Chapter 3
Central Tendency

Likan Zhan
Beijing Language and Culture University

2021-09-27

https://likan.info
zhanlikan@blcu.edu.cn

Table of Contents

Table of Contents

Central tendency

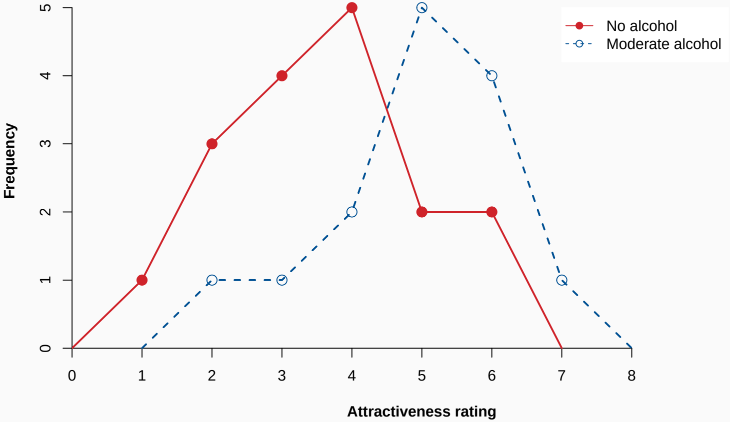
Central tendency

- The general purpose of descriptive statistical methods is to organize and summarize a set of scores.

Central tendency

- The general purpose of descriptive statistical methods is to organize and summarize a set of scores.
- Perhaps the most common method for summarizing and describing a distribution is to find a single value that defines the average score and can serve as a typical example to represent the entire distribution.

Alcohol and attractiveness



Central tendency

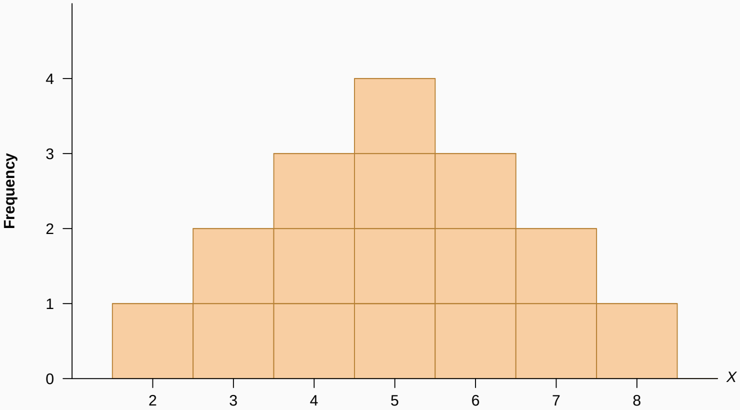
Central tendency

- **Central tendency** (集中趋势) is a statistical measure to determine a single score that defines the center of a distribution.

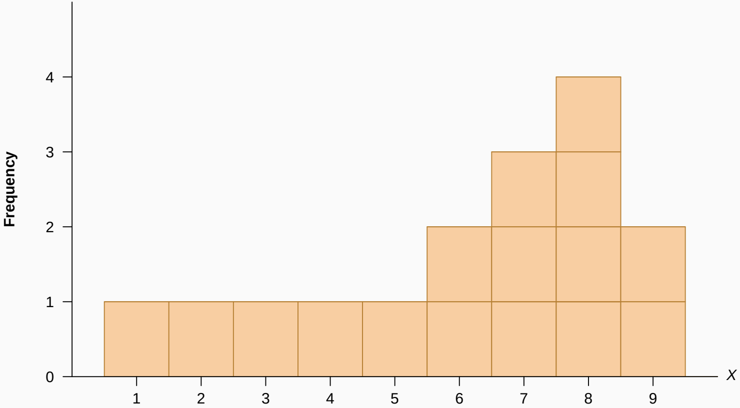
Central tendency

- **Central tendency** (集中趋势) is a statistical measure to determine a single score that defines the center of a distribution.
- The goal of central tendency is to find the single score that is most typical or most representative of the entire group.

