

# Applied Biostatistics Lecture

Research Diploma in Cancer Epidemiology  
(2<sup>nd</sup> 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](mailto:a-rafei@razi.tums.ir)

Website: <http://rafeistat.ir/>



**Descriptive statistics**  
Statistical inference  
ANOVA and modeling  
Categorical data analysis

**Introduction**  
Frequency tables  
Statistical charts  
Statistical indices

## Descriptive Statistics:

Summarizing data by

- tables
- graphs
- indices

## Frequency tables:

For categorical variables:

Stage of Female Breast Cancer Patients					
		Frequency ( $f_j$ )	Relative Frequency ( $r_j$ )	Cumulative Frequency ( $F_j$ )	Cumulative Frequency ( $R_j$ )
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		

## Types of frequencies:

- Raw frequency ( $f_j$ )
- Relative frequency ( $r_j$ )
- Cumulative frequency ( $F_j$ )
- Relative cumulative frequency ( $R_j$ )

## 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		

<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	Introduction Frequency tables <b>Statistical charts</b> Statistical indices
---	--

**Statistical charts:**

For categorical variables:

- Dot diagram
- Bar chart
- Pie chart
- Line chart
- ...

TUMS Research Diploma      July 1<sup>th</sup>, 2015      7 / 32

<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	Introduction Frequency tables <b>Statistical charts</b> Statistical indices
---	--

**Dot diagram:**

Typically for limited observations:

Age (Year)	Number of Patients
45	0
46	1
47	0
48	1
49	0
50	1
51	2
52	3
53	4
54	5
55	10
56	5
57	4
58	3
59	2
60	1
61	0
62	1
63	2
64	1
65	1
66	2
67	1
68	1
69	1

Dot chart of sample age among female breast cancer patients

TUMS Research Diploma      July 1<sup>th</sup>, 2015      8 / 32

**Descriptive statistics**  
 Statistical inference  
 ANOVA and modeling  
 Categorical data analysis

**Introduction**  
 Frequency tables  
**Statistical charts**  
 Statistical indices

## Bar chart:

Car type	% with good record
Domestic	~25
Foreign	~85

Response Category	Drug A (%)	Drug B (%)
0	~55	~50
1	~20	~15
2	~10	~12
3	~5	~8
4	~3	~10

Region	marriages	divorces
NE	~400	~100
N. Central Canada region	~500	~300
South	~900	~300
West	~500	~300

