DATA DESCRIPTION

I. PURPOSE

- Primarily describe specific characteristics of data

- Find out abnormal observations, outliers and mistakes /errors. Then clean the data before doing further analysis

- Inverstigate remarkable features of data, using those features to choose suitable model for data analysis

SIMPLE METHODS USED IN DATA DESCRIPTION

A. Describing 1 qualitative variable

A qualitative variable with ***k*** values corresponding to ***k*** groups of observations in data

$$K_1, K_2, \ldots, K_k,$$

the variable has one same value for all observations in each group → Data description is that to compare numbers of observations in those groups.

→ Data can be represented by

i) Frequency/Percentage table

ii) Bar chart

iii) Pie chart

i) Frequency/percentage table

Qualitative variable with ***k*** values classifies n observations of a study sample into ***k*** groups with $n_1, n_2, \ldots, n_k$ observations respectively ($n_1 + n_2 + \ldots + n_k = n$). The variable can be represenred by a table with ***k*** columns:

	Group 1	Group 2		Group k
N	n_1	n_2	. . .	n_k
%	(n_1 / n) * **100%**	(n_2 / n) * **100%**		(n_1 / n) * **100%**

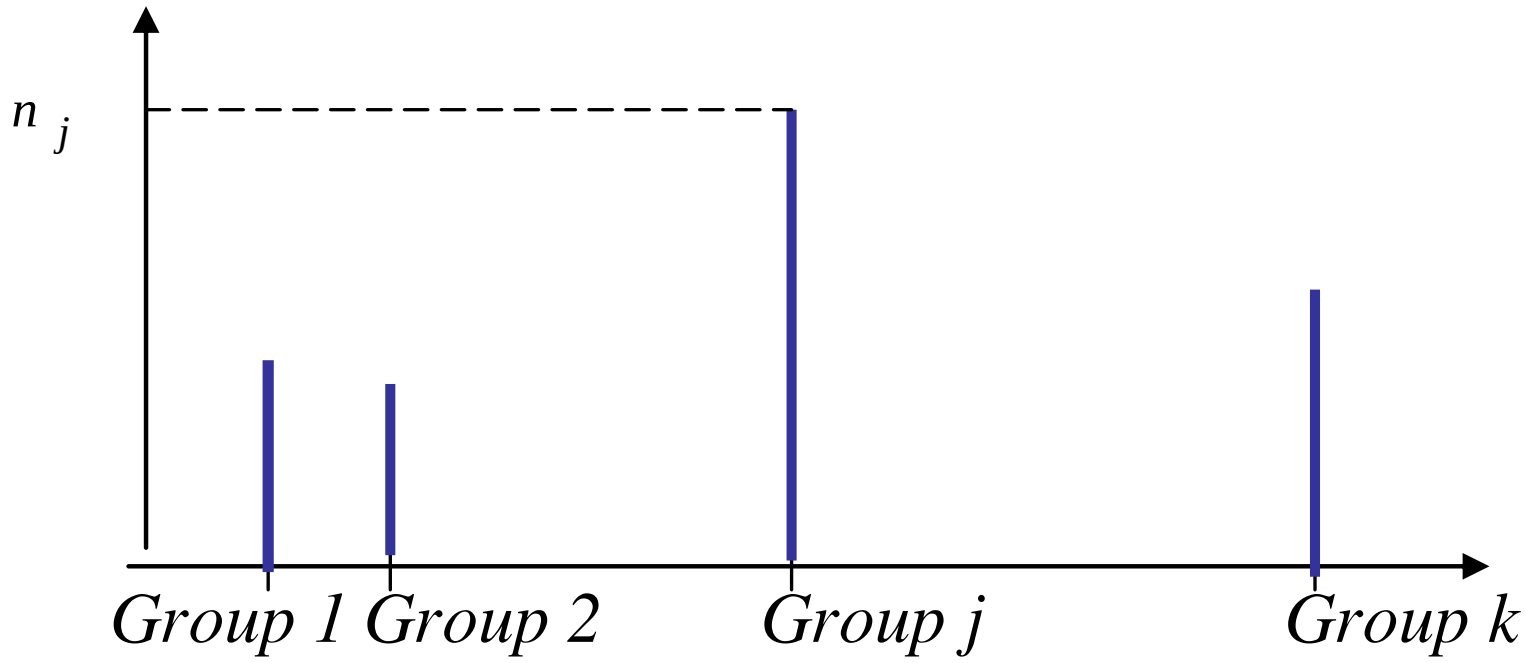
The table gives primary information:
- Frequency (amount of observations) in each group
- Distribution of data: proportion of observations number of each group, …

Example 1. To interview question “How often do you go to theater?”, from 148 interviewee, 47 answered “Never”, 71 “Rarely”, 24 “Sometime” and 6 “Frequently”. The data can be presented by frequency table:

