List of Topics

- Types of Variables
  - Categorical/Qualitative versus Quantitative
- Quantitative Variable Types
  - Continuous versus Discrete
- Important Terms
  - Population, Sample, Subject, Variable, Parameter, Statistic
- Notation
  - \( \mu, \sigma, p \) versus \( \bar{x}, s, \hat{p} \)
- Descriptive versus Inferential Statistics
- Sampling Methods
  - Simple Random Sample, Systematic Sample, Convenience Sample, Stratified Sample, Cluster Sample
- Frequencies and Percentages
- Charts for Categorical Variables
  - Bar Graph, Pareto Graph, Pie Chart
- Charts for Quantitative Variables
  - Dot Plot, Histogram, Stem and Leaf Plot
- Distribution Shape
  - Symmetric, Skewed Left, Skewed Right
- Outliers
  - Interquartile Range Rule
- Mean, Median and Standard Deviation
- Five Figure Summary
  - Minimum, First Quartile, Median, Third Quartile, Maximum
- Empirical Rule for Bell Shaped Distributions
- Standardizing Data
- Percentiles
- Relative Risk
- Correlation
- Regression
  - Explanatory Variable, Response Variable, Lurking Variables, Residuals, Extrapolation
Practice Problems

1. Which of these are categorical or quantitative? For the latter, which are discrete or continuous?
   - Distance a car travels in 30 minutes. 
     - Quantitative, Continuous
   - Conference a school team is in. 
     - Categorical
   - Literary genre. 
     - Categorical
   - Number of times in one month the Creswell fire alarm goes off. 
     - Quantitative, Discrete

2. A college dean wants to know the average age of the faculty at her university. She takes a random sample of 10 faculty members and averages their ages. Identify what is the
   - Population
     - Sample
   - Subject
- Parameter

- Statistic

3. To estimate the percentage of older people (50+) in the U.S. who intended to purchase a laptop in the coming year, a simple random sample was selected using voter registration records. 33% of the people sampled said they intend to purchase a laptop in the coming year.

Which of the following statements are true?

I. 33% of all older people in the U.S. intend to purchase a laptop
II. 33% is a statistic
III. 33% is a parameter

II is true. I and III are about the population (which we do not know).

4. A survey was conducted to determine the effects of education on salary. Employees provided their salary (in dollars), highest level of education (high school degree, college degree, graduate degree, none) and weight (in pounds).

Which of the following statements are true?

I. Salary is a quantitative variable
II. Education is a categorical variable
III. Weight is a continuous variable

All three statements are true.

5. Identify the sampling method.

- A researcher takes 3 possible classifications of companies, each of which contains 1000 businesses, and draws 100 random subjects from all three. What type of sampling is this?
- Suppose instead she draws 200 businesses at random from the whole population of companies. What type of sampling is this?

- The same researcher instead randomly selects 2 of the 3 possible classifications and then surveys all businesses in those groups. What type of sampling is this?

- Suppose instead she gets an alphabetical list of all these companies, starts with #4, and selects every 100th after that.

6. We asked 200 families how many children a family has had.

<table>
<thead>
<tr>
<th># Children</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>61</td>
<td>64</td>
<td>50</td>
<td>25</td>
</tr>
</tbody>
</table>

Answer the following

- What proportion of families have 2 children

- What percentage of families have less than 2 children

- What proportion of families have more than 2 children

- What percentage of families do not have children
7. Consider the following table of 240 car brands we saw downtown:

<table>
<thead>
<tr>
<th>Toyota</th>
<th>Chevrolet</th>
<th>Honda</th>
<th>Ford</th>
<th>Nissan</th>
</tr>
</thead>
<tbody>
<tr>
<td>83</td>
<td>19</td>
<td>90</td>
<td>28</td>
<td>20</td>
</tr>
</tbody>
</table>

- What proportion of cars were not Toyota or Honda?

- If we were to make a pie chart, which car brands would have the largest and smallest slices?

- Find the mean, median, mode, and range from this data.

