Precept 1: Probability, Simulations, Working with Data

Soc 500: Applied Social Statistics

Simone Zhang

Princeton University

September 2016

Support Resources

- Office hours
- Math camp materials
- Piazza
- Email (please CC both of us)
- Google is your best friend!

Learning objectives

- Create an R Markdown document

Learning objectives

- Create an R Markdown document
- Translate information provided in word problems into probability statements

Learning objectives

- Create an R Markdown document
- Translate information provided in word problems into probability statements
- Run simulations (loops, functions, replicate)

Learning objectives

- Create an R Markdown document
- Translate information provided in word problems into probability statements
- Run simulations (loops, functions, replicate)
- Perform basic data manipulations, create summary tables, and graphs

Learning objectives

- Create an R Markdown document
- Translate information provided in word problems into probability statements
- Run simulations (loops, functions, replicate)
- Perform basic data manipulations, create summary tables, and graphs

Learning objectives

- Create an R Markdown document
- Translate information provided in word problems into probability statements
- Run simulations (loops, functions, replicate)
- Perform basic data manipulations, create summary tables, and graphs

Acknowledgements: These slides draw on materials developed by past preceptors, Elisha Cohen and Clark Bernier. Thanks!

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 from a 2 X 2 Table

- Imagine that someone on GradCafe posts that they have just been admitted to all ten of the top 10 sociology programs. Is this claim plausible?
- Consider the following table of grad school applicants:

	Princeton		
Stanford?	Yes	No	Total
Yes	15	10	25
No	15	760	760
Total	30	770	800

Probability from a 2 X 2 Table

- What is the sample space here?

	Princeton		
Stanford?	Yes	No	Total
Yes	15	10	25
No	15	760	760
Total	30	770	800

Probability from a 2 X 2 Table

- What is the sample space here?
  Admissions outcomes for people who applied to Princeton and Stanford

	Princeton		
Stanford?	Yes	No	Total
Yes	15	10	25
No	15	760	760
Total	30	770	800

Probability from a 2 X 2 Table

- What is the sample space here?
  Admissions outcomes for people who applied to Princeton and Stanford
- What is the probability that a randomly selected student got into both Stanford and Princeton?

	Princeton		
Stanford?	Yes	No	Total
Yes	15	10	25
No	15	760	760
Total	30	770	800

Probability from a 2 X 2 Table

- What is the sample space here?
  Admissions outcomes for people who applied to Princeton and Stanford
- What is the probability that a randomly selected student got into both Stanford and Princeton?
  $\Pr(P = Y, S = Y) = 15/800$

	Princeton		
Stanford?	Yes	No	Total
Yes	15	10	25
No	15	760	760
Total	30	770	800

Probability from a 2 X 2 Table

- What is the sample space here?
  Admissions outcomes for people who applied to Princeton and Stanford
- What is the probability that a randomly selected student got into both Stanford and Princeton?
  $\Pr(P = Y, S = Y) = 15/800$
- Given that a student got into Princeton, what is the probability that they got into Stanford?

	Princeton		
Stanford?	Yes	No	Total
Yes	15	10	25
No	15	760	760
Total	30	770	800

Probability from a 2 X 2 Table

- What is the sample space here?
  Admissions outcomes for people who applied to Princeton and Stanford
- What is the probability that a randomly selected student got into both Stanford and Princeton?
  $\Pr(P = Y, S = Y) = 15/800$
- Given that a student got into Princeton, what is the probability that they got into Stanford?
  $\Pr(S = Y \mid P = Y) = 15/30$

	Princeton		
Stanford?	Yes	No	Total
Yes	15	10	25
No	15	760	760
Total	30	770	800

Probability from a 2 X 2 Table

- Is getting into Stanford independent of getting into Princeton?

	Princeton		
Stanford?	Yes	No	Total
Yes	15	10	25
No	15	760	760
Total	30	770	800

Probability from a 2 X 2 Table

- Is getting into Stanford independent of getting into Princeton?
- Recall that events A and B are independent if knowing that A occurred provides no information about whether B occured
  $Pr(A,B) = Pr(A)Pr(B) \implies A \perp B$
  $Pr(A|B) = Pr(A)$ and $Pr(B|A) = P(B)$

	Princeton		
Stanford?	Yes	No	Total
Yes	15	10	25
No	15	760	760
Total	30	770	800

Probability from a 2 X 2 Table

- Is getting into Stanford independent of getting into Princeton?
- Recall that events A and B are independent if knowing that A occurred provides no information about whether B occured
  $Pr(A,B) = Pr(A)Pr(B) \implies A \perp B$
  $Pr(A|B) = Pr(A)$ and $Pr(B|A) = P(B)$
- Applying that here:
  $Pr(P{=}Y, S{=}Y) = 15/800 = 0.01875$
  $Pr(P{=}Y)Pr(S{=}Y) = (30/800)(25/800) = 0.00117$
  Getting into Princeton and getting into Stanford are not independent

