Stat 200A, Fall 2018
A. Adhikari and Z. Bartha

WEEK 4 EXERCISES

You are expected to do all these problems, but for **Homework 4** please turn in **only Problems 2, 3, 5, and 6** on **Thursday September 20 at the start of lecture**.

1. Flattened Die

Let $p \in (0,1)$ and let X be the number of spots showing on a flattened die that shows its six faces according to the following chances:

- $P(X=1) = P(X=6)$
- $P(X=2) = P(X=3) = P(X=4) = P(X=5)$
- $P(X = 1 \text{ or } 6) = p$

(a) Find $E(X)$.

$\mathbb{P}(X=1) = \mathbb{P}(X=6) = p/2$ and $\mathbb{P}(X=2) = \ldots = \mathbb{P}(X=5) = (1-p)/4$. Hence,

$$E(X) = \sum_{k=1}^{6} \mathbb{P}(X=k) \cdot k = \frac{p}{2} \cdot 1 + \frac{1-p}{4} \cdot 2 + \frac{1-p}{4} \cdot 3 + \frac{1-p}{4} \cdot 4 + \frac{1-p}{4} \cdot 5 + \frac{p}{2} \cdot 6 = 3.5,$$

which you we also see by symmetry.

(b) Find $SD(X)$. Explain algebraically and also by an intuitive argument why the answer is an increasing function of p.

$$E(X^2) = \sum_{k=1}^{6} \mathbb{P}(X=k) \cdot k^2 = \frac{p}{2} \cdot 1^2 + \frac{1-p}{4} \cdot 2^2 + \frac{1-p}{4} \cdot 3^2 + \frac{1-p}{4} \cdot 4^2 + \frac{1-p}{4} \cdot 5^2 + \frac{p}{2} \cdot 6^2 = 5p + 13.5,$$

hence

$$SD(X) = \sqrt{E(X^2) - (EX)^2} = \sqrt{5p + 13.5 - 3.5^2} = \sqrt{5p + 1.25}.$$

The fact that it is increasing in p is intuitively explained by that the larger p is, the more probability the distribution has on its extremal values (1 and 6) and the less on the intermediate ones.

2. Poisson Expectations

Let X have the Poisson (μ) distribution. Find

(a) $E(X+1)$

$E(X+1) = E(X) + 1 = \mu + 1$

(b) $E(1/(X+1))$

$$\begin{aligned} E(1/(X+1)) &= \sum_{k=0}^{\infty} \mathbb{P}(X=k) \cdot \frac{1}{k+1} = \sum_{k=0}^{\infty} \frac{\mu^k}{k!(k+1)} e^{-\mu} \\ &= \frac{1}{\mu} \sum_{k=0}^{\infty} \frac{\mu^{k+1}}{(k+1)!} e^{-\mu} = \frac{1}{\mu} \sum_{l=1}^{\infty} \frac{\mu^l}{l!} e^{-\mu} = \frac{1}{\mu}(1 - e^{-\mu}) \end{aligned}$$

3. Collecting Distinct Values

(a) A fair die is rolled n times. Find the expected number of times the face with six spots appears.

With $A_i = \{6 \text{ appears on the } i\text{th roll}\}$, $i = 1, 2, \ldots, n$, we have

$$E(\#\text{ of 6's in } n \text{ rolls}) = E\left[\sum_{i=1}^{n} \mathbf{1}_{A_i}\right] = \sum_{i=1}^{n} \mathbb{P}(A_i) = \frac{n}{6}.$$

(b) A fair die is rolled n times. Find the expected number of faces that *do not* appear, and say what happens to this expectation as n increases.

Now with $B_i = \{\text{face } i \text{ does not appear on any of the } n \text{ rolls}\}$, $i = 1, 2, \ldots, 6$, we have

$$E(\#\text{ of faces that do not appear}) = E\left[\sum_{i=1}^{6} \mathbf{1}_{B_i}\right] = \sum_{i=1}^{6} \mathbb{P}(B_i) = 6\left(\frac{5}{6}\right)^n.$$

This decreases exponentially with n.

