ICEF HSE

Probability Theoryand Statistics

Lecture 2: Introduction to Statistics. Organization of a statistical study

Populations and samples.

- A population is a complete set of all items that interest a statistician.

  Examples.

  - all people living in a country
  - grades of all students in a university
  - history of temperature observations in a city

- A sample is a subset of population available for analysis

  Examples.

  - a group of selected people
  - marks of 2 students from each academic group
  - temperature during the last year

A population is often too large for analysis and in this case the analysis is performed on a sample. A sample can accurately represent the whole population if the sample is chosen in a right way.

1. Types of statistical studies

Two different types of statistical studies:

- Observational studies
  Only collect data and analyze variables and relations between them

- Experiments
  Create differences in variables and analyze the result of such differences

Observational study *vs.* Experiment.

Observational study	**Experiment**
No assignment of actions	Choose groups and assign actions
Processing	Processing
Conclusions: only relationships	Conclusions: cause-and-effect relationships

Examples.

Observational studies	**Experiments**
Conduct a survey of worker's satisfaction	Try different bonus compensation scheme for workers
Collect data on demand curve	Offer discounts and evaluate the increase in demand
Collect statistics on patients' recovery time	Try different treatment schemes

2. Part I. Observational studies

The goal of an observational study is to describe some characteristics of the whole population (often, using for analysis a sample instead of the whole population).

Census vs. sample study.

A census is an observational study where each element of a population is studied .

In a sample study, only some subset of a population is analyzed.

Problems of a census.

1. Can be extremely cost and time consuming
   Example: national population censuses

2. The whole population may be not available
   Example: we cannot test new medicine on all the people – some are not ill at the moment.

3. Impossible if enumerating an element destroys it
   Example: check of electronic appliances lifetime

4. May contain errors due to poor qualification of the personnel

3. Sample studies

Sampling methods.

1. Simple random sampling (our main method).

Simple random sampling (SRS) is a method of data collection in which every possible sample of the desired size has an equal chance of being selected.

A SRS is identified with $X_1, \ldots, X_n$ i.i.d. random variables.

Disadvantages of SRS.

1. It may be too difficult to obtain the full list of a population
   Examples:
   a) all customers of a supermarket
   b) people with a specific disease (some don't know they have it)

2. Some elements of a sample may be difficult to reach, which will make the sample study too difficult and expensive
   Example:
   choosing a few elements from the population of a country will typically result in elements spread throughout the country

2. Systematic sampling.

Choose the first element randomly, then pick every tenth/hundredth/thousandth/. . . .

Example: a survey of supermarket's customers. Choose the first customer randomly, then every tenth entering the supermarket.

Advantage: can be performed when the full list of the population is not available or difficult to obtain

3. Stratified sampling.

1. The population is divided into K disjoint groups, called strata, of size N_1, N_2, $\dots$, N_K

2. Then a SRS of size n_k is chosen from the each stratum k

If all n_k are proportionate to N_k, it is called proportionate stratification. If not, it is called disproportionate stratification

Example: divide the population of students into academic groups and choose 2 students from each group.

Advantages of stratified sampling.

1. Can provide better precision ⇒ smaller sample size is required
2. Can guard against an “unrepresentative” sample (e.g., an all-male sample from a mixed-gender population)
3. Can be used to make study cheaper (include a fewer number of elements from strata difficult to reach)
4. Data can be used for a separate analysis of any subgroup

4. Cluster sampling

1. The population is divided in to K groups called clusters
2. Several clusters are chosen randomly

- One-stage cluster sampling: all elements from the selected clusters are included in the sample
  Example: choose randomly 2 of 10 academic groups
- Multi-stage cluster sampling Divide each selected cluster into new subclusters and select some of them randomly, . . . (repeat a required number of times)
  Example: choose random cities, then choose random districts of these cities, then choose random houses and finally SRS of tenants.

Difference between stratified and cluster sampling.

Stratified sampling: the sample includes elements from each stratum

Cluster sampling: the sample includes elements from selected clusters

Advantages of cluster sampling.

1. Can be cheaper than other methods
2. No need of a complete list of population, only a list of clusters
   Example: a list of houses in a town instead of a list of all residents

5. Quota sampling

Attempts to obtain a representative sample by specifying quotas on certain specified characteristics (age, gender, social class).

The (approximate) distribution of such characteristics in the population is required in order to replicate it in the sample.

