Problem Sheet 6 [1]

Optional Revision Problems

Exercise 1. A book has n typos. Two proofreaders, Prue and Frida, independently read the book. Prue catches each typo with probability p_1 and misses it with probability $q_1 = 1-p_1$, independently, and likewise for Frida, who has probabilities p_2 of catching and $q_2 = 1 - p_2$ of missing each typo. Let X_1 be the number of typos caught by Prue, X_2 be the number caught by Frida, and X be the number caught by at least one of the two proofreaders.

1. Find the distribution of X.
2. For this part only, assume that $p_1 = p_2$. Find the conditional distribution of X_1 given that $X_1 + X_2 = t$.

Solution 1. 1. Denote the events in which Prue and Frida catch a typo with T_1 and T_2, respectively. Then using De Morgan's laws and that they are catching typos independently, the probability that at least one of them is catching a typo is

$$P(T_1 \cup T_2) = 1 - P((T_1 \cup T_2)^c) = 1 - P(T_1^c \cap T_2^c) = 1 - P(T_1^c)P(T_2^c) = 1 - q_1 \cdot q_2$$

.

Alternatively, by independence and by the inclusion exclusion formula

$$P(T_1 \cap T_2) = P(T_1) + P(T_2) - P(T_1 \cap T_2) = P(T_1) + P(T_2) - P(T_1)P(T_2) = p_1 + p_2 - p_1 \cdot p_2.$$

(If you substitute in $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$, you can get that these two are the same).

Hence, for each typo, a "trial" is performed and a typo is caught at least by one of them with probability $1-q_1 \cdot q_2$. We have n many typos/trials in total that are i.i.d $Bernoulli(1-q_1 \cdot q_2)$, so it follows that $X \sim Binom(n, 1 - q_1 \cdot q_2)$.

2. We want to find the probability $P(X_1 = k|X_1 + X_2 = t)$. As Prue cannot find more typos than the combined numbers between her and Frida, and she cannot find a negative number of typos $P(X_1 = k|X_1 + X_2 = t) = 0$ for $k < 0$ and $k > t$.

Consider $0 \leq k \leq t$. By the definition of conditional probability

$$P(X_1 = k|X_1 + X_2 = t) = \frac{P((X_1 = k) \cap (X_1 + X_2 = t))}{P(X_1 + X_2 = t)}.$$

The event $\{(X_1 = k) \cap (X_1 + X_2 = t)\} = \{(X_1 = k) \cap (X_2 = t - k)\}$ as if Prue found k many typos and together combined they found t many typos, then Frida must have found

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

the remaining $t-k$ typos. As they are finding typos independently it follows that $P(X_1 = k) \cap (X_2 = t-k)) = P(X_1 = k)P(X_2 = t-k)$.

Each of them is performing n independent typo checks with success rates p_1 and p_2 respectively, so it follows that $X_1 \sim Binom(n, p_1)$ and $X_2 \sim Binom(n, p_2)$. As $p_1 = p_2$ it follows that $X_1 + X_2 \sim Binom(n+n, p_1)$ (see digital whiteboard page 30). Now we can finally substitute back to the previous expression:

$$P(X_1 = k|X_1 + X_2 = t) = \frac{P(X_1 = k)P(X_2 = t-k)}{P(X_1 + X_2 = t)}$$
$$= \frac{\binom{n}{k}p^k(1-p)^{n-k}\binom{n}{t-k}p^{t-k}(1-p)^{n-t+k}}{\binom{2n}{t}p^t(1-p)^{2n-t}} = \frac{\binom{n}{k}\binom{n}{t-k}}{\binom{2n}{t}},$$

that means that $(X_1|X_1 + X_2 = t) \sim HGeom(n, n, t)$.

(Observe the similarity to Exercise S4E9. It is the same problem with a different story, hence the same result)

Exercise 2. Let X, Y, Z be discrete r.v.s such that X and Y have the same conditional distribution given Z, i.e., for all a and z we have

$$P(X = a|Z = z) = P(Y = a|Z = z).$$

Show that X and Y have the same distribution (unconditionally, not just when given Z).

Solution 2. From the law of total probability, it follows that

$$P(X = a) = \sum_{z \in \mathcal{Z}} P(X = a|Z = z)P(Z = z),$$

where $\mathcal{Z}$ stands for all the possible values the random variable Z can take. Given that the two conditional probabilities $P(X = a|Z = z) = P(Y = a|Z = z)$ are equal, and using the law of total probability for Y, we have

$$P(X = a) = \sum_{z \in \mathcal{Z}} P(X = a|Z = z)P(Z = z) = \sum_{z \in \mathcal{Z}} P(Y = a|Z = z)P(Z = z) = P(Y = a),$$

for all a, that is, the PMF of X and Y are the same, so they have the same distribution.

