homework2

Alexander Van Roijen

January 17, 2019

Problem 1

Problem 1.1

```
vals = c(-5, -2, -1, -1, 0, 0, 2, 3, 4, 4, 5, 5, 6, 6, 11)
dft = length(vals)
sampleMean = mean(vals)
sampleSD = sd(vals)
se = sampleSD/sqrt(dft)
widthFactor = qt(0.975,dft-1)
upper = sampleMean+widthFactor*se
lower = sampleMean - widthFactor*se
print(upper)
```

```
## [1] 4.700176
```

```
print(lower)
```

```
## [1] 0.2331574
```

For this hypothesis test, we are checking the following $H_0 : \mu = 0$ if $qt(0.975, 15) = \frac{|\bar{X}-0|}{\frac{s}{\sqrt{(15)}}}$ where $s = se$ from above and $\bar{X} = sampleMean$ from above. plugging in everything we get

```
print(sampleMean/(sampleSD/sqrt(15)))
```

```
## [1] 2.368682
```

```
print(widthFactor)
```

```
## [1] 2.144787
```

We see this still agrees that we should reject the null that the mean is 0 as our CI does not contain 0 and our tabulated Z is greater than the critical value for a t distribution with a alpha of 0.05 and 15 degrees of freedom. This agrees with our solutions found in problem 2.2

Problem 1.2

```
numRepl=100
samples = replicate(numRepl,runif(15,min=-7,max=7))
allMeans = apply(samples, MARGIN=2, FUN = mean)
numOff=0
lower = -se*2.14
upper = se*2.14
for(i in 1:numRepl)
{
  if(allMeans[i]<lower || allMeans[i]>upper)
  {
    numOff = numOff+1
```

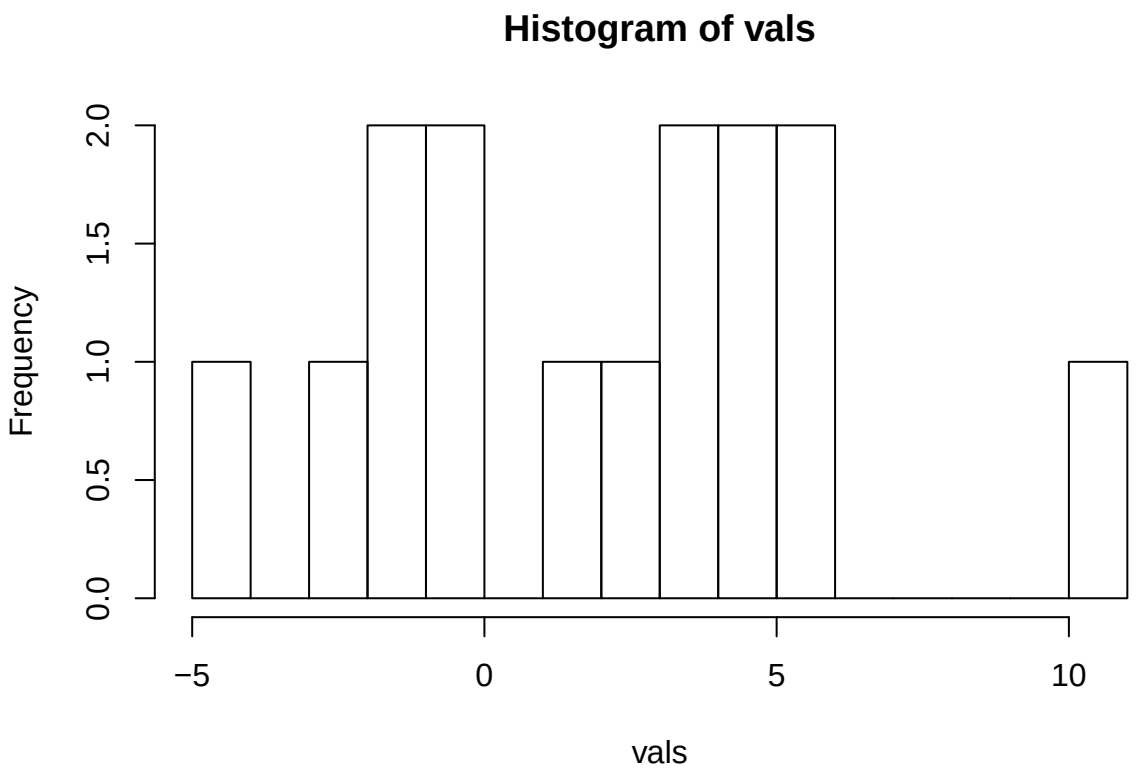
```
  }
}
print(numOff/numRepl)
```

```
## [1] 0.04
```

