CHAPTER 2

DESCRIPTIVE STATISTICS

After measurements are collected then we should describe them. There are two ways:

- Graphically.
- Numerically.

Graphically

Each graph should provide with the following information:

- What values of variable have been measured.
- How often those values have occurred.

Stem & Leaf Graph

To create complete the following steps:

1. Divide the measurements into two parts.
2. List the stems in a column.
3. Record the leaves of each stem.
4. Sort the leaves from the lowest to the highest.

Example: Measure the show size of 15 people.

- 34, 31, 30, 38, 30, 41, 42,
- 36, 43, 40, 36, 46, 45, 41, 39.

3|4 1 0 8 0 6 6 9	→	3|0 0 1 4 6 6 8 9	Unit of leaf = 1.
4|1 2 3 0 6 5 1	→	4|0 1 1 2 3 5 6	Unit of stems = 10.

Frequency Histogram

To create, complete the following steps:

1. Number of classes $\approx \sqrt{n}$, number is always rounded up.
2. Length of classes $\approx \frac{Largest\ measurement\ -\ Smallest\ one}{Number\ of\ classes}$, number is always rounded up.
3. If measurements are discrete then each one of them can be taken as one class.
4. Locate the classes' boundaries
    a. First boundary = Minimum measurement.
    b. Second boundary = First boundary + Length.
    c. Third boundary = Second boundary + Length.
5. Construct a statistical table.
    a. Example:

Classes	Boundaries	Frequency	Relative Frequency
1 2 . . . k	First boundary → Secondary boundary Secondary boundary → Third boundary . . . …	f_1 f_2 . . . f_k n	f_1/n f_2/n . . . f_k/n 1

Shape of Distribution

- Symmetric.
  - If it forms a mirror image about the middle class.
- Skewed to the right.
  - If the greater portion of the measurement lie on the right hand side of the class of the peak.
- Skewed to the left.
  - If the greater portion of the measurement lie on the left hand side of the class of the peak.

Describing Central Tendency

Parameter is a number calculated using the measurement in the population.

The Mean (Average)

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

Point Estimator

A point estimator is a one number value computed based on a sample of size n to approximate or estimate a parameter of interest of the population. Example:

- We have the annual income of 5 canadians
  - \$25,000, \$50,000, \$35,000, \$90,000, \$35,000
  - $\overline{x} = \frac{25000 + 50000 + 35000 + 90000 + 35000}{5} = \$47,000$

Median

It is the number that is larger than or equal to 50% of the measurement. Steps include:

1. Order the measurements from the lowest to the highest
2. If n is an odd number $M_d = x_{(\frac{n+1}{2})}$
3. If n is an even number the $M_d = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$

Example (odd):

- Find M_d
  - 8, 8, 10, 5, 5, 6, -1
  - -1, 5, 5, 6, 8, 8, 10
- $M_d = x_{(\frac{7+1}{2})} = x_4 = 6$

Example (even):

- Find M_d
  - 8, 8, 10, 5, 5, 6, -1, -2
  - -2, -1, 5, 5, 6, 8, 8, 10
- $M_d = \frac{x_{(\frac{8}{2})} + x_{(\frac{8}{2}+1)}}{2} = \frac{5+6}{2} = 5.5$

Mode

The mode is the category(s) that occur more often comparing to the rest (M_o). Example:

- 1, 1, -3, -1, 3, 1, 15, 25, 3
- $M_o = 1$ & 3

When there is one mode, the distribution is called unimodal, and when there is two modes it is called bimodal.

Measures of Variation

The range of a set of data is the difference between the largest and smallest values:

- $Range = Largest\ measure - Smallest\ measure$

Variance is the measure of how far a set of numbers is spread out:

- $Sample\ variance = s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Standard deviation shows how much variance from the mean exists:

- $Standard\ deviation = \sqrt{\sigma^2}$

Chebyshev's Theorem

For a given number $k \geq 1$ and a set of measurements of size n, then at least $1 - \frac{1}{k^2}$ will lie inside the interval $[\bar{x} - ks,\ \bar{x} + ks]$.

Empirical Rule

If the distribution of measurements is mound shaped (bell shaped), Chebyshev's theorem is the only thing that can be used.

Z-Score

$Z - Score = \frac{x - \bar{x}}{s}$

$If\ |Z - Score| > 3,\ then\ x\ is\ an\ outlier$. Example:

- $\bar{x} = \frac{-1 + 10 + 15 + 9}{4} = 8.25$

  $\sum_{i=1}^{4} x_i = 33,\ \sum_{i=1}^{4} x_i^2 = 407$

  $s^2 = \frac{1}{3}[407 - \frac{(33)^2}{4}] = 44.91$

  $s = \sqrt{44.91} \approx 6.7$

  $Z - Score\ for\ -1 = \frac{-1 - 8.25}{6.7} = -1.38\ = |-1.38| = 1.38$

  $1.38 < 3\ and\ therefore\ is\ not\ an\ outlier$.

Coefficient of Variation Percentiles

In the ordered (from the lowest to the highest) the $100th \cdot p$ percentile is the number that is larger than or equal to $(100 \cdot p)\%$ of the measurement and smaller than $(100 \cdot (1 - p))\%$ of them.

How to find p*th* percentile

The location of it is $p(n + 1)\ then\ x_{(p(n+1))}$. Example:

- Find the 80th percentile
  - 2, 3, 15, 17, 22, 23, 31, 35, 38
  - $0.8(9 + 1) = 8$

    $x_8 = 35$
- The 80th percentile = 35

Quartile

There are three quartiles:

- Q_1 is the lower quartile and is the 25th percentile $(p = 0.25 = 1/4)$
- Q_2 is the second quartile and is the 50th percentile $(p = 0.5 = 2/4)$, also known as the median
- Q_3 is the upper quartile and is the 75th percentile $(p = 0.75 = 3/4)$

When $p(n + 1)$ is not an integer for the required quartile:

- The required quartile = The just below value + Remainder x (The just above value - The just below value)

Five Number Summary

The numbers include:

- Min
- Lower quartile
- Median
- Upper quartile
- Max

Box and Whisker (Box Plot) Display

$Interquartile\ range = IQR = Q_3 - Q_1$

$Inner\ fence = Q_1 - 1.5(Q_3 - Q_1) = Q_1 - 1.5(IQR)$

$Outer\ fence = Q_3 + 1.5(Q_3 - Q_1) = Q_3 + 1.5(IQR)$

Describing Qualitative (Categorical) Data

You can describe these qualities using bar charts.