DATA 500
Lecture 4

Graphical Methods I: Plots

Pie charts, Bar charts

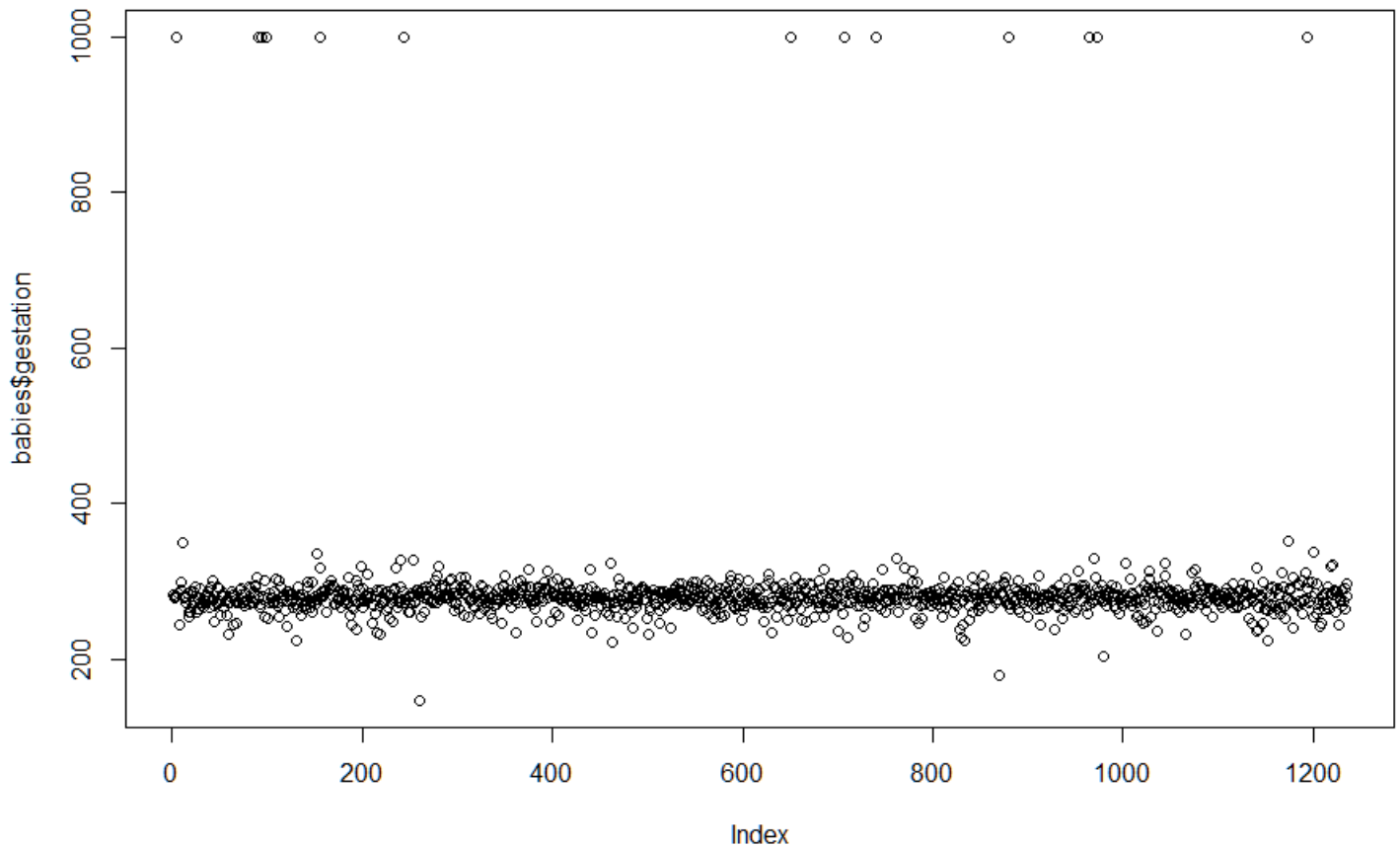
Load the data in babies dataframe

```
babies <- read.csv("http://math.mercyhurst.edu/~sousley/STAT_139/data/babies.csv", header=T);
str(babies) # get the structure
'data.frame':    1236 obs. of  23 variables:
 $ id      : int  15 20 58 61 72 100 102 129 142 148 ...
 $ pluralty: int  5 5 5 5 5 5 5 5 5 5 ...
 $ outcome : int  1 1 1 1 1 1 1 1 1 1 ...
 $ date    : int  1411 1499 1576 1504 1425 1673 1449 1562 1408 1568 ...
 $ gestdays: int  284 282 279 999 282 286 244 245 289 299 ...
 $ sex     : int  1 1 1 1 1 1 1 1 1 1 ...
 $ bwt     : int  120 113 128 123 108 136 138 132 120 143 ...
 $ parity  : int  1 2 1 2 1 4 4 2 3 3 ...
 $ mrace   : int  8 0 0 0 0 0 7 7 0 0 ...
 $ mage    : int  27 33 28 36 23 25 33 23 25 30 ...
 $ med     : int  5 5 2 5 5 2 2 1 4 5 ...
 $ mht     : int  62 64 64 69 67 62 62 65 62 66 ...
 $ mwt     : int  100 135 115 190 125 93 178 140 125 136 ...
 $ frace   : int  8 0 5 3 0 3 7 7 3 0 ...
 $ fage    : int  31 38 32 43 24 28 37 23 26 34 ...
 $ fed     : int  5 5 1 4 5 2 4 4 1 5 ...
 $ fht     : int  65 70 99 68 99 64 99 71 70 99 ...
 $ fwt     : int  110 148 999 197 999 130 999 192 180 999 ...
 $ marital : int  1 1 1 1 1 1 1 1 0 1 ...
 $ inc     : int  1 4 2 8 1 4 98 2 2 2 ...
 $ msmoke  : int  0 0 1 3 1 2 0 0 0 1 ...
 $ time    : int  0 0 1 5 1 2 0 0 0 1 ...
 $ number  : int  0 0 1 5 5 2 0 0 0 4 ...
```

gestation

We can do plots of gestation data

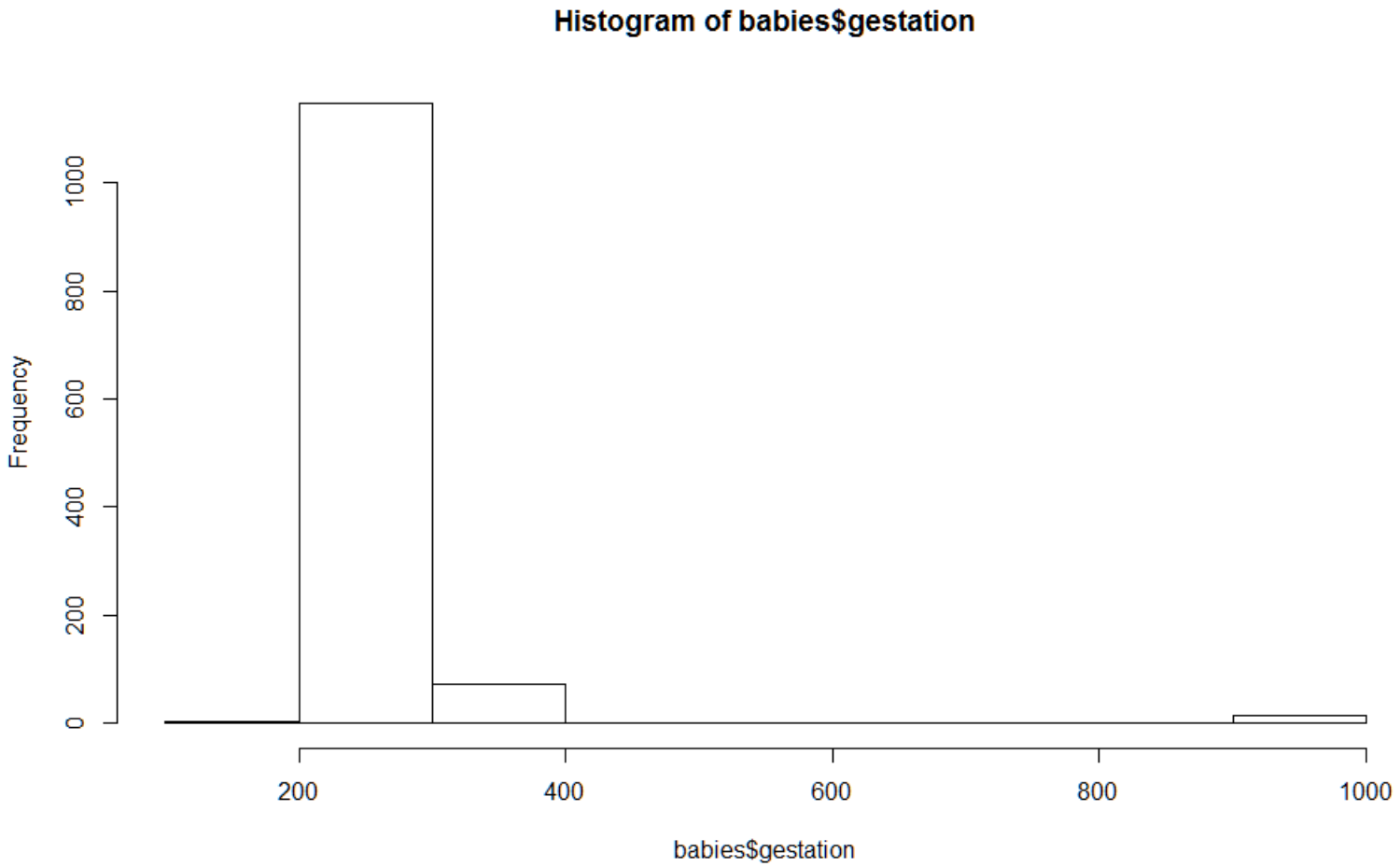
```
plot(babies$gestdays)
```



gestation

We can do plots of gestation data: histogram

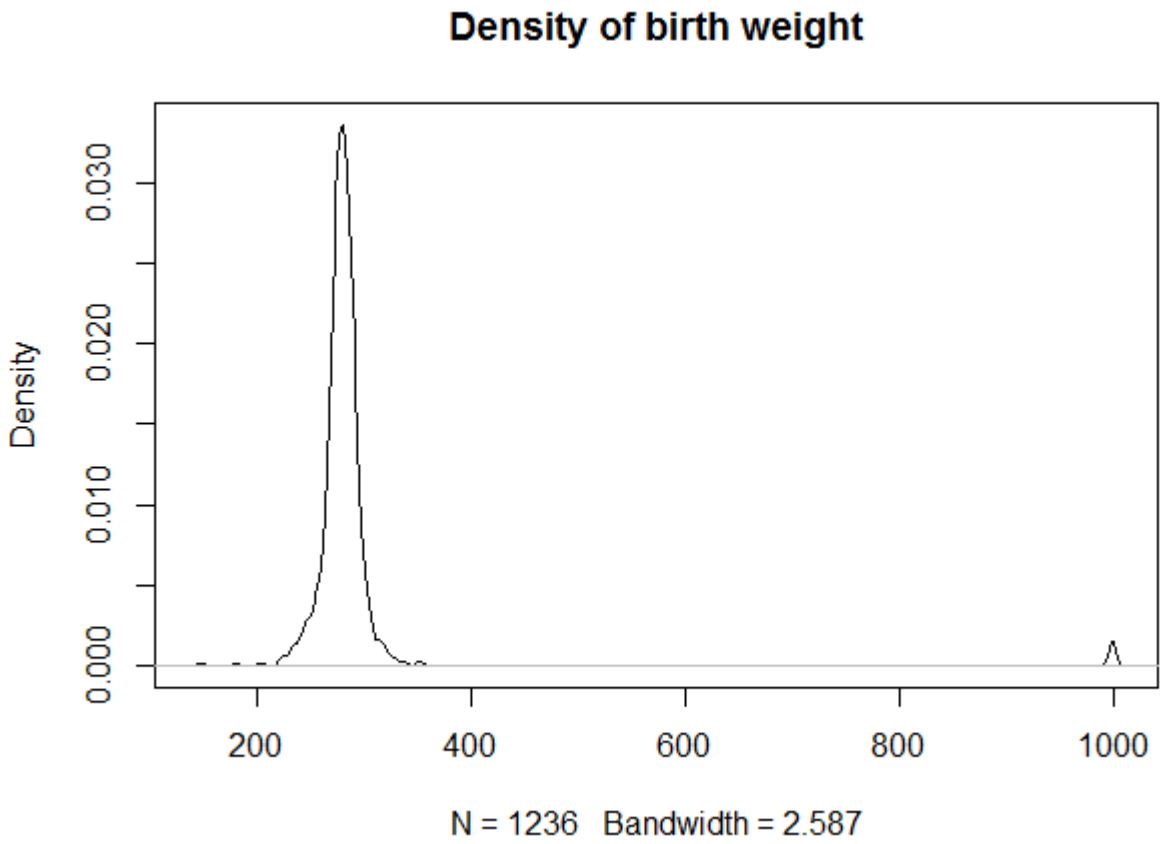
```
hist(babies$gestdays)
```



gestation

We can do a (smoothed) density plot of gestation data

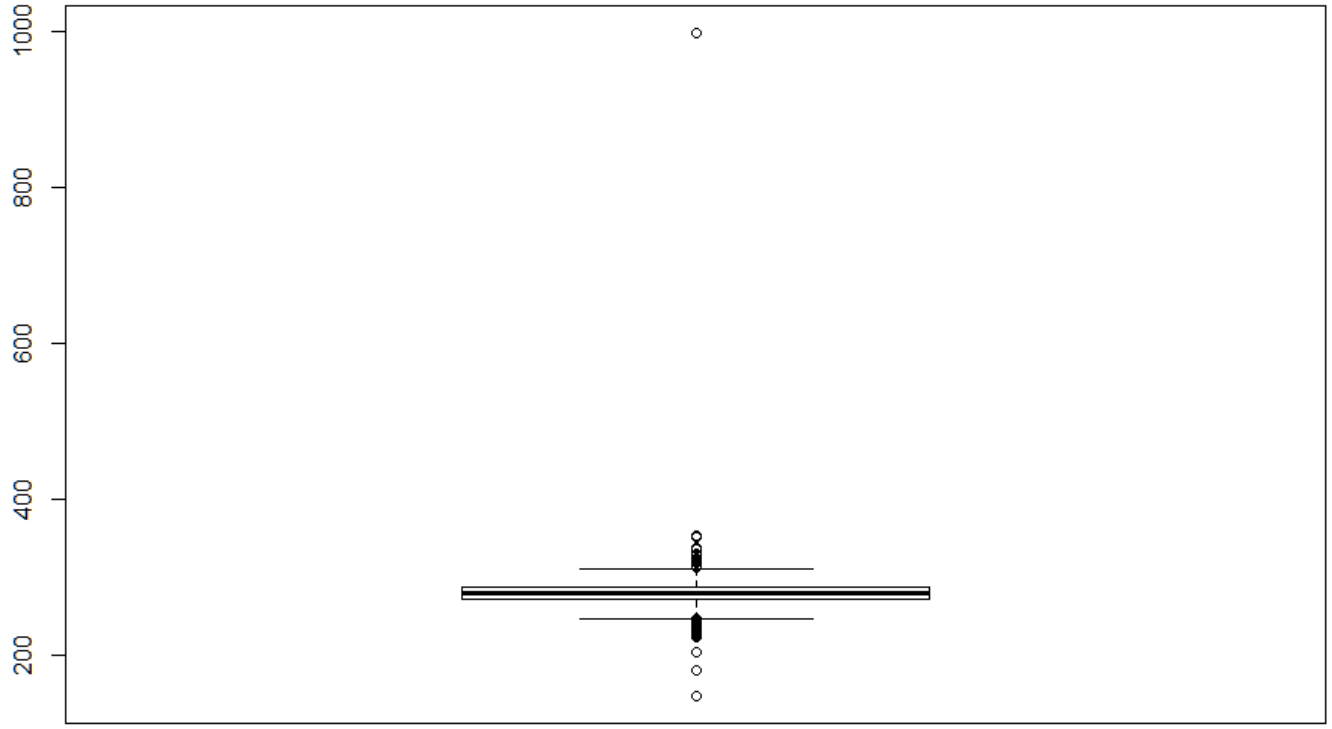
```
plot(density(babies$gestdays), main = "Density of birth weight")
```



gestation

We can do plots of gestation data: boxplot

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



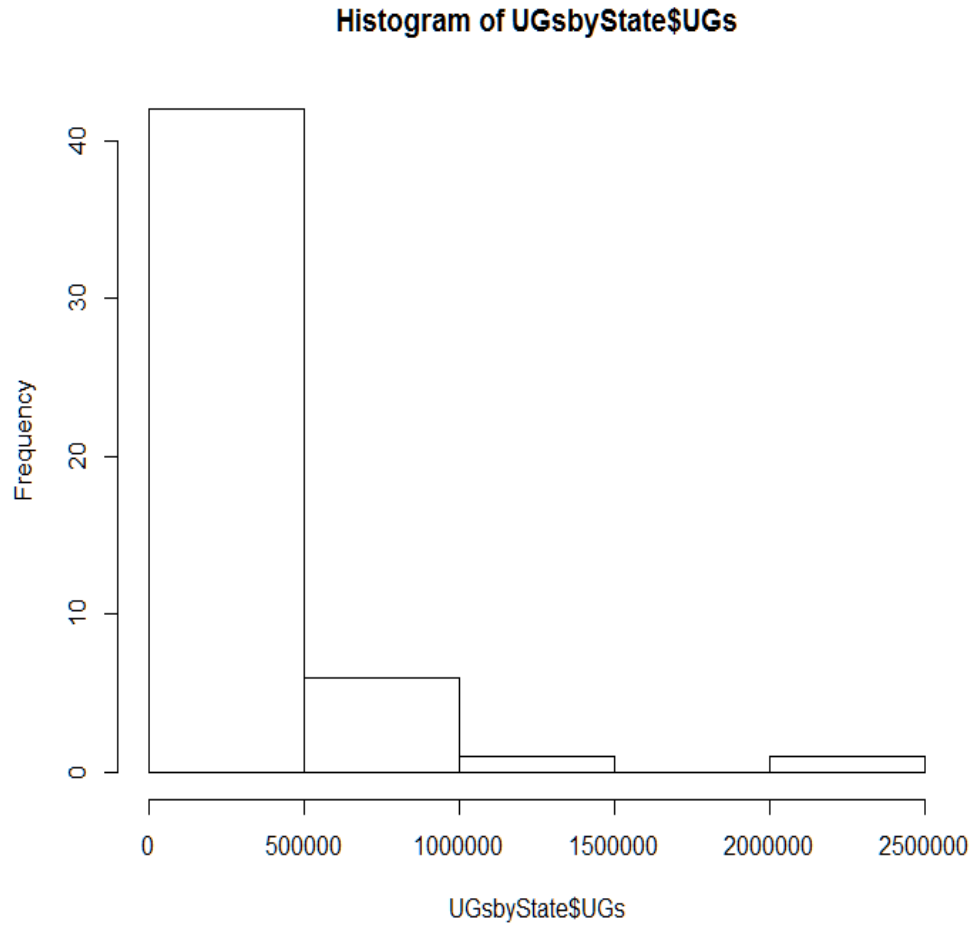

```
summary(babies$gestdays)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  148.0   272.0   280.0   286.9   288.0   999.0
```

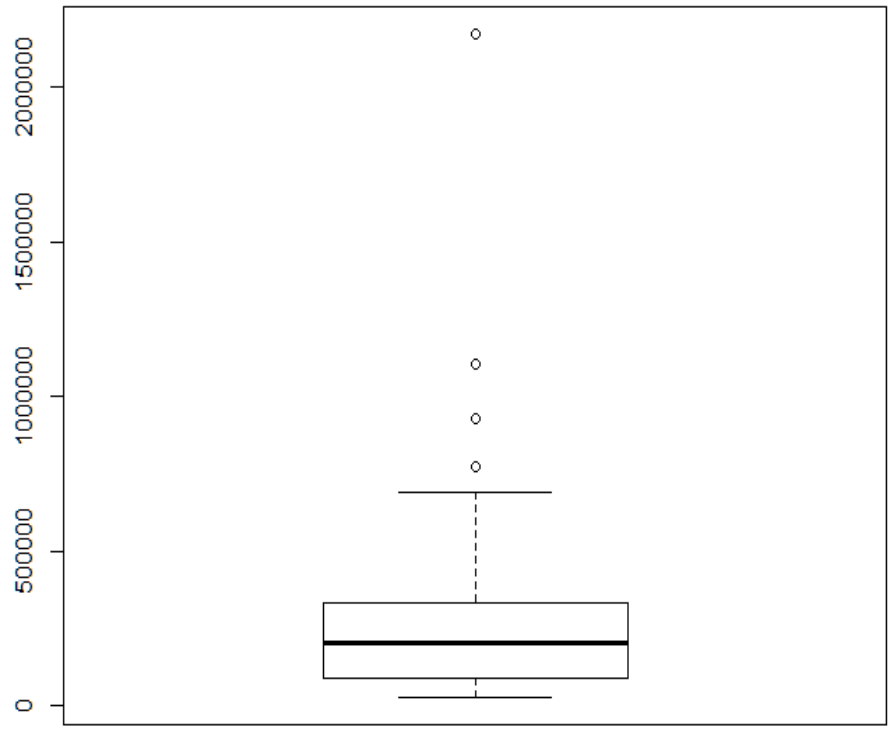
The undergraduate data

```
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

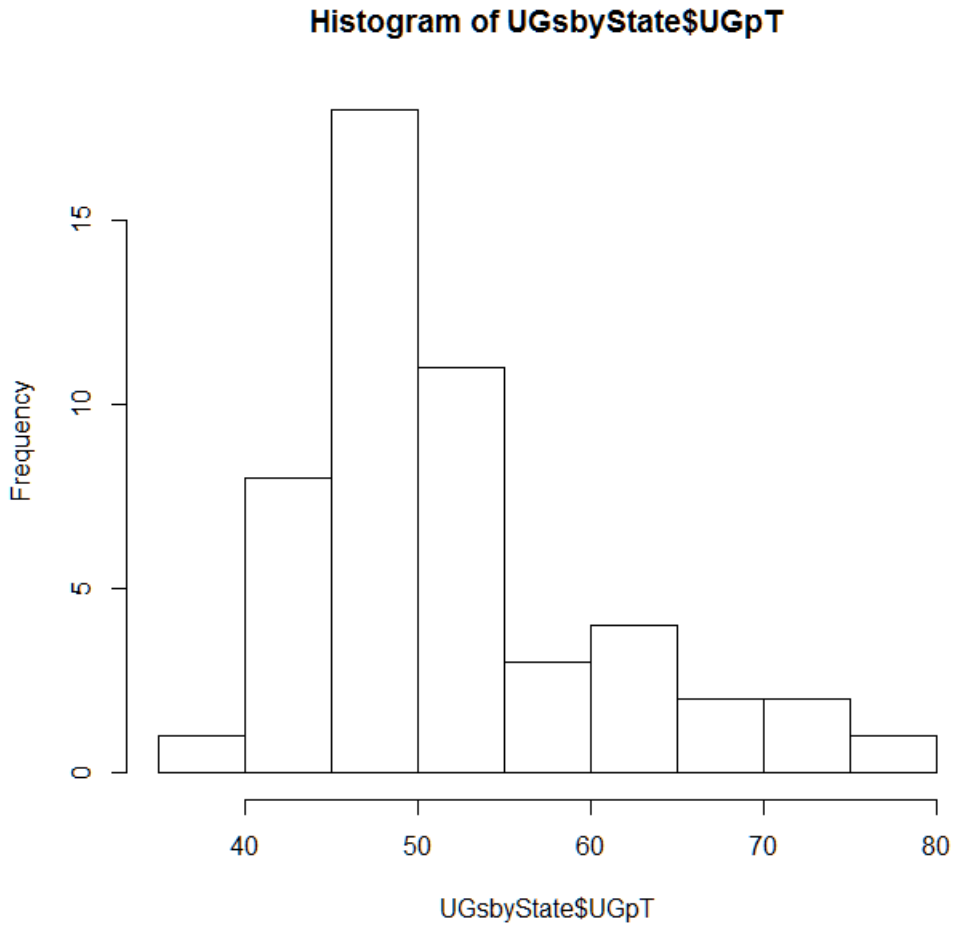
The undergraduate data

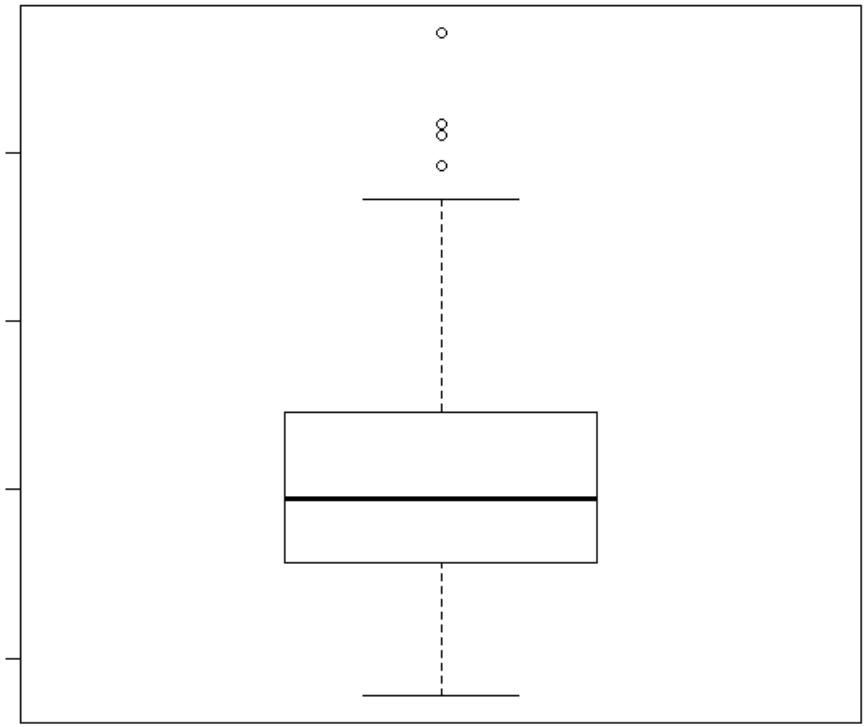





```
boxplot(UGsbyState$UGs)
hist(UGsbyState$UGs)
```

The undergraduate data: per 1000





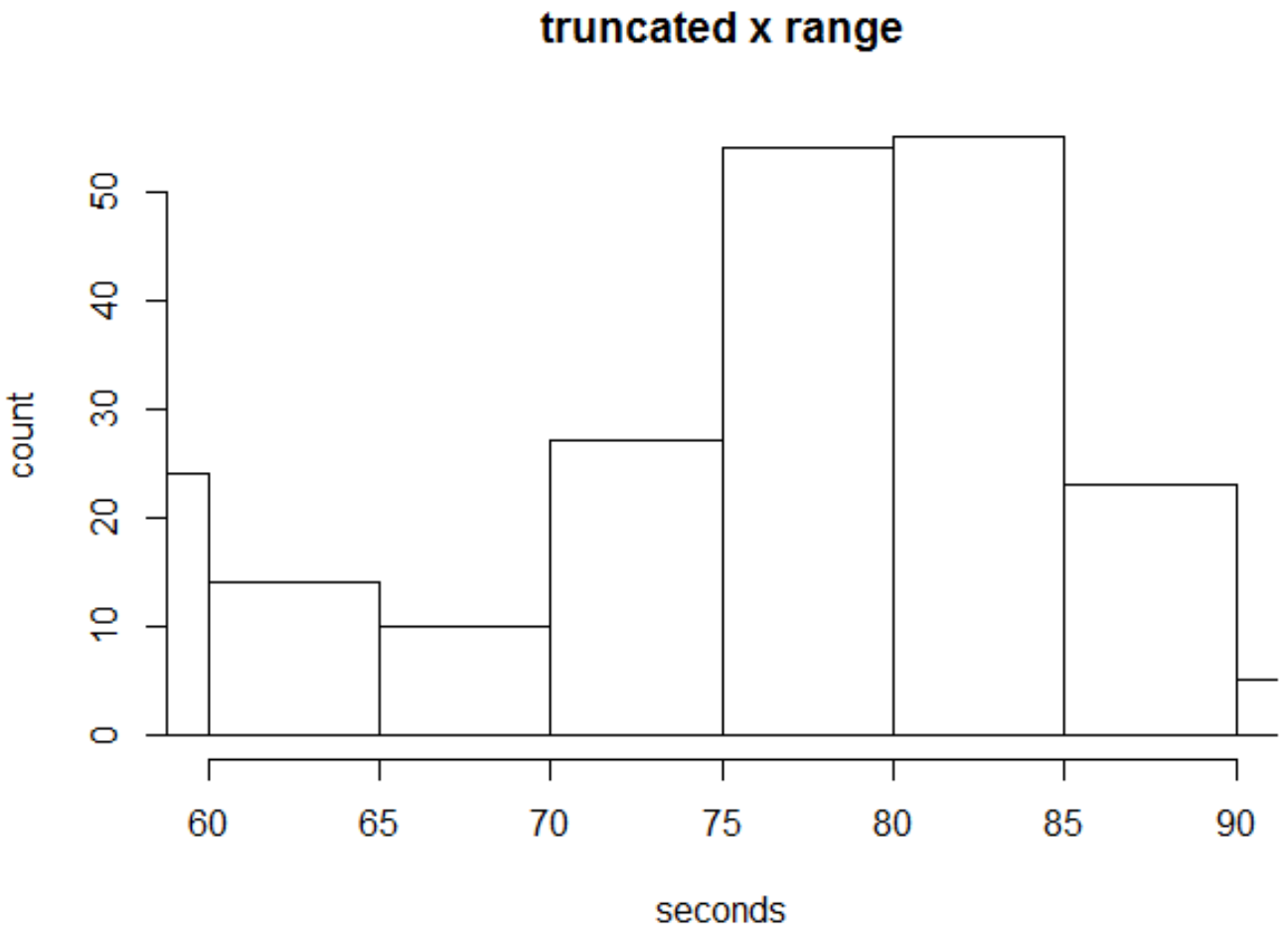
```
boxplot(UGsbyState$UGpT)
hist(UGsbyState$UGpT)
```

Extra arguments (specs) for plots

Argument	Description
`xlim`	Set x coordinate range.
`ylim`	Set y coordinate range.
`xlab`	Set label for x axis.
`ylab`	Set label for y axis.
`main`	Set main title.
`pch`	Adjust plot symbols (`?pch`).
`cex`	Adjust size of text and symbols on a graphic.
`col`	Adjust color of objects drawn (`?colors`).
`lwd`	Adjust width of lines drawn.
`lty`	Adjust how line is drawn. Can be `"blank"`, `"solid"`, `"dashed"`, `"dotted"`, `"dotdash"`, etc.
`bty`	Adjust box type, if drawn. One of `"o"`, `"l"`, `"7"`, `"c"`, `"u"`, or `"]"`.

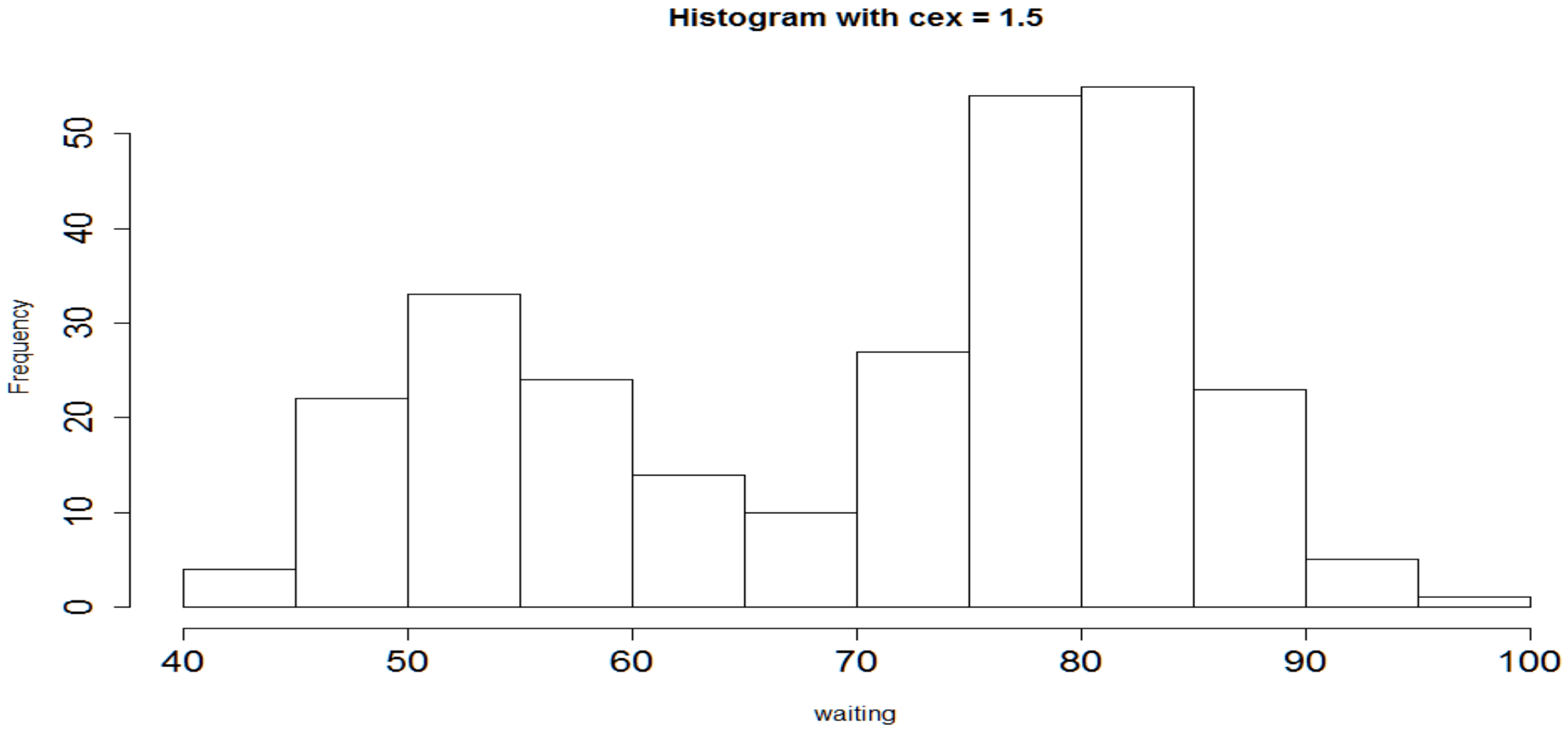
Table 2.4: Standard plotting arguments to modify a graphic.

Extra arguments (specs) for plots




```
# old faithful data ;  attach() lets us refer to column names directly; faithful$waiting
attach(faithful);
hist(waiting); # generic, defaults;
hist(waiting, main = 'truncated x range', xlab = 'seconds', ylab = 'count', xlim = c(60,90) );
```

Extra arguments (specs) for plots

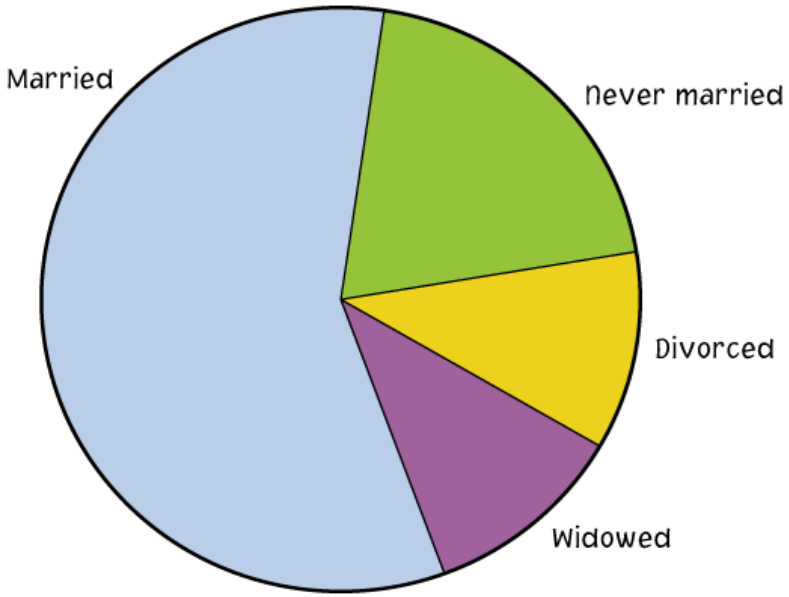



```
# cex magnifies the text size, 1 = 100%, 1.5 = 150% of normal size;
hist(waiting, main = 'Histogram with cex = 1.5', cex.axis = 1.5 );
```

Categorical Variables

The **distribution of a categorical variable** lists the categories and gives the **count** or **percent** of individuals who fall into each category.

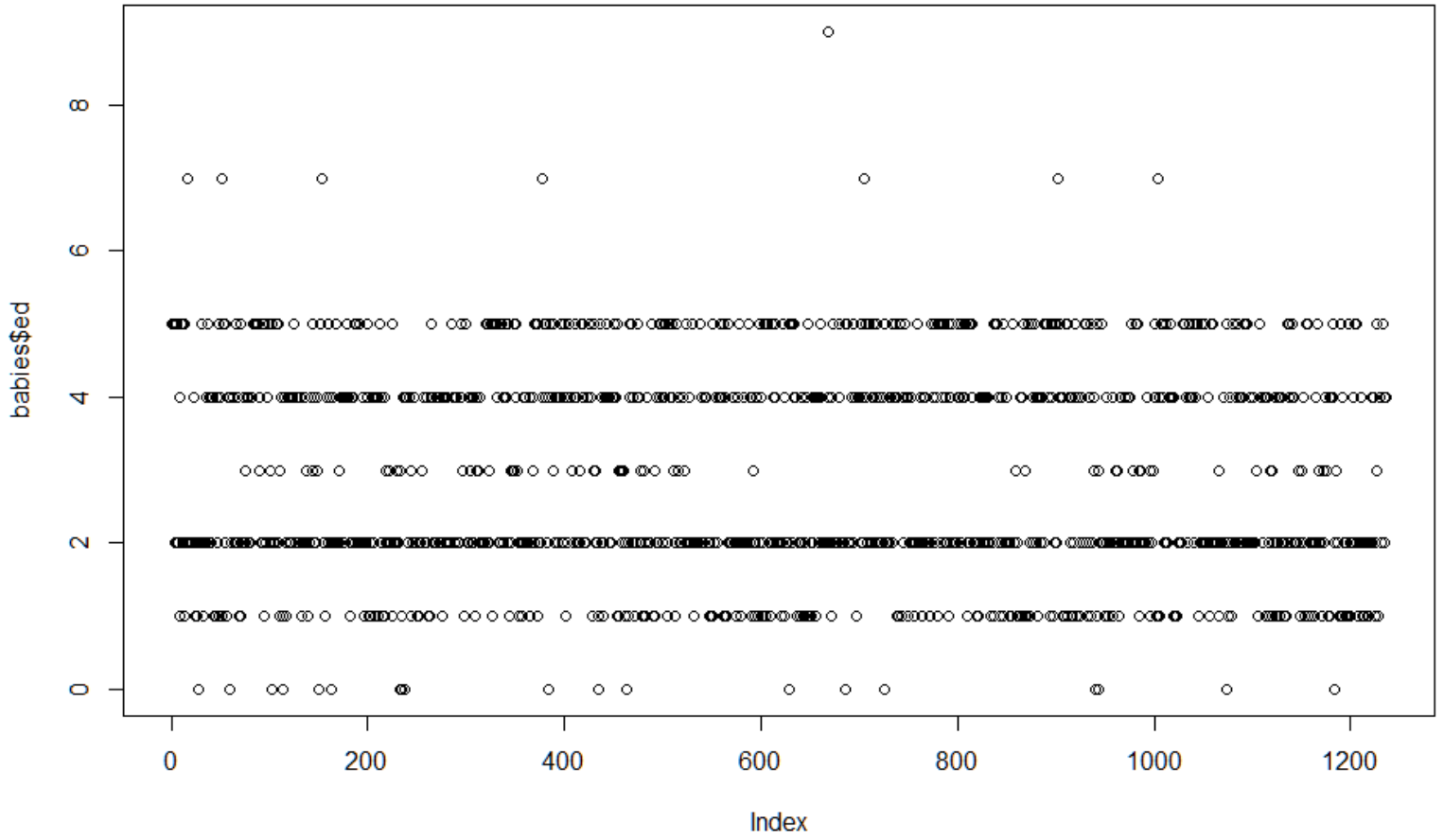
- **Pie charts** show the distribution of a categorical variable as a “pie” whose slices are sized by the counts or percents for the categories.
- **Bar graphs** represent **categories** as bars whose heights show the category counts or percents.



Categorical data: mother's education (med)

We can do plots of education codes

```
plot(babies$med)
```

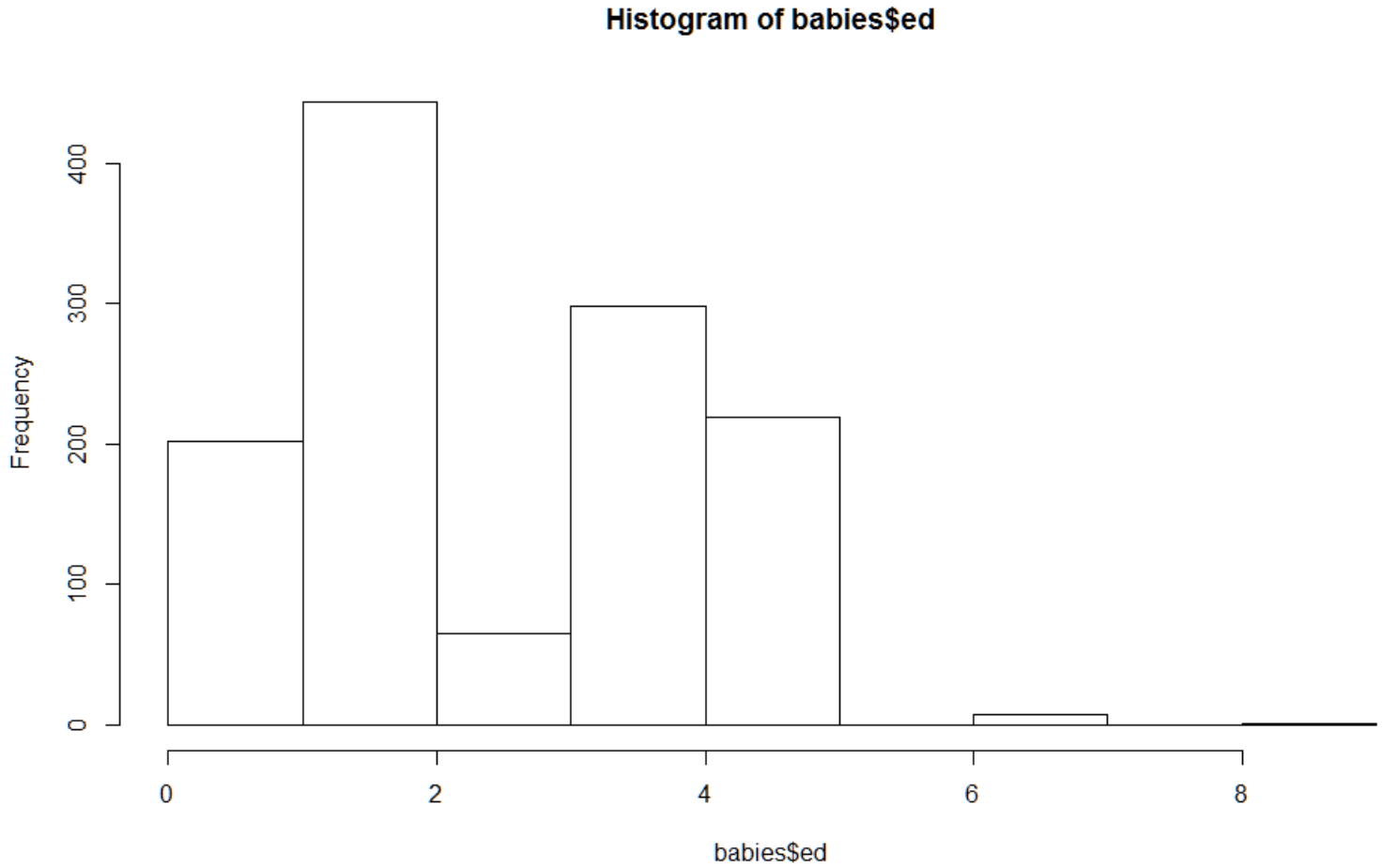


education

We can do plots of education codes

A histogram is best for numeric data - these are categories.

```
hist(babies$med)
```



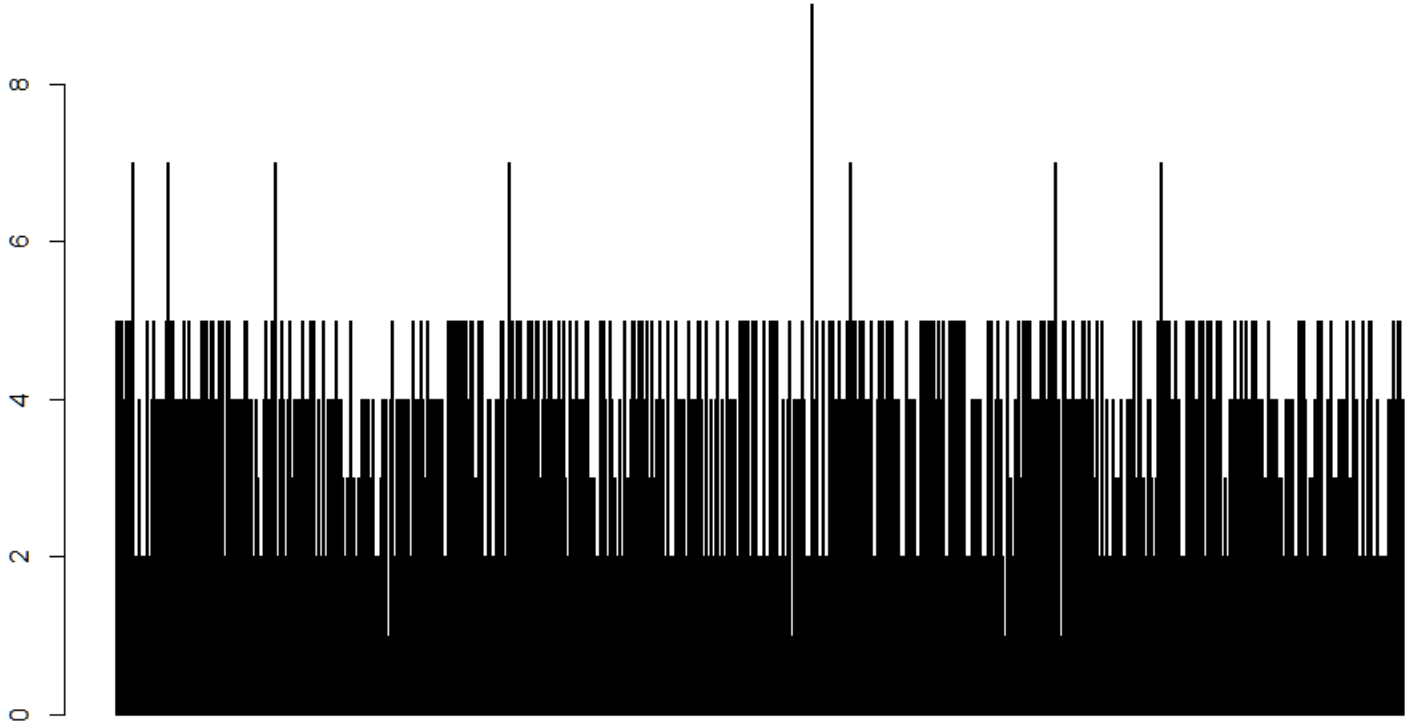
education

We can do plots of education codes

- but they are categories, not measurements

so we want a bar chart

```
barplot(babies$med)
```



education: bar chart

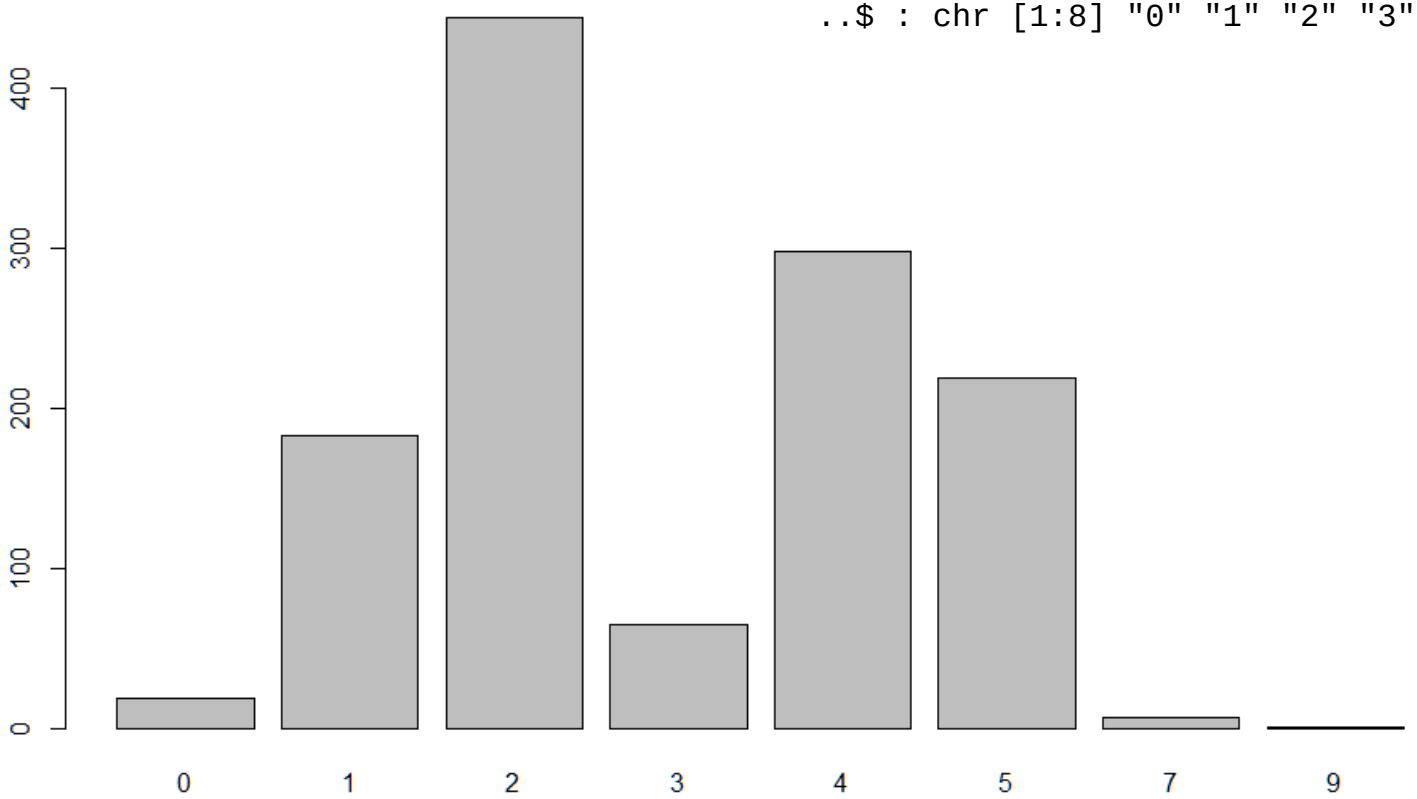
We need the counts of each education code for a bar chart

```
bcounts <- table(babies$med)
bcounts
```

```
  0   1   2   3   4   5   7   9
 19 183 444  65 298 219   7   1
```

```
barplot(bcounts)
```

```
str(bcounts)
 'table' int [1:8(1d)] 19 183 444 65 298 219 7 1
 - attr(*, "dimnames")=List of 1
  ..$ : chr [1:8] "0" "1" "2" "3" ...
```



Pie Charts:
Always ask why, just say NO!

The final sample which we used in our survey included 441 radiographs of patients aged between 5.5 and 14.5 years (218 girls and 223 boys, Fig.1).

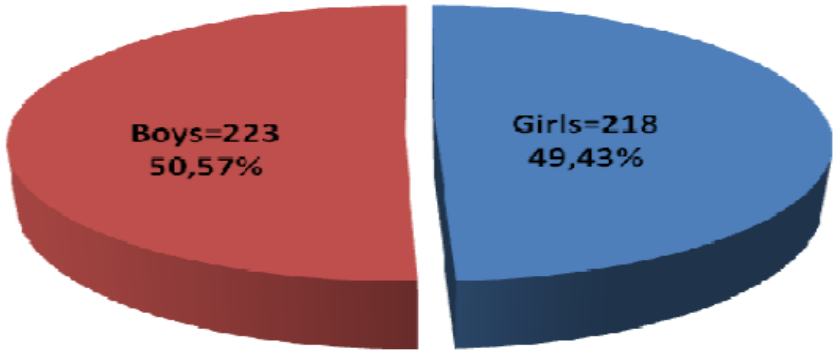


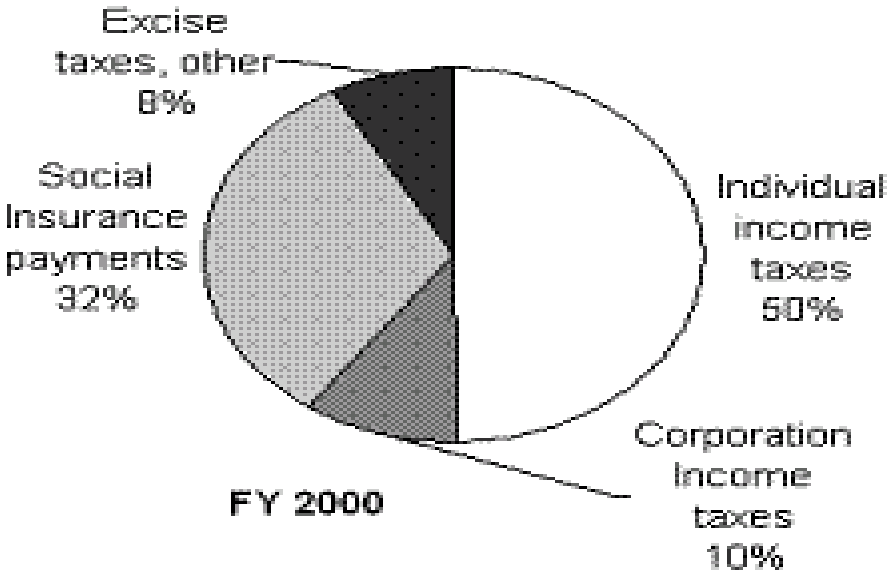
Fig.1 The distribution of children by gender

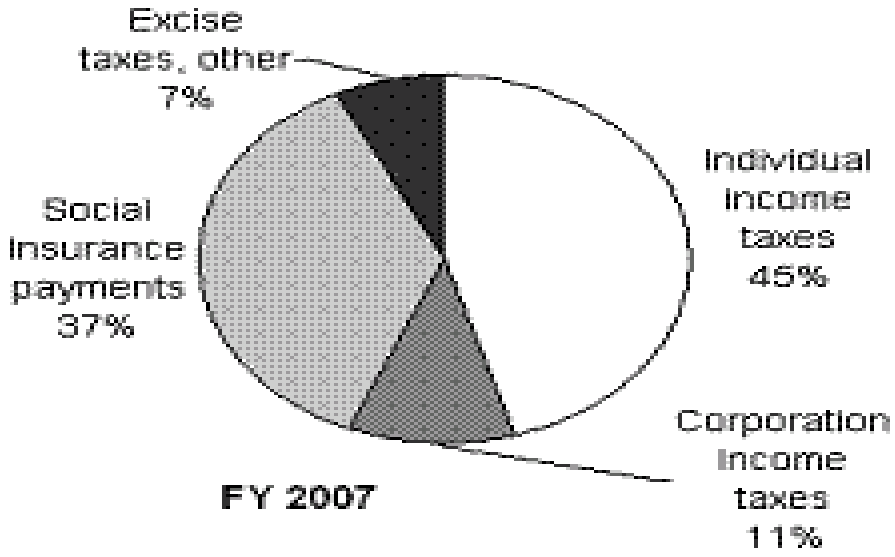
For 218 girls, average age is *10.03* with standard deviation of *2.32* and the 95% confidence interval for the chronologic age is (5.57, 14.49) years (Fig.2).

For 223 boys, average age is 9.73 with standard deviation of 2.14 and 95% confidence interval for the chronologic age is (5.45,14.01)years (Fig.2).

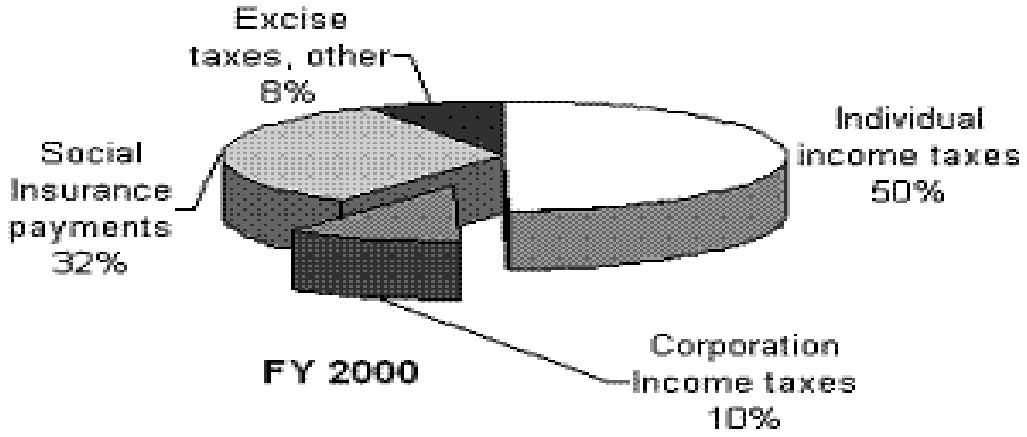
(Ogodescu et al. 2011)

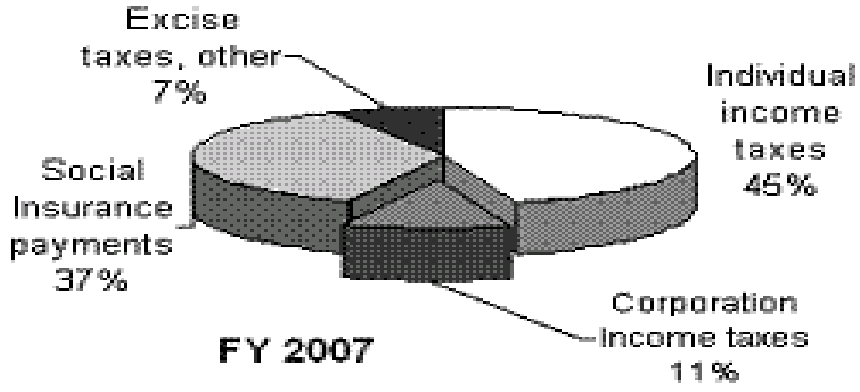
Federal Government Receipts by Source





source: 2007 US Budget, Historical Tables

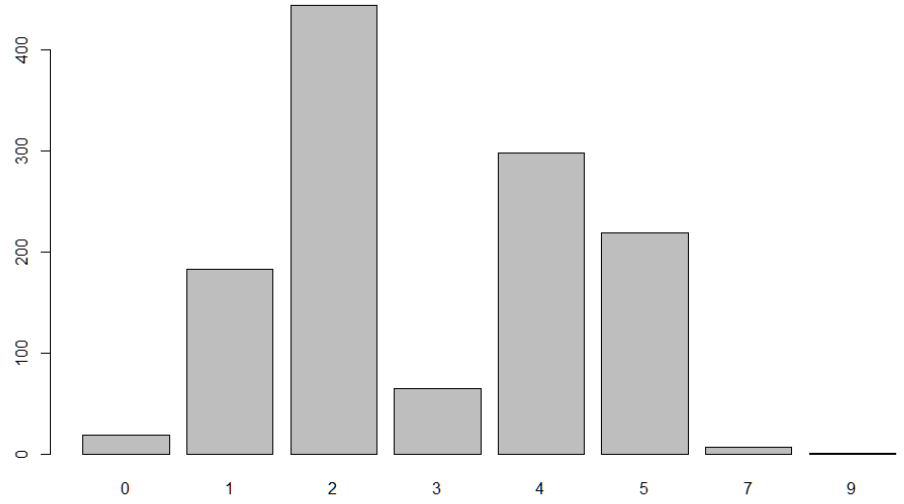


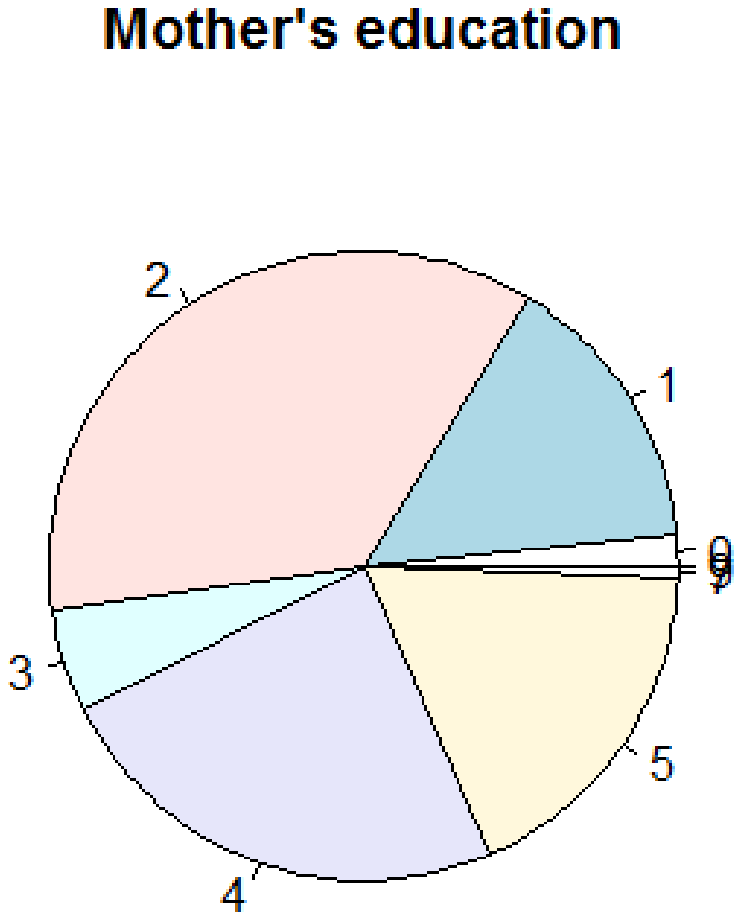


education: pie chart

We need the counts of each education code for a bar chart

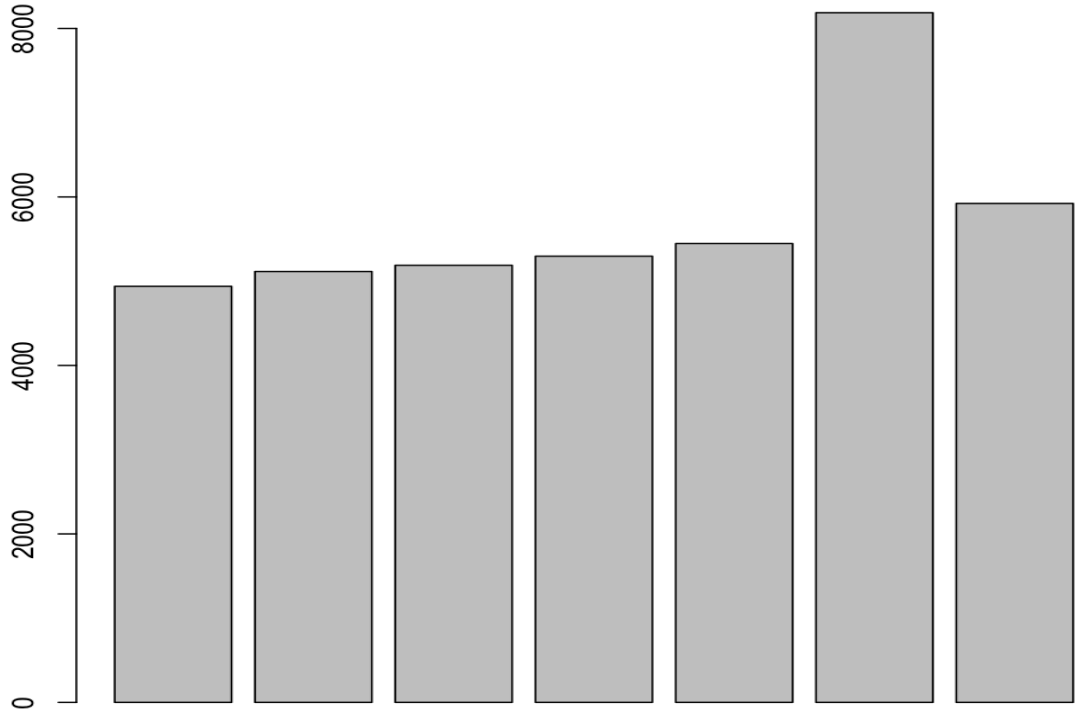
```
bcounts
  0   1   2   3   4   5   7   9
 19 183 444  65 298 219   7   1
pie(bcounts, main = "Mother's education")
```

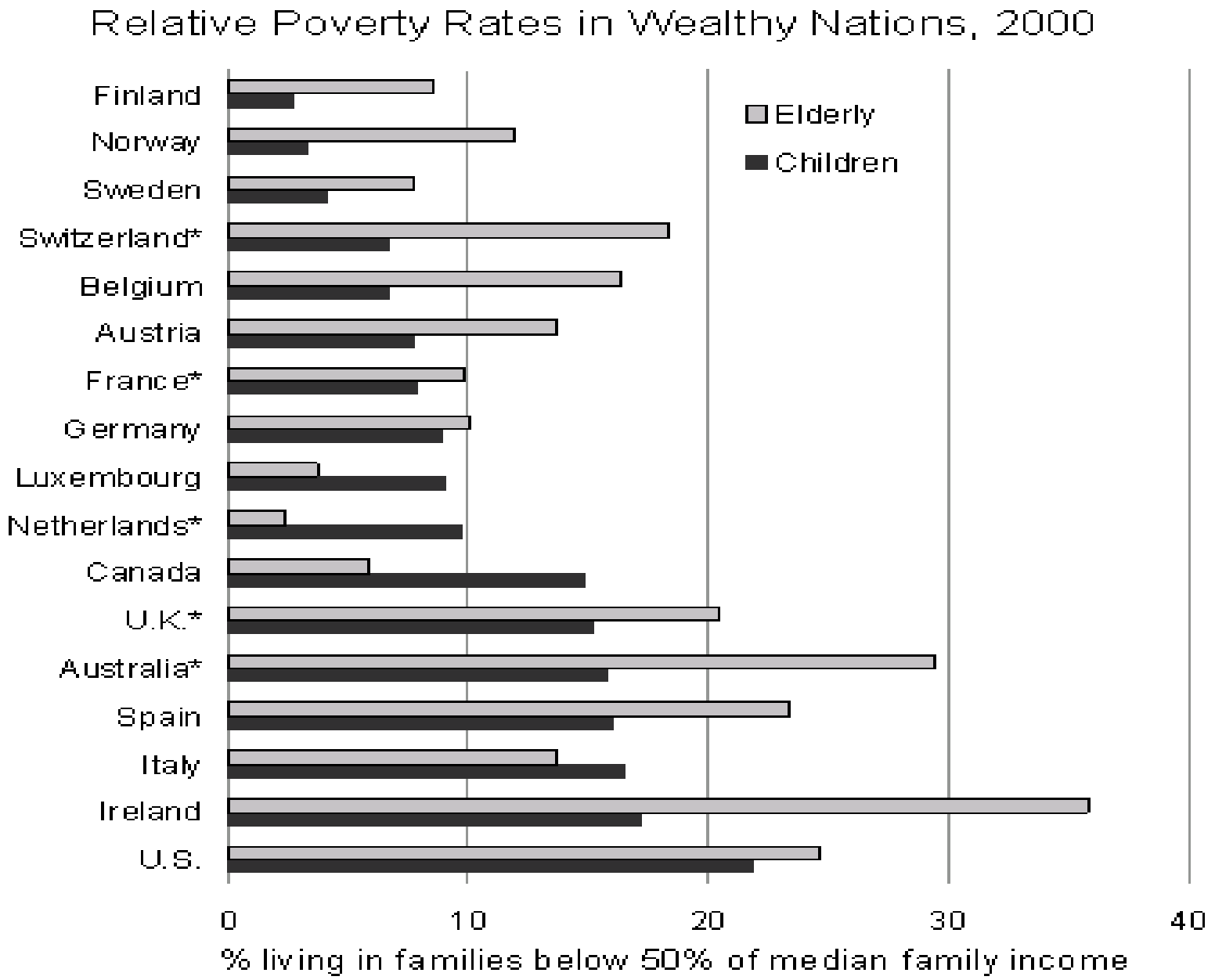




Bar Plots: A Waste of Time and …

ink? (and pixels?)
Simple data is better in a table
Uncertainty in estimates is unclear



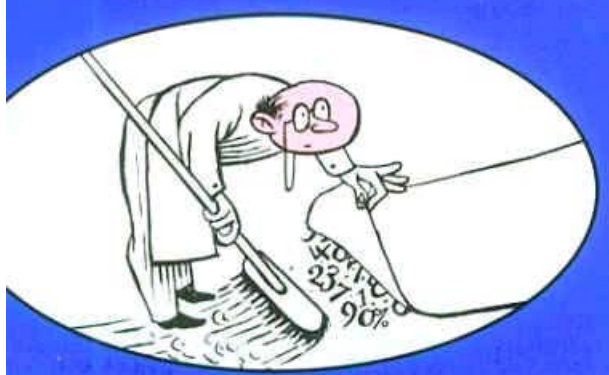


* most recent year
source: Luxembourg Income Study

HOW TO LIE WITH STATISTICS

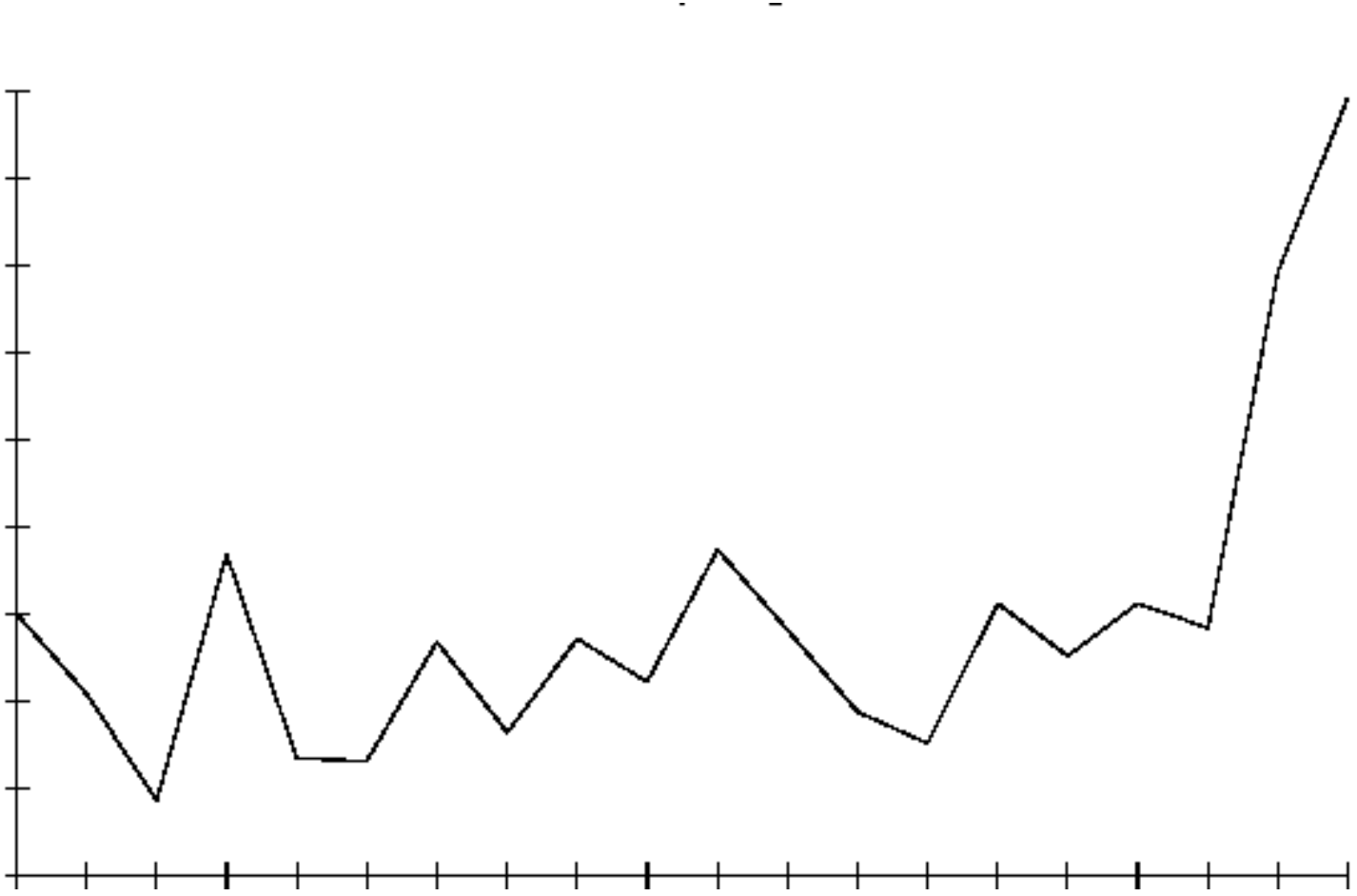
Darrell Huff

Illustrated by Irving Geis

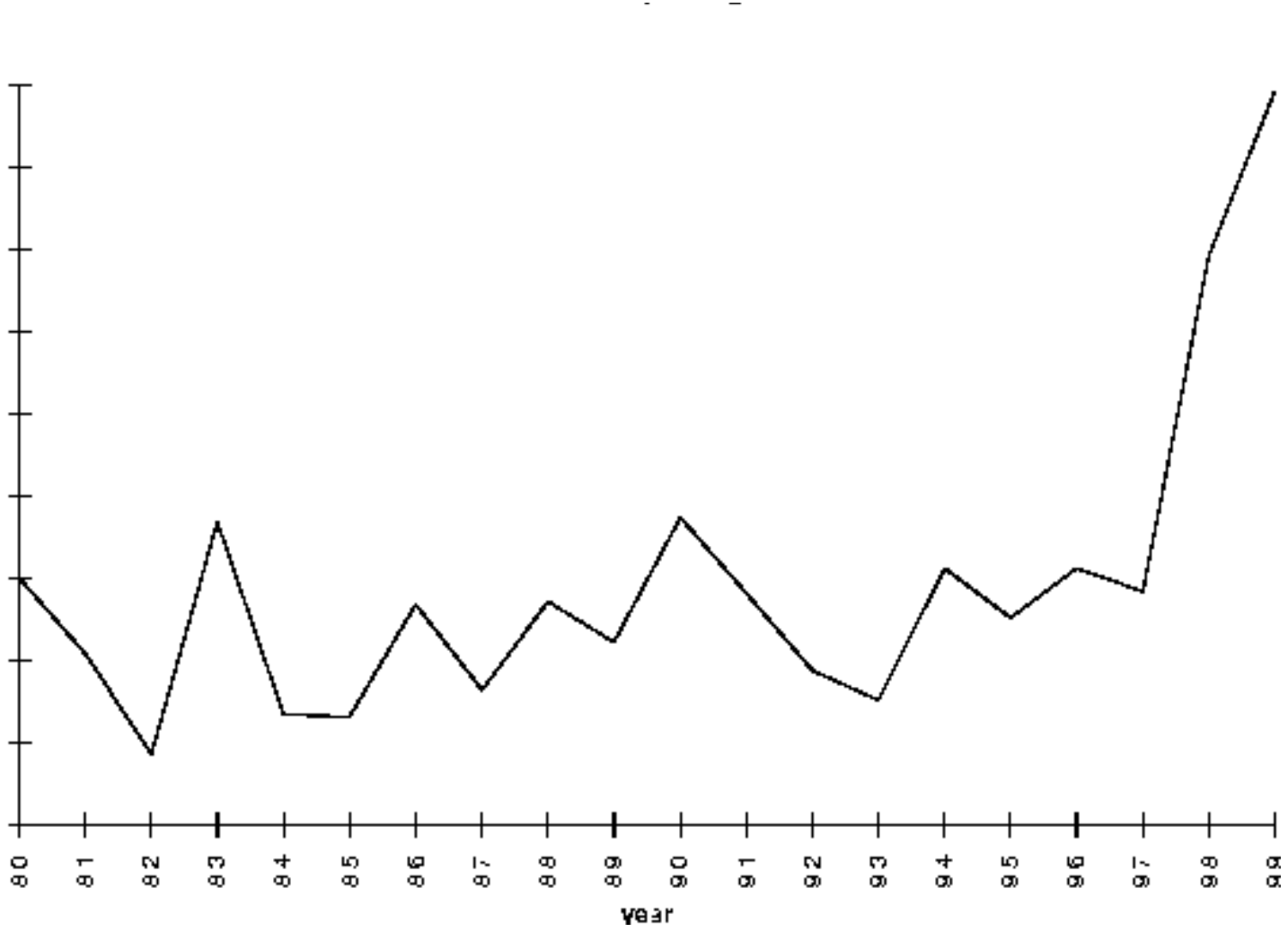


Over Half a Million Copies Sold—
An Honest-to-Goodness Bestseller

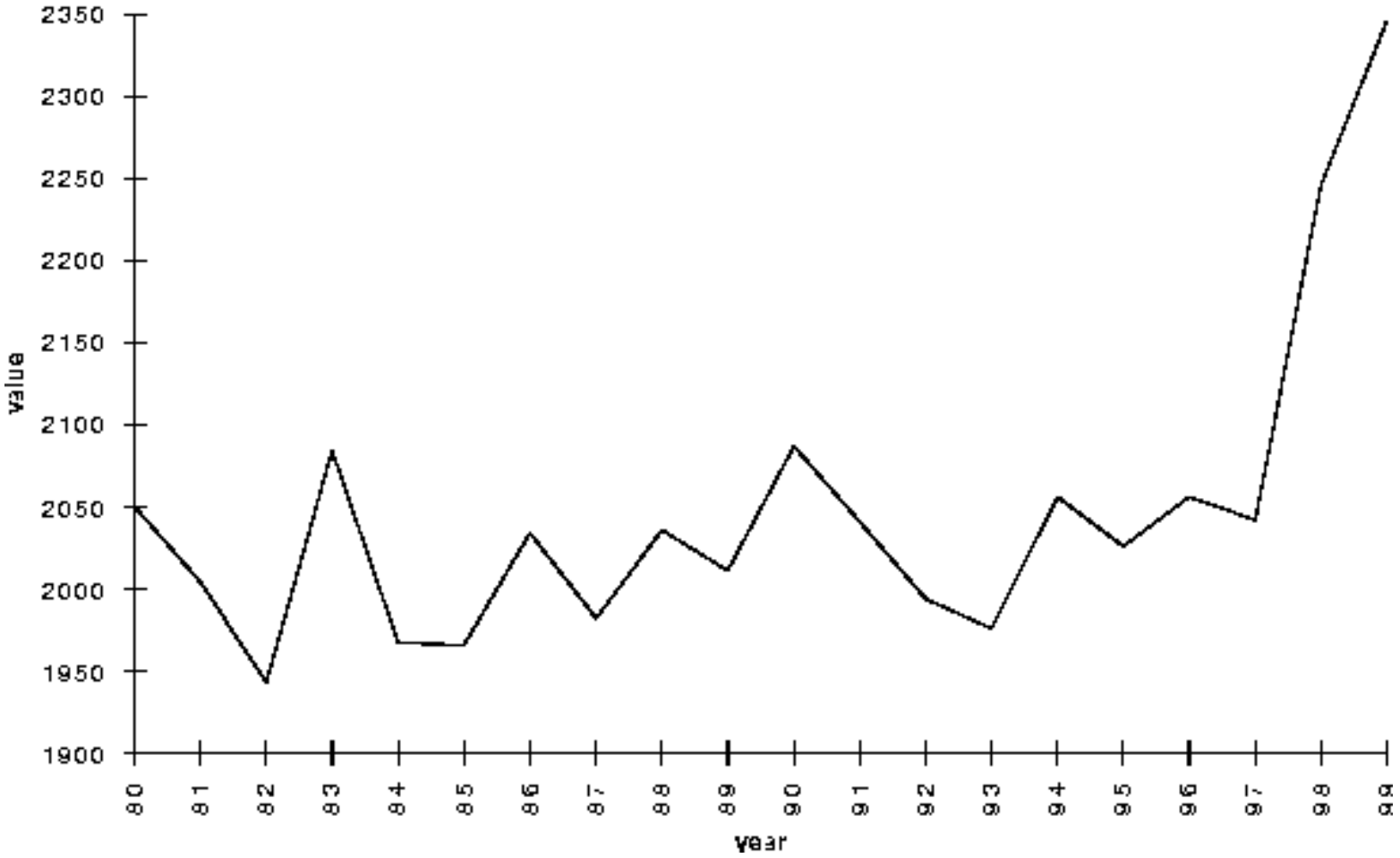
A sudden and dramatic jump! (?)



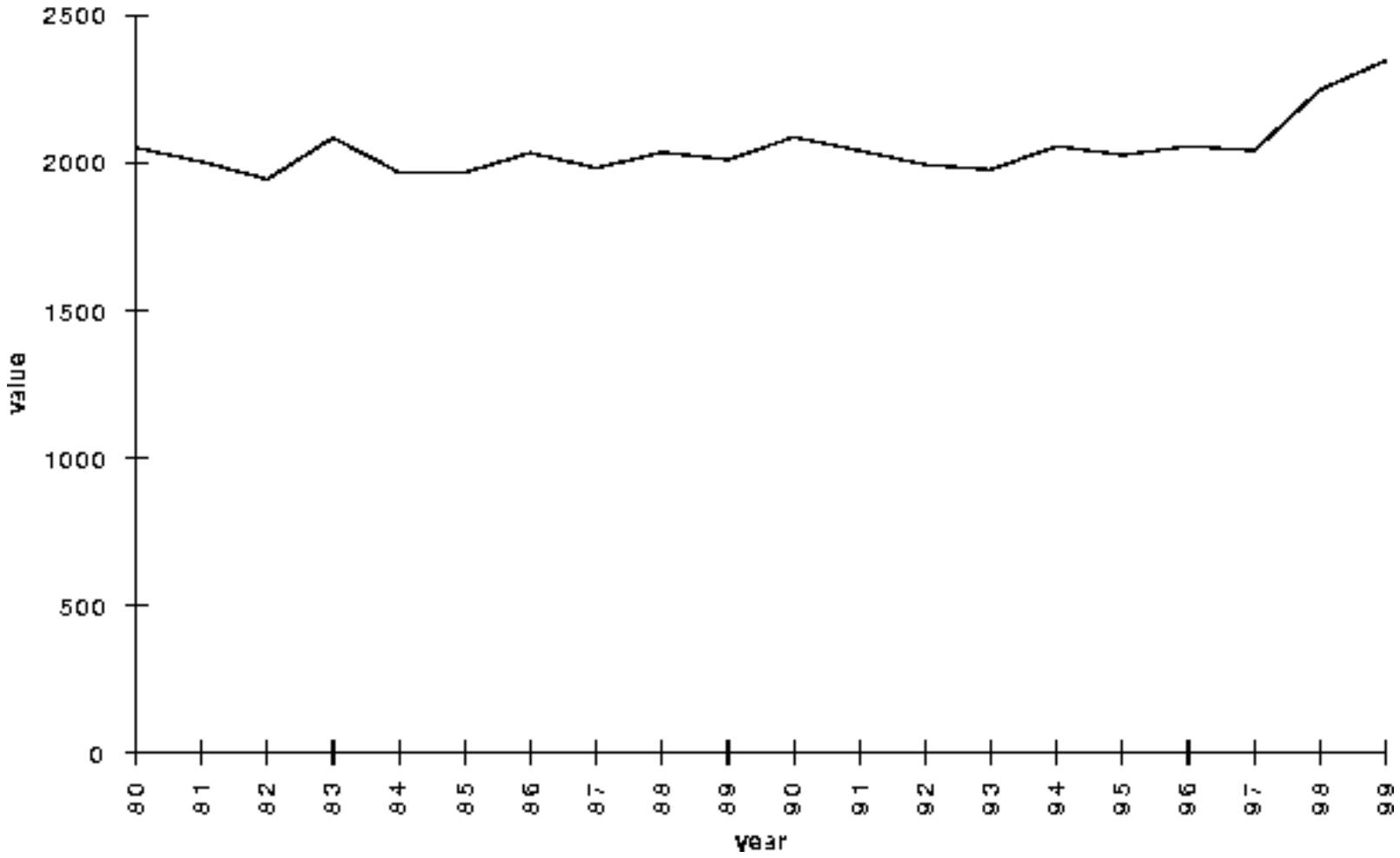
A sudden and dramatic jump! (?)



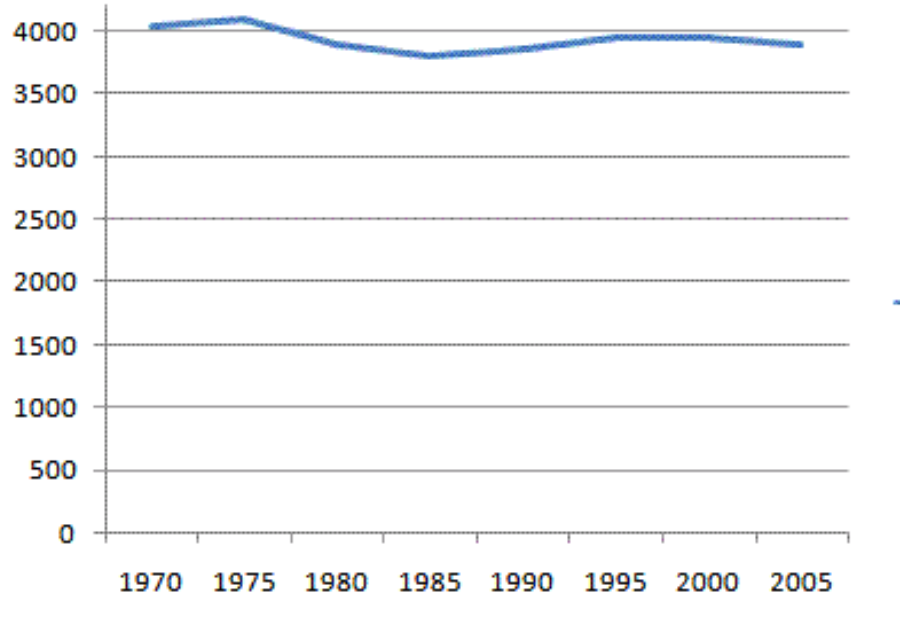
A sudden and dramatic jump?

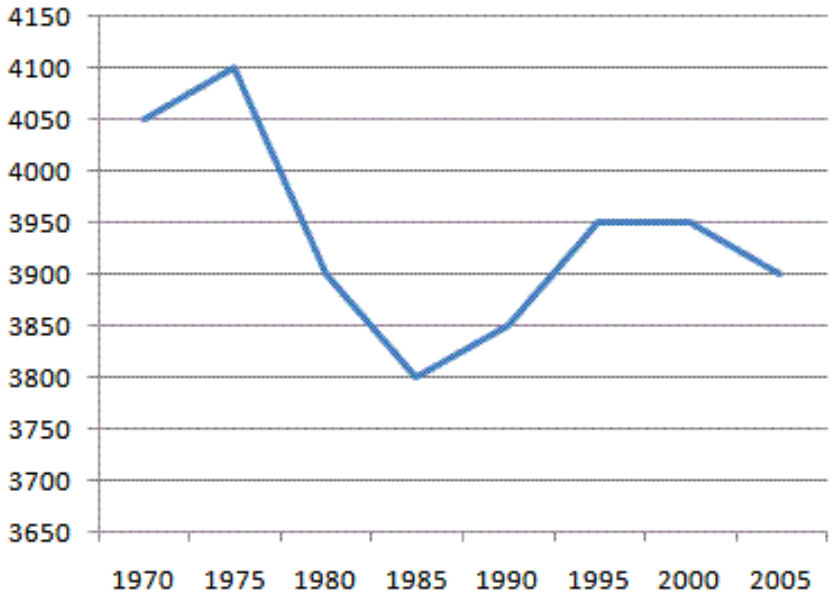


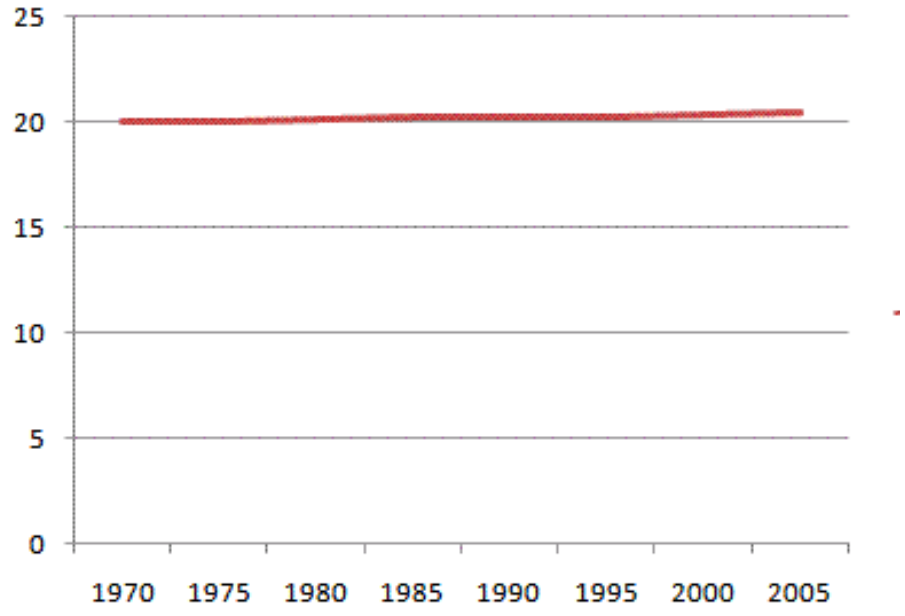
A sudden and dramatic jump?

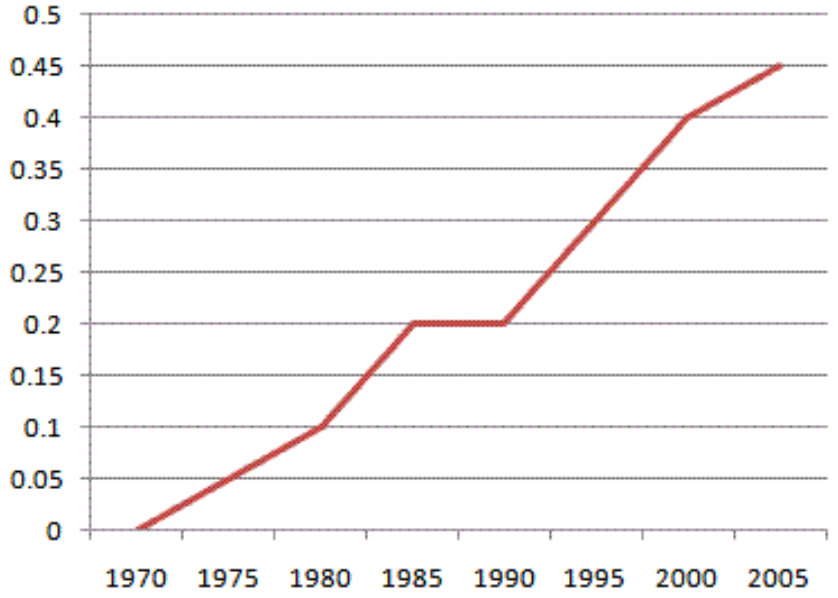


The scale is important

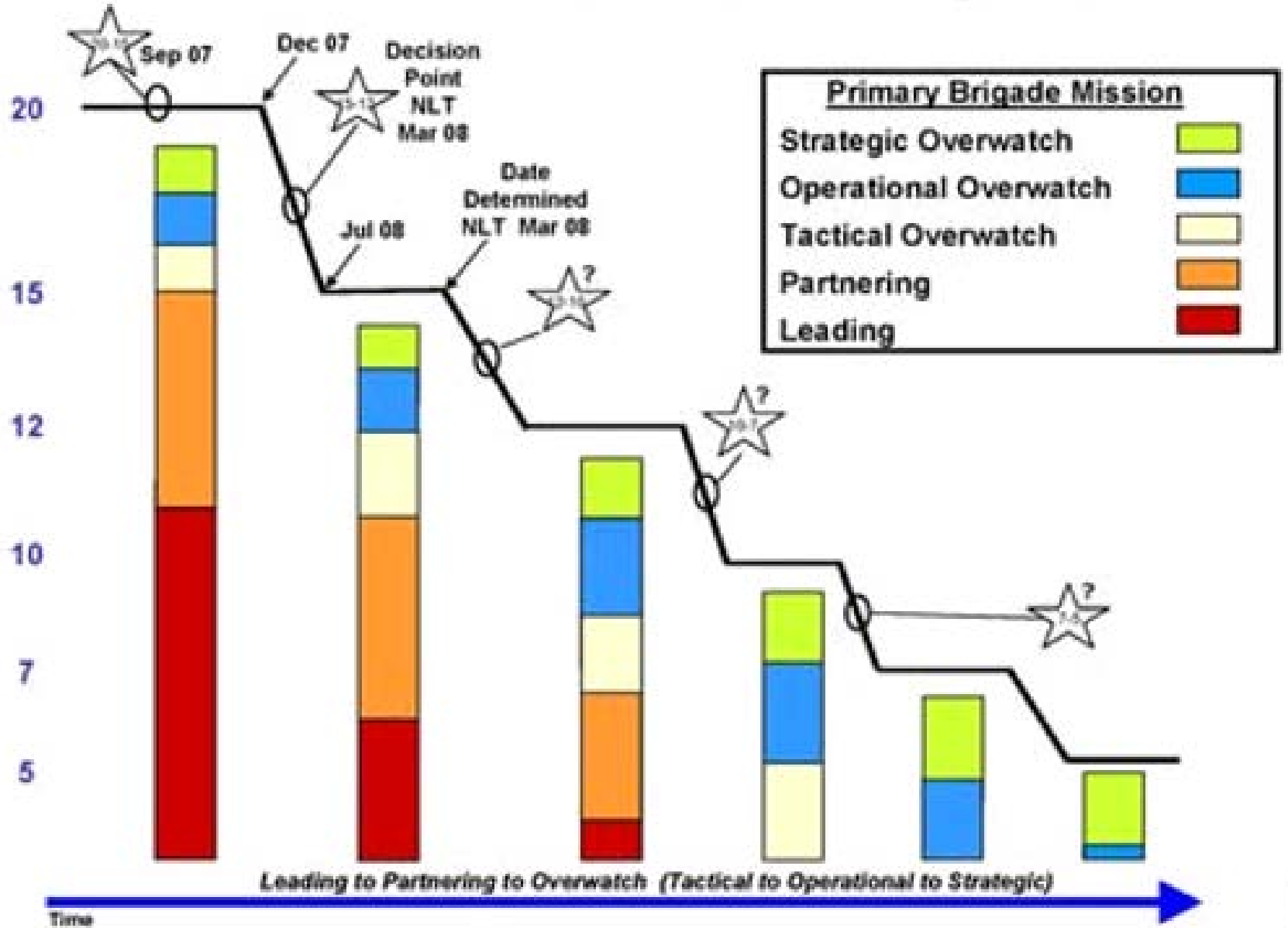




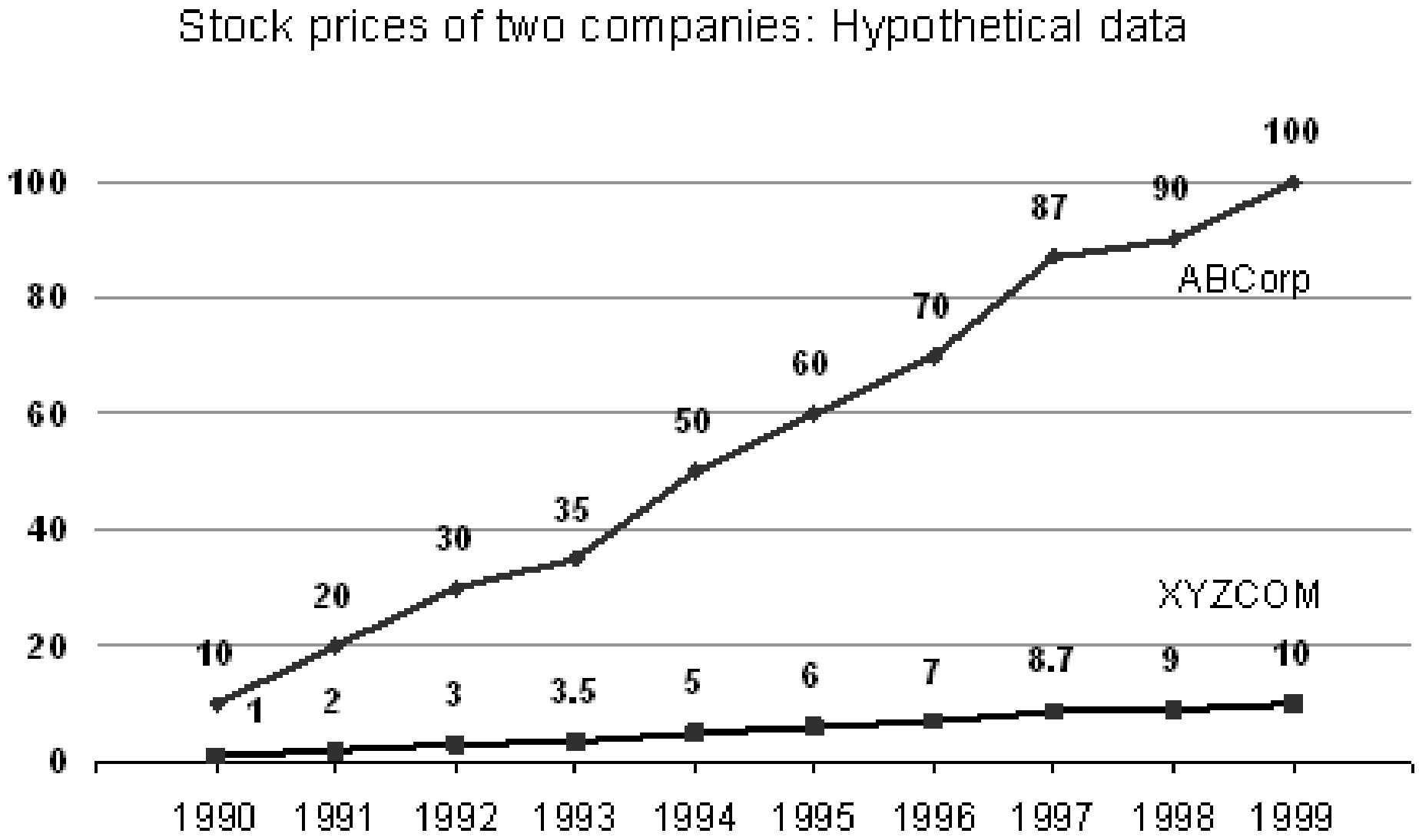




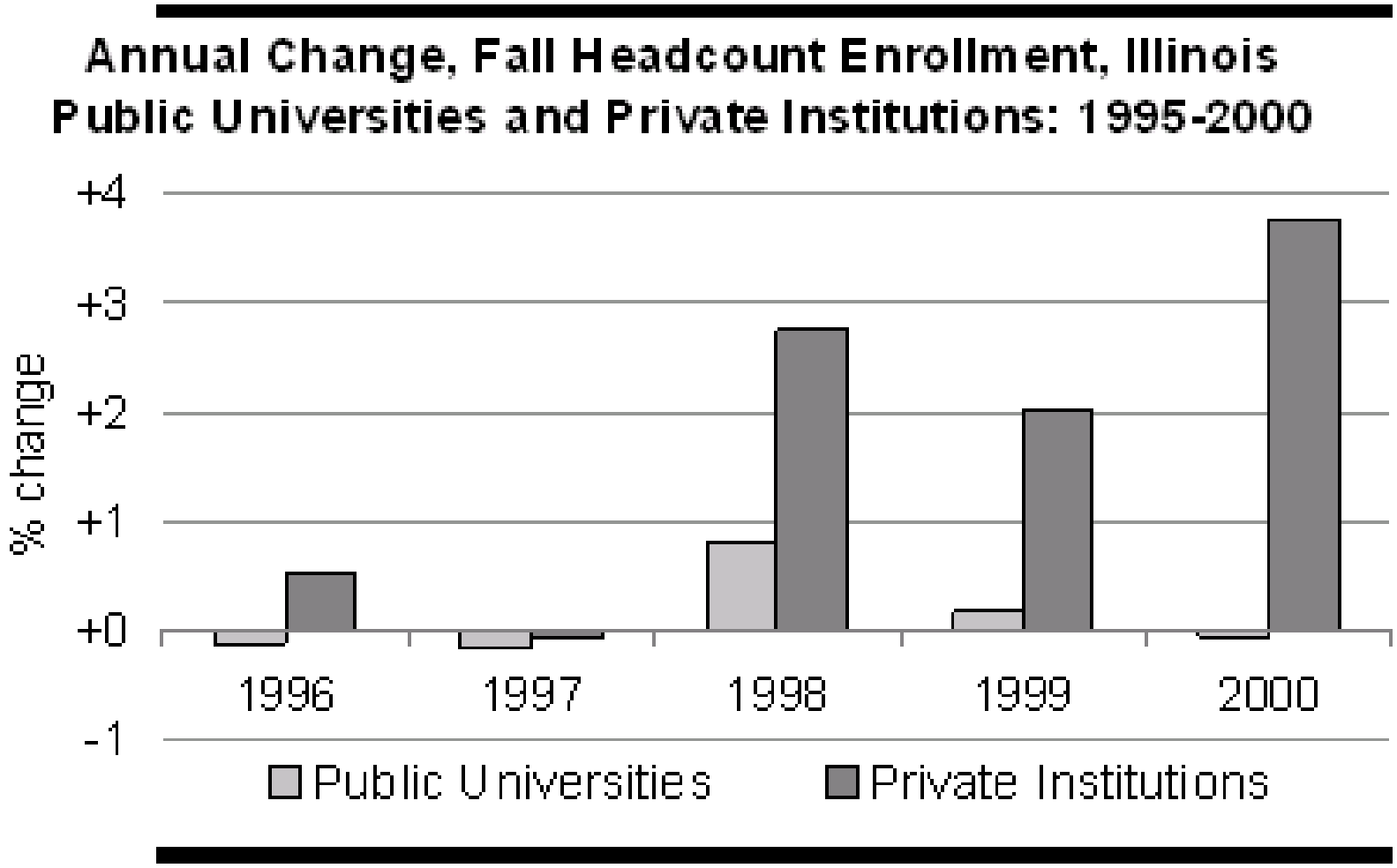
Recommended Force Reductions/Mission Shift



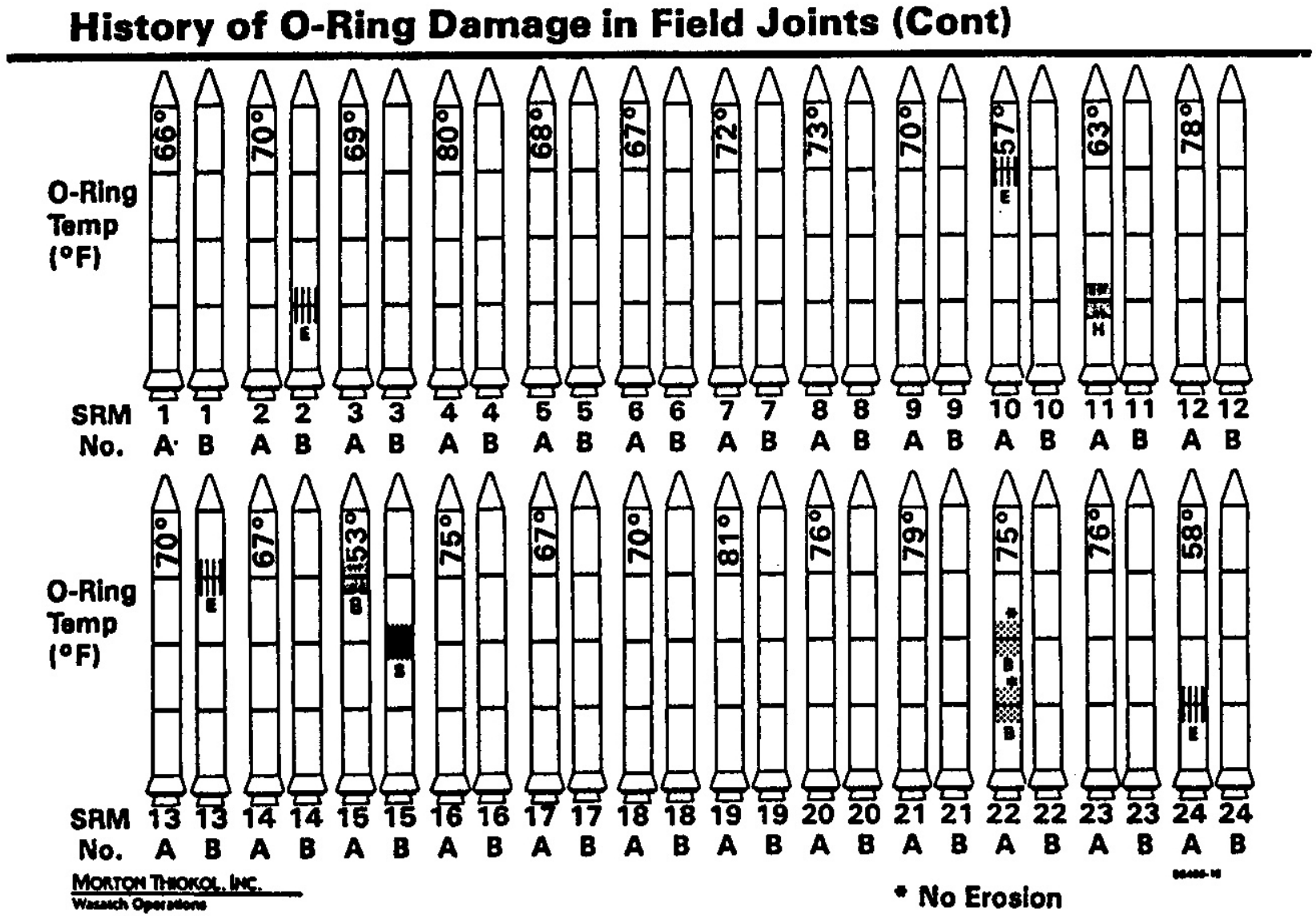
Which is a better investment?



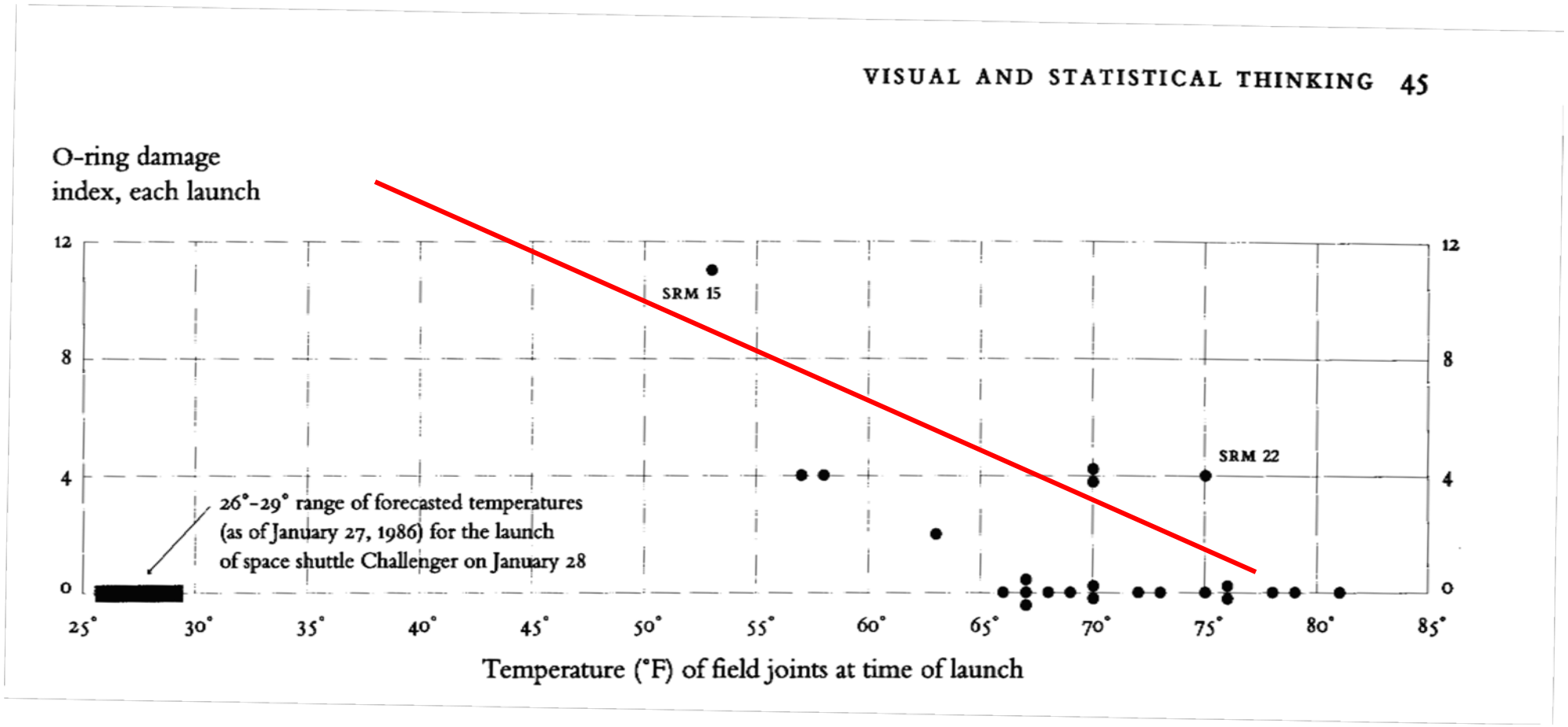
Deceiving Proportions



The Challenger Disaster



The Challenger Disaster



from Tufte, "The Visual Display of Quantitative Information"

Next Week

30	Graphical Methods 2	Ch2-Verzani-2014:80-87;
Feb		
1	Numerical Summaries of data	Ch1-PSDS;
		Ch2-Verzani-2014:50-70;