

3-2 Distribution Shape: Skewness

Last class, we learned measures of location and measures of variability. We know that they can be used to characterize a data set. In fact, a fundamental task in many statistical analyses is to characterize the **location** and **variability** of a data set. However, even if two data sets are the same in terms of the location and variability, they may still have different distributions. Therefore, we need some other measures to further characterize a data set. One of the measures is skewness. It provides some insight into the distribution shape of a data set.

Skewness is a measure of asymmetry, or, the lack of symmetry. The histogram can help to learn the skewness of a data set. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. A skewed (asymmetric) distribution is a distribution that's lack of the symmetry.

- For a sample x_1, x_2, \dots, x_n , the formula for skewness is:

$$skewness = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)s^3},$$

where \bar{x} is the mean, s is the standard deviation, and n is the number of data points.

- Any symmetric data should have a skewness zero, i.e. it's not skewed.
- 0 skewness implies mean and median are equal.
- Negative values for the skewness indicate the data are skewed left. By skewed left, we mean that the left tail is long relative to the right tail. Here's a left skewed histogram, i.e. the skewness is negative.
- Positive values for the skewness indicate data are skewed right. Skewed right means that the right tail is long relative to the left tail. Here's a right skewed histogram, i.e. the skewness is positive.

Z-Scores

An observation's z-score is a measure of the relative location of the observation in a data set.

It denotes the number of standard deviations a data value x_i is from the mean.

$$z_i = \frac{x_i - \bar{x}}{s}$$

In other words, it measures the difference between a data value and the mean in units of the standard deviation.

- The z-score is often called the standardized value.
- A data value less than the sample mean will have a z-score less than zero.
- A data value greater than the sample mean will have a z-score greater than zero.
- A data value equal to the sample mean will have a z-score of zero.

Z-Scores can be used to detect outliers

Outliers are extreme values, or unusually small or unusually large values in a data set. :observations "far away" from the rest of the data

- A data value with a z-score less than -3 or greater than +3 might be considered an outlier.

It might be:

- an incorrectly recorded data value,
- a data value that was incorrectly included in the data set,
- a correctly recorded data value that belongs in the data set.

Measures of Association between two variables

Covariance: the covariance is a measure of linear association between two variables. For a sample of size n with the n observations on two variables x and y , the sample covariance is defined as follows:

Sample covariance

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

Population covariance

$$\sigma_{xy} = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{N}$$

Compare the formulas with those of variances?

- A positive value indicates a positive association between x and y , i.e., as the value of x increases, the value of y increases. (scatter diagram for number of commercials and sales)
- A negative value indicates a negative association between x and y , i.e., as the value of x increases, the value of y decreases.
- 0 indicates no linear association between x and y , i.e., as the value of x increases, the value of y may increase as well as decrease.

Correlation coefficient

Correlation coefficient is a numerical measure that indicates the degree of correlation between two sets of data.

The **correlation coefficient** between two variables or two sets of data equals their covariance divided by the product of their standard deviations

Correlation coefficient for a sample

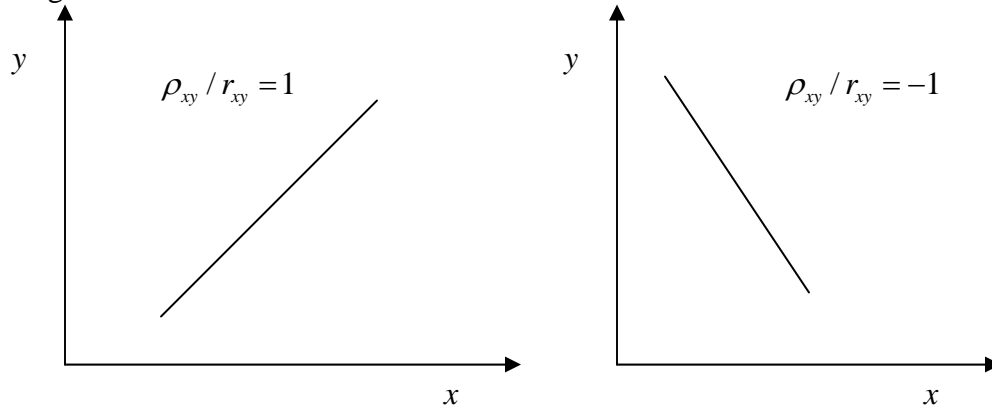
$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Correlation coefficient for a population

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- The correlation coefficient can take on values between -1 and +1
- Correlation coefficient has the same sign as covariance, since the standard deviation is always positive
- A positive correlation coefficient indicates a positive relationship; a negative correlation coefficient indicates a negative relationship, while correlation coefficient equal to 0 indicates no linear relationship between x and y .

- Particularly, correlation coefficient =1 indicates a perfect positive correlation between x and y; correlation coefficient =-1 indicates a perfect negative correlation between x and y. perfect correlation means y is a linear function of x. $Y=a+bx$. In a scatter diagram for x and y, you will observe all the points fall on a straight line.



- Strong positive linear association: The points should lie roughly on a straight line that slopes upwards to the right. The value of r should be 1 or close to 1.
- Strong negative linear association: The points should lie roughly on a straight line that slopes downwards to the right. The value of r should be -1 or close to -1.
- No linear association: The points could be scattered all over the grid with no pattern at all. The value of r should be zero or close to zero. Or the points could be in a pattern that has a strong shape other than linear, for example a circle, or a strong curve, which could have an r value close to zero.
- $r = 1$. Points exactly on a straight line sloping upwards to the right. $r = -1$. Points exactly on a straight line sloping downwards to the right. $r = 0$. Points randomly scattered over the grid.
- If the outlier is close to the other points, the scatter plot may still look roughly linear. If the outlier is far away, the line will be pulled away from the original points. The r value will be close to 1 if the outlier is close to the original points, but will be farther away from 1 if the outlier is farther away.