3251 Statistics for Data Science

Data Science Fundamentals Certificate

Module 2
PROBABILITY

Course Roadmap

Module / Week	Title
1	Introduction to Statistics for Data Science
2	Probability
3	Distribution of Random Variables
4	Inference
5	Model Building
6	Linear Regression
7	Multiple Linear Regression
8	Logistic Regression
9	Introduction to Bayesian Inference
10	Multi-level Models
11	Markov Chain Monte Carlo
12	Presentations
13	Final Exam

Module 2: Learning Objectives

- Define Probability
- Understand random variables
- Work with conditional probabilities
- Describe the difference between Frequentist and Bayesian definitions of probability

INTRODUCTION TO PROBABILITY

History of Probability

Image: pixabay.com

The Introduction of Probability Theory

Gerolamo Cardano, School of Mathematics and Statistics, University of St Andrews

1564

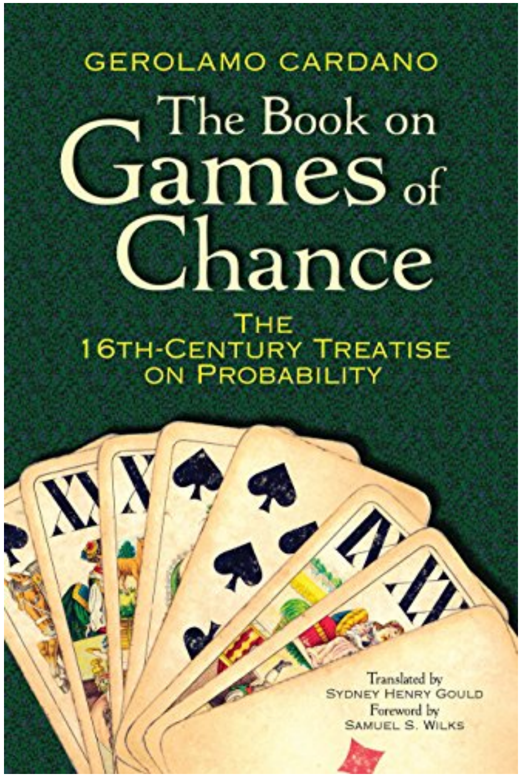


Image: amazon.com

Blaise Pascal and Pierre de Fermat

https://en.wikipedia.org/wiki/
File:Blaise_pascal.jpg

Blaise Pascal

1654

Image: pexels.com

http://www-groups.dcs.st-
and.ac.uk/~history/PictDisplay/Fermat.html

Pierre de Fermat

Probability Discussion

Image: pixabay.com

What is the chance of getting heads? … or tails?

What does it mean to say $P = 0.5$?

Probability and Decision-Making Process

More Probability Examples

What is the chance of getting 1 when rolling a die?

$P(1) = 1/6$

What is the chance of getting 1 or 6?

Image: pexels.com

$P(1 \text{ or } 6\ ) = 1/3$

Random Variables Definition

- A variable or process with a random outcome is called a **random variable**
- Usually a random variable is represented with a capital letter such as X, Y or Z

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

MECE

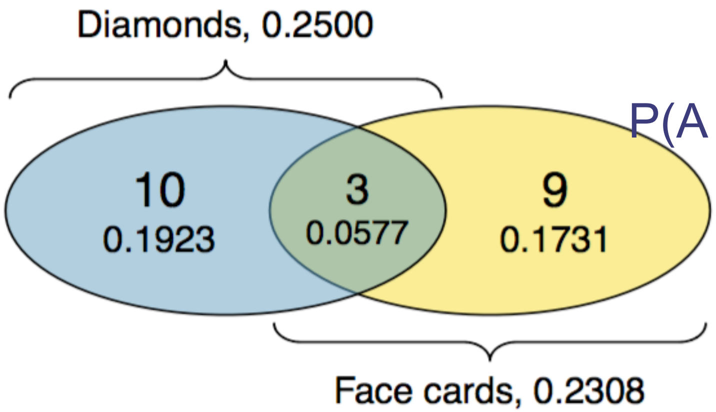
- Mutually exclusive
- Collectively exhaustive

Source: “Disorder in the Court” – Public Domain movie

When Events Aren't Disjoint

- Consider an example: regular deck of 52 cards.
- What is the probability that a randomly selected card is a diamond?
- What is the probability that a randomly selected card is a face card?

