Basic Course on R: Hypothesis Testing and Confidence Intervals 1 Practical Answers

Elizabeth Ribble*

18-24 May 2017

Contents

*emcclel3@msudenver.edu

1 Baby Data

1. Read in the data "R_data_January2015.csv" with a header and row names from the first column. Assign it to the object `babydata` and allow strings be converted to factors. Attach the data to the environment.

```
babydata <- read.csv("R_data_January2015.csv", header = T,row.names = 1)
attach(babydata)
```

2. What are the dimensions of `babydata`? What is the class? Answer these questions separately with two functions and then together with one function.

```
dim(babydata)

## [1] 190  19

class(babydata)

## [1] "data.frame"

str(babydata)

## 'data.frame': 190 obs. of  19 variables:
##  $ Status                : Factor w/ 2 levels "intellectual disability",..: 1 2 1
##  $ iodine_deficiency     : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 2 2 2 2 ...
##  $ BMI                   : int  32 23 29 22 22 24 24 28 33 32 ...
##  $ educational_level     : Factor w/ 3 levels "high","intermediate",..: 2 2 3 3 3
##  $ alcohol               : Factor w/ 2 levels "no","yes": 1 2 1 1 2 2 2 2 2 1 ...
##  $ smoking               : Factor w/ 2 levels "no","yes": 1 2 1 1 1 1 1 1 2 1 ...
##  $ medication            : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ birthweight           : int  2618 3541 2619 3810 4136 4030 3377 2500 4255 2952
##  $ pregnancy_length_weeks: int  38 40 38 40 42 41 40 37 42 39 ...
##  $ pregnancy_length_days : int  4 2 3 5 3 1 4 1 0 2 ...
##  $ SAM                   : num  54.5 84 61 43 83 69 79 71.5 56 42.5 ...
##  $ SAH                   : num  14.8 23.6 18.7 23.2 17.1 19.6 22.4 18 20 23.4 ...
##  $ homocysteine          : num  18.8 15.6 15.2 16.5 19.5 17.5 14.9 22.2 19.1 16 .
##  $ cholesterol           : num  16.5 17.5 16.4 16.4 16.9 15.9 16.9 16 18.6 16.7 .
##  $ HDL                   : num  26.1 26.7 26.2 25.9 26.7 ...
##  $ triglycerides         : num  8.84 7.78 7.54 8.95 7.57 7.35 7.63 7.38 8.25 8.27
##  $ vitaminB12            : int  303 370 533 346 389 611 604 518 288 520 ...
##  $ folicacid_serum       : num  26.4 37.8 33.7 35.1 29 28.3 33.8 31.1 27.7 33.4 .
##  $ folicacid_erys        : num  1132 1467 1528 1539 1178 ...
```

3. Answer the following questions pertaining to the variable `SAH`:

   (a) What are the 20% quantiles of `SAH`?

```
quantile(SAH,seq(0,1,.2))
```

```
##    0%   20%   40%   60%   80%  100%
##  9.40 14.80 16.46 18.00 19.82 28.50
```

   (b) What are the mean, median, variance and standard deviation of `SAH`?

```
summary(SAH)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.40   15.22   17.30   17.59   19.08   28.50
```

```
## or
mean(SAH)
```

```
## [1] 17.59053
```

```
median(SAH)
```

```
## [1] 17.3
```

```
vs <- var(SAH)
vs
```

```
## [1] 11.26901
```

```
sqrt(vs)
```

```
## [1] 3.356935
```

   (c) Create a stem and leaf plot of `SAH`.

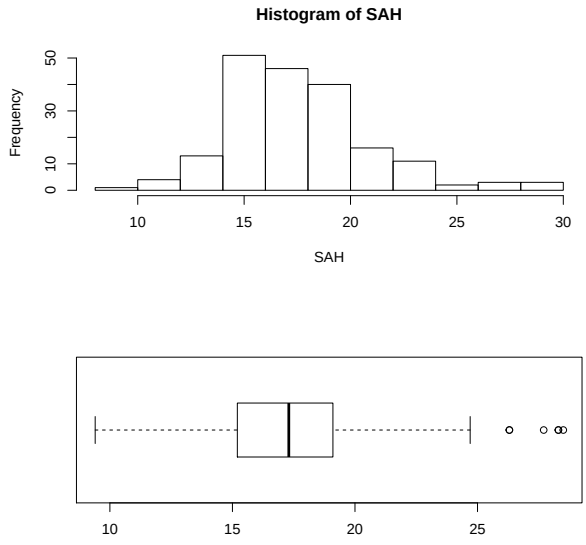
```
stem(SAH)
```

```
##
##   The decimal point is at the |
##
##    9 | 4
##   10 | 7
##   11 | 355
##   12 | 135
##   13 | 146677999
##   14 | 011112233345566677788899
```

```
##    15 | 0000123333455777889999
##    16 | 000001123334557788999
##    17 | 00111122333345677788889
##    18 | 000001112222334444555666789
##    19 | 000011222368889
##    20 | 0022245567899
##    21 | 2669
##    22 | 014
##    23 | 011124567
##    24 | 67
##    25 |
##    26 | 33
##    27 | 7
##    28 | 335
```

