Exploratory Data Analysis (EDA)[©]

John Tukey has developed a set of procedures collectively known as EDA. Two of these procedures that are especially useful for producing initial displays of data are: 1. the Stem-and-Leaf Display, and 2. the Box-and-Whiskers Plot.

To illustrate EDA, consider the following set of pulse rates from 96 people:

66	60	64	64	64	76	82	70	60	78
92	82	90	70	62	60	68	68	70	70
68	76	68	72	98	60	80	80	104	70
92	80	90	64	78	60	60	70	66	76
70	52	74	78	70	68	66	80	62	56
58	68	60	48	78	86	68	90	76	70
94	90	64	68	68	80	70	72	80	60
68	99	60	74	56	86	64	86	64	68
76	74	70	77	80	72	88	94	78	70
78	78	55	62	74	58				

Stem and Leaf Display

You first decide how wide each row (class interval) will be. I decided on an interval width of 5, that is, I'll group together on one row all scores of 40-44; on another, 45-49; on another, 50-54, etc. I next wrote down the **leading digits** (most significant digits) for each interval, starting with the lowest. These make up the stem of the display. Next I tallied each score, placing its **trailing digit** (rightmost, least significant digit) in the appropriate row to the right of the stem. These digits (each one representing one score) make up the leaves of the display. Here is how the display looks now:

4	
4	8
5	2
5	68658
6	0444020040020400442
6	68888686888888
7	0000200040002440204
7	6868688667888
8	220000000
8	6668
9	20200404
9	89
10	4
10	

Notice that the leaves in each row are in the order that I encountered them when reading the unordered raw data in rows from left to right. A more helpful display is one where the trailing digits are ordered from low to high, left to right, within each row of the display. Here is how the display looks with such an ordering:

```
  1      4   8
  2      5   2
  7      5   56688
 26      6   0000000002224444444
 40      6   66688888888888
(19)     7   0000000000002224444
 37      7   6666678888888
 24      8   000000022
 15      8   6668
 11      9   00002244
  3      9   89
  1     10   4
```

Notice that I have entered an additional column to the left of the stem. It gives the **depth** of each row, that is, how many scores there are in that row and beyond that row to the closer tail. For example, the 7 in the third row from the top indicates that counting from the lowest score up to and including the scores in the third row, there are seven scores. The 24 in the eighth row indicates that there are 24 scores of 80 and higher. The row that has a number in parentheses is the row that contains the median. The number within the parentheses is the number of scores in that row.

Rotate the display 90 degrees counter-clockwise. Now you see a histogram. But this histogram is made out of the scores themselves, so you can always find out how many times any particular score occurs. For example, if you want to know how many 76's there are, go to the row with the 70's and count the number of trailing 6's. There are five 76's.

We could have grouped our data with interval-widths of 10. For these data such a display follows:

```
  1      4   8
  7      5   256688
 40      6   000000000222444444466688888888888
(32)     7   00000000000022244446666678888888
 24      8   0000000226668
 11      9   0000224489
  1     10   4
```

Do you notice any of the trailing digits that seem odd when compared to the others? What does this suggest about the way the pulse rates were probably determined? Might these odd scores represent errors, or just deviations in the way the pulse rate was determined for a couple of subjects?

Box and Whiskers Plot

Another handy EDA technique is the box-and-whiskers plot. One first completes the **median location**, which is equal to (N + 1)/2, where N = the total number of scores. For our data, the median location = (96 + 1)/2 = 48.5. That means that the median is the 48.5th score from the top (or the bottom). The 48.5th score is the mean of the 48th and the 49th scores. To find the 48th score look at the second stem-and-leaf display on this handout [it is at the top of page 2]. Look at the row with (19) in the depth column. The median is in there somewhere. There are 40 scores lower than 70 [

the depth of the row just above is 40], so we count in 8 scores, making the 48th score a 70. The next score over, the 49th score, is also a 70, and the mean of 70 and 70 is 70, so the median is 70.

Now we find the **hinge location**, which is equal to (Median Location + 1)/2. Drop any decimal on the median location when using it in the formula for the hinge location. For our data, hinge location = (48 + 1)/2 = 24.5. Now, the **upper hinge** is the 24.5th score from the upper end of the distribution. Scan the depth column from the bottom up the stem-and-leaf display. Note the row with a depth of 24. There are 24 scores of 80 or more. The 25th score from the top is a 78, so the upper hinge is 79. This is essentially the same thing as the third quartile, Q_3.

