Homework 1

1a. $P(\text{lose laptop}) = 1 - P(\text{win laptop})$

* after n times $= \left(1 - \frac{1}{8n}\right)^n$

$P(\text{win laptop}) = 1 - P(\text{lose laptop})$

$= 1 - \left(1 - \frac{1}{8n}\right)^n$

Homework_01

January 27, 2025

Probability for Data Science

UC Berkeley, Spring 2025

Michael Xiao and Ani Adhikari

[4]:
```
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 (Due Monday, January 27th at 5 PM)

1.0.1 Instructions

Your homeworks will generally 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. Start the work for the written portions of each section on a new page. 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 lab are submitted completely and properly to Gradescope. **Make sure to assign each page of your pdf to the correct question. 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 We will not grade assignments which do not have pages selected for each question.

1.2 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 $\frac{1}{8n}$ 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.

[5]:
```
# Answer to 1e

1 - np.e**(-1/8 - 1/3)
```

[5]:
```
0.36766333781375016
```

1.3 2. 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?

[6]:
```
# calculations for Ex 2

# Work done on paper
```

1.4 3. Two IID Random Variables

Let $N \geq 5$ be a fixed integer. Let X_1 and X_2 be independent and identically distributed (i.i.d.) random variables, and let $P(X_1 = i) = p_i$ where $0 \leq i \leq N$ and $\sum_{i=0}^{N} p_i = 1$.

Provide (with justification) expressions for the following probabilities, in terms of $p_0, p_1, \ldots, p_N$.

a) $P(X_1 = X_2)$

b) $P(X_1 = X_2 \mid X_1 > 3)$

c) $P(X_1 \neq X_2 \mid X_1 > 3)$

d) $P(X_1 + X_2 \leq N \mid X_1 = 2)$

e) $P(X_1 + X_2 \leq N \mid X_1 > 2)$

1.5 4. Extrema and Tails

THIS IS A HUGELY IMPORTANT EXERCISE or I wouldn't be using bold all-caps. Make sure you understand the logic and especially the suggested visualization; that's what's going to be applied later. Note also how the logic can be applied or modified in variations of the initial problem setting.

The maximum and minimum of a random sample of numbers are called the *extrema* of the sample. Distributions of extrema are best described using the left or right hand tail probabilities. In this exercise you will see how.

Fix a positive integer n. In data science, a *sample* is a sequence of random variables $X_1, X_2, \dots, X_n$ defined on an outcome space, and n is called the *sample size*.

Suppose our sample is positive integer valued, as follows. Fix a positive integer N and suppose each X_i has possible values in the set $\{1, 2, 3, \dots, N\}$.

Let $V_n = \min\{X_1, X_2, \dots, X_n\}$ be the *sample minimum* and let $W_n = \max\{X_1, X_2, \dots, X_n\}$ be the *sample maximum*.

As you know from Data 8, each of V_n and W_n is a statistic.

a) [Describing an event] The event that a sample maximum is "small" is straightforward to describe in terms of the individual elements of the sample. To see this, fill in the blank with an appropriate mathematical symbol or English phrase, making no further assumptions about the random variables involved. Justify your answer.

Fix an integer k such that $1 \le k \le N$. The event "$W_n \le k$" is the same as the event "each of $X_1, X_2, \dots, X_n$ is _____k".

