(P1)

Chapter 1 Introduction to data analysis

(I)

Basic type of data
- (i) qualitative data - open-ended questions
- (ii) quantitative data

(i) qualitative data:

e.g. (a) give reasons why you like or dislike this course.

(b) give comments on this book

(ii) quantitative data:

e.g. height, sbp, dbp, age ...

In a scientific research, concepts in which the research is interested are refered to as variables.

A variable is something that varies or take different values.

(II) Independent vs dependent variables

Independent variables are variables that predict (or have effects) dependent variables.

e.g. (a) Effect of smoking on incidence of Lung Cancer

(b) Effect of beta blockers on blood pressure

(c) Effect of diet & excercise on Cholesterol level

(III) Level of measurement for Numeric data

(a) Nominal measurement - using numbers simply as labels to classify attributes into different categories.

Variables that use nominal measurement refer to as categorical variable.

e.g. $$\text{religion} = \begin{cases} 1. \text{ Catholics} \\ 2. \text{ Protestant} \\ 3. \text{ Islam} \end{cases}$$

$$\text{gender} = \begin{cases} 1 & \text{male} \\ 2 & \text{female} \end{cases}$$

$$\text{province} = \begin{cases} 1 & \text{Nova Scotia} \\ 2 & \text{New Brunswick} \\ 3 & \text{Ontario} \\ 4 & \cdots \end{cases}$$

(b) Ordinal Measurement - use of numbers to designate ordering or an attribute.

e.g. education level $\begin{cases} 1. & \text{elementary school} \\ 2. & \text{high school} \\ 3. & \text{college/university} \\ 4. & \text{post graduate} \end{cases}$

but it does not tell us anything about the distance between categories.

e.g. statistical background : { 1. excellent, 2. very good, 3. neutral, 4. poor, 5. very poor

(c) Interval measurement - assigning numbers that indicate both the ordering of an attribute and the distance between different amounts of the attribute

e.g. temperature

It provides information about not only rank ordering but also magnitude of difference between different values on the scale.

(d) Ratio measurement : it is an interval measurement and also there is a presence of a natural, meaningful zero point.

e.g. dosage = { 0 No dosage, 1 mg etc..

length = 0 cm.

Interval measurement do NOT have a rational zero point. e.g. 0°C does NOT indicate total absence of heat.

(IV) Discrete & Continuous Data :

set 1 = {3, 5, 2, 4, 4, 8, 7, 6}

set 2 = {4.31, 2.22, 11.38, 12.07, 7.26}

(a) Discrete data :
First set of data consists of counting numbers or integers, we call it discrete data.

