IRIS EDA Part 2

First, we load the Iris Data Set

```
data(iris)
```

Let's first get a little feel for the data.

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

Now we are going to look at the empircal probability mass function for the Iris variables. In order to do this, we need to load the EnvStats library

```
library(EnvStats)
```

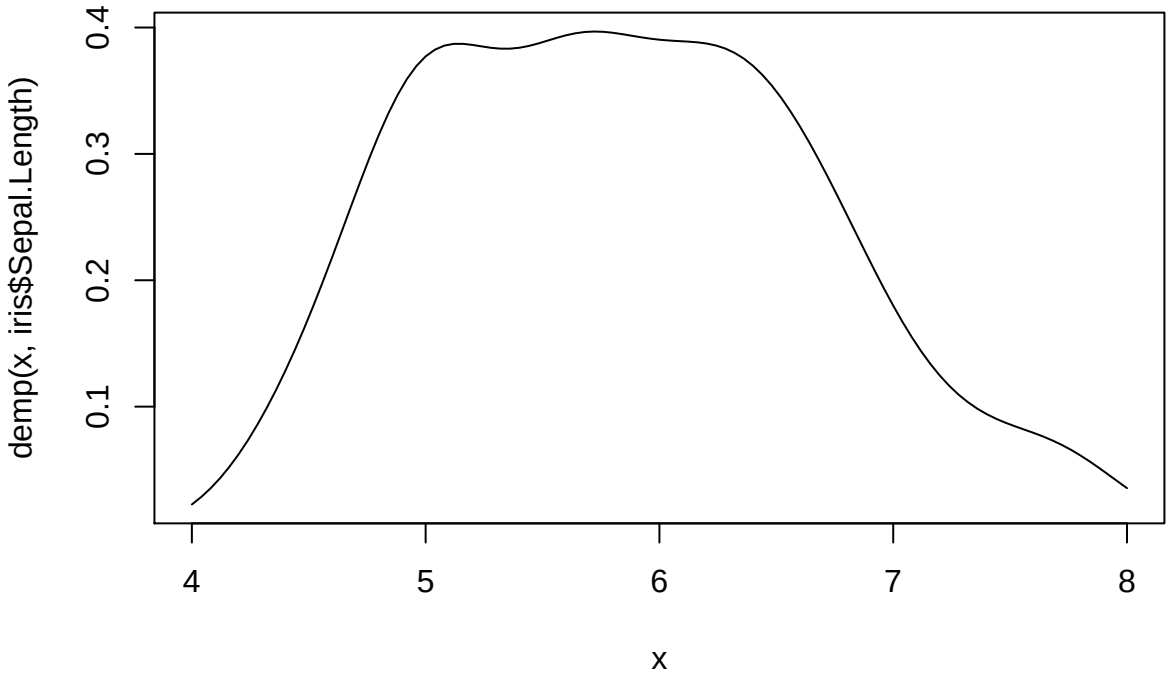
```
##
## Attaching package: 'EnvStats'
```

```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
## The following object is masked from 'package:base':
##
##     print.default
```

