Statistical Inference

Y. Zheng

Thursday, October 13th, 2016

Verifying Central Limit Theorem for Exponential Distribution

Overview

We wish to investigate the validity of Central Limit Theorem: *The arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined (finite) expected value and finite variance, will be approximately normally distributed*

1000 simulations of **40** *exponential distributed random variables* were conducted and the eventual distribution was compared against the *normal distribution*.

Density plots of simulation results were compared against theoretical predicted probability density functions.

The following properties of the exponential distribution is assumed:

- The pdf, f(x) = $\lambda e^{(-\lambda x)}$
- Mean = $\frac{1}{\lambda}$
- Standard Deviation(sd) = $\frac{1}{\lambda}$

For this project specifically: $\lambda = 0.2$

```
lambda <- 0.2
```

Simulations

1000 simulations were functionally iterated and the means of 40 random samples of exponentially distributed random variables were caquired. The final result is stored in a data.frame called means with a single column labelled x.

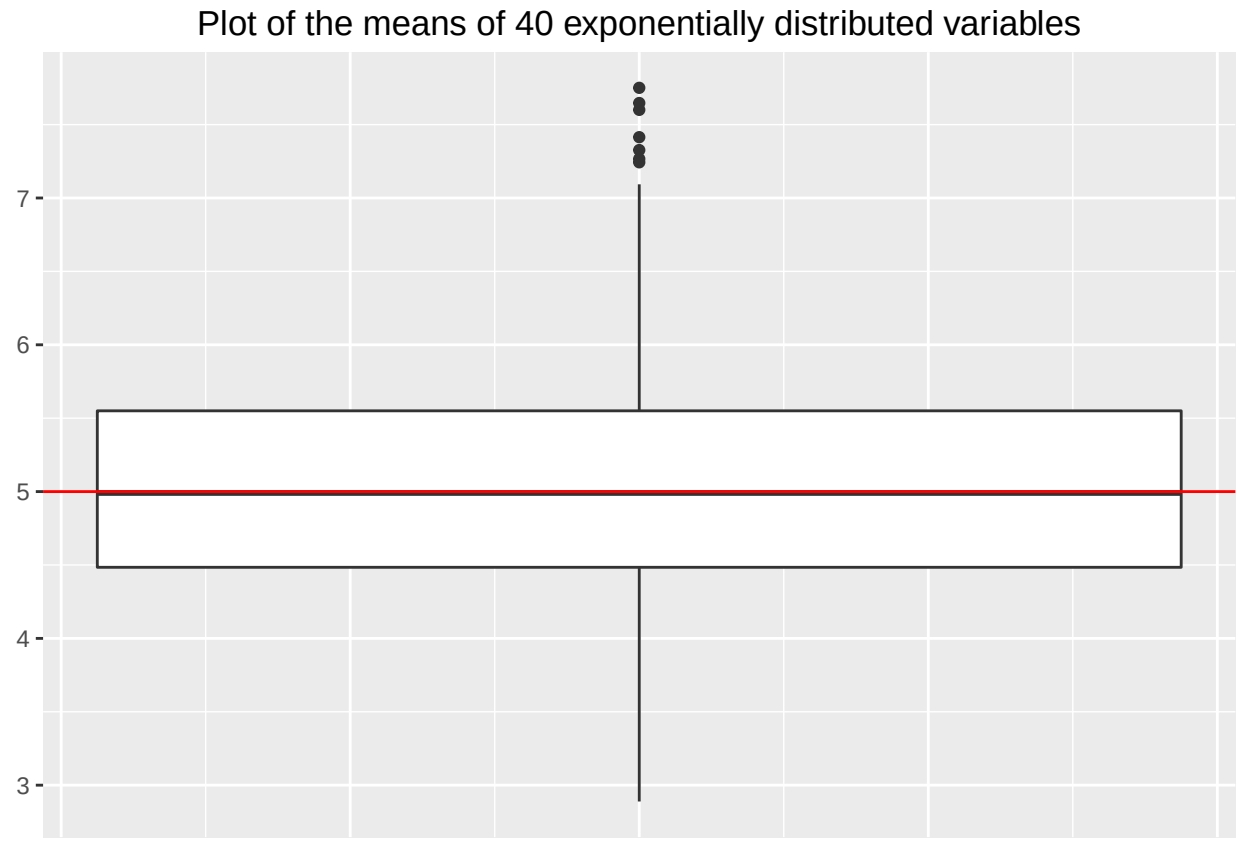
In the spirit of reproducibility, the random seed provided corresponds to the day this investigation was conducted - 1310 (13th of October).

```
set.seed(1310)

means <- data.frame(
    x = sapply(1:1000, function(x){
      mean(rexp(40, 0.2))
    })
)

head(means)
```

```
##          x
## 1 4.610897
## 2 4.510691
## 3 3.505286
## 4 6.108270
## 5 5.470979
## 6 6.402285
```

