STAT 513 hw 5

1. Let $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$ be independent random samples from the Normal(μ_1, σ^2) and Normal(μ_1, σ^2) distributions, respectively, where μ_1, μ_2, and σ^2 are unknown. Suppose it is of interest to test the hypotheses H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$.

   (a) Give the likelihood function $L(\mu_1, \mu_2, \sigma^2; X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2})$ for $X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}$. *Hint: It is the product of the likelihood functions of the two samples.*

$$
\begin{aligned}
&L(\mu_1, \mu_2, \sigma^2; X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}) \\
&\quad = \textstyle\prod_{i=1}^{n_1} (2\pi\sigma^2)^{-1/2} \exp[-(X_{1i} - \mu_1)^2/(2\sigma^2)] \prod_{j=1}^{n_2} (2\pi\sigma^2)^{-1/2} \exp[-(X_{2j} - \mu_1)^2/(2\sigma^2)] \\
&\quad = (2\pi\sigma^2)^{-(n_1+n_2)/2} \exp[-(\textstyle\sum_{i=1}^{n_1} (X_{1i} - \mu_1)^2 + \sum_{j=1}^{n_2} (X_{2j} - \mu_2)^2)/(2\sigma^2)].
\end{aligned}
$$

   (b) Give the log-likelihood function $\ell(\mu_1, \mu_2, \sigma^2; X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2})$.

$$
\begin{aligned}
&\ell(\mu_1, \mu_2, \sigma^2; X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}) \\
&\quad = -[(n_1 + n_2)/2] \log(2\pi) - [(n_1 + n_2)/2] \log \sigma^2 - [\textstyle\sum_{i=1}^{n_1} (X_{1i} - \mu_1)^2 + \sum_{j=1}^{n_2} (X_{2j} - \mu_2)^2]/(2\sigma^2).
\end{aligned}
$$

   (c) Find the maximum likelihood estimators $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\sigma}^2$ of μ_1, μ_2, and σ^2, respectively; that is, find

$$(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2) = \text{argmax}_{\mu_1, \mu_2, \sigma^2} L(\mu_1, \mu_2, \sigma^2; X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}).$$

   Hint: Use calculus methods on the log-likelihood function.

   We find $\hat{\mu}_1 = \bar{X}_1$, $\hat{\mu}_2 = \bar{X}_2$, and

$$\hat{\sigma}^2 = [\textstyle\sum_{i=1}^{n_1} (X_{1i} - \hat{\mu}_1)^2 + \sum_{j=1}^{n_2} (X_{2j} - \hat{\mu}_2)^2]/(n_1 + n_2)$$

   (d) Under H_0, we have $\mu_1 = \mu_2 = \mu$, say, where μ denotes the common mean. Let $\hat{\mu}_0$ and $\hat{\sigma}_0^2$ be

$$(\hat{\mu}_0, \hat{\sigma}_0^2) = \text{argmax}_{\mu, \sigma^2} L(\mu, \mu, \sigma^2; X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}).$$

   Find expressions for $\hat{\mu}_0$ and $\hat{\sigma}_0^2$.

We get

$$\hat{\mu}_0 = (\sum_{i=1}^{n_1} X_{1i} + \sum_{j=1}^{n_2} X_{2j})/(n_1+n_2) = (n_1\bar{X}_1 + n_2\bar{X}_2)/(n_1+n_2)$$

and

$$\hat{\sigma}_0^2 = [\sum_{i=1}^{n_1}(X_{1i}-\hat{\mu}_0)^2 + \sum_{j=1}^{n_2}(X_{2j}-\hat{\mu}_0)^2]/(n_1+n_2).$$

(e) Show that the likelihood ratio

$$\mathrm{LR}(X_{11},\ldots,X_{1n_1},X_{21},\ldots,X_{2n_2}) = \frac{L(\hat{\mu}_0,\hat{\mu}_0,\hat{\sigma}_0^2;X_{11},\ldots,X_{1n_1},X_{21},\ldots,X_{2n_2})}{L(\hat{\mu}_1,\hat{\mu}_2,\hat{\sigma}^2;X_{11},\ldots,X_{1n_1},X_{21},\ldots,X_{2n_2})}$$

can be simplified to

$$\mathrm{LR}(X_{11},\ldots,X_{1n_1},X_{21},\ldots,X_{2n_2}) = \left[\frac{\sum_{i=1}^{n_1}(X_{1i}-\hat{\mu}_1)^2 + \sum_{j=1}^{n_2}(X_{2j}-\hat{\mu}_2)^2}{\sum_{i=1}^{n_1}(X_{1i}-\hat{\mu}_0)^2 + \sum_{j=1}^{n_2}(X_{2j}-\hat{\mu}_0)^2}\right]^{(n_1+n_2)/2}.$$

