Precept 1: Probability, Simulations, Working with Data

Soc 500: Applied Social Statistics

Shay O'Brien

Princeton University

September 13, 2018

Support Resources

- Office hours
- Piazza
- Email
- Google is your best friend!

Office hours

Please raise your hand if you absolutely cannot make any of these office hours:

- Mondays 10am - 12pm (S + Z)
- Mondays 4:30 - 6:30pm (A + S + Z)
- Wednesdays 10am - 12pm (A + S + Z)
- Thursdays 1pm - 3pm (A + S)
- Thursdays 5pm - 7pm (A + Z)

Learning objectives

- Apply the Law of Total Probability and the Bayes' Rule

Learning objectives

- Apply the Law of Total Probability and the Bayes' Rule
- Translate information provided in word problems into probability statements

Learning objectives

- Apply the Law of Total Probability and the Bayes' Rule
- Translate information provided in word problems into probability statements
- Create an R Markdown document

Learning objectives

- Apply the Law of Total Probability and the Bayes' Rule
- Translate information provided in word problems into probability statements
- Create an R Markdown document
- Run simulations (loops, functions, replicate)

Learning objectives

- Apply the Law of Total Probability and the Bayes' Rule
- Translate information provided in word problems into probability statements
- Create an R Markdown document
- Run simulations (loops, functions, replicate)
- Perform basic data manipulations and create summary tables

Learning objectives

- Apply the Law of Total Probability and the Bayes' Rule
- Translate information provided in word problems into probability statements
- Create an R Markdown document
- Run simulations (loops, functions, replicate)
- Perform basic data manipulations and create summary tables
- Learn an easy way to check your code style

Learning objectives

- Apply the Law of Total Probability and the Bayes' Rule
- Translate information provided in word problems into probability statements
- Create an R Markdown document
- Run simulations (loops, functions, replicate)
- Perform basic data manipulations and create summary tables
- Learn an easy way to check your code style

Learning objectives

- Apply the Law of Total Probability and the Bayes' Rule
- Translate information provided in word problems into probability statements
- Create an R Markdown document
- Run simulations (loops, functions, replicate)
- Perform basic data manipulations and create summary tables
- Learn an easy way to check your code style

Acknowledgements: These slides were most recently edited by Ziyao Tian, and they draw heavily on materials developed by past preceptors Shay O'Brien, Simone Zhang, Elisha Cohen, and Clark Bernier. Thanks!

Probability from a Contingency Table

- Contingency tables show frequency counts for each combination of two categorical variable values in a population or sample space.

Probability from a Contingency Table

- Contingency tables show frequency counts for each combination of two categorical variable values in a population or sample space.
- They are often used to represent the probability of two events.

Probability from a Contingency Table

- Contingency tables show frequency counts for each combination of two categorical variable values in a population or sample space.
- They are often used to represent the probability of two events.
- We're going to make a simply two-by-two contingency table together.

Probability from a Contingency Table

- Contingency tables show frequency counts for each combination of two categorical variable values in a population or sample space.
- They are often used to represent the probability of two events.
- We're going to make a simply two-by-two contingency table together.
- Our two events or variables are:

Probability from a Contingency Table

- Contingency tables show frequency counts for each combination of two categorical variable values in a population or sample space.
- They are often used to represent the probability of two events.
- We're going to make a simply two-by-two contingency table together.
- Our two events or variables are:
  - Is this your first Sociology course in Princeton (F)?

Probability from a Contingency Table

- Contingency tables show frequency counts for each combination of two categorical variable values in a population or sample space.
- They are often used to represent the probability of two events.
- We're going to make a simply two-by-two contingency table together.
- Our two events or variables are:
  - Is this your first Sociology course in Princeton (F)?
  - Have you done any R project before (R)?

