STATISTICS

Unit 1

1. Introduction
2. Descriptive Statistics

   2.01 Descriptive Data Measures

   Mean, Median, Mode, Quartiles, Percentiles

   2.02 Central tendency & variability

   2.03 Other Descriptive Data Measures

   2.04 Types of measures of dispersion

   2.041 Positional Measures

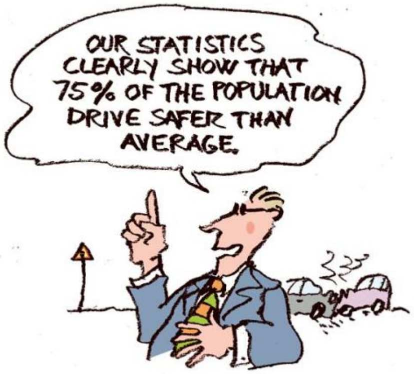
   Range, Quartile Deviation, Mean Deviation, Standard Deviation

   2.042 Relative Measure

   Coefficient of Variation, Skewness, Moments, Kurtosis

1 Introduction

- The statistics most people are familiar with from the news, sports and books are usually about describing data.
- Statistics is also using data to predict things that are unknown. We can be 95% confident more people will vote for Candidate A than B.
- We start with questions like which candidate A or B will win in an election and attempt to answer them with data. After gathering enough data we make predictions.



1 Introduction – continued

- We deal with two types of statistics: Descriptive & Inferential.
- Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way.
- This results in patterns emerging from the data.
- Descriptive statistics is the discipline of quantitatively describing the main features of a collection of information.
- Examples of descriptive statistics:

1. *Frequency distributions*
2. *Measures of central tendency (mean, median, and mode)*
3. *Graphs like pie charts, bar charts that describe data*

1 Introduction – continued

- Inferential statistics is concerned with making predictions or inferences about a population on the evidence of sample data.
- To address this issue of generalization, we have tests of significance.
- A Chi-square or T-test, for example, can tell us the probability that the results of our analysis on the sample are representative of the population that the sample represents.
- Examples:

1. *Linear regression analysis*
2. *ANOVA*
3. *Correlation analysis*

2 Descriptive Statistics

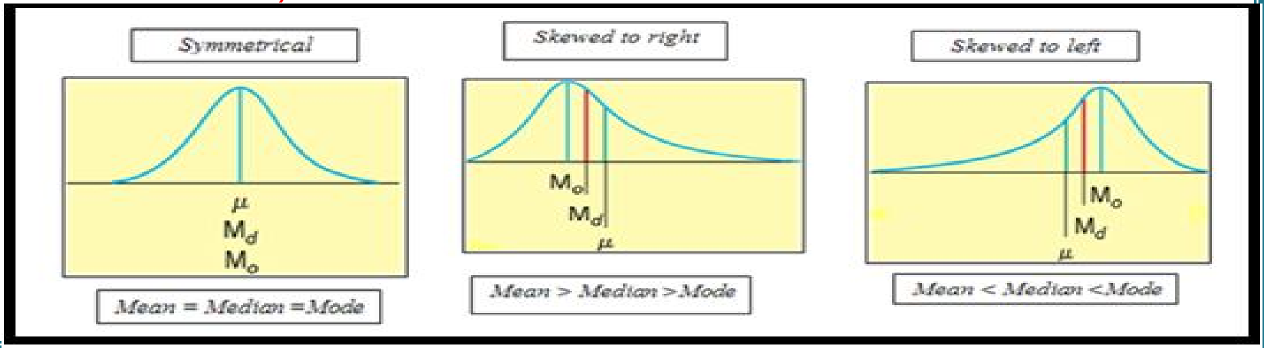
2.01 Descriptive data measures

- Measures of central tendency* (includes: mean, median & mode)
  - *For example, we would like to know the overall performance of 100 students in an examination.*
  - *We can describe this central position using a number of statistics, including mode, 1median and mean.*

Note

*a) * A central tendency is a single figure that represents whole of the distribution*

b) The median lies between the mean and mode and is usually approximately one third of the way from the mean to the mode.