(d) Create a histogram and a horizontal boxplot of `SAH` in one graphics window where the plot of the histogram is above the boxplot.

```
par(mfrow=c(2,1))
hist(SAH)
boxplot(SAH,horizontal=TRUE)
```



(e) Utilize all 3 graphs to describe the shape of the distribution of `SAH`.

The distribution is unimodal and slightly skewed to the right.

(f) Log-transform SAH (assign it to `logSAH`).

```
logSAH <- log(SAH)
```

(g) What are the 20% quantiles of `logSAH`?

```
quantile(logSAH,seq(0,1,.2))
```

```
##       0%      20%      40%      60%      80%     100%
## 2.240710 2.694627 2.800929 2.890372 2.986689 3.349904
```

(h) What are the mean, median, variance and standard deviation of `logSAH`?

```
summary(logSAH)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.241   2.723   2.851   2.850   2.948   3.350
```

```
## or
mean(logSAH)
```

```
## [1] 2.849985
```

```
median(logSAH)
```

```
## [1] 2.850707
```

```
vls <- var(logSAH)
vls
```

```
## [1] 0.03460558
```

```
sqrt(vls)
```

```
## [1] 0.1860257
```

(i) Create a stem and leaf plot of `logSAH`.

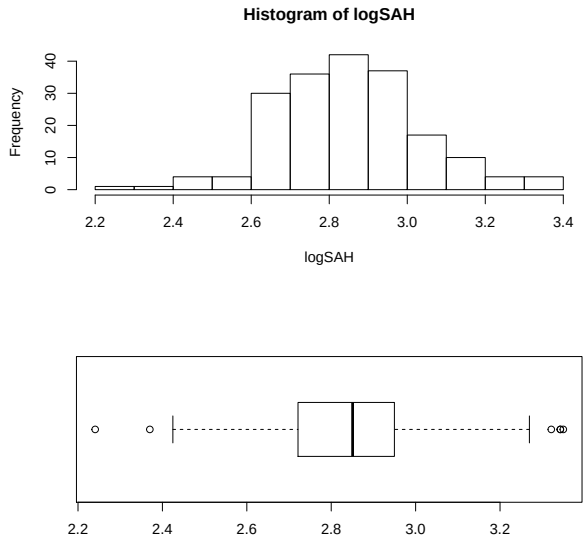
```
stem(logSAH)
```

```
##
##   The decimal point is 1 digit(s) to the left of the |
##
##   22 | 4
##   23 | 7
##   24 | 2449
```

```
##    25 | 137
##    26 | 01122333455555566677788889999999
##    27 | 00111112333334455566777777777889999
##    28 | 000022223333344444455556677777888899999
##    29 | 00000001111112222222334444455555689999
##    30 | 001112223334457799
##    31 | 01444445667
##    32 | 0177
##    33 | 2445
```

(j) Create a histogram and a horizontal boxplot of `logSAH` in one graphics window where the plot of the histogram above the boxplot.

```
par(mfrow=c(2,1))
hist(logSAH)
boxplot(logSAH,horizontal=TRUE)
```

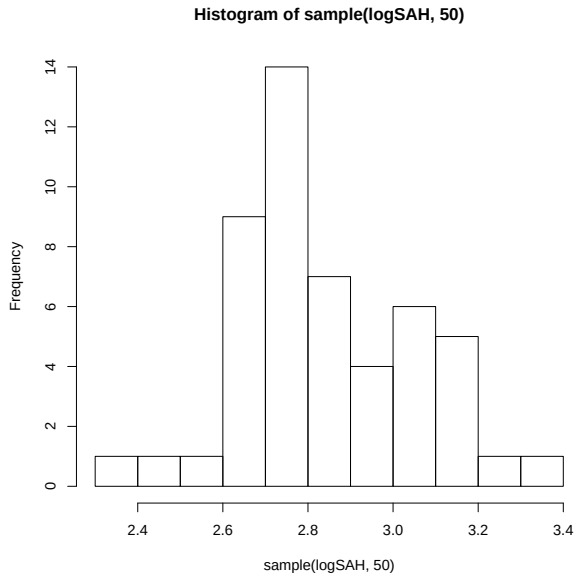


(k) Utilize all 3 graphs to describe the shape of the distribution of `logSAH`.
The distribution is unimodal, symmetric and appears normal.

