Statistical Methods - Assignment 2

Michel Mooiweer (1866761) Thomas Webbers (2560695) Eirik Kultorp (2544992)

November 2016

Exercise 2.1

a)

shift	fail rate	production proportion
early	0.015	0.4
late	0.021	0.35
night	0.024	0.25

Chance of failure is the sum of the products of each shift's production proportion and fail rate: `0.40*0.015 + 0.35 * 0.021 + 0.25 * 0.024 = 0.01914 ~ 0.019`

b)

Bayes' theorem states that `P(A|B)=(P(B|A)*P(A))/P(B)`. Adapted for our context, we have

```
P(night|flaw) = (P(flaw|night)*P(night))/P(flaw)
              = (0.024*0.25)/0.01914
              = 0.3134796238244514
              ~ 0.313
```

Exercise 2.2

a)

value	p
1	0.6
0	0.4

b)

value	p
0	0.16
1	0.24
1	0.24
2	0.36

c)

The expected number of heads in one coin toss is the sum of `value * p` over all rows. `(0 * 0.4) + (1 * 0.6) = 0.6`

d)

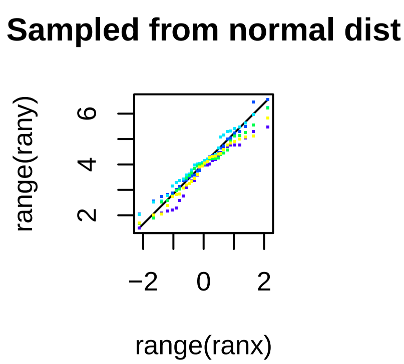
The formula for standard deviation is: `sqrt(1*0.6*(1-0.6)) = 0.489 ~ 0.49`

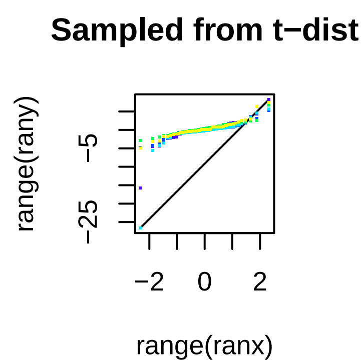
e)

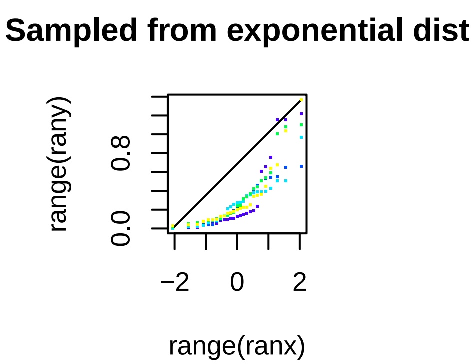
Consider a million tosses as a large value of n. Mean `1000000 * 0.6 = 600000` heads so the mean number heads per coin toss is `600000/1000000 = 0.6` Expectation `1000000 * 0.6 = 600000`. The standard deviation is: `sqrt(1000000*0.6(1-0.6)) = 490`

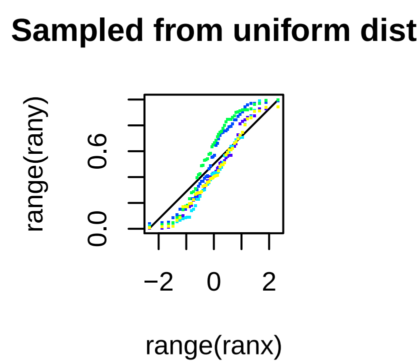
Exercise 2.3

a)



