Problem 1

1)

$P(A) = \frac{N_1}{N}$ — chance of sampling male from total population

$P(B) = \frac{T_0 + T_1}{N}$ — chance of sampling either male, tall OR female, tall from total population

$P(A|B) = \frac{T_1}{T_0 + T_1}$ — chance of sampling male from all the tall people

$P(B|A) = \frac{T_1}{N_1}$ — chance of sampling tall male from all the males

$P(A \cap B) = \frac{T_1}{N}$ — chance of sampling tall male from total population

2)

$P(A \cap B) = \frac{T_1}{N}$

$P(A)P(B|A) = \frac{N_1}{N} * \frac{T_1}{N_1} = \frac{T_1}{N} = P(A \cap B)$✔

$P(B)P(A|B) = \frac{T_0 + T_1}{N} * \frac{T_1}{T_0 + T_1} = \frac{T_1}{N} = P(A \cap B)$✔

3)

$P(B) = \frac{T_0 + T_1}{N}$

$P(A)P(B|A) = \frac{N_1}{N} * \frac{T_1}{N_1} = \frac{T_1}{N}$

$P(A^C)P(B|A^C) = \frac{N_0}{N} * \frac{T_0}{N_0} = \frac{T_0}{N}$

$P(A)P(B|A) + P(A^C)P(B|A^C) = \frac{T_1}{N} + \frac{T_0}{N} = \frac{T_0 + T_1}{N} = P(B)$ ✔

4)

We have from Part (2) that $P(A \cap B) = P(B)P(A|B)$

So, $P(A|B) = \frac{P(A \cap B)}{P(B)}$. ✔

We also have from Part (2) that $P(A \cap B) = P(A)P(B|A)$,

and have from Part (3) that $P(B) = P(A)P(B|A) + P(A^C)P(B|A^C)$

Substituting these in, we get:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^C)P(B|A^C)}$$ ✔

Take home message:

Probability = population proportion

Conditional probability = proportion within subpopulation

5)

Approximate using the sampled counts and using sampling size instead of population:

$P(A) \approx \frac{n_1}{n}$

$P(B) \approx \frac{t_0 + t_1}{n}$

$P(A|B) \approx \frac{t_1}{t_0 + t_1}$

$P(B|A) \approx \frac{t_1}{n_1}$

$P(A \cap B) \approx \frac{t_1}{n}$

Take home message:

Probability (fixed) ≈ frequency (fluctuating)

Conditional probability (fixed) ≈ relative frequency (fluctuating)

6)

If we repeat random sampling from the same population n times independently, then there will be N^n equally likely sequences.
The answers in Part (5) are only approximations of the actual probabilities since there are fluctuations between different equally likely sequences within the hyper-population.

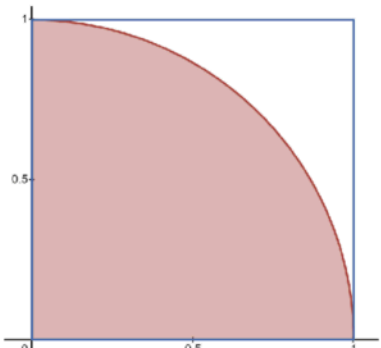
As the number of samples increases, the amount of sequences in the hyper-population that are representative of the theoretical probability (i.e., give close approximations of the theoretical probability) increases. So, the approximations become more accurate as the overall chance of significant fluctuations in the empirical frequencies decreases. As the number of samples approaches ∞, the empirical frequency will approach the theoretical probability.

Note: Flipping fair coin corresponds to N = 2. Rolling fair die corresponds to N = 6.

Problem 2

1)

The sample space can be represented like so:



The area of a circle is πr^2, so a quarter-circle with a radius of 1 is $\frac{\pi}{4}$ square units.

To find $P(X^2 + Y^2 \leq 1)$ we can then divide that area by the total area of the sample space, 1, to get $P(X^2 + Y^2 \leq 1) = \frac{\pi}{4}$.

2)

For large n, $X^2 + Y^2 \leq 1$ about 78.5% ($\frac{\pi}{4} \approx 0.785$) of the time since the long-run frequency is similar to the theoretical probability.

