MATH 323 - Tutorial 5 Questions

Tyrel Stokes

October 2020

1. Political Party Preference. Suppose you are Vice President of Surveys and Sampling for the Generic McGill Political Science Club. There are $N = 100$ members in the club. Suppose that there are 3 main parties denoted Party A, B, and C respectfully and that in the Generic McGill Political Science club there are $n_A = 45$, $n_B = 30$, and $n_c = 25$ people who prefer Party A,B, and C respectfully. As Vice President of Surveys and Sampling, you are task will randomly surveying $R = 10$ members of the club. Sampling is with replacement.

a) What is the probability that exactly 4 of them prefer party A?

b) Write down a pmf for the number of students surveyed which prefer party A. What is this distribution called?

c) Now suppose, you want to take this survey national, sampling now 100 students from Generic Political Science Clubs all across the country. The problem is that you do not know how many total members there are of generic political science clubs in Canada, but you can reasonably assume that it is a large number. Suppose that $p_a = 0.5$, $p_b = 0.3$, and $p_c = 0.2$ are the probabilities that Generic Political Science members prefer the 3 parties nationally.

Justify an expression for the probability that at least 40 of the 100 randomly surveyed prefer Party B. What is the key assumption for this expression?

Solutions:

a) Here, we can divide the club members into two disjoint groups, those that prefer part A, A, and those that prefer either B or C, BC. Let $X :=$ "The number of students sampled which belong to set A". There are 45 elements in group A and 55 elements in $B \cup C$.

Notice that we are sampling with replacement randomly and thus there are several ways to set up a counting space. Here we will consider the sample space to be all ways of sampling 10 different students from 100.

$$\begin{aligned} P(X = 4) &= \frac{|\{X = 4\}|}{|S|} \\ &= \frac{\binom{45}{4}\binom{55}{6}}{\binom{100}{10}} \end{aligned}$$

Where the numerator follows from the sausage rule. Similarly, we could use the random sampling with replacement to justify that X follows a hyper-geometric distribution with the above parameters.

b) Using the fact that we have a hyper-geometric as justified above,

$$P(X = x) = \frac{\binom{45}{x}\binom{55}{10-x}}{\binom{100}{10}}, \quad x \in \{0, 1, 2, \ldots, 10\}$$

Implicitly when x is not an integer from 0 to 10, the the value of the pmf is 0.

c) Here the key is that we need to justify the distribution. In reality, the sampling mechanism has not changed. That is, there is a fixed population and a fixed number of people which belong to now set B and $A \cup C$, but we do not know the number of total people and thus the number of people in each group. Sampling from such a population randomly with replacement in reality still produces a hypergeometric distribution. However, we know from class that if the overall population is large enough, then the random variable of interest, $Y :=$ "Number of people sampled which prefer party B", is approximately $Bin(100, 0.3)$. Why: Bernoulli trials which are approximately independent when the population is large.

$$\begin{aligned} P(Y \geq 40) &= 1 - P(Y \leq 39) \\ &= 1 - \sum_{y=0}^{39} \binom{100}{y} p_b^y (1 - p_b)^{100-y} \end{aligned}$$

2. Consider a simple lottery, where two digits are drawn (with replacement) from the set $\{0, 1, 2, \ldots 9\}$. Players can buy tickets of two unordered numbers. If a player owns a winning ticket, they get the prize. This lottery occurs every week and the numbers are drawn identically and independently week to week.

Consider a player who plays the same strategy every week. They always buy all the tickets corresponding to the same digit twice. Suppose they play every week until they win. Let $Y :=$ "Number of weeks until the player wins" .

a) What values can Y take with positive probability, i.e the set $\{y : P(Y = y) > 0\}$ (the support)?

Solution

$\{y : P(Y = y) > 0\} = \{y : y \in \mathcal{N}, y \geq 1\} = \{y : 1, 2, 3, \ldots, \ldots\}$
b) What is $P(Y \leq 2)$?

Solution

For clarity, I will define the sample space.

