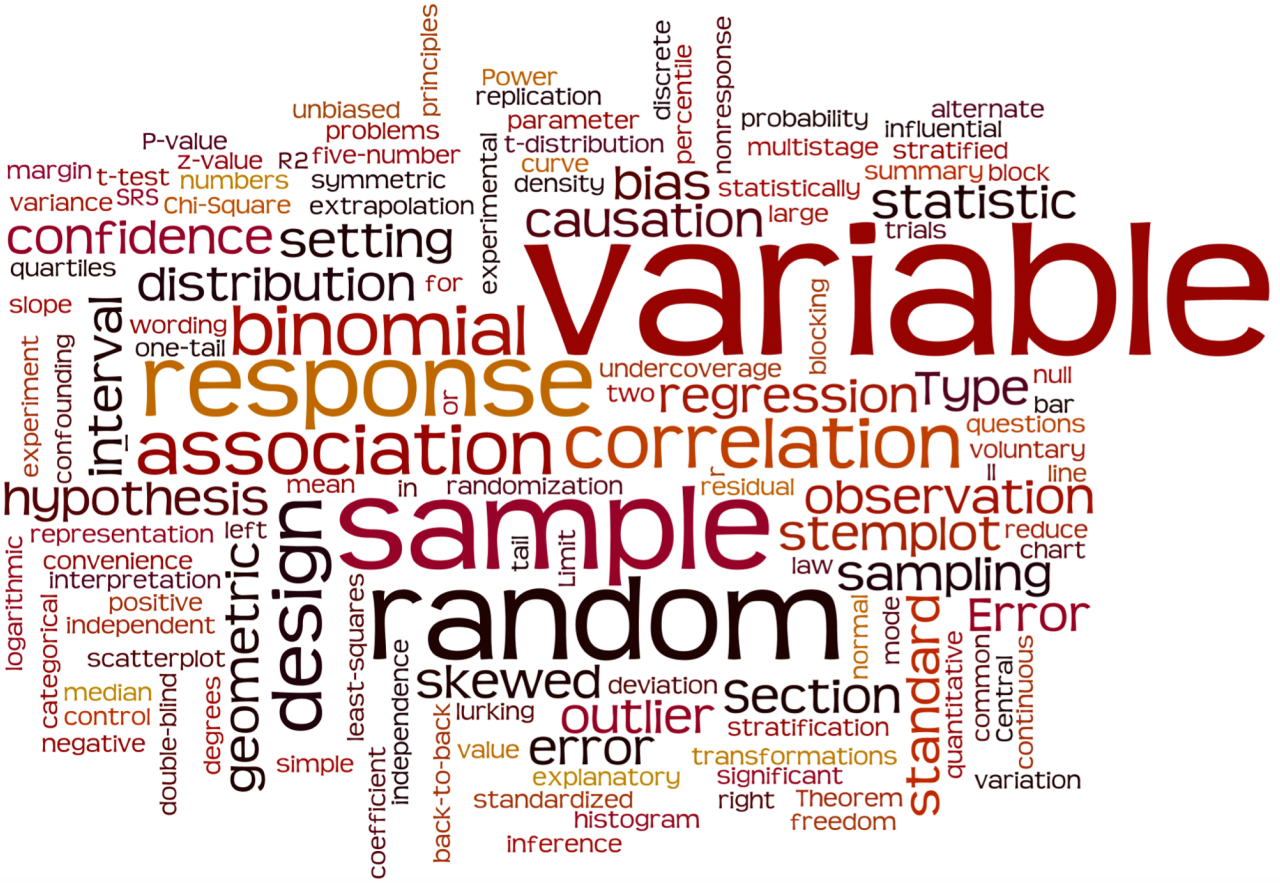
Introduction to Quantitative Methods



Kristin Eccles

Outline

- Sampling
  - Samples vs. populations
  - Research question... Scope
  - Sampling design
  - Bias, sample reliability
- Inference
  - Sampling distributions
  - Central limit theorem
  - Standard error
  - Confidence interval
- Making a database
  - Developing a questionnaire
  - Storing your data
  - Creating a codebook

Study Design

Find a gap in the literature

- This will require a literature review
- Review articles are helpful!

- Developing a research question
  - Clear, focused, tractable, feasible, interesting, original
  - What is the population you are studying?
  - What is the scope of the study?
  - Be as specific as possible

Study scope

- What process/phenomenon are you investigating?
- What controls/influences it?
- Where is it manifested?
- What scale is it manifested at?
  - Is there a spatial and temporal scale?
- Scope the study to account for all of these
  - Change the research question or sampling scheme accordingly
  - Sample in the right place at the right time
  - Control factors that may influence your variables

Samples vs. Populations- again

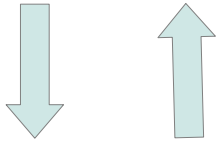
- Not practical to measure everything, everywhere
- So the sample must be:
  - Practical to carry out (logistically feasible)
  - Statistically robust/meaningful
- Sampling is **almost always** the part of the study that takes the most effort
  - Sampling takes careful consideration and a lot of planning a head of time.
- The sample must be **representative of the population** to infer its characteristics

Population to Sample

Population

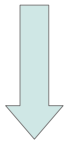
Target population (conceptual)

all individuals & measurements relevant to a study



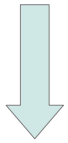
Sample frame (operational)

defines where, when & what the samples will be



Sampling design

How samples will be acquired



Samples (*n*=enough)

