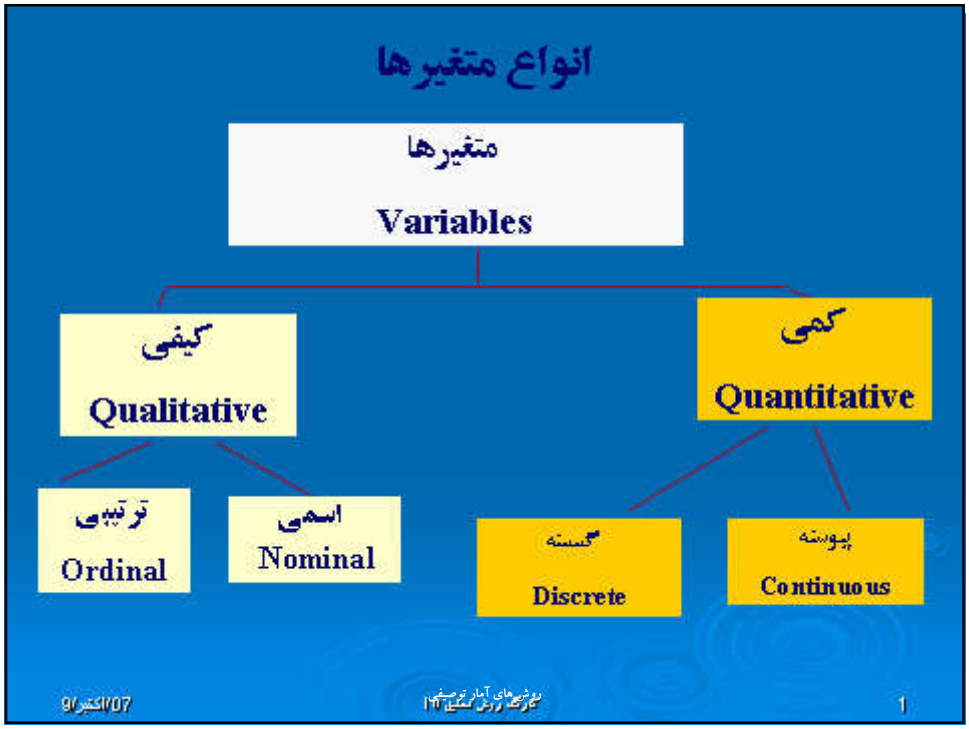
مروری بر

روش های آمار توصیفی

دکتر یدا... محرابی

دانشکده بهداشت- دانشگاه علوم پزشکی شهید بهشتی

ymehrabi@gmail.com



Numerical Summaries

- **Center of the data**
  - mean
  - Median
- **Variation**
  - range
  - quartiles (interquartile range)
  - variance
  - standard deviation

Mean or Average

- Traditional measure of center
- Sum the values and divide by the number of values

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median (*M*)

- A *resistant measure* of the data's center
- At least half of the ***ordered*** values are less than or equal to the median value
- At least half of the ***ordered*** values are greater than or equal to the median value
- If *n* is odd, the median is the middle ordered value
- If *n* is even, the median is the average of the two middle ordered values

Median (*M*)

Location of the median: $L(M) = (n+1)/2$,

where *n* = sample size.

Example: If 25 data values are recorded, the Median would be the

(25+1)/2 = 13th ordered value.

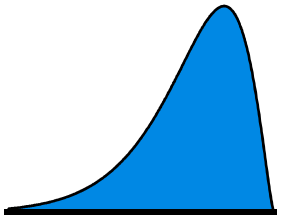
Median

- Example 1 data: 2 4 6
  Median (*M*) = 4
- Example 2 data: 2 4 6 8
  Median = 5 (ave. of 4 and 6)
- Example 3 data: 6 2 4
  Median ≠ 2
  (**order** the values: 2 4 6 , so Median = 4)

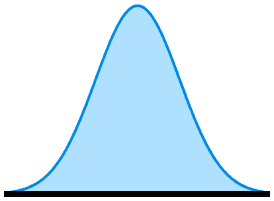
Comparing the Mean & Median

- The mean and median of data from a symmetric distribution should be close together. The actual (true) mean and median of a symmetric distribution are exactly the same.
- In a skewed distribution, the mean is farther out in the long tail than is the median [the mean is 'pulled' in the direction of the possible outlier(s)].

