Homework 7 Solution

Math 445, Spring 2017

Chapter 8, Exercise 24

(a)

$p = P(X > 9|\lambda = 0.25) = \int_9^\infty 0.25e^{-0.25x}dx = 0.1054$ Let Y be the number of data values greater than 9. Then the probability of rejecting H_0 when the null is true is $\alpha = P(Y \geq 3|\lambda = 0.25) = \sum_{k=3}^{10} \binom{10}{k} 0.1054^k(1-0.1054)^{10-k} = 0.08$

(b)

If $\lambda = 0.15$, then

$$p_1 = P(X > 9|\lambda = 0.15) = \int_9^\infty 0.15e^{-0.15x}dx = 0.259$$

so the power is

$$P(Y \geq 3|\lambda = 0.15)\sum_{k=3}^{10} \binom{10}{k} 0.259^k(1-0.259)^{10-k} = 0.5$$

Chapter 8, Exercise 30

The Sidak correction discussed by the book is $\alpha^* = 1 - 0.9^{1/12} = 0.0087$. The Bonferroni correction is given by $\alpha^* = .1/12 = 0.0083$.

Chapter 8, Exercise 32

(a)

$T = \frac{\frac{1}{2}x^{-1/2}}{\frac{1}{4}x^{-3/4}} = 2x^{1/4} < c$, which can be rewritten as $x < c$ (note that this is a different c now, so you could use c' to differentiate, but that isn't practically important). To finish the problem, we need to find the value of c that defines a size $\alpha = 0.05$ test and state the rejection region:

$$0.05 = \int_0^k \frac{1}{2}x^{-1/2}dx = k^{1/2} \Longrightarrow c = 0.0025$$

So we will reject H_0 if $x < 0.0025$ and fail to reject otherwise.

(b)

$1 - \beta = P(x < 0.0025|\theta = 1/4) = \int_0^{.0025} \frac{1}{4}x^{-3/4} = 0.224$

Chapter 8, Exercise 37

Under H_A the MLE estimates are $\hat{\mu}_A = \bar{x}$ and $\hat{\sigma}_A^2 = n^{-1}\sum(x_i - \bar{x})^2$. The latter is smaller than s^2 by a factor $(n-1)/n$. Under H_0 the MLE estimates are $\hat{\mu}_0 = \mu_0$ and $\hat{\sigma}_0^2 = (1/n)\sum(x_i - \mu_0)^2 = \hat{\sigma}_A^2 + (\bar{x} - \mu_0)^2$. The likelihood ratio test statistic is

$$T(x) = \frac{L(\bar{x}, \hat{\sigma}_A^2)}{L(\mu_0, \hat{\sigma}_0^2)}.$$

The log of the numerator is

$$\begin{aligned} \log(L(\bar{x}, \hat{\sigma}_A^2)) &= -(n/2)\log(2 \cdot \pi \cdot \hat{\sigma}_A^2) - (1/2)\sum((x_i - \bar{x})^2)/\hat{\sigma}_A^2 \\ &= -(n/2)\log(2 \cdot \pi \cdot \hat{\sigma}_A^2) - (n/2). \end{aligned}$$

The second term reduces to $-n/2$. Similarly, the log of the denominator simplifies to

$$\begin{aligned} \log(L(\mu_0, \hat{\sigma}_0^2)) &= -(n/2)\log(2 \cdot \pi \cdot \hat{\sigma}_0^2) - (1/2)\sum((x_i - \mu_0)^2)/\hat{\sigma}_0^2 \\ &= -(n/2)\log(2 \cdot \pi \cdot \hat{\sigma}_0^2) - (n/2). \end{aligned}$$

With cancellation, the log of the test statistic simplifies to

$$\begin{aligned} \log(T(x)) &= -(n/2)\log(\hat{\sigma}_0^2) + (n/2)\log(\hat{\sigma}_A^2) \\ &= -(n/2)\log(\frac{\hat{\sigma}_A^2 + (\bar{x} - \mu_0)^2}{\hat{\sigma}_A^2} \\ &= -(n/2)\log(1 + (\bar{x} - \mu_0)^2/\hat{\sigma}_A^2) \\ &= -(n/2)\log(1 + (n/(n-1))(\bar{x} - \mu_0)^2/s^2) \\ &= -(n/2)\log(1 + (n/(n-1))nt^2) \end{aligned}$$

where $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$ is the usual t statistic. For $t < 0$ there is a one-to-one relationship between t and the LRT test statistic, so tests based on these statistics are equivalent.

