STA 321
Spring 2016

Lecture 5

Thursday, January 28

What is Statistics?

Methods for Collecting, Describing, Analyzing, and Drawing Conclusions from Data

These methods are used for…

Design
- Planning research studies
- How best to obtain the required data

Description
- Summarizing data
- Exploring patterns in the data
- Extract/condense information
- Graphical pictures of the data

Inference
- Make predictions based on the data
- "Infer" from sample to population
- Generalize

Basic Terminology I

- **Population**
  - total set of all subjects of interest
  - the entire group of people, animal or things about which we want information
- **Elementary Unit**
  - any individual member of the population
- **Sample**
  - subset of the population from which the study actually collects information
  - used to draw conclusions about the whole population

Basic Terminology II

- **Variable**
  - a characteristic of a unit that can vary among subjects in the population/sample
  - Examples: gender, nationality, age, income, hair color, height, disease status, company rating, grade in STA 321, state of residence
- **Sampling Frame**
  - listing of all the units in the population
- **Parameter**
  - numerical characteristic of the **p**opulation
  - calculated using the whole population
- **Statistic**
  - numerical characteristic of the **s**ample
  - calculated using the sample

Recall Example from Lecture 5

- The Current Population Survey of about 60,000 households in the United States indicated that 5.3% of married couple households in the United States have annual income below the poverty level.
- Is this number a statistic or a parameter?

Modified Example

- A census of all households in Lexington indicated that 6.2% of married couple households in Lexington have annual income below the poverty level.

- Is this number a statistic or a parameter?

Univariate vs Multivariate

- Univariate data set
  - Consists of observations on a single attribute
- Multivariate data
  - Consists of observations on several attributes
- Special case: Bivariate data
  - Two attributes collected per observation

Why is it important to distinguish between different types of data?

- Some statistical methods only work for quantitative variables, others are designed for qualitative variables.

Nominal - Ordinal - Interval

Qualitative (Categorical) — Quantitative

Lowest level — Highest Level

- most information
- best statistical methods

Discrete and Continuous

- If a variable can take only a finite number of values, it is discrete
- Examples: number of children, number of teeth, etc
- Qualitative (categorical) variables are *always* discrete

Discrete and **Continuous**

- Continuous variables can (in theory) take an *infinite continuum* of possible real number values
- Example: time spent on STA 321 homework
  - can be 63 min. or 85 min.
    or 27.358 min. or 27.35769 min. or ...
  - can be **subdivided**
  - therefore **continuous**

Discrete or Continuous

- Quantitative variables can be discrete or continuous
- How about age, income, height?
- **It depends** on the scale
- Age is potentially continuous, but usually measured in years (discrete)

- NB: The distinction between discrete and continuous is not as important as the one between nominal/ordinal/interval

Statistical model

Statistical model consists of the following elements:

- Identification of random variables of interest
- Specification of joint distributions
- Specification for unknown parameter θ

Example

Height and age are each probabilistically distributed over humans.

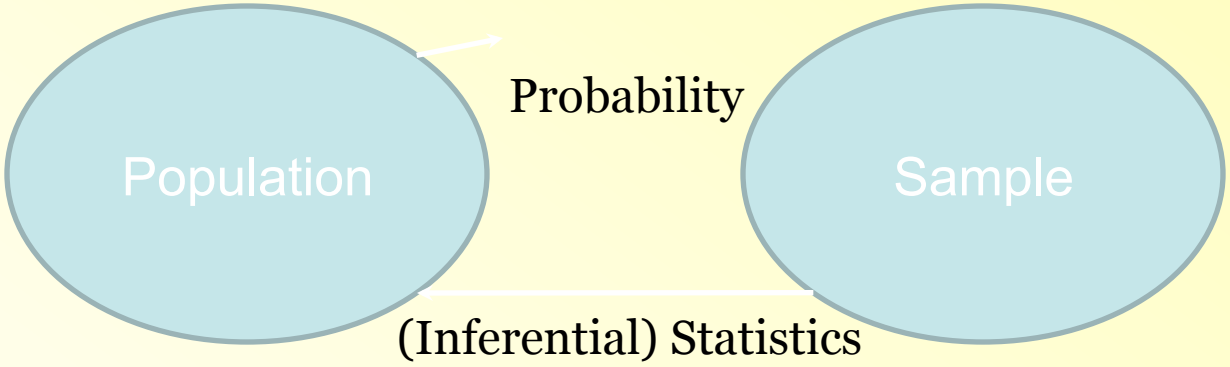
when we know that a person is of age 10, this influences the chance of the person being 6 feet tall. We could formalize that relationship in a linear regression model with the following form: $\text{height}_i = b_0 + b_1\text{age}_i + \varepsilon_i$, where b_0 is the intercept, b_1 is a parameter that age is multiplied by to get a prediction of height, ε is the error term, and i identifies the person. This implies that height is predicted by age, with some error.

