Final Project Part II

Part I

Author: Davut Emrah Ayan Now in the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package.

1. Load the ToothGrowth data and perform some basic exploratory data analyses

```
library(datasets)
data(ToothGrowth)
```

In the data set, there are 3 variables. len and dose are numerical, supp is factor variable with two levels.

```
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

2. Provide a basic summary of the data.

It is good that there is no missing value in the data. Factors in supp variable is evenly distributed. Min-max, mean and summary statistics of the numerical variables are shown below.

```
library(skimr)
skim(ToothGrowth)
```

Table 1: Data summary

Name	ToothGrowth
Number of rows	60
Number of columns	3
Column type frequency:	
factor	1
numeric	2
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
supp	0	1	FALSE	2	OJ: 30, VC: 30

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
len	0	1	18.81	7.65	4.2	13.07	19.25	25.27	33.9	
dose	0	1	1.17	0.63	0.5	0.50	1.00	2.00	2.0	

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
ToothGrowth %>%
    group_by(supp, dose) %>%
    summarise(mean_length = mean(len),
              .groups = "drop")
```

```
## # A tibble: 6 x 3
##   supp   dose mean_length
##   <fct> <dbl>       <dbl>
## 1 OJ      0.5        13.2
## 2 OJ      1          22.7
## 3 OJ      2          26.1
## 4 VC      0.5         7.98
## 5 VC      1          16.8
## 6 VC      2          26.1
```

3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering)

by supp Here we can see the difference in means of length between OJ and VC group.

```
ToothGrowth %>%
    group_by(supp) %>%
    summarise(mean_length = mean(len),
              .groups = "drop")
```

```
## # A tibble: 2 x 2
##   supp  mean_length
##   <fct>       <dbl>
## 1 OJ           20.7
## 2 VC           17.0
```

I test the hypotheses of equality in means of length. I don't have information about the groups tested, whether they are the same group tested twice or randomized two independent groups. For this reason I will test with paired and unpaired twice.

Below test results assume unpaired groups and constant population variance.

T-statistics of mean equality is 1.9153, with 58 degrees of freedom. 95% confidence interval for mean

difference is [-0.17 7.57]. Since CI contains 0 and p-value = 0.06, we can conclude that we can not reject the null hypotheses of means are equal at 5% significance level.

```
t.test(len ~ supp, data = ToothGrowth, paired = FALSE, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 58, p-value = 0.06039
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  -0.1670064  7.5670064
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

Below test results assume paired groups and constant population variance.

T-statistics of mean equality is 3.3026, with 29 degrees of freedom. 95% confidence interval for mean difference is [1.41 5.99]. Since CI does not contain 0 and p-value = 0.003, we can conclude that we reject the null hypotheses of means are equal at 5% significance level.

```
t.test(len ~ supp, data = ToothGrowth, paired = TRUE, var.equal = T)
```

```
##
##  Paired t-test
##
## data:  len by supp
## t = 3.3026, df = 29, p-value = 0.00255
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.408659 5.991341
## sample estimates:
## mean of the differences
##                     3.7
```

by dose Here we can see there is clear differences in means of length by dose. Mean length of tooth grow increase by dose.

```
ToothGrowth %>%
    group_by(dose) %>%
    summarise(meanlen = mean(len))
```

```
## # A tibble: 3 x 2
##    dose meanlen
##   <dbl>   <dbl>
## 1   0.5    10.6
## 2   1      19.7
## 3   2      26.1
```

```
dose05 = ToothGrowth[ToothGrowth$dose==0.5, "len"]
dose1 = ToothGrowth[ToothGrowth$dose==1, "len"]
dose2 = ToothGrowth[ToothGrowth$dose==2, "len"]
```

T-test results for the equality of means of dose 0.5 and dose 1 shows below that either paired or unpaired we reject the null hypotheses at very low significance level.

```r
t.test(dose05, dose1, paired = T, var.equal = T)
```

```
##
##  Paired t-test
##
## data:  dose05 and dose1
## t = -6.9669, df = 19, p-value = 1.225e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.872879  -6.387121
## sample estimates:
## mean of the differences
##                   -9.13
```

```r
t.test(dose05, dose1, paired = F, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  dose05 and dose1
## t = -6.4766, df = 38, p-value = 1.266e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983748  -6.276252
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

Similarly, I can reject the null hypothesis that mean lengths are equal when dose 0.5 and 2 with very small p-values.

```r
t.test(dose05, dose2, paired = T, var.equal = T)
```

```
##
##  Paired t-test
##
## data:  dose05 and dose2
## t = -11.291, df = 19, p-value = 7.19e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.3672 -12.6228
## sample estimates:
## mean of the differences
##                  -15.495
```

```r
t.test(dose05, dose2, paired = F, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  dose05 and dose2
## t = -11.799, df = 38, p-value = 2.838e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15352 -12.83648
## sample estimates:
```

```
## mean of x mean of y
##    10.605    26.100
```

Finally, I can reject the equality of mean lengths when dosage are 1 and 2 at very low significance levels.

```
t.test(dose1, dose2, paired = T, var.equal = T)
```

```
##
##  Paired t-test
##
## data:  dose1 and dose2
## t = -4.6046, df = 19, p-value = 0.0001934
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.258186 -3.471814
## sample estimates:
## mean of the differences
##                  -6.365
```

```
t.test(dose1, dose2, paired = F, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  dose1 and dose2
## t = -4.9005, df = 38, p-value = 1.811e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.994387 -3.735613
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

4. State your conclusions and the assumptions needed for your conclusions.

I tested several hypothesis of mean equalities among groups of dose and supp.

By dose levels, all three group means are found to be significantly different, when I assume paired and unpaired but equal variances for all tests.

By supp, I fail to reject the mean equality when I assume unpaired (two sample test). It did not change the decision when I assume equal or unequal Variances.

But when I assumed one sample (paired), I can reject the mean equality.

So, I will need more information about how data is collected to make better decisions.