We get

$$\begin{aligned}\mathrm{LR}(X_{11},\ldots,X_{1n_1},X_{21},\ldots,X_{2n_2}) &= (\hat{\sigma}^2/\hat{\sigma}_0^2)^{(n_1+n_2)/2} \\ &= \left[\frac{\sum_{i=1}^{n_1}(X_{1i}-\hat{\mu}_1)^2 + \sum_{j=1}^{n_2}(X_{2j}-\hat{\mu}_2)^2}{\sum_{i=1}^{n_1}(X_{1i}-\hat{\mu}_0)^2 + \sum_{j=1}^{n_2}(X_{2j}-\hat{\mu}_0)^2}\right]^{(n_1+n_2)/2}\end{aligned}$$

(f) Show that for any $c \in [0,1]$, there exists a c_1 such that the likelihood ratio test

$$\text{Reject } H_0 \text{ iff } \mathrm{LR}(X_{11},\ldots,X_{1n_1},X_{21},\ldots,X_{2n_2}) < c$$

is equivalent to the test

$$\frac{|\bar{X}_1 - \bar{X}_2|}{S_{\text{pooled}}\sqrt{1/n_1 + 1/n_2}} > c_1.$$

Note: Please just attempt this part. It is quite tricky. You will get points for trying.

We have the following equivalencies:

$$
\begin{aligned}
&\mathrm{LR}(X_{11},\ldots,X_{1n_1},X_{21},\ldots,X_{2n_2}) < c \\
&\iff \frac{\sum_{i=1}^{n_1}(X_{1i}-\hat{\mu}_1)^2+\sum_{j=1}^{n_2}(X_{2j}-\hat{\mu}_2)^2}{\sum_{i=1}^{n_1}(X_{1i}-\hat{\mu}_0)^2+\sum_{j=1}^{n_2}(X_{2j}-\hat{\mu}_0)^2} < c^{2/(n_1+n_2)} \\
&\iff \frac{\sum_{i=1}^{n_1}(X_{1i}-\hat{\mu}_1)^2+\sum_{j=1}^{n_2}(X_{2j}-\hat{\mu}_2)^2}{\sum_{i=1}^{n_1}(X_{1i}-\hat{\mu}_1)^2+n_1(\hat{\mu}_1-\hat{\mu}_0)^2+\sum_{j=1}^{n_2}(X_{2j}-\hat{\mu}_2)^2+n_2(\hat{\mu}_2-\hat{\mu}_0)^2} < c^{2/(n_1+n_2)} \\
&\iff \frac{n_1(\hat{\mu}_1-\hat{\mu}_0)^2+n_2(\hat{\mu}_2-\hat{\mu}_0)^2}{\sum_{i=1}^{n_1}(X_{1i}-\hat{\mu}_1)^2+\sum_{j=1}^{n_2}(X_{2j}-\hat{\mu}_2)^2} > c^{-2/(n_1+n_2)}-1 \\
&\iff \frac{n_1n_2^2(\hat{\mu}_1-\hat{\mu}_2)^2/(n_1+n_2)^2+n_1^2n_2(\hat{\mu}_2-\hat{\mu}_1)^2/(n_1+n_2)^2}{\sum_{i=1}^{n_1}(X_{1i}-\hat{\mu}_1)^2+\sum_{j=1}^{n_2}(X_{2j}-\hat{\mu}_2)^2} > c^{-2/(n_1+n_2)}-1 \\
&\iff \frac{(\hat{\mu}_1-\hat{\mu}_2)^2}{\sum_{i=1}^{n_1}(X_{1i}-\hat{\mu}_1)^2+\sum_{j=1}^{n_2}(X_{2j}-\hat{\mu}_2)^2} > \frac{c^{-2/(n_1+n_2)}-1}{(n_1n_2)/(n_1+n_2)} \\
&\iff \frac{(\hat{\mu}_1-\hat{\mu}_2)^2}{(n_1+n_2-2)S^2_{\text{pooled}}} > \frac{c^{-2/(n_1+n_2)}-1}{(n_1n_2)/(n_1+n_2)} \\
&\iff \frac{(\hat{\mu}_1-\hat{\mu}_2)^2}{S^2_{\text{pooled}}(1/n_1+1/n_2)} > \frac{(n_1+n_2-2)(c^{-2/(n_1+n_2)}-1)}{(1/n_1+1/n_2)(n_1n_2)/(n_1+n_2)} \\
&\iff \frac{|\hat{\mu}_1-\hat{\mu}_2|}{S_{\text{pooled}}\sqrt{1/n_1+1/n_2}} > \underbrace{\sqrt{\frac{(n_1+n_2-2)(c^{-2/(n_1+n_2)}-1)}{(1/n_1+1/n_2)(n_1n_2)/(n_1+n_2)}}}_{c_1}.
\end{aligned}
$$

