Sullivan Section 3.4

Gene Quinn

Measures of Position

In this section, we introduce a new group of measures.

Measures of Position

In this section, we introduce a new group of measures.

Measures of central tendency are used to describe the "average" or typical data value.

Measures of Position

In this section, we introduce a new group of measures.

Measures of central tendency are used to describe the "average" or typical data value.

Measures of dispersion are used to describe the degree to which data values are "spread out".

Measures of Position

In this section, we introduce a new group of measures.

Measures of central tendency are used to describe the "average" or typical data value.

Measures of dispersion are used to describe the degree to which data values are "spread out".

We now introduce **measures of position**, which are used to indicate the relative position of a certain data value within the entire set of data values.

Z Scores

The most natural way to indicate relative position within a population would seem to be the deviation from the population mean:

$$x - \mu$$

Z Scores

The most natural way to indicate relative position within a population would seem to be the deviation from the population mean:

$$x - \mu$$

One drawback of this measure is that datasets with different dispersion will have different values for the same relative position:

In a population with a small standard deviation, a data value with a relatively small deviation may in fact be at an extreme in terms of the overall distribution of data.

Z Scores

The most natural way to indicate relative position within a population would seem to be the deviation from the population mean:

$$x - \mu$$

One drawback of this measure is that datasets with different dispersion will have different values for the same relative position:

In a population with a small standard deviation, a data value with a relatively small deviation may in fact be at an extreme in terms of the overall distribution of data.

On the other hand, in a population with a large standard deviation, a data value with a relatively large deviation might be relatively close to the middle.

Z Scores

The Z-score attemts to adjust for this by dividing by the standard deviation.

As with the mean and standard deviation, there are two kinds of Z-scores.

Population Z Scores

Definition:

The **population** Z-score for a data value x is defined as

$$z = \frac{x - \mu}{\sigma}$$

Data Value Z Scores

Definition:

The **sample** Z-score for a data value x is defined as

$$z = \frac{x - \overline{x}}{s}$$

Some Facts About Z Scores

The same symbol, z is used for both the population and sample Z-scores.

Some Facts About Z Scores

The same symbol, z is used for both the population and sample Z-scores.

Because of the way it is calculated, the mean of the Z-scores for either a population or a sample is always 0.

Some Facts About Z Scores

The same symbol, z is used for both the population and sample Z-scores.

Because of the way it is calculated, the mean of the Z-scores for either a population or a sample is always 0.

The standard deviation of the Z-scores from a population or sample is always 1.

Some Facts About Z Scores

The same symbol, z is used for both the population and sample Z-scores.

Because of the way it is calculated, the mean of the Z-scores for either a population or a sample is always 0.

The standard deviation of the Z-scores from a population or sample is always 1.

The term **standardized scores** is often used to data values that have been converted to Z-scores.

Percentiles

The **median** is the value X with the property that 50% of the data values fall below it and 50% fall above it.

Percentiles

The **median** is the value X with the property that 50% of the data values fall below it and 50% fall above it.

This idea can be generalized by allowing any percentage to be the dividing line.

These more general measures are called **percentiles**.

Percentiles

The **median** is the value X with the property that 50% of the data values fall below it and 50% fall above it.

This idea can be generalized by allowing any percentage to be the dividing line.

These more general measures are called **percentiles**.

Definition: The k^{th} **percentile** P_k is the number that divides the lower $k\%$ of the data set from the upper $(100 - k)\%$.

Percentiles

The number k can be any whole number from 1 to 99, inclusive.

Percentiles

The number k can be any whole number from 1 to 99, inclusive.

The 1^{st} percentile P_1 divides the bottom 1% of the data from the top 99%.

Percentiles

The number k can be any whole number from 1 to 99, inclusive.

The 1^{st} percentile P_1 divides the bottom 1% of the data from the top 99%.

The 12^{th} percentile P_{12} divides the bottom 12% of the data from the top 88%.

Percentiles

The number k can be any whole number from 1 to 99, inclusive.

The 1^{st} percentile P_1 divides the bottom 1% of the data from the top 99%.

The 12^{th} percentile P_{12} divides the bottom 12% of the data from the top 88%.

The 92^{nd} percentile P_{92} divides the bottom 92% of the data from the top 8%.

Percentiles

One of the most common applications of percentiles is to measure the relative standing.