Venn Diagram and General Addition Rule



P(A or B) = P (♦ or face card)

= P (♦) + P (face card) − P (♦ and face card)

= 13/52 + 12/52 − 3/52

= 22/52 = 11/26

General Addition Rule

If A and B are any two events, disjoint or not, then the probability that at least one of them will occur is

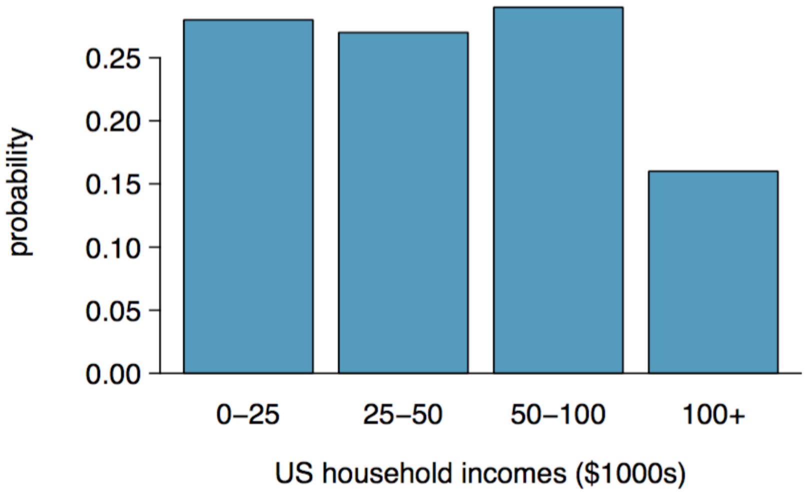
P (A or B) = P (A) + P (B) − P (A and B),

where P(A and B) is the probability that both events occur.

Plotting Probability Distributions

- Example: Table below suggests three distributions for household income in the United States. Only one is correct. Which one must it be?

Income range ($1000s)	0-25	25-50	50-100	100+
(a)	0.18	0.39	0.33	0.16
(b)	0.38	-0.27	0.52	0.37
(c)	0.28	0.27	0.29	0.16



- Probabilities of (a) do not sum up to 1
- (b) contains negative probability
- (c) is correct

M&M's Example

Image: pixabay.com

Monty Hall problem

Both Images: pixabay.com

Definitions of Probability

FREQUENTIST AND BAYESIAN DEFINITIONS OF PROBABILITY

Defining Probability

- The likelihood that an event will occur
- Objectively: A measure of some objective physical phenomenon such as the proportion of times the outcome would occur if we observed the random process an infinite number of times
- Subjectively: A degree of belief in a predicate (i.e. a statement which can be true or false)

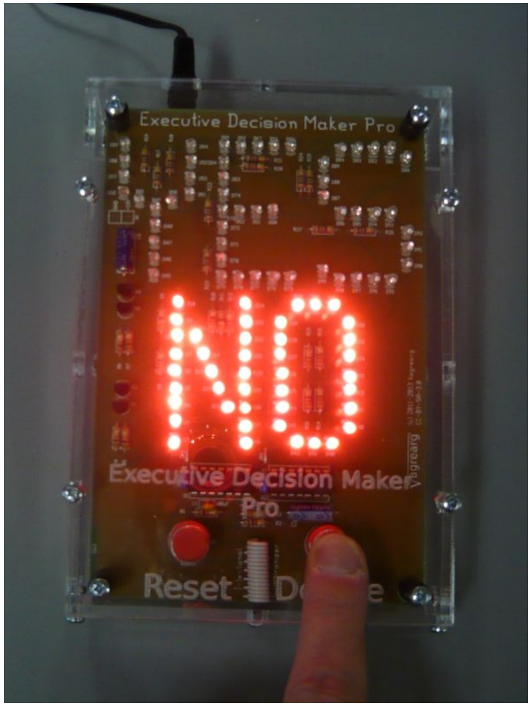
The Frequentist Definition

- Frequentist definition of probability of an event is its relative frequency in a large number of trials
- For example, if you can repeat flipping a coin indefinitely and count how many heads you get and divide that number by the number of flips, the value you obtain should be 0.5

