Meteorites

Bradley Baker

Statistical Inference Final Project : Meteorites

Inroduction

Vignette

I spent last summer in the desert, working for the Mind Research Network in Albuquerque, New Mexico on a problem in distributed fMRI data analysis (see the folder labelled djica in my portfolio). Though that particular job found me wrapped almost entirely *in silico*, on my offtime, I had the opportunity to embed myself in various outdoor locations in the southwest. Though most of this involved hiking and exploration of the mountains, wood, and desert areas, on one night toward the end of the trip, I turned my head up to the stars. When a colleague and I visited a meeting of the Albuquerque Astronomical Society, we were first amazed by the huge turn-out to a solitary location far in the mountains. A clearing in the woods teamed with casual and professional astronomers, some setting up expensive telescopes, and others just embracing the yawning blanket of stars above us. One guide, a nucleus of authority surrounded by a cell of interested casual observers, gestured excitedly with small handheld laser-pointer, marking out the locations of constellations, planets, nebulae, and more.

My colleague and I, being entirely foreign to the society, mostly hung around the edges of larger groups, listening to the most knowledgeable members of the society describe the night sky with fantastic names, describing the phenomena, sometimes providing historical epithets regarding the particular astronomer known for first doing what they did now. Toward the end of the night, my colleague and I struck up a conversation with one of the owners of one of the largest telescopes set up in the clearing. It turns out that he had also worked as a data-scientist, and though he had focused mainly on robotics and artifical intelligence, he recounted the few exciting days he once worked for NASA, wistfully claiming that he realized far too late his true interest was hanging far above the earth.

That night, I experienced a moment of crystallization in the field I had, up to this point, been somewhat blindly pursuing, because the opportunities had led to it, because I was good at it. Data science really is everywhere - even in the stars - and though my own personal dreams of becoming an astronomer or astrophysicist were probably long gone at this point, my studies of applied math, machine learning, and data mining had given me tools which would allow me to explore, at least in some way, some of the objects of which I had once only dreamed.

The Data

Anyone with even a casual interest in astronomy will regularly encounter statistics regarding cosmological phenomena, which aim to infer information about the behavior of said phenomena, perhaps for the purpose of aiding in prediction of these phenomena, for describing their behavior.

Interested in the kind of statistical analyses which might be useful for tracking cosmological phenomenon, I came across a possible project investigating data taken on meteorites, that is, meteors which have fallen to earth. Particularly, I found myself asking questions regarding the rates at which meteorites have fallen throughout the past few decades, regarding whether or not certain locations seem to experience a far greater number of meteorite impacts, and others.

Thus to the end of answering these initial guiding questions, in this project, I explore data from the NASA's online databases. Namely, I investigate the meteorite landings dataset available online. This dataset included 45,717 individual records of meteorites and meteorite fragments, identified to a time period spanning **2500 B.C.E to 2013 A.D.E**. It represents data collected by the meteorological society, and though the NASA website claims that the Meteorological society has an updated version of this dataset, I could not find it available online without some serious webscraping involved.

The original dataset included ten variables with the following labels: name (string) - the given name of the meteorite id (integer) - the Identification number used in the dataset nametype (string) - whether or not the name has been recognized as valid or **relict**(i.e. meteorites "which are dominantly (>95%) composed of secondary minerals formed on the body on which the object was found"Guidelines for meteorite nomenclature, §1.2c) recclass (string) - a classification of the meteorite which gives information about its chemical composition, structure, etc mass (g) (numeric) - the mass of the object in grams year (string) - in the format MM/DD/YY 00:00:00 AM. Most entries just give the date of 01/01/YY 12:00:00 AM. reclat - recovery latitude reclong - recovery longitude Geolocation - a touple of (reclat, reclong)

Initially, this dataset needs **a lot** of cleaning. Many records are missing, and many others are just unclear or not useful.

First, though - here are my external source files and working directory setups

```
## Warning: package 'stringr' was built under R version 3.0.3

## Warning: package 'beepr' was built under R version 3.0.3

## Warning: package 'knitr' was built under R version 3.0.3

## Loading required package: rjags

## Warning: package 'rjags' was built under R version 3.0.3

## Loading required package: coda

## Warning: package 'coda' was built under R version 3.0.3
```

```
## Linked to JAGS 3.4.0
## Loaded modules: basemod,bugs

## Warning: package 'boot' was built under R version 3.0.3

## Warning: package 'sandwich' was built under R version 3.0.3

## Warning: package 'e1071' was built under R version 3.0.3
```

The dataset was downloaded as a CSV, and cast into a data frame.

```
## 'data.frame':    45716 obs. of  10 variables:
##  $ name       : Factor w/ 45716 levels "Ã�sterplana 002",..: 68 69 73 77
473 484 496 497 502 521 ...
##  $ id         : int  1 2 6 10 370 379 390 392 398 417 ...
##  $ nametype   : Factor w/ 2 levels "Relict","Valid": 2 2 2 2 2 2 2 2 2 2
...
##  $ recclass   : Factor w/ 466 levels "Acapulcoite",..: 333 197 85 1 339 85
360 190 339 242 ...
##  $ mass..g.   : num  21 720 107000 1914 780 ...
##  $ fall       : Factor w/ 2 levels "Fell","Found": 1 1 1 1 1 1 1 1 1 1 ...
##  $ year       : Factor w/ 270 levels "","01/01/1583 12:00:00 AM",..: 124
195 196 221 146 163 193 59 174 164 ...
##  $ reclat     : num  50.8 56.2 54.2 16.9 -33.2 ...
##  $ reclong    : num  6.08 10.23 -113 -99.9 -64.95 ...
##  $ GeoLocation: Factor w/ 17101 levels "","(-1.002780, 37.150280)",..:
16779 16983 16923 9106 844 14808 16496 16453 784 721 ...

##                name             id           nametype        recclass
##  Ã�sterplana 002:    1   Min.   :    1   Relict:   75   L6     : 8285
##  Ã�sterplana 003:    1   1st Qu.:12689   Valid :45641   H5     : 7142
##  Ã�sterplana 004:    1   Median :24262                  L5     : 4796
##  Ã�sterplana 005:    1   Mean   :26890                  H6     : 4528
##  Ã�sterplana 006:    1   3rd Qu.:40657                  H4     : 4211
##  Ã�sterplana 007:    1   Max.   :57458                  LL5    : 2766
##  (Other)        :45710                                  (Other):13988
##     mass..g.          fall                          year
##  Min.   :       0   Fell : 1107   01/01/2003 12:00:00 AM: 3323
##  1st Qu.:       7   Found:44609   01/01/1979 12:00:00 AM: 3046
##  Median :      33                 01/01/1998 12:00:00 AM: 2697
##  Mean   :   13278                 01/01/2006 12:00:00 AM: 2456
##  3rd Qu.:     203                 01/01/1988 12:00:00 AM: 2296
##  Max.   :60000000                 01/01/2002 12:00:00 AM: 2078
##  NA's   :131                      (Other)               :29820
##      reclat          reclong                          GeoLocation
##  Min.   :-87.37   Min.   :-165.43                           : 7315
##  1st Qu.:-76.71   1st Qu.:   0.00   (0.000000, 0.000000)    : 6214
##  Median :-71.50   Median :  35.67   (-71.500000, 35.666670) : 4761
##  Mean   :-39.12   Mean   :  61.07   (-84.000000, 168.000000): 3040
##  3rd Qu.:  0.00   3rd Qu.: 157.17   (-72.000000, 26.000000) : 1505
##  Max.   : 81.17   Max.   : 354.47   (-79.683330, 159.750000):  657
##  NA's   :7315     NA's   :7315      (Other)                 :22224
```

I can get rid of some of the columns from the original dataset. Really, only the name, mass, year, location, and the kind of meteorite are useful. The validity of the name doesn't seem to be something I'd want to measure. I also drop the toupled GeoLocation column, because it will be easier to parse the individual columns, rather than a touple. I also change some of the column names for simplicity's sake. Finally, I clean up the **year** column of the data, such that the levels for that column are only the years themselves, and we don't have to deal with inconsistently collected times-of-day+days+months.

```
#changing column names
colnames(raw_dataset)[1] <- 'name'
colnames(raw_dataset)[5] <- 'mass'

#subset of the data, limited for useful columns
limited_dataset <-
raw_dataset[,c('name','recclass','mass','year','reclat','reclong')]

#parsing the year column to extract just the year
for (date in levels(limited_dataset$year)){#data is a string of format
"MM/DD/YYYY HH:MM:SS AM"
  if (date != "" && date != "NA"){ #some dates are empty or NAs
    new_date <- "NA"  #make sure all empties become NAs
  }
  else{ #the date is there
    new_date <- unlist(str_split(date,"/"))[3] #split on the / in the date,
and take whatever follows the second split
    new_date <- unlist(str_split(new_date," "))[1] #and then split on the
remaining space, and take the date before the time
    #print(new_date)
  }
  levels(limited_dataset$year)[levels(limited_dataset$year) == date] <-
new_date #wherever we were, update it
}
limited_dataset$year <- as.numeric(as.character(limited_dataset$year)) #and
cast it as a numeric

## Warning: NAs introduced by coercion
```

Data Extension, further cleaning

Thanks to the meteorological institute, we can expand some of the information from the recclass label. The label corresponds with certain information regarding the composition and structure of the meteorite. This requires some minor webscraping.

```
#a vector of the unique classes in the classifications
recclass_factors <- unique(limited_dataset$recclass)

#attaching the classification to the url pulls up a webpage with the
interesting information
url_prefix <- "http://www.lpi.usra.edu/meteor/metbullclass.php?sea="
```

```
#the unique extensions are the extended information for the unique classification - the cut extension is that information compacted into a more usable format
if (!file.exists("cut_extensions.csv")){#simply comment this if you want to write the file anyway
  unique_extensions <- vector(length=length(recclass_factors))
  cut_extensions <- vector(length=length(recclass_factors))
  for (f in 1:length(recclass_factors)){#for loop for scraping
    url_full <- paste(url_prefix,recclass_factors[f],sep="")
    url_full <- gsub(" ","",url_full)
    #print(recclass_factors[f],max.levels=0)
    webpage <- readLines(url_full)
    html_extract <- webpage[grep(recclass_factors[f],webpage)][1]
    plain_extract <- html_strip(html_extract) #helper function from meteor_helper.R
    remove_this <- paste("The recommended classification ", recclass_factors[f], " means:\"",sep="")
    unique_extensions[f] <- plain_extract
    cut_extensions[f] <- extract_between(plain_extract,remove_this,"\\.") #another helper function from meteor_helper.R
  }

write.table(cut_extensions,file="cut_extensions.csv",sep=",")
}else{
  cut_extensions <- read.csv("cut_extensions.csv",header=FALSE)
}

# This code was used originally to help diagnose and fix some holes which were appearing in early iterations of the method above. Perhaps useful if further changes are made.
if (FALSE){
  fill_ins <- vector()
  for (f in 1:length(cut_extensions)){
    if (cut_extensions[f] == ""){
      tmp <- paste(url_prefix,recclass_factors[f],sep="")
      fill_ins <- c(fill_ins,gsub(" ","",tmp))
      print(paste(f,": ",url_prefix,recclass_factors[f],sep=""))
    }
  }
}
```

Ultimately, I fixed up the data in excel, and was forced to make some adjustments to some of the categories to avoid an explosion in dimensionality. I'll describe that process more in the end.

```
fixed_extensions <- read.table('fixed_extensions.csv',header=FALSE, sep=",", stringsAsFactors=FALSE)
recclass_sorted <- recclass_factors[fixed_extensions$V1]
str(fixed_extensions)
```

```
## 'data.frame':    466 obs. of  8 variables:
##  $ V1: int  4 283 343 387 115 23 26 333 162 117 ...
##  $ V2: chr  "Acapulcoite" "Acapulcoite/Lodranite" "Acapulcoite/lodranite"
"Achondrite-prim" ...
##  $ V3: chr  "achondrite" "achondrite" "achondrite" "achondrite" ...
##  $ V4: chr  "sec:primitive" "sec:primitive" "sec:primitive"
"sec:primitive" ...
##  $ V5: chr  "family:acapulcoite-lodranite" "family:acapulcoite-lodranite"
"family:acapulcoite-lodranite" "" ...
##  $ V6: chr  "" "" "" "" ...
##  $ V7: chr  "" "" "" "" ...
##  $ V8: chr  "" "" "" "" ...
```

The scraping provides a whole new wealth of information which will allow for interesting analyses of the meteorite dataset. and now, we can generate a table for the extensions, which will give us some awesome variables: (1) Meteorite Class (all entries),(factor) (2) Secondary Class (only some entries),(factor) (3) group (a further subsetting tool within classes,(factors) (4) family (only some entries),(factors) (5) chemical group (Iron meteorites only),(factors) (6) petrologic type (Chondrites only),(integer:1-7) (7) is breccia (all entries),(binary:0-1) (8) petrologic class (Mesosiderites only),(factor) (9) metamorphic grade (Mesosiderites only),(integer:1-4) (10) martian type (Martian only),(factor) (11) type of lithologies present (Lunar only),(factor) (12) type of melting present (all entries),(factor) most of these designations may allow for subsetting and classification, perhaps more useful in future projects.

