Assignment 9 = Exam 2: statistics
Due December 5, 2023

John McCuan

Recall the the two data sets with simulated probabilities from Problem 8 of Assignment 8:

```
 [1] 0.57 0.59 0.68 0.60 0.61 0.69 0.62 0.54 0.57 0.58 0.66 0.62 0.62 0.61 0.57
[16] 0.69 0.62 0.50 0.64 0.58 0.55 0.59 0.55 0.56 0.57 0.52 0.64 0.61 0.54 0.66
[31] 0.62 0.59 0.64 0.64 0.64 0.63 0.63 0.60 0.58 0.58 0.57 0.59 0.54 0.71 0.66
[46] 0.54 0.58 0.58 0.64 0.60 0.61 0.60 0.60 0.67 0.59 0.68 0.52 0.63 0.61 0.61
[61] 0.62 0.57 0.58 0.55 0.55 0.62 0.62 0.60 0.65 0.70 0.58 0.48 0.65 0.56 0.57
[76] 0.65 0.59 0.54 0.61 0.59 0.60 0.62 0.58 0.63 0.59 0.62 0.65 0.62 0.58 0.66
```

and

```
 [1] 0.52 0.58 0.55 0.66 0.54 0.59 0.50 0.62 0.61 0.56 0.72 0.71 0.64 0.57 0.63
[16] 0.61 0.60 0.59 0.59 0.58 0.56 0.66 0.62 0.67 0.52 0.64 0.60 0.70 0.61 0.57
[31] 0.64 0.60 0.58 0.59 0.61 0.46 0.61 0.62 0.59 0.55 0.59 0.59 0.57 0.57 0.46
[46] 0.63 0.62 0.57 0.62 0.66 0.61 0.65 0.61 0.67 0.61 0.60 0.55 0.63 0.58 0.61
[61] 0.53 0.62 0.70 0.55 0.67 0.56 0.60 0.52 0.60 0.60 0.60 0.71 0.61 0.61 0.58
[76] 0.56 0.69 0.57 0.64 0.72 0.67 0.60 0.61 0.62 0.57 0.58 0.57 0.65 0.52 0.57
```

one of which was generated using

```
round(rnorm(90,0.6,0.05),digits = 2)
```

and the other of which was generated using

```
rbinom(90,100,0.6)/100.
```

Problem 1 (Problem 8 of Assignment 8) Which of the following would be useful in identifying which data was generated by which command?

(i) More samples, i.e., replacing `90` with a larger number in each command.

(ii) More decimal places, e.g, using `digits = 3` in the normal sampling.

(iii) Knowing the seed I used to produce one of the data sets.

Problem 2 (Problem 8 of Assignment 8)

(a) Use `R` to plot the PMF associated with the sampling

`rbinom(90,100,0.6)/100`.

(b) Use `R` to plot the MDF associated with the sampling

`rnorm(90,0.6,0.05)`.

(c) Display both plots on the same axes with statistical range $[0, 1]$ and using different symbols for the two different plots.

Problem 3 ()

(a)

(b)

(c)

Problem 4 ()

(a)

(b)

(c)

(d)

Problem 5 ()

(a)

(b)

(c)

Problem 6 (Markov's lemma) The following assertion and "proof" have been attributed to Markov:

> For a continuous random variable Y with $P(Y = y) = f(y)$ and finite, positive real number a, $P(Y \geq a) \leq E[Y]/a$.

Proof:

$$E[Y] = \int_{-\infty}^{\infty} y f(y)\, dy \tag{1}$$

$$\geq \int_{a}^{\infty} y f(y)\, dy \tag{2}$$

$$\geq \int_{a}^{\infty} a f(y)\, dy \tag{3}$$

$$= a \int_{a}^{\infty} f(y)\, dy \tag{4}$$

$$= a \ * \ P(Y \geq a). \tag{5}$$

Dividing both sides by a yields $P(Y \geq a) \leq E[Y]/a$.

