Homework1

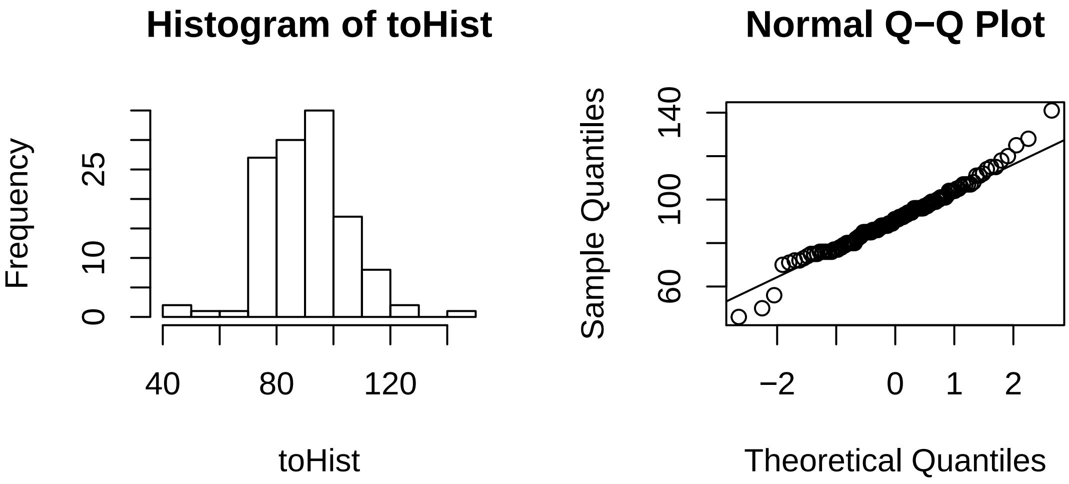
Alexander Van Roijen

January 10, 2019

Problem 1

Important note: " You might wonder why this is necessary. Isn't the coverage probability always (1-??) = 0.95? No, that is only true when the population is normally distributed (which is never true in practice) or the sample sizes are large enough that you can invoke the Central Limit Theorem. "

```
par(mfrow=c(1,2))
toHist = iqData$IQ
hist(toHist)
qqnorm(toHist)
qqline(toHist)
```



Looking at these graphs, we see it looks decently normal, with perhaps some concern for the heavy-ish tails

Problem 2

```
iqs = iqData$IQ
print(mean(iqs))
```

```
## [1] 91.08065
```

```
print(sqrt(var(iqs)))
```

```
## [1] 14.40393
```

It appears that the sample SD is not too far off from the true SD, but the average is 10% off from what is expected to be the true average IQ. There are a few explanations as to why this may be, but lets consider these two possibilities.
1) We have a biased sampling, with a bias towards students who preform worse on the iq test
2) We have too small a sample size, and thus do not accurately reflect the true population.

Problem 3

```
SET = 15/sqrt(n)
negD = -1.96 * SET
posD = 1.96 * SET
lowerB = mean(iqs)+negD
upperB = mean(iqs)+posD
print(lowerB)
```

```
## [1] 88.44045
```

```
print(upperB)
```

```
## [1] 93.72084
```

As we can see, 100 s not within this 95% confidence interval on the sample mean using a standard error calculated using the theoretical SD. This is significant, as even if we consider the possibiity of a unlucky sampling, the variance and our standard error show that we are 95% confident that the true mean is not within this interval. Since 100 is outside of this interval, we can be concerned that perhaps there is something wrong with the children we have sampled outside of random chance.

Problem 4

The equation for the 95% confidence interval is $\bar{x} \pm se * 1.96$

This means our width is $2 * se * 1.96$, which must satisfy ≤ 30

Pluggin in $se = \frac{SD}{\sqrt{n}}$ and solving for n we get $n \geq (\frac{2*SD*1.96}{30})^2 => n \geq 1.96^2$

This implies n must only be greater than or equal to 4 in order to achieve this desired width at a minimum.

