Untitled

Anvita Sampathirao

6/29/2019

R Markdown

#1 Rolling

```
rolling<-function(k,r){
    rolls<- sample(1:k,r,replace=TRUE)
  return(sum(rolls))
}
n<-1000
x<- replicate(n,rolling(12,1))
y<- replicate(n,rolling(6,2))
z<- replicate(n,rolling(4,3))
Summ_stat<- data.frame(minimum= c(min(x),min(y),min(z)),
                       maximum= c(max(x),max(y),max(z)),
                       meanofdice= c(mean(x),mean(y),mean(z)),
                       stddevofdice= c(sd(x),sd(y),sd(z)),
                       row.names= c("roll1","roll2","roll3"))
Summ_stat
```

```
##       minimum maximum meanofdice stddevofdice
## roll1       1      12      6.414     3.478894
## roll2       2      12      7.053     2.392889
## roll3       3      12      7.494     1.986935
```

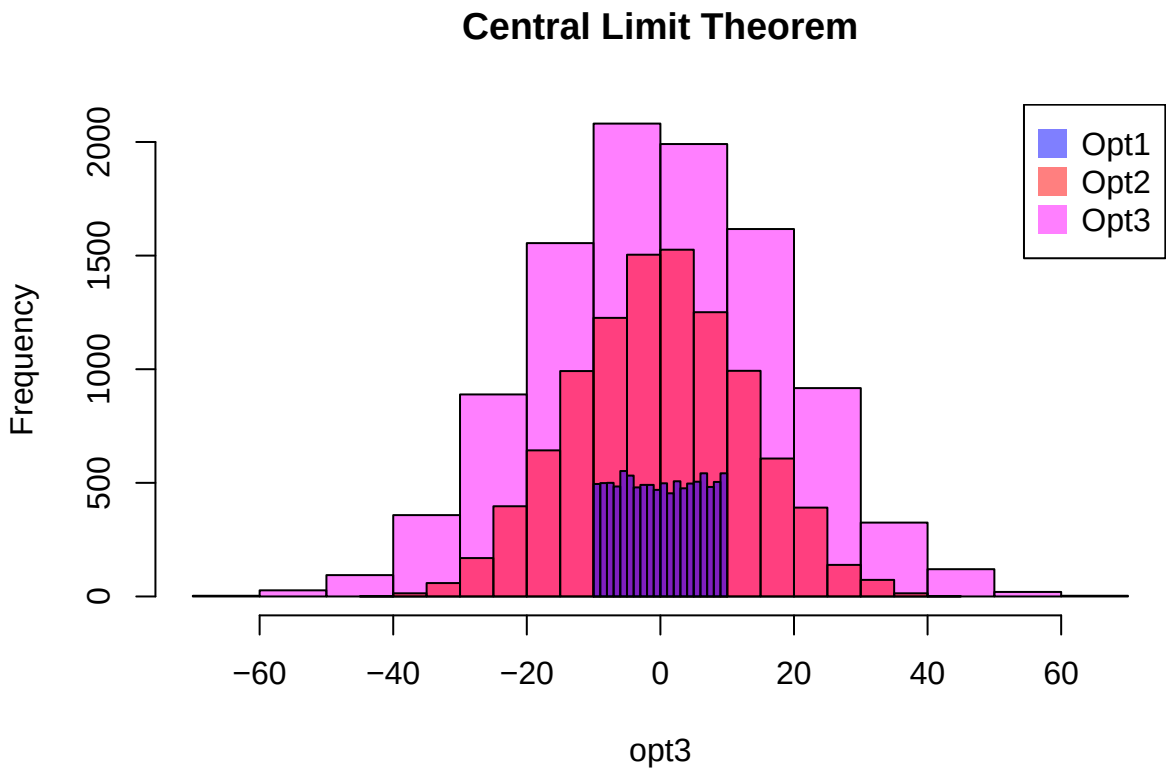
#2 Central Limit Theorem

```
library(ggplot2)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
vadd<- function(k,n){
  x_1<- rep(0,n) #Initializing null vector 1 of length n
  x_2<- rep(0,n)  #Initializing null vector 2 of length n
  for(i in 1:k){
      x_1<- runif(n,min=-10,max=10)
#Generating random values for vector 1 that follow a uniform distribution
      x_2<- x_1+x_2
#Storing the state of the vector and adding the new state of the randomly generated vector
  }
  return(x_2) #Returning the sum of k vectors of length n
}
```

```r
opt1<-vadd(1,10000)
opt2<- vadd(5,10000)
opt3<- vadd(10,10000)
hist(opt3, col=rgb(1,0,1,0.5), main="Central Limit Theorem" )
hist(opt2, col=rgb(1,0,0,0.5), add = T )
hist(opt1, col=rgb(0,0,1,0.5), add = T)
legend("topright", legend=c("Opt1","Opt2", "Opt3"),
       col= c(rgb(0,0,1,0.5),rgb(1,0,0,0.5),
              rgb(1,0,1,0.5)),
       pt.cex=2, pch=15 )
```



