Descriptive Statistics (part 2)

The most important characteristic of the data in a sample is *location*, which answers the question

What is the magnitude of a typical data value?

Descriptive Statistics (part 2)

The most important characteristic of the data in a sample is *location*, which answers the question

What is the magnitude of a typical data value?

The mean and median are useful measures for answering this question.

Descriptive Statistics (part 2)

The most important characteristic of the data in a sample is *location*, which answers the question

What is the magnitude of a typical data value?

The mean and median are useful measures for answering this question.

The next most important characteristic in most cases is *variability*, also called *dispersion*

Descriptive Statistics (part 2)

The most important characteristic of the data in a sample is *location*, which answers the question

What is the magnitude of a typical data value?

The mean and median are useful measures for answering this question.

The next most important characteristic in most cases is *variability*, also called *dispersion*

Measures of dispersion answer the question

How far away from the center or mean is a typical data value?

Descriptive Statistics (part 2)

The obvious measure of dispersion, average distance from the mean, is completely useless because its value is zero for any sample.

Descriptive Statistics (part 2)

The obvious measure of dispersion, average distance from the mean, is completely useless because its value is zero for any sample.

That is, for any collection of data $x_1, x_2, \ldots, x_n$,

$$\frac{\sum_{i=1}^{n} (x_i - \overline{x})}{n} = 0$$

Descriptive Statistics (part 2)

The obvious measure of dispersion, average distance from the mean, is completely useless because its value is zero for any sample.

That is, for any collection of data $x_1, x_2, \ldots, x_n$,

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})}{n} = 0$$

The reason is that the total positive and negative deviations from the mean always cancel each other out.

Descriptive Statistics (part 2)

One way around this is to square the deviations before we add them up.

Then there is no cancellation because the numbers are all positive.

Descriptive Statistics (part 2)

One way around this is to square the deviations before we add them up.

Then there is no cancellation because the numbers are all positive.

The most common measure of variation for a sample is the **sample variance**

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

Descriptive Statistics (part 2)

One way around this is to square the deviations before we add them up.

Then there is no cancellation because the numbers are all positive.

The most common measure of variation for a sample is the **sample variance**

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

For theoretical reasons that will become clear later, we divide by $n-1$ rather than n

Descriptive Statistics (part 2)

One way around this is to square the deviations before we add them up.

Then there is no cancellation because the numbers are all positive.

The most common measure of variation for a sample is the **sample variance**

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

For theoretical reasons that will become clear later, we divide by $n-1$ rather than n

Although this causes some confusion, for a large sample it makes very little difference in the computed value.

Descriptive Statistics (part 2)

Because the deviations are squared, they tend to be on a wider scale than the original measures.

Descriptive Statistics (part 2)

Because the deviations are squared, they tend to be on a wider scale than the original measures.

To make the measurement scales closer, a related measure called the **sample standard deviation** is often used. It is defined by

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

Descriptive Statistics (part 2)

Because the deviations are squared, they tend to be on a wider scale than the original measures.

To make the measurement scales closer, a related measure called the **sample standard deviation** is often used. It is defined by

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

The notation s reflects the fact that the sample standard deviation is the (positive) square root of the sample variance s^2.

Descriptive Statistics (part 2)

We noted that outliers, observations very different from the rest of the sample, can greatly impact the sample mean $\overline{x}$.

Descriptive Statistics (part 2)

We noted that outliers, observations very different from the rest of the sample, can greatly impact the sample mean $\overline{x}$.

The situation is even worse with the sample variance s^2 and sample standard deviation s

Descriptive Statistics (part 2)

We noted that outliers, observations very different from the rest of the sample, can greatly impact the sample mean $\overline{x}$.

The situation is even worse with the sample variance s^2 and sample standard deviation s

As with measures of location, there are measures of variation based on the relative *ordering* of the data values rather than their magnitude. In general these are much less sensitive to outliers.

Descriptive Statistics (part 2)

We noted that outliers, observations very different from the rest of the sample, can greatly impact the sample mean $\overline{x}$.

The situation is even worse with the sample variance s^2 and sample standard deviation s

As with measures of location, there are measures of variation based on the relative *ordering* of the data values rather than their magnitude. In general these are much less sensitive to outliers.

