MATH1. Part II

Probability and Statistics

Chapter 2

Description of Samples and Populations

2.1 Introduction

Variable

- **Variable:** a variable is a characteristic of a person or a thing that can be assigned a number or a category.
  - For example, blood type (A, B, AB, O) and age are two variables we might measure on a person.
- A **categorical variable** is a variable that records which of several categories a person or thing is in.
- A **numeric variable** records the amount of something.
  - A **continuous variable** is a numeric variable that is measured on a continuous scale.
  - Some types of numeric variables are not continuous but fall on a discrete scale, with spaces between the possible values. A **discrete variable** is a numeric variable for which we can list the possible values.

2.1 Introduction

Variable

Examples of variables

- Discuss with your classmates and give examples of following variable.
  - categorical variable
  - numeric variable
    - continuous variable
    - discrete variable

2.1 Introduction

Variable

Examples of variables

- categorical variable
  - Blood type of a person: A, B, AB, O
  - Sex of a fish: male, female
  - Color of a flower: red, pink, white
  - Shape of a seed: wrinkled, smooth
- numeric variable
  - continuous variable
    - Weight of a baby
    - Cholesterol concentration in a blood specimen
    - Optical density of a solution
  - discrete variable
    - Number of bacteria colonies in a petri dish
    - Number of cancerous lymph nodes detected in a patient
    - Length of a DNA segment in basepairs

2.1 Introduction

- **Observational units:** When we collect a sample of n persons or things and measure one or more variables on them, we call these persons or things **observational units** or cases.

Sample	Variable	Observational unit
150 babies born in a certain hospital	Birthweight (kg)	A baby
73 *Cecropia* moths caught in a trap	Sex	A moth
81 plants that are a progeny of a single parental cross	Flower color	A plant
Bacterial colonies in each of six petri dishes	Number of colonies	A petri dish

- **Notation for Variables and Observations**
  - **Y** = birthweight (the variable-uppercase letters)
  - **y** = 7.9 lb (the observation value-lowercase letters)

2.3 Descriptive statistics

- **Descriptive statistics** are statistics that describe a set of data.
- The sample **median** is the value that most nearly lies in the middle of the sample

Example 2.3.1 Weight Gain of Lambs

- The following are the 2-week weight gains (lb) of young lambs of the same breed that had been raised on the same diet:
- If the ordered observations are: 1 2 10 10 11 13 19, what is the median weight gain?
- If the ordered observations are: 1 2 10 11 13 19, what is the median weight gain?

2.3 Descriptive statistics

- **Descriptive statistics** are statistics that describe a set of data.
- The sample **median** is the value that most nearly lies in the middle of the sample

Example 2.3.1 Weight Gain of Lambs

- The following are the 2-week weight gains (lb) of young lambs of the same breed that had been raised on the same diet:
- If the ordered observations are: 1 2 10 10 11 13 19, what is the median weight gain?
  - The median weight gain is: $\tilde{y}$ = 10 lb
- If the ordered observations are: 1 2 10 11 13 19, what is the median weight gain?
  - The median weight gain is: $\tilde{y}$ = (10 + 11)/2 = 10.5 lb

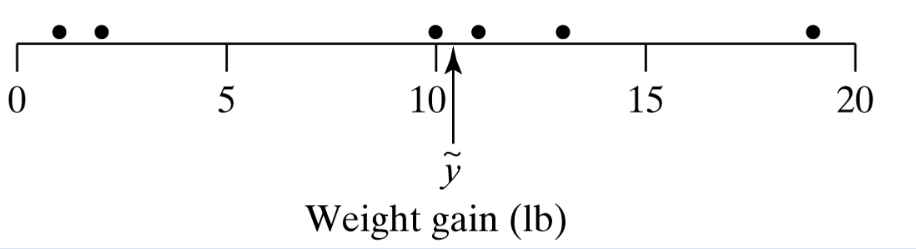


Figure 2.3.1 Plot of the lamb weight-gain data

2.3 Descriptive statistics

- **Descriptive statistics** are statistics that describe a set of data.
- The **mean** of a sample (or "the sample mean") is the sum of the observations divided by the number of observations.

THE SAMPLE MEAN The general definition of the sample mean is

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

where the y_i's are the observations in the sample and n is the sample size (that is, the number of y_i's).

Example 2.3.1 Weight Gain of Lambs

- The following are the 2-week weight gains (lb) of young lambs of the same breed that had been raised on the same diet:
- If the ordered observations are: 1 2 10 11 13 19, what is the mean weight gain?

2.3 Descriptive statistics

- **Descriptive statistics** are statistics that describe a set of data.
- The **mean** of a sample (or "the sample mean") is the sum of the observations divided by the number of observations.

