Chapter 1. Overview and Descriptive Statistics

Weiqi Luo (骆伟祺)
School of Data & Computer Science
Sun Yat-Sen University
Email：luoweiqi@mail.sysu.edu.cn

- **Textbook:**

  Jay L. Devore, Probability and statistics for engineering and the sciences (the 8^{th} Edition), 2010

- **References:**

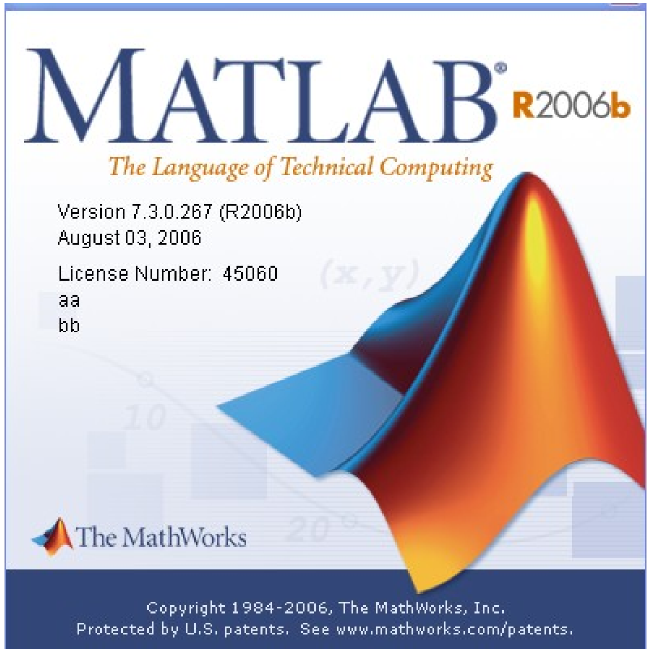
  1. Miller and Freund, "Probability and Statistics for Engineers" (the 7^{th} Edition), Publishing House of Electronics Industry, 2005.
  2. 盛骤、谢式千、潘承毅，《概率论与数理统计》第4版，高等教育出版社，2008

  Kai Lai Chung, "A Course in Probability Theory", (the 3^{rd} Edition), China Machine Press, 2010.

- MATLAB

A powerful software with various toolboxes, including

- **Statistics Toolbox**
- Image Processing Toolbox
- Signal processing Toolbox
- Robust Control Toolbox
- Curve Fitting Toolbox
- Fuzzy Logic Toolbox



…

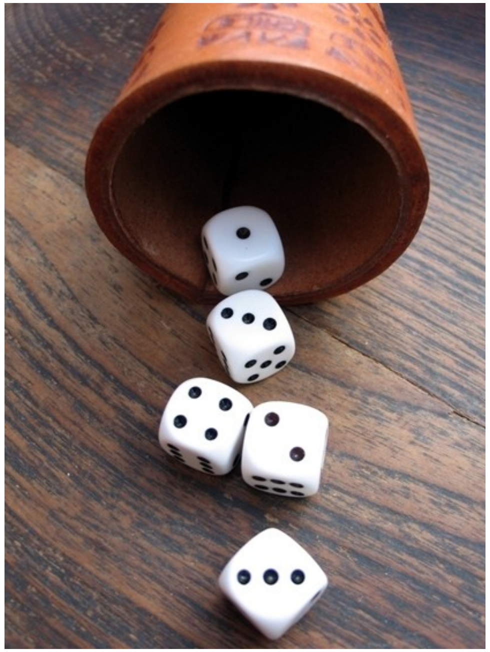
- Prerequisite Courses
  - SE-101 Advanced Mathematics
  - SE-103 Linear Algebra

- Successive Courses
  - SE-328 Digital Signal Processing
  - SE-343 Digital Image Processing
  - SE-352 Information Security
  - Pattern Recognition & Machine learning
  - etc.

What is Uncertainty?

- Uncertainty

  It can be assessed informally using the language such as "it is unlikely" or "probably".



玩 法	期号	最新开奖结果
双色球	2011099期	10 12 13 21 26 27 14
3 D	2011231期	8 2 7
36选7	2011231期	11 14 16 17 21 31 20
好彩1	2011231期	20 羊 秋 北
26选5	2011099期	03 04 14 15 19
快乐十分	2011082584期	17 12 18 16 20 04 07 11

This science came of gambling in 7^{th} century

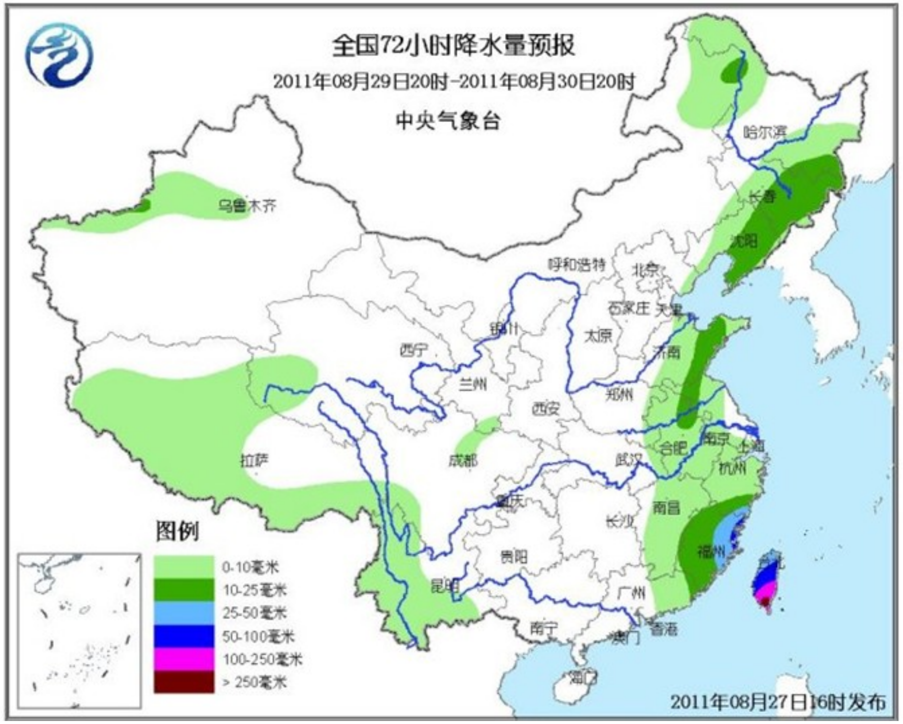
Why Study Probability & Statistics?

- **Probability** measures uncertainty formally, quantitatively. It is the mathematical language of uncertainty.

- **Statistics** show some useful information from the uncertain data, and provide the basis for making decisions or choosing actions.

