Stat S-110 Homework 2 Solutions, Summer 2015

The following problems are from Chapters 2-3 of the book.

1. (BH 2.41) You are the contestant on the Monty Hall show. Monty is trying out a new version of his game, with rules as follows. You get to choose one of three doors. One door has a car behind it, another has a computer, and the other door has a goat (with all permutations equally likely). Monty, who knows which prize is behind each door, will open a door (but not the one you chose) and then let you choose whether to switch from your current choice to the other unopened door.

Assume that you prefer the car to the computer, the computer to the goat, and (by transitivity) the car to the goat.

(a) Suppose for this part only that Monty always opens the door that reveals your less preferred prize out of the two alternatives, e.g., if he is faced with the choice between revealing the goat or the computer, he will reveal the goat. Monty opens a door, revealing a goat (this is again for this part only). Given this information, should you switch? If you do switch, what is your probability of success in getting the car?

(b) Now suppose that Monty reveals your less preferred prize with probability p, and your more preferred prize with probability $q = 1 - p$. Monty opens a door, revealing a computer. Given this information, should you switch (your answer can depend on p)? If you do switch, what is your probability of success in getting the car (in terms of p)?

Solution:

(a) Let C be the event that the car is behind the door you originally chosen, and let M_{goat} be the event that Monty reveals a goat when he opens a door. By Bayes' rule,

$$P(C|M_{\text{goat}}) = \frac{P(M_{\text{goat}}|C)P(C)}{P(M_{\text{goat}})} = \frac{(1)(1/3)}{2/3} = 1/2,$$

where the denominator comes from the fact that the two doors other than your initial choice are equally likely to have $\{\text{car}, \text{computer}\}, \{\text{computer}, \text{goat}\}$, or $\{\text{car}, \text{goat}\}$, and only in the first of these cases will Monty not reveal a goat.

So you should be indifferent between switching and not switching; either way, your conditional probability of getting the car is $1/2$. (Note though that the *unconditional* probability that switching would get you the car, before Monty revealed the goat, is $2/3$ since you will succeed by switching if and only if your initial door does not have the car.)

(b) Let C, R, G be the events that the car is behind the door you originally chosen is a car, computer, goat, respectively, and let M_{comp} be the event that Monty reveals a computer when he opens a door. By Bayes' rule and LOTP,

$$P(C|M_{\text{comp}}) = \frac{P(M_{\text{comp}}|C)P(C)}{P(M_{\text{comp}})} = \frac{P(M_{\text{comp}}|C)P(C)}{P(M_{\text{comp}}|C)P(C) + P(M_{\text{comp}}|R)P(R) + P(M_{\text{comp}}|G)P(G)}.$$

We have $P(M_{\text{comp}}|C) = q, P(M_{\text{comp}}|R) = 0, P(M_{\text{comp}}|G) = p$, so

$$P(C|M_{\text{comp}}) = \frac{q/3}{q/3 + 0 + p/3} = q.$$

Thus, your conditional probability of success if you follow the switching strategy is p. For $p < 1/2$, you should not switch, for $p = 1/2$, you should be indifferent about switching, and for $p > 1/2$, you should switch.

2. (BH 2.45) A gambler repeatedly plays a game where in each round, he wins a dollar with probability 1/3 and loses a dollar with probability 2/3. His strategy is "quit when he is ahead by \$2", though some suspect he is a gambling addict anyway. Suppose that he starts with a million dollars. Show that the probability that he'll ever be ahead by \$2 is less than 1/4.

Solution: This is a special case of the gambler's ruin problem. Let A_1 be the event that he is successful on the first play and let W be the event that he is ever ahead by \$2 before being ruined. Then by the law of total probability, we have

$$P(W) = P(W|A_1)P(A_1) + P(W|A_1^c)P(A_1^c).$$

Let a_i be the probability that the gambler achieves a profit of \$2 before being ruined, starting with a fortune of \$$i$. For our setup, $P(W) = a_i$, $P(W|A_1) = a_{i+1}$ and $P(W|A_1^c) = a_{i-1}$. Therefore,

$$a_i = a_{i+1}/3 + 2a_{i-1}/3,$$

with boundary conditions $a_0 = 0$ and $a_{i+2} = 1$. We can then solve this difference equation for a_i (directly or using the result of the gambler's ruin problem):

$$a_i = \frac{2^i - 1}{2^{2+i} - 1}.$$

This is always less than 1/4 since $\frac{2^i-1}{2^{2+i}-1} < \frac{1}{4}$ is equivalent to $4(2^i - 1) < 2^{2+i} - 1$, which is equivalent to the true statement $2^{2+i} - 4 < 2^{2+i} - 1$.