(g) Provide the value c_1 such that the test in the previous part has size α for any $\alpha \in (0,1)$.

Use $c_1 = t_{n_1+n_2-2,\alpha/2}$.

2. Let $X_1,\ldots,X_n$ be a random sample from the Gamma(α,β) distribution with density

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}\exp(-x/\beta)\mathbb{1}(x>0),$$

where α and β are unknown.

(a) Give the likelihood function $L(\alpha,\beta;X_1,\ldots,X_n)$ for the sample $X_1,\ldots,X_n$.

The likelihood function is given by

$$L(\alpha,\beta;X_1,\ldots,X_n) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} X_i^{\alpha-1}\exp(-X_i/\beta)$$
$$= \Gamma(\alpha)^{-n}\beta^{-n\alpha}\left(\prod_{i=1}^{n} X_i\right)^{\alpha-1}\exp\left(-\sum_{i=1}^{n} X_i/\beta\right)$$

(b) Give the log-likelihood function $\ell(\alpha,\beta;X_1,\ldots,X_n)$ for the sample $X_1,\ldots,X_n$.

The log-likelihood function is given by

$$\ell(\alpha,\beta;X_1,\ldots,X_n) = -n\log\Gamma(\alpha) - n\alpha\log\beta + (\alpha-1)\sum_{i=1}^{n}\log X_i - \sum_{i=1}^{n} X_i/\beta$$

(c) For any $\alpha \geq 0$, let $\hat{\beta}(\alpha)$ be the value of β which maximizes $L(\alpha,\beta;X_1,\ldots,X_n)$. Get an expression for $\hat{\beta}(\alpha)$.

We get $\hat{\beta}(\alpha) = \bar{X}_n/\alpha$.

(d) Consider testing the hypotheses H_0: $\alpha = \alpha_0$ versus H_1: $\alpha \neq \alpha_0$ and let $\hat{\alpha}$ be the maximum likelihood estimator for α. Then the likelihood ratio is given by

$$\mathrm{LR}(X_1,\ldots,X_n) = \frac{\sup_{\{\alpha,\beta:\alpha=\alpha_0,\beta\geq 0\}} L(\alpha,\beta;X_1,\ldots,X_n)}{\sup_{\{\alpha,\beta:\alpha\geq 0,\beta\geq 0\}} L(\alpha,\beta;X_1,\ldots,X_n)}$$
$$= \frac{L(\alpha_0,\hat{\beta}(\alpha_0);X_1,\ldots,X_n)}{L(\hat{\alpha},\hat{\beta}(\hat{\alpha});X_1,\ldots,X_n)}.$$

Show that $-2\log\mathrm{LR}(X_1,\ldots,X_n)$ can be simplified to

$$-2\log\mathrm{LR}(X_1,\ldots,X_n)$$
$$= -2\left[n\log\left(\frac{\Gamma(\hat{\alpha})}{\Gamma(\alpha_0)}\right) + n(\hat{\alpha}-\alpha_0)\left(\log\bar{X}_n - n^{-1}\sum_{i=1}^{n}\log X_i + 1\right) + n\alpha_0\log\alpha_0 - n\hat{\alpha}\log\hat{\alpha}\right].$$

(e) The following R code stores in the vector `X` the survival times of several guinea pigs from the point in time at which they were infected with virulent tubercle bacilli and computes on these data the maximum likelihood estimators $\hat{\alpha}$ and $\hat{\beta}$ for the Gammma(α,β) distribution. The data are taken from Bjerkedal (1960).

```
X <-   c(12, 15, 22, 24, 24, 32, 32, 33, 34, 38, 38, 43, 44, 48, 52,
         53, 54, 54, 55, 56, 57, 58, 58, 59, 60, 60, 60, 60, 61, 62,
         63, 65, 65, 67, 68, 70, 70, 72, 73, 75, 76, 76, 81, 83, 84,
         85, 87, 91, 95, 96, 98, 99, 109, 110, 121, 127, 129, 131,
         143, 146, 146, 175, 175, 211, 233, 258, 258, 263, 297, 341, 341, 376)

library(MASS) # pull in library of functions including the fitdistr() function
fitdistr(X,"gamma") # gives alpha.hat and 1/beta.hat
```

Compute the value of $-2\log \mathrm{LR}(X_1, \ldots, X_n)$ for these data when testing the hypotheses H_0: $\alpha = 1$ versus $H_0 : \alpha \neq 1$.

