Statistical Inferences and simulations

Shaun

February 19, 2017

Part 1

Overview

For the purpose of this assignment, we'll be exploring the estimation of population parameters from samples by using the Central Limit Theorem. It will be shown that, given enough samples, the mean of will be approximately normal and act as a good estimate for the population.

Simulations

The following code generates a dataframe with a 1000 rows of 40 randomly generated numbers (40000 in total) from an exponential distribution. It then calculates the mean and standard deviation of each row.

```
# Load required libraries
library(ggplot2)
library(gridExtra)
set.seed(42)        # Set random number generator seed for reproducibility
n <- 40             # Number of exponentials to generate per simulation
lambda <- 0.2       # Rate parameter for rexp function
mu <- 1/lambda      # mean
sig <- 1/lambda     # standard deviation
vars <- sig^2       # calculate the variance
sim.count <- 1000   # Number of simulation to run
dis.matrix <- NULL  # Initialising data frame

# Run simulation to generate a sim.count by n (1000 x 40) data frame
for(i in 1:sim.count){
  dis.matrix <- rbind(dis.matrix,rexp(n, lambda))
}
dis.matrix.stacked <- as.data.frame(stack(as.data.frame(dis.matrix))$values)

# Calculate the row means
rms <- as.data.frame(rowMeans(dis.matrix))
colnames(rms) <- "mean"

# Calculate the row standard deviations
sds <- as.data.frame(apply(dis.matrix, 1, sd))
colnames(sds) <- "std"
```

