Relative risk: RR

Risk factor

- something (a factor) that
  - if present
    - increase the probability of development of a disease

- ex. smoking is a risk factor for lingual cancer
- ex. periodontosis is a risk factor for nephropathy

The contingency table

	B With disease	non(B) Without disease	Total
A Risk factor positive	a	b	a+b
non A Risk factor negative	c	d	c+d
Total	a+c	b+d	n

The contingency table

- a frequency table
- show the relationship between
  - a dependent and an independent variable

Disease / Risk factor	B With illness	non(B) Without illness	Total
Risk factor + A	a	b	a+b
Risk factor – non A	c	d	c+d
Total	a+c	b+d	n

- The risk of disease when the factor is present = The probability of disease (event A) occurrence given that the factor (event B) is already happen
  - $P(A|B) = P(A \text{ and } B) / P(B) = a/(a+b)$

- The risk of disease when the factor is absent = The probability of occurrence of disease (event A) given that the factor (event B) is not happen
  - $P(A|\bar{B}) = P(A \text{ and } \bar{B}) / P(\bar{B}) = c/(c+d)$

Disease / Risk factor	B With illness	non(B) Without illness	Total
Risk factor + A	a	b	a+b
Risk factor – non A	c	d	c+d
Total	a+c	b+d	n

Relative risk (RR)

= the ratio between the probability of the event occurring in the group with the risk factor (exposed) to the probability of the event occurring in the group without the presence of the risk factor (unexposed)

$$RR = \frac{P(B|A)}{P(B|\bar{A})} = \frac{P(Disease|Risk\ factor)}{P(Disease|\overline{Risk\ factor})} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Non A

RR=1

No risk

RR>1

Risk

RR<1

Protection

EXAMPLE Relative risk

- Probability of diabetes at obese persons 50/100 = 0.5 (50%)
- Probability of diabetes at non-obese person 100/1000 = 0.1 (10%)
- RR = Pr(B | A) / Pr(B | non A) = 0.5 / 0.1 = 5 (times high)
-

	with diabetes	without diabetes	Total
obese	50	50	100
non obese	100	1000	1100

EXAMPLE

- Probability of heart attack at an active person 100/1000 = 0.10
- Probability of heart attack at a sedentary person 300/1000 = 0.30
- RR = Pr(B | A) / Pr(B | non A) = 0.10 / 0.30 = 0.33 (<1 protective factor)
-

	with heart attack	without heart attack	Total
active	100	900	1000
sedentary	300	700	1000

EXAMPLE

- Probability of fracture at an active person 100/1000 = 0.10
- Probability of fracture at a sedentary person 100/1000 = 0.10
- RR = Pr(B | A) / Pr(B | non A) = 0.10 / 0.10 = 1 (no risk, no protective factor)
-

	with fracture	without fracture	Total
active	100	900	1000
sedentary	100	900	1000

Bayes' theorem

Dependent events

Conditional probability—the probability of an outcome depending on an earlier outcome.

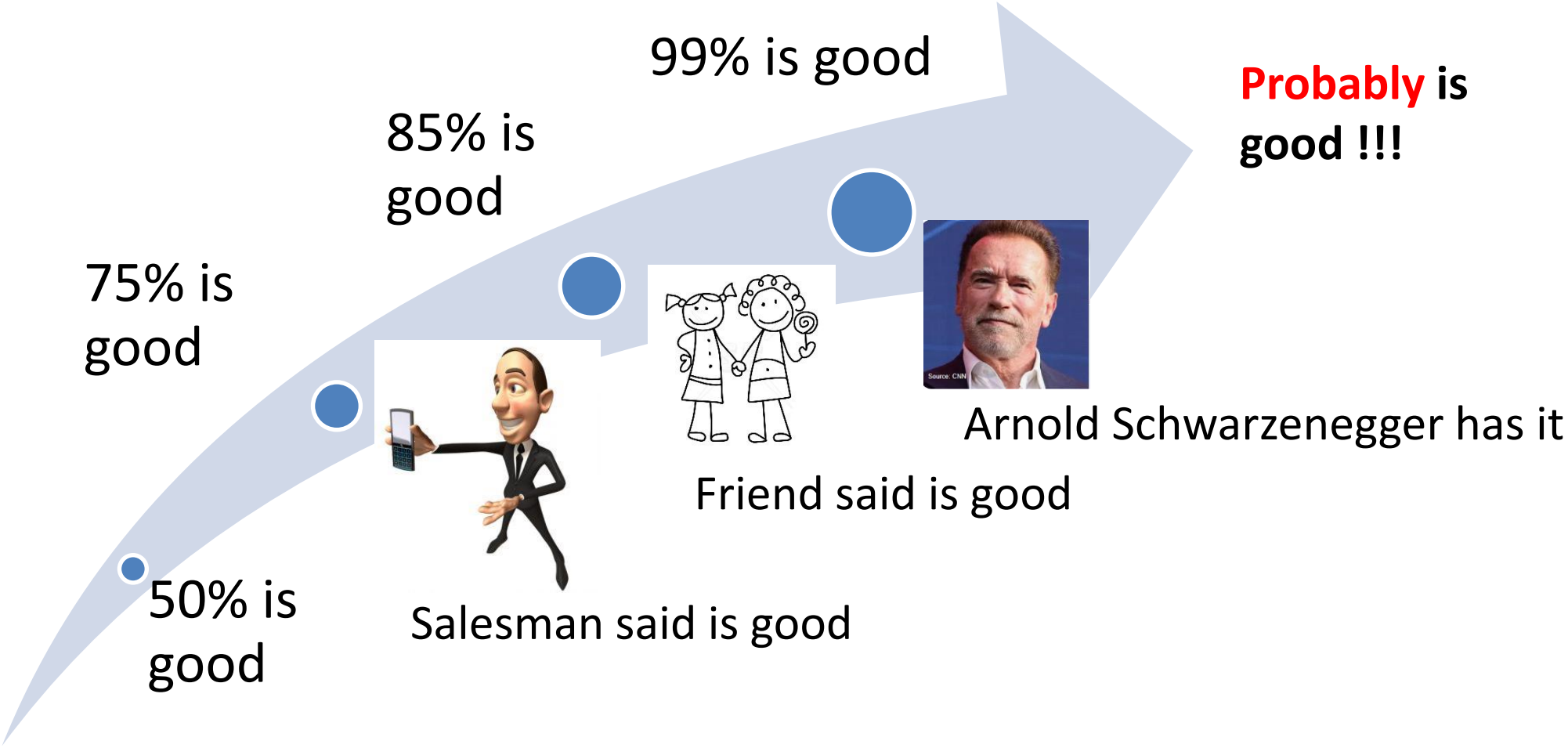
$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)}.$$

Involved in making medical decision – reasoning process - interpreting diagnostic procedures.

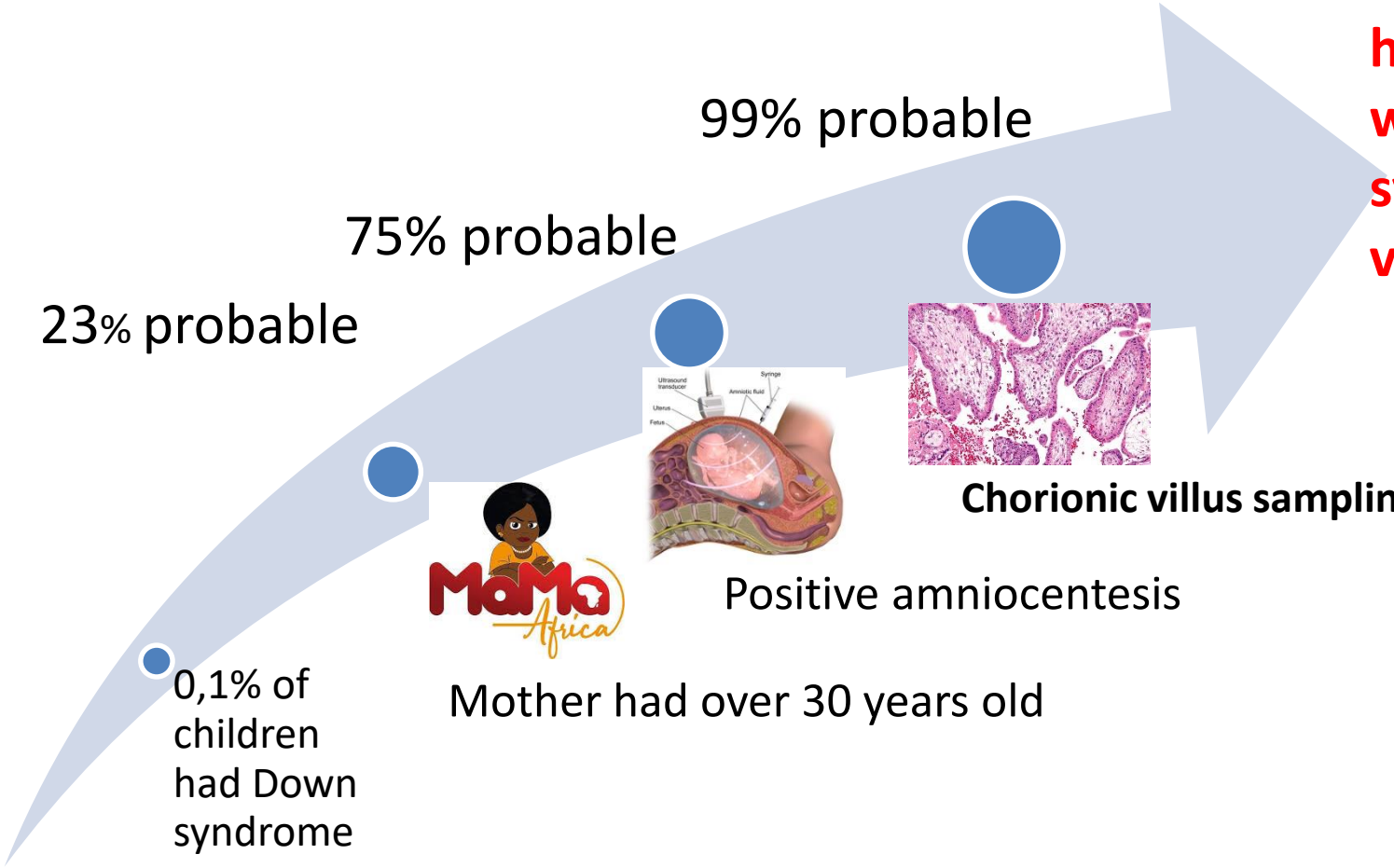
Bayes' theorem - to improve what we know after something happen

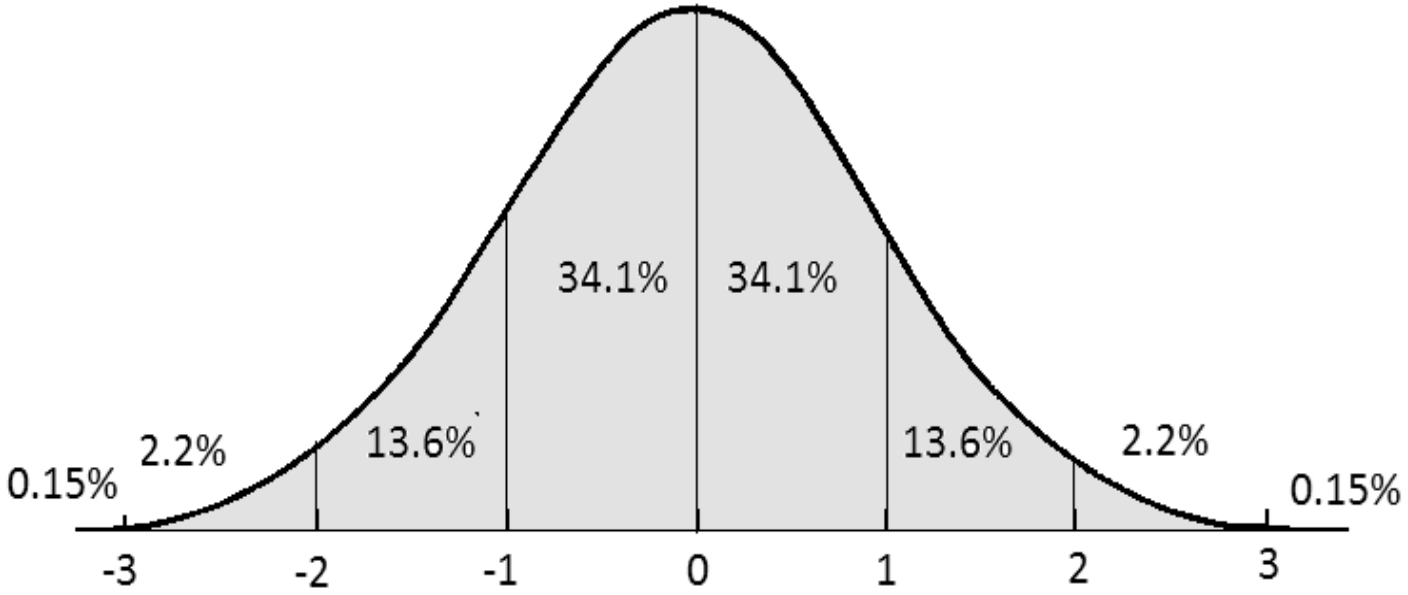
- That’s how medical doctors put a diagnostic
- That’s how we and computers learn (AI)
- That’s how spam filter (email) works

I want to buy an Apple smart mobile phone
Is it good or is it not?