$$p = \lim_{n \to \infty} \frac{k}{n}$$

The probability of an event is defined as the proportion of times the event K occurs and number of trials n when n goes to infinity

An Executive Decision-Maker



Source: http://www.vagrearg.org/content/edm-pro

Let's Try it 100 Times

- Let's say we pushed the button 100X and observed it showed "Yes" 55 times
- What is p?
- The frequentist answer: pick the value of p that makes the observation of 55 Yes's the most probable result of our test i.e. p = 0.55
- This is called the Maximum Likelihood Estimate

The Bayesian Framework

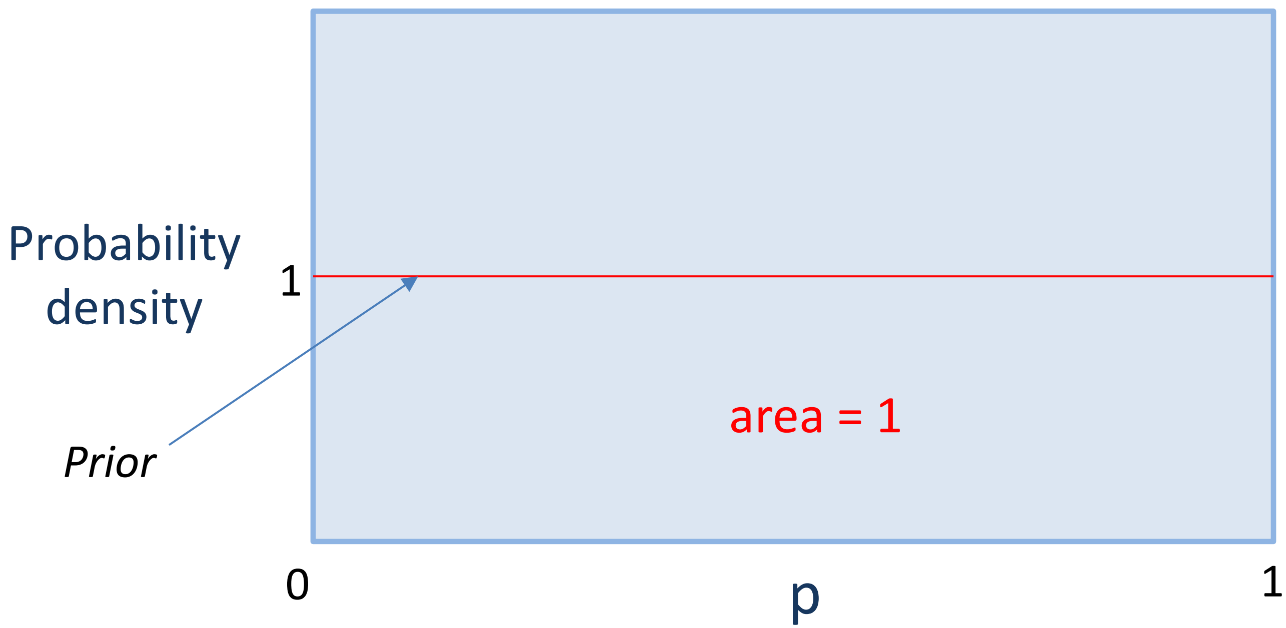
- The Bayesian approach assumes that we always have a prior distribution even though the prior may be very vague, equiprobable, or even outright wrong
- When we obtain new data, we update the prior distribution in light of the new data to get an updated probability distribution called the posterior distribution
- The posterior distribution reflects our state of knowledge after collecting data

What if We Can Only Do an Experiment Once?

