STA 210: HW 1

Jackson Hubbard

September 12th, 2018

Question 1: Ex. #2.16

Based on the summaries and t-test below, the sumamry statements from the textbook are true for both the intrinsic and extrinsic categories.

```
#use case0101 data set from the Sleuth3 library
library("Sleuth3")
q1 <- case0101
intrinsic <- as.data.frame(q1 %>% filter(q1$Treatment == "Intrinsic")) %>% select(Score

extrinsic <- as.data.frame(q1 %>% filter(q1$Treatment == "Extrinsic")) %>% select(Score
intrinsic %>% summarise(n=n(), mean = mean(Score), sd = sd(Score))
```

```
##    n     mean       sd
## 1 24 19.88333 4.439513
```

```
extrinsic %>% summarise(n=n(), mean = mean(Score), sd = sd(Score))
```

```
##    n     mean       sd
## 1 23 15.73913 5.252596
```

```
t.test(intrinsic, extrinsic, alternative = "two.sided")
```

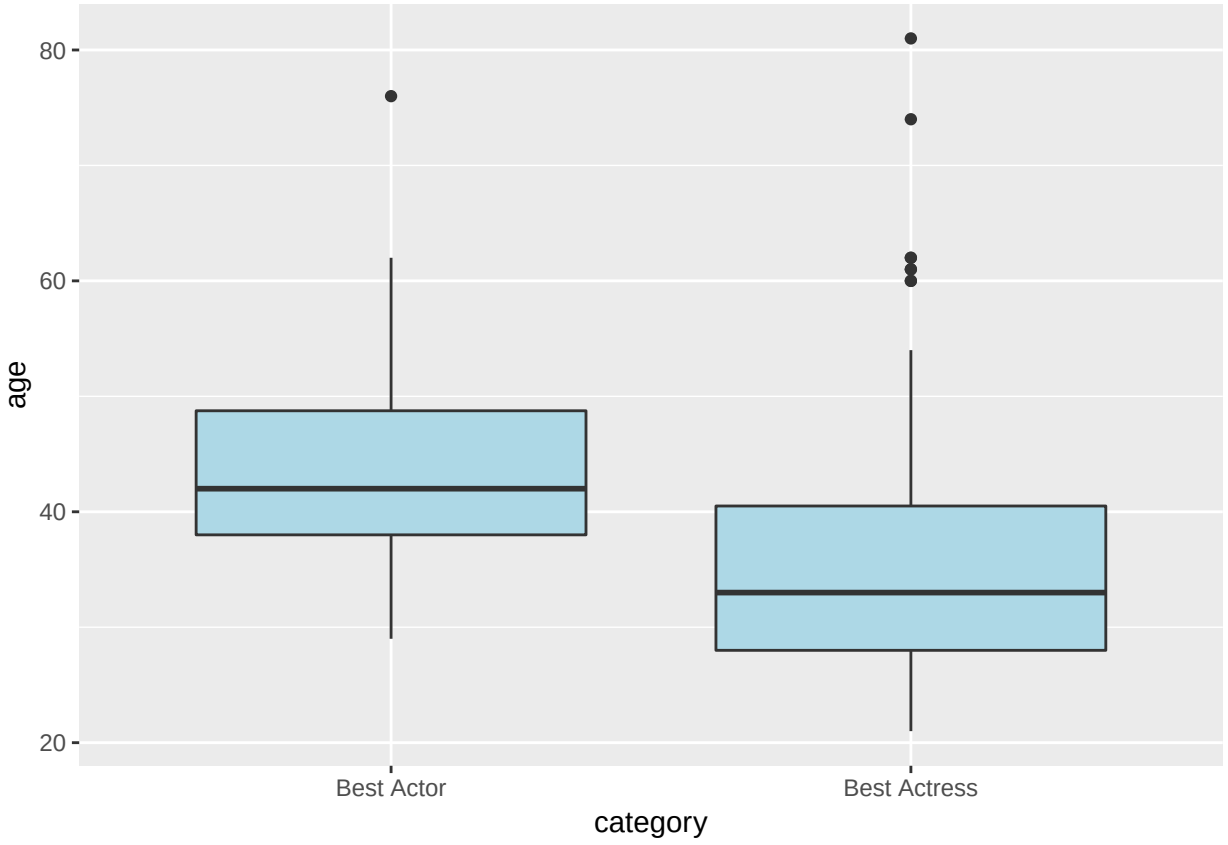
```
##
##  Welch Two Sample t-test
##
## data:  intrinsic and extrinsic
## t = 2.9153, df = 43.108, p-value = 0.005618
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.277603 7.010803
## sample estimates:
## mean of x mean of y
##  19.88333  15.73913
```