(c) Use your answer to Part **b** to find the expected number of distinct faces that *do* appear in n rolls of a die.

$$E(\#\text{ of faces that do appear}) = E(6 - \#\text{ of faces that do not appear}) = 6 - 6\left(\frac{5}{6}\right)^n.$$

(d) Find the expected number of times you have to roll a die till you have seen all of the faces. This is a version of what is known as the *collector's problem*. The collector is waiting to get a complete set.

Let $T_1 = 1$, T_2 be the number of rolls we need after the first one to see a new face, T_3 be the number of further rolls we need to see a third face, and so on until T_6. Then the total number of rolls until we get all the faces at least once is $T = T_1 + T_2 + T_3 + T_4 + T_5 + T_6$.

Now notice that $T_2 \sim Geo(5/6)$, $T_3 \sim Geo(4/6), \ldots, T_6 \sim Geo(1/6)$. Therefore

$$E(T) = \sum_{i=1}^{6} E(T_i) = 1 + \frac{6}{5} + \frac{6}{4} + \frac{6}{3} + \frac{6}{2} + \frac{6}{1} = 14.7.$$

4. Aces and Face Cards

A standard deck consists of 52 cards of which 4 are aces, 4 are kings, and 12 (including the four kings) are "face cards" (Jacks, Queens, and Kings).

Cards are dealt at random without replacement from a standard deck till all the cards have been dealt.

Find the expectation of the following. None of them requires a long calculation.

(a) the number of aces among the first 5 cards

With $A_i = \{\text{the } i\text{th card is an ace}\}$, $i = 1, 2, \ldots, 5$,

$$E(\#\text{ of aces among the first 5 cards}) = E\left[\sum_{i=1}^{5} \mathbf{1}_{A_i}\right] = \sum_{i=1}^{5} \mathbb{P}(A_i) = 5 \cdot \frac{4}{52} = \frac{5}{13}.$$

(b) the number of face cards that *do not* appear among the first 13 cards

With $B_i = \{\text{the } i\text{th card is a face card}\}$, $i = 1, 2, \ldots, 13$,

$$E(\#\text{ of face cards among the first 13 cards}) = E\left[\sum_{i=1}^{13} \mathbf{1}_{B_i}\right] = \sum_{i=1}^{13} \mathbb{P}(B_i) = 13 \cdot \frac{12}{52} = 3.$$

Now

$$\begin{aligned} E(\#\text{ of face cards } \textit{not} \text{ among the first 13 cards}) &= E(12 - \#\text{ of face cards among the first 13 cards}) \\ &= 12 - 3 = 9. \end{aligned}$$

(c) the number of aces among the first 5 cards minus the number of kings among the last 5 cards

This expectation is

$$E(\#\text{ of aces among the first 5 cards}) - E(\#\text{ of kings among the last 5 cards}) = 0$$

by the symmetric role of the first / last 5 cards, and that of the aces / kings.

(d) the number of cards before the first ace

Let $D_i = \{\text{card } i \text{ comes before the first ace}\}$, $i = 1, 2, \ldots, 48$ with some arbitrary numbering of the 48 non-ace cards. Then

$$E(\#\text{ of cards before the first ace}) = E\left[\sum_{i=1}^{48} \mathbf{1}_{D_i}\right] = \sum_{i=1}^{48} \mathbb{P}(D_i) = \frac{48}{5},$$

since there are 5 possible positions of card i relative to the four aces, all of which occur with equal probability, so $\mathbb{P}(D_i) = 1/5$.

(e) the number of cards strictly in between the first ace and the last ace

$$\begin{aligned} &E(\#\text{ of cards between the first ace and the last ace}) \\ &\quad = E(50 - \#\text{ of cards before the first ace} - \#\text{ of cards after the last ace}) \\ &\quad = 50 - E(\#\text{ of cards before the first ace}) - E(\#\text{ of cards after the last ace}) \\ &\quad = 50 - \frac{48}{5} - \frac{48}{5} \end{aligned}$$

by part (d) and the symmetric role of the cards before the first ace / cards after the last ace.

