Hickman Problem Set 9

Keith Hickman

October 29, 2017

Problem 1.

a) Treating the data as a simple random sample, find a 95% confidence interval for the percentage of all U.S. adults who support same-sex marriage.

We can assume that $n = 1025$ is a large enough sample for the CLT to kick in and approximate a normal distribution.

```
n <- 1025
x_bar <- .61
sd_error <- sqrt(x_bar * (1-x_bar) / n)

upper_bound <- x_bar + qnorm(.975) * sd_error
lower_bound <- x_bar - qnorm(.975) * sd_error
upper_bound
```

```
## [1] 0.6398596
```

```
lower_bound
```

```
## [1] 0.5801404
```

Thus, we have a 95% confidence that the actual value is between 58% and 63.9%.

b) How large of a simple random sample would we need? To calculate n, we need to solve for $n = 2q\sqrt{\frac{p(1-p)}{n}}$. Using the values from (a) and using algebra to solve for n, we need a sample size of $n = 9046$.

```
q <- qnorm(.975)
L <- .02
n = 9046
nL2 <- 2 * q *(sqrt(x_bar * (1-x_bar) / n))
nL2
```

```
## [1] 0.02010236
```

Problem 2

Trosset 10.a.1 a) If we observe x that results in $\overline{x} = 3.194887$ and $s^2 = 104.0118$, then what is the value of the test statistic?

We should use a t-test because we don't know the variance σ^2 and we have a moderate sample size of $n = 400$.

We can approximate the distribution of μ of our distribution by approximating the Error $\overline{X} - \mu \sim Normal(0, \sigma^2/n)$, converting to standard units by dividing by $\sigma/\sqrt{n}$, we can approximate a Normal(0,1) distribution.
Our test statistic then is:

```
x_bar <- 3.19
mu <- 0
s <- 104.0118
```

```
n <- 400

t.stat <- (x_bar - mu)/(s/sqrt(n))
t.stat
```

```
## [1] 0.6133919
```

b) Choice `v.` is the best, as we're using the t-test vs. a `pnorm` function

c) Here, the p-value is close to alpha, so we may want to consider more evidence before rejecting the null. We have to consider several factors, including; what our objects are being measured, from what population they're measured, what measurements were taken, and what other random variables are relevant to the inferences we need to make. Assuming we're satisfied that the answers to the above questions will safely allow us to reject the null hypothesis, a significance probability of .03 though close to, is less than .05 and we could reject H_0 in favor of the alternative. Tricky!

Problem 3

a) The experimental unit here is individuals (patients) with diabetes. The measurement taken on the patient is the change in glycemic index between time A and time B given the presence of coffee.

b) Given that we are interested in measuring any change in the glycemic index and not necessarily an increase or decrease, the null and alternative hypotheses can be stated as follows: $H_0 : \Delta = 11.5$ and $H_1 : \Delta \neq 11.5$ where Δ is the change in glycemic index.

The test statistic can be calculated as follows. Given that we do not know the population variance, we can approximate

```
x_bar <- 11.5
mu <- 0
s <- 21
n <- 10

t.stat <- (x_bar - mu)/(s/sqrt(n))
t.stat
```

```
## [1] 1.731723
```

Our t-stat is 1.73.

c) The P-value (significance probability) was calculated to be 0.12, so the null hypothesis was not rejected. From this and the other information given, is it correct to conclude that we are sure that on average, dates have the same glycemic index with or without coffee? Explain.

My concern with not rejecting the null is the very small sample size and and assumption of normality. The results are certainly interesting and the significance probability was well above the p-value (assuming either .05 or .1), but it does appear that there was at least some change between the two samples that might warrant further investigation. With such a small sample, we could have drawn samples that gave a skewed distribution.

Problem 4

A one-sample t-test will assume that the distribution is approximately normal. Considering the small sample size here, this might be a risky assumption. We should investigate using either a boxplot or qqplot.

