Problem Sheet 4 [1]

Optional Revision Problems

Exercise 1 (Conditional Independence)**.** Two different diseases cause a certain weird symptom; anyone who has either or both of these diseases will experience the symptom. Let D_1 be the event of having the first disease, D_2 be the event of having the second disease, and W be the event of having the weird symptom. Suppose that D_1 and D_2 are independent with $P(D_j) = p_j$, and that a person with neither of these diseases will have the weird symptom with probability w_0. Let $q_j = 1 - p_j$, and assume that $0 < p_j < 1$.

1. Find $P(W)$.
2. Find $P(D_1|W)$, $P(D_2|W)$, and $P(D_1, D_2|W)$.
3. Determine algebraically whether or not D_1 and D_2 are conditionally independent given W.

Solution 1. 1. By the law of total probability and the independence between D_1 and D_2

$$\begin{aligned}
P(W) &= P(W|D_1^c, D_2^c)P(D_1^c, D_2^c) + P(W|D_1, D_2^c)P(D_1, D_2^c) \\
&\quad + P(W|D_1^c, D_2)P(D_1^c, D_2) + P(W|D_1, D_2)P(D_1, D_2) \\
&= P(W|D_1^c, D_2^c)P(D_1^c)P(D_2^c) + P(W|D_1, D_2^c)P(D_1)P(D_2^c) \\
&\quad + P(W|D_1^c, D_2)P(D_1^c)P(D_2) + P(W|D_1, D_2)P(D_1)P(D_2) \\
&= w_0 q_1 q_2 + 1 \cdot p_1 q_2 + 1 \cdot q_1 p_2 + 1 \cdot p_1 p_2.
\end{aligned}$$

2. By Bayes' rule

$$P(D_1|W) = \frac{P(W|D_1)P(D_1)}{P(W)} = \frac{1 \cdot p_1}{w_0 q_1 q_2 + 1 \cdot p_1 q_2 + 1 \cdot q_1 p_2 + 1 \cdot p_1 p_2}.$$

Similarly,

$$P(D_2|W) = \frac{1 \cdot p_2}{w_0 q_1 q_2 + 1 \cdot p_1 q_2 + 1 \cdot q_1 p_2 + 1 \cdot p_1 p_2},$$

and by using independence of D_1 and D_2

$$P(D_1, D_2|W) = \frac{P(W|D_1, D_2)P(D_1)P(D_2)}{P(W)} = \frac{1 \cdot p_1 p_2}{w_0 q_1 q_2 + 1 \cdot p_1 q_2 + 1 \cdot q_1 p_2 + 1 \cdot p_1 p_2}.$$

[1] Exercises are based on the coursebook Statistics 110: Probability by Joe Blitzstein

3. If $w_0q_1q_2+p_1q_2+q_1p_2+p_1p_2=1$, that is equivalent to $w_0=\frac{1-p_1q_2-q_1p_2-p_1p_2}{q_1q_2}$, the events D_1 and D_2 are conditionally independent given W. You can see this by multiplying the solutions from Part 2, as if $w_0q_1q_2+p_1q_2+q_1p_2+p_1p_2=1$, then $P(D_1|W)\cdot P(D_2|W)=p_1p_2=P(D_1,D_2|W)$. For intuition, note that $w_0q_1q_2+p_1q_2+q_1p_2+p_1p_2=P(W)=1$, i.e. this implies that everyone develops the weird symptom. Then knowing if someone developed the symptom carries no extra information, as we knew beforehand that everyone has it, and as the diseases themselves are independent they are also conditionally independent.

   If $w_0q_1q_2+p_1q_2+q_1p_2+p_1p_2\neq 1$, $P(D_1|W)P(D_2|W)\neq P(D_1,D_2|W)$, hence D_1 and D_2 are not conditionally independent given W.

Exercise 2 (Monty Hall)**.** Consider the Monty Hall problem, except that Monty enjoys opening door 2 more than he enjoys opening door 3, and if he has a choice between opening these two doors, he opens door 2 with probability p, where $\frac{1}{2}\leq p\leq 1$.

