Statistical Inference Project

Kevin Tham

April 23, 2018

Simulation Exercise

```
if (!require("pacman")) install.packages("pacman")
pacman::p_load(knitr, dplyr, ggplot2, tidyr, rcompanion)
```

We set the seed to ensure that the following analysis is reproducible

```
set.seed(123)
```

We prepare a thousand repetitions of 40 samples drawn from an exponential distribution.

```
lambda <- 0.2
n <- 40
rep <- 1000
rawdat <- rexp(40000, rate = lambda)
matdat <- matrix(rawdat, rep, n)
```

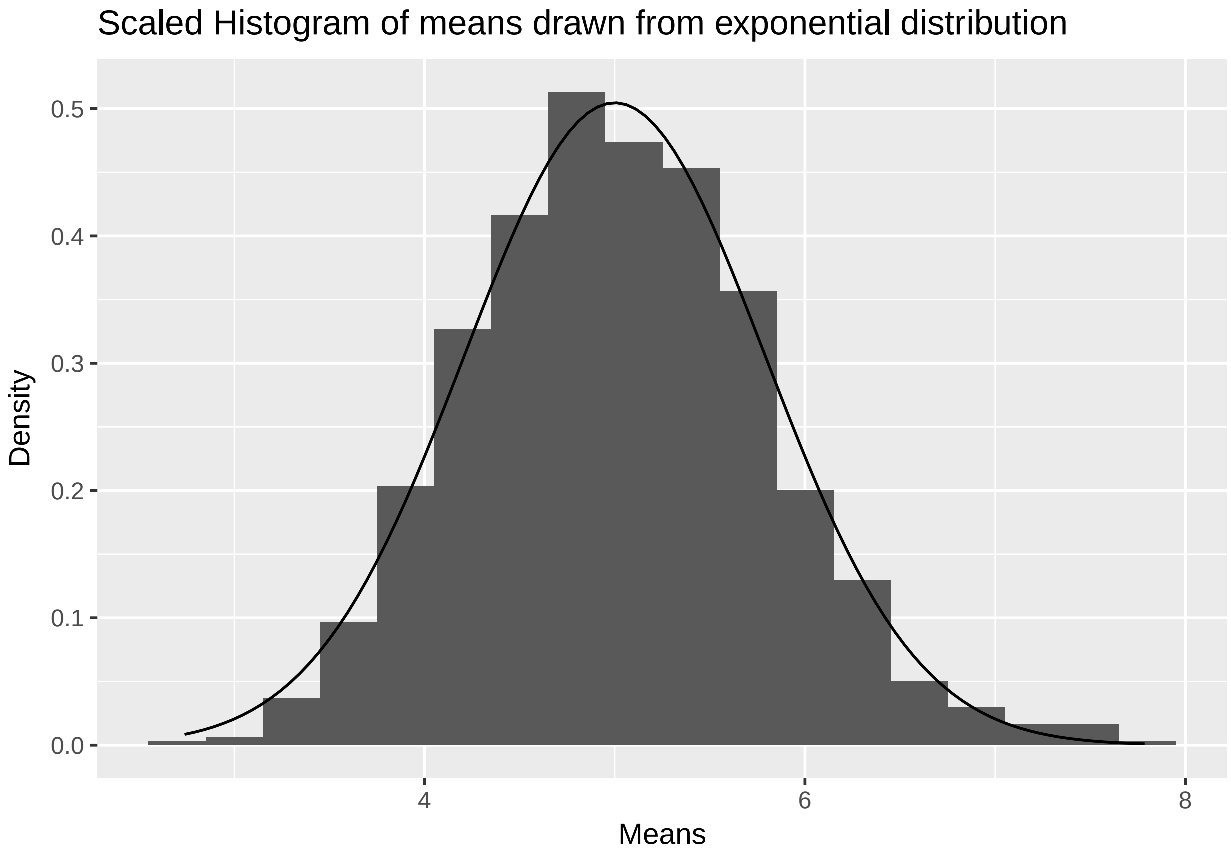
```
th_mean <- 1/lambda
th_var <- (1/lambda)^2/n
th_sd <- sqrt(th_var)
```

