Coursera Statistical Inference Project

Father Abraham

13/09/2014

Part 1

The first task of the project is to “[i]llustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponential(0.2)s”.

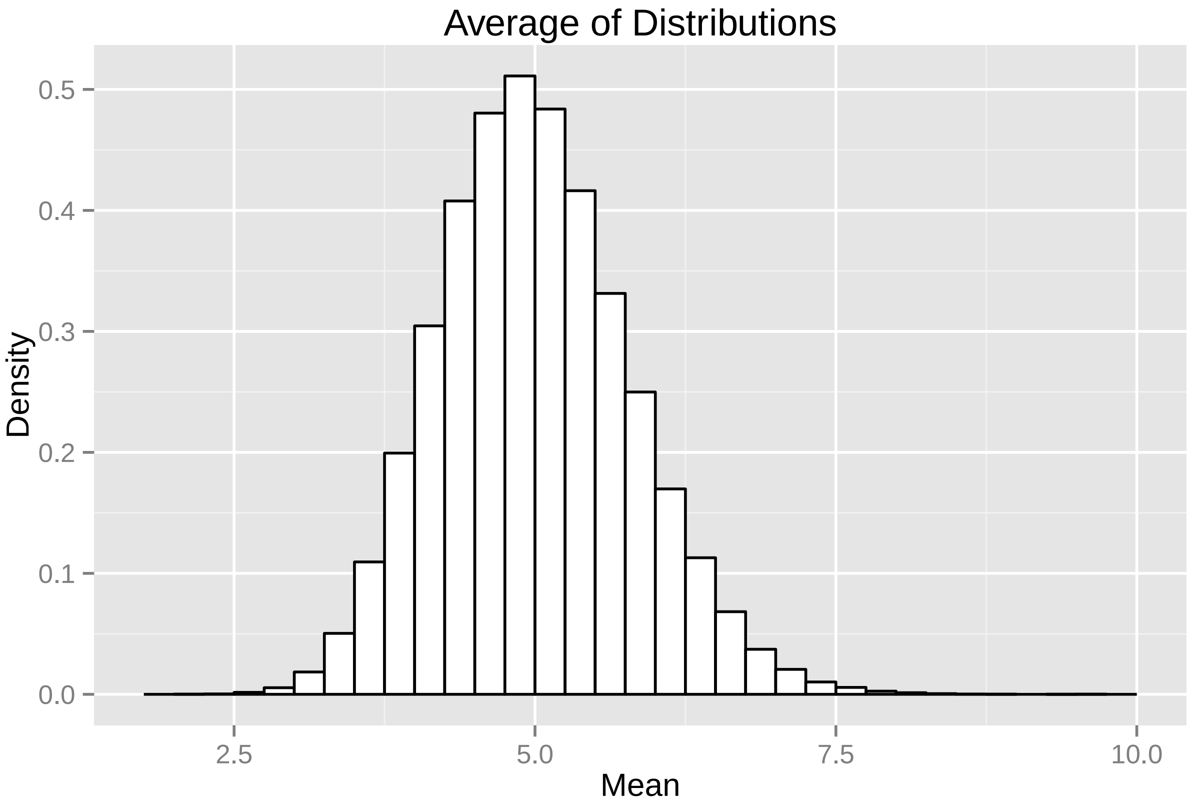
```r
##set seed for predictable random numbers (!)
set.seed(23)

##satisfy project requirements
sims=1000
lambda=0.2
obs=40

##load and format data
dat1000 <- as.data.frame(replicate(sims, mean(rexp(obs, lambda))))
names(dat1000) <- c("Means")
```