Problem 5

```
SET = sqrt(var(iqs))/sqrt(n)
negD = -1.96 * SET
posD = 1.96 * SET
lowerB = mean(iqs)+negD
upperB = mean(iqs)+posD
print(lowerB)
```

```
## [1] 88.54536
```

```
print(upperB)
```

```
## [1] 93.61593
```

We can see that the confidence interval is slightly smaller in this section, as the variance and according standard deviation are smaller than what is expected.

Problem 6

```
numSamples=1000
samples = replicate(numSamples,rnorm(n,100,15))
allSEs = apply(samples,FUN=confInterval,MARGIN=2,doTrueVar = TRUE)
#print(allSEs)
res = getResults(allSEs,100)
```

```
#allSEs=lapply(allSEs, "[[", 1)
#print(allSEs[,1]$lower)
#percentOfTrue = sapply(allSEs,FUN=withinInterval,100)
print(mean(res)*100)
```

```
## [1] 94.9
```

As we can see, there are no instances in which our interval includes the "true" population mean. I chose 100 as it seemed appropriate.

Problem 7

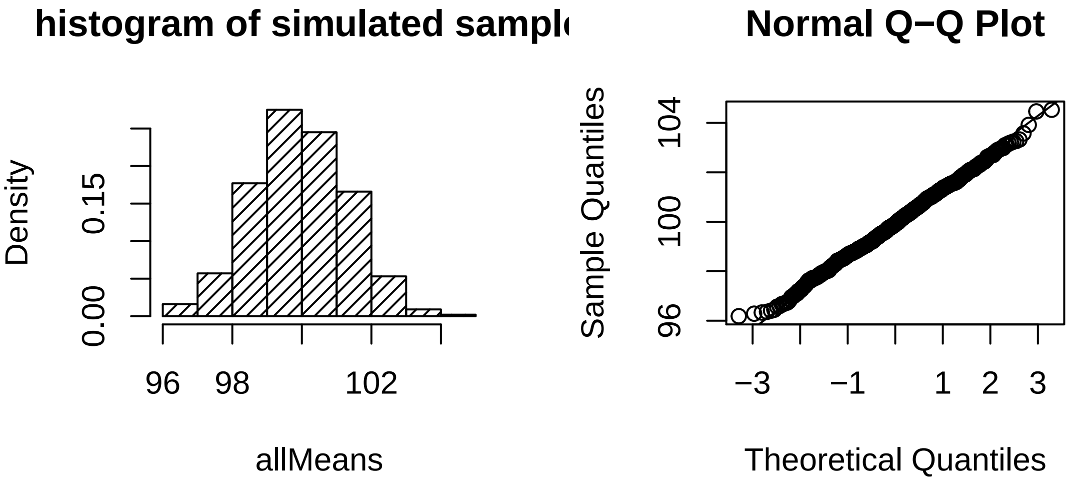
```
numSamples = 1000
samples = replicate(numSamples,rnorm(n,100,15))
allSEs = apply(samples,FUN=confInterval,MARGIN=2,doTrueVar = FALSE)
#print(allSEs)
res = getResults(allSEs,100)
#allSEs=lapply(allSEs, "[[", 1)
#print(allSEs[,1]$lower)
#percentOfTrue = sapply(allSEs,FUN=withinInterval,100)
print(mean(res)*100)
```

```
## [1] 95.2
```

As we have already noticed, the Sample variance of the data is lower than the true variance, and thus the interval is smaller, and even more likely not to hold the true mean within it.

