Welcome to STT 201!

I. **Highlights of syllabus**

- Professor: Vince Melfi

- TAs:
  - Don DeGroat (Sections 5,6)
  - Fang Li (Sections 1,3)
  - Xiang Li (Sections 2,4)
  - Santhosh Srinivasan (Sections 9,10)
  - Jing Wang (Sections 7,8)

- Getting help
  - Ask questions during lecture and labs.
  - Ask questions after lecture (I'll stay after lecture every day).
  - Help Room (C100 Wells Hall) staffed by TAs. Hours announced soon. Ask questions about readings, lecture, labs, exercises, etc.
  - TA office hours (announced soon).
- Web site: Go to `www.stt.msu.edu/~melfi` and follow the link for STT 201.

- Grading
  - Thirteen five point quizzes given on Wednesdays. Quizzes are cumulative. Low three dropped. No make-ups.
  - Thirteen five point computer lab projects. Low three dropped. No make-ups.
  - One thirty point exam.
  - Points possible: 50 + 50 + 30 = 130. Compute percentage. Grading scale on syllabus.
- Exercises
  - Many exercises assigned from each chapter.
  - Not collected, but essential for understanding.
  - Work together if you'd like. (I encourage this, but make sure you all understand.)

- Computer Labs
  - You'll learn more about policies (due dates, etc.) during the first lab.
  - You'll learn the statistical package Minitab.
  - ***Lab meets this week!!***
  - ***Lab is open but optional next (Labor Day) week.*** Come if you need more time to finish the first project (unlikely) or if you want extra guidance and help with Minitab.
- Readings
  - Reading assignments will be announced in class
  - You're responsible for understanding the material in the readings, even if it isn't explicitly covered in lecture.
  - Feel free to skip all "Computer Tips" since we'll cover computers in lab.

- Attendance and courtesy
  - No (explicit) penalty for missing class.
  - If you miss class, you're responsible for learning the material
    - Read friend's notes
    - Read my notes on web
    - Get help in help room
  - If you arrive late or leave early, sit near an exit
  - Be attentive and quiet during lecture.

II. **This week's reading and exercises**

- Chapter 1: Read Section 1.1 for vocabulary. Skip or skim the rest of Chapter 1.
- Chapter 2: Read Sections 2.1–2.4. Skip "Time Series Data" on pp. 58–61. Skip "Polygons" on pp. 85–86.
- Exercises
  - None from Chapter 1
  - Chapter 2 (and 3) exercises on syllabus.

III. **Population, sample, et. al. (Section 1.1)**

Example: Want to know the proportion of people who will vote for a school bond issue. Choose 250 registered voters at random and ask whether they'll support the bond. The proportion who say YES is our estimate of the proportion who will support the bond.

- The people who will vote is the *target population*. (The group we're interested in.)
- All registered voters is the *sampled population*. (The group we choose from.)
- The 250 people we ask is the *sample*.

Example: An MSU researcher wants to know the effectiveness of a smoking cessation program. Ask for students who smoke to volunteer for the program, and monitor their smoking behavior.

- (Probably) all smokers is the target population.
- All MSU student smokers is the sampled population.
- Those who volunteer comprise the sample.

Important points:

1. The group of interest (target population) may be different from the group we choose from (sampled population). In both cases this may affect the validity of our conclusions.
   - Many registered voters won't vote, especially in a school bond election.
   - The program may work well for students but not for others.
   - In one case the target population is larger than the sampled population. In the other it is smaller.

2. In the smoking study the sample is not random. May introduce *bias* in the conclusions. Maybe those who are willing to volunteer are different in important ways from general smokers.

3. Even if the target and sampled populations are the same, and the sample is chosen at random, it's possible that it will not be representative of the population. One of the main things we'll learn is how to assess the likelihood of such an event.

4. If the sample consists of the entire population, it's called a census.
   - Is the U.S. census a census???
     - Not really. Some people aren't counted. (Homeless, poor, etc.)
     - In fact, there is major controversy about proposed "statistical adjustment" of the census!!

IV. Types of Data (Section 2.1)

We can't treat all data alike! It's important to recognize various types of data and to use methods appropriate for the data type.

Example: Exercise 2.22 gives the following data for the eye color of a random sample of college students.

Eye color	Blue	Brown	Green	Hazel
Number	124	150	15	103

Figure 1: Eye color of college students from Exercise 2.22

It clearly doesn't make sense to talk about the mean eye color! On the other hand, a pie chart of the data would be informative. If the data were numerical, maybe a mean would be reasonable and a pie chart would not.