Applications

- Weather Forecast



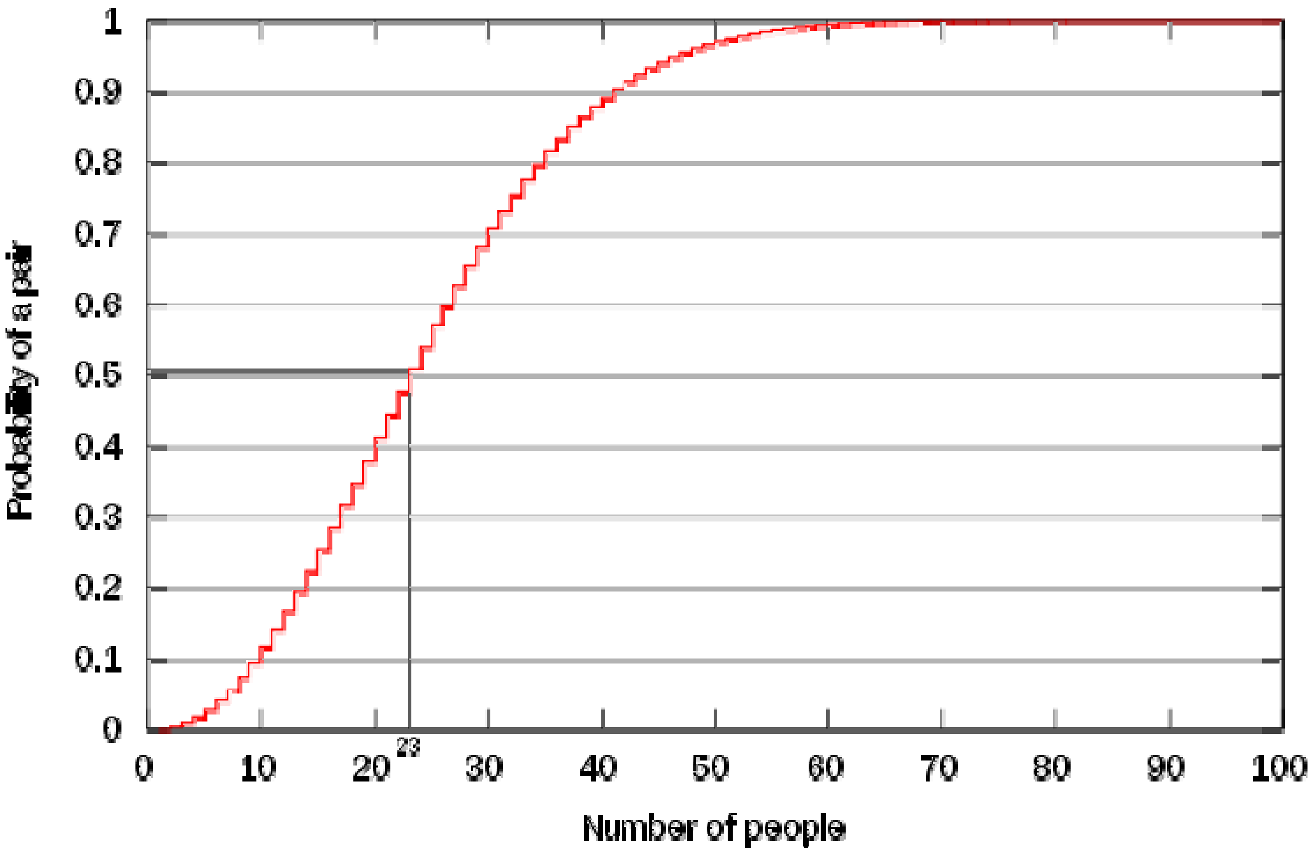
Applications

- In medical treatment

e.g. Relationship between smoking and lung cancer

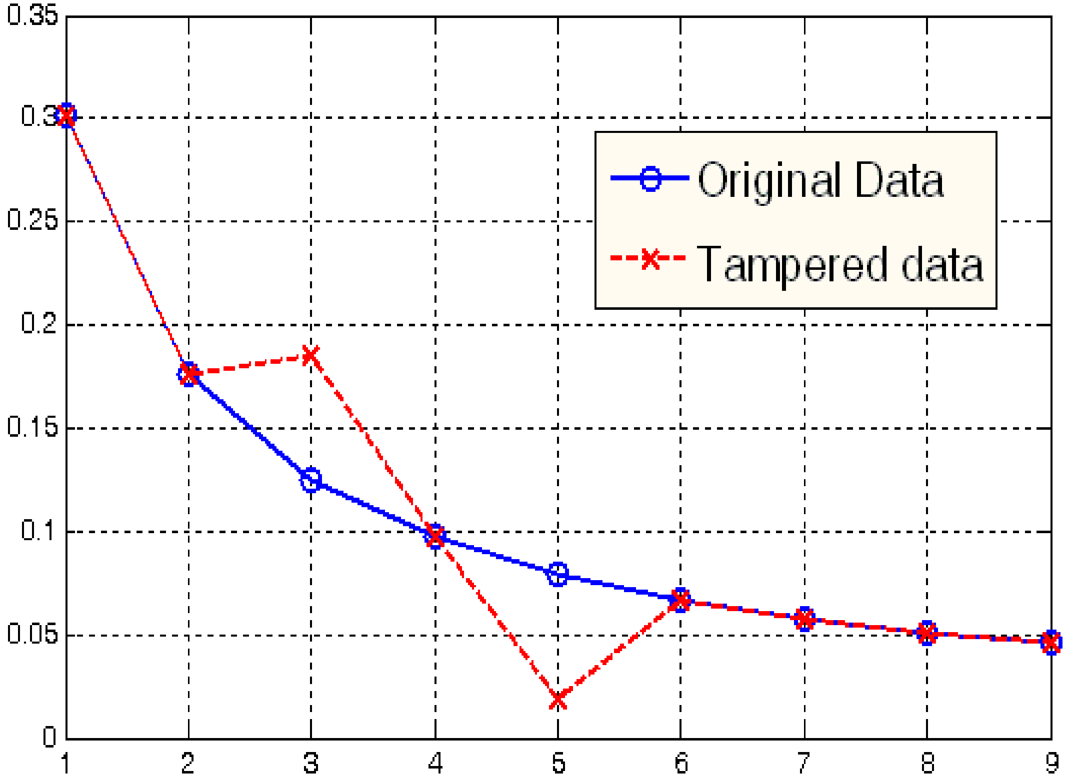
Applications

- Birthday Paradox (from Wikipedia)



Applications

- Benford's Law/ First Digit Law (from Wikipedia)



Accounting Forensics

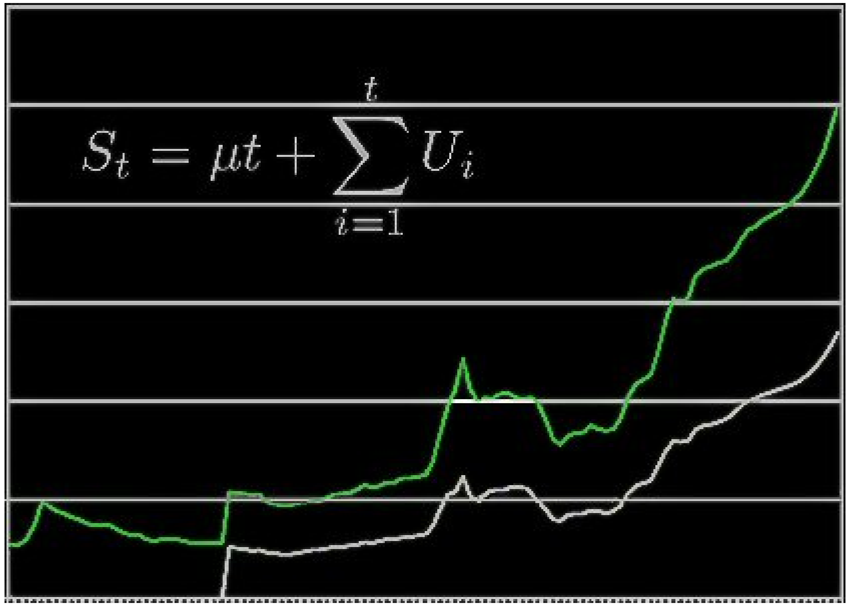
Multimedia Forensics

…

$$P(d) = \log_{10}(1+\frac{1}{d}), d = 1, 2, \ldots 9$$

Applications

- Time Series Analysis



•Economic Forecasting
•Sales Forecasting
•Budgetary Analysis
•Stock Market Analysis
•Process and Quality Control
•Inventory Studies
etc.

Applications

- More interesting applications in real life

Millon 2 one (概率知多少):
https://www.youtube.com/watch?v=3RngSBNw1AE

Chapter 1: Overview & Descriptive Statistics

- 1.1. Populations, Samples, and Processes
- 1.2. Pictorial and Tabular Methods in Descriptive Statistics
- 1. 3 Measures of Location
- 1.4. Measures of Variability

1.1. Populations, Samples, and Processes

- Population

  An investigation will typically focus on a ***well-defined*** collection of objects (units). A population is the set of all objects of interest in a particular study.

- Variables

  Any characteristic whose value (categorical or numerical) may change from one object to another in the population.

1.1. Populations, Samples, and Processes

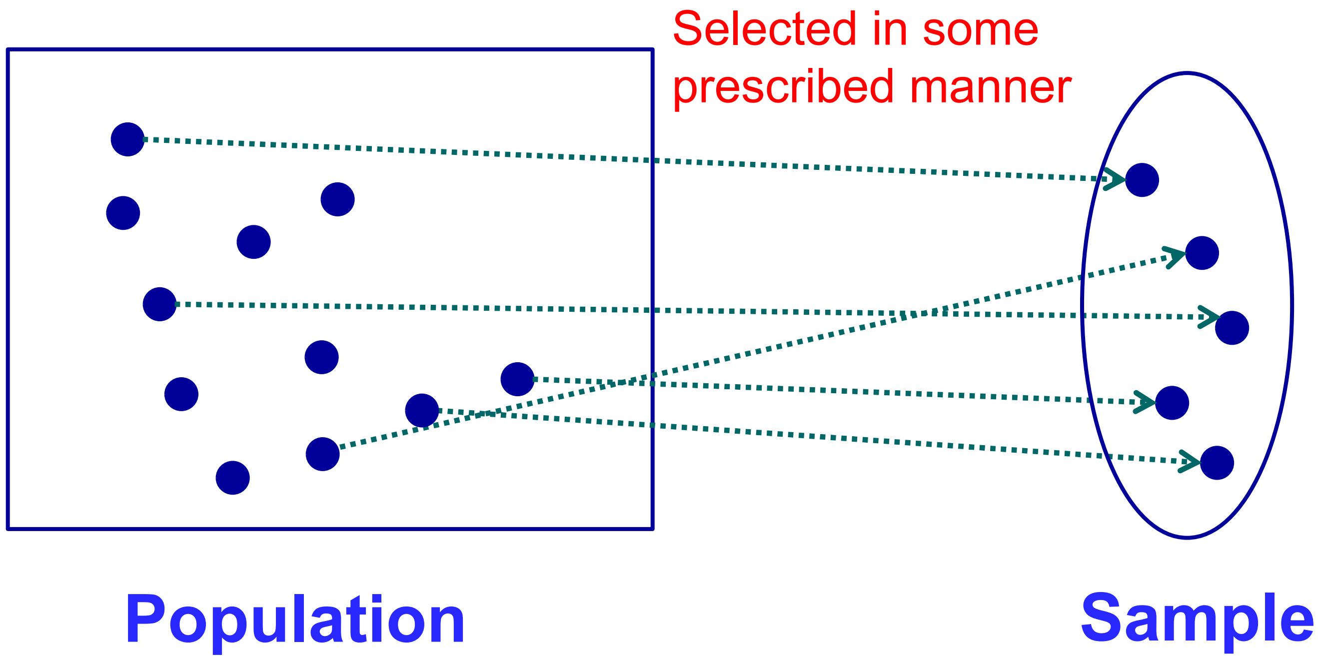
Examples of Populations, Objects and variables

