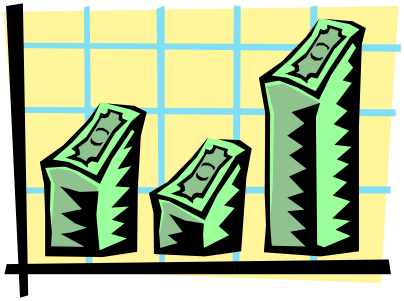
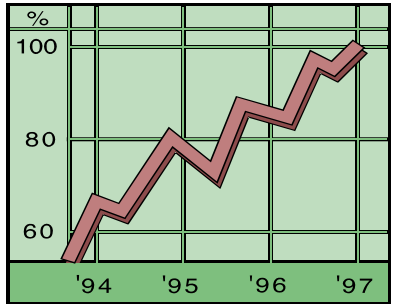


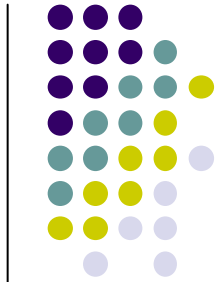


Introduction to Probability and Statistics

Slides 1 – Chapter 1

Prof. Ammar M. Sarhan,
asarhan@mathstat.dal.ca
Department of Mathematics and Statistics,
Dalhousie University
Fall Semester 2010

Course outline



- Overview & Descriptive Statistics
- Probability
- Discrete Random variables and Distributions
- Continuous Random variables and Distributions
- Joint Probability Distributions & Random Samples
- Point Estimation
- Statistical Intervals Based on a single Sample
- Tests of Hypotheses Based on a single Sample
- Inferences Based on Two Samples

Overview and Descriptive Statistics

- Introduction
- Populations and Samples
- Pictorial and Tabular Methods in Descriptive Statistics
- Measures of Location
- Measures of Variability

Overview

Statistical concepts and methods are not only useful but indeed indispensable in understanding the world around us.

They provide ways of gaining new insights into the behavior of many phenomena that we will encounter in our chosen fields of specializations in sciences or engineering.

The discipline of statistics teaches us how to make a right decision in the presence of *uncertainty* and *variation*.

> Without *uncertainty* or *variation*, there would be little need for statistical methods or statisticians. Example: If all students had the same level of ability to understand statistics, then a single observation would reveal all desired information.

How can statistical techniques be used to gather information and draw conclusion?

Descriptive Statistics

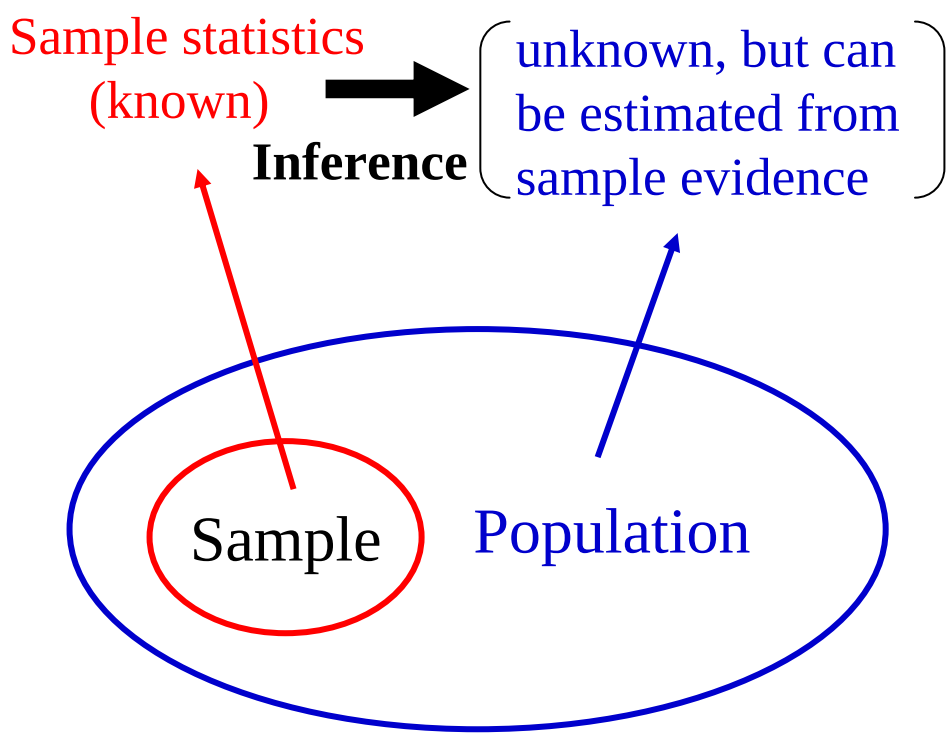
Population and Samples

Population:

- A population is a well-defined collection of objects.
- When desired information is available for all objects in the population, we have what is called a ***census***.
- Constraints on time, money, and other scarce resources usually make a ***census*** impractical or infeasible.

Sample:

- A sample is a subset of the population.

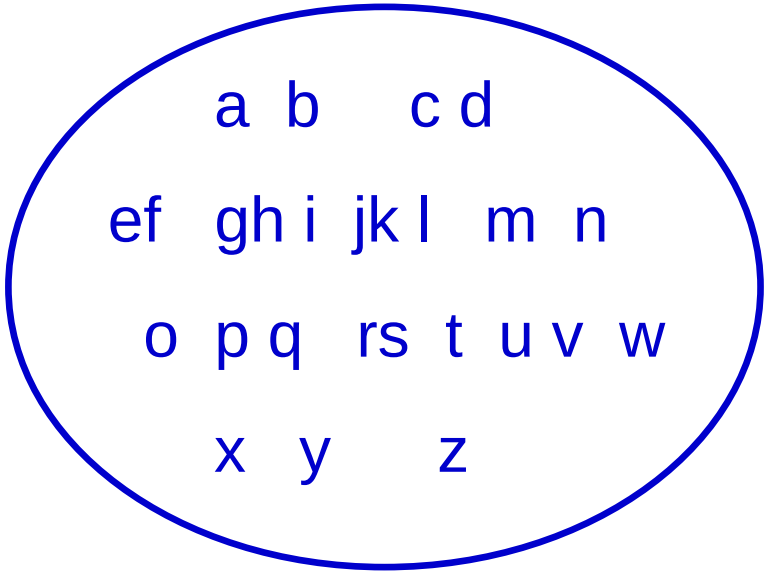


Key Definitions

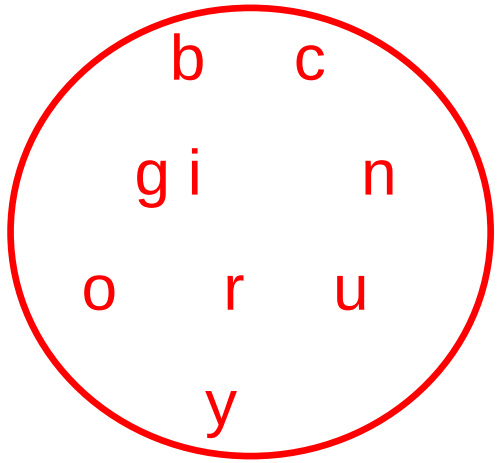
- A **population** is the entire collection of things under consideration and referred to as the frame.
  - The sampling unit is each object or individual in the frame.
  - A **parameter** is a summary measure computed to describe a characteristic of the population.
- A **sample** is a subset of the population selected for analysis.
  - A **statistic** is a summary measure computed to describe a characteristic of the sample drawn from the population.

Population vs. Sample

Population



Sample



Why Sample?

- Less time consuming than a census.
- Less costly to administer than a census.
- It is possible to obtain statistical results of a sufficiently high precision based on samples.

Strive for representative samples to reflect the population of interest accurately!

Data and Observations

- A *variable* is a characteristic of an individual or object in the population whose value may change from one object to another.
- We shall denote variables by lowercase letters from the end of our alphabet.

Such as: x, y, z, w.

Examples (of variables for human beings):

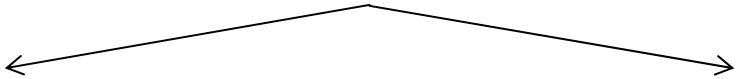
Age, Weight, Height, Eye colour, Marital Status, Blood Type, Household size,… etc.

Different Types of Variables

Quantitative :

Like the time for a person to finish a task or the person's age, or the lifetime of a machine, or number of people in your household, etc.

Types of quantitative variables



Discrete

A variable whose possible values form a finite (or countable) set.
e.g., number of people, Household size.

Continuous

A variable whose possible values form some interval of numbers.
e.g., time, age.

Qualitative :

As the person's nationality or the person's preferred sport or blood type, Marital Status, ..., etc.