Let $\Omega_i = \{\text{Sample Space of the ith Week}\} = \{(i, j) : i, j \in \{0, 1, 2, \dots, 9\}\}$, then $\Omega = \Omega_1 \times \Omega_2 \times \Omega_3 \times \cdots \times \dots$

Let $A_i = \{\text{Winning in the ith week}\} = \{(j, j) \in \Omega_i : j \in \{0, 1, 2, \dots, 9\}\}$

$$\begin{aligned}
P(Y \leq 2) &= P(Y = 1) + P(Y = 2) \\
P(Y = 1) &= P(A_1) \\
&= \frac{|A_1|}{|\Omega_1|} \\
&= \frac{10}{10 \times 10} = 0.1 \\
P(Y = 2) &= P(A_1^c \cap A_2) \\
&\overset{ind.}{=} P(A_1^c)P(A_2) \\
&\overset{ident.}{=} (1 - P(A_1))P(A_1) \\
&= (0.9)(0.1) \\
&= .09 \\
P(Y \leq 2) &= .1 + 0.09 \\
&= .19
\end{aligned}$$

c) What is the p.m.f of Y? What kind of distribution does Y follow?

Solution

From part a), we know that we need to find $P(Y = y)$ for $y \in \{1, 2, \dots\}$

$$\begin{aligned}
P(Y = k) &= P(A_1^c \cap A_2^c \cap \cdots \cap A_{k-1}^c \cap A_k) \\
&\overset{ind.}{=} (\prod_{i=1}^{k-1} P(A_i^c))P(A_k) \\
&\overset{ident.}{=} (1 - P(A_1))^{k-1}P(A_1) \\
&= (0.9)^{k-1}(0.1)
\end{aligned}$$

Therefore

$$P(Y = y) = \begin{cases} (0.9)^{y-1}(0.1), & y \in \{1, 2, \dots\} \\ 0, & \text{else} \end{cases}$$

Notice that this is precisely a geometric distribution with $p = 0.1$.

d) What is $E[Y]$?

Solution

$$E[Y] = \frac{1}{p} = \frac{1}{.1} = 10$$

The above uses the fact that $Y \sim geometric(0.1)$. See your notes for the derivation.

e) The player changes strategy. If the week is an odd week, they only buy the odd duplicates $\{(i,i) : i \in \{1,3,5,7,9\}\}$ and if the week is even then they only buy the even duplicates. What is the pmf of Y now? What kind of distribution is this?

Solution

The sample space is still defined as Ω from part b.

Let

$$\begin{aligned} B_i &= \{\text{Winning with odd even doubles strategy in week i}\} \\ &= \begin{cases} \{(j,j) \in \Omega_i : j \in \{0,2,4,6,8\}\}, & i \text{ even} \\ \{(j,j) \in \Omega_i : j \in \{1,3,5,7,9\}\}, & i \text{ odd} \end{cases} \end{aligned}$$

Notice that $P(B_i) = P(B_{i+1}) = \frac{5}{100} = 0.05$. Since the lottery in each week is random the only thing that determines the probability is the size of the event $|B_i|$. Although, the way that the person wins depends on whether the week is odd or even, the probability that they win is constant across the weeks. In other words the events $B_1, B_2, \ldots,$ are identically distributed. Following our derivation in part c.

$$\begin{aligned} P(Y = y) &= P((\cap_{i=1}^{y-1} B_i^c) \cap B_y) \\ &\overset{ind.,iden.}{=} (P(B_1^c))^{y-1} P(B_1) \\ &= (.95)^{y-1}(0.05) \end{aligned}$$

$$P(Y = y) = \begin{cases} (0.95)^{y-1}(0.05), & y \in \{1,2,\ldots\} \\ 0, & \text{else} \end{cases}$$

This is also a geometric distribution, but with $p = 0.05$.

f) Challenge Question: Suppose we don't know what week it is. Let the week be a random variable I, with support in $\{1,2,\ldots,n\}$, where n is a fixed end date that is known to be even. Let W_i be a random variable equal to 1 if player wins on the ith week and 0 otherwise. Let W_I be the random variable that depends on unknown week I. Suppose that during the unknown week we do know that one of the winning numbers is odd, but we don't know the other one. What is the probability that the player wins in that week?