In this example, I used a uniform distribution. However, in the previous problem I sampled from an incorrect distribution! (it didnt satisfy the variance requirement). Accordingly, I achieved a simulated type I error of 12%, while in this one it is much lower. Consequently ,this uniform distribution is doing a better job at simulating the data sampled. We can say this test is quite more valid as it simulates a type one error much closer to the true data we have.

Problem 1.3

```
{
hist(vals,breaks=seq(-5,11,1))
ss=15
numIteration=100
allSamples = numeric(numIteration)
allSamples=replicate(numIteration,mean(sample(c(-2.5,5),size=ss,replace=TRUE)))
SE = sqrt(15/ss)
lower = -1.96*SE
upper = 1.96*SE
numOutside=0
for(i in 1:60)
{
  if(allSamples[i]>upper || allSamples[i]<lower)
  {
    numOutside=numOutside+1
  }
}
print((numOutside/100))
}
```




```
## [1] 0.13
```

I justified my sampling here using the visuals from the histogram. I "eye-balled" it to be around an even sampling from an average of -2.5 and 5. As we can see, and as was aforementioned, this was a bit aggressive and does not produce a very good type I error and is likely not a good distribution to emulate the original. However, I will say that I wasnt too far off.

Problem 1.4

```
vals = c(-5, -2, -1, -1, 0, 0, 2, 3, 4, 4, 5, 5, 6, 6)
dft = length(vals)
sampleMean = mean(vals)
sampleSD = sd(vals)
se = sampleSD/sqrt(dft)
widthFactor = qt(0.975,dft-1)
upper = sampleMean+widthFactor*se
lower = sampleMean - widthFactor*se
print(upper)
```

```
## [1] 3.816546
```

```
print(lower)
```

```
## [1] -0.1022603
```

```
print(sampleMean/(sampleSD/sqrt(15)))
```

```
## [1] 2.119488
```

```
print(widthFactor)
```

```
## [1] 2.160369
```

Clearly, if we assume the 11 was in fact an error, we see that we can no longer reject the null hypothesis, both due to removing a high value away from zero and reducing our degrees of freedom by 1 once more. Thus extending the range of our confidence interval to include 0, and our hypothesis test falling below the critical value.

Problem 2

Problem 2.1

We shall assume in this problem that the CLT applies as our sample size is quite large. In this case, a rejection region was defined as follows $\frac{|\hat{p}-.99|}{\frac{\sigma}{\sqrt{n}}} > C$ where C is our critical Value. If we were to solve for X, we would have $n * |\bar{X}| < 494$. We notice this is a one sided test. solving for $\bar{X}$ we get $\hat{p} = .988$. We plug this in to get our SE $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.004869497$. With all this calculated, and plugging back in, we get the following crticial value

```
numSample = 500
critP = 494/numSample
p_knot = .99
trueSD = sqrt((p_knot*(1-p_knot)/numSample))
ZVAL = (critP-p_knot)/trueSD
print(ZVAL)
```

```
## [1] -0.4494666
```

```
print(pnorm(ZVAL))
```

```
## [1] 0.3265476
```