(l) What did the log transformation do to the values of `SAH`?
Dampened the effect of skewness to make the distribution more normal and symmetric.

(m) Take a random sample of size 50 from `logSAH` and make a histogram. Does this distribution have a similar shape compared to that of all `logSAH` values?

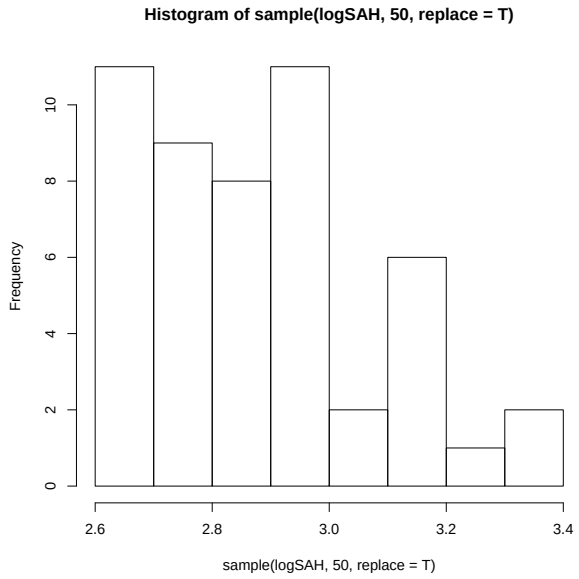
```
set.seed(1234)
hist(sample(logSAH, 50))
```



The distribution looks somewhat similar to that of the complete data.

(n) Take a random sample of size 50 with replacement from `logSAH` and make a histogram. Does this distribution have a similar shape compared to that of all `logSAH` values?

```
set.seed(1234)
hist(sample(logSAH, 50, replace=T))
```



The distribution looks quite different from the original distribution.

4. Answer the following questions pertaining to the variable `medication`:

   (a) Use a function to create frequency table of the number of mothers taking medication and not taking medication.

```
table(medication)

## medication
##  no yes
## 159  31
```

   (b) Calculate the percent of the mothers are taking medication; what is the percentage?

```
table(medication)/length(medication)

## medication
##        no       yes
## 0.8368421 0.1631579
```

   So 16.3% of mothers are taking medication.

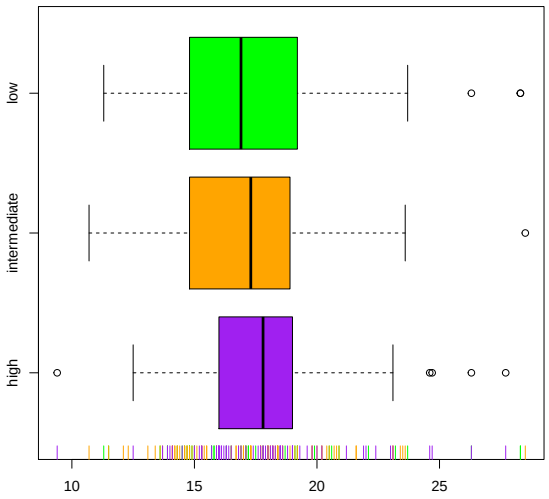
5. Answer the following questions pertaining to the variable `educational_level`:

   (a) Create a frequency table of the number of mothers in each education level.

```
table(educational_level)

## educational_level
##         high intermediate          low
##           63           82           45
```

