Homework 5

Alexander Van Roijen

February 8, 2019

Problem 1

Authors: Alexander Van Roijen, Frank Chen, John Mahoney, Sam Miller, Vivek Kumar

First, we decided to look at all possible 2 sample t-tests to get some potential intuition on the differences between groups.

```
processData = read.csv('defects.csv')
print(names(processData))
```

```
## [1] "Sample"  "Method"  "Defects" "Weight"
```

```
chars = c('A','B','C','D')
as = processData[processData$Method=='A',][,'Defects']
bs = processData[processData$Method=='B',][,'Defects']
cs = processData[processData$Method=='C',][,'Defects']
ds = processData[processData$Method=='D',][,'Defects']
for(i in 1:4)
{
  j = i+1
  Data1 = processData[processData$Method==chars[i],][,'Defects']
  while(j<5)
  {
    print(chars[i])
    print(chars[j])
    Data2 = processData[processData$Method==chars[j],][,'Defects']

    print(t.test(Data1,Data2,var.equal=F))
    j=j+1
  }
}
```

```
## [1] "A"
## [1] "B"
##
##  Welch Two Sample t-test
##
## data:  Data1 and Data2
## t = -3.2589, df = 101.78, p-value = 0.001521
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.29941360 -0.07283865
## sample estimates:
##  mean of x  mean of y
## 0.04054054 0.22666667
##
## [1] "A"
## [1] "C"
```

```
##
##  Welch Two Sample t-test
##
## data:  Data1 and Data2
## t = -2.0919, df = 78.185, p-value = 0.03969
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.286812032 -0.007106887
## sample estimates:
##  mean of x  mean of y
## 0.04054054 0.18750000
##
## [1] "A"
## [1] "D"
##
##  Welch Two Sample t-test
##
## data:  Data1 and Data2
## t = -2.7728, df = 67.717, p-value = 0.007173
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.29406843 -0.04792741
## sample estimates:
##  mean of x  mean of y
## 0.04054054 0.21153846
##
## [1] "B"
## [1] "C"
##
##  Welch Two Sample t-test
##
## data:  Data1 and Data2
## t = 0.4638, df = 124.57, p-value = 0.6436
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1279715  0.2063048
## sample estimates:
## mean of x mean of y
## 0.2266667 0.1875000
##
## [1] "B"
## [1] "D"
##
##  Welch Two Sample t-test
##
## data:  Data1 and Data2
## t = 0.19531, df = 115.97, p-value = 0.8455
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1382847  0.1685411
## sample estimates:
## mean of x mean of y
## 0.2266667 0.2115385
##
```

```
## [1] "C"
## [1] "D"
##
##  Welch Two Sample t-test
##
## data:  Data1 and Data2
## t = -0.27443, df = 113.79, p-value = 0.7843
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1975645  0.1494875
## sample estimates:
## mean of x mean of y
## 0.1875000 0.2115385
```

We can see that their is quite a difference between group A and the other groups, but no real indication between all other pairings. Now let us conduct an F-test and see what comes up.

