

# Week 1

Monday, 2 August 2021 23:36

## BIOSTATISTICS

IMPORTANT FOR PUBLIC HEALTH EDUCATION,  
RESEARCH & PRACTICE

STUDYING SUB-SET OF A LARGER  
POPULATION OR PROCESS

STUDY DESIGN PRACTICES  
TYPES OF DATA WE WILL ENCOUNTER

STATISTICS  
DATA SCIENTISTS  
BIG DATA ENGINEER

} SEXIEST JOBS  
OF THE 21<sup>ST</sup> C.

DATA IS EVERYWHERE  
DATA IS THE NEW OIL

PLANNING  
DESIGN OF STUDY  
DATA COLLECTION  
DATA ANALYSIS  
PRESENTATION  
INTERPRETATION

DATA  
PIPELINE  
[ RESEARCH  
PROCESS ]

## • SAMPLES vs POPULATION



IMPERFECT REPRESENTATION  
OF SOME LARGER POPULATIONS

SUBSET OF POPULATION FROM  
WHICH INFO IS COLLECTED

"IMPERFECT REPRESENTATION"

WE CANNOT OBSERVE POPULATION DIRECTLY

ERROR ASSOCIATED WITH CHOOSING THE SAMPLE

SAMPLE SHOULD BE REPRESENTATIVE

### SIMPLE RANDOM SAMPLING

EVERY POSSIBLE SUBSET OF SIZE  $N$   
FROM POPULATION IS EQUALLY LIKELY  
TO BE SELECTED

IF SAMPLE NOT REPRESENTATIVE → BIAS

SOMETIME THIS IS NECESSARY

EX DRUG USER IN MUMBAI INDIA (THERE IS NO LIST OR REGISTRY)

EX WOMEN ARE LESS LIKELY TO BE OPEN ABOUT THEIR DRUG USE

RANDOM DIGIT DIALING  
CONVENIENCE SAMPLING  
RESPONDENT DRIVEN SAMPLING

USED WHEN  
SAMPLE IS  
DIFFICULT TO  
OBTAIN

NOT ALL ELEMENTS OF A STUDY CAN BE SAMPLED

## STATISTICS

HOW TO DEAL WITH THE IMPERFECT ?  
REPRESENTATION OF OUR POPULATION

HOW CAN WE ACCOUNT FOR THE UNCERTAINTY  
IN OUR ESTIMATE ?

SAMPLING PROCEDURE ?

## • STUDY DESIGN

### PROSPECTIVE COHORT STUDIES

(DEFINING OUTCOME)



RANDOMIZED  
CONTROLLED  
STUDY

RANDOMLY ASSIGN  
SAMPLE MEMBERS  
TO EXPOSURE  
GROUPS

"SMOKE CAUSES X"



ONLY SYSTEMATIC  
DIFFERENCE IS  
EXPOSED OF  
INTEREST

INTERNAL



OBSERVATIONAL  
COHORT  
STUDY

GROUP MEMBERSHIP,  
MEASURING EFFECT  
ON DIFFERENT  
GROUPS

"SMOKERS VS NON-SMOKERS  
PERFORM DIFFERENTLY" X

### SELF-SELECTION

METHOD MAY LEADS TO  
SYSTEMATIC DIFFERENCES

"MAYBE SMOKERS ARE MORE LIKELY TO  
DRINK ALCOHOL AND BOTH ARE LINKED  
TO CARDIOVASCULAR DISEASES"

PROTECTION  
AGAINST SELF  
SELECTION BIASES

TREATMENT DECIDED  
BY RESEARCHERS

COMPLEX RANDOMIZING SCHEMA

(NOT ALWAYS POSSIBLE)

POTENTIAL  
CONFOUNDERS

FACTORS RELATED TO BOTH  
THE OUTCOME & EXPOSURE  
OF INTEREST

CASE/CONTROL STUDIES (AFTER OUTCOME)

IN PRECEDENT STUDIES SUBJECT WERE ASSIGNED  
AN EXPOSURE STATUS OR WERE SELECTED &  
THEIR EXPOSED STATUS WAS CLASSIFIED

EXPOSURE W/ KNOWN OUTCOMES

SELECTED PEOPLE ON WHETHER THEY HAVE AN  
OUTCOME OR NOT (MOSTLY A RARE DISEASE)

ANALYTICAL ISSUE

CAN'T ESTIMATE RISK OF OUTCOME

INFLUENCE BY HOW RESEARCHERS  
CHOOSE CONTROLS & CASES

EX LINK BETWEEN SMOKE & LUNG-CANCER

POTENTIAL CONFOUNDERS

THERE CAN BE SYSTEMATIC DIFFERENCES  
BESIDE EXPOSURE DIFFERENCES

WHAT CONFOUNDERS HAVE NOT BEEN ADDRESSED?

RECALL BIAS

PATIENT INTERVIEW (THEY TEND TO EXAGGERATE)

• DATA TYPES & SUMMARIZATION

CONTINUOUS DATA (INCREMENTAL) mmHg,  $[NO_2]$ , €/year

## BINARY DATA

YES-NO

## CATEGORICAL DATA

(NOMINAL & ORDINAL)

RACE, GENDER, RELIGION (NOMINAL)

↓  
NO ORDER

↓  
ORDER (STANDARD RANKING)

BINARY IS A SPECIAL CASE OF CATEGORICAL DATA (2 LEVELS)

## TIME-TO-EVENT DATA

# OF TIMES, PLACE, TIME

## COMPARING BLOOD PRESSURE?

- ESTIMATE  $\mu$  MEAN DIFFERENCE IN BP  
OF EACH OF THE GROUP
- ESTIMATE A 95% CONFIDENCE  
INTERVAL & USE A T-TEST  
FOR POPULATION LEVEL DIFFERENCES
- CONFIDENCE INTERVALS
- PVALUE
- T, Z,  $\chi^2$ , F TESTS
- KAPLAN MEIER
- LOG-RANK TEST
- PROPORTIONS, PROP TEST

DIFFERENT APPROACHES TO  
ANALYZE DIFFERENT DATA-TYPES

DATATYPES &  
STUDY-DESIGNS