Homework_01

August 29, 2022

Probability for Data Science

UC Berkeley, Fall 2022

Ani Adhikari

[1]:
```
from prob140 import *
from datascience import *
import numpy as np
from scipy import special

import matplotlib.pyplot as plt
%matplotlib inline
import matplotlib
matplotlib.style.use('fivethirtyeight')
```

1 Homework 1

1.0.1 Instructions

Your homeworks have two components: a written portion and a portion that also involves code. Written work should be completed on paper, and coding questions should be done in the notebook. You are welcome to LaTeX your answers to the written portions, but staff will not be able to assist you with LaTeX related issues. It is your responsibility to ensure that both components of the homework are submitted completely and properly to Gradescope. Refer to the bottom of the notebook for submission instructions.

1.0.2 How to Do Your Homework

The point of homework is for you to try your hand at using what you've learned in class. The steps to follow:

- Go to lecture and sections, and also go over the relevant text sections before starting on the homework. This will remind you what was covered in class, and the text will typically contain examples not covered in lecture. The weekly Study Guide will list what you should read.
- Work on some of the practice problems before starting on the homework.

- Attempt the homework problems by yourself with the text, section work, and practice materials all at hand. Sometimes the week's lab will help as well. The two steps above will help this step go faster and be more fruitful.
- At this point, seek help if you need it. Don't ask how to do the problem — ask how to get started, or for a nudge to get you past where you are stuck. Always say what you have already tried. That helps us help you more effectively.
- For a good measure of your understanding, keep track of the fraction of the homework you can do by yourself or with minimal help. It's a better measure than your homework score, and only you can measure it.

1.0.3 Rules for Homework

- Every answer should contain a calculation or reasoning. For example, a calculation such as $(1/3)(0.8) + (2/3)(0.7)$ or `sum([(1/3)*0.8, (2/3)*0.7])`is fine without further explanation or simplification. If we want you to simplify, we'll ask you to. But just $\binom{5}{2}$ by itself is not fine; write "we want any 2 out of the 5 frogs and they can appear in any order" or whatever reasoning you used. Reasoning can be brief and abbreviated, e.g. "product rule" or "not mutually exclusive."
- You may consult others (see "How to Do Your Homework" above) but you must write up your own answers using your own words, notation, and sequence of steps.
- We'll be using Gradescope. You must submit the homework according to the instructions at the end of homework set.

1.1 1. Playing to Win

This exercise is a workout in the following problem-solving skills.

To find the exact chance of an event that involves multiple trials:

- Start by asking, "What does the first trial have to be?" and then "What does the second trial have to be?". If the answers are clear, such as "win, then lose," then the multiplication rule might do the job directly.
- But if the answers aren't clear, such as, "Well, the first trial could be a win, or not, but then the second trial should be a win, or not, but then …" then try partitioning the event into simpler pieces (also known as *listing the ways*), or look at the complement. Maybe one of these methods will help you find a solution. Almost always, one of these two is simpler than the other, but which one is simpler depends on the problem.

To find an exponential approximation, deeply internalize the subsection headings 1.5.1 through 1.5.4 of Section 1.5, and don't forget that x^m is a product when m is a positive integer.

Try out these moves in the following setting.

A gambler plays two games of chance. Every time she plays Game A, she has chance $\frac{1}{6n}$ of winning a laptop, regardless of the outcomes of all other games. Every time she plays Game B, she has chance $\frac{1}{3n}$ of winning a smartphone, regardless of the results of all other games.

She has decided on the following strategy.

- She keeps playing Game A until either she wins a laptop or she has played Game A n times and has lost every time.
- If she wins a laptop she stops playing.

- If she loses all n times on Game A, she starts playing Game B. She keeps playing until either she wins a smartphone or she has played Game B n times and has lost every time.

a) Find the chance that the gambler wins a laptop.

b) Assume n is large, and find an exponential approximation to the chance in Part **a**.

c) Find the chance that the gambler wins a laptop or a smartphone.

d) Assume n is large, and find an exponential approximation to the chance in Part **c**.

e) In the cell below, write an expression the evaluates to your answer in Part **d**. Use `np.e` for e.

a)

[ ]:
```
# Answer to 1e
answer_1e = (1 - np.e**(-1/6)) + (np.e**(-1/6)) * (1 - np.e**(-1/3))
print(f"There is a {round(answer_1e*100, 3)}% chance that the player wins␣
  ↪either the laptop or smartphone")
```

#newpage

1.2 2. Fair Coin

One of the fundamental models of probability theory is for n tosses of a fair coin, where n is a fixed positive integer. The model says that all sequences that have length n and consist only of heads (H) and tails (T) are equally likely.

