MAST20004 Probability
Semester 1, 2021
Assignment 2: Questions

Due 3 pm, Friday 16 April 2021

April 9, 2021

Name: James La Fontaine

Student ID: 1079860

Important instructions:

(1) This assignment contains 5 questions, **two** of which will be randomly selected to be marked. Each marked question is worth 10 points and each unmarked question with substantial working is worth 1 point.

(2) To complete this assignment, you need to write your solutions into the blank answer spaces following each question in this assignment PDF.

- If you have a printer (or can access one), then you must print out the assignment template and handwrite your solutions into the answer spaces.
- If you do not have a printer but you can figure out how to annotate a PDF using an iPad/Android tablet/Graphics tablet or using Adobe Acrobat, then annotate your answers directly onto the assignment PDF and save a copy for submission.

Failing both of these methods, you may handwrite your answers as normal on blank paper and then scan for submission (but note that you will thereby miss valuable practice for the exam process). In that case, however, your document should have the same length as the assignment template otherwise Gradescope will reject your submission. So you will need to add as many blank pages as necessary to reach that criterion.

Scan your assignment to a PDF file using your mobile phone (we recommend Cam - Scanner App), then upload by going to the Assignments menu on Canvas and submit the PDF to the **GradeScope tool** by first selecting your PDF file and then clicking on 'Upload PDF'.

Note that here you do not need to submit any Matlab code with your assignment.

(3) A poor presentation penalty of 10% of the total available marks will apply unless your submitted assignment meets all of the following requirements:

- it is a single pdf with all pages in correct template order and the correct way up, and with any blank pages with additional working added only at the end of the template pages;
- has all pages clearly readable;
- has all pages cropped to the A4 borders of the original page and is imaged from directly above to avoid excessive 'keystoning'.

These requirements are easy to meet if you use a scanning app on your phone and take some care with your submission - please review it before submitting to double check you have satisfied all of the above requirements.

(4) Late submission within 20 hours after the deadline will be penalised by 5% of the total available marks for every hour or part thereof after the deadline. After that, the Gradescope submission channel will be closed, and your submission will no longer be accepted. You are strongly encouraged to submit the assignment a few days before the deadline just in case of unexpected technical issues. If you are facing a rather exceptional/extreme situation that prevents you from submitting on time, please contact the tutor coordinator **Robert Maillardet** with formal proofs such as medical certificate.

(5) Working and reasoning must be given to obtain full credit. Clarity, neatness, and style count.

Q1. A random variable N has pmf $p_N(k) = \frac{c}{k(k+1)}$ where $k = 2, 3, 4, \ldots$, and c is a constant.

(a) Derive the value of constant c.

$\sum_{R=2}^{\infty} P_N(R) = 1$ (pmf property), $R \geq 2$

$\lim_{n\to\infty} \sum_{R=2}^{n} \frac{c}{R(R+1)} = 1$

$= c \sum_{R=2}^{n} \frac{1}{R(R+1)}$ $\left(\frac{1}{R(R+1)} = \frac{1}{R} - \frac{1}{R+1},\ R \neq 0, -1\right)$

$\Rightarrow c\left(\frac{1}{2\cdot 3} + \frac{1}{3\cdot 4} + \frac{1}{4\cdot 5} + \ldots + \frac{1}{(n+1)n}\right) = 1$

$\Rightarrow c\left(\frac{1}{2} - \frac{1}{3} + \frac{1}{3} - \frac{1}{4} + \frac{1}{4} - \frac{1}{5} + \ldots + \frac{1}{n} - \frac{1}{n+1}\right) = 1$

$\Rightarrow \frac{c}{2} - \frac{c}{n+1} = 1$ (a)

$\Rightarrow \lim_{n\to\infty} \frac{c}{2} - \frac{c}{n+1} = 1$

$\Rightarrow \frac{c}{2} = 1$

$\Rightarrow c = 2$

(b) Find a simple expression for the cdf $F_N(x)$ of N (that is, you should simplify any potential sum appearing in the expression).