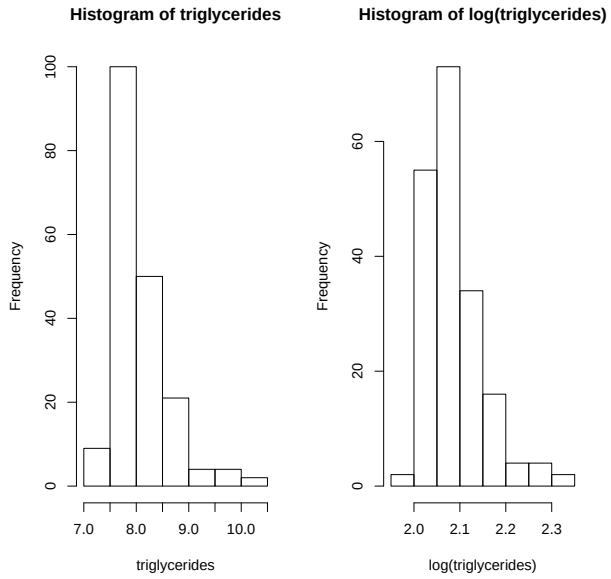
(b) Create a horizontal boxplot of the `SAH` values for the different levels of education and color each box a different color. Add a rug plot of the values where the ticks for each group have the same color as their corresponding box.

```
boxplot(SAH ~ educational_level, horizontal=TRUE,
        col=c("purple","orange","green"))
rug(SAH[educational_level=="low"], col="green")
rug(SAH[educational_level=="intermediate"], col="orange")
rug(SAH[educational_level=="high"], col="purple")
```



(c) Are `triglycerides` normally distributed (make a plot to answer this question)? If not, log-transform them. Are the log-transformed values normal?

```
par(mfrow=c(1,2))
hist(triglycerides)
hist(log(triglycerides))
```

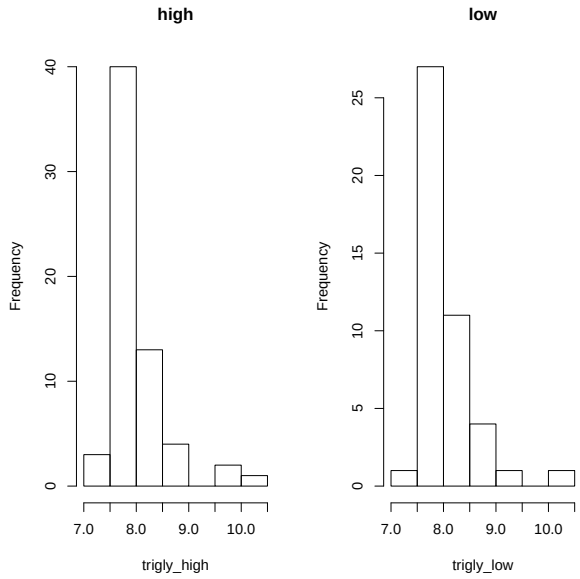


The values are skewed to the right and are still skewed after a log transformation.

(d) Is the average triglyceride level for highly educated mothers different from that of mothers with a low education level? Formulate a hypothesis, test it, and make a decision about whether or not you can reject the null hypothesis. Can you use a t-test (either on the raw or log-transformed data)? Why or why not (hint: how are the data distributed)?

We cannot use the t-test because the data are not normally distributed, even after a log transformation! However, since the shapes of the distributions are similar, as evidence by the below plot, we can perform a Wilcoxon Rank Sum test.

```
trigly_high <- triglycerides[educational_level=="high"]
trigly_low <- triglycerides[educational_level=="low"]
par(mfrow=c(1,2))
hist(trigly_high, main = "high")
hist(trigly_low, main = "low")
```



Our null hypothesis is that the difference in location between the low and high education levels is 0. The alternative hypothesis is that there is a difference.

```
wilcox.test(trigly_high, trigly_low)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  trigly_high and trigly_low
## W = 1328, p-value = 0.5791
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is not less than 0.05 so we fail to reject the null and conclude that we do not have enough evidence to show a statistically significant difference between the centers of the high and low mothers' triglycerides.

(e) Now re-do the test and make your decision to reject/not reject the null based on the confidence interval. Challenge: extract the confidence interval from the test output and use logical operators to answer the question of whether the interval contains the null value.

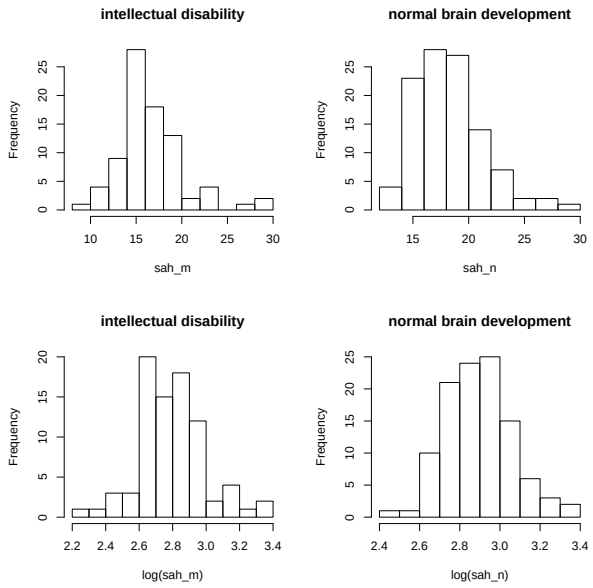
```
wtest <- wilcox.test(trigly_high, trigly_low, conf.int=TRUE)
ciw <- wtest$conf.int
0 > ciw[1] & 0 < ciw[2]

