

STA 321

Spring 2015

Lecture 1

Thursday, January 15

WELCOME to STA 321!

- Syllabus: on the course website
<http://polytopes.net/courses/STA321S15/>
- Class Instructor: Dr. Ruriko Yoshida
- ruriko.yoshida@uky.edu
- "Probability and Statistics" (4th Edition) by Morris H. DeGroot and Mark J. Schervish.
- Office Hours: 10:00-10:50 on Tues/Thurs
- 325D Multi-disciplinary Sciences Building
- Note: Course material on the course website—usually not handouts in class

Course Information

- Lecture in *CB 233*
Tuesday & Thursday, 14:00 – 15:15
- Grading
There will be two in-class exams, graded homeworks, and a final exam: these will count 40%, 30%, and 30% of your grade, respectively.

Topics

- Hypothesis testing (Ch 9)
- Analysis of categorical data (Ch 10)
- Regression and correlation (Ch 11)
- Analysis of variance (Ch 11)
- Nonparametric methods (Ch 10)

Three Important Concepts

- Sampling Distribution (Ch 8)
- Confidence Interval (Ch 8)
- P-value of a Statistical Test (Ch 9)

Grade Calculation

- Midterm Exams (Feb 26/April 23)..... **40%**
- Final Exam (May 7)..... **30%**
- Weekly Homework Assignments.... **30%**

Letter Grades

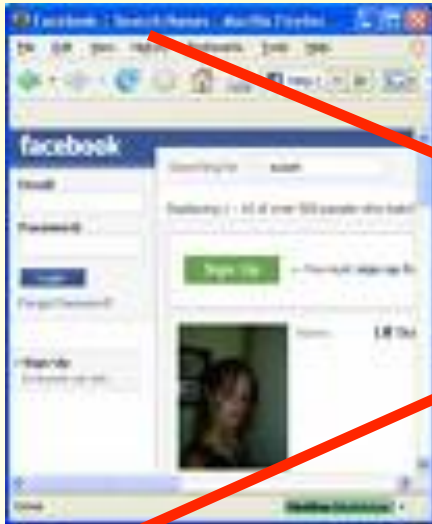
A: 90-100%, **B:** 80-89%, **C:** 70-79%,

D: 60-69%, **E:** 0-59%

Please...



Also Please...



Why Statistics?

- **Research** in the sciences is getting more quantitative (look at research journals)
- Computers make even complex statistical methods easier to use
 - danger of using inappropriate methods
 - vital to understand a method before using it
- Job market: Most graduates need to be familiar with basic statistical methodology
- *“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.”*
Herbert George Wells (1866–1946)

Why Statistics? (contd)

- **Everyday** Newspaper, advertising, surveys, internet and other media:
Many statements contain statistical arguments or techniques
- Recent examples...

<http://www.bls.gov/news.release/empsit.t01.htm>

What does it take to understand the STA 321 material?

- Logical thinking
- Perseverance
- ...+ see Syllabus
(attend lectures and labs, obtain material when absent, do homework yourself, etc.)
- Don't procrastinate 😊

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

Descriptive Statistics, e.g.

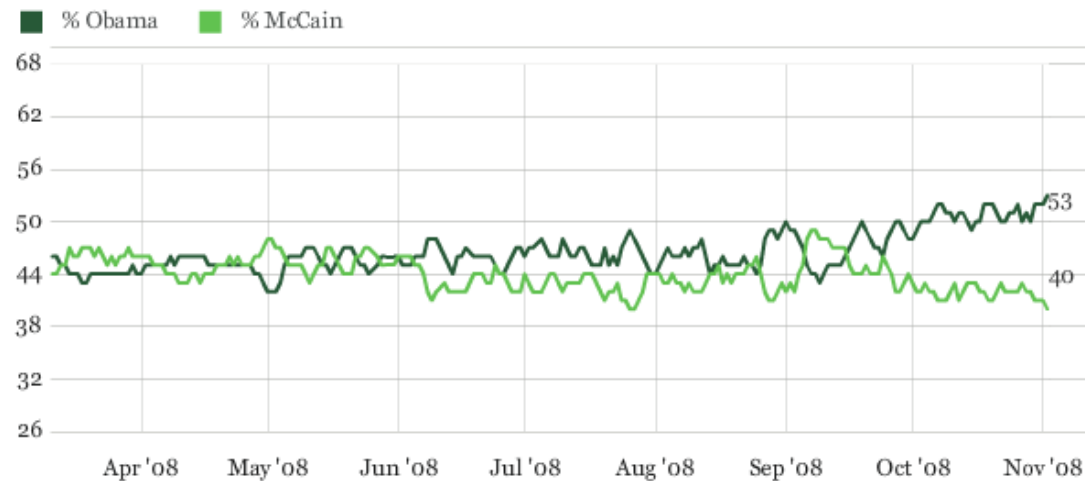
Frequency Distribution

STA 321 Grade	Frequency
A	31
B	37
C	10
E	2

Time Plot

Gallup Daily: Election 2008

Results based on a five-day rolling average through June 8 and a three-day average since June 9



Gallup Daily election tracking reports the percentage of registered voters who say they would support each candidate if the presidential election were held today.