- What if we were only allowed to push the Yes/No button once and it came up "Yes"?
  - The Max Likelihood Estimate would be p = 1
  - Would you use this device to play Russian Roulette?
  - Does p = 1 (i.e. 100% certain it will always come up Yes) make sense?
  - Is it reasonable to give a single answer?

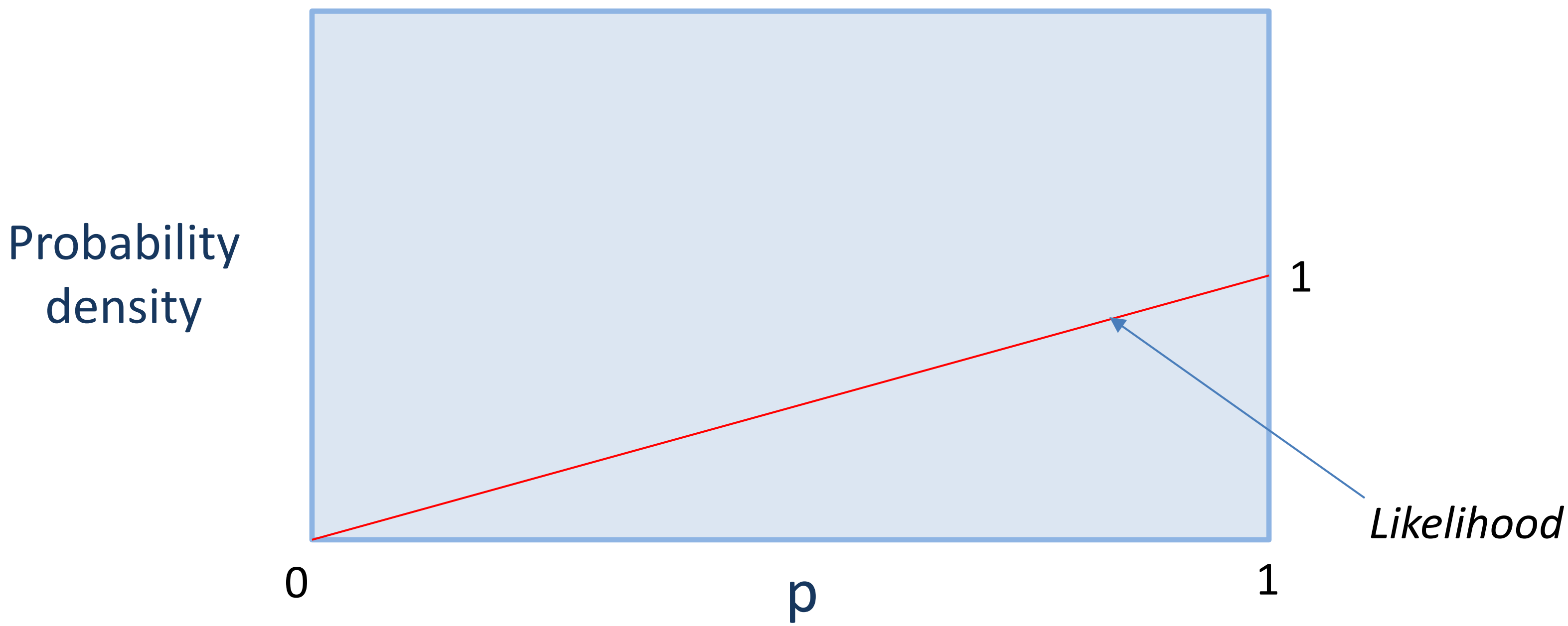
How should our beliefs evolve as we gather new information?