Types of data

- Numerical: Examples are weight, distance, age in years, number of children in a family. Numerical variables are divided into two types.
  - Discrete: Have only a finite (or countable) number of possible values. Number of children and age in years are discrete.
  - Continuous: Not discrete. Examples are weight and distance.

  (The distinction is sometimes blurry. For example, in reality weight is measured to something like the nearest pound, and is discrete. But it's useful to ignore this and treat weight as continuous.)
- Categorical: Examples are gender, major, disease status (ill or well), marital status.

Example: Exercise 2.6

- Gender is categorical
- Age is numerical and continuous.
- Marital status is categorical.
- Occupation is categorical.
- Type of surgery needed is categorical.
- Urgency of surgery is probably categorical.

Displaying data

Choosing good methods for displaying data can be very important. Here are two examples taken from a beautiful book about displaying data called “The visual display of quantitative information” by E. Tufte.

Example: Devastating outbreaks of cholera occurred in the 19th century in Europe. The cause and mode of transmission of cholera were unkown at the time. Dr. John Snow, however, believed that cholera was spread through contaminated water. (This was later proved to be true.) In support of his theory, he prepared the following very convincing graphic. On the graphic, an “x” marks the location of a pump where water was available, and small dots represent cholera cases.

From the graphic, it’s pretty clear that most of the cholera cases in this outbreak were clustered around the Broad Street pump. In fact, once the pump was disabled the outbreak subsided.

Example: Of course graphics can be misleading. Here is a particularly egregious example related to fuel economy standards set by congress in 1978. The first graphic appeared in the New York Times. It misleads the reader by making the line for the year 1985, which is supposed to represent a standard of 27.5 mpg, about 8.3 times as long as the line for the year 1978, which is supposed to represent a standard of 18 mpg. The second graphic is a more honest view of the standards.

V. **Displaying categorical data (Sec 2.2)**

- We'll learn about *frequency tables*, *bar graphs*, and *pie charts*.

Example: The next page contains a frequency table of the racial composition of Ingham County, according to the 2000 census. Note that the *relative frequency* of a category is just the proportion of the data that are in that category.

Race	Frequency	Relative Frequency
White	221935	0.795
Black or Afr. Am.	30340	0.109
Am. Indian or Alaska Native	1528	0.005
Asian	10273	0.037
Native Hawaiian etc.	143	0.001
Other	6746	0.024
Two or More races	8355	0.029
Total	279320	1.0

Table 1: Racial composition of Ingham County according to the 2000 census.