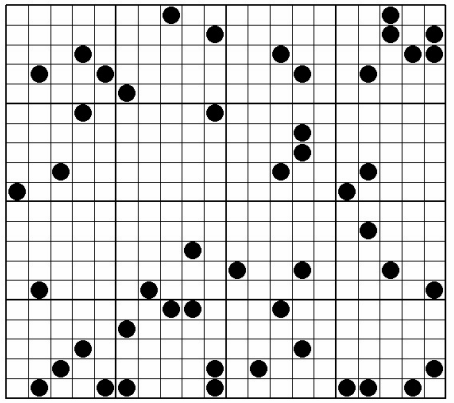
Sampling design

- Non-probabilistic
  - Cannot guarantee a representative sampling
  - Not to be used for inferential statistics
  - Great for other purposes (see GEOG 2005)
  - Snowball, quota, judgmental, convenience
- Probabilistic
  - Can estimate how representative the sample is or how much error there is

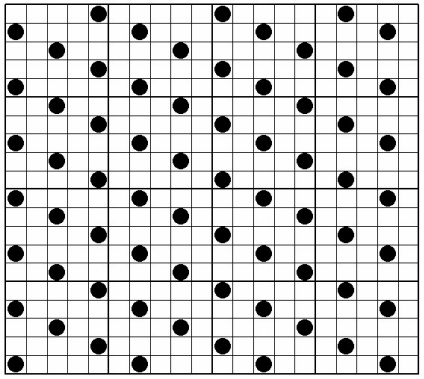
Probabilistic

- Simple random
  - Fully randomized across space & time
  - Inefficient – but the *most probabilistic*
- Systematic
  - Regular spacing, starting at an *arbitrary* point (grid, transect, lists)
  - Very effective but potential to miss info (aliasing)
- Stratified random
  - Divide sample frame into strata and sample each stratum randomly
  - Proportional or disproportional
  - Good for heterogeneous sample frames
- Combinations/hybrids of the above

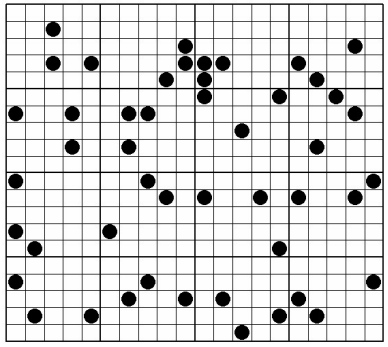
SIMPLE RANDOM SAMPLE



SYSTEMATIC SAMPLE



STRATIFIED SAMPLE

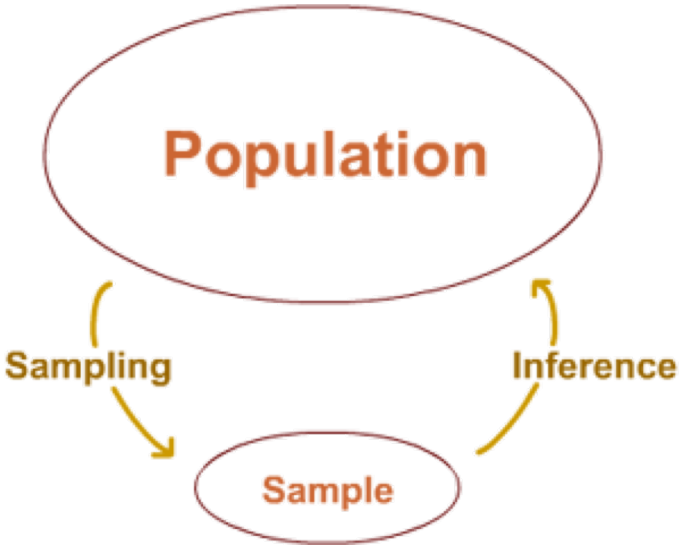


Comparing samples (and resamples)

- Remember that population parameters are constant and unknowable
- Sample statistics depend on which individuals/values are in the sample (not to mention measurement error)
- No two samples are exactly the same – so everyone will get a different mean based on the sample they chose!
  - Even if they take large samples

Representative vs biased samples

- Bias – unrepresentative samples will impact the outcome of the study
  - Sample selection, method, spatial or temporal
- Causes:
  - Poor planning/understanding
  - Constraints
  - Scientific fraud
- Primary vs. secondary data



- Representative samples reflect the relevant characteristics of the population

Sampling and the law of large numbers

- Your aim is to sample until your statistics become close enough to the population parameters
- The law of large numbers guarantees this
  - If you have a random/probabilistic sample design
  - And each sample is independent of each other
  - And you have a suitably large number of observations in your sample!
- The sampling error will decrease as sample size increases
- How do we know we are close?

How reliable is the sample?

- 2 factors:
  - Variability in the sample (s) → less is better
  - Sample size (n) → bigger is better

- The sampling error, or the unreliability, for a sample mean is:

$$unreliability = \frac{s}{\sqrt{n}}$$

Central Limit Theorem (CLT) Review

- In probability theory the CLT states that when a variable is independent and random the sample average will tend to form a normal curve
- This implies that probabilistic and statistical methods work for normal distributions
- https://www.youtube.com/watch?v=JNm3M9cqWyc

