Aditya Joshi

Data Science Interview Prep

Question 1 – Unfair Coin

This problem was asked by Facebook.

There is a fair coin (one side heads, one side tails) and an unfair coin (both sides tails). You pick one at random, flip it 5 times, and observe that it comes up as tails all five times. What is the chance that you are flipping the unfair coin?

Solution:

This problem can be solved using Bayes Rule. We are asked to calculate the following quantity: $P(\text{Unfair} \mid \text{TTTTT})$.

We have that

$$\begin{aligned}
P(\text{Unfair} \mid \text{TTTTT}) &= \frac{P(\text{TTTTT} \mid \text{Unfair}) \cdot P(\text{Unfair})}{P(\text{TTTTT})} \\
&= \frac{P(\text{TTTTT} \mid \text{Unfair}) \cdot P(\text{Unfair})}{P(\text{TTTTT} \mid \text{Unfair}) \cdot P(\text{Unfair}) + P(\text{TTTTT} \mid \text{Fair}) \cdot P(\text{Fair})} \\
&= \frac{1^5 \cdot 0.5}{1^5 \cdot 0.5 + 0.5^5 \cdot 0.5} \\
&= \frac{0.5}{0.5 + 0.015625} \\
&= \boxed{0.9697}
\end{aligned}$$

Question 2 – Flips until two heads

This problem was asked by Lyft.

What is the expected number of coin flips needed to get two consecutive heads?

Solution:

This is a slighly more complicated version of the classic problem – expected number of coin flips needed to get heads (which is simply the expected value of the geometric random variable with $p = 0.5$)

We can represent this sytem as a Markov chain as follows:

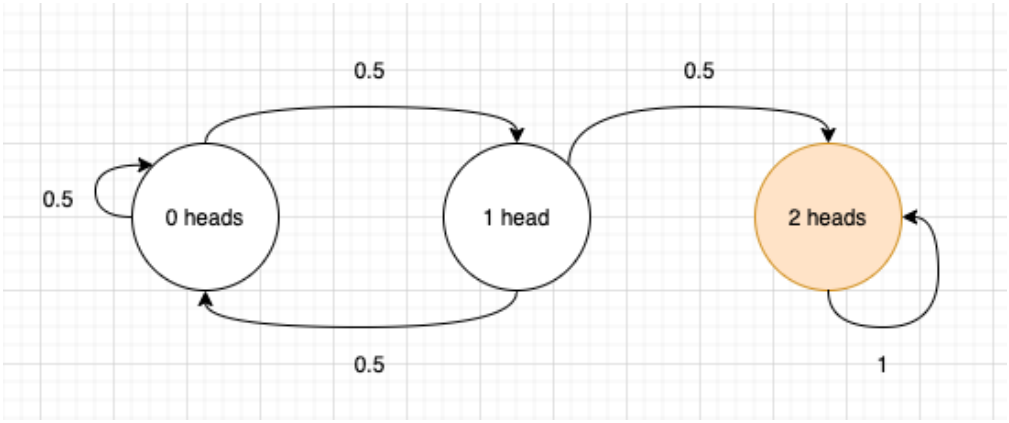


Figure 1: We start with 0 heads. With 0.5 probability, we see 1 head. After we have seen 1 head, if we see another head (which happens with $p = 0.5$), we are in the absorbing state 2. If we see a tails ($p = 0.5$), we go back to state 0.

If we are in state 2, the expected time to see two heads is $E[2] = 0$ as we have already seen two heads.

If we are in state 1, the expected time to see two heads is given by (using the law of total expectation):

$$E[1] = 1 + \frac{1}{2} \cdot E[0] + \frac{1}{2} \cdot E[2] \tag{1}$$

$$= 1 + \frac{1}{2} \cdot E[0] \tag{2}$$

If we are in state 0, the expected time to see two heads is given by:

$$E[0] = 1 + \frac{1}{2} \cdot E[0] + \frac{1}{2} \cdot E[1] \tag{3}$$

Substituting (2) in (3), we have:

$$E[0] = 1 + \frac{1}{2} \cdot E[0] + \frac{1}{2} \cdot \left(1 + \frac{1}{2} \cdot E[0]\right) \tag{4}$$

$$E[0] = \boxed{6} \tag{5}$$

Thus, if we are in state 0, that is when we start the experiment, the expected number of flips to see 2 heads is 6.