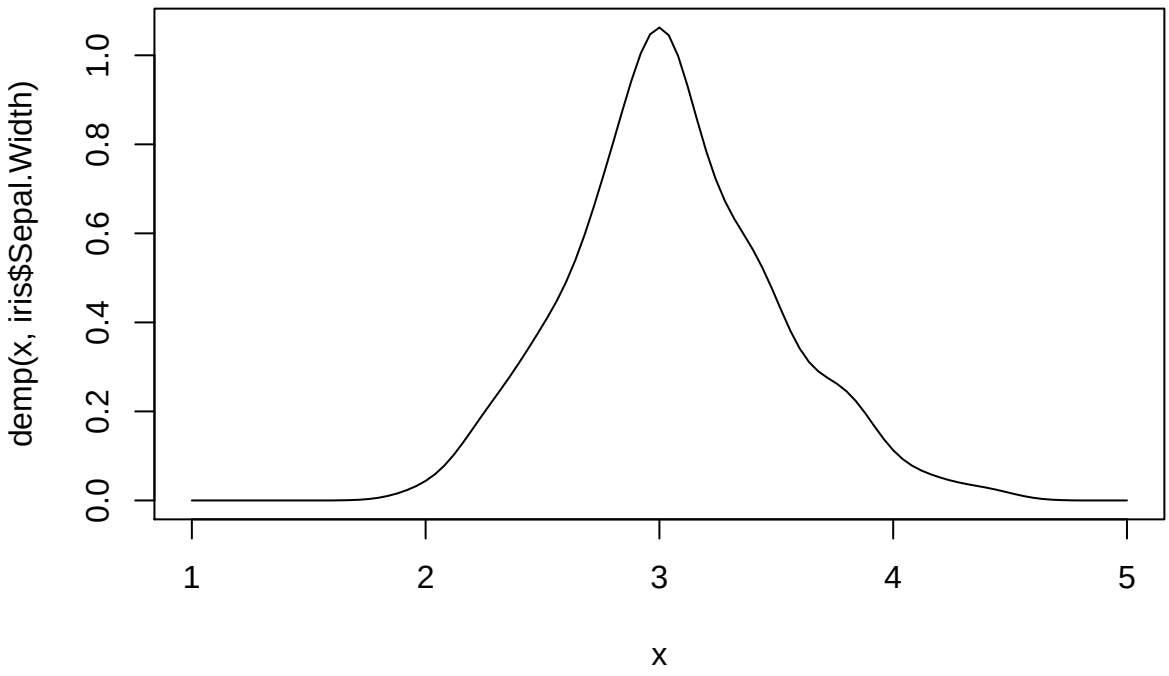
We know from the summary that sepal length has a minimum of 4.3 and a max of 7.9, so let's look at the empirical probability mass function over the range 4 to 8

```
curve(demp(x, iris$Sepal.Length), from=4, to=8)
```