It might help to draw the number line, mark the integers 1 through N, and put a special mark on the integer k. For the maximum to be at or to the left of k, where do all the X's have to be?

b) [Using a probability model] In Data 8, you found empirical distributions of statistics by simulation. Now you will start finding the exact distribution of a statistic by probability theory. For this, you need a probability model. Let's start with a simple one. Suppose n draws are made at random with replacement from the numbers $\{1, 2, 3, \dots, N\}$. Let X_i be the number that appears on the ith draw. Use Part **a** to find $P(W_n \le k)$, for each k in the range $1 \le k \le N$.

c) [Using tail probabilities] Continue under the assumptions in Part **b**. Use your sketch in Part **a** to express the event $\{W_n = k\}$ in terms of any subset of the events $\{W_n \le 1\}, \{W_n \le 2\}, \dots, \{W_n \le N\}$. Use this and Part **b** to find $P(W_n = k)$ for $1 \le k \le N$ and to show algebraically that $\sum_{k=1}^{N} P(W_n = k) = 1$. Congratulations – you have found the distribution of the sample maximum for a sample of independent uniform draws from the integers 1 through N.

d) [Changing the statistic] Continue under the assumptions in Part **b**. Modify Parts **a** through **c** to find the distribution of the sample minimum V_n, as follows. For the event that the sample minimum is "large", fill in the blank with an appropriate mathematical symbol or English phrase.

The event "$V_n > k$" is the same as the event "each of $X_1, X_2, \dots, X_n$ is _____k".

Use this observation and the ideas of the previous parts to find $P(V_n = k)$ for $1 \le k \le N$.

e) [Changing the probability model] Now assume $N > n$ and suppose $X_1, X_2, \ldots, X_n$ are draws made at random without replacement from the numbers $\{1, 2, 3, \ldots, N\}$. Apply the logic in Parts **a** and **c** to find the distribution of the sample maximum W_n.

In this part is is very important that you think carefully about the possible values of W_n before calculating chances.

1.6 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.

1.6.1 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.
- If you used LaTeX to do the written portions, you do not need to do any scanning; you can just download the whole notebook as a PDF via LaTeX.

1.6.2 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 post a follow-up on the general Homework 1 Ed thread.

1.6.3 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 are having difficulties scanning, uploading, or submitting your work, please read the Ed Thread on this topic and post a follow-up on the general Homework 1 Ed thread.

[ ]:

1b. $p = \left(1 - \frac{1}{8n}\right)^n$

$\log(p) = n \log\left(1 - \frac{1}{8n}\right)$

$\log(p) = \cancel{n} \cdot -\frac{1}{8\cancel{n}}$

$e^{\log(p)} = e^{-\frac{1}{8}}$

$p = e^{-\frac{1}{8}}$

$\therefore P(\text{win laptop}) = 1 - e^{-\frac{1}{8}}$

1c. P(lose smartphone after n times) = 1 - P(win smartphone)

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

P(win smartphone) = 1 - P(lose smartphone)

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

P(win laptop OR smartphone) = P(win laptop) + P(lose laptop) * P(win smartphone)