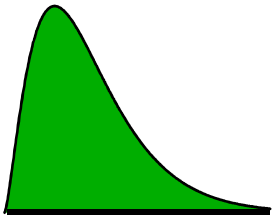
Skewness



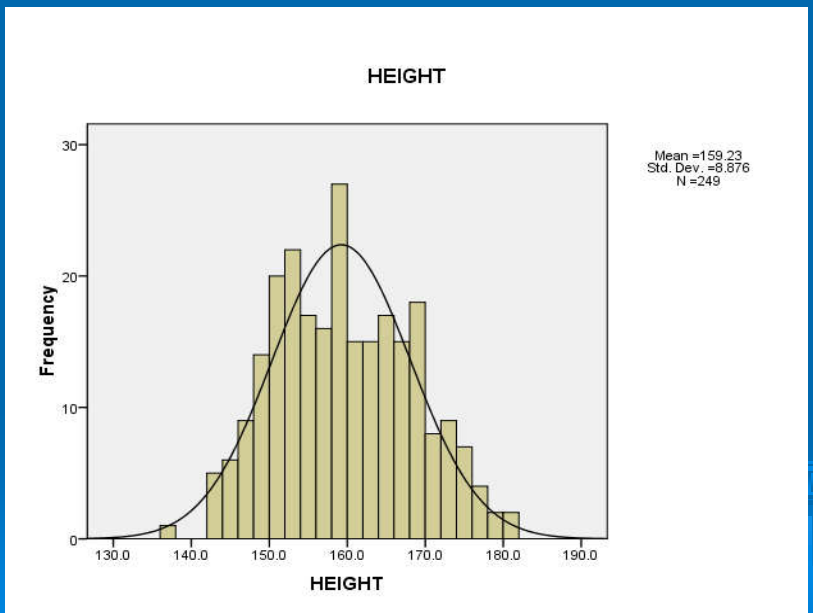
Negatively Skewed

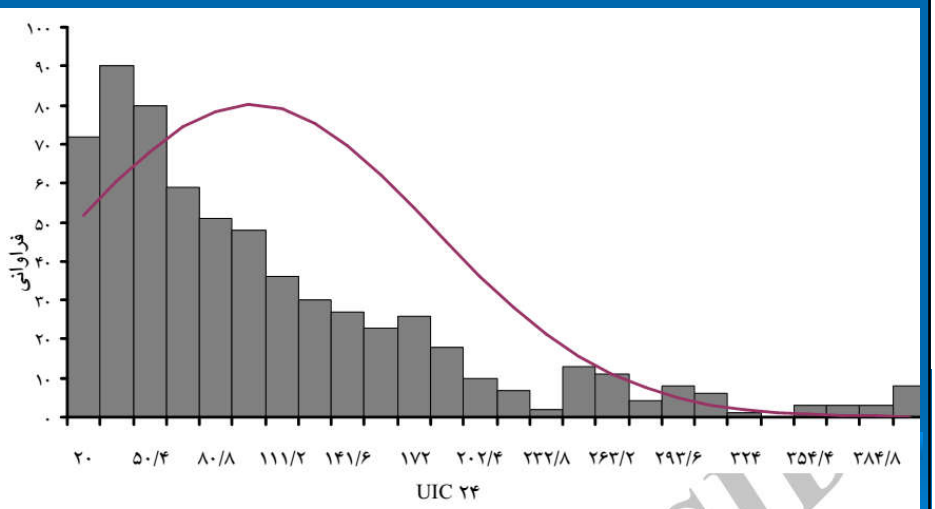


Symmetric (Not Skewed)



