Statistical Inference: a look at the exponential distribution data

Larry Colucci

The first part of this report looks at the exponential distribution; we will take the average of 40 samples and look at whether or not that distribution is vaguely normal, and compare that with the theoretical distribution. The second part will do basic exploration of the ToothGrowth data, which has information on tooth growth in guinea pigs based on six treatments (Orange Juice vs Vitamins, and dose levels of .5, 1.0, and 2.0)

Simulations

We look at the exponential distribution with lambda = 0.2; for this we know the theoretical mean is 1/lambda (5), and the standard deviation is also 1/lambda. We set up a simulation which will take the average of 40 samples, and simulate 1000 times. This is done with plyr and the rexp

```
library(plyr)
#lambda is the rate parameter

#mean of exponential distribution is 1/lambda
#standard deviation is 1/lambda
lambda = 0.2
sd = 1/lambda
n = 40
nsim = 1000

df <- rdply(nsim, mean(rexp(n, lambda)), .progress="none", .id = NA)
mean(df$V1)
```

```
## [1] 5.054114
```

Sample Mean versus Theoretical Mean

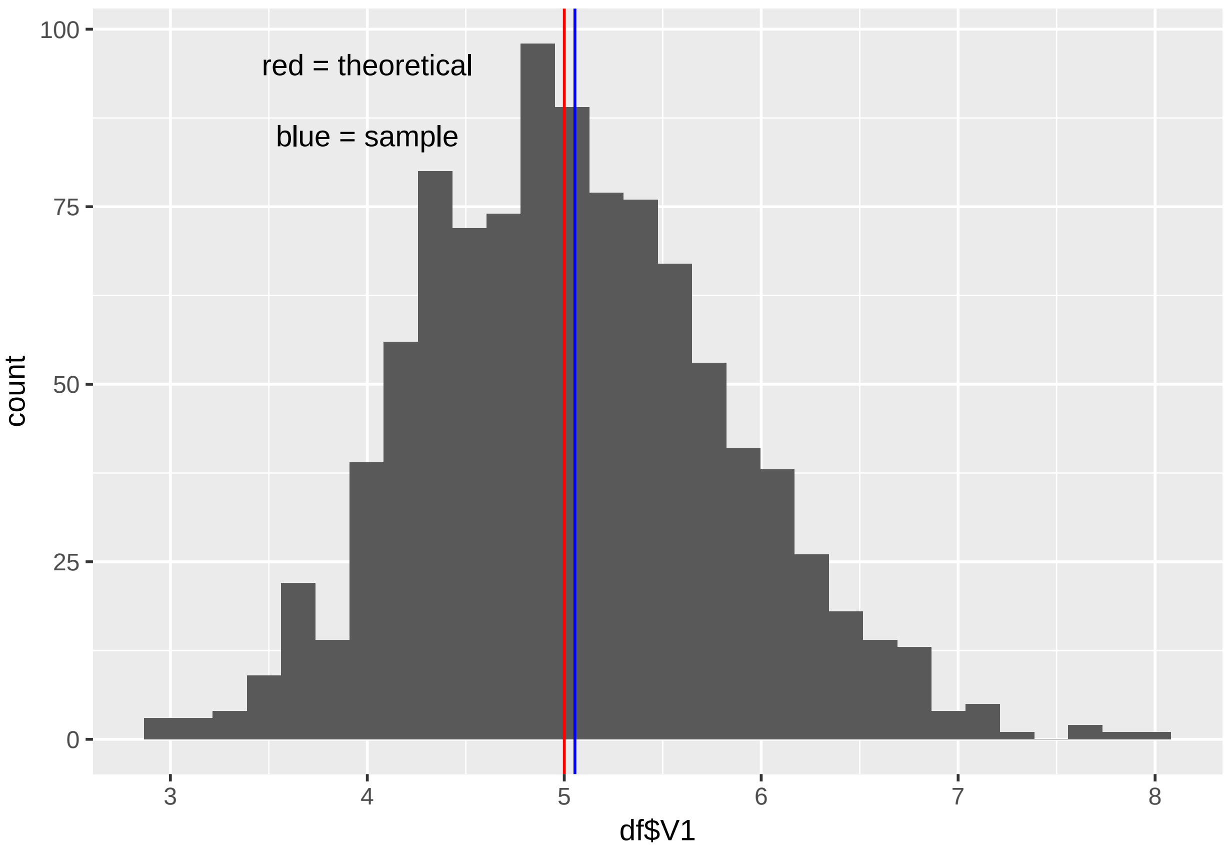
We plot the histogram of the distribution of the samples; the theoretical mean is shown in red and the sample mean is shown in blue.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
#plot
ggplot(data=df, aes(df$V1)) + geom_histogram() + geom_vline(xintercept = 1/lambda,
                                                             linetype = 1,
