Descriptive Statistics

Measures of central tendency

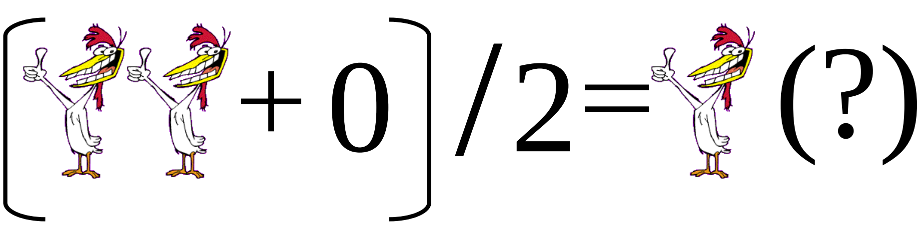
Measures of variability / dispersion

Prof. Giuseppe Verlato

Unit of Epidemiology & Medical Statistics
Department of Diagnostics & Public Health
University of Verona

TRILUSSA's dilemma

**«According to current statistics,
you have got a chicken per year:
if you don't find this chicken in your
expenses, it is easy to explain:
another guy is eating two chickens»**



Trilussa's original poem

«Me spiego: da li conti che se fanno
seconno le statistiche d'adesso
risurta che te tocca un pollo all'anno:
e, se nun entra ne le spese tue,
t'entra ne la statistica lo stesso
perché c'è un antro che ne magna due»

Statistics can assess not only how many chickens are eaten on average by the population under study, but also whether the chickens are equally or unequally distributed within the population

SYNTHESIS

MEASURES OF CENTRAL TENDENCY

MEASURES OF VARIABILITY

Statistical Synthesis

A data set is fully described by three main properties:

- **Central tendency or location**
- **Variability** or **dispersion** or **spread**
- **Shape**

These synthetic measures, which can adequately summarize a data set, are named:

- **statistics**, expressed with Latin letters, when computed on a sample
- **parameters**, expressed with Greek letters, when computed on a population

