

1.1

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 referred 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 & exercise 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. religion = $\begin{cases} 1. \text{ Catholics} \\ 2. \text{ Protestant} \\ 3. \text{ Islam} \end{cases}$

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

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

(b) Ordinal Measurement - use of numbers to designate ordering on 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 = $\begin{cases} 0 & \text{No dosage} \\ 1 \text{ mg etc.} \end{cases}$

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 = $\{1, 1.5, 2, \dots, 6\}$.

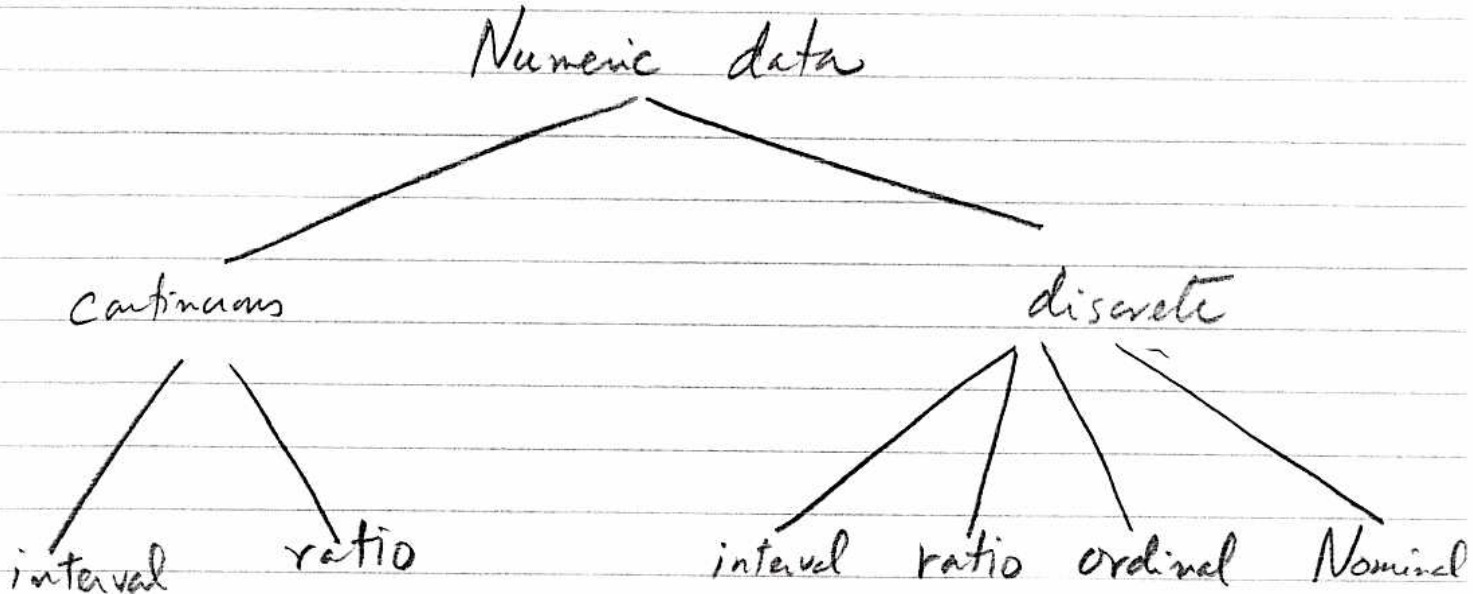
Nevertheless, such data have 'GAPS' in their possible values. e.g. # of beds in hospitals
of children in your family

(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.

Classification of numeric data:



④ 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.

① 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?

② What is the average age of our class?

(B) Inferential statistics.

When researchers use data from a sample to make inferences or draw conclusions about a population, they use inferential statistics.

(C) Univariate, Bivariate and Multivariate Statistics.

Univariate statistics involve one variable at a time
e.g. % of men and women in the sample

Bivariate statistics involve two variables examined simultaneously. If the researcher compared the average heart rate of men versus women.

e.g. relationship between blood pressure and salt intake.

Multivariate statistics involve three or more variables in the sample analysis.

e.g. We would like to use: gender, weight, age and amount of exercise to predict blood pressure.

1.7 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).