In this section, I apply this extended data to the old data.

```r
if (!file.exists("data_full.csv")){
  #empty dataframe with the correct columns to be added to the old data
  empty_df <- data.frame(ID=recclass_sorted,
                        MeteorClass=fixed_extensions$V3,

SecondClass=vector(mode='character',length=length(recclass_sorted)),

Group=vector(mode='character',length=length(recclass_sorted)),

Family=vector(mode='character',length=length(recclass_sorted)),

ChemGroup=vector(mode='character',length=length(recclass_sorted)),

PetroType=vector(mode='character',length=length(recclass_sorted)),

Breccia=vector(mode='character',length=length(recclass_sorted)),

PetroClass=vector(mode='character',length=length(recclass_sorted)),

MetaGrade=vector(mode='character',length=length(recclass_sorted)),

MarsType=vector(mode='character',length=length(recclass_sorted)),
```

```
Lithol=vector(mode='character',length=length(recclass_sorted)),

Melt=vector(mode='character',length=length(recclass_sorted)),

Other=vector(mode='character',length=length(recclass_sorted)),
                         stringsAsFactors=FALSE)
  ext_df <- empty_df
  # now, to fill it up
  for (irow in 1:nrow(fixed_extensions)){
    cat(paste("Row:",irow,"\n",sep=""))
    for (icol in 4:ncol(fixed_extensions)){
      key <-
substr(fixed_extensions[irow,icol],1,regexpr(":",fixed_extensions[irow,icol])
[1]-1)
      val <-
substr(fixed_extensions[irow,icol],regexpr(":",fixed_extensions[irow,icol])[1
]+1,nchar(fixed_extensions[irow,icol]))
      if (key != ""){
        switch(key,
               sec={ext_df[irow,]$SecondClass<-val},
               group={ext_df[irow,]$Group<-val},
               petrologictype={ext_df[irow,]$PetroType<-val},
               family={ext_df[irow,]$Family<-val},
               chemicalgroup={ext_df[irow,]$ChemGroup<-val},
               breccia={ext_df[irow,]$Breccia<-val},
               petrologicclass={ext_df[irow,]$PetroClass<-val},
               metamorphicgrade={ext_df[irow,]$MetaGrade<-val},
               type={ext_df[irow,]$MarsType<-val},
               lithologies={ext_df[irow,]$Lithol<-val},
               melt={ext_df[irow,]$Melt<-val},
               other={ext_df[irow,]$Other<-val}
        )
      }
    }
  }
  ext_df <- data.frame(lapply(ext_df,as.factor),stringsAsFactors=TRUE)

  # Now, to add this information to the original dataset...
  new_extension <- data.frame()
  for (irow in 1:nrow(limited_dataset)){
    cat(paste("row:",irow,"\n",sep=""))
    rows <-ext_df[ext_df$ID == limited_dataset[irow,]$recclass,]
    new_extension <- rbind(new_extension,rows[1,])
  }
  beep()
  data_full <- cbind(limited_dataset,new_extension)
  write.table(new_full_extension,"data_full.csv",sep=",",row.names=FALSE)
}else{
  data_full <- read.csv("data_full.csv",header=TRUE)
}
```

```
str(data_full)

## 'data.frame':    45716 obs. of  20 variables:
##  $ name       : Factor w/ 45716 levels "Ã�sterplana 002",..: 77 964 1493
1940 2243 2316 2369 3766 6322 6330 ...
##  $ recclass   : Factor w/ 466 levels "Acapulcoite",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ mass       : num  1914 7.9 8.6 3.87 40 ...
##  $ year       : int  1976 1984 1977 1977 1981 1981 1981 1988 2003 2000 ...
##  $ reclat     : num  16.9 -76.7 -76.7 -76.7 -76.7 ...
##  $ reclong    : num  -99.9 159.3 159.7 159.7 159.3 ...
##  $ ID         : Factor w/ 455 levels "Acapulcoite",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ MeteorClass: Factor w/ 13 levels "achondrite","chondrite",..: 1 1 1 1 1
1 1 1 1 1 ...
##  $ SecondClass: Factor w/ 7 levels "","carbonaceous",..: 6 6 6 6 6 6 6 6 6
6 ...
##  $ Group      : Factor w/ 29 levels "angrite","aubrite",..: 27 27 27 27 27
27 27 27 27 27 ...
##  $ Family     : Factor w/ 2 levels "","acapulcoite-lodranite": 2 2 2 2 2 2
2 2 2 2 ...
##  $ ChemGroup  : Factor w/ 21 levels "","ES","IAB",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ PetroType  : Factor w/ 25 levels "","1","1&2","1|2",..: 1 1 1 1 1 1 1 1
1 1 ...
##  $ Breccia    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PetroClass : Factor w/ 4 levels "","A","B","C": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MetaGrade  : Factor w/ 6 levels "","1","2","3",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ MarsType   : Factor w/ 4 levels "","chassignite",..: 1 1 1 1 1 1 1 1 1
1 ...
##  $ Lithol     : Factor w/ 9 levels "","anorthositic",..: 1 1 1 1 1 1 1 1 1
1 ...
##  $ Melt       : Factor w/ 4 levels "","breccia","impact",..: 1 1 1 1 1 1 1
1 1 1 ...
##  $ Other      : Factor w/ 8 levels "","basaltic clasts",..: 1 1 1 1 1 1 1
1 1 1 ...

summary(data_full)

##                 name           recclass          mass                 year
##  Ã�sterplana 002:    1   L6      : 8285   Min.   :        0   Min.   : 301
##  Ã�sterplana 003:    1   H5      : 7142   1st Qu.:        7   1st Qu.:1987
##  Ã�sterplana 004:    1   L5      : 4796   Median :       33   Median :1998
##  Ã�sterplana 005:    1   H6      : 4528   Mean   :    13278   Mean   :1992
##  Ã�sterplana 006:    1   H4      : 4211   3rd Qu.:      203   3rd Qu.:2003
##  Ã�sterplana 007:    1   LL5     : 2766   Max.   :60000000   Max.   :2501
##  (Other)        :45710   (Other):13988   NA's   :131         NA's   :288
##      reclat           reclong              ID              MeteorClass
##  Min.   :-87.37   Min.   :-165.43   L6      : 8340   chondrite    :42167
```

```
##  1st Qu.:-76.71   1st Qu.:   0.00   H5      : 7164   achondrite  : 1837
##  Median :-71.50   Median :  35.67   L5      : 4817   iron        : 1070
##  Mean   :-39.12   Mean   :  61.07   H6      : 4530   mesosiderite:  187
##  3rd Qu.:  0.00   3rd Qu.: 157.17   H4      : 4221   lunar       :  165
##  Max.   : 81.17   Max.   : 354.47   LL5     : 2766   martian     :  119
##  NA's   :7315     NA's   :7315      (Other):13878   (Other)     :  171
##        SecondClass          Group                              Family
##              : 3397   H         :17873                            :45612
##  carbonaceous: 1582   L         :15841   acapulcoite-lodranite:  104
##  enstatite   :  530   LL        : 5876
##  kakangari   :    3   ungrouped: 2477
##  ordinary    :39901   eucrite  :  681
##  primitive   :  171   CM        :  460
##  R           :  132   (Other)  : 2508
##    ChemGroup        PetroType         Breccia          PetroClass MetaGrade
##         :44803   6       :15212   Min.   :0.00000    :45673        :45686
##  IIIAB  :  292   5       :15069   1st Qu.:0.00000   A:   18     1  :   10
##  IIAB   :  119   4       : 5985   Median :0.00000   B:   16     2  :   10
##  IAB    :  107           : 4296   Mean   :0.02957   C:    9     3  :    3
##  IAB-MG :   84   3       : 2914   3rd Qu.:0.00000               3|4:    1
##  IVA    :   75   2       :  569   Max.   :1.00000               4  :    6
##  (Other):  236   (Other): 1671
##         MarsType              Lithol              Melt
##             :45601              :45592            :45535
##  chassignite:    2   anorthositic:   69   breccia  :   97
##  nakhlite   :   14   feldspathic :   27   impact   :   44
##  shergottite:   99   basalic     :   16   secondary:   40
##                      gabbroic    :    6
##                      basaltic    :    2
##                      (Other)     :    4
##                          Other
##                             :45664
##  cumulatae                  :   26
##  unusually rich in olivine  :    9
##  sec:enstatite-rich         :    6
##  contains magnesian pyroxene:    4
##  fusion crust               :    4
##  (Other)                    :    3
```

Some further cleaning in excel was needed even after all of this. The provided dataset data_full.csv is what we need.

Analysis

Before the analysis is done, I turn back to some of the original problems I posed in the beginning. Specifically, I look at what kind of questions we can pose. There are many, many interesting questions we could ask about this dataset. Each of these questions will require specific subsetting, preprocessing, and analysis.

For now I just focus on a few possible questions: ### Investigating Impacts over Time: (Q1.1) Has one of the three centuries present experienced significantly more impacts? I have four centuries' worth of data - I might as well look if there's a relationship. I keep in mind that data-gathering techniques have changed significantly as well.

Initial General Hypothesis: The 20th century has experienced significantly more frequent meteorite impacts than other centures. (This is based on an expectation that data collection has been significantly better in this century than in others, and on the fact that the 21st century just hasn't lastedf as long - I want to demonstrate this possible bias of the data)

Given that the data regarding the frequency of impacts is a rare-event, I can expect something like a poisson distribut ion from the frequencies. First, the data needs to be properly subset and cleaned of rows with no date.

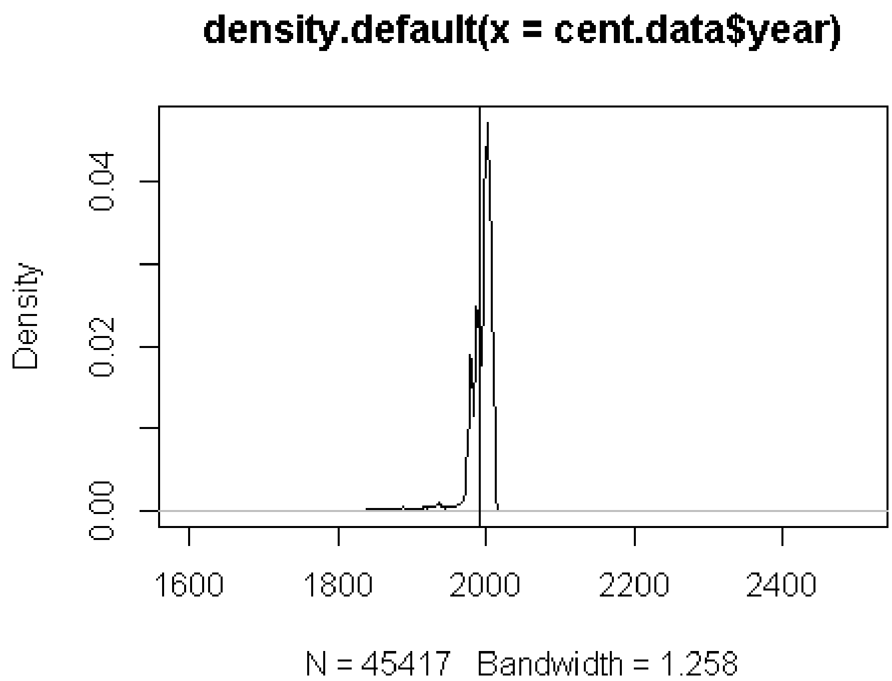
```
cent.data <- data_full[!is.na(data_full$year),]
cent.data <- cent.data[cent.data$year > 1599,] #there are some records of older meteorites, but we don't want them
cent.17 <- cent.data[cent.data$year < 1700,]
cent.18 <- cent.data[cent.data$year < 1800 & cent.data$year >= 1700,]
cent.19 <- cent.data[cent.data$year < 1900 & cent.data$year >= 1800,]
cent.20 <- cent.data[cent.data$year < 2000 & cent.data$year >= 1900,]
cent.21 <- cent.data[cent.data$year >= 2000 & cent.data$year,]
```

Frequency over the entire century needs to be counted, each century vector will have elements corresponding to individual years. The frequency for one year is just the number of impacts in that year.