Week 6 Exercises

Exercise 3. Find the mean and variance of a Discrete Uniform r.v. on $1, 2, \ldots, n$.

Hint: See the math appendix for some useful facts about sums.

Solution 3. We defined the expectation of a discrete r.v. X as

$$E(X) = \sum_k k \cdot P(X = k).$$

Since in this exercise X is Discrete Uniform on $1, 2, \ldots, n$, $P(X = k) = 1/n$ if $k \in \{1, 2, \ldots, n\}$ and 0 otherwise. So

$$E(X) = \sum_{k=1}^{n} k(1/n) = \frac{\sum_{k=1}^{n} k}{n} = \frac{(n+1)n}{2n} = \frac{n+1}{2},$$

where in the last but one step we used that $\sum_{k=1}^{n} k = (n+1)n/2$ (example of Gauss summing 1 to 100 from Lecture 5, or Math Appendix).

The variance can be calculated either as $Var(X) = E((X - E(X))^2)$ or as $Var(X) = E(X^2) - E(X)^2$. Arguably the latter is easier, so we continue with that approach. By the law of the unconscious statistician (LOTUS), for a Discrete Uniform X

$$E(X^2) = \sum_{k=1}^{n} k^2 P(X = k) = \sum_{k=1}^{n} k^2 (1/n) = \frac{\sum_{k=1}^{n} k^2}{n}.$$

From the Math Appendix Section A.8.4 we know that

$$\sum_{k=1}^{n} k^2 = n(n+1)(2n+1)/6.$$

Therefore, the variance of X is

$$Var(X) = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{(n+1)(2(2n+1) - 3(n+1))}{12} = \frac{(n+1)(n-1)}{12} = \frac{n^2-1}{12}.$$

Exercise 4. 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.

1. 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.

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

Solution 4. 1. Denote with R_F a child's birth rank, who was chosen randomly from a randomly chosen family, and denote with F the size of the randomly chosen family. We want to find the PMF first, i.e. $P(R_F = k)$, that is non-zero only if $k \in \{1, 2, 3\}$ and zero otherwise. By the law of total probability

$$\begin{aligned} P(R_F = k) = {} & P(R_F = k|F = 1)P(F = 1) + P(R_F = k|F = 2)P(F = 2) \\ & + P(R_F = k|F = 3)P(F = 3). \end{aligned}$$

Observe that $P(R_F = k|F = j) = 0$ for $k > j$, as you cannot have higher birth rank than the size of your family, e.g. you cannot be thirdborn if in your family there are only two children (including yourself). If $k \leq j$ then $P(R_F = k|F = j) = 1/j$ by the naive definition of probability, e.g. if you pick a family with two children, you can pick each of the children with equal probability of 1/2. Finally, by the naive definition of probability, you can pick at random a family of size 1,2, and 3, with probabilities 30/100, 50/100 and 20/100 respectively.

Putting everything together

$$\begin{aligned}
P(R_F = 1) &= P(R_F = 1|F = 1)P(F = 1) + P(R_F = 1|F = 2)P(F = 2) \\
&\quad + P(R_F = 1|F = 3)P(F = 3) \\
&= 1 \cdot \frac{3}{10} + \frac{1}{2} \cdot \frac{5}{10} + \frac{1}{3} \cdot \frac{2}{10} = \frac{37}{60} \\
P(R_F = 2) &= P(R_F = 2|F = 2)P(F = 2) + P(R_F = 2|F = 3)P(F = 3) \\
&= \frac{1}{2} \cdot \frac{5}{10} + \frac{1}{3} \cdot \frac{2}{10} = \frac{19}{60} \\
P(R_F = 3) &= P(R_F = 3|F = 3)P(F = 3) \\
&= \frac{1}{3} \cdot \frac{2}{10} = \frac{4}{60},
\end{aligned}$$

and $P(R_F = k) = 0$ for all other values of k. By the definition of expectation

$$E(R_F) = \sum_{k=1}^{3} k \cdot P(R_F = k) = 1 \cdot \frac{37}{60} + 2 \cdot \frac{19}{60} + 3 \cdot \frac{4}{60} = \frac{87}{60} \approx 1.45.$$

The variance of R_F is $Var(R_F) = E(R_F^2) - E(R_F)^2$. By LOTUS,

$$E(R_F^2) = \sum_{k=1}^{3} k^2 \cdot P(R_F = k) = 1 \cdot \frac{37}{60} + 2^2 \cdot \frac{19}{60} + 3^3 \cdot \frac{4}{60} = \frac{149}{60},$$

therefore,

$$Var(R_F) = \frac{149}{60} - \left(\frac{87}{60}\right)^2 = \frac{1371}{3600} \approx 0.381.$$

2. Denote the a randomly chosen child's birth rank with R_C. We know that there are 20 families with 3 children, so 20 thirdborn children, in addition 50 families with 2 children so in total $50 + 20 = 70$ secondborn children, and 30 families with only one child, so $30 + 50 + 20 = 100$ first born children. Therefore, we have $100 + 70 + 20 = 190$ children in total. By the naive definition of probability the PMF is

$$\begin{aligned}
P(R_C = 1) &= \frac{100}{190} \\
P(R_C = 2) &= \frac{70}{190} \\
P(R_C = 3) &= \frac{20}{190},
\end{aligned}$$

and $P(R_C = k) = 0$ for all other values of k.

