# Statistical Inference Course Project

## Part 1

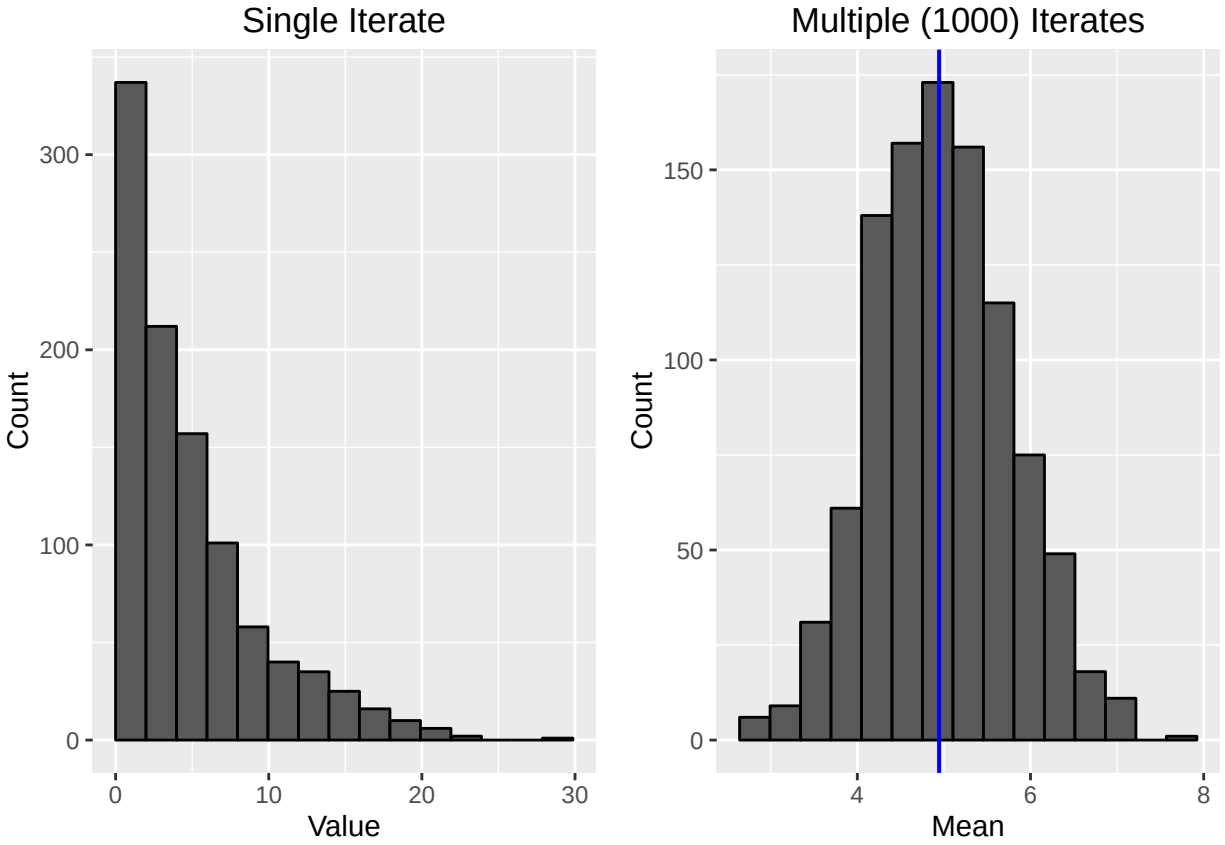
Part 1 of the assignment requires us to explore the understanding of **Central Limit Theorem (CLT)**. Basically, regardless of the underlying distribution of the data, the mean of a large number of iterates (repeated sampling) of **independent and identically distributed (iid)** variables will always be approximately normally distributed.

Here, we will look at a large number of exponentially distributed data (1000) and see if CLT applies to it. The theoretical mean and sd are given in the assignment page.

We show that the theoretical mean and variance are similar to the sample mean and variance.

Sample Mean: 4.98
Theoretical Mean: $1/0.2 = 5$
Sample Variance: 25.04
Theoretical Variance: $(1/0.2)\hat{}2 = 25$

Next, we will plot the distribution to check whether the exponentially distributed data are approximately normally distributed if large number of iterates are ran.




While for single iterate, the plot looks exponentially distributed, the multiple (1000) iterates plot reflects a typical bell-shape indicative of a normal distribution. The blue line marks the median of the multiple iterates.

Thus we can say that the CLT applies for exponentially distributed data.