TUMS Research Diploma
July 1<sup>th</sup>, 2015
9 / 32

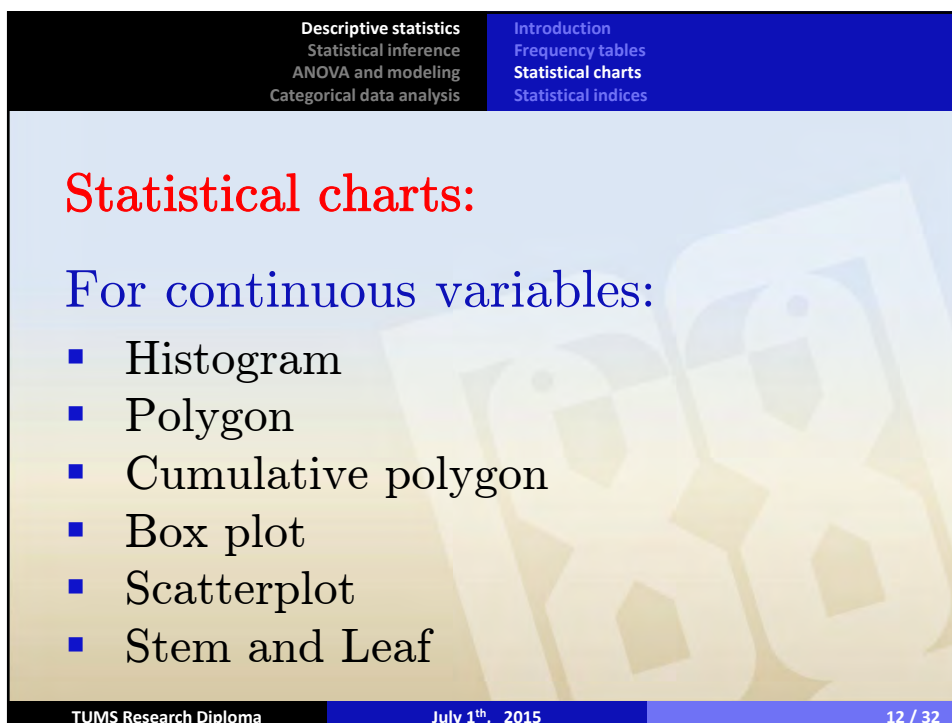
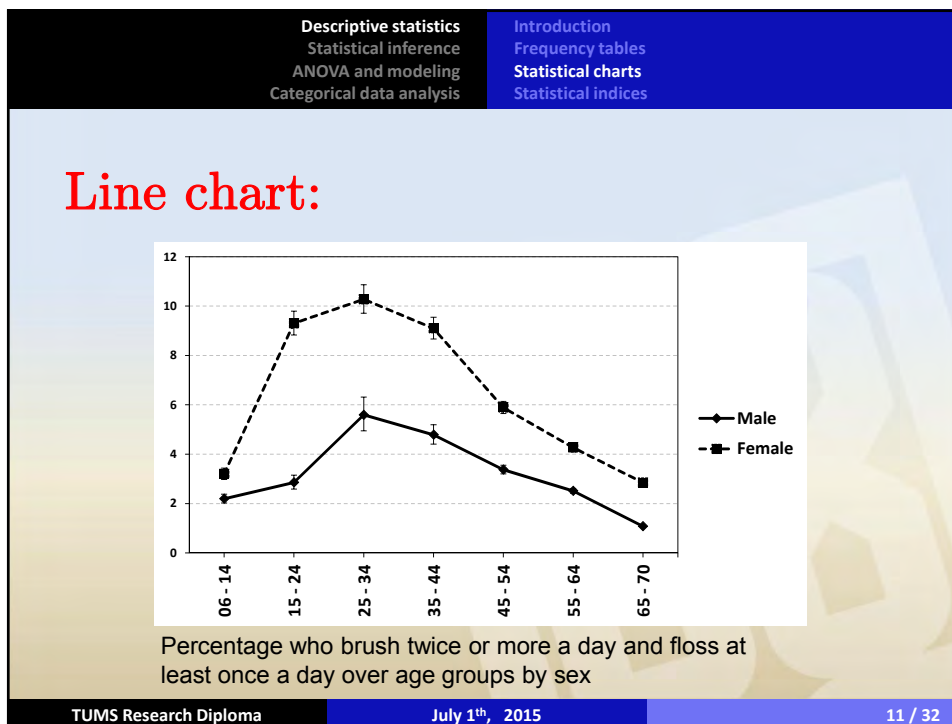
**Descriptive statistics**  
 Statistical inference  
 ANOVA and modeling  
 Categorical data analysis

**Introduction**  
 Frequency tables  
**Statistical charts**  
 Statistical indices

## Pie chart:

Cause of Death	Percentage
Heart disease	22.3%
Cancer	18.7%
Stroke	5.1%
Chronic respiratory disease	4.2%
Accidents	3.7%
Diabetes	2.5%
Alzheimer's	2.2%
Flu & pneumonia	2%
Kidney disease	1.5%
Infection	1.1%

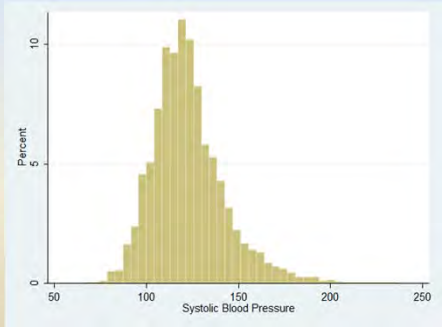
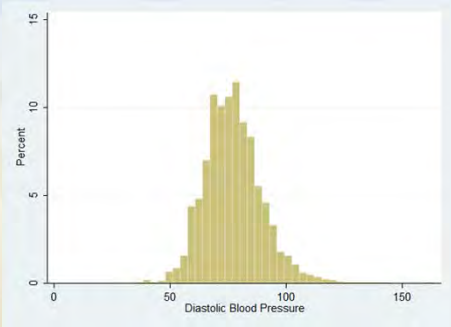
TUMS Research Diploma
July 1<sup>th</sup>, 2015
10 / 32