3. (BH 2.51)

(a) There are two crimson jars (labeled C_1 and C_2) and two mauve jars (labeled M_1 and M_2). Each jar contains a mixture of green gummi bears and red gummi bears. Show by example that it is possible that C_1 has a much higher percentage of green gummi bears than M_1, and C_2 has a much higher percentage of green gummi bears than M_2, yet if the contents of C_1 and C_2 are merged into a new jar and likewise for M_1 and M_2, then the combination of C_1 and C_2 has a lower percentage of green gummi bears than the combination of M_1 and M_2.

(b) Explain how (a) relates to Simpson's paradox, both intuitively and by explicitly defining events A, B, C as in the statement of Simpson's paradox.

Solution:

(a) As an example, let C_1 have 9 green, 1 red; M_1 have 50 green, 50 red; C_2 have 30 green, 70 red; M_2 have 1 green, 9 red.

(b) This is a form of Simpson's paradox since which jar color is more likely to provide a green gummy bear flips depending on whether the jars get aggregated. To match this example up to the notation used in the statement of Simpson's paradox, let A be the event that a red gummi bear is chosen in the random draw, B be the event that it is drawn from a crimson jar, and C be the event that it is drawn from a jar with index 1. With the numbers from the solution to (a), $P(C|B) = 1/11$ is much less than $P(C|B^c) = 10/11$, which enables us to have $P(A|B) < P(A|B^c)$ even though the inequalities go the other way when we also condition on C or on C^c.

4. (BH 2.54) Fred decides to take a series of n tests, to diagnose whether he has a certain disease (any individual test is not perfectly reliable, so he hopes to reduce his uncertainty by taking multiple tests). Let D be the event that he has the disease, $p = P(D)$ be the prior probability that he has the disease, and $q = 1 - p$. Let T_j be the event that he tests positive on the jth test.

(a) Assume for this part that the test results are conditionally independent given Fred's disease status. Let $a = P(T_j|D)$ and $b = P(T_j|D^c)$, where a and b don't depend on j. Find the posterior probability that Fred has the disease, given that he tests positive on all n of the n tests.

(b) Suppose that Fred tests positive on all n tests. However, some people have a certain gene that makes them *always* test positive. Let G be the event that Fred has the gene. Assume that $P(G) = 1/2$ and that D and G are independent. If Fred does *not* have the gene, then the test results are conditionally independent given his disease status. Let $a_0 = P(T_j|D, G^c)$ and $b_0 = P(T_j|D^c, G^c)$, where a_0 and b_0 don't depend on j. Find the posterior probability that Fred has the disease, given that he tests positive on all n of the tests.

Solution:

(a) Let $T = T_1 \cap \cdots \cap T_n$ be the event that Fred tests positive on all the tests. By Bayes' rule and LOTP,

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)} = \frac{pa^n}{pa^n + qb^n}.$$

(b) Let T be the event that Fred tests positive on all n tests. Then

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)} = \frac{pP(T|D)}{pP(T|D) + qP(T|D^c)}.$$

Conditioning on whether or not he has the gene, we have

$$P(T|D) = P(T|D,G)P(G|D) + P(T|D,G^c)P(G^c|D) = \frac{1}{2} + \frac{a_0^n}{2},$$
$$P(T|D^c) = P(T|D^c,G)P(G|D^c) + P(T|D^c,G^c)P(G^c|D^c) = \frac{1}{2} + \frac{b_0^n}{2}.$$

Thus,

$$P(D|T) = \frac{p(1+a_0^n)}{p(1+a_0^n) + q(1+b_0^n)}.$$

5. (BH 2.55) A certain hereditary disease can be passed from a mother to her children. Given that the mother has the disease, her children independently will have it with probability 1/2. Given that she doesn't have the disease, her children won't have it either. A certain mother, who has probability 1/3 of having the disease, has two children.

