1. Types of Data

1) Qualitative data

: data is not given numerically, and can be observed but not measured.

i) Nominal data

- no inherent order in categories

- place of birth, favorite color, type of car, eye colour, ethnicity, diagnosis, gender

ii) Ordinal data

- categories have inherent order

- job grade, age groups

2) Quantitative data

: data can be measured, and is given numerically.

i) Count data

- data that can only take specific numeric values

- number of sisters, number of car in a parking lot

ii) Continuous data

data that take any numerical values

- height, length, weight, temperature, age

[Note] it is important to understand that appropriate statistical methods for correct statistical inferences depend on the type of data being used.

2. Dependent vs. independent variable

- Independent variables are those that are manipulated whereas dependent variables are only measured or registered.

- The terms dependent and independent variable apply mostly to experimental research where some variables are manipulated, and in this sense they are "independent" from the initial reaction patterns, features, intentions, etc. of the subjects. Some other variables are expected to be "dependent" on the manipulation or experimental conditions. That is, they depend on "what the subject will do" in response.

For example, suppose that a vaccine against a contagious disease is introduced. One is interested in comparing the disease incidence rate among people taking this vaccine to that among people without using this vaccine. Here, incidence rate is a variable, and vaccine is an variable.

- These terms are also used in studies where we do not literally manipulate independent variables, but only assign subjects to "experimental groups" based on some pre-existing properties of the subjects. For example, if in an experiment, males are compared with females regarding their white cell count (WCC), Gender could be called the variable and WCC the variable.

- Independent variable (X) : explanatory variable, predictor variable, factor

- Dependent variable (Y): response variable

3. Types of Studies

- Importance of study design: a well designed study is simple to analyze; while a poorly designed study often can not be saved even with the most sophisticated analysis.

  (Vaccine example) Suppose that a vaccine against a contagious disease (e.g., flu) is introduced. How should the investigators test its effectiveness? A basic method is to compare the disease incidence rate among people taking this vaccine to that among people without using this vaccine.

  How should the study be conducted?

  Design 1) 200 kids were recruited. The investigators got parents permission for 80 kids and vaccine were given to them; and for the remaining 120 kids, no vaccine were given. At the end of the year, they observed:

  Design 2) 200 school kids were recruited for the study. 100 of the kids were <u>randomly chosen</u> and assigned to the <u>treatment group</u> who were given the vaccine; while the other 100 kids were assigned to the <u>control group</u> and they were given placebos. At the end of the year, the disease occurrence status for each kid was recorded which showed:

Design 1			Design 2		
Treatment	Size	Incidence rate	Treatment	Size	Incidence rate
Vaccine	80	20%	Vaccine	100	28%
No vaccine	120	85%	No vaccine	100	71%

The main difference between two designs :

Randomization of treatments to experimental units (kids) occurs in Design 2), not Design 1).

The randomization would tend to balance out the differences in other independent variables, leaving the observed differences attributable to Vaccine under study. Thus, we might know if there is a cause-and-effect relationship between Vaccine and Incidence rate.

- Observational study : a study where the researchers have no control over exposures and instead observe what happens to groups of people.

  An observational study usually can only answer whether there is an association between the explanatory variable(factor) and the response variable. In general, external evidence is required to rule out possible alternative explanations for a cause-and-effect relationship.

  (e.g.) Design 1)

- Experimental study : a study where the conditions are under the direct control of the researcher. Usually this involves giving a group of people an intervention that would not have occurred naturally. Such studies are often used to test the effects of a treatment in people and usually involve comparison with a group who do not get the treatment.

An experimental study aims to answer the question: whether there is a cause-and-effect relationship between the explanatory variable(factor) and the response variable

(e.g.) Design 2)

- Why observational studies at all then? It is sometimes impossible to do randomization (i.e., to do experiments on the experimental units), due to many practical reasons(say ethnical reasons, practical feasibility, etc).

  (e.g.) the study of effects of smoking on lung cancer incidence rate; the study of effects of education on income

- Be cautious on your conclusions/interpretations: sometimes the data and even the statistical analysis could be very similar for an experimental study and for an observational study. However, the interpretations of the results are usually different. You usually can only make a statement on association for an observational study and should be very cautious on drawing any conclusions about cause-and-effect relationship.

4. Regression and Analysis of Variance

1) Regression(Chapter 11)

- Learn more about the relationship between one or more independent or predictor variables and a dependent variable.

: For example, a real estate agent might record for each listing the size of the house (in square feet), the number of bedrooms, the average income in the respective neighborhood according to census data, and a subjective rating of appeal of the house. Once this information has been compiled for various houses it would be interesting to see whether and how these measures relate to the price for which a house is sold.

: In the social and natural sciences regression procedures are very widely used in research. In general, regression allows the researcher to ask the general question "what is the best predictor of ...". For example, educational researchers might want to learn what are the best predictors of success in high-school. Sociologists may want to find out which of the multiple social indicators best predict whether or not a new immigrant group will adapt and be absorbed into society.

2) Analysis of Variance(ANOVA) (Chapter 12 and 13)

   - Test for significant differences between means from several populations.

   - t-test(comparison of means from two population) is a special case of ANOVA(comparison of means from several populations)

   : For example, a manager wishes to determine whether the mean times required to complete a certain task differ for the three levels of employee training. He randomly selected 10 employees with each of the three levels of training (Beginner, Intermediate and Advanced).

   - we are actually comparing (i.e., analyzing) variances in order to test for statistical significance between means(so we call this ANOVA)

3) Regression and ANOVA are linear models.