We calculate the expectation as in part 1.

$$E(R_C) = \sum_{k=1}^{3} k \cdot P(R_C = k) = 1 \cdot \frac{100}{190} + 2 \cdot \frac{70}{190} + 3 \cdot \frac{20}{190} = \frac{300}{190} \approx 1.579.$$

Similarly, $Var(R_C) = E(R_C^2) - E(R_C)^2$ and by LOTUS

$$E(R_C^2) = \sum_{k=1}^{3} k^2 \cdot P(R_C = k) = 1 \cdot \frac{100}{190} + 2^2 \cdot \frac{70}{190} + 3^2 \cdot \frac{20}{190} = \frac{560}{190},$$

then

$$Var(R_C) = \frac{560}{190} - \left(\frac{300}{190}\right)^2 = \frac{16400}{36100} \approx 0.454$$

(Mildly) interestingly, if we are selecting a family first and then the child from the selected family, on average the rank of the selected child is lower, then if we are selecting a child at random. However, the variance of the rank is lower in the former case. So the imprecise definition of "random selection" can make a difference.

Exercise 5. Let $X \sim Bin(100, 0.9)$. 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.

1. Is it possible to have $Y \sim Pois(0.01)$ with $P(X \geq Y) = 1$?
2. **Optional Challenging Exercise:** Is it possible to have $Y \sim Bin(100, 0.5)$ with $P(X \geq Y) = 1$?

Solution 5. 1. No it is not possible. X can only take values in $\{0, \ldots, 100\}$ while Y can take any non-negative integer values, $0, 1, 2, \ldots$. Therefore

$$P(X \geq Y) = 1 - P(X < Y) \leq 1 - P(100 < Y).$$

As Y can take any non-negative integer values $P(100 < Y)$ is non-zero (even though it is very small), therefore $P(X \geq Y) < 1$.

2. **This is a more rigorous argument for why the example works, skip if you just want to see the example for which $P(X \geq Y) = 1$ holds.**

In Section 3.3 of the book, it was discussed that the random variable distributed as $Bin(n, p)$ can be thought of as a sum of n independent Bernoulli random variables, each having a success probability p. So let us rewrite X and Y as $X = X_1 + X_2 \ldots X_{100}$ and $Y = Y_1 + Y_2 \ldots Y_{100}$, where all X_i and Y_i are i.i.d $Bern(0.9)$ and $Bern(0.5)$ respectively (the X_i-s are not necessarily independent from the Y_i-s). Note that the event $\{X \geq Y\} = \{(X_1 + X_2 \ldots X_{100}) \geq (Y_1 + Y_2 \ldots Y_{100})\}$ is implied by the intersection of the events $\{X_1 \geq Y_1\} \cap \{X_2 \geq Y_2\} \cdots \cap \{X_{100} \geq Y_{100}\}$. In words, if for two sums of the same length, the elements of one of the sums are always greater than or equal to the corresponding elements of the other sum, then the first sum itself will be greater than or equal to the second sum (the converse is not true, $6 + 2 \geq 3 + 4$, but $6 \geq 3$, $2 \leq 4$). Therefore, $P(X \geq Y) \geq P(\{X_1 \geq Y_1\} \cap \{X_2 \geq Y_2\} \cdots \cap \{X_{100} \geq Y_{100}\})$. Note that as X_i-s are i.i.d and the Y_i-are i.i.d. also the events $\{X_i \geq Y_i\}$ are independent from each other, hence $P(\{X_1 \geq Y_1\} \cap \{X_2 \geq Y_2\} \cdots \cap \{X_{100} \geq Y_{100}\}) = P(\{X_1 \geq Y_1\})P(\{X_2 \geq Y_2\}) \ldots P(\{X_{100} \geq Y_{100}\})$. Thus if we can define X_i and Y_i such that $P(X_i \geq Y_i) = 1$, then $P(\{X_1 \geq Y_1\}) \cdot P(\{X_2 \geq Y_2\}) \cdot \cdots \cdot P(\{X_{100} \geq Y_{100}\}) = 1 \leq P(X \geq Y)$, and as probabilities are between 0 and 1, it follows that $P(X \geq Y) = 1$

