Sullivan Section 3.2

Gene Quinn

Measures of Dispersion

By "measures of central tendency" we mean:

measures that numerically describe the **average** or **typical** data value.

Measures of Dispersion

By "measures of central tendency" we mean:

measures that numerically describe the **average** or **typical** data value.

Central tendency measures tell only part of the story.

Usually, we also need to know how data values "spread out" or disperse around an average.

Measures of Dispersion

By "measures of central tendency" we mean:

measures that numerically describe the **average** or **typical** data value.

Central tendency measures tell only part of the story.

Usually, we also need to know how data values "spread out" or disperse around an average. Measures of this characteristic of data are known as **measures of dispersion**

Measures of Dispersion

In most applications of statistics, both measures of central tendency and measures of dispersion are needed.

Measures of Dispersion

In most applications of statistics, both measures of central tendency and measures of dispersion are needed.

Measures of dispersion give an indication of how precise or how reliable an average is.

Measures of Dispersion

In most applications of statistics, both measures of central tendency and measures of dispersion are needed.

Measures of dispersion give an indication of how precise or how reliable an average is.

Stating an average by itself is much less useful than stating an average together with some measure of its reliability.

The Range

The simplest measure of dispersion is the **range**.

Definition: The **range R** of a variable is the difference between the largest data value and the smallest:

$$\text{Range} = R = \text{Largest data value} - \text{Smallest data value}$$

The Range

Example:

If we measured the weights of a group of people and obtained the following data,

$$146, 185, 157, 225, 120, 190, 216$$

The Range R would be:

$$\text{Range} = R = \text{Largest data value} - \text{Smallest data value}$$

$$= 225 - 120 = 105$$

Deviation About the Mean

Because we are interested in how data spreads out around the mean, we define dispersion in terms of the **deviation about the mean** or deviation from the mean.

Deviation About the Mean

Because we are interested in how data spreads out around the mean, we define dispersion in terms of the **deviation about the mean** or deviation from the mean.

Since we have two kinds of means, the population mean μ and the sample mean $\overline{x}$, we also have two kinds of deviations about the mean.

Deviation About the Mean

For the deviation of the i^{th} data value about the *population* mean, we compute

$$x_i - \mu$$

Deviation About the Mean

For the deviation of the i^{th} data value about the *population* mean, we compute

$$x_i - \mu$$

For the deviation of the i^{th} data value about the *sample* mean, we compute

$$x_i - \overline{x}$$

Deviation About the Mean

Since we are interested in the **average** deviation about μ for a population,

or the **average** deviation about $\overline{x}$ for a sample,

in either case, the average deviation about the appropriate mean would seem to be the natural choice.

Deviation About the Mean

Since we are interested in the **average** deviation about μ for a population,

or the **average** deviation about $\overline{x}$ for a sample,

in either case, the average deviation about the appropriate mean would seem to be the natural choice.

Unfortunately, there is a problem with this approach.

In both cases, the average deviation about the mean is **always** zero, because of the way that the means are calculated:

Deviations can be positive or negative, and the positive and negative deviations cancel each other out.

Deviation About the Mean

One way to prevent this cancellation is to **square** the deviations before we add them up.

(Remember, when you square a positive or negative number, the result is always positive)

Deviation About the Mean

One way to prevent this cancellation is to **square** the deviations before we add them up.

(Remember, when you square a positive or negative number, the result is always positive)

This technique used in a measure of dispersion called the **variance**.

Deviation About the Mean

One way to prevent this cancellation is to **square** the deviations before we add them up.

(Remember, when you square a positive or negative number, the result is always positive)

This technique used in a measure of dispersion called the **variance**.

As with means, we will need to define two variances: a *population* variance and a *sample* variance.

The Population Variance

Definition:

The **population variance** of a variable, denoted by the symbol σ^2, is the sum of the squared deviations from the population mean, taken over the entire population:

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}$$

The Population Variance

Definition:

The **population variance** of a variable, denoted by the symbol σ^2, is the sum of the squared deviations from the population mean, taken over the entire population:

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}$$

Using summation notation, the formula for the population variance can be written more compactly as

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

The Population Variance

Definition:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

In this formula, μ is the population mean, and $x_1, x_2, \ldots, x_N$ are the N observations in the population.