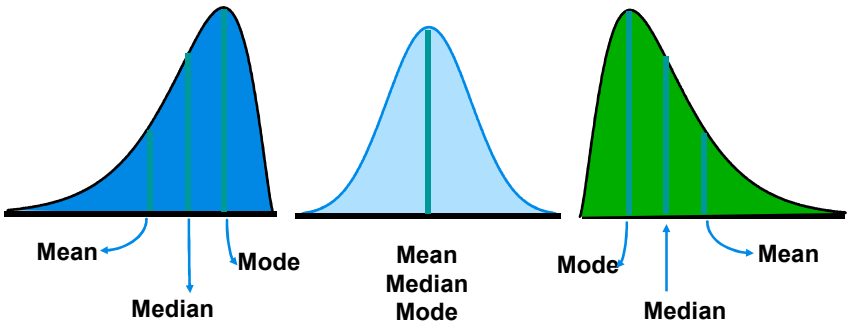
Positively Skewed

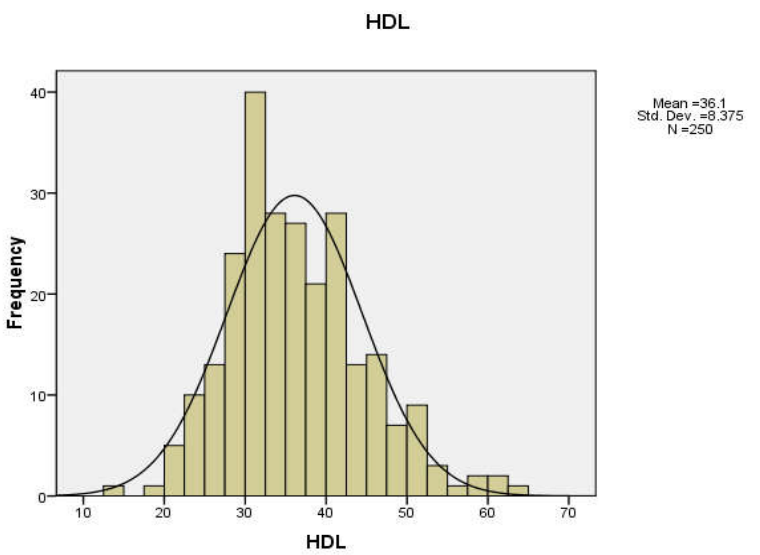


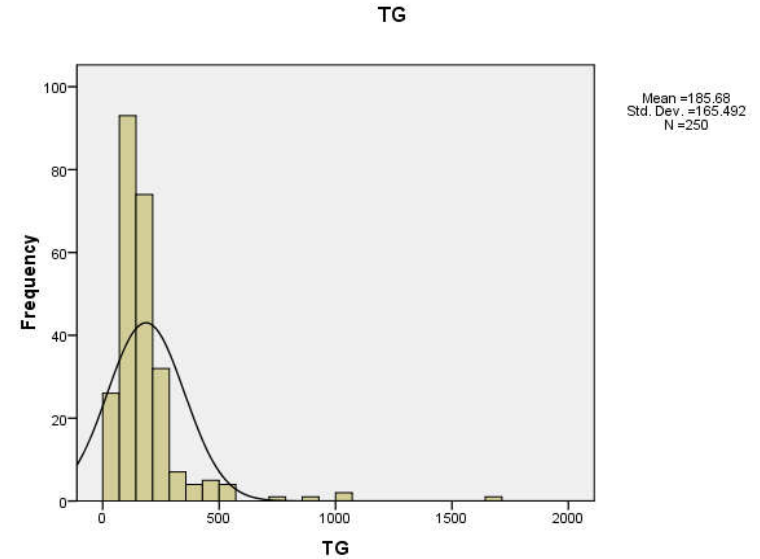


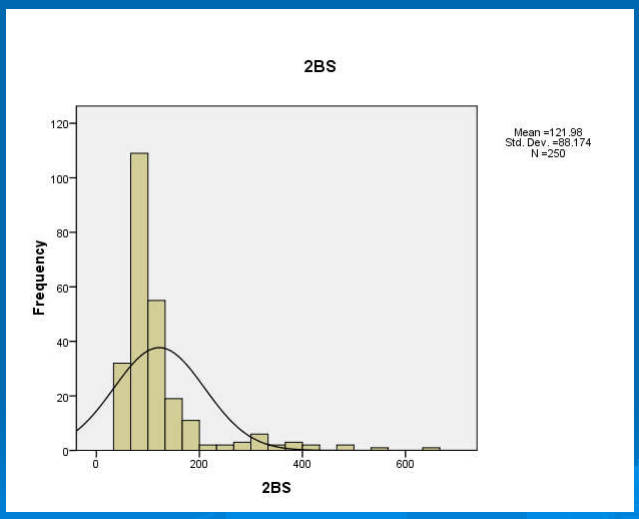
نمودار ۱- هیستوگرام غلظت ید ادرار ۲۴ ساعته

Skewness



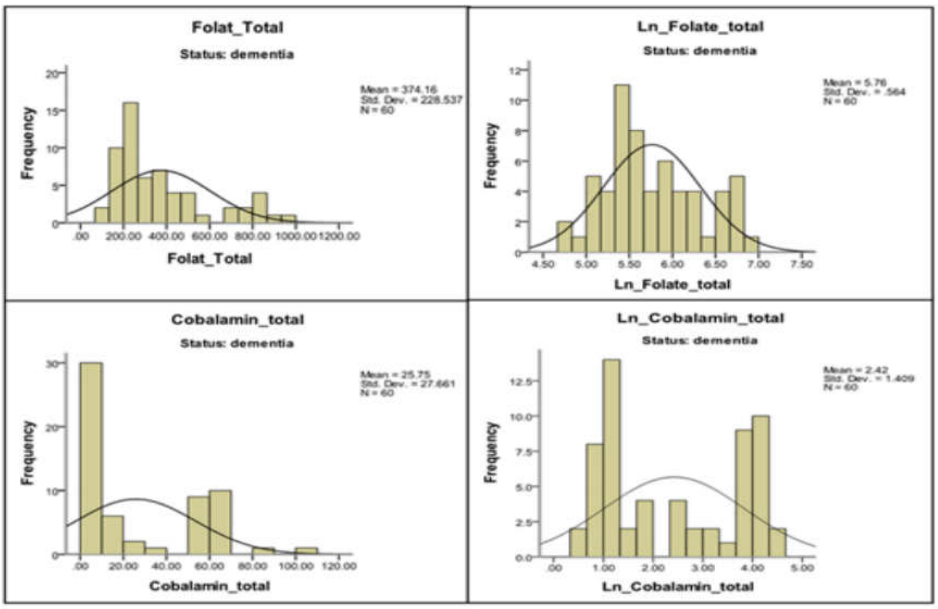

HDL
Mean =36.1
Std. Dev. =8.375
N =250
Frequency
HDL


TG
Mean =185.68
Std. Dev. =165.492
N =250
Frequency
TG



مثال

	WEIGHT	HEIGHT	HDL	TG	2hrBS
N Valid	250	249	250	250	250
Mean	**67.5**	159.2	36.1	185.6	121.9
Median	**67.2**	158.5	35.0	148.5	97.0
Std. Deviation	12.40	8.8	8.3	165.5	88.2
Range	63.5	44.0	49	1645	619
Minimum	38.5	137.0	14	43	34
Maximum	102.0	181.0	63	1688	653



