Topic for the class-Measures of central tendency, variation, quartiles and percentiles

Unit _3 : Title-Descriptive statistics

Date & Time : 23.8.24 10.00 AM – 10.50 AM

Dr. Bhramaramba Ravi

Professor

Department of Computer Science and Engineering

GITAM School of Technology (GST)

Visakhapatnam – 530045

Email: **bravi@gitam.edu**

Unit3-syllabus

- **UNIT 3 Descriptive statistics 9 hours, P - 2 hours**
- Measures of Central Tendency – Measures of Variation – Quartiles and Percentiles –

  Moments – Skewness and Kurtosis. Exploratory Data Analytics Descriptive Statistics – Mean,

  Standard Deviation, Skewness and Kurtosis – Box Plots – Pivot Table – Heat Map – Correlation Statistics – ANOVA, Random variable, Variance, covariance, and correlation- Linear transformations of random variables, Regression.
- https://www.coursera.org/learn/data-visualization-r

Measures of central tendency

- Of the various ways in which a variable can be summarized, one of the most important is the value used to characterize the center of the set of values it contains.
- It is useful to quantify the middle or central location of a variable, such as its average, around which many of the observations values for that variable lie.
- There are several approaches to calculating this value and which is used can depend on the classification of the variable.
- The following sections describe some common descriptive statistical approaches for calculating the central location: the *mode*, the *median*, and the *mean*.

Mode

- **Mode**
- The *mode* is the most commonly reported value for a particular variable.
- The mode calculation is illustrated using the following variable whose values (after being ordered from low to high) are

  3, 4, 5, 6, 7, 7, 7, 8, 8, 9
- The mode would be the value 7 since there are three occurrences of 7 (more than any other value).
- The mode is a useful indication of the central tendency of a variable, since the most frequently occurring value is often toward the center of the variable's range.
- When there is more than one value with the same (and highest) number of occurrences, either all values are reported or a midpoint is selected.
- For example, for the following values, both 7 and 8 are reported three times:

  3, 4, 5, 6, 7, 7, 7, 8, 8, 8, 9
- The mode may be reported as *{*7, 8*}* or 7.5.
- Mode provides the only measure of central tendency for variables measured on a nominal scale; however, the mode can also be calculated for variables measured on the ordinal, interval, and ratio scales.

Median

- The *median* is the middle value of a variable, once it has been sorted from low to high.
- The following set of values for a variable will be used to illustrate:

  3, 4, 7, 2, 3, 7, 4, 2, 4, 7, 4
- Before identifying the median, the values must be sorted:

  2, 2, 3, 3, 4, 4, 4, 4, 7, 7, 7
- There are 11 values and therefore the sixth value (five values above and five values below) is selected as the median value, which is 4:

  2, 2, 3, 3, 4, *4*, 4, 4, 7, 7, 7
- For variables with an even number of values, the average of the two values closest to the middle is selected (sum the two values and divide by 2).
- The median can be calculated for variables measured on the ordinal, interval, and ratio scales and is often the best indication of central tendency
- for variables measured on the ordinal scale. It is also a good indication of the central value for a variable measured on the interval or ratio scales since, unlike the mean, it will not be distorted by extreme values.

Mean

- The *mean*—commonly referred to as the average—is the most commonly used summary of central tendency for variables measured on the interval or ratio scales.
- It is defined as the sum of all the values divided by the number of values.
- For example, for the following set of values:

  3, 4, 5, 7, 7, 8, 9, 9, 9
- The sum of all nine values is (3 + 4 + 5 + 7 + 7 + 8 + 9 + 9 + 9) or 61.
- The sum divided by the number of values is 61 ÷ 9 or 6.78.
- $\bar{x}$

Mean contd.

The sum of all nine values is $(3 + 4 + 5 + 7 + 7 + 8 + 9 + 9 + 9)$ or 61. The sum divided by the number of values is $61 \div 9$ or 6.78.

For a variable representing a subset of all possible observations (x), the mean is commonly referred to as $\bar{x}$. The formula for calculating a mean, where n is the number of observations and x_i is the individual values, is usually written:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The notation $\sum_{i=1}^{n}$ is used to describe the operation of summing all values of x from the first value ($i = 1$) to the last value ($i = n$), that is $x_1 + x_2 + \cdots + x_n$.