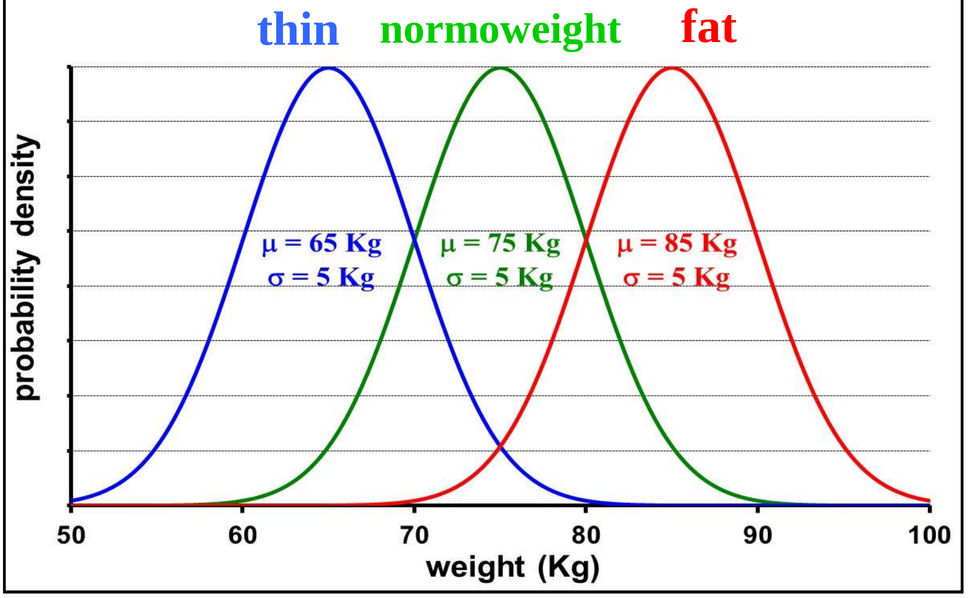
Measures of central tendency

- MEAN
- MEDIAN
- MODE

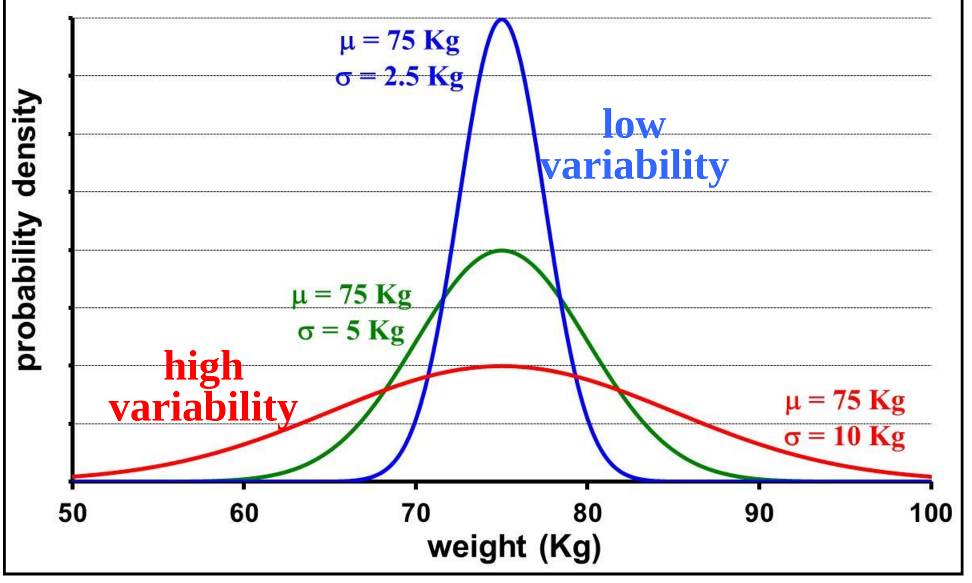
Measures of variability

- RANGE and INTERQUARTILE RANGE
- SUM OF SQUARES → VARIANCE → STANDARD DEVIATION → COEFFICIENT of VARIATION

These three distributions differ only for the mean μ (measure of central tendency)



These three distributions differ only for the standard deviation σ (measure of variability)



EXAMPLE

Which are the main MEASURES OF CENTRAL TENDENCY in the following data set ?

x_i	3	15	11	4	5	8	6	4	4
Absolute rank	*1*	*3*	*3*	*3*	*5*	*6*	*7*	*8*	*9*
Ordered series ($x_{(i)}$)	**3**	**4**	**4**	**4**	**5**	**6**	**8**	**11**	**15**

MODE, most frequent value

MEDIAN, middle value in an ordered series

MEAN
($\sum_i x_i / n$)
$= 60/9 = 6.67$

Most biological variables (weight, height, diastolic pressure, heart rate) have a normal distribution, where mean, median and mode are the same.

Some variables (reaction time, survival time, number of metastatic lymph nodes, serum concentrations of triglycerides) have a skewed (asymmetric) distribution, where mean, median and mode differ.

Fictitious example:

During the Nineties 7 physicians were working in a hospital unit: 2 specializing doctors, 2 assistants, 2 senior physicians and 1 director. Their income was respectively **2, 2, 3, 3, 4, 4** e **25** millions lire per month. Which measure of central tendency is most suited to summarize this data set ?

mean = $\Sigma x/n = 43/7 =$ **6.14 millions per month**
median = value of the 4^{th} observation in the ordered series = **3 millions per month**

The measure of central tendency, which best summarizes these physicians' income, is the median not the mean.

Exercise on the median

Age in years: 39 25 18 14 69 81 42

1) Data are sorted in ascending order:
14 18 25 39 42 69 81

2) The rank of the median is computed:
n=7 (odd) rank = (n+1)/2 = (7+1)/2 = 8/2

3) The value of the 4° observation is assessed:
14 18 25 39 42 69 81

MEDIAN = 39 years

Exercise on the median

Age in years: 81 72 16 42 38 8

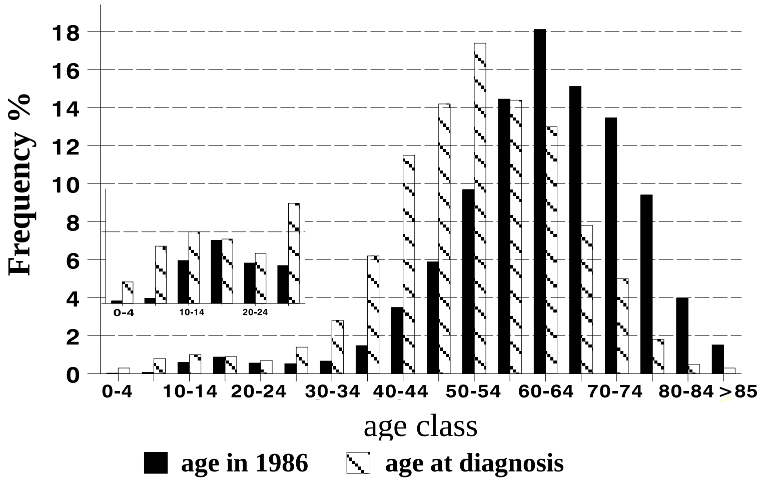
1) Data are sorted in ascending order:
8 16 38 | 42 72 81