Three purposes for a statistical model

- Predictions
- Extraction of information
- Description of stochastic structures

Probability

- Abstract but necessary because this is the mathematical theory underlying all statistical inference



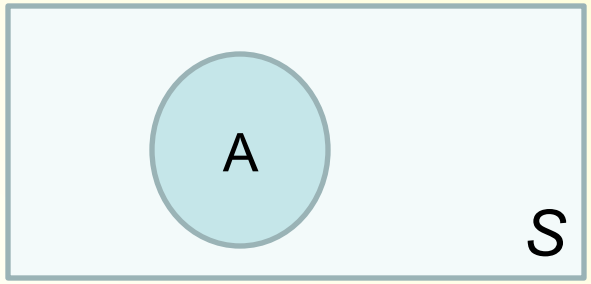
- Fundamental concepts that are very important to understanding *Sampling Distribution*, *Confidence Interval*, and *P-Value*
- Our goal is to learn the rules involved with assigning probabilities to events

Probability: Basic Terminology

- **Experiment**: Any activity from which an outcome, measurement, or other such result is obtained.
- **Random (or Chance) Experiment**: An experiment with the property that the outcome cannot be predicted with certainty.
- **Outcome**: Any possible result of an experiment.
- **Sample Space**: The collection of all possible outcomes of an experiment.
- **Event**: A specific collection of outcomes.
- **Simple Event**: An event consisting of exactly one outcome.

Complement

- Let A denote an event.
- The **complement** of an event A: All the outcomes in the sample space S that do not belong to the event A. The complement of A is denoted by A^c



$$P(A^c) = 1 - P(A)$$

Law of Complements

Union and Intersection

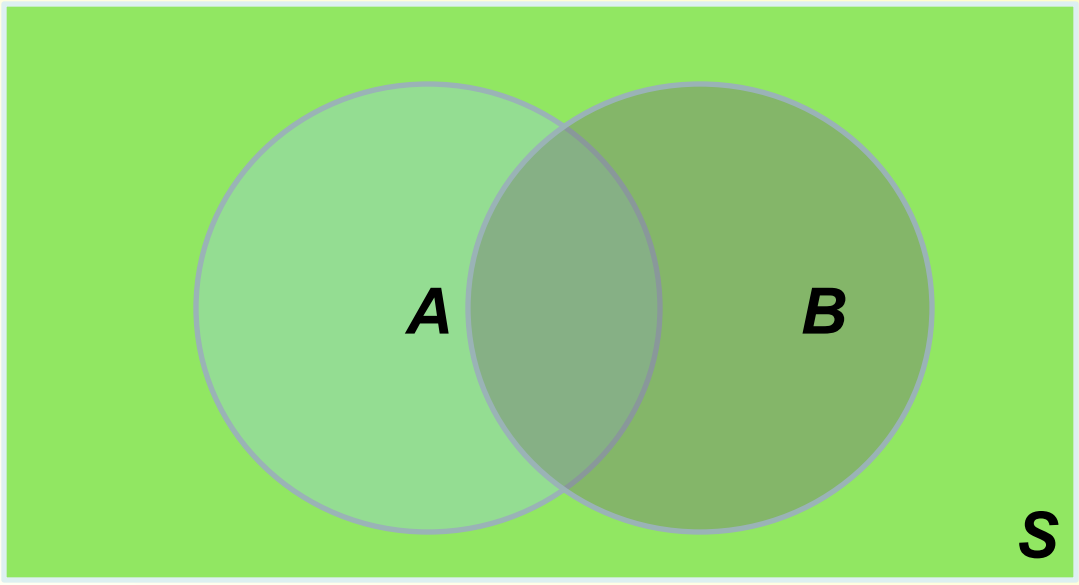
- Let A and B denote two events.

- The **union** of two events: All the outcomes in S that belong to at least one of A or B. The union of A and B is denoted by $A \cup B$
- The **intersection** of two events: All the outcomes in S that belong to both A and B. The intersection of A and B is denoted by $A \cap B$