Problem 8

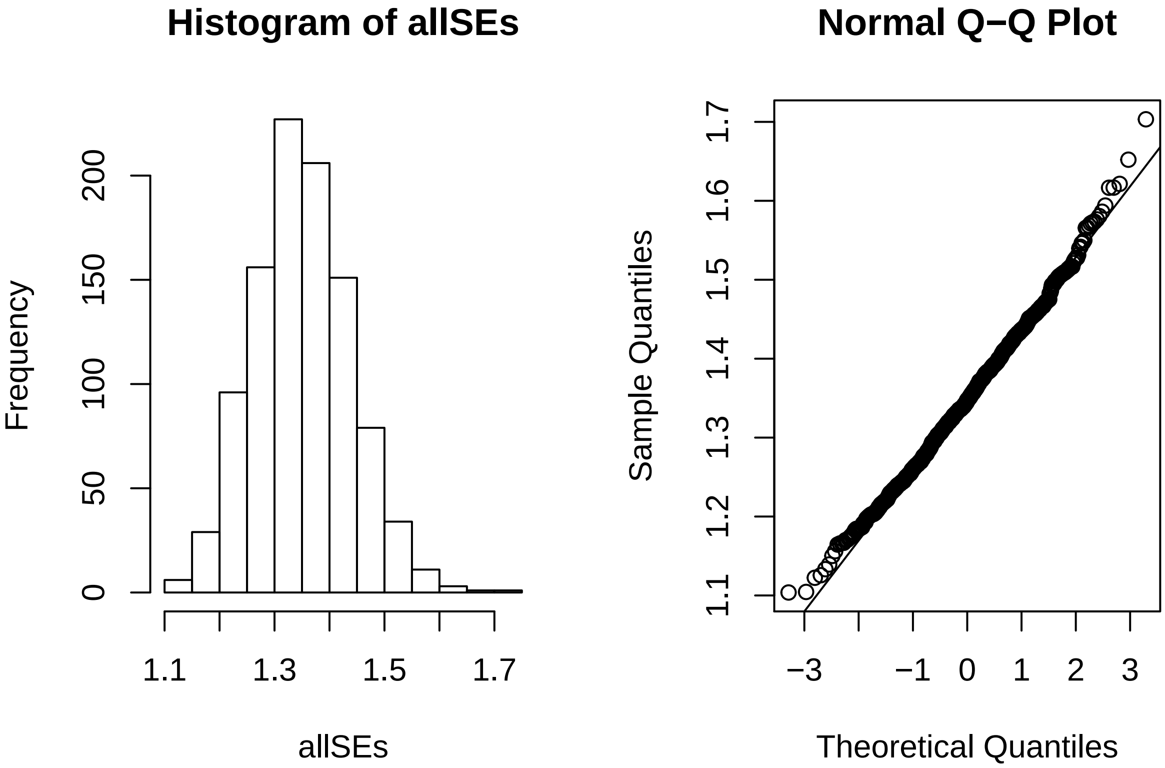
```
par(mfrow=c(1,2))
allMeans = apply(samples,FUN=mean,MARGIN=2)
hist(allMeans,density=20,prob=T,main="histogram of simulated samples")
qqnorm(allMeans)
qqline(allMeans)
```



It appears that the means are quite normally distributed, wth a mean around 100, as to be expected.

Problem 9

```r
par(mfrow=c(1,2))
allSEs = calcSEs(samples,n)
hist(allSEs)
qqnorm(allSEs)
qqline(allSEs)
```



```r
allMeans = apply(samples,FUN=mean,MARGIN=2)
print(sd(allMeans))
```

```
## [1] 1.34979
```

```r
print(mean(allSEs))
```

```
## [1] 1.350016
```

From each sample, we have a calculated standard error, looking at the mean of those standard errors, we expect that given our sample size from the population, each sampled mean should vary from each other by about 1.28 points. or at least, that is what most of the samples indicated. the standard deviation amongst the means of the samples is also indicative of this fact, representing from our samples, what is actually the case. Since they are close, we can see that our standard errors are quite indicative of the true variance between samples that is to be expected.