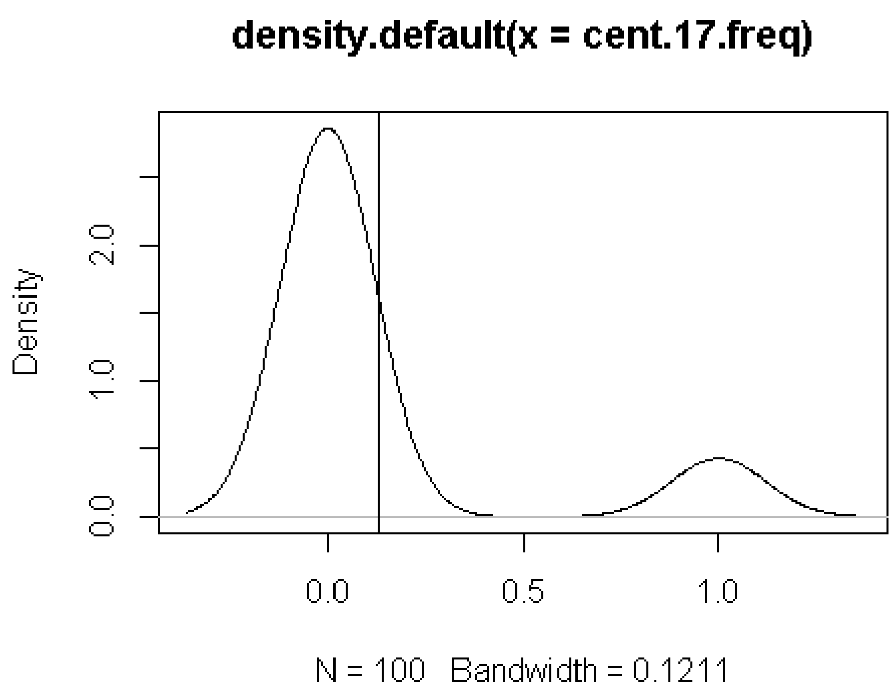
```
freq.count <- function(y,z){return(unlist(lapply(y,function(x){sum(x == z)})))}

cent.17.freq <- freq.count(1600:1699,cent.17$year)
cent.18.freq <- freq.count(1700:1799,cent.18$year)
cent.19.freq <- freq.count(1800:1899,cent.19$year)
cent.20.freq <- freq.count(1900:1999,cent.20$year)
cent.21.freq <- freq.count(2000:2015,cent.21$year)
```

Now, I check the distributions of these frequencies.

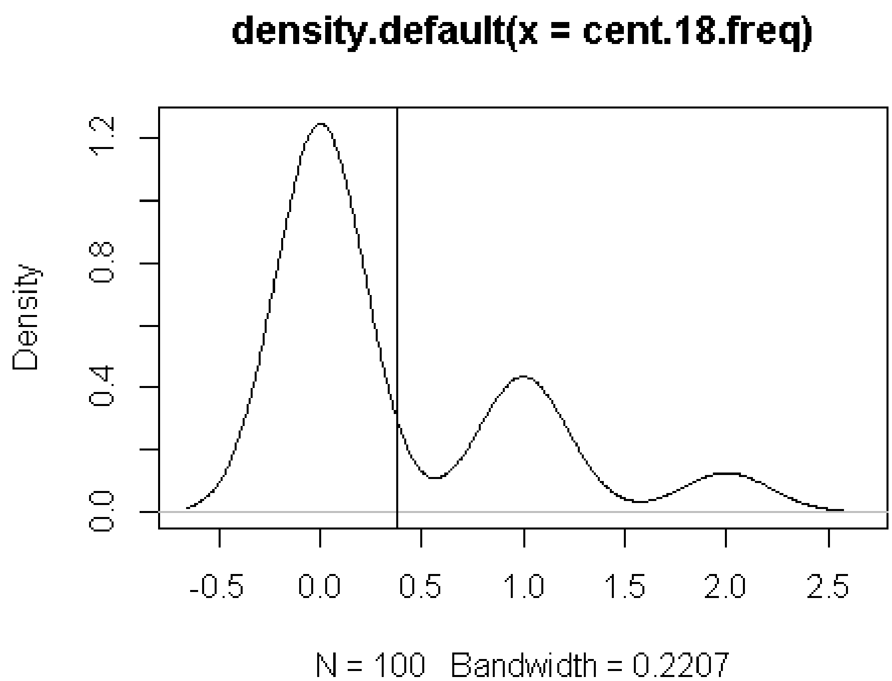
```
plot(density(cent.data$year))
abline(v=mean(cent.data$year))
```



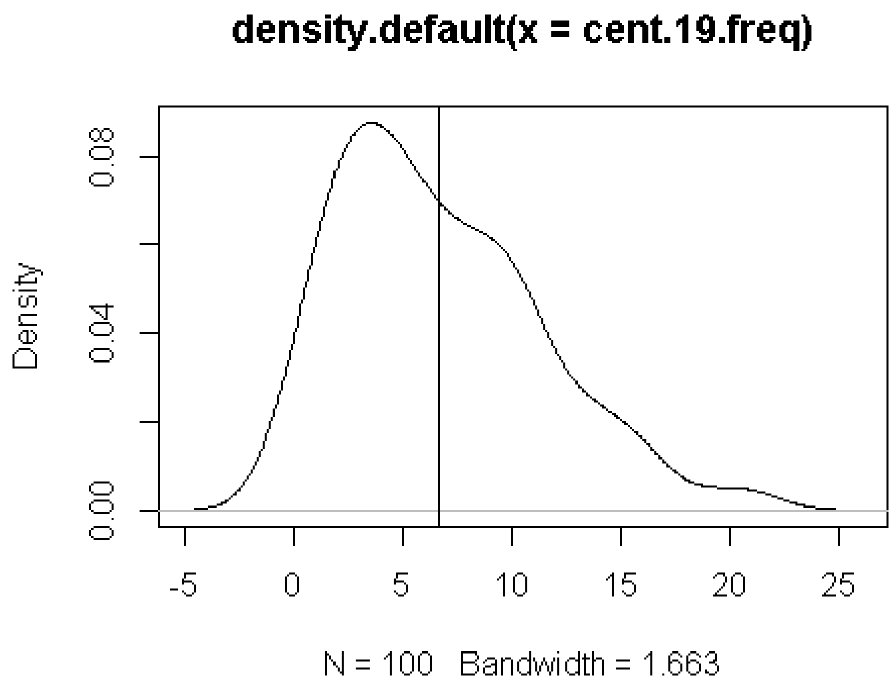

```
plot(density(cent.17.freq))
abline(v=mean(cent.17.freq))
```



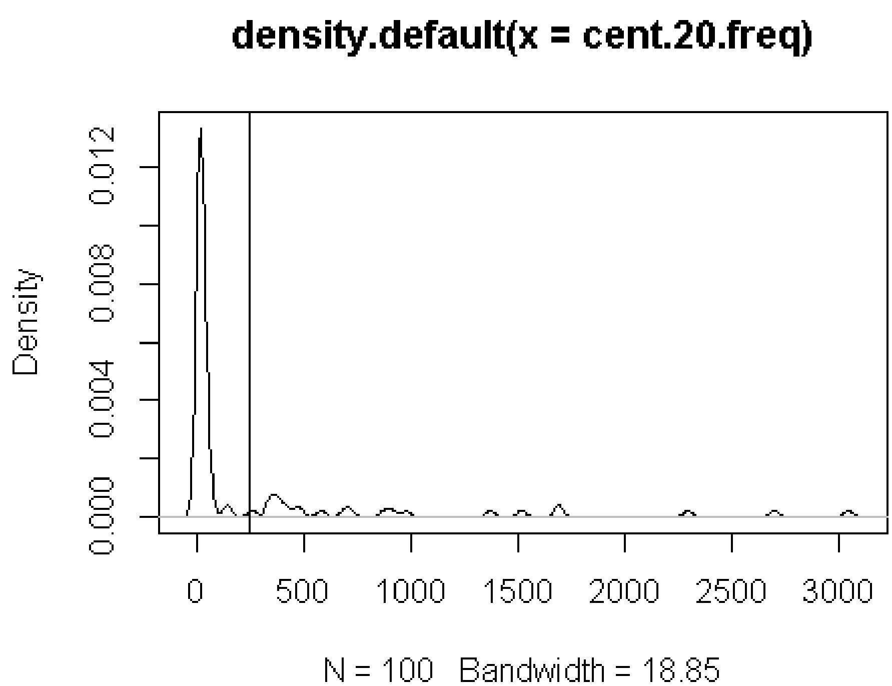
```r
plot(density(cent.18.freq))
abline(v=mean(cent.18.freq))
```



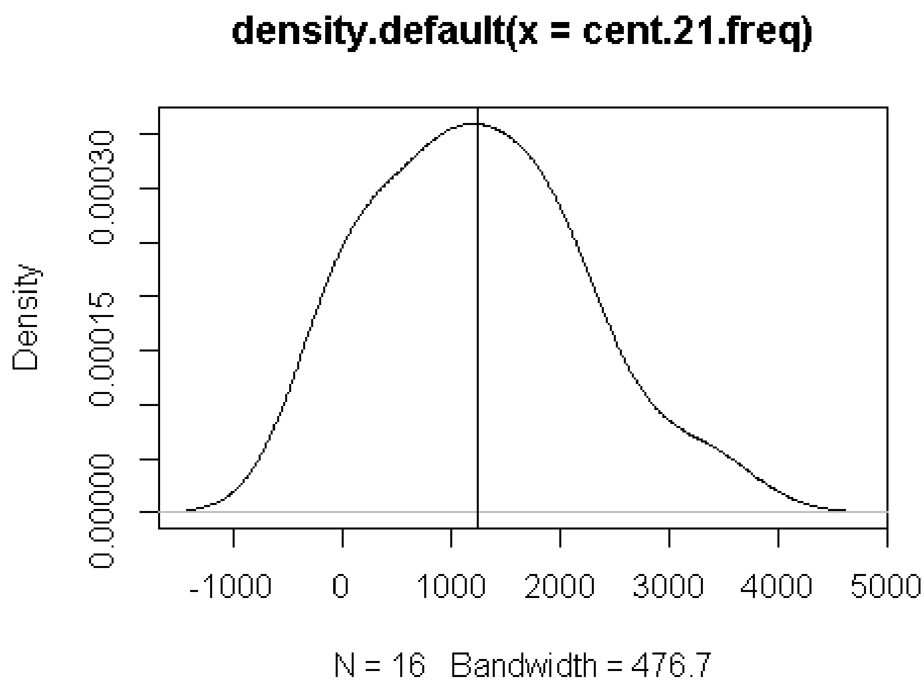

```r
plot(density(cent.19.freq))
abline(v=mean(cent.19.freq))
```




```
plot(density(cent.20.freq))
abline(v=mean(cent.20.freq))
```



```
plot(density(cent.21.freq))
abline(v=mean(cent.21.freq))
```