To determine the central tendency



To determine the central tendency



To determine the central tendency

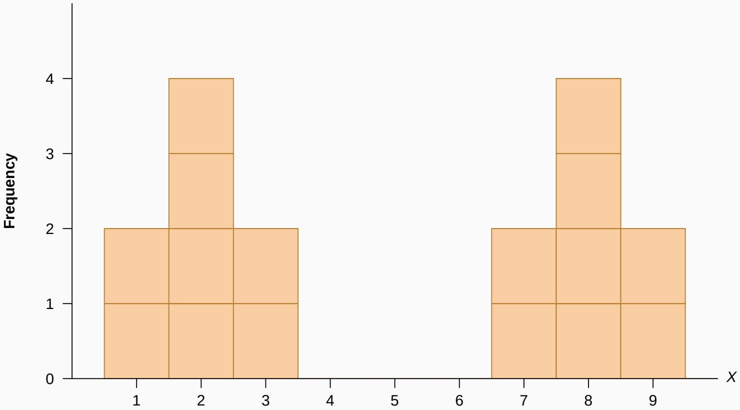


Table of Contents

The Mean

The Mean

- The **mean** (平均值) or **arithmetic average** (算术平均值) for a distribution is the sum of the scores divided by the number of scores.

The Mean

- The **mean** (平均值) or **arithmetic average** (算术平均值) for a distribution is the sum of the scores divided by the number of scores.
- The formula for the *population mean* is

$$\mu = \frac{\sum X}{N}$$

The Mean

- The **mean** (平均值) or **arithmetic average** (算术平均值) for a distribution is the sum of the scores divided by the number of scores.
- The formula for the *population mean* is

$$\mu = \frac{\sum X}{N}$$

- The formula for the *sample mean* is

$$\text{sample mean} = M = \overline{X} = \frac{\sum X}{n}$$

The Mean

The Mean

- Use `sum(X)` and `length(X)` to calculate the mean

```
X <- c(3, 7, 4, 6)
sum(X) / length(X)

## [1] 5
```

The Mean

- Use `sum(X)` and `length(X)` to calculate the mean

```
X <- c(3, 7, 4, 6)
sum(X) / length(X)
## [1] 5
```

- Use `mean(X)` to calculate mean

```
mean(X)
## [1] 5
```

Alternative Definitions for the Mean

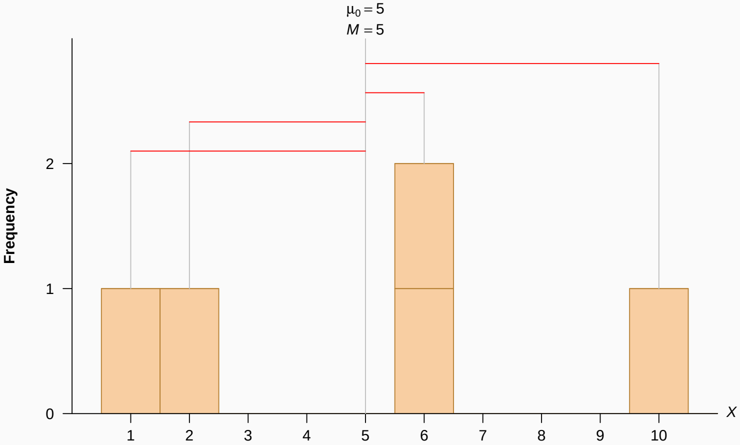
Alternative Definitions for the Mean

- *Dividing the Total Equally*.

Alternative Definitions for the Mean

- *Dividing the Total Equally*.
- *The Mean as a Balance Point* (重心). The distances above the mean are exactly balanced by the distances below the mean.

The Mean as a Balance Point



The Weighted Mean

The Weighted Mean

- To calculate the overall mean of two combined samples, we need two values:

The Weighted Mean

- To calculate the overall mean of two combined samples, we need two values:
- The overall sum of the scores for the combined group ($\sum X$), and The total number of scores in the combined group (n).

The Weighted Mean

- To calculate the overall mean of two combined samples, we need two values:
- The overall sum of the scores for the combined group ($\sum X$), and The total number of scores in the combined group (n).
- With these two values, we can compute the mean using the basic equation:

$$\text{overall mean} = \frac{\sum X \text{ (overall sum for the combined group)}}{n \text{ (total number in the combined groupd)}} = \frac{\sum X_1 + \sum X_2}{n_1 + n_2}$$

The Weighted Mean

The Weighted Mean

- Given the following two samples:

```r
Mean1 <- 6; n1 <- 12
Mean2 <- 7; n2 <- 8
```

The Weighted Mean

- Given the following two samples:

```
Mean1 <- 6; n1 <- 12
Mean2 <- 7; n2 <- 8
```

- The overall mean of the two samples are:

```
(Mean1 * n1 + Mean2 * n2) / (n1 + n2)

## [1] 6.4
```

An Alternative Procedure

An Alternative Procedure

- Logically, when these two samples are combined, the larger sample will make a greater contribution to the combined group than the smaller sample.

