Fit Distributions to Somerville City Data

City Wages

Somerville wages: https://data.somervillema.gov/Finance/City-Of-Somerville-Weekly-Payroll-Gross-Wages-Over/q35q-jc9v

```
SomWage<-read.csv("./Data/City_Of_Somerville_Weekly_Payroll_Gross_Wages_Over_50K_2016.csv")
head(SomWage)
```

```
##                 Name Dept.            Title Total.Gross.Calc X5.2.BB Perf.Att
## 1           REDACTED 43010   Police Officer         236854.0 1307.04       NA
## 2    CABRAL, MICHAEL 43010          Captain         222849.1 2779.20      500
## 3  WARD, CHRISTOPHER 43010 Lieutenant + 25%         201717.4      NA      400
## 4      FALLON, DAVID 43010     Police Chief         195816.9      NA       NA
## 5  LAVEY JR, RICHARD 43010  Lieutenant +25%         192693.0   55.78       NA
## 6 CICERONE, FERNANDO 43010   Police Officer         191351.0 1629.96       NA
##   Auto.Allow Birthday Clothing Court.Time CPR Defib.Pay   Detail Shift.Diff DT
## 1         NA       NA       NA    7611.54  NA        NA 89364.20    7840.96 NA
## 2         NA       NA       NA         NA  NA        NA 63137.25    7135.44 NA
## 3         NA       NA       NA    4963.68  NA        NA 51347.10    7551.54 NA
## 4         NA       NA       NA         NA  NA        NA       NA         NA NA
## 5         NA       NA       NA    5187.99  NA        NA 33242.50    7752.56 NA
## 6         NA       NA       NA   10321.73  NA        NA 64585.86    6888.96 NA
##   Educ.Inc Election.Pay EMD EMT Ex.Other Fluen.Bon FSD Gas.Allowance    O.G
## 1       NA           NA  NA  NA       NA        NA  NA            NA     NA
## 2       NA           NA  NA  NA       NA        NA  NA            NA     NA
## 3       NA           NA  NA  NA       NA        NA  NA            NA 531.73
## 4       NA           NA  NA  NA       NA        NA  NA            NA     NA
## 5       NA           NA  NA  NA       NA        NA  NA            NA     NA
## 6       NA           NA  NA  NA       NA        NA  NA            NA     NA
##   Grant.OT Haz.Duty     Hol Matron.Pay Meals Master.FAO Master.FF       OT OT2
## 1  2998.36       NA 5381.94         NA    NA         NA        NA 40191.59  NA
## 2  1179.84       NA 9176.60         NA    NA         NA        NA  9635.44  NA
## 3  1004.16       NA 7394.45         NA    NA         NA        NA 20297.99  NA
## 4       NA       NA      NA         NA    NA         NA        NA       NA  NA
## 5   669.44       NA 7809.86         NA    NA         NA        NA 21069.95  NA
## 6  5535.48       NA 5381.94         NA    NA         NA        NA 17126.12  NA
##   OtherNR Personal      Reg Retro Sick Sick.Buy.Back StipR StipNR Uniform
## 1      NA       NA  74729.2    NA   NA            NA    NA     NA      NA
## 2      NA       NA 127418.7    NA   NA            NA    NA     NA      NA
## 3      NA       NA 102974.9    NA   NA            NA    NA     NA      NA
## 4      NA       NA 191443.2    NA   NA            NA    NA     NA      NA
## 5      NA       NA 108441.3    NA   NA            NA    NA     NA      NA
## 6      NA       NA  74729.2    NA   NA            NA    NA     NA      NA
##   Vacation Weapons WMD Extra.Hol.NR Senior.Longevity Longevity Service.Recog
## 1       NA     500 500           NA             3200        NA            NA
## 2       NA     600 500           NA               NA        NA            NA
## 3       NA     600 500           NA               NA        NA            NA
```

```
## 4       NA      NA  NA       373.77              NA        NA            NA
## 5       NA     600 500           NA              NA        NA            NA
## 6       NA     500 500           NA              NA        NA            NA
##   Adm.Leave Furlough Training.OT Snow.OT Snow.OT2 Snow.DT Gym.Reim OptI    OptF
## 1        NA       NA     3229.17      NA       NA      NA       NA   NA      NA
## 2        NA       NA      786.57      NA       NA      NA       NA   NA      NA
## 3        NA       NA     4151.83      NA       NA      NA       NA   NA      NA
## 4        NA       NA          NA      NA       NA      NA       NA   NA 3999.96
## 5        NA       NA     7363.63      NA       NA      NA       NA   NA      NA
## 6        NA       NA     4151.76      NA       NA      NA       NA   NA      NA
##   Trans
## 1    NA
## 2    NA
## 3    NA
## 4    NA
## 5    NA
## 6    NA
```

```
str(SomWage)
```

```
## 'data.frame':    640 obs. of  59 variables:
##  $ Name            : Factor w/ 636 levels "ACCAPUTO, LAURA",..: 489 66 614 186 322 93 344 240 141 60
##  $ Dept.           : int  43010 43010 43010 43010 43010 43010 43010 42010 43010 43010 ...
##  $ Title           : Factor w/ 300 levels "Accountant","Accountant/Business Analyst",..: 205 30 180 1
##  $ Total.Gross.Calc: num  236854 222849 201717 195817 192693 ...
##  $ X5.2.BB         : num  1307 2779.2 NA NA 55.8 ...
##  $ Perf.Att        : int  NA 500 400 NA NA NA NA 300 NA 500 ...
##  $ Auto.Allow      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Birthday        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Clothing        : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Court.Time      : num  7612 NA 4964 NA 5188 ...
##  $ CPR             : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Defib.Pay       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Detail          : num  89364 63137 51347 NA 33243 ...
##  $ Shift.Diff      : num  7841 7135 7552 NA 7753 ...
##  $ DT              : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Educ.Inc        : logi  NA NA NA NA NA NA ...
##  $ Election.Pay    : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ EMD             : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ EMT             : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Ex.Other        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Fluen.Bon       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ FSD             : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Gas.Allowance   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ O.G             : num  NA NA 532 NA NA ...
##  $ Grant.OT        : num  2998 1180 1004 NA 669 ...
##  $ Haz.Duty        : int  NA NA NA NA NA NA NA 9000 NA NA ...
##  $ Hol             : num  5382 9177 7394 NA 7810 ...
##  $ Matron.Pay      : logi  NA NA NA NA NA NA ...
##  $ Meals           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Master.FAO      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Master.FF       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ OT              : num  40192 9635 20298 NA 21070 ...
##  $ OT2             : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ OtherNR         : int  NA NA NA NA NA NA NA NA NA NA ...
```