Measures of variation

- **Range**
- The range is a simple measure of the variation for a particular variable. It is calculated as the difference between the highest and lowest values.
- The following variable will be used to illustrate:

  2, 3, 4, 6, 7, 7, 8, **9**
- The range is 7 calculated from the highest value (9) minus the lowest value (2). Ranges can be used with variables measured on an ordinal, interval, or ratio scale.

Variance

- The *variance* describes the spread of the data and measures how much the values of a variable differ from the mean.
- For variables that represent only a sample of some population and not the population as a whole, the variance formula is

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

The sample variance is referred to as s^2. The actual value (x_i) minus the mean value ($\bar{x}$) is squared and summed for all values of a variable. This value is divided by the number of observations minus 1 ($n - 1$).

The following example illustrates the calculation of a variance for a particular variable:

Variance

$$3, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9$$

where the mean is

$$\bar{x} = \frac{3 + 4 + 4 + 5 + 5 + 5 + 6 + 6 + 6 + 7 + 7 + 8 + 9}{13}$$

$$\bar{x} = 5.8$$

Table 2.3 is used to calculate the sum, using the mean value of 5.8.

To calculate s^2, we substitute the values from Table 2.3 into the variance formula:

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

$$s^2 = \frac{34.32}{13 - 1}$$

$$s^2 = 2.86$$

Variance

- The variance reflects the average squared deviation and can be calculated for variables measured on the interval or ratio scale.

TABLE 2.3 Variance Intermediate Steps

x	$\bar{x}$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
3	5.8	−2.8	7.84
4	5.8	−1.8	3.24
4	5.8	−1.8	3.24
5	5.8	−0.8	0.64
5	5.8	−0.8	0.64
5	5.8	−0.8	0.64
6	5.8	0.2	0.04
6	5.8	0.2	0.04
6	5.8	0.2	0.04
7	5.8	1.2	1.44
7	5.8	1.2	1.44
8	5.8	2.2	4.84
9	5.8	3.2	10.24
			Sum = 34.32

Standard deviation

- The *standard deviation* is the square root of the variance. For a sample
- from a population, the formula is

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

where s is the sample standard deviation, x_i is the actual data value, $\bar{x}$ is the mean for the variable, and n is the number of observations. For a calculated variance (e.g., 2.86) the standard deviation is calculated as $\sqrt{2.86}$ or 1.69.

The standard deviation is the most widely used measure of the deviation of a variable. The higher the value, the more widely distributed the variable's data values are around the mean. Assuming the frequency distribution is approximately normal (i.e., a bell-shaped curve), about 68% of all observations will fall within one standard deviation of the mean (34% less than and 34% greater than). For example, a variable has a mean value of 45 with a standard deviation value of 6. Approximately 68% of the observations should be in the range 39–51 (45 ± one standard deviation) and approximately 95% of all observations fall within two standard deviations

Standard deviation

- of the mean (between 33 and 57). Standard deviations can be calculated for variables measured on the interval or ratio scales.

Quartiles

- Quartiles divide a continuous variable into four even segments based on the number of observations. The first quartile (Q1) is at the 25% mark,
- the second quartile (Q2) is at the 50% mark, and the third quartile (Q3) is at the 75% mark. The calculation for Q2 is the same as the median value (described earlier).
- The following list of values is used to illustrate how quartiles are calculated:

  3, 4, 7, 2, 3, 7, 4, 2, 4, 7, 4
- The values are initially sorted:

  2, 2, 3, 3, 4, 4, 4, 4, 7, 7, 7
- Next, the median or Q2 is located in the center:

  2, 2, 3, 3, 4, **4**, 4, 4, 7, 7, 7
- We now look for the center of the first half (shown underlined) or Q1:

  2, 2, **3**, 3, 4, **4**, 4, 4, 7, 7, 7

The value of Q1 is recorded as 3.

Quartiles

- Finally, we look for the center of the second half (shown underlined) or Q3:

  The value of Q3 is identified as 7.

- When the boundaries of the quartiles do not fall on a specific value, the quartile value is calculated based on the two numbers adjacent to the boundary.
- The *interquartile range* is defined as the range from Q1 to Q3.
- In this example it would be 7 − 3 or 4.

Box plots

- Box plots provide a succinct summary of the overall frequency distribution of a variable.
- Six values are usually displayed: the lowest value, the lower quartile (Q1), the median (Q2), the upper quartile (Q3), the highest value, and the mean.
- In the conventional box plot displayed in Figure 2.8, the box in the middle of the plot represents where the central 50% of observations lie.
- A vertical line shows the location of the median value and a dot represents the location of the mean value.
- The horizontal line with a vertical stroke between “lowest value” and “Q1” and “Q3” and “highest value” are
- the “tails”—the values in the first and fourth quartiles.