Question 2: Oscars and Age

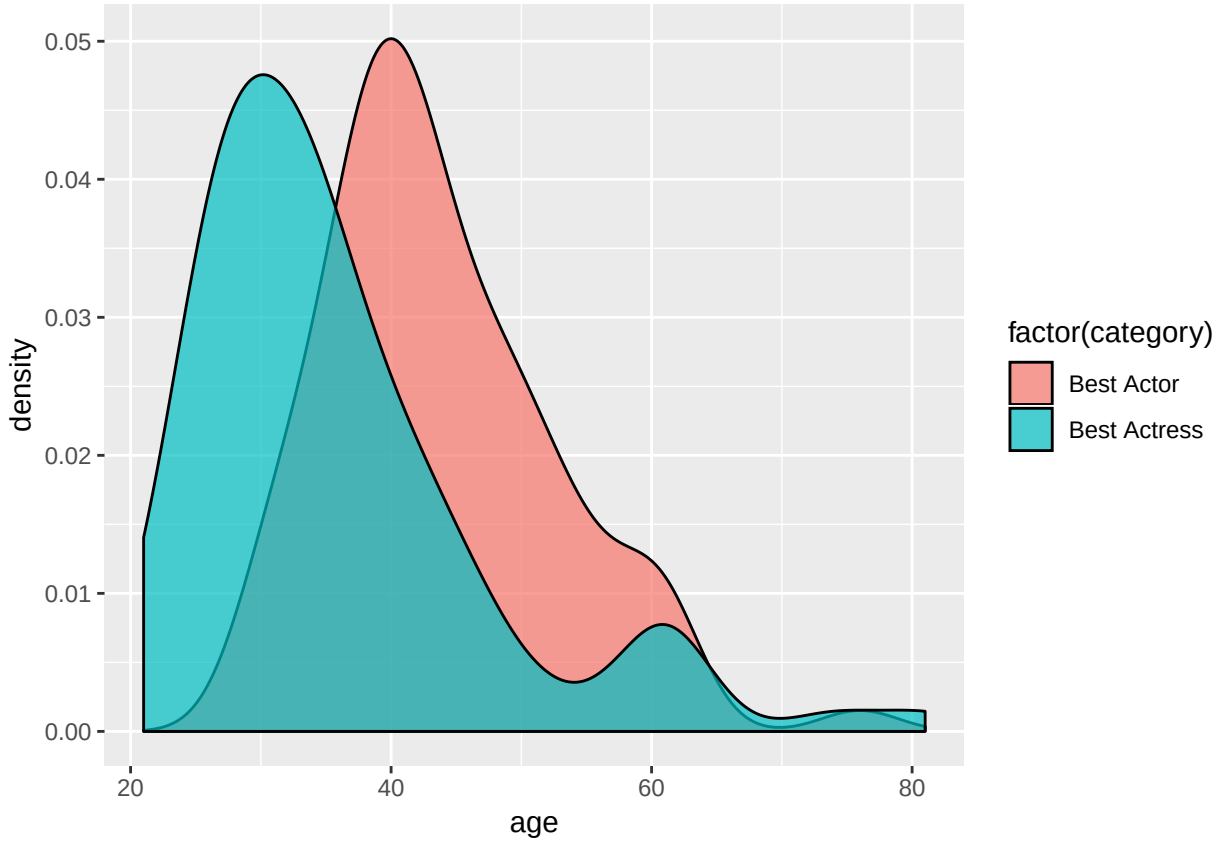
a

Consider the distributions of age for the Best Actor and Best Actress winners. Based on the distributions of age, is it appropriate to use the two-sample t inference methods? Looking at the data, we see that the sample size for each category (actors and actresses) is equal at 90 cases each. The age distribution for actors is slightly higher than for actresses. Both distributions are skewed to the right due to the outliers of older actors/actresses winning the oscars. Further, the two data sets have approximately the same shape, as well as only slightly different standard deviations. This means that only assumption that is broken is the skew to the right and the slightly unequal variance . However, this is made up for by the large and equal sample size. Also, the independence assumption is not broken since there is mo cluster or serial effetcs.

```
# Use the oscar_winners data set
best_actor <- as.data.frame(oscar_winners %>% filter(category=="Best Actor") %>% select
best_actress <- as.data.frame(oscar_winners %>% filter(category=="Best Actress") %>% se
ggplot(oscar_winners, aes(category, age))+ geom_boxplot(fill="light blue")
```



```
ggplot(oscar_winners, aes(age)) +geom_density(aes(fill = factor(category)),alpha=0.7)
```




```
best_actor %>% summarise(n=n(), mean=mean(age), median = median(age), s=sd(age))
```

```
##    n     mean median        s
## 1 90 43.82222     42 8.878174
```

```
best_actress %>% summarise(n=n(), mean=mean(age), median = median(age), s=sd(age))
```

```
##    n     mean median        s
## 1 90 36.02222     33 11.58358
```

b

Before doing the t-test, we want to set the parameter paired= FALSE because we want a two sample t-test and not a paired hypothesis test. The var.equal parameter is FALSE because as seen above, the standard deviations of the two categories are different. After doing the hypothesis test on the average difference in age of actors and actress oscar winners, we