SOC 3811

BASIC SOCIAL STATISTICS

Professor:	David Knoke
Teaching Assistants:	Kyungmin Baek
	Yu-Ju Chien
	Jasmine Harris

Department of Sociology

SPRING 2011

Chapter 1
The Social Research Process

1.1 Ideas into Research Projects

1.2 Concepts in Propositions

1.3 Variables into Hypotheses

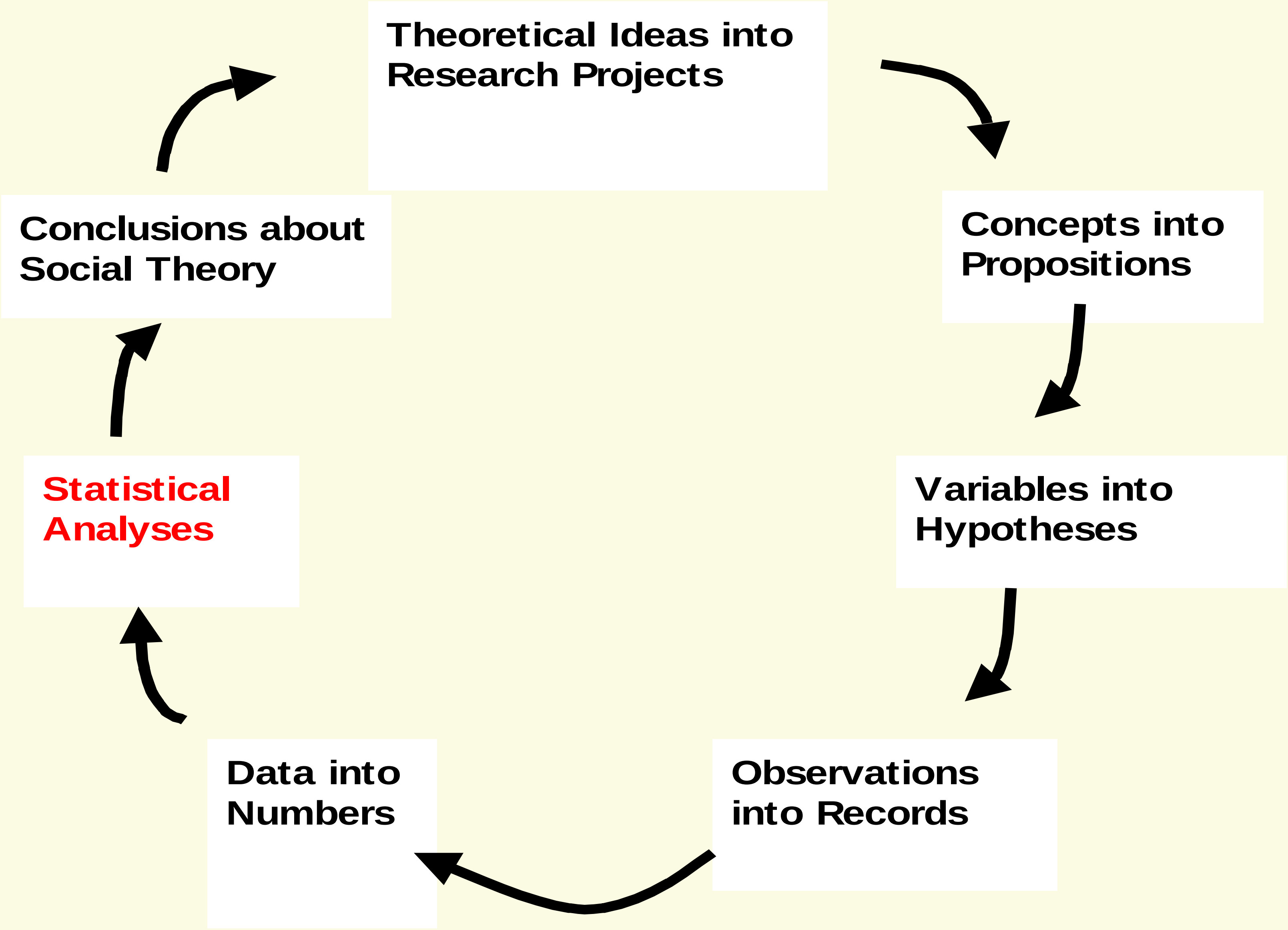
1.4 Observations into Records

1.5 Data into Numbers

1.6 Statistical Analysis

1.7 The General Linear Model

The Research Cycle



Theoretical Ideas into Research Projects

Concepts into Propositions

Variables into Hypotheses

Observations into Records

Data into Numbers

Statistical Analyses

Conclusions about Social Theory

TYPES OF VARIABLES

Variable: Any characteristic or attribute of persons, objects, or events that can take on different numerical values

To test a research hypothesis, choose a test statistic that is appropriate for a specific type of variable

A Typology of Variables	
Nonorderable Variables	**Orderable Variables**
Discrete	**Discrete**
Polychotomous (many categories)	Polychotomous (many categories)
Dichotomous (2 categories)	Dichotomous (2 categories)
	Continuous

No nonorderable continuous variables!

Discrete Variables

Discrete variable: classifies persons, objects, or events according to the kind or quality of their attributes

Discrete variables may have many categories (polychotomous) or only two (dichotomous)

Nonorderable discrete: the sequence of categories *cannot* be meaningfully ordered

Eye color: black, blue, brown, green, grey, purple, …
Nationality: Indonesian, Iraki, Iranian, Japanese, Kenyan, …
Music: blues, classical, country, hip hop, jazz, rap, rock, …

Orderable discrete: the sequence of categories *can* be meaningfully ordered (numbers show low-high sequence)

Life-stage: infant, toddler, child, adolescent, young adult, ...
Social class: lower, working, middle, upper
Company size: small, medium, large, very large
Attitude: strongly disagree, disagree, agree, strongly agree

Dichotomous variable: a discrete measure with two categories that may or may not be ordered

Which of these dichotomies are ordered? Why?

Gender: Female/Male

Wealth: Poor/Rich

Vote: McCain/Obama

Age: Young/Old

Height: Small/Large

Mascot: Gopher/Hawkeye