$$F_N(x) = P[X \le x] = \sum_{\substack{k \in S_N \\ k \le x}} P_N(k)$$

$$= \sum_{k=2}^{x} \frac{2}{k(k+1)}$$

$$\overset{(a)}{=} 1 - \frac{2}{x+1}$$

$$\Rightarrow F_N(x) = 1 - \frac{2}{x+1}, \quad x \neq -1$$

(c) Find the mean of N, if it exists. If it does not exist, prove that it does not exist.

$$E[N] = \sum_{k \in S_N} k \cdot p_N(k)$$

$$= \lim_{n\to\infty} \sum_{k=2}^{n} \frac{2k}{k(k+1)}$$

$$= \lim_{n\to\infty} \sum_{k=2}^{n} \frac{2}{k+1}$$

$$= 2 \cdot \lim_{n\to\infty} \sum_{k=2}^{n} \frac{1}{k} = \infty$$ (harmonic series diverges)

$\Rightarrow$ the sum does not converge absolutely

$\Rightarrow E[N]$ does not exist

(d) Find the standard deviation of N, if it exists. If it does not exist, prove that it does not exist.

$Sd(N) = \sqrt{V(N)}$

$V(N) = E(N^2) - E(N)^2$

$E(N)$ does not exist from (c)

$\Rightarrow V(N)$ does not exist

$\Rightarrow Sd(N)$ does not exist

(e) Let $Y = (-1)^N N$. What are the possible values of Y? Does the mean of Y exist? Justify your answer.

$S_N = 2, 3, 4$

$S_Y = (-1)^2 \cdot 2, (-1)^3 \cdot 3, (-1)^4 \cdot 4, \ldots$

$S_Y = 2, -3, 4$

Therefore,

$$E[Y] = \sum_{y \in S_Y} y \cdot P_Y(y)$$

$$\Rightarrow \sum_{y \in S_Y} |y| \cdot P_Y(y) = \infty \quad \text{as } |y| \text{ is increasing infinitely}$$

$\Rightarrow$ the series does not converge absolutely

$\Rightarrow E[Y]$ does not exist

Q2 A random variable $0 < X < 1$ has density $f_X(x) = C_a x^a(1-x)$ for some constant C_a, where a is a real positive parameter.

(a) What is the value of the constant C_a?

$$\int_0^1 C_a x^a(1-x)\,dx = 1, \quad \text{pdf property}$$

$$= C_a \int_0^1 x^a(1-x)\,dx$$

$$= C_a \int_0^1 x^a - x^{a+1}\,dx$$

$$= C_a \left[\frac{x^{a+1}}{a+1} - \frac{x^{a+2}}{a+2}\right]_0^1 \quad (a)$$

$$= C_a \left[\frac{1}{a+1} - \frac{1}{a+2}\right]$$

$$= C_a \left(\frac{1}{(a+1)(a+2)}\right)$$

$$\Rightarrow \frac{C_a}{(a+1)(a+2)} = 1 \Rightarrow C_a = (a+1)(a+2)$$

(b) What is the CDF of X?

$$F_X(x) = \int_0^x f(x)\,dx$$

$$= C_\alpha \int_0^x x^\alpha (1-x)\,dx$$

$$\overset{(a)}{=} C_\alpha \left[\frac{x^{\alpha+1}}{\alpha+1} - \frac{x^{\alpha+2}}{\alpha+2}\right]_0^x$$

$$= \frac{(\alpha+1)(\alpha+2)\,x^{\alpha+1}}{\alpha+1} - \frac{(\alpha+1)(\alpha+2)\,x^{\alpha+2}}{\alpha+2}$$

$$= x^{\alpha+1}(\alpha+2) - x^{\alpha+2}(\alpha+1)$$

$$F_X(x) = \begin{cases} x^{\alpha+1}(\alpha+2) - x^{\alpha+2}(\alpha+1), & x \geq 0 \\ 0, & x < 0 \end{cases}$$

(c) For $k = 0, 1, \ldots$, find an expression for $E(X^k)$ in terms of a and k.

