Aashwin Vats

Homework#3

1. **Practice log-transforming data and reporting result of an analysis on log-transformed data.**

a. *Read in the data set [see Lab 3, item 5 for an example of reading a .csv file into R]. Examine the structure of the data frame using head(). Turn in your R code and output.*

R Code:

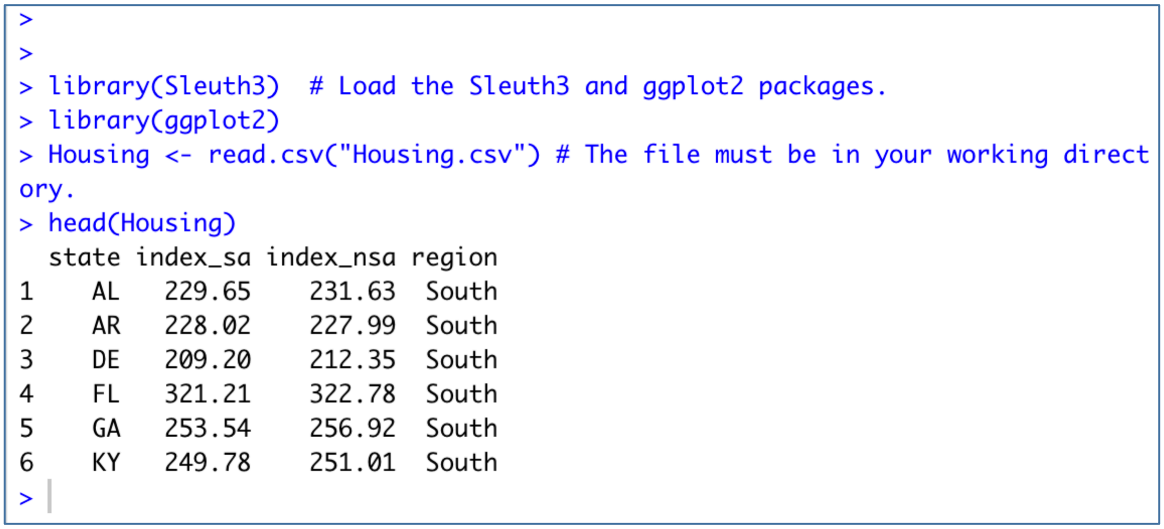
After setting the working directory, reading the data of .csv file into R:

```
Housing <- read.csv("Housing.csv")
```

To examine the structure of the data frame using head:

```
head(Housing)
```

Output:



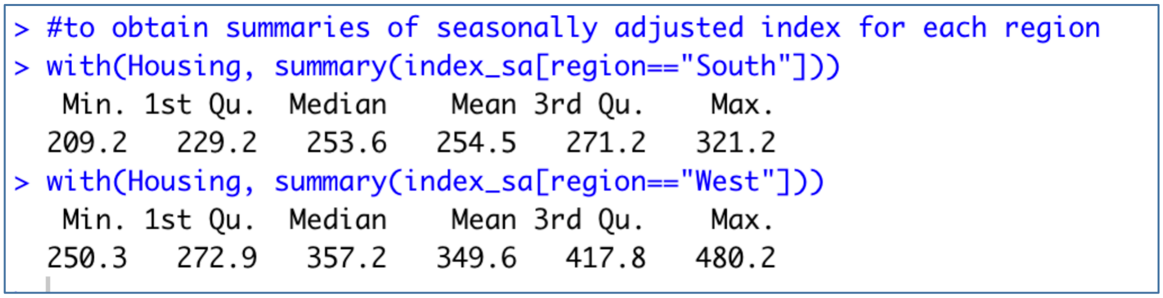
```
>
>
> library(Sleuth3)  # Load the Sleuth3 and ggplot2 packages.
> library(ggplot2)
> Housing <- read.csv("Housing.csv") # The file must be in your working direct
ory.
> head(Housing)
  state index_sa index_nsa region
1    AL   229.65    231.63  South
2    AR   228.02    227.99  South
3    DE   209.20    212.35  South
4    FL   321.21    322.78  South
5    GA   253.54    256.92  South
6    KY   249.78    251.01  South
> 
```

b. *Obtain "summaries" of the seasonally adjusted index for each region.*

R Code:

```
with(Housing, summary(index_sa[region=="South"]))
with(Housing, summary(index_sa[region=="West"]))
```