Data

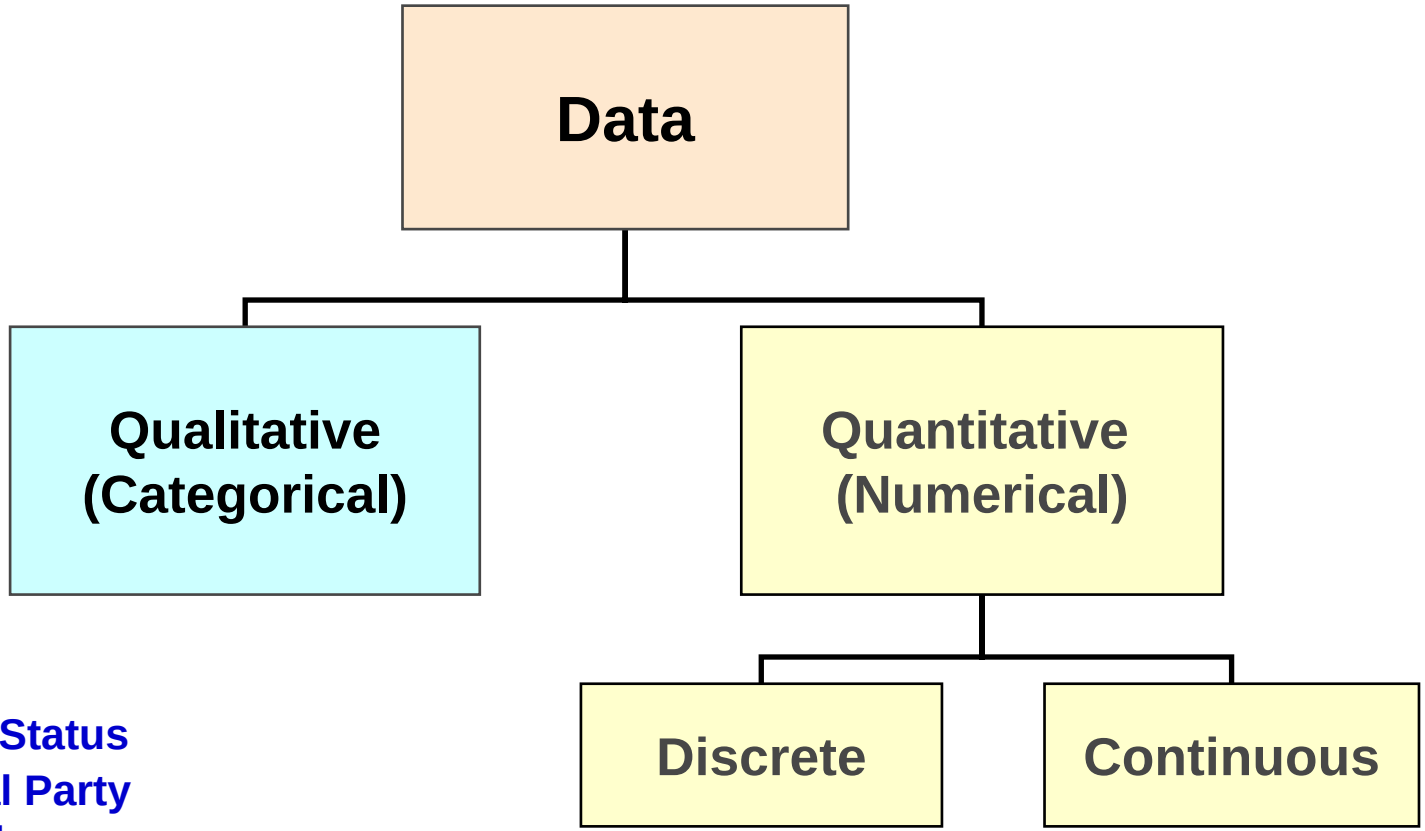
Consists of the values of a variable for one or more people or things . That is, the ***data*** is the information collected, organized, and analyzed by statisticians.

- *Univariate data* consists of observations on a single variable.
- Example:
- The blood type of 4 persons .
  The data is: A O A+ AB,
- The lifetime of 5 lights.
  The data set is 12.3, 10.9, 21.1, 8.3, 31.5 hr.

- *Multivariate data* consists of observations on more than two variables.
- Example: the height and weight of 4 basketball players on a team.
  The data is (72, 169), (75, 176), (77,180), (81,190).
  This is a *bivariate* data set (observations).

How to collect the data: (p. 7)

Data Types



Examples:

- **Marital Status**
- **Political Party**
- **Eye Color**

(Defined categories)

Examples:

- **Number of Children**
- **Defects per hour**

(Counted items)

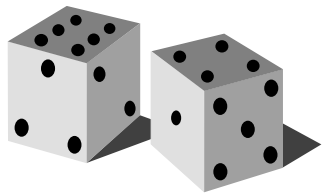
Examples:

- **Weight**
- **Voltage**

(Measured characteristics)

Simple Random Sampling

- Every possible sample of a given size has an equal chance of being selected.
- Selection may be with replacement or without replacement.
- The sample can be obtained using a table of random numbers or computer random number generator.



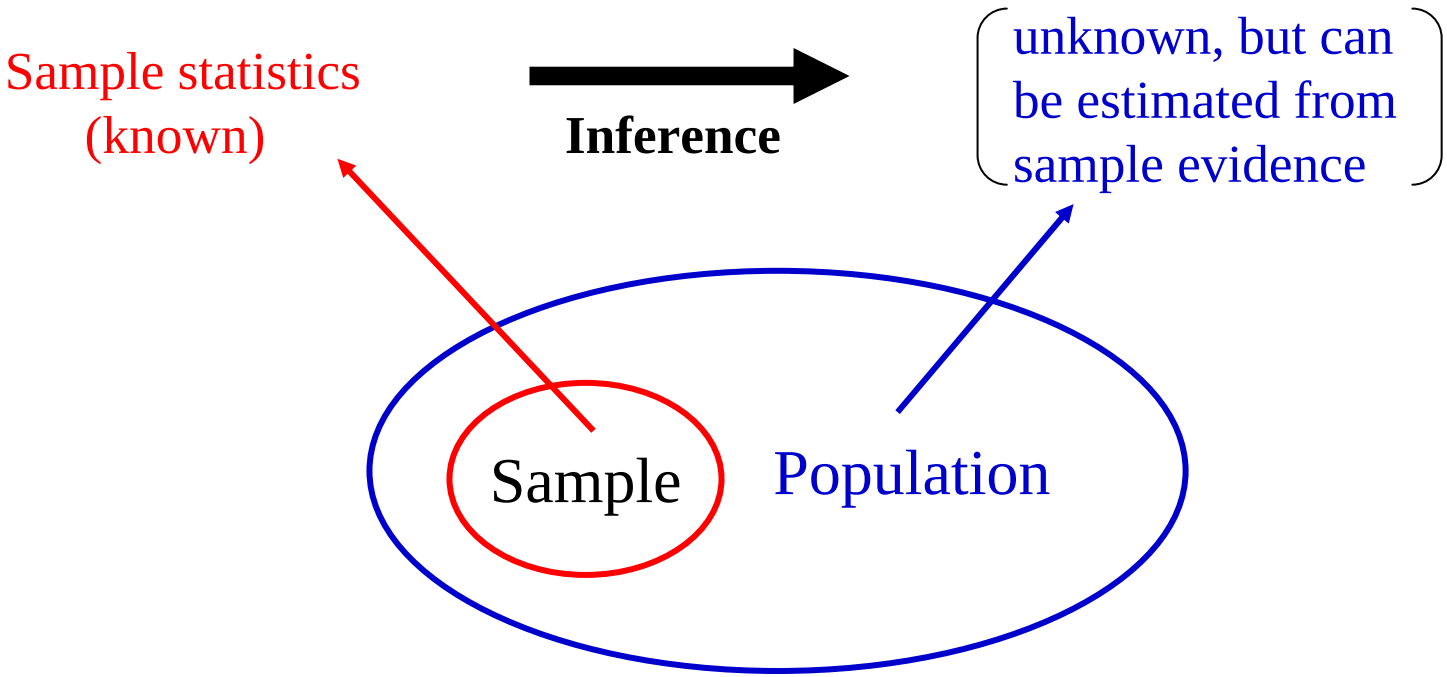
Branches of Statistics

Descriptive Statistics

- Provides numerical and graphic procedures to summarize the information of the data in a clear and understandable way.

Inferential Statistics

- Provides procedures to draw inferences about a population from a sample.