                                                             color = "red") +
                                               geom_vline(xintercept = mean(df$V1),
                                                            linetype = 1,
                                                            color = "blue")+
  annotate("text", x=4, y = 95, label = "red = theoretical") +
  annotate("text", x=4, y = 85, label = "blue = sample")
```

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



Sample Variance Versus Theoretica Variance

We compute the theoretical variance (var), and the sample variance (sample.var). If we subtract the sample from the theoretical, we get a difference that is small enough to be within sample error, so we can conclude that the sample is a good approximation.

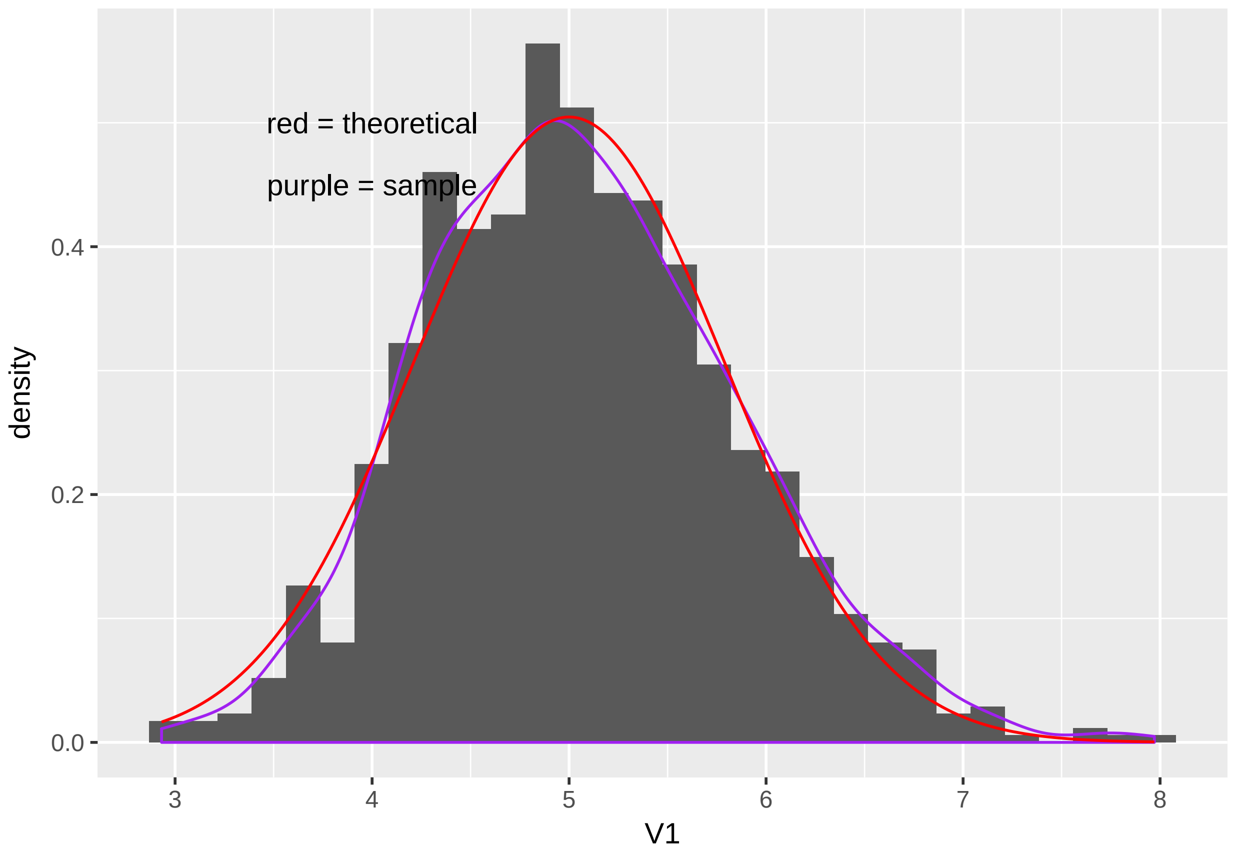
```
var <- sd^2/n

sample.var <- var(df$V1)
```

Distribution

Observationally we can see from the figure that the distribution is approximately normal; centered around 5 with approximately equal distribution on either sides of the mean. We overlay the normal distribution (red) to show how close the two are.

```
ggplot(df, aes(x=V1)) + geom_histogram(aes(y=..density..)) +
    geom_density(color = "purple") +
  stat_function(fun=dnorm, n=1000, args=list(mean=1/lambda, sd=sqrt(var)), color = "red")+
  annotate("text", x=4, y = .5, label = "red = theoretical") +
  annotate("text", x=4, y = .45, label = "purple = sample")
```

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


red = theoretical
purple = sample
density
0.0
0.2
0.4
3
4
5
6
7
8
V1