1. Suppose $X_1, \ldots, X_n$ are i.i.d Poisson($\theta$) random variables.

a) Write down the likelihood function $L(\lambda)$ and the log-likelihood function $l(\theta)$. Given a model specification, the likelihood function represents all of the information about the unknown parameters provided by the data. Show that, for an independent Poisson sample, the shape of the log-likelihood function depends on the data only through the sufficient statistic $S = \sum X_i$.

b) To prove explicitly that $S$ is a sufficient statistic for $\theta$, you must show that the conditional distribution of $X_1, \ldots, X_n$, given $S = s$, does not depend on $\theta$. This will be true if and only if the “factorization theorem” requirement holds (8.8.1 in Rice). Give some intuition for this, and argue that the MLE must depend on the data only through a sufficient statistic. Note that any 1:1 function of a sufficient statistic is also sufficient.

c) Find the conditional distribution of $X_1, \ldots, X_n$, given $S = s$, and describe how you could generate $X_i$’s from this distribution for given values of $n$ and $s$ (e.g., $n = s = 10$). Explain how this could be used to generate “replicate” Poisson samples (i.e., random samples of size $n$ from a Poisson($\theta$) distribution that would all have the same $s$).

d) The number of turnovers in an NFL game (interceptions and fumbles lost by both teams) is thought to follow an approximate Poisson distribution. A random sample of $n = 16$ NFL games in 2009 yielded the following counts (sorted smallest to largest):

$$0, 1, 1, 2, 2, 2, 2, 2, 4, 4, 4, 5, 5, 5, 6, 7$$

Graph the likelihood function for the mean number of turnovers per game, and identify the MLE. To check whether a Poisson model is appropriate, generate 1000 new Poisson samples with the same sufficient statistic. For each simulated sample, compute the sample variance and look at their distribution. Compare the observed sample variance to this distribution to see if the sample data seems unusual.

e) The expected value of the second derivative of the negative log-likelihood function is called the Fisher Information. Theory shows that the asymptotic variance of the MLE is the inverse Fisher Information. Find the Fisher Information in the Poisson context, and compare it to the variance of the MLE. Find the estimated variance of the MLE for the NFL turnover data.

2. Bayesian Inference. One way to incorporate prior information into an analysis is to specify a prior density for the unknown parameter. For example, with the NFL turnover data, we could use data from the previous season to see that the mean number of turnovers per game is about $\theta = 3$. We can then specify a prior density with mean 3, but with some variability to represent our uncertainty about $\theta$ in the current season.

a) The Gamma($\alpha, \lambda$) density is conjugate to the Poisson distribution because it has the same functional form as the likelihood function for $\theta$. Suppose we specify the prior distribution $\theta \sim$ Gamma($\alpha, \lambda$), with $\alpha = 3$ and $\lambda = 1$. Find the posterior density for $\theta$, given the $n = 16$ turnover counts. Plot the prior density, the likelihood function and the posterior density all on the same graph. Note that each is proportional to a Gamma
density, so you can use `dgamma(x, α, λ)` in R to get functions scaled to have the same integral (likelihood functions are not guaranteed to have a finite integral, but this one does).

b) Write out the posterior mean of θ as a weighted average of the prior mean and the MLE. Note how the weights depend on the sample size and the prior parameters. Also identify the posterior mode, the value of θ that has the highest posterior density, and a 95% posterior interval estimate for θ (in R, `qgamma(p, α, λ)` returns the value for which the Gamma(α, λ) CDF takes the value p).

c) If we want to inject minimal information into an analysis, we can use “objective” Bayes methods. This involves identifying an appropriate “non-informative” prior density (often improper). One approach is to identify a transformation \( φ = g(θ) \) for which the Fisher information is constant with respect to φ. All data sets of the same size provide the same information about φ, regardless of the value of φ. This is a scale where it makes sense to say “non-informative” means “all values of φ are equally likely,” and specify a Uniform prior density. Show that this is true for the Poisson model with \( g(θ) = \sqrt{θ} \). That is, find the Fisher information for \( φ = \sqrt{θ} \), and show that it does not depend on the value of φ.

