STOR 155 Introductory Statistics

Lecture 1: Overview; Displaying Distributions with Graphs
Tip: Strategy for Success

• Stay active/involved in class.
• Ask questions during class (especially if you do not understand something).
• Answer questions to help other students if you can.
• Keep pace with the lectures, review daily, do homework after each lecture to help understand the materials.
• Make effective use of office hours (Instructor), open tutorial sessions, UNC Learning Center.
  – Help you answer questions about homework and lectures
  – Private time vs. public time
What is Statistics?

**Statistics:** the science of collecting, organizing, analyzing and interpreting *data* (= information)
SAT Scores

• Some parents and teachers have been concerned about the trend of declining SAT scores …

• Question: effect of classroom atmosphere (*strict* or *liberal*)?

• To answer the question, 50 students (24 males and 26 females) participated in a study on their performance, as measured by SAT scores at the end of the school year.

• The students were divided into two groups of 25 each (12 males and 13 females), with Group 1 to study under a *strict* atmosphere while Group 2 under a very *permissive* atmosphere.

• They were matched according to socio-economic background.
SAT Scores

- After 9 months, all students were given the same standardized tests: verbal and math.

<table>
<thead>
<tr>
<th>Student</th>
<th>Group</th>
<th>Gender</th>
<th>SATMath</th>
<th>SATVer</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Strict</td>
<td>F</td>
<td>670</td>
<td>700</td>
</tr>
<tr>
<td>B</td>
<td>Strict</td>
<td>M</td>
<td>700</td>
<td>680</td>
</tr>
<tr>
<td>C</td>
<td>Liberal</td>
<td>F</td>
<td>750</td>
<td>730</td>
</tr>
<tr>
<td>D</td>
<td>Liberal</td>
<td>M</td>
<td>690</td>
<td>750</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>
SAT Scores

• This example involves *data collection, data analysis, and statistical inference*.
  – How?

• Questions:
  – Does stricter classroom atmosphere increase the average score?
  – Why “matched according to socio-economic background”?
  – Why “12 males and 13 females per group”?
  – Is the group size 50 large enough to make a confident conclusion?
Fundamental Concepts

• **Population**: the entire group of individuals that we want information about.
  – Students (who are about to take SAT)

• **Sample**: a part of the population that we actually examine in order to gather information.
  – those students selected into the study

• **Sample size**: number of observations/individuals in a sample.
  – 50

• **Statistical inference**: to make an inference about a population based on the information contained in a sample.
  – Based on the data from the study, to infer whether a stricter classroom atmosphere increases SAT scores in general.
Fundamental Concepts

• A *parameter* is a value that describes the population. It’s fixed but unknown in practice.
  – the average SAT score of all the students, who are about to take SAT.

• A *statistic* is a value that describes a sample. It’s known (calculated) from the sample.
  – the average SAT score of all the students, who are selected into the study.
  – a sample analogue of the parameter.
Practice Exercise

• Suppose you are interested in finding the average SAT score of UNC unders,
  -- SAT scores of all UNC unders in STOR155 (sample)
  -- SAT scores of all UNC unders (population)
• Suppose you are interested in finding the average SAT score of US unders,
  – SAT scores of all UNC unders ( )
  – SAT scores of all US unders ( )
Summary

• **Statistics is the science of data:**
  – Collecting
  – Organizing and analyzing
  – Decision making

  = Information processing

• **Fundamental concepts:**
  – Population, parameter, sample, statistic, sample size

• **You can do a LOT with statistics … what?**
Take home message

- Interested in population, but it’s too large to become known completely
- Statisticians work on sample, which is a smaller and observable "proxy"
- There is uncertainty in this transition, hence errors are inevitable …
- That’s why statistical methods are needed …
Chapter 1: Looking at Data - Distributions

1.1 Displaying Distributions with Graphs

1.2 Displaying Distributions with Numbers

1.3 Density Curves and Normal Distributions
Data

Data contain

- **Individuals**: the subjects described by the data;
- **Variables**: any characteristic of an individual. A variable can take different values for different individuals.
Categorical & Quantitative Variables

• A **categorical variable** places an individual into one of several groups or categories.

• A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense.
NBA Draft 2005

<table>
<thead>
<tr>
<th>Name</th>
<th>Team</th>
<th>Nationality</th>
<th>Weight</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Bogut</td>
<td>Milwaukee</td>
<td>Australia</td>
<td>245</td>
<td>7-0</td>
</tr>
<tr>
<td>M. Williams</td>
<td>Atlanta</td>
<td>US</td>
<td>230</td>
<td>6-9</td>
</tr>
<tr>
<td>D. Williams</td>
<td>Utah</td>
<td>US</td>
<td>210</td>
<td>6-3</td>
</tr>
<tr>
<td>C. Paul</td>
<td>New Orleans</td>
<td>US</td>
<td>175</td>
<td>6-0</td>
</tr>
<tr>
<td>R. Felton</td>
<td>Charlotte</td>
<td>US</td>
<td>198</td>
<td>6-1</td>
</tr>
</tbody>
</table>

...
NBA Draft 2005

• Variables:
  – Team & Nationality - Categorical
  – Weight & Height - Quantitative

• How many teams in the draft? How many players drafted by each team?
• How many players higher than 6-9? How many players between 200 and 250 pounds?
• Equivalently, what is the distribution for each variable?
Distributions of Variables

