1.3 – Descriptive Data Story of a Data Set

You know how to apply modeling for descriptive statistics, but now we need to build the toolbox with respect to the statistics and visualizations we can use to model data.

1 Population vs. Sample

Before we understand how to calculate the statistics or visualize the data we first need to understand the difference between a population and a sample. All of a specific group or thing, is called a *population*[2]. A subset, or smaller group of the population, is called a sample [2]. Often, collecting data on a population is infeasible and so data is only available for a sample of the population. For example, if we were interested in determining the average grade point average (GPA) of all undergraduate mechanical engineering students in the United States for the last academic year, the population is every mechanical engineering student in the US. Collecting data on every single student is time-consuming and impractical. However, we can collect a sample of the data, say the GPA of all of the mechanical engineering students at just a few universities, that we can analyze and draw conclusions from. Unless you know you have the data for an entire population, you are going to analyze the data as a sample.

2 Graphical Methods

Graphical methods of descriptive statistics, or data visualization, aim to improve the communication of numerical information by graphing it [3]. One such method is the histogram, which helps us understand how data is distributed. A histogram plots a single variable by grouping its values into equally spaced intervals, called bins, and displaying the frequency, or number of observations, within each bin [3]. When analyzing a histogram, we are looking for patterns or trends that describe the data's shape, center, spread, and any unusual features.

Figure 1 shows a histogram of the deadlift data used in the previous reading. From this histogram we can make several observations:

- Shape: The data appears fairly balanced overall, with a peak near the middle deadlift values. However, the lower end of the graph (lighter deadlift weights) extends slightly farther out and has a few smaller bars, suggesting there are more low-end outliers than on the higher end.
- Outliers: Similar to the shape observation, there seem to be a few exceptionally strong cadets that have maximum deadlifts in the 380-400 range.
- Center: Most cadets seem to have maximum deadlifts between 220 and 280 lbs, with the highest number of observations between 240 and 260 lbs. This suggests the *mode* is somewhere between 240 and 260 lbs. We will discuss two other measures of center shortly.
- Spread: The values range from 100 lbs to 400 lbs, a difference of 300 lbs. This shows high variability in cadet strength.

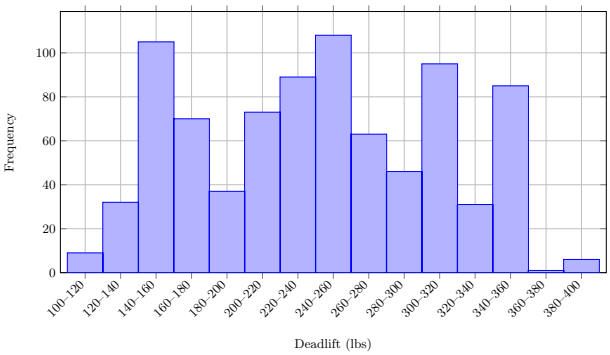

Figure 1: Histogram of Cadet Maximum Deadlift in Pounds

Definition 1.3.1 (Outlier)
An *outlier* is data that falls noticeably outside the main cluster of the data. [1]

3 Numerical Methods

A histogram can provide a lot of information and help us visualize the shape, center, spread, and unique attributes of a data set, but it doesn't provide numerical information that we can use to make inferences [4]. We need to calculate and analyze numerical summary measures that serve to characterize the data and provide more insight than the histogram [2]. This section explains how to do this.

3.1 Measures of Center

The most common way to measure the center of a dataset is using the *mean*.

Definition 1.3.2 (Mean)
The *mean* of a sample of n observations $(x_1, x_2, ..., x_n)$ is given by

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

The corresponding population mean is denoted μ. [4]

The symbol $\bar{x}$, read "x bar," refers to a sample mean. As discussed in Section 1, collecting the data for a population is normally infeasible, so calculating μ, the population mean, is impossible [4]. However, we can use $\bar{x}$ as an estimation for μ. An example calculation for sample mean is given in Problem 1.

1.3 – Descriptive Data Story of a Data Set

Let's explore how the mean represents location. We can use a physical representation of $\bar{x}$ to show how it measures the location of center by thinking of the sample mean as the balance point of a bar with equal weight placed at every observation [2]. Figure 2 shows this for the data found in Problem 1.

