DATA DESCRIPTION

Road Map

- Why data description?
- Describing 1 qualitative variable
- Describing relation between 2 qualitative variables
- Describing 1 quantitative variable
- Describing relation between 2 quantitative variables

Why data description?

- Primarily describe specific characteristics of data
- Summaries of data presented in a form that is easy to understand
- Investigate remarkable features of data, using those features to choose suitable model for data analysis
- Find out abnormal observations, outliers, mistakes or errors. Then clean the data before doing further analysis

Simple data description methods

- (a) Describe 1 qualitative variable
  - Qualitative variable with ***k*** values classifies the observations into ***k*** groups: K_1, K_2, …, K_k
  - Observations in each group have one same value of the variable

→ Numbers of observations in those groups represent the main feature of the data

Simple data description methods

- Summarizing Categorical Data
  - 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, …, n_k observations respectively ($n_1 + n_2 + … + n_k = n$).
- The variable can be represented 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_k/N) * 100%

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

i) Frequency/percentage table

- Relative Frequency Distribution:
  - The relative frequency of a class is the fraction or proportion of the total number of data items belonging to the class
  - A relative frequency distribution is a tabular summary of a set of data showing the relative frequency for each class

- Percent Frequency Distribution
  - The percent frequency of a class is the relative frequency multiplied by 100
  - A percent frequency distribution is a tabular summary of a set of data showing the percent frequency for each class

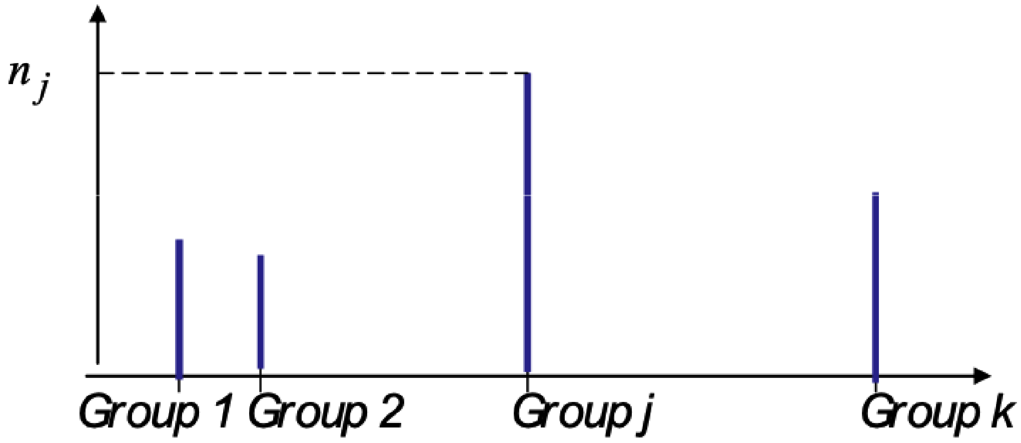
i) Frequency/percentage table

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

	Never	Rarely	Sometimes	Frequently	Total
N	47	71	24	6	148
%	31.8	48	16.2	4	100%

ii) Bar chart

- Provides evident picture of qualitative variable distribution:



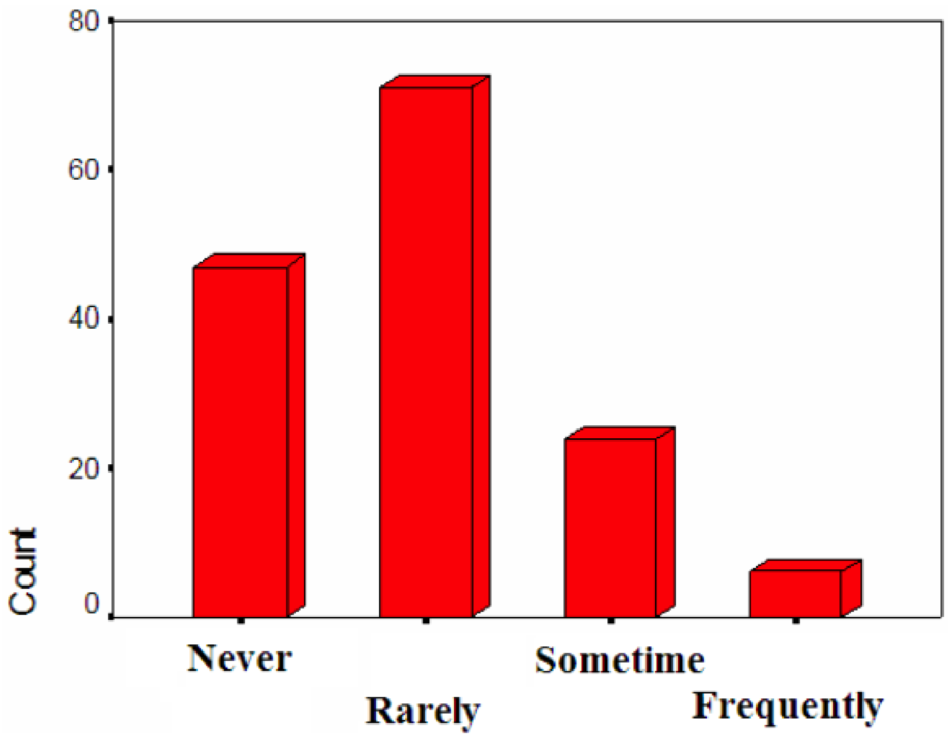
- In the graph, the height of each bar is proportional to observation number of the corresponding group

ii) Bar chart

- A bar chart is a graphical display for depicting qualitative data
- On one axis (usually the horizontal axis), we specify the labels that are used for each of the classes
- A frequency, relative frequency, or percent frequency scale can be used for the other axis (usually the vertical axis)
- Using a bar of fixed width drawn above each class label, we extend the height appropriately
- The bars are separated to emphasize the fact that each class is a separate category

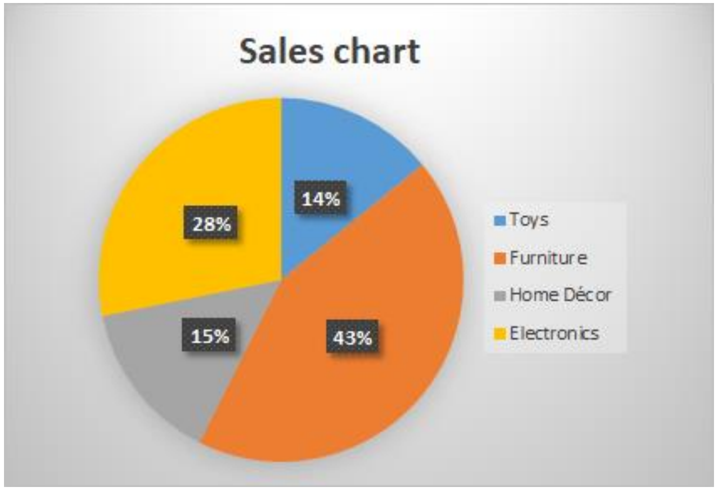
ii) Bar chart

- Bar chart of Example 1:



iii) Pie chart

- Presents proportions (percentages) of observation numbers of groups in total number of all observations in the sample



- Area of each part in the chart is proportional to the observation number of the corresponding group

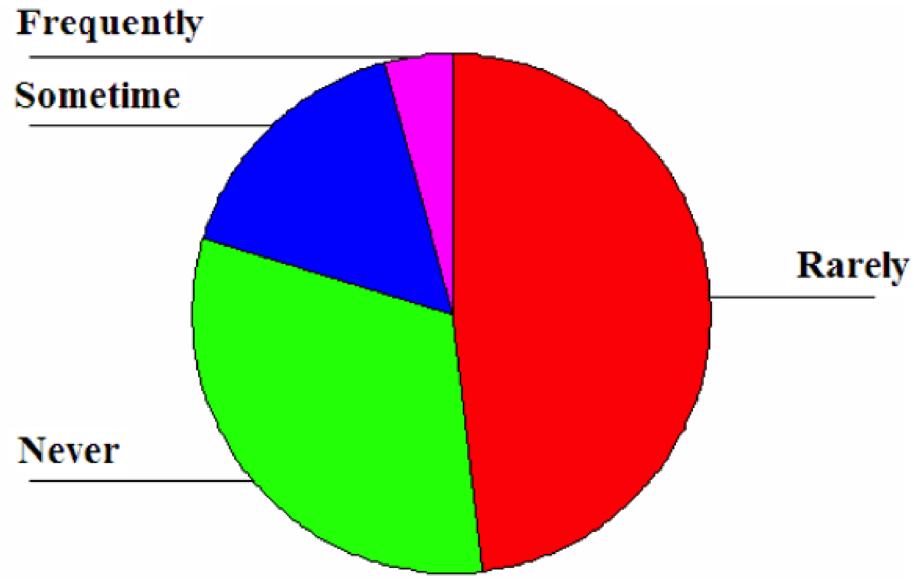
iii) Pie chart

- The pie chart is a commonly used graphical display for presenting relative frequency and percent frequency distributions for categorical data

- First draw a circle; then use the relative frequencies to subdivide the circle into sectors that correspond to the relative frequency for each class

iii) Pie chart

- Pie chart of Example 1:

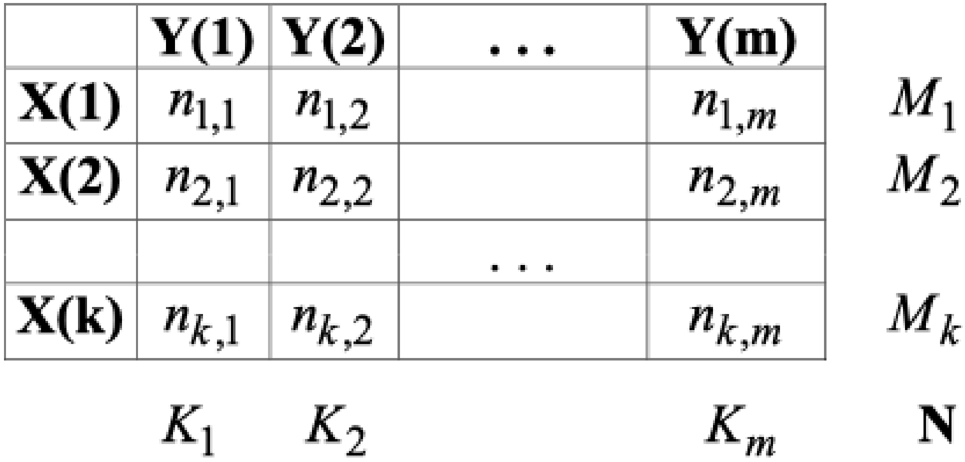


Using Excel to Describe Categorical Variables

- Excel has functions for computing the parameters of categorical variables and draw graphs:
  - SUM, COUNT, COUNTIF used to compute the frequencies
  - BAR, PIE used to draw the bar charts and pie charts

Simple data description methods

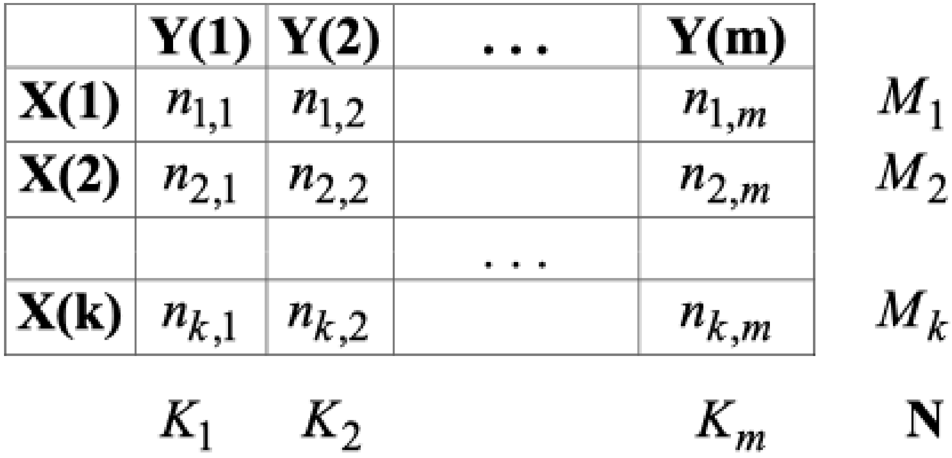
- (b) Describe relation between 2 qualitative variables
  - Cross table with levels of one variable in rows, levels of the second variable in columns:



	Y(1)	**Y(2)**	**. . .**	**Y(m)**	
X(1)	$n_{1,1}$	$n_{1,2}$		$n_{1,m}$	M_1
X(2)	$n_{2,1}$	$n_{2,2}$		$n_{2,m}$	M_2
			. . .		
X(k)	$n_{k,1}$	$n_{k,2}$		$n_{k,m}$	M_k
	K_1	K_2		K_m	**N**

