Module 2: Exploring Data with Graphs and Numerical Summaries

A primary goal of statistics is to visualize, summarize, and understand **variability**. A **variable** is any characteristic observed in a study.

Types of variables

- A variable is **categorical** if each observation belongs to one of a set of categories. e.g., class status: freshmen, sophomore, junior, senior, other
- A variable is **quantitative** if observations take on numerical values that represent different magnitudes of the variable.
  - A quantitative variable is **discrete** if its possible values form a set of numbers, and is usually a count. e.g., number of pets in a household
  - A quantitative variable is **continuous** if it can take on a continuum of infinitely many values, and is usually a measurement. e.g., time (in minutes).

A **frequency table** (typically used for qualitative variables) is a listing of possible values for a variable, together with the number of observations (i.e., the frequency) for each value. In addition, one may include

- **proportion** - frequency / total number of observations
- **percentage** - the proportion multiplied by 100

Proportions and percentages are also called **relative frequencies**

Example

Let's look at some examples in *R*.

Graphical summaries of categorical data

- A **pie chart** has a "slice of pie" for each category. The size of each slice corresponds to the percentage of observations in each category.
- A **bar graph** displays a vertical bar for each category. The height of the bar is the frequency or relative frequency of observations in the corresponding category. There should be a space between vertical bars. A **Pareto chart** is a bar chart where bars are arranged from tallest to shortest.

Example

Let's look at some examples in *R*.

Graphical summaries of quantitative variables

- A **histogram** uses bars to portray the frequencies or relative frequencies when observations are grouped into classes or *bins*

Example

Let's look at some examples in *R*

Describing the distribution of a variable

The **distribution** of a dataset is indicated by a frequency table or graphical display of one. There are three questions one often asks about a distribution.

1. Is there a gap in the data? Are there one or more 'outlier' observations?
2. Does the data have a single mound or **mode**. A distribution with *one* mode is called **unimodal**. A distribution with *two* modes is called **bimodal**
3. Is the distribution *symmetric* or *skewed*? A distribution is **symmetric** if its two sides are mirror images of each other. A distribution is **skewed** if one side stretches out longer than the other side.

Measuring the Center of Quantitative Data

Calculating the mean

The **mean** is the sum of the observations divided by the number of observations. When people use the word average, they usually (but not always!) are talking about the mean.
Let $n =$ the sample size. The mean has the formula $\bar{x} = \frac{\sum x}{n}$

Example

Suppose the height of a group of students in inches is 65, 59, 63, 61, 72. Find the mean:
The mean height in inches is $(65 + 59 + 63 + 61 + 72)/5 = 64$

Calculating the median

The **median** is the middle value of the observations when the observations are ordered from the smallest to largest. If there are two middle values, the median is the mean of these two values.

Example

Suppose the height of a group of students in inches is 65, 59, 63, 61, 72. Find the median:

1. Arrange the numbers in order from smallest to largest
   59, 61, 63, 65, 72
2. If n is odd, the median is the middle value; otherwise, the median is the mean of the two middle numbers
   Since $n = 5$ which is odd, the median is the middle number, at the $(n+1)/2$ position, which is 63 in this example