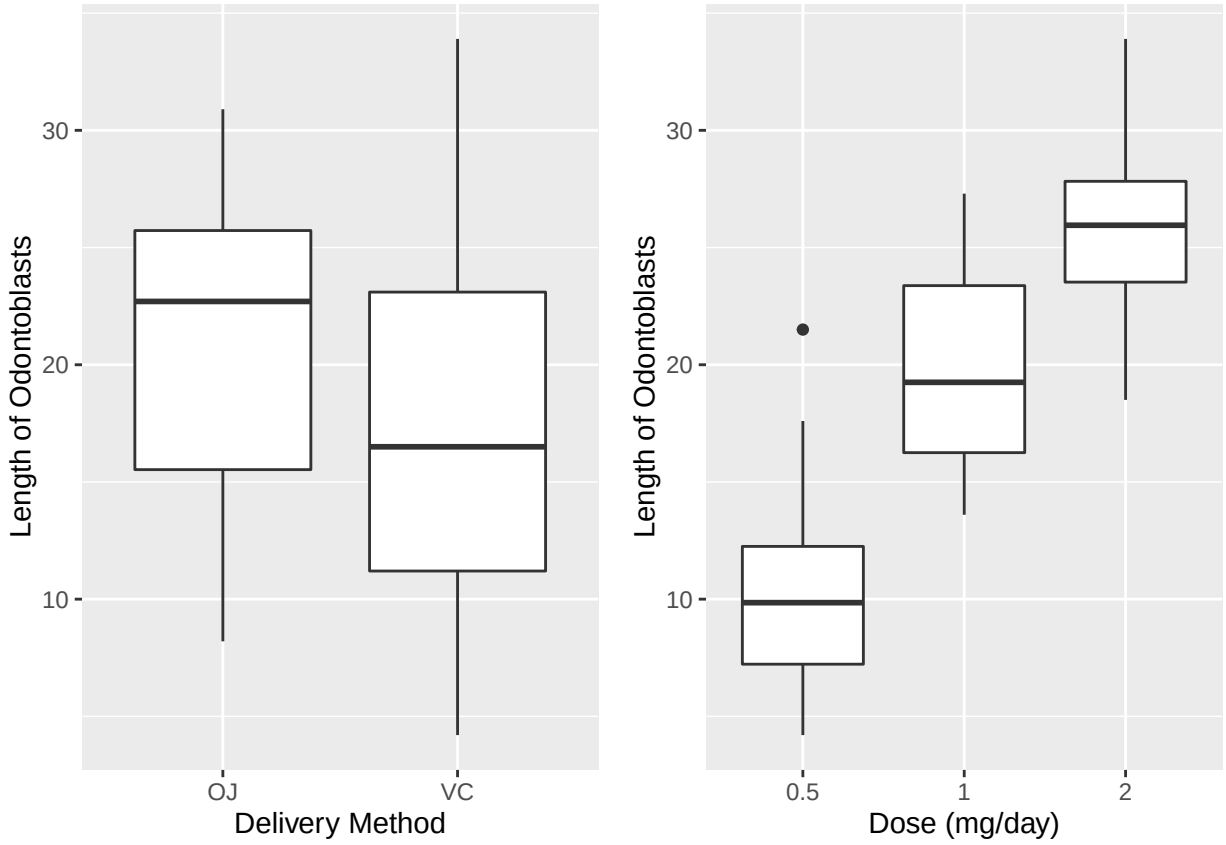
## Part 2

Part 2 of the assignment requires the comparison between the tooth growth by different delivery methods (supp) and dosage level (dose; mg/day). More information about dataset can be found in `?ToothGrowth`.

Briefly, guinea pigs were given different dosage of vitamin C (0.5, 1.0 and 2.0 mg/day) via two different delivery methods (orange juice or ascorbic acid). The length of odontoblasts were then measured in response to the supplement given. Since the response was measured at a single time point, we will be using **Two Sample t-test** to determine if the difference between the dosage level or the delivery method causes a genuine difference in the length of the odontoblasts.

```
##       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
```

```
## '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 ...
```




When we use **Two Sample t-test**, we need to know whether the variances between the two samples are equal. Since **F-test** is not taught in this series of lecture, and Brian Caffo mentioned to assume **unequal**

**variance** when in doubt, the **Two Sample t-test** will be performed under the assumption of **unequal variance**.

For hypothesis testing, our null hypothesis states that there is no difference between the delivery methods and the alternative hypothesis states that there is a true difference between the delivery methods.

$H_0$: $\mu_{oj} = \mu_{vc}$, $H_a$: $\mu_{oj} =/= \mu_{vc}$ where $\mu$ is the difference in length of odontoblasts

```
##
##  Welch Two Sample t-test
##
## data:  oj$len and vc$len
## 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 of x mean of y
##  20.66333  16.96333
```

Note that the p-value $> 0.05$, and the confidence interval passes the 0 value, thus there is no difference in length of odontoblasts by different delivery methods.

