STAT 352: Data Modeling and Inference

Homework 5. Due Friday, February 25th. I encourage you to type all of your solutions, though this is not necessary. However, you must scan (or photograph) any handwritten portions and upload the files to Canvas. For questions that require `R` code, you must turn in your `R` code on Canvas. Your code must in a .Rmd file.

Question 1 (Computation): *Preamble:* You have already partially analyzed (or read about partial analyses) of several datasets in this class. We are going to revisit five of them and you are going to perform a likelihood ratio test with the data. The null hypothesis will always be a point or subspace null hypothesis. You are to write the appropriate MLE function for the alternative hypothesis. Better, and easier, would be to find the appropriate function in the helper code and modify it for your current purposes. The last problem will be a new model, but one you should have seen in Stat 350, the Gaussian linear model.

The Gamma density (HW 4) is given by

$$f(y;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1}\exp(-\beta y)$$

where the data $y > 0$ and the parameters $\alpha, \beta > 0$ and $\Gamma(a)$ is the gamma function. The parameter α controls the behavior of the density near 0. The Exponential distribution is a special case of the Gamma distribution with $\alpha = 1$.

The (shifted and scaled) Type III logistic density (HW 3) is given by

$$f(y;\mu,s,\alpha) = \frac{1}{\text{Beta}(\alpha,\alpha)}\frac{1}{\delta}\left(\exp\left(\frac{y-\mu}{2\delta}\right) + \exp\left(-\frac{y-\mu}{2\delta}\right)\right)^{-2\alpha},$$

where the data $y \in \mathbb{R}$ and the parameters $\alpha, s > 0$ and $\text{Beta}(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the Beta function. The parameter α is a shape parameter that controls the tails of the distribution. The logistic distribution is a special case of the Type III logistic distribution with $\alpha = 1$.

The Generalized-Gamma density (Lab 5) is given by

$$f(y;d,k,s) = \frac{k}{\Gamma(d/k)s^d} y^{d-1}\exp\left(-\left(\frac{y}{s}\right)^k\right)$$

where the data $y > 0$ and the parameters $d, k, s > 0$ and $\Gamma(a)$ is the gamma function. The parameter k controls the right tail and the parameter d controls the behavior near 0. The Weibull distribution is a special case of the Generalized-Gamma Distribution with $d = k$.

The Hyperbolic density (Lab 3) is given by

$$f(y;\mu,\delta,\alpha) = \frac{1}{2\delta K_1(\alpha)}\exp\left(-\alpha\sqrt{1+\frac{(y-\mu)^2}{\delta^2}}\right)$$

there the parameters $\delta, \alpha > 0$ and K_1 is the modified Bessel function of the second kind with index 1. This distribution is used to model heavy tails and provide a robust measure of center. The usual hypothesis tested when using this distribution is whether μ is equal to some specific value μ_0.

The Gaussian Linear Model (Shiny App 1) has a density for y that is specified in terms of linear function of a covariate x and is given by

$$f(y|x;\alpha,\beta,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(y-\alpha-\beta x)^2\right)$$

where the parameter $\sigma > 0$ and we are assuming that for each observed y there is a known value of x (but, of course, these x values vary over the sample and the population). The usual hypothesis tested is whether β is equal to a specific value β_0, usually β_0 is taken to be 0 and we are testing whether a linear relationship between the x and y values exists.

Useful functions: The file `S352_HW5_help_code_Sp_2021.R` contains the necessary density and MLE functions to fit each of the alternative hypotheses below. It also contains code for accessing the necessary datasets. You job will be to write a version of the MLE function that makes the necessary restriction for the null hypothesis and to perform a Likelihood Ratio Test.

Do the following:

a) Perform a likelihood ratio test at significance level 0.04 on the `melanoma$thickness` data. The assumption is that the data are Gamma distributed and the null hypothesis is that the shape parameter α is 1 and the alternative hypothesis is that it is not.

b) Perform a likelihood ratio test at significance level 0.03 on the log of the `melanoma$thickness` data. The assumption is that the data are Logistic Type III distributed and the null hypothesis is that the shape parameter α is 1 and the alternative hypothesis is that it is not.

c) Perform a likelihood ratio test at significance level 0.05 on the a subset of the `faithful$eruptions` data subsetted for the observations where `faithful$waiting>71`. The assumption is that the data are Generalized-Gamma distributed and the null hypothesis is the Weibull distribution. The null hypothesis is that the shape parameters d and k are equal and the alternative hypothesis is that they are not.

d) Perform a likelihood ratio test at significance level 0.01 on the `acme$acme` data. The assumption is that the data are Hyperbolically distributed and the null hypothesis is that the mean parameter μ is 0 and the alternative hypothesis is that it is not.

e) Perform a likelihood ratio test at significance level 0.001 on the `penguins` data where the response variable is `y=penguins$flipper_length_mm` and the covariate is `penguins$body_mass_g`. The assumption is that the data follow a Gaussian Linear Model and the null hypothesis is that the slope parameter β is 0 and the alternative hypothesis is that it is not.

