1a. fair coin : C_F : f

biased coin : $C_T = C_H$: 1 - f

$\therefore P(C_H) = \frac{1}{2}(1-f)$

1b. $P(H_1) = P(H_1 | C_F)P(C_F) + P(H_1 | C_T)P(C_T) + P(H_1 | C_H)P(C_H)$

$P(H_1 | C_F) = 0.5$

$P(H_1 | C_T) = 0.1$

$P(C_1 | C_H) = 0.9$

$P(H_1) = (0.5)(f) + (0.1)(\frac{1-f}{2}) + (0.9)(\frac{1-f}{2})$

$= 0.5f + 0.05(1-f) + 0.45(1-f)$

$= 0.5f + 0.05 - 0.05f + 0.45 - 0.45f$

$= 0.5$

$P(H_1) = 0.5$, → symmetry ensures that the probability of landing heads remains 50%, as the biases of the coins are symmetrically distributed, causing the bias effects to cancel out.

1c. $P(C_H | H_1) > P(C_H)$

Given that the first toss lands heads, the chance that the coin is biased towards heads increases, as a heads-biased coin has a higher chance of landing heads with 90% chance than a tails-biased coin with 10% chance or fair coin with 50%. observing heads makes it more likely that the coin is biased towards heads, as this outcome is more consistent with the behaviors of a heads-biased coin.

1d. $P(C_H \mid H_1) = \frac{P(H_1 \mid C_H)\,P(C_H)}{P(H_1)}$

$= \frac{(0.9)(\frac{1}{2}(1-f))}{0.5}$

$= \frac{0.9(1-f)}{2(0.5)}$

$= 0.9(1-f)$

$\therefore P(C_H \mid H_1) > P(C_H)$

$0.9(1-f) > 0.5(1-f)$

1e. $P(H_2) = P(H_2 \mid C_F)P(C_F) + P(H_2 \mid C_T)P(C_T) + P(H_2 \mid C_H)P(C_H)$

$=$

$P(H_2 \mid C_F) = 0.5$

$P(H_2 \mid C_T) = 0.1$

$P(H_2 \mid C_H) = 0.9$

$P(H_2) = (0.5)(f) + (0.1)(\frac{1-f}{2}) + (0.9)(\frac{1-f}{2})$

$= 0.5f + 0.05 - 0.05f + 0.45 - 0.45f$

$= 0.5$

$P(H_1 H_2) = P(H_2 \mid H_1)\,P(H_1)$

$= P(H_2 \mid C_F)P(H_1 \mid C_F)P(C_F) + P(H_2 \mid C_T)P(H_1 \mid C_T)P(C_T)$

$+ P(H_2 \mid C_H)P(H_1 \mid C_H)P(C_H)$

$= (0.5)(0.5)(f) + (0.1)(0.1)(\frac{1-f}{2}) + (0.9)(0.9)(\frac{1-f}{2})$

$= 0.25f + 0.005(1-f) + 0.405(1-f)$

$= 0.25f + 0.005 - 0.005f + 0.405 - 0.405f$

$= -0.16f + 0.41$

$P(H_1)\,P(H_2) = (0.5)(0.5)$

$= 0.25$

1c (cont.)

$f = 0 \longrightarrow P(H_1, H_2) = -0.16(0) + 0.41$

$= 0.41$

$f = 1 \longrightarrow P(H_1, H_2) = -0.16(1) + 0.41$

$= 0.25$

$\therefore P(H_1, H_2) \neq P(H_1)P(H_2)$

If. $P(H_2 \mid H_1) > P(H_2)$

$$P(H_2 \mid H_1) = \frac{P(H_1, H_2)}{P(H_1)}$$

$$= \frac{-0.16f + 0.41}{0.5}$$

$$= -0.32f + 0.82$$

$f = 0 \longrightarrow P(H_2 \mid H_1) = -0.32(0) + .82$

$= 0.82$

$f = 1 \longrightarrow P(H_2 \mid H_1) = -0.32(1) + 0.82$

$= 0.5$

Observing heads in first toss increases the chance of seeing heads in second toss, meaning the two events are not independent.

2a. $P(X_a = 3, X_b = 3, X_c = 3, X_d = 3) = \dfrac{\binom{6}{3}\binom{6}{3}\binom{6}{3}\binom{6}{3}}{\binom{24}{12}}$

Since we take a simple random sample of 12 individuals, each category will have exactly 3 individuals. 6 individuals exist in each category, so the total is 24 individuals. Hence, $\binom{6}{3}$ represents the probability of choosing 3 individuals among the 6 individuals in each of the four categories over $\binom{24}{12}$, which represents the chosen 12 individuals among the total population of 24. We use hypergeometric distribution because it is without replacement (simple random sample) with finite population.

