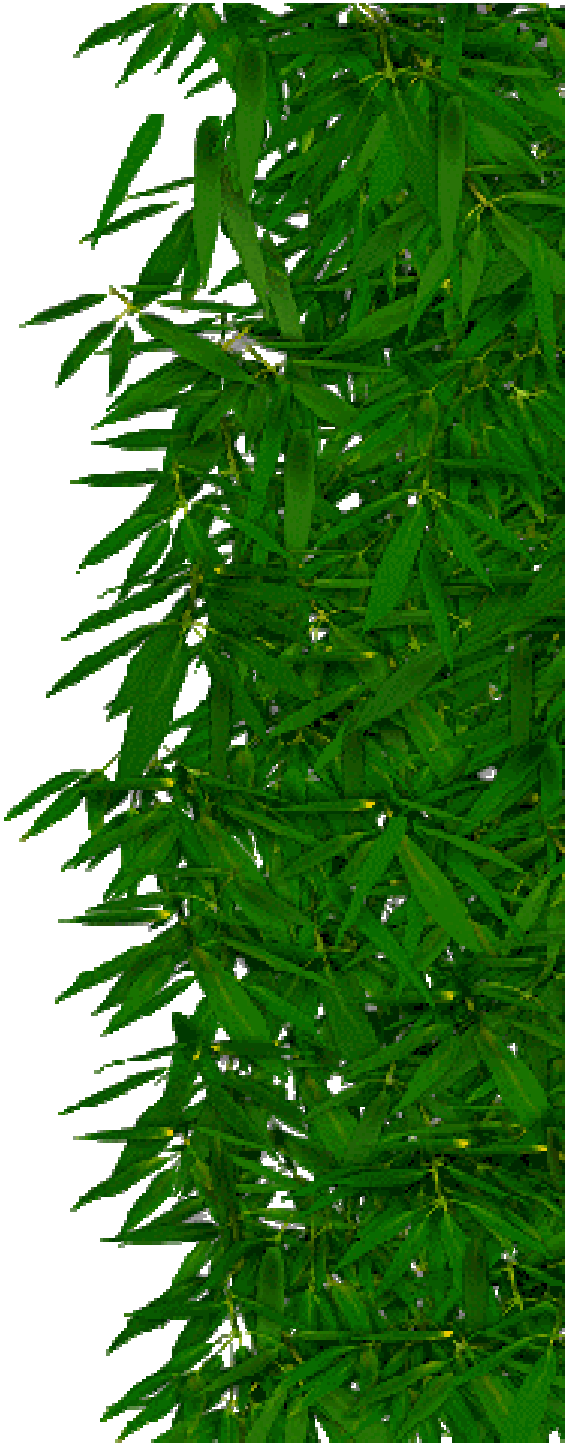
STA103

Probability/Statistical Inference



Jenise's contact info

- Instructor: Jenise Swall
- Office: 221 Old Chem Bldg.
- Phone: 684-4608
- Office hours: Tue. 12:45PM-2:00PM, Wed. 3:30PM-4:45PM
- jenise@stat.duke.edu

TA contact info

Christine Kohnen-Mickelson

- Office: 212 Old Chem Bldg.
- Phone: 684-4365
- Office hours: TBA
- cnk@stat.duke.edu

Merrill Liechty

- Office: 211B Old Chem Bldg.
- Phone: 684-8088
- Office hours: TBA
- merrill@stat.duke.edu

CourseInfo page

- Course web page is primary reference point for schedules, assignments, etc.
- Using CourseInfo system to maintain the site and provide security so students can view their grades online
- You must be enrolled in STA103 to make full use of the CourseInfo page

Using CourseInfo

- First, obtain your CourseInfo userid and password. Detailed instructions are on the "public" STA103 page: www.stat.duke.edu/courses/Fall00/sta103
- Relevant info will be on both pages until add/drop ends, for the convenience of those waitlisted.