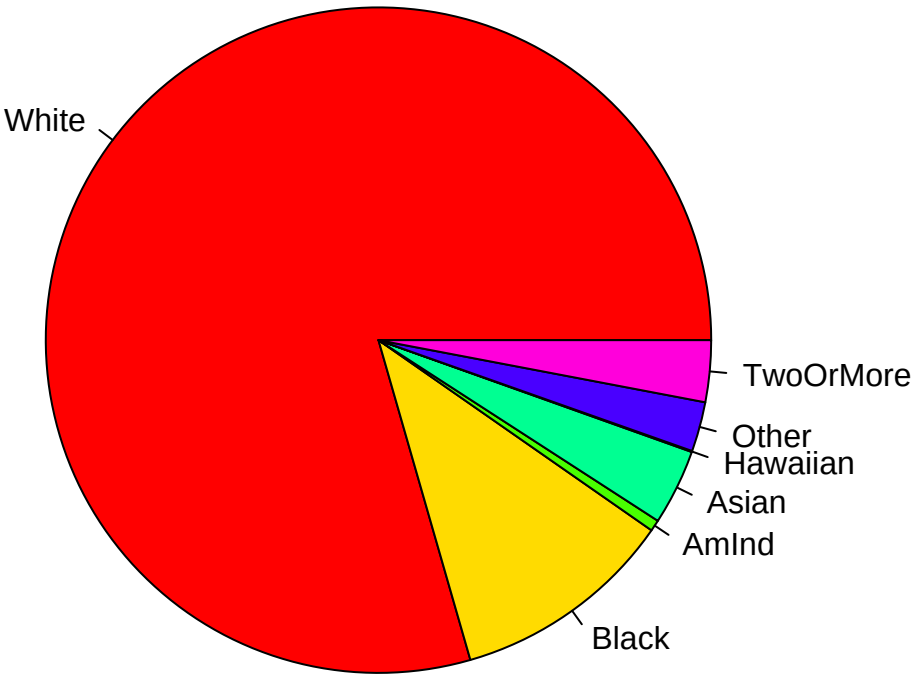

Figure 2: Pie chart of the races of Ingham county residents from the 2000 census

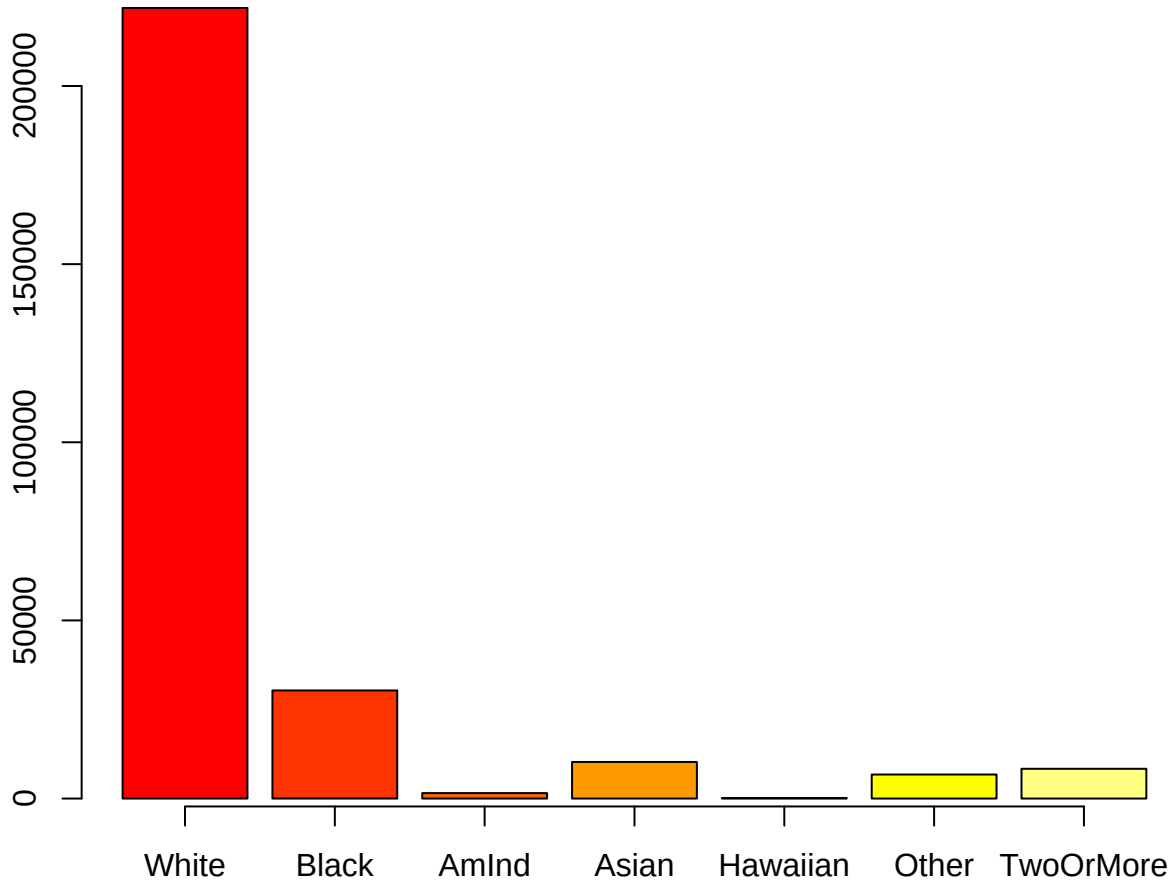

Figure 3: Bar chart of the races of Ingham county residents from the 2000 census

- In this case either display gives a reasonable summary of the data.
- In recent years it has been found that people have a hard time judging relative areas and hence have a hard time interpreting pie charts. But for good or bad they are very widely used.

VI. **Displaying numerical data**

- Goal: Graphical summary of data that informs about center, variability, shape, etc.
- Stem and Leaf Plots: Quick display for small datasets.
- Histograms: Useful for any size dataset. Easy to construct with a computer.

Example: Here are Babe Ruth's home run totals for the 15 years he played for the Yankees.

```
54 59 35 41 46 25 47
60 54 46 49 46 41 34 22
```

Here is a stem and leaf plot of these data.

```
2 | 25
3 | 45
4 | 1166679
5 | 449
6 | 0
```

From it we get a quick picture of the distribution of the data values. For example, we see that about half the years his total was in the 40s.

