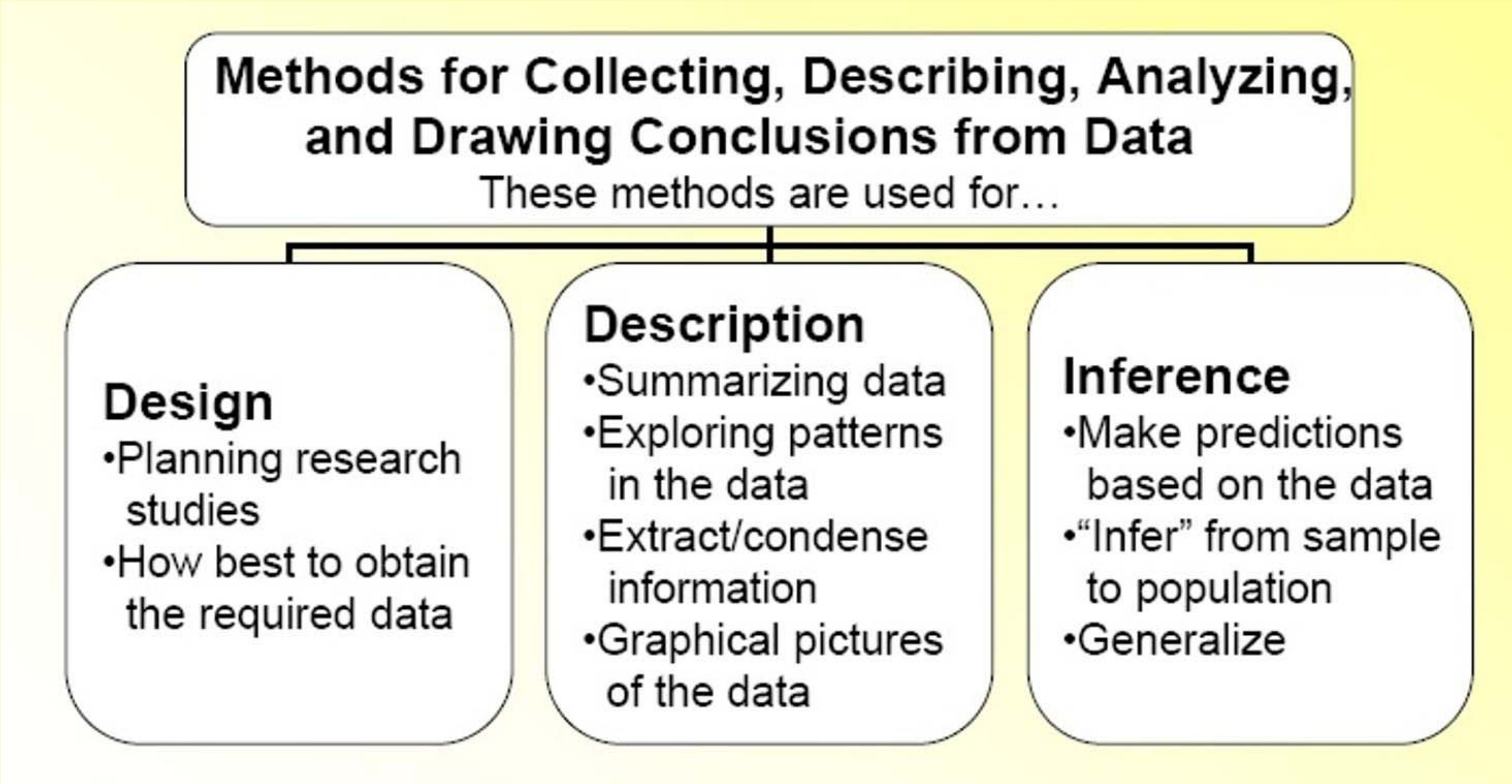
STA 291
Spring 2009

LECTURE 2
TUESDAY,
20 JANUARY

Administrative

- ***Start watching for*** **real** ***homework 1.***
- Suggested problems from the textbook (not graded, but useful as exam preparation): 1.1 – 1.8
- Watch the web page for news about the lab/ recitation sessions
- Tomorrow is last add day—check web page for override policy ☹

What is Statistics?



