Unit 3 Homework: Summarizing Distributions

w203: Statistics for Data Science

Applied Practice

Best Game in the Casino

You flip a fair coin 3 times, and get a different amount of money depending on how many heads you get.

- For 0 heads, you get \$0.
- For 1 head, you get \$2.
- For 2 heads, you get \$4.

Another piece of information is also flashing at you as you stand before the lights:

- Overall, across all possible outcomes your expected winnings from the game are \$6.

1. (1 point) How much do you get paid if the coin comes up heads 3 times?
2. (3 points) Write down a complete expression for the cumulative probability function for your winnings from the game.

The Warranty is Worth It

Suppose the life span of a particular (shoddy) server is a continuous random variable, T, with a uniform probability distribution between 0 and 1 year. The server comes with a contract that guarantees you money if the server lasts less than 1 year. In particular, if the server lasts t years, the manufacturer will pay you $g(t) = \$100(1-t)^{1/2}$. Let $X = g(T)$ be the random variable representing the payout from the contract.

1. (1 points) Compute the expected payout from the contract, $E(X) = E(g(T))$. Given the nature of the function, you might have to use integration by parts, or help from an integral solver.

Great Time to Watch Async

Suppose your waiting time in minutes for the Caltrain in the morning is uniformly distributed on $[0, 5]$, whereas waiting time in the evening is uniformly distributed on $[0, 10]$. Each waiting time is independent of all other waiting times.

a. (1 point) If you take the Caltrain each morning and each evening for 5 days in a row, what is your total expected waiting time?

b. (1 point) What is the variance of your total waiting time?

c. (1 point) What is the expected value of the total evening waiting time (all 5 days) minus the total morning waiting time (all 5 days)?

d. (1 point) What is the variance of the total evening waiting time (all 5 days) minus the total morning waiting time (all 5 days)?

Proof Practice

Maximizing Correlation

Correlation is a measure of linear dependence. Then, what are the possible values for correlation when one random variable is a linear function of another? To fix terms, suppose that X and Y are random variables with $V[X] > 0$, and a is a constant where $a \neq 0$ and b is any constant in $\mathbb{R}$. Furthermore, suppose that Y is a function of X that takes the following form:

$$Y = aX + b$$

(3 points) Prove that the possible values of $\rho(X, Y)$, the correlation between X and Y, are -1 and 1.

Notice that like the proof from last week, there is some sense of cases involved in this week's proof.

Optional Advanced Proof:

(0 points total) Heavy Tails

This challenge question cannot increase a student's final homework score above 100%. If you were to miss two points somewhere else in the homework, but produced a full answer for this question, you would have earned those two deduction points back.)

One reason to study the mathematical foundation of statistics is to recognize situations where common intuition can break down. An unusual class of distributions are those we call **heavy-tailed**. The exact definition[1] varies, but we'll say that a heavy-tailed distribution is one for which not all moments are finite.

Consider a random variable M with the following pmf:

$$p_M(x) = \begin{cases} \frac{c}{x^3}, & x \in 1, 2, 3, \ldots \\ 0, & otherwise, \end{cases}$$

where c is a constant (you can calculate its value if you like, but this is not crucial for the proof).

1. (1 bonus point) Is $E(M)$ finite?
2. (1 bonus point) Is $V(M)$ finite?

   Heavy-tailed distributions may seem odd, but they're not as rare as you might suspect.

   Researchers argue that the distribution of wealth is heavy-tailed; so is the distribution of computer file sizes, insurance payouts, and area burned by forest fires. These random variables are problematic in that a lot of common statistical techniques don't work on them.

   In this class, we won't cover heavy tailed distributions in depth, but we want you to become alert to their possibility.

Note: Maximum score on any homework is 100%

[1] This is a point that we make at places in the async and the live session: Arguments based on definitions are not particularly interesting arguments. *Chicago-style deep dish "pizza" is just lasagna. Change my mind.*