Course materials

- Required text: *Mathematical Statistics with Applications* by Wackerly, Mendenhall, Scheaffer (5th edition)
- Optional text: Student solutions manual for the textbook
- Calculator capable of logs, exponentiation, powers, etc. for quizzes/exams

Homework

- Suggested problems (and solutions) will be posted on the web site as we go along
- Intended to help you gauge your progress and review the material
- Will not be graded

Sections

- Supervised by TAs
- Quizzes administered each week
- Some computer exercises will be incorporated, mostly in the last portion of the course (using S-Plus software)

Quizzes

- Quizzes administered weekly, but may be cumulative in nature
- Students must take quizzes in their assigned sections
- Lowest quiz grade will be dropped (can be used as one unexcused absence)

Quiz/exam regrades

- You have 2 weeks after test/quiz date to request a regrade
- Submit a note detailing the nature of the grading error along with the quiz/exam to your TA
- Papers submitted for regrade may be examined in their entirety; either net gain or net loss possible

Absences

- Dean's excuse must be presented to be excused from quizzes or to reschedule exams
- Athletic team schedules, illness, or other less official excuses will not be accepted in place of a Dean's excuse

Overview

- Covers basic statistical ideas needed for further study in econ or quantitative social science
- Fairly mathematical approach requires calculus at the level of MTH31
- Best way to learn the material is through practice – suggested homework problems are crucial

Descriptive statistics

- Statistics that we usually see in the media and other everyday events are *descriptive* statistics
- These are just summaries of data
- They include charts, graphs, summary statistics (mean, standard deviation, etc.), and other such displays

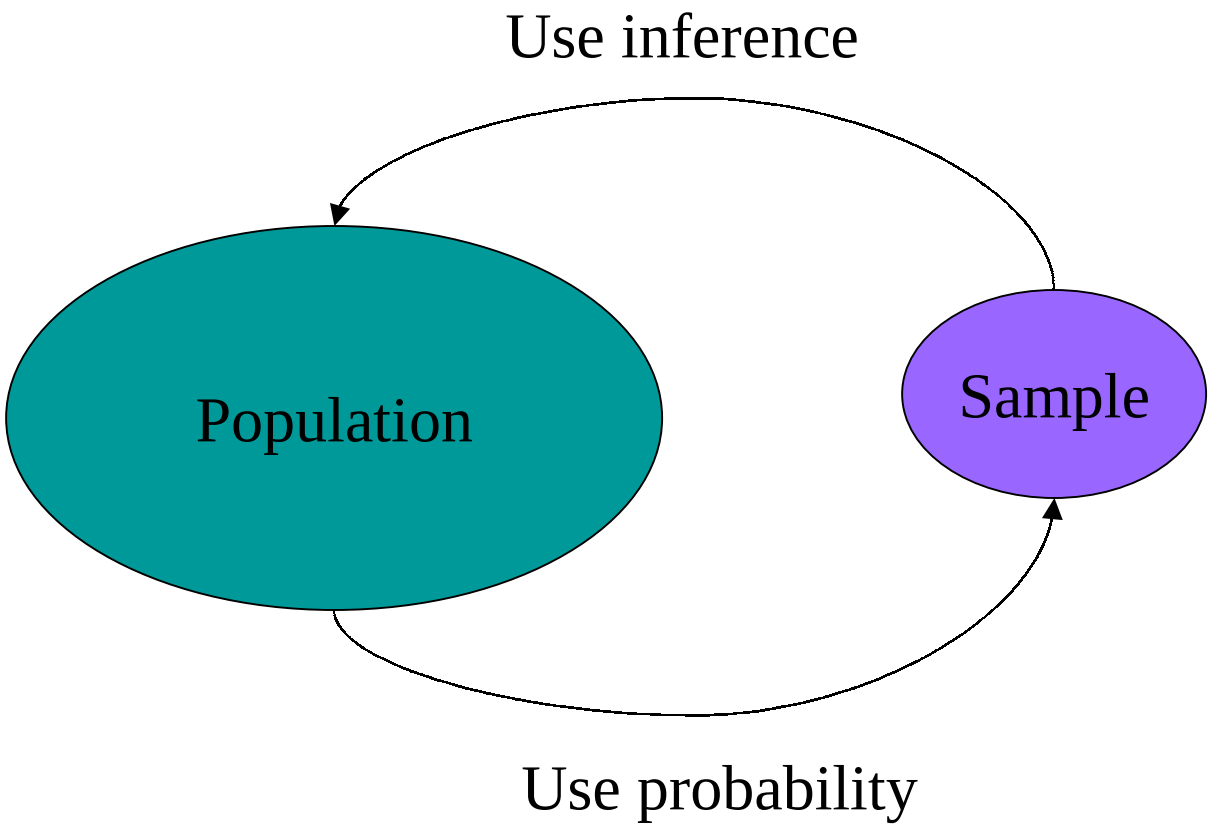
Probability & sampling

- Because the population is large, the sample is a useful way of understanding it
- If we know what the make-up of the population is, then we can calculate the probability of obtaining a certain sample
- Caution must be exercised when choosing a sampling scheme to avoid *bias*