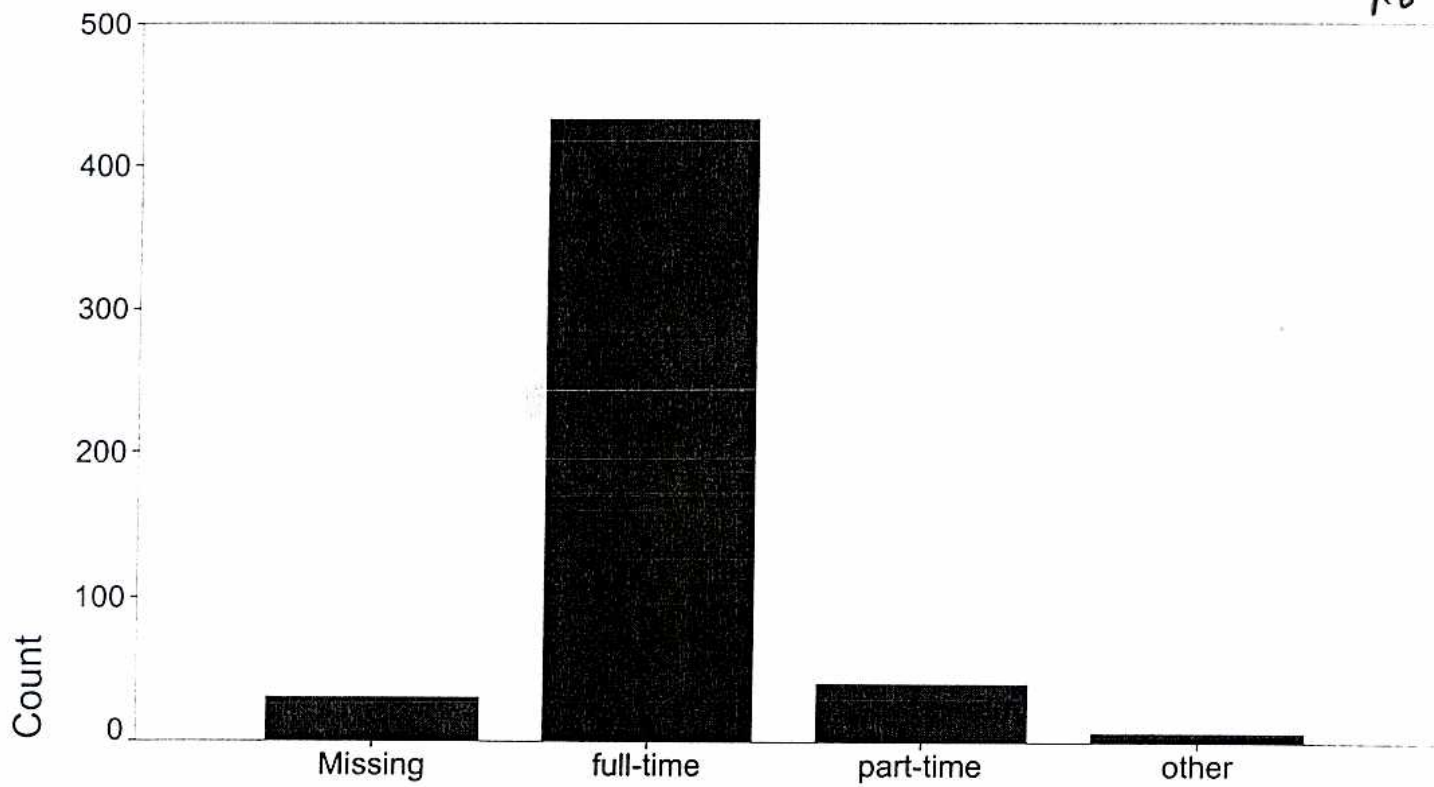
(II) Grouped frequency distributions

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.

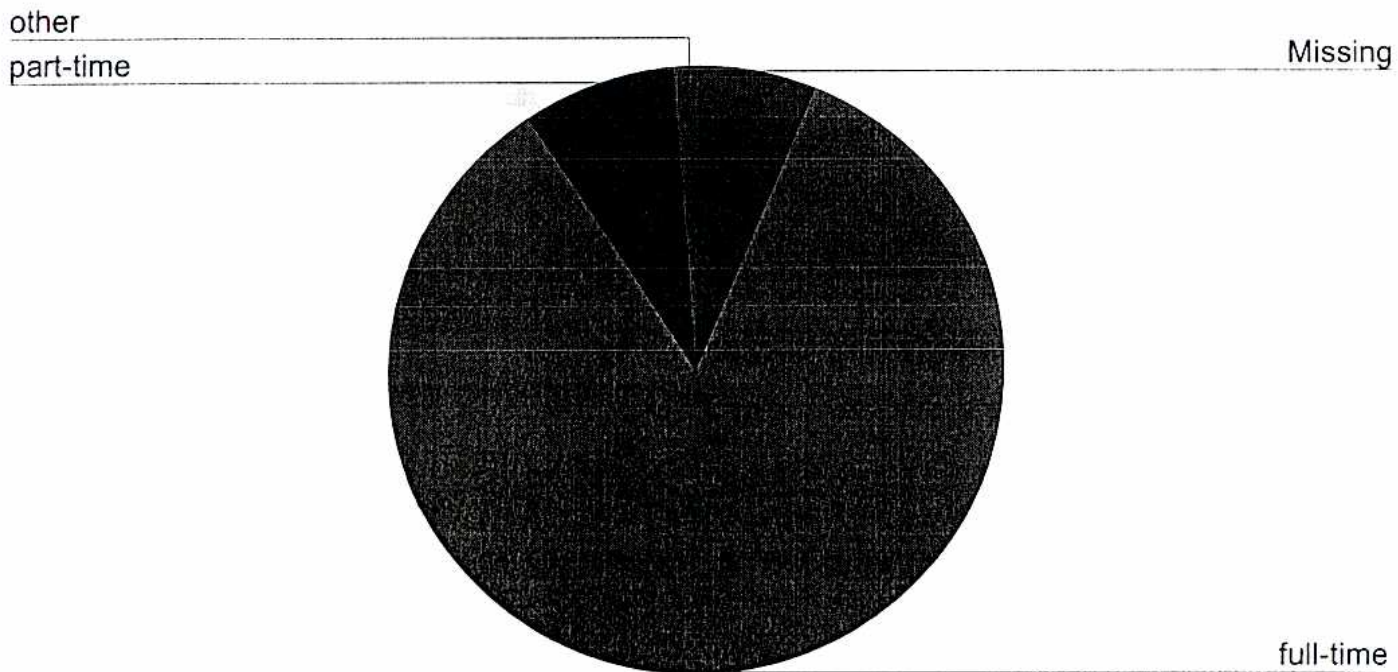
(a) Bar Graphs and Pie Charts (see SPSS output)

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 corresponding to the relevant percentages.



