The Central Limit Theorem and the Exponential Distribution and Tooth Growth in Guinea Pigs

T. Bennett

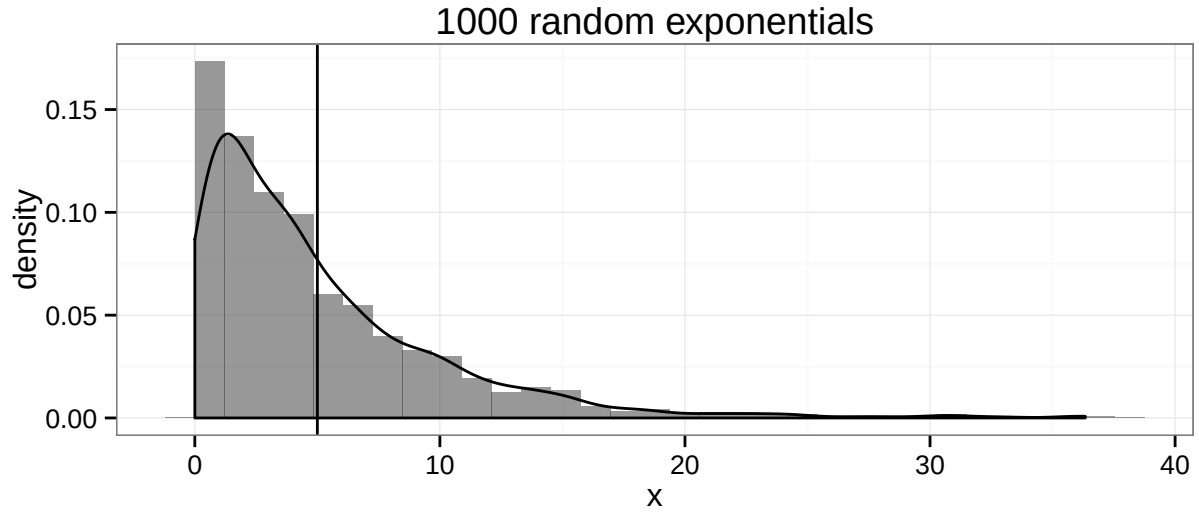
Monday, May 18, 2015

Synopsis

This report shows that the central limit theorem applies for the exponential distribution, i.e. the means of sets of random exponentials tend to be normally distributed with their mean at the population mean. Tooth growth is associated with both vitamin C dose and delivery method.

Simulations

Here we generate 1000 random exponentials with $\lambda = 0.2$ and display the `x` values using a histogram with an overlain density curve.



The theoretical mean of the distribution should be $1/\lambda = 5$ and the theoretical variance should be $(1/\lambda)^2 = 25$. The observed values are quite close to the theoretical values.

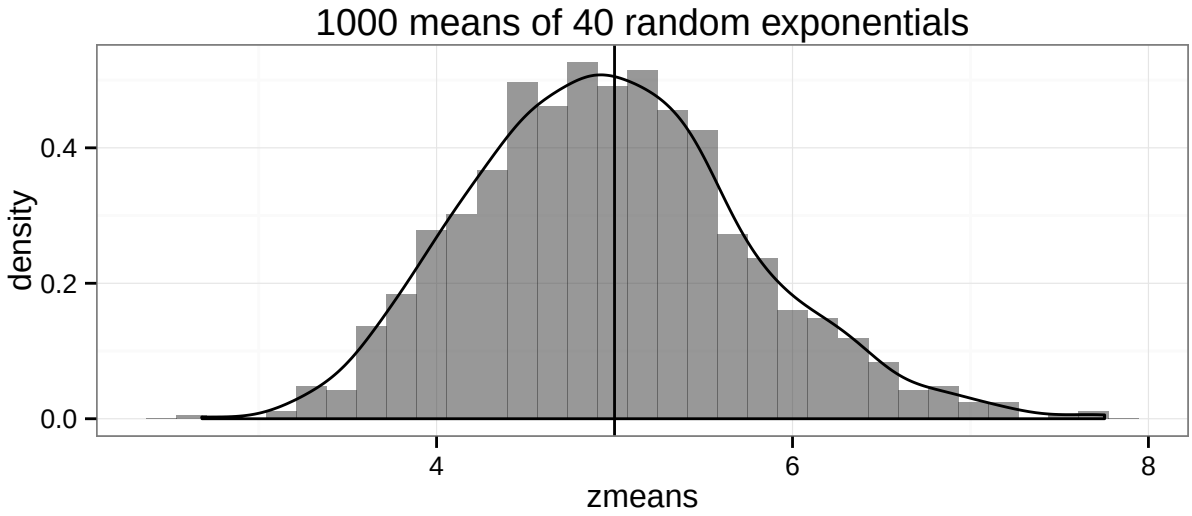
```
mean(x)
```

```
## [1] 4.991419
```

```
var(x)
```

```
## [1] 25.53088
```

