Statistical Inference Course Project

John Cardona

2022-11-30

Description

The first part of this report will descriptively and graphically elucidate the differences between a distribution of a large collection of exponential variables (40000) and a distribution of 1000 averages of 40 exponential variables. The second part of this report will evaluate whether toothlength is affected by supplementation (vitamin C vs orange juice) or dosage (.5, 1, or 2 mg/day).

```
library(dplyr) library(tidyr)
library(ggplot2)
library(ggpubr)
```

Simulation of the averages of 1000 sets of 40 exponential variables (lambda = .2):

```
#This simulation creates a vector 'output" with the averages of #1000 sets of 40
exponential variables with lambda = .2
set.seed(15) output_mean <- vector("numeric",
length(1:1000)) for (i in 1:1000) { output_mean[[i]] <-
mean(rexp(40,.2))
}

set.seed(15)
output_sd <- vector("numeric", length(1:1000)) for (i in
1:1000) { output_sd[[i]] <- sd(rexp(40,.2))
}
set.seed(15)
output_var <- vector("numeric", length(1:1000)) for (i in
1:1000) { output_var[[i]] <- var(rexp(40,.2))
}
```

Summary statistics for 1000 averages of 40 exponential variables (lambda = .2):

```
sample_mean <- mean(output_mean)
sample_sd <- mean(output_sd) sample_var <-
mean(output_var) sample_mean
```

```
## [1] 4.980535
```

```
sample_sd
```

```
## [1] 4.875578
```

```
sample_var
```

```
## [1] 24.87989
```

Notice that the sample mean, 4.9805352, is near the theoretical mean (1/lambda), 5.

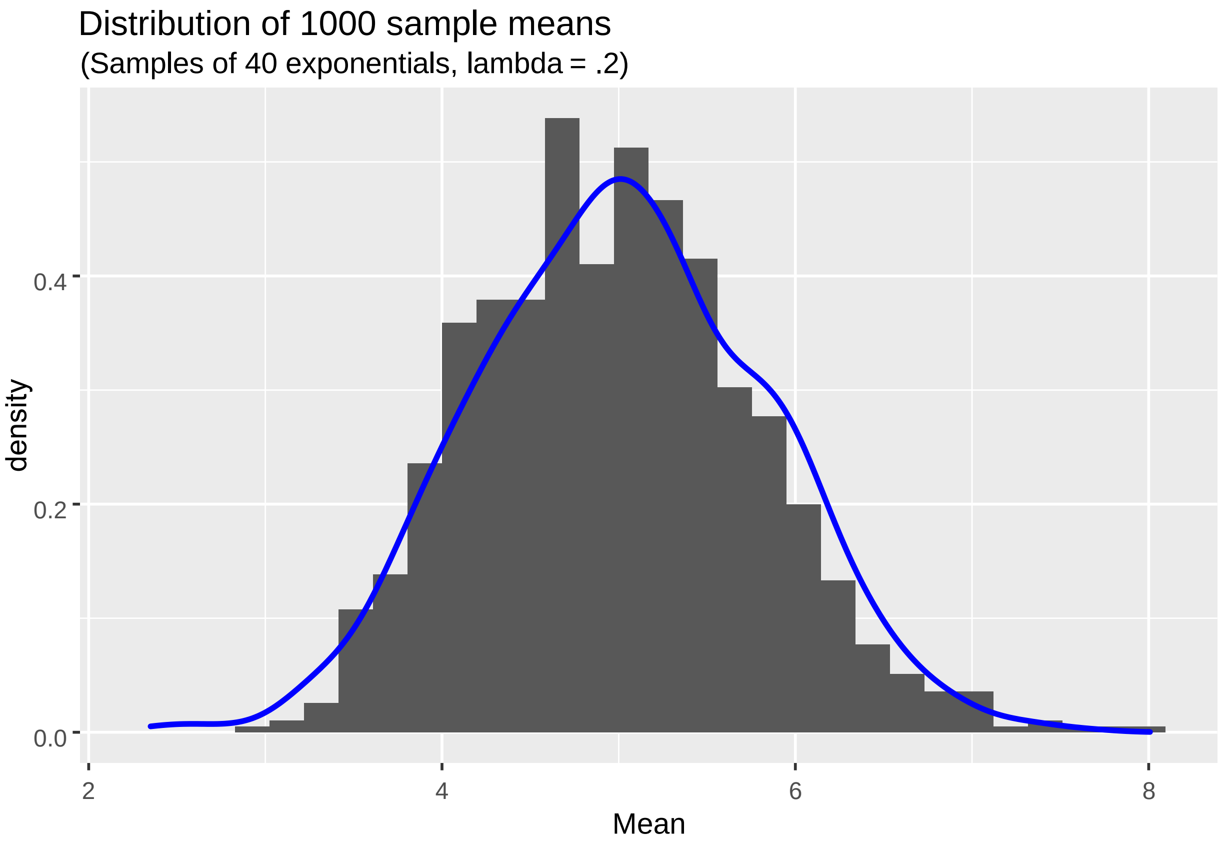
Likewise, the sample variance, 24.8798938 is near the theorectical variance (1/lambdaˆ2), 25.