(a) Find the probability that neither child has the disease.

(b) Is whether the elder child has the disease independent of whether the younger child has the disease? Explain.

(c) The elder child is found not to have the disease. A week later, the younger child is also found not to have the disease. Given this information, find the probability that the mother has the disease.

Solution:

(a) Let M, A, B be the events that the mother, elder child, and younger child have the disease (respectively). Then

$$P(A^c, B^c) = P(A^c, B^c|M)P(M) + P(A^c, B^c|M^c)P(M^c) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{3} + \frac{2}{3} = \frac{3}{4}.$$

(b) These events are conditionally independent given the disease status of the mother, but they are not independent. Knowing whether the elder child has the disease gives information about whether the mother has the disease, which in turn gives information about whether the younger child has the disease.

(c) By Bayes' rule,

$$P(M|A^c, B^c) = \frac{P(A^c, B^c|M)P(M)}{P(A^c, B^c)} = \frac{\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{3}}{\frac{3}{4}} = \frac{1}{9}.$$

Alternatively, we can do the conditioning in two steps: first condition on A^c, giving

$$P(M|A^c) = \frac{P(A^c|M)P(M)}{P(A^c)} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + \frac{2}{3}} = \frac{1}{5}.$$

Then do further conditioning on B^c, giving

$$P(M|A^c, B^c) = \frac{P(B^c|M, A^c)P(M|A^c)}{P(B^c|A^c)} = \frac{\frac{1}{2} \cdot \frac{1}{5}}{\frac{1}{2} \cdot \frac{1}{5} + \frac{4}{5}} = \frac{1}{9},$$

which agrees with the result of the one-step method.

6. (BH 3.2)

(a) Independent Bernoulli trials are performed, with probability 1/2 of success, until there has been at least one success. Find the PMF of the number of trials performed.

(b) Independent Bernoulli trials are performed, with probability 1/2 of success, until there has been at least one success and at least one failure. Find the PMF of the number of trials performed.

Solution:

(a) Let X be the number of trials (including the success). Then $X = n$ says that there were $n - 1$ failures followed by a success, so

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

for $n = 1, 2, \ldots$. (This is the *First Success* distribution, a Geometric shifted to start at 1.)

(b) Let Y be the number of trials performed. The support of Y is $\{2, 3, \ldots\}$. Let A be the event that the first trial is a success. The PMF of Y is

$$P(Y = n) = P(Y = n|A)P(A) + P(Y = n|A^c)P(A^c) = 2\left(\frac{1}{2}\right)^{n-1}\frac{1}{2} = \left(\frac{1}{2}\right)^{n-1},$$

for $n = 2, 3, \ldots$.

7. (BH 3.16) Let $X \sim \text{DUnif}(C)$, and B be a nonempty subset of C. Find the conditional distribution of X, given that X is in B.

Solution: The conditional PMF of X is

$$P(X = x|X \in B) = \frac{P(X \in B|X = x)P(X = x)}{P(X \in B)} = \begin{cases} \frac{1}{|B|} \text{ if } x \in B \\ 0 \text{ if } x \notin B, \end{cases}$$

since $P(X \in B) = |B|/|C|$ and for $x \in B$, $P(X \in B|X = x) = 1, P(X = x) = 1/|C|$. So the conditional distribution of X, given that X is in B, is Discrete Uniform over B. This makes sense intuitively since the values in B were equally likely before we knew that X was in B; just learning that X is in B rules out values not in B but should not result in some values in B being more likely than others.

8. (BH 3.20) Suppose that a lottery ticket has probability p of being a winning ticket, independently of other tickets. A gambler buys 3 tickets, hoping this will triple the chance of having at least one winning ticket.

(a) What is the distribution of how many of the 3 tickets are winning tickets?