• The distribution of a variable indicates what values a variable takes and how often it takes these values.
  – For a categorical variable, distribution: categories + count/percent for each category
  – For a quantitative variable, distribution: pattern of variation of its values
# Highest Level of Education for People Aged 25-34

<table>
<thead>
<tr>
<th>Education</th>
<th>Count (millions)</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than high school</td>
<td>4.6</td>
<td>11.8</td>
</tr>
<tr>
<td>High school graduate</td>
<td>11.6</td>
<td>30.6</td>
</tr>
<tr>
<td>Some college</td>
<td>7.4</td>
<td>19.5</td>
</tr>
<tr>
<td>Associate degree</td>
<td>3.3</td>
<td>8.8</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>8.6</td>
<td>22.7</td>
</tr>
<tr>
<td>Advanced degree</td>
<td>2.5</td>
<td>6.6</td>
</tr>
</tbody>
</table>
Exploratory Data Analysis (EDA)

- Use statistical tools and ideas to help us examine data
- Goal: to describe the main features of the data
- NEVER skip this
- EDA
  - Displaying distributions with graphs
  - Displaying distributions with numbers
Basic Strategies for EDA

- **Strategy I**
  1. One variable at a time
  2. Relationships among the variables

- **Strategy II**
  1. Graphical visualizations
  2. Numerical summaries
Graphic Techniques for Categorical Variables

- **Bar Graph** uses bars to represent the frequencies (or relative frequencies) such that the height of each bar equals the frequency or relative frequency of each category.
  - Frequencies: counts
  - Relative frequencies: percent
  - height indicates count or percent

- **Pie Chart** is a circle divided into a number of slices that represent the various categories such that the size of each slice is proportional to the percentage corresponding to that category.
  - area = relative %
  - Note: Pie chart requires to include all the categories that make up a whole.
### Highest Level of Education for People Aged 25-34

<table>
<thead>
<tr>
<th>Education</th>
<th>Count (millions)</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than high school</td>
<td>4.6</td>
<td>11.8</td>
</tr>
<tr>
<td>High school graduate</td>
<td>11.6</td>
<td>30.6</td>
</tr>
<tr>
<td>Some college</td>
<td>7.4</td>
<td>19.5</td>
</tr>
<tr>
<td>Associate degree</td>
<td>3.3</td>
<td>8.8</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>8.6</td>
<td>22.7</td>
</tr>
<tr>
<td>Advanced degree</td>
<td>2.5</td>
<td>6.6</td>
</tr>
</tbody>
</table>
Graphic Techniques for Quantitative Variables

- Stemplot (Stem-and-Leaf Plot)
- Histogram
- Time plot
Stemplot

• Separate each observation into a stem consisting of all but the final (rightmost) digit and a leaf, the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.

• Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.

• Write each leaf in the row to the right of its stem, in increasing order out from the stem.
# of Home Runs per Season

- **Babe Ruth (New York Yankees): 1920-1934**
  - 54 59 35 41 46 25 47 60 54 46 49 46 41 34 22

- **Mark McGwire (St. Louis Cardinals): 1986:2001**
  - 3 49 32 33 39 22 42 9 9 39 52 58 70 65 32 29

- **Question:** (see Ex 1.7 p9 –11 1st …)
  - Work out the stem-plot of McGwire
  - back-to-back stem-plot of the two players
Example: Midterm Scores of STOR 151

The following data set contains the midterm exam scores of STOR 151.

| 74 | 76 | 78 | 88 | 87 | 87 | 53 | 95 | 82 | 79 | 79 | 78 |
| 62 | 80 | 77 | 70 | 60 | 60 | 84 | 95 | 85 | 93 | 79 | 84 |
| 71 | 85 | 100| 77 | 72 | 95 | 79 | 83 | 97 | 87 | 73 | 84 |
| 74 | 83 | 85 | 95 | 62 | 50 | 86 | 83 | 86 | 36 |   |   |
Splitting & Trimming Stems

• For a moderate number of obs,
  – Split each stem into two: one with leaves 0-4 and the other with leaves 5-9
  – Increase # of stems, reduce # of leaves

• Trimming:
  – If the observed values have too many digits, you can trim them by rounding to a certain digit.

• Disadvantage of stemplots
  – Awkward for large data sets
Example: A study on litter size

- **Data**: (170 observations)

```plaintext
4  6  5  6  7  3  6  4  4  6  4  4  9  5 10  6  6  5  6  6  
8  2  7  7  9  3  7  5  7  7  4  5  5  6  7  6  7  8  
6  6  7  6  6  7  5  4  5  6  6  1  3  4  7  5  4  7  5  
8  8  5  6  8  5  5  4  9  6  7  3  7  7  5  4  6  9  6  
7  7  5  7  3  7  6  5  3  7 10  5  6  8  7  5  5  7  5  
5  8  9  7  5  7  5  5  5  6  3  7  8  7  7  6  3  4  4  
4  7  2  7  8  5  8  6  6  5  6  4  7  5  5  6  9  3  5  
4  8  3  9  8  3  6  5  4  7  8  4  8  6  8  5  6  4  3  
8  8  6  9  5  5  6  6  7  6  8  6 11  6  5  6  6  3  
```
Stem-and-leaf plot for pups

0|1223333333333333344… (35)
0|5555555555555555555555555… (132)
1| 001
Take Home Message

- **Data:**
  - Individuals
  - Variables
    - Categorical variables
    - Quantitative variables

- **Distribution of variables**

- **Graphical tools for categorical data**
  - Bar graph
  - Pie chart

- **Graphical tools for quantitative data**
  - Stemplot