Methods for Collecting, Describing, Analyzing, and Drawing Conclusions from Data

These methods are used for…

Design

- Planning research studies
- How best to obtain the required data

Description

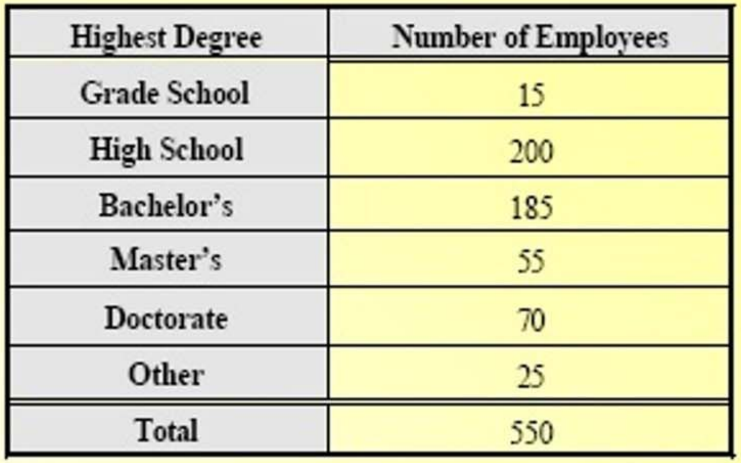
- Summarizing data
- Exploring patterns in the data
- Extract/condense information
- Graphical pictures of the data

Inference

- Make predictions based on the data
- "Infer" from sample to population
- Generalize

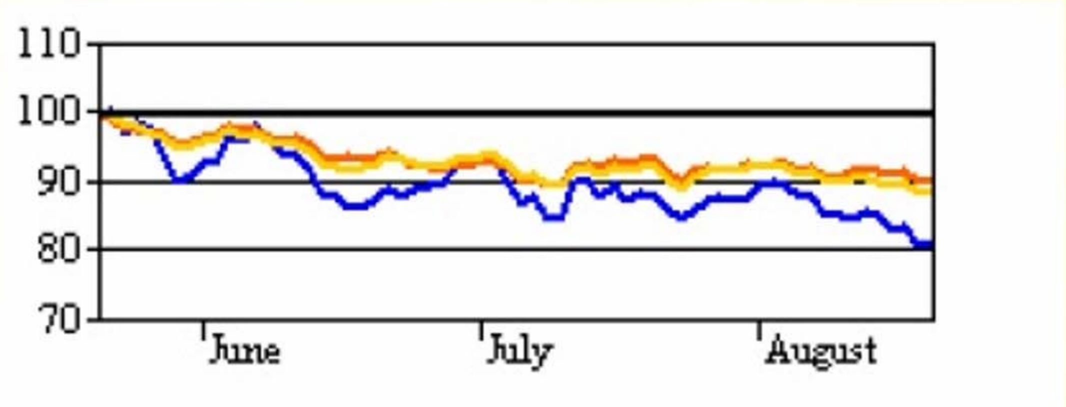
Example Descriptive Statistics

Frequency Distribution



Highest Degree	Number of Employees
Grade School	15
High School	200
Bachelor’s	185
Master’s	55
Doctorate	70
Other	25
Total	550

Time Plot



Basic Terminology One

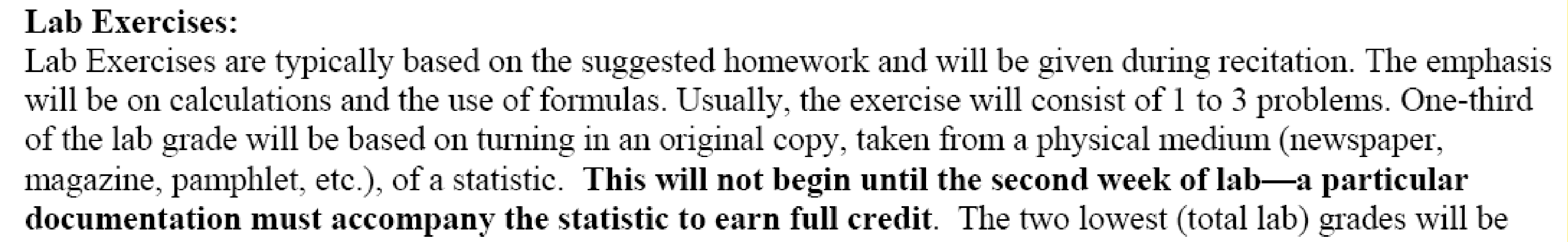
- **Population**
  - total set of all subjects of interest
  - the entire group of people, animal or things about which we want information
- **Elementary Unit**
  - any individual member of the population
- **Sample**
  - subset of the population from which the study actually collects information
  - used to draw conclusions about the whole population