From our simulated data we can obtain the sample mean:

```
means <- apply(matdat, 1, mean)
sample_mean <- mean(means)
variance <- var(means)

print(variance)
```

```
## [1] 0.6088292
```

```
bw = 0.3

df_means <- data.frame('means'= means)

ggplot(df_means, aes(means)) + geom_histogram(aes(y=..density..), binwidth=bw) +
  stat_function(fun=function(means) dnorm(means, mean=th_mean, sd=th_sd)) +
  labs(title='Scaled Histogram of means drawn from exponential distribution',
       x = 'Means', y = 'Density')
```



Inferential Data Analysis

In this section, we will analyse the dataset named ‘ToothGrowth’ in the R datasets package. We will undertake a statistical study of the data in order to test the effect of Vitamin C on the tooth growth of guinea pigs.

Exploratory Analysis

First we begin by loading the dataset and calling the `head()` and `str()` functions to give us a brief overview of the data.

```
data(ToothGrowth)
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
str(ToothGrowth)
```

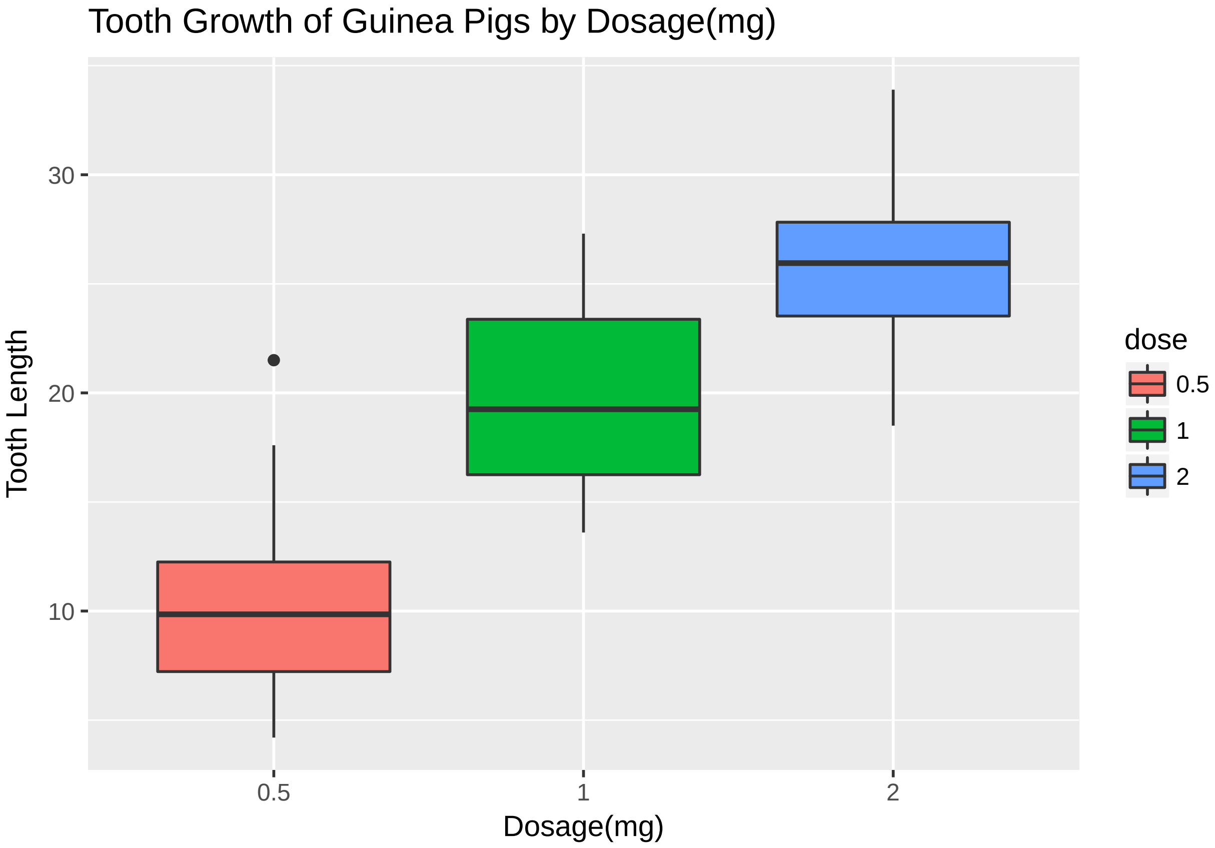
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