Down syndrome during pregnancy





Trainer: PhD. MsC. Bondor Cosmina-Ioana

Probability distributions

A ALWAYS

S SEEK

K KNOWLEDGE

Objectives

Population, sample

Normal distribution

Sampling distribution

Population – representative sample

Population

- from statistical point of view
  - a set of elements with the same characteristic

  - medical dentistry
    - patients
      - dental office

- medicine
  - patients
    - hospital

we have a question

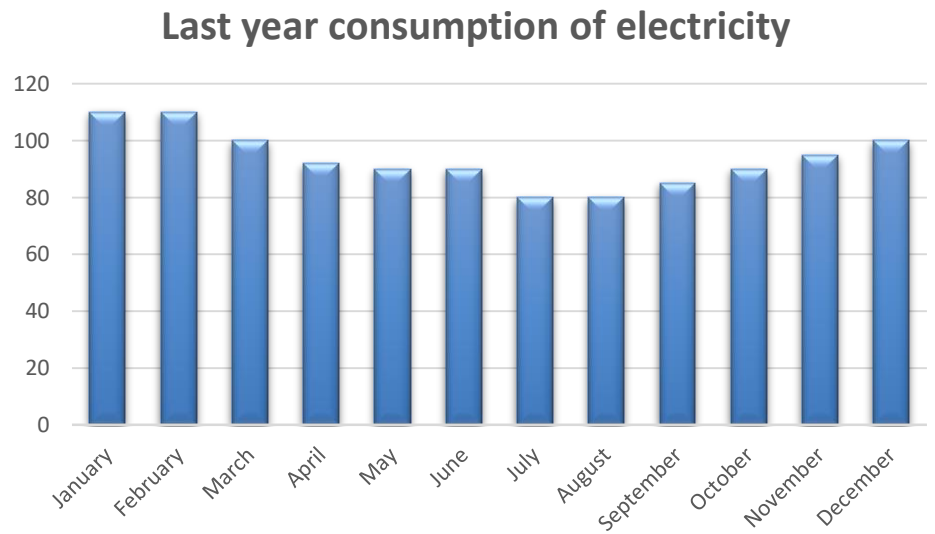
- we should follow **whole population** during their life time to find it?
  - we cannot
    - too much money, too many employee, too many data, a lot of errors
- we should follow **2000 people** during life time?
  - we can
    - cost relative small, few staff, few errors
- we should follow **100 people** during life time?
  - we can, but the results will be not reliable

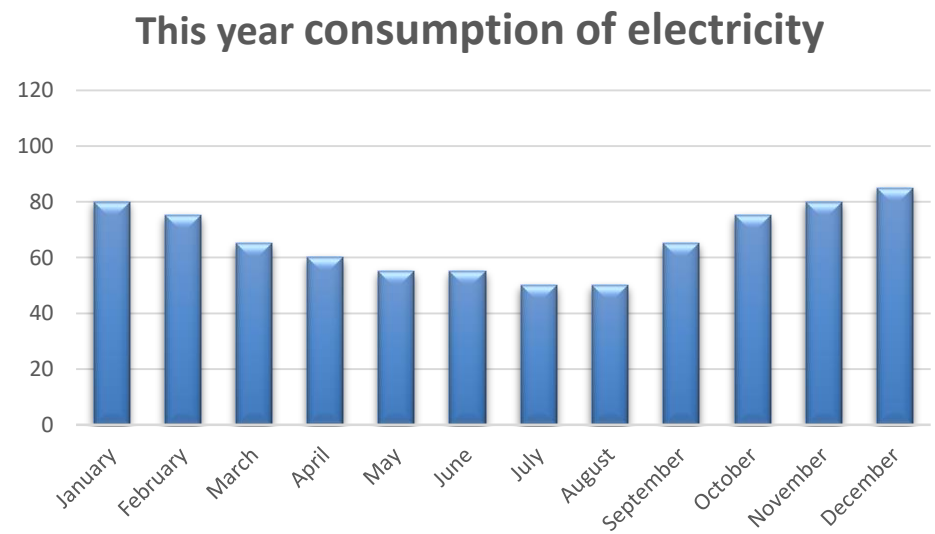
Conclusion: we need a sample

Sample

- a subset of a population
- give as a glimpse of whole population
- give as an ideea of the shape of the data

- why shape is important? if we want to compare two groups





Why to study samples instead of whole population?

- More quickly
- Less expansive
- Less dangerous
- More accurate conclusions

- **Less dangerous** – if we should test a new drug and this drug is dangerous, than only few people will be affected, not whole population
- **More accurate conclusions** – if we investigate a phenomenon or we put a diagnostic or interpret a radiography and we have several investigators (whole population – we need more than one investigator), each one will have a different subjective approach, so we will have a lot of errors, not all of them understand the process of investigation.

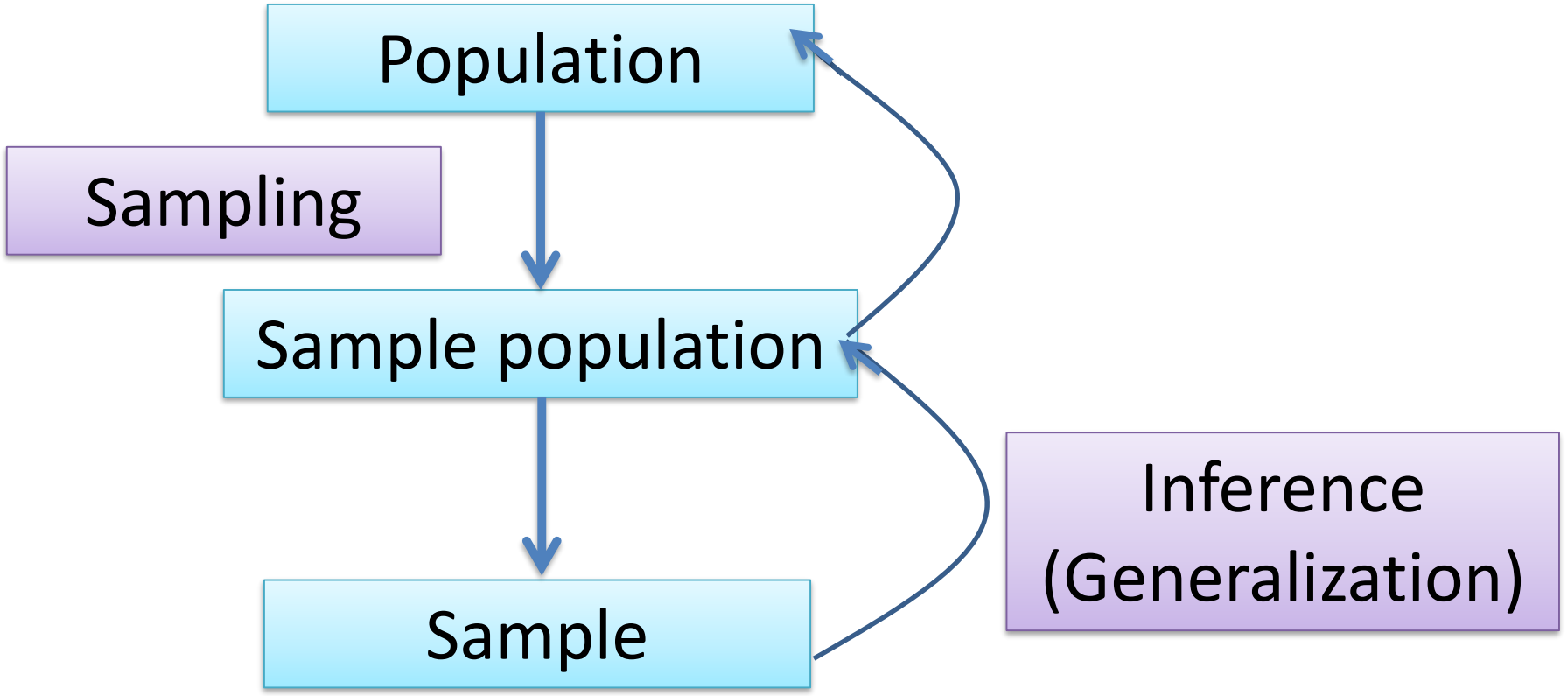
Clinical research

Generalizing results:

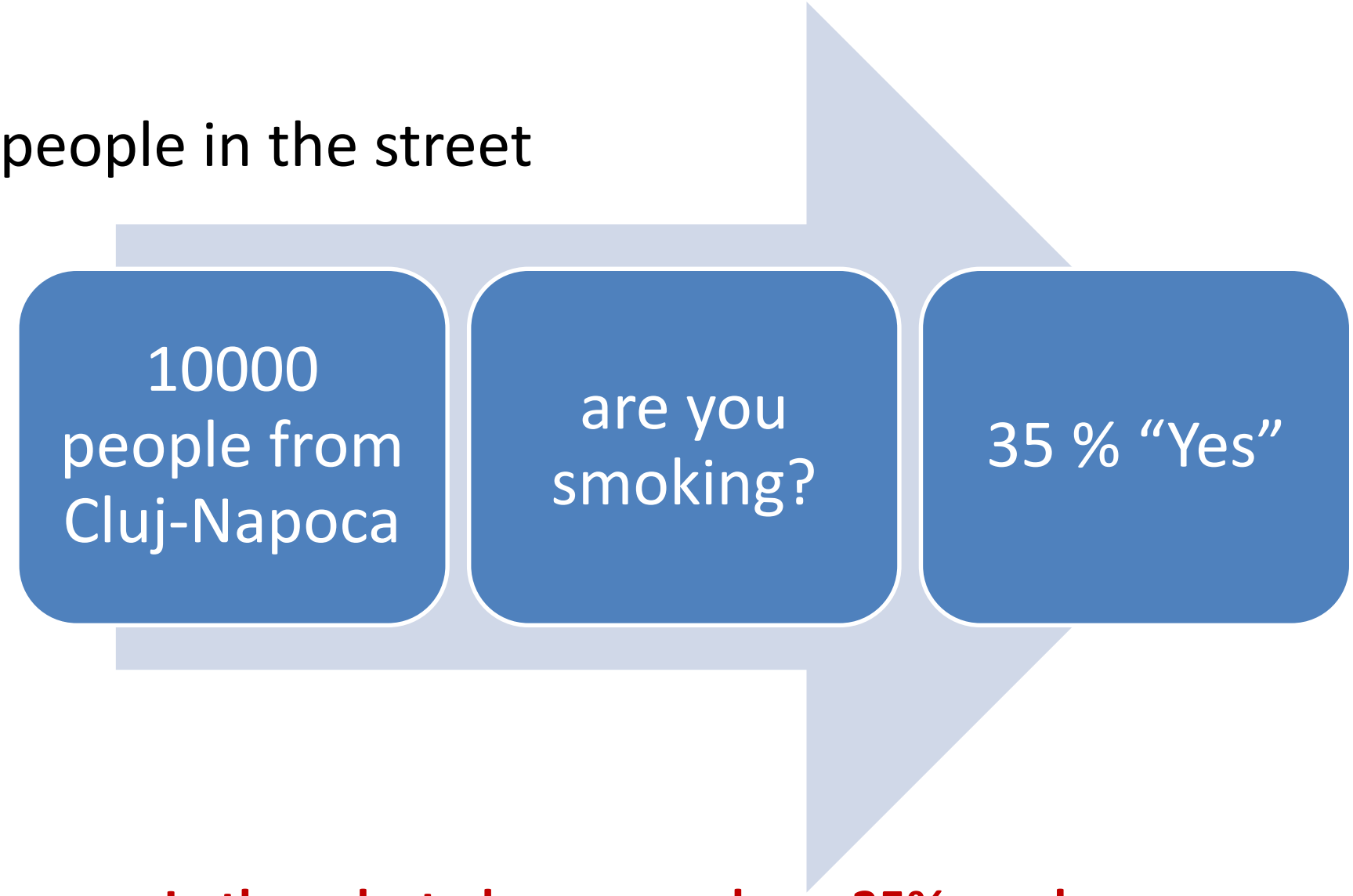
Group of patients (sample) → similar subjects

Sample → population (should be similar)

Sampling


Population
Sampling
Sample population
Sample
Inference
(Generalization)