see that the p-value is 5.242e-07. This p-value is very small, which means that there is sufficient evidence to reject the null hypothesis of there being no difference in average age of actor winners and actress winners. Further, there is sufficient evidence that the average age of oscar actor winners is higher than actress oscar winners.

```r
t.test(best_actor, best_actress, alternative = "greater", paired= FALSE, var.equal = FAL
```

```
##
##  Welch Two Sample t-test
##
## data:  best_actor and best_actress
## t = 5.0702, df = 166.74, p-value = 5.242e-07
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  5.255416      Inf
## sample estimates:
## mean of x mean of y
##  43.82222  36.02222
```

c

No my friend is not correct in saying "the probability that actors and actresses have the same average age is 5.242e-07". The correct interpretation of the p-value is that given that the null hypothesis is true, the probability of observing a difference in average ages of actor oscar winners and actress winners that is as or more extreme than the sample is 5.242e-07.

d

After perforiming this confidence interval, we see that the interval is (5.255, 10.344). This means that this interval will capture the true difference in means between the age of actor and actress oscar winners 90% of the time.

```r
t.test(best_actor, best_actress, alternative= "two.sided", var.equal=FALSE, conf.level =
```

```
##
##  Welch Two Sample t-test
##
## data:  best_actor and best_actress
## t = 5.0702, df = 166.74, p-value = 1.048e-06
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##   5.255416 10.344584
## sample estimates:
## mean of x mean of y
##  43.82222  36.02222
```

e

No it is not reasonable to make conclusions about the differences in average age between movie actors and actresses based off of this dataset. This is beacuse this data set only included data points from oscar winners. Thus, this data set is not representative of Actors' and actresses' age distribution as a whole.

