Statistical Inference Project

Suhas PK

2023-06-18

Introduction

This project is part of Coursera John Hopkins Data Science Course: Statistical Inference. This project consists of two parts: 1. A simulation exercise. 2. Basic inferential data analysis

Project 1 : A simulation exercise

In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. We will investigate the distribution of averages of 40 exponentials. Note that we will need to do a thousand simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

As a motivating example, compare the distribution of 1000 random uniforms.

```
hist(runif(1000))
```

and the distribution of 1000 averages of 40 random uniforms

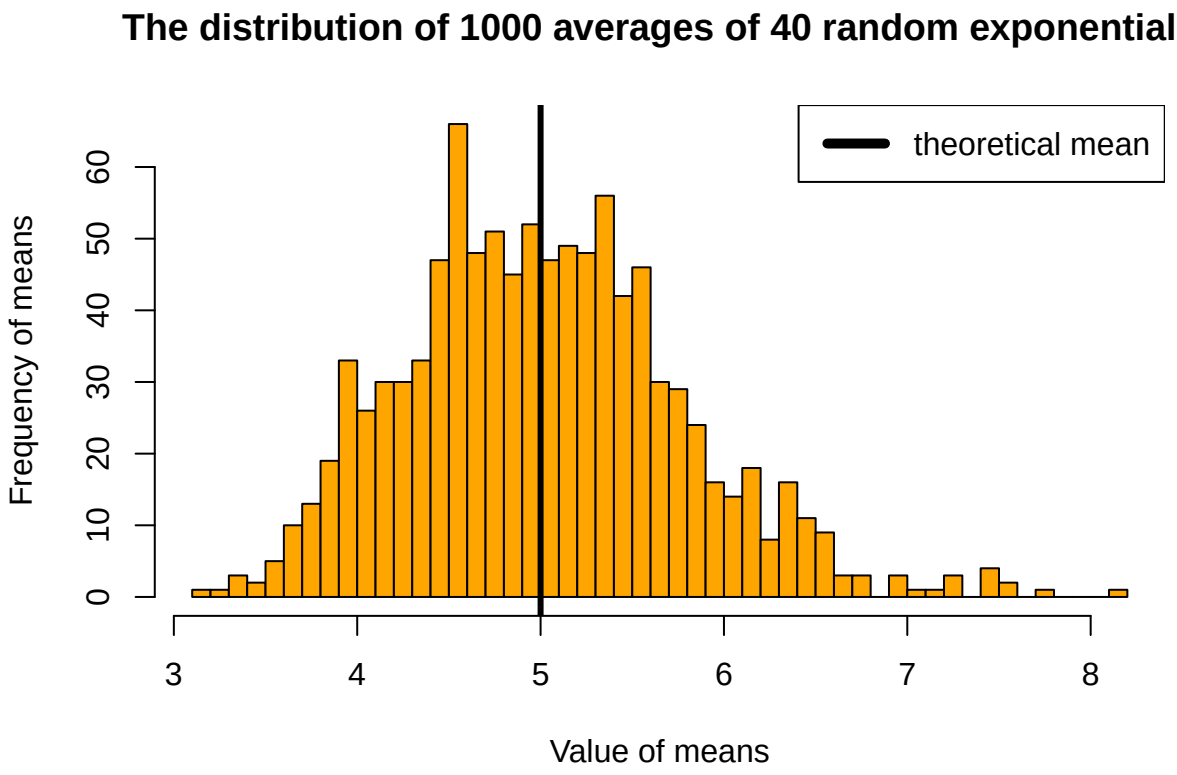
```
mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(runif(40)))
hist(mns)
```

This distribution looks far more Gaussian than the original uniform distribution!