Example:
we select people in the street

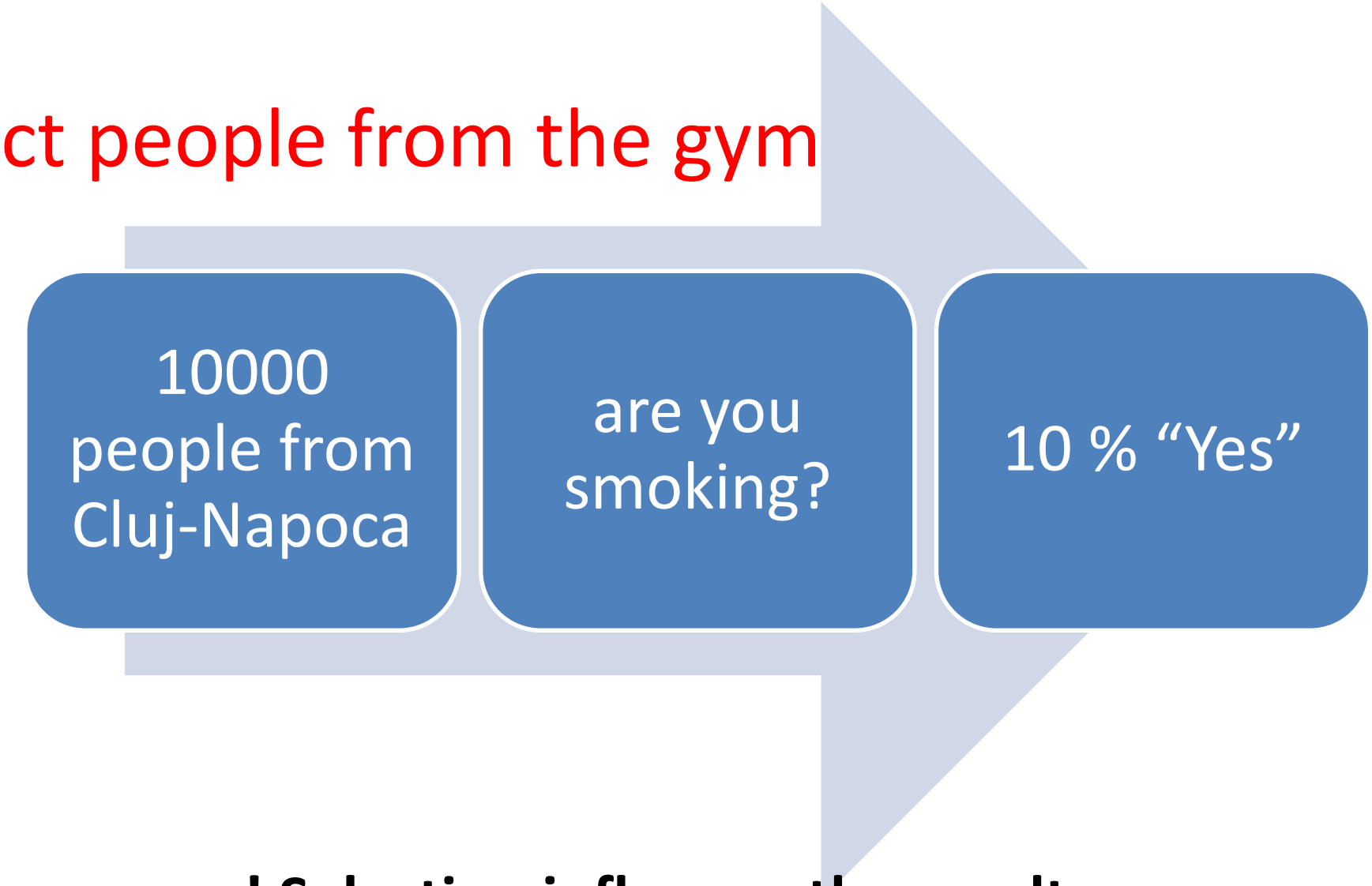


In the selected group we have 35% smokers

When can we make generalizations?

Example of generalization to population: The probability of smoking in Cluj-Napoca is 35%

but,
if we select people from the gym



! Selection influence the results

The selected sample should be reprezentative for the population

Representative sample for the population

- have the same distribution of important characteristics as the population

where:

  - important characteristics = in connection with the studied characteristic

select random → high probability to select a representative sample

select random → find 30% frequency of smoking

↓

Generalization (estimate the frequency) to the population:

Frequency of smoking in Cluj-Napoca is 30%

Population

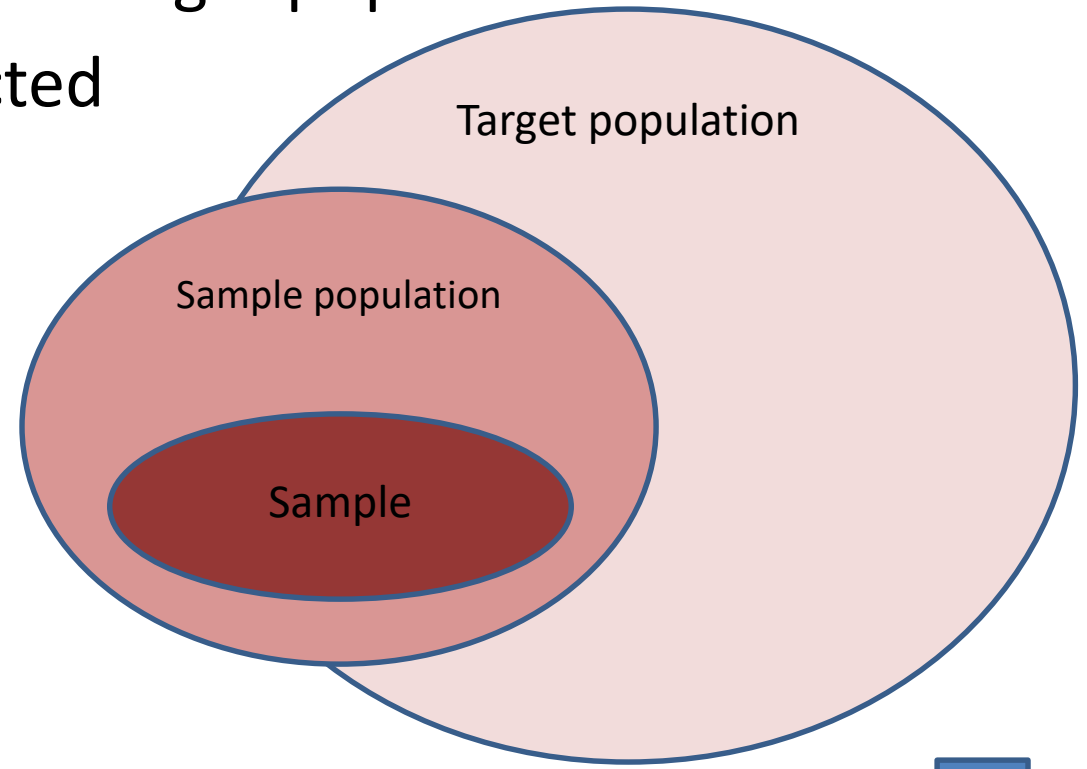
Random selection

- **Random selection** – each individual from the population have the same probability to be selected in the sample

- Ex. Go to town hall take all ID numbers of the people who live in the city. Extract randomly ID's

Erorr (bias) of selection

- to select wrong
  - sample population is not a subset of the target population
  - a part of the population was not selected



Error - Bias of selection

Example

- to select from the gym in the example with the prevalence of smokers
- to select people from the ski resort only if you want to appreciate the number of fractures in Cluj-Napoca
- to select only students from Medicine if you want to appreciate the distance between the Faculty of Dentistry and student homes

Even if we select random there is still probability of errors

- there is only a small probability to select with errors if we select random
- ex. prevalence of smoking – the are chances that even if we randomly select, all the people selected to play a sport
- What to do? we **replicate** the study ! Only one study cannot demonstrate a hyphothesis

Probabilistic - Sampling methods

Probabilistics – the probability of each subject to be selected is known

- **Random** sampling
  - each subject has the same probability to be selected
- **Systematic** sampling
  - each k^{th} subject is selected
- **Stratified** sampling
  - population is divided into subgroups and a random sample is selected from each subgroups
- **Cluster** sampling
  - population is divided into clusters and a random sample is selected from each cluster (cluster = geographic zones)

Non-probabilistic - Sampling methods

Non-probabilistic – the probability of each subject to be selected is unknown

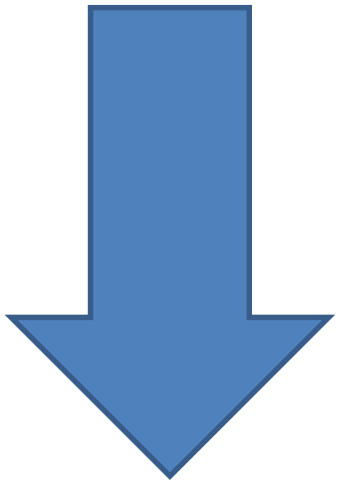
- **Snowball** sample – respondent are asked to help recruit people who fit the criteria of inclusion
- **Convenient sample** – subjects are selected because they are available to the researcher

Census

- **Census** – all population are selected
  - in this case we do not need inferential statistics

Condiţia inferenței eşantion → populaţie

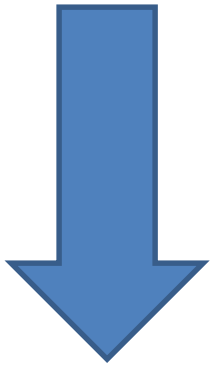
Selecţia aleatoare



Probabilitatea mare de a selecta un eșantion reprezentativ

When we talk about a sample

- we suppose it was randomly selected



- it is representative for the population on which we want to made the conclusion of the research

Sample with random selection

Measurements

random result

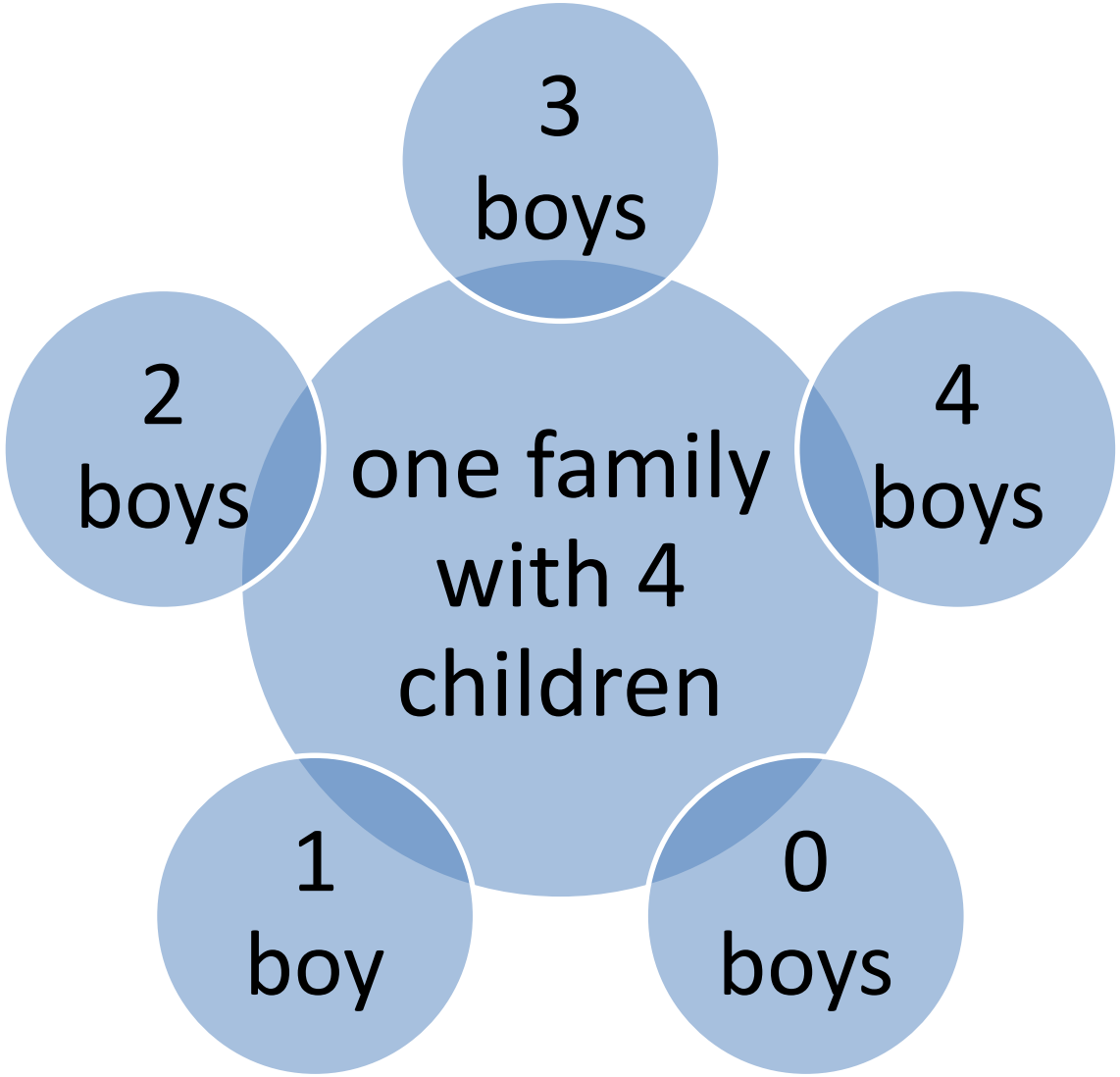
Result = random variable

Probability distribution

- The values of a random variable can be summarized in a frequency distribution which we call **probability distribution**

Example: Experiment

Probability of a child born being male ≈0.5 (50% of the cases)



In **100 family with 4 children selected randomnly**?