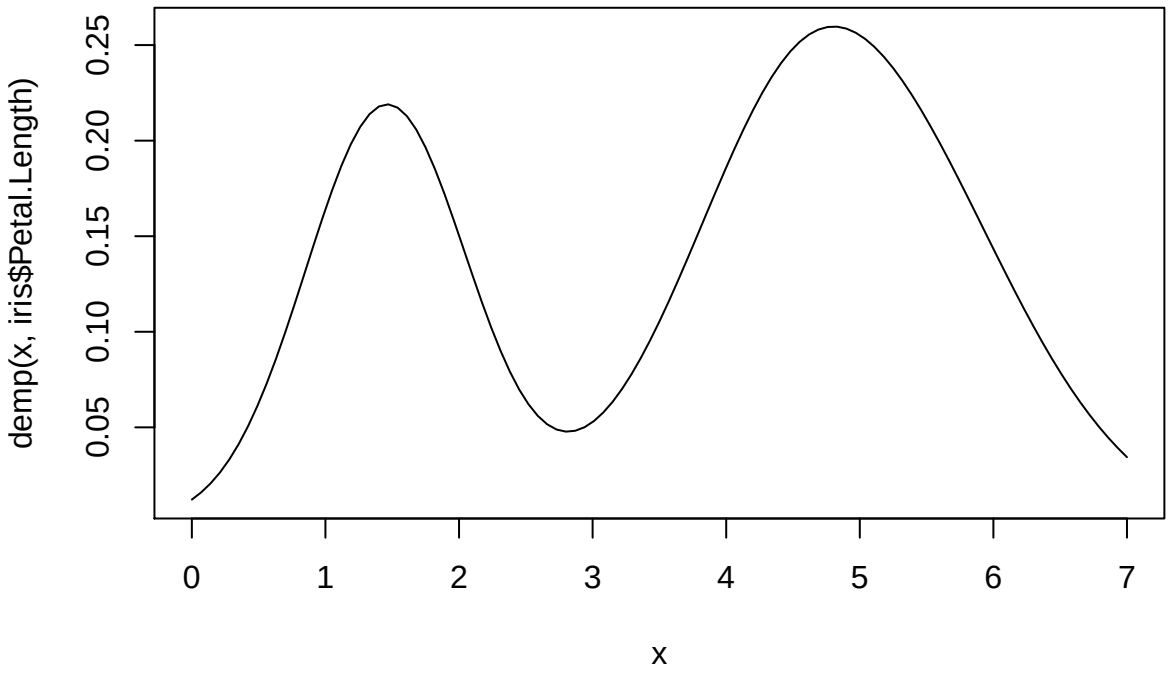


Let's repeat this for The other three variables

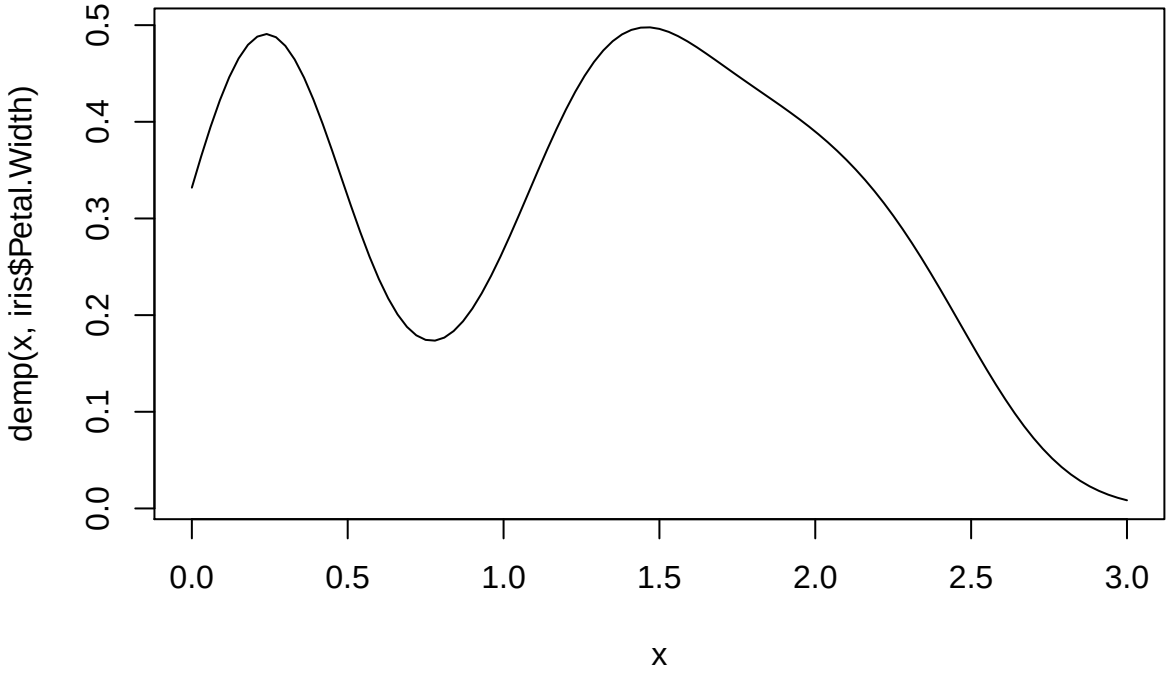
```
curve(demp(x, iris$Sepal.Width), from=1, to=5)
```




```
curve(demp(x, iris$Petal.Length), from=0, to=7)
```

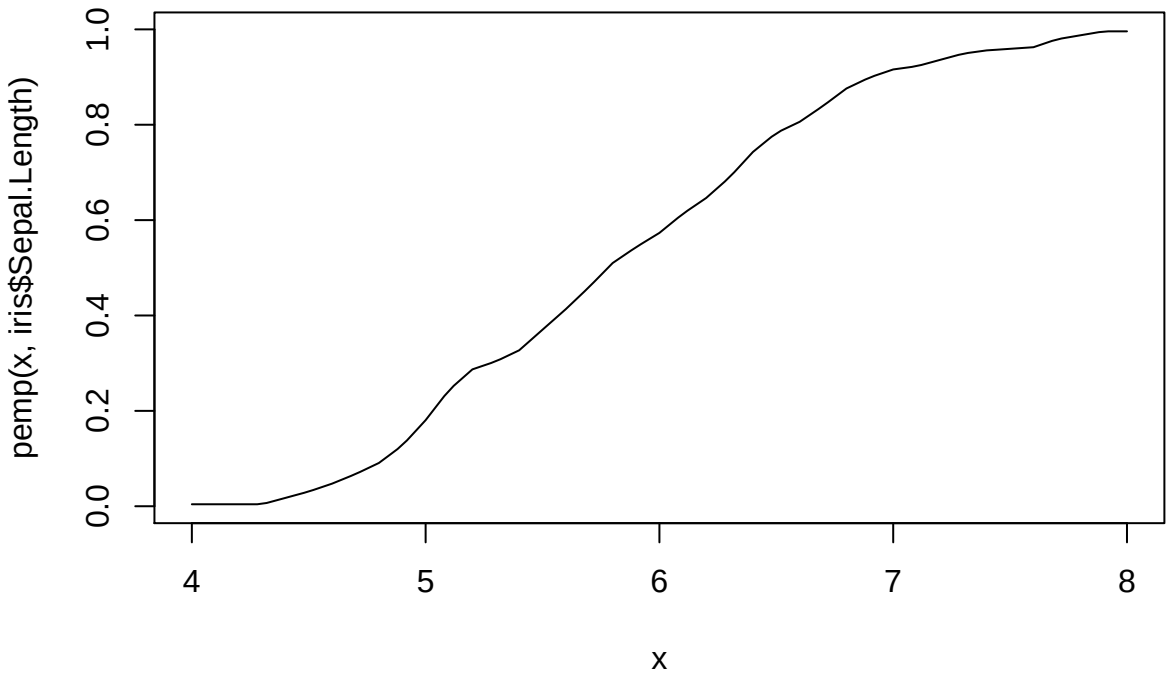



```
curve(demp(x, iris$Petal.Width), from=0, to=3)
```