Problem 10

```r
numSamples=100
iqSamples = replicate(numSamples,rnorm(n,90,15))
allSEs = apply(iqSamples,FUN=confInterval,MARGIN=2,doTrueVar = T)
```

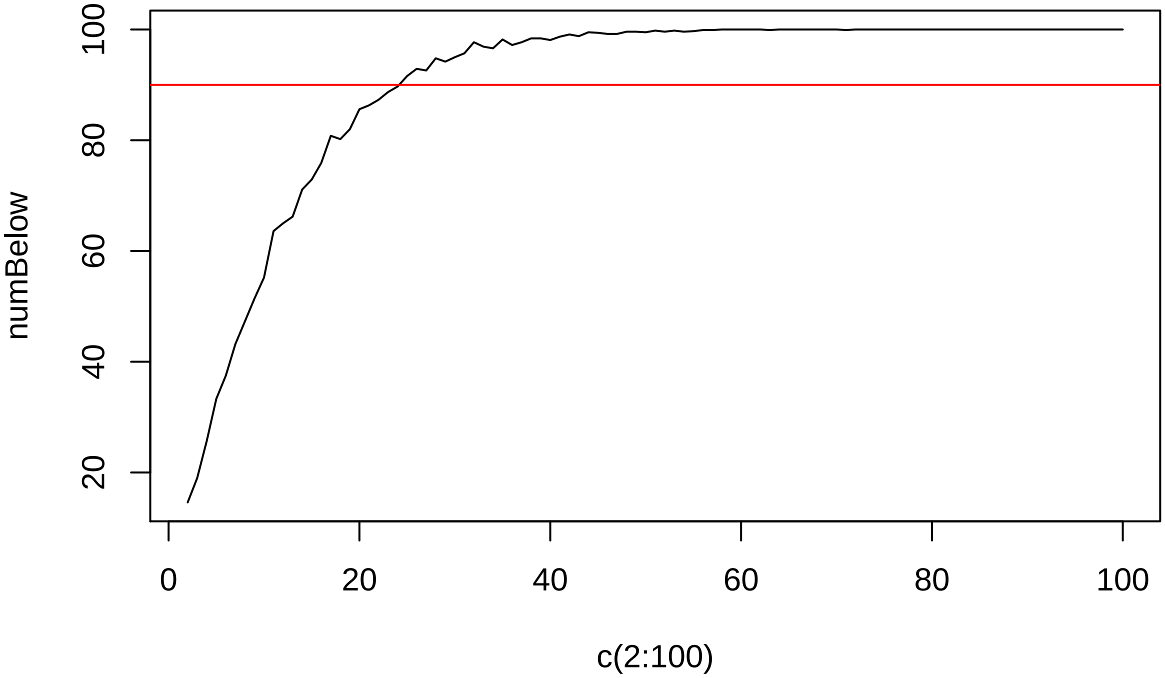
```r
res = getAltResults(allSEs,100)
percentBelow = mean(res)*100
print(percentBelow)
```

```
## [1] 100
```

Considering the size of our sample, it would take almost 8 standard erros to get up to 100 on the upper bound assuming a mean around 90. Basically, a lot would have to go wrong from this sampling to have an upper bound above 100.

Problem 11

```r
numBelow=numeric(99)
for(i in 2:100)
{
  numSamples=1000
  iqSamples = replicate(numSamples,rnorm(i,90,15))
  allSEs = apply(iqSamples,FUN=confInterval,MARGIN=2,doTrueVar = T)
  res = getAltResults(allSEs,100)
  percentBelow = mean(res)*100
  numBelow[i-1] = (percentBelow)
}
plot(x=c(2:100),y=numBelow,type="l")
abline(h=90,col='red')
```



We can study this, and we see that a sample size of about 25 or so will be satisfactor for a 90% below rate.