Unless otherwise specified, coins in this course are assumed to be fair.

This exercise is a series of quick observations. Before you start, look over the reference in the Basic Counting section of the Math Prerequisites page.

Suppose you toss a coin n times and note down the sequence of heads (H) and tails (T).

Fix an integer k such that $0 \leq k \leq n$.

a) In total, how many possible sequences are there? How many sequences have k heads?

[That means exactly k heads, now and throughout the course. To answer the second question, it might help to imagine that there are n empty spaces and you have to write the letter H in k of them.]

b) What is the chance that you get k heads in your n tosses? Why?

c) Does your answer in **b** make sense in the cases $k = 0$ and $k = n$? Explain.

d) `SciPy` is a Python library for scientific computing. You will be using it a lot in this course. In particular, the `special` module of `SciPy` computes combinatorial terms and has been imported in this notebook.

To calculate $\binom{n}{k}$, use `special.comb(n, k)` as in the example below.

[ ]:
```
# 10 choose 2

special.comb(10, 2)
```

Each student in a probability class is asked to toss a coin 20 times and note the number of heads. Six students do this exercise during office hours. What is the chance that none of these six students notes 10 heads?

[Review Part **a** for what "10 heads" means, and do not import any further libraries.]

[ ]:
```
# Answer to 2d

def complement(p):
    "Returns the complement of probability `p`"
    return 1 - p
def tosses(tosses, heads):
    """
    Returns the chance in proportion form that out of:
    `n` tosses of a fair coin
    there are _exactly_ `k` heads
    """
    possible = 2**tosses
    satisfied = special.comb(tosses, heads)
    return satisfied/possible

one_student = complement(tosses(tosses = 20, heads = 10))
answer = one_student**6
print(f"There is a {round(answer*100, 3)}% chance that none of the 6 students␣
 ↪in office hours get 10 heads form 20 rolls")
```

e) The calculation of $\binom{n}{k}$ involves factorials, and factorials get large very quickly. *Stirling's approximation* says that for large n

$$n! \sim \sqrt{2\pi n} \cdot (n/e)^n$$

where the symbol $\sim$ is read as "is asymptotically equivalent to" and means that the ratio of the two sides goes to 1 as n tends to ∞.

Let m be a positive integer. Use Stirling's formula to approximate the chance of getting m heads in $2m$ tosses of a fair coin, and say what the limit is as $m \to \infty$.

f) The Law of Averages for fair coins implies that if you keep tossing, then in the long run you get about half heads and half tails. Explain briefly why your answer to **e** doesn't contradict this statement.

#newpage

1.3 3. Combining Proportions

The Pew Research Foundation is a "nonpartisan fact tank" that conducts numerous careful surveys both nationally and internationally. The data below are from various Pew surveys in 2018 and 2019.

Remember the advice to draw diagrams. For the arithmetic, you can use the cell below the question.

a) In 2018, the adult population of the US was about 8.5 times the adult population of Canada. The percent of adults who didn't own a cell phone was 25% in Canada and only 6% in the US. Suppose you had picked one person at random from the combined adult population in the US and Canada in 2018. Pick the correct option below; if you pick (iii), fill in the blank with the chance.

Given that the selected person didn't own a cell phone, the chance that the person was from the US is

(i) $\frac{6}{6+25} \approx \frac{1}{5}$

(ii) not possible to find based on the information given

(iii) neither of the above; the chance is equal to _____

b) In 2019, the Pew Foundation surveyed US adults to ask if they had read books in print or digital formats in the past 12 months. Of the surveyed adults, - 1% did not respond - 27% responded that they had not read a book in any format in the past 12 months - 65% responded that they had read a print book in the past 12 months - 35% responded that they had read a digital book in the past 12 months

Suppose you picked one of the surveyed adults at random. Find the chance that the selected person responded that they had read both a print book and a digital book in the past 12 months, if it is possible to find it based on the information given. If this is not possible, explain why not.

c) The bar chart below summarizes some other results from the survey in Part **b**. For example, among the surveyed adults who were 50+ years old, 31% had not read a book in any format in the past 12 months.