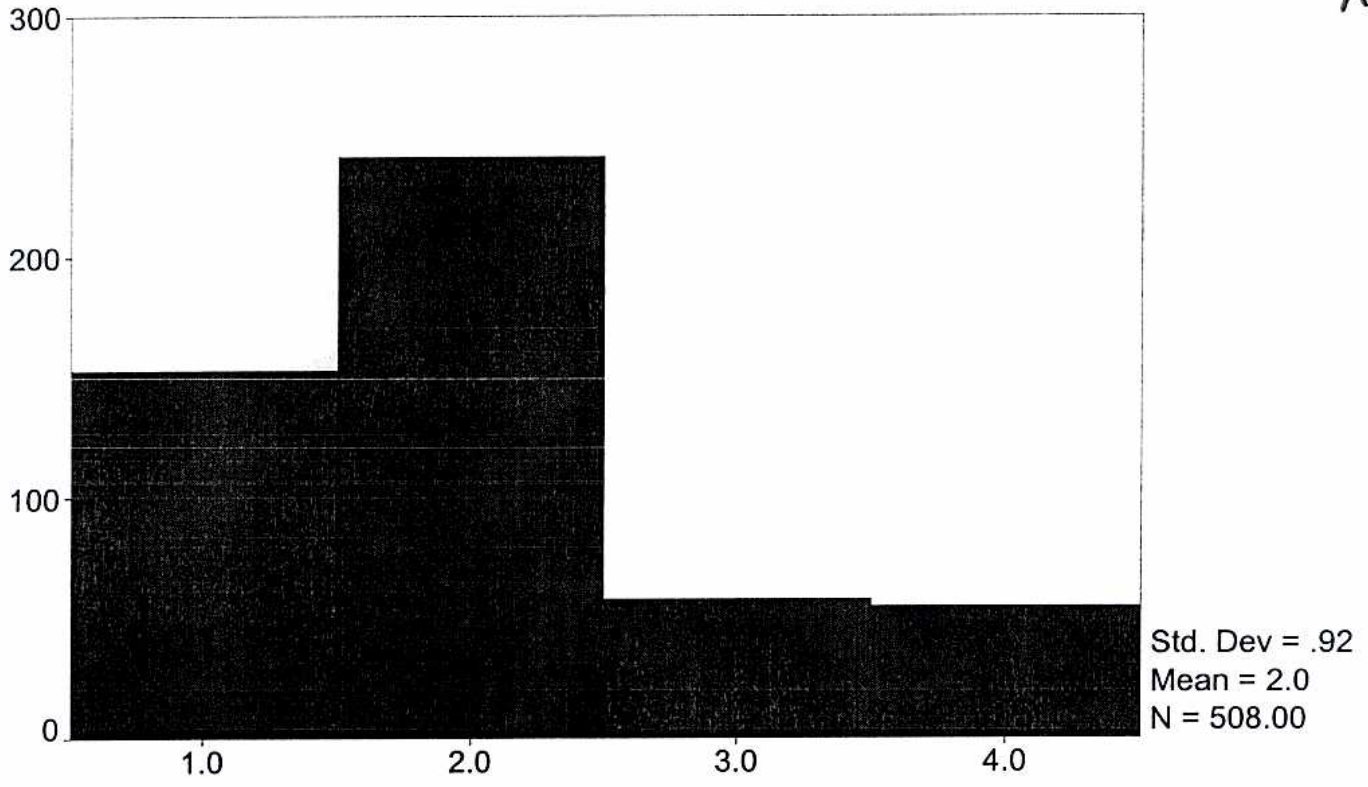
Are you a full-time or part-time student

Bar graph



Pie chart

1.9



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.

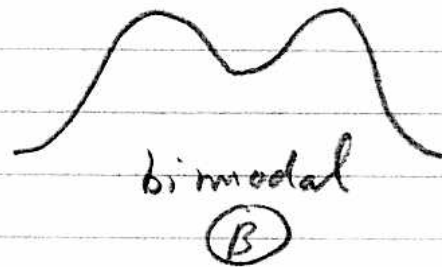
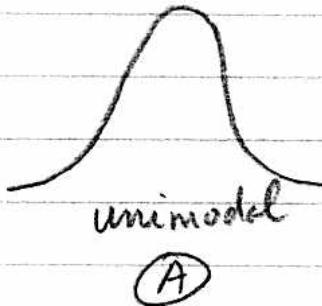
(IV) Shapes of Distributions

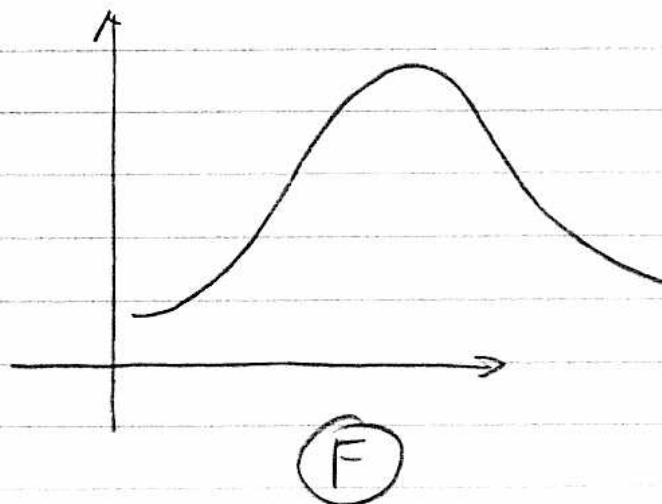
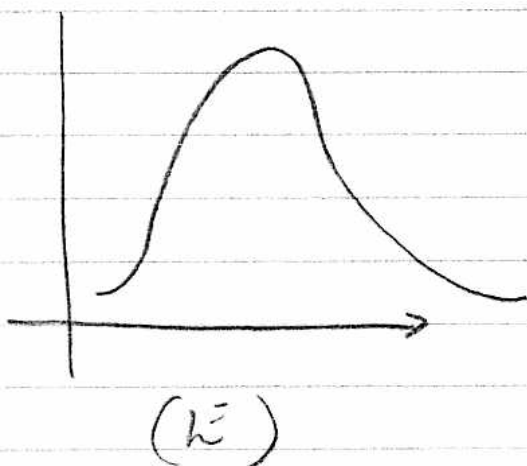
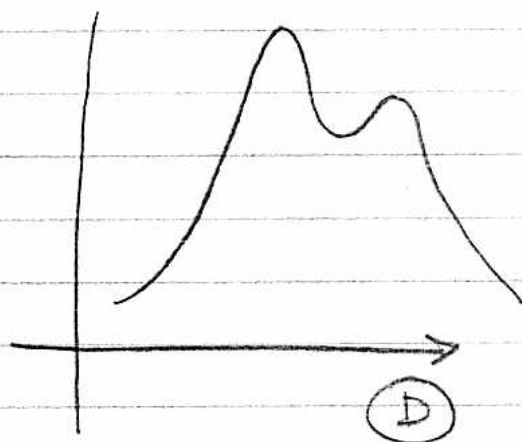
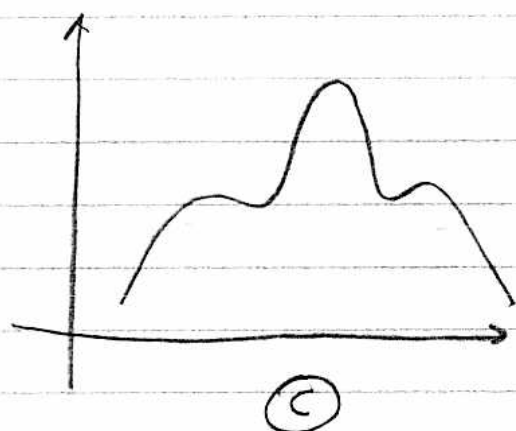
(a) Modality

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 referred to as unimodal distribution.

