From the Expert: The Statistical Process

From the Expert: The Statistical Process

The statistical process begins with formulating the problem, collecting the data, analyzing the data, and end with conclusions.

Formulation

Using complicated or inappropriate data analysis without fully understanding the objectives can lead to meaningless results. Thus, it is essential to formulate the problems correctly. This involves understanding objectives, understanding the background and subject areas, understanding the need of the clients, and translating the problem into statistical terms.

Data Collection

Data can be divided into primary data, and secondary data. Primary data are data collected to address the problems currently under study or research. Secondary data are data collected for different purposes but may be related to investigating problems.
Primary data can be obtained from surveys, observational study, and experimental study.

- Surveys gather information from people such as marketing surveys and exit pools. There are various ways to do the survey such as personal interview, phone interview, self-administered questionnaire (e.g. Internet, by person, by mail). The proportion of selected people who complete the survey is called response rate. Some important principles for designing the survey questions are questions should be simple, short and clear, start with demographic data for the comfortable start, use yes/no or multiple choice questions, open-ended questions should use with cautious and avoid leading questions, keep questionnaire short, and think about how you can collect the data from questionnaire.
- Experimental study – Subject conditions are controlled by the investigators. Typically, treatments are used to observe the response.
- Observational study – There is no attempt to control or influence the subjects of interest.

Secondary data can be obtained from internal sources (e.g. company database) or external data sources (e.g. government, non-government, industries publications, Newsgroup, Newsletters, Journals, Periodicals, Internet). Some advantages of using secondary data are low cost, less effort, and available immediately. Disadvantages are inaccuracy, lack of relevance, and lack of availability.

As a reminder, population is the complete set of data and ***parameters*** are terms to summarize characteristics of populations. Sample is the subset of the population and ***statistics*** are terms to summarize characteristics of samples.

Descriptive statistics are methods to make conclusions on the presented data but not beyond using tables, graphs, numerical summaries of central tendency measures (e.g.

Descriptive statistics are methods to make conclusions on the presented data but not beyond using tables, graphs, numerical summaries of central tendency measures (e.g. mean, mode, median), variability measures (e.g. standard deviation), or position measure (e.g. quartiles, percentiles). Inferential statistics (e.g. t-test) utilize sample data to make conclusions about the populations.

In statistical inference, the conclusion about the population can be drawn based on samples. Usually, a subset of the population is selected or sampled due to practical reason of time and cost. For example, it is not reasonable to survey 10 million stream viewers. Sampling is a procedure for choosing the representative subset of data. There are several types of sampling methods.

Sampling methods

- Simple random sampling
  The sample is selected from the population, where every possible sample has an equal chance to be chosen. For example, a random sample of 10 cars from the car factory. This is the most common method.

- Sampling with replacement
  Put the sample back in the population after being selected. Thus, a single sample can be selected more than once.

- Sampling without replacement
  Do not put the sample back in the population after being selected. So, there is only one chance of a sample to be selected.

- Stratified random sampling
  The population is divided into mutually exclusive sets, or strata (groups). Then, do simple random sampling from each stratum. This method, typically, is employed when all sub-populations are needed in the sample.

- Cluster sampling
  First divide the population into clusters, then simple random sample from the clusters. This method is used when the populations are widely dispersed geographically. It may increase sampling error because of similarity among the cluster menbers.

- Systematic sampling
  Select one of the first kth sample, then subsequently select every kth sample.

- Convenient sampling
  Use data that is already available or convenient to obtain.

Sources of Errors

Errors can come from sampling or estimation errors and non-sampling errors. Sampling error occurs when samples do not give a complete picture of population. Non-sampling errors include response errors (e.g. poor questionnaire questions, poor survey process, interview bias), non-response errors (e.g. incomplete answers or no response questions), processing errors (e.g. coding error, computation errors), and analysis errors (e.g. poor choices of statistical test), etc.

Data preprocessing

Real life data is dirty, noisy, inconsistent, and may be missing. The original data form may not suitable for analyzing or learning. Thus, some preprocessing techniques are necessary. Generally, the preprocessing techniques consist of:

- Data cleaning – clean the data by filling missing values, resolving inconsistencies, removing outliers, etc. This may have to be done manually and usually it takes a lot of time.

- Missing values: there are several ways to handle the missing values such as ignore records with missing values, fill in the missing values with attribute mean, attribute mean of the same class, predicted value by a predictor (e.g. kNN), etc. The process of estimating or deriving values where data is missing is called

imputation.There is bias in the data for these methods. But, small noise may be a good compromise than to exclude all the examples.

- Noisy data -- Noise happens when there is error in measured variables. Here are some methods for smooth or remove noise such as binning (e.g. group closer data together), clustering (similar data are organized into groups, outside groups may be outliers), regression (smooth data with a function), and manually inspect the data.
- Data integration – multiple database and files may be combined. Issues such as schema integration, different data types, different units, different names (e.g. Bob and Robert), and redundancy need to be resolved.
- Data transformation – the difference in measurement units can make one feature dominate the other feature, normalizing data into the range [0,1] will help the analysis techniques work better.
- Data reduction – massive datasets may be reduced using techniques such as dimensionality reduction (e.g. principle component), and the predictive power still remain.