From the R code

```
a.hat <- fitdistr(X,"gamma")$estimate[1]
a.0 <- 1
n <- length(X)

minus2ll <- -2*(n*log(gamma(a.hat)/gamma(a.0))+n*(a.hat-a.0)*
            (log(mean(X))-mean(log(X))+1)+n*a.0*log(a.0)-n*a.hat*log(a.hat))
```

we get the value 18.38911.

(f) Report the p-value for testing the hypotheses in the previous question, using the asymptotic distribution of $-2\log \mathrm{LR}(X_1, \ldots, X_n)$ under the null hypothesis.

We compute the area under the χ^2_1 distribution to the right of the value 18.38911. It is `1-pchisq(minus2ll,1)` $= 1.800847 \times 10^{-5}$.

(g) Consider testing H_0: $\alpha = \alpha_0$ versus H_1: $\alpha \neq \alpha_0$ using the guinea pig data. Find an interval such that you fail to reject H_0 at the 0.01 significance level for all α_0 in the interval. *Hint: Compute* $-2\log \mathrm{LR}(X_1, \ldots, X_n)$ *over many values of* α_0 *and find those values of* α_0 *(search, say, between* $1/2$ *and* 4*) for which* $-2\log \mathrm{LR}(X_1, \ldots, X_n) < \chi^2_{1,0.01}$.

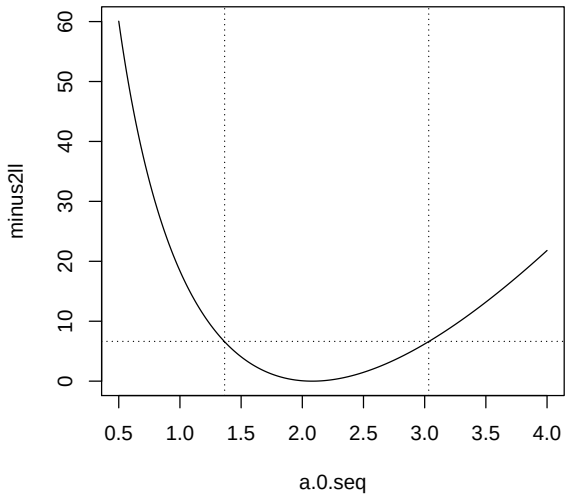
The following R code finds such and interval and also plots the $-2\log \mathrm{LR}(X_1, \ldots, X_n)$ across values of α_0 from $1/2$ to 4.

```
a.0.seq <- seq(1/2,4,length=1000)
minus2ll <- - 2 * (n*log( gamma(a.hat)/gamma(a.0.seq)) + n*(a.hat-a.0.seq)*
    (log(mean(X))-mean(log(X))+1)+n*a.0.seq*log(a.0.seq)-n*a.hat*log(a.hat))

plot(minus2ll~a.0.seq,type="l")
abline(h=qchisq(.99,1),lty=3)

which(minus2ll < qchisq(.99,1))
lower <- min(a.0.seq[which(minus2ll < qchisq(.99,1))])
upper <- max(a.0.seq[which(minus2ll < qchisq(.99,1))])
```

```
abline(v=lower,lty=3)
abline(v=upper,lty=3)
```



We have that $-2 \log \text{LR}(X_1, \ldots, X_n) < \chi^2_{1,0.01}$ when $\alpha_0 \in (1.37, 3.03)$. Note that the interval is approximate.

(h) Give an interpretation of this interval.

This is a 99% confidence interval for α.

(i) Based on these results, do you think it would be reasonable to model these data using the Exponential(β) distribution?

The Exponential(β) distribution is the Gamma(α, β) distribution when $\alpha = 1$, and we reject H_0: $\alpha = 1$ at all significance levels greater than the p-value 1.800847×10^{-5}. Thus the evidence is quite strong that α is not equal to 1.

References

Bjerkedal, T. (1960). Acquisition of Resistance in Guinea Pigs infected with Different Doses of Virulent Tubercle Bacilli. *American Journal of Hygiene*, 72(1), 130-48.