Stat 021 Homework 1

Chongkyung Kim

Due: Fri, Sept 13

Q 2) Design your own experimental study (made up) on a population of your choice. Provide details on what is the population under study, how will you collect a sample, what are potential biases keeping your sample from truely being random and define different treatment/control groups and describle how you will randomly assign treatments to your observational units. (5 points)

Does healthy soil help Swarthmore trees grow?

Population: trees on the Swarthmore campus

Sample Selection: Use random generator (using python, for example) to select 200 trees.

Random Assignment: Use random generator and select 100 trees; they are the treatment group, they are provided with healthy soil, engineered from a lab. Rest of the 100 trees do no receive any treatment and we let them grow in their natural setting and state.

Treatment Group: the randomly selected trees that grow in the artificially engineered, healthy soil

Control Group: the other 100 trees that just grow in their natural, original setting

Potential Biases: There's a chance that really healthy trees, that perhaps might not have needed the help of healthy soil, were assigned to the treatment group, thus keeping our sample from being truely random. There's also a chance that the very unhealthy, weak trees were assigned to the control group, which could actually have benefited from the healthy soils. These types of unexpected factors might have prevented us from having a truely random sample.

Q 3) Access the data set called *sleep* in R. Note this data set is loaded into R automatically (you do not need to import it or install a package to access it). (10 points)

1. Define a new variable called *group1.sleep* that includes only the values of the variable *extra* for individuals from group 1.
2. Perform a t-test on "group1.sleep" to test of the extra hours slept by group 1 is smaller than or equal to 0.5 hours at an $\alpha = 0.1$ level. I.e. Test the hypothesis $H_0 : \mu = 0.5$ vs. $H_1 : \mu \neq 0.5$ at an $\alpha = 0.1$ significance level.
3. Report and interpret the 90% CI for the average extra hours of sleep for group 1.

We can be 90 percent confident that the true mean of the extra hours of sleep by group 1 is in the interval of (-inf , 1.53427]

4. Form a new categorical variable called *extra1.cat* that categorizes the variable *extra* into two groups, the first where extra hours slept is at least zero hours and the second where extra hours slept is less than zero hours. Print a table that counts the total number of observations in each group. (You may want to use the *ifelse()* function.)
5. Produce two boxplots for the variable *extra*, one corresponding to each group. Make sure each plot has a title. (You may find the function *grid.arrange()* in the package *gridExtra* useful for displaying two plots at once.)

```
data(sleep)

new_sleep <- sleep %>% filter(group==1) %>% mutate(group1.sleep = extra)

new_sleep
```

```
##    extra group ID group1.sleep
## 1    0.7     1  1          0.7
## 2   -1.6     1  2         -1.6
## 3   -0.2     1  3         -0.2
## 4   -1.2     1  4         -1.2
## 5   -0.1     1  5         -0.1
## 6    3.4     1  6          3.4
## 7    3.7     1  7          3.7
## 8    0.8     1  8          0.8
## 9    0.0     1  9          0.0
## 10   2.0     1 10          2.0
```

```
t.test(new_sleep$group1.sleep, alternative= "less" , mu=0.5,conf.level = .9)
```

```
##
##  One Sample t-test
##
## data:  new_sleep$group1.sleep
## t = 0.4419, df = 9, p-value = 0.6655
## alternative hypothesis: true mean is less than 0.5
## 90 percent confidence interval:
##      -Inf 1.532427
## sample estimates:
## mean of x
##      0.75
```

```
categorized_sleep1 <-sleep %>% filter(extra <0) %>% mutate(extra1.cat = "less than zero")
categorized_sleep2 <- sleep %>% filter(extra>=0) %>% mutate(extra1.cat ="at least zero")

categorized_sleep = rbind(categorized_sleep1,categorized_sleep2)

categorized_sleep
```

