# Exploratory Analysis of Health Index data

*A.J.Nicholson*

## Introduction

This documents records the exploratory analysis of the health index data. The purpose of this analysis is:

- Explore the shape and facets of the data,
- Produce exploratory plots
- Identify and highlight any issues in the data

This analysis follows from the end of "CleaningData.rmd".

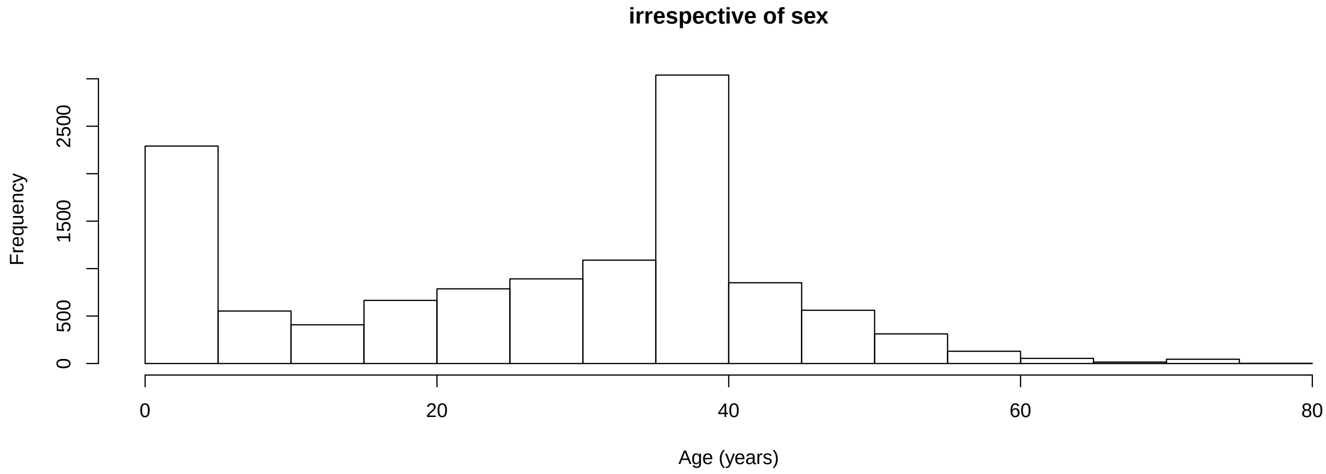
```
cols<-read.table("data/cols.txt",colClasses="character");cols<-as.character(cols[2,])
HIdata<-read.table("data/HIdata.txt",header=TRUE,colClasses=cols)
```

## Part 1 Demographic fields

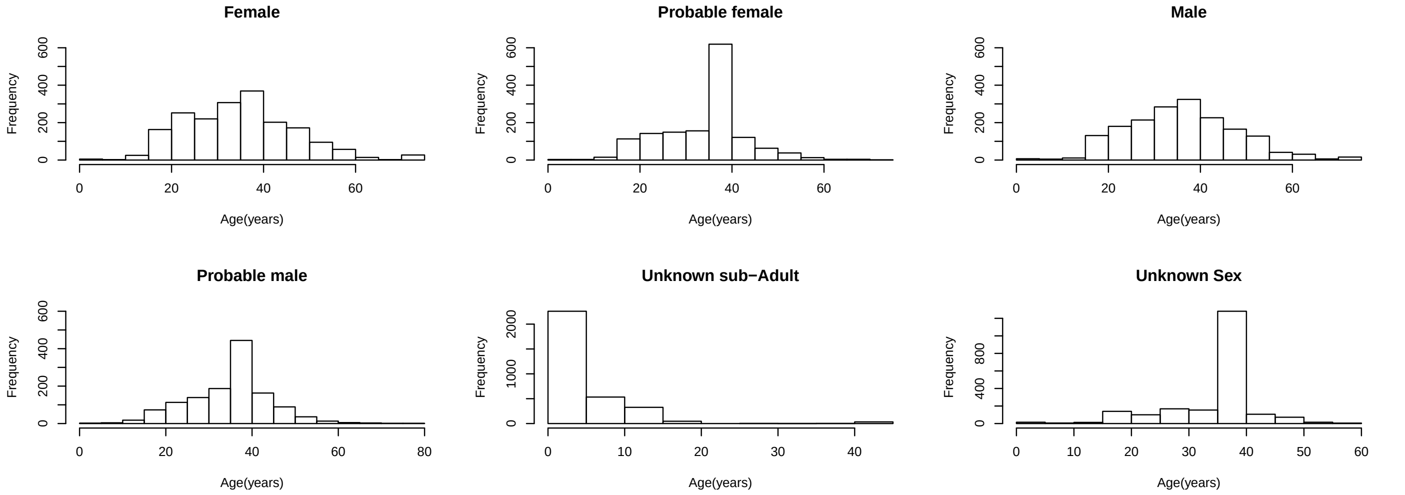
### Age Against Sex

A check for potential bias in the distrubution of age across the different age categories

```
hist(HIdata$Age, xlab="Age (years)",main="irrespective of sex")
```



```
par(mfrow=c(2,3))
hist(HIdata$Age[HIdata$Sex==1],main="Female",xlab="Age(years)",ylim=c(0,600))
hist(HIdata$Age[HIdata$Sex==2],main="Probable female",xlab="Age(years)",ylim=c(0,600))
hist(HIdata$Age[HIdata$Sex==3],main="Male",xlab="Age(years)",ylim=c(0,600))
hist(HIdata$Age[HIdata$Sex==4],main="Probable male",xlab="Age(years)",ylim=c(0,600))
hist(HIdata$Age[HIdata$Sex==5],main="Unknown sub-Adult",xlab="Age(years)")
hist(HIdata$Age[HIdata$Sex==6],main="Unknown Sex",xlab="Age(years)")
```



There is no obvious bias. The strong peak around 40 in the age distribution irrespective of sex seem to be mostly comprised of unknown and probable sex determinations.

These histogram highlight two potential issues in the data: Some juvenile remains not recorded as category 5 ("Sex is undetermined because the individual is less than 15 years of age and sex determination would be uncertain"). And some recorded as category 5 that are not juvenile.

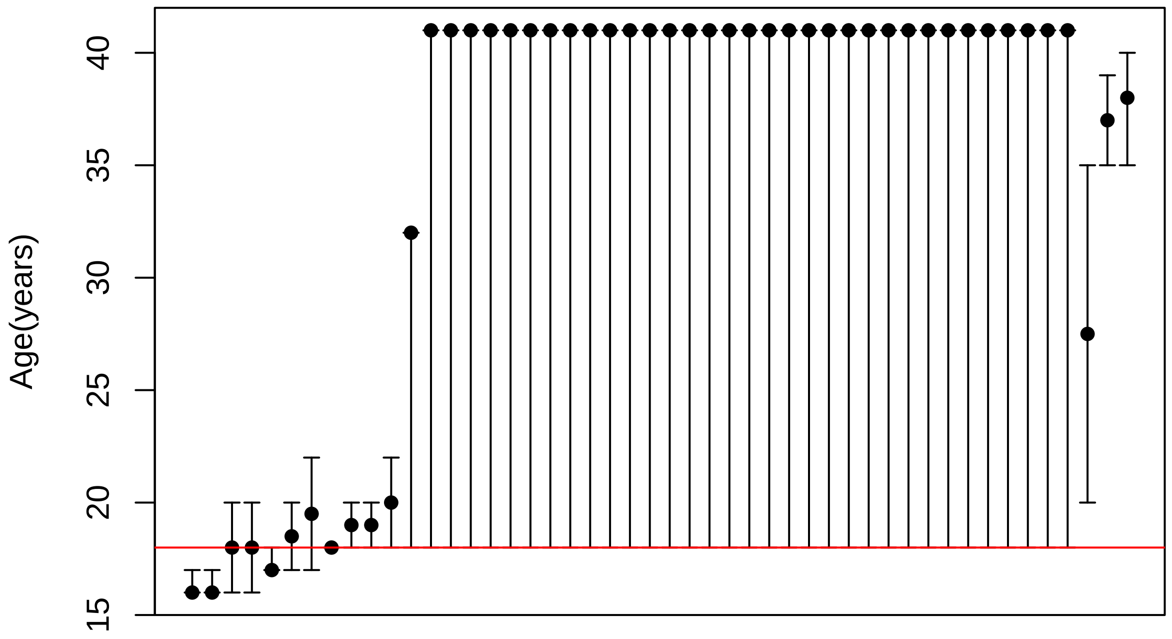
### Not Category 5

```
#select problematic data
Age<-HIdata[,c("Age","MinAge","MaxAge","cID","Sex")]
Age<-Age[Age$Sex=="5",];Age<-Age[!(is.na(Age$Age)),]
Issue<-Age[Age$Age>15,]

#Look at those with no MinAge
UnknownMin<-Issue[is.na(Issue$MinAge),]
UnknownMin$Age
```

```
##  [1] 16.0 15.5 16.0 16.0 16.0 16.0 17.0 16.0 30.0 16.0 16.0
```

With the exception of the record with an age of 30 these could all be reasonably considered juvenile (despite being over 15).

```
#Look at those with ranges where the minimum age is over 15
Issue<-Issue[!(Issue$cID %in% UnknownMin$cID),]
Issue<-Issue[Issue$MinAge>15,];I<-Issue

#create a plot
library(Hmisc)
for (i in 1:length(Issue[,1])){if(is.na(Issue$MaxAge[i])){Issue$MaxAge[i]<-Issue$Age[i]}}
Issue<-Issue[order(Issue$MinAge,Issue$MaxAge),];n<-length(Issue$cID)
errbar(x=1:n, y= Issue$Age, yplus=Issue$MaxAge, yminus=Issue$MinAge,xaxt='n',ylab="Age(years)",xlab="")
abline(h=18, col="red")# red line at 18 years old
```



The last three point on the graph have minimum ages of 35 and 20, these are very clearly not juveniles. Equally there are seven observation which given the age range could be under 18, and reasonably considered juvenile. The next four observation along would be categorized today as young adults (18-22). Although these should be old enough to sex given that many epiphysis closure age range include this range, the presence of open epiphyses may have lead the researcher to legitimately consider the remains to be non-adult.

The remaining observations plotted all lack a maximum age and have minimum ages of 18. The report ages for these observations (32 and 41) suggest clearly that they are not juvenile and that the minimum age of 18 may have come from the research deciding the remains are adult (i.e. all epiphyses are closed so they must be over 18).

To keep track of issues and errors in the data I'm going to create a dataframe called Issues that will store the IDs and a description of problem.

```
Issue<-I
Not5<-c("301_28001","TL2_208","BU1_5002","SA2_512B",Issue[is.na(Issue$MaxAge),"cID"])
Issues<-data.frame(ID=Not5, issue=rep("Not sex 5"))
```

### Should Be category 5

The Category 5 exist in this dataset because it is recognized that sex determinations on juvenile remains are not accurate. Therefore any individual that is juvenile (or at the very least under 15) should be recorded in this category. This is however not the case:

```
TooYoung<-HIdata[,c("cID","Sex","Age","MinAge","MaxAge")]
TooYoung<-TooYoung[TooYoung$Age<=15,]; TooYoung<-TooYoung[!(is.na(TooYoung$Age)),]
S<-summary(TooYoung$Sex)
S
```

```
##    1    2    3    4    5    6
##   32   21   23   24 3118   34
```

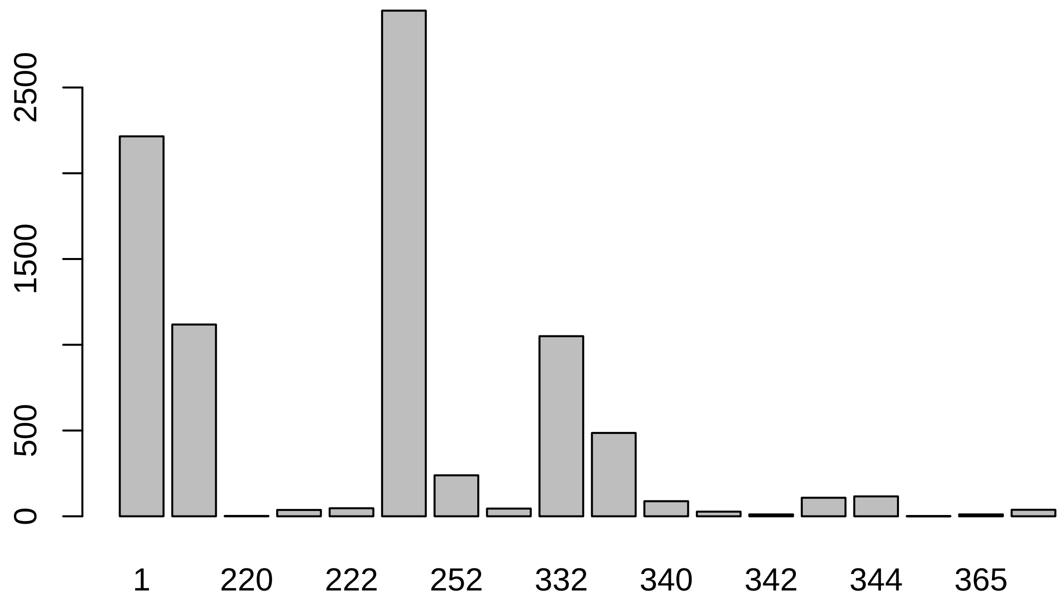
From this you can see that 100 of 3252 individuals under 15 have been sexed and 34 are in the other unknown category (unknown adult). These should all be added to the issues data frame.

```
TooYoung<-TooYoung[TooYoung$Sex!=5,]
TooYoung<-data.frame(ID=TooYoung$cID, issue=rep("too young to sex"))
Issues<-rbind(Issues,TooYoung)
```

### Social Status

The social status code is a three digit number which gives number of apparent divisions within the society and that individuals position within that social structure. See http://global.sbs.ohio-state.edu/docs/data.pdf for details.

```
plot(HIdata$SOC)
```



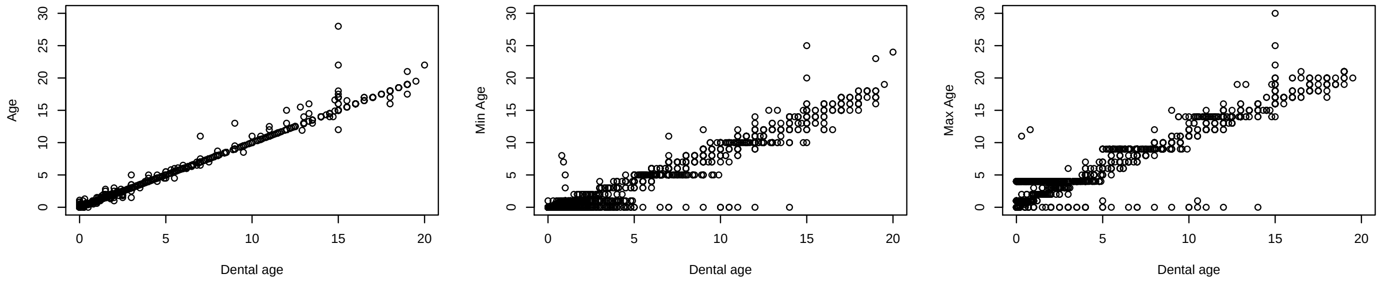
The code associated with the highest number of individuals is 232. This represent a ranked society (with groups or individuals having clear differential access to luxury or exotic goods) that have 3 social strata with the individual belonging the middle group.