Problem 5

Clouding seeding is the practice of using an airplane to spray certain chemicals into a cloud. The hope is to increase the amount of rain that falls from the cloud. It is difficult to evaluate the effectiveness of cloud seeding. The following data are from one of the best studies.

This study was conducted in southern Florida between 1968 and 1972. The weather was watched daily. If the cloud conditions were considered suitable, then that day was included in the experiment. A total of 52 days were considered suitable. Each suitable day, one target cloud was arbitrarily chosen and randomly assigned to one of two treatments:

seed: fly an airplane through the cloud and spray the seed chemical

control: fly an airplane through the cloud but the sprayer was not loaded with the seed chemical.

The total rain that fell from the target cloud was measured. This experiment was double blind (neither the pilots nor rainfall observers knew which treatment was used). The file `rainfall.csv` on the class web site contains two variables: treatment and rain for each of the 52 clouds in the study.

(a)

Do you expect a problem with the assumption of independent groups?

No, we don't expect a problem with the assumption of independent groups, since the treatments were arbitrarily assigned.

(b)

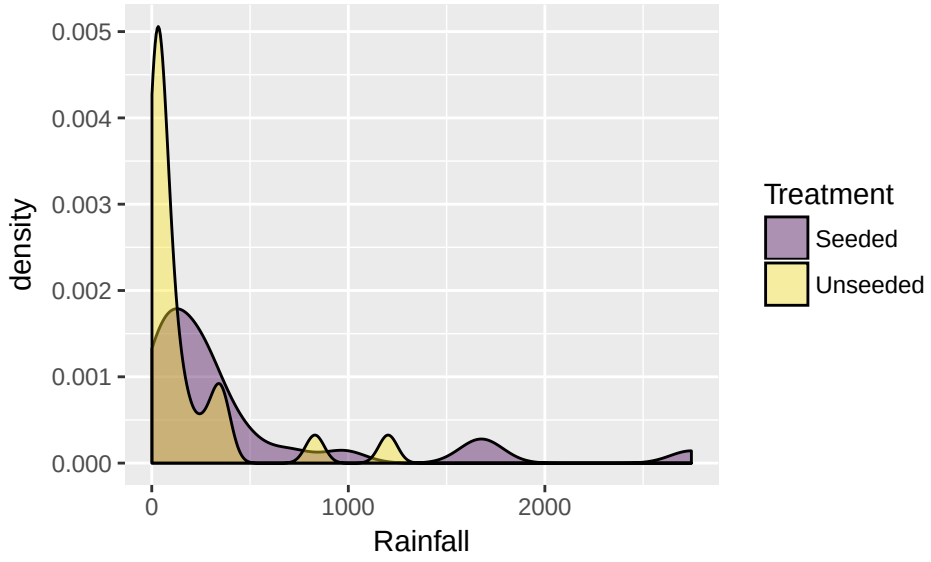
Consider the untransformed rainfall values by treatment. Is a normal distribution reasonable? Explain why or why not.

No, the data are not consistent with the assumption of normality. You can assess this assumption using histograms or Q-Q plots of rainfall by group.