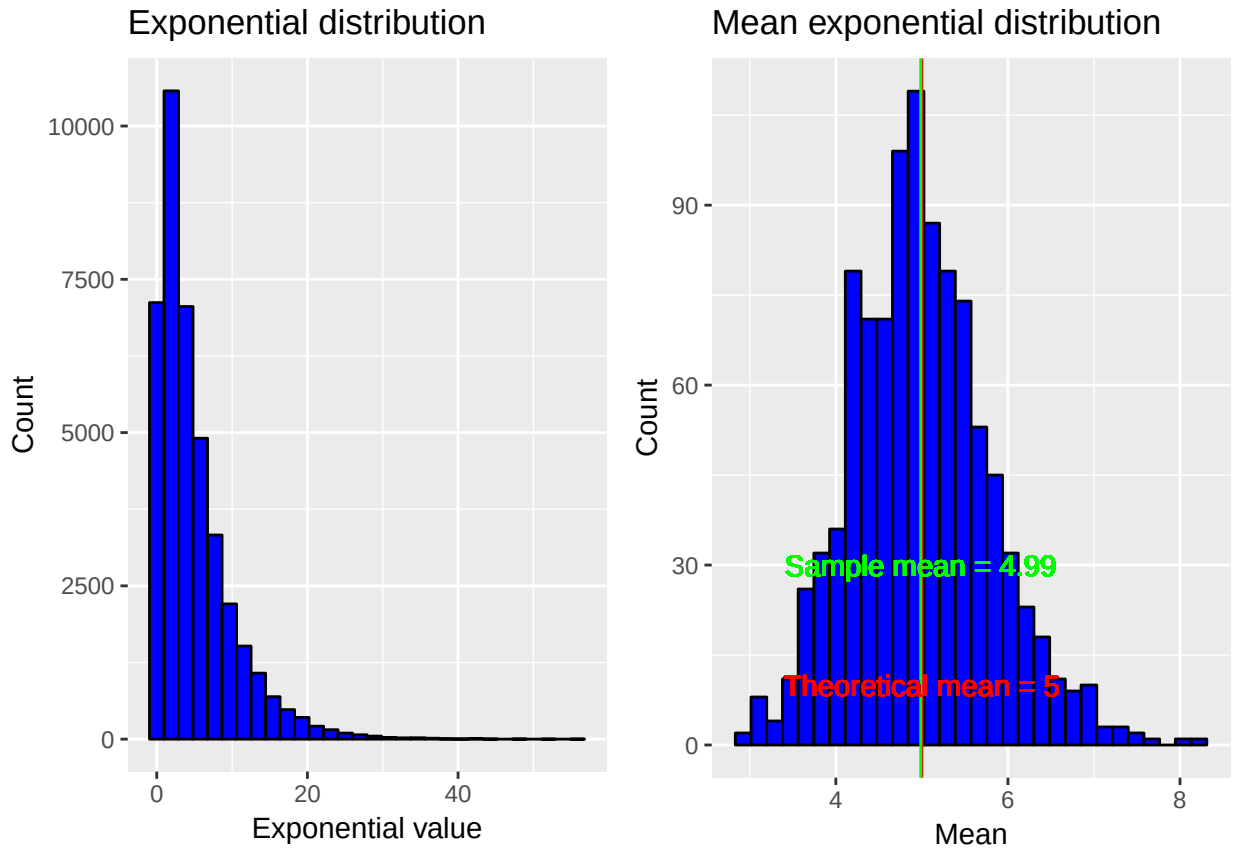
Sample Mean versus Theoretical Mean

The first graph represents a histogram of the random samples generated from a exponential distribution. The second graph represents the mean of 40 exponential generated samples using the same parameters. The mean of the mean sample distribution has a 95% confidence interval of 4.9370807, 5.035936 where the theoretical mean is 5.

```
mean.plot <- ggplot(rms, aes(x= rms))+
  geom_histogram(fill="blue", colour="black", bins = 30)+
  scale_x_continuous()+
  labs(title="Mean exponential distribution", x="Mean", y="Count")+
  geom_vline(xintercept = mu, colour="red")+
  geom_vline(xintercept = mean(as.matrix(rms)), colour="green")+
  geom_text(aes(x=mu, label=paste0("Theoretical mean = ", mu), y=10), colour="red")+
  geom_text(aes(x=mean(as.matrix(rms)), label=paste0("Sample mean = ", round(mean(as.matrix(rms)),2)), y

exp.plot <- ggplot(dis.matrix.stacked, aes(x= dis.matrix.stacked))+
  geom_histogram(fill="blue", colour="black", bins = 30)+
  scale_x_continuous()+
  labs(title="Exponential distribution", x="Exponential value", y="Count")

grid.arrange(exp.plot, mean.plot, ncol = 2)
```