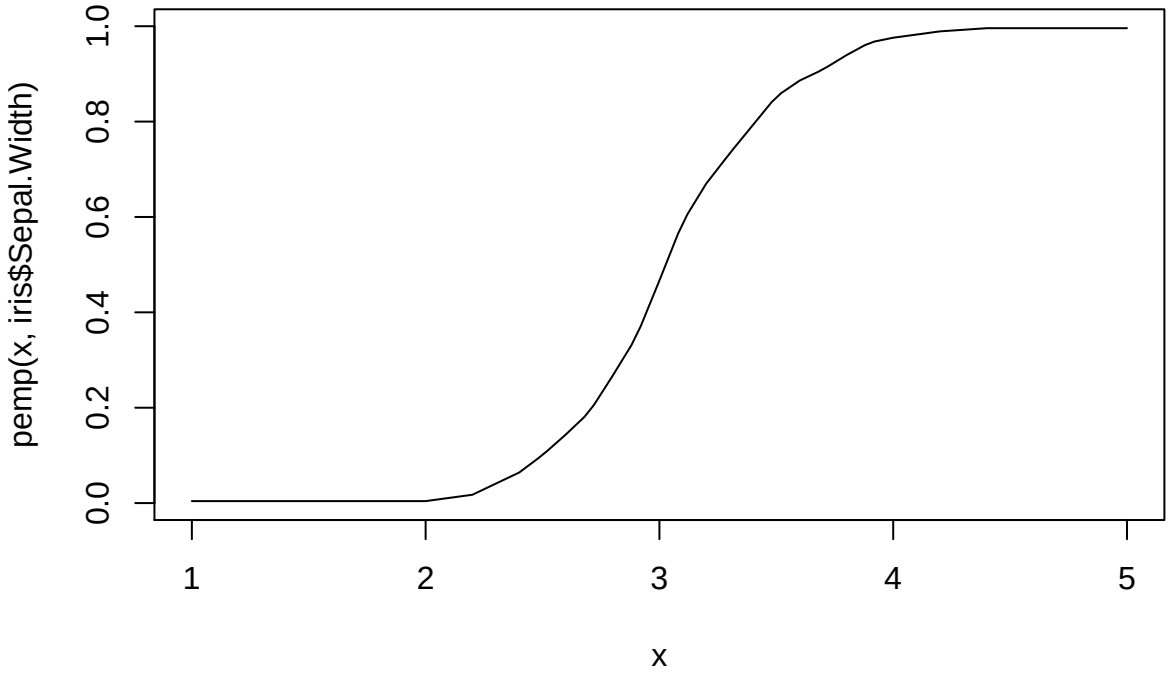


Now, Let's compute the empiracal cumulative distribution of each variable

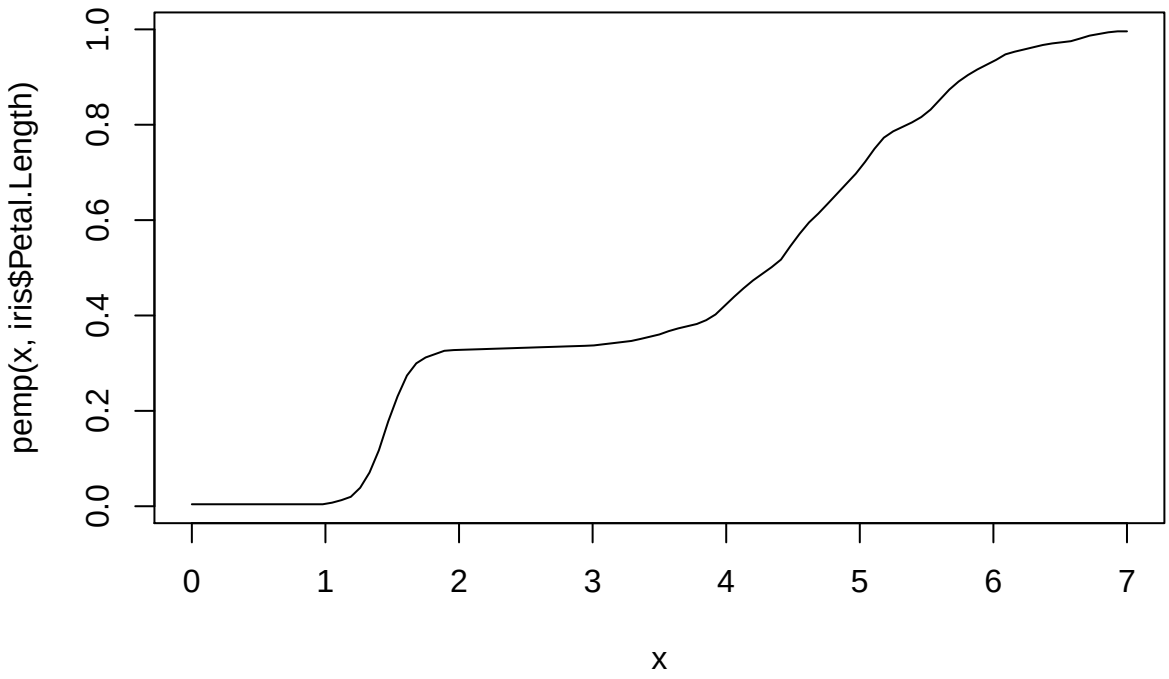
```
curve(pemp(x, iris$Sepal.Length), from=4, to=8)