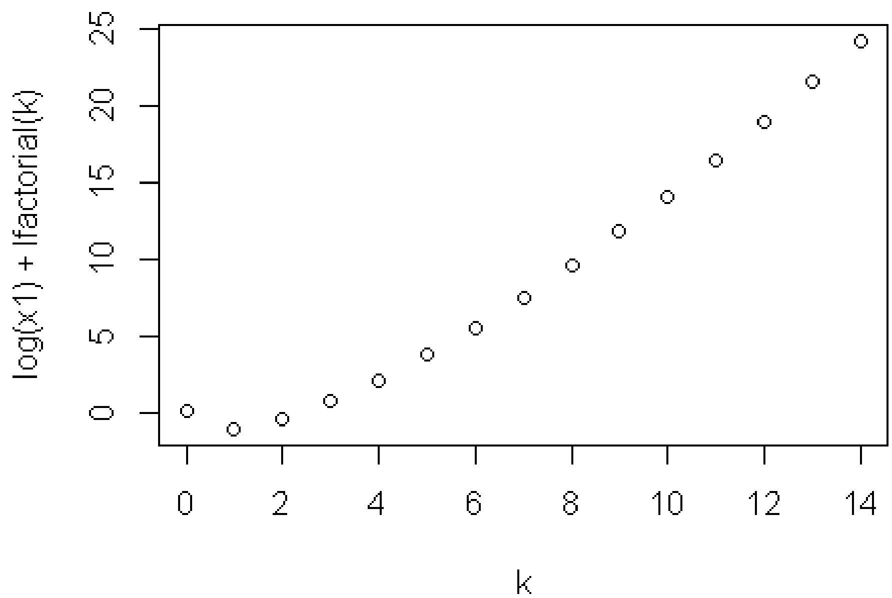
We can see from the plot over all years measured that a heavy frequency is unnormally clustered in the late 20th and early 21st centuries. If we treat the incidence of meteorite strikes as a poisson process, we see that the distribution of frequencies across each century looks somewhat poisson-like. Indeed, these look like rough poisson distributions with each century taking a different value for lambda. Unfortunately, just rough shape isn't enough to confirm **poisson-ness**. I use, instead, a goodness of fit test found in a Hoaglin book, and do 2 bootsraps: David C. Hoaglin (1980), "A Poissonness Plot", The American Statistician Vol. 34, No. 3 (Aug.,), pp. 146-149

and

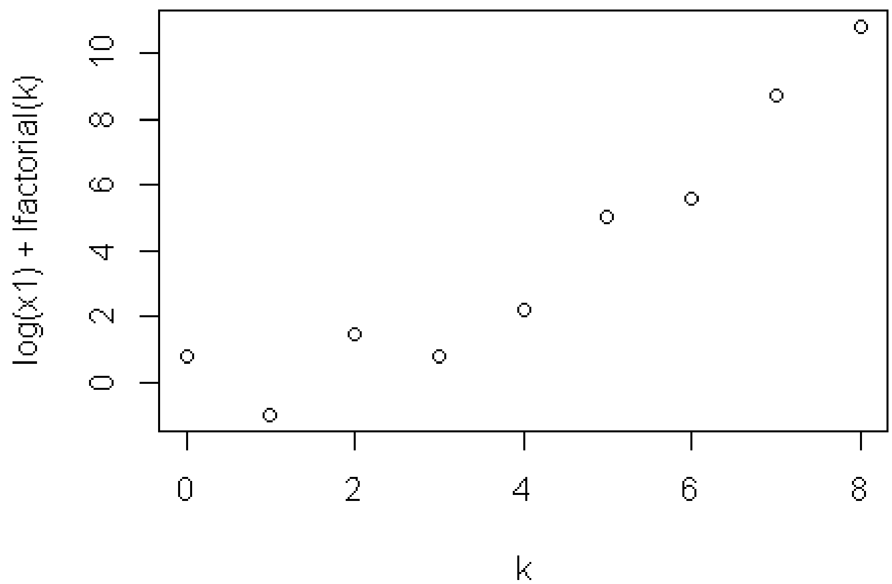
Hoaglin, D. and J. Tukey (1985), "9. Checking the Shape of Discrete Distributions", Exploring Data Tables, Trends and Shapes, (Hoaglin, Mosteller & Tukey eds) John Wiley & Sons

21st century:

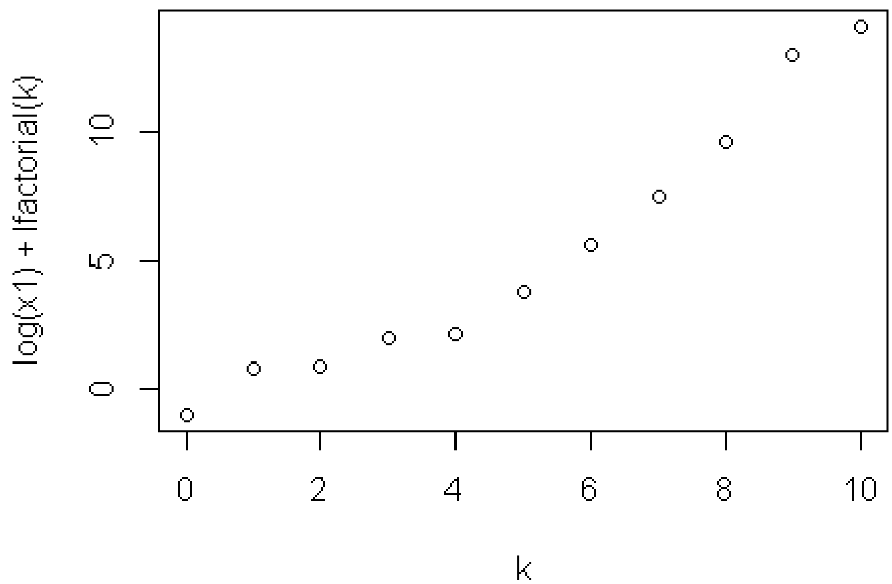
```
cent.21.boot <- boot(cent.21.freq,poissonness_plot ,R=2)

##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
## 
##    0   11  234  713  875  957 1005 1189 1497 1650 1792 1940 2078 2456 3323 
##    2    1    1    1    1    1    1    1    1    1    1    1    1    1    1
```




```
## [1] 0.5383004
##  [1] 12 12  3  8  3 16 14 10 15  1 15 12  6  4 10  4
##
##    0   11  713  875 1189 1497 1792 2078 3323
##    3    1    3    1    1    2    1    2    2
```




```
## [1] 1.403693
##  [1]  2  1  7 16  3  3  5  6  9 13 13 13 12  6 10 12
##
##    0  234  713  875  957 1497 1650 1792 1940 2078 2456
##    1    3    2    2    1    1    1    1    1    2    1
```



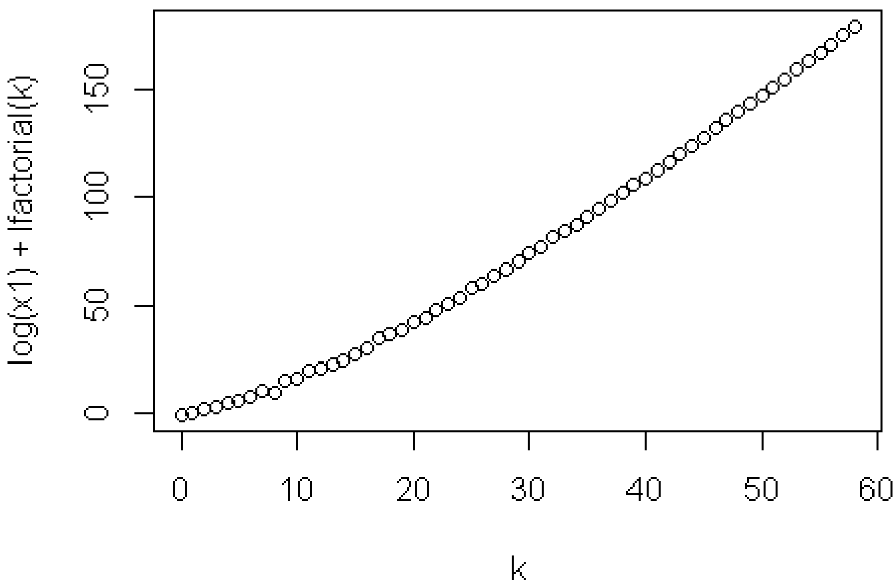

```
## [1] 0.7474243
```

So, it looks like the model may be slightly overfit for the 21st century dataset.

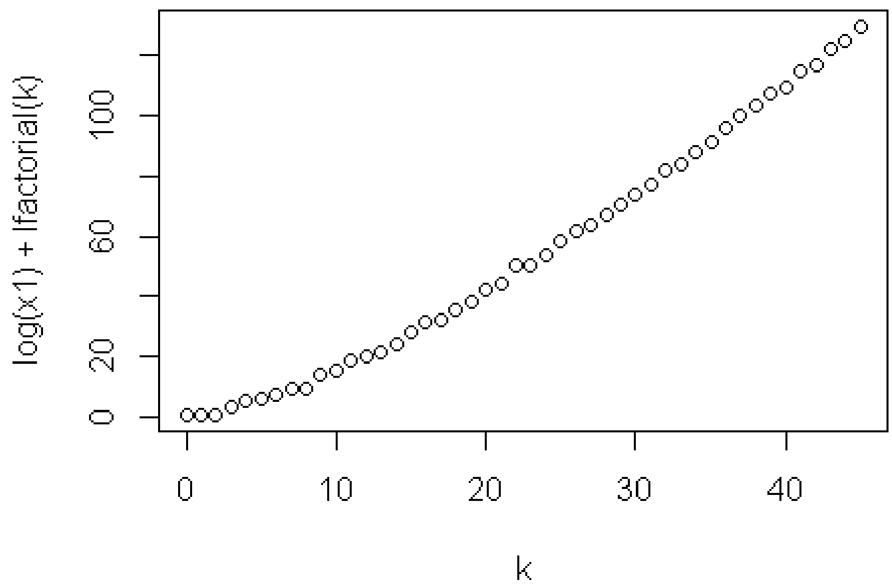
20th Century

```
cent.20.boot <- boot(cent.20.freq,statistic=poissonness_plot ,R=2)
```

