KBTU | KAZAKH-BRITISH TECHNICAL UNIVERSITY

DATA ANALYSIS

Christian Capone
Kazakh-British Technical University
Almaty, Kazakhstan

Christian Capone

- Educational Background: Aerospace Engineering, MBA
- 20+ years working experience at executive level
- Information and Communication Technology Direction Responsibility
- Operational Management Responsibility
- Excellent business development track record
- Presently active as:
  - Docent, at the Kazakh-British Technical University (Organisational Behaviour, Project Management, Oil & Gas Project Management for MBA, Information Systems for MBA, Risk Analysis & Management for MSc, Operations Management for MSc and MBA, Supply Chain Management for MSc)
  - Consultant, for European Automotive Companies
  - Owner, in Kazakh Business Analysis & Software Development Company

Contact Infos

- WhatsApp: +7 701 0711 825
- christian.kbtu@gmail.com
- Receiving students at room #354, upon appointment

Students Conduct

- Students are required:
  - to be respectful to the teacher and other students;
  - to silent mobile phones and not using them during classes; in case of urgent calls they are required to leave the class;
  - to meet the deadlines;
  - to come to classes prepared and actively participate in classroom work;
  - to enter the room before the teacher starts the lesson; students who will enter 20 minutes after the starting of the lessons will count as absent;
  - to attend all classes: the University's regulation states that students will be excluded if they miss more than 10 scheduled sessions;
  - No make-up tests/exams are organised unless there is a reason that the instructor considers valid for missing them;
  - any type of cheating won't be tolerated and it will lead to "0" grade;
- Students are encouraged to:
  - consult the teacher on any issues related to the course;
  - make any proposals on improvement of the academic process;
  - monitor their continuous assessment throughout the semester.

Course Infos

- Dr. Kirill Yakunin, Senior Lecture
  - Lectures
  - Tutorials
- Course Objectives: in the syllabus
- Intended Learning Outcomes: in the syllabus
- Grading Policy: in the syllabus
- Grading Breakdown: in the syllabus

Descriptive Statistic

how to describe a dataset

Dataset

- Dataset are a collection of numbers or a list of things
- The number of data in the sample is called sample size
- Often the dataset is a sample of observation from some reference
- In this case we might want to infer something about the population considering the dataset in the sample

Data Types

- DISCRETE: they can take only particular values and they can be numeric or categorical
  - e.g. hair color
- CONTINUOUS: they occupy any value over a continuous range
  - e.g. student's weight

Describe Discrete Data

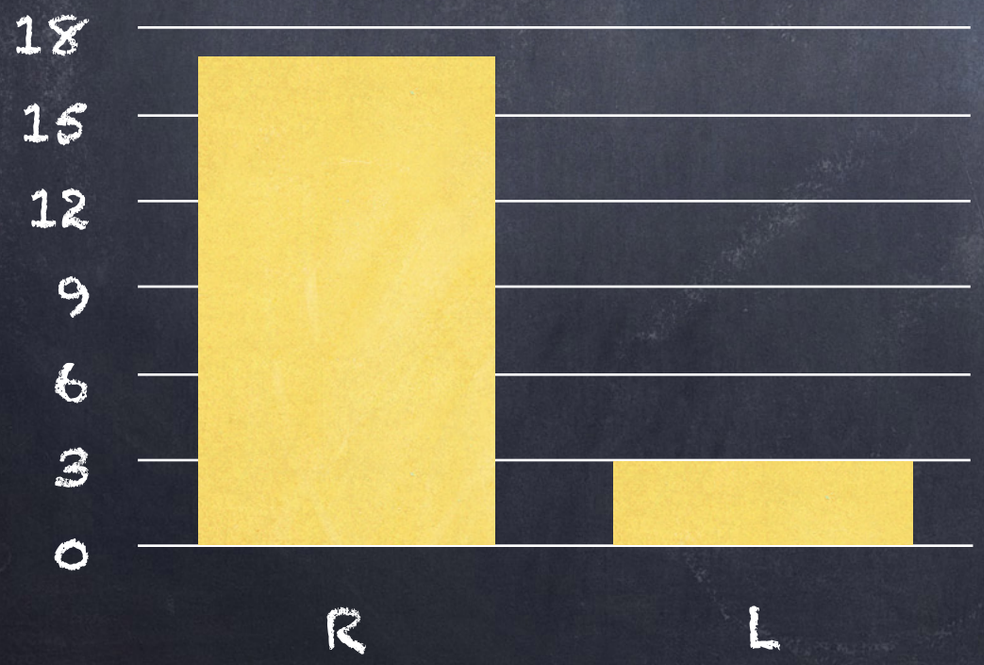
- Suppose I ask to a class of 20 students their stronger hand (R=Right, L=Left)