8. The following histogram shows the MPG from a random sample of cars

Answer the following questions using the above histogram.

- How many total cars sampled?

- Which MPG has the highest frequency? What is the frequency?
- How many cars sampled get between 20 and 30 MPG?

9. The following is a stem and leaf plot for the prices of randomly selected textbooks (in dollars)

<table>
<thead>
<tr>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>199</td>
</tr>
<tr>
<td>200</td>
</tr>
<tr>
<td>210</td>
</tr>
<tr>
<td>223</td>
</tr>
<tr>
<td>225</td>
</tr>
<tr>
<td>225</td>
</tr>
<tr>
<td>228</td>
</tr>
<tr>
<td>232</td>
</tr>
<tr>
<td>235</td>
</tr>
</tbody>
</table>

- What are the sell prices of the textbooks

- How many textbooks are in the sample

10. The following table shows the number of subways 200 randomly sampled New Yorkers ride to get to work

<table>
<thead>
<tr>
<th>Subways Taken</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>102</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
</tr>
<tr>
<td>4</td>
<td>32</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>200</strong></td>
</tr>
</tbody>
</table>

- Compute the mean

- Compute the median
11. A random sample of employees were selected. The five figure summary of the salaries is shown below.

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>46360</td>
<td>69693</td>
<td>77020</td>
<td>96750</td>
<td>129420</td>
</tr>
</tbody>
</table>

- Compute the interquartile range

- The individual with the salary of $129,420 receives a raise and now has a salary of $150,000. Which of the following is true?
  I. The mean would increase
  II. The median would decrease
  III. The range would change

I and III are true. The median does not change when a maximum value is changed.

12. Compute Drive Use (in kilobytes)

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>256</td>
<td>530</td>
<td>1105</td>
<td>320,000</td>
</tr>
</tbody>
</table>

- Is this bell shaped or skewed?

- Use the 1.5 * IQR rule to test for outliers.
13. The average SAT mathematics score of an incoming UGA student is bell-shaped with an average of 610 and a standard deviation of 50.

- Give an interval within which about 95% of the data fall.

\[ \mu = 610 \text{ and } \sigma = 50 \]

95% means we go 2 standard deviations to the right and left of the center.

Lower Limit:

\[ \mu - 2 \times \sigma = 610 - (2 \times 50) = 510 \]

Upper Limit:

\[ \mu + 2 \times \sigma = 610 + (2 \times 50) = 710 \]

So the interval is (510, 710).

- Approximately what percentage of the data is between 560 and 660?

Notice 660 - 610 = 50 and 610 - 560 = 50

We have therefore gone out 50 units, which is 1 deviation from the mean. By the Empirical Rule, 1 deviation has about 68% of the data.

- Find the score of an incoming student who is three standard deviations above the mean.

\[ \mu + 3 \times \sigma = 610 + (3 \times 50) = 760 \]

14. The distribution of the weight of an Irish Setter is bell-shaped with an average of 70 pounds and a standard deviation of 12 pounds.

- Approximately what percentage of the data is between 58 and 94?

We have 68% between 58 and 82 (so 16% in each tail)

We have 95% between 46 and 94 (so 2.5% in each tail)

That means we have 16% below 58 and 2.5% above 94, which leaves 81.5% in the middle

15. Two types of exams are offered in a statistics class. Type 1 is from a population with a mean of 68 and a standard deviation of 3 inches. Type 2 is from a population with a mean of 205 and a standard deviation of 35.

- We randomly select an individual who made a 74 on the Type 1 exam and we randomly select another individual who made a 220 on the Type 2 exam. Which student scored relatively higher?
16. In a sample of 261 individuals, the mean height was 65.8 inches and the standard deviation was 3.0 inches. The shortest person in this sample had a height of 56 inches.

- Calculate the standardized sample score for this person.

\[ Z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{56 - 65.8}{3.0} = -3.267 \]

- Interpret the standardized score.

