Types of data
Random variables
Probability

Fundamental Rule of Data Analysis

Different types of data require different statistical analyses.

Types of Data

- Continuous
- Categorical
- Count
- Time-to-event
- Qualitative

Types of Data: Continuous

- Continuous → Any number in a range or interval →
  Examples of continuous data:
  - Height
  - Weight
  - Age
  - IQ
  - Systolic blood pressure
- Categorical
- Count
- Time-to-event
- Qualitative

Types of Data: Categorical

- Continuous
- Categorical
- Count
- Time-to-event
- Qualitative

Ordinal: Ordered categories
Nominal: Unordered categories
    Also called "binary" or "dichotomous" when there are only two categories

Examples of ordinal data:
    Highest level of education
        (high school / associate's degree / bachelor's degree / graduate school)
    Likert scale
        (agree / neutral / disagree)

Examples of nominal data:
    Race
    Gender
    Alive/dead
    Yes/no
    Religion

Transforming Continuous Data

- Continuous measurements are occasionally transformed into categorical measurements.
- Example: In a study of the effects of maternal smoking on infant birth weight, weight could be categorized as <1500 grams, 1500 to 2499 grams, or ≥2500 grams.
- The analysis may be simplified when continuous measurements are categorized, but some information is lost.
- It is best to record data with as much precision as possible; measurements can always be collapsed or transformed later.

Types of Data: Count

- Continuous
- Categorical
- Count
- Time-to-event
- Qualitative

Count → Numerical data that can only be non-negative integers →

Examples of count data:

- Number of children each mother has
- Number of lung cancer deaths in Lexington each year
- Number of doctor visits each person made in 2018

Types of Data: Time-to-Event

- Continuous
- Categorical
- Count
- Time-to-event → The time it takes for something to happen →
- Qualitative

Examples of time-to-event data:

Time to death
Time to cancer remission
Time to breast cancer diagnosis
Time to hospital readmission

How is this different from continuous data?

- Some subjects may not experience the event at all
- Some subjects may leave the study before the event occurs
- These are types of **censoring**

Types of Data: Qualitative

- Continuous
- Categorical
- Count
- Time-to-event
- Qualitative

Anything that cannot be represented as a number or category; based on observations, interviews, or evaluations

Examples of qualitative data:

- Answers to an open-ended question
- Notes from focus groups
- Descriptive statements about subjects' well-being

Types of Data

- Continuous
- Categorical (focus on binary)
- Count
- Time-to-event
- Qualitative

DISCLAIMER: This course will *not* prepare you to analyze every type of data you encounter.

Variables

- A **variable** is a descriptor or condition that can take multiple **values** (also called observations or measurements)
- A variable can be thought of as a question about a subject/participant, to which the value is their answer.
  - Example: How old is this participant? Answer: 38 years old

    The variable is:

    The value is:

Variables

ID	Age	Sex	Height
001	38	F	65
002	45	F	63
003	44	M	67
004	51	M	70
005	51	M	75
006	38	F	60
007	39	M	72
008	38	F	69
009	50	F	66
010	37	M	66

- It may be helpful to think of an Excel spreadsheet
- Columns are **variables**
- Rows are **participants**
- Inner cells are **values/ observations/measurements**

Variables

- A variable is oftentimes represented by a capital letter
- Example: Let X represent age. For subject 001, X=38. For subject 005, X=51.

ID	Age	Sex	Height
001	38	F	65
002	45	F	63
003	44	M	67
004	51	M	70
005	51	M	75
006	38	F	60
007	39	M	72
008	38	F	69
009	50	F	66
010	37	M	66

Events

- An **event** is a potential outcome
  - The assignment of a value or a set of values to a variable
  - The result of an experiment or observation
  - A declarative statement (a statement that can be true or false)
- Examples:
  - $X=1$
  - $Y \geq 10$
  - a coin flip lands on heads
  - the subject is older than 40

Probability

- The way we express uncertainty
- The probability of an event is its relative frequency of occurrence
- Proportion of times the event occurs in a specific number of trials
- Denoted with $P(\ )$
- Always lies between 0 and 1 (inclusive)

Probability Example

- Example: We are interested in the probability that a child born in the United States has low birth weight, defined as weight <2500 grams (or equivalently, 5 lb 8 oz). In 2004, among 4,112,052 registered live births, 331,772 were low birth weight infants.

Conditional Probability

- **Conditional probability** is the probability that an event occurs, given that another event occurred
- Restricts your "population" to the people for which the second event occurred
- Examples
  - Probability of getting the flu this year given you got the flu shot
  - Probability of being diagnosed with breast cancer given you have the BRCA1 gene mutation
  - Probability of falling asleep while watching this lecture given you got 4 hours of sleep last night

Conditional Probability Example

- Example: In 2004, among 4,112,052 registered live births, 331,772 were low birth weight infants. Among all live births, 3,972,558 were singletons and the other 139,494 were from a multiple birth (twins, triplets, etc.). Among the 139,494 multiple births, 91,648 were low birth weight infants. What is the probability of a newborn being low birth weight given they are from a multiple birth?