Basic Terminology Two

- **Variable**
  - a characteristic of a unit that can vary among subjects in the population/sample
  - Examples: gender, nationality, age, income, hair color, height, disease status, company rating, grade in STA 291, state of residence
- **Sampling Frame**
  - listing of all the units in the population
- **Parameter**
  - numerical characteristic of the **population**
  - calculated using the whole population
- **Statistic**
  - numerical characteristic of the **sample**
  - calculated using the sample

Statistic of the Week

- From the Syllabus:



Lab Exercises:

Lab Exercises are typically based on the suggested homework and will be given during recitation. The emphasis will be on calculations and the use of formulas. Usually, the exercise will consist of 1 to 3 problems. One-third of the lab grade will be based on turning in an original copy, taken from a physical medium (newspaper, magazine, pamphlet, etc.), of a statistic. **This will not begin until the second week of lab—a particular documentation must accompany the statistic to earn full credit**. The two lowest (total lab) grades will be

- "Particular documentation" must include:
  - Your name
  - Citation (link on web page for info)
  - Population
  - Sample
  - Parameter
  - Statistic

Data Collection and Sampling Theory

Why not measure all of the units in the population? Why not take a **census?**

Problems:

- *Accuracy: May not be able to list them all—* may not be able to come up with a **frame.**
- *Time: Speed of Response*
- *Expense: Cost*
- *Infinite Population*
- *Destructive Sampling or Testing*

Flavors of Statistics

- **Descriptive Statistics**
  - Summarizing the information in a collection of data

- **Inferential Statistics**
  - Using information from a sample to make conclusions/predictions about the population

Example 1

University Health Services at UK conducts a survey about alcohol abuse among students. Two-hundred of the 30,000 students are sampled and asked to complete a questionnaire. One question is “Have you regretted something you did while drinking”?

- What is the population? Sample?

For the 30,000 students, of interest is the percentage who would respond “yes”.

- Is this value a parameter or a statistic?

The percentage who respond “yes” is computed for the students sampled.

- Is this a parameter or a statistic?

Example 2

The Current Population Survey of about 60,000 households in the United States in 2002 distinguishes three types of families: Married-couple (MC), Female householder and no husband (FH), Male householder and no wife (MH).

- It indicated that 5.3% of “MC”, 26.5% of “FH”, and 12.1% of “MH” families have annual income below the poverty level.
- Are these numbers statistics or parameters?

The report says that the percentage of all “FH” families in the USA with income below the poverty level is at least 25.5% but no greater than 27.5%.

- Is this an example of descriptive or inferential statistics?

Modified Example

- A census of all households in Lexington indicated that 6.2% of married couple households in Lexington have annual income below the poverty level.

- Is this number a statistic or a parameter?

Univariate versus Multivariate

- Univariate data set
  - Consists of observations on a single attribute
- Multivariate data
  - Consists of observations on several attributes
- Special case: Bivariate data
  - Two attributes collected per observation

Textbook Section 2.1

Scales of Measurement

– Qualitative and Quantitative

– Nominal and Ordinal

– Discrete and Continuous

Nominal or Ordinal, Difference

- **Nominal: gender, nationality, hair color,** state of residence
- Nominal variables have a **scale of unordered categories**
- It does not make sense to say, for example, that green hair is greater/higher/better than orange hair

Nominal or Ordinal, Difference

- **Ordinal:** Disease status, company rating, grade in STA 291

- Ordinal variables have a scale of ordered categories. They are often treated in a quantitative manner (A=4.0, B=3.0,…)

- One unit can have more of a certain property than does another unit

Nominal or Ordinal, What in Common?

- They’re all categorical and therefore ***qualitative*** variables.

If not Qualitative, then what?