**Descriptive statistics**  
 Statistical inference  
 ANOVA and modeling  
 Categorical data analysis
 

**Introduction**  
 Frequency tables  
**Statistical charts**  
 Statistical indices

## Histogram:

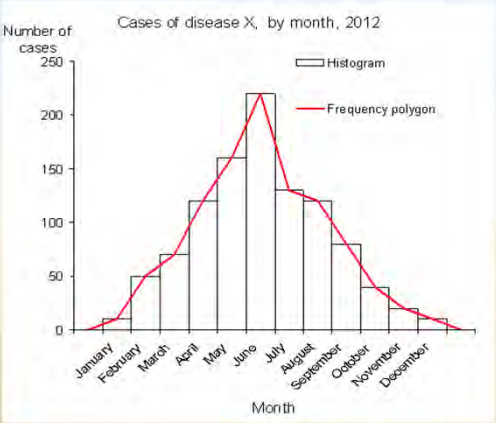
TUMS Research Diploma
July 1<sup>th</sup>, 2015
13 / 32

**Descriptive statistics**  
 Statistical inference  
 ANOVA and modeling  
 Categorical data analysis
 

**Introduction**  
 Frequency tables  
**Statistical charts**  
 Statistical indices

## Polygon:

Cases of disease X, by month, 2012



TUMS Research Diploma
July 1<sup>th</sup>, 2015
14 / 32

[Introduction](#)  
[Frequency tables](#)  
**[Statistical charts](#)**  
[Statistical indices](#)

**Descriptive statistics**  
[Statistical inference](#)  
[ANOVA and modeling](#)  
[Categorical data analysis](#)

## Cumulative polygon:

**Relative Cumulative Frequency Graph**

Bacterial Level	Relative Cumulative Frequency
0	0.00
2	0.60
5	0.85
10	0.92
15	0.95
20	0.96
25	0.97
30	0.98
35	0.99
40	1.00

TUMS Research Diploma
July 1<sup>th</sup>, 2015
15 / 32

[Introduction](#)  
[Frequency tables](#)  
**[Statistical charts](#)**  
[Statistical indices](#)

**Descriptive statistics**  
[Statistical inference](#)  
[ANOVA and modeling](#)  
[Categorical data analysis](#)

## Box chart:

**Box chart of sbp\_ed**

Category	Min	Q1	Median	Q3	Max
undrwght	75	100	110	120	150
nrlwght	80	110	120	130	160
ovrwght	80	115	125	140	175
obs	85	120	130	145	185

TUMS Research Diploma
July 1<sup>th</sup>, 2015
16 / 32



**Descriptive statistics**  
 Statistical inference  
 ANOVA and modeling  
 Categorical data analysis

Introduction  
 Frequency tables  
**Statistical charts**  
 Statistical indices

## 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

TUMS Research Diploma
July 1<sup>th</sup>, 2015
17 / 32

**Descriptive statistics**  
 Statistical inference  
 ANOVA and modeling  
 Categorical data analysis

Introduction  
 Frequency tables  
**Statistical charts**  
 Statistical indices

## Scatter plot:

TUMS Research Diploma
July 1<sup>th</sup>, 2015
18 / 32

<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	Introduction Frequency tables <b>Statistical charts</b> Statistical indices
---	--

## Normal distribution:

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

Normal Distribution

TUMS Research Diploma      July 1<sup>th</sup>, 2015      19 / 32

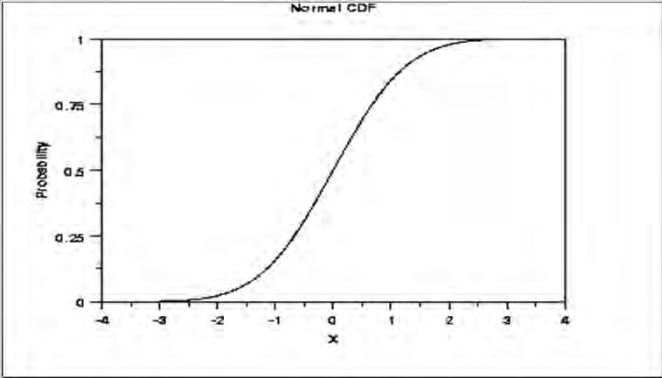
<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	Introduction Frequency tables <b>Statistical charts</b> Statistical indices
---	--

