Homework 5

Alexander Van Roijen

February 8, 2019

Problem 1, found in seperate drop box.

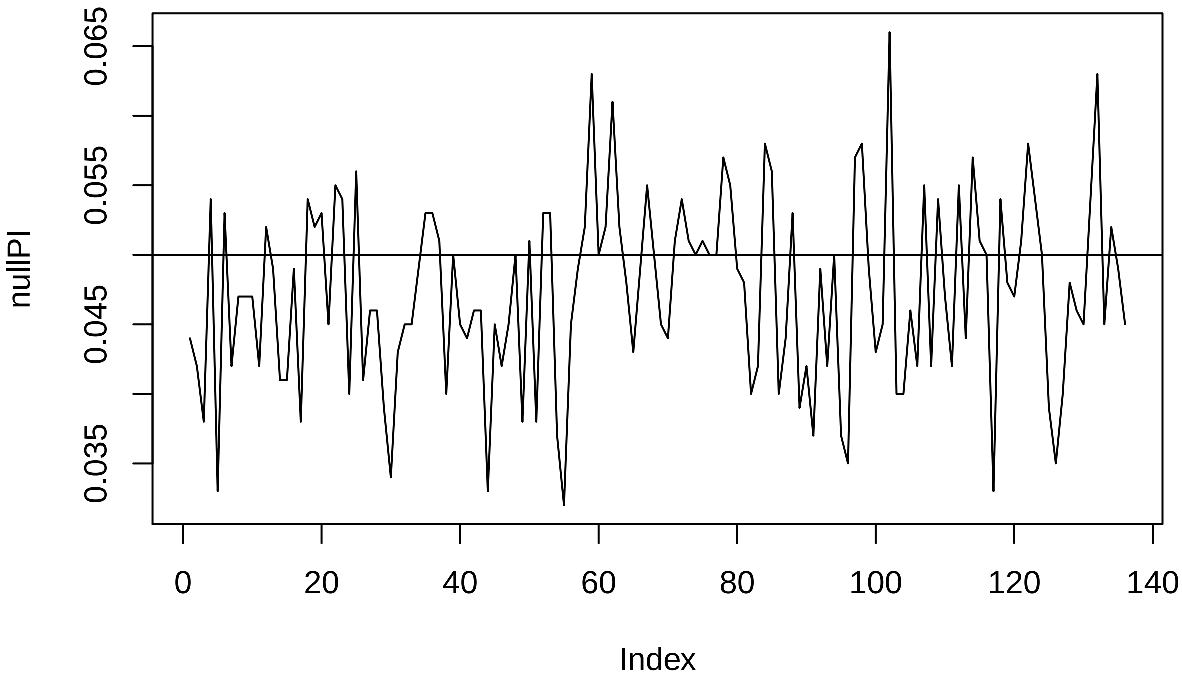
Problem 2

Simulations will be conducted below to determine appropriate sample sizes to achieve all 3 requirements.