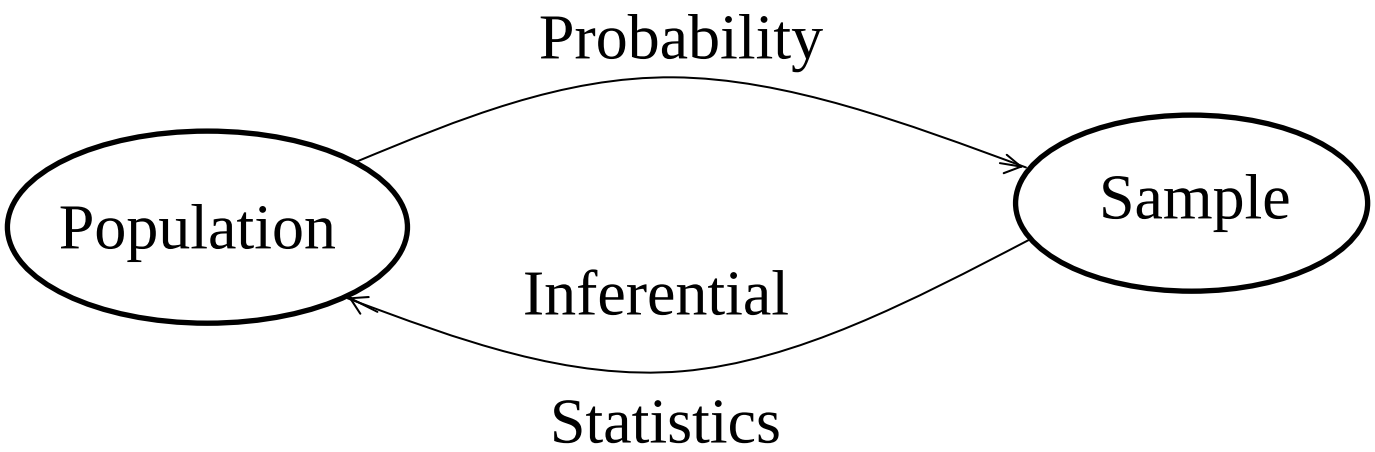
Relationship Between Probability and Inferential Statistics

Probability

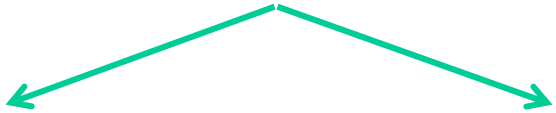
- In probability problems, properties of the population under study are assumed known (e.g., in a numerical population, some specified distribution of the population values may be assumed) and question regarding a sample taken from the population are posed and answered.

Statistics

- In Statistics problems, characteristics of a sample are available to the experimenter, and this information enables the experimenter to draw conclusion about the population.



Descriptive statistics



Visual techniques
Sect. 1.2

Numerical techniques
Sect. 1.3 and 1.4

Notation

The ***sample size*** is the number of observation in a single sample. It will be denoted by *n*. So that *n = 4* for the sample of 4 persons.

If two samples are simultaneously under consideration, either *m* and *n* or n_1 and n_2 can be used to denote the numbers of observations.

Given a data set with size *n* on some variable *x*, the individual observations will be denoted by $x_1, x_2, \cdots, x_n$.

1.2 Pictorial and Tabular Methods in Descriptive Statistics

Pictorial (*Visual*) Methods :

1) **Stem-and- Leaf Display**
2) **Dotplot Display**
3) **Histogram**

1) Stem-and- Leaf Display

Assume a numerical data set $x_1, x_2, \cdots, x_n$ for which each x consists of at least two digits. A quick way to obtain an informative visual representation of the data is to construct a *stem-and-leaf display*.

Steps for constructing a Stem-and-Leaf Display

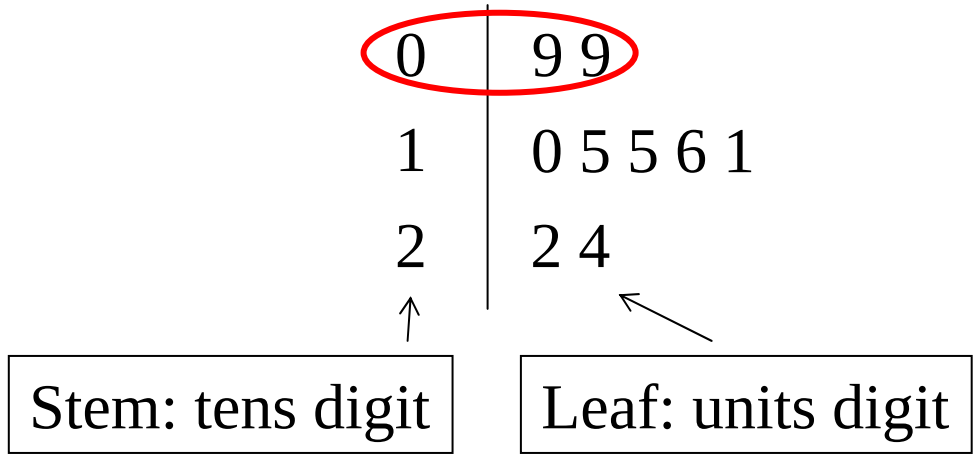
1. Select one or more leading digits for the stem values. The trailing digits become the leaves. Ex. the obs. *45* has stem *4* and leaf *5*.
2. List stem values in a vertical column.
3. Record the leaf for every observation.
4. Indicate the units for the stem and leaf on the display.

Example 1: Construct the Stem-and-Leaf display of the following data

9, 10, 15, 22, 9, 15, 16, 24,11

Solution:

9, 9, 10, 11, 15, 15, 16, 22, 24



Stem-and- Leaf Displays

- Typical value
- Spread about a value
- Presence of gaps
- Symmetry
- Number and location of peaks
- Presence of outlying values

Note:

The "*leaf*" is usually the last digit of the number and the other digits to the left of the "leaf" form the "*stem*".

Example 2:

The number 123 would be split as leaf =3 and *stem = 12*.

The *stem* in the number 6433 can be chosen as:

A single digit *6*, (thousands digit) or

too few stems.

Three digits *643*, (thousands, hundreds and tens digits) or

too many stems.

These would yield an uninformative display

Two digits *64* (thousands and hundreds digits) .

This would yield an informative display

Example 3: Construct the Stem-and-Leaf display of the following data

6435, 6464, 6433, 6470, 6526, 6527, 6506, 6583, 6605, 6694, 6614, 6790, 6770, 6700, 6798, 6770, 6745, 6713, 6890, 6870, 6873, 6850, 6900, 6927, 6936, 6904, 7051, 7005, 7011, 7040, 7050, 7022, 7131 , 7169, 7168, 7105, 7113, 7165, 7280, 7209

Solution:

(a)

64	35 64 33 70
65	26 27 06 83
66	05 94 14
67	90 70 00 98 70 45 13
68	90 70 73 50
69	00 27 36 04
70	51 05 11 40 50 22
71	31 69 68 05 13 65
72	80 09

Stem: Thousands and hundreds digits
Leaf: Tens and ones digits

(a) two-digit leaves

(b)

```
Stem-and-leaf of yardage   N   = 40
Leaf Unit = 10
    4    64 3367
    8    65 0228
   11    66 019
   18    67 0147799
   (4)   68 5779
   18    69 0023
   14    70 012455
    8    71 013666
    2    72 08
```

The middle value interval

(b) Display from Minitab with truncated one-digit leaves

Questions:

Use the display (b) to find:

- The smallest and largest observations.
- How many observations having the value of 6690.
- The number of observations that are greater than 7000.
- Probability that $X < 6690$ or Proportion of the observations that do not exceed 6690.
- Probability that $6500 < X < 7000$ or Proportion of the observations lie in the interval (6500, 7000).
- The middle value interval.

Dotplots

❖Represent data with dots.

❖It is an attractive summary of numerical data when the data set is reasonably *small* or there are *relatively few distinct values*.

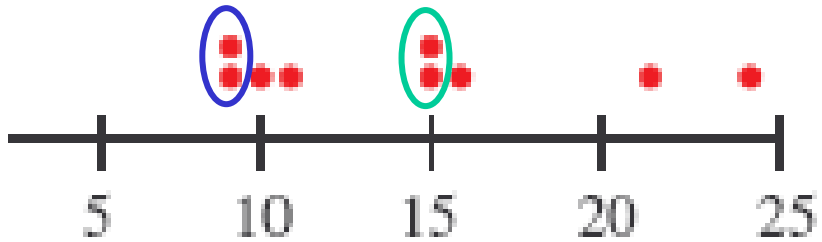
Steps for constructing a dotplot

1. Each observation is represented by a dot above the corresponding location on a horizontal measurement scale.
2. When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically.

As with a stem-and-leaf display, a ***dotplot*** gives information about location, spread, extremes, and gaps.

Example 4: Consider the data 9, 10, 15, 22, 9, 15, 16, 24,11

9, 9, 10, 11, 15, 15, 16, 22, 24



Note

If the data set consists of 50 or 100 observations (large size), it will have much more cumbersome to construct a dotplot.