Additive Law of Probability

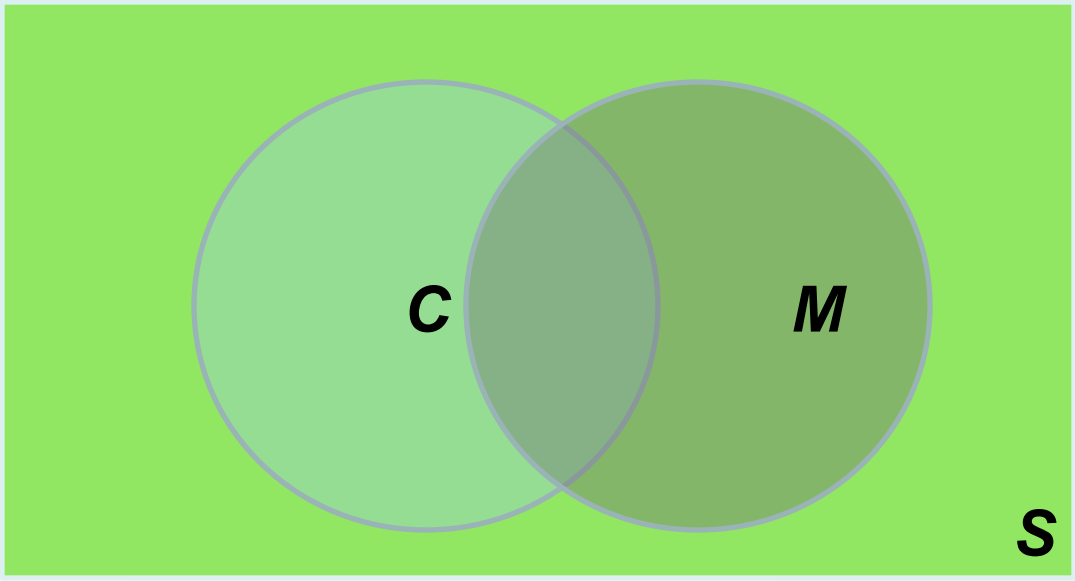
- Let A and B be *any* two events in the sample space S. The probability of the union of A and B is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



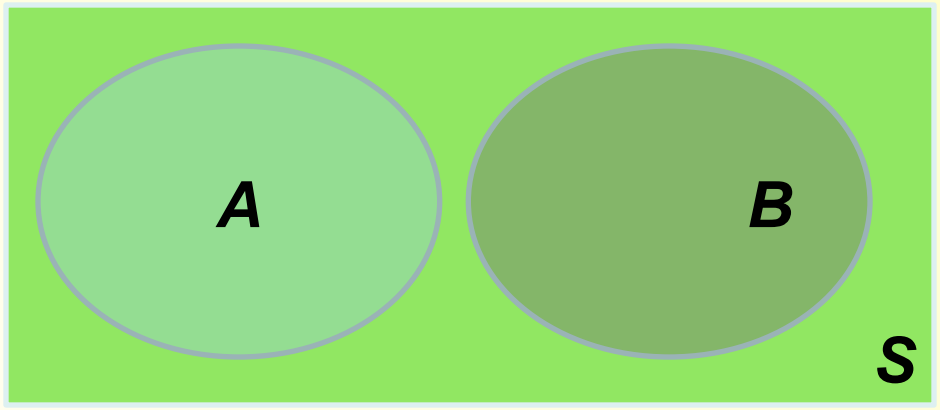
Additive Law of Probability

Example: At a large University, all first-year students must take chemistry and math. Suppose 85% pass chemistry, 88% pass math, and 78% pass both. Suppose a first-year student is selected at random. What is the probability that this student passed at least one of the courses?



Disjoint Sets

- Let *A and B denote two events*.
- **Disjoint (mutually exclusive) events**: *A* and *B* are said to be *disjoint* if there are no outcomes common to both *A* and *B*.
- The notation for this is written as $A \cap B = \{\ \} = \phi$
- Note: The last symbol denotes the null set or the empty set.



Assigning Probabilities to Events

- The probability of an event is a value between 0 and 1.

- In particular:
  - 0 implies that the event will never occur
  - 1 implies that the event will always occur

- How do we assign probabilities to events?

Assigning Probabilities to Events

- There are different approaches to assigning probabilities to events
- Objective
  - **equally likely outcomes (classical approach)**
  - **relative frequency**
- Subjective

Probabilities of Events

Let A be the event $A = \{o_1, o_2, \ldots, o_k\}$, where o_1, o_2, …, o_k are k different outcomes. Then

$$P(A)=P(o_1)+P(o_2)+\ldots+P(o_k)$$

Problem: The number on a license plate is any digit between 0 and 9. What is the probability that the first digit is a 3? What is the probability that the first digit is less than 4?

Conditional Probability & the Multiplication Rule

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \text{ provided } P(B) \neq 0$$

- Note: $P(A|B)$ is read as "the probability that A occurs given that B has occurred."
- Multiplied out, this gives *the multiplication rule:*

$$P(A \cap B) = P(B) \times P(A \mid B)$$

Multiplication Rule Example

- The multiplication rule:

$$P(A \cap B) = P(B) \times P(A \mid B)$$

- Ex.: A disease which occurs in .001 of the population is tested using a method with a false-positive rate of .05 and a false-negative rate of .05. If someone is selected and tested at random, what is the probability they are positive, and the method shows it?