2) The rank of the median is computed
n=6 (even) rank = (n+1) / 2 = 7/2 = 3.5

3) The median is the mean of the third and fourth observations:
8 16 38 42 72 81

MEDIAN = (38+42)/2 = 40 years

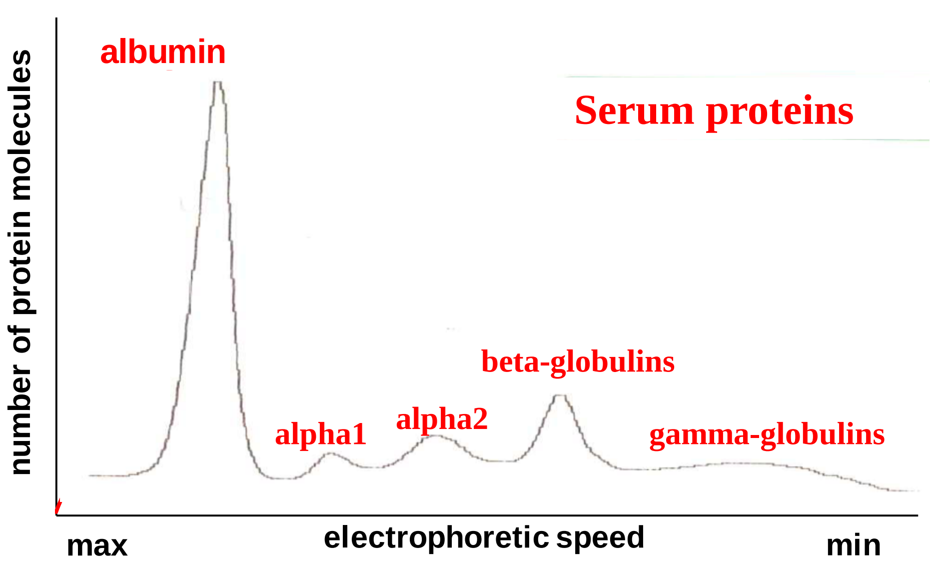
EXAMPLE of BIMODAL distribution
DIABETIC MEN in VERONA on the 31.12.1986

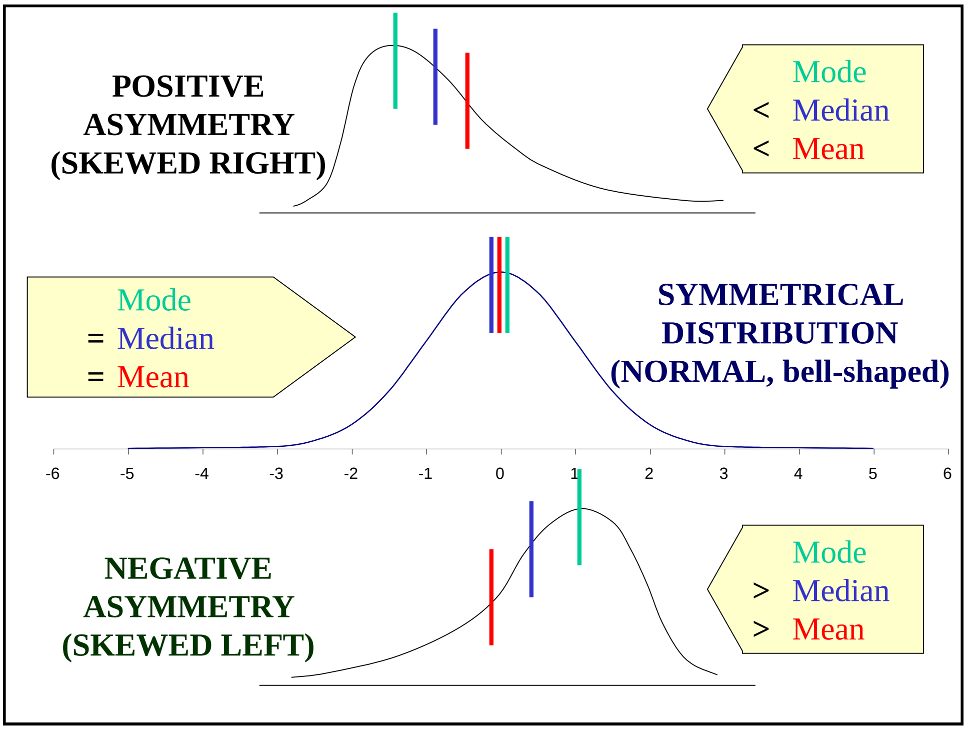


N.B. : 100% = all diabetic men

Muggeo M, Verlato G, …, de Marco R (1995) The Verona Diabetes Study: a population-based survey on known diabetes mellitus prevalence and 5-year all-cause mortality. *Diabetologia*, 38: 318-325