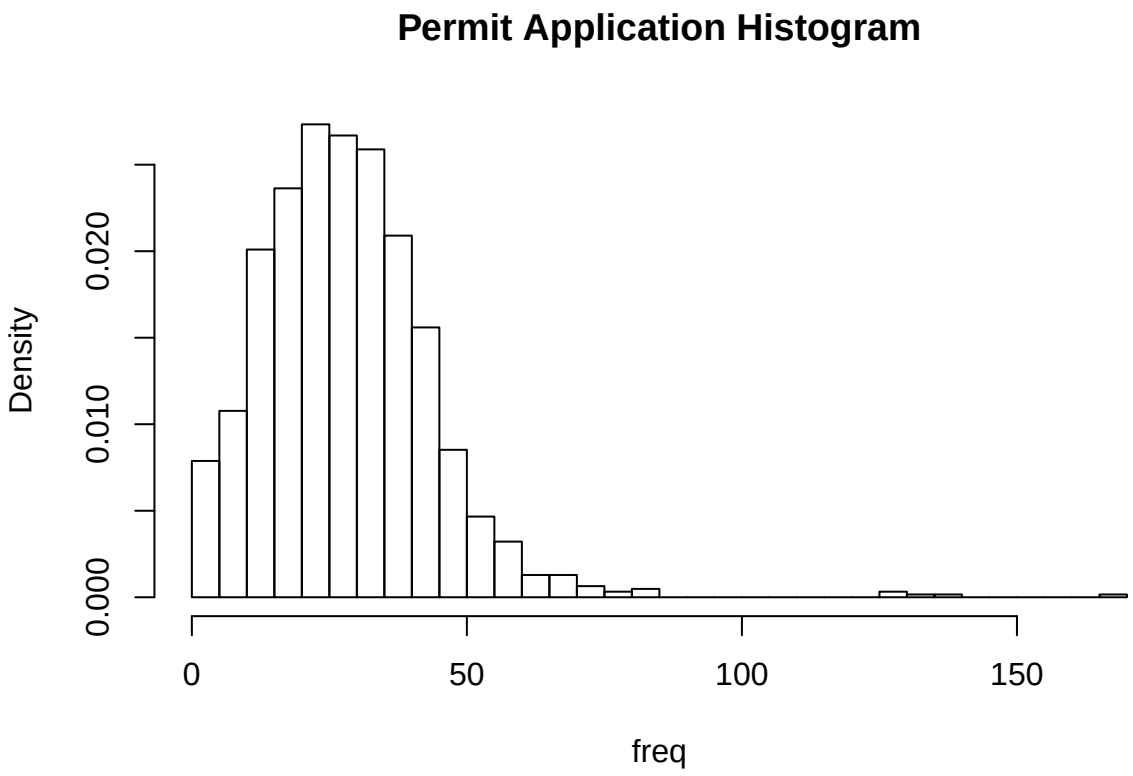
```
##  $ Personal        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Reg             : num  74729 127419 102975 191443 108441 ...
##  $ Retro           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Sick            : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Sick.Buy.Back   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ StipR           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ StipNR          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Uniform         : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Vacation        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Weapons         : int  500 600 600 NA 600 500 475 NA 600 600 ...
##  $ WMD             : int  500 500 500 NA 500 500 500 NA 500 500 ...
##  $ Extra.Hol.NR    : num  NA NA NA 374 NA ...
##  $ Senior.Longevity: int  3200 NA NA NA NA NA NA NA NA NA ...
##  $ Longevity       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Service.Recog   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Adm.Leave       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Furlough        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Training.OT     : num  3229 787 4152 NA 7364 ...
##  $ Snow.OT         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Snow.OT2        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Snow.DT         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Gym.Reim        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ OptI            : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ OptF            : num  NA NA NA 4000 NA ...
##  $ Trans           : int  NA NA NA NA NA NA NA NA NA NA ...
```

```
SomWage[1,]
```

```
##       Name Dept.          Title Total.Gross.Calc X5.2.BB Perf.Att Auto.Allow
## 1 REDACTED 43010 Police Officer           236854 1307.04       NA         NA
##   Birthday Clothing Court.Time CPR Defib.Pay  Detail Shift.Diff DT Educ.Inc
## 1       NA       NA    7611.54  NA        NA 89364.2    7840.96 NA       NA
##   Election.Pay EMD EMT Ex.Other Fluen.Bon FSD Gas.Allowance O.G Grant.OT
## 1           NA  NA  NA       NA        NA  NA            NA  NA  2998.36
##   Haz.Duty     Hol Matron.Pay Meals Master.FAO Master.FF       OT OT2 OtherNR
## 1       NA 5381.94         NA    NA         NA        NA 40191.59  NA      NA
##   Personal     Reg Retro Sick Sick.Buy.Back StipR StipNR Uniform Vacation
## 1       NA 74729.2    NA   NA            NA    NA     NA      NA       NA
##   Weapons WMD Extra.Hol.NR Senior.Longevity Longevity Service.Recog Adm.Leave
## 1     500 500           NA             3200        NA            NA        NA
##   Furlough Training.OT Snow.OT Snow.OT2 Snow.DT Gym.Reim OptI OptF Trans
## 1       NA     3229.17      NA       NA      NA       NA   NA   NA    NA
```

```
#Will use Total Gross. Subtract 50,000 because that is cut off
SomWage.TGC<-SomWage$Total.Gross.Calc-50000

SomWage.hist<-hist(SomWage.TGC,breaks="FD",probability=TRUE,main="Wage Histogram")
```