(f) the number of face cards before the first ace

Similarly to part (d), let $F_i = \{\text{face card } i \text{ comes before the first ace}\}$, $i = 1, 2, \ldots, 12$ with some arbitrary numbering of the 12 face cards. Then

$$E(\#\text{ of face cards before the first ace}) = E\left[\sum_{i=1}^{12} \mathbf{1}_{F_i}\right] = \sum_{i=1}^{12} \mathbb{P}(F_i) = \frac{12}{5}.$$

5. The Gamma Densities

In this problem you will start with some calculus exercises and then develop one of the fundamental families of densities.

(a) The *Gamma function* of mathematics is defined by

$$\Gamma(r) \;=\; \int_0^\infty t^{r-1}e^{-t}dt, \qquad r > 0$$

That letter is the upper case Greek letter Gamma. You can assume that the integral converges and that therefore $\Gamma(r)$ is a positive number.

Use integration by parts to show that

$$\Gamma(r+1) \;=\; r\Gamma(r), \quad r > 0$$

$$\Gamma(r+1) = \int_0^\infty t^r e^{-t}dt = \left[t^r(-e^{-t})\right]_{t=0}^\infty + \int_0^\infty rt^{r-1}e^{-t}dt = r\Gamma(r).$$

(b) Use Part **a** and induction to show that if r is a positive integer then $\Gamma(r) = (r-1)!$. This is an indication that the Gamma function is a continuous extension of the factorial function.

For $r = 1$:

$$\Gamma(1) = \int_0^\infty e^{-t}dt = 1 = 0!$$

so the formula $\Gamma(r) = (r-1)!$ is correct for $r = 1$. The inductive step is shown by part (a).

(c) Let X have density given by

$$f_X(t) = \begin{cases} \frac{1}{\Gamma(r)}t^{r-1}e^{-t}, & t > 0 \\ 0 & \text{otherwise} \end{cases}$$

We say that X has the *gamma* $(r, 1)$ density. In the case $r = 1$, this should be a density that you recognize. Provide its name and the appropriate parameters.

Exp(1)

(d) Now fix $\lambda > 0$ and consider the function

$$f(t) = \begin{cases} \frac{\lambda^r}{\Gamma(r)}t^{r-1}e^{-\lambda t}, & t > 0 \\ 0 & \text{otherwise} \end{cases}$$

Show that this function is a density. It is called the gamma (r, λ) density.

f is clearly non-negative, so we only have to show that it integrates to 1. This is shown by

$$\int_0^\infty \lambda^r t^{r-1}e^{-\lambda t}dt = \int_0^\infty (\lambda t)^{r-1}e^{-\lambda t}\lambda dt = \int_0^\infty s^{r-1}e^{-s}ds = \Gamma(r)$$

with the substitution $s = \lambda t$.

(e) Use Part **d** to fill in the blank: For every $r > 0$,

$$\int_0^\infty t^{r-1}e^{-\lambda t}dt \;=\; \underline{\qquad \frac{\Gamma(r)}{\lambda^r} \qquad}$$

Now let Y have the gamma (r, λ) density. Use what you have shown in the previous parts to find $E(Y)$. Provide the numerical value of $E(Y)$ in the case $r = 2.2$ and $\lambda = 1.1$. You don't need a computer.

$$E(Y) = \int_0^\infty \frac{\lambda^r}{\Gamma(r)} t^{r-1} e^{-\lambda t} t \, dt = \frac{\lambda^r}{\Gamma(r)} \int_0^\infty t^r e^{-\lambda t} dt = \frac{\lambda^r}{\Gamma(r)} \cdot \frac{\Gamma(r+1)}{\lambda^{r+1}} = \frac{r}{\lambda}$$

(f) Let Y have gamma (r, λ) density. Find $SD(Y)$.

$$E(Y^2) = \int_0^\infty \frac{\lambda^r}{\Gamma(r)} t^{r-1} e^{-\lambda t} t^2 \, dt = \frac{\lambda^r}{\Gamma(r)} \int_0^\infty t^{r+1} e^{-\lambda t} dt = \frac{\lambda^r}{\Gamma(r)} \cdot \frac{\Gamma(r+2)}{\lambda^{r+2}} = \frac{(r+1)r}{\lambda^2},$$

hence

$$SD(Y) = \sqrt{E(Y^2) - (EY)^2} = \sqrt{\frac{(r+1)r}{\lambda^2} - \frac{r^2}{\lambda^2}} = \sqrt{\frac{r}{\lambda^2}} = \frac{\sqrt{r}}{\lambda}.$$

6. Bounds

A random variable X, not necessarily non-negative, has $E(X) = 20$ and $SD(X) = 4$. For the bounds below, use 0 and 1 only if you feel you can't do better than those.