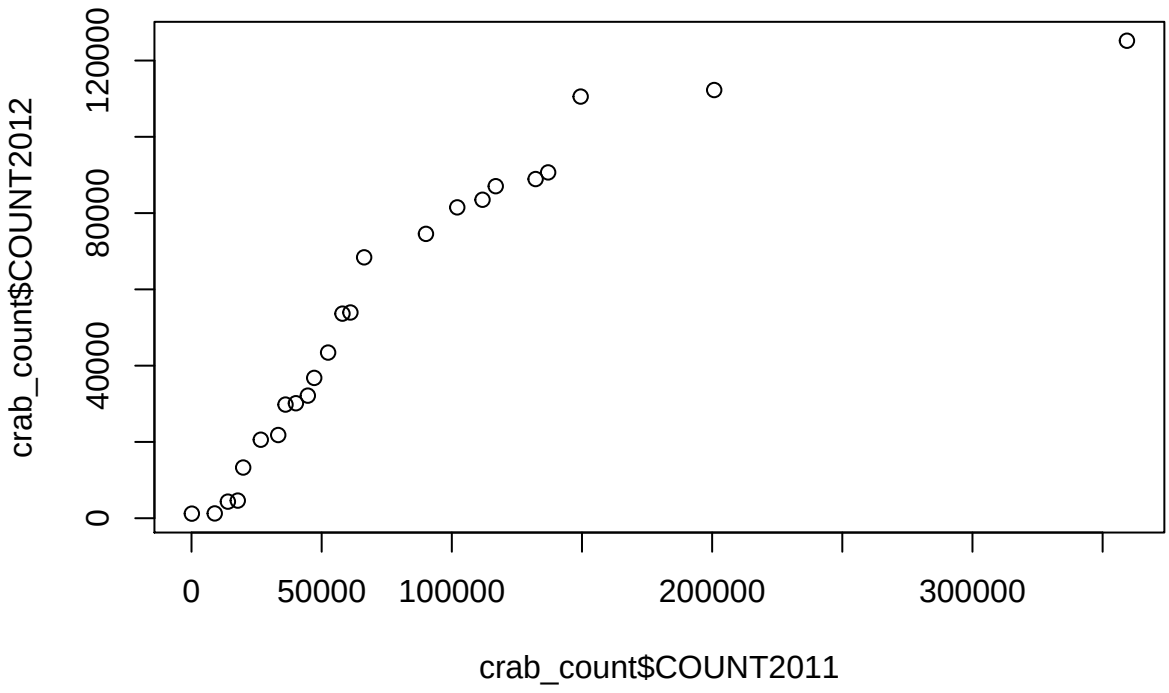
```
library(readr)
crab_count <- read_delim("C:/Users/khickman/Desktop/Personal/IUMSDS/StatsS520/Module10/crab-count.txt",
```

```
## Warning: Missing column names filled in: 'X2' [2], 'X4' [4], 'X6' [6],
## 'X8' [8]
```

```
## Parsed with column specification:
## cols(
##   BEACH = col_character(),
##   X2 = col_character(),
##   COUNT2011 = col_integer(),
##   X4 = col_integer(),
##   COUNT2012 = col_integer(),
##   X6 = col_integer(),
##   CHANGE = col_integer(),
##   X8 = col_character()
## )
```

```
## Warning: 1 parsing failure.
## row # A tibble: 1 x 5 col     row   col  expected    actual expected   <int> <chr>     <chr>      <ch
```

```
crab_count <- crab_count[,c(1,3,5,7)]
qqplot(crab_count$COUNT2011, crab_count$COUNT2012)
```



Here, we have a non-normal distribution, including several outliers. This distribution does not follow a straight line, and therefore indicates that the distribution would not be suitable for a t-test. We should instead be using the sign test, which does not require normality.

Problem 5

a) It won't make much of a difference whether we analyze the natural log vs. the ratios - we could use either. Both exhibit non-normal characteristics. The values are small enough and the length of the dataset short enough that analyzing the log of each number won't add much normality to the distribution. Additionally, the numbers are already ratios, thus there is some division/normalizing taking place and taking the logarithm wouldn't transform the ratios in any meaningful way.

```
shoshoni <- read.csv("shoshoni.csv",header=FALSE)
shoshoni
```