MULTI-MODAL DISTRIBUTION





Mean	Median	Mode
The most used measure of central tendency	The most suited measure with asymmetrical distributions (reaction time, survival time)	The most suited measure when a value has a high relative frequency (number of fingers in the right hand)
Easy to mathematically handle	the 50th percentile	The most frequently occurring value
It is based on all available information ($\Sigma x/n$)		
A weighted value is easy to compute: $\bar{x} = (\bar{x}_1 n_1 + \bar{x}_2 n_2) / (n_1+n_2)$		
1st property of the mean: the sum of the deviations from the mean is zero: $\Sigma(x - \bar{x}) = 0$	the sum of distances is the lowest when computed from the median $\Sigma\lvert x - me\rvert = \min$	
the sum of squared deviations is the lowest when computed from the mean: $\Sigma(x - \bar{x})^2 = \min$		

WEIGHTED MEAN

Sample 1 (horse riders)
$n_1=30$ $\bar{x}_1=50$ kg

Sample 2 (Sumo wrestlers)
$n_2=10$ $\bar{x}_2=150$ kg

Overall mean

$\cancel{\bar{x} = (50+150)/2 = 100 \text{ kg}}$

Wrong computation: overall mean should be closer to the mean of the largest sample

$$\bar{x} = (n_1 \bar{x}_1 + n_2 \bar{x}_2) / (n_1 + n_2) = (30*50 + 10*150) / (30+10) =$$
$$(1500 + 1500) / 40 = 3000/40 = 75 \text{ kg}$$

	Chickens per month	Reference value	Deviation	Deviation^2
	1		-5	25
	6	6	0	0
	11	mean	5	25
Total	18		0	50
Deviations are computed from values other than the mean				
	1		-4	16
	6	5	1	1
	11		6	36
Total	18		3	53
	1		-7	49
	6	8	-2	4
	11		3	9
Total	18		-6	62

1° property: the algebraic sum of the deviations from the mean is zero

2° property: the sum of squared deviations is the lowest when computed from the mean

Arithmetic sequence:

Number = previous number + k

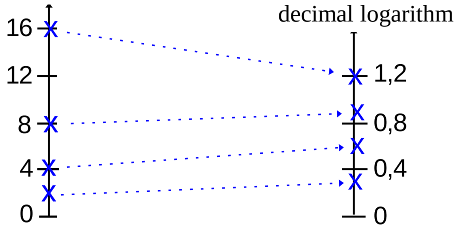
- 3 4 5 6 7 8
- 5 7 9 11

Geometric sequence:

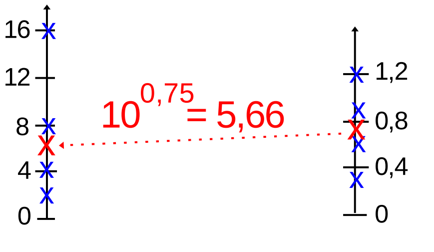
Number = previous number * k

- 4 8 16 32 64 128
- 6 12 24 48 96
- 1/2 1/4 1/8 1/16 1/32

Geometric mean = anti-log of the mean of log-transformed values



$$\frac{0,3+0,6+0,9+1,2}{4}=0,75$$



WEIGHTED MEAN

Hospital stay (days) after surgery for hemorrhoids in a given hospital

Days of hospital stay	Number of patients	Overall days
1	9	1*9 = 9
2	15	2*15 = 30
3	12	3*12 = 36
4	9	4* 9 = 36
5	5	5*5 = 25
TOTAL	50	136

MEAN = $\sum$ nx / $\sum$n = 136/50 = 2.72 days

MODE and MEDIAN in a frequency distribution

mode = 2 days

Days of hospital stay	Number of patients	Cumulative abs. frequency
1	9	9
2	**15**	24
3	12	**36**
4	9	45
5	5	50
TOTAL	50	