The next technique (Histograms) is well suited for such data situations.

Histograms

(1) Discrete Data (2) Continuous Data → Equal Class Widths / Unequal Widths

(1) **Discrete Data**

Notations

Frequency: The frequency (*f*) of a value is the number of times that value occurs in the data set.

Relative frequency: The relative frequency (*rf*) of a value is the fraction of times the value occurs. That is,

$$rf = \frac{f}{n}$$

Percentage: Multiplying the *rf* by 100 gives the percentage.

The *rf*s , or percentages are usually of more interest than the frequencies.

Theoretically, the *rf*s should sum to 1, but in practice the sum may differ slightly from 1 because of rounding.

Frequency distribution: It is a tabulation of the frequencies and/or relative frequencies. (Sometimes called ***frequency table***)

Constructing a Histogram for Discrete Data

1. Determine the *distinct values (d.v.)* in the data.
2. Determine the *frequency* and *relative frequency* for each *d.v.*
3. Then mark *distinct values* on a *horizontal* scale.
4. Above each distinct value, draw a rectangle whose *height* is the *relative frequency* (or *frequency*) of that value.

Example 5: The Journal of Marketing Research published the results of a study in which 22 consumers reported the number of times that they had purchased a particular brand of a product during the previous 48-week period. The results were as follows.

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

Solution:

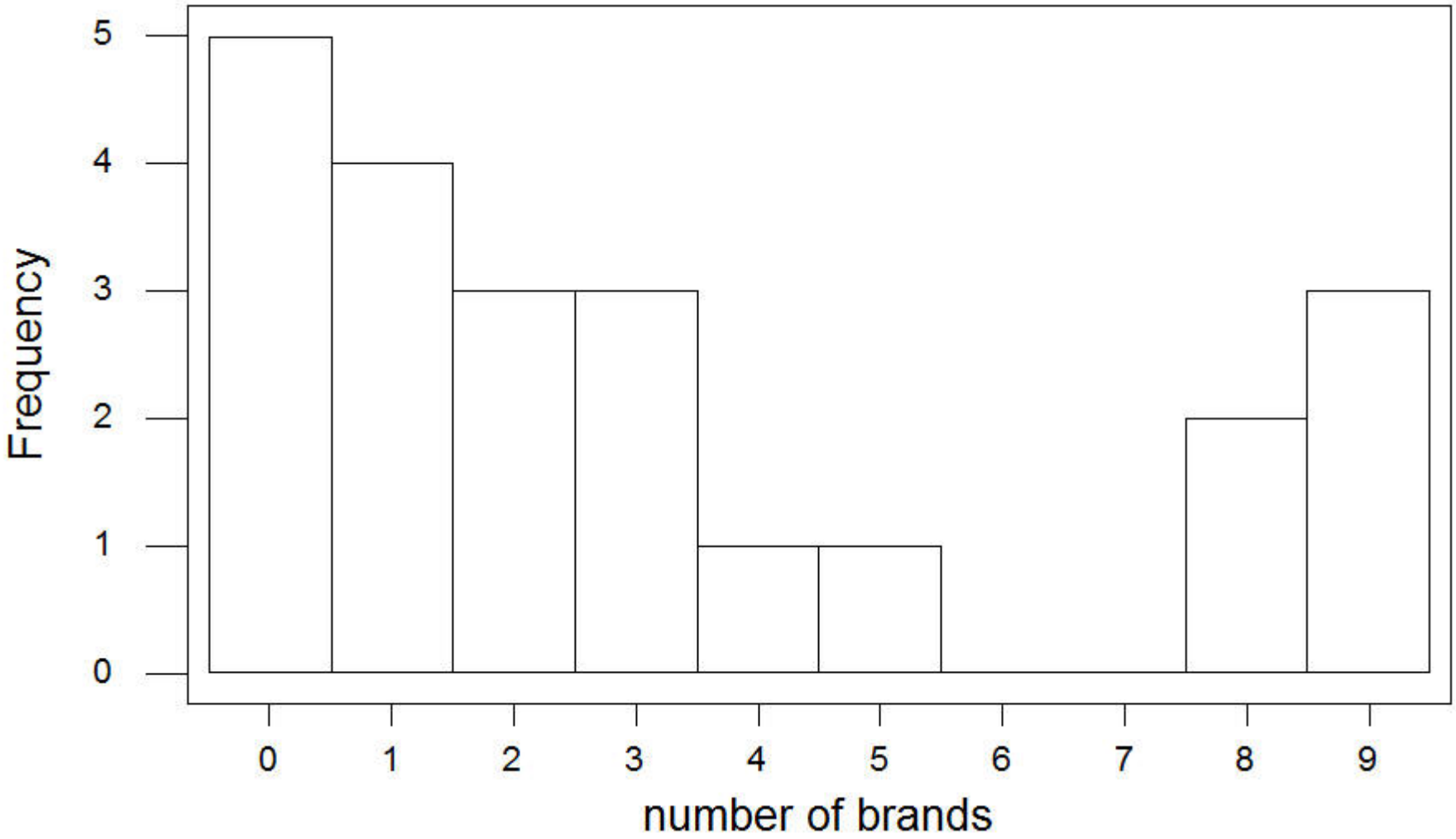
The distinct values (*d.v.*) are 0, 1, 2, 3, 4, 5, 8, and 9.

The sample size is 22.

Frequency distribution

d.v.	0	1	2	3	4	5	8	9
f.	5	4	3	3	1	1	2	3
r.f.	5/22	4/22	3/22	3/22	1/22	1/22	2/22	3/22

Histogram of the number of brands bought



Frequency Distribution

# of times items purchased (*d.v.*)	Frequency	Relative Frequency	Percentage
0	5	5/22	22.73%
1	4	4/22	18.18%
2	3	3/22	13.64%
3	3	3/22	13.64%
4	1	1/22	4.55%
5	1	1/22	4.55%
8	2	2/22	9.09%
9	3	3/22	13.64%
Total	22	1.00	100%

From the frequency distribution, one can find out:

The number of consumers who never bought the product is 5.

The relative frequency of shoppers in this study that never bought the brand under investigation is 5/22 = 0.2273.

The proportion of shoppers in this study that never bought the brand under investigation is 22.73%.

The percentage of shoppers in this study that bought between 1 and 3 brands is 18.18+13.64+13.64 = 45.46%.

That is, roughly 45.5% of shoppers bought between 1 and 3 brands.

(2) Continuous Data

Histograms

Continuous Data: Equal Class Widths

The main point in this case is to **group** the data.

Grouping the Data

- The first step to group data is to decide on the **classes.** Usually number of classes should be between 5 and 20 **.**
- One convenient way is to group by using the classes a-<*b*, b-<c, ...
- The symbol -< is short hand for "*up to, but not including*". Ex., the class 10-<20 means 10 up to, but not including, 20.
- Determine the number of elements in each class (frequency of the class).

Guidelines for Grouping:

- ❑Number of classes should be small enough to provide an effective summary but large enough to display the relevant characteristics.
- ❑Each observation must belong to one and, and only one, class.

Example 6: Here are some test scores from a Math 2060 class (40 obs.).

65	91	85	76	85	87	79	93
82	75	99	70	88	78	83	59
87	69	89	54	74	89	83	80
94	67	77	92	82	70	94	84
96	98	46	70	90	96	88	72

