Basic numerical and graphical summaries

A variable may be classified as:

- **Quantitative:** numeric or integer
- **Ordinal:** ordered, like integers
- **Qualitative:** categorical, nominal, or factors

A data frame may contain variables of different types. Also, data may have additional structure. For example, when data are collected over time we have time series data, which has a time index. Similarly, when data are collected in space we have spatial data, which has information about the location.

Basic numerical and graphical summaries

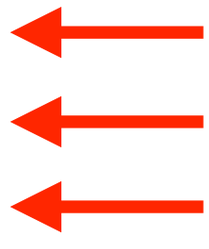
Example: body and brain size of mammals (two quantitative variables)

`?mammals` **gives you information about these data**

```
> library(MASS)
> head(mammals)
                    body brain
Arctic fox         3.385  44.5
Owl monkey         0.480  15.5
Mountain beaver    1.350   8.1
Cow              465.000 423.0
Grey wolf         36.330 119.5
Goat              27.660 115.0
```

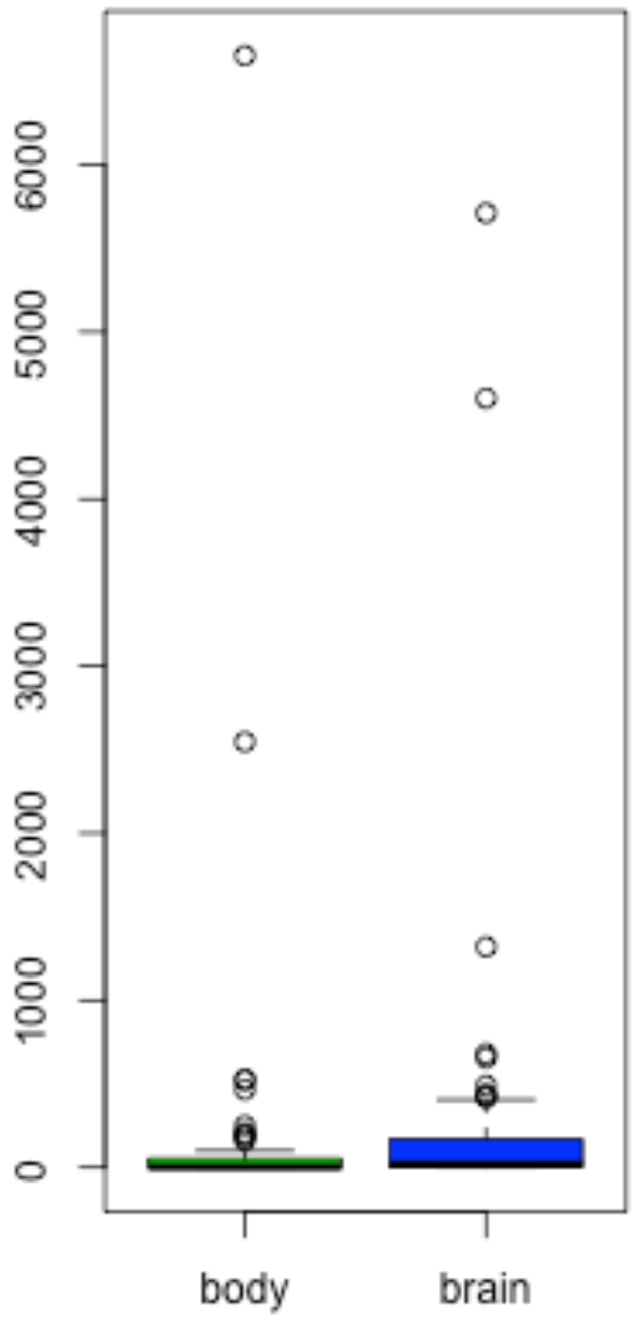
```
> summary(mammals)
      body                brain
 Min.   :   0.005   Min.   :   0.14
 1st Qu.:   0.600   1st Qu.:   4.25
 Median :   3.342   Median :  17.25
 Mean   : 198.790   Mean   : 283.13
 3rd Qu.:  48.203   3rd Qu.: 166.00
 Max.   :6654.000   Max.   :5712.00
```

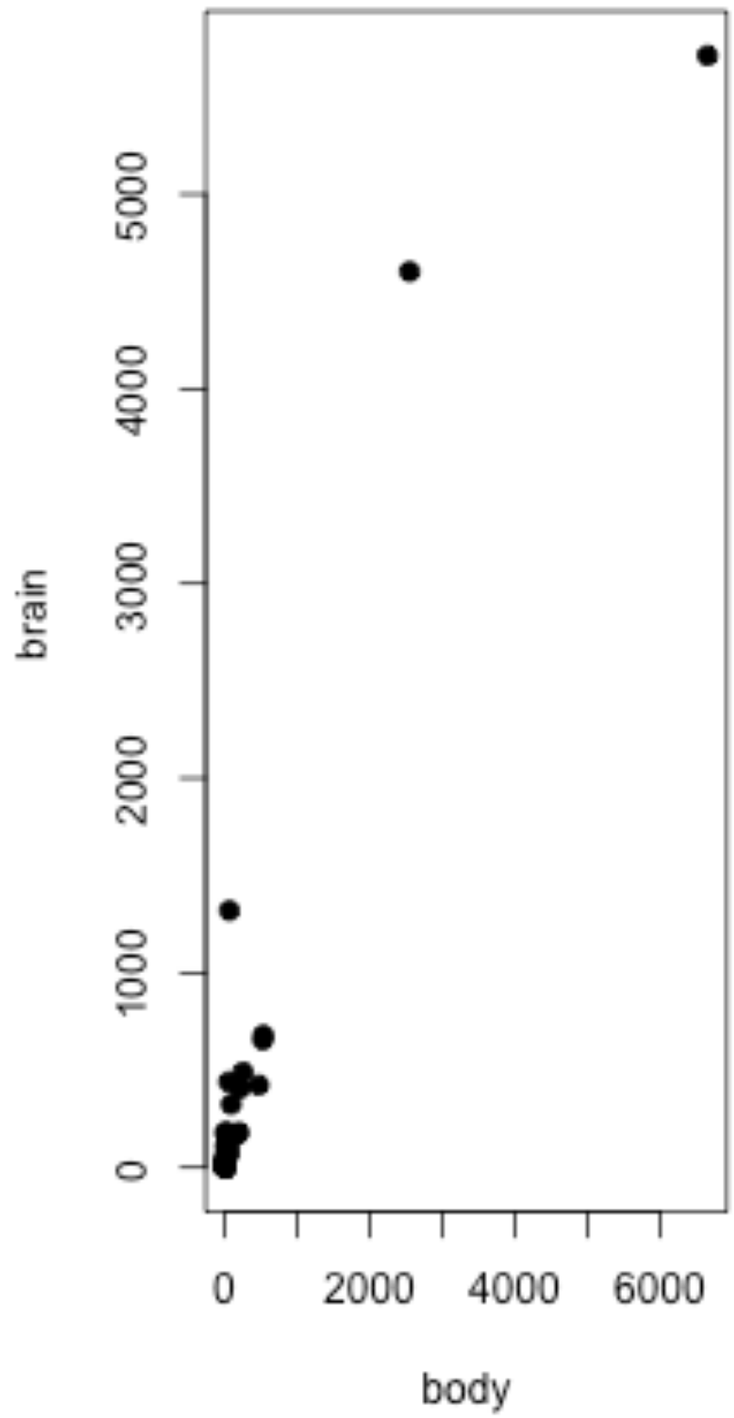
Extreme observations!



Basic numerical and graphical summaries

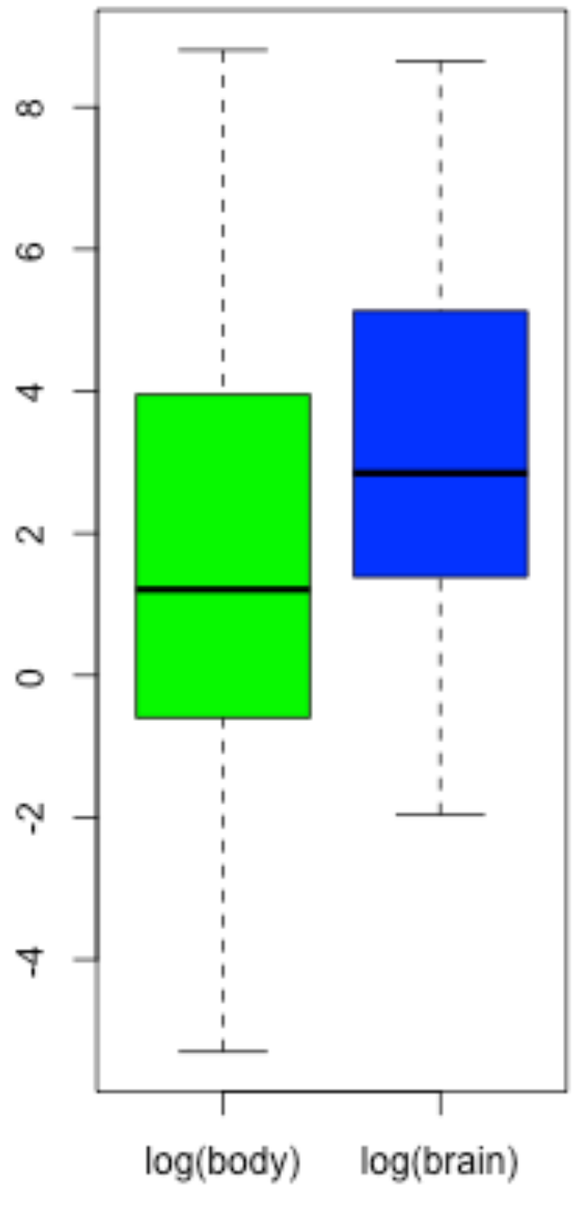
```
> boxplot(mammals,col=c('green','blue'))
> plot(mammals,pch=19)
```

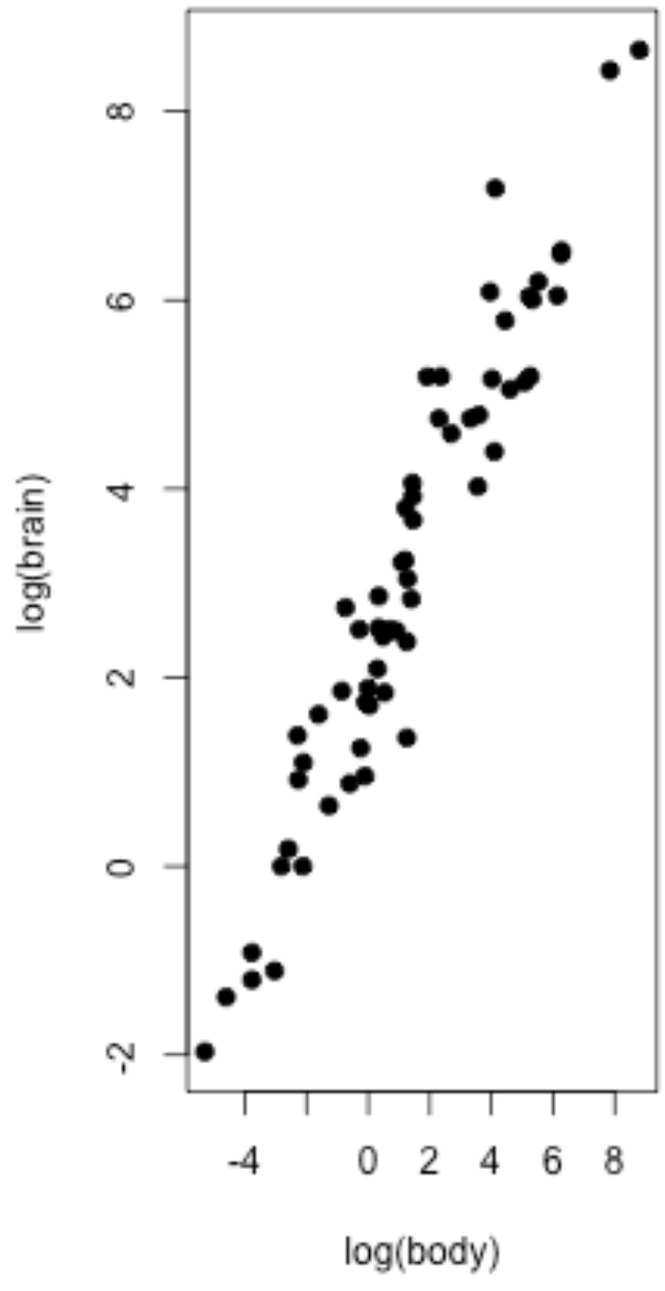




Basic numerical and graphical summaries

```
>boxplot(log(mammals),names=c(“log(body)","log(brain)"),
col=c('green','blue'))
>plot(log(mammals$body),log(mammals$brain),pch=19,
ylab="log(brain)",xlab="log(body)")
```





Basic numerical and graphical summaries

- Correlation

$$\rho_{X,Y} = \frac{\operatorname{Cov}(X,Y)}{\operatorname{SD}(X) \times \operatorname{SD}(Y)} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Sample correlation:

$$r_{XY} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

```
> cor(mammals)
             body     brain
body   1.0000000 0.9341638
brain  0.9341638 1.0000000
> cor(log(mammals))
             body     brain
body   1.0000000 0.9595748
brain  0.9595748 1.0000000
```

Basic numerical and graphical summaries

Example: analysis of bivariate data by group

Table 2.1 IQ of twins separated near birth. The data is given in three columns in the file "twinIQ.txt".

Foster	Biological	Social	Foster	Biological	Social	Foster	Biological	Social
82	82	high	71	78	middle	63	68	low
80	90	high	75	79	middle	77	73	low
88	91	high	93	82	middle	86	81	low
108	115	high	95	97	middle	83	85	low
116	115	high	88	100	middle	93	87	low
117	129	high	111	107	middle	97	87	low
132	131	high				87	93	low
						94	94	low
						96	95	low
						112	97	low
						113	97	low
						106	103	low
						107	106	low
						98	111	low

Foster: IQ for twin raised with foster parents
Biological: IQ for twin raised with biological parents
Social: Social status of biological parent

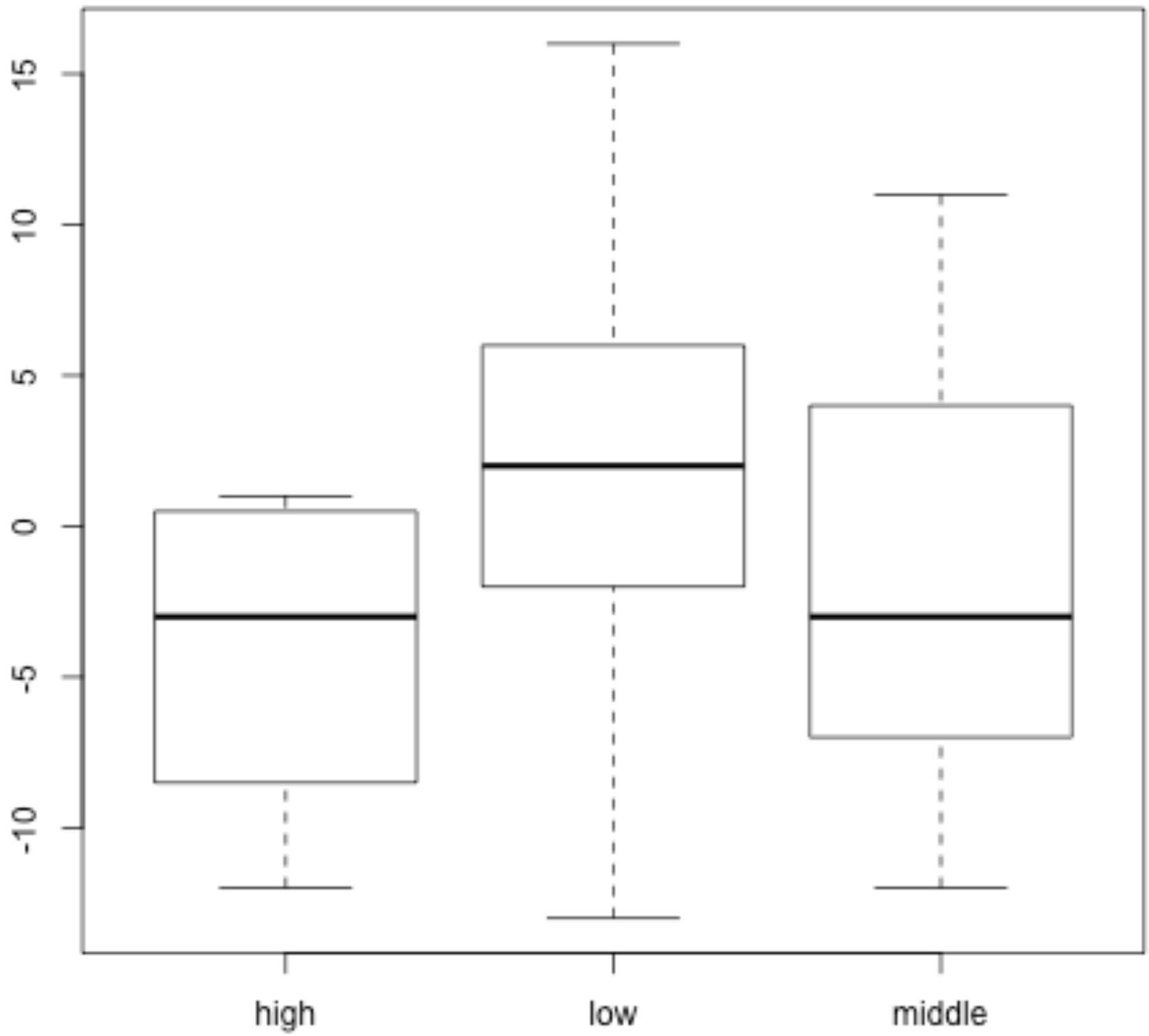
Basic numerical and graphical summaries

```
> summary(twins)
     Foster        Biological       Social
 Min.   : 63.00   Min.   : 68.0   high  : 7
 1st Qu.: 84.50   1st Qu.: 83.5   low   :14
 Median : 94.00   Median : 94.0   middle: 6
 Mean   : 95.11   Mean   : 95.3
 3rd Qu.:107.50   3rd Qu.:104.5
 Max.   :132.00   Max.   :131.0
```

- Boxplots of the difference in IQ scores by social status:

```
> boxplot(Foster - Biological ~ Social, twins)
```

Basic numerical and graphical summaries

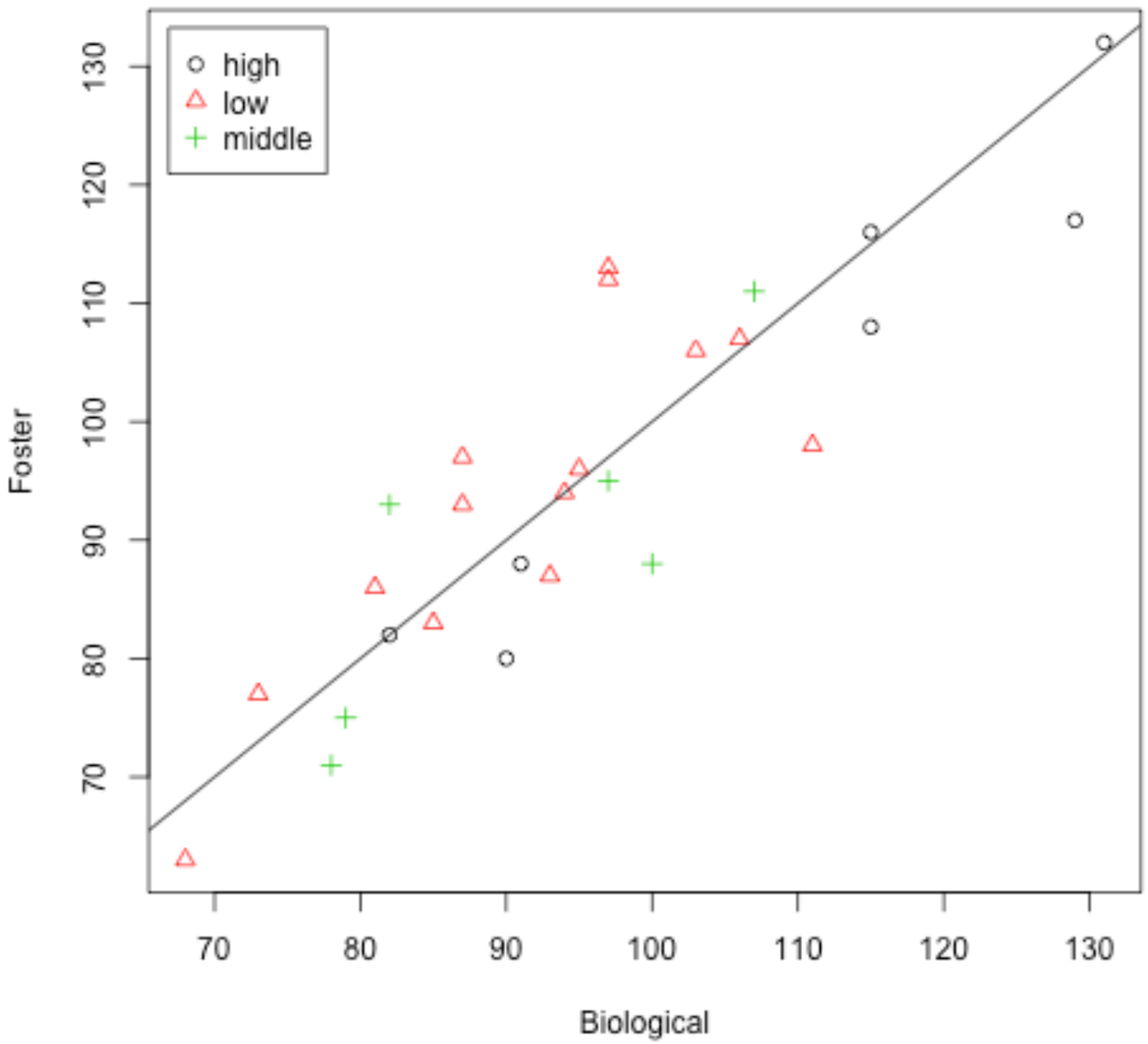


Basic numerical and graphical summaries

- Another way to display these data is using a scatterplot:

```
> plot(Foster ~ Biological, data=twins,pch=status,col=status)
> legend("topleft", c("high","low","middle"), pch=1:3,
col=1:3,inset=.02)
> abline(0,1)
```

no dramatic differences, but high social status may correspond to higher IQ scores for twins with biological parents



Basic numerical and graphical summaries

- **Conditional plots:** the function coplot displays several plots on the same scale. Syntax: `y ~ x|a` means plots of y vs x are conditional on a

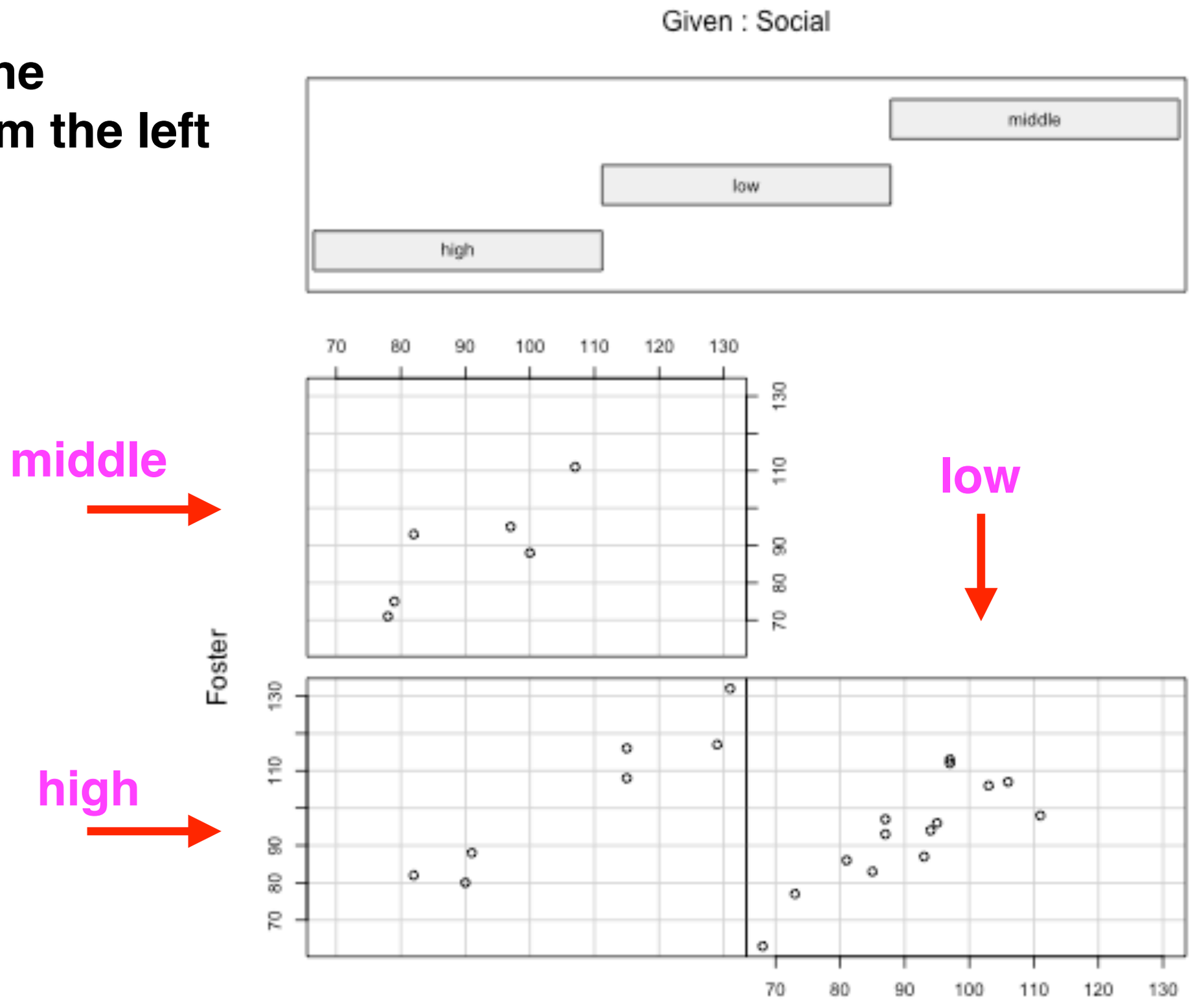
```
>coplot(Foster ~ Biological|Social, data=twins)
```

- The library `lattice` also allows us to do conditional plots (nicer because one does not have to remember the order…)

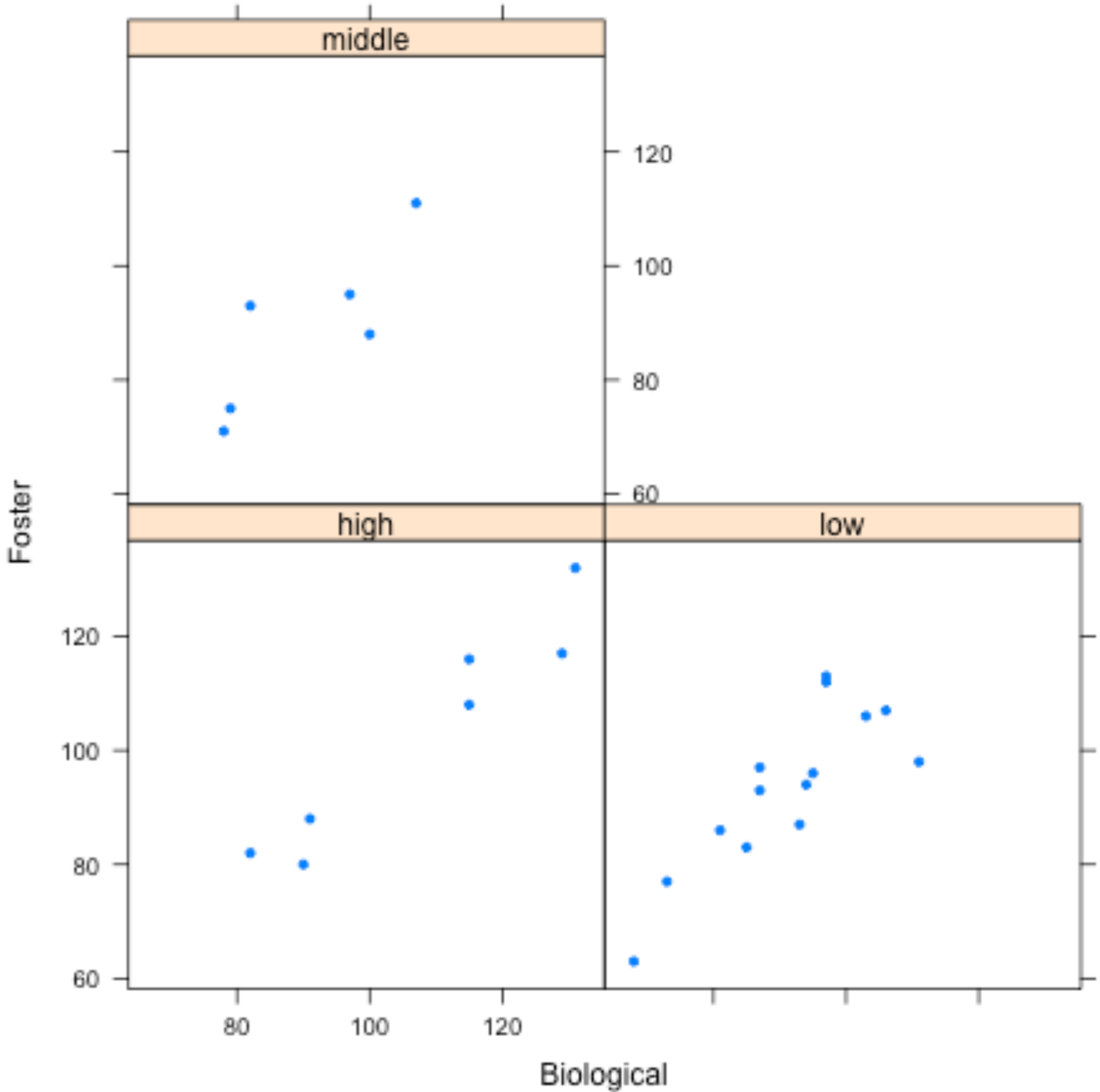
```
> xyplot(Foster~Biological|Social, data=twins)
```

Basic numerical and graphical summaries

Order is from the bottom and from the left



Basic numerical and graphical summaries



Basic numerical and graphical summaries

Multivariate data: Several quantitative variables

Example: data from a study comparing brain size and intelligence; 40 individuals; brain size measured by MRI; 8 variables:

Variable	Description
Gender	Male or Female
FSIQ	Full Scale IQ scores based on four Wechsler (1981) subtests
VIQ	Verbal IQ scores based on four Wechsler (1981) subtests
PIQ	Performance IQ scores based on four Wechsler (1981) subtests
Weight	Body weight in pounds
Height	Height in inches
MRI_Count	total pixel Count from the 18 MRI scans

Basic numerical and graphical summaries

```
> summary(brain)
    Gender          FSIQ            VIQ            PIQ          Weight     
 Female:20   Min.   : 77.00   Min.   : 71.0   Min.   : 72.00   Min.   :106.0  
 Male  :20   1st Qu.: 89.75   1st Qu.: 90.0   1st Qu.: 88.25   1st Qu.:135.2  
             Median :116.50   Median :113.0   Median :115.00   Median :146.5  
             Mean   :113.45   Mean   :112.3   Mean   :111.03   Mean   :151.1  
             3rd Qu.:135.50   3rd Qu.:129.8   3rd Qu.:128.00   3rd Qu.:172.0  
             Max.   :144.00   Max.   :150.0   Max.   :150.00   Max.   :192.0  
                                                               NA's   :2      
     Height         MRI_Count        
 Min.   :62.00   Min.   : 790619  
 1st Qu.:66.00   1st Qu.: 855918  
 Median :68.00   Median : 905399  
 Mean   :68.53   Mean   : 908755  
 3rd Qu.:70.50   3rd Qu.: 950078  
 Max.   :77.00   Max.   :1079549  
 NA's   :1
```

table

missing values

missing values

Basic numerical and graphical summaries

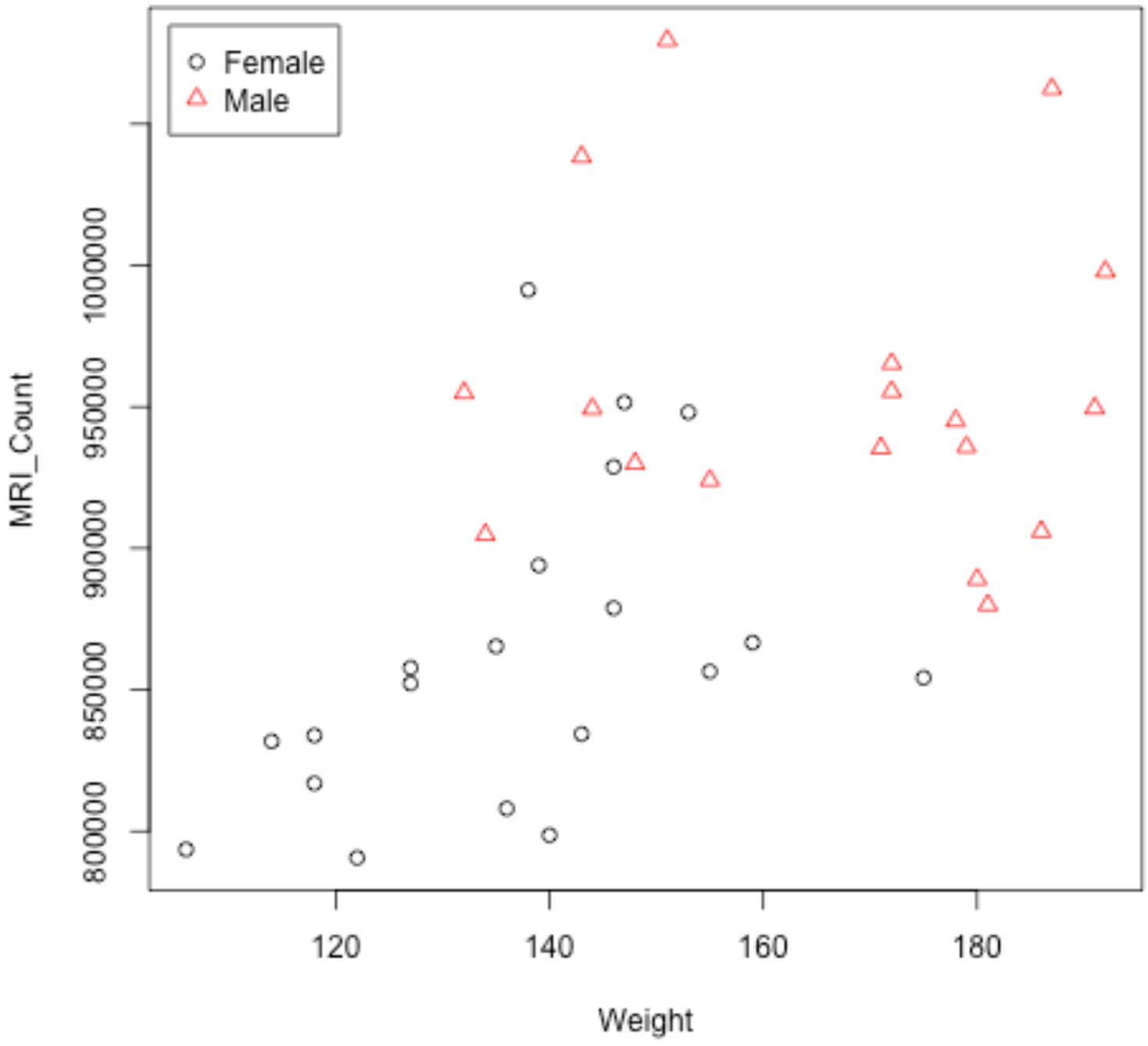
- Missing values:

```
> mean(brain$Weight)
[1] NA
> mean(brain$Weight, na.rm=TRUE)
[1] 151.0526
```

- The function by allows us to consider summaries by group:

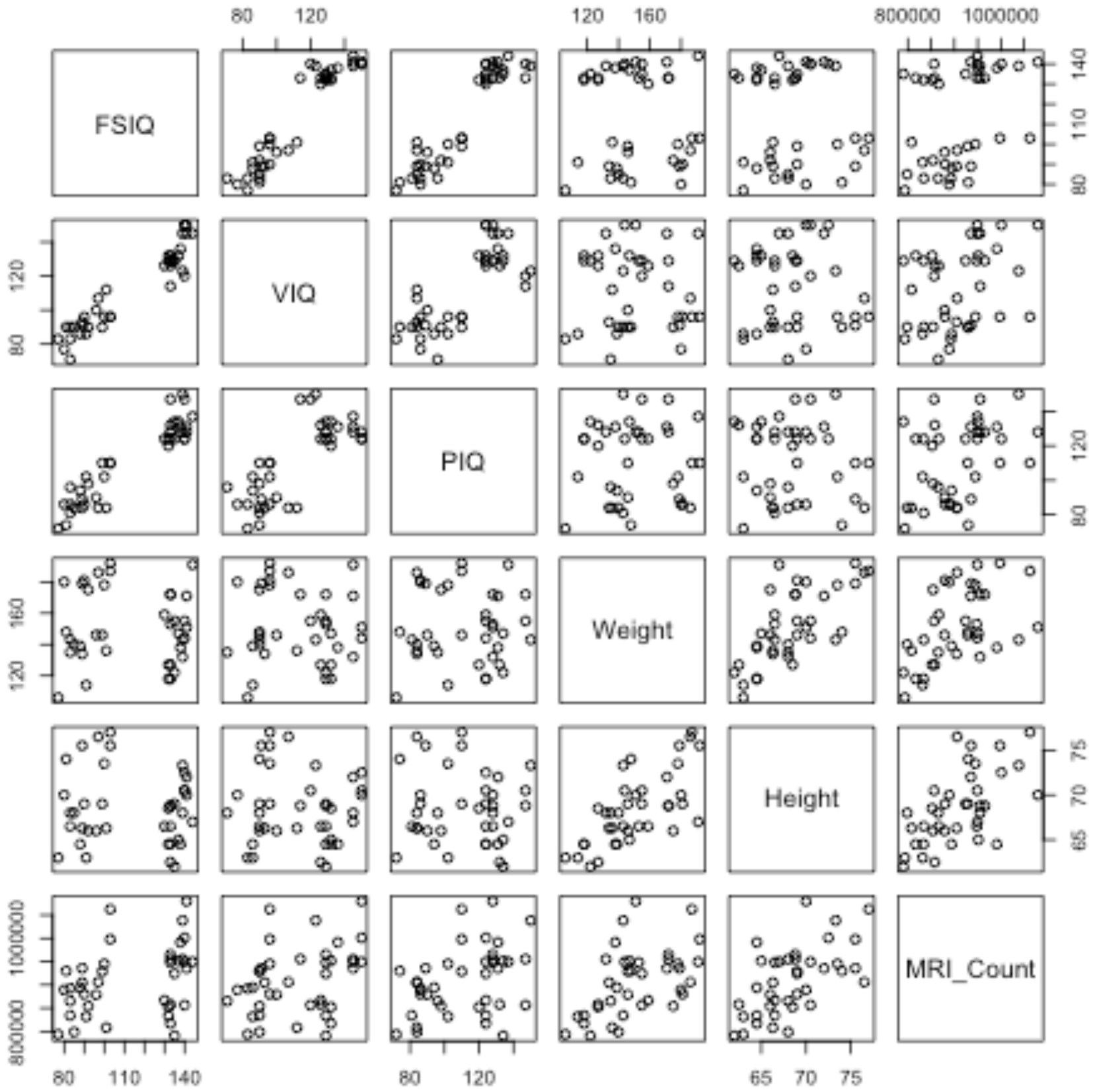
```
> by(data=brain[, 2], INDICES=brain$Gender,
FUN=mean, na.rm=TRUE)
brain$Gender: Female
[1] 111.9
------------------------------------------------
brain$Gender: Male
[1] 115
```

Basic numerical and graphical summaries



- The function pairs can be used to display scatterplots for all pairs of variables…

Basic numerical and graphical summaries



Basic numerical and graphical summaries

- The variables FSIQ, VIQ and PIQ have positive correlation:

```
> round(cor(brain[, 2:7]), 2)
             FSIQ  VIQ  PIQ Weight Height MRI_Count
FSIQ         1.00 0.95 0.93     NA     NA       0.36
VIQ          0.95 1.00 0.78     NA     NA       0.34
PIQ          0.93 0.78 1.00     NA     NA       0.39
Weight         NA   NA   NA      1     NA         NA
Height         NA   NA   NA     NA      1         NA
MRI_Count    0.36 0.34 0.39     NA     NA       1.00
> round(cor(brain[, 2:7], use="pairwise.complete.obs"), 2)
             FSIQ   VIQ   PIQ Weight Height MRI_Count
FSIQ         1.00  0.95  0.93  -0.05  -0.09      0.36
VIQ          0.95  1.00  0.78  -0.08  -0.07      0.34
PIQ          0.93  0.78  1.00   0.00  -0.08      0.39
Weight      -0.05 -0.08  0.00   1.00   0.70      0.51
Height      -0.09 -0.07 -0.08   0.70   1.00      0.60
MRI_Count    0.36  0.34  0.39   0.51   0.60      1.00
```

Basic numerical and graphical summaries

- Controlling for body size (measured as weight):

```
> mri = MRI_Count / Weight
> cor(FSIQ, mri, use="pairwise.complete.obs")
[1] 0.235308
```

- Identifying missing data:

```
 > which(is.na(brain), arr.ind=TRUE)
      row col
[1,]   2   5
[2,]  21   5
[3,]  21   6
> brain[21,]
   Gender FSIQ VIQ PIQ Weight Height MRI_Count
21   Male   83  83  86     NA     NA    892420
```

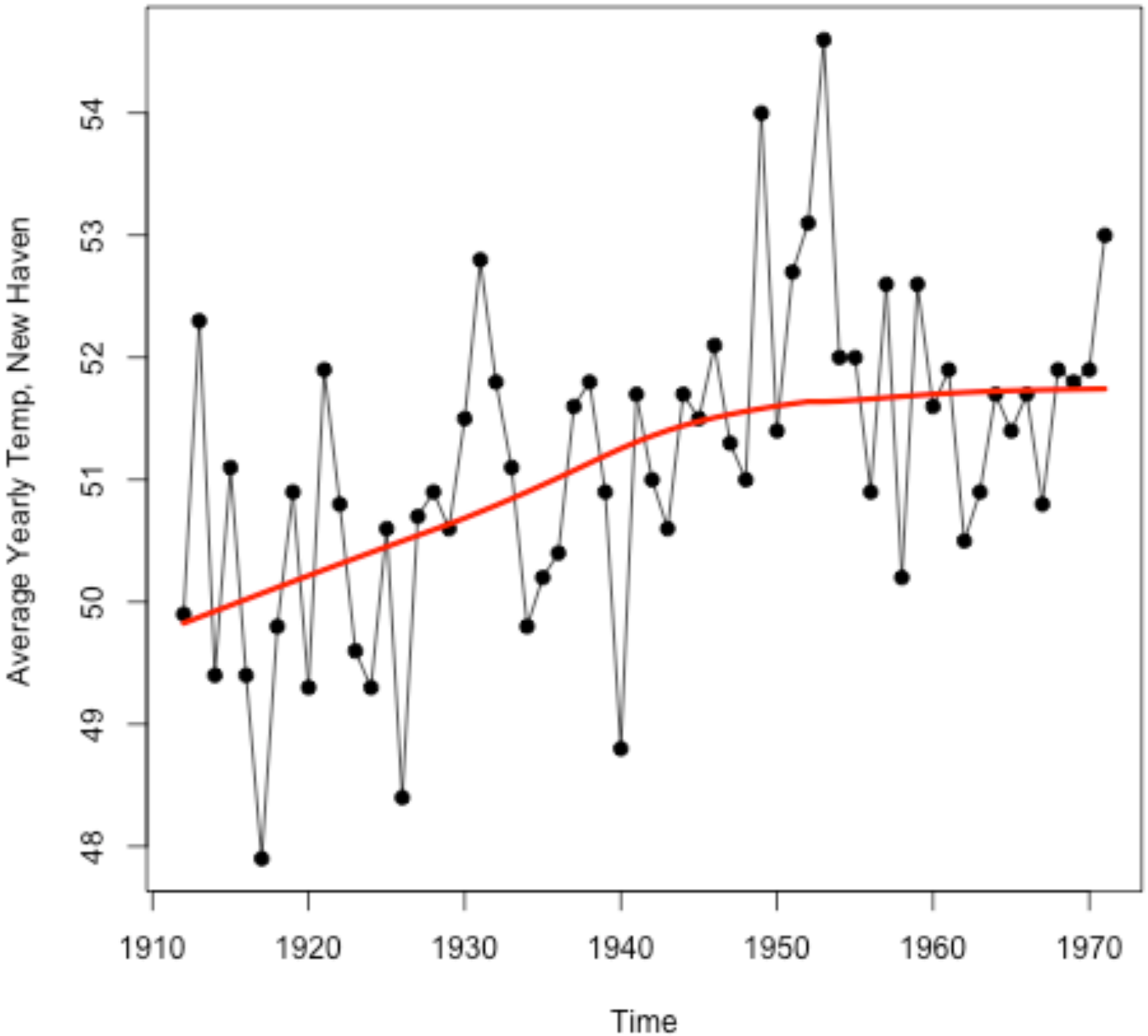
Basic numerical and graphical summaries

- Time series data: Average yearly temperatures for New Haven from 1912-1971:

```
>nhtemp
Time Series:
Start = 1912
End = 1971
Frequency = 1
 [1] 49.9 52.3 49.4 51.1 49.4 47.9 49.8 50.9 49.3 51.9
50.8 49.6 49.3 50.6 48.4 50.7 50.9 50.6
[19] 51.5 52.8 51.8 51.1 49.8 50.2 50.4 51.6 51.8 50.9
48.8 51.7 51.0 50.6 51.7 51.5 52.1 51.3
[37] 51.0 54.0 51.4 52.7 53.1 54.6 52.0 52.0 50.9 52.6
50.2 52.6 51.6 51.9 50.5 50.9 51.7 51.4
[55] 51.7 50.8 51.9 51.8 51.9 53.0

>plot(nhtemp)
```

Basic numerical and graphical summaries

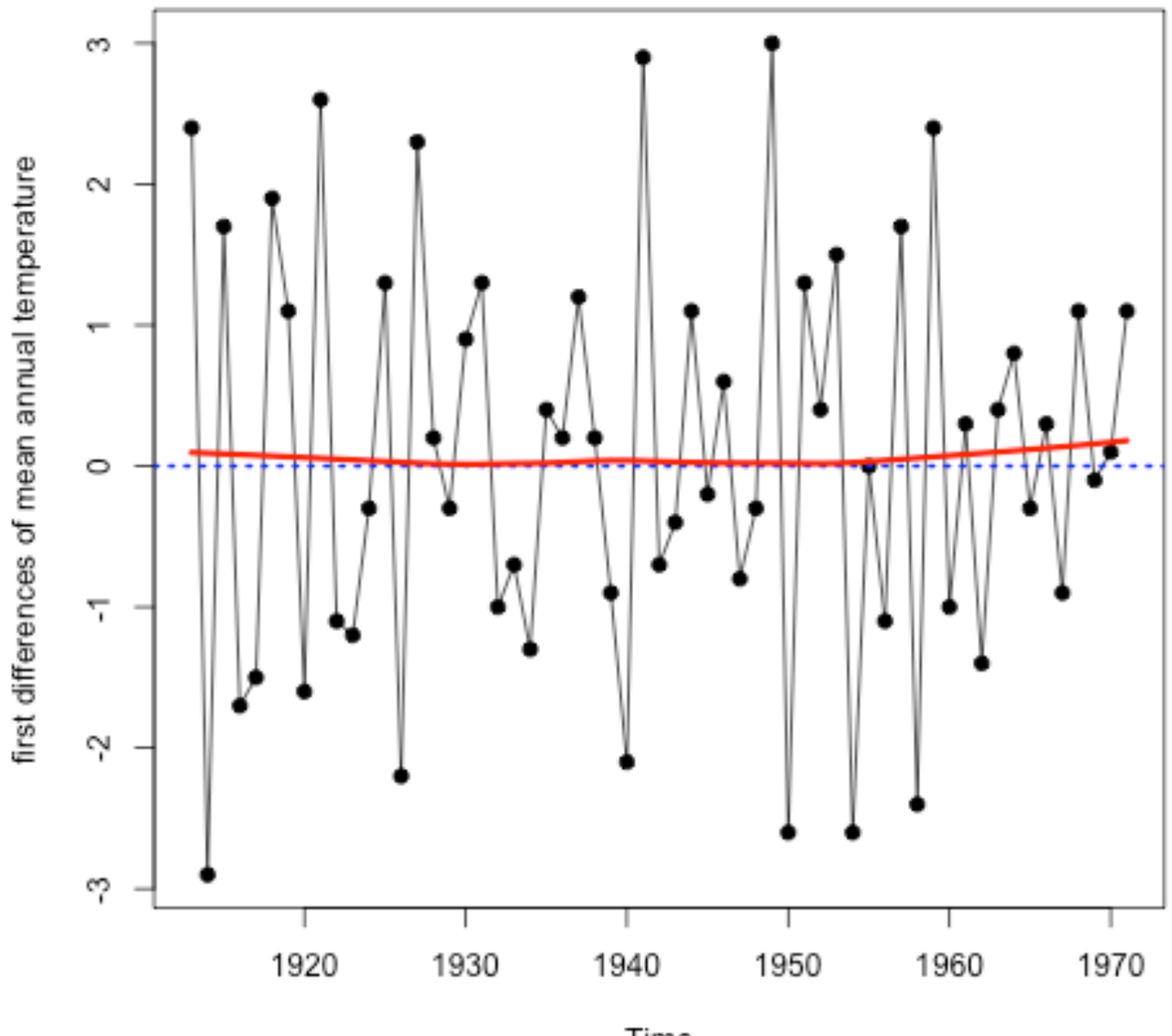


- Is there a trend?
- Seasonal pattern?
- Extreme observations?

Basic numerical and graphical summaries

- First differences:

```
> d=diff(nhtemp)
> plot(d,ylab="first differences of mean annual temperature")
> points(d,pch=19)
> abline(h=0,lty=3,lwd=2,col='blue')
> lines(lowess(d),lwd=3,col='red')
```



Basic numerical and graphical summaries

The Central Limit Theorem

The data frame randu contains 400 triples of successive random numbers that were generated using the Fortran function RANDU. If the numbers are truly from a U(0,1) their expected value should be 1/2 and their variance 1/12.

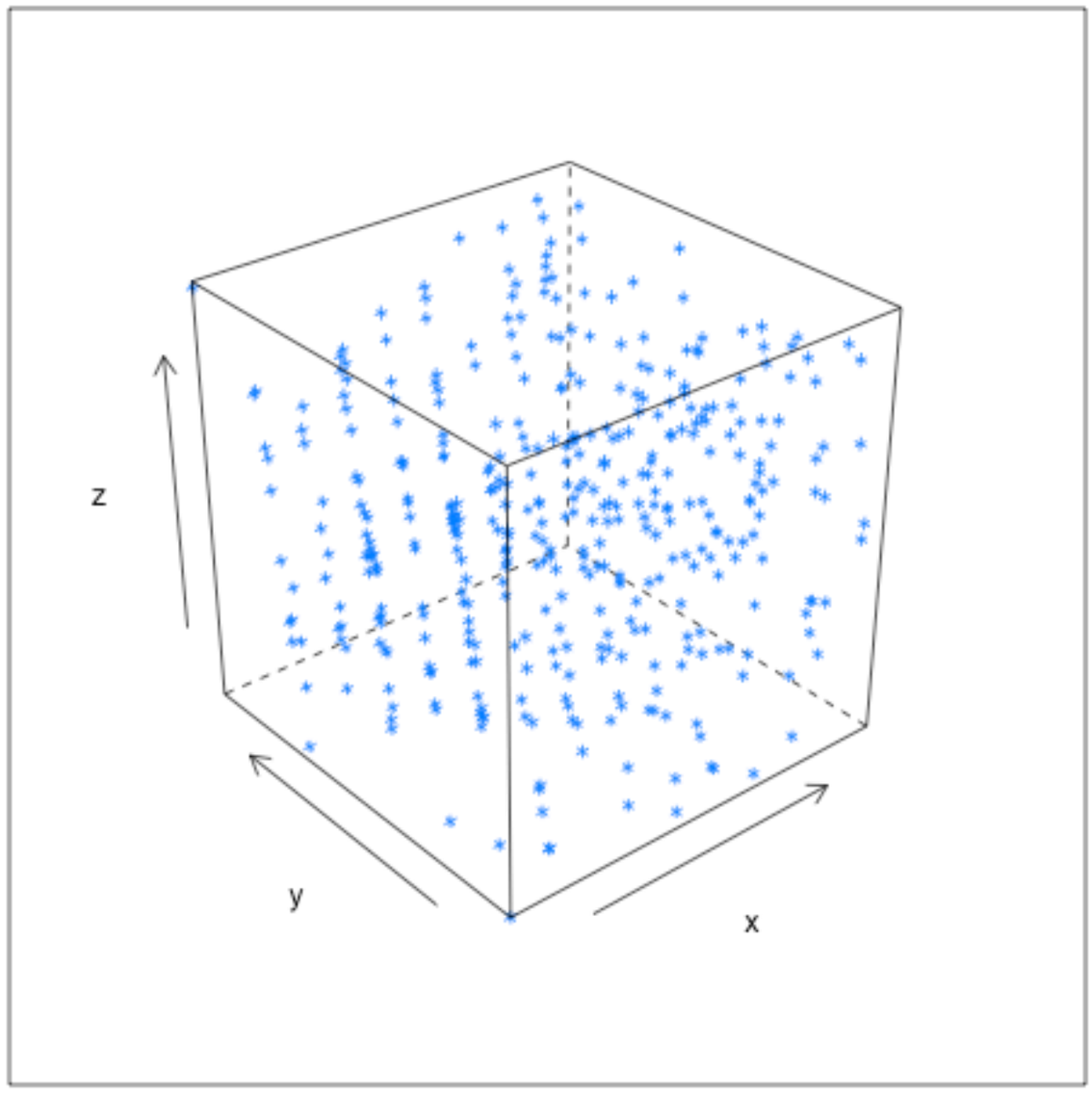
```
> apply(randu,2,mean)
        x         y         z
0.5264293 0.4860531 0.4809547

> apply(randu,2,var)
         x          y          z
0.08123189 0.08627021 0.07786043
> 1/12
[1] 0.08333333
```

Basic numerical and graphical summaries

```
> cor(randu)
             x           y           z
x  1.00000000 -0.04847127  0.05831454
y -0.04847127  1.00000000 -0.06281830
z  0.05831454 -0.06281830  1.00000000
```

