer, tails are fatter. Peak is higher and sharper than Mesokurtic, which means that data are heavy-heavy tailed or profusion of outliers
- Outliers stretch the horizontal axis of the histogram graph, which makes the bulk of the data appear in a narrow ('skinny') vertical range, thereby giving the "skinniness" of a leptokurtic distribution

Measures of Kurtosis IV

- **Platykurtic:** ($Kurtosis < 3$): Distribution is shorter, tails are thinner than the normal distribution. The peak is lower and broader than Mesokurtic, which means that data are light-tailed or lack of outliers.
- The reason for this is because the extreme values are less than that of the normal distribution.
- The **moment coefficient of kurtosis** of a dataset is computed almost the same way as the coefficient of skewness

Computation of Kurtosis I

- Kurtosis:

$$a_4 = \frac{m_4}{m_2^2}$$

and

- Excess kurtosis:

$$g_2 = a_4 - 3$$

where

$$m_4 = \frac{\sum (x - \bar{x})^4}{n}$$

and

$$m_2 = \frac{\sum (x - \bar{x})^2}{n}$$

- Sample excess kurtosis: $G_2 = \frac{n-1}{(n-2)(n-3)} [(n+1)g_2 + 6]$