```



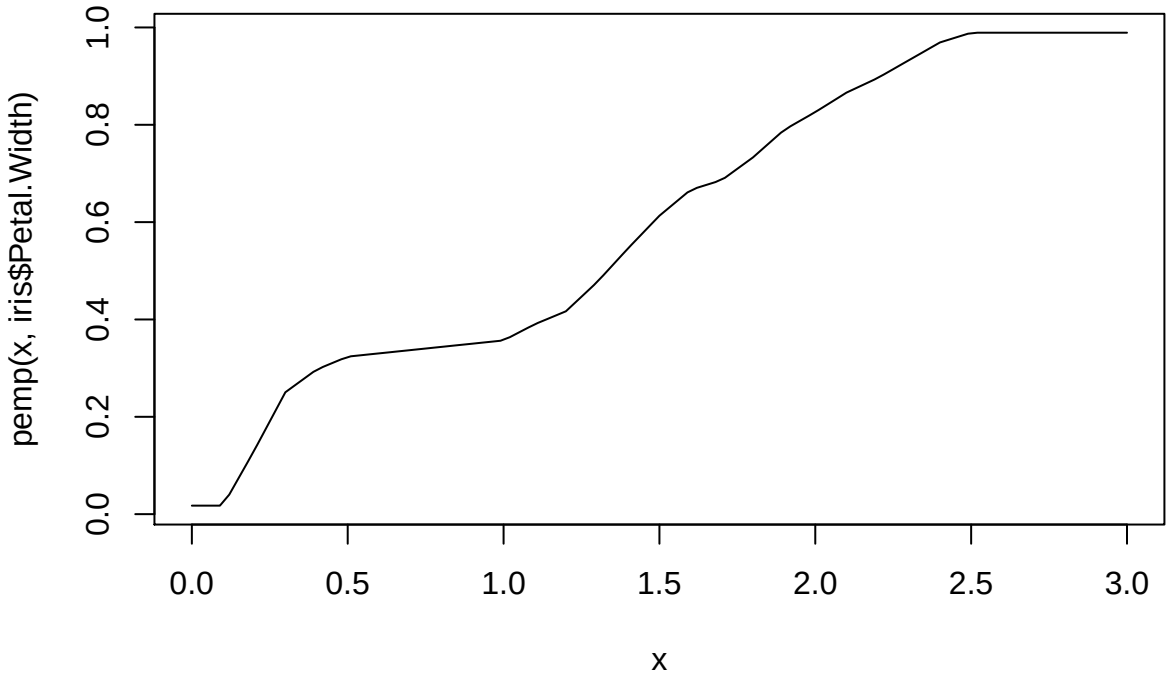

```
curve(pemp(x, iris$Sepal.Width), from=1, to=5)
```




```
curve(pemp(x, iris$Petal.Length), from=0, to=7)
```