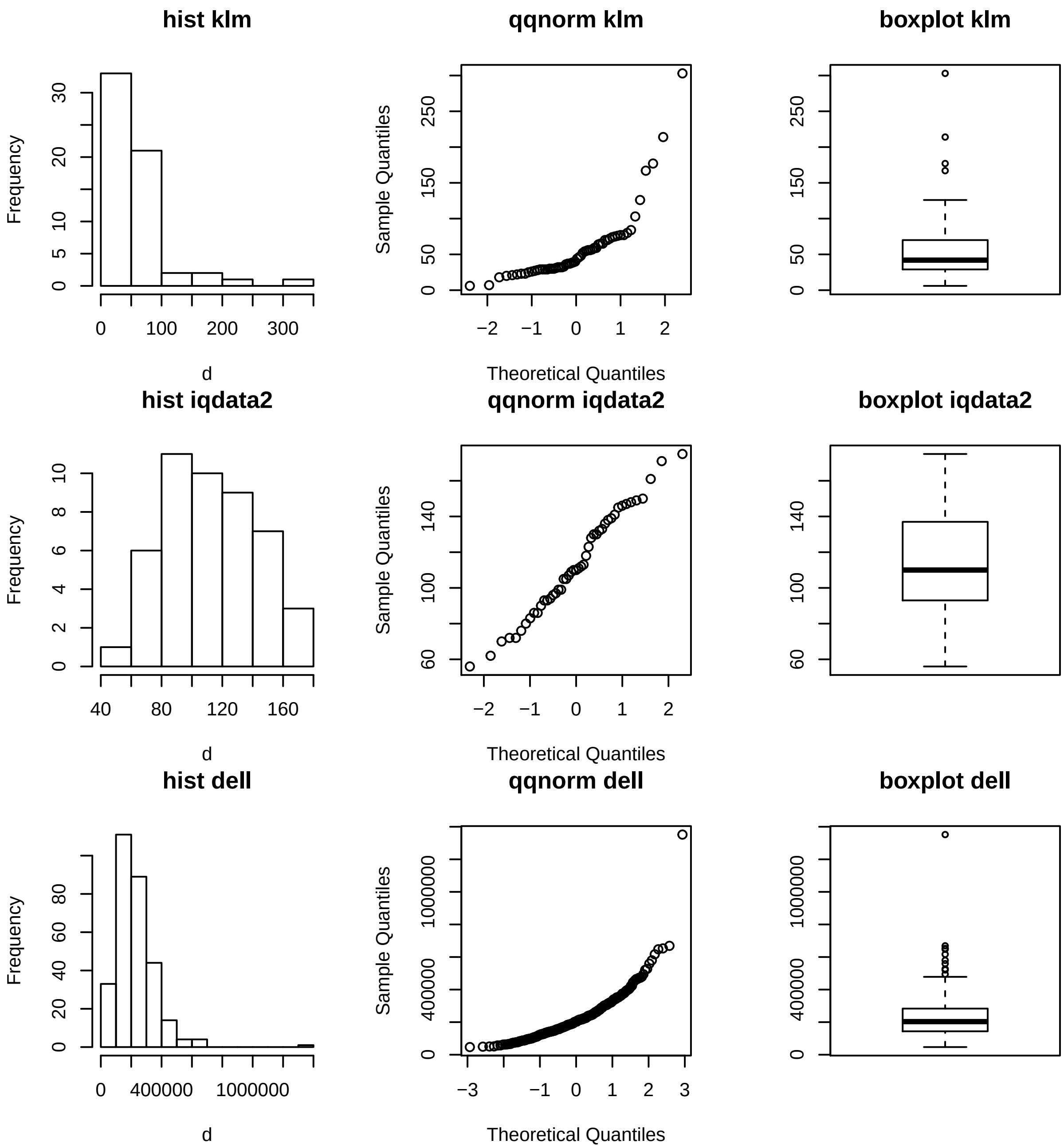


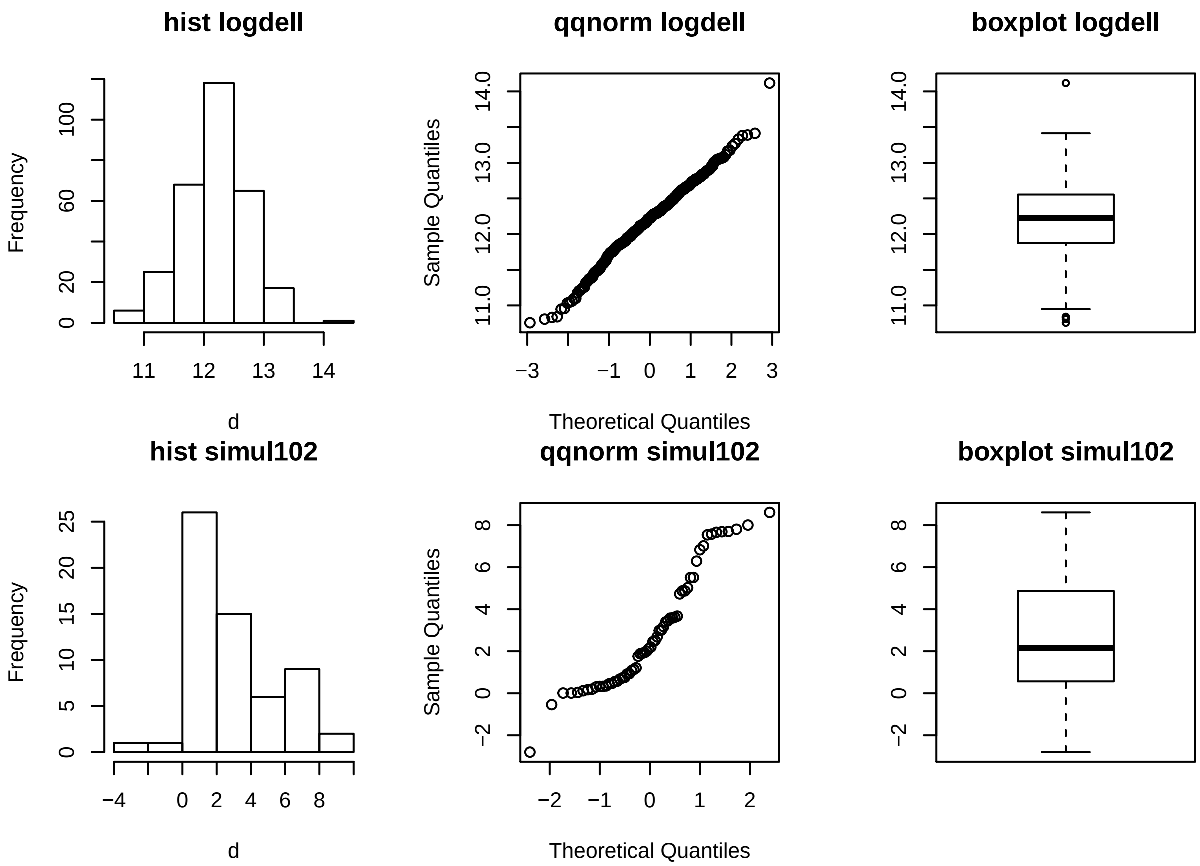




Because of the small sample sizes, we repeat the sample drawings many times, so we can account for randomness in our analysis. We see that, as expected, the samples drawn from a normal distributions tend to be approximately normal, samples drawn from a t-distribution tend to be long heavy-tailed, the exponential sample is of course right-skewed, and the samples from uniform distributions is light-tailed.

b)

