Review 1

Chapter 1:

1. Elements, Variable, and Observations:
2. Type of Data: Qualitative Data and Quantitative Data
   - Qualitative data may be nonnumeric or numeric.
   - Quantitative data are always numeric.
   - Arithmetic operations are only meaningful with quantitative data.

Chapter 2:

1. Summarizing qualitative data:
   - Frequency distribution, relative frequency distribution, and percent frequency distribution.
   - Bar plot and pie plot.
2. Summarizing quantitative data:
   - Frequency distribution, relative frequency distribution, percent frequency distribution, cumulative frequency distribution, cumulative relative frequency distribution, cumulative percent frequency distribution
   - Histogram, ogive, and stem-and leaf display.

Chapters 3:

- Measures of Location
- Measures of Dispersion
- Exploratory Data Analysis
- Measure of Relative Location
- Weighted and Grouped Mean and Variance

Equations:

$$s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}=\frac{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}{n-1}, CV=\frac{s}{\bar{x}}\cdot 100$$

$$\bar{x}_g=\frac{\sum_{k=1}^{m}f_kM_k}{\sum_{k=1}^{m}f_k}=\frac{\sum_{k=1}^{m}f_kM_k}{n}, s_g^2=\frac{\sum_{k=1}^{m}f_k\left(M_k-\bar{x}_g\right)^2}{n-1}=\frac{\sum_{k=1}^{m}f_kM_k^2-n\bar{x}_g^2}{n-1}$$

Example 1:

A magazine surveyed a sample of its subscribers. Some of the responses from the survey are shown below.

Subscriber ID	Gender	Age	Income ($1000)
0006	F	22	45
4798	M	21	53
2291	F	33	82
4988	M	38	30

(a) How many elements are in the data set? Write them down.

(b) How many variables are in the data set? Write them down.

(c) How many observations are in the data set? Write them down.

(d) Which of the above (Sex, Age, Annual Household Income) are qualitative and which are quantitative?

(e) Are the data time series or cross-sectional?

[Solution:]

(a) 4 **elements, subscribers:** $0006, 4798, 2291,$ **and** $4988.$

(b) 3 **variables, Gender, Age, and Income.**

(c) 4 **observations,** $(\mathrm{F}, 22, 45), (\mathrm{M}, 21, 53), (\mathrm{F}, 33, 82)$ **and** $(\mathrm{M}, 38, 30).$

(d) Quantitative: Age and Income; Qualitative: Gender.

(e) The data are cross-sectional.

Example 2:

For the following data, $2, 1, 0, 2, 0, 2, 1, 2, 0, 2, 1, 2.$

(a) Compute the mean.

(a) The standard deviation.

(c) The coefficient of variation.

(d) The $(100/3)th$ **percentile.**

(e) The $82th$ **percentile**

(f) The mode.

(g) The interquartile range.

(h) The five number summary for the data.

(i) The box plot.

[Solution:]

(a)

$$\overline{x} = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{2 + 1 + \cdots + 1 + 2}{12} = 1.25.$$

(b)

$$s = \sqrt{\frac{\sum_{i=1}^{12} (x_i - \overline{x})^2}{12 - 1}} = \sqrt{\frac{(2 - 1.25)^2 + \cdots + (2 - 1.25)^2}{11}} = 0.866.$$

(c)

$$CV = \frac{s}{\overline{x}} \cdot 100 = \frac{0.866}{1.25} \cdot 100 = 69.28.$$

(d)

1. The data are

0	0	0	1	1	1	2	2	2	2	2	2

2.

The index is

$$12 \cdot \frac{\left(^{100}/_{3}\right)}{100} = 4.$$

Thus,

$$\frac{1+1}{2} = 1$$

is the $(100/3)th$ percentile.

(e) The index is

$$12 \cdot \frac{82}{100} = 9.84.$$

Thus, the $10th$ data in (d), 2, is the $82th$ percentile.

(f) The mode is 2.