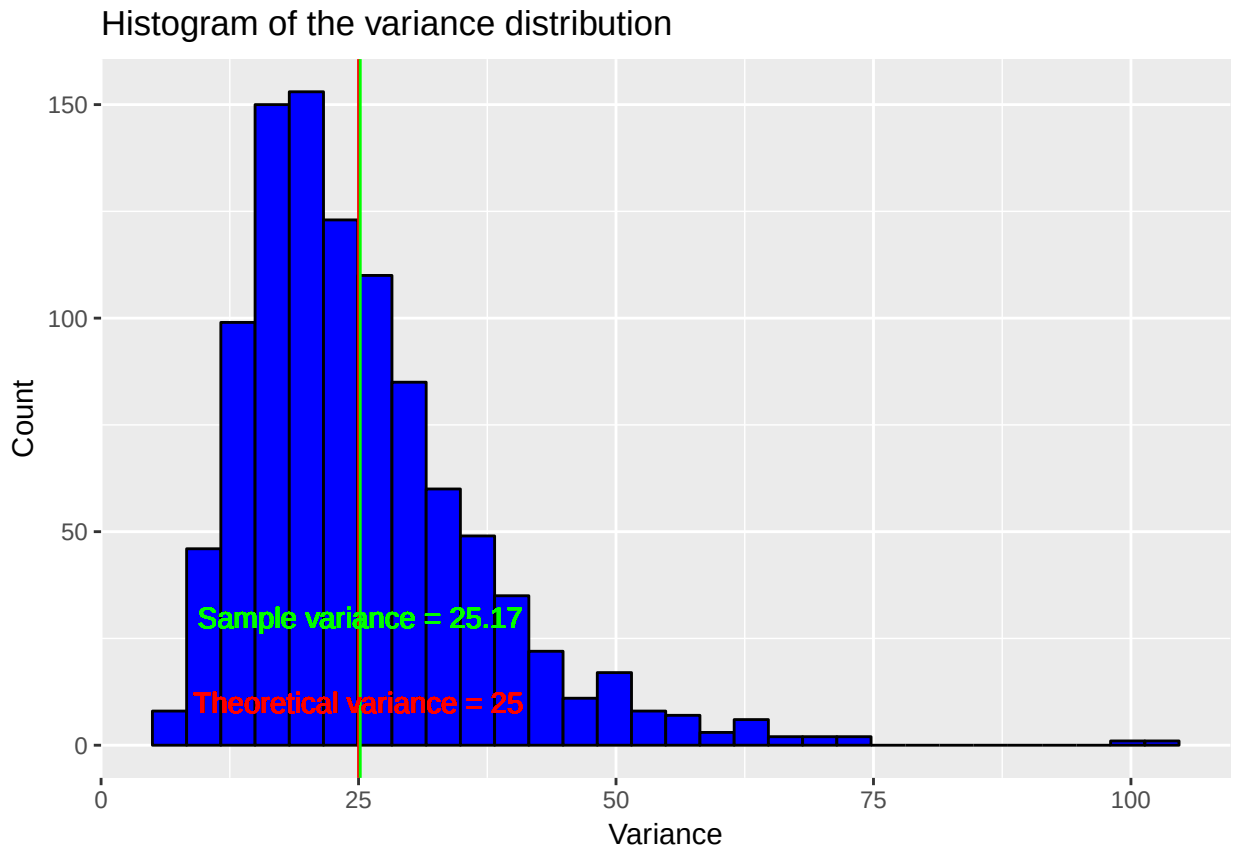


Sample Variance versus Theoretical Variance

The following graph represents the variance of the means of the exponential distribution. Unlike the distributions of the mean, it will not always approximate to another distribution. The difference between the theoretical distribution and the sample distribution is -0.165914.

```
ggplot(sds, aes(x = sds^2))+
  geom_histogram(fill="blue", colour="black", bins=30)+
  scale_x_continuous()+
```