- Let's say we start out with no opinion about what p should be i.e. a uniform distribution



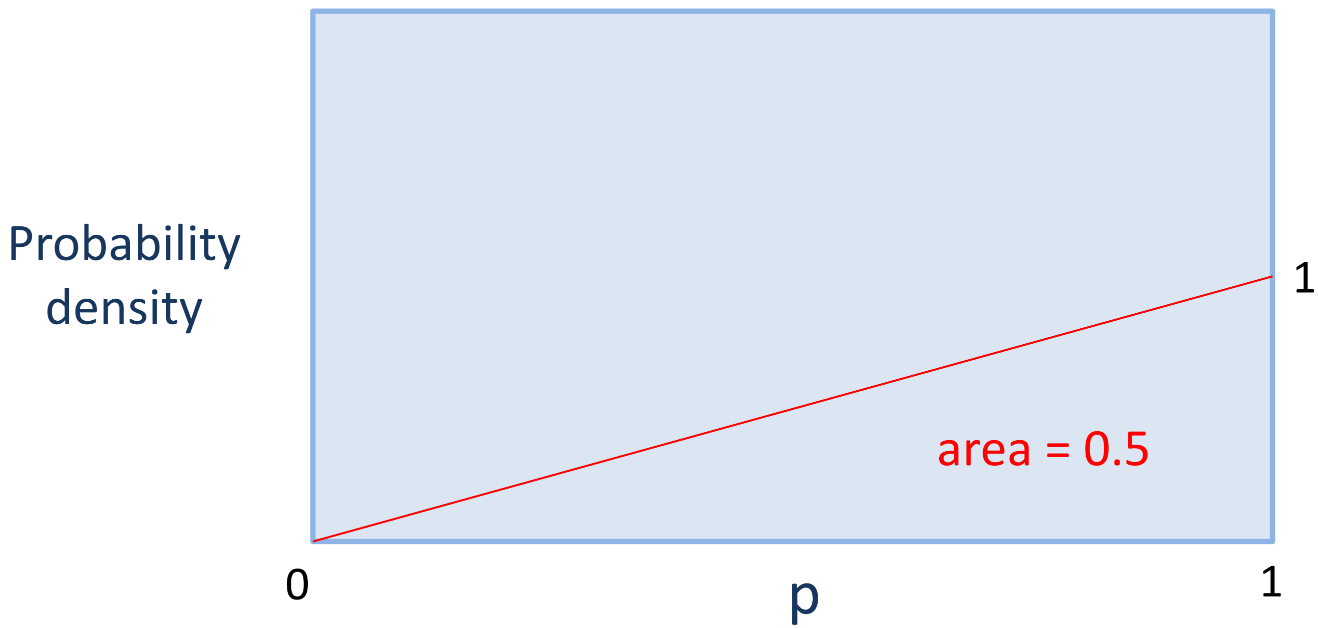
Say we push the button and get Yes

- We now know p = 1 is the most likely and p cannot be 0, but it could still be something between and the farther from 1 the less likely