Sometimes a back-to-back stem and leaf plot allows us to quickly compare two datasets. Here is a back-to-back stem and leaf plot of Babe Ruth's HR totals and Mickey Mantle's HR totals.

```
MANTLE        RUTH
 9853  | 1 |
73321  | 2 | 25
75410  | 3 | 45
   20  | 4 | 1166679
   42  | 5 | 449
       | 6 | 0
```

Histograms

To draw a histogram:

- Divide numerical data into classes.
- Count the number of observations in each class.
- Draw a bar graph.

Usually we'll let Minitab do this for us!

Example: Data were collected on mercury concentrations (parts per million) in 52 Florida lakes. Some of the data are `1.23, 7.00, 6.00, 0.44`. Here are the data divided into classes:

Category	Number of Lakes
0 to 2	20
2 to 4	0
4 to 6	2
6 to 8	4
8 to 10	1
10 to 12	10
12 to 14	13
14 to 16	2

Here is a histogram of the mercury data.

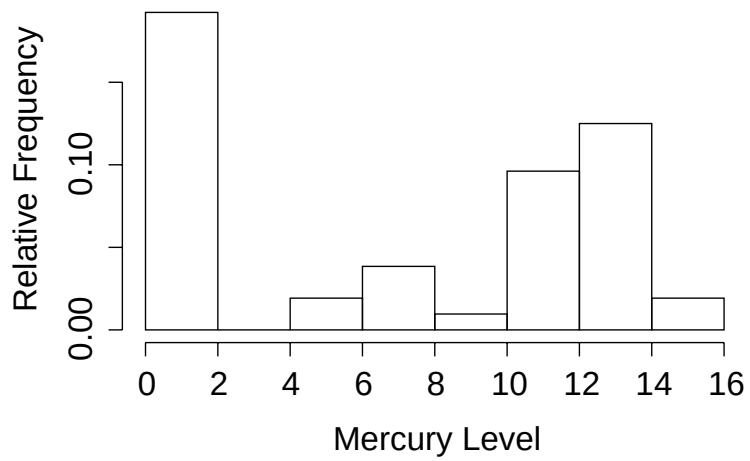

Figure 4: Histogram of mercury concentrations in 52 Florida Lakes

What we learn:

- Non-symmetric shape
- Many lakes with low mercury level; many lakes with high level; few in middle.
- Levels are all between 0 and 16 ppm.

Some issues:

- Different choices for classes lead to different looking histograms. More on this later.
- Endpoints.
  - Q: Should 6.00 go into the class 6 to 8 or the class 4 to 6?
  - A: Just be consistent; if it goes into 6 to 8, then 10.00 should go into 10 to 12.
- This histogram is drawn using a relative frequency scale. Sometimes a frequency scale is used.

Example: Chest sizes of 5738 scottish militiamen from early 19th century.

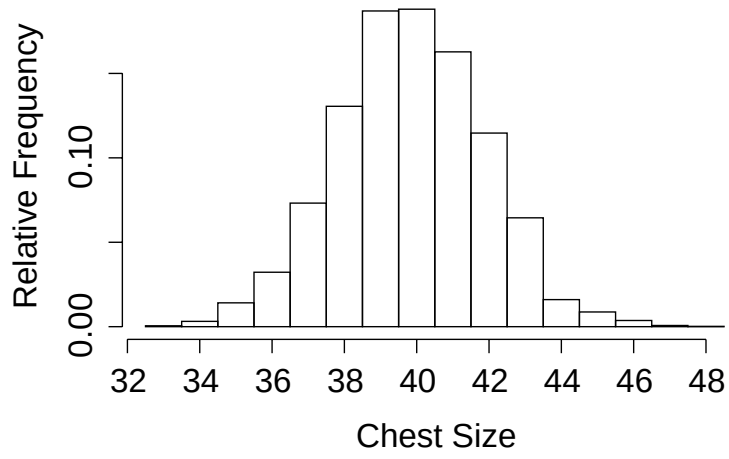

Figure 5: Histogram chest sizes of scottish militiamen

- Center around 40 inches.
- Symmetric and bell-shaped.
- Range between about 34 and 46 inches.

Example: Who wrote the disputed *Federalist Papers*? Hamilton or Madison? Look at rates of use of various words. We'll look at the word *by*. See Figure 6.

- Histograms suggest that Madison wrote the disputed papers.