An Alternative Procedure

- Logically, when these two samples are combined, the larger sample will make a greater contribution to the combined group than the smaller sample.
- An Alternative Procedure for Finding the Weighted Mean

```
(n1 / (n1 + n2)) * Mean1 +
(n2 / (n1 + n2)) * Mean2

## [1] 6.4
```

From a Frequency Distribution Table

From a Frequency Distribution Table

- Suppose a frequency distribution table,

X	f
10	1
9	2
8	4
7	0
6	1

From a Frequency Distribution Table

- Suppose a frequency distribution table,

X	f
10	1
9	2
8	4
7	0
6	1

- Calculate the mean using the frequency

```
sum(X * f) / sum(f)

## [1] 8.25
```

Characteristics of the Mean

Characteristics of the Mean

- **Changing a Score.**

```
(X <- c(9, 8, 7, 5, 1))

## [1] 9 8 7 5 1

mean(X)

## [1] 6
```

Characteristics of the Mean

- **Changing a Score.**

```r
(X <- c(9, 8, 7, 5, 1))
## [1] 9 8 7 5 1
mean(X)
## [1] 6
```

```r
X[5] <- 8; X
## [1] 9 8 7 5 8
mean(X)
## [1] 7.4
```

Characteristics of the Mean

Characteristics of the Mean

- **Introducing a New Score or Removing a Score.**

```r
n <- 5
Mean <- 7
X <- 13
(n * Mean + X) / (n + 1)
```

```
## [1] 8
```

Characteristics of the Mean

Characteristics of the Mean

- Adding or Subtracting a Constant from Each Score.

```
(X <- c(4, 2, 3, 3, 2 , 3))

## [1] 4 2 3 3 2 3

mean(X)

## [1] 2.833333
```

Characteristics of the Mean

- Adding or Subtracting a Constant from Each Score.

```r
(X <- c(4, 2, 3, 3, 2 , 3))
## [1] 4 2 3 3 2 3
mean(X)
## [1] 2.833333
```

```r
(X1 <- X + 1)
## [1] 5 3 4 4 3 4
mean(X1)
## [1] 3.833333
```

Characteristics of the Mean

Characteristics of the Mean

- **Multiplying or Dividing Each Score by a Constant.**

```
(X <- c(10, 9, 12, 8, 11))

## [1] 10  9 12  8 11

mean(X)

## [1] 10
```

Characteristics of the Mean

- **Multiplying or Dividing Each Score by a Constant.**

```r
(X <- c(10, 9, 12, 8, 11))

## [1] 10  9 12  8 11

mean(X)

## [1] 10
```

```r
(X1 <- X * 2.54)

## [1] 25.40 22.86 30.48 20.32 27.94

mean(X1)

## [1] 25.4
```

Table of Contents

Median

Median

- If the scores in a distribution are listed in order from smallest to largest, the **median** (中数) is the midpoint of the list.

Median

- If the scores in a distribution are listed in order from smallest to largest, the **median** (中数) is the midpoint of the list.
- More specifically, the median is the point on the measurement scale below which 50% of the scores in the distribution are located.

Median

- If the scores in a distribution are listed in order from smallest to largest, the **median** (中数) is the midpoint of the list.
- More specifically, the median is the point on the measurement scale below which 50% of the scores in the distribution are located.
- Defining the median as the midpoint of a distribution means that that the scores are being divided into two equal-sized groups.

Median

Median

- To find the median, list the scores in order from smallest to largest.

Median

- To find the median, list the scores in order from smallest to largest.
- Begin with the smallest score and count the scores as you move up the list.

Median

- To find the median, list the scores in order from smallest to largest.
- Begin with the smallest score and count the scores as you move up the list.
- The median is the first point you reach that is greater than 50% of the scores in the distribution.