Output:



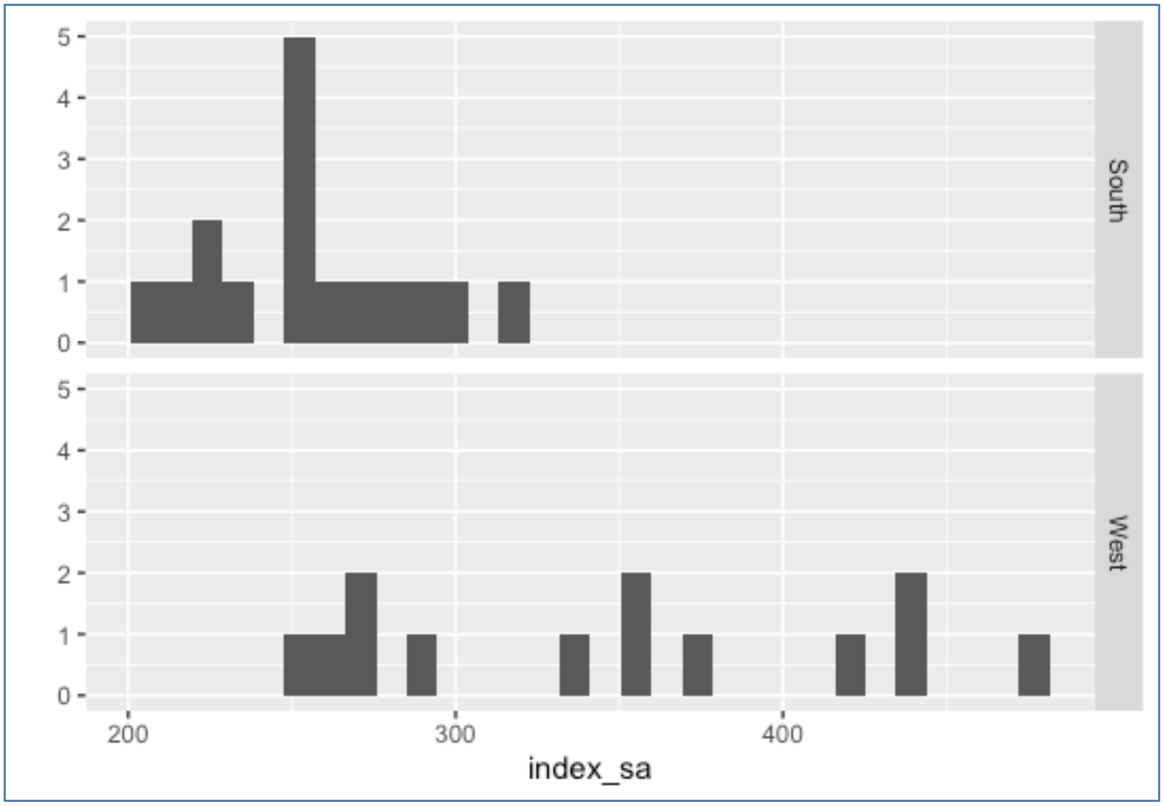
```
> #to obtain summaries of seasonally adjusted index for each region
> with(Housing, summary(index_sa[region=="South"]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  209.2   229.2   253.6   254.5   271.2   321.2 
> with(Housing, summary(index_sa[region=="West"]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  250.3   272.9   357.2   349.6   417.8   480.2 
```

c. *Use the example code on page 8 of Outline 3 to produce histograms for the two regions. Turn in your R code and graph.*

R Code:

qplot(index_sa, data =Housing, geom="histogram") + facet_grid(region ~ .)

Output:



d. *Log-transform the seasonally-adjusted index. Obtain summaries as in (b) and histograms as in (c) using the logged data. Turn in your R code, output, and graph*.

