Fundamentals of Statistical Models

STA 610 - Applied Statistics for Health Professions

Andrew DiLernia

## Learning Objectives

* Variable types
* Variable roles



*Image from* [*http://www.imstat.org/wp-content/uploads/bulletin/42\_05-George-E-P-Box.jpg*](http://www.imstat.org/wp-content/uploads/bulletin/42_05-George-E-P-Box.jpg)

“Essentially, all models are wrong, but some are useful.” -George Box



*Image from* [*https://techjobstour.com/wp-content/uploads/2017/09/kathypham.jpg*](https://techjobstour.com/wp-content/uploads/2017/09/kathypham.jpg)

“Today’s computer scientists write code with the potential to affect billions of people’s privacy, security, equality, and well-being.” -Kathy Pham

## Statistical Models

Statistical models characterize real-world processes using mathematical equations and assumed frameworks, and are integral in areas such as economics, ecology, social sciences, the health sciences, and numerous other domains. Across these areas, statistical models are useful tools for describing patterns, relationships, and variability within observed data sets and to improve the understanding of stochastic processes.

**Examples**:

1. Modeling the selling value of a house based on characteristics e.g. size, number of bedrooms, and age 🏠
2. Modeling the national unemployment rate based on historical data 📈
3. Predicting if an e-mail is spam or not based on the number of consecutive capital letters 💻
4. Modeling disease outcome (progression or remission) based on patient age, gender, and treatment status (active treatment or placebo) 💊
5. Creating an algorithm for recommending new videos for Netflix subscribers to maximize watch time based on a subscriber’s watch history and the behavior of other subscribers 📺
6. Model or forecast the spread of an infectious disease in a population of interest based on vaccination rates in a population, the time of year, and other factors 🦠

## Key Terms

* **Observational unit**: people, objects, or cases on which data are recorded.
* **Variables**: characteristics that are measured or recorded about each observational unit.

➡️ What are the observational units for each of the scenarios above?

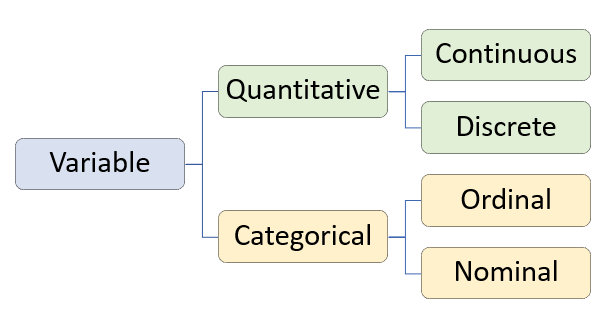
1. *Each individual house/apartment*
2. *Country*
3. *Each individual email*
4. *Each individual patient*
5. *Each netflix subscriber*
6. *Record of the infectious disease level on each day.*

➡️ What are the variables for each of the scenarios above?

1. *Selling price of house (dollars), number of bedrooms, age of the house.*
2. *The number of people unemployed/ unemployment rate*
3. *Spam status, number of CAPITAL letters in an email etc.*
4. *Patient age, gender, and treatment status (active treatment or placebo)*
5. *Subscriber’s watch history and the behavior of other subscribers*
6. *Vaccination rates in a population, the time of year, and other factors*

## Variable Types

When using statistical models, it is important to correctly identify the types of variables we are working with since different models are appropriate for certain types of variables.



Although there are more specific subtypes as well, generally variables fall into two categories: quantitative and categorical variables.

#### Quantitative variables

**Quantitative variable**: measures numbers describing the observational units such that arithmetic is sensible (also called numeric variable).

* Continuous quantitative variables can always be measured more precisely
* Discrete quantitative variables must be recorded in whole numbers

**Examples:**

* Height in inches (continuous) 📏
* Weight in pounds (continuous) ⚖️
* Age in years (continuous) 👵
* Number of pets a person owns (discrete) 🐈 🐶
* Concentration of a protein in a sample measured in milligrams per millimeter (continuous) 🧪
* Number of mutations in a genomic region (discrete) 🧬

#### Categorical variables

**Categorical variable**: records a category designation of the observational units.

* Ordinal categorical variables have a particular ordering to the different levels or categories of the variable
* Nominal categorical variables have no particular ordering of its categories / groups
* When a categorical variable has only two possible categories, it is said to be **binary**.

**Examples:**

* Netflix subscriber being yes or no (nominal, binary) 📺
* Class standing being a freshman, sophomore, junior, or senior (ordinal) 📚
* Genotype being either homozygous dominant; AA, heterozygous; Aa, or homozygous recessive; aa (ordinal) 🧬

➡️ What are the types and subtypes of each of the variables for the three scenarios above?

## Concept Check

➡️ What are the types and subtypes of the following variables?