Now, the **lower hinge** is the 24.5th score from the lower end of the distribution. Our stem-and-leaf display shows a depth of 26 for the fourth row, so the 26th score from the lowest is a 64. The 25th and 24th are also 64's, so the lower hinge is 64. This is essentially the same as Q_1, the first quartile.

The **H-spread** is the difference between the upper hinge and the lower hinge. For our data, 79 - 64 = 15. This is a trimmed range, a range where we have eliminated some percentage of extreme scores. In this case we eliminated all but the middle 50%. The H-spread is essentially the same thing as the interquartile range.

Next we find the upper and lower inner fences. The **upper inner fence** is the upper hinge plus 1.5 H-spreads. For our data, 79 + 1.5(15) = 101.5. The **lower inner fence** is the lower hinge minus 1.5 H-spreads, 64 - 1.5(15) = 41.5.

The **adjacent values** are actual scores that are no more extreme than the inner fences [not lower than the lower inner fence and not higher than the higher inner fence]. For our data the adjacent values run out to 48 on the low side and up to 99 on the high side. All scores that are not adjacent values are **outliers**. If possible, one should investigate outliers. Why are these scores so extreme? They might be errors.

A second pair of fences, the **outer fences**, is located 3 H-spreads below the lower hinge and 3 H-spreads above the upper hinge. For our data the outer fences are located at 79 + 3(15) = 124 and at 64 - 3(15) = 19.

Now we can draw the plot. First a box is drawn from the lower to the upper hinge. A vertical line or a cross marks the median within the box. Sometimes the mean is also marked with some symbol. Whiskers are drawn out from the box to the lowest adjacent value and to the highest adjacent value (not necessarily all the way to the inner fences). Finally, outliers are located with asterisks or another prominent symbol.

Outliers that are not only beyond the inner fences but also beyond the outer fences are **way-outliers**. These are plotted with a symbol different from those outliers that within the outer fences. For our data there are no way-outliers.

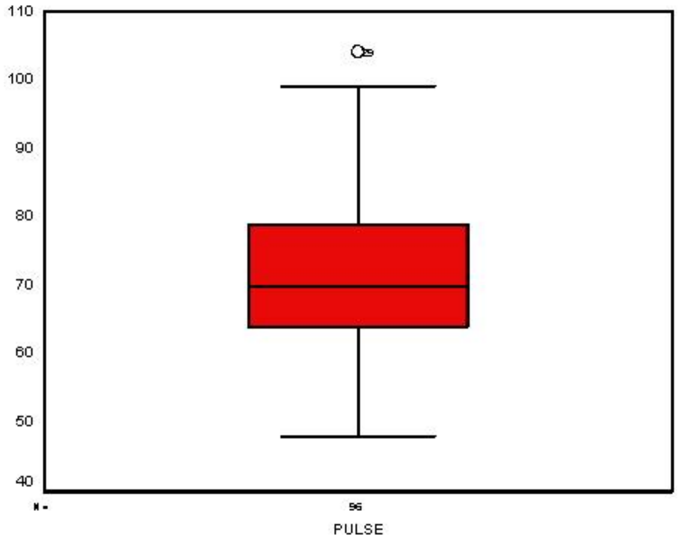
Stem and leaf and box and whiskers plots are most conveniently prepared using statistical software. The data from this handout are available in the file EDA.SAV on my SPSS data files page and in the file EDA.DAT on my StatData page. Here are the plots as prepared by SPSS for Windows, (both SPSS and SAS produce a Frequency column rather than a Depth column):

```
PULSE Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     1.00        4 .  8
     6.00        5 .  256688
    33.00        6 .  000000000222444444466688888888888
    32.00        7 .  00000000000022244444666667888888
    13.00        8 .  0000000226668
    10.00        9 .  0000224489
     1.00 Extremes    (>=104)

 Stem width:   10
 Each leaf:       1 case(s)
```



Here are the plots, as produced by SAS.

```
Stem Leaf                     #             Boxplot
  10 4                        1                0
   9 89                       2                |
   9 00002244                 8                |
   8 6668                     4                |
   8 000000022                9                |
   7 6666678888888           13             +-----+
   7 0000000000002224444     19             *--+--*
   6 66688888888888          14             |     |
   6 0000000002224444444     19             +-----+
   5 56688                    5                |
   5 2                        1                |
   4 8                        1                |
     ----+----+----+----+
 Multiply Stem.Leaf by 10**+1
```

The original boxplots, called “skeletal box and whiskers plots,” were more simple than those I described above. They consisted of a box representing the middle 50% of the scores, bisected by a line at the median, and with whiskers drawn all the way out to the lowest data value and the highest data value. Tukey called the sort of boxplot I have described above a “schematic plot,” but most people simply refer to them as boxplots.

A Silly Way to Conceptualize Box and Whisker Plots

Think of the box as a hen house and the scores as hens. At any time one half of the hens are inside the hen house. Most of the other hens are going to be adjacent to the hen house. The farmer has an electric fence to keep the hens from wandering off. This is the **inner fence**. This farmer is very modern, and his electric fence is invisible, like that in the classic science fiction move, Forbidden Planet (1956), which was inspired by Shakespeare’s “The Tempest.” Watch the forbidden planet way-outlier trying to break back through the invisible fence. This monster, by the way, is the Id (updated a bit since the time of Sigmund Freud).

Every once and a while one of hens gets a wild feather up her cloaca and runs full speed through the inner fence. Electrocuted by the force field, it falls to the ground in a smoldering heap. The plotted symbol for the outlier is just the smoldering carcass of the hen.

The farmer is an avid motorsports fan, and has constructed a race track that encircles the area where the hen house is located, but well beyond the inner fences. Think of the race track as being the **outer fence**. Every once in a blue moon a hen runs through the inner fence with such great abandon that it makes it all the way to the track where it collides with a race car (usually one driven by Kyle Bush, whose Toyota does not have a functioning brake system and has an accelerator that is stuck on full throttle). The plotted symbol for the way-outlier is the flattened carcass of such a hen.

I know this is really silly, but sometimes really silly stories help you remember important points about things that otherwise would be really boring.

A Quick Review of What You Should Know About Box and Whiskers Plots

From 1983 to 2010, 624 of my students in PSYC 2101, Psychological Statistics, answered the following question: “**On a 0-10 scale, how frightened of this statistics course are you at this moment?** (0 indicates not at all, 10 indicates extreme sympathetic arousal -- a racing heart, dry mouth, sweaty palms, queasy stomach, etc.)” Here are some basic statistics and then a box and whiskers plot of the data.

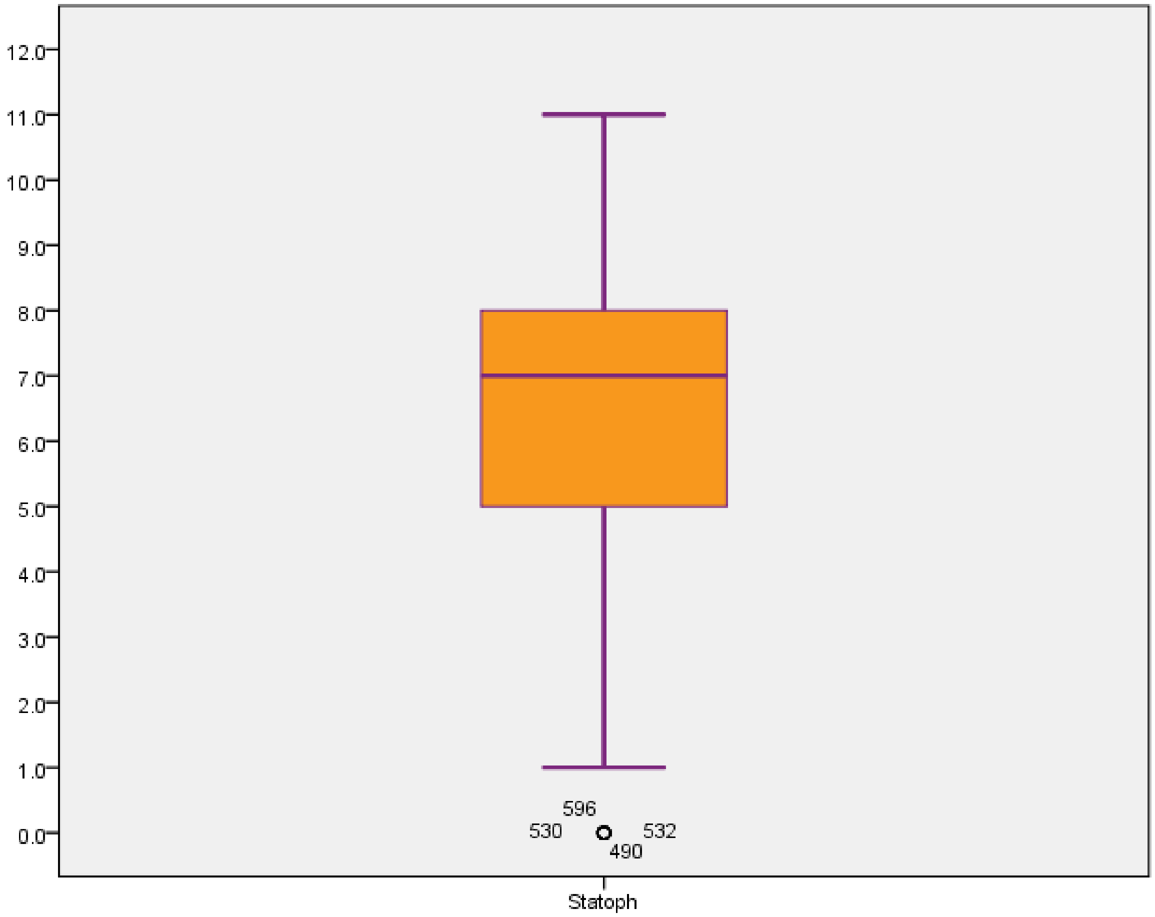
Descriptives

			Statistic	Std. Error
Statoph	Mean		6.357	.0920
	95% Confidence Interval for Mean	Lower Bound	6.177	
		Upper Bound	6.538	
	5% Trimmed Mean		6.458	
	Median		7.000	
	Variance		5.285	
	Std. Deviation		2.2990	
	Minimum		.0	

	Maximum	11.0	
	Range	11.0	
	Interquartile Range	3.0	
	Skewness	-.600	.098
	Kurtosis	-.010	.195

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	Statoph	2.000	3.000	5.000	7.000	8.000	9.000	10.000
Tukey's Hinges	Statoph			5.000	7.000	8.000		

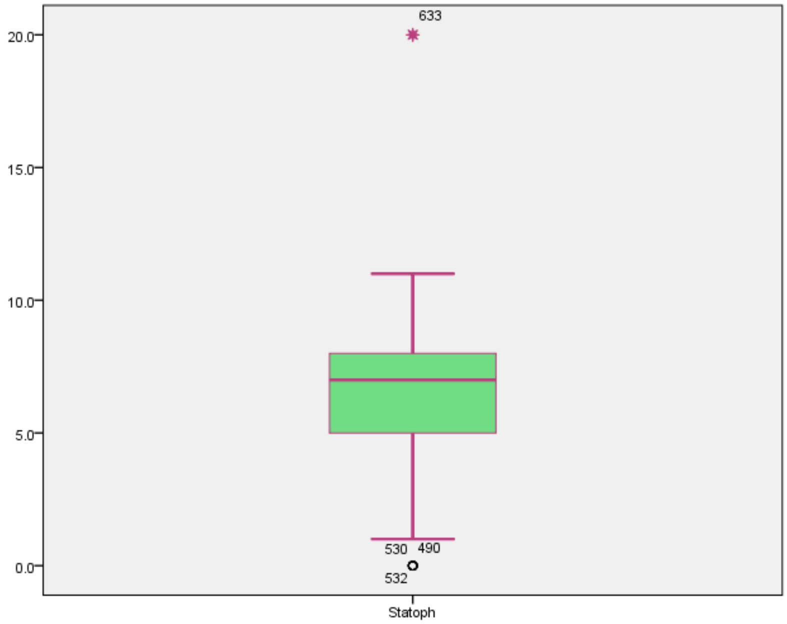


Key Points You Should Know

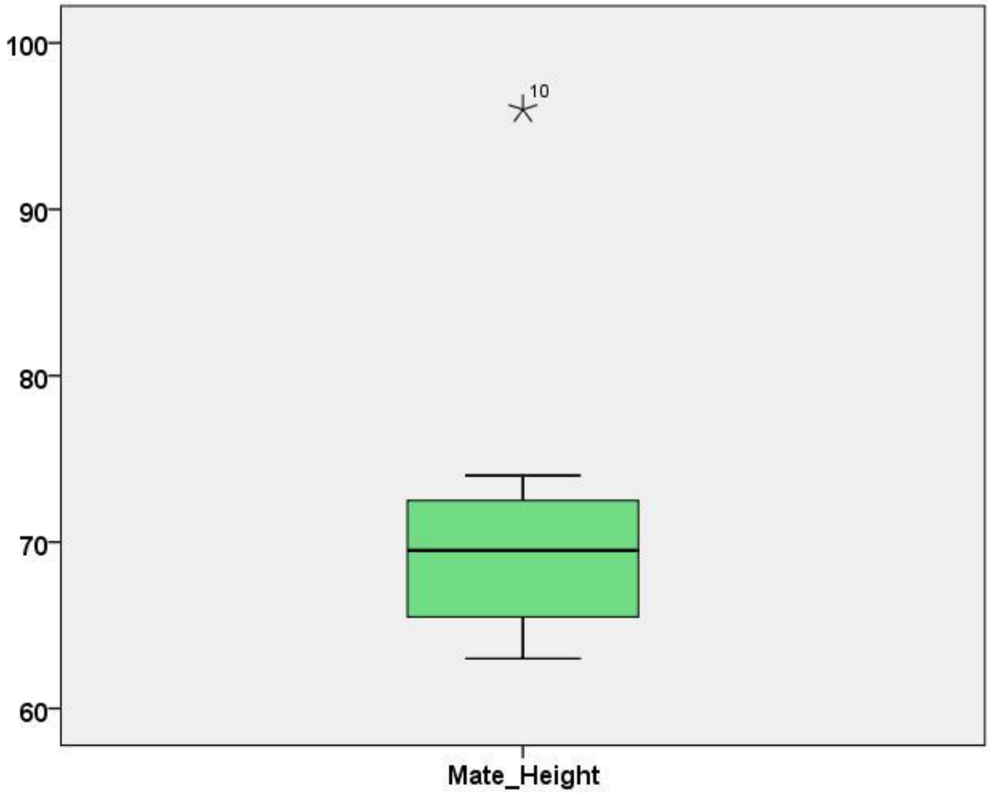
- The box is drawn from the **lower hinge** to the **upper hinge**, which are, for all practical purposes, the same as **the first and the third quartiles**. For these data they are 5 and 8.
- The line bisecting the box locates the median, 7 in this case.

- The **Hinge Spread** is the difference between the upper hinge and the lower hinge, which is essentially the same things as the **interquartile range**, the range of the middle 50% of the scores. For these data the hinge spread is 8-5 = 3.
- The inner fences are invisible boundaries between the adjacent values and the outliers. The upper inner fence is located at 1.5 hinge spreads above the upper hinge. For these data the upper inner fence is located at 8+1.5(3) = 12.5. The lower inner fence is 1.5 hinge spreads below the lower hinge. For these data the lower inner fence is 5-1.5(3) = 0.5
- Scores between the box and the fences are called **adjacent values**.
- Scores outside of the fences are called **outliers**. The symbol used to plot them varies from program to program. SPSS has used a circle.
- The **whiskers** are the lines drawn from the hinges to the most extreme adjacent values. For these data there are no outliers on the upper end. The highest score in the distribution is an 11, and it is within the upper fence (12.5). The line is drawn up to that highest adjacent value, 11. Do notice that the line does not go all the way to the fence.
- There are some **outliers on the lower end of this distribution**, the scores less than the lower fence (0.5). The plot shows that there are three such values (all zeros), and the **case numbers** of those scores are 490, 530, 532, and 596. Knowing these case numbers allows us to go back to the data file and identify the cases with such unusual scores. They should be investigated, as there is an enhanced probability that they represent some sort of error. One actual such error in the "height of ideal mate" data was a score of 4 inches. The student, who had a hearing impairment, had misunderstood the question. I teased him by asking if his ideal mate was named Patty <apologies to Christine O'Donnell>. He was a good chap, and took my teasing in good humor.
- There is a second set of fences, beyond the inner fences, called the **outer fences**. Each is 3 hinge spreads beyond the box. For these data the upper outer fence is located at 8+3(3) = 17, and the lower outer fence is at 5-3(3) = -4. Scores that fall beyond the outer fences are called "**way-outliers**." They are highly suspect of representing errors. We have no way-outliers in the data here.

Just for fun, I added to the data a score of 20. Such as score would be a way-outlier, probably due to somebody not understanding the instructions when answering the question, being silly, or an error on the part of the person entering the data into the file. Here is the boxplot that results. SPSS used an * as the symbol for the way-outlier, but I edited the plot and replaced it with a more prominent star. The graphics editing capabilities in SPSS are good and easy to use.



Here is an example of the utility of boxplots. My Teaching Assistant entered the Introductory Questionnaire data in an Excel file for me first summer session, 2014. I created a boxplot for the responses to the question “how tall, in inches, is your ideal mate?” Here is the boxplot:



Holy moly, look at that WAY-OUTLIER. One case has an ideal mate that is 69 inches tall (eight feet). That is one helluva tall mate. The tallest players in the history of the NBA were only 7 foot 7 inches. I immediately suspected a data entry error here. I went back to the original data sheets and was able to identify the case based on the responses to the other items. The correct score was a 69, not a 96.

Links

- ABC’s of EDA – classic by Velleman and Hoaglin (1981)
- Ask Dr. Math – an illustration of how to produce a box and whiskers plot.
- Wuensch’s Statistics Lessons