e.g. In a particular question

1	2	3	4	5
SA	Agree	Neutral	Disagree	SD

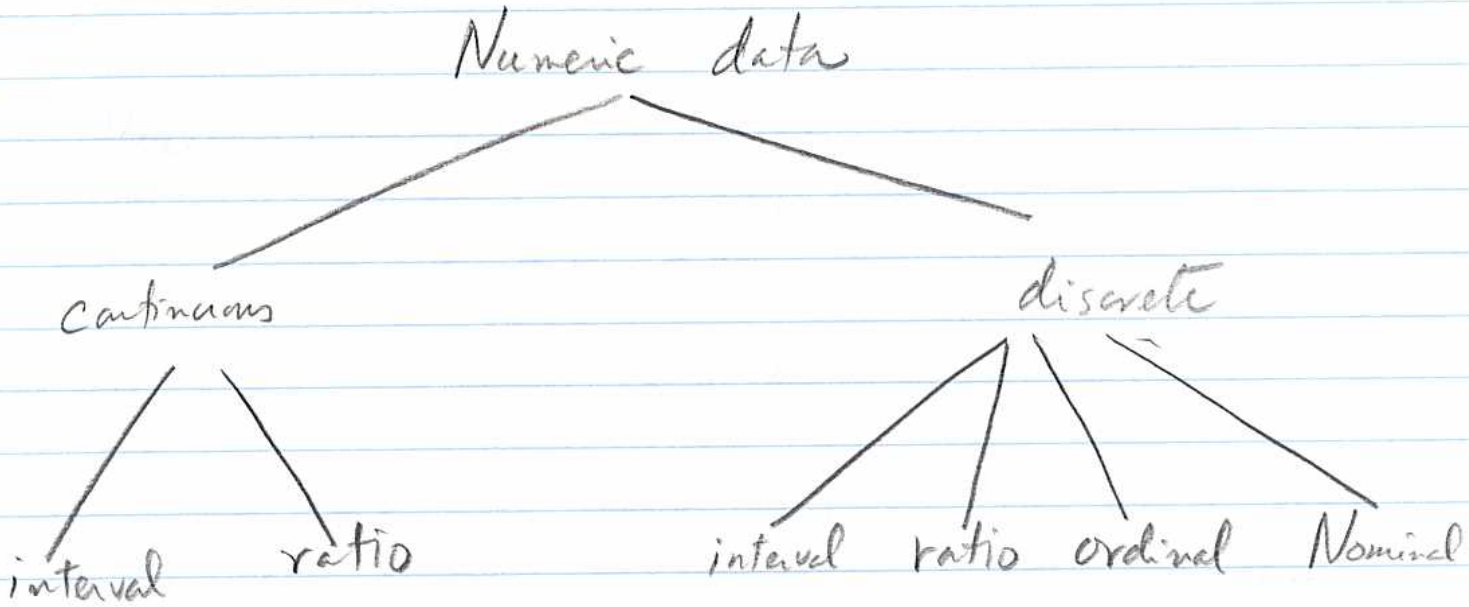
Note that discrete data may contain a decimal point
e.g. We throw a single dice twice and the average, we have = {1, 1.5, 2, ... 6}.

Nevertheless, such data have 'GAPS' in their possible values. e.g. # of beds in hospitals
# of children in your family

(b) Continuous data :
2nd set of data may represent the weights of 5 parcels in a postal office. Any value over some particular range is possible.
e.g. height data, weight data.

(P5)

Classification of numeric data:



(IV) The use of Data Analysis

Once the research variables have been measured, the resulting quantitative data can be analyzed in a variety of ways, and the analysis can serve many different purposes. The purposes can be categorized and illustrated along three dimensions.

(A) Descriptive statistics

These are used by researchers to describe and summarize data, help to make data readily comprehensible.

A wide variety of descriptive statistics is available to researchers. Averages and percentages are examples of descriptive statistics.

e.g. ① What percent of the sample is female?

Univariate statistics: Tabulating and Display Data

(I) Frequency Distributions

A systematic arrangement of data values - usually from lowest to highest. # of times each value was observed in the data (see SPSS output).