Question 2 (Theory): *Preamble:* We are going to think about the LRT in the context of the five different models for different kinds of data from HW 3 and HW4. For each question, I will give you the density (mass) function for a single data point, the log-likelihood for an observed data vector, and the MLE. You will notice that these are all examples where the sample mean is sufficient for the parameter, as evidenced by the part of the log-likelihood function that I have sectioned off as $n \times [\cdots]$. Your task is to compute the LRT test statistic λ (given by twice the difference in log-likelihoods) for a point null hypothesis and write it in terms of the MLE and the null value.

Useful formulas and facts:

$$\begin{aligned}
\exp(a+b) &= \exp(a)\exp(b) \\
\exp(\textstyle\sum a_i) &= \textstyle\prod \exp(a_i) \\
\log(ab) &= \log(a) + \log(b) \\
\log(\textstyle\prod a_i) &= \textstyle\sum \log(a_i) \\
\log(a^b) &= b\log(a) \\
\exp(a)^b &= \exp(ab) \\
\log(\exp(b)) &= b \\
\exp(\log(a)) &= a \\
k! &= 1 \times 2 \times 3 \times \cdots \times k \\
\binom{n}{k} &= \frac{n!}{k!(n-k)!}
\end{aligned}$$

The general formula for the LRT test statistic for testing a point null hypothesis $H_0 : \theta = \theta_0$ versus and alternative hypothesis $H_A : \theta \neq \theta_0$ (for a single parameter family of distributions, based on observed data $y_1, \ldots, y_n$) is twice the differences in log likelihoods evaluated at the MLEs

$$\lambda(y_1, \ldots, y_n) = 2 \times \left[\ell\left(\hat{\theta}; y_1, \ldots, y_n\right) - \ell\left(\theta_0; y_1, \ldots, y_n\right)\right]$$

Do the following: I highly recommend you look at Lab 5 Section 2.1 for an example of how to do these problems in (probably) the easiest way possible.

a) A random variable Y follows a Bernoulli(p) distribution if Y can take values only 0 or 1 and $p = P(Y = 1)$. This is the most basic model for $0-1$ events like disease diagnosis or whether a machine will work properly when you turn it on. The mass function for a Bernoulli(p)

$$f(y; p) = p^y (1-p)^{1-y}.$$

The parameter p must be between 0 and 1. Let $Y_1, \ldots, Y_n \overset{iid}{\sim}$ Bernoulli(p) with realizations $y_1, \ldots, y_n$. The log-likelihood for p is

$$\ell(p; y_1, \ldots, y_n) = n \times [\bar{y}\log(p) + (1-\bar{y})\log(1-p)]$$

and the MLE is $\hat{p} = \bar{y}$. If the null hypothesis is $H_0 : p = p_0$ and the alternative hypothesis is $H_A : p \neq p_0$, show that the LRT test statistic is

$$\lambda(y_1, \ldots, y_n) = 2n \times \left[\hat{p}\log\left(\frac{\hat{p}}{p_0}\right) + (1-\hat{p})\log\left(\frac{1-\hat{p}}{1-p_0}\right)\right].$$

b) A random variable K follows a Negative-Binomial(r, p) distribution if K counts the number successes are observed before the r-th failure in independent and sequential Bernoulli(p) trials. The random variable K can take any non-negative integer value and r is a fixed positive integer. This is often used to model the number of times a certain type of machine will function properly in sequential tasks before you might want to think about replacing or repairing it. The mass function for a Negative-Binomial(r, p) random variable is

$$f(k; p) = \binom{r+k-1}{k}(1-p)^r p^k.$$

The parameter p must be between 0 and 1. Let $K_1, \ldots, K_n \overset{iid}{\sim}$ Negative-Binomial(r, p) with realizations $k_1, \ldots, k_n$. The log-likelihood for p is

$$\ell(p; k_1, \ldots, k_n) = \sum_{i=1}^{n} \log\left(\binom{r+k_i-1}{k_i}\right) + n \times \left[\bar{k}\log(p) + r\log(1-p)\right]$$

and the MLE is $\hat{p} = \bar{k}/(r+\bar{k})$. If the null hypothesis is $H_0 : p = p_0$ and the alternative hypothesis is $H_A : p \neq p_0$, show that the LRT test statistic is

$$\begin{aligned}\lambda(k_1, \ldots, k_n) &= 2n \times \left[\frac{r\hat{p}}{1-\hat{p}}\log\left(\frac{\hat{p}}{p_0}\right) + r\log\left(\frac{1-\hat{p}}{1-p_0}\right)\right] \\ &= 2n\frac{r}{1-\hat{p}} \times \left[\hat{p}\log\left(\frac{\hat{p}}{p_0}\right) + (1-\hat{p})\log\left(\frac{1-\hat{p}}{1-p_0}\right)\right].\end{aligned}$$