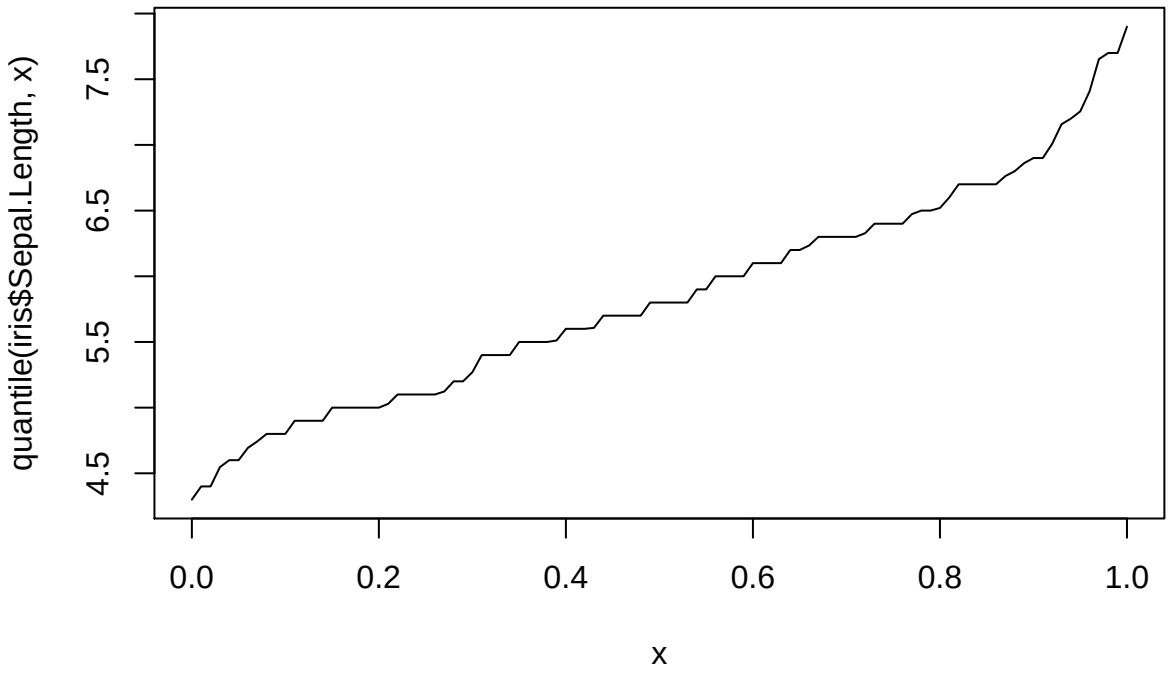



```
curve(pemp(x, iris$Petal.Width), from=0, to=3)
```