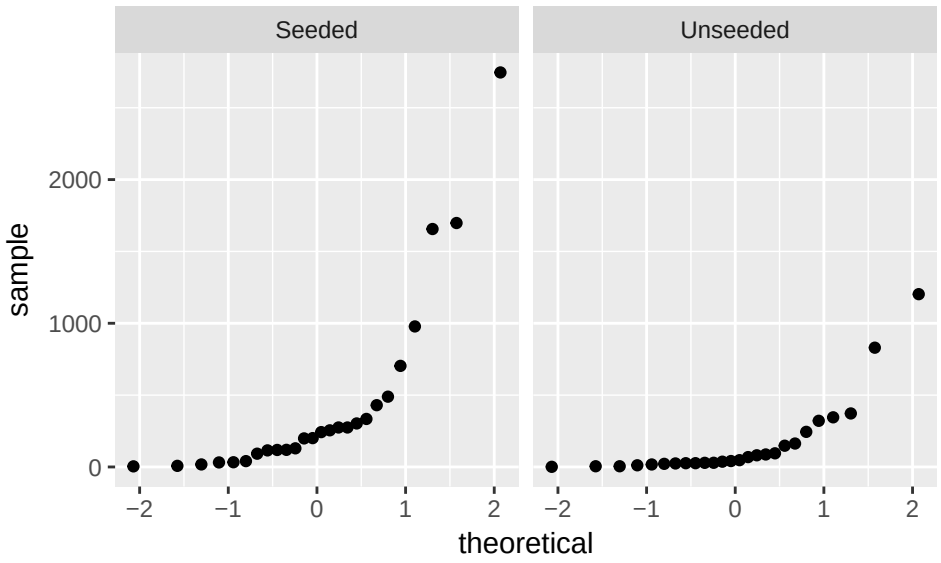
```
rainfall <- read.csv("https://raw.githubusercontent.com/math445-LU/sp17_assets/master/data/rainfall.csv
library(ggplot2)
library(viridis)
```

```
## Loading required package: viridisLite
```

```
ggplot(data = rainfall, mapping = aes(x = Rainfall, fill = Treatment)) +
  geom_density(alpha = 0.4) +
  scale_fill_viridis(discrete = TRUE)
```




```
ggplot(data = rainfall, mapping = aes(sample = Rainfall)) +
  stat_qq() +
  facet_wrap(~Treatment, ncol = 2)
```

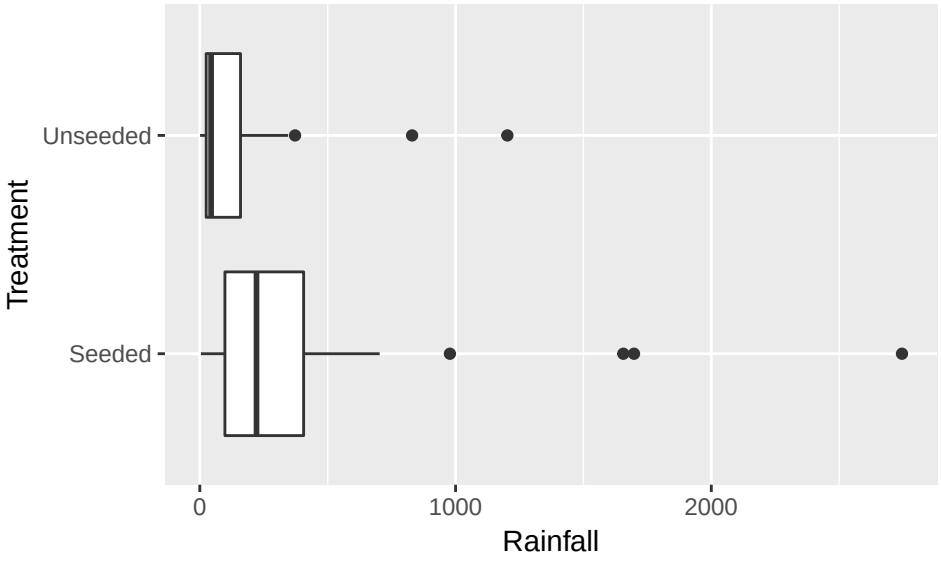


(c)

For the untransformed rainfall values, use a graphical method to assess the assumption of equal variances. Is the assumption reasonable? Explain why or why not.

The assumption of equal variances does not seem warranted. Based on the boxplots below, the seeded treatment group has a larger IQR, which is a robust estimate of the standard deviation.

```
ggplot(data = rainfall, mapping = aes(x = Treatment, y = Rainfall)) +
  geom_boxplot() +
  coord_flip()
```