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





(A), (E), (F) unimodal

(B), (D) bimodal - exactly have two peaks.

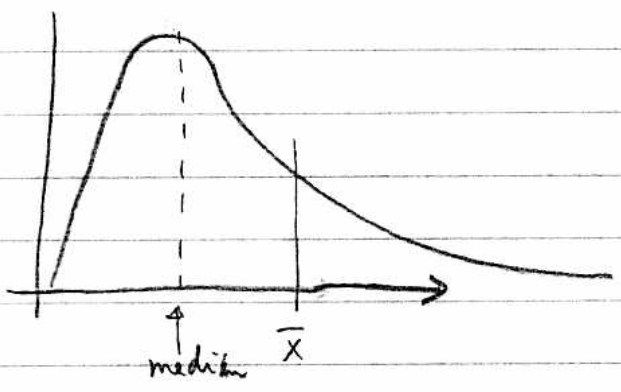
(B), (C), (D) multimodal

(b) Symmetry and Skewness

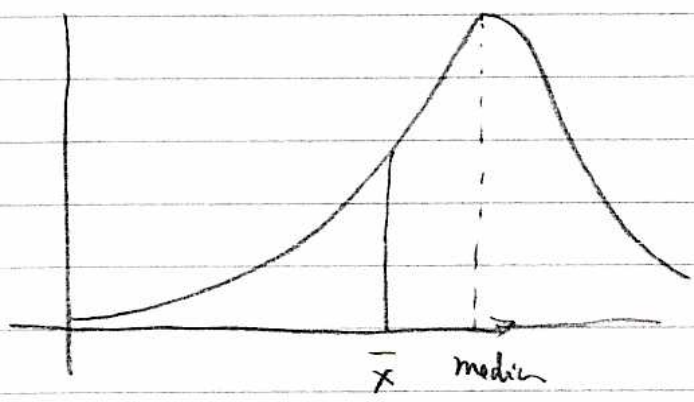
A distribution can be described as symmetric if the distribution could be split down the middle to form two halves that are mirror images of one another.

Ⓐ Ⓑ are symmetric.

$$\text{skewness} = 3 \frac{(\bar{X} - \text{Median})}{S}$$



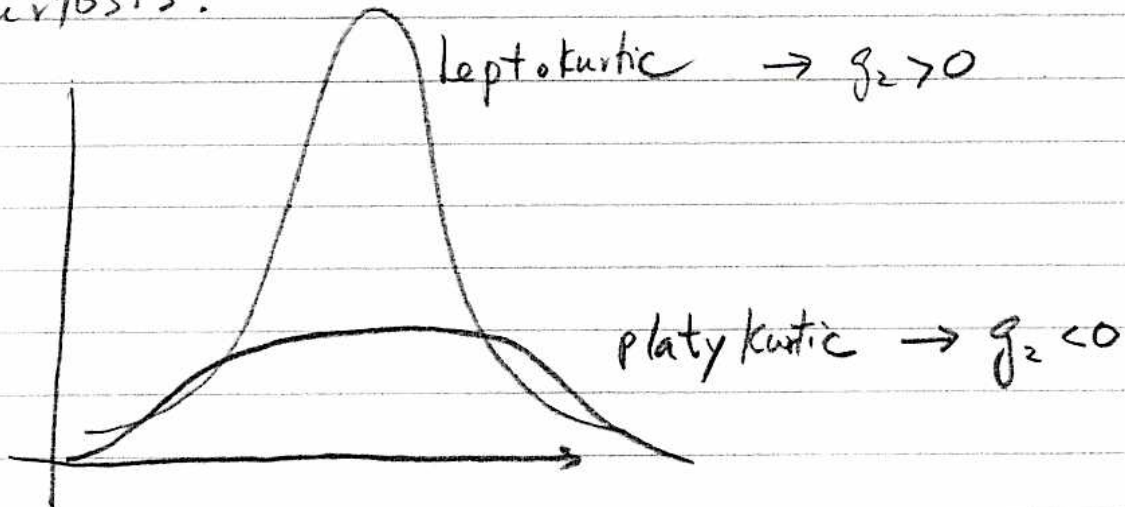
positively skewed skew to the right



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

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.

Chapter 2

Summary statistics. Measures of Location and Dispersion

Defⁿ: Parameters

Numbers that describes population characteristics are called parameters

Defⁿ: Estimate and Estimator

Values calculated from a sample of data that are used to estimate population parameters are called estimates. The formula used to calculate an estimate is called an estimator.

An estimator is a function, while an estimate is a specific value.

The mean, median and mode are three different numerical measures used to describe the centre of a data set. The range, the variance, the standard deviation, and the mean absolute deviation are four different measures used to measure the amount of spread, dispersion, or variability in a set of data.

2.1 The mean, the median and the mode

Defⁿ: Population mean

The population mean is denoted by μ and is calculated by using the following formula where N is the number of observations of the population

$$\mu = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\sum_{i=1}^N x_i}{N}$$

Defⁿ: Sample mean

Sample mean is denoted by \bar{x} . Given the sample of n observations x_1, x_2, \dots, x_n the sample mean \bar{x} is

$$\bar{x} = \frac{x_1 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

Defⁿ: Median

The median is the middle value of data ordered from lowest to highest. If the sequence contains an odd number of observations, the median is the middle value in the ordered sequence. If the sequence contains an even number of observations, the median is the arithmetic average of the two central values.

When a population of data is available, we calculate the population median. The median calculated from a sample of data is called the sample median and is an estimate of the population median.

e.g. On a sample of 7 days, # of patients at a walk-in clinic

$$x_1=92, x_2=84, x_3=70, x_4=76, x_5=66, x_6=80, x_7=71$$

$$\text{sample mean } \bar{x} = \frac{\sum x_i}{7} = \frac{539}{7} = 77$$

Order the observations from lowest to highest

$$66 \quad 70 \quad 71 \quad 76 \quad 80 \quad 84 \quad 92 \quad \text{low}$$

obs. = 7 which is odd, so the median is the middle or fourth value in ordered sequence.

