STAT 352: Data Modeling and Inference

Homework 6. Due Friday, March 4th. 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 invert the likelihood ratio test to form confidence intervals.

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 $y > 0$ 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 $d = k$. An alternative parameterization that we are going to use today uses $\alpha = d/k$ and its density is given by

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

and the Weibull is given by the restriction that $\alpha = 1$.

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 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 `HW6_352Fa2022_helper_code.R` contains the necessary density and MLE functions to fit each of the null hypotheses for the tests we want to invert. It also contains code for accessing the necessary datasets. You job will be to write a function that outputs the Likelihood Ratio Test statistic and invert that function in order to form a confidence interval. Recall that there were MLE functions for the alternative hypotheses provided with HW5. Except for problem c), which has been reparameterized using $\alpha = d/k$, the parameterizations of these functions coincide between HW5 and HW6.

Do the following:

a) Invert the Likelihood Ratio Test to form a confidence interval with confidence level $1 - 0.04 = 0.96$ for the shape parameter α when a Gamma distribution is used to model the `melanoma$thickness` data.

b) Invert the Likelihood Ratio Test to form a confidence interval with confidence level $1 - 0.03 = 0.97$ for the shape parameter α when a Logistic Type III distribution is used to model the log of the `melanoma$thickness` data.

c) Invert the Likelihood Ratio Test to form a confidence interval with confidence level $1 - 0.05 = 0.95$ for the shape parameter α when a Generalized-Gamma distribution is used to model a subset of the `faithful$eruptions` data subsetted for the observations where `faithful$waiting>71`.

d) Invert the Likelihood Ratio Test to form a confidence interval with confidence level $1 - 0.01 = 0.99$ for the location parameter μ when a Hyperbolic distribution is used to model the `acme$acme` data.

e) Invert the Likelihood Ratio Test to form a confidence interval with confidence level $1 - 0.001 = 0.999$ for the slope parameter β when a Gaussian Linear Model is used to model the `penguins` data where the response variable is `y=penguins$flipper_length_mm` and the covariate is `penguins$body_mass_g`.

Question 2 (Theory): *Preamble:* We are going to think about inverting the LRT to form a confidence interval in the context of the five different models for different kinds of data from HW 3, HW 4, and HW 5. For each question, I will give you the density (mass) function for a single data point, the LRT test statistic, 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 write a function that takes in an observed data vector and confidence level and inverts the LRT to form a confidence interval. In contrast to the computational question, you are not to use MLE functions. Rather, you for this theory question you must hard code the function λ.

Prototype Example: Here is an example for the scale parameter λ of a Weibull distribution when the shape parameter k is known. The test is of the null $\lambda = \lambda_0$ vs the alternative $\lambda \neq \lambda_0$. Recall, the density for a single Weibull random variable is

$$f(z;\lambda) = k\frac{z^{k-1}}{\lambda^k}\exp\left(-\frac{z^k}{\lambda^k}\right).$$

The LRT test statistic from data $z_1, \ldots, z_n$ (written explicitly as a function of the null value) is

$$\lambda(\lambda_0; \mathbf{z}) = 2n \times \left[k\log\left(\frac{\lambda_0}{\hat{\lambda}}\right) + \frac{\hat{\lambda}^k}{\lambda_0^k} - 1\right]$$