DATASET

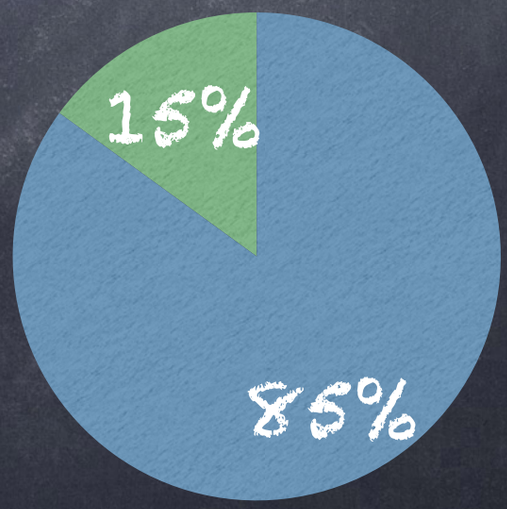
L	R	R	R	R	L	R	R	R	R
R	R	R	R	L	R	R	R	R	R

the SAMPLE DISTRIBUTION

R	L
17	3



Sample Proportion



Exercise

- Obtain distribution and proportion of the following sample

DATASET

0	1	1	1	1	0	1	0	1	1
1	1	1	1	0	1	0	1	1	1
1	0	1	0	0	0	0	0	0	1
0	0	1	1	0	1	0	1	0	0

Describe Continuos Data

- Stem Leaf Plot

Etruscan Head Size (mm)

141	148	132	138	154	142	150	146	155	158	150
140	147	148	144	150	149	145	149	158	143	141
144	144	126								

Sample Distribution

12	6
13	28
14	182607849593144
15	4058008

Stem Class

Leaves

Better Classification

12	6
13	2
13	8
14	12043144
14	8678959
15	4000
15	588

Exercise

- Obtain a stem leaf plot of the following sample

Data									
14	117	77	81	205	21	22	157	134	69
193	8	162	0	156	194	17	100	50	53
235	29	191	81	167	29	158	105	171	2
8	89	82	11	247	149	106	61	18	172

Distribution Classifications

Symmetric (single mode)

```
Low:   49

     6 : 4
     6 : 78
     7 : 14
     7 : 556788
     8 : 0122334
     8 : 67799
     9 : 01122223334
     9 : 55556666777888899999
    10 : 000000001122223444
    10 : 568889
    11 : 000001134
    11 : 599
    12 : 123
    12 : 89
    13 : 2
    13 : 56
    14 : 0

High:  161
```

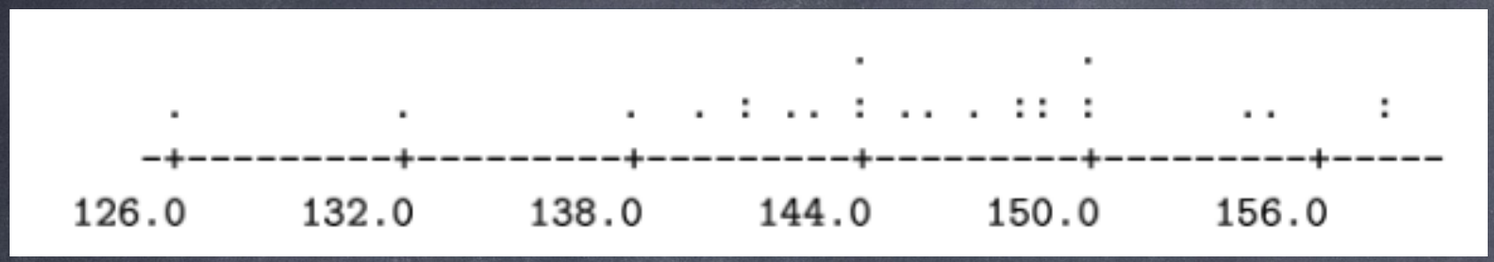
Symmetric (bi-modal)

