Math507M: Statistics for Medical and Public Health Researchers

Instructor: Nan LIN
nlin@wustl.edu

Department of Mathematics, College of Arts and Sciences
Division of Biostatistics, School of Medicine
Washington University in St. Louis

Acknowledgement: These slides are made based on course materials previously used by Jeff Gill and Han Liang Gan.
Target audiences

Target audiences of this class are

- medical and public health researchers
- who do not have much background in statistics
- but need statistics for their research work.
Motivation of the course

- **Prevalence:** Statistics is widely used in every data-oriented field, including public health and medicine.
- **Evidence:** In a recent study of three top UK medical journals over a 1-year period, 66% of the articles contained some statistical analysis of data.
- **Understanding the Literature:** The material in this course is important for reading reports, articles, and book-length treatments of health/medical analysis.
- **Contributing to the Literature:** You cannot publish in these areas without a basic understanding (or better) of statistical analysis. Even with sophisticated coauthors.
- **Purpose:** Statistics is used in these areas to make assertions about biological and social relationships and the strength of evidence of those relationships.
- **A Pervasive Problem:** Statistics are often used misleadingly (on purpose or accidentally), and we want enough sophistication to make judgments about such work.
Research question: is fluorine in tap water related to cancer mortality?

Burke & Yiamouyannis (1975):
- 10 fluoridated and 10 non-fluoridated US towns between 1950 and 1970
- Cancer deaths increased 20% in fluoridated and 10% in non-fluoridated.

Reanalysis by Oldham & Newell (1977):
- Controlled for changes in age and ethnicity in these towns
- Cancer deaths increased 1% in fluoridated and 4% in non-fluoridated.

Real Story: in fluoridated towns there was more out-migration of young and white residents (confounding variables).
Example: Fluoridated Water Supplies (page 4)

- Research question: is fluorine in tap water related to cancer mortality?
- Burke & Yiamouyannis (1975):
  - 10 fluoridated and 10 non-fluoridated US towns
  - between 1950 and 1970
  - cancer deaths increased 20% in fluoridated and 10% in non-fluoridated.
Research question: is fluorine in tap water related to cancer mortality?

Burke & Yiamouyannis (1975):
- 10 fluoridated and 10 non-fluoridated US towns
- between 1950 and 1970
- cancer deaths increased 20% in fluoridated and 10% in non-fluoridated.

Reanalysis by Oldham & Newell (1977):
- controlled for changes in age and ethnicity in these towns
- cancer deaths increased 1% in fluoridated and 4% in non-fluoridated.
Example: Fluoridated Water Supplies (page 4)

- Research question: is fluorine in tap water related to cancer mortality?
- Burke & Yiamouyannis (1975):
  - 10 fluoridated and 10 non-fluoridated US towns
  - between 1950 and 1970
  - cancer deaths increased 20% in fluoridated and 10% in non-fluoridated.
- Reanalysis by Oldham & Newell (1977):
  - controlled for changes in age and ethnicity in these towns
  - cancer deaths increased 1% in fluoridated and 4% in non-fluoridated.
- Real Story: in fluoridated towns there was more out-migration of young and white residents (confounding variables).
Power Difference: Aspirin For Non-Fatal Myocardial Infarction (page 110)

- The treatment in question: 300 mg aspirin three times daily.
- Elwood and Sweetnam (1979, *Lancet*) studied 1239 patients just after a myocardial infarction, and found the mortality difference at one year to be:

  \[ p_{aspirin} = 0.080, p_{placebo} = 0.107, CI_{0.95} = [-0.005 : 0.06]. \]

- Later a study by the *Persantine-Aspirin Reinfarction Research Study Group* (1980, *Circulation*) of 6292 patients under almost exactly the same circumstances found:

  \[ p_{aspirin} = 0.092, p_{placebo} = 0.115, CI_{0.95} = [0.008 : 0.038]. \]

- So the first confidence interval contains zero, and the second does not.
- What is going on here?
Course Objectives

