Introductory Statistics

2024 Lectures
Part 8 - Sampling Techniques

Institute of Economic Studies
Faculty of Social Sciences
Charles University in Prague

Sample

- sample is a subset of the population
- sampled population is the population from which the sample is drawn
- we collect data to make an inference and/or answer to research question about a population - selecting a sample

Example 38: A tire manufacturer is considering producing a new tire designed to provide an increase in mileage over firm's current line of tires. To estimate the mean useful life of the new tires, the manufacturer produced a sample of 120 tires for testing. The test results provided a sample mean of 36 500 miles. Hence an estimate of the mean useful life for the population of new tires was 36 500 miles.

Sampling error

- sample results provide only estimates of the values of the corresponding population characteristics
- sample contains only a proportion of the population thus our estimates contain sampling error

Example 38 cont.: We do not expect the sample mean of 36 500 miles to be exactly equal to the mean mileage for the population of all such tires ever produced.

- sampling error is not a result of a wrong computation, it is a central element of statistics
- statistics provide theory for quantification of such errors
- for a finite sampled population one can easily construct a frame - the list of all elements of the sampled population
- for an infinite sampled population (in Example 38 that is all the tires that could have been made by the production process) it is impossible to construct a frame to draw the sample from

Example: Electronics Associates

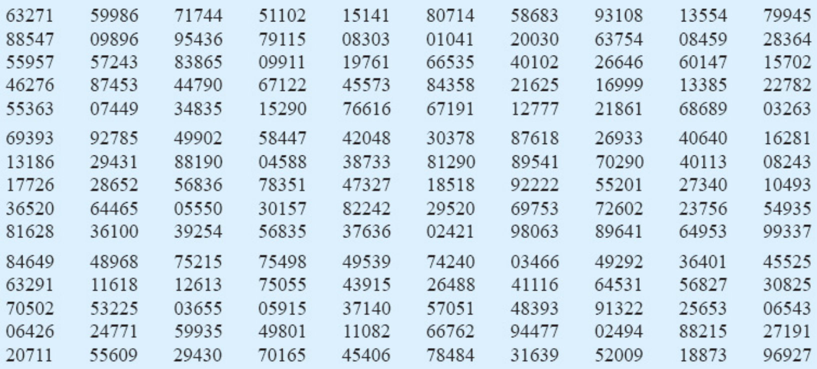
Example 39: Electronics Associates

- task: to develop a profile of the company's 2500 managers
- characteristics to be identified: mean annual salary and the proportion of managers having completed the company's management training program
- the 2500 managers is the population for this study; having the data of all 2500 managers in firm's personnel records one can find parameters of the population
  - population mean $\mu = \$51800$
  - population standard deviation $\sigma = \$4000$
  - proportion of managers that completed the training $p = 0.6$
- and now, assume that the necessary information is not in the company's database - how to obtain estimates of the population parameters by using a sample of managers instead of 2500 in the population?
- if the personnel director is assured that a sample of 30 managers would provide adequate information, how can we identify a sample of 30 managers?

Simple random sample (finite population)

- a simple random sample of size n from a finite population of size N is a random sample selected such that each possible sample of size n has the same probability of being selected
- for example, a computer-generated (pseudo-)random numbers can be used to implement the random sample selection process:
  - label managers from 1 to 2500 (anyhow)
  - take the table of random numbers
  - as 2500 is a 4-digit number, take 4-digit numbers from the table such that if it is less or equal to 2500 then include it in the sample else discard it
  - repeat until the simple random sample of the prescribed size is obtained
  - it can happen that same number is selected repeatedly - ignoring repeatedly selected numbers results in sampling without replacement; else we speak of sampling with replacement