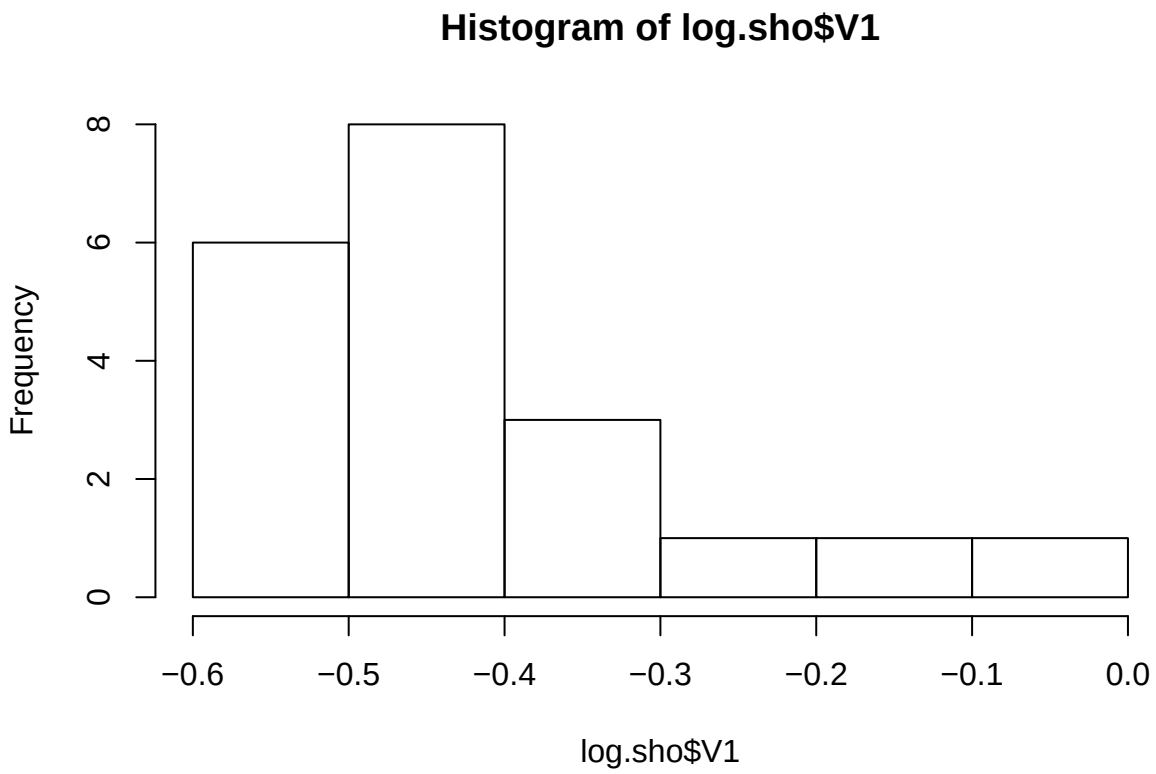
```
##       V1
## 1  0.693
## 2  0.662
## 3  0.690
## 4  0.606
## 5  0.570
## 6  0.749
## 7  0.672
## 8  0.628
## 9  0.609
## 10 0.844
## 11 0.654
## 12 0.615
## 13 0.668
## 14 0.601
## 15 0.576
## 16 0.670
## 17 0.606
## 18 0.611
## 19 0.553
## 20 0.933
```

```
mean(shoshoni$V1)
```