2 Descriptive Statistics – continued

2.01 Descriptive data measures – continued

- Measures of central tendency* (includes: mean, median & mode)
  - *For example, we would like to know the overall performance of 100 students in an examination.*
  - *We can describe this central position using a number of statistics, including mode, 1median and mean.*

Student #	Scores	Remarks
1	50	
2	54	
3	54	Mode
4	54	
5	55	Median
6	65	
7	70	
8	80	
9	90	
Total	**572**	
Arithmetic Mean	**63.6**	= 572 / 9 Mean
Mode	**54**	Most occuring score
Median	**55**	Middle number is in the position = 5^{th} = (9 + 1) / 2

Note

*a) * A central tendency is a single figure that represents whole of the distribution*

b) The median lies between the mean and mode and is usually approximately one third of the way from the mean to the mode.

c) With a symmetrical distribution, the mean, median, and mode will all be the same.

d) If the distribution is skewed to the right, the mean will be to the right of the mode; if t is skewed to the left, the mean will be to the left of the mode;

2 Descriptive Statistics – continued

2.01 Descriptive data measures – continued

Arithmetic Mean

$$\text{mean} = \frac{1}{n}\sum_{i=0}^{n} a_i \qquad \text{mean} = \frac{\text{sum of elements in set}}{\text{number of elements in set}}$$

Given a set of **n** elements from a_1 to a_n

$$\text{mean} = \frac{a_1 + a_2 + a_3 + a_4 + \ldots.. + a_{n-1} + a_n}{n}$$

If f_1, f_2,..f_n are the weights of a_1,a_2,..a_n, then its arithmetic mean is given in the picture.

$$\text{mean} = \frac{1}{\mathbf{N}}\sum_{i=0}^{n} a_i\,\mathbf{f}_i \qquad \mathbf{N} = \sum_{i=0}^{n} \mathbf{f}_i$$

Find the Arithmetic Mean of the following data:

Age in years (x)	8	10	12	15	18	Σ	
Number of workers (f)	5	7	12	6	10	40	Σ(f)
fx	40	70	144	90	180	524	Σ fx
					AM	13.1	$= \frac{524}{40}$

2 Descriptive Statistics – continued

2.01 Descriptive data measures – continued

Geometric Mean

- Geometric mean is a type of average which indicates the central tendency or typical value of a set of numbers by using the product of their values.

The geometric mean of a data set $\{a_1, a_2, \ldots, a_n\}$ is given by:

$$\left(\prod_{i=1}^{n} a_i\right)^{1/n} = \sqrt[n]{a_1 a_2 \cdots a_n}.$$

The geometric mean is defined as the nth root of the product of n numbers.

- The geometric mean is more appropriate than the arithmetic mean for describing proportional growth (constant proportional growth) and varying growth.
- In business the geometric mean of growth rates is known as the compound annual growth rate (CAGR).

2 Descriptive Statistics – continued

2.01 Descriptive data measures – continued

Geometric Mean

- Suppose an orange tree yields 100 oranges in one year and then 180, 210 and 300 the following years.
- By using the arithmetic mean, we calculate the (linear) average growth of 46.5079% – i.e., (80 % + 16.6666% + 42.85261 % divided by 3). However, if we start with 100 oranges and let it grow 46.5079% each year, the result is 314 oranges, not 300 oranges.
- The growth rate is 1.8, 1.16666, 1.428571. By using geometric mean, $\sqrt[3]{1.80 \times 1.166666 \times 1.428571} = 1.442249$

  We get 1.442249 as the average growth per year resulting in 300 oranges for the third year.

2 Descriptive Statistics – continued

2.01 Descriptive data measures – continued

Harmonic Mean

- Harmonic mean is one of the several kinds of averages.

The harmonic mean H of the positive real numbers $x_1, x_2, \ldots, x_n > 0$ is

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Typically, it is appropriate when the average of rates is desired.

The harmonic mean is defined as the reciprocal of the arithmetic mean of the reciprocal of n numbers.