Population	Unit / Object	Variables / Characteristics
All students currently in the class	Student	•Height •Weight •Hours of work per week •Right/left – handed
All Printed circuit boards manufactured during a month	Board	•Type of defects •Number of defects •Location of defeats
All campus fast food restaurants	Restaurant	•Number of employees •Seating capacity •Hiring/not hiring
All books in library	Book	•Replacement cost •Frequency of checkout •Repairs needs

1.1. Populations, Samples, and Processes

- Sample

  A subset of the population



1.1. Populations, Samples, and Processes

- According to the number of the variables under investigation, we have

- **Univariate** : a single variable, ***e.g.***
  the type of transmission, automatic or manual, on cars
- **Bivariate** : two variables, ***e.g.***
  the height & weight of the students
- **Multivariate** : more than two variables, ***e.g.***
  systolic blood pressure, diastolic blood pressure and serum cholesterol level for each patient

1.1. Populations, Samples, and Processes

- Descriptive statistics

  An investigator who has collected data may wish simply to summarize and describe important features of the data. (**descriptive statistics**)

- Visual techniques **(Sec. 1.2)**, ***e.g.***

  Stem-and-Leaf display, Dotplot & histograms

- Numerical summary measures **(Sec. 1.3, 1.4)**, ***e.g.***

  means, standard deviations & correlations coefficients

1.1. Populations, Samples, and Processes

- Example 1.1.

  Here is data on fundraising expenses as a percentage of total expenditures for a random sample of 60 charities:

6.1	12.6	34.7	1.6	18.8	2.2	3.0	2.2	5.6	3.8
2.2	3.1	1.3	1.1	14.1	4.0	21.0	6.1	1.3	20.4
7.5	3.9	10.1	8.1	19.5	5.2	12.0	15.8	10.4	5.2
6.4	10.8	83.1	3.6	6.2	6.3	16.3	12.7	1.3	0.8
8.8	5.1	3.7	26.3	6.0	48.0	8.2	11.7	7.2	3.9
15.3	16.6	8.8	12.0	4.7	14.7	6.4	17.0	2.5	16.2

Without any organization, it is difficult to get a sense of the data's most prominent features

1.1. Populations, Samples, and Processes

- Inferential statistics

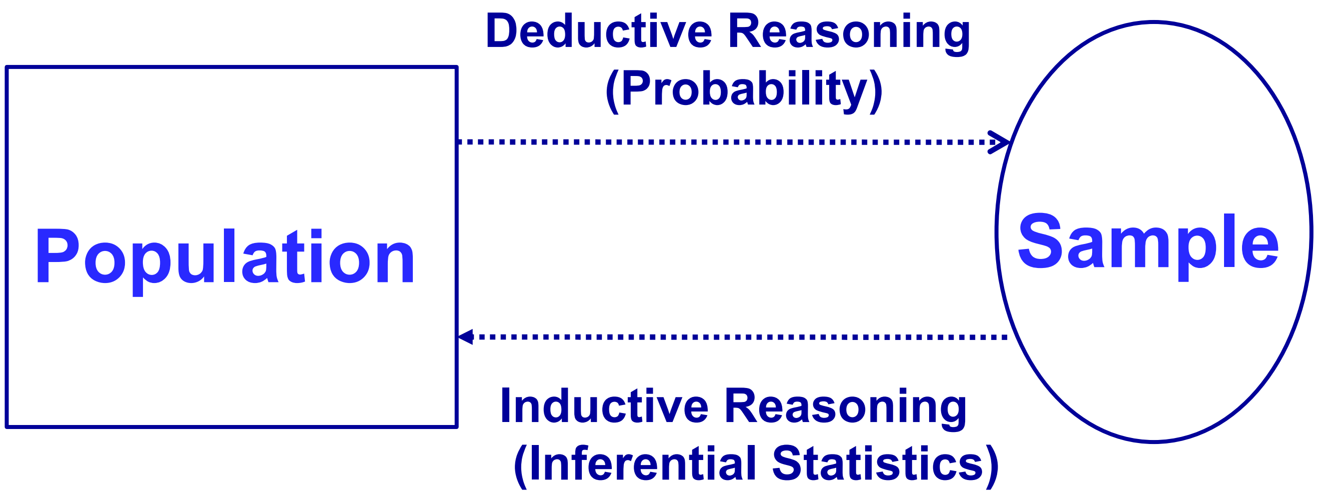
  Use sample information to draw some type of conclusion (make an inference of some sort) about the population.

- Point Estimation ---- Chapter 6
- Hypothesis testing ---- Chapter 8
- Estimation by confidence interval --- Chapter 7

  …

1.1. Populations, Samples, and Processes

- Probability & Statistics



The mathematical language is "Probability"

1.1. Populations, Samples, and Processes

- Collecting Data

  If data is not properly collected, an investigator may not be able to answer the questions under consideration with a reasonable degree of confidence.

- Methods for collecting data

➢ **Random sampling**: any particular subset of the specified size has the same chance of being selected

➢ **Stratified sampling**: entails separating the population units into non-overlapping groups and taking a sample from each one.

So on and so forth

- Descriptive Statistics

- Visual techniques **(Sec. 1.2)**

1. Stem-and-Leaf Displays
2. Dotplots
3. Histogram

- Numerical summary measures **(Sec. 1.3 & 1.4)**

1. Measures of location
2. Measure of variability

- Notation

Sample size: The number of observations in a single sample will often be denoted by $\boldsymbol{n}$.

Given a data set consisting of $\boldsymbol{n}$ observations on some variable x, the individual observations will be denoted by $\boldsymbol{x_1, x_2, x_3, \ldots, x_n}$

Stem-and-Leaf Displays

Suppose we have a numerical data set $x_1, x_2, x_3, \ldots, x_n$ for which each x_i consists of at least two digits.

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***.
2. List possible stem values in a vertical column.
3. Record the leaf for every observation beside the corresponding stem value.
4. Indicate the units for stems and leaves someplace in the display.

- Example:

Observations: 16%, 33%, 64%, 37%, 31% …

Stem-and-Leaf Display

Stem | Leaf
1 | 6
3 | 3 7 1 [or 3 | 1 3 7]
6 | 4

Stem: tens digit

Leaf: ones digit

- Example 1.6

0	4
1	1345678889
2	1223456666777889999
3	0112233344555666677777888899999
4	111222223344445566666677788888999
5	00111222223345566666777888899
6	01111244455666778

Stem: tens digit
Leaf: ones digit

Figure 1.4 Stem-and-leaf display for the percentage of binge drinkers at each of the 140 colleges

- A stem-and-leaf display conveys information about the following aspects of the data:

  - Identification of a typical or representative value
  - Extent of spread about the typical value
  - Presence of any gaps in the data
  - Extent of symmetry in the distribution of values
  - Number and location of peaks
  - Presence of any outlying values

- Example

Stem: Thousands and hundreds digits

Leaf: Tens and ones digits

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

⇨

6	|	435 464 433 470 … 904
7	|	051 005 011 040 … 209

Stem: Thousands digits

Leaf: Hundreds, tens and ones digits

- Example (repeated stems)

5H	5
5L	242330
4H	768896
4L	21421414444
3H	9696656

Stem: tens digit

Leaf: ones digit

=

5	242330 5
4	21421414444 768896
3	9696656

Stem: tens digit

Leaf: ones digit

Note: L: the leafs are 0, 1 , 2, 3 or 4
H: the leafs are 5, 6, 7, 8 or 9

- Dotplot

  the data set is reasonably small or there are relatively few distinct data values

- Each observation is represented by a dot above the corresponding location on a horizontal measurement scale.
- 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 & gaps.**

- Example 1.8

10.8	6.9	8.0	8.8	7.3	3.6	4.1	6.0	4.4	8.3
8.1	8.0	5.9	5.9	7.6	8.9	8.5	8.1	4.2	5.7
4.0	6.7	5.8	9.9	5.6	5.8	9.3	6.2	2.5	4.5
12.8	3.5	10.0	9.1	5.0	8.1	5.3	3.9	4.0	8.0
7.4	7.5	8.4	8.3	2.6	5.1	6.0	7.0	6.5	10.3

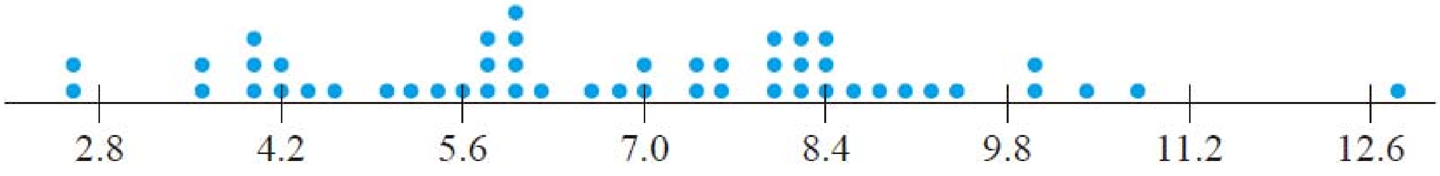


Figure 1.6 A dotplot of the data from Example 1.8

- Histogram

  Types of variables:

- **Discrete variable:** A variable is discrete if its set of possible values either is finite or else can be listed in an infinite sequence.
- **Continuous variable:** A variable is continuous if its possible values consist of an entire interval on the number line.