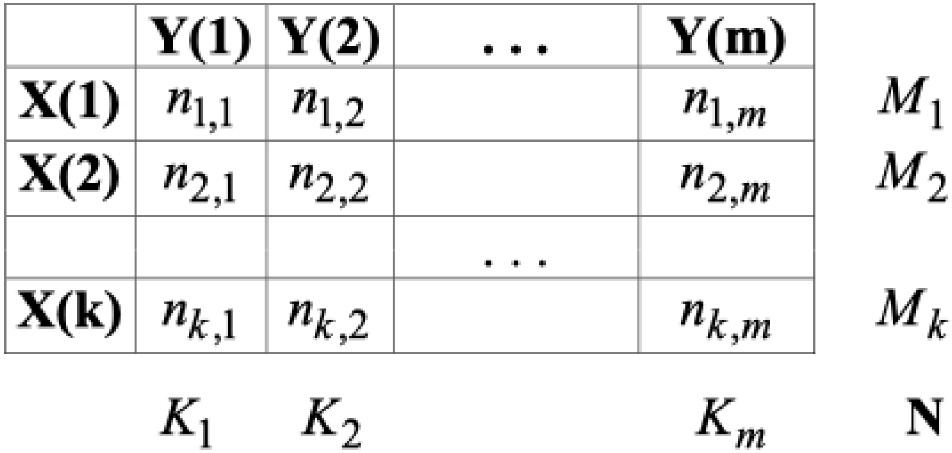
Cross table of two qualitative variables

- Usually:
  - the first variable (rows) is a independent (describing, cause, input) variable
  - And, the second variable (columns) is a dependent (descriptive, result, output) variable



	Y(1)	Y(2)	. . .	Y(m)	
X(1)	$n_{1,1}$	$n_{1,2}$		$n_{1,m}$	M_1
X(2)	$n_{2,1}$	$n_{2,2}$		$n_{2,m}$	M_2
			. . .		
X(k)	$n_{k,1}$	$n_{k,2}$		$n_{k,m}$	M_k
	K_1	K_2		K_m	N

Cross table of two qualitative variables



	Y(1)	Y(2)	. . .	Y(m)	
X(1)	$n_{1,1}$	$n_{1,2}$		$n_{1,m}$	M_1
X(2)	$n_{2,1}$	$n_{2,2}$		$n_{2,m}$	M_2
			. . .		
X(k)	$n_{k,1}$	$n_{k,2}$		$n_{k,m}$	M_k
	K_1	K_2		K_m	N

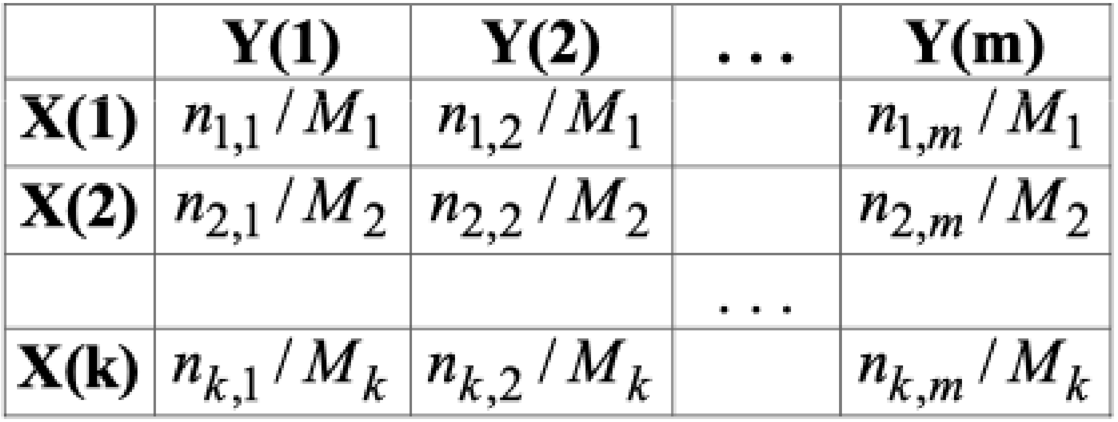
- In cell *ij* of table $n_{i,j}$: number of observations belonging simmutaneously to the level *i* of the first variable and to the the level *j* of the second variable,
- M_i : sum of all numbers in the row *i* (number of observations in *i-th* level of the first variable),
- K_j : sum of all numbers in the column *j* (number of observations in *j-th* level of the second variable),
- *N* : total number of all observations in the sample

Crosstabulation: Row or Column Percentages

- Converting the entries in the table into row percentages or column percentages can provide additional insight about the relationship between the two variables.

Crosstabulation: Row or Column Percentages

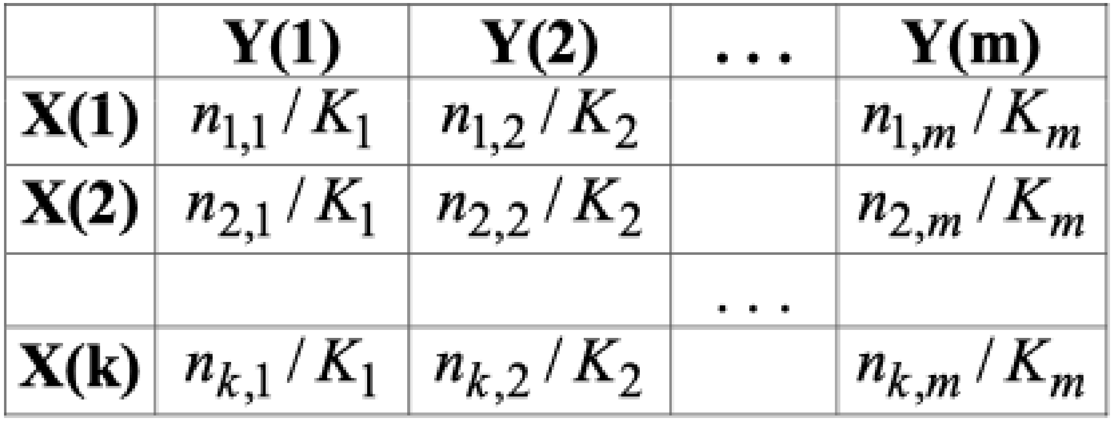
- Table with percentages % across rows: $n_{i,j} / M_i$ gives information about distribution of “output” variable Y in each level of “input” variable X



	Y(1)	Y(2)	. . .	Y(m)
X(1)	$n_{1,1} / M_1$	$n_{1,2} / M_1$		$n_{1,m} / M_1$
X(2)	$n_{2,1} / M_2$	$n_{2,2} / M_2$		$n_{2,m} / M_2$
			. . .	
X(k)	$n_{k,1} / M_k$	$n_{k,2} / M_k$		$n_{k,m} / M_k$

Crosstabulation: Row or Column Percentages

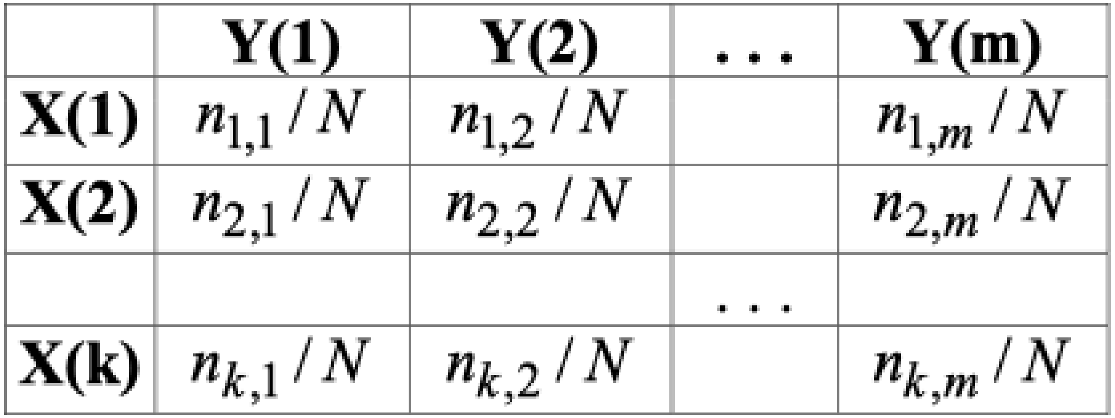
- Table with percentages % across columns: $n_{i,j}$ / K_j gives information about distribution of “input” variable X in each level of “output” variable Y



	Y(1)	**Y(2)**	**. . .**	**Y(m)**
X(1)	$n_{1,1} / K_1$	$n_{1,2} / K_2$		$n_{1,m} / K_m$
X(2)	$n_{2,1} / K_1$	$n_{2,2} / K_2$		$n_{2,m} / K_m$
			. . .	
X(k)	$n_{k,1} / K_1$	$n_{k,2} / K_2$		$n_{k,m} / K_m$

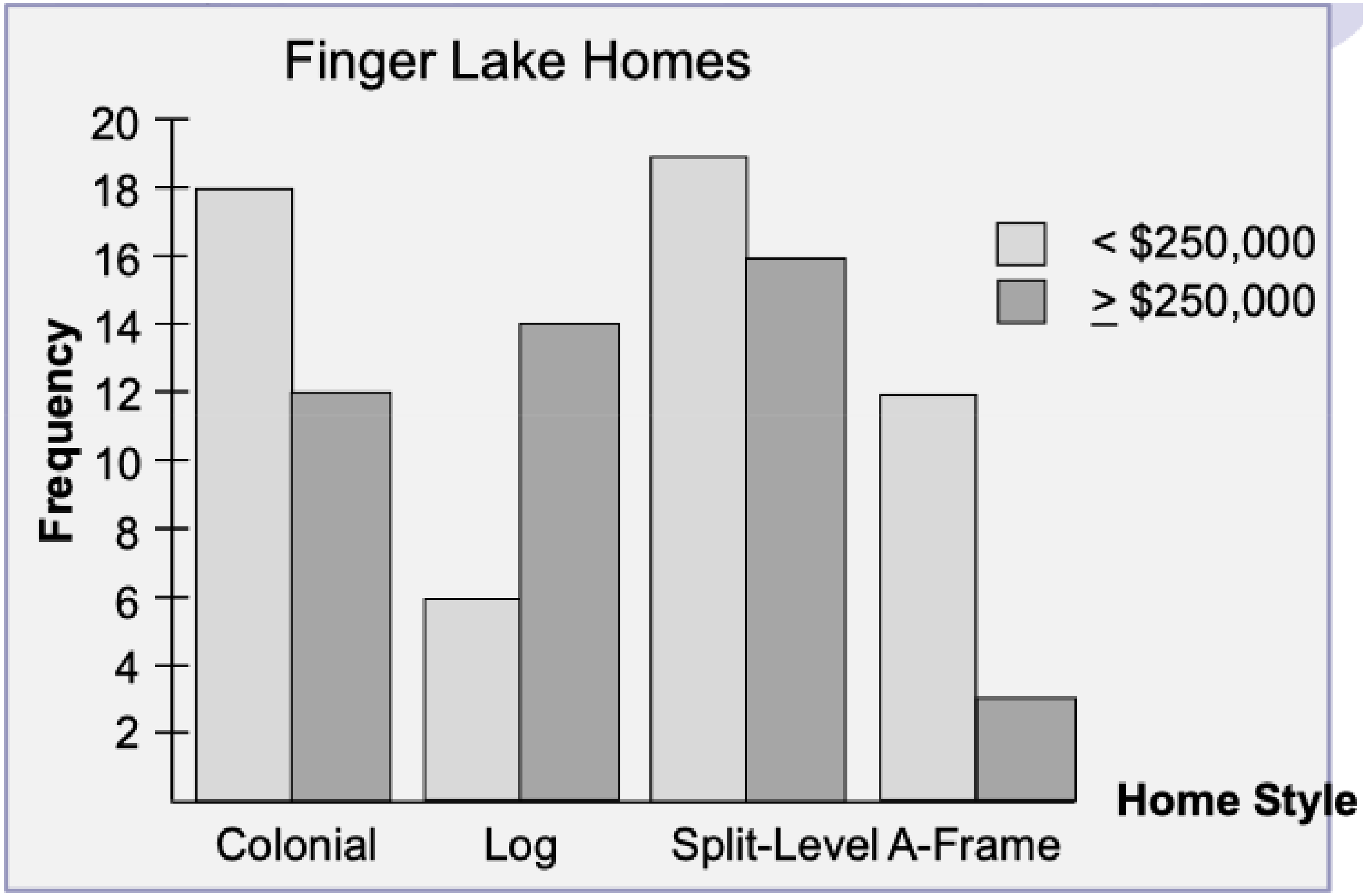
Crosstabulation: Row or Column Percentages

- Table with percentages % in whole sample: $n_{i,j} / N$ gives information about total distribution in sample



