Stat 343 Midterm Review

Midterm Structure and Coverage

The midterm will have a few conceptual multiple choice and short answer questions at be beginning, and a couple of more "do-the-math" type questions.

The midterm covers material discussed up through Feb. 27, problem sets 1 through 5. I won't ask about Newton's method on the midterm, but I do hope you understand how it works.

Conceptual Topics

Here are some definitions to know and things to understand for the conceptual part:

- Definition of a simple random sample
- The difference between:
  - Mean and variance of values in a population
  - Mean and variance of values in a sample
  - Mean and variance of a statistic (like the sample mean)
  - You do **not** need to memorize any formulas that we derived about the mean and variance of the sample mean from a finite population
- What a sampling distribution is
- What the Central Limit Theorem says
- Estimators vs estimates (estimators are random variables, estimates are realized values based on sample data)
- Bias, variance, and mean squared error of an estimator:
  - Definitions of each. You should be able to state them mathematically and understand what they mean at a more intuitive level
  - For the second part, you should be able to find the bias, variance, and mean squared error of an estimator
- Confidence intervals:
  - Definition and interpretation
  - Before observing data, end points are random variables; can make probability statements
  - After observing data, we have a realized value of the confidence interval; can no longer make probability statements
  - Definitions of coverage probability and nominal coverage probability
- Maximum likelihood estimation
  - Describe what the likelihood function is and why we would want to maximize it
  - For the second part, you will be asked to find a maximum likelihood estimator/estimate
- Bayesian inference
  - Explain what prior and posterior distributions represent
  - Definition of conjugate prior
  - Interpret credible intervals; understand where they come from
  - For the second part, you will be asked to show that a prior is a conjugate prior and find the parameters of the posterior.

Example Conceptual Problems

There are a huge variety of possible conceptual problems. Here are a few examples.

Problem 1. A 90% confidence interval for the average number of children per household based on a simple random sample is found to be (0.7, 2.1). Because the average number of children per household, μ, is some fixed number in the population (at least, at a particular moment in time when we conduct the study), it doesn't make any sense to claim that $P(0.7 \leq \mu \leq 2.1) = 0.90$. What do we mean, then, by saying that this is a "90% confidence interval"? Can we ever make probability statements about confidence intervals?

Problem 2. What is a posterior distribution in a Bayesian analysis? If I know the posterior distribution for a model parameter, how can a 95% posterior credible interval be formed? (You should answer the first question with a written sentence. For the second question, you could write a sentence and/or draw a picture to illustrate.)

Problem 3. What is the mean squared error of an estimator (you can answer with either a formula or a written sentence explaining the intuition)? Why is an estimator with low mean squared error preferred to an estimator with high mean squared error?

Example Worked Problems

The midterm will have problems roughly similar in content to the examples below.

Problem 1

The EPA conducts occasional reviews of its standards for airborne asbestos. During a review, the EPA examines data from several studies (denote the number of studies by s). Different studies keep track of different groups of people; different groups have different exposures to asbestos. Let n_i be the number of people in the i'th study, let x_i be the asbestos exposure for people in that study, and let y_i be the number of people who developed lung cancer in that study. The EPA's model is $Y_i \sim \text{Poisson}(\lambda_i)$, where $\lambda_i = n_i x_i \lambda$ and where λ is the typical rate at which asbestos causes cancer. The n_i's and x_i's are known constants; the Y_i's are random variables. Because the different studies involve different sets of people in different locations, they model the Y_i's from different studies as being independent (but not identically distributed since the λ_i's are different!). The EPA wants to estimate λ.

In answering the questions below, you may use the following facts about the Poisson and Gamma distributions:

Suppose $X \sim$ Poisson(λ)

p.f.	$f(x\|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$
Mean	λ
Variance	λ

Suppose $X \sim$ Gamma(α, β)

p.f.	$f(x\|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$
Mean	$\frac{\alpha}{\beta}$
Variance	$\alpha\beta^2$

(a) Find the pdf of the joint distribution of $Y_1, \ldots, Y_s|\lambda$ (this will also involve the constants $x_1, n_1, \ldots, x_s, n_s$).

(b) Find the maximum likelihood estimator of λ.

(c) Is the maximum likelihood estimator an unbiased estimator of λ?

(d) Find the variance of the maximum likelihood estimator.

(e) Find the mean squared error of the maximum likelihood estimator.

(f) Suppose the EPA uses this model to estimate λ by combining data from $s = 3$ studies with data recorded in the table below. Find an expression for the maximum likelihood estimate of λ. Your answer should involve only numbers, no symbols; but you do not need to simplify your expression.

Study Number (i)	Sample Size (n_i)	Exposure Level (x_i)	Cancer Case Count (y_i)
1	10	0.3	1
2	25	0.2	3
3	100	0.5	15

(g) Suppose the analysts adopt a prior of $\Lambda \sim \text{Gamma}(\alpha, \beta)$, where α and β are known constants they choose to reflect their prior knowledge about λ. Find the posterior distribution for Λ. You should arrive at a specific form for the posterior distribution, with parameters involving α, β, $x_1, \ldots, x_s$, $n_1, \ldots, n_s$, and $y_1, \ldots, y_s$.

(h) Again, the EPA uses this model to estimate λ by combining data from three studies with data recorded in the table below. They use a prior of $\Lambda \sim \text{Gamma}(1, 3)$. Find expressions for the parameters of the posterior distribution for Λ. Your answer should involve only numbers, no symbols; but you do not need to simplify your expression.

Study Number (i)	Sample Size (n_i)	Exposure Level (x_i)	Cancer Case Count (y_i)
1	10	0.3	1
2	25	0.2	3
3	100	0.5	15

Problem 2. From independent surveys of two populations, 90% confidence intervals for the population means μ_1 and μ_2 will be constructed. Denote the first interval, which is an estimate of μ_1, by $[L_1, U_1]$ and the second interval, which is an estimate of μ_2, by $[L_2, U_2]$. We have not taken our sample yet, so L_1, U_1, L_2, and U_2 are random variables. What is the probability that both of these confidence intervals will contain their respective population means?

Problem 3. Two suveys were independently conducted to estimate a population mean μ. Denote the estimators from the independent surveys and their variances by $\hat{\mu}_1$, with variance $\sigma_1^2 > 0$ and $\hat{\mu}_2$, with variance $\sigma_2^2 > 0$. Assume that both $\hat{\mu}_1$ and $\hat{\mu}_2$ are unbiased. For some constants α and β, the two estimators can be combined to give a new estimator $Y = \alpha\hat{\mu}_1 + \beta\hat{\mu}_2$.

(a) Find a condition on α and β so that the combined estimator Y is unbiased.

(b) What choice of α and β minimizes the variance of Y, subject to the condition of unbiasedness?