1 Overview and Descriptive Statistics

- 1.1. Populations, Samples, and Processes

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 defects
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

- 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.***

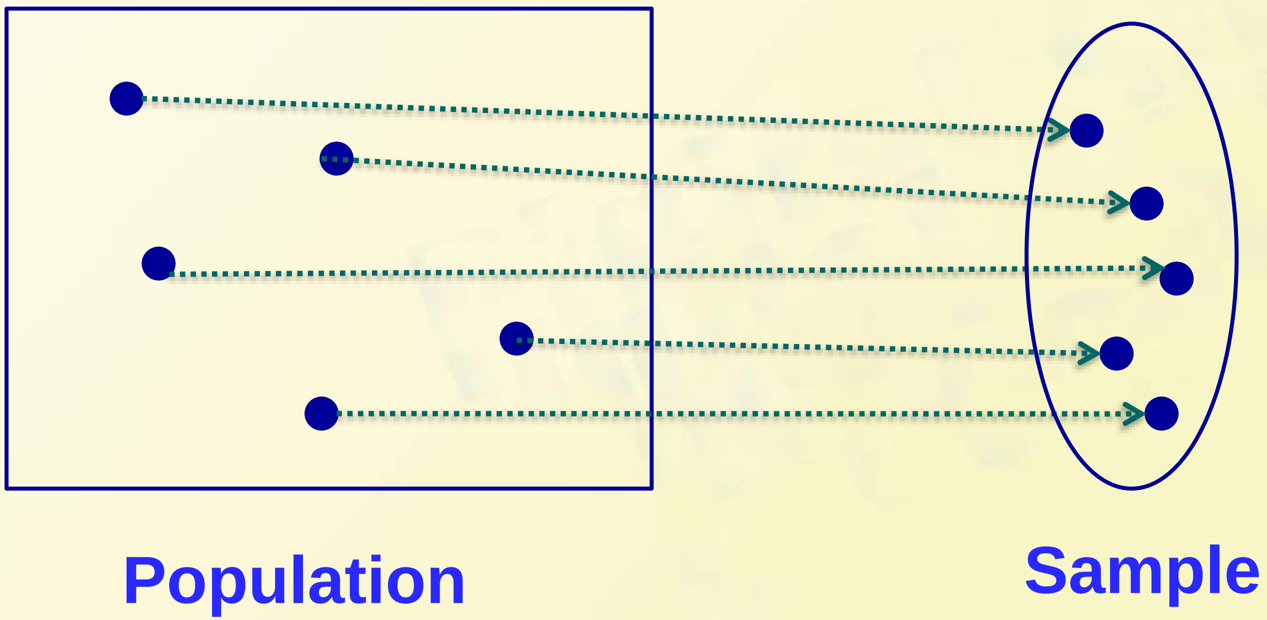
  systolic blood pressure, diastolic blood pressure and serum cholesterol level for each patient

1.1. Populations, Samples, and Processes

- **Sample**

  A **subset** of the population

  A sample is selected from the population in some prescribed manner



- **Example 1**

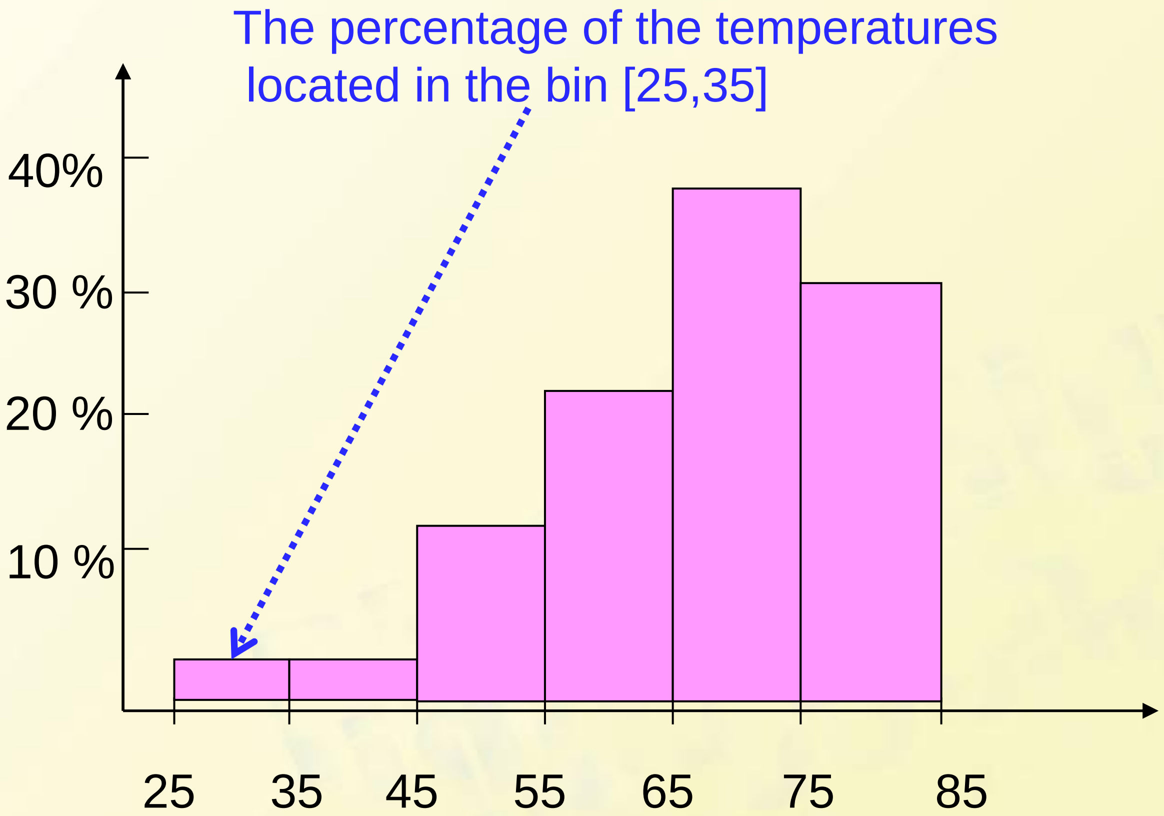
Here is data consisting of observations on x = O-ring temperature for each test firing or actual launch of the shuttle rocket engine.

