# Part 1: Simulation Exercise

# Overview: This part is going to execute simulations and data analysises to illustrate application of the central limit theorem. R programming will be the major tool to realize the mentioned goal.

## Question 1: Show the sample mean and compare it to the theoretical mean of the distribution.

knitr::opts\_chunk$set(echo = TRUE) lambda <- 0.2 simData <- matrix(rexp(1000\*40, lambda), nrow = 1000, ncol = 40) distMean <- apply(simData, 1, mean) hist(distMean, breaks = 50, main = “The distribution of 1000 averages of 40 random exponentials”, xlab = “Value of means”, ylab = “Frequency of means”, col = “pink”) abline(v = 1/lambda, lty = 1, lwd = 5, col = “red”) legend(“topright”, lty = 1, lwd = 5, col = “red”, legend = “theoretical mean”)

#The simulated sample means are normally distributed with a center very close to the theoretical mean.

## Question 2: Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

distVar <- apply(simData, 1, var) hist(distVar, breaks = 50, main = “The distribution of 1000 variance of 40 random exponentials”, xlab = “Value of variances”, ylab = “Frequency of variance”, col = “light blue”) abline(v = (1/lambda)^2, lty = 1, lwd = 5, col = “blue”) legend(“topright”, lty = 1, lwd = 5, col = “blue”, legend = “theoretical variance”)

#The simulated sample variances are almost normally distributed with a center near the theoretical variance.

## Question 3: Show that the distribution is approximately normal.

par(mfrow = c(3, 1)) hist(simData, breaks = 50, main = “Distribution of exponentials with lambda equals to 0.2”, xlab = “Exponentials”, col = “yellow”) hist(distMean, breaks = 50, main = “The distribution of 1000 averages of 40 random exponentials”, xlab = “Value of means”, ylab = “Frequency of means”, col = “pink”) simNorm <- rnorm(1000, mean = mean(distMean), sd = sd(distMean)) hist(simNorm, breaks = 50, main = “A normal distribution with theoretical mean and sd of the exponentials”, xlab = “Normal variables”, col = “light green”)

### Results

#### 1. The first histogram is the distribution of the exponentials with lambda equals to 0.2. The second histogram is the distribution of 1000 averages of 40 random exponentials

#### 2. The third histogram is a real normal distribution with a mean and standard deviation equals to the second histogram’s.

#### 3. Comparing the first with the second histogram, we can see the distrubution becames normal as the means were taken from each groups.

#### 4. It is a result of the central limit theorem.

#### 5. Comparing the second and the third histogram, we can see the distribution of the means is similar to a real normal distribution with the same mean and standard deviation.

# Part 2:Analyze the ToothGrowth data in the R datasets package

# Overview: In this part we will do some statistical data analysises about the Toothlength data. ###Question 1: Load the ToothGrowth data and perform some basic exploratory data analyses.

library(stats) data(ToothGrowth) library(ggplot2) qplot(dose, len, data = ToothGrowth, color = supp, geom = “point”) + geom\_smooth(method = “lm”) + labs(title = “ToothGrowth”) + labs(x = “Dose of supplements”, y = “Length of teeth”)

### From the plot we can get

#### 1. the length of teeth goes up as the dose of supplements increases, which indicates that the supplements may help teeth growth.

#### 2. At the same dose, OJ seems to incur a higher increase of teeth growth than VC.

#### 3. the slope of OJ is not as steep as the slope of VC, meaning an increase in VC may make a larger increase in teeth length than in OJ.