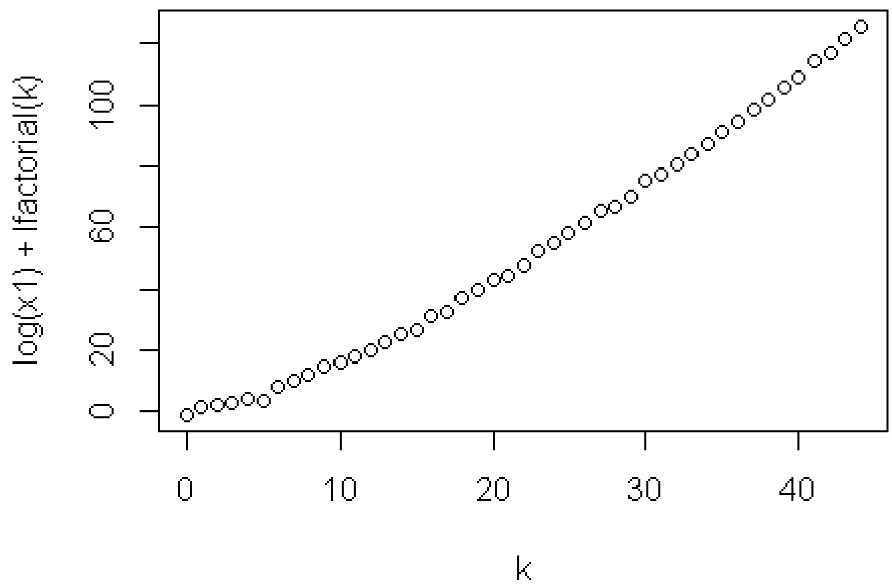
```
##   [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
##  [18]  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34
##  [35]  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51
##  [52]  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68
##  [69]  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85
##  [86]  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
## 
##    9   10   11   12   13   14   15   16   17   18   19   20   22   23   24 
##    1    2    4    3    5    3    3    5    1    7    3    7    3    2    1 
##   25   26   27   30   31   32   33   34   35   36   37   40   45   48   49 
##    1    1    3    2    1    2    1    1    1    1    2    1    1    1    1 
##   50   52   54   70  136  152  262  337  344  360  372  378  402  421  463 
##    1    1    2    1    1    1    1    1    1    1    1    1    1    1    1 
##  487  583  691  719  877  916  979 1375 1518 1691 1696 2296 2697 3046 
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1
```




```
## [1] 1.099633
##   [1]  68  20  55  76  67  96  21  77   3  94  41  53   7  87   7  57  54
##  [18]  22  50  84  81  93  41  31  67  27  47  38  65  98  92  64  60  56
##  [35]  96  89  92  73  57  21  21  90  25  37   5  17  36  79  12  11  51
##  [52]   2  27  98  82  68  19  38  54  52   1  44 100  80  94  71   8  25
##  [69]  30  41  59  44  49  80  12  91  16  17  10  68  74  83  64  57   9
##  [86]  77  86  82   7   9  30  29   3  54  48  70  12  32  91  83
##
##    9   10   11   12   13   14   15   16   17   18   19   20   22   23   24
##    3    3    2    5    9    4    3    3    1    5    2    6    2    1    1
##   26   27   31   32   33   34   35   37   40   48   52   54   70  136  152
##    2    3    1    1    1    2    1    6    1    1    2    2    1    1    1
##  262  337  344  360  372  378  463  487  877  979 1375 1518 1691 1696 2296
##    1    1    2    1    1    1    2    2    2    2    1    2    1    2    1
## 3046
##    2
```




```
## [1] 1.344028
##   [1]  86  81  33  13  51  66  89  72  36  83  77   9  60  83  30  67  67
##  [18]  83  36  72  33  39   6  12  49  16  12  64  20   3  43  99  15  59
##  [35]  22  26   9  12  64 100  91  14 100  70  36  87  74  93  14  77  13
##  [52]  57  32  57  20  88  28  98  63  24  10  99  68  84  31  40   4  79
##  [69]  42  60  14  63  52  82  65  64  36  31   2  70  22  66  16  70  50
##  [86]  69  85  73  68  89  45  68  72  92  32  46  96  38  36  29
##
##   10   11   12   13   14   15   16   18   19   20   22   23   25   26   27
##    1    5    6    4    3    1    5    5    5    6    3    2    2    2    2
##   31   32   33   34   36   37   40   45   49   50   52   54   70  152  262
##    1    2    1    3    2    3    1    1    3    2    2    2    3    1    1
##  344  360  372  378  402  463  487  877  916 1375 1518 1691 1696 2296 2697
##    3    1    1    1    1    1    1    1    1    1    1    2    1    2    2
```



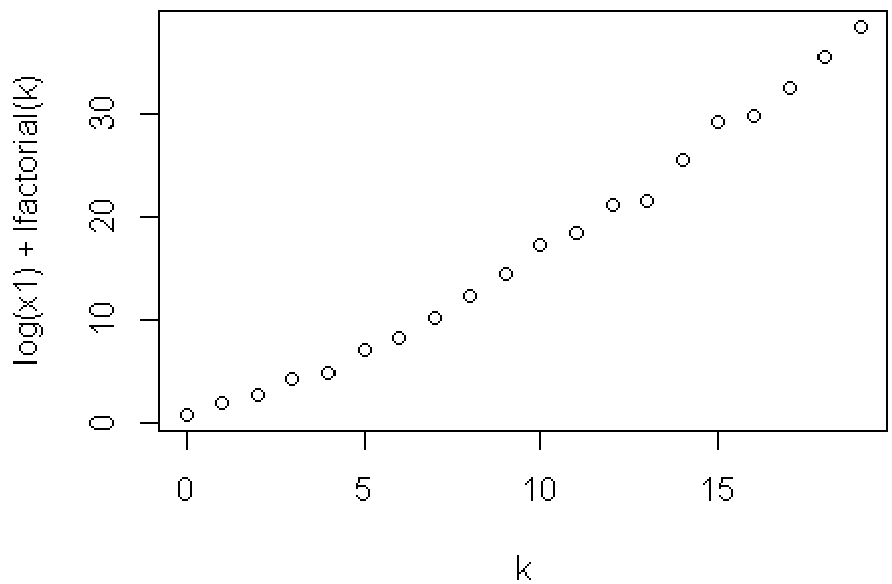

```
## [1] 1.287172
```

It looks like this century is well-modelled by a poisson distribution. I'm going to run this statistic through boot-strapping.

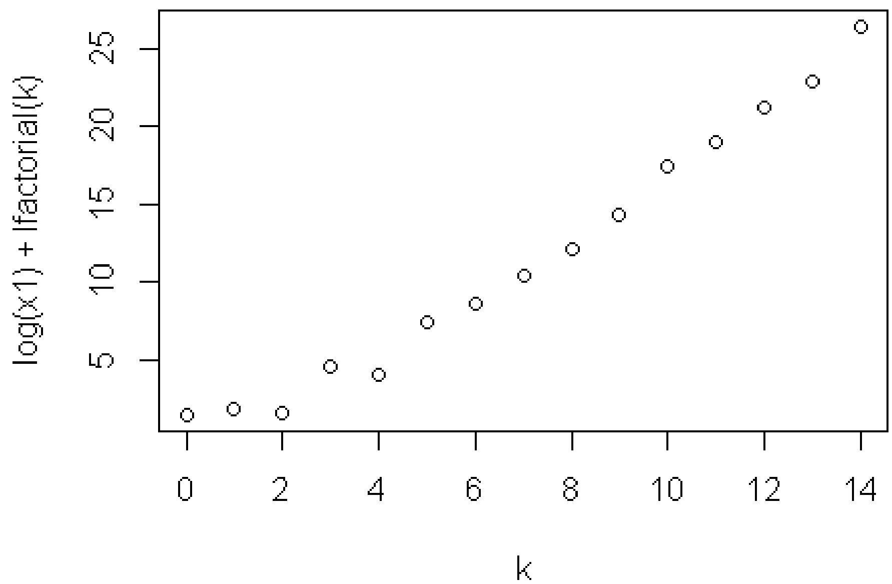
19th century

```
#19th century
cent.19.boot <- boot(cent.19.freq,statistic=poissonness_plot ,R=2)
```

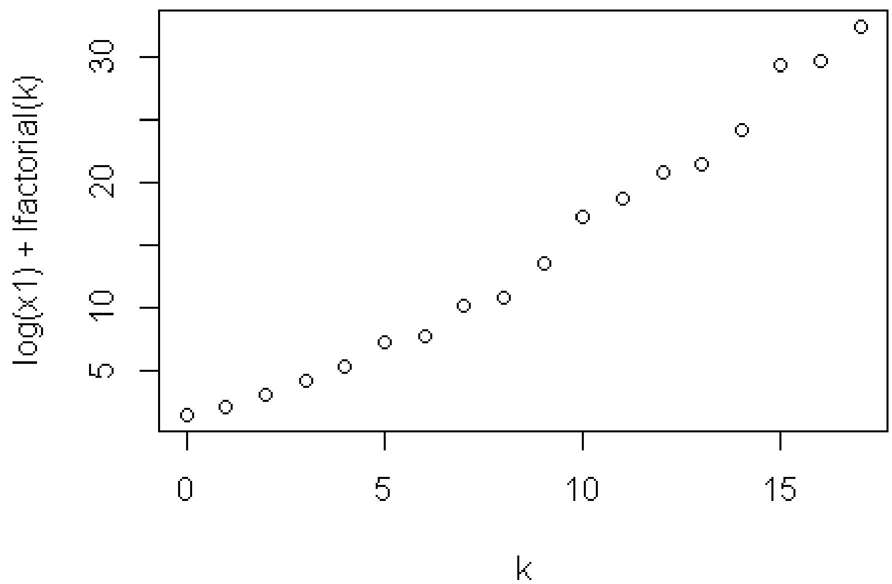
```
##   [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
##  [18]  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34
##  [35]  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51
##  [52]  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68
##  [69]  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85
##  [86]  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
## 
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 20 21 
##  3  8  8 13  6 11  6  6  6  6  9  3  4  1  2  4  1  1  1  1
```




```
## [1] 1.780111
##   [1]  9 38 94 86 51  4 26 91 47  3 52 70 57  9 91 27  1 50  7 75  2 66 86
##  [24]  3 33 84 30 93 91 48 78 51 98  3 97 91 58  2 84 19 60 78 92 71 96 52
##  [47] 82 65 93 69  5  1 48 77 55 87 29 12 57 57 95 93 93 27 39 43 73 46 28
##  [70] 42 34 58 61 63  9 78 43 25 36 37 34  9 44 59 11 57 50 69  4 27 72 48
##  [93] 59 70 19  2 44 36 20 74
##
##  0  1  2  3  4  5  6  7  8  9 10 11 12 16 20
##  5  7  3 16  3 15  8  7  5  5 11  5  4  2  4
```




```
## [1] 2.472334
##   [1]  16  50  67  69  84   2  53  81  86  64  90  22  89  14  36  16  63
##  [18]  42  65  28  37  74  30  29  85  13  74  17  30  11  76  41  48  43
##  [35]  52  39  10  28   3  42   3  82  61  45  91 100   7  93  72  49  10
##  [52]   7   6  95  92  86  46  94  78  52  74  87   3  18   6  85   3  15
##  [69]  48  14  29  12  14  71  40  28  38  32  98  11   2  37  47  84  62
##  [86]  43  51  29  11  62  21  14  89  56  42  97  48  56  20  71
## 
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 20 
##  5  9 11 12  9 13  4  6  2  3 10  4  3  1  1  5  1  1
```



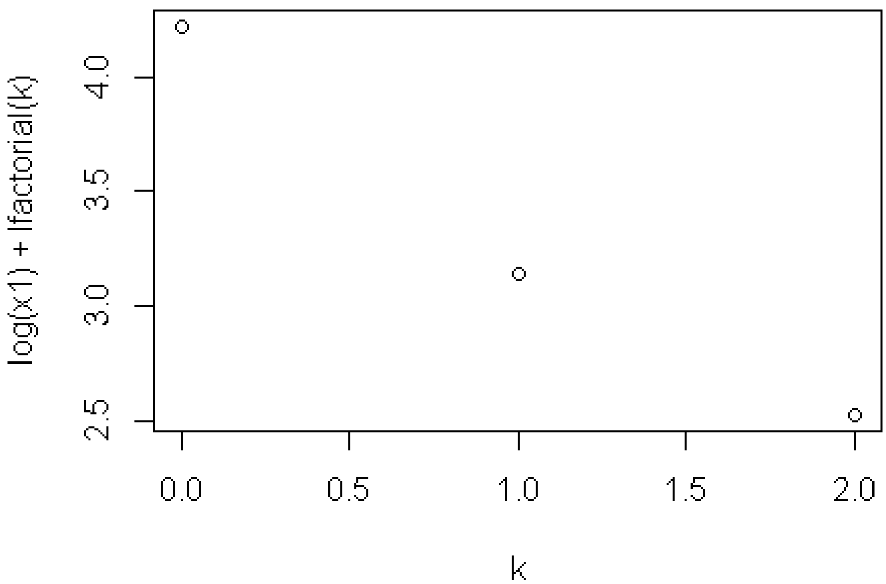

```
## [1] 2.466503
```

The 19th century looks somewhat underfit by the model.

