Lecture 2

1. EDA (EXPLORATORY DATA ANALYSIS by J.Tukey).
Learn how to explore data and find valuable information, structures and relationships among variables. Find the structure of the majority of the data but also detect exceptional observations, rare events. Read the article “The future of Data Analysis” J W Tukey (1952), Annals of Probability & Statistics.

2. STEM AND LEAF DISPLAYS (Tukey).
A number has about 3 significant digits. Split numbers into STEM + LEAF

| 22.5 | 22 . 5 |
| 0.00000917 | 0.0000091 | 7

If we find a set of numbers decomposed such as common + stem + leaf + noise

| 0.000007884482 | 0.0000078 | 84 | 4 | 82 |
| 0.00000788237 | 0.0000078 | 88 | 2 | 37 |
| 0.000007884282 | 0.0000078 | 84 | 2 | 82 |
| 0.000007883434 | 0.0000078 | 83 | 4 | 34 |
| 0.000007884150 | 0.0000078 | 84 | 1 | 50 |

- **ROUNDING:**
  - Rounding is important: 84 5, 88 2, 84 3, 83 4
  - Round 5’s to the even number. 84 1 50 rounds to 84 2

- **Constructing a STEM AND LEAF diagram.**
  Put the stems in one column and the leaves are added in rows:
  GROUP LINES BY TWO DIGITS:

| Stems: 83 to 88 |
| 83 4 |
| 84 235 |
| 85 |
| 86 |
| 87 |
| 88 2 |

GROUP LINES BY TWO DIGITS:

| 0-1 2-3 4-5 6-7 8-9 |
| 8 3 |
| 8 444 |
| 8 |
| 8 8 |

Or we group lines by 5 digits: 0-4 5-9

| 8 3444 |
| 8 8 |

- **R example**
  In R we use the stem function:
  ```r
  > data(faithful)
  > attach(faithful)
  > stem(waiting,scale=1)
  The decimal point is 1 digit(s) to the right of the |
  4 | 3
  4 | 55566667778889999999
  ```
SAS example

In SAS this is part of PROC UNIVARIATE.
THIS IS AN EXAMPLE OF WHAT SAS DOES.

```
OPTIONS PS=55 LS=80;
DATA CRIME;
INFILE 'crime.dat';
INPUT MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY AUTOTHFT REGION $;
run;
proc univariate plot;
var murder;
run;
```

Stem Leaf                    #
Boxplot
15 3                        1                |
14 6                        1                |
13                                           |
12 27                       2                |
11 156778                   6                |
10 17                       2                |
  9 24456                    5             +-----+
  8 488                      3             |     |
  7 79                       2             |     |
  6 2269                     4             *---+-*
  5 3779                     4             |     |
  4 6688                     4             |     |
  3 02245568                 9             +-----+
  2 000                      3             |
  1 359                      3             |
  0 5                        1             |

Depth, Ranks, Order of data values

DEPTH OF A VALUE = MINIMUM OF ASCENDING AND DESCENDING RANKS.
WE MAY ADD DEPTH TO THE STEM AND LEAF DIAGRAM

```r
depth = function(x) pmin( rank(waiting), length(x) +1-rank(x))
cbind(waiting,depth=depth(waiting), rank=rank(waiting))
```

waiting  depth  rank
[1,]    79   97.5  175.5
[2,]    54   49.0  49.0
[3,]    74  123.5 123.5 .....
3. Histograms

```r
par(mfrow=c(1,2))
hist(waiting)
hist(waiting,50)
```

![Histogram of waiting](image1)

![Histogram of waiting](image2)

For large samples letter values are hard to read but histograms are suitable.

- **CHOOSING THE NUMBER OF LINES OF A STEM & LEAF OR A HISTOGRAM**
  
  \[ L = \lceil 10 \times \log_{10}(n) \rceil \]
  
  where \( \lceil x \rceil \) is the integer part of \( x \).
  
  **EXAMPLE:** If \( n = 25 \) \( L = \lceil 10 \times \log_{10}(25) \rceil = \lceil 10 \times 1.39794 \rceil = 13 \)

- **CLASS WIDTH:** \round{\text{RANGE} / L}
  
  The range for ROBBERY IS 435.7 so 435.7 / 13 = 33.5
  
  We choose width=50.
  
  The range for MURDER IS 13.3 so 13.3 / 13 = 1.02
  
  We choose width= 1 or 2.

- **OTHER RULES**
  
  **Sturges:** \( L = \lceil 1 + \log_2(n) \rceil \)
  
  **RootN:** \( L = \lceil 2 \sqrt{n} \rceil \)
  
  Table 1. comparing \( \log_{10} \), RootN and Sturges rules.

<table>
<thead>
<tr>
<th>N</th>
<th>\log_{10}</th>
<th>\text{RootN}</th>
<th>\text{Sturges}</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10.00</td>
<td>6.32</td>
<td>4.32</td>
</tr>
<tr>
<td>20</td>
<td>13.01</td>
<td>8.94</td>
<td>5.32</td>
</tr>
<tr>
<td>30</td>
<td>14.77</td>
<td>10.95</td>
<td>5.90</td>
</tr>
<tr>
<td>40</td>
<td>16.02</td>
<td>12.64</td>
<td>6.32</td>
</tr>
<tr>
<td>50</td>
<td>16.98</td>
<td>14.14</td>
<td>6.64</td>
</tr>
<tr>
<td>60</td>
<td>17.78</td>
<td>15.49</td>
<td>6.90</td>
</tr>
<tr>
<td>70</td>
<td>18.45</td>
<td>16.73</td>
<td>7.12</td>
</tr>
<tr>
<td>80</td>
<td>19.03</td>
<td>17.88</td>
<td>7.32</td>
</tr>
<tr>
<td>90</td>
<td>19.54</td>
<td>18.97</td>
<td>7.49</td>
</tr>
<tr>
<td>n</td>
<td>F-D</td>
<td>SCOTT</td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>-----</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.017</td>
<td>1.619</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.881</td>
<td>1.285</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0.803</td>
<td>1.123</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>0.749</td>
<td>1.020</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.709</td>
<td>0.947</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>0.678</td>
<td>0.891</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>0.652</td>
<td>0.846</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>0.630</td>
<td>0.809</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>0.611</td>
<td>0.778</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.595</td>
<td>0.751</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.494</td>
<td>0.596</td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>0.443</td>
<td>0.521</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>0.409</td>
<td>0.473</td>
<td></td>
</tr>
</tbody>
</table>
3. LETTER VALUES

Start with a batch of data \( X_1, \ldots, X_n \)

Order Statistics: SORT IT and it becomes \( X_{(1)}, \ldots, X_{(n)} \)

DEFINE Ranks: Upward and Downward for each observation.

DEFINE depth: Min of Upward and Downward rank.

- **MEDIAN**
  
  \[
  \text{Median} = \begin{cases} 
  \frac{X_{(k)} + X_{(k+1)}}{2} & \text{if } n = 2k \\
  X_{(k+1)} & \text{if } n = 2k + 1
  \end{cases}
  \]

  This definition has some bad properties.

- **FOURTHS**
  
  \[
  \text{depth(Fourth)} = \frac{[\text{depth(Median)}] + 1}{2}
  \]

- **5-NUMBER SUMMARIES**
  
  Min Fourth Median Fourth Max

- **LETTER VALUES**
  
  \[
  \text{depth(Eight)} = \frac{[\text{depth(FOURTH)}] + 1}{2}
  \]

  AND SO ON.

  \[
  (\text{Prev depth} + 1)/2
  \]

  Letters as Tags: M F E D C B A Z Y X

- **MID VALUES** = (UPPER LV + LOWER LV)/

- **SPREAD** = UPPER LV - LOWER LV

  Example: Fourth Spread = (Upper Fourth - Lower Fourth)

- **NICE FORM LETTER VALUE DISPLAYS**

<table>
<thead>
<tr>
<th>M</th>
<th>Depth of the Median</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Depth of Fourth</td>
<td>Lower Fourth</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>Lower Extreme</td>
</tr>
</tbody>
</table>

Figure 1. Letter values for the Standard Normal Distribution
Table 3. Letter Values for the standard Normal

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>E</th>
<th>D</th>
<th>C</th>
<th>B</th>
<th>A</th>
<th>Z</th>
<th>Y</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.67</td>
<td>1.15</td>
<td>1.53</td>
<td>1.86</td>
<td>2.15</td>
<td>2.42</td>
<td>2.66</td>
<td>2.89</td>
<td>3.10</td>
</tr>
</tbody>
</table>

Table 4. Letter Spreads for the T-10, T5, T-2 and T-1 and Ratios by std Normal spreads

<table>
<thead>
<tr>
<th></th>
<th>NORMAL</th>
<th>T-10</th>
<th>RAT</th>
<th>T-5</th>
<th>RAT</th>
<th>T-2</th>
<th>RAT</th>
<th>T-1</th>
<th>RAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSPREAD</td>
<td>1.35</td>
<td>1.40</td>
<td>1.04</td>
<td>1.45</td>
<td>1.08</td>
<td>1.63</td>
<td>1.21</td>
<td>2.00</td>
<td>1.48</td>
</tr>
<tr>
<td>ESPREAD</td>
<td>2.30</td>
<td>2.44</td>
<td>1.06</td>
<td>2.60</td>
<td>1.13</td>
<td>3.21</td>
<td>1.39</td>
<td>4.83</td>
<td>2.10</td>
</tr>
<tr>
<td>DSPREAD</td>
<td>3.07</td>
<td>3.35</td>
<td>1.09</td>
<td>3.68</td>
<td>1.20</td>
<td>5.11</td>
<td>1.67</td>
<td>10.05</td>
<td>3.28</td>
</tr>
<tr>
<td>CSPREAD</td>
<td>3.73</td>
<td>4.19</td>
<td>1.13</td>
<td>4.78</td>
<td>1.28</td>
<td>7.62</td>
<td>2.05</td>
<td>20.31</td>
<td>5.45</td>
</tr>
<tr>
<td>BSPREAD</td>
<td>4.31</td>
<td>5.01</td>
<td>1.16</td>
<td>5.93</td>
<td>1.38</td>
<td>11.05</td>
<td>2.56</td>
<td>40.71</td>
<td>9.45</td>
</tr>
<tr>
<td>ASPREAD</td>
<td>4.84</td>
<td>5.82</td>
<td>1.20</td>
<td>7.19</td>
<td>1.49</td>
<td>15.81</td>
<td>3.27</td>
<td>81.47</td>
<td>16.85</td>
</tr>
<tr>
<td>ZSPREAD</td>
<td>5.32</td>
<td>6.63</td>
<td>1.25</td>
<td>8.57</td>
<td>1.61</td>
<td>22.49</td>
<td>4.23</td>
<td>162.97</td>
<td>30.63</td>
</tr>
<tr>
<td>YSPREAD</td>
<td>5.77</td>
<td>7.46</td>
<td>1.29</td>
<td>10.12</td>
<td>1.75</td>
<td>31.91</td>
<td>5.53</td>
<td>325.95</td>
<td>56.48</td>
</tr>
<tr>
<td>XSPREAD</td>
<td>6.19</td>
<td>8.32</td>
<td>1.34</td>
<td>11.85</td>
<td>1.91</td>
<td>45.19</td>
<td>7.29</td>
<td>651.90</td>
<td>105.24</td>
</tr>
</tbody>
</table>

Figure 2. log₂(spread ratios of T with df=10,5,2,1 over Normal spreads ) Vs -log₂(Tail Prob)

Estimated slopes: 0.047 0.1046 0.328 0.782

Figure 3. DF approximation: 1/2 + 1/(2.25*slope)
R-CODE
probs= 2^(-2:10) # tail prob for 9 letter values
u = -qnorm(probs) # Letter values for Standard Normal
x = c(t(cbind(-u,-u,NA,u,u,NA))) # Lines for the graph
yy = seq(0.05,0.4,length=9) # Lines for the graph
y = c(t(cbind(0,yy,NA,0,yy,NA))) # Lines for the graph
x1 = c(t(cbind(-u+0.07,u-0.09,NA))) # Lines for the graph
y1 = c(t(cbind(yy-0.005,yy-0.005,NA))) # Lines for the graph
# Graph commands start here
par(cex=0.8) # Set character size at 80%
plot(z <- ((-400):400)/100,dnorm(z),type="l",xlab="") #Plots the bell curve
lines(x,y,lwd=2,col=2) # draws vertical lines
L=c("F","E","D","C","B","A","Z","Y","X") # vector of letters
text(-u-.05,yy+0.02,L,col=3) # Draws letters
text(u+.03,yy+0.02,L,col=3) # Draws letters
lines(x1,y1,col=4) # Draw lines
par(cex=0.6) # lower character size
text(0,yy+0.006,paste(L,"sp=" ,round(2*u,2), sep="")) #text over lines
v = qnorm(probs) # normal quantiles
v = -2*cbind(v,qt(probs,df=10),qt(probs,df=5),qt(probs,df=2),qt(probs,df=1)) # combine it with t quantiles # for df= 10,5,2,1
matplot(1:9,log2(v/v[,1])) # Plot all the ratios
lsfit(1:9, log2(v/v[,1])[,2])$coef[2] -> sl   # calculate
lsfit(1:9, log2(v/v[,1])[,3])$coef[2] -> sl[2]  # individual
lsfit(1:9, log2(v/v[,1])[,5])$coef[2] -> sl[4]  # all lines

lsfit(1/sl, c(10,5,2,1))$coef  # line fit
plot(1/sl, c(10,5,2,1), pch=16, col=2, ylab="DFs")  # graph of
abline(0.5, 1/2.25, col=3, lwd=2)  # fit
round(1/2 + 1/(sl*2.25))  # final estimate of df’s from line

LETTER VALUE DISPLAYS USING SAS IML CODE

We write a program in SAS PROC IML. This is more like a programming language.
/* This is just an example of PROC IML*/
OPTIONS LS=80 PS=55;
PROC IML;
START LETVAL;
/* DATA IS IN X, M IS THE NUMBER OF LETTER VALUES*/
M = 5;
N = NROW(X);
/* SORT X */
A = X;
X[RANK(X),] = A;
/* SET THE TABLE OF LETTER VALUES */
LV = REPEAT(0, M, 4);

/* D IS THE LENGTH OF THE SUBSAMPLE FOR THE LV */
D = N;
DO I=1 TO M;
    IF(D > 0) THEN DO;
        /* CREATE THE RIGHT SUBSET */
        B = X[1:D,];
        /* SORT THE RIGHT SUBSET */
        A = B;
        B[RANK(B),] = A;
        /* CREATE THE LEFT SUBSET */
        C = X[(N-D+1):N,];
        /* SORT THE LEFT SUBSET */
        A = C;
        C[RANK(C),] = A;
        /* THE MEDIANS*/
        DD = INT((D+1)/2);
        IF( D = (2*DD -1)) THEN DO;
            LV[I,1] = B[DD];
            LV[I,3] = C[DD];
            END;
        ELSE DO;
            LV[I,1] = (B[DD]+B[DD+1])/2;
            LV[I,3] = (C[DD]+C[DD+1])/2;
            END;
        /*CALCULATE THE MID VALUES*/
    LV[I, 2] = (LV[I, 1] + LV[I, 3])/2;
        /*CALCULATE THE SPREAD*/
LV[I, 4] = LV[I, 3] - LV[I, 1];
D = DD;
END;
END;
FINISH;

/*DEFINE X*/
X = {1,3,4,10,25,66,89,99,120,140,160,175};

/* RUN THE CODE */
RUN LETVAL;

/* PRINT THE TABLE */
R = {"M" "F" "E" "D" "C"} ;
C = {"Lower" "Mid" "Upper" "Spread"};
PRINT LV[ ROWNAME=R COLNAME=C];
QUIT;

• OUTPUT FROM SAS:

<table>
<thead>
<tr>
<th></th>
<th>Lower</th>
<th>Mid</th>
<th>Upper</th>
<th>Spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>77.5</td>
<td>77.5</td>
<td>77.5</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>7</td>
<td>68.5</td>
<td>130</td>
<td>123</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
<td>81.5</td>
<td>160</td>
<td>157</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>84.75</td>
<td>167.5</td>
<td>165.5</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>88</td>
<td>175</td>
<td>174</td>
</tr>
</tbody>
</table>

• R: LETTER VALUE FUNCTION.

```r
letval <- function(x, k = 4) {
  out <- array(NA, c(k, 4))
  lx <- rx <- sort(x)
  dimnames(out) <- list(LV[1:k],c("LOWER","UPPER","MID","SPREAD"))
  for(i in 1:k) {
    out[i, 1:2] <- c(median(lx), median(rx))
    nn <- (length(lx) + 1)/2
    lx <- lx[1:nn]
    rx <- rev(rev(rx)[1:nn])
  }
  out[, 3] <- (out[, 1] + out[, 2])/2
  out[, 4] <- out[, 2] - out[, 1]
  out
}
```

A more complicated version:

```r
letval2 <- function(x, k = 4) {
  out <- array(NA, c(k, 6))
  lx <- rx <- sort(x)
  dimnames(out) <- list(LV[1:k],c("LOWER","UPPER","DEPTH","MID","SPREAD","TAIL"))
  for(i in 1:k) {
    out[i, 1:2] <- c(median(lx), median(rx))
    nn <- (length(lx) + 1)/2
    lx <- lx[1:nn]
    rx <- rev(rev(rx)[1:nn])
    out[i, 3] <- nn
  }
  out[, 4] <- (out[, 1] + out[, 2])/2
  out[, 5] <- out[, 2] - out[, 1]
  out
}
```
out[, 6] <- c(0, out[-1, 5]/2/qnorm(1 - 1/2^(2:k)))
out
}

> data(stackloss)
> letval2(stack.loss, 7)

<table>
<thead>
<tr>
<th>LOWER</th>
<th>UPPER</th>
<th>DEPTH</th>
<th>MID</th>
<th>SPREAD</th>
<th>TAIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>M 15.0</td>
<td>15.0</td>
<td>11.0</td>
<td>15.00</td>
<td>0.0</td>
<td>0.000000</td>
</tr>
<tr>
<td>F 11.0</td>
<td>19.0</td>
<td>6.0</td>
<td>15.00</td>
<td>8.0</td>
<td>5.930409</td>
</tr>
<tr>
<td>E  8.0</td>
<td>32.5</td>
<td>3.5</td>
<td>20.25</td>
<td>24.5</td>
<td>10.648939</td>
</tr>
<tr>
<td>D  8.0</td>
<td>37.0</td>
<td>2.0</td>
<td>22.50</td>
<td>29.0</td>
<td>9.451669</td>
</tr>
<tr>
<td>C  7.5</td>
<td>39.5</td>
<td>1.5</td>
<td>23.50</td>
<td>32.0</td>
<td>8.589535</td>
</tr>
<tr>
<td>B  7.0</td>
<td>42.0</td>
<td>1.0</td>
<td>24.50</td>
<td>35.0</td>
<td>8.124892</td>
</tr>
<tr>
<td>A  7.0</td>
<td>42.0</td>
<td>1.0</td>
<td>24.50</td>
<td>35.0</td>
<td>7.238706</td>
</tr>
</tbody>
</table>

4. HOMEWORK:

1. Run the letter value program in SAS and in R for the following data.

   stack.loss
   42 37 37 28 18 18 19 20 15 14 14 13 11 12 8 7 8 9 15 15

   Air Flow
   80 80 75 62 62 62 62 62 58 58 58 58 58 58 50 50 50 50 50 50 56 70

   Water Temp
   27 27 25 24 22 23 24 24 23 18 18 17 18 19 18 19 19 20 20 20

   Acid Conc.
   89 88 90 87 87 87 93 93 87 80 89 88 82 93 89 86 72 79 80 82 91

2. Change the SAS program so it will do a proper letter value table for M more than 5 letter values.

3. For a log-normal distribution and a chi-square with df=1,3,10 find the degrees of freedom of corresponding t-distribution with similar tail. Use the code posted above and make small modifications. First you need to generate a sample from the distribution, then work with the letval2 function and use the last column with the rations as input for the regression. The slope of the regression is the input for the equation

   This is an example for the log normal distribution:

   \[ x = \exp(rnorm(1000)) \]
   \[ xlet = letval2(x, 9) \]
(beta <- lsfit(1:(nrow(xlet)-1), log(xlet[-1,6]))$coef[2])
  0.1267003
.5 + 1/(3.25*beta)
  2.928506