84	49	61	40	83	67	45	66	70	69	80	58
68	60	67	72	73	70	57	63	70	78	52	67
53	67	75	61	70	81	76	79	75	76	58	31

Without any organization, it is different to get a sense of what a typical or representative temperature might be!

1.1. Populations, Samples, and Processes

- Normalized Histogram



According the histogram, we can find how the values of temperature are distributed along the measurement scale.

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.** This entails using methods from **descriptive statistics**

- **Graphical methods (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

- **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

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. Techniques for generalizing from a sample to a population are called **inferential statistics**

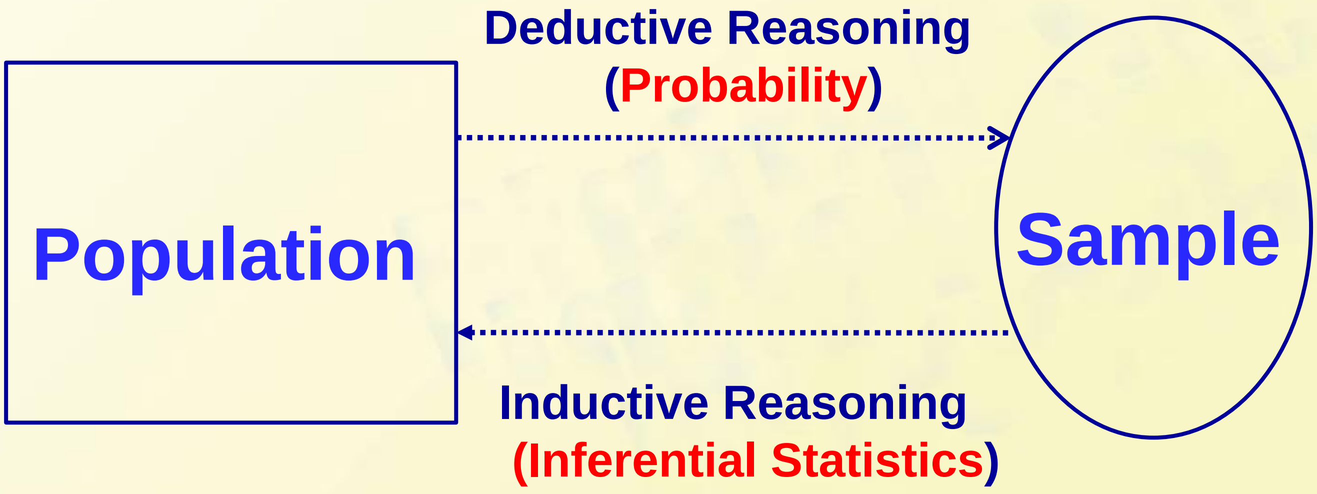
- Point Estimation ---- Chapter 6
- Hypothesis testing ---- Chapter 8
- Estimation by confidence interval --- Chapter 7

…

1.1. Populations, Samples, and Processes

- **Relation between Probability and Statistics**

Probability reasons from the population to the sample(deductive reasoning), whereas inferential statistics reasons from the sample to the population



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**: each element of population has an equal chance of been selected
- **Stratified sampling**: the population is divided into subpopulation(Strata) and random samples are taken of each stratum.

So on and so forth

1.2 Pictorial and Tabular Method in Descriptive Statistics

1.2 Pictorial and Tabular Method in Descriptive Statistics

- Notation

Sample size: The number of observations in a single sample will often **be denoted by *n*.**

Given a data set consisting of ***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 2:**

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

- **Exercise 1:**

Draw the stem-and-leaf display for data:

18，8，10，43，5，30，10，22，6，27，25，58，14，18，30，41；

Stem	Leaf
0	5 6 8
1	0 0 4 8 8
2	2 5 7
3	0 0
4	1 3
5	8

Stem: tens digit

Leaf: ones digit

- Example 4

Given some four digits, draw stem-and-left display

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

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

Stem: Thousands digits

Leaf: Hundreds, tens and ones digits

Notice that a stem choice here of either a single digit (6 or 7) or three digits (647,…,728) would yield an uninformative display.

- **Exercise 2:**

Draw the stem-and-leaf display for data:

7435 6328 7439 5468

- 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

- **Dotplot**

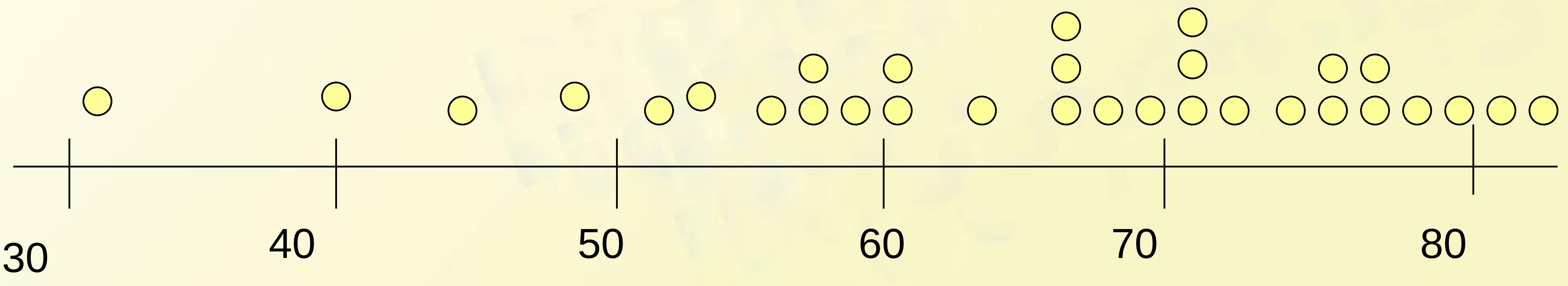
  the data set is reasonably small or there are relatively few distinct data values

- 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 5

84	49	61	40	83	67	45	66	70	69	80	58
68	60	67	72	73	70	57	63	70	78	52	67
53	67	75	61	70	81	76	79	75	76	58	31



- 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.

