Statistical Inference Course Project

Adekunle Adeniran

2023-11-07

PART 1 Simulation Exercise

```
## Required libraries
library(ggplot2)
library(kableExtra)
```

Overview

To simulate the distribution of the mean of 40 exponentials, I generated 10,000 samples, each consisting of 40 exponential random variables. The sample mean closely approximated the theoretical mean of the exponential distribution, and the variance of the sample means was close to the theoretical variance calculated as the population variance divided by the sample size. Additionally, a histogram of the sample means showed a bell-shaped, approximately normal distribution, in line with the Central Limit Theorem, confirming the tendency toward normality as the sample size increases.

Sample Exponential Distribution

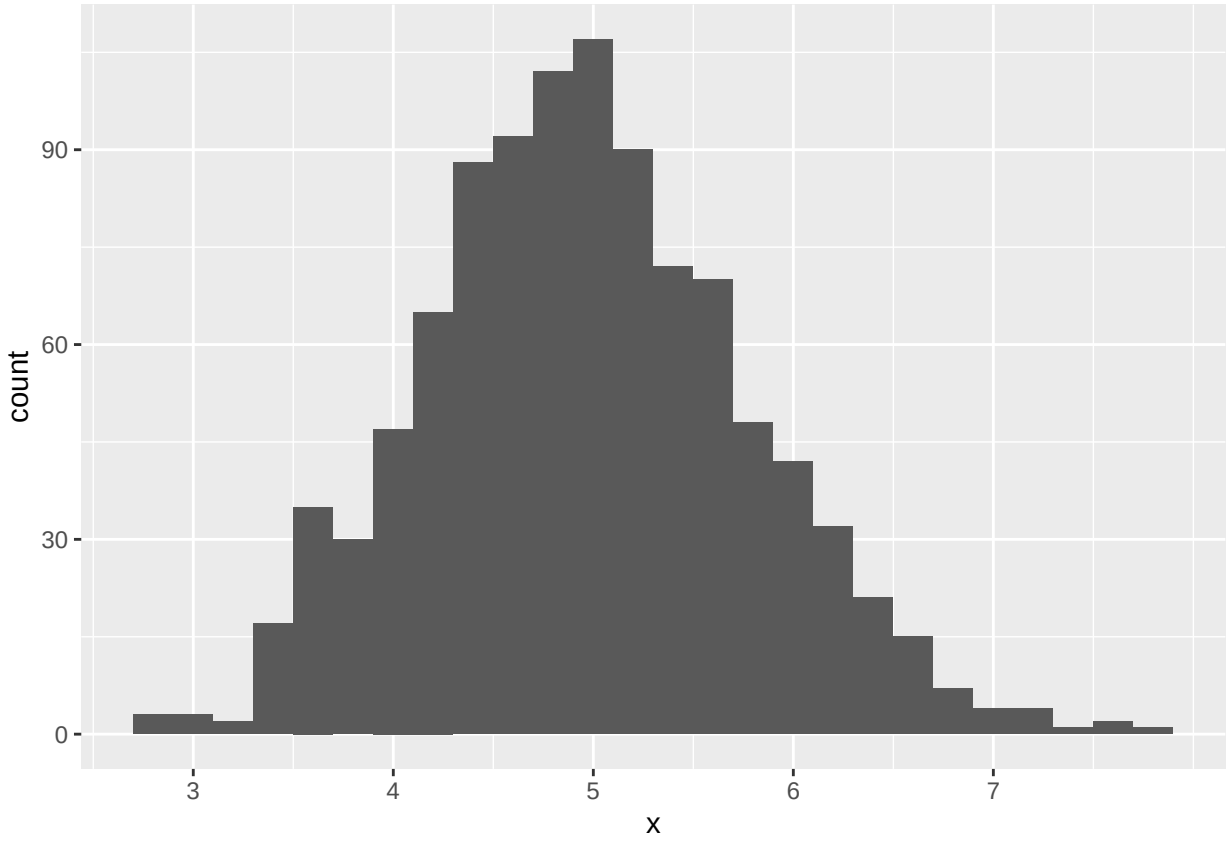
```
## setting seed for reproducibility
set.seed(2023)
lambda <- 0.2 # This is rate parameter
n <- 40 # number of exponentials
noSim <- 1:1000 # number of simulations
```

```
meanexpdist <- data.frame(x = sapply(noSim, function(x) {mean(rexp(n, lambda))}))
head(meanexpdist)
```

```
##          x
## 1 4.265533
## 2 4.269001
## 3 4.805971
## 4 5.010245
## 5 4.374267
## 6 5.625569
```

