DATA 500
Lecture 5

Shape of continuous distributions

Normal or Skewed?

Kurtosis

Histograms and Density Plots in R

Boxplots

Shape of distributions

Measure just about anything...

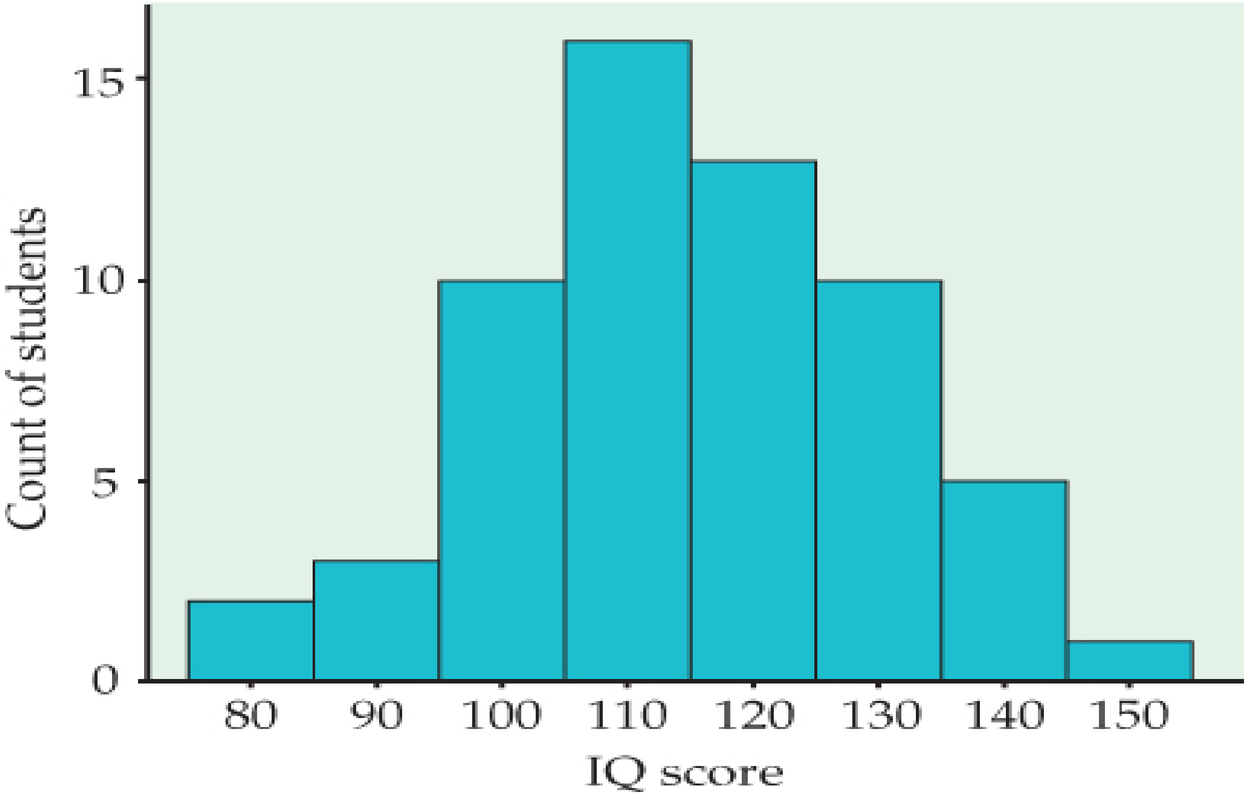
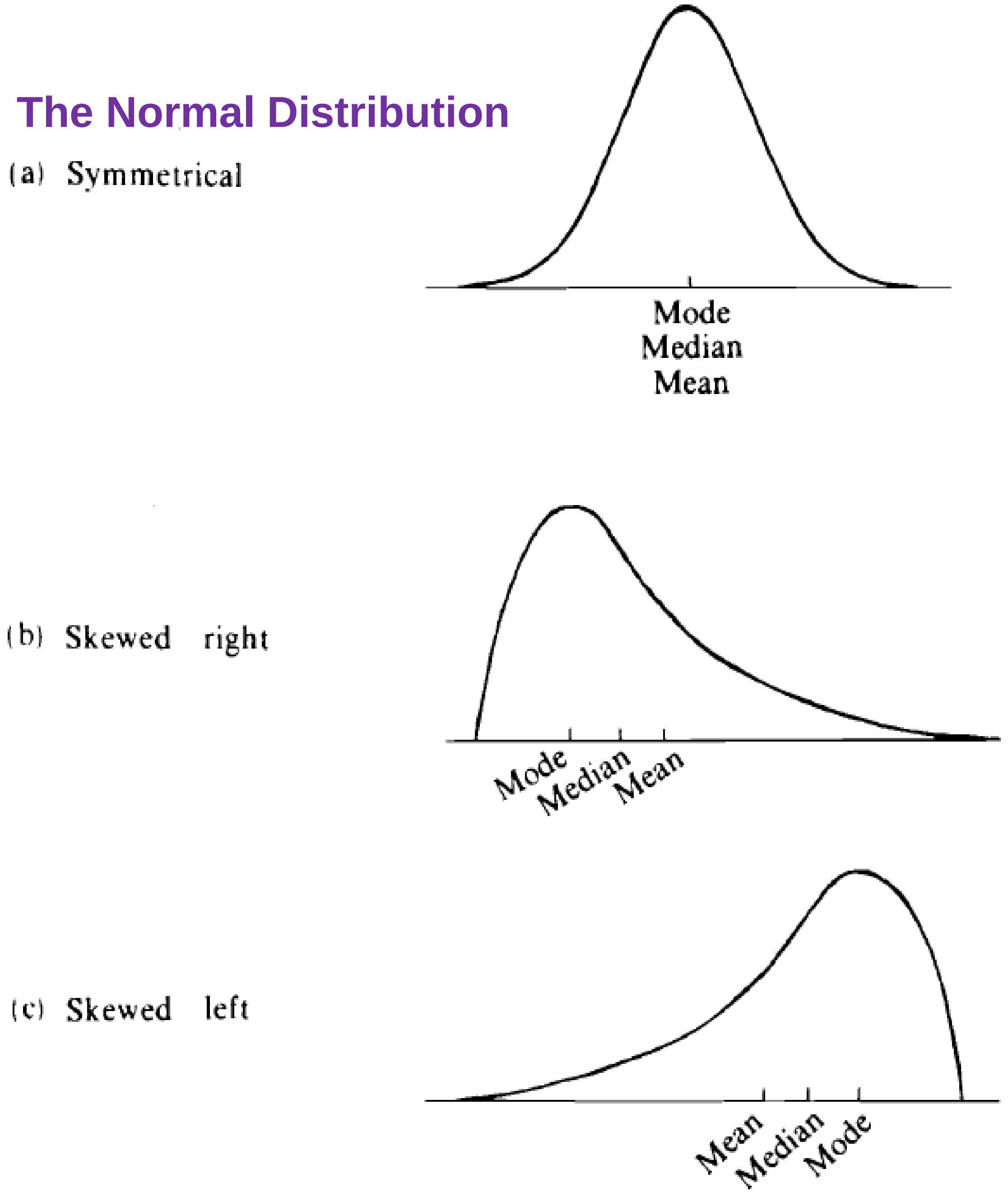


Figure 1.9, INTRODUCTION to the PRACTICE of STATISTICS, © 2014 W. H. Freeman

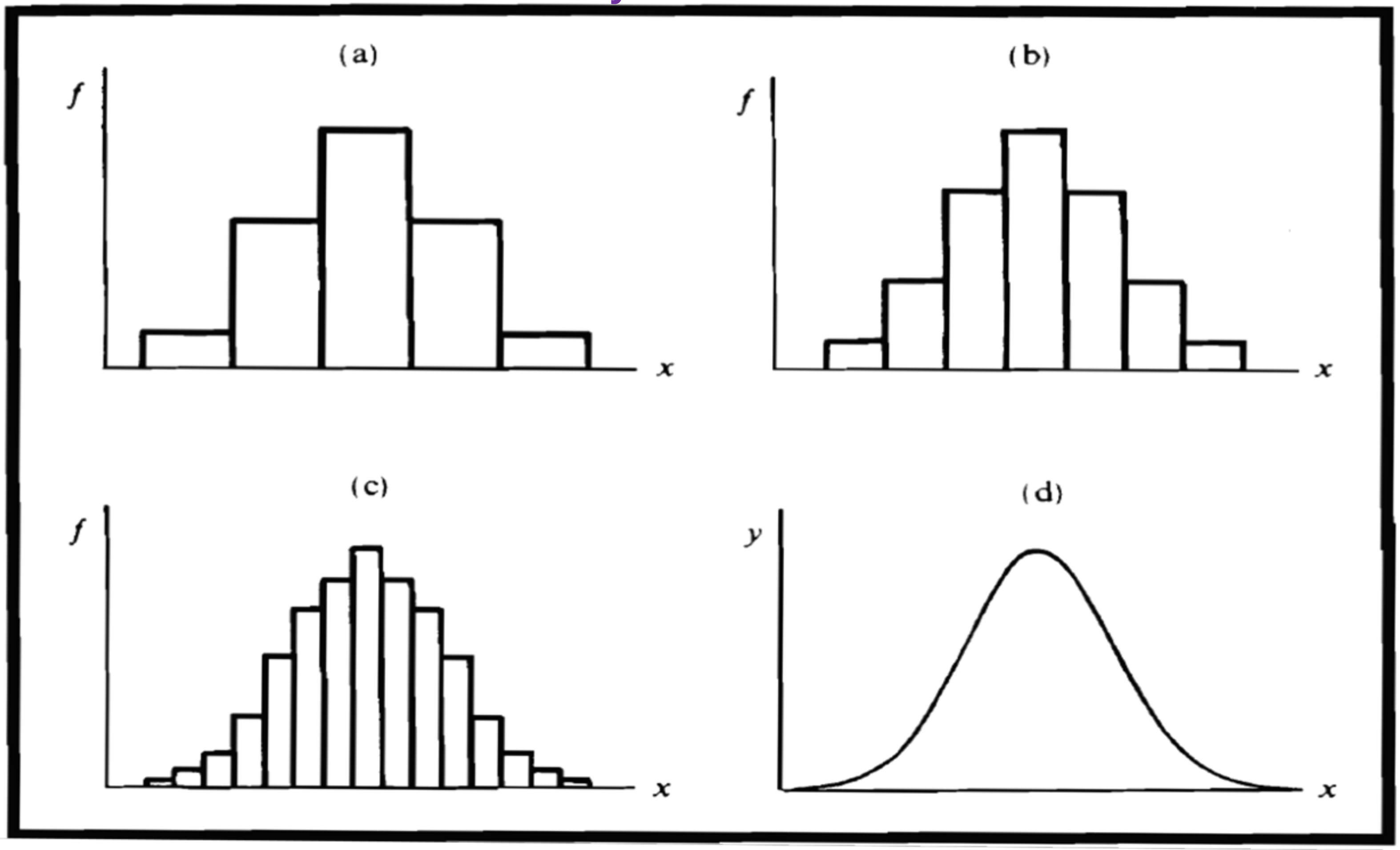
The Normal Distribution



MacDonnell Finger Lengths

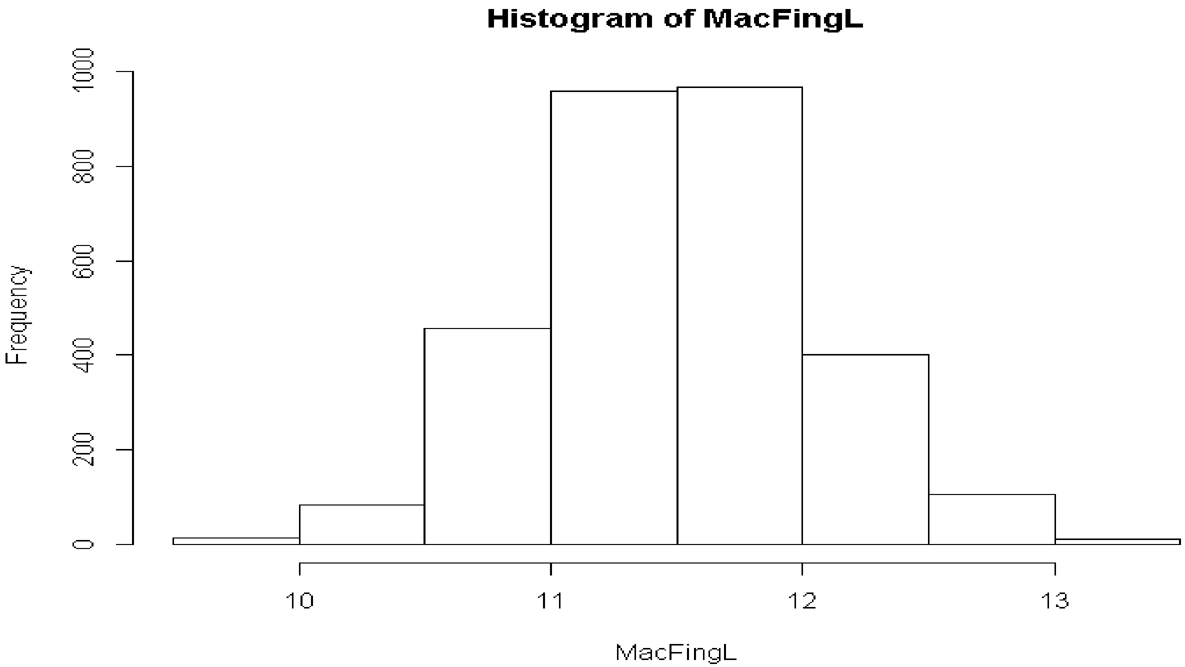
```
# Look at MacDonell finger length data, finger lengths from 3,000 men;
# scan is a special function for a vector;

MacFingL <- scan("http://math.mercyhurst.edu/~sousley/STAT_139/data/MacFingL.vec");


# what is the structure of the data?;
str(MacFingL);
num [1:3000] 10 10.3 9.9 10.2 10.2 10.3 10.4 10.7 10 10.1 ...

# get the first few records;
head(MacFingL);
[1] 10.0 10.3 9.9 10.2 10.2 10.3


# list the data;
MacFingL
...
[2863] 12.0 12.1 12.1 12.1 12.1 12.1 12.1 12.1 12.1 12.1 12.1 12.2 12.2 12.2 12.2 12.2 12.2 12.2
[2881] 12.2 12.3 12.3 12.3 12.3 12.3 12.3 12.4 12.4 12.4 12.4 12.4 12.4 12.4 12.5 12.5 12.5 12.5
[2899] 12.5 12.5 12.6 12.6 12.6 12.6 12.6 12.6 12.6 12.6 12.7 12.7 12.7 12.7 12.7 12.7 12.7 12.7
[2917] 12.8 12.8 12.8 12.8 12.8 12.8 12.8 12.8 13.0 11.6 11.6 11.7 11.7 11.8 11.9 11.9 11.9 12.0
[2935] 12.0 12.1 12.1 12.1 12.1 12.2 12.3 12.3 12.3 12.3 12.3 12.4 12.5 12.5 12.5 12.5 12.5 12.5
[2953] 12.5 12.5 12.6 12.6 12.6 12.6 12.6 12.6 12.7 12.7 12.8 12.8 12.8 12.8 12.8 12.9 13.2 13.2
[2971] 13.2 11.9 12.0 12.0 12.1 12.2 12.6 12.6 12.6 12.7 12.7 12.8 12.8 12.8 12.9 13.0 13.0 13.3
[2989] 11.6 12.5 12.5 12.6 12.8 13.0 13.5 12.4 12.6 12.8 13.3 11.2
```

Normally distributed data



with different bin widths

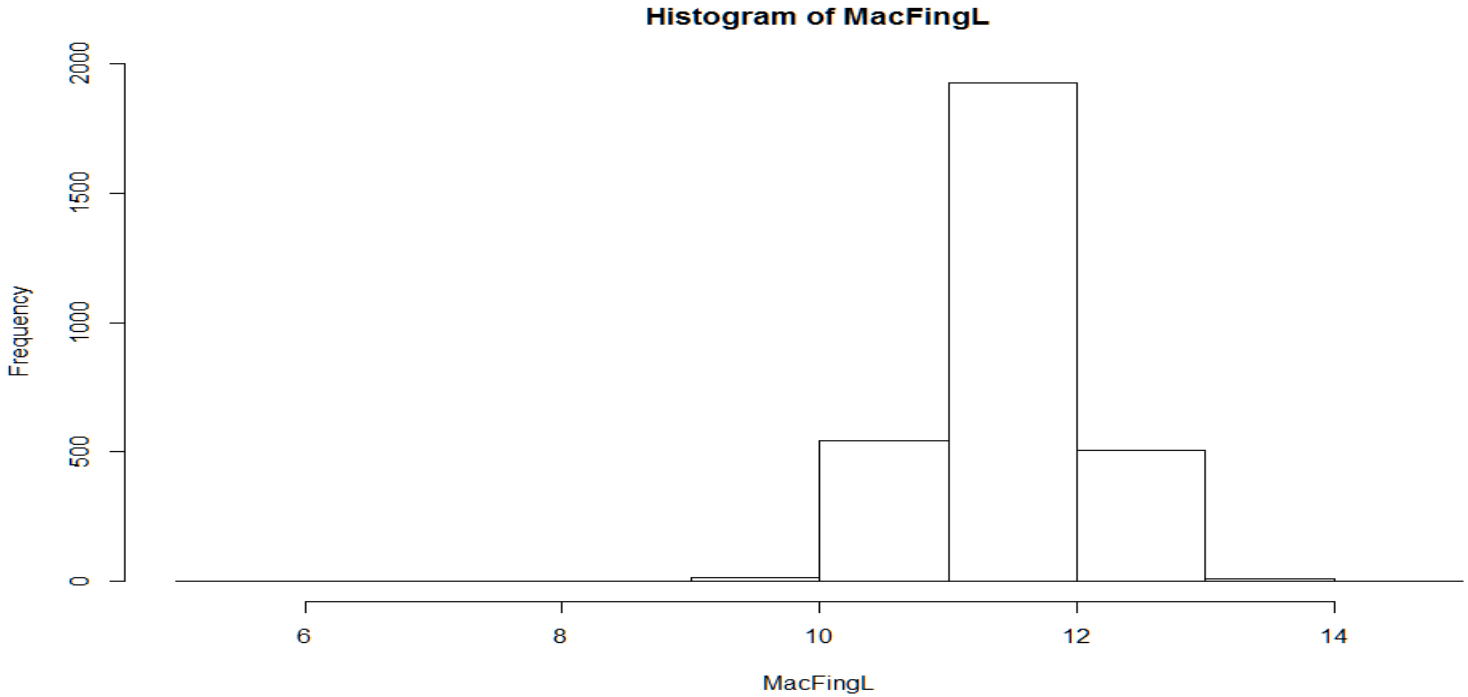
The default bins and bin width look good



`hist(MacFingL)`

`To change, use`
`hist(MacFingL, sequence of bin intervals such as 5,6,7,8,9,10,11,12,13,14,15)`

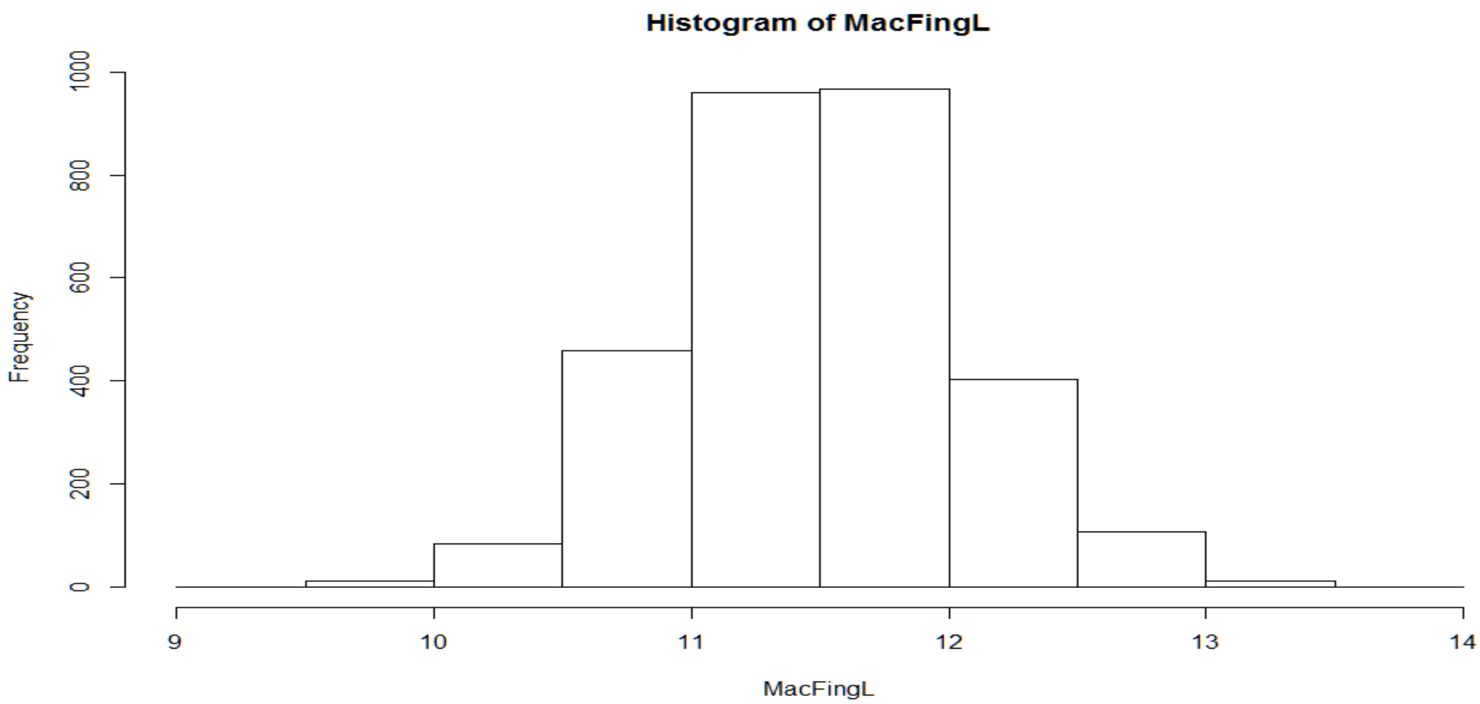
Changing the bins: the `breaks` argument




```
hist(MacFingL, breaks = c(5,6,7,8,9,10,11,12,13,14,15) )
```

```
Another way of getting a sequence: seq(min, max, increment)
seq(5,15,1)

hist(MacFingL, seq(5,15,1) )
```