Question 3 – Drawing normally

This problem was asked by Quora.

You are drawing from a normally distributed random variable $X \sim \mathcal{N}(0, 1)$ once a day. What is the approximate expected number of days until you get a value of more than 2?

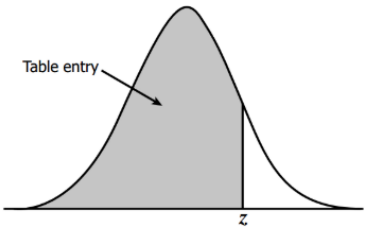
Solution

We can look at this problem as follows. Each day we carry out an experiment in which we draw from a standard unit normal. If the value sampled is greater than 2, then the experiment is successful. We want to know the average number of days in which we can expect a success.

The second part of the problem – average number of days in which we can expect a success sounds like a geometric random variable with parameter $Y \sim \text{Geom}(p)$ All we need to do is to find the value of parameter p and then the mean of the geometric random variable is $\frac{1}{p}$ which will give us the average number of days until we see the first "success" as we have defined it.

Now, to find p, consider the experiment that we perform every day. For it to be a "success", i.e., have a value more extreme than 2, it needs to be 2 standard deviations above the mean. In other words, the $z-$score is 2.0. We can look up the value for a $z-$score of 2.0, (from a $z-$table such as 2) which will give us the probability of getting a value less than 2.0. We can then subtract this from 1 to give us the probability of getting a value more extreme than 2.0.

The value here is 0.9772. Thus the probability we want is $1 - 0.9772 \approx 0.0228$. Thus, we have that $Y \sim \text{Geom}(0.0228)$. Thus, the average number of days that we have to wait are $\frac{1}{0.0228} \approx \boxed{43.86.}$



Find values on the right of the mean in this z-table. Table entries for z represent the area under the bell curve to the left of z. Positive scores in the Z-table correspond to the values which are greater than the mean.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Figure 2: We need to find the value for $z-$score equal to 2.0. Figure is from here.

Question 4 – Is this coin biased?

This problem was asked by Google.

A coin was flipped 1000 times, and 550 times it showed up heads. Do you think the coin is biased? Why or why not?

Solution

We could approach this the frequentist way, since the question demands a specific answer, that is, is the coin biased or not. We can formulate the following hypotheses:

$$H_0 : \text{Coin is not biased } (p = 0.5)$$
$$H_1 : \text{Coin is biased } (p \neq 0.5)$$

Let the random variable X denote the number of heads obtained where $X \sim \text{Binom}(p)$. The PMF looks like so:

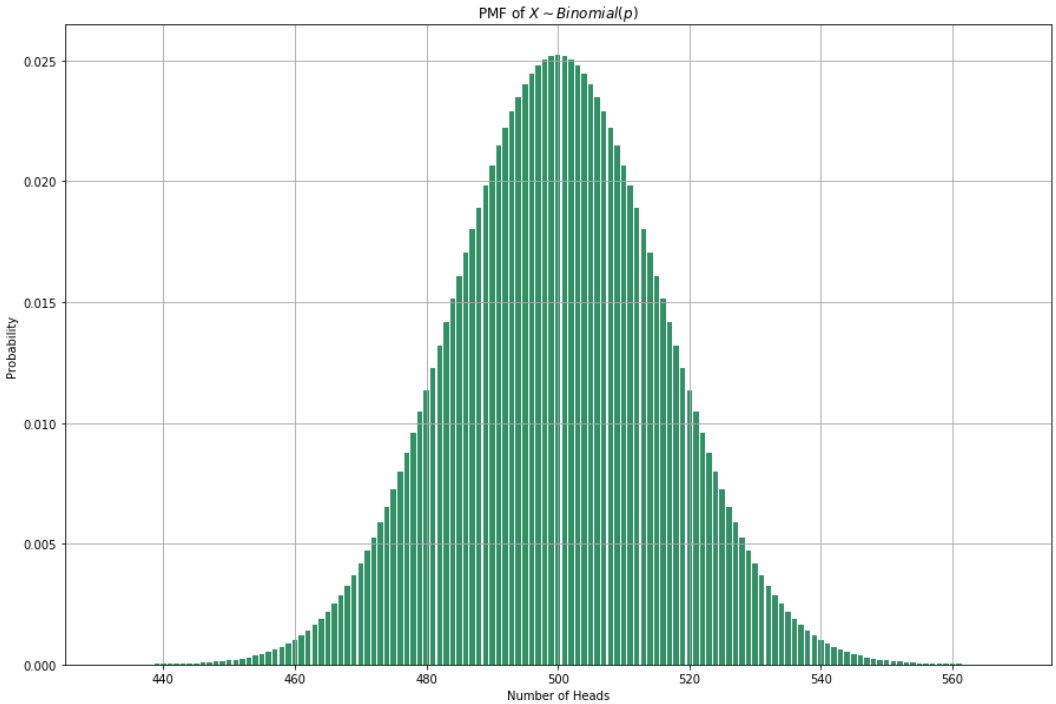


Figure 3: PMF of $X \sim \text{Binom}(0.5)$

We can formulate this as a one-sided test, that is, we want to find the value of $p(X \geq 550 \mid p = 0.5)$ This is the shaded area (in red) in 4.

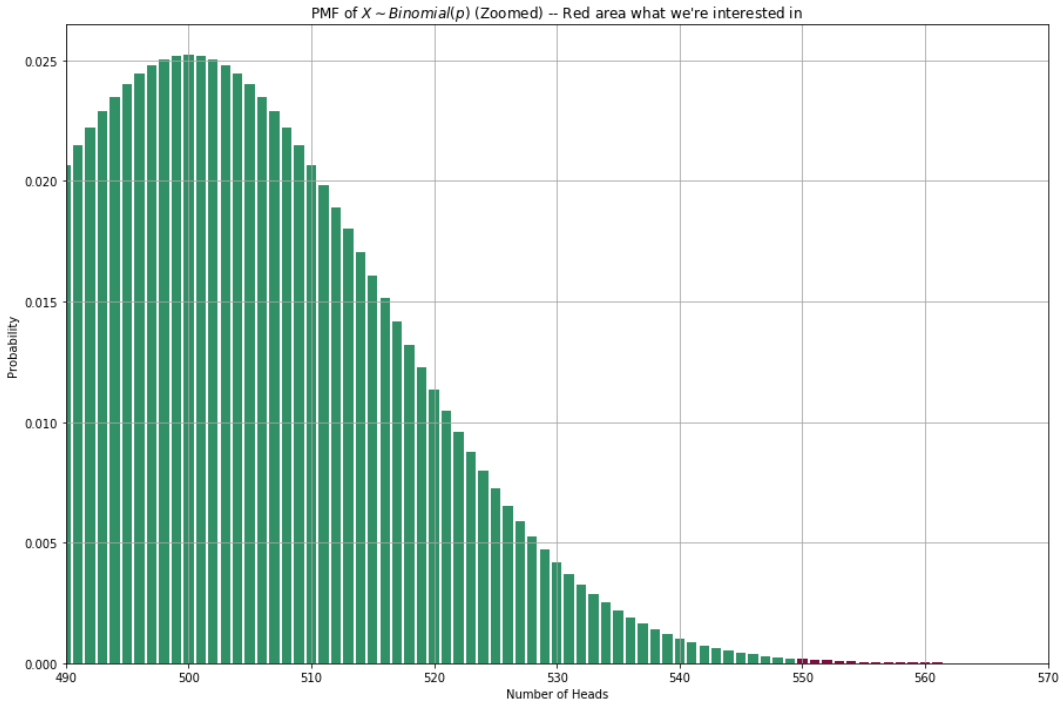


Figure 4: PMF of $X \sim \text{Binom}(0.5)$. We want to find the value of the red shaded area.

