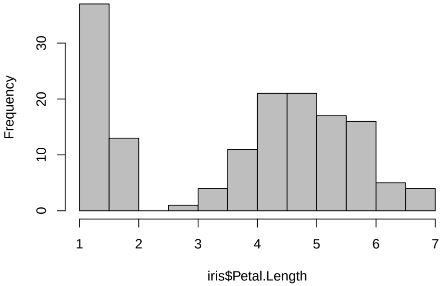
Load Some Data

```
> read.table.ISTA370<-function(filename){
   dataURL<-"http://www.sista.arizona.edu/~cohen/ISTA%20370/D
   # Reads a data frame from a URL path rooted at ISTA370 dat
   read.table(paste(dataURL,filename,sep=""))
 }
>
> # taheri<-read.table.ISTA370("taheri1.txt")
> # iris<-read.table.ISTA370("iris.txt")
> # heightC<-read.table.ISTA370("heightC.txt")
> # treering<-read.table.ISTA370("treering1.txt")
> # blast<-read.table.ISTA370("blastSummary.txt")
> # kinect<-read.table.ISTA370("onemovie.txt")
> # readability<-read.table.ISTA370("readability.txt")
```

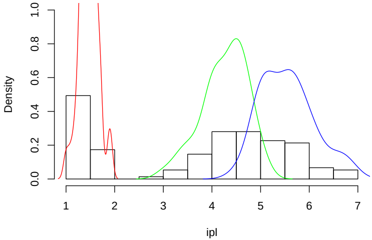
What Do You See? What Does it Mean?

```
> hist(iris$Petal.Length,col="grey",main=NULL)
```



What Do You See? What Does it Mean?

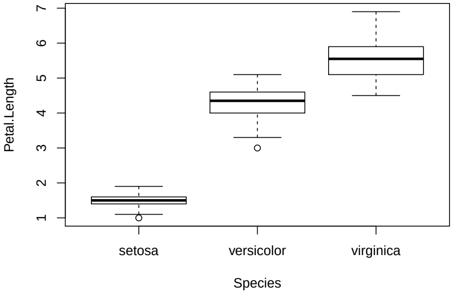
```
> ipl<-iris$Petal.Length
> hist(ipl,prob="true",ylim=c(0,1),main=NULL)
> lines(density(ipl[iris$Species=="setosa"]),col="red")
> lines(density(ipl[iris$Species=="versicolor"]),col="green")
> lines(density(ipl[iris$Species=="virginica"]),col="blue")
```



Looking at density curves for each species, we see that the histogram did indeed indicate two or more separate populations (species).

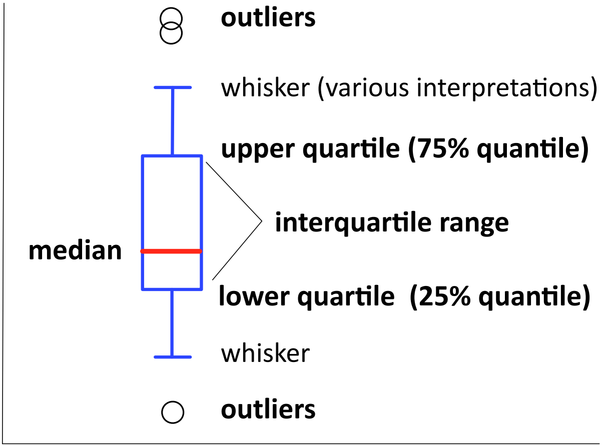
What Do You See? What Does it Mean?

```
> boxplot(iris$Petal.Length~iris$Species,
        ylab="Petal.Length",xlab="Species")
```



A boxplot by species confirms, and summarizes the petal length statistics for each species.

Boxplots


outliers
whisker (various interpretations)
upper quartile (75% quantile)
interquartile range
median
lower quartile (25% quantile)
whisker
outliers

Median, Quartiles, Interquartile Range

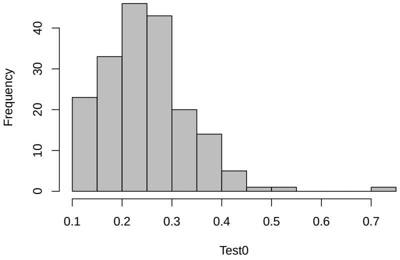
- If you sort the values in a sample from lowest to highest, the median is the middle value, or the average of the two middle values when the sample contains an even number of points.
- The median is the 50th quantile, or the value for which 50% of the values are greater.
- The lower quartile is the 25th quantile, above which 75% of the values are found.
- The upper quartile is the 75th quantile, above which 25% of the values are found.
- The interquartile range is a measure of variability and is the difference between the upper and lower quartiles.