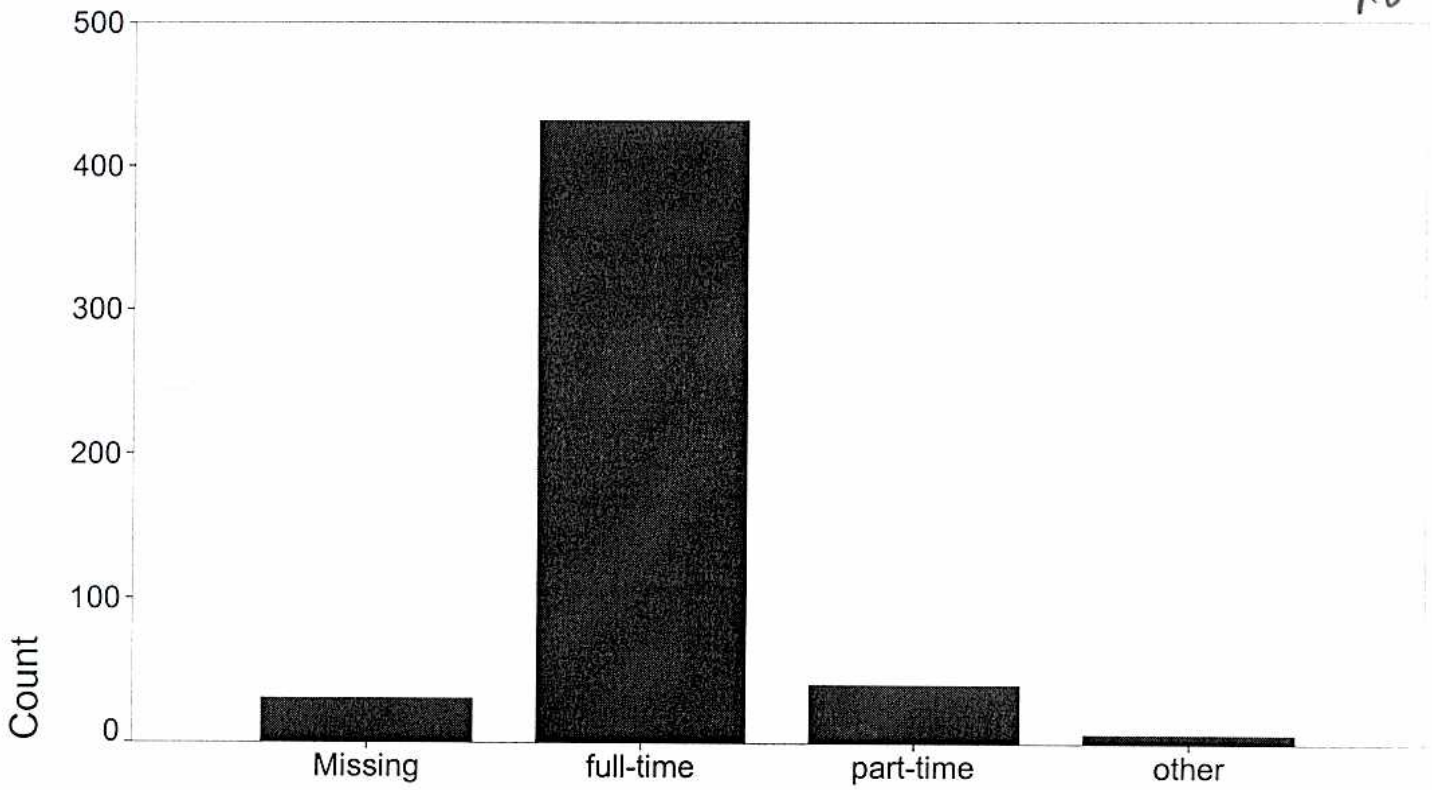
(II) Grouped frequency distributions

When a variable has many possible values, researchers often construct a grouped frequency distribution. Such a distribution involves grouping together values into sets, called class intervals. (see SPSS output).

(III) Graphic Displays of Frequency Distributions.

