ToothGrowth data analysis

Load the ToothGrowth data and perform some basic exploratory data analysis

The below data plot shows that average tooth length has a linear relationship with amount of dose. Higher the dose, longer the tooth.

```
library(ggplot2)
lenBYdose <- NULL
data(ToothGrowth)
lenBYdose <- aggregate(ToothGrowth$len, by=list(ToothGrowth$dose), FUN=mean)
names(lenBYdose) <- c("len", "dose")

ggplot(lenBYdose, aes(dose, len, label=round(dose,1))) +
   geom_line() +
   ylab("Average tooth length") +
   xlab("Dose in milligrams") +
   ggtitle("Average Tooth Length by Dose") +
   geom_text()
```

The following plot is a different view of the positive relationship between tooth length and the dose levels of Vitamin C for both delivery methods. The difference in tooth length is significant when dose amount is 0.5mm or 1mm. At 2mm, the difference becomes less significant.

```
ggplot(ToothGrowth, aes(as.factor(dose), len, fill=supp, label=round(len,1))) +
    geom_bar(stat="identity") +
    facet_grid(. ~ supp) +
    xlab("Dose in milligrams") +
    ylab("Tooth length") +
    guides(fill=guide_legend(title="Supplement type"))
```
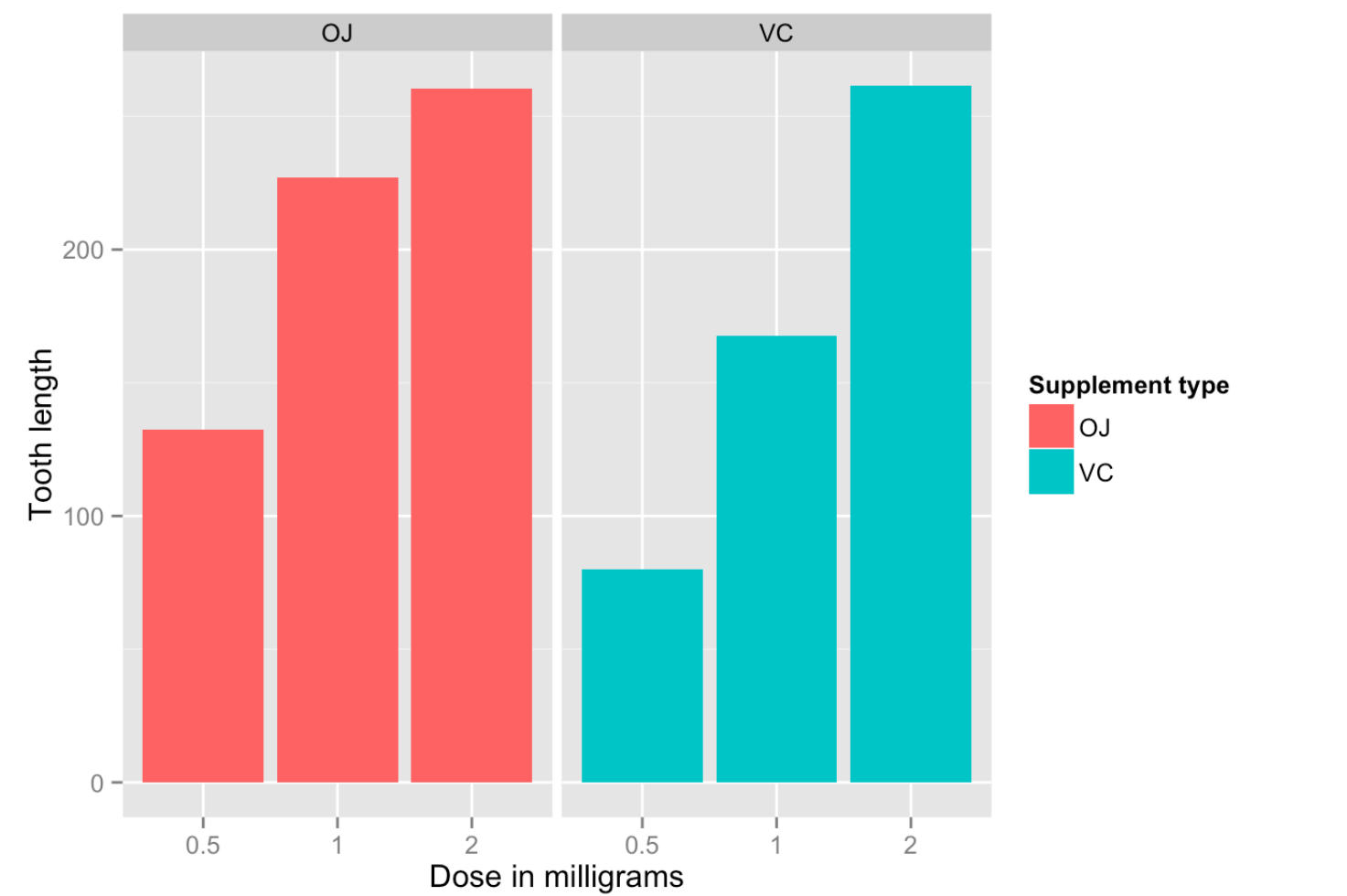

