Introduction to Statistics

Dr. Edgar Acuna

http://academic.uprm.edu/eacuna

DEPARTAMENTO DE CIENCIAS MATEMATICAS
UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGUEZ

Why to study Statistics/Data Science/Data Mining/Bigdata

1. Data is available everywhere.
2. Regardless your field of interest, very frequently you have to take decisions based on data. Therefore, learning statistical methods to analyze data help us to make decisions more effectively.
3. To work with a small sample may be more efficient that working with the whole data (Big data). In particular, when the large data has not been cleaned up.

Why to study Statistics/Data Science/Data Mining/Bigdata

Data science: Modern Statistics closer to Computer Science than Math. It deals with data of several types not only numerical as in Statistics. Analysis of Big data.

Data Mining: Statistics applied to large datasets. Besides Statistical methods, it includes Machine Learning algorithms to analyze data. It uses high performance computing for analyzing large databases.

Big Data Analysis. Modern Data Mining with intensive use of modern computational tools (Hadoop, Spark) and powerful hardware (clusters).

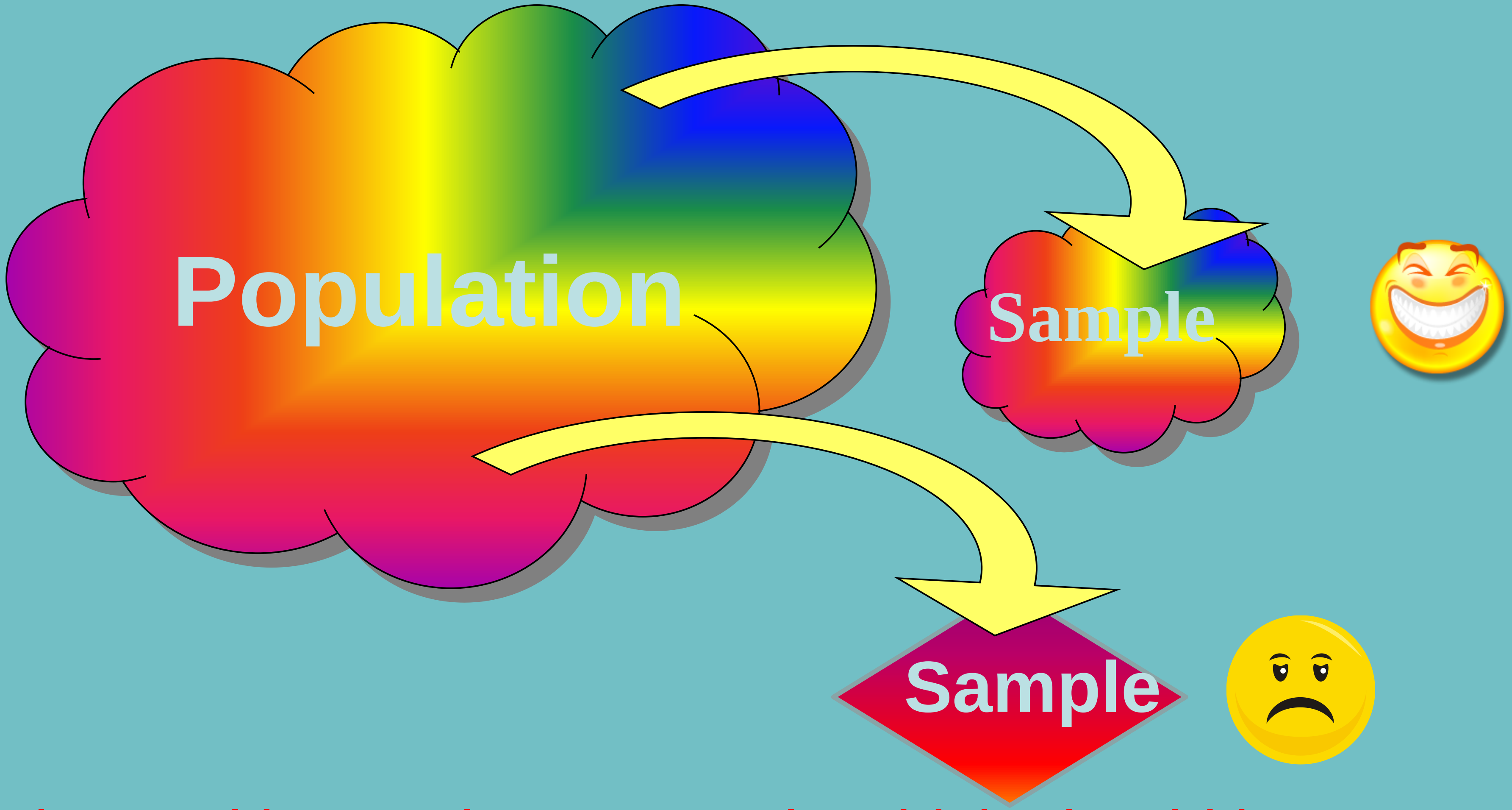
Basic Statistical concepts

a) **Population:** It is a set of subjects or objects having the characteristic which is the target of the study. More precisely, a population *is the set of measurements* of some characteristic taken in all the subjects or objects having such characteristic.

b) **Sample:** It is the set of measurements that are really gathered. The process of drawing the sample is a very important step since from there conclusions about the population are reached. If the process to draw the sample is a simple one, then the sample size should be about 10% of the population size.

c) **Random Sample:** It is a sample that very well represents the population. Each element of the population has the same chance to be selected in the sample. The conclusions based on a random sample are reliable.

A sample without/with bias



The goal is to select a sample which should be representative of the population.

The sampling bias

Sampling bias occurs when the method used to select a sample generates sample does not represent the population.

- When there exists sampling bias, then the conclusions about the population drawn from the sample are unreliable.

The power of random Sampling