	Princeton		
Stanford?	Yes	No	Total
Yes	15	10	25
No	15	760	760
Total	30	770	800

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

- 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?
  $P(M|A), P(A|M, G), P(G|M), P(A|G', M)$

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

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

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

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?
  $P(M|A), P(A|M, G), P(G|M), P(A|G', M)$
- 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 $Pr(B|A)$, but require $Pr(A|B)$ instead.
- When this happens, always think **Bayes' Rule**
- Bayes' rule: if $Pr(B) > 0$

$$\Pr(\text{A} \mid \text{B}) = \frac{Pr(B|A)Pr(A)}{Pr(B)}$$

Bayes' Rule

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

$$\Pr(\text{A} \mid \text{B}) = \frac{Pr(B|A)Pr(A)}{Pr(B)}$$

- Also recall from the definition of conditional probability:

$$\Pr(\text{A, B}) = \Pr(\text{B} \mid \text{A})\Pr(\text{A})$$

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$

$$P(A|G, M) = 1/2$$

$$P(G|M) = 1/5$$

$$P(A|G', M) = 1/10$$

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

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$

$$P(A|G, M) = 1/2$$

$$P(G|M) = 1/5$$

$$P(A|G', M) = 1/10$$

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

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

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$

$$P(A|G, M) = 1/2$$

$$P(G|M) = 1/5$$

$$P(A|G', M) = 1/10$$

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

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

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

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$

$$P(A|G, M) = 1/2$$

$$P(G|M) = 1/5$$

$$P(A|G', M) = 1/10$$

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

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

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

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

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$

$$P(A|G, M) = 1/2$$

$$P(G|M) = 1/5$$

$$P(A|G', M) = 1/10$$

How do we find P(A | M)?

Recall Law of Total Probability:

$$P(X) = P(X \mid Y)P(Y) + P(X \mid Y')P(Y')$$

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$

$$P(A|G, M) = 1/2$$

$$P(G|M) = 1/5$$

$$P(A|G', M) = 1/10$$

How do we find P(A | M)?

Recall Law of Total Probability:

$$P(X) = P(X \mid Y)P(Y) + P(X \mid Y')P(Y')$$

Applying here:

$$P(A \mid M) = P(A|G,M)P(G|M) + P(A|G', M)P(G'|M)$$

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$

$$P(A|G, M) = 1/2$$

$$P(G|M) = 1/5$$

$$P(A|G', M) = 1/10$$

Putting it all together:

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

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$

$$P(A|G, M) = 1/2$$

$$P(G|M) = 1/5$$

$$P(A|G', M) = 1/10$$

Putting it all together:

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

$$= \frac{(.5)(.2)}{(.5)(.2) + (.1)(1 - 0.2)}$$

Prosecutor's Fallacy

$$P(M|A) = 1/1000$$

$$P(A|G, M) = 1/2$$

$$P(G|M) = 1/5$$

$$P(A|G', M) = 1/10$$

Putting it all together:

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

$$= \frac{(.5)(.2)}{(.5)(.2) + (.1)(1 - 0.2)}$$

$$= 0.556$$

Prosecutor's Fallacy

- What does this mean for our defendant?

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 colour?

- 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 colour?

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

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 colour?