Median

Median

- The median can be equal to a score in the list or it can be a point between two scores.

Median

- The median can be equal to a score in the list or it can be a point between two scores.
- Notice that the median is not algebraically defined (there is no equation for computing the median), which means that:

Median

- The median can be equal to a score in the list or it can be a point between two scores.
- Notice that the median is not algebraically defined (there is no equation for computing the median), which means that:
- There is a degree of subjectivity in determining the exact value.

Find the Median for Most Distributions

```
(X <- c(10, 11, 5, 3, 8))

## [1] 10 11  5  3  8
```

Find the Median for Most Distributions

```r
(X <- c(10, 11, 5, 3, 8))

## [1] 10 11  5  3  8
```

```r
(X1 <- sort(X))

## [1]  3  5  8 10 11
```

Find the Median for Most Distributions

```r
(X <- c(10, 11, 5, 3, 8))
```

```
## [1] 10 11  5  3  8
```

```r
(X1 <- sort(X))
```

```
## [1]  3  5  8 10 11
```

```r
(pos <- (length(X1) + 1) / 2)
```

```
## [1] 3
```

Find the Median for Most Distributions

```
X1[pos]

## [1] 8
```

Find the Median for Most Distributions

```
X1[pos]
```

```
## [1] 8
```

```
median(X)
```

```
## [1] 8
```

Find the Median for Most Distributions

```
(X <- c(4, 5, 7, 1, 1, 8))

## [1] 4 5 7 1 1 8

(X1 <- sort(X))

## [1] 1 1 4 5 7 8
```

Find the Median for Most Distributions

```
(X1[length(X) / 2] +
 X1[length(X) / 2 + 1]) / 2

## [1] 4.5
```

Find the Median for Most Distributions

```
(X1[length(X) / 2] +
 X1[length(X) / 2 + 1]) / 2

## [1] 4.5
```

```
median(X)

## [1] 4.5
```

Precise Median for a Continuous Variable

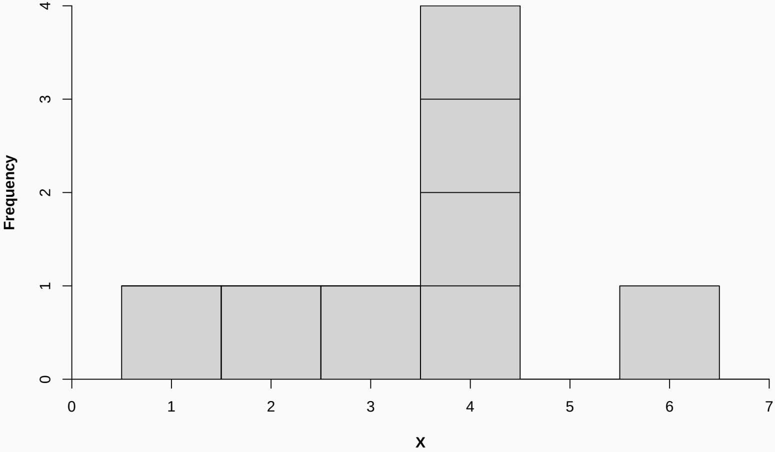
Precise Median for a Continuous Variable

- The simple technique of listing and counting scores is sufficient to determine the median for most distributions and is always appropriate for discrete variables.

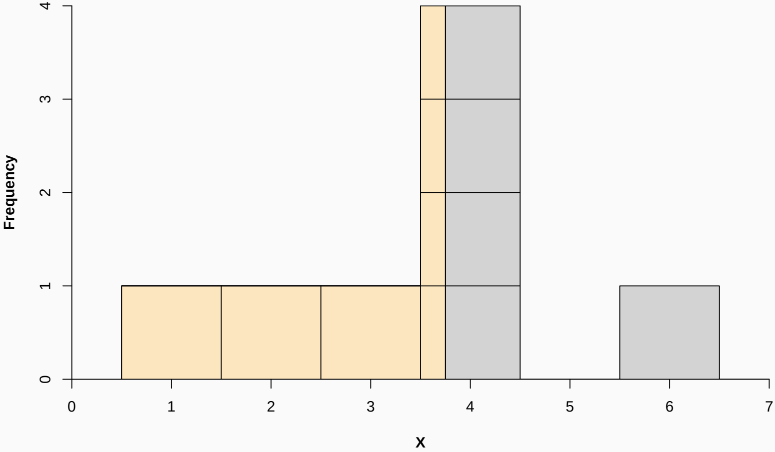
Precise Median for a Continuous Variable

- The simple technique of listing and counting scores is sufficient to determine the median for most distributions and is always appropriate for discrete variables.
- With a continuous variable, however, it is possible to divide a distribution precisely in half so that exactly 50% of the distribution is located below (and above) a specific point.