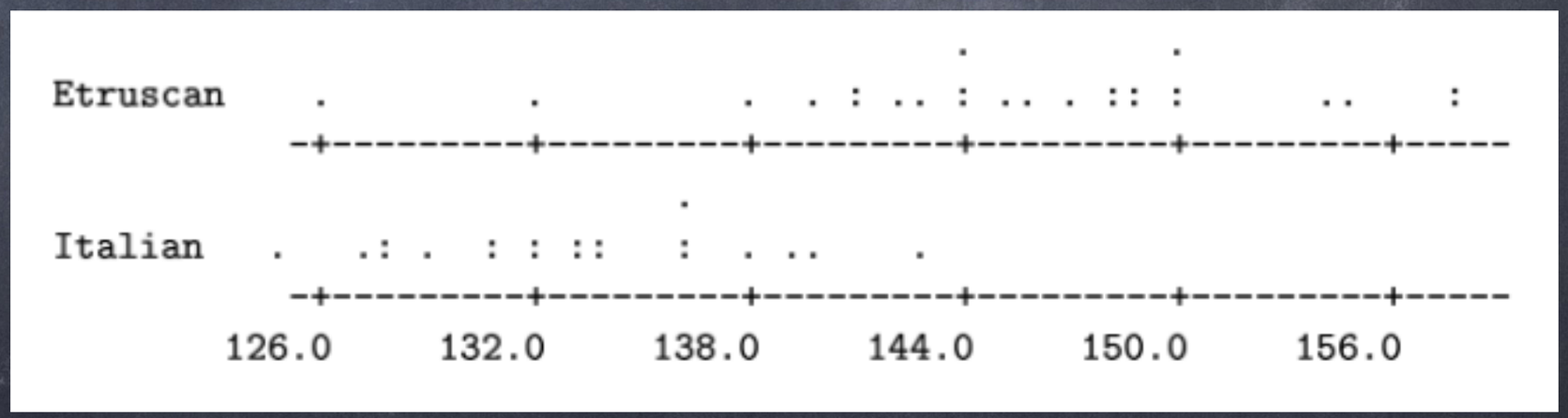
```
-1 : 2
-0 :
 0 : 2
 1 : 5
 2 : 5669
 3 : 1125677
 4 : 223445556699
 5 : 011123334444566667788888999
 6 : 00112224444444555555567788899
 7 : 014445566778899
 8 : 0122334455677799
 9 : 011122223334455556666777788889999
10 : 000000001122223444568889
11 : 00000011123334599
12 : 0122389
13 : 256
14 : 0
15 :
16 : 1
```

Asymmetric (skewed)

```
0 : 1223444444
0 : 5566666677777777888999
1 : 0011111123333344
1 : 5555566788889999
2 : 011222333334
2 : 56666789999
3 : 0114
3 : 668
4 : 02
4 : 58
5 : 02
```

Dot Plot


126.0
132.0
138.0
144.0
150.0
156.0


Etruscan
Italian
126.0
132.0
138.0
144.0
150.0
156.0

Exercise

- Obtain stem leaf plots and comparison dot plots for the following 3 samples

Sample 1

```
      76 183 125  24   8 59  25 179  29 101
      55 108  68 128   5 12  35  25 122  39
      59  91  90  81 66 20 178 111 186  26
      5  123 124  45 13 79 158  20  92  23
```

Sample 2

```
      66   9 62 21 11 39 21  24  21  19
      67  71 67  0  4 82 32  91 152 124
      20 108  5 63  1 10 23 125  59  25
```

Sample 3

```
      59 54 19 79 22 81 18 67 61 53
      71 14 10 87 76 49 21 16 35 11
       7 77 90  6 79 55 83 28 11 60
      55 43  9 65 25
```

5 Basic Descriptive Statistic for Continuous Data

- MINIMUM: the smallest sample value
- MAXIMUM: the largest measure
- DATA RANGE: Maximum - Minimum
  - it measures the sample scale / dispersion / noise
- Measure of the center: MEDIAN (Q_2). It is the middle ordered data point if the sample size is off, the average of the middle ordered data points otherwise. Median is very robust (not much influenced but outliers)
- QUARTILES:
  - 1st Quartile (Q_1): the median of the first half of the data (25% of the data is less than or equal to Q_1)
  - 2nd Quartile (Q_3): the median of the second half of the data (75% of the data is less than or equal to Q_3)

Exercise

- Calculate min, max, data range, median, quartiles for the following stem leaf plot

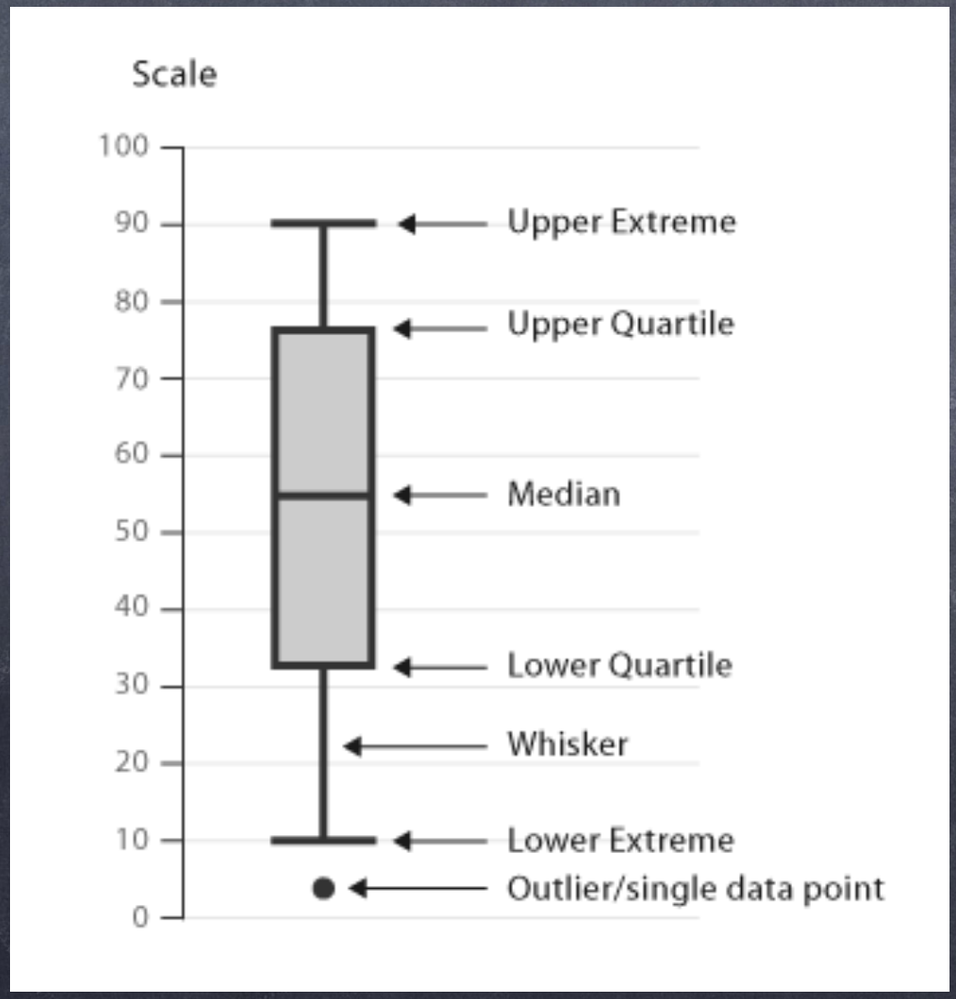
Stem	Leaves	f	F	FTB
12	87859	5	5	
13	31123442	8	13	
13	66698	5	18	7
14	03	2	20	2

