Applied Biostatistics Lecture

Research Diploma in Cancer Epidemiology

(2nd session)

Lecturer: **Ali Rafei, MSc.**
Research assistant
Cancer Research Center, Cancer Institute of I.R. IRAN
Tehran University of Medical Sciences

Email address: a-rafei@razi.tums.ir
Website: http://rafeistat.ir/

Cancer Research Center
Cancer Institute of I.R. Iran

Descriptive Statistics:

Summarizing data by

- tables
- graphs
- indices

Frequency tables:

For categorical variables:

Stage of Female Breast Cancer Patients					
		Frequency (f_i)	Relative Frequency (r_i)	Cumulative Frequency (F_i)	Cumulative Frequency (R_i)
Valid	IA	30	9.2	30	9.8
	IB	2	.6	32	10.5
	IIA	74	22.7	106	34.8
	IIB	75	23.0	181	59.3
	IIIA	62	19.0	243	79.7
	IIIB	31	9.5	274	89.8
	IIIC	31	9.5	305	100.0
	Total	305	93.6		
Missing	System	21	6.4		
Total		326	100.0		

Descriptive statistics
Statistical inference
ANOVA and modeling
Categorical data analysis
Introduction
Frequency tables
Statistical charts
Statistical indices

Types of frequencies:

- Raw frequency (f_i)
- Relative frequency (r_i)
- Cumulative frequency (F_i)
- Relative cumulative frequency (R_i)

Example:

A physician met 20 cancer patients with following blood groups:

B A O A A A O O A A B B AB O AB AB O O

Blood groups	x_i	f_i	r_i	F_i	R_i
A	1	6	0.30	6	0.30
B	2	3	0.15	9	0.45
AB	3	4	0.20	13	0.65
O	4	7	0.35	20	1.00
Total		20	1.00		

Frequency tables:

For continuous variables:

BMI categories of female breast cancer patients

BMI categories		Frequency (f_i)	Relative Frequency (r_i)	Cumulative Frequency (F_i)	Cumulative Frequency (R_i)
Valid	[0-18.5)	2	.6	2	.7
	[18.5-25)	71	21.8	73	24.7
	[25-30)	100	30.7	173	58.4
	[30-60]	123	37.7	296	100.0
	Total	296	90.8		
Missing	System	30	9.2		
Total		326	100.0		

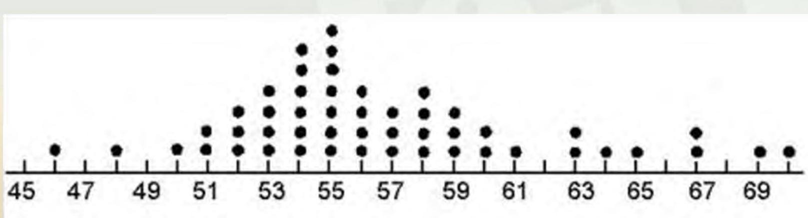
Statistical charts:

For categorical variables:

- Dot diagram
- Bar chart
- Pie chart
- Line chart
- …

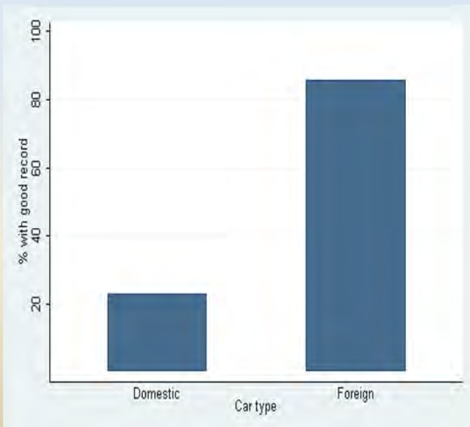
Dot diagram:

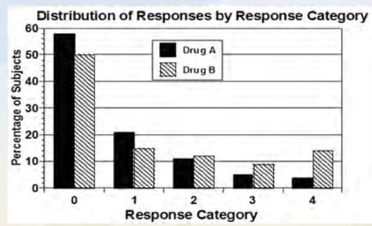
Typically for limited observations:



