Stat S-110 Homework 3 Solutions, Summer 2015

The following problems are from Chapter 4 of the book.

1. (BH 4.7) A certain small town, whose population consists of 100 families, has 30 families with 1 child, 50 families with 2 children, and 20 families with 3 children. The *birth rank* of one of these children is 1 if the child is the firstborn, 2 if the child is the secondborn, and 3 if the child is the thirdborn.

(a) A random family is chosen (with equal probabilities), and then a random child within that family is chosen (with equal probabilities). Find the PMF, mean, and variance of the child's birth rank.

(b) A random child is chosen in the town (with equal probabilities). Find the PMF, mean, and variance of the child's birth rank.

Solution:

(a) Let X be the child's birth rank. Using LOTP to condition on how many children are in the random family, the PMF of X is

$$\begin{aligned} P(X=1) &= 0.3 \cdot 1 + 0.5 \cdot 0.5 + 0.2 \cdot (1/3) \approx 0.617, \\ P(X=2) &= 0.3 \cdot 0 + 0.5 \cdot 0.5 + 0.2 \cdot (1/3) \approx 0.317, \\ P(X=3) &= 0.3 \cdot 0 + 0.5 \cdot 0 + 0.2 \cdot (1/3) \approx 0.0667. \end{aligned}$$

So

$$\begin{aligned} E(X) &= P(X=1) + 2P(X=2) + 3P(X=3) = 1.45, \\ E(X^2) &= 1^2 P(X=1) + 2^2 P(X=2) + 3^2 P(X=3) \approx 2.483, \\ \mathrm{Var}(X) &= E(X^2) - (EX)^2 \approx 0.381. \end{aligned}$$

(b) Let Y be the child's birth rank. There are $30 + 50 \cdot 2 + 20 \cdot 3 = 190$ children, of which 100 are firstborn, 70 are second born, and 20 are thirdborn. The PMF of Y is

$$\begin{aligned} P(Y=1) &= 100/190 \approx 0.526, \\ P(Y=2) &= 70/190 \approx 0.368, \\ P(Y=3) &= 20/190 \approx 0.105. \end{aligned}$$

It makes sense intuitively that $P(X=1)$ from (a) is less than $P(Y=1)$, since the sampling method from (a) gives all families equal probabilities, whereas the sampling

method from (b) gives the children equal probabilities, which effectively gives higher probabilities of being represented to larger families. Then

$$\begin{aligned} E(Y) &= P(Y=1) + 2P(Y=2) + 3P(Y=3) = 1.579, \\ E(Y^2) &= 1^2P(Y=1) + 2^2P(Y=2) + 3^2P(Y=3) \approx 2.947, \\ \mathrm{Var}(Y) &= E(Y^2) - (EY)^2 \approx 0.454. \end{aligned}$$

2. (BH 4.8) A certain country has four regions: North, East, South, and West. The populations of these regions are 3 million, 4 million, 5 million, and 8 million, respectively. There are 4 cities in the North, 3 in the East, 2 in the South, and there is only 1 city in the West. Each person in the country lives in exactly one of these cities.

(a) What is the average size of a city in the country? (This is the arithmetic mean of the populations of the cities, and is also the expected value of the population of a city chosen uniformly at random.)

Hint: Give the cities *names* (labels).

(b) Show that without further information it is impossible to find the variance of the population of a city chosen uniformly at random. That is, the variance depends on how the people within each region are allocated between the cities in that region.

(c) A region of the country is chosen uniformly at random, and then a city within that region is chosen uniformly at random. What is the expected population size of this randomly chosen city?

Hint: First find the selection probability for each city.

(d) Explain intuitively why the answer to (c) is larger than the answer to (a).

Solution:

(a) Let x_i be the population of the ith city, for $1 \le i \le 10$, with respect to some labeling of the cities. The sum of the city populations equals the sum of the region populations, so the average size of a city is

$$\frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10}(3+4+5+8) \text{ million} = 2 \text{ million}.$$

(b) The variance is

$$\frac{1}{10}\sum_{i=1}^{10} x_i^2 - \left(\frac{1}{10}\sum_{i=1}^{10} x_i\right)^2.$$

The second term is $(2 \cdot 10^6)^2$ by (a), but the first term depends on how people are allocated to cities. For example, if the two cities in the South each have 2.5 million people, then the South contributes $2(2.5 \cdot 10^6)^2 = 1.25 \cdot 10^{13}$ to the sum of x_i^2, but if one city in the South has 5 million people and the other is completely deserted, then the South contributes $(5 \cdot 10^6)^2 = 25 \cdot 10^{12}$ to the sum of x_i^2.