The processed data is ready for data analysis. After the cleaning process, we will get a *Tidy data set*. Here, each variable is in one column, each observation is in different row, etc.

Initial Data Analysis

This step may looks trivial but it should always be performed. Exploratory data analysis (EDA) goal is to understand the data using visual techniques such as plots and graphs. Data exploration examples using R can be found at:

Practical Regression and Anova using R

https://web.archive.org/web/20190925232817/https://sux13.github.io/DataScienceSpCourseNotes/3_GETDATA/Getting_and_Cleaning_Data_Course_Notes.html

Experimental Design or Design of Experiments (DOE)

DOE is a study to determine the effects of one or more variables (known as explanatory variables or factors) on another variable (known as response variable). The data is collected from subjects or experimental units. Treatments are conditions assigned on the explanatory variables. Three main types of experiments are:

1. Completely randomized design. Each experimental unit is randomly assigned to a treatment.
2. Randomized complete block design. The subjects are divided into homogeneous group or block, and are randomly assigned to a treatment group.
3. Factorial design. The effects of factors and their interactions are studied.

Research Design

Design research study is one of the important areas in statistics. Being familiar with the custom and practice of your profession is a necessity such as what type of study can answer research questions or is appropriate for a particular data type. Research design is a large subject. This is the introduction to common types of designs. A perfect design is not always possible. In general, there is always a compromise between what researchers want to do and the feasibility (e.g. resource, ethic). The choice should be guided by research objectives, standards and traditional practices in the field.
There are 3 types of research designs:

Experimental

In experimental design, subjects are randomly assigned to groups. For instance, in drug administration studies, subjects are assigned to experimental and control groups randomly. The control group is considered the strongest type of research design and a gold standard for evidence; however, it is not always practical or possible to conduct this

randomly. The control group is considered the strongest type of research design and a gold standard for evidence, however, it is not always practical or possible to conduct this study.

Quasi-experimental design

Here, the subjects are not randomly assigned to the groups. The pre-existing groups such as classroom, hospital, etc. are used instead of random assignment. The group with no intervention is called the comparison group. This design is popular in fields of research such as education and social science, where data is collected in their normal environmental setting rather than in experimental setting.

Observational

Subjects are observed in real situations without any intervention. This study is conducted when experimental study is not possible. For example, a study about the effects of smoking on health can be done only as an observational study since it is not ethical to assign people to smoke tobacco, which may be harmful to their health.

Basic vocabulary

- *Factor* - an independent variable (also known as predictor variable) that has influence on the dependent variable (outcome variable)
- *Control group* – participants who are not exposed to the experimental condition
- *Experimental group* – participants who are in experimental condition
- *Experimental condition* – a set of conditions created by investigators to test the impact of different levels of the independent variable
- *Single-blinded* – the participants do not know what group (treatment or control) he/she has been assigned to
- *Double-blinded* – Both the participants and the treatment allocator do not know what group the participant belongs to
- *Triple-blinded* – The participant, treatment allocator, and data gatherer have no idea what group the participant has been assigned to
- *Cohort study (prospective)* – an observational study that follows subjects to collect information about them for analysis into future (forward in time). This usually deals with large groups of individuals
- Case control design is often used to study diseases that take a long time to develop (e.g. 20 – 30 years or longer). A study begins with people with the diseases (called cases), and people without disease (called controls), who are similar to the cases in many ways but without disease. The goal is finding a factor (factors) that can explain why diseases happened with cases but not the controls. This is considered as observational study.
- Ecological study is the study where unit of analysis is not individuals but larger units such as school, country, etc.
- A cross-sectional design involves a snap shot of the state of individuals at a particular time.

The role of statistics in experimentation can be divided into 1) project planning involves what to measure, what are the factors, and how large are the likely variations 2) experimental design relates to controlling known sources of variation, estimating the uncontrollable variation size, and investing the suitable models 3) statistical analysis phase includes make inferences on factors, suggest the appropriate and subsequent design.

Cook & Campbell developed the notation for research design. The notation is simple and flexible. It continues to be popular among researchers.

`X` = intervention (or independent variable); `O` = observation (measurement)
`R` = random assignment or random selection (control); `N` = non-random (comparison)
Subscripts – the order of interventions or observations
Dash line – groups without randomization

Example:

Randomized pretest-posttest design

In this notation, subjects are randomly assigned to treatment and control groups. Measurements are performed from both groups, then a treatment (or intervention) is given to the first group but not the second group (aka control group), and finally, the measurements are performed on both groups again. This design is typically used in medical studies. The intervention for the first group can be drug or other treatments. The control group is administered with no treatment (placebo) or standard treatment.

Quasi-experimental pretest-posttest design

In this notation, subjects are not randomly assigned. Instead, preexisting groups such as classrooms, hospitals, or regions are used. Measurements are performed on both groups, initially. Then, the intervention is given to the first group or the experimental group, whereas, the second group or comparison group receives no intervention. Finally, the measurements are performed on both groups again.

References:

Mason, L.R., Gunst, F.R., and Hess, L, J. (2003) Statistical Design and Analysis of Experiments, 2nd Edition, Wiley.

Faraway. J.J (2002). Practical regression and ANOVA using R. Retrieved from: https://web.archive.org/web/20190925232851/https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

Activity Details

Task: View this topic