Constructing a Histogram for Discrete Data

Three Steps:

1. Determine the frequency (or relative frequency) of each x value.
2. Mark possible x values on a horizontal scale.
3. Draw a rectangle whose height is the frequency (or relative frequency) of the value.

- **Example**

Suppose that our data set consists of **200 observations** on x = the number of major defects in a new car of a certain type. **If 70 of these x are 1**, then

Frequency of the x value 1 : 70

Relative frequency of the x value 1: 70 / 200 = 0.35

Note:

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

Example

Given the scores of students, draw the histogram.

72,65,75,85,89,95,100,63,73,78

- Example 1.9

hits/game	number of games	relative frequency	hits/game	number of games	relative frequency
0	20	0. 001	14	569	0. 0294
1	72	0. 0037	15	393	0. 0203
2	209	0. 0108	16	253	0. 0131
3	527	0. 272	17	171	0. 0088
4	1048	0. 541	18	97	0. 005
5	1457	0. 752	19	53	0. 0027
6	1988	0. 1026	20	31	0. 0016
7	2256	0. 1164	21	19	0. 001
8	2403	0. 124	22	13	0. 0007
9	2256	0. 1164	23	5	0. 0003
10	1967	0. 1015	24	1	0. 0001
11	1509	0. 0779	25	0	0
12	1230	0. 0635	26	1	0. 0001
13	834	0. 043	27	1	0. 0001

1.2 Pictorial and Tabular Method in Descriptive Statistics

- Example 1.9

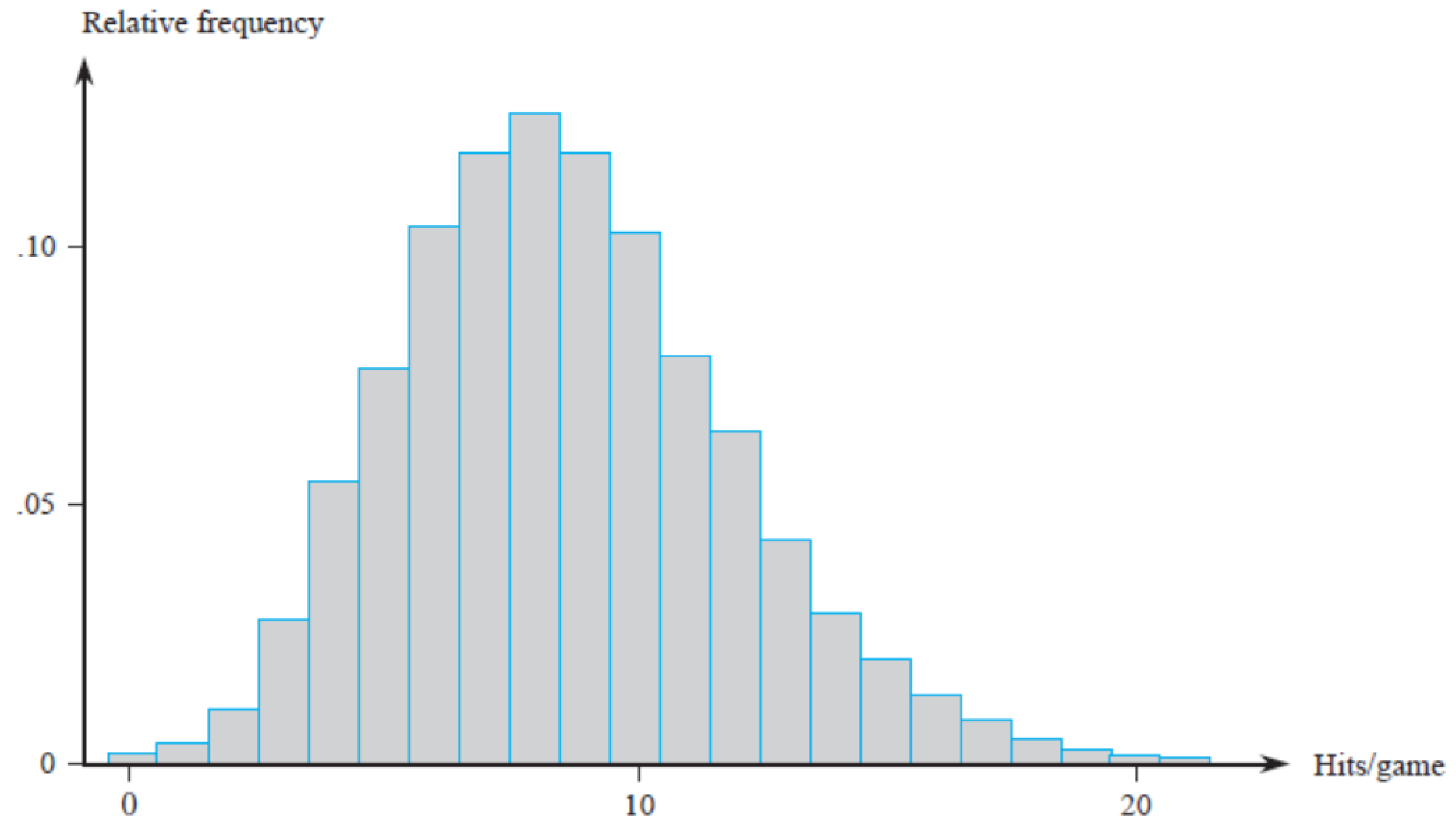


Figure 1.7 Histogram of number of hits per nine-inning game

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

Similar to the discrete case

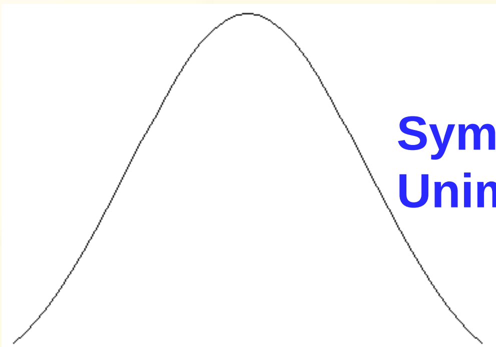
Make sure that:

class width $\times$ rectangle height (density)