(c) Let x_i be the population of the ith city, with respect to an ordering where cities $1, \ldots, 4$ are in the North, $5, \ldots, 7$ are in the East, $8, 9$ are in the South, and 10 is in the West. The probability that the ith city is chosen is $1/16$ for $1 \leq i \leq 4$, $1/12$ for $5 \leq i \leq 7$, $1/8$ for $8 \leq i \leq 9$, and $1/4$ for $i = 10$. So the expected population size is

$$\frac{x_1 + \cdots + x_4}{16} + \frac{x_5 + x_6 + x_7}{12} + \frac{x_8 + x_9}{8} + \frac{x_{10}}{4} = \left(\frac{3}{16} + \frac{4}{12} + \frac{5}{8} + \frac{8}{4}\right) \cdot 10^6 \approx 3.146 \times 10^6.$$

(d) It makes sense intuitively that the answer to (c) is greater than the answer to (a), since in (a) all cities are equally likely, whereas in (c) the city in the West, which is very populous, is more likely to be chosen than any other particular city.

3. (BH 4.19) Let $X \sim \text{Bin}(100, 0.9)$. For each of the following parts, construct an example showing that it is possible, or explain clearly why it is impossible. In this problem, Y is a random variable on the same probability space as X; note that X and Y are not necessarily independent.

(a) Is it possible to have $Y \sim \text{Pois}(0.01)$ with $P(X \geq Y) = 1$?

(b) Is it possible to have $Y \sim \text{Bin}(100, 0.5)$ with $P(X \geq Y) = 1$?

(c) Is it possible to have $Y \sim \text{Bin}(100, 0.5)$ with $P(X \leq Y) = 1$?

Solution:

(a) No, since there is a positive probability that Y will exceed 100, whereas X is always at most 100.

(b) Yes. To construct such an example, consider the same sequence of trials, except with two definitions of "success": a less stringent definition for X and a more stringent definition for Y (any a success for Y is a success for X). Specifically, we can let the experiment consist of rolling a fair 10-sided die (with sides labeled 1 through 10) 100 times, and let X and Y be the numbers of times the value of the die was at most 9 and at most 5, respectively.

(c) No, this is impossible, since if $X \leq Y$ holds with probability 1, then $E(X) \leq E(Y)$, but in fact we have $E(X) = 90 > 50 = E(Y)$.

4. (BH 4.25) Nick and Penny are independently performing independent Bernoulli trials. For concreteness, assume that Nick is flipping a nickel with probability p_1 of Heads and Penny is flipping a penny with probability p_2 of Heads. Let $X_1, X_2, \ldots$ be Nick's results and $Y_1, Y_2, \ldots$ be Penny's results, with $X_i \sim \text{Bern}(p_1)$ and $Y_j \sim \text{Bern}(p_2)$.

(a) Find the distribution and expected value of the first time at which they are simultaneously successful, i.e., the smallest n such that $X_n = Y_n = 1$.

Hint: Define a new sequence of Bernoulli trials and use the story of the Geometric.

(b) Find the expected time until at least one has a success (including the success).

Hint: Define a new sequence of Bernoulli trials and use the story of the Geometric.

(c) For $p_1 = p_2$, find the probability that their first successes are simultaneous, and use this to find the probability that Nick's first success precedes Penny's.

Solution:

(a) Let N be the time this happens. Then N has a First Success distribution with parameter p_1p_2, so the PMF is $P(N = n) = p_1p_2(1 - p_1p_2)^{n-1}$ for $n = 1, 2, \ldots$, and the mean is $E(N) = 1/(p_1p_2)$.

(b) Let T be the time this happens, and let $q_1 = 1 - p_1, q_2 = 1 - p_2$. Define a new sequence of Bernoulli trials by saying that the jth trial is a success if at least one of the two people succeeds in the jth trial. These trials have probability $1 - q_1q_2$ of success, which implies that $T - 1 \sim \text{Geom}(1 - q_1q_2)$. Therefore, $E(T) = 1/(1 - q_1q_2)$.

(c) Let T_1 and T_2 be the first times at which Nick and Penny are successful, respectively. Let $p = p_1 = p_2$ and $q = 1 - p$. Then

$$P(T_1 = T_2) = \sum_{n=1}^{\infty} P(T_1 = n|T_2 = n)P(T_2 = n) = \sum_{n=1}^{\infty} p^2 q^{2(n-1)} = \frac{p^2}{1 - q^2} = \frac{p}{2 - p}.$$

By symmetry,

$$1 = P(T_1 < T_2) + P(T_2 < T_1) + P(T_1 = T_2) = 2P(T_1 < T_2) + P(T_1 = T_2).$$

So

$$P(T_1 < T_2) = \frac{1}{2}\left(1 - \frac{p}{2 - p}\right) = \frac{1 - p}{2 - p}.$$

5. (BH 4.34) Each of $n \geq 2$ people puts his or her name on a slip of paper (no two have the same name). The slips of paper are shuffled in a hat, and then each

person draws one (uniformly at random at each stage, without replacement). Find the average number of people who draw their own names.