This plot also highlights a potential issue. There are 6146 observations were the social status is recorded as 1. As previously mentioned this should be a three digit number.

```
ID<-HIdata[,c("cID","SOC")];ID<-ID[!(is.na(HIdata$SOC)),];ID<-ID[ID$SOC=="1",]
singledigit<-data.frame(ID=ID$cID,issue=rep("Single digit social code"))
Issues<-rbind(Issues,singledigit)
```

## Comparison of dental age and other Age fields

```
par(mfrow=c(1,3))
plot(HIdata$DentalAge, HIdata$Age,
         xlab="Dental age", ylab="Age",ylim=c(0,30))
plot(HIdata$DentalAge, HIdata$MinAge,
    xlab="Dental age", ylab="Min Age",ylim=c(0,30))
plot(HIdata$DentalAge, HIdata$MaxAge,
      xlab="Dental age", ylab="Max Age",ylim=c(0,30))
```



## Age and Age range

```
Age<-HIdata[,c("cID","Age","MaxAge","MinAge")]
Age$Range<-Age$MaxAge-Age$MinAge
Age<-Age[!(is.na(Age$Range)),]

summary(Age$Range)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -6.000   3.000   4.000   6.184   6.000  64.000
```

The minimum value for range is -6 this should not be possible and means that the maximum age must be lower than the minimum age.

```
LowMax<-Age[Age$Range<0,]
LowMax<-data.frame(ID=LowMax$cID, issue=rep("MaxAge lower than MinAge"))
Issues<-rbind(Issues,LowMax)
```

### Age within range?

```
Age$inrange<-Age$Age<=Age$MaxAge & Age$Age>=Age$MinAge
summary(Age$inrange)
```

```
##    Mode   FALSE    TRUE    NA's
## logical     346    8730     120
```

There are 466 instance where the Age is not between the range given by the minimum and maximum ranges.

```
OutRange<-Age[Age$inrange==FALSE,]; OutRange<-OutRange[!(is.na(OutRange$inrange)),]
```

There are also a number of NA in the inrange column this is where there is a range recorded but not age.

```
NoAge<-Age[is.na(Age$Age),"cID"]; NoAge<-data.frame(ID=NoAge, issue=rep("Max and Min but no Age"))
Issues<-rbind(Issues,NoAge)
```

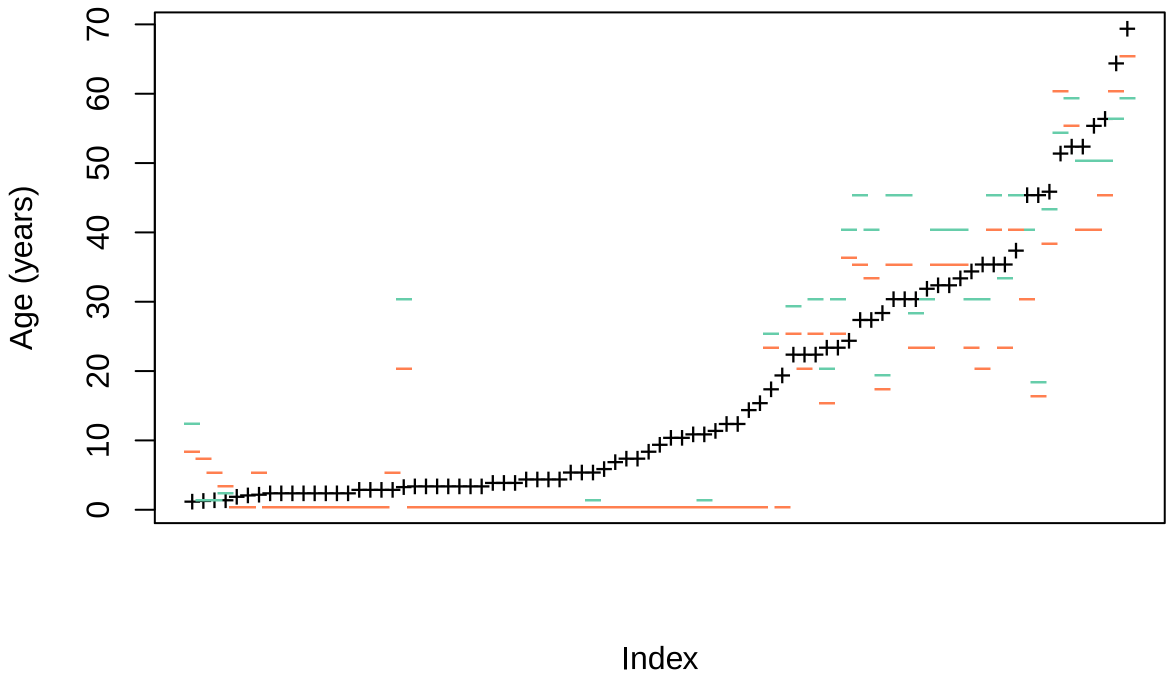
The age variable is recorded to one decimal place whereas the range values (MaxAge and MinAge) are recorded without decimal places. This could conceivably put the age within range.

```
OutRange$Extended<-OutRange$Age<=(OutRange$MaxAge +1) & OutRange$Age>=(OutRange$MinAge-1)
summary(OutRange$Extended)
```

```
##    Mode   FALSE    TRUE    NA's
## logical      85     261       0
```

This puts 261 out of the 346 within range. The remaining can be plotted to get a better idea of the source of the issue.

```
OR<-OutRange[OutRange$Extended==FALSE,]
OR<-OR[order(OR$Age),]
par(mfrow=c(1,1))
plot(OR$Age,ylab="Age (years)",xaxt="n",pch="+")
points(y=OR$MaxAge, x=1:85, pch="-",col="aquamarine3")
points(y=OR$MinAge, x=1:85, pch="-", col="coral")
```

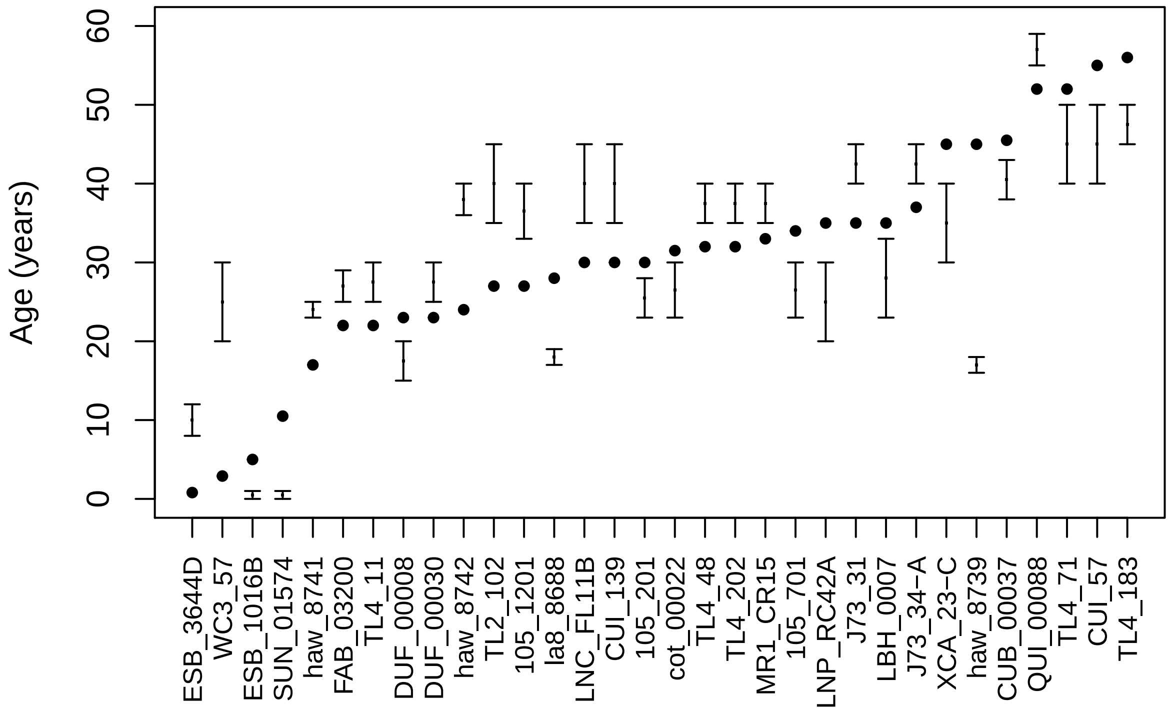


There are a large number where the here both the min and max are both 0. This could be a substitute for unknown info.

```
Zero<-OR;Zero<-Zero[Zero$MaxAge==0,]
Zero<-data.frame(ID=Zero$cID, issue=rep("Min and Max age should be NA"))
Issues<-rbind(Issues,Zero)

OR<-OR[OR$MaxAge!=0,]
OR<-OR[OR$Range>0,]
```

This leaves 32 problematic observations, plotted below:

```
plot(OR$Age, ylab="Age (years)",xlab=NA,xaxt="n",pch=20, ylim=c(0,60))
axis(side=1,at=1:32,labels=OR$cID, las=2,cex.axis=0.8)

OR$mid<-(OR$MaxAge+OR$MinAge)/2
errbar(x=1:32, y= OR$mid, yplus=OR$MaxAge, yminus=OR$MinAge, add=TRUE, pch=".")
```



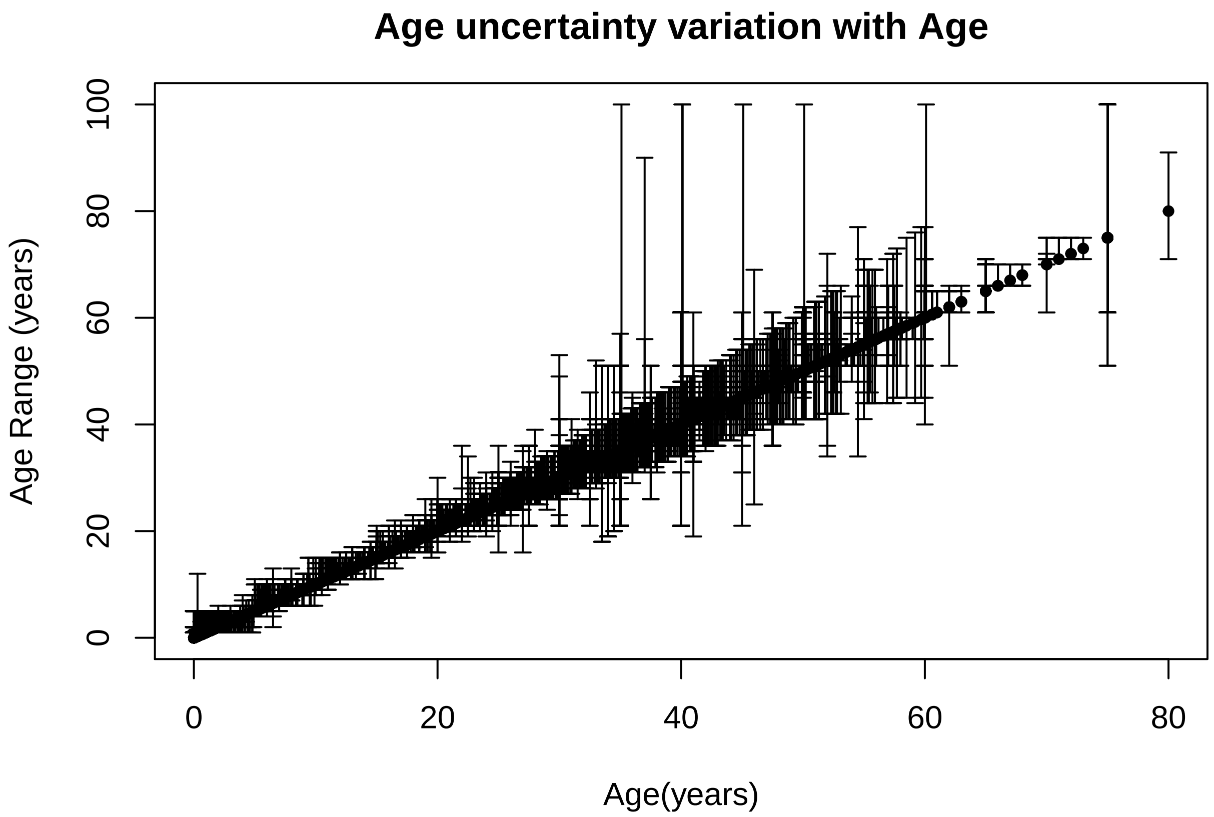
```
OR<-data.frame(ID=OR$cID, issue=rep("age outside age range"))
Issues<-rbind(Issues,OR)
```

### Age versus Age range

Taking the observations where there is no known issue with the age range, Age is plotted against the age range to see if the level of uncertainty in age is related to the age.

```
Age<-Age[!(is.na(Age$Age)),]; Age<-Age[Age$Range>=0,]
Age$Extended<-Age$Age<=(Age$MaxAge +1) & Age$Age>=(Age$MinAge-1)
Age<-Age[Age$Extended==TRUE,]
Age<-Age[order(Age$Age),]
par(mar=c(4.5,4.1,3,1))
```

```
errbar(x=Age$Age,y=Age$Age,yplus=Age$MaxAge+1,yminus=Age$MinAge+1,
       ylab="Age Range (years)",xlab="Age(years)",pch=20)
title(main="Age uncertainty variation with Age")
```



This plot shows a general increase in uncertainty with age, this mimic the results of previous research; and the know osteological problem of decreased accuracy of age estimation among older individuals.

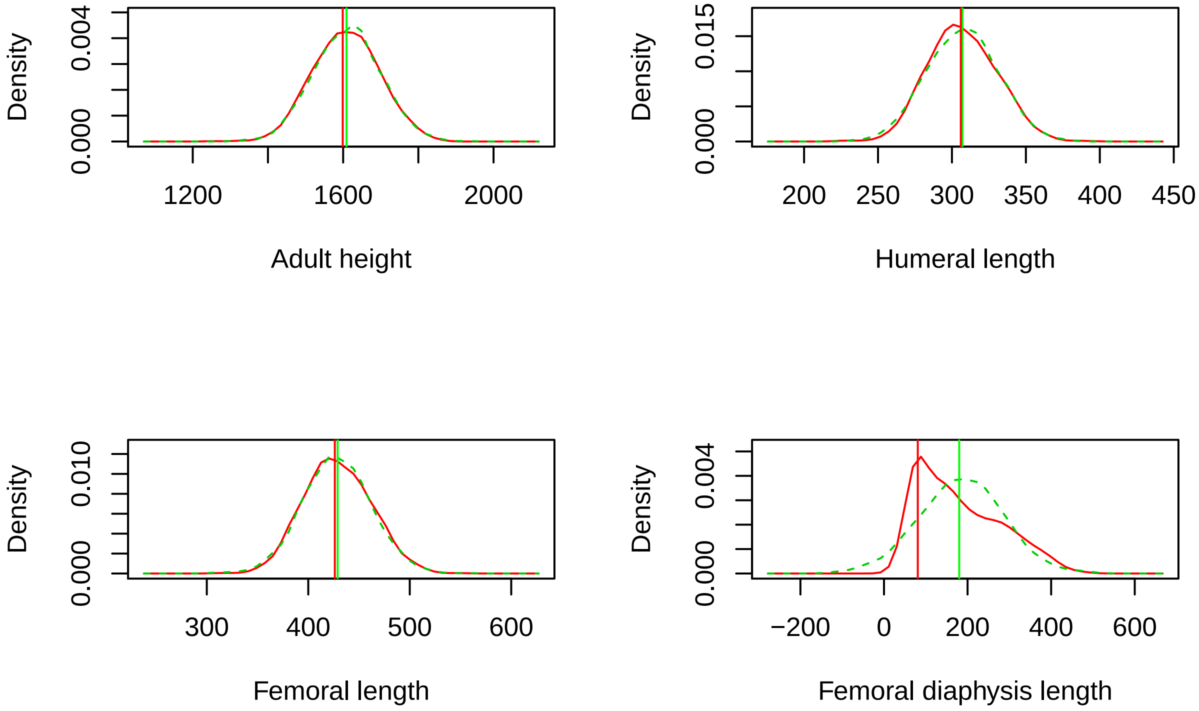
# Part 2 Continuous Data

## Normal distribution

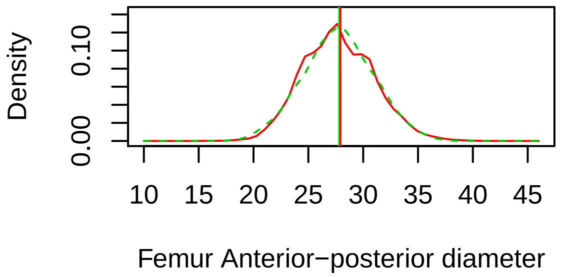
Each of the variables is plotted (in red) against a theoretical normal distribution (in green) using the following function:

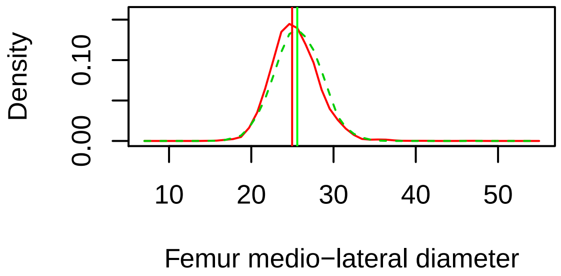
```
NormalDensityComp<-function(x,lab){
  x<-x[!(is.na(x))]
  density<-density(x)
  mean<-mean(x)
  norm<-rnorm(length(x),mean=mean(x),sd=sd(x))
  norm<-data.frame(value=norm,group="Theoretical Normal")
  x<-data.frame(value=x,group=lab)
  x<-rbind(x,norm)
  library(sm)
  sm.density.compare(x$value,x$group,xlab=lab)
  abline(v=mean,col="green")
  abline(v=density$x[which.max(density$y)],col="red")
}
```

