Stats 117 Problem Set 7

Due: Monday, May 18 5:00 p.m. on Gradescope

Please show your work for each exercise. If you collaborate with someone else—this is fine—be sure to note that in your homework submission. You must each write up separate answer sets. Any starred exercise is optional: they are extra challenging theoretical exercises for further developing your mastery.

Question 7.1 (Conditional expectations and gambling)**:** Consider the following simple gambling game, based on a random walk: at each time step t, we wager \$1, and with probability $\frac{1}{2}$ we win \$2, and with probability $\frac{1}{2}$ we lose our money, independently of all previous bets and results. Let $X_t \in \{-1, 1\}$ be the net profit in step t, and $S_n = \sum_{t=1}^{n} X_t$ be our profit after the nth bet (which may be negative). We decide to "bet until we are ahead" and walk away: for a value m, we stop betting once our profit reaches m, defining $N = \min\{n \in \mathbb{N} \mid S_n \geq m\}$.[1] The casino cuts off betting after T bets regardless of the outcomes, giving total profit

$$S_{\min\{T,N\}} = \sum_{t=1}^{\min\{T,N\}} X_t.$$

(a) Show how to write the event $\{N \geq t\}$ as a function of $X_1, \ldots, X_{t-1}$.

(b) Use conditional expectations to show that

$$\mathbb{E}[S_{\min\{T,N\}}] = 0.$$

Hint. It may be useful to rewrite $S_{\min\{T,N\}} = \sum_{t=1}^{T} \mathbf{1}\{N \geq t\} X_t$.

(c) Interpret the result of part (b).

Question 7.2: A random variable X has c.d.f.

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2} + \frac{x^2}{2} & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1. \end{cases}$$

(a) Does X have a p.d.f.?

(b) Give the median of X, that is, the value m such that $P(X \leq m) = \frac{1}{2}$.

(c) Write $X = 0 \cdot \mathbf{1}\{B = 0\} + Y\mathbf{1}\{B = 1\}$ for independent random variables B and Y. What distributions should B and Y have to give the CDF above?

(d) What is the density of Y?

(e) Give the 75th percentile of X.

Question 7.3 (Blitzstein and Huang, Ex. 5.1)**:** Let X have the *Rayleigh distribution*, meaning that it has density

$$f(x) = xe^{-x^2/2} \text{ for } x \geq 0,$$

and $f(x) = 0$ otherwise. Let $Y = 2X$.

[1] An argument similar to that for problem 6.5 shows that N is finite with probability 1.

(a) Find $P(1 \leq Y \leq 3)$.

(b) Find the first quartile, median, and third quartile of Y; these are the values q_1, q_2, q_3 (respectively) satisfying $P(X \leq q_j) = \frac{j}{4}$ for $j = 1, 2, 3$.

Question 7.4 (Variant of The Art of Chance, Ex. 18.4)**:** Suppose we have quantitative data, such as stock prices or country populations. What does the distribution of first digits look like? That is, what percentage of observations do you expect to start with the digit 1? What about the digit 9? If you've never tried this, look up a list of stock prices or country populations and count how many start with a 1. It may be more than you expect! This phenomenon is called *Benford's Law*.

Here is one model that explains Benford's Law. Suppose the quantitative data can be modeled by a random variable X with p.d.f.

$$f(x) = \begin{cases} \frac{c}{x^3} & \text{if } x \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

(This is one of a family called the *power law* distributions.)

(a) Determine the value of c that makes this a valid p.d.f.

(b) Calculate $P(\text{first digit of } X \text{ is } 1)$. *Hint.* You will have to calculate the probability of disjoint intervals. These probabilities form a geometric series.

(c) Calculate $P(\text{first digit of } X \text{ is } 9)$. Compare your answer to the previous part.

Question 7.5 (The Art of Chance, Ex. 20.6)**:** A common problem in statistics is to determine whether data is too large to have plausibly come from a distribution with CDF F. One way to do this is to calculate the probability of observing x or greater, $p = 1 - F(x)$, and if this p-**value** is small (say, less than .05), then we conclude that the data likely did not come from that distribution. (In the language of scientific experimentation, this arises when we reject a null hypothesis.)

(a) Suppose that the data X is a random variable that really does have CDF F, where F is continuous. What is the distribution of the p-value? *Hint.* To get full marks on this question, you should use the inverse $F^{-1}(u) = \min\{x \in \mathbb{R} \mid F(x) \geq u\}$, defined for $u \in (0, 1)$.

(b) Suppose the data X is a random variable that has CDF F, but F need not be continuous. Which of the following relationships between the distribution of a uniform $U \sim \mathsf{Uni}[0, 1]$ random variable and $F(X)$ is true?

i. For all $0 < u < 1$, $P(U \leq u) \leq P(F(X) \leq u)$
ii. For all $0 < u < 1$, $P(U \leq u) \geq P(F(X) \leq u)$
iii. For all $0 < u < 1$, $P(U \leq u) = P(F(X) \leq u)$.

(c) In our study, we use the p-value

$$p = 1 - F(X).$$

We say this is *conservative* if it is less likely to be small than a uniform: $P(p \leq u) \leq P(U \leq u)$ for all $u \in (0, 1)$. Is the p-value conservative?