STAT 500 - Fall 2019

Unit 1 - Introduction

What is Statistics?

- Dictionary definitions:
  - Branch of mathematics dealing with the collection, analysis, interpretation and presentation of data
  - Art and science of drawing justifiable conclusions from data

Statistics as a Mathematical Science

- Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $\epsilon_i \sim$ iid $N(0, \sigma^2)$, and $i = 1, 2, \cdots, n$.

- Model parameters are β_0, β_1 and σ^2

- We will find estimators for these parameters and derive their properties.

Statistics as a Mathematical Science

- The simple linear regression model in matrix form:
  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad$ where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$

- The matrix formulation has

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \text{and} \quad E(\mathbf{Y}) = X\boldsymbol{\beta} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

- The unknown parameters are $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ and σ^2

Statistics as a Mathematical Science

We have the following results:

- The least squares estimator

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix}$$

  is the minimum variance linear unbiased estimator for $\boldsymbol{\beta}$

- $\text{Var}(\widehat{\boldsymbol{\beta}}) = \mathbf{V} = \sigma^2\,(\mathbf{X}^T\mathbf{X})^{-1}$

- $\mathbf{c}^T\widehat{\boldsymbol{\beta}} \sim N(\mathbf{c}^T\boldsymbol{\beta},\ \mathbf{c}^T\mathbf{V}\mathbf{c})$

- Test H_0: $\mathbf{c}^T\boldsymbol{\beta} = 0$ using $t = \dfrac{\mathbf{c}^T\widehat{\boldsymbol{\beta}} - 0}{\sqrt{\mathbf{c}^T V\,\mathbf{c}}}$