Now,

$$p(X \geq 550 \mid p = 0.5) = \sum_{i=550}^{1000} \binom{1000}{i} \cdot 0.5^{1000}$$
$$\approx 0.00086$$

For a reasonable threshold of $\alpha = 0.05$, since $0.00086 < \alpha = 0.05$, we can reject the null hypothesis and **conclude that the coin is biased.** Note that the value of 0.00086 was obtained computationally and we could also obtain this value analytically using the normal approximation to the binomial since our sample size is large.

To find the value using the normal approximation, we find the $z-$score of X. The standard deviation of the binomial distribution is $\sqrt{n \cdot p \cdot (1-p)} = \sqrt{1000 \cdot 0.5 \cdot 0.5} \approx 15.8113$ and thus the $z-$score is $\dfrac{550 - 500}{15.8113} \approx 3.162$. The probability is then $1 - 0.9992 \approx 0.0008$ (the value 0.9992 is obtained by looking up the value of 3.16 in a $z-$table like the one in 2) which is very close to the computational value that we obtained.

Question 5 – Rolls to see all sides

What is the expected number of rolls needed to see all 6 sides of a fair die?

Solution

During the first roll, we are guaranteed to see an unseen side. For the second roll, there is a probability of $\frac{5}{6}$ to see an unseen side. The expected number of rolls to see an unseen side is a geometric random variable $X \sim \text{Geom}\left(\frac{5}{6}\right)$ with a mean value of $\frac{6}{5}$. Similary, for the third unseen side, there is a probability of $\frac{4}{6}$ to see an unseen side. The expected number of rolls to see an unseen side is again a geometric random variable $X \sim \text{Geom}\left(\frac{4}{6}\right)$ with a mean value of $\frac{6}{4}$. This goes on for all the unseen sides until we've seen all of them which gives us:

$$\begin{aligned} E[Y] &= 1 + \frac{6}{5} + \frac{6}{4} + \cdots + \frac{6}{1} \\ &= \boxed{14.7} \end{aligned}$$

Thus, on average we will have to roll the die 14.7 times to see all the sides.

Question 6 – Picking between two dice games

This problem was asked by Facebook.

There are two games involving dice that you can play. In the first game, you roll two die at once and get the dollar amount equivalent to the product of the rolls. In the second game, you roll one die and get the dollar amount equivalent to the square of that value. Which has the higher expected value and why?

Solution

Consider the first game. Let the outcomes of the two rolls be represented by the random variables X and Y. The quantity to be computed here is $E[XY]$. Since the two die rolls are independent the expected value of the product of random variables is equal to the product of their expectation. Individually, X and Y are discrete random variables distributed as $X \sim \text{Uniform}(1, 6)$ and $Y \sim \text{Uniform}(1, 6)$.

$$\begin{aligned} E[XY] &= E[X] \cdot E[Y] \\ &= 3 \cdot 3 \\ &= 9 \end{aligned}$$

For the second game, let the outcome of the first die roll be denoted by the random variable Z which is also a discrete uniform random variable distributed as $Z \sim \text{Uniform}(1, 6)$. For a discrete uniform random variable with parameters a and b, the variance of Z is given by:

$$\begin{aligned} \text{Var}(Z) &= \frac{(b - a + 1)^2 - 1}{12} \\ &= \frac{(6 - 1 + 1)^2 - 1}{12} \\ &= \frac{35}{12} \\ &\approx 2.92 \end{aligned}$$

The quantity of interest here is $E[Z^2]$. We have the following relation:

$$\begin{aligned} \text{Var}(Z) &= E[Z^2] - (E[Z])^2 \\ E[Z^2] &= \text{Var}(Z) + (E[Z])^2 \\ &= 2.92 + 9 \\ &= 11.92 \end{aligned}$$

Thus, the expected value of the second game (11.92) is higher than that of the first game (9).

Question 7 – Fair odds from unfair coin

This problem was asked by Airbnb.

Say you are given an unfair coin, with an unknown bias towards heads or tails. How can you generate fair odds using this coin?