	Y(1)	Y(2)	. . .	Y(m)
X(1)	$n_{1,1} / N$	$n_{1,2} / N$		$n_{1,m} / N$
X(2)	$n_{2,1} / N$	$n_{2,2} / N$		$n_{2,m} / N$
			. . .	
X(k)	$n_{k,1} / N$	$n_{k,2} / N$		$n_{k,m} / N$

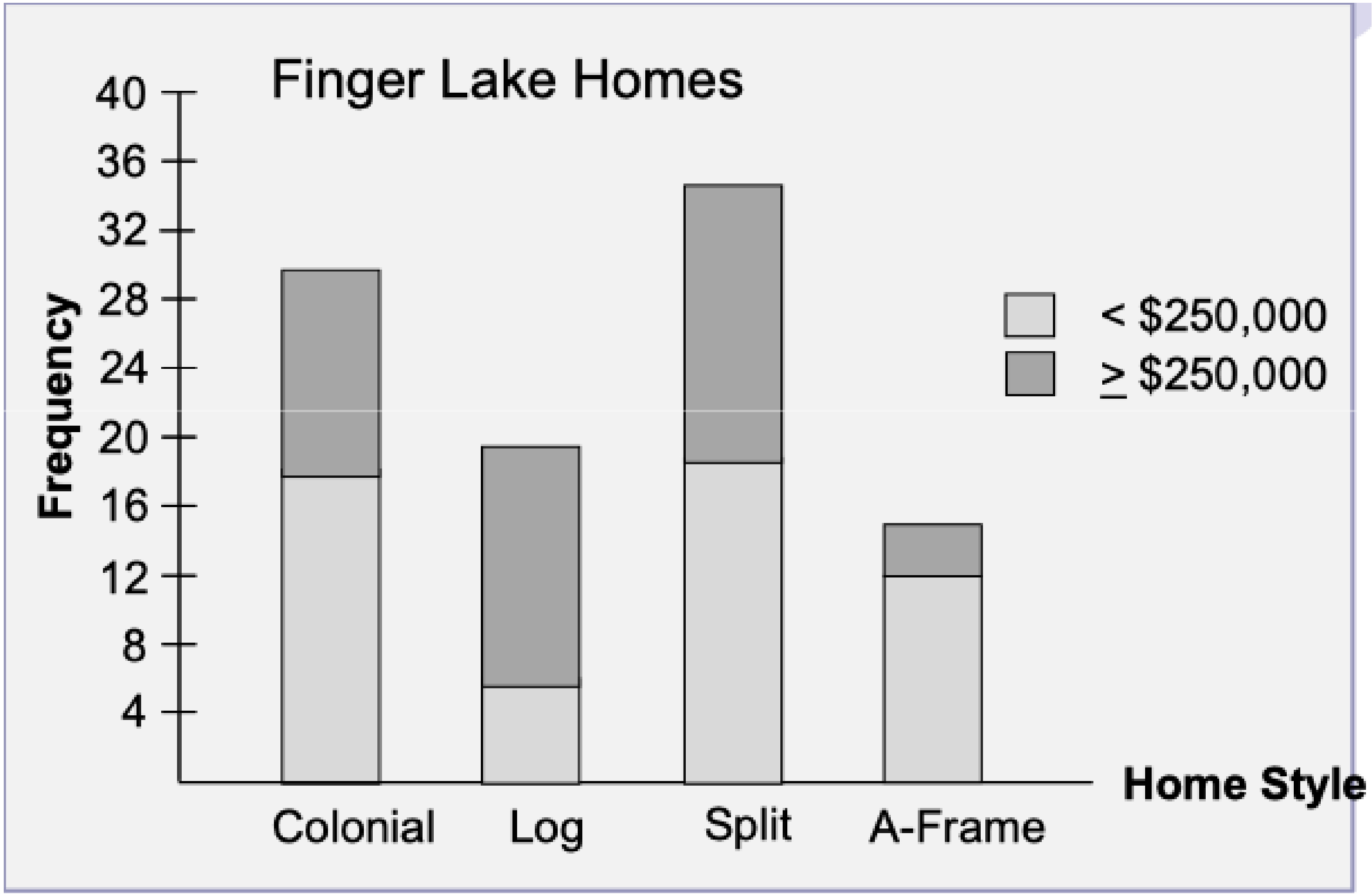
Side-by-Side Bar Chart



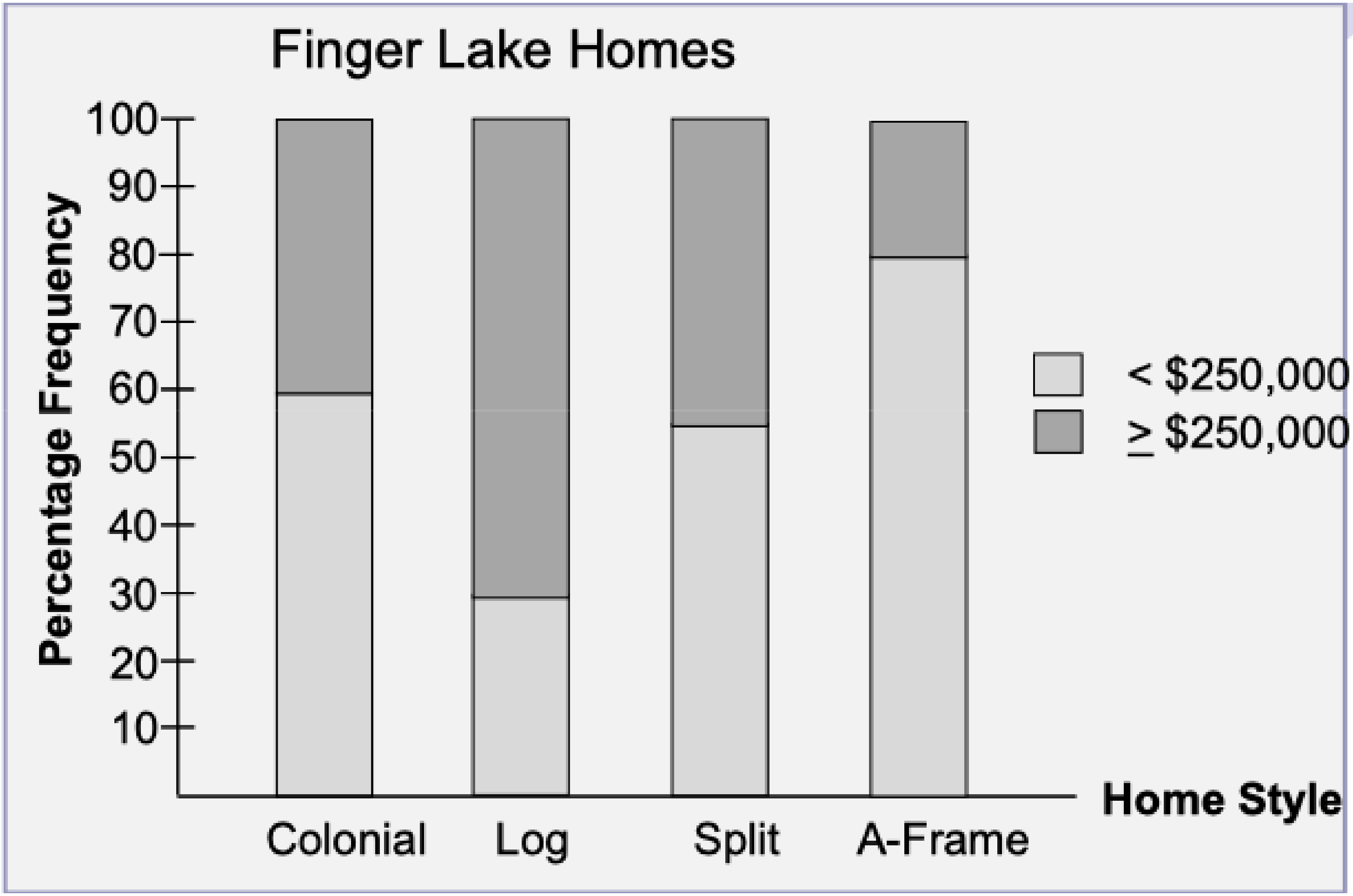
Stacked Bar Chart

- A stacked bar chart is another way to display and compare two variables on the same display
- It is a bar chart in which each bar is broken into rectangular segments of a different color
- If percentage frequencies are displayed, all bars will be of the same height (or length), extending to the 100% mark

Stacked Bar Chart



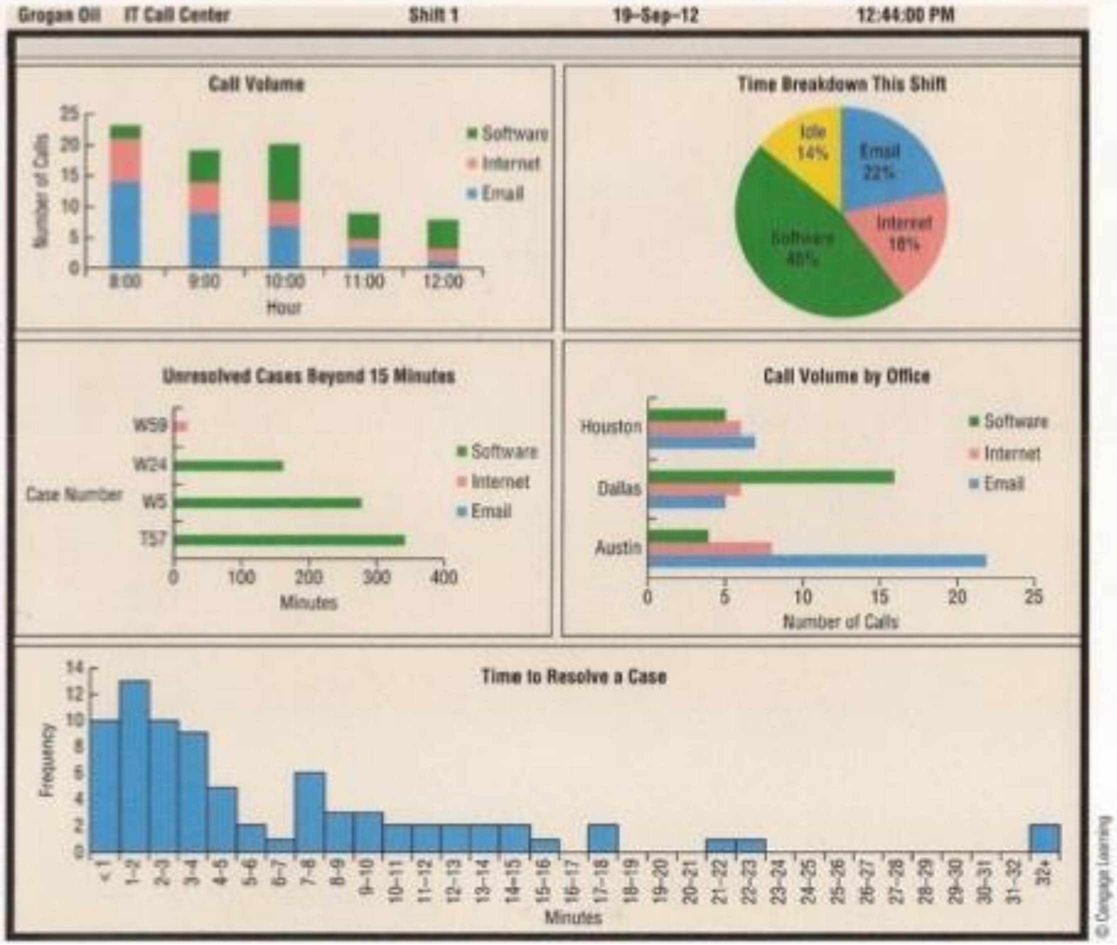
Stacked Bar Chart



Choosing the Type of Graphical Display

- Displays used to make comparisons:
  - Side-by-Side Bar Chart to compare two variables
  - Stacked Bar Chart to compare the relative frequency or Percent frequency of two categorical

Choosing the Type of Graphical Display: Example



Using Excel to Describe Categorical Variables

- Excel has functions for computing the parameters of categorical variables and draw graphs:
  - SUM, COUNT, COUNTIF, COUNTIFS used to compute the frequencies
  - BAR used to draw the bar charts

Simple data description methods

- (c) Describing a quantitative variable
  - For a quantitative variable X with the sample of n observations.