$$E[X^k] = \int_0^1 x^k \cdot C_a x^a (1-x)\, dx, \quad k \geq 0$$

$$= C_a \int_0^1 x^k (x^a - x^{a+1})\, dx$$

$$= C_a \int_0^1 x^{k+a} - x^{k+a+1}\, dx$$

$$= C_a \left[\frac{x^{k+a+1}}{k+a+1} - \frac{x^{k+a+2}}{k+a+2} \right]_0^1$$

$$= (a+1)(a+2) \left[\frac{1}{(k+a+1)(k+a+2)} \right]$$

$$= \frac{(a+1)(a+2)}{(a+k+1)(a+k+2)}$$

Q3 The lifetime (in weeks) of a certain system component is thought to have an *exponential* distribution, whose density function is given by

$$f(x) = \lambda e^{-\lambda x}, \qquad x \geq 0,$$

for some strictly positive parameter λ.

(a) The *reliability* of a component is defined as the probability that the component will not have failed by a specified time. If the reliability of the system component at 10.5 weeks is 0.9, find the reliability at 10 weeks.

X = # weeks until failure $\stackrel{d}{=} \exp(\lambda)$, $P(X>x) = e^{-\lambda x}$

$P(X>10.5) = 0.9$

$\Rightarrow e^{-10.5\lambda} = 0.9$

$\Rightarrow \lambda = \dfrac{\log_e(0.9)}{-10.5}$

$\Rightarrow P(X>10) = e^{-10\lambda}$

$= 0.9045 \approx 0.905$

(b) One hundred components of this type are put in a new system. All components that have failed are replaced at 20 week intervals, and none are replaced at other times.

If R is the number of components that have to be replaced at the end of the first interval, state assumptions and compute the mean and variance of R.

Explain why this result holds for any such interval, and not just the first.

X = weeks until failure $\stackrel{d}{=} \exp(\lambda)$

R = # Components that have failed in 20 weeks

$R \stackrel{d}{=} Bi(100, p)$ where $p = P(X \leq 20)$

$= 1 - e^{-20\lambda} \approx 0.181831$

$\Rightarrow E[R] = np = 100 \cdot 0.181831 \approx 18.1831$

$\Rightarrow V[R] = np(1-p) = 18.1831 \cdot 0.818169$

≈ 14.8768

This result holds for any such interval as the probability of a component failing in 20 weeks, p, is determined by the exponential random variable X, so X is memoryless and for every real a, b we have $P(X > a+b \mid X > a) = P(X > b)$ and $P(X \leq b) = 1 - P(X > b)$. This means that everytime we check if a component failed in the last 20 weeks, it doesn't matter whether it has failed before this 20 week interval or not.

Q4 Let X be a continuous non-negative random variable with distribution function $F(x)$ and probability density function $f(x)$ which is such that $f(x) = F'(x)$ for all $x \in [0, \infty)$. Define the *failure rate function* (also called *hazard function*) as

$$r(x) = \frac{f(x)}{1 - F(x)},$$

and the *survival function* as

$$G(x) = 1 - F(x).$$

(a) Show that for a non negative continuous random variable X with failure rate function r, X has survival function

$$G(x) = \exp\left(-\int_0^x r(u)du\right).$$

$F(\infty) = 1$, $P[0 < x < \infty] = 1$

$\Rightarrow F(\infty) - F(0) = 1 \Rightarrow F(0) = 0$

$G(x) = e^{-\int_0^x \frac{f(u)}{1-F(u)}\,du}$ let $v = F(u)$, $\frac{dv}{du} = f(u)$

$\Rightarrow G(x) = e^{-\int_{F(0)}^{F(x)} \frac{1}{1-v}\,dv}$

$= e^{-[-\log_e|1-v|]_{F(0)}^{F(x)}}$

$= e^{-[-\log_e(1-f(x)) + \log_e(1-F(0))]}$, $0 \le F(x) \le 1$

$= e^{\log_e(1-f(x))}$

$= 1 - f(x)$

$= G(x)$ ✓

(b) Consider a certain type of organ in the human body that degrades gradually at a slow but steady rate. To model this, we suppose that its failure rate function is linearly increasing $r(x) = ax$ for some $a > 0$. It has been observed that the median lifetime of the organ is 60 years. What is the probability that such an organ lasts for more than 100 years?