The frequency is $\frac{m}{n}$, which fluctuates around the theoretical probability:

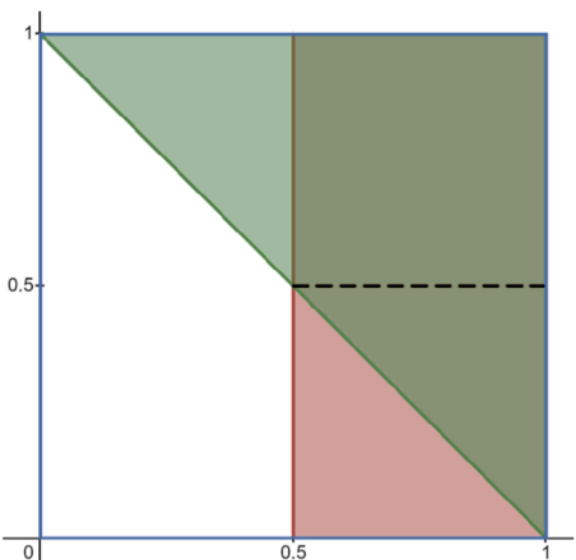
$$\frac{m}{n} \approx \frac{\pi}{4}$$

$$\pi \approx 4\frac{m}{n}$$

3)

$P(X \geq 0.5) = 0.5$, since that interval covers exactly half of the possible values of X.

To find $P(X \geq 0.5 | X + Y \geq 1)$ we can use a geometric representation of the sample space:



Green is the condition $X + Y \geq 1$, red is the condition $X \geq 0.5$

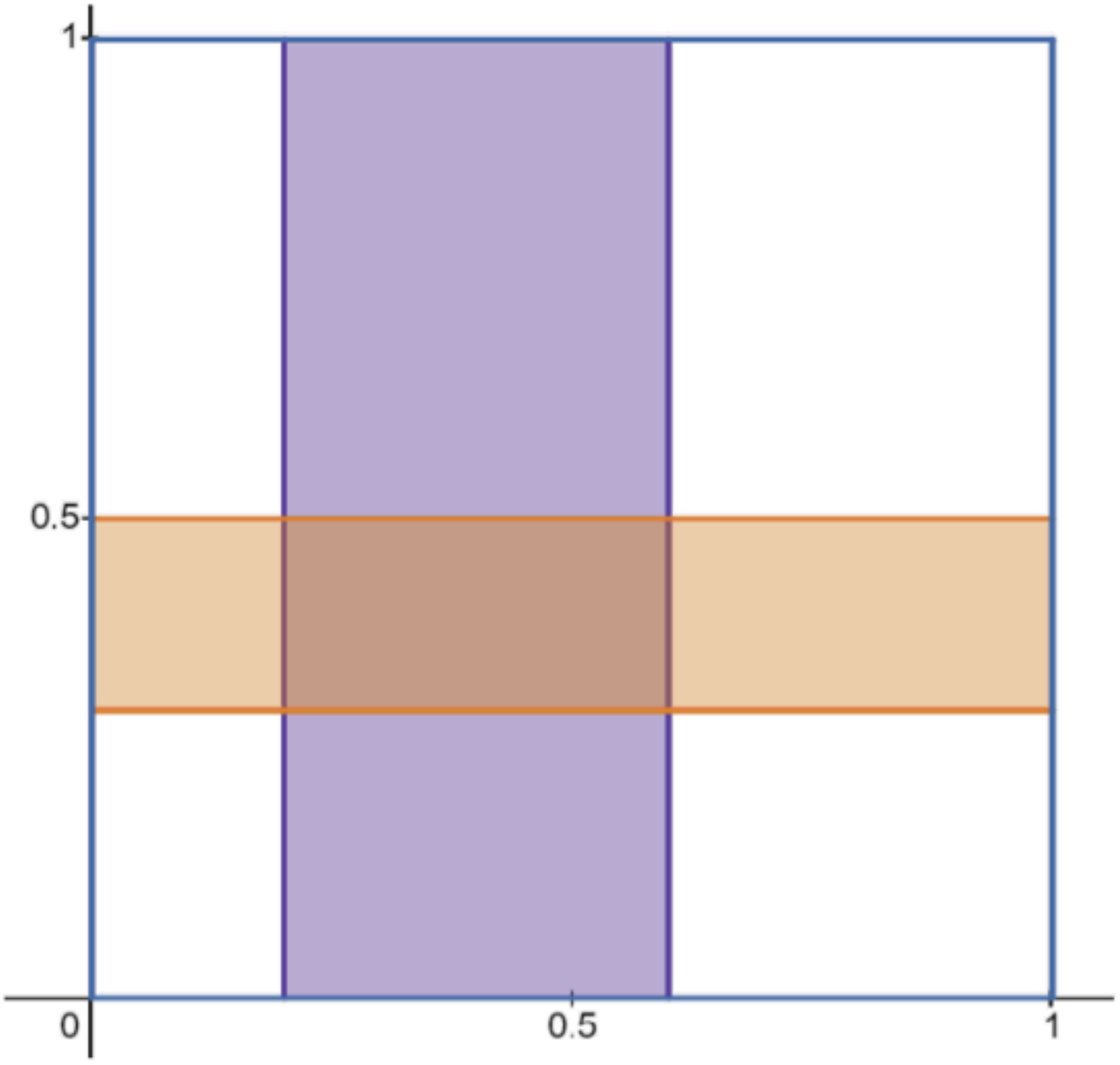
We want to calculate the joint area of red and green and then divide it by the total green area to find this probability since $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