شکل ۲- نمودار هیستوگرام فولات و کوبالامین توتال در افراد دمانس

چند ک ها Quantiles

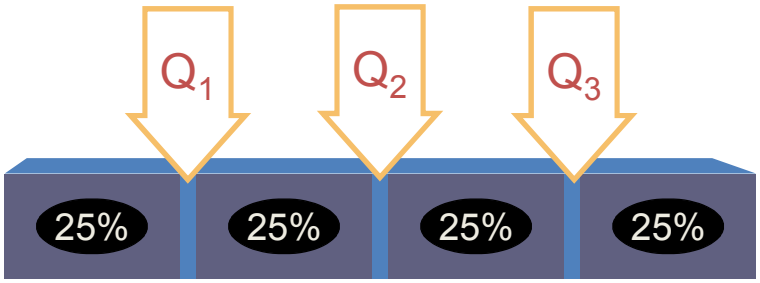
- صد ک ها Centiles
- دهک ها Tentiles
- پنجک ها Quintiles
- چار ک ها Quartiles
- ....

Quartiles

- Three numbers which divide the ordered data into four equal sized groups.

- Q_1 has 25% of the data below it.
- Q_2 has 50% of the data below it. (Median)
- Q_3 has 75% of the data below it.

Quartiles



Cholesterol

L(M)=(53+1)/2=27

$L(Q_1)=(26+1)/2=13.5$

100	124	148	170	185	215
101	125	150	170	185	220
106	127	150	172	186	260
106	128	152	175	187	
110	130	155	175	192	
110	130	157	180	194	
119	133	165	180	195	
120	135	165	180	203	
120	139	165	180	210	
123	140	170	185	212	

Cholesterol Quartiles

- Q_1= 127.5
- Q_2= 165 (Median)
- Q_3= 185

Five-Number Summary

- minimum = 100
- Q_1 = 127.5
- M = 165
- Q_3 = 185
- maximum = 260

Interquartile Range (IQR)
= $Q_3 - Q_1$
= 57.5

IQR gives spread of middle 50% of the data

ارتباط مصرف فرآورده‌های لبنی با سندرم متابولیک و اجزای آن در نوجوانان: مطالعه قند و لیپید تهران

لطفا مفهوم اعداد گزارش شده برای تری گلیسرید را بیان کنید.

جدول ۱- ویژگی‌های آمارنگاری و اجزای سندرم متابولیک افراد شرکت‌کننده در مطالعه در دو گروه سالم و مبتلا به سندرم متابولیک

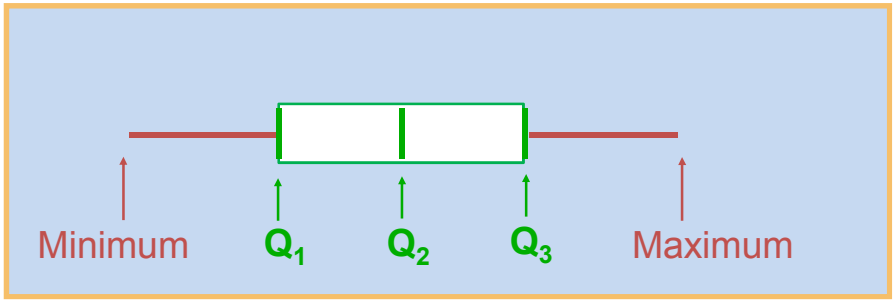
	سالم	مبتلا به سندرم متابولیک	P†
	۶۱۱ نفر	۱۷۴ نفر	
سن (سال)	۱۴/۹±۲/۹	۱۴/۲±۲/۸	<۰/۰۰۱
نمایه‌ی توده‌ی بدن (کیلوگرم بر متر مربع)	۲۱/۱±۴/۲	۲۵/۲±۴/۴	<۰/۰۰۱
گلوکز ناشتای خون (میلی‌گرم در صد میلی‌لیتر)	۸/۱± ۹۱/۸	۹۴/۶±۷/۸	<۰/۰۰۱
تری‌گلیسیرید خون (میلی‌گرم در صد میلی‌لیتر)‡	۷۴ (۶۰-۹۴)	۱۳۱ (۱۱۳-۱۶۶)	<۰/۰۰۱
کلسترول ـ HDL خون (میلی‌گرم در صد میلی‌لیتر)	۵۲/۵±۱۰/۳	۴۰/۹±۶/۹	<۰/۰۰۱
دورکمر (سانتی‌متر)	۷۴/۲±۱۱/۵	۸۶/۳±۱۰/۹	<۰/۰۰۱
فشار خون سیستولی (میلی‌متر جیوه)	۱۰۰±۱۱/۶	۱۰۷±۱۳/۰	<۰/۰۰۱
فشار خون دیاستولی (میلی‌متر جیوه)	۶۶/۲±۹/۴	۷۰/۳±۱۱/۶	<۰/۰۰۱

