Before starting the main subject of this week, let's have a quick review of central statistical values including: mean, median, mode. These values are related to the center of data.

Definition 1

To calculate the mean or average of data, add all values and divide them to the number of values.

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

where $\bar{x}$ is the mean, n is the total number of values, $x_1, x_2, \ldots, x_n$ are observations.

Definition 2

Median is the middle value of data. To calculate median, first sort data. The middle value or the average of middle values – depend on number of data – is median.

Definition 3

Mode is the value with the most frequency. A dataset may have more than one mode.

Example 1

Find the mean, median and mode of: 2,4,3,6,3,4,3,5,3,2

Answer:

Mean: $\bar{x} = \frac{2+4+3+6+3+4+3+5+3+2}{10} = \frac{35}{10} = 3.5$

Median: Sort data- 2,2,3,3,3,3,4,4,5,6. There are two values 3 and 3 in the middle. The average or median of 3 and 3 is (3+3)/2=3.

Mode: There are four 3s in values. So, 3 is the mode.

Variance and Standard Deviation are dispersion or variation values. These values show the deviation from mean.

Definition 4

Range is the difference of maximum and minimum values.

Definition 5

Formula of standard deviation:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{\text{n-1}}}$$

where μ is mean and n is the total number of data. Low standard deviation means the values are all closed to average and high standard deviation means data are far from average.

Definition 6

Variance is the square of standard deviation. Variance is expectation of the squared deviation from mean value.

Example 2

Find the range, standard deviation and variance of: 2,4,3,6,3,4,3,5,3,2. Interpret the results.

Answer:

Range: 6-1=4

Standard Deviation: $\sigma = \sqrt{\frac{2*(2-3.5)^2+4*(3-3.5)^2+2*(4-3.5)^2+(5-3.5)^2+(6-3.5)^2}{10-1}} = \sqrt{\frac{14.5}{9}} \sim \sqrt{1.6} \sim 1.3$

Variance: 1.6

Mean is 3.5, Range is 4 and standard deviation is 1.3. In range 4, 1.3 is a not a big difference. So, it shows that data are close to average.

Three more statistical valued that we use in this course are: z score, quantile and percentile.

Definition 7

Z score is a value to show the normality of one value. It shows that if one value is in a usual area or not. If z is between -2 and 2 for value x, then x falls in the usual area.

$$z = \frac{x - \mu}{\sigma}$$

Example 3

The mean blood sugar is 120 mg/dl (milligrams per deciliter) and the standard deviation is 80. Does someone with a 100 level of blood sugar has diabetics or low-level sugar?

Answer: $z = \frac{100-120}{80} = -.25 > -2$, So, this person has the normal level of blood sugar.

Definition 7

Confidence interval is a range of values with true values included in that range.

$$\bar{x} - z\frac{S}{\sqrt{n}} \leq CI \leq \bar{x} + z\frac{S}{\sqrt{n}}$$

Corresponding z scores to CI are in table 1. We normally want to find the values in 95% confidence interval.

Confidence Interval	Z
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

Table 1: Corresponding z scores to CI

Example 4

Find the confidence interval values for 100 women, with height's mean 160 and S=15 and 99% confidence.

Answer: $\bar{x} - z\frac{S}{\sqrt{n}} \leq CI \leq \bar{x} + z\frac{S}{\sqrt{n}}$. Then 160-2.576*15/10≤CI≤160+2.576*15/10. So, 156.136≤CI≤163.864. It means the confidence range of women's height almost is between 156 and 164.

Three more statistical valued that we use in this course are: z score, quantile and percentile.

Definition 8

Quantile is splitting data to n parts. So, each part has 1/n length. The last value of each part shows the quantile. The set should be sort increasing or decreasing.

Example 5

Find 3 quantiles of set {2,5,6,8,12,15,20}.

Answer: Q_1=5, Q_2=8, Q_3=15.

Definition 9

The n^{th} percentile is the lowest score in a dataset that is greater than a percentage n of all scores. For example, for .25 percentile is the lowest score greater than 25% of scores in the dataset. To calculate the percentile of n, divide the number of values less than n by size of dataset and multiply by 100.

$$\mathbf{Percentile(n)} = \frac{\mathbf{number\ of\ values\ less\ than\ n}}{\mathbf{size\ of\ dataset}} * \mathbf{100}$$

Example 6

Find the percentile (8) in set {2,5,6,8,12,15,20}.

Answer: Percentile (8)=3/7*100 ~ 42.8%. It means almost 43% of data are less than 8.

Definition 10

There is a relationship between percentile and quantile. Q_0 = 0^{th} percentile, Q_1=25^{th} percentile, Q_2=50^{th} percentile, Q_3=75^{th} percentile, Q_4=100^{th} percentile.

Discussion 1:

- **Make a supervised and unsupervised learning of features of data of week1.**
- **How to split data to train and test data?**