(g) Since

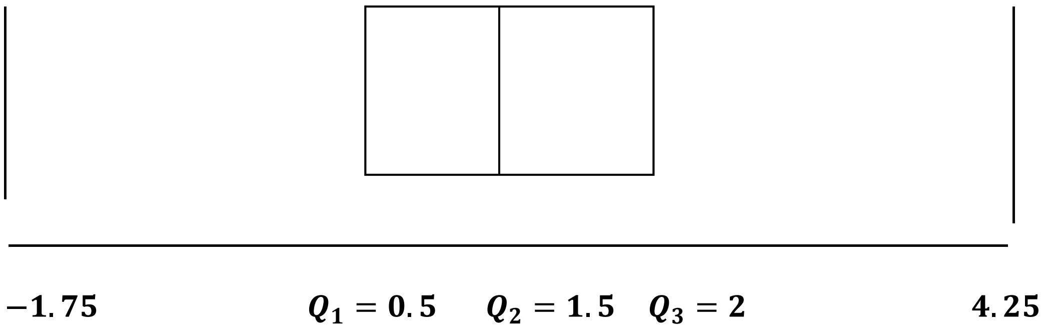
$$Q_1 = \frac{0+1}{2} = 0.5, Q_3 = \frac{2+2}{2} = 2,$$

$$IQR = Q_3 - Q_1 = 2 - 0.5 = 1.5.$$

(h)

Minimum	Q_1	Q_2	Q_3	Maximum
0	0.5	1.5	2	2

(i)



Example 3:

Suppose we have the following data:

Rent	420 ~439	440 ~459	460 ~479	480 ~499	500 ~519
Frequency	8	17	12	8	7
Rent	520 ~539	540 ~559	560 ~579	580 ~599	600 ~619
Frequency	4	2	4	2	6

What are the mean rent and the sample variance for the rent?

[Solution:]

f_k is the frequency of class k, M_k is the midpoint of class k and n is the sample size. Then,

Rent	420 ~439	440 ~459	460 ~479	480 ~499	500 ~519
f_k	8	17	12	8	7
M_k	429.5	449.5	469.5	489.5	509.5
Rent	520 ~539	540 ~559	560 ~579	580 ~599	600 ~619
f_k	4	2	4	2	6
M_k	529.5	549.5	569.5	589.5	609.5

Thus,

$$\sum_{k=1}^{10} f_k M_k = 34525 \Longrightarrow \overline{x}_g = \frac{34525}{70} = 493.21.$$

For the sample variance,

$$s_g^2 = \frac{\sum_{k=1}^{10} f_k \left(M_k - \overline{x}_g\right)^2}{70-1} = \frac{208234.287}{69} = 3017.89.$$

Example 4:

(a) Consider a sample with data values of $10, 20, 12, 17$ and 16. Compute the z-score for each of the five observations.

(b) Suppose the data have a bell-shaped distribution with a mean of 20 and a standard deviation of 5. Use both Chebyshev's theorem and the empirical rule to determine the percentage of data within the range $10-30$.

[Solution:]

(a)

Since

$$\overline{x} = \frac{10+20+12+17+16}{5} = 15$$

and

$$s = \sqrt{\frac{(10-15)^2+(20-15)^2+(12-15)^2+(17-15)^2+(16-15)^2}{5-1}}$$

$$= \sqrt{\frac{64}{4}} = 4,$$

$$\boldsymbol{x_1 = 10: z = \frac{x_1 - \bar{x}}{s} = \frac{10 - 15}{4} = -1.25,}$$

$$\boldsymbol{x_2 = 20: z = \frac{x_2 - \bar{x}}{s} = \frac{20 - 15}{4} = 1.25,}$$

$$\boldsymbol{x_3 = 12: z = \frac{x_3 - \bar{x}}{s} = \frac{12 - 15}{4} = -0.75,}$$

$$\boldsymbol{x_4 = 17: z = \frac{x_4 - \bar{x}}{s} = \frac{17 - 15}{4} = 0.5,}$$

$$\boldsymbol{x_5 = 16: z = \frac{x_5 - \bar{x}}{s} = \frac{16 - 15}{4} = 0.25.}$$

(b)

$$\boldsymbol{[10, 30] = 20 \pm 10 = \bar{x} \pm 2s.}$$

Thus, by Chebyshev's theorem, within 2 standard deviation, there is at least

$$\boldsymbol{\left(1 - \frac{1}{2^2}\right) \cdot 100\% = 75\%}$$

of data within the range $\boldsymbol{10 - 30}$.

By empirical rule, there are approximately 95% of the data values will be within this interval.