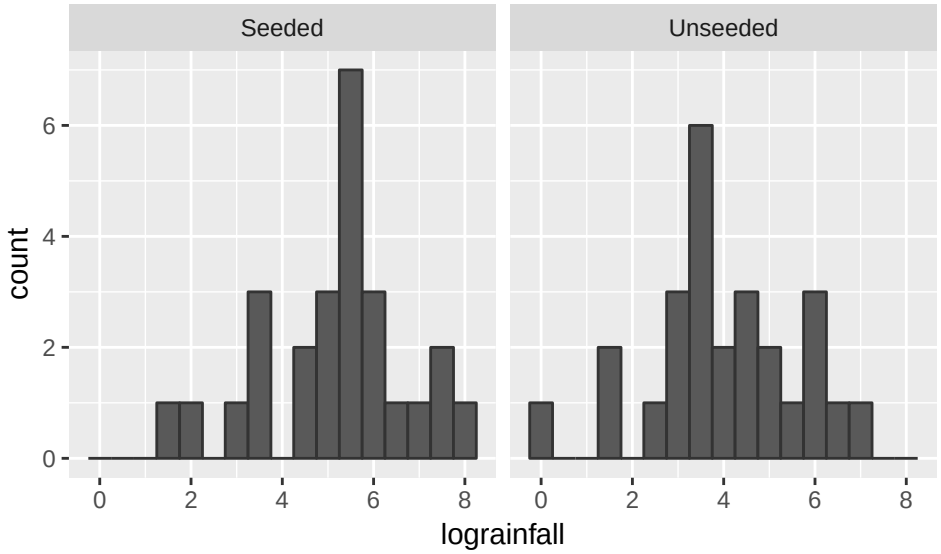
(d)

Create a new variable in the data set, `lograinfall`, that applies a natural log transformation to the rainfall values. Is a normal distribution reasonable for the log transformed values?

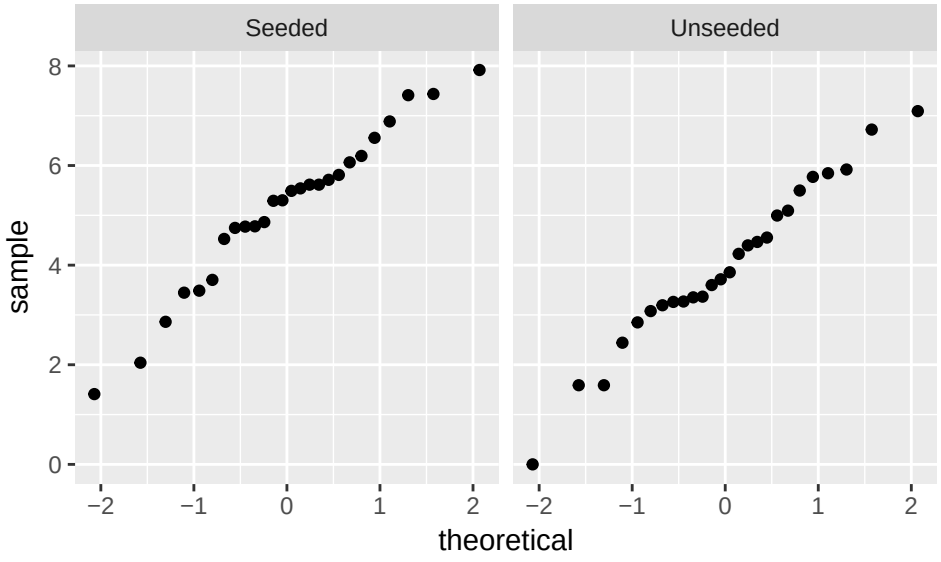
Yes, the log-transformed data seem approximately normal. This is harder to see in the histograms, but the normal Q-Q plots show no substantial deviations from the diagonal.

```
library(dplyr)
rainfall <- mutate(rainfall, lograinfall = log(Rainfall))

ggplot(data = rainfall) +
  geom_histogram(mapping = aes(x = lograinfall), colour = "gray20", binwidth = .5) +
  facet_wrap(~Treatment, ncol = 2)
```




```
ggplot(data = rainfall) +
  stat_qq(mapping = aes(sample = lograinfall)) +
  facet_wrap(~Treatment, ncol = 2)
```

