Descriptive Statistics 2

Measures of Central Tendency and Variability

Dr. Gordon Wright

Descriptive statistics

Introduction to Descriptive Statistics

Descriptive statistics are fundamental tools in data analysis, providing a way to summarize and describe data in a meaningful and concise manner. When confronted with a new dataset, one of the first tasks is to find ways of condensing the information into easily understood summaries. This is the essence of descriptive statistics, as opposed to inferential statistics which aims to draw conclusions about a population based on sample data.

In this document, we'll explore various measures of central tendency and variability, discuss their calculations and interpretations, and consider when each measure is most appropriate.

Measures of Central Tendency

Measures of central tendency aim to identify the "center" or "typical" value of a dataset. The three primary measures are the mean, median, and mode.

Mean

The mean, often referred to as the average, is the sum of all values divided by the number of values.

As a psychology student, understanding the mean is crucial for interpreting research findings and conducting your own studies. In psychology, we often deal with measures like test scores, reaction times, or survey responses. The mean helps us summarize these data points into a single, representative value. For instance, when studying the effectiveness of a new therapy, you might compare the mean depression scores of a treatment group to those of a control

group. Or in cognitive psychology, you might examine the mean reaction times in different experimental conditions to understand how certain factors affect cognitive processing.

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

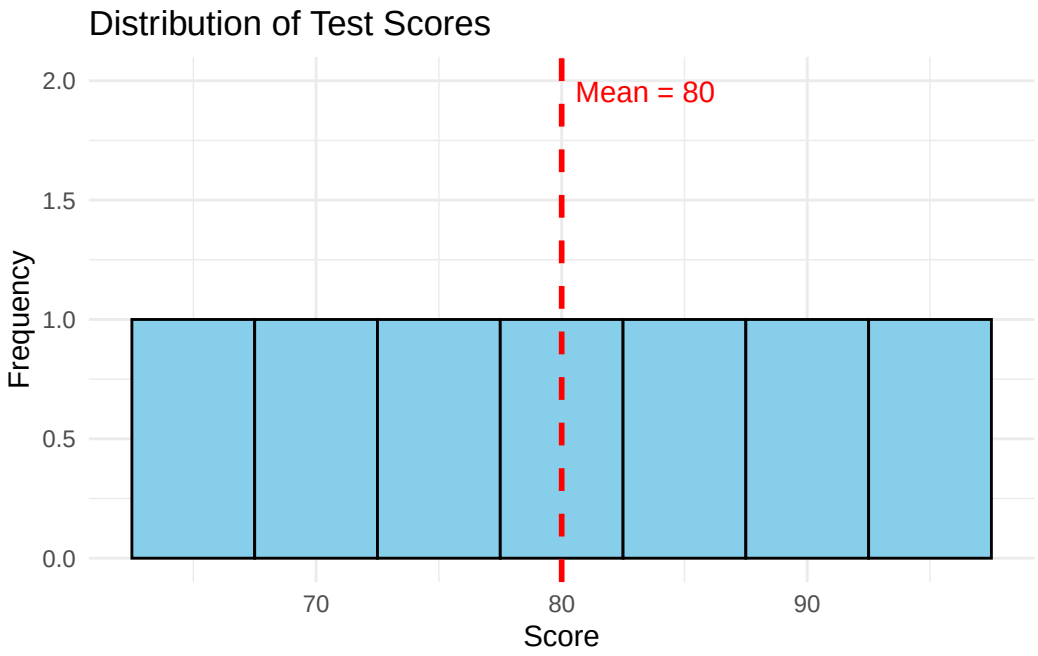

Figure 1: Distribution of test scores with mean highlighted

```
Formula for mean: X = ( X) / n

Where X is the mean,  X is the sum of all scores, and n is the number of scores.

Example calculation:

Test scores: 65, 70, 75, 80, 85, 90, 95

Mean score: 80
```

Formula

For a dataset X with n observations, the mean ($\bar{X}$) is calculated as:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Interpretation

The mean represents the balance point of the data. It's useful for interval and ratio data and is sensitive to all values in the dataset.

Considerations

In Psychology, it's crucial to understand the limitations of the mean. While it's a powerful and commonly used measure, the mean can be heavily influenced by extreme values or outliers, which may not always be desirable. This sensitivity to extreme values can sometimes lead to a misrepresentation of the typical score in your data.