- Relative frequency of a value

Suppose, for example, that our data set consists of 200 observations on of courses a college student is taking this term. If 70 of these x values are 3, then

$$\text{relative frequency of a value} = \frac{\text{number of times the value occurs}}{\text{number of observations in the data set}}$$

frequency of the x value 3: 70

Relative frequency of the x value 3: $\frac{70}{200} = .35$

1.2 Pictorial and Tabular Method in Descriptive Statistics

Constructing a Histogram for Discrete Data

First, determine the frequency and relative frequency of each x value. Then mark possible x values on a horizontal scale. Above each value, draw a rectangle whose height is the relative frequency (or alternatively, the frequency) of that value.

- Example 1.9

Table 1.1 Frequency Distribution for Hits in Nine-Inning Games

Hits/Game	Number of Games	Relative Frequency	Hits/Game	Number of Games	Relative Frequency
0	20	.0010	14	569	.0294
1	72	.0037	15	393	.0203
2	209	.0108	16	253	.0131
3	527	.0272	17	171	.0088
4	1048	.0541	18	97	.0050
5	1457	.0752	19	53	.0027
6	1988	.1026	20	31	.0016
7	2256	.1164	21	19	.0010
8	2403	.1240	22	13	.0007
9	2256	.1164	23	5	.0003
10	1967	.1015	24	1	.0001
11	1509	.0779	25	0	.0000
12	1230	.0635	26	1	.0001
13	834	.0430	27	1	.0001
				19,383	1.0005

Why not 1?

1.2 Pictorial and Tabular Method in Descriptive Statistics

- Example 1.9

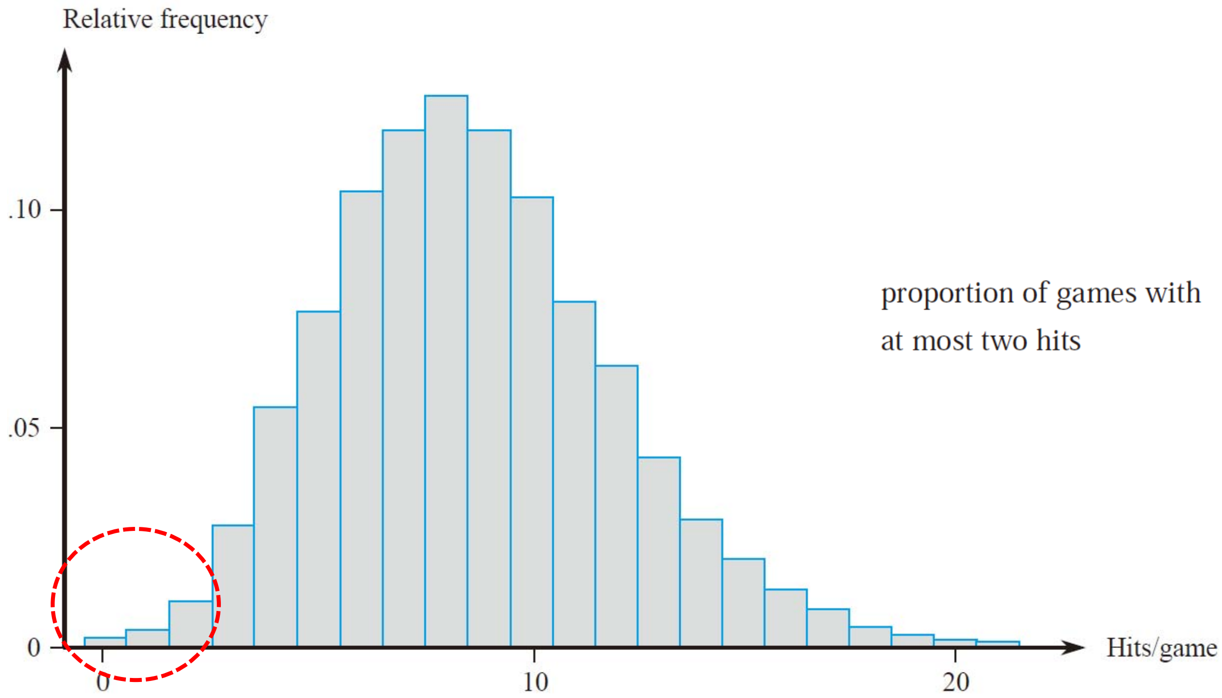


Figure 1.7 Histogram of number of hits per nine-inning game

$$\text{proportion of games with at most two hits} = \begin{matrix}\text{relative}\\ \text{frequency}\\ \text{for } x = 0\end{matrix} + \begin{matrix}\text{relative}\\ \text{frequency}\\ \text{for } x = 1\end{matrix} + \begin{matrix}\text{relative}\\ \text{frequency}\\ \text{for } x = 2\end{matrix}$$

$$= .0010 + .0037 + .0108 = .0155$$

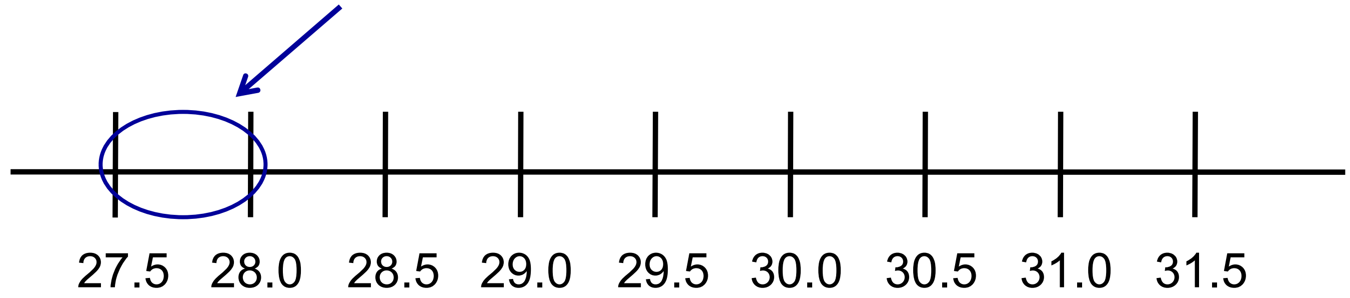
1.2 Pictorial and Tabular Method in Descriptive Statistics

- Continuous Case

p17. Support that we have 50 observations on x=fuel efficiency of an automobile (mpg), the smallest of which is 27.8 and the largest of which is 31.4

Equal or Unequal width

Class intervals : Continues → Discrete



Each observation is contained in exactly one class

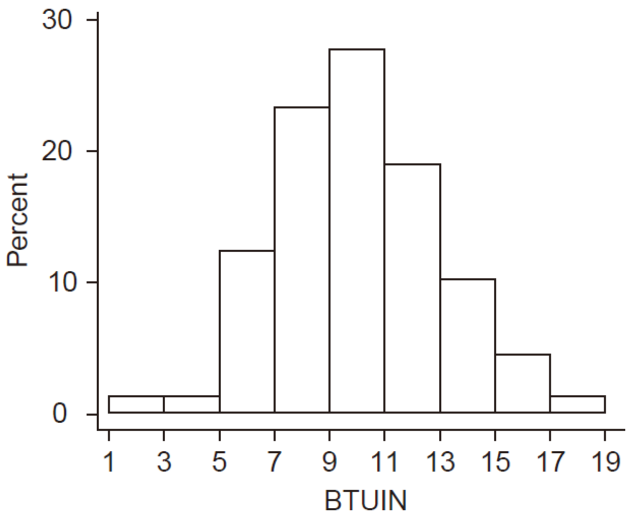
$$\text{number of classes} \approx \sqrt{\text{number of observations}}$$

Constructing a Histogram for Continuous Data: Equal Class Widths

Determine the frequency and relative frequency for each class. Mark the class boundaries on a horizontal measurement axis. Above each class interval, draw a rectangle whose height is the corresponding relative frequency (or frequency).