R code:

To log transform seasonally-adjusted index:

```
Housing$log.index_sa <- log(Housing$index_sa)
```

To obtain summaries:

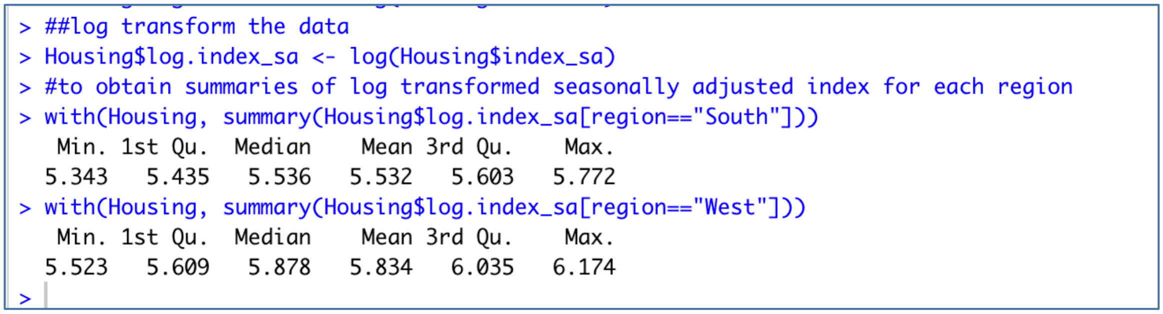
```
with(Housing, summary(Housing$log.index_sa[region=="South"]))
with(Housing, summary(Housing$log.index_sa[region=="West"]))
```

To obtain histograms:

```
qplot(Housing$log.index_sa, data =Housing, geom="histogram" ) +
facet_grid(region ~ .)
```

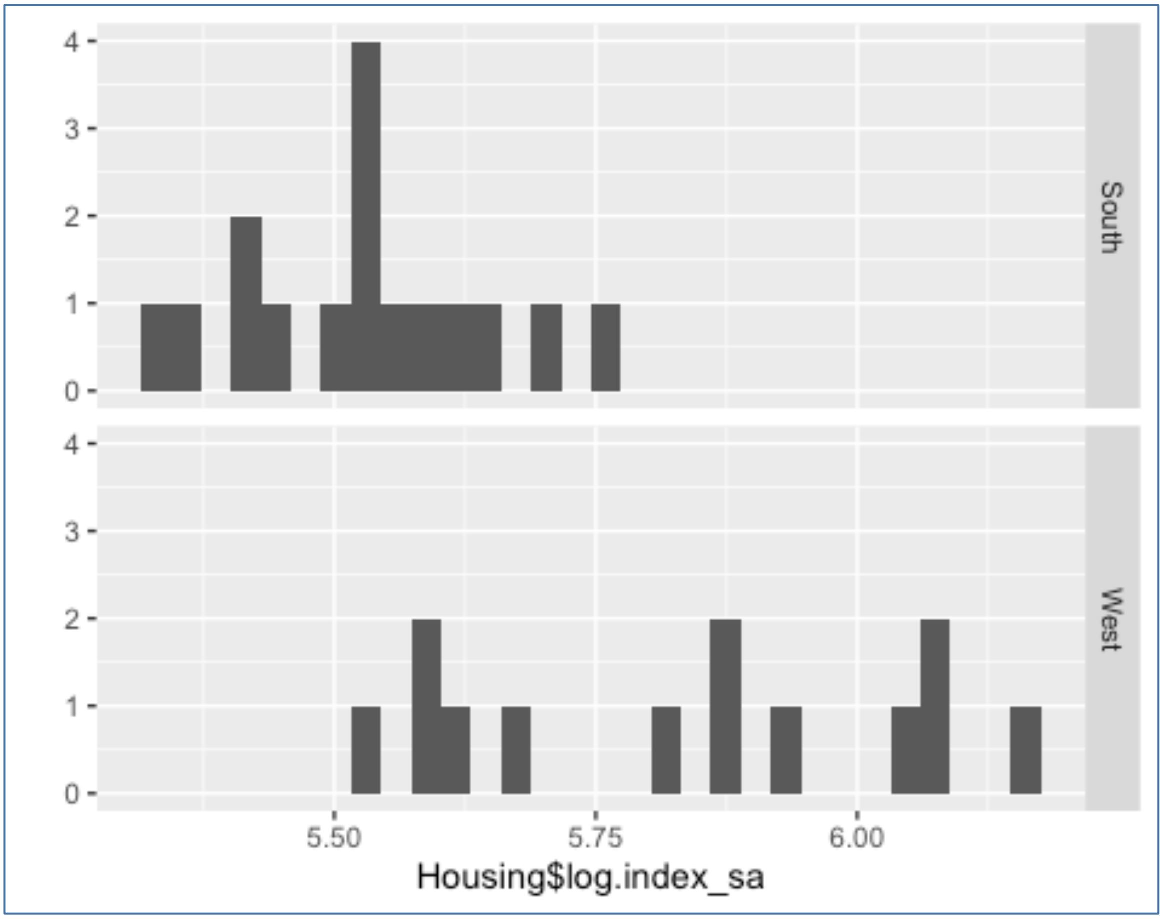
Output:

Summaries:



```
> ##log transform the data
> Housing$log.index_sa <- log(Housing$index_sa)
> #to obtain summaries of log transformed seasonally adjusted index for each region
> with(Housing, summary(Housing$log.index_sa[region=="South"]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  5.343   5.435   5.536   5.532   5.603   5.772 
> with(Housing, summary(Housing$log.index_sa[region=="West"]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  5.523   5.609   5.878   5.834   6.035   6.174 
> 
```

Histogram:



e. *State the three assumptions needed to use the t-tools. For each assumption, state your opinion whether it is reasonable for the untransformed data, then for the transformed data. Give a reason for your opinion (for example, what do the plots or summaries above suggest?). You may be very brief here. One or two sentences per assumption will be enough.*

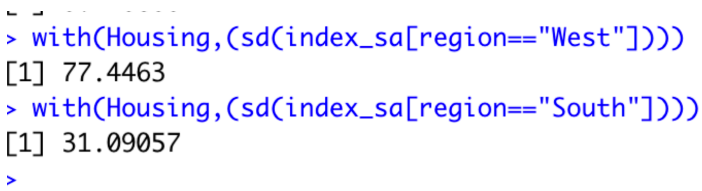
Three assumptions needed for t-test and t confidence interval:

1. **Assumption 1:** The population is normally distributed for both regions.
    - For untransformed data, we can observe from the histogram that the population for South seems to be normally distributed. However, not much can be said about the population of West looking at the histogram. If we look at the summaries obtained for untransformed data, we can see that mean ≈ median, that indicates a symmetric distribution.
    - For transformed data, we observe that the population for South seems to be normally distributed. Again, we can't say

about the behavior of West. If we look at the summaries obtained for transformed data, we can see that mean ≈ median, that indicates a symmetric distribution.

2. **Assumption 2:** The two populations have same variance or standard deviation.

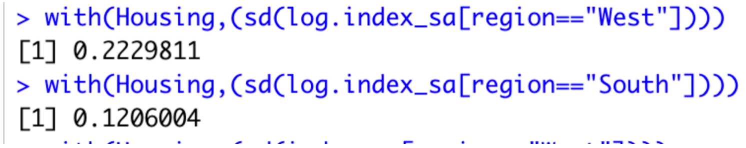
   - **Getting SD for untransformed data:**



```
> with(Housing,(sd(index_sa[region=="West"])))
[1] 77.4463
> with(Housing,(sd(index_sa[region=="South"])))
[1] 31.09057
>
```

There is substantial difference between the standard deviation of untransformed data.

   - **Getting SD for transformed data:**



```
> with(Housing,(sd(log.index_sa[region=="West"])))
[1] 0.2229811
> with(Housing,(sd(log.index_sa[region=="South"])))
[1] 0.1206004
```

There is a big improvement in the difference of SD with transformed data.

