Statistics

|Outliers – Trimmed mean

Prof. Dr. Mostafa Elhosseini

https://youtube.com/drmelhosseini

Outliers

- Sometimes a sample may contain a few points that are much larger or smaller than the rest. Such points are called outliers
- Sometimes outliers result from data entry errors
- Outliers should always be scrutinized, and any outlier that is found to result from an error should be corrected or deleted. Not all outliers are errors

Outliers

- **Outliers** are a **real problem** for data analysts. For this reason, when people see outliers in their data, they sometimes try to **find a reason, or an excuse, to delete them**.
- An outlier **should not be deleted**, however, unless there is **reasonable certainty** that it results from an error.
- If a population truly contains outliers, but they are deleted from the sample, the sample **will not characterize the population correctly**.

Sample Median

- The **median, like the mean**, is a **measure of center**.
- To compute the median of a sample, **order** the values from smallest to largest.
- The sample **median** is the **middle number**.
- If the sample size is an **even number**, it is customary to take the sample median to be the **average of the two middle numbers**

If n numbers are ordered from smallest to largest:

- If n is odd, the sample median is the number in position $\frac{n+1}{2}$.
- If n is even, the sample median is the average of the numbers in positions $\frac{n}{2}$ and $\frac{n}{2}+1$.

Sample Median

- A simple random sample of five men is chosen from a large population of men, and their heights are measured. The five heights (in inches) are 65.51, 72.30, 68.31, 67.05, and 70.68. Find the sample median for the height data

Sample Median

- The five heights, arranged in increasing order, are **65.51, 67.05, 68.31, 70.68, 72.30**.
- The sample median is the middle number, which is **68.31**

Mean or Median?

- The median is often used as a measure of center for samples that contain outliers
- To see why, consider the sample consisting of the values **1, 2, 3, 4, and 20**. The mean is **6**, and the **median is 3**.
- It is reasonable to think that the median is more representative of the sample than the mean is

FIGURE 1.3 When a sample contains outliers, the median may be more representative of the sample than the mean is.

Trimmed Mean

- Like the **median**, the **trimmed mean** is a **measure of center that is designed to be unaffected by outliers**.
- The trimmed mean is computed by arranging the sample values in order, “**trimming**” an **equal number of them from each end**, and computing the mean of those remaining.
- If p% of the data are trimmed from each end, the resulting trimmed mean is called the “**p% trimmed mean**.”
- There are **no hard-and-fast rules** on how many values to trim. The most commonly used trimmed means are the **5%, 10%, and 20%** trimmed means

Trimmed Mean

In the article "Evaluation of Low-Temperature Properties of HMA Mixtures" (P. Sebaaly, A. Lake, and J. Epps, *Journal of Transportation Engineering*, 2002: 578–583), the following values of fracture stress (in megapascals) were measured for a sample of 24 mixtures of hot-mixed asphalt (HMA).

30	75	79	80	80	105	126	138	149	179	179	191
223	232	232	236	240	242	245	247	254	274	384	470

Compute the mean, median, and the 5%, 10%, and 20% trimmed means.

Trimmed Mean

- The **mean** is found by averaging together all 24 numbers, which produces a value of **195.42**.
- The **median** is the **average of the 12th and 13th** numbers, which is (191 + 223)/2 = **207.00**.
- To compute the **5% trimmed mean**, we must drop 5% of the data from each end. This comes to **(0.05)(24) = 1.2** observations. We round 1.2 to 1, and trim one observation off each end. The 5% trimmed mean is the average of the remaining 22 numbers:

Trimmed Mean

- The **mean** is found by averaging together all 24 numbers, which produces a value of **195.42**.
- The **median** is the **average of the 12th and 13th** numbers, which is (191 + 223)/2 = **207.00**.
- To compute the **5% trimmed mean**, we must drop 5% of the data from each end. This comes to **(0.05)(24) = 1.2** observations. We round 1.2 to 1, and trim one observation off each end. The 5% trimmed mean is the average of the remaining 22 numbers:

Trimmed Mean

$$\frac{75 + 79 + \cdots + 274 + 384}{22} = 190.45$$

To compute the 10% trimmed mean, round off $(0.1)(24) = 2.4$ to 2. Drop 2 observations from each end, and then average the remaining 20:

$$\frac{79 + 80 + \cdots + 254 + 274}{20} = 186.55$$

To compute the 20% trimmed mean, round off $(0.2)(24) = 4.8$ to 5. Drop 5 observations from each end, and then average the remaining 14:

$$\frac{105 + 126 + \cdots + 242 + 245}{14} = 194.07$$