Question 3: Ex #3.33

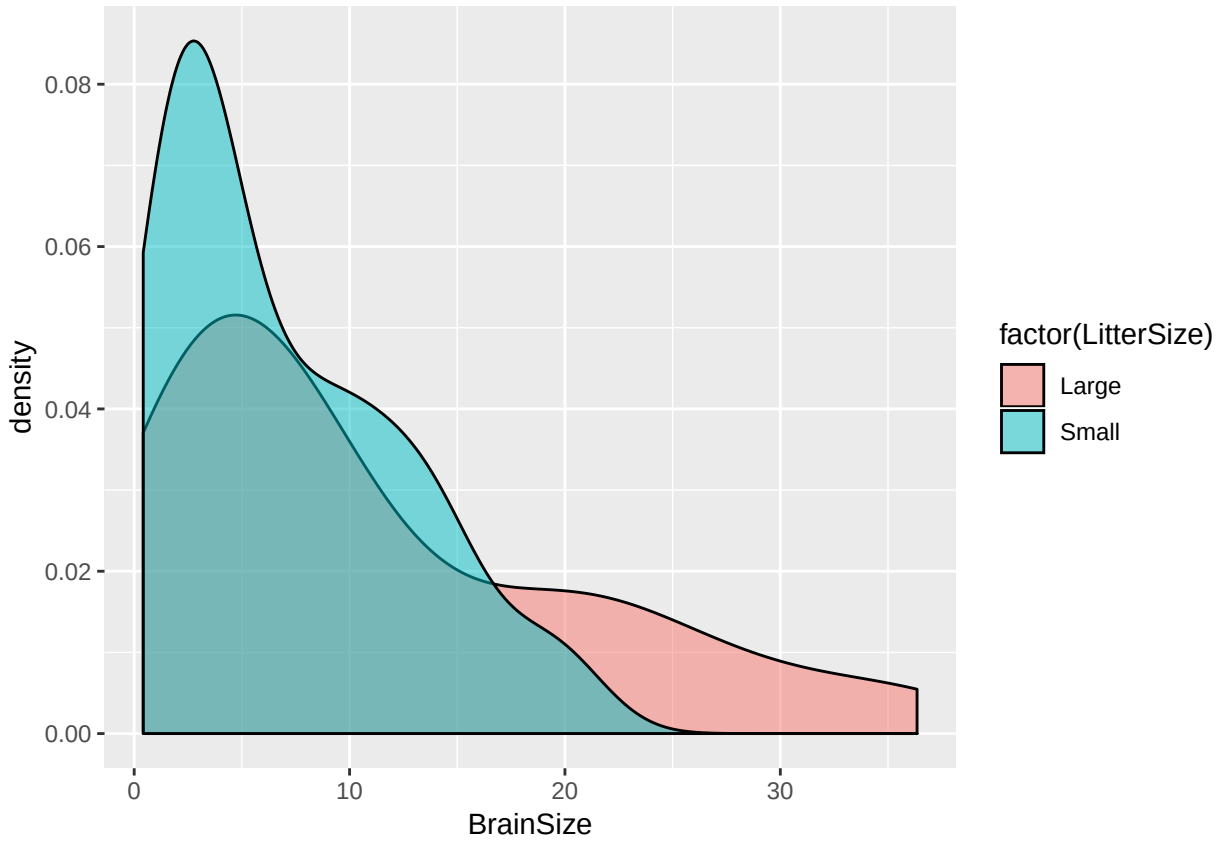
```
# Use the ex0333 dataset from the Sleuth3 package
q3 <- ex0333
large_litter <- as.data.frame(q3 %>% filter(LitterSize == "Large") %>% select(BrainSize
small_litter <- as.data.frame(q3 %>% filter(LitterSize == "Small") %>% select(BrainSize



small_litter %>% summarise(n=n(), mean=mean(BrainSize),
median = median(BrainSize), s=sd(BrainSize))
```

```
##    n     mean median        s
## 1 51 6.885882      5 5.460298
```

```
large_litter %>% summarise(n=n(), mean=mean(BrainSize),
median = median(BrainSize), s=sd(BrainSize))
```