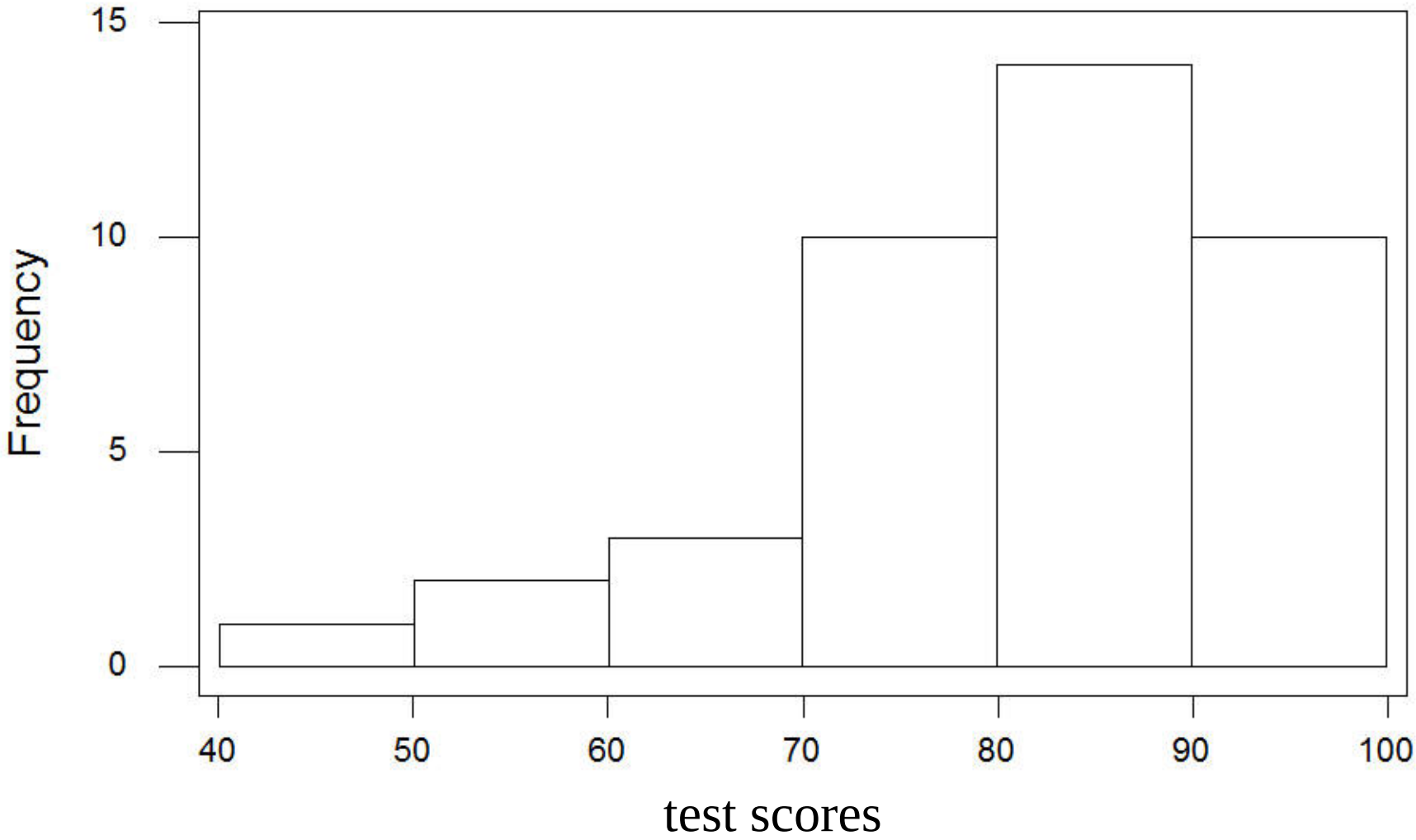
- Note how it's hard to get a feel for this data in its current format because it is unorganized.
- To group the data, we should first identify the lowest and highest values.
- We do this because we want to be sure that each value in the list fits into one of our classes.
- The lowest value here is 46, and the highest is 99. A set of classes that would work here is 40 -<50, 50 -<60, 60 -<70, 70 -<80, 80 -< 90, and 90 -< 100. So there are 6 classes.
- Determine the frequency of each class.

Frequency Distribution →

We can now see that the biggest numbers of tests were between 80 and 90, and most of the tests were between 70 and 99.

Class	*f.*	*r.f.*
40 -< 50	1	1/40
50 -< 60	2	2/40
60 -< 70	3	3/40
70 -< 80	10	10/40
80 -< 90	14	14/40
90 -< 100	10	10/40

Date file : C:\Ammar\in Canada 2008\In Canada 2008\mat 2060\Final\example6.tex



Histograms

Continuous Data: Unequal Class Widths (Also works for Equal widths)

Constructing a Histogram:

- Determine the frequency and relative frequency for each class.
- Calculate the *height* of each rectangle using the formula

$$\textit{rectangle height} = \frac{\text{relative frequency of the class}}{\text{class width}}$$

The resulting rectangle heights are usually called *densities* and the vertical scale is the *density scale*.

Note: This prescription works when the class widths are equal.

- The area of each rectangle is the relative frequency of the corresponding class.
- Furthermore, since the sum of relative frequencies must be **1.0**, the total area of all rectangles in a density histogram is **1**.

Example 7: (1.11 p. 20)

Consider the following 48 observations on measured bond strength:

11.5	12.1	9.9	9.3	7.8	6.2	6.6	7.0	13.4	17.1	9.3	5.6
5.7	5.4	5.2	5.1	4.9	10.7	15.2	8.5	4.2	4.0	3.9	3.8
3.6	3.4	20.6	25.5	13.8	12.6	13.1	8.9	8.2	10.7	14.2	7.6
5.2	5.5	5.1	5.0	5.2	4.8	4.1	3.8	3.7	3.6	3.6	3.6

The sample size is 48.

Frequency table

Class	2-<4	4-<6	6-<8	8-<12	12-<20	20-<30
f.	9	15	5	9	8	2
r.f.	.1875	.3125	.1042	.1875	.1667	.0417
Density	.094	.156	.052	.047	.021	.004

f_i / n

$r.f._i / \Delta_i$

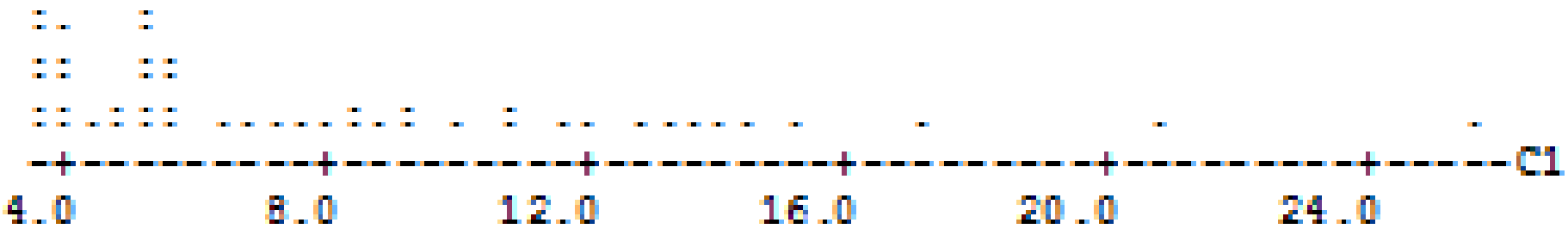
the width of class i.