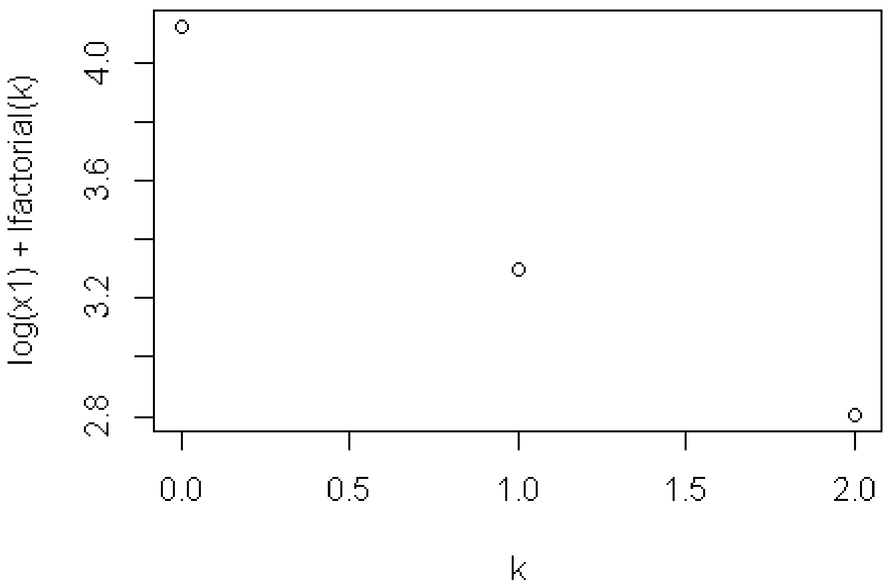
18th century

```
#18th century
cent.18.boot <- boot(cent.18.freq,statistic=poissonness_plot ,R=2)
```

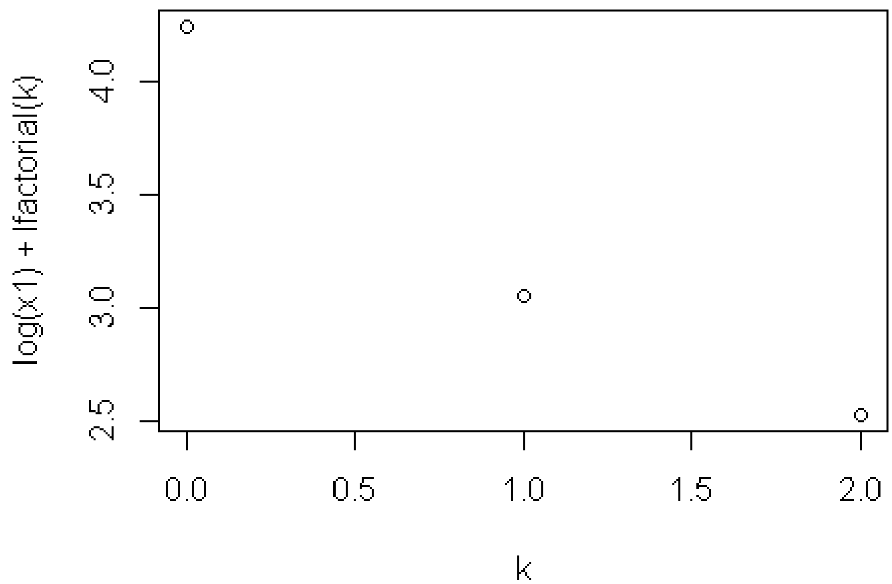
```
##   [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
##  [18]  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34
##  [35]  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51
##  [52]  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68
##  [69]  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85
##  [86]  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
## 
##  0  1  2 
## 69 24  7
```




```
## [1] 3.384186
##   [1]  76   6  26  13   9  81  81  25  17  24  74   3  79  37  50  96  32
##  [18]  48  62  27  60  54  43  97  22  77  10  88  14  87  39  51  43  24
##  [35]  53  47  60  41  77  82  10 100  86  81  55  23  70  19  31  33  90
##  [52]   3  15  82  23  97  86  96   1  20   3  87   8  83  86   4  26  61
##  [69]  63  24  30  84  37  91  53  60  34  85  86  52  20  91  27  87  96
##  [86]  85  67  17  35  21  78  95  51  66  21  67  11  27  52  60
##
##  0  1  2
## 63 28  9
```




```
## [1] 2.812128
##   [1] 71 64 71 20 86 37 85 61 34 32 46 10 77 28 12 54 39 33 85 73 12 31 46
##  [24] 30  1 62 80 31 32 17 88 60 44 84 83 34 18  5 18 33 76  8 97 86 67 86
##  [47] 51 93 64 49 28 68 83 92 24 95 12 10 75 71 22 93 35 64 10 81 61 19 84
##  [70] 17 73 73 79 64 21  4 18 32 80 22 38 64  9 25 49 52 23 39 60 22 53 96
##  [93] 73 50 33 33 50 10 97 24
##
##  0  1  2
## 71 22  7
```



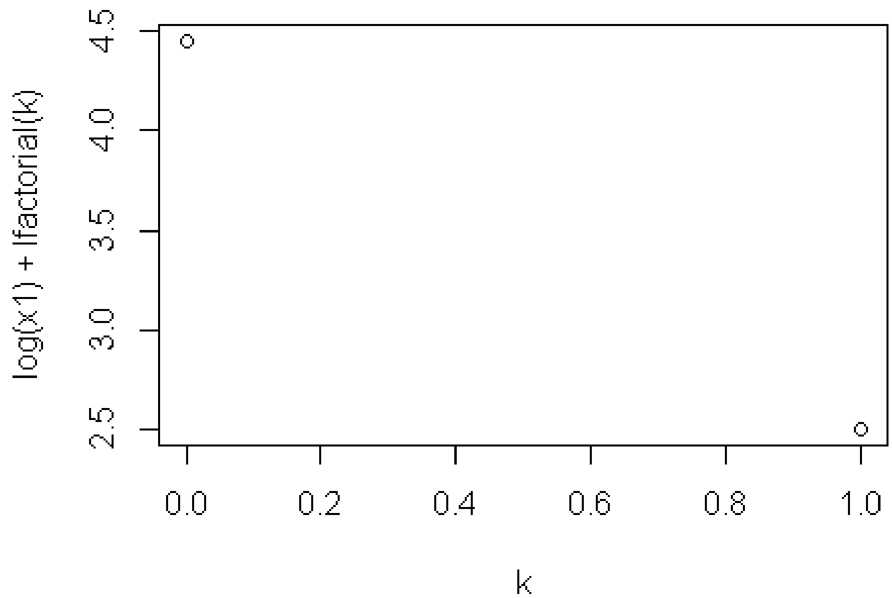

```
## [1] 6.146189
```

The 18th century looks grossly underfit by the model.

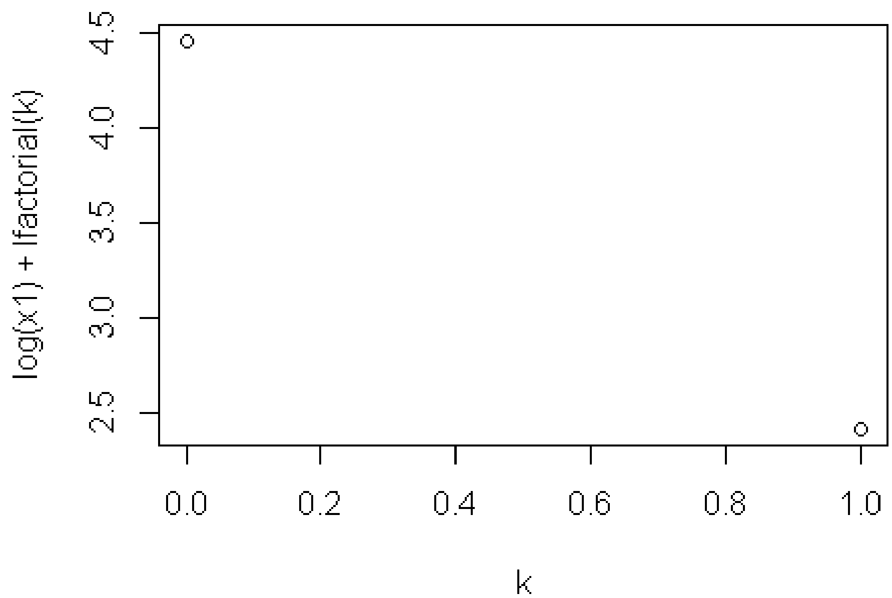
17th century

```
cent.17.boot <- boot(cent.17.freq,statistic=poissonness_plot ,R=2)
```

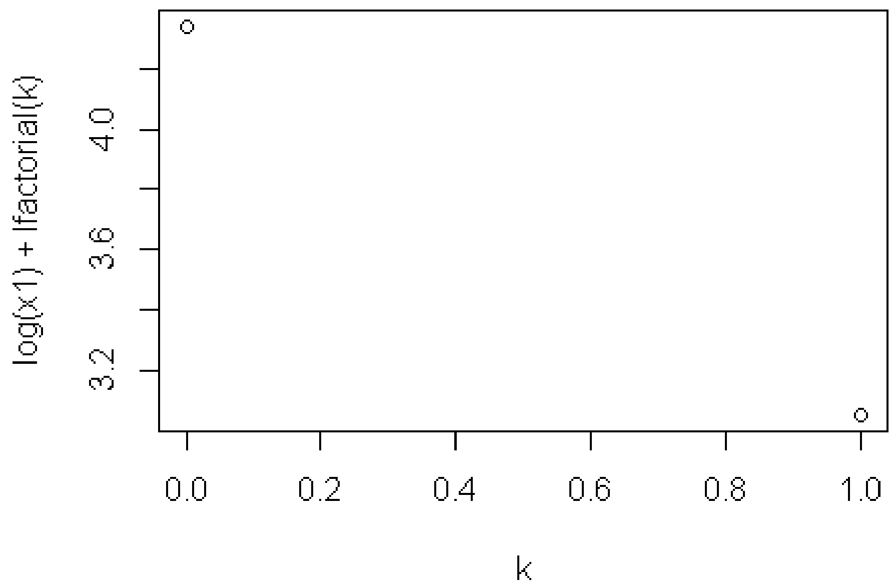
```
##   [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
##  [18]  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34
##  [35]  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51
##  [52]  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68
##  [69]  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85
##  [86]  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
## 
##  0  1 
## 87 13
```




```
## [1] 0
##   [1]  21  39  40  33  98  75  24  23  63  86   7  32  86  29  46  50   8
##  [18]  15  39  76  60  53  22  67 100  79  89  68  23  27  12  43  96  45
##  [35]  81  78  63  54  61  40  27   6  53  60  30  58  82  26  86   8  35
##  [52]  43   1  39   8  75  96  83  41  86  86  36  65  79  50  53  81  45
##  [69]  48  26  60  73  48  93  16  64  53   8  84 100  13  77  67  91  35
##  [86]  20  65  87  79  32  83  45  92  25  89  82  84  75  89  71
## 
##  0  1 
## 88 12
```




```
## [1] 0
##   [1]  99  87  14  59  18  22  17  76  38  43  10  54  63  25  63   6  32
##  [18] 100  17  17  59  43  24  42  16  99  36  22  34  82  80  53  31  37
##  [35]  20  72  12  36  74  43  17  71   9  37  83  70  38  18  23  94  37
##  [52]  60  16  62  51  23  37  53  75  21  31  46  40  62  63  84  83  51
##  [69]  30  78  64  45  89  84  79  35  17  41  29  37  57  26  31  35  29
##  [86]  89   2  11  54  48  24  69  25  29  54  99  14  50  36  19
##
##  0  1
## 78 22
```




```
## [1] 0
```

The 17th century also looks grossly underfit by the model.

It looks like we might not lose too much if we try a poisson test between the 21st,20th,and 19th centuries. Indeed, the models wiggle slightly under bootstrapping, but not by much.

