Project 1, Part 2 of Statistical Inference

Jin-Keat Lim

Friday, October 24, 2014

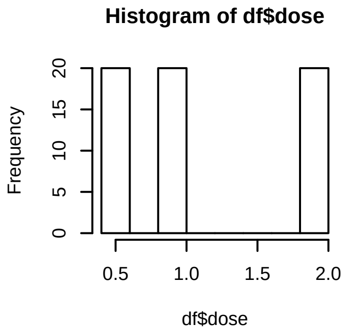
Introduction

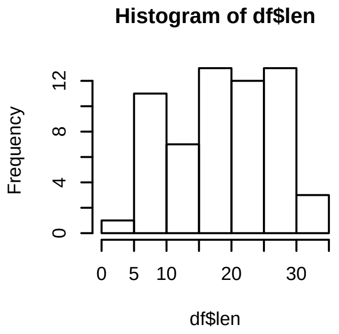
Before beginning the project, I simply loaded the ToothGrowth dataset into a dataframe. Note I hid the warnings that both "ggplot2" and "xtable" packages were built under R version 3.1.1.

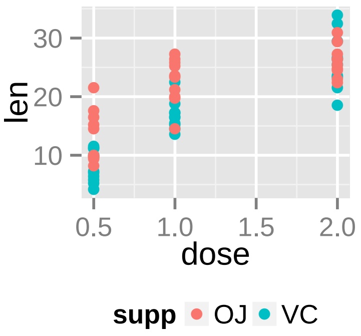
```r
library(ggplot2)
library(xtable)
df <- data.frame(ToothGrowth)
```

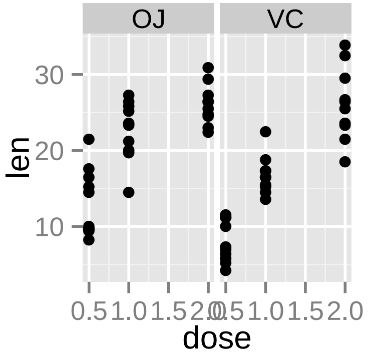
Part 1 of the project - basic Exploratory Data Analysis of the dataset

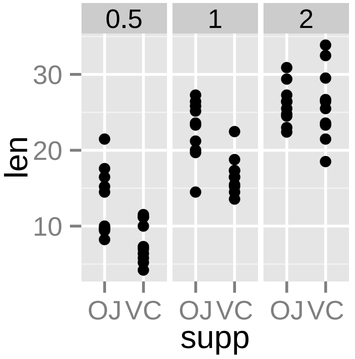
```r
hist(df$dose)
hist(df$len)

print(qplot(dose, len, data=df, color=supp, fill=supp) + theme(legend.position="bottom"))
print(qplot(dose, len, data=df, facets = .~supp))
print(qplot(supp, len, data=df, facets = .~dose))
```











Using various histograms and breakdowns, it is easy to see that the doses are separated in 3 categories. Further analysis will be needed to evaluate the effect of **supp** and **dose** on the ToothGrowth **len**.

Part 2 of the project - a basic summary of the data

```
options(xtable.comment = FALSE)
print(xtable(summary(df), caption="Summary of entire dataset"))
```

	len	supp	dose
1	Min. : 4.20	OJ:30	Min. :0.500
2	1st Qu.:13.07	VC:30	1st Qu.:0.500
3	Median :19.25		Median :1.000
4	Mean :18.81		Mean :1.167
5	3rd Qu.:25.27		3rd Qu.:2.000
6	Max. :33.90		Max. :2.000

Table 1: Summary of entire dataset

```
print(xtable(summary(df[df$supp=="OJ",]), caption="Summary of OJ supp"))
```

	len	supp	dose
1	Min. : 8.20	OJ:30	Min. :0.500
2	1st Qu.:15.53	VC: 0	1st Qu.:0.500
3	Median :22.70		Median :1.000
4	Mean :20.66		Mean :1.167
5	3rd Qu.:25.73		3rd Qu.:2.000
6	Max. :30.90		Max. :2.000