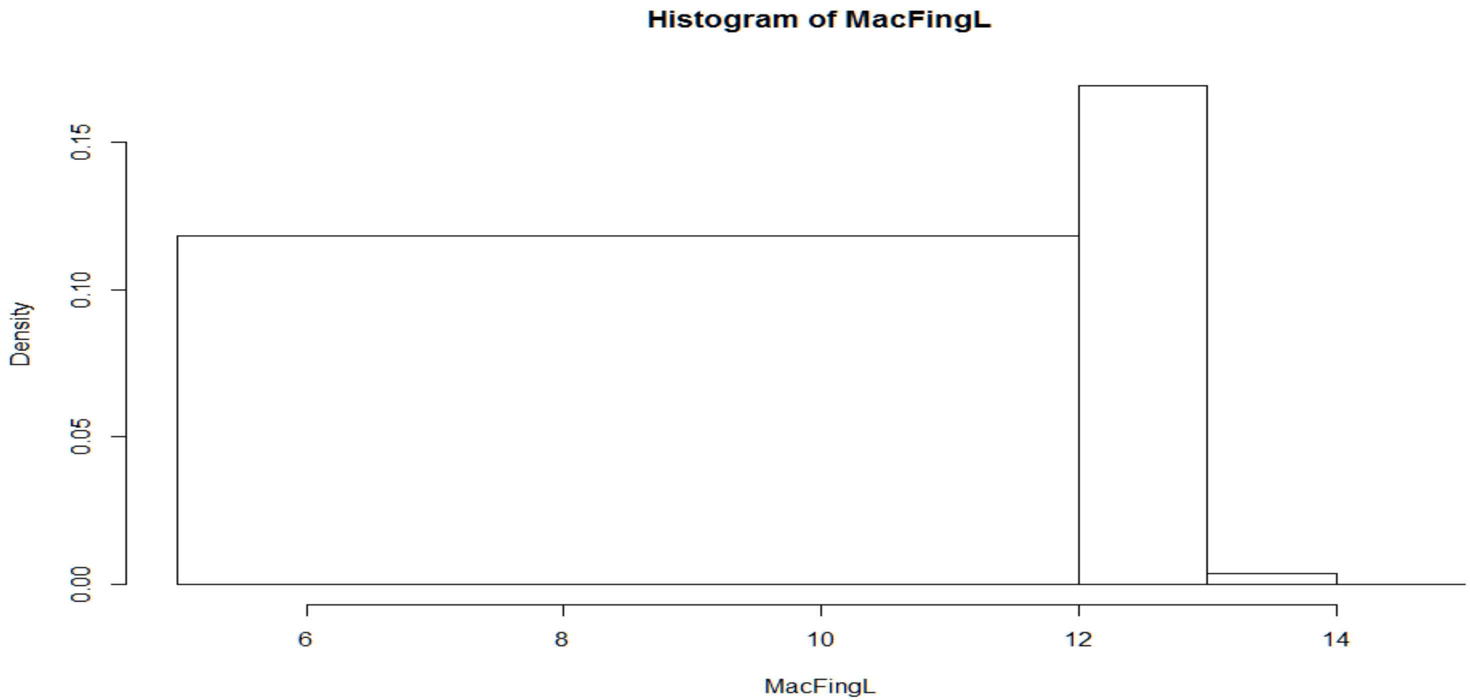
Changing the bins



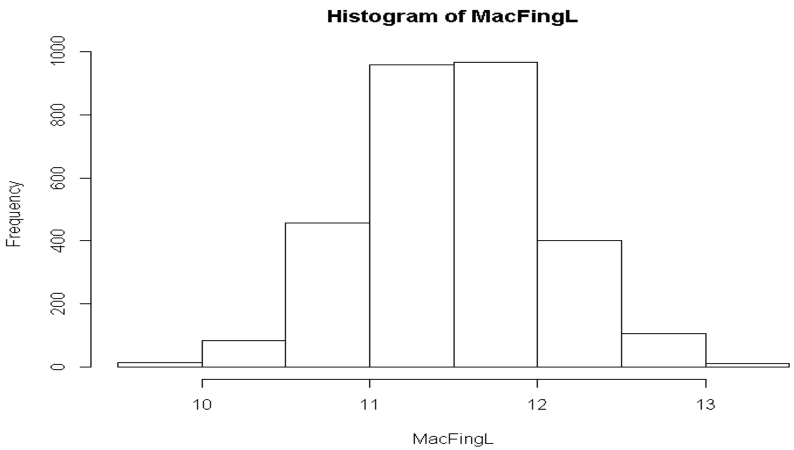

```
# Default bin width in this case = 0.5?
# use seq() function
hist(MacFingL, breaks = seq(9,14,0.5) )
```

Another way of getting a sequence: seq(min, max, increment)
seq(5,15,2)

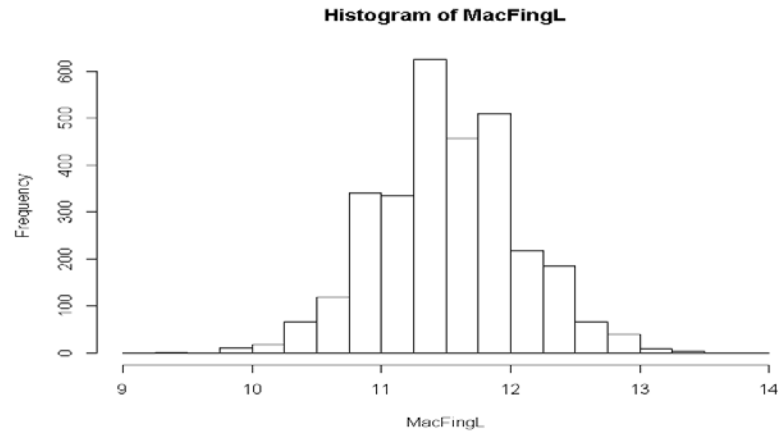
Make a funky histogram



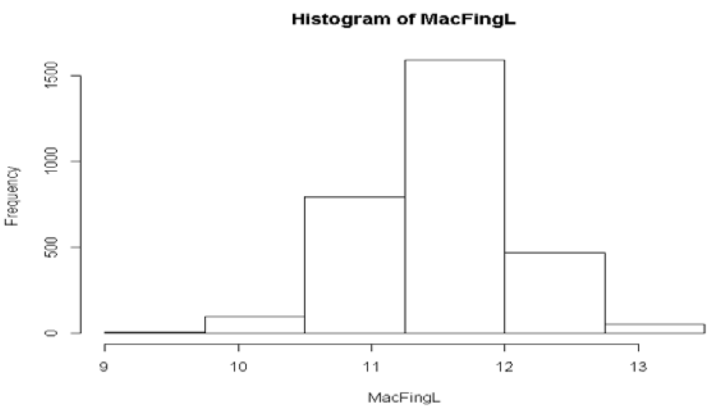

```
# Use unequal bin widths
hist(MacFingL, breaks = c(5,12,13,14,15) )
```



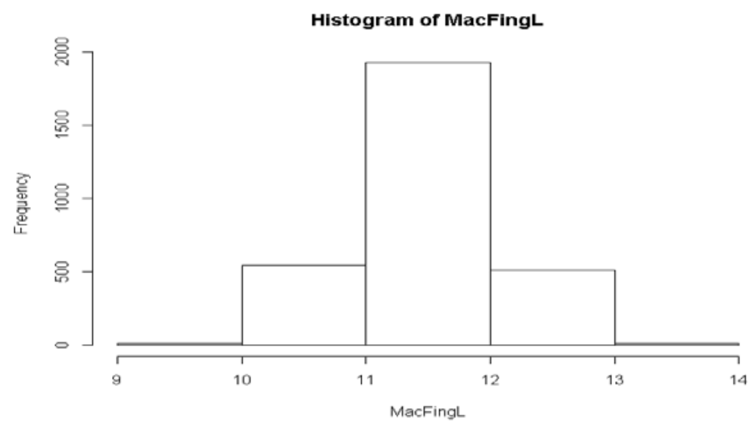

```
hist(MacFingL)
```



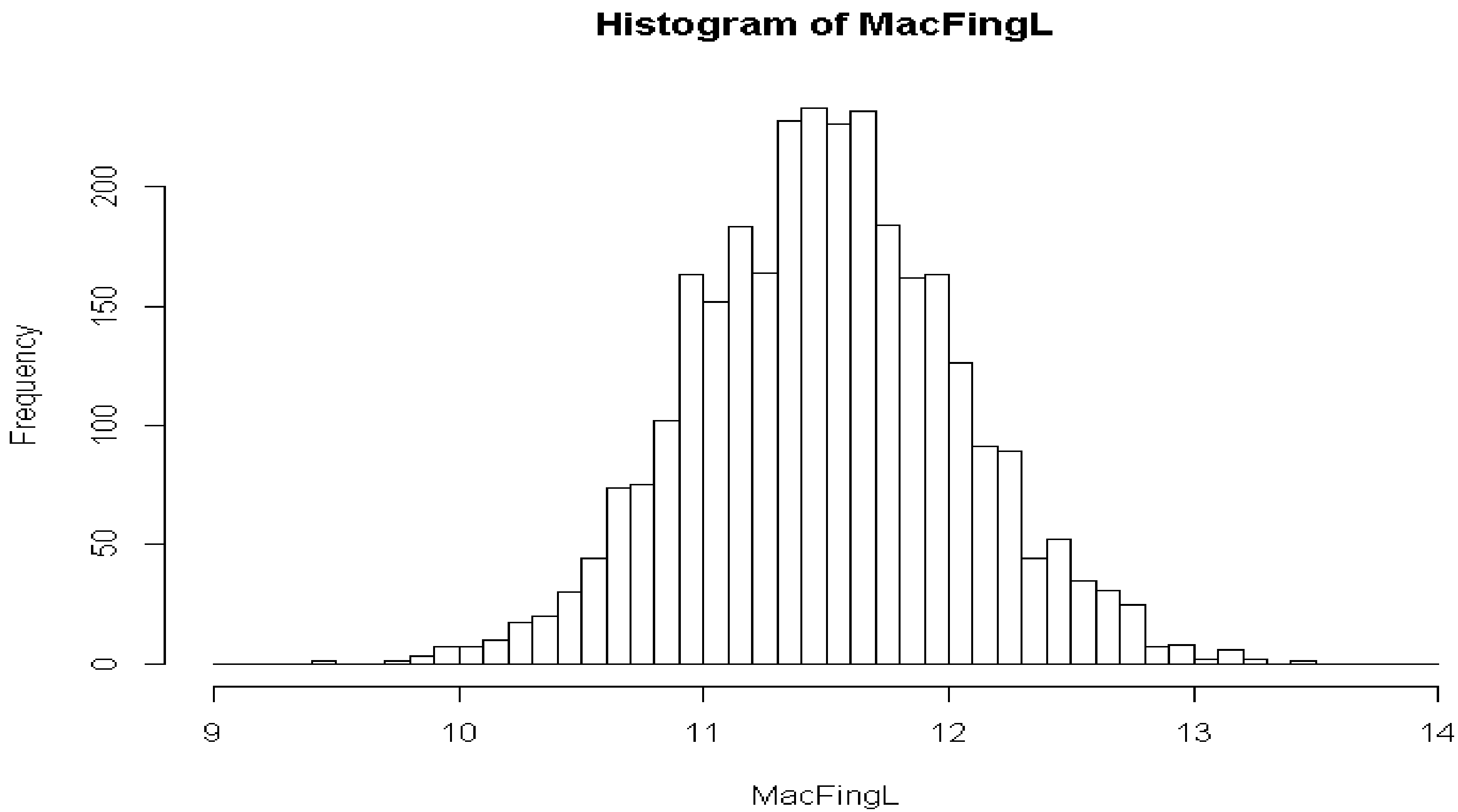

```
# breaks = is optional
hist(MacFingL, seq(9,14,0.25))
```




```
hist(MacFingL, seq(9,14,0.75))
```




```
hist(MacFingL, seq(9,14,1))
```



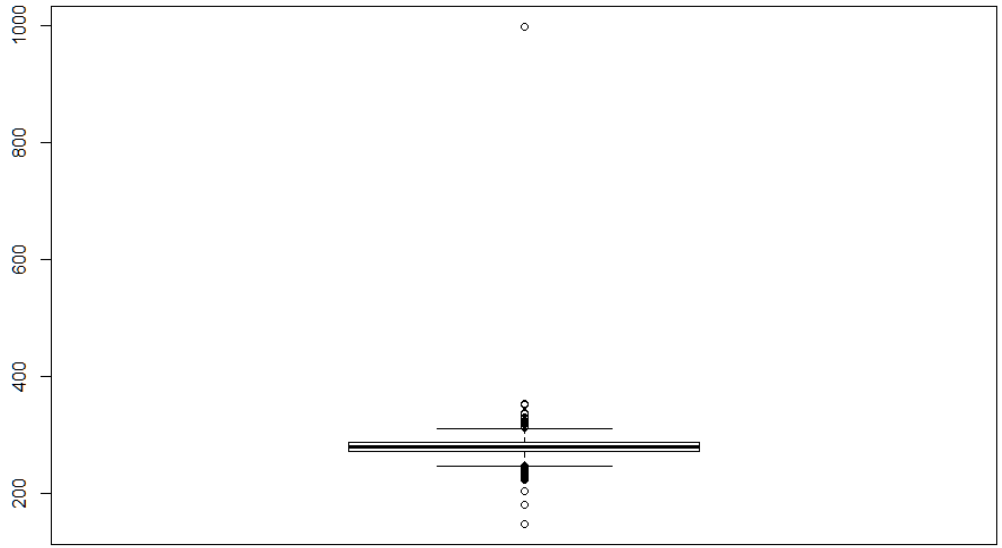
hist(MacFingL, seq(9,14,0.1))

babies: gestation

```
babies <- read.csv("http://math.mercyhurst.edu/~sousley/STAT_139/data/babies.csv", header=T);
```

We can do plots of gestation data: boxplot

```
boxplot(babies$gestdays);
```

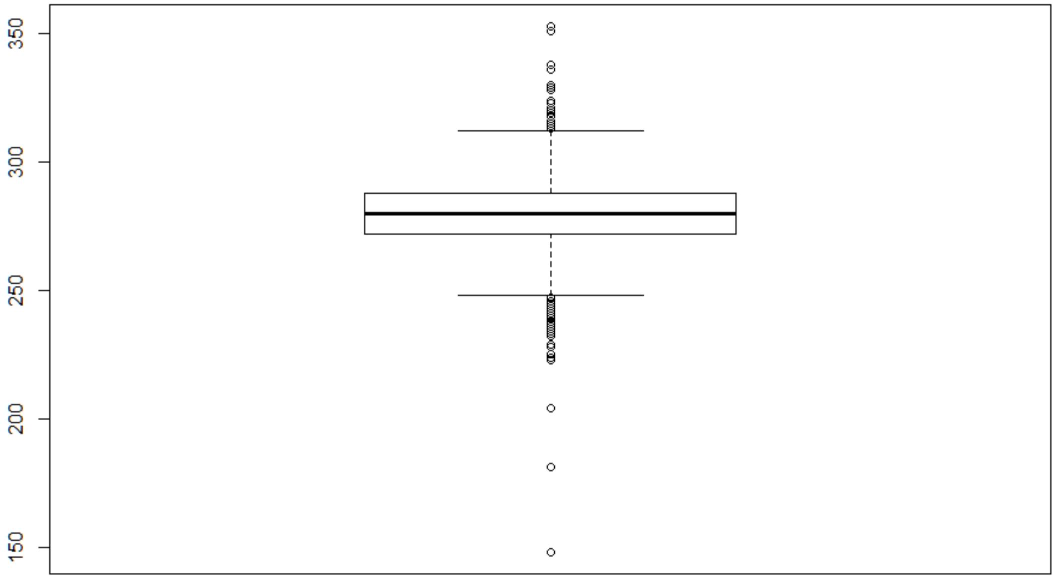



```
summary(babies$gestdays)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  148.0   272.0   280.0   286.9   288.0   999.0
```

gestation

Leaving out gestdays = 999

```
babies$gestdays < 999 # returns 1,236 Booleans (TRUE/FALSE)
which(babies$gestdays < 999) # returns 1,223 record #s that meet condition
boxplot(babies[which(babies$gestdays < 999),"gestdays"])
```




```
summary(babies[which(babies$gestdays < 999),"gestdays"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  148.0   272.0   280.0   279.3   288.0   353.0 
```

Some data are NOT normally distributed

TABLE 1.2 Service Times (Seconds) for Calls to a Customer Service Center

77	289	128	59	19	148	157	203
126	118	104	141	290	48	3	2
372	140	438	56	44	274	479	211
179	1	68	386	2631	90	30	57
89	116	225	700	40	73	75	51
148	9	115	19	76	138	178	76
67	102	35	80	143	951	106	55
4	54	137	367	277	201	52	9
700	182	73	199	325	75	103	64
121	11	9	88	1148	2	465	25

Some data are NOT normally distributed (skewed, kurtotic)