There are a number of errors here attributed to Markov. Let us begin by noting that the basic assertion

$$P(Y \geq a) \leq E[Y]/a \tag{6}$$

is false under the stated assumptions. In order to extract the falsehood clearly, we will perhaps need to give some reasonable meaning to the symbols used in (6). This is accomplished by taking

$$P(Y \geq a) = \int_{a}^{\infty} f(y)\, dy$$

and

$$E[Y] = \int_{-\infty}^{\infty} y\, f(y)\, dy.$$

It may in fact be noted that both of these reasonable meanings are used in the first and last lines (1) and (5) in the proof. If the symbols still seem mysterious to you, do not worry, we will come back to consider them in more detail below. For now, note simply that the main assertion (6) amounts to the assertion

$$\int_{a}^{\infty} f(y)\, dy \leq \frac{1}{a} \int_{-\infty}^{\infty} y\, f(y)\, dy \tag{7}$$

for a positive real number a and a probability density $f : \mathbb{R} \to [0, \infty)$.

(a) Take $f : \mathbb{R} \to [0, \infty)$ by

$$f(y) = \frac{1}{\sqrt{\pi}} e^{-(y+1)^2},$$

and show the reverse inequality holds in (7) for any $a > 0$. Show in fact,

$$P(Y \geq a) = \int_a^\infty f(y)\, dy > 0 > -\frac{2}{a\sqrt{\pi}} \int_{-1}^\infty e^{-(y+1)^2}\, dy = \frac{1}{a} E[Y]. \tag{8}$$

Hint: Write

$$\int_{-\infty}^\infty y\, f(y)\, dy = \int_{-\infty}^{-1} y\, f(y)\, dy + \int_{-1}^\infty y\, f(y)\, dy.$$

Use a u-substitution with $u = -2 - y$ in the first integral, and then recombine the integrals to obtain the (negative) quantity on the right in (8).

(b) You might find the calculation of part **(a)** somewhat disturbing. If it is any consolation, Professor Markov would undoubtedly find that such an assertion has been attributed to him rather disturbing as well. The good news is that with some additional hypotheses (and some additional corrections) one can extract a correct assertion out of the statement above. It may be difficult for you to discern what the appropriate hypothesis might be, but I am going to give you a chance to think about it, and maybe come up with it on your own, in this part of the problem:

> What condition may be imposed on the MDF $f : \mathbb{R} \to [0, \infty)$ so that the asserted inequality (7) does hold?

If you can't figure it out, do not worry, the answer is contained in Problem 9 below. For the moment let's attempt to clear up some of the easy points of confusion in the statement attributed to Markov.

(c) The statement starts with the introduction of a "continuous random variable Y." It is perhaps worth understanding what someone means by a statement like this. What is intended are two things: First there is postulated the existence of an MDF (mass density function) on the real line. Here it is called $f : \mathbb{R} \to [0, \infty)$. For consistency, let us call this function $\delta : \mathbb{R} \to [0, \infty)$. The use of the word "continuity" does not mean the function δ is continuous, but rather that $\delta = f$ is integrable with respect to Lebesgue measure and that $\alpha : \mathfrak{M} \to [0, 1]$ by

$$\alpha(A) = \int_A \delta$$

defines a "nice" integral measure with $\alpha(\mathbb{R}) = 1$. The measure α is a probability measure in particular. This is all perfectly mathematical. You may recall that where the use of the term "random" starts (and the mention of a "random variable" in particular appears) mathematical considerations have come to an end. The person is talking about something different, but the person is talking about something that can be understood in some kind of vague non-mathematical way. You'll notice that the magical "P" symbol for "probability" is soon to appear as well, and this plays a role, but not a very flattering one in this instance I'm afraid. In any case, keep in mind the two non-mathematical entities "P" and "Y," and I will attempt to tell you what is meant by a random variable. Oh, and don't forget the mathematical part which is the integral measure α with its MDF δ.