$$P(X \geq 0.5 | X + Y \geq 1) = \frac{Area(top\ right\ red-green\ square) + Area(red-green\ triangle)}{Area(total\ green\ triangle)}$$

$$P(X \geq 0.5 | X + Y \geq 1) = \frac{0.5*0.5 + 0.5*0.5*0.5}{1*1*0.5} = \frac{0.25 + 0.125}{0.5} = 0.75$$

4)

We can represent the sample space as a square again:



A is purple, B is orange

$P(A \cap B) = Area(purple\ \&\ orange\ rectangle) = 0.4 * 0.2 = 0.08$

$P(A)P(B) = Area(purple\ rectangle) * Area(orange\ rectangle)$

$P(A)P(B) = (0.4 * 1) * (1 * 0.2) = 0.08 = P(A \cap B)$✔

Problem 3

In the simplest case, we assume that the length of the needle and the size of the gap between the parallel lines are equal to 1. Also assume that the lines are horizontal.

There are 2 variables that define every possibility for how a needle could end up:

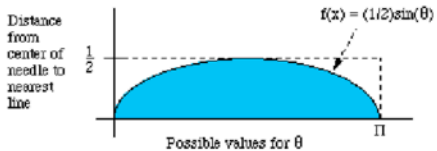
- D, the distance from the center of the needle to the nearest line
- θ, the angle the needle makes with the parallel lines

These variables are uniformly sampled from a respective interval:

- D is sampled from [0, 1/2] since this distance can at most be 1/2 (when the center of the needle is exactly in the middle of the two lines)
- θ is sampled from [0, π] since the max angle that can be made is 180 degrees

For the needle to cross a line, D must not exceed the vertical distance from the center of the needle to one of the tips. The entire vertical range of the needle is $sin\theta$ (using trigonometry), so the vertical distance from the center of the needle to one of the tips is $\frac{1}{2}sin\theta$. Taking this all together, P(needle crosses line) = P($D \leq \frac{1}{2}sin\theta$).

We can calculate this probability within a geometric representation of the sample space:



(Pulled from University of Illinois article)

To find the probability of the needle crossing the line we can divide the blue area by the total area:

$$\text{P(needle crosses line)} = \frac{\int_0^{\pi}(1/2)sin\theta}{(1/2)\pi} = \frac{2}{\pi}\frac{1}{2}\int_0^{\pi} sin\theta = -\frac{1}{\pi}(cos(\pi) - cos(0))$$

$$\text{P(needle crosses line)} = -\frac{1}{\pi}(-2) = \frac{2}{\pi}$$

From here, you can estimate π by generating samples of needle drops and approximating the theoretical probability (what we calculated) through the empirical probability (frequency that the needle crossed the lines in the experiment).

$$\frac{m}{n} \approx \frac{2}{\pi}$$

$$\pi \approx 2\frac{n}{m}$$

Lazzarini's result claiming to use this setup to estimate π is suspicious because the number he got is *exactly* equal to a fraction that is a good estimate for π. According to him, he dropped 3408 needles and 1808 of them crossed a line. This gives an estimate of $\pi \approx 3.141592\ldots$, exactly matching 6 digits. We expect the fluctuation of m/n with n = 3408 is much bigger than an accuracy that matches 6 digits.

STATS 100A HW2 Solutions

Problem 1: Coin Flipping

Suppose we flip a fair coin 100 times independently, so $X \sim \text{Binomial}(100, 1/2)$, where X is the number of heads.