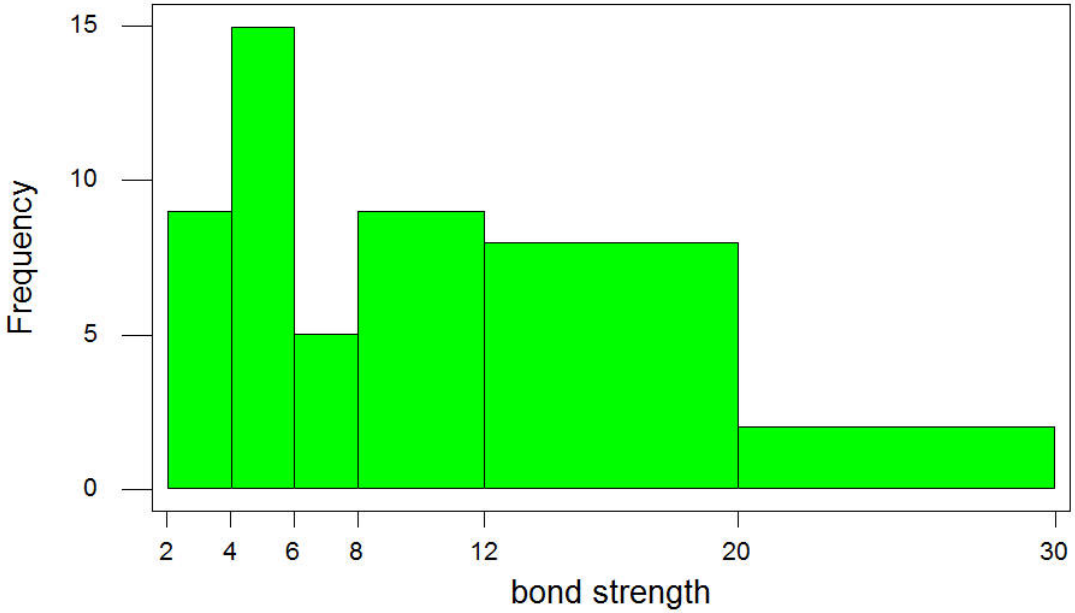
From the above table,

Proportion of observations less 8 $\approx$ .1875+ .3125+ .1042 = 0.0.6042

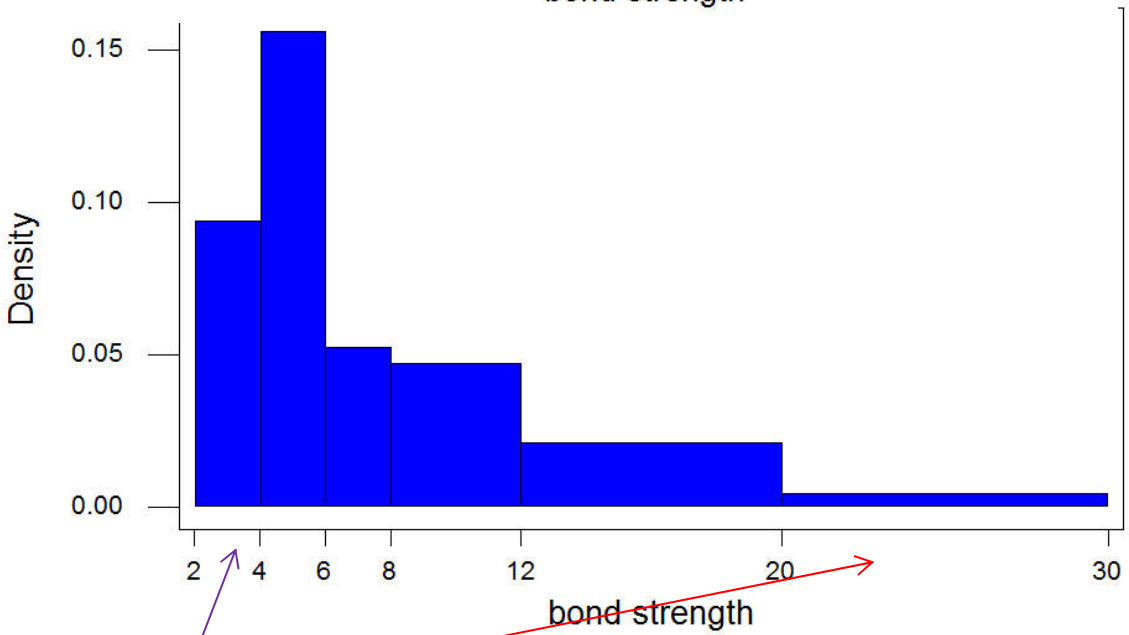
Date file : C:\Ammar\in Canada 2008\In Canada 2008\mat 2060\Final\example7.tex

Dotplot





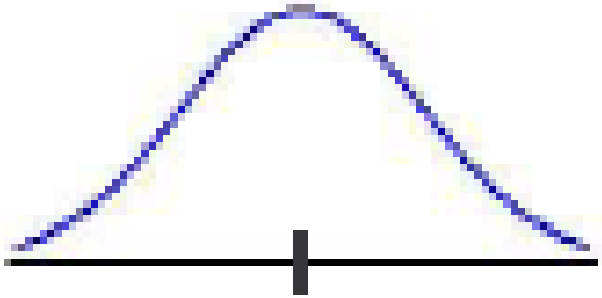
Minitab Histogram for the bond strength data.



Minitab density histogram for the bond strength data.

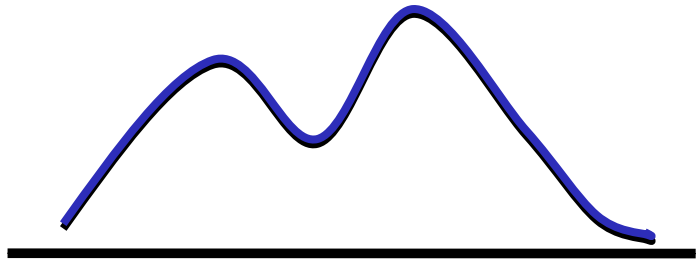
The right (upper) tail stretches out much farther than does the left (lower) tail.

Histogram Shapes



Symmetric

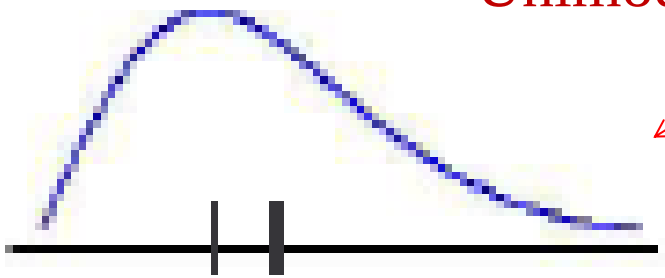
The left half is a mirror image of the right half.



Bimodal

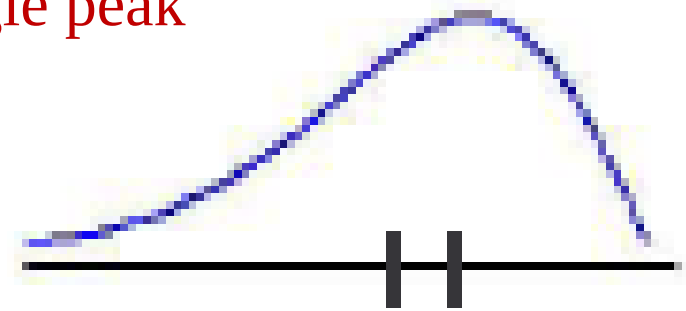
It has two peaks.

Unimodal (rises) a single peak



positively skewed

The right (upper) tail is stretched out compared with the left (lower) tail.



Negatively skewed

The stretched is to left.

Qualitative Data

Both a frequency distribution and a histogram can be constructed when the data set is *qualitative* (categorical).

The *classes* are the *different categories* of the corresponding variable.

Count the number of time for each category, which is the *frequency*.

Example 8: Twenty-five army inductees were given a blood test to determine their blood type. The data set is as follows: A B B AB O O O B AB B B B O A O A O O O AB AB A O B A.

Construct a frequency distribution for the data.

Solution:

The four blood types are the classes for the distribution.

Count the number of times each blood type appear.

The frequency distribution

Blood Type	*f.*	*r.f.*	Percent
A	5	5/25 = 0.20	20
B	7	7/25 = 0.28	28
O	9	9/25 = 0.36	36
AB	4	4/25 = 0.16	16
Total	25	1.00	100

For the sample, more people have type O blood than any other type.

1.3 Measures of Location (Central Tendency):

Let $x_1, x_2, \ldots, x_n$ be the sample values (numbers).

What features of such sample are of most interest and deserve emphasis?

One important characteristic of this sample is its location, and in particular its centre.

Now, we present methods for describing the location of a data set.

The Mean

The *sample mean* denoted $\bar{x}$ is $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

The *population mean*, μ, is sum of N population values / N (Unknown).

$\bar{x}$ gives an estimate of μ.

The *sample mean* satisfies the following property

$$\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 0$$

Example 9: (Ex. 1.13, p. 29)

$x_1 = 16.1$ $x_2 = 9.6$ $x_3 = 24.9$ $x_4 = 20.4$ $x_5 = 12.7$ $x_6 = 21.2$ $x_7 = 30.2$

$x_8 = 25.8$ $x_9 = 18.5$ $x_{10} = 10.3$ $x_{11} = 25.3$ $x_{12} = 14.0$ $x_{13} = 27.1$ $x_{14} = 45.0$

$x_{15} = 23.3$ $x_{16} = 24.2$ $x_{17} = 14.6$ $x_{18} = 8.9$ $x_{19} = 32.4$ $x_{20} = 11.8$ $x_{21} = 28.5$

The *sample mean* is $\bar{x} = \dfrac{444.8}{21} = 21.18$

The *point estimate of the population mean* is 21.18

Note:-

The *mean* suffers from one deficiency that makes it an inappropriate measure of center under some circumstances.

It greatly affected by the *outliers* (small or large observation).

In Ex. 9, the value 45.0 is obviously an outlier. Without this obs, $\bar{x} = 399.8/20 = 19.99$

That is, the outlier increases the mean by more than 1 μ m.

If 45.0 were replaced by 295.0, a rally extreme outlier, then $\bar{x} = 694.8/21 = 33.09$, which is larger than all but one of the obs. !

The Median

The ***sample median*** is the middle value in a set of data that is arranged in ascending order.

The symbol $\tilde{x}$ will be used to represent the sample median.

How to compute the median:

1) Ordering the observations from smallest to largest (with any repeated values included, so that every sample observation appears in the ordered list). Assume the ordered values are $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$
2) Then

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if n is odd} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{if n is even} \end{cases}$$

The ***population median***, denoted by $\tilde{\mu}$, is the middle value in the population (Unknown).

$\tilde{x}$ gives an estimate of $\tilde{\mu}$.