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

where x_i is the value of X at observation i

Methods to describe a quantitative variable

- i) Extreme values of variable
- ii) Parameters measuring central tendency of data
- iii) Parameters measuring variability of data
- iv) Histogram
- v) Quantile
- vi) Percentile
- vii) Stem-leaf plot
- viii) Box plot

i) Extreme values of variable

Max(X) – the largest value of data,

Min(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 measuring central tendency of data

- 1) Mean value of the variable

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

- 2) Average number of two extreme values

$$ME(X) = (min(X) + max(X)) / 2$$

ii) Parameters measuring central tendency of data

- 3) Mode of the sample: *Mode(X)*
  - There exist data value whose frequency is higher than frequency of any neighbourhood value of data.
  - The mode of a data set is the value that occurs with greatest frequency.
  - The greatest frequency can occur at two or more different values.
  - If the data have exactly two modes, the data are bimodal.
  - If the data have more than two modes, the data are multimodal.

ii) Parameters measuring central tendency of data

- 4) Median of the sample: *Med(X)*
  - Value whose cumulative frequency equals (approximately) 50%: the point of value dividing the sample into two “equal” parts, ½ lying in the left and ½ lying in the right hand side of this point.
  - If n elements of data are arranged in order:

$$x_1 \leq x_2 \leq \ldots \leq x_n$$

then Med(X) = $x_{(n+1)/2}$ if n is odd,

and

Med(X) = $x_{n/2}$ if n is even

Med(X) = $(x_{n/2} + x_{n/2+1})$

ii) Parameters measuring central tendency of data

- 4) Median of the sample: *Med(X)*
  - The median of a data set is the value in the middle when the data items are arranged in ascending order.
  - Whenever a data set has outlier extreme values, the median is the preferred measure of central location.
  - The median is the measure of location most often reported for annual income and property value data.
  - A few extremely large incomes or property values can inflate the mean.

ii) Parameters measuring central tendency of data

- 4) Median of the sample: *Med(X)*
  - Median example:

    Med({1, 2, 5}) = 2

    Med({1, 3, 3, 3}) = 3

    Med({1, 2, 5, 7}) = 2

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

iii) Parameters measuring variability of data

- 1) Range
- 2) Interquartile range
- 3) Variance
- 4) Standard deviation

iii) Parameters measuring variability of data

- 1) Range
  - The range of a data set is the difference between the largest and smallest data values:

$$\text{Range}(X) = \text{Largest value} – \text{Smallest value}$$

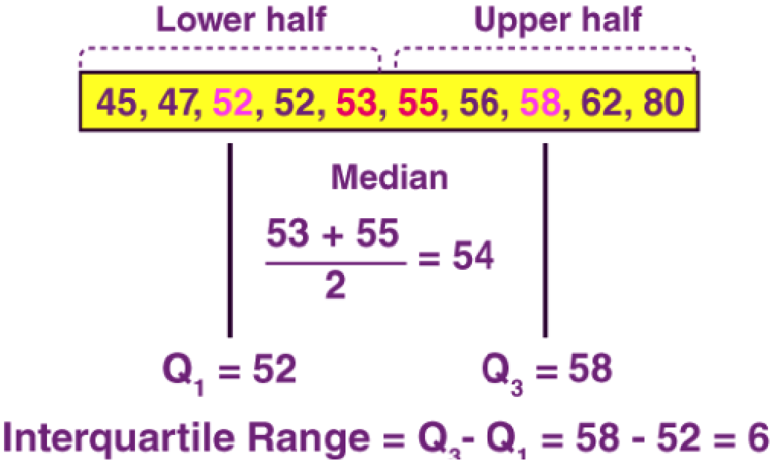
  - It is the simplest measure of variability.
  - It is very sensitive to the smallest and largest data values.

iii) Parameters measuring variability of data

- 2) Interquartile range
  - The interquartile range of a data set is the difference between the third quartile and the first quartile.

    Interquartile range(X) = Third Quartile – First Quartile

    = Q3 – Q1
  - It is the range for the middle 50% of the data.
  - It overcomes the sensitivity to extreme data values.



Lower half

Upper half

45, 47, 52, 52, 53, 55, 56, 58, 62, 80

Median

$$\frac{53 + 55}{2} = 54$$

$Q_1 = 52$

$Q_3 = 58$

Interquartile Range = $Q_3 - Q_1 = 58 - 52 = 6$

iii) Parameters measuring variability of data

3) Variance

- The variance defines a measure of the spread or dispersion within a set of data. It is calculated as the average of the squared differences between each data value and the mean.
- Sample variance measures the dispersion in the sample

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

n is sample size, diving by n-1 corrects the bias in estimation

- Population variance measures the dispersion in the entire population

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

N is population size

iii) Parameters measuring variability of data

- 4) Standard deviation
  - The standard deviation of a data set is the positive square root of the variance.
  - Data sets with a small standard deviation are tightly grouped around the mean, whereas a larger standard deviation indicates the data is more spread out.
  - Standard deviation is measured in the same units as the data, making it more easily interpreted than the variance.
  - Sample standard deviation:

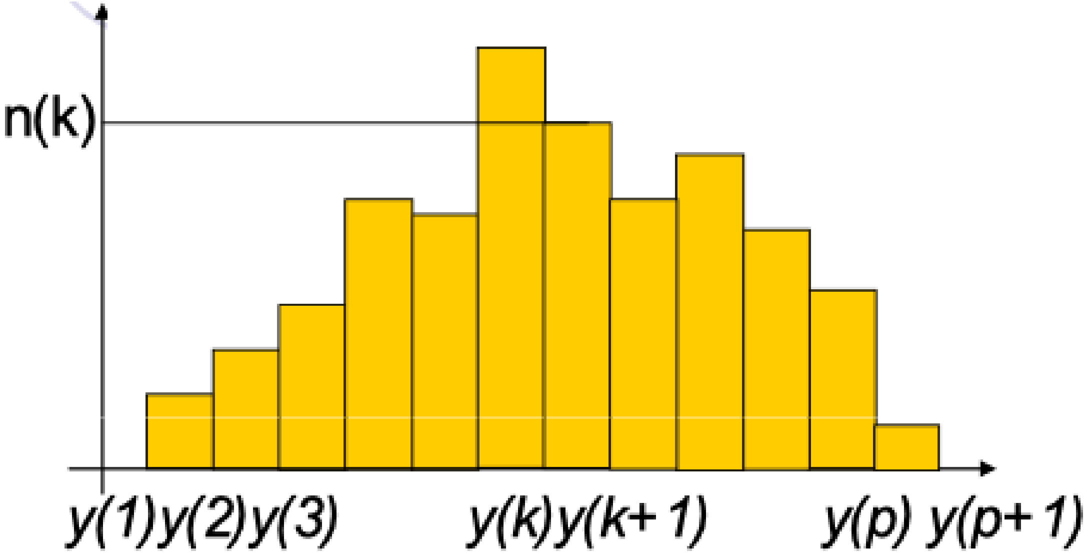
$$s = \sqrt{s^2}$$

  - Population standard deviation:

$$\sigma = \sqrt{\sigma^2}$$

iv) Histogram

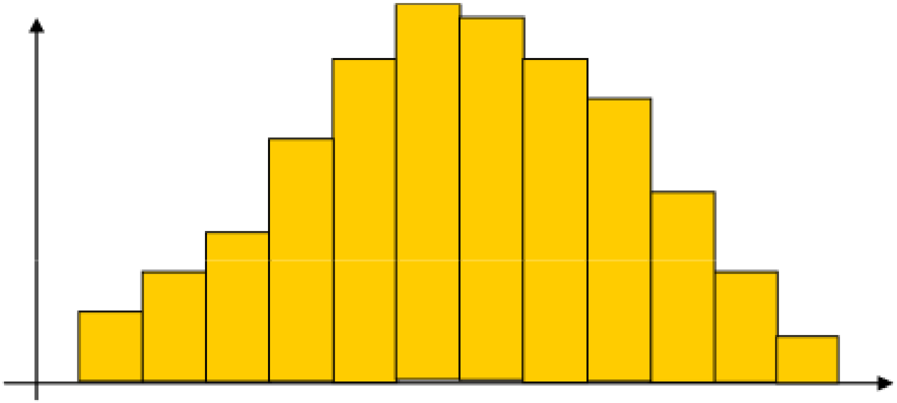
- Let y(1) = min(X), y(p+1) = max(X), and set A = [y(1), y(p+1)]
- Divide A into p equal intervals
- Determine n(k) as frequency of values of X belonging to the k-th interval
- The height of k-th rectangle is taken proportionally to n(k)



iv) Histogram

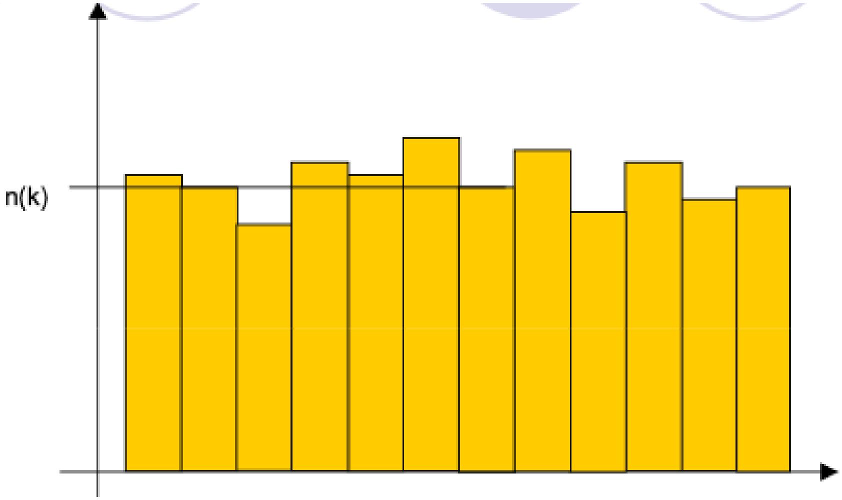
- Histogram types:
  - 1) Symmetric unimodal histogram
  - 2) Uniform histogram
  - 3) Asymmetric unimodal histogram
  - 4) Bi- or multimodal histogram

1) Symmetric unimodal histogram



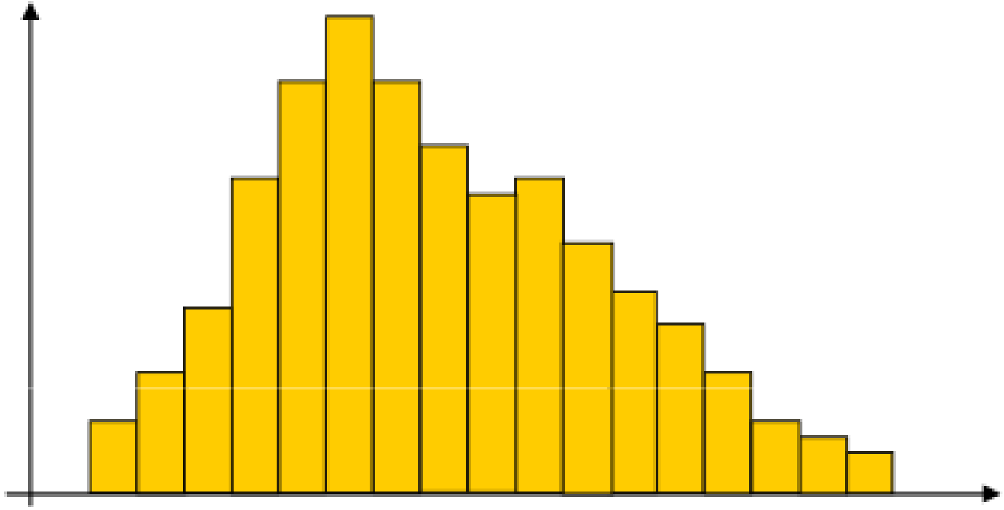
- Properties:
  - Mode, mean and median values are close each to another
  - The sample can be represented by two parameters: mean value and standard deviation: mean(X) and sd(X)

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(X)