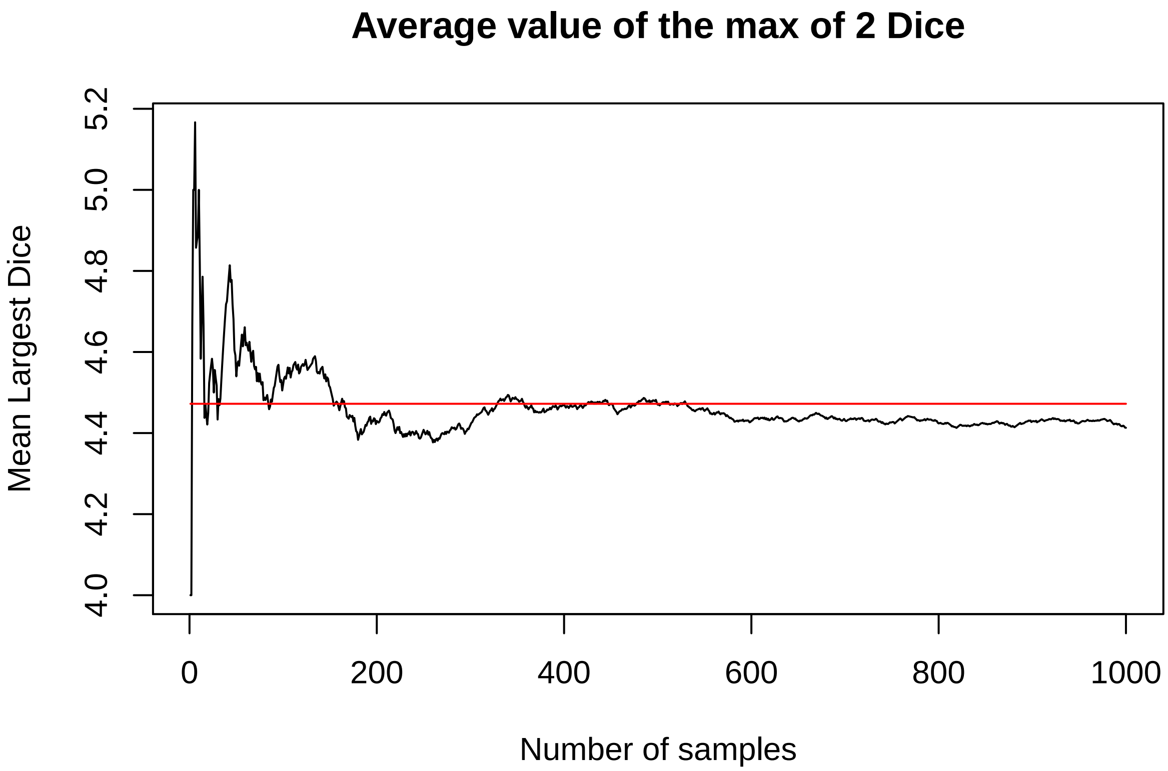

hist klm
Frequency
0
5
10
20
30
0
100
200
300
d
qqnorm klm
Sample Quantiles
0
50
150
250
−2
−1
0
1
2
Theoretical Quantiles
boxplot klm
0
50
150
250
hist iqdata2
Frequency
0
2
4
6
8
10
40
80
120
160
d
qqnorm iqdata2
Sample Quantiles
60
100
140
−2
−1
0
1
2
Theoretical Quantiles
boxplot iqdata2
60
100
140
hist dell
Frequency
0
20
40
60
80
0
400000
1000000
d
qqnorm dell
Sample Quantiles
0
400000
1000000
−3
−1
0
1
2
3
Theoretical Quantiles
boxplot dell
0
400000
1000000



dataset	can exclude normality	cannot exclude normality	reasoning
klm	1		not straight diagonal qqnorm,skewed
iqdata		1	fairly diagonal,fairly symmetric
dell	1		skewed,has long tail on one side
logdell		1	has all the features of a normal distribution
simul102	1		not straight diagonal qqnorm

Exercise 2.4

a)



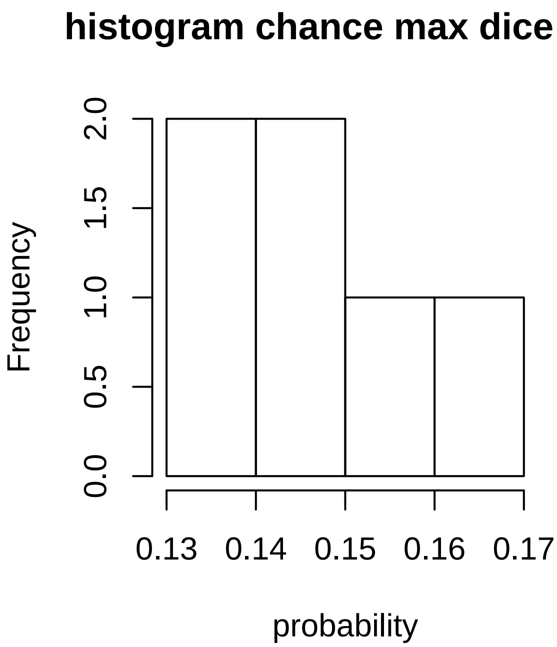
b)