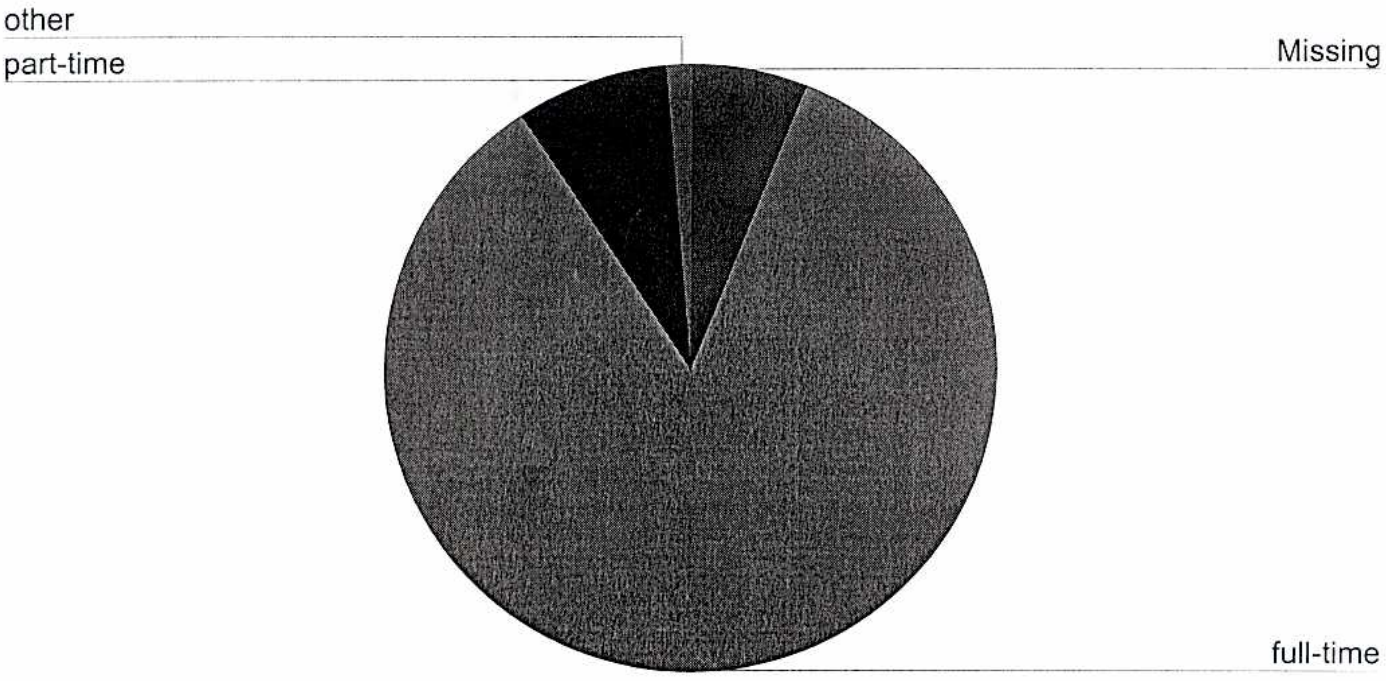
(a) Bar Graphs and Pie charts (see SPSS output)

When the variable of interest is measured on a nominal scale (and sometimes when it is measured on an ordinal scale with a limited number of values), researchers can construct a bar graph to display frequency information. An alternative to a bar graph is a pie chart which is a circle divided into pie-shaped wedges corresponds to the relevant percentages.