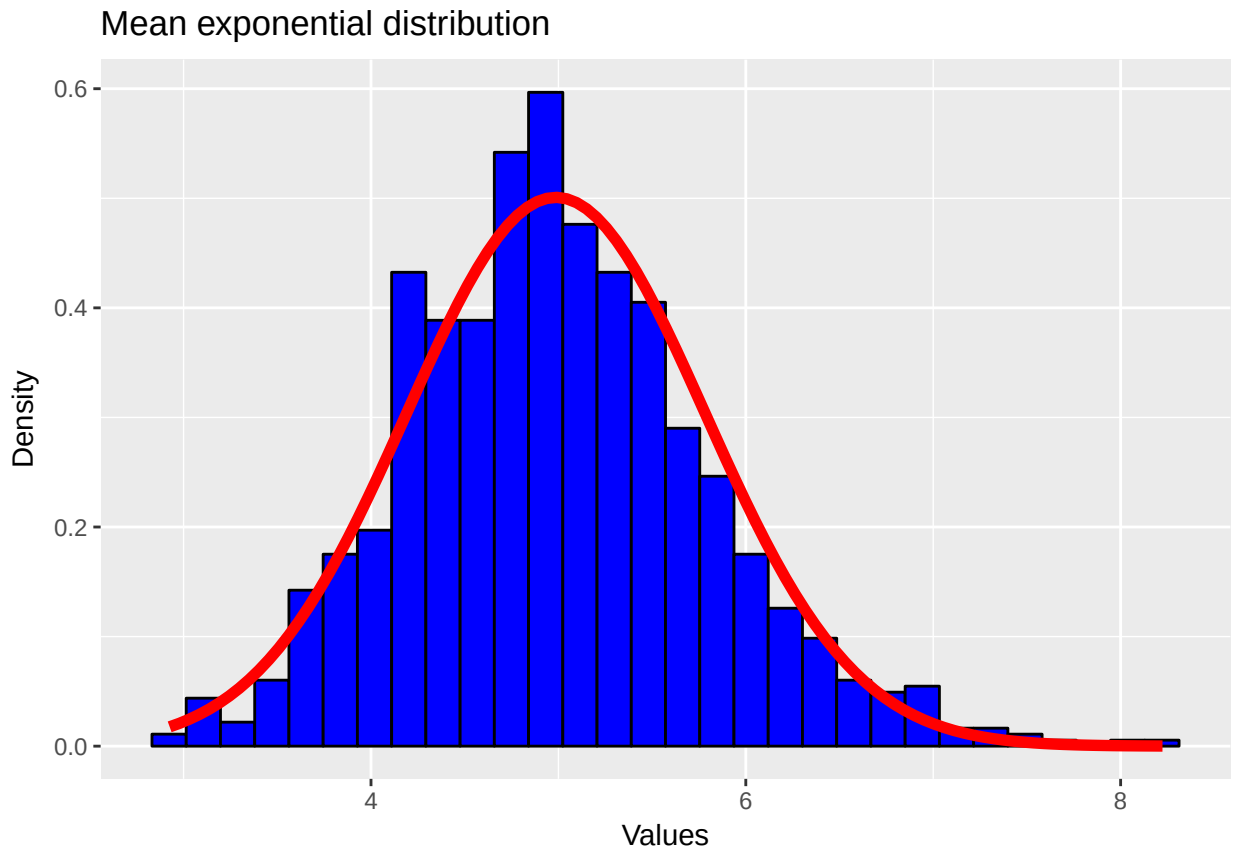
```
  labs(title="Histogram of the variance distribution", x = "Variance", y = "Count")+
  geom_vline(xintercept = vars, colour = "red")+
  geom_vline(xintercept = mean(sds^2), colour="green")+
  geom_text(aes(x = vars, label=paste0("Theoretical variance = ", vars), y=10), colour="red")+
  geom_text(aes(x = mean(sds^2), label=paste0("Sample variance = ", round(mean(sds^2),2)), y=30), colou
```



Distribution

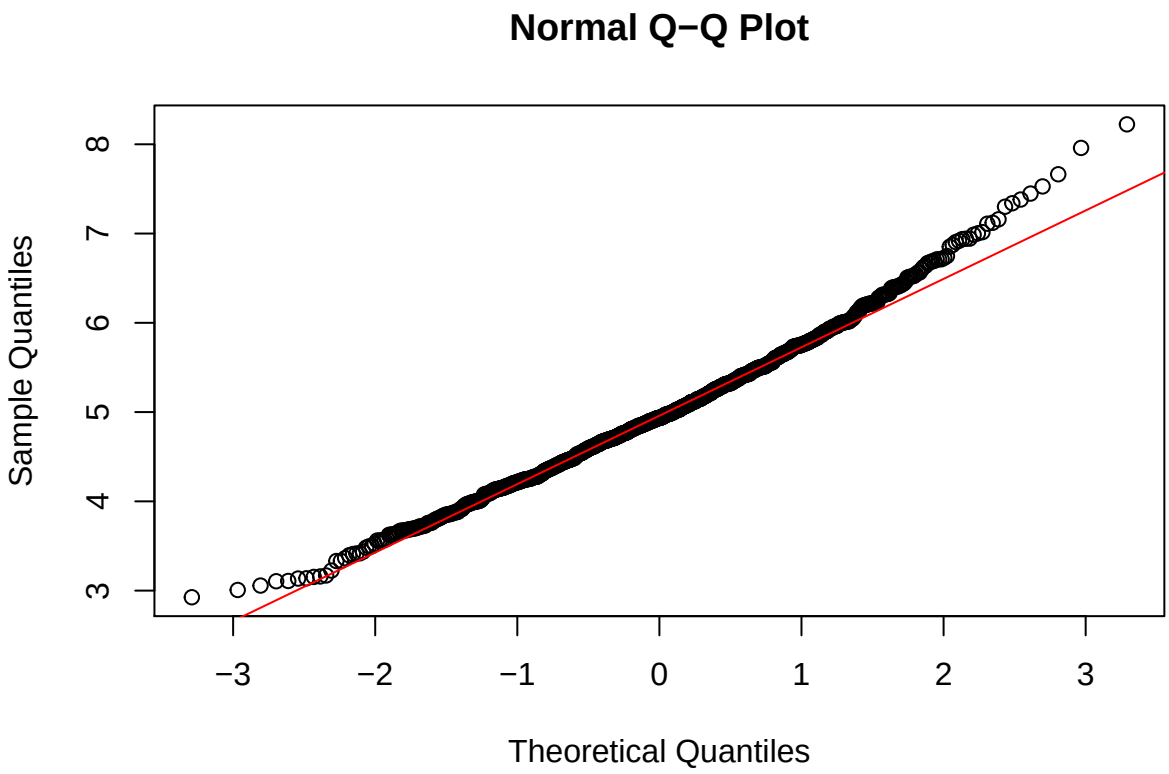
We can see that the distribution appears approximately normally distributed as a result of the Central Limit Theorem. The red graph below represents a normal distribution with the parameters of the simulation results N ~(mean = 4.9865083, sd = 0.7965177).

```
ggplot(rms, aes(x= rms))+
  geom_histogram(aes(y=..density..), fill="blue", colour="black", bins = 30)+
  stat_function(fun = dnorm, args = list(mean = mean(rms$mean), sd = sd(rms$mean)), lwd = 2, col = 'red
  scale_x_continuous()+
  labs(title="Mean exponential distribution", x="Values", y="Density")
```



