CS114 Deep Dive 1

March 3, 2022

Part I

CS114. Assignment 1: Discrete Random Variables

Problem 1. Coin Spins (#Probability)

1. I agree with both. If you spin a coin 30 times, any sequence has the same probability of $\frac{1}{2^{30}}$, which is extremely unlikely. So, since we consider each sequence independently, there is no difference between them : no matter what the sequnce actually looks like, they are equally likely or unlikely. This logic holds if we assume that the coin is fair. However, if initially we have a biased Tails coin, it makes it more likely to get 30 tails in a row, because the probability of Tails is 1. Thus, the fact that we have 30 tails in a row does not tell us anything if the coin is biased or not, but if it were a biased coin, it would be a better explanation for the observed data since the probability of getting 30 tails would be 1. I agree with Yueh Han that the outcome would be extremely unlikely with fair coins, but I don't agree that they 'must be using trick coins' since obtaining such a sequence is still possible with a fair coin.

2. To find the posterior that coins are fair given 30 tails out of 30 trials, we could use Bayes theorem:

$P(F|T) = \frac{P(T|F) \cdot P(F)}{P(T)}$, where

F - the event that the coin is fair

T - the event of getting 30 Tails out of 30

$P(F)$ - prior probability that the coin is fair - p

$P(T|F)$ - likelihood, probability of getting 30 tails out of 30 given a fair coin.

$P(T)$- probability of getting 30 Tails out of 30

To find the likelihood, we should use the binomial distribution:

$P(T|F) = binomial(30, 0.5) = 1 \cdot 0.5^{30} = 0.5^{30}$

It means that each of the flips of a fair coin, there is a 0.5 chance of having either Tails or Heads. Since we are flipping it 30 times, and every flip is 0.5 chance of getting Tails, getting 30 Tails in a row is has a 0.5^{30} probability. We also multiply this value by the number of ways we can flip 30 tails out of 30 flips, which is just 1, giving us 0.5^{30} as a final answer. We can use binomial distribution becauseeach coin flip is independent, there aare only 2 outcomes : heads ot tails, and the probability is constant - 0.5.

To find $P(T)$, we should use the law of total probabilties. To implement that, we should find the complements of $P(F)$ and $P(T|F)$

$P(F^c) = 1 - p$, since $P(F) = p$, and the total probability is 1

$P(T|F^c) = 1$, since the complement of having a fair coin is double-tailed coin, and if we flip double-tailed coin, it will only give us Tails any number of times.

Now we can apply the law:

$P(T) = P(T|F) \cdot P(F) + P(T|F^c) \cdot P(F^c) = 0.5^{30} \cdot p + 1 \cdot (1 - p)$

Now we can substitute all values in the Bayes rule to find the posterior:

$P(F|T) = \frac{0.5^{30} \cdot p}{0.5^{30} \cdot p + 1 - p}$

3. To answer this question, we should iuse the expression for posterior probability we found in previous part, and solve for p when the expression is greater than 0.5:

$\frac{0.5^{30} \cdot p}{0.5^{30} \cdot p + 1 - p} > 0.5$

$0.5^{30}p > 0.5^{31}p - 0.5p + 0.5$

$0.5^{30}p - 0.5^{31}p + 0.5p > 0.5$

$p(0.5^{30} - 0.5^{31} + 0.5) > 0.5$

$p > \frac{0.5}{0.5^{30} - 0.5^{31} + 0.5}$

$p > 0.9999999991$ - the prior should be to make a fair-model coin more probable.

Problem 2. Medical test (#Probability)

1. Let TP be the event that Omer tests positive on all n tests. Then, using the hint, we can express the probabiltiies of him testing positive on all n tests given the disease and not:

$P(TP|D) = a_0^n$

$P(TP|D^c) = b_0^n$

The above likelihoods are given by a binomial distribution $binomial(n, a_0^n/b_0^n) = comb(n, n) \cdot a_0^n/b_0^n \cdot (1 - a_0^n/b_0^n)^{n-n} = a_0^n/b_0^n$ because each test result is independent, the probability of testing positive is constant, and there are only 2 outcomes. The slash sign in the above equation indicates that the distribution works for both a_0 and b_0 probabilities.