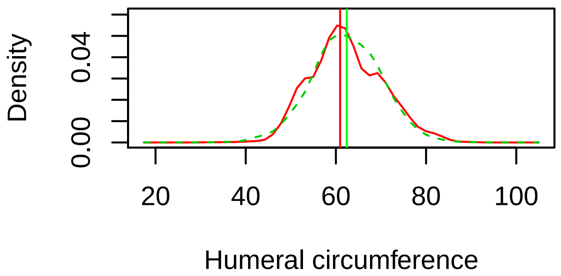
```
par(mfrow=c(2,2))
NormalDensityComp(HIdata$HEIGT,"Adult height")
NormalDensityComp(HIdata$HLEN,"Humeral length")
NormalDensityComp(HIdata$FLEN,"Femoral length")
NormalDensityComp(HIdata$FDIAP,"Femoral diaphysis length")
```



```
NormalDensityComp(HIdata$FMIDA,"Femur Anterior-posterior diameter")
NormalDensityComp(HIdata$FMIDM,"Femur medio-lateral diameter")
NormalDensityComp(HIdata$HCIR,"Humeral circumference")
```



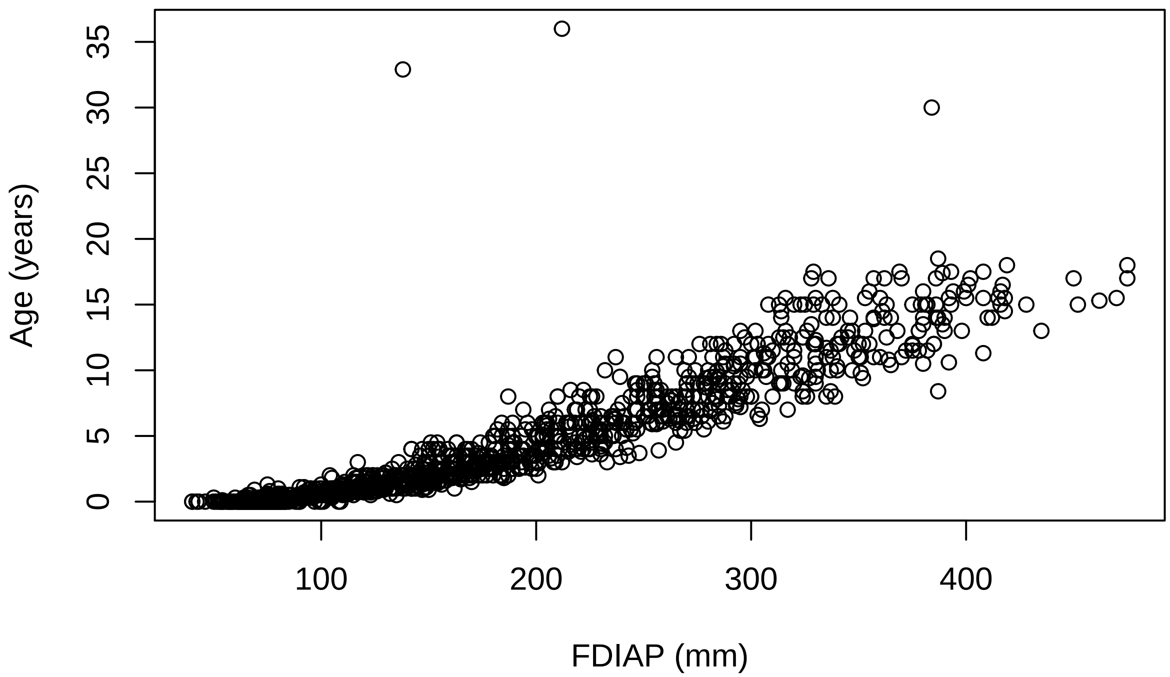




## FDIAP and Age

FDIAP (femoral diaphysis length)is a measure of growth for juvenile remains. It should show a relationship to age.

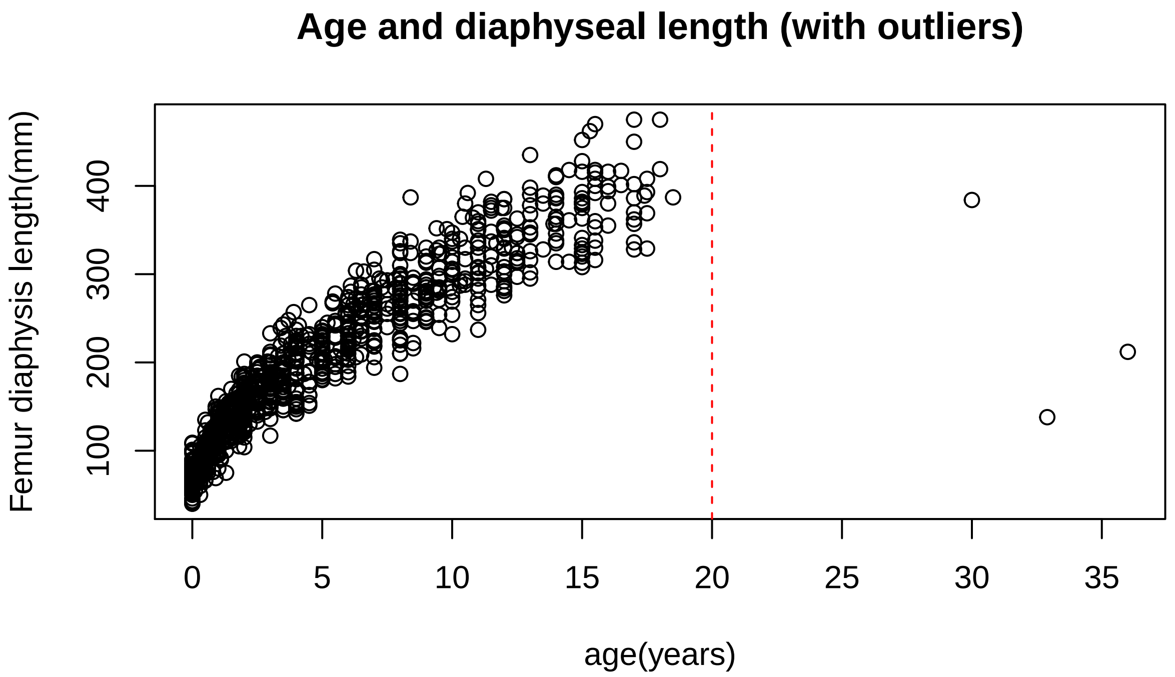
```
plot(HIdata$FDIAP[!(is.na(HIdata$FDIAP))],HIdata$Age[!(is.na(HIdata$FDIAP))],
     xlab="FDIAP (mm)",ylab="Age (years)")
```

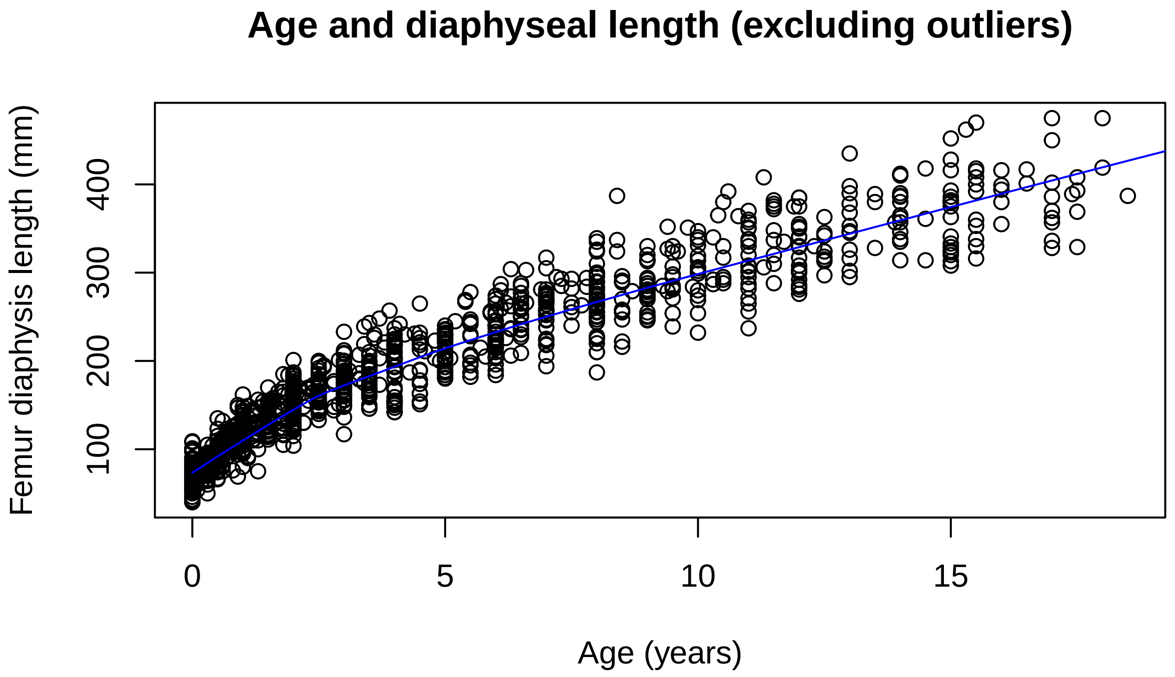


There seems to be an roughly exponential relationship to age. As well as an increase in variability with age.

There also appear to be a few observation where the age is above a level which would be considered juvenile. Femur diaphyseal length can only be accurately measure before the epiphysis is completely fused.

```r
Growth<-HIdata[,c("Age","FDIAP","cID","MinAge")]
Growth<-Growth[!(is.na(Growth$FDIAP)),]
Growth<-Growth[!(is.na(Growth$Age)),]

par(mfrow=c(2,1))
plot(Growth$Age,Growth$FDIAP,
     xlab="age(years)", ylab="Femur diaphysis length(mm)",main="Age and diaphyseal length (with outliers)")
abline(v=20, lty=2, col=34)

Issue<-Growth[Growth$Age>20,]
Exclude<-Issue[,"cID"]

plot(Growth$Age[!(Growth$cID %in% Exclude)],Growth$FDIAP[!(Growth$cID %in% Exclude)],
     xlab="Age (years)",ylab="Femur diaphysis length (mm)",main="Age and diaphyseal length (excluding outlier
lines(lowess(Growth$Age,Growth$FDIAP), col="blue")
```


Age and diaphyseal length (with outliers)

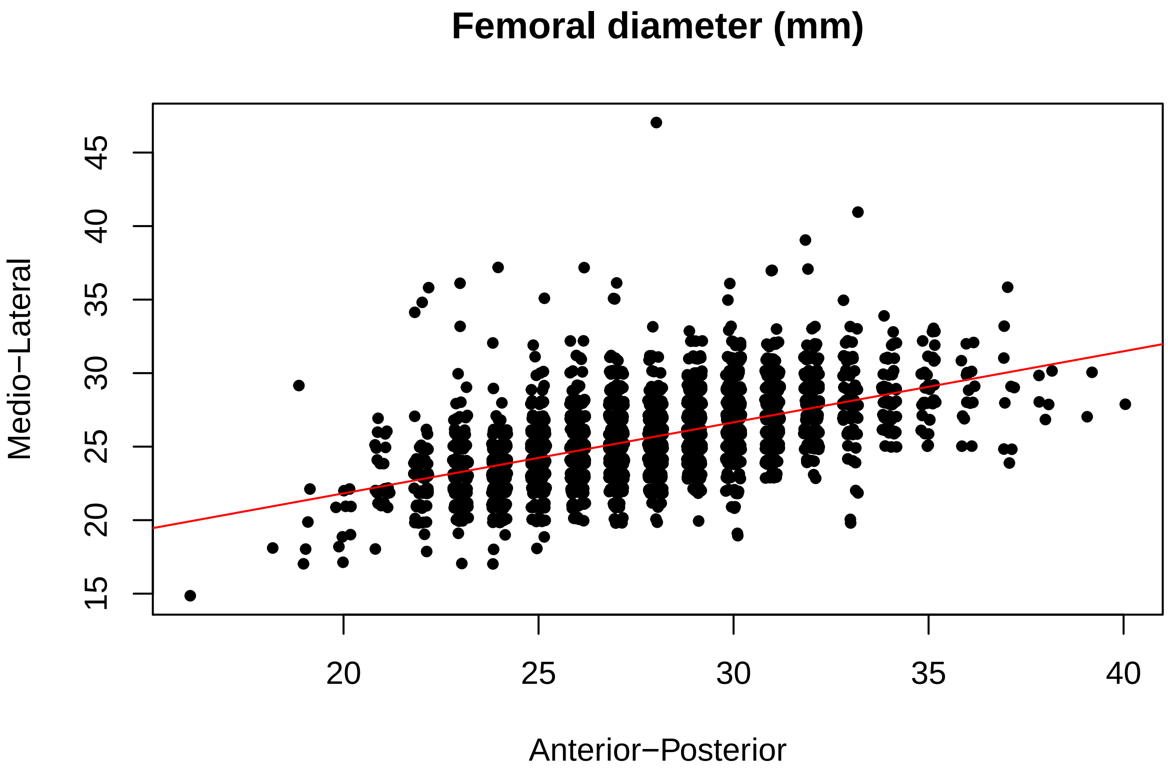

Age and diaphyseal length (excluding outliers)

```
Exclude<-data.frame(ID=Exclude, issue=rep("Non Juvenile femoral diaphysis", length(Exclude)))
Issues<-rbind(Issues,Exclude)
```

### Femoral diameters

There are two recorded measures of femoral diameter at mid-shaft, one anterior-posterior (FMIDA) and one medio-lateral(FMIDM). Common sense would suggest there should be a correlation between the two.

```r
plot(jitter(HIdata$FMIDA),jitter(HIdata$FMIDM),
     xlab="Anterior-Posterior",ylab="Medio-Lateral",main="Femoral diameter (mm)"
     ,pch=20)
abline(lm(HIdata$FMIDM~HIdata$FMIDA), col=10)# simple linear model
```



There does seem to be some positive trend but there is a lot of spread around the line.

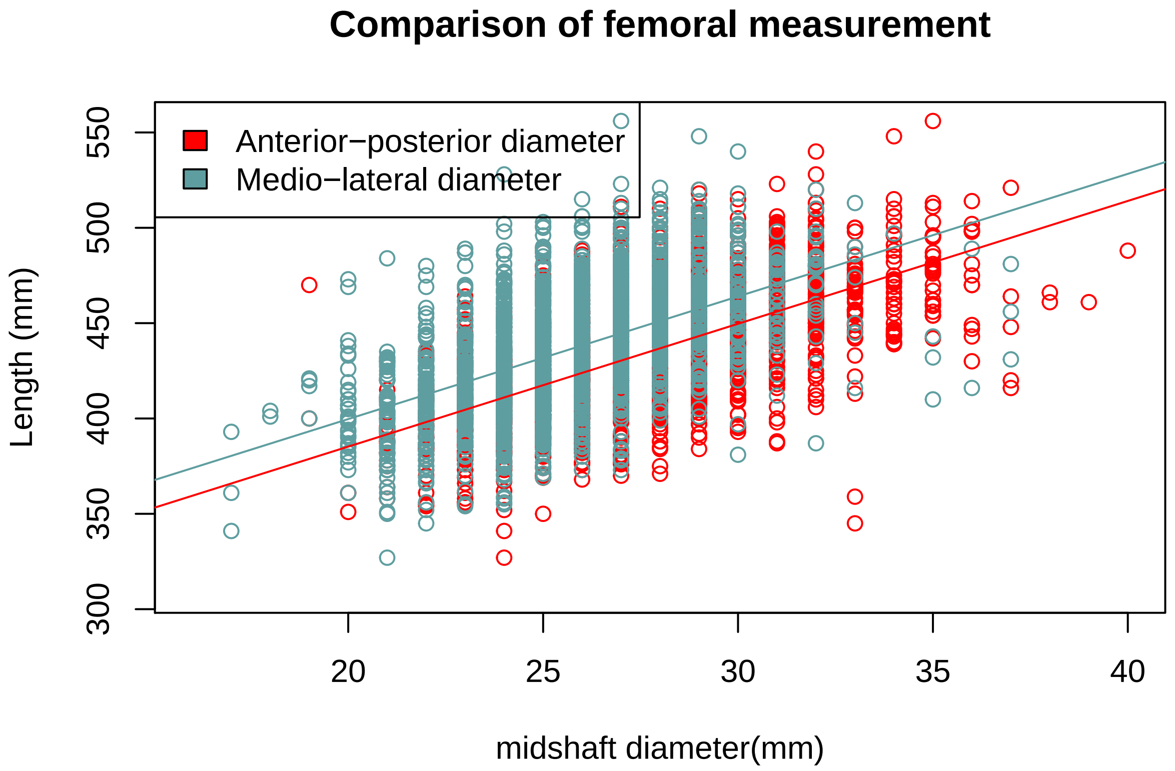
```r
summary(lm(HIdata$FMIDM~HIdata$FMIDA))
```

```
##
## Call:
## lm(formula = HIdata$FMIDM ~ HIdata$FMIDA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1027 -1.6533 -0.2039  1.3130 21.3130
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   12.1589     0.3836   31.70   <2e-16 ***
## HIdata$FMIDA   0.4832     0.0137   35.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.378 on 2952 degrees of freedom
##   (9564 observations deleted due to missingness)
## Multiple R-squared:  0.2963, Adjusted R-squared:  0.2961
## F-statistic:  1243 on 1 and 2952 DF,  p-value: < 2.2e-16
```

The p values suggest a significant relationship but the low adjusted r^2 attest to the poor fit (large spread) and hence poor predictive ability of this relationship.