	Distance	Speed (kmph)	
Route 1	50	60	x
Route 2	50	40	y
Average speed	= 2 / ((1/x)+(1/y))	48	kmph

Example: If a vehicle travels a certain distance at a speed x (say 60 kmph) and then the same distance at a speed y (say 80 kmph), then its average speed is the harmonic mean of x and y (48 kmph).

2 Descriptive Statistics – continued

2.01 Descriptive data measures – continued

Quartiles

For a set of observations arranged in ascending order of magnitude,

Q_1 = the value of $((N+1)/\ 4)$th observation

Q_2 = the value of $((N+1)/\ 2)$th observation

Q_3 = the value of $(3(N+1)/\ 4)$th observation

In a frequency distribution,

$Q_i = li + (((N/4) - m) / fi) * c;\ I = 1,2,3....$

Where

l_i = lower limit of the Q_i class

m_i = cumulative frequency of the preceding class

f_i = Frequency of the Q_i class

c = Class interval

N = Total frequency

Median is called the second quartile.

2 Descriptive Statistics – continued

2.01 Descriptive data measures – continued

Percentiles

- *Percentiles are like quartiles, except that percentiles divide the set of data into 100 equal parts while quartiles divide the set of data into 4 equal parts.*
- *Percentiles measure position from the bottom.*
- *Percentiles are most often used for determining the relative standing of an individual in a population or the rank position of the individual. Some of the most popular uses for percentiles are connected with test scores and graduation standings.*
- *Percentiles ranks are an easy way to convey an individual's standing at graduation relative to other graduates.*
- *A percentile rank is the percentage of scores that fall at or below a given score.*

2 Descriptive Statistics – continued

2.01 Descriptive data measures – continued

Percentiles formula – 1

To find the percentile rank of a score x, out of a set of n scores, where x is included:

((X + 0.5 Y) / n) * 100 = Percentile rank

Where X = Number of scores below x

Y = Number of scores equal to x

n = Number of scores

Consider the average marks of ten students in a school final exam:

50,55,60,65,70,70,80,85,90,95

Percentile rank for mark 70= ((4 + 0.5 * 2) / 10) * 100 = 50

Percentile rank for mark 80= ((6 + 0.5 * 1) / 10) * 100 = 65

Read more : http://www.ehow.com/how_2310404_calculate-percentiles.html

2 Descriptive Statistics – continued

2.01 Descriptive data measures – continued

Percentiles formula – *Newest Rank method*

- *The P-th percentile (0 <= P <= 100) of a list of N ordered values (sorted from least to greatest) is the smallest value in the list such that P percent of the data is less than or equal to that value.*
- *The ordinal rank n is calculated as $n = (P/100) * N$*
- *Where P is the percentile and N is the total number of observations*
- *Consider the ordered list (15,20,35,40,50). What is the 80th percentiles of this list using the Newest Rank method?*
- *$n = (80/100)*5 = 4$. The value 40 is the element in the list corresponding to 80th percentile.*

Read more :http://en.wikipedia.org/wiki/Percentile

2 Descriptive Statistics – continued

2.01 Descriptive data measures – continued

Mode

In the case of a frequency distribution of a continuous variable, mode is obtained by giving weights to the frequencies of the modal class and pre-modal class.

Mode		=	$l + ((f_1 - f_0) / (2f_1 - (f_0 + f_2))) * c$
Where	l	=	*lower limit of the modal class*
	f_1	=	*Frequency of the modal class*
	f_0	=	*Frequency of the pre-modal class*
	f_2	=	*Frequency of the post-modal class*
	c	=	*Class interval*