$P(D^c) = 1 - p$, since we know $P(D) = p$ and the total probability is 1

To find $P(TP)$ we can apply the law of total probabilities:

$P(TP) = P(TP|D) \cdot P(D) + P(P|D^c) \cdot P(D^c) = a_0^n \cdot p + b_0^n n \cdot (1-p)$

Now, to find posterior, we can apply Bayes theorem:

$P(D|TP) = \frac{P(TP|D) \cdot P(D)}{P(TP)} = \frac{a_0^n p}{a_0^n p + b_0^n (1-p)}$

2. Since we are given that having a gene guarantees testing positive on all n tests, we have

$P(TP|D, G) = 1$

$P(TP|D^c, G) = 1$

Given the probabilties we are given in case when he does not have the gene, we can find the probability of testing positive on all n tests given the conditions of no gene having or not the disease:

$P(TP|D, G^c) = a_1^n$

$P(TP|D^c, G^c) = b_1^n$

Given that having a gene or not is independent from having the disease or not, all probabilities of having a gene or not will be equal to 0.5 no matter if he has the diease or not:

$P(G|D) = P(G) = 0.5$

$P(G^c|D) = P(G^c) = 0.5$

$P(G|D^c) = P(G) = 0.5$

$P(G^c|D^c) = P(G^c) = 0.5$

Now again, we apply the law of total probabilties to find probability of being tested positive on all tests:

$P(TP|D) = P(TP|D, G) \cdot P(G|D) + P(TP|D, G^c) \cdot P(G^c|D) = 1 \cdot 0.5 + a_1^n \cdot 0.5 = \frac{1+a_1^n}{2}$

$P(TP|D^c) = P(TP|D^c, G) \cdot P(G|D^c) + P(TP|D^c, G^c) \cdot P(G^c|D^c) = 1 \cdot 0.5 + b_1^n \cdot 0.5 = \frac{1+b_1^n}{2}$

$P(TP) = P(TP|D) \cdot P(D) + P(TP|D^c) \cdot P(D^c) = \frac{p(1+a_1^n)}{2} + \frac{(1-p)(1+b_1^n)}{2}$

Finally, we can apply Bayes rule to find the posterior:

$P(D|TP) = \frac{P(TP|D) \cdot P(D)}{P(TP)} = \frac{p(1+a_1^n)}{2} \cdot \frac{2}{p(1+a_1^n)+(1-p)(1+b_1^n)} = \frac{p(1+a_1^n)}{p(1+a_1^n)+(1-p)(1+b_1^n)}$

3. As we already said, having a gene guarantees testing positive on all n tests. Thus,

$P(P|G) = 1$

Using Bayes rule, we find the new posterior:

$P(G|TP) = \frac{P(TP|G) \cdot P(G)}{P(TP)} = \frac{0.5}{p(1+a_1^n)+(1-p)(1+b_1^n)}$

Problem 3. Proofreading (#Distributions, #ModelSelection)

1. We can use indicator random variables to show the probabilties of different events. Let I_s be the indicator random variable of the event that Sho finds a typo and I_h - that Haruna finds a typo. So,

$I_s = 1$ if Sho finds a typo

$I_s = 0$ if Sho does not find a typo

$I_h = 1$ if Haruna finds a typo

$I_h = 0$ if Haruna does not find a typo

We can assign probabilities we are given to the events represeted by indicator random variables:

$P(I_s = 1) = p_1$

$P(I_s = 0) = 1 - p_1$

$P(I_h = 1) = p_2$

$P(I_h = 0) = 1 - p_2$

To find the probability of a typo being found by either of guys (union), we first need to find the probability that a typo is found by both of them (intersection):

$P(I_s = 1, I_h = 1) = P(I_s = 1) \cdot P(I_h = 1) = p_1 p_2$ - because both students read the book independently

Now we can find the union:

$P(I_s or I_h) = P(I_s = 1) + P(I_h = 1) - P(I_s = 1, I_h = 1) = p_1 + p_2 - p_1 p_2$. We subtract the intersection because by summing first 2 probabilties, we count the intersection twice.

To decide what distribution to use, we should think about the independence of events, probabilities, and outcomes. Here we have the situation with a binary outcome : either they find a typo (success) or don't (failure). The probability of finding each typo stays constant all the time. And the events are independent: finding one typo does not impact in any way finding another. Because of these conditions, we should use binomial distribution to represent a random variable X:

$X \sim Binomial(n, p_1 + p_2 - p_1 p_2)$, where X is a random variable representing the number of typos found by at least one of the students, n is the total number of typos, and $p_1 + p_2 - p_1 p_2$ is the probability of at least one of the students finding a typo.

2. We are asked to find the probability of Sho finding a certain number of typos given the total number of typos they found, which is $P(X_1 = x | X_1 + X_2 = t)$. To calculate it, we will use a formula for conditional probability :

$P(X_1 = x | X_1 + X_2 = t) = \frac{P(X_1 = x, X_1 + X_2 = t)}{P(X_1 + X_2 = t)}$

To find the intersection $P(X_1 = x, X_1 + X_2 = t)$, we can use the formula

$P(X_1 = x, X_1 + X_2 = t) = P(X_1 = x) P(X_1 + X_2 = t)$ because the events $X_1 = x$ and $X_1 + X_2 = t$ are independent. That is, the total number of typos found by both proofreaders tells us nothing about how many typos were found only by the first one, and X_1 and X_2 are also independent.