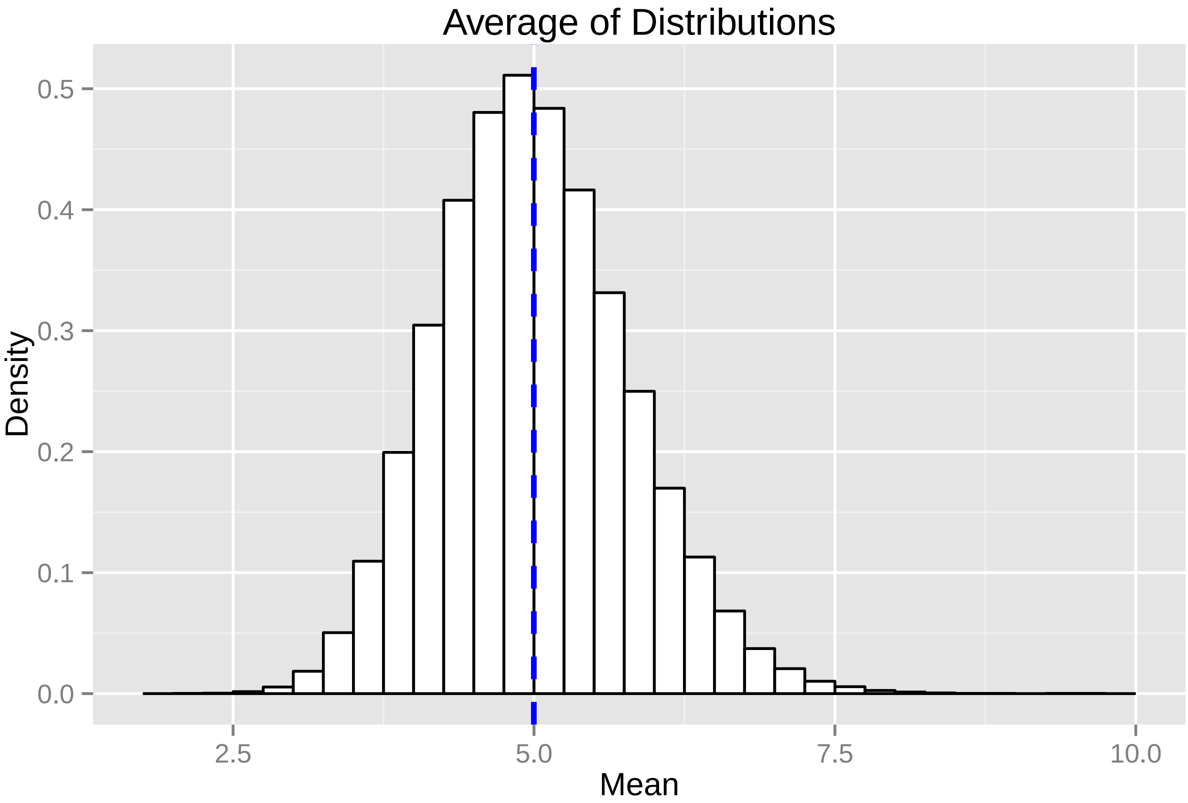
With the data in place we will now plot the distribution as a histogram using the following code:

```r
library(ggplot2)
g <- ggplot(dat, aes(x=Means)) + geom_histogram(aes(y=..density..), binwidth=.25,  colour="black", fill
g <- g + ggtitle("Average of Distributions")
g <- g + xlab("Mean") + ylab("Density")
g
```



We add a line to show where the distribution is centered - very close to 5.

```
g <- g + geom_vline(aes(xintercept=mean(dat$Means)), color="blue", linetype="dashed", size=1)
g
```



Let's have a look at some other properties of the original distribution.

The variance:

```
var(dat1000$Means)
```

```
## [1] 0.6862
```

The theoretical variance is:

???