THE SAMPLE MEAN The general definition of the sample mean is

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

where the y_i's are the observations in the sample and n is the sample size (that is, the number of y_i's).

Example 2.3.1 Weight Gain of Lambs

- The following are the 2-week weight gains (lb) of young lambs of the same breed that had been raised on the same diet:
- If the ordered observations are: 1 2 10 11 13 19, what is the mean weight gain?
  - The mean weight gain is: $\bar{y} = (11 + 13 + 19 + 2 + 10 + 1)/6 = 9.33$ lb

2.3 Descriptive statistics

Mean vs. Median

- **Median is more robust than the mean.**
  - **Robustness**. A statistic is said to be robust if the value of the statistic is relatively unaffected by changes in a small portion of the data, even if the changes are dramatic ones.
  - The median is a robust statistic, but the mean is not robust because it can be greatly shifted by changes in even one observation.
- **Mean is more efficient than the median.**
  - **Efficiency** is a technical notion in statistical theory; roughly speaking, a method is efficient if it takes full advantage of all the information in the data.
  - Partly because of its efficiency, the mean has played a major role in classical methods in statistics.

2.6 Measures of dispersion

Dispersion: how spread out the distribution is.

- **Range:** The sample range is the **difference** between the largest and smallest observations in a sample.

Example 2.6.1 Blood Pressure

- The systolic blood pressures (mm Hg) of seven middle-aged men were given in Example 2.4.1 as follows: 113 124 124 132 146 151 170
- What is the range of above data?

2.6 Measures of dispersion

Dispersion: how spread out the distribution is.

- **Range:** The sample range is the **difference** between the largest and smallest observations in a sample.

Example 2.6.1 Blood Pressure

- The systolic blood pressures (mm Hg) of seven middle-aged men were given in Example 2.4.1 as follows: 113 124 124 132 146 151 170
- What is the range of above data?
    - For these data, the sample range is 170 - 113 = 57 mm Hg
- The range is easy to calculate, but it is very sensitive to extreme values; that is, **range is not robust**.
- If the maximum in the blood pressure sample had been 190 rather than 170, the range would have been changed from 57 to 77.

2.6 Measures of dispersion

Dispersion: how spread out the distribution is.

The standard deviation

- **Deviation** is the difference between an observation and the sample mean:

$$\text{deviation} = \text{observation} - \bar{y}$$

- The sample **standard deviation** is denoted by **s** and is defined by the following formula:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

- So, to find the standard deviation of a sample, first find the deviations. Then
- 1. square; 2. add; 3. divide by n – 1; 4. take the square root

2.6 Measures of dispersion

Dispersion: how spread out the distribution is.

Interpretation of the definition of SD (s)

- For large n, SD can be interpreted approximately as

$$s \approx \sqrt{\text{sample average value of } (y_i - \bar{y})^2}$$

- Thus, it is roughly appropriate to think of the SD as a “typical” distance of the observations from their mean.

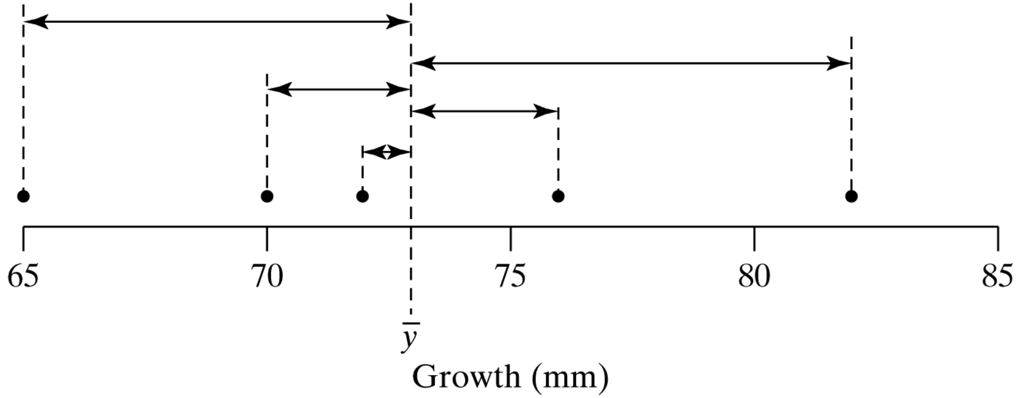


Figure 2.6.1 Plot of chrysanthemum growth data with deviations indicated as distances

2.6 Measures of dispersion

Dispersion: how spread out the distribution is.

