WELCOME TO STAT 2331!!!

Before we embark on our journey through the land of Statistics, let us first convince ourselves why do we really need to go through it!

Q: Why we study statistics?

Statistics is the science of making sense of data and of how to gather data.

Reason to study Statistics:
- To be an informed “information consumer”.
- Decision making, including personal decisions.

The key idea of this course:
- Suppose we have 2 groups of patients, 300 each.
- 183 patients are ‘cured’ in the standard treatment group.
- 214 patients are ‘cured’ in the new treatment group.
- Is the new treatment better?

**Method of Generating Data**

**Experiment**: Active - Control which units get which treatments
**Sampling**: Passive
**Anecdote**: Haphazard

**Example**: *Is there a connection between olive oil consumption and avoiding cancer?*

**Experiment**: Two groups of rats, one gets olive oil, other doesn’t. See which group has higher rate to get cancer.

**Sampling**: Look at large group of people, randomly chosen from the population. Ask them about olive oil consumption and try to relate that to cancer.

**Anecdote**: My friend’s grandfather drank a pint of olive oil every morning and lived to 110!!
Chapter 1: Examining Distributions

Goal: Learn to examine a set of data.

A data set consists of individuals characterized by variables.

- **Individuals** - objects being described by the data, e.g. Students in Stat2331 class
- **Variables** - characteristic of an individual. It varies from individual to individual, e.g. Height of student in Stat2331 class

Different kinds of variables:

  a) **categorical** - not expressed as number, but as an adjective or category, e.g., Gender, race  
     → percentages make sense
  b) **quantitative** - expressed as number, e.g., Height, weight  
     → averaging makes sense

**Example:** Fuel Economy of some 1998 model motor vehicles

<table>
<thead>
<tr>
<th>Make &amp; Model</th>
<th>Vehicle Type</th>
<th>Transmission</th>
<th>Number Cylinders</th>
<th>City MPG</th>
<th>Highway MPG</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMW 318I Subcompact</td>
<td>Automatic</td>
<td>4</td>
<td>22</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>BMW 318I Subcompact</td>
<td>Manual</td>
<td>4</td>
<td>23</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Buick Century Midsize</td>
<td>Automatic</td>
<td>6</td>
<td>20</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>Chevy Blazer 4Wheel Drive</td>
<td>Automatic</td>
<td>6</td>
<td>16</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

Q: What are the individuals in this data set?

Q: Which variables are categorical? Quantitative?
**Distribution** of a variable tells us what values the variable can take and how often it takes these values.

**Example:** In last quarter there were 60 Stat2331 students. Of those, suppose 18 had blue eyes, 36 had brown eyes and 6 had green eyes. What is the distribution of eye color for this class?

**Values:**

**Frequency:**

**Percent:**

**Issue:** How do we examine and describe the main features of a data set?

We do *exploratory data analysis*. It involves
- examining each variable
- examining relationships between variables

This is accomplished using
1) graphs  
2) numerical summaries
Section 1.1: Summarizing Data Graphically

*Graphing Distributions of Categorical Data*

Distribution of categorical data involves

- listing all possible levels or categories
- counting or obtaining percentages of individuals who fall into each category

Two typical graphs used for categorical data are

1) Bar graphs  
2) Pie Charts

The main drawback of the pie chart is that you need to include *all categories* that make up a whole. That is, the percentages of all categories **must sum to 100%**. This is not required for a bar chart.

**Example:** What type of graph(s) can be used to summarize the ethnic origins of the students in the statistics department?

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Caucasian</th>
<th>Asian</th>
<th>African-American</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>50</td>
<td>40</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>
Example: The data summarized below are the % of females among people earning doctorates in 1993 in several fields of study.

<table>
<thead>
<tr>
<th>Field</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer science</td>
<td>14.4%</td>
</tr>
<tr>
<td>Life sciences</td>
<td>39.9%</td>
</tr>
<tr>
<td>Education</td>
<td>59.2%</td>
</tr>
<tr>
<td>Engineering</td>
<td>9.6%</td>
</tr>
<tr>
<td>Physical sciences</td>
<td>21.9%</td>
</tr>
<tr>
<td>Psychology</td>
<td>61.2%</td>
</tr>
</tbody>
</table>

What type of graph(s) can be used to summarize the data?

