Introduction to Probability

STA 032: Gateway to data science Lecture 12

Jingwei Xiong

April 28, 2023

Reminders

- HW 3 due May 3 12pm.
  - Please start the homework as soon as possible.
- **Midterm 1** (Open book, take home, approximate 2 hours, time limit 5 hours)
  - **Due April 29 midnight, cover lecture 1-12**
  - Preparing guides: Finish Homework 1-3, be familiar to the lecture slides.
  - You can copy any your own codes in homework 1-3 to finish the open book exam.
- The PDF version of slides can be found on Canvas.
  - You can use search on it to find function examples.

Today

- Statistical inference overview
- Introduction to probability
- Permutations and combinations

Statistics definition

> A statistic or *sample statistic* is any quantity computed from values in a sample which is considered for a statistical purpose.

- Recall: Population vs. sample
  - A **sample** is a portion or **subset** of the larger **population**
  - **Population**: the entire group we would like to make conclusions about
  - **Sample**: specific group we have collected data from
  - Population **parameter**, e.g., population mean
  - This is a fixed quantity
  - Sample **statistic**, e.g., sample mean
  - depends on sample selected.

> Specifically, statistic is a **function** of a sample.

Statistic inference background setting (formal definition)

Overview | Population distribution | Sample, statistics

- Population follows some **statistical model**: e.g: normal distribution.
  - This is called **population distribution**.
  - This is an assumption!
- Population parameter is the population distribution parameter. e.g: μ and σ of normal distribution.
- Sample are random variables drawn from the population distribution
- Statistics are functions of a sample.
  - e.g: Sample mean, sample variance.

Statistic inference background setting (formal definition)

Overview | Population distribution | Sample, statistics

> Any statistical inference requires some assumptions. The population is something we don't know and wants to infer,

- We need assumptions about population distribution: **population distribution**
  - We pick a statistic model, and assume the population follows that model (distribution)
  - The population distribution will have (population) parameters. *Assumed* **Unknown!**
  - We will cover those distribution details later.

Population	Model	Parameters	Math form
Student heights	Normal distribution	μ, σ	$N(\mu, \sigma)$
Heads in **fixed** n coin toss	Binomial distribution	p (Is n a parameter?)	$Binomial(n, p)$
Waiting time of a service center	Exponential distribution	θ	$exponential(\theta)$

- What happens if we use a wrong population distribution?

Statistic inference background setting (formal definition)

Overview | Population distribution | Sample, statistics

- Samples are drawn from the population model. Denoted as:$X_1,...,X_n \sim Population \ model$
- Sample size is how many samples you get from the population. n here.
- Statistics is a function of the sample:
  - Sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - Sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
  - Sample maximum: $max(X)$
- Special case: The first sample X_1 is also a function of the sample, thus it is a statistic! (Though it make no sense)

Statistical Inference

- **Statistical inference** allows us to draw conclusions about the larger population. It is the process of using sample data to make conclusions about the underlying population the sample came from.

Similar to tasting a spoonful of soup while cooking to make an inference about the entire pot.

- This is one goal of statistics: to answer a question, by making **inferences** about a **population** based on data in one or more **samples**.
- The validity of our inferences depends on a variety of factors, including the **representativeness** of the sample

Statistical inference

- In order to draw principled conclusions from our data, we rely on a formal probabilistic framework that allows us to **quantify uncertainty**.
- **Statistical inference** is built upon the foundation of **probability theory**.
- Many topics in statistical inference
  - Fundamentals: probability, distributions, random variables, ...
  - Sampling
  - Hypothesis testing
  - Point estimates and confidence intervals
  - Modeling: Linear regression, analysis of variance, nonparametric models, machine learning, ...

Probability

- The **probability** of an event tells us how likely an event is to occur, and it can take values from 0 to 1, inclusive.
- Coin flipping: What is the chance of getting heads?
- Die rolling: What is the chance of rolling a 1?
- Y chromosomes in the US population: 51.2% of births are to babies with Y chromosomes, and 48.8% of babies do not have Y chromosomes. Thus the probability of having a baby with a Y chromosome is 0.512.