#3 Robocalls #3a. X=Unknown Number X'=Not an Unknown Number Y=Robocall Y'=Not a Robocall We have to find: P(Y|X) By Bayes Theorem,

$$\mathrm{P(Y|X)} = \frac{\mathrm{P(X|Y)*P(Y)}}{\mathrm{P(X)}}$$

P(X|Y)=P(Unknown Number|Robocall)=1 P(Y)= 1 robocall a day= 1/3 P(X)= 2 out of 3 calls are from unknown numbers= 2/3 Therefore,

$$P(Y|X) = \frac{1 * 1/3}{2/3} = \frac{1}{2}$$

#3b.

```r
lambda<-1
r<-2
1-ppois(r,lambda,lower.tail=TRUE)
```

```
## [1] 0.0803014
```

#4 Fuel Efficiency #4a.

$$\mu_0 = 24 \text{ (population average of previous model)}$$

$$H_0 : \mu = 24$$

$$H_1 : \mu > 24$$

#4b. We observe from the alternate hypothesis that it is a right tailed test. Thus to test the hypothesis we can use a t-statistic, which is:

$$t - stat = \frac{\bar{x} - \mu_0}{se}$$

where

$$se = \frac{sd}{\sqrt{n}}$$

Also,

$$sd = 5$$

$$\bar{x} = 27$$

$$n = 200$$

Thus,

$$\text{degree of freedom} = n - 1 = 199$$

```
serr<-5/sqrt(200)
t_Val<-(27-24)/serr
t_Val
```

```
## [1] 8.485281
```

```
uppertail<-qt(0.05,199,lower.tail = FALSE)#right tail critical value in t distribution
uppertail
```

```
## [1] 1.652547
```

Therefore, rejection region is:

$$RR : (1.65, \infty)$$

Because, our test statistic for the hypothesis lies in the rejection region of t distribution, we can reject the null hypothesis. Also, to verify:

```
pt(t_Val,199,lower.tail=FALSE)
```

```
## [1] 2.445748e-15
```

Since, p-value is less than 0.05, it is confirmed, we can reject the null hypothesis, i.e. the data is not statistically significant to determine if the new SUV model is more fuel efficient than the previous model.

```
#To verify
set.seed(1)
data_sample<- rnorm(200, mean=27, sd=5)
t.test(data_sample, mu=24)
```

```
##
##  One Sample t-test
##
## data:  data_sample
## t = 9.6738, df = 199, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 24
## 95 percent confidence interval:
##  26.52994 27.82546
## sample estimates:
## mean of x
##   27.1777
```

#5 SAT #5a.

$$n_{\text{NJ}} = 100$$
$$\bar{x_{\text{NJ}}} = 58$$
$$sd_{\text{NJ}} = 8$$

To calculate a confidence interval:

$$CI_{(\mu_{\text{NJ}})} : [(\bar{x_{\text{NJ}}}) \pm t_c * se]$$

```
xbarNJ<-58
stderr<-8/sqrt(100) #standard error of New Jersey
alpha<-0.05
lowert<- qt(alpha/2,99)
uppert<- qt(1-alpha/2,99)
lowerboundCI<-xbarNJ+(lowert*stderr)
upperboundCI<-xbarNJ+(uppert*stderr)
c(lowerboundCI,upperboundCI) # 95% Confidence interval for New Jersey mean score
```

```
## [1] 56.41263 59.58737
```

Thus, the 95% confidence interval for the mean score of all third grade New Jersey students is

$$[56.41263, 59.58737]$$

#5b.