In **randomnly selected 100 families with 4 children**?

Frequency distribution of variable X the number of times occur the possible values of the variable X :

Number of boys	0	1	2	3	4	Total
No. of family	4	29	40	24	9	100

- Number of boys in a family – random **variable**

Probability distribution

Number of boys	0	1	2	3	4	Total
No. of family	4	29	40	24	9	100
Probability	**0.04**	**0.29**	**0.40**	**0.24**	**0.09**	**1**

Probability distribution

We call **probability distribution** of variable X the number of times occur the possible values of the variable X divide by the total

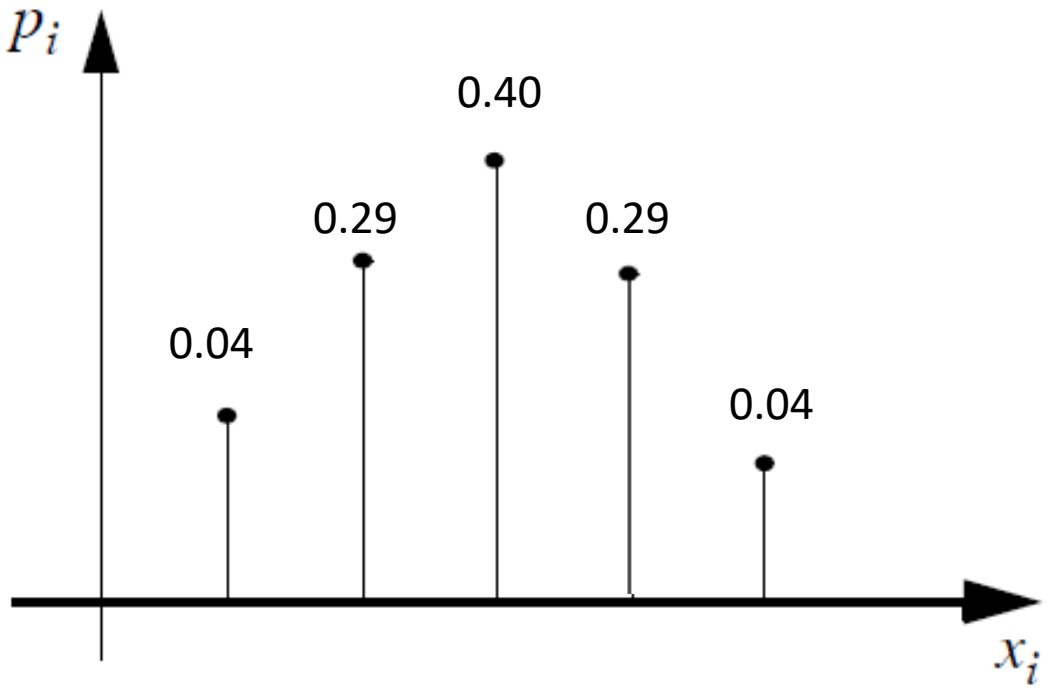
Number of boys	0	1	2	3	4	Total
Probability	0.04	0.29	0.40	0.24	0.09	1

Sum of probabilities in a probability distribution is 1

- Number of boys in a family – random **variab**

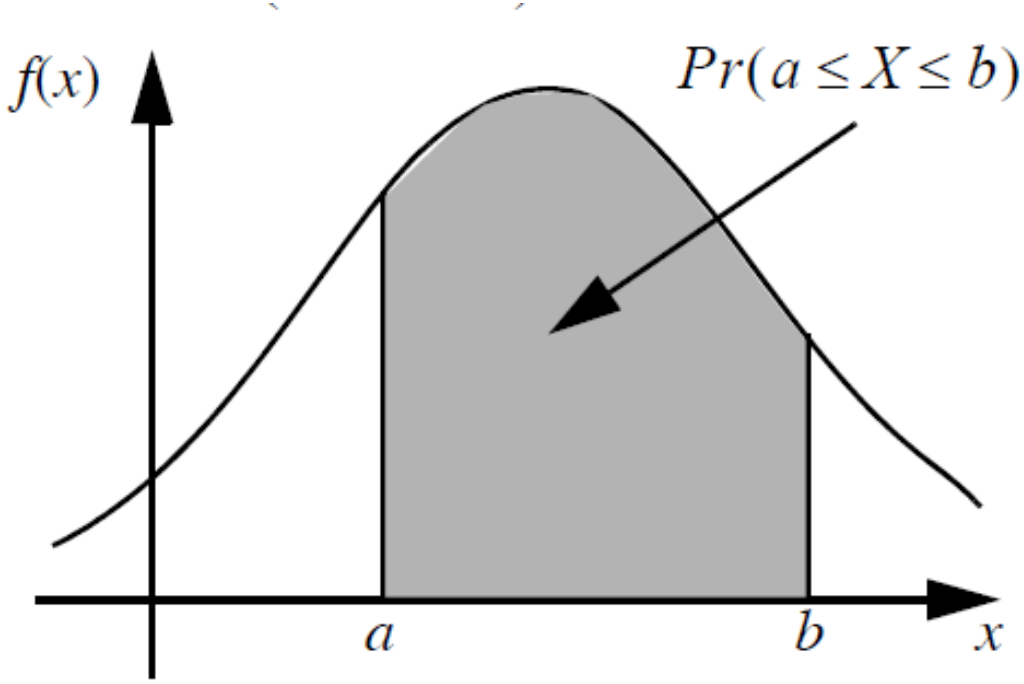
Random variable

Finite



Infinite

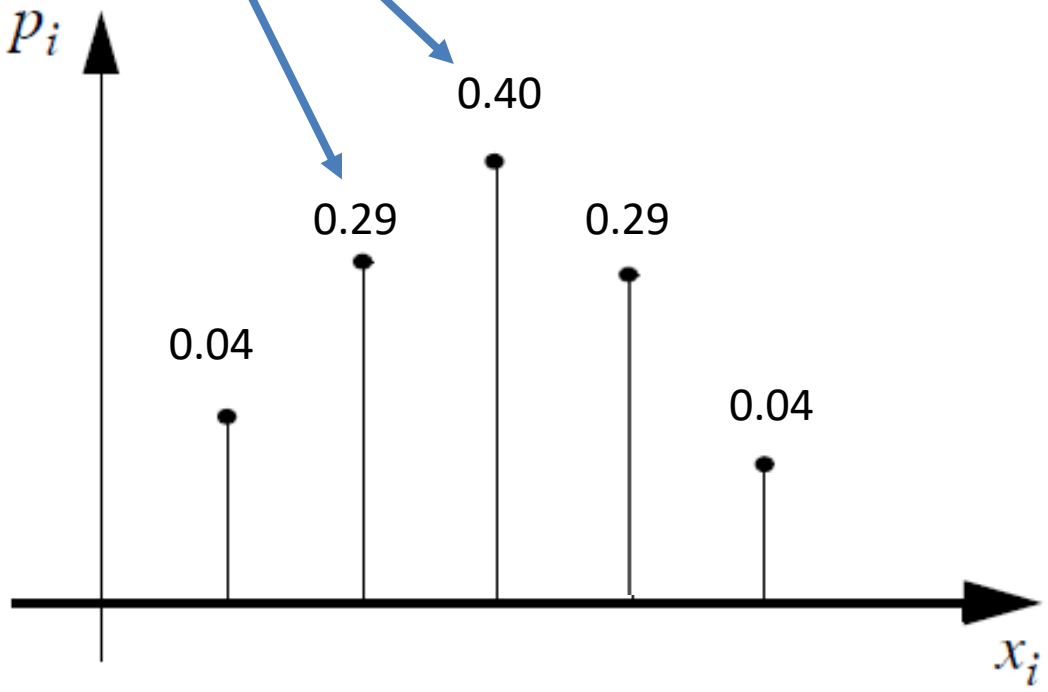
Area under the curve



Random variable

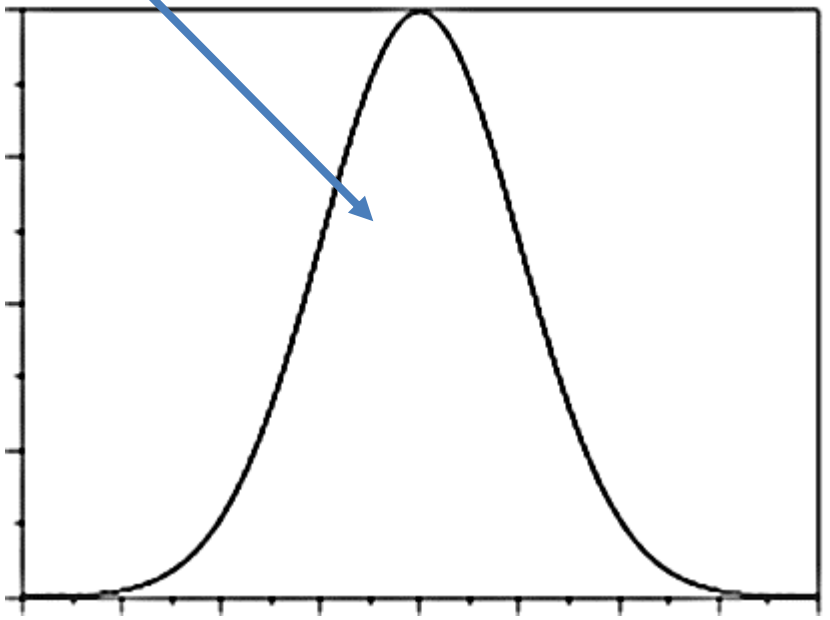
Finite

Sum of probabilities =1



Infinite

Area under the curve =1



How we compute probability distribution?

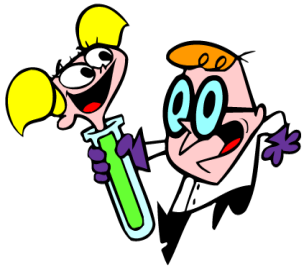
If the variable is finit

- Empirical – experiment – probability distribution

If is ∞?

! we are lucky – we found

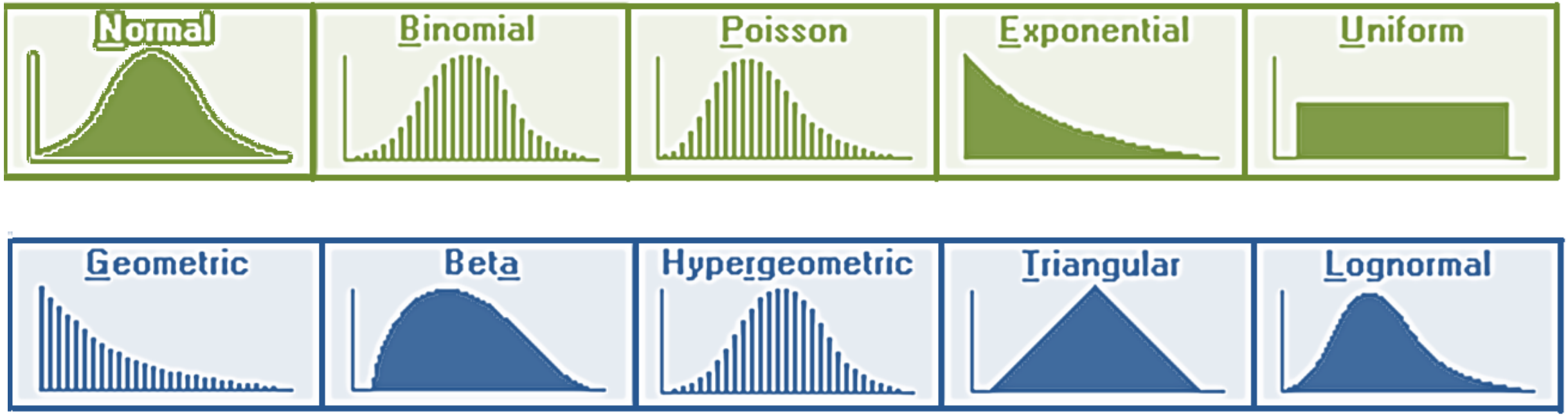
- Formula
- Rule



If we are not lucky:

- We model (approximate) after a theoretical probability distribution (a known one – one we were lucky with)