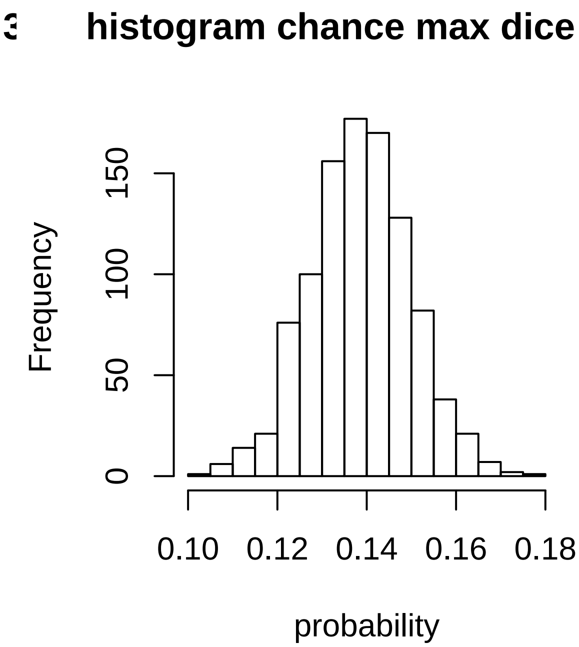
We can compute the exact expected value by taking the mean of outcomes of all combinations of rolls. The complexity of this is exponential, O(6^num_dices). With 5 dices this 20 dices this is 3656 trillion possible outcomes. That's not feasible, and so a statistical approach makes sense: we run a single trial 1000 times, and define the expected val as the mean of all the outcomes. We also include the expected value from 2 dices, so we can compare with the exact expected value that we computed in `b)`.

```
## [1] "Expected value for m=2: 4.47"
## [1] "Expected value for m=5: 5.43"
## [1] "Expected value for m=10: 5.82"
## [1] "Expected value for m=20: 5.97"
```

c)

The assignment asks for us to use at most 6 repetitions. This hides away interesting information. Above to the left is with 6 trials, and to the right is with 1000 trials. To the right we observe normality, but we can't from the left. We see that the distribution has most of its values within 0.12:0.16, indicating that the probability of getting 3 as the highest value of 2 dices has a likelyhood of approximately 0.14.