Formalizing Probability

- **Random process** gives rise to an **outcome** with an associated probability
  - Flip coin: heads or tails
  - Roll die: 1, 2, 3, 4, 5 or 6
- The **probability** of an outcome is the proportion of times the outcome would occur if we observed the random process an *infinite* number of times.
- An outcome or set of outcomes are called **events**
- E.g., A is the event that a die roll results in 1 or 2
  - In set notation: $A = \{1, 2\}$

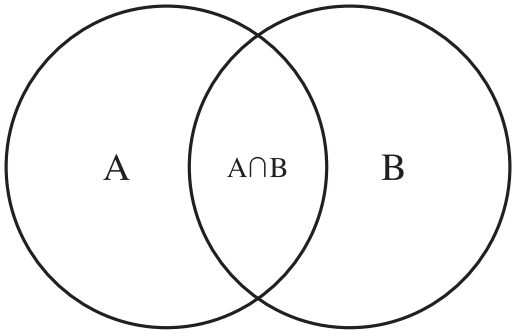
Formalizing Probability

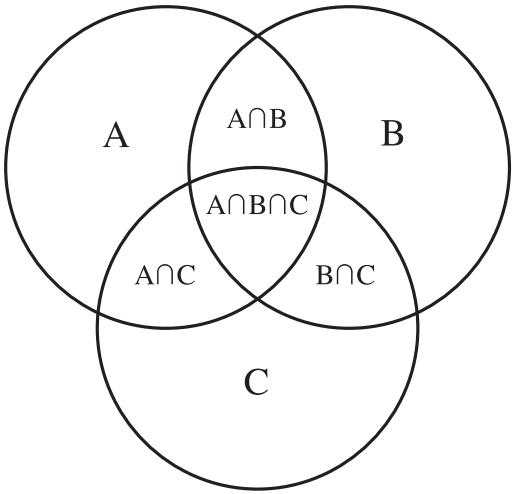
- A **Sample space** is the set of all possible outcomes, often denoted by $\mathcal{S}$
  - Flipping a coin: {H, T}
  - Rolling a die: {1, 2, 3, 4, 5, 6}
- **Probability** is usually denoted by p, P, or $\mathbb{P}$, e.g., $\mathbb{P}(A) = 1$, where A is the event
- Always takes values between 0 and 1 (inclusive); sometimes represented as a percentage, e.g., 50%
- So we have $\mathbb{P}(\text{flipping H}) = .5$

Operations on Events

- The **union** of A and B, denoted $A \cup B$, is the event that A, or B, or both A and B, occur. E.g.,:
  - A is the event that a person smokes
  - B is the event that a person identifies as female
  - $A \cup B$ is the event that a person is either a smoker, or identifies as female, or both.
  - "or" is inclusive, i.e., "and/or" in everyday language
- The **intersection** of A and B, denoted $A \cap B$, is the event that both A and B occur. Using the same example,
  - $A \cap B$ is the event that a person both smokes and identifies as female

Venn Diagram





- $A \cup B$ includes both circles
- Will sometimes see a rectangle encapsulating the Venn diagram; this denotes the sample space

Disjoint or mutually exclusive events

- A and B are **disjoint** or **mutually exclusive** if $A \cap B = \emptyset$
  - A and B cannot occur simultaneously
  - Flipping heads and flipping tails are disjoint events
  - Smoking and identifying as female are *not* disjoint; they can happen at the same time
- The **complement** of A, denoted A^c or $\overline{A}$, is the event A does not occur.
- In other words, A^c represents all outcomes in the sample space that are not in A
- A and A^c are **disjoint**.

Probability Rules

- The probability of any event in the sample space is between 0 and 1, inclusive.
- The probability of the entire sample space is 1.
- If we know the probability of A, often denoted $P(A)$, it is easy to calculate the probability of A^c as $P(A^c) = 1 - P(A)$.
- This is called the **complement rule**: $P(A) + P(A^c) = 1$.