(a) Find upper and lower bounds for $P(0 < X < 40)$.

The best upper bound is 1, since there is a random variable with the given properties that makes this probability 1. For example, if $\mathbb{P}(X = 16) = \mathbb{P}(X = 24) = 0.5$.

For a lower bound we can use Chebyshev's inequality:

$$\mathbb{P}(0 < X < 40) = \mathbb{P}(|X - 20| < 20) = 1 - \mathbb{P}(|X - 20| \geq 20) \geq 1 - \frac{Var(X)}{20^2} = 1 - \frac{16}{400} = 1 - \frac{1}{25}.$$

(b) Find upper and lower bounds for $P(10 < X < 40)$.

The same example as above shows that the best upper bound is 1. For a lower bound, aiming for the application of Chebyshev's inequality again:

$$\mathbb{P}(10 < X < 40) \geq \mathbb{P}(10 < X < 30) = \mathbb{P}(|X-20| < 10) = 1-\mathbb{P}(|X-20| \geq 10) \geq 1-\frac{Var(X)}{10^2} = 1-\frac{16}{100} = 1-\frac{4}{25}.$$

(c) Find an upper bound for $P(X \geq 40)$.

$$\mathbb{P}(X \geq 40) \leq \mathbb{P}(X \geq 40 \text{ or } X \leq 0) = \mathbb{P}(|X - 20| \geq 20) \leq \frac{Var(X)}{20^2} = \frac{1}{25},$$

as in (a).

(d) Find an upper bound for $P(X^2 \geq 900)$.

By Markov's inequality

$$\mathbb{P}(X^2 \geq 900) \leq \frac{E(X^2)}{900} = \frac{Var(X) + (EX)^2}{900} = \frac{16 + 400}{900} = 0.4622.$$

Alternatively, we can use Chebyshev's inequality in the following way to get a better bound:

$$\mathbb{P}(X^2 \geq 900) = \mathbb{P}(X \geq 30 \text{ or } X \leq -30) \leq \mathbb{P}(X \geq 30 \text{ or } X \leq 10) = \mathbb{P}(|X - 20| \geq 10) \leq \frac{Var(X)}{10^2} = \frac{16}{100}.$$

7. Randomized Response

Survey respondents understandably don't like to answer questions about sensitive topics such as illegal drug use. If data scientists want to estimate the proportion of illegal drug users in a population, they have to devise methods of getting the information they need while maintaining the privacy of the individual respondents.

Randomized response schemes are often used in such situations. In one such scheme, each surveyed person is given a coin and asked to answer YES or NO after following these instructions out of sight of the surveyor:

- Toss the coin.
- If it lands heads, then truthfully answer, "Do you use illegal drugs?"
- If it lands tails, then toss it again and answer, "Did the second toss land heads?"

This way each respondent answers YES or NO but the surveyor doesn't know which question was answered. The data scientists then have to estimate the proportion of illegal drug users based on the overall proportion of YES answers, which includes the YES answers to the second question.

Let the unknown proportion of illegal drug users in a large population be p, and suppose a random sample of size n is surveyed using the scheme above. You can assume that the sampling is equivalent to drawing at random with replacement.

(a) Let X be the proportion of sampled people who answer YES. Find $E(X)$.

Let A_i be the event that the ith person in the sample answers YES, $i = 1, 2, \ldots, n$. Then $\mathbb{P}(A_i) = \frac{1}{2} \cdot p + \frac{1}{2} \cdot \frac{1}{2}$, separating the cases according to the first toss. Therefore,

$$E(X) = \frac{1}{n} E(nX) = \frac{1}{n} E(\#\text{ of people who answered YES}) = \frac{1}{n} E\left[\sum_{i=1}^{n} \mathbf{1}_{A_i}\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(A_i) = \frac{p}{2} + \frac{1}{4}.$$

(b) Use X to construct an unbiased estimate of p.