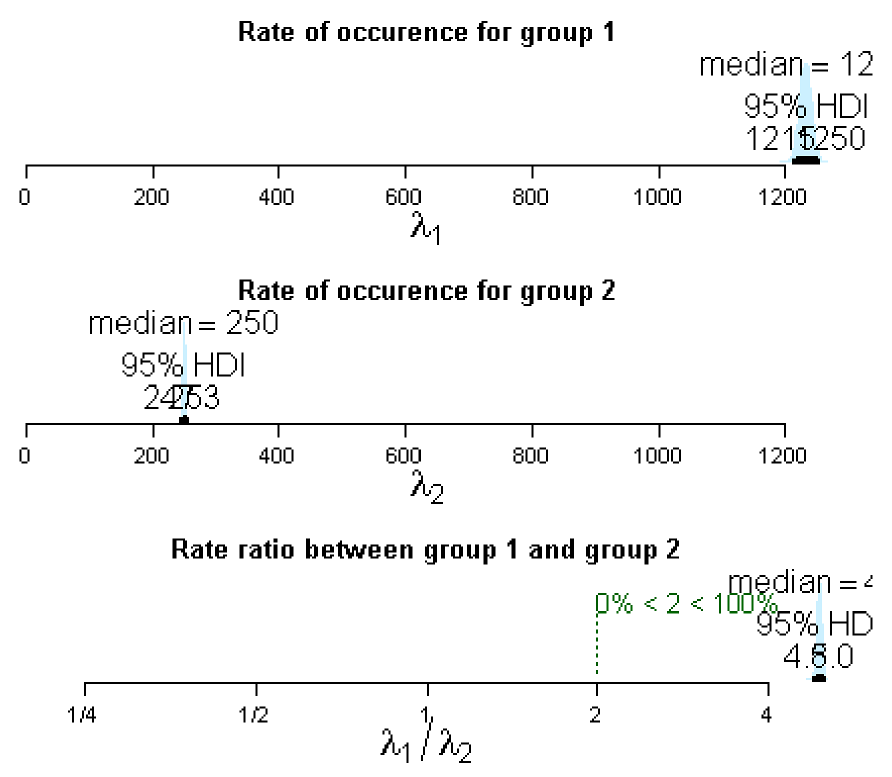
I start by testing the null hypothesis that the poisson-rate of meteorite impacts in the 21st is less than 2x than of the 20th

```
poisson.test(c(sum(cent.21.freq),sum(cent.20.freq)),c(length(cent.21.freq),le
ngth(cent.20.freq)),r=2,alternative="greater")
```

```
##
##  Comparison of Poisson rates
##
## data:  c(sum(cent.21.freq), sum(cent.20.freq)) time base:
c(length(cent.21.freq), length(cent.20.freq))
## count1 = 19720, expected count1 = 6621.63, p-value < 2.2e-16
## alternative hypothesis: true rate ratio is greater than 2
## 95 percent confidence interval:
##  4.857781      Inf
## sample estimates:
## rate ratio
##   4.934737
```

With this test alone, we might reject the null hypothesis, and say there is no evidence that the incidence rate is less than 2x of that in the 20th century.

```
bayes.poisson.test(c(sum(cent.21.freq),sum(cent.20.freq)),c(length(cent.21.fr
eq),length(cent.20.freq)),r=2)

##
##  Bayesian Fist Aid poisson test - two sample
##
## number of events: 19720 and 24976, time periods: 16 and 100
##
##   Estimates [95% credible interval]
## Group 1 rate: 1232 [1215, 1249]
## Group 2 rate: 250 [246, 253]
## Rate ratio (Group 1 rate / Group 2 rate):
##              4.9 [4.8, 5.0]
##
## The event rate of group 1 is more than 2 times that of group 2 by a
probability
## of >0.999 and less than 2 times that of group 2 by a probability of <0.001
.

plot(bayes.poisson.test(c(sum(cent.21.freq),sum(cent.20.freq)),c(length(cent.
21.freq),length(cent.20.freq)),r=2))
```



All of this gives us a strong indication that the incidence rate of meteor strikes in the 21st century is more than 2 times that of the 20th. I could experiment with some different rates to settle on a more exact relationship between the incidence rates, but this gives me enough to reject the initial hypothesis that the 20th century would in general have more incidences.

I test similar null hypotheses for the 21st century vs other centuries. **21st vs 19th**

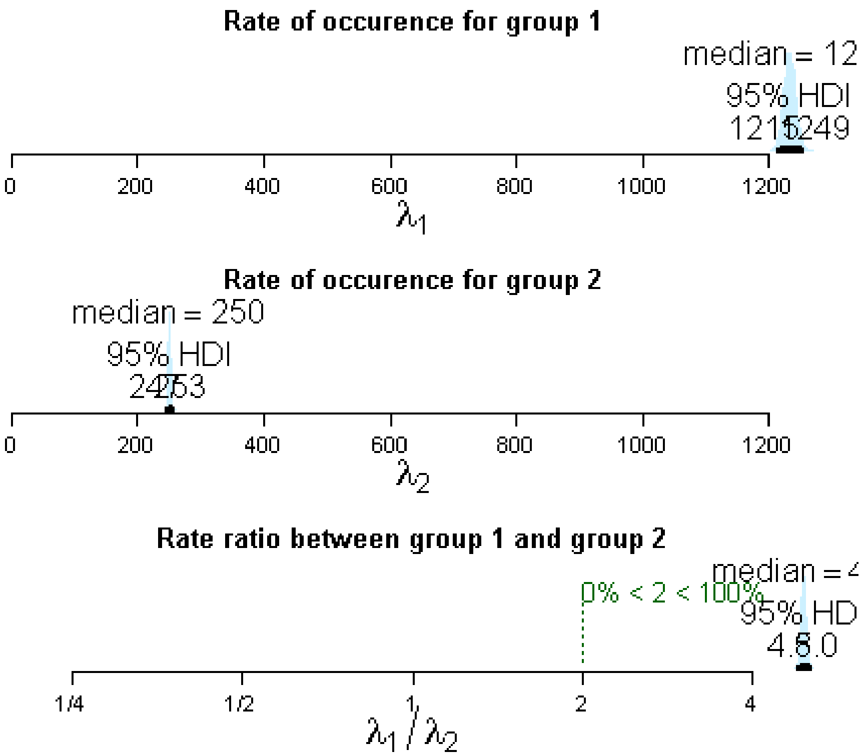
```
poisson.test(c(sum(cent.21.freq),sum(cent.19.freq)),c(length(cent.21.freq),le
ngth(cent.19.freq)),r=2,alternative="greater")
```

```
##
##  Comparison of Poisson rates
##
## data:  c(sum(cent.21.freq), sum(cent.19.freq)) time base:
c(length(cent.21.freq), length(cent.19.freq))
## count1 = 19720, expected count1 = 3020.444, p-value < 2.2e-16
## alternative hypothesis: true rate ratio is greater than 2
## 95 percent confidence interval:
##  172.8865      Inf
## sample estimates:
## rate ratio
##    184.506
```

```
bayes.poisson.test(c(sum(cent.21.freq),sum(cent.20.freq)),c(length(cent.21.fr
eq),length(cent.20.freq)),r=2)
```

```
##
##  Bayesian Fist Aid poisson test - two sample
##
## number of events: 19720 and 24976, time periods: 16 and 100
##
##   Estimates [95% credible interval]
## Group 1 rate: 1233 [1215, 1250]
## Group 2 rate: 250 [247, 253]
## Rate ratio (Group 1 rate / Group 2 rate):
##              4.9 [4.8, 5.0]
##
## The event rate of group 1 is more than 2 times that of group 2 by a
probability
## of >0.999 and less than 2 times that of group 2 by a probability of <0.001
.
```

```
plot(bayes.poisson.test(c(sum(cent.21.freq),sum(cent.20.freq)),c(length(cent.
21.freq),length(cent.20.freq)),r=2))
```

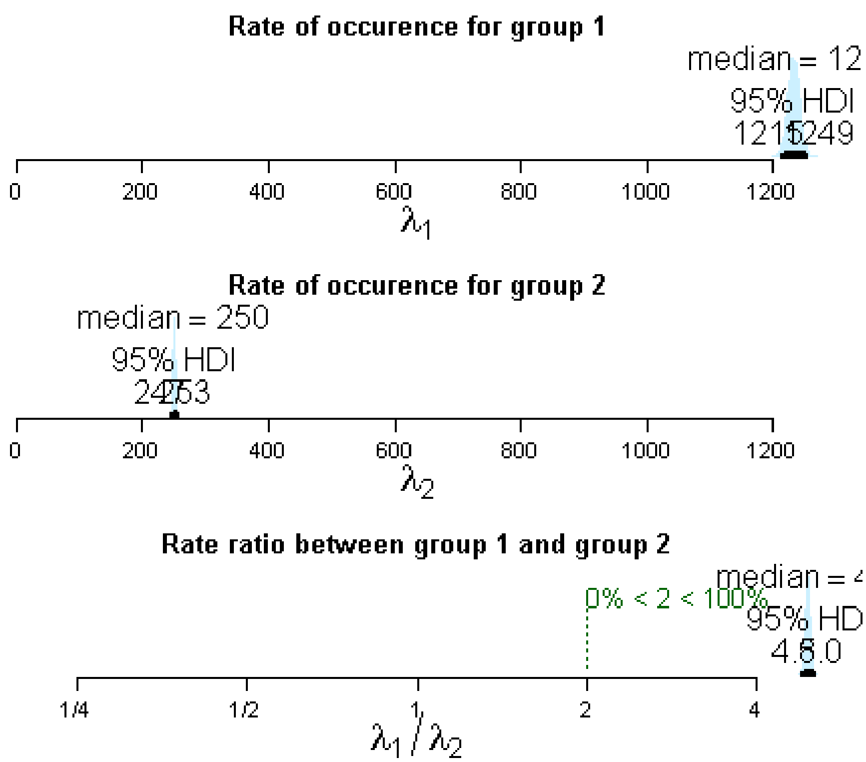


This indicates a strong rejection of the null hypothesis.

And also, just throw another test in of the 20th vs the 19th. **20th vs 19th**

```
poisson.test(c(sum(cent.20.freq),sum(cent.19.freq)),c(length(cent.20.freq),le
ngth(cent.19.freq)),r=1.5,alternative="greater")

##
##  Comparison of Poisson rates
##
## data:  c(sum(cent.20.freq), sum(cent.19.freq)) time base:
c(length(cent.20.freq), length(cent.19.freq))
## count1 = 24976, expected count1 = 15386.4, p-value < 2.2e-16
## alternative hypothesis: true rate ratio is greater than 1.5
## 95 percent confidence interval:
##  35.04272      Inf
## sample estimates:
## rate ratio
##   37.38922

bayes.poisson.test(c(sum(cent.21.freq),sum(cent.20.freq)),c(length(cent.21.fr
eq),length(cent.20.freq)),r=2)

##
##  Bayesian Fist Aid poisson test - two sample
##
## number of events: 19720 and 24976, time periods: 16 and 100
##
```

```
##   Estimates [95% credible interval]
## Group 1 rate: 1232 [1215, 1250]
## Group 2 rate: 250 [247, 253]
## Rate ratio (Group 1 rate / Group 2 rate):
##                4.9 [4.8, 5.0]
##
## The event rate of group 1 is more than 2 times that of group 2 by a
probability
## of >0.999 and less than 2 times that of group 2 by a probability of <0.001
.

plot(bayes.poisson.test(c(sum(cent.21.freq),sum(cent.20.freq)),c(length(cent.
21.freq),length(cent.20.freq)),r=2))
```



Again, a strong rejection of the null hypothesis.

Conclusions

the results of running Poisson tests between the frequencies of impacts between the 19th-21st centuries indicates that there is no evidence that the 20th century demonstrated a greater rate of incidence than has been shown in the 21st thus far. Indeed, there is a 99% probability that the rate of incidence in the 21st century is greater than the rate of incidence in the 20th century, and a 99% probability that the rate of incidence in the 20th century in the is greater than the rate of incidience in the 19th. More than anything, these results indicate a growing sophistication in the cataloguing of meteorite impacts.

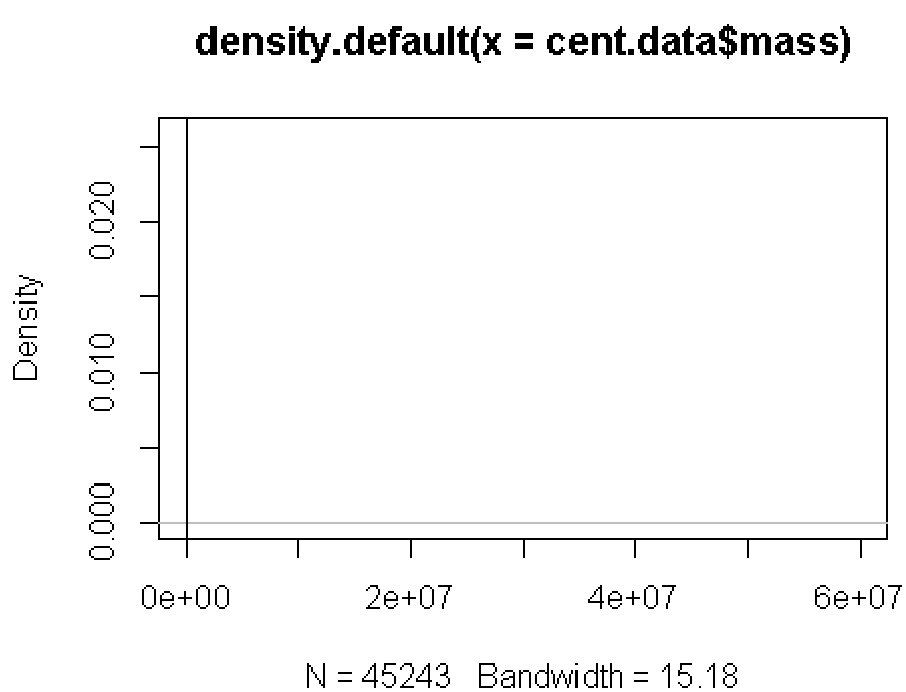
Investigating Mass of Impacts over Time:

(Q1.2) "Is there a correlation between time and mass of impacts?" Investigating whether or not there is a temporal trend between time and the mass of impacts. Have meteorite impacts been less massive as time has gone on, or more massive?