Example: It is known that approximately 15% of visitors of certain restaurant are vegeterians. You want to know customers' opinion about chef's new course taking into account proportion of vegetarians.

Advantages of quota sampling:.

1. Accuracy - use known information about population features.

2. Time constains - if quotas are set some of them can be easier to fill.

Summary on sampling methods.

We use SRS because it gives i.i.d. random variables which are easy to work with mathematically.

There are other sampling methods, which may be more useful (but the mathematical theory for them is more difficult).

4. Review: observational studies

In an observational study we only observe characteristics of data.

Steps

1. Choose a sample
2. Study data
3. Make conclusions. What can we learn from a sample:
   (a) The structure of the sample (and the population)
   For example, what proportion of the objects belong to a particular group
   (b) Relationships between variables
   For example, the salary of a person and their education

5. Part II. Experiments

An experiment is a controlled statistical study.

In an experiment, a researcher manipulates (controls) one or more variables, and studies how they affect a response variable.

The goal of an experiments is to study causal relationships between the manipulated variables and the response variable.

Examples.

1. Salesmen compensation scheme
   $p\%$ of sales volume or a bonus if sold more than $\$x$?
2. Discount on products
   50% off or 2-for-1?
3. Medical studies
   Which flu medicine is more effective?

Parts of an experiment.

1. Experimental units – objects on which the experiment is performed. If units are people, they are often called participants.

2. Explanatory variables (also called factors; independent variables)
   The manipulated variables.

3. Response variable (also dependent variable)
   The affected variables, the effect on which is which are measured.

4. Treatments
   Combinations of independent variables to be tested.

6. Experimental design

A statistical experiment consist of the following steps:

1. Choosing experimental units
2. Diving experimental units into groups which will receive different treatments
3. Applying treatments
4. Comparing results and making conclusions

Experimental design is a plan for assigning treatments to experimental units (step 2).

7. Randomization: eliminating confounding factors

Confounded variables (or confounded factors) – when it is not clear which one causes the effect.

Example – testing two medicines

1. Group 1: young patients receiving Medicine A
2. Group 2: elder patients receiving Medicine B

We cannot say which treatment is better: the treatment and the age are confounded. We also say that the age is a confounding variable (or confounding factor).

Randomization is a strategy to reduce the possible influence of confounding variables.

A general rule: groups for different treatments should be maximally similar (ideally, they should differ only by the applied treatments).

1. Completely randomized design – participants are distributed between the treatments randomly:
   Example – test two treatments on 600 subjects

Treatment A	Treatment B
300 subjects	300 subjects

- Typically equal sizes of the groups (but not necessary)
- There may be more than 2 groups (if several treatments are tested)

2. Randomized block design
   (a) Experimental units are divided into subgroups, called blocks
   (b) Experimental units within each block are randomly assigned to the treatment groups

	Treatment A	Treatment B
Male	100	100
Female	100	100
Child	100	100

Ideally, each combination of possibly confounding variables should be in a separate block (this may be be too difficult, if the number of them is large).

3. Matched design

Each block contains only 2 experimental units – 1 for Treatment A and one for Treatment B (or n experimental units if n treatments are considered).

The elements of each block are chosen to be the most similar. Often, they are the same experimental unit but at different moments of time.

Pair	Treatment A	Treatment B
1	1	1
2	1	1
...	...	...
300	1	1

8. Reading

Newbold, Carlson, Thorne: Chapter 17.

9. Bias in sample studies

Bias (in the general meaning of this word) is a result of a poorly designed study which leads to incorrect interpretation of the results.

Types of bias:

- Selection bias – "choosing a wrong sample"
- Response bias – "asking wrong questions" (in sample surveys)

10. Selection bias

Examples

1. An internet survey on a website ignores those who don't visit the website (undercoverage bias)
2. Survey of students performed during a lecture ignores those who don't attend lectures (undercoverage bias)
3. Surveys sent by e-mail (non-response bias)
   Those who don't have time will not reply
4. Call-in TV or radio shows (voluntary response)
   Often only people with active position will participate
5. Many founders of large corporations dropped out of university. Should average students do that? (survivorship bias)

11. Response bias

Response bias can occur when the participants of a survey are asked wrong questions.

1. Non-anonymous surveys
   Participants do not want to show their point of view in controversial topics
2. Non-neutral questions
   First persuade, then ask a question.
3. Fear of consequences