Statistical Inference Course Project

C. Davis Buenger

August 27, 2016

The following is a write-up for the Statistical Inference course in the John Hopkins data science specialization through Coursera. We investigate two problems. The first studies applying the central limit theorem to i.i.d. random variables which follow an exponential distribution with mean 5. The second investigates the ToothGrowth data in the R datasets package. Specifically, we use statistical techniques to assess significance of the effect of various treatments on tooth growth.

The Central Limit Theorem Applied to the Exponential Distribution

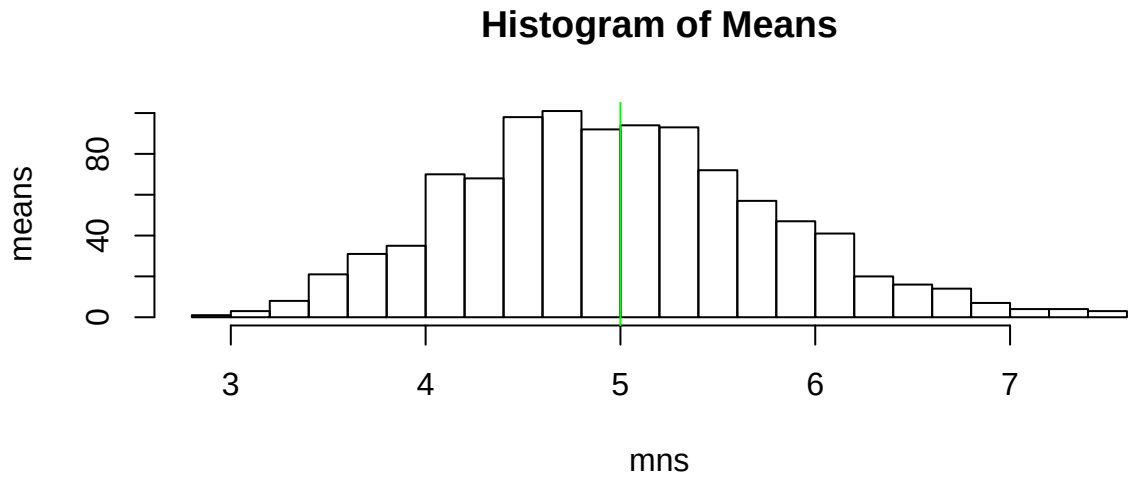
The classical central limit theorem states that for any i.i.d. random variables $x_1, x_2, \ldots$ with $Ex_i = \mu < \infty$ and $\text{Var}(x_i) = \sigma^2 < \infty$, we have that

$$\sqrt{n}\left(\frac{1}{n}\sum_i x_i - \mu\right) \to \text{N}(0, \sigma^2),$$

where convergence in this case is convergence in distribution and $\text{N}(0, \sigma^2)$ is the normal distribution centered at 0 with variance σ^2. Since the exponential distribution has finite mean and variance, we know that it follows the central limit theorem. Let us demonstrate this through simulation.

The following code simulates collections of 40 random variables with exponential distribution where $\lambda = 0.2$. We then plot a histogram of the average value of these simulations. The mean of a random variables with exponential distribution where $\lambda = 0.2$ is 5 and the variance is 25. By the CLT the average value of these collections should be approximately distributed like a normal distribution with mean 5 and variance 25/40.

```
 mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(rexp(40,0.2)))
hist(mns,breaks=20,ylab="means",main = "Histogram of Means")
abline(v=5,col="green")
```

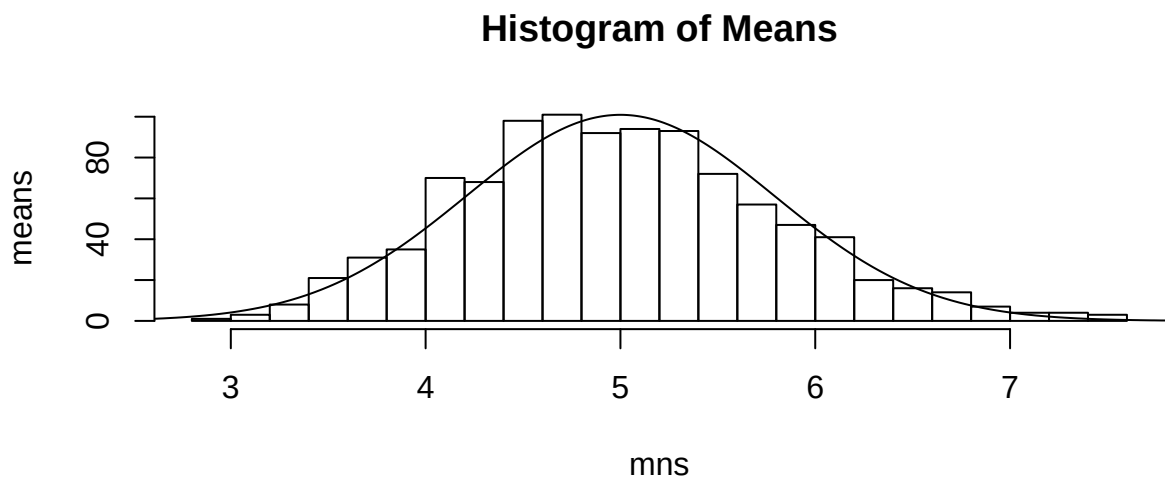


```
varmns<-var(mns);varmns
```

```
## [1] 0.6376319
```