$$F(x) = 1 - G(x)$$
$$= 1 - e^{-\int_0^x au\,du}$$
$$= 1 - e^{-a\left[\frac{u^2}{2}\right]_0^x}$$
$$= 1 - e^{\frac{-ax^2}{2}}$$

$F(60) = 0.5$ as median is 60

$$\Rightarrow 1 - e^{\frac{-a(60)^2}{2}} = 0.5$$
$$\Rightarrow a = \frac{-2\log_e(0.5)}{3600}$$

$$P(X > 100) = 1 - P(X \leq 100)$$
$$= 1 - F(100)$$
$$= 1 - (1 - e^{-(100^2)\cdot a}$$
$$= 1 - (0.8542)$$
$$= 0.1458$$

(c) Find the probability density functions of *non negative continuous random variables* with failure rate functions

(i) $r(x) = x^3$,

(ii) $r(x) = 1/(1+x)$.

(i) $r(x) = x^3$

$f(x) = r(x) \cdot (1 - F(x))$ (i)

$= x^3 \cdot G(x)$

Now $G(x) = e^{-\int_0^x u^3 \, du}$

$= e^{-\left[\frac{u^4}{4}\right]_0^x}$

$= e^{\frac{-x^4}{4}}$

$\Rightarrow f(x) = \frac{x^3}{e^{x^4/4}}$

(ii) $f(x) \overset{(i)}{=} \frac{1}{1+x} \cdot G(x)$

Now $G(x) = e^{-\int_0^x \frac{1}{1+u} \, du}$

$= e^{-[\log_e(1+x) - \log_e(1)]}$

$= e^{-\log_e(1+x)}$, $x \geq 0$

$= \frac{1}{1+x}$

$\Rightarrow f(x) = \frac{1}{1+x} \cdot \frac{1}{1+x} = \frac{1}{(1+x)^2}$

Q5 A bag contains n tokens numbered from 1 to n. A random sample of n tokens is selected from the bag, one at a time (the order matters). A *match* occurs if the token numbered i is selected on the ith draw.

In this exercise, we are going to investigate if selecting the tokens *with replacement* or selecting *without replacement* does affect the probability of at least one match. We will also try to understand to which extent the number of tokens in the bag affects the probability of at least one match.

The Matlab program `Assignment2_Q5_2021.m` simulates `nreps` repetitions of the experiment of drawing n tokens, with and without replacement. Read through the program and make sure you understand what each line is computing.

(a) If the draws are done *with replacement*,

(i) What is the theoretical probability that there is at least one match?

X = # matches in the sample of n tokens

P (of match) $= \frac{1}{n}$, equal probability

$$P(X \geq 1) = 1 - P(X < 1)$$
$$= 1 - P(X = 0)$$
$$= 1 - (1-p)^n$$
$$= 1 - \left(1 - \frac{1}{n}\right)^n$$

(ii) What is the theoretical distribution of the number of matches?

$X \overset{d}{=} Bi(n, \frac{1}{n})$

n Bernoulli trials with $P(\text{success or match}) = \frac{1}{n}$

$\Rightarrow P_X(x) = \binom{n}{x}\left(\frac{1}{n}\right)^x\left(1-\frac{1}{n}\right)^{n-x}$

(iii) If n is large, what distribution can be used to approximate the number of matches? Justify your answer.