Inference

- Inferences are the conclusions made about the population after considering the sample
- Inferences usually concern quantifiable facts that we are interested in about the population (mean, variance, etc.)
- Since inferences are rarely exactly correct, we also want to estimate how close we can expect ours to be

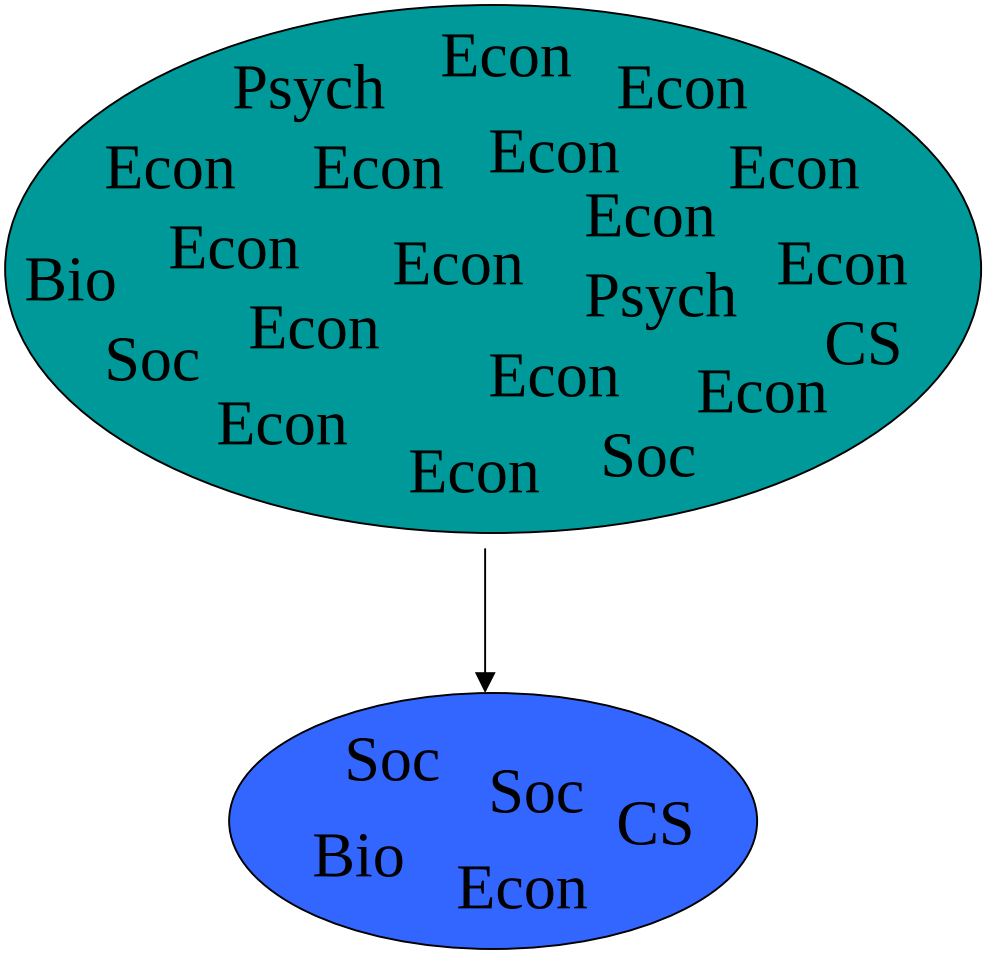
Probability/inference



Simple example

- Question of interest: In At Duke, what percentage of students are econ majors?
- Method 1: Ask each undergraduate about his/her major, tally results, and find the answer
- Method 2: Ask a sample about their majors, tally results, and make an estimate