### Femur Diameter vs. Length

```r
plot(HIdata$FMIDA,HIdata$FLEN,
     xlab="midshaft diameter(mm)",ylab="Length (mm)",main="Comparison of femoral measurement",col="red")
points(HIdata$FMIDM, HIdata$FLEN, col="cadetblue")
abline(lm(HIdata$FLEN~HIdata$FMIDA),col="red")
abline(lm(HIdata$FLEN~HIdata$FMIDM),col="cadetblue")
legend("topleft",c("Anterior-posterior diameter","Medio-lateral diameter"),
       fill=c("red","cadetblue"))
```



There is a clear overlap between two clusters. And there doesn't seem to be a good a fit for either line.

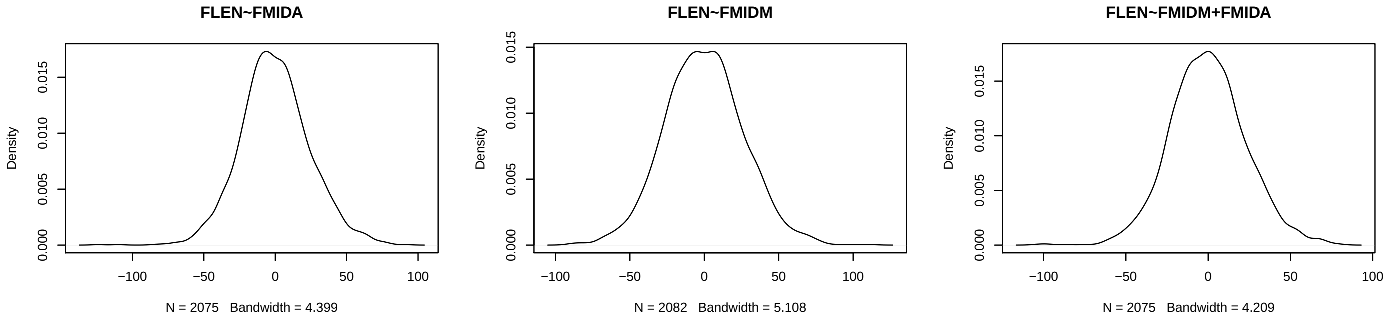
```r
x<-data.frame(model=c("Anterior-posterior","Medio-lateral","Combined"),
              interceptP=NA,slopeP=NA, AdjustedR2=NA)

#Model 1: Femur length against Anterior-Posterior diameter
A<-summary(lm(HIdata$FLEN~HIdata$FMIDA))
x[1,c("interceptP","slopeP")]<-signif(A[4][[1]][,4],3)
x[1,"AdjustedR2"]<-signif(A[9][[1]],3)

#Model 2: Femur length against Medio-Lateral diamter
M<-summary(lm(HIdata$FLEN~HIdata$FMIDM))
x[2,c("interceptP","slopeP")]<-signif(M[4][[1]][,4],3)
x[2,"AdjustedR2"]<-signif(M[9][[1]],3)

#Model 3: Femur length against the combination of both Diameters
C<-summary(lm(HIdata$FLEN~HIdata$FMIDM+HIdata$FMIDA))
x[3,c("interceptP")]<-signif(C[4][[1]][1,4],3)
x[3,c("slopeP")]<-paste("FMIDM:",signif(C[4][[1]][2,4],3)," FMIDA:",signif(C[4][[1]][2,4],3))
x[3,"AdjustedR2"]<-signif(C[9][[1]],3)
```

```r
x$interceptP<-as.character(x$interceptP)
kable(x,row.names = NA)
```

| model | interceptP | slopeP | AdjustedR2 |
|---|---|---|---:|
| Anterior-posterior | 0 | 5.51e-237 | 0.406 |
| Medio-lateral | 0 | 2.39e-160 | 0.295 |
| Combined | 1.98e-255 | FMIDM: 1.33e-47 FMIDA: 1.33e-47 | 0.463 |

Each of these models is significant but have low Adjusted R squared. The slightly higher Adjust R squared for the combined model doesn't necessarily equate to a better fix as adding components to a model always inflates the R-squared value.

```r
ARes<-lm(HIdata$FLEN~HIdata$FMIDA)$residuals
MRes<-lm(HIdata$FLEN~HIdata$FMIDM)$residuals
CRes<-lm(HIdata$FLEN~HIdata$FMIDM+HIdata$FMIDA)$residuals

par(mfrow=c(1,3))
plot(density(ARes), main="FLEN~FMIDA")
plot(density(MRes), main="FLEN~FMIDM")
plot(density(CRes), main="FLEN~FMIDM+FMIDA")
```

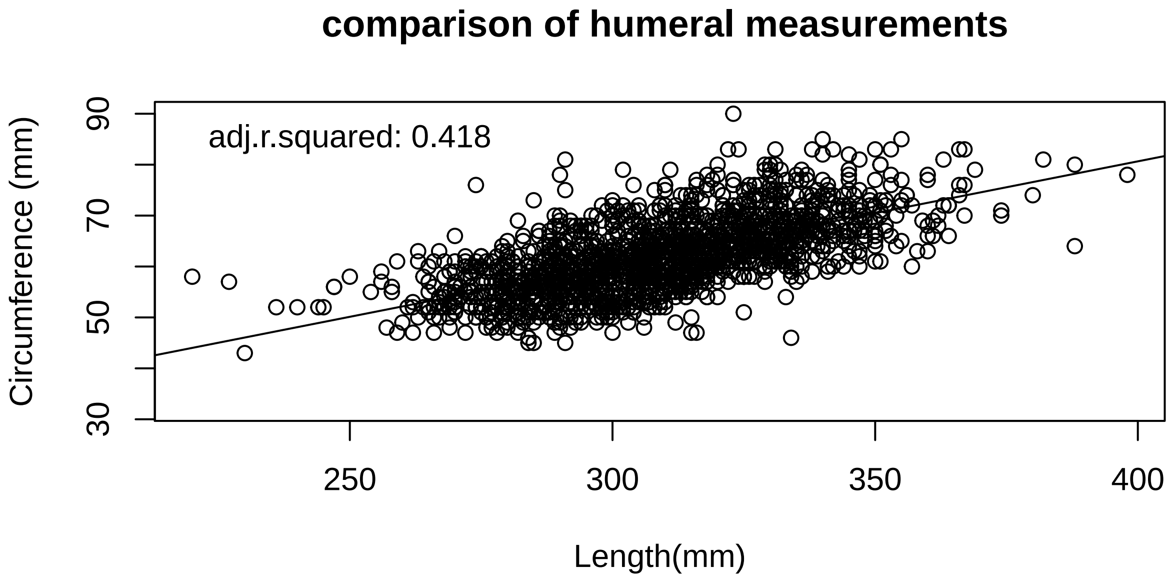


The residuals are all roughly normally distributed around zero.

## Humeral Measurements

There are two measurments of the humerus recorded, mid-shaft circumference(HCIR) and humeral length(HLEN).

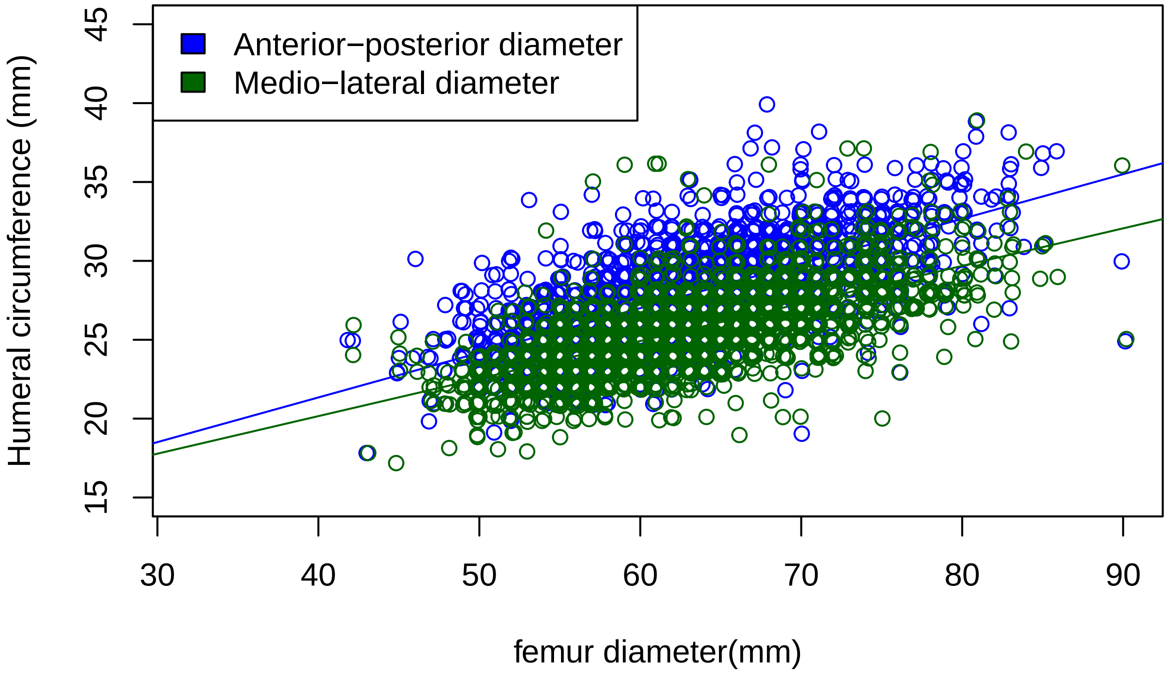
```r
plot(HIdata$HLEN,HIdata$HCIR,
     xlab="Length(mm)",ylab="Circumference (mm)", main=" comparison of humeral measurements")
abline(lm(HIdata$HCIR~HIdata$HLEN))
ARS<-summary(lm(HIdata$HCIR~HIdata$HLEN))$adj.r.squared
text(250,85,paste("adj.r.squared:",signif(ARS,3)))
```



### Comparison of Robusticity measures

Humeral circumference and femur diameter measures are all measures of robusticity so should theoretically correlate.

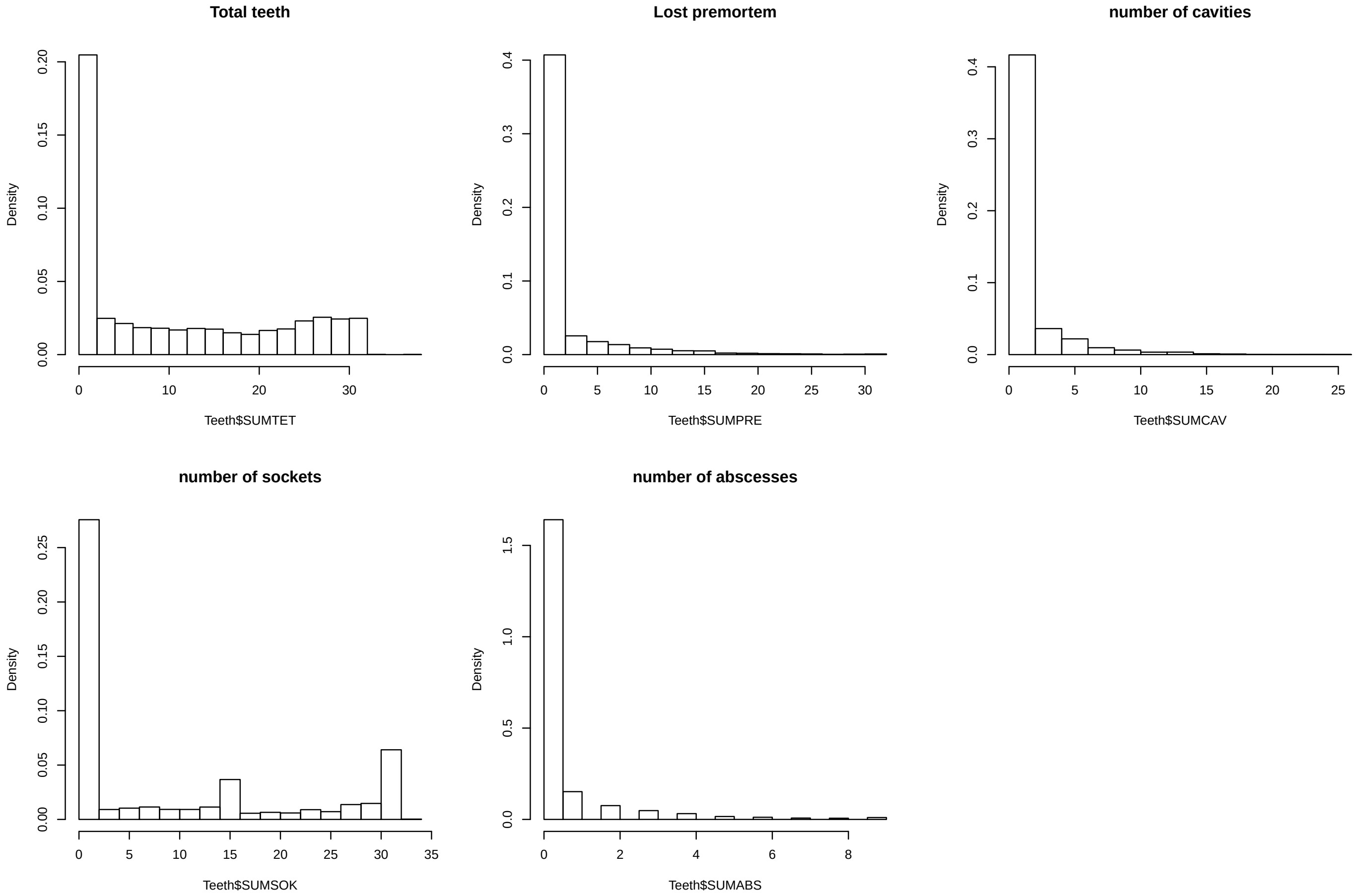
```
plot(jitter(HIdata$HCIR),jitter(HIdata$FMIDA),col="blue",
     xlab="femur diameter(mm)",ylab="Humeral circumference (mm)",ylim=c(15,45))
abline(lm(HIdata$FMIDA~HIdata$HCIR),col="blue")
points(jitter(HIdata$HCIR),jitter(HIdata$FMIDM),col="dark green")
abline(lm(HIdata$FMIDM~HIdata$HCIR),col="dark green")
legend("topleft",c("Anterior-posterior diameter","Medio-lateral diameter"),
       fill=c("blue","dark green"))
```



## Part 3: Count Data

All of the dental health measurements are represented as count data.

```
Teeth<-HIdata[,c("cID","SUMTET","SUMPRE","SUMCAV","SUMSOK","SUMABS","Age")]
par(mfrow=c(2,3))
hist(Teeth$SUMTET,main="Total teeth",prob=TRUE)
hist(Teeth$SUMPRE,main="Lost premortem",prob=TRUE)
hist(Teeth$SUMCAV,main="number of cavities",prob=TRUE)
hist(Teeth$SUMSOK,main="number of sockets",prob=TRUE)
hist(Teeth$SUMABS, main="number of abscesses",prob=TRUE)
```



All of these have a high zero counts.

```
kable(summary(Teeth[2:7]))
```

| SUMTET | SUMPRE | SUMCAV | SUMSOK | SUMABS | Age |
|---|---|---|---|---|---|
| Min. : 0.00 | Min. : 0.000 | Min. : 0.000 | Min. : 0.000 | Min. :0.0000 | Min. : 0.00 |
| 1st Qu.: 0.00 | 1st Qu.: 0.000 | 1st Qu.: 0.000 | 1st Qu.: 0.000 | 1st Qu.:0.0000 | 1st Qu.:11.20 |
| Median : 6.00 | Median : 0.000 | Median : 0.000 | Median : 0.000 | Median :0.0000 | Median :32.00 |
| Mean :10.71 | Mean : 1.766 | Mean : 1.238 | Mean : 9.468 | Mean :0.4611 | Mean :26.52 |
| 3rd Qu.:21.00 | 3rd Qu.: 1.000 | 3rd Qu.: 1.000 | 3rd Qu.:17.000 | 3rd Qu.:0.0000 | 3rd Qu.:36.00 |
| Max. :38.00 | Max. :32.000 | Max. :25.000 | Max. :33.000 | Max. :9.0000 | Max. :80.00 |
| NA's :1916 | NA's :2635 | NA's :2338 | NA's :2152 | NA's :2559 | NA's :826 |

There are a large number of missing values (NA) in each field. There is a potential issue in that the maximum values for SUMTET (total teeth present) and SUMSOK (number of sockets observed) are both above 32.