Variance

- The sample **variance**, denoted by **s^2**, is simply the standard deviation squared:

$$\text{variance} = s^2$$

- Typical Percentages: The Empirical Rule
  - For “nicely shaped” distributions—that is, unimodal distributions that are not too skewed and whose tails are not overly long or short—we usually expect to find
  - about 68% of the observations within $\pm$ 1 SD of the mean.
  - about 95% of the observations within $\pm$ 2 SDs of the mean.
  - >99% of the observations within $\pm$ 3 SDs of the mean.

2.6 Measures of dispersion

Dispersion: how spread out the distribution is.

Example 2.6.2 Chrysanthemum Growth

- The stem elongation (mm in 7 days) of five plants grown on the same green-house bench. The results were as follows: 76 72 65 70 82
- What is the mean, SD and variance?

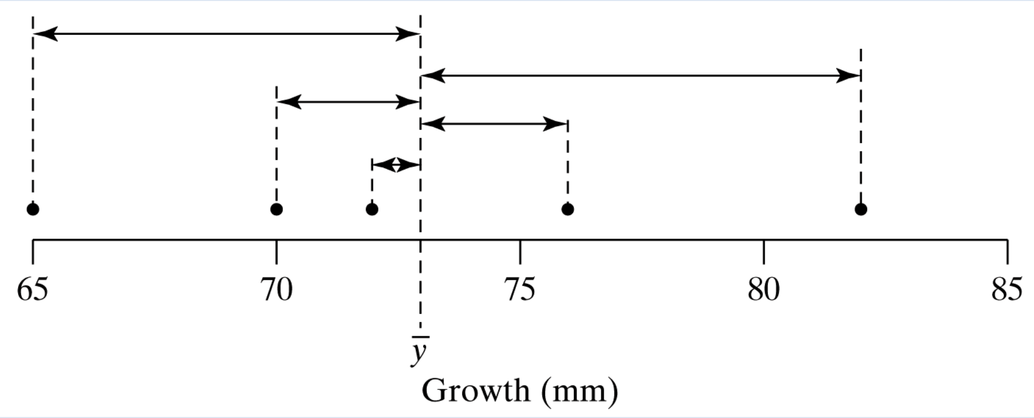
2.6 Measures of dispersion

Dispersion: how spread out the distribution is.

Example 2.6.2 Chrysanthemum Growth

- The stem elongation (mm in 7 days) of five plants grown on the same green-house bench. The results were as follows: 76 72 65 70 82
- What is the mean, SD and variance?
  - $\bar{y} = 365/5 = 73$ mm
  - $\sum_{i=1}^{n}(yi - \bar{y})^2 = 164$
  - s = $\sqrt{164/4} = \sqrt{41} = 6.4\ mm$
  - $s^2 = 41\ mm^2$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$



2.2 Frequency Distribution

- A **frequency distribution** is simply a display of the frequency, or number of occurrences, of each value in the data set.
- A **dotplot** is a simple graph that can be used to show the distribution of a numeric variable when the sample size is small.

Example 2.2.4 Litter size of sows

- A group of thirty-six 2-year-old sows of the same breed were bred to Yorkshire boars. The number of piglets surviving to 21 days of age was recorded for each sow.
- The results are given in Table 2.2.4 and displayed as a dotplot in Figure 2.2.4.

Table 2.2.4 Number of surviving piglets of 36 sows

Number of piglets	Frequency (number of sows)
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2
Total	36

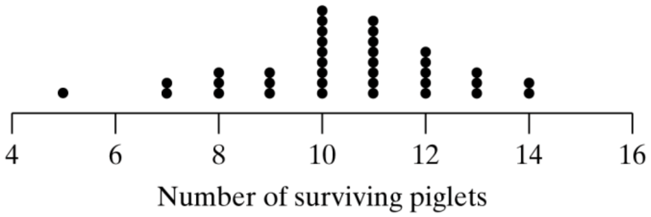


Figure 2.2.4 Dotplot of number of surviving piglets of 36 sows

2.2 Frequency Distribution

- A **frequency distribution** is simply a display of the frequency, or number of occurrences, of each value in the data set.
- A **bar chart** is a graph of categorical data showing the number of observations in each category.

