Submitted by- Swayanshu shanti Pragnya

Exercise 1

This article is an introduction of adverse impact of methodology to P value fallacy, a mistaken ideation that a single element or number can capture long run outcomes for an experiment. Their suggestion needs higher understanding of reasoning types. The process to know the underlying reasons of observed world is Inferential reasoning which includes 2 logical types i.e. deductive inference and inductive inference.

Deductive inference is very similar with a statement or hypothesis for example, how a nature works, then seeing the result that if hypothesis were meant to be true. Its objective because what we will see are always true if the hypothesis is true.

Example- "In weekdays 1 o'clock is the traffic jam time", its true for some cases as for most of the companies it's the lunch break hence more traffic

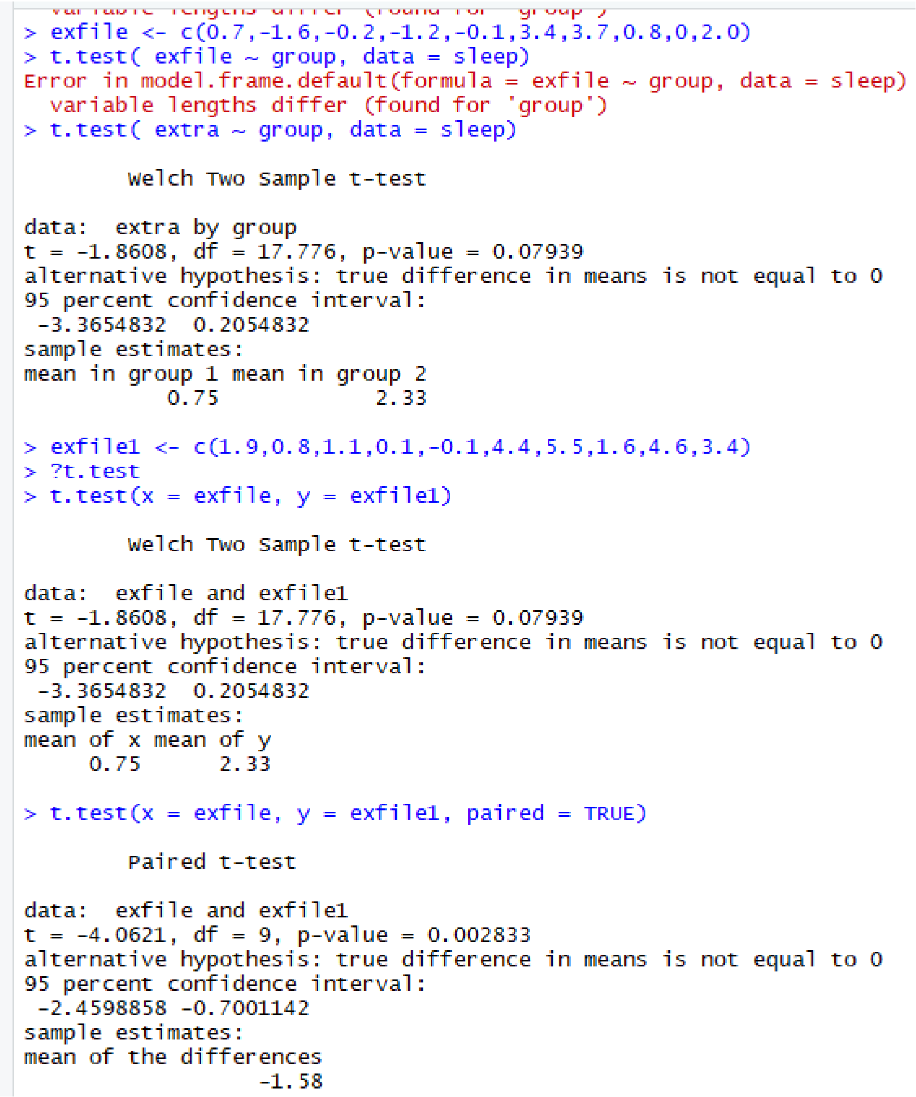
The limitation of this inference is we cant extend the knowledge beyond the hypothesis.

In other hand, Inductive inference is total opposite to deductive. What we see, we evaluate for which the hypothesis is defensible. It is a measure which reflects from observations to actual underlying truth. Benefit is it creates a broad generalization which allows to expand in all possible directions.

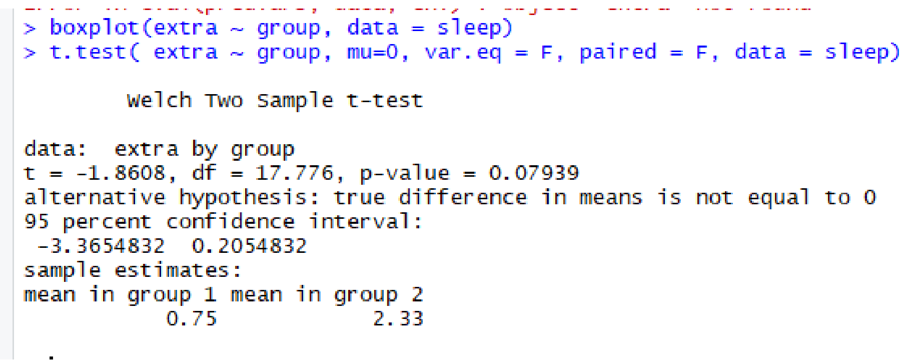
Ideation behind P value fallacy is, the event can be viewed concurrently both Long run and short-run perspective. Using both types of inference we can evaluate and conclude which is like solving an issue in different ways even if it's a contradiction. Fallacy can be treated as, A result cannot be substitutable member in long run view and an unique member in short run view.

In general if confidence interval does not contain any kind of null hypothesis values, then the results are statistically notable. But in paper they have mentioned if some p value is higher that means the discrepancy in between sample mean and null hypothesis is weighty. In similar way lower P value can be a significant result up to some percentage.