For example, in clinical psychology research, you might encounter a patient with an unusually severe symptom score. This single extreme value could significantly skew the mean, potentially leading to incorrect conclusions about the typical severity of symptoms in your sample. Similarly, in cognitive psychology experiments, a participant with an exceptionally long reaction time (perhaps due to distraction) could distort the mean reaction time for a condition.

In such cases, you might need to consider using other measures of central tendency, like the median, or use statistical techniques to handle outliers. Always critically evaluate whether the mean is the most appropriate measure for your specific research question and dataset.

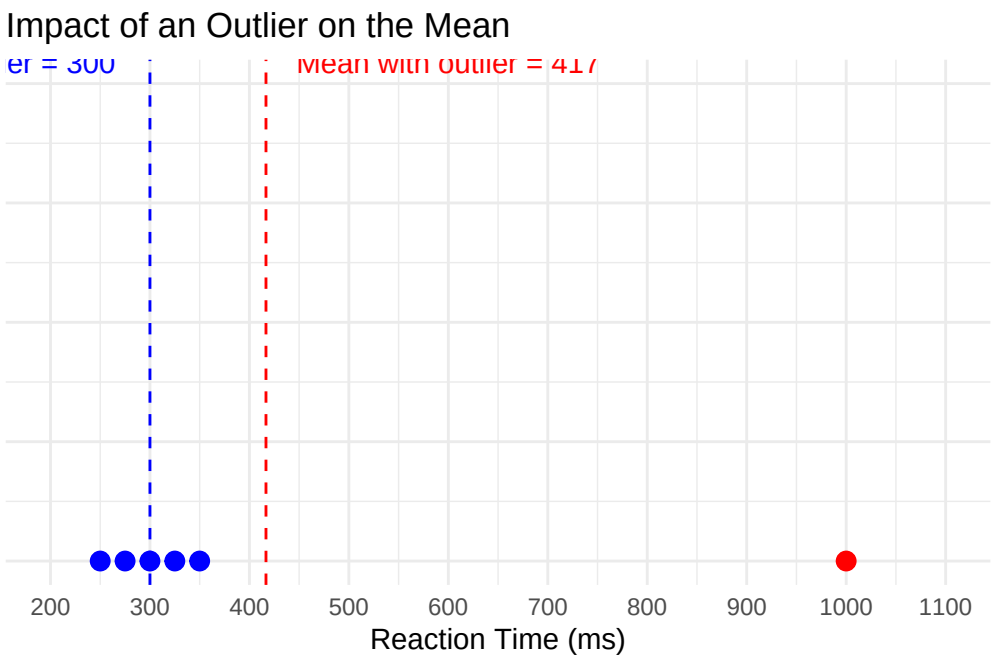

Figure 2: Impact of an outlier on the mean

```
Mean without outlier: 300 ms
```

```
Mean with outlier: 416.6667 ms
```

```
Percentage increase in mean due to outlier: 38.9 %
```

Median

The median is the middle value when the data is ordered from lowest to highest. As a psychology student, understanding the median is crucial because it provides a measure of central tendency that is less affected by extreme values or outliers compared to the mean.

In psychological research, you'll often encounter situations where the median is more appropriate than the mean. For example:

1. When studying income levels in relation to mental health, income distributions are often skewed (many people with lower incomes, fewer with very high incomes). The median income gives a better representation of the "typical" income in this case.

2. In clinical psychology, when measuring the severity of symptoms, you might have a few patients with extremely severe symptoms. The median symptom score would be less influenced by these extreme cases than the mean.

3. In developmental psychology, when looking at developmental milestones (e.g., age at first word), some children might be very early or very late. The median age would give a more robust measure of the typical age for reaching the milestone.

4. In psychophysics experiments, reaction times often have a skewed distribution with some very long responses. The median reaction time is often used as it's less sensitive to these occasional long responses.

Remember, the median divides the data into two equal halves: 50% of the data points are below the median, and 50% are above. This property makes the median particularly useful for describing the "typical" value in skewed distributions.