Percentiles

One of the most common applications of percentiles is to measure the relative standing.

Scores on standardized tests like the SAT are reported together with a percentile.

Percentiles

One of the most common applications of percentiles is to measure the relative standing.

Scores on standardized tests like the SAT are reported together with a percentile.

In fact, results of the vast majority of psychoeducational tests are reported as percentiles.

Quartiles

The most commonly used percentiles are called **quartiles**.

Quartiles

The most commonly used percentiles are called **quartiles**.

Quartiles divide the data into two parts which are even multiples of 25%.

Quartiles

The most commonly used percentiles are called **quartiles**.

Quartiles divide the data into two parts which are even multiples of 25%.

Quartiles

- **1)** The first quartile is the 25^{th} percentile P_{25} which is also given the symbol Q_1 and called the **first quartile**.

Quartiles

- **1)** The first quartile is the 25^{th} percentile P_{25} which is also given the symbol Q_1 and called the **first quartile**.
- **2)** The second quartile is the 50^{th} percentile P_{50} which is the same as the median M. The second quartile is denoted by Q_2.

Quartiles

- **1)** The first quartile is the 25^{th} percentile P_{25} which is also given the symbol Q_1 and called the **first quartile**.
- **2)** The second quartile is the 50^{th} percentile P_{50} which is the same as the median M. The second quartile is denoted by Q_2.
- **3)** The third quartile is the 75^{th} percentile P_{75} which is also given the symbol Q_3.

The Interquartile Range

Quartiles are used to determine a measure called the *interquartile range*.

The Interquartile Range

Quartiles are used to determine a measure called the *interquartile range*.

Definition:

The **interquartile range** or **IQR** is the difference between the first and third quartiles,

$$\text{IQR} = Q_3 - Q_1$$

Outliers

Observations that fall at extreme values are called **outliers**.

Outliers

Observations that fall at extreme values are called **outliers**.

There are various ways to define outliers. One common way is to define as an outlier any observation that is

- 1.5 times the IQR below the first quartile Q_1 or
- 1.5 times the IQR above the third quartile Q_3

Outliers

Observations that fall at extreme values are called **outliers**.

There are various ways to define outliers. One common way is to define as an outlier any observation that is

- 1.5 times the IQR below the first quartile Q_1 or
- 1.5 times the IQR above the third quartile Q_3

The cutoff points for outliers are called **fences**.

Outliers

The procedure to check for outliers is the following:

- Determine the first and third quartiles Q_1 and Q_3

Outliers

The procedure to check for outliers is the following:

- Determine the first and third quartiles Q_1 and Q_3
- Compute the IQR which is Q_3-Q_1

Outliers

The procedure to check for outliers is the following:

- Determine the first and third quartiles Q_1 and Q_3
- Compute the IQR which is Q_3-Q_1
- Determine the **fences**:
  - The lower fence is $Q_1 - 1.5 \cdot (IQR)$
  - The upper fence is $Q_3 + 1.5 \cdot (IQR)$

Outliers

The procedure to check for outliers is the following:

- Determine the first and third quartiles Q_1 and Q_3
- Compute the IQR which is Q_3-Q_1
- Determine the **fences**:
  - The lower fence is $Q_1 - 1.5 \cdot (IQR)$
  - The upper fence is $Q_3 + 1.5 \cdot (IQR)$
- If a data value is less than the lower fence or grater than the upper fence, it is considered an outlier.

Outliers

The procedure to check for outliers is the following:

- Determine the first and third quartiles Q_1 and Q_3
- Compute the IQR which is Q_3-Q_1
- Determine the **fences**:
  - The lower fence is $Q_1 - 1.5 \cdot (IQR)$
  - The upper fence is $Q_3 + 1.5 \cdot (IQR)$
- If a data value is less than the lower fence or grater than the upper fence, it is considered an outlier.

Outliers

The procedure to check for outliers is the following:

- Determine the first and third quartiles Q_1 and Q_3
- Compute the IQR which is Q_3-Q_1
- Determine the **fences**:
  - The lower fence is $Q_1 - 1.5 \cdot (IQR)$
  - The upper fence is $Q_3 + 1.5 \cdot (IQR)$
- If a data value is less than the lower fence or grater than the upper fence, it is considered an outlier.