Additive Rule of Probability

- When events are mutually exclusive (cannot occur together), $P(A \cup B) = P(A) + P(B)$. In a Venn diagram, the circles representing A and B do not intersect.
- When two events can occur simultaneously (think about the overlapping sections in the Venn diagrams), then we need to avoid double-counting when calculating the probability either of two events will occur.
- The general **additive rule of probability** is therefore $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $A \cap B$ is part of the event A and part of the event B, we need to avoid double-counting it.
    - Note that if two events A and B are mutually exclusive, then $P(A \cap B) = 0$.

Computing Probabilities

Intuitively, we can think of the probability of an outcome (or set of outcomes) as the proportion of times the outcome (or set of outcomes) would occur if we observed the random process infinitely many times.

If all the outcomes in our random process (sample space $\mathcal{S}$) are **equally likely**, then for some event E,

$$P(E) = \frac{\#\text{ of outcomes in } E}{\#\text{ of total outcomes in } \mathcal{S}}.$$

- Rolling die:
  - Let A be the event that we roll a 1 or a 6
  - $P(A) = \frac{2}{6}$

Vaccine Hesitancy in the UK

Contents lists available at ScienceDirect

Brain Behavior and Immunity

journal homepage: www.elsevier.com/locate/ybrbi

Predictors of COVID-19 vaccine hesitancy in the UK household longitudinal study

Elaine Robertson [a,1], Kelly S. Reeve [b,1], Claire L. Niedzwiedz [c], Jamie Moore [b], Margaret Blake [d], Michael Green [a], Srinivasa Vittal Katikireddi [a,e,*,1], Michaela J. Benzeval [b,1]

We will consider this UK study on vaccine hesitancy, which explored hesitancy across a variety of UK ethnic groups.

The study collected information on 10,000 people. This was a sample in the original study, but for our purposes (to illustrate probability concepts) we are going to treat this as a population.

Vaccine Hesitancy

Ethnicity	Vaccine Hesitant	Not Hesitant
White British or Irish	1362	7368
Other white background	71	199
Mixed	55	115
Asian or Asian British - Indian	37	143
Asian or Asian British - Pakistani/Bangladeshi	85	115
Asian or Asian British - other	15	95
Black or Black British	136	54
Other Ethnic Group or Not Specified	31	119

Practicing with Probabilities

Define events A = vaccine hesitant and B = Asian or Asian British-Indian. Calculate the following probabilities for a randomly-selected person from the 10,000 people

- $P(A)$
- $P(B)$
- $P(A \cap B)$
- $P(A \cup B)$
- $P(A \cup B^c)$

Vaccine Hesitancy

Ethnicity	Vaccine Hesitant	Not Hesitant
White British or Irish	1362	7368
Other white background	71	199
Mixed	55	115
Asian or Asian British - Indian	37	143
Asian or Asian British - Pakistani/Bangladeshi	85	115
Asian or Asian British - other	15	95
Black or Black British	136	54
Other Ethnic Group or Not Specified	31	119

- A = vaccine hesitant; B = Asian or Asian British-Indian
- Want $P(A)$, $P(B)$, $P(A \cap B)$, $P(A \cup B)$, $P(A \cup B^c)$

Vaccine Hesitancy

Ethnicity	Vaccine Hesitant	Not Hesitant
White British or Irish	1362	7368
Other white background	71	199
Mixed	55	115
Asian or Asian British - Indian	37	143
Asian or Asian British - Pakistani/Bangladeshi	85	115
Asian or Asian British - other	15	95
Black or Black British	136	54
Other Ethnic Group or Not Specified	31	119

- A = vaccine hesitant; B = Asian or Asian British-Indian
- $P(A) = \frac{1362+71+55+37+85+15+136+31}{10000} = \frac{1792}{10000} = .1792$
- $P(B) = \frac{37+143}{10000} = .018$
- $P(A \cap B) = \frac{37}{10000} = .0037$
- $P(A \cup B) = \frac{1362+71+55+37+85+15+136+31+143}{10000} = .1935$
- $P(A \cup B^c) = \frac{1935-143+7368+199+115+115+95+54+119}{10000} = .9875$
  - Or, $1 - \frac{143}{10000}$

Probability distribution

- A **probability distribution** is a table of all **disjoint** outcomes and their associated probabilities.