Sample Mean

```
# Useful plotting library loaded
library(ggplot2)

# Boxplot of the means of 40 exponentially distributed variables in 1000 simulations
qplot(x= 1, y= means$x, geom = "boxplot") + geom_hline(aes(yintercept = 5), color = "red") +
    theme(axis.title.x=element_blank(), axis.text.x=element_blank(),
          axis.ticks.x=element_blank(), axis.title.y=element_blank()) +
    labs(title = "Plot of the means of 40 exponentially distributed variables")
```



Sample Mean calculation:

```
mean(means$x)
```

```
## [1] 5.035438
```

The **red line** corresponds to the theoretical mean: $\frac{1}{\lambda} = \mathbf{5}$ From the plot above, we can see the sample mean (Given by the thick black line) roughly corresponds to the theoretical mean. This suggests evidence for the **Law of Large Number (LLN)**.

Sample Variance

The sample variance is evaluated in the below code.

```r
var(means$x)
```

```
## [1] 0.6216815
```

The theoretical variance for a *exponentially distributed random variable*, $\sigma^2 = \frac{1}{\lambda^2}$.
We therefore expect the variance of the mean of 40 such variables, ${\sigma_{40}}^2 = \frac{\sigma^2}{40} = \frac{1}{40\lambda^2}$

```r
1/(40*lambda^2)
```

