Course: Probability and Statistics

Faculty: Dr. Kota Venkata Ratnam
Module 1: Introduction to Statistics
Lesson 2: Understanding the Data

Topic: Understanding the Data

Reading Objectives:

In this reading, you will learn about the spread or variability of the data values, which is accomplished by the sample variance. You will be introduced to Chebyshev's inequality and the one-sided Chebyshev's inequality. You will also look into the normal data set and skewness of data.

Main Reading Section:

1. Sample variance and sample standard deviation

A statistic that could be used for this purpose would be one that measures the average value of the squares of the distances between the data values and the sample mean.

Definition 1: The sample variance, call it s^2, of the data set $x_1, x_2, \ldots\ldots, x_n$ is defined by

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

Definition 2: The quantity *s*, defined by $s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{(n-1)}}$ is called the sample standard deviation.

Example 3.1: Find the sample variance and standard deviation of the data sets 3, 4, 6, 7, 10.

Solution: As the sample mean for data set A is $\bar{x} = (3 + 4 + 6 + 7 + 10)/5 = 6$.

Then, the sample variance is $s^2 = [(-3)^2 + (-2)^2 + (0)^2 + (1)^2 + (4)^2]/4 = 7.5$.

The standard deviation of the data is $s = \sqrt{7.5} = 2.738$.

Definition 3: The sample 100p percentile is that data value such that at least 100p percent of the data are less than or equal to it and at least 100(1 − p) percent are greater than or equal to it. If two data values satisfy this condition, then the sample 100p percentile is the arithmetic average of these two values.

Definition 4: The sample 25 percentile is called the first quartile; the sample 50 percentile is called the sample median or the second quartile; the sample 75 percentile is called the third quartile.

2. Chebyshev's inequality

Let $\bar{x}$ and s be the sample mean and sample standard deviation of the data set consisting of the data $x_1, x_2, \ldots\ldots, x_n$, where s > 0. Let:

S_k = {i, 1 ≤ i ≤ n : $|x_i - \bar{x}| < ks$}; and

let $|S_k|$ be the number of elements in the set S_k. Then, for any $k \geq 1$

$$\frac{|S_k|}{n} \geq 1 - \frac{n-1}{n k^2} > 1 - \frac{1}{k^2}$$

3. The one-sided Chebyshev inequality

Let $\bar{x}$ and s be the sample mean and sample standard deviation of the data set consisting of the data $x_1, x_2, \ldots\ldots, x_n$. Suppose $s > 0$ and let $N(k)$ = number of i: $x_i - \bar{x} \geq ks$. Then, for any $k > 0$,

$$\frac{N(k)}{n} \leq \frac{1}{1+k^2}$$

Example 1: Table lists the 10 top-selling passenger cars in the United States in the month of June 2013. Find the interval in which data values are expected to lie.

Top Selling Vehicles June 2013 Sales (in thousands of vehicles)	
Ford F Series	68.0
Chevrolet Silverado	43.3
Toyota Camry	35.9
Chevrolet Cruze	32.9
Honda Accord	31.7
Honda Civic	29.7
Dodge Ram	29.6
Ford Escape	28.7
Nissan Altima	26.9
Honda CR – V	26.6

Solution:

A simple calculation yields that the sample mean and sample standard deviation of these data are

$\bar{x}$= 35.33 and s = 11.86.

Thus, Chebyshev's inequality states that at least 100(5/9) = 55.55 percent of the data lies in the interval

$(\bar{x} - \frac{3}{2}s, \bar{x} + \frac{3}{2}s) = (17.54, 53.12)$

whereas, in actuality, 90 percent of the data falls within these limits.

4. Normal data sets and skewness of data

- Many of the large data sets observed in practice have histograms that are similar in shape. These histograms often reach their peaks at the sample median and then decrease on both sides of this point in a bell-shaped symmetric fashion. Such data sets are said to be **normal** and their histograms are called **normal histograms**.
- If the histogram of a data set is close to being a normal histogram, then we say that the data set is **approximately normal**.
- Any data set that is not approximately symmetric about its sample median is said to be **skewed**.
- It is "skewed to the right" if it has a long tail to the right and "skewed to the left" if it has a long tail to the left.

Scatter plot and correlation coefficient

- A useful way of portraying a data set of paired values is to plot the data on a two-dimensional graph, with the *x*-axis representing the *x* value of the data and the *y*-axis representing the *y* value. Such a plot is called a **scatter diagram**.

Definition 5: Consider the data pairs (x_i, y_i), $i = 1,\ldots,n$ and let s_x and s_y denote, respectively, the sample standard deviations of the *x* values and the *y* values. The sample correlation coefficient, call it *r*, of the data pairs (x_i, y_i), $i = 1,\ldots, n$ is defined by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, s_x s_y}$$

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

When $r > 0$, we say that the sample data pairs are positively correlated, and when $r < 0$, then we say that they are negatively correlated.

Properties of *r*

- $-1 \le r \le 1$
- If for constants a and b, with b > 0, $y_i = a + b\, x_i$, $i = 1,\ldots,n$ then $r = 1$.

- If for constants a and b, with b < 0, y_i = a + b x_i, i = 1,…,n then r = −1.
- If r is the sample correlation coefficient for the data pairs x_i, y_i, i = 1,…,n then it is also the sample correlation coefficient for the data pairs
  a + b x_i, c + d y_i, I = 1,…,n provided that b and d are both positive or both negative.

Reading Summary

In this reading, you have learned the following:

- Statistics that describe the spread or variability of the data values
- The Chebyshev's inequality and the one-sided Chebyshev's inequality
- The normal data sets and its skewness
- How the data variables are correlated using the correlation coefficient