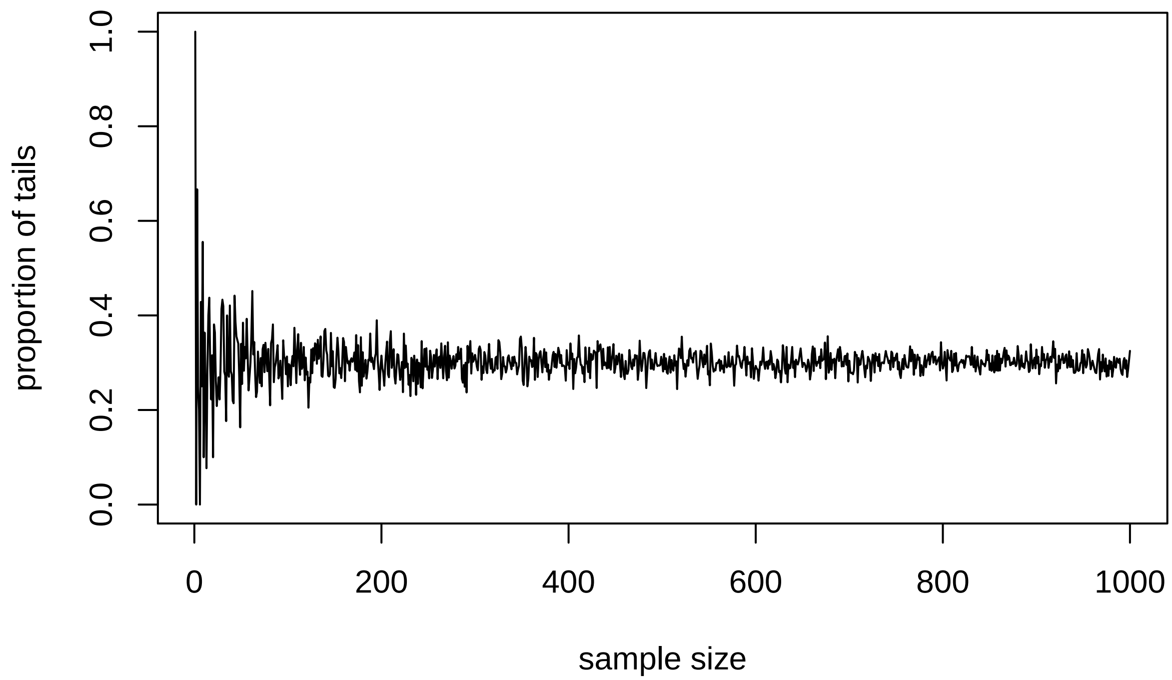
Exercise 2.5

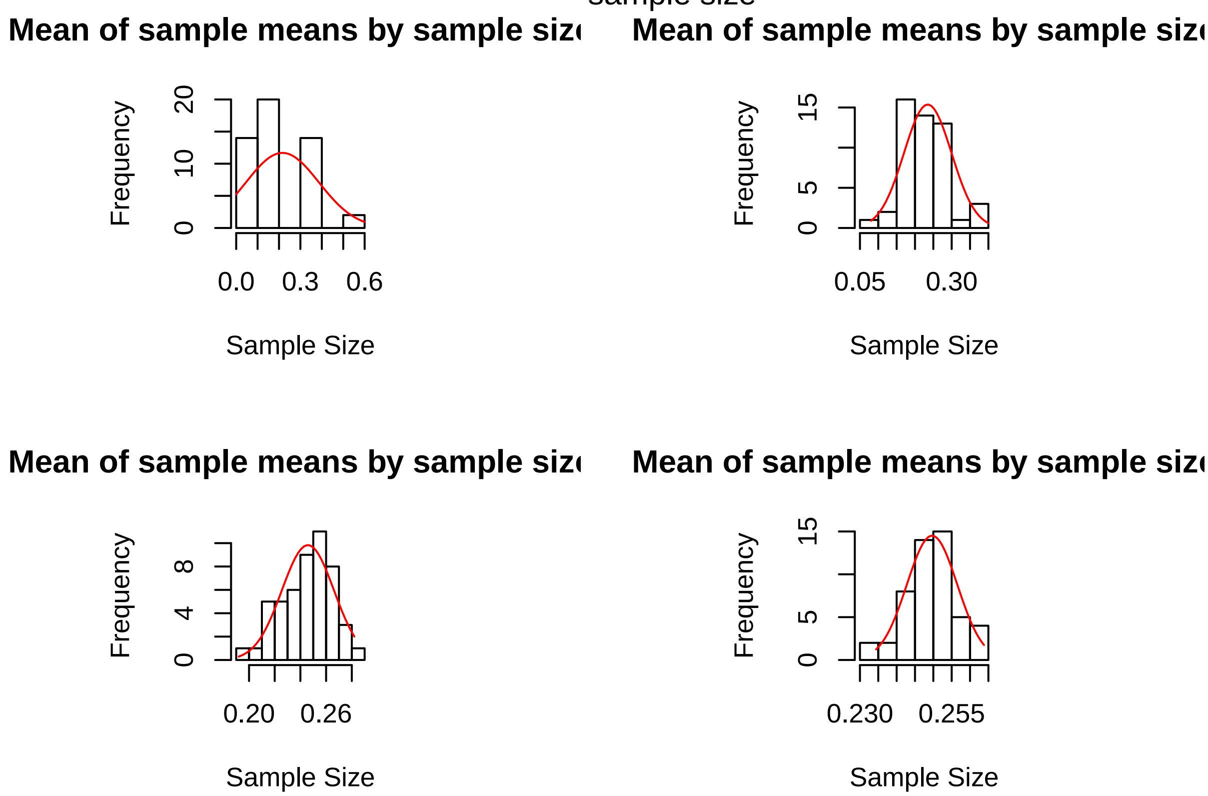
a)

Please see the appendix.

b)

Below we show that values approach some point, which is the expected value (which is the probability `p` for the desired result, which we input to the cointoss function. Below we use 0.3, so our coin is biased).





c)

According to Wikipedia,

> The Central Limit Theorem states that given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined (finite) expected value and finite variance, will be approximately normally distributed, regardless of

the underlying distribution. The theorem is a key concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions.

In other words, the sample mean approaches the population mean as the sample size increases.

The red line shows the normal distribution. The histograms show the distribution with increasing numbers of iterations. As can be clearly observed from the histograms at 2.5b) the distribution of the histograms converges with the redline of the normal distribution the larger the amount of iterations becomes.