Plotting the means

```
ggplot(data = meanexpdist, aes(x = x)) +
  geom_histogram(binwidth=0.2) +
  scale_x_continuous(breaks=round(seq(min(meanexpdist$x), max(meanexpdist$x), by=1)))
```



Comparing Expected and Sample Values of Mean, Standard deviation and Variance

```
simumean <- apply(meanexpdist, 2, mean)
```

```
emean <- 1/lambda
```

```
simsd <- sd((meanexpdist$x))
```

```
esd <- (1/lambda)/sqrt(n)
```

```
simvar  <-  var(meanexpdist$x)
```

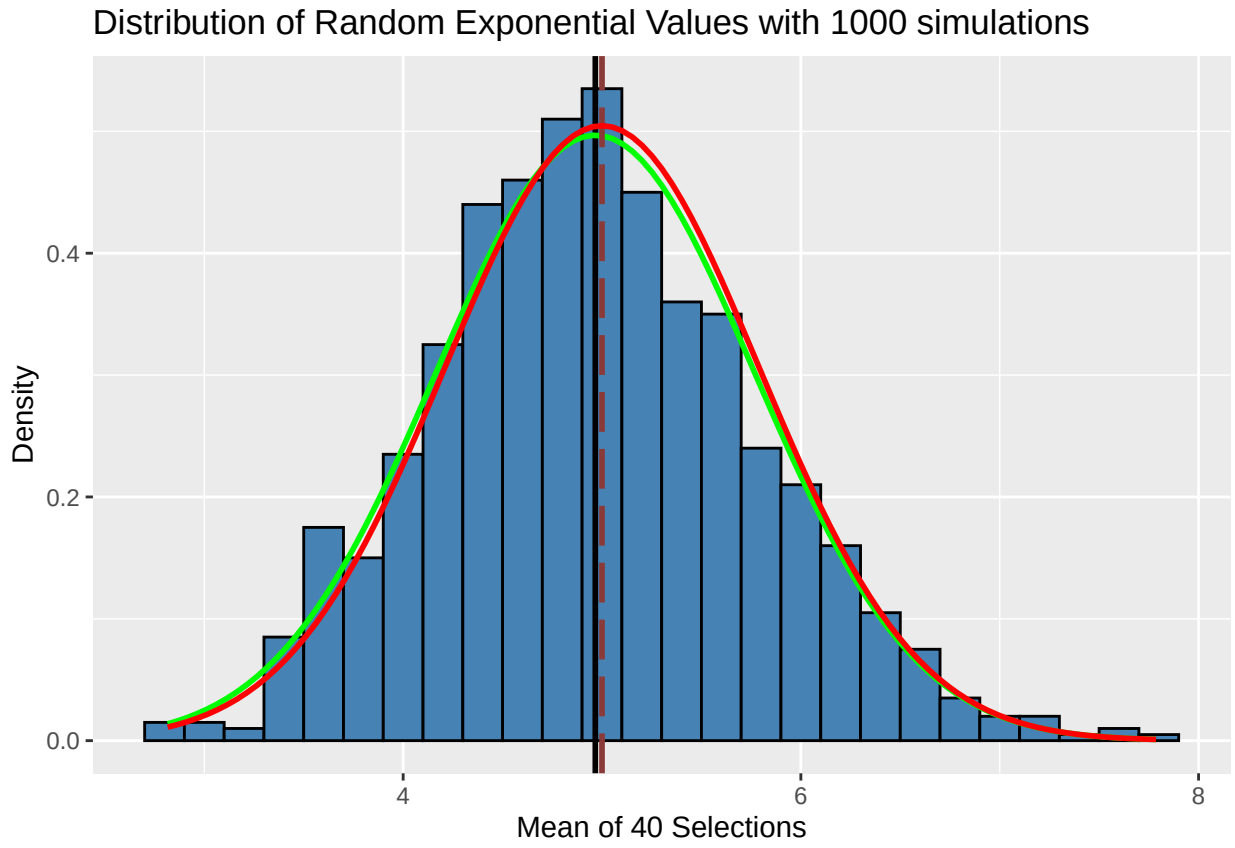
```
evar = esd^2
```

```
Sample <- c(simumean, simsd, simvar)
Expected <- c(emean, esd, evar)
Diff <-
  c(abs(emean - simumean),
    abs(esd - simsd),
    evar - simvar)
names <- c("Mean", "Std", "Variance")
data.frame(Sample,
           Expected,
           Diff,
           row.names =  c("Mean", "Std", "Variance"))
```

```
##            Sample  Expected        Diff
## Mean     4.966410 5.0000000  0.03359007
## Std      0.802972 0.7905694  0.01240255
## Variance 0.644764 0.6250000 -0.01976398
```

