Introduction to Statistics

Lecture 1: Introduction

By Mr. Thuo

Brief overview

What *are* statistics?

What *is* statistics?

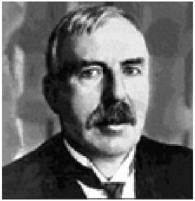
Why you *should care about* statistics?

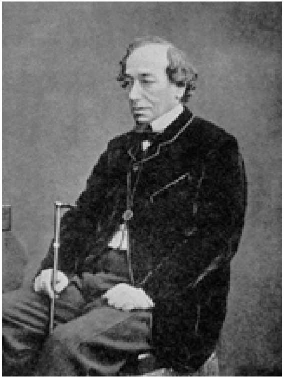
What kinds of questions can statistics help to answer?

Some opinions of statistics

"If your experiment needs statistics, you should have done a better experiment."

Ernest Rutherford





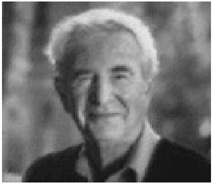
"There are three types of lies: lies, damn lies, and statistics!"

Benjamin Disraeli

Some opinions of statistics

"To call in a statistician after the experiment is done may be no more than asking him to perform a postmortem examination; he may be able to say what the experiment died of."

Sir Ronald Fisher



"The purpose of models is not to fit the data, but to sharpen the questions."

Samuel Karlin

Objectives of Statistics course

- Understand the fundamental principles of Descriptive statistics and statistical inference.
- To analyze data and apply appropriate statistical methods
- Know the assumptions of common tests and understand impact of violations.
- Learn to understand statistical results (understand statistics in all publications)
- Learn to design experiments for research and clinical studies
- To judge statistical results from a critical point of view
- Learn to use R, a free software environment for statistical computing and graphics

What is meant by statistics?

Literally, summary of information (data) in a meaningful fashion, and its appropriate presentation

More broadly, the discipline of drawing conclusions from data

- designing studies or experiments
- estimating unknown quantities
- quantifying uncertainty
- developing and applying a formal framework for drawing conclusions
- communicating and explaining results

This page was intentionally left blank By Mr Thuo

What is Statistics?

Collecting data: design of studies and instruments

Summarizing data: numerical and graphical representations of data

Analyzing data: concern of most of this course

Uses of statistics:

- Collect and use empirical data efficiently to gain the most value with the least cost
- Use empirical data to describe the world around us.
- Interpretation of data
- Characterize replicable processes
- Distinguish random noise from pattern

It all starts with DATA!

Obtain DATA

Analyze DATA

Present DATA

A few thoughts

- Beware it is often difficult to express a property as a number;
- make sure you define your data precisely or accurately;
- your data are only as good as your measurement process
- record your data accurately and unambiguously – a '?' in your data collection form is not helpful!
- There are different types of data, and you can't analyze data without knowing what type you have!
- Good data can be analyzed and summarized to provide useful information
- Bad data can be analyzed and summarized to provide incorrect/ harmful/non-informative information

Steps in Research

- Planning/design of study
- Data collection
- Data analysis
- Presentation
- Interpretation

Statistics Issues

Planning/design of studies

- Primary question(s) of interest:
  - Quantifying information about a single group?
  - Comparing multiple groups?
- Sample size
  - How many subjects needed total?
  - How many in each of the groups to be compared?
- Selecting study participants
  - Randomly chosen from "master list?"
  - Selected from a pool of interested persons?
  - Take whoever shows up?
- If group comparison of interest, how to assign to groups?

Statistics Issues

- Data collection
- Data analysis
  - What statistical methods are appropriate given the data collected?
  - Dealing with variability (both natural and sampling related):
    - Important patterns in data are obscured by variability
    - Distinguish real patterns from random variation
  - Inference: using information from the single study coupled with information about variability to make statement about the larger population/process of interest

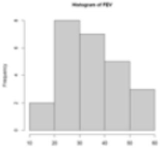
Statistics Issues

- Presentation
  - What summary measures will best convey the ”main messages” in the data about the primary (and secondary) research questions of interest
  - How to convey/ rectify uncertainty in estimates based on the data
- Interpretation
  - What do the results mean in terms of practice, the program, the population etc.?

Classification of statistical methods

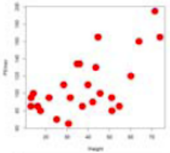
Univariate methods

Each variable is considered individually



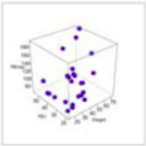
Bivariate methods

Relation between 2 variables is studied



Multivariate methods

Relation between >2 variables is studied



Definitions

Parameter: a numerical measurement describing some characteristic of a population

Statistic: a numerical measurement describing some characteristic of a sample Population Sample Parameter Statistic

Sample vs. Population

Parameters and Statistics - Experimenters normally use sample statistics as estimates of population parameters.

Population parameters are written with Greek letters; sample statistics with Latin letters.

Two main types of Statistics:

- Descriptive Statistics summarize or describe the important characteristics of a known set of population data
- Inferential Statistics use sample data to make inferences (or generalizations) about a population

Descriptive statistics

- Collection, organization, summarization, and presentation of data.
- Used to describe the main features of a collection of data in quantitative terms.
- Aims to quantitatively summarize a data set
- Some statistical summaries common in descriptive analyses.
  - Frequency Distribution
  - Central Tendency
  - Dispersion
  - Association

Definition

A *Variable* is an attribute that describes a person, place, thing, or idea. Its value can "vary" from one entity to another.

Random Variable: Variable whose values are determined by chance. Can be thought of as an unknown value that may change every time it is inspected

Dependent or response or outcome variable: a variable that depend on other variable/s. .

Independent or predictor or explanatory variable: variable that does not depend on other variable/s

Confounding variable: variable/s that correlates (directly or inversely) with both the dependent and independent variable.

Inferential statistics

- Used to make an inference, on the basis of data, about the (non)existence of a relationship between the independent and dependent variables.
- Used to generalize from samples to populations using probabilities.
- Performing hypothesis testing, determining relationships between variables, and making predictions.

Bias

- In survey sampling, bias refers to the tendency of a sample statistic to systematically over- or under-estimate a population parameter,
- Types:
  - Selection bias
  - Nonresponse bias
  - Under-coverage bias
  - Voluntary response bias
  - Measurement bias
  - Leading questions
  - Social desirability

Misuse of statistical analysis

- Obsession with statistical recipes, in particular, hypothesis testing ? demanding statistical significance test;
- Use of statistical techniques as a ?black-box?, or cook-book recipe (standard example is disregard of serial correlation).
- Misunderstanding or misinterpreting the names (e.g. p-values as probability of hypotheses)
- Use of sophisticated techniques... There is sometimes unwarranted expectation of miracle-like results from very advanced techniques.

Abuses of Statistics

- Bad Samples
- Small Sample
- Loaded Questions
- Misleading Graphs
- Distorted percentages
- etc

What statistics can and can’t do

Can

- provide objective criteria for evaluating hypotheses
- synthesize information (not without information loss… keep your raw data!)
- help detect patterns in messy data
- help optimize effort
- help you critically evaluate arguments

Can’t

- tell the truth (probabilistic conclusions only!)
- compensate for poor design
- indicate biological significance: statistical significance *does not* mean biological significance, nor *vice versa*!