* Rating of agreement with 5 different levels from “strongly disagree” to “strongly agree” 👎👍 *→ Ordinal Categorical variable*
* Roller coaster intensity from 1 to 5 🎢 *→ Ordinal Categorical variable*
* Probability of each NFL football team winning the Super Bowl this year 🏈 *→ Continuous quantitative variable*

## Variable Roles

* **Response variable**: the variable describing the outcome of interest.
* **Explanatory variable**: the variable whose relationship with the response variable is being studied. Explanatory variables in regression analyses are also referred to as predictors or predictor variables.

**Response:** In a study exploring whether a student is accepted to a university or not *(Categorical binary)*

**Explanatory:** GPA *(quantitative continuous)* and ACT or SAT scores *(quantitative discrete)* are recorded, what are the response and explanatory variables? 🎓

➡️ What are the response and explanatory variables in the following scenarios?

1. Estimating selling value of a house based on characteristics e.g. size, number of bedrooms, and age 🏠

***Response:*** *Selling value of house*

***Explanatory:*** *size, number of bedrooms, age of house, location etc.*

1. Using historical unemployment rates to predict next month’s national unemployment rate 📈

***Response:*** *Monthly (next month) unemployment rate*

***Explanatory:*** *Previous month’s unemployment rate*

1. Determining if an e-mail is spam or not knowing the number of consecutive capital letters 💻

***Response:*** *Spam status (binary categorical)*

***Explanatory:*** *no. of capital letters in the email for example (discrete quantitative)*

1. Predicting disease outcome (progression or remission) based on patient age, gender, and treatment status (active treatment or placebo) 💊

***Response:*** *disease outcome (active or remission)*

***Explanatory:***  *age, gender and treatment status(treatment active or placebo)*

## Study Designs

Another important consideration when using statistical models is the consider how the data set we are analyzing was collected. The inferences one can make based on the analysis of a data set greatly depend on the type of study design employed.

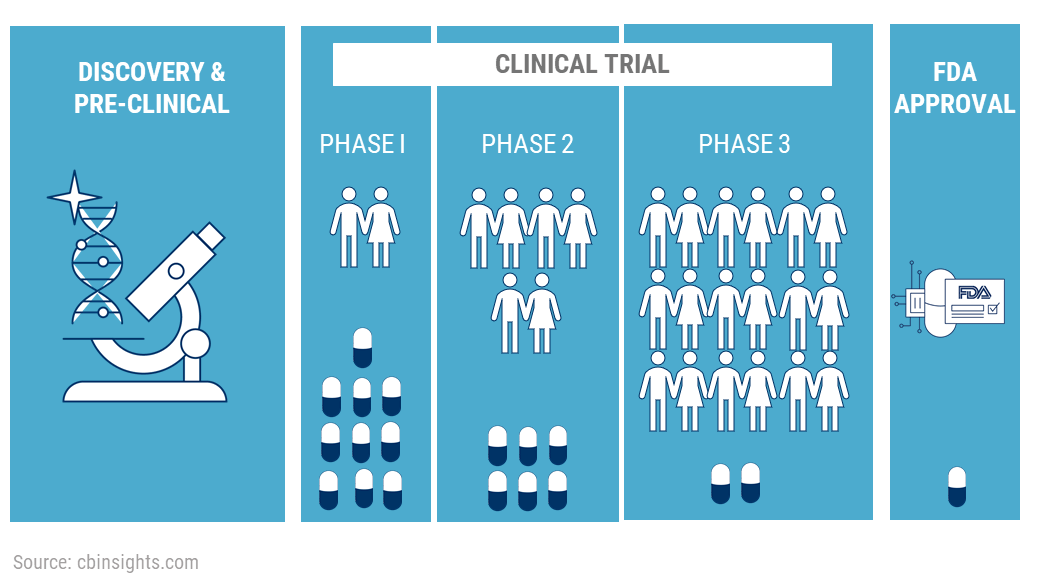
* **Controlled experiment**: a study in which explanatory variables are manipulated or assigned.
* **Observational study**: A study in which explanatory variables are not assigned, but passively observed.

➡️ For each of the following scenarios, indicate whether the setting corresponds to a controlled experiment or observational study.

1. A study in which we record students’ heights and whether or not they like basketball. 🏀 *→ observational study*



1. A clinical trial in which some patients are given an active drug for lowering blood-pressure, and other patients are given a placebo 💊. The change in patient blood-pressures is recorded. *→ controlled experiment*



1. A study in which we record how many hours per week students study and their final exam scores 📚. *→ observational study*
2. A case-control study in which we recruit 50 people diagnosed with Alzheimer’s disease and 50 matched healthy controls. We measure the level of amyloid beta () proteins in each participants brain 🔬. *→ observational study*