Education: Noncollege/College

Continuous variable: a variable that, in theory, can take on *all* possible numerical values in a given interval

Ideally, precise intervals (distances) should be measured, as in the natural sciences. In practice, social variables usually allow only a limited number of values on a underlying continuous scale.

How many total categories for each continuous variable?

Education: 0, 1, 2, 3, …. 20 years of schooling

Age: 1, 2, 3, 4, 5, 6, …. 125 years old

Annual Income: <$500; $501-999; $1,000-1,999; $2,000-2,999, …

Pres. Obama's job rating: Poor, good, fair, excellent

Attitude: Strongly disagree, Disagree, Agree, Strongly agree

Reasonable people may disagree on whether the last two examples are discrete orderable or continuous variables. How about you?

We should feel comfortable in treating a variable as continuous only if a statistical test we apply is robust: it's insensitive to small departures from assumptions on which it depends; e.g., continuous measurement

Identify These Variables' Types

Classify the following. Keep in mind that a variable's type may be ambiguous, reflecting imprecision of social measurements:

Class Role: **(Instructor, TA, Student)** ______________________

Desserts Eaten: **(None, One, Two, ….)** ______________________

Social Class: **(Lower, Working, Middle, Upper)** ______________________

Textbook Price: **(dollars and cents)** ______________________

Religion: **(Christian, Nonchristian)** ______________________

Marital Status: **(Single, Married, Widowed)** ______________________

Residential Area: **(ZIP code)** ______________________

Population growth: **(Down, Unchanged, Up)** ______________________

Workplace Accidents: **(Number of injuries)** ______________________

Industrialization: **(Pre-industrial, Industrial)** ______________________

Descriptive vs. Inferential Stats

The field of statistics breaks down into two broad categories – descriptive and inferential statistics

Descriptive statistics are concerned with summarizing the properties of a <u>sample</u> of observations

The Gallup Poll's final survey of 2010 found that 48% of the 1,500 respondents said they approved how Pres. Obama was doing his job, with a "margin of sampling error" = ±3 per cent.