Probability from a Contingency Table

- Contingency tables show frequency counts for each combination of two categorical variable values in a population or sample space.
- They are often used to represent the probability of two events.
- We're going to make a simply two-by-two contingency table together.
- Our two events or variables are:
  - Is this your first Sociology course in Princeton (F)?
  - Have you done any R project before (R)?
- Let's find out:

Probability from a Contingency Table

- Contingency tables show frequency counts for each combination of two categorical variable values in a population or sample space.
- They are often used to represent the probability of two events.
- We're going to make a simply two-by-two contingency table together.
- Our two events or variables are:
  - Is this your first Sociology course in Princeton (F)?
  - Have you done any R project before (R)?
- Let's find out:
  - What is the probability that a randomly selected person is taking Sociology course for the first time and has never done any R project before? $P(F, R^c)$?

Probability from a Contingency Table

- Contingency tables show frequency counts for each combination of two categorical variable values in a population or sample space.
- They are often used to represent the probability of two events.
- We're going to make a simply two-by-two contingency table together.
- Our two events or variables are:
    - Is this your first Sociology course in Princeton (F)?
    - Have you done any R project before (R)?
- Let's find out:
    - What is the probability that a randomly selected person is taking Sociology course for the first time and has never done any R project before? $P(F, R^c)$?
    - Given that a person has done R programming project before, what is the probability that SOC400 is this person's first Sociology course? $P(F|R)$?

A Card Player's Example

If we randomly draw two cards from a standard 52 card deck and define the events
$A = \{\text{Ace on Draw 1}\}$ and $B = \{\text{Ace on Draw 2}\}$, then

A Card Player's Example

If we randomly draw two cards from a standard 52 card deck and define the events
$A = \{\text{Ace on Draw 1}\}$ and $B = \{\text{Ace on Draw 2}\}$, then

- $P(A) =$

A Card Player's Example

If we randomly draw two cards from a standard 52 card deck and define the events
$A = \{\text{Ace on Draw 1}\}$ and $B = \{\text{Ace on Draw 2}\}$, then

- $P(A) = 4/52$

A Card Player's Example

If we randomly draw two cards from a standard 52 card deck and define the events
$A = \{\text{Ace on Draw 1}\}$ and $B = \{\text{Ace on Draw 2}\}$, then

- $P(A) = 4/52$
- $P(B|A) =$

A Card Player's Example

If we randomly draw two cards from a standard 52 card deck and define the events
$A = \{\text{Ace on Draw 1}\}$ and $B = \{\text{Ace on Draw 2}\}$, then

- $P(A) = 4/52$
- $P(B|A) = 3/51$

A Card Player's Example

If we randomly draw two cards from a standard 52 card deck and define the events
$A = \{\text{Ace on Draw 1}\}$ and $B = \{\text{Ace on Draw 2}\}$, then

- $P(A) = 4/52$
- $P(B|A) = 3/51$
- $P(A, B) = P(A) \times P(B|A) =$

A Card Player's Example

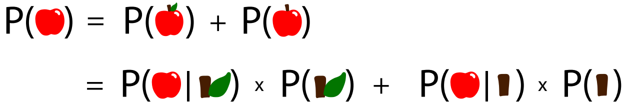
If we randomly draw two cards from a standard 52 card deck and define the events
$A = \{\text{Ace on Draw 1}\}$ and $B = \{\text{Ace on Draw 2}\}$, then

- $P(A) = 4/52$
- $P(B|A) = 3/51$
- $P(A, B) = P(A) \times P(B|A) = 4/52 \times 3/51 \approx .0045$

Law of Total Probability (LTP)

With 2 Events:

$$
\begin{aligned}
P(B) &= P(B, A) + P(B, A^c) \\
&= P(B|A) \times P(A) + P(B|A^c) \times P(A^c)
\end{aligned}
$$



Recall, if we randomly draw two cards from a standard 52 card deck and define the events $A = \{\text{Ace on Draw 1}\}$ and $B = \{\text{Ace on Draw 2}\}$, then

- $P(A) = 4/52$
- $P(B|A) = 3/51$
- $P(A, B) = P(A) \times P(B|A) = 4/52 \times 3/51$

Question: $P(B) = ?$

Confirming Intuition with the LTP

Confirming Intuition with the LTP

$$
\begin{aligned}
P(B) &= P(B, A) + P(B, A^c) \\
&= P(B|A) \times P(A) + P(B|A^c) \times P(A^c)
\end{aligned}
$$

Confirming Intuition with the LTP

$$\begin{aligned} P(B) &= P(B, A) + P(B, A^c) \\ &= P(B|A) \times P(A) + P(B|A^c) \times P(A^c) \end{aligned}$$

$$\begin{aligned} P(B) &= 3/51 \times 1/13 + 4/51 \times 12/13 \\ &= \frac{3+48}{51 \times 13} = \frac{1}{13} = \frac{4}{52} \end{aligned}$$

Example: Voter Mobilization

Suppose that we have put together a voter mobilization campaign and we want to know what the **probability of voting** is after the campaign: $P(\text{vote})$.