2 Descriptive Statistics – continued

2.01 Descriptive data measures – continued

Example

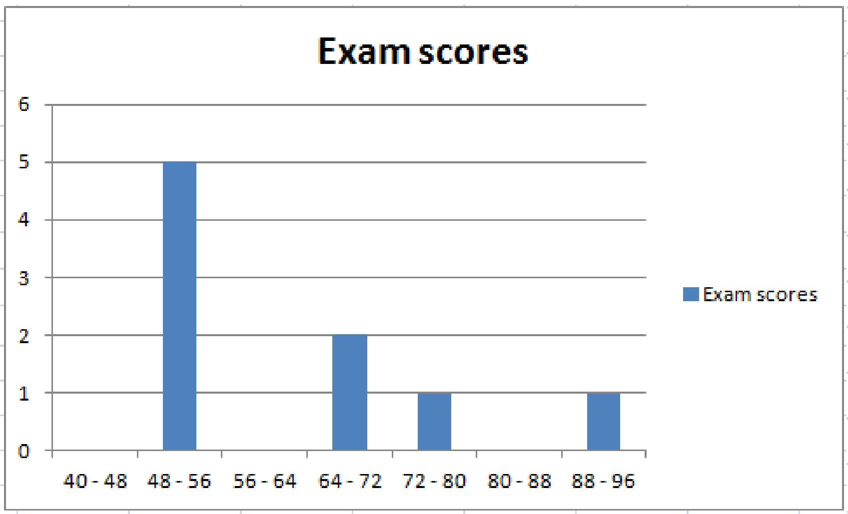
In the case of a frequency distribution of a continuous variable, mode is obtained by giving weights to the frequencies of the modal class and pre-modal class.

Value		10 -20	20 -30	30 -40	40 -50	50 -60	60 -70	70 -80	
Frequency		5	12	25	20	24	10	4	100
cum. Freq		5	17	42	62	86	96	100	
N /2 =		=100/2 =	50						
c	10								
l	40								
m	42								
f	20								
Median	=l + (((N/2) - m)/f)*c								
	= 40 + ((50 - 42)/20)*10	**44**	**Median**						
Mode	= l + (((f_1 - f_0)/ (2f_1 - (f_0+f_2)))*c								
l	30								
f_0	12								
f_1	25								
f_2	20								
c	10								
	= 30 + (((25 - 12)/(2*25 - (12+20)))*10								
	=30 + ((13/ (50 - 32)*10								
	= 30 + (13/18)*10								
	=30 + 7.22	**37.22**	**Mode**						

2 Descriptive Statistics – continued

2.01 Descriptive data measures – continued

- *In this case, the frequency distribution and pattern of the marks scored by 100 students from the lowest to the highest.*



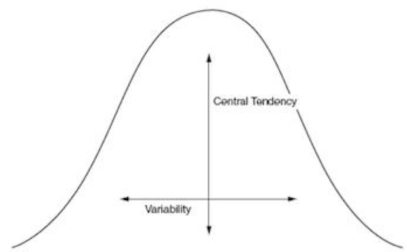
Frequncy distribution

Scores	Frequency
40 - 48	0
48 - 56	5
56 - 64	0
64 - 72	2
72 - 80	1
80 - 88	0
88 - 96	1
Total	**9**

2 Descriptive Statistics – continued

2.02 Central tendency & Variability

- Central tendency describes the central point in a data set. Variability describes the spread of the data.
- Measures of central tendency use a single value to describe the center of a data set. (Examples: Mean, Median and Mode)
- Variability describes the spread of the data.



2 Descriptive Statistics – continued

2.03 Other Descriptive data measures

- Measures of variability or dispersion* (includes: standard deviation, range, kurtosis and skewness)
- These are the ways of summarizing a group of data by describing how spread out the scores are.
  - *For example, the mean score of 100 students may be 70 out of 100.*
  - *However, not all students will have scored 70 marks. Rather, their scores will be spread out; some scores lower and some scores higher.*
  - *To describe this spread, a number of statistics are available to us, including the range, quartiles, standard deviation.*

*Note * The degree to which numerical data tend to spread about an average value or among itself is called variation or dispersion of data.*

2 Descriptive Statistics – continued

2.04 Types of measures of dispersion

1. **Absolute Dispersion**

   When we are interested in determining the degree of variability in a given distribution, we make use of absolute dispersion.

2. **Relative Dispersion**

   If we are interested in comparing the degree of variability in two or more distributions we make use of relative dispersion.