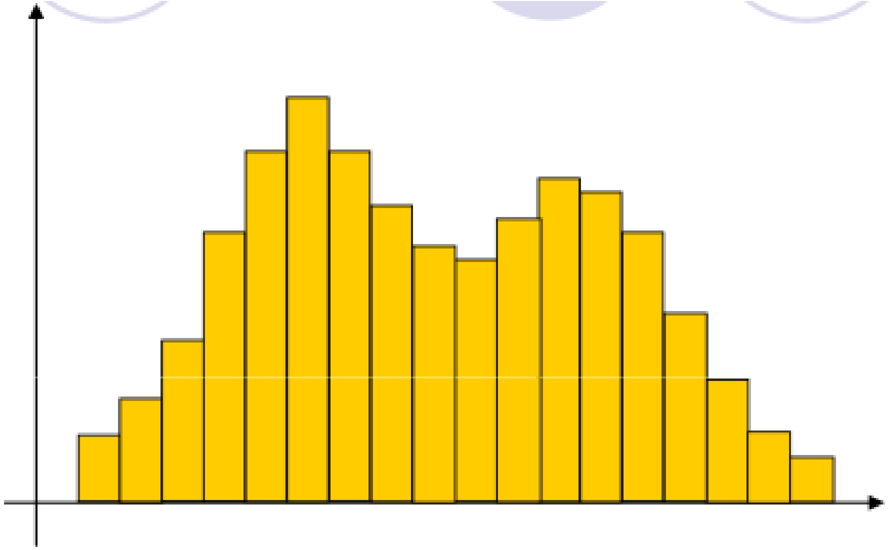
3) Asymmetric unimodal histogram



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

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

4) Asymmetric unimodal histogram

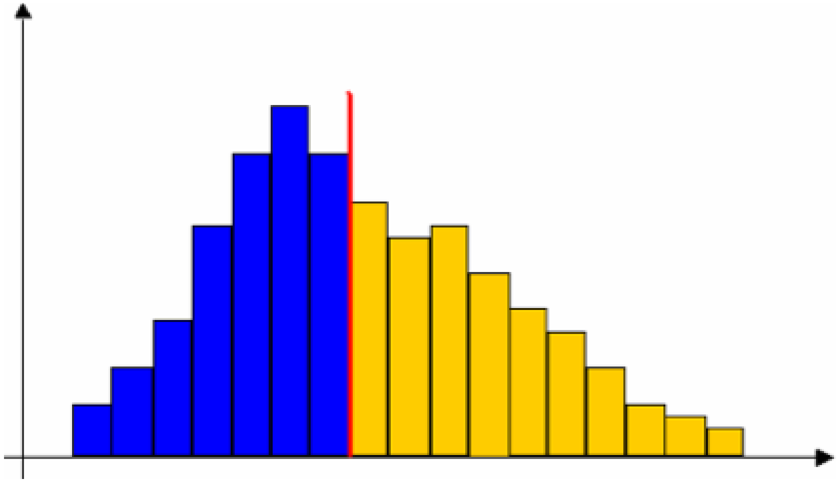


- With multi-modal histogram, the data should be non-homogenous, may be a compound of several subpopulations

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

v) Quantile

- Quantiles are values that split sorted data or a probability distribution into equal parts. A p-quantile divides sorted data into p equal sizes.
- x% quantile is the value dividing sample units into two parts: the left part contains x% amount of all observations in sample (then the right part contains (100-x)% amount of observations in the sample).
- Median = 50% quantile, dividing the sample into 2 equal parts, each contains ½ amount of sample units.



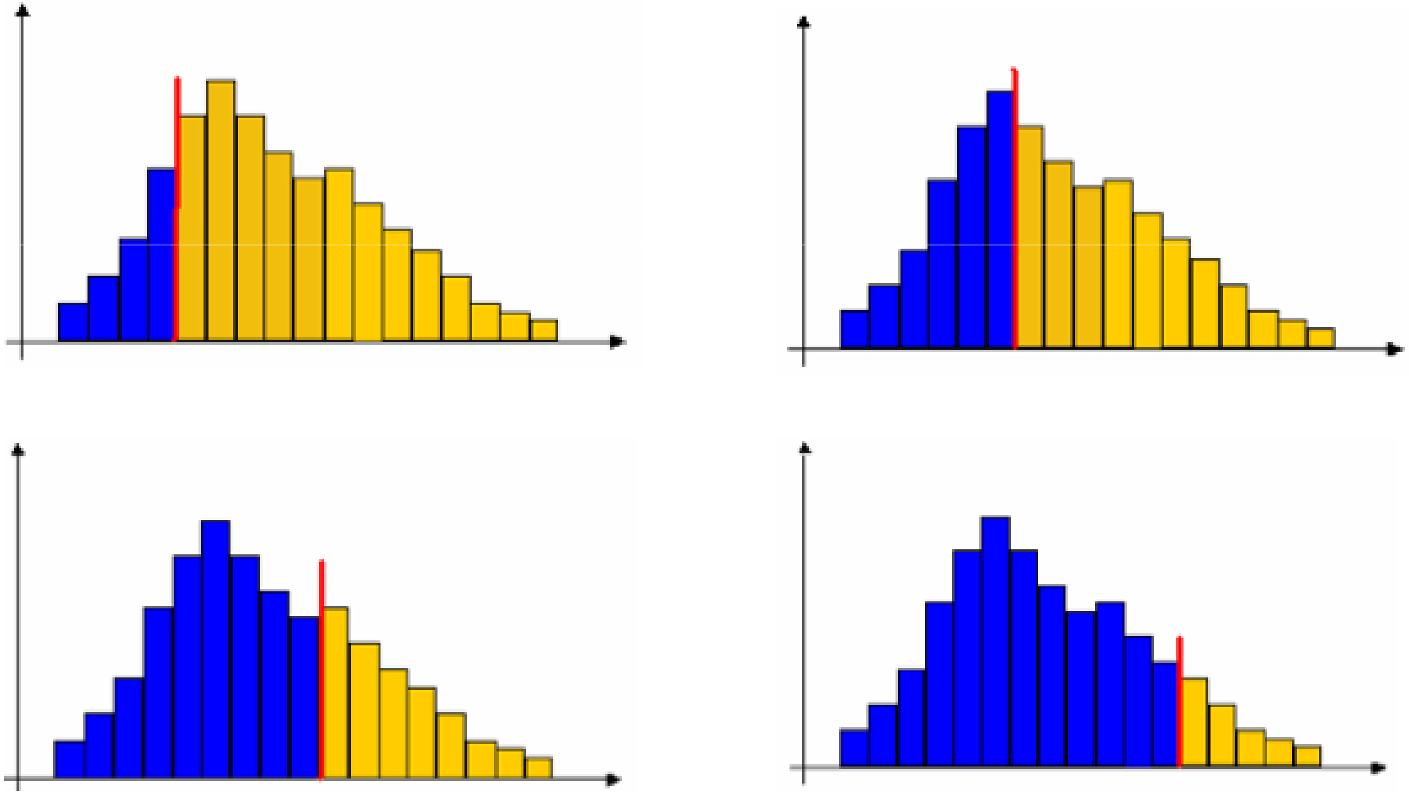
vi) Percentile

- The p_th percentile of a dataset is a value such that at least p percent of the observations are less than or equal to this value.
- Arrange the data in ascending order, a percentile provides information about how the data are spread over the interval from the smallest value to the largest value.
- Compute the location of the pth percentile, L_p as follows:

$$L_p = p/100\ (n+1)$$

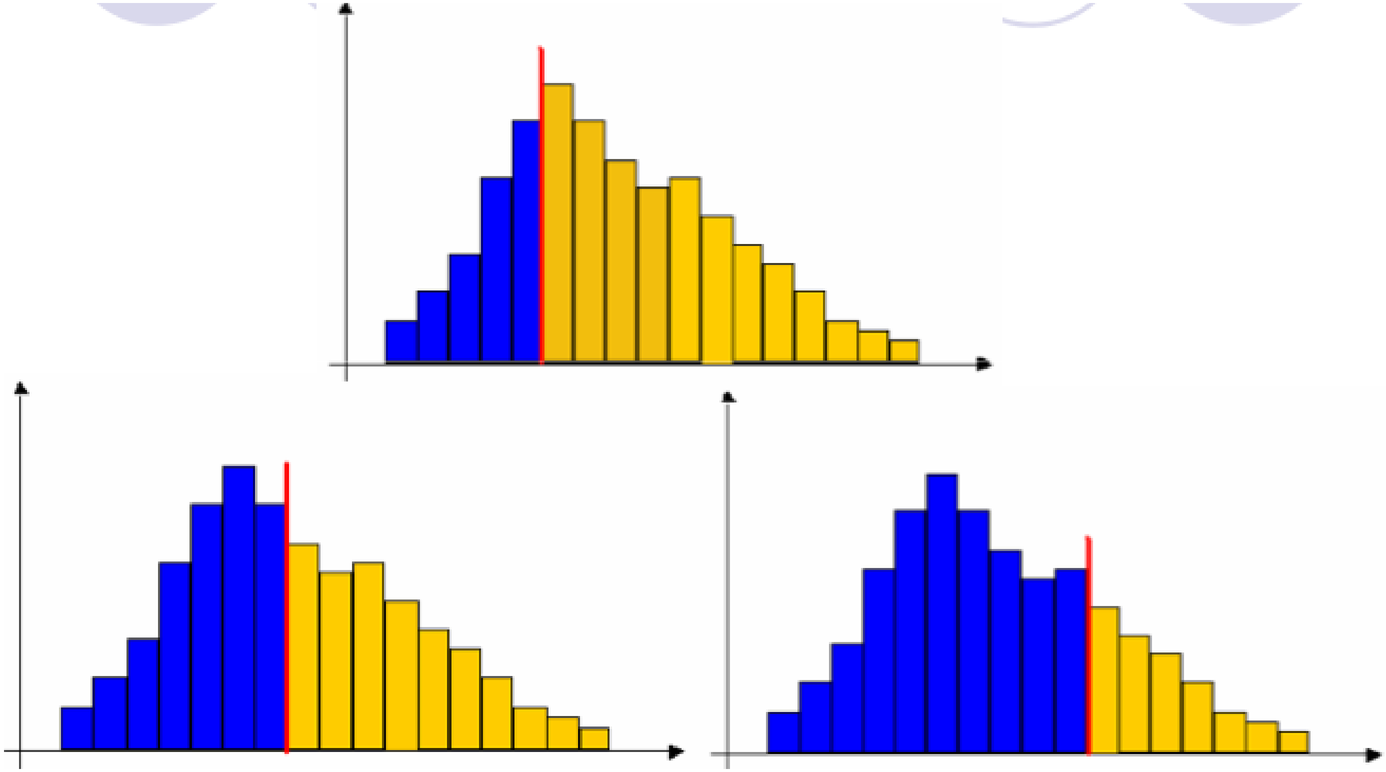
Special cases

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



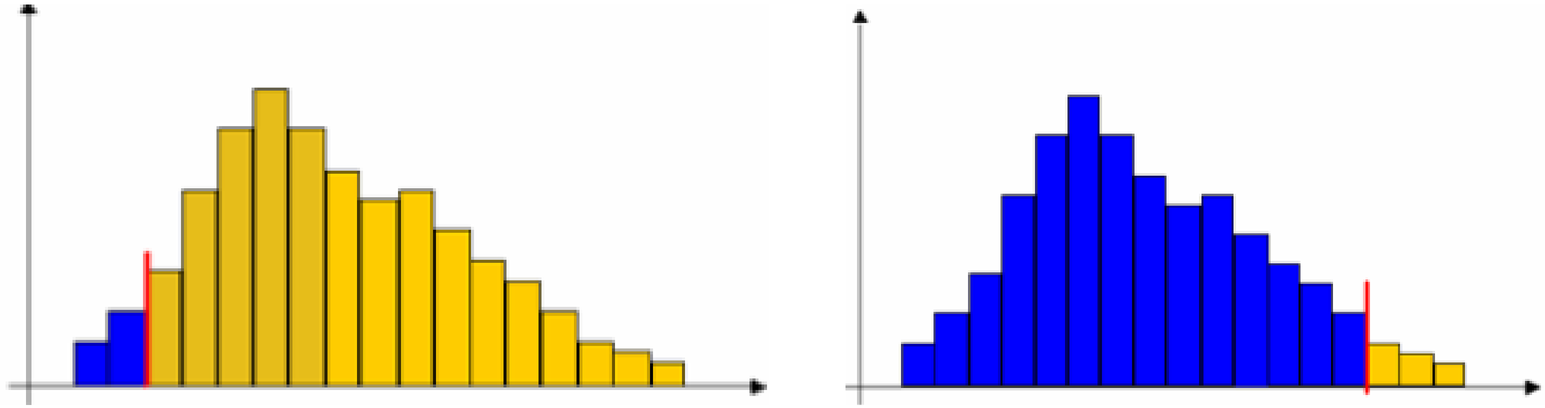
Special cases

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



Special cases

- 5% and 95% percentiles

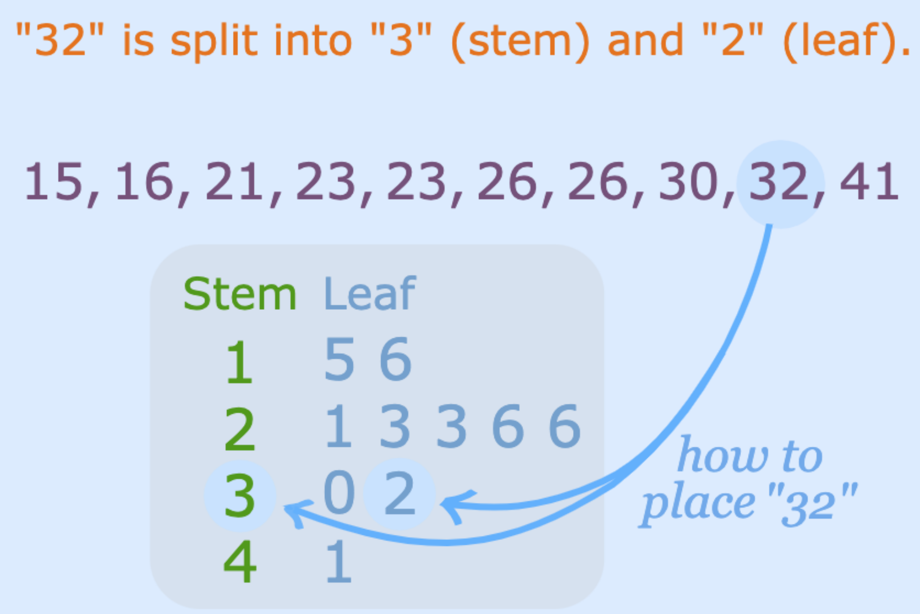


Quartiles vs. Quantiles

- Quartiles are specific percentiles
  - First Quartile = 25th Percentile
  - Second Quartile = 50th Percentile = Median
  - Third Quartile = 75th Percentile

vii) Stem-leaf plot

- A Stem and Leaf Plot is a special table where each data value is split into a "stem" (the first digit or digits) and a "leaf" (usually the last digit). It allows the grouping of the data as well as showing all the original data



