

DATA 606: Statistical Methods in Data Science

— Introduction of Survey

Wenjun Jiang

Department of Mathematics & Statistics
The University of Calgary

Lecture 1



UNIVERSITY OF
CALGARY

Course Introduction

- ▶ Time: Mon & Wed, 17:00–19:45. Location: ICT 517
- ▶ References:
 - *Sampling: Design and Analysis*, 2nd edition, Sharon Lohr, Duxbury press.
 - *Categorical Data Analysis*, 3rd edition, Alan Argenti, Wiley.
- ▶ Course site: [DATA 606 L01-\(Winter 2020\)-Statistical Methods in Data Science](#).

Instructor info

- ▶ [Personal website](#)
- ▶ Office: MS 544
- ▶ Email address: wenjun.jiang@ucalgary.ca
- ▶ Office hours: Tue & Thu, 14:00-15:30 or by appointment.

Note: if you email me for any inquiries regarding this course, please include DATA 606 in the title.

Grading

Components	Weighting	Date
Assignment 1	15%	Jan 16, 2020
Assignment 2	15%	Jan 30, 2020
Quiz 1	10%	Jan 22, 2020
Quiz 2	10%	Feb 5, 2020
Project & Presentation	50%	

Intro of survey

A sample controversy

In Shere Hite's book *Women and Love: A Cultured Revolution in Progress*, there are some results:

- ▶ 84% of women are “not satisfied emotionally with their relationships”.
- ▶ 95% of women “report forms of emotional and psychological harassment from men with whom they are in love relationships”.
- ▶ 84% of women report forms of condescension from the men in their love relationships.

Intro of survey

A sample controversy

In Shere Hite's book *Women and Love: A Cultured Revolution in Progress*, there are some results:

- ▶ 84% of women are “not satisfied emotionally with their relationships”.
- ▶ 95% of women “report forms of emotional and psychological harassment from men with whom they are in love relationships”.
- ▶ 84% of women report forms of condescension from the men in their love relationships.

How much do you believe these results?

Intro of survey

A sample controversy

In Shere Hite's book *Women and Love: A Cultured Revolution in Progress*, there are some results:

- ▶ 84% of women are “not satisfied emotionally with their relationships”.
- ▶ 95% of women “report forms of emotional and psychological harassment from men with whom they are in love relationships”.
- ▶ 84% of women report forms of condescension from the men in their love relationships.

How much do you believe these results?

The impact: the book was widely criticized in newspaper and magazine articles throughout the US, saying these results are “dubious” and “of limited value”.

Intro of survey

Unsuitable characteristics

- ▶ The sample was self-selected—that is, recipients of questionnaires decided whether they would be in the sample or not. Hite mailed 100,000 questionnaires; of these, 4.5% were returned (there are huge non-responses).
- ▶ The questionnaires were mailed to such organizations as professional women's groups, counseling centers, church societies, and senior citizens' centers.
- ▶ The survey has 127 essay questions, and most of the questions have several parts. [Who will tend to return such a survey?](#)

Intro of survey

Unsuitable characteristics (cont.)

- ▶ Many of the questions are vague, using words such as “love.”
 - The concept of love probably has as many interpretations as there are people, making it impossible to attach a single interpretation to any statistic purporting to state how many women are “in love.”
- ▶ Many of the questions are leading—they suggest to the respondent which response she should make.
 - For instance: “Does your husband/lover see you as an equal? Or are there times when he seems to treat you as an inferior? Leave you out of the decisions? Act superior?”

Intro of survey

Conclusion: the final sample is not representative of women in the United States, and the statistics can only be used to describe women who would have responded to the survey.

Intro of survey

Conclusion: the final sample is not representative of women in the United States, and the statistics can only be used to describe women who would have responded to the survey.

Question: what sample is good? Any requirements?

Intro of survey

Conclusion: the final sample is not representative of women in the United States, and the statistics can only be used to describe women who would have responded to the survey.


Question: what sample is good? Any requirements?

Principle: a good sample should be *representative* in the sense that characteristics of interest in the population can be estimated from the sample with a known degree of accuracy.

Intro of survey

Some notions

- ▶ *Observation unit*: An object on which a measurement is taken.
- ▶ *Target population*: The complete collection of observations we want to study.
- ▶ *Sample*: A subset of a population.
- ▶ *Sampled population*: The population from which the sample was taken¹.
- ▶ *Sampling unit*: A unit that can be selected for a sample.
- ▶ *Sampling frame*: A list, map, or other specification of sampling units in the population from which a sample may be selected.

¹In an ideal survey, the sampled population will be identical to the target population. 

Intro of survey

Example 1 (Hite's book)

The research interest was the percentage of women who are harassed in their relationship. An individual woman was an *observation unit*. The *target population* was all adult women in the United States. Hite's *sampled population* was women belonging to women's organizations who would return the questionnaire. Consequently, inferences can only be made to the sampled population, not to the population of all adult women in the United States.

Selection bias

Selection bias occurs when

- ▶ some part of the target population is not in the sampled population;
- ▶ some population units are sampled at a different rate than intended by the investigator.

Example 2

A sample of convenience is often biased, since the units that are easiest to select or that are most likely to respond are usually not representative of the harder-to-select or nonresponding units.

Selection bias

Some ways in which selection bias may occur

- ▶ Judgment sample: deliberately or purposively selecting a “representative” sample.
- ▶ Misspecifying the target population.
- ▶ Undercoverage: failing to include all of the target population in the sampling frame.
- ▶ Overage: including population units in the sampling frame that are not in the target population.

Measurement error

Measurement error occurs when a response in the survey differs from the true value.

Example 3 (Unavoidable error)

In the North American Breeding Bird Survey, observers stop every one-half mile on designated routes and count all birds heard or seen during a 3-minute period within a quarter-mile radius. The count of birds for that point is almost always an underestimate of the number of birds in the area.

Measurement error

Measurement error particularly occurs in surveys of people:

- ▶ People sometimes do not tell the truth.
- ▶ People do not always understand the questions.
- ▶ People forget.
- ▶ People give different answers to different interviewers.
- ▶ A particular interviewer may affect the accuracy of the response.

Questionnaire design

Principle: Decide what you want to find out, be precise.

Some useful techniques:

- ▶ Always test your questions before taking the survey.
- ▶ Keep it simple and clear.
- ▶ Use specific questions instead of general ones.
- ▶ Relate your questions to the concept of interests.
- ▶ Decide whether to use open or closed questions.
- ▶ Avoid using leading questions.

Sampling and nonsampling error

Sampling error : the error that results from taking one sample instead of examining the whole population.

Example: If we took a different sample, we would most likely obtain a different sample percentage of persons who visited the public library last week.

Nonsampling errors : any errors that cannot be attributed to the sample-to-sample variability.

Example: Selection bias and measurement error.

Why use sampling?

Three main justifications:

- ▶ Sampling can provide reliable information at far less cost than a census.
- ▶ Data can be collected more quickly, so estimates can be published in a timely fashion.
- ▶ Finally, and less well known, estimates based on sample surveys are often more accurate than those based on a census because investigators can be more careful when collecting data.