Suppose one of the surveyed adults was picked at random. Answer the following question if it's possible to do so *based on the bar chart alone*. If it's not possible, explain why not. You can assume that everyone's age was recorded in completed years, and that adults are defined as people aged 18+.

Given that the selected person had not read a book in any format in the past 12 months, what is the chance that the person was in the 18-49 age group?

[1]:
```
# calculations for Ex 3
# PART A
top = .895*.06
bottom = top + (.105*.25)
part_a = top/bottom
print(part_a)
```

```
0.6716697936210132
```

1.4 Submission Instructions

Many assignments throughout the course will have a written portion and a code portion. Please follow the directions below to properly submit both portions.

1.4.1 Written Portion

- Scan all the pages into a PDF. You can use any scanner or a phone using applications such as CamScanner. Please **DO NOT** simply take pictures using your phone.
- Please start a new page for each question. If you have already written multiple questions on the same page, you can crop the image in CamScanner or fold your page over (the old-fashioned way). This helps expedite grading.
- It is your responsibility to check that all the work on all the scanned pages is legible.

1.4.2 Code Portion

- Save your notebook using File > Save and Checkpoint.
- Generate a PDF file using File > Download as > PDF via LaTeX. This might take a few seconds and will automatically download a PDF version of this notebook.
  - If you have issues, please make a follow-up post on the general HW 1 Ed thread.

1.4.3 Submitting

- Combine the PDFs from the written and code portions into one PDF. Here is a useful tool for doing so.
- Submit the assignment to Homework 1 on Gradescope.
- **Make sure to assign each page of your pdf to the correct question.**
- **It is your responsibility to verify that all of your work shows up in your final PDF submission.**

If you have questions about scanning or uploading your work, please post a follow-up to the Ed thread on this topic.

Homework 1

① a) L : the event that the player wins the laptop

$P(L) = 1 - P(\bar{L})$

$P(\bar{L})$: player plays all games and loses all games

$= \left(1 - \frac{1}{6n}\right)^n$

$\therefore P(L) = 1 - \left(1 - \frac{1}{6n}\right)^n$

b) $P(L) = 1 - \left(1 - \frac{1}{6n}\right)^n$

$\log(1 - P(L)) = n \log\left(1 - \frac{1}{6n}\right)$

$\log(1 - P(L)) \sim -\frac{1}{6}$

$1 - P(L) \sim e^{-1/6}$

$P(L) \sim 1 - e^{-1/6}$

(Via lemma 1)

c) W : Player wins laptop or smartphone

S : Player wins smartphone

$P(W) = P(L) + P(S)$

From process in (1a), if player plays game 2, they win with chance $1 - \left(1 - \frac{1}{3n}\right)^n$.

$P(S)$ = player plays game 2 · player wins game 2

$P(S) = P(\bar{L}) \cdot \left(1 - \left(1 - \frac{1}{3n}\right)^n\right)$

$P(S) = \left(1 - \frac{1}{6n}\right)^n \cdot \left(1 - \left(1 - \frac{1}{3n}\right)^n\right)$

$P(W) = 1 - \left(1 - \frac{1}{6n}\right)^n + \left(1 - \frac{1}{6n}\right)^n \cdot \left(1 - \left(1 - \frac{1}{3n}\right)^n\right)$

d) $P(W) = P(L) + P(S)$

$P(S) = \left(1 - \frac{1}{6n}\right)^n \cdot \left(1 - \left(1 - \frac{1}{3n}\right)^n\right)$

$P(S_1) = \left(1 - \frac{1}{6n}\right)^n$

$P(S_2) = \left(1 - \frac{1}{3n}\right)^n$

$P(S) = P(S_1) \cdot (1 - P(S_2))$

$\log P(S_1) = n \log\left(1 - \frac{1}{6n}\right)$

$P(S_1) \sim e^{-1/6}$ (via lemma 1)

$\log(P(S_2)) \sim n \log\left(1 - \frac{1}{3n}\right)$

$P(S_2) \sim e^{-1/3}$ (via lemma 1)

$P(S) = e^{-1/6} \cdot (1 - e^{-1/3})$

$P(L) \sim 1 - e^{-1/6}$ (via 1b)

$\Rightarrow P(W) = 1 - e^{-1/6} + e^{-1/6} \cdot (1 - e^{-1/3})$

e) See Notebook

Lemma 1

$P(A) = \left(1 - \frac{1}{an}\right)^n$

$\log(P(A)) = n \log\left(1 - \frac{1}{an}\right)$

$\sim n \cdot -\frac{1}{an}$

$\log(P(A)) \sim -\frac{1}{a}$

$P(A) \sim e^{-1/a}$

$\therefore \left(1 - \frac{1}{an}\right)^n \sim e^{-1/a}$

② a) For n tosses, 2^n possible sequences exist.

