2023Fall Probability & Mathematical Statistics 2023/11/05

Homework 5

Professor: Ziyu Shao & Dingzhu Wen *Due: 2023/11/12 10:59pm*

1. (a) The Cauchy distribution has PDF

$$f(x) = \frac{1}{\pi\left(1+x^2\right)}$$

for all x. Find the CDF of a random variable with the Cauchy PDF. Hint: Recall that the derivative of the inverse tangent function $\tan^{-1}(x)$ is $\frac{1}{1+x^2}$.

(b) The Pareto distribution with parameter $a > 0$ has PDF

$$f(x) = \frac{a}{x^{a+1}}$$

for $x \geq 1$ (and 0 otherwise). This distribution is often used in statistical modeling. Find the CDF of a Pareto r.v. with parameter a; check that it is a valid CDF.

2. The Exponential is the analog of the Geometric in continuous time. This problem explores the connection between Exponential and Geometric in more detail, asking what happens to a Geometric in a limit where the Bernoulli trials are performed faster and faster but with smaller and smaller success probabilities.

Suppose that Bernoulli trials are being performed in continuous time; rather than only thinking about first trial, second trial, etc., imagine that the trials take place at points on a timeline. Assume that the trials are at regularly spaced times 0, Δt, $2\Delta t$, ..., where Δt is a small positive number. Let the probability of success of each trial be $\lambda\Delta t$, where λ is a positive constant. Let G be the number of failures before the first success (in discrete time), and T be the time of the first success (in continuous time).

(a) Find a simple equation relating G to T. Hint: Draw a timeline and try out a simple example.

(b) Find the CDF of T. Hint: First find $P(T > t)$.

(c) Show that as $\Delta t \to 0$, the CDF of T converges to the Expo(λ) CDF, evaluating all the CDFs at a fixed $t \geq 0$.

3. Let X be a Pois (λ) random variable, where λ is fixed but unknown. Let $\theta = e^{-3\lambda}$, and suppose that we are interested in estimating θ based on the data. Since X is what we observe, our estimator is a function of X, call it $g(X)$. The bias of the estimator $g(X)$ is defined to be $E(g(X)) - \theta$, *i.e.*, how far off the estimate is on average; the estimator is unbiased if its bias is 0.

(a) For estimating λ, the random variable X itself is an unbiased estimator. Compute the bias of the estimator $T = e^{-3X}$. Is it unbiased for estimating θ?

(b) Show that $g(X) = (-2)^X$ is an unbiased estimator for θ. (In fact, it turns out to be the only unbiased estimator for θ.)

(c) Explain intuitively why $g(X)$ is a silly choice for estimating θ, despite (b), and show how to improve it by finding an estimator $h(X)$ for θ that is always at least as good as $g(X)$ and sometimes strictly better than $g(X)$. That is,

$$|h(X) - \theta| \leq |g(X) - \theta|$$

with the inequality sometimes strict.

4. Elk dwell in a certain forest. There are N elk, of which a simple random sample of size n is captured and tagged (so all $\binom{N}{n}$ sets of n elk are equally likely). The captured elk are returned to the population, and then a new sample is drawn. This is an important method that is widely used in ecology, known as capture-recapture. If the new sample is also a simple random sample, with some fixed size, then the number of tagged elk in the new sample is Hypergeometric.

   For this problem, assume that instead of having a fixed sample size, elk are sampled one by one without replacement until m tagged elk have been recaptured, where m is specified in advance (of course, assume that $1 \leq m \leq n \leq N$). An advantage of this sampling method is that it can be used to avoid ending up with a very small number of tagged elk (maybe even zero), which would be problematic in many applications of capture-recapture. A disadvantage is not knowing how large the sample will be.

   (a) Find the PMFs of the number of untagged elk in the new sample (call this X) and of the total number of elk in the new sample (call this Y).

   (b) Find the expected sample size $E[Y]$ using symmetry, linearity, and indicator random variables.

   Hint: We can assume that even after getting m tagged elk, they continue to be captured until all N of them have been obtained; briefly explain why this can be assumed. Express $X = X_1 + \cdots + X_m$, where X_1 is the number of untagged elk before the first tagged elk, X_2 is the number between the first and second tagged elk, *etc.* Then find $E[X_j]$ by creating the relevant indicator random variable for each untagged elk in the population.

   (c) Suppose that m, n, N are such that $E[Y]$ is an integer. If the sampling is done with a fixed sample size equal to $E[Y]$ rather than sampling until exactly m tagged elk are obtained, find the expected number of tagged elk in the sample. Is it less than m, equal to m, or greater than m (for $n < N$)?

5. The legendary Caltech physicist Richard Feynman and two editors of *The Feynman Lectures on Physics* (Michael Gottlieb and Ralph Leighton) posed the following problem about how to decide what to order at a restaurant. You plan to eat m meals at a certain restaurant, where you have never eaten before. Each time, you will order one dish.

   The restaurant has n dishes on the menu, with $n \geq m$. Assume that if you had tried all the dishes, you would have a definite ranking of them from 1 (your least favorite) to n (your favorite). If you knew which your favorite was, you would be happy to order it always (you never get tired of it).

   Before you've eaten at the restaurant, this ranking is completely unknown to you. After you've tried some dishes, you can rank those dishes amongst themselves, but don't know how they compare with the dishes you haven't yet tried. There is thus an *exploration-exploitation trade-off*: should you try new dishes, or should you order your favorite among the dishes you have tried before?

   A natural strategy is to have two phases in your series of visits to the restaurant: an exploration phase, where you try different dishes each time, and an exploitation phase, where you always order the best dish you obtained in the exploration phase. Let k be the length of the exploration phase (so $m - k$ is the length of the exploitation phase). Your goal is to maximize the expected sum of the ranks of the dishes you eat there (the rank of a dish is the "true" rank from 1 to n that you would give that dish if you could try all the dishes). Show that the optimal choice is

$$k = \sqrt{2(m+1)} - 1$$

   or this rounded up or down to an integer if needed. Do this in the following steps:

   (a) Let X be the rank of the best dish that you find in the exploration phase. Find the expected sum of the ranks of the dishes you eat, in terms of $E[X]$.

   (b) Find the PMF of X, as a simple expression in terms of binomial coefficients.

   (c) Show that

$$E[X] = \frac{k(n+1)}{k+1}.$$

   (d) Use calculus to find the optimal value of k.

6. (Optional Challenging Problem)

   (a) What is the probability that four points selected uniformly at random on a circle lie on the same semicircle?

   (b) What is the probability that $n \geq 2$ points selected uniformly at random on a circle lie on the same semicircle?

(c) Suppose $n \geq 2$ points selected uniformly at random on the surface of $d \geq 3$-dimension unit sphere, what is the probability that all points lie on the same hemisphere?