Notes:
(1) Bar-charts can be made by using counts or % but for pie-charts we need %s.
(2) Keep gaps between the bars in a bar-chart.
Graphing Distributions of Quantitative Variables

Two types of graphs are typically used for quantitative data.

1) Stemplots
   → Work best for a small number of observations that are all greater than or equal to 0.

To make a stemplot:

a) **Sort the data.** Separate each observation into a leaf, which is the final digit, and the stem, which consists of the remaining digits.

b) Write the stems in a column in consecutive order, with the smallest at the top, and draw a vertical line to the right of this column.

c) Write each leaf in the row to the right of its stem, with the leaves in consecutive order, smallest to largest.

Example: Initial blood pressure readings are taken for a group of subjects taking part in an experiment to see if calcium in the diet lowers blood pressure.

Calcium Group

| 112 | 111 | 107 | 136 | 107 | 112 | 102 | 123 | 129 | 110 |

Placebo Group

| 117 | 123 | 112 | 110 | 112 | 109 | 114 | 98  | 119 | 102 | 130 |

The distribution of the calcium group using a stemplot looks like:
Back-to-back stemplots are useful to compare two sets of data.

Example: "Initial blood pressure readings" (contd.)

When all the leaves would fall on just a few stems, we can split stems to double the number of stems. A stem that consists of data from 0 to 9 may be split into two stems: 0 to 4 and 5 to 9.

Example: "Initial blood pressure readings" (contd.)
2) **Histograms**

→ Useful for large data sets

A histogram breaks up quantitative variables into intervals and displays the count or percentage of the observations that fall into each interval.

**Example:** On a 20-pt. quiz, a section of 40 students had these scores: *(Each class interval includes the right endpoint but not the left endpoint of the interval)*

<table>
<thead>
<tr>
<th>Score</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exactly 0</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>0 - 5</td>
<td>3</td>
<td>3/40=7.5%</td>
</tr>
<tr>
<td>5-10</td>
<td>5</td>
<td>5/40=12.5%</td>
</tr>
<tr>
<td>10-15</td>
<td>12</td>
<td>30%</td>
</tr>
<tr>
<td>15-20</td>
<td>20</td>
<td>50%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>40</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

**Notes:**

(1) Histograms look like bar charts without the gaps between bars.

(2) If frequencies (same as counts) are used and the intervals are all of the same width, then the sum of the heights of the bars must equal the total number of observations.

(3) If %s are used, then the sum of the %s of all the bars must equal 100%.

**Comparison of histogram and stemplot**

Histograms do not display the actual observed values while stemplot do.

→ You can reconstruct the original data set from a stemplot.
Describing Distributions

Overall pattern of a distribution is described by its

(1) shape

a) symmetry
   • symmetric - area to the left and right of the midpoint of the distribution are mirror images of each other.

   ![Symmetric Distribution](image1)

   • skewed - area is concentrated on one side and there is long tail on the other side.

   ![Skewed Distributions](image2)

b) # of modes (most frequent observation) or peaks of the distribution e.g. unimodal (one peak) or bimodal (two peaks)

   ![Unimodal and Bimodal Distributions](image3)
(2) **center** - midpoint of the distribution

![Diagram showing center, median, and mean]

(3) **spread** - range (smallest and largest values) of the distribution

Deviation from the overall pattern
- **outliers** - values that are far from the bulk of data.

**Example**: Draw a histogram of the starting salaries (in $1000's) of last year's graduating students of SMU.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>14</td>
<td>47</td>
<td>26</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Use the histogram to answer the following questions:
Q: Describe the shape of the distribution?

Q: What is the range of the salaries in this data set?

Q: Where do most of the salaries seem to be centered?

Q: What percentage of students had starting salaries between $25,000 and $45,000?

Q: In which class interval does the third quartile lie?