# Lecture 1
# Formulating Research Problems and Constructing Hypotheses

Ioannis Pavlidis
Dinesh Majeti

Computational Physiology Lab

*ipavlidis@uh.edu*
*dmajeti@uh.edu*

January 19, 2018

# Inductive vs. Deductive Methods

- An **inductive** method aims at generating a new theory from the data
  - Exploring new phenomena
  - Completely open minded - no preconceived ideas

- A **deductive** method aims at testing a theory using the data
  - Hypothesis driven with emphasis on association

# Examples of Inductive and Deductive Research

- **Inductive Research:** Data mining
  - Bottom-up process that discovers patterns in big data
    - Explore massive genomic data bases for a defective gene
    - Discovering customer preferences in online purchases

- **Deductive Research:** Effectiveness of a new drug
  - Randomized control trial on a patient sample
    - Half receiving drug and the other half receiving placebo
    - Test if the drug sample fared better than the placebo sample

# Hypothesis-Driven Research

- Hypothesis-driven research is most appropriate when a researcher is trying to decide between a small number of mutually exclusive cases
  - Finding which of two network configurations is more effective

- Hypothesis-driven research is NOT appropriate when the researcher is trying to decide among a vast number of possible cases
  - Testing the hypothesis that the density of a rock is $1 \text{ g/cm}^3$

# Null and Alternative Hypothesis

- **Control** or **Independent** variable: Something that for sure will change during the course of experimental measurements

- **Response** or **Dependent** variable: Not sure if it will change

- **Null Hypothesis**: Changes in the independent variable will not lead to any changes in the dependent variable

- **Alternative Hypothesis**: Changing the independent variable does change the dependent variable

# Statistics

- **Descriptive Statistics** is concerned with the properties of the observed data
  - No assumptions about the underlying population

- **Inferential Statistics** is concerned with the underlying population of a sample
  - First select a statistical model and then draw a statistical inference

# Statistical Model

- A **statistical model** is a set of assumptions concerning the generation of the observed data
  - **Fully Parametric Models:** The probability distributions involve a finite number of unknown parameters
    - e.g., normal with unknown mean and variance
  - **Non-Parametric Models:** Minimal assumptions
    - e.g., sample median
  - **Semi-Parametric Models:** Somewhere in between

# Statistical Inference

- **Statistical Inference** is the process of deducing properties of an underlying distribution via data analysis

- The outcome of statistical inference is a statistical proposition
  - Point estimate (for a parameter of interest)
  - Interval estimate (for a parameter of interest)
  - Credible interval (e.g., 95% of posterior belief)
  - Rejection of hypothesis
  - Classification (grouping of data points)

# Experiments

- **Comparative Experiments:** Designed to measure changes in an environment
  - Comparison of fertilizers, computer networks, or medications

- **Absolute Experiments:** Involve the measurement of quantities that are assumed to be constant
  - Velocity of light or mass of an electron

# Observations

- **Observation** is the active acquisition of information from a primary source
  - In living beings, observation employs the senses
  - In science, observation can involve the recording of data via instruments

# Observational Studies

- An **observational study** draws inferences about the possible effect of a treatment on entities, where the assignment of entities into a treated group vs. a control group is outside the control of the investigator
  - **Sample survey:** Information about a population at a specific time
  - **Prospective study:** Longitudinal study
  - **Retrospective study:** Sample survey + recollection

  - **Comparative study:** Two or more methods are compared
    - Compare three educational systems based on cost effectiveness
  - **Descriptive study:** Characterize a population based on attributes
    - Characterize the computer skills of teenagers attending public schools

- No cause-and-effect relationships can be found; only associations
  - Confounding factors due to lack of control

## Sample Survey Example

- Product preference poll related to Dell vs. Apple
  - What was the population of interest?
  - Was the sample in fact selected from this population?
  - What questions were asked and how were the questions phrased?

# Study Variables

- Explanatory variables

- Response variables

- Confounding variables or covariates
  - In an observational study we cannot control covariates

# The Scientific Method

1. Asking a question about a natural phenomenon
2. Making observations of the phenomenon
3. Hypothesizing an explanation for the phenomenon
4. Predicting a logical consequence of the hypothesis
5. Testing the hypothesis by an experiment or an observational study
   - 5.1 Identifying the variable(s) of interest
   - 5.2 Choosing an appropriate design for the study
   - 5.3 Collecting the data
   - 5.4 Summarizing the data
   - 5.5 Analyzing the data
6. Creating a conclusion with data gathered in the experiment/study, or forming a revised/new hypothesis and repeating the process

## An Example Project Formulation

1. Assess the public's perception of the city's bus system

2. Identify variables that can increase the number of people using buses
   2.1 Cleanliness, proximity & frequency of service, on-time arrival, cost
   2.2 Choose Likert scale

3. Sample survey or controlled experiment?
   3.1 If survey, choose covariates: age, financial status, and others
   3.2 If experiment, decide on variable to vary (regular vs. reduced price)

# Science is ... Measurements

- Measuring a few numbers underlies many of the most expensive projects in science
  - Particle Physics is devoted to finding masses and changes of a small number of elementary particles
  - The GENOME project aimed to find a long sequence composed of four letters, which one can think as a single very large number in base 4
  - Government agencies like to fund projects of this type for the simple reason that the success of the project is almost guaranteed and there is specific deliverable!

- Measuring even a single number is more difficult than it seems
  - Random errors
  - Systematic errors