The person means two (more) things—the non-mathematical parts:

1. There is some kind of "process" or physical real world happening with outcomes modeled by real numbers. If the "process" is **hypothetically** executed (you can call this a "trial" or an "experiment," or whatever) then the outcome is symbolically represented by "Y." The symbol "Y" is not a real number itself; it is not the value of a function; it is not a function nor any other mathematical object. It is a vague symbol representing the **hypothetical** assignment of a real number to the hypothetical execution of a "process" or experiment.
2. There is some kind of relation—a sort of magical relation—between the hypothetical execution of this "process" and the probability measure α associated with δ. This magical relation is called "probability." A person who is talking about "random variables" definitely believes in "probability," not in the mathematical sense of a probability measure but in the magical sense of the "probability" symbol "P." This symbol again, does not represent a function or a number or anything mathematical, though mathematical looking notation is used in connection with it. In particular, symbolic expressions like

   $$P(Y = \omega) \qquad \text{and} \qquad P(Y \geq \omega)$$

   are used and assumed to have some kind of meaning: The "probability" that the "process" represented by "Y" has outcome corresponding to the value ω and the "probability" that the "process" represented by "Y" has

outcome corresponding to some value greater than or equal to ω respectively. Once you have the "process" and the "probability" doing their magical things together with the symbol "P," the magical relation with the measure is

$$P(Y \in A) = \int_A \delta \qquad \text{for every } A \in \mathfrak{M}. \tag{9}$$

This relation is often expressed simply in the case where $A = (\omega_1, \omega_2)$ is an interval as

$$P(\omega_1 < Y < \omega_2) = \int_{(\omega_1,\omega_2)} \delta.$$

Finally, I'm ready to get to the statement of part **(c)** of my problem:

(i) Given the belief in the magical relation of "random variables," "probabilities," and integral measures, what do you make of the hypothesis

$$P(Y = \omega) = f(\omega)$$

or $P(Y = \omega) = \delta(\omega)$ attributed to Markov?

(ii) What is the person actually trying to say with this "phrase?"

(d) Let's go through the calculation leading to the false Markov assertion above.

(i) Equality (1) is a definition. This is the definition of the "expectation of a random variable." In principle, one can't complain too much about mathematical definitions. The complaint, of course, is with the "random variable." This is why I defined the **expectation of a function with respect to a measure**.

- What function is this integral in line (1) the expectation of?
- What other name does this particular expectation go by?
- What would the expectation of a different function $g : \mathbb{R} \to \mathbb{R}$ be?

(ii) The inequality in (2) is the error leading to the overall incorrect stated result.

- Explain why this inequality is just wrong.
- Does (2) become correct with your added hypothesis from part **(b)** above?

(iii) The inequality in line (3) is correct. Explain why.

(iv) The equality in line (4) is also correct. This is just the linearity of the integral.

(v) Equality (5) is again just a "definition" or more properly a magical mystical interpretation. Explain how the definition in (5) is derived from the magic relation (9).

Problem 7 (weather probabilities and powers of a matrix; inspired by Uma Anand's probability project) Given that the weather today (or any particular day) is "sunny," assume the "probability" that it will also be sunny tomorrow is p_s, i.e., p_s is the probability that it continues to be sunny. Assume the (only) alternative to sunny weather is "rainy" weather. Thus, given that it is sunny today, the probability that it is rainy tomorrow is $1 - p_s$. Similarly, if the weather on a certain day is rainy, then assume the probability it will be rainy on the following day is p_r.

(a) Four very easy questions:

(i) If it is sunny today, what is the probability it will be sunny tomorrow?

(ii) If it is sunny today, what is the probability it will be rainy tomorrow?

(iii) If it is rainy today, what is the probability it will be sunny tomorrow?

(iv) If it is rainy today, what is the probability it will be rainy tomorrow?

(b) Four slightly less easy questions:

(i) If it is sunny today, what is the probability it will be sunny the day after tomorrow?