Example 10: (Ex. 1.14, p. 31)

$x_1 = 15.2$ $x_2 = 9.3$ $x_3 = 7.6$ $x_4 = 11.9$ $x_5 = 10.4$ $x_6 = 9.7$

$x_7 = 20.4$ $x_8 = 9.4$ $x_9 = 11.5$ $x_{10} = 16.2$ $x_{11} = 9.4$ $x_{12} = 8.3$

The list of ordered values is

7.6 8.3 9.3 9.4 9.4 9.7 10.4 11.5 11.9 15.2 16.2 20.4

The size n = 12 (even) $x_{(\frac{12}{2})}$ $x_{(\frac{12}{2}+1)}$

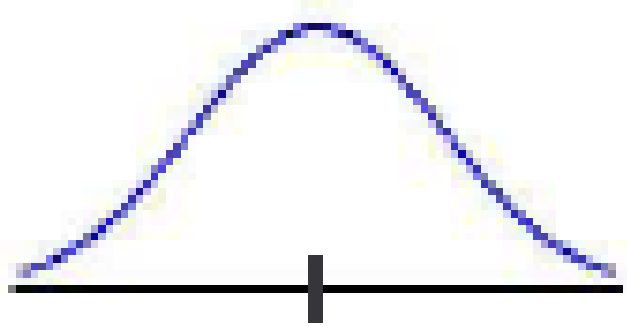
The *sample median* is $\tilde{x} = \dfrac{9.7 + 10.4}{2} = 10.05$

The *point estimate of the population median* is 10.05

Notice: if the largest observation, 20.4 had not appeared in the sample, The resulting sample median for n=11 obs would be $x_{(6)}$= 9.7.
The sample mean $\bar{x} = 11.61$, which is somewhat larger than the median because of the *outliers*, 15.2, 16.2, and 20.4.

The *sample median* is ***very insensitive*** to a number of extremely small or extremely large data values.

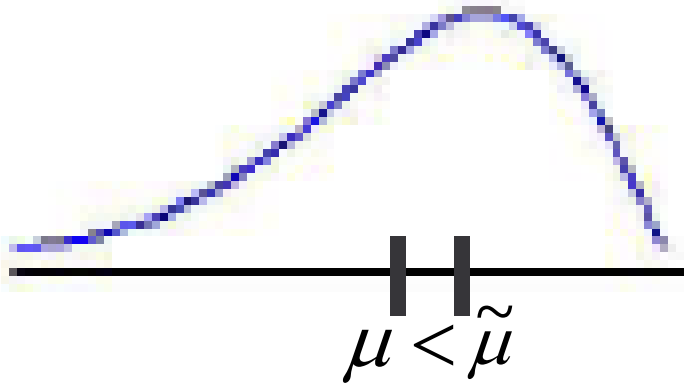
Three Different Shapes for a Population Distribution



$\tilde{\mu} = \mu$

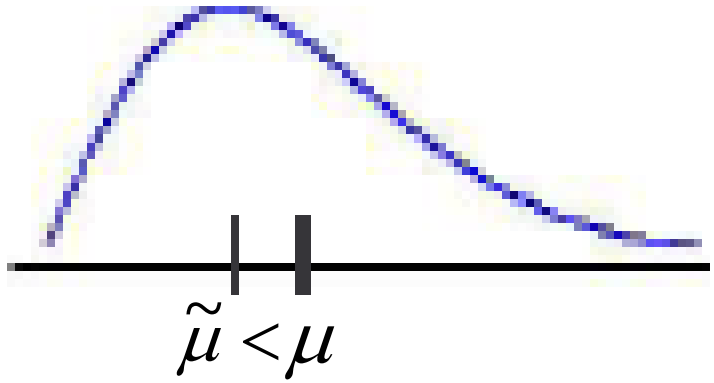
Symmetric

The population mean and median will not generally be identical.



$\mu < \tilde{\mu}$

Negative skewed



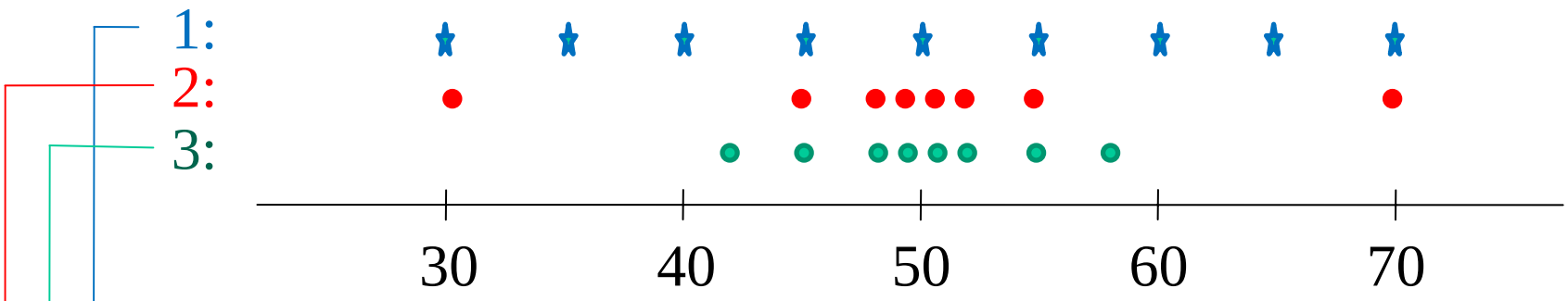
$\tilde{\mu} < \mu$

Positive skewed

The relation between the population mean and median determines the shape of the Population Distribution.

1.4 Measures of Variability:

- Reporting a measure of centre gives *only partial information* about the sample.
- Different samples or population may have identical measures of centre yet differ from one another in other important ways.
- The following figure shows dotplots of three samples with the *same mean* and *median*, yet the extent of spread about the centre is different for all three samples.



has the largest amount of variability.

has the smallest amount of variability.

is intermediate to the other two in this respect.

Measures of Variability

- Range
- Variance
- Standard Deviation

The range:

It is the simplest measure of variability. The range = $x_{(n)} - x_{(1)}$

Notice that the range of sample 1 is much larger than it is for sample 3, reflecting more variability in the first sample than in the third.

A ***defect*** of the range is that it depends only on the two most extreme observations and disregards the positions of the remaining values. Samples 1 and 2 have the same range, but there is much less variability in the second sample than in the first.

So, ***the Range is not Enough.***

The sample variance

The variance takes into account the deviation around the mean of the data.

The formula for the ***sample variance*** , denoted by s^2, is as follows

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

n-1 is the ***degrees of freedom*** (***df***).

The ***sample standard deviation***, denoted by ***s***, is $s = \pm\sqrt{s^2}$

The *Standard deviation* is a measure of the spread of the data using the same units as the data.

Example [similar to 1.16 (p. 37) 6th edition or 1.15 (p. 33) 7th edition]

0.684 2.540 0.924 3.130 1.038 0.598 0.483 3.520 1.285 2.650 1.497

Data file: C:\Ammar\in Canada 2008\In Canada 2008\mat 2060\Final\example1_16.tex

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
0.684	-1.02709	1.05492
2.540	0.82891	0.68709
0.924	-0.78709	0.61951
3.130	1.41891	2.01330
1.038	-0.67309	0.45305
0.598	-1.11309	1.23897
0.483	-1.22809	1.50821
3.520	1.80891	3.27215
1.285	-0.42609	0.18155
2.650	0.93891	0.88155
1.497	0.25891	0.06703
Sum → 18.822		11.9773 ← S_{xx}

Mean $\bar{x}$ =18.822/11= 1.66809

Variance: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1} = \frac{11.9773}{11-1} = 1.19773$

Standard deviation: $S = \sqrt{11.9773} = 1.09441$

```
MTB >  set 'C:\Ammar\in Canada 2008\In Canada 2008\mat 2060\Final\example1_16.tex' c2
Entering data from file: C:\Ammar\in Canada 2008\In Canada 2008\mat
2060\Final\example1_16.tex
MTB > let k1 = mean(c2)
MTB > let c3 = c2-k1
MTB > let c4 = sqr(c3)
MTB > let c4 = c3**2
MTB > let k2 = sum(c4)
MTB > let k3=k2/(n(c2)-1)
MTB > name c2 'x' c3 'x-xb' c4 '(x-xb)^2'
MTB > name k1 'mean' k2 'sum squares' k3 'sample variance' k4 'sample st.dv.'
MTB > let k4 = sqrt(k3)
MTB > print k1-k4
```

Data Display

mean	1.66809
sum squares	11.9358
sample variance	1.19358
sample st.dv.	1.09251

S_{xx}

$\frac{S_{xx}}{n-1}$

x	x-xb	(x-xb)^2
0.684	-0.98409	0.96843
2.540	0.87191	0.76023
0.924	-0.74409	0.55367
3.130	1.46191	2.13718
1.038	-0.63009	0.39701
0.598	-1.07009	1.14509
0.483	-1.18509	1.40444
3.520	1.85191	3.42957
1.285	-0.38309	0.14676
2.650	0.98191	0.96415
1.497	-0.17109	0.02927

Formula for s^2

An alternative expression for the numerator of s^2 *is*

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = \sum x_i^2 - n\bar{x}^2$$

Previous Example

```
MTB > let c5 = c2**2
MTB > name c5 'x^2'
MTB > let k5 = sum(c5)
MTB > name k5 'sum x^2'
MTB > let k6 = sum(c2)
MTB > name k6 'sum x'
MTB > let k7 = k5 - k6**2/n
MTB > let k7 = k5 - k6**2/n(c2)
MTB > name k7 'sample variance 2'
MTB > print k5-k7
```

Data Display

sum x^2	42.5436
sum x	18.3490
sample variance 2	11.9358

x	x^2
0.684	0.4679
2.540	6.4516
0.924	0.8538
3.130	9.7969
1.038	1.0774
0.598	0.3576
0.483	0.2333
3.520	12.3904
1.285	1.6512
2.650	7.0225
1.497	2.2410
18.3490	42.5436

```
MTB > desc c2
```

Descriptive Statistics: x

Variable	N	Mean	Median	TrMean	StDev	SE Mean
x	11	1.668	1.285	1.594	1.093	0.329

Variable	Minimum	Maximum	Q1	Q3
x	0.483	3.520	0.684	2.650

Note on Minitab:

If you need to compute a specific sample statistic for the sample saved in C1, you can use subcommands:

Desc C1;

- Mean; for sample mean
- Vari; for sample variance
- Stde; for sample standard deviation
- Rang; for sample range
- Mini; for sample minimum
- Maxi; for sample maximum
- Median; for sample median
- n. for sample size.