The Q-Q plot below indicates that the results are approximately distributed. Only the extreme values tend to deviate from the expected normal distribution results.

```
qqnorm(rms$mean)
qqline(rms$mean, col="red")
```



Part 2

Overview

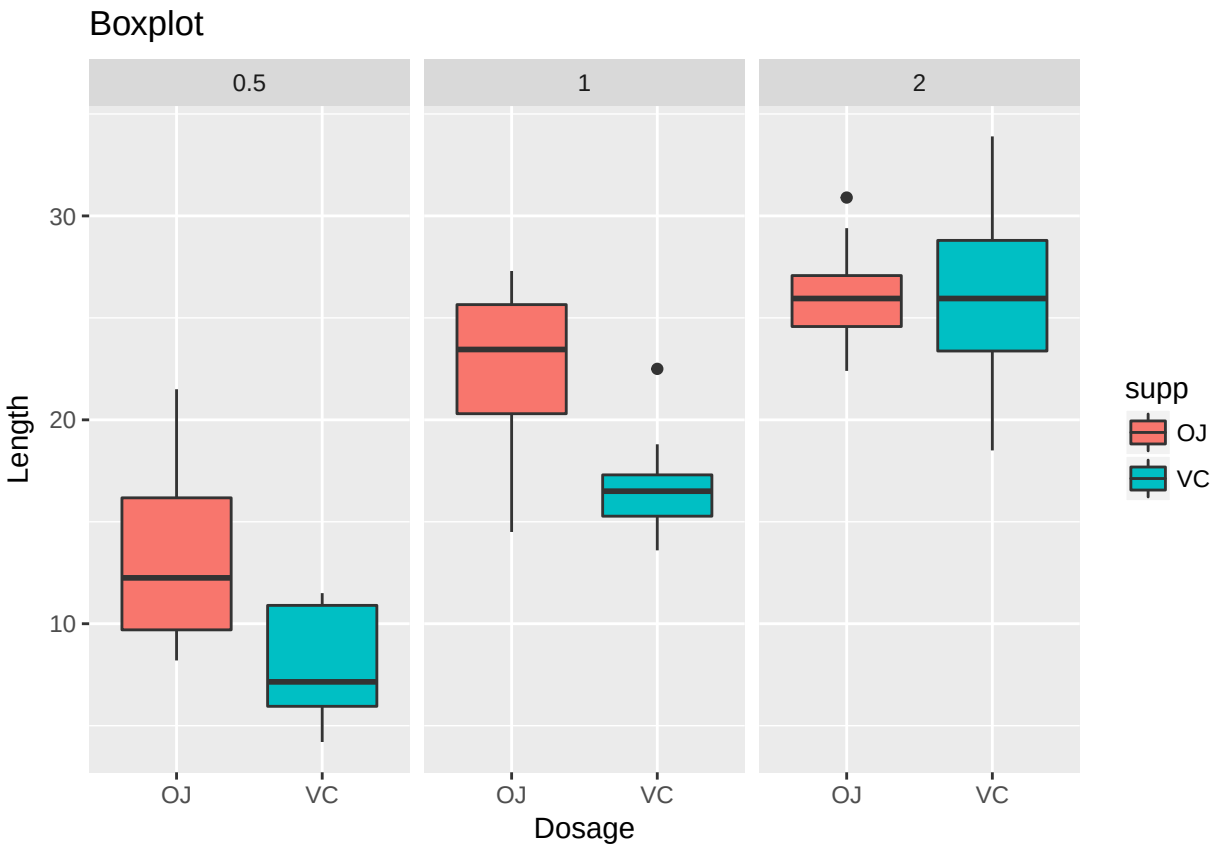
For Part 2 of this assignment, we'll perform an exploratory analysis on the impact that vitamin C has on the length of odontoblasts in guinea pigs via two delivery methods (orange juice and ascorbic acid) at varying dosage levels.