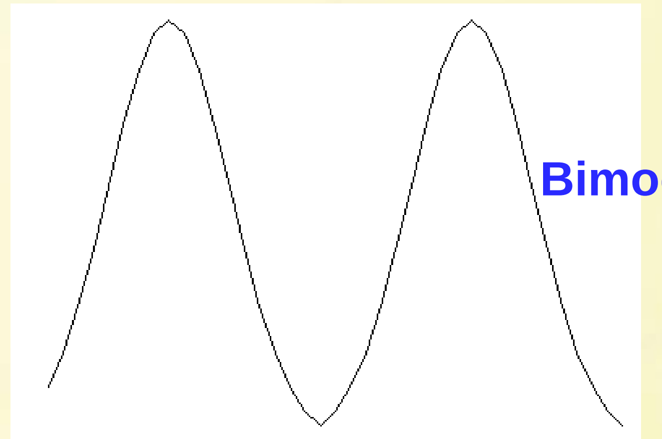
= relative frequency of the class

1.2 Pictorial and Tabular Method in Descriptive Statistics

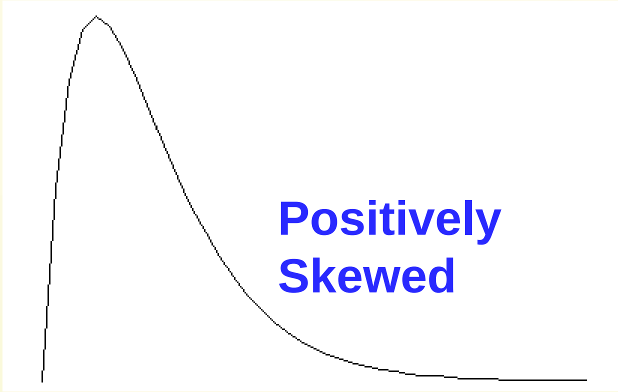
- ## Typical Histogram Shapes



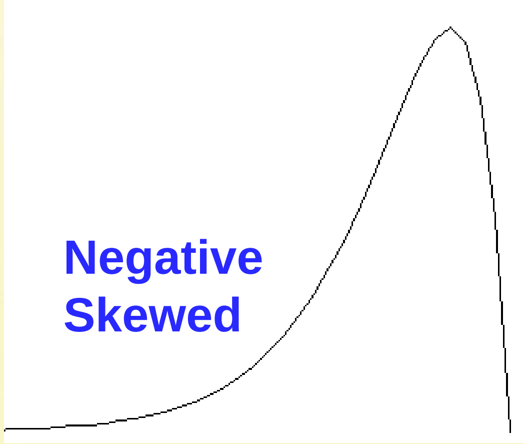
Symmetric Unimodal



Bimodal



Positively Skewed



Negative Skewed

- 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.

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.

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

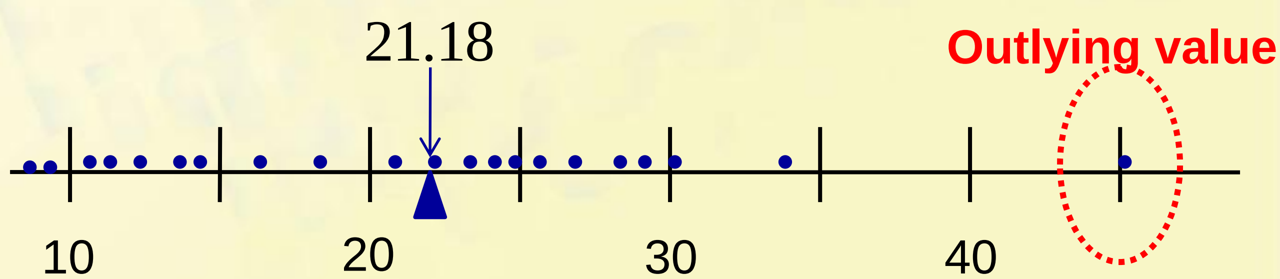
1.3 Measures of Location

- 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**

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



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.

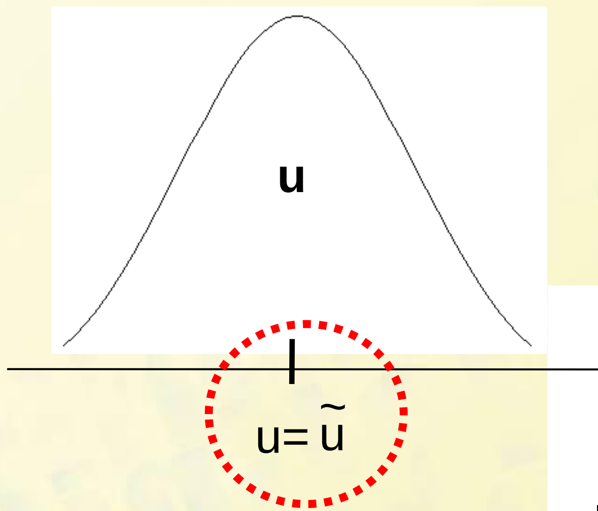
1.3 Measures of Location

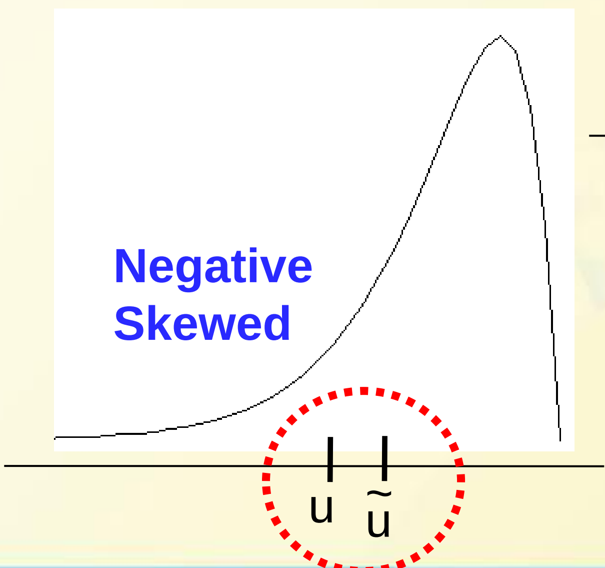
- Three different sharps for a population distribution