Basic Terminology I

- **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 II

- **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 321, state of residence
- **Sampling Frame**
 - listing of all the units in the population
- **Parameter**
 - numerical characteristic of the **p**opulation
 - calculated using the whole population
- **Statistic**
 - numerical characteristic of the **s**ample
 - calculated using the sample

Statistic vs. Parameter

- **Statistics are based on a sample**
(even if they are used to describe a population)
- **Parameters are calculated using the whole population**

Data Collection and Sampling Theory

Why not measure all of the 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.
- 200 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”.

This value is computed for the students sampled.
Is this a parameter or a statistic?

Example 2

- Polls in the United Kingdom, France, and Germany indicated that (Gallup, July 23, 2008) majorities of citizens in these countries preferred Obama over McCain.
- In the U.K., 60% of those surveyed favored Obama, 15% favored McCain, and 25% did not know or refused to answer this question.
- France: 64%, 4%, 32%
- Germany: 62%, 10%, 27%
- Are these numbers statistics or parameters?
- The report says that the percentage of all adults in the U.K. who favored Obama was at least 57%, but no greater than 63%.
- Is this an example of descriptive or inferential statistics?

Univariate vs 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

Scales of Measurement

- Qualitative and Quantitative
 - Nominal and Ordinal
 - Discrete and Continuous
-
- **Recall:**
 - A **Variable** is a characteristic of a unit that can vary among subjects in the population/sample

Qualitative Variables (=Categorical Variables) Nominal or Ordinal

- **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

Qualitative (Categorical) Variables

Nominal or Ordinal

- **Ordinal:** Disease status, company rating, grade in STA 321
- 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

Quantitative Variables

- **Quantitative:** age, income, height
- Quantitative variables are measured numerically, that is, for each subject, a number is observed
- The scale for quantitative variables is called **interval scale**

Example 1

- Vigild “Oral hygiene and periodontal conditions among 201 institutionalized elderly”,
Gerodontology, 4:140-145
- Variables measured
 - Nominal: Requires Assistance from Staff?
Yes / No
 - Ordinal: Plaque Score
No Visible Plaque - Small Amounts of Plaque -
Moderate Amounts of Plaque - Abundant Plaque
 - Interval: Number of Teeth

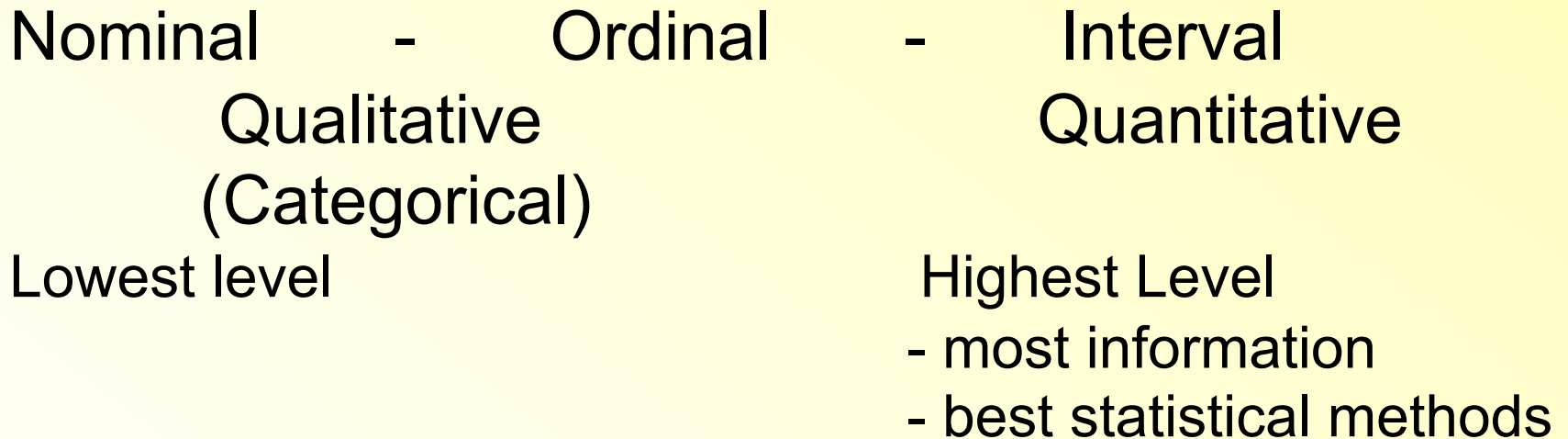
Example 2

- 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.



You **can not** use statistical methods for quantitative data to analyze qualitative data.

You **can** treat variables in a less quantitative manner.

- Example.

- Height: Quantitative variable, interval scale,
measured in cm (or ft/in)

- Can be treated as ordinal
short, average, tall

- Can even be treated as nominal
180cm-200cm, all others

- Try to measure variables at the highest possible level
- Higher-level variables can be analyzed with a greater variety of methods

Caution: Sometimes, ordinal variables are treated as quantitative