Statistical Inference Course Project

Simulation Exercise

The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also also 1/lambda. Set lambda = 0.2 for all of the simulations. In this simulation, you will investigate the distribution of averages of 40 exponential(0.2)s. Note that you will need to do a thousand or so simulated averages of 40 exponentials.

Generating data Data is generated from random exponential distribution function (`rexp`). Seeds are set for reproduceability.

```
#1. Generate 1000 (random) seeds
set.seed(100001); seeds <- sample(1:100000, 1000, replace=F)

#2. Generate 1000 samples of 40 exp(0.2) data
mns = NULL
for (i in 1 : 1000) {
  set.seed(seeds[i]); mns = c(mns, mean(rexp(40,0.2)))
}

sumstat <- summary(mns); sumstat
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.897   4.454   4.955   5.018   5.558   8.043
```

Plot Below plot depicts the distribution of 1000 means, marked mean (of the 1000 means), and normal distribution around mean.

```
library(ggplot2)

ggplot(data.frame(mns), aes(x=mns)) +
  # Histogram with density instead of count on y-axis
  geom_histogram(aes(y=..density..), binwidth=.5, colour='black', fill='white') +
  # Overlay with transparent density plot
  geom_density(alpha=.2, fill='#FF6666') +
  # Add axis labels
  xlab('Mean') + ylab('Density') +
  # Add 'mean' line
  geom_vline(aes(xintercept=mean(mns, na.rm=T)), color='red', linetype='dashed', size=1) +
  # Add with normal distribution around 'mean'
  stat_function(fun=dnorm, args=list(mean=mean(mns, na.rm=T), sd=sd(mns,na.rm=T)),
                color='blue', linetype='dashed')
```

Analysis The ***theoritical* center** of the distribution is 1/lambda = 1/0.2 = 5, and the ***actual mean* from simulation** above is 5.018.

The ***theoritical* variance** is 1/lambda = 1/0.2 = 5, and the ***actual mean* from simulation** above is `var(mns)`= 0.6338409.

The **blue dash line** in the plot shows the normal distribution with mean = 5.018 (i.e. the calculated mean). As we can see from the plot, the distribution of the means is *not very far* from normal distribution.

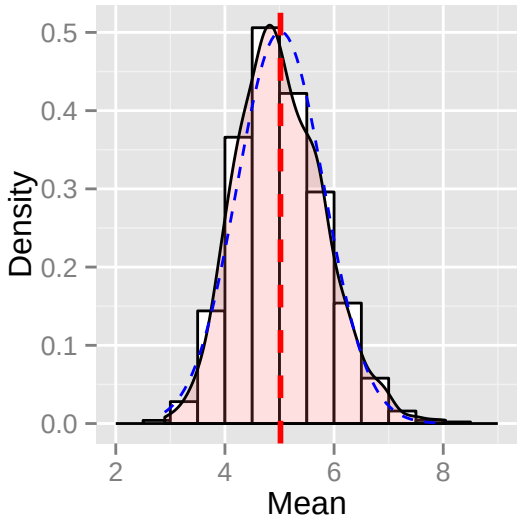


Figure 1: Mean of Random Exponential Distribution (lambda=0.2)

Analysis Exercise: The Effect of Vitamin C on Tooth Growth in Guinea Pigs

Load the ToothGrowth data & Summary of Data Data structure: [,1] **len** (*numeric*) Tooth length, [,2] **supp** (*factor*) Supplement type (VC or OJ), [,3] **dose** (*numeric*) Dose in milligrams.