2b. $P(X_1 = 5, X_2 = 4, X_3 = 2, X_4 = 1) = \dfrac{\binom{6}{5}\binom{6}{4}\binom{6}{2}\binom{6}{1}}{\binom{24}{12}} \times 4!$

$\binom{6}{k}$ represents the k chosen individuals for each category. We multiply this by $4!$ to take into account the number of ways to assign the individuals into the four categories.

2c. $P(X_1 = 4, X_2 = 4, X_3 = 3, X_4 = 1) = \frac{\binom{6}{4}\binom{6}{4}\binom{6}{3}\binom{6}{1}}{\binom{24}{12}} * \binom{4}{2} * 2$

$\binom{6}{k}$ represents the k chosen individuals in each category. We multiply by $\binom{4}{2}$ because it is the number of ways of choosing 2 categories out of 4 to assign the 4 individuals in. We also multiply by 2 because there are 2 ways of specifying the assignment of individuals into the remaining two categories

2d. $P(X_1 = 6, X_2 = 2, X_3 = 2, X_4 = 2) = \frac{\binom{6}{6}\binom{6}{2}\binom{6}{2}\binom{6}{2}}{\binom{24}{12}} * \binom{4}{1}$

We multiply by $\binom{4}{1}$ because we have to choose one category where all 6 individuals will be placed. Since there can only be 6 individuals in one category, the remaining three categories must have 2 individuals to have equal numbers, hence $\binom{6}{2}^3$.

3a. $X_1 \sim \text{hypergeometric}(N = 24, G = 6, n = 6)$

$$P(X_1 = k) = \frac{\binom{6}{k}\binom{18}{6-k}}{\binom{24}{6}}$$

Since students are assigned at random into breakout rooms, we use hypergeometric dist. to find how many freshmen end up in a specific room, specifically breakout room 1. $\binom{6}{K}$ represents K number of freshmen chosen into the breakout room of six students.

3b. $(X_1, X_4) \sim \text{hypergeometric}(24, 6, n_1 = n_4 = 6)$

$$P(X_1 = k_1, X_4 = k_4) = \frac{\binom{6}{k_1}\binom{6}{k_4}\binom{12}{6-k_1-k_4}}{\binom{24}{12}}$$

Similar to above, $\binom{6}{k_1}$ and $\binom{6}{k_4}$ represent k_1 and k_4 freshmen that are in breakout rooms 1 and 4. Since we are selecting 12 students among the total 24, the total number of freshmen chosen cannot exceed 6, as there are only 6 in the zoom meeting, hence $\binom{12}{6-k_1-k_4}$

3c. $P\left(\bigcup_{i=1}^{4} A_i\right) = \sum_{i=1}^{4} P(A_i) - \sum\sum_{1 \le i < j \le 4} P(A_i A_j)$

$$P(A_i) = \frac{\binom{6}{6}\binom{18}{0}}{\binom{24}{6}} \times 4$$

$$P(A_i A_j) = \frac{\binom{6}{6}\binom{6}{6}\binom{12}{0}}{\binom{24}{6}\binom{18}{6}} \times 4 \times 3$$

$$= \frac{\binom{6}{6}\binom{6}{6}\binom{12}{0}}{\binom{24}{6}\binom{18}{6}} \times 12$$

$$\therefore P\left(\bigcup_{i=1}^{4} A_i\right) = \underbrace{4\,\frac{\binom{6}{6}\binom{18}{0}}{\binom{24}{6}}}_{P(A_i)} - \underbrace{12\,\frac{\binom{6}{6}\binom{6}{6}\binom{12}{0}}{\binom{24}{6}\binom{18}{6}}}_{P(A_i A_j)}$$

$P(A_i)$ finds the probability that a specific breakout room has all 6 students from the same year. We multiply by 4 to take into account the 4 possible years (freshmen, sophomore, juniors, seniors). $P(A_i A_j)$ is similar to $P(A_i)$, but takes into account the remaining 3 years for the second breakout room after assigning a year to the first breakout room. Using inclusion-exclusion rule

4a. $P(\text{at least one late}) = 1 - P(\text{no late})$

$$P(\text{no late}) = 1 - 0.01$$
$$= (0.99)^{10}$$

$$P(\text{at least one late}) = 1 - (0.99)^{10}$$
$$P\left(\bigcup_{i=1}^{10} A_i\right) \approx 0.0956 \rightarrow \text{lower bound}$$

$$\sum_{i=1}^{10} P(A_i) = P(A_1) + P(A_2) + \ldots + P(A_{10})$$
$$= 10 \times 0.01$$
$$= 0.1$$

$$\therefore \quad 0.0956 \leq P(\text{at least one late}) \leq 0.1$$

Since we cannot assume independence for all 10 flights, we cannot determine if one flight being late makes another flight less likely to be late. So, we use Boole's inequality to combat the lack of knowledge of dependency.