Known distribution law of probability



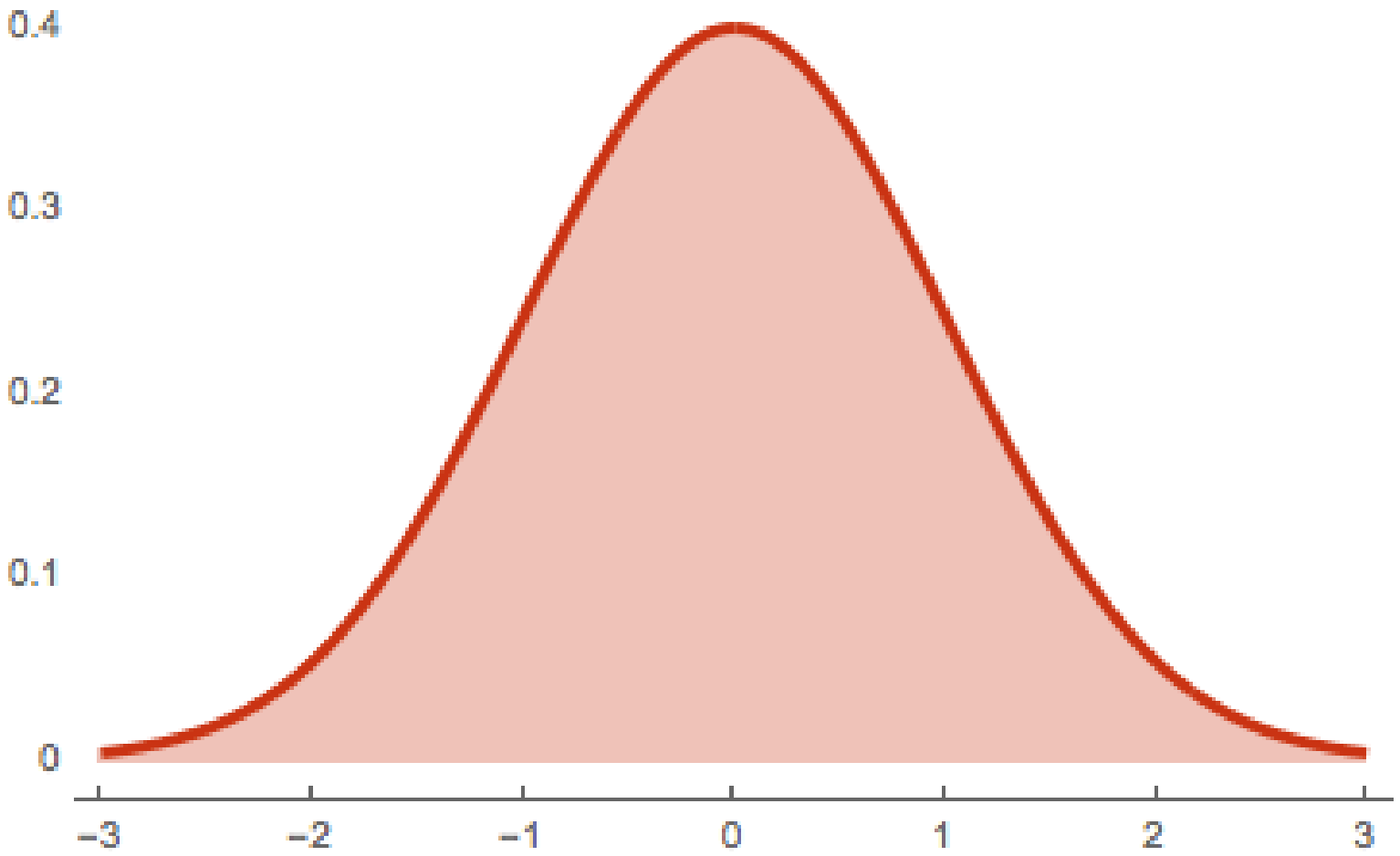
Probability distributions

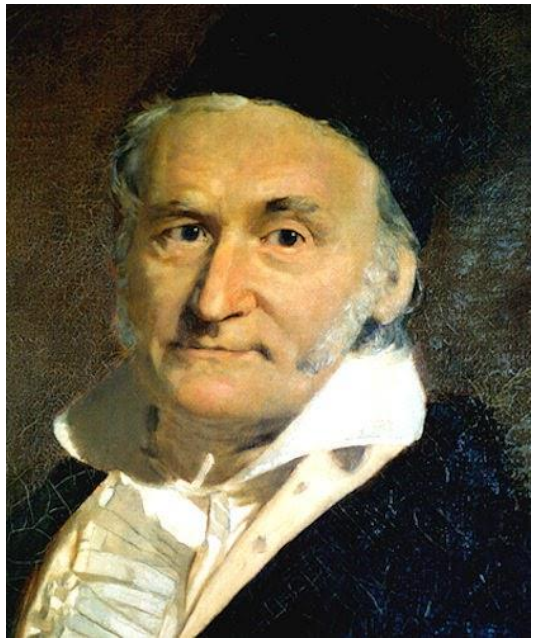
- **Binomial distribution** is used to determine **the probability of yes/no events**—the number of times a given outcome occurs in a given number of attempts.
- **Poisson distribution** is used to determine **the probability of rare events**.
- **Normal distribution** is used to find the probability that an outcome occurs when we measure a numerical observations - have a bell-shaped distribution.

Normal distribution

Normal distribution (bell shape, Gauss)

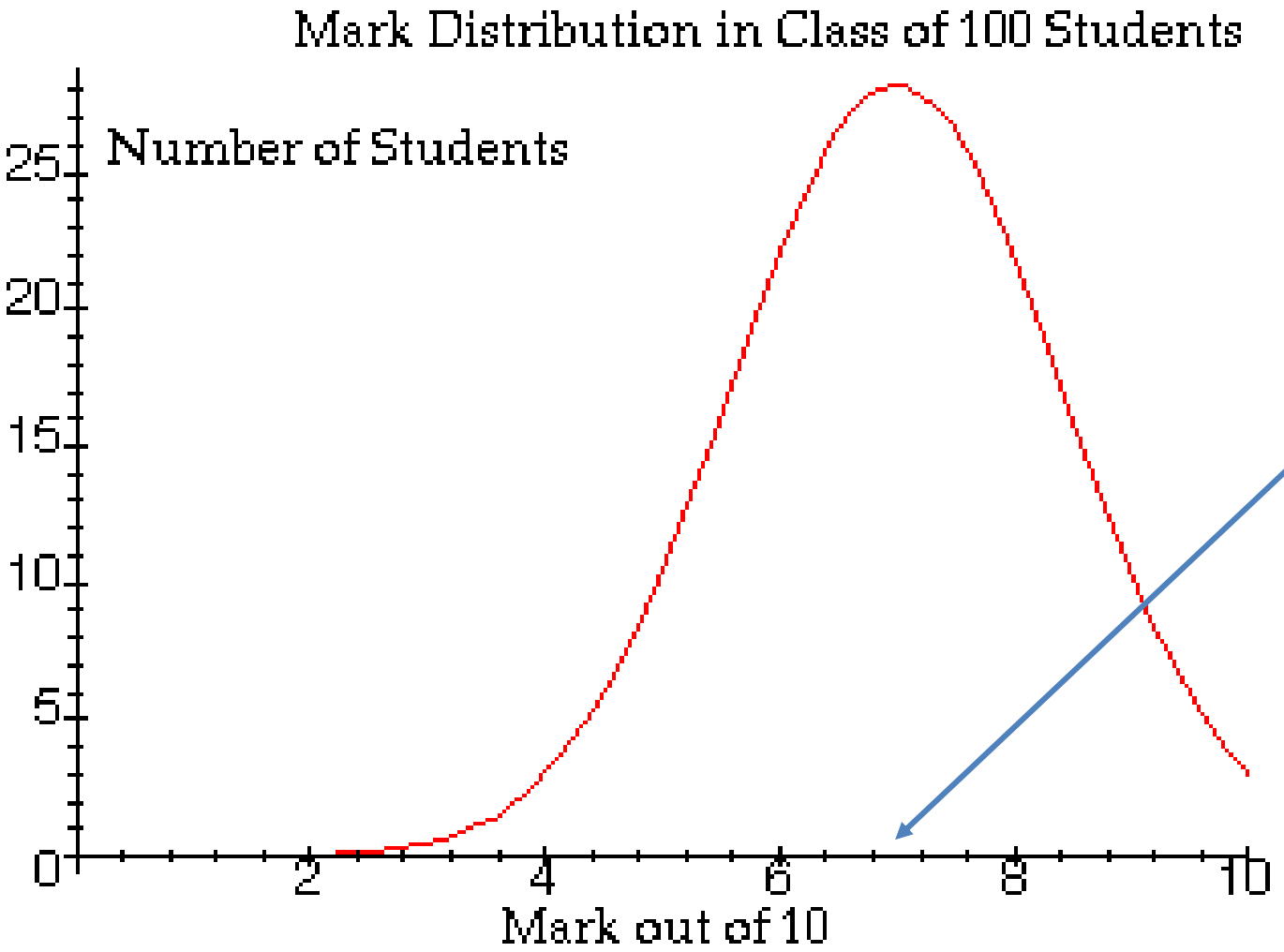
It is “normal” to happen like this





Carl Friedrich Gauss
(30 April 1777 – 23 February 1855)

The normality – most of the student get 7 (the middle of the mark between 4 to 10)



less student get the extrem values 4 and 10

Normal distribution

Random variable **X is normal N(μ, σ)** if the distribution depend on two parameters: mean μ and standard deviation **σ**

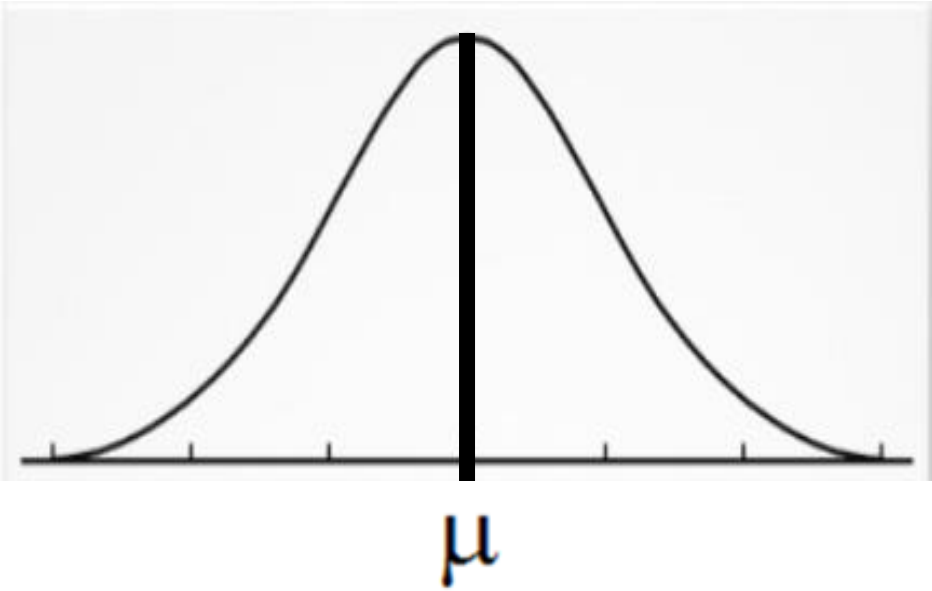
Formula:

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

we can check if a series is normally distributed

Rule: A serie of numbers is normal distributed if

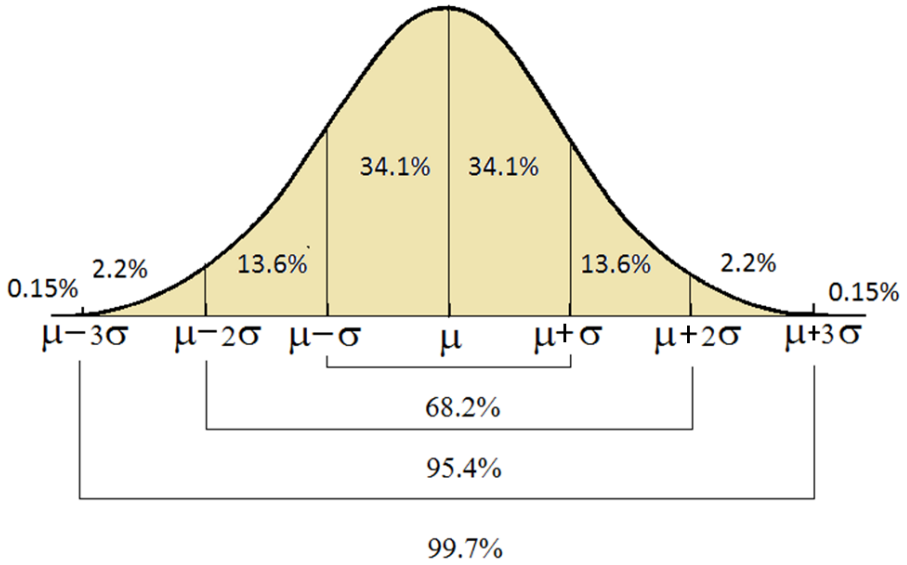
- ✓ Arithmetic mean = Median = Mode (or near equal)
- ✓ Quartile 1, Quartile 3 are simetrical with the mean (or near simetrical)
- ✓ Skewness ≈ 0 (between -1 to 1)
- ✓ Kurtosis ≈ 0 (between -1 to 1)



