Chapter 1

Exploratory Data Analysis - Alternative Measures of Centrality

Contents

1.1 Alternatice Measures of Centrality

1.1.1 The Weighted Mean

- The weighted mean (or weighted average) is an arithmetic mean in which each value is weighted according to its importance in the overall group.

- The formulas for the population, and sample weighted means are identical:

$$\mu_w = \bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$$

1.1.2 The Winsorized mean

- The winsorized mean is less sensitive to outliers because it replaces them with less influential values. This method of averaging is similar to the trimmed mean; however, instead of eliminating data, observations are altered, allowing for a degree of influence.

- Let's calculate the first winsorized mean for the following data set: 1, 5, 7, 8, 9, 10, 14. Because the winsorized mean is in the first order, we replace the smallest and largest values with their nearest observations. The data set now appears as follows: 5, 5, 7, 8, 9, 10, 10. Taking an arithmetic average of the new set produces a winsorized mean of 7.71 ((5+5+7+8+9+10+10) / 7).

1.1.3 The Trimean

- The trimean is computed by adding the 25th percentile plus twice the 50th percentile plus the 75th percentile and dividing by four.
- What follows is an example of how to compute the trimean. The 25th, 50th, and 75th percentile of the dataset "Example 1" are 51, 55, and 63 respectively.
- Therefore, the trimean is computed as: The trimean is almost as resistant to extreme scores as the median.

1.1.4 Trimean

The **trimean** (TM) is a measure of central location defined as a weighted average of the distribution's median and its two quartiles:

$$TM = \frac{Q_1 + 2Q_2 + Q_3}{4}$$

This is equivalent to the average of the median and the midhinge:

$$TM = \frac{1}{2}\left(Q_2 + \frac{Q_1 + Q_3}{2}\right)$$

$$TM = \frac{Q_1 + 2Q_2 + Q_3}{4}$$

For the sample data, the trimean is computed as

$$TM = \frac{25 + (2 \times 27) + 31}{4} = \frac{110}{4} = 27.5$$

1.1.5 The trimmed mean

The trimmed mean looks to reduce the effects of outliers on the calculated average. This method is best suited for data with large, erratic deviations or extremely skewed distributions. A trimmed mean is stated as a mean trimmed by X%, where X is the sum of the percentage of observations removed from both the upper and lower bounds.

For example, a figure skating competition produces the following scores: 6.0, 8.1, 8.3, 9.1, 9.9. A mean trimmed 40% would equal 8.5 ((8.1+8.3+9.1)/3), which is larger than the arithmetic mean of 8.28. To trim the mean by 40%, we remove the lowest 20% and highest 20% of values, eliminating the scores of 6.0 and 9.1. As shown by this example, trimming the mean can reduce the effects of outlier bias in a sample.

A trimmed mean is calculated by discarding a certain percentage of the lowest and the highest scores and then computing the mean of the remaining scores. For example, a mean trimmed 50% is computed by discarding the lower and higher 25% of the scores and taking the mean of the remaining scores.A trimmed mean is obviously less susceptible to the effects of extreme scores than is the arithmetic mean. Trimmed means are often used in Olympic scoring to minimize the effects of extreme ratings possibly caused by biased judges.

Statistics

var1

N	Valid	11.000
	Missing	.000
Mean		10.182
Std. Error of Mean		1.828
Median		9.000
Mode		1.000[a]
Std. Deviation		6.063
Variance		36.764
Skewness		.042
Std. Error of Skewness		.661
Kurtosis		-1.501
Std. Error of Kurtosis		1.279
Range		18.000
Minimum		1.000
Maximum		19.000
Sum		112.000
Percentiles	25	5.000
	50	9.000
	75	16.000

a. Multiple modes exist. The smallest value is shown

Figure 1.1: SPSS Descriptive Statistics.

1.1.6 Midhinge

The midhinge is a measure of central location, determined as the average of the first and third quartiles.

$$\text{midhinge} = \frac{Q_1 + Q_3}{2}$$

For the sample data, the midhinge is computed as

$$\text{midhinge} = \frac{25 + 31}{2} = \frac{56}{2} = 28$$

1.1.7 Trimmed Means

For certain data sets, an option is available to trim the extreme values from the distribution of values of a variable. For example, we can trim (i.e., remove) the lowest 5% and the highest 5% from the distribution of values. The mean of the trimmed distribution of values is referred to as a "trimmed mean".