To recap: there are three doors, behind one of which there is a car (which you want), and behind the other two of which there are goats (which you don't want). Initially, all possibilities are equally likely for where the car is. You choose a door, which for concreteness we assume is door 1. Monty Hall then opens a door to reveal a goat, and offers you the option of switching. Assume that Monty Hall knows which door has the car, will always open a goat door and offer the option of switching, and as above assume that if Monty Hall has a choice between opening door 2 and door 3, he chooses door 2 with probability p, (with $\frac{1}{2}\leq p\leq 1$).

1. Find the unconditional probability that the strategy of always switching succeeds (unconditional in the sense that we do not condition on which of doors 2 or 3 Monty opens).

2. Find the probability that the strategy of always switching succeeds, given that Monty opens door 2.

3. Find the probability that the strategy of always switching succeeds, given that Monty opens door 3.

Solution 2. 1. Let C_j be the event that the car is hidden behind door j and let W be the event that we win using the switching strategy. Using the law of total probability, we can find the unconditional probability of winning:

$$\begin{aligned}P(W)&=P(W|C_1)P(C_1)+P(W|C_2)P(C_2)+P(W|C_3)P(C_3)\\&=0\cdot 1/3+1\cdot 1/3+1\cdot 1/3=2/3.\end{aligned}$$

2. Let D_i be the event that Monty opens Door i. Note that we are looking for $P(W|D_2)$, which is the same as $P(C_3|D_2)$ as we first choose Door 1 and then switch to Door 3. By Bayes' rule and the law of total probability,

$$\begin{aligned}P(C_3\mid D_2)&=\frac{P(D_2\mid C_3)P(C_3)}{P(D_2)}\\&=\frac{P(D_2\mid C_3)P(C_3)}{P(D_2\mid C_1)P(C_1)+P(D_2\mid C_2)P(C_2)+P(D_2\mid C_3)P(C_3)}\\&=\frac{1\cdot\frac{1}{3}}{p\cdot\frac{1}{3}+0\cdot\frac{1}{3}+1\cdot\frac{1}{3}}\\&=\frac{1}{1+p}.\end{aligned}$$

3. The structure of the problem is the same as part 2. Imagine repainting doors 2 and 3, reversing which is called which. By part 2. with $1-p$ in place of p, $P(C_2|D_3) = \frac{1}{1+(1-p)} = \frac{1}{2-p}$.

Exercise 3 (Simpson's paradox)**.** The book *Red State, Blue State, Rich State, Poor State* by Andrew Gelman discusses the following election phenomenon: within any U.S. state, a wealthy voter is more likely to vote for a Republican than a poor voter, yet the wealthier states tend to favor Democratic candidates!

1. Assume for simplicity that there are only 2 states (called Red and Blue), each of which has 100 people, and that each person is either rich or poor, and either a Democrat or a Republican. Make up numbers consistent with the above, showing how this phenomenon is possible, by giving a 2×2 table for each state (listing how many people in each state are rich Democrats, etc.). So within each state, a rich voter is more likely to vote for a Republican than a poor voter, but the percentage of Democrats is higher in the state with the higher percentage of rich people than in the state with the lower percentage of rich people.

2. In the setup of part 1. (not necessarily with the numbers you made up there), let D be the event that a randomly chosen person is a Democrat (with all 200 people equally likely), and B be the event that the person lives in the Blue State. Suppose that 10 people move from the Blue State to the Red State. Write P_{old} and P_{new} for probabilities before and after they move. Assume that people do not change parties, so we have $P_{\text{new}}(D) = P_{\text{old}}(D)$. Is it possible that both $P_{\text{new}}(D \mid B) > P_{\text{old}}(D \mid B)$ and $P_{\text{new}}(D \mid B^c) > P_{\text{old}}(D \mid B^c)$ are true? If so, explain how it is possible and why it does not contradict the law of total probability $P(D) = P(D \mid B)P(B) + P(D \mid B^c)P(B^c)$; if not, show that it is impossible.

Solution 3. 1. Here are the two tables as desired:

Red	Dem	Rep	**Total**
Rich	5	25	30
Poor	20	50	70
Total	25	75	100

Blue	Dem	Rep	**Total**
Rich	45	15	60
Poor	35	5	40
Total	80	20	100

In these tables, within each state a rich person is more likely to be a Republican than a poor person; but the richer state has a higher percentage of Democrats than the poorer state. Of course, there are many possible tables that work.