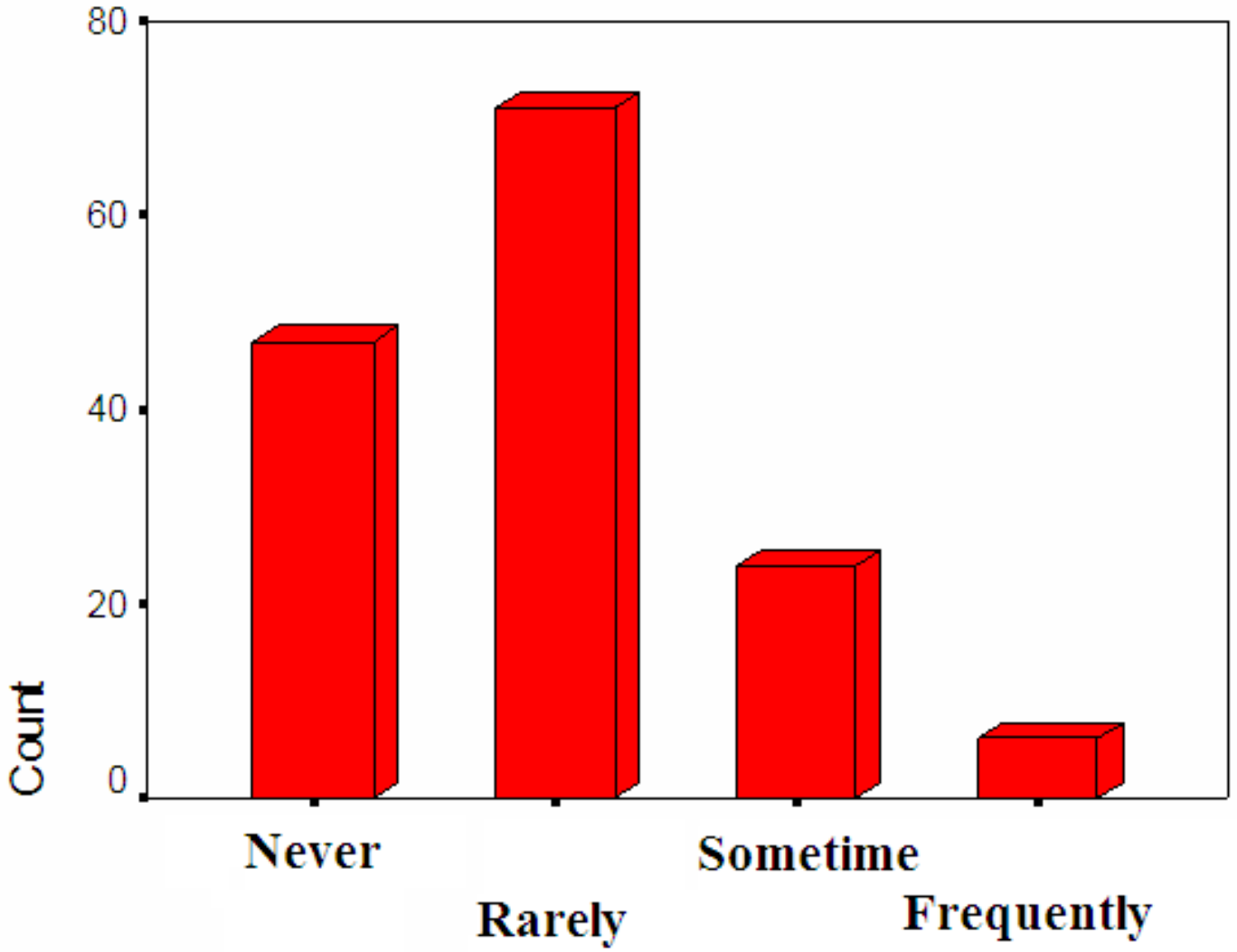
	Never	Rarely	Sometime	Frequently	Total
N	**47**	**71**	**24**	**6**	**148**
%	**31.8 %**	**48.0 %**	**16.2 %**	**4.1 %**	**100.0 %**

ii) Bar chart

Provides evident picture of qualitative variable distribution:

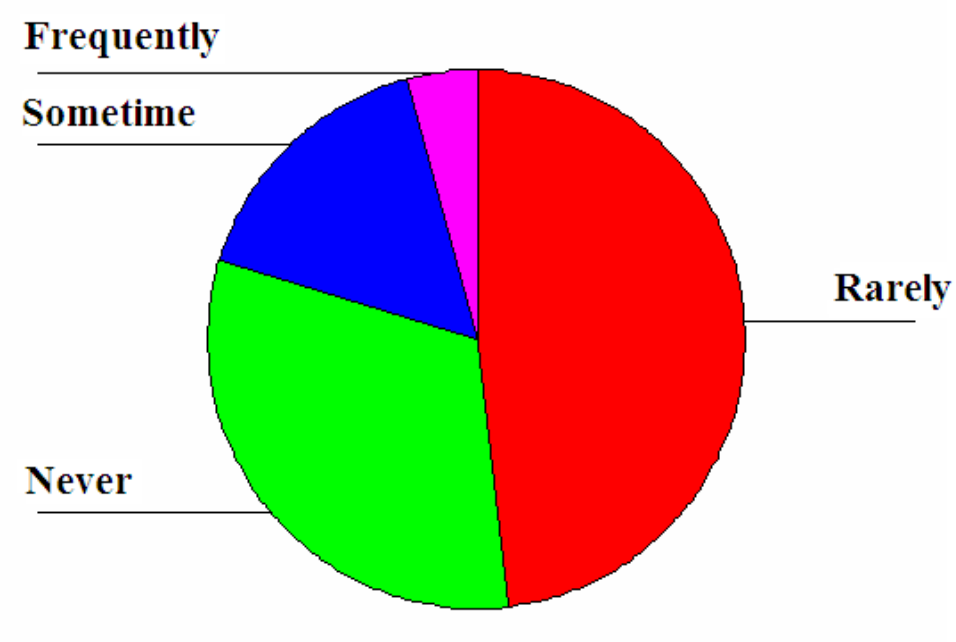


In the graph, the height of each bar is proportional to observation number of the corresponding group



iii) Pie chart

Presents proportions (percentages) of observations numbers of groups in total numberof all observations in the sample



Area of each part in the chart is proportional to the observations number of corresponding group.

B. Describing a quantitative variable

For a quantitative variable X with the sample of n observations

$$\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\},$$

where x_i is the value of X at observation i. Then several methods can be used to describe the variable:

i) Extremal values of variable

ii) *Parameters measuring central tendency of data*

iii) Parameters measuring variability of data

iv) Histogram

v) Percentiles

vi) Stem-leaf plot

vii) Box plot

i) Extremal values of variable

Max($\boldsymbol{X}$) - the largest value of data,
min($\boldsymbol{X}$) - the smallest value of data

Knowing the largest and the smallest values of data one can have some conclusions, i.g.