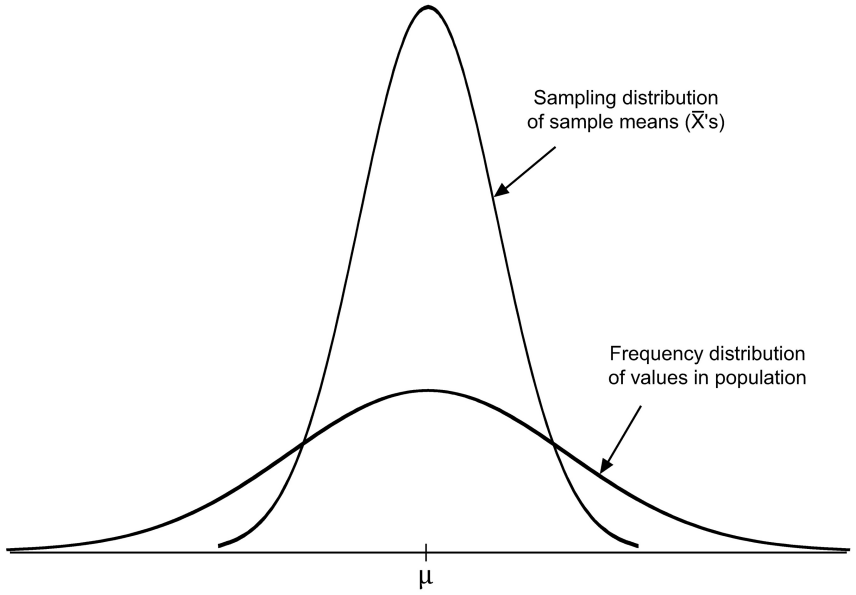


FIGURE 8.1
Sampling Distribution of Sample Means and Frequency Distribution of Population Values

Central Limit Theorem (CLT) Review

- Imagine all possible ways to sample a population with a given sample size (*n*)
- The mean of each of these individual samples is itself a variable with a distribution (called a **sampling distribution**)
- The *mean of the sampling distribution* ($\bar{x}$) is centered on the population mean (μ)
- The *variance of the sampling distribution* (σ_s^2) is the population variance (σ^2) divided by the sample size (*n*)
- As *n* becomes larger, the sampling distribution becomes *more normal*
  - *Regardless* of the distribution of the underlying variable!

CLT - Conditions/Assumptions

- Sampled values must be statistically independent
- Sampling must be random and unbiased
- The sample size (n) must be sufficiently large
  - $n \geq 30$ is recommended but should be increased if the variable's distribution is far from normal (e.g., highly skewed)
- n must be < 10% of the population size N if samples are not replaced (otherwise they are not independent)
- There must be some variability in the population

CLT – further

- Some conditions/assumptions can be relaxed a bit:
  - Independence → mostly independent
  - Random sample → probabilistic sampling
- Remember: the sampling distribution is a mathematical/theoretical construct, nobody looks at this in real life
- Will a bigger sample size make the underlying data normally distributed??
  - No, variables are normally distributed if they are influenced by random independent factors
  - A greater n will only make the sampling distribution more normal

CLT – what does this mean?

- When you sample your mean sits in a normal distribution (the sampling distribution).
- You know the sampling distribution is centered on the population mean but not where the mean of your sample is.
- BUT you can estimate the standard deviation (σ) of the sampling distribution and determine probabilities using the normal distribution.
- The CLT is the theory behind using the normal distribution as the 'probability portal' to infer population parameters from sample statistics

Standard error of the mean

- The standard deviation of the sampling distribution is variance / *n*
- This is also known as the **standard error of the mean** (SEM or se or $se_{\bar{x}}$):

$$SEM = \frac{s}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$

- Since σ is unknown, estimate it by substituting s
- The SEM is a measure of uncertainty / unreliability / error
- The smaller the SEM the better

Sample mean to population mean

- We will never know μ
- But, because the sampling distribution is:
  - Normally distributed
  - Centered on μ
  - With a standard deviation ≈ SEM
- Using the normal curve we can determine probabilities:
  - P($\bar{x}$ is under the normal curve) = 1 → trivial and useless
  - P($\bar{x}$ is > μ) or P($\bar{x}$ is < μ) = 0.5 → also not helpful
  - P($\bar{x}$ is within 1 standard deviation of μ) = 0.68
  - P($\bar{x}$ – SEM < μ < $\bar{x}$ + SEM) = 0.68
  - P($\bar{x}$ is within 2 standard deviations of μ) = 0.954
  - P($\bar{x}$ – 2 x SEM < μ < $\bar{x}$ + 2 x SEM) = 0.954

The Normal Curve

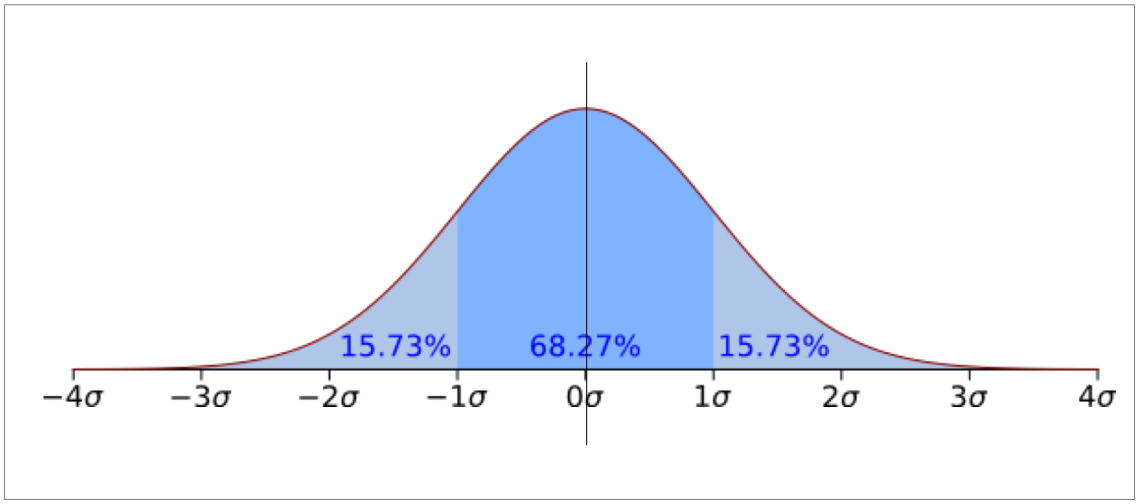
68% of the values are within one standard deviation of the mean...