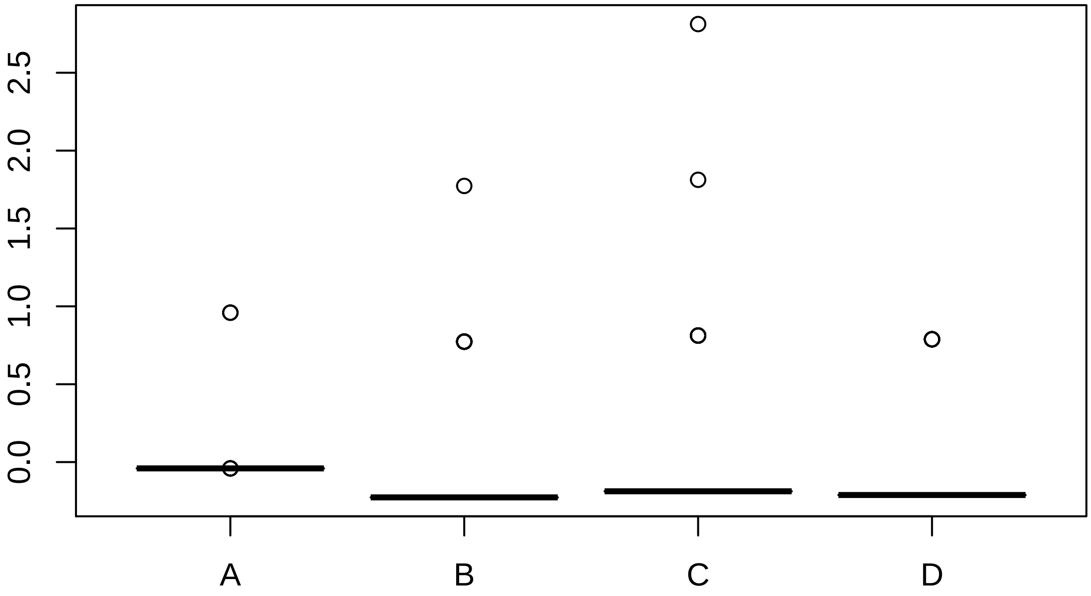
```
ftest = aov(Defects ~ Method,data=processData)
print(summary(ftest))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Method        3   1.57  0.5248   3.082  0.028 *
## Residuals   261  44.45  0.1703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see, the F-test deems there is a significant enough differences between our population means.

However, was this test valid? We will assume there is independence between samples and within samples, however that is not guaranteed. Next thing we can look at is if the data have equal variances and then finally if they have normally distributed values.

```
boxplot(ftest$residuals~Method,data=processData)
```



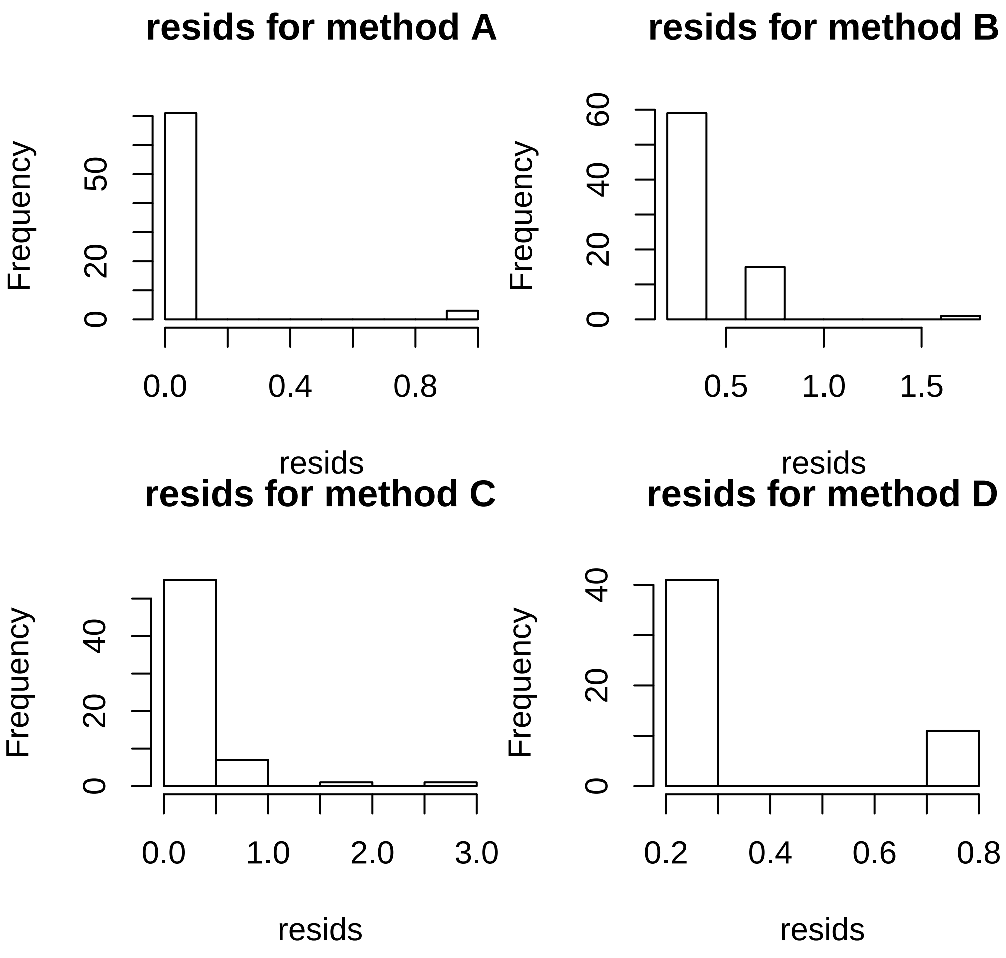
```
print(tapply(processData$Defects,processData$Method,sd))
```