I will fit gross wages of Somerville city employees to a parametric distribution. The data is truncated at 50000. It looks like it could follow an exponential distirbution.

```
m<-mean(SomWage.TGC);m
```

```
## [1] 37765.71
```

```
v<-var(SomWage.TGC);v
```

```
## [1] 992228559
```

```
lambda<-1/m;lambda
```

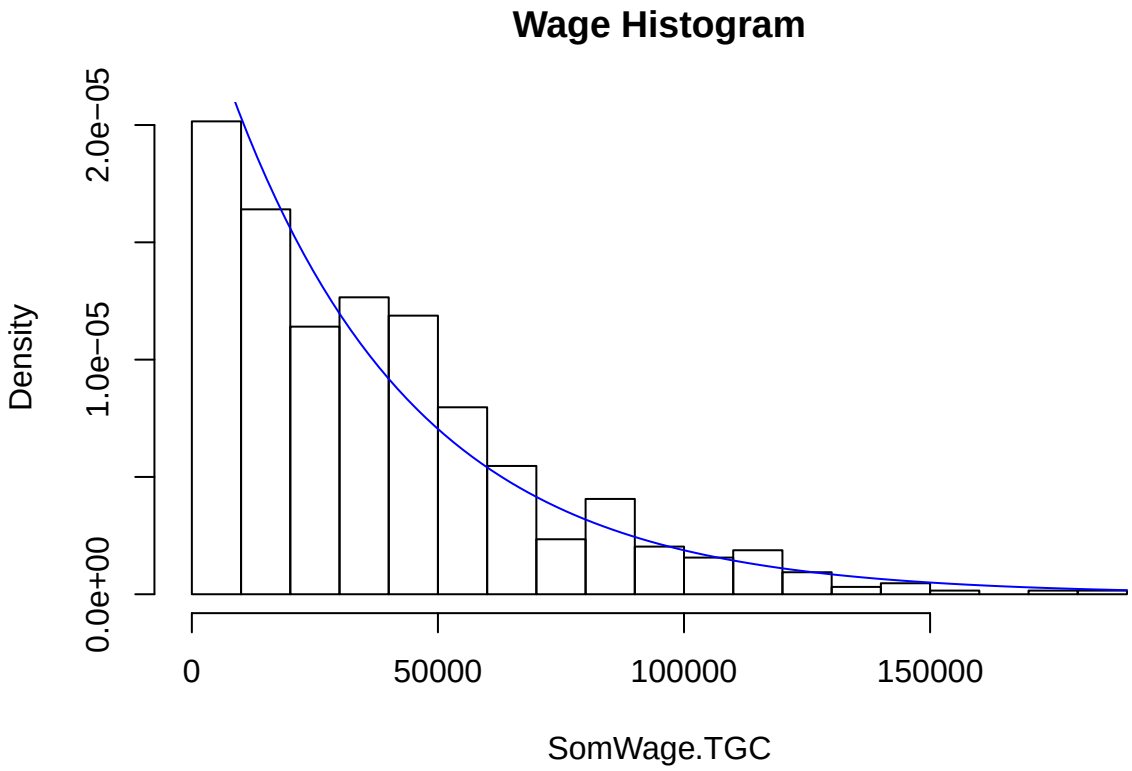
```
## [1] 2.647904e-05
```

```
#check variance
((m^2 - v) /v);m^2
```

```
## [1] 0.43742
```

```
## [1] 1426249196
```

```
plot(SomWage.hist,freq=FALSE, main="Wage Histogram")
curve(dexp(x,lambda), col = "blue", add= TRUE)
```



The fit looks pretty good, but maybe a Gamma would be better.