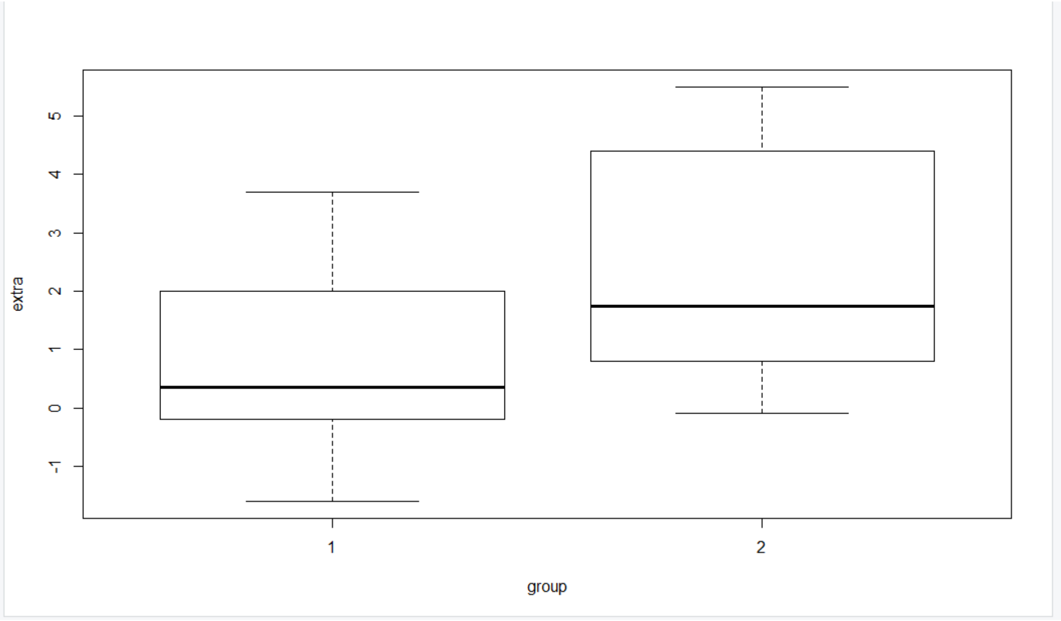
Exercise *2*: Revisiting the t-test



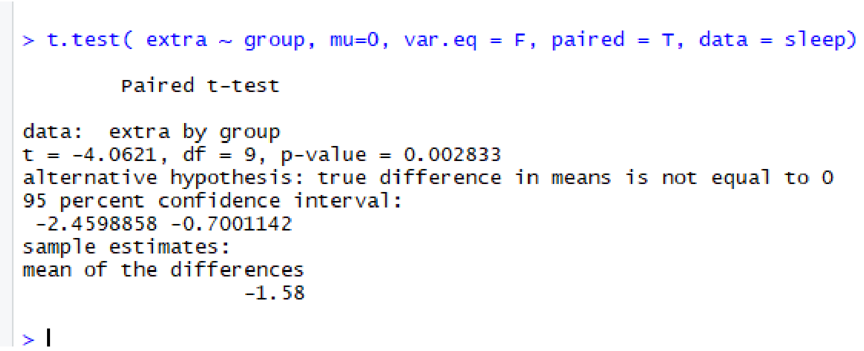
```
  variable lengths differ (found for 'group')
> exfile <- c(0.7,-1.6,-0.2,-1.2,-0.1,3.4,3.7,0.8,0,2.0)
> t.test( exfile ~ group, data = sleep)
Error in model.frame.default(formula = exfile ~ group, data = sleep)
  variable lengths differ (found for 'group')
> t.test( extra ~ group, data = sleep)

        Welch Two Sample t-test

data:  extra by group
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3654832  0.2054832
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33

> exfile1 <- c(1.9,0.8,1.1,0.1,-0.1,4.4,5.5,1.6,4.6,3.4)
> ?t.test
> t.test(x = exfile, y = exfile1)

        Welch Two Sample t-test

data:  exfile and exfile1
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3654832  0.2054832
sample estimates:
mean of x mean of y
     0.75      2.33

> t.test(x = exfile, y = exfile1, paired = TRUE)

        Paired t-test

data:  exfile and exfile1
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean of the differences
                  -1.58
```



```
> boxplot(extra ~ group, data = sleep)
> t.test( extra ~ group, mu=0, var.eq = F, paired = F, data = sleep)

        Welch Two Sample t-test

data:  extra by group
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3654832  0.2054832
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```



2b. Conduct a paired t-test that correctly assumes the data is paired. Write a brief, but complete, summary for your conclusions.



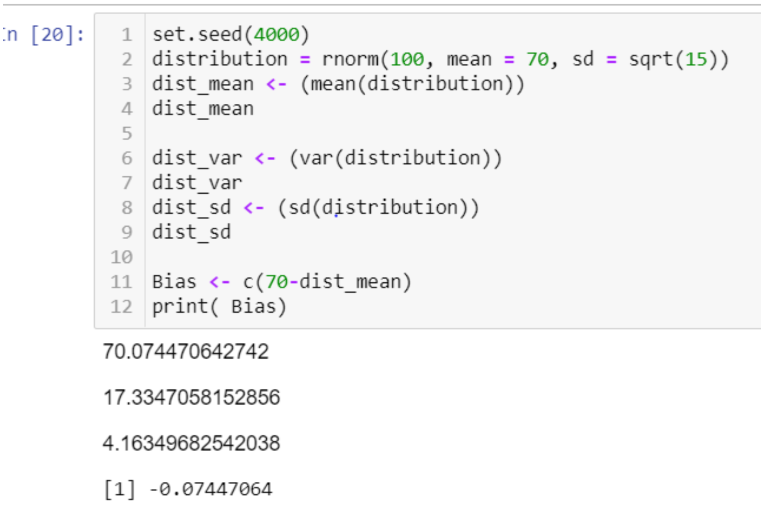
```
> t.test( extra ~ group, mu=0, var.eq = F, paired = T, data = sleep)

        Paired t-test

data:  extra by group
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean of the differences
                  -1.58