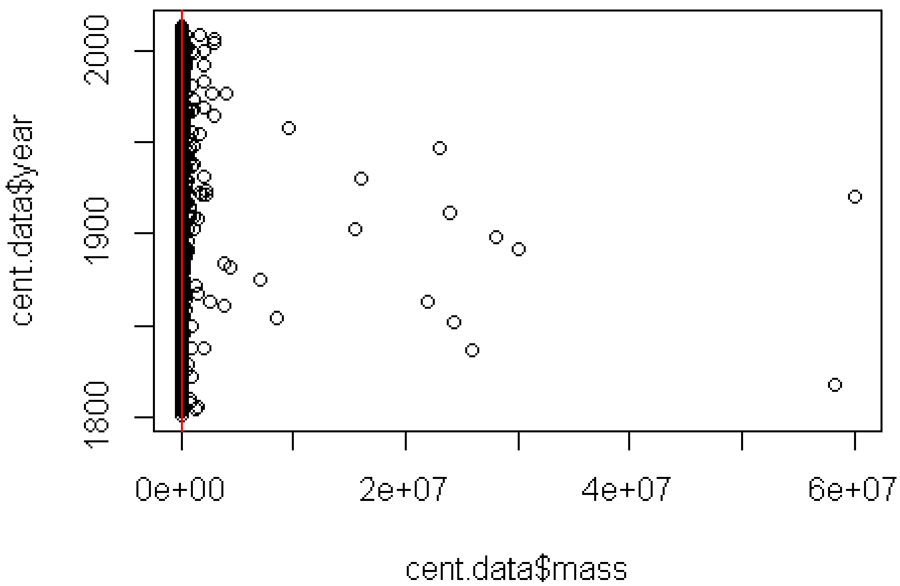
General Hypothesis: There is no correlation.

I need to make sure that I'm working with complete data

```
data.nona <- data_full[!is.na(data_full$mass),]
data.nona <- data.nona[data.nona$mass > 0,]
# I redo this from before, because we're not only looking for more frequent
impacts, but more massive impacts
cent.data <- data.nona[!is.na(data.nona$year),]
cent.data <- cent.data[cent.data$year > 1799 & cent.data$year < 2015,] #there
are some records of older meteorites, but we don't want

summary(cent.data$mass)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        0        7       32    11750      200 60000000

plot(density(cent.data$mass))
abline(v=mean(cent.data$mass))
```




```
plot(cent.data$year ~ cent.data$mass)
abline(v=mean(cent.data$mass),col="red")
```

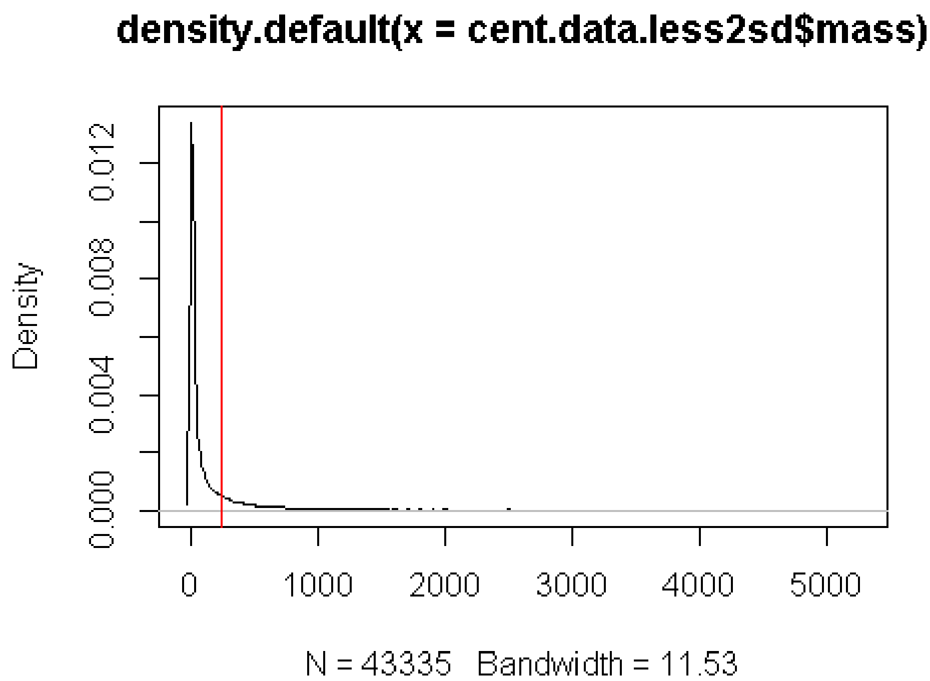


Looking at these plots, it seems clear that there isn't a simple correlation between the year of the impacts and their mass; however, I may be able to remove some of the really massive outliers and find something worthwhile. I remove data greater than one-hundreth of the standard deviation away from the mean, and attempt to fit a linear regression to the trend.

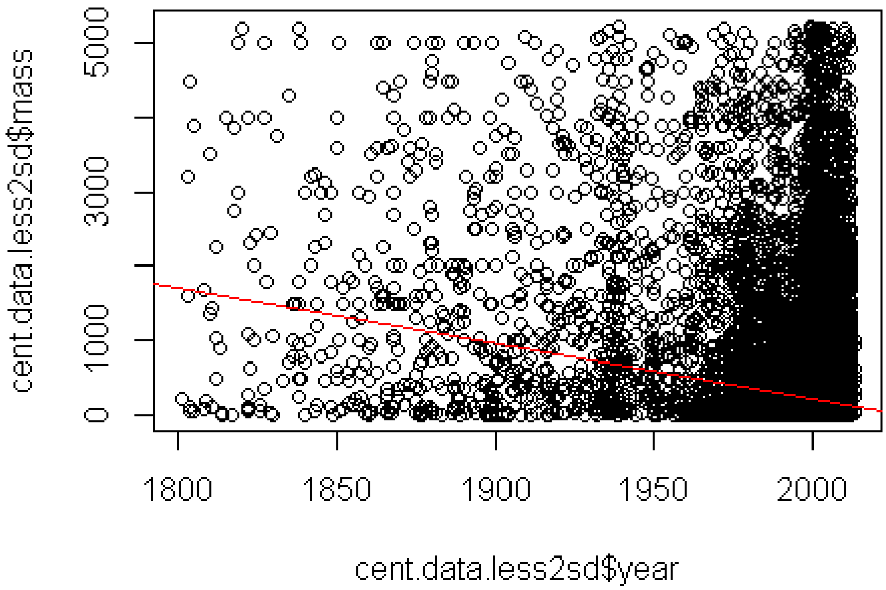
```
cent.data.less2sd <- cent.data[cent.data$mass < sd(cent.data$mass)/100,]

summary(cent.data.less2sd$mass)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.01    6.70   28.10  246.10  151.90 5230.00
```

```
plot(density(cent.data.less2sd$mass))
abline(v=mean(cent.data.less2sd$mass),col="red")
```

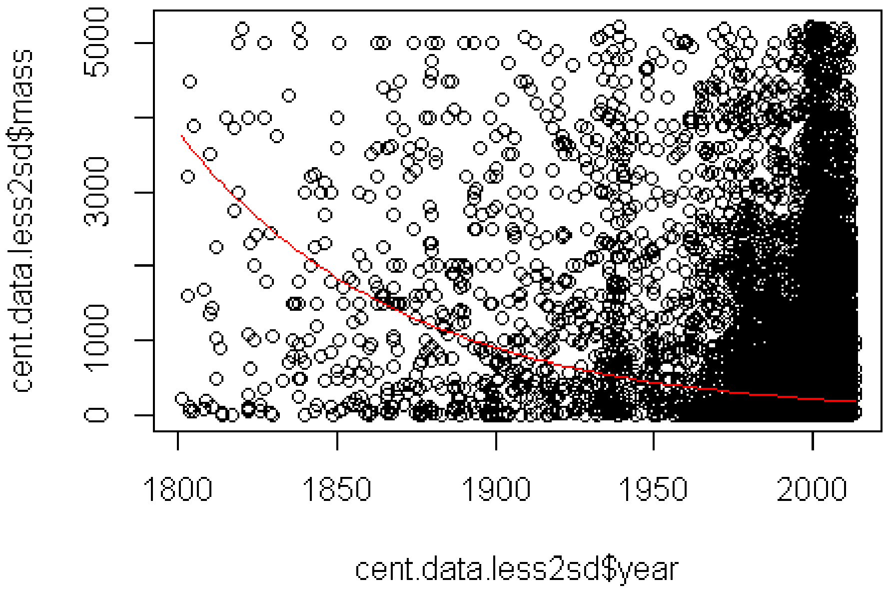



```
plot(cent.data.less2sd$mass ~ cent.data.less2sd$year)
fit <- lm(mass ~ year,data = cent.data.less2sd)
abline(fit,col="red")
```



Obviusly, the relationship is not linear. We try a general linear model.

```
# Quasi-Poisson
plot(cent.data.less2sd$mass ~ cent.data.less2sd$year)
fit <- glm(mass ~ year,data = cent.data.less2sd,family=quasipoisson)
curve(predict(fit,data.frame(year=x),type="resp"),add=TRUE,col="red")
```




```
cov.fit <- vcovHC(fit, type="HC0")
std.err <- sqrt(diag(cov.fit))
r.est <- cbind(Estimate= coef(fit), "Robust SE" = std.err,
               "Pr(>|z|)" = 2 * pnorm(abs(coef(fit)/std.err),
lower.tail=FALSE),
               LL = coef(fit) - 1.96 * std.err,
               UL = coef(fit) + 1.96 * std.err)
with(fit, cbind(res.deviance = deviance, df = df.residual,
               p = pchisq(deviance, df.residual, lower.tail=FALSE)))

##      res.deviance    df p
## [1,]     29925230 43333 0
```

Again, this curve can't really well-model the relationship, it seems. Indeed, it seems that though there is an increase in the frequency of more massive meteorites, this can be chalked up to the increase of frequency of impacts over time (as a function, likely, of better catalouging) and not so much to a direct relationship between mass and time.

Conclusion

It is not clear from the evidence that there is a relationship between the year of the impact and the mass. It is not likely that we have been experiencing more massive impacts as time has increased.

Projects for later

Unfortunately, I am busy with a huge workload, including a continuation of my work with the Mind Research Network. I am primarily a coder and applied mathematician - my experience with statistics is limited; however, I think the problems I have investigated here are at least interesting in their descriptive value, and I think the cleaning and extension I've done of the original dataset will make any future investigations easier. Given the opportunity, I would like to continue working in greater detail with this dataset, but for the time being, I list a number of questions which might be investigated at a future time:

Investigating Impacts over Locations:

(Q3.1) has any location experienced significantly more frequent impacts?

(Q3.2) has any location experienced significantly more massive impacts? ### Investigating Correlation between Mass and Frequency of Impact: (Q4.1) Is there a correlation between the mass and the frequency of impact? (e.g. if we decrease mass, do we increase frequency of impact)

Investigating questions of classification

(QC1) Can you predict whether a meteorite is brecciated or unbrecciated based on its mass? (QC2) Can we reliably classify meteorites by their mass?