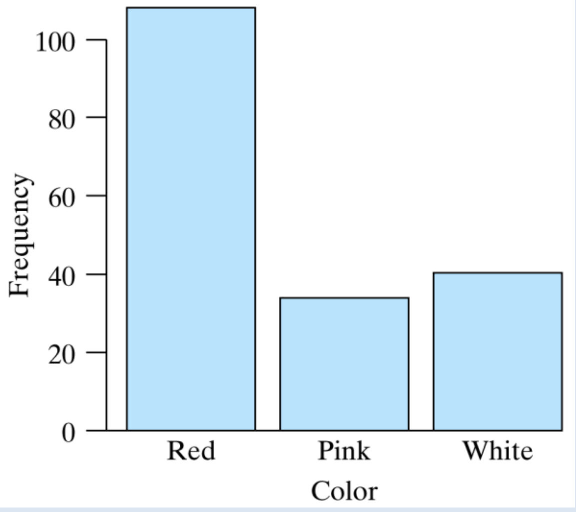
Example 2.2.1 Color of Poinsettias

- Poinsettias can be red, pink, or white. In one investigation of the hereditary mechanism controlling the color, 182 progeny of a certain parental cross were categorized by color.
- The bar graph in Figure 2.2.1 is a visual display of the results given in Table 2.2.1

Table 2.2.1 Color of 182 poinsettias

Color	Frequency (number of plants)
Red	108
Pink	34
White	40
Total	182

Figure 2.2.1 Bar chart of color of 182 poinsettias



2.2 Frequency Distribution

Grouped frequency distributions

- For many data sets, it is necessary to group the data in order to condense the information adequately

Example 2.2.6 Serum CK

Table 2.2.6 Serum CK values for 36 men

121	82	100	151	68	58
95	145	64	201	101	163
84	57	139	60	78	94
119	104	110	113	118	203
62	83	67	93	92	110
25	123	70	48	95	42

Table 2.2.7 Frequency distribution of serum CK values for 36 men

Serum CK (U/l)	Frequency (number of men)
[20,40)	1
[40,60)	4
[60,80)	7
[80,100)	8
[100,120)	8
[120,140)	3
[140,160)	2
[160,180)	1
[180,200)	0
[200,220)	2
Total	36

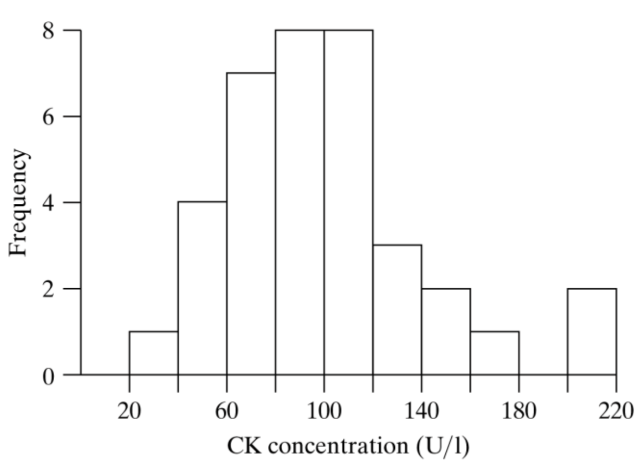


Figure 2.2.7 Histogram of serum CK concentrations for 36 men

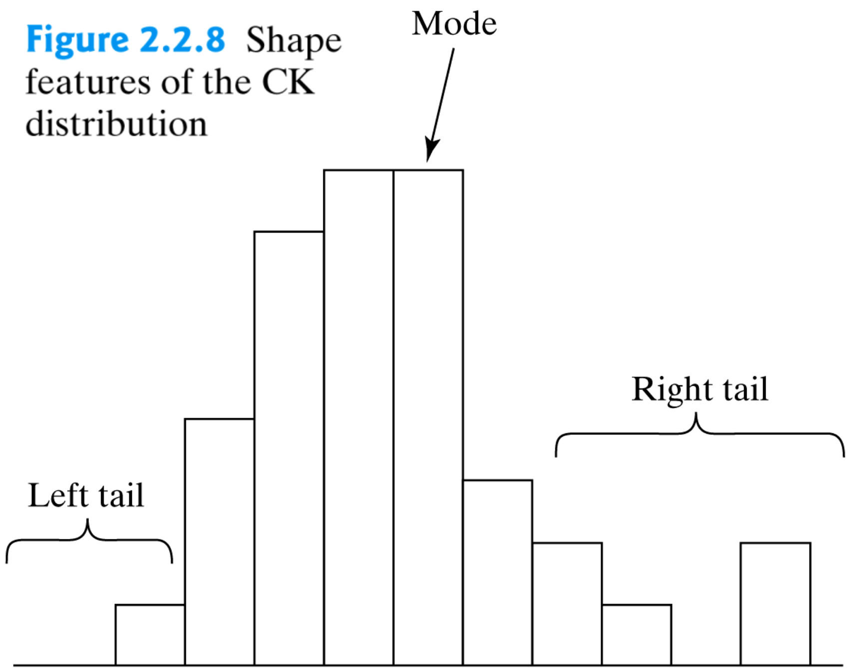
2.2 Frequency Distribution

Grouped frequency distributions

- For many data sets, it is necessary to group the data in order to condense the information adequately