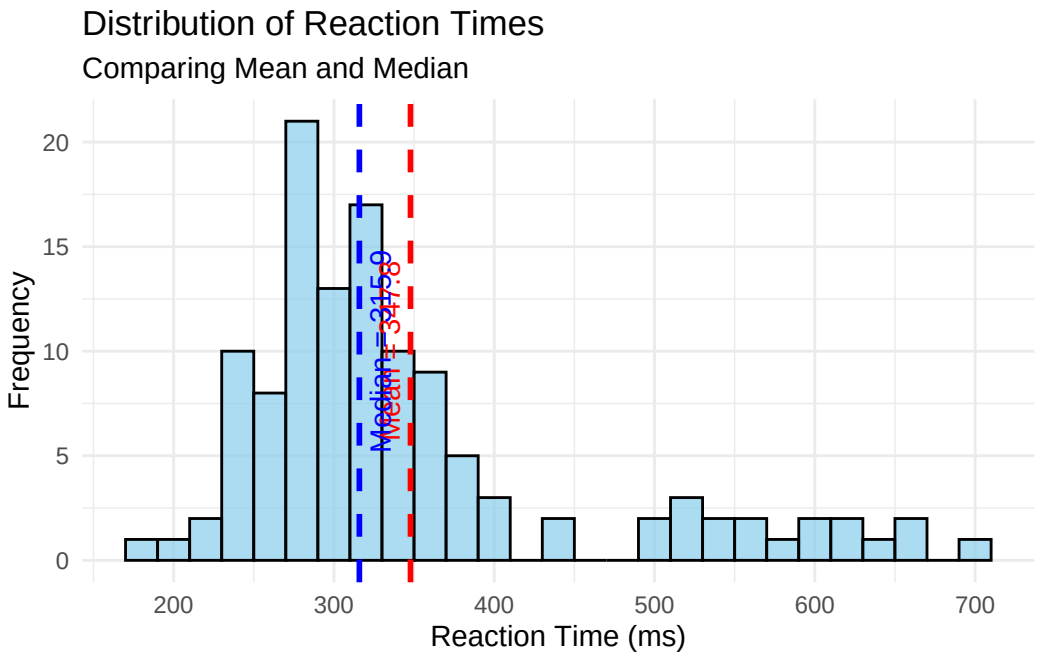

Figure 3: Comparison of Mean and Median in a Skewed Distribution

```
Mean reaction time: 347.8 ms
```

```
Median reaction time: 315.9 ms
```

```
Difference (Mean - Median): 31.9 ms
```

```
The mean is higher than the median, indicating a right-skewed distribution.
This suggests some unusually long reaction times are pulling the mean up.
```

Calculation

1. Order the data from lowest to highest.
2. If n is odd, the median is the middle value.
3. If n is even, the median is the average of the two middle values.

Interpretation

The median represents the 50th percentile of the data. It's less sensitive to extreme values compared to the mean.

Considerations

The median is particularly useful for skewed distributions or when dealing with ordinal data. As a psychology student, you'll encounter many situations where the median is more appropriate than the mean. However, it's crucial to understand both the strengths and limitations of the median, especially when it comes to the practice of median splits.

Strengths of the median: 1. Robustness to outliers: Unlike the mean, the median is not heavily influenced by extreme values, making it useful for skewed data often found in psychological research (e.g., reaction times, income levels). 2. Appropriate for ordinal data: For Likert scales or ranked data, the median can be more meaningful than the mean. 3. Easy interpretation: The median represents the middle value, with 50% of observations above and 50% below.

However, the use of median splits on continuous data is a controversial practice in psychology that you should approach with caution:

Median Split: This is the practice of dividing a continuous variable into two groups based on the median value. For example, splitting participants into "high anxiety" and "low anxiety" groups based on whether their anxiety scores are above or below the median.

Problems with median splits: 1. Loss of information: Converting a continuous variable to a dichotomous one discards potentially valuable information about individual differences. 2. Reduced statistical power: This can make it harder to detect real effects in your data. 3. Increased risk of Type I and Type II errors: You might find spurious relationships or miss true relationships. 4. Difficulty in replication: The median in one sample might not be the same in another, making cross-study comparisons challenging. 5. Arbitrary grouping: Participants just above and below the median might be more similar to each other than to others in their respective groups. 6. Assumption of linearity: Median splits assume a linear relationship between variables, which may not always be true.

Instead of median splits, consider these alternatives: - Use the full continuous variable in your analyses (e.g., regression techniques). - If categorization is necessary, use established cut-off points based on theory or previous research. - Consider more sophisticated statistical techniques that can handle non-linear relationships.

```
`geom_smooth()` using formula = 'y ~ x'
```