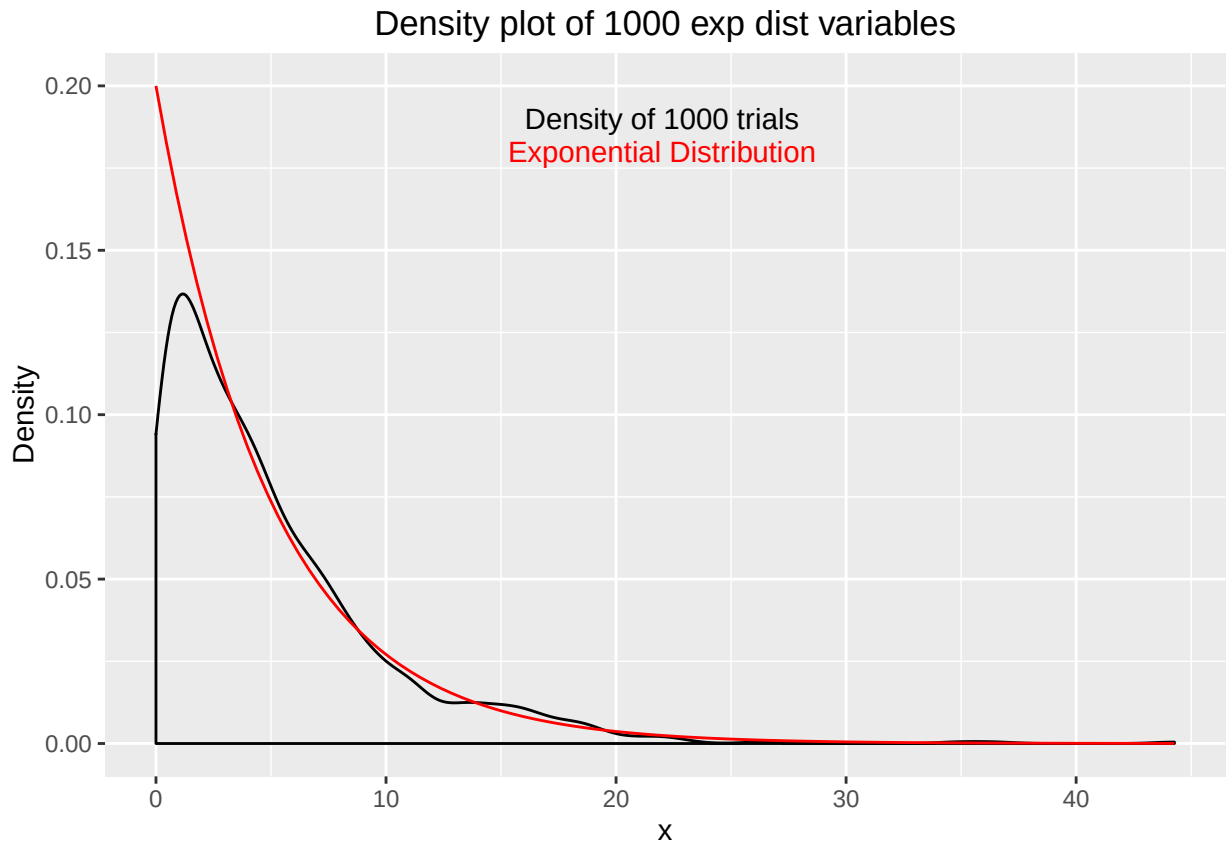
```
## [1] 0.625
```

We are therefore able to see the correspondence between theoretical and sample variance.
However, they do not correspond exactly. The disparity may be explained by the variability of the samples we are taking. Since we are taking a sample of 40 random variables, there exists a non-zero probability of selecting variables such that the mean of 40 samples does not correspond to the exact mean. Pedantically speaking, it is unlikely we are able to recover the exact variance and disparities are well within expectations.

Distribution

We will now show that the means of the 40 exponentially distributed random variables does indeed correspond to the normal distribution. The graphical method is preferred as it takes into account higher order variability such as *skew* and *ketosis*.

```r
# Plot of density function of 1000 exponential distributed random variables
# compared to theoretical probability density function
ggplot(data.frame(x = rexp(1000, 0.2)), aes(x)) + geom_density() +
    labs(y = "Density", title = "Density plot of 1000 exp dist variables") +
    stat_function(fun = dexp, colour = "red", args = list(rate = 0.2)) +
    annotate("text", x = 22, y = 0.18, label = "Exponential Distribution", color = "red") +
    annotate("text", x = 22, y = 0.19, label = "Density of 1000 trials")
```