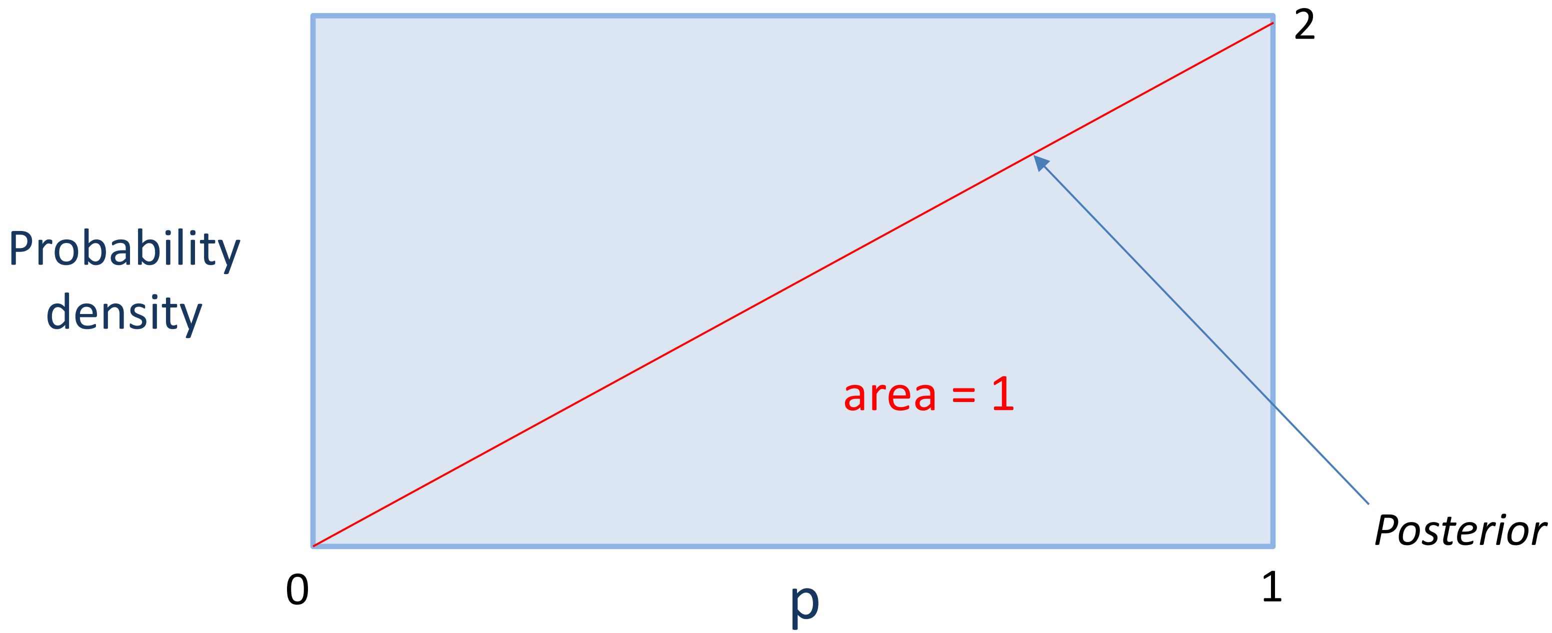


The rule is: we multiply those two distributions together

- 1 times the Likelihood is this line, but the area under the curve isn't 1 like we need

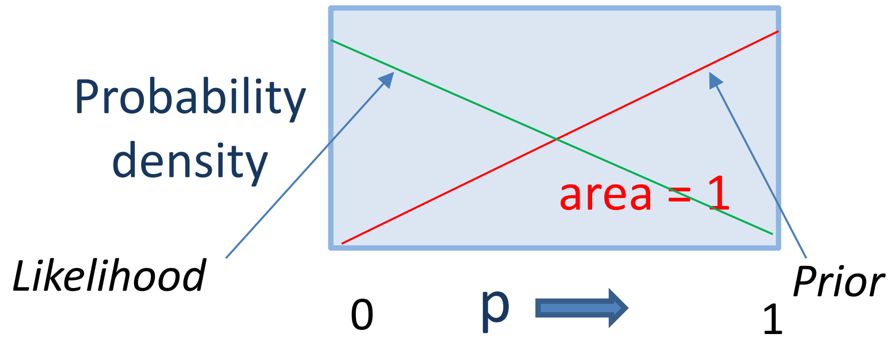


Then adjust so the area under the curve is 1 again

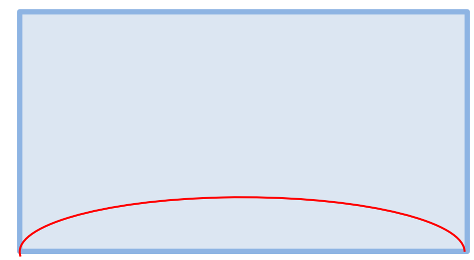


Suppose now we get a “No”

