Name: ______________________________

Instructions: Turn off your phone. No calculators or electronics of any kind are allowed. Make sure your exam has 6 pages. You may write on the back of the pages. Do as many problems as you can. If you get stuck on one, move on to the next. You have until 4:30pm to complete the exam.

Academic honesty: Sign on the line below when you are finished with the exam.

I have neither given nor received aid in this exam. All the work below is entirely my own.

______________________________ ____________

Signature Date

Problem 1 How many people inject heroin in New Haven? (20 points)

Suppose we wish to estimate N, the number of people who inject heroin in New Haven. The Drug Enforcement Agency and the local hospital work together to keep track of all patients admitted for a heroin overdose. The hospital reports that n unique persons have suffered at least one overdose in the last year. Let X_i, $i = 1, \ldots, n$ be the number of overdoses reported for person i in the last year. The hospital does not report any subjects who have had zero overdoses, since these people are not admitted to the hospital. Assume that the number of times a heroin injector overdoses in a single year has Poisson(λ) distribution, independent of other individuals. Further assume that *every* person who experiences a heroin overdose is taken to the hospital. Since our data come from the hospital, we observe only positive counts $X_1, \ldots, X_n$, since a heroin injector who has never experienced an overdose is not observed in the data – there are no zero counts.

a. (5 points) Show that the likelihood of the observed data is

$$L(\lambda) = (e^{\lambda} - 1)^{-n} \times \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!}$$

b. (10 points) Devise a Newton's method algorithm for finding the MLE of λ.

c. (5 points) Suppose you have an estimate $\hat{\lambda}$ of λ. Derive an estimate N.

[Scratch]

Problem 2 Disease alleles (20 points)

Everyone has exactly two copies of a certain gene in their genome. There are two variants (alleles) of the gene, called A and a. A person can have one of three possible genotypes (pairs of alleles): AA, Aa, or aa. Under certain assumptions about the flow of alleles in large populations, these genotypes have population frequencies p^2, $2p(1-p)$, and $(1-p)^2$ respectively, where p is the population frequency of the A allele. The A allele is dominant: anyone having at least one A allele has a certain disease, while aa individuals are healthy. This disease is only caused by having an A allele – there is no other cause. The disease is the same regardless of whether the person afflicted has genotype Aa or AA. In a random sample of size n from the population, we find n_h individuals are healthy and n_d individuals have the disease, where $n_h + n_d = n$. We do not measure the subjects' genotypes, but we wish to estimate p.

genotype	population frequency	phenotype
AA	p^2	Disease
Aa	$2p(1-p)$	Disease
aa	$(1-p)^2$	Healthy

a. (2 points) What is the probability that a given subject has the disease, in terms of p?

b. (3 points) What is the probability that someone has genotype AA, given that they have the disease?

c. (5 points) Show that the likelihood is

$$L(p) = [p^2 + 2p(1-p)]^{n_d}(1-p)^{2n_h}.$$

d. (10 points) Derive an EM algorithm to estimate p, the frequency of the A allele in the population. Invent the "missing data" and give the update expressions for p and the missing data variable.

[Scratch]

[Scratch]