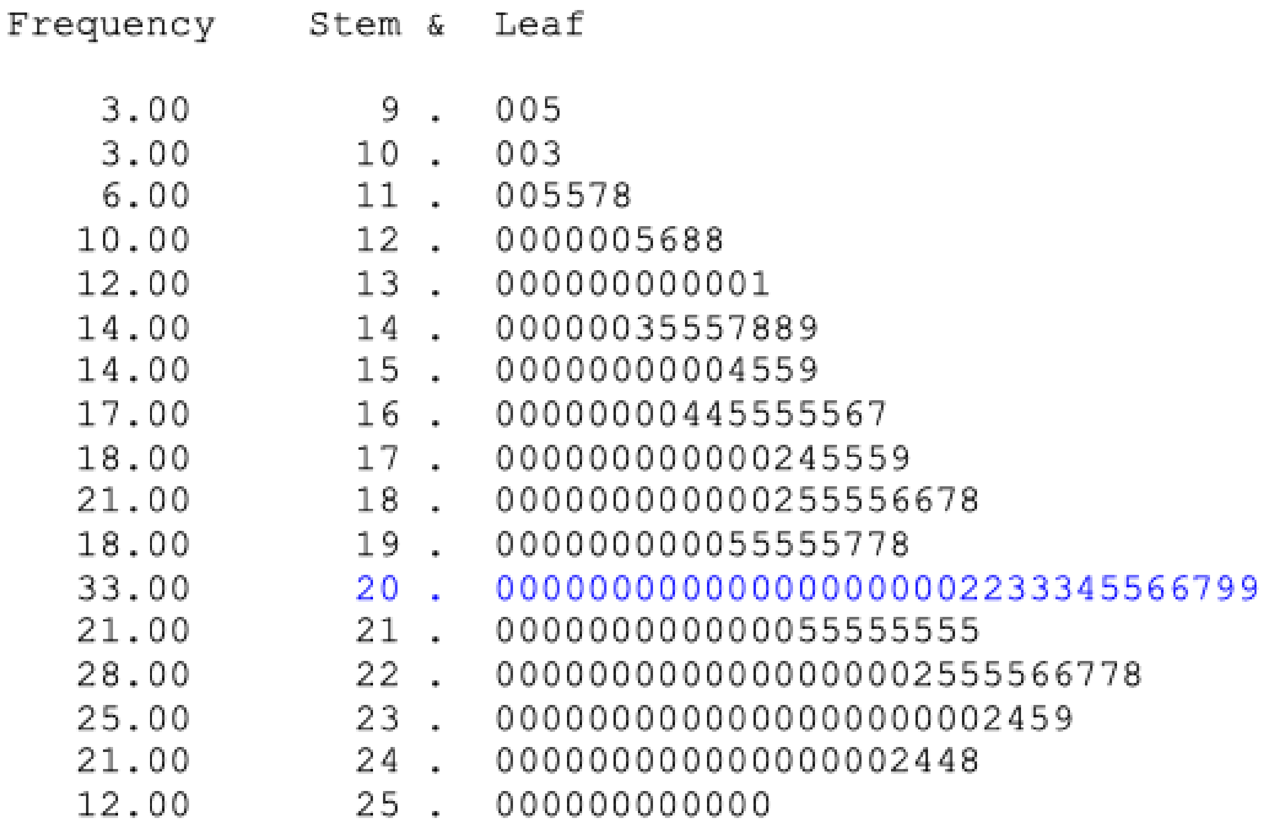
"32" is split into "3" (stem) and "2" (leaf).

15, 16, 21, 23, 23, 26, 26, 30, 32, 41

Stem	Leaf
1	5 6
2	1 3 3 6 6
3	0 2
4	1

vii) Stem-leaf plot

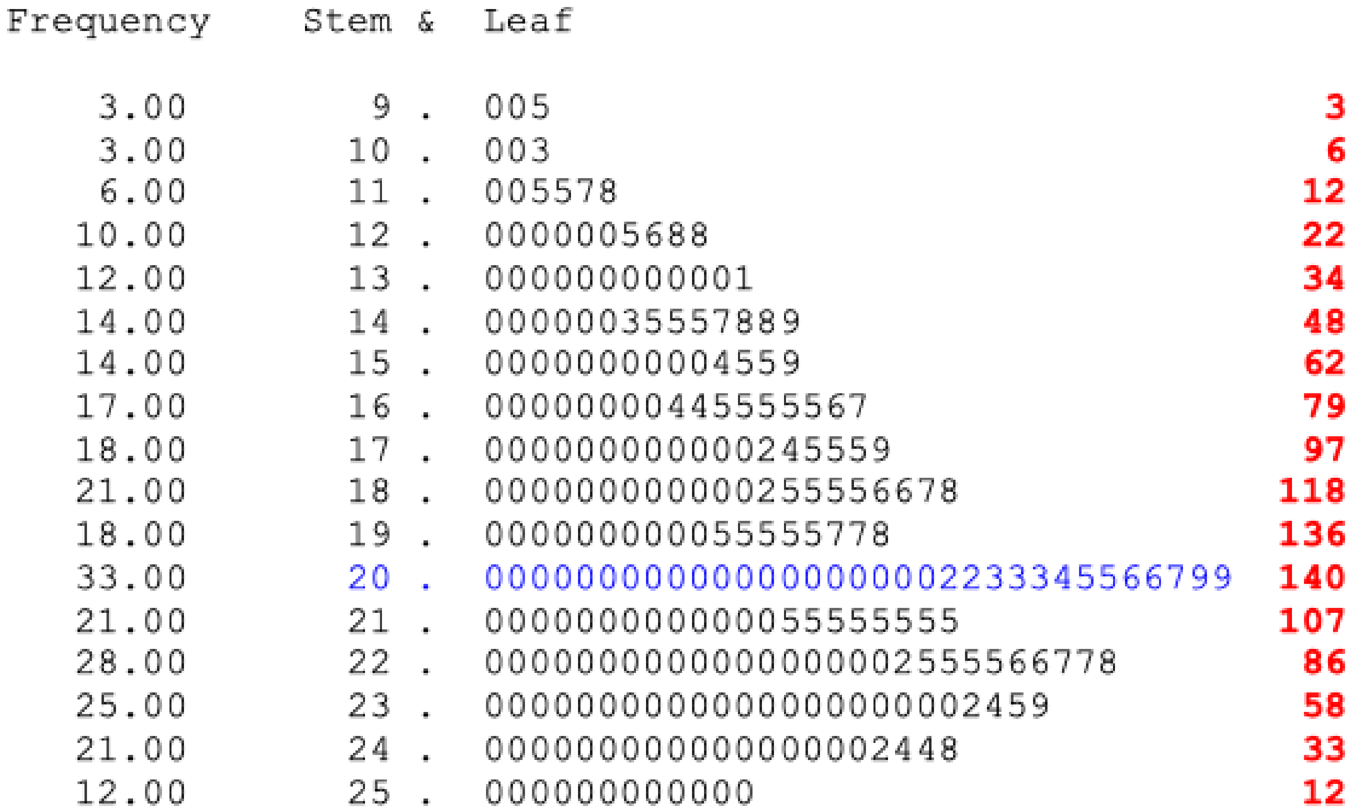
- Example: given weight of children in Uong Bi hospital, create weight Stem-and-Leaf Plot.



```
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 .  000000000000245559
   21.00       18 .  000000000000255556678
   18.00       19 .  000000000055555778
   33.00       20 .  000000000000000000002233345566799
   21.00       21 .  000000000000055555555
   28.00       22 .  0000000000000000002555566778
   25.00       23 .  0000000000000000000002459
   21.00       24 .  000000000000000002448
   12.00       25 .  000000000000
```

vii) Stem-leaf plot

- Example: given weight of children in Uong Bi hospital, create weight Stem-and-Leaf Plot.



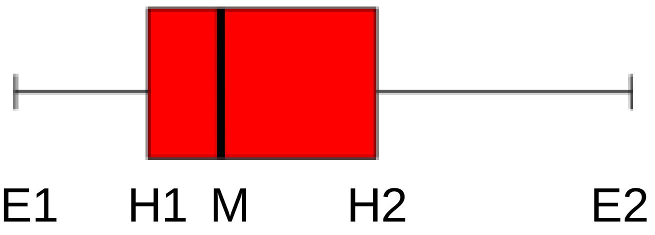
```
Frequency     Stem &  Leaf

     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 .  000000000000245559                   97
    21.00       18 .  000000000000255556678               118
    18.00       19 .  000000000055555778                  136
    33.00       20 .  000000000000000000002233345566799   140
    21.00       21 .  000000000000055555555               107
    28.00       22 .  0000000000000000000255566778         86
    25.00       23 .  0000000000000000000002459            58
    21.00       24 .  000000000000000002448                33
    12.00       25 .  000000000000                         12
```

viii) Box plot

- To quickly summarize large quantities of data, we usually use summary statistics and easy-to-draw graphs.
- Two tools that accomplish this are five-number summaries and box plots
- Five-number summary are:
  - First quartile
  - Median
  - Third quartile
  - Smallest value
  - Largest value

viii) Box plot

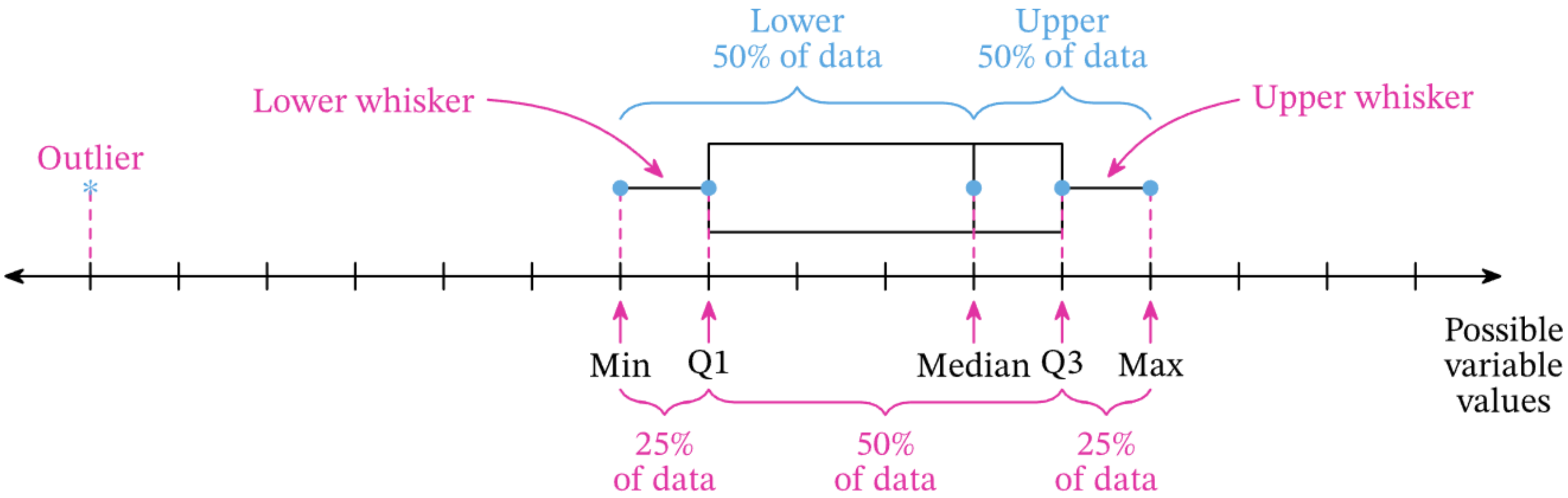
- A box plot is a graphical summary of data that is based on a five-number summary.
- A key to the development of a box plot is the computation of the median and the quartiles Q_1 and Q_3