The above example is a form of Simpson's paradox: aggregating the two tables seems to give different conclusions than conditioning on which state a person is in. Letting D, W, B be the events that a randomly chosen person is a Democrat, wealthy, and from the Blue State (respectively), for the above numbers we have $P(D|W,B) < P(D|W^c,B)$ and $P(D|W,B^c) < P(D|W^c,B^c)$ (controlling for whether the person is in the Red State or the Blue State, a poor person is more likely to be a Democrat than a rich person),

but $P(D|W) > P(D|W^c)$ (stemming from the fact that the Blue State is richer and more Democratic).

2. Yes, it is possible. Suppose with the numbers from part 1. that 10 people move from the Blue State to the Red State, of whom 5 are Democrats and 5 are Republicans. Then $P_{\text{new}}(D|B) = 75/90 > 80/100 = P_{\text{old}}(D|B)$ and $P_{\text{new}}(D|B^c) = 30/110 > 25/100 = P_{\text{old}}(D|B^c)$. Intuitively, this makes sense since the Blue State has a higher percentage of Democrats initially than the

Red State, and the people who move have a percentage of Democrats in between these two values.

This result does not contradict the law of total probability since these weights $P(B), P(B^c)$ change from $P_{\text{old}}(B) = 90/200$, while $P_{\text{new}}(B) = 1/2$. The same calculation applies in reverse if 10 people (with the same distribution) had also moved from the Red State to the Blue State (so that $P(B)$ is kept constant).

Week 4 Exercises

Exercise 4. 1. 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.

2. 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 4. 1. Denote with X the number of trials until the first success, including the last trial with the positive outcome. Then the possible values for X are, i.e. the support is $\{1, 2, \dots\}$. Since each trial is independent, the probability of $k-1$ failures followed by a success is

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

for $k \in \{1, 2, \dots\}$ and 0 otherwise.

2. Denote with Y the number of trials until there has been at least one success and at least one failure. Then as we must perform at least 2 trials, the possible values for Y are $\{2, 3, \dots\}$. The event corresponding to $Y = k$ can happen if $k-1$ failures is followed by one success or if $k-1$ successes are followed by one failure. Hence,

$$P(Y = k) = \left(\frac{1}{2}\right)^{k-1} \cdot \frac{1}{2} + \left(\frac{1}{2}\right)^{k-1} \cdot \frac{1}{2} = \left(\frac{1}{2}\right)^{k-1},$$

for $k \in \{2, 3, \dots\}$ and 0 otherwise.

Exercise 5. *Benford's law* states that in a very large variety of real-life data sets, the first digit approximately follows a particular distribution with about a 30% chance of a 1, an 18% chance of a 2, and in general

$$P(D = j) = \log_{10}\left(\frac{j+1}{j}\right), \text{ for } j \in \{1, 2, 3, \dots, 9\},$$

where D is the first digit of a randomly chosen element. Check that this is a valid PMF (using properties of logs, not with a calculator).

Solution 5. The function $P(D = j)$ is nonnegative and the sum over all values is

$$\sum_{j=1}^{9} \log_{10}\left(\frac{j+1}{j}\right) = \sum_{j=1}^{9} \left(\log_{10}(j+1) - \log_{10}(j)\right).$$

All terms cancel except $\log_{10} 10 - \log_{10} 1 = 1$ (this is a *telescoping series*). Since the values add to 1 and are nonnegative, $P(D = j)$ is a PMF.

Exercise 6. There are 100 prizes, with one worth \$1, one worth \$2, . . . , and one worth \$100. There are 100 boxes, each of which contains one of the prizes. You get 5 prizes by picking random boxes one at a time, without replacement. Find the PMF of how much your most valuable prize is worth (as a simple expression in terms of binomial coefficients).

Solution 6. Denote with X the value of the most valuable prize won. Then

$$P(X \leq k) = \frac{\binom{k}{5}}{\binom{100}{5}},$$

as the 5 prizes can be chosen from the 100 options $\binom{100}{5}$ ways, but if we restrict ourselves to the prizes from 1 to k, then we have $\binom{k}{5}$ options. Moreover, the event $\{X \leq k\}$ can be rewritten as the disjoint union of the events $\{X = k\}$ and $\{X \leq k-1\}$, hence the probability of interest is

$$P(X = k) = P(X \leq k) - P(X \leq k-1) = \frac{\binom{k}{5}}{\binom{100}{5}} - \frac{\binom{k-1}{5}}{\binom{100}{5}},$$

for $k \in \{5, 6, \ldots, 100\}$ and 0 otherwise.