Example: Voter Mobilization

Suppose that we have put together a voter mobilization campaign and we want to know what the **probability of voting** is after the campaign: $P(\text{vote})$. We know the following:

Example: Voter Mobilization

Suppose that we have put together a voter mobilization campaign and we want to know what the **probability of voting** is after the campaign: $P(\text{vote})$. We know the following:

- $P(\text{vote}|\text{mobilized}) = 0.75$

Example: Voter Mobilization

Suppose that we have put together a voter mobilization campaign and we want to know what the **probability of voting** is after the campaign: $P(\text{vote})$. We know the following:

- $P(\text{vote}|\text{mobilized}) = 0.75$
- $P(\text{vote}|\text{not mobilized}) = 0.15$

Example: Voter Mobilization

Suppose that we have put together a voter mobilization campaign and we want to know what the **probability of voting** is after the campaign: $P(\text{vote})$. We know the following:

- $P(\text{vote}|\text{mobilized}) = 0.75$
- $P(\text{vote}|\text{not mobilized}) = 0.15$
- $P(\text{mobilized}) = 0.6$ and so $P(\text{not mobilized}) = 0.4$

Example: Voter Mobilization

Suppose that we have put together a voter mobilization campaign and we want to know what the **probability of voting** is after the campaign: $P(\text{vote})$. We know the following:

- $P(\text{vote}|\text{mobilized}) = 0.75$
- $P(\text{vote}|\text{not mobilized}) = 0.15$
- $P(\text{mobilized}) = 0.6$ and so $P(\text{not mobilized}) = 0.4$

Note that mobilization **partitions** the data. Everyone is either mobilized or not.

Example: Voter Mobilization

Suppose that we have put together a voter mobilization campaign and we want to know what the **probability of voting** is after the campaign: $P(\text{vote})$. We know the following:

- $P(\text{vote}|\text{mobilized}) = 0.75$
- $P(\text{vote}|\text{not mobilized}) = 0.15$
- $P(\text{mobilized}) = 0.6$ and so $P(\text{not mobilized}) = 0.4$

Note that mobilization **partitions** the data. Everyone is either mobilized or not. Thus, we can apply the LTP:

Example: Voter Mobilization

Suppose that we have put together a voter mobilization campaign and we want to know what the **probability of voting** is after the campaign: $P(\text{vote})$. We know the following:

- $P(\text{vote}|\text{mobilized}) = 0.75$
- $P(\text{vote}|\text{not mobilized}) = 0.15$
- $P(\text{mobilized}) = 0.6$ and so $P(\text{not mobilized}) = 0.4$

Note that mobilization **partitions** the data. Everyone is either mobilized or not. Thus, we can apply the LTP:

$$\begin{aligned} P(\text{vote}) =& P(\text{vote}|\text{mobilized})P(\text{mobilized}) + \\ & P(\text{vote}|\text{not mobilized})P(\text{not mobilized}) \end{aligned}$$

Example: Voter Mobilization

Suppose that we have put together a voter mobilization campaign and we want to know what the **probability of voting** is after the campaign: $P(\text{vote})$. We know the following:

- $P(\text{vote}|\text{mobilized}) = 0.75$
- $P(\text{vote}|\text{not mobilized}) = 0.15$
- $P(\text{mobilized}) = 0.6$ and so $P(\text{not mobilized}) = 0.4$