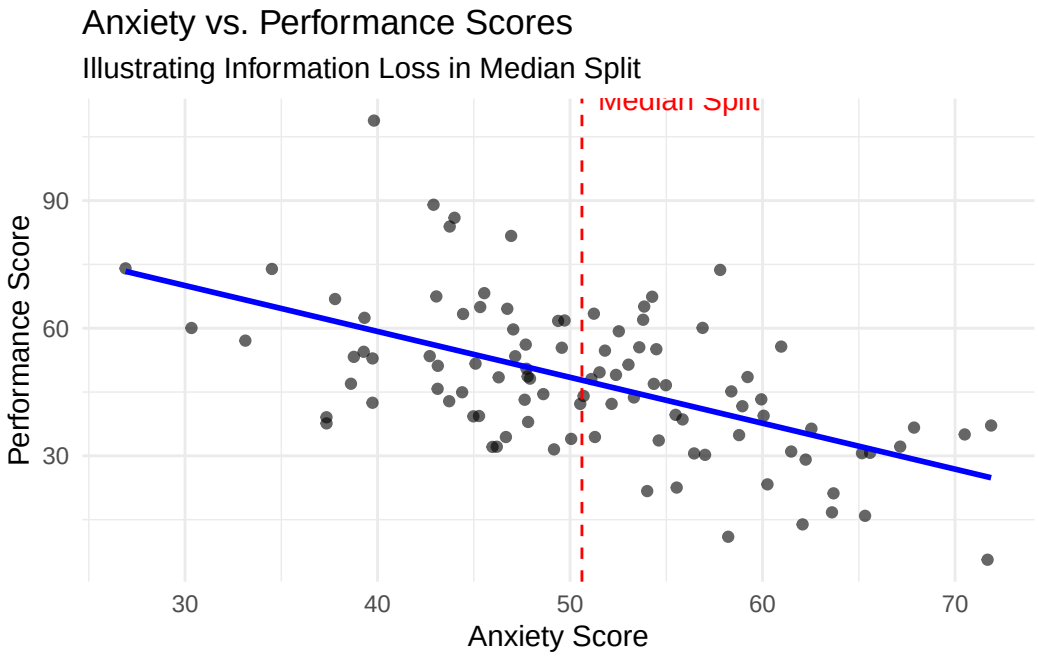

Figure 4: Illustration of Information Loss in Median Split

```
Correlation using full continuous data: -0.562
```

```
Correlation after median split: -0.424
```

```
Information loss: 43.1 %
```

Mode

The mode is the most frequently occurring value in a dataset. As a psychology student, understanding the mode is crucial, especially when dealing with categorical data or when you want to identify the most common response or behavior in a study.

Calculation

To find the mode: 1. List all unique values in the dataset. 2. Count the frequency of each value. 3. Identify the value(s) with the highest frequency.

Unlike the mean and median, the mode doesn't require numerical data, making it versatile for various types of psychological data.

Interpretation

The mode represents the most common value(s) in the dataset. Datasets can be: - Unimodal: One mode (most common in many psychological measures) - Bimodal: Two modes (might indicate two distinct groups in your sample) - Multimodal: More than two modes (could suggest complex patterns in your data)

In psychology, the mode is particularly useful for: 1. Analyzing categorical data (e.g., most common diagnosis in a clinical sample) 2. Understanding preferences (e.g., most popular response in a survey) 3. Identifying typical behaviors (e.g., most frequent reaction in a social psychology experiment)

Considerations

1. Nominal data: The mode is the only measure of central tendency that's meaningful for nominal data (e.g., categories with no intrinsic order).
2. Discrete data: For discrete numerical data (e.g., Likert scales), the mode can provide insights into the most common response.
3. Continuous data: The mode is less useful for continuous data, as exact repetitions are less likely. However, you can group continuous data into intervals and find the modal interval.
4. Multiple modes: When data is bimodal or multimodal, it often suggests underlying patterns or subgroups in your sample, which may warrant further investigation.
5. Stability: The mode can be unstable in small datasets, where small changes can lead to a different mode.
6. Completeness: Unlike the mean or median, the mode doesn't take into account the entire distribution of the data.

