Stat 111 Spring 2025 Week 5: Likelihood Functions (Rice Chapter 8)

1. **The likelihood Function and Sufficient Statistics** (8.5, 8.8)

   a) Define the likelihood function $L(\theta)$ as the probability (or probability density) of the observed data values for a given value of the unknown parameter θ. Explain why the value $\hat{\theta}$ that maximizes $L(\theta)$ (the *maximum likelihood estimate*) is a natural estimate for θ. Point out how very different likelihood functions could all give the same maximum likelihood estimates.

   b) We sometimes encounter situations with discrete parameter spaces (e.g., in the context of hypothesis testing). Suppose a coin is known to either be fair ($\theta = 0.5$) or 2-headed ($\theta = 1$). Write out the likelihood function $L(\theta)$ following two independent trials with the coin landing heads 0, 1 or 2 times. For outcomes with 2 heads, how much more likely is $\theta = 1$ than $\theta = 0.5$?

   c) Now consider $\theta \in [0, 1]$ to be a continuous probability parameter, and graph $L(\theta)$ and $l(\theta) = \log(L(\theta))$ for $0, 1$ or 2 successes in 2 independent trials. Explain why it is reasonable to rescale $L(\theta)$ to have maximum value 1, and recenter $l(\theta)$ to have maximum value 0 (e.g., when comparing $\theta = 1$ and $\theta = 0.5$).

   d) For a given probability model, the likelihood function contains all information from the data about the unknown parameter θ. For a sample $X_1, \ldots, X_n$ there is often a lower dimensional summary statistic $T(X_1, \ldots, X_n)$ that retains all of the information from the n data values. Define sufficient statistics and state the factorization theorem. Explain how a sufficient statistic is a data summary that allows you to graph the likelihood function, scaled to have a maximum of 1, or the log-likelihood function, recentered to have a maximum of 0. As an example, for the $n = 272$ regular season games in the 2024 NFL season, the average winning margin for the home team was $\bar{x} = 1.95$. Assuming the winning margins are approximately iid $N(\theta, \sigma^2)$ random variables with known standard deviation $\sigma = 14.0$, graph the likelihood function for θ.

   e) Define the Exponential family of distributions, and show how the k-parameter family includes only distributions for which there is a sufficient statistic of dimension k. Show how $N(\mu, \sigma^2)$ is included but Unif$(\theta - 1/2, \theta + 1/2)$ is not.

2. **Maximum Likelihood Estimates, Information** (8.5.2, 8.7)

   a) Define the *information* as the expectation of the squared first derivative of the log-likelihood function. Also show the equivalent second derivative formula that works for exponential families. Explain why more information suggests a more precise maximum likelihood estimate.

   b) Using $N(\theta, (14)^2)$ and Unif$(0, 2\theta)$ as examples, show likelihood graphs with large and small samples that give roughly the same information about the parameters.

   c) Explain how information gives us a large-sample estimate for the variance of the maximum likelihood estimate. Compare to the exact variances for the mle's based on samples from $N(\theta, (14)^2)$ and Unif$(0, 2\theta)$.

   d) Define the Cramer-Rao lower bound on the variance of an unbiased estimate, and its connection to information.

   e) Give the large-sample approximate distribution for the maximum likelihood estimates for exponential family distributions. Show this approximation is exact for iid $N(\theta, (14)^2)$ data, reasonably good for iid Exponential$(1/\theta)$ data with $n = 100$, and not appropriate for iid Unif$(0, 2\theta)$ data.

3. **Unbiased estimates and Rao Blackwell** (8.3, 8.4, 8.8)

   a) Define the notion of an unbiased estimate and an asymptotically unbiased estimate. For $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 2\theta)$, show the mle is biased, but asymptotically unbiased.

   b) Define expected (mean) squared error and its relationship to bias and variance. Explain how a biased estimate might be preferable to an unbiased estimate, based on expected squared error.

   c) For $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 2\theta)$, show the mle $\hat{\theta} = X_{(n)}/2$ is biased, but has lower expected squared error than the unbiased method of moments estimate $\bar{X}$.

   d) State the Rao-Blackwell theorem and demonstrate how to get an improved unbiased estimator for the $\text{Unif}(0, 2\theta)$ data by Rao-Blackwellizing $\bar{X}$ (or X_1).

   e) Give a proof of the Rao-Blackwell theorem, and explain the take-away with regards to sufficient statistics.