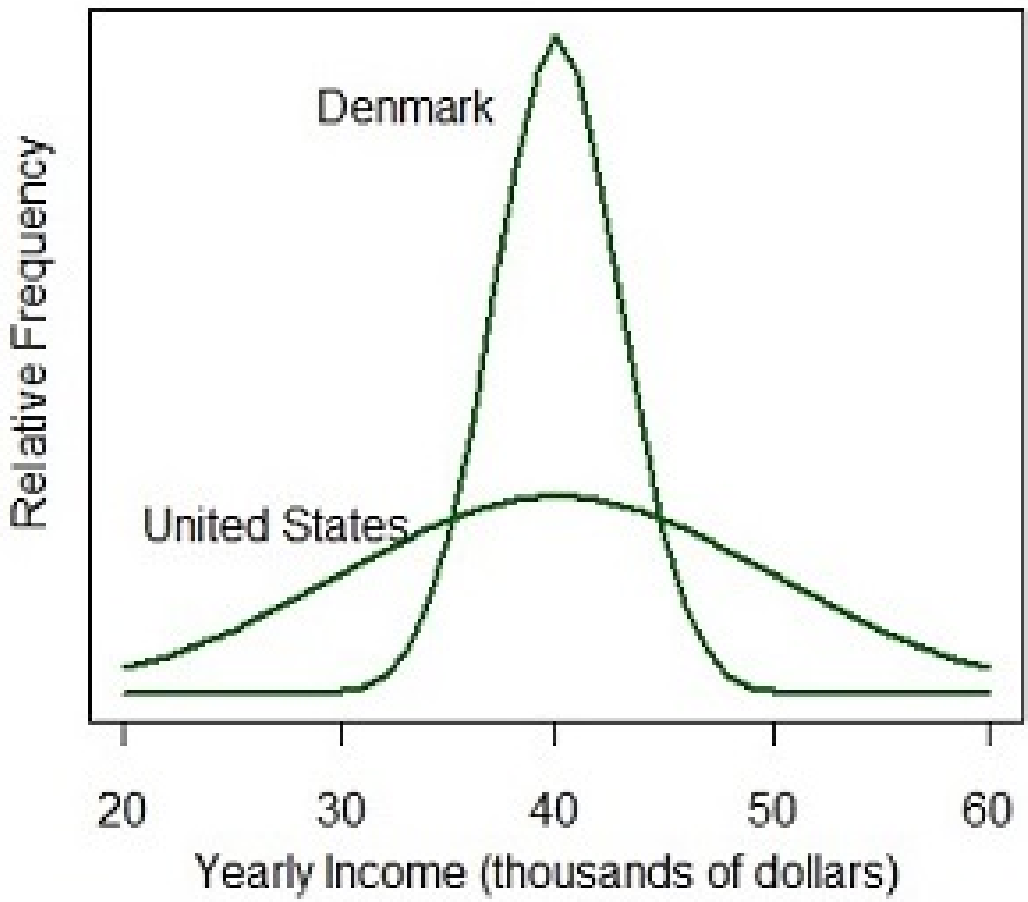
The **mode** is the value that occurs most frequently. This is often useful information. However, the mode is not necessarily near the 'center' of the data.

We will now do some exercises using R to gain further insight into Measures of Center for Quantitative Data.

Measuring the Variability of Quantitative Data

A measure of center is not enough to adequately describe the distribution of a quantitative variable

Hypothetical distribution of public school music teacher income



In the previous slide, both income distributions have the same mean (about \$40,000). However, the incomes are more similar in Denmark and more *variable* in the United States.

The variance (s^2) of a dataset is a measure of how spread out the data is around its mean. Usually, the standard deviation (s) is used to measure variability. The standard deviation is roughly the average (mean) distance of an observation from its mean. Specifically, s is equal to the square root of the mean squared deviation.

This is discussed in more detail on next few slides.

Deviation

The **deviation** of an observation x from its mean $\bar{x}$ is equal to $(x - \bar{x})$.

How might we use these deviations to measure variability?

Example

When asked "How many children do you think is ideal for a family?", the answers for 7 men were the following:
$0, 0, 0, 2, 4, 4, 4$

Value	Deviation	Squared Deviation
0	0 - 2 = -2	4
0	0 - 2 = -2	4
0	0 - 2 = -2	4
2	2 - 2 = 0	0
4	4 - 2= 2	4
4	4 - 2 =0	4
4	4-2 = 0	4
Total	0	24

Standard deviation formula

The **standard deviation** s of n observations is equal to

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{\text{sum of squared deviations}}{\text{sample size} - 1}}$$

and this is the square root of the **variance**, which is denoted s^2, which is the average square deviation of an observation from its mean,

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Interpretation: s represents a typical (or type of average) distance of an observation from the mean

Example (continued)

In the previous example, we found that the sum of the squared deviations was equal to 24.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{24}{6}} = \sqrt{4} = 2.0$$

Interpretation: The typical distance of an observation from its mean is 2.

In other words, Although the mean number of children was 2 for men, on average their responses tended to be 2 higher (4 children) or two lower (0 children) than that.