u: Population mean

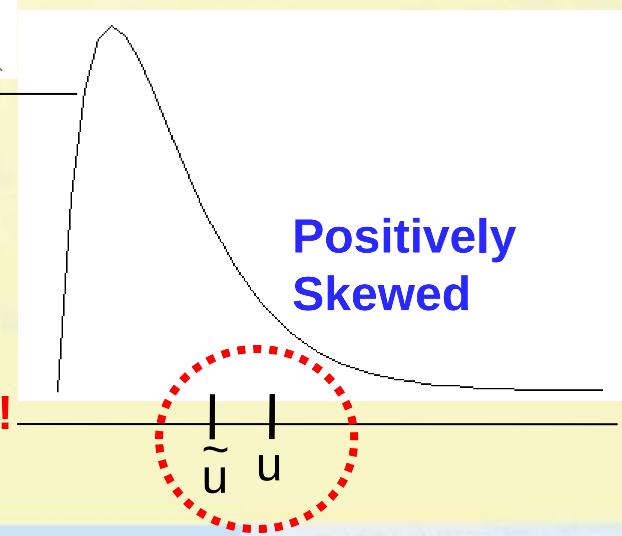
$\tilde{u}$: Population median

Symmetric Unimodal





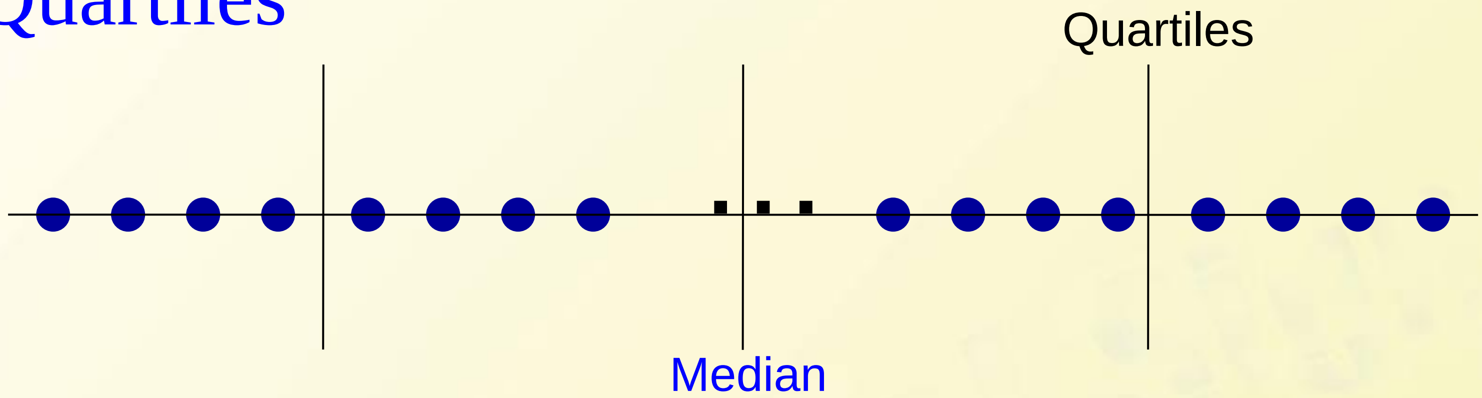
u and $\tilde{u}$ will not generally be identical!