Random table



63271	59986	71744	51102	15141	80714	58683	93108	13554	79945
88547	09896	95436	79115	08303	01041	20030	63754	08459	28364
55957	57243	83865	09911	19761	66535	40102	26646	60147	15702
46276	87453	44790	67122	45573	84358	21625	16999	13385	22782
55363	07449	34835	15290	76616	67191	12777	21861	68689	03263
69393	92785	49902	58447	42048	30378	87618	26933	40640	16281
13186	29431	88190	04588	38733	81290	89541	70290	40113	08243
17726	28652	56836	78351	47327	18518	92222	55201	27340	10493
36520	64465	05550	30157	82242	29520	69753	72602	23756	54935
81628	36100	39254	56835	37636	02421	98063	89641	64953	99337
84649	48968	75215	75498	49539	74240	03466	49292	36401	45525
63291	11618	12613	75055	43915	26488	41116	64531	56827	30825
70502	53225	03655	05915	37140	57051	48393	91322	25653	06543
06426	24771	59935	49801	11082	66762	94477	02494	88215	27191
20711	55609	29430	70165	45406	78484	31639	52009	18873	96927

- 6327 is larger than 2500, thus discard
- 1599 is less than 2500, thus the manager number 1599 will be included
- 8671 is discarded, 7445 is discarded, 1102 will be included in the sample, etc. till sample of 30 managers is obtained

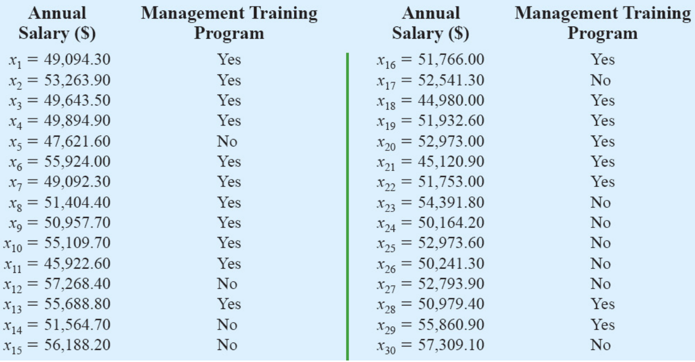
Example: customers of fast food restaurant

Example 40: Customers of fast food restaurant

- consider a customer population of a fast food restaurant
- a customer arrival to the restaurant is an ongoing process
- how to sample from all customers of the restaurant - from infinite population?
- sampling from infinite population are required to satisfy
  - that each element is selected from the population of the same characteristics
  - that each element is selected independently
- E.g. to determine the sample of customers, discard all people that do not make a purchase (e.g. come just to use the toilet) and select customer such that the selection of one does not influence selection of any other (e.g. not ask all customers that came as one group, avoid asking just one particular age group, etc.)- to prevent the selection bias

Sample characteristics

Example 39 cont:

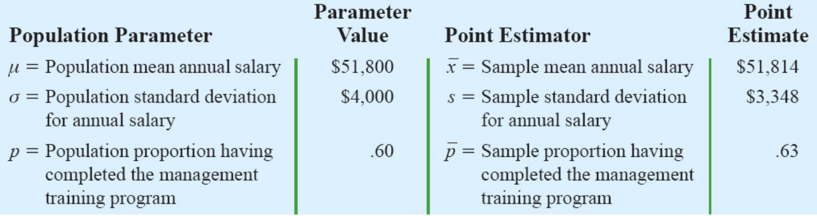


Annual Salary ($)	Management Training Program	Annual Salary ($)	Management Training Program
$x_1 = 49,094.30$	Yes	$x_{16} = 51,766.00$	Yes
$x_2 = 53,263.90$	Yes	$x_{17} = 52,541.30$	No
$x_3 = 49,643.50$	Yes	$x_{18} = 44,980.00$	Yes
$x_4 = 49,894.90$	Yes	$x_{19} = 51,932.60$	Yes
$x_5 = 47,621.60$	No	$x_{20} = 52,973.00$	Yes
$x_6 = 55,924.00$	Yes	$x_{21} = 45,120.90$	Yes
$x_7 = 49,092.30$	Yes	$x_{22} = 51,753.00$	Yes
$x_8 = 51,404.40$	Yes	$x_{23} = 54,391.80$	No
$x_9 = 50,957.70$	Yes	$x_{24} = 50,164.20$	No
$x_{10} = 55,109.70$	Yes	$x_{25} = 52,973.60$	No
$x_{11} = 45,922.60$	Yes	$x_{26} = 50,241.30$	No
$x_{12} = 57,268.40$	No	$x_{27} = 52,793.90$	No
$x_{13} = 55,688.80$	Yes	$x_{28} = 50,979.40$	Yes
$x_{14} = 51,564.70$	No	$x_{29} = 55,860.90$	Yes
$x_{15} = 56,188.20$	No	$x_{30} = 57,309.10$	No