This person’s height is 3.267 standard deviations below the mean (because it is negative). It is less than -3, so it is an unusual observation (outlier).

- What is the standardized score for someone whose height is 2.4 standard deviations below the mean?

\[ z = -2.4 \] (negative because below mean)

- Find the height corresponding to the above standardized score.

\[ -2.4 = \frac{x - 65.8}{3.0} \Rightarrow x = (-2.4)(3.0) + 65.8 = 58.6 \text{ in.} \]

17. We have a sample of how much time 200 subjects watch TV.

- How many people are above the 45th percentile

\[ 200 \times 0.45 = 90 \text{ fall below the 45th percentile} \]

\[ \text{Therefore } 200 - 90 = 110 \text{ are above} \]

\[ \text{Alternatively, } 200 \times (1 - 0.45) = 110 \text{ (just a different method)} \]

- For 200 subjects, what is the percentile for the person who’s 52nd from the top?

The 52nd from the top is the 200 - 52 = 148th person, so \[ \frac{148}{200} = 0.74 \]. Therefore it’s the 74th percentile.

- For 200 subjects, what is the percentile for the person who’s 52nd from the top?
18. We have a left skewed distribution. Answer the following.

- What percentage of the data are above the 61st percentile?

   By definition, 61% of the data are below the 61st percentile, so 39% of the data are above the 61st percentile.

- What percentage of the data are between the median and the 89th percentile?

   The median is the 50th percentile so 89% - 50% = 39%. 39% of the data fall between the median and the 89th percentile.

19. The following shows a breakdown of students based on football game attendance and happiness.

<table>
<thead>
<tr>
<th></th>
<th>Happy</th>
<th>Unhappy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attend</td>
<td>72</td>
<td>14</td>
<td>86</td>
</tr>
<tr>
<td>Do Not Attend</td>
<td>28</td>
<td>45</td>
<td>73</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>59</td>
<td>159</td>
</tr>
</tbody>
</table>

We want to examine if a student who attends football games is more or less likely to be happy.

- Find the proportion of students who are happy if they attend games.

  \[
  \frac{72}{86} = 0.83721
  \]

- Find the proportion of students who are happy if they do not attend games.

  \[
  \frac{28}{73} = 0.38356
  \]

- Find the relative risk of being happy for all students based on football game attendance.

  \[
  \frac{0.83721}{0.38356} = 2.18272
  \]

- Describe what these results tell you

  Students that attend football games are 2.18272 times more likely to be happy than students that do not attend games.

20. We want to predict average monthly car insurance payments (y), given the number of accidents (x) the client has had within the past three years.

  \[
  \hat{y} = 137.11 + 39.82x
  \]
- What’s the predicted payment for someone who has had 2 accidents

- Interpret the slope and intercept

- Is correlation positive or negative?

- Suppose someone with 2 accidents had an actual payment of $201. Compute this person’s residual.

- The model is based on people with between 0 and 6 accidents. Can we use it to predict the payment for someone with 13 recent accidents?

21. A coffee shop owner wants to assign a new price for a large cup of coffee. She is curious how the price per cup ($x$, in dollars) affects the number sold per day ($y$). She studies previous years’ data and gets $\hat{y} = 145 - 18x$.

- Interpret the slope

- Interpret the intercept
22. We want to predict the number of misprints \((y)\) in a novel that is \(x\) pages long (in hundreds). For instance, \(x = 2.5\) is a 250 page novel. The regression equation is \(\hat{y} = 5.1 + 3.2x\).

- Interpret the intercept (choose the best answer):
  (a) For every additional 100 pages, the predicted number of misprints goes up by 5.1.
  (b) The number of misprints in a novel 0 pages long is about 5.1.
  (c) The intercept has no practical interpretation.

- Interpret the slope (choose the best answer):
  (a) For every additional 3.2 pages, the predicted number of misprints goes up by 1.
  (b) A novel 400 pages long can be expected to have 3.2 more misprints than a novel 300 pages long.
  (c) The slope has no practical interpretation.