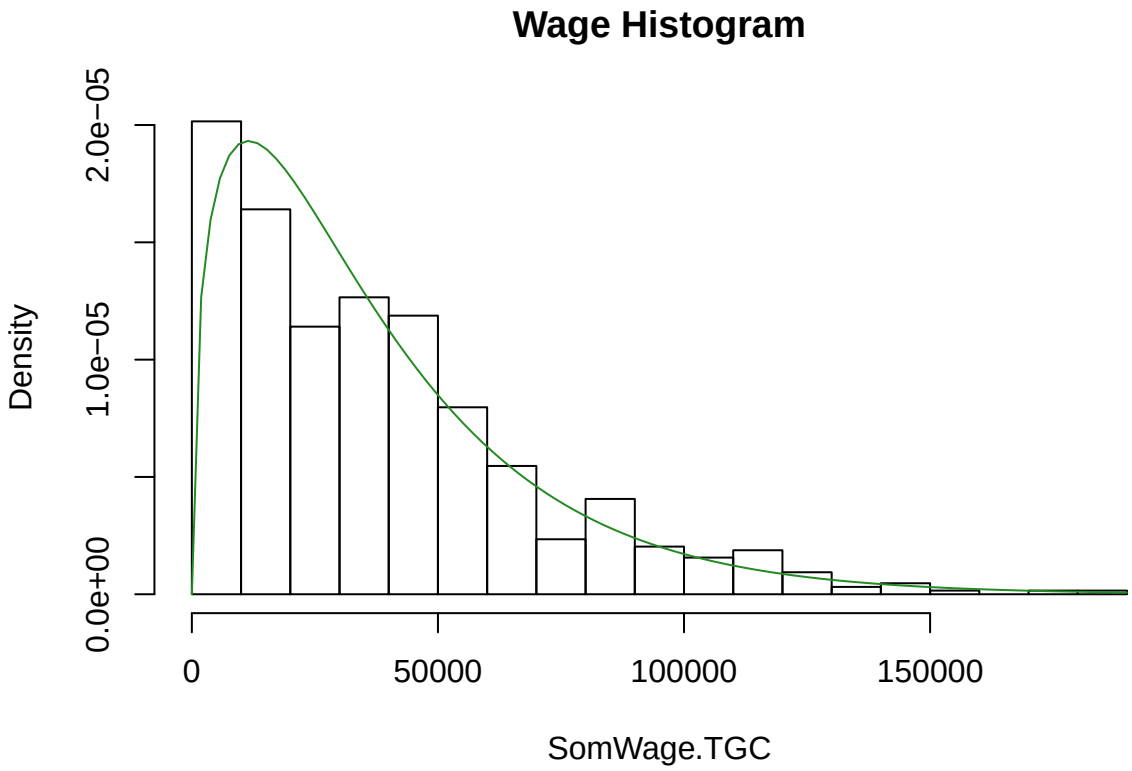
```
beta<-m/v;beta
```

```
## [1] 3.806151e-05
```

```
alpha<-beta*m;alpha
```

```
## [1] 1.43742
```

```
plot(SomWage.hist,freq=FALSE, main="Wage Histogram")
curve(dgamma(x,shape=alpha, rate=beta), col = "forest green", add= TRUE)
```



The Gamma starts at a low frequency while the dataset starts at a high frequency so the fit is not good.

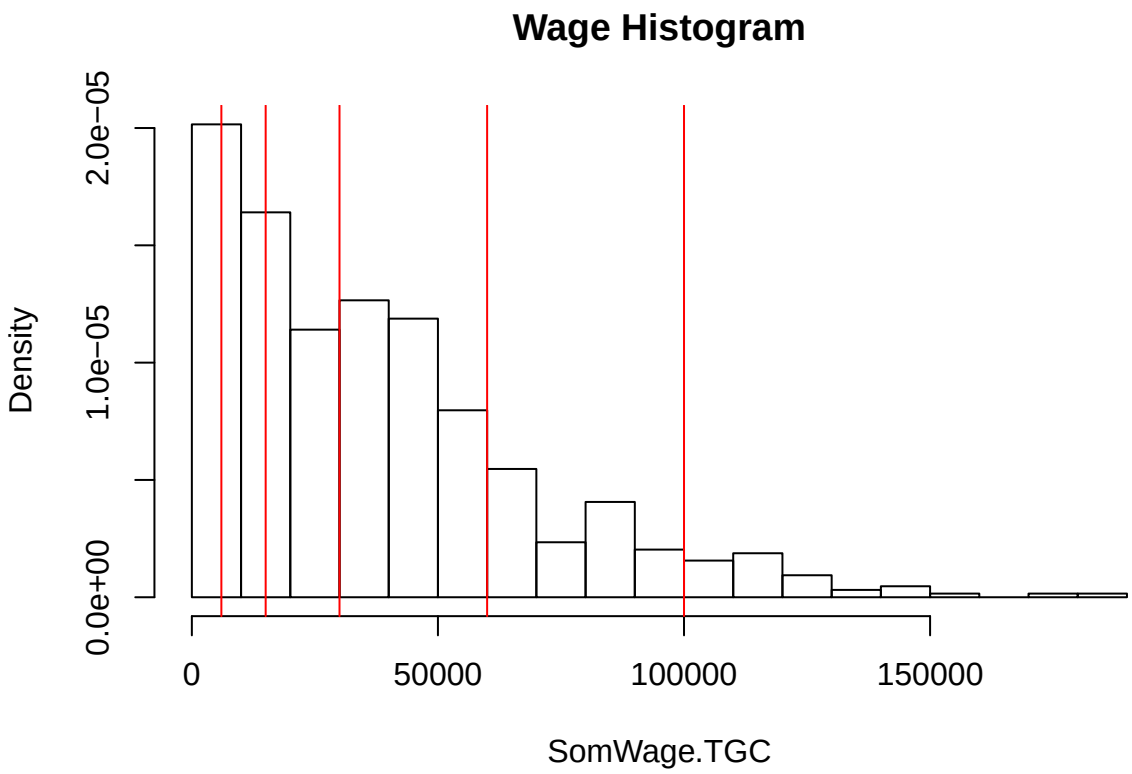
Now I use a χ^2-test for goodness of fit.

```
#choose bins
min(SomWage.TGC);max(SomWage.TGC)
```

```
## [1] 78.95
```

```
## [1] 186854
```

```
plot(SomWage.hist,freq=FALSE, main="Wage Histogram")
abline(v = 6000, col = "red")
abline(v = 15000, col = "red")
abline(v =30000, col = "red")
abline(v =60000, col = "red")
abline(v =100000, col = "red")
```




```
#Obs table
ob1<-sum(SomWage.TGC<=6000)
ob2<-sum(SomWage.TGC<=15000)
ob3<-sum(SomWage.TGC<=30000)
ob4<-sum(SomWage.TGC<=60000)
ob5<-sum(SomWage.TGC<=100000)
ob6<-sum(SomWage.TGC>100000)

Observed<-c(ob1,ob2-ob1,ob3-ob2,ob4-ob3,ob5-ob4,ob6);Observed
```

```
## [1]  82  97 128 208  89  36
```

```
#expected
expf<-function(x) dexp(x,lambda)
eb1<-integrate(expf,0,6000)$value
eb2<-integrate(expf,6000,15000)$value
eb3<-integrate(expf,15000,30000)$value
eb4<-integrate(expf,30000,60000)$value
eb5<-integrate(expf,60000,100000)$value
eb6<- 1-integrate(expf,0,100000)$value

length(SomWage.TGC)
```

```
## [1] 640
```

```
Expected <- 640*c(eb1,eb2,eb3,eb4,eb5,eb6); Expected
```

```
## [1]  94.01368 115.77242 141.02015 158.51715  85.36497  45.31163
```

```r
#chi sq
Chi<-sum((Observed-Expected)^2/Expected);Chi #23.30
```

```
## [1] 23.29621
```