> 
```

2c. Briefly compare your conclusions for 2a and 2b.

By definition Two-sample *t*-test is used where the data of two samples are statistically independent. The paired *t*-test is used when data is in the form of matched pairs.
By seeing the changes in P-value two sample test has alternative differences compared to paired t test as there is a significance change in confidence interval.

Exercise 3: Properties of Estimators: Bias, Consistency, and Efficiency

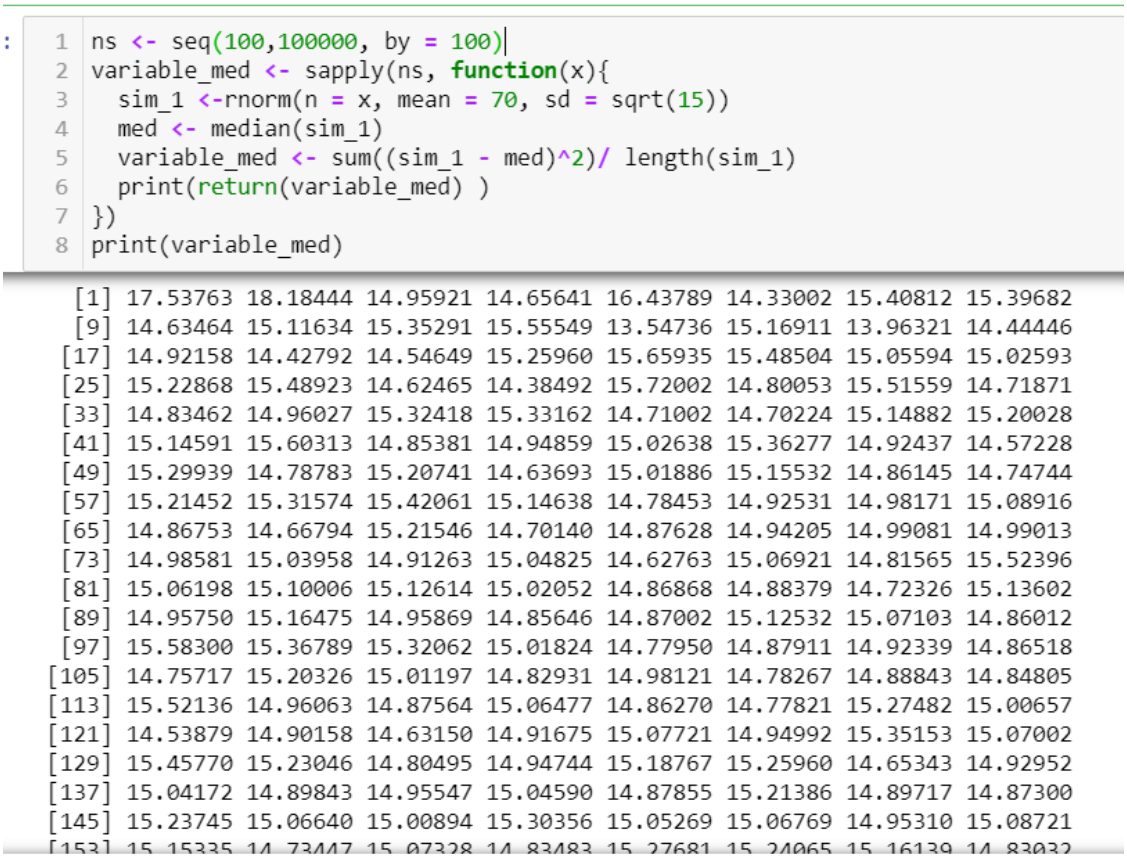
3.a



```
set.seed(4000)
distribution = rnorm(100, mean = 70, sd = sqrt(15))
dist_mean <- (mean(distribution))
dist_mean

dist_var <- (var(distribution))
dist_var
dist_sd <- (sd(distribution))
dist_sd

Bias <- c(70-dist_mean)
print( Bias)
```

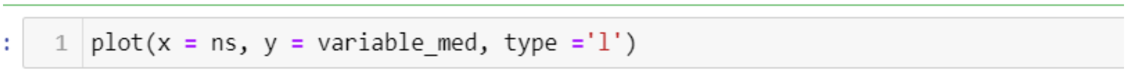
```
70.074470642742

17.3347058152856

4.16349682542038

[1] -0.07447064
```

3.b

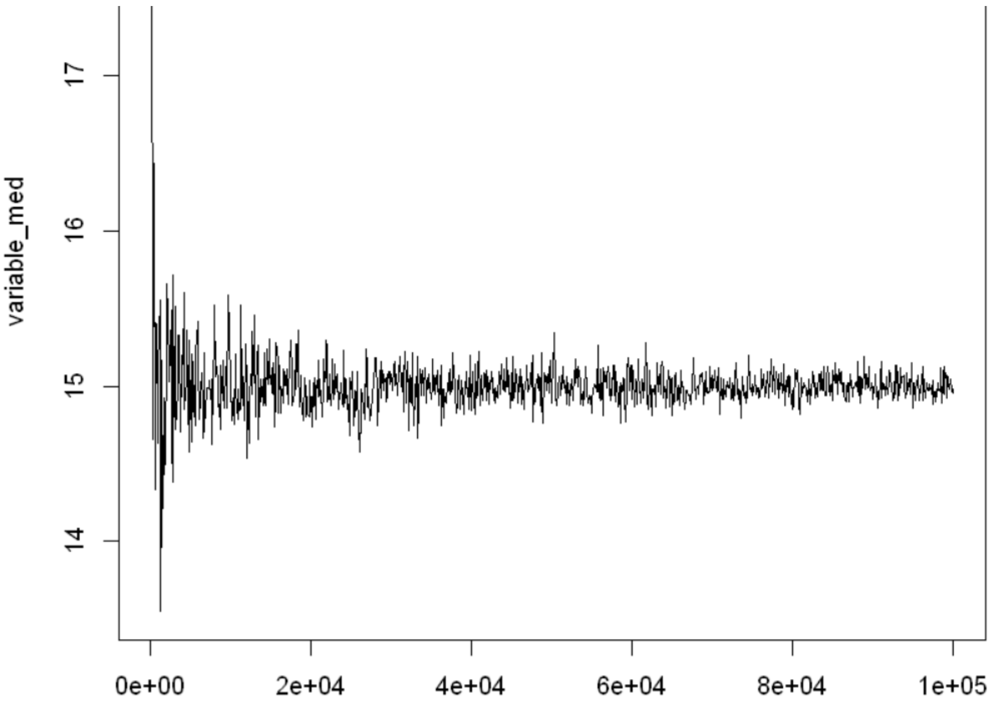


```
ns <- seq(100,100000, by = 100)
variable_med <- sapply(ns, function(x){
  sim_1 <-rnorm(n = x, mean = 70, sd = sqrt(15))
  med <- median(sim_1)
  variable_med <- sum((sim_1 - med)^2)/ length(sim_1)
  print(return(variable_med) )
})
print(variable_med)
```