To find $P(X_1 = x)$ and $P(X_1 + X_2 = t)$, we use pmf of binomial distribution as in previous tasks because the events are independent, the probability is constant and there are only 2 outcomes: found typo or not. Thus, the overall formula looks like this:

$P(X_1 = x | X_1 + X_2 = t) = \frac{P(X_1 = x, X_1 + X_2 = t)}{P(X_1 + X_2 = t)} = \frac{P(X_1 = x) P(X_2 = t - x)}{P(X_1 + X_2 = t)} \frac{comb(n,x) \cdot p^x \cdot (1-p)^{n-x} \cdot comb(n, t-x) \cdot p^{t-x} \cdot (1-p)^{n-t+x}}{comb(2n, t) \cdot p^t \cdot (1-p)^{2n-t}} = c \frac{omb(n,x) \cdot comb(n, t-x)}{comb(2n, t)}$

The resultant equation resembles the pmf of hypergeometric distribution with the following parameters: $N = 2n$ - total number of words, $K = n$ - the number of typos found by Sho (the only one we are interested in), and $n = t$ - total number of typos found by both readers.

```
In [5]: import scipy.stats as sts
        import matplotlib.pyplot as plt
        import numpy as np

        # define null hypothesis as hypergeomtric distribution with corresponding parameters
        null = sts.hypergeom(200000, 100000, 613)

        #define the range of the plot and plot the distribution under null model
        x = np.arange(250, 370)
        plt.bar(x, null.pmf(x), label='Distribution Under Null Model')

        # plot chosen test statistic as a line
        plt.axvline(314, color='red', label='Real Test Statistic')
        plt.legend()
        plt.show()
```

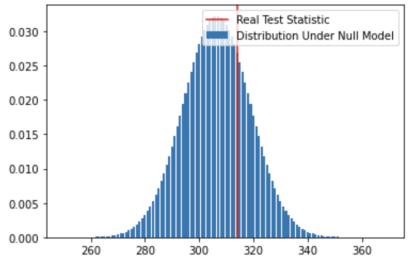

Figure 1: Distribution over test statistic under the null model

```
In [7]: #computes p-value
        p_value = 1 - null.cdf(314 - 1)
        print('p value:', p_value)

        p value: 0.28560476124170264
```

Figure 2: p-value

Thus,

$X_1|X_1 + X_2 = t \sim hypergeometric(N = 2n, K = n, n = t)$

3. To answer the question if Haruna finding more typos than Sho is just a coincidence, we take Sho's typos distribution as a null hypothesis : we would expect the number of typos found by Haruna to be withing the distribution of Sho's typos if we are wrong about how the data were generated. We take our null hypothesis as a hypergeometric distribution as defined in part 2 as the number of Sho's typos is conditional on the total number of typos:

Null hypothesis : $X_1|X_1 + X_2 = t \sim hypergeometric(N = 200000, K = 100000, n = 613)$, where 200000 is the total number of words, 100000 is the number of words only for one reader we are interested in (Sho), and 613 is the total number of typos found by both readers (299+314)

As a test statistic, we can use the number of typos found by Haruna - 314

Significance level will be standrad 0.05

To find the distribution over the test statistic given the null model we write a simulation in Python (Figure 1).

Now we find the corresponding p-value (Figure 2).

Since the p-value is greater than significance level : 0.29>0.05, we can not reject the null hypothesis and must admit that it might be a coincidence that Haruna found more typos than Sho.

Problem 4. Randomized response surveys (#ParameterEstimation)

First, we define the events:

P - a person gets a paper that says "I have used illegal drugs"

Y - a person says 'yes' to the question on the paper

D - a person actually has used illegal drugs

To define probabiltiies, we need to state an important assumption that participants don't lie : if they did take illigal drugs they will answer 'yes' to the paper "I have used illegal drugs"

$P(Y|P, D) = 1$, because if the person did drugs and is asked about it, they will always say the truth - 'yes'

$P(Y|P^c, D^c) = 1$, because if they did not do drugs and are asked about it, they will always tell the truth that they did not do it.

$P(Y|P, D^c) = 0$

$P(Y|P^c, D) = 0$

$P(P, D) = P(P)P(D) = pd$, because getting a paper with "I have used illegal drugs" is independent from having used illigal drugs

Following the same logic,

$P(P^c, D^c) = P(P^c)P(D^c) = (1 - p)(1 - d)$

To find P(Y), we apply the law of total probabilities:

$P(Y) = y = P(Y|P, D) \cdot P(P, D) + P(Y|P, D^c) \cdot P(P, D^c) + P(Y|P^c, D) \cdot P(P^c, D) + P(Y|P^c, D^c) \cdot P(P^c, D^c) = pd + (1 - p)(1 - d)$

2. Given the previous equation we found and our interest in d, the worst choice of p would be the one that would lead to the elimination of d from the equation. This is the case for $p = 0.5$:

$y = pd + (1 - p)(1 - d) = 0.5d + 0.5 - 0.5d = 0.5$

As we can see, half of people will say 'yes' given that half of the papers say 'I used illigal drugs' no matter what the proportion of drug users is since the equation does not depend on d in this case. Thus, we can not find d when $p = 0.5$.

