Chapter 1

Summary and Display of Univariate Data (contd.)

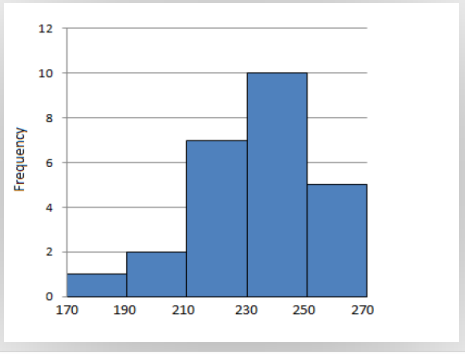
Lecture 3

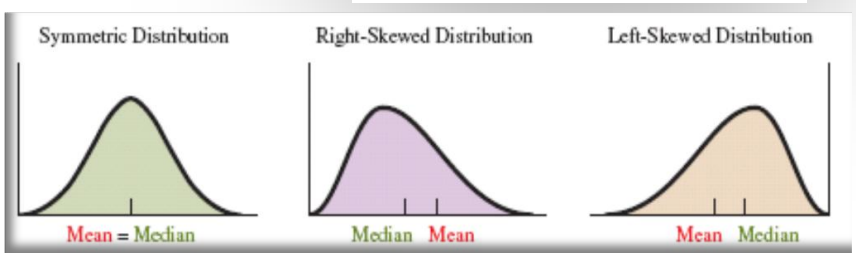
Histograms

Describe a distribution

Measure of center

Dr. Lasantha Premarathna





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

Histograms can be useful to describe data.

Example: Here are the data (number of hours worked) for 25 students in a particular semester. Construct a histogram from the following data.

175, 192, 207, 212, 213, 214, 218, 225, 229, 230,
231, 235, 235, 237, 240, 240, 240, 242, 248, 250,
253, 257, 260, 265, 265

- Nice graph
- excl. empty

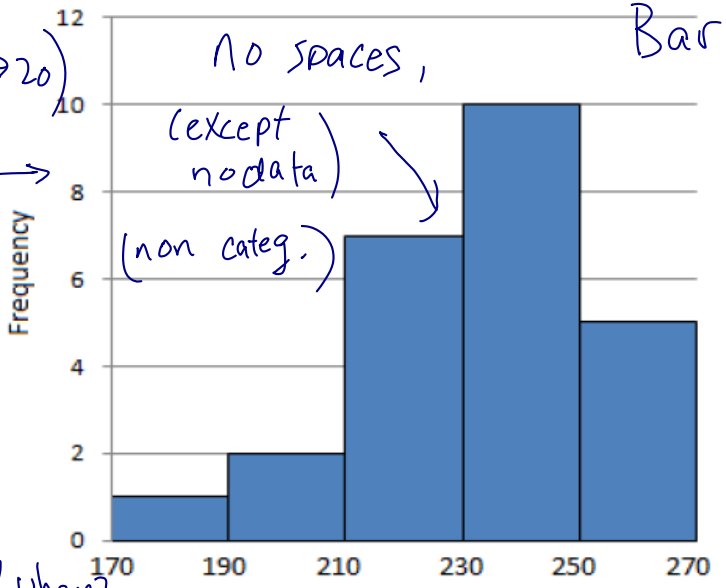
Range = 265-175 = 90 (1)
Number of intervals needed = 5 (2) (smaller (5–15 intervals, typically), larger depending on dataset.)
Width of an interval = 90/5 = 18 (3) → (typically → 20 round)

Create the frequency table

Interval(hours)	Frequency
170-190	1
190-210	2
210-230	7
230-250	10
250-270	5

175
192, 207
212, 213, 214, 218, 225, 229

Histogram of Hours worked ← Bar graf.



No spaces,
(except no data)
(non categ.)

scale →
Name →

230 where?