(1) The probability of getting 50 heads:

$$P(X = 50) = \binom{100}{50} \left(\frac{1}{2}\right)^{50} \left(\frac{1}{2}\right)^{50} = \binom{100}{50} \left(\frac{1}{2}\right)^{100}$$

(2) The probability that the number of heads is between 40 and 60 (inclusive):

$$P(40 \leq X \leq 60) = \sum_{x=40}^{60} \binom{100}{x} \left(\frac{1}{2}\right)^{x} \left(\frac{1}{2}\right)^{100-x} = \sum_{x=40}^{60} \binom{100}{x} \left(\frac{1}{2}\right)^{100}$$

(3) Let $Z_i = 1$ if the i-th flip is a head, and $Z_i = 0$ if it is a tail. Then:

$$X = Z_1 + Z_2 + \cdots + Z_{100}$$

Here, X is the total number of heads, expressed as the sum of the indicator variables Z_i over all 100 flips.

Problem 2: Survey Sampling

We have a population of 10 million people, with 4 million having a college degree. Let $p = 4,000,000/10,000,000 = 0.4$.

(1) The probability that a randomly sampled person has a college degree:

$$P(\text{college degree}) = \frac{4,000,000}{10,000,000} = 0.4$$

(2) Sampling 100 times independently, $X \sim \text{Binomial}(100, 0.4)$, where X is the number of sampled people with a college degree.

$$P(X = 40) = \binom{100}{40} (0.4)^{40} (0.6)^{60}$$

$$P(35 \leq X \leq 45) = \sum_{x=35}^{45} \binom{100}{x} (0.4)^{x} (0.6)^{100-x}$$

(3) Sampling 10,000 times independently, $X \sim \text{Binomial}(10,000, 0.4)$. We need $P(X/n \in [0.39, 0.41])$, i.e., $P(0.39 \leq X/10,000 \leq 0.41)$, or equivalently $P(3900 \leq X \leq 4100)$:

$$P\left(\frac{X}{10,000} \in [0.39, 0.41]\right) = P(3900 \leq X \leq 4100) = \sum_{x=3900}^{4100} \binom{10,000}{x} (0.4)^{x} (0.6)^{10,000-x}$$

Problem 3: Monte Carlo

X and Y are generated independently from a uniform distribution over $[0, 1]$, so (X, Y) is a random point in the unit square $[0, 1]^2$.

(1) Calculate $P(X^2+Y^2 \leq 1)$, the probability that the point lies within the unit circle centered at the origin. This corresponds to the area of a quarter-circle with radius 1 (since $X, Y \in [0, 1]$) over the total area of the unit square (area = 1):

$$P(X^2 + Y^2 \leq 1) = \frac{\text{Area of quarter-circle}}{\text{Area of unit square}} = \frac{\frac{1}{4}\pi \cdot 1^2}{1} = \frac{\pi}{4}$$

(2) Repeating the experiment $n = 10,000$ times, let m be the number of times $X^2+Y^2 \leq 1$. Then $m \sim \text{Binomial}(10,000, \pi/4)$. We need $P(m/n \in [\pi/4 - 0.01, \pi/4 + 0.01])$, i.e., $P(\pi/4 - 0.01 \leq m/10,000 \leq \pi/4 + 0.01)$, or:

$$P\left(\frac{m}{10,000} \in \left[\frac{\pi}{4} - 0.01, \frac{\pi}{4} + 0.01\right]\right) = P\left(10,000\left(\frac{\pi}{4} - 0.01\right) \leq m \leq 10,000\left(\frac{\pi}{4} + 0.01\right)\right)$$

Approximating $\pi \approx 3.14159$, $\pi/4 \approx 0.7854$, so the interval is $[0.7754, 0.7954]$, and:

$$P(7754 \leq m \leq 7954) = \sum_{m=7754}^{7954} \binom{10,000}{m} \left(\frac{\pi}{4}\right)^m \left(1 - \frac{\pi}{4}\right)^{10,000-m}$$

Problem 4: Galton Board

The Galton board consists of a series of rows of pegs, with a ball dropped from the top that bounces left or right at each peg with equal probability ($p = 1/2$). The ball's final position corresponds to a bin at the bottom. Assume there are n rows of pegs, and the bins are numbered from 0 (leftmost) to n (rightmost), where the bin number represents the total number of right turns.

Let X be the number of right turns (or the bin number where the ball lands). At each of the n rows, the ball independently goes right with probability $1/2$ or left with probability $1/2$ (a left turn contributes 0 to X). Thus, X follows a binomial distribution:

$$X \sim \text{Binomial}(n, 1/2)$$

The probability of landing in bin x (i.e., taking exactly x right turns out of n moves) is:

$$P(X = x) = \binom{n}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{n-x} = \binom{n}{x}\left(\frac{1}{2}\right)^n$$

As n increases, the distribution of balls across the bins approximates a bell-shaped curve (due to the Central Limit Theorem), resembling a normal distribution centered at $n/2$.