Precise Median for a Continuous Variable



Precise Median for a Continuous Variable



Using Interpolation to Locate the 50th Percentile

X	f	cf	cf%
6	1	8	100%
5	0	7	87.5%
4	4	7	87.5%
3	1	3	37.5%
2	1	2	25%
1	1	1	12.5%

Using Interpolation to Locate the 50th Percentile

Using Interpolation to Locate the 50th Percentile

- These values are shown in the following table:

	Scores (X)	Percentages	
Top	4.5	87.5%	
	?	50%	$\leftarrow$ Intermediate value
Bottom	3.5	37.5%	

Using Interpolation to Locate the 50th Percentile

- These values are shown in the following table:

	Scores (X)	Percentages	
Top	4.5	87.5%	
	?	50%	← Intermediate value
Bottom	3.5	37.5%	

- It can be calculated as

$$\frac{4.5-?}{4.5-3.5} = \frac{87.5\% - 37.5\%}{50\% - 37.5\%}$$

Using Interpolation to Locate the 50th Percentile

- These values are shown in the following table:

	Scores (X)	Percentages	
Top	4.5	87.5%	
	?	50%	← Intermediate value
Bottom	3.5	37.5%	

- It can be calculated as

$$\frac{4.5-?}{4.5-3.5} = \frac{87.5\% - 37.5\%}{50\% - 37.5\%}$$

- The 50^{th} percentile (The Median) is $X = 3.75$.

The Median function in R

```
X <- c(1, 2, 3, rep(4, 4), 6)
median(X)

## [1] 4
```

The Median, the Mean, and the Middle

The Median, the Mean, and the Middle

- The mean is located in the middle of the distribution if you use the concept of distance to define the "middle".

The Median, the Mean, and the Middle

- The mean is located in the middle of the distribution if you use the concept of distance to define the "middle".
- The median is located in the middle of the distribution, provided that the term "middle" is defined by the number of scores.

The Median, the Mean, and the Middle

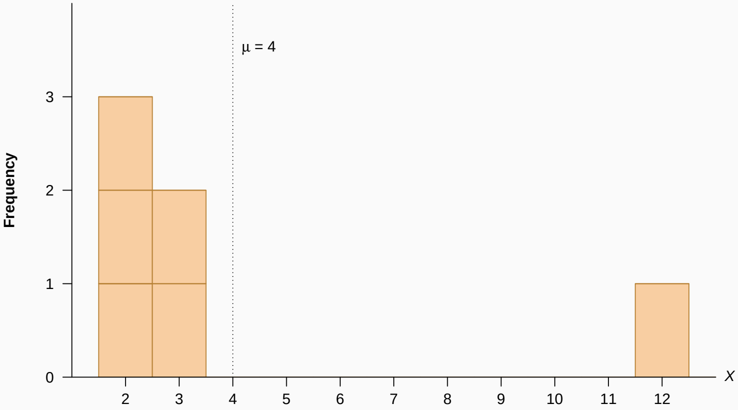


Table of Contents

The Mode

The Mode

- In a frequency distribution, the **mode** (众数) is the score or category that has the greatest frequency.

The Mode

- In a frequency distribution, the **mode** (众数) is the score or category that has the greatest frequency.
- In a frequency distribution graph, the greatest frequency will appear as the tallest part of the figure.

The Mode

- In a frequency distribution, the **mode** (众数) is the score or category that has the greatest frequency.
- In a frequency distribution graph, the greatest frequency will appear as the tallest part of the figure.
- To find the mode, you simply identify the score located directly beneath the highest point in the distribution.

The Mode

- In a frequency distribution, the **mode** (众数) is the score or category that has the greatest frequency.
- In a frequency distribution graph, the greatest frequency will appear as the tallest part of the figure.
- To find the mode, you simply identify the score located directly beneath the highest point in the distribution.
- The mode also can be useful because it is the only measure of central tendency that corresponds to an actual score in the data.