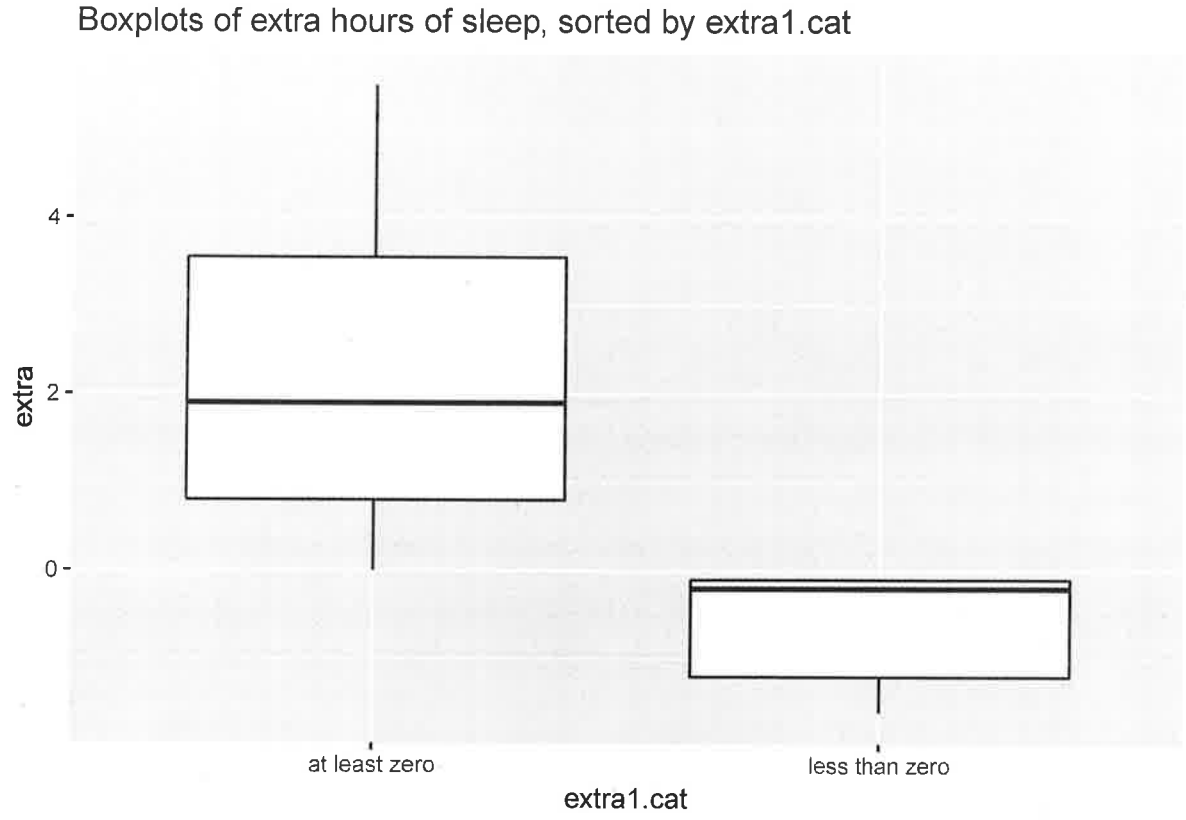
```
##    extra group ID     extra1.cat
## 1   -1.6     1  2 less than zero
## 2   -0.2     1  3 less than zero
## 3   -1.2     1  4 less than zero
## 4   -0.1     1  5 less than zero
## 5   -0.1     2  5 less than zero
## 6    0.7     1  1  at least zero
## 7    3.4     1  6  at least zero
## 8    3.7     1  7  at least zero
## 9    0.8     1  8  at least zero
## 10   0.0     1  9  at least zero
## 11   2.0     1 10  at least zero
## 12   1.9     2  1  at least zero
## 13   0.8     2  2  at least zero
## 14   1.1     2  3  at least zero
## 15   0.1     2  4  at least zero
## 16   4.4     2  6  at least zero
## 17   5.5     2  7  at least zero
## 18   1.6     2  8  at least zero
## 19   4.6     2  9  at least zero
## 20   3.4     2 10  at least zero
```

```
categorized_sleep %>% count(extra1.cat)
```

```
## # A tibble: 2 x 2
##   extra1.cat         n
##   <chr>          <int>
## 1 at least zero     15
## 2 less than zero     5
```

```
count_sleep <- categorized_sleep %>% count(extra1.cat)

p1 <- ggplot(data=categorized_sleep)+geom_boxplot(mapping = aes(extra1.cat,extra)) +
       labs(title = "Boxplots of extra hours of sleep, sorted by extra1.cat")
p1
```



Definition that $E(x) = \mu$ and $Var(x) = \sigma^2$.

then,

$$E(z) = E\left(\frac{X-\mu}{\sigma}\right) = E\left(\frac{x}{\sigma} - \frac{\mu}{\sigma}\right)$$

$$= \frac{E(x)}{\sigma} - \frac{\mu}{\sigma}$$

$$= \frac{\mu}{\sigma} - \frac{\mu}{\sigma} \quad (\text{since } E(x) = \mu)$$

$$= 0$$

$$Var(z) = Var\left(\frac{x-\mu}{\sigma}\right) = Var\left(\frac{X}{\sigma} - \frac{\mu}{\sigma}\right)$$

$$= \frac{1}{\sigma^2} Var(x) \quad (\text{since } Var(x) = \sigma^2)$$

$$= 1.$$