- given a sample, we can estimate the value of a population parameter - compute sample characteristics (sample statistics, point estimates)
  - sample mean $\bar{x} = \$51814$
  - sample standard deviation $s = \$3348$
  - sample proportion $\hat{p} = 0.63$

Sample characteristics

Example 39 cont:

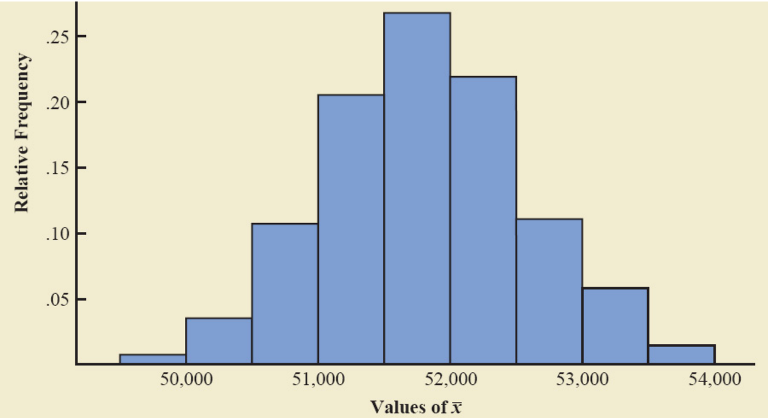


Population Parameter	Parameter Value	Point Estimator	Point Estimate
μ = Population mean annual salary	\$51,800	$\bar{x}$ = Sample mean annual salary	\$51,814
σ = Population standard deviation for annual salary	\$4,000	s = Sample standard deviation for annual salary	\$3,348
p = Population proportion having completed the management training program	.60	$\bar{p}$ = Sample proportion having completed the management training program	.63

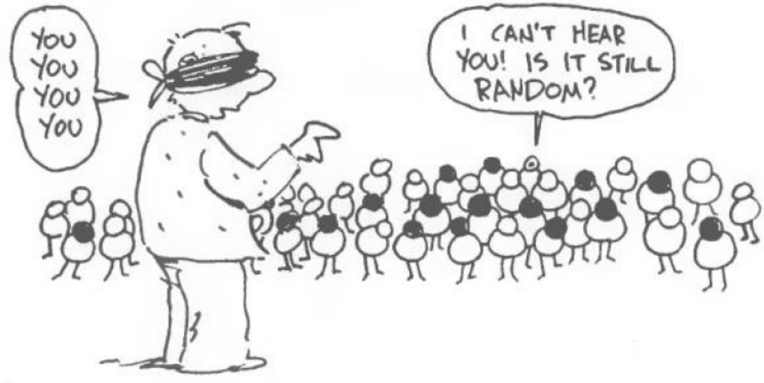
- these point estimates differ from the corresponding population parameter - we used sample, not census

Repeated sampling

- various possible values of $\bar{x}$ result from different random samples
- each such point estimate is realized with the same probability (simple random samples)
- we can visualize the sampling distribution of $\bar{x}$



