Topic for the class-skewness and kurtosis
Unit _3 : Title-Descriptive statistics
Date & Time : 2.9.24 11.00 AM – 11.50 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

Shape

- We know how to visualize frequency distributions.
- In addition to these visualizations, there are methods for quantifying the lack of symmetry or *skewness* in the distribution of a variable.
- For asymmetric distributions, the bulk of the observations are either to the left or the right of the mean.
- For example, in Figure 2.12 the frequency distribution is asymmetric and more of the observations are to the left of the mean than to the right; the right tail is longer than the left tail.
- This is an example of a positive, or right skew.
- Similarly, a negative, or left skew would have more of the observations to the right of the mean value with a longer tail on the left.

Frequency distribution showing a positive skew

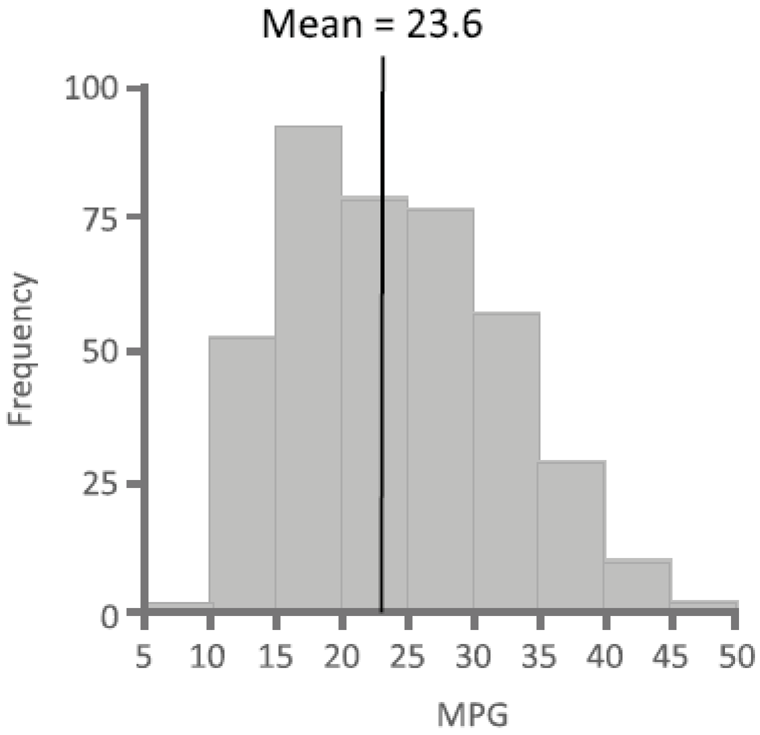


FIGURE 2.12 Frequency distribution showing a positive skew.

Skewness

- It is possible to calculate a value for skewness that describes whether the variable is positively or negatively skewed and the degree of skewness.
- One formula for estimating skewness, where the variable is *x* with individual values *xi*, and *n* data values is

$$skewness = \left(\frac{\sqrt{n \times (n-1)}}{n-2} \right) \times \frac{{}^{1}\!/_{n} \times \sum_{i=1}^{n} (x_i - \bar{x})^3}{\left({}^{1}\!/_{n} \times \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{3/2}}$$

Skewness

- A skewness value of zero indicates a symmetric distribution.
- f the lower tail is longer than the upper tail the value is positive; if the upper tail is longer than the lower tail, the skewness score is negative.
- Figure 2.13 shows examples of skewness values for two variables.
- The variable *alkphos* in the plot on the left has a positive skewness value of 0.763, indicating that the

  majority of observations are to the left of the mean, whereas the negative skewness value for the variable *mcv* in the plot on the right indicates that the majority are to the right of the mean.
- That the skewness value for *mcv* is closer to zero than *alkphos* indicates that *mcv* is more symmetric than

  alkphos.