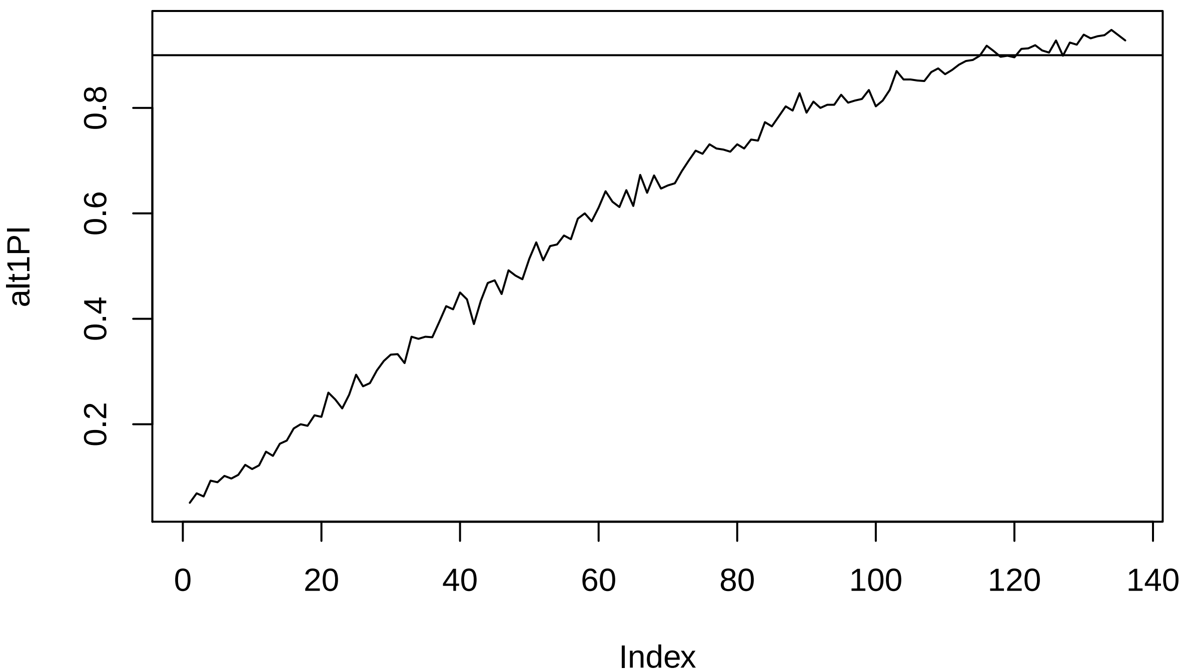
```
numSimu = 1000
sizes = c(5:140)
meanLL=3
nullrate=1/meanLL
nullPI = numeric(length(sizes))
alt1PI = numeric(length(sizes))
alt2PI = numeric(length(sizes))
min=sizes[1]
for(i in sizes)
{
  ss=i
  batNULLDataFrame = data.frame(matrix(ncol = 2,nrow = ss*3))
  batAlt1DataFrame = data.frame(matrix(ncol = 2,nrow = ss*3))
  batAlt2DataFrame = data.frame(matrix(ncol = 2,nrow = ss*3))
  names(batNULLDataFrame)=c('batType','lifeTime')
  names(batAlt1DataFrame)=c('batType','lifeTime')
  names(batAlt2DataFrame)=c('batType','lifeTime')
  types=c(rep('A',ss),rep('B',ss),rep('C',ss))
  batNULLDataFrame$batType=types
  batAlt1DataFrame$batType=types
  batAlt2DataFrame$batType=types
  nullResult=numeric(numSimu)
  alt1Result=numeric(numSimu)
  alt2Result=numeric(numSimu)
  for(j in 1:numSimu)
  {
    nullas=rexp(n=ss,rate=nullrate)
    nullbs=rexp(n=ss,rate=nullrate)
    nullcs=rexp(n=ss,rate=nullrate)
    alt1as=rexp(ss,1/3)
    alt1bs=rexp(ss,1/3)
    alt1cs=rexp(ss,1/2)
    alt2as=rexp(ss,1/2)
    alt2bs=rexp(ss,1/3)
    alt2cs=rexp(ss,1/4)
    nullvals=c(nullas,nullbs,nullcs)
    alt1vals=c(alt1as,alt1bs,alt1cs)
    alt2vals=c(alt2as,alt2bs,alt2cs)
    batNULLDataFrame$lifeTime=nullvals
    batAlt1DataFrame$lifeTime=alt1vals
    batAlt2DataFrame$lifeTime=alt2vals
```

```
    nullRes = summary(aov(lifeTime ~ batType,data=batNULLDataFrame))
    alt1Res = summary(aov(lifeTime ~ batType,data=batAlt1DataFrame))
    alt2Res = summary(aov(lifeTime ~ batType,data=batAlt2DataFrame))
    nullpval = nullRes[[1]][["Pr(>F)"]][1]
    if(nullpval<=0.05)
    {
      nullResult[j]=1
    }
    else
    {
      nullResult[j]=0
    }
    alt1pval = alt1Res[[1]][["Pr(>F)"]][1]
    if(alt1pval<=0.05)
    {
      alt1Result[j]=1
    }
    else
    {
      alt1Result[j]=0
    }
    alt2pval = alt2Res[[1]][["Pr(>F)"]][1]
    if(alt2pval<=0.05)
    {
      alt2Result[j]=1
    }
    else
    {
      alt2Result[j]=0
    }
  }
  #print(nullResult[1:100])
  nullPI[i-min+1]=(mean(nullResult))
  alt1PI[i-min+1]=(mean(alt1Result))
  alt2PI[i-min+1]=(mean(alt2Result))
}
```