Alternative sampling methods

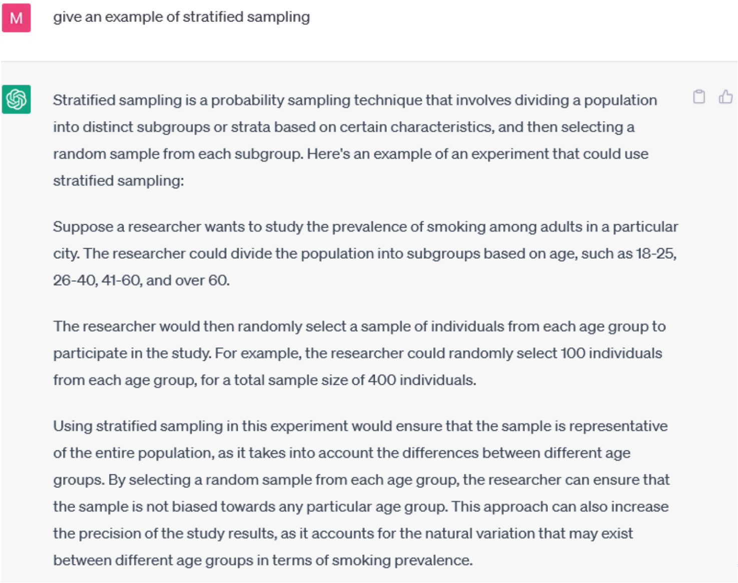


- there are alternatives to simple random sampling
- two main classes of sampling methods: probability sampling techniques and non-probability sampling techniques

Probability techniques: Stratified random sampling

- elements in the population are first divided into groups - strata - e.g. by location, age, industry type etc.
- when strata are formed, a simple random sample is taken from each stratum
- there are formulas for stratified random sampling to combine results from each individual stratum samples into one estimate of the population parameter of interest
- improves representativeness of the sample by reducing sample error
- works well when elements of each stratum are as alike as possible (strata with low variance), lower sample size compared to simple random sampling is required

Probability techniques: Stratified sampling



give an example of stratified sampling

Stratified sampling is a probability sampling technique that involves dividing a population into distinct subgroups or strata based on certain characteristics, and then selecting a random sample from each subgroup. Here's an example of an experiment that could use stratified sampling:

Suppose a researcher wants to study the prevalence of smoking among adults in a particular city. The researcher could divide the population into subgroups based on age, such as 18-25, 26-40, 41-60, and over 60.

The researcher would then randomly select a sample of individuals from each age group to participate in the study. For example, the researcher could randomly select 100 individuals from each age group, for a total sample size of 400 individuals.

Using stratified sampling in this experiment would ensure that the sample is representative of the entire population, as it takes into account the differences between different age groups. By selecting a random sample from each age group, the researcher can ensure that the sample is not biased towards any particular age group. This approach can also increase the precision of the study results, as it accounts for the natural variation that may exist between different age groups in terms of smoking prevalence.

Probability techniques: Cluster sampling

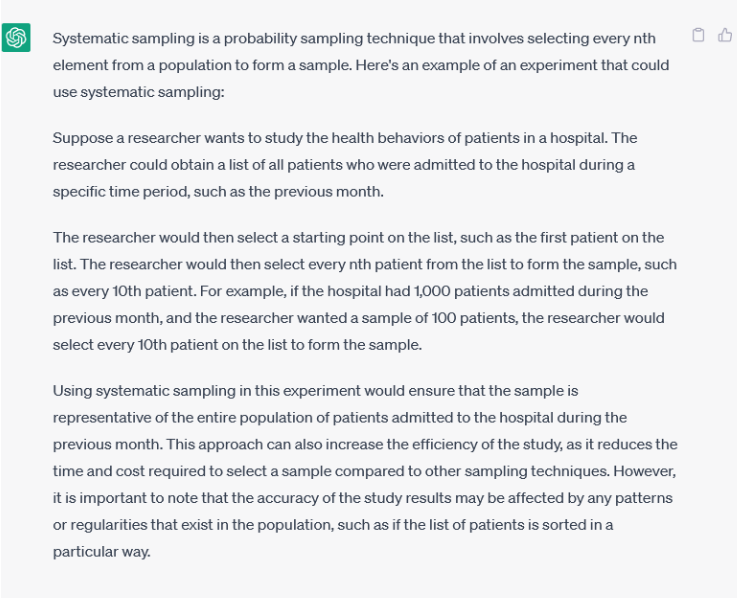
- elements in the population are first divided into separate groups - clusters
- when clusters are formed, a simple random sample is taken of clusters
- then take a simple random sample (or take every element) from sampled clusters
- works well when elements of each cluster are as heterogenous as possible (to achieve homogeneity between cluster means), i.e. each cluster is a small-scale version of the entire population
- if clusters are alike, sampling small number of clusters will provide good estimates of the population parameters - e.g. area sampling of city blocks
- often larger sample size compared to simple random sample or stratified sample, but may result in significant cost savings