- Example 1.10

2.97	4.00	5.20	5.56	5.94	5.98	6.35	6.62	6.72	6.78
6.80	6.85	6.94	7.15	7.16	7.23	7.29	7.62	7.62	7.69
7.73	7.87	7.93	8.00	8.26	8.29	8.37	8.47	8.54	8.58
8.61	8.67	8.69	8.81	9.07	9.27	9.37	9.43	9.52	9.58
9.60	9.76	9.82	9.83	9.83	9.84	9.96	10.04	10.21	10.28
10.28	10.30	10.35	10.36	10.40	10.49	10.50	10.64	10.95	11.09
11.12	11.21	11.29	11.43	11.62	11.70	11.70	12.16	12.19	12.28
12.31	12.62	12.69	12.71	12.91	12.92	13.11	13.38	13.42	13.43
13.47	13.60	13.96	14.24	14.35	15.12	15.24	16.06	16.90	18.26



Class	1–<3	3–<5	5–<7	7–<9	9–<11	11–<13	13–<15	15–<17	17–<19
Frequency	1	1	11	21	25	17	9	4	1
Relative frequency	.011	.011	.122	.233	.278	.189	.100	.044	.011

1.2 Pictorial and Tabular Method in Descriptive Statistics

Equal-width classes may not be a sensible choice if there are some regions of the measurement scale that have a high concentration of data values and other parts where data is quite sparse.

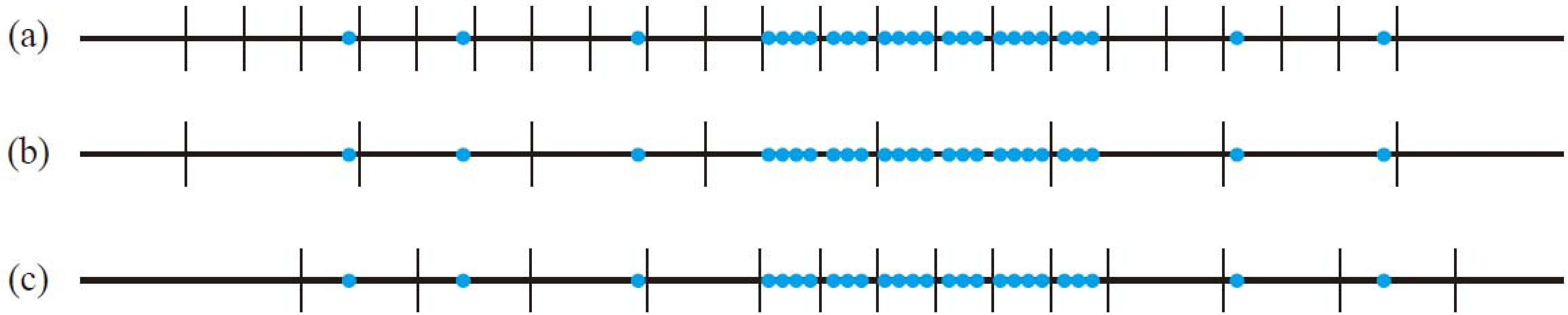


Figure 1.9 Selecting class intervals for “varying density” data: (a) many short equal-width intervals; (b) a few wide equal-width intervals; (c) unequal-width intervals

- Constructing a Histogram for Continuous Data : Equal (or Unequal) Class Widths

  Make sure that:

  $$\text{class width} \times \text{rectangle height (density)}$$
  $$= \text{relative frequency of the class}$$

- ✓ That is, the area of each rectangle is the relative frequency of the corresponding class.
- ✓ Furthermore, since the sum of relative frequencies should be 1, the total area of all rectangles in a density histogram is 1.

Constructing a Histogram for Continuous Data: Unequal Class Widths

After determining frequencies and relative frequencies, calculate the height of each rectangle using the formula

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

The resulting rectangle heights are usually called *densities,* and the vertical scale is the **density scale.** This prescription will also work when class widths are equal.

- Example 1.11

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

Class	2–<4	4–<6	6–<8	8–<12	12–<20	20–<30
Frequency	9	15	5	9	8	2
Relative frequency	.1875	.3125	.1042	.1875	.1667	.0417
Density	.094	.156	.052	.047	.021	.004

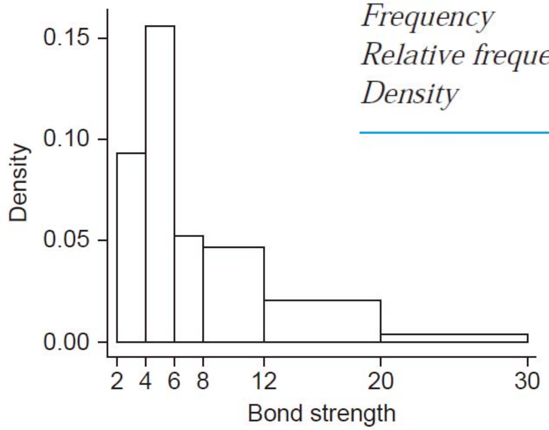


Figure 1.10 A Minitab density histogram for the bond strength data of Example 1.11

- Typical Histogram Shapes

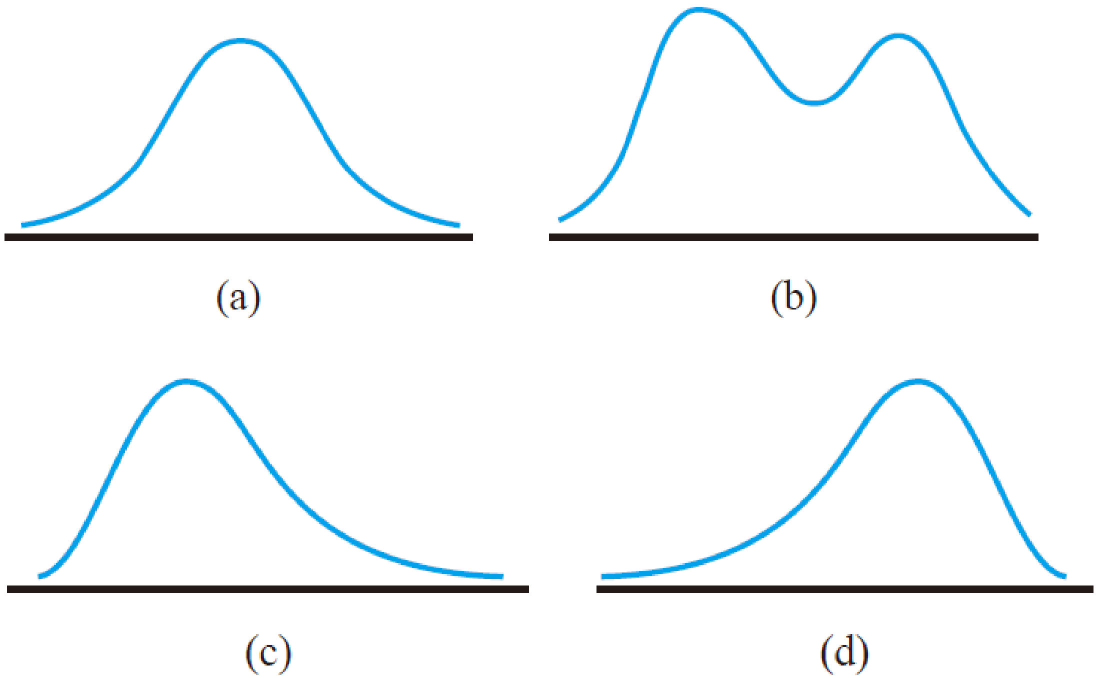


Figure 1.12 Smoothed histograms: (a) symmetric unimodal; (b) bimodal; (c) positively skewed; and (d) negatively skewed

- Qualitative Data

- ✓ Both a frequency distribution and a histogram can be constructed when the data set is *qualitative* (categorical) in nature.
- ✓ In some cases, there will be a natural ordering of classes—for example, freshmen, sophomores, juniors, seniors, graduate students
- ✓ In other cases the order will be arbitrary—for example, Catholic, Jewish, Protestant, and the like.
- ✓ With such categorical data, the intervals above which rectangles are constructed should have equal width

- Example 1.13

Table 1.2 Frequency Distribution for the School Rating Data

Rating	Frequency	Relative Frequency
A	478	.191
B	893	.357
C	680	.272
D	178	.071
F	100	.040
Don't know	172	.069
	2501	1.000

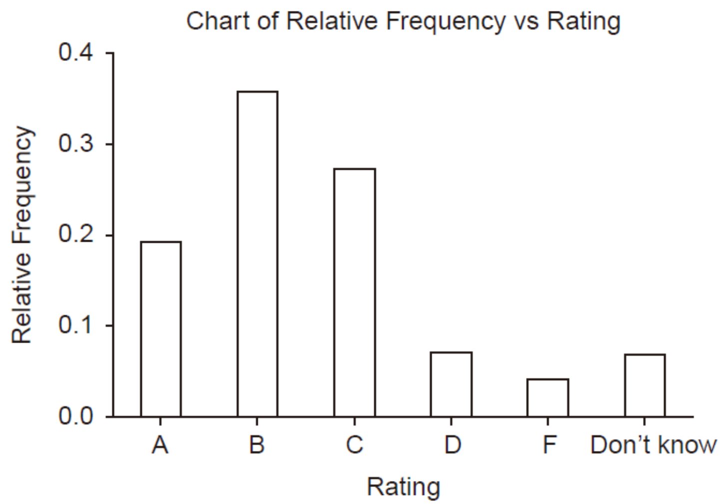


Figure 1.13 Histogram of the school rating data from Minitab

- Multivariate Data

The above mentioned techniques have been exclusively for situations in which each observation in a data set is either a single number or a single category.