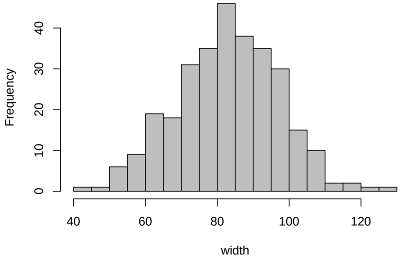
Median, Quartiles, Interquartile Range

- The median is *robust against outliers*; the mean is not.
- Suppose 100 families in a neighborhood each make \$40,000/year. When a millionaire moves in the mean jumps from \$40,000 to \$49,504/year. What happens to the median?
- Before the millionaire arrived, the variance in income was zero. Afterwards the variance is over *nine billion*!!! What happens to the interquartile range?
- Suppose you have a sorted sample of 9 elements; the median is the fifth element. If you add another element, what will the median be? By how many locations in the sorted distribution can the median shift?

Symmetry and Skew

```
> with(blast, hist(Test0,breaks=20,col="grey",main=NULL))
> with(treering, hist(width,breaks=20,col="grey",main=NULL))
```

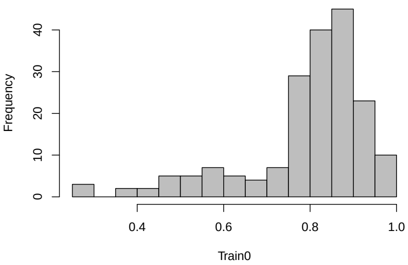


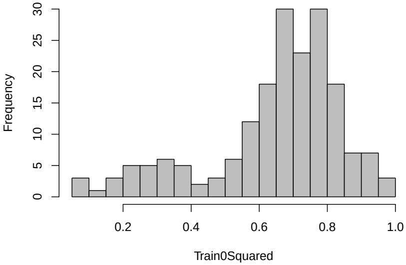


Test0 is skewed to the right, meaning it has a long tail of higher values, while Treering is nearly symmetric.

Transformations

```
> attach(blast)
> hist(Train0,breaks=20,col="grey",main=NULL)
> Train0Squared<-Train0^2  #square the Train0 data
> hist(Train0Squared,breaks=20,col="grey",main=NULL)
```

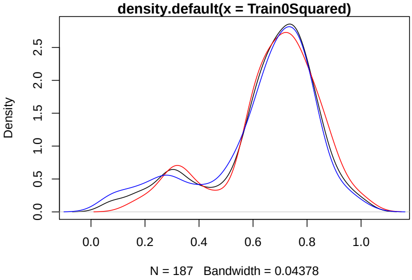




A simple transformation amplifies an otherwise hidden feature

Transformations

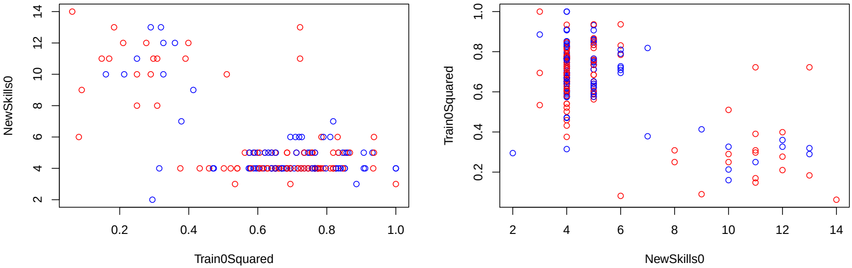
```
> Train0Squared<-with(blast,Train0^2)
> with(blast,plot(density(Train0Squared)))
> with(blast,lines(density(Train0Squared[gender=="female"]),c
> with(blast,lines(density(Train0Squared[gender=="male"]),col
```



Does gender explain the bump?

What Explains the Bump New Topics

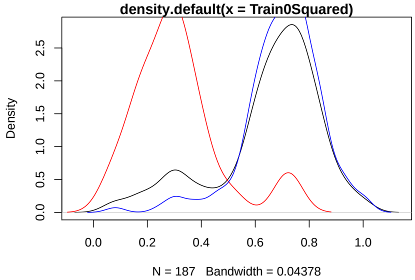
```
> plot(Train0Squared,NewSkills0,col=gender)
> plot(NewSkills0,Train0Squared,col=gender)
```



The number of topics to which students were exposed (NewSkills0) seems to explain the bump, but gender doesn't.

What Explains the Bump
New Topics

```
> precocious<-NewSkills0>8
> plot(density(Train0Squared))
> lines(density(Train0Squared[precocious=="TRUE"]),col="red")
> lines(density(Train0Squared[precocious=="FALSE"]),col="blue
```



So the students who saw too many subjects account for the bump.

Boxplots instead of density plots

```
> boxplot(TrainOSquared~precocious,
          xlab="precocious",
          ylab="proportion training items correct"
          )
```