Solution: Label the people as $1, 2, \ldots, n$, let I_j be the indicator of person j getting their own name, and let $X = I_1 + \cdots + I_n$. By symmetry and linearity,

$$E(X) = nE(I_1) = n \cdot \frac{1}{n} = 1.$$

6. (BH 4.36) In a sequence of n independent fair coin tosses, what is the expected number of occurrences of the pattern HTH (consecutively)? Note that overlap is allowed, e.g., $HTHTH$ contains two overlapping occurrences of the pattern.

Solution: Let I_j be the indicator of the pattern HTH occurring starting at position j, for $1 \leq j \leq n-2$. By symmetry, linearity, and the fundamental bridge, the expected number of occurrences of HTH is $(n-2)/8$ (assuming, of course, that $n \geq 3$; if $n < 3$, then there are no occurrences).

7. (BH 4.43) You are being tested for psychic powers. Suppose that you do not have psychic powers. A standard deck of cards is shuffled, and the cards are dealt face down one by one. Just after each card is dealt, you name any card (as your prediction). Let X be the number of cards you predict correctly. (See Diaconis (1978) for much more about the statistics of testing for psychic powers.)

(a) Suppose that you get no feedback about your predictions. Show that no matter what strategy you follow, the expected value of X stays the same; find this value. (On the other hand, the *variance* may be very different for different strategies. For example, saying "Ace of Spades" every time gives variance 0.)

Hint: Indicator r.v.s.

(b) Now suppose that you get partial feedback: after each prediction, you are told immediately whether or not it is right (but without the card being revealed). Suppose you use the following strategy: keep saying a specific card's name (e.g., "Ace of Spades") until you hear that you are correct. Then keep saying a different card's name (e.g., "Two of Spades") until you hear that you are correct (if ever). Continue in this way, naming the same card over and over again until you are correct and then switching to a new card, until the deck runs out. Find the expected value of X, and show that it is very close to $e - 1$.

Hint: Indicator r.v.s.

(c) Now suppose that you get complete feedback: just after each prediction, the card is revealed. Call a strategy "stupid" if it allows, e.g., saying "Ace of Spades" as a

guess after the Ace of Spades has already been revealed. Show that any non-stupid strategy gives the same expected value for X; find this value.

Hint: Indicator r.v.s.

Solution:

(a) Think of the cards as labeled from 1 to 52, so each card in the deck has a "codename" (this is just so we can use random variables rather than random cards). Let $C_1, \ldots, C_{52}$ be the labels of the cards in the shuffled deck (so the card on top of the deck has label C_1, etc.). Let $Y_1, \ldots, Y_{52}$ be your predictions, and let I_j be the indicator of your jth prediction being right. Then $X = I_1 + \cdots + I_{52}$. Since you do not have psychic powers (by assumption) and there is no feedback, $(C_1, \ldots, C_{52})$ is independent of $(Y_1, \ldots, Y_{52})$. Thus, $P(I_j = 1) = 1/52$ and

$$E(X) = E(I_1) + \cdots + E(I_{52}) = 1.$$

(b) To simplify notation, assume that the strategy is to keep saying "Card 1" until you hear that you are correct, then start saying "Card 2", etc. Let X_j be the indicator of your guessing the card with label j correctly when it is dealt. So $X = X_1 + \cdots + X_{52}$. We have $X_1 = 1$ always. For X_2, note that $X_2 = 1$ if and only if Card 1 precedes Card 2 in the deck (if Card 2 precedes Card 1, you will miss out on Card 2 while being fixated on Card 1). Then $X_3 = 1$ if and only if Cards 1, 2, and 3 are in that order, and similarly for the other X_j's. Thus,

$$E(X) = 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{52!} \approx 1.718.$$

The sum is *extremely close* to $e - 1$, by the Taylor series of e^x and since factorials grow so fast; the difference is about $2 \cdot 10^{-70}$, so $e - 1$ is an incredibly accurate approximation! Note that the answer to (a) does not depend on the number of cards in the deck, and the answer to this part depends on the number of cards only very slightly, as long as there are, say, at least 10 cards.