3. Given the new conditions,

$P(Y|P, D) = 1$

$P(Y|P, D^c) = 0$

staey the same because here we are still dealing with the paper 'I have used illigal drugs'.

However,

$P(Y|P^c, D^c) = P(Y|P^c) = 0.25$

$P(Y|P^c, D) = P(Y|P^c) = 0.25$

because if we have the paper 'I was born in winter', it does not matter if the person did or did not do drugs: the person will always answer 'yes' if they are born in winter, and the proportion of such people is 0.25.

In [4]:
```python
import numpy as np
import matplotlib.pyplot as plt

n = 10000 # number of trials
participants = 100
use_drugs_number = [] # list that stores the number of drug users for each trial, given 20 'yes' answers

#generates trials
for i in range(n):
    answer_yes = 0
    use_drugs_truefalse = [] # list that stores drug users cases as True or False for each trial and participant

    #generates each participant
    for i in range(participants):
        #simulates people born in winter and not using relevant fraction
        if np.random.uniform() < 0.25:
            born_in_winter = True
        else:
            born_in_winter = False
        #simulates people who use drugs and not using relevant fraction
        if np.random.uniform() < 0.15:
            drug_user = True
        else:
            drug_user = False
        #simulates papers with question about drugs and about birth date using relevant fraction
        if np.random.uniform() < 0.5:
            drug_question = True
        else:
            drug_question = False

        use_drugs_truefalse.append(drug_user) #collects info on the usage of drugs

        #counts the number of total 'yes' answers
        if drug_question==True and drug_user==True or drug_question==False and born_in_winter==True:
            answer_yes += 1
    #calculate the number of drug users for the trial with 20 'yes' answers
    if answer_yes == 20:
        use_drugs_number.append(sum(use_drugs_truefalse))

plt.hist(use_drugs_number)
plt.title('Bayesian Posterior over Drug Users')
plt.ylabel('Frequency')
plt.xlabel('Fraction of drug users (out of 100)')
plt.show()
```

Figure 3: Python Simulation

Since getting of a certain paper is independent of using illigal drugs, we get

$P(P, D) = P(P)P(D) = pd$

$P(P^c, D^c) = P(P^c)P(D^c) = (1-p)(1-d)$

$P(P, D^c) = P(P) \cdot P(D^c) = p(1-d)$

$P(P^c, D) = P(P^c) \cdot P(D) = d(1-p)$

Applying the law of total probabilities, we get

$P(Y) = y = P(Y|P, D) \cdot P(P, D) + P(Y|P, D^c) \cdot P(P, D^c) + P(Y|P^c, D) \cdot P(P^c, D) + P(Y|P^c, D^c) \cdot P(P^c, D^c) = pd + 0.25d(1-p) + 0.25(1-p)(1-d) = 0.75dp + 0.25d + 0.25 - 0.25p - 0.25d + 0.25pd = pd - 0.25p + 0.25$

From here, we can isolate d :

$pd = y + 0.25p - 0.25$

$d = \frac{y+0.25p-0.25}{p}$

4. We are given that the number of 'yes' responses is 20 out of 100, which makes $P(Y) = y = 0.2$. We are also given that $p = 0.5$. Using the equation from part 3, we can find d:

$d = \frac{0.2+0.25 \cdot 0.5-0.25}{0.5} = 0.15$- probability that a participant used illegal drugs.

As we can see from the histogram (Figure 3 and 4), its mode is 15 drug users out of 100, which supports our numerical calculations.

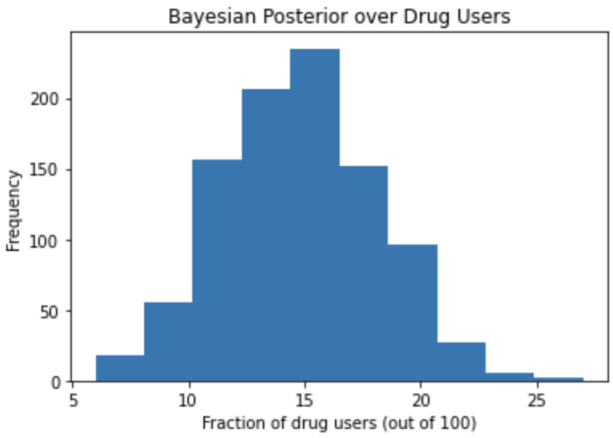

Figure 4: Bayesian Posterior histogram