$= 1 - \left(1 - \frac{1}{8n}\right)^n + \left[\left(1 - \frac{1}{8n}\right)^n * \left(1 - \left(1 - \frac{1}{3n}\right)^n\right)\right]$

3b. $P(X_1 = X_2 \mid X_1 > 3) = \dfrac{P((X_1 = X_2) \cap (X_1 > 3))}{P(X_1 > 3)}$

$$P((X_1 = X_2) \cap (X_1 > 3)) = \sum_{i=4}^{N} P(X_1 = i, X_2 = i)$$
$$= \sum_{i=4}^{N} P(X_1 = i)\, P(X_2 = i)$$
$$= \sum_{i=4}^{N} P_i^2$$

$$P(X_1 > 3) = \sum_{i=4}^{N} P(X_1 > 3)$$
$$= \sum_{i=4}^{N} P_i$$

$$\therefore P(X_1 = X_2 \mid X_1 > 3) = \frac{\sum_{i=4}^{N} P_i^2}{\sum_{i=4}^{N} P_i}$$

3c. $P(X_1 \neq X_2 \mid X_1 > 3) = 1 - P(X_1 = X_2 \mid X_1 > 3)$

$$= 1 - \frac{\sum_{i=4}^{\infty} p_i^2}{\sum_{i=4}^{\infty} p_i}$$

3d. $P(X_1 + X_2 \leq N \mid X_1 = 2) = \dfrac{P(X_1 + X_2 \leq N,\ X_1 = 2)}{P(X_1 = 2)}$

$P(X_1 + X_2 \leq N,\ X_1 = 2) = P(X_1 = 2) \times P(2 + X_2 \leq N)$

$P(X_2 \leq N - 2) = \sum_{i=0}^{N-2} p_i$

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

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

$\log(P) = n \cdot -\frac{1}{3n}$

$e^{\log(P)} = e^{-\frac{1}{3}}$

$P = e^{-\frac{1}{3}}$

$\therefore P(\text{win smartphone}) = 1 - e^{-\frac{1}{3}}$

$\therefore P(\text{win laptop OR smartphone}) = P(\text{win laptop}) + P(\text{lose laptop}) \times P(\text{win smartphone})$

$= 1 - e^{-\frac{1}{8}} + e^{-\frac{1}{8}} \times \left(1 - e^{-\frac{1}{3}}\right)$

$= 1 - e^{-\frac{1}{8}} + e^{-\frac{1}{8}} - e^{-\frac{1}{8}-\frac{1}{3}}$

$= 1 - e^{-\frac{1}{8}-\frac{1}{3}}$

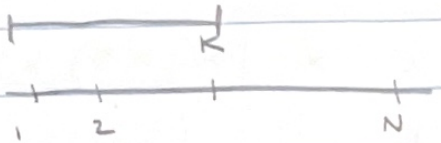
3a. $P(X_1 = X_2) = \sum_{i=0}^{2M} P(X_1 = i, X_2 = i)$

$= \sum_{i=0}^{2M} P(X_1 = i) P(X_2 = i)$

$= \sum_{i=0}^{2M} p_i \times p$

$= \sum_{i=0}^{2M} p_i^2$

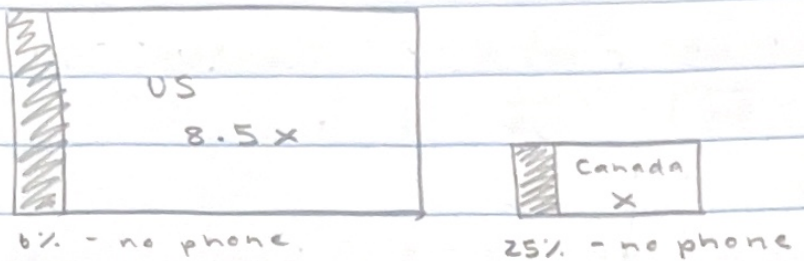
4 a.



each x_i ($i = 1, 2, \dots n$) must satisfy $x_i \leq k$

$w_n \leq k$ is the same as the event 'each of $x_1, x_2, \dots x_n$ is $\leq k$'

2a



$$P(US \mid \text{no phone}) = \frac{P(\text{no phone} \mid US) \cdot P(US)}{P(\text{no phone})}$$

$$= \frac{0.06 \times \frac{8.5x}{8.5x + x}}{\left(.06 \times \frac{8.5x}{8.5x + x}\right) + \left(.25 \times \frac{x}{8.5x + x}\right)}$$

$$= \frac{.06 \times \frac{8.5}{9.5}}{\left(.06 \times \frac{8.5}{9.5}\right) + \left(.25 \times \frac{1}{9.5}\right)}$$

2b. $P(\text{print and digital}) = P(\text{print}) + P(\text{digital}) - P(\text{print or digital})$

$P(\text{print or digital}) = 1 - P(\text{no reading})$

$= 1 - 0.27$

$= 0.73$

$\therefore P(\text{print and digital}) = 0.65 + 0.35 - 0.73$

$= 0.27$

4b. $P(W_n \leq k) = P(X_1 \leq k) * P(X_2 \leq k) * \dots * P(X_n \leq k)$

$P(X_i \leq k) = \frac{k}{N}$ (random w/ replacement, equal chance)

$\therefore P(W_n \leq k) = \left(\frac{k}{N}\right)^n$

2c. $P(18-49 \mid \text{no reading}) = \frac{P(\text{no reading} \mid 18-49) \times P(18-49)}{P(\text{no reading})}$

$\therefore P(\text{no reading}) = P(\text{no reading} \mid 18-49) \times P(18-49) + P(\text{no reading} \mid 50+) \times P(50+)$

since population proportions for ages 18-49 and 50+ are unknown...

$P(18-49) = p$

$P(50+) = 1-p$

$\therefore P(\text{no reading}) = .22 \times p + .31 \times (1-p)$

$.27 = .22p + .31 - .31p$

$0.09p = 0.04$

$p \approx 0.444$

$P(18-49) = 0.444$

$P(50+) = 0.56$

$\therefore P(18-49 \mid \text{no reading}) = \frac{.22 \times .444}{.27}$

$\approx .362$

36.2% chance that person was in the 18-49 age group