- The data values are contained in a reasonable interval or not?
- If there is some thing implying meaningless of the data?
- etc.

ii) Parameters measusing "central" tendency of data

1. Mean value of variable

$$\text{Mean}(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) ,$$

2. Average number of two extremal values

$$\text{ME}(\boldsymbol{X}) = \{\min(\boldsymbol{X}) + \text{Max}(\boldsymbol{X})\} / 2$$

3. Mode of sample: Mod ($\boldsymbol{X}$)

A data value whose frequency is higher than frequency of any neighbourhood value of data

4. *Median of sample*: Med($\boldsymbol{X}$)

Value whose cumulative frequency equals (approximately) 50%: the point of value dividing the sample into two "equal" parts, *1/2 lying in the left* and *1/2 lying in the right* hand side of this point.

If n elements of data are arranged in order:

$$x_1 \le x_2 \le \ldots \le x_n \ .$$

Then Med($\boldsymbol{X}$) = $x_{(n+1)/2}$ if n is odd, and

Med($\boldsymbol{X}$) = $[\, x_{n/2} + x_{(n/2)+1} \,]/2$ if n is even

Example:

$$\text{Med}(\{1, 2, 5\}) = 2 ,$$

$$\text{Med}(\{1, 3, 3, 3\}) = 3 ,$$

$$\text{Med}(\{1, 2, 5, 7\}) = 3.5$$

iii) Parameters measuring variability of data (sample)

1. Variance and Standard Deviation

a. *Variance*: $Var(\boldsymbol{X}) = \frac{1}{n}\sum_{k=1}^{n}(x_k - Mean(X))^2$

(sometime: $Var(\boldsymbol{X}) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - Mean(X))^2$)

b. *Standard Deviation*: $SD(\boldsymbol{X}) = \sigma(\boldsymbol{X}) = \sqrt{Var(X)}$.

iii) Parameters measuring variability of data (sample)

2. Median deviation:

$$EC(\boldsymbol{X}) = \frac{1}{n}\sum_{k=1}^{n} \left| x_k - Med(X) \right| .$$

3. Range of sample

$$w(\boldsymbol{X}) = \{\text{Max}(\boldsymbol{X}) - \min(\boldsymbol{X})\}$$

4. Mean deviation:

$$MD(\boldsymbol{X}) = \frac{1}{n}\sum_{k=1}^{n} \left| x_k - Mean(X) \right| .$$

Remarks

+ If

$$\text{Var}(\boldsymbol{X}) = 0$$

or $$\text{EC}(\boldsymbol{X}) = 0$$

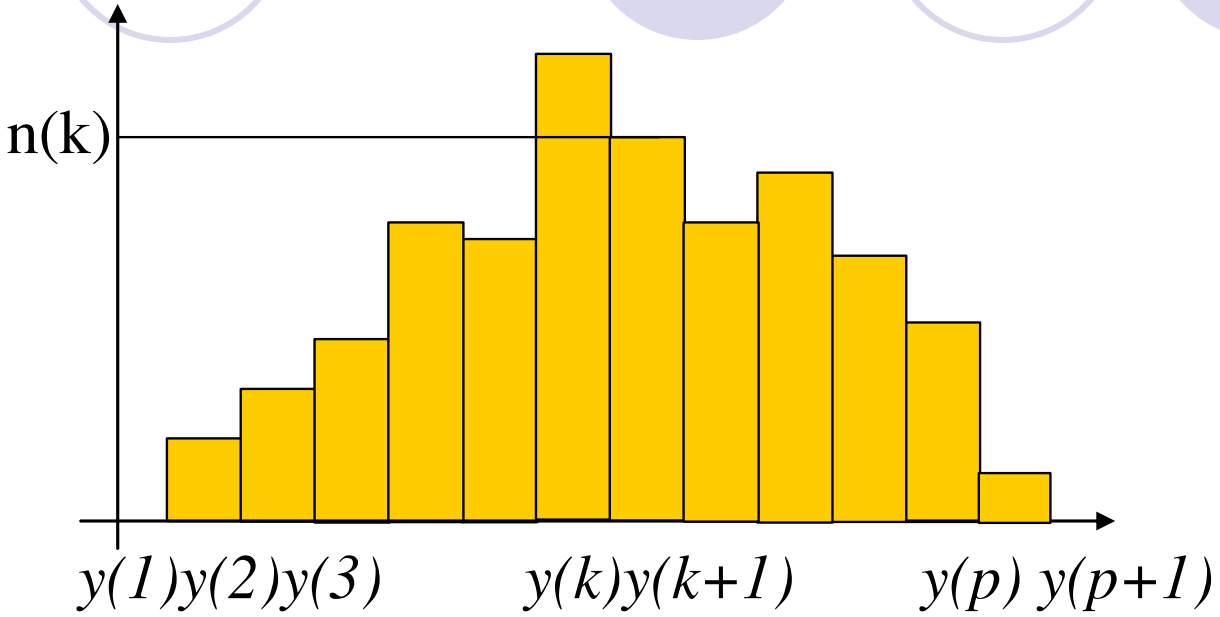
or $$w(\boldsymbol{X}) = 0$$

or $$\text{MD}(\boldsymbol{X}) = 0 \ ,$$

then all n elements of $\boldsymbol{X}$ are equal

+ Parameters $\text{Var}(\boldsymbol{X})$, $\text{MD}(\boldsymbol{X})$, $\text{EC}(\boldsymbol{X})$ and $w(\boldsymbol{X})$ measuring variability of sample are depent on scale of variable X