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

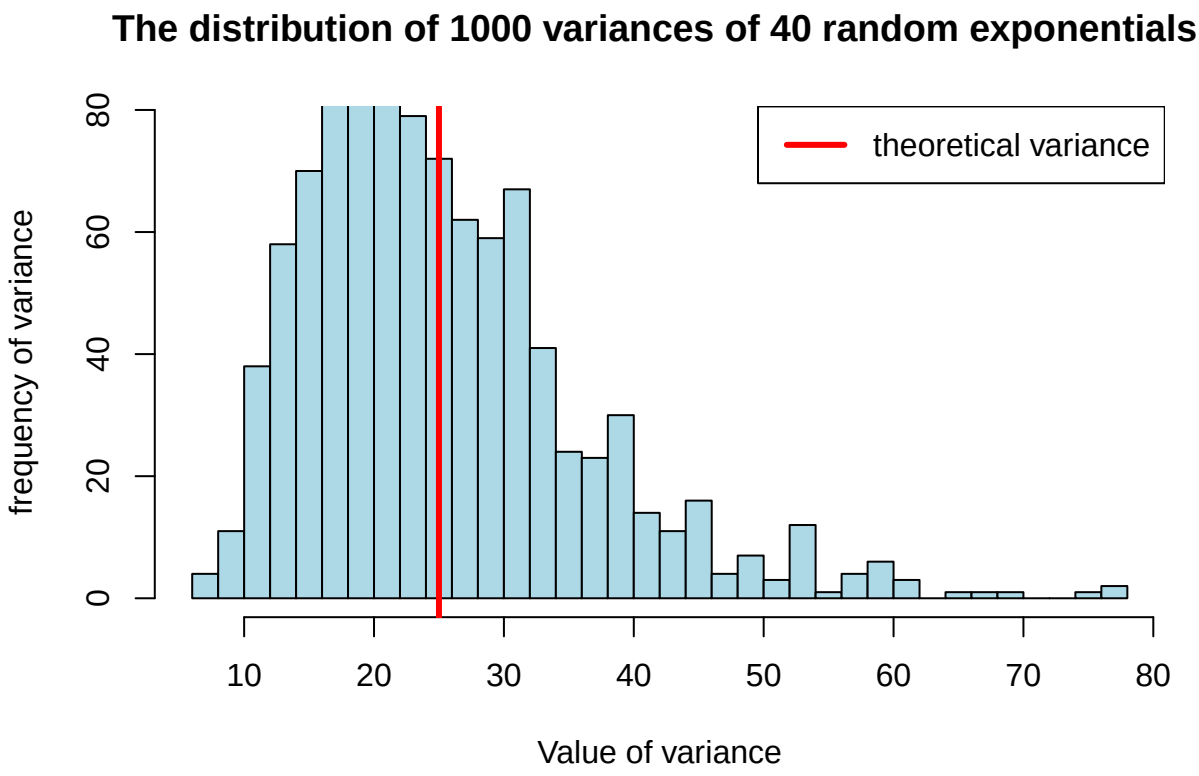
```
lambda <- 0.2
sim_data <- matrix(rexp(1000*40, lambda),
                   nrow = 1000, ncol = 40)

dist_mean <- apply(sim_data, 1, mean)

#Plotting
```

```r
hist(dist_mean, breaks = 50,
     main = " The distribution of 1000 averages of 40 random exponentials.",
     xlab = "Value of means",
     ylab = "Frequency of means", col = "orange")
abline(v=1/lambda, lty = 1,
       lwd=3, col = "black")
legend("topright", lty = 1,
       lwd = 5, col = "black", legend = "theoretical mean")
```