95% of the area under the curve is within 1.96 standard deviations of the mean

99% of the area under the curve is within 2.58 standard deviations of the mean

99.7% of the area... within 3 standard deviations of the mean

Find your own 'within' using the table or computer



- Remember: *within* implicates the area on **both** sides of the mean
- Draw a diagram!

Confidence intervals of the mean (CI)

- The probability that the interval surrounding the sample mean ($\bar{x}$) actually contains the mean of the population (μ)
- Often expressed as a percent:
  - 95% and 99% confidence intervals are standard
- Can calculate any confidence interval:

$$P\left(\bar{x} - \left|z_{\alpha/2}\right| \times SEM < \mu < \bar{x} + \left|z_{\alpha/2}\right| \times SEM\right) = 1 - \alpha$$

- Where α (the **alpha value**) is the complement of the confidence interval probability
- $\alpha = 1 - (CI/100)$
- Assumes a large sample size ($n > 30$)

Confidence Interval Continued

$$ME = |z_{\alpha/2}| \times SEM$$

- In practice, the confidence interval can be understood as the mean $\pm$ the Margin of Error (ME)
- Or, the region between the **lower bound** (mean – ME) and the **upper bound** (mean + ME)
- It is where you expect to find the *population mean* (μ), but remember: you cannot be certain of this, so you need to assign a probability
- =CONFIDENCE.NORM(α, stdev, n)

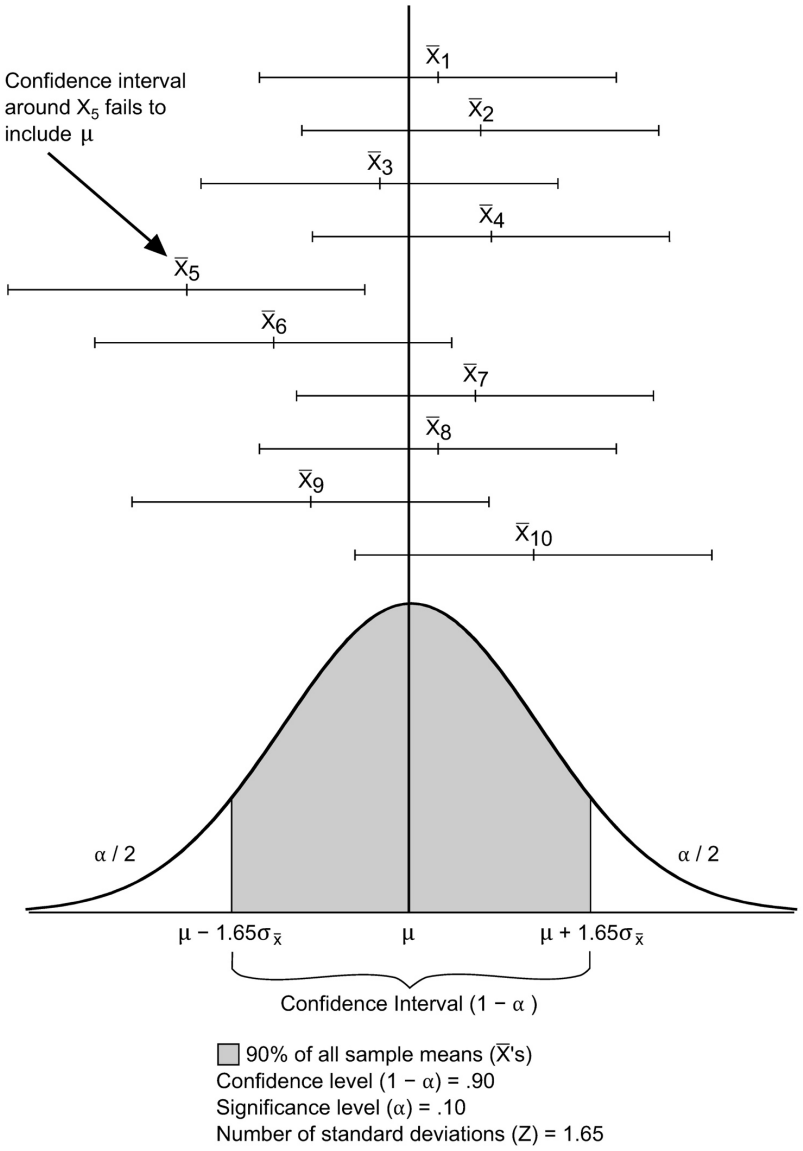


FIGURE 8.5
Distribution of Sample Means and the Confidence Interval Concept

Confidence interval of a proportion

- Suppose you conducted a poll "Are you concerned about the environment?"
- 345 said 'yes' out of *n*=500; the rest said 'no'
- The proportion (x-tilde) of 'yes' is 345/500 or 0.69
- What is the **proportion (*p*)** of people who are concerned about the environment in the **population**?
- Use the formula to find the SEM of a proportion the rest is the same as the previous example

$$SEM = \sqrt{\frac{p(1-p)}{n}}$$

- Note that proportion is a binomial variable and it follows the binomial distribution. In this example we are avoiding this by using the normal approximation of the binomial distribution which is ok so long as n is big