```
  [1] 17.53763 18.18444 14.95921 14.65641 16.43789 14.33002 15.40812 15.39682
  [9] 14.63464 15.11634 15.35291 15.55549 13.54736 15.16911 13.96321 14.44446
 [17] 14.92158 14.42792 14.54649 15.25960 15.65935 15.48504 15.05594 15.02593
 [25] 15.22868 15.48923 14.62465 14.38492 15.72002 14.80053 15.51559 14.71871
 [33] 14.83462 14.96027 15.32418 15.33162 14.71002 14.70224 15.14882 15.20028
 [41] 15.14591 15.60313 14.85381 14.94859 15.02638 15.36277 14.92437 14.57228
 [49] 15.29939 14.78783 15.20741 14.63693 15.01886 15.15532 14.86145 14.74744
 [57] 15.21452 15.31574 15.42061 15.14638 14.78453 14.92531 14.98171 15.08916
 [65] 14.86753 14.66794 15.21546 14.70140 14.87628 14.94205 14.99081 14.99013
 [73] 14.98581 15.03958 14.91263 15.04825 14.62763 15.06921 14.81565 15.52396
 [81] 15.06198 15.10006 15.12614 15.02052 14.86868 14.88379 14.72326 15.13602
 [89] 14.95750 15.16475 14.95869 14.85646 14.87002 15.12532 15.07103 14.86012
 [97] 15.58300 15.36789 15.32062 15.01824 14.77950 14.87911 14.92339 14.86518
[105] 14.75717 15.20326 15.01197 14.82931 14.98121 14.78267 14.88843 14.84805
[113] 15.52136 14.96063 14.87564 15.06477 14.86270 14.77821 15.27482 15.00657
[121] 14.53879 14.90158 14.63150 14.91675 15.07721 14.94992 15.35153 15.07002
[129] 15.45770 15.23046 14.80495 14.94744 15.18767 15.25960 14.65343 14.92952
[137] 15.04172 14.89843 14.95547 15.04590 14.87855 15.21386 14.89717 14.87300
[145] 15.23745 15.06640 15.00894 15.30356 15.05269 15.06769 14.95310 15.08721
[153] 15.15335 14.73447 15.07328 14.83483 15.27681 15.24065 15.16139 14.83032
```



```
plot(x = ns, y = variable_med, type ='l')
```



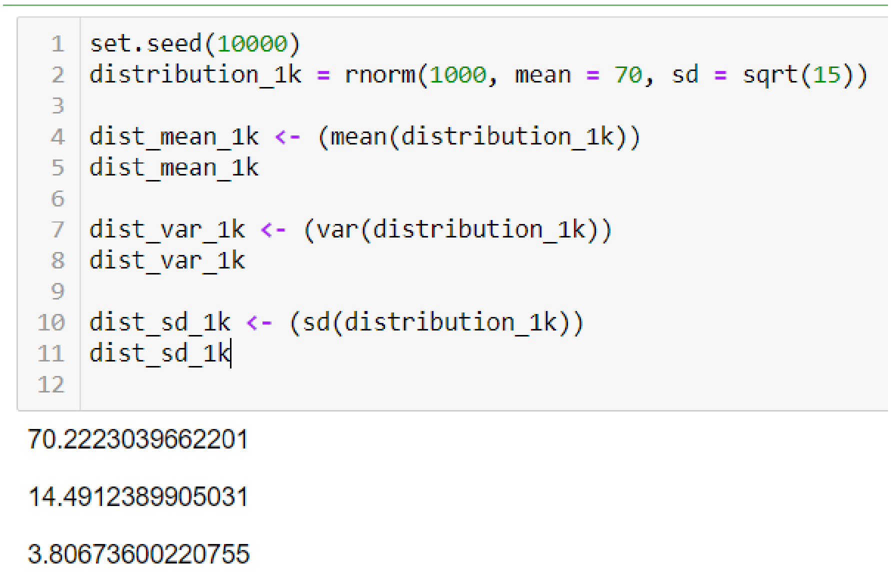
***3*c.**

How does the variance of the data wrt the median estimator change as the sample size increases?

I do not see any markable changes as variance is almost similar with some random variations even after changing the sample size

***3*d.**

Dr. Billy bets Dr. Bob that the sample mean is more efficient (i.e. less variable about the population mean) than the sample median. To compare the relative efficiency of estimators, simulate 10,000 normal distributions with sample size n=1000, population mean=70 inches, and variance=15 inches2. Calculate the median and mean for each simulation. Then compare the variance of the set of sample medians to the variance of the set of sample means. Using the results of your simulation, which estimator is more efficient?



```
set.seed(10000)
distribution_1k = rnorm(1000, mean = 70, sd = sqrt(15))