iv) Histogram



+ Let $y() = min(X)$, $y(p+1) = Max(X)$ and set A = $[y(1),y(p+1))$

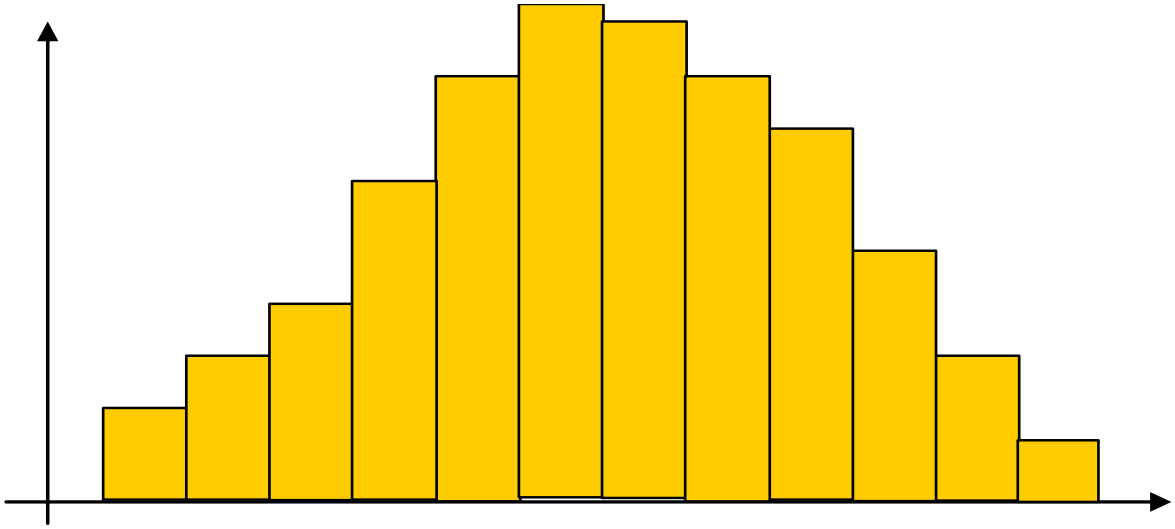
+ Divide A into p equal intervals

+ Determine n(k) as frequency of values of X belonging to the k-th interval

+ The height of k-th rectangle is taken proportionally to n(k)

Histogram types

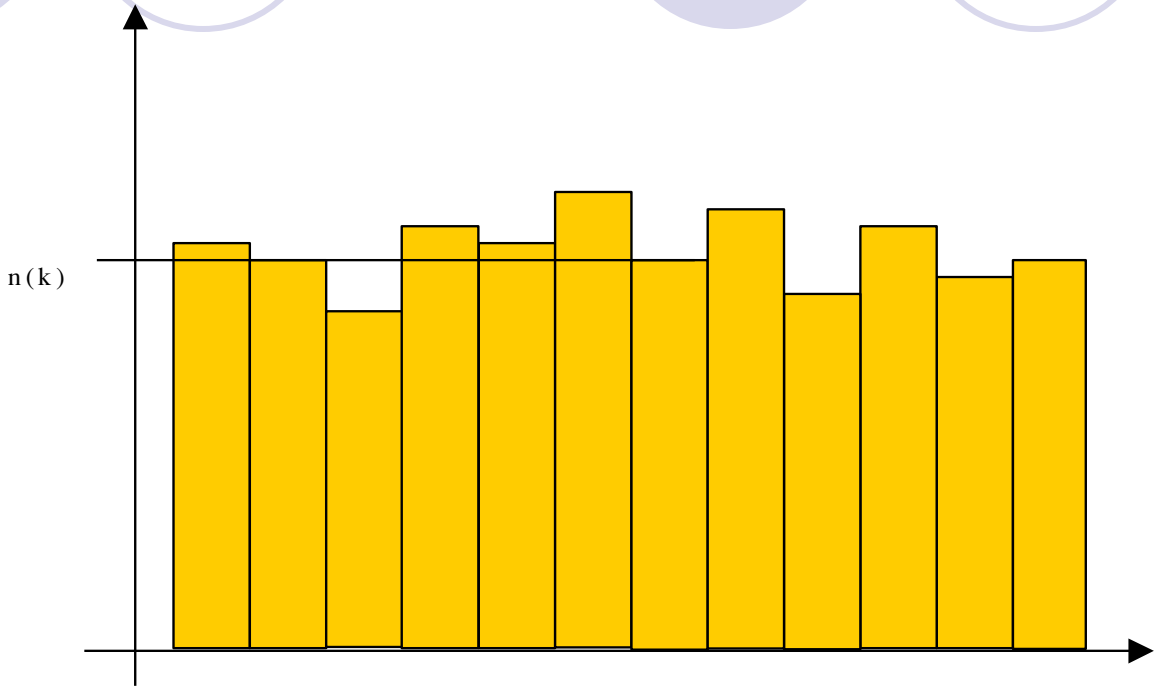
(1) Symmetric unimodal histogram



Properties:

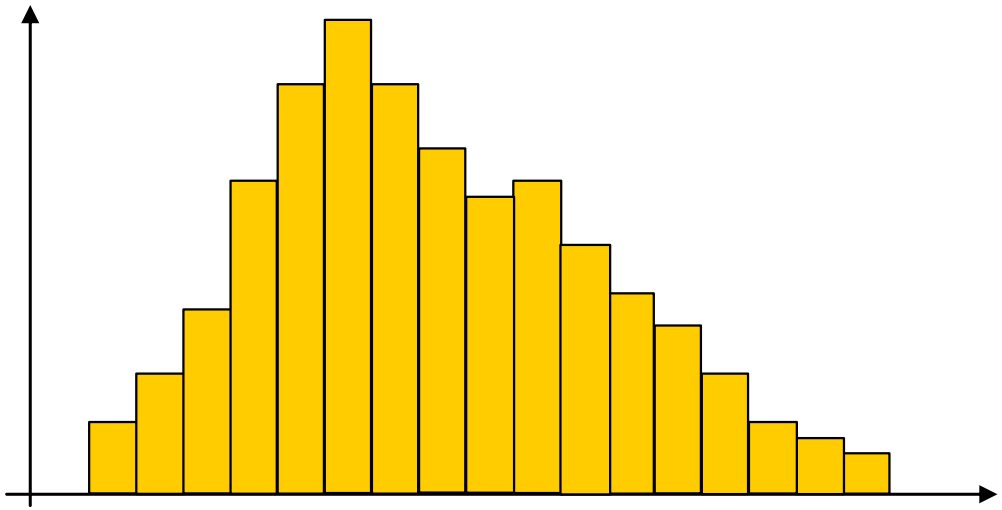
- Mode, mean and median values are close each to another
- The sample can be represented by two parameters: mean value Mean($\boldsymbol{X}$) and standard deviation σ ($\boldsymbol{X}$).