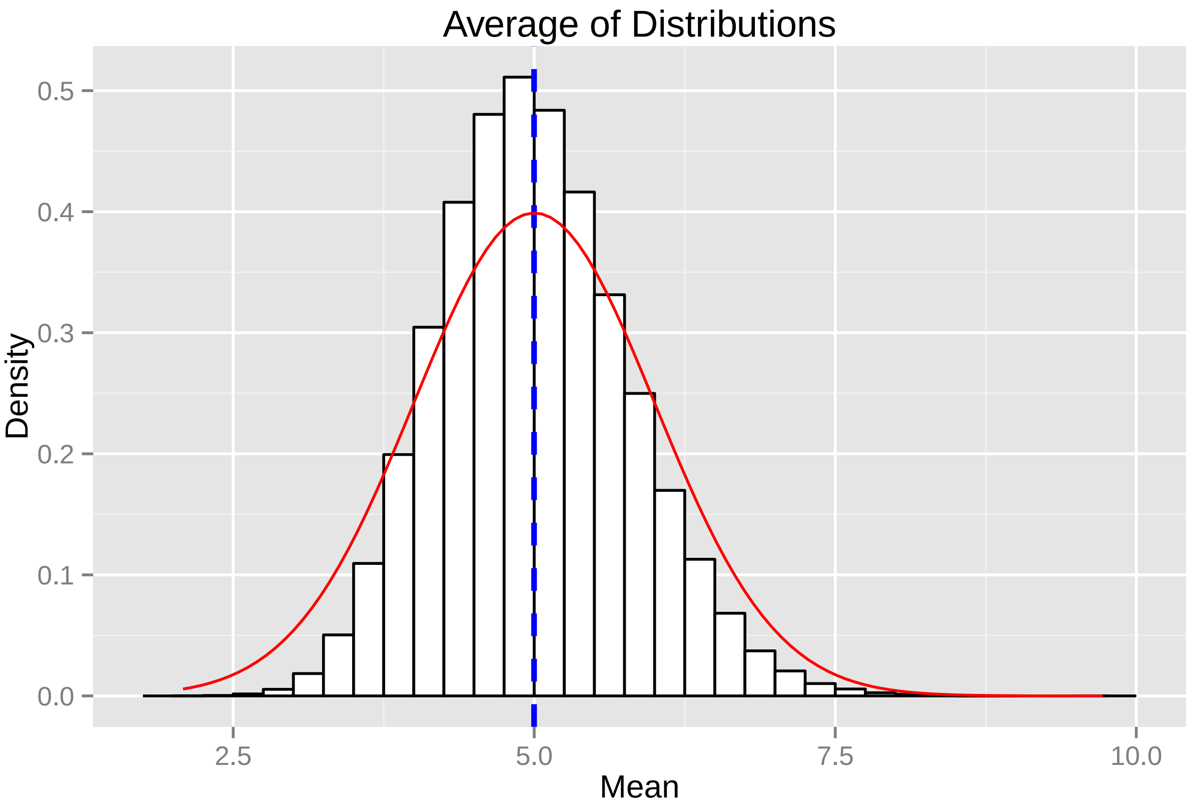
The standard deviation:

```
sqrt(mean(dat1000$Means))
```

```
## [1] 2.239
```

Next, we plot the normal distribution on top, we know the mean is 5 from 1/lambda, so we will use that (i.e. we won't *fully* standardize), and for a standard normal distribution the standad deviation is 1 (Source: Wikipedia)

```
g <- g + stat_function(fun = dnorm, colour = "red", arg = list(mean=5, sd =1))
g
```



Clearly, the distribution is very close to a normal distribution, and we have demonstrated that the Central Limit Theorem applies here.

I don't yet know how to create the 95% Confidence Level. . . .

Part 2

In this part of we're going to analyze the ToothGrowth data in the R datasets package.

Load the datasets package and do some exploratory data analysis:

```
data(ToothGrowth)
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 ...
```

```
summary(tg)
```

```
##       len        supp         dose      
##  Min.   : 4.2   OJ:30   Min.   :0.50  
##  1st Qu.:13.1   VC:30   1st Qu.:0.50  
##  Median :19.2           Median :1.00
##  Mean   :18.8           Mean   :1.17
##  3rd Qu.:25.3           3rd Qu.:2.00
##  Max.   :33.9           Max.   :2.00
```

```
names(tg)
```

```
## [1] "len"  "supp" "dose"
```

```
head(tg)
```

```
##    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
```

```
## Warning: unable to load shared object '/Library/Frameworks/R.framework/Resources/modules//R_X11.so':
##   dlopen(/Library/Frameworks/R.framework/Resources/modules//R_X11.so, 6): Library not loaded: /opt/X
##   Referenced from: /Library/Frameworks/R.framework/Resources/modules//R_X11.so
##   Reason: image not found
```

```
sqldf("select count(*) from tg where supp='VC'")
```

```
##   count(*)
## 1       30
```