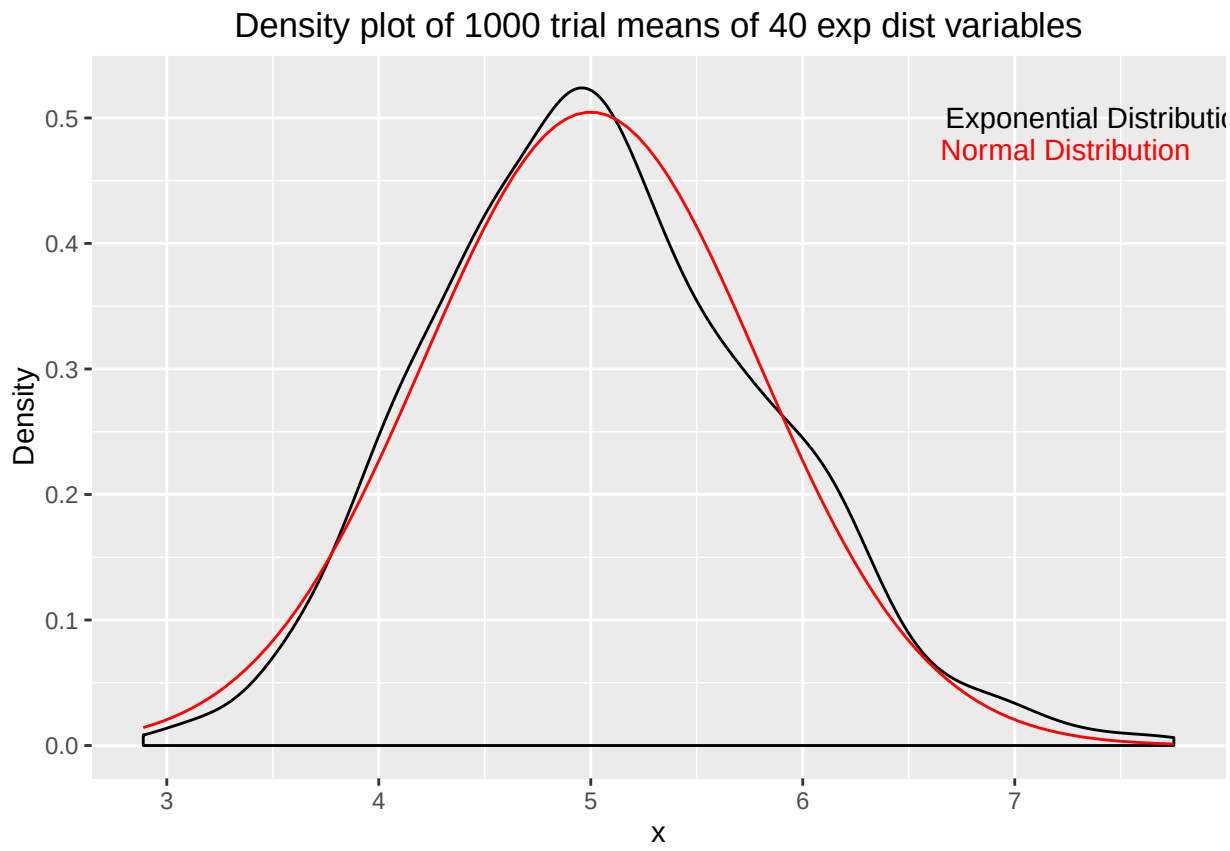


The plot above is used to verify 2 properties we are interested in:

1. The density of 1000 random variables does correspond to the exponential distribution.
2. It does not look anything like the normal distribution.

Property 2 is particularly useful as it depicts the power of **Central Limit Theorem (CLT)**. Despite how drastically different the exponential distribution is from the normal distribution, the means of 40 exponentially distributed random variables still resembles the normal distribution (as we will soon see in the plot below).

```
# Plot of density function of 1000 trial of the means of 40 exponential distributed random variables
# compared to predicted normal distribution pdf according to Central Limit Theorem
ggplot(means, aes(x)) + geom_density() +
    labs(y = "Density", title = "Density plot of 1000 trial means of 40 exp dist variables") +
    stat_function(fun = dnorm, colour = "red", args = list(mean = 5, sd = 5/sqrt(40))) +
    annotate("text", x = 7.24, y = 0.475, label = "Normal Distribution", color = "red") +
    annotate("text", x = 7.4, y = 0.5, label = "Exponential Distribution")
```



The means of a conservative sample size of 40 have already produced such striking resemblence to the normal distribution. We can therefore see by example that CLT works.

Tooth Growth Analysis

Loading Data

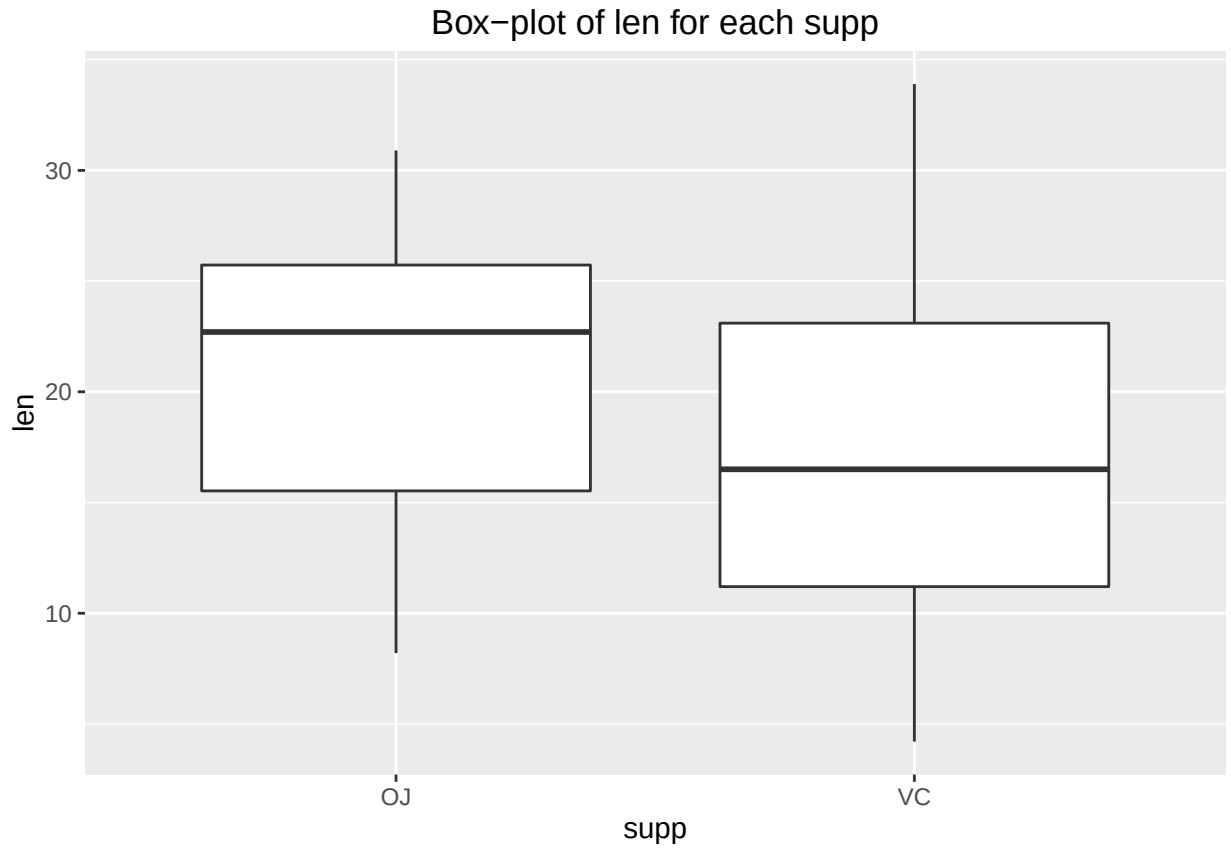
```
library(datasets)
data(ToothGrowth)

# Quick look at the ToothGrowth dataframe
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 ...
```