Research Example: First Nations Food, Nutrition and Environment Study

Website: www.fnfnes.ca

Many sources of pollution threaten traditional food systems

What We Know

Traditional food:

- ✓Traditional food is an important source of many nutrients that are not consumed in sufficient amounts
- ✓Diets are healthier when traditional food is eaten than if just market foods are eaten

Study Objectives and Questions

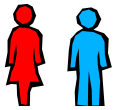
- There remains a gap in knowledge at the national and regional level on nutritional composition and the environmental safety of foods consumed by First Nations peoples living on reserve lands south of 60th parallel across Canada.
- There is a lack of knowledge on the baseline levels of environmental pollutants in the traditional foods across Canada.
- There remains a gap in knowledge on the total diet of First Nations across Canada.
- Research Questions:
  - How much of an issue is chemical contamination of traditional food?
  - What traditional foods are eaten and how much?
  - What level of contaminants are present in traditional food?

FNFNES Study Design

FNFNES used a 3 stage **random sampling design:**

1. In each region, First Nations communities were randomly chosen.

2. In participating communities, households were randomly selected to participate.



3. In each randomly selected household, the person asked to participate was the adult (aged 19+) with the next birthday

Sampling Stage 1:

- **12 First Nations communities in the Atlantic were randomly selected**
- **2 were pre-selected**
- **4 declined, 3 alternate communities invited; 1 no alternate community**

Abegweit
Lennox Island
Elsipogtog First Nation
Buctouche
Esgenoopetitj First Nation
Madawaska Maliseet First Nation
Eel Ground First Nation
Eel River Bar First Nation
Fort Folly
Indian Island
Kingsclear
Oromocto
Pabineau
Metepenagiag Mi'kmaq Nation
Saint Mary's
Tobique
Woodstock First Nation

Acadia
Paqtnkek First Nation
Annapolis Valley
Bear River
Potlotek First Nation
Eskasoni First Nation
Pictou Landing First Nation
Shubenacadie
Membertou First Nation
Millbrook
Wagmatcook
Waycobah First Nation
Glooscap First Nation
Miawpukek First Nation

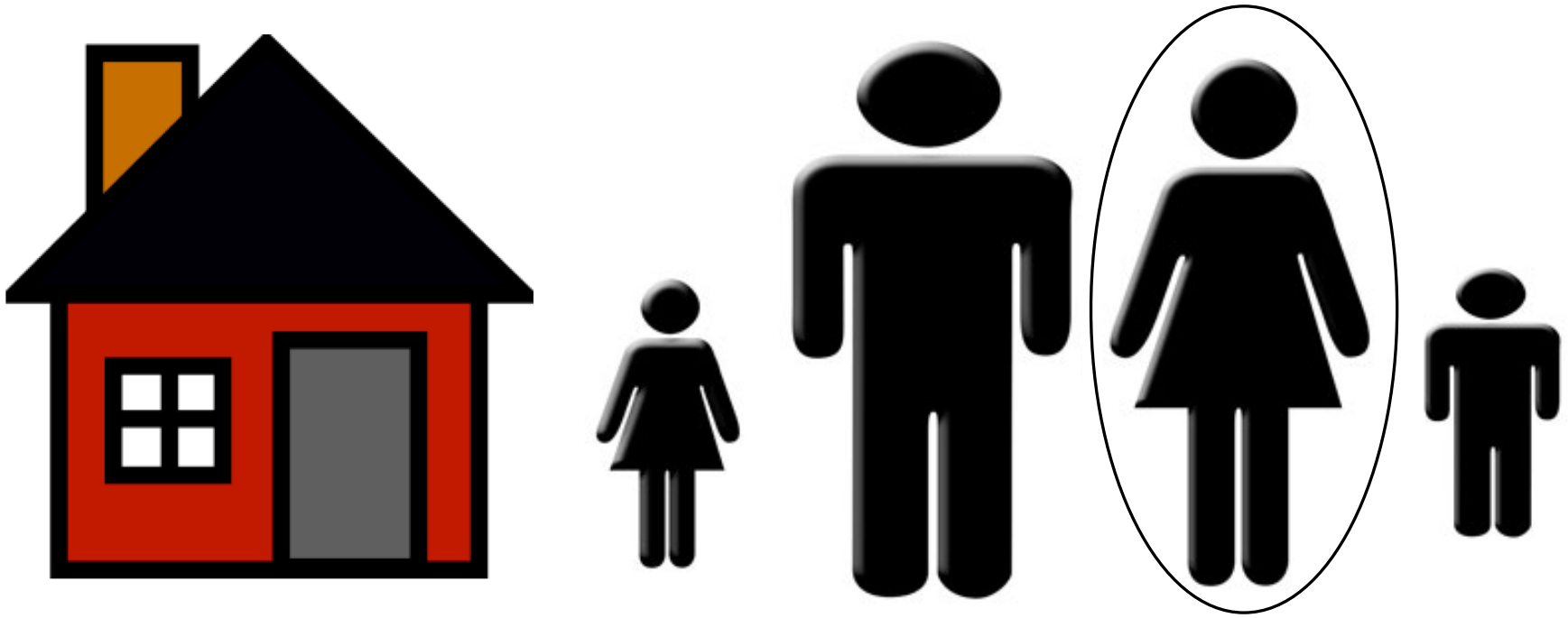
Total: 11 communities participated

Sampling Stage 2:

In participating communities, 125 households were randomly selected to participate.