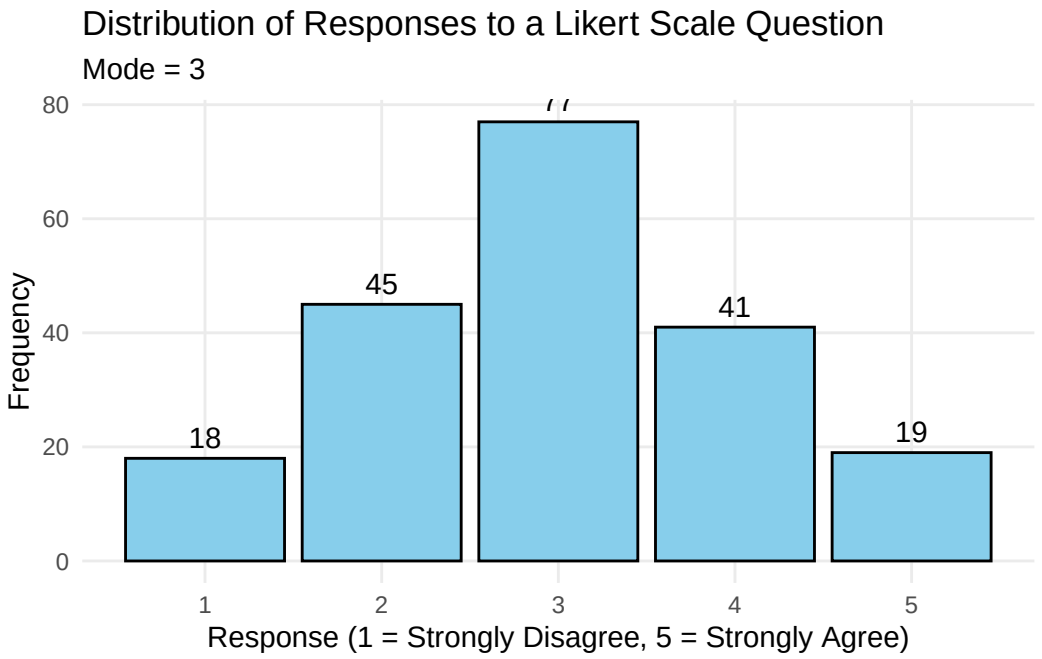

Figure 5: Distribution of Responses in a Psychological Survey

```
Mode: 3

Interpretation: The most common response was 3 on the 5-point scale.

The distribution appears to be unimodal, with a clear single most common response.


Comparison with other measures of central tendency:

Mean: 2.99

Median: 3


While the mean and median provide the 'average' and 'middle' responses,

the mode specifically identifies the most frequent response, which can be

particularly informative for Likert scale data in psychological research.
```

Measures of Variability

Measures of variability describe how spread out the data are in a dataset. They complement measures of central tendency by providing crucial information about the distribution of values. In psychological research, understanding variability is essential for interpreting individual differences, assessing the reliability of measurements, and making inferences about populations.

Range

The range is the simplest measure of variability, representing the difference between the largest and smallest values in a dataset.

Formula

$$\text{Range} = \max(X) - \min(X)$$

Interpretation

The range provides a quick snapshot of the total spread of the data. However, it's highly sensitive to outliers and doesn't provide information about the distribution of values between the extremes.

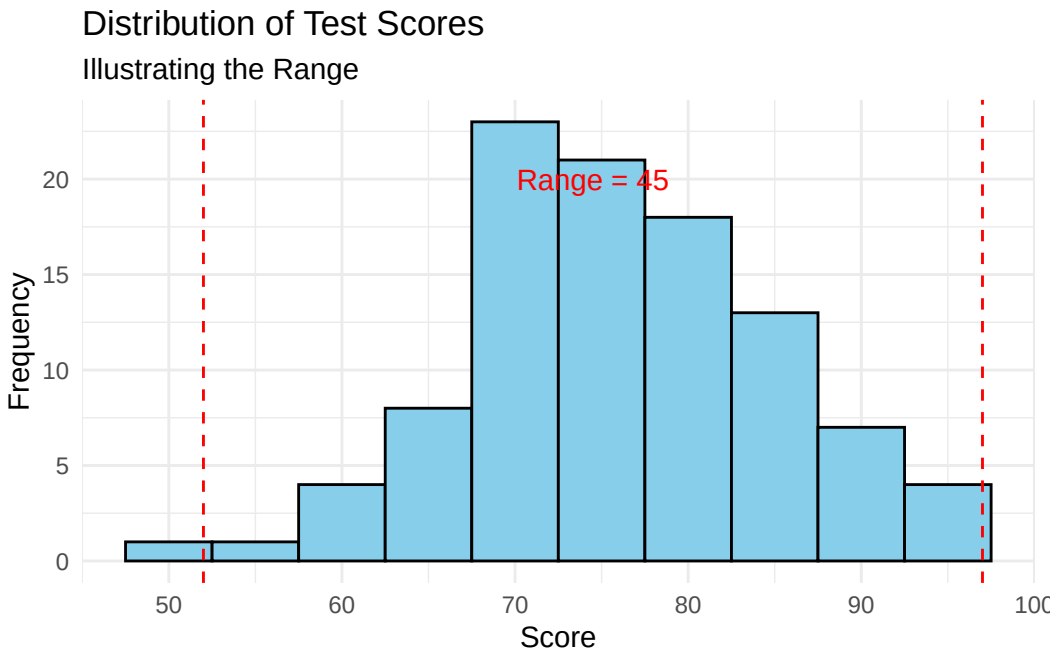

Figure 6: Illustration of Range in Test Scores

```
Minimum score: 52
```

```
Maximum score: 97
```

```
Range: 45
```

Interquartile Range (IQR)

The IQR represents the range of the middle 50% of the data, making it less sensitive to outliers than the full range.