## Too Many teeth

```
TooMany<-Teeth[Teeth$SUMTET>32,]
TooMany<-TooMany[!(is.na(TooMany$SUMTET)),]
kable(TooMany)
```

| | cID | SUMTET | SUMPRE | SUMCAV | SUMSOK | SUMABS | Age |
|---|---|---|---|---|---|---|---|
| 228 | 3La_00081 | 38 | 0 | 2 | 0 | 0 | 22.5 |

|  | cID | SUMTET | SUMPRE | SUMCAV | SUMSOK | SUMABS | Age |
|---|---|---|---|---|---|---|---|
| 847 | 41D_0862 | 34 | 0 | 2 | 2 | 0 | 33.3 |
| 1414 | 41D_1512 | 33 | 0 | 1 | NA | NA | 20.4 |
| 4684 | TL2_65 | 33 | 0 | 1 | 32 | 0 | 33.0 |
| 9946 | haw_8739 | 34 | 0 | 0 | 32 | 0 | 45.0 |

None of the individuals are juvenile so this isn't accidental recording of deciduous teeth. Hyperdontia is a possibility but it is more likely that these are recording errors.

```
TooMany<-data.frame(ID=TooMany$cID, issue=rep("Too many teeth"))
Issues<-rbind(Issues,TooMany)
```

## No teeth

Checking to see if there are any interesting pattern or issues among the records where SUMTET is either zero or NA.

```
NoTeeth<-Teeth[Teeth$SUMTET==0,]
kable(summary(NoTeeth[3:7]))
```

| SUMPRE | SUMCAV | SUMSOK | SUMABS | Age |
|---|---|---|---|---|
| Min. : 0.0000 | Min. :0.0000 | Min. : 0.0000 | Min. :0.0000 | Min. : 0.00 |
| 1st Qu.: 0.0000 | 1st Qu.:0.0000 | 1st Qu.: 0.0000 | 1st Qu.:0.0000 | 1st Qu.: 6.00 |
| Median : 0.0000 | Median :0.0000 | Median : 0.0000 | Median :0.0000 | Median :36.00 |
| Mean : 0.6217 | Mean :0.0259 | Mean : 0.9162 | Mean :0.0602 | Mean :25.94 |
| 3rd Qu.: 0.0000 | 3rd Qu.:0.0000 | 3rd Qu.: 0.0000 | 3rd Qu.:0.0000 | 3rd Qu.:36.00 |
| Max. :32.0000 | Max. :9.0000 | Max. :32.0000 | Max. :9.0000 | Max. :80.00 |
| NA's :2277 | NA's :2340 | NA's :1981 | NA's :2297 | NA's :2175 |

The SUMCAV summary presents a potential problem. The max value isn't zero, and you can't have cavities if you don't have teeth.

```
NoTeeth<-NoTeeth[!(is.na(NoTeeth$SUMCAV)),]
NoTeeth<-NoTeeth[NoTeeth$SUMCAV>0,]
```

There are 36 records with no teeth but cavities. There are two possible explanation for this:

1. The teeth were present but not recorded.
2. There were actually no cavities, it's a typing error.

There is no way to know for sure which of these is true for each case. However SUMTET could be approximated by take 32-the teeth lost antemortem(SUMPRE). Any records that still have no teeth recorded will be assumed to be cases of erroneous cavities.

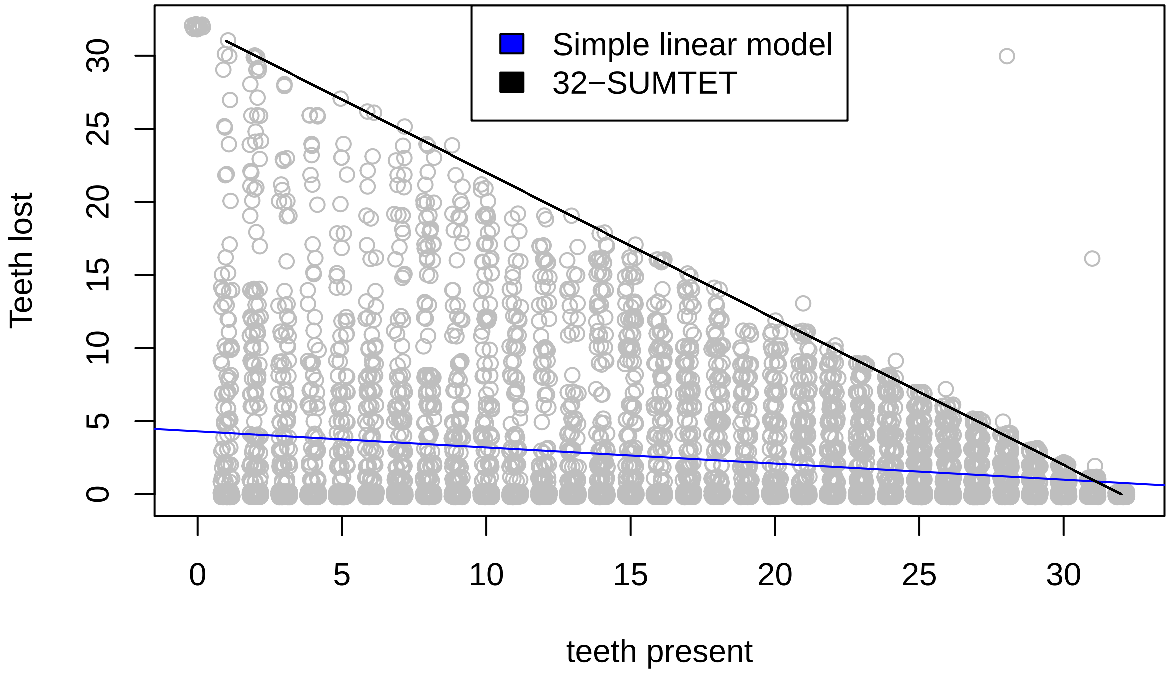
```
NoTeeth$Present<-32-NoTeeth$SUMPRE
kable(NoTeeth[NoTeeth$Present<=0,])
```

|  | cID | SUMTET | SUMPRE | SUMCAV | SUMSOK | SUMABS | Age | Present |
|---|---|---|---|---|---|---|---|---|
| 2162 | HPK_00071 | 0 | 32 | 1 | 32 | 0 | 65 | 0 |

```
NoCavities<-data.frame(ID=NoTeeth$cID[NoTeeth$Present<=0],issue="Erroneous Cavities")
Issues<-rbind(Issues,NoCavities)

NoTeeth<-NoTeeth[!(NoTeeth$Present<=0),]
NoTeeth<-data.frame(ID=NoTeeth$cID,issue="Cavities but no teeth")
Issues<-rbind(Issues,NoTeeth)
```

### Teeth present vs. lost pre-mortem

```
#remove unrecorded fields
Known<-Teeth
Known$SUMTET[is.na(Known$SUMTET)]<-0;Known<-Known[(Known$SUMTET>0),]
AllLost<-Teeth;AllLost<-AllLost[!(is.na(AllLost$SUMPRE)),]
AllLost<-AllLost[(AllLost$SUMPRE==32),]
Known<-rbind(Known,AllLost);Known<-Known[Known$SUMTET<=32,]

plot(jitter(Known$SUMTET),jitter(Known$SUMPRE),xlab="teeth present",ylab="Teeth lost",col="grey")
abline(lm(Known$SUMPRE~Known$SUMTET),col="blue")
lines(Known$SUMTET,32-Known$SUMTET)
legend("top",c("Simple linear model","32-SUMTET"),
       fill=c("blue","black"))
```



The black line represent the theoretical maximum possible value of ante-mortem tooth loss given the number of teeth present. Therefore all points should be under this line, is isn't the case.

```
Known$SUMPRE[is.na(Known$SUMPRE)]<-0
Known$total<-Known$SUMTET+Known$SUMPRE
summary(Known$total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   10.00   21.00   18.93   28.00   58.00
```

```
TooMany<-Known[!(is.na(Known$total)),]
TooMany<-TooMany[TooMany$total>32,]
kable(TooMany)
```

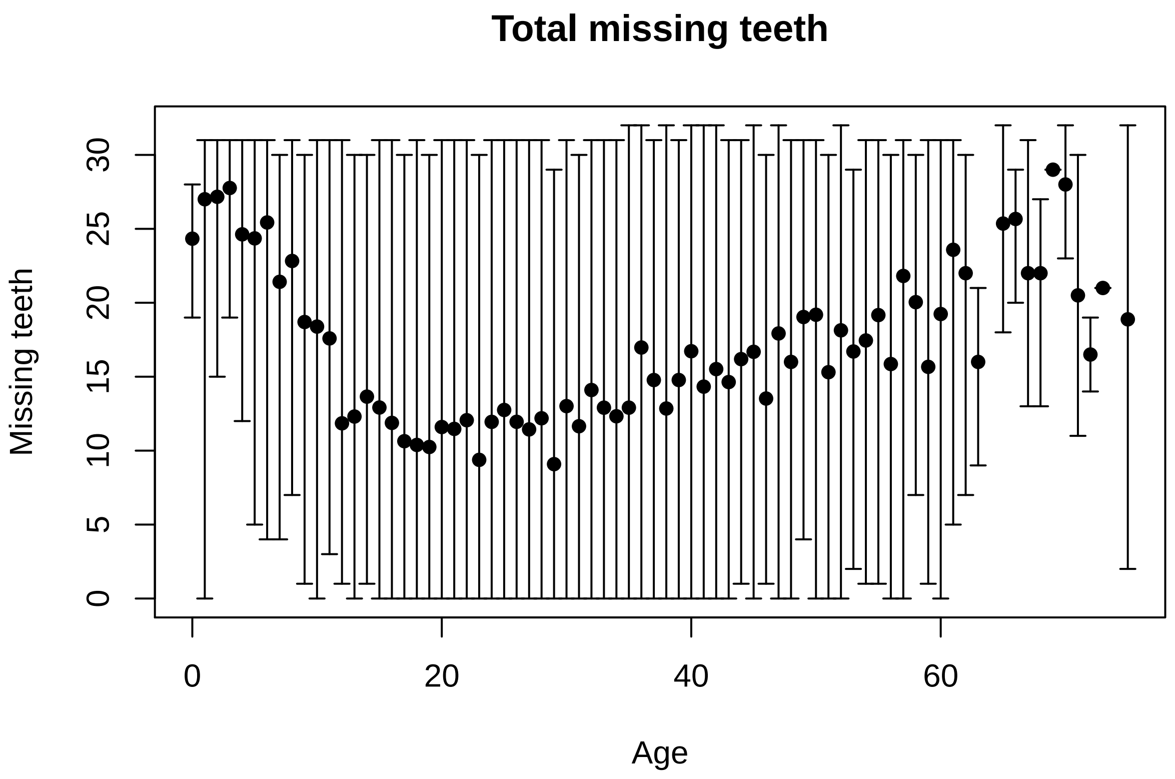
| | cID | SUMTET | SUMPRE | SUMCAV | SUMSOK | SUMABS | Age | total |
|---|---|---|---|---|---|---|---|---|
| 2483 | CUB_00044 | 28 | 5 | 1 | 30 | 4 | 28.5 | 33 |
| 3464 | STI_03063 | 28 | 30 | 6 | 3 | 1 | 27.0 | 58 |
| 3948 | SUN_00874 | 31 | 16 | 2 | 30 | 0 | 41.0 | 47 |
| 4658 | TL2_20 | 26 | 7 | 0 | 32 | 0 | 47.0 | 33 |
| 4846 | TL4_21 | 31 | 2 | 8 | 27 | 2 | 28.0 | 33 |
| 4883 | TL4_62 | 9 | 24 | 2 | 32 | 2 | 44.0 | 33 |
| 5002 | CO1_61-I | 24 | 9 | 0 | 8 | 2 | 32.0 | 33 |
| 10202 | la8_8724 | 21 | 13 | 1 | 30 | 2 | 47.0 | 34 |

The observation where the total is 33 and one where it is 34, could be a result of miss counting. As it is the most likely to have been calculated from the other value (so most likely the cause of the miscalculation), I shall assume the error is in the SUMPRE value. The other two values are "very wrong".

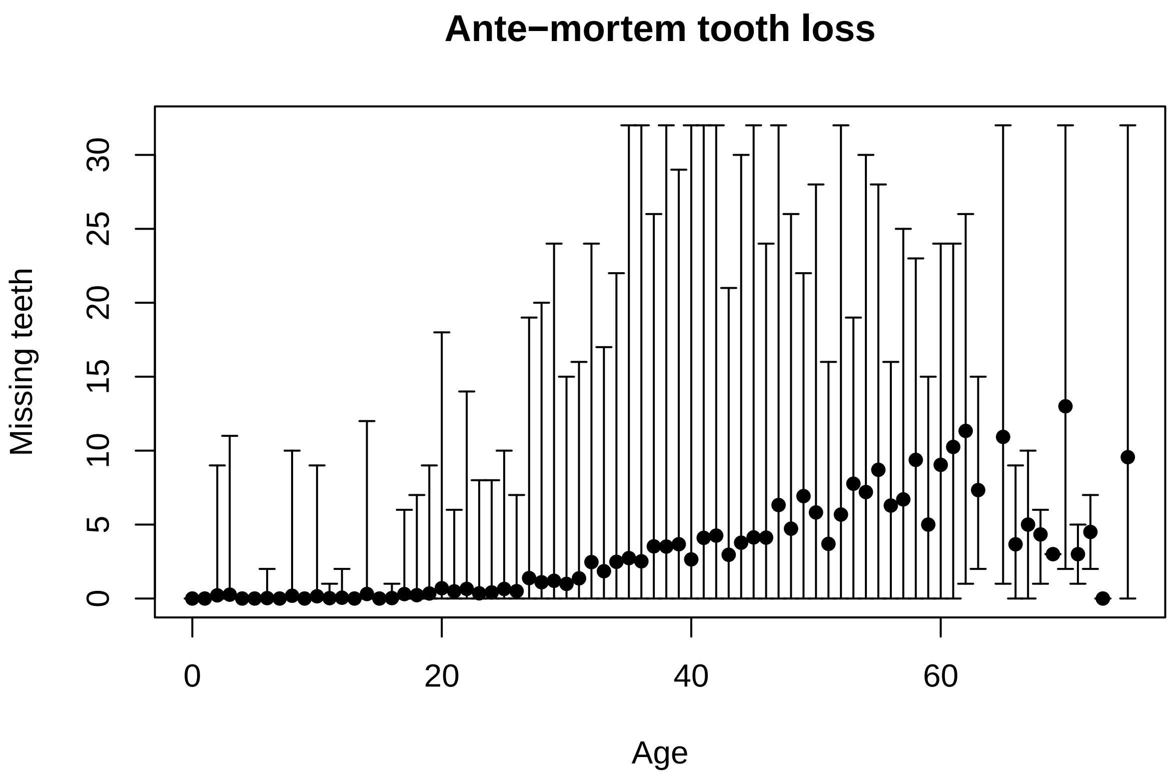
```
Miscount<-TooMany[TooMany$total<35,]
TooMany<-TooMany[TooMany$total>35,]
TooMany<-data.frame(ID=TooMany$cID, issue=rep("SUMTET+SUMPRE >32"))
Miscount<-data.frame(ID=Miscount$cID,issue=rep("SUMPRE miscounted"))
Issues<-rbind(Issues,TooMany,Miscount)

Known<-Known[Known$total<=32,]
```