Sampling Stage 3:

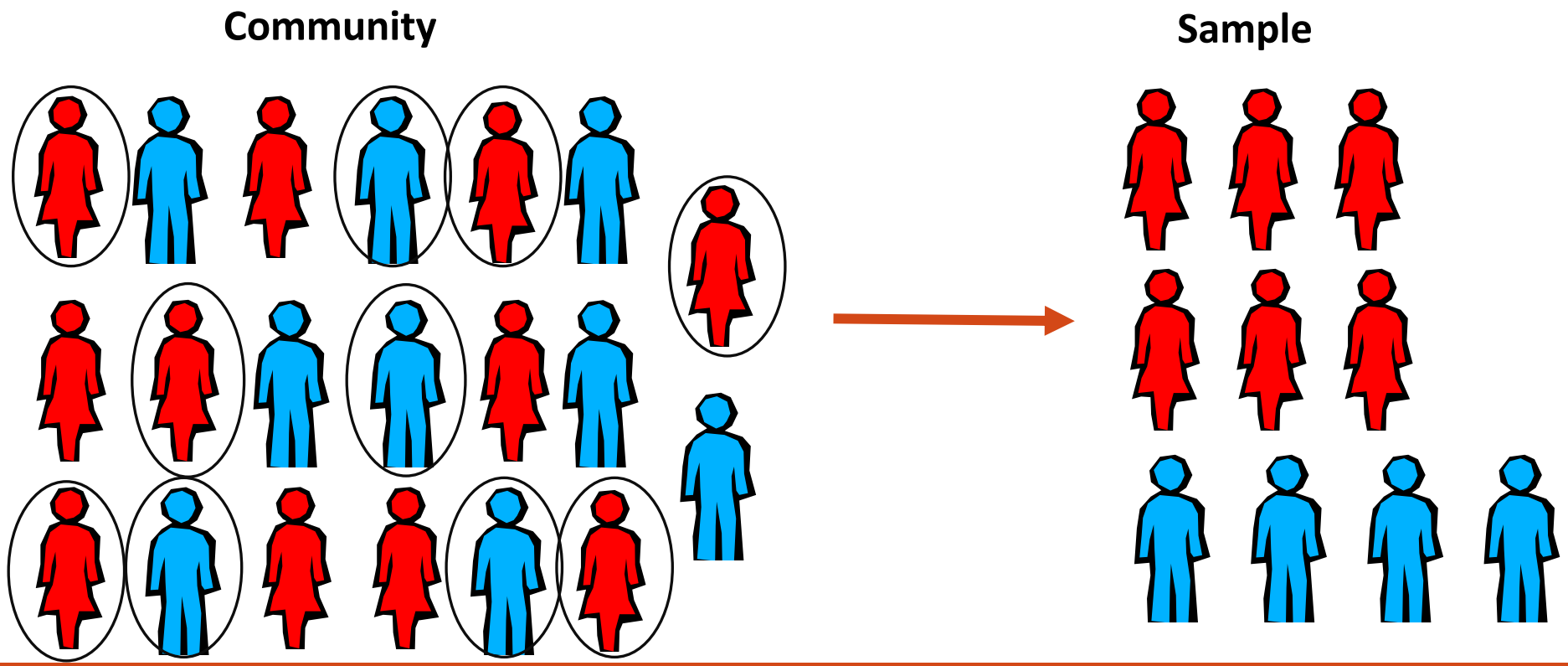
In each randomly selected household, the person asked to participate was the adult (aged 19+) with the next birthday