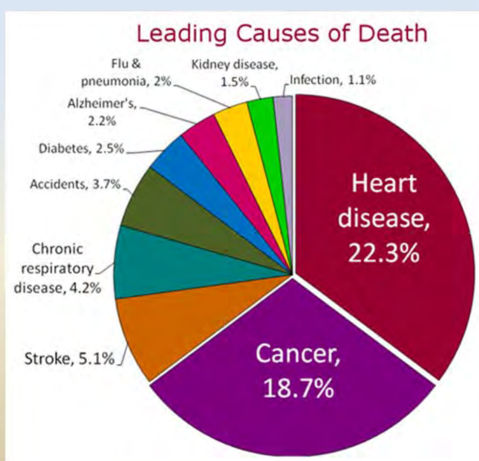
Dot chart of sample age among female breast cancer patients

Bar chart:

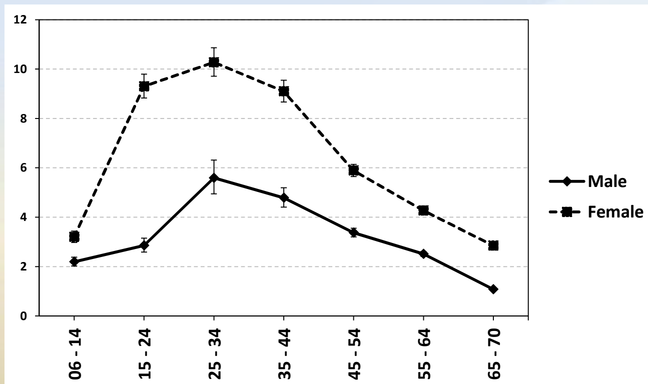




Pie chart:



Line chart:



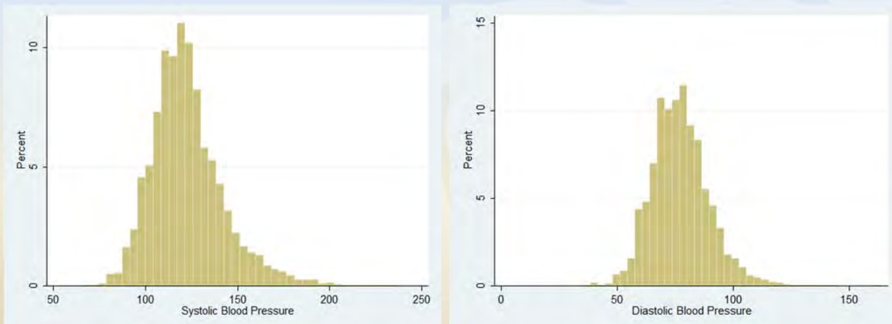
Percentage who brush twice or more a day and floss at least once a day over age groups by sex

Statistical charts:

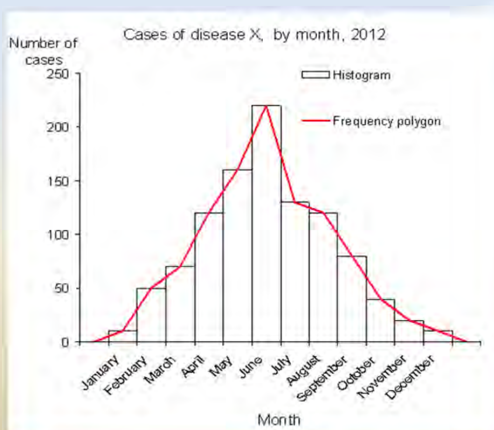
For continuous variables:

- Histogram
- Polygon
- Cumulative polygon
- Box plot
- Scatterplot
- Stem and Leaf