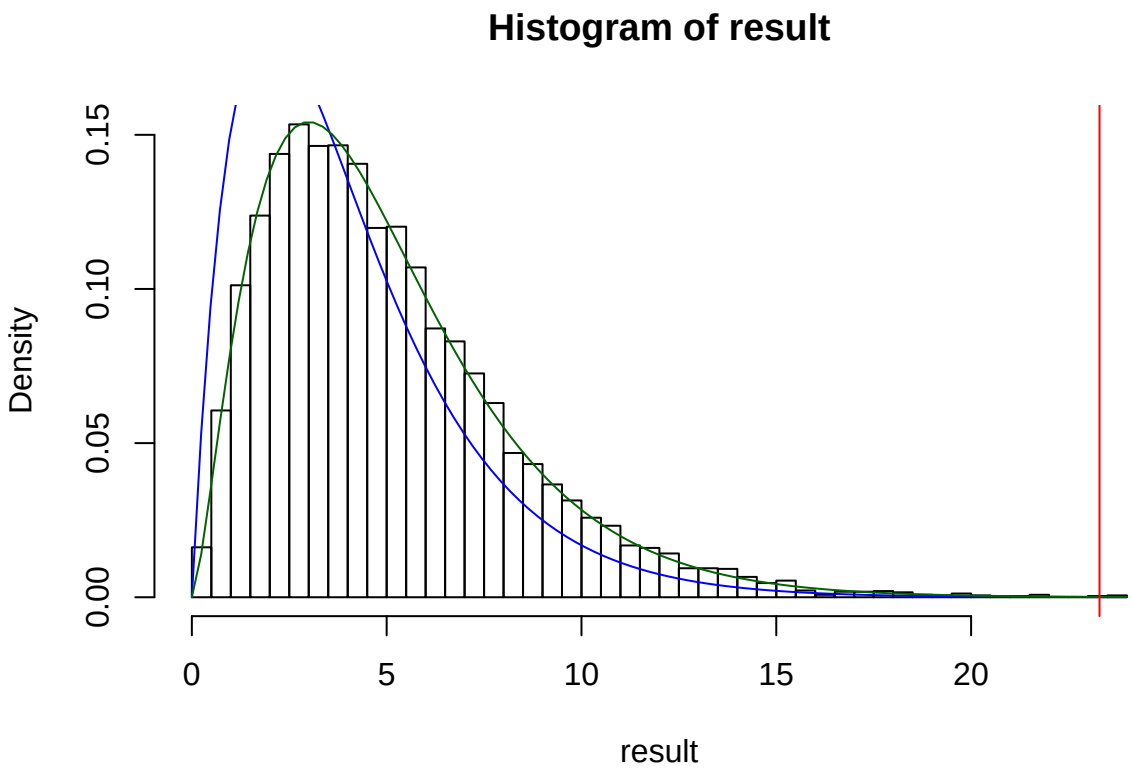
```r
#how probable get this large of a test stat
#df = 6-1-1=4  6 buckets, calculated lambda from data
Pval<-pchisq(Chi,4,lower.tail=FALSE);Pval #.0001105
```

```
## [1] 0.0001104887
```

```r
#very low
```

I can also perform a similar test but through simulation instead of using a χ^2 distribution directly

```r
#draw 10k samples from the exp distribution
N = 10^4; result = numeric(N)
for (i in 1:N) {
  expData = rexp(640,lambda)
  Counts=numeric(6)
  Counts[1] = sum(expData <= 6000)
  Counts[2] = sum((expData > 6000) & (expData <= 15000))
  Counts[3] = sum((expData > 15000) & (expData <= 30000))
  Counts[4] = sum((expData > 30000) & (expData <= 60000))
  Counts[5] = sum((expData > 60000) & (expData <= 100000))
  Counts[6] = sum(expData > 100000)
  result[i] = sum((Counts-Expected)^2/Expected)
}
hist(result, breaks = "FD",probability =TRUE)
curve(dchisq(x, df=4), col = "blue", add= TRUE)
#calculated lambda from the data so should have 4 df but df=5 has a much better fit to simulation
curve(dchisq(x, df=5), col = "Dark Green", add= TRUE)
abline(v = Chi, col = "red")
```




```
sum(result >= Chi)/N;
```

```
## [1] 3e-04
```

In both approaches, I get a high test statistic and low probability of getting this distribution from an exponential with this lambda, indicating a poor fit.