d) To illustrate the difference, evaluate the log-likelihood values for \( θ \) and for \( φ \) on a grid (say, \( θ = 0.01 \) to \( θ = 16 \), and \( φ = 0.01 \) to \( φ = 4 \)). Do this once for the value of s from the NFL data, and again using a value half as large (say, if you were estimating the mean number of turnovers for a single team in a game). Graph the two log-likelihood functions for \( θ \) on the same axis, and the two for \( φ \) on the same axis (subtract off the maximum value for each log-likelihood vector so that they all have maximum value 0). Note that the curvatures differ for the two \( θ \) log-likelihoods, but are the same for the two \( φ \) log-likelihoods. Also compare the likelihood functions and show that they also have the same second derivative at the mode.

e) Assuming an improper Uniform prior density \( p_φ(φ) \propto c \), find the form of the implied prior density \( p_θ(θ) \), where \( θ = φ^2 \). Use this to get a new posterior density for \( θ \), and add it to the graph you made in part a.

f) A prior density that satisfies the requirements we applied here (specifying a Uniform prior density on a transformation of \( θ \) that has constant information) is called “Jeffreys’ prior” (Lord Jeffreys once wrote: “It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude.”) The problem simplifies to setting \( p(θ) \) proportional to the square root of the Fisher Information. Show that this yields the same prior density you found in part e.

3. A very common model for data is \( X_1, \ldots, X_n \sim N(μ, σ^2) \). For example, in NFL games, the distribution of the total points scored (by both teams) is approximately Normal. For the \( n = 256 \) regular season games in 2009, the average and standard deviation of total points were \( \bar{x} = 42.9 \) and \( s = 13.5 \).

a) Write down the joint likelihood function for \( μ \) and \( σ^2 \) and find the joint maximum likelihood estimates \( \hat{μ} \) and \( \hat{σ^2} \). Make a contour plot of the joint likelihood function for the NFL point data, and identify the MLE’s.
b) Show that the MLE for $\sigma^2$ is the square of the MLE for $\sigma$. In general, the MLE is invariant under 1:1 transformations ($g(\bar{\theta}) = g(\hat{\theta})$). Also note that the MLE is not always unbiased, but is asymptotically unbiased.

c) Show that $\mu$ and $\sigma$ are jointly sufficient for $\mu$ and $\sigma$. Find the conditional distribution of $X_1, \ldots, X_n \mid \bar{x}, s$ and describe how you might go about simulating replicate data sets with the same sample mean and standard deviation.

d) Find the inverse Fisher Information values and the true variances for $\hat{\mu}$ and $\hat{\sigma}^2$.

e) For Normal data the MLE’s for $\mu$ and $\sigma^2$ are independent, so we consider the the marginal likelihood functions to derive independent Jeffreys’ priors for $\mu$ and $\sigma^2$. That is, find the Jeffreys’ prior for each, and multiply them to get a non-informative joint prior density. Find the joint posterior density for $\mu$ and $\sigma^2$. Make a contour plot for the NFL data.

f) Find the marginal posterior distribution for $\sigma^2 \mid \bar{x}, s$, and the conditional posterior distribution of $\mu \mid \sigma, \bar{x}, s$. Describe how you could use these distributions to simulate $\mu, \sigma^2$ pairs from their joint posterior distribution.

A broad collection of probability distributions belong to the exponential family of distributions (Rice, p. 308-309). Most of the distributions we use belong to this family, and there are nice properties guaranteed for the behavior of the likelihood function and the MLE when using such distributions. The data distribution in the following example does not satisfy the requirements (why not?) and we see some of the problems that arise.