Basic summary

```
summary(ToothGrowth)
```

```
##       len           supp          dose
##  Min.   : 4.20    OJ:30    Min.   :0.500
##  1st Qu.:13.07    VC:30    1st Qu.:0.500
##  Median :19.25             Median :1.000
##  Mean   :18.81             Mean   :1.167
##  3rd Qu.:25.27             3rd Qu.:2.000
##  Max.   :33.90             Max.   :2.000
```

ToothGrowth is a dataset of type data.frame with:

number of observations: 60

number of columns: 3

column names: len, supp, dose

Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. State your conclusions

We will us use T-Tests to compare tooth growth by supplement type and dosage amount.

First, we will compare tooth growth by supplement type

```
t.test(len ~ supp, data=ToothGrowth)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##          20.66333          16.96333
```

The p-value of 0.06 indicates that if we were to repeat our experiment many, many times (each time we select 30 observations for each of supplement type at random and compute the sample mean) then 6 times out of 100 we can expect to see a sample mean greater than mean of the group for each supplement type.

The 95% confidence interval shows that there is a 95% chance that the confidence interval -0.1710156, 7.5710156 contains the true population mean. If repeated samples were taken and the 95% confidence interval computed for each sample, 95% of the intervals would contain the population mean.

Next, we will compare tooth growth by dosage amount by subsetting data in 3 pairs of doses.

```
t.test(len ~ dose, data=subset(ToothGrowth, dose %in% c(0.5, 1.0)))
```

```
##
##   Welch Two Sample t-test
##
## data:   len by dose
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   -11.983781  -6.276219
## sample estimates:
## mean in group 0.5    mean in group 1
##             10.605             19.735
```

The p-value of 1.268300710^{-7} indicates that if we were to repeat our experiment many, many times (each time we select 30 observations for each of supplement type at random and compute the sample mean) then 0 times out of 100 we can expect to see a sample mean greater than mean of the group for each supplement type.

The 95% confidence interval shows that there is a 95% chance that the confidence interval -11.9837813, -6.2762187 contains the true population mean. If repeated samples were taken and the 95% confidence interval computed for each sample, 95% of the intervals would contain the population mean.

```
t.test(len ~ dose, data=subset(ToothGrowth, dose %in% c(1.0, 2.0)))
```

```
##
##   Welch Two Sample t-test
##
## data:   len by dose
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   -8.996481 -3.733519
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

The p-value of 1.906429510^{-5} indicates that if we were to repeat our experiment many, many times (each time we select 30 observations for each of supplement type at random and compute the sample mean) then 0 times out of 100 we can expect to see a sample mean greater than mean of the group for each supplement type.

The 95% confidence interval shows that there is a 95% chance that the confidence interval -8.9964805, -3.7335195 contains the true population mean. If repeated samples were taken and the 95% confidence interval computed for each sample, 95% of the intervals would contain the population mean.

```
t.test(len ~ dose, data=subset(ToothGrowth, dose==c(0.5, 2.0)))
```

```
##
##   Welch Two Sample t-test
##
## data:   len by dose
## t = -7.3335, df = 17.635, p-value = 9.362e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   -19.72833 -10.93167
## sample estimates:
## mean in group 0.5   mean in group 2
##             10.63              25.96
```

The p-value of 4.39752510^{-14} indicates that if we were to repeat our experiment many, many times (each time we select 30 observations for each of supplement type at random and compute the sample mean) then 0 times out of 100 we can expect to see a sample mean greater than mean of the group for each supplement type.

The 95% confidence interval shows that there is a 95% chance that the confidence interval -18.1561665, -12.8338335 contains the true population mean. If repeated samples were taken and the 95% confidence interval computed for each sample, 95% of the intervals would contain the population mean.

State the assumptions needed for your conclusions

1. Samples are random, independent, and come from normally distributed population with unknown but equal variances.
2. Populations follow a Gaussian distribution.

end of document