Outliers

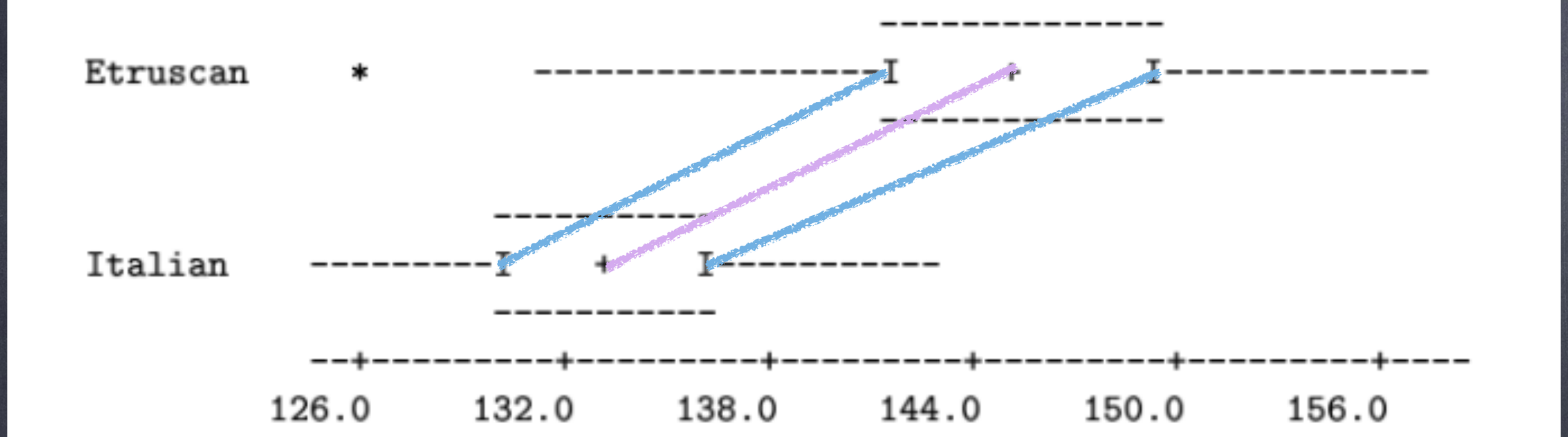
- Frequently the most interesting points of a dataset are the ones that don't seem to belong to it (e.g. level of ozone over Antartica)
- They are points worthy of investigation in order to understand why they differ: they are named OUTLIERS
- We consider POTENTIAL OUTLIERS all the data that are outside +/-1.5 times the interquartile range (IQR = Q_3 - Q_1)
- If h = 1.5IQR, LIF = Q_1 - h and UIF = Q_3 + h
- The data that are very close to the inner fences are called ADJACENT POINTS

Boxplot

- The boxplot is a graphic representation of the basic five descriptive statistic



Comparing Datasets



- Location Problem: difference of the 2 median (146-133 = 13) with samples with similar scale and noise
- We can say that Etruscan heads size shifted up of 13mm from the Italian heads size
- Be careful to do so since we will have also to check the sample error

Exercise

Ten batteries from each of three brands (A, B, and C) were put on test to determine their lifetimes (in hours). Obtain comparison dotplots. Use these dotplots to obtain the 5 basic descriptive statistics for each brand. Bigger means better here. Which brand seems best, if any?

A:	41	289	214	102	38
	94	179	87	116	155
B:	39	65	22	64	22
	191	99	32	142	317
C:	24	95	139	122	41
	360	318	34	43	18

Other Statistic Measures

- Three Types:
  - Measure of the Center (e.g. Q_2)
  - Measure of the Scale or Noise (e.g. Q_1, Q_3)
  - Measure of the Relationships

