Stats 11 (Fall 2004) Lecture Note Instructor: Hongquan Xu
Introduction to Statistical Methods for Business and Economics

Chapter 1: What is Statistics?

- Statistics is the science of DATA.
- Statistics deals with the collection, organization and interpretation of data.
- **Course Goal:** Learn various tools for using data to gain understanding and make sound decisions.

The Subject of Statistics (Section 1.4)

Statistics is concerned with the process of finding out about the world and how it operates in the face of variation and uncertainty by collecting and then making sense of data.

The investigative process

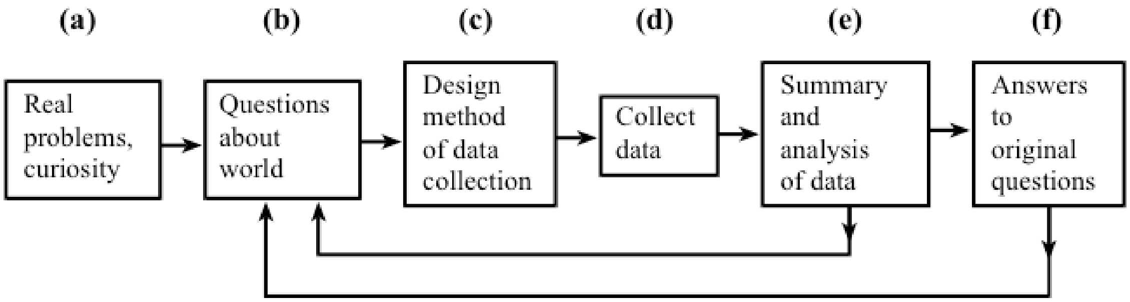

(a)
(b)
(c)
(d)
(e)
(f)
Real problems, curiosity
Questions about world
Design method of data collection
Collect data
Summary and analysis of data
Answers to original questions

Figure 1.4.1 The investigative process.

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

(a) Real problem: Who will be the next US President?

(b) Question: Who do you vote for the President: Bush or Kerry?

(c) Design method of data collection:

(d) Collect data:

(e) Summary and analysis of data:

(f) Answers to original questions:

Example: US Presidential Race 2004

An in-class poll: Bush _______________ Kerry _______________

	Bush	Kerry
In-class poll		
ABC/Post		
CBS/NYTimes		
CNN/USA/Gallup		

Which one do you trust? Why?

Three fundamental kinds of statistical investigations

- sampling or surveys (Section 1.1)
- experiments (Section 1.2)
- observational studies (Section 1.3)

Section 1.1 Polls and Surveys

Surveys study part to gain information about the whole.

Jargon describing surveys

- Target population: complete set of individuals, objects, or units that we want to information about.
- Study population: complete set of units that might possibly be included in the study.
- Sample: subset of units about which we actually collect info.
- Census: attempt to study every individual in the entire population.
- Variable: a characteristic of each unit that we measure.
- Parameter: a numerical characteristic of the population
- Statistic: a numerical characteristic of the sample.

A census often takes a long time and is expensive.
A carefully conducted survey is often more ________________ than a census.

Often we don't know the population parameter and want to ________________ it.
We take a sample, compute the sample statistic and use it to ________________ the population parameter.

Random sampling

Draw units from the study population at random, e.g., using a lottery. Every subject in the study population has some chance to be selected.

- simple random sample (SRS): every individual has equal chance to be selected.
- cluster sampling: a cluster of individuals are used as a sampling unit, rather than individuals.

- stratified sampling: divide the population into **strata** (groups of similar individuals) and choose a SRS from each strata.
- multi-stage sampling: a combination of SRS, cluster sampling and stratified sampling.

Why do we sample at random?

- To avoid subjective and other ________________ .
- To allow the calculation of ________________ (i.e., the likely size of the error in our sample estimates).

Sources of error in surveys

- Random sampling leads to **sampling errors**.
- **Nonsampling errors** can be much larger than the sampling errors.

The chance errors in survey estimates are generally smaller if we take ________________ samples.