Independence

- If events A and B are independent, then the events A and B have no influence on each other.

- So, the probability of A is unaffected by whether B has occurred.

- Mathematically, if A is independent of B, we write: $P(A|B) = \mathrm{P}(A)$

Multiplication Rule and Independent Events

Multiplication Rule for Independent Events: Let A and B be two independent events, then

$$P(A \cap B) = P(A)P(B).$$

Examples:

- Flip a coin twice. What is the probability of observing two heads?
- Flip a coin twice. What is the probability of getting a head and then a tail? A tail and then a head? One head?
- Three computers are ordered. If the probability of getting a "working" computer is 0.7, what is the probability that all three are "working" ?

Conditional Probabilities—Another Perspective

Example: Smoking and Lung Disease I

Joint Probabilities	Lung Disease	Not Lung Disease	*Row Totals*
Smoker	.12	.19	.31
Nonsmoker	.03	.66	.69
Column Totals	.15	.85	1.00

Conditional Probabilities—Another Perspective

Example: Smoking and Lung Disease I

Joint Probabilities	Lung Disease	Not Lung Disease	*Row Totals*
Smoker	.12	.19	.31
Nonsmoker	.03	.66	.69
Column Totals	.15	.85	1.00

Example: Smoking and Lung Disease II

Conditional Row Probabilities	Lung Disease	Not Lung Disease	*Row Totals*
Smoker	.12/.31 =.39	.19/.31 =.61	.31/.31 =1.00
Nonsmoker	.03/.69 =.04	.66/.69 =.96	.69/.69 =1.00
Smokers and Nonsmokers	.15	.85	1.00

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Conditional Probabilities—Another Perspective

Example: Smoking and Lung Disease I

Joint Probabilities	Lung Disease	Not Lung Disease	*Row Totals*
Smoker	.12	.19	.31
Nonsmoker	.03	.66	.69
Column Totals	.15	.85	1.00

Example: Smoking and Lung Disease III

Conditional Column Probabilities	Lung Disease	Not Lung Disease	*Lung Disease and Not Lung Disease*
Smoker	.12/.15 =.80	.19/.85 =.22	.31
Nonsmoker	.03/.15 =.20	.66/.85 =.78	.69
Column Totals	.15/.15 =1.00	.85/.85 =1.00	1.00

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Law of total probability

- For E and F are evenets,

$$P(E) = P(E \cap F) + P(E \cap F^c)$$

- Example: *A machine produces parts that are either good (90%), slightly defective (2%), or obviously defective (8%). Now assume that a one-year warranty is given for the parts that are shipped to customers. Suppose that a good part fails within the first year with probability 0.01, while a slightly defective part fails within the first year with probability 0.10. What is the probability that a customer receives a part that fails within the first year and therefore is entitled to a warranty replacement?*

Partition

- A collection of events $\{A_1, A_2, \ldots, A_k\}$ to be said a partition of a sample space S if $A_i \cap A_j$ is empty set.

Example: A is any event. Then $\{A, A^c\}$ is a partition.

Example: *A machine produces parts that are either good (90%), slightly defective (2%), or obviously defective (8%).*

Bayes Rule

- For a given partition of S into sets $F_1, \ldots, F_n$, we want to know the probability that some particular case, F_j, occurs, given that some event E occurs. We can compute this easily using the definition
$P(F_j|E) = P(F_j \cap E)/P(E)$

- This is called Bayes Formula. By applying the Law of Total Probability, we can rewrite the denominator:
$P(E) = \Sigma^n_{i=1} P(E|F_i)P(F_i)$.

- Thus, $P(F_j|E) = P(F_j)\, P(E|F_j) / \Sigma^n_{i=1} P(E|F_i)P(F_i)$.

- $P(F_j)$ is called prior and $P(F_j|E)$ is called posterior distributions.

Example

- *Urn 1 contains 5 white balls and 7 black balls. Urn 2 contains 3 whites and 12 black. A fair coin is flipped; if it is Heads, a ball is drawn from Urn 1, and if it is Tails, a ball is drawn from Urn 2. Suppose that this experiment is done and you learn that a white ball was selected. What is the probability that this ball was in fact taken from Urn 2? (i.e., that the coin flip was Tails)*

Terminology

- $P(A \cap B) = P(A,B)$ joint probability of A and B (of the intersection of A and B)
- $P(A|B)$ conditional probability of A given B
- $P(A)$ marginal probability of A