- Before the 2008's US presidential election, the Gallup Poll drew a ***random sample*** of 2,847 voters out of approx. 150 millions voters. It found that 52% of **voters** support Obama.
- The result of the election was 53% voters in favor of Obama.
- Before the 2016's election, Gallup gave a 5% advantage to Clinton over Trump.

2016 a bad year for Stats

The Brexit Referendum: June 26
Prediction: 4 days before the event: 52% remain in EU, 48% leave the EU.
Result: 52% leave, 48% remain.

The FARC (Colombia) peace deal Referendum: October 2
Prediction:, one week before the event: 66% in favor to accept the peace deal.
Result: 50.2% reject the peace deal.

US Presidential Elections: November 1
Predictions, three days before the event: Clinton wins by 4-5 %
Result: Clinton lost.

The Trump's case: Why the polls failed ?

Days before the election 18 of 20 polls predicted Clinton's victory over Trump. Some of them gave a winning margin of more than 15%. Only two polls, one from USC/LA Times and IBD/TIPP predicted the right result. Among the reasons for the failure are:

1-the models lost power of prediction as the time pass.
2-Most of the polls tend to lead toward the current trend
3- The irrational behavior of the voters was not measured.
4-Look for information beyond the one given by the polls.
5- Not appropriate sample size.
6-The social network effect (Sentiment Analysis-Twitter)

Statistical Basic Concepts (cont.)

d) Variable/Feature: It is the characteristic under study: Age, Weight, Salary, Temperature, etc.

e) Datum: It is a particular value of the variable.

f) Parameter: It is a value that caracterizes to a population The parameter has a constant value, which in general is unknown.

g) Statistic Value: It is a value that is computed using the data available in the sample.. Its value varies according to the sample that was taken, and it is used to estimate the parameter. Th statistic value is always known.

Statistical Basic Concepts

h) Census: It is a list of one o more characteristic of all the elements of a population. There are several type of census, the most well-known is the Census of Population that is performed each 10 years.

i) Survey: It is a list of one o more characteristic of all the elements of a sample.

Definition of Statistics

Statistics is the science where we learn about a population from the information gathered in a sample drawn from such population. Statistics involves the methods used to collect the sample, to organize and present the collected data, and to reach general conclusions based on the collected data.