$$n_I = 200$$
$$\bar{x_I} = 62$$
$$sd_I = 11$$

To calculate a confidence interval:

$$CI_{(\mu_{\text{diff}})} : [(\bar{x_I} - \bar{x_{\text{NJ}}}) \pm t_c * se_{\text{diff}}]$$

```
xbardiff<-62-58 #difference in mean scores of Iowa & New Jersey
stderrb<-11/sqrt(200) #std error of Iowa
stderrdiff<-sqrt(stderr^2+stderrb^2)
#combined std error of new jersey and iowa
alpha1<-0.10
dfdiff<- (stderrdiff^4)/((stderr^4/99)+(stderrb^4/199))
#combined degree of freedom of new jersey and iowa
lowertb<- qt(alpha1/2,dfdiff)
uppertb<- qt(1-alpha1/2,dfdiff)
lowerboundCIb<-xbardiff+(lowertb*stderrdiff)
upperboundCIb<-xbardiff+(uppertb*stderrdiff)
c(lowerboundCIb,upperboundCIb)
```

```
## [1] 2.1581 5.8419
```

```
#90% Confidence Interval for difference in  mean scores for Iowa and New Jersey
```

Thus, the 90% confidence interval for the mean score difference between third grade Iowa students and New Jersey students is:

$$[2.1581, 5.8419]$$

#5c.

$$H_0 : \mu_{\text{NJ}} - \mu_I = 0$$

$$H_1 : \mu_{\text{NJ}} - \mu_I \neq 0$$

```
tTest<- function(h0,xbar1,xbar2,sigma1,sigma2,n1,n2,alphafun){
  mu<-h0
  xbar<- xbar1-xbar2
  se1<- sigma1/sqrt(n1)
  se2<- sigma2/sqrt(n2)
  sed<- sqrt(se1^2+se2^2)
  dof<- sed^4/((se1^4/(n1-1))+(se2^4)/(n2-1))
  t_calc<- (xbar-mu)/sed
  t_crit<- qt(1-alphafun/2,dof)
  decision<- ifelse(abs(t_calc)>=t_crit, "Reject H0", "Can't reject H0")
  output<- paste("Decision: At significance level of", alphafun,"we",decision)
  return(output)
}

#For alpha = 0.1
tTest(0, 62, 58, 11, 8, 200, 100, 0.10)
```

```
## [1] "Decision: At significance level of 0.1 we Reject H0"
```

```
#For alpha = 0.05
tTest(0, 62, 58, 11, 8, 200, 100, 0.05)
```

```
## [1] "Decision: At significance level of 0.05 we Reject H0"
```

```
#For alpha = 0.01
tTest(0, 62, 58, 11, 8, 200, 100, 0.01)
```

```
## [1] "Decision: At significance level of 0.01 we Reject H0"
```

```
#To verify
NewJersey<-rnorm(100,mean=58,sd=8)
Iowa<-rnorm(200,mean=62,sd=11)
t.test(Iowa,NewJersey,conf.level = 0.90)
```

```
##
##  Welch Two Sample t-test
##
## data:  Iowa and NewJersey
## t = 3.2454, df = 267.5, p-value = 0.001322
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  1.882679 5.779689
## sample estimates:
## mean of x mean of y
##  62.06857  58.23739
```

```
t.test(Iowa,NewJersey,conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  Iowa and NewJersey
## t = 3.2454, df = 267.5, p-value = 0.001322
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.506922 6.155446
## sample estimates:
## mean of x mean of y
##  62.06857  58.23739
```

```
t.test(Iowa,NewJersey,conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  Iowa and NewJersey
## t = 3.2454, df = 267.5, p-value = 0.001322
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  0.7685621 6.8938051
## sample estimates:
## mean of x mean of y
##  62.06857  58.23739
```

Yes, because the respective p values for significance levels 90%, 95% and 99% are less than 0.05, we can reject the null hypothesis which means, the population means for Iowa and New Jersey are different.