- Box plot is defined by 5 characteristic values of data:
  - Median M
  - Quartiles H1 (25%) and H2 (75%)
  - Smallest value (E1) and largest value (E2)

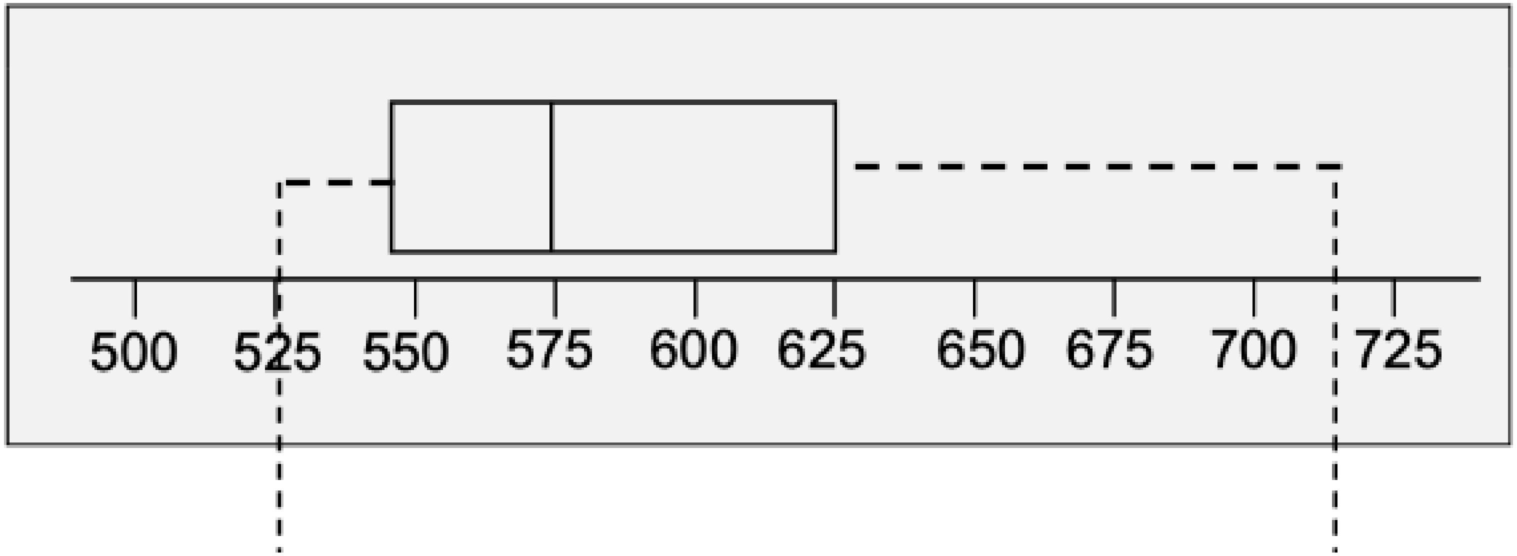
viii) Box plot

- Limits are located (not drawn) using the interquartile range (IQR) multiplied by 1.5
- Data outside these limits are considered outliers
- The locations of each outlier is shown with the symbol *



viii) Box plot

- Whiskers (dashed lines) are drawn from the ends of the box to the smallest and largest data values inside the limits.

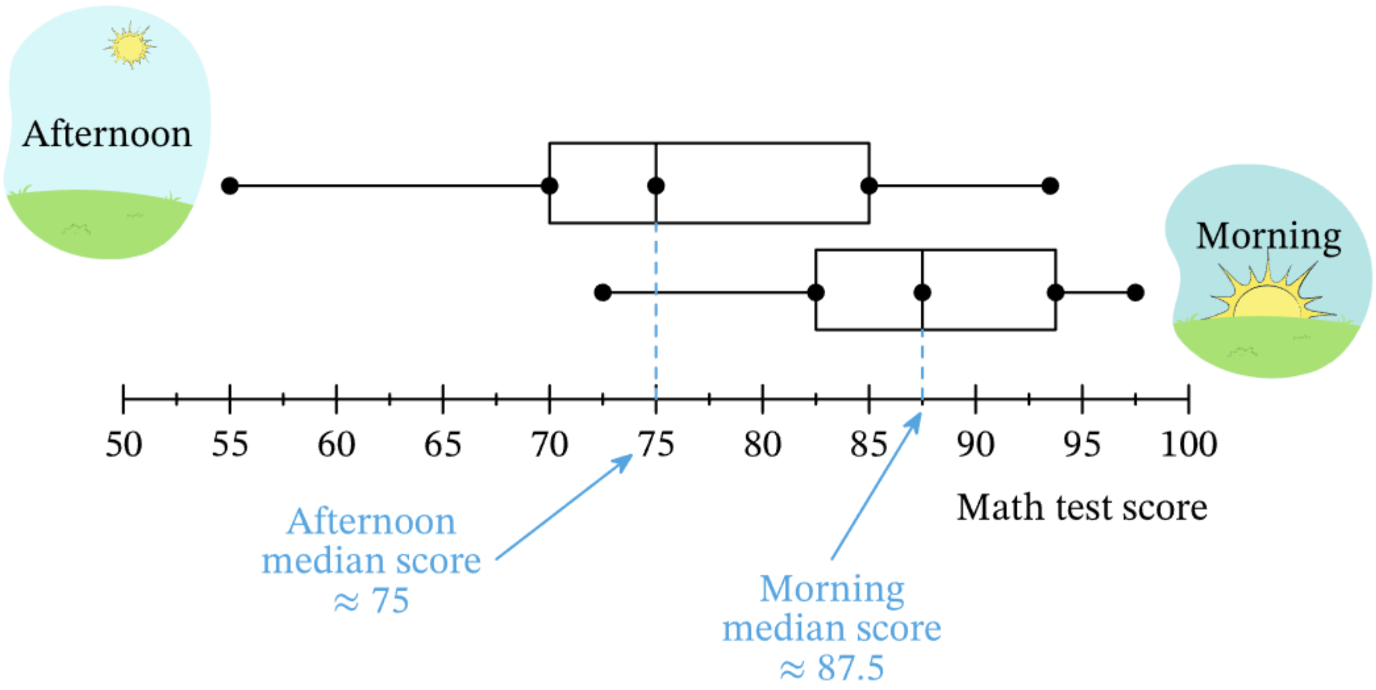


Smallest value
inside limits = 525

largest value
inside limits = 715

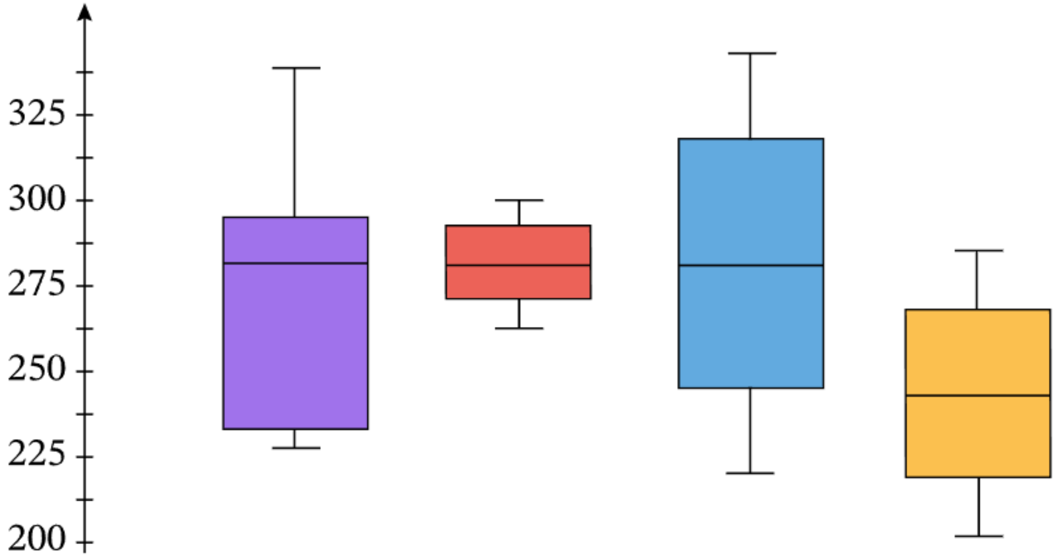
viii) Box plot

- 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



viii) Box plot

- Compare populations:



Using Excel to Describe Quantitative Variables

- MIN, MAX used to compute the extreme values.
- AVERAGE, MODE , MEDIAN used to compute the Mean, Mode and median parameters.
- VARP, VAR, STDEVP, STDEV used for the Variance and Standard Deviation parameters.
- PERCENTILE, QUARTILE used for the Percentile and Quartile parameters.
- HISTOGRAM used to draw the Histograms.

Simple data description methods

- (d) Describe relation between 2 quantitative variables
  - For primary describing relation between 2 quantitative variables we can use:
    - Scatter plot
    - Covariance
    - Linear Correlation Coefficient

Simple data description methods

- (d) Describe relation between 2 quantitative variables
  - For primary describing relation between 2 quantitative variables we can use:
    - Scatter plot and trend-line
    - Covariance
    - Linear Correlation Coefficient

Scatter plot

- A scatter diagram is a graphical presentation of the relationship between two quantitative variables.
- The general pattern of the plotted points suggests the overall relationship between the variables.
- One variable is shown on the horizontal axis and the other variable is shown on the vertical axis.
- A trend-line provides an approximation of the relationship

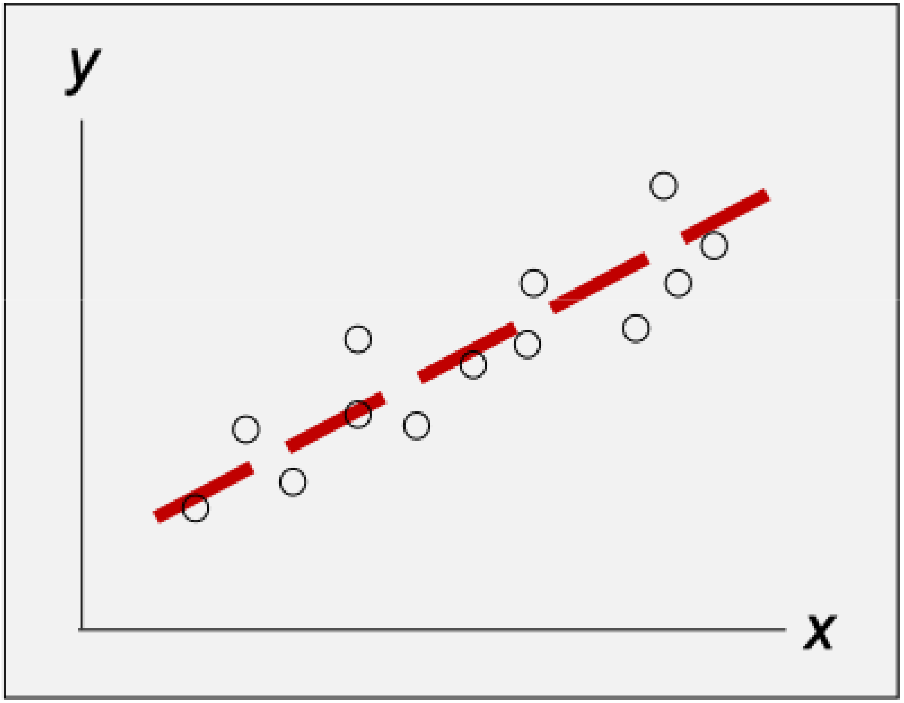
Scatter plot

- Scatter plot providing two-dimensional picture of data represents distribution of data. In that plot we can see concentration area of data, see if there are some outliers, abnormal points, etc.