(c) If j cards have been revealed, then by symmetry, any non-stupid guess for the next card has probability $1/(52 - j)$ of success (for $0 \leq j \leq 51$). Using indicator r.v.s as in (a), the expected value is

$$\frac{1}{52} + \frac{1}{51} + \cdots + \frac{1}{2} + 1 \approx 4.538.$$

The expected value is the sum of the first 52 terms of the harmonic series; for an n-card deck with n large, this is approximately $\ln(n) + \gamma$, where $\gamma \approx 0.577$. For

$n = 52$, this approximation gives 4.528, which is quite close to the true value. Note that the answers to (a), (b), (c) are in increasing order, which is very sensible since with more information it should be possible to do better (or at least not do worse!).

8. (BH 4.66) Use Poisson approximations to investigate the following types of coincidences. The usual assumptions of the birthday problem apply, such as that there are 365 days in a year, with all days equally likely.

(a) How many people are needed to have a 50% chance that at least one of them has the same birthday as *you*?

(b) How many people are needed to have a 50% chance that there are two people who not only were born on the same day, but also were born at the same *hour* (e.g., two people born between 2 pm and 3 pm are considered to have been born at the same hour).

(c) Considering that only 1/24 of pairs of people born on the same day were born at the same hour, why isn't the answer to (b) approximately $24 \cdot 23$? Explain this intuitively, and give a simple approximation for the factor by which the number of people needed to obtain probability p of a birthday match needs to be scaled up to obtain probability p of a birthday-birthhour match.

(d) With 100 people, there is a 64% chance that there are 3 with the same birthday (according to R, using `pbirthday(100,classes=365,coincident=3)` to compute it). Provide two different Poisson approximations for this value, one based on creating an indicator r.v. for each triplet of people, and the other based on creating an indicator r.v. for each day of the year. Which is more accurate?

Solution:

(a) Let k be the number of people, and for each create an indicator for that person having the same birthday as you. These indicator r.v.s are independent, each with probability 1/365 of being 1. By Poisson approximation, the probability of at least 1 match is approximately $1 - e^{-k/365}$. The smallest k for which this is at least 0.5 is

$$k = 253.$$

This turns out to be *exactly* the right number, as seen using the exact probability of at least one match, which is $1 - (1 - 1/365)^k$.

(b) This is the birthday problem, with $365 \cdot 24$ types of people rather than 365. A Poisson approximation gives $1 - e^{-\binom{k}{2}/(365 \cdot 24)}$ as the approximate probability of at least one match. The smallest k for which this is at least 0.5 is

$$k = 111.$$

This again turns out to be exactly right, as seen using `qbirthday(0.5,classes=365*24)` to compute it in R.

(c) What matters is the number of *pairs* of people, which grows quadratically as a function of the number of people. Let k be the number of people in the birthday-birthhour problem and m be the number of people in the birthday problem. To obtain about the same probability p for both problems, we need

$$e^{-\binom{k}{2}/(365\cdot 24)} \approx e^{-\binom{m}{2}/365},$$

which reduces to

$$k \approx \sqrt{24}m,$$

using the approximations $\binom{k}{2} = k(k-1)/2 \approx k^2/2$ and $\binom{m}{2} \approx m^2/2$. For $p = 0.5$, this correctly suggests using $\sqrt{24} \cdot 23 \approx 113$ rather than $24 \cdot 23 = 552$ as an approximate way to convert from the birthday problem to the birthday-birthhour problem.

(d) Creating an indicator for each triplet of people gives

$$1 - e^{-\binom{100}{3}/365^2} \approx 0.70$$

as the Poisson approximation for the probability of at least one triple match. An alternative method is to create an indicator for each day of the year: let I_j be the indicator for at least 3 people having been born on the jth day of the year, and $X = I_1 + \cdots + I_{365}$. Then $X = 0$ is the event that there is no triple birthday match. We have

$$E(I_j) = P(I_j = 1) = 1 - \left(\frac{364}{365}\right)^{100} - 100 \cdot \left(\frac{364}{365}\right)^{99} \cdot \left(\frac{1}{365}\right) - \binom{100}{2} \cdot \left(\frac{364}{365}\right)^{98} \cdot \left(\frac{1}{365}\right)^2,$$

and by linearity $E(X) = 365E(I_1)$. We then have the Poisson approximation

$$P(X > 0) = 1 - P(X = 0) \approx 1 - e^{-E(X)} \approx 0.63.$$

The latter is closer to 0.64 (the correct value, as stated in the problem and found using `pbirthday(100,coincident=3)` in R). An intuitive explanation for why the former approximation is less accurate is that there is a more substantial dependence in the indicators in that method: note that if persons 1, 2, and 3 all have the same birthday, say October 31, then if person 4 is born on October 31 that will automatically result in the triplets $\{1, 2, 4\}$, $\{1, 3, 4\}$, and $\{2, 3, 4\}$ also being triple birthday matches.