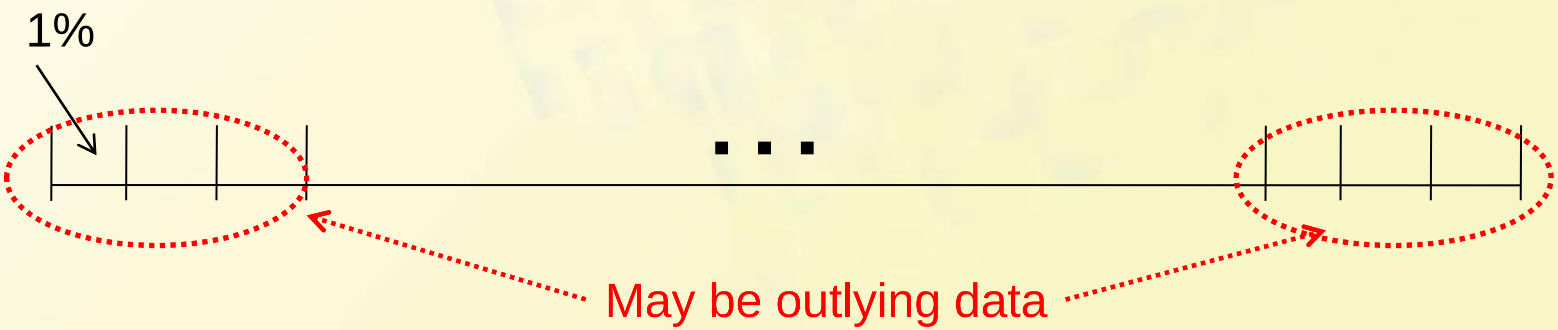
1.3 Measures of Location

- **Other Measures of Location**

Quartiles



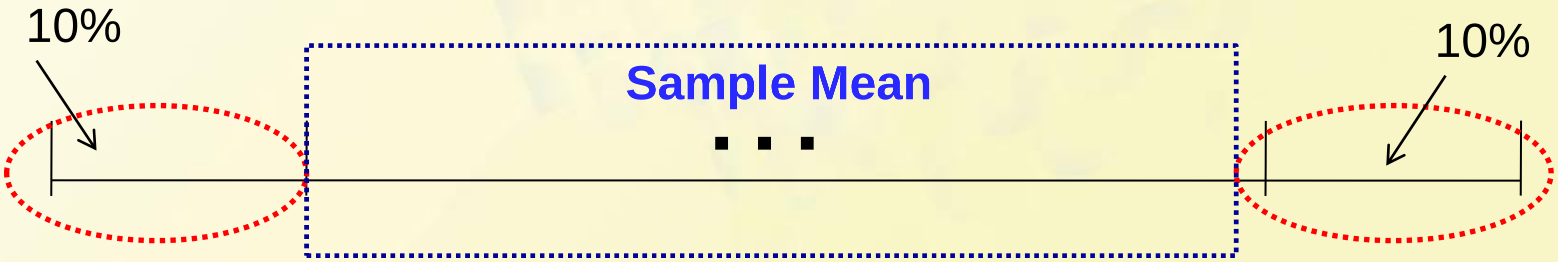
Percentiles



▪ 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

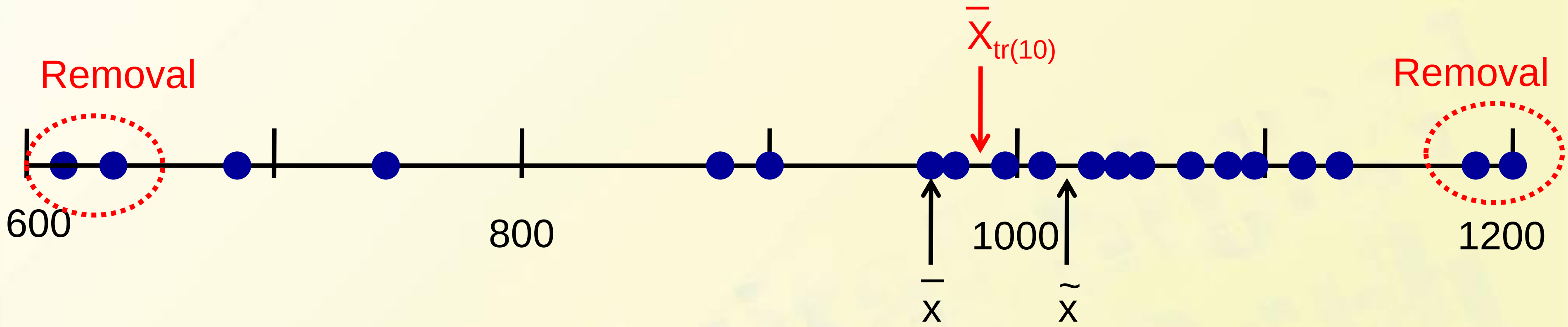
612 623 666 744 883 898 964 970 983 1003

1016 1022 1029 1058 1085 1088 1122 1135 1197 1201

Find mean, median and 10% trimmed Means

1.4 Measures of Location

Solution:



Note: Trimming proportion: 5%~25%

1.4 Measures of Variability

- **The Range**

  The difference between the largest and smallest sample values.

- **Deviations from the mean**

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

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

Measure 2:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0$$

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

Considering unbiased estimatem, here divide by n-1 Artificially

1.4 Measures of Variability

Example

Given 11 data:

0.684
2.54
0.924
3.13
1.038
0.598
0.483
3.52
1.285
2.65
1.497

Find sample variance and sample standard deviation

1.4 Measures of Variability

- **Solution:**

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

$$\sum x_i = 18.349$$

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

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

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

$$= 11.9359$$

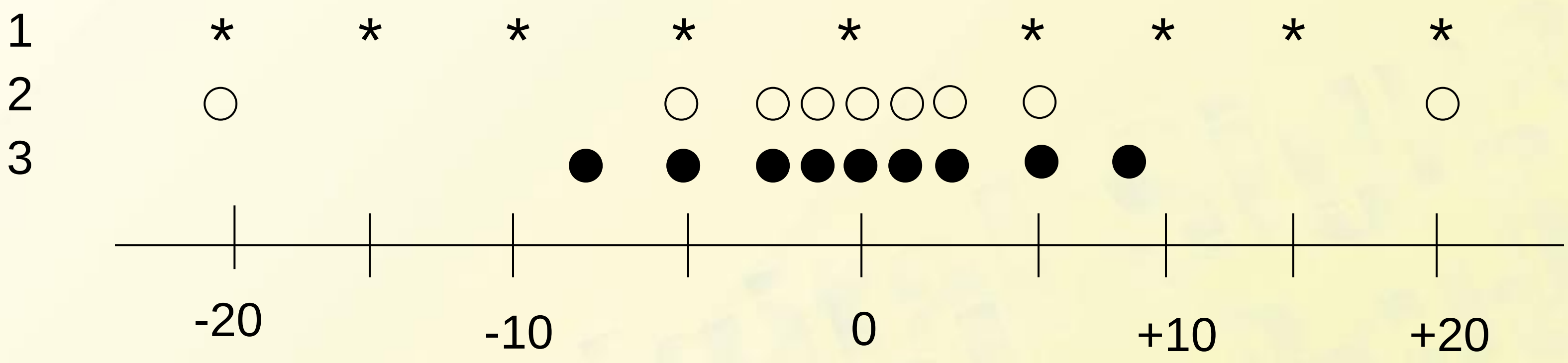
$$s^2 = \frac{S_{xx}}{n-1} = \frac{11.9359}{11-1} = 1.19359$$

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

1.4 Measures of Variability

- Time error for three type of watches

  9 observations for each type



Q: Which type is the best ? And why?

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 estimate!**

1.4 Measures of Variability

A Computing Formula for s^2

It is best to obtain s^2 from statistical software or else use a calculator that allows you to enter data into memory and then view s^2 with a single keystroke. If your calculator does not have this capability, there is an alternative formula for S_{xx} that avoids calculating the deviations.