Problem 12

```r
count = sum(100>iqs)
estProp = count/n
```

```
estSE = sqrt(estProp*(1-estProp))/sqrt(n)
print(estProp-1.96*estSE)
```

```
## [1] 0.6826859
```

```
print(estProp+1.96*estSE)
```

```
## [1] 0.8334431
```

we see that our confidence interval does not inclue .5, which is siginifigant, as that means the proportion of students above and below 100, which would be expected to be 50/50 for a normal centered around 100, is likely not the case for this sample of students.

Problem 13

Bernoulli sampling

```
numSamples=1000
props = replicate(numSamples,rbinom(n,1,.5))
allSEs = calcPropSEs(props,n,numSamples)
allProps = apply(props, MARGIN=2, FUN = mean)
numPass = testIntervals(allProps,allSEs,numSamples,0.5)
print((numPass/numSamples)*100)
```

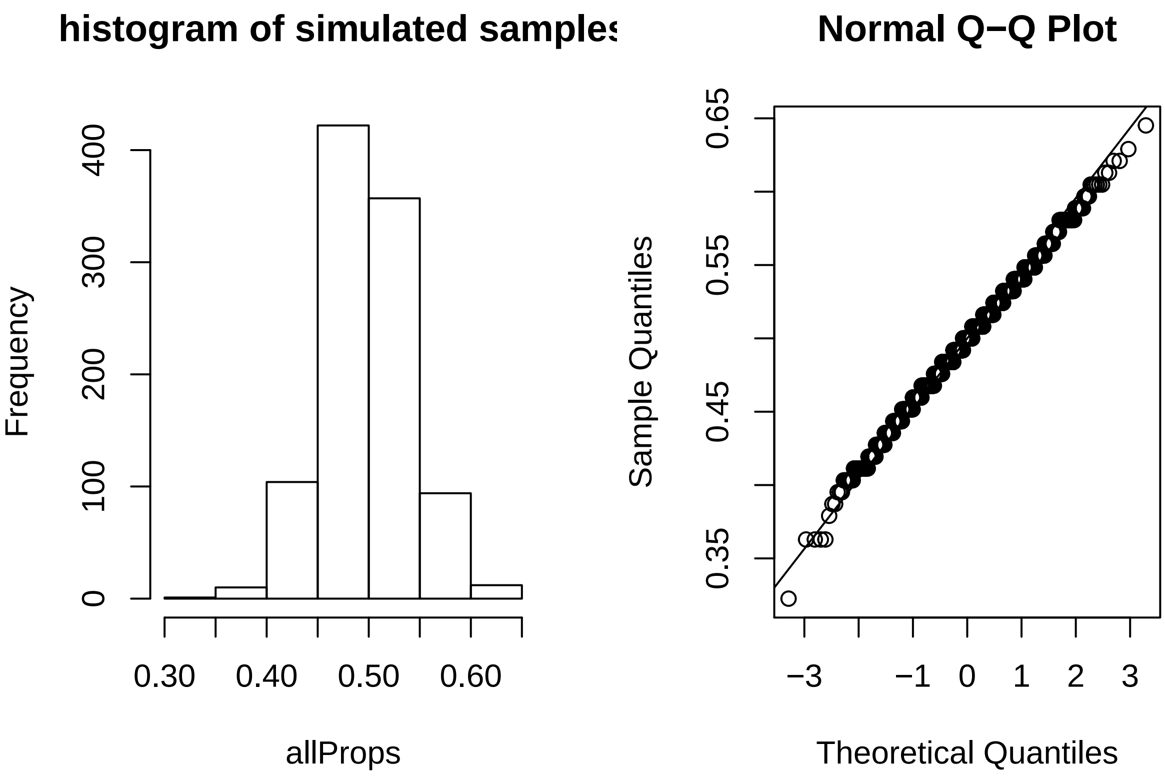
```
## [1] 94.2
```

```
#trueSEs = apply(props,margin=2,FUN=sd)
#trueSEs = calcPropSEs(props,n)
```

As we can see, we do just about achieve that 95% confidence interval, as we would expect.