dist_mean_1k <- (mean(distribution_1k))
dist_mean_1k

dist_var_1k <- (var(distribution_1k))
dist_var_1k

dist_sd_1k <- (sd(distribution_1k))
dist_sd_1k

```

```
70.2223039662201

14.4912389905031

3.80673600220755
```

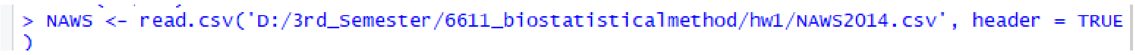
The result is slightly varying based on the sample size but as we know variance of the sample mean is inversely proportional to sample size so both estimators has less differentiation.
Calculating mean and median with each simulation is a better as well time consuming estimator so based on the result requirement we can choose the estimator.

3e. Extra Credit: What is the Cramer-Rao Lower Bound, and why does it relate to this exercise?

(CRLB)Cramer-Rao Lower Bound is an unbiased estimator. It gives lower estimation for variances of an unbiased estimator.
In the following exercise we worked on variance estimation so one if the method CRLB can be a better method to choose the estimator.

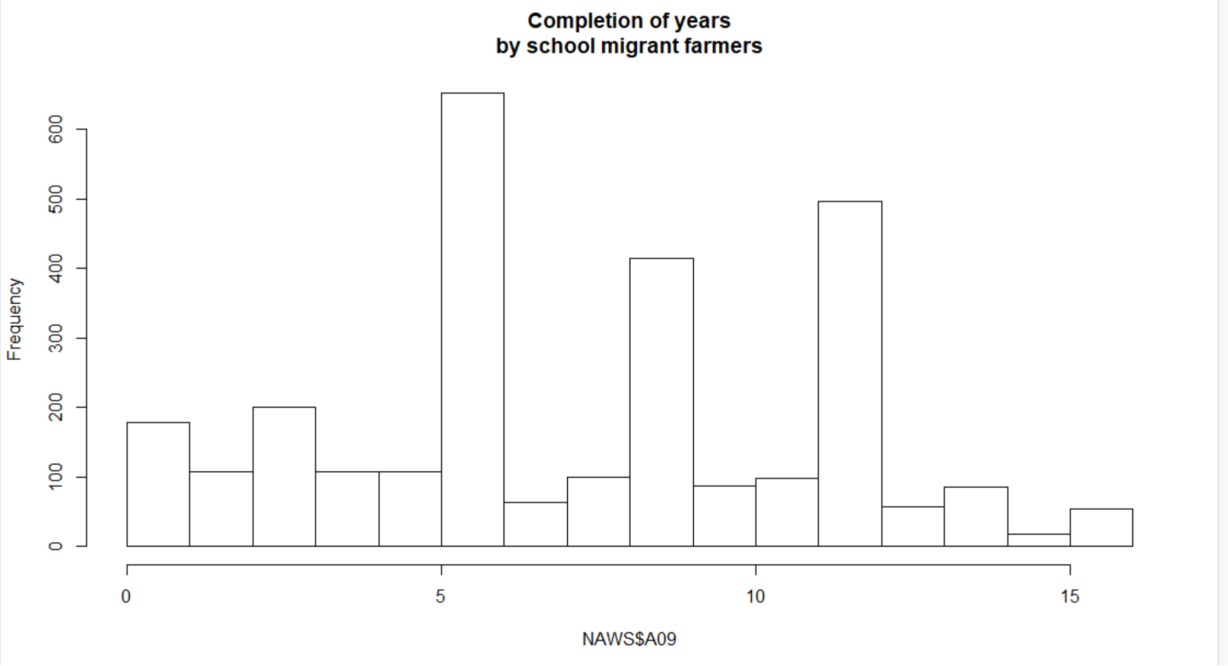
Exercise 4: NAWS Farm Worker Survey Simulation

4a



```
> NAWS <- read.csv('D:/3rd_semester/6611_biostatisticalmethod/hw1/NAWS2014.csv', header = TRUE
)
```

4b



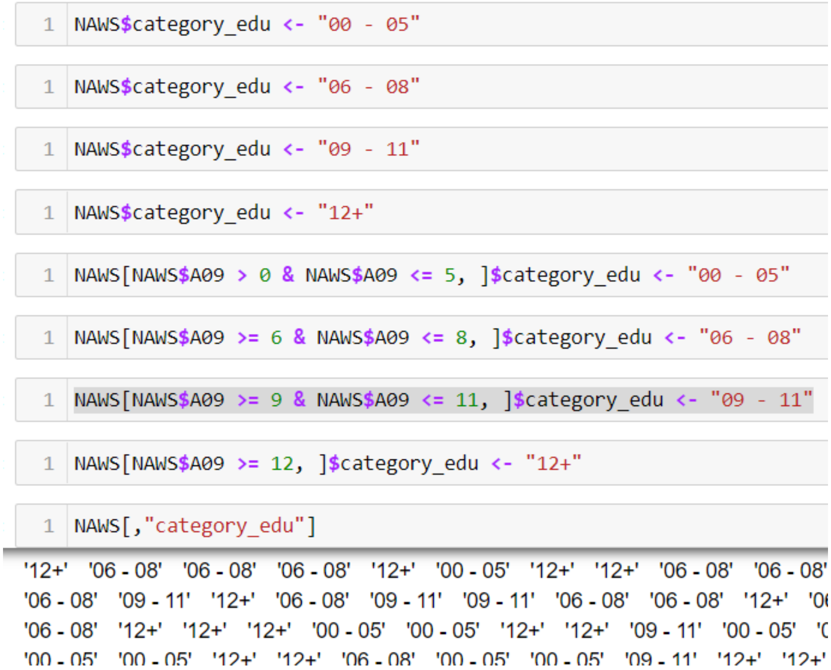
4c

In your opinion, does reporting just this average tell the whole story? Why or why not? (Feel free to speculate on why you think the histogram has its unique shape.)

In my opinion reporting the average is not enough as based on the dataset there can be other factors than just A09.

This information “ average educational attainment among migrant farmers is 8 years” can be just an over all statement but if we see the histogram there is a fluctuation based on frequency that clearly indicates that we can not simply rely on this.

4d.



```
NAWS$category_edu <- "00 - 05"
```

```
NAWS$category_edu <- "06 - 08"
```

```
NAWS$category_edu <- "09 - 11"
```

```
NAWS$category_edu <- "12+"
```

```
NAWS[NAWS$A09 > 0 & NAWS$A09 <= 5, ]$category_edu <- "00 - 05"
```

```
NAWS[NAWS$A09 >= 6 & NAWS$A09 <= 8, ]$category_edu <- "06 - 08"
```

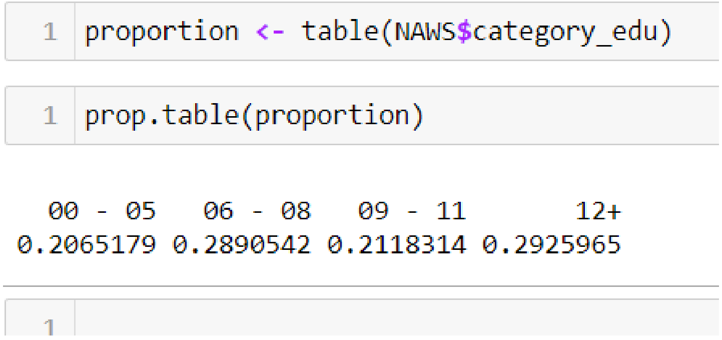
```
NAWS[NAWS$A09 >= 9 & NAWS$A09 <= 11, ]$category_edu <- "09 - 11"
```