Here we generate 40 exponentials with $\lambda = 0.2$ **1000 times** and display the mean of each group of 40 exponentials using a histogram with an overlain density curve.



The theoretical mean of the distribution of 1000 means should be the population mean $1/\lambda = 5$ and the observed mean is quite close to that. The standard error of the mean should be $\sqrt{((1/\lambda)^2)/n} = 0.016$ and the sample value is just a bit larger.

```
mean(zmeans)
```

```
## [1] 4.972126
```

```
sd(zmeans)
```

```
## [1] 0.7737079
```

The above figure shows that the distribution of the means of 1000 groups of 40 exponentials is approximately normal, much more so than the single group of 1000 exponentials. The mean is very near the population mean, and the variance is symmetric.

ToothGrowth Analysis

Loading the dataset and checking its structure

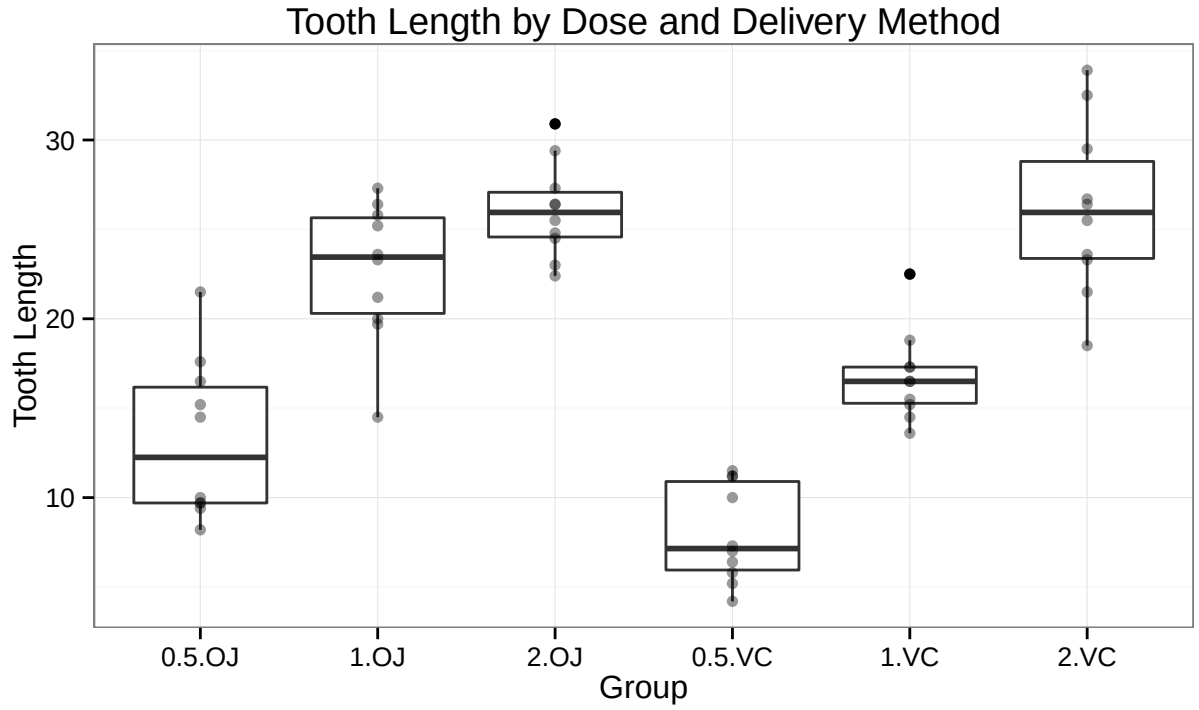
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
##       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
```

Summary of data

60 observations of 3 variables, no missing data

Design: 10 guinea pigs, 3 different doses (`dose`) of vitamin C
Delivery method: `supp` = orange juice or ascorbic acid
Response: `len` = length of teeth



Visually, tooth length appears to be associated with vitamin C dose and perhaps delivery method. Those hypotheses are tested below. I assume in both cases that the variances are unequal and that a paired test is appropriate because the same 10 guinea pigs are being tested.