Thus we can see that we have an approximate 32% probability of ahcieveing a type I error, where we reject the Null hypothesis when it is in fact true. This is overall not very good

Problem 2.2

No we want type I error very close to 0.05, so we must satisfy $P(\frac{|\hat{p}-.99|}{\frac{\sigma}{\sqrt{n}}} < -1.644854)$

```
step1 = qnorm(0.05)*trueSD
step2 = step1+.99
step3 = step2*numSample
print(step3)
```

```
## [1] 491.3404
```

Accordingly, we see that we need to see more misses in our test in order to achieve this type I error. More precisely, we need to miss 491 out of 500 samples in order to have this confidence in rejecting the null hypothesis and only being wrong about 5% of the time.

Problem 2.3

If we got 490 out of 500 samples correctly identified, we would reject this hypothesis. We can calculate the Z score below

```
numSample = 500
critP = 490/numSample
p_knot = .99
trueSD = sqrt((p_knot*(1-p_knot)/numSample))
ZVAL = (critP-p_knot)/trueSD
print(ZVAL)
```

```
## [1] -2.247333
```

```
print(pnorm(ZVAL))
```

```
## [1] 0.01230938
```

This further confirms our statement as our p value is below the alpha level of 0.05.

Problem 2.4

With a reduced sample size, we would expect to see a wider interval for our rejection region as our standard error will increase

```
numSample = 100
critP = 98/numSample
p_knot = .99
trueSD = sqrt((p_knot*(1-p_knot)/numSample))
step1 = qnorm(0.05)*trueSD
step2 = step1+.99
step3 = step2*numSample
print(step3)
```

```
## [1] 97.36339
```

```
ZVAL = (critP-p_knot)/trueSD
print(ZVAL)
```

```
## [1] -1.005038
```

```
print(pnorm(ZVAL))
```

```
## [1] 0.1574393
```

As Expected, we see that our region proportion changed from less than .982 to .973, which is wider than our previous rejection region. Thus, our .98 proportion tested in 2.3 wont fall in the rejection and we do not have enough evidence to reject the null hypothesis. As aforementioned, this is due to the decreased number of samples causing or calculated standard error to increase and thus expanding the amount of allowable values in our acceptance region to maintain the same alpha level.

Problem 2.5

The p value represents the probability that we would see a value of equal or greater in value than what we have already realized. We already calculated the values above, but to restate the calculation will be done below

```
numSample = 100
numSample2 = 500
truep = .99
samplep = 0.98
sdtop = truep*(1-truep)
```

```
sd1 = sqrt(sdtop/numSample)
sd2 = sqrt(sdtop/numSample2)
testDiffs = samplep-truep
zval1 = testDiffs/sd1
zval2 = testDiffs/sd2
print(pnorm(zval1))
```

```
## [1] 0.1574393
```

```
print(pnorm(zval2))
```

```
## [1] 0.01230938
```

As stated previously, we see that our p values are below the type I error probability and not below it resepectively on their sample size. With a large sample size, we can be equally confident with the same typer I error within a smaller acceptance region, thus allowing to reject .98. However, with a 5 times decrease in our sample size, we can no long have such a narrow interval. These changing widths in interval widths reflect on our pvalues as our probability of seeing more extreme values than .98 are directly impacted by the width of these regions. The more narrow the region, the probability of seeing the same value or greater is much lower.

Problem 3

```
iqData = read.csv('/home/bdvr/Documents/GitHub/Data557/Week1/Homework/iq.csv')
n = length(iqData$IQ)
trueSD = 15
iqs = iqData$IQ
```

We are solving $H_0 : \mu = 100$ if $qt(0.975, n) = \frac{|\bar{X}-100|}{\frac{s}{\sqrt{(n)}}}$ for $\bar{X}$ mu is not working?