## Tooth loss and age

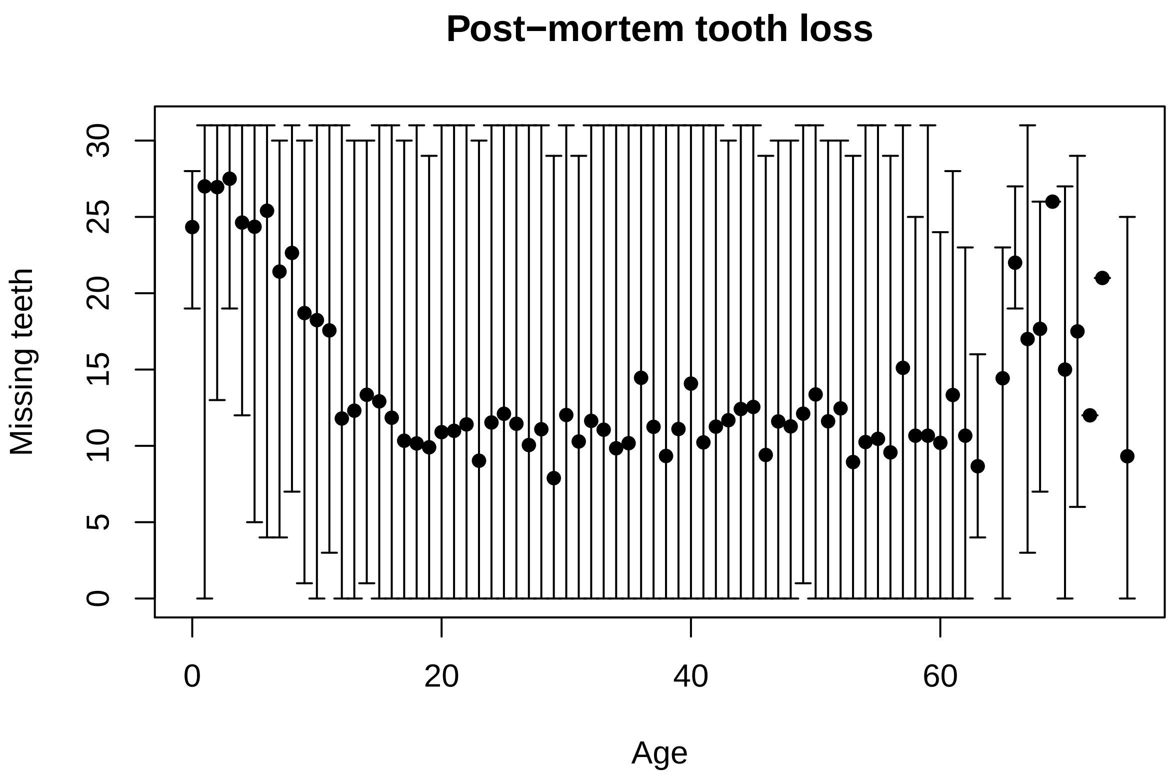
```
Known$SUMPOS<-32-Known$total# postmortem loss
Known$Missing<-32-Known$SUMTET# total missing
Known$Age<-round(Known$Age)

library(plyr)
Missing<-ddply(Known,"Age",summarise,max=max(Missing),min=min(Missing),mean=mean(Missing))
errbar(x=Missing$Age, y= Missing$mean, yplus=Missing$max, yminus=Missing$min,
       xlab="Age",ylab="Missing teeth")
title(main="Total missing teeth")
```



```
SUMPRE<-ddply(Known,"Age",summarise,max=max(SUMPRE),min=min(SUMPRE),mean=mean(SUMPRE))
errbar(x=SUMPRE$Age, y= SUMPRE$mean, yplus=SUMPRE$max, yminus=SUMPRE$min,
       xlab="Age",ylab="Missing teeth")
title(main="Ante-mortem tooth loss")
```



```
SUMPOS<-ddply(Known,"Age",summarise,max=max(SUMPOS),min=min(SUMPOS),mean=mean(SUMPOS))
errbar(x=SUMPOS$Age, y= SUMPOS$mean, yplus=SUMPOS$max, yminus=SUMPOS$min,
       xlab="Age",ylab="Missing teeth")
title(main="Post-mortem tooth loss")
```



### Abscesses and sockets

```
Abcess<-Teeth[,c("cID","SUMSOK","SUMABS","Age")]
Abcess$SUMSOK[is.na(Abcess$SUMSOK)]<-0
Abcess$SUMABS[is.na(Abcess$SUMABS)]<-0
```

Abscesses are recognized by a clear drainage passage leading from the tooth root(s) to the external surface of either maxilla or mandible. In order to record Abscesses they have to be socket to observe, so records with no sockets recorded shouldn't have Abscesses.

```
NoSocket<-Abcess[Abcess$SUMSOK==0,]
NoSocket<-NoSocket[NoSocket$SUMABS>0,]
```

However there are 245 observations were this is not the case.32-SUMPRE could be used as a substitute for sockets, as judging antemortem loss requires evidence of healing(i.e. erosion of the socket).

```
NoSocket<-Teeth[Teeth$cID %in% NoSocket$cID,]

NOrecord<-NoSocket[is.na(NoSocket$SUMPRE),]
ApproxSOK<-NoSocket[!(NoSocket$cID %in% NOrecord$cID),]

NOrecord<-data.frame(ID=NOrecord$cID,issue=rep("Abscesses should be NA"))
ApproxSOK<-data.frame(ID=ApproxSOK$cID, issue=rep("No sockets but Abscesses"))
Issues<-rbind(Issues,NOrecord,ApproxSOK)
```

The other problem with the socket recording is the instances with more than 32 sockets.

```r
TooManyS<-Abcess[Abcess$SUMSOK>32,]

kable(TooManyS)
```

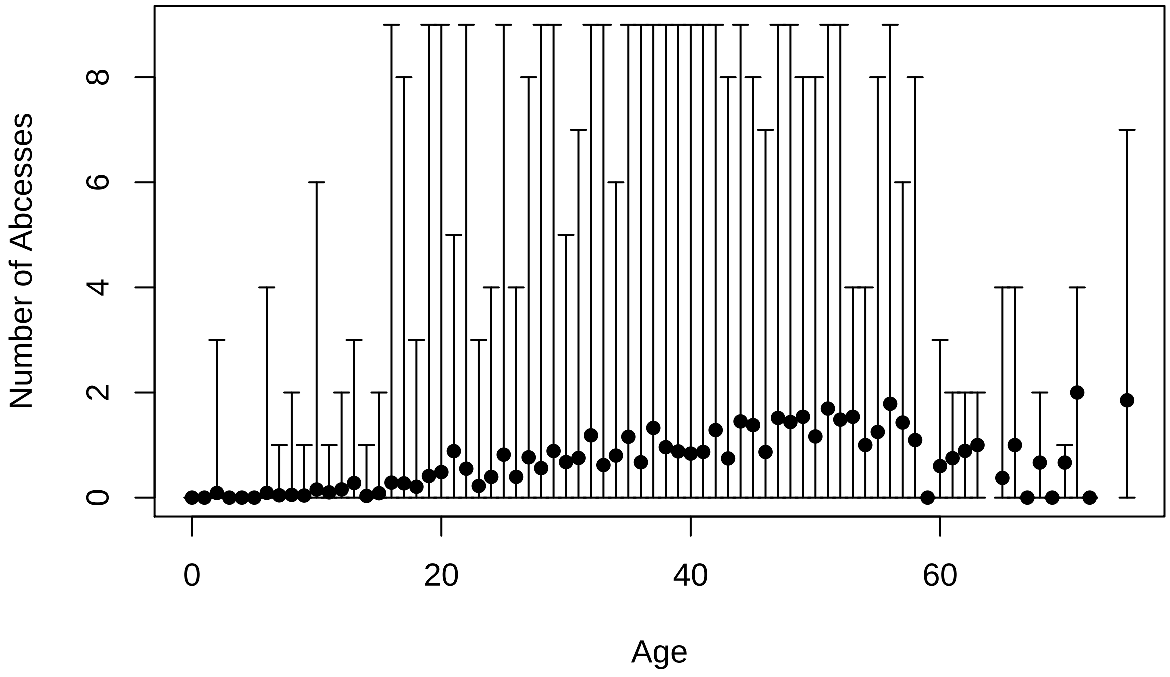
| | cID | SUMSOK | SUMABS | Age |
|---|---|---|---|---|
| 2399 | osg_00011 | 33 | 0 | 27.5 |
| 4429 | CUI_7 | 33 | 2 | 35.0 |
| 5016 | CO1_74-C | 33 | 0 | 17.0 |
| 5134 | CO2_259 | 33 | 1 | 27.0 |
| 8519 | WO7_2759 | 33 | 0 | 19.5 |
| 8565 | WO7_2805 | 33 | 0 | 25.6 |

For each of these there are 33 recorded sockets. This might be miscounting.

```r
TooManyS<-data.frame(ID=TooManyS$cID,issue=rep("33 sockets"))
Issues<-rbind(Issues,TooManyS)
```

### Age and Abscesses

```r
Sockets<-Abcess[Abcess$SUMSOK>0,]
Sockets<-Sockets[Sockets$SUMSOK<=32,]
Sockets$Age<-round(Sockets$Age)

s<-ddply(Sockets,"Age",summarise,max=max(SUMABS),min=min(SUMABS),mean=mean(SUMABS))

errbar(x=s$Age, y= s$mean, yplus=s$max, yminus=s$min,
       xlab="Age",ylab="Number of Abcesses")
```



# Part 4: Categorical data

## Trauma

```
Trauma<-HIdata[,c("cID","TRARM","TRLEG","TRNAS","TRFAC","TRSKUL","TRHAN","TRWEAP")]
kable(summary(Trauma[,-1]))
```

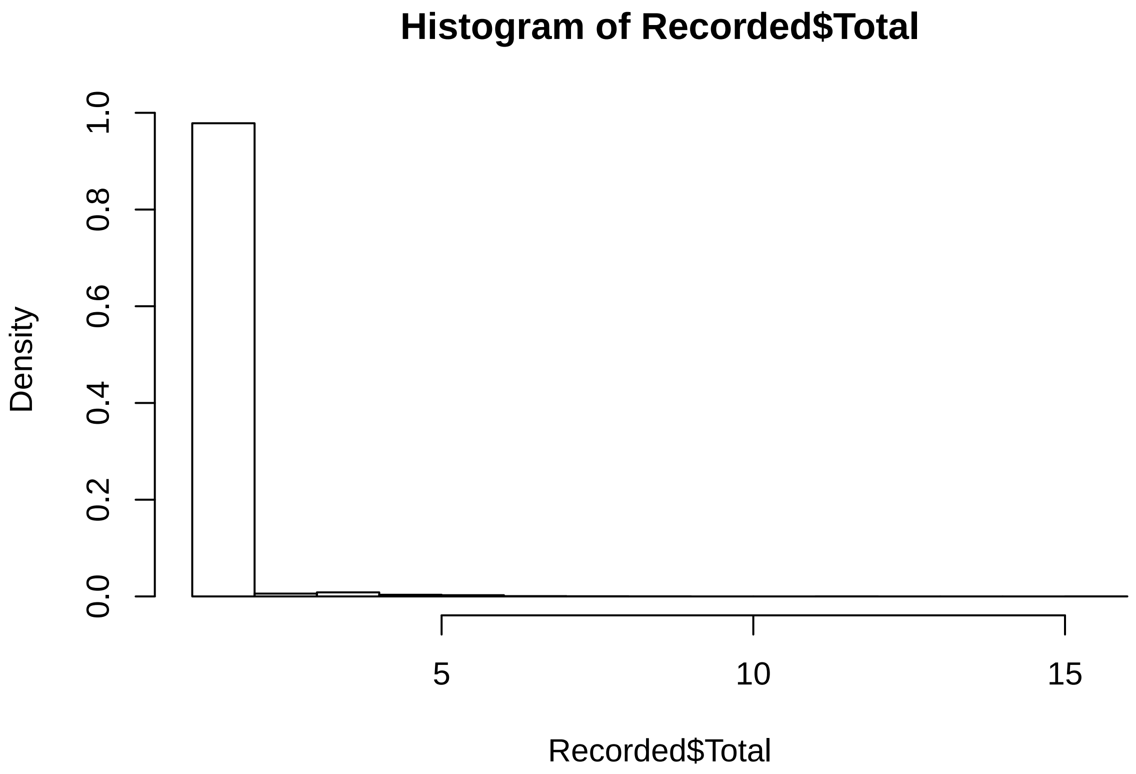
| TRARM | TRLEG | TRNAS | TRFAC | TRSKUL | TRHAN | TRWEAP |
|---|---|---|---|---|---|---|
| 0 :7545 | 0 :6963 | 0 :8977 | 0 :8644 | 0 :7033 | 0 :9815 | 0 : 1229 |
| 1 :4777 | 1 :5393 | 1 :3477 | 1 :3787 | 1 :5164 | 1 :2630 | 1 :10739 |
| 2 : 107 | 2 : 97 | 2 : 56 | 2 : 79 | 2 : 313 | 2 : 64 | 2 : 190 |
| 3 : 52 | 3 : 36 | NA's: 8 | NA's: 8 | NA's: 8 | NA's: 9 | NA's: 360 |
| 4 : 2 | 4 : 8 | NA | NA | NA | NA | NA |
| 5 : 27 | 5 : 12 | NA | NA | NA | NA | NA |
| NA's: 8 | NA's: 9 | NA | NA | NA | NA | NA |

There are NA's in every field; as 0 =no bones to be observed in most cases it is unclear what circumstance would need a NA. The only field that the codebook doesn't specify a 0 value is the weapon trauma (TRWEAP).However in the data there are 1229 instances of 0 in this column (along with 360 NA's).

### Total trauma

```
Ttrauma<-function(x){
  #this is a function to calculate total trauma given a database with trauma values
  #it output the enter dataframe with an additional column (total)
  #in which 0= all unobservable, 1= no trauma present, >1= total of present scored trauma

  Variables<-c("TRARM","TRLEG","TRNAS","TRFAC","TRSKUL","TRHAN","TRWEAP")
  x$Total<-NA # blank column to fill

  for (a in 1:length(x[,1])){# loops through all records
    sum<-sum(x[a,Variables])
    if(is.na(sum)){sum<-0}
    Adj<-0 #blank adjustment
    for (b in 1:7){
      if(!(is.na(x[a,Variables[b]]))){
        if (x[a,Variables[b]]==1){Adj<-Adj+1}}
    }
    if (Adj==0|sum==1){x$Total[a]<-sum
    }else{
      if((sum-Adj)==0){x$Total[a]<-1
      }else{x$Total[a]<-sum-Adj}}
  }
  x
}

Fact2Num<- function(x){
  # function from HIBasicClean.R
  a<-as.character(x)
  a<-as.numeric(a)
  a
}
```

```
Trauma[,2:8]<-apply(Trauma[,2:8],2,FUN=Fact2Num)
Trauma<-Ttrauma(Trauma)
Recorded<-Trauma[Trauma$Total>0,]
hist(Recorded$Total, xlim=c(1,16),breaks=17,prob=TRUE)
```



Mostly at 1= no recorded trauma.

```
Recorded$Total<-as.factor(Recorded$Total)
kable(t(summary(Recorded$Total)))
```

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 12 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|
| 10561 | 615 | 67 | 96 | 39 | 29 | 7 | 4 | 2 | 1 | 1 |

Largest number where there was recorded trauma are at 2 (this is a single instance of trauma).The maximum possible score is 20 (TRARM=5,TRLEG=5,TRWEAP=2,TRNAS=2,TRFAC=2,TRSKUL=2,TRHAN=2). The maximum recorded here is 16.

Scores of 3, 4 and 5 can represent single instances of trauma on either the arm (TRARM) or leg(TRLEG); which are scored on a 5 pont scale:

1. not fractured
2. healed fracture with acceptable alignment
3. healed and poorly aligned
4. healed with fusion of the joint
5. healed fracture with alignment unknown.