```
## mean of the differences
##                     3.7
```

```
## [1] 1.408659 5.991341
## attr(,"conf.level")
## [1] 0.95
```

The above test shows that orange juice is associated with more tooth growth than ascorbic acid.

```
## mean of the differences
##                 -15.495
```

```
## [1] -18.3672 -12.6228
## attr(,"conf.level")
## [1] 0.95
```

I excluded the middle dose (1) to make the t-test work, but lower dose is associated with less tooth growth when the `0.5` and `2` groups are compared. Linear regression would be a better way to do this.

Appendix

R code - run above, not run here

Set random number seed and load packages

```
set.seed(1234)
library(ggplot2)
library(dplyr)
```

Generate one set of 1000 random exponentials and show that histogram

```
x <- (rexp(1000, 0.2))
y <- as.data.frame(x)
p <- ggplot(y, aes(x = x)) + theme_bw()
p + geom_histogram(aes(y = ..density..), alpha=0.5) + geom_density(size = 0.5, alpha = 0.2) + geom_vlin
```

Generate 40 exponentials **1000 times** and show that histogram

```
nosim <- 1000
n <- 40
z <- matrix(rexp(nosim * n, 0.2), nrow = nosim)
zmeans <- apply(z, 1, mean)
zmdf <- as.data.frame(zmeans)
p2 <- ggplot(zmdf, aes(x = zmeans)) + theme_bw()
p2 + geom_histogram(aes(y = ..density..), alpha=0.5) + geom_density(size = 0.5, alpha = 0.2) + geom_vli
```

Load the toothgrowth data and generate some simple dataset summaries Loading the dataset and checking its structure

```
library(datasets)
data(ToothGrowth)
str(ToothGrowth)
summary(ToothGrowth)
```

Generate the necessary analysis variables and the exploratory tooth growth box plot

```
data <- ToothGrowth %>%
    mutate(Group = interaction(as.factor(dose), supp))
p <- ggplot(data, aes(Group, len)) + theme_bw()
p + geom_boxplot() + geom_point(alpha=0.4) + labs(title = "Tooth Length by Dose and Delivery Method", y
```

Perform the t-tests on the tooth growth data

```
t.test(len ~ supp, data = data, paired=TRUE, var.equal=FALSE)$estimate
t.test(len ~ supp, data = data, paired=TRUE, var.equal=FALSE)$conf
```

```
datasm1 <- data %>%
    filter(dose!=1)
t.test(len ~ dose, data = datasm1, paired=TRUE, var.equal=FALSE)$estimate
t.test(len ~ dose, data = datasm1, paired=TRUE, var.equal=FALSE)$conf
```