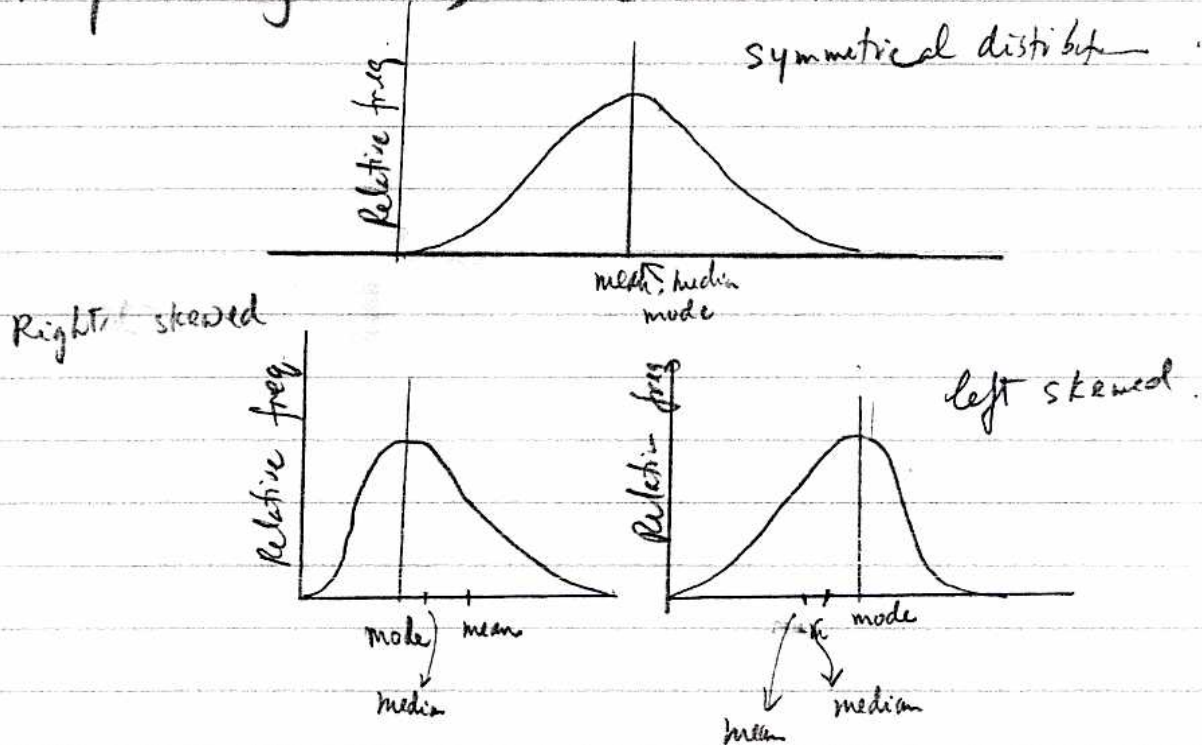
Sample median = 76

Defⁿ: Mode

The mode of a set of observations is the value that occurs with the greatest frequency. The mode is NOT necessarily unique.

e.g. Our walk-in clinic example, each of the values occurs exactly once. Because each value has the same frequency, each value could be a mode. In this problem, the mode conveys NO useful information.

Relationship among mean, mode and median



When a distribution is unimodal and symmetric like the bell-shaped normal distribution, the mean, median and mode all coincide. If the distribution is skewed to the right, the mean is larger ^{than} the mode because it is affected by the large values in the tail, whereas

The mode is influenced only by the most frequently occurring values, which are small. The median is less influenced by the right-hand tail than is the mean, so the median will be less than the mean. Furthermore, the peak of a right-skewed distribution will be left of the value that divides the area under the curve into two equal parts, and so the median will exceed the mode.

- i.e. (a) skewed to the right:

$$\text{mode} < \text{median} < \text{mean}$$
- (b) skewed to the left:

$$\text{mean} < \text{median} < \text{mode}.$$

Calculate the mean of Grouped Data

Let k denote # of groups or classes, let M_1, M_2, \dots, M_k denote the midpoints of the k classes. Let f_1, f_2, \dots, f_k denote the frequencies of the k classes. The mean is estimated using the formula

$$\bar{x} = \frac{\sum f_i M_i}{n} = \frac{f_1 M_1 + \dots + f_k M_k}{n}$$

Calculate the median of Grouped Data

- ① Determine which class contains the median
- ② Denote the lower boundary of this class by A & upper boundary by B
- ③ Let c denote # of observations less than A , and let d denote # of observations between A and B
- ④ Let n denote the sample size

If the data are in frequency distribution, estimate the median using the following formula:

Estimate the Median from a Frequency Distribution

$$\text{Median} = A + \frac{n/2 - c}{d} (B - A)$$

e.g.

Monthly Salary	Frequency (f_i)	Midpoint (M_i)	Product ($f_i M_i$)
1300 - under 1400	8	1350	10,800
1400 - under 1500	11	1450	15,950
1500 - under 1600	17	1550	26,350
1600 - under 1700	19	1650	31,350
1700 - under 1800	10	1750	17,500
1800 - under 1900	6	1850	11,100
1900 - under 2000	4	1950	7,800
2000 - under 2100	5	2050	10,250
TOTAL	80	1,0000	131,100

$$\sum f_i M_i = 131,100,$$

$$\text{The estimated mean } \bar{x} = \frac{\sum f_i M_i}{n} = \frac{131,100}{80} = 1638.75$$

$$\text{Median} = 1600 + \frac{\frac{80}{2} - 36}{19} (1700 - 1600)$$

$$= 1600 + 0.2105 \times 100 = 1621.05$$

Measure of dispersion

Summary statistics that measured the amount of variation in a data set are called measure of dispersion

(a) Range

The range of a set of observations is the difference between the largest & the smallest value.

(b) Interquartile Range

For any population or sample of values, the interquartile range is $Q_3 - Q_1$.

Q_3 : 75th percentile (upper Quartile)

Q_1 : 25th " " (lower Quartile).