Building Permits

Somerville building permits: https://data.somervillema.gov/City-Services/ISD-Building-Permit-Daily-Applications/q3yh-mp87

I wanted a dataset that could be a good fit for a normal distribution. Each row in this dataset is an application. I'm expecting that the distribution of the number of permit applications each day is approximately normal.

```
Permit<-read.csv("./Data/ISD_Building_Permit_Daily_Applications.csv")
head(Permit)
```

```
##       File.     Permit.             PermitType           Address
## 1 19-006907  G19-000319 Residential - Existing    11  BOWDOIN ST
## 2 19-006874  P19-000323 Residential - Existing  31  IRVINGTON RD
## 3 19-006898 DP19-000090        Dumpster Permit      52  DOVER ST
## 4 19-006901 DP19-000091        Dumpster Permit 29  PARTRIDGE AVE
## 5 19-006903  P19-000326 Residential - Existing     9  MEDFORD ST
## 6 19-006905  P19-000327 Residential - Existing    11  BOWDOIN ST
##                          Applicant ApplicantAddress    ApplicantCityStZip
## 1     Redmond plumbing and heating      9A grove st Westborough  MA 01581
## 2      ken mcconnell plg & htg inc    19 chardon rd     medford  ma 02155
```

```
## 3                  Shakeel Hossain  52 Dover Street   Somerville MA 02144
## 4                      Brandon Roy 29 Partridge Ave   SOMERVILLE MA 02145
## 5 Wambolt Plumbing & Heating, Corp 11 Wadsworth Ave      Waltham MA 02453
## 6     Redmond plumbing and heating      9A grove st Westborough  MA 01581
##                                                                        ProjectName
## 1                                                                repiping gas fixtures
## 2                                                              adding 3/4 bath to attic
## 3                                                                   Dumpster on Street
## 4 Dumpster on Street for removal of some pavement in back yard at my single family home
## 5                                              Replacing 2 sinks and 2 garbage disposals
## 6                                                        remodel kitchen and bathrooms
##          ApplicationDate              IssueDate         ExpirationDate Status
## 1 04/18/2019 12:00:00 AM 04/18/2019 12:00:00 AM 10/18/2019 12:00:00 AM Issued
## 2 04/18/2019 12:00:00 AM 04/18/2019 12:00:00 AM 10/18/2019 12:00:00 AM Issued
## 3 04/18/2019 12:00:00 AM 04/18/2019 12:00:00 AM                        Issued
## 4 04/18/2019 12:00:00 AM 04/18/2019 12:00:00 AM                        Issued
## 5 04/18/2019 12:00:00 AM 04/18/2019 12:00:00 AM 10/18/2019 12:00:00 AM Issued
## 6 04/18/2019 12:00:00 AM 04/18/2019 12:00:00 AM 10/18/2019 12:00:00 AM Issued
##   CloseDate PermitAmount AmountPaid Latitude Longitude PermitTypeDetail
## 1                    70         70 42.37912 -71.10090              Gas
## 2                    70         70 42.41560 -71.12938         Plumbing
## 3                   100        100 42.39532 -71.12551         Dumpster
## 4                   100        100 42.39385 -71.10354         Dumpster
## 5                    70         70 42.39072 -71.10131         Plumbing
## 6                   130        130 42.37912 -71.10090         Plumbing
```

```
str(Permit)
```

```
## 'data.frame':    35598 obs. of  18 variables:
##  $ File.             : Factor w/ 31840 levels "14-000002","14-000004",..: 31840 31835 31836 31837 318
##  $ Permit.           : Factor w/ 35598 levels "B14-000001","B14-000002",..: 29374 34603 14311 14312 3
##  $ PermitType        : Factor w/ 51 levels "","Commercial",..: 33 33 13 13 33 33 33 3 33 33 ...
##  $ Address           : Factor w/ 9400 levels "0  ALEWIFE BROOK PKWY",..: 571 4584 6918 4266 9042 571
##  $ Applicant         : Factor w/ 8612 levels "",".",".J.jerry .Mmazziotta",..: 6601 4467 7196 1119 83
##  $ ApplicantAddress  : Factor w/ 8230 levels "","&& Emery Rd.",..: 7667 2159 5696 3654 615 7667 5291
##  $ ApplicantCityStZip: Factor w/ 2327 levels "","2127","2143",..: 2169 1139 1839 1851 2085 2169 1988
##  $ ProjectName       : Factor w/ 28877 levels "","- 200 amp 2 spot meter bank- (2) 100 amp 24 positi
##  $ ApplicationDate   : Factor w/ 1747 levels "01/01/2015 12:00:00 AM",..: 540 540 540 540 540 540 534
##  $ IssueDate         : Factor w/ 1244 levels "01/02/2015 12:00:00 AM",..: 371 371 371 371 371 371 371
##  $ ExpirationDate    : Factor w/ 1774 levels "","01/01/2015 12:00:00 AM",..: 1418 1418 1 1 1418 1418
##  $ Status            : Factor w/ 12 levels "","Approved",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ CloseDate         : Factor w/ 1087 levels "","01/02/2018 12:00:00 AM",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ PermitAmount      : num  70 70 100 100 70 ...
##  $ AmountPaid        : num  70 70 100 100 70 ...
##  $ Latitude          : num  42.4 42.4 42.4 42.4 42.4 ...
##  $ Longitude         : num  -71.1 -71.1 -71.1 -71.1 -71.1 ...
##  $ PermitTypeDetail  : Factor w/ 8 levels "Building","Certificate of Occupancy",..: 6 7 4 4 7 7 7 7
```

```
min(as.Date(Permit$IssueDate,format = "%m/%d/%Y"))
```

```
## [1] "2014-05-21"
```

```
max(as.Date(Permit$IssueDate,format = "%m/%d/%Y"))
```