2. Show how variable the sample is (using variance) and compare it to the theoretical variance of the distribution.

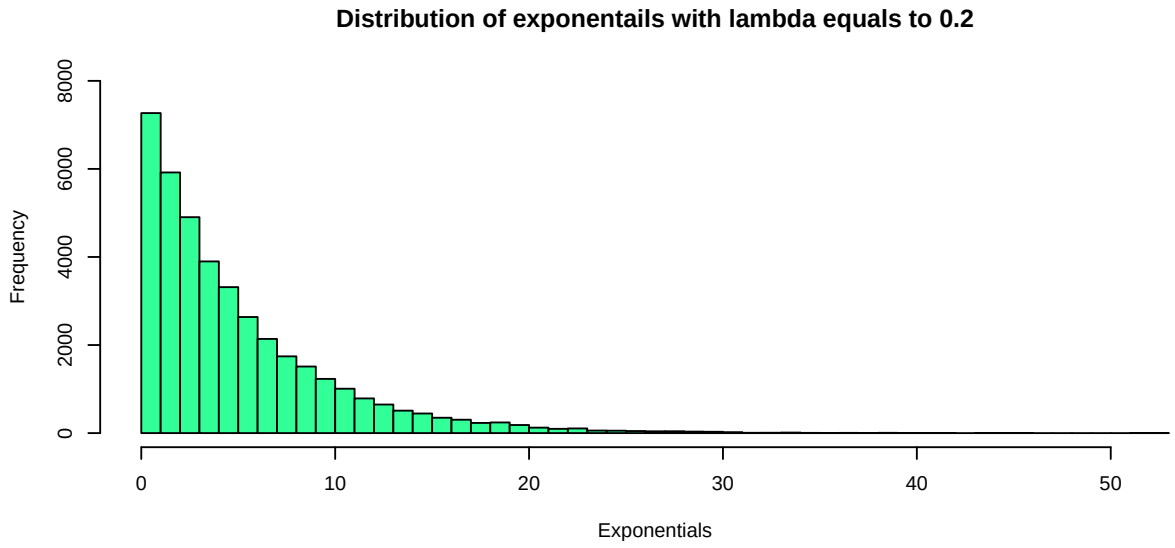
```r
variance_dist <- apply(sim_data, 1, var)

#plotting
hist(variance_dist, breaks = 50,
     main = "The distribution of 1000 variances of 40 random exponentials.",
     xlab = "Value of variance",
     ylim = c(0,max(variance_dist)),
     ylab = "frequency of variance",
     col = "lightblue")
abline(v = (1/lambda)^2, lty = 1,
       lwd = 3, col = "red")
legend("topright",lty = 1,
       lwd = 3,
       col = "red",
       legend = "theoretical variance",)
```