Example 2.2.6 Serum CK

Figure 2.2.8 Shape features of the CK distribution



- The histogram shows the shape of the **distribution**.
- Note that the CK values are piled up around a central peak, or **mode**.
- On either side of this mode, the frequencies decline and ultimately form the **tails** of the distribution.
- The CK distribution is not symmetric but is a bit **skewed to the right**, which means that the right tail is more stretched out than the left.*

2.2 Frequency Distribution

Interpreting Areas in a Histogram

- The area of each bar is proportional to the corresponding frequency.
- The area of one or several bars can be interpreted as expressing the number of observations in the classes represented by the bars.

Example 2.2.6 Serum CK

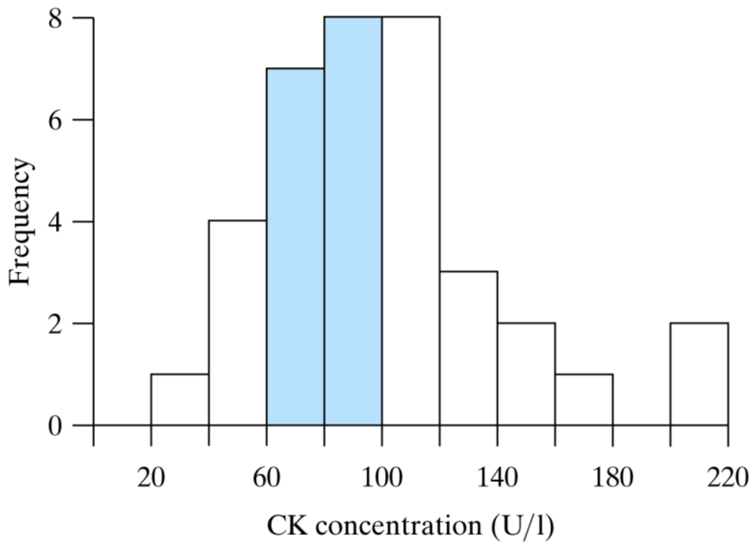


Figure 2.2.12 Histogram of CK distribution. The shaded area is 42% of the total area and represents 42% of the observations.

2.2 Frequency Distribution

Shapes Of Distributions

- A common shape for biological data is **unimodal** (has one mode) and is somewhat skewed to the right, as in (c).
- Approximately bell-shaped distributions, as in (a), also occur.
- Sometimes a **Bimodality** (two modes), as in (f), can indicate the existence of two distinct subgroups of observational units.