Now let’s visual our simulated sample means:

```
sim <- as_tibble(output_mean) norm_curve <- rnorm(1000,
sample_mean, sd(output_mean)) sim$norm <- norm_curve

sim %>% ggplot() + geom_histogram(aes(x = value, y = ..density..))
   + geom_density(aes(x=norm), col = "blue", lwd = 1) + labs(
      x = "Mean", title = "Distribution of 1000 sample means", subtitle =
      "(Samples of 40 exponentials, lambda = .2)"
   )
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



As you can see, the sample means follow a normal distribution (Central Limit Theorem). The blue line shows a density curve of 1000 normal variables with an equal mean and sd to our simulated set of sample means.

Now let’s conceptualize the difference between the mean of a large sample of exponentials and a sample of means of exponentials:

Simulation of 40000 exponentials (lambda = .2):

```
#Sample distribution of 40000 exponential with lambda = .2
set.seed(12) output2 <-
rexp(40000, .2) head(output2)
```

```
## [1] 10.9460810 3.1777520 0.5869159 14.2862432 9.0828720 1.4166537
```

Summary statistics for distribution of 40000 exponentials (lambda = .2):

```
mean(output2)
```

```
## [1] 5.010015
```

```
sd(output2)
```

```
## [1] 4.970864
```

```
var(output2)
```

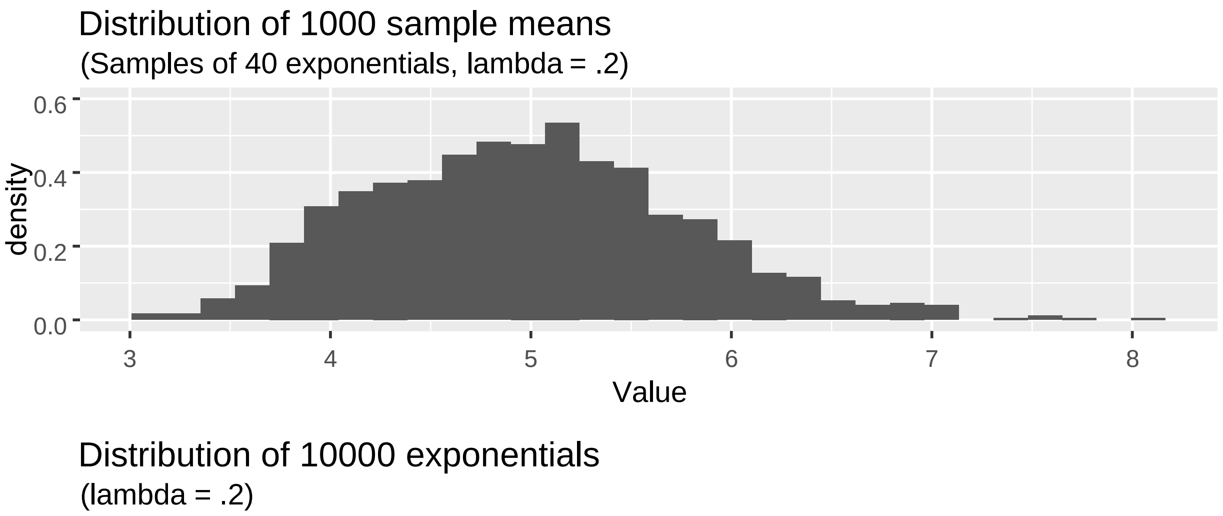
```
## [1] 24.70949
```

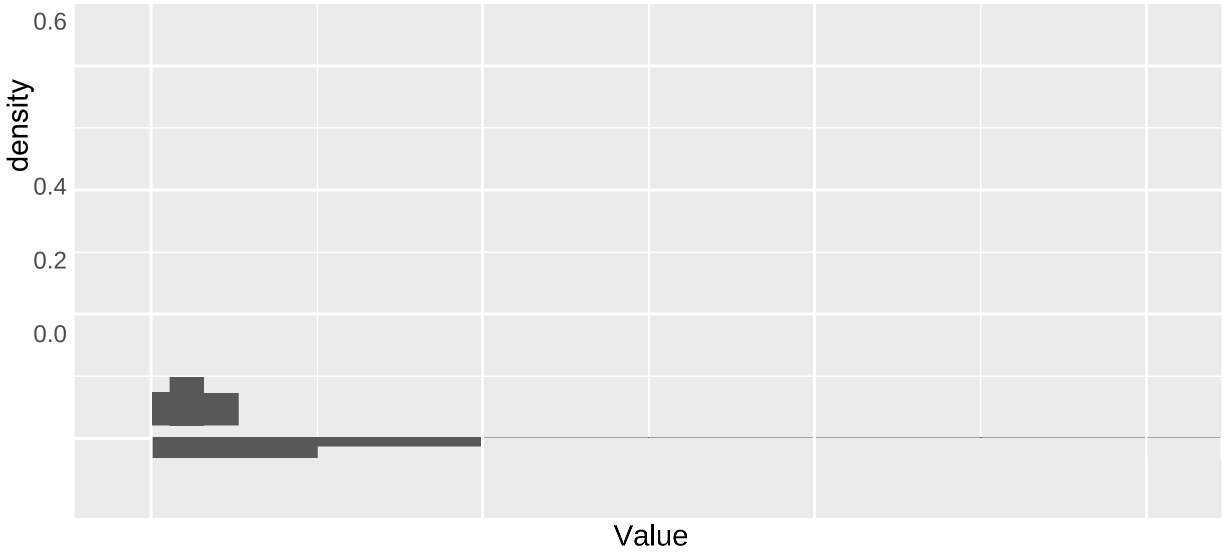
Notice that the means of 40000 exponentials and the mean of the averages of 1000 sets of 40 exponentials are similar but the standard deviation (and variance) of the single sample (40000 exponentials) is quite larger.

To show the variability between the two distributions are different, let's take a look the following graphs:

```
a <- as_tibble(output_mean) %>% ggplot() +
    geom_histogram(aes(x = value, y = ..density..)) +
    scale_y_continuous(limits = c(0,.6)) + labs(
        x = "Value", title = "Distribution of 1000 sample means", subtitle =
        "(Samples of 40 exponentials, lambda = .2)"
    ) b <- as_tibble(output2) %>% ggplot() +
geom_histogram(aes(x = value, y = ..density..)) +
scale_y_continuous(limits = c(0,.6)) + labs(
        x = "Value", title = "Distribution of 10000 exponentials",
        subtitle = "(lambda = .2)"
    ) ggarrange(a,b, ncol = 1, nrow = 2)
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```





As you can see, the variability of the single large sample is greater than the variability of the distribution of sample means. Notice that the distribution on the bottom is exponentially distributed rather than normally distributed.

Now let’s take a look at the ‘ToothGrowth’ dataset:

```
TG <- ToothGrowth str(TG)
```

```
## ’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 ...
```

This dataset contains 60 observations on 3 variables - Tooth length, supplement type (orange juice of vitamin C), and dose (mg/day).

```
#Tables of categorical predictors table(TG$dose)
```

```
##
## 0.5   1   2
## 20 20 20
```

```
table(TG$supp)
```

```
##
## OJ VC
## 30 30
```

Let’s take a look at the effect of supplement type and dose on tooth length by using a linear regression model.

```
y <- lm(data = TG, len ~ as.factor(dose) + supp + factor(dose) * supp) summary(y)
```

```
##
## Call:
```

```
## lm(formula = len ~ as.factor(dose) + supp + factor(dose) * supp,
##         data = TG)
##
## Residuals:
##      Min       1Q Median      3Q     Max
## -8.20 -2.72 -0.27                 2.65     8.27
##
## Coefficients: (2 not defined because of singularities)
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  13.230         1.148 11.521 3.60e-16 ***
## as.factor(dose)1              9.470         1.624          5.831 3.18e-07 ***
## as.factor(dose)2             12.830         1.624          7.900 1.43e-10 ***
## suppVC                       -5.250         1.624 -3.233 0.00209 **
## factor(dose)1                     NA            NA          NA           NA
## factor(dose)2                     NA            NA          NA           NA
## suppVC:factor(dose)1         -0.680         2.297 -0.296 0.76831
## suppVC:factor(dose)2          5.330         2.297        2.321 0.02411 *
## ---
## Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.631 on 54 degrees of freedom
## Multiple R-squared: 0.7937, Adjusted R-squared: 0.7746 ## F-statistic:
41.56 on 5 and 54 DF, p-value: < 2.2e-16
```