The Sample Variance

Definition:

The **sample variance** of a variable, denoted by the symbol s^2, is the sum of the squared deviations from the sample mean, taken over the sample:

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n - 1}$$

The Sample Variance

Definition:

The **sample variance** of a variable, denoted by the symbol s^2, is the sum of the squared deviations from the sample mean, taken over the sample:

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

Using summation notation, the formula for the sample variance can be written more compactly as

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

The Sample Variance

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

In this formula, $\overline{x}$ is the sample mean, and $x_1, x_2, \ldots, x_n$ are the n observations in the population.

The Sample Variance

Note that the divisor in the sample variance is $n - 1$ and not n.

The reason for this is not obvious, but it relates to the fact that usually we use the sample variance to get an estimate of the population variance.

The Sample Variance

Note that the divisor in the sample variance is $n-1$ and not n.

The reason for this is not obvious, but it relates to the fact that usually we use the sample variance to get an estimate of the population variance.

It can be shown mathematically that if we use $n-1$, the sample variance will be free of **bias** as an estimator of the population variance.

We say that an estimator has no bias, or is unbiased, if it has no tendency to either consistently overestimate or consistently underestimate the population variance.

The Sample Variance

If we used n as the divisor in computing s^2, we would obtain a **biased** estimate of the population variance σ^2 (s^2 would consistently underestimate σ^2). The value $n - 1$ in the sample variance is referred to as the number of **degrees of freedom**

Again, the reason for this is not obvious.

The Sample Variance

If we used n as the divisor in computing s^2, we would obtain a **biased** estimate of the population variance σ^2 (s^2 would consistently underestimate σ^2). The value $n - 1$ in the sample variance is referred to as the number of **degrees of freedom**

Again, the reason for this is not obvious.

This term originates from the fact that, if we have n numbers and their mean, one of the numbers is superfluous.

The Sample Variance

If we used n as the divisor in computing s^2, we would obtain a **biased** estimate of the population variance σ^2 (s^2 would consistently underestimate σ^2). The value $n-1$ in the sample variance is referred to as the number of **degrees of freedom**

Again, the reason for this is not obvious.

This term originates from the fact that, if we have n numbers and their mean, one of the numbers is superfluous.

That is, if I know the sample mean and $n-1$ of the n sample values, I can always figure out what the n^{th} sample value is.

Another way to say this is that, once I have $n-1$ data values and the mean, I have no freedom in the choice of the n^{th} data value.

The Standard Deviation

Most types of statistical analyses use a measure called the **standard deviation** to measure dispersion.

As usual, we have two standard deviations, one for the population, and one for a sample.

The Standard Deviation

Most types of statistical analyses use a measure called the **standard deviation** to measure dispersion.

As usual, we have two standard deviations, one for the population, and one for a sample.

The **population standard deviation**, denoted by σ, is the square root of the population variance:

$$\sigma = \sqrt{\sigma^2}$$

The Standard Deviation

Most types of statistical analyses use a measure called the **standard deviation** to measure dispersion.

As usual, we have two standard deviations, one for the population, and one for a sample.

The **population standard deviation**, denoted by σ, is the square root of the population variance:

$$\sigma = \sqrt{\sigma^2}$$

The **sample standard deviation**, denoted by s, is the square root of the sample variance:

$$s = \sqrt{s^2}$$

The Standard Deviation

The standard deviation is a measure of dispersion in the sense that, the larger the standard deviation, the more the data values are dispersed around the mean.

The Empirical Rule

If data has a distribution that is roughly bell shaped, the following statements regarding the *population* mean and standard deviation will be true:

The Empirical Rule

If data has a distribution that is roughly bell shaped, the following statements regarding the *population* mean and standard deviation will be true:

Approximately 68 percent of the data will lie within 1 standard deviation of the population mean (that is, between $\mu - \sigma$ and $\mu + \sigma$).