4b. $P(\text{at least one suit missing}) \leq P(\text{all missing})$

$$P(\text{at least one suit missing}) = \frac{\binom{39}{13}}{\binom{52}{13}}$$

$$P(\text{all missing}) = \frac{\binom{37}{13}}{\binom{52}{13}} \times 4$$

$$\frac{\binom{39}{13}}{\binom{52}{13}} \leq P(\text{at least one suit missing}) \leq 4 \times \frac{\binom{39}{13}}{\binom{52}{13}}$$

4b (cont) We use bounds and not exact probability because it becomes difficult to count all possible hands where at least one suit is missing.

4c. P(at least one slot unchosen) $\leq$ P(all slots chosen)

P(at least one slot unchosen) $= (1 - \frac{1}{5})$

$= (1 - \frac{1}{5})^5$

P(all slots chosen) $= 1 - P$(at least one slot unchosen)

$= 1 - (1 - \frac{1}{5})^5 \times 5$

$(1 - \frac{1}{5})^5 \leq$ P(at least one slot unchosen) $\leq 1 - (1 - \frac{1}{5})^5 \times 5$

We use bounds instead of exact probability because the chance of each slot being chosen is determined by the others. They can all choose the same room, all choose different rooms, or a combination of both.

4d. P(rain every day) = P(sun) × P(mon) × ... × P(sat)

$= (0.9)(0.95)(0.95)(0.7)(0.9)(0.85)(0.8)$

≈ 0.447

We use exact probability over bounds because the chance of rain is given, so we can assume independence in that rain on any day does not affect the chance of rain on any other day.

$X \sim$ # of successes

$$P(X \geq 10) = 1 - \sum_{k=0}^{9} \binom{30}{k} \left(\frac{1}{3}\right)^k \left(1 - \frac{1}{3}\right)^{30-k}$$

5a. `1 - stats.binom.cdf(9, 30, 1/3)`

The binomial cumulative distribution function gives the probability of getting at most 9 successes, but subtracting from 1 ensures we get at least 10 successes. Probability mass function only gives us 10 successes

5b. $Y \sim$ # of draws to get 10 successes

$$P(Y \geq 10) = \sum_{10}^{30} \binom{30}{k} \left(\frac{1}{3}\right)^{10} \left(1 - \frac{1}{3}\right)^{k-10}$$

`1 - stats.binom.cdf(9, 30, 1/3)`

This answer relates to part a because the probability of at least 1/3 of the 30 draws being successful and getting 10 successes with the 30 draws is the same idea and calculation. Both require you to figure out the chance of at least 10 successes

5c. $X \sim$ # successes I make in 30 draws

$Y \sim$ # successes friend makes in 30 draws

$$P(X > Y) = \sum_{k=0}^{20} P(X > k) \cdot P(Y = k)$$

$P(X > Y)$ considers all possible Y values and find the X values that exceeds those values. I use binomial pmf to find the exact k successes for my friend, and binomial cdf to find the probability that I get more than k successes

6a. exact : Binomial (1000, 0.003)

approx : Poisson (1000 × 0.003)

: Poisson (3)