Problem 14

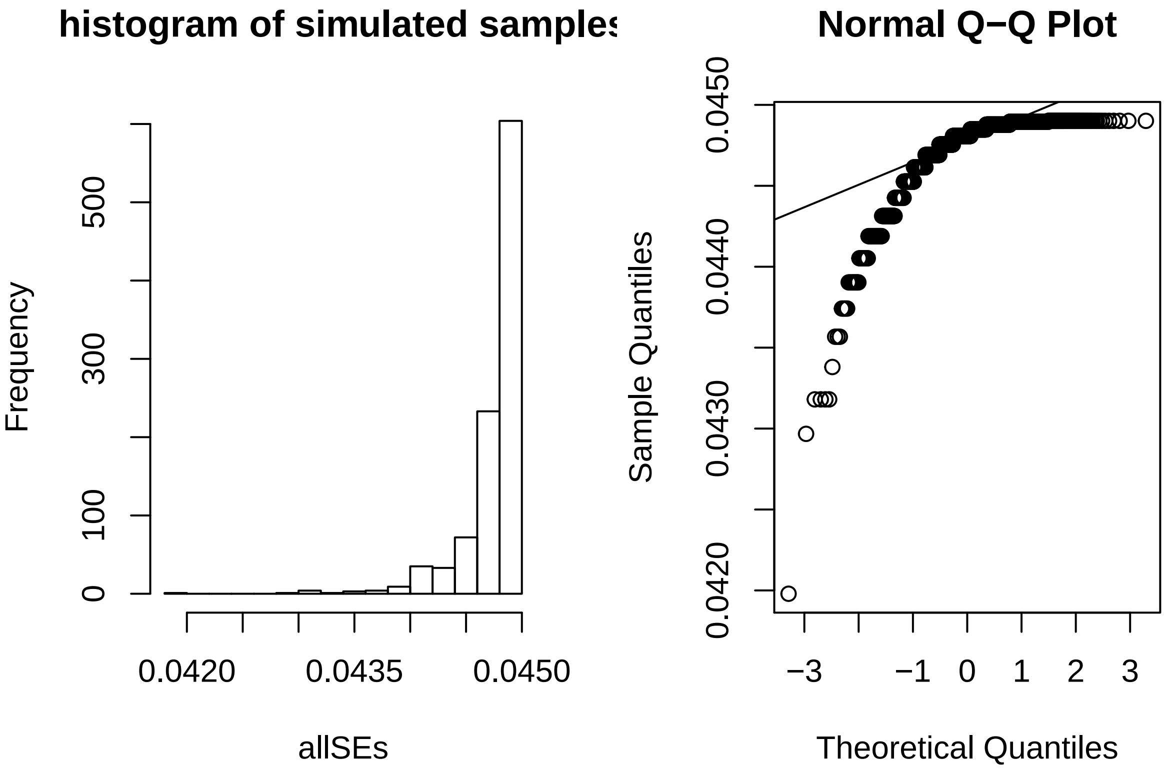
```
par(mfrow=c(1,2))
hist(allProps,main="histogram of simulated samples")
qqnorm(allProps)
qqline(allProps)
```



This is more in line with what I would expect, a quite normal looking distribution, with perhaps a bit of skew. I may want to consider increasing the number of samples done.

Problem 15

```
par(mfrow=c(1,2))
hist(allSEs,main="histogram of simulated samples")
qqnorm(allSEs)
qqline(allSEs)
```



```
print(mean(allSEs))
```

```
## [1] 0.04472477
```