Distribution

```
plot <- ggplot(data = meanexpdist, aes(x = x)) +
    geom_histogram(aes(y=after_stat(density)), binwidth = 0.20, fill="steelblue", col="black")
plot <- plot + labs(title="Distribution of Random Exponential Values with 1000 simulations", x="Mean of
plot <- plot + geom_vline(xintercept=simumean,linewidth=1.0, color="black")
plot <- plot + stat_function(fun=dnorm,args=list(mean=simumean, sd=simsd),color = "green", linewidth=1.
plot <- plot+ geom_vline(xintercept=emean,linewidth=1.0,color="indianred4",linetype = "longdash")
plot <- plot + stat_function(fun=dnorm,args=list(mean=emean, sd=esd),color = "red", linewidth=1.0)
plot
```



Conclusion

As depicted in the graph, the distribution of means derived from randomly sampled exponential distributions shows a significant alignment with the normal distribution, particularly in relation to the anticipated values determined by the provided lambda.

```
##
## ...
```

PART 2 Basic Inferential Data Analysis

Loading and take a peek at the dataset

```
library(stats)
data("ToothGrowth")
head(ToothGrowth)
```

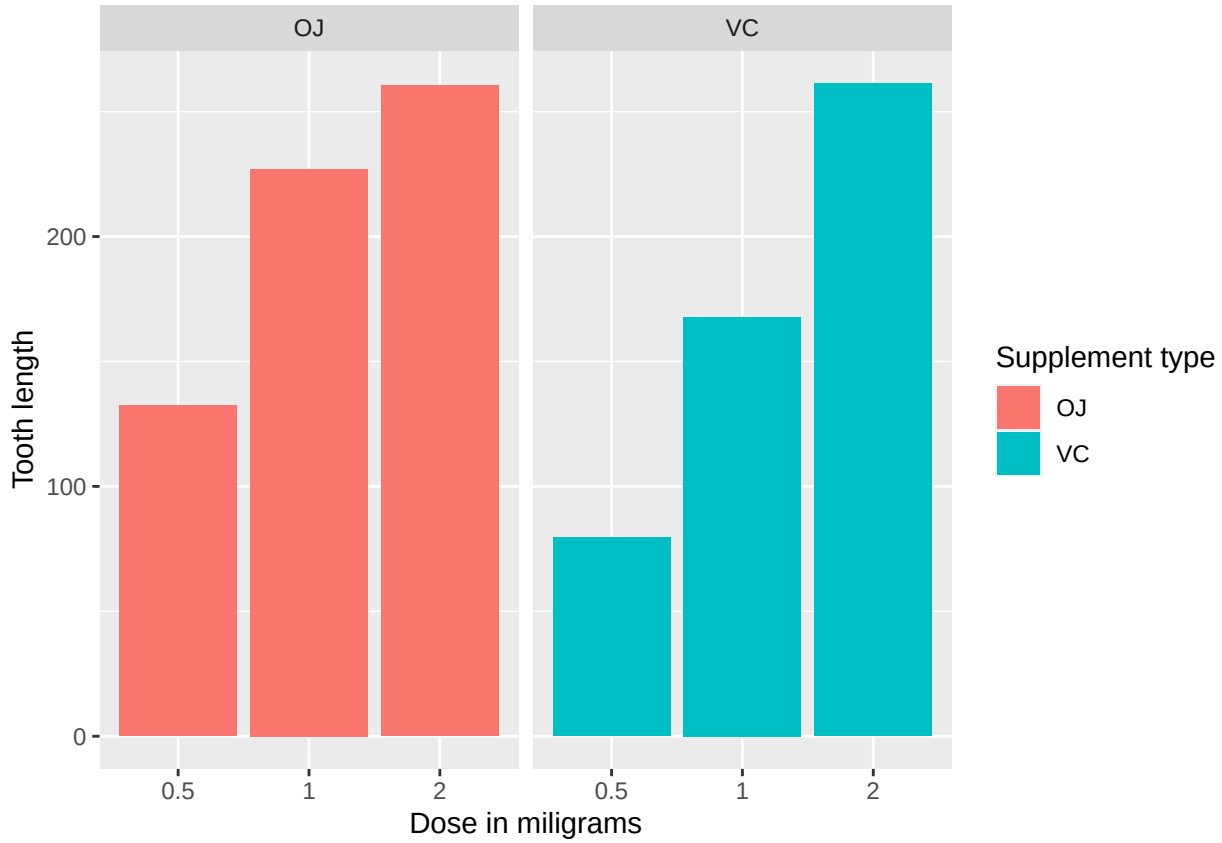
```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

Summary of Data

```
summary(ToothGrowth)
```

```
##       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
```

```
library(ggplot2)
ggplot(data=ToothGrowth, aes(x=as.factor(dose), y=len, fill=supp)) +
    geom_bar(stat="identity",) +
    facet_grid(. ~ supp) +
    xlab("Dose in miligrams") +
    ylab("Tooth length") +
    guides(fill=guide_legend(title="Supplement type"))
```



Using Confidence Intervals to compare tooth growth by supp and dose.

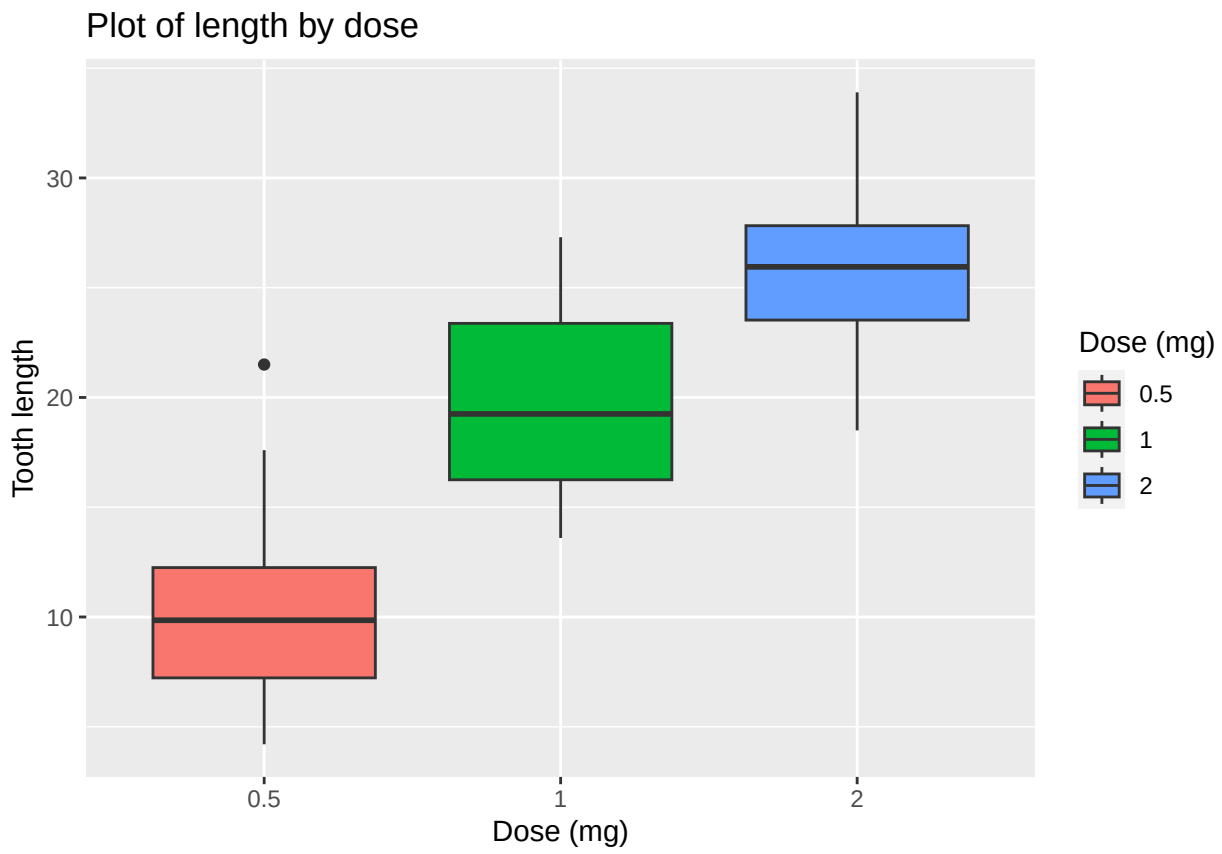
```
unique(ToothGrowth$dose)
```