Assume that I is independent of W_i for all $i \in \{1, 2, \dots, n\}$. Similarly I is independent of whether or not one of the outcomes is a week is odd. We will also assume that I is independent of the event $W_i \cap \{\text{One of the winning numbers is odd}\}$ for all $i \in \{1, 2, \dots, n\}$

Solution

The difficult part of this question conceptually is that W_I depends on a random variable. The secret to dealing with unknown random variables is to condition on them. For the purposes of this question, the only thing that matters is whether I is odd or even, so we can directly condition on that event.

First define the event:

$$\begin{aligned} L_i &= \{\text{One of the winning numbers in week i is odd}\} \\ &= \{(i,j) \in \Omega_i : \{i \in \{1,3,5,7,9\}\} \cup \{j \in \{1,3,5,7,9\}\}\} \end{aligned}$$

$$\begin{aligned} P(W_I|L_I) &\overset{Cond.LTP}{=} P(W_I|L_I, \{I \text{ is odd}\})P(\{I \text{ is odd}\}|L_I) + P(W_I|L_I, \{I \text{ is even}\})P(\{I \text{ is even}\}|L_I) \\ &= P(W_{odd}|L_{odd}, \{I \text{ is odd}\})P(\{I \text{ is odd}\}|L_I) + P(W_{even}|L_{even}, \{I \text{ is even}\})P(\{I \text{ is even}\}|L_I) \end{aligned}$$

Once we condition on whether or not I is odd or even, we can replace the unknown subscripts with known subscripts indicating whether they are odd or even.

To simply, the randonly draw week I is independent of the outcomes in each week. Therefore, $P(\{I \text{ is even}\}|L_I) = P(\{I \text{ is even}\}) = \frac{1}{2}$ and $P(\{I \text{ is odd}\}|L_I) = P(\{I \text{ is odd}\}) = \frac{1}{2}$.

We also know that $P(W_{even}|L_{even}, \{I \text{ is even}\}) = 0$. This is because in even weeks player only wins if one of the double evens is a winning number. If there is even 1 odd number, it is not possible for the player to win.

Therefore, the only remaining probability to figure out is $P(W_{odd}|L_{odd}, \{I \text{ is odd}\})$. We know that we are going to have to find a clever way of using the independence assumptions we haven't used.

$$
\begin{aligned}
P(W_{odd}|L_{odd}, \{I \text{ is odd}\}) &= \frac{P(W_{odd} \cap L_{odd} \cap \{I \text{ is odd}\})}{P(L_{odd} \cap \{I \text{ is odd}\})} \\
&= \frac{P(W_{odd} \cap L_{odd}|\{I \text{ is odd}\})P(\{I \text{ is odd}\})}{P(L_{odd} \cap \{I \text{ is odd}\})} \\
&\overset{ind.}{=} \frac{P(W_{odd} \cap L_{odd})P(\{I \text{ is odd}\})}{P(L_{odd} \cap \{I \text{ is odd}\})} \\
&\overset{ind.}{=} \frac{P(W_{odd} \cap L_{odd})P(\{I \text{ is odd}\})}{P(L_{odd})P(\{I \text{ is odd}\})} \\
&= \frac{P(W_{odd} \cap L_{odd}))}{P(L_{odd})} \\
&= P(W_{odd}|L_{odd})
\end{aligned}
$$

Now we just have to figure out $P(W_{odd}|L_{odd})$ which is a much simpler problem.