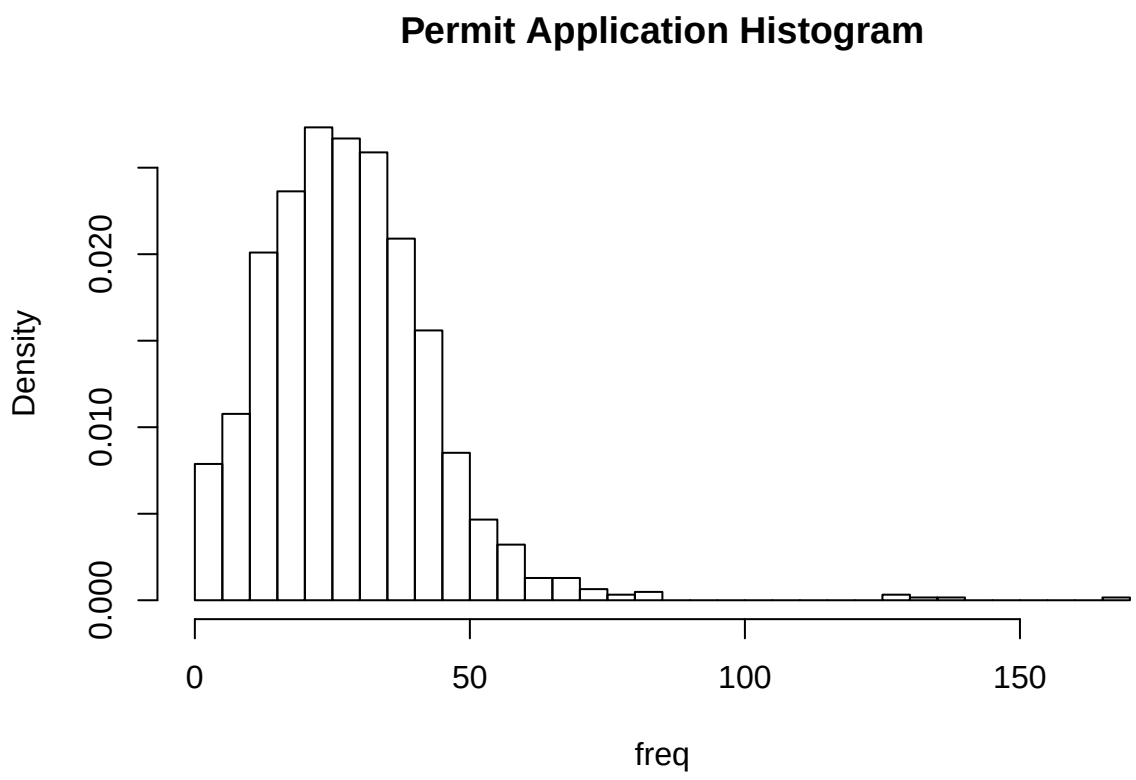
```
## [1] "2019-04-18"
```

```
min(as.numeric(as.Date(Permit$IssueDate,format = "%m/%d/%Y")))
```

```
## [1] 16211
```

```
IssueDate<-as.numeric(as.Date(Permit$IssueDate,format = "%m/%d/%Y"))-16211
freq<-table(IssueDate)

Permit.hist<-hist(freq,breaks="FD",probability=TRUE,main="Permit Application Histogram")
```



This appears to be more Gamma than Guassian, but it could work as a truncated normal distribution.

```
m<-mean(freq);m
```

```
## [1] 28.61576
```

```
sd<-sd(freq);sd
```

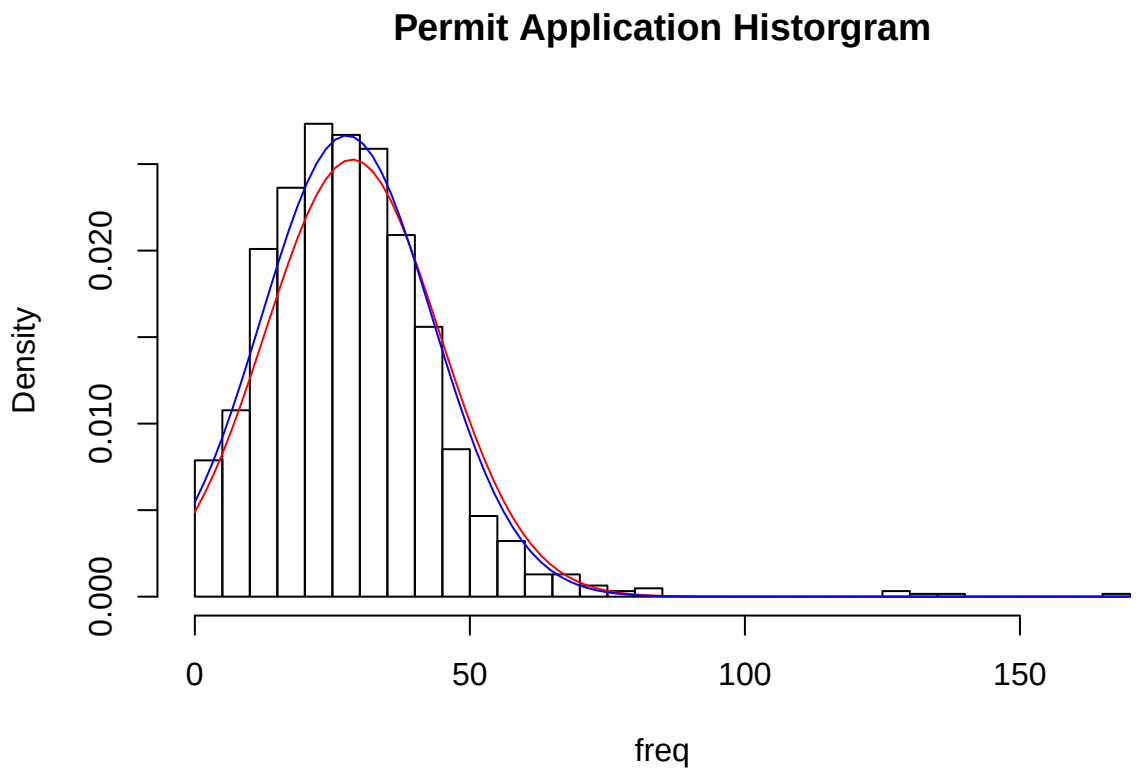
```
## [1] 15.78797
```

```
ln<-length(freq)
plot(Permit.hist, freq=FALSE, main="Permit Application Historgram")
curve(dnorm(x, m, sd), add = TRUE, col = "red")
#actually not bad if we consider a truncated normal distribution

#for truncated normal P(X|X>0)=P(XnX>0)/P(X>0)
fnN<-function(x) dnorm(x,m,sd)
XGT0<-integrate(fnN,0,Inf)$value;trunc #P(X>0)
```

```
## function (x, ...)  .Primitive("trunc")
```

```
XLT0<-1-XGT0
m2<-m*XGT0
sd2<-sd*sqrt(XGT0)
curve(dnorm(x, m2, sd2)/XGT0, add = TRUE, col = "blue")
```