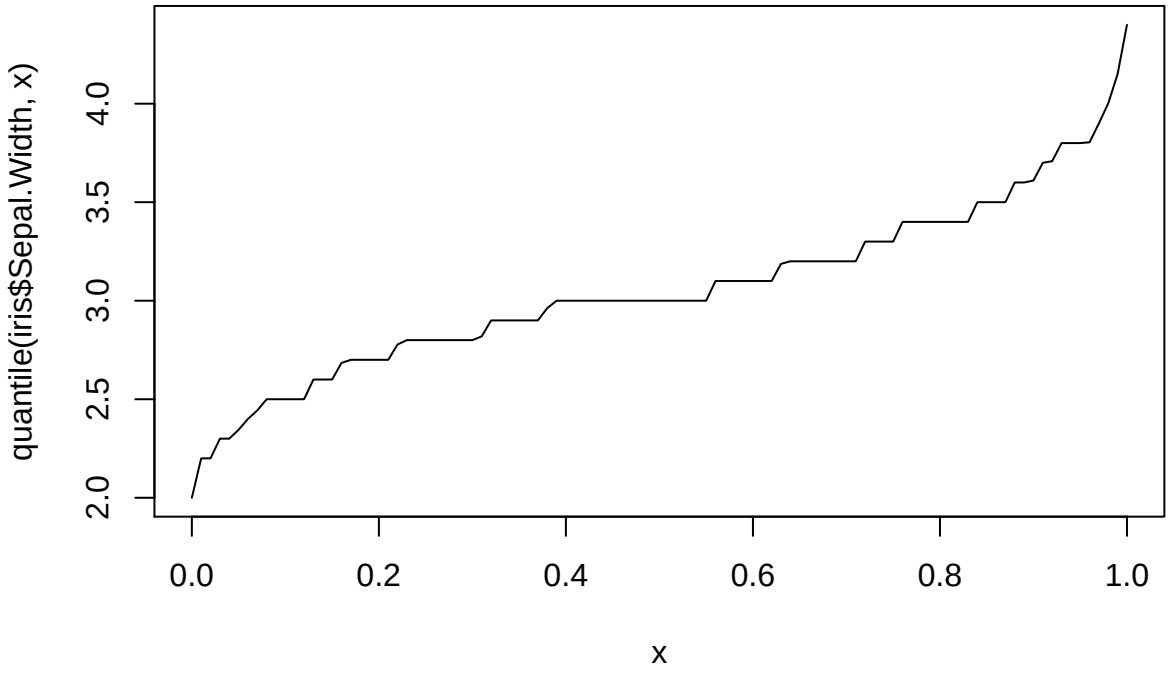


We can also observe the inverse cumulative distribution functions of each

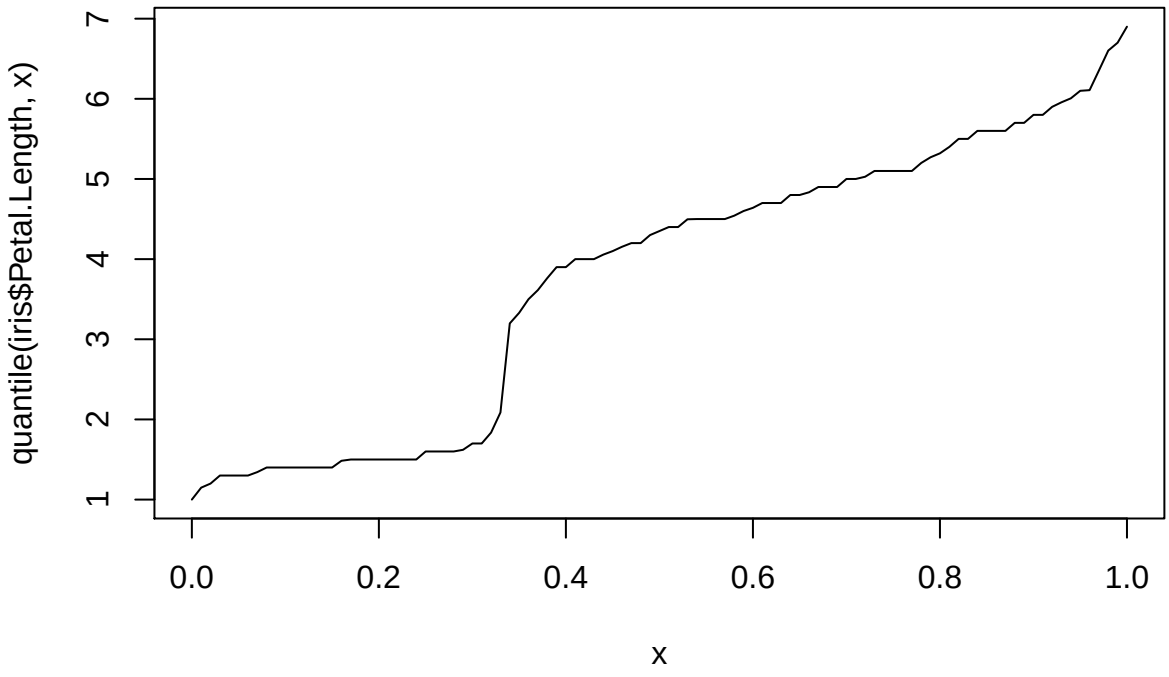
```
curve(quantile(iris$Sepal.Length, x), from=0, to=1)
```