Although the author uses slightly different terminology, these are usually called *quartiles*

Descriptive Statistics (part 2)

The idea of the *first quartile* Q_1 is that it represents the data value for which is larger than one quarter of the values in the sample and smaller than the other three quarters.

Descriptive Statistics (part 2)

The idea of the *first quartile* Q_1 is that it represents the data value for which is larger than one quarter of the values in the sample and smaller than the other three quarters.

The *third quartile* Q_3 on the other hand is larger than three quarters of the data values and smaller than one quarter.

Descriptive Statistics (part 2)

The idea of the *first quartile* Q_1 is that it represents the data value for which is larger than one quarter of the values in the sample and smaller than the other three quarters.

The *third quartile* Q_3 on the other hand is larger than three quarters of the data values and smaller than one quarter.

As with the median, there are minor details regarding whether there are an even or odd number of data values, but you can largely disregard these if you are using automation (which you should be).

Descriptive Statistics (part 2)

The idea of the *first quartile* Q_1 is that it represents the data value for which is larger than one quarter of the values in the sample and smaller than the other three quarters.

The *third quartile* Q_3 on the other hand is larger than three quarters of the data values and smaller than one quarter.

As with the median, there are minor details regarding whether there are an even or odd number of data values, but you can largely disregard these if you are using automation (which you should be).

You can think of the median as the second quartile. Basically the quartiles divide the ordered data into four equal parts.

Descriptive Statistics (part 2)

The *range* of the sample data is the difference between the largest and smallest data values.

Descriptive Statistics (part 2)

The *range* of the sample data is the difference between the largest and smallest data values.

Like the sample variance, the sample range is sensitive to outliers.

Descriptive Statistics (part 2)

The *range* of the sample data is the difference between the largest and smallest data values.

Like the sample variance, the sample range is sensitive to outliers.

A good alternative is the *interquartile range* defined as

$$\mathsf{I} = Q_3 - Q_1$$

Descriptive Statistics (part 2)

The *range* of the sample data is the difference between the largest and smallest data values.

Like the sample variance, the sample range is sensitive to outliers.

A good alternative is the *interquartile range* defined as

$$\mathrm{I} = Q_3 - Q_1$$

The *five number summary* shows the max, min, median, Q_1, and Q_3. Some computer implementations such as R also include the mean.

Descriptive Statistics (part 2)

A useful graphical device for summarizing data is the **box plot**. The usual definition is that a boxplot is a rectangle extending from Q_1 to Q_3. A line is drawn at the median. Thin lines called "whiskers" extend from the center of the rectangle along the x-axis to the largest and smallest data values.

Descriptive Statistics (part 2)

A useful graphical device for summarizing data is the **box plot**. The usual definition is that a boxplot is a rectangle extending from Q_1 to Q_3. A line is drawn at the median. Thin lines called "whiskers" extend from the center of the rectangle along the x-axis to the largest and smallest data values.

There are minor variations in the details depending on which implementation you use.

Descriptive Statistics (part 2)

A useful graphical device for summarizing data is the **box plot**. The usual definition is that a boxplot is a rectangle extending from Q_1 to Q_3. A line is drawn at the median. Thin lines called "whiskers" extend from the center of the rectangle along the x-axis to the largest and smallest data values.

There are minor variations in the details depending on which implementation you use.

In general you do not need to be concerned with these.

Descriptive Statistics (part 2)

A useful graphical device for summarizing data is the **box plot**. The usual definition is that a boxplot is a rectangle extending from Q_1 to Q_3. A line is drawn at the median. Thin lines called "whiskers" extend from the center of the rectangle along the x-axis to the largest and smallest data values.

There are minor variations in the details depending on which implementation you use.

In general you do not need to be concerned with these.

A Few Useful R Functions

mean(x)	the sample mean
median(x)	the sample median
var(x)	the sample variance
sd(x)	the sample standard deviation
max(x)	the maximum value
min(x)	the minimum value
summary(x)	the five number summary
boxplot(x)	a boxplot of the data
IQR(x)	the interqartile range