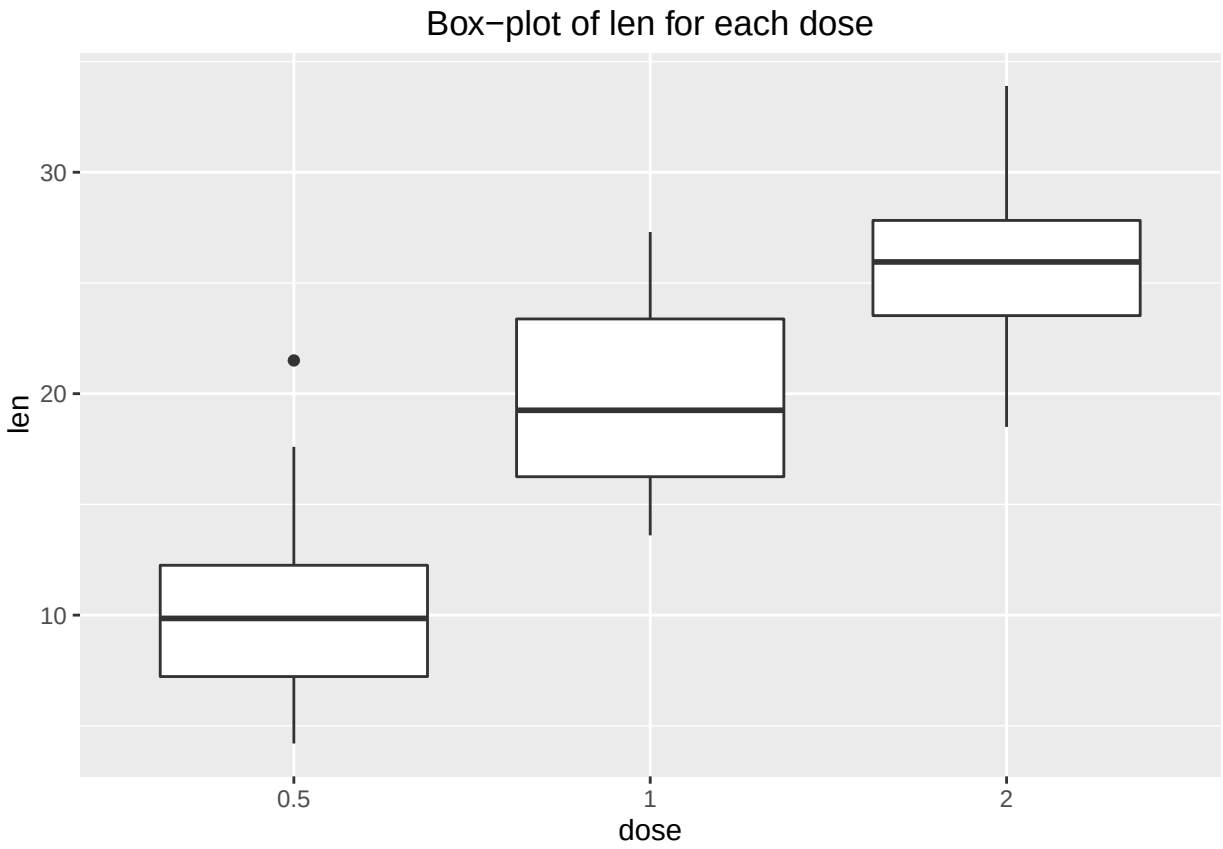
Summary of the data

We investigate the relationships between len and each supp as well as len and each dose via exploratory data analysis.

```r
# Look at how len varies for each supp
qplot(supp, len, data = ToothGrowth, geom = "boxplot") + labs(title = "Box-plot of len for each supp")
```




```r
# Look at how len varies for each dose
ToothGrowth$dose = factor(ToothGrowth$dose)
qplot(dose, len, data = ToothGrowth, geom = "boxplot") + labs(title = "Box-plot of len for each dose")
```



From exploratory data analysis, we expect that OJ is the more effective supplement for tooth growth. We also expect higher dosage to correspond to increased tooth growth (len).