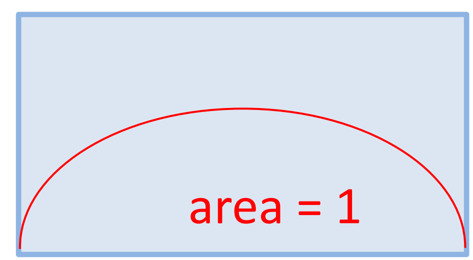
- Our posterior is now our new prior and our new evidence is the green line



- Multiply

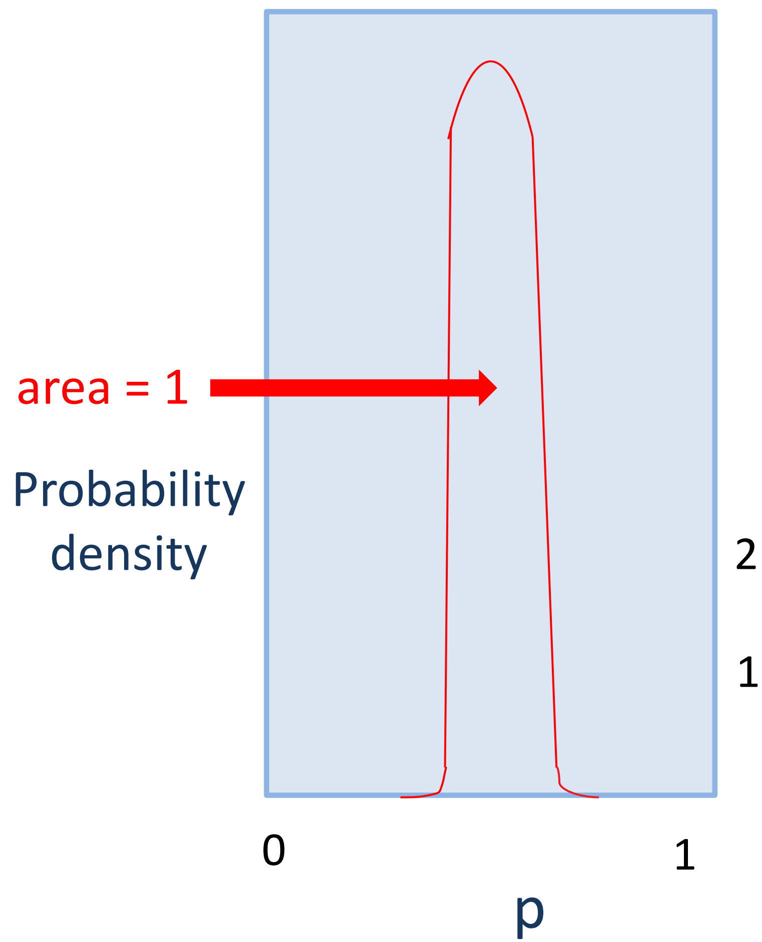


- Then adjust the area again



Repeat…

- After 55 Yes's and 45 No's we would get a posterior distribution that has its peak at 0.55 and a clear sense of how much we should trust that estimate
- Note that the initial uniform assumption has been essentially "washed away" by the data



