Sample Surveys

Think about data collected in class.

What group of people is described?

Young adults?

Young adults in Canada?

In Lower Mainland?

SFU undergraduates?

This class?

Only the people who filled out forms?

Goal of a survey or experiment:

Extend conclusions beyond subjects actually studied.

Must distinguish between:

Population: target group of individuals we are studying.

Sample: group of individuals from population on which we have data.

In order to generalize conclusions from sample to population:

Must know the relation between the two groups of people.

Methods for picking a sample: (called **sampling designs**).

1) Convenience samples: stand on corner and ask passers by.

2) Voluntary response: invite audience to go on-line to fill in questionnaire.

3) Quota sampling: give interviewers quotas: so many white, so many male, ...

4) Probability samples use impartial chance device to select.

4a) Simple random sampling: pick people from pop one at a time with each person having same chance of being drawn at random.

4b) Stratified random sampling: make separate lists for each province, say, and then do 10 simple random samples.

4c) Cluster samples: make list of households (cluster) and then put everyone in selected households in sample.

4d) Multistage sampling: make list of towns, pick some towns by SRS, say. In each selected town pick some neighbourhoods by SRS. In each neighbourhood pick houses by SRS. Then pick people from each house by SRS.

4e) Probability proportional to size: pick farms to survey by picking points on map at random. So large farms more likely to be picked.

Government surveys usually combine 4b,c,d.

Probability designs good; others bad.

Explanation via examples:

Example: US presidential election of 1936. Landon (Republican) vs Roosevelt (Democrat).

Survey carried out by *Literary Digest*:

Surveyed 10 million people; got 2.4 million answers.

Predicted election result: Landon 57%, Roosevelt 43%.

Survey carried out by *Gallup*:

Surveyed 50,000 people.

Predicted election result: Landon 44%, Roosevelt 56%

Actual result: Landon 38%, Roosevelt 62%

Detailed criticisms of surveys along with **jargon**:

Literary digest **frame** (list of people in population) had 10 million people: car owners, telephone owners, club membership lists.

Frame did not cover the population. Serious **undercoverage** issues — missed those less wealthy.

Then 3/4 of those surveyed did not respond: very large potential **non-response bias**. Non-respondents tend to be different from respondents.

Gallup: used quota sampling. Had lower non-response rate, better coverage.

BUT: suffered from selection bias. Interviewers free to select from among people meeting quotas. Easier to interview Republicans within each quota group.

Another example:

US presidential election of 1948: Truman (Democrat) vs Dewey (Republican).

Predictions from pollsters:

	Polling Organization			
Candidate	*Crossley*	*Gallup*	*Roper*	Results
Truman	45	44	38	50
Dewey	50	50	53	45


Chicago Daily Tribune
DEWEY DEFEATS TRUMA
GOP Sweep Indicated in State, Bogle Leads in

Why the failures?

Quota sampling: send interviewer to some city. Set quota.

One example: survey 13 people (in St Louis)

- 6 suburbs, 7 central city.
- 7 men, 6 women
- 3 men under 40, 4 over 40.
- 1 black male, 6 white

and so on.

Problem: otherwise interviewers free to pick as best they can. Leads to **selection bias**: better dressed, eg, easier to find, easier to approach (?) in each category.

Other sampling problems:

1) **Response bias**: ask questions about race – answers gathered by black interviewers differ from those gathered by white. Respondents tailor answers to questioner.

2) **Questionnaire design**: exact wording matters. Example from text: "support for poor" more popular than 'welfare'.

A Canadian survey:

The labour force survey. Used to estimate unemployment rate.

Stratified: by province, economic area, ...

Multistage: pick towns / neighbourhoods, then households.

Cluster: collect data on everyone in selected household.

Panel: once household selected into sample it stays there for 6 consecutive months.

First month: Computer Assisted Personal Interview.

Next 5 months: Computer Assisted Telephone Interview.

Issues:

1) definition of unemployment.

2) universe / target population: excluded: military, inmates, residents of Reserves, residents of Territories.

3) panel fatigue: do respondents modify answers as time goes by to shorten process?

4) different results from CATI / CAPI?

5) Response rate: low 90% range. Extraordinarily good. NOTE: response is legally required!

6) panel structure permits accurate estimate of **change** in unemployment rate.

Recent Canadian Example: National Household Survey

Replaced Census Long Form.

Census every 5 years – 8 to 10 questions.

Long form to "random" sample of 1 in 5 households, many more questions.

Cancelled by Harper government – invasion of privacy.

Replaced by NHS – sample of 1 in 3 households.

But response rate down from about 96% – perhaps 80%.

Non-respondents are different – younger, smaller households, more in some areas than others.