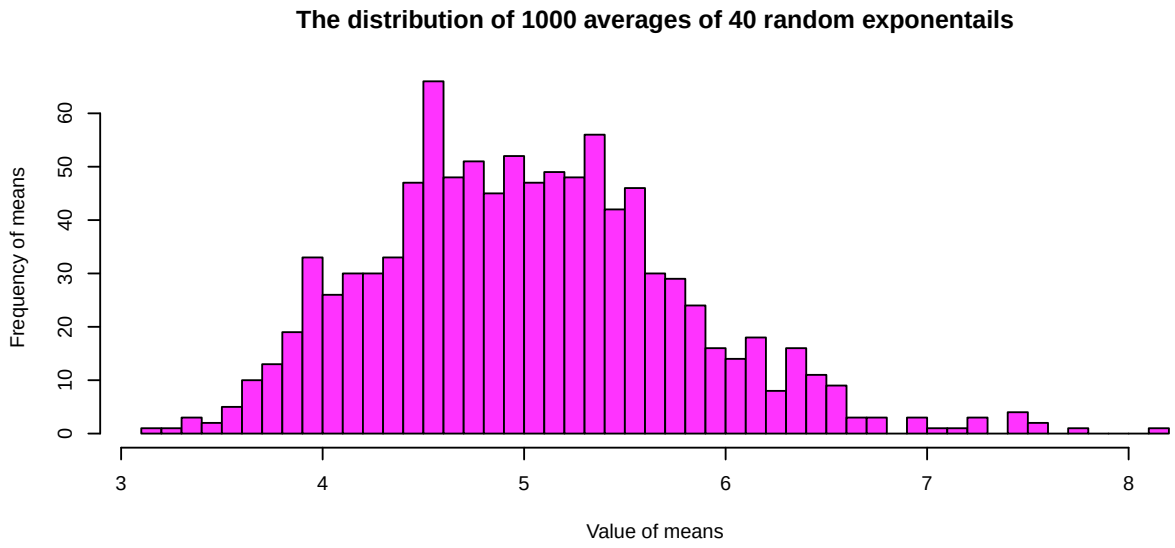


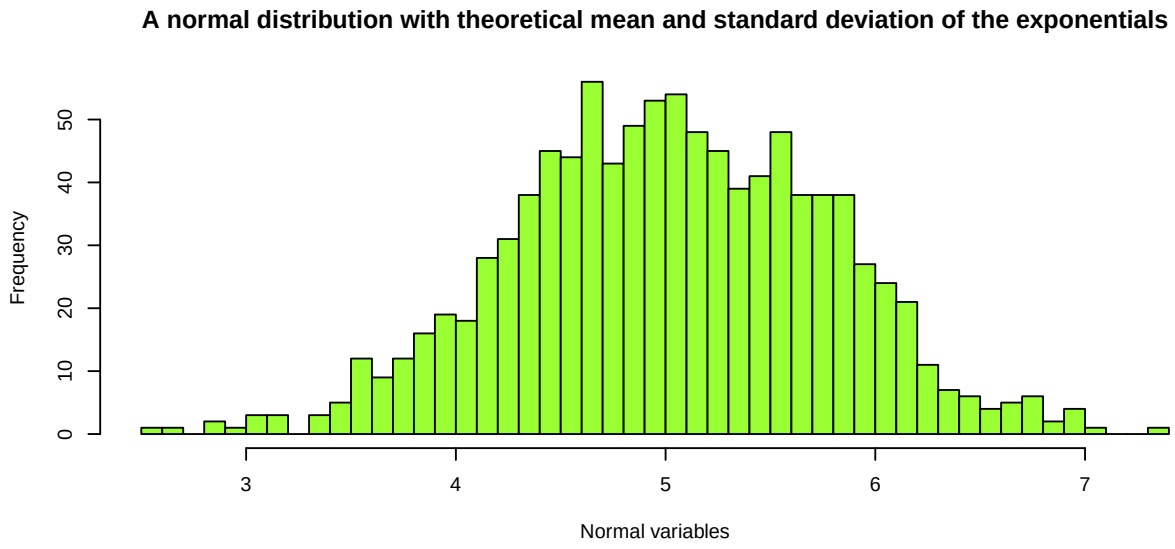
The simulated sample variances are almost normally distributed with a center near the theoretical values.

3. Show that the distribution is approximately normal.

```
# plotting
par(mfrow = c(3,1))

hist(sim_data, breaks = 50,
     main = "Distribution of exponentails with lambda equals to 0.2",
     xlab = "Exponentials",
     ylim = c(0,8000),
     col = "#33ff99")

hist(dist_mean, breaks = 50,
     main = "The distribution of 1000 averages of 40 random exponentails",
     xlab = "Value of means", ylab = "Frequency of means", col="#ff33ff")

normal_sim <- rnorm(1000, mean = mean(dist_mean), sd = sd(dist_mean))
hist(normal_sim, breaks = 50,
     main = "A normal distribution with theoretical mean and standard deviation of the exponentials",
     xlab = "Normal variables",
     col = "#99ff33")
```


Distribution of exponentails with lambda equals to 0.2
Frequency
0
2000
4000
6000
8000
0
10
20
30
40
50
Exponentials


The distribution of 1000 averages of 40 random exponentails
Frequency of means
0
10
20
30
40
50
60
3
4
5
6
7
8
Value of means


A normal distribution with theoretical mean and standard deviation of the exponentials
Frequency
0
10
20
30
40
50
3
4
5
6
7
Normal variables

The first histogram is the distribution of exponential with lambda equals to 0.2. The second histogram is the distribution of 100 averages of 40 random exponential. The third histogram is a real normal distribution with a mean and standard deviation equals to the second histograms. Comparing the first with the second histogram, we can see the distribution becomes normal as the means were taken from each groups. It is a result of the **Central Limit Theorem**.Comparing the second and the third histogram, we can see the distribution with the same mean and standard deviation.

Project 2 : Basic inferential data analysis

Now in the second project, we're going to analyze the ToothGrowth data in the R datasets package.

1. Load the ToothGrowth data and perform some basic exploratory data analyses
2. Provide a basic summary of the data.
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)
4. State your conclusions and the assumptions needed for your conclusions.

About Tooth growth dataset.

The **Tooth growth data set** is one of the standard learning data set included in R. The tooth growth data set is the length of the **ondoblasts (teeth)** in each of 10 guinea pigs at three Vitamin-C dosage levels (0.5,1,and 2 mg) with two delivery methods (orange juice or ascorbic acid).