Hypothesis Testing

We will divide our test into 2 portions: one for supplement and one for dosage.

Supplement Analysis

Definitions:

- μ_{OJ} = population mean of len using OJ supplement
- μ_{VC} = population mean of len using VC supplement

From our exploratory plots, we make the following hypothesis:
$H_0 : \mu_{OJ} - \mu_{VC} \leq 0$
$H_1 : \mu_{OJ} > \mu_{VC}$

```
t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data = ToothGrowth, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
```

```
## t = 1.9153, df = 55.309, p-value = 0.03032
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4682687       Inf
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

p-value = 0.03032
95% confidence interval = [0.4682687, ∞]

The 95% confidence interval is entirely positive.
The p-value perspective tells us that the ¶$(Observation|H_0) = 0.03032 < 0.05$.
We have sufficient evidence to reject H_0 in preference for $H_1 : \mu_{OJ} > \mu_{VC}$.

Dosage Analysis

Definitions:

- $\mu_{0.5}$ = population mean of len with 0.5 dose
- μ_1 = population mean of len with 1 dose
- μ_2 = population mean of len with 2 dose

Similarly, we expect 1 dose has greater len than 0.5 dose:
$H_0 : \mu_{0.5} - \mu_1 \geq 0$
$H_1 : \mu_{0.5} - \mu_1 < 0$

```
t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = ToothGrowth[ToothGrowth$dose!=2,], alterna
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 6.342e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.753323
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

p-value = 6.342e-08

The probability of us observing the data given H_0 as prior is vanishingly small.

We have sufficient evidence to reject H_0 in preference for $H_1 : \mu_{0.5} - \mu_1 < 0$.

From the exploratory plot, we expect 2 dose has greater len than 1 dose: $H_0 : \mu_1 - \mu_2 \geq 0$ $H_1 : \mu_1 - \mu_2 < 0$

```
t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = ToothGrowth[ToothGrowth$dose!=0.5,], alter
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  len by dose
## t = -4.9005, df = 37.101, p-value = 9.532e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -4.17387
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

p-value = 9.532e-06

The probability of us observing the data given H_0 as prior is vanishingly small.
We have sufficient evidence to reject H_0 in preference for $H_1 : \mu_1 - \mu_2 < 0$.

By the *transitive property of inequality*, we conclude that $\mu_2 > \mu_1 > \mu_{0.5}$.

Conclusions and Assumptions

Assumptions

The above analysis makes the following assumptions:

- len is normally distributed
- Each sample is an independent and random observation
- Variables between factors are uncorrelated - this justifies the use of unpaired t-test

We use weighted variances and did not assume equal variance.
This applies to equal variances without loss of generality (WLOG).

Conclusions

- **Dosage effectiveness in terms of tooth growth is in this order: 0.5 dose < 1 dose < 2 dose**
- **Supplement OG is more effective than supplement VC in assisting tooth growth**

Session Details

Specification of software and hardware used.

```
sessionInfo()
```

```
## R version 3.3.1 (2016-06-21)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 14393)
##
## locale:
## [1] LC_COLLATE=English_Singapore.1252  LC_CTYPE=English_Singapore.1252
## [3] LC_MONETARY=English_Singapore.1252 LC_NUMERIC=C
## [5] LC_TIME=English_Singapore.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
```

```
## other attached packages:
## [1] ggplot2_2.1.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.7      digest_0.6.10    grid_3.3.1       plyr_1.8.4
##  [5] gtable_0.2.0     magrittr_1.5     evaluate_0.9     scales_0.4.0
##  [9] stringi_1.1.1    rmarkdown_1.0    labeling_0.3     tools_3.3.1
## [13] stringr_1.1.0    munsell_0.4.3    yaml_2.1.13      colorspace_1.2-6
## [17] htmltools_0.3.5  knitr_1.14
```