(2) Uniform histogram



All rectangles have almost the same height. Then the sample can be resumed by values of *min(X)*, *Max(X)* and the range *w(X)*

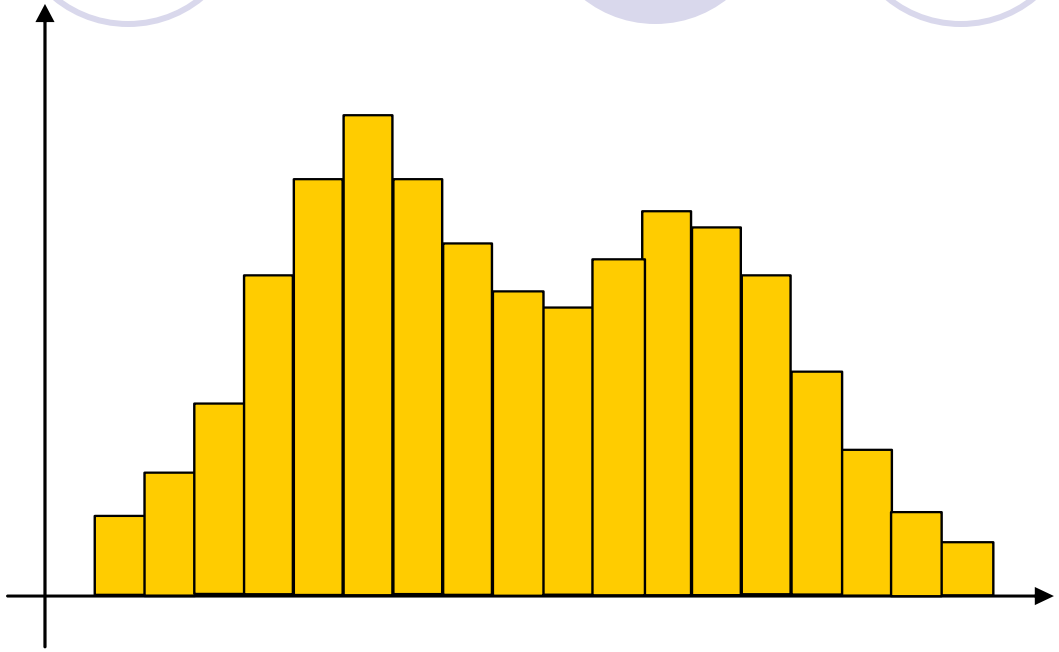
(3) Asymmetric unimodal histpgram



- Mode, median and mean values are different. The sample can not be resummed by mean value and standard deviation

→ Use some transformation for X (i.e. log(X)) to make (if possible) a variable with symmetric form

(4) Bi- or multimodal histogram



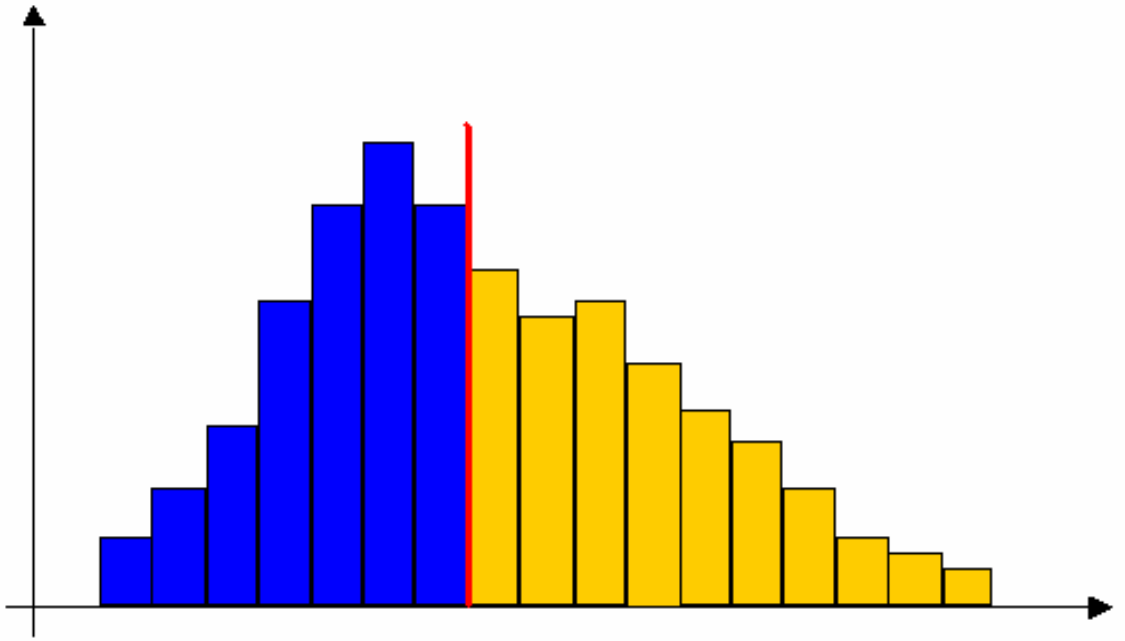
With multi-modal histogram, the data should be non-homogenous, may be a compound of several subpopulations

→ Separate the sample to two or many smaller sub-samples to study separately

v) Percentile

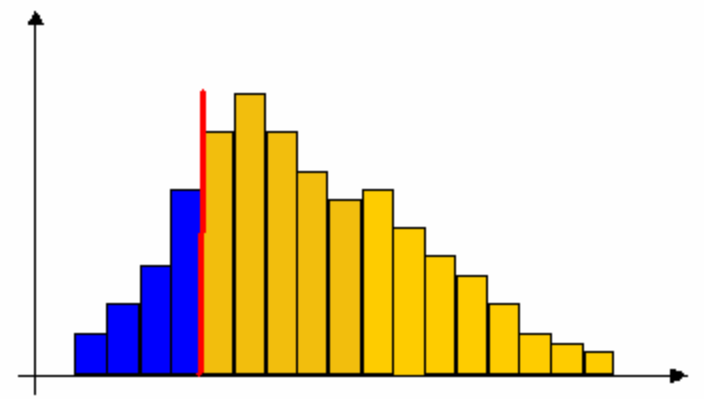
Percentile a% : - point dividing sample units into two parts: the left part contains a% amount of all observations in sample (then the right part contains (100-a)% amount of observations)

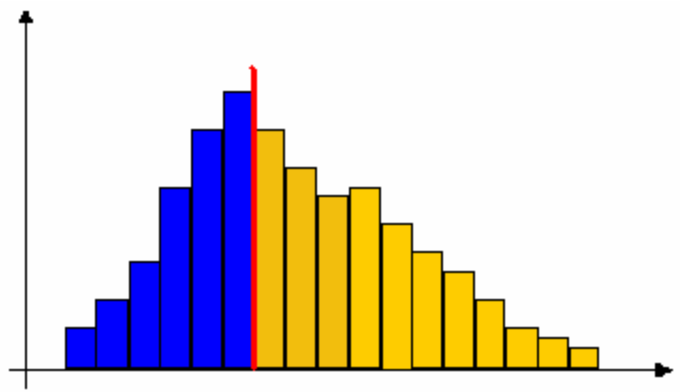
Median = percentile 50%, dividing the sample to 2 equal parts, each contains 1/2 amount of sample units

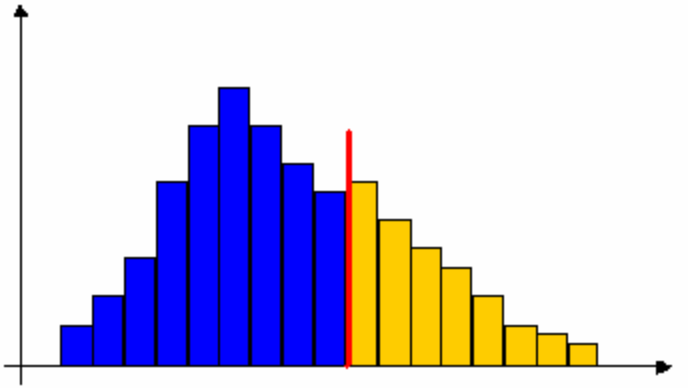


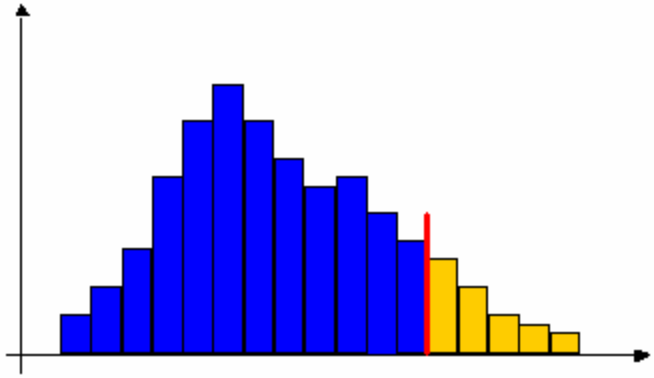
Special cases

Quintiles: percentiles 20%, 40%, 60% and 80%, dividing the sample into 5 equal parts

