Stanford Stats 116 Final Examination

Open book and open notes
Duration: 180 minutes

Spring 2022

Name: _____

Student ID Number: _____

By taking this exam, you agree to be bound by the Stanford Honor Code, meaning specifically in this context that you

- will not give or receive aid in examinations;
- will do your share and take an active part in seeing to it that others as well as yourself uphold the spirit and letter of the Honor Code.

Signature: _____

Question F.1 (Properties of probabilities (11 pts)): Let S be a sample space and $A, B \subset S$, as well as $A_1, \ldots, A_n \subset S$, where $P(A) > 0$ and $P(B) > 0$.

(a) (2 pts) Assume the A_i are disjoint. Give $P(A_1 \cup A_2 \cup \ldots \cup A_n)$.

(b) (2 pts) Give $P(B \mid A)$.

(c) (3 pts) Suppose the A_i are independent. Give a formula for $P(A_1 \cup A_2 \cup \ldots \cup A_n)$ that *is not* inclusion/exclusion.

(d) (2 pts) Let A_n be conditionally independent of $A_1, \ldots, A_{n-2}$ given A_{n-1}. Show that $P(A_n \mid A_1, \ldots, A_{n-1}) = P(A_n \mid A_{n-1})$.

(e) (2 pts) Let A_i be independent of A_j for $j = 1, \ldots, i-2$ conditional on A_{i-1}. Give a formula for $P(A_1, A_2, \ldots, A_n)$ involving only $P(A_1)$ and $P(A_i \mid A_{i-1})$, $i = 2, 3, \ldots, n$.

Question F.2 (Randomized response (10 pts))**:** In private data analysis a technique called *randomized response* allows researchers to get unbiased estimates of quantities that may be sensitive, such as the proportion of a population using drugs. Assume you have an i.i.d. sample of individuals $i = 1, 2, \ldots, n$, and let $X_i \in \{0, 1\}$ be the indicator variable that individual i is currently using drugs. (So $\mathbb{P}(X_i = 1) = p$, though p is unknown, and X_i are i.i.d.)

In randomized response, for each individual i, the scientist asks "Are you currently using drugs?" Instead of answering truthfully, the individual flips a biased coin C with sides representing "Answer truthfully" ($C = \mathrm{T}$) and "Answer falsely" ($C = \mathrm{F}$), where the scientist *does not observe* the result of the flip, and then the individual answers the question either truthfully (if the coin is T) or falsely (if the coin is F). This coin has $P(T) = q$ and $P(F) = 1 - q$, where $q > \frac{1}{2}$ is known. (Alternatively, the individual spins a spinner, where different areas underneath the spinner are marked T/F, and the relative areas of each are q and $1 - q$.) Let Z_i be individual i's response, so that

$$Z_i = \begin{cases} X_i & \text{if } C = \mathrm{T} \\ 1 - X_i & \text{if } C = \mathrm{F}. \end{cases}$$

(a) (2 pts) Give $\mathbb{E}[Z_i \mid X_i]$.

(b) (3 pts) Give scalars $a, b \in \mathbb{R}$ such that $Y_i = aZ_i + b$ is unbiased for X_i, that is,

$$\mathbb{E}[Y_i \mid X_i] = X_i.$$

(c) (5 pts) You decide to use $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ as your estimate of the prevalence of drug use in the community, where $\mathbb{E}[X_i] = p$ (but of course, you don't know p). Give

$$\mathbb{E}[\overline{Y}_n] \quad \text{and} \quad \text{Var}(\overline{Y}_n)$$

in terms of n, p, and q.

Question F.3 (Concentration and inequalities (9 pts))**:** Let $X_1, X_2, \ldots, X_n$ be random variables. For each of the following sets of assumptions, give the sharpest bound you can on the probability

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \quad \text{for a fixed } t \geq 0.$$

That is, give a function $f(t)$ such that $\mathbb{P}(\sum_{i=1}^{n} X_i \geq t) \leq f(t)$, where $f(t)$ is as small as possible. Note that the assumptions on the X_i reset between each of parts (a), (b), (c).