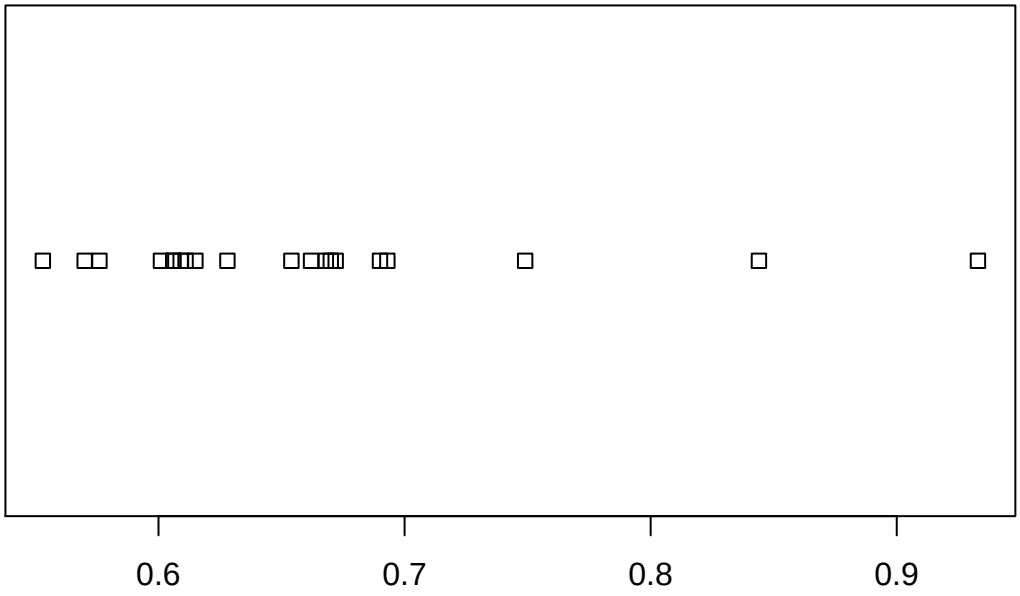
```
## [1] 0.6605
```

```
log.sho <- log(shoshoni)
hist(shoshoni$V1)
```