The Empirical Rule

If data has a distribution that is roughly bell shaped, the following statements regarding the *population* mean and standard deviation will be true:

Approximately 68 percent of the data will lie within 1 standard deviation of the population mean (that is, between $\mu - \sigma$ and $\mu + \sigma$).

Approximately 95 percent of the data will lie within 2 standard deviations of the population mean (that is, between $\mu - 2\sigma$ and $\mu + 2\sigma$).

The Empirical Rule

If data has a distribution that is roughly bell shaped, the following statements regarding the *population* mean and standard deviation will be true:

Approximately 68 percent of the data will lie within 1 standard deviation of the population mean (that is, between $\mu - \sigma$ and $\mu + \sigma$).

Approximately 95 percent of the data will lie within 2 standard deviations of the population mean (that is, between $\mu - 2\sigma$ and $\mu + 2\sigma$).

Approximately 99.7 percent of the data will lie within 3 standard deviations of the population mean (that is, between $\mu - 3\sigma$ and $\mu + 3\sigma$).

Percentiles

We defined the **median** M of a measure for a population or a sample to be the value for which 50% of the population or sample have a lower value than M, and 50% have a higher value than M.

Percentiles

We defined the **median** M of a measure for a population or a sample to be the value for which 50% of the population or sample have a lower value than M, and 50% have a higher value than M.

We can generalize this idea to percentages other than 50%. These generalizations are called **percentiles**.

Percentiles

We defined the **median** M of a measure for a population or a sample to be the value for which 50% of the population or sample have a lower value than M, and 50% have a higher value than M.

We can generalize this idea to percentages other than 50%. These generalizations are called **percentiles**.

We say P_{90} is the 90^{th} percentile of a measure for a population if 90% of the individuals in the population have a value of P_{90} or less.

Percentiles

We defined the **median** M of a measure for a population or a sample to be the value for which 50% of the population or sample have a lower value than M, and 50% have a higher value than M.

We can generalize this idea to percentages other than 50%. These generalizations are called **percentiles**.

We say P_{90} is the 90^{th} percentile of a measure for a population if 90% of the individuals in the population have a value of P_{90} or less.

The median M is the same as the 50^{th} percentile, P_{50}.

Quartiles

Percentiles for which the percentage is a multiple of 25 are called **quartiles**.

Quartiles

Percentiles for which the percentage is a multiple of 25 are called **quartiles**.

The 25^{th} percentile P_{25} is called the **first quartile** Q_1.

Quartiles

Percentiles for which the percentage is a multiple of 25 are called **quartiles**.

The 25^{th} percentile P_{25} is called the **first quartile** Q_1.

The 50^{th} percentile P_{50} is called the **second quartile** Q_1. (and is also the median M)

Quartiles

Percentiles for which the percentage is a multiple of 25 are called **quartiles**.

The 25^{th} percentile P_{25} is called the **first quartile** Q_1.

The 50^{th} percentile P_{50} is called the **second quartile** Q_1. (and is also the median M)

The 75^{th} percentile P_{75} is called the **third quartile** Q_3.

Interquartile Range

The difference between the 75^{th} percentile and the 25^{th} percentile is called the **interquartile range**

Interquartile Range

The difference between the 75^{th} percentile and the 25^{th} percentile is called the **interquartile range**

The interquartile range is a measure of dispersion, like the variance and standard deviation.

Interquartile Range

The difference between the 75^{th} percentile and the 25^{th} percentile is called the **interquartile range**

The interquartile range is a measure of dispersion, like the variance and standard deviation.

The interquartile range is much less sensitive to outliers or extreme values in the dataset than the variance and standard deviation.

Interquartile Range

The difference between the 75^{th} percentile and the 25^{th} percentile is called the **interquartile range**

The interquartile range is a measure of dispersion, like the variance and standard deviation.

The interquartile range is much less sensitive to outliers or extreme values in the dataset than the variance and standard deviation.