(a) (3 pts) Assume that $X_i \geq 0$ for each i and that $\mathbb{E}[X_i]$ is finite.

(b) (3 pts) Assume that $\mathbb{E}[X_i] = 0$, that $\mathrm{Var}(X_i) = \sigma^2$, and that the X_i are uncorrelated, so $\mathbb{E}[X_i X_j] = 0$ when $i \neq j$.

(c) (3 pts) Assume the X_i are independent, mean zero, and that $\log(\mathbb{E}[e^{\lambda X_i}]) \leq \frac{\lambda^2 \sigma^2}{2}$ for all $\lambda \in \mathbb{R}$.

Question F.4 (15 pts)**:** A disease is transmitting through a population, and any time an infected individual interacts with another, there is a probability $p \in [0, 1]$ of transmitting the disease. Interactions happen at times $t = 1, 2, 3, 4, \ldots$, and $N(t)$ is the number of new infections at time t. At time t, each infected individual interacts with 2 additional individuals, and we begin at time $t = 0$ with patient 0, where $N(0) = 1$. See Figure 1 for a diagram. Note: an individual infected at time t is recovered at time $t + 1$ and can never infect anyone else.

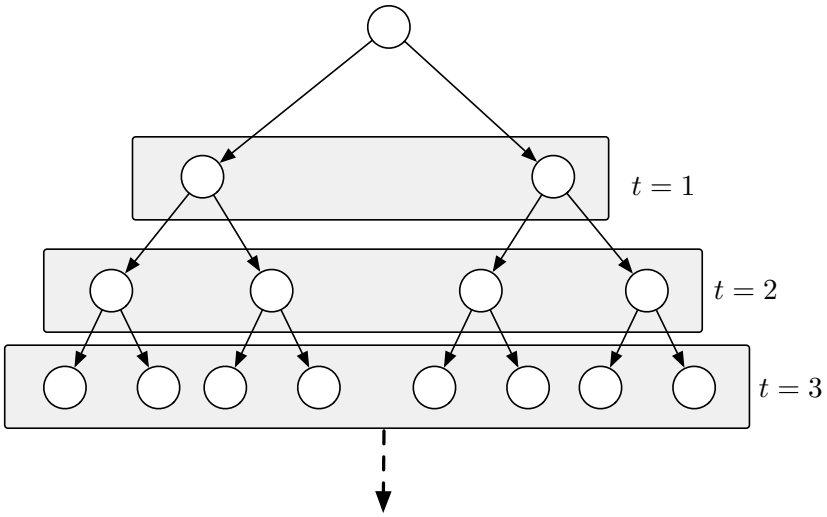

Figure 1. Interactions with potential infected individuals at times $t = 1, 2, 3, \ldots$. The random variable $N(t)$ is the number of infected individuals in a time block t. (So if each individual in the last row of the graph above is infected, $N(3) = 8$.)

(a) (2 pts) Give the conditional expectation $\mathbb{E}[N(t) \mid N(t-1)]$.

(b) (2 pts) Assuming $N(0) = 1$, give $\mathbb{E}[N(t)]$, the expected number of infections at time t using part (a).

(c) (2 pts) If $p < \frac{1}{2}$, show that the number of infections converges in probability to 0, that is, for all $\epsilon > 0$

$$\lim_{t \to \infty} \mathbb{P}(N(t) \geq \epsilon) = 0.$$

(d) (3 pts) Give a formula for $\mathrm{Var}(N(t))$ in terms of $\mathrm{Var}(N(t-1))$ and $\mathbb{E}[N(t-1)]$.