(c) Deviation from the Mean.

For any value x , the deviation from the mean is difference $(x - \mu)$ if the population mean is known or $(x - \bar{x})$ if the sample mean is used.

(d) Mean absolute deviation

$$M.A.D. = \frac{\sum |x_i - \bar{x}|}{n}$$

e.g. On a sample of 7 days, # of patients in a walk in clinic:

92, 84, 70, 76, 66, 80, 71

$$\bar{x} = 77$$

$$M.A.D. = \frac{|92-77| + |84-77| + |70-77| + |76-77| + |66-77| + |80-77| + |71-77|}{7}$$

$$= 7.143$$

(e) Variance and Standard Deviation

Defⁿ: Population variance is calculated using

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}, \quad \begin{array}{l} \mu: \text{population mean} \\ N: \text{population size} \end{array}$$

Defⁿ: sample variance is calculated using

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad \begin{array}{l} \text{where } \bar{x}: \text{sample mean} \\ n: \text{sample size} \end{array}$$

$$= \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

The use of $n-1$ rather than n in the denominator of s^2 should be explained. Recall that \bar{x} and s^2 are calculated from sample data and are estimates of the population parameters μ and σ^2 . In general, the sample will not exactly reproduce the characteristics of the population, and it is unlikely that \bar{x} will be exactly equal to μ .

When estimating a population parameter, we want to use a procedure that has no tendency to over- or underestimate the population value on the average.

If an estimator equals the population value on the average (i.e. overestimate & underestimate balance out), then we have an unbiased estimator.

s^2 has $n-1$ rather than n as its denominator so that it will be unbiased estimator of σ^2 .

If n were used as the denominator, on the average s^2 would be smaller than σ^2 and thus be a biased estimator. The reason for this bias

is that, on the average, the sample squared deviations $(x_i - \bar{x})^2$ are slightly smaller than the population squared deviations $(x_i - \mu)^2$. The use of $(n-1)$ in the denominator corrects this bias. The number $(n-1)$ is sometimes called degree of freedom.

e.g.

x	$x - \bar{x}$	$(x - \bar{x})^2$
92	$92 - 77 = 15$	225
84	$84 - 77 = 7$	49
70	$70 - 77 = -7$	49
76	$76 - 77 = -1$	1
66	$66 - 77 = -11$	121
80	$80 - 77 = 3$	9
71	$71 - 77 = -6$	36
Total 539	0	490

sample
variance, $S^2 = \frac{490}{6} = 81.667$,

sample
s.d. $S = \sqrt{81.667} = 9.037$

Calculate the Standard Deviation using Grouped Data

Let k denote # of classes in the frequency distribution;
let M_1, M_2, \dots, M_k denote k class midpoints;
& let f_1, f_2, \dots, f_k denote k class frequencies.

The sample variance is

$$S^2 = \frac{\sum f_i (M_i - \bar{x})^2}{n-1} = \frac{\sum f_i M_i^2 - n\bar{x}^2}{n-1}$$

The population variance σ^2 for grouped data

$$\sigma^2 = \frac{\sum f_i (M_i - \mu)^2}{N} = \frac{\sum f_i M_i^2 - N\mu^2}{N}$$

e.g.

# of cars owned	frequency	$f_i M_i$	$f_i M_i^2$
0	9	0	0
1	39	39	39
2	48	96	192
3	15	45	135
4	6	24	96
5	3	15	75
TOTAL	120	219	537

$\bar{x} = \frac{219}{120} = 1.825$

$$S^2 = \frac{\sum f_i M_i^2 - n \bar{x}^2}{n-1} = \frac{537 - 120(1.825)^2}{119} = 1.154$$

$$S = \sqrt{1.154} = 1.074$$

e.g.

Monthly Salary	Frequency f_i	Mid point (M_i)	Product $f_i M_i$	$f_i M_i^2$
1300 - under 1400	8	1350	10800	14580 000
1400 - under 1500	11	1450	15950	23127 500
1500 - under 1600	17	1550	26350	40842 500
1600 - under 1700	19	1650	31350	51727 500
1700 - under 1800	10	1750	17500	30625 000
1800 - under 1900	6	1850	11100	20535 000
1900 - under 2000	4	1950	7800	15210 000
2000 - under 2100	5	2050	10250	21012 500
TOTAL	80		131,100	217,660,000

$$S^2 = \frac{\sum f_i (M_i)^2 - n(\bar{x})^2}{n-1} = \frac{217660000 - 80 \times (1638.75)^2}{79}$$

$$= \frac{217660000 - 214840125}{79} = 35694.62$$

$$S = 188.93$$

Quartiles and Percentiles:

Definition: Percentiles

Suppose the observations x_1, x_2, \dots, x_n , have been arranged in ascending order. The p th percentile is the value x_p such that p percent of the observations are less than or equal to x_p and $(1-p)$ percent of the observations are greater than or equal to x_p .

Lower quartile(or first quartile) is the 25th percentile of the data. i.e. 25% of the data below the first quartile and thus 75% above it. The first quartile is denoted by Q_1 .

Upper quartile(or third quartile) is the 75th percentile of the data. i.e. 75% of the data below the third quartile and thus 25% above it. The third quartile is denoted by Q_3 .

Median(or second quartile) is the 50th percentile of the data. i.e. 50% of the data below the median and 50% lie above it.

Procedure for Calculating Percentiles-Ungrouped Data:

1. Arrange the n observations in ascending order. The smallest observation is given rank 1, the second smallest observation is given rank 2, and so forth.
2. Calculate the index $i = (p \times n) / 100$, where p is the percentile of interest and n is the sample size.
3. If i is an integer, the p th percentile is the arithmetic average of the values having ranks i & $(i+1)$.
4. If i is not an integer, the next integer value greater than i denotes the rank of the value which is the p th percentile.

For example:

1. Calculate the 70th & 80th percentiles for the following 15 values:

18, 14, 45, 32, 65, 43, 25, 41, 83, 51, 26, 36, 40, 55, 20

Solution:

Arrange the 15 values in ascending order:

14, 18, 20, 25, 26, 32, 36, 40, 41, 43, 45, 51, 55, 65, 83

For the 70th percentiles, the index $I : (70 \times 15) / 100 = 10.5$

Because I is not an integer, the position of the 70th percentile is the next integer greater than 10.5, which is 11, that is, the 11th position. The 70th percentile corresponds to the 11th data value is 45.