Downloading website:

http://its.dal.ca/services/computer_services/downloads/

Properties of s^2

Let $x_1, x_2, \ldots, x_n$ be any sample and c be any nonzero constant.

1) If $y_1 = x_1 \pm c, \ldots, y_n = x_n \pm c$, then $s_y^2 = s_x^2$

2) If $y_1 = c\,x_1, \ldots, y_n = c\,x_n$, then $s_y^2 = c^2\,s_x^2$, $s_y = |c|\,s_x$

where

s_x^2 is the sample variance of the x's and

s_y^2 is the sample variance of the y's.

Two more examples will be given in the class.

Upper and Lower Fourths

After the *n* observations in a sample are ordered from smallest to largest, the *lower* (*upper*) *fourth is the median of the smallest* (*largest*) half of the data, where the median is included in both halves if *n* is odd.

A measure of the spread that is resistant to outliers is the *fourth spread*, given by f_s = *upper fourth – lower fourth.*

Outliers

Any observation farther than 1.5*fs* from the closest fourth is an ***outlier***.

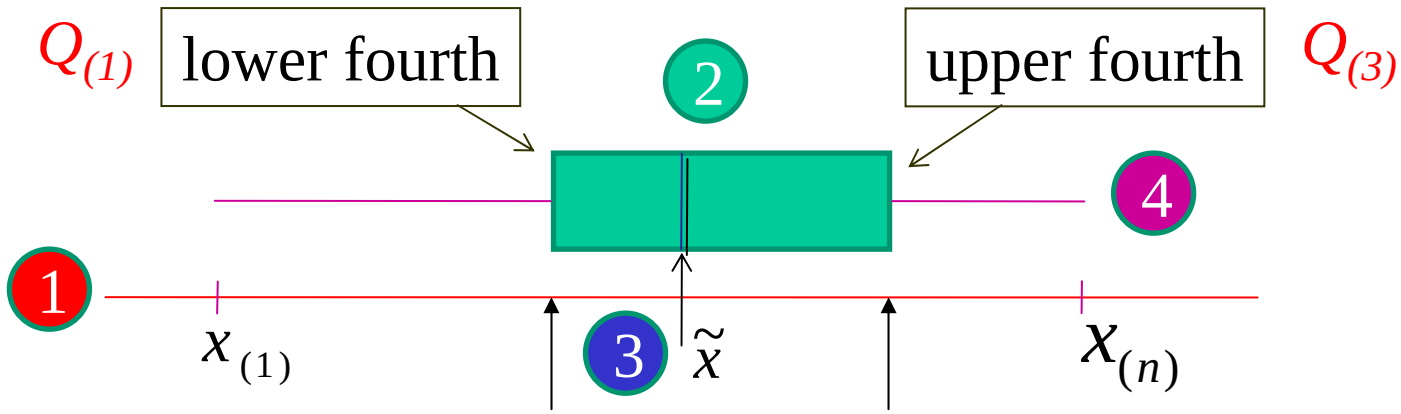
An outlier is ***extreme*** if it is more than *3fs* from the nearest fourth, and it is ***mild*** otherwise.

How to construct the Boxplot

The *boxplot* is based on the following five-number summary:

Smallest x_i	lower fourth	median	upper fourth	largest x_i
$X_{(1)}$	$Q_{(1)}$	$\tilde{x}$	$Q_{(3)}$	$X_{(n)}$

1. Draw a horizontal measurement scale.
2. Place a rectangle above this axis; the left edge of the rectangle is at the lower fourth, and the right edge is at the upper fourth. So box width = f_s.
3. Place a vertical line segment inside the rectangle at the location of $\tilde{x}$
4. Draw lines out from either end of the rectangle to the smallest and largest observations.



Example 1.18 (p. 41) Draw the boxplot and check if there is any outlier of the following data set (n=19)

The largest half

40 52 55 60 70 75 85 85 90 90 92 94 94 95 98 100 115 125 125

The smallest half

lower fourth (70+75)/2=72.5

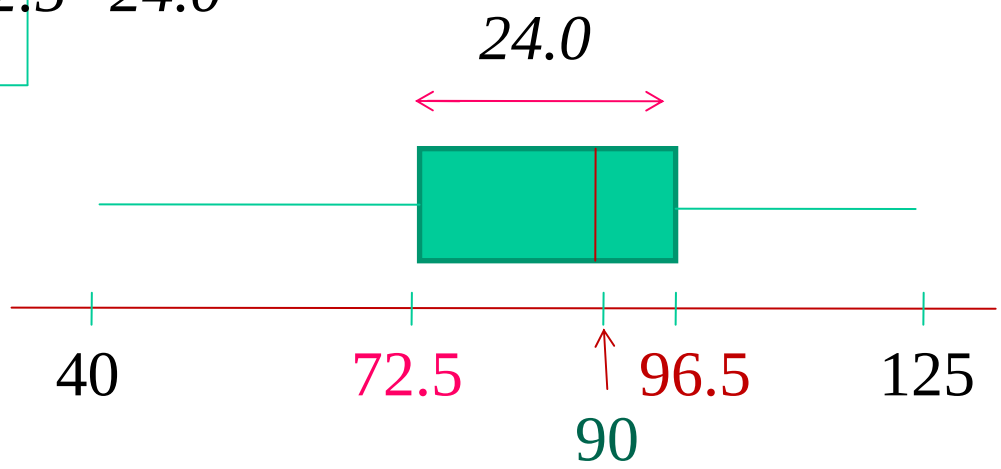
upper fourth (94+95)/2=96.5

The *fourth spread*, $f_s = 96.5 - 72.5 = 24.0$

$\tilde{x} = x_{(10)} = 90$

$x_{(1)} = 40$

$x_{(19)} = 125$



The outliers

$1.5\ f_s = 1.5 \times 24 = 36$

Lower fourth-$1.5\ f_s = 72.5 - 36 = 36.5$

Upper fourth $+1.5\ f_s = 96.5 + 36 = 132.5$

Therefore, **there is no outliers**.

Example 1.19 (p. 42) (n=25)

5.3 8.2 13.8 74.1 85.3 88.0 90.2 91.5 92.4 92.9 93.6 94.3 94.8
94.9 95.5 95.8 95.9 96.6 96.7 98.1 99.0 101.4 103.7 106.0 113.5

Find the outliers, and decide if there are either mild or extreme outliers.

Solution: $\tilde{x} = x_{(13)} =$ 94.8

The smallest half (13)

5.3 8.2 13.8 74.1 85.3 88.0 90.2 91.5 92.4 92.9 93.6 94.3 94.8

lower fourth = 90.2

The largest half

94.8 94.9 95.5 95.8 95.9 96.6 96.7 98.1 99.0 101.4 103.7 106.0 113.5

upper fourth =96.7

The *fourth spread*, $f_s = 96.7 - 90.2 = 6.5$ ⇒ $1.5\ f_s = 1.5 \times 6.5 = 9.75$

Lower fourth-$1.5\ f_s$ =90.2 - 9.75 = 80.45 ⇒ 5.3 8.2 13.8 74.1

Upper fourth +$1.5\ f_s$ = 96.7 + 9.75 = 106.45 ⇒ 113.5

outliers

The extreme outliers:

$3\ f_s = 3 \times 6.5 = 19.50$

Lower fourth-$3\ f_s$ =90.2 - 19. 5 = 70.7 ⇒ 5.3 8.2 13.8

Upper fourth +$3\ f_s$ = 96.7 + 19.5 = 116.2 ⇒ No obs. >116.2

extreme outliers

mild outliers are 74.1 113.5

Boxplot: