Descriptive Statistics

September 10, 2013

Functions for calculating summary statistics of vector elements

Entry	Package	Description
min()	base	the minimum value of the numeric vector
max()	base	the maximum value of the numeric vector
range()	base	the range of the numeric vector
mean()	base	the arithmetic mean of the numeric vector
median()	stats	the median of a numeric vector
quantile()	stats	various sample quantiles of a numeric vector
IQR()	stats	the inter-quartile range of a numeric vector
fivenum()	stats	Tukey's five-number summary
sd()	stats	the standard deviation of a numeric vector
var()	stats	the variance of a numeric vector
pmin()	base	the parallel minima of two or more numeric vect
pmax()	base	the parallel maxima of two or more numeric vect
weighted.mean()	stats	the weighted mean of a numeric vector
mad()	stats	the median absolute difference of a numeric vect
rank()	base	the sample ranks of the values of a vector

Functions for calculating summary statistics of vector elements

Entry	Package	Description
smean.sd()	Hmisc	the mean and standard deviation of a numeric vector
wtd.mean()	Hmisc	the weighted mean of a numeric vector
wtd.var()	Hmisc	the weighted variance of a numeric vector
wtd.quantile	Hmisc	the weighted quantiles of a numeric vector
ecdf()	stats	the empirical CDF (ECDF) of a numeric vector
wtd.ecdf	Hmisc	the weighted ECDF of a numeric vector
wtd.rank	Hmisc	the weighted ranks of a numeric vector, using mid-ra
describe	Hmisc	concise statistical description of vector, matrix, data
cor()	stats	the correlation between two numeric vectors or matri
cov()	stats	the covariance between two numeric vectors or matri
cov2cor()	stats	Scales a covariance matrix into the corresponding co
density()	stats	the kernel density estimates of a numeric vector

Examples

Suppose, CEO yearly compensations are sampled and the following are found (in millions).
12, 0.4, 5, 2, 50, 8, 3, 1, 4, 0.25

```
> sals = scan() # read in with scan
1: 12 .4 5 2 50 8 3 1 4 0.25
11:
Read 10 items
> mean(sals) # the average
[1] 8.565
> var(sals) # the variance
[1] 225.5145
> sd(sals) # the standard deviation
[1] 15.01714
> median(sals) # the median
[1] 3.5
> fivenum(sals) # min,lower hinge, Median, upper hinge, max
[1] 0.25 1.00 3.50 8.00 50.00
> summary(sals)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.250 1.250 3.500 8.565 7.250 50.000
> mean(sals,trim=1/10) # This computes the 10%trimmed mean
[1]  4.425
> IQR(sals)
[1] 6
```

Examples

```
> data=c(10, 17, 18, 25, 28, 28)
> summary(data)
Min. 1st Qu. Median Mean 3rd Qu. Max.
10.00 17.25 21.50 21.00 27.25 28.00
> quantile(data,.25)
25%
17.25
> quantile(data,c(.25,.75)) # two values of p at once
25% 75%
17.25 27.25
```

Example

```
> x<-c(2,4,5,6,4,5,6,7,8,9,12,23)
> y<-c(5,1,3,5,6,7,8,9,3,21,13,21)
> pmin(x,y)
 [1]  2  1  3  5  4  5  6  7  3  9 12 21
> pmax(x,y)
 [1]  5  4  5  6  6  7  8  9  8 21 13 23
> range(x)
[1]  2 23
> smean.sd(x)
Error: could not find function "smean.sd"

> library(Hmisc)
> x
 [1]  2  4  5  6  4  5  6  7  8  9 12 23
> smean.sd(x)
    Mean       SD
7.583333 5.517877
```

Functions for calculating summary statistics of vector elements

`summary()`	Summary statistics of each column; type of statistics depends on data type
`apply()`	Apply a function to each column, works best if all columns are the same data type
`tapply()`	Divide the data into subsets and apply a function to each subset, returns an array
`by()`	Similar to tapply(), return an object of class by
`ave()`	Similar to tapply(), returns a vector the same length as the argument vector
`aggregate()`	Similar to tapply(), returns a dataframe
`sweep()`	Sweep "out" a summary statistic from a dataframe, matrix or array

Boxplot

A boxplot is a way of summarizing a set of data measured on an interval scale. It is often used in exploratory data analysis. It is a type of graph which is used to show the shape of the distribution, its central value, and variability. The picture produced consist five number summaries. The median for each dataset is indicated by center line, and the first and third quartiles are the edges of the box. The extreme values (within 1.5 times the inter-quartile range from the upper or lower quartile) are the ends of the lines extending from the IQR. Points at a greater distance from the median than 1.5 times the IQR are plotted individually as asterisks. These points represent potential outliers.

```
>x=c(24,58,61,67,71,73,76,79,82,83,85,87,88,88,92,93,94,97)
>boxplot(x, main="Boxplot of test scores", col=2)
> arrows(1,24,1.2,30)
> text(1.4,31,"This is an Outlier")
```

Example:

Link below provides the number of Atlantic hurricane from 1870 to 2010
http://biostatistics.it/Didattica/Dati/SilwoodWeather.txt
We will import the subject data and plot a boxplot for monthly data

```
>temperature="http://biostatistics.it/Didattica/Dati/SilwoodWeather.txt
>weather=read.table(temperature,header=T)
>attach(weather)
> names(weather)
[1] "upper" "lower" "rain" "month" "yr"
```

Before we can plot the data we need to declare month to be a factor. At the moment, R just thinks it is a number.

```
>month<-factor(month)
>plot(month,upper)
```