Since t = 1.915 $<$ threshold = 2.004 necessary for rejection of null hypothesis, we conclude that there is no difference between the length of odontoblasts by different delivery methods.
The 95% confidence interval ojvc_ci = -0.171, 7.571 also passes the 0 value, thus confirming that there is no difference between the length of odontoblasts by different delivery methods.
We also note the confidence interval manually calculated and the confidence interval calculated via R's `t.test(oj$len, vc$len)` are equal.

We perform the same tests with the dosage levels.

$H_0$: $\mu_{0.5} = \mu_{1.0} = \mu_{2.0}$, $H_a$: $\mu_{0.5} =/= \mu_{1.0} =/= \mu_{2.0}$

```
##                       2.5%      97.5%
## Dose 0.5 v 1.0 -11.983781 -12.833833
## Dose 0.5 v 2.0  -6.276219  -8.996481
## Dose 1.0 v 2.0 -18.156167  -3.733519
```

Since all confidence intervals lie within the negative range, we can conclude that lower dosage of vitamin C, regardless of the delivery method, results in lower length of odontoblasts.

## Appendix (Codes)

```
# To calculate the mean and variance of a exponential distribution
nosim <- 1000
expo <- matrix(rexp(40 * 1000, rate = 0.2), nosim)
expoMean <- apply(expo, 1, mean)

expo_mean <- round(mean(expo), 2)
expo_var <- round(sd(expo)^2, 2)
```

```r
# To plot the single and multiple iterations of exponential distribution
library (ggplot2)
expo2 <- rexp(1000, rate = 0.2)
expodf <- data.frame(expoMean, expo2)

require(gridExtra)
g1 <- ggplot(data = expodf, aes(x = expo2)) +
        geom_histogram(bins = 15, color = "black", boundary = 0) +
        labs(title = "Single Iterate", x = "Value", y = "Count")
g2 <- ggplot(data = expodf, aes(x = expoMean)) +
        geom_histogram(bins = 15, color = "black") +
        geom_vline(xintercept = median(expoMean), color = "blue", size = 0.7) +
        labs(title = "Multiple (1000) Iterates", x = "Mean", y = "Count")
grid.arrange(g1, g2, ncol = 2)
```

```r
# Exploratory data analysis for the tooth growth data set
tg <- ToothGrowth
# summary of the data
summary(tg)
# structure of the data
str(tg)
# though dosage level is numeric, we can treat as a categorial variable
tg$dose <- as.factor(tg$dose)
t1 <- qplot(data = tg, x = supp, y = len,
            geom = "boxplot",
            xlab = "Delivery Method",
            ylab = "Length of Odontoblasts")
t2 <- qplot(data = tg, x = dose, y = len,
            geom = "boxplot",
            xlab = "Dose (mg/day)",
            ylab = "Length of Odontoblasts")
grid.arrange(t1, t2, ncol = 2)
```

```r
# Two-sample t-test for delivery methods
oj <- subset(tg, supp == "OJ")
vc <- subset(tg, supp == "VC")

t.test(oj$len, vc$len)
```

```r
# Manual calculation for the two-sample t-test
# to perform the manual calculations of two sample t-test with unequal variance
# compute the approximate degree of freedom - d
oj_sd <- sd(oj$len)
vc_sd <- sd(vc$len)
oj_n <- length(oj$len)
vc_n <- length(vc$len)
d_num <- (oj_sd^2 / oj_n + vc_sd^2 / vc_n)^2
d_den <- (oj_sd^2 / oj_n)^2 / (oj_n - 1) + (vc_sd^2 / vc_n)^2 / (vc_n - 1)
d <- d_num / d_den

# determine the test statistic threshold
threshold <- round(qt(0.975, d), 3)
```

```r
# compute the test statistic
oj_mean <- mean(oj$len)
vc_mean <- mean(vc$len)
t <- round((oj_mean - vc_mean) / sqrt(oj_sd^2 / oj_n + vc_sd^2 / vc_n), 3)

# here, we also calculate the 95% confidence interval of the difference
# in length of odontoblasts by different delivery methods
ojvc_ci <- round(oj_mean - vc_mean + c(-1, 1) *
                       qt(0.975, d) *
                       sqrt(oj_sd^2 / oj_n + vc_sd^2 / vc_n), 3)
```

```r
# To calculate the confidence intervals for tooth growth with different
# dosages
group1 <- subset(tg, dose == "0.5")
group2 <- subset(tg, dose == "1")
group3 <- subset(tg, dose == "2")
g12 <- t.test(group1$len, group2$len)$conf.int
g13 <- t.test(group1$len, group3$len)$conf.int
g23 <- t.test(group2$len, group3$len)$conf.int
matrix(c(g12, g13, g23),
       ncol = 2,
       dimnames = list(c("Dose 0.5 v 1.0", "Dose 0.5 v 2.0", "Dose 1.0 v 2.0"),
       c("2.5%", "97.5%"))
       )
```