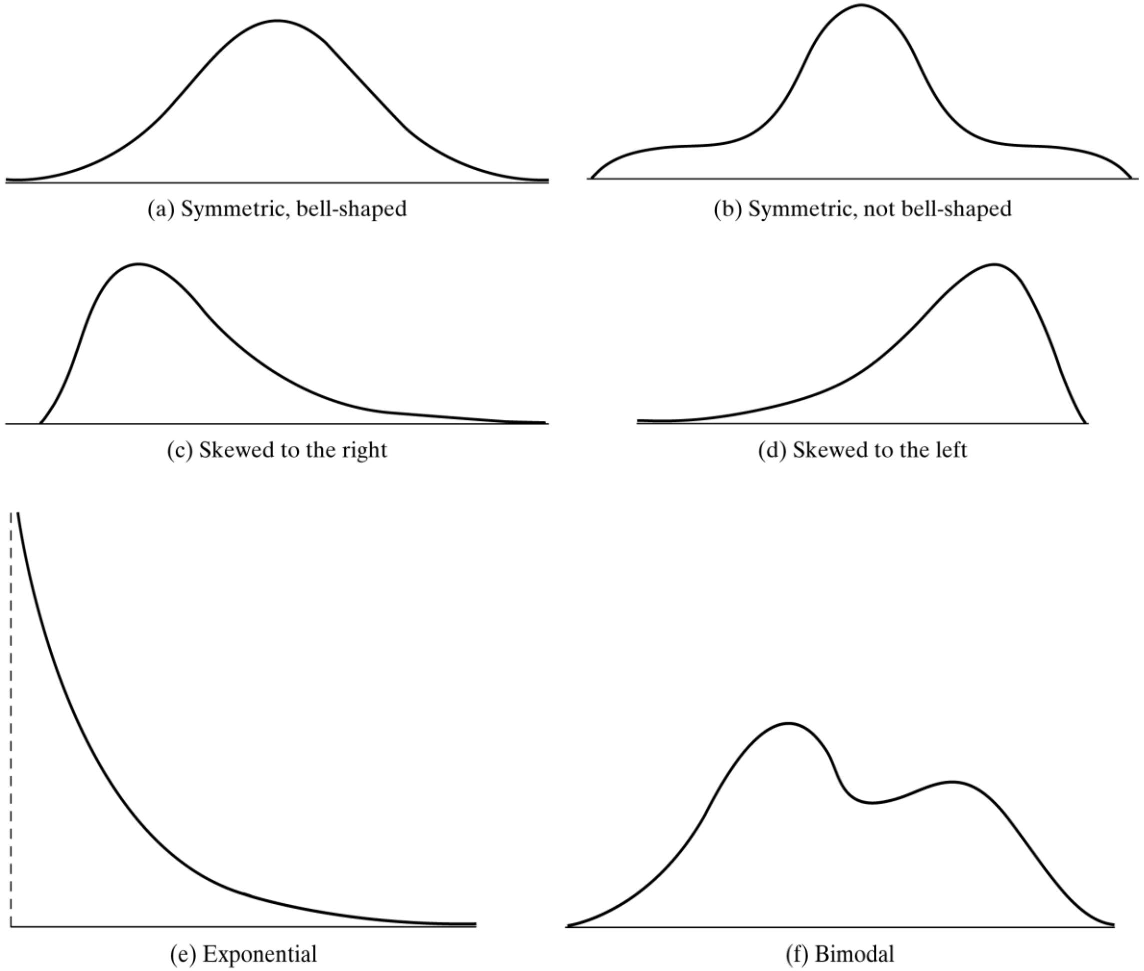


Figure 2.2.14 Shapes of distributions

To be continued ….

Tomorrow

2.4 Boxplot

- One of the most efficient graphics, both for examining a single distribution and for making comparisons between distributions.

Quartiles

Quartiles are the values that divide a list of numbers into quarters.

- The **first quartile**, denoted by Q_1 ,is the median of the data values in the lower half of the data set.
- The **third quartile**, denoted by Q_3 , is the median of the data values in the upper half of the data set.
- The **interquartile range (IQR)** is the difference between the first and third quartiles and is abbreviated as **IQR**: IQR = $\mathbf{Q_3}$ - $\mathbf{Q_1}$.

2.4 Boxplot

Example 2.4.1 Blood Pressure

- The systolic blood pressures (mm Hg) of seven middle-aged men were as follows:
  - 151, 124, 132, 170, 146, 124, 113
- What are Q1, Q3 and IQR?

Example 2.4.2 Pulses

- The pulses of 12 college students were measured:
  62, 64 , 68 70, 70, 74 74, 76, 76 78, 78, 80
- What are Q1, Q3 and IQR?

2.4 Boxplot

Example 2.4.1 Blood Pressure

- The systolic blood pressures (mm Hg) of seven middle-aged men were as follows:
  - 151, 124, 132, 170, 146, 124, 113
- What are Q1, Q3 and IQR?
  - Putting these values in rank order, the sample is:

113 124 124 132 146 151 170

first quartile
Q_1 = 124

median
132

third quartile
151

IQR = 151-124 = 27

Example 2.4.2 Pulses

- The pulses of 12 college students were measured:

62, 64 , 68 70, 70, 74 74, 76, 76 78, 78, 80

- What are Q1, Q3 and IQR?

first quartile
Q_1 = (68+70)/2 = 69

median
(74+74)/2 = 74

third quartile
Q_3 = (76+78)/2 = 77

IQR = 77-69 = 8

2.4 Boxplot

- One of the most efficient graphics, both for examining a single distribution and for making comparisons between distributions.

Outliers

A data point differs so much from the rest of the data that it doesn't seem to belong with the other data.

- The lower fence of a distribution is **lower fence = Q_1 - 1.5 x IQR**
- The upper fence of a distribution is **upper fence = Q_3 + 1.5 x IQR**
- Definition of "**outlier**" in statistical practice: An outlier is a data point that falls outside of the fences.

data point < Q_1 - 1.5 x IQR <u>**or**</u> **data point > Q_3 + 1.5 x IQR**

2.4 Boxplot

Example 2.4.4 Radish Growth in Light

- Students grew 14 radish seedlings in constant light. The observations, in order, are
  3 5 5 7 7 8 9 10 10 10 10 14 20 21
- What are Q1, Q3 and IQR?
- What are the lower fence and upper fence?
- Is there any outlier?