Appendix

2.3.a

```
gend <- function(key,samples){
  res = list()
  for (i in seq(1,samples)){
    r=0
    if (key==1){r=rnorm(30,4,1)}
    else if (key==2){r=rt(50,2)}
    else if (key==3){r=rexp(25,3)}
    else if (key==4){r=runif(50,0,1)}
    res[[i]] = qqnorm(r,plot.it=FALSE)
  }
  return(res)
}

plotem <- function(qs,title){
  ranx = c()
  rany = c()
  for (q in qs){
    ranx = c(ranx,q$x)
    rany = c(rany,q$y)
  }
  cols = topo.colors(length(qs))
  plot(range(ranx), range(rany), type = "l",main=title)
  points(qs[[1]],col=cols[1],pch='.')
  for (i in seq(2,length(qs))){
    points(qs[[i]],pch='.',col=cols[i])
  }
}
titles = c("Sampled from normal dist","Sampled from t-dist","Sampled from exponential dist", "Sampled f
par(mfrow=c(2,2),pty="s")
for (i in seq(1,4)){plotem(gend(i,5),titles[i])}
```

2.3.b

```
graphical_summary <- function(d,key){
  hist(d,main=paste('hist',key))
  qqnorm(d,main=paste('qqnorm',key))
  boxplot(d,main=paste('boxplot',key))
```

```
}

keys = list("klm","iqdata2","dell","logdell","simul102")
par(mfrow=c(1,3),pty="s")
for (p in keys){
  d = scan(paste0(p,'.txt'),'r',what=double())
  graphical_summary(d,p)
}
```

2.4.a

```
source("function2.txt")

diceThrows=c()
meanMaxOfDice = 0
diceThrows = maxdice(n=1000, m = 2)
for (i in (1:1000)){
  meanMaxOfDice[i] = mean(diceThrows[1:i])
}

expected_val_max2dice <- function(){
  possibles=c()
  for (x in 1:6){for (y in 1:6){possibles[length(possibles)+1] = max(x,y)}}
  return(mean(possibles))
}

lln_plot <- function(res,expected_value=FALSE){
  plot(res,pch='.',type='l', main = 'Average value of the max of 2 Dice', ylab= 'Mean Largest Dice', xla
  if (expected_value!=FALSE){
    li=rep(expected_value,length(res))
    lines(li,col='red',type='l')
  }
}
lln_plot(meanMaxOfDice,expected_val_max2dice())
```

2.4.b

```
statistical_expected_val <- function(m,n){
  sum = 0
  for (i in seq(1:1000)){sum=sum+ maxdice(1,m)}
  v=sum/1000
  return(mean(maxdice(m,n)))
}

for (m in c(2,5,10,20)){print(paste(paste0('Expected value for m=',m,':'),round(statistical_expected_va
```

2.4.c

```
par(mfrow=c(1,2),pty='s')
thing <- function(trials){
  cc = c()
  for (i in 1:trials){cc[i]=mean(maxdice(1000,2)==3)} # the proportion of trials where the highest dice
  hist(cc, main = "histogram chance max dice is 3",xlab='probability')
}
thing(6)
thing(1000)
```

2.5.a

```
cointoss <- function(n,p){
  return(sample(c(0,1), size = n, replace = TRUE, prob = c(p, 1 - p)))
}
```

2.5.b

```
p=0.7

tosses = c()

for (n in 1:1000){
  tosses[n]=mean(cointoss(n,p))
}

plot(tosses,ylab='proportion of tails',xlab='sample size',type='line')

par(mfrow=c(2,2),pty='s')

for (n in c(5,50,500,5000)){
  means = c()
  for (i in 1:50){
    means[i]=mean(cointoss(n,0.75))
  }
  g=means
  # ty http://stackoverflow.com/questions/20078107/
  h<-hist(g, xlab="Sample Size", main="Mean of sample means by sample size")
    xfit<-seq(min(g),max(g),length=40)
    yfit<-dnorm(xfit,mean=mean(g),sd=sd(g))
    yfit <- yfit*diff(h$mids[1:2])*length(g)
    lines(xfit, yfit, col="red")
}
```