Welcome to STAT 401, Section F

Instructor: Petrutza Caragea

Monday, August 20-th, 2007
Day One

• Me
• You: on the supplied notecard, please provide
  1. Name
  2. Hometown
  3. Field of Research/Major and Major professor (if known)
  4. Height (in feet/inches)
  5. Operation system most familiar with (Windows, Mac, Linux)
  6. Statistical software you used before: SAS, JMP, R, Splus, SPSS, STATA, Other (specify)
  7. Something fun that you did this summer

• Syllabus
http://www.public.iastate.edu/~pcaragea/stat401

The course web page contains

- Reading assignments
- Homework assignments (assign/collect on Wednesdays)
- Course handouts/lecture notes
- Computer programs and data sets
- Course assistant information
- Final project instructions
- Important updates
We will use two software packages: "SAS" and "R"

- They are used in other STAT courses
- Familiarity with these increases marketability
- R better for graphs and exploratory analyses
- R is free to use for everyone but you need a license to use SAS
- SAS works under Windows, R works under all op. systems
Statistics is an information science

- a tool to draw conclusions from data with \textit{variability}
- scientific study of how to
  - collect data (design) $\leftarrow$ STAT 402
  - summarize data (description) $\leftarrow$ STAT 401
  - draw conclusions from “incomplete” data (inference) $\leftarrow$ STAT 401

$\Rightarrow$ often interested in using a small data \textit{sample} to answer a larger question.

Let’s see an example.
Allied intelligence reports on German production of tanks and other war materials were somewhat unreliable during World War II.

Statisticians set out to improve estimates of German tank production when it was discovered that German Mark V tanks were labeled with consecutive serial numbers.

Capturing a tank was like drawing a number from the sequence \( 1, 2, ..., N \).

The hope was to use the serial numbers from a sample of \( n \) captured tanks to obtain a reliable estimate of \( N \), the total number of Mark V tanks that had been produced by the Germans.
It can be shown that one of the best estimators for $N$ is:

$$\hat{N} = \frac{n + 1}{n} \text{MAX} - 1$$

where $n$ is the size of the sample (i.e. number of captured tanks) and MAX is the largest value in the sample (i.e. the largest serial number recorded).
Did the *statistical answer* get closer to the truth than the intelligence agency estimate?
German Mark V tank estimates

<table>
<thead>
<tr>
<th>Month</th>
<th>Intelligence Agency Estimate</th>
<th>Serial Number Estimate</th>
<th>Actual Number of Tanks</th>
</tr>
</thead>
<tbody>
<tr>
<td>June, 1940</td>
<td>1000</td>
<td>169</td>
<td>122</td>
</tr>
<tr>
<td>June, 1941</td>
<td>1550</td>
<td>244</td>
<td>271</td>
</tr>
<tr>
<td>Sept., 1942</td>
<td>1550</td>
<td>327</td>
<td>342</td>
</tr>
</tbody>
</table>

Example of inference

“Conclusions that patterns in the data are present in some broader context” (page 8, textbook).