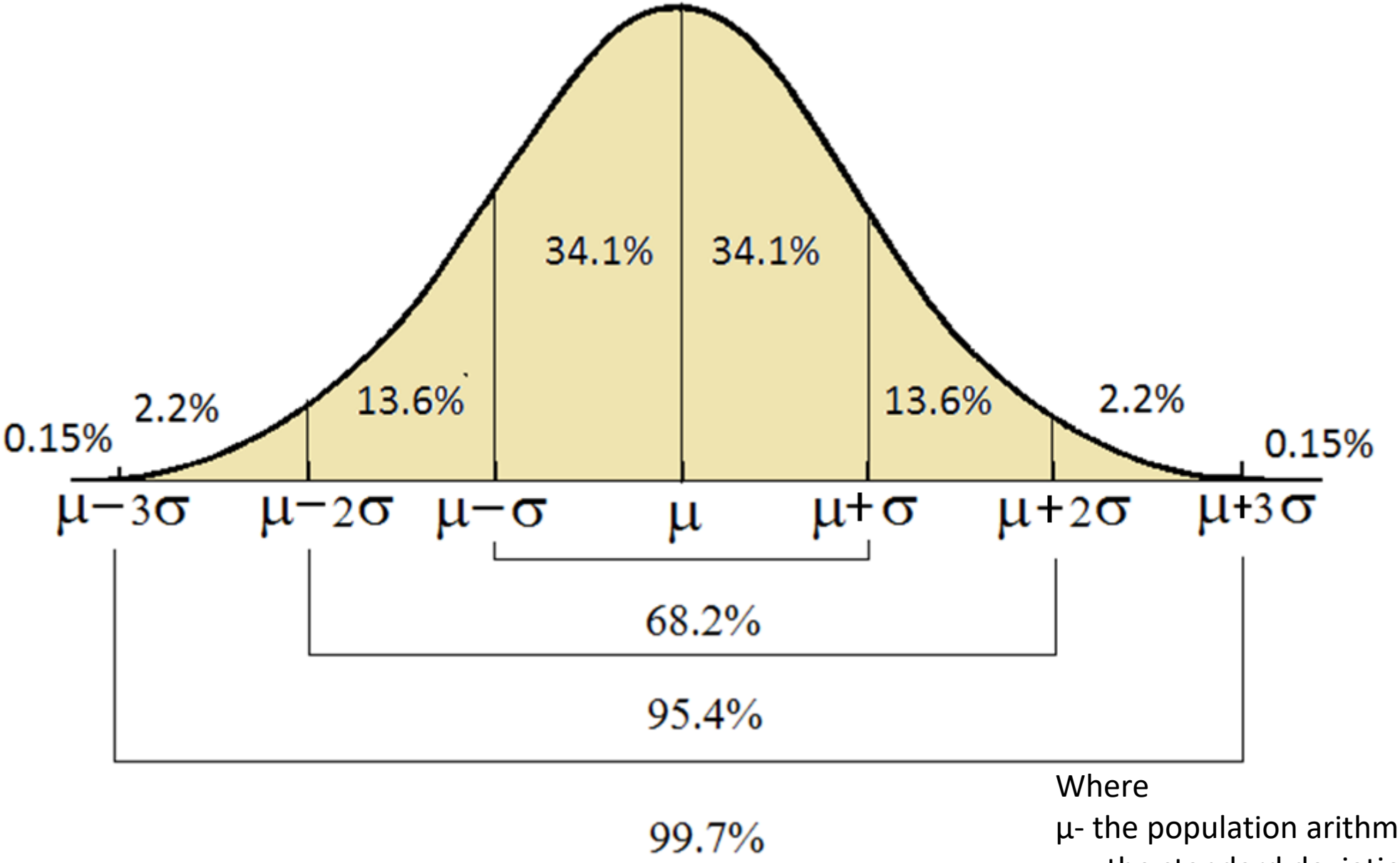
Other properties of normal distribution

- ✓ In the interval: mean ± st.dev. there are minimum 68.2% of data
- ✓ In the interval: mean ±2* st.dev. there are minimum 95.4% of data
- ✓ In the interval: mean ±3* st.dev. there are minimum 99.7% of data

Where
st.dev. – standard deviation
mean – arithmetic mean
mean ± st.dev. – the interval between mean-st.dev and mean+st.dev.
μ- the population arithmetic mean
σ – the standard deviation



Normal distribution

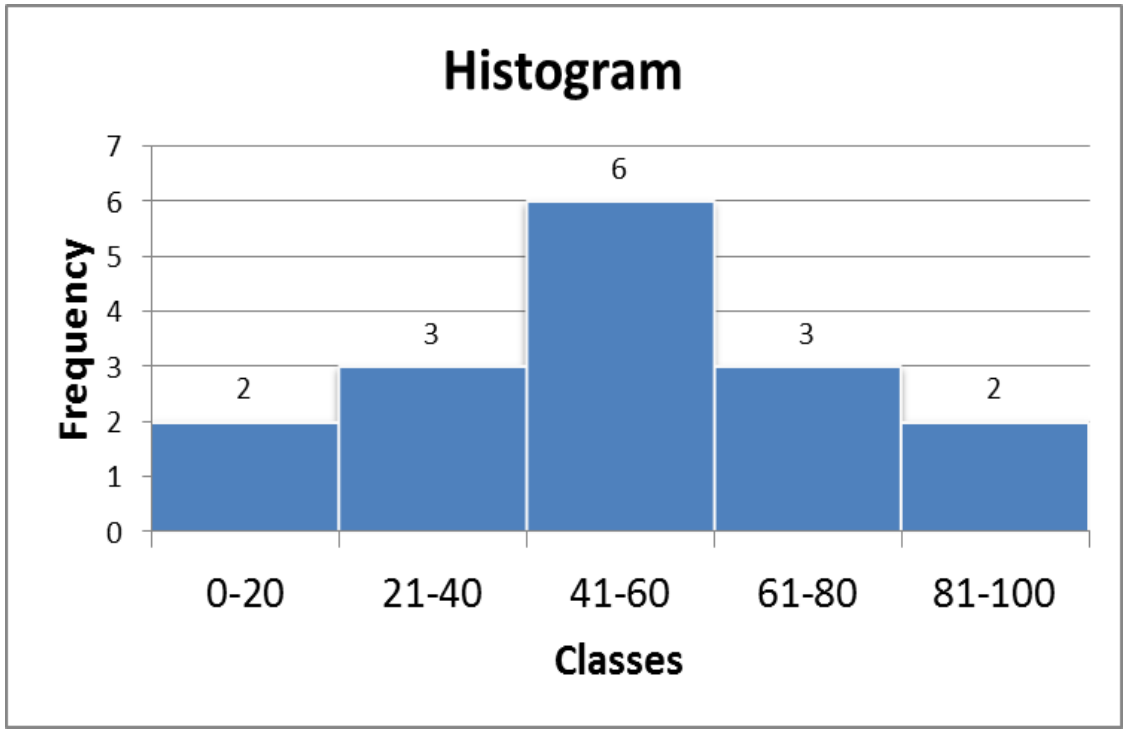


Where
μ- the population arithmetic mean
σ – the standard deviation

Series 3

1
11
24
29
36
41
45
49
51
55
59
64
71
76
88
100

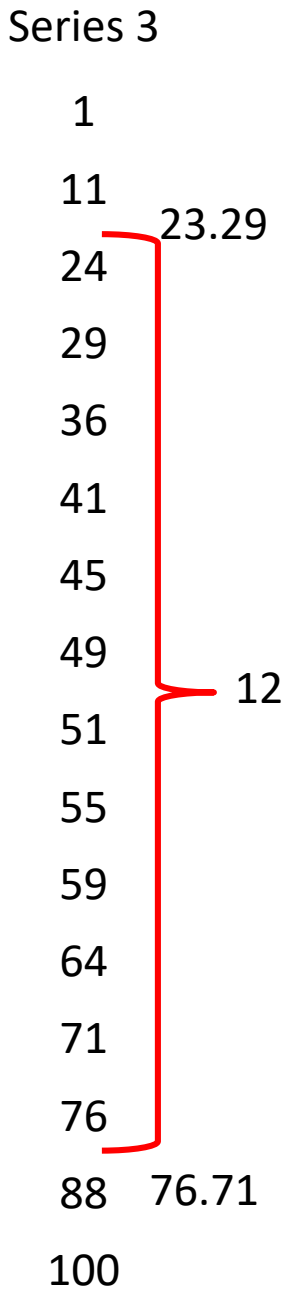
- Arithmetic mean = 50
- Median = 50
- Mode = there is no
- Standard deviation = 26,71
- Quartile 1 = 34,25
- Quartile 3 = 67,75
- Skewness = 0,01
- Kurtosis = -0,23
- arithmetic mean = median ✓
- skewness between -1 to 1 ✓
- kurtosis between -1 to 1 ✓
- quartiles are symmetrical to the median ✓



From the histogram:

- symetrical, ✓
- arithmetic mean = median ✓
- quartiles are symmetrical to the median ✓

Conclusion: The Series 3 is normal distributed from this points of view



For normal distribution in the interval

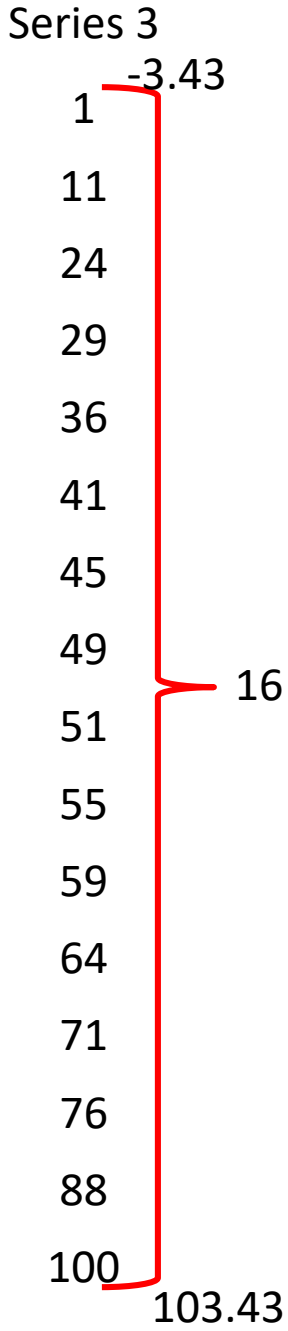
Mean±st.dev should be minimum 68.2% of data

mean = arithmetic mean

st.dev. = standard deviation

- Arithmetic mean = 50 Standard deviation = 26.71
- mean±st.dev = [50 - 26.71; 50 + 26.71]= [23.29; 76.71]
- in the interval [23.29; 76.71] are 12 values
- 12 values from 16 = 12/16*100 = **75%** of data
- in the interval [23.29; 76.71] are 75% of data

Conclusion: The Series 3 is normal distributed from this point of view



For normal distribution in the interval

Mean± 2*st.dev should be minimum 95.4% of data ✓

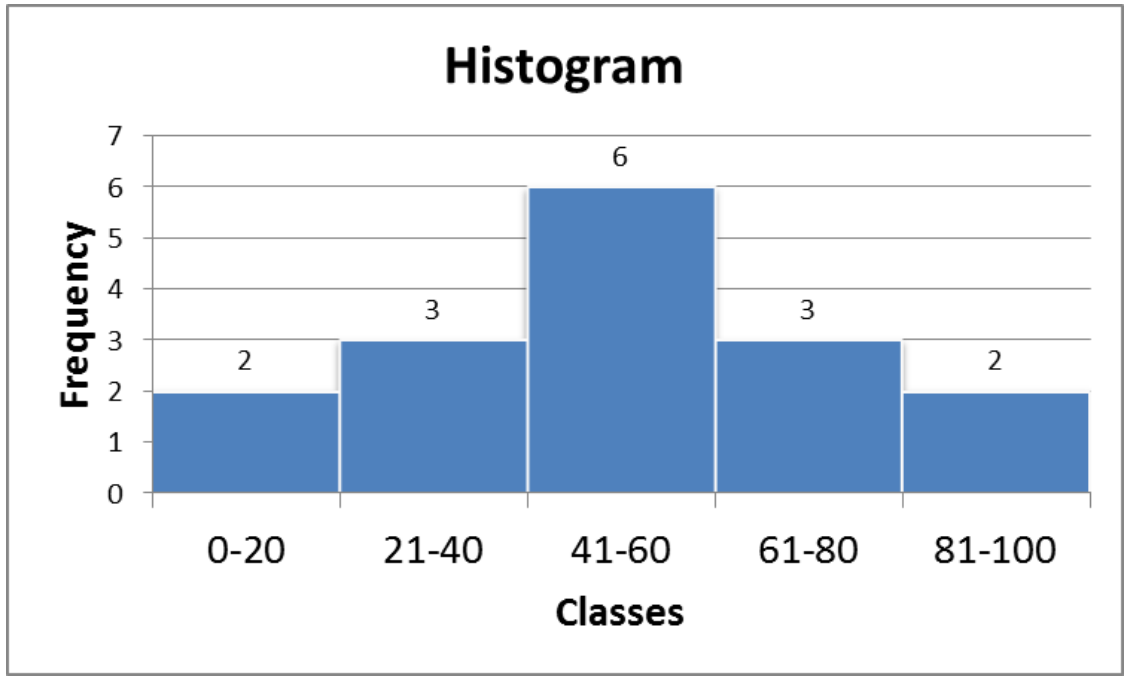
Mean± 3*st.dev should be minimum 99.7% of data ✓

- Arithmetic mean = 50 Standard deviation = 26.71
- In the interval mean ± 2*st.dev = [50 - 2 * 26.71; 50 + 2 * 26.71] = [-3.43; 103.43] are 16 values (16 from 16) i.e. **100%** of data
- In the interval mean ± 3*st.dev = [50 -3 * 26.71; 50 + 3 * 26.71] = [-30.15; 130.15] are 16 values, i.e. **100%** of data