For the 80th percentile, the index $I : (80 \times 15) / 100 = 12.0$

Because I is an integer, the 80th percentile is the average of the data values in positions 12 & 13. Thus the 80th percentile is $(51 + 55) / 2 = 53$

2. The following 24 observations have been ordered from lowest to highest:

6	8	12	17	20	22	25	26	26	28	30
32	34	37	40	50	61	63	65	67	69	80
82	86									

Find the first and third quartiles.

Solution:

For the lower quartile, calculate the index with $p=25$ and $n=24$. Using the formula $I=(p \times n)/100$, we get $I=6.0$. Because I is an integer, Q_1 is the average of observations having ranks 6 and 7. Because the 6th observation is 22 and the 7th observation is 25, the lower quartile is $Q_1=(22+25)/2=23.5$.

For the upper quartile, calculate the index with $p=75$ and $n=24$, we have $I=18.0$. Because I is an integer, Q_3 is the average of observations having ranks 18 and 19. The 18th observation is 63 and 19th observation is 65. The upper quartile Q_3 is $(63+65)/2=64.0$.

Procedure for Calculating Percentiles-Grouped Data:

If the data have been grouped as a frequency distribution, the p th percentile can only be approximated because the individual observations are not known. The p th percentile is approximated by using the following procedure:

1. Find the interval that contains the p th percentile.
2. Let A and B denote the lower and upper boundaries of this interval.
3. Let c denote the number of observations less than A .
4. Let d denote the number of observations between A and B .
5. Let n denote the sample size and let p denote the percentile desired. Let $m=n \times p/100$.
6. The p th percentile can be approximated as the value

$$P_{th} \text{ percentile} = A + (m-c)(B-A)/d$$

The formula is the result of interpolation between points A and B .

There are several methods of calculating percentiles under the Procedure-Examine of SPSS-please see my SPSS handouts. For the sake of simplicity & popularity, we are using **empirical method**.

Example:

1. Frequency distribution of data:

Monthly Salary	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
1300 – under 1400	8	.1000	8	.1000
1400-under 1500	11	.1375	19	.2375
1500-under 1600	17	.2125	36	.4500
1600-under 1700	19	.2375	55	.6875
1700-under 1800	10	.1250	65	.8125
1800-under 1900	6	.0750	71	.8875
1900-under 2000	4	.0500	75	.9375
2000-under 2100	5	.0625	80	1.0000

Total 80 1.000

Estimate the lower quartile.

Solution: To find the class containing the lower quartile, examine the cumulative relative frequencies in the table. The lower quartile must fall in the 1500-1600 interval, because 23.75% of the observations are less than 1500 and 45% of the observations are less than 1600. Thus, we have $A=1500$, $B=1600$, $m=80$ * $25/100=20$, $c=19$, and $d=17$. The lower quartile is then estimated as:

$$Q1=A+(m-c)*(B-A)/d=1500+(20-19)*(1600-1500)/17=1505.88$$

The Stem-and-Leaf Plot:

Whenever we are presented with a set of numerical observations, it is natural for us to want to condense or simplify the data in some way so that we can recognize general characteristics of the data. For example, most people would like to know the following things:

1. How are the data distributed? What is the general shape of the distribution? Where are most of the observations concentrated?
2. Where is the center of the data?
3. How spread out, or dispersed, are the data about this central value?
4. Are there any very large or very small values that appear to be quite different from the other observations?

The most frequently used technique for answering these questions is to construct a frequency or relative frequency distribution and to calculate some summary statistics. Another closely related way of answering these questions is to construct what is called a **stem-and-leaf diagram**, a technique proposed by John Tukey. The purpose of the **stem-and-leaf diagram** are as follows:

1. To show the range of the data, that is, to make it easy to determine the difference between the largest observation and the smallest observation;
2. To show how the data are distributed;
3. To show in a general way where the center of the data is;
4. To show in a general way how spread out the data are; and
5. To show whether there are any extremely high or low values, called outliers.

Example: Construction of a Stem-and-leaf Diagram:

The data below show the ages of 60 employees in a company:

43	34	21	26	57	64	51	30	38	60
19	26	41	47	58	50	42	32	25	18
36	27	38	37	24	49	56	61	20	39
32	51	60	52	45	44	33	25	29	27
31	35	39	19	20	30	46	47	32	24
29	22	20	37	34	29	60	55	30	41

In a stem-and-leaf diagram, the first or leading digit serves as the stem and the trailing digit serves as the leaf. The leading digit determines the row in which an observation is placed; the trailing digit is written in this row to the right of a vertical bar that separates the stems from the leaves. In this example, the stems are the first digits for each person's age; thus the stems are the digits 1 through 6. For example, for a person aged 53, the stem is 5 and the leaf is 3. To improve the appearance of the diagram, it is useful to arrange the leaves in each row from lowest to highest. The stem-and-leaf diagram for the age data above is shown:

1	8 9 9
2	0 0 0 1 2 4 4 5 5 6 6 7 7 9 9 9
3	0 0 0 1 2 2 2 3 4 4 5 6 7 7 8 8 9 9
4	1 1 2 3 4 5 6 7 7 9
5	0 1 1 2 5 6 7 8
6	0 0 0 1 4

The stem-and-leaf diagram allows us to pick out several important characteristics of the data. In this example, the smallest observation is 18 and the largest is 64. There are 18 leaves follow the 3 stem and 16 leaves following the 2 stem, indicates that 18 people were in their thirties and 16 employees were in their twenties. These are biggest classes. Few people under 20 or over 60 are hired by this company.

Advantages : it allows you to reconstruct the original data set and it lists the observations in order of magnitude.

Disadvantages: suitable for relatively small data sets. Little flexibility in the choice of stems and thus in the number of classes.