## Normal distribution:

TUMS Research Diploma      July 1<sup>th</sup>, 2015      20 / 32

<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	<b>Introduction</b> Frequency tables <b>Statistical charts</b> Statistical indices
---	---

## Normal distribution:



The graph shows the Normal Cumulative Distribution Function (CDF). The x-axis is labeled 'x' and ranges from -4 to 4 with major ticks every 1 unit. The y-axis is labeled 'Probability' and ranges from 0 to 1 with major ticks at 0, 0.25, 0.5, 0.75, and 1. The curve starts near 0 at x = -4, passes through (0, 0.5), and approaches 1 as x increases towards 4.

TUMS Research Diploma	July 1 <sup>th</sup> , 2015	21 / 32
-----------------------	-----------------------------	---------

<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	<b>Introduction</b> Frequency tables Statistical charts <b>Statistical indices</b>
---	---

## Statistical indices:

- Central tendency
- Variation
- Skewness
- Kurtosis

TUMS Research Diploma	July 1 <sup>th</sup> , 2015	22 / 32
-----------------------	-----------------------------	---------

<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	<b>Introduction</b> Frequency tables Statistical charts <b>Statistical indices</b>
---	---

## Central tendency indices:

- Mean
- Median
- Quantiles
- Mode

TUMS Research Diploma July 1<sup>th</sup>, 2015 23 / 32

<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	<b>Introduction</b> Frequency tables Statistical charts <b>Statistical indices</b>
---	---

## Mean:

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

TUMS Research Diploma July 1<sup>th</sup>, 2015 24 / 32

<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	Introduction Frequency tables Statistical charts <b>Statistical indices</b>
---	--

**Median:**

- For raw data
- For classified data

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

TUMS Research Diploma      July 1<sup>th</sup>, 2015      25 / 32

<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	Introduction Frequency tables Statistical charts <b>Statistical indices</b>
---	--

**Mode:**

Most frequent value in a data set

Sum	Frequency
2	1
3	3
4	4
5	5
6	12
7	7
8	7
9	4
10	4
11	2
12	1

Bimodal Distribution

TUMS Research Diploma      July 1<sup>th</sup>, 2015      26 / 32

<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	Introduction Frequency tables Statistical charts <b>Statistical indices</b>
---	--

## Median:

- For raw data
- For classified data

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

TUMS Research Diploma      July 1<sup>th</sup>, 2015      27 / 32

<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	Introduction Frequency tables Statistical charts <b>Statistical indices</b>
---	--

## Measures of spread:

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

Method	Value (mg/dL)
Autoanalyzer method	177
Autoanalyzer method	193
Autoanalyzer method	195
Autoanalyzer method	209
Autoanalyzer method	226
Microenzymatic method	192
Microenzymatic method	197
Microenzymatic method	202
Microenzymatic method	209
Mean (x̄)	200

TUMS Research Diploma      July 1<sup>th</sup>, 2015      28 / 32

<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	<b>Introduction</b> Frequency tables Statistical charts <b>Statistical indices</b>
---	---

## Measures of spread:

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

TUMS Research Diploma      July 1<sup>th</sup>, 2015      28 / 32

<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	<b>Introduction</b> Frequency tables Statistical charts <b>Statistical indices</b>
---	---

## 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 1<sup>th</sup>, 2015      28 / 32





<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	Introduction Frequency tables Statistical charts <b>Statistical indices</b>
---	--

## 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)

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

TUMS Research Diploma
July 1<sup>th</sup>, 2015
28 / 32

<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	Introduction Frequency tables Statistical charts <b>Statistical indices</b>
---	--

## 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

TUMS Research Diploma
July 1<sup>th</sup>, 2015
28 / 32

<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	Introduction Frequency tables Statistical charts <b>Statistical indices</b>
---	--

## 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. This makes these two indicators 'comparable'.

TUMS Research Diploma      July 1<sup>st</sup>, 2015      28 / 32

<b>Descriptive statistics</b> Statistical inference ANOVA and modeling Categorical data analysis	Introduction Frequency tables Statistical charts <b>Statistical indices</b>
---	--

## Coefficient of variation

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

- 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.