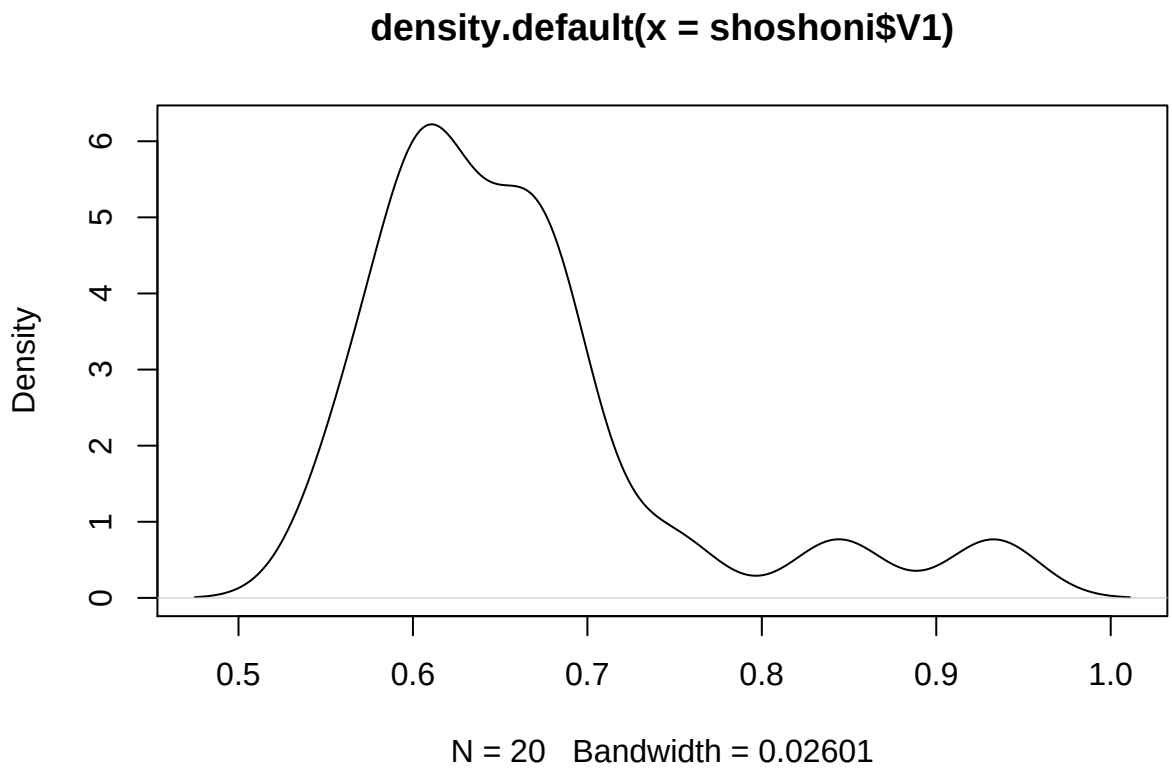

```
hist(log.sho$V1)
```



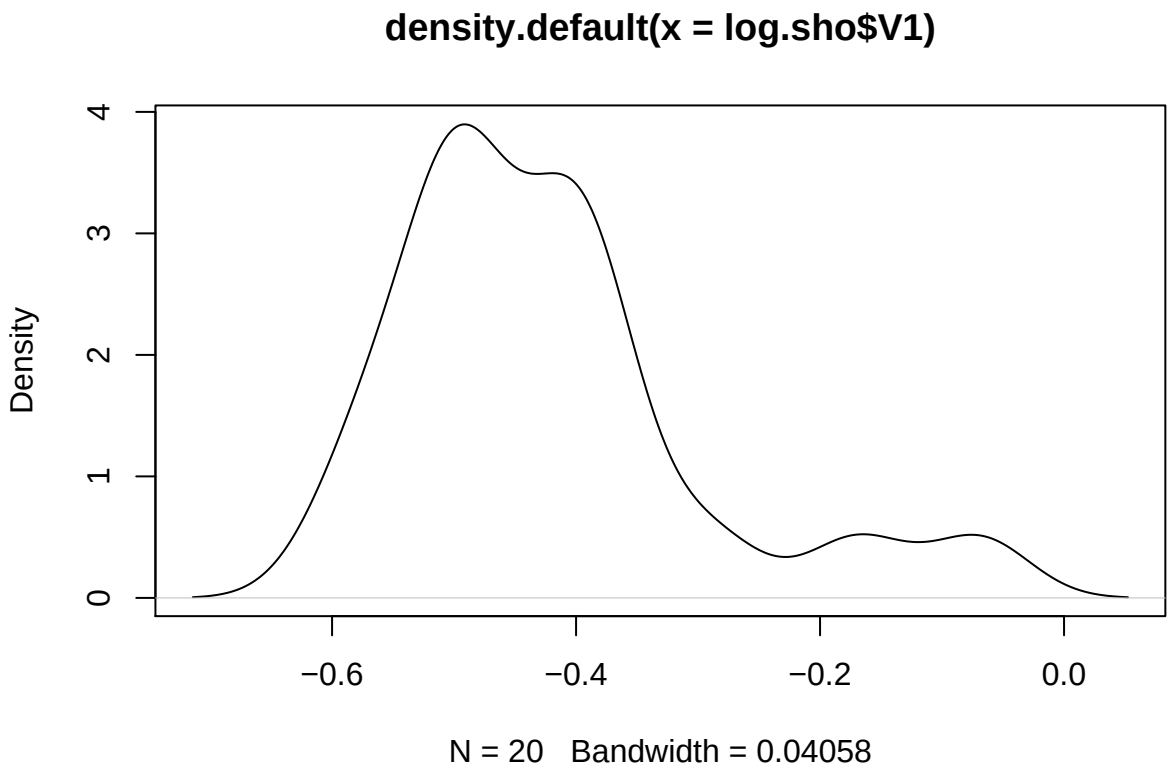

```
plot(shoshoni)
```




```
log.sho <- log(shoshoni)
plot(density(shoshoni$V1))
```




```
plot(density(log.sho$V1))
```



b) Here, we will use the natural log to develop and test our hypothesis. The scientist likely wants to show that the Shoshoni used the golden ratio, so we will state that as our alternative hypothesis: $H_0 : \mu_0 \neq -.41$ and $H_1 : \mu_1 = -.41$ A two-tailed hypothesis is suitable for our t-test. To begin, we'll calculate a t-statistic as follows:

```
x_bar <- mean(log.sho$V1)
n <- length(log.sho$V1)
s <- sd(log.sho$V1)

sho.t.stat <- (x_bar / (s/sqrt(n)))
sho.t.stat
```

```
## [1] -14.69797
```

```
2 * pt(sho.t.stat, df=n-1)
```

```
## [1] 7.867447e-12
```

Yikes - I got an extremely small p-value of 7.32e-12, which is obviously smaller than our significance value of .05. (Not sure I did that right). Again, the small number of observations could produce skewed results, unless that's the entirety of the population, in which case we can say that there is compelling evidence to reject the null in favor of the alternative.