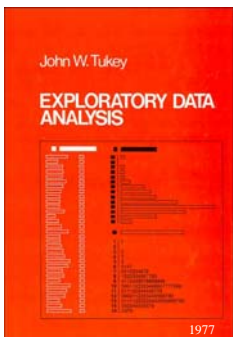


### Exploratory Data Analysis

- Outline -

- Overview
- Stem-and-leaf plots
- Hinges and 5-number series
- Box-and-whisker plots
- Graphing data




---

---

---

---

---

---

---

---

Key to Success:




---

---

---

---

---

---

---

---

### Exploratory Data Analysis

Recommended Text:  
 Tukey, J. W. 1977. Exploratory data analysis.  
 Addison-Wesley Publishing, Reading, MA. 499 p.

Recommended Website:  
 Engineering Statistics Handbook  
<http://www.itl.nist.gov/div898/handbook/eda/eda.htm>

Exploratory vs. Confirmatory Statistics

Emphasis on plots, schematics, re-expression, smoothing, fitting, shaping ⇒ EXPLORING

---

---

---

---

---

---

---

---

### Goals of EDA

Methods which permit easy visual assessment of:

- where the values are centered
- how widely the values are spread
- possible group separation (multi-modality)
- asymmetry or trailing off (skewness)
- unexpectedly popular values (kurtosis & modes)
- unusual or outlying values

---

---

---

---

---

---

---

---

### 3 Graphical Techniques of EDA

1. Stem-and-leaf plots
2. Hinges and 5-number series
3. Box-and-whisker plots

---

---

---

---

---

---

---

---

### Stem-and-leaf plots

Analog to Biological Tree:

Reduce data to a trunk and leaves.

In other words, remove largest divisor by division & reduce each observation to a single entry.

---

---

---

---

---

---

---

---

### Stem-and-leaf plots

-Example-

Batch of 17 obs

250 1166  
 150 688  
 795 1333  
 895 895  
 695 1775  
 1699 895  
 1499 1895  
 1099 795  
 1693

Follow these procedures:

1. Round the last digit of all observations to zero
2. Truncate the trailing zero
3. Find the smallest observation and divide by the largest power of ten that leaves only 1 decimal place
4. Use values to the left of the decimal as your stem (left of line)
5. Use values to the right of decimal as your leaves

---

---

---

---

---

---

---

---

---

---

### Stem-and-leaf plots

-Example (first 4 variates)-

				Step-4&5
				1 5
				2 5
				3
				4
				5
		Step-3		6 98
		2.5		7 99
	Step-1	1.5		8 999
	250	7.9		9
				10 9
Data	150	8.9		11 6
250	790			12
150	890	Step-2		13 3
795		25		14 9
895		15		15
		79		16 99
		89		17 7
				18 9

---

---

---

---

---

---

---

---

---

---

### Stem & Leaf Plot using R

```
> stem (data)
The decimal point is 3
digit(s) to the right of the |
```

```
0 | 23
0 | 7788999
1 | 123
1 | 57789
```

```
> stem(data, scale=5)
The decimal point is 2 digit(s) to the
right of the |
1 | 5
2 | 5
3 |
4 |
5 |
6 | 9
7 | 0
8 | 00
9 | 000
10 |
11 | 07
12 |
13 | 3
14 |
15 | 0
16 | 9
17 | 08
18 |
19 | 0
```

---

---

---

---

---

---

---

---

---

---

### Hinges and 5-number series

Take an ordered batch of numbers  
and "bend them"  
into quartiles (25%*s*)

Thus, one will clearly visualize the:

Minimum value	(0%)
Lower quartile	(25%)
Median	(50%)
Upper quartile	(75%)
Maximum value	(100%)

---

---

---

---

---

---

---

---

### Hinge Plots

-Example-

Sorted batch:  
-3.2, -1.7, -0.4, 0.1, 0.3, 1.2, 1.5, 1.8, 2.4, 3.0, 4.3, 6.4, 9.8

Hinge plot:

The hinge plot shows the sorted batch of numbers: -3.2, -1.7, -0.4, 0.1, 0.3, 1.2, 1.5, 1.8, 2.4, 3.0, 4.3, 6.4, 9.8. Five red circles highlight the minimum value (-3.2), the lower quartile (1.5), the median (0.1), the upper quartile (3.0), and the maximum value (9.8).

---

---

---

---

---

---

---

---

### 5-number series

-Example-

Sorted batch:  
-3.2, -1.7, -0.4, 0.1, 0.3, 1.2, 1.5, 1.8, 2.4, 3.0, 4.3, 6.4, 9.8

5-number series:

#13		
M7	1.5	
H4	0.1	3.0
1	-3.2	9.8

---

---

---

---

---

---

---

---

### 5-Number Series using R

```
> batch<-c(-3.2, -1.7, -0.4, 0.1, 0.3, 1.2, 1.5, 1.8,
2.4, 3.0, 4.3, 6.4, 9.8)

> fivenum(batch)
[1] -3.2 0.1 1.5 3.0 9.8

Recall: Summary function
> summary(batch)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-3.200  0.100  1.500  1.962  3.000  9.800
```

---

---

---

---

---

---

---

---

### Box-and-whisker Plots

Essentially a graphical extension  
of the 5-number series  
that permits easier  
interpretation  
and  
cross-comparison

(Universally used in statistical software.)

---

---

---

---

---

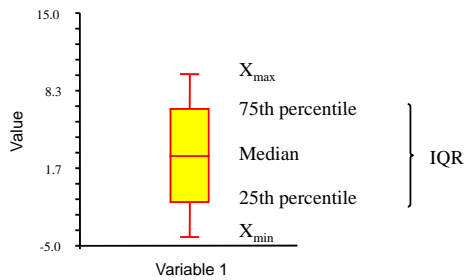
---

---

---

### Simple box-and-whisker plot

(using hinge data set; showing extremes)




---

---

---

---

---

---

---

---

### Box-and-whisker Plots

Recall that a weakness of using the range ( $Y_{\max} - Y_{\min}$ ) to express dispersion is the extreme sensitivity to outliers.

Thus, B&W plots rarely employ  $Y_{\min}$  and  $Y_{\max}$  as terminal values in the graphic.

An alternate method of construction is typically employed...

---

---

---

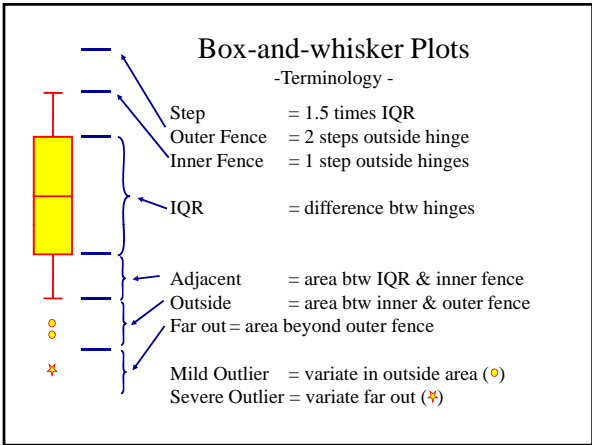
---

---

---

---

---




---

---

---

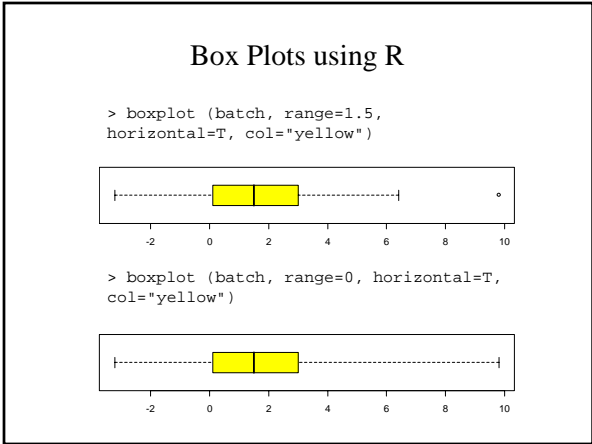
---

---

---

---

---




---

---

---

---

---

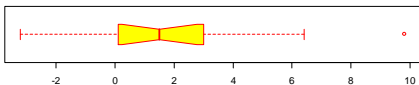
---

---

---

### Box Plots using R

```
> boxplot (batch, range=1.5,
horizontal=T, col="yellow", border="red",
notch=T)
```



How does R determine whisker length?

---

---

---

---

---

---

---

---

---

---

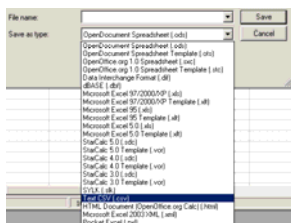
---

---

	G1	G2	G3	G4
1	1200	1900	1800	1200
2	1100	1800	1700	1300
3	1300	1200	1800	1400
4	1400	1400	1600	1500
5	1500	1500	16000	1600
6	1100	1900	1800	1200
7	1300	1800	1700	1300
8	1400	1200	1800	1400
9	1500	1400	1600	1500
10	1200	1500	1600	1600

### Box Plots for EDA Data Screening

Enter 4x10 dataframe  
in spreadsheet and  
save as CSV file.