4. **Prior and Posterior distributions, Conjugate families** (8.6)

   a) Explain the role of the likelihood function in Bayesian inference. Show that the posterior density is always proportional to the likelihood function multiplied by the prior density. Also show that, when considering two parameter values θ_o and θ_1, the posterior odds equals the prior odds multiplied by the likelihood ratio:

$$\frac{f_{\theta|y}(\theta_o|y)}{f_{\theta|y}(\theta_1|y)} = \frac{p(\theta_o)}{p(\theta_1)} \frac{L(\theta_o)}{L(\theta_1)}$$

   b) Show that conditioning on a sufficient statistic is equivalent to conditioning on all of the data. As an example, suppose $X_1, \ldots, X_n|\theta \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$, with prior distribution $\theta \sim \text{Gamma}(\alpha, \alpha/\mu)$. Show that $Y = \sum X_i$ is sufficient and that the distribution of $\theta|x_1, \ldots, x_n$ is the same as the distribution of $\theta|y$.

   c) Define a conjugate family of prior distributions. For example, explain how we can see from part b that Gamma is conjugate for the Poisson rate. Show that the marginal distribution for $Y = \sum X_i$ is Negative Binomial.

   d) Show how to identify a conjugate family from the form of the likelihood function. Demonstrate for the binomial likelihood function and show that $\text{Beta}(a, b)$ is a conjugate family of prior distributions for the success probability. Show how a and b act like prior successes and failures when computing the posterior mean.

   e) Show that reciprocal-Gamma is conjugate for the variance of a Normal distribution. For $X|V \sim N(0, V)$ with prior distribution $V \sim \text{Inv-Gamma}(\frac{\nu}{2}, \frac{\nu\zeta}{2})$, show using representation that the marginal distribution for $T = \zeta^{-1/2}X$ is $T \sim t_{(\nu)}$.

5. **Objective Bayesian inference and Jeffrey's prior** (8.6 + notes)

Bayesian methods allow us to make inference about very complicated problems, but require that we specify a prior density for the unknown parameters. A prior distribution could incorporate expert information about the likely values of an unknown parameter, but it is also valuable to have a "reference" prior distribution that adds as little information as possible. Objective Bayes methods make use of non-informative priors in an attempt to construct Bayesian estimates that have good frequentist properties, regardless of the true parameter values. Jeffreys' prior is a common choice for a non-informative prior distribution.

a) For $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, \sigma^2)$, we have seen that the prior distribution $\theta \sim N(\mu, \tau^2)$ leads to a Normal posterior density (and is therefore conjugate) with $\theta|\bar{x} \sim N(B\mu + (1-B)\bar{x},\ (1-B)\frac{\sigma^2}{n})$, for $B = \frac{\sigma^2/n}{\sigma^2/n+\tau^2}$. For example, with the 2024 NFL data on home team winning margin, we might assume $\theta \sim N(2.5, (0.5)^2)$ as a prior distribution on the mean winning margin for the home team. Find a 95% posterior interval estimate for θ, having observed $\bar{x} = 1.95$ with $n = 272$ and $\sigma = 14.0$ assumed known.

b) Compare the prior density, likelihood function and posterior density graphically. Show what happens as the prior variance τ^2 increases. Show that the limit of the posterior density as $\tau \to \infty$ is the normalized likelihood function, and is a proper Normal distribution. Note that the implied limit of the prior distribution is $p(\theta) \propto c$ and is improper (it does not have a finite integral). This is an example of Jeffreys' non-informative prior.

c) Show that the information for θ does not depend on the value of θ. Now consider the variance σ^2 to be unknown, and note how the information for σ^2 increases as the σ^2 decreases. Explain how a constant prior density seems more appropriate for θ than for σ^2. Define $\phi = \log(\sigma^2)$ and show the information for ϕ does not depend on ϕ. Show that an improper constant prior density for ϕ implies $p_{\sigma^2}(\sigma^2) \propto (\sigma^2)^{-1}$.

d) Jeffreys' prior density for a parameter θ is the square root of the information for θ. If the information is constant in θ (as with the Normal mean) then $p(\theta) \propto c$. Otherwise, $p(\theta)$ is the prior density implied by a constant prior density for $\phi = g(\theta)$, where ϕ is chosen so that the information for ϕ is constant. Show this is true for the Normal variance. That is, show Jeffreys' prior is $p(\sigma^2) \propto (\sigma^2)^{-1}$.

e) Show how Jeffreys' prior leads to Bayes and Frequentist symmetries. Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, \sigma^2)$. With σ known, $Z = \frac{\bar{X}-\theta}{\sigma/\sqrt{n}}|\theta \sim N(0,1)$ if we treat θ as fixed and $\bar{X}$ as random. Assuming $p(\theta) \propto c$, we have $Z = \frac{\bar{x}-\theta}{\sigma/\sqrt{n}}|\bar{x} \sim N(0,1)$, treating $\bar{x}$ as fixed and θ as random. With joint prior density $p(\theta, \sigma^2) \propto 1/\sigma^2$, the t statistic follows a $t_{(n-1)}$ distribution whether data or parameters are considered random.