```
## [1] 0.5 1.0 2.0
```

There are 3 dose groups: 0.5, 1 and 2

The Graph below shows the relationship between Tooth length and Dose

```
ggplot(aes(x=factor(dose), y=len), data=ToothGrowth) + geom_boxplot(aes(fill= factor(dose))) + ggtitle(
```



T-test for dose 0.5 mg:

```
t.test(len ~ supp, ToothGrowth[ToothGrowth$dose == .5, ])
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

T-test for dose 1 mg:

```
t.test(len ~ supp, ToothGrowth[ToothGrowth$dose == 1, ])
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
```

```
## mean in group OJ mean in group VC
##            22.70            16.77
```

T-test for dose 2 mg:

```
t.test(len ~ supp, ToothGrowth[ToothGrowth$dose == 2, ])
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

```
dose <- c(0.5, 1.0, 2.0)
p_value <- c(0.0064, 0.0010, 0.9639)
conf.int <-c("1.72, 8.78","2.80, 9.06","-3.80, 3.64")
decision <- c("Reject null", "Reject null", "Do not reject null")

knitr::kable(data.frame(dose, conf.int, p_value, decision), align = "cccc")
```

dose	conf.int	p_value	decision
0.5	1.72, 8.78	0.0064	Reject null
1.0	2.80, 9.06	0.0010	Reject null
2.0	-3.80, 3.64	0.9639	Do not reject null

As anticipated, the p-values for doses 0.5 and 1.0 are expected to be very low due to the substantial mean differences between them.

Consequently, for doses 0.5 and 1.0, since the p-values fall below 0.5, we can reject the null hypotheses asserting that there is no difference in tooth growth among the supplement types. However, for dose 2.0 mg/day, the null hypothesis can be retained as the p-value exceeds 0.5.

Conclusion

The fundamental assumption underlying the results is that the sample is a representative depiction of the population, and the variables are independent and identically distributed (IID) random variables.

Regarding the t-test, two key assumptions are taken into account:

1. The data is not paired, indicating independence.
2. The variances are unequal.

Given these considerations, upon reviewing the t-test results, it is observed that supplement type OC proves to be more effective than VC for doses below 1.0. However, at a dose of 2.0 mg/day, there is no discernible difference between the supplement types.