Problem 3.1

```
trueMean=100
twoSides = qt(0.975,n-1)
dataSD = sd(iqs)
print(dataSD)
```

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

```
print(twoSides)
```

```
## [1] 1.979439
```

Plugging in our previous value and solving for $\bar{X}$ we get $\bar{X} > 1.97928 * \frac{14.40393}{\sqrt{124}} + 100$ and $\bar{X} < -1.97928 * \frac{14.40393}{\sqrt{124}} + 100$ we get

```
lower = trueMean- ((twoSides*dataSD)/sqrt(n))
upper = trueMean+ ((twoSides*dataSD)/sqrt(n))
print(lower)
```

```
## [1] 97.43957
```

```
print(upper)
```

```
## [1] 102.5604
```

```r
sampleMean = mean(iqs)
print(sampleMean)
```

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

```r
zval = (sampleMean - trueMean) / (dataSD/sqrt(n))
print(zval)
```

```
## [1] -6.895462
```

```r
print(2*pt(zval,n-1))
```

```
## [1] 2.486475e-10
```

SHOULD WE USE SAMPLE OR REAL STANDARD DEVIATION?

We see that our mean of iqs is outside of the rejection region at the 0.05 signifigance level and thus say we have enough evidence to reject the null hypothesis. Accordingly, our pvalue is also very very small, which makes sense as our mean is very far away from the rejection region. the probability of seeing a more extreme value is quite small.

Problem 3.2

WE are now going to assume that our sample mean is the null hypothesis in a way, and establishing our confidence interval around it and see what is the 2 sided 95% confidence interval that our true mean falls within the region.

```r
lower = sampleMean- ((twoSides*dataSD)/sqrt(n))
upper = sampleMean+ ((twoSides*dataSD)/sqrt(n))
print(lower)
```

```
## [1] 88.52022
```

```r
print(upper)
```

```
## [1] 93.64107
```

As we can see, under the 95% CI, we believe that the true mean should fall within this range, however, the mean of 100 does not fall within this range, which agrees with out previous conclusion.

Problem 3.3

By increasing the requirement of our confidence interval and similarly decreasing the signifigance level, we expect to see a wider range of our acceptance region and our confidence interval, as we are requiring the probability of us rejecting the null hypothesis when it is true to be even smaller. Similarly, for hte confidence interval, it will be larger as we want ot expand the chance of us finding the true mean within the interval.

```r
twoSides = qt(0.995,n-1)
lower = trueMean- ((twoSides*dataSD)/sqrt(n))
upper = trueMean+ ((twoSides*dataSD)/sqrt(n))
print(paste("Lower rejection region bound:" ,lower))
```

```
## [1] "Lower rejection region bound: 96.6156688598188"
```

```r
print(paste("upper rejection region bound:" ,upper))
```

```
## [1] "upper rejection region bound: 103.384331140181"
```

```r
sampleMean = mean(iqs)
zval = (sampleMean - trueMean) / (dataSD/sqrt(n))
print(paste("P value:" ,pnorm(zval)))
```

```
## [1] "P value: 2.6845019106064e-12"
```

```r
lower = sampleMean- ((twoSides*dataSD)/sqrt(n))
upper = sampleMean+ ((twoSides*dataSD)/sqrt(n))
print(paste("Lower confidence interval bound:" ,lower))
```

```
## [1] "Lower confidence interval bound: 87.6963140211091"
```

```r
print(paste("Upper confidence interval bound:" ,upper))
```

```
## [1] "Upper confidence interval bound: 94.4649763014715"
```

Looking at the output above, as exepected, the regions have expanded, but not enough to prevent us from rejecting the null hypothesis, suggesting a pvalue that is greater than the signifigance level, or having a confidence interval that does not have a high probability of including the true mean which does not contain 100.