Stem-and-leaf plot from SPSS :

AGE

Frequency	Stem &	Leaf
3.00	1 .	899
16.00	2 .	0001244556677999
18.00	3 .	000122234456778899
10.00	4 .	1123456779
8.00	5 .	01125678
5.00	6 .	00014
Stem width:	10.00	
Each leaf:	1 case(s)	

The Boxplot:

Both the histogram and the stem-and-leaf plot provide useful information about the distribution of observed values. We can see how tightly cases cluster together. We can see if there is a single peak or several peaks. We can determine if there are extreme values.

A display that further summarizes information about the distribution of the values is the boxplot. Instead of plotting the actual values, a boxplot displays summary statistics for the distribution. It plots the median, the 25th percentile, the 75th percentile, and values that are far removed from the rest. Following is an example of a boxplot from the lower boundary of the box is the 25th percentile, and the upper boundary is 75th percentile. A line in the box represents the median. Fifty percent of the cases have values within the box. The length of the box corresponds to the interquartile range (IQR), which is the difference between the 75th and 25th percentile.

The next step is to decide how far away from the median a measurement must lie before it is classified as an outlier. The decision is made by constructing two sets of fences. The **inner fences** are located 1.5 IQR below Q1 and 1.5 IQR above Q3. The **inner fences** are defined as follows:

Lower inner fence = $Q1 - 1.5(IQR)$

Upper inner fence = $Q3 + 1.5(IQR)$

The second set of fences, the **outer fences**, are located 3 IQR below Q1 and 3 IQR above Q3:

Lower outer fence = $Q1 - 3(IQR)$

Upper outer fence = $Q3 + 3(IQR)$

Values lying between an inner fence and the corresponding outer fence are considered to be **'mild'** outliers; values outside the outer fences are considered to be **'extreme'** outliers.

Example:

Here is a set of data on ages of 50 patients:

39 37 32 40 40 37 39 35 42 43
 40 41 31 38 38 37 24 34 42 47
 40 51 33 37 36 29 39 44 49 37
 34 37 39 38 35 36 43 45 38 42
 41 41 56 40 42 48 37 42 39 41

```

-> get file='d:\stat601.14\pl33.sav'.
-> examine variables=age/percentiles(25,50,75)=aempirical/
-> plot=boxplot stemleaf.

```

AGE

Valid cases: 50.0 Missing cases: .0 Percent missing: .0

Mean	39.3000	Std Err	.7663	Min	24.0000	Skewness	.3059
Median	39.0000	Variance	29.3571	Max	56.0000	S E Skew	.3366
5% Trim	39.2222	Std Dev	5.4182	Range	32.0000	Kurtosis	2.0199
95% CI for Mean	(37.7602, 40.8398)			IQR	5.0000	S E Kurt	.6619

Frequency	Stem &	Leaf
2.00	Extremes	(24), (29)
1.00	3 *	1
2.00	3 t	23
4.00	3 f	4455
9.00	3 s	667777777
9.00	3 .	888899999
9.00	4 *	000001111
7.00	4 t	2222233
2.00	4 f	45
1.00	4 s	7
2.00	4 .	89
2.00	Extremes	(51), (56)

Stem width: 10.00
Each leaf: 1 case(s)

Percentiles

Percentiles	25.0000	50.0000	75.0000
Aempirical	37.0000	39.0000	42.0000
Tukey's Hinges	37.0000	39.0000	42.0000

Hi-Res Chart # 7:Boxplot of age

$$\text{lower inner fence} = 37 - 1.5 \times 5 = 29.5$$

$$\text{upper inner fence} = 42 + 1.5 \times 5 = 49.5$$

$$\text{lower outer fence} = 37 - 3 \times 5 = 22$$

$$\text{upper outer fence} = 42 + 3 \times 5 = 57$$

Further explanation of Boxplot example(i.e. age of 50 patients):

OBS	AGE
1.00	39.00
2.00	37.00
3.00	32.00
4.00	40.00
5.00	40.00
6.00	37.00
7.00	39.00
8.00	35.00
9.00	42.00
10.00	43.00
11.00	40.00
12.00	41.00
13.00	31.00
14.00	38.00
15.00	38.00
16.00	37.00
17.00	24.00
18.00	34.00
19.00	42.00
20.00	47.00
21.00	40.00
22.00	51.00
23.00	33.00
24.00	37.00
25.00	36.00
26.00	29.00
27.00	39.00
28.00	44.00
29.00	49.00
30.00	37.00
31.00	34.00
32.00	37.00
33.00	39.00
34.00	38.00
35.00	35.00
36.00	36.00
37.00	43.00
38.00	45.00
39.00	38.00
40.00	42.00
41.00	41.00
42.00	41.00
43.00	56.00
44.00	40.00
45.00	42.00
46.00	48.00
47.00	37.00
48.00	42.00
49.00	39.00
50.00	41.00

Sort ages:

ORDER	OBS	AGE
1.00	17.00	24.00
2.00	26.00	29.00
3.00	13.00	31.00
4.00	3.00	32.00
5.00	23.00	33.00
6.00	18.00	34.00
7.00	31.00	34.00
8.00	8.00	35.00
9.00	35.00	35.00
10.00	25.00	36.00
11.00	36.00	36.00
12.00	2.00	37.00
13.00	6.00	37.00
14.00	16.00	37.00
15.00	24.00	37.00
16.00	30.00	37.00
17.00	32.00	37.00
18.00	47.00	37.00
19.00	14.00	38.00
20.00	15.00	38.00
21.00	34.00	38.00
22.00	39.00	38.00
23.00	1.00	39.00
24.00	7.00	39.00
25.00	27.00	39.00
26.00	33.00	39.00
27.00	49.00	39.00
28.00	4.00	40.00
29.00	5.00	40.00
30.00	11.00	40.00
31.00	21.00	40.00
32.00	44.00	40.00
33.00	12.00	41.00
34.00	41.00	41.00
35.00	42.00	41.00
36.00	50.00	41.00
37.00	9.00	42.00
38.00	19.00	42.00
39.00	40.00	42.00
40.00	45.00	42.00
41.00	48.00	42.00
42.00	10.00	43.00
43.00	37.00	43.00
44.00	28.00	44.00
45.00	38.00	45.00
46.00	20.00	47.00
47.00	46.00	48.00
48.00	29.00	49.00
49.00	22.00	51.00
50.00	43.00	56.00

→ 29.5 lower inner fence

→ 49.50 upper inner fence