*داده‌ها به استثنای تری‌گلیسیرید صورت میانگین±انحراف معیار است. † مقادیر P به استثنای تری‌گلیسیرید با استفاده از آزمون آماری تی به دست آمد.
‡ برای تری‌گلیسیرید میانه و دامنه میان چارکی گزارش شده است.

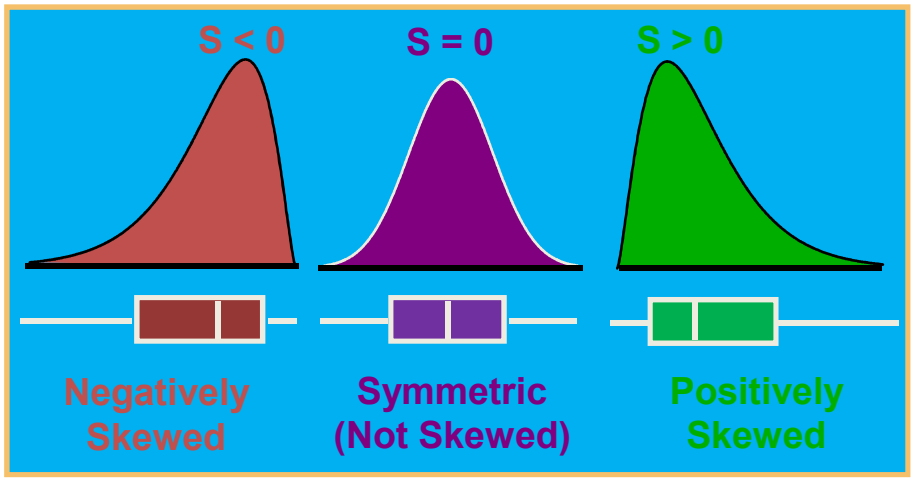
Boxplot

- Central box spans Q_1 and Q_3.
- A line in the box marks the median M.
- Lines extend from the box out to the minimum and maximum.

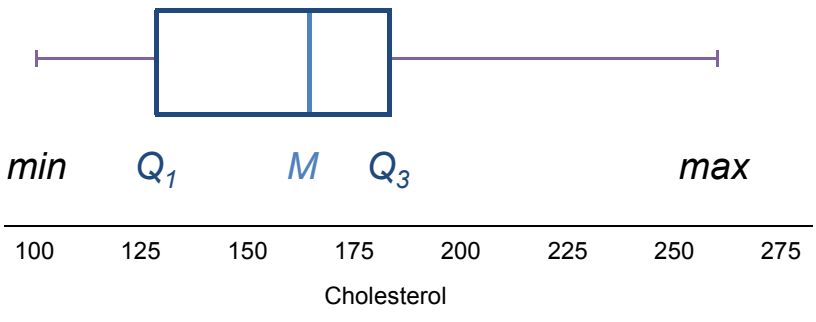
Box Plot



Skewness: Box Plots, and Coefficient of Skewness



Cholesterol: Boxplot



Identifying Outliers

- The central box of a boxplot spans Q_1 and Q_3; recall that this distance is the Interquartile Range (*IQR*).
- We call an observation a suspected **outlier** if it falls more than $1.5 \times IQR$ above the third quartile or below the first quartile.

چندک ها

Quantiles

		WEIGHT	HEIGHT	SYSTOLIC BP	TG	CHOLESTEROL
Median		67.250	158.500	120.00	148.50	192.50
Percentiles	5	47.000	145.750	95.00	63.00	126.10
	10	51.550	148.000	100.00	71.00	146.10
	20	56.000	151.000	105.00	94.00	157.00
	25	58.000	152.000	110.00	102.00	163.00
	30	60.000	153.000	110.00	110.30	168.00
	40	64.000	156.000	110.00	129.40	179.40
	50	67.250	158.500	120.00	148.50	192.50
	60	70.000	161.000	120.00	169.00	200.60
	70	74.000	164.000	130.00	198.00	219.70
	75	76.125	166.000	136.25	209.00	226.00
	80	78.000	168.000	140.00	223.00	231.00
	90	83.900	171.000	150.00	302.10	251.90
	95	90.000	174.250	170.00	434.50	271.00

لطفا مفهوم اعداد گزارش شده برای کلسیم را بیان کنید.