Alternative Solution: Set the box with the highest value to k. Then you have to pick the four remaining boxes from the set $\{1, \ldots, (k-1)\}$, that you can do $\binom{k-1}{4}$ ways. Hence the probability of interest is

$$P(X = k) = \frac{\binom{k-1}{4}}{\binom{100}{5}},$$

for $k \in \{5, 6, \ldots, 100\}$ and 0 otherwise.

In fact, you can use the two solutions as a story proof to show

$$\binom{k-1}{l-1} = \binom{k}{l} - \binom{k-1}{l},$$

for $1 \leq l < k$.

Exercise 7. An airline overbooks a flight, selling more tickets for the flight than there are seats on the plane (figuring that it's likely that some people won't show up). The plane has 100 seats, and 110 people have booked the flight. Each person will show up for the flight with probability 0.9, independently. Find the probability that there will be enough seats for everyone who shows up for the flight

Solution 7. As the individuals are showing up independently, we can consider this scenario as 110 independent Bernoulli trials, that is a known distribution, the *Binomial distribution*, with a known PMF $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, where $n = 110$ and $p = 0.9$. The event that everyone who shows up has a seat corresponds to the event that less than or equal to 100 passengers showed up for the flight. Hence the probability of interest is

$$P(X \leq 100) = \sum_{k=0}^{100} \binom{110}{k} 0.9^k 0.1^{110-k},$$

or alternatively, as we know that the PMF should sum up to 1,

$$P(X \leq 100) = 1 - \sum_{k=101}^{110} \binom{110}{k} 0.9^k 0.1^{110-k} \approx 0.671.$$

Exercise 8. 1. In the World Series of baseball, two teams (call them A and B) play a sequence of games against each other, and the first team to win four games wins the series. Let p be the probability that A wins an individual game, and assume that the games are independent. What is the probability that team A wins the series?

2. Give a clear intuitive explanation of whether the answer to 1. depends on whether the teams always play 7 games (and whoever wins the majority wins the series), or the teams stop playing more games as soon as one team has won 4 games (as is actually the case in practice: once the match is decided, the two teams do not keep playing more games).

Solution 8. 1. Let $q = 1 - p$. First let us do a direct calculation:

$$\begin{aligned} P(A \text{ wins}) =& P(A \text{ wins in 4 games}) + P(A \text{ wins in 5 games}) \\ &+ P(A \text{ wins in 6 games}) + P(A \text{ wins in 7 games}) \\ &= p^4 + \binom{4}{3} p^4 q + \binom{5}{3} p^4 q^2 + \binom{6}{3} p^4 q^3. \end{aligned}$$

To understand how these probabilities are calculated, note for example that

$$\begin{aligned} P(A \text{ wins in 5}) &= P((A \text{ wins the 5th game}) \cap (A \text{ wins 3 out of first 4})) \\ &= P(A \text{ wins 3 out of first 4}) \cdot P(A \text{ wins 5th game} \mid A \text{ wins 3 out of first 4}) \\ &= \binom{4}{3} p^3 q \cdot p, \end{aligned}$$

where in the first equality we used that A winning in 5 games implies that A won exactly 3 games out of the first 4, and in the second line we used the definition of conditional probability.

(Each of the 4 terms in the expression for $P(A \text{ wins})$ can also be found using the PMF of a distribution known as the Negative Binomial, which is introduced in Chapter 4.)

A neater solution is to use the fact (explained in the solution to Part 2.) that we can assume that the teams play all 7 games no matter what. Let X be the number of wins for team A, so that $X \sim \text{Binom}(7, p)$. Then

$$\begin{aligned} P(X \geq 4) &= P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7) \\ &= \binom{7}{4} p^4 q^3 + \binom{7}{5} p^5 q^2 + \binom{7}{6} p^6 q + p^7, \end{aligned}$$

which looks different from the above but is actually identical as a function of p (as can be verified by simplifying both expressions as polynomials in p).