Skewness

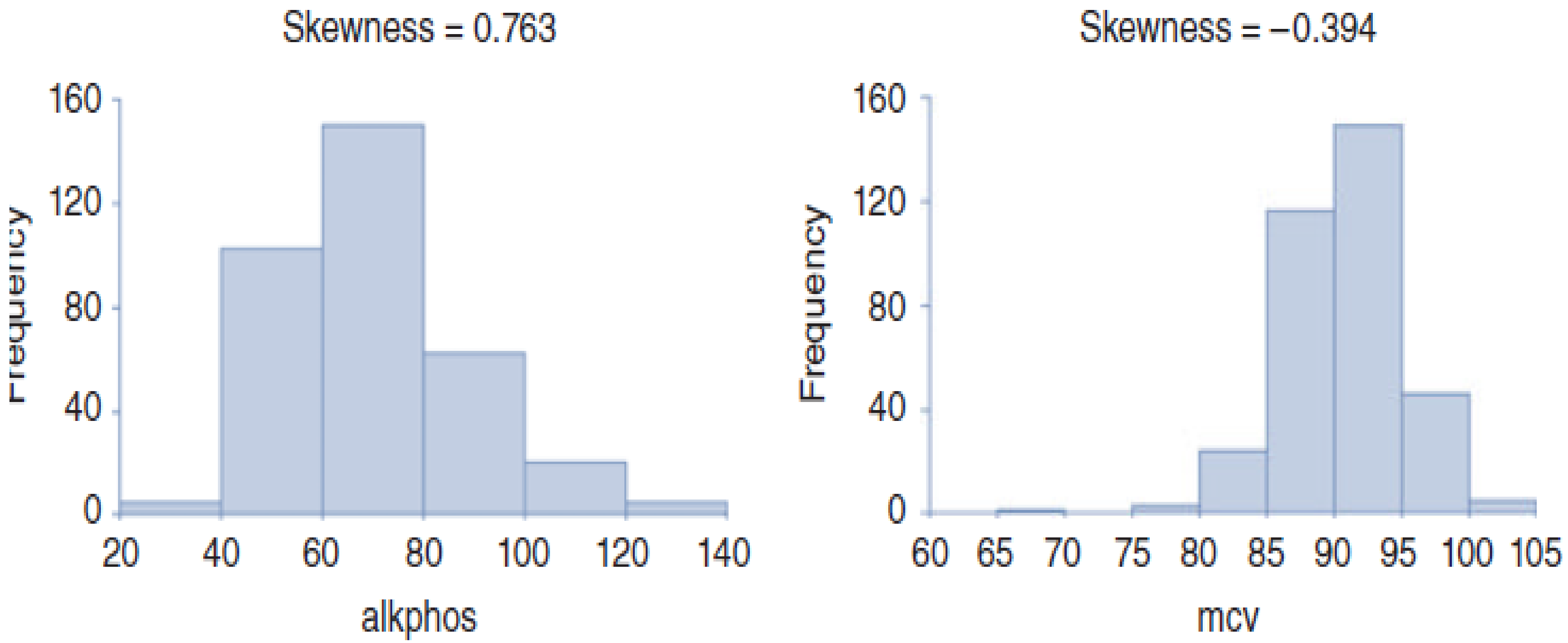


FIGURE 2.13 Skewness estimates for two variables.

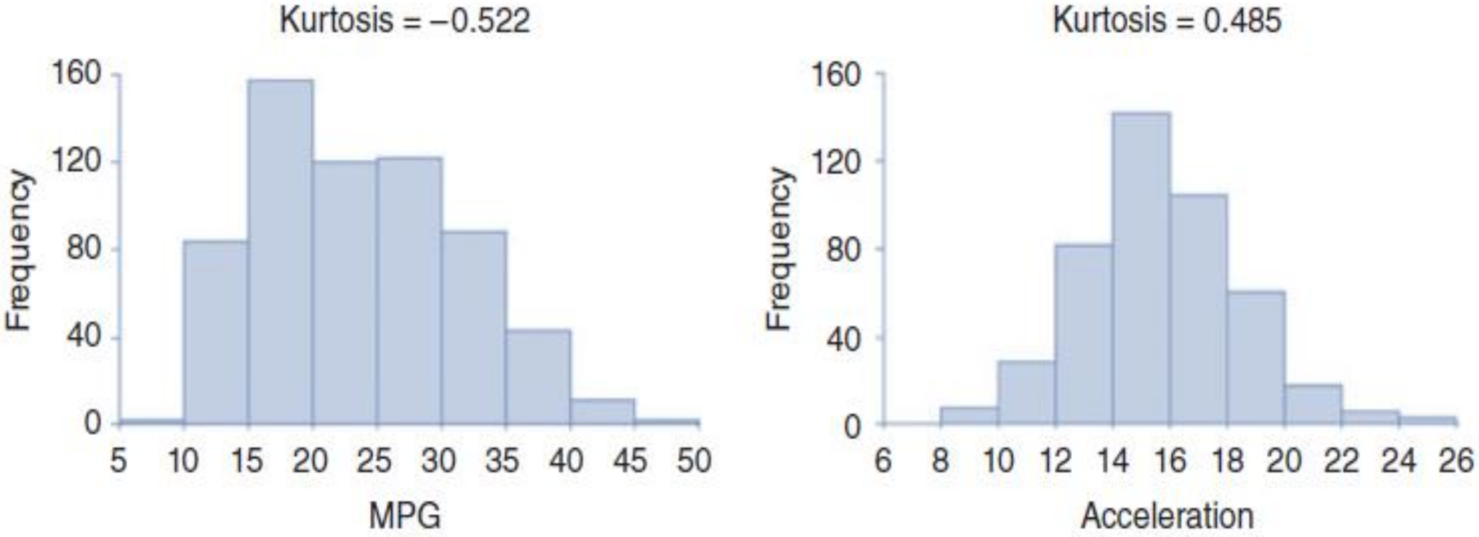
Kurtosis

- In addition to the symmetry of the distribution, the type of peak thet distribution has should be considered and it can be characterized by a measurement called *kurtosis*.
- The following formula can be used for calculating kurtosis for a variable *x*, where *xi* represents the individual values, and *n* the number of data values:

$$kurtosis = \frac{n-1}{(n-2)\times(n-3)} \times \left((n+1) \times \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4 / n}{\left(\sum_{i=1}^{n} (x_i - x)^2 / n \right)^2} - 3 \right) + 6$$

Kurtosis

- Variables with a pronounced peak near the mean have a high kurtosis score while variables with a flat peak have a low kurtosis score.
- Figure 2.14 illustrates kurtosis scores for two variables.
- Fig. Kurtosis Estimates for 2 Variables



Kurtosis

- It is important to understand whether a variable has a normal distribution, since a number of data analysis approaches require variables to have this type of frequency distribution.
- Values for skewness and kurtosis close to zero indicate that the shape of a frequency distribution for a variable
- approximates a normal distribution which is important for checking assumptions in certain data analysis methods.

THANK YOU