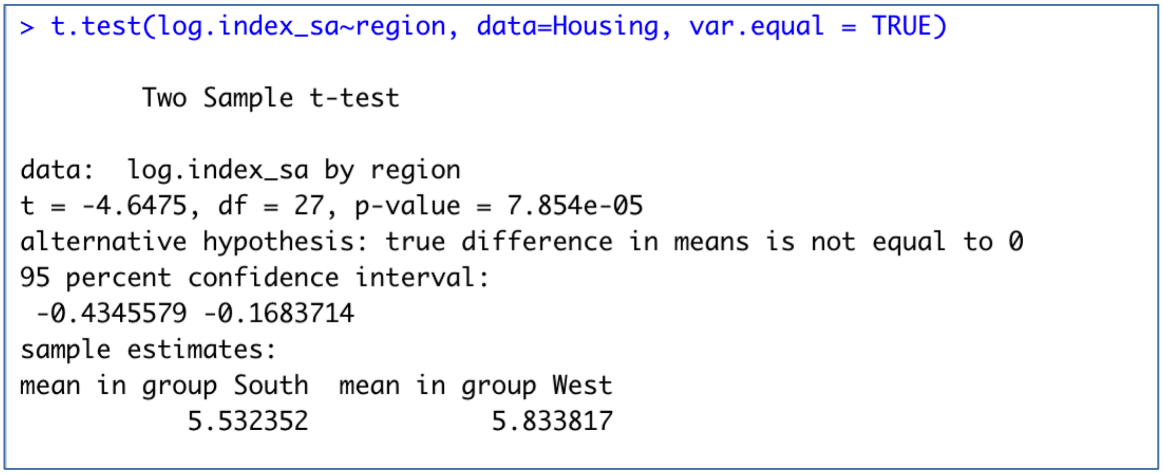
3. **Assumption 3:** Observations are independent.
   Observations are independent if you have a random sample. It can be observed from transformed & untransformed data that members of the grouping variable 'Region' are not assigned randomly (from the dataset).

f. *The t-tools are robust to many departures from the assumptions, so even if you don't believe the assumptions are met, perform a two-sample t-test using R on the logged data* ***to test the null hypothesis*** *that population mean log seasonally-adjusted housing index for the two regions is the same vs. the alternative that they are not the same. Submit your R code and output.*

R Code:

```
t.test(log.index_sa~region, data=Housing, var.equal = TRUE)
```

Output:



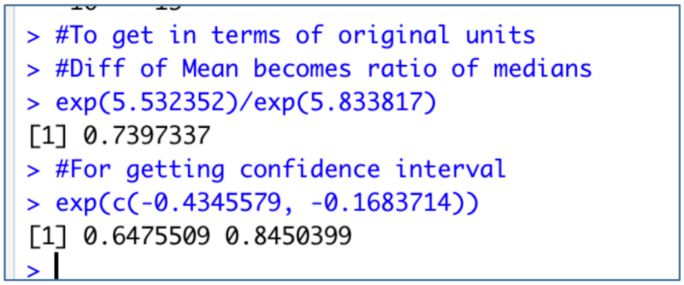
```
> t.test(log.index_sa~region, data=Housing, var.equal = TRUE)

	Two Sample t-test

data:  log.index_sa by region
t = -4.6475, df = 27, p-value = 7.854e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4345579 -0.1683714
sample estimates:
mean in group South  mean in group West 
           5.532352            5.833817 
```

g. *Give a "statistical conclusion" reporting the results of your test in part (f). A statistical conclusion should be in terms of the original units, not log units.*

To get in terms of original units:



```
> #To get in terms of original units
> #Diff of Mean becomes ratio of medians
> exp(5.532352)/exp(5.833817)
[1] 0.7397337
> #For getting confidence interval
> exp(c(-0.4345579, -0.1683714))
[1] 0.6475509 0.8450399
> 
```

There is strong evidence that the ratio of population median of seasonally adjusted housing index for the two regions is not 1. (two-sided t-test p-value= 0.0000784). We estimate that the population median for seasonally adjusted housing index for South is 0.739 times the population median for seasonally adjusted housing index for West.

h. *Obtain a two-sided 95% confidence interval for the difference in population means, using the logged data, then back-transform the endpoints of the interval.*

95% CI for the difference in population means, using the logged data:

-0.435 to -0.168

Now, back-transforming the end points of the interval:

exp(c(-0.4345579, -0.1683714))

0.6475509 to 0.8450399

i. *Give a "statistical conclusion" reporting your back-transformed confidence interval from (h).*

There is strong evidence that the ratio population median seasonally adjusted housing index for the two regions is not 1. (two-sided t-test p-value= 0.0000784). We estimate that the population median for seasonally adjusted housing index for South is 0.739 times population median for seasonally adjusted housing index for West. (95% CI 0.6475509 to 0.8450399 for ratio of population medians).