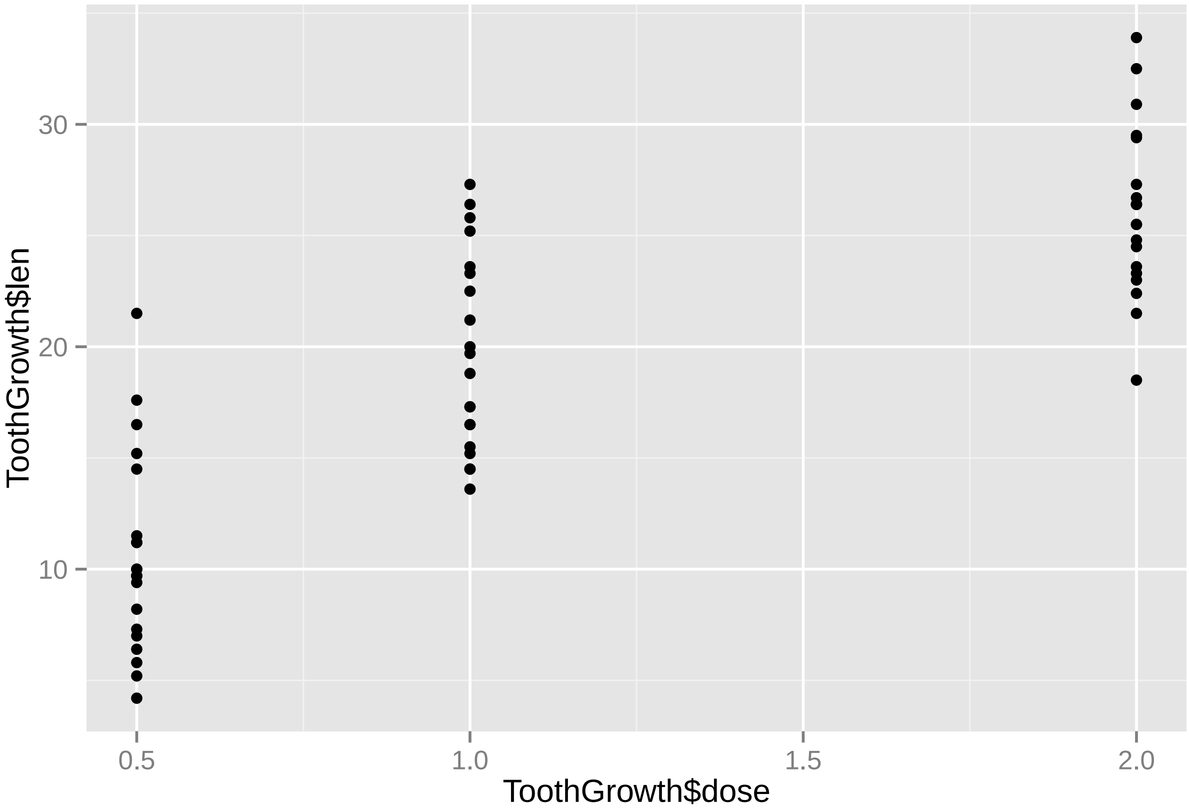
```
sqldf("select count(*) from tg where supp='OJ'")
```

```
##   count(*)
## 1       30
```

The last two results are useful, we have the same sample size, so we can use a t.test and replicate the sleep study.