جدول ۳- میانگین دریافت مواد مغذی بر اساس چارک‌های دریافت کل لبنیات*

	چارک‌های دریافت لبنیات†				
	۱	۲	۳	۴	P‡
انرژی دریافتی(کیلوکالری در روز)	۲۹۲۳±۵۷/۹	۲۵۶۱±۵۷/۶	۲۵۴۵±۵۷/۸	۲۸۹۵ ±۵۸/۰	<۰/۰۰۱
دانسیته‌ی انرژی	۸۲/۴±۱/۳	۸۸/۳±۱/۳	۹۲/۰±۱/۳	۹۶/۵±۱/۳	<۰/۰۰۱
کربوهیدرات (درصد از انرژی دریافتی)	۶۰/۵±۱/۰	۵۸/۲±۱/۰	۵۷/۲±۱/۰	۵۶/۹±۱/۰	۰/۰۴۶
پروتئین (درصد از انرژی دریافتی)	۱۳/۰±۱/۰	۱۳/۹±۱/۰	۱۴/۴±۱/۰	۱۷/۴±۱/۰	۰/۰۱۲
چربی (درصد از انرژی دریافتی)	۲۹/۶±۲/۳	۳۱/۳±۲/۳	۳۱/۲±۲/۳	۳۷/۴±۲/۳	۰/۰۷۹
چربی‌اشباع (درصد از انرژی دریافتی)	۸/۶±۲/۳	۹/۹±۲/۳	۱۰/۸±۲/۳	۱۷/۲±۲/۳	۰/۰۳۷
فیبر غذایی (گرم/۱۰۰۰کیلوکالری)	۱۹/۱±۲/۵	۱۸/۳±۲/۵	۱۶/۸±۲/۵	۲۰/۸±۲/۵	۰/۷۱۵
کلسترول (میلی‌گرم/۱۰۰۰کیلوکالری)	۸۱/۵±۳/۹	۹۱/۵±۳/۹	۹۸/۸±۳/۹	۱۱۲±۳/۹	<۰/۰۰۱
کلسیم (میلی‌گرم/۱۰۰۰کیلوکالری)	۴۳۰±۱۰/۵	۵۳۹±۱۰/۵	۵۹۵±۱۰/۵	۷۴۳±۱۰/۶	<۰/۰۰۱
فسفر (میلی‌گرم/۱۰۰۰کیلوکالری)	۵۷۱±۶/۲	۶۳۱±۶/۱	۶۸۸±۶/۱	۷۸۹±۶/۲	<۰/۰۰۱

*داده‌ها به صورت میانگین±انحراف معیار است. مقادیر با استفاده از آنالیز کوواریانس (ANCOVA) پس از تعدیل برای سن و جنس به دست آمد. † چارک‌های دریافت کل لبنیات به ترتیب از کمترین به بیشترین عبارت بود از: <۳۱۱/۳، ۳۱۱/۳- ۴۵۲/۵، ۴۵۲/۵- ۶۲۲/۸، ≥۶۲۲/۸ گرم در روز، ‡ مقدار P<۰/۰۵ از نظر آماری معنی‌دار است.

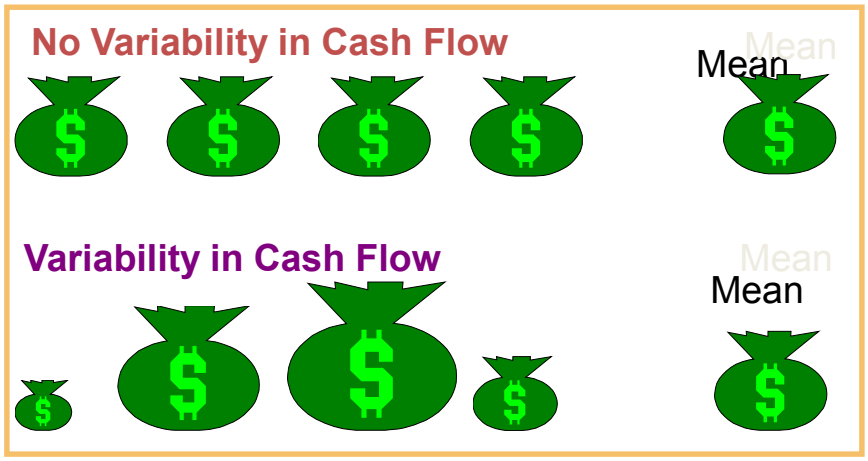
خلاصه سازی داده ها

شاخصهای پراکندگی

- دامنه تغییرات Range
- $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$
- واریانس Variance
- انحراف معیار Standard Deviation
- ضریب تغییرات Coefficient of Variation
- $CV = \left(\frac{S}{\bar{X}}\right) \cdot 100\%$

Variance
and
Standard Deviation

Variability



Variance: a measure of how data points differ from the mean

- Data Set 1: 3, 5, 7, 10, 10
  Data Set 2: 7, 7, 7, 7, 7

What is the mean and median of the above data set?