The Mode

The Mode

- A distribution with two modes is said to be bimodal, and a distribution with more than two modes is called multimodal. Occasionally, a distribution with several equally high points is said to have no mode.

The Mode

- A distribution with two modes is said to be bimodal, and a distribution with more than two modes is called multimodal. Occasionally, a distribution with several equally high points is said to have no mode.
- Technically, the mode is the score with the absolute highest frequency. However, the term mode is often used more casually to refer to scores with relatively high frequencies —that is, scores that correspond to peaks in a distribution even though the peaks are not the absolute highest points.

The Mode

- A distribution with two modes is said to be bimodal, and a distribution with more than two modes is called multimodal. Occasionally, a distribution with several equally high points is said to have no mode.
- Technically, the mode is the score with the absolute highest frequency. However, the term mode is often used more casually to refer to scores with relatively high frequencies —that is, scores that correspond to peaks in a distribution even though the peaks are not the absolute highest points.
- When two modes have unequal frequencies, researchers occasionally differentiate the two values by calling the taller peak the major mode, and the shorter one the minor mode.

The Mode

The Mode

- Favorite restaurants named by a sample of $n = 100$ students.

Restaurant	Frequency
穆斯林	5
Hope	16
一心	42
韩国菜	18
南阳菜	7
食堂	7

A frequency distribution

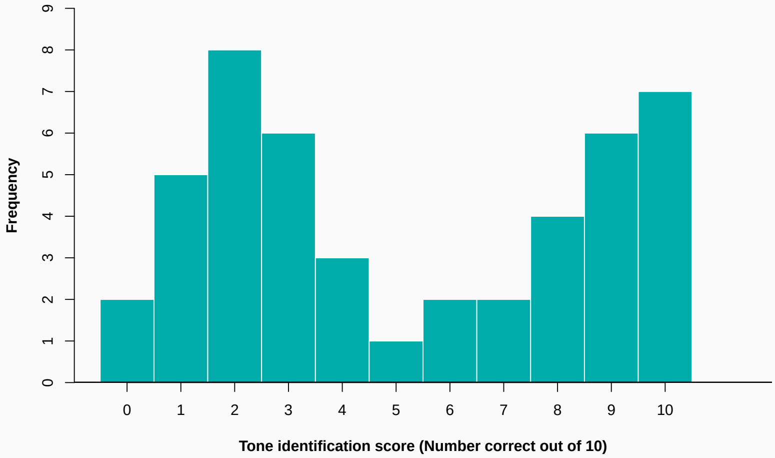
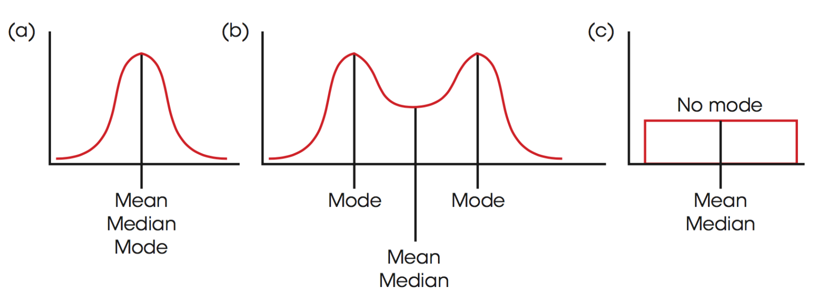


Table of Contents

Symmetrical Distributions



Skewed Distributions

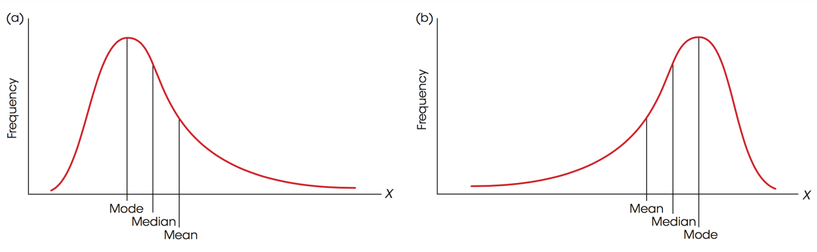


Table of Contents

When to Use the Mean

When to Use the Mean

- The mean is usually the preferred measure of central tendency whenever the scores are numerical values (interval or ratio scale).