CONDITIONAL PROBABILITY

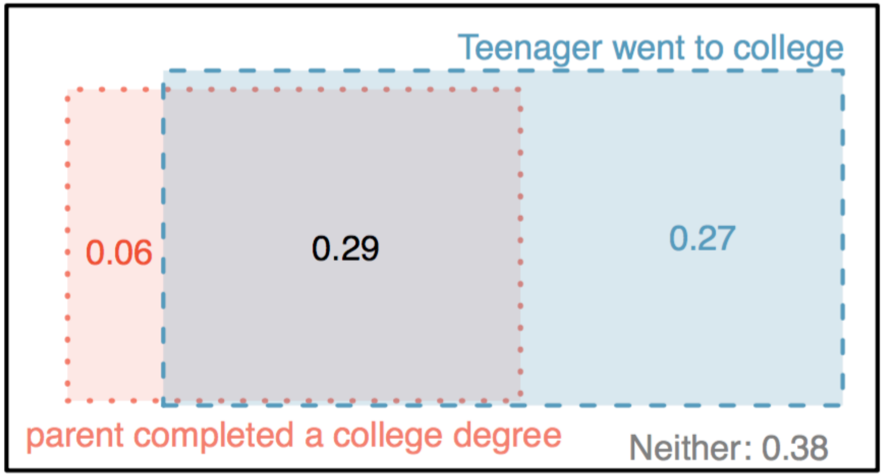
Conditional Probabilities

- You have these socks in a drawer:
  - 2 blue socks
  - 4 red socks
  - 3 yellow socks
  - 6 grey socks
- What is the probability of pulling out two blue socks (with replacement)
- Without replacement?
- Are the draws independent?

Conditional Probability Example

The *family_college* data set contains a sample of 792 cases with two variables, *teen* and *parents*, and is summarized in the table below. The *teen* variable is either *college* or *not*, where the *college* label means the teen went to college immediately after high school. The *parents* variable takes the value *degree* if at least one parent of the teenager completed a college degree.

		parents		
		degree	not	Total
teen	college	231	214	445
	not	49	298	347
	Total	280	512	792



Marginal and Joint Probabilities

- If a probability is based on a single variable, it is a **marginal** probability:

  a probability based solely on the teen variable is a marginal probability:

  P (teen college) = 445 / 792 = 0.56

- The probability of outcomes for two or more variables or processes is called a **joint** probability:

  P (teen college and parents not) = 214 / 792 = 0.27

Defining Conditional Probability

- There is a connection between education level of parents and of the teenager: a college degree by a parent is associated with college attendance of the teenager
- The probability that a random teenager from the study attended college and that one of the teen's parents has a college degree - only 280 cases:

  P (teen college given parents degree) = 231 / 280

  = 0.825
- This is a **conditional** probability because we computed the probability under a condition: a parent has a college degree

Conditional Probability Definition

- The conditional probability of the outcome of interest A given condition B is computed as the following:

$$P(A|B) = P(A \text{ and } B) / P(B)$$

Independence

- Two events are called **independent** when:

$$P\ (A|B) = P\ (A)$$

- Then:

$$P(A \text{ and } B) = P(A)P(B)$$

Bayes' Theorem

Bayes' Theorem: inverting probabilities

- Consider the following conditional probability for variable 1 and variable 2:

  P (outcome A_1 of variable 1 | outcome B of variable 2)

- Bayes' Theorem states that this conditional probability can be identified as the following fraction:

  $P(B|A_1)P(A_1) / [P(B|A_1)P(A_1)+P(B|A_2)P(A_2)+\cdots+P(B|A_k)P(A_k)]$

  where A_3, A_3, ..., and A_k represent all other possible outcomes of the first variable (B)

Resources

- OpenIntro Statistics Chap. 2
- Westfall & Henning, Understanding Advanced Statistical Methods, CRC Press, 2013. Chaps. 1-6
- E. T. Jaynes, Probability Theory: The Logic of Science, Cambridge Press, 2003
- Murray Spiegel, Probability and Statistics, Shaum’s Outlines

Resources (Cont'd)

- Sheldon Ross, Probability and Statistics for Engineers and Scientists, Fifth Edition, Academic Press, 2014. Chaps. 1-4
- Frederik Palk Schoenberg, Introduction to Probability with Texas Hold'em Examples, 2^{nd} Edition, CRC Press, 2016.

Resources (Cont'd)

- Article "The Odds, Continually Updated", applications of Bayesian statistics - https://www.nytimes.com/2014/09/30/science/the-odds-continually-updated.html?_r=1
- "Frequentist vs Bayesian statistics - a non-statisticians view" - http://www.met.reading.ac.uk/~sws97mha/Publications/Bayesvsfreq.pdf

Next Class

- Commonly used distributions
- In preparation:
  - Read OpenIntro Statistics Chapter 3: Distributions of Random Variables