

Stat 111 Spring 2011 Week 4: The Likelihood Function

1. Suppose X_1, \dots, X_n are i.i.d Poisson(θ) 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 θ , you must show that the conditional distribution of X_1, \dots, X_n , given $S = s$, does not depend on θ . 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, \dots, 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(θ) 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, 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 the distribution of the simulated sample variances. Compare the observed sample variance to this distribution to see if the sample data seems unusual.

- e) With *Bayesian inference* it is possible to incorporate prior information into an analysis by specifying 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 could then specify a prior density with mean 3, but with some variability to represent our uncertainty about θ in the current season. The Gamma(α, λ) density is *conjugate* to the Poisson distribution because it has the same functional form as the likelihood function $L(\theta)$. Suppose we specify the prior distribution $\theta \sim \text{Gamma}(\alpha, \lambda)$, with $\alpha = 3$ and $\lambda = 1$. Find the posterior density for θ , 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, alpha, lambda)` in R to get functions scaled to have the same integral.
- f) Express 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, alpha, lambda)` returns the value for which the Gamma(α, λ) CDF takes the value p).

2. A very common model for data is $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. 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 2010, the average and standard deviation of total points were $\bar{x} = 44.1$ and $s = 14.0$.
- Write down the joint likelihood function for μ and σ^2 and find the joint maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}^2$. Make a contour plot of the joint likelihood function for the NFL point data, and identify the MLE's.
 - Show that the MLE for σ^2 is the square of the MLE for σ . In general, the MLE is invariant under 1 : 1 transformations ($g(\hat{\theta}) = g(\hat{\theta})$). Also note that the MLE is not always unbiased, but is asymptotically unbiased.
 - Show that the MLEs $\hat{\mu}$ and $\hat{\sigma}$ are jointly sufficient for μ and σ . Find the conditional distribution of $X_1, \dots, X_n | \bar{x}, s$ and describe how you might go about simulating replicate data sets with the same sample mean and standard deviation.
 - The expected value of the second derivative of the negative log-likelihood function is called the *Fisher Information*. For many data distributions, the asymptotic variance of the MLE is the inverse Fisher Information. Find the Fisher Information for the Normal mean and variance and compare the inverse information matrix to the true covariance matrix for the MLEs. Explain the intuition for why the second derivative of the log-likelihood should be related to the variance of the MLE.
 - Last time we showed that, assuming the improper prior density $p(\mu, \sigma^2) \propto 1/\sigma^2$, the marginal posterior density for $\mu | x$ is related to the $t_{(n-1)}$ distribution. Instead of integrating over σ^2 , integrate over μ to find the marginal posterior density for $\sigma^2 | x$ and the conditional posterior distribution for $\mu | \sigma^2, x$. Explain how to use these two distributions to simulate μ and σ^2 from their joint posterior distribution. Demonstrate for the NFL points data.

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.

3. 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 θ , where θ corresponds to the depth where the species actually went extinct.
- Write down the likelihood function for θ based on X_1, \dots, X_n , the distances in meters each find is above the 0 point. Sketch this likelihood function and identify the MLE. Explain why the Fisher information is not useful in this problem.
 - Identify a sufficient statistic for θ . Also find two unbiased estimates for θ , one based on the MLE and one 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 θ^* 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 θ . Compute this interval for $n = 4$ and $x_{(4)} = 100$.
- e) Suppose that you have some “prior” information about the parameter θ and specify an Inverse-Gamma(α, λ) prior distribution. Find the posterior density for $\theta | X_1, \dots, X_n$.
- 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

1. Suppose you wish to generate X and Y from a bivariate Normal distribution with correlation ρ and arbitrary means and variances.
 - a) Show that, for $Z_1, Z_2 \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$, the random variable $Z_{12} = \rho Z_1 + \sqrt{1 - \rho^2} Z_2$ is also $N(0, 1)$, with $\text{Cor}(Z_1, Z_{12}) = \rho$.
 - b) Define \mathbf{Z} to be a 2×1 vector composed of two independent standard Normal random variables. Set $\mathbf{Y} = \mathbf{AZ}$, where $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1 - \rho^2} \end{pmatrix}$. The vector \mathbf{Y} is a linear transformation of the MVN vector \mathbf{Z} , so \mathbf{Y} is also MVN. What are the mean and covariance matrix of \mathbf{Y} ?
 - c) Let $\mathbf{X} = \mathbf{YY}'$. The mean of a matrix random variable is the matrix of mean values for each element of the matrix. Find the mean of \mathbf{X} and relate this to the matrix \mathbf{A} . (Think about how you would describe the *variance* of a matrix random variable).
2. Simulate 1000 (X, Y) pairs with $X \sim N(11, 1)$ and $Y \sim N(68, 4)$, with correlation $\rho = 0.7$ (this is a simulation based on my height vs. shoe length data).

- a) First use the method of problem 1a:

```
# create independent vectors of standard Normal deviates.
z1 = rnorm(1000); z2 = rnorm(1000)
```

```
# define z12 to have correlation 0.7 with z1.
z12 = 0.7*z1 + sqrt(1-0.7^2)*z2
```

```
# create X and Y to have the desired means and standard deviations.
x = 11 + z1
y = 68 + 4*z12
```

Check your sample means, standard deviations and correlation using `mean(x)`, `sd(x)`, and `cor(x,y)`. Graph y vs. x by typing

```
plot(x,y, xlab="shoe length", ylab="height", main="Height vs.
Shoe Lengths (in inches)")
```

To add the least squares line and summarize the fit, type

```
out = lm(y ~ x); abline(out); summary(out)
```

- b) Now consider the general problem of generating correlated vectors of arbitrary dimension. To specify a covariance matrix \mathbf{V} , we must compute a matrix square root $\mathbf{V}^{1/2}$, such that $(\mathbf{V}^{1/2})'\mathbf{V}^{1/2} = \mathbf{V}$. Symmetric square roots require finding the eigenvalues, and an alternative is to find the upper-triangular Choleski decomposition.

We can re-simulate the values in part a using vector and matrix operations. In general, this would work for deviates from any distribution, and for a covariance matrix larger than 2×2 .

```
# create a 1000 x 2 matrix of independent standard Normal deviates.
```

```
z = matrix(rnorm(2000), ncol=2)
```

```
# create the desired mean vectors and covariance matrix.
```

```
muxy = cbind(rep(11,1000), rep(68,1000))
```

```
# compute the covariance between x and y and assemble the covariance matrix.
```

```
sxy = 0.7*4*1
```

```
V = matrix(c(1^2, sxy, sxy, 4^2), ncol=2, byrow=T)
```

```
# Find a matrix square root for V using the upper-triangle Choleski decomposition
```

```
#
```

```
rtV = chol(V)
```

```
xy = muxy + z %*% rtV
```

```
x = xy[,1]; y = xy[,2]
```

```
plot(x,y)
```

Generate similar data with deviates from a Gamma(10,1) distribution (e.g.), redefine z as follows:

```
z1 = (rgamma(1000,10,1)-10)/sqrt(10); z2 = (rgamma(1000,10,1)-10)/sqrt(10)
```

```
z = cbind(z1,z2)
```

Print the two graphs and comment on how they differ.