where the MLE under the alternative is

$$\hat{\lambda} = \left(\frac{1}{n}\sum_{i=1}^{n} z_i^k\right)^{\frac{1}{k}}.$$

Below is code for computing a confidence interval using the LRT test statistic and a given confidence level. It also includes code for simulating data from a specific true value of the parameter in order to estimate the coverage of the confidence interval construction. This code is also in the file `HW6_352Fa2022_helper_code.R` on Canvas.

```
CI_weibull_scale_fixed_shape = function(z,k=2,conf=0.95){ #shape k is fixed, defaults to 2
  n = length(z) #getting sample size
  cutoff = qchisq(1-conf,1,lower.tail = FALSE)  #or qchisq(conf,1,lower.tail = TRUE)
  lambda_hat = (mean(z^k))^(1/k)  #the MLE under the alternative, does not change
  lrt_stat_minus_cutoff = function(lambda_0){
    if(lambda_0 <=0){  #checking boundaries, lambda must be positive
      return(Inf)
    }else{
      lrt_stat = 2*n*(k*log(lambda_0/lambda_hat)+lambda_hat^k/lambda_0^k-1)
      return(lrt_stat - cutoff)
    }
  }
  interval = c(0,lambda_hat)  #for the lower bound, using the constraint and the MLE
  lower = uniroot(lrt_stat_minus_cutoff,interval,extendInt = "downX",tol = 1e-10,maxiter=1000
  interval = c(0,10)+lambda_hat  #for the upper bound, MLE to MLE+10 (+10 was arbitrary and
  upper = uniroot(lrt_stat_minus_cutoff,interval,extendInt = "upX",tol = 1e-10,maxiter=10000
  return(c(estimate=lambda_hat,lower=lower$root,upper=upper$root))
```

```
}

 #a simulation to test the coverage - just for fun
N_sim = 1000  #number of simulations
out = rep(NA,N_sim)  #container for output
lambda_true = 10  #true scale parameter
k = 2  #fixed shape
n = 100  #fixed sample size
conf = 0.95
for(i in 1:N_sim){  #for loop for simulation
  z = rweibull(n,shape=k,scale=lambda_true)  #random sample
  ci = CI_weibull_scale_fixed_shape(z,k=k,conf=conf)  #getting confidence interval for scale
  out[i] = (ci["lower"]<lambda_true && ci["upper"]>lambda_true)  #TRUE/FALSE for whether int
}
coverage_est = mean(out)  #estimated probability that intervals (that are random because of
 #making a 0.99 confidence interval for coverage using the CLT and the fact the N_sim is lar
sd_coverage_est = sqrt(conf*(1-conf)/N_sim)
coverage_ci = coverage_est + c(-1,1)*qnorm(0.995)*sd_coverage_est
print(coverage_ci)
```

Do the following:

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 $H_A : p \neq p_0$, then the LRT test statistic is

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

Write a function to compute the confidence interval for p by inverting the LRT. Make sure your function takes in a data vector and a confidence level. Use the data `set.seed(345854); y=rbinom(142,1,0.3)` to test your function and form a 0.95 confidence interval for p (whose true value is 0.3). Interpret the interval you formed.

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 $H_A : p \neq p_0$, then the LRT test statistic is

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

Write a function to compute the confidence interval for p by inverting the LRT. Make sure your function takes in a data vector and a confidence level as well as the parameter r which is fixed. Use the data
`set.seed(345854); k=rnbinom(142,10,1-0.3)` to test your function and form a 0.95 confidence interval for p (whose true value is 0.3) when $r = 10$. Interpret the interval you formed.

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 $H_A : \lambda \neq \lambda_0$, then the LRT test statistic is

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

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?).

Write a function to compute the confidence interval for *lambda* by inverting the LRT. Make sure your function takes in a data vector and a confidence level. Use the data
`set.seed(345854); x=rexp(142,6.4)` to test your function and form a 0.95 confidence interval for λ (whose true value is 6.4). Interpret the interval you formed.

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 $H_A : \lambda \neq \lambda_0$, then the LRT test statistic is

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

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?).

Write a function to compute the confidence interval for λ by inverting the LRT. Make sure your function takes in a data vector and a confidence level as well as the parameter m which is fixed. Use the data `set.seed(345854); g=rgamma(142,12,6.4)` to test your function and form a 0.95 confidence interval for *lambda* (whose true value is 6.4) when $m = 12$. Interpret the interval you formed.

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 $H_A : \lambda \neq \lambda_0$, then the LRT test statistic is

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

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?).

Write a function to compute the confidence interval for λ by inverting the LRT. Make sure your function takes in a data vector and a confidence level. Use the data `set.seed(345854); c=rpois(142,6.4)` to test your function and form a 0.95 confidence interval for λ (whose true value is 6.4). Interpret the interval you formed.