This is the example:

$P(X_i \geq Y_i) = 1$ if $Y_i = 1 \implies X_i$. Thus we can define Y_i as

$$X_i = \begin{cases} 1, & \text{with probability 1 if } Y_i = 1, \\ 1, & \text{with probability } p \text{ if } Y_i = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Now we just have to find p, such that $X_i \sim Bern(0.9)$. We know that $P(Y_i = 1) = 0.5$ and by the law of total probability

$$P(X_i = 1) = P(X_i = 1|Y_i = 1)P(Y_i = 1) + P(X_i = 1|Y_i = 0)P(Y_i = 0) = 1 \cdot 0.5 + p \cdot 0.5.$$

Thus $0.9 = 0.5 + 0.5p$ so it follows that $p = 0.8$

With this specific structure of dependence, we can guarantee that each Bernoulli trial X_i is a success whenever Y_i is a success, therefore the total number of successes for X is greater than or equal to the total number of successes for Y (equal whenever $Y_i = 1$ for all i), such that the marginal distributions for X and Y are as specified in the exercise.

Exercise 6. Ten million people enter a certain lottery. For each person, the chance of winning is one in ten million, independently.

1. Find a simple, good approximation for the PMF of the number of people who win the lottery.
2. Congratulations! You won the lottery. However, there may be other winners. Assume now that the number of winners other than you is $W \sim Pois(1)$, and that if there is more than one winner, then the prize is awarded to one randomly chosen winner. Given this information, find the probability that you win the prize (simplify).

Solution 6. 1. Let X be the number of people who win. Then X has a distribution of $Binom(n, p)$, where $n = 1/p = 10^7$. Then

$$E(X) = np = \frac{10^7}{10^7} = 1.$$

A Poisson approximation is very good here since X is the number of "successes" for a very large number of independent trials where the probability of success on each trial is very low. So X is approximately $Pois(1)$, and for k a nonnegative integer, $P(X = k) \approx \frac{1}{e \cdot k!}$

2. Let A be the event that you win the prize, and condition on W:

$$P(A) = \sum_{k=0}^{\infty} P(A|W = k)P(W = k) = \frac{1}{e}\sum_{k=0}^{\infty} \frac{1}{k+1}\frac{1}{k!}$$
$$= \frac{1}{e}\sum_{k=0}^{\infty} \frac{1}{(k+1)!} = \frac{e-1}{e} = 1 - \frac{1}{e}.$$

(If the last but one equation is unclear, see the Math Appendix A.8.3: Taylor series for e^x.)

Exercise 7. In a group of 90 people, find a simple, good approximation for the probability that there is at least one pair of people such that they share a birthday and their biological mothers share a birthday (i.e. both of the two people have their birthdays at *date x* and both of the moms at *date y*). Assume that no one among the 90 people is the biological mother of another one of the 90 people, nor do two of the 90 people have the same biological mother. Express your answer as a fully simplified fraction in the form a/b, where a and b are positive integers and $b \leq 100$.

Make the usual assumptions as in the birthday problem. To simplify the calculation, you can use the approximations $365 \approx 360$ and $89 \approx 90$, and the fact that $e^x \approx 1 + x$ for $x \approx 0$.

Solution 7. For each pair of people, they have the same birthday with probability 1/365. Similarly, for each pair of mothers, they have the same birthday with probability 1/365. Since the (calendar) date of the daughters/sons and the mothers are independent, it follows that the probability for each pair that both the "descendants" and the mothers share their birthdays is $1/365 \cdot 1/365$. As they are $\binom{90}{2}$ pairs of people, and the pairs of people also determine the pairs of mothers (everyone has exactly one biological mother), from the Poisson paradigm, the number X of people and mother birthday matches is approximately distributed as $Pois(\lambda)$, where $\lambda \approx \frac{90 \cdot 90}{2} \frac{1}{360 \cdot 360} = \frac{1}{32}$, where we used the approximations given in the exercise. The probability that at least one pair of people and the corresponding pair of biological mothers share their birthdays is

$$P(X \geq 1) = 1 - P(X = 0) \approx 1 - e^{-\lambda}.$$

By the second approximation given in the exercise $e^{-\lambda} \approx 1 - \frac{1}{32}$, hence

$$P(X \geq 1) \approx 1 - (1 - \frac{1}{32}) = \frac{1}{32} = 0.03125.$$

.

Using `R` we can get an exact answer for this question by `pbirthday(90, classes=365*365)`, as we choose a pair from 90 people and we have 365^2 classes induced by the composition of "descendants" and biological mothers. This gives 0.02962109, so we are not too far off, given that we used approximations at multiple steps, and the Poisson distribution was an approximation in itself.