The number of sequences with k heads is the number of possible orderings of k heads and $n-k$ tails.

This is just $\binom{n-k+k}{k} = \binom{n}{k}$

or $\boxed{\frac{n!}{(n-k)!k!}}$

b) If (from 2a), there are $\binom{n}{k}$ ways to toss k heads in n tosses, an 2^n possible tosses, there are then if A is the event that there are k heads in n tosses then

$$P(A) = \frac{\binom{n}{k}}{2^n} = \frac{\frac{n!}{(n-k)!k!}}{2^n} =$$

$$\boxed{\frac{n!}{2^n k!(n-k)!}}$$

c) If $k=0$ or $k=n$, there is only one possible toss that satisfy each. These are all tails and all heads, respectively. So, in both, $P(A) = \frac{1}{2^n}$. Let us verify from part b to validate:

$k=0 \Rightarrow P(A) = \frac{n!}{2^n \cdot 1 \cdot n!} = \frac{1}{2^n}$

$k=n \Rightarrow P(A) = \frac{n!}{2^n \cdot n! \cdot 1} = \frac{1}{2^n}$

$\therefore$ The formula from (b) holds

d) See notebook

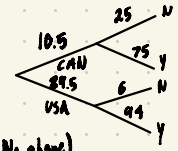
e) $$\binom{2m}{m} = \frac{(2m)!}{(m!)^2 2^{2m}} \sim \frac{\sqrt{2\pi 2m}\left(\frac{2m}{e}\right)^{2m}}{2^{2m}\left(\sqrt{2\pi m}\left(\frac{m}{e}\right)^m\right)^2} \sim \frac{\sqrt{4\pi m} \cdot 2^{2m} \cdot \frac{1}{e^{2m}}}{2^{2m} 2\pi m \cdot m^{2m} \cdot \frac{1}{e^{2m}}} \sim \frac{\sqrt{4\pi m}}{2\pi m \cdot m^{2m}} \sim \frac{\sqrt{\pi m}}{\pi m}$$

$$\sim (\pi m)^{-1/2}$$

$$\lim_{m\to\infty} (\pi m)^{-1/2} = \lim_{m\to\infty} \frac{\sqrt{m}}{m} = \boxed{\lim_{m\to\infty} \frac{1}{2\sqrt{m}} = 0}$$

f) The calculation in part e is the chance that we get EXACTLY half heads, while the law of large averages approaches ABOUT half, but not exactly. As m increases, the chance that there is a small deviation from EXACTLY half increases as well, thus: There is no contradiction, as Law of large Averages and 2e relate to slightly different events.

③ a)



$$P(\text{from US} \mid \text{No plane}) = \frac{(.895)(.06)}{(.895)(.06)+(.105)(.25)}$$

$$\approx 67.17\%$$

iii) neither of the above, the chance is equal to:

$$= \frac{(.895)(.06)}{(.895)(.06)+(.105)(.25)}$$

$$\approx 67.17\%$$

b) 1% : No response

27% : No books

$\Rightarrow$ 72% read books

65% print

35% digital

If 65% read print and 35% read digital, and only 72% read anything $65+35-72 = 100-72 = 28\%$ must have read both print and digital.

28% chance

c) Adults: 27%

18-49 : 22%

50+ : 31%

Young % : Young count + Old % · old count

$$22 \cdot x + 31 \cdot y = 27$$

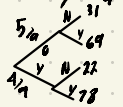
$x + y = 1$ $\quad x$: % of sample under 50

$x = 1 - y$ $\quad y$: % of sample above 50

$$22 + 9y = 27$$

$$9y = 5$$

$$y = \frac{5}{9} \Rightarrow x = \frac{4}{9}$$



$$P(\text{under 50} \mid \text{no reading}) = \frac{\frac{4}{9} \cdot .22}{\left(\frac{4}{9} \cdot .22\right)+\left(\frac{5}{9} \cdot .31\right)}$$

$$\approx 36.21\%$$