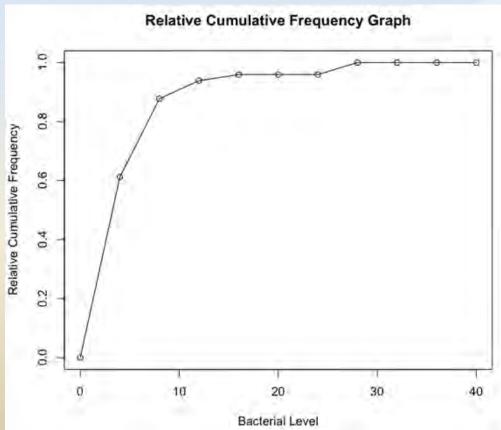
Histogram:



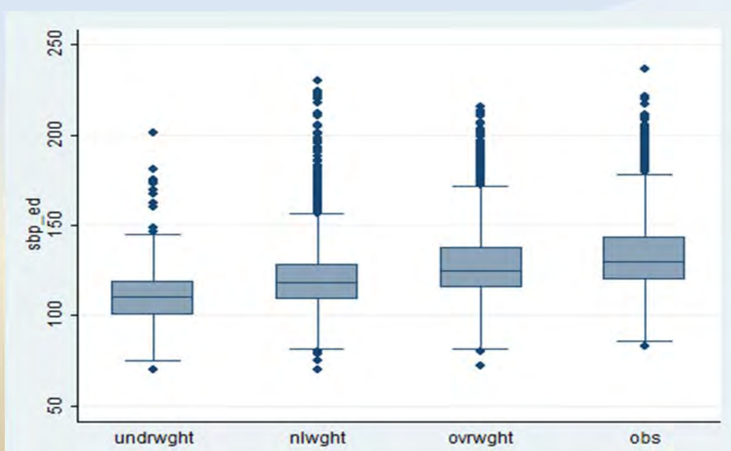
Polygon:



Cumulative polygon:



Box chart:

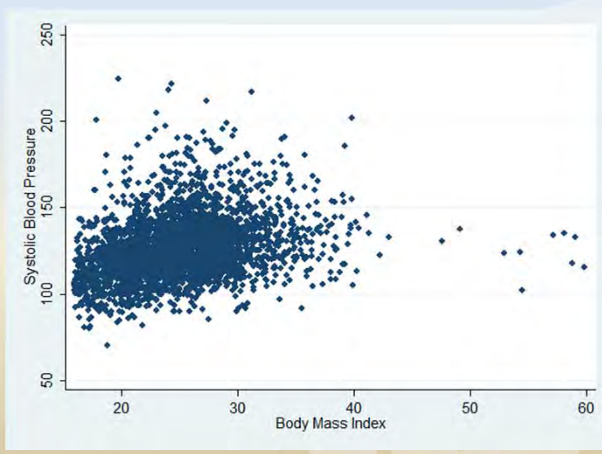


Stem and Leaf diagram:

stem	leaf
0	1, 1, 2, 2, 3, 4, 4, 4, 4, 5, 8
1	0, 0, 0, 1, 1, 3, 7, 9
2	5, 5, 7, 7, 8, 8, 9, 9
3	0, 1, 1, 1, 2, 2, 2, 4, 5
4	0, 4, 8, 9
5	2, 6, 7, 7, 8
6	3, 6

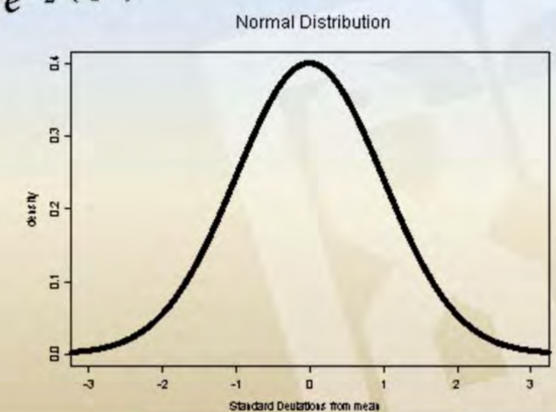
Key: 6|3 = 63 years old

Scatter plot:

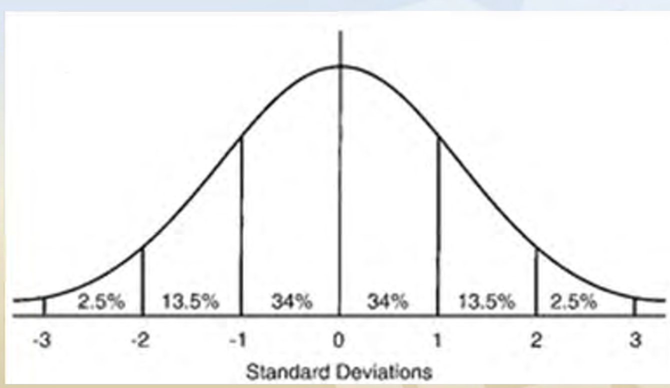


Normal distribution:

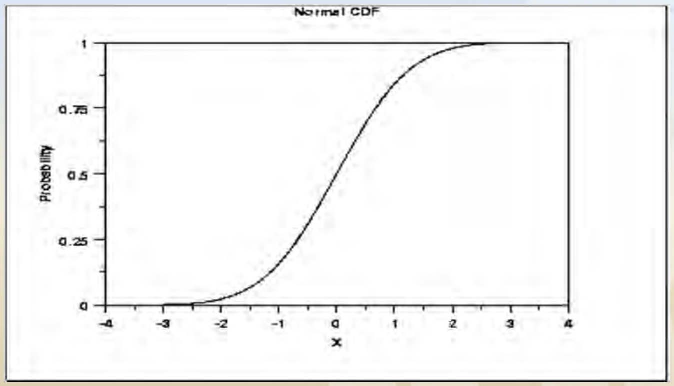
$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Normal distribution:



Normal distribution:



Statistical indices:

- Central tendency
- Variation
- Skewness
- Kurtosis

Central tendency indices:

- Mean
- Median
- Quantiles
- Mode

Mean:

- Arithmetic mean ($\bar{x}$)
- Geometric mean (G)
- Harmonic mean (H)
- Root square mean (M)
- Trimmed mean($\overline{T_k}$)
- ...

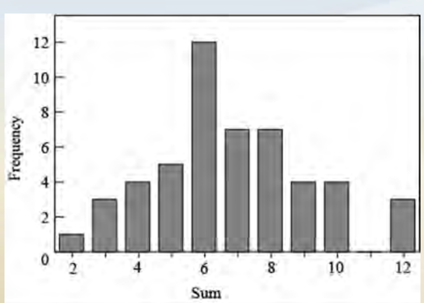
Median:

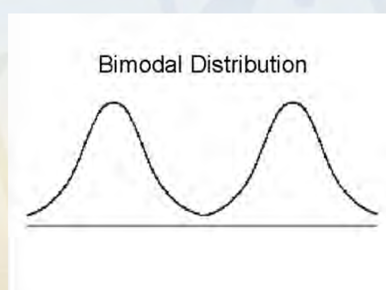
- For raw data
- For classified data

$$Median = L_1 + \left(\frac{\frac{N+1}{2} - F_{j-1}}{N_j} \right) \times h$$

Mode:

Most frequent value in a data set





Median:

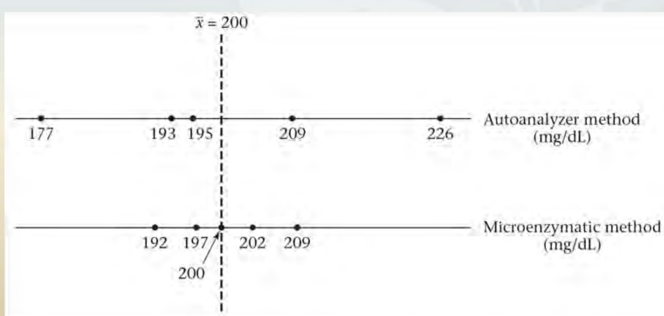
- For raw data

- For classified data

$$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

Measures of spread:

Spread, **dispersion** and **variation** all refer to a measure of the way a data set is distributed around a central value.