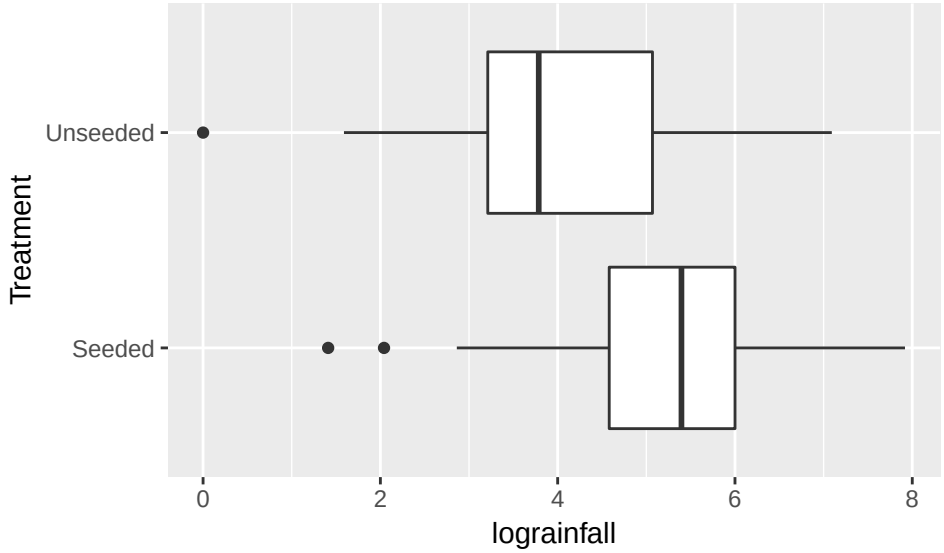


(e)

It is reasonable to assume that the log-transformed values have equal variances? You can use a graphical approach to make this assessment.

Boxplots provide an informal check of this assumption, and the IQRs do not appear to be substantially different.

```
ggplot(data = rainfall) +
  geom_boxplot(mapping = aes(x = Treatment, y = lograinfall)) +
  coord_flip()
```



(f)

Use an appropriate two-sample t-test on the log-transformed rainfall to determine whether seeding effects rainfall.

```
lseeded <- subset(rainfall, select = lograinfall, subset = Treatment == "Seeded", drop = TRUE)
lunseeded <- subset(rainfall, select = lograinfall, subset = Treatment == "Unseeded", drop = TRUE)

test_results <- t.test(x = lseeded, y = lunseeded, var.equal = TRUE)
test_results
```

```
##
##  Two Sample t-test
##
## data:  lseeded and lunseeded
## t = 2.5444, df = 50, p-value = 0.01408
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.240865 2.046697
## sample estimates:
## mean of x mean of y
##  5.134187  3.990406
```

Based on the test statistic $t = 2.54$ with associated p-value 0.0141, we find strong evidence that seeding effects rainfall. Notice that since we conducted a two-sided test, this test does not provide further information.

Problem 6

You are interested in determining who sends more text messages: graduate students or undergraduate students. Results of a survey that measured the variables Y = texts, the number of texts sent per day, and X = grad, a binary (0 or 1) variable indicating whether a student is a grad students (grad=1) or an undergrad (grad=0)

for $n = 91$ Harvard students are in the file `textsHW7.csv`. Let $Y_i \overset{\text{iid}}{\sim} \text{LogNormal}(\mu = \beta_0 + \beta_1 X_i,\ \sigma^2)$. The lognormal distribution has PDF

$$f(y|\mu, \sigma^2) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(\log y - \mu)^2}{(2\sigma^2)}\right\},\ y \in (-\infty, \infty),\ \mu \in (-\infty, \infty),\ \sigma > 0,$$

and $\text{E}(Y) = \exp\left\{\mu + (\sigma^2/2)\right\}$.

(a)

Derive the log-likelihood function. The lognormal log-likelihood was derived on homework XX, the only change that needs to be made is to substitute $\beta_0 + \beta_1 X_i$ for μ.

$$\begin{aligned}
L(\beta_0, \beta_1, \sigma^2|X_1, \ldots, X_n) &= \prod_{i=1}^{n} \frac{1}{Y_i\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2}\left[\log(Y_i) - (\beta_0 + \beta_1 X_i)\right]^2\right\} \\
&= \left(2\pi\sigma^2\right)^{-n/2} \left(\prod_{i=1}^{n} \frac{1}{Y_i}\right) \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}\left[\log(Y_i) - (\beta_0 + \beta_1 X_i)\right]^2\right\} \\
\ell(\beta_0, \beta_1, \sigma^2) &= -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}\log(Y_i) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\log(Y_i) - (\beta_0 + \beta_1 X_i)\right]^2 \\
&= -\frac{n}{2}\left[\log(2\pi) + \log(\sigma^2)\right] - \sum_{i=1}^{n}\log(Y_i) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\log(Y_i) - (\beta_0 + \beta_1 X_i)\right]^2
\end{aligned}$$

(b)

Write a function in R (call it something like `lognormal.loglik` that computes the log-likelihood for a lognormal distribution given three parameters (call it `theta` where `theta[1]` is β_0, `theta[2]` is β_1, and `theta[3]` is σ^2) and the data vectors `y` and `x`.

```
lognormal.loglik <- function(theta, x, y) {
  sum(dlnorm(y, meanlog = theta[1] + theta[2] * x, sdlog = sqrt(theta[3]), log = TRUE))
}
```