If n is large, a Poisson distribution can be used to approximate the number of matches with

$\lambda = np = 1$

$$\lim_{n\to\infty} P_X(x) = \lim_{n\to\infty}\binom{n}{x}\left(\frac{1}{n}\right)^x\left(1-\frac{1}{n}\right)^{n-x}$$

$$= \lim_{n\to\infty}\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \lim_{n\to\infty}\frac{n!}{n^k(n-k)!}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}$$

$$= \frac{\lambda^x}{x!} \times 1 \times e^{-\lambda} \times 1 = \frac{\lambda^x e^{-\lambda}}{x!}$$

$\Rightarrow X \overset{d}{=} Po(\lambda = 1)$ for large n

(iv) Explain what Lines 24, 26 and 38 are computing.

24 - compute number of matches in each experiment with replacement

26 - tracks occurrences of each number of matches (number of times $X = x$ occurs in the experiment with replacement)

38 - divides total occurrences of each number of matches by the number of reps of the experiment to obtain empirical probabilities

(v) Uncomment Line 42 and replace the "??" with the right expression in terms of `pmf_1` to display the empirical probability that there is at least one match. Copy your answer here.

?? → $1 - \text{pmf\_1}(1) = 1 - P(X=0)$

$= 0.692$

(vi) Uncomment Lines 48 and 49 and replace the "??' in Line 48 by the expression you found in (a). Copy your answer here.

$P_1 = 1 - (1 - \frac{1}{n})^n$

≈ 0.672

(b) If the draws are done *without replacement*, it can be shown that

$$\begin{aligned} P(\text{at least one match}) &= 1 - \left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^n}{n!}\right) \\ &= 1 - \sum_{k=0}^{n} \frac{(-1)^k}{k!}. \end{aligned}$$

It is much harder to describe the full distribution of the number of matches in this case.

Lines 51–60 compute and display the theoretical probability that there is at least one match without replacement using the above formula.

Uncomment Line 43 and replace the "??' with the right expression in terms of `pmf_2` to display the empirical probability that there is at least one match. Copy your answer here.

?? $\rightarrow 1 - pmf\_2(1) = 1 - P(X=0)$

≈ 0.633

(e) Draw a table comparing the empirical and theoretical values of the probability that there is at least one match for $n = 5$, $n = 20$, $n = 100$, and $n = 500$, both with and without replacement. You can adjust the value of `nreps` to increase the accuracy. What do you observe as n increases?

P(X>1) nreps = 10000

n	W/ replacement		W/o replacement	
	Emp.	Theor.	Emp.	Theor.
5	0.621	0.622	0.634	0.633
20	0.639	0.642	0.638	0.632
100	0.628	0.634	0.643	0.632
500	0.635	0.632	0.631	0.632

As n increases, probability (at least 1 match) in experiment with replacement → probability (at least 1 match) in experiment without replacement

(d) Prove that, as n increases, the probability that there is at least one match converges to $1 - 1/e$ *in both cases* (with and without replacement). Provide an interpretation of why this is the case.

w/ replacement:

$$\lim_{n\to\infty} P(X \geq 1) = \lim_{n\to\infty} \left(1 - \left(1 - \frac{1}{n}\right)^n\right)$$

$$= 1 - e^{-1}, \quad \left(e^x = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n\right)$$

$$= 1 - \frac{1}{e}$$

w/o replacement:

$$\lim_{n\to\infty} P(X \geq 1) = 1 - \sum_{k=0}^{n} \frac{(-1)^k}{k!}$$

$$= 1 - \lim_{n\to\infty} \sum_{k=0}^{n} \frac{(-1)^k}{k!}$$

$$= 1 - \sum_{k=0}^{\infty} \frac{(-1)^k}{k!}$$

$$= 1 - e^{-1}, \quad \left(e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}\right)$$

$$= 1 - \frac{1}{e}$$

As n becomes larger, sampling with replacement becomes sampling w/o replacement essentially as the probability of selecting the same token more than once becomes very low.

(e) For $n = 100$, display a comparison of the first five values of the approximating pmf in (c) and of the empirical `pmf_1`.

	(a)(iii)	empirical pmf_1
$P(X=0)$	0.3679	0.3622
$P(X=1)$	0.3679	0.3644
$P(X=2)$	0.1839	0.1932
$P(X=3)$	0.0613	0.0639
$P(X=4)$	0.0153	0.0130