Diagnostic Testing

- Consider a situation where people either truly have some disease (D+) or don't have the disease (D-), but we don't know which. We've developed a test where each subject will either test positive (T+) or test negative (T-)
- Ideally, we want the D+ people to test positive (T+) and the D- people to test negative (T-) as often as possible
- Diagnostic tests are rarely perfect. They'll inevitably classify some people wrong (false positives or false negatives)

Diagnostic Testing

- **Sensitivity**: $P(\mathrm{T}+ \mid \mathrm{D}+)$
  - Probability of testing positive given you have the disease
  - Among those with the disease, it's the probability of testing positive
- **Specificity**: $P(\mathrm{T}- \mid \mathrm{D}-)$
  - Probability of testing negative given you do not have the disease
  - Among those without the disease, it's the probability of testing negative

- Want sensitivity and specificity to be as high as possible

Diagnostic Testing Example

- Example: Researchers have developed a new screening test for asthma. They take a random sample of 1000 American adults. In this sample, 80 people truly have asthma and 920 truly do not have asthma. Among the 80 who truly have asthma, 72 test positive for asthma and 8 test negative for asthma. Among the 920 who truly do not have asthma, 184 test positive for asthma and 736 test negative for asthma. What is the sensitivity and specificity of the new screening test?

Diagnostic Testing

- **Positive predictive value**: $P(\mathrm{D}+ \mid \mathrm{T}+)$
  - Probability of having the disease given you tested positive
  - Among those who test positive, it's the probability of having the disease
- **Negative predictive value**: $P(\mathrm{D}— \mid \mathrm{T}—)$
  - Probability of not having the disease given you tested negative
  - Among those who test negative, it's the probability of not having the disease

- Want positive predictive value and negative predictive value to be as high as possible

Diagnostic Testing Example

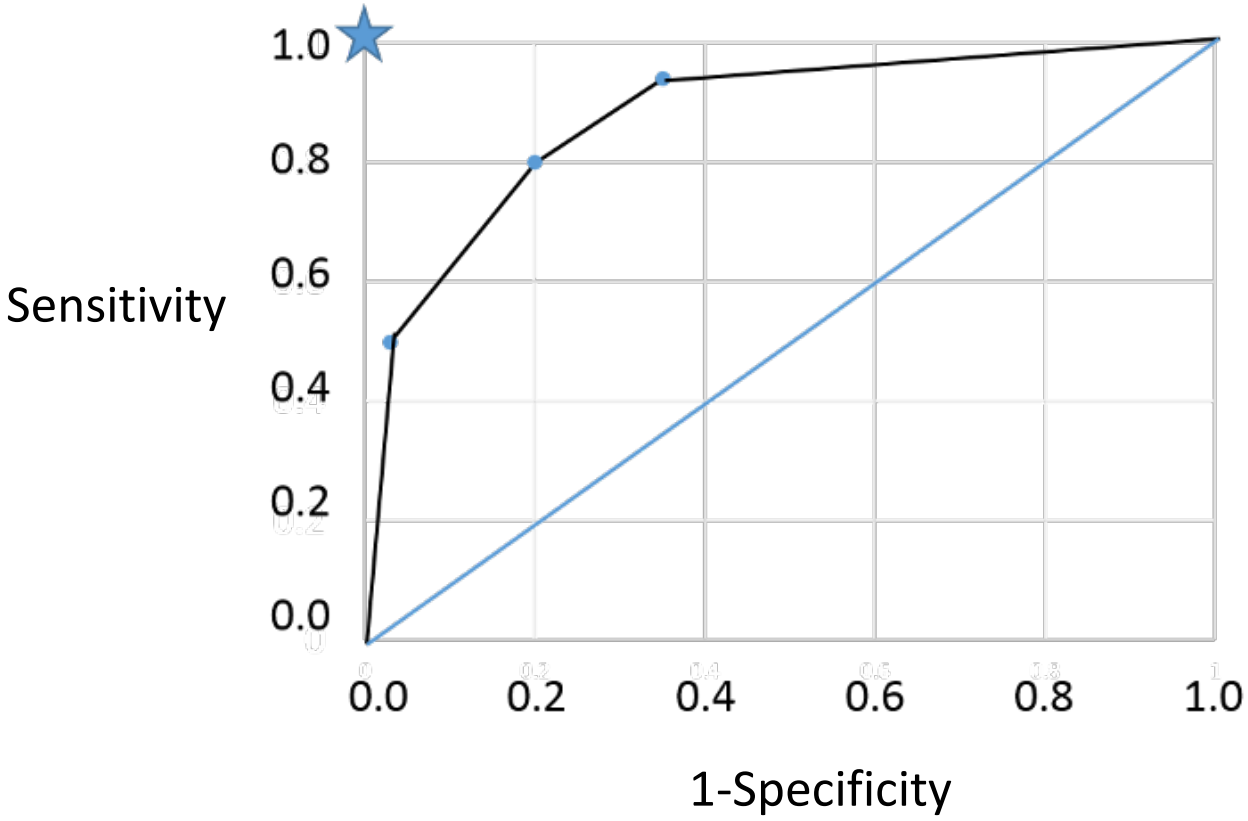
- Example: Researchers have developed a new screening test for asthma. They take a random sample of 1000 American adults. In this sample, 80 people truly have asthma and 920 truly do not have asthma. Among the 80 who truly have asthma, 72 test positive for asthma and 8 test negative for asthma. Among the 920 who truly do not have asthma, 184 test positive for asthma and 736 test negative for asthma. What is the positive and negative predictive value of the new screening test?