```
NAWS[NAWS$A09 >= 12, ]$category_edu <- "12+"
```

```
NAWS[,"category_edu"]
```

```
'12+'   '06 - 08'   '06 - 08'   '06 - 08'   '12+'   '00 - 05'   '12+'   '12+'   '06 - 08'   '06 - 08'
'06 - 08'   '09 - 11'   '12+'   '06 - 08'   '09 - 11'   '09 - 11'   '06 - 08'   '06 - 08'   '12+'   '0
'06 - 08'   '12+'   '12+'   '12+'   '00 - 05'   '00 - 05'   '12+'   '12+'   '09 - 11'   '00 - 05'   '
'00 - 05'   '00 - 05'   '12+'   '12+'   '06 - 08'   '00 - 05'   '00 - 05'   '09 - 11'   '12+'   '12+'
```

4e

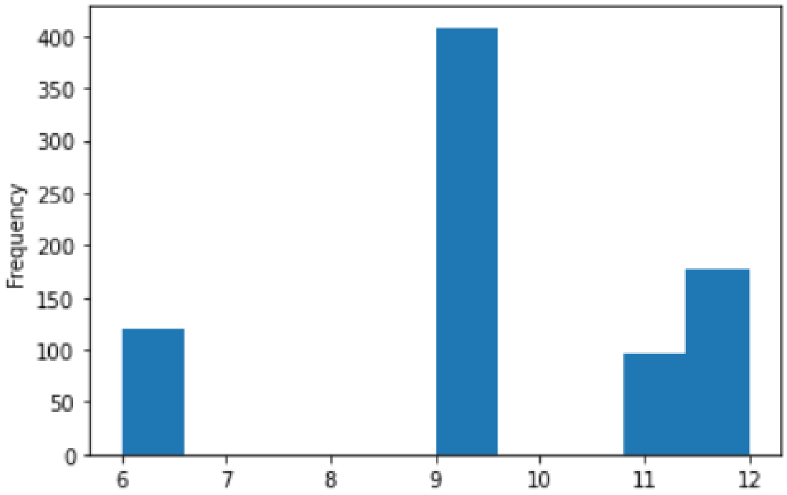


```
proportion <- table(NAWS$category_edu)
```

```
prop.table(proportion)
```

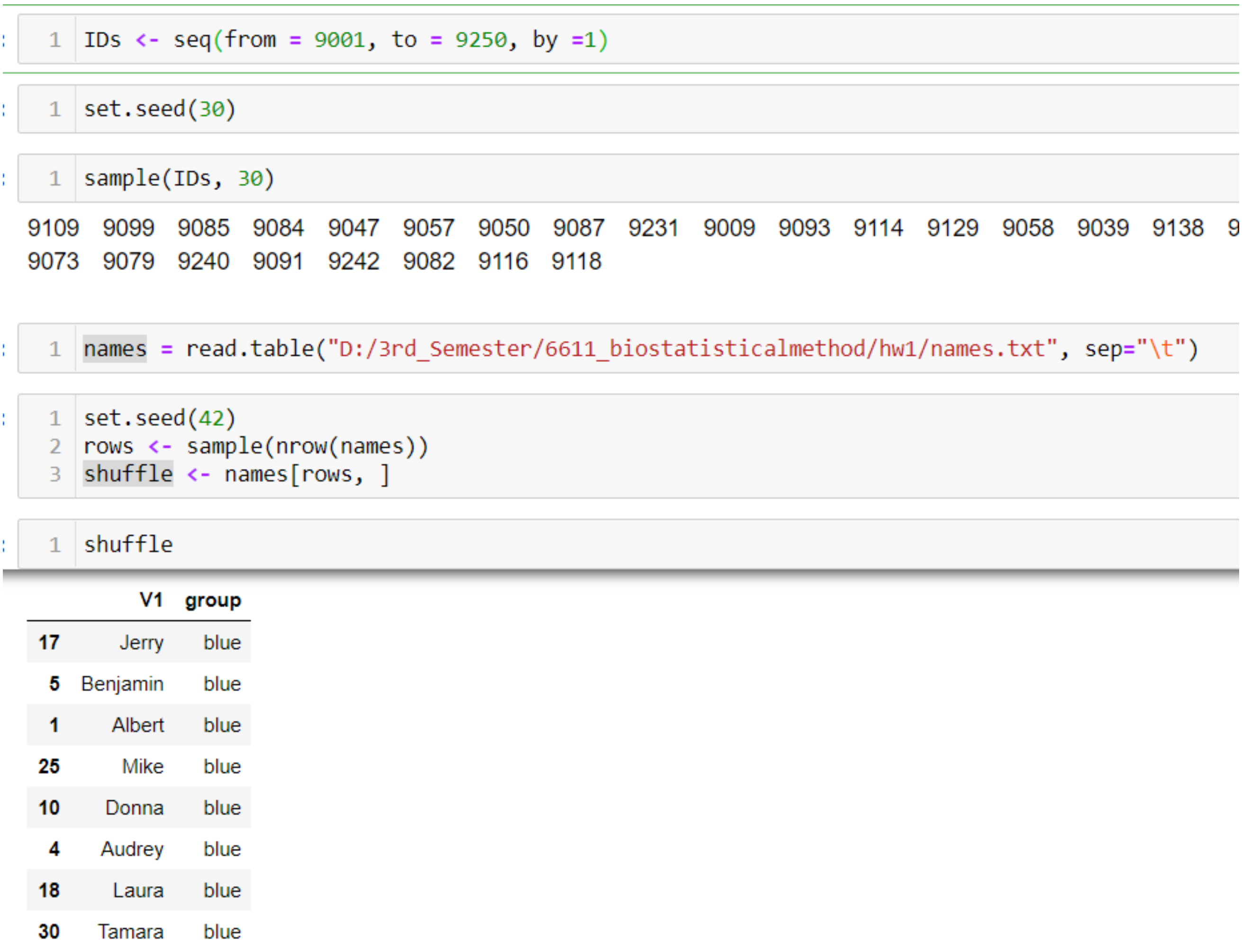
```
  00 - 05   06 - 08   09 - 11       12+ 
0.2065179 0.2890542 0.2118314 0.2925965 
```

4f.



Exercise 5

5a.