(ii) If it is sunny today, what is the probability it will be rainy the day after tomorrow?

(iii) If it is rainy today, what is the probability it will be sunny the day after tomorrow?

(iv) If it is rainy today, what is the probability it will be rainy the day after tomorrow?

(c) One question with four answers: Let $n \in \mathbb{N}$ be a natural number. Assuming the state of the weather today is known, there are four natural questions to ask about the weather n days from now. What are those four questions?

(d) There are various ways to arrange the numbers (probabilities) p_s, $1-p_s$, p_r and $1-p_r$ in a 2×2 matrix. Here are three of those ways.

$$A_1 = \begin{pmatrix} p_s & 1-p_s \\ p_r & 1-p_r \end{pmatrix}, \qquad A_2 = \begin{pmatrix} p_s & p_r \\ 1-p_s & 1-p_r \end{pmatrix}, \qquad \text{and} \qquad A_3 = \begin{pmatrix} p_s & 1-p_r \\ 1-p_s & p_r \end{pmatrix}.$$

Compute the squares of these matrices: A_1^2, A_2^2 and A_3^2. How do the entries compare to your calculations in part **(b)** above?

(e) One of the matrices from part **(d)** above should answer the questions from part **(b)**. Call this matrix A.

(i) Make a conjecture about the answers to the four questions from part **(c)** above.

(ii) Can you explain what probabilities are expressed by the entries in the matrices A_j^2 for $A_j \neq A$?

Problem 8 (weather probabilities and powers of a matrix; Problem 7 above) Let A be the matrix from part **(e)** of Problem 7 above. I'm going to introduce a (new) arrow notation that looks like this $X \xrightarrow{k} Y$ where X and Y are symbols from among S and R denoting "sunny" and "rainy" and k is some natural number. The meaning is that starting in state X (on a given day) one has the hypothetical outcome Y in k days. Thus, $S \xrightarrow{3} R$ means it is sunny one day and rainy on the third day after that. As you should have discovered in Problem 7 above the entries in the matrix $A^2 = (b_{ij})$ give the probabilities answering the four questions from part **(b)** of Problem 7 above. It may be further observed that these answers arise according to the following scheme:

$$\begin{array}{cc} & S \quad \leftarrow \text{start} \rightarrow \quad R \\ \begin{array}{c} S \\ \uparrow \\ \text{end weather} \\ \downarrow \\ R \end{array} & \begin{pmatrix} b_{11} & b_{12} \\ & \\ b_{21} & b_{22} \end{pmatrix} \end{array}$$

That is, b_{11} is the probability of $S \xrightarrow{2} S$ and b_{12} is the probability of $R \xrightarrow{2} S$ with the column number corresponding to starting conditions and the row index corresponding to ending weather conditions.

(a) Notice that $S \xrightarrow{2} S$ can also be represented as

$$\begin{array}{ccccc} & & S & \longrightarrow & S \\ & \nearrow & & & \\ S & & & & \\ & \searrow & & & \\ & & R & \longrightarrow & S \end{array} \tag{10}$$

where each arrow represents $\xrightarrow{1}$. Explain how to use the diagram (10) in conjunction with the law of conditional probability (Bayes' rule) and the law of total probability to compute the probability b_{11} of $S \xrightarrow{2} S$. Note the last symbolic phrase here should be read "it starts sunny one day and is sunny two days later."

(b) Draw a diagram using only arrows representing $\xrightarrow{1}$ to illustrate $S \xrightarrow{3} S$ and show your diagram is equivalent to

$$\begin{array}{ccccc} & & S & \longrightarrow & S \\ & \overset{2}{\nearrow} & & & \\ S & & & & \\ & \underset{2}{\searrow} & & & \\ & & R & \longrightarrow & S \end{array} \tag{11}$$

(c) Use the diagram (11) in conjunction with the laws of conditional and total probability to express the probability c_{11} of $S \xrightarrow{3} S$ as an entry in the product of one row in A^2 and one column in A. Hint(s): Given that it starts rainy (today) the probabilities it will be sunny and/or rainy respectively in two days are given by the entries in the second column

$$\begin{pmatrix} b_{12} \\ b_{22} \end{pmatrix}$$

of A^2. The probabilities that it turns out rainy tomorrow given that it is sunny and/or rainy respectively today are given by the second row

$$(1 - p_s \quad p_r)$$

in A. What would you do with this column and row to calculate the probability of $R \xrightarrow{3} R$?