Conclusion: The Series 3 is normal distributed from this points of view

Series 3

1
11
23.29
24
29
36
41
45
49
12
51
55
59
64
71
76
88
76.71
100



For normal distribution:

Minimum 68.3% of data

Minimum 95.4% of data

Minimum 99.7% of data

This serie is normal distributed

- Arithmetic mean = 50
- Standard deviation = 26.71
- Mean±st.dev= [50-26.71;50+26.71]= [23.29;76.71] are 12 values **75%** of data
- Mean±2*st.dev = [-3.43;103.43] are **100%** of data
- Mean±3*st.dev = [-30.15;130.15] are **100%** of data

The Series 3 is normal distributed

Example 2

Series 1

1
1
2
3
5
6
6
7
93
94
94
95
97
98
98
100

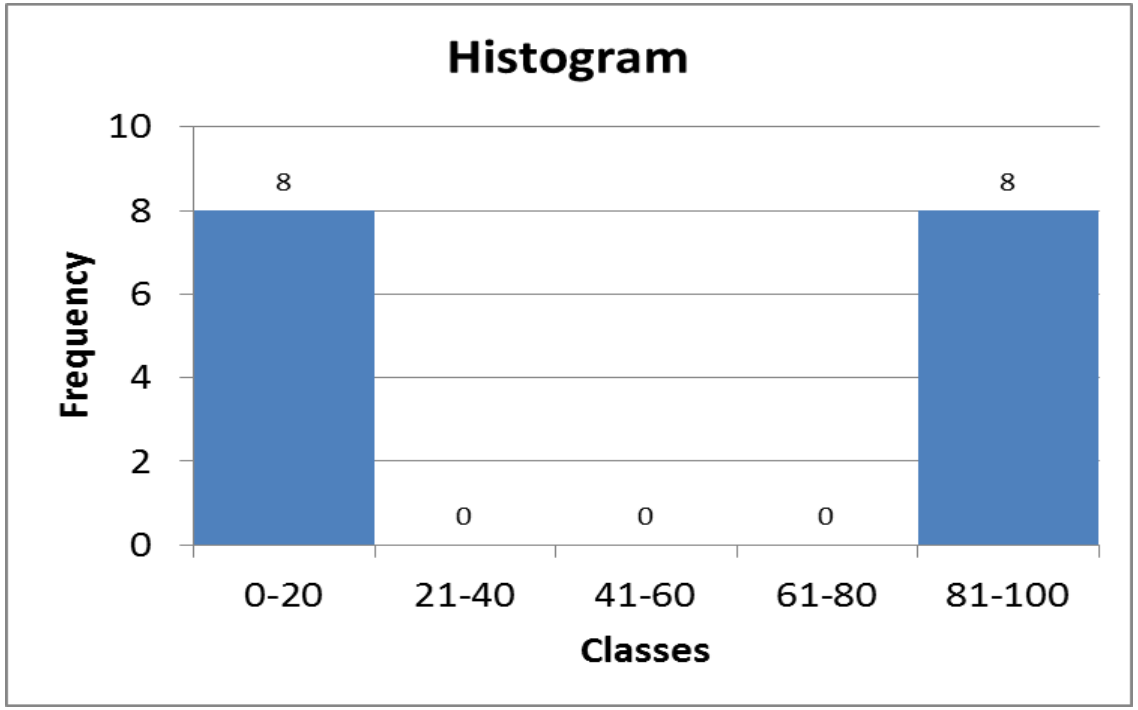
- Arithmetic mean = 50
- Median=50
- Mode – there is no
- Standard deviation = 47.70
- Quartile 1 = 4.5
- Cuartile 3 = 95.5
- Skewness =0,0002
- Kurtosis = -2,29

This series is not normal distributed

Series 1

1
1
2
3
5
6
6
7
93
94
94
95
97
98
98
100

- Arithmetic mean = 50
- Median=50
- Mode – there is no
- Standard deviation = 47.70
- Quartile 1 = 4.5
- Cuartile 3 = 95.5
- Skewness =0,0002
- Kurtosis = -2,29
- arithmetic mean = median ✓
- skewness between -1 to 1 ✓
- kurtosis between -1 to 1 X
- quartiles are symmetrical to the median ✓



From the histogram:

- symetrical ✓
- arithmetic mean = median ✓
- quartiles are symmetrical to the median ✓

Conclusion: The Series 1 is not normal distributed

Series 1

1
1
2
3
5
6
6
7
93
94
94
95
97
97.70
98
98
100

2.30

10

- Arithmetic mean = 50
- Standard deviation = 47.70
- Mean±st.dev =[50 - 47.7; 50 + 47.7] =[2.30; 97.70]

Series 1

1
1
2
3
5
6
6
7
93
94
94
95
97
98
98
100

2.30

10

97.70

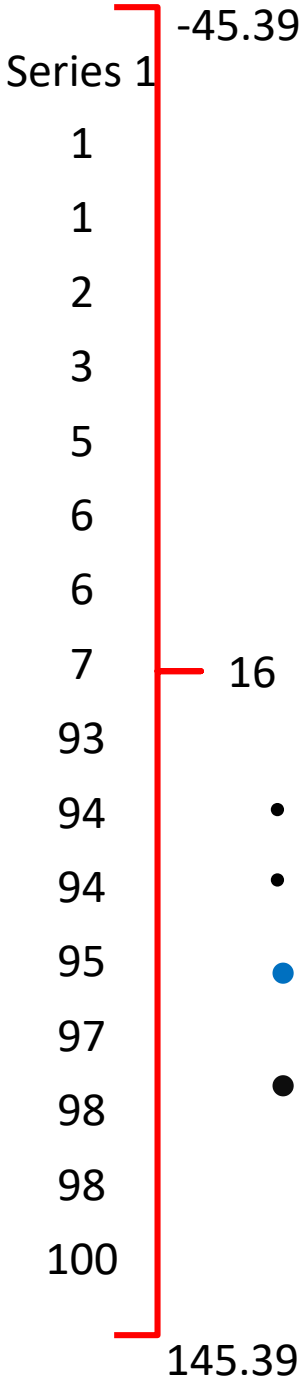
For normal distribution:

First vicinity of the mean - Minimum 68.2% of data

Should be 68.3 or higher to be normal distribution

- Arithmetic mean = 50 Standard deviation = 47.70
- Mean±st.dev =[50-47.7;50+47.7] =[2.30;97.70] are 10 values, e.g. 10/16=**62.5%** of data X

Conclusion: The Series 1 is not normal distributed



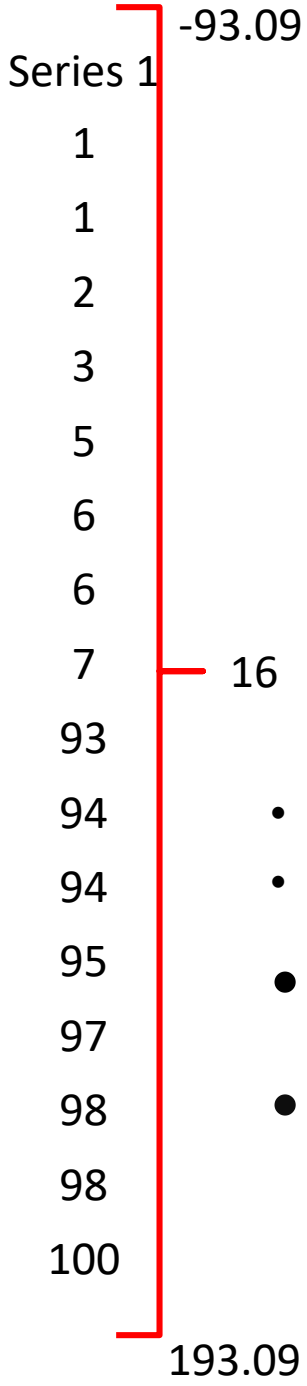
For normal distribution:

Minimum 68.2% of data

Second vicinity of the mean - Minimum 95.4% of data

Minimum 99.7% of data

- Arithmetic mean = 50
- Standard deviation = 47.70
- Second vicinity of the mean
- Mean±2*st.dev = [50-2*47.7;50+2*47.7] = [-45.39;145.39]
  - are 16 values, e.g. 16/16=100% of data ✓



For normal distribution:

Minimum 68.2% of data

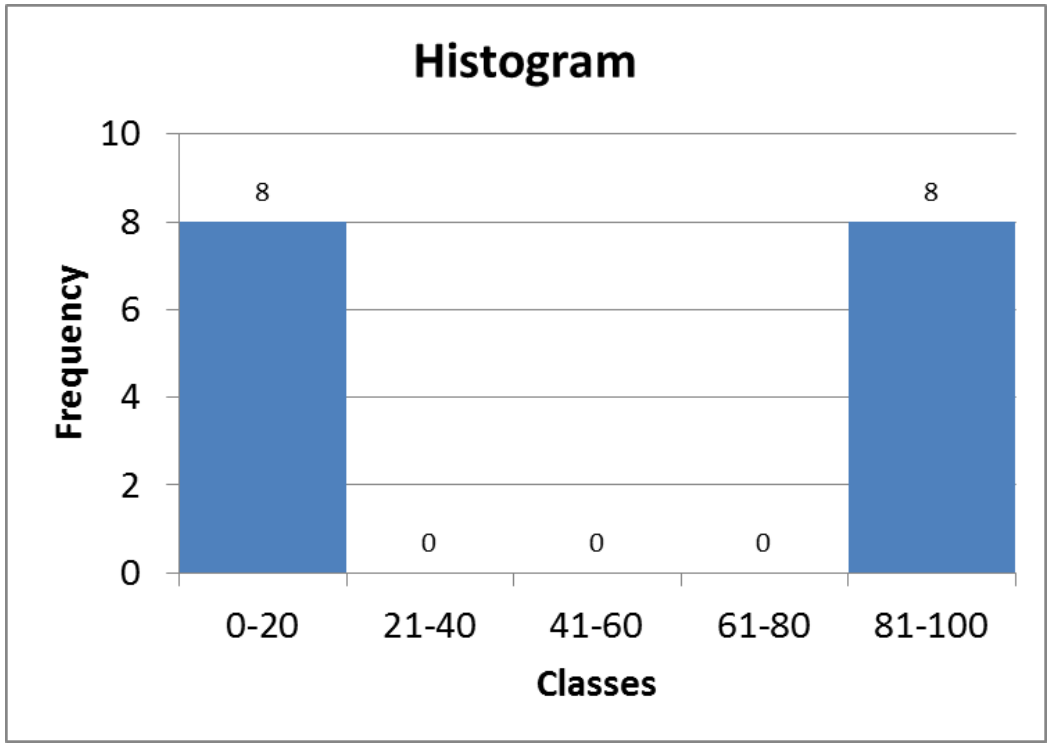
Minimum 95.4% of data

Third vicinity of the mean - Minimum 99.7% of data

- Arithmetic mean = 50
- Standard deviation = 47.70
- Third vicinity of the mean
- Mean±3*st.dev = [50-3*47.7;50+3*47.7] = [-93.09;193.09]
  - 16 values, e.g. 16/16=100% of data ✓

Series 1

1
1
2
3
5
6
6
7
93
94
94
95
97
98
98
100



For normal distribution:

Minimum 68.2% of data

Minimum 95.4% of data

Minimum 99.7% of data

- Arithmetic mean = 50
- Standard deviation = 47.70

- Mean±st.dev =[50-47.7;50+47.7] =[2.30;97.70] are 10 values, e.g. 10/16=**62.5%** of data
- Mean±2*st.dev = [50-2*47.7;50+2*47.7] = [-45.39;145.39] are 16 values, e.g. 16/16=100% of data
- Mean±3*st.dev = [50-3*47.7;50+3*47.7] =[-93.09;193.09] are 16 values, e.g. 16/16=100% of data