When to Use the Mean

- The mean is usually the preferred measure of central tendency whenever the scores are numerical values (interval or ratio scale).
- Because the mean uses every score in the distribution, it typically produces a good representative value.

When to Use the Mean

- The mean is usually the preferred measure of central tendency whenever the scores are numerical values (interval or ratio scale).
- Because the mean uses every score in the distribution, it typically produces a good representative value.
- Besides being a good representative, the mean has the added advantage of being closely related to variance and standard deviation, the most common measures of variability.

When to Use the Median

When to Use the Median

- Extreme Scores or Skewed Distributions: The median is the preferred measure of central tendency when a distribution has a few extreme scores that displace the value of the mean.

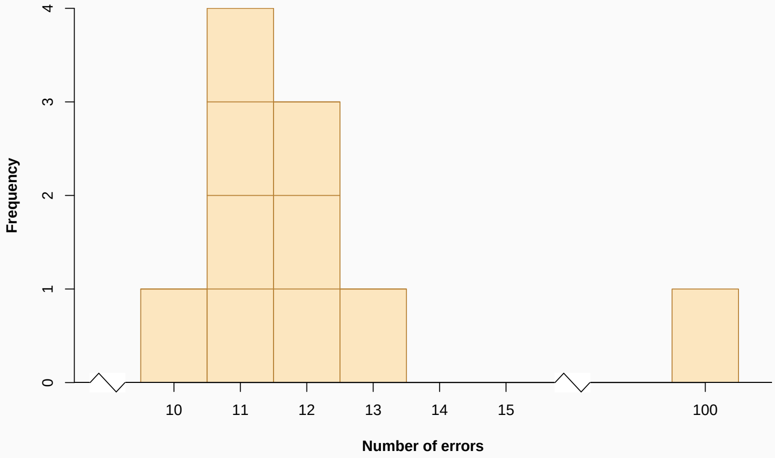
When to Use the Median

- Extreme Scores or Skewed Distributions: The median is the preferred measure of central tendency when a distribution has a few extreme scores that displace the value of the mean.
- Undetermined Values, Open-ended Distributions: The median also is used when there are undetermined (infinite) scores that make it impossible to compute a mean.

When to Use the Median

- Extreme Scores or Skewed Distributions: The median is the preferred measure of central tendency when a distribution has a few extreme scores that displace the value of the mean.
- Undetermined Values, Open-ended Distributions: The median also is used when there are undetermined (infinite) scores that make it impossible to compute a mean.
- Ordinal Scale: Finally, the median is the preferred measure of central tendency for data from an ordinal scale.

Extreme Scores or Skewed Distributions



Undetermined Values

Undetermined Values

- Number of minutes needed to assemble a wooden puzzle

Person	Time(Min.)
1	8
2	11
3	12
4	13
5	17
6	Never finished

Open-ended Distributions

Open-ended Distributions

- A distribution is said to be open-ended when there is no upper limit (or lower limit) for one of the categories.

Number of Pizzas	Frequency
5 or more	3
4	2
3	2
2	3
1	6
0	4

When to Use the Mode

When to Use the Mode

- Nominal Scales: The mode is the only option for describing central tendency for nominal data.

When to Use the Mode

- Nominal Scales: The mode is the only option for describing central tendency for nominal data.
- Discrete Variables: In many situations, especially with discrete variables, people are more comfortable using the realistic, whole-number values produced by the mode.

When to Use the Mode

- Nominal Scales: The mode is the only option for describing central tendency for nominal data.
- Discrete Variables: In many situations, especially with discrete variables, people are more comfortable using the realistic, whole-number values produced by the mode.
- Describing Shape: The value of the mode (or modes) gives an indication of the shape of the distribution as well as a measure of central tendency.

Table of Contents

Reporting Measures of Central Tendency

Reporting Measures of Central Tendency

- The APA style uses the letter *M* as the symbol for the sample mean. Thus, a study might state:

Reporting Measures of Central Tendency

- The APA style uses the letter M as the symbol for the sample mean. Thus, a study might state:
- The treatment group showed fewer errors ($M = 2.56$) on the task than the control group ($M = 11.76$).