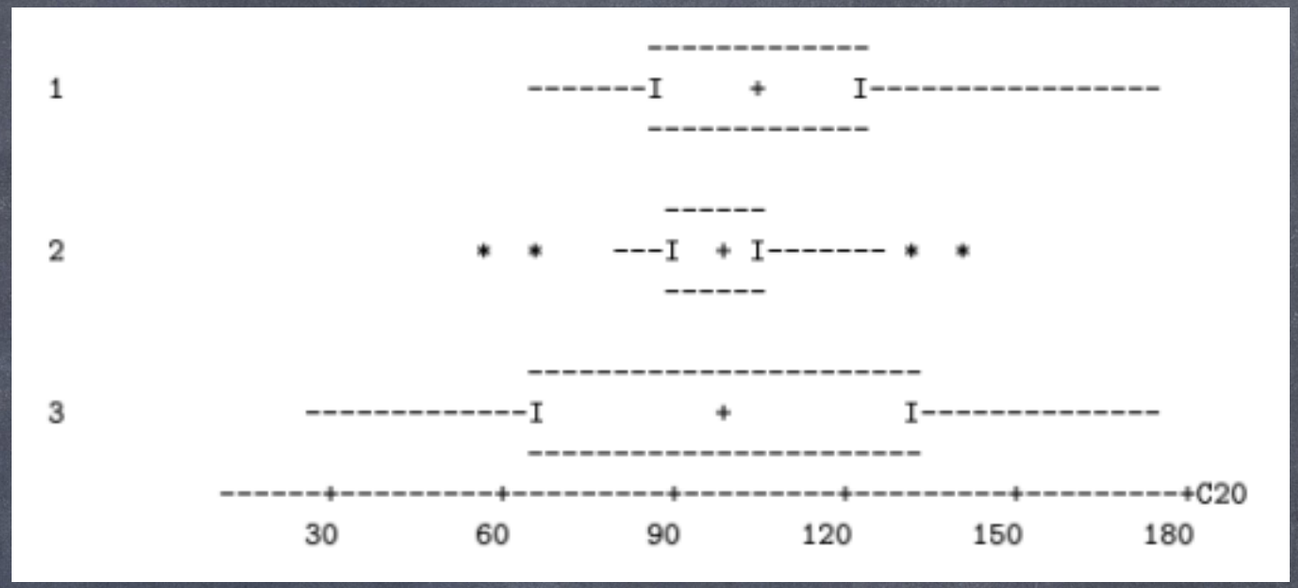
Measure of Center: Mean

- Sample Mean (arithmetic average of the sample)
  - it is consider as the center of gravity of the histogram which is representing the sample
  - it is quite sensitive to the outliers (it is not robust statistic)

	Data						median	mean	IQ
Set 1:	11	18	6	4	8	15	22	11	12	12
Set 2:	11	18	6	4	8	15	72	11	19.1	12
Set 3:	11	18	6	4	8	15	720	11	112	12
Set 4:	11	18	6	4	8	15	2200	11	323	12
Set 5:	11	18	6	4	8	15	7200	11	1037	12
Set 6:	11	18	6	4	8	15	72000	11	10295	12

Measure of the Noise

	Samp.1	Samp.2	Samp.3
1	88	119	91
2	166	116	98
3	143	92	117
4	110	94	62
5	86	86	51
6	108	81	40
7	133	133	57
8	105	65	74
9	114	82	65
10	126	90	60
11	87	86	26
12	99	98	81
13	72	58	133
14	98	106	174
15	73	99	134
16	137	102	120
17	109	93	119
18	82	101	171
19	122	100	132
20	174	101	88
21	65	126	154
22	99	103	154
23	109	142	94
24	105	103	121
25	79	105	131
Median	105	100	98
Mean	108	99	102




```
 6  5
 7  239
 8  2678
 9  899
10  55899
11  04
12  26
13  37
14  3
15
16  6
17  4
```

```
 5  8
 6  5
 7
 8  1266
 9  023489
10  01123356
11  69
12  6
13  3
14  2
```

```
0  3
0  45
0  66677
0  8999
1  0
1  22223333
1  55
1  77
```

You MUST plot the data

Sample Variance and Standard Deviation

$$x_1 - \bar{x}, x_2 - \bar{x}, ..., x_n - \bar{x}$$

$$s = \sqrt{\frac{\text{Sum}(x - \bar{x})^2}{(n-1)}}$$

	Data						median	mean	IQ	s
Set 1:	11	18	6	4	8	15	22	11	12	12	6.61
Set 2:	11	18	6	4	8	15	72	11	19.1	12	23.8
Set 3:	11	18	6	4	8	15	720	11	112	12	268
Set 4:	11	18	6	4	8	15	2200	11	323	12	828
Set 5:	11	18	6	4	8	15	7200	11	1037	12	2717
Set 6:	11	18	6	4	8	15	72000	11	10295	12	27210

The Sample Standard Deviation is not Robust

Exercise

Using the LDL levels of quail a drug compound (call it A) was put on test. In the experiment, 30 quail were randomly chosen and 20 were assigned to a placebo and the other 10 to the treatment using Drug A. The drug was mixed in their food. Other than this, though, the quail were treated the same. At the end of the treament period, the Low Density Lipid levels of the quail were measured and are given below. Here smaller is definitely better. The data are real.

```
Placebo:  64  49  54  64  97  66  76  44  71  89
          70  72  71  55  60  62  46  77  86  71

 Drug A:  40  31  50  48 152  44  74  38  81  64
```

(a) Obtain comparison dot plots of the data and try to decide if the drug A was effective.

(b) Obtain the descriptive statistics for each data sets. Which (difference in means, difference in medians, difference in HL) seem more appropriate here? Why?


Q&A