Components of Statistcs

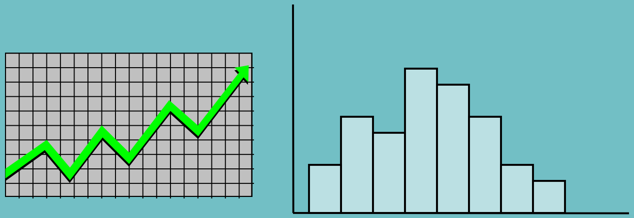
- **Descriptive Statistics:** It includes techniques and methods used to collect, to organize, and to present numerical information in tables and graphs. Also, it includes the computation of statistical measures of central tendency and variability.

- **Inferential Statistics:** It includes techniques and methods used to draw general conclusions about a population using the information given by a sample collected from the population.

Descriptive Statistics

- Data gathering
  - Ex. Surveys

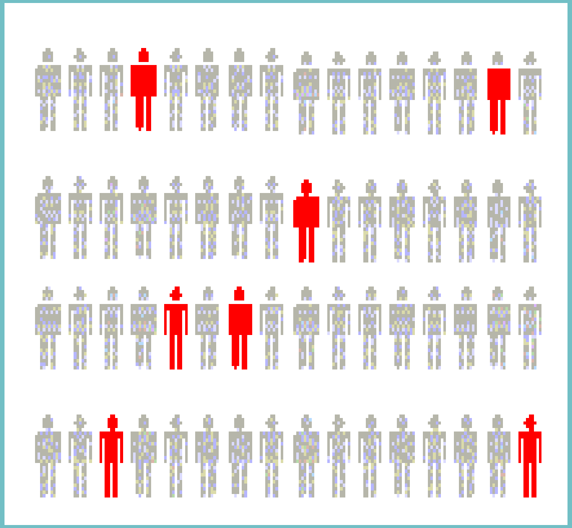
- Data presentation
  - Ex.Tables and graphs