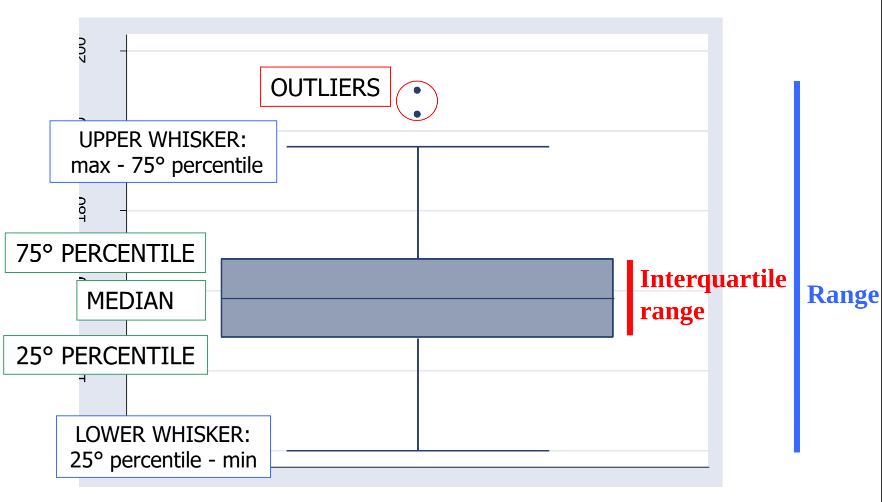
1 1 1 1 1 1 1 1 1 2
2 2 2 2 2 2 2 2 2 2
2 2 2 2 **3** **3** 3 3 3 3
3 3 3 3 3 3 4 4 4 4
4 4 4 4 4 5 5 5 5 5

MEDIAN = (3 + 3) / 2 = 3 days

Measures of variability

Italian name	English name
Campo di variazione	Range
Distanza interquartile	Interquartile range
Devianza	Sum of squares (SSq)
Varianza	Mean Square (MSq) / variance
Deviazione standard	Standard deviation
Coefficiente di variazione	Variation coefficient

HEIGHT DISTRIBUTION AMONG 1° CLASS MEDICAL STUDENTS
BOX-and-WHISKERS PLOT (GRAFICO SCATOLA E BAFFI)



Whisker maximum length = 1.5 * interquartile range (box height)

Range

$$\boxed{\textbf{Range} = X_{max} - X_{min}}$$

The range is the simplest measure of variation to find: it is simply the highest value minus the lowest value.

Disadvantages

- The range only uses the extreme values, without considering intermediate values
- It tends to increase with increasing number of observations
- It is largely affected by outliers

Interquartile range

$$\mathbf{IQR = Q_3 - Q_1}$$

Difference between the 3rd quartile (*75° percentile*) and the 1st quartile (*25° percentile*)

Features

- This interval comprises half of the values, which represent the middle 50% of the distribution.
- It is not affected by outliers or extreme values (**robust statistic**).
- It is suited to express variability in skewed distributions.

EXAMPLE: DESCRIPTION OF A SERIES OF GASTRIC CANCER PATIENTS

In the series of 921 patients, the total number of dissected lymph nodes was 23,288, with an average of 25.3 ± 16.3 (mean ± SD) dissected nodes per case (median 21, range 1-108). The mean number of metastatic nodes was 4.3 ± 7.5 (median 1, range 0-74) in the overall series and 8.3 ± 8.7 (median 5, range 1-74) in pN+ patients.

Bibliografia

De Manzoni G, Verlato G, Roviello F, Morgagni P, Di Leo A, Saragoni L, Marrelli D, Kurihara H, Pasini F, for the Italian Research Group for Gastric Cancer (2002) The new TNM classification of lymph node metastasis minimizes stage migration problems in gastric cancer patients. Brit J Cancer , 87: 171-174

Table 3. Allergy parameters in subjects without self-reported allergic rhinitis and in subjects with perennial, seasonal and perennial+seasonal rhinitis. Absolute frequencies with percentage in brackets are reported for all variables but total IgE, which is expressed as median (interquartile range).