Now I do the chi-square test on the truncated normal distribution, bucketing with deciles.

```
#10% in each
#.1=P(0<X<a)/P(X>0)
#.1*P(X>0)+P(X<0)=P(X<a)
dec <- qnorm(seq(0.0, 1, by = 0.1)*XGT0+XLT0, m2, sd2);dec
```

```
##  [1] -0.4956724 10.2521888 16.0520165 20.5549132 24.5228065 28.2951854
##  [7] 32.1088279 36.2203261 41.0601701 47.8047938        Inf
```

```
Expected<-rep(ln/10,10); Expected
```

```
##  [1] 124.4 124.4 124.4 124.4 124.4 124.4 124.4 124.4 124.4 124.4
```

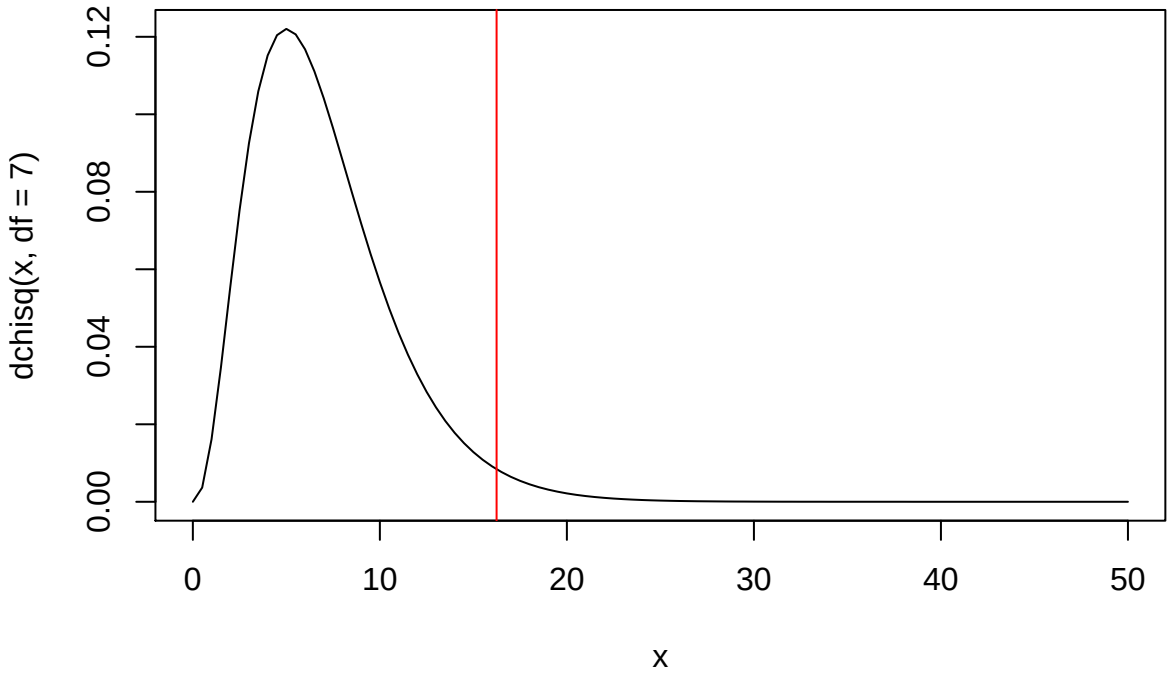
```
obs<-numeric(10)
for (i in 1:10)
  obs[i] <- sum((freq >= dec[i]) & (freq <= dec[i+1]) ); obs
```

```
##  [1] 116 150 122 142 129 129 120 128  98 110
```

```
chi<-sum((obs-Expected)^2/Expected);chi  #16.24
```

```
## [1] 16.24116
```

```r
#df=10-1-2=7
curve(dchisq(x, df = 7),from=0,to=50)
abline(v=chi, col = "red")
```



This shows that once again the fitted model is not very good.