---

---

---

---

---

---

---

---

---

---

---

---

```
> setwd("C:/TEMPR/")
> data<-read.csv("BP.csv")
> data
> data
```

	G1	G2	G3	G4
1	1200	1900	1800	1200
2	1100	1800	1700	1300
3	1300	1200	1800	1400
4	1400	1400	1600	1500
5	1500	1500	16000	1600
6	1100	1900	1800	1200
7	1300	1800	1700	1300
8	1400	1200	1800	1400
9	1500	1400	1600	1500
10	1200	1500	1600	1600

Dataframe BP.csv has  
4 columns and 10  
rows.

Assume each column  
is same variable (G)  
measured from 4  
samples.

---

---

---

---

---

---

---

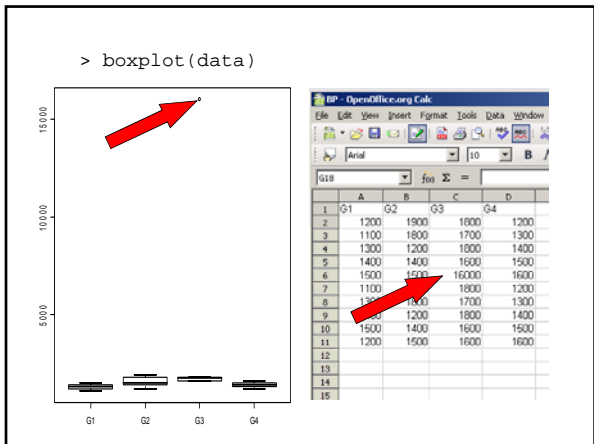
---

---

---

---

---




---

---

---

---

---

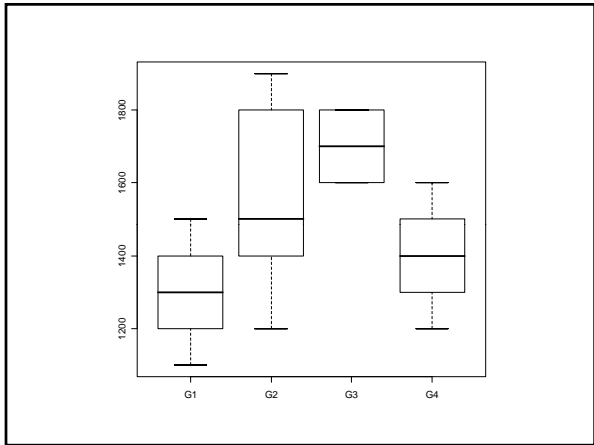
---

---

---

---

---




---

---

---

---

---

---

---

---

---

---

### Box-and-whisker Plots

- Caveats -

All software apps use unique algorithms. It is the responsibility of the user to understand what the program is doing!

SigmaPlot is considered an "industry standard" scientific graphing program (but its calculation of Box Plots is VERY unusual; although acceptable for its purpose).

Recall the first data set we used to examine the basic notion of box-and-whisker plots...

**USER BEWARE!**

---

---

---

---

---

---

---

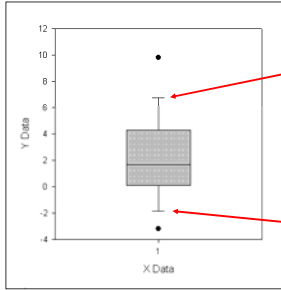
---

---

---

### Box-and-whisker Plots

- SigmaPlot -



Outlier  
90th percentile  
75th percentile  
50th percentile  
25th percentile  
5th percentile  
Outlier

---

---

---

---

---

---

---

---

### Graphing Data

- General Practices -

Classic references on topic:

Tufte, E.R. 1983. The visual display of quantitative information. Graphics Press, Cheshire, CT.

Cleveland, W.S. 1994. The elements of graphing data. Hobart Press, Summit, NJ.

---

---

---

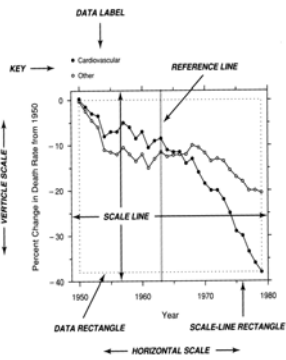
---

---

---

---

---



### Graphics Terminology and Graph Construction

Source: Cleveland 1994

---

---

---

---

---

---

---

---

**Rules for Quality Graph Construction**

1. Use 4 scale lines to encapsulate data (not 2)
2. Data space should NEVER intersect scale space
3. Keep data labels outside scale space
4. Keep scale line tick marks to outside
5. Do not overdo tick marks
6. Use a reference line to mark data separation

cont...

---

---

---

---

---

---

---

---

**Rules for Quality Graph Construction**

7. Captions need to be comprehensive & informative  
(Figures must "stand alone")
8. Always use error bars when appropriate  
(Specify what error bars are used)
9. Be careful with aspect ratio (1:1 often best)
10. Minimize scale and maximize data space  
(do not insist on a 0,0 origin)
11. Use log scale to emphasize exponential change

---

---

---

---

---

---

---

---