	No rhinitis	Subjects with self-reported allergic rhinitis			
	(n=745)	Perennial (n=19)	Seasonal (n=50)	Perennial + seasonal (n=87)	P value
Parental allergy	120/736 (16)	5/19 (26)	21/48 (44)	30/87 (34)	<0.001
Pos. specific IgE					
D.pteronyssinus	56/623 (9)	6/15 (40)	7/43 (16)	19/70 (27)	<0.001
Cat	17/623 (3)	2/15 (13)	4/43 (9)	12/70 (17)	---
Timothy grass	57/623 (9)	3/15 (20)	26/43 (60.5)	39/70 (56)	<0.001
Cl.herbarum	3/623 (0.5)	1/15 (7)	1/43 (2)	3/70 (4)	---
Pariet. judaica	29/623 (5)	1/15 (7)	16/43 (37)	32/70 (46)	<0.001
Total IgE	**36.1 (13.2-101)**	**110.5 (11.6-217.5)**	**87 (38-214.5)**	**106 (50.5-240)**	**<0.001**

Significance of differences was evaluated by chi-squared test for categorical variables and by one-way ANOVA for total IgE after logarithmic transformation. Significance was not evaluated by chi-squared test (---) when cells with expected value<5 exceeded 25%. NS = not significant

Olivieri M, Verlato G, Corsico A, Lo Cascio V, Bugiani M, Marinoni A, de Marco R, for the Italian ECRHS group (2002) Prevalence and features of allergic rhinitis in Italy. Allergy, 57:600-606

In the example dealing with gastric cancer the **range** is used as measure of variability to describe a series as a whole.

In the example dealing with allergic rhinitis the **interquartile range** is used to **compare** variability among groups with **very different size**: indeed, the group with perennial allergic rhinitis comprises only 19 subjects, while the group without allergic rhinitis includes 745 subjects.

	Chickens per month	Mean	Deviation	Deviation^2	
	5		-1	1	
	6	**6**	0	0	
	7		1	1	
Total	18		0	2	sum of squares
	1		-5	25	
	6	**6**	0	0	
	11		5	25	
Total	18		0	50	sum of squares

Sum of squares (SSq) ≡ $\Sigma(x - \bar{x})^2$
Sum of squared deviations of single values from the mean

5
6 } Sum of squares = 2
7

Sum of squares doubles even if variability remains constant

5
6
7 } Sum of squares = 4
5
6
7

Variance was created to take into account **sample size**!

Variance = sum of squares / n

However, if one considers a sample of only one subject eating 6 chickens/month…

Mean	Sum of squares	Uncorrected variance	Corrected variance
6	0	0/1 = 0	0/0 = ?

If one divides sum of squares by **n-1** rather than by n, variance is undetermined, better reflecting the real situation.

Mean	Sum of squares	Corrected variance
6 chickens/mo	2 $\text{chickens}^2/\text{mo}^2$	1 $\text{chickens}^2/\text{mo}^2$
6 chickens/mo	50 $\text{chickens}^2/\text{mo}^2$	25 $\text{chickens}^2/\text{mo}^2$

However, $\text{chickens}^2/\text{month}^2$ is a unit of measurement difficult to understand !

To solve this difficulty, standard deviation was developed !

Standard deviation = √ variance

Mean	Corrected variance	Standard deviation
6 chickens/mo	1 $\text{chickens}^2/\text{mo}^2$	1 chickens/mo
6 chickens/mo	25 $\text{chickens}^2/\text{mo}^2$	5 chickens/mo

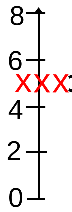
Low variability: 6 ± 1 chickens/month (mean ± SD)

High variability: 6 ± 5 chickens/month (mean ± SD)

Variability

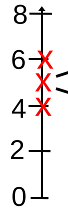
Σx Sum of values

Σx^2 Sum of squared values



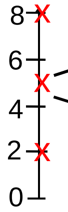
$\Sigma x = 5 + 5 + 5 = 15$

$\Sigma x^2 = 25+25+25 = 75$



$\Sigma x = 4 + 5 + 6 = 15$

$\Sigma x^2 = 16+25+36 = 77$



$\Sigma x = 2 + 5 + 8 = 15$

$\Sigma x^2 = 4+25+64 = 93$

As variability increases, the difference between sum of squared values and sum of values also increases