Measures of spread:

- Range
- Quantiles and Interquartile range
- Variance
- Standard deviation
- Coefficient of variation

Descriptive statistics
Statistical inference
ANOVA and modeling
Categorical data analysis
Introduction
Frequency tables
Statistical charts
Statistical indices

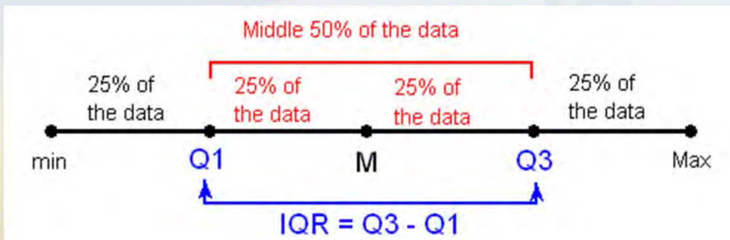
Range:

- The simplest measure of dispersion
- Calculated by finding the difference between the greatest and the least values of the data
- Useful since it is the easiest to understand
- Affected by extreme data
- The range of values 1, 2, 4, 6, 9, 11, 15, 25 is $25 - 1 = 24$

TUMS Research Diploma July 1th, 2015 28 / 32

Interquartile range:

- Quartiles divide a set of ordered data into four groups with equal numbers of values



- The interquartile range is Q1 – Q3, which is the range of the middle of the data

Interquartile range:

- omits the upper and lower 25% of scores
- eliminates the effect of extreme scores
- trimmed samples
- loss of information
- groups with equal numbers of values

Data Set I:
8, 8, 9, 10, 11, 11, 12, 12, 12, 12, 12, 13, 13, 13, 14,14, 15, 15, 16, 17

Range = 9 Interquartile Range = 3

Data Set II:
1, 2, 3, 10, 11, 11, 12, 12, 12, 12, 12, 13, 13, 13, 14,14, 15, 21, 25, 30

Range = 29 Interquartile Range = 3

Mean deviation

Population Deviation

$$= x - \mu$$

Sample Deviation

$$= x - \bar{x}$$

- Larger the deviation = greater the spread in data (the further the data is from the centre)

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Variance

Population variance

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

Sample variance

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

- how the elements are spread around the mean
- large variance means the data is widely spread around the mean

Standard Deviation

Population variance

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Sample variance

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

- The standard deviation is in the **same** **'scale'** as the mean is. The makes these two indicators 'comparable'.

Coefficient of variation

$$CV = 100 \times S/\bar{x}$$

- As $\bar{x}$ increases, S tends to be increased. So CV would be better in some cases than standard deviation.

- CV has **no scale** which makes it comparable among variables with **different** scales.

Coefficient of variation:

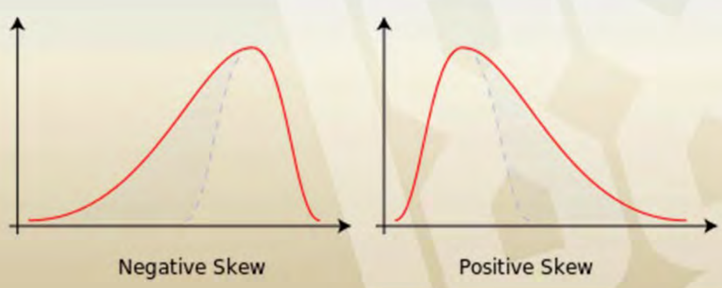
Reproducibility of cardiovascular risk factors in children, Bogalusa Heart Study, 1978–1979

	n	Mean	*sd*	*CV* (%)
Height (cm)	364	142.6	0.31	0.2
Weight (kg)	365	39.5	0.77	1.9
Triceps skin fold (mm)	362	15.2	0.51	3.4
Systolic blood pressure (mm Hg)	337	104.0	4.97	4.8
Diastolic blood pressure (mm Hg)	337	64.0	4.57	7.1
Total cholesterol (mg/dL)	395	160.4	3.44	2.1
HDL cholesterol (mg/dL)	349	56.9	5.89	10.4