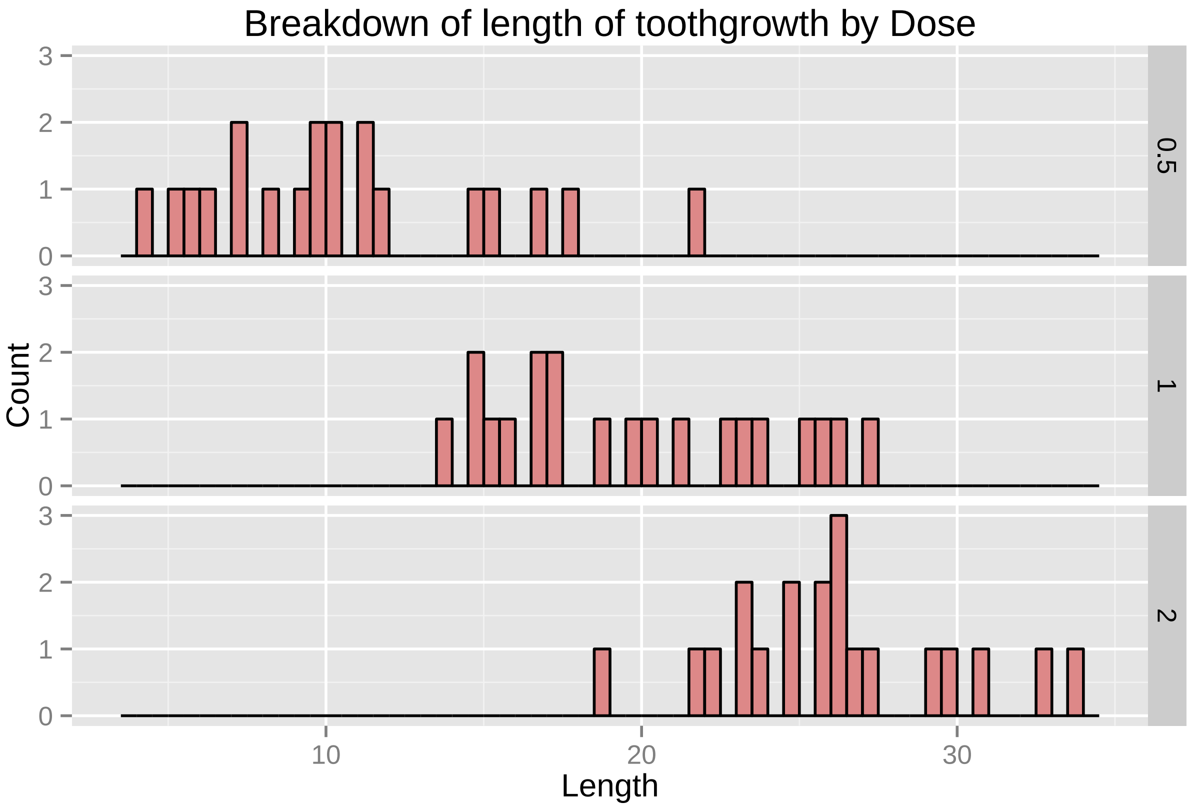
Look at dosage and growth:

```
qplot(ToothGrowth, y=ToothGrowth$len, x=ToothGrowth$dose)
```