- be consistant (upper vs lower)

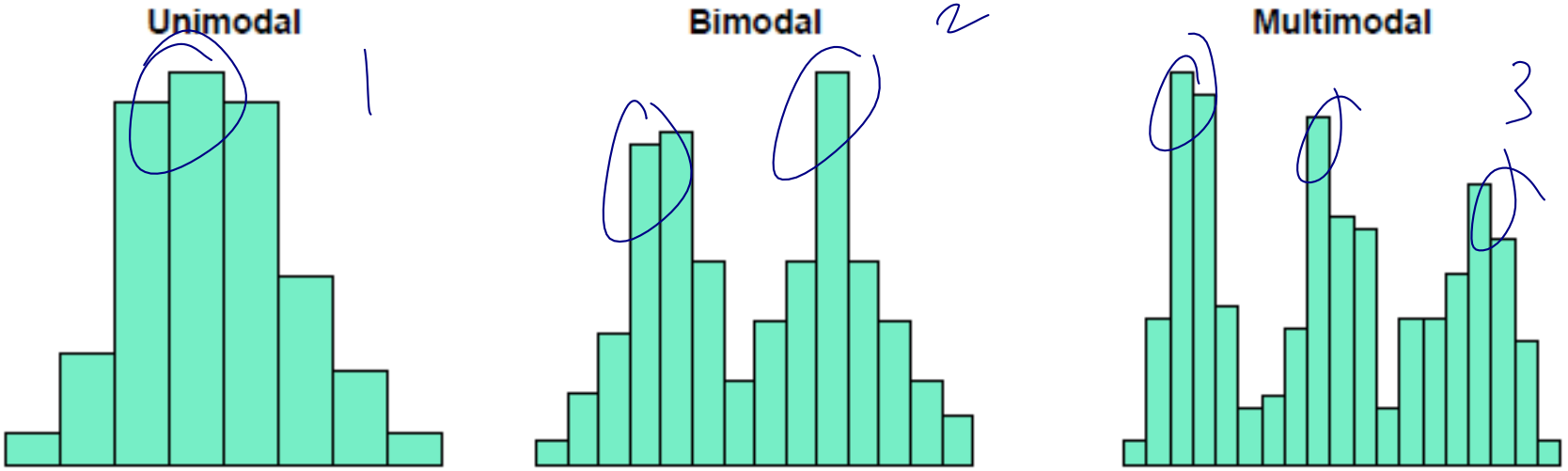
Constructing a Histogram

- Divide the range of the data into intervals of equal width
- Count the number of observations in each interval, creating a frequency table
- On the horizontal axis, label the values or the endpoints of the intervals.
- Draw a bar over each value or interval with height equal to its frequency (or percentage), values of which are marked on the vertical axis.
- Label axes and provide proper headings

Describing a distribution

Type of Mound

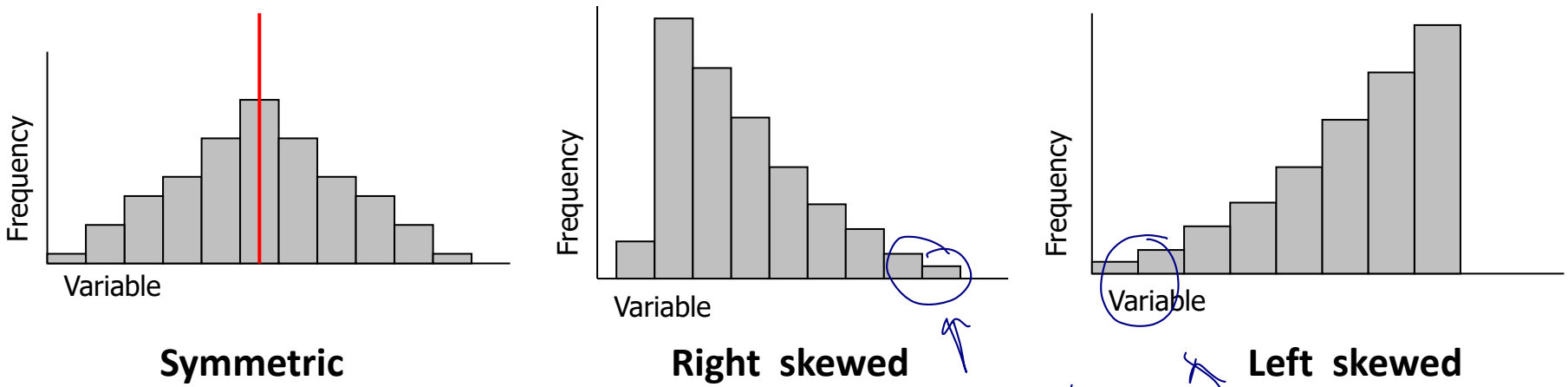
- Unimodal - one clear peak
- Bimodal - 2 peaks
- Multimodal - more than 2 peaks