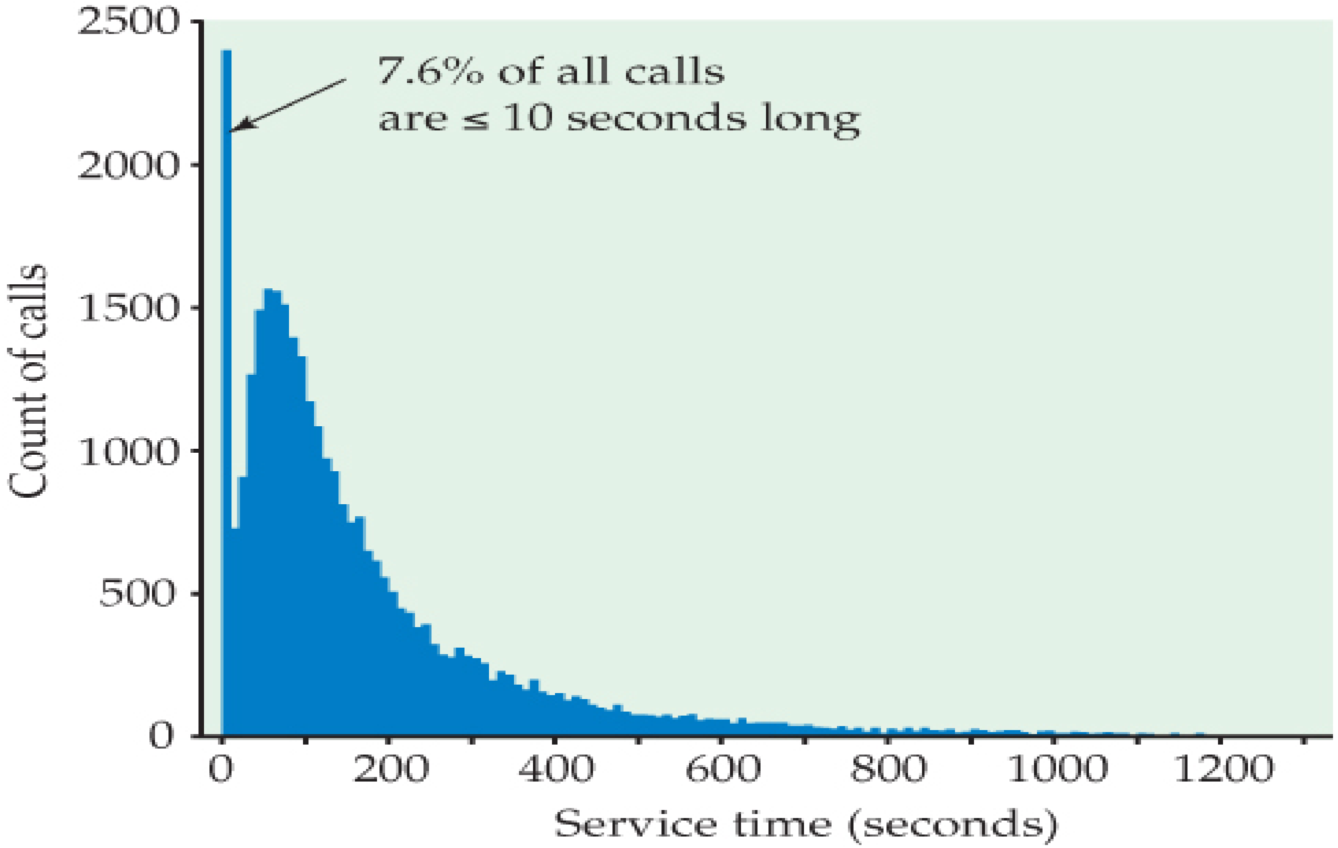


Figure 1.10, INTRODUCTION to the PRACTICE of STATISTICS, © 2014 W. H. Freeman

Some data are NOT normally distributed

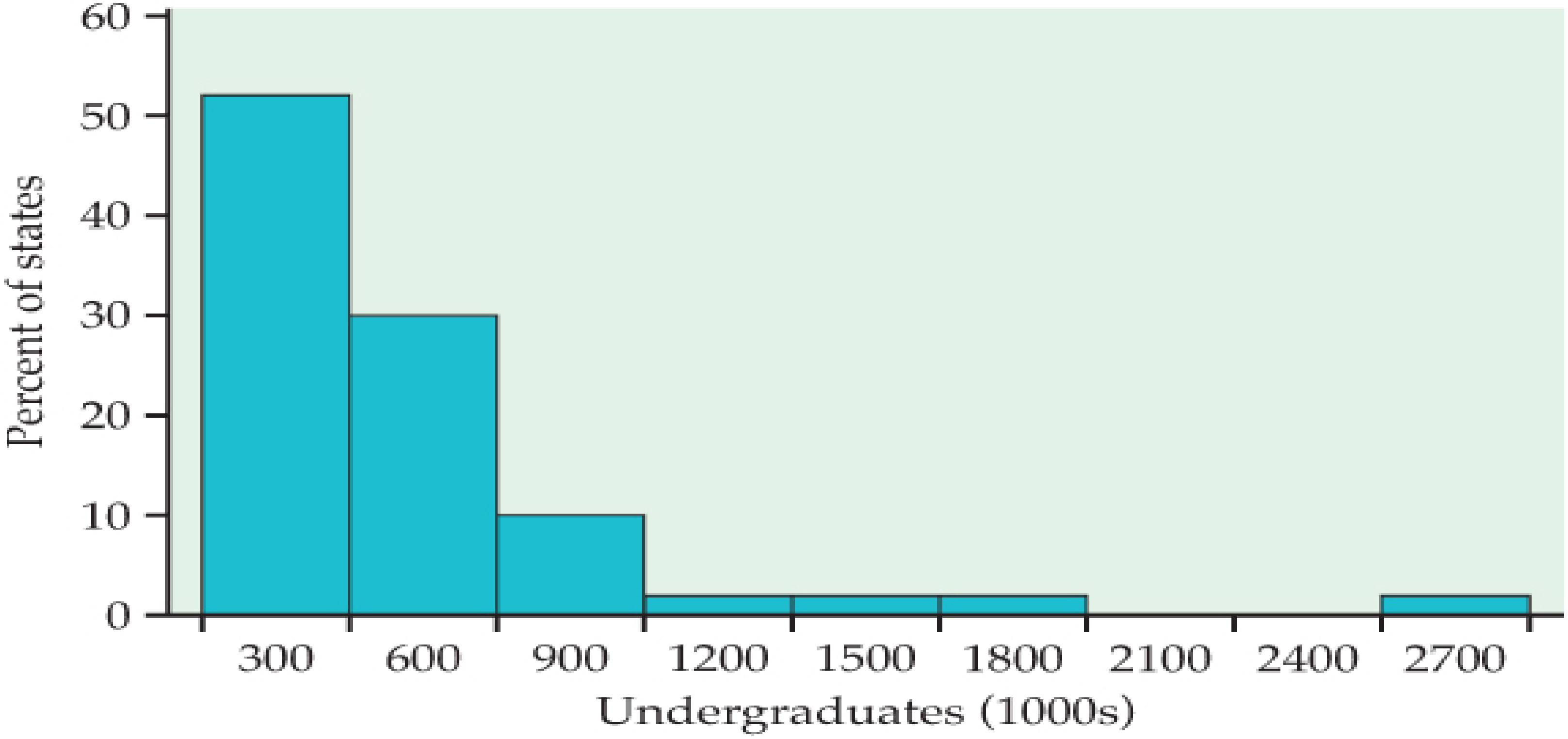


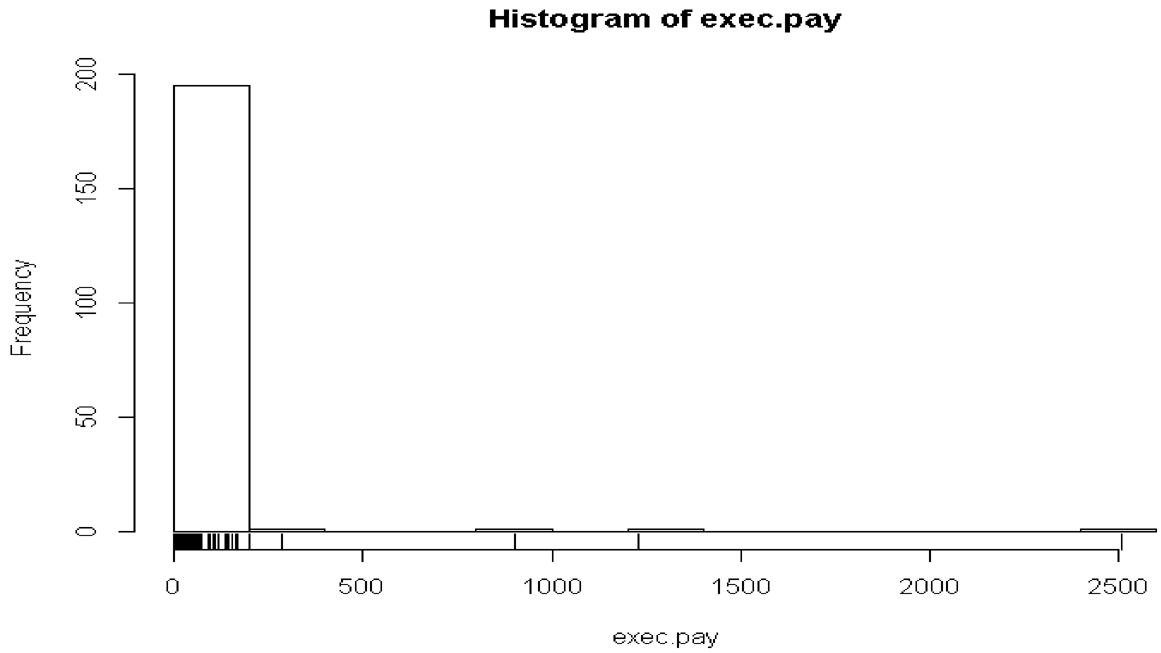
Figure 1.11, INTRODUCTION to the PRACTICE of STATISTICS, © 2014 W. H. Freeman

Some data are NOT normally distributed

Look at executive pay:

```
exec.pay2 <- read.csv("http://math.mercyhurst.edu/~sousley/STAT_139/data/exec.pay2.csv", header=T);
```

```
mean(exec.pay2$salary)
[1] 59.88945

median(exec.pay2$salary)
[1] 27

hist(exec.pay2$salary)

rug(exec.pay2$salary)
```



Because salaries are skewed, we use the median instead

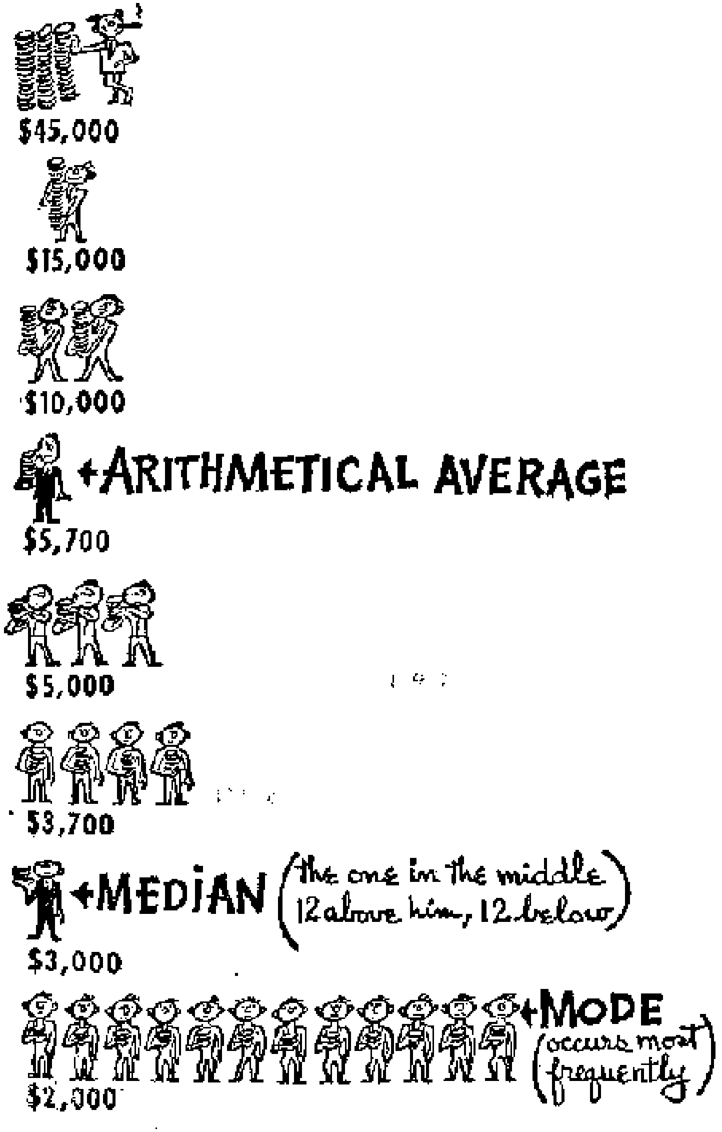
HOW TO LIE WITH STATISTICS

Darrell Huff

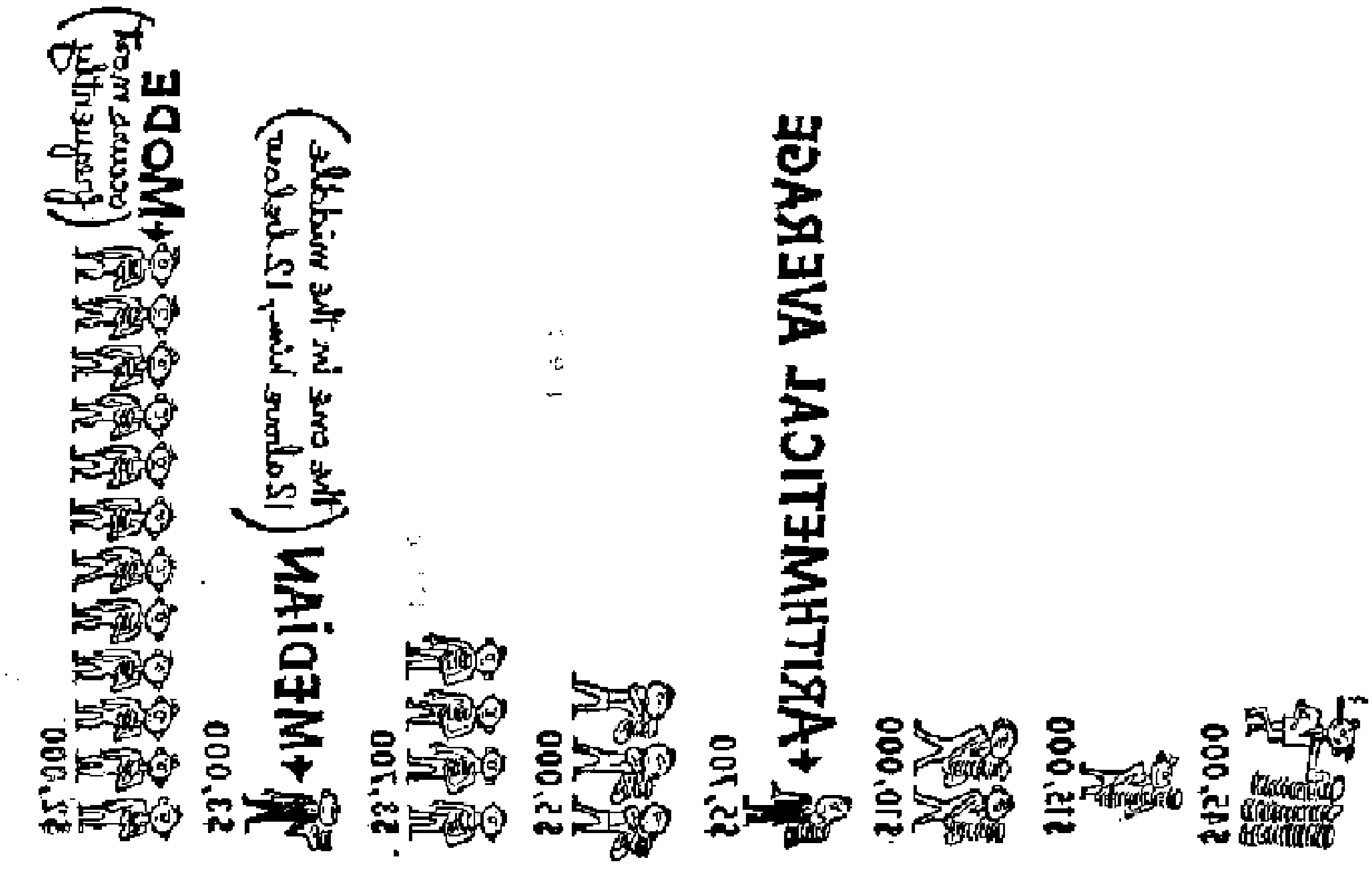
Illustrated by Irving Geis

Over Half a Million Copies Sold—
An Honest-to-Goodness Bestseller

Mean, Median, Mode: Income

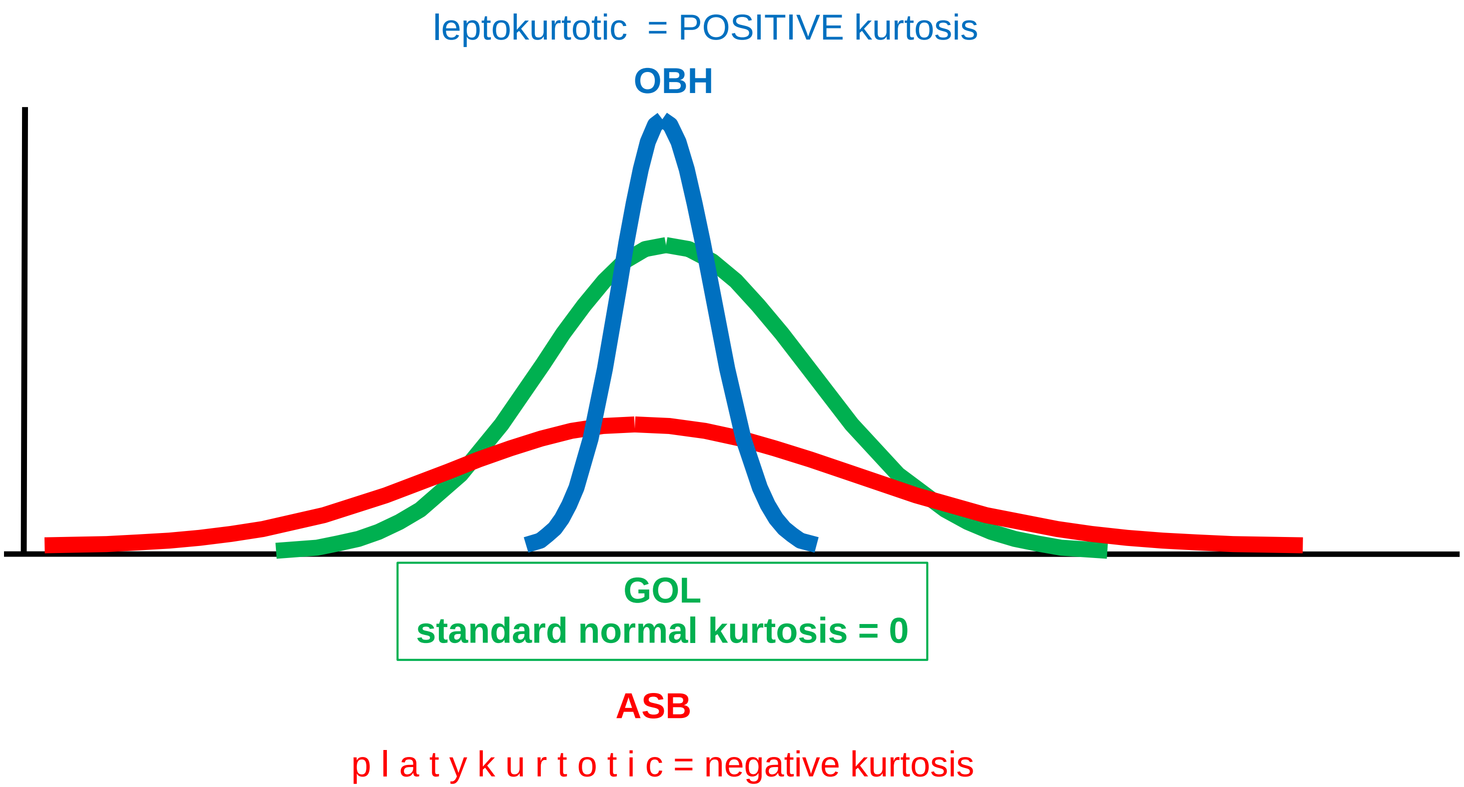


Mean, Median, Mode: Income



Some data are NOT normally distributed - in other ways

Kurtosis (concentration)



Density plots

- are smoothed probability density lines
- are "layered" (added) onto a plot, like a rug plot
- the histogram must be set to show probability (not freq)

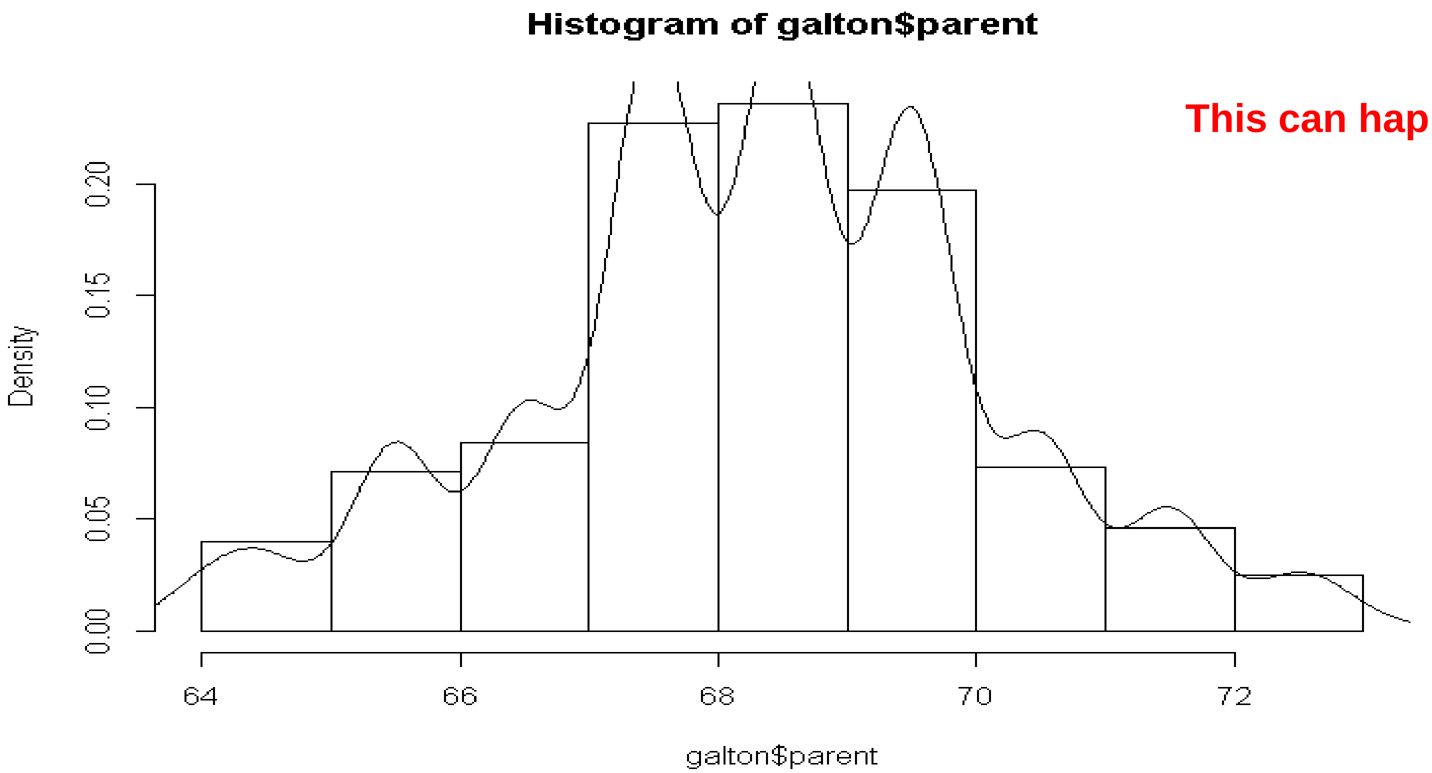
```
prob = T
```

- draw (add) density using lines

```
lines(density(<data> ) )
```

```
# Galton's data; height of parents and their children in inches;
# Remember, EVERYTHING is case-sensitive!;

galton <- read.csv("http://math.mercyhurst.edu/~sousley/STAT_139/data/galton.csv", header=T);

# draw a histogram;
hist(galton$parent, prob = T);

# ADD density lines ;
lines(density(galton$parent));
```

Density plots

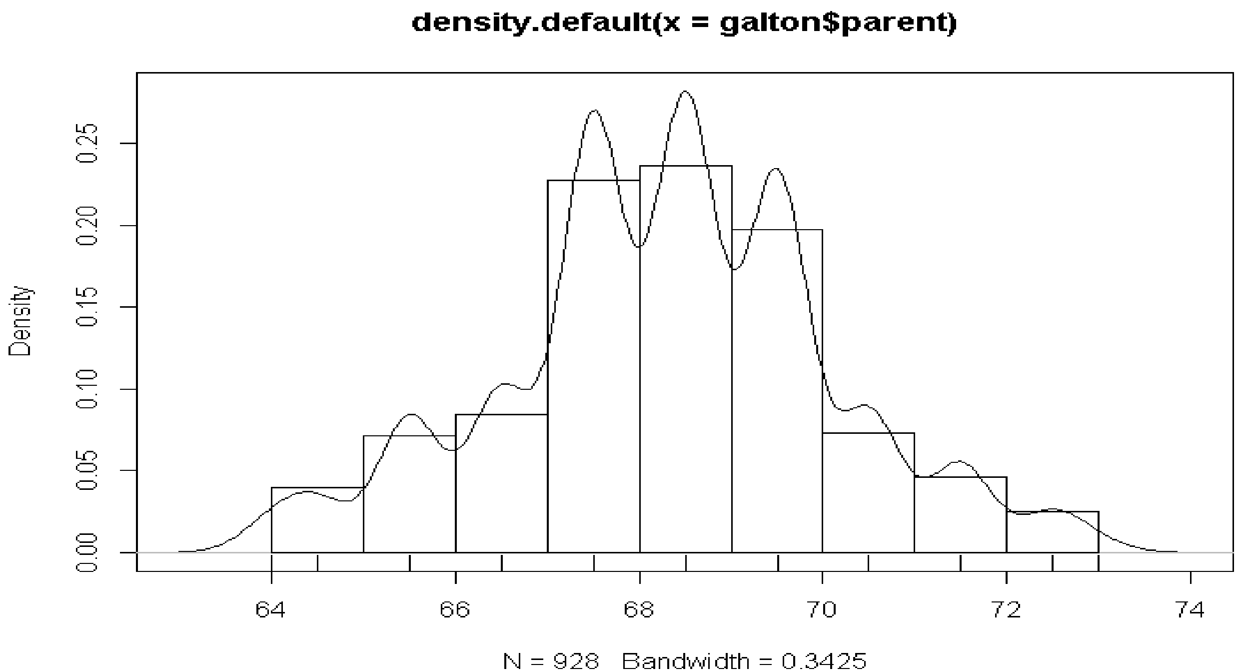


This can happen!

```
hist(galton$parent, prob = T);
lines(density(galton$parent));
```

Density plots

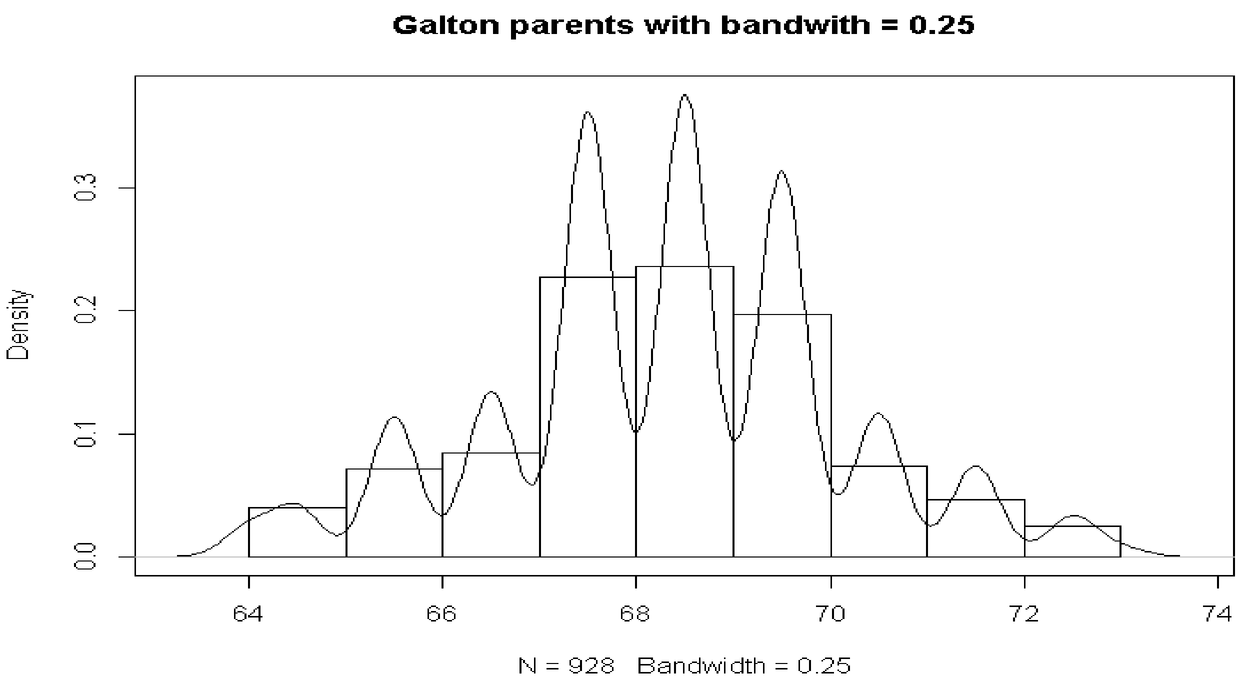
If some graphics are cut off, reverse the order
- the first **plot** sets the limits (**`plot`**)
- the second layer is drawn on top of the first, use **`add=T`**




```
plot(density(galton$parent));
hist(galton$parent, prob = T, add=T);
rug(galton$parent )  # but little information added - rounded;
```

Density plots

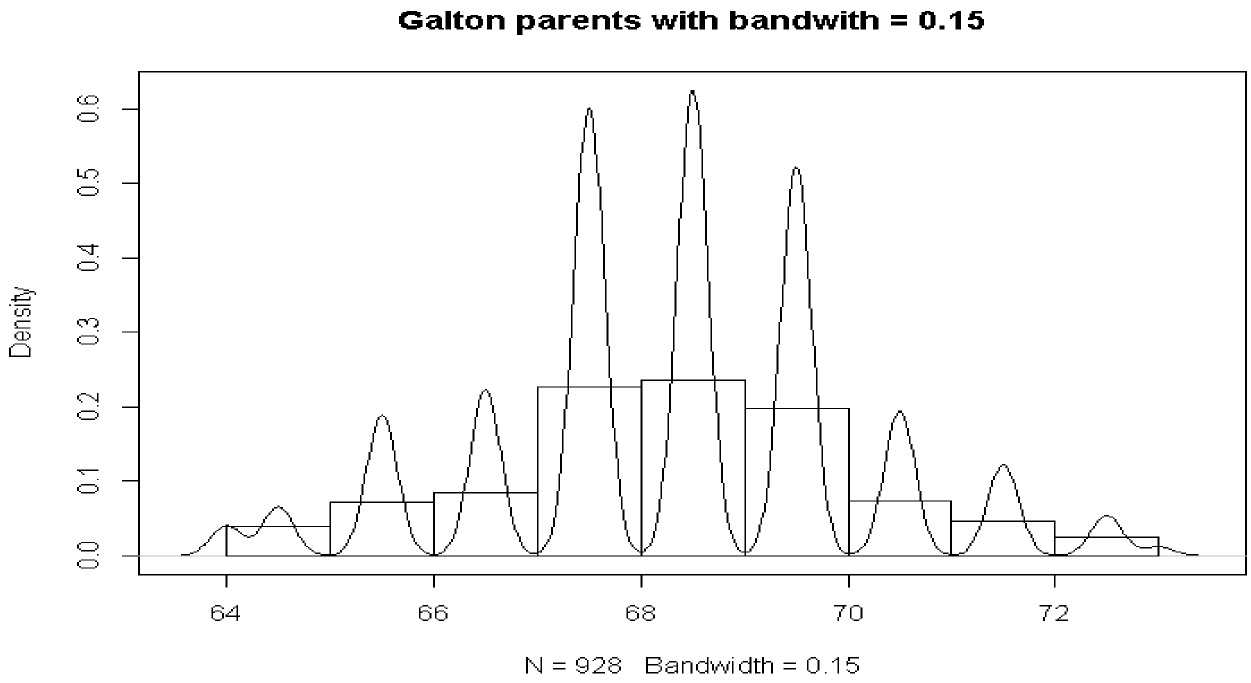
The **bandwidth** (**bw**) functions like the bin width of histograms
- larger bandwidth means more smoothing
- smaller bandwidth means more jagged lines




```
plot(density(galton$parent, bw=0.25), main = "Galton parents with bandwith = 0.25" );
hist(galton$parent, prob = T, add=T);
```

Density plots

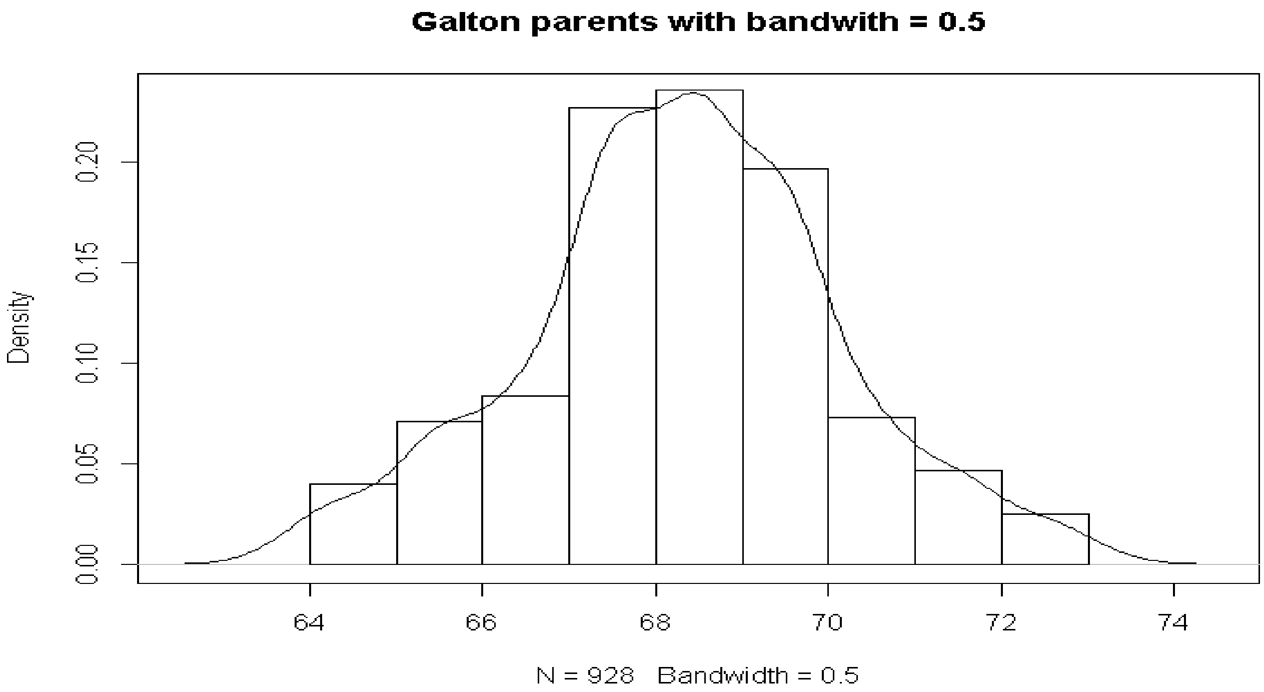
The **bandwidth** (**bw**) functions like the bin width of histograms
- larger bandwidth means more smoothing
- smaller bandwidth means more jagged lines




```
plot(density(galton$parent, bw=0.15), main = "Galton parents with bandwith = 0.15" );
hist(galton$parent, prob = T, add=T);
```

Density plots

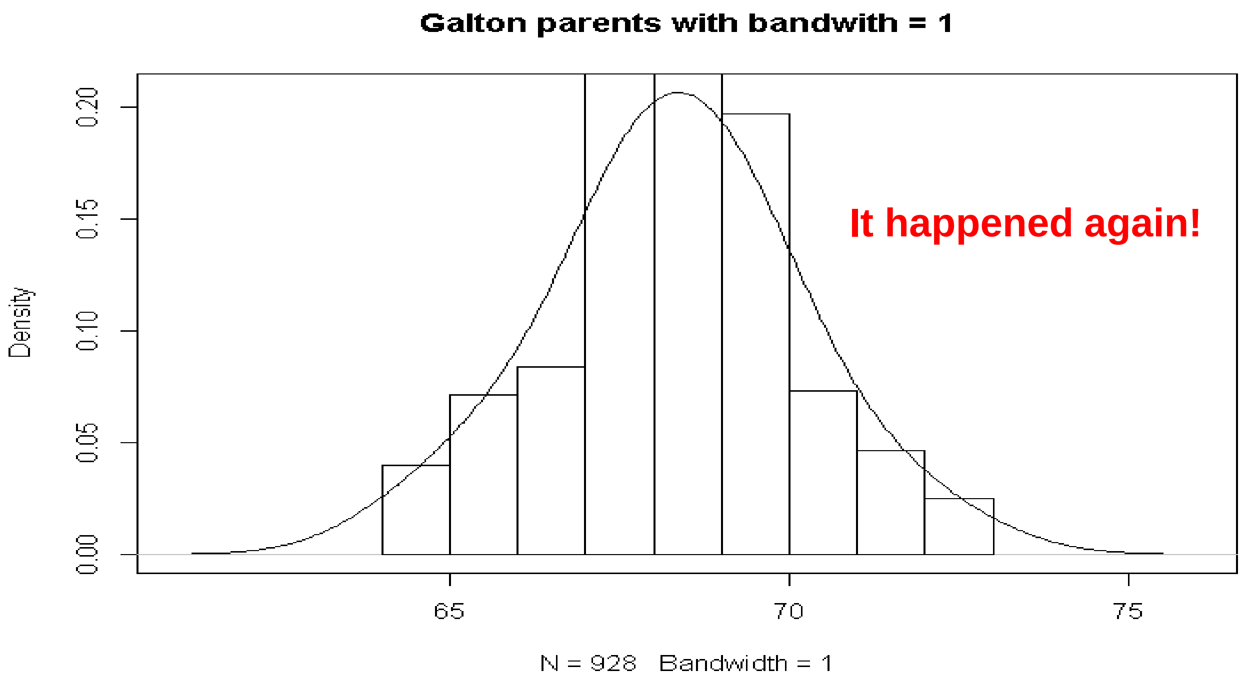
The **bandwidth** (**bw**) functions like the bin width of histograms
- larger bandwidth means more smoothing
- smaller bandwidth means more jagged lines




```
plot(density(galton$parent, bw=0.5), main = "Galton parents with bandwith = 0.5" );
hist(galton$parent, prob = T, add=T);
```

Density plots

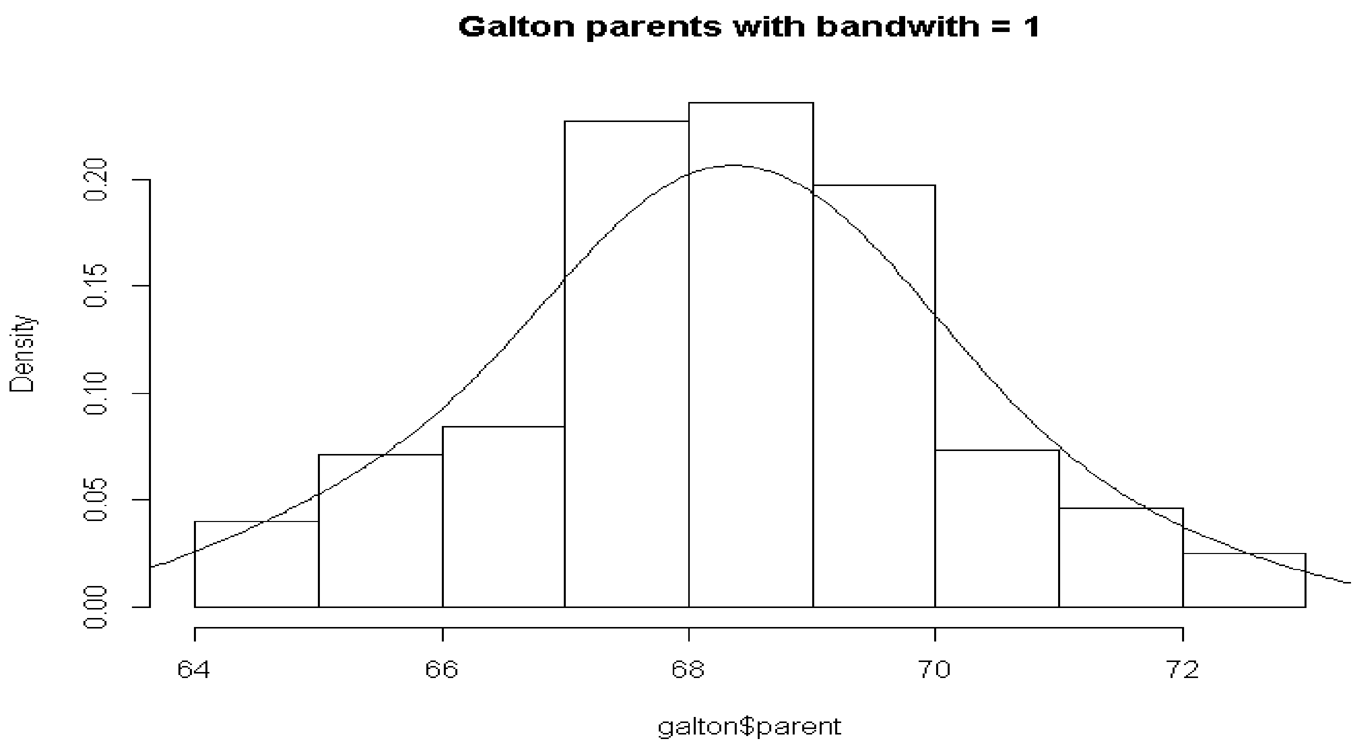
The bandwidth functions like the bin width of histograms
- larger bandwidth means more smoothing
- smaller bandwidth means more jagged lines




```
plot(density(galton$parent, bw=1), main = "Galton parents with bandwith = 1" );
hist(galton$parent, prob = T, add=T);
```

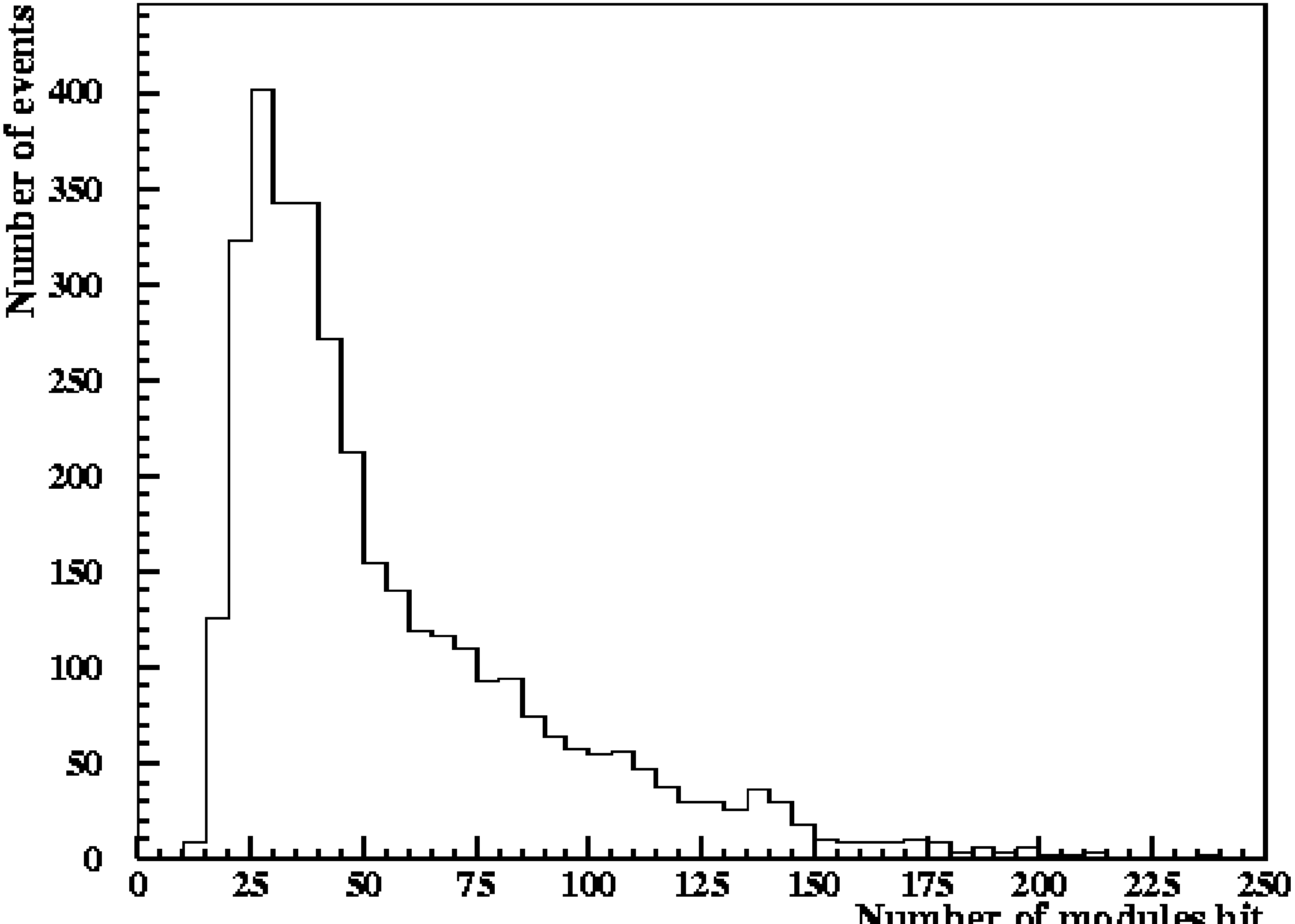
Density plots

Changing the main title (**`main=`**)
- it goes with the plot statement

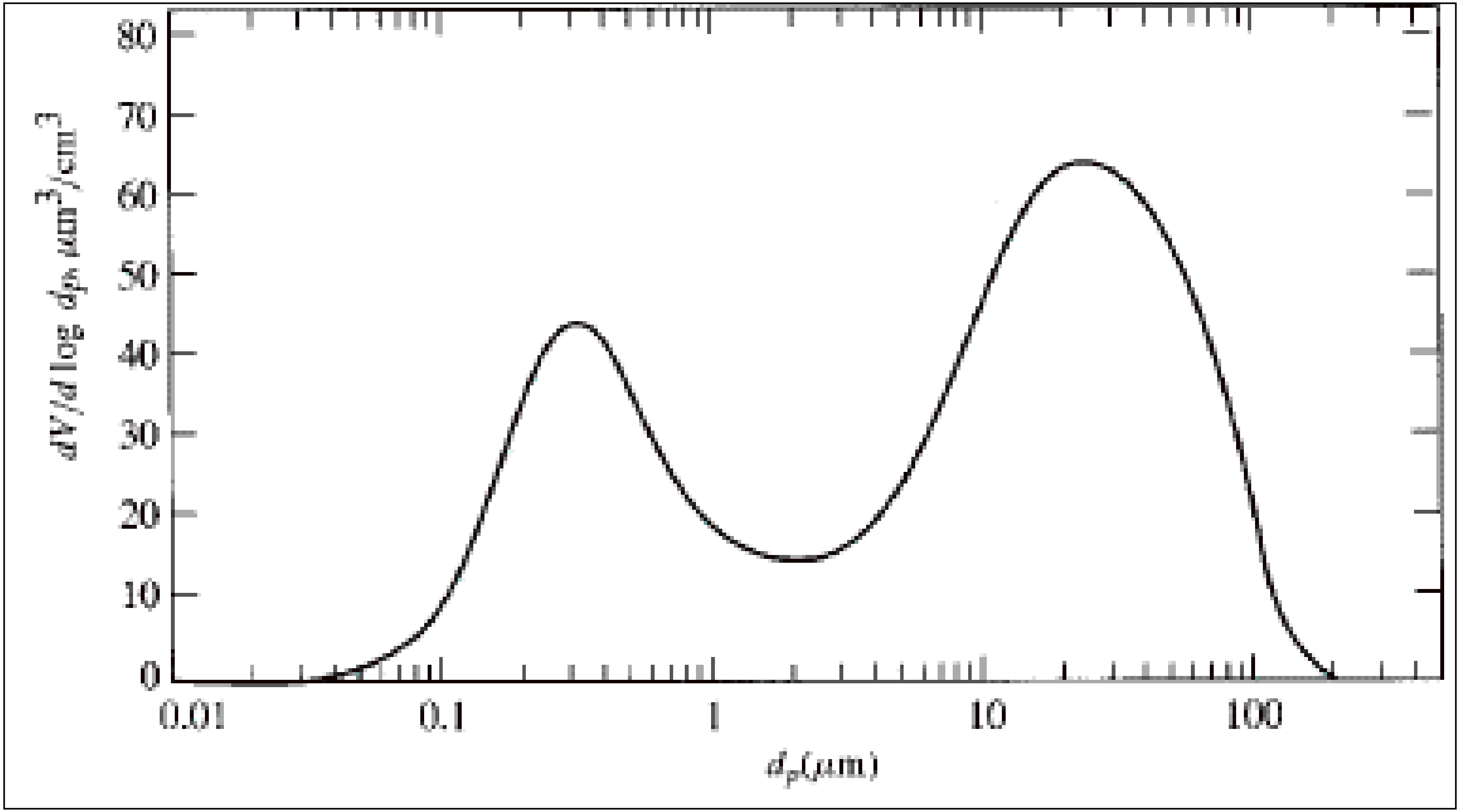



```
# This time draw the histogram first to set the scaling
hist(galton$parent, prob = T, main = "Galton parents with bandwith = 1");
lines(density(galton$parent, bw=1, add=T) );
```

Describing data distributions graphically (Skewness and kurtosis)

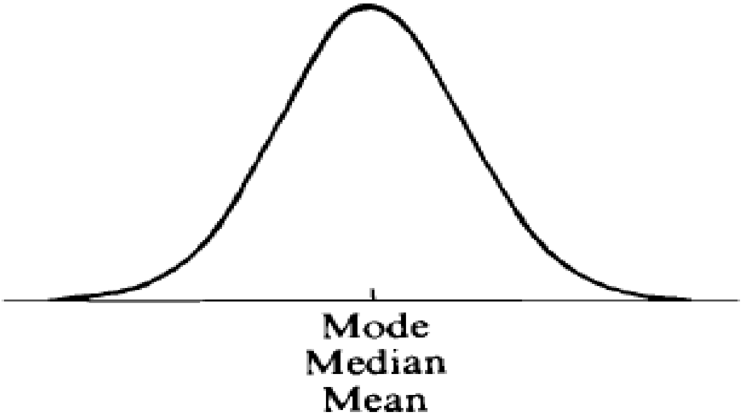


Describing data distributions graphically (Bimodal? Multimodal?)

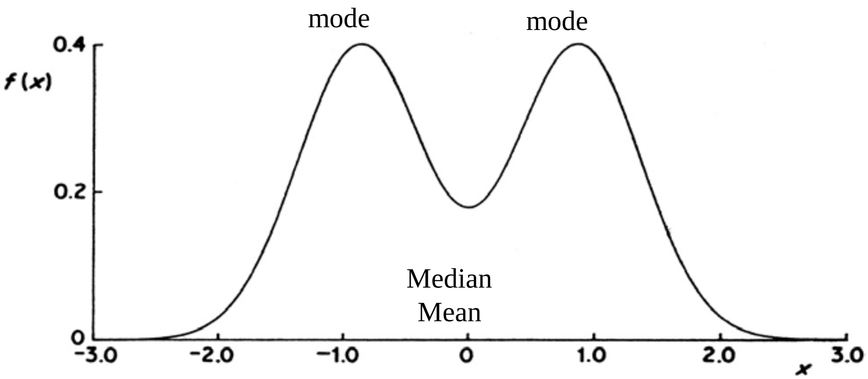


Normally distributed data

Mean = sum of all values / n
Median: (midpoint; 50^{th} percentile)
Mode: most common value

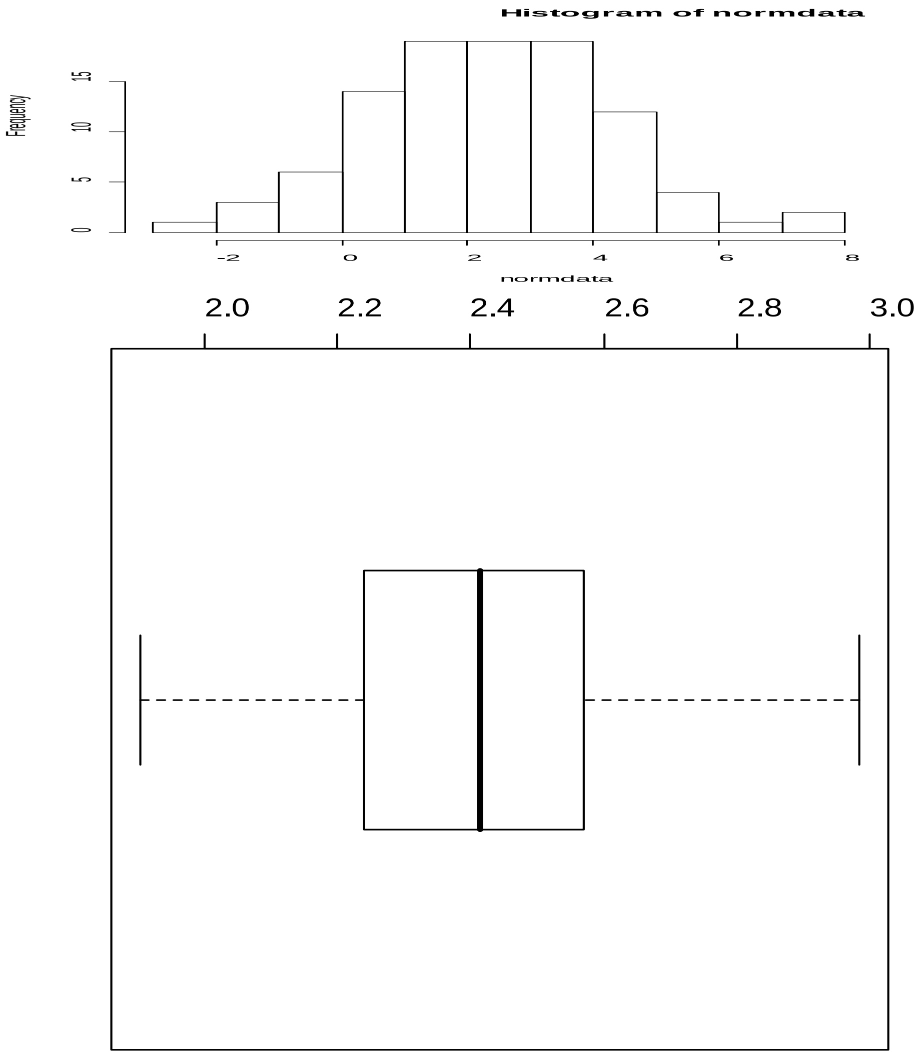


The Normal Distribution



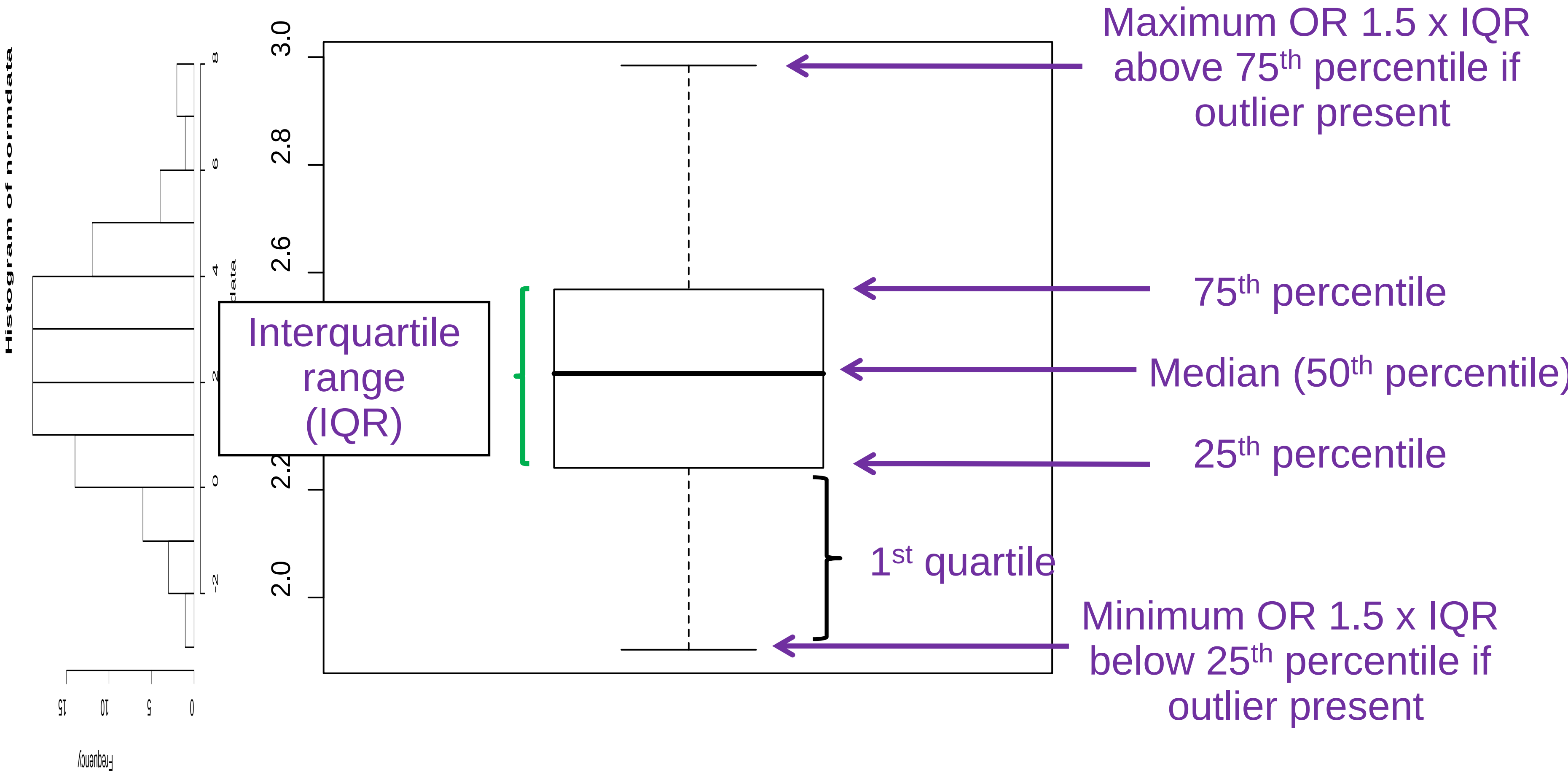
Bimodal Distribution

The anatomy of a box and whisker plot



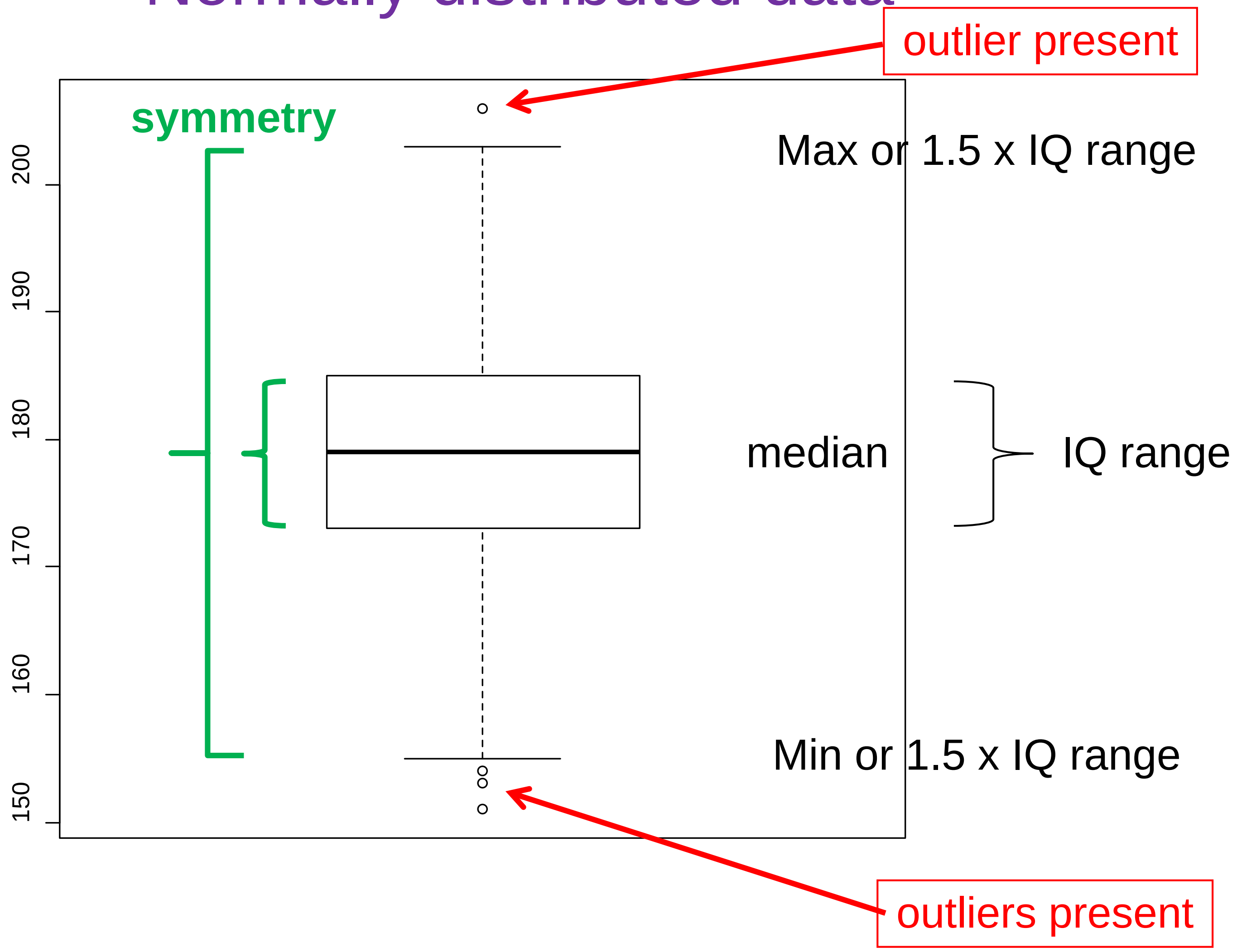
Quartiles: 25th, 50th (median), 75th percentiles
IQR: between the 25th and 75th percentiles

The anatomy of a box and whisker plot

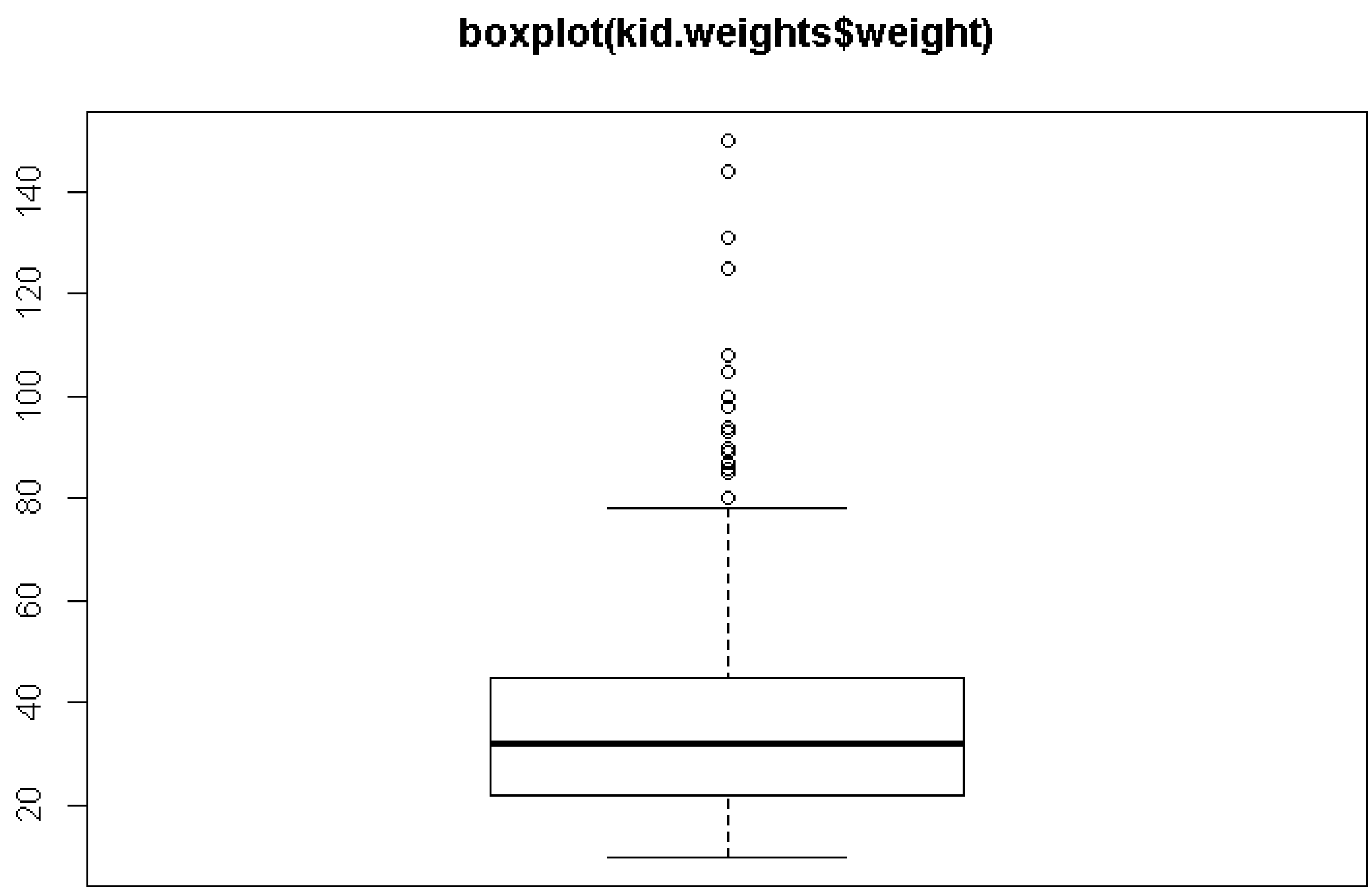


Quartiles: 25th, 50th (median), 75th percentiles
IQR: between the 25th and 75th percentiles

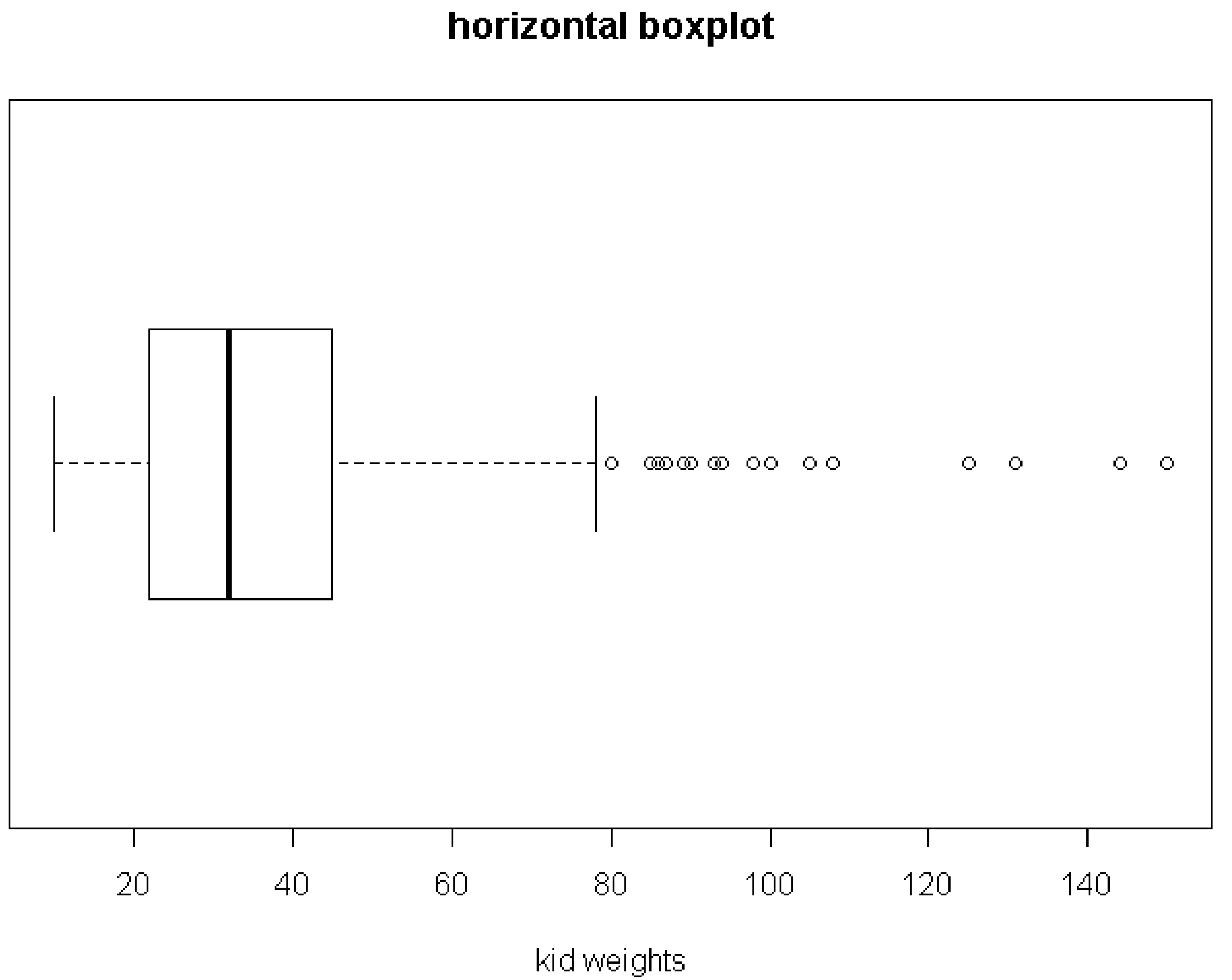
The box and whisker plot: Normally distributed data



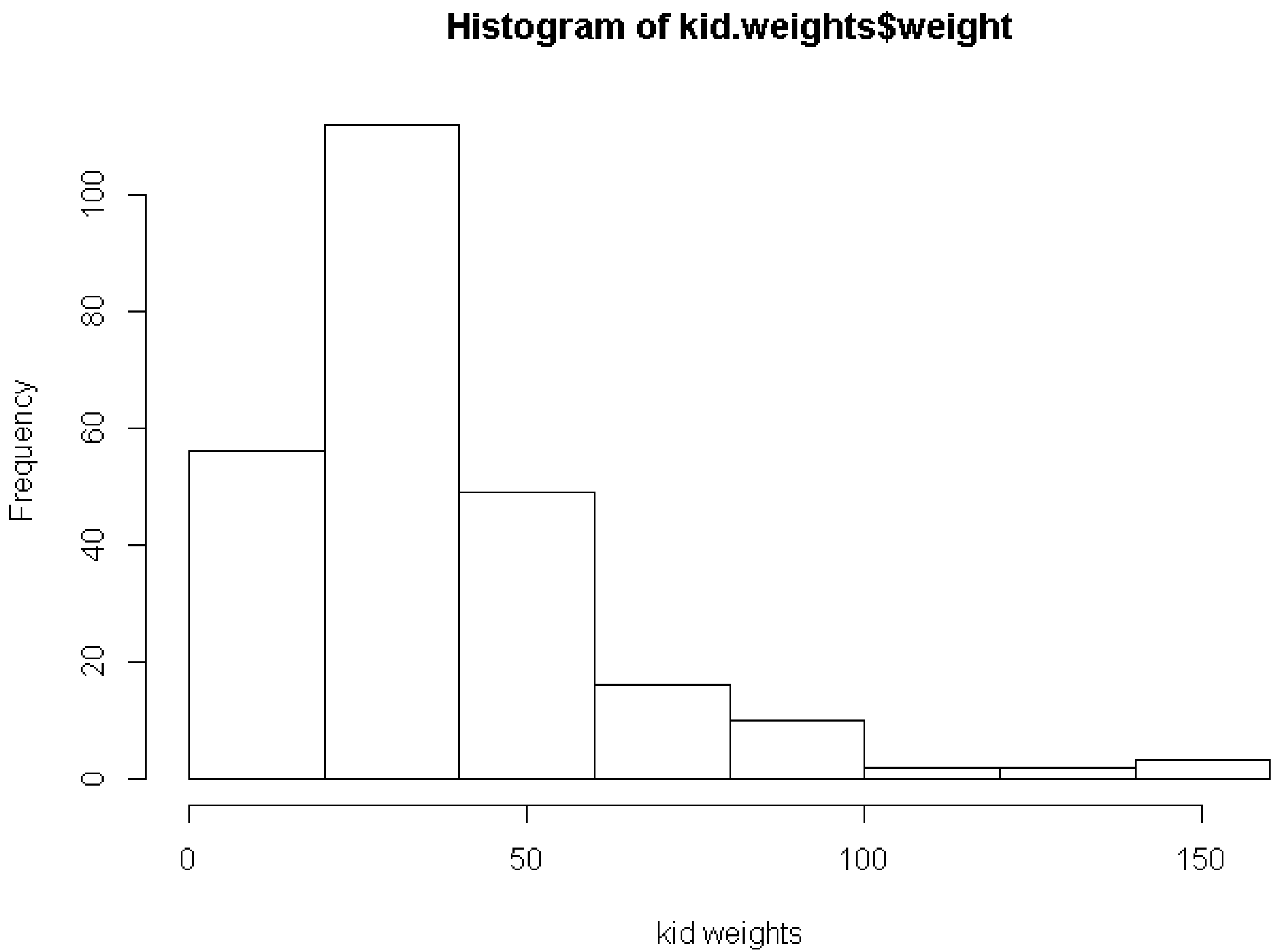

```
Howells <- read.csv("http://math.mercyhurst.edu/~sousley/STAT_139/data/Howells.csv", header = T);
boxplot(Howells$GOL);
```




```
kid.weights <- read.csv("http://math.mercyhurst.edu/~sousley/STAT_139/data/kid.weights.csv", header = T);

boxplot(kid.weights$weight, main = 'boxplot(kid.weights$weight)');
```

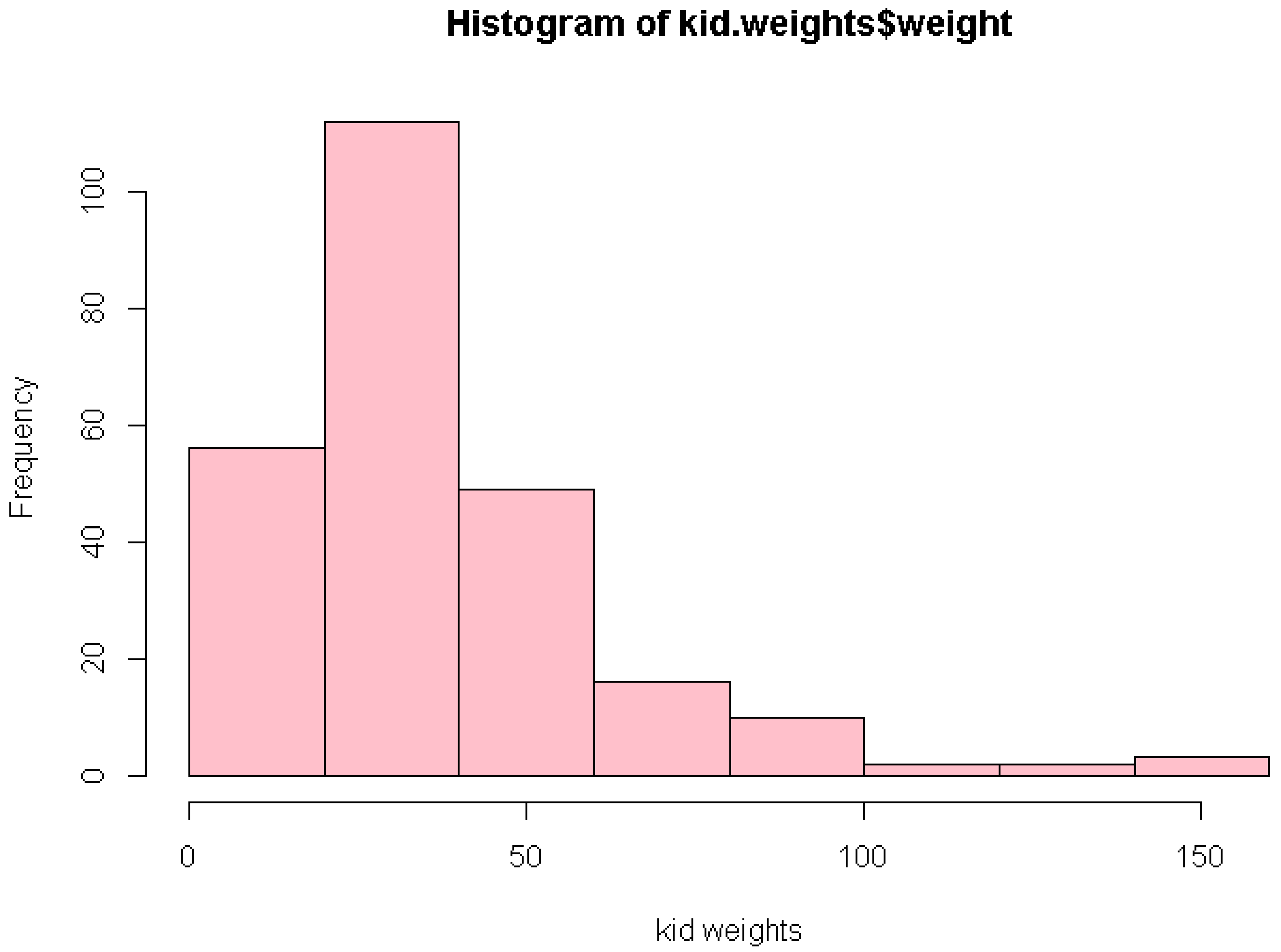



```
boxplot(kid.weights$weight, main = 'horizontal boxplot', horizontal = T, xlab = 'kid weights')
```

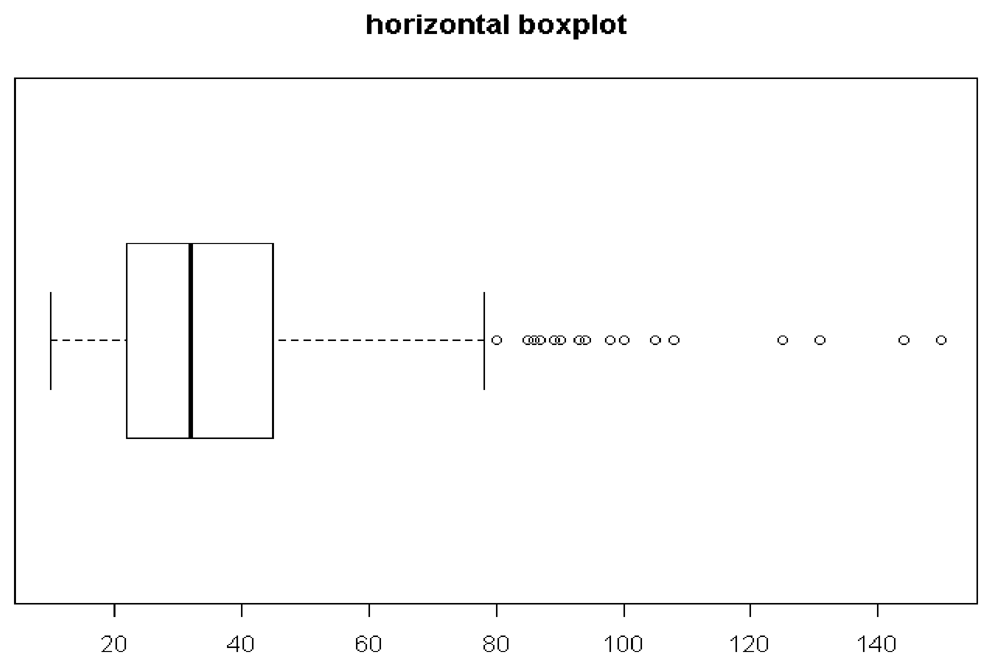


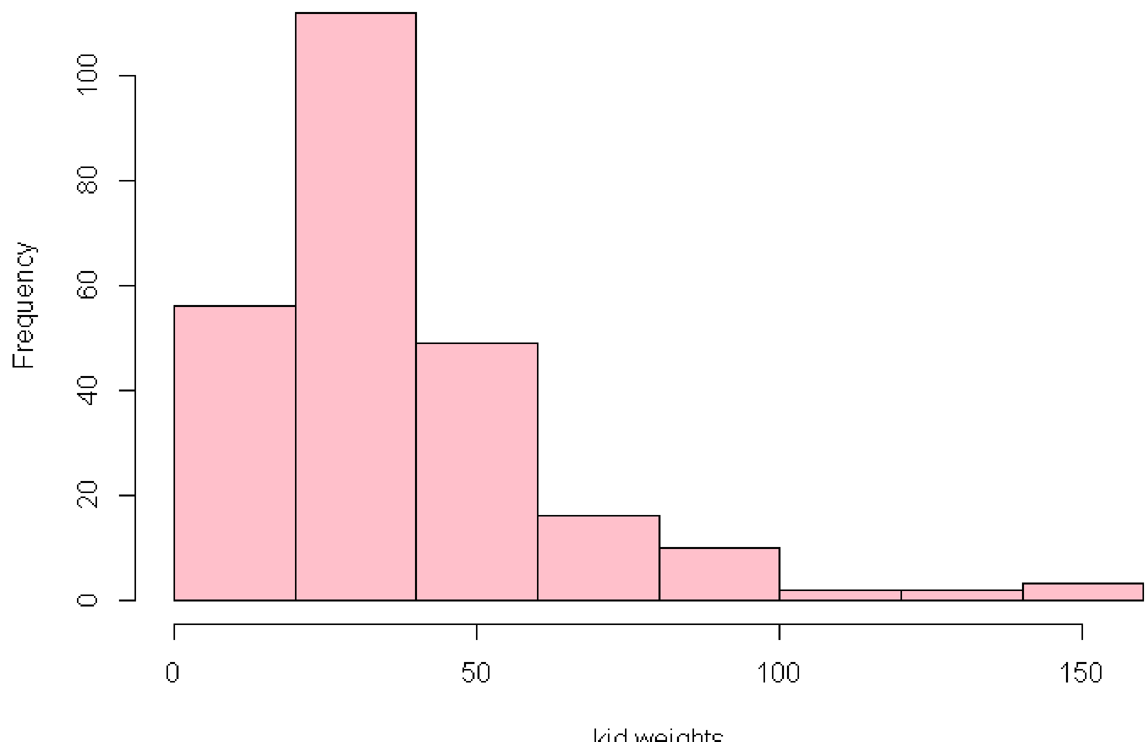
`hist(kid.weights$weight, xlab = 'kid weights')`

Add some color

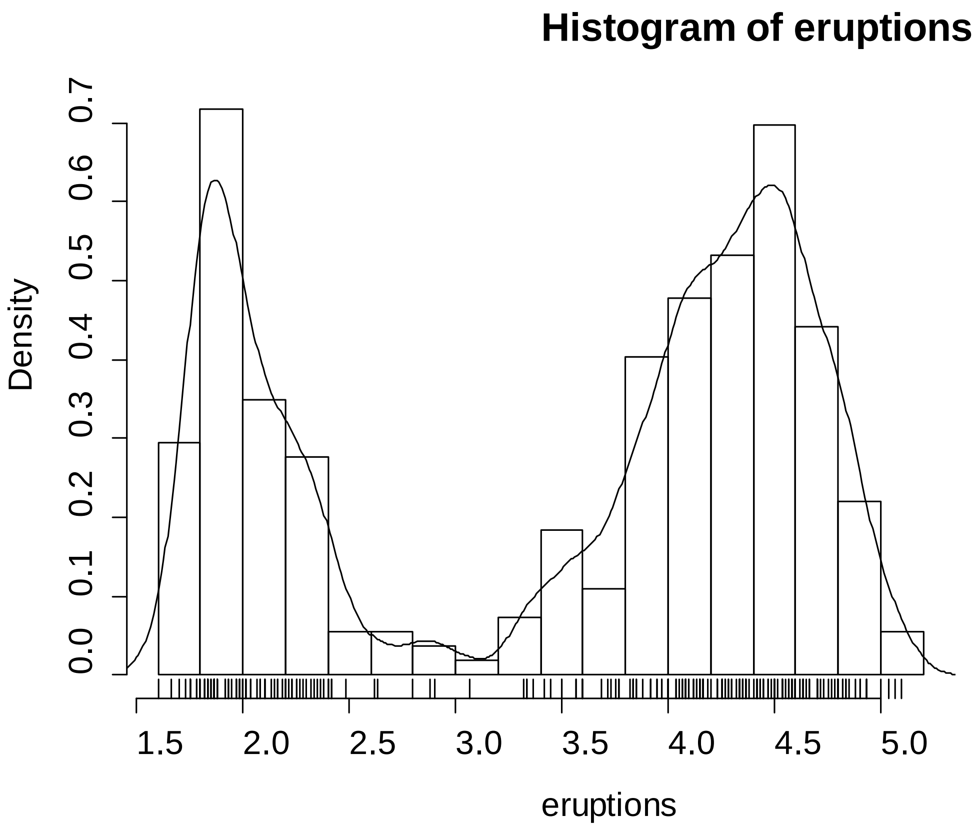


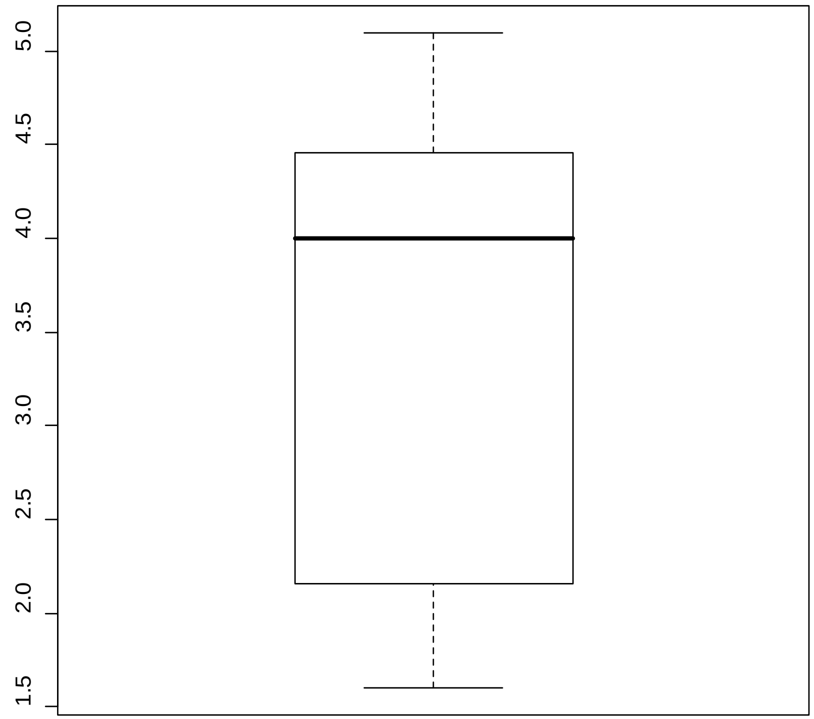

```
hist(kid.weights$weight, xlab = 'kid weights', col = 'pink')
```


horizontal boxplot
20
40
60
80
100
120
140


Frequency
0
20
40
60
80
100
0
50
100
150
kid weights

Some data are NOT normally distributed - so the mean and sd don't mean much

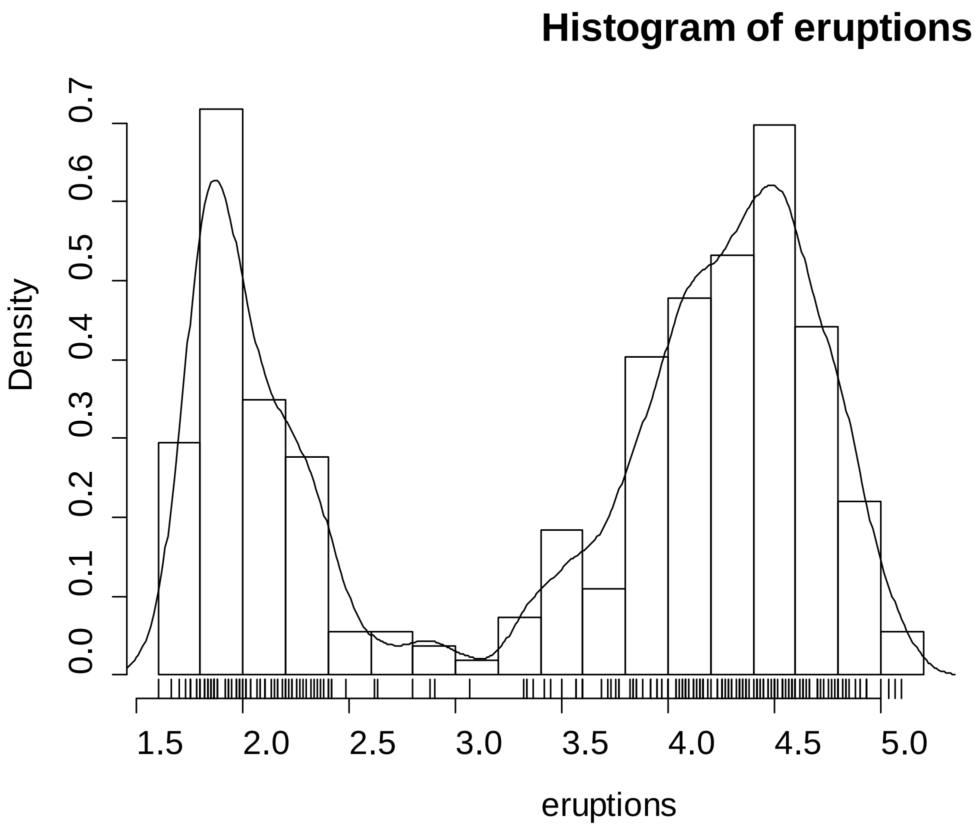


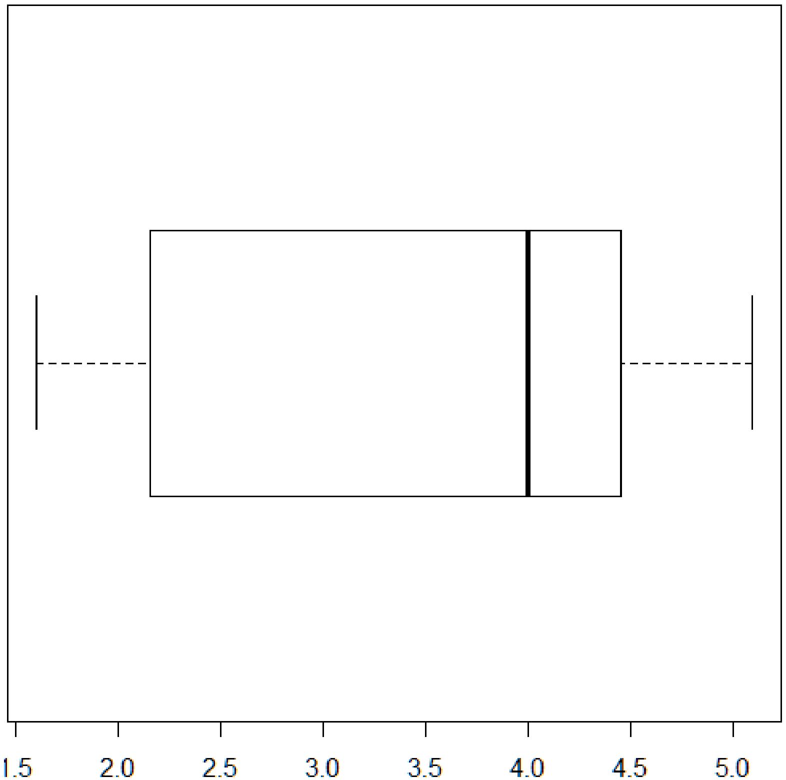



```
# old faithful data;
attach(faithful);
# eruption time (duration) in minutes;
summary(eruptions) ;
hist(eruptions, seq(1.6, 5.2, 0.2), prob=TRUE);
lines(density(eruptions, bw=0.1)) # prob density;
rug(eruptions) # tick marks for data points;
```

```
boxplot(eruptions)
```

Some data are NOT normally distributed - you need both histogram and boxplot

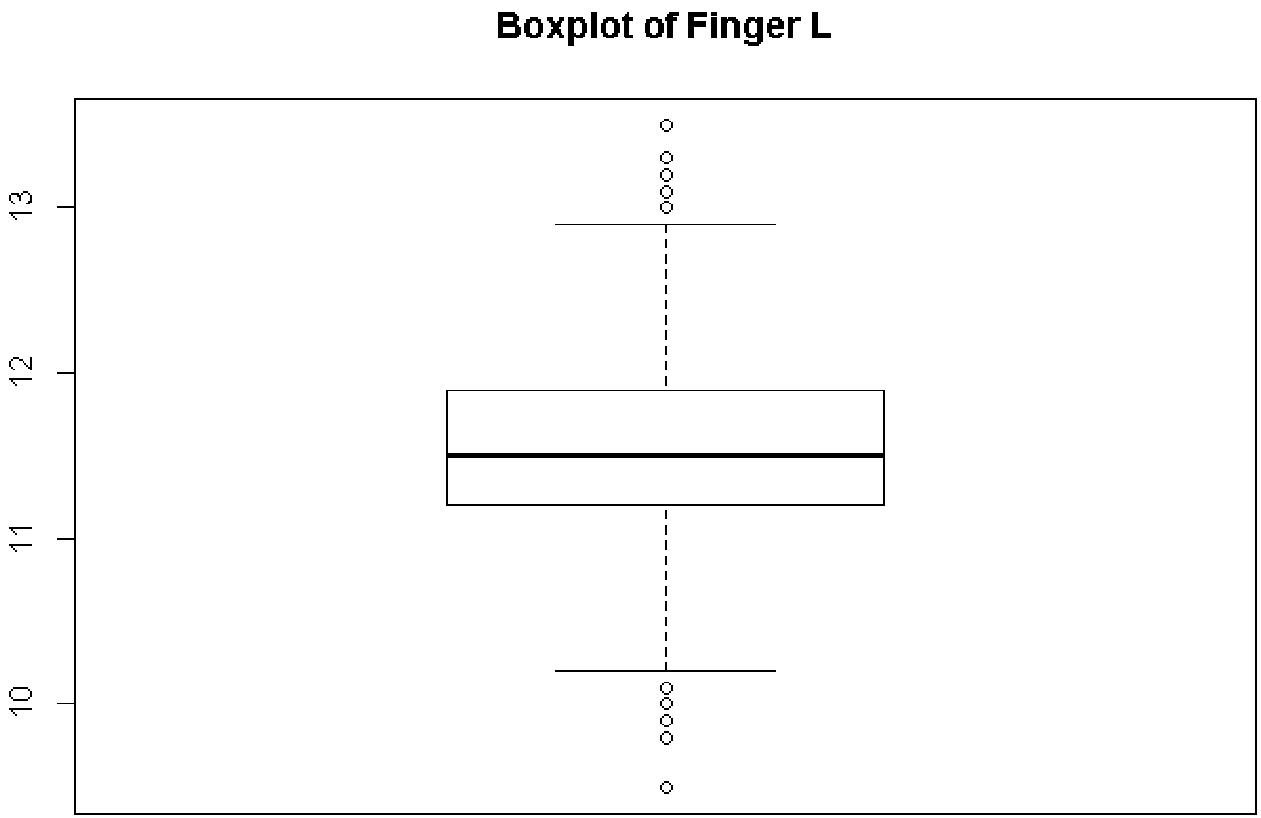




boxplot(eruptions)

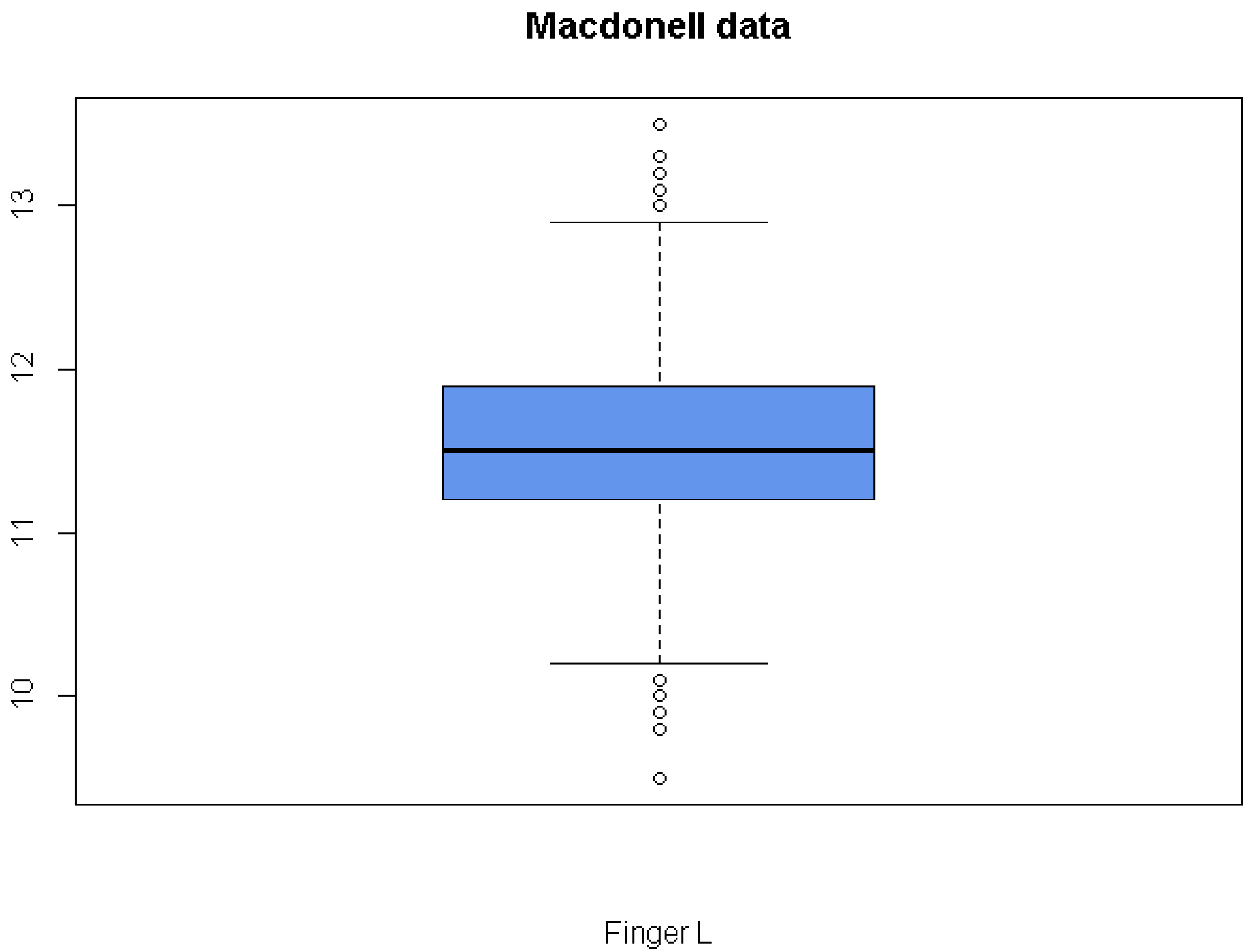
```
# old faithful data;
attach(faithful);
# eruption time (duration) in minutes;
summary(eruptions);
hist(eruptions, seq(1.6, 5.2, 0.2), prob=TRUE);
lines(density(eruptions, bw=0.1)) # prob density;
rug(eruptions) # tick marks for data points;
```

Boxplot



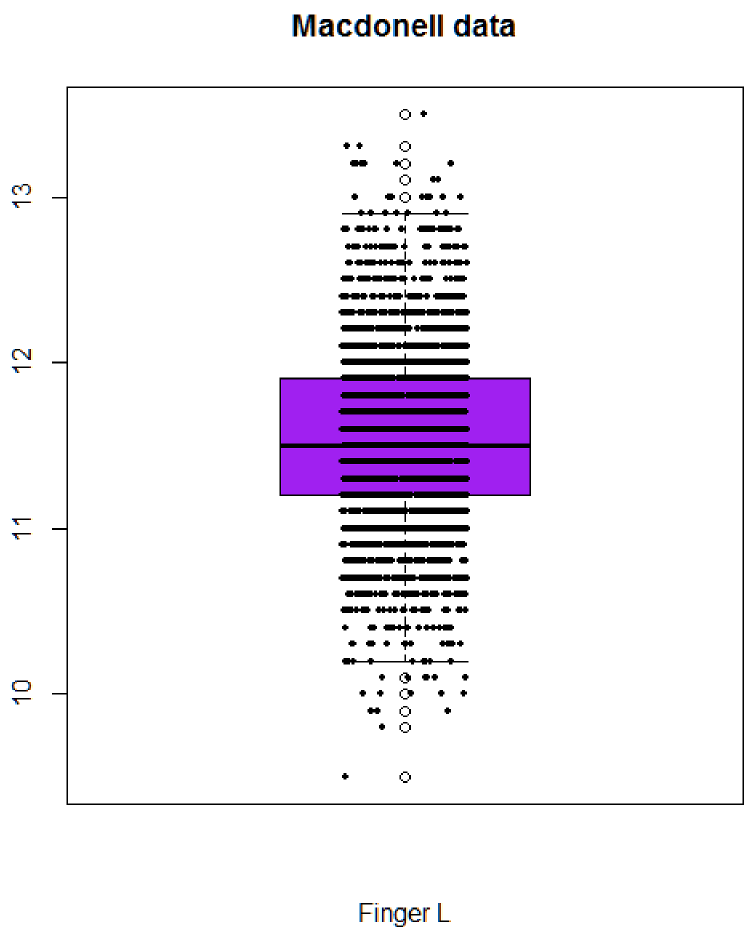

```
# draw a box and whisker plot
boxplot(MacFingL, main = 'Boxplot of Finger L');
```

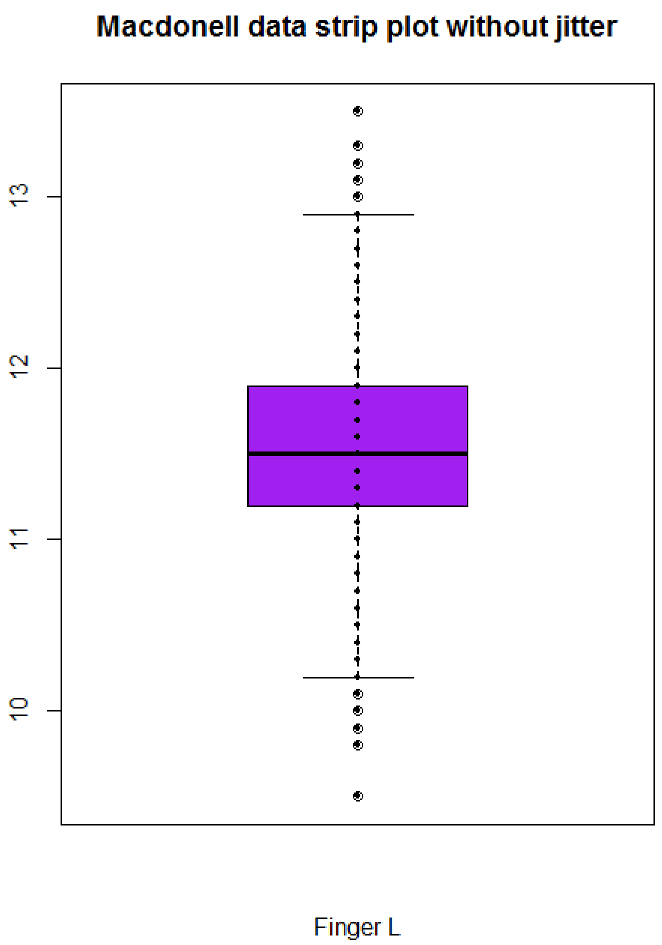
Box plot color




```
boxplot(MacFingL, main = 'Macdonell data',xlab = 'Finger L', col=c("cornflowerblue") ) ;
```