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

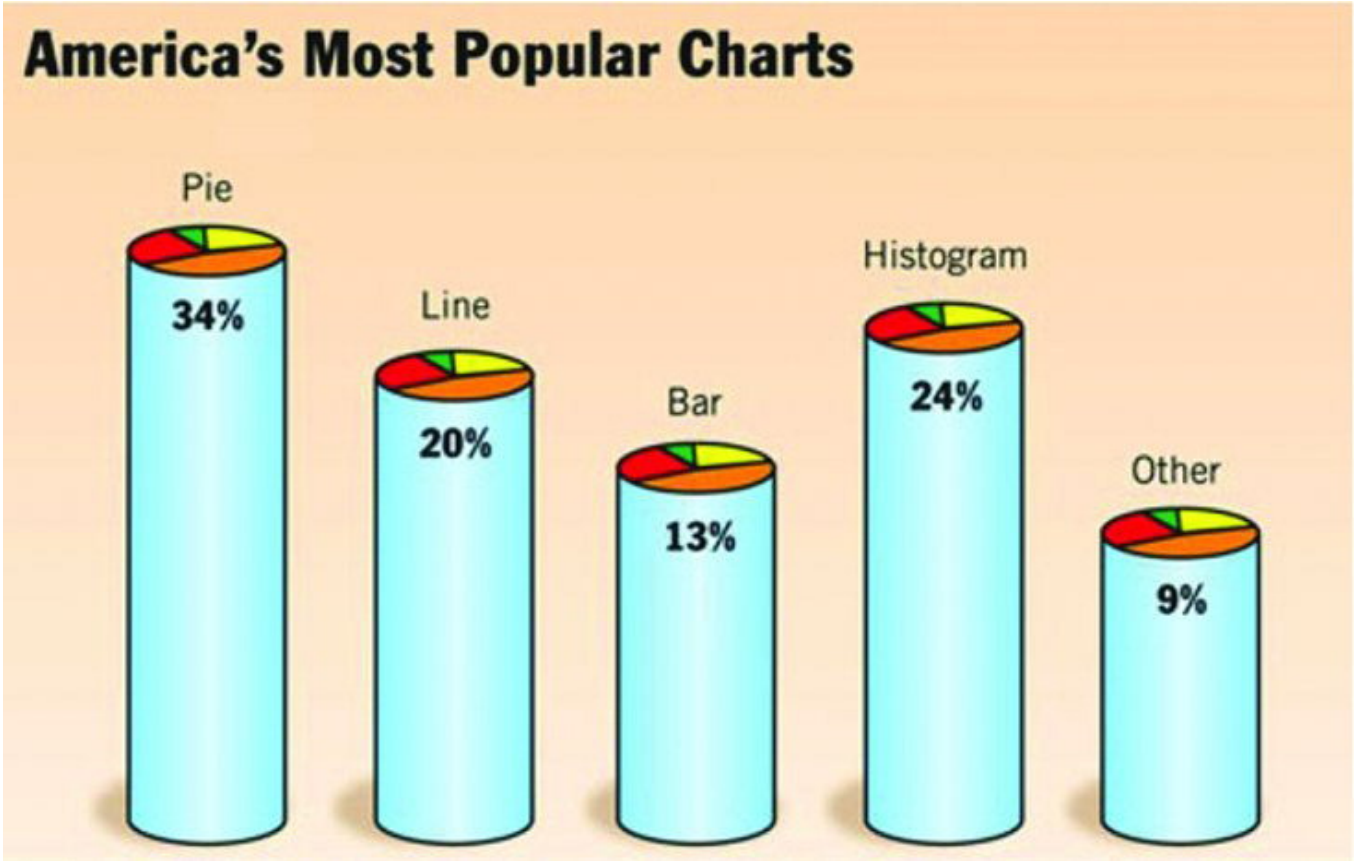
```
> my.function <- function(x,y,z){
+   out <- (x + y)/z
+   return(out)
+ }
> ## use the function
> my.function(1, 5, 2)
```

```
[1] 3
```

Data Manipulation and Tables

See 3_Data Manipulations and Tables.Rmd

Graphics[1]



[1] http://www.theonion.com/graphic/americas-most-popular-charts-7492

Graphics[2]

Goals of Data Visualization (i.e. why use graphics?)

- Discovery (exploratory)

[2] Gelman, Andrew, and Antony Unwin. "Infovis and statistical graphics: different goals, different looks." Journal of Computational and Graphical Statistics 22.1 (2013): 2-28.

Graphics[2]

Goals of Data Visualization (i.e. why use graphics?)

- Discovery (exploratory)
  - qualitative overview, looking for patterns, outliers, scale of data

[2] Gelman, Andrew, and Antony Unwin. "Infovis and statistical graphics: different goals, different looks." Journal of Computational and Graphical Statistics 22.1 (2013): 2-28.

Graphics[2]

Goals of Data Visualization (i.e. why use graphics?)

- Discovery (exploratory)
  - qualitative overview, looking for patterns, outliers, scale of data
- Communication (presentation)

[2] Gelman, Andrew, and Antony Unwin. "Infovis and statistical graphics: different goals, different looks." Journal of Computational and Graphical Statistics 22.1 (2013): 2-28.

Graphics[2]

Goals of Data Visualization (i.e. why use graphics?)

- Discovery (exploratory)
  - qualitative overview, looking for patterns, outliers, scale of data
- Communication (presentation)
  - displaying information from the data in an accessible way

[2] Gelman, Andrew, and Antony Unwin. "Infovis and statistical graphics: different goals, different looks." Journal of Computational and Graphical Statistics 22.1 (2013): 2-28.

Graphics[2]

Goals of Data Visualization (i.e. why use graphics?)