c) A random variable $X > 0$ follows an Exponential(λ) distribution if it is memoryless, which means that $P(X > t+s|X > t) = P(X > s)$. As a waiting time to an event, this means that if you have already waited for t minutes then the probably of waiting a further s minutes is the same as the probability of waiting s minutes if you are just starting to wait. It is often used as a simple first model for waiting times for things like internet queues. The density function for an Exponential(λ) random variable is

$$f(x; \lambda) = \lambda \exp(-\lambda x).$$

The parameter λ must be positive. Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Exponential(λ) with realizations $x_1, \ldots, x_n$. The log-likelihood for λ is

$$\ell(\lambda; x_1, \ldots, x_n) = n \times [\log(\lambda) - \lambda\bar{x}]$$

and the MLE is $\hat{\lambda} = 1/\bar{x}$. If the null hypothesis is $H_0 : \lambda = \lambda_0$ and the alternative hypothesis is $H_A : \lambda \neq \lambda_0$, show that the LRT test statistic is

$$\lambda_{test}(x_1, \ldots, x_n) = 2n \times \left[\log\left(\frac{\hat{\lambda}}{\lambda_0}\right) + \frac{\lambda_0}{\hat{\lambda}} - 1\right].$$

Notice that I subscripted the λ for the test statistic with $_{test}$ so as to not confuse it with the parameter λ. If you are uncomfortable with this, then change the letter representing either the test statistic (the book uses W to invoke Wilks) or of the parameter (I don't know, maybe θ is better?).

d) A random variable $G > 0$ follows a Gamma(m, λ) distribution if G is the total amount of waiting time for m independent and sequential events to occur where the waiting time for each event is distributed as Exponential(λ). An example would be something like the total amount of time for m customers to get through a queue. The density function for an Gamma(m, λ) random variable is

$$f(g; \lambda) = \frac{\lambda^m}{(m-1)!} g^{m-1} \exp(-\lambda g).$$

The parameter λ must be positive and m is a fixed positive integer. Let $G_1, \ldots, G_n \overset{iid}{\sim}$ Gamma(m, λ) with realizations $g_1, \ldots, g_n$. The log-likelihood for λ is

$$\ell(\lambda; g_1, \ldots, g_n) = (m-1) \sum_{i=1}^{n} \log(g_i) - n \log((m-1)!) + n \times [m \log(\lambda) - \lambda \bar{g}]$$

and the MLE is $\hat{\lambda} = m/\bar{g}$. If the null hypothesis is $H_0 : \lambda = \lambda_0$ and the alternative hypothesis is $H_A : \lambda \neq \lambda_0$, show that the LRT test statistic is

$$\lambda_{test}(g_1, \ldots, g_n) = 2n \times \left[m \log\left(\frac{\hat{\lambda}}{\lambda_0}\right) + \frac{m\lambda_0}{\hat{\lambda}} - m \right] = 2nm \times \left[\log\left(\frac{\hat{\lambda}}{\lambda_0}\right) + \frac{\lambda_0}{\hat{\lambda}} - 1 \right].$$

Notice that I subscripted the λ for the test statistic with $_{test}$ so as to not confuse it with the parameter λ. If you are uncomfortable with this, then change the letter representing either the test statistic (the book uses W to invoke Wilks) or of the parameter (I don't know, maybe θ is better?).

e) A random variable C follows a Poisson(λ) distribution if C counts the number of sequential events that are observed in a fixed window of time if the waiting times between events are independent and follow an Exponential(λ) distribution. This is often used as a basic model for the number of customers that get through a queue in an hour. The random variable C can take any non-negative integer value. The mass function for a Poisson(λ) random variable is

$$f(c; \lambda) = \frac{\lambda^c}{c!} \exp(-\lambda).$$

The parameter λ must be positive. Let $C_1, \ldots, C_n \overset{iid}{\sim}$ Poisson(λ) with realizations $c_1, \ldots, c_n$. The log-likelihood for λ is

$$\ell(\lambda; c_1, \ldots, c_n) = -\sum_{i=1}^{n} \log(c_i!) + n \times [\log(\lambda)\bar{c} - \lambda]$$

and the MLE is $\hat{\lambda} = \bar{c}$. If the null hypothesis is $H_0 : \lambda = \lambda_0$ and the alternative hypothesis is $H_A : \lambda \neq \lambda_0$, show that the LRT test statistic is

$$\lambda_{test}(c_1, \ldots, c_n) = 2n \times \left[\hat{\lambda} \log\left(\frac{\hat{\lambda}}{\lambda_0}\right) + \lambda_0 - \hat{\lambda} \right] = 2n\hat{\lambda} \times \left[\log\left(\frac{\hat{\lambda}}{\lambda_0}\right) + \frac{\lambda_0}{\hat{\lambda}} - 1 \right].$$

Notice that I subscripted the λ for the test statistic with $_{test}$ so as to not confuse it with the parameter λ. If you are uncomfortable with this, then change the letter representing either the test statistic (the book uses W to invoke Wilks) or of the parameter (I don't know, maybe θ is better?).