- Absolute Dispersion
  - Range
  - Quartile Deviation (QD)
  - Mean Deviation (MD)
  - Standard Deviation (SD)
- Relative Dispersion
  - Coefficient of QD
  - Coefficient of MD
  - Coefficient of variation

Goals

1. *To determine the reliability of an average*
2. *To serve as a basis for the control of variability*
3. *To compare two or more series with regard to their variability*
4. *To facilitate the use of other statistical measure*

2 Descriptive Statistics – continued

2.041 Positional measures

Range and quartile deviations are positional measures of dispersion, wherein all the observations are not taken into account in the calculation.

a. Range

- *Durell Huff has pointed out "Place little faith in an average when important figures are missing. Otherwise you are as blind as man choosing a camp-site from a report of mean temperature alone. You can freeze or roast if you ignore the range."*
- *This explains the importance of range.*
- *Range is defined as the difference between the highest and lowest value of the variable in the series.*
- *It does not take into account of the frequency of different variates.*

- *For the scores set: 50,54,54,54,55,65,70,80,90*

  Minimum value (M) = 50 ; Highest value (H) = 90
  Range = H – M = 90 – 50 = 40
  Coefficient of Range = (H – M) / (H + M) = (90 – 50) / (90 +50) = 0.2857

2 Descriptive Statistics – continued

b. Quartile Deviation

	Determining Quartiles								
1	Order the data from smallest to largest								
2	Find the place that occupies every quartile using the expression (k*N) / 4, k = 1,2,3, in the cumulative frequency tables.								
Odd number of data	2	3	4	6	7	8	9		
		Q_1		Q_2		Q_3			
				Median					
					$Q_3 - Q_1$	= 8 - 3	= 5	Inter Quartile Range	
					$(Q_3 - Q_1)/2$		= 2.5	QD	
					$(Q_3 - Q_1)/$ (Q3 + Q1)	= 5 / 11 =	0.45455	Coefficient of QD	
Even number of data	2	3	4	6	7	8	9	10	
		(3+4)/2 = 3.5		(6+7)/2 = 6.5		(8+9)/2 = 8.5			
		Q_1		Q_2		Q_3			
				Median					
					$Q_3 - Q_1$	= 8.5- 3.5	= 5	Inter Quartile Range	
					$(Q_3 - Q_1)/2$		= 2.5	QD	
					$(Q_3 - Q_1)/$ (Q3 + Q1)	= 5 / 12 =	0.41667	Coefficient of QD	

Quartiles - Divides the distribution into quarters

Q_1 Q_2 Q_3

Inter Quartile Range = $Q_3 - Q_1$

Quartile Deviation (QD) = $(Q_3 - Q_1) / 2$

Coefficient of QD = $(Q_3 - Q_1)/(Q_3 + Q_1)$

2 Descriptive Statistics – continued

c. Mean Deviation

Mean deviation (MD) is defined as an average or mean of the deviations of the values from the central tendency. The central tendency used can be either arithmetic mean or median or mode.

Example:

Value		10 -20	20 -30	30 -40	40 -50	50 -60	60 -70	70 -80	Σ
Frequency		5	12	25	20	24	10	4	100
cum. Freq		5	17	42	62	86	96	100	
N /2 =		=100/2 =	50						
c	10								
l	40								
m	42								
f	20								
Median	=l + (((N/2) - m)/f)*c								
	= 40 + ((50 - 42)/20)*10	**44**	**Median**						
Mid Value (x)		**15**	**25**	35	45	55	65	75	
dx	= $\lvert x - 44 \rvert$	**29**	**19**	**9**	**1**	**11**	**21**	**31**	121
fdx		**145**	**228**	**225**	**20**	**264**	**210**	**124**	1216
Mean Deviation (MD)	Σ fdx / Σ dx	**10.05**	= 1216 / 121						
Coefficient of MD	= Mean Deviation / MD	**= 10.05 /44**	**0.2284**						

2 Descriptive Statistics – continued

d. Standard Deviation

- *Standard Deviation (SD), denoted by σ (sigma). is widely used in business, production, quality control techniques, etc.*
- *In mean deviation we take the sum of deviations after ignoring their plus and minus signs.*
- *In standard deviation, we square up all the deviations and take their square root of their arithmetic mean.*
- *For a set of observations* $x_1, x_2, x_3, \ldots x_n$