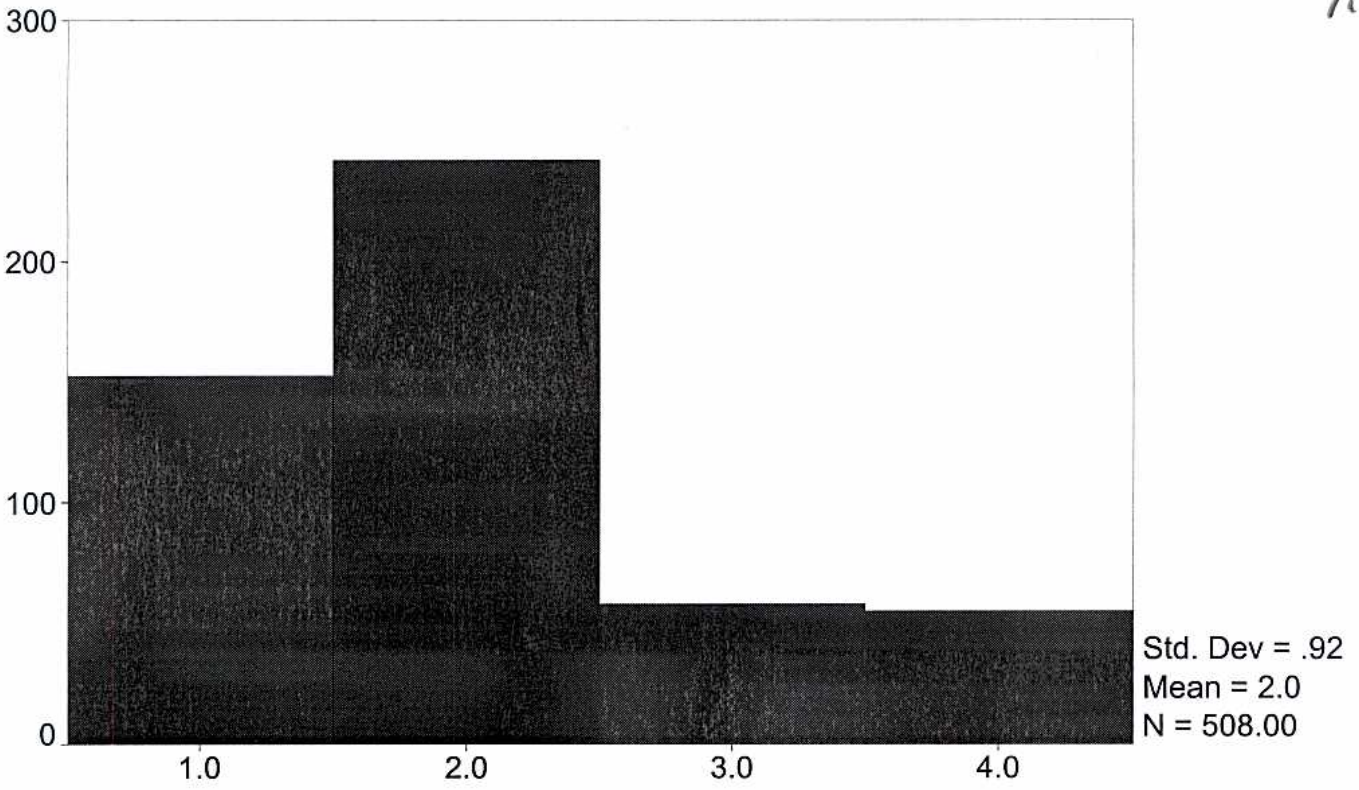


Are you a full-time or part-time student

Bar graph



Pie chart



Hours spent per day using a computer

Histogram

(b) Histograms (see spss output)

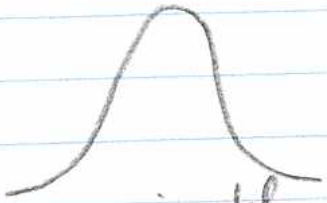
Frequency information for interval-level and ratio-level variables can be displayed in a histogram, which is a graphic display similar to a bar graph.

(IV) Shapes of Distributions

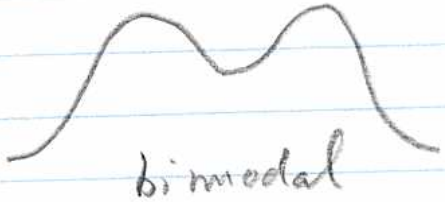
(a) Modality

Modality of a distribution concerns how many peaks or high points there are.
A distribution with a single high peak - that is, a single value with a high frequency - is refered to as unimodal distribution.

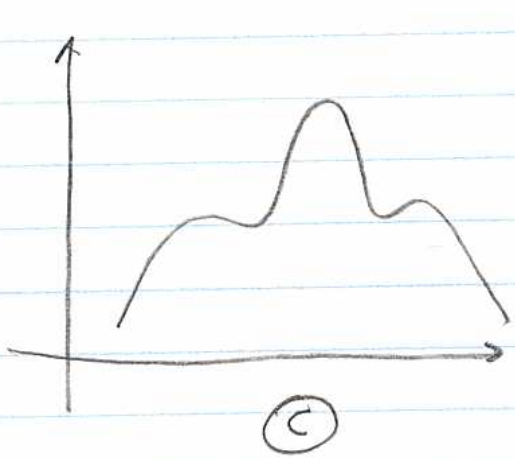
Some distributions have 2 or more peaks, are refered to as multimodal distributions. (When there are exactly 2 peaks, the distribution is called bimodal.