A score of 4 or 5 could also represent a double instance of trauma. Anything above a 5 must be multiple instances of trauma. And any odd number must involve trauma to either the arm or the leg.

```
kable(Recorded[Recorded$Total %in% 10:20,])
```

| | cID | TRARM | TRLEG | TRNAS | TRFAC | TRSKUL | TRHAN | TRWEAP | Total |
|---|---|---|---|---|---|---|---|---|---|
| 3327 | QUI_00557 | 5 | 5 | 1 | 1 | 1 | 2 | 1 | 12 |
| 3343 | QUI_00403 | 5 | 5 | 1 | 2 | 2 | 2 | 1 | 16 |

## DJD

Degenerative joint disease.

```
DJD<-HIdata[,c("cID","DJSH","DJHK","DJCER","DJTHO","DJLUM","DJTMJ","DJWR","DJHAN","Age")]
kable(summary(DJD[,-1]))
```

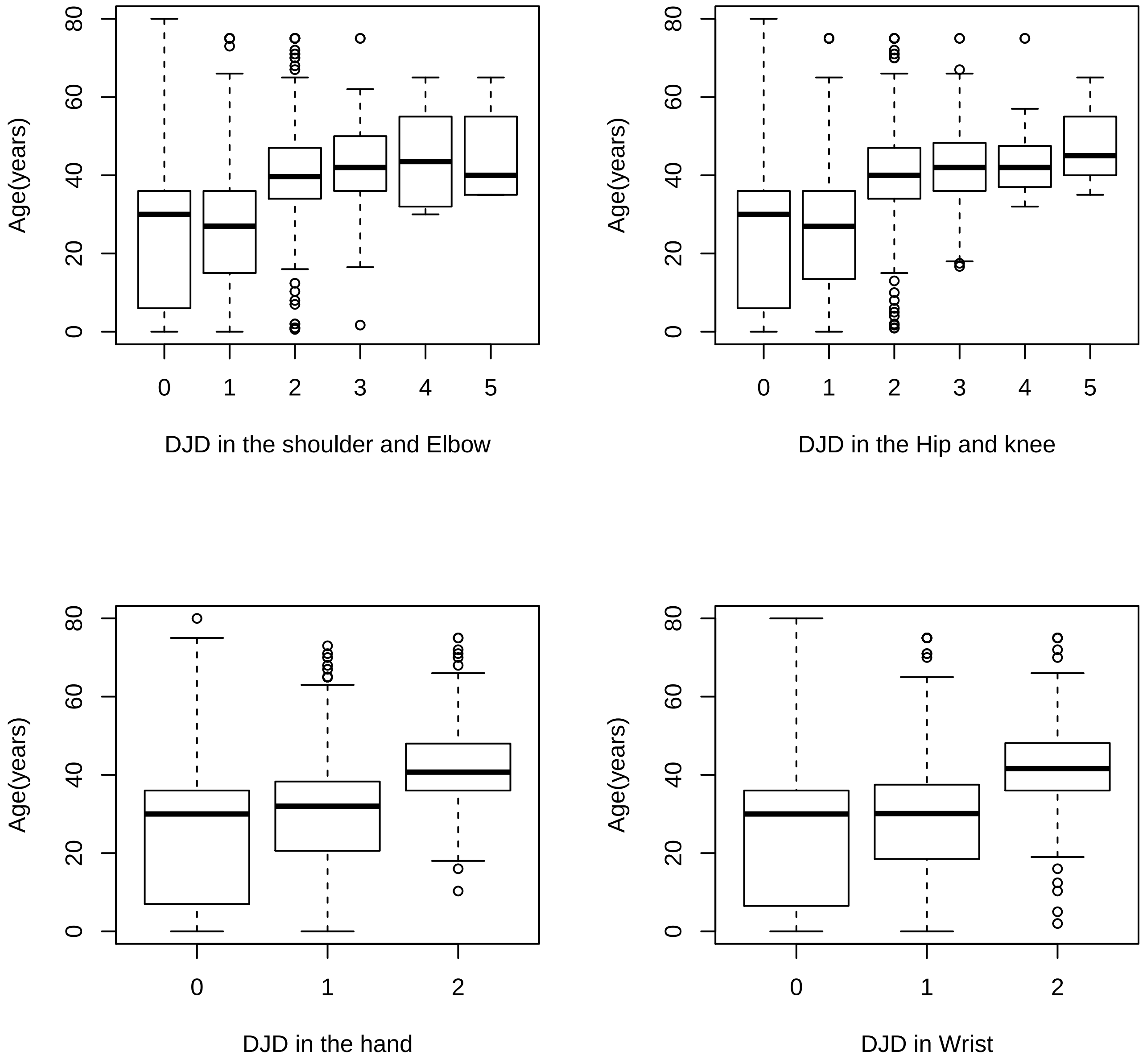
| DJSH | DJHK | DJCER | DJTHO | DJLUM | DJTMJ | DJWR | DJHAN | Age |
|---|---|---|---|---|---|---|---|---|
| 0 :7946 | 0 :7981 | 0 :9062 | 0 :9312 | 0 :9182 | 0 :9222 | 0 :8927 | 0 :9686 | Min. : 0.00 |
| 1 :3030 | 1 :2918 | 1 :2371 | 1 :2032 | 1 :1905 | 1 :2524 | 1 :2807 | 1 :2326 | 1st Qu.:11.20 |
| 2 :1193 | 2 :1239 | 2 : 658 | 2 : 781 | 2 : 695 | 2 : 690 | 2 : 701 | 2 : 424 | Median :32.00 |
| 3 : 246 | 3 : 277 | 3 : 276 | 3 : 256 | 3 : 539 | NA's: 82 | NA's: 83 | NA's: 82 | Mean :26.52 |
| 4 : 14 | 4 : 17 | 4 : 69 | 4 : 55 | 4 : 116 | NA | NA | NA | 3rd Qu.:36.00 |
| 5 : 4 | 5 : 3 | NA's: 82 | NA's: 82 | NA's: 81 | NA | NA | NA | Max. :80.00 |
| NA's: 85 | NA's: 83 | NA | NA | NA | NA | NA | NA | NA's :826 |

As with the trauma field there are a number of NAs despite a score of 0 representing un-recordable data.

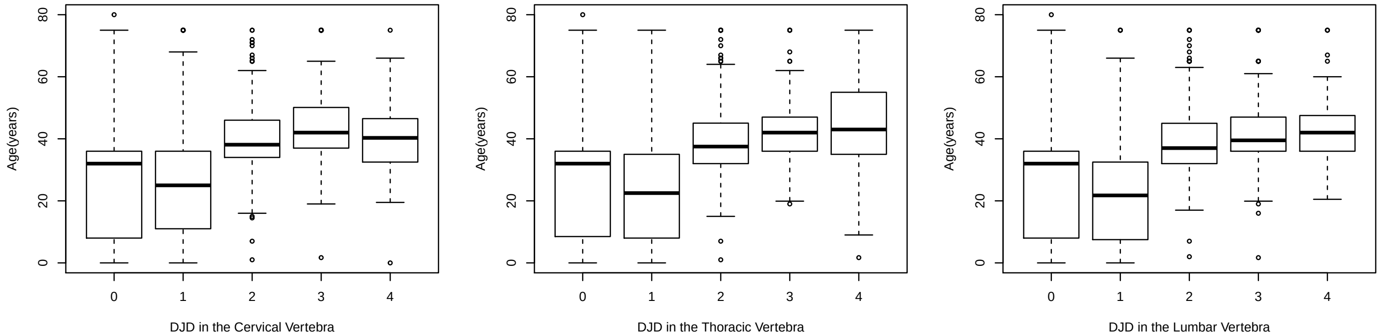
### DJD and Age

Although degenerative joint disease is use to look at activity patterns (chronic stress on the joints eventually damages the cartilaginous surfaces and, when sufficiently advanced, also the bone surface beneath.); It is often more correlated to age.

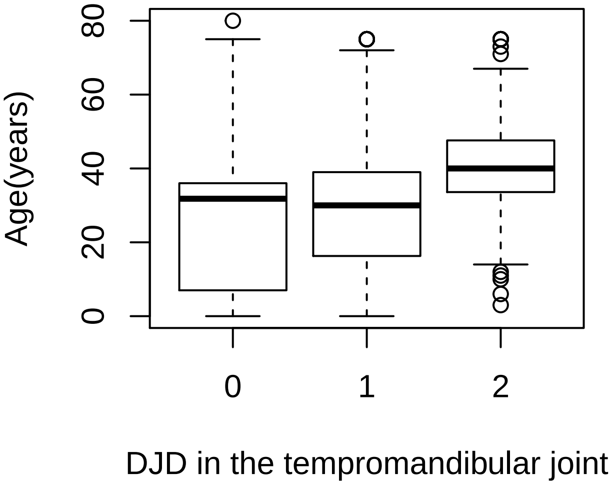
```
Aged<-DJD[!(is.na(DJD$Age)),]
par(mfrow=c(2,2))
plot(Aged$DJSH,Aged$Age,xlab="DJD in the shoulder and Elbow", ylab="Age(years)")
plot(Aged$DJHK,Aged$Age,xlab="DJD in the Hip and knee", ylab="Age(years)")
plot(Aged$DJHAN,Aged$Age,xlab="DJD in the hand", ylab="Age(years)")
plot(Aged$DJWR,Aged$Age,xlab="DJD in Wrist", ylab="Age(years)")
```



```r
par(mfrow=c(1,3))
plot(Aged$DJCER,Aged$Age,xlab="DJD in the Cervical Vertebra", ylab="Age(years)")
plot(Aged$DJTHO,Aged$Age,xlab="DJD in the Thoracic Vertebra", ylab="Age(years)")
plot(Aged$DJLUM,Aged$Age,xlab="DJD in the Lumbar Vertebra", ylab="Age(years)")
```



```r
par(mfrow=c(1,1))
plot(Aged$DJTMJ,Aged$Age,xlab="DJD in the tempromandibular joint", ylab="Age(years)")
```



For all these graphs the presences of DJD (moving from a score of 1 to a score of 2) does have a clear impact on average age. There also appear to be an effect on spread. There is however less evidence that the severity has any impact on age: n.b. 3 and 4 represent more severe forms of DJD (including major osteophyte growth or degeneration of the joint and immobilization); whereas 5 represent systemic degenerative disease such as rheumatoid arthritis, Alkaptonuria, etc.

## Hypoplasia

Linear enamel hypoplasia is recorded on incisors and canines, both deciduous and permanent.

```r
LEH<-HIdata[,c("cID","LDI","LDC","LPI","LPC","Age")]
kable(summary(LEH[,-1]))
```

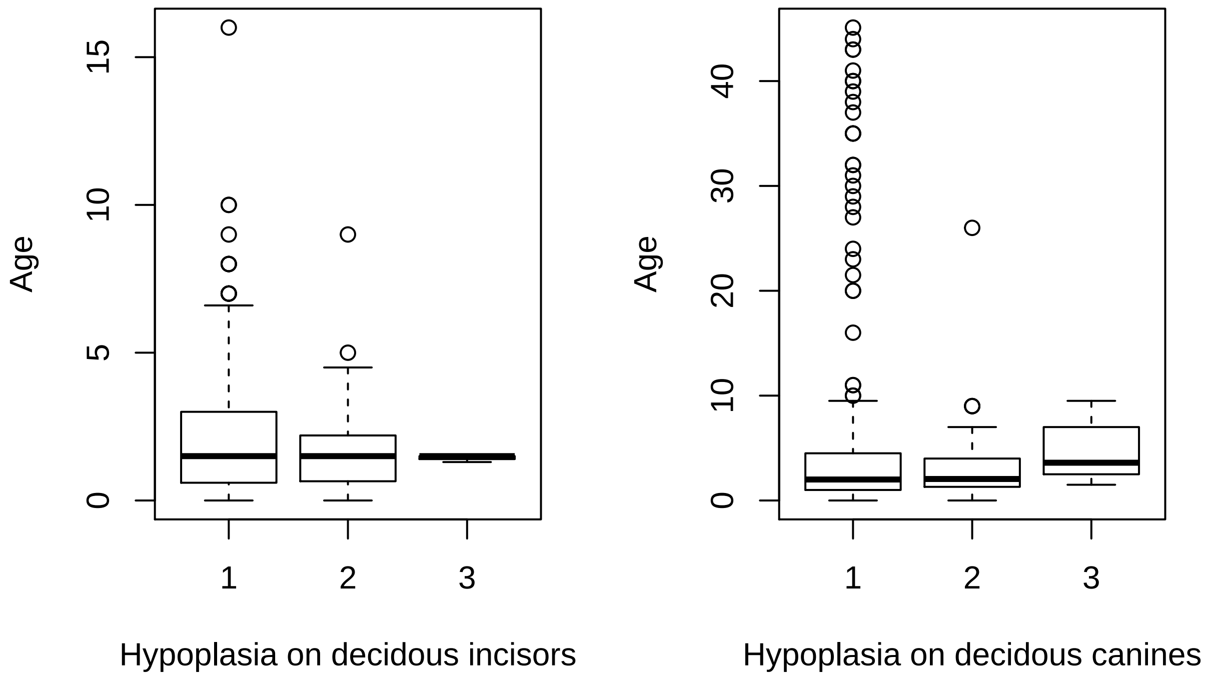
| LDI | LDC | LPI | LPC | Age |
|---|---|---|---|---|
| 0 :9886 | 0 :9822 | 0 :8237 | 0 :7343 | Min. : 0.00 |
| 1 : 556 | 1 : 613 | 1 :1896 | 1 :2025 | 1st Qu.:11.20 |
| 2 : 28 | 2 : 34 | 2 : 484 | 2 : 945 | Median :32.00 |
| 3 : 3 | 3 : 5 | 3 : 271 | 3 : 572 | Mean :26.52 |
| NA's:2045 | NA's:2044 | NA's:1630 | NA's:1633 | 3rd Qu.:36.00 |

| LDI | LDC | LPI | LPC | Age |
|---|---|---|---|---|
| NA | NA | NA | NA | Max. :80.00 |
| NA | NA | NA | NA | NA's :826 |

Again there are many NA despite 0 meaning missing data.

#### Age and deciduous teeth

```
Decidous<-LEH[,c("cID","LDI","LDC","Age")]
Decidous<-Decidous[!(is.na(Decidous$LDI)),];Decidous<-Decidous[!(is.na(Decidous$LDC)),]
Decidous<-Decidous[!(is.na(Decidous$Age)),]

par(mfrow=c(1,2))
plot(as.factor(as.character(Decidous$LDI[Decidous$LDI!=0])),Decidous$Age[Decidous$LDI!=0],
     xlab="Hypoplasia on decidous incisors",ylab="Age")
plot(as.factor(as.character(Decidous$LDC[Decidous$LDC!=0])),Decidous$Age[Decidous$LDC!=0],
     xlab="Hypoplasia on decidous canines",ylab="Age")
```



There is a potential issue on the canine graph: a number of outliers that are of adult age (adults don't have deciduous teeth).

```
Issue<-Decidous[Decidous$LDC!=0,]; Issue<-Issue[Issue$Age>18,]
summary(Issue)[,4]
```

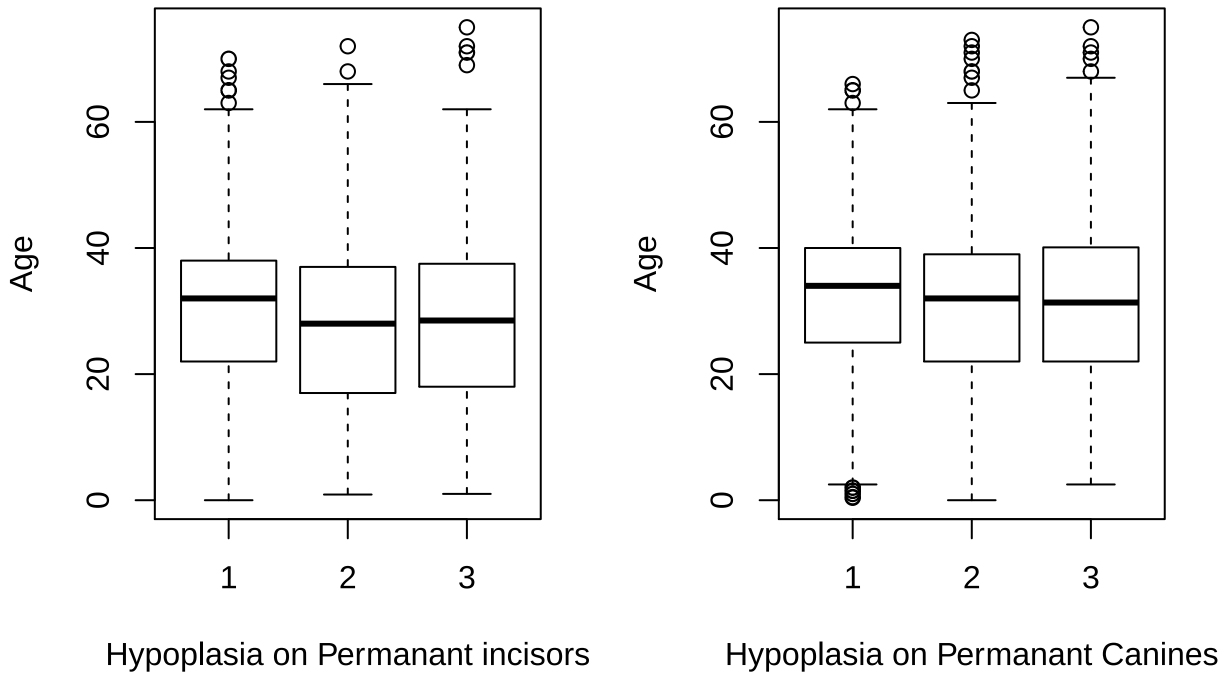
```
##
## "Min.   :20.00  " "1st Qu.:27.25  " "Median :33.50  " "Mean   :33.02  "
##
## "3rd Qu.:39.75  " "Max.   :45.10  "
```