Measure of symmetry

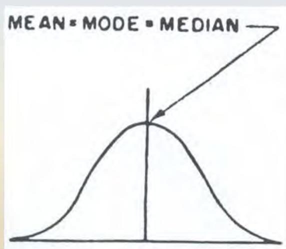
Skewness:

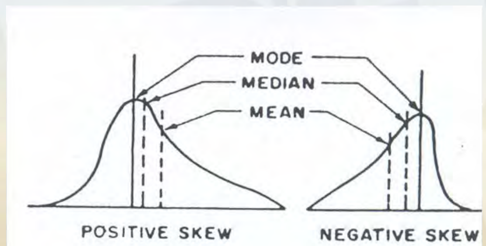
- Symmetric distribution
- Positively skewed distribution
- Negatively skewed distribution



Measure of symmetry

- If Mean > Mode, the skewness is positive.
- If Mean < Mode, the skewness is negative.
- If Mean = Mode, the skewness is zero.





$$\text{Pearson's coefficient of skewness} = \frac{\text{Mean - Mode}}{\text{Standard deviation}}$$

Descriptive statistics
Statistical inference
ANOVA and modeling
Categorical data analysis
Introduction
Frequency tables
Statistical charts
Statistical indices

Measure of symmetry

- Moment skewness coefficient

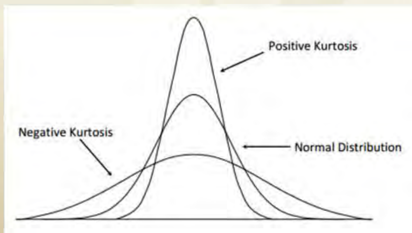
$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^{3/2}},$$

- If $b_1>0$ the distribution is right-skewed
- If $b_1<0$ the distribution is left-skewed

Measure of kurtosis

- Moment kurtosis coefficient

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3$$



Mean, Median, Mode?

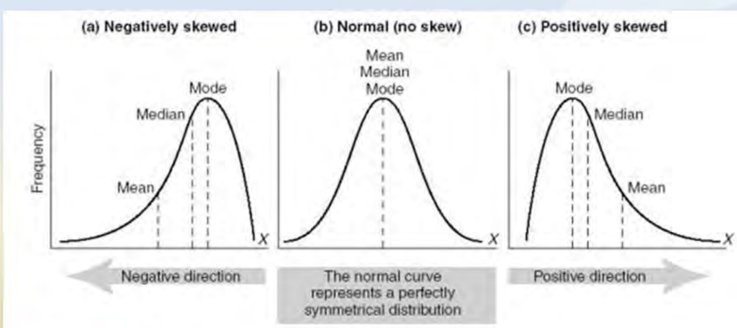


FIGURE 15.6 Examples of normal and skewed distributions

Transformation

if y_i defined as:

$$y_i = ax_i + b \quad ; \quad a \neq 0$$

Then

$$\bar{y} = a\bar{x} + b$$

$$S_y^2 = a^2 S_x^2$$

$$S_y = |a| S_x$$

Standardization

Suppose x_1, x_2, … , x_n be a sample of size n
Define z_i as follows:

$$z_i = \frac{x_i - \bar{x}}{s}$$

Then

$$\bar{z} = 0$$

$$S_z = 1$$

Standardization

Suppose X=SBP; Y=Cholesterol:

i	1	2	3	**4**	5	6	7	8	9
X	116	133	93	**117**	125	115	113	128	137
Y	193	209	114	**228**	107	191	191	189	193

$$\bar{x} = 120\ ;\ \ S_x = 13; \quad \Rightarrow \quad z_{x4} = \frac{117-120}{13} = -0.2$$

$$\bar{y} = 179\ ;\ \ S_y = 41; \quad \Rightarrow \quad z_{y4} = \frac{228-179}{41} = 1.2$$