As we can see, the dataset contains three variables and 60 observations. The variables correspond to:

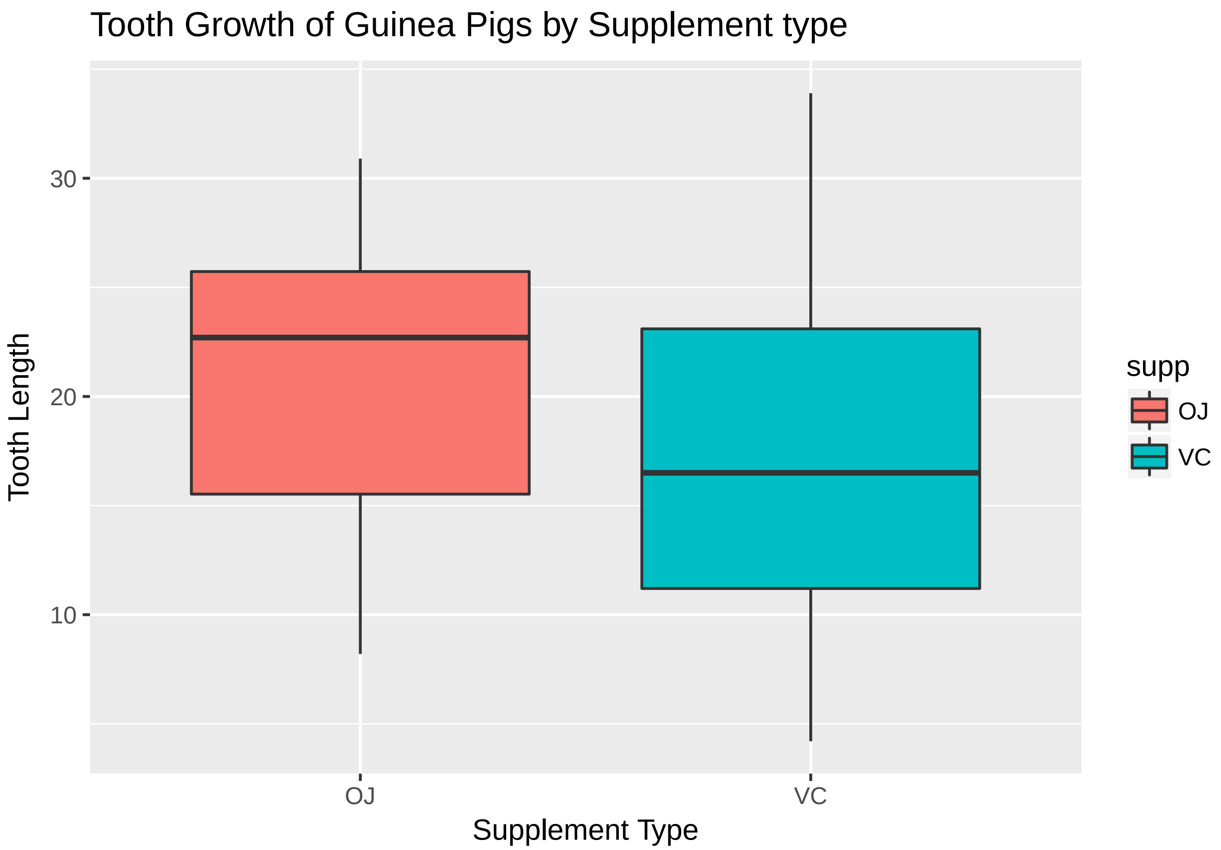
1. `len`: Tooth length
2. `supp`: Supplement type (VC:ascorbic acid, OJ:orange juice)
3. `dose`: Dose in milligrams/day

Since we want to compare numeric variables (`len`) against categorical data (`supp` and `dose`), we can utilise boxplots to display our dataset.