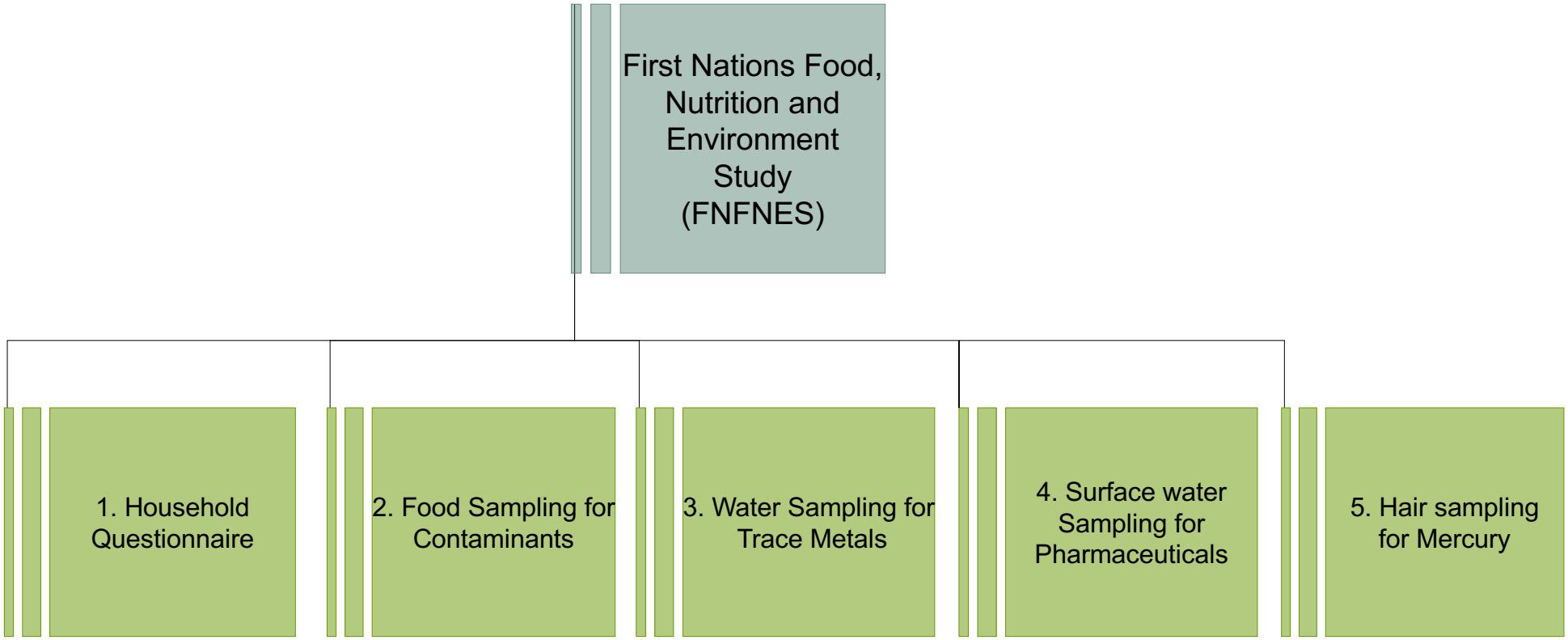
Target: 100 participants per community

Sample

In each community, the people who participated in FNFNES were a **sample** of all the people (population) of that First Nation.

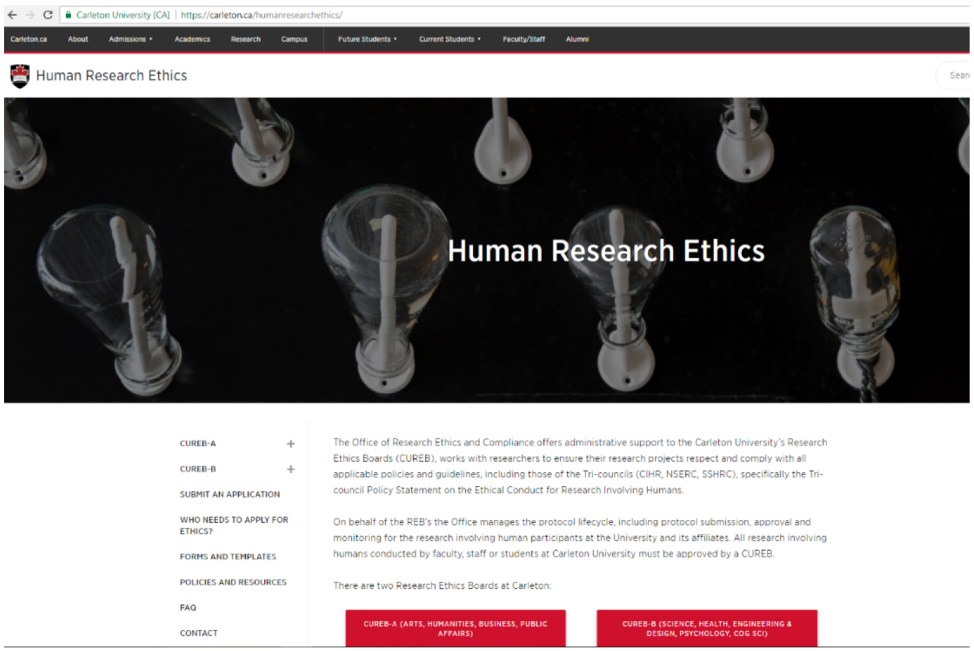


First Nations Food, Nutrition and Environment Study



Ethics

- Do you need ethical approval?
  - Are you working with humans or animals?
  - Usually from a university research board



https://carleton.ca/humanresearchethics/

FNFNES Consent Form

First Nations Food, Nutrition and Environment Study
Individual Consent Form

Invitation to Participate: You are invited to participate in the First Nations Food, Nutrition and Environment Study (FNFNES). This is a study being done in collaboration with the Assembly of First Nations (AFN) (Dr. Diego Garcia), Health Canada (Dr. Harold Schwartz), the University of Ottawa (Dr. Laurie Chan and Dr. Malek Batal) and the Université de Montréal (Dr. Olivier Receveur). Funding is provided through Health Canada.

Purpose of the Study: Canada has been conducting health and total diet studies of the general Canadian population to understand more about food safety and how changes in diet relates to chronic conditions like heart disease and diabetes but First Nations people living on-reserve have not been included in these studies. This study will gather information on current consumption of traditional and store-bought food and test traditional foods for nutrient content and hazardous environmental chemicals, such as mercury. In order to determine if these foods increase exposure to contaminants and contribute to health risks, samples of hair will be analysed for the presence of mercury. The study will also test samples of drinking water for trace metals and surface water for pharmaceuticals.

Participation: If you agree to participate, it will require about one to two hours of your time, mainly to answer questions about the food you eat. We would also like to measure your height and weight and take a small sample of hair for mercury analysis. In some randomly selected households, you may be asked if we can collect a water sample from the kitchen sink.

Risks: There is no physical harm anticipated for participating in the project.

Benefits: You will have a chance to provide input into the benefits and risks of traditional and commercial food use. Benefits to each community include the development of baseline of exposure to contaminants through food, greater knowledge of levels of exposure of mercury in the environment and state of knowledge of regional risks associated with food such as the reliance on traditional foods and the importance of maintaining traditional foods in the daily diet.

Confidentiality and anonymity: All information you provide in this interview will be treated with respect and held in confidence including information you share with the interviewer. All information from the study will be kept strictly confidential and your name will not be associated with any information except for the mercury in hair data. After the results of the hair analysis has been returned to you, all information linking your name to the survey is destroyed. All hardcopies of the questionnaires collected will be kept in a secured room at the University of Ottawa until the final report of this project is complete or a maximum of 5 years.