```
##         A         B         C         D
## 0.1985695 0.4524209 0.5307975 0.4123837
```

The box plot doesn't help much as defects are discrete and mainly fall around zero. Looking at their standard deviations, we can see A has much lower standard deviation than the rest which may indicate non equal variances.

Now lets look at distributions of their residuals to assess our normality distribution.

```
par(mfrow=c(1,1))
residFrame = data.frame(matrix(ncol = length(chars),nrow=75))
colnames(residFrame) = chars
for(i in 1:4)
{
  par(mfrow=c(1,1))
  Data1 = c(processData[processData$Method==chars[i],][,'Defects'])
  resids = (abs(Data1-mean(Data1)))
  hist(resids,main=paste("resids for method ",chars[i],sep=""))
}
```



Clearly, we can see that the residuals do not follow a normal distribution, it appears to be either a poisson distribution, exponential distribution, or perhaps a folded normal distribution. Lets simulate this to verify.

```
numSimu = 2000
results = numeric(numSimu)
aMean = mean(as)
bMean = mean(bs)
cMean = mean(cs)
dMean = mean(ds)
```

```
asd = sd(as)
bsd = sd(bs)
csd = sd(cs)
dsd = sd(ds)
lens = tapply(processData$Defects,processData$Method,length)
alen = lens[1]
blen = lens[2]
clen = lens[3]
dlen = lens[4]
netMean = (aMean+bMean+cMean+dMean)/4
#print(bsd)
netMean = (bsd)
#print(netMean)
totsd=0
for(i in 1:numSimu)
{
  simData = processData
  aSim = rpois(n=alen,netMean)
  bSim = rpois(blen,netMean)
  cSim = rpois(clen,netMean)
  dSim = rpois(dlen,netMean)
  totsd=totsd+(sd(aSim))
  simData$Defects[simData$Method=='A']=aSim
  simData$Defects[simData$Method=='B']=bSim
  simData$Defects[simData$Method=='C']=cSim
  simData$Defects[simData$Method=='D']=dSim
  result = aov(Defects~Method,data=simData)
  store = summary(result)
  pval = store[[1]][["Pr(>F)"]][1]
  if(pval<=0.05)
  {
    results[i]=1
  }
  else
  {
    results[i]=0
  }

}
#print(totsd/numSimu)
print(mean(results))
```

```
## [1] 0.039
```

We can see, that despite drawing from a poisson distribution, we still have some validity in the type I error of our F-test as it it quite close to 0.05. However, there is a bit of an issue here as I was unable to provide each population with their respective sample variance due to how the poisson distribution is structured. However, I assumed despite these the poisson distribution was the best fit.

There are concerns still about the variance, as the F-test assumes equal variance, which we have incorporated into our distribution, but may not necessarily be true for the underlying population. Our sample SDs seem to indicate this, but it could have been one bad sample. As a result, we decided to stick with the closeness of our simulation here to accurately reflect the validity of our F-test.

Overall, given all these facts and simulations, We believe the F-test is still an appropriate test for this study.

I will note that many of the members studied different types of distributions, including Normal and exponential, and found some conflicting results which may indicate there is some more concern here to be had. However, to avoid being too verbose, the work is left out and you are encouraged to ask members questions about our reservations and thoughts on the matter.