(b) Show that the probability that at least 1 of the 3 tickets is winning is $3p - 3p^2 + p^3$, in two different ways: by using inclusion-exclusion, and by taking the complement of the desired event and then using the PMF of a certain named distribution.

(c) Show that the gambler's chances of having at least one winning ticket do not quite triple (compared with buying only one ticket), but that they do *approximately* triple if p is small.

Solution:

(a) By the story of the Binomial, the distribution is $\text{Bin}(3, p)$.

(b) Let A_i be the event that the ith ticket wins, for $i = 1, 2, 3$. By inclusion-exclusion and symmetry, we have

$$P(A_1 \cup A_2 \cup A_3) = 3P(A_1) - \binom{3}{2}P(A_1 \cap A_2) + P(A_1 \cap A_2 \cap A_3) = 3p - 3p^2 + p^3.$$

The Binomial PMF yields the same result: for $X \sim \text{Bin}(3, p)$,

$$P(X \geq 1) = 1 - P(X = 0) = 1 - (1-p)^3 = 1 - (1 - 3p + 3p^2 - p^3) = 3p - 3p^2 + p^3.$$

(c) For any $p \in (0,1)$, we have $p^3 < 3p^2$ (since $p < 3$), so $3p - 3p^2 + p^3 < 3p$. So the probability does not triple. But if p is small, then $3p^2$ and p^3 are *very* small, and then $3p - 3p^2 + p^3 \approx 3p$.

9. (BH 3.22) There are two coins, one with probability p_1 of Heads and the other with probability p_2 of Heads. One of the coins is randomly chosen (with equal probabilities for the two coins). It is then flipped $n \geq 2$ times. Let X be the number of times it lands Heads.

(a) Find the PMF of X.

(b) What is the distribution of X if $p_1 = p_2$?

(c) Give an intuitive explanation of why X is *not* Binomial for $p_1 \neq p_2$ (its distribution is called a *mixture* of two Binomials). You can assume that n is large for your explanation, so that the frequentist interpretation of probability can be applied.

Solution:

(a) By LOTP, conditioning on which coin is chosen, we have

$$P(X = k) = \frac{1}{2}\binom{n}{k}p_1^k(1-p_1)^{n-k} + \frac{1}{2}\binom{n}{k}p_2^k(1-p_2)^{n-k},$$

for $k = 0, 1, \dots, n$.

(b) For $p_1 = p_2$, the above expression reduces to the $\text{Bin}(n, p_1)$ PMF.

(c) A mixture of two Binomials is *not* Binomial (except in the degenerate case $p_1 = p_2$). Marginally, each toss has probability $(p_1 + p_2)/2$ of landing Heads, but the tosses are *not* independent since earlier tosses give information about which coin was chosen, which in turn gives information about later tosses.

Let n be large, and imagine repeating the entire experiment many times (each repetition consists of choosing a random coin and flipping it n times). We would expect to see *either* approximately np_1 Heads about half the time, and approximately np_2 Heads about half the time. In contrast, with a $\text{Bin}(n, p)$ distribution we would expect to see approximately np Heads; no fixed choice of p can create the behavior described above.

10. (BH 3.23) There are n people eligible to vote in a certain election. Voting requires registration. Decisions are made independently. Each of the n people will

register with probability p_1. Given that a person registers, he or she will vote with probability p_2. Given that a person votes, he or she will vote for Kodos (who is one of the candidates) with probability p_3. What is the distribution of the number of votes for Kodos (give the PMF, fully simplified, or the name of the distribution, including its parameters)?

Solution: Let X be the number of votes for Kodos. By the story of the Binomial, $X \sim \text{Bin}(n, p_1p_2p_3)$. The PMF is $P(X = k) = \binom{n}{k}p^kq^{n-k}$ for $k \in \{0, 1, \ldots, n\}$, with $p = p_1p_2p_3$ and $q = 1 - p$.

11. (BH 3.24) Let X be the number of Heads in 10 fair coin tosses.

(a) Find the conditional PMF of X, given that the first two tosses both land Heads.

(b) Find the conditional PMF of X, given that at least two tosses land Heads.