$$
\begin{aligned}
P(W_{odd}|L_{odd}) &= P(W_1|L_1) && (1) \\
&= P(B_1|L_1) && (2) \\
&= \frac{P(B_1 \cap L_1)}{P(L_1)} && (3) \\
&\overset{subset}{=} \frac{P(B_1)}{P(L_1)} && (4) \\
&\overset{countingspace}{=} \frac{\frac{|B_1|}{|\Omega_1|}}{\frac{|L_1|}{|\Omega_1|}} && (5) \\
&= \frac{|B_1|}{|L_1|} && (6) \\
&= \frac{5}{55} && (7) \\
&= \frac{1}{11} && (8)
\end{aligned}
$$

To find the cardinality of L_1 you can break it into 3 cases. Either the first number is odd and the second is even ($5 \times 5 = 25$) or the second is odd and the first is even ($5 \times 5 = 25$) or both numbers are odd which is just even B_1.

Now that we have all the pieces, we can put it all together.

$$
\begin{aligned}
P(W_I|L_I) &= P(W_{odd}|L_{odd}, \{I \text{ is odd}\})P(\{I \text{ is odd}\}|L_I) + P(W_{even}|L_{even}, \{I \text{ is even}\})P(\{I \text{ is even}\}|L_I) && (9)\\
&= P(W_{odd}|L_{odd})P(\{I \text{ is odd}\}) + 0 * P(\{I \text{ is even}\}) && (10)\\
&= \frac{1}{11}\frac{1}{2} && (11)\\
&= \frac{1}{22} && (12)
\end{aligned}
$$

This is conceptually a very difficult question and certainly harder than what you would be expected to see in this course. The reason I put a hard question like this in the problem set is that it can push your understanding a little bit, sometimes expose concepts that are not quite clear. The other thing is that when you follow the solutions you can see that the difference between a "hard" question and an easier one is not so much about fancier tools, but developing the skill of breaking difficult questions into pieces that you can manage more easily. All of the individual manipulations are things that you are familiar with from the course. Don't be intimidated if this question was very difficult for you, but maybe as you follow the solution make note of the steps that were difficult.

3. Suppose that the number of goals scored by the home and away team in a hockey match are well-modelled by independent Poisson random variables, $Y_{home} \sim Poisson(\lambda_{home})$ and $Y_{away} \sim Poisson(\lambda_{away})$ respectively.

a) Find an expression for the probability that the home team scores strictly more than x goals in a game?

Solution
(Implicitly, x here is a positive integer)

$$
\begin{aligned}
P(Y_{home} > x) &= \sum_{i=x+1}^{\infty} P(Y_{home} = i)\\
&= \sum_{i=x+1}^{\infty} \frac{exp(-\lambda_{home})\lambda_{home}^{i}}{i!}
\end{aligned}
$$

b) Challenge: Thinking of part a as a hint and making use of the independence assumption. Find an expression for the probability that the home team wins. Remember the home team wins if and only if they score strictly more goals than their opponent.

$$
\begin{aligned}
P(Y_{home} > Y_{away}) &\overset{TLP}{=} \sum_{x=0}^{\infty} P(Y_{home} > Y_{away} | Y_{away} = x) P(Y_{away} = x) \\
&\overset{Independence}{=} \sum_{x=0}^{\infty} P(Y_{home} > x) P(Y_{away} = x) \\
&= \sum_{x=0}^{\infty} (\sum_{i=x+1}^{\infty} \frac{exp(-\lambda_{home})\lambda_{home}^{i}}{i!}) P(Y_{away} = x) \\
&= \sum_{x=0}^{\infty} (\sum_{i=x+1}^{\infty} \frac{exp(-\lambda_{home})\lambda_{home}^{i}}{i!}) \frac{exp(-\lambda_{away})\lambda_{away}^{x}}{!x}
\end{aligned}
$$

The key to this question is to use the values that Y_{away} can take as a partition and then condition on the amount of goals the away team scores. Then by indepedence we are able to simplify the conditional part of the expression and plug in what we had in part a. This is super convenient and makes our life much easier in deriving an expression for the probability we want. In real life, it is not always realistic to assume the number of goals scored by the home and away team are independent of each other. If you are familiar with sports, it might be worth thinking about why that might be.