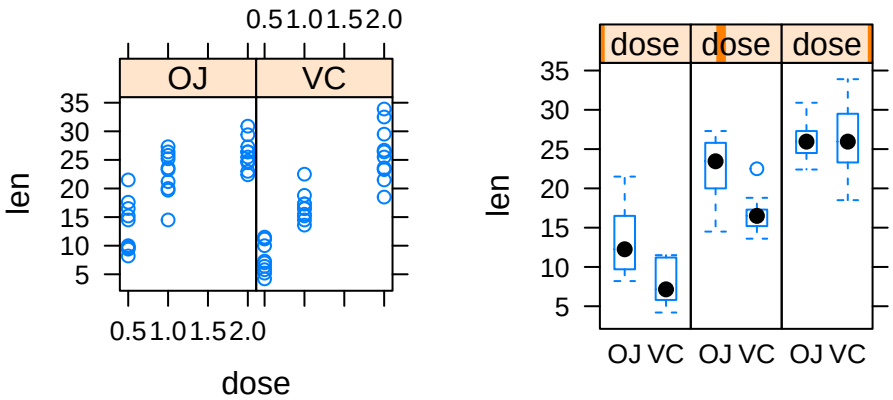
```
data(ToothGrowth)
```

```
summary(ToothGrowth)
```

```
##       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
```

Basic exploratory data analyses Explore tooth growth by supplement and dose:

```
library(grid); library(gridExtra); library(lattice)
plot1 <- xyplot(len ~ dose | supp, data=ToothGrowth)
plot2 <- bwplot(len ~ supp | dose, data=ToothGrowth)
grid.arrange(plot1, plot2, ncol=2)
```



At glance, the above table shows that higher doses leads to longer teeth for both supplements and Orange Juice (`OJ`) seems to generate longer teeth at lower dosage (0.5mg and 1mg) than arsobic acid (`VC`)

Confidence Intervals & Hypothesis Tests to compare tooth growth by supp and dose. (Only the techniques from class) Null Hypotheis *(**H0**)*: No difference between Orange Juice (`OJ`) and arsobic acid (`VC`) ; Alternative Hypothesis *(**HA**)*: There is difference.

- Dose Level = 0.5mg t-test for dose=0.5 as follow:

```
#subset the data for dose=0.5
oj05 <- ToothGrowth[which(ToothGrowth$supp=='OJ' & ToothGrowth$dose==0.5), 'len']
vc05 <- ToothGrowth[which(ToothGrowth$supp=='VC' & ToothGrowth$dose==0.5), 'len']
```

```
#test the means, dose=0.5
test05 <- t.test(oj05, vc05, var.equal=TRUE); test05$p.value;
```

```
## [1] 0.005303661
```

```
#confidence interval, dose=0.5
test05$conf.int[c(1,2)]
```

```
## [1] 1.770262 8.729738
```

- Dose Level = 1.0mg Data for dose=1.0mg is subset as per above, with value 1.0 instead of 0.5.

```
test10 <- t.test(oj10, vc10, var.equal=TRUE);  test10$p.value;
```

```
## [1] 0.0007807262
```

```
#confidence interval, dose=1.0
test10$conf.int[c(1,2)]
```

```
## [1] 2.840692 9.019308
```

- Dose Level = 2.0 Data for dose=1.0mg is subset as per above, with value 2.0 instead of 0.5 or 1.0.

```
test20 <- t.test(oj20, vc20, var.equal=TRUE);  test20$p.value;
```

```
## [1] 0.9637098
```

```
#confidence interval, dose=2.0
test20$conf.int[c(1,2)]
```

```
## [1] -3.722999  3.562999
```

Conclusions & the assumptions Based on the above results on supplement for tooth growth, at 95% confidence we reject Null Hypothesis (H0) at dose level=0.5mg and 1.0mg as the p-value is below 5%. This means, at dose=0.5mg and 1.0mg we accept that Orange Juice is more effective than asorbic acid (Vit. C). However, at dose=2.0mg, we can't reject H0 as p-value is very large; it means that there isn't enough evidence to differentiate effectiveness of Orange Juice or asorbic acid (Vit. C) at dose=2.0mg - also, CI contains 0.

The t-test performed with default values with the following inherent assumptions: Two-sided test (i.e. estimated diff in means is evaluated in both directions), 95% confidence level, Equal variance (i.e. pooled variance value is used).

References

R `Toothgrowth` dataset