Descriptive Statistics

Over the last 15 years, data storage capacity has literally exploded.

Descriptive Statistics

Over the last 15 years, data storage capacity has literally exploded.

In 1990, datacenters with a terabyte (1,024 gigabytes) of storage were rare.

Descriptive Statistics

Over the last 15 years, data storage capacity has literally exploded.

In 1990, datacenters with a terabyte (1,024 gigabytes) of storage were rare.

Today one can purchase a single hard drive with that much capacity off the shelf for less than $200.

Descriptive Statistics

Over the last 15 years, data storage capacity has literally exploded.

In 1990, datacenters with a terabyte (1,024 gigabytes) of storage were rare.

Today one can purchase a single hard drive with that much capacity off the shelf for less than $200.

Naturally, inexpensive data storage creates a demand for data analysis.

Descriptive Statistics

Over the last 15 years, data storage capacity has literally exploded.

In 1990, datacenters with a terabyte (1,024 gigabytes) of storage were rare.

Today one can purchase a single hard drive with that much capacity off the shelf for less than $200.

Naturally, inexpensive data storage creates a demand for data analysis.

The first step in data analysis is to summarize the important features of the data using **descriptive statistics**

Descriptive Statistics

The two features a collection of data most often sumarized are *location* and *variability* or *dispersion*.

Descriptive Statistics

The two features a collection of data most often sumarized are *location* and *variability* or *dispersion*.

By *location*, we mean the magnitude of a typical data value, usually represented by something like an average or mean.

Descriptive Statistics

The two features a collection of data most often sumarized are *location* and *variability* or *dispersion*.

By *location*, we mean the magnitude of a typical data value, usually represented by something like an average or mean.

Definition: The **sample mean**, denoted by $\overline{x}$, of a set of *observations*

$$x_1, x_2, \ldots, x_n$$

is defined as the sum of the data values, divided by the number of values n.

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Descriptive Statistics

The two features a collection of data most often sumarized are *location* and *variability* or *dispersion*.

By *location*, we mean the magnitude of a typical data value, usually represented by something like an average or mean.

Definition: The **sample mean**, denoted by $\overline{x}$, of a set of *observations*

$$x_1, x_2, \ldots, x_n$$

is defined as the sum of the data values, divided by the number of values n.

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Descriptive Statistics

The two features a collection of data most often sumarized are *location* and *variability* or *dispersion*.

By *location*, we mean the magnitude of a typical data value, usually represented by something like an average or mean.

Definition: The **sample mean**, denoted by $\overline{x}$, of a set of *observations*

$$x_1, x_2, \ldots, x_n$$

is defined as the sum of the data values, divided by the number of values n.

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Measures of Location

The formula for the mean can also be written using summation notation as

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Measures of Location

The formula for the mean can also be written using summation notation as

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

If it is understood that the sum is to be taken over the entire sample, this is often abreviated as

$$\overline{x} = \frac{\sum x}{n}$$

Measures of Location

The formula for the mean can also be written using summation notation as

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

If it is understood that the sum is to be taken over the entire sample, this is often abreviated as

$$\overline{x} = \frac{\sum x}{n}$$

The mean is the most commonly used measure of location.

Measures of Location

The only shortcoming of the mean is that it can be heavily influenced by a single data value that is very different from the others.

Measures of Location

The only shortcoming of the mean is that it can be heavily influenced by a single data value that is very different from the others.

Suppose there are seven houses in a certain neighborhood.

We will assign the house prices, in thousands, to an array in R using the following statement:

$$x < -c(320, 340, 295, 410, 450, 377, 610)$$

Measures of Location

The only shortcoming of the mean is that it can be heavily influenced by a single data value that is very different from the others.

Suppose there are seven houses in a certain neighborhood.

We will assign the house prices, in thousands, to an array in R using the following statement:

$$x < -c(320, 340, 295, 410, 450, 377, 610)$$

We can display the results by typing:

$$x$$

Measures of Location

The only shortcoming of the mean is that it can be heavily influenced by a single data value that is very different from the others.

Suppose there are seven houses in a certain neighborhood.