Please refer to Chapters 11-14 for analyzing multivariate data sets.

Homework

- Ex. 14, 19, 23, 27

1.3 Measures of Location

- The Mean

➢ **Sample mean**: The sample mean of observations x_1, x_2, … , x_n is given by

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

➢ **Sample median**: The sample media is obtained by first ordering the n observations from smallest to largest. Then

$$\tilde{x} = \begin{cases} (\frac{n+1}{2})^{th}\ orderd\ value, & n\ is\ odd \\ ave.\ of\ (\frac{n}{2})^{th}\ \&\ (\frac{n}{2}+1)^{th}\ orded\ values, & n\ is\ even \end{cases}$$

1.3 Measures of Location

- Example 1.14 (Sample mean)

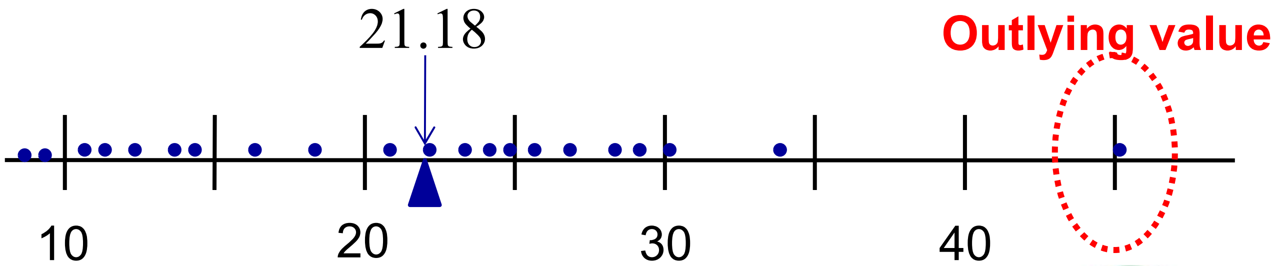
x_1=16.1 x_2=9.6 x_3=24.9 x_4=20.4 x_5=12.7 x_6=21.2 x_7=30.2

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

0H	**96 89**
1L	**27 03 40 46 18**
1H	**61 85**
2L	**49 04 12 33 42**
2H	**58 53 71 85**
3L	**02 24**
3H	
4L	
4H	**50**

$$\bar{x} = \frac{\sum x_i}{n} = \frac{444.8}{21} = 21.18$$



1.3 Measures of Location

- Example (Median)

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 valued 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

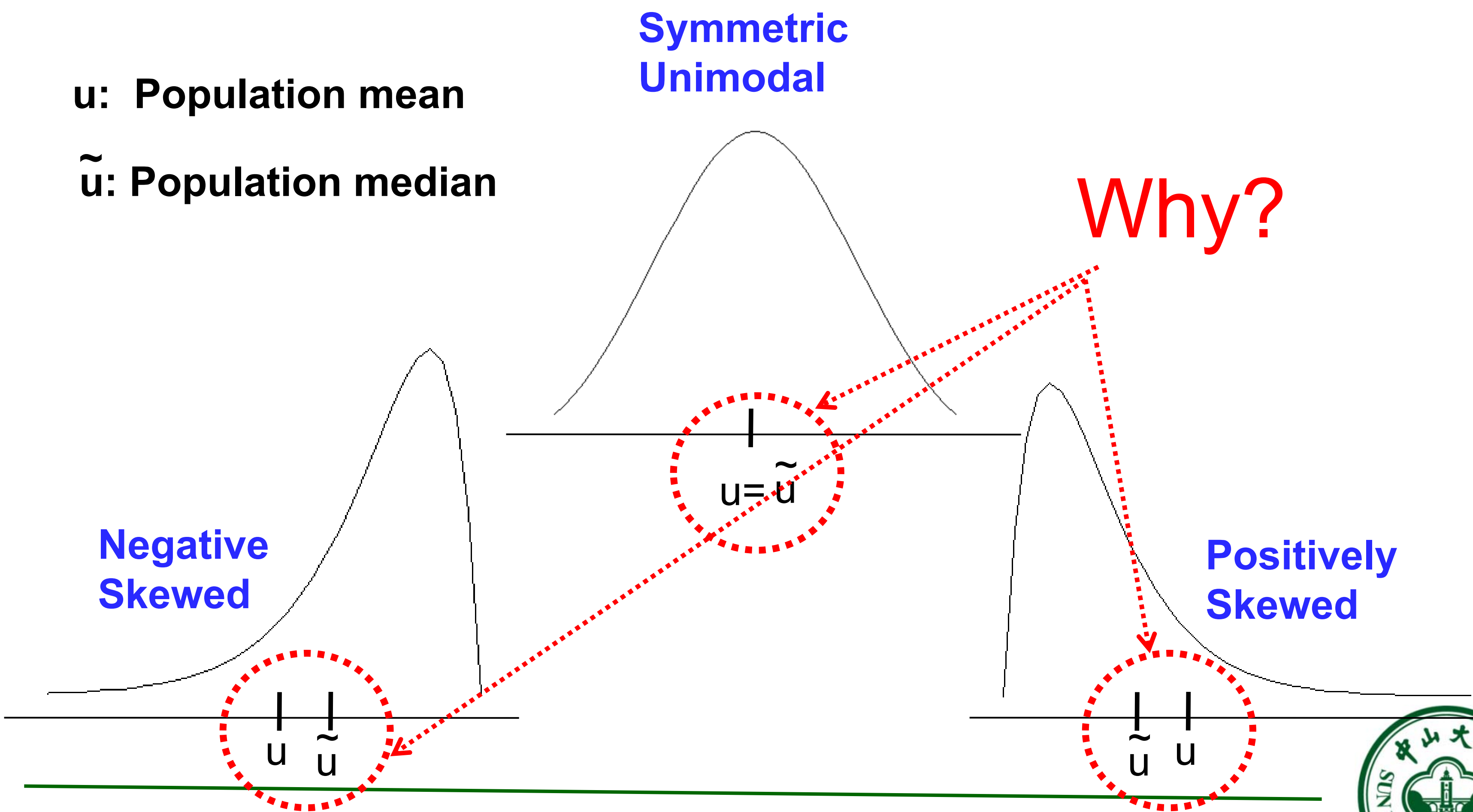
n = 12 is even, then the sample median is

$$(9.7 + 10.4) / 2 = 10.05$$

Note: the sample mean here is 139.3/12 = 11.61.

1.3 Measures of Location

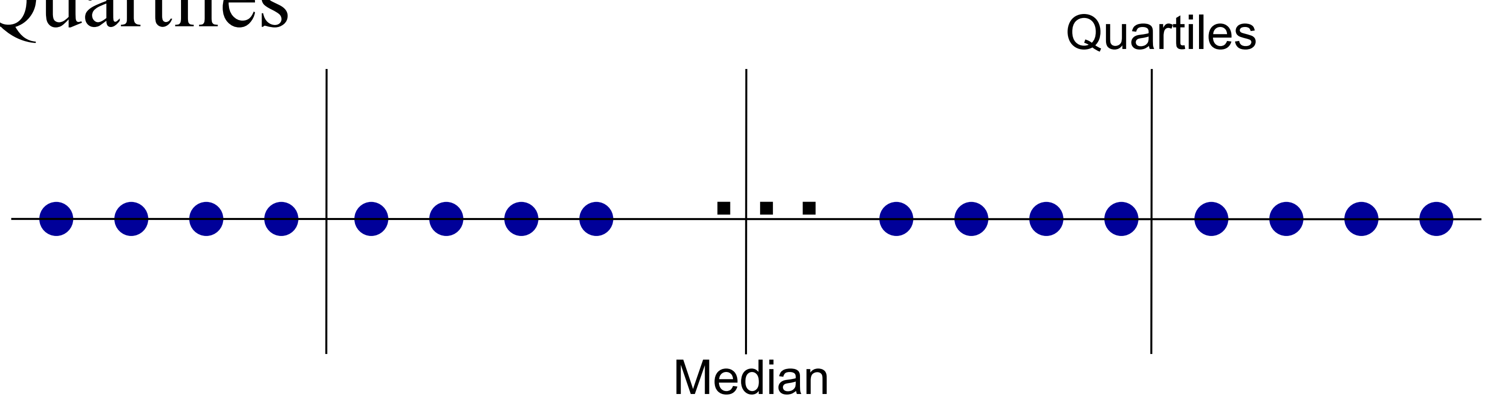
- Three different sharps for a population distribution