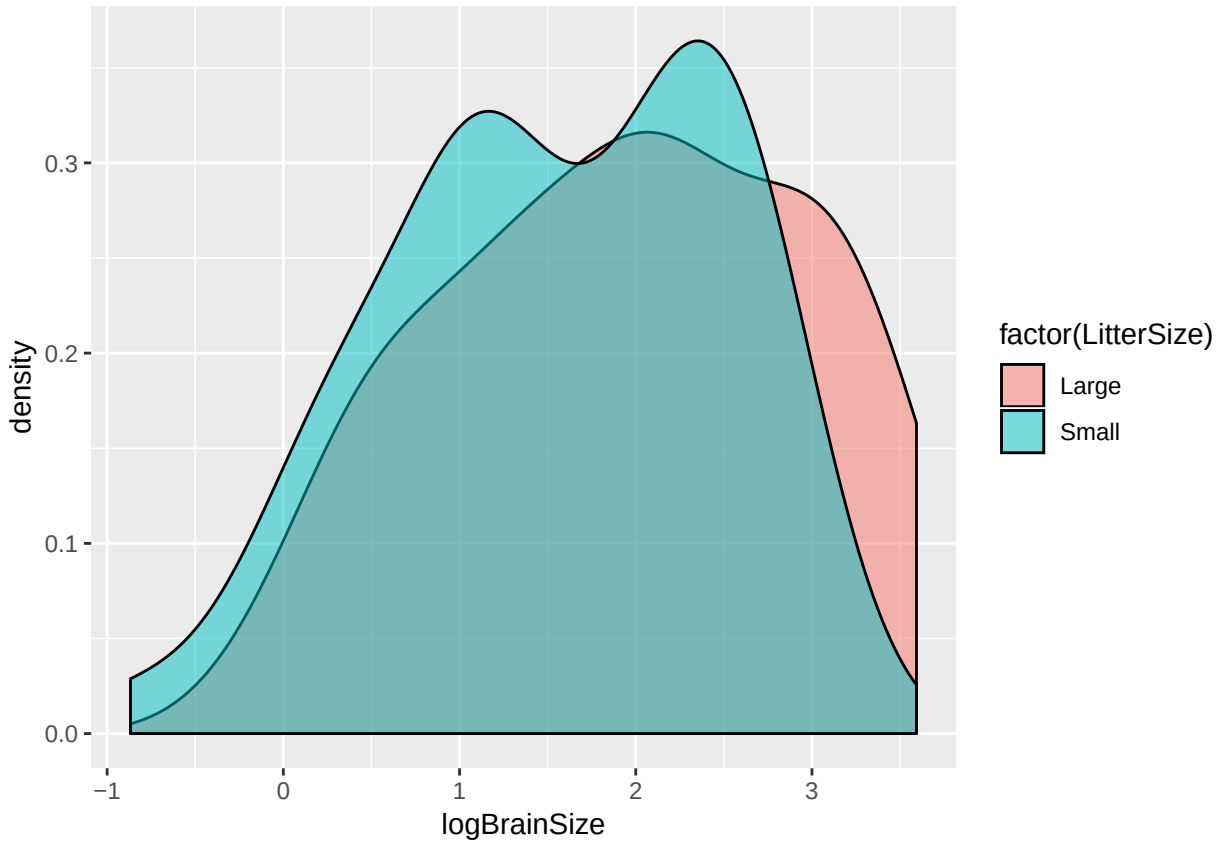
```
##    n     mean median        s
## 1 45 10.96844   7.97 9.836924
```

```
ggplot(q3, aes(BrainSize)) +
geom_density(aes(fill = factor(LitterSize)),alpha=0.5)
```



Looking at the summary statistics and boxplot distribution of brain size for mammalls who have small vs large litter sizes, it is clear that the standard deviations of the two categories are not the same. Further, mammals with large litter sizes have a larger brain size on average and also have a higher variation. Next, looking at the density graphs, we see that the data violates the normality assumption of a two sample t-test. They both have a skew to the right and do not have a high enough sample size to counter this. As a result, applying a log transformation would make this data easier to work with.

```
q3 <- q3 %>% mutate(logBrainSize = log(BrainSize))
ggplot(q3, aes(logBrainSize)) + geom_density(aes(fill = factor(LitterSize)),alpha=0.5)
```



Now that the data is more normally distributed, we can redo the anylsis.

```
q3 %>% filter(LitterSize=="Small") %>% summarise(n=n(), mean=mean(logBrainSize), median
```

```
##    n     mean   median         s
## 1 51 1.552458 1.609438 0.9522342
```

```
q3 %>% filter(LitterSize=="Large") %>% summarise(n=n(), mean=mean(logBrainSize), median
```

```
##    n     mean   median         s
## 1 45 1.949426 2.075684 1.016293
```

Looking at the summary statistics, the standard deviations are now approximately similar and the same size is (still) the same. Further, the independence assumption is also not violated since the outcome of one observation does not seem to influence another observation. There are no cluster or serial effects either. With no violations, now I can perform a t-test.

mu1 = average brain size of animals with large litter mu2 = average brain size of animals with small litter

```
small_litter2 <- q3 %>% filter(LitterSize=="Small")
large_litter2 <- q3 %>% filter(LitterSize=="Large")

t.test(small_litter2$logBrain, large_litter2$logBrain, alternative="two.sided", paired=F
```

```
##
##  Welch Two Sample t-test
##
## data:  small_litter2$logBrain and large_litter2$logBrain
## t = -1.9669, df = 90.684, p-value = 0.05225
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.797879510  0.003942989
## sample estimates:
## mean of x mean of y
##  1.552458  1.949426
```

Looking at the t-test, we see that the p value is 0.05225. This is not smaller than the standard threshold of 0.05, which means that there is not sufficient evidence to reject the null hypothesis. This can also be seen in the confidence interval on logBrainSize which inlcudes 0. Transforming this confidence interval from logBrainSize to BrainSize results in an interval of (0.450, 1.004).