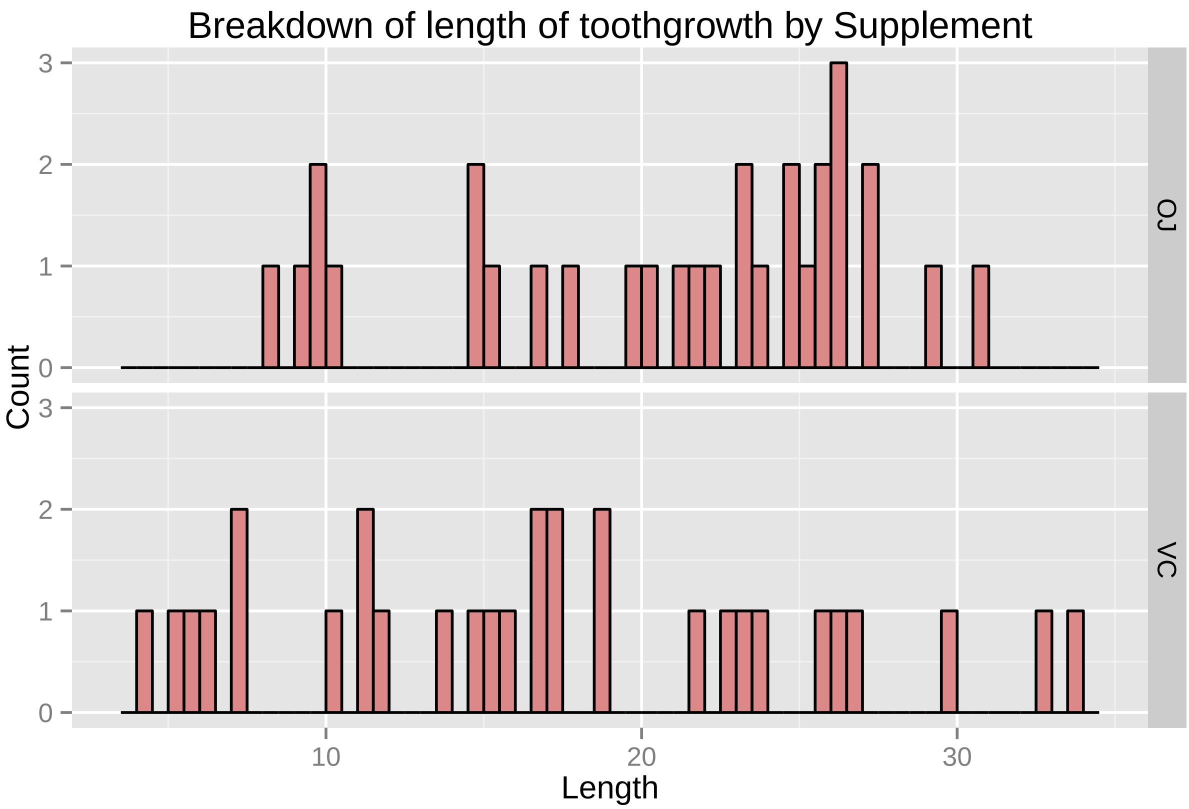
Draw facets for the distribution of length by dose:

```
g <- ggplot(data=tg, aes(x=len)) + geom_bar(colour="black", stat="bin", fill="#DD8888", binwidth=0.5) +
g <- g + ggtitle("Breakdown of length of toothgrowth by Dose")
g
```



Draw facets for the distribution of length by supplement:

```
g <- ggplot(data=tg, aes(x=len)) + geom_bar(colour="black", stat="bin", fill="#DD8888", binwidth=0.5) +
g <- g + ggtitle("Breakdown of length of toothgrowth by Supplement")
g
```



Provide a basic summary of the data:

```
summary(tg)
```

```
##       len        supp         dose      
##  Min.   : 4.2   OJ:30   Min.   :0.50  
##  1st Qu.:13.1   VC:30   1st Qu.:0.50  
##  Median :19.2           Median :1.00  
##  Mean   :18.8           Mean   :1.17  
##  3rd Qu.:25.3           3rd Qu.:2.00  
##  Max.   :33.9           Max.   :2.00  
```

In simple terms, it is a small dataset of two supplements at different dosages and an documents tooth growth on these two dimensions. Exploratory data analysis suggests that dosage or the supplement may be influential.

Let us create a null hypothesis H0: Dosage has no influence on tooth growth. We then use the t.test() function to evaluate:

```
t.test(tg$len, tg$dose)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  tg$len and tg$dose
## t = 17.81, df = 59.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  15.66 19.63
## sample estimates:
## mean of x mean of y
##    18.813     1.167
```

With such a low p-value, the null hypothesis must go.

Let us create another null hypothesis H0: The supplement has no influence on tooth growth.

```
t.test(len ~ supp, data=tg)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  len by supp
## t = 1.915, df = 55.31, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.171  7.571
## sample estimates:
## mean in group OJ mean in group VC
##            20.66            16.96
```

The p-value is higher than 0.05 - suggesting this hypothesis holds; just.

Conclusions