Data Ownership: Your community is considered the owner of all data collected from your community and will be provided with a copy of the dataset upon completion of the study. The data will be kept in a secure manner in a locked room at the University of Ottawa until the completion of the study. The AFN will securely store a backup copy of the data on behalf of your community and will not use or provide it to anyone unless explicitly directed to do so by your community.

http://www.fnfnes.ca/docs/MB%20Reports/Individual_consent_form_032613.pdf

Developing a Questionnaire

- You designed your research question

PARTICIPANT ID __ __ / __ / __ __ __

I. TRADITIONAL FOOD AND WATER QUESTIONNAIRE

Community Number ____

Participant's gender (1=female, 2=male) _____ Household number _____

Participant's age _____

Age group: 19-30 years old _____
31-50 years old _____
51-70 years old _____
71+ years old _____

For WOMEN only:
Currently pregnant: Yes □ No □
Currently breastfeeding: Yes □ No □

Interview language: English □ Other, please specify _____

Interviewer's Initials _____

Date of interview (dd/mm/yyyy) _____

This section contains 2 parts. The first part asks about the traditional foods that you have eaten in the past year and how often you ate them. The second part asks about the sources of water in your house, your average consumption of beverages and soups likely to contain water and the different sources water in your home used to make the beverages and soups.

A. TRADITIONAL FOOD

This part is about traditional food – that is, food harvested within the local environment. It can be in any form – for example: dried, smoked, fermented, fresh, frozen...
I will begin by asking about fish that were harvested within the past year.

Participant ID: __ __ / __ / __ __ __

11. In general, compared to other people of your age, would you say your health is:
    a. Excellent
    b. Very good
    c. Good
    d. Fair
    e. Poor

12. Which of the following statements best describes your activities for most days when you are in the community?
    a. I am usually sitting and do not walk around very much.
    b. I stand or walk around quite a lot, but I do not have to carry or lift things very often.
    c. I usually lift or carry light loads or I have to climb stairs or walk up hills often.
    d. I do heavy work or carry heavy loads.

13. In general, compared to other people of your age, are you physically:
    a. More active
    b. Less active
    c. About average
    d. Don't know

14a. Did you smoke cigarettes yesterday? YES ☐ NO ☐

14b. ***[IF YES ABOVE, ASK]*** How many? _____

15. Have you ever been told by a health care provider that you have:
    a. diabetes YES ☐ NO ☐
    b. If yes to 15a, how long ago were you diagnosed? _____# years
       _____ don't know
    c. If yes to 15a, circle type if known: Type 1 Type 2 Unknown

Storing your data

- What types of data are you storing
  - Nominal/ Ordinal
  - Interval/Ratio
- Does this require coding?
  - E.g. Male=1, Female=2
  - E.g. 0=no, 1=yes
- Developing a database
  - Excel
  - MS Access

Variable Names (column headings)

- A unique, unambiguous name should be given to each variable.
- Variables names should be long enough to be meaningful, but short enough to be easy to handle
- The codebook is the place to clarify what is coded in each variable in more detail
- Variables names **MUST** consist of one string only
  - Consisting of letters, numbers, and underscores (_)
    - BUT do not start a variable name with a number
  - **Spaces are not allowed** in variables names in most statistical programs, even if data entry programs like Excel or Access will allow this.
  - Camel case: thisIsTheVariableName
  - Snake case: this_is_the_variable_name

Variable Names (column headings)

id	community_no	gender	household_no	age_group	pregnant

Writing a codebook

- Sometimes referred to as a “cookbook”
- Writing a codebook is an important step in the management of any data
- Codebook will serve as a reference for the research team
- Help newcomers to the project to rapidly understand the data
- This the place to clarify what is coded in each variable in more detail
- Can be included in a word document or on another sheet in an excel file

Code Book Example

Section Number	Question Number	Full Question	Database name	Variable Type	Value numeric	Value labels	Notes
I		Participant ID	id	numeric			
I		Community Number	community_no	numeric			
I		Participant's gender	gender	category	0	Male	
					1	Female	
I		Household number	household_no	numeric			
I		Participant's age	age_group	category	1	19-30 years	
					2	31-50 years	
					3	51-70 years	
					4	71+ years	
I		Currently pregnant	pregnant	category	0	No	For women only
					1	Yes	

Other database creation tips

- Dealing with missing data
  - Use impossible values should be used, e.g. -88, -99, -999
  - E.g. males cannot be pregnant
  - Leave them blanks
  - What ever you do make sure it is in the code book and it is systematic
- Always start with an ID
- Other how to's on creating a good database:
  - http://www.medicine.mcgill.ca/epidemiology/joseph/pbelisle/CodebookCookbook/CodebookCookbook.pdf
- Other examples of data and code books:
  - National health and Nutrition Examination Survey (NHANES)
  - https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&CycleBeginYear=2013

Summary

- Sampling:
  - Think before you do!
  - Representative, probabilistic for more stats!
  - Large n if possible
- Inference:
  - Population parameter from sample statistic
  - Sampling distribution – theoretical construct
  - Central limit theorem via normal distribution → probability
  - SEM and CI
- Research in practice
  - Once you design you study and collect your data you must take care to store and organize it
  - Proper database development
  - Proper coding/ documentation

Resources

Confidence interval of a mean (explained in a different way):

https://www.khanacademy.org/math/probability/statistics-inferential/confidence-intervals/v/confidence-interval-1

Confidence interval of a proportion (explained in a different way):

https://www.khanacademy.org/math/probability/statistics-inferential/confidence-intervals/v/confidence-interval-example