1.1.8 The Geometric Mean

- The Geometric mean is a specialized measure used to calculate the average proportional changes.
- Geometric mean formula
  $G = \sqrt[n]{(1+p_1) \times (1+p_1) + \ldots + (1+p_n)}$
- The price of a commodity changes by the following percentages over a period of four years.
- Compute the average price change.
- How to find the $n-th$ root using your calculator.
- What is the n-th root of the number X
  $\sqrt[n]{X} \; = \; X^{\frac{1}{n}}$
- for example
  $\sqrt[5]{11} \; = \; 11^{\frac{1}{5}} \; = \; 11^{0.2}$

	Year 1	Year 2	Year 3	Year 4
Change	10%	5%	-8%	12%

The four terms we are going to multiply are 1.10 , 1.05 0.92 and 1.12.

$$= (1.190112)^{0.25} = 1.044472 \tag{1.1}$$

Mean Absolute Deviation (MAD)

The mean absolute deviation (MAD) uses the absolute values of the deviations from the mean and perhaps gives us a more intuitively understandable measure of deviation than variance and standard deviation.

- The mean absolute deviation, or MAD, is based on the absolute value of the difference between each value in the data set and the mean of the group.
- It is sometimes called the average deviation.
- The mean average of these absolute values is then determined.
- The absolute values of the differences are used because the sum of all of the plus and minus differences (rather than the absolute differences) is always equal to zero.

- Thus the respective formulas for the population and sample MAD are

$$\text{Population MAD } = \frac{\sum |x_i - \mu|}{N}$$

$$\text{Sample MAD } = \frac{\sum |x_i - \bar{x}|}{n}$$

1.1.9 The Coefficient of Variation

The coefficient of variation is the ratio of the standard deviation to the mean, and is a useful statistic for comparing the degree of variation from one data series to another, particularly if the means are drastically different from each other.
When comparing different distributions are compared, the distribution with the highest CV value has the greatest dispersion.

- This is a statistical measure of the dispersion of data points in a data series around the mean.
- It is calculated as follows:

$$\text{CV} = \frac{\sigma}{\mu}$$

$$\text{CV} = \frac{s}{\bar{x}}$$

- The coefficient of variation represents the ratio of the standard deviation to the mean
- The coefficient of variation is a useful statistic for comparing the degree of dispersion from one data set to another, even if the means are very different from each other.
- The coefficient of variation, CV, indicates the relative magnitude of the standard deviation as compared with the mean of the distribution of measurements, as a percentage.
- What happens if you have two sets of data with two different means and two different standard deviations? How do you decide which set is more spread out? Remember the size of the standard deviation is relative to the mean it is associated with.
- The coefficient of variation (cv) is often used to compare the relative dispersion between two or more sets of data. It is formed by dividing the standard deviation by the mean and is usually expressed as a percentage i.e. (multiplied by 100).
- Again we distinguish between the population and sample coefficient of variation. Thus, the formulas are

$$\text{Population : } CV = \frac{\sigma}{\mu} \times 100$$

$$\text{Sample : } CV = \frac{s}{\bar{x}} \times 100$$

The coefficient of variation is useful when we wish to compare the variability of two data sets relative to the general level of values (and thus relative to the mean) in each set. The coefficient of variation for different distribution are compared and the distribution with the largest CV value has the greatest spread.

1.2 Computing The Coefficient of Variation

The coefficient of variation (CV) is calculated by dividing the sample standard deviation (s) by the absolute value of the sample mean ($|\bar{x}|$). The coefficient of variation is normally expressed as a percentage.
The formula is as follows:

$$CV = \frac{s}{|\bar{x}|} \times 100\%$$

N.B. Remember that the standard deviation is the square root of the variance.

Example 1

Data on the delivery time (in hours) of purchases were collected for a random sample of similar orders sent to a building supplies company. The following summary statistics are available:

mean	median	variance	maximum	minimum	IQR
12	13	16	20	2	8

Use the data to compute the coefficient of variation.

Example 2

In 2000, the average donation to a charity was €225 with a standard deviation of €45.
In 2001, the average donation to the same charity was €400 with a standard deviation of €60.
What are the coefficients of variation for both years?

In which year do the donations show a more dispersed distribution?

1.3 Coefficient of variation

The coefficient of variation, CV, indicates the relative magnitude of the standard deviation as compared with the mean of the distribution of measurements, as a percentage. Thus, the formulas are

$$\text{Population}: CV = \frac{\sigma}{\mu} \times 100$$
$$\text{Sample}: CV = \frac{s}{\bar{x}} \times 100$$

The coefficient of variation is useful when we wish to compare the variability of two data sets relative to the general level of values (and thus relative to the mean) in each set.

1.3.1 Computing the Skewness Coefficient

- n Sample Size
- s Sample Standard Deviation
- $\bar{x}$ Sample Mean
-

Computing the Skewness Coefficient

$$S_k = \frac{(n-1)(n-2)}{s} \times \frac{\sum (x_i - \bar{x})^3}{x}$$

1.4 Skewness

Computing the Skewness Coefficient

- n Sample Size
- s Sample Standard Deviation
- $\bar{x}$ Sample Mean
-