- Data summarization
  - Ex. Sample mean= $\dfrac{\sum X_i}{n}$

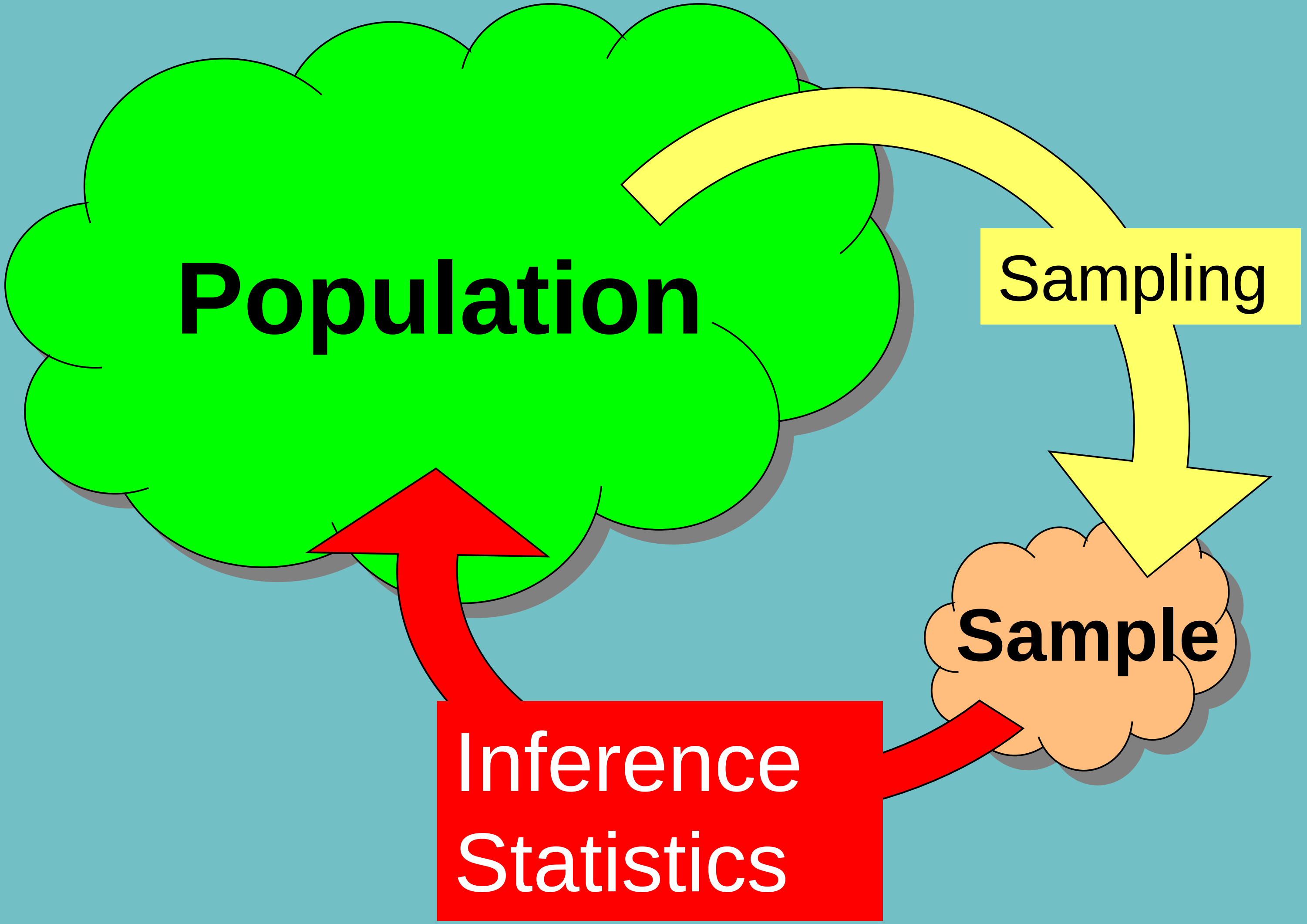
Inferential Statistics

- Estimation
  - Ex. Estimate the population mean salary using the sample mean.
- Hypothesis Testing
  - Ex. Test the clain that the population mean salary is greater than 20k per year.



Inference is the process of drawn conclusions or take decisions about a population based on the results of a sample taken from the population.

How Statistics works



Type of Data

A. **Quantitative Data:** data coming from taking measurements or from countings.

There are two types:

A1. Discrete Data: data coming from counting and in general assumes integer values.

A2. Continuous Data: data coming from measurements and asumes any real value numbre.