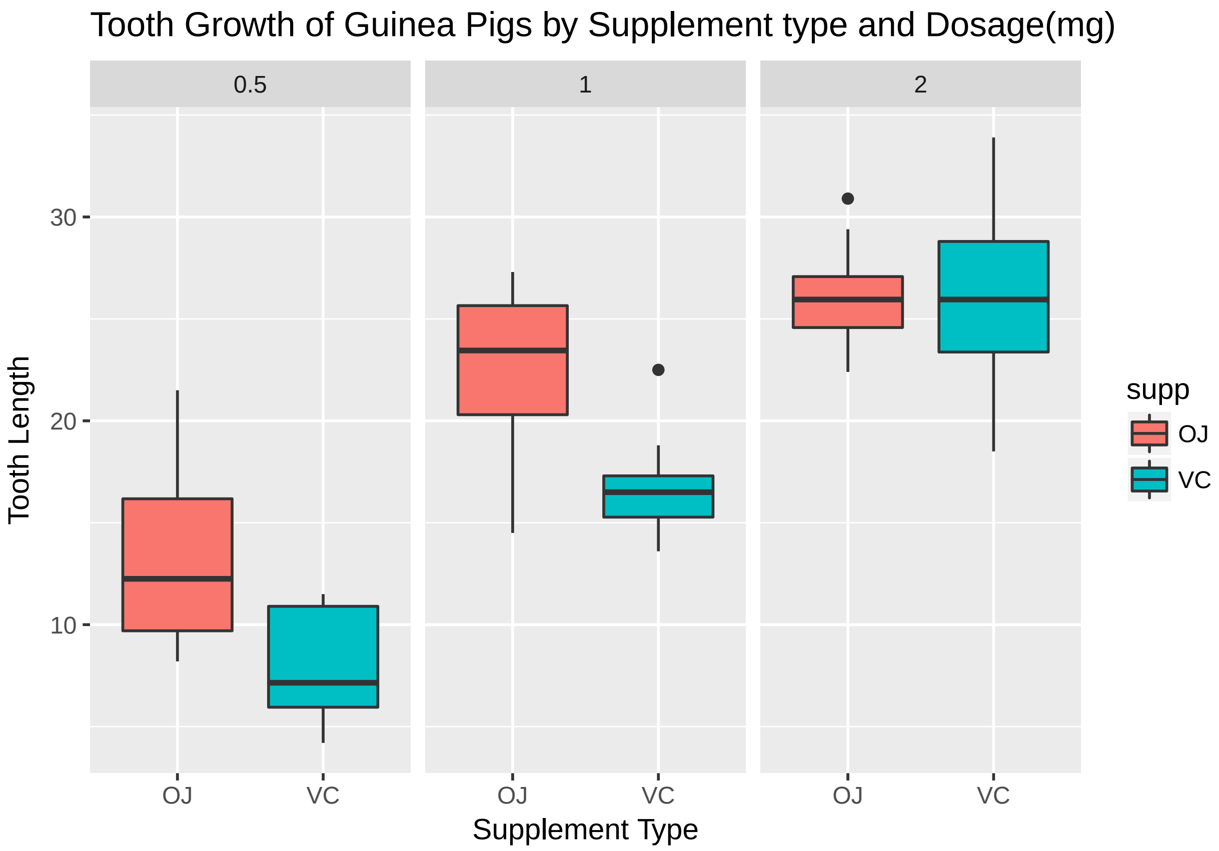
```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
ggplot(ToothGrowth, aes(x=dose,y=len)) +
  geom_boxplot(aes(fill=dose)) +
  labs(title="Tooth Growth of Guinea Pigs by Dosage(mg)",
       x="Dosage(mg)", y="Tooth Length")
```