```
IDs <- seq(from = 9001, to = 9250, by =1)
```

```
set.seed(30)
```

```
sample(IDs, 30)
```

```
9109  9099  9085  9084  9047  9057  9050  9087  9231  9009  9093  9114  9129  9058  9039  9138  9
9073  9079  9240  9091  9242  9082  9116  9118
```

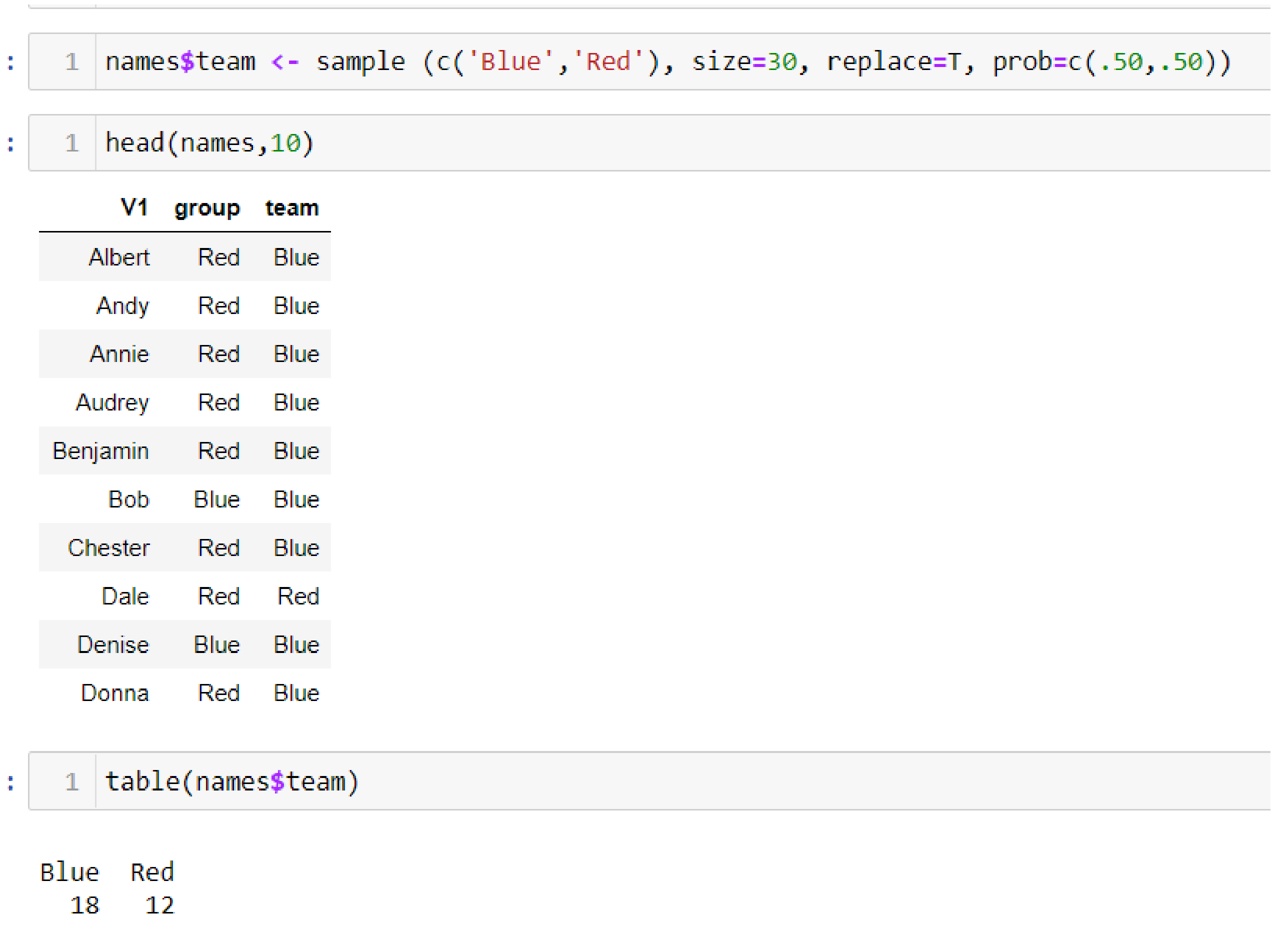
```
names = read.table("D:/3rd_Semester/6611_biostatisticalmethod/hw1/names.txt", sep="\t")
```

