Stats 116 Problem Set 3

Due: Wednesday, April 20 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. Questions are either from Ross's *A First Course in Probability* or our are home-cooked.

Question 3.1 (Ross problem 3.81)**:** A and B play a series of games. Each game is independently won by A with probability p and by B with probability $1 - p$. They stop when the total number of wins of one of the players is two greater than that of the other player. The player with the greater number of total wins is declared the winner of the series.

(a) Find the probability that a total of 4 games are played.

(b) Find the probability that A is the winner of the series.

Question 3.2 (Ross problem 4.25)**:** Two coins are to be flipped. The first coin will land on heads with probability .6, the second with probability .7. Assume that the results of the flips are indpendent, and let X equal the total number of heads that result.

(a) Find $P(X = 1)$.

(b) Determine $\mathbb{E}[X]$.

Question 3.3 (Ross problem 4.28)**:** A sample of 3 items is selected at random from a box containing 20 items, of which 4 are defective. Find the expected number of defective items in the sample.

Question 3.4 (Ross problem 4.68)**:** Each of 500 soliders in an army company independently has a certain disease with probability $1/10^3$. This disease will show up in a blood test, and to facilitate matters, blood samples from all 500 soldiers are pooled and tested.[1]

(a) What is the (approximate) probability that the blood test will be positive (that is, at least one person has the disease)? *Hint.* You may either use formulae for binomial distributions or inclusion/exclusion for this. An answer accurate to a few digits suffices.

Suppose now that the blood test yields a positive result.

(b) What is the probability, under this circumstance, that more than one person has the disease?

Now, suppose one of the 500 people is Jones, who knows that he has the disease.

(c) Given that Jones has the disease, what is the (approximate) probability that more than 1 person has the disease?

Because the pooled test was positive, the testing facility decies to test each individual separately. The first $(i - 1)$ of these tests are negative, and test i is Jones, which is positive.

[1] This type of pooling of samples allows faster turnaround time for, e.g., Color COVID tests at Stanford as long as positivity rates remain low.

(d) Given the preceding scenario, what is the probability, as a function of i, that any of the remaining people have the disease?

Question 3.5 (Ross Theoretical exercise 4.5)**:** Let N be a nonnegative integer-valued random variable. For nonnegative values $a_j, j = 1, 2, \ldots$, show that

$$\sum_{j=1}^{\infty}(a_1 + \cdots + a_j)P(N = j) = \sum_{i=1}^{\infty} a_i P(N \geq i).$$

Then show that

$$\mathbb{E}[N] = \sum_{i=1}^{\infty} P(N \geq i)$$

and

$$\mathbb{E}[N(N+1)] = 2\sum_{i=1}^{\infty} iP(N \geq i).$$

Question 3.6 (Interviewing for a job at Google)**:** In interviewing for a job at Google[2], the interviewer asked John the following question: "You have an iPod with r songs on it. The iPod is on shuffle mode, so that each time you listen to a new song, it picks one at random from its library (with replacement) and plays it for you. About how long do you expect it to take to hear all the songs on the iPod at least once?" We'll answer a few versions of this question here, which are a bit harder than the particular question in the job interview. Let T be the amount of time until you hear each song at least once (so that $T \geq r$ always). We will show that (more or less) T will never be less than $2r \log r$.

(a) Show that

$$\mathbb{P}(T > 2r \log r) \leq \frac{1}{r}.$$

(b) Show that

$$\mathbb{P}(T > 2r \log r) \geq \frac{1}{r} - \frac{2 \log r}{r^2} - \frac{1}{2r^2}.$$

We will still give full credit if you show that $\mathbb{P}(T > 2r \log r) \geq \frac{1}{r} - \mathrm{err}(r)$, where $\mathrm{err}(r)$ is any error term that satisfies $r \cdot \mathrm{err}(r) \to 0$ as $r \to \infty$.

Hint. Use example 4.1e from the Ross book (page 122 of the 10th edition). You may also find it useful that for the inclusion/exclusion principle and any events $A_1, A_2, \ldots, A_r$, one has

$$\mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_r) \leq \sum_{i=1}^{r} \mathbb{P}(A_i) \quad \text{and} \quad \mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_r) \geq \sum_{i=1}^{r} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i, A_j).$$

Feel free to use the two inequalities $e^x \geq 1 + x$, valid for all $x \in \mathbb{R}$ (especially $e^{-1/r} \geq 1 - \frac{1}{r}$ and $e^{-2/r} \geq 1 - \frac{2}{r}$), as well as the inequality $e^{-1/r - 1/r^2} \leq 1 - \frac{1}{r}$ for all $r \geq 2$.

[2]this is a true story!

(c) ***Extra credit (and super extra challenging):** Show that if $\epsilon > 0$, we have

$$\mathbb{P}(T > (1+\epsilon) r \log r) \to 0$$

as $r \to \infty$, that is, as the number r of songs gets large, the probability that the number of plays until you hear each song at least once is essentially never (much) larger than $r \log r$. Additionally, show that

$$\mathbb{P}(T > r \log r - \gamma r) \approx 1 - \exp(-e^{\gamma})$$

for any fixed $\gamma > 0$ as $r \to \infty$. With these two results, you have then shown that

$$\mathbb{P}((1-\epsilon) r \log r < T \leq (1+\epsilon) r \log r) \to 1$$

as $r \to \infty$, valid for any $\epsilon > 0$. *Hint.* You may use the approximation

$$\sum_{i=1}^{k} (-1)^{1+i} \binom{r}{i} \left(1 - \frac{i}{r}\right)^n \approx \sum_{i=1}^{k} (-1)^{1+i} \frac{r^i}{i!} \exp\left(-\frac{n \cdot i}{r}\right),$$

valid for fixed (finite) k and $n \leq r \log r$ as r tends to ∞.