TUMS Research Diploma      July 1<sup>st</sup>, 2015      28 / 32

	<i>n</i>	Mean	<i>sd</i>	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

## Coefficient of variation:

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

	<i>n</i>	Mean	<i>sd</i>	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:

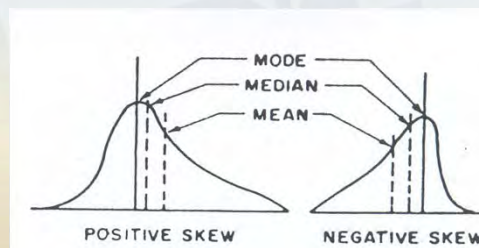
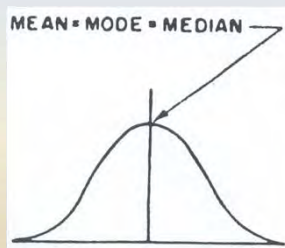
- Symmetric distribution
- Positively skewed distribution
- Negatively skewed distribution

Negative Skew

Positive Skew

## Measure of symmetry

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



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

## Measure of symmetry

- Moment skewness coefficient

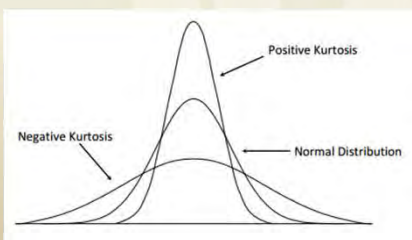
$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[ \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}}$$

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

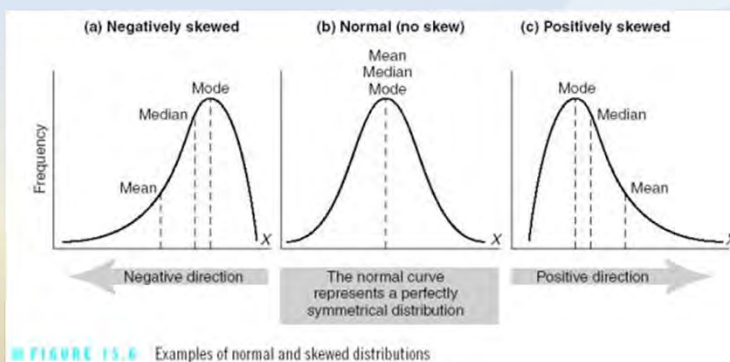
## Measure of kurtosis

- Moment kurtosis coefficient

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



## Mean, Median, Mode?



Descriptive statistics Statistical inference ANOVA and modeling Categorical data analysis	Introduction Frequency tables Statistical charts <b>Statistical indices</b>
--	--

## Transformation

if  $y_i$  defined as:

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

Then

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

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

$$S_y = |a| S_x$$

TUMS Research Diploma	July 1 <sup>th</sup> , 2015	31 / 32
-----------------------	-----------------------------	---------

Descriptive statistics Statistical inference ANOVA and modeling Categorical data analysis	Introduction Frequency tables Statistical charts <b>Statistical indices</b>
--	--

## Standardization

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

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

Then

$$\bar{z} = 0$$

$$S_z = 1$$

TUMS Research Diploma	July 1 <sup>th</sup> , 2015	29 / 32
-----------------------	-----------------------------	---------

	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

$\bar{x} = 120 ; S_x = 13; \Rightarrow z_{x4} = \frac{117 - 120}{13} = -0.2$   
 $\bar{y} = 179 ; S_y = 41; \Rightarrow z_{y4} = \frac{228 - 179}{41} = 1.2$

TUMS Research Diploma July 1<sup>th</sup>, 2015 29 / 32

Thank you for your attention

TUMS Research Diploma July 1<sup>th</sup>, 2015 32 / 32