Problem 3.4

```r
n=124
twoSides = qt(0.975,n-1)
lower = trueMean- ((twoSides*dataSD)/sqrt(n))
upper = trueMean+ ((twoSides*dataSD)/sqrt(n))
numSimulations = 500
allConclusions = numeric(500)
tempMean=0
for(i in 1:numSimulations)
{
  tempMean = mean(rnorm(n,100,15))
  #print(tempMean)
  if((tempMean < lower) | (tempMean>upper))
  {
    allConclusions[i]=1
  }
  else
  {
    allConclusions[i]=0
  }
}
print(mean(allConclusions))
```

```
## [1] 0.068
```

SHOULD THIS BE AT THE 0.05 SIGNIFIGANCE LEVEL OR 0.01?

As expected, we see that the percentage of means that fall outside of our rejection region matches very closely with the signifigance level defined.

Problem 3.5

```
twoSides = qt(0.975,n-1)
lower = trueMean- ((twoSides*dataSD)/sqrt(n))
upper = trueMean+ ((twoSides*dataSD)/sqrt(n))
numSimulations = 500
allConclusions = numeric(500)
for(i in 1:numSimulations)
{
  tempMean = mean(rnorm(n,95,15))
  if(tempMean > lower & tempMean<upper)
  {
    allConclusions[i]=1
  }
  else
  {
    allConclusions[i]=0
  }
}
print(1-mean(allConclusions))
```

```
## [1] 0.972
```

We have a power of 95.2% in the case of an alternative hypothesis of mean 95. This makes sense with our previous findings of the confidence interval and rejection region including only as low as 97, which menas many of these means will fall outside of this region, as to be expected.

Problem 3.6

```
twoSides = qt(0.975,n-1)
lower = trueMean- ((twoSides*dataSD)/sqrt(n))
upper = trueMean+ ((twoSides*dataSD)/sqrt(n))
numSimulations = 500
allConclusions = numeric(500)
altMean = 90
power=1
while(power >=0.9)
{
  for(i in 1:numSimulations)
  {
    tempMean = mean(rnorm(n,altMean,15))
    if(tempMean > lower & tempMean<upper)
    {
      allConclusions[i]=1
    }
    else
    {
      allConclusions[i]=0
    }
  }
  power=(1-mean(allConclusions))
  if(power<.99)
  {
```

```
    print(paste("Alternative Mean: ",altMean))
    print(paste("Power: ",power))
  }
  altMean=altMean+0.1
}
```

```
## [1] "Alternative Mean:  94.2999999999998"
## [1] "Power:  0.98"
## [1] "Alternative Mean:  94.3999999999997"
## [1] "Power:  0.978"
## [1] "Alternative Mean:  94.4999999999997"
## [1] "Power:  0.986"
## [1] "Alternative Mean:  94.5999999999997"
## [1] "Power:  0.978"
## [1] "Alternative Mean:  94.6999999999997"
## [1] "Power:  0.986"
## [1] "Alternative Mean:  94.7999999999997"
## [1] "Power:  0.974"
## [1] "Alternative Mean:  94.8999999999997"
## [1] "Power:  0.972"
## [1] "Alternative Mean:  94.9999999999997"
## [1] "Power:  0.958"
## [1] "Alternative Mean:  95.0999999999997"
## [1] "Power:  0.972"
## [1] "Alternative Mean:  95.1999999999997"
## [1] "Power:  0.962"
## [1] "Alternative Mean:  95.2999999999997"
## [1] "Power:  0.93"
## [1] "Alternative Mean:  95.3999999999997"
## [1] "Power:  0.932"
## [1] "Alternative Mean:  95.4999999999997"
## [1] "Power:  0.918"
## [1] "Alternative Mean:  95.5999999999997"
## [1] "Power:  0.92"
## [1] "Alternative Mean:  95.6999999999997"
## [1] "Power:  0.898"
```

It appears that the closest whole number we can approximate as an alternative mean with power of at least .9 would be 96 or 95. If we allow for decimal places, we can say a mean of about 95.7 is the largest alternative mean we can achieve with a power of at least .9