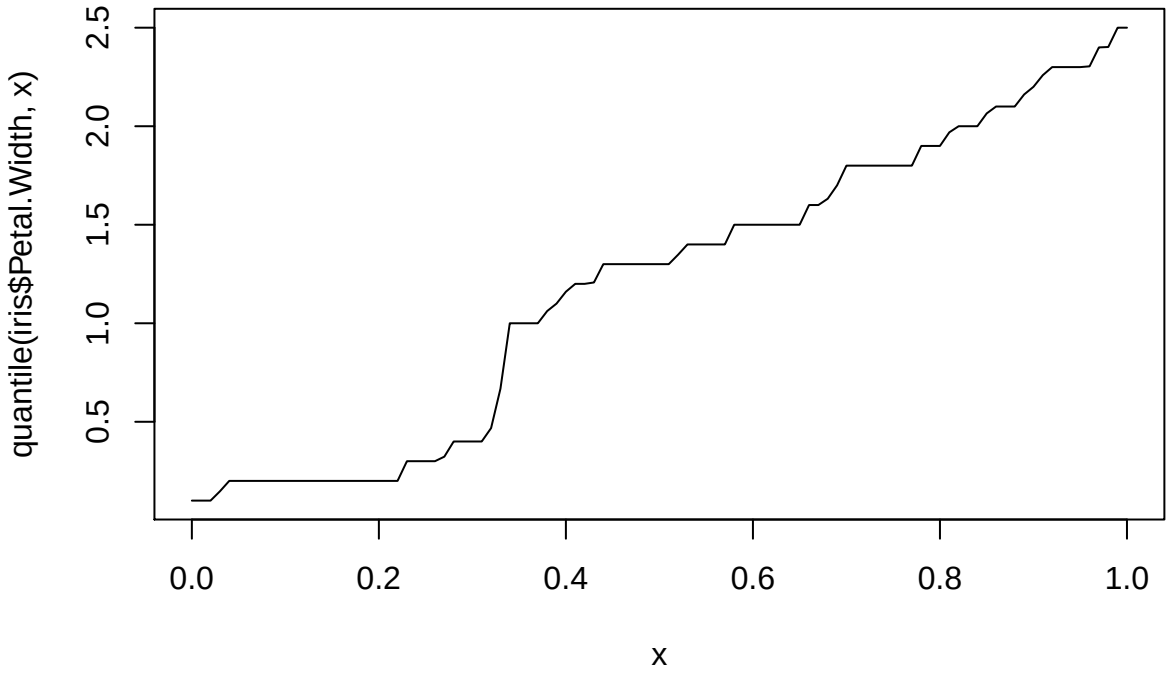

```
curve(quantile(iris$Sepal.Width, x), from=0, to=1)
```




```
curve(quantile(iris$Petal.Length, x), from=0, to=1)
```




```
curve(quantile(iris$Petal.Width, x), from=0, to=1)
```



We can also observe the variance of the variables

```
c(var(iris$Sepal.Length),
  var(iris$Sepal.Width),
  var(iris$Petal.Length),
  var(iris$Petal.Width)) -> s
s
```

```
## [1] 0.6856935 0.1899794 3.1162779 0.5810063
```

And, of course, we can compute the standard deviation

```
sqrt(s) -> s2
s2
```

```
## [1] 0.8280661 0.4358663 1.7652982 0.7622377
```

We'll throw in mean at no extra charge

```
c(mean(iris$Sepal.Length),
  mean(iris$Sepal.Width),
  mean(iris$Petal.Length),
  mean(iris$Petal.Width)) -> mu
mu
```

```
## [1] 5.843333 3.057333 3.758000 1.199333
```

Now, we can do a little multivariate analysis. Let's take a look at the covariance matrix for the iris data:

```
cov(iris[1:4])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
## Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
## Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
## Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063
```

And why not take a look at the correlation coeffecients as well?

```
cor(iris[1:4])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
## Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
## Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
## Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

Discuss: What does it mean when things are well correlated for the purposes of modeling the data?