  $\sigma = \sqrt{\Sigma (xi - \overline{x})^2 / n}$
- *For a set of observations* $x_1, x_2, x_3, \ldots x_n$ *with frequencies* $f_1, f_2, \ldots f_n$

  $\sigma = \sqrt{\Sigma f_i (xi - \overline{x})^2 / N}$; $N = \Sigma f_i$

2 Descriptive Statistics – continued

2.042 Relative Measures

a. Coefficient of Variation

- Coefficient of variation (CV) or relative measure is defined as $CV = (\sigma / \overline{x}) * 100$
- In order to decide which of the two distributions is more variable, we compare the coefficient of variation.
- The distribution with greater CV is said to be more variable.
- Consider the scores by batsmen A & B in 12 matches..

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
											n =	12	Σ	AM	σ	CV	
A	74	75	78	72	78	77	79	81	79	76	72	71	912	76		$CV = (\sigma / \overline{x}) * 100$	
B	87	84	80	88	89	85	86	82	82	79	86	80	1008	84			
A - Mean	-2	-1	2	-4	2	1	3	5	3	0	-4	-5					
Sq(A- Mean)	4	1	4	16	4	1	9	25	9	0	16	25	114	SD=sq rt(114/12)	3.082207	4.06	
B - Mean	3	0	-4	4	5	1	2	-2	-2	-5	2	-4					
Sq(B- Mean)	9	0	16	16	25	1	4	4	4	25	4	16	124	SD=sq rt(124/12)	3.21455	3.83	

2 Descriptive Statistics – continued

b. Measure of Skewness

- Average measures the central point of distribution, but gives no information about the shape of the frequency curve.
- Measures of dispersion give some idea of the spread of the variable about its average.
- Both these measures do not study whether the distribution is symmetrical or not.
- Skewness is a measure to study this aspect of the statistical distribution.
- If a distribution is not symmetrical distribution, we say it is skewed.
- If a symmetrical distribution, the values of the variable equi-distant from their mean have equal frequencies.

2 Descriptive Statistics – continued

b. Measure of Skewness – continued

- In a perfectly symmetrical distribution mea n, median and mode coincide. If this is not the case, the distribution is said to be skewed.
- Pearson's coefficient of skewness = (Mean – Mode) / SD, if mode is well defined
- Pearson's coefficient of skewness = 3(Median – Mode) / SD, if mode is not well defined
- Bowley's coefficient of skewness = (Q_3 + Q_1 – @M) / (Q_3 – Q_1); whereQ_1,Q_3 are first and third quartiles and M is the median.

negative skew: The left tail is longer; the mass of the distribution is concentrated on the right of the figure.

positive skew: The right tail is longer; the mass of the distribution is concentrated on the left of the figure.

2 Descriptive Statistics – continued

b. Measure of Skewness – Example

Bowley's measure of skewness

	Payment of commission	Number of Salesman - f	Cum. Frequency cf
	100 - 120	4	4
	120 - 140	10	14
	140 - 160	16	30
	160 - 180	29	59
Q_1 Class	180 - 200	52	111
Median Class	200 - 220	80	191
Q_3 Class	220 - 240	42	233
	240 - 260	23	256
	260 - 280	17	273
	280 - 300	7	280

N = 280
N / 2 = 140
N / 4 = 70
3N / 4 = 210

Q1 = l1 + (((N/4) - m1)/f1)*c)
= 180 + ((70 - 59)/52)*20 = 184.23

Q2 = l2 + (((N/2) - m2)/f2)*c)
= 200 + ((140 -80)/111)*20 = 207.25

Q3 = l3 + (((3N/4) - m3)/f3)*c)
= 220 + ((210 -191)/42)*20 = 229.05

Bowley's coefficient of skewness = $(Q_3-Q_1-2M)/(Q_3-Q_1)$
= ((229.05 - 184.23 - 2*207.25))/ (229.05 - 184.23)
= -.03

There is a very small negative skewness! Distribution is almost symmetrical

2 Descriptive Statistics – continued

c. Moments

- Mean and the variance provide information on the location and variability (spread, dispersion) of a set of numbers; thus providing some information on the appearance of the distribution of the numbers.
- The mean and variance are the first two statistical moments, and the third and fourth moments also provide information on the shape of the distribution.