Calculation

1. Calculate the first quartile (Q1, 25th percentile) and third quartile (Q3, 75th percentile).
2. IQR = Q3 - Q1

Interpretation

The IQR provides a measure of spread for the central portion of the data, giving insight into the variability of the "typical" values in a dataset.

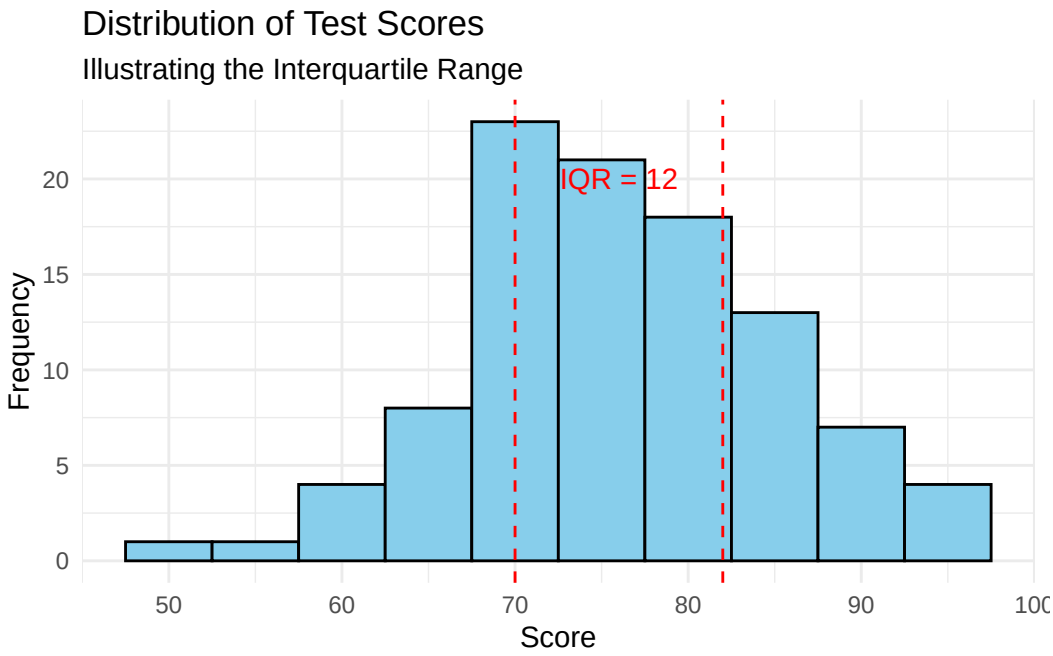

Figure 7: Illustration of Interquartile Range in Test Scores

```
First Quartile (Q1): 70

Third Quartile (Q3): 82

Interquartile Range (IQR): 12
```

Variance

The variance quantifies the average squared deviation from the mean, providing a comprehensive measure of variability that takes all data points into account.

Formula

For a sample:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

For a population:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

Where s^2 is the sample variance, σ^2 is the population variance, $\bar{X}$ is the sample mean, and μ is the population mean.

Interpretation

The variance provides a measure of the overall spread of the data. However, it's expressed in squared units, which can make it difficult to interpret directly.