Describing a distribution

Shape

- **Symmetric** Distributions: if both left and right sides of the histogram are mirror images of each other
- A distribution is **skewed to the left** if the left tail is longer than the right tail
- A distribution is **skewed to the right** if the right tail is longer than the left tail

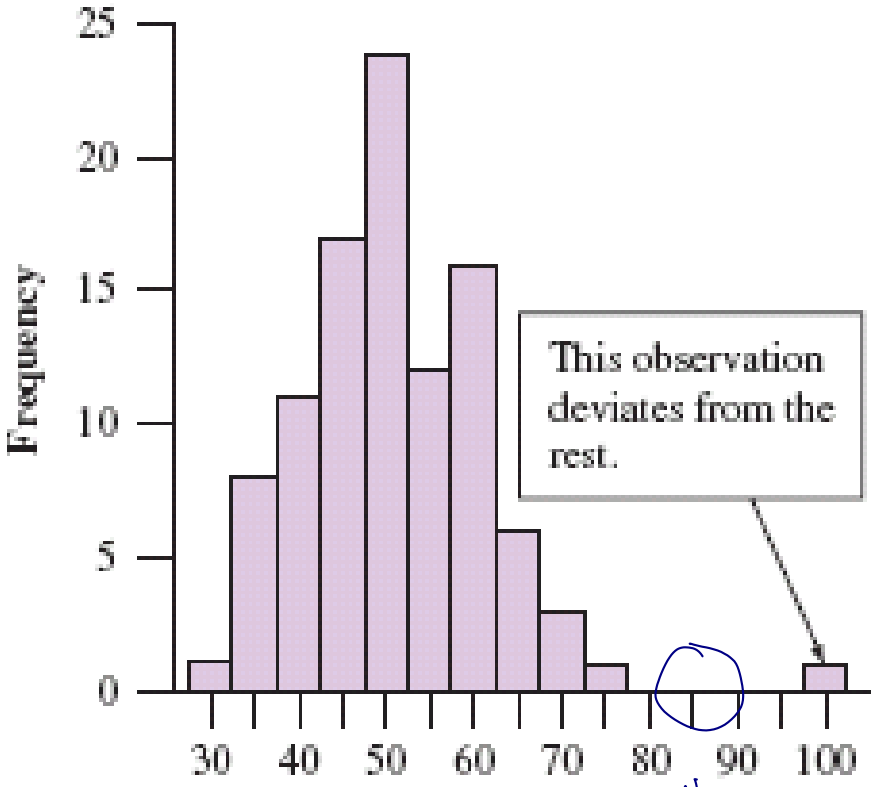


Center: where do the observation cluster about?

Spread: Assess the spread of a distribution.

Describing a distribution

Outlier: an outlier falls far from the rest of the data
unusually large or small observation



Measures of Center

Mean

The mean is the sum of the observations divided by the number of observations. Sample mean is

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

e.g. Number of hours spent studying per week for 5 students are 4, 6, 8, 7, 5.

Find the mean number of hours spent studying/week.

$$\bar{x} = \frac{4 + 6 + 8 + 7 + 5}{5} = 6 \text{ hours}$$

Measures of Center

1. Sort first

Median

- The median is the midpoint of the observations when they are ordered from the smallest to the largest (ascending order)
- If the number of observations is:
  - Odd : median is the middle observation; i.e. $\left(\frac{n+1}{2}\right)^{th}$ observation
  - Even: median is the average of the two middle observations