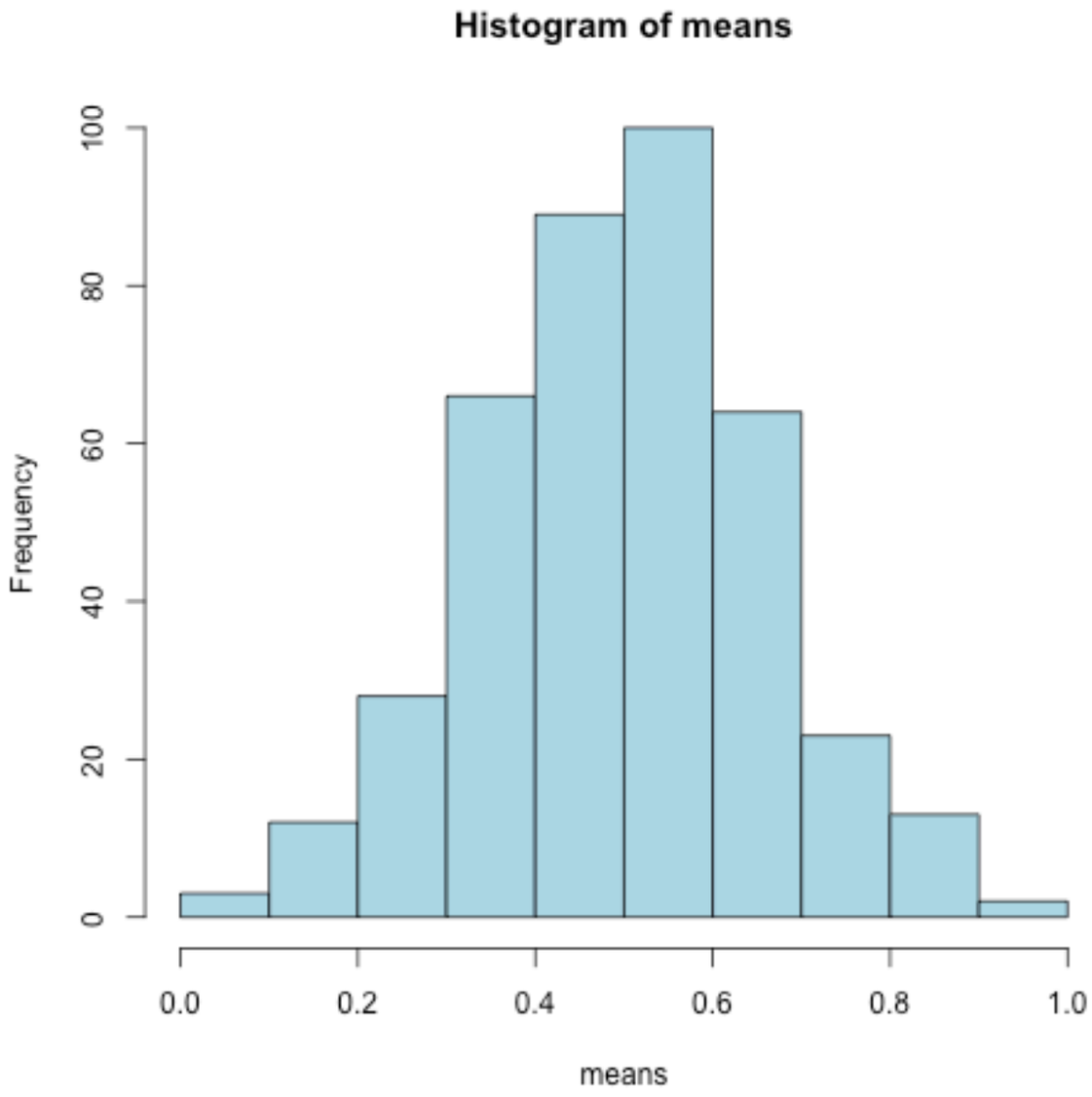
```
>library(lattice)
>cloud(z~x+y,data=randu)
```



Basic numerical and graphical summaries

- Each row is assumed to be a sample of size 3 from U(0,1):

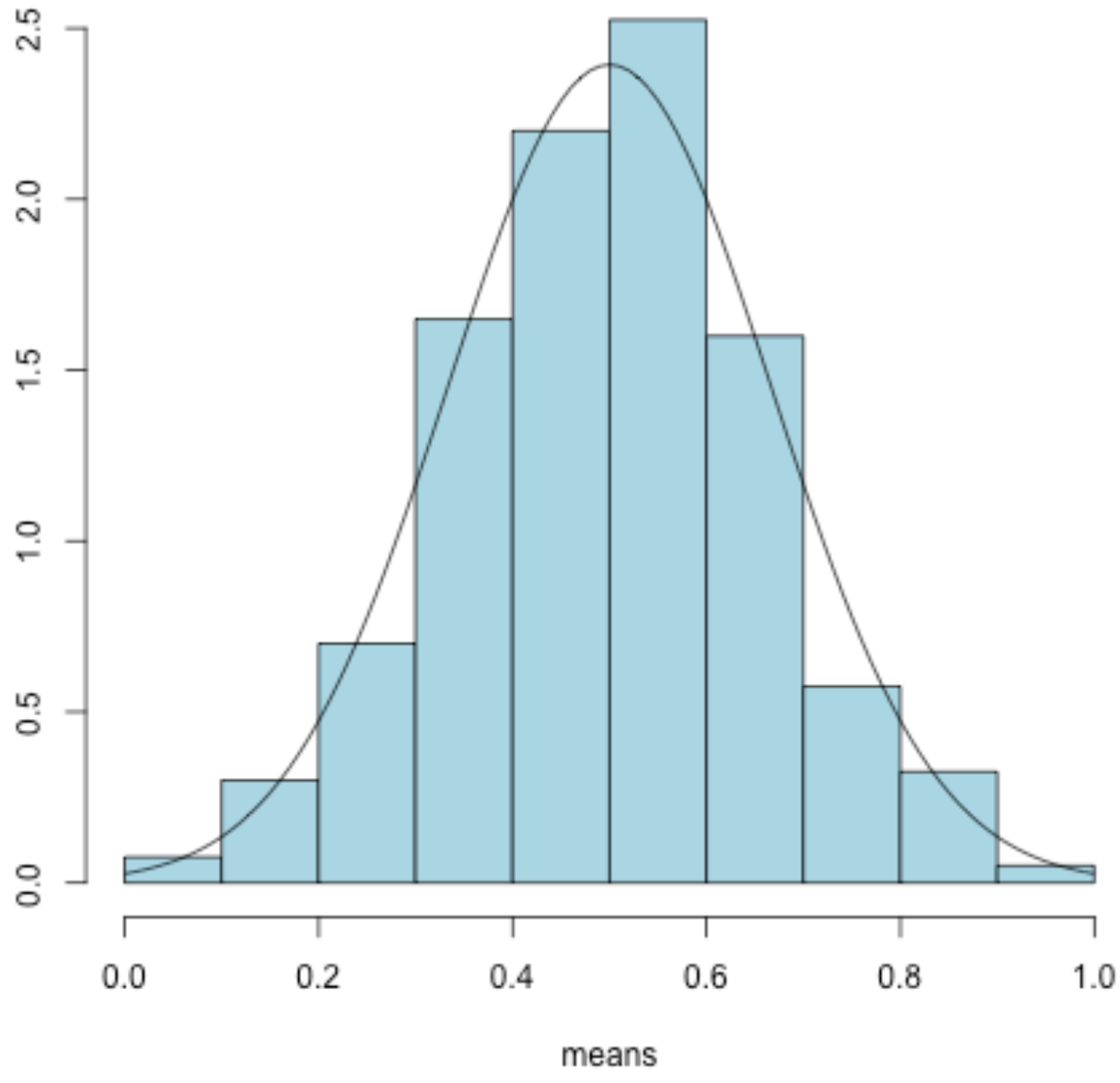
```
>means=apply(randu,1,mean)
>hist(means,col='lightblue')
```



How does this distribution compare to the N(1/2, 1/36)?

Basic numerical and graphical summaries

```
>truehist(means,col='lightblue')
>curve(dnorm(x,1/2,sd=sqrt(1/36)),add=TRUE)
```



Basic numerical and graphical summaries

```
> qqnorm(means)
> qqline(means)
```