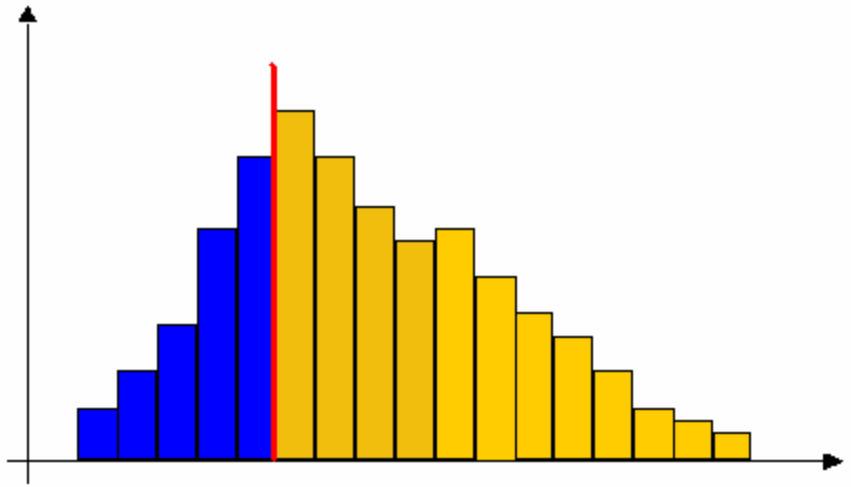


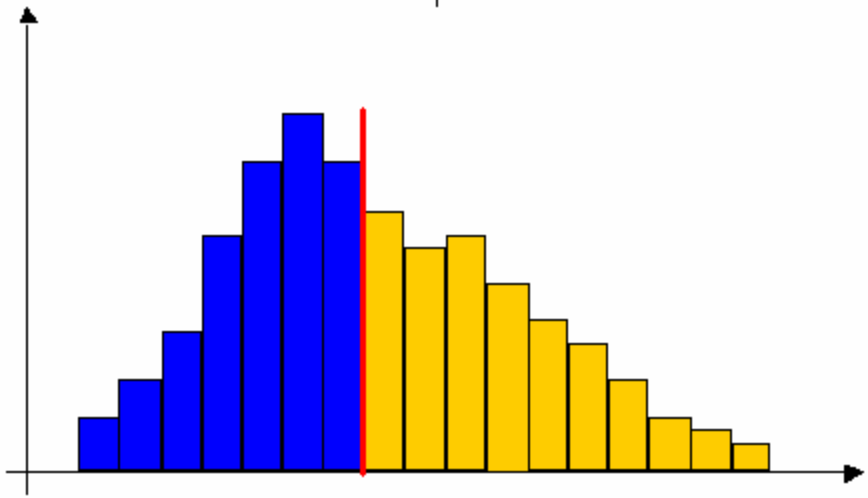


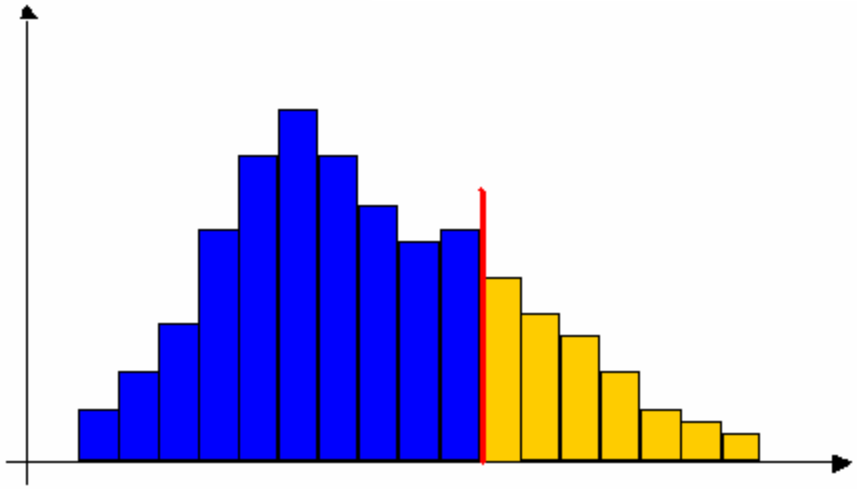




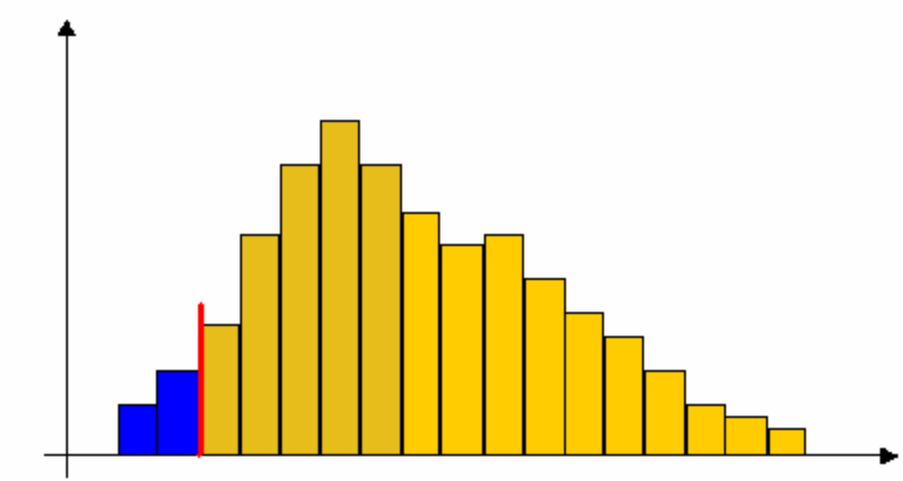
Quartiles: percentiles 25%, 50% and 75%, divide the sample into 4 equal parts

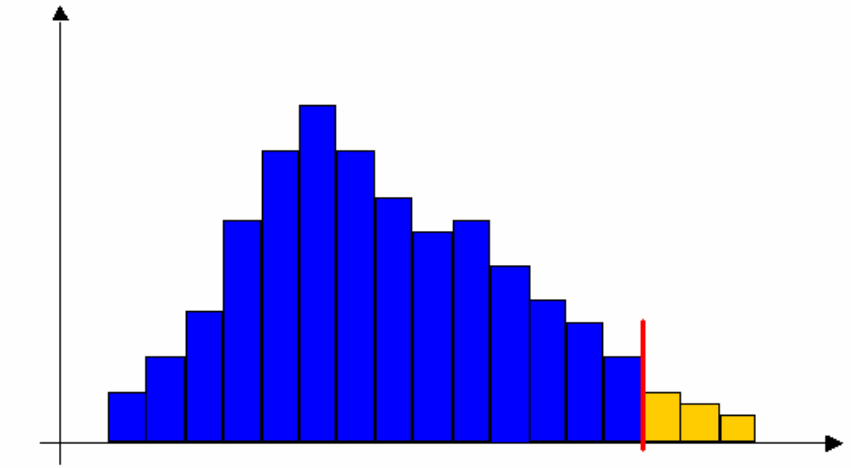






Percentiles 5% and 95%





vi) Stem-leaf Plot

Example: Weight of children in Uong Bi hospital

```
  Weight Stem-and-Leaf Plot

 Frequency     Stem &  Leaf

      3.00        9 .  005
      3.00       10 .  003
      6.00       11 .  005578
     10.00       12 .  0000005688
     12.00       13 .  000000000001
     14.00       14 .  00000035557889
     14.00       15 .  00000000004559
     17.00       16 .  00000000445555567
     18.00       17 .  00000000000245559
     21.00       18 .  000000000000255556678
     18.00       19 .  00000000055555778
     33.00       20 .  000000000000000000022333455667799
     21.00       21 .  000000000000055555555
     28.00       22 .  0000000000000000002555566778
     25.00       23 .  0000000000000000000002459
     21.00       24 .  000000000000000002448
     12.00       25 .  000000000000
```