The percentage of respondents choosing a response is a descriptive sample statistic. But, why should we care about the opinions of those 1,500 unknown people, from an adult population of more than 240 million persons?

Inferential statistics apply the mathematical theory of probability to make decisions about the likely properties of populations based on the sample evidence

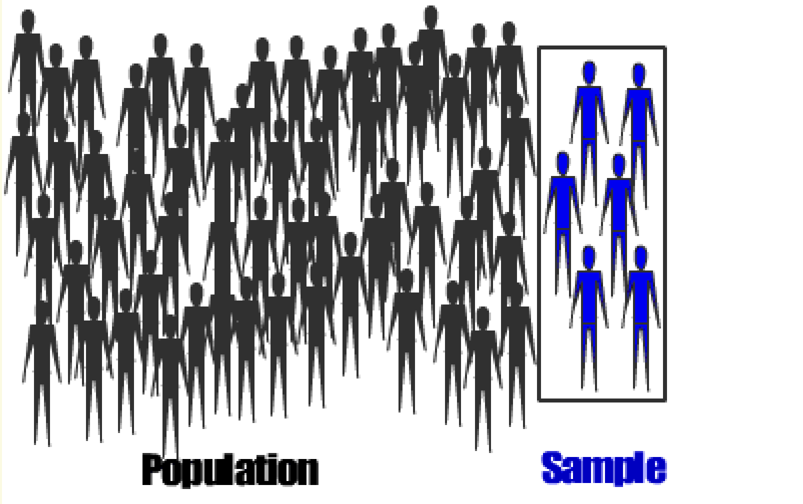
The *Gallup Poll* stated that the "margin of error is ±3%" (plus or minus three percent) around the respondents' responses to the item. What does this phrase mean?

The t-test, an inferential statistic that you will learn, allows us to infer (make conclusions or generalize) about the *probable* value of the population parameter.

The American public had a 95% "confidence interval," from 45% to 51%, in approving Pres. Obama's job.

Statistical Significance

A **statistical significance test** allows us to make a statement about the probability that a population with a hypothesized parameter value could have produced the observed value of a sample statistic.



When **random sampling** assures us of a sample that highly represents the population, then we can make inferences about likely population parameters with a high level of confidence (but not with complete certainty).

During this course, you will learn how to combine descriptive and inferential statistics to test the truth-value of research hypotheses about population parameters, using statistics based on sample data.