#6 Plants and Caffeine #6a.

```
plant_data<-read.csv("plants.csv")
a<- aggregate(plant_data$days ~ plant_data$treatment, data=plant_data, mean)
b<- aggregate(plant_data$days ~ plant_data$treatment, data=plant_data, sd)
c<- aggregate(plant_data$days ~ plant_data$treatment, data=plant_data, length)
summ_plant_1 <- merge(a,b,by.x="plant_data$treatment",by.y="plant_data$treatment")
summ_plant<- merge(summ_plant_1,c,by.x="plant_data$treatment",by.y="plant_data$treatment")
colnames(summ_plant)<- c("Groups","Mean","Std-Dev","Count")
summ_plant
```

```
##      Groups     Mean  Std-Dev Count
## 1    coffee 45.19167 17.29914   240
## 2  dietCoke 38.32245 19.61803   245
## 3 justWater 57.52713 15.02065   258
```

#6b. To test the difference between means for more than 2 groups, we test using the F statistic Groups: C: Coffee D: Diet Coke W: Water

$$H_0 : \mu_C = \mu_D = \mu_W$$

$$H_A : \text{At least one mean is different}$$

$$BetweenVariance = \sum_{i=1}^{3} \frac{n_i * (\bar{y_i} - \bar{y})^2}{g-1}$$

$$WithinVariance = \sum_{i=1}^{3} \frac{(n_i - 1) * s_i^2}{N-g}$$

$$F = \frac{BetweenVariance}{WithinVariance}$$

```
G<-3
N<-sum(summ_plant$Count)
ybar<-mean(summ_plant$Mean)
BV<- (summ_plant$Count %*% (summ_plant$Mean - ybar)^2)/(G-1)
WV <- ((summ_plant$Count-1) %*% summ_plant$`Std-Dev`^2 )/(N-G)

fstat<- BV/WV
fstat #Calculated F value
```

```
##          [,1]
## [1,] 79.19659
```

```
df_N<- G-1 #Numerator degree of freedom
df_D<- N-G #Denominator degree of freedom

f_Th<- qf(1-0.10,df_N,df_D)
f_Th #Threshold F value
```

```
## [1] 2.309765
```

Since, Calculated F value is greater than Threshold F value, we reject the null hypothesis which states that the population means of treatments with coffee, dietcoke and water are the same.

```
1-pf(fstat,df_N,df_D)
```

```
##      [,1]
## [1,]    0
```

As we notice that p-value is approximately 0 and less than 0.05, we can thus confirm that we can reject the null hypothesis at a significance level of 90%.

To verify

```
aov.ex<- aov(plant_data$days ~ plant_data$treatment, data=plant_data)
summary(aov.ex)
```

```
##                         Df Sum Sq Mean Sq F value Pr(>F)
## plant_data$treatment   2  47792   23896   79.15 <2e-16 ***
## Residuals            740 223415     302
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#7 Memory and Language Let, X= Ability to speak more than one language Y=Memory

$$H_0 : P(X\&Y) = P(X) * P(Y)$$

$$H_1 : P(X\&Y) \neq P(X) * P(Y)$$

To test the independence of 2 categorical variables, we use use the chi square test.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where O is observed frequency E is the expected frequency

$$E = \frac{(RowTotal)(ColumnTotal)}{OverallTotal}$$

Also, Degree of freedom is:

$$df = (R-1)(C-1)$$

```
Observed<- data.frame(Monlingual=c(10,58,12),
                      Atleast_Bilingual=c(10,7,3),
                      row.names=c("About Avg Memory",
                                  "Avg Memory",
                                  "Below Avg Memory"))
Observed
```

```
##                   Monlingual Atleast_Bilingual
## About Avg Memory          10                10
## Avg Memory                58                 7
## Below Avg Memory          12                 3
```

```
rowtotal<-rowSums(Observed)
coltotal<-colSums(Observed)
Ovralltotal<-sum(Observed)
Expected<- outer(rowtotal,coltotal)/Ovralltotal
Expected
```

```
##                   Monlingual Atleast_Bilingual
## About Avg Memory          16                 4
## Avg Memory                52                13
## Below Avg Memory          12                 3
```

```
chi_sq<- sum((Observed-Expected)^2/Expected)
chi_sq #Calculated chi square value
```

```
## [1] 14.71154
```

```r
df<- (3-1)*(2-1)
qchisq(0.95,df) #Threshold chi square value
```

```
## [1] 5.991465
```

As the calculated chi square value is greater than the Threshold chi square value, we can reject the null hypothesis.

```r
1-pchisq(chi_sq,df)
```

```
## [1] 0.0006388958
```

As p-value is less than 0.05, we can conclude that the data is not statistically significant to say that memory and ability to speak more than one language are independent. Thus, we reject the null hypothesis.

```r
#To verify
chisq.test(Observed)
```

```
## Warning in chisq.test(Observed): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Observed
## X-squared = 14.712, df = 2, p-value = 0.0006389
```