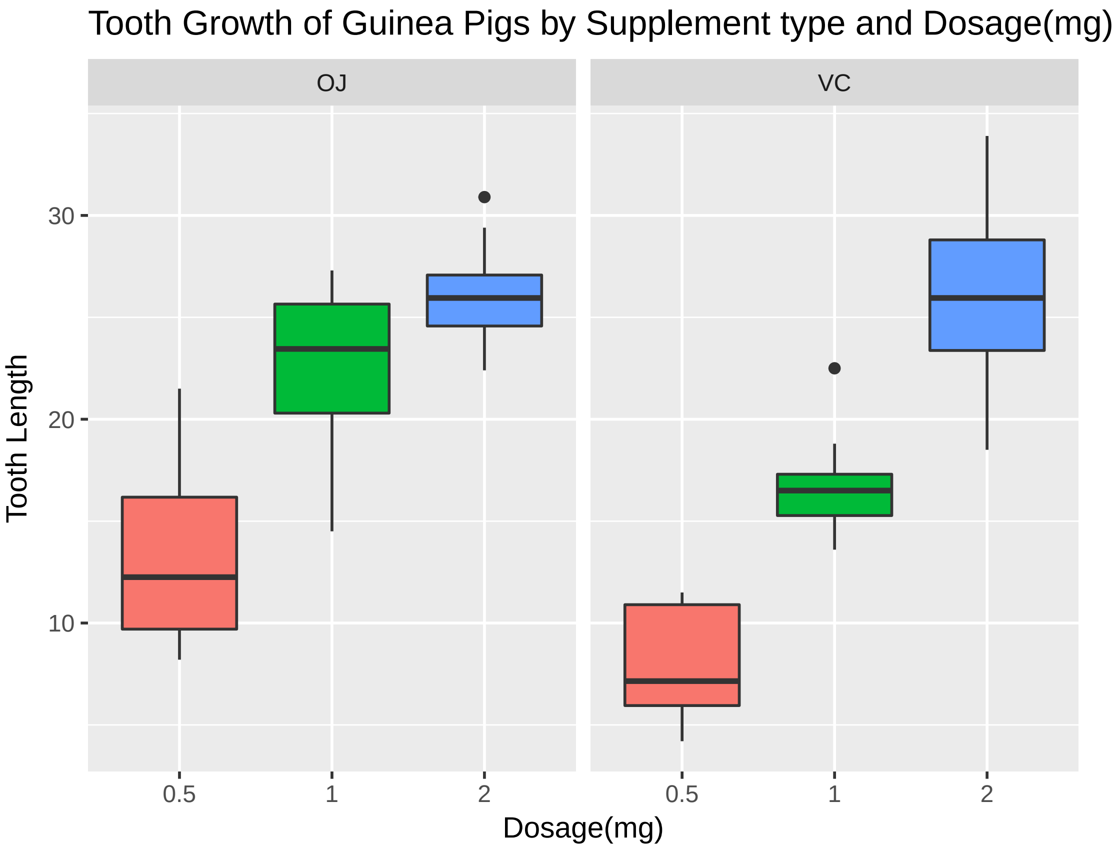
```
ggplot(ToothGrowth, aes(x=supp,y=len)) +
  geom_boxplot(aes(fill=supp)) +
  labs(title="Tooth Growth of Guinea Pigs by Supplement type",
       x="Supplement Type", y="Tooth Length")
```



```
ggplot(ToothGrowth, aes(x=supp,y=len)) +
  geom_boxplot(aes(fill=supp)) +
  facet_grid(.~ dose) +
  labs(title="Tooth Growth of Guinea Pigs by Supplement type and Dosage(mg)",
       x="Supplement Type", y="Tooth Length")
```



```
ggplot(ToothGrowth, aes(x=dose,y=len)) +
  geom_boxplot(aes(fill=dose)) +
  facet_grid(.~ supp) +
  labs(title="Tooth Growth of Guinea Pigs by Supplement type and Dosage(mg)",
       x="Dosage(mg)", y="Tooth Length")
```



The boxplot highlights certain features of the data:

1. Increasing dosage appears to be correlated with increasing tooth length.
2. For 0.5mg and 1.0mg dosage, orange juice appears to result in larger toothlength than ascorbic acid, however for 2.0mg dosage it appears that the medians of the distributions are very similar.

Hypothesis Testing

In order to determine if the supplement type or dosage has an effect on the tooth growth of guinea pigs, and if possible, the magnitude of the effect. Therefore we will utilise hypothesis testing in order to achieve this.

In hypothesis testing, we have a null hypothesis (H_0, representing status quo or independence from variable being tested) and an alternative hypothesis (H_1, representing a positive effect). The aim is employ statistics to decide whether the null hypothesis H_0 should be rejected or not.

This is done by taking the difference in the means of each population being tested (for eg. splitting the data based on supplement type). The difference in means is treated as a random variable and the probability of it being zero (variable has no effect), the p-value, is calculated. If it is below a certain level, which we specify to be 0.05, then we reject the null hypothesis as the probability of the null hypothesis being true is low.

The following assumptions are made:

- Tooth growth follows a normal distribution (may or may not have different means and variance for different variables)
- Within each variable, tooth growth is independent and identically distributed (i.i.d.)