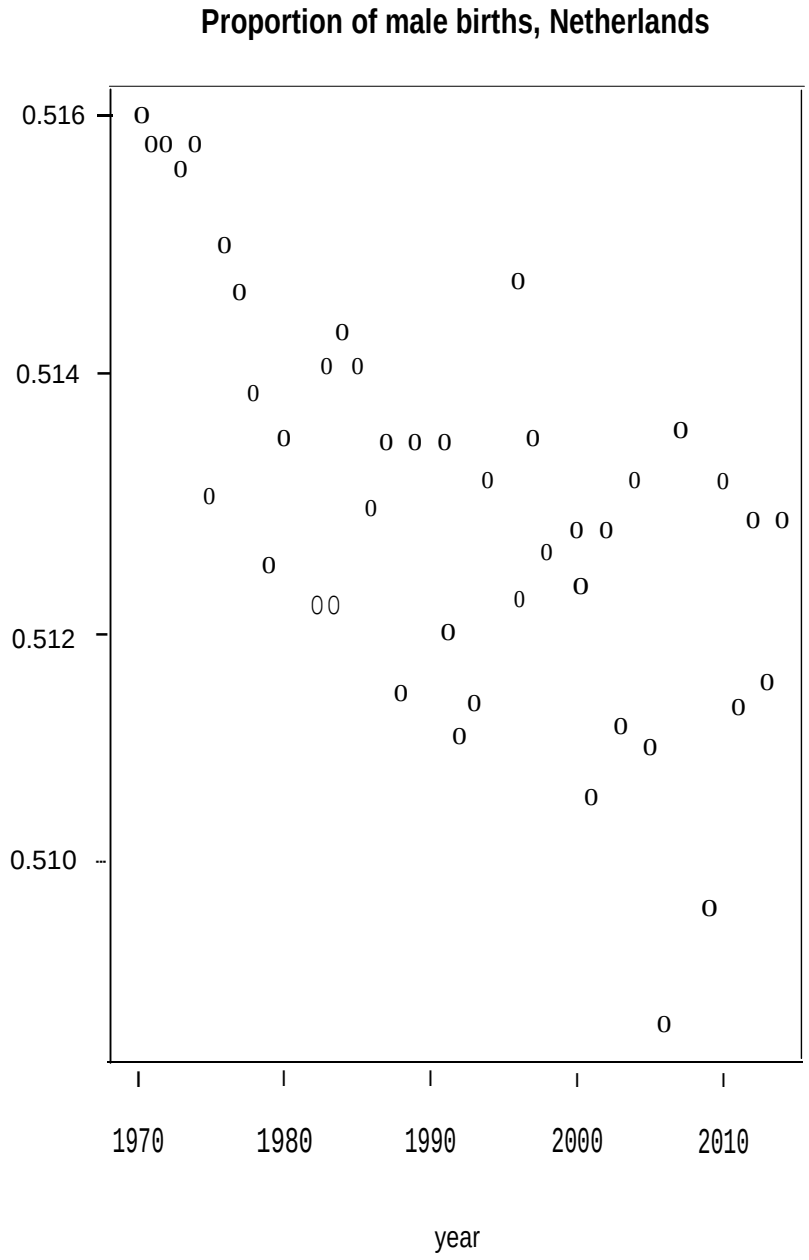
Statistics as Art

Suppose a researcher says to you:

> *I have data on percentages of male births in the Netherlands from 1970 to 2013. They seem to be declining. I need to forecast the expected change in the next ten years (2014 - 2023). You're a statistician; can you help me?*

Some issues:

- What type of model can be used to address the question?
- How can you determine if a proposed model is reasonable?


Proportion of male births, Netherlands
0.516
0.514
0.512
0.510
1970
1980
1990
2000
2010
year

Statistics as Art

- What types of inference can be made?
  - Did a significant decline in male births occur?
  - Will trends continue into the future?
  - What types of predictions can be made?
  - How can accuracy or reliability of predictions be assessed?
- Usually more than one reasonable approach.
- Models are almost never "true", but some may be useful?

Statistics as Science

Statistics is the science of using information to make decisions and quantify uncertainty inherent to those decisions.

There are four basic steps in the statistical problem solving process (Deming):

- Define the questions to be answered (Plan)
- Gather appropriate data (Do)
- Analyze the data (Study)
- Interpret the results (Act)

Define Questions

- Researchers define study questions.
- As statisticians, sometimes we play a role in question development, sometimes not.
- Questions drive rest of statistical investigation.

Data Collection

- Experiments: Researchers impose an intervention on members of some population
  - Planned intervention (prospective)
    - Researcher changes the level of at least one factor in order to observe a response
  - Causal inferences are possible
    - Hold the levels of all other factors constant
    - Random assignment of experimental units to treatment groups (randomized experiment)
      - Randomization eliminates uncontrolled sources of bias
      - Randomization provides a basis for inference

Data Collection

- Observational studies: Members of some population are observed as they naturally exist
  - Census: Observe all members of some population
  - Haphazard sample
  - Representative random sample

- May be able to make inferences about associations, but causal inferences are usually not be possible.

Data Collection

Experiments and observational studies are performed to obtain useful information

- Industrial experiments
  - improve process yields and quality
  - reduce production costs and increase profits
  - reduce sensitivity to uncontrolled factors (variation)

- Agriculture and food sciences
  - improve crop yields
  - improve disease resistance and manage pests
  - develop new food products

- Human health studies
  - establish effectiveness of treatments
  - identify causes of disease
  - assess the role of nutrition in long term health status

- Business and economics
  - model consumer behavior
  - inventory control
  - improve management processes (business analytics)

- Ecology
  - Monitor population growth
  - Examine competition among species
  - Monitor water or air quality
  - Examine climate change

Data Analysis

- Methods depend on
  - Type of data collected
  - How data were collected
  - Research Questions
- Typically what outsiders consider **Statistics**

Interpret Results

- Answer research questions based on results from data analysis

- Conclusions must match scope of data collection methods
  - Ex. Experiment - generalize results to a population
  - Ex. Observational study - no causal inference

Statistics is ...

- Part mathematics
- Part art
- Part science

STAT 500 will include all three aspects.