Solution

Let the coin be biased with a probability of heads equal to p. Thus, the probability of tails is $(1-p)$. Instead of flipping it once, consider flipping it twice. There are four possible outcomes:

Sequence	**Probability**
HH	p^2
HT	$p \cdot (1-p)$
TH	$(1-p) \cdot p$
TT	$(1-p)^2$

Table 1: Four outcomes with their corresponding probability

We use a form of **rejection sampling.** The probability of HT and TH is the same, equal to $p \cdot (1-p)$. Thus, we change our experiment to flip the coin twice instead of once. If the outcome is HH or TT, we repeat the experiment. Otherwise, we can arbitrarily denote HT to denote heads and TH to denote tails before the experiment and record the outcome.

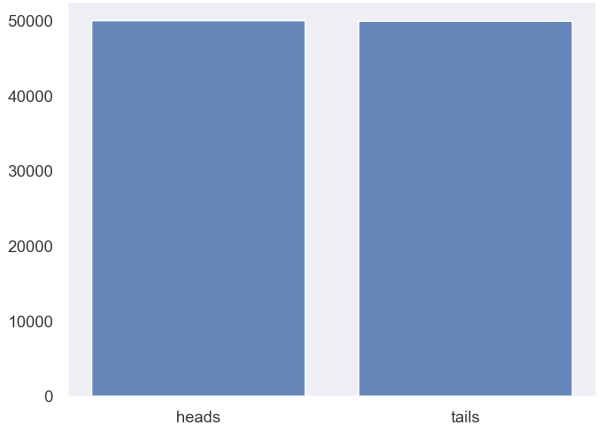


Figure 5: We ran a simulation of this technique with $p = 0.7$ and found 50036 heads and 49964 tails in 100000 experiments. These appear to be roughly equal.

Question 8 – Ant Collision

This problem was asked by Facebook.

Three ants are sitting at the corners of an equilateral triangle. Each ant randomly picks a direction and starts moving along the edge of the triangle. What is the probability that none of the ants collide? Now, what if it is k ants on all k corners of an equilateral polygon?

Solution

The ants will not collide if they all go in the same direction. They can either all go left or all go right. As each ant independently randomly picks a direction, the probability of this happening is $0.5^3 + 0.5^3 = 0.25$.

For k ants on k corners, the probability is $0.5^k + 0.5^k$.

Question 9 – Classification metrics

This problem was asked by Uber.

Say you need to produce a binary classifier for fraud detection. What metrics would you look at, how is each defined, and what is the interpretation of each one?

Solution

Consider the following table for classification:

		True	
		Fradulent	Authentic
Predicted	Fradulent	True Positive (TP)	False Positive (FP)
	Authentic	False Negative (FN)	True Negative (TN)

Table 2: Binary Confusion Matrix

True positives and true negatives are when we correctly identify frauds and geniune transactions respectively. From a business perspective, we would want to minimize False Negatives, or when we incorrectly classify a fradulent transaction as authentic. These could cost the company a lot of money. False positives typically would be less expensive (but still important, for example for the reputation of the company and to maintain a positive customer relationship), as these are instances when we incorrectly classify a geniune transaction as a fraud. In a realistic scenario, most of the transactions would be authentic and only a small number would be fradulent. As a consequence, we would expect a smaller number of False Positives, and we can have additional measures of safety like asking the user for additional documentation or adding a human in the loop to verify the transaction. Considering the above, the following metrics could be useful:

- **Accuracy:** The overall accuracy of the model, that is how often is the model making correct predictions (identifying both fradulent and authentic transactions). This is defined as $\frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$.
- **False Negative Rate:** This is defined as $\frac{\text{FN}}{\text{TP} + \text{FN}}$, that is, what proportion of fradulent transactions were (incorrectly) classified as authentic.
- **Recall (True Positive Rate):** This is defined as $\frac{\text{TP}}{\text{TP} + \text{FN}}$, that is, what proportion of fradulent transactions were correctly classified as fradulent.

- **Precision:** This is defined as $\frac{\text{TP}}{\text{TP} + \text{FP}}$, that is, what proportion of transactions that were classified as fradulent are actually fradulent.