- Scatter plot can be used to compare several populations: draw several samples (differently colored) on a common plot

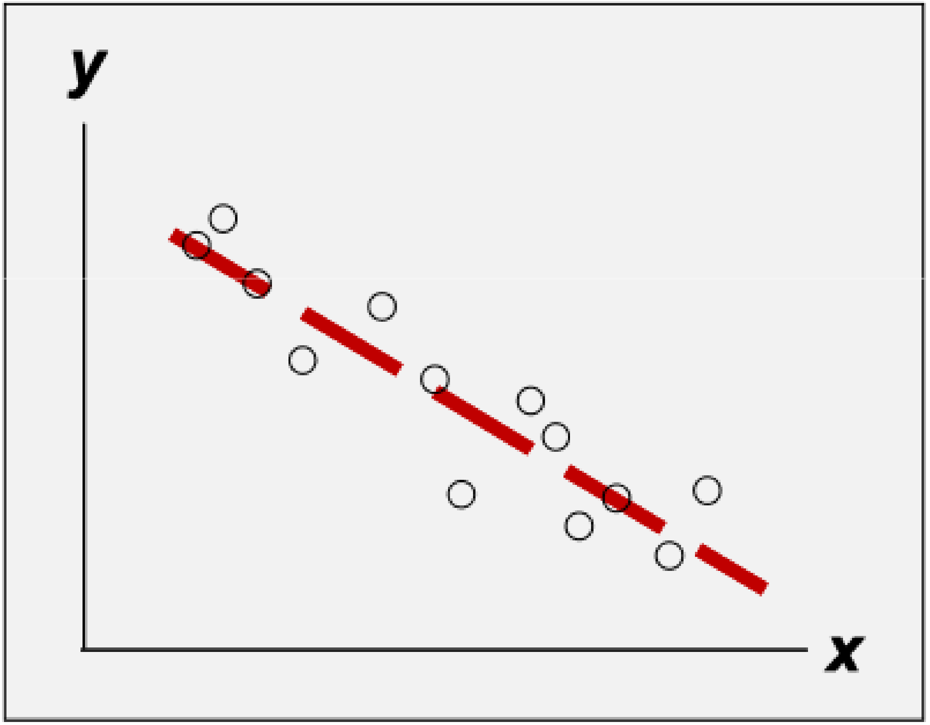
Scatter plot

- A scatter diagram:
  - A positive relationship



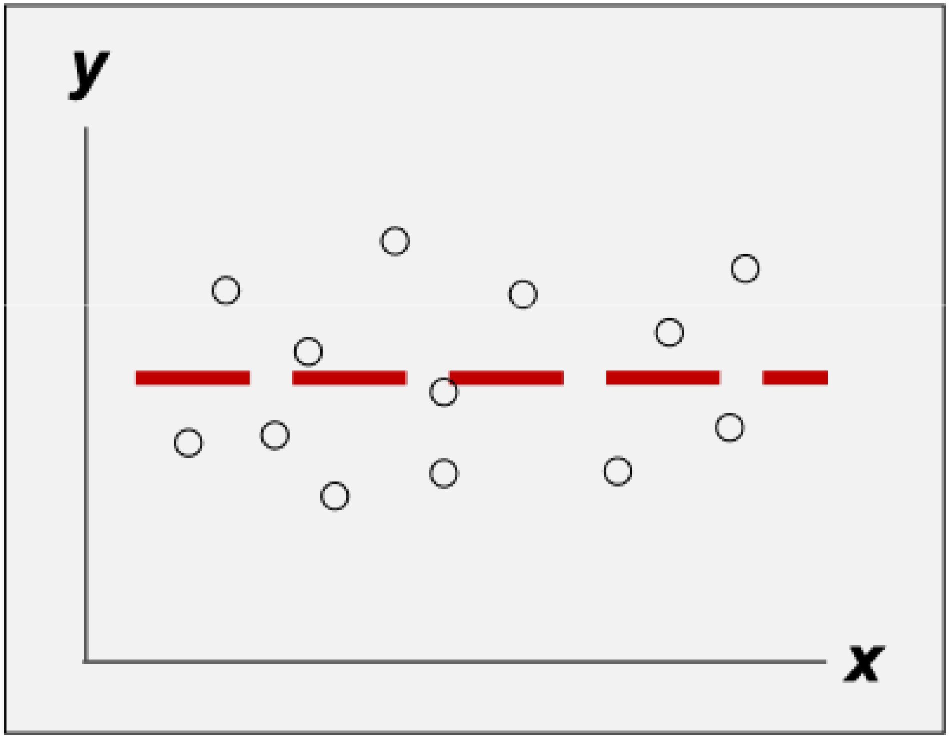
Scatter plot

- A scatter diagram:
  - A negative relationship



Scatter plot

- A scatter diagram:
  - A No Apparent relationship



Measures of Association Between Two Variables

- Often a manager or decision maker is interested in the relationship between two variables.

- Two descriptive measures of the relationship between two variables are covariance and correlation coefficient.

Covariance

- For presentation of relationship between two quantitative variables, we can use covariance between those variables:

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{k=1}^{n}(x_k - Mean(X)).(y_k - Mean(Y))$$

- Covariance is a measure of the linear association between two variables.
  - Positive values indicate a positive relationship, or positively correlated
  - Negative values indicate a negative relationship, or negatively correlated
  - Zero values indicate no relationship, or uncorrelated

Covariance

- Property of covariance:
  - Symmetric: Cov(X,Y) = Cov(Y,X)
  - Depends on measure scale of X and Y:

    Cov(a.X,b.Y) = a.b.Cov(Y,X), where a, b are constant

For samples: $$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

For populations: $$\sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

Linear correlation coefficient

- Linear correlation coefficient measures the linear dependence between two variables :

$$r(X,Y) \; = \; Cov(X,Y) \,/\, (\sigma\,(\mathbf{X}).\sigma(\mathbf{Y}))$$

$$-1 \leq r(X,Y) \leq 1$$

- (a) if X and Y are completely linearly dependent:

  r(X,Y) = 1 if and only if Y = aX + b with a > 0,

  r(X,Y) = -1 if and only if Y = aX + b with a < 0.

- (b) if r(X,Y) close to 1 (or -1) then X and Y are very strongly related, can have some linear correlation
- (c) if r(X,Y) close to 0 then X and Y are linearly independent, there is not linear relation between them

Linear correlation coefficient

- The correlation coefficient can take on values between -1 and +1
- Values near -1 indicate a strong negative linear relationship
- Values near +1 indicate a strong positive linear relationship
- The closer the correlation is to zero, the weaker the relationship

Linear correlation coefficient

- Property of linear correlation coefficient:
  - ❑ Symmetric: r(X,Y) = r(Y,X)
  - ❑ Not depend on measure scale of variables: For all numbers a,b different from 0 and a.b > 0, we have

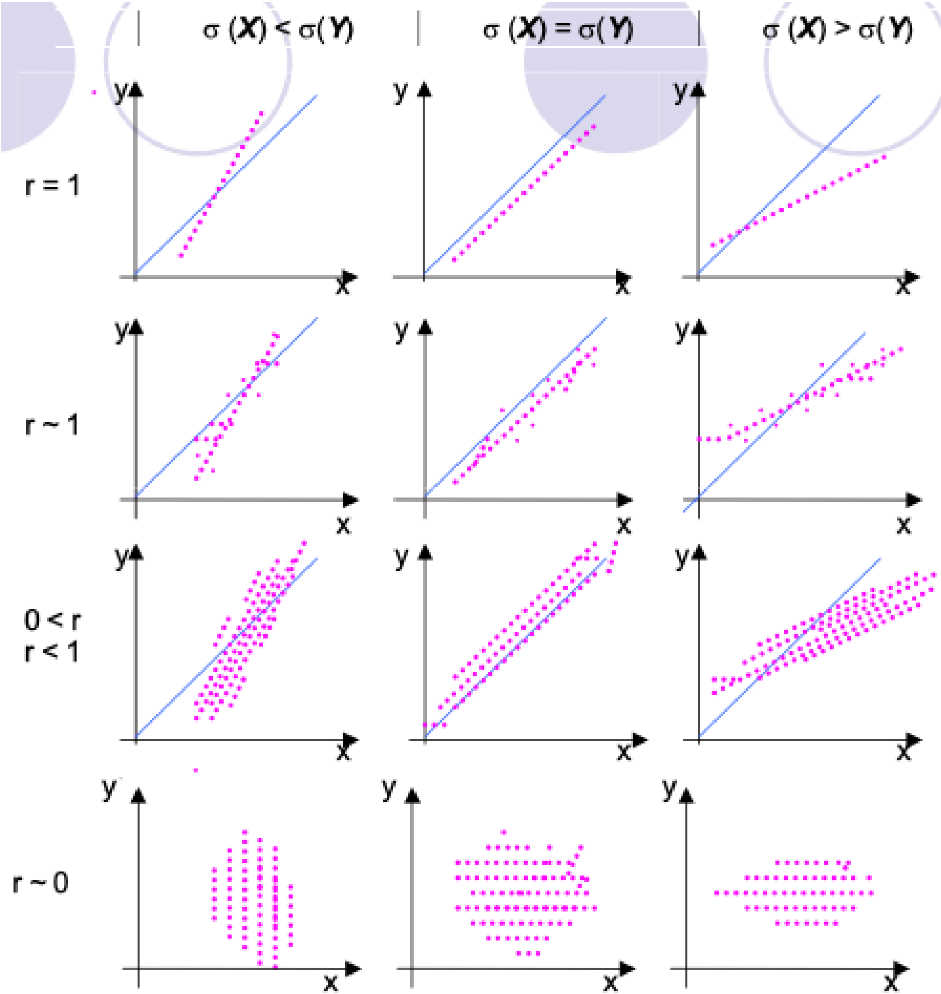
    r(aX,bY) = r(X,Y)

For samples: $r_{xy} = \frac{s_{xy}}{s_x s_y}$

For populations: $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

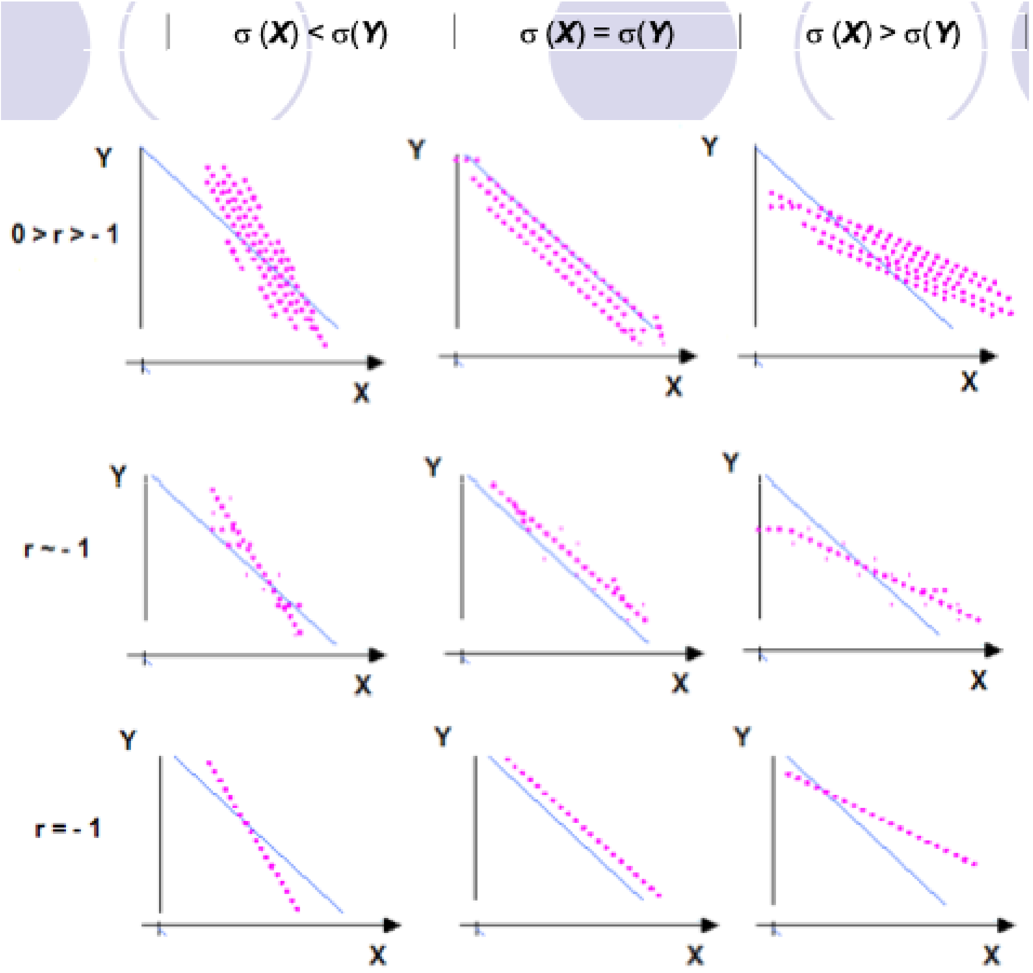
Linear correlation coefficient

- Example:



Linear correlation coefficient

- Example:



Using Excel to Describe Quantitative Variables

- Excel has functions for computing the parameters of quantitative variables and draw graphs:
  - MIN, MAX, MED, MODE, QUARTILE, PERCENTILE used to compute the Minimum, Maximum, Median, Mode, Quartiles, Percentiles.
  - AVERAGE, AVERAGEIF, AVERAGEIFS used to compute the Mean Values.
  - VAR, VARP, STDEV, STDEVP used to compute the Sample Variance, Population Variance, Sample Standard Deviation, Population Standard Deviation.
  - COVAR, CORREL used to compute the Covariance, Correlation Coefficient.
  - HISTOGRAM, SCATTER used to draw the Histogram and Scatter Plot