2. The answer to part 1. does not depend on whether the teams play all seven games no matter what. Imagine telling the players to continue playing the games even after the match has been decided, just for fun: the outcome of the match won't be affected by this, and this also means that the probability that A wins the match won't be affected by assuming that the teams always play 7 games!

Exercise 9. A certain company has n + m employees, consisting of n women and m men. The company is deciding which employees to promote.

1. Suppose for this part that the company decides to promote t employees, where $1 \leq t \leq n+m$, by choosing t random employees (with equal probabilities for each set of t employees). What is the distribution of the number of women who get promoted?

2. Now suppose that instead of having a predetermined number of promotions to give, the company decides independently for each employee, promoting the employee with probability p. Find the distributions of the number of women who are promoted, the number of women who are not promoted, and the number of employees who are promoted.

3. In the set-up from part 2., find the conditional distribution of the number of women who are promoted, given that exactly t employees are promoted.

Solution 9. 1. Denote the number of women promoted among the t promoted employees with W^*. Translate this problem to the urn "story" introduced for the Hypergeometric distribution during the lecture. For each female employee add a white ball to the urn, and for each male employee add a black ball. Then since the promotions are given at random, we can assign promotion to the event that the ball corresponding to them is drawn from the urn. In addition, since everyone can be promoted at most once, we will draw the balls from the urn *without* replacement. This scenario exactly the one that was used for the Hypergeometric distribution. Therefore $W^* \sim HGeom(n, m, t)$ and,

$$P(W^* = k) = \frac{\binom{n}{k}\binom{m}{t-k}}{\binom{n+m}{t}},$$

for $k \in \{0, 1, \ldots, t\}$ and 0 otherwise.

2. Denote the number of promoted women with W. If each of the promotions are decided independently with probability p, then this corresponds to $n+m$ independent Bernoulli trials with success probability p. If we are only interested in the women promoted, then we perform n trials, hence $W \sim Binom(n, p)$ and,

$$P(W = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

for $k \in \{0, 1, \ldots, n\}$ and 0 otherwise. If k many women are promoted, then $n - k$ women are not promoted and vice versa, meaning so, if (with a slight abuse of notation) we denote the number of non-promoted women with W^c, then

$$P(W^c = k) = P(W = n - k) = \binom{n}{n-k} p^{n-k} (1-p)^k,$$

for $k \in \{0, 1, \ldots, n\}$ and 0 otherwise. Since $\binom{n}{k} = \binom{n}{n-k}$, it follows that $W^c \sim Binom(n, 1-p)$. This result can also be derived by noting that if we are counting the non-promotions, then the "success" in the Bernoulli trial is not receiving a promotion, which happens with probability $1 - p$.

Denote with T the total number of promotions. Then by the same reasoning as before $T \sim Binom(n + m, p)$ and,

$$P(T = k) = \binom{n+m}{k} p^k (1-p)^{n+m-k},$$

for $k \in \{0, 1, \ldots, n + m\}$ and 0 otherwise.

3. By the definition of conditional probability for $k < t$

$$P(W = k|T = t) = \frac{P(W = k \cap T = t)}{P(T = t)}.$$

Denote the number of male employees promoted with M. Then event $(W = k) \cap (T = t)$ is equivalent to $(W = k) \cap (M = t - k)$, as if t many employees are promoted and out of them k were women, then as this company only has male and female employees, the remaining promoted individuals must be men. Moreover, as the promotions are conducted independently, $P((W = k) \cap (M = t - k)) = P(W = k)P(M = t - k)$. Using the findings from part 2., and noting that the male promotions are distributed as $Binom(m, p)$

$$\begin{aligned} P(W = k|T = t) &= \frac{\binom{n}{k}p^k(1-p)^{n-k} \cdot \binom{m}{t-k}p^{t-k}(1-p)^{m-t+k}}{\binom{n+m}{t}p^t(1-p)^{n+m-t}} \\ &= \frac{\binom{n}{k}\binom{m}{t-k}}{\binom{n+m}{t}}, \end{aligned}$$

for $k \in \{0, 1, \ldots, min(n, t)\}$ and 0 otherwise; meaning that $W|T = t \sim HGeom(n, m, t)$.

This is exactly as in part 1., it does not matter whether we select t employees at random, or we condition on the fact that t employees were promoted, and then each of the promotions are determined independently, the number of women promoted will be distributed the same way.