All undergraduate students



Small sample of undergrads

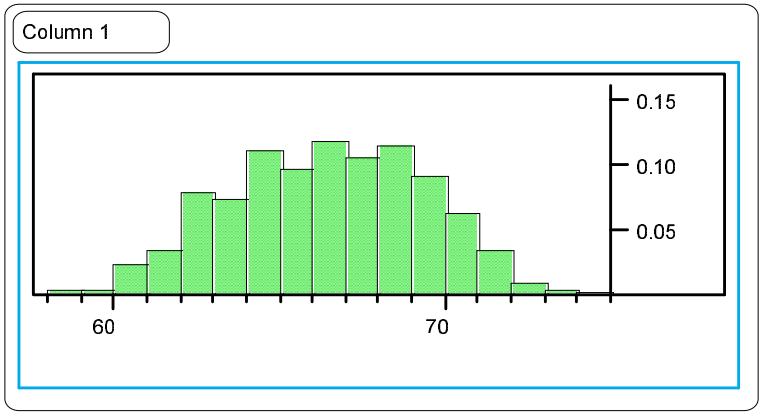
Simple example (cont.)

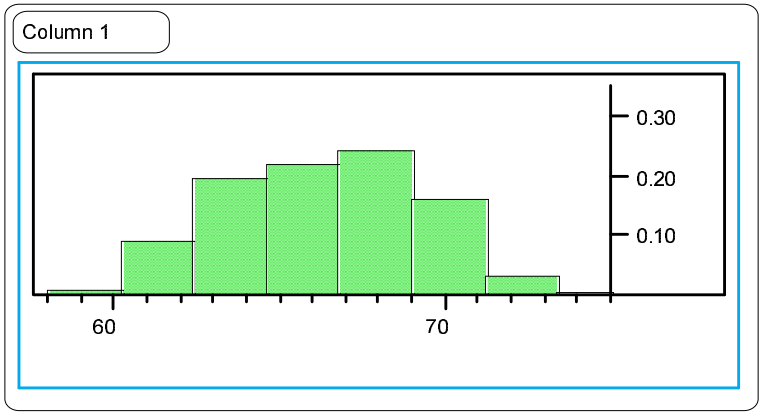
- It is unlikely to get a sample like the one obtained in the previous diagram, but we have to recognize that it's possible
- Probability and inference concepts work together to help us determine whether to rely on our estimates

Graphical summaries

- Command of descriptive stats (both graphical and numerical) needed for further study
- Most common graphical summary for us will be the *histogram*
- Histogram bars have area in proportion to the number of data points that fall in the interval they cover

Histogram intervals





- Note that choice of interval affects the shape of the plot and audience's perception
- Care should be taken to choose intervals in an appropriate way

Numerical descriptions

- Measures of central tendency
  - "Where's the middle of the data?"
  - Two major ones are mean and median
- Measures of dispersion
  - "How much variation is there in the data?"
  - Major ones are variance, standard deviation, and inter-quartile range (IQR)

Mean

- Sum data points and divide by number of data points (the average)
- Provides “balance point” of data
- Easily influenced by outliers
- Easy to calculate and interpret for audiences
- Population mean: μ, sample mean: $\bar{x}$

Median

- Sort data points, median is the middle data point (interpolate between the middle two data points if the number of data points is even)
- Not very susceptible to outliers
- Not as easy to interpret for audiences, often less useful mathematically

Percentiles

- Median: 50th percentile, boundary value separating bottom and top halves of population
- xth percentile separates the bottom x% from the top (100-x)%
- First quartile (Q1): 25th percentile (marks boundary for lower 4th)
- Third quartile (Q3): 75th percentile (marks boundary for upper 4th)

Variance

- Might want to measure dispersion as avg distance of each data point from the mean as $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)$
- But this is always 0, so look at the avg squared distance $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$
- Standard deviation is sq root of variance
- Sample variance differs in minor ways; more about uses for this later

Variance (cont.)

- Pros
  - Mathematically useful
  - Frequently used
  - More understandable to audiences
- Cons
  - Easily influenced by outliers
  - Difficult to calculate by hand

Inter-quartile range

- Also called IQR, is Q3 – Q1
- “How far is the median of the top half from the median of the bottom half?”
- Not as susceptible to outliers as the variance
- Not as commonly used by general audiences