There are 26 records where deciduous teeth have been recorded for adults.

```
Issue<-data.frame(ID=Issue$cID, issue=rep("Non-juvenile deciduous teeth"))
Issues<-rbind(Issues,Issue)
```

### Age and permanent teeth hypoplasia

```
Permanant<-LEH[,c("cID","LPI","LPC","Age")]
Permanant<-Permanant[!(is.na(Permanant$LPI)),];Permanant<-Permanant[!(is.na(Permanant$LPC)),]
Permanant<-Permanant[!(is.na(Permanant$Age)),]

par(mfrow=c(1,2))
plot(as.factor(as.character(Permanant$LPI[Permanant$LPI!=0])),Permanant$Age[Permanant$LPI!=0],
     xlab="Hypoplasia on Permanant incisors",ylab="Age")

plot(as.factor(as.character(Permanant$LPC[Permanant$LPC!=0])),Permanant$Age[Permanant$LPC!=0],
     xlab="Hypoplasia on Permanant Canines",ylab="Age")
```



There is a slight reduction in average age of those with LEH but the distribution of age across the groups are very similar. ##Infection

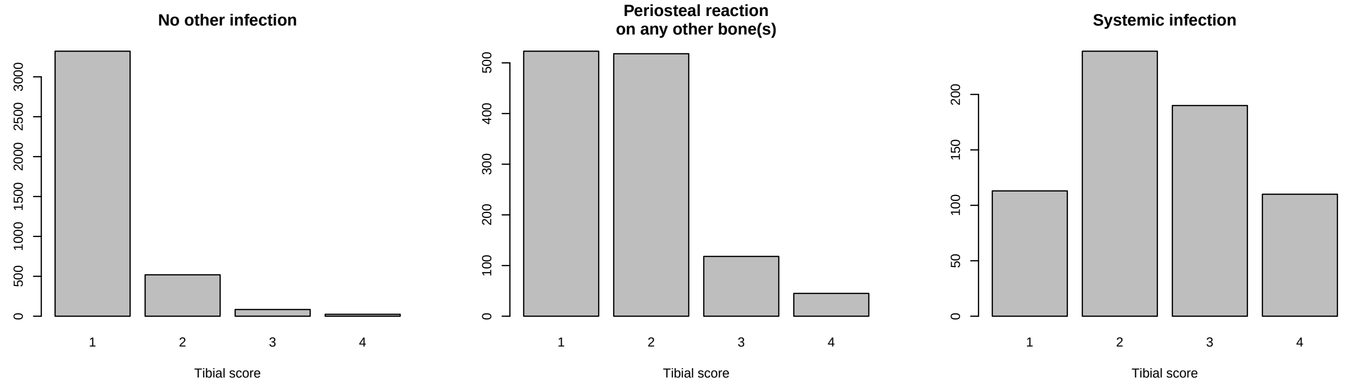
```
Infection<-HIdata[,c("cID","TIBINF","SKELINF")]
kable(summary(Infection[,-1]))
```

| TIBINF | SKELINF |
|---|---|
| 0 :6543 | 0 :9718 |
| 1 :4099 | 1 :1470 |
| 2 :1286 | 2 : 751 |

| TIBINF | SKELINF |
|---|---|
| 3 : 398 | NA's: 579 |
| 4 : 184 | NA |
| NA's: 8 | NA |

Same problem with NAs. n.b. for SKELINF 0= "no periosteal reaction on any other bone than the tibiae", not un-recordable. But for TIBINF 0 ="no tibia(e) present for scoring"

```
I<-Infection[Infection$TIBINF!=0,]; I$TIBINF<-factor(I$TIBINF)
par(mfrow=c(1,3))
plot(I$TIBINF[I$SKELINF==0],main="No other infection", xlab="Tibial score")
plot(I$TIBINF[I$SKELINF==1],main="Periosteal reaction
on any other bone(s)", xlab="Tibial score")
plot(I$TIBINF[I$SKELINF==2],main="Systemic infection", xlab="Tibial score")
```



Tibial scores:

1. no infectious lesions of the tibia(e) with at least one tibia available for observation
2. slight, small discrete patch(s) of periosteal reaction involving less than one quarter of the tibia(e) surface on one or both tibiae;
3. moderate periosteal reaction involving less than one half of the tibia(e) surface on one or both tibiae
4. severe periosteal reaction involving more than one-half of the tibia(e) surface (osteomyelitis is scored here)

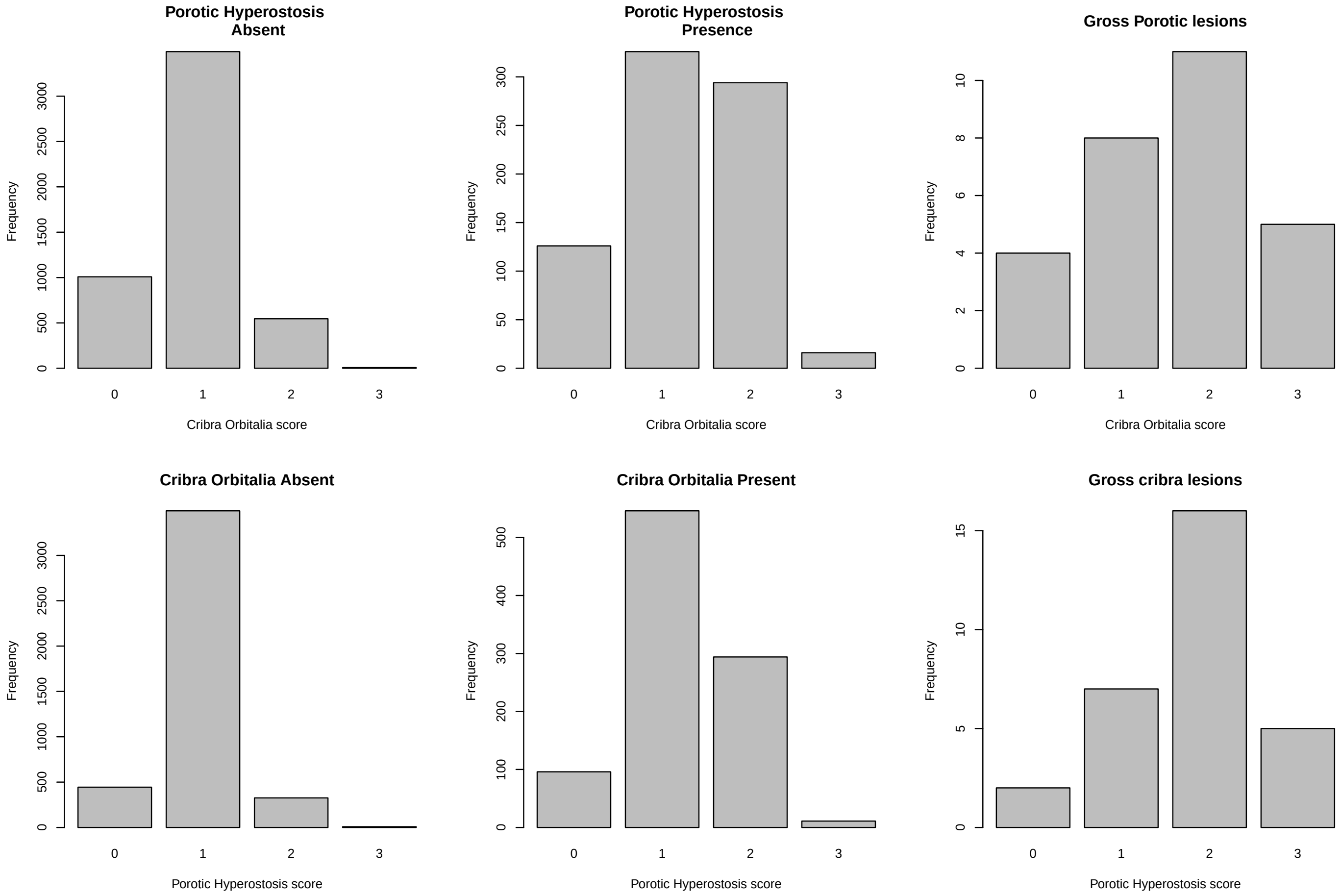
## Anemia/stress

```
Anemia<-HIdata[,c("cID","CROB","PORHY")]
kable(summary(Anemia[,-1]))
```

| CROB | PORHY |
|---|---|
| 0 :7256 | 0 :6659 |
| 1 :4270 | 1 :5054 |
| 2 : 955 | 2 : 762 |
| 3 : 30 | 3 : 28 |
| NA's: 7 | NA's: 15 |

NAs present again. 0 = "no orbits to be observed" /"no parietals to be observed"

```
par(mfrow=c(2,3))
plot(Anemia$CROB[Anemia$PORHY==1], xlab="Cribra Orbitalia score",ylab="Frequency",
     main="Porotic Hyperostosis
     Absent")
plot(Anemia$CROB[Anemia$PORHY==2], xlab="Cribra Orbitalia score",ylab="Frequency",
     main="Porotic Hyperostosis
     Presence")
plot(Anemia$CROB[Anemia$PORHY==3], xlab="Cribra Orbitalia score",ylab="Frequency",
     main="Gross Porotic lesions")
plot(Anemia$PORHY[Anemia$CROB==1], xlab="Porotic Hyperostosis score",ylab="Frequency",
     main="Cribra Orbitalia Absent")
plot(Anemia$PORHY[Anemia$CROB==2], xlab="Porotic Hyperostosis score",ylab="Frequency",
     main="Cribra Orbitalia Present")
plot(Anemia$PORHY[Anemia$CROB==3], xlab="Porotic Hyperostosis score",ylab="Frequency",
     main="Gross cribra lesions")
```



With increasing category of either the chance of being in a higher category for the other increases.

## Auditory Exostosis

```
kable(t(summary(HIdata$AUDEX)))
```

| 0 | 1 | 2 | NA's |
|---:|---:|---:|---:|
| 8080 | 2988 | 144 | 1306 |

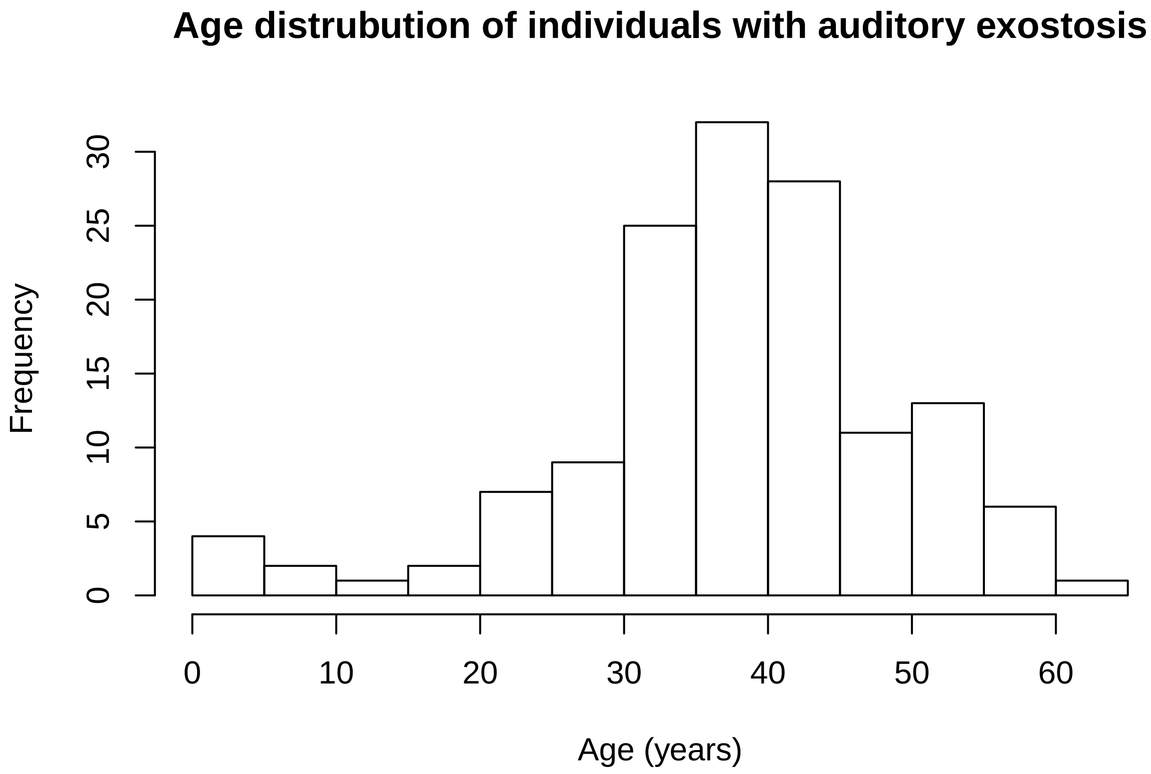
Na present again.

```
Exostosis<-HIdata[HIdata$AUDEX==2,]; Exostosis<-Exostosis[!(is.na(Exostosis$AUDEX)),]
kable(summary(Exostosis[,c("Sex","Age","Ancestry","SOC")]))
```

| Sex | Age | Ancestry | SOC |
|---|---|---|---|
| 1:38 | Min. : 0.00 | 1:139 | 111 : 23 |
| 2:17 | 1st Qu.:32.50 | 2: 3 | 343 : 2 |
| 3:49 | Median :38.00 | 3: 2 | 1 : 1 |
| 4:29 | Mean :38.14 | 5: 0 | 340 : 1 |
| 5: 7 | 3rd Qu.:45.00 | 6: 0 | 344 : 1 |
| 6: 4 | Max. :62.00 | NA | (Other): 0 |
| NA | NA's :3 | NA | NA's :116 |

Those with exostoses are almost entirely Native American (ancestry =1). There are slightly more male (3,4) than female (1,2).

```
hist(Exostosis$Age, xlab="Age (years)",main="Age distrubution of individuals with auditory exostosis")
```



There are some children in this sample with Auditory exostosis this is contrary to previous research that has found children and infant's ear tend to be lesion free (Eastman & Rodning, 2001,Archaeological Studies of Gender in the Southeastern United States).

```
Exostosis$Site<-factor(Exostosis$Site)
summary(Exostosis$Site)
```

```
## 3La CO1 CO2 COY CUI DUF FAB gua J73 lat LNP MR1 rea sal SF1 STI TL2 TL3
##   1   2  16   1   4   1   1   1   1   3  20  10   2   1   1   2  22   3
## TL4 WLE WW7
##  49   2   1
```

The TL site are from Tlaltico, a pre-Columbian Mexican culture centered around lake Texcoco. Co is Cholula, pre-Columbian Mexican city by the Atoyac river (and near lake texcoco). MRI is from a coastal Chilean site. And LNP is from Shell mounds in Brazil, a fisher-hunter-gather society on the southeastern coast.

## Next step

This concludes the exploratory analysis of the data. The next stage is to attempt to correct/ compensate for the issues highlighted. (see "CorrectData.rmd")

```
summary(Issues$issue)
```

```
##                    Not sex 5               too young to sex
##                           38                            134
##       Single digit social code       MaxAge lower than MinAge
##                         2215                              7
##        Max and Min but no Age  Min and Max age should be NA
##                          120                             44
##          age outside age range Non Juvenile femoral diaphysis
##                           32                              3
##                Too many teeth             Erroneous Cavities
##                            5                              1
##          Cavities but no teeth               SUMTET+SUMPRE >32
##                           35                              2
##             SUMPRE miscounted         Abscesses should be NA
##                            6                              4
##       No sockets but Abscesses                     33 sockets
##                          241                              6
##    Non-juvenile deciduous teeth
##                           26
```

```
write.table(Issues,"data/Issues.txt",row.names=FALSE)
```