The file contains 60 observations of 3 variables, - len : Tooth length - supp : Supplement type (VC or OJ) - dose : Dose in milligrams

Data analysis of the Tooth Growth data

```
## Loading required package: ggplot2
```

```
## Loading required package: ggdark
```

```
head(ToothGrowth, 10)
```

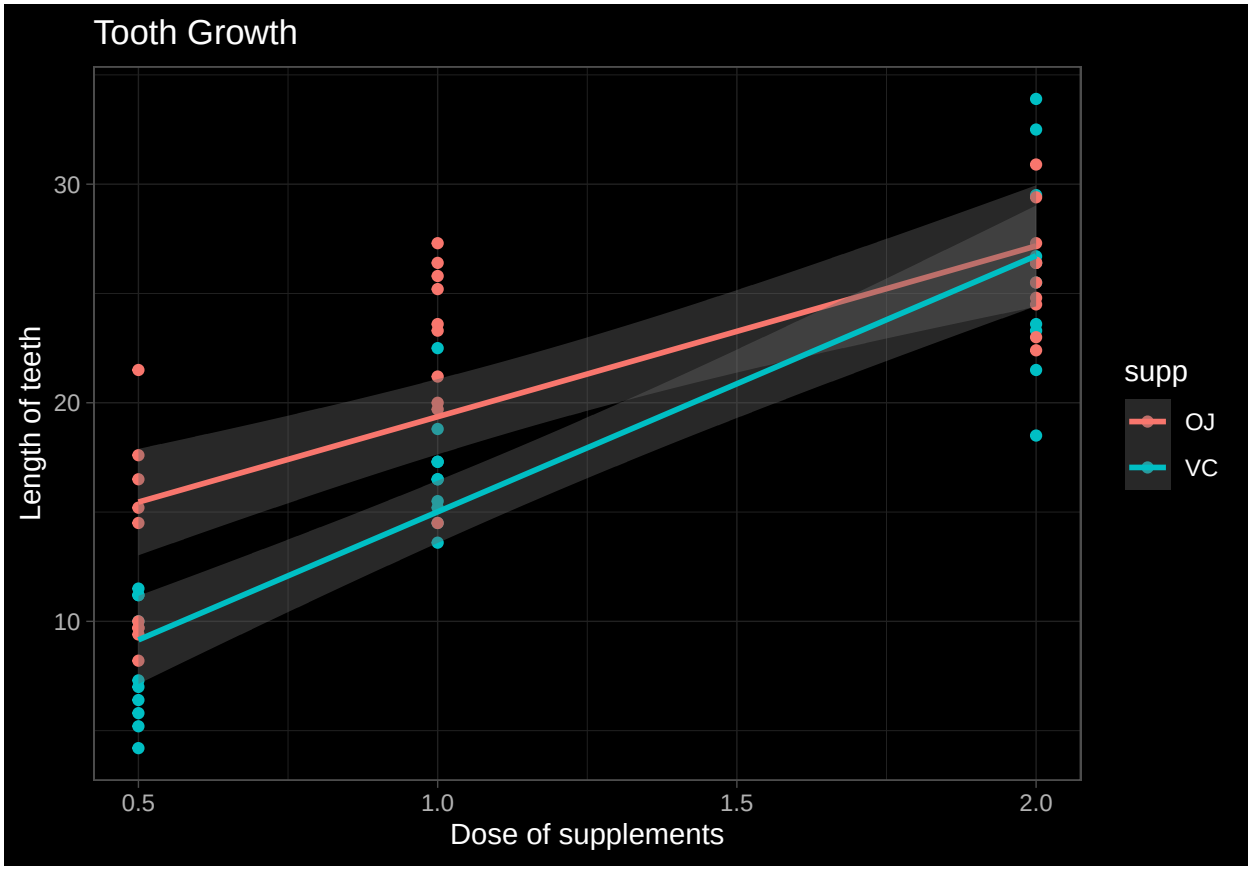
```
##     len supp dose
## 1   4.2   VC  0.5
## 2  11.5   VC  0.5
## 3   7.3   VC  0.5
## 4   5.8   VC  0.5
## 5   6.4   VC  0.5
## 6  10.0   VC  0.5
## 7  11.2   VC  0.5
## 8  11.2   VC  0.5
## 9   5.2   VC  0.5
## 10  7.0   VC  0.5
```

