Unit 5: Introduction to Probability

1. An excursion to probability

Stat 140 - 02

Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

Outline

1. Announcement

2. Review

3. This week: probability and inference

4. Main ideas
    1. Definitions of probability
    2. Basic probability
    3. A card example

5. Summary

Resources

- I have uploaded three R command summaries: *data wrangling*, *visualization* and *regression* on Piazza
- I have also uploaded an *example project proposal* on Piazza
- Let me show you how to load data in Rstudio
  - Kaggle (read_csv)
  - FiveThirtyEight: library(fivethirtyeight)

Outline

1. Announcement

2. Review

3. This week: probability and inference

4. Main ideas
    1. Definitions of probability
    2. Basic probability
    3. A card example

5. Summary

Review

In recent classes, we have laid the foundations for statistical inference

1. Confidence interval → estimate population parameter
2. Hypothesis testing → make a decision

1. **Hypothesis:** Start with two hypotheses about the population: the null hypothesis (H_0) and the alternative hypothesis (H_A).
2. **Collect data:** Choose a (representative) sample, compute the test statistic
3. **Judge the evidence:** Figure out how likely it is to see data like what we observed, if the null hypothesis were in fact true.
4. **Make a decision:** If our data would have been extremely unlikely under the null hypothesis, then we reject it and deem the alternative claim worthy of further study. Otherwise, we cannot reject the null claim.

Think hypothesis testing as a jury trial

Does the reasoning of hypothesis tests seem backward? That could be because we usually prefer to think about getting things right rather than getting them wrong. You have seen this reasoning before because it's the logic of jury trials.

Step 1: null hypothesis

Let's suppose a defendant has been accused of robbery. The null hypothesis is that the defendant is innocent. Instructions to juries are quite explicit about this.

Step 2: collect data

How is the null hypothesis tested? The prosecution first collects evidence. ("If the defendant were innocent, wouldn't it be remarkable that the police found him at the scene of the crime with a bag full of money in his hand, a mask on his face, and a getaway car parked outside?") For us, the data is the evidence.

Step 3: judge the evidence

The jury considers the evidence in light of the *presumption* of innocence and judges whether the evidence against the defendant would be plausible *if the defendant were in fact innocent*.

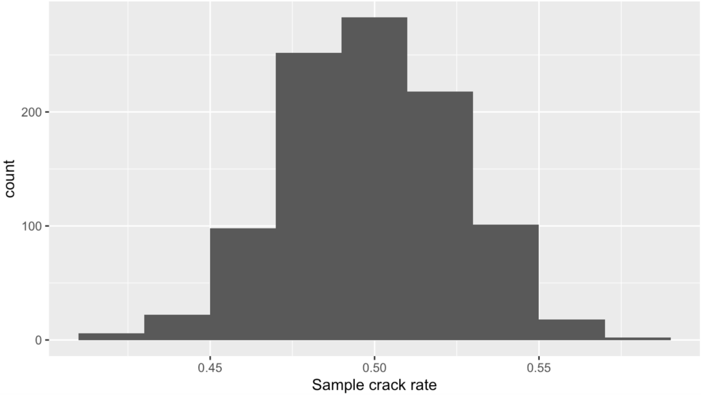
Step 4: make a decision

The standard of "beyond a reasonable doubt" is wonderfully ambiguous because it leaves the jury to decide the degree to which the evidence contradicts the hypothesis of innocence. Juries don't explicitly use probability to help them decide whether to reject that hypothesis. But when you ask the same question of your null hypothesis, you have the advantage of being able to quantify exactly how surprising the evidence would be if the null hypothesis were true.

Are the data surprising, given the null hypothesis?

Simulation approach

1. Sample data under the null hypothesis
2. Look at the distribution of their sample statistic
3. Proportion of times we get a sample like ours or more extreme?



Outline

1. Announcement

2. Review

3. This week: probability and inference

4. Main ideas
    1. Definitions of probability
    2. Basic probability
    3. A card example

5. Summary

Outline

1. Announcement

2. Review

3. This week: probability and inference

4. Main ideas
    1. Definitions of probability
    2. Basic probability
    3. A card example

5. Summary

Outline

1. Announcement

2. Review

3. This week: probability and inference

4. Main ideas
    1. Definitions of probability
    2. Basic probability
    3. A card example

5. Summary

(1) Definitions of probability

There are two ways to define the probability of an event.

A **frequentist** says that the probability of event A (or $P[A]$) is the proportion of times that A occurs in a infinite sequence (or very long run) of separate tries.[1]

$$P[A] = \lim_{n \to \infty} \frac{\text{\# times A happens}}{n}$$

A **Bayesian** can pick whatever number they prefer for $P[A]$, based on their own personal experience and intuition, provided that number is consistent with all of the other probabilities they choose in life.

[1] John Maynard Keynes (1883-1946) commented on this: In the long run, we are all dead.

Imagine tossing a coin, and let A denote the event of getting a head.

A **frequentist** must define

$$P[\text{head in a coin toss}]$$

as the limit of the proportion of heads in n tosses.