## [1] TRUE
```

Since the confidence interval contains 0, we fail to reject the null.

6. Answer the following questions pertaining to the variable `Status`:

   (a) Are the average `SAH` values for the two levels of `Status` (normal brain development or intellectual disability) different? Formulate a hypothesis, test it, and make a decision about whether or not you can reject the null hyopthesis. Can you use a t-test (either on the raw or log-transformed data)? Why or why not (hint: check distributions with plots)?
   We can use the t-test on the log-transformed data, since they are normally distributed, as shown in the plots below.

```
sah_m <- SAH[Status=="intellectual disability"]
sah_n <- SAH[Status=="normal brain development"]
par(mfrow=c(2,2))
hist(sah_m, main = "intellectual disability")
hist(sah_n, main = "normal brain development")
hist(log(sah_m), main = "intellectual disability")
hist(log(sah_n), main = "normal brain development")
```



   Our null hypothesis is that the difference in mean `SAH` between the two groups is 0. The alternative hypothesis is that there is a difference in means.

```
sstest <- t.test(log(SAH)~Status)
sstest

##
##  Welch Two Sample t-test
##
```

```
## data:  log(SAH) by Status
## t = -3.8522, df = 152.24, p-value = 0.0001721
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.15753466 -0.05072509
## sample estimates:
##  mean in group intellectual disability
##                                2.790795
## mean in group normal brain development
##                                2.894925
```

```
sstest$p.value
```

```
## [1] 0.0001720599
```

```
sstest$conf.int
```

```
## [1] -0.15753466 -0.05072509
## attr(,"conf.level")
## [1] 0.95
```

The p-value is less than 0.05 (and 0 is not in the CI) so we can reject the null hypothesis that the difference in means between the two groups is 0. We conclude that there is a statistically significant difference between the average log-SAH of mothers with intellectually disabled children and the average log-SAH of those women whose children have normal brain development.

(b) What is the fold change of log-SAH between the 2 groups? Calculate it two ways: use the output from the previous test and also use the data itself (function `mean` plus logical operators).

```
sstest$estimate[2]/sstest$estimate[1]
```

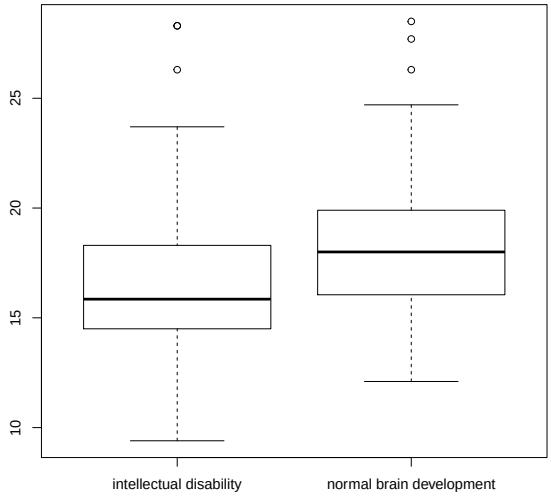
```
## mean in group normal brain development
##                                1.037312
```

```
mean(log(SAH)[Status=="normal brain development"])/
  mean(log(SAH)[Status=="intellectual disability"])
```

```
## [1] 1.037312
```

(c) Make a boxplot of the SAH values of the 2 groups and calculate the fold change of SAH between the 2 groups. Does the difference seem clinically relevant? Why or why not?

```
boxplot(SAH ~ Status)
```




```
mean(SAH[Status=="normal brain development"])/
  mean(SAH[Status=="intellectual disability"])

## [1] 1.101694
```

The fold change is really close to one and there is actually quite a bit of overlap in the boxplots. The statistically significant difference of the log-transformed data is quite subtle and may not be relevant, though further investigation of confounding factors may elucidate the true relationship between `SAH` and brain development.