Probability techniques: Systematic sampling

- selection of elements is performed from an ordered sampling frame based on a predefined procedure
- e.g. the skip - start by randomly selecting an element and choose every kth element, $k = \frac{N}{n}$, in the frame (N population size, n sample size) - equal probability method
- useful when the population is homogeneous and the choice of the sampling interval does not hide a pattern in data

EXCUSE ME... WOULD YOU MIND ANSWERING FIFTY OR SIXTY QUESTIONS?
TOLL 47¢
47¢

Probability techniques: Systematic sampling



Systematic sampling is a probability sampling technique that involves selecting every nth element from a population to form a sample. Here's an example of an experiment that could use systematic sampling:

Suppose a researcher wants to study the health behaviors of patients in a hospital. The researcher could obtain a list of all patients who were admitted to the hospital during a specific time period, such as the previous month.

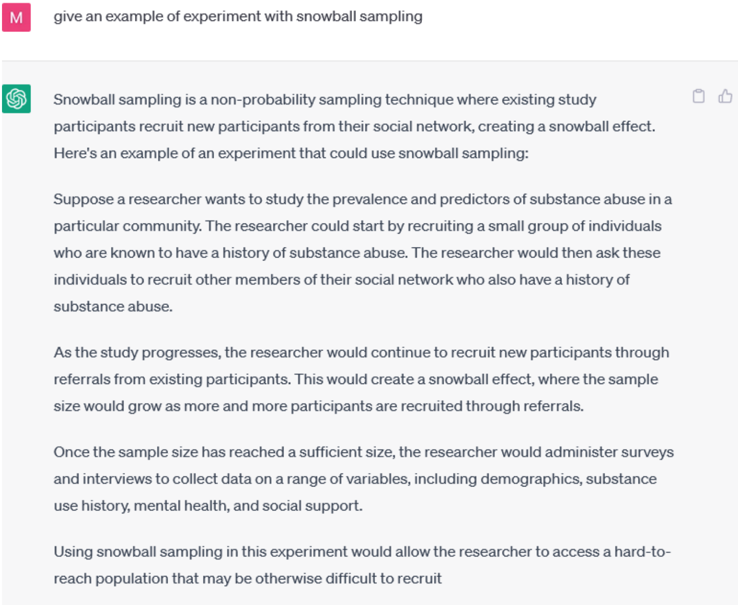
The researcher would then select a starting point on the list, such as the first patient on the list. The researcher would then select every nth patient from the list to form the sample, such as every 10th patient. For example, if the hospital had 1,000 patients admitted during the previous month, and the researcher wanted a sample of 100 patients, the researcher would select every 10th patient on the list to form the sample.

Using systematic sampling in this experiment would ensure that the sample is representative of the entire population of patients admitted to the hospital during the previous month. This approach can also increase the efficiency of the study, as it reduces the time and cost required to select a sample compared to other sampling techniques. However, it is important to note that the accuracy of the study results may be affected by any patterns or regularities that exist in the population, such as if the list of patients is sorted in a particular way.

Non-probability technique: Convenience sampling

- with non-probability sampling method some elements of the population may have no chance of selection
- convenience sample is identified primarily by convenience or proximity of researcher
- e.g. wildlife captures, volunteers as subject of research, Facebook friends
- advantage: easy (and comfortable) sample selection and data collection
- price: with non-probability samples we are unable to establish representativeness of the population and unable to perform a probability analysis and inference about the quality of the sample results

Non-probability techniques: Snowball sampling



give an example of experiment with snowball sampling

Snowball sampling is a non-probability sampling technique where existing study participants recruit new participants from their social network, creating a snowball effect. Here's an example of an experiment that could use snowball sampling:

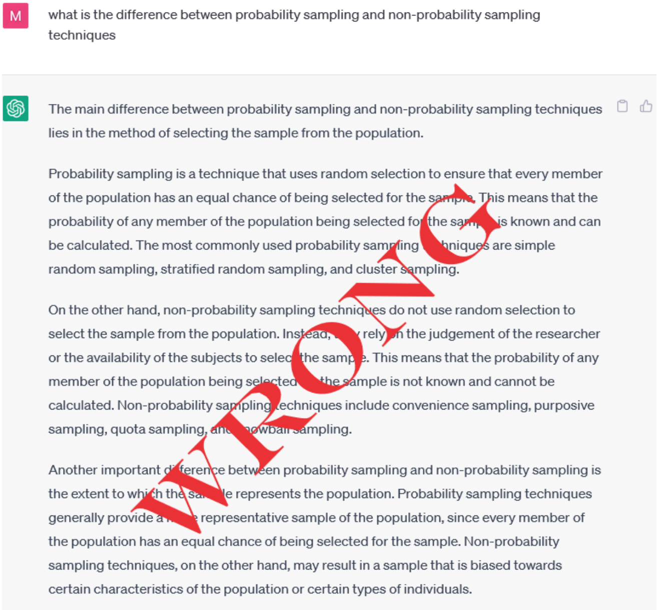
Suppose a researcher wants to study the prevalence and predictors of substance abuse in a particular community. The researcher could start by recruiting a small group of individuals who are known to have a history of substance abuse. The researcher would then ask these individuals to recruit other members of their social network who also have a history of substance abuse.

As the study progresses, the researcher would continue to recruit new participants through referrals from existing participants. This would create a snowball effect, where the sample size would grow as more and more participants are recruited through referrals.

Once the sample size has reached a sufficient size, the researcher would administer surveys and interviews to collect data on a range of variables, including demographics, substance use history, mental health, and social support.

Using snowball sampling in this experiment would allow the researcher to access a hard-to-reach population that may be otherwise difficult to recruit

Probability vs Non-probability techniques



what is the difference between probability sampling and non-probability sampling techniques

The main difference between probability sampling and non-probability sampling techniques lies in the method of selecting the sample from the population.

Probability sampling is a technique that uses random selection to ensure that every member of the population has an equal chance of being selected for the sample. This means that the probability of any member of the population being selected for the sample is known and can be calculated. The most commonly used probability sampling techniques are simple random sampling, stratified random sampling, and cluster sampling.

On the other hand, non-probability sampling techniques do not use random selection to select the sample from the population. Instead, they rely on the judgement of the researcher or the availability of the subjects to select the sample. This means that the probability of any member of the population being selected for the sample is not known and cannot be calculated. Non-probability sampling techniques include convenience sampling, purposive sampling, quota sampling, and snowball sampling.

Another important difference between probability sampling and non-probability sampling is the extent to which the sample represents the population. Probability sampling techniques generally provide a more representative sample of the population, since every member of the population has an equal chance of being selected for the sample. Non-probability sampling techniques, on the other hand, may result in a sample that is biased towards certain characteristics of the population or certain types of individuals.

Statistical inference

Example 41: A company develops new type of a lightbulb. The population are all such lightbulbs produced. Study is conducted, inspecting 200 such lightbulbs (sample) and the number of hours each lightbulb operated before filament burnout is recorded. Suppose that the sample average lifetime for the lightbulbs is 76 hours. This sample statistics can be used to estimate the average lifetime for the lightbulbs in the population. Statistician usually provide statement of precision of such estimate, for example a margin of error ± 4 hours. Thus, an interval estimate of the average lifetime of new type of lightbulbs is 72 hours to 80 hours. The statistician can also state how confident he or she is that this interval contains the population average.

- details in course Statistics JEB105

Ethical guidelines

- be aware of the possibility of data errors in statistical studies which can result in misleading information
- errors can be limited using checks for internal consistency or via reviewing data with unusually large or small values
- be aware of the source as well as purpose and objectivity of the statistics provided

Ethical guidelines

- avoid unethical behaviour which can include:
  - improper sampling - e.g. running multiple tests until a desired result is obtained
  - inappropriate analysis of the data - e.g. discarding part of data to improve statistics
  - use of inappropriate summary statistics and misleading graphs
  - tendency to slant statistical work towards predetermined outcomes via unrepresentative samples
  - biased interpretation of the statistical results
- cf. American Statistical Association report “Ethical Guidelines for Statistical Practice”, 1999 (67 guidelines).