Variability

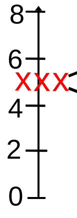
Σx

Σx^2

St.dev.= $\sqrt{(SSq/(n-1))}$

St.dev. = $\sqrt{(\text{variance})}$

St.dev. = $\sqrt{(0/(n-1))} = 0$

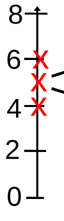


$\Sigma x = 5 + 5 + 5 = 15$

$\Sigma x^2 = 25+25+25 = 75$

Sum of Squares = $(5-5)^2 + (5-5)^2 + (5-5)^2 = 0^2 + 0^2 + 0^2 = 0$

$\Sigma x^2 - (\Sigma x)^2/n = 75 - 15^2/3 = 75 - 225/3 = 75-75 = 0$

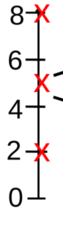


$\Sigma x = 4 + 5 + 6 = 15$

$\Sigma x^2 = 16+25+36 = 77$

Sum of Squares = $(4-5)^2 + (5-5)^2 + (6-5)^2 = (-1)^2 + 0^2 + 1^2 = 2$

$\Sigma x^2 - (\Sigma x)^2/n = 77 - 15^2/3 = 77 - 225/3 = 77-75 = 2$



$\Sigma x = 2 + 5 + 8 = 15$

$\Sigma x^2 = 4+25+64 = 93$

Sum of Squares = $(2-5)^2 + (5-5)^2 + (8-5)^2 = (-3)^2 + 0^2 + 3^2 = 18$

$\Sigma x^2 - (\Sigma x)^2/n = 93 - 15^2/3 = 93 - 225/3 = 93-75 = 18$

Sum of Squares - SSq

- It is a measure of variability around a center
- It is the basis, the starting point, to compute all the other measures of variability, used in parametric statistics.
- **Variance**, S**tandard Deviation, Coefficient of Variation** are computed from Sum of Squares

Heuristic equation **Empirical equation**

$$\sum_{k=1}^{N}(x_k - \bar{x})^2 \longrightarrow \sum_{k=1}^{N}(x_k)^2 - \frac{\left(\sum_{k=1}^{N} x_k\right)^2}{N}$$

Variance or Mean Square (MSq)

- It is the **average squared deviation from the mean**, i.e. the sum of squares divided by the number of observations in the sample (***n***) or in the population (***N***)

In the population **In the sample (*corrected variance!*)**

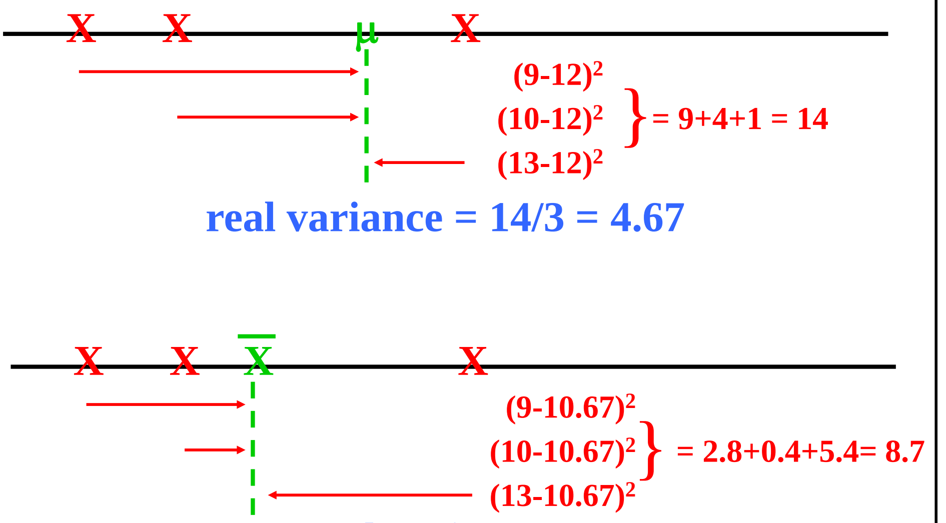
$$\sigma^2 = \frac{\sum_{k=1}^{N}(x_k - \mu)^2}{N} \qquad S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Sigma squared

Number of observations

Degrees of freedom (**df**)

Sample: 9, 10, 13 **μ = 12**



real variance = 14/3 = 4.67

uncorrected variance = 8.67/3 = 2.89

corrected variance = 8.67/2 = 4.33

Variance

- It takes into account all observations, and hence it is largely affected by outliers. For this reason, variance is suited only for symmetric distributions.
- Variance is the most important measure of variability in statistical theory.
- To compute sum of squares, deviations were squared as well as their unit of measurement. Variance is also expressed in squared units, and cannot be directly compared with the mean or other measures of central tendency. For this reason, variance is usually not reported in biomedical scientific literature.
- **Degrees of freedom (df)** represent the number of independent observation in the sample under study (n -1), as a statistic (the mean) has already been computed from available data.

Standard Deviation - SD

• <u>(Positive) square root</u> of the **Variance**

<u>In the sample</u>

$$\sqrt{\frac{\sum_{k=1}^{N}(x_k - \bar{x})^2}{n-1}}$$

Main features of Standard Deviation

- It measures the <u>distance from the mean.</u> Remember that the deviation is positive or negative, while the distance is an absolute number. It measures the **variability** of a random variable <u>around the mean.</u>
- It is directly <u>comparable</u> with the <u>mean</u>, as they are computed using the same unit of measurement. For this reason the standard deviation is the <u>most widely used</u> measure of variability in the biomedical scientific literature.
- However it is <u>less important</u> than v**ariance** in <u>statistical theory</u>.