Solution:

(a) Let X_2 and X_8 be the number of Heads in the first 2 and last 8 tosses, respectively. Then $X = X_2 + X_8$, where $X_2 \sim \text{Bin}(2, 1/2)$ and $X_8 \sim \text{Bin}(8, 1/2)$. The conditional PMF of X given $X_2 = 2$ is

$$\begin{aligned}
P(X = k|X_2 = 2) &= P(X_2 + X_8 = k|X_2 = 2) \\
&= P(X_8 = k - 2|X_2 = 2) \\
&= P(X_8 = k - 2) \\
&= \binom{8}{k-2}\left(\frac{1}{2}\right)^{k-2}\left(\frac{1}{2}\right)^{8-(k-2)} \\
&= \frac{1}{256}\binom{8}{k-2},
\end{aligned}$$

for $k = 2, 3, \ldots, 10$, and $P(X = k|X_2 = 2) = 0$ for all other values of k.

(b) The conditional PMF of X given $X \geq 2$ is

$$\begin{aligned}
P(X = k|X \geq 2) &= \frac{P(X = k, X \geq 2)}{P(X \geq 2)} \\
&= \frac{P(X = k)}{1 - P(X = 0) - P(X = 1)} \\
&= \frac{\binom{10}{k}\left(\frac{1}{2}\right)^{10}}{1 - \left(\frac{1}{2}\right)^{10} - 10\left(\frac{1}{2}\right)^{10}} \\
&= \frac{1}{1013}\binom{10}{k},
\end{aligned}$$

for $k = 2, 3, \ldots, 10$, and $P(X = k | X \geq 2) = 0$ for all other values of k.

12. (BH 3.34) There are n students at a certain school, of whom $X \sim \text{Bin}(n, p)$ are Statistics majors. A simple random sample of size m is drawn ("simple random sample" means sampling without replacement, with all subsets of the given size equally likely).

(a) Find the PMF of the number of Statistics majors in the sample, using the law of total probability (don't forget to say what the support is). You can leave your answer as a sum (though with some algebra it can be simplified, by writing the binomial coefficients in terms of factorials and using the binomial theorem).

(b) Give a story proof derivation of the distribution of the number of Statistics majors in the sample; simplify fully.

Hint: Does it matter whether the students declare their majors before or after the random sample is drawn?

Solution:

(a) Let Y be the number of Statistics majors in the sample. The support is $0, 1 \ldots, m$. Let $q = 1 - p$. By LOTP,

$$\begin{aligned} P(Y = y) &= \sum_{k=0}^{n} P(Y = y | X = k) P(X = k) \\ &= \frac{1}{\binom{n}{m}} \sum_{k=0}^{n} \binom{k}{y} \binom{n-k}{m-y} \binom{n}{k} p^k q^{n-k}, \end{aligned}$$

for $0 \leq y \leq m$ (note that any term with $y > k$ or $m - y > n - k$ is 0).

We can simplify the sum algebraically as follows (not required for this part). Despite all the exclamation points, we are much more enthusiastic about the method in (b)!

Writing the binomial coefficients in terms of factorials, we have

$$
\begin{aligned}
P(Y=y) &= \frac{m!(n-m)!}{n!}\sum_{k=y}^{n-m+y}\frac{k!(n-k)!n!}{y!(k-y)!(m-y)!(n-m-k+y)!k!(n-k)!}p^k q^{n-k}\\
&= \frac{m!}{(m-y)!y!}p^y q^{m-y}\sum_{k=y}^{n-m+y}\frac{(n-m)!}{(k-y)!(n-m-(k-y))!}p^{k-y}q^{n-m-(k-y)}\\
&= \binom{m}{y}p^y q^{m-y}\sum_{j=0}^{n-m}\binom{n-m}{j}p^j q^{n-m-j}\\
&= \binom{m}{y}p^y q^{m-y}.
\end{aligned}
$$

(b) The distribution of Y has nothing to do with when the majors were declared, so we can let them be made after drawing the sample. Then by the story of the Binomial, we have $Y \sim \text{Bin}(m, p)$.