▶ design sound experiments providing valid and precise conclusions
▶ conduct simple exploratory data analyses
▶ construct appropriate statistical models
▶ effectively use R in a statistical analysis
▶ appreciate nature of variability and how statisticians deal with it
▶ understand the role and purpose of modelling in statistical data analysis
A typical process

- Form a research question
- Design a study to collect data
- Analyze the data
- Make inferences
- Make conclusions
Types of Data

- **Levels of Measurement:**
  - nominal (categorical)
  - ordinal (categorical)
  - count (quantitative)
  - continuous (quantitative)

- We need to understand these categories in detail since they determine which appropriate statistical procedure should be used.
Examples

- **NOMINAL**: dichotomous, polychotomous, categorical
  - old/young
  - gender
  - blood type
  - eye color
  - diabetic/non-diabetic
  - marital status
  - religions of the world

- **ORDINAL**: ranks, ordering
  - lower/middle/upper class
  - disease stage
  - social class
  - education level
Examples

- **COUNT**: number of pregnancy  
  annual number of death in a hospital  
  number of decayed teeth  
  number of cigar smoked per day

- **CONTINUOUS**: takes any value within a given range  
  body mass index  
  blood pressure  
  income  
  survival time
Scale of measurements

When comparing two values, does the difference matter or the ratio?

- **INTERVAL**: the difference matters
- **RATIO**: the ratio matters

In a ratio scale, the value of zero has real meaning, whereas in an interval scale, the position of zero is arbitrary.
Data types with regard to time:
- cross-sectional: collected at the same time
- time-series: collected over time to evaluate trends
- longitudinal versus time-series

Typical data sources: patient records, hospital/clinic records, government agencies, receipts, patient demographics, academic reports, clinical trials, physical devices...

Population data analyzed with descriptive statistics.
Sample data analyzed with inferential statistics.

Quality assessment of data collection is critical.
Soya food intake and risk of endometrial cancer among women in Shanghai

Xu et al (2004, BMJ), Table 1
Different data collection methods

There are typically two types of studies.

- **Experimental studies**: This is where the investigator assigns the exposure/treatment to some of the individuals with the aim to compare those exposed/treated to those who are not.

- **Observational studies**: The investigator will often select individuals, some of which will have had the exposure/treatment, and others not, and observe an outcome
Basic Terminology

- **Data** can be obtained from either *observational* or *experimental* sources.
- **Random Variable** a stochastic future observation from data, not constant, not necessarily uniform. Data are observed random variables.
- **Explanatory Variable** a variable used as an explanation for the behavior of another ("independent variable", "predictor", "input variable").
- **Outcome Variable** a variable of primary interest whose behavior is analyzed relative to levels of explanatory variable ("dependent variable", "response variable", "output variable").
- **Control Variable** a variable held constant at more than one level to observe the relative effects of other variables.
Population and Sample

We typically have a research question about a general population. To find the correct answer, we could sample every single person in the population but for practical reasons this is unlikely to be possible. As a result, typically we take a sample of a population, and using the sample, try to draw conclusions about the general population. This is only valid if our sample is representative of the general population.

In general, our conclusions can only be applied to the population that the sample represents.
More Terminology

- **Inference** using sample data and a model to make claims (estimates) about population parameters.
- **Prediction** using sample data and a model to make claims about future observations.
- **Statistic** a descriptive measure based on a population or sample data that does not depend on a parameter.
- **Model** a necessarily unrealistic picture of nature, a formal representation and simplification using symbology and assumptions.
Introduction to R

- Download R: [https://cran.r-project.org/](https://cran.r-project.org/)
- Download RStudio: [https://www.rstudio.com/](https://www.rstudio.com/)
- Install R package Rcmdr
  ```r
  install.packages("Rcmdr", dependencies = TRUE)
  ```
- Load the R package Rcmdr
  ```r
  library(Rcmdr)
  ```