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

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

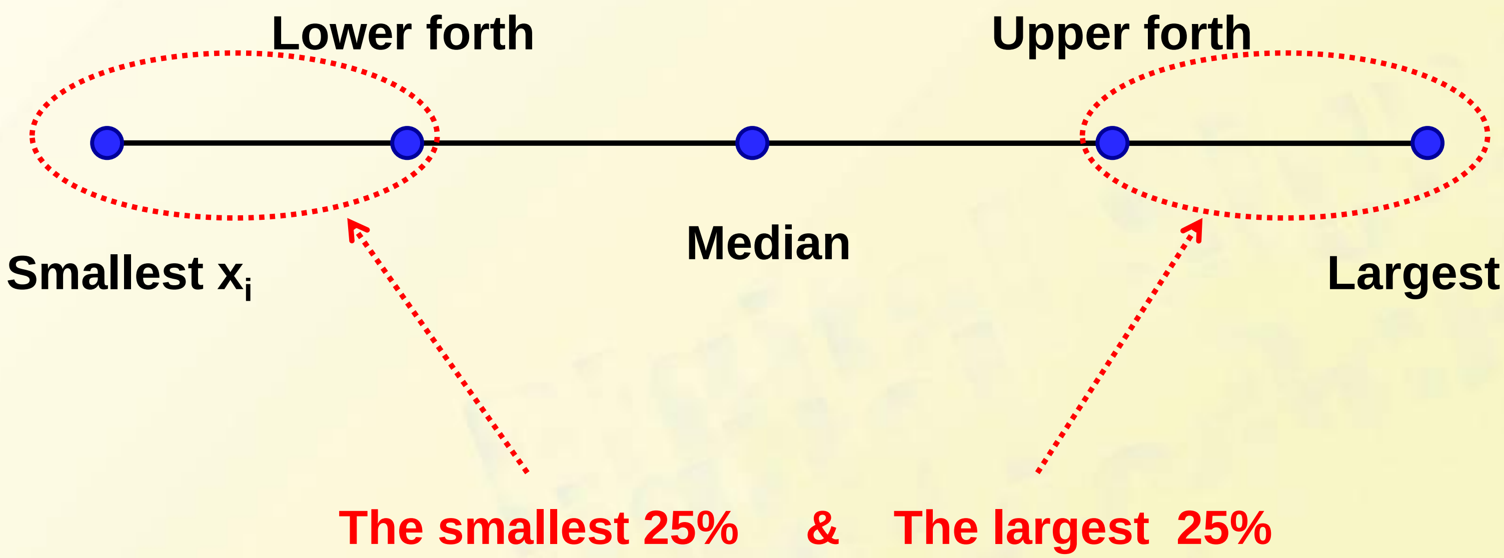
- Fourth Spread

  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



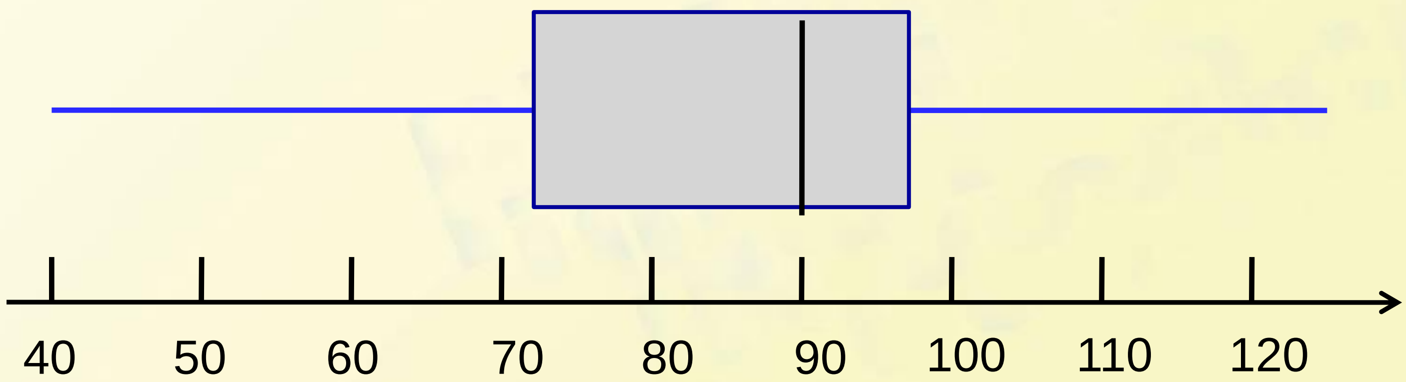
1.4 Measures of Variability

- Example 1.19

$\underbrace{40\ \ 52\ \ 55\ \ 60\ \ 70\ \ 75\ \ 85\ \ 85\ \ 90}\ \overbrace{90\ \ 92\ \ 94\ \ 94\ \ 95\ \ 98\ \ 100\ \ 115\ \ 125\ \ 125}$

Smallest x_i: 40   lower fourth = 72.5 median= 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 f_s from the closest fourth is an outlier.
- **Extreme:** An outlier is extreme if it is more than 3 f_s from the nearest fourth
- **Mild:** An outlier is mild if it is in the range of (1.5 f_s ,3 f_s] from the nearest fourth.