- Discovery (exploratory)
  - qualitative overview, looking for patterns, outliers, scale of data
- Communication (presentation)
  - displaying information from the data in an accessible way
  - telling a story, reporting results

[2] Gelman, Andrew, and Antony Unwin. "Infovis and statistical graphics: different goals, different looks." Journal of Computational and Graphical Statistics 22.1 (2013): 2-28.

Graphics[2]

Goals of Data Visualization (i.e. why use graphics?)

- Discovery (exploratory)
  - qualitative overview, looking for patterns, outliers, scale of data
- Communication (presentation)
  - displaying information from the data in an accessible way
  - telling a story, reporting results
  - grab your audience and keep them interested

[2] Gelman, Andrew, and Antony Unwin. "Infovis and statistical graphics: different goals, different looks." Journal of Computational and Graphical Statistics 22.1 (2013): 2-28.

Graphics using `ggplot2()`[3]

`ggplot2()` conceptually:

- each graphic is made up of different layers of components

[3] Wickham, Hadley. ggplot2: elegant graphics for data analysis. 2009

Graphics using `ggplot2()`[3]

`ggplot2()` conceptually:

- each graphic is made up of different layers of components
  - start with layer plotting raw data

[3] Wickham, Hadley. ggplot2: elegant graphics for data analysis. 2009

Graphics using `ggplot2()`[3]

`ggplot2()` conceptually:

- each graphic is made up of different layers of components
  - start with layer plotting raw data
  - add annotations

[3] Wickham, Hadley. ggplot2: elegant graphics for data analysis. 2009

Graphics using `ggplot2()`[3]

`ggplot2()` conceptually:

- each graphic is made up of different layers of components
  - start with layer plotting raw data
  - add annotations
  - add statistical summaries

[3] Wickham, Hadley. ggplot2: elegant graphics for data analysis. 2009

Graphics using `ggplot2()`[3]

`ggplot2()` conceptually:

- each graphic is made up of different layers of components
  - start with layer plotting raw data
  - add annotations
  - add statistical summaries
- highly customizable

[3] Wickham, Hadley. ggplot2: elegant graphics for data analysis. 2009

Graphics using ggplot2

grammar of `ggplot2()` is composed of:

- **data** that you want to visualize
  - set of aesthetic **mappings**
- **geoms**: geometric shapes – points, lines, polygons, etc.
- **stats**: statistical transformations e.g. binning and counting for histogram
- **scales**: map data values to aesthetical values – color, shape, size, and legend
- **coord**: coordinate system – how data is mapped to coordinate; provides axes and gridlines
- **facet**: how to break up the data into subsets

diamonds data

- easy way to start plotting is to use `qplot()`, short for **q**uick plot

Show distribution of 1 variable:

```
> qplot(carat, data = diamonds, geom = "histogram")
> qplot(carat, data = diamonds, geom = "density")
```

qplots

Change `binwidth` argument:

```
> qplot(carat, data = diamonds, geom = "histogram",
+       binwidth = 1, xlim = c(0,3))
> qplot(carat, data = diamonds, geom = "histogram",
+       binwidth = 0.1, xlim = c(0,3))
```

qplots

To compare different subgroups (diamonds of different color groups) use an aesthetic mapping:

```
> qplot(carat, data = diamonds, geom = "histogram",
+       fill = color)
```

ggplot()

Reproduce the same histogram using full ggplot()

```
> p <- ggplot(diamonds, aes(x = carat))
> p + geom_histogram()
```

ggplot()

Change binwidth:

```
> p <- ggplot(diamonds, aes(x = carat))
> p + geom_histogram(binwidth = 0.1)
```

ggplot()

Group counts by diamond color:

```
> p <- ggplot(diamonds, aes(x = carat))
> p + geom_histogram(aes(fill = color))
```

Plots with more options

More complicated qplot:

```
> qplot(carat, data = diamonds,
+       geom = "histogram",
+       binwidth = 0.1,
+       main = "Histogram for Carat",
+       xlab = "Carat",
+       fill=I("green"),
+       col=I("red"),
+       alpha=I(.2), # transparency
+       xlim=c(0,4))
```

Plots with more options

Same plot, but using `ggplot()` specification:

```
> p1 <- ggplot(data = diamonds, aes(x = carat))
> p1 + geom_histogram(binwidth = 0.1,
+                     col = "red",
+                     fill = "green",
+                     alpha = .2) +
+   labs(title = "Histogram for Carat") +
+   labs(x = "Carat", y = "Count") +
+   xlim(c(0,4))
```

Resources

- 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`
- Kosuke Imai's textbook contains lots of sample R code!