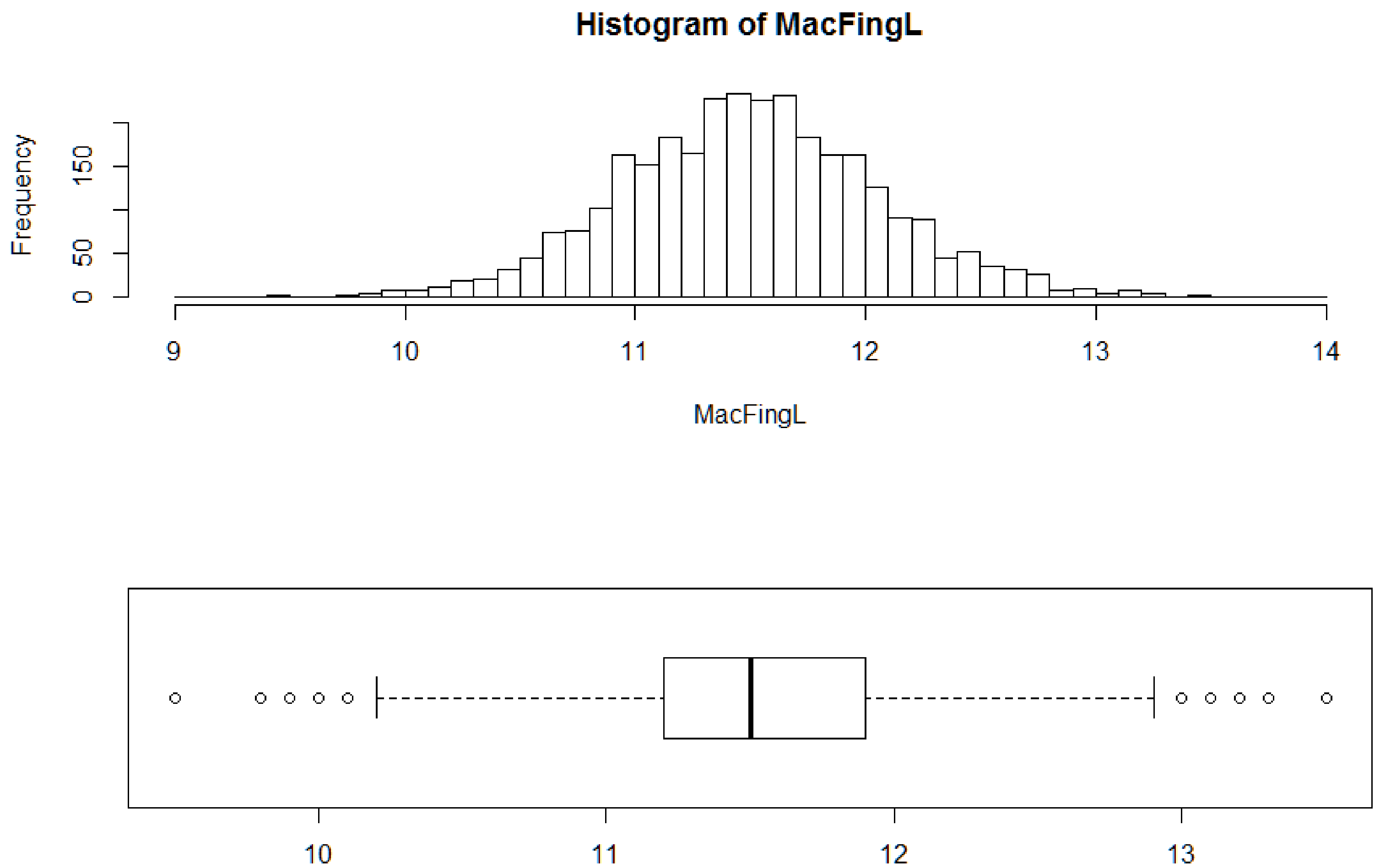
Box plot color and strip charts





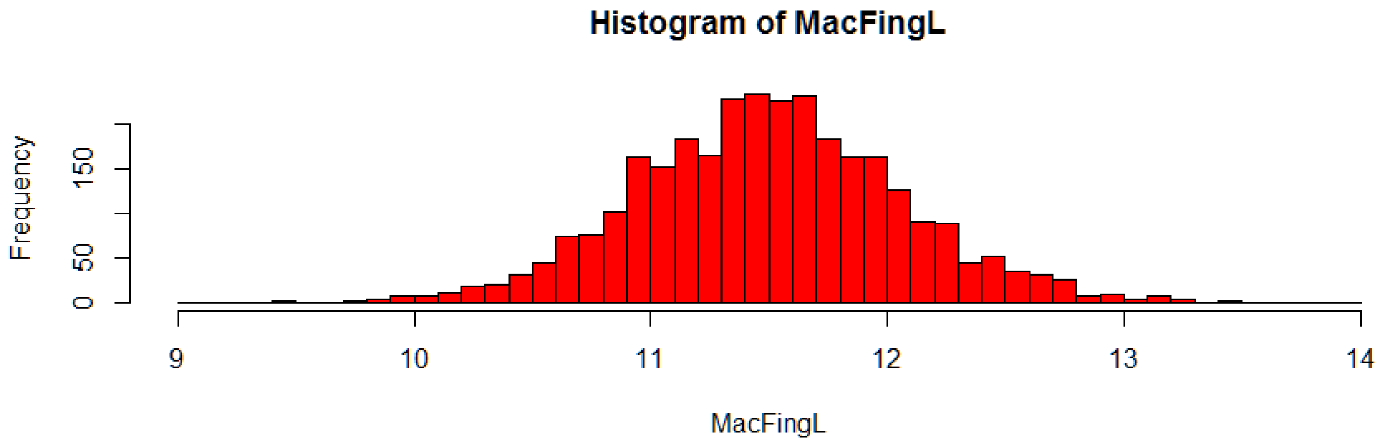

```
boxplot(MacFingL, main = 'Macdonell data', xlab = 'Finger L', col=c("purple") ) ;

# pch chooses symbol (16); jitter scatters points; ces scales symbol ;
# add = T: add to current plot
stripchart(MacFingL,vertical=T,pch=16,method="jitter",cex=0.5,add=T) ;
```

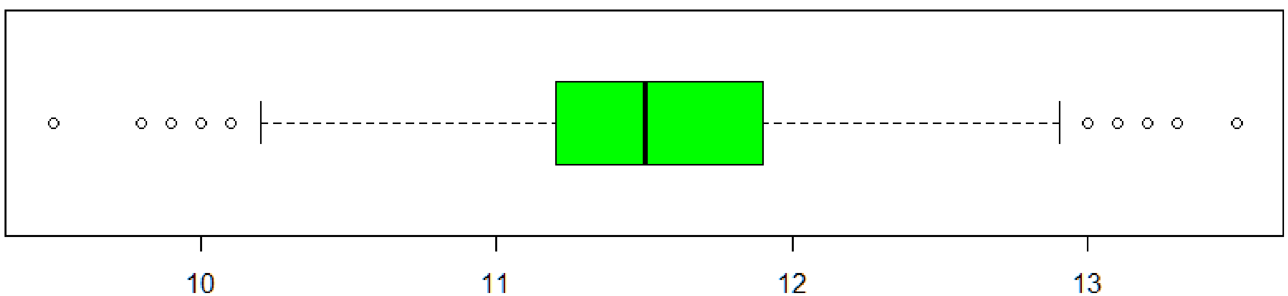
Printing two plots in the same graph



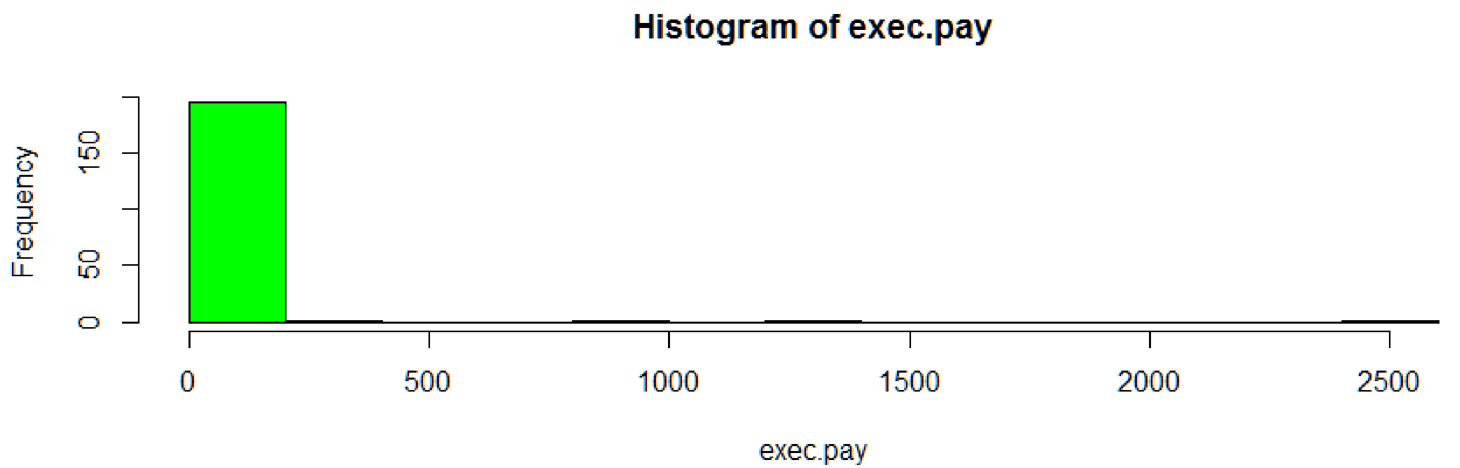

```
# Set up graph pane of 1 row, 2 columns;
par(mfrow=c(2,1));
hist(MacFingL, seq(9,14,0.1));
boxplot(MacFingL,horizontal=T );
# reset to whole graph pane;
par(mfrow=c(1,1));
```

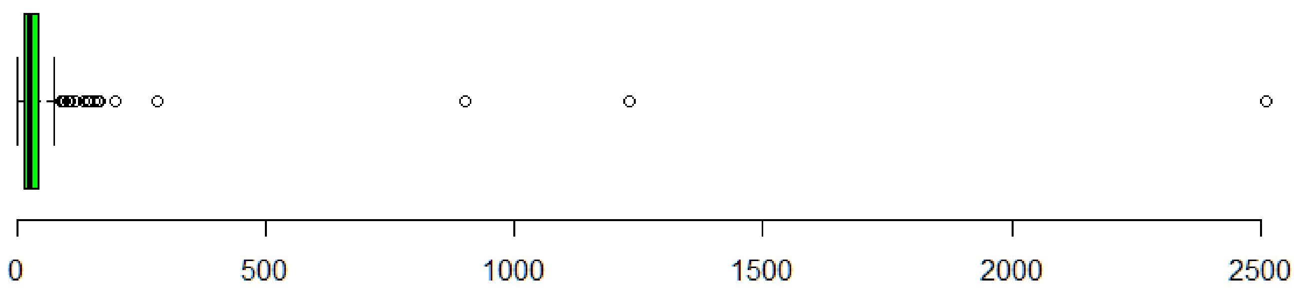
Printing two plots in the same graph





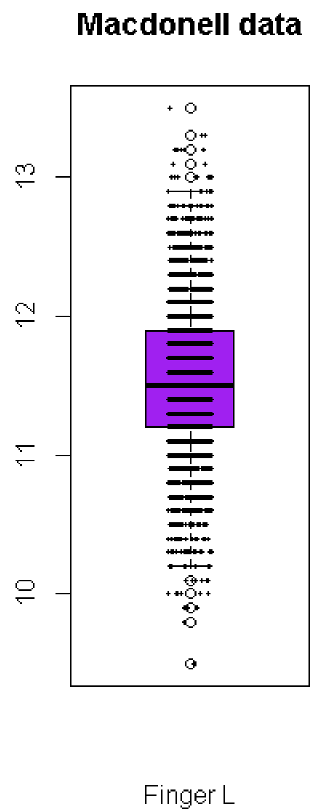

```
#set up graph pane;
par(mfrow=c(2,1));
hist(MacFingL, seq(9,14,0.1), col = "red");
boxplot(MacFingL,horizontal=T, col = "green" );
#reset;
par(mfrow=c(1,1));
```

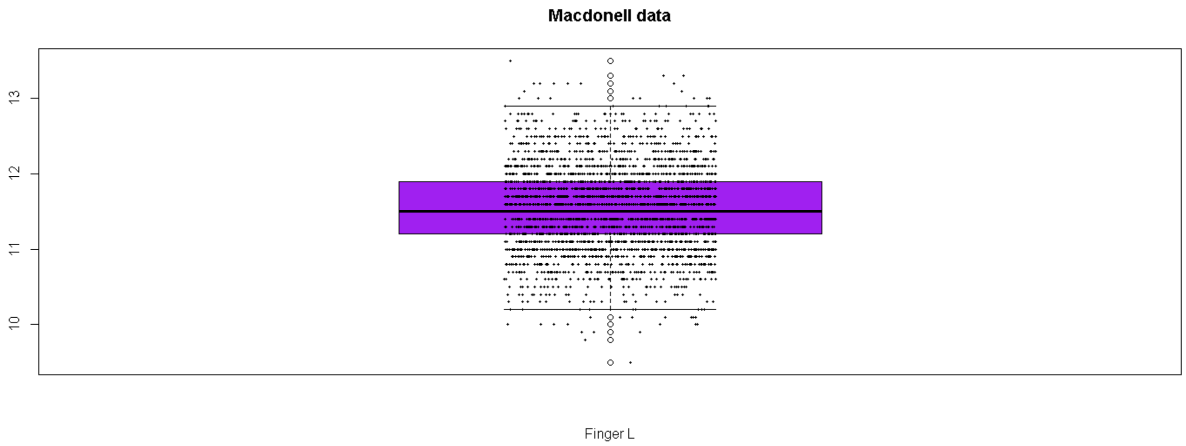





```
par(mfrow=c(2,1));
hist(exec.pay2$salary,col = "green");
boxplot(exec.pay2, horizontal=TRUE,  outline=TRUE, frame=F, col = "green", width = 10 );
#reset;
par(mfrow=c(1,1));
```

Copying and pasting a boxplot





R Studio lets you change the dimensions of the plot before copying.

In MS Office programs you can simply paste OR paste as a picture to preserve font sizes, etc.

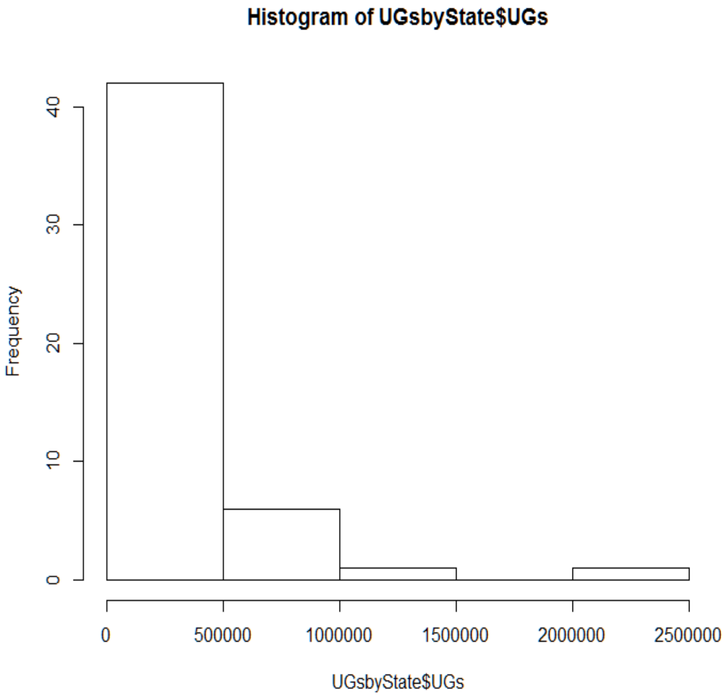
The undergraduate data

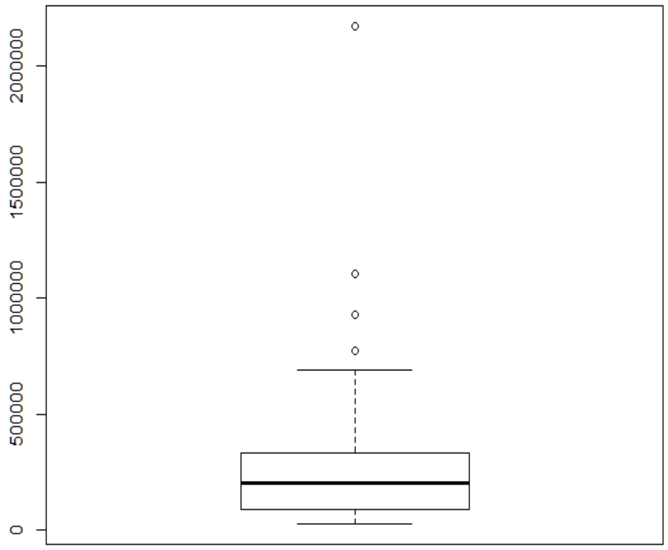
```
UGsbyState <- read.csv("http://math.mercyhurst.edu/~sousley/STAT_139/data/UGsbyState.csv", header = T);
```

UGsbyState

	State	UGs	Pop	UGpT
1	New Jersey	326358	8640218	37.77196
2	Nevada	100760	2484196	40.56041
3	Alaska	27463	676301	40.60766
4	Georgia	378947	9318715	40.66516
5	Connecticut	142926	3487896	40.97771
6	Tennessee	250974	6068306	41.35816
7	Florida	775171	18019093	43.01942
8	South Carolina	187254	4324799	43.29773
9	Maine	58512	1313355	44.55155
10	Hawaii	57527	1275264	45.10988
11	New Hampshire	59405	1308824	45.38807
12	Montana	42990	945428	45.47147
13	Maryland	255933	5602258	45.68390
14	Louisiana	194567	4243634	45.84915
15	Oregon	170742	3680968	46.38508
16	Mississippi	134699	2896713	46.50064
17	Ohio	533652	11458390	46.57304
18	Arkansas	132112	2804199	47.11221
19	Pennsylvania	585006	12388055	47.22339
20	Texas	1104529	23367534	47.26767
21	**New York**	928563	19367028	**47.94556**
22	Alabama	220520	4587564	48.06908
23	West Virginia	87292	1806760	48.31411
24	Idaho	70754	1461183	48.42241
25	North Carolina	436662	8845343	49.36632
26	Washington	314862	6360529	49.50249
27	Delaware	42488	850366	49.96437
28	Indiana	317963	6294124	50.51743
29	Virginia	387593	7628347	50.80957
30	Oklahoma	182340	3568132	51.10237
31	Massachusetts	335511	6443424	52.07030
32	Kentucky	219194	4199440	52.19601
33	Missouri	306201	5832977	52.49481
34	Colorado	255412	4751474	53.75427
35	Illinois	688043	12759673	53.92325
36	Wisconsin	300932	5568505	54.04179
37	Michigan	545001	10083878	54.04677
38	South Dakota	42985	787380	54.59245
39	Minnesota	289018	5143134	56.19492
40	Vermont	34923	620196	56.30962
41	New Mexico	115875	1937916	59.79361
42	Nebraska	105611	1759779	60.01379
43	**California**	**2172354**	36121296	**60.14053**
44	Wyoming	30928	512573	60.33872
45	Kansas	168244	2756267	61.04053
46	Rhode Island	71175	1058991	67.21020
47	North Dakota	44042	636453	69.19914
48	Utah	183518	2585155	70.98917
49	Iowa	212715	2967270	71.68711
50	Arizona	476547	6178251	**77.13299**

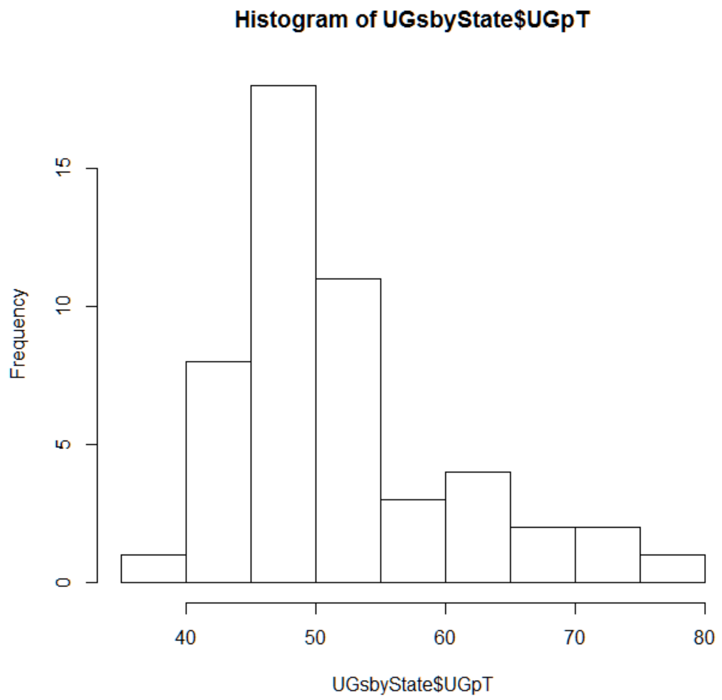
The undergraduate data

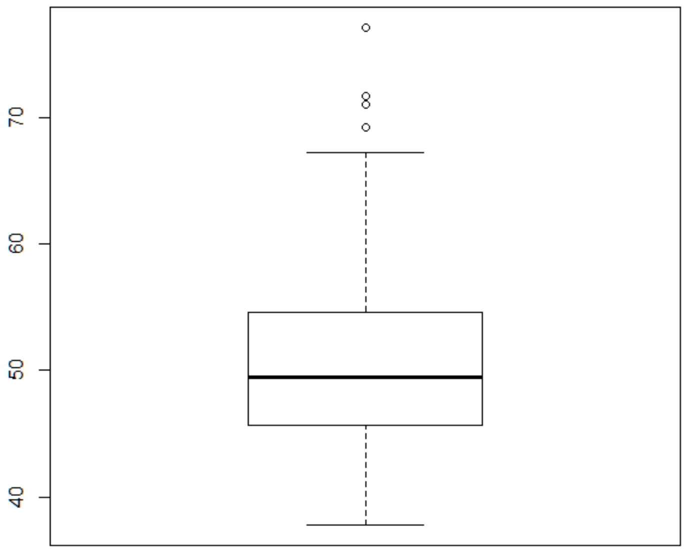





```
boxplot(UGsbyState$UGs)
hist(UGsbyState$UGs)
```

The undergraduate data: per 1000






```
boxplot(UGsbyState$UGpT)
hist(UGsbyState$UGpT)
```

Homework 3

Verzani CH2, Pages 81, 82

2.31 2.32 2.33 2.36

Give your plots new and relevant main titles

Data sets are in same location as other data sets and are vectors, except for normtemp.csv, which is a dataframe.

Vector example: bumpers.vec (scan)

```
bumpers <- scan("http://math.mercyhurst.edu/~sousley/STAT_139/data/bumpers.vec");
```

Dataframe example: normtemp.csv (read.csv)

```
normtemp <- read.csv("http://math.mercyhurst.edu/~sousley/STAT_139/data/normtemp.csv", header=T);
```

and...

Homework 3 (continued)

3.5 For the data sets bumpers, firstchi, math, pi2000, normtemp$temperature, and paradise make boxplots. Give each plot a specific and relevant title and make each in a different color. Are they consistent with the histograms of each you made for the previous homework?

3.6 Add a stripchart to the pi2000 and bumpers boxplots. What additional information do they reveal?

Due on Feb 6 before class

- same Email and Word format as before.

- I will try to get back Homework 1 very soon