    average of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ observations

Example1 : 12, 14 ,15, 17, 20, 24, 24, 27, 29 ; *n = 9*

Median is the (9+1)/2 th observation , *median = 20*

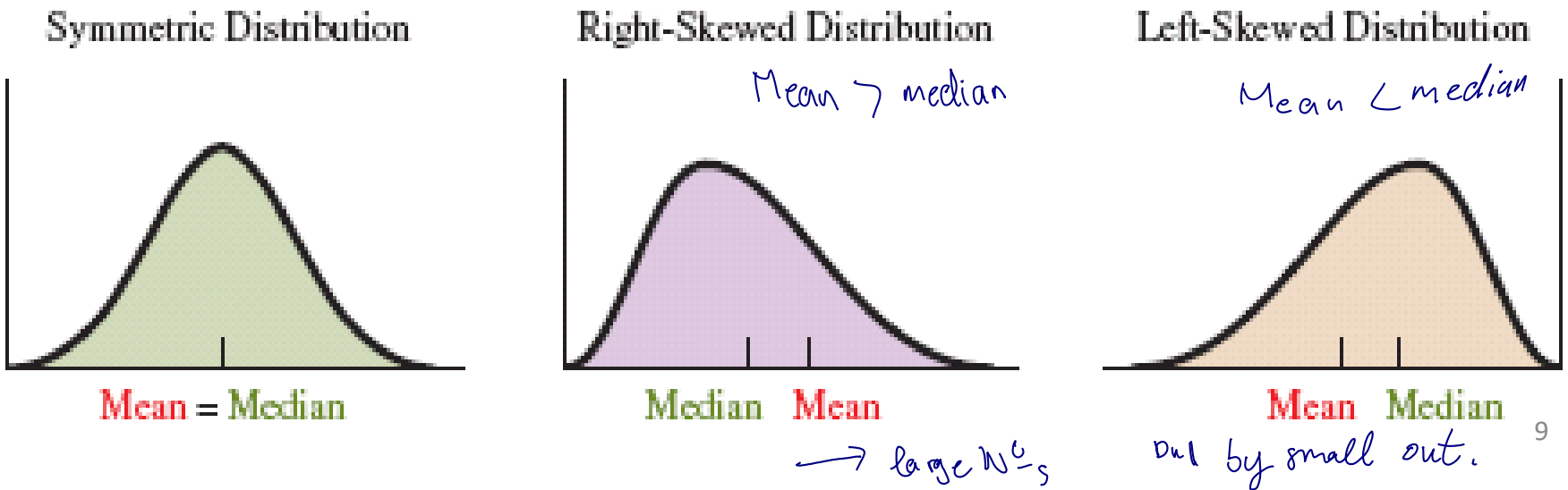
Example 2 : 12, 14 ,15, 17, 20, 24, 24, 27, 29, 30 ; *n = 10*

Median is the 5 th and 6 th observation , *median = (20+24)/2 = 22*

Histogram → Distr.
smoothing

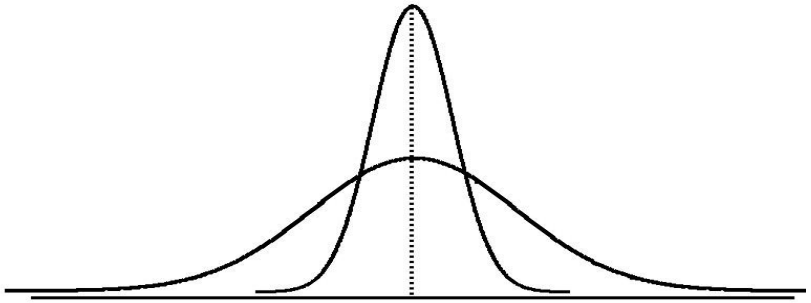
Comparing the Mean and Median

- When data nearly symmetric $mean \approx median$
- In a skewed distribution, the mean is farther out in the long tail than is the median
  - When data have long right tail $mean > median$
  - When data have long left tail $mean < median$
  - For skewed distributions the median is preferred because it is better representative of a typical observation



Measures of variability

Measures of variation give information on the **spread** or **variability** or **dispersion** of the data values



Same center, different variation

- **Range**
  - Difference between the largest and the smallest values

$$\textbf{Range} = X_{largest} - X_{smallest}$$

  - The range is strongly affected by outliers

e.g. Data : 70, 46, 62, 64, 15, 78, 56, 64, 69, 49

Range = 78 – 15 = 63

Summary

- Construction Histograms
- Describe a distribution
- Measure of centre
- Measure of variability (this will continue in the next class)

Before the next class

- Review the lecture 3 and related sections in the text book
- Register to iClicker Cloud, if not done already

Next Class:

- Chapter 1 : Summary and Display of Univariate Data (contd)
  - Measures of variability
  - Boxplots