Note that mobilization **partitions** the data. Everyone is either mobilized or not. Thus, we can apply the LTP:

$$\begin{aligned} P(\text{vote}) =& P(\text{vote}|\text{mobilized})P(\text{mobilized}) + \\ & P(\text{vote}|\text{not mobilized})P(\text{not mobilized}) \\ =& 0.75 \times 0.6 + 0.15 \times 0.4 \end{aligned}$$

Example: Voter Mobilization

Suppose that we have put together a voter mobilization campaign and we want to know what the **probability of voting** is after the campaign: $P(\text{vote})$. We know the following:

- $P(\text{vote}|\text{mobilized}) = 0.75$
- $P(\text{vote}|\text{not mobilized}) = 0.15$
- $P(\text{mobilized}) = 0.6$ and so $P(\text{not mobilized}) = 0.4$

Note that mobilization **partitions** the data. Everyone is either mobilized or not. Thus, we can apply the LTP:

$$\begin{aligned}
P(\text{vote}) =& P(\text{vote}|\text{mobilized})P(\text{mobilized}) + \\
& P(\text{vote}|\text{not mobilized})P(\text{not mobilized}) \\
=& 0.75 \times 0.6 + 0.15 \times 0.4 \\
=& .51
\end{aligned}$$

Prosecutor's Fallacy

A woman has been murdered, and her husband is accused of having committed the murder. It is known that the man abused his wife repeatedly in the past, and the prosecution argues that this is important evidence pointing towards the man's guilt. The defense attorney says that the history of abuse is irrelevant, as only 1 in 1000 women who experience spousal abuse are subsequently murdered.

Assume that the defense attorney's 1 in 1000 figure is correct, and that half of men who murder their wives previously abused them. Also assume that 20% of murdered women were killed by their husbands, and that if a woman is murdered and the husband is not guilty, then there is only a 10% chance that the husband abused her. What is the probability that the man is guilty? Is the prosecution right that the abuse is important evidence in favor of guilt?

Prosecutor's Fallacy

A woman has been murdered, and her husband is accused of having committed the murder. It is known that the man abused his wife repeatedly in the past, and the prosecution argues that this is important evidence pointing towards the man's guilt. The defense attorney says that the history of abuse is irrelevant, as only 1 in 1000 women who experience spousal abuse are subsequently murdered.

Assume that the defense attorney's 1 in 1000 figure is correct, and that half of men who murder their wives previously abused them. Also assume that 20% of murdered women were killed by their husbands, and that if a woman is murdered and the husband is not guilty, then there is only a 10% chance that the husband abused her. What is the probability that the man is guilty? Is the prosecution right that the abuse is important evidence in favor of guilt?

Prosecutor's Fallacy

- Let's define our events

Prosecutor's Fallacy

- Let's define our events
  M => woman is murdered
  A => woman has previously experienced abuse
  G => woman's husband is guilty

Prosecutor's Fallacy

- Let's define our events
  M => woman is murdered
  A => woman has previously experienced abuse
  G => woman's husband is guilty
- What do we know?

Prosecutor's Fallacy

Table: What Do We Know?

Statements	We know	We also know
1 in 1000 women who experience spousal abuse are subsequently murdered	$P(?\|?) = 0.001$	$P(?\|?) = 1 - 0.001$
half of men who murder their wives previously abused them	$P(?\|?) = 0.5$	$P(?\|?) = 1 - 0.5$
20% of murdered women were killed by their husbands	$P(?\|?) = 0.2$	$P(?\|?) = 1 - 0.2$
woman is murdered and the husband is not guilty, then there is only a 10% chance that the husband abused her	$P(?\|?) = 0.1$	$P(?\|?) = 1 - 0.1$

Prosecutor's Fallacy

Table: What Do We Know?