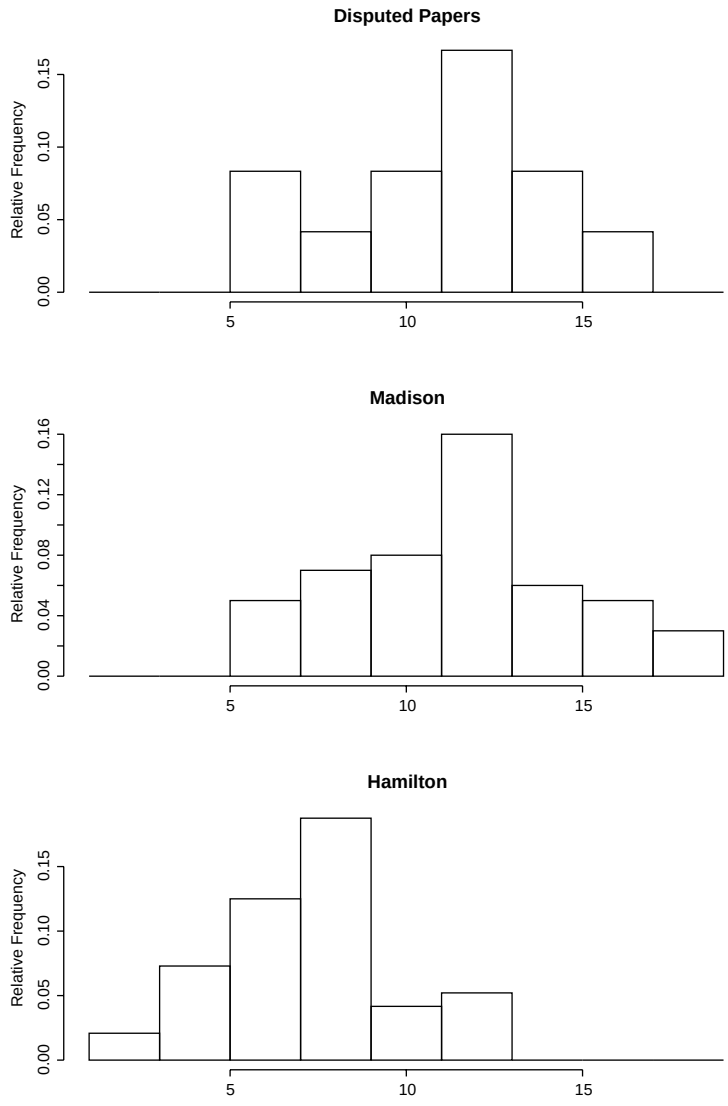

Figure 6: Rate of use (per 1000 words) of *by* in Hamilton's work, Madison's work, and the disputed papers.

VII. **Numerical measures of location**

- Histograms and stem and leaf plots are useful, but usually a few numbers summarizing the location, variability, etc. of the data are desired.
- Measures of center include the *mean*, *median*, *trimmed mean*.

Example: (Only partly true) Net worth of 10 randomly chosen residents of Washington state (in thousands of dollars): `$100, $1000, $250, $25, $750, $575, $2500, $3200, $670, $320`.

- Mean worth (in thousands of dollars) is

$$\frac{100 + 1000 + \cdots + 320}{10} = 939.$$

- Median worth (in thousands of dollars) is (arrange in order; pick 2 middle numbers; compute their mean)

$$\frac{575 + 670}{2} = 622.5.$$

What happens if we add an *outlier*: Bill Gates' net worth of \$40.5 billion dollars, which is \$40500000 thousands of dollars?

- The new mean is

$$\frac{100 + 1000 + \cdots + 40500000}{11} = 3682672.$$

- The new median is 670.
- Outliers can change the mean dramatically! The median is relatively unaffected and hence is called *resistant*.

Trimmed mean For many reasons people like the mean as a measure of location. But people don't like its lack of resistance to outliers. Trimmed means provide a compromise between the mean and median.

Example: 10% trimmed mean for wealth data.

- Arrange the data in order.
- Remove the smallest 10% of data and largest 10% of data, and compute the mean of what remains.
- In our case 10% of 11 is 1.1. We'll round up to 2. So remove smallest 2 and largest 2 numbers.
- Left with 250, 320, 575, 670, 750, 1000, 2500. Mean is approximately 866.

Notation for the mean:

- If we're computing the mean of the whole population, we denote it by μ, the Greek letter "mu."
- If we're computing the mean of a sample we denote it by $\overline{y}$ or $\overline{x}$ or ...

The book has notation for the median, but you need not know it.

VIII. **Help Room Hours**

- Tuesday, 9:00–11:20, Srinivasan.
- Tuesday, 1:40–4:00, Fang Li.
- Tuesday, 4:00–5:00, Xiang Li.
- Wednesday, 2:50–5:00, DeGroat.
- Thursday, 9:00–11:20, Wang.
- Thursday, 11:20–12:30, Xiang Li.