Statistics 134 – Takehome Final
Due: Wednesday, 13 August

Instructions: There are five questions, numbered 0-4. You **must** do problem #0, and **three** out of the remaining four questions, numbered 1-4. Each question is worth 25 points. Turning in more problems will not get you extra credit. If a problem depends on the answer to a previous, unanswered problem, leave that answer as a variable. (for example: *"Let x be the answer to part (a). Then, . . . "*)

Show calculations, or give reasons, on all parts of all problems. Each step in your solutions should be **explained**, and your explanations should be **clear** — use **complete sentences** when not unreasonable. You should **state completely** any results from class or from the book that you use, and say why they apply. *Note: you won't lose points for incorrect grammar or spelling.*

You may discuss these problems with other students, but your write-up should be your own — do not compare final write-ups. Finally, **please tell me what other students you worked with**.

Good luck!

Do this problem:

(0) Explain in complete sentences, without formulas, why the following statements are true. A brief example: "If X and Y are independent Geometric(p) random variables, then $X + Y$ is Negative Binomial(2,p): *Suppose we conduct a sequence of independent trials, in which the probability of success is p. The distribution of the number of failures before the first success is Geometric(p), and the distribution of the number of failures between the first and the second success is Geometric(p) as well, independently of the previous number of failures. The sum of these two values is the number of failures before the second success, which has the Negative Binomial(2,p) distribution.*"

(a) If X is Binomial(n,p) and Y is Binomial(m,p), and X and Y are independent, then $X+Y$ is Binomial($n+m$,p).

(b) If X is Gamma(r, λ) and Y is Gamma(s, λ), and X and Y are independent, then $X+Y$ is Gamma($r+s$,λ).

(c) Consider a Poisson(λ) process. Let N_1 be the number of hits that occur in the first 6 minutes; let N_2 be the number that occur between the 6^{th} minute and the 10^{th} minute; and let N_3 be the number that occur between the 10^{th} minute and the 12^{th} minute. If $N = N_1 + N_2 + N_3$ is the total number of hits that occur in the first 12 minutes, then the distribution of (N_1, N_2, N_3) conditioned on $N = 20$ is Multinomial(20,$\frac{1}{2}$,$\frac{1}{3}$,$\frac{1}{6}$).

(d) If $X_1, X_2, \ldots, X_{10}$ are independent and have the same continuous distribution with density $f(x)$ and cumulative distribution function $F(x)$, then the density of $X_{(5)}$ at x is given by: 10 times $f(x)$ times the probability that a Binomial(9, $F(x)$) random variable takes the value 4.

Do three out of the following four problems.

(1) My (fictional) dog has fleas. They arrive on him mysteriously at the arrival times of a Poisson process – an average of five fleas per day. So, every day at 6pm, I remove fleas – each flea is removed and squished with probability p, independently of the other fleas, or how long the flea has been on my dog. Suppose that he had no fleas at all at 6pm on May 1^{st}. The fleas do not reproduce, or leave on their own.

(a) What is the distribution of the number of fleas on my dog at 6pm (immediately before flea removal), n days after May 1^{st}?

(b) What is the distribution of the number of fleas that I will remove from my dog on the n^{th} day after May 1^{st}?

(c) Suppose that on May 10^{th} I removed 10 fleas. What's the probability that there were k more fleas that I didn't remove?

(2) A seagull is dropping small snails on a parking lot. Taking the point she is aiming for as the center of the coordinate system, the locations that the snails hit are independent, and the x-coordinate and the y-coordinate are independent Gaussian random variables with mean 0 and standard deviation 1 foot. After she drops 10 snails, for $k \in \{1, 2, \ldots, 10\}$ let $R_{(k)}$ be the distance of the k^{th} closest snail to the target, and let $(X_{(k)}, Y_{(k)})$ be the location of that snail. For instance, $(X_{(1)}, Y_{(1)})$ is the location of the snail that has landed closest to the target, and $R_{(1)} = \sqrt{X_{(1)}^2 + Y_{(1)}^2}$.

(a) What is the distribution of $R_{(1)}^2$?

(b) What is the distribution of $(X_{(1)}, Y_{(1)})$? Justify your answer carefully.

(c) Compute the expected value of $R_{(2)}^2$.

(3) A strand of DNA consists of two paired strands of n nucleotides each. Call the two strands the "left" and the "right" strands. For $k \in \{1, 2, \ldots, n\}$, the k^{th} nucleotide in the left strand is paired with the k^{th} nucleotide in the right strand, so that there are n pairs of nucleotides in total. If only one nucleotide in a given pair mutates, then the error can be corrected, but if both mutate, it cannot. Suppose that m different nucleotides are chosen uniformly at random to mutate. (so, the first mutation is chosen uniformly from the $2n$ nucleotides; the second is chosen uniformly from the remaining $2n - 1$; and so forth)

(a) What's the expected number of nucleotide pairs in which both nucleotides have mutated?

(b) What's the variance of the number of nucleotide pairs in which both nucleotides have mutated?

(c) Given that k pairs of nucleotides had both nucleotides mutate, what's the distribution of the number of pairs in which only the nucleotide on the "left" strand mutated?

(4) Suppose that a randomly chosen tree in a forest with 10,000 trees has height H and width W, where H is exponentially distributed with mean 3 meters, and $W = \frac{UH}{20}$, where U is independent of H and uniformly distributed on $(0, 1)$. Suppose that the dimensions of each tree are independent of all other trees.

(a) Find a h so that the distribution of the number of trees that are taller than h is approximately Poisson(10).

(b) What is the distribution of the **locations** of all trees of height taller than h, where h is the answer to part (a)? Justify your answer carefully.

(c) We cut down 100 randomly chosen trees with height greater than 5 meters. What is the approximate distribution of the sum of the widths of those trees?