(c)

In R, read in the data set and calculate the mean and variance of the number of texts for each group (undergraduate students and graduate students). How many of each type of students are in the sample?

```
# Read in the data and obtain basic summary stats
texts <- read.csv("https://raw.githubusercontent.com/math445-LU/sp17_assets/master/data/textsHW7.csv")

# remember 0 = undergraduate; 1 = graduate
library(dplyr)
texts %>%
  group_by(grad) %>%
  summarize(mean = mean(texts), var = var(texts), n = n())
```

```
## # A tibble: 2 x 4
##    grad     mean      var     n
##   <int>    <dbl>    <dbl> <int>
## 1     0 43.43478 2567.985    69
## 2     1 26.06818 1342.531    22
```

(d)

Use your function from part (b) and the `optim` function to calculate the maximum likelihood estimates for β_0, β_1, and σ^2. What is the value of the log-likelihood function at the ML estimates?

```
MLE <- optim(par = c(1, 1, 1), fn = lognormal.loglik,
             control = list(fnscale = -1), x = texts$grad, y = texts$texts)

MLE
```

```
## $par
## [1]  3.3268154 -0.9447229  1.1510077
##
## $value
## [1] -417.4568
##
## $counts
## function gradient
##      112       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

```
# ML esimates
MLE$par
```

```
## [1]  3.3268154 -0.9447229  1.1510077
```

```
# Value of the likelihood
MLE$value
```

```
## [1] -417.4568
```

(e)

Based on this model, what is the estimated average number of text messages for undergraduates from this model? How about for graduate students?

```
# Compute the average expected number of texts for undergrads.
# You do this using the formula for expected value of a lognormal RV
ugrad_average_texts <- exp(MLE$par[1] + MLE$par[3] / 2); ugrad_average_texts
```

```
## [1] 49.51716
```

```
grad_average_texts <- exp(MLE$par[1] + MLE$par[2] + MLE$par[3] / 2); grad_average_texts
```

```
## [1] 19.25164
```

(f)

In preparation to test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$, create a new R function, similar to the one in part (b), which passes just two parameters in `theta` (now, `theta[1]` is β_0 and `theta[2]` is σ^2). Notice that we no longer include β_1, as we will assume $\beta_1 = 0$ under the null hypothesis. Name this function `lognormal.loglik.null`.

```r
# log-likelihood where beta1 = 0
lognormal.loglik.null <- function(theta, x, y) {
  sum(dlnorm(y, meanlog = theta[1], sdlog = sqrt(theta[2]), log = TRUE))
}
```

(g)

Test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ using a likelihood ratio test. Be sure to clearly show your steps in calculating the test statistic and the p-value (using R will be necessary here), and state your conclusion for the test. The results from part (d) and the function in part (f) will be helpful.

```r
# LRT of H_0: beta1 = 0 vs. H_a: beta1 != 0
# First maximize lognormal.loglik.null
MLE.beta0 <- optim(par = c(1, 1), fn = lognormal.loglik.null,
                   control = list(fnscale = -1), x = texts$grad, y = texts$texts)

MLE.beta0
```

```
## $par
## [1] 3.098117 1.314844
##
## $value
## [1] -423.5076
##
## $counts
## function gradient
##       67       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

```r
# Now compute the LRT stat
lrt_stat <- -2 * (MLE.beta0$value - MLE$value ); lrt_stat
```

```
## [1] 12.10151
```

```r
# Calculate the p-value using the Chi-square dsn w/ 1 df
pvalue <- pchisq(lrt_stat, df = 1, lower.tail = FALSE); pvalue
```

```
## [1] 0.0005038101
```

By Wilk's Theorem (I didn't give the name of the theorem in class, but there it is), -2 times the log of the likelihood ratio follows a chi-square distribution, asymptotically. Using R, we find that this statistic is 12.1 and the pvalue is 0.0005038. The rejection rule for this chi-square test is to reject is the LRT statistic is greater than 3.84.

(h)

Interpret the results of the hypothesis test in part (g). What does this mean for the difference in text messaging between undergraduate and graduate students? Given the tiny p-value, we reject the assertion that the number of texts sent by undergrads and grads is equivalent.