(e) (4 pts) Using your answer to part (d), show that if $p \neq \frac{1}{2}$,

$$\mathrm{Var}(N(t)) = \frac{(2p)^{2t} - (2p)^t}{2p-1}(1-p)N(0).$$

(f) (2 pts) Now assume that $N(0)$ may be larger than 1. Show that

$$\frac{\mathrm{Var}(N(t))}{\mathbb{E}[N(t)]^2} = \frac{(1-p)(1-(2p)^{-t})}{(2p-1)N(0)}.$$

As the number of initial infections $N(0)$ gets large, if $p > \frac{1}{2}$ and $u > 0$, what happens to

$$\mathbb{P}\big(N(t) \leq (1-u)\mathbb{E}[N(t)]\big)?$$

Question F.5 (Disease testing (5 pts))**:** Say that the prevalance of a disease (i.e., the prior probability an individual has the disease) is $p = \frac{1}{10000} = .0001$. You have a high quality test for the disease with low false positive and false negative rates. Call T the outcome of the test, where $T \in \{+, -\}$ (positive and negative). Then

$$\mathbb{P}(T = + \mid \text{no disease}) = .0098\overline{0098}\ldots = \frac{98}{9999}, \quad \mathbb{P}(T = - \mid \text{disease}) = .02.$$

So the false positive rate is under 1/100 and the false negative rate is 1/50. You (a random individual from the population) test positive for the disease. What is the probability you actually have the disease?

Question F.6 (Testing the effect of antibiotics on cows (8 pts))**:** Farmers (unfortunately) often feed their animals antibiotics as it causes the animals to gain weight, meaning the farmers can sell more.[1] Farmer John hears about this and wonders if including antibiotics in the feed for his Holsteins (dairy cows) will increase their milk production. He has 5000 cows and divides them randomly into two groups, each of 2500 cows, including a specified dose of antibiotics in the feed for the first group. He measures the total milk production for the year for each of his cows, letting X_i be the amount of milk that cow i in the first group produces, Y_i be the amount of milk that cow i in the second group produces (these are different cows, but the range for both is $i = 1, \ldots, 2500$).

Let $D_i = X_i - Y_i$ be the difference in milk production between cows in each group. Farmer John has a *null hypothesis* that there is no difference between the distributions of milk production in the groups; he would like to know if that hypothesis is false.

(a) (2 pts) What is the mean $\mathbb{E}[D_i]$ under Farmer John's null hypothesis?

(b) (4 pts) The milk any single cow can produce over a year is of course bounded[2], so $\mathbb{E}[D_i^2] < \infty$. Give a statistic T_n, which is a function of all of the D_i, which is approximately $\mathsf{N}(0, 1)$ distributed under Farmer John's null hypothesis (and which should be large if antibiotics increase milk production). Justify (in a sentence or two) why your statistic is approximately normal.

[1] This is a major contributor to the rise of multi-drug resistant bacteria.

[2] This is not important for the question, but the average Holstein produces roughly 10,000kg of milk per year.

(c) (2 pts) Farmer John will start adding antibiotics to his feed for all the cows if T_n is larger than a threshold t, where $\mathbb{P}(T_n \geq t) \leq .01$. What value for t should he choose? What if he wants $\mathbb{P}(T_n \geq t) \leq .001$? The following Z-scores may be useful, where $Z \sim \mathsf{N}(0, 1)$ in the table.

z	0	1	2	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3
$P(Z \leq z)$	.5	.84	.977	.982	.986	.989	.992	.994	.995	.996	.997	.998	.999

Question F.7 (A few limit distributions (8 pts)): Let $X_i \overset{\text{iid}}{\sim} \mathsf{Poisson}(\lambda)$ and $S_n = \sum_{i=1}^n X_i$ be their sum.

(a) (2 pts) Give the distribution of S_n.

(b) (2 pts) Let $\epsilon > 0$. Give $\lim_{n\to\infty} \mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq \epsilon \mathbb{E}[S_n])$.

(c) (2 pts) Does $\overline{X}_n = \frac{1}{n} S_n$ have a limiting distribution as $n \to \infty$? If so, what is it? If not, why not?

(d) (2 pts) Give an approximate distribution of $Z_n = \frac{1}{\sqrt{n}}(S_n - n\lambda)$. What happens as $n \to \infty$?

Scratch