Table 2: Summary of OJ supp

```
print(xtable(summary(df[df$supp=="VC",]), caption="Summary of VC supp"))
```

	len	supp	dose
1	Min. : 4.20	OJ: 0	Min. :0.500
2	1st Qu.:11.20	VC:30	1st Qu.:0.500
3	Median :16.50		Median :1.000
4	Mean :16.96		Mean :1.167
5	3rd Qu.:23.10		3rd Qu.:2.000
6	Max. :33.90		Max. :2.000

Table 3: Summary of VC supp

Part 3 of the project - using confidence intervals and hypothesis tests to compare tooth growth by supp and dose.

```
VC_df <- df[df$supp=="VC",]
OJ_df <- df[df$supp=="OJ",]

#note I am using T-CI and the qt() function rather than using qnorm()

##T-CI by supp
OJ_len_CI <- mean(OJ_df$len) + c(-1,1) * qt(0.975, length(OJ_df$len)-1) +
              sd(OJ_df$len) / sqrt(length(OJ_df$len))
VC_len_CI <- mean(VC_df$len) + c(-1,1) * qt(0.975, length(VC_df$len)-1) +
              sd(VC_df$len) / sqrt(length(VC_df$len))

CI_df <- t(data.frame(OJ_len_CI, VC_len_CI))
colnames(CI_df) <- c("lower limit", "upper limit")
xtable(CI_df, caption="Summary of T-CI lengths by supp")
```

	lower limit	upper limit
OJ_len_CI	19.82	23.91
VC_len_CI	16.43	20.52

Table 4: Summary of T-CI lengths by supp

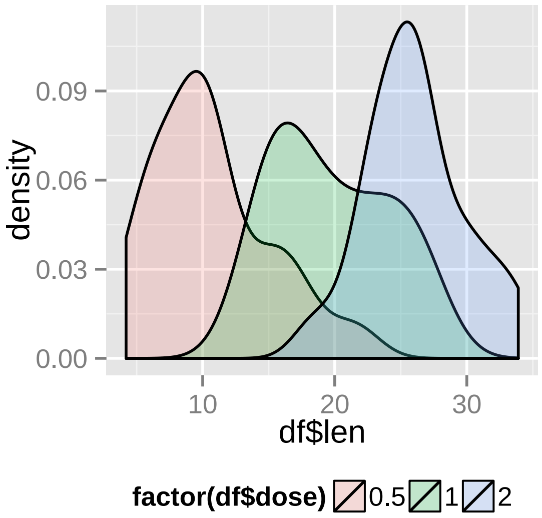
```
##T-CI by dose
dose1_df <- df[df$dose==0.5,];dose2_df <- df[df$dose==1,];dose3_df <- df[df$dose==2,]

dose1_len_CI <- mean(dose1_df$len) + c(-1,1) * qt(0.975, length(dose1_df)-1) +
  sd(dose1_df$len) / sqrt(length(dose1_df$len))
dose2_len_CI <- mean(dose2_df$len) + c(-1,1) * qt(0.975, length(dose2_df)-1) +
  sd(dose2_df$len) / sqrt(length(dose2_df$len))
dose3_len_CI <- mean(dose3_df$len) + c(-1,1) * qt(0.975, length(dose3_df)-1) +
  sd(dose3_df$len) / sqrt(length(dose3_df$len))

CI_dose_df <- t(data.frame(dose1_len_CI, dose2_len_CI, dose3_len_CI))
colnames(CI_dose_df) <- c("lower limit", "upper limit")
xtable(CI_dose_df, caption="Summary of T-CI lengths by dose")
```