As you can see, the means of our simulations are roughly centered around the green line at 5 which is the mean of our random variables and the variance of our simulations is 0.6376319 which is only 0.0126319 from our predicted variance of 0.625. Finally, let us fit a normal distribution to our histogram and observe the fit. (Note that we had to re-scale our normal function because the width of our histogram bins was not equal to 1.)

```
hist(mns,breaks=20, ylab="means",main = "Histogram of Means")
x   <- seq(0,10,length=1000)
y   <- 200*dnorm(x,mean=5, sd=5/(sqrt(40)))
lines(x,y, lwd=1)
```



A Statistical Analysis of the ToothGrowth Data Set

The ToothGrowth data set records tooth growth in 60 guinea pigs in three variables. The varaible len measures the length of odontoblasts (cells responsible for tooth growth). Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (coded as OJ or VC). These are measured in the variables dose and supp respectfully. We use statistical infference to deduce significance of the differences in effect of dose and delivery method on tooth growth.

First, let us examine the data.

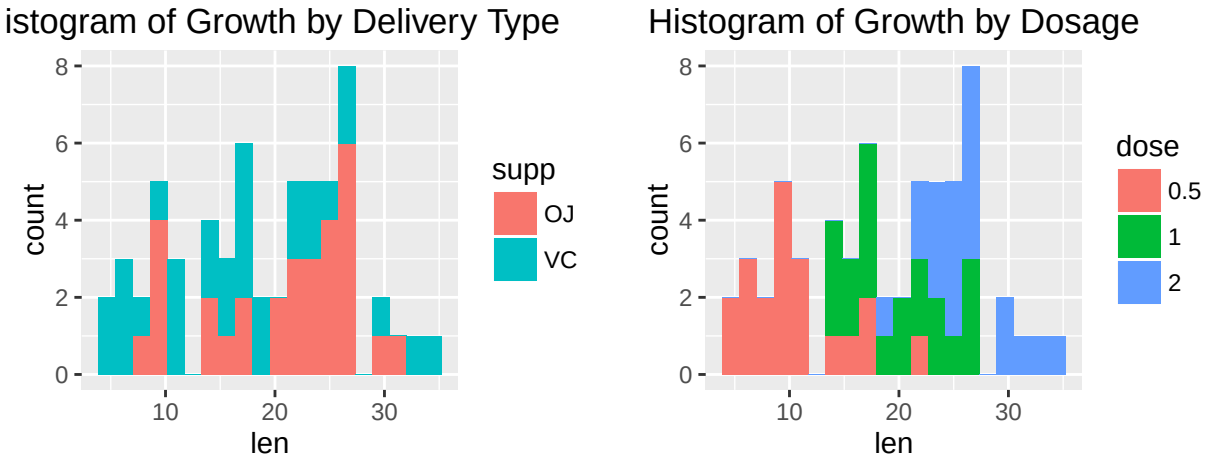
```
data("ToothGrowth")
ToothGrowth$dose<-as.factor(ToothGrowth$dose)
summary(ToothGrowth)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
```

```
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

As advertised, the first variable measures tooth growth as a numeric which ranges from 4.2 to 33.9, and the second two variables reflect treatment type. Second, let us explore the shape of the data when broken down by delivery type and by dosage.

```
library(ggplot2)
library(gridExtra)
plot1 <-qplot(len,data= ToothGrowth, fill=supp, bins=20, main = "Histogram of Growth by Delivery Type")
plot2 <-qplot(len,data= ToothGrowth, fill=dose, bins=20, main = "Histogram of Growth by Dosage")
grid.arrange(plot1, plot2, ncol=2)
```



As you can see the, the growth under the OJ treatment seems slightly higher than that VC, and the growth seems to be larger as we increase dosage. Let us test the statistical significance of these claims against the null hypothesis that treatment differences have no effect.

First, we compare orange juice and ascorbic acid. We performed an unpaired t-test on the two subgroups.

```
OJ<-subset(ToothGrowth, supp=="OJ")
VC<-subset(ToothGrowth, supp=="VC")
test1<-t.test(OJ$len,VC$len,paired= FALSE)
test1$p.value
```

```
## [1] 0.06063451
```

```
test1$conf.int
```

```
## [1] -0.1710156  7.5710156
## attr(,"conf.level")
## [1] 0.95
```

The p value of the test is larger than .5, so the difference is not statistically significant at the 95% level. This can also be inferred by the fact that 0 fell into the t-tests 95% confidence interval.

Second, we perform three unpaired t-test comparing each combination of two of the three dosage treatments.

```
dose.5<-subset(ToothGrowth, dose==0.5)
dose1<-subset(ToothGrowth, dose==1.0)
dose2<-subset(ToothGrowth, dose==2.0)
test2<-t.test(dose.5$len,dose1$len,paired= FALSE)
test2$p.value
```

```
## [1] 1.268301e-07
```

```
test3<-t.test(dose1$len,dose2$len,paired= FALSE)
test3$p.value
```

```
## [1] 1.90643e-05
```

```
test4<-t.test(dose.5$len,dose2$len,paired= FALSE)
test4$p.value
```

```
## [1] 4.397525e-14
```

The p-values in each of the three test are less than 0.001. Thus, there is a statistically significance between the three dosages treatments. Under the assumption that the difference of the two sets follows a t-distribution with 38 degrees of freedom, we can say that tooth growth increases as dosage increases with a probability of at least 99.9%.