It will be shown that the delivery method does not impact on the length of odontoblast but the dosage level does.

Visual exploration

```
library(datasets)
suppressMessages(library(dplyr))
data("ToothGrowth") #loading the required dataset
ToothGrowth$dose <- as.factor(ToothGrowth$dose) #setting dosage variable to factor

ggplot(ToothGrowth, aes(y = len, x = supp))+
  geom_boxplot(aes(fill = supp))+
  facet_wrap(~ dose)+
  labs(title="Boxplot", x="Dosage", y="Length")
```

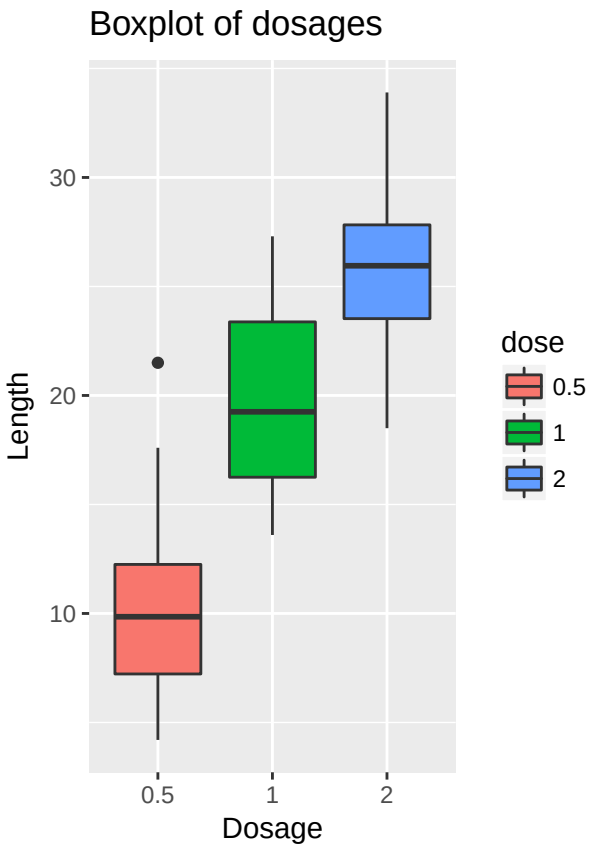


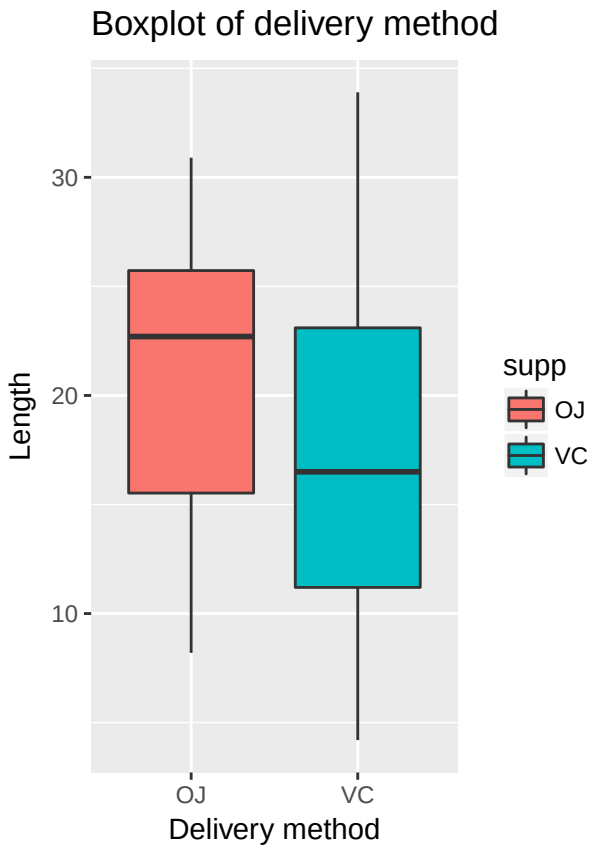
The above graph represents the box-plots of different dosage rates for each delivery method. It appears that there is an increase of legnth in odontoblasts as there is an increase of dosage rates.

```
dosage.graph <- ggplot(ToothGrowth, aes(y = len, x = dose))+
  geom_boxplot(aes(fill=dose))+
  labs(title="Boxplot of dosages", x="Dosage", y="Length")

supp.graph <- ggplot(ToothGrowth, aes(y = len, x = supp))+
  geom_boxplot(aes(fill=supp))+
  labs(title="Boxplot of delivery method", x="Delivery method", y="Length")

grid.arrange(dosage.graph, supp.graph, ncol=2)
```





From the graphs above, it appears that there are distinct groups if you were to seperate them by dosage levels. The seperation does not appear to be as clear between delivery methods.

Numerical analysis

Summary of dose rate:

```
dose.summary <- tapply(ToothGrowth$len, ToothGrowth$dose, summary)
dose.summary
```

```
## $`0.5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.200   7.225   9.850  10.600  12.250  21.500
##
## $`1`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.60   16.25   19.25   19.74   23.38   27.30
##
## $`2`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.50   23.52   25.95   26.10   27.83   33.90
```

Summary of the delivery method:

```
supp.summary <- tapply(ToothGrowth$len, ToothGrowth$supp, summary)
supp.summary
```

```
## $OJ
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.20   15.52   22.70   20.66   25.72   30.90
##
## $VC
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20   11.20   16.50   16.96   23.10   33.90
```

```
supp.tresults <- t.test(data = ToothGrowth, len ~ supp, paired = F, var.equal = F)
supp.tresults
```

```
##
##  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
```

From the t-test above, we can see that there is insufficient evidence to reject the null hypothesis that the means between the two groups are equal as the p-value $= 0.0606345 > 0.05$.

```
dose.tresults <- pairwise.t.test( ToothGrowth$len, ToothGrowth$dose, p.adj = "bonf", paired = F, pool.s
dose.tresults
```

```
##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  ToothGrowth$len and ToothGrowth$dose
##
##   0.5     1
## 1 3.8e-07 -
## 2 1.3e-13 5.7e-05
##
## P value adjustment method: bonferroni
```

From the pairwise t-test for the dosage levels with a bonferroni correction, we can see that the null hypothesis that the means are equal can be rejected with alpha $= 0.05$. Each subgroup is statistically different.

Conclusions and assumptions

We can conclude that dosage rates have a significant impact on the length of odontoblasts for guinea pigs. The following assumptions were made:
* The population are independant
* The populations have unequal variances