Data Set 1: mean = 7, median = 7
Data Set 2: mean = 7, median = 7

But we know that the two data sets are not identical! The **variance** shows how they are different.

We want to find a way to represent these two data set numerically.

How to Calculate sd?

	Score X	$X - \overline{X}$	$(X - \overline{X})^2$
1	3		
2	5		
3	7		
4	10		
5	10		
Totals	35		

The mean is 35/5=7.

How to Calculate sd?

	Score X	$X - \overline{X}$	$(X - \overline{X})^2$
1	3	3-7=-4	
2	5	5-7=-2	
3	7	7-7=0	
4	10	10-7=3	
5	10	10-7=3	
Totals	35		

How to Calculate sd?

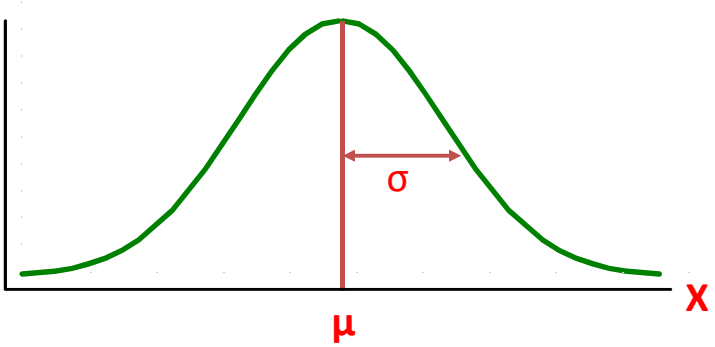
	Score X	$X - \overline{X}$	$(X - \overline{X})^2$
1	3	3-7=-4	16
2	5	5-7=-2	4
3	7	7-7=0	0
4	10	10-7=3	9
5	10	10-7=3	9
Totals	35		38

How to Calculate sd?

	Score X	$X-\bar{X}$	$(X-\bar{X})^2$
1	3	3-7=-4	16
2	5	5-7=-2	4
3	7	7-7=0	0
4	10	10-7=3	9
5	10	10-7=3	9
Totals	35		38

$$sd = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}} = \sqrt{\frac{38}{5-1}} = \sqrt{9.5} = 3.1$$

The Population Mean and Standard Deviation



$$\sigma^2 = \frac{\sum\left(x - \bar{X}\right)^2}{N}$$ For population variance

$$s^2 = \frac{\sum\left(x - \bar{X}\right)^2}{n-1}$$ For sample variance

Example: Age of Patients

Male	Female
28	27
22	27
21	28
26	6
18	27

Find the mean, median, range?

mean	**23**	**23**
median	**22**	**27**
range	**10**	**22**

What can be said about this data?

Due to the outlier, the median is more typical of overall performance.

Which diver was more consistent?

Dive	Male Age X	$X-\overline{X}$	$(X-\overline{X})^2$
1	28	5	25
2	22	-1	1
3	21	-2	4
4	26	3	9
5	18	-5	25
Totals	115	0	64

$$Male\ age\ sd = \sqrt{\frac{64}{5-1}} = \sqrt{16} = 4$$

$$Female\ age\ sd = \sqrt{\frac{362}{5-1}} = \sqrt{90.5} = 9.5$$

Example: Sample Standard Deviation

- Square root of the sample variance

X	$X-\overline{X}$	$(X-\overline{X})^2$
2,398	625	390,625
1,844	71	5,041
1,539	-234	54,756
1,311	-462	213,444
7,092	0	663,866

$$S^2 = \frac{\sum (X-\overline{X})^2}{n-1} = \frac{663,866}{3} = 221,288.67$$

$$S = \sqrt{S^2} = \sqrt{221,288.67} = 470.41$$

Example

Metabolic rates of 7 men (cal./24hr.) :

1792 1666 1362 1614 1460 1867 1439

$$\bar{x} = \frac{1792+1666+1362+1614+1460+1867+1439}{7}$$

$$= \frac{11{,}200}{7}$$

$$= 1600$$

Example

Observations x_i	Deviations $x_i - \bar{x}$	Squared deviations $(x_i - \bar{x})^2$
1792	1792–1600 = 192	$(192)^2$ = 36,864
1666	1666 –1600 = 66	$(66)^2$ = 4,356
1362	1362 –1600 = -238	$(-238)^2$ = 56,644
1614	1614 –1600 = 14	$(14)^2$ = 196
1460	1460 –1600 = -140	$(-140)^2$ = 19,600
1867	1867 –1600 = 267	$(267)^2$ = 71,289
1439	1439 –1600 = -161	$(-161)^2$ = 25,921
	sum = 0	sum = 214,870