Reporting Measures of Central Tendency

- The APA style uses the letter M as the symbol for the sample mean. Thus, a study might state:
- The treatment group showed fewer errors ($M = 2.56$) on the task than the control group ($M = 11.76$).
- The median can be reported using the abbreviation *Mdn*, as in $Mdn = 8.5$ errors, or it can simply be reported in narrative text, as follows:

Reporting Measures of Central Tendency

- The APA style uses the letter M as the symbol for the sample mean. Thus, a study might state:
- The treatment group showed fewer errors ($M = 2.56$) on the task than the control group ($M = 11.76$).
- The median can be reported using the abbreviation *Mdn*, as in $Mdn = 8.5$ errors, or it can simply be reported in narrative text, as follows:
- The median number of errors for the treatment group was 8.5, compared to a median of 13 for the control group.

Presenting Means and Medians in Graphs

Presenting Means and Medians in Graphs

- Graphs also can be used to report and compare measures of central tendency.

Presenting Means and Medians in Graphs

- Graphs also can be used to report and compare measures of central tendency.
- The value of a graph is that it allows several means (or medians) to be shown simultaneously so it is possible to make quick comparisons between groups or treatment conditions.

Rules to construct a graph

Rules to construct a graph

- The height of a graph should be approximately two-thirds to three-quarters of its length.

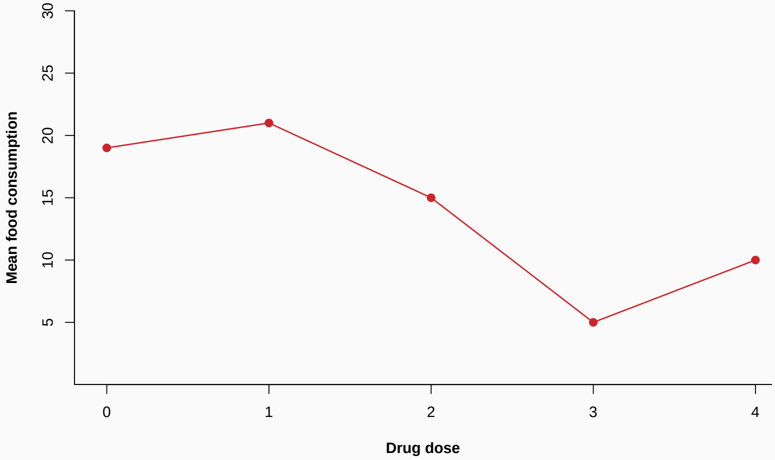
Rules to construct a graph

- The height of a graph should be approximately two-thirds to three-quarters of its length.
- Normally, you start numbering both the X-axis and the Y-axis with zero at the point where the two axes intersect.

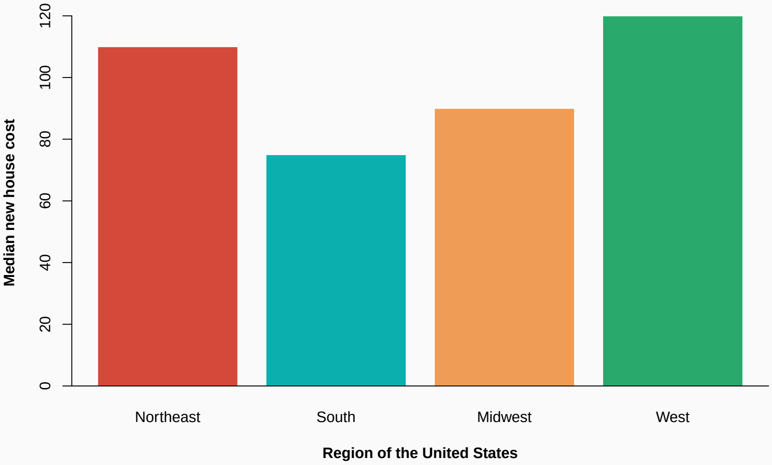
Rules to construct a graph

- The height of a graph should be approximately two-thirds to three-quarters of its length.
- Normally, you start numbering both the X-axis and the Y-axis with zero at the point where the two axes intersect.
- However, when a value of zero is part of the data, it is common to move the zero point away from the intersection so that the graph does not overlap the axes.

Presenting Means and Medians in Graphs



Presenting Means and Medians in Graphs



Questions?