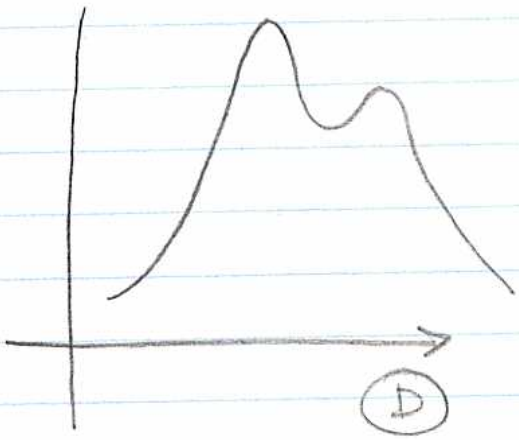
unimodal
(A)



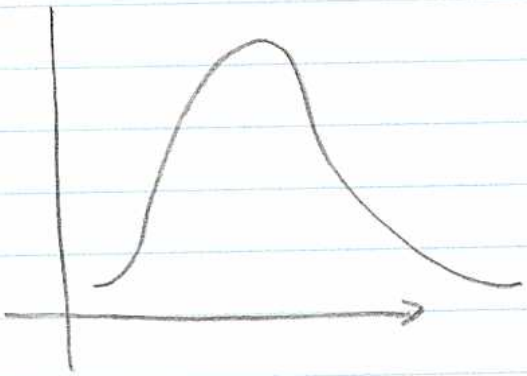
bimodal
(B)



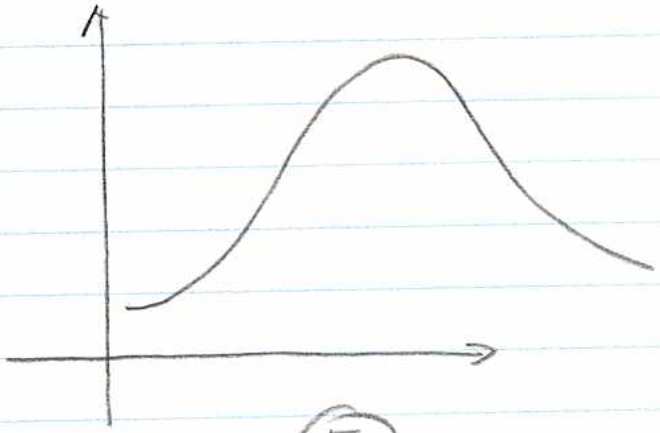
C



D



E



F

A, E, F unimodal

B D bimodal – exactly have two peaks.

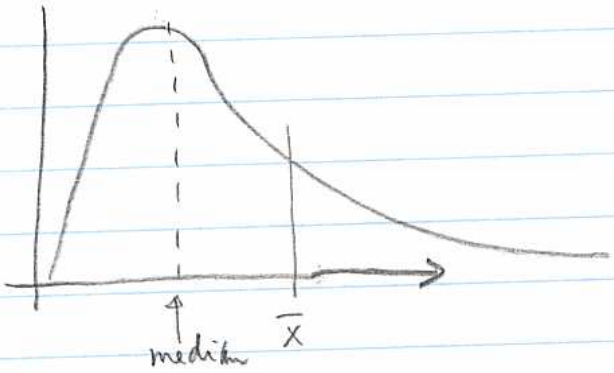
B C D multimodal

(b) Symmetry and Skewness

A distribution can be described as symmetric if the distribution could be split down the middle to form two halves that are mirror images of one another.