- Then they’re **Quantitative**
- Quantitative variables are measured numerically, that is, for each subject, a number is observed
- The scale for quantitative variables is called **interval scale**

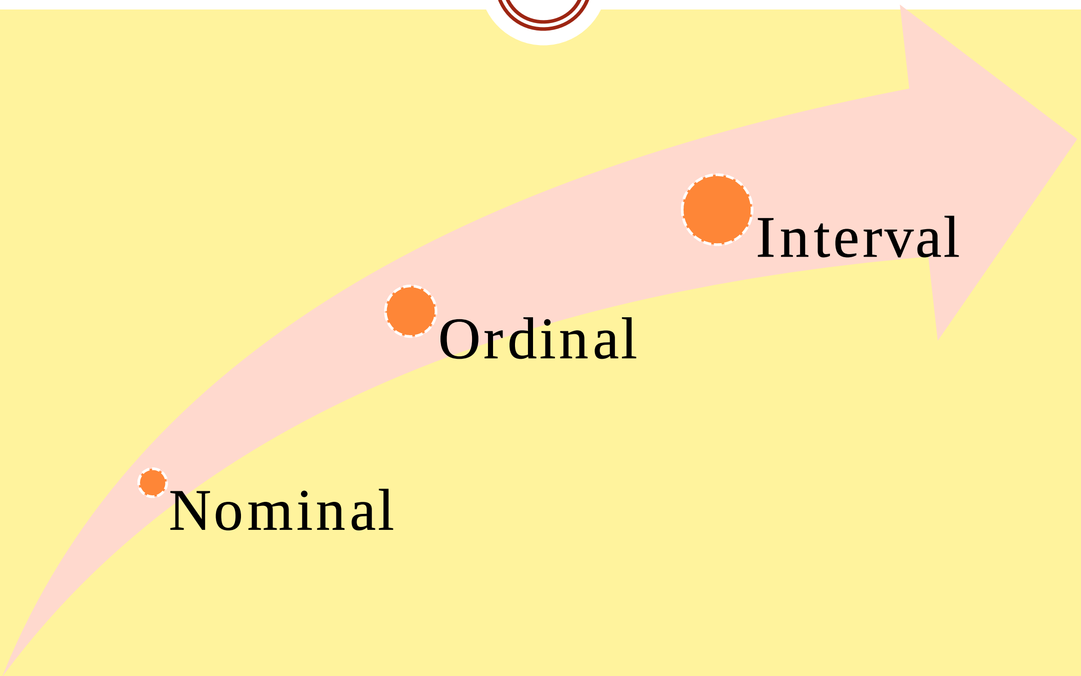
Scale of Measurement Example

The following data are collected on newborns as part of a birth registry database:

- Ethnic background: African-American, Hispanic, Native American, Caucasian, Other
- Infant’s Condition: Excellent, Good, Fair, Poor
- Birthweight: in grams
- Number of prenatal visits

What are the appropriate scales?

Why is it important to distinguish between different types of data?



Some statistical methods only work for quantitative variables, others are designed for qualitative variables. The higher the level, the more information and the better statistical methods we may use.

Discrete versus Continuous

- A variable is **discrete** if it has a finite number of possible values
- ***All*** qualitative (categorical) variables are discrete.
- *Some* quantitative (numeric) variables are discrete—which are not?
- A variable is **continuous** if it can take all the values in a continuum of real values.

Discrete versus Continuous, Which?

- Discrete versus Continuous for quantitative variables:

  - discrete quantitative variables are (almost) always counts

  - continuous quantitative variables are everything else, but are usually physical measures such as time, distance, volume, speed, etc.

Simple Random Sample

- Each possible sample has the same probability of being selected.
- The sample size is usually denoted by *n*.

SRS Example

- Population of 4 students: Adam, Bob, Christina, Dana
- Select a simple random sample (SRS) of size n=2 to ask them about their smoking habits
- 6 possible samples of size n=2:

  (1) A & B, (2) A & C, (3) A & D

  (4) B & C, (5) B & D, (6) C & D

How to choose a SRS?

- Old way: use a random number table.

- A little more modern: http://www.randomizer.org
  (first lab exercise)

How **not to choose a SRS?**

- Ask Adam and Dana because they are in your office anyway
  - “convenience sample”
- Ask who wants to take part in the survey and take the first two who volunteer
  - “volunteer sampling”

Problems with Volunteer Samples

- The sample will poorly represent the population
- Misleading conclusions
- BIAS
- Examples: Mall interview, Street corner interview

Attendance Survey Question #2

- On an index card
  - Please write down your name and section number
  - Today’s Questions:

1. What were the main points of today’s lecture?

2. What was *least* clear about today’s lecture?