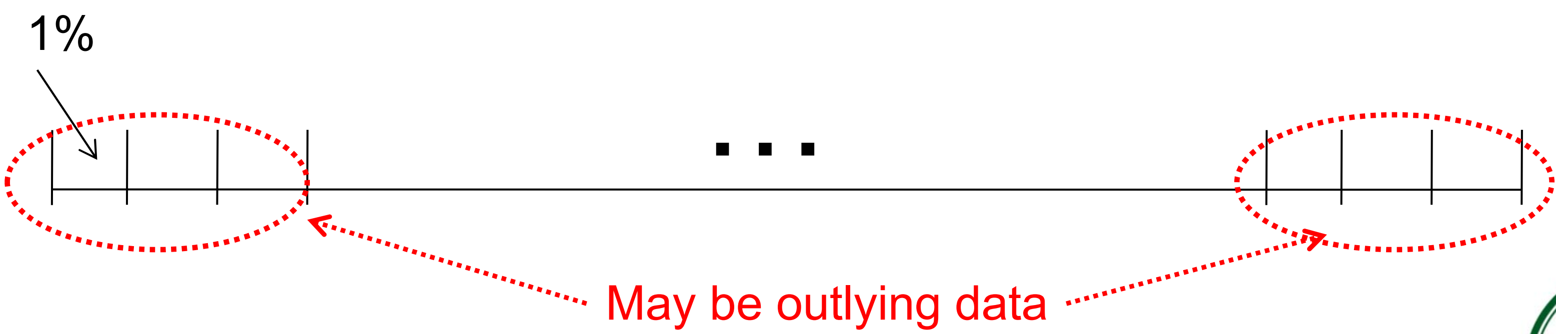
1.3 Measures of Location

- Other Measures of Location

Quartiles



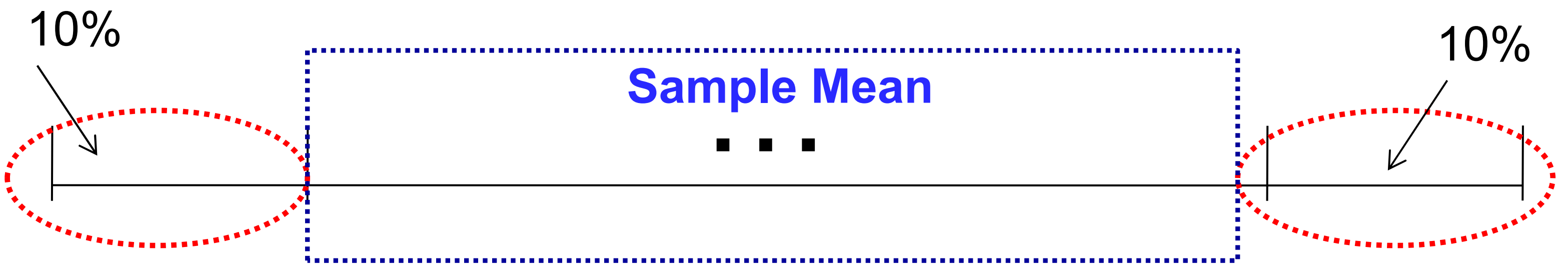
Percentiles



1.4 Measures of Location

- Trimmed Means

A trimmed mean is a compromise between **sample mean & sample median**. A 10% trimmed mean, for example, would be computed by eliminating the smallest 10% and the largest 10% of the sample and then averaging what is left over.

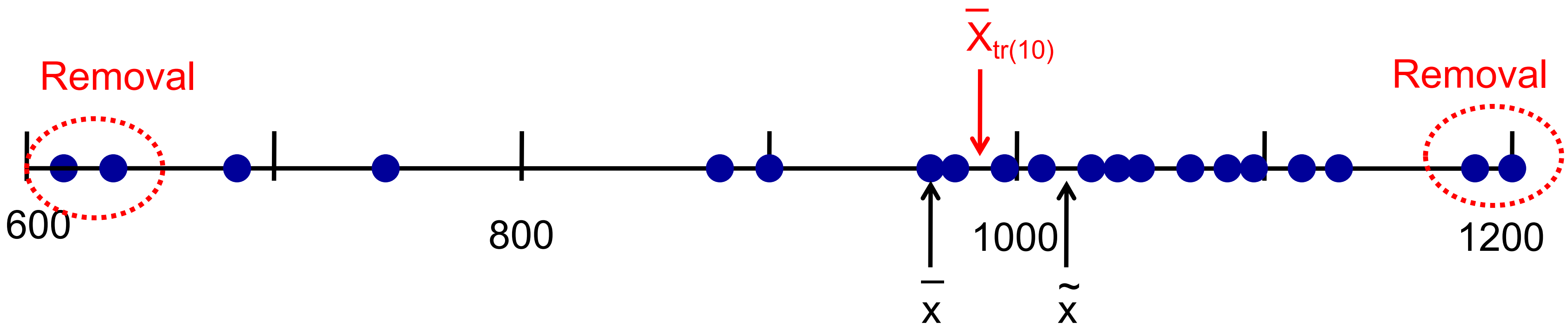


1.4 Measures of Location

- Example

612 623 666 744 883 898 964 970 983 1003

1016 1022 1029 1058 1085 1088 1122 1135 1197 1201



Note: Trimming proportion: 5%~25%

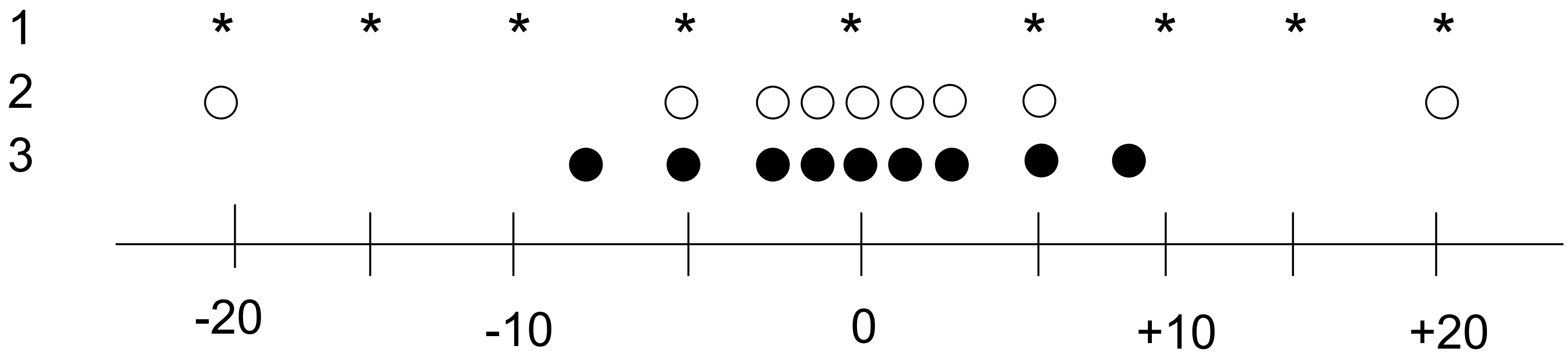
Homework

- Ex. 36, 40, 41

1.4 Measures of Variability

- Time error for three types of watches

  9 observations for each type



Q: Which type is the best ? And why?

1.4 Measures of Variability

- The Range

  The difference between the largest and smallest sample values. Refer to the previous example, type 1 and 2 have identical ranges, however, there is much less variability in the second sample than in the first.

- Deviations from the mean

  Measure 1: x_1-mean, x_2-mean, …, x_n-mean, then for all cases

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

1.4 Measures of Variability

- Sample variance

The sample variance, denoted by s^2, is given by

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

The sample standard deviation, denoted by s, is the square root of the variance s=sqrt(s^2).

Q1: $(x_i - \bar{x})^2$ vs. $|x_i - \bar{x}|$

Q2: n-1 vs. n

1.4 Measures of Variability

- Example

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0.684	0.9841	0.9685
2.54	0.8719	0.7602
0.924	-0.7441	0.5537
3.13	1.4619	2.1372
1.038	-0.6301	0.3970
0.598	-1.0701	1.1451
0.483	-1.1851	1.4045
3.52	1.8519	3.4295
1.285	-0.3831	0.1468
2.65	0.9819	0.9641
1.497	-0.1711	0.0293

$$\sum x_i = 18.349$$

$$\bar{x} = \frac{18.349}{11} = 1.6681$$

$$\sum \left(x_i - \bar{x}\right) = -0.0001 \approx 0$$

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

$$= 11.9359$$

$$s^2 = \frac{S_{xx}}{n-1} = \frac{11.9359}{11-1} = 1.19359$$

$$s = \sqrt{1.19359} = 1.0925$$

1.4 Measures of Variability

- Population variance

We will use σ^2 to denote the population variance and σ to denote the population standard deviation. When the population is finite and consists of N values,

$$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2 / N$$

1.4 Measures of Variability

- Consider a population with just 3 elements $\{1,2,3\}$
- The mean of the population is $\mu = \frac{1+2+3}{3} = 2$
- And the variance

$$\sigma^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = \frac{2}{3}$$

- Suppose all we can take is a sample of 2 elements taken with repetition to learn about the population.
  - We would like the sample to accurately estimate the mean and variance values of the population.