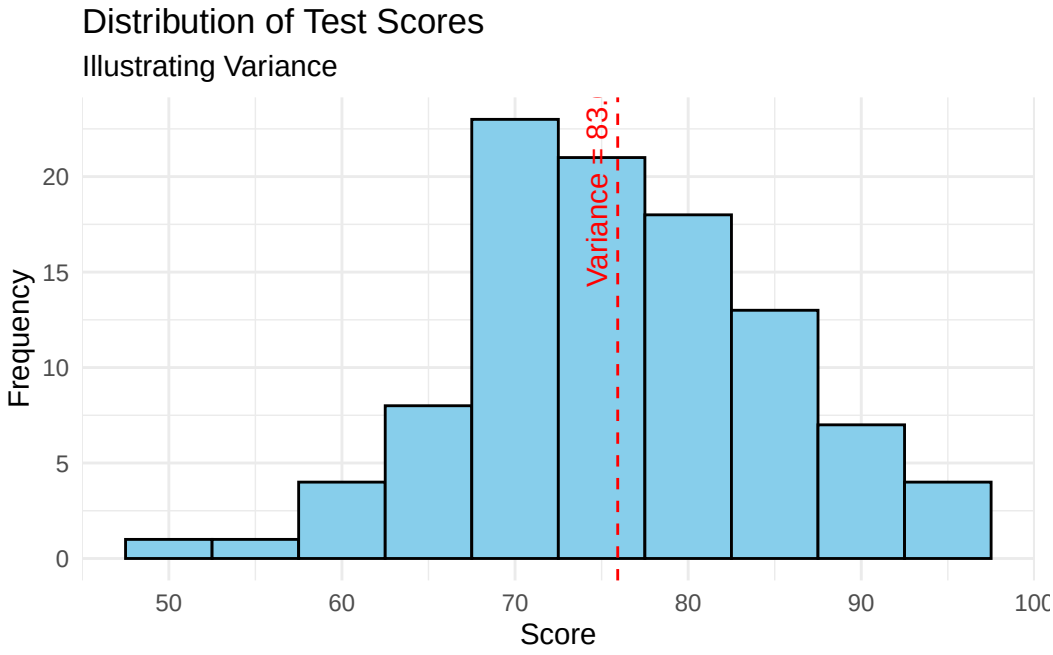

Figure 8: Illustration of Variance in Test Scores

```
Mean score: 75.93
```

```
Variance: 83.62
```

Standard Deviation

The standard deviation is the square root of the variance, providing a measure of spread in the original units of the data.

Formula

For a sample:

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

For a population:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$$

Interpretation

The standard deviation is more interpretable than the variance as it's expressed in the same units as the original data. In a normal distribution:

- Approximately 68% of the data falls within one standard deviation of the mean.
- Approximately 95% falls within two standard deviations.
- Approximately 99.7% falls within three standard deviations.

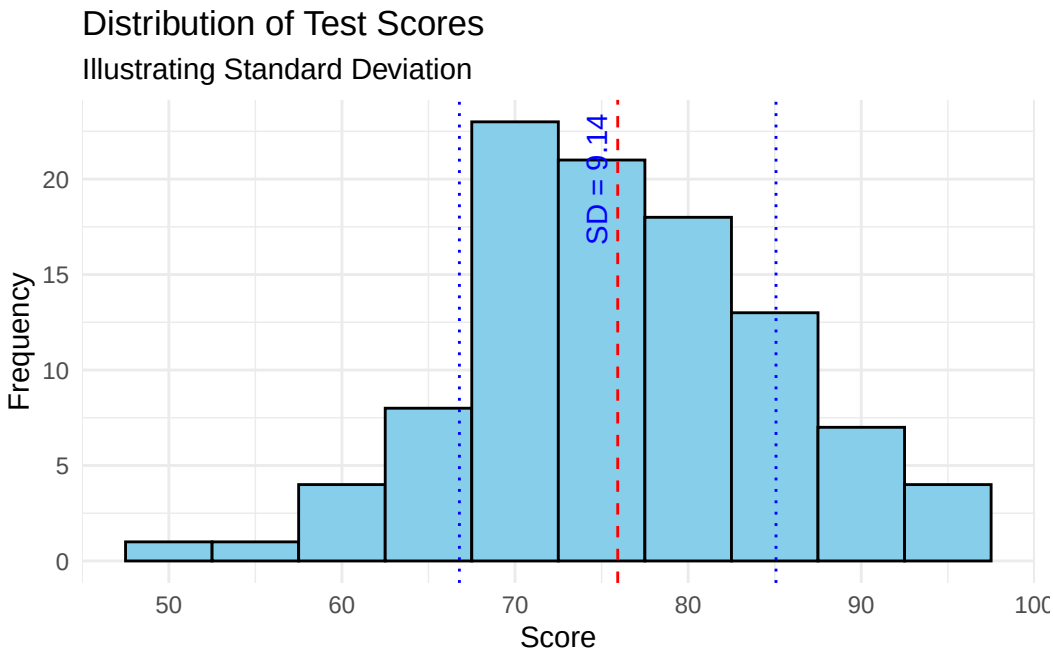

Figure 9: Illustration of Standard Deviation in Test Scores

```
Mean score: 75.93

Standard Deviation: 9.14

68% of scores fall between: 66.79 and 85.07
```

Mean Absolute Deviation

The Mean Absolute Deviation (MAD) is a robust measure of variability that calculates the average of the absolute differences between each value and the mean. It's particularly useful in psychological research when dealing with data that may contain outliers or when the distribution is non-normal.