(d) Draw four diagrams like (11) illustrating that $(c_{ij}) = AA^2$ gives the probabilities associated with $X \xrightarrow{3} Y$ according to the same scheme organizing the probabilities in A and A^2.

(e) Consider the diagrams

$$\begin{array}{ccc} & & S \\ & \overset{n-1}{\nearrow} & \\ S & & \\ & \underset{n-1}{\searrow} & \\ & & R \end{array} \qquad \text{and} \qquad \begin{array}{ccc} & & S \\ & \overset{n-1}{\nearrow} & \\ R & & \\ & \underset{n-1}{\searrow} & \\ & & R \end{array} \tag{12}$$

(i) What matrix represents the four probabilities associated with the diagrams in (12)?

(ii) Draw four diagrams illustrating the probabilities in the product matrix AA^{n-1}.

(iii) How can these diagrams be modified to illustrate the product $A^{n-1}A$?

Problem 9 (inadequate Markov lemma; Problem 6) In mathematical notation the assertion of the false Markov lemma from Problem 6 above is

$$\int_{\omega\in\mathbb{R}} \omega\,\delta(\omega) \geq a \int_{(a,\infty)} \delta. \tag{13}$$

Assume the following:

(i) $a > 0$ and

(ii) $\delta : \mathbb{R} \to [0,\infty)$ is the MDF of a probability measure with statistical range in $[0,\infty)$. That is,

$$\{\omega \in \mathbb{R} : \delta(\omega) > 0\} \subset [0,\infty). \tag{14}$$

Under these assumptions prove (13). Hint(s): Complete steps (2-4) above using the notation of (13) and correct reasoning.

Problem 10 (McMarkov lemma) The statement of Markov's lemma in Problem 6 was given in the hopes that it could be used to prove Chebyshev's inequality. In that application the assumption(s) of the correction in Problem 9 associated with (14) are too restrictive. One really needs to allow δ to have unrestricted statistical range. In fact, the statement of Markov's lemma in Problem 6 (even if it were correct) is not adequate to use in a (correct) proof of Chebyshev's inequality. Here is a generalization that is adequate:

Lemma 1 *Given a non-negative measurable function $x : \mathbb{R} \to [0,\infty)$,* **any** *probability MDF $\delta : \mathbb{R} \to [0,\infty)$ for an integral measure, and a constant $a \geq 0$, there holds*

$$\int_{\mathbb{R}} x\,\delta \geq a \int_{\{\tau\in\mathbb{R}:x(\tau)\geq a\}} \delta. \tag{15}$$

(a) Prove (15).[1]

(b) Translate (15) into hocus pocus language using the magical symbols "E" and "P" and a "random variable" called "X." (A full answer here should start with an introduction of the random variable X with an appropriate magical relation like (9).)

[1]If you find this difficult, you may wish to consult Problem 3 in the final assignment where the proof of the special case needed to prove Chebyshev's inequality is given in detai.

(c) Notice a crucial difference between the integral

$$\int_{\omega \in \mathbb{R}} \omega \, \delta(\omega)$$

on the left in (13) and the integral

$$\int_{\omega \in \mathbb{R}} x(\omega) \, \delta(\omega)$$

on the left in (15). Compare the incorrect assertion (13) of Problem 6 to the correct assertion (15):

- What measurable function $x : \mathbb{R} \to \mathbb{R}$ features in (13)?
- Why does this function fail to satisfy the hypotheses of the McMarkov lemma (Lemma 1) stated above?