Types:

a. Moments about mean – Central moments
b. Moments about an arbitrary origin (Non-central /raw moments)

Definition: *Assume we have a set of observations $y_1, y_2, .. y_N$*

The rth population moment about mean is denote by μ_r is

$$\mu_r = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^r}{N} \quad \text{where } r=1, 2, \ldots$$

2 Descriptive Statistics – continued

c. Moments– continued

The rth population moment about mean is denote by μ_r is

$$\mu_r = \frac{\sum f_i \ (y_i - \bar{y})^r}{N}$$

where $r=1, 2, \ldots$ $\sum f_i = N$ *For a grouped distribution*

$$\mu_r' = \frac{\sum f_i \ (y_i - A)^r}{N}$$

where $r=1, 2, \ldots$ $\sum f_i = N$

A is the assumed mean

c1. Relationship between Raw and Central Moments

Note that the first moment is zero $\mu_1 = 0.$

The second, third and fourth central moments can be expressed in terms of the raw moments as follows:

$$\mu_2 = \mu_2' - \mu^2$$

$$\mu_3 = \mu_3' - 3\mu\mu_2' + 2\mu^3$$

$$\mu_4 = \mu_4' - 4\mu\mu_3' + 6\mu^2\mu_2' - 3\mu^4$$

2 Descriptive Statistics – continued

d. Kurtosis

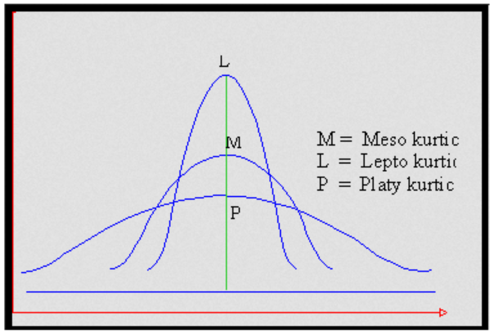
- It is a measure of platness or peakedness of a distribution.

Kurtosis can be formally defined as the standardized fourth population moment about the mean,

$\beta_2 = \frac{\mu_4}{\sigma^4}$ where μ is the mean,

μ_4 is the fourth moment about the mean, and

σ is the standard deviation



- Excess Kurtosis = β_2 – 3
- Normal curve is Mesokurtic; (β_2= 3; Excess Kurtosis = 0)
- More pateopped than a Normal curve – Platykurtic ($\beta_2 < 3$; Excess Kurtosis is negative)
- More peaked than a Normal curve – Leptokurtic ($\beta_2 > 3$; Excess Kurtosis is positive)

2 Descriptive Statistics – continued

d. Kurtosis – continued

Example: Find if the following distribution is symmetrical

										Σ		
x	0	1	2	3	4	5	6	7	8	36		
f	1	8	28	56	70	56	28	8	1	256		
fx	0	8	56	168	280	280	168	56	8	1024	$\bar{x}$	4
$d = x - \bar{x}$	-4	-3	-2	-1	0	1	2	3	4	0	N	256
fd	-4	-24	-56	-56	0	56	56	24	4	0		
fd^2	16	72	112	56	0	56	112	72	16	512		
fd^3	-64	-216	-224	-56	0	56	224	216	64	0		
fd^4	256	648	448	56	0	56	448	648	256	2816		

$$\mu_r = \frac{\sum f_i d^r}{N}$$

$$\sum f_i = N$$

$\mu_1' = 0$

$\mu_2 = 512 / 256 = 2$

$\mu_3 = 0 / 256 = 0$

$\mu_4 = 2816 / 256 = 11$

$$\beta_2 \; \frac{\mu_4}{\mu_2{}^2} = \frac{11}{2 * 2} = 2.8$$

Given distribution is almost mesokurtic