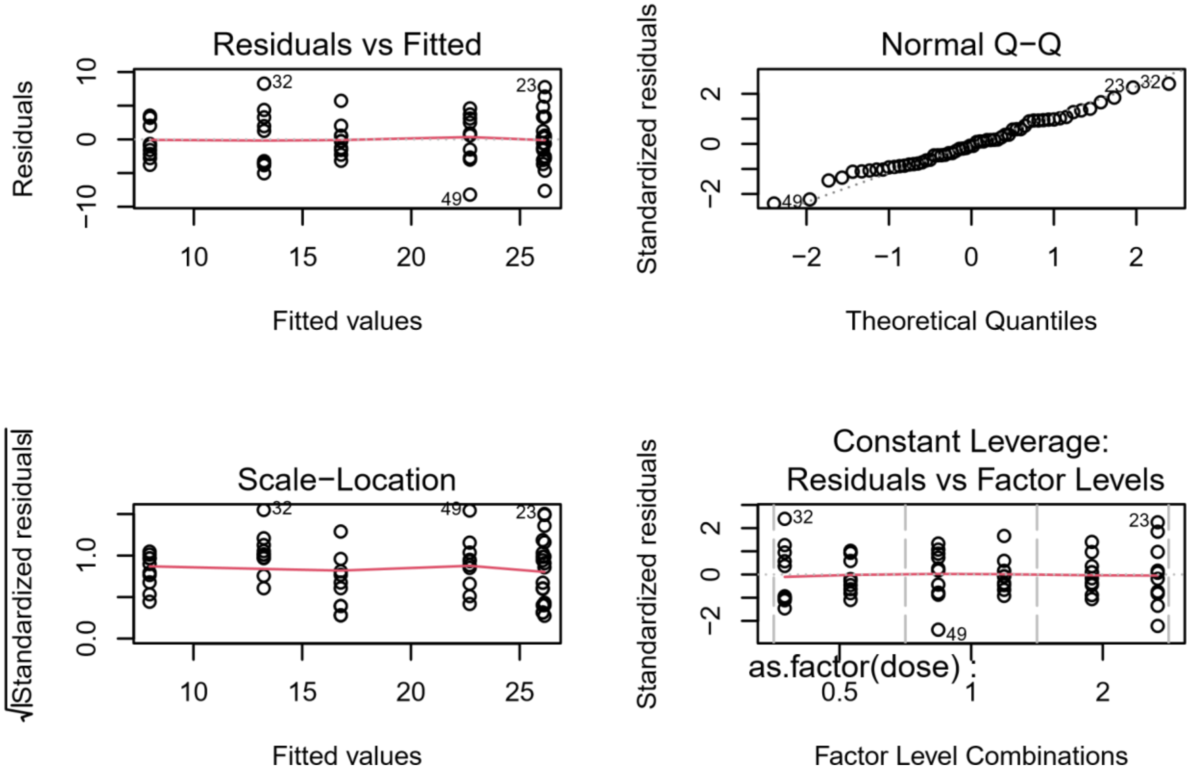
```
confint(y)
```

```
##                                         2.5 %        97.5 %
## (Intercept)                   10.9276907 15.532309
## as.factor(dose)1               6.2140429 12.725957
## as.factor(dose)2               9.5740429 16.085957
## suppVC                        -8.5059571 -1.994043
## factor(dose)1                         NA              NA
## factor(dose)2                         NA              NA
## suppVC:factor(dose)1 -5.2846186 3.924619
## suppVC:factor(dose)2 0.7253814 9.934619
```

As you can see by the output, the higher dosages increase tooth length in a dose-dependent manner. Additionally, we see that guinea pigs given orange juice had longer teeth than guinea pigs given vitamin C. There also appears to be an interaction between supplement and dosage at 2 mg/day.

Let’s assess whether our model is a good fit to data:

```
par(mfrow = c(2,2)) plot(y)
```



It should be noted that this model assumes that the relationship between tooth length and the independent variables is linear. Furthermore, the model assumes that data on all variables that significantly influence tooth length were collected.