Sources of nonsampling errors

- Selection bias: Arises when the population sampled is not exactly the population of interest.
- Self-selection: People themselves decide whether or not to be surveyed. Results akin to severe non-response.
- Non-response bias: Non-respondents often behave or think differently from respondents. Low response rates can lead to huge biases.
- Question wording effects: Even slight differences in question wording can produce measurable differences in how people respond.
- Interviewer effects: Different interviewers asking the same questions can tend to obtain different answers.
- Survey format effects: Factors such as question order, questionnaire layout, self-administered questionnaire or interviewer, can effect the results.

Dealing with errors

- Statistical methods are available for estimating the likely size of ________________ errors.
- All we can do with nonsampling errors is to try to minimize them at the ________________ stage.
- **Pilot survey**: One tests a survey on a relatively small group of people to try to identify any problems with the survey design before conducting the survey properly.

Some questions for thinking:

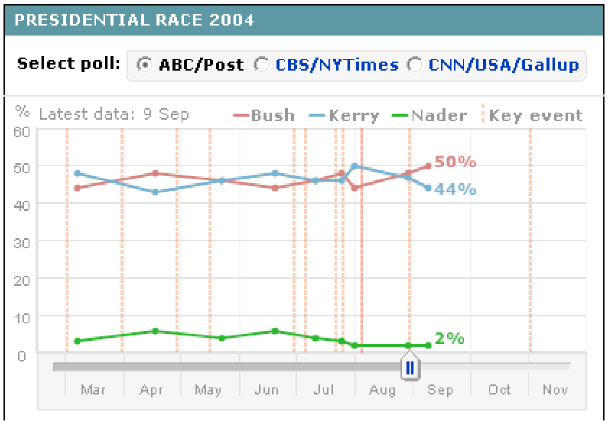
Q: Why are many magazine surveys suspect?

A:

Q: Is interviewing people at random on the street a good survey? Why?

A:

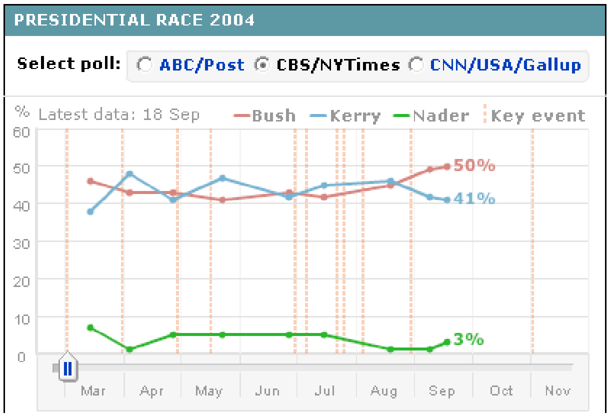
US Presidential Race 2004



PRESIDENTIAL RACE 2004

Select poll: ● ABC/Post ○ CBS/NYTimes ○ CNN/USA/Gallup

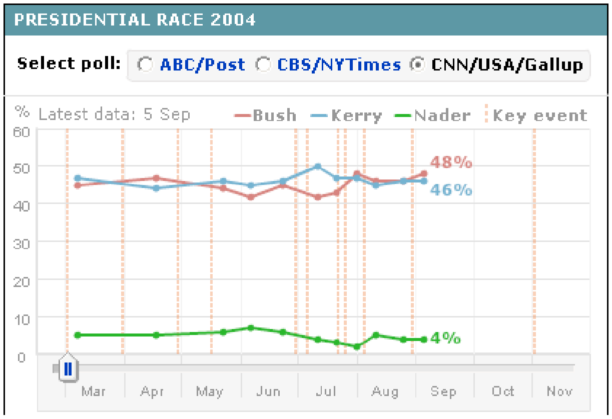
Latest data: 9 Sep



PRESIDENTIAL RACE 2004

Select poll: ○ ABC/Post ● CBS/NYTimes ○ CNN/USA/Gallup

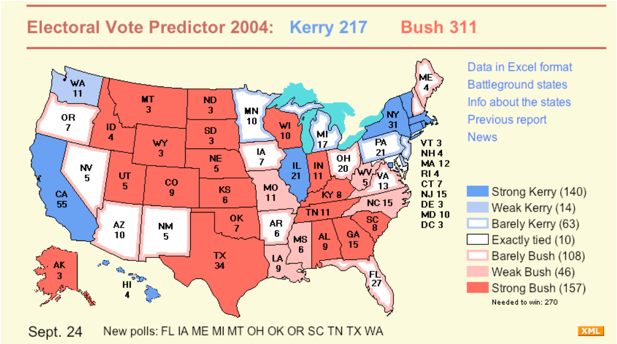
Latest data: 18 Sep



PRESIDENTIAL RACE 2004

Select poll: ○ ABC/Post ○ CBS/NYTimes ● CNN/USA/Gallup

Latest data: 5 Sep



Electoral Vote Predictor 2004: Kerry 217 Bush 311

Sept. 24 New polls: FL IA ME MI MT OH OK OR SC TN TX WA

Source: http://news.bbc.co.uk/ and http://www.electoral-vote.com/

Which one do you trust? Why?

On September 17, 2004, the CNN/USA/Gallup Poll shows Bush (55%) and Kerry (42%) amongst likely voters. Read an interesting article about this poll:
“Why You Should Ignore The Gallup Poll This Morning – And Maybe Other Gallup Polls As Well?”
at http://www.theleftcoaster.com/archives/002806.html

Section 1.2 Experimentation

Experiments deliberately impose some treatment on individuals in order to observe their response.

Principles of Experimental Design

- **Blocking:** group or match homogenous experimental units.
- **Randomization:** use chance to assign experimental units to treatments.
- **Replication:** apply treatments to many experimental units.

Blocking ensures ________________ comparisons with respect to factors known to be important. Randomization tries to obtain comparability with respect to ________________ factors and allows the calculation of estimation error.

Example: Typing efficiency of two keyboards: A and B. Experiment: Six different manuscripts are randomly assigned to numbers 1–6 and given to the same typist. Each manuscript is first typed on keyboard A then on keyboard B in the following order:

1. A, B, 2. A, B, 3. A, B, 4. A, B, 5. A, B, 6. A, B.

How are the principles used?

- Blocking:
- Randomization:
- Replication:

What are the strength and weakness of this design?

Consider another design:

1. A, B, 2. B, A, 3. A, B, 4. B, A, 5. A, B, 6. A, B.

Any weakness?

Consider a third design:

1. A, B, 2. A, B, 3. A, B, 4. B, A, 5. B, A, 6. B, A.

Any weakness?

What design do you recommend?

A good advice is: ________________ **what you can and** ________________ **what you cannot.**

Jargon describing experiments

- Control group: group of experimental units that is given no treatment. Treatment effect is estimated by comparing each treatment group with control group
- Blinding: Preventing people involved in experiment from knowing which experimental subjects have received which treatment. One may be able to blind subjects themselves, people administering the treatments, people measuring the results.
- Double blind: Both the subjects and those administering the treatments have been blinded.
- Placebo: An inert dummy treatment.
- Placebo effect: Response caused in human subjects by the idea that they are being treated

Section 1.3 Observational Studies

An observational study does not attempt to influence the response.

Essential differences between observational studies and experiments:

- In an experiment, the experimenter ________________ which subjects (experimental units) receive which treatments
- In an observational study, we simply compare subjects that ________________ have received each of the treatments.
- Observational studies widely used for identifying possible causes of effects but ________________ reliably establish causation.
- Only properly designed and executed experiments can reliably demonstrate ________________ .

Example: Is use of cellular phone associated with risk of brain cancer?

Study: A group of 469 people who have brain cancer. For each cancer patient, match a person of the same sex, age, and race who do not have brain cancer. Then ask them about the use of cell phones.

Observational Study or Experiment?

Example: Does exercise raise metabolic rate for as long as 12 hours?

Study: Subjects were asked to walk briskly on treadmill for several hours. Their metabolic rates were measured before, immediately after, and 12 hours after the exercise.

Observational Study or Experiment?

Section 1.5 Summary. Read it!