1.4 Measures of Variability

- Example

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

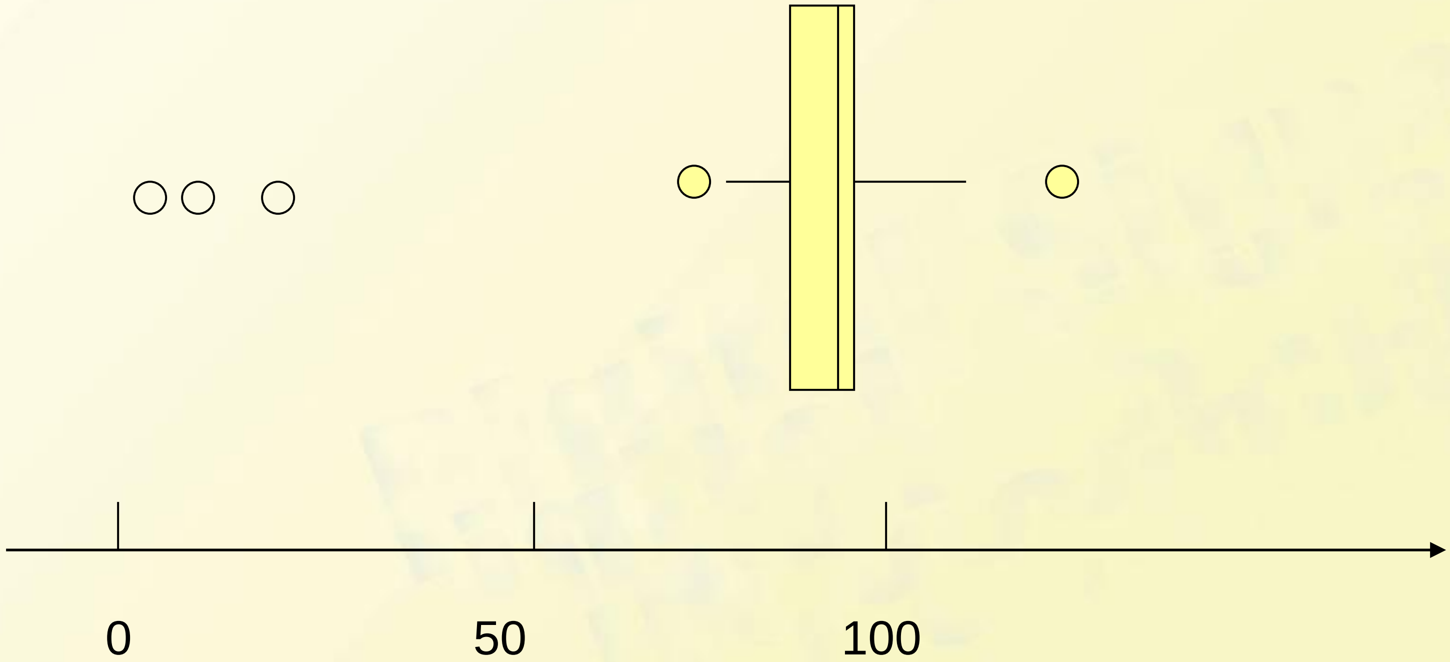
Relevant quantities

median = 94.8 lower fourth =90.2 upper fourth=96.7

f_s=6.5 1.5f_s=9.75 3f_s=19.5

1.4 Measures of Variability

- A boxplot of the pulse width data showing mild and extreme outliers



Comparative Boxplots

- A comparative or side-by-side boxplot is a very effective way of revealing similarities and differences between two or more data sets consisting of observations on the same variable—fuel efficiency observations for four different types of automobiles, crop yields for three different varieties, and so on.

Example 1.21

- In recent years, some evidence suggests that high indoor radon concentration may be linked to the development of childhood cancers, but many health professionals remain unconvinced.

- Houses in the second sample had no recorded cases of childhood cancer. Fig. 1.20 presents a stem-and-leaf display of the data.

1. Cancer		2. No cancer
9683795	0	95768397678993
86071815066815233150	1	12271713114
12302731	2	99494191
8349	3	839
5	4	
7	5	55
	6	
	7	Stem: Tens digit
HI: 210	8	5 Leaf: Ones digit

Stem-and-leaf display for Example 1.20

Figure 1.20

Numerical summary quantities are as follows:

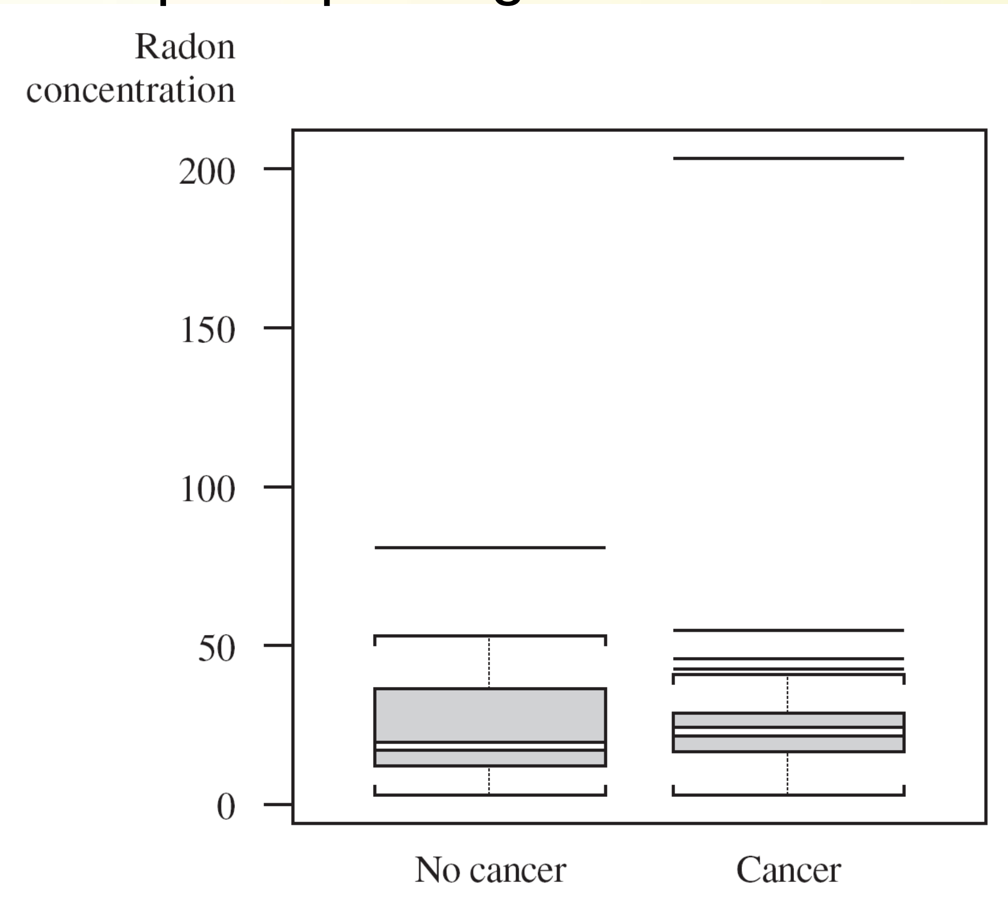
	$\bar{x}$	$\tilde{x}$	s	f_s
Cancer	22.8	16.0	31.7	11.0
No cancer	19.2	12.0	17.0	18.0

- The values of both the mean and median suggest that the cancer sample is centered somewhat to the right of the no-cancer sample on the measurement scale.

The mean, however, exaggerates the magnitude of this shift, largely because of the observation 210 in the cancer sample.

- **The values of *s* suggest more variability in the cancer sample than in the no-cancer sample.**

▪Figure 1.23 shows a comparative boxplot from the S-Plus computer package.



A boxplot of the data in Example 1.21, from S-Plus

Figure 1.23

▪**The no-cancer box is stretched out compared with the cancer box (*fs* = 18 vs. *fs* = 11), and the positions of the median lines in the two boxes show much more skewness in the middle half of the no-cancer sample than the cancer sample.**