By the above, $p = 2(E(X) - 1/4) = E(2X - 1/2)$, thus $2X - 1/2$ is an unbiased estimate of p.

8. Indicators and the Inclusion-Exclusion Formula

We guessed the general inclusion-exclusion formula (see Section 5.2 of the Prob 140 textbook) but we never proved it. Let's get that done.

(a) Let $x_1, x_2, \ldots, x_n$ be numbers. Expand the product $(1-x_1)(1-x_2)$ and then expand $(1-x_1)(1-x_2)(1-x_3)$ by using the expansion you got for $(1 - x_1)(1 - x_2)$. Now guess a formula for the expansion of the product

$$\prod_{i=1}^{n} (1 - x_i)$$

and use induction to prove it. The induction shouldn't take many steps. It consists of just two observations, both of which can be expressed in English without complicated notation.

$$\begin{aligned}
(1 - x_1)(1 - x_2) &= 1 - (x_1 + x_2) + x_1 x_2 \\
(1 - x_1)(1 - x_2)(1 - x_3) &= 1 - (x_1 + x_2 + x_3) + (x_1 x_2 + x_1 x_3 + x_2 x_3) - x_1 x_2 x_3 \\
&\vdots \\
\prod_{i=1}^{n} (1 - x_i) &= 1 - \sum_{i=1}^{n} x_i + \sum_{1 \le i < j \le n} x_i x_j - \sum_{1 \le i < j < k \le n} x_i x_j x_k + \ldots + (-1)^n x_1 x_2 \ldots x_n
\end{aligned}$$

To prove the general formula we need the following inductive step:

$$\prod_{i=1}^{n+1} (1 - x_i) = (1 - x_{n+1}) \prod_{i=1}^{n} (1 - x_i) = \prod_{i=1}^{n} (1 - x_i) - x_{n+1} \prod_{i=1}^{n} (1 - x_i), \tag{1}$$

that is, expanding by the last factor $1 - x_{n+1}$. If we plug in the formula for the n-fold product into the right-hand side of (1), $\prod_{i=1}^{n}(1 - x_i)$ will give all the terms that do not contain x_{n+1}, with the correct sign, and $-x_{n+1} \prod_{i=1}^{n}(1 - x_i)$ will give all the terms that contain x_{n+1}, with the correct sign.

(b) Let $A_1, A_2, \ldots, A_n$ be events. For each i in the range 1 through n let I_i be the indicator of A_i. Let I be the indicator of $\cup_{i=1}^{n} A_i$. Explain why

$$I \;=\; 1 - \prod_{i=1}^{n}(1 - I_i)$$

$$1 - I = \mathbf{1}_{(\cup_{i=1}^n A_i)^c} = \mathbf{1}_{\cap_{i=1}^n A_i^c} = \prod_{i=1}^{n} \mathbf{1}_{A_i^c} = \prod_{i=1}^{n}(1 - I_i)$$

(c) Use Parts **a** and **b** to establish the inclusion-exclusion formula.

$$\begin{aligned}
\mathbb{P}(\cup_{i=1}^n A_i) = E(I) &\overset{\text{part (b)}}{=} E\left[1 - \prod_{i=1}^{n}(1 - I_i)\right] = 1 - E\left[\prod_{i=1}^{n}(1 - I_i)\right] \\
&\overset{\text{part (a)}}{=} 1 - E\left[1 - \sum_{i=1}^{n} I_i + \sum_{1\le i<j\le n} I_i I_j - \sum_{1\le i<j<k\le n} I_i I_j I_k + \ldots + (-1)^n I_1 I_2 \ldots I_n\right] \\
&= 1 - 1 + \sum_{i=1}^{n} E(I_i) - \sum_{1\le i<j\le n} E(I_i I_j) + \sum_{1\le i<j<k\le n} E(I_i I_j I_k) - \ldots + (-1)^{n+1} E(I_1 I_2 \ldots I_n) \\
&= \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{1\le i<j\le n} \mathbb{P}(A_i \cap A_j) + \sum_{1\le i<j<k\le n} \mathbb{P}(A_i \cap A_j \cap A_k) - \ldots + (-1)^{n+1} \mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n)
\end{aligned}$$