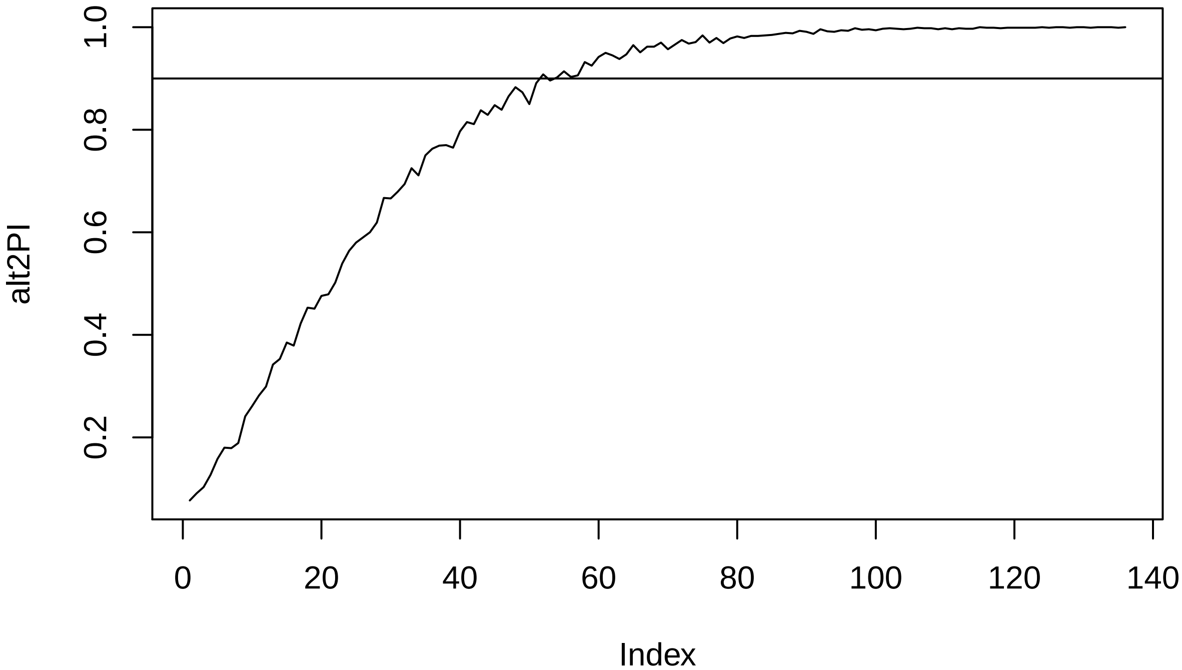
```
plot(nullPI,type='l')
abline(h=0.05)
```




```r
plot(alt1PI,type='l')
abline(h=0.9)
```




```r
plot(alt2PI,type='l')
abline(h=0.9)
```



We can see from the above that to satisfy all three requirements, we would need a sample size of approximately 128 in order to satisfy the power requirements and signfigance level under the null hypothesis. This makes sense as the difference between all three means is reduced in the first alternative hypothesis, and thus would require more samples to have confidence in stating their inequality. The alpha level of our null hypothesis expresses some concerns as it fluctuates by about 1% around the 0.05 mark. I chalk this up due to the use of the exponential distribution and violating the normal distribution assumption.

Problem 3

I have a wide variety of interests, ranging from education, sports, food, health and safety, and much more. I looked through kaggle to see if I couldn't find some interesting data sets that hone in to this point. I managed to find this dataset: https://www.kaggle.com/noriuk/us-education-datasets-unification-project This dataset highlights test scores, income levels, and enrollment at different states in their public education system. There is more granular data available through the census bureau, mainly the american community survey, but this is higher level which may be good for narrowing down the scope of a project. Ideally, I wanted to see what schools in what years and what states differ most from their fellow states and then go into further detail, ideally on income data of the households, as well as availability of sports involvement. However, I understand that may be too adventurous,so I would be more than satisfied with a good analysis on test score differences, including paired t-tests on 4th to 8th grade exam scores in different states to determine if there is improvement, stagnation, or decreasing performance.