4. A common problem in the field of paleo-biology is to estimate when a species went extinct. The depth at which a fossil is found can be converted to a date, so it is sufficient to consider the depths as a measure of age. Suppose $n$ fossils of a particular species have been found, and that the closest to the surface was $x(n)$ meters above the deepest dig (consider that to be the 0 level). Assume that the fossil finds were uniformly distributed between 0 and $\theta$, where $\theta$ corresponds to the depth/time where the species actually went extinct.

a) Write down the likelihood function for $\theta$ based on $X_1, \ldots, X_n$, the distances in meters each find is above the 0 point. Sketch this likelihood function.

b) Identify a sufficient statistic for $\theta$, and the MLE. Also find two unbiased estimates for $\theta$, one based on the MLE and on based on $\bar{X}$ (a “method of moments” estimate). Which do you prefer? Will the sampling distribution of either estimate become approximately Normal when $n$ is large?

c) The Rao-Blackwell theorem (Rice, p. 310) says that any unbiased estimate with a finite variance can be improved by taking its conditional expectation given the value of a sufficient statistic. Show that in this example, ‘Rao-Blackwellizing’ the estimate based on $\bar{X}$ yields the estimate based on the MLE.

d) Write down the CDF for $X(n)$ and draw a sketch. Determine the value $\theta^*$ for which the observed $X(n)$ is at the 10th percentile of its sampling distribution. Explain why the interval $(x(n), \theta^*)$ represents a 90% Confidence Interval (CI) for $\theta$. Compute this interval for $n = 4$ and $x(4) = 100$. 

e) Suppose that you have some “prior” information about the parameter $\theta$ and specify an Inverse-Gamma($\alpha, \lambda$) prior distribution. Find the posterior density for $\theta \mid X_1, \ldots, X_n$.

f) Graph the posterior density and find a 90% posterior interval based on $n = 4$ and $x(4) = 100$, and assuming $\alpha = 3$ and $\lambda = 600$. A scale-invariant improper prior density has $\alpha = \lambda = 0$. Compare the likelihoods and your intervals for these two prior specifications, and repeat the comparison with $n = 50$.

(Problem to turn in) Suppose you observe $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Bin}(1, \theta)$.

a) Write out the likelihood function and log-likelihood function for $\theta$, and show using the factorization theorem that $S = \sum X_i$ is a sufficient statistic for $\theta$. Graph the likelihood function for $n = 20$ and $s = 5$, and for $n = 100$ and $s = 25$.

b) Find the conditional distribution of $X_1, \ldots, X_n$, given $S = s$, and verify that it does not depend on $\theta$.

c) If you assume $\theta \sim U(0, 1)$, what is the marginal distribution of $S = \sum X_i$? What is the conditional distribution of $\theta$ given $S = s$?

d) The Uniform density is also a Beta(1,1) density. Work out the posterior distribution for $\theta$ when the prior distribution is Beta($a, b$). Show that the posterior mean is a weighted average of the mean of the Beta distribution and the MLE.

e) You want to estimate a player’s probability $\theta$ of making a free-throw. You are pretty sure $\theta$ is larger than 0.4, and that it isn’t too close to 1, and settle on $\theta \sim \text{Beta}(4, 2)$. After $n = 10$ free-throws, the player made $x = 5$ shots. After $n = 100$ tries, the player made $x = 50$ shots. Find the posterior mean and standard deviation in each case ($n = 10$ and $n = 100$).

f) Graph the prior density, likelihood function, and posterior density on the same axis for each $n$ value. Here is some R code to help you make the graphs. You’ll need to add/modify commands to include the posterior density.

```r
# create a vector of theta values ranging from 0 to 1 in steps of 0.01.
n=10; x=5; theta = seq(0, 1, 0.01)
pttheta = dbeta(theta, 4, 2)
Ltheta = dbeta(theta, x+1, n-x+1)
# note that the likelihood function is proportional to a Beta density.
plot(theta, Ltheta, type="l", xlab = "theta", ylab="")
lines(theta, ptheta, col="red")
# add lines to an existing graph.
legend(0,2, c("prior", "likelihood"), fill=c("red","black"))
# position a legend with the upper-left corner at (0,2).
```