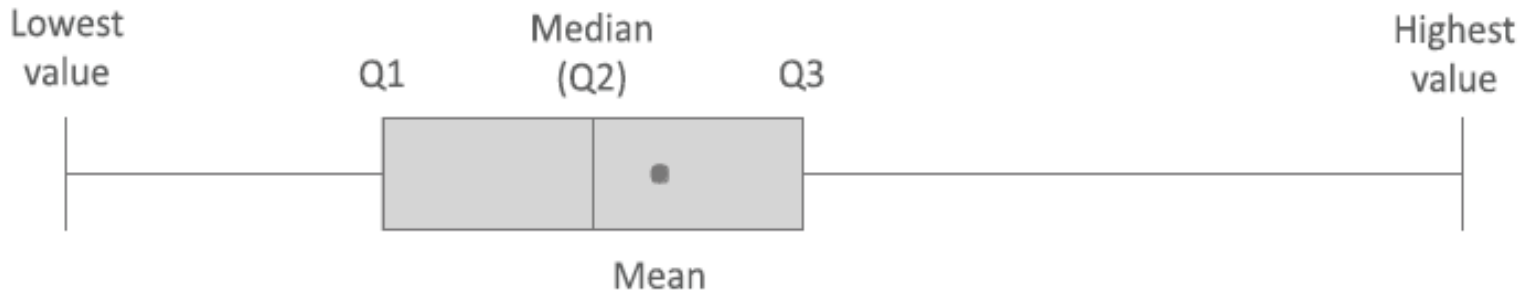

FIGURE 2.8 Overview of elements of a box plot.

Box plots contd.

- Figure 2.9 provides an example of a box plot for one variable (*MPG*).
- The plot visually displays the lower (9) and upper (46.6) bounds of the variable.
- Fifty percent of observations begin at the lower quartile (17.5)

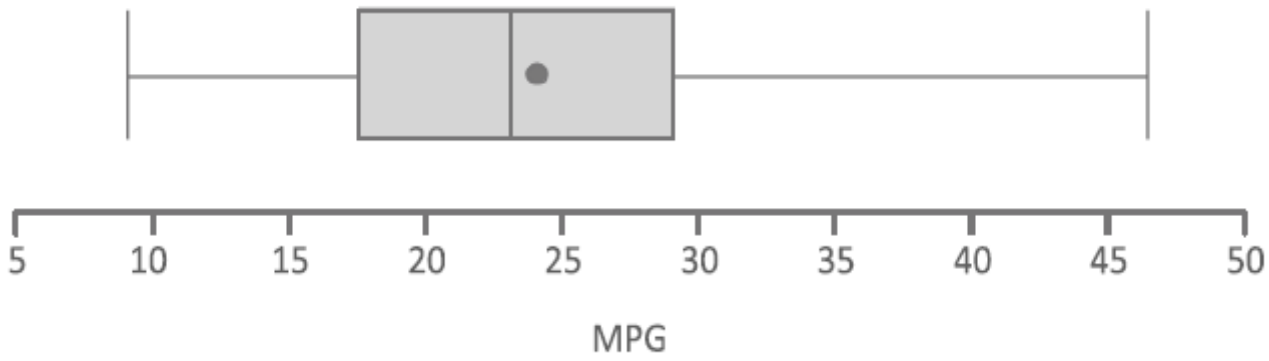

FIGURE 2.9 Box plot for the variable *MPG*.

Box plots contd.

- and end at the upper quartile (29). The median and the mean values are close, with the mean slightly higher (around 23.6) than the median (23).

Percentile

- In statistics, a percentile is a term that describes how a score compares to other scores from the same set. While there is no universal definition of percentile, it is commonly expressed as the percentage of values in a set of data scores that fall below a given value.
- Percentiles are a type of quantiles, obtained adopting a subdivision into 100 groups. The 25th percentile is also known as the first *quartile* (Q_1), the 50th percentile as the *median* or second quartile (Q_2), and the 75th percentile as the third quartile (Q_3). For example, the 50th percentile (median) is the score *below* (or *at or below*, depending on the definition) which 50% of the scores in the distribution are found.

THANK YOU