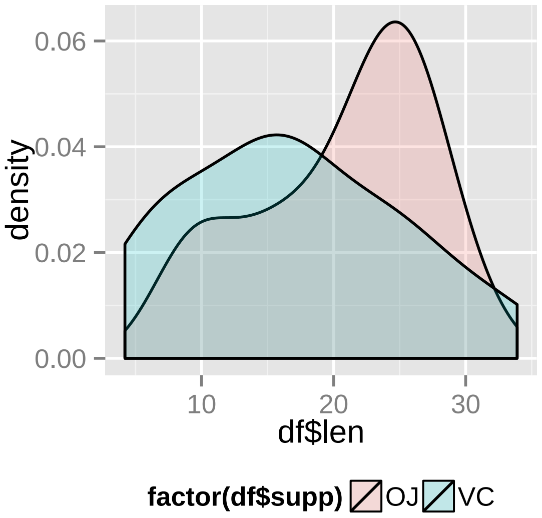
	lower limit	upper limit
dose1_len_CI	7.31	15.91
dose2_len_CI	16.42	25.02
dose3_len_CI	22.64	31.25

Table 5: Summary of T-CI lengths by dose

For easier visualization and to also help determine a relevant hypothesis test, I plotted the density of the ToothGrowth **len** broken down by **supp** and **dose**:

```
#taking a look at the density of the tooth growth broken down by dose
plot_dose <- ggplot(df, aes(x=df$len, fill=factor(df$dose))) +
        geom_density(alpha=0.2, binwidth=1) +
        theme(legend.position="bottom")
print(plot_dose)

#taking a look at the density of the tooth growth broken down by supp
plot_supp <- ggplot(df, aes(x=df$len, fill=factor(df$supp))) +
  geom_density(alpha=0.2, binwidth=1) +
  theme(legend.position="bottom")
print(plot_supp)
```





Based on these plots, we can subjectively see that the **dose** affects the **len**, and the effect of the **supp** on the **len** is less clear. I decided to first construct a hypothesis test to test the effect of **supp** on **len**:

Null hypothesis:the mean length of group **supp** OJ is equal to the mean length of group **supp** VC

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

```
##
##  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 is 6.06%, and is the probability of the null being true. Therefore, this can be considered a failure to reject the null hypothesis (assuming alpha is 5% using a standard 95% CI).

Because there are three discrete **dose** values, constructing a hypothesis test to test the effect of **dose** on **len** require testing of the below null hypotheses:

- H_A: mean length of group **dose** 0.5 is equal to the mean length of group **dose** 1.0
- H_B: mean length of group **dose** 1.0 is equal to the mean length of group **dose** 2.0
- H_C: mean length of group **dose** 0.5 is equal to the mean length of group **dose** 2.0

```
#testing Ha, Hb, Hc, respectively
t.test(len ~ dose, paired=FALSE, var.equal=FALSE, data= subset(df, dose %in% c(0.5,1)))
t.test(len ~ dose, paired=FALSE, var.equal=FALSE, data= subset(df, dose %in% c(1,2)))
t.test(len ~ dose, paired=FALSE, var.equal=FALSE, data= subset(df, dose %in% c(0.5,2)))
```

```
##
##  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
##
##
##  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
##
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
```

All 3 of these tests show small p-values, and the null hypotheses H_{A}, H_{B}, H_{C} can all be rejected. We can conclude the alternative hypotheses that the **dose** administered does make a difference in the **len**.

Part 4 - Conclusion and Assumptions

For all hypothesis testing, an alpha of 0.05 was assumed (standard 95% CI). In addition, each **t.test()** assumed unequal variances between each corresponding group.

When testing the effect of **supp** on **len**, the p-value 0.0606 > alpha 0.05, resulting in a failure to reject the null hypothesis. There is not sufficient evidence to say that the **supp** administered makes no difference in the **len**.

When testing the effect of **dose** on **len**, the H_{A}, H_{B}, H_{C} p-values << alpha 0.05, resulting in a rejection of the null hypothesis. We can state there is sufficient evidence to reject the null hypotheses, and accept the alternate hypotheses that the **dose** administered makes a difference in the **len**. This is supported by the density plot observed earlier.