1.4 Measures of Variability

Possible Samples of Size Two	Sample mean $\bar{x}$	s^2 using n =2	s^2 using n – 1 = 1
{1,1}	1	0/2	0/1
{2,2}	2	0/2	0/1
{3,3}	3	0/2	0/1
{1,2}	1.5	.5/2 = .25	.5/1 = .5
(2,1)	1.5	.5/2 = .25	.5/1 = .5
{1,3}	2	2/2 = 1.0	2/1 = 2
(3,1)	2	2/2 = 1.0	2/1 = 2
{2,3}	2.5	.5/2 = .25	.5/1 = .5
(3,2)	2.5	.5/2 = .25	.5/1 = .5
Average of Sample Statistics	2	1/3	2/3 Better estimation!

1.4 Measures of Variability

- An alter expression for the numerator of s^2

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

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

Be care of the rounding errors when using the two different expressions

- If $y_1 = x_1 + c$, $y_2 = x_2 + c, \ldots, y_n = x_n + c$, then $s_y^2 = s_x^2$
- If $y_1 = cx_1, y_2 = cx_2, \ldots, y_n = cx_n$, 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.

1.4 Measures of Variability

- Boxplots

  Describe several of a data set's most prominent features:

- center;
- spread;
- extent and nature of any departure from symmetry ;
- identification of "outliers ", observations that lie unusually far from the main body of the data.

1.4 Measures of Variability

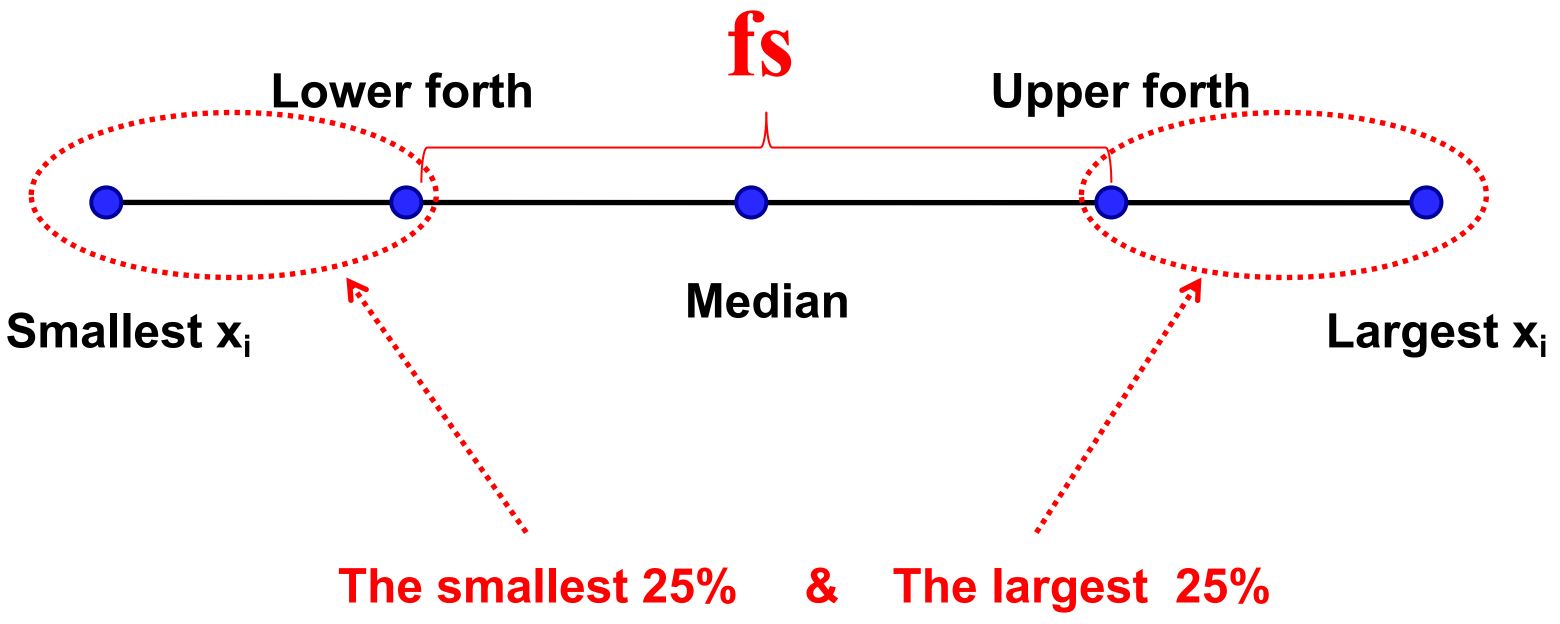
- Fourth Spread

  Order the n observations from smallest to largest and separate the smallest half from the largest half; the median is included in both halves if n is odd. Then the lower fourth is the median of the smallest half and the upper fourth is the median of the largest half. A measure of spread that is resistant to outliers is the fourth spread f_s, given by

$$f_s = \text{upper fourth} - \text{lower fourth}$$

1.4 Measures of Variability

- The simplest boxplot is based on the 5-number summary



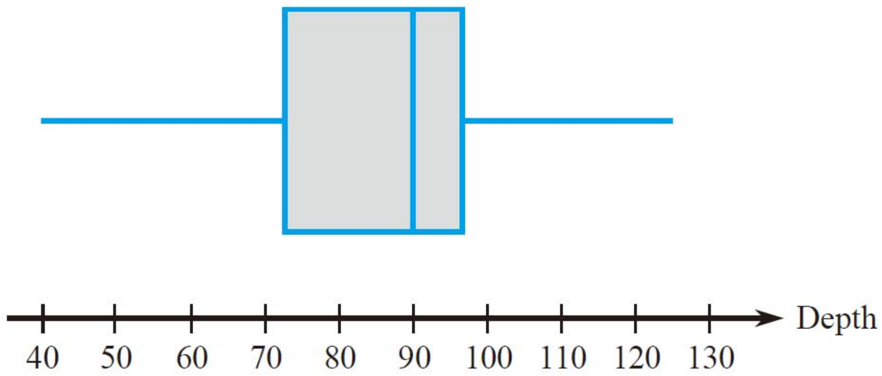
1.4 Measures of Variability

- Example 1.19

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

The five-number summary is as follows:

smallest $x_i = 40$ lower fourth $= 72.5$ $\tilde{x} = 90$ upper fourth $= 96.5$
largest $x_i = 125$



1.4 Measures of Variability

- A boxplot can be embellished to indicate explicitly the presence of outliers.

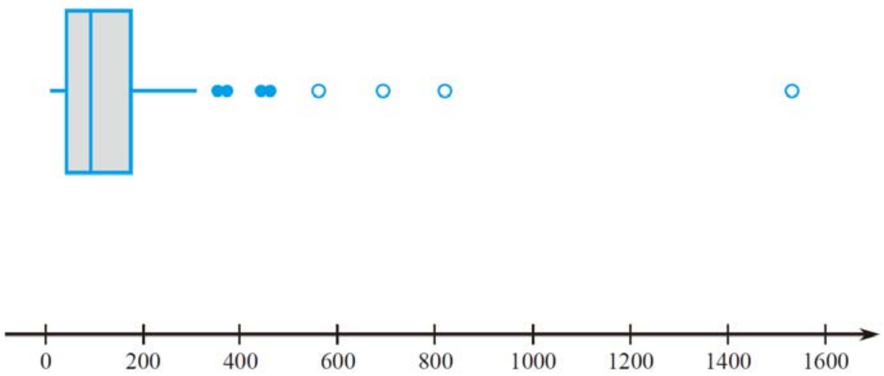
- **Outlier:** Any observation father than 1.5 fs from the closest fourth is an outlier.
- **Extreme:** An outlier is extreme if it is more than 3 fs from the nearest fourth
- **Mild:** An outlier is mild if it is in the range of (1.5fs , 3fs] from the nearest fourth.

- Example 1.20

9.69	13.16	17.09	18.12	23.70	24.07	24.29	26.43
30.75	31.54	35.07	36.99	40.32	42.51	45.64	48.22
49.98	50.06	55.02	57.00	58.41	61.31	64.25	65.24
66.14	67.68	81.40	90.80	92.17	92.42	100.82	101.94
103.61	106.28	106.80	108.69	114.61	120.86	124.54	143.27
143.75	149.64	167.79	182.50	192.55	193.53	271.57	292.61
312.45	352.09	371.47	444.68	460.86	563.92	690.11	826.54
1529.35							

$\tilde{x} = 92.17$ lower $4^{th} = 45.64$ upper $4^{th} = 167.79$

$f_s = 122.15$ $1.5f_s = 183.225$ $3f_s = 366.45$



Homework

- Ex. 44, 54