2.4 Boxplot

Example 2.4.4 Radish Growth in Light

- Students grew 14 radish seedlings in constant light. The observations, in order, are
  3 5 5 7 7 8 9 10 10 10 10 14 20 21
- What are Q1, Q3 and IQR?
- What are the lower fence and upper fence?
- Is there any outlier?
  - 3 5 5 7 7 8 9 10 10 10 10 14 <u>20</u> <u>21</u>

first quartile $Q_1 = 7$

median $(9+10)/2 = 9.5$

third quartile $Q_3 = 10$

$\boxed{IQR = 10-7 = 3}$

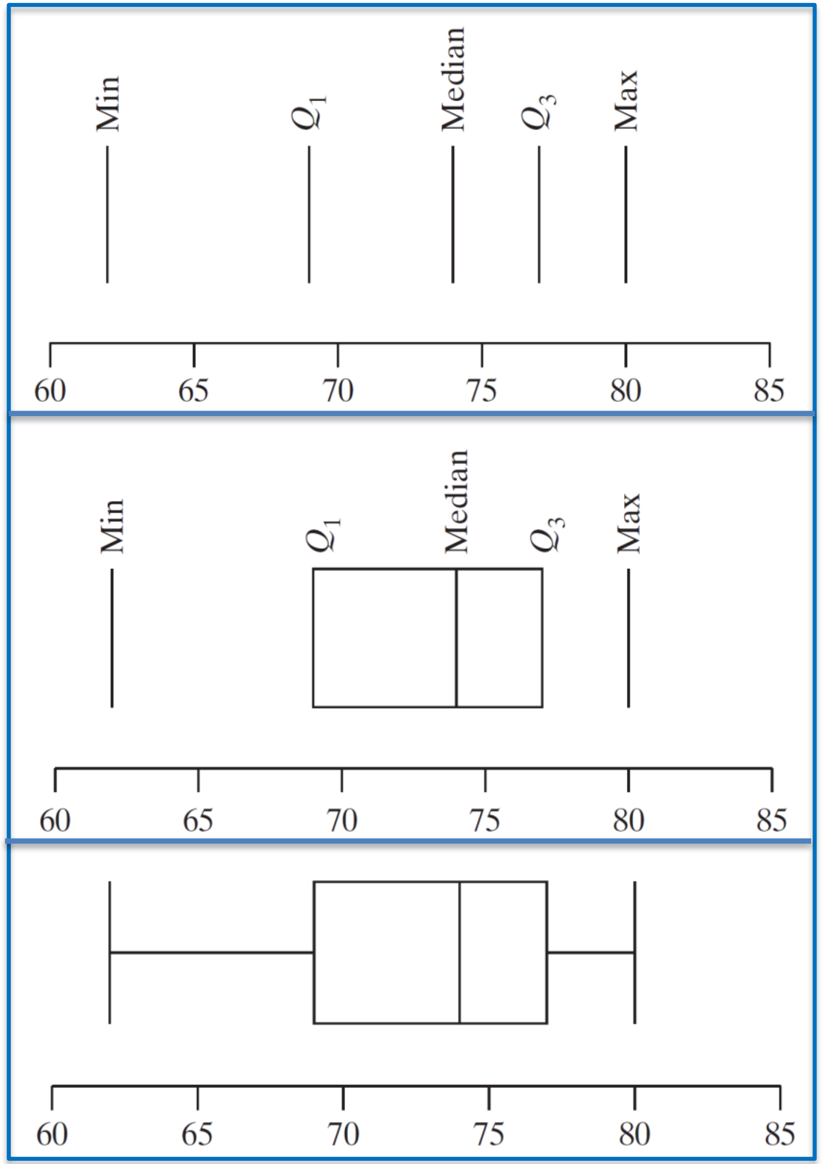
- lower fence = Q_1 - 1.5 x IQR = 7 - 1.5 x 3 = 2.5
- upper fence = Q_3 + 1.5 x IQR = 10 + 1.5 x 3 = 14.5
- Thus, outliers are 20 and 21.

2.4 Boxplot

Boxplots for data <u>with no outliers</u>

- To make a **boxplot** for a data set with no outliers, we first make a number line; then we mark the positions **minimum**, $\mathbf{Q_1}$, the **median**, $\mathbf{Q_3}$, and the **maximum**;
- Next, we make a box connecting the quartiles:

  ** Note that the interquartile range is equal to the length of the box.*
- Finally, provided there are **no outliers*** we extend "whiskers" from Q_1 down to the minimum and from Q_3 up to the maximum:



2.4 Boxplot

Boxplots for data with no outliers

Example 2.4.5 Radish Growth

- Students kept their radish seed bags in total darkness for 3 days and then measured the length, in mm, of each radish shoot at the end of the 3 days.
- Here are the data in order from smallest to largest:
  - 8 10 11 15 15 20 20 | 22 25 29 30 33 35 37

first quartile
$Q_1 = 15$

median

third quartile
$Q_3 = 30$

- minimum = 8, Q_1 = 15, median $\tilde{y}$= 21, Q_3 = 30, maximum = 37.
- Draw the box plot of above data.