Problem 1.3.1: Sample Mean

Given the following 20 observations of the deadlift data, determine the sample mean.

390, 200, 300, 200, 260, 260, 300
340, 220, 260, 300, 260, 150, 260
180, 300, 240, 240, 150, 200

Solution:

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

$$\bar{x} = \frac{1}{20}(x_1 + x_2 + x_3 + ... + x_{18} + x_{19} + x_{20})$$

$$\bar{x} = \frac{1}{20}(390 + 200 + 300 + ... + 240 + 150 + 200)$$

$$\bar{x} = 250.5$$

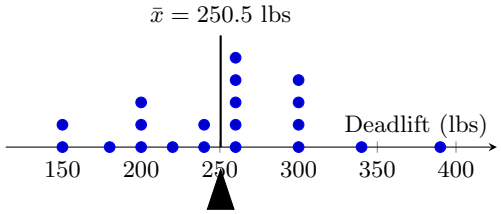

Figure 2: The sample mean as a balance point for deadlift data.

The *median* value of data is another way to measure center. It is the middle value once observations are ordered from smallest to largest, including repeated valuesn Then the median is the single middle value if the number of observations are odd and the average of the two middle values if the number of observations are even.

Definition 1.3.3 (Median)
The *median* of a sample of n observations $(x_1, x_2, ..., x_n)$ is given by

$$\tilde{x} = \begin{cases} \left(\frac{n+1}{2}\right)^{\text{th}} \text{ ordered value} & , n \text{ is odd} \\ \text{average of the } \left(\frac{n}{2}\right)^{\text{th}} \text{ and } \left(\frac{n}{2}+1\right)^{\text{th}} \text{ ordered values} & , n \text{ is even} \end{cases}$$

The corresponding population median is denoted $\tilde{\mu}$. [2]

The symbol $\tilde{x}$, read "x tilde," represents the sample median. Just as we can with mean, we can use the sample median to estimate the population median. An example of how to calculate sample median is in Problem 2.

Problem 1.3.2: Sample Median

Given the following 20 observations of the deadlift data, determine the sample median.

390, 200, 300, 200, 260, 260, 300
340, 220, 260, 300, 260, 150, 260
180, 300, 240, 240, 150, 200

Solution:

Reorder values in order from lowest to highest.

150, 150, 180, 200, 200, 200, 220
240, 240, 260, 260, 260, 260, 260
300, 300, 300, 300, 340, 390

$n = 20$ is even, therefore,

$(\frac{20}{2})^{th}$ term averaged with $(\frac{20}{2}+1)^{th}$ term.
10^{th} term $= 260$
11^{th} term $= 260$

$$\tilde{x} = \frac{260 + 260}{2}$$

$$\tilde{x} = 260$$

If we compare the results from Problem 1 to the results from Problem 2, we can see that there is a 9.5lb difference from the mean to the median with $\bar{x} < \tilde{x}$. This is because the *median is insensitive to outliers* [2]. This means there are more cadets that lift less weight within our data set which pulls the mean value down. This supports our initial analysis in Section 2 where we noticed the left side seemed to stretch farther.

Rule of Thumb: Choosing Mean or Median

- Use the **mean** when the data is *symmetrical* and has no extreme values or you want a value that reflects the total magnitude of the mathematical average. The mean represents the "balance point" of the data and works well when every data point contributes evenly.
- Use the **median** when the data is *not symmetrical* or contains *outliers*. The median gives the middle value, and it is not affected by extreme highs or lows. It is a typical value rather than a mathematically precise average.

Let's think about this in a slightly different way. Imagine there are 20 cadets deadlifting in a room and recording their max weight. Then one Olympic lifter walks in with a 700 lb deadlift. The average max weight of the room would increase, even though nothing about the "typical" cadet has changed. The median would stay the same. This is why the median can be a more useful measure of center.

1.3 – Descriptive Data Story of a Data Set

3.2 Measure of Spread

The measure of center may not tell the full story of the data set. Different samples may have similar measures of center, but may still be very different.