## Question 2: Provide a basic summary of the data.

### Overally, the spread of the length data is as follows, with a mean of 18.81 and a standard deviation of 7.65.

summary(ToothGrowth)

sd(ToothGrowth$len)

### The summary of the length data while taking OJ as supplement is as follows, with a mean of 20.66 and a standard deviation of 6.61.

summary(ToothGrowth[ToothGrowth$supp == “OJ”, ])

sd(ToothGrowth[ToothGrowth$supp == "OJ", ]$len)

### The summary of the length data while taking VC as supplement is as follows, with a mean of 16.96 and a standard deviation of 8.27.

summary(ToothGrowth[ToothGrowth$supp == “VC”, ])

sd(ToothGrowth[ToothGrowth$supp == "VC", ]$len)

### To see a better spread of the data, we may look at the boxplots under different values of dose and different kinds of supplements.

g <- ggplot(ToothGrowth, aes(dose, len, group = supp)) g + geom\_boxplot() + facet\_grid(supp ~ dose) + labs(x = “Dose of supplements”, y = “Length of teeth growth”)

## Question 3: Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there’s other approaches worth considering)

### If we assume the data is normally distributed, we have a null hypothesis that there is no difference between the mean under each kind of supplements, or each dose of the supplements:

t.test(x = ToothGrowthconf.

### We will be able to construct a confidence interval that 95% of the time, an interval between 16.84 and 20.79 will contain the true mean of the population.

###Then we calculate the mean under each kind of supplements, and each dose of the supplements:

summary(ToothGrowth[ToothGrowth$supp == "OJ", ]$len)[4]

summary(ToothGrowth[ToothGrowth$supp == "VC", ]$len)[4]

### Thus the mean of teeth growth after taking OJ is 20.66; the mean of teeth growth after taking VC is 16.96. Both the two are within the confidence interval. We fail to reject the null hypothesis that there is not a difference in teeth growth after taking the two kinds of supplements.

summary(ToothGrowth[ToothGrowthlen)[4]

summary(ToothGrowth[ToothGrowthlen)[4]

summary(ToothGrowth[ToothGrowthlen)[4]

#### Thus the mean of teeth growth after taking dose of 0.5 is 10.6; the mean of teeth growth after taking dose of 1.0 is 19.74; the mean of teeth growth after taking dose of 2.0 is 26.1. We are able to reject the null hypotheis, and there is a difference in teeth growth between each dose of supplements.

## Now we know the data is not normally distributed under each dose, along with this conclusion, we may assume the data is normally distributed within each dose. Then we will be able to compare the teeth growth between each supplements under each dose

dose05 <- ToothGrowth[ToothGrowthlen, paired = FALSE, conf.level = 0.95)$conf.

mean(dose05[dose05$supp == "VC", ]$len)

mean(dose05[dose05$supp == "OJ", ]$len)

### Thus under the dose of 0.5, there are 95% of the time that a confidence interval between 8.50 and 12.71 will contain the true population mean. We also know that the mean of teeth growth after taking 0.5 dose of VC is 7.98 and the mean of teeth growth after taking 0.5 dose of OJ is 13.23. We rejected the null hypo.

dose10 <- ToothGrowth[ToothGrowthlen, paired = FALSE, conf.level = 0.95)$conf.

mean(dose10[dose10$supp == "VC", ]$len)

mean(dose10[dose10$supp == "OJ", ]$len)

### Thus under the dose of 1.0, there are 95% of the time that a confidence interval between 17.67 and 21.80 will contain the true population mean. We also know that the mean of teeth growth after taking 1.0 dose of VC is 16.77 and the mean of teeth growth after taking 1.0 dose of OJ is 22.7. We rejected the null hypo.

dose20 <- ToothGrowth[ToothGrowthlen, paired = FALSE, conf.level = 0.95)$conf.

mean(dose20[dose20$supp == "VC", ]$len)

mean(dose20[dose20$supp == "OJ", ]$len)

### Thus under the dose of 2.0, there are 95% of the time that a confidence interval between 24.33 and 27.87 will contain the true population mean. We also know that the mean of teeth growth after taking 2.0 dose of VC is 26.14 and the mean of teeth growth after taking 2.0 dose of OJ is 26.06. Both of them are within the confidence interval. We fail to reject the null hypothesis.

## Question 4: State your conclusions and the assumptions needed for your conclusions.

#### The conclusion is when the dose is 0.5 or 1.0 there is a difference between the teeth growth after taking OJ and VC, while when the dose is 2.0, there is no difference between the teeth growth after taking OJ and VC. The assumption needed is we first assumed the whole population is normally distributed, then we assumed the population is normally distributed under each dose.