Balance between Sensitivity and Specificity

- What would a test with 100% sensitivity look like?
- What would a test with 100% specificity look like?
- When the screening test is a continuous measurement, we have a choice of where to put the “cutoff” for positive/negative results
- High sensitivity may be more important than high specificity (or vice versa) depending on the situation. How serious is a false positive vs. false negative?
  - Example of high sensitivity/low specificity test: mammograms
  - Example of low sensitivity/high specificity test: screening test for illicit drug use

ROC Curves

- Calculate sensitivity and specificity for different cutoffs
- Plot sensitivity on y-axis, 1-specificity on x-axis
- Helpful for identifying what cutoff to use for indicating positive/negative test results
- Area under the curve sometimes used to quantify test performance



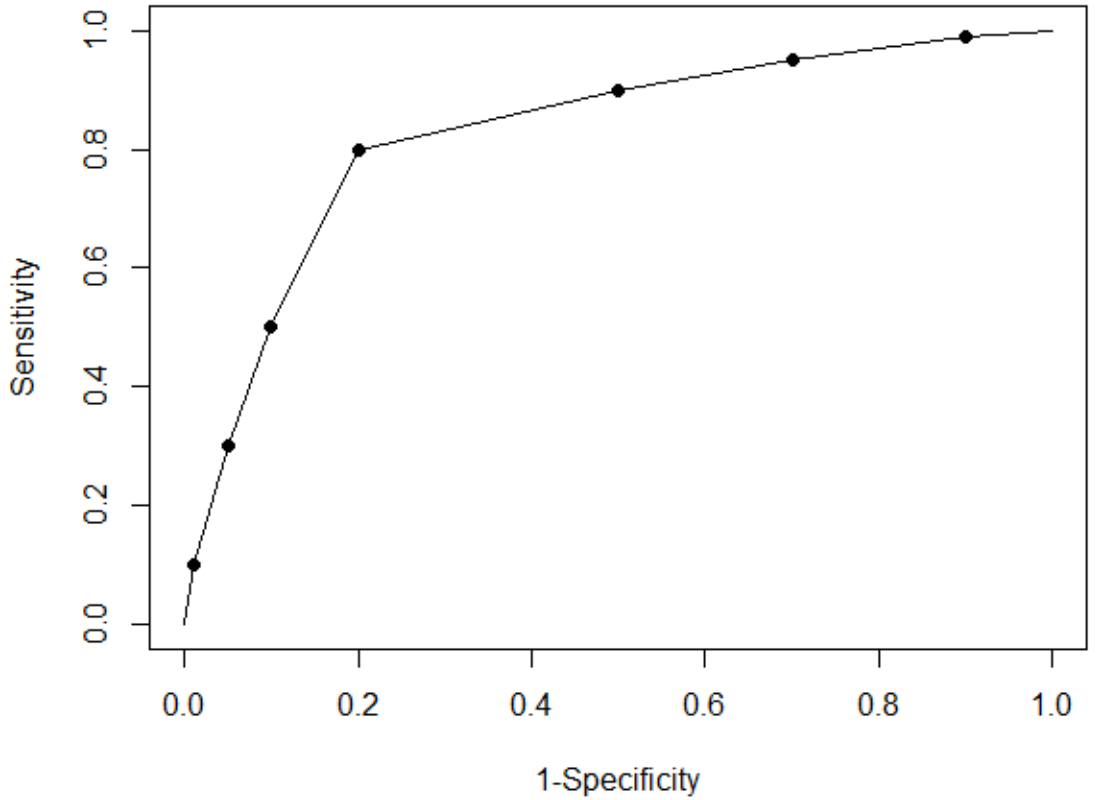
ROC Curve Example

- Example: Suppose we use peak expiratory flow rate (PEFR) as a screener for asthma. What PEFR cutoff should we use to indicate a positive test for asthma?

ROC Curve Example

- Example: Suppose we use peak expiratory flow rate (PEFR) as a screener for asthma. What PEFR cutoff should we use to indicate a positive test for asthma?

T+ Cutoff	Sensitivity	Specificity
PEFR < 100 L/min	0.10	0.99
PEFR < 200 L/min	0.30	0.95
PEFR < 300 L/min	0.50	0.90
PEFR < 400 L/min	0.80	0.80
PEFR < 500 L/min	0.90	0.50
PEFR < 600 L/min	0.95	0.30
PEFR < 700 L/min	0.99	0.10



Important Points

- Identify different types of data
- Concepts of a variable and an event
- Basic probability calculations
- Sensitivity, specificity, positive predictive value, negative predictive value, and ROC curves