Ethnicity	Vaccine Hesitant	Not Hesitant
White British or Irish	$\frac{1362}{10000}$	$\frac{7368}{10000}$
Other white background	$\frac{71}{10000}$	$\frac{199}{10000}$
Mixed	$\frac{55}{10000}$	$\frac{115}{10000}$
Asian or Asian British - Indian	$\frac{37}{10000}$	$\frac{143}{10000}$
Asian or Asian British - Pakistani/Bangladeshi	$\frac{85}{10000}$	$\frac{115}{10000}$
Asian or Asian British - other	$\frac{15}{10000}$	$\frac{95}{10000}$
Black or Black British	$\frac{136}{10000}$	$\frac{54}{10000}$
Other Ethnic Group or Not Specified	$\frac{31}{10000}$	$\frac{119}{10000}$

Rules for Probability Distributions

A probability distribution is a list of the possible outcomes with corresponding probabilities that satisfies three rules:

1. The outcomes listed must be disjoint.
2. Each probability must be between 0 and 1.
3. The probabilities must total 1.

The fundamental principle of Counting

The fundamental counting principle states that if there are p ways to do one thing, and q ways to do another thing, then there are $p \times q$ ways to do both things.

Example: Build your own Pizza: 2 crusts, 4 toppings, 3 sizes. How many possible combinations?

> $2 * 4 * 3 = 24$

Permutation

A permutation is an ordering of a collection of objects. The number of permutations of n objects is:

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1$$

- n factorial: the product of the integers from 1 to n. $1! = 1, 0! = 1$

Example: 8 students sitting in the first row of class. How many ways can I arrange the 8 students?

> $8! = 40320$

```
factorial(8)
```

```
[1] 40320
```

Permutations for different items

- n items
- Select k of those n items (without replacement)
- consider ABC different from BCA (order counts)

$$ {}_nP_k = \frac{n!}{(n-k)!} $$

Example: 12 songs in playlists. How many different ordering of 3 songs?

> $$ {}_{12}P_3 = \frac{12!}{(12-3)!} = 12 \times 11 \times 10 $$

Permutation for identical items (not all unique)

- n items, not all unique.
- Total r groups, group $i, (1 \leq i \leq r)$ has k_i subjects.
- How many orders do we have?

$$\frac{n!}{k_1!k_2!\ldots k_r!}$$

Example: How many ways can the letters in STATISTICS be arranged?

> n =10, S,T,A,I,C: $k_1 = 3, k_2 = 3, k_3 = 1, k_4 = 2, k_5 = 1$

$$\frac{10!}{3!3!1!2!1!} = 50400$$

Combinations

Combination: Each distinct group of objects that can be selected, without regard to order.

- Select k items out of n items
- ABC is the same as BCA< CBA, ACB. (order does not count)

The number of combinations of k items from n items is:

$$ {}_nC_k = \binom{n}{k} = \frac{n!}{(n-k)!k!} $$

Example: A state lottery involves the random selection of 6 different numbers between 1 and 26. If you select one six-number combination, what is the probability that it will be the winning number?

$$ n = 26, k = 6, {}_{26}C_6 = \binom{26}{6} = \frac{26!}{(20)!6!} = 230230 $$

$$ P(win) = \frac{1}{230230} $$

We assume all ways have equal probability to be chosen.

Summary

- Descriptive statistics vs. inferential statistics
- Statistical inference and probability theory
- Introduction to probability
  - Events, sample space
  - Probability rules
    - Complement rule: $P(A) + P(A^c) = 1$
    - Additive rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - Probability distributions
  - Counting rules
- Please pay attention to the Theory part today. There will be 15 points in the midterm 1 for today's problem.