We will assign the house prices, in thousands, to an array in R using the following statement:

$$x < -c(320, 340, 295, 410, 450, 377, 610)$$

We can display the results by typing:

$$x$$

which produces: [1] 320 340 295 410 450 377 610

Measures of Location

To compute the mean $\overline{x}$, the R command is:

mean(x)

Measures of Location

To compute the mean $\overline{x}$, the R command is:

```
mean(x)
```

In this case the result is:

```
[1] 400.2857
```

Measures of Location

To compute the mean $\overline{x}$, the R command is:

```
mean(x)
```

In this case the result is:

```
[1] 400.2857
```

which seems reasonably representative of this collection of seven numbers:

```
[1] 320 340 295 410 450 377 610
```

Measures of Location

Now suppose one of the houses is torn down and replaced by a mansion costing $7 million.

Measures of Location

Now suppose one of the houses is torn down and replaced by a mansion costing $7 million.

We will represent this in the data by replacing the value of $x[1]$, the first element of the x array:

```
x[1]<-7000
```

Measures of Location

Now suppose one of the houses is torn down and replaced by a mansion costing $7 million.

We will represent this in the data by replacing the value of $x[1]$, the first element of the x array:

x[1]<-7000

Now recalculate the mean with mean(x)

Measures of Location

Now suppose one of the houses is torn down and replaced by a mansion costing $7 million.

We will represent this in the data by replacing the value of $x[1]$, the first element of the x array:

x[1]<-7000

Now recalculate the mean with mean(x)

This time the result is

[1] 1354.571

Measures of Location

Now suppose one of the houses is torn down and replaced by a mansion costing $7 million.

We will represent this in the data by replacing the value of $x[1]$, the first element of the x array:

```
x[1]<-7000
```

Now recalculate the mean with mean(x)

This time the result is

```
[1] 1354.571
```

This value is not very representative of either the new expensive house, or the original houses.

Measures of Location

A data value that is very different from the others is called an *outlier*

Measures of Location

A data value that is very different from the others is called an *outlier*

The new $7 million dollar house is an outlier because it is very different from the values of the original houses.

Measures of Location

A data value that is very different from the others is called an *outlier*

The new \$7 million dollar house is an outlier because it is very different from the values of the original houses.

It is very important to detect and explain outliers, because they can strongly influence the interpretation of the data.

Measures of Location

A data value that is very different from the others is called an *outlier*

The new $7 million dollar house is an outlier because it is very different from the values of the original houses.

It is very important to detect and explain outliers, because they can strongly influence the interpretation of the data.

Example: In 1976, the U.S. Enviromnental Protection Agency was studying the concentrations of heavy metals in shellfish dredged up in the Baltimore Canyon.

Measures of Location

A data value that is very different from the others is called an *outlier*

The new $7 million dollar house is an outlier because it is very different from the values of the original houses.

It is very important to detect and explain outliers, because they can strongly influence the interpretation of the data.

Example: In 1976, the U.S. Enviromnental Protection Agency was studying the concentrations of heavy metals in shellfish dredged up in the Baltimore Canyon.

Sea scallops swim by rapidly opening and closing their shells.

Measures of Location

If the dredge happens to catch the scallop during the brief time when its shell is open, it may be dragged along the ocean floor like a scoop.

Measures of Location

If the dredge happens to catch the scallop during the brief time when its shell is open, it may be dragged along the ocean floor like a scoop.

This results in a scallop that is solidly packed with sand.

Measures of Location

If the dredge happens to catch the scallop during the brief time when its shell is open, it may be dragged along the ocean floor like a scoop.

This results in a scallop that is solidly packed with sand.

The chemists noticed this, and asked the investigators if they should wash the sand out before digesting the scallop in acid for metals analysis

Measures of Location

If the dredge happens to catch the scallop during the brief time when its shell is open, it may be dragged along the ocean floor like a scoop.

This results in a scallop that is solidly packed with sand.

The chemists noticed this, and asked the investigators if they should wash the sand out before digesting the scallop in acid for metals analysis

The investigators told them not to, because they thought it might wash out the pollutants they were trying to detect.

Measures of Location

If the dredge happens to catch the scallop during the brief time when its shell is open, it may be dragged along the ocean floor like a scoop.

This results in a scallop that is solidly packed with sand.

The chemists noticed this, and asked the investigators if they should wash the sand out before digesting the scallop in acid for metals analysis

The investigators told them not to, because they thought it might wash out the pollutants they were trying to detect.

Because the sand contained a lot of iron, this particular shellfish was reported as being 4% iron, way out of line with the other specimens where we were dealing with parts per million.