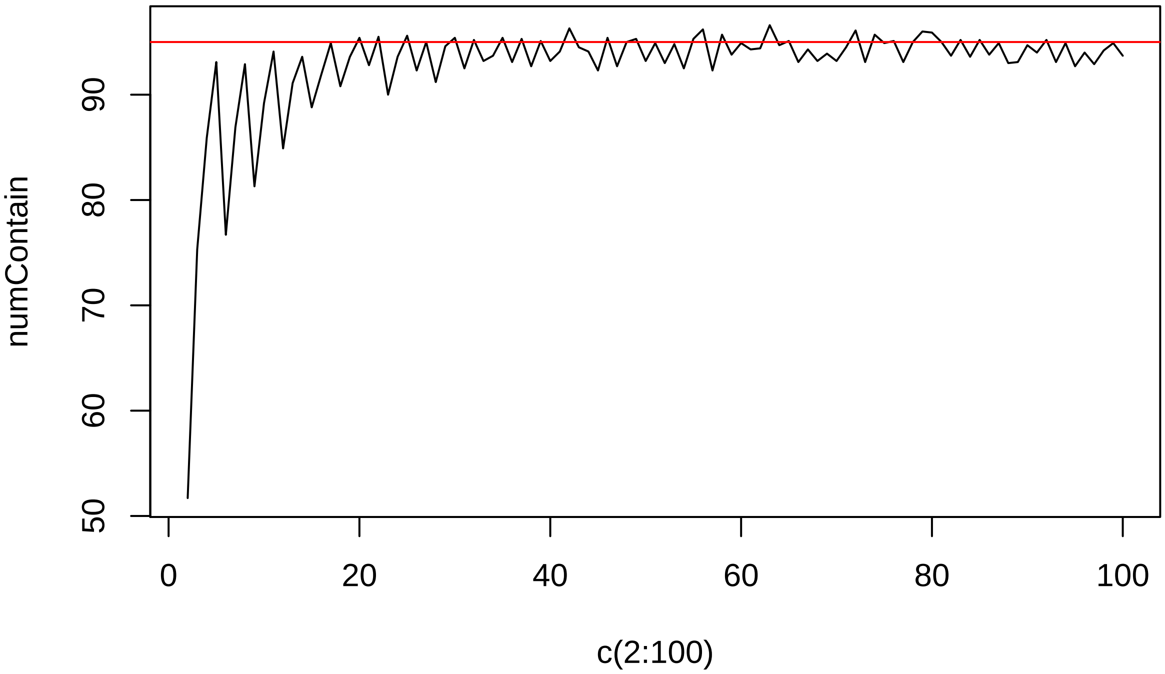
```
print(sd(allProps))
```

```
## [1] 0.04420908
```

Well, this is not what I was expecting! This is a heavily skewed distribution, and clearly not even close to normal. Reminds me of a log scale. However, the standard deviation of all the simulated proportions and the mean of the estimated SEs are quite close. Which is important, for the same reasons as listed in problem 9

Problem 16

```
numContain=numeric(99)
for(i in 2:100)
{
  numSamples=1000
  props = replicate(numSamples,rbinom(i,1,.5))
  allSEs = calcPropSEs(props,i,numSamples)
  allProps = apply(props, MARGIN=2, FUN = mean)
  numPass = testIntervals(allProps,allSEs,numSamples,0.5)
  percentBelow = (numPass/numSamples)*100
  numContain[i-1] = (percentBelow)
}
plot(x=c(2:100),y=numContain,type="l")
abline(h=95,col='red')
```



It appears that a sample size of size 30 satisifies this requirement

Problem 17

WEll we understand that the formula to calculate the standard error for a sample proportion is $\frac{\sqrt{\hat{p}*(1-\hat{p})}}{\sqrt{n}} = SE$. In order for our width, which is $2 * SE$, to have a width of .1 or less, we must satisfy $2 * \frac{\sqrt{\hat{p}*(1-\hat{p})}}{\sqrt{n}} \leq 0.1$. Solving for n, we get the following $n \geq \frac{10}{\hat{p}*(1-\hat{p})} = \frac{10}{\hat{p}-\hat{p}^2}$. If $\hat{p} = .5$ then $n \geq 40$, as n approaches 0 or 1, this number approaches inifity.