Formula

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^{n} |X_i - \bar{X}|$$

Where: - n is the number of observations - X_i is the ith value in the dataset - $\bar{X}$ is the mean of the dataset

Interpretation

Like the standard deviation, the MAD provides a measure of variability in the original units of the data. It's less sensitive to outliers than the standard deviation, making it a more robust measure of dispersion. In psychological research, this can be particularly useful when:

1. Dealing with small sample sizes where outliers might have a disproportionate effect
2. Analyzing data from clinical populations where extreme scores are common but meaningful
3. Exploring reaction time data, which often includes occasional very long responses

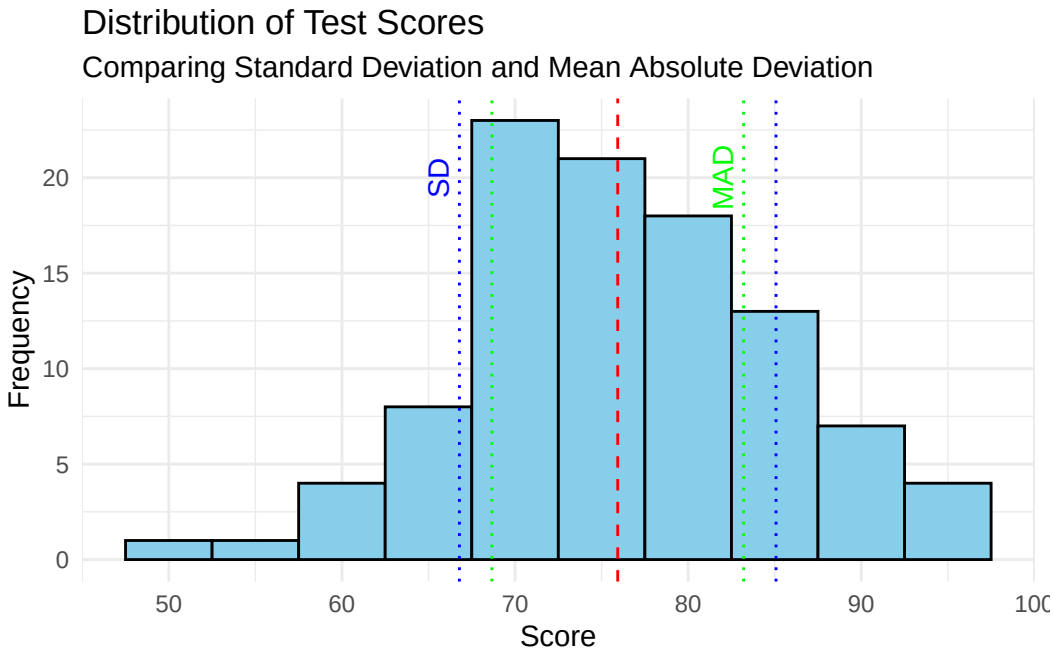

Figure 10: Comparison of Standard Deviation and Mean Absolute Deviation

```
Standard Deviation: 9.14
```

```
Mean Absolute Deviation: 7.27
```

Choosing Appropriate Descriptive Statistics

The choice of which descriptive statistics to use depends on several factors:

1. **Data Type**:

- Nominal: Mode
- Ordinal: Median, Mode
- Interval/Ratio: Mean, Median, Mode, all variability measures

2. **Distribution Shape**:
   - Symmetric: Mean, Standard Deviation
   - Skewed: Median, IQR
3. **Presence of Outliers**:
   - With outliers: Median, IQR
   - Without outliers: Mean, Standard Deviation
4. **Sample Size**:
   - Small samples: Be cautious with mean and standard deviation
   - Large samples: All measures become more reliable
5. **Research Question**:
   - The specific goals of your analysis should guide your choice of statistics

Conclusion

Descriptive statistics are essential tools for summarizing and understanding datasets. They provide a foundation for more advanced analyses and help communicate key features of the data to others. However, it's important to remember that these summary measures can sometimes obscure important details in the data. Therefore, they should often be used in conjunction with data visualizations and more detailed analyses when appropriate.

When reporting descriptive statistics, always consider the context of your data and the audience you're communicating to. The goal is not just to calculate numbers, but to tell a meaningful story about your data that others can understand and act upon.