6b. $Y \sim$ # of failures $\quad q = 1-p$

$$P(Y=k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

$X \sim$ # successes

$X = n - Y$

$$P(X=k) = P(Y=n-k)$$

$$= \frac{e^{-\lambda}\lambda^{n-k}}{(n-k)!}$$

Homework_02

February 5, 2025

Probability for Data Science

UC Berkeley, Spring 2025

Michael Xiao and Ani Adhikari

[5]:
```
from prob140 import *
from datascience import *
import numpy as np
from scipy import special
from scipy import stats

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

1 Homework 2 (Due Monday, February 3rd 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.**

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."

1.1 1. Random Coin

This exercise is an introduction to experiments in which randomizing a parameter can affect dependence and independence. The setting is simple but powerful: it demonstrates some fundamental ideas of Bayesian prediction. We will use the same ideas in more complex settings later in the course.

Suppose you have a coin that has a fixed (non-random) probability p of heads. Under this assumption the tosses are independent, that is, knowing the results of some tosses doesn't change probabilities for how other tosses will come out. So for example the chance of two heads in two tosses is $p \cdot p$.

Now let's see what happens when the coin is picked randomly from a bunch of different coins. In this situation, the chance of heads is random.

In a bag of coins, a proportion f of the coins are fair; assume $0 < f < 1$. Of the remaining coins, half are biased towards tails and land heads with chance 0.1; the other half land heads with chance 0.9.

A coin is picked at random from the bag and tossed twice. Define the following events.

- C_F is the event that the selected coin is fair, C_T is the event that the selected coin is biased towards tails, and C_H is the event that the selected coin is biased towards heads.
- H_1 is the event that the first toss lands heads and H_2 is the event that the second toss lands heads.

a) Find $P(C_H)$.

b) Calculate $P(H_1)$. Simplify the answer as much as possible and then explain it by symmetry.

c) Without calculation, fill in the blank with one of the symbols $>$, $<$, or $=$. Explain your reasoning.

$P(C_H \mid H_1)$ _____ $P(C_H)$

d) Now calculate $P(C_H \mid H_1)$ and show that it is consistent with your answers to Parts **a** and **c**.

e) Find $P(H_2)$ and (carefully!) $P(H_1H_2)$. Is $P(H_1H_2) = P(H_1)P(H_2)$?

f) Fill in the blank with one of the symbols $>$, $<$, or $=$. Justify your choice by calculation and also explain it intuitively.

$P(H_2 \mid H_1)$ _____ $P(H_2)$

1.2 2. Counting Categories

In each part below, remember to **explain your answer. Don't just write down a formula.** You don't have to provide decimal answers.

Note: See the video right at the bottom of Section 5.4 (below the last Quick Check) for an effective approach.

A population consists of 6 individuals in each of 4 categories A, B, C, and D. A simple random sample of 12 individuals is chosen from the population.

a) Find the chance that the sample contains equal numbers of individuals in the four categories.

b) Find the chance that the sample contains 5 individuals in one category, 4 in another, 2 in a third, and 1 in the remaining category.

c) Find the chance that the sample contains 4 individuals in each of two categories, 3 individuals in another, and 1 in the remaining category.

d) Find the chance that the sample contains all 6 individuals in one category and equal numbers in the other three.

1.3 3. Breakout Rooms

Among 24 students in a Zoom meeting, there are six each of freshmen, sophomores, juniors, and seniors. The 24 students are split at random into four breakout rooms of six students each.

a) What is the distribution of the number of freshmen in Breakout Room 1? **Why?**

b) What is the joint distribution of the numbers of freshmen in Breakout Rooms 1 and 4? **Why?**

[Note: You don't have to write out a joint distribution table. A more compact way is to provide a formula. Remember to specify the possible values appropriately.]

c) What is the chance that in at least one of the breakout rooms all the students are in the same year? **Explain your calculation.**

1.4 4. Exact Value or Bound

In each part below, find the exact value of the probability if it is possible to do so with the information given. If it is not possible, provide the best lower and upper bounds you can. **Explain your choices.**

a) the chance that at least one of 10 flights at an airport is late, if each flight at the airport has a 1% chance of being late

b) the chance that not all suits appear in a bridge hand of 13 cards dealt at random without replacement from a standard deck

[A standard deck consists of 13 cards in each of 4 suits, making 52 cards in all.]

c) the chance that all s office hours slots are selected, if each of $g \geq s$ GSIs selects one of the s slots at random without being influenced by the choices of others

d) the chance that it rains every day next week if the daily chances of rain are given by

Sun	Mon	Tue	Wed	Thu	Fri	Sat
0.9	0.95	0.95	0.9	0.9	0.85	0.8

1.5 5. College Degrees

In the U.S., 38% of adults aged 25 and over have a four-year college degree.

In each part below, write a math expression for the chance and provide a brief justification. Please use the appropriate summation, not "...". Then use the appropriate code cell to find the numerical value of the chance, using `stats.binom.pmf`, `stats.binom.cdf`, and arithmetic. See the textbook for a reference. The `stats` library of `scipy` has been imported in the top cell of this notebook.

In what follows, we will use the term *population* to mean US adults aged 25 and over, and a *successful draw* to mean a draw that results in a person who has a four-year college degree.

a) Suppose I draw from the population at random with replacement. What is the chance that at least one-third of my first 30 draws are successful?

b) Suppose I draw from the population at random with replacement, till 10 of my draws are successful. What is the chance that I draw at most 30 times? How is this answer related to the answer in Part **a**?

c) Suppose my friend and I both draw from the population at random with replacement, independently of each other. Suppose I make 30 draws and my friend makes 20 draws. What is the chance that I get more successful draws than my friend?

[6]:
```
# Answer to a (you can use more than one line of code)