Variance and Standard Deviation
Example from Text

$$s^2 = \frac{214{,}870}{7-1} = 35{,}811.67$$

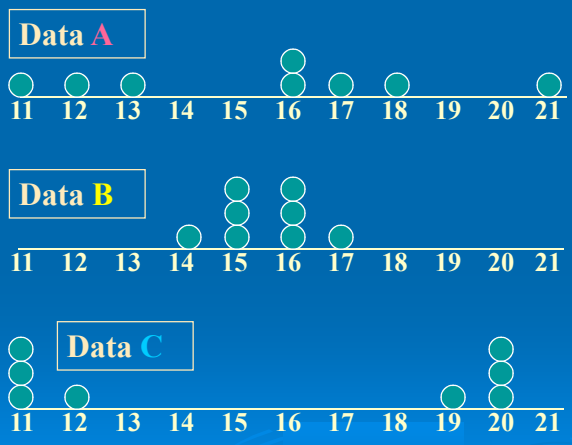
$$s = \sqrt{35{,}811.67} = 189.24 \text{ calories}$$

Standard Deviation انحراف معیار

- مهم ترین شاخص پراکندگی

$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

مقایسه انحراف معیار داده ها

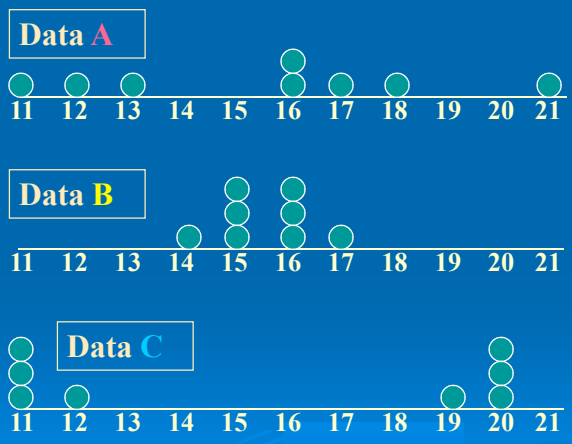


Mean = 15.5

Mean = 15.5

Mean = 15.5

مقایسه انحراف معیار داده ها



Mean = 15.5
sd = 3.338

Mean = 15.5
sd = .9258

Mean = 15.5
sd = 4.57

SYSTOLIC BP

Age Groups	N	Mean	Std. Deviation	Minimum	Maximum	Range
<35	54	109.54	11.421	90	140	50
35-44	60	112.08	17.303	85	180	95
45-54	52	125.87	20.430	100	185	85
55-64	54	136.30	22.554	100	180	80
65+	30	134.33	23.034	100	210	110
Total	250	122.30	21.848	85	210	125

Coefficient of Variation ضریب تغییرات

- پراکندگی نسبت به میانگین را نشان می دهد
- مقایسه پراکندگی دو یا چند گروه
- مقایسه پراکندگی دو متغیر با واحد اندازه گیری متفاوت

$$CV = \left(\frac{S}{\bar{X}}\right) \cdot 100\%$$

Coefficient of Variation ضریب تغییرات

$\bar{X}=3000 \quad Sd=300$

وزن نوزادن

$\bar{X}=40Kg \qquad Sd=2$

وزن افراد ۱۲ ساله

Coefficient of Variation ضریب تغییرات

$\bar{X}=3000 \qquad Sd=300$

$\bar{X}=3000\,gr \qquad Sd=300\,gr \qquad \mathrm{CV}=\frac{300}{3000}\times 100=10\%$

وزن نوزادن

$\bar{X}=40Kg \qquad Sd=2$

وزن افراد ۱۲ ساله

$\bar{X}=40kg \qquad Sd=2kg \qquad \mathrm{CV}=\frac{2}{40}\times 100=5\%$

ضریب تغییرات داده های زیر را محاسبه کنید. کدام یک پراکندگی بیشتری دارد؟

SYSTOLIC BP

Age Groups	N	Mean	Std. Deviation
<35	54	109.54	11.421
35-44	60	112.08	17.303
45-54	52	125.87	20.430
55-64	54	136.30	22.554
.=65	30	134.33	23.034
Total	250	122.30	21.848

ضریب تغییرات داده های زیر را محاسبه کنید. کدام یک پراکندگی بیشتری دارد؟

SYSTOLIC BP

Age Groups	N	Mean	Std. Deviation	CV%
<35	54	109.54	11.421	10.43
35-44	60	112.08	17.303	15.44
45-54	52	125.87	20.430	16.23
55-64	54	136.30	22.554	16.55
.=65	30	134.33	23.034	17.15
Total	250	122.30	21.848	17.86

Weighted Mean میانگین وزنی

n_1	n_2	.....	n_k	تعداد در گروه j
$\overline{X}_1$	$\overline{X}_2$	.....	$\overline{X}_k$	میانگین گروه j

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{K} n_j \overline{X}_j$$

$$n = \sum_{j=1}^{k} n_j$$

نسبت وزنی

n_1	n_2	.....	n_k	تعداد در گروه j
p_1	p_2	.....	p_k	نسبت در گروه j

$$P = \frac{1}{n}\sum_{j=1}^{K} n_j p_j$$

$$n = \sum_{j=1}^{k} n_j$$