Weight Stem-and-Leaf Plot

Frequency	Stem	&	Leaf	
3.00	9	.	005	3
3.00	10	.	003	6
6.00	11	.	005578	12
10.00	12	.	0000005688	22
12.00	13	.	000000000001	34
14.00	14	.	00000035557889	48
14.00	15	.	00000000004559	62
17.00	16	.	00000000445555567	79
18.00	17	.	00000000000245559	97
21.00	18	.	000000000000255556678	118
18.00	19	.	00000000055555778	136
33.00	20	.	000000000000000000002233345566799	140
21.00	21	.	000000000000055555555	107
28.00	22	.	0000000000000000002555566778	86
25.00	23	.	0000000000000000000002459	58
21.00	24	.	000000000000000002448	33
12.00	25	.	000000000000	12

Notes

Stem-leaf plot is very practical and provides a lot of information like:

- Range of data,
- Distribution shape of data,
- Sample is symmetric or not,
- Where the data is concentrated,
- If there are some outliers of data,
- Smallest, largest values of data, ...

In the plot, data has been arranged in a order and performs a figure look like a histogram

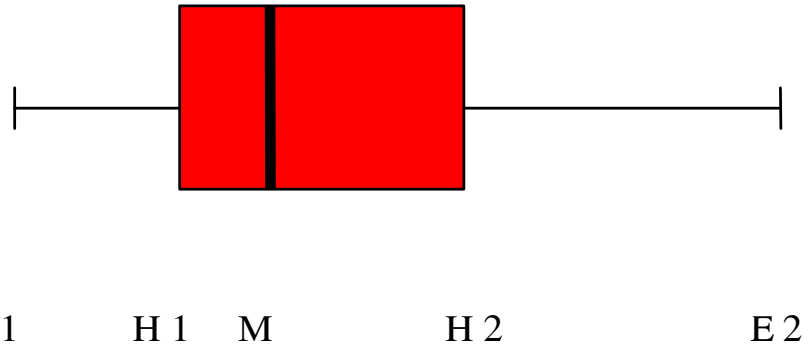
How to draw stem-leaf plot

Step 1. Primarily determine how many digits contained in each value (number) of data. Then separate the digits in each number to 2 part: *heading digits* and *driving digits*

Step 2. Write out in column *heading* in increasing (or decreasing) order, perform **stem** of "tree"

Step 3 For each value of data, write *driving digits* on the row of corresponding heading digits, perform **leaves** of "tree"

vii) Box plot



Box plot is defined by 5 characteristic values of data:

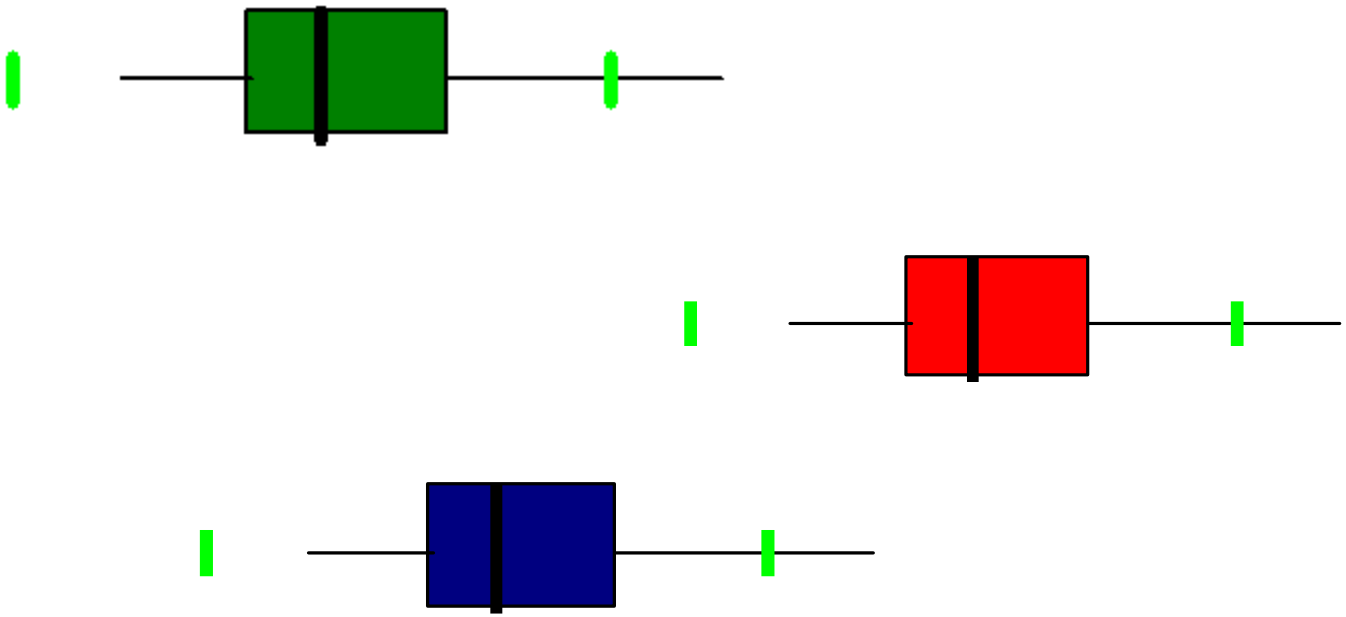
- Median M

- Quartiles H1 (25%) and H2 (75%),

- Octiles E1 (12.5%), E2 (87.5%).

2) Compare populations

Setting several box plots or stem-leaf plots each beside other, we can compare correspondent populations to see if there is any difference between populations



Excercise

Use SPSS , EXCEL to describe qualitative and quantitative variables by tables, charts, plot