2.4 Boxplot

Boxplots for data with no outliers

Example 2.4.5 Radish Growth

- Students kept their radish seed bags in total darkness for 3 days and then measured the length, in mm, of each radish shoot at the end of the 3 days.
- Here are the data in order from smallest to largest:
  - 8 10 11 15 15 20 20 | 22 25 29 30 33 35 37

first quartile $Q_1 = 15$

median

third quartile $Q_3 = 30$

- minimum = 8, Q_1 = 15, median $\tilde{y}$= 21, Q_3 = 30, maximum = 37.
- Draw the box plot of above data.

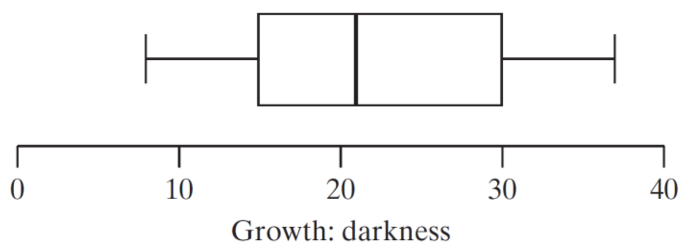


Figure 2.4.1 Boxplot of data on radish growth in darkness

2.4 Boxplot

Boxplots for data <u>with outliers</u>

If there are outliers

- extend a “whisker” from Q_3 up to the largest data point that is **not** an outlier.
- extend a “whisker” from Q_1 down to the smallest observation that is **not** an outlier.

  * *Note that the interquartile range is equal to the length of the box.*
- Draw circles for outliers

Example 2.4.4 Radish Growth in Light

- Students grew 14 radish seedlings in constant light. The observations, in order, are

3 5 5 7 7 8 9 10 10 10 10 14 20 21

outliers: 20 21

first quartile $Q_1 = 7$

median $(9+10)/2 = 9.5$

third quartile $Q_3 = 10$

Draw a boxplot of these data.

2.4 Boxplot

Boxplots for data <u>with outliers</u>

If there are outliers

- extend a "whisker" from Q_3 up to the largest data point that is **<u>not</u>** an outlier.
- extend a "whisker" from Q_1 down to the smallest observation that is **<u>not</u>** an outlier.

  * *Note that the interquartile range is equal to the length of the box.*
- Draw circles for outliers

Example 2.4.4 Radish Growth in Light

- Students grew 14 radish seedlings in constant light. The observations, in order, are

3 5 5 7 7 8 9 10 10 10 10 14 **20 21** (outliers)

first quartile $Q_1 = 7$

median $(9+10)/2 = 9.5$

third quartile $Q_3 = 10$

A boxplot gives a quick visual summary of the distribution.

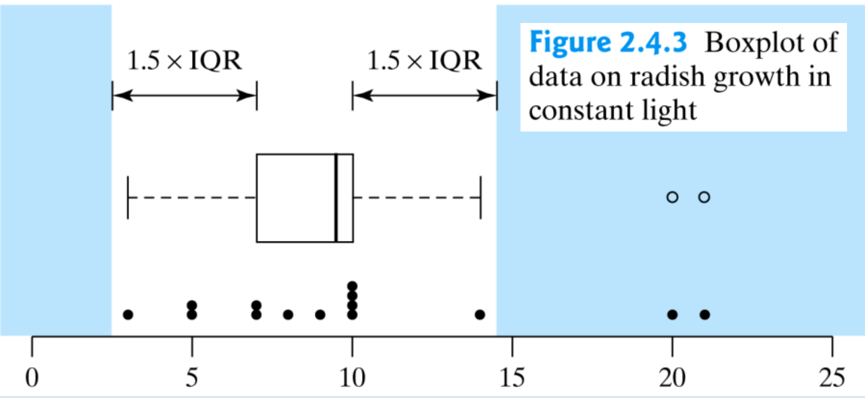


Figure 2.4.3 Boxplot of data on radish growth in constant light

Summary

Chapter 2

- 2.1 Introduction
- 2.3 Descriptive Statistics: Measures of Center
- 2.6 Measures of Dispersion
- 2.2 Frequency Distributions
- 2.4 Boxplots

Homework

Chapter 2

- 2.1.2; 2.1.3
- 2.2.1
- 2.3.3; 2.3.13; 2.3.14
- 2.4.2
- 2.6.4