Sullivan Section 3.5

Gene Quinn

The 5-Number Summary and Boxplots

In this section, we introduce a pair of devices used in **exploratory data analysis**.

The 5-Number Summary and Boxplots

In this section, we introduce a pair of devices used in **exploratory data analysis**.

The goal of exploratory data analysis is to discover interesting or unexpected results from the data.

The 5-Number Summary and Boxplots

In this section, we introduce a pair of devices used in **exploratory data analysis**.

The goal of exploratory data analysis is to discover interesting or unexpected results from the data.

We have noted that some measures of central tendency and dispersion can be easily distorted by the presence of extreme values or outliers in the data.

Exploratory data analysis makes use of measures that are relatively insensitive to outliers;

That is, they are not changed very much if extreme values are included or excluded.

The 5-Number Summary and Boxplots

In this section, we introduce a pair of devices used in **exploratory data analysis**.

The goal of exploratory data analysis is to discover interesting or unexpected results from the data.

We have noted that some measures of central tendency and dispersion can be easily distorted by the presence of extreme values or outliers in the data.

Exploratory data analysis makes use of measures that are relatively insensitive to outliers;

That is, they are not changed very much if extreme values are included or excluded.

This is a very desirable quality for statistical measures to have.

The 5-Number Summary

Among the measures of central tendency we have studied, the most resistant to extreme values is the median.

The 5-Number Summary

Among the measures of central tendency we have studied, the most resistant to extreme values is the median.

For example, let's suppose that we measure 5 numbers and the results are:

$$1, 2, 3, 4, 5$$

The 5-Number Summary

Among the measures of central tendency we have studied, the most resistant to extreme values is the median.

For example, let's suppose that we measure 5 numbers and the results are:

$$1, 2, 3, 4, 5$$

Both the mean $\overline{x}$ and the median X are 3.

The 5-Number Summary

Instead, suppose we had the following sample:

$$1, 2, 3, 4, 900$$

The 5-Number Summary

Instead, suppose we had the following sample:

$$1, 2, 3, 4, 900$$

The median of this sample is still 3, but the mean is now about 165.

The 5-Number Summary

Instead, suppose we had the following sample:

$$1, 2, 3, 4, 900$$

The median of this sample is still 3, but the mean is now about 165.

In fact, the last number in the list can be made as large as you want and the median will still be 3.

The 5-Number Summary

Instead, suppose we had the following sample:

$$1, 2, 3, 4, 900$$

The median of this sample is still 3, but the mean is now about 165.

In fact, the last number in the list can be made as large as you want and the median will still be 3.

The median of this sample is still 3:

$$1, 2, 3, 4, 6 \text{ trillion}$$

Exploratory Data Analysis

The measures of dispersion we studied are, if anything, even more sensitive to outliers than the measures of central tendency.

Exploratory Data Analysis

The measures of dispersion we studied are, if anything, even more sensitive to outliers than the measures of central tendency.

However, the measure of dispersion we introduced later, the interquartile range $Q_3 - Q_1$, is not sensitive to outliers.

Exploratory Data Analysis

The measures of dispersion we studied are, if anything, even more sensitive to outliers than the measures of central tendency.

However, the measure of dispersion we introduced later, the interquartile range $Q_3 - Q_1$, is not sensitive to outliers.

In exploratory data analysis, the goal is to summarize the important characteristics of the dataset in terms of central tendency, dispersion, and relative position.

Exploratory Data Analysis

The measures of dispersion we studied are, if anything, even more sensitive to outliers than the measures of central tendency.

However, the measure of dispersion we introduced later, the interquartile range $Q_3 - Q_1$, is not sensitive to outliers.

In exploratory data analysis, the goal is to summarize the important characteristics of the dataset in terms of central tendency, dispersion, and relative position.

By design, exploratory data analysis makes use of measures that are resistant to extreme values.

The 5-Number Summary

One technique for exploratory data analysis called the **Five Number Summary** uses the following five measures to characterize a dataset:

MINIMUM $\quad Q_1 \quad M \quad Q_3 \quad$ MAXIMUM

Box Plots

Another device used to visually summarize data in a robust way is the **box plot**.

Box Plots

Another device used to visually summarize data in a robust way is the **box plot**.

The numbers used in the box plot are similar to the Five-Number summary, except we use the upper and lower fences,

$$\text{Lower Fence} \quad = \quad Q_1 - 1.5 \cdot (IQR)$$

and

$$\text{Upper Fence} \quad = \quad Q + 3 + 1.5 \cdot (IQR)$$

instead of the min and max.

Drawing Box Plots

The first step in drawing a Box Plot is:

Determine the lower and upper fences:

$$\begin{aligned} \text{Lower Fence} &= Q_1 - 1.5 \cdot IQR \\ \text{Upper Fence} &= Q_3 + 1.5 \cdot IQR \end{aligned}$$

Drawing Box Plots

Once we have the upper and lower fences, the remaining steps are:

- Draw vertical lines at Q_1, M, and Q_3, and enclose them in a box.

Drawing Box Plots

Once we have the upper and lower fences, the remaining steps are:

- Draw vertical lines at Q_1, M, and Q_3, and enclose them in a box.
- Label the upper and lower fences.

Drawing Box Plots

Once we have the upper and lower fences, the remaining steps are:

- Draw vertical lines at Q_1, M, and Q_3, and enclose them in a box.
- Label the upper and lower fences.
- Draw a line from Q_1 to the smallest data value larger than the lower fence.

Drawing Box Plots

Once we have the upper and lower fences, the remaining steps are:

- Draw vertical lines at Q_1, M, and Q_3, and enclose them in a box.
- Label the upper and lower fences.
- Draw a line from Q_1 to the smallest data value larger than the lower fence.
- Draw a line from Q_3 to the largest data value smaller than the lower fence.

Drawing Box Plots

Once we have the upper and lower fences, the remaining steps are:

- Draw vertical lines at Q_1, M, and Q_3, and enclose them in a box.
- Label the upper and lower fences.
- Draw a line from Q_1 to the smallest data value larger than the lower fence.
- Draw a line from Q_3 to the largest data value smaller than the lower fence.
- Mark any data values beyond the fences as outliers with an asterisk (*)