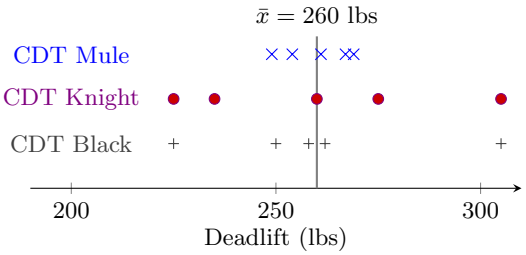

Figure 3: A plot of CDTs Mule, Knight, and Black's five deadlift attempts.

For example, CDTs Mule, Knight, and Black record their maximum deadlift weight over the course of five days. From Figure 3 we can see that CDT Mule is pretty consistent over the course of the five days, with lifts closely clustered around the mean of 260 lbs. CDTs Knight and Black, however, though they have the same mean as CDT Mule, were less consistent. The difference between these three data sets are best described using a measure of spread versus a measure of center.

The simplest measure of spread is the *range*, which is the difference between the largest and smallest sample size [2]. CDT Mules's range is less than that of CDT Knight's and CDT Black's in Figure 3. However, CDT Knight and CDT Black have the same range, but there is still a difference with how variable the data is.

The primary method of measuring variation is how far data deviates, or differs, from the mean [2]. A positive deviation means an observation is above the mean, while a negative deviation means the observation is below the mean. You may think we should be able to add these deviations together to determine valuable insight into a dataset's variation. However, the positive and negative deviations cancel each other out when added together, so simply adding the deviations does not provide a helpful measure [2]. Instead we calculate the *standard deviation* which is the square root of the average squared deviation from the mean. This gives a measure of spread in the same units as the data. Problem 3 provides an example for the data in Figure 3.

Definition 1.3.4 (Standard Deviation)
The *standard deviation* of a sample of n observations $(x_1, x_2, ..., x_n)$ is given by

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

The corresponding population standard deviation is denoted σ. [2]

Problem 1.3.3: Sample Standard Deviation

Given the data for CDTs Mule, Knight, and Black and that the mean for each data set is $\bar{x} = 260$lbs, calculate the sample standard deviation for each cadet.

CDT Mule:	`249, 254, 261, 269, 267`
CDT Knight:	`225, 235, 260, 275, 305`
CDT Black:	`225, 250, 258, 262, 305`

Solution:

$$n = 5, \qquad \bar{x} = 260$$

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

CDT Mule:

$$s_m = \sqrt{\frac{(249-260)^2 + ... + (267-260)^2}{5-1}}$$

$$s_m = \sqrt{\frac{121 + 36 + 1 + 81 + 49}{4}}$$

$$s_m = \sqrt{\frac{288}{4}}$$

$$s_m = 8.49$$

CDT Knight:

$$s_k = \sqrt{\frac{(225-260)^2 + ... + (305-260)^2}{5-1}}$$

$$s_k = \sqrt{\frac{1225 + 625 + 0 + 225 + 2025}{4}}$$

$$s_k = \sqrt{\frac{4100}{4}}$$

$$s_k = 32.02$$

CDT Black:

$$s_c = \sqrt{\frac{(225-260)^2 + ... + (305-260)^2}{5-1}}$$

$$s_c = \sqrt{\frac{1225 + 100 + 4 + 4 + 2025}{4}}$$

$$s_c = \sqrt{\frac{3358}{4}}$$

$$s_c = 28.97$$

Interpreting the standard deviations from Problem 3 shows that CDT Mule's standard deviation from the mean is 8.49 lb, CDT Knight's standard deviation is 32.02 lbs from the mean, and CDT Black has a 28.97 lb standard deviation. From this we can see that CDT Black is more consistent than CDT Knight as the standard deviation is lower. So even though the range is the same, the standard deviation allows us to see the difference.

References

[1] Richard D. De Veaux, Paul F. Velleman, and David E. Bock. *Intro Stats*. Pearson, 2009.

[2] Jay Devore. *Probability & Statistics for Engineering and the Sciences*. Brooks/Cole, 2012.

[3] Frederick Hillier et al. *MA103 Mathematical Modeling: Introduction to Management, Science, & Business Analytics with Connect*. McGraw-Hill, 2024.

[4] Dennis Wackerly, William III Mendenhall, and Richard Scheaffer. *Mathematical Statistics with Applications*. Brooks/Cole, 2008.