Type of data

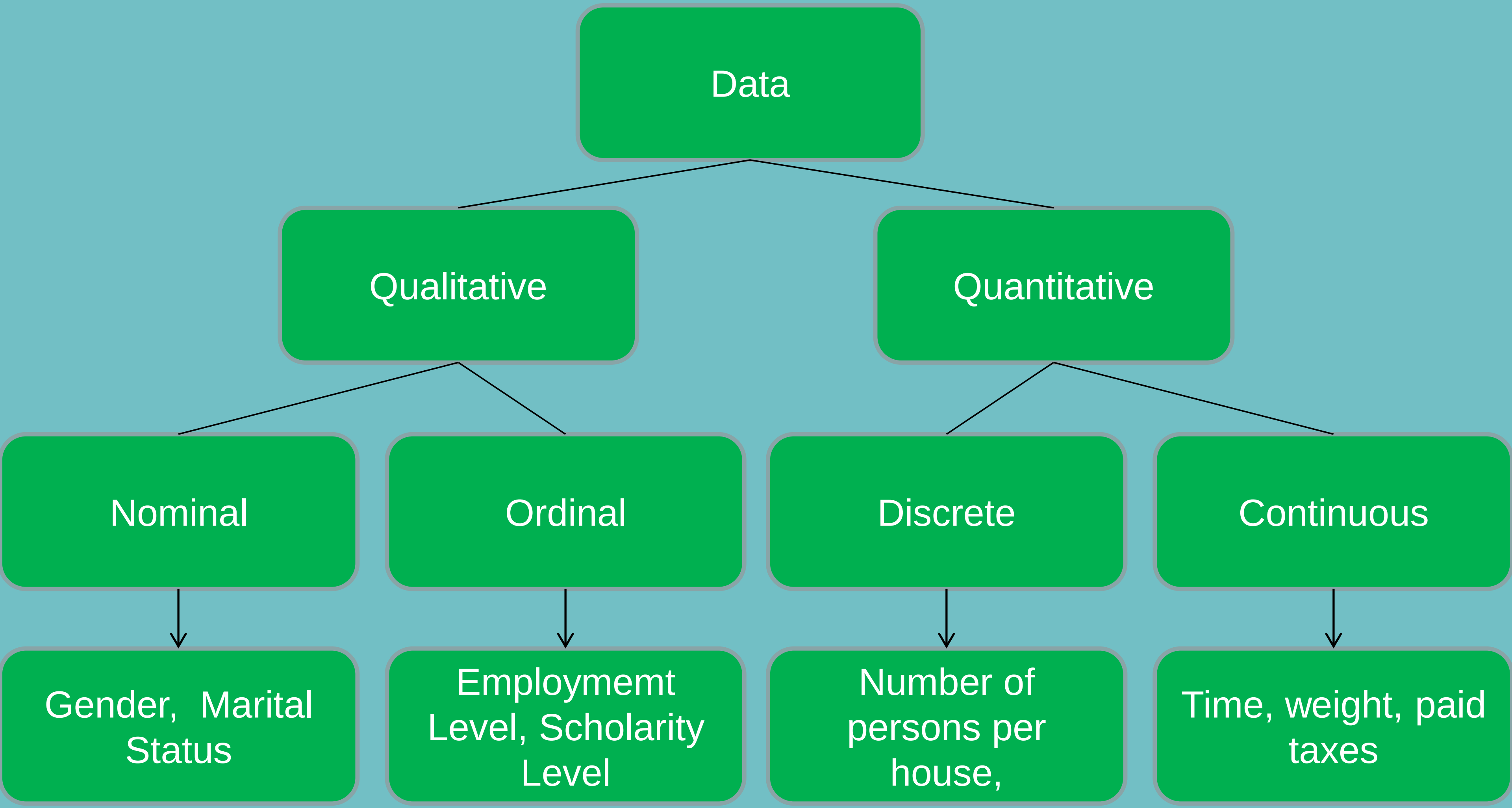
B. Qualitative (Categorical) Data: represents categories. In order to facilitate the statistical analysis this data is commonly numerically coded.

This codification generates two types of categorical data:

B1. Nominal Data: In this type of data there is an arbitrary relationship between the value of the categorical value and its corresponding numerical value.

B2. Ordinal Data: There exists a correspondence between the value of the categorical attribute and its assigned numerical value .

Type of Data



Data

Qualitative

Quantitative

Nominal

Ordinal

Discrete

Continuous

Gender, Marital Status

Employmemt Level, Scholarity Level

Number of persons per house,

Time, weight, paid taxes

Sampling Techniques

a) **Random Sampling**. Each element of the population has the same chance to be selected in the sample.

b) **Stratified Sampling**. Used when it is known beforehand that the population is divided in strata (categories) that usually do not have the same size. Then, from each strata we randomly collect data in such way that in the sample there are data from the different strata in the same proportion as in the population.

c) **Sampling by Clusters.** In here, the popuation is divided in several groups called clusters. A certain number of clusters is chosen randomly and then all the elements of the clusters are chosen to be part of the sample. It is used in surveys for market research.

d) **Sistematic Sampling.** It is used only when the members of the population can be uniquely identified with an ID. Assuming that the elements of the population are ordered by its ID, only the first observation of the sample is chosen randomly among the first elements of the population, the next elements are chosen following an equally spaced scheme.

Ways to collect Data

a) Personal interviews. Perhaps it is the most effective method but it is expensive and it requires to spend a lot of time.

b) Interviews by phone. The major disadvantage is that most of the time one cannot get sincere answers from the person beeing interviewed.

c) Questionaries by mail. It is expensive and usually no more than 30% of questionaries are retuned. (Alternatives: SurveyMonkey and Google Forms).

d) By direct observation.

e) Through the internet. Plenty repositories of data.

f) Using computer simulation