Statements	We know	We also know
1 in 1000 women who experience spousal abuse are subsequently murdered	$P(M\|A) = 0.001$	$P(M^c\|A) = 1 - 0.001$
half of men who murder their wives previously abused them	$P(A\|M, G) = 0.5$	$P(A^c\|M, G) = 1 - 0.5$
20% of murdered women were killed by their husbands	$P(G\|M) = 0.2$	$P(G^c\|M) = 1 - 0.2$
woman is murdered and the husband is not guilty, then there is only a 10% chance that the husband abused her	$P(A\|G^c, M) = 0.1$	$P(A^c\|G^c, M)$=1-0.1

Prosecutor's Fallacy

- What we know $P(M|A)$, $P(A|M,G)$, $P(G|M)$, $P(A|G^c,M)$

Prosecutor's Fallacy

- What we know $P(M|A)$, $P(A|M, G)$, $P(G|M)$, $P(A|G^c, M)$
- What do we want to know?
  $P(G|M, A)$

Prosecutor's Fallacy

- What we know $P(M|A)$, $P(A|M,G)$, $P(G|M)$, $P(A|G^c,M)$
- What do we want to know?
  $P(G|M,A)$
- What can we use to get our quantity of interest?

Prosecutor's Fallacy

- What we know $P(M|A), P(A|M,G), P(G|M), P(A|G^c,M)$
- What do we want to know?
  $P(G|M,A)$
- What can we use to get our quantity of interest?
  Bayes' Rule

Bayes' Rule

- Often we have information about $P(B|A)$, but require $P(A|B)$ instead.
- When this happens, always think **Bayes' Rule**
- Bayes' rule: if $P(B) > 0$

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

Bayes' Rule

- Often we have information about $P(B|A)$, but require $P(A|B)$ instead.
- When this happens, always think **Bayes' Rule**
- Bayes' rule: if $P(B) > 0$

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

- Also recall from the definition of conditional probability:

$$\mathrm{P(A, B) = P(B \mid A)P(A)}$$

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$
$$P(A|G, M) = 1/2$$
$$P(G|M) = 1/5$$
$$P(A|G^c, M) = 1/10$$

$$P(G|M, A) = \frac{P(M, A|G)P(G)}{P(M, A)}$$

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$
$$P(A|G, M) = 1/2$$
$$P(G|M) = 1/5$$
$$P(A|G^c, M) = 1/10$$

$$\begin{aligned} P(G|M, A) &= \frac{P(M, A|G)P(G)}{P(M, A)} \\ &= \frac{P(M, A, G)}{P(M, A)} \end{aligned}$$

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$
$$P(A|G, M) = 1/2$$
$$P(G|M) = 1/5$$
$$P(A|G^c, M) = 1/10$$

$$P(G|M, A) = \frac{P(M, A|G)P(G)}{P(M, A)}$$
$$= \frac{P(M, A, G)}{P(M, A)}$$
$$= \frac{P(A|G, M)P(G|M)P(M)}{P(A|M)P(M)}$$

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$
$$P(A|G, M) = 1/2$$
$$P(G|M) = 1/5$$
$$P(A|G^c, M) = 1/10$$

$$\begin{aligned}
P(G|M, A) &= \frac{P(M, A|G)P(G)}{P(M, A)} \\
&= \frac{P(M, A, G)}{P(M, A)} \\
&= \frac{P(A|G, M)P(G|M)P(M)}{P(A|M)P(M)} \\
&= \frac{P(A|G, M)P(G|M)}{P(A|M)}
\end{aligned}$$

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$
$$P(A|G, M) = 1/2$$
$$P(G|M) = 1/5$$
$$P(A|G^c, M) = 1/10$$

How do we find P(A | M)?

Recall Law of Total Probability:

$$P(X) = P(X|Y)P(Y) + P(X|Y^c)P(Y^c)$$

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$
$$P(A|G, M) = 1/2$$
$$P(G|M) = 1/5$$
$$P(A|G^c, M) = 1/10$$

How do we find P(A | M)?