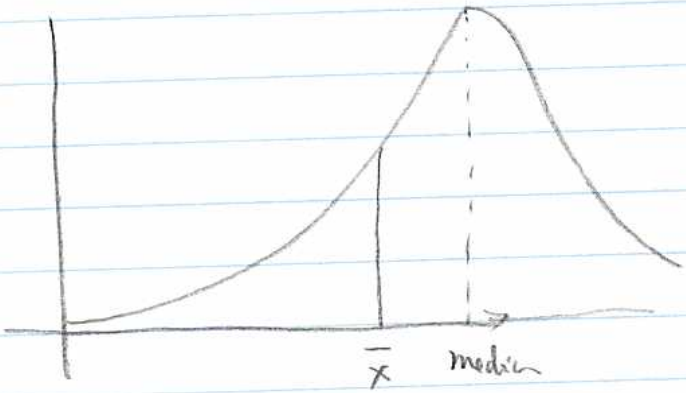
(A) (B) are symmetric.

$$\text{skewness} = 3\,\frac{(\bar{x} - \text{Median})}{s}$$



+vely skewed

skew to the right

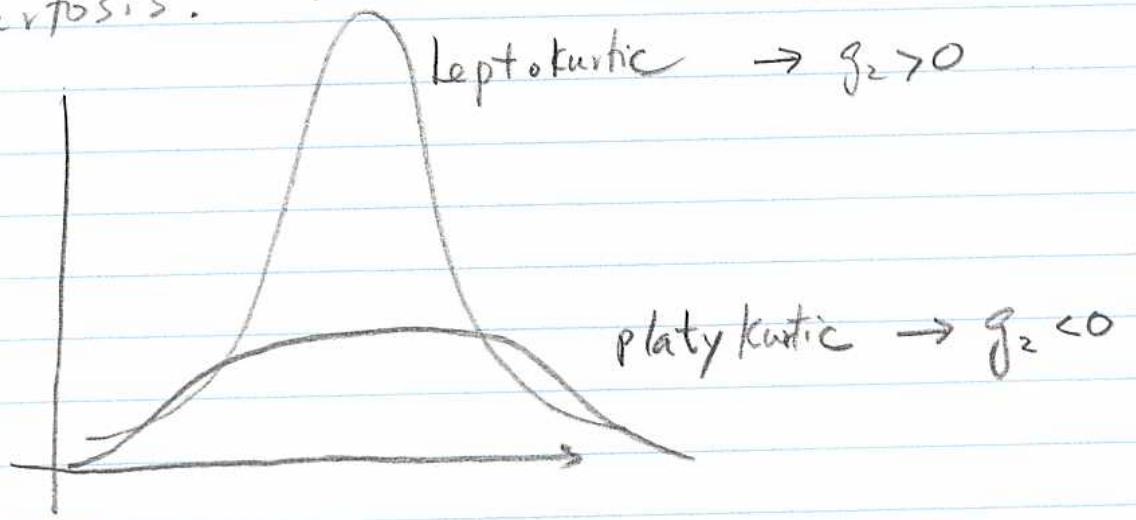


-ve skewed

skew to the left

(c) Kurtosis

A third aspect of a distribution's shape concerns how pointed or flat its peak is, which is referred to as the distribution's kurtosis.



The most acceptable measure of kurtosis is obtained from the second and fourth moments and is defined as

$$g_2 = \frac{m_4}{m_2^2} - 3$$

where $m_4 = \frac{\Sigma (X - \bar{X})^4}{N}$, $m_2 = \frac{\Sigma (X - \bar{X})^2}{N}$

When g_2 is 0, the distribution is a particular type of symmetrical distribution known as a normal distribution. When g_2 is less than 0, the distribution is flatter on top than the normal distribution. When g_2 is greater than 0, the distribution is more peaked than the normal distribution.