Conclusion: The Series 1 is not normal distributed

Example 3

Series 2

1

44

45

46

48

48

49

50

50

51

52

52

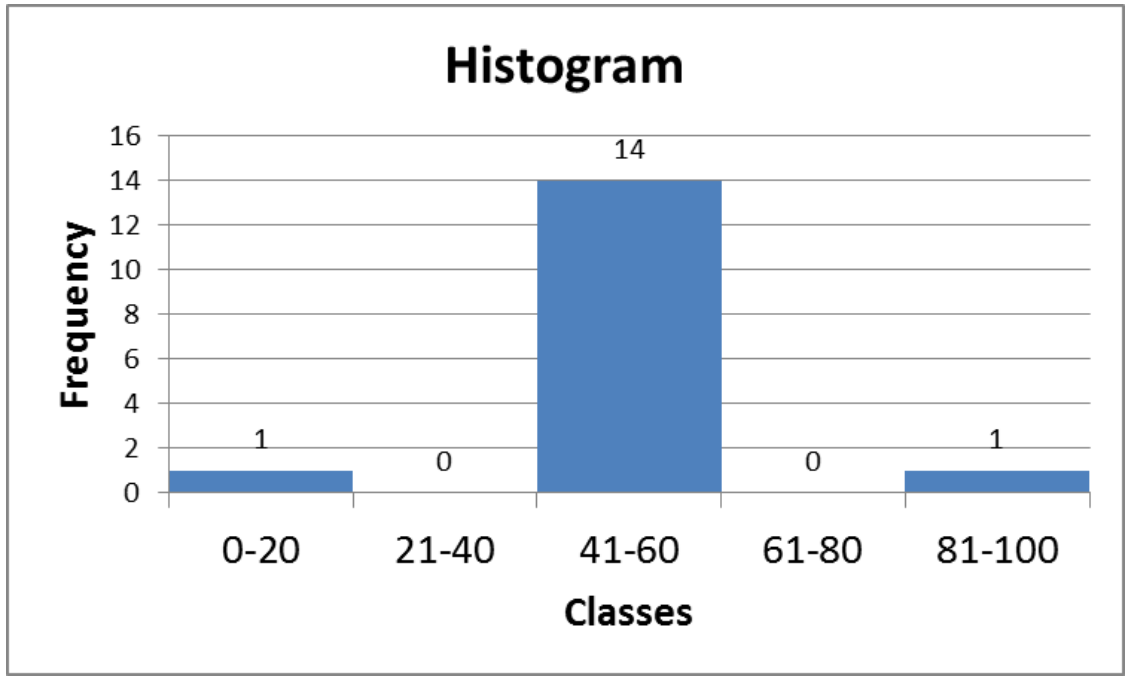
54

55

55

100

- Arithmetic mean = 50
- Median=50
- Mode = 50
- Standard deviation = 18,37
- Quartile 1 = 47.5
- Cuartile 3 = 52.5
- Skewness =0,09
- Kurtosis = 6,81
- arithmetic mean = median ✓
- skewness between -1 to 1 ✓
- kurtosis between -1 to 1 X
- quartiles are symmetrical to the median ✓



From the histogram:

- symetrical ✓
- arithmetic mean = median ✓
- quartiles are symmetrical to the median ✓

Conclusion: The Series 2 is not normal distributed

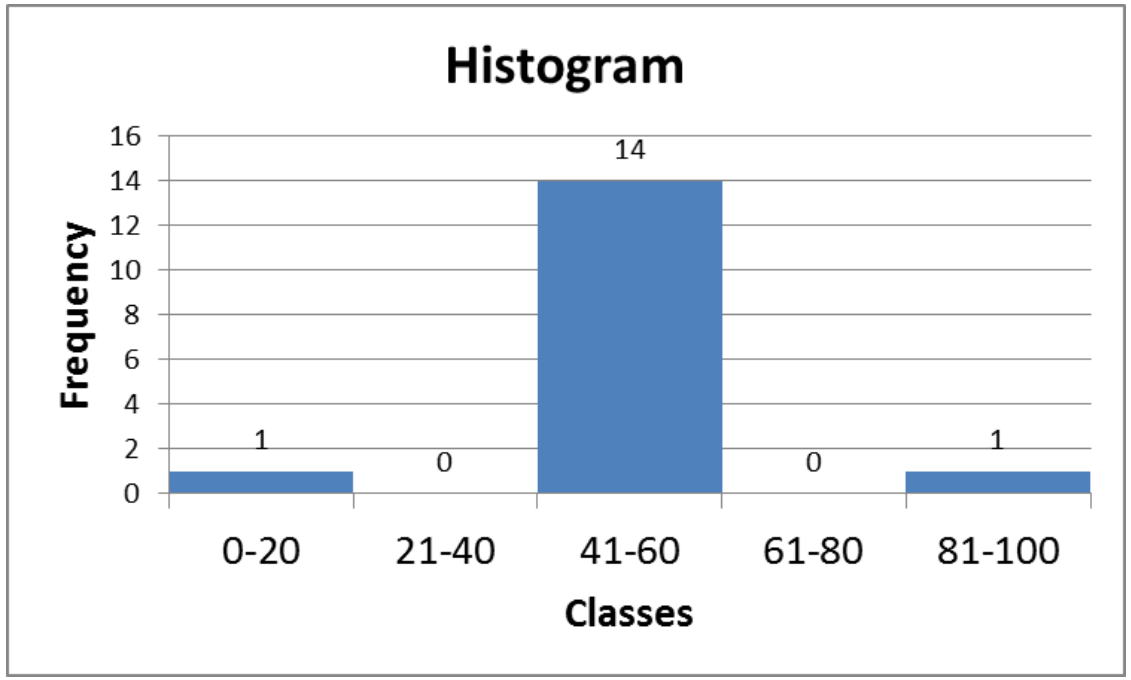
Series 2

1
44
45
46
48
48
49
50
50
51
52
52
54
55
55
100

31.63

14

68.37



For normal distribution:

Minimum 68.2% of data

Minimum 95.4% of data

Minimum 99.7% of data

- Arithmetic mean = 50
- Standard deviation = 18.37
- Mean±st.dev = [50-18.37;50+18.37] = [31.63;68.37]
  14 values, e.g. 14/16=**87.5%** of data
- Mean±2*st.dev = [50-2*18.37;50+2*18.37] = [13.26;86.74]
  14 values, **87.5%** of data
- Mean±3*st.dev= [-5.11;105.11]
  - 16 values, **100%** of data

Conclusion: The Series 2 is not normal distributed

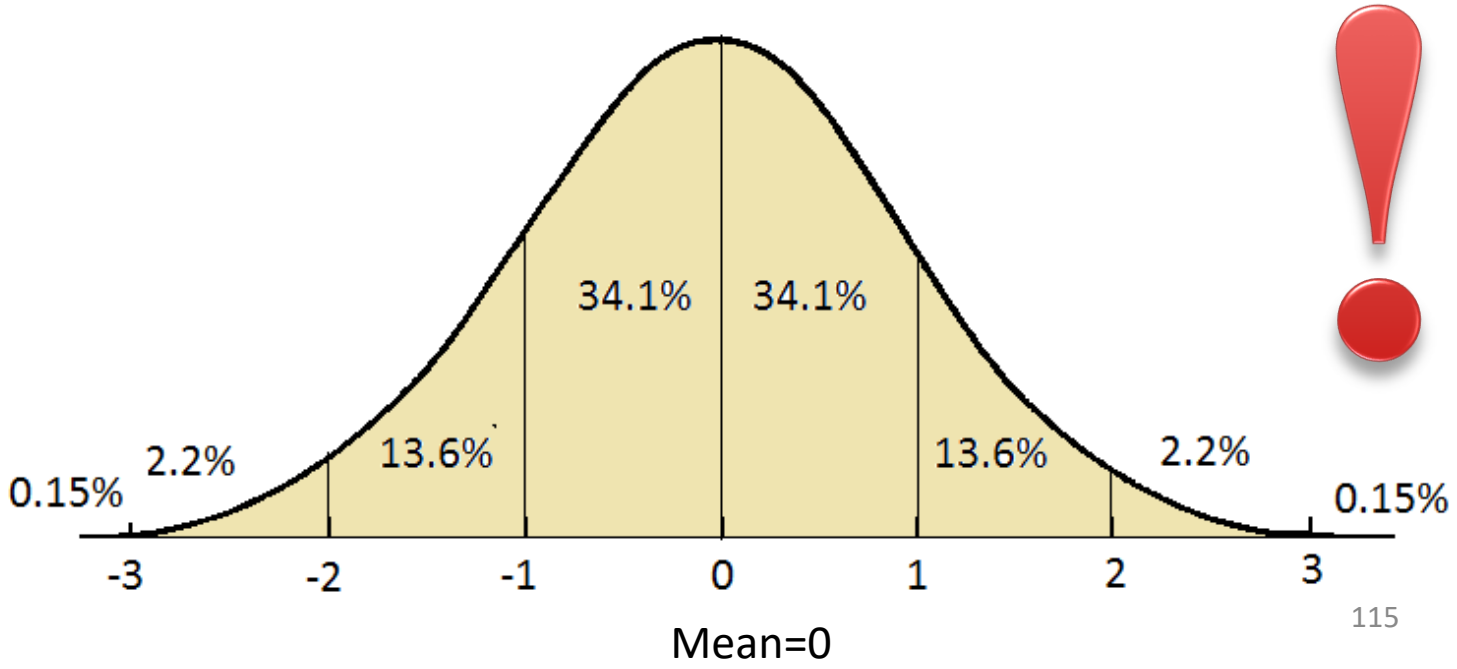
The Standard Normal Distribution

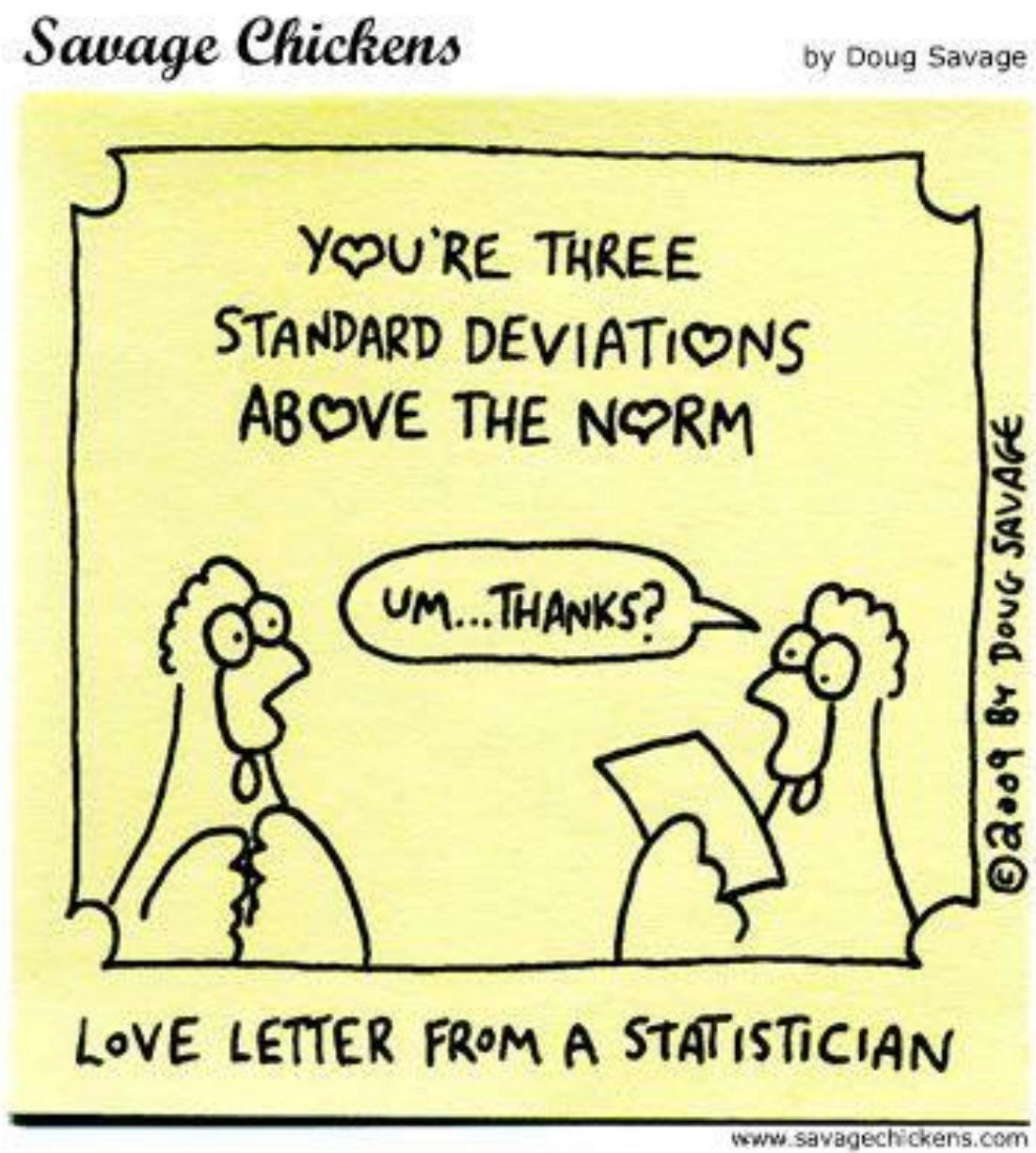
A normal distribution with $\mu=0$ and $\sigma=1$. We change the variable with:

$$Z = \frac{X - \mu}{\sigma}$$

Formula:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$




Savage Chickens
by Doug Savage
YOU'RE THREE STANDARD DEVIATIONS ABOVE THE NORM
UM...THANKS?
©2009 BY DOUG SAVAGE
LOVE LETTER FROM A STATISTICIAN
www.savagechickens.com

*DBP is normal distributed with μ=80 and σ=5 mmHg. Which value of DBP divide the area under the curve in 5% and 95% in the sample distribution of the mean? (Z_{α}=-1.645)

A. 88.225

B. 71.775

C. 86.333

D. 80

E. 5

F. -1.645

*DBP is normal distributed with μ=80 and σ=5 mmHg. Which value of DBP divide the area under the curve in 5% and 95% in the sample distribution of the mean? (Z_{α}=-1.645)

A. 88.225

B. 71.775

C. 86.333

D. 80

E. 5

F. -1.645

- Thank you !!!