hw7_student

April 7, 2020

1 Homework 7: Random Variables

1.1 Due Date: 11:59pm Monday, April 13

1.1.1 Submission

You will turn in this homework by uploading your answers in PDF format to Gradescope. You may turn in your answer as a scan or good quality camera phone picture of handwritten sheets, or you may turn it in as a PDF generated from typeset math (e.g. using LaTeX or Microsoft Word).

1.1.2 Collaboration Policy

Data science is a collaborative activity. While you may talk with others about the homework, we ask that you **write your solutions individually**. If you do discuss the assignments with others please **include their names** at the top of your submission.

1.2 Score Breakdown

Question	Points
Question 1a	1
Question 1b	2
Question 1c	1
Question 1d	1
Question 1e	1
Question 2a	2
Question 2b	2
Question 3	4
Question 4a	2
Question 4b	2
Question 4c	2
Question 4d	2
Total	22

1.3 Question 1: Guessing At Random

A multiple choice test has 100 questions, each with five possible answers of which one is right. The grading scheme is as follows:

- 4 points are awarded for each right answer
- For each other answer (wrong, missing, etc), one point is taken off; that is, -1 points are awarded.

A student hasn't studied at all and therefore guesses each answer uniformly at random, independently of all the other answers.

Define the following random variables:

- R: the number of answers the student gets right
- W: the number of answers the student does not get right
- S: the student's score on the test

1.3.1 Question 1a

What is the distribution of R? Either state the possible values and provide a formula for the probabilities, or provide the name and parameters of the appropriate distribution. Explain your answer.

Write your answer here, replacing this text.

1.3.2 Question 1b

Find $\mathbb{E}(R)$ and $\mathbb{SD}(R)$.

Write your answer here, replacing this text.

1.3.3 Question 1c

True or False: $\mathbb{SD}(R) = \mathbb{SD}(W)$. Explain your answer.

Write your answer here, replacing this text.

1.3.4 Question 1d

Find $\mathbb{E}(S)$.

Write your answer here, replacing this text.

1.3.5 Question 1e

Find $\mathbb{SD}(S)$.

Write your answer here, replacing this text.

1.4 Question 2: Correlation

Recall that the correlation between random variables X and Y is defined as

$$r(X, Y) \;=\; \mathbb{E}(X_{su} Y_{su})$$

where X_{su} is X in standard units and Y_{su} is Y in standard units.

In this exercise you will show that $-1 \le r(X, Y) \le 1$.

It will help to recall properties of random variables in standard units, so please do that before getting started.

Correlation is an expected product. We are much better at working with sums and squares than with products. So recall an elementary fact that connects sums, squares, and products: $(a+b)^2 = a^2 + 2ab + b^2$.

Hint: Refer to Slide 33 of Lec 20 to recall that $\mathbb{E}(X_{su}^2) = \mathbb{E}(Y_{su}^2) = 1$.

1.4.1 Question 2a

Consider the non-negative random variable $V = (X_{su} + Y_{su})^2$. Find $\mathbb{E}(V)$ and hence show that $r(X, Y) \ge -1$.

Write your answer here, replacing this text.

1.4.2 Question 2b

Use $W = (X_{su} - Y_{su})^2$ to show that $r(X, Y) \le 1$.

Write your answer here, replacing this text.

1.5 Question 3: Modified Robots

A company that makes robots has 12 new robots all designed for the same task.

The company times all the robots as they complete their task. Then it modifies each robot's mechanism. After the modificiation, it times the robots again as they complete their tasks.

Assume that the first and second times for Robot i are (X_i, Y_i) and that the pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_{12}, Y_{12})$ are i.i.d. That means each pair is an independent copy of all other pairs. Remember that *within* each pair, the distribution of X and Y might be different.

You can also assume that time is measured with enough precision that no two times come out exactly equal.

Nine out of the 12 robots performed faster after modification. Come up with hypotheses that you can test to see whether the modifications did nothing or whether they made the robots faster. Perform the test at the 5% level and provide your conclusion. The test is called the *sign test* because it is based on the signs of the differences $D_i = Y_i - X_i$.

You might find this Data 8 webpage helpful: Assessing Models.

Write your answer here, replacing this text.

1.6 Question 4: 10,000 Repetitions

When you run a simulation, you have to decide how many times to simulate your random variable. The number 10,000 is often suggested and used (the Packers did win 28-23). Let's see why.

First, 10,000 a nice round number and it feels like a lot. But there's more to it than that, as you will discover in this exercise.

1.6.1 Question 4a

Let I have the Bernoulli (p) distribution. Think of $\mathbb{V}ar(I)$ as a function of $p \in (0, 1)$. At what value of p is the function at its maximum, and what is the corresponding value of $\mathbb{V}ar(I)$ for that value of p? You can answer by drawing a graph or by using algebra or calculus.

Write your answer here, replacing this text.

1.6.2 Question 4b

Suppose you run n independent simulations of an experiment, and suppose that on each single simulation the chance that the experiment is a success is p. Let X be the proportion of successful experiments among the n experiments in your simulation. You can assume n is sufficiently big to apply CLT.

Fill in the blank with a function of n and p, and explain your answer: $P(X \in p \pm$_____$) \approx 0.95$

Write your answer here, replacing this text.

1.6.3 Question 4c

Now fill in the blank below with a function of n only. Note that the approximation in Part **b** has been replaced by an inequality.

Your answer shouldn't involve p. Often we don't know the value of p, so it's useful to have an interval width that doesn't depend on p.

Fill in the blank with a function of n, and explain your answer: $P(X \in p \pm$ _____$) \geq 0.95$

Hint: Use part **a**.

Write your answer here, replacing this text.

1.6.4 Question 4d

Suppose you are going to simulate a random variable Y and plot the empirical histogram based on your simulations.

Let $[a, b)$ be one of your histogram bins, and let $p = P(Y \in [a, b))$. That is, p is the true probability of your random variable falling in the interval $[a, b)$.

Suppose you run 10,000 independent simulations of Y. Let X be the proportion of simulations in which the simulated value falls in the interval $[a, b)$. Fill in the blank with a numerical value w and explain your answer:

$P(X \in p \pm w) \ \geq 0.95$

Since the bin $[a, b)$ was arbitrary, this calculation shows that if you make 10,000 simulations, then for every bin in your histogram there is at least a 95% chance that your observed proportion is within w of the true probability of falling into that bin.

Your value of w should have come out nice and simple (even simpler if you convert it to a percent). This makes the conclusion easy to remember.

Write your answer here, replacing this text.