In contrast, a **Bayesian** might declare their personal belief about the coin based on symmetry, or knowledge of the integrity of the coin's owner, or divine inspiration.

The Bayesian's view must:

- conform to all other personal opinions
- change as new data arise according to Bayes' Rule

Kolmogorov's Axioms

Whether one is frequentist or Bayesian, all probabilities must obey Kolmogorov's Axioms:

- $0 \leq P[A] \leq 1$
- $P[\text{some possible event happens}] = 1$
  (one of the possible outcomes must occur).
- If A and B are incompatible (disjoint) events, then
  $P[A \text{ or } B] = P[A] + P[B]$.

Two events A and B are **disjoint** if it is impossible for both A and B to happen at the same time; e.g., you cannot throw a head and tail on the same toss.

Outline

1. Announcement

2. Review

3. This week: probability and inference

4. Main ideas
    1. Definitions of probability
    2. Basic probability
    3. A card example

5. Summary

These are the main definitions:

1. A and B are **independent**:

   $$P[A \text{ and } B] = P[A]P[B]$$

   Having information on A does not tell us anything about B (and vice versa).

2. A and B are **disjoint** (or incompatible, or mutually exclusive):

   $$P[A \text{ and } B] = 0$$

   A and B cannot happen at the same time.

Independent and disjoint do not mean the same thing.

Probability rules

These are the main rules:

1. complement rule

$$P[\text{not } A] = 1 - P(A)$$

2. inclusive or

$$P[A \text{ or } B] = P[A] + P[B] - P[A \text{ and } B]$$

3. conditional probability

$$P[A \text{ given that } B \text{ occurs}] = P[A|B] = P[A \text{ and } B]/P[B]$$

Are the data surprising, given the null hypothesis?

Probability approach

1. Imagine all samples under the null hypothesis
2. Look at the distribution of their sample statistic
3. Compute probability that we get a sample like ours or more extreme?

P-value

The probability of seeing data like these (or something even less likely) given that the null hypothesis is true

$$P(\text{data} \mid H_0)$$

1. Announcement

2. Review

3. This week: probability and inference

4. Main ideas
    1. Definitions of probability
    2. Basic probability
    3. A card example

5. Summary

A card example

A standard deck of 52 cards (13×4) contains 4 types:

$$\text{clubs}(\clubsuit), \text{diamonds}(\diamondsuit), \text{hearts}(\heartsuit), \text{spades}(\spadesuit)$$

Poll question

What's the probability of drawing a $\heartsuit$ card?

- a $1/2$
- b $1/4$
- c $1/52$

A card example

A standard deck of 52 cards (13×4) contains 4 types:

$$\text{clubs}(\clubsuit), \text{diamonds}(\diamondsuit), \text{hearts}(\heartsuit), \text{spades}(\spadesuit)$$

Poll question

What's the probability of drawing a red card?

a. $1/2$
b. $1/4$
c. $1/52$

A standard deck of 52 cards (13×4) contains 4 types:

$$\text{clubs}(\clubsuit), \text{diamonds}(\diamondsuit), \text{hearts}(\heartsuit), \text{spades}(\spadesuit)$$

Poll question

Let A be the event of drawing a $\heartsuit$ card, and let B be the event of drawing a red card? Are A and B independent?

a Yes

b No

Hint: What is $P(A \text{ and } B)$? What is $P(A)P(B)$?

A card example

A standard deck of 52 cards (13×4) contains 4 types:

$$\text{clubs}(\clubsuit), \text{diamonds}(\diamondsuit), \text{hearts}(\heartsuit), \text{spades}(\spadesuit)$$

Poll question

Let A be the event of drawing a $\heartsuit$ card, and let B be the event of drawing a red card? Are A and B disjoint?

a. Yes
b. No

A card example

A standard deck of 52 cards (13×4) contains 4 types:

$$\text{clubs}(\clubsuit), \text{diamonds}(\diamondsuit), \text{hearts}(\heartsuit), \text{spades}(\spadesuit)$$

Poll question

What is the probability you don't draw a heart?

a $1/2$

b $1/4$

c $3/4$

d $1/52$

e $51/52$

A standard deck of 52 cards (13×4) contains 4 types:

$$\text{clubs}(\clubsuit), \text{diamonds}(\diamondsuit), \text{hearts}(\heartsuit), \text{spades}(\spadesuit)$$

Poll question

What is the probability of you draw a red card or a heart?

a) $1/2$

b) $1/4$

c) $3/4$

A standard deck of 52 cards (13×4) contains 4 types:

$$\text{clubs}(\clubsuit), \text{diamonds}(\diamondsuit), \text{hearts}(\heartsuit), \text{spades}(\spadesuit)$$

Poll question

What is the probability of you draw a heart given that the card is red?

a. $1/2$
b. $1/4$
c. $3/4$

Outline

1. Announcement

2. Review

3. This week: probability and inference

4. Main ideas
    1. Definitions of probability
    2. Basic probability
    3. A card example

5. Summary

Summary of main ideas

1. Definitions of probability
2. Basic probability
3. A card example