Load the data set.

```
data(ToothGrowth)
```

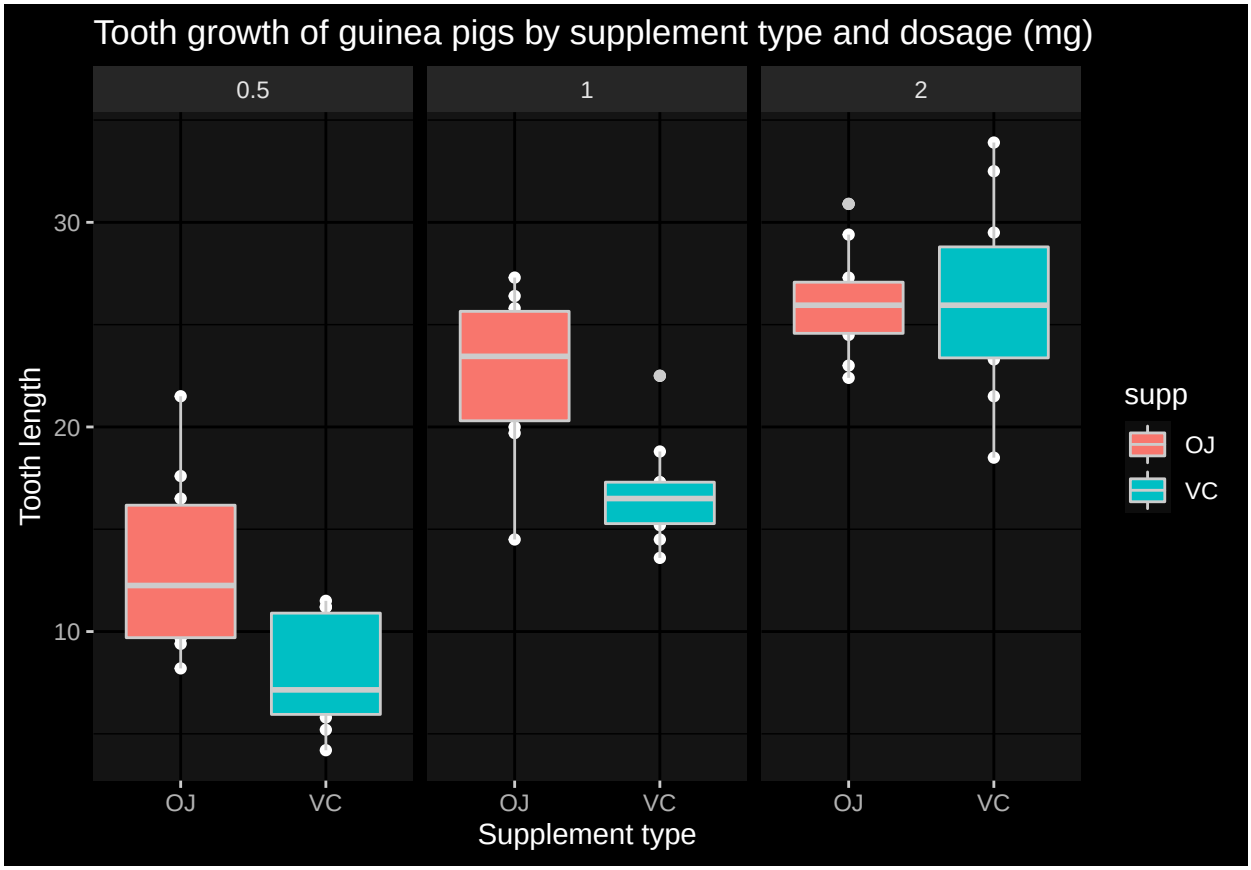
```
## Inverted geom defaults of fill and color/colour.
## To change them back, use invert_geom_defaults().
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Analysis from the plot 1. The length of teeth goes up as the does of supplements increase, which indicates that the supplements may help teeth growth. 2. At the same dose, OJ seems to incur a higher 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.