1 - stats.binom.cdf(9, 30, 1/3)
```

[6]: 0.56825564442147991

[7]:
```
# Answer to b (you can use more than one line of code)

1 - stats.binom.cdf(9, 30, 1/3)
```

[7]: 0.56825564442147991

[10]:
```
# Answer to c (you can use more than one line of code)

total_prob = 0

for i in np.arange(21):
    prob_x = 1 - stats.binom.cdf(i, 30, 1/3)
    prob_y = stats.binom.pmf(i, 20, 1/3)
    total_prob += prob_x * prob_y

total_prob
```

[10]: 0.80227680420107106

1.6 6. Poisson Approximation at Both Ends

Consider n independent Bernoulli (p) trials.

a) Fill in the blanks with names of distributions along with parameters in parentheses: If $n = 1000$ and $p = 0.003$, the distribution of the number of successes is exactly _____ (_____) and approximately _____ (_____).

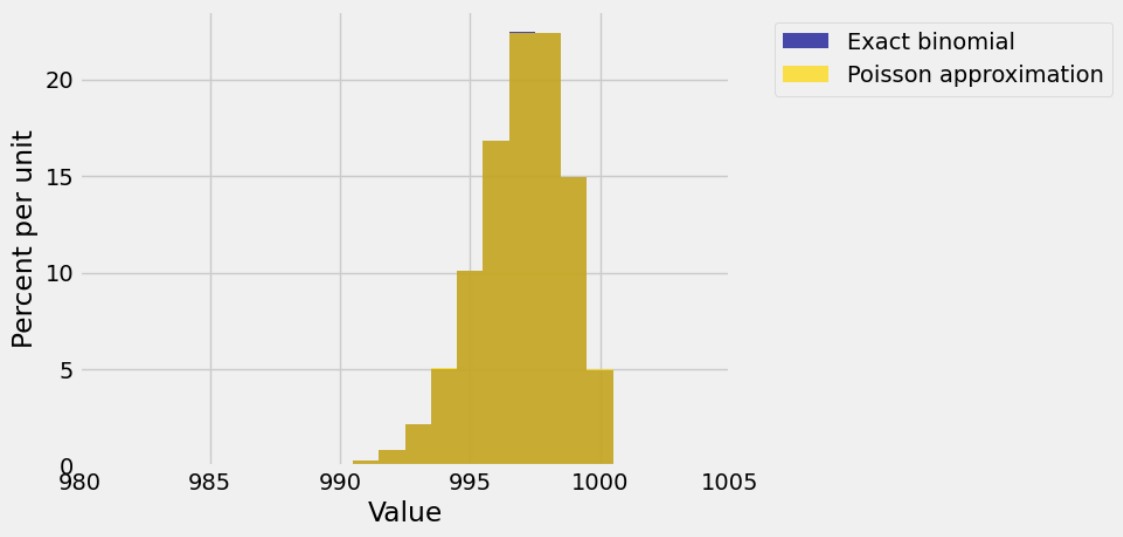
b) Let n be large and let p be close to 1. Find a Poisson approximation to p_k (the chance of k successes) by an appropriate use of the Poisson approximation to the binomial derived in the textbook.

Note: Don't try to derive a new limit from scratch. Just use the limit already derived in the textbook, but appropriately.

c) Plot the probability histogram of the binomial (1000, 0.997) distribution, and overlay your Poisson approximation from part (b). For computing Poisson probabilities, see the textbook. Please don't plot the entire range of the binomial. Choose an informative range of values on the horizontal axis.

[11]:
```
# Answer to c

n = 1000
p = 0.997

k = np.arange(1001)                # array of possible values
binomial_probs = stats.binom.pmf(k, n, p)    # array of exact binomial␣
 ↪probabilities

def poisson_approximation_pmf(j):
    """Returns the Poisson approximation to the
    exact binomial probability of j successes"""
    return stats.poisson.pmf((n-j), n*(1-p))

exact_binomial = Table().values(k).probabilities(binomial_probs)
poisson_approximation = Table().values(k).
 ↪probability_function(poisson_approximation_pmf)

Plots("Exact binomial", exact_binomial, "Poisson approximation",␣
 ↪poisson_approximation)
plt.xlim(980, 1005)
```

[11]:
```
(980.0, 1005.0)
```



1.7 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.

1.7.1 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.

1.7.2 Code Portion

- Save your notebook using `File > Save Notebook`.
- Generate a PDF file using `File > Save and Export Notebook As > PDF`. 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 2 Ed thread.

1.7.3 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 2 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 2 Ed thread.

1.8 We will not grade assignments which do not have pages selected for each question.