Recall Law of Total Probability:

$$P(X) = P(X|Y)P(Y) + P(X|Y^c)P(Y^c)$$

Applying here:

$$P(A|M) = P(A|M, G)P(G|M) + P(A|M, G^c)P(G^c|M)$$

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$
$$P(A|G, M) = 1/2$$
$$P(G|M) = 1/5$$
$$P(A|G^c, M) = 1/10$$

Putting it all together:

$$P(G|M, A) = \frac{P(A|G, M)P(G|M)}{P(A|M, G)P(G|M) + P(A|M, G^c)P(G^c|M)}$$

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$
$$P(A|G, M) = 1/2$$
$$P(G|M) = 1/5$$
$$P(A|G^c, M) = 1/10$$

Putting it all together:

$$P(G|M, A) = \frac{P(A|G, M)P(G|M)}{P(A|M, G)P(G|M) + P(A|M, G^c)P(G^c|M)}$$
$$= \frac{(.5)(.2)}{(.5)(.2) + (.1)(1 - 0.2)}$$

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$
$$P(A|G, M) = 1/2$$
$$P(G|M) = 1/5$$
$$P(A|G^c, M) = 1/10$$

Putting it all together:

$$\begin{aligned} P(G|M, A) &= \frac{P(A|G, M)P(G|M)}{P(A|M, G)P(G|M) + P(A|M, G^c)P(G^c|M)} \\ &= \frac{(.5)(.2)}{(.5)(.2) + (.1)(1 - 0.2)} \\ &= 0.556 \end{aligned}$$

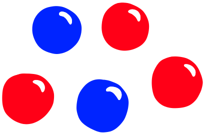
Prosecutor's Fallacy

- What does this mean for our defendant?

R Markdown

- install.packages("knitr")
- File - New File - R Markdown
- Preferences - Under Sweave set "Weave Rnw files with" to "knitr"
- See 1_Sample Markdown Document.Rmd

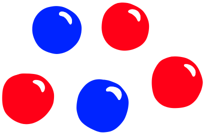
Probability by Simulation



Problem: You have a bag of five marbles. Three are red and two are blue. You draw one marble. Without replacing it, you then draw another marble.
What is the probability that the two marbles are the same color?

- We could do this analytically:

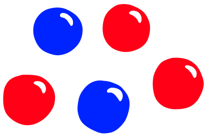
Probability by Simulation



Problem: You have a bag of five marbles. Three are red and two are blue. You draw one marble. Without replacing it, you then draw another marble.
What is the probability that the two marbles are the same color?

- We could do this analytically:
  P(Same color)
  $=P(D1 = R)P(D2 = R \mid D1 = R) + P(D1 = B)P(D2 = B \mid D1 = B)$
  $= (3/5)(2/4) + (2/5)(1/4)$
  $= 2/5$

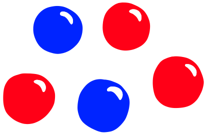
Probability by Simulation



Problem: You have a bag of five marbles. Three are red and two are blue. You draw one marble. Without replacing it, you then draw another marble.
What is the probability that the two marbles are the same color?

- We could do this analytically:
  P(Same color)
  $= P(D1 = R)P(D2 = R \mid D1 = R) + P(D1 = B)P(D2 = B \mid D1 = B)$
  $= (3/5)(2/4) + (2/5)(1/4)$
  $= 2/5$
- Or we can run a simulation!
  See 2_Simulation example.R

Writing Functions



- We've already used many built in R functions: `mean()`, `head()`, etc.
- We can also define our own functions:

Define a function that takes 3 arguments; it will add the first two and divide by the third:

```
> myfunction <- function(x, y, z){
+   out <- (x + y) / z
+   return(out)
+ }
> ## use the function
> myfunction(1, 5, 2)

[1] 3
```

Data Manipulation and Tables

See 3_Data Manipulations and Tables.Rmd

Fun with Lint!!

adv-r.had.co.nz/Style.html

RStudio → Preferences →
Code → Diagnostics →
"Provide R style diagnostics"

Resources

- R Style Guide: `http://adv-r.had.co.nz/Style.html`
- R Cookbook: `http://www.cookbook-r.com/`
- ggplot cheatsheet: `https://www.rstudio.com/wp-content/uploads/2015/08/ggplot2-cheatsheet.pdf`
- dplyr cheatsheet: `https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf`
- Probability cheatsheet: `https://static1.squarespace.com/static/54bf3241e4b0f0d81bf7ff36/t/55e9494fe4b011aed10e48e5/1441352015658/probability_cheatsheet.pdf`
- Probability and statistics visualizations: `http://students.brown.edu/seeing-theory/index.html`
- Kosuke Imai's textbook contains lots of sample R code!