```
qplot(supp,len,data=ToothGrowth,
      facets=~dose, main="Tooth growth of guinea pigs by supplement type and dosage (mg)",
      xlab="Supplement type",
      ylab="Tooth length") +
    geom_boxplot(aes(fill = supp)) + dark_theme_grey()
```



There is a positive effect of the dosage, as the dosage increases the tooth growth increase. In the specific case of the VC, the tooth growth has a linear relationship with dosage. The higher dosage (2.0 mg) has less improvement in tooth growth with the OJ supplement. However, the OJ supplement generally induces tooth more growth than VC except at higher dosage (2.0 mg).

Hypothesis Testing

Assumptions - The variables must be independent and identically distributes (iid). - Variances of tooth growth are different when using different supplement and dosage. - Tooth growth follows a normal distribution.

Hypothesis for the supplement OJ vs VC

Let our null hypothesis to be there is no difference in tooth growth when using the supplement OJ and VC.

Let our alternate hypothesis to be there more tooth growth when using supplement OJ than VC.

```
OJ <- ToothGrowth$len[ToothGrowth$supp == 'OJ']
VC <- ToothGrowth$len[ToothGrowth$supp == 'VC']
```

Let us perform a t-test following the indications of the work to be evaluated.
One-tailed independent t-test with unequal variance.

```
t.test(
    OJ,VC,
    alternative = "greater",
    paired = FALSE,
    var.equal = FALSE,
    conf.level = 0.95
)
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  OJ and VC
## 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 of x mean of y
##  20.66333  16.96333
```

As the p-value (0.03032) is lower than 0.05 (the default value for the tolerance of the error alpha), then, we **reject the null hypothesis**. This can be interpreted as there is approximately 3% of chance to obtain an extreme value for the difference in mean of the tooth growth.

Based on this low p-value, it can be concluded that it is very likely that supplement OJ, the greater the effect on tooth growth than supplement VC.

Hypothesis for dosage.

The null hypothesis is that there is no difference in tooth growth between dosage. Our alternate hypothesis is that there are most tooth growth when the dosage increases.

Extract the tooth growth by dosage.

```
dose_half <- ToothGrowth$len[
    ToothGrowth$dose == 0.5
]
dose_one <- ToothGrowth$len[
    ToothGrowth$dose == 1
]
dose_two <- ToothGrowth$len[
    ToothGrowth$dose == 2
]
```

One-tailed independent t-test with unequal variance.

```
t.test(dose_half,
       dose_one,
       alternative = "less",
       paired = FALSE,
       var.equal = FALSE,
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  dose_half and dose_one
## 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 of x mean of y
##    10.605    19.735
```

As the p-value (6.342e-08) is lower than 0.05 (the default value for the tolerance of the error alpha), then, we **reject the null hypothesis**. That can be interpreted as there is almost null chances of obtain an extreme value for the difference in mean of those dosages (dose_half < dose_one) on the tooth growth.

```
t.test(
    dose_one,dose_two,
    alternative = "less",
    paired = FALSE,
    var.equal = FALSE,
    conf.level = 0.95
)
```

```
##
##  Welch Two Sample t-test
##
## data:  dose_one and dose_two
## 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 of x mean of y
##    19.735    26.100
```

The conclusion is similar than the previous, the p-value is 9.532e-06, close to 0. Then we reject the null hypothesis. That can be interpreted as there is almost null chances of obtain an extreme value for the difference in mean of those dosages (dose_one < dose_two) on the tooth growth. The value is extreme (that's what we reject the null hypothesis)

Based on these low p-values, we can conclude that it is very likely that dosage has effect, and a higher dosage higher tooth growth.

Hypothesis for the supplement OJ vs VC at dosage 2.0 mg

Two-tailed independent t-test with unequal variance.

```
t.test(OJ2, VC2,
       alternative = "two.sided",
       paired = FALSE,
       var.equal = FALSE,
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ2 and VC2
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

The p-value (0.9639) confirm what we suspect, that we can't reject the null hypothesis (p-value is higher than 0.05 (the default value for the tolerance of the error alpha). Then, there is insufficient evidence to show that there is a difference in tooth growth when using supplement OJ and VC at dosage 2.0 mg.

CONCLUSION

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.