Properties of the standard deviation s

- The greater the variability of the data (with respect to the mean), the larger the value of s
- If $s = 0$, then all observations are the same and the data do not vary!
- The standard deviation (which is based on the mean), can be sensitive to outliers

Example

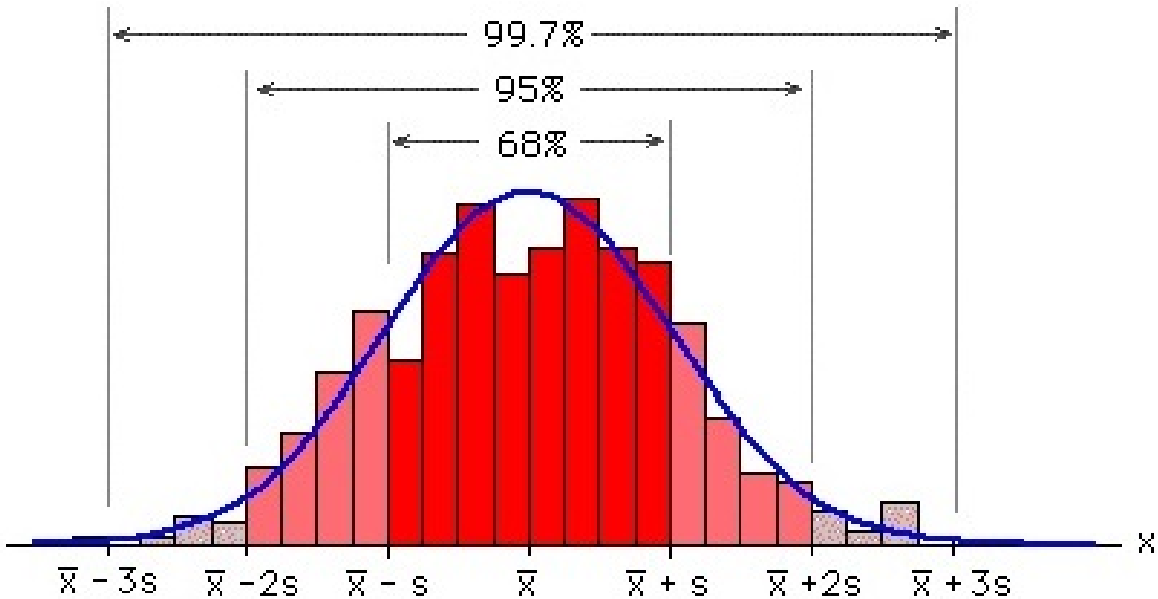
The file heights.csv contains data for the heights of males (coded as 0) and females (coded as 1). Let's use *R* to analyze the data.

1. Who is taller, on average, males or females?
2. Which gender has more variable heights? Males or females? Justify your answer

The empirical rule

If a distribution is **bell shaped**, then approximately

- 68% of observations fall within 1 standard deviation of the mean, that is, between the values of $\bar{x} - s$ and $\bar{x} + s$ (denoted $\bar{x} \pm s$).
- 95% of observations fall within 2 standard deviations of the mean ($\bar{x} \pm 2s$).
- All or nearly all observations fall within 3 standard deviations of the mean ($\bar{x} \pm 3s$).



Quartiles and other Percentiles

The p**th percentile** is a value such that p percent of the observations fall below or at that value.

Examples: a child's height is at the 43rd percentile, an SAT score is at the 90th percentile.

Special cases are

- The median or second quartile (Q2) is the 50th percentile ($p = 50$)
- The first quartile (Q1) is the 25th percentile ($p = 25$)
- The third quartile (Q3) is the 75th percentile ($p = 75$)

Measuring variability and detecting outliers

The **interquartile range** is the distance between the third and first quartiles, with **IQR = Q3-Q1**.

An observation is a potential outlier if it falls more than $1.5 \times$IQR below the first quartile or more than $1.5 \times$IQR above the third quartile.

Five-number summary

A graphical display of the data can be obtained by summarizing positions and describing the center and variability of the data. This is based on the IQR and the **five-number summary**:

1. Minimum value
2. Maximum value
3. First quartile (Q1)
4. Median
5. Third quartile (Q3)

Constructing a Box Plot

- Draw a box connecting Q1 with Q3
- Draw a line inside at the box indicating the median
- Draw a line from Q3 to the highest observation that is not an outlier (i.e., that is not $>$ Q3 $+$ 1.5$\times$ IQR), and draw a line from Q1 to the lowest observation that is not an outlier (i.e., that is not $<$ Q1 - 1.5$\times$ IQR). These lines are called *whiskers*. Use an asterisk or special notation to denote any outliers that are present.

Example

Boxplot example in R

1. Estimate the median height of males.
2. Estimate the median height of females.
3. Who is taller, on average?
4. How many outliers are there when looking at females?

Recognizing and Avoiding Misuses of Graphical Summaries

Always make sure you understand the underlying data! Are the graphs a valid representation of this data?

3D graphs or graphs based on complex figures often make differences seem larger than they really are.

Axes should be clearly labeled and vertical axis should generally start at 0.