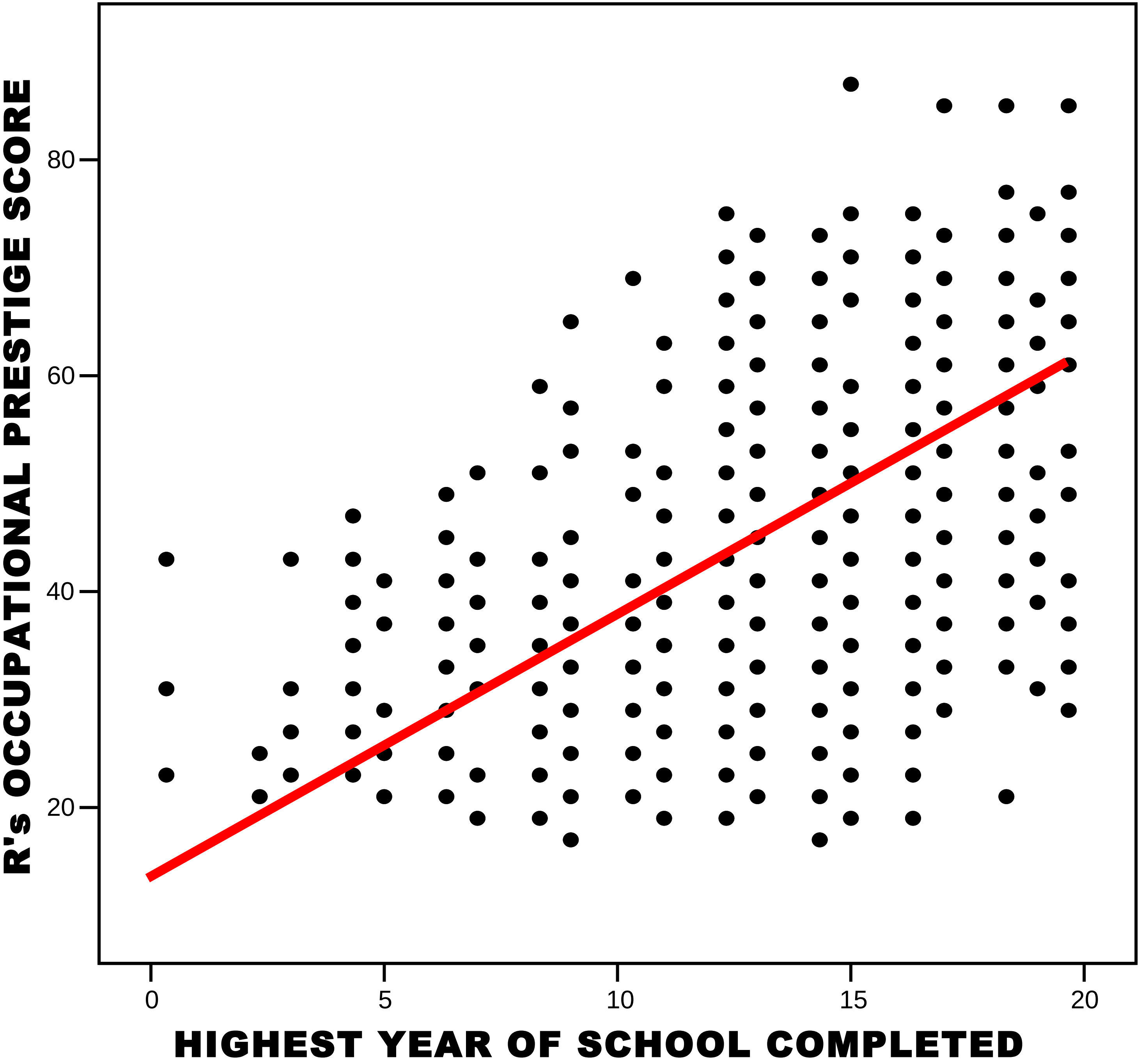
We’ll learn why an apparently small Gallup sample allows us to be very confident that Obama’s true rating in 2009 was probably closer to 50% or to 52%, than it was to 47% or 55%. (And also that the most probable parameter value was the Gallup Poll’s point-estimate sample statistic: 51%)

The General Linear Model

General linear model: assumes the relationships among independent and dependent measures basically vary according to straight line patterns

The regression line in the figure on next slide shows that, as respondents' years of formal education increase, expected occupational prestige scores increase in a linear pattern.

We'll learn how to program the computer to calculate and graph the "best-fitting line" through a scatterplot showing the relationship between two variables.


R's OCCUPATIONAL PRESTIGE SCORE
80
60
40
20
0
5
10
15
20
HIGHEST YEAR OF SCHOOL COMPLETED

GSS & SPSS

General Social Surveys annual/biennial samples of U.S. adults interviewed since 1972 on diverse social, economic, political topics. An online codebook with question wordings & response categories is available at:

http://www.norc.org/GSS+Website/

Many examples in this course come from statistical analyses of the 2008 GSS and earlier GSS surveys.

SPSS is a computer software package for calculating descriptive and inferential statistics. We'll use it to analyze GSS data files whose variables have mnemonic names.