EXERCISE

	x_i	x_i^2		$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	3	9		3-6= -3	9
	5	25		5-6= -1	1
	6	36	$\bar{x}$ = 30/5 =6	6-6= 0	0
	7	49		7-6= +1	1
	9	81		9-6= +3	9
total	30	200		0	20

Sum of squares = $\Sigma(x - \bar{x})^2 = 20$

or

Sum of squares = $\Sigma x^2 - (\Sigma x)^2/n = 200 - 30^2/5 =$

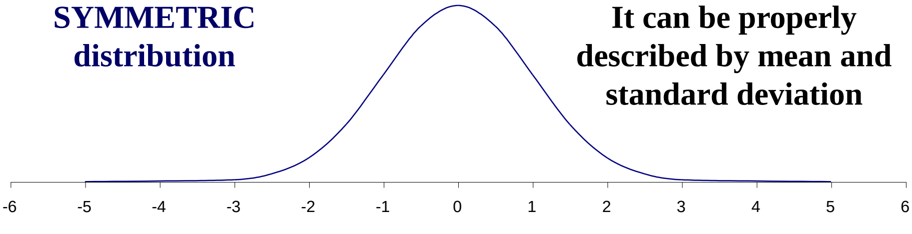
$= 200 - 900/5 = 200 - 180 = 20$

Variance = SSq/(n-1) = 20/(5-1) = 20/4 = 5

Standard deviation = $\sqrt{5}$ = 2.24

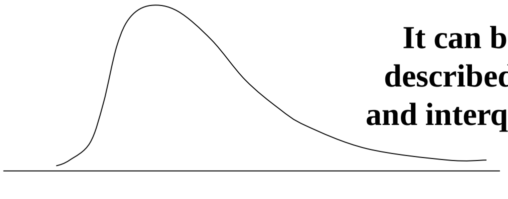
6 ± 2.24 (mean ± SD)

SYMMETRIC distribution



It can be properly described by mean and standard deviation

SKEWED distribution



It can be properly described by median and interquartile range

Coefficient of Variation (CV) - 1

SAME variable but very different means

Three newborns weigh respectively **3, 4 and 5 Kg** (mean ± SD: 4 ± **1 Kg**).
Three one-year-old infants weigh **10, 11 and 12 Kg** (mean ± SD: 11 ± **1 Kg**).
The standard deviation is the same in both groups, but common sense suggests that weight variability could be higher in the newborn group.

Two DIFFERENT variables

In 91 female 1st class medical students at Verona University in 1995/96,
weight was 55.1 ± **5.7 Kg** (mean ± SD) with a range of **45-70 Kg,**
height was 166.1 ± **6.1 cm** (mean±SD) with a range of **150-182 cm.**
Which is higher ? the variability of weight or the variability of height ?

Coefficient of Variation (CV) - 2

To answer these questions one has to compute the **coefficient of variation:**
CV = (standard deviation / mean) * 100.
In other words, standard deviation is expressed as percentage of the mean.

	Mean	Standard deviation	CV
Newborns	4 Kg	1 Kg	**25 %**
One-year-old infants	11 Kg	1 Kg	**9.1 %**

Weight variability is higher in newborns.

	Mean	Standard deviation	CV
Weight	55.1 Kg	5.7 Kg	**10.3 %**
Height	166.1 cm	6.1 cm	**3.7 %**

Weight variability is higher than height variability.

Measures of Shape

Measures of symmetry

1) Galton skewness = [(Q3-Q2) - (Q2-Q1)] / (Q3-Q1)
where Q3, Q2, Q1 =75th, 50th and 25th percentile

For example, if we consider both men and women attending the 1st class of Medical School at Verona University in 1995:

Galton skewness = [(174.5-169)-(169-164)] / (174.5-164] =
= [5.5-5] / 10.5 = 0.5 cm / 10.5 cm = 4.8%

A small positive asymmetry is detected.

2) Pearson's coefficient of skewness = (mean - mode) / st.dev.

Measure of flattening

1) Kurtosis = a measure of the concentration of the distribution around its mean. It indicates whether the distribution is flattened or has a peak around the mean.
Kurtosis = $[\Sigma(x-\bar{x})^4/n] / [\Sigma(x-\bar{x})^2/n]^2$