```
set.seed(42)
rows <- sample(nrow(names))
shuffle <- names[rows, ]
```

```
shuffle
```

	V1	group
17	Jerry	blue
5	Benjamin	blue
1	Albert	blue
25	Mike	blue
10	Donna	blue
4	Audrey	blue
18	Laura	blue
30	Tamara	blue

5b.



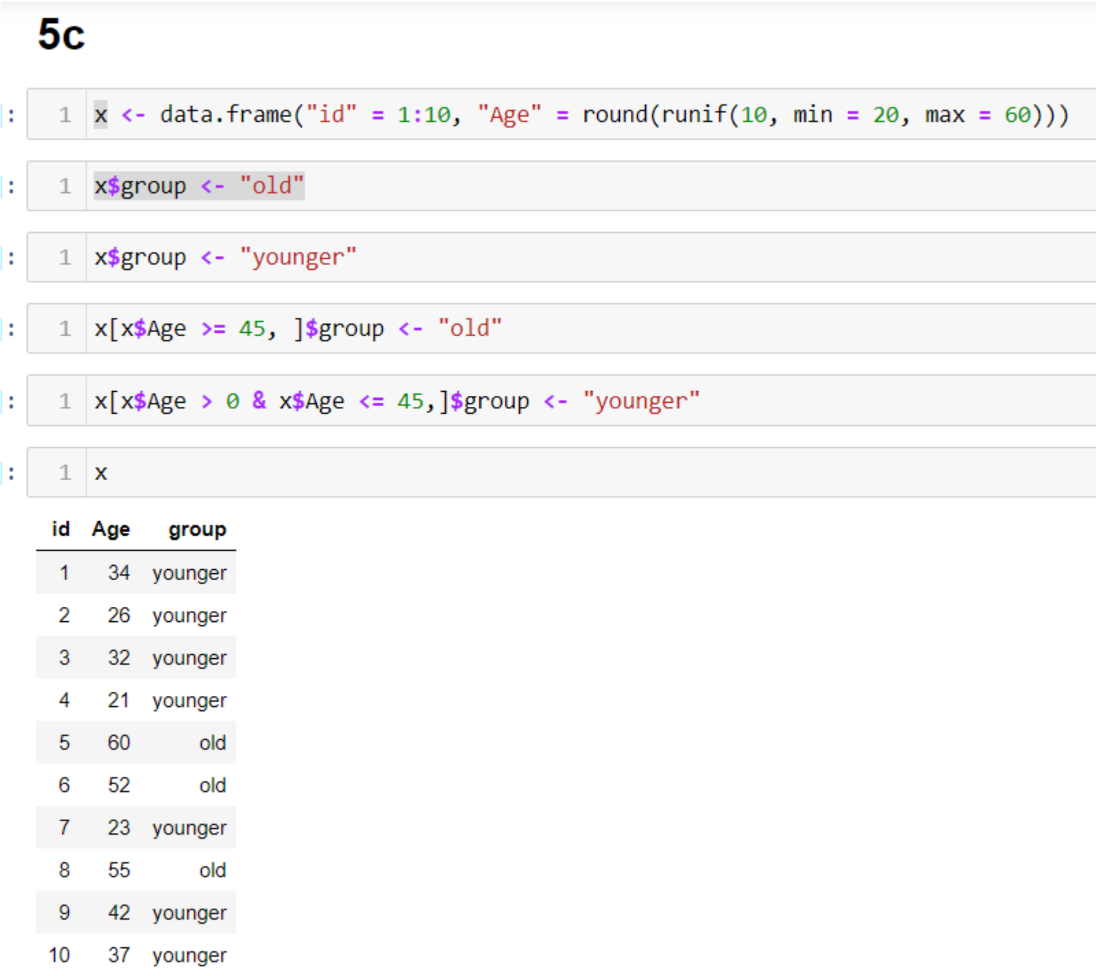
```
names$team <- sample (c('Blue','Red'), size=30, replace=T, prob=c(.50,.50))
```

```
head(names,10)
```

V1	group	team
Albert	Red	Blue
Andy	Red	Blue
Annie	Red	Blue
Audrey	Red	Blue
Benjamin	Red	Blue
Bob	Blue	Blue
Chester	Red	Blue
Dale	Red	Red
Denise	Blue	Blue
Donna	Red	Blue

```
table(names$team)
```

```
Blue  Red 
  18   12 
```



5c

```
x <- data.frame("id" = 1:10, "Age" = round(runif(10, min = 20, max = 60)))
```

```
x$group <- "old"
```

```
x$group <- "younger"
```

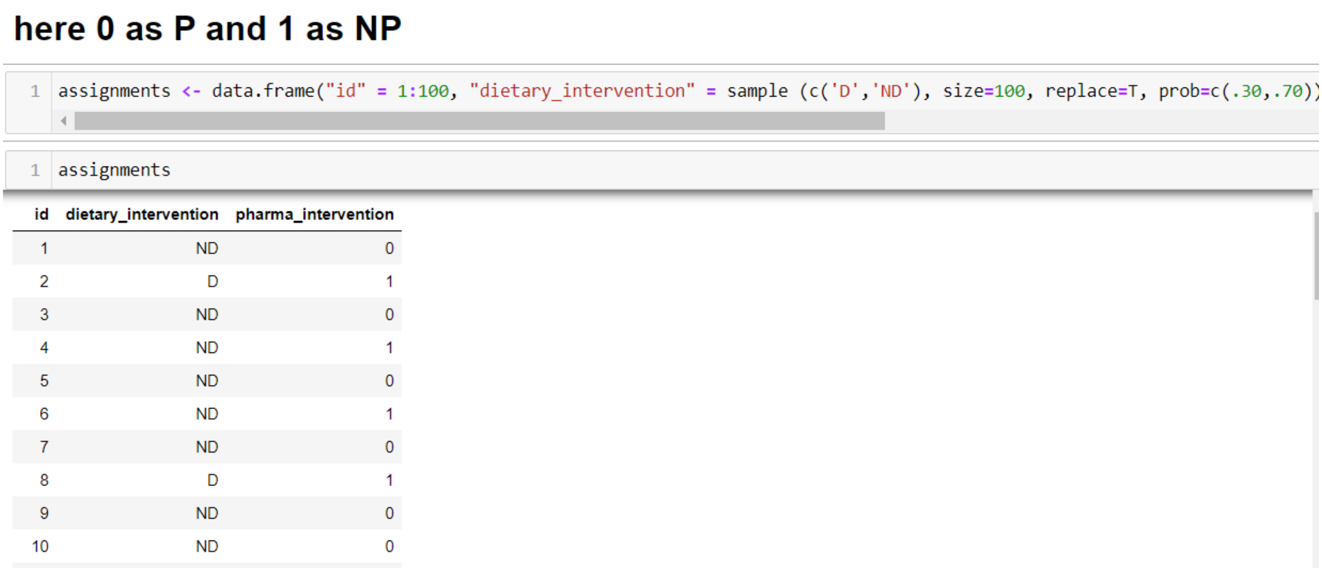
```
x[x$Age >= 45, ]$group <- "old"
```

```
x[x$Age > 0 & x$Age <= 45,]$group <- "younger"
```

```
x
```

id	Age	group
1	34	younger
2	26	younger
3	32	younger
4	21	younger
5	60	old
6	52	old
7	23	younger
8	55	old
9	42	younger
10	37	younger

5d.



here 0 as P and 1 as NP

```
assignments <- data.frame("id" = 1:100, "dietary_intervention" = sample (c('D','ND'), size=100, replace=T, prob=c(.30,.70)
```

```
assignments
```

id	dietary_intervention	pharma_intervention
1	ND	0
2	D	1
3	ND	0
4	ND	1
5	ND	0
6	ND	1
7	ND	0
8	D	1
9	ND	0
10	ND	0