First look at the data

Introduction to Data Science

6th lecture

Izv. prof. dr. sc. Ana Sović Kržić

2025/2026

Contents

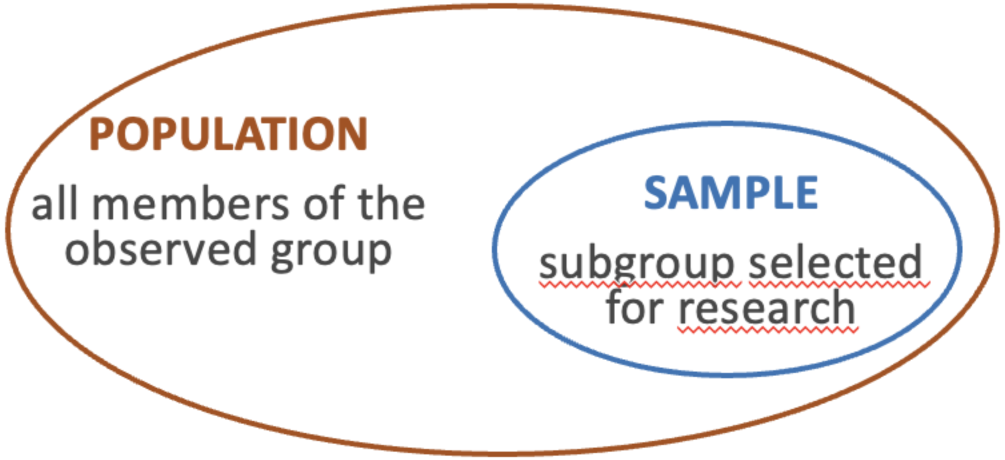
- Sample and population
- Measurament scale
- Descriptive statistics
- Inferential statistics

Sample and population

- Sample and population
- Fraction of the sample
- Selection of a random sample
- Types of random samples
- Types of non-random samples
- Sample size

Sample and population

- **statistical unit** (element) – unit on which measurement is performed (e.g. person, group of people, class, product, factory...)
- **sample** – a smaller number of defined statistical units or elements that make a larger whole (population)
    - represents the nature of the population in all important features with respect to measurement
- **population** = all statistical units



Sampling fraction

- the population has **N statistical units** or elements of the population
- we randomly select **n statistical units into the sample**
- **sampling fraction**:

$$f = \frac{n}{N}$$

Example: the population has 5000 elements, we randomly select 150 of them in the sample, the sampling fraction is f = 150/5000 = 0.03 = 3%

Types of samples

- **random sample** – has a normal distribution (theoretically)
- **biased sample** – when a sample has a higher chance of being selected
- we can use a **non-random sample**, but we cannot calculate the error in relation to the entire population, it is used for practical reasons (e.g. students, patients, voluntary research participants)

And this is the last you will read

You will read this first

Then you will read this

Then this

Wrong random sample selection

- we often unconsciously prefer some numbers (e.g. 3 or 7) - so all numbers between 1 and 1000 do not have an equal chance
- take names from the list that catch your eye - possible influence of the length or familiarity of the name

Selection of a random sample

- by throwing a **10-sided "dice"**, the digits from the "dice" are recorded (random numbers obtained from the population of numbers 0 to 9, occurrence probability of each p = 0.1)
- "**table of random numbers**" - open the table at any place - read the numbers in order (by rows, columns, diagonals) in groups of digits as big as we need - if the number is too big or if a number that we already had appears again, it is skipped
- using a **computer** (check if the frequency of each individual number deviates from the theoretical frequency is made using the chi-square test (p=0.1 for each digit))

7766	7520	1607	6048	2771	4733	8558	8681	5204	3806
9627	5293	3539	0457	4426	2857	3666	9156	6931	6157
4594	2563	6826	8102	2543	4032	3897	2012	0945	0709
6668	4104	4018	4544	8117	7664	5270	3014	0420	4232
8874	0822	0949	8697	7550	4154	9697	9045	4916	1235
8009	5708	7072	8045	8451	5777	1613	0399	2069	7909
7271	5633	6025	0745	9804	3333	7160	5150	7743	5221
6450	6850	0602	9518	2275	9221	6441	8899	4640	7742
0598	0564	9655	3988	5620	3286	6319	6392	5743	1111
6546	4417	4453	5125	1356	6011	5965	9253	1486	7503
5806	6217	4278	3170	1626	1746	9731	9289	7667	5209
6901	9464	6302	6404	8049	3653	8101	4498	8558	6238
3625	0749	5025	7327	3984	1635	5963	0970	7357	2033
2222	9942	1706	2907	6304	8022	7972	7852	6242	6269
7224	3014	3943	5982	4052	4243	5306	1530	7537	3233
7160	6043	0767	0230	6082	3637	4556	5564	8972	9697
7965	7435	3397	9741	6207	2297	6491	7961	0243	6897
6708	0600	2765	1911	0813	2268	3554	7976	4102	0414
4159	6804	3838	4255	9664	7044	3067	6720	7416	4748
6592	1846	2269	9136	7107	0676	9782	8016	2715	3932
2805	7999	3743	1655	7812	7223	0954	4397	7427	9120
9501	0400	8056	4148	5585	7497	7421	0640	6695	6127
3346	6596	1997	9417	0164	9718	5671	9765	7091	1920
4447	3427	6134	9130	4763	2301	2892	4251	4491	5772
0610	4363	0705	0969	4684	4202	5274	6660	0468	1814
2131	4792	1418	0080	9763	7306	0167	9688	6959	2250
9569	9416	5681	9632	8505	8948	6475	2934	6046	9640
1412	7690	5615	1776	8568	7209	9907	3541	8847	8752
5064	7408	1951	1033	7817	2626	2441	3795	3275	1319
4193	2082	0412	5519	4108	3333	5546	0177	9345	5260
6414	5111	4003	3695	2976	4939	7555	7374	2913	2705
2672	8618	7005	5736	0172	7472	2033	6308	8779	1270
0758	3869	9288	2397	6264	8352	8617	7869	2459	8591
4502	2535	2434	5018	1202	9081	2674	2467	2532	9689
4823	3965	2801	6179	8592	6763	6567	1016	5801	9288
3011	0939	7162	4443	3849	9142	2922	9191	6029	7631
6611	9238	2160	9339	8177	2180	3905	2977	9234	3434
0378	8311	0623	4299	2335	7044	5855	0186	5895	5642
9905	4972	6907	5633	6548	3412	8469	0559	8878	8671
9424	4750	8325	3871	1831	7268	1863	9963	1905	7484
7004	3469	1159	4841	8681	8751	9214	1145	4394	1160
5658	2963	5798	4691	8653	7427	7826	9971	2622	9886
9327	2129	3459	1165	1011	4805	1821	7999	2136	9308
1161	2217	1797	3906	5304	4087	6766	3063	1747	3836
6002	3340	3648	3765	1565	8483	6353	8232	4942	5721
4311	3087	1756	6612	3277	1269	6573	3096	0898	1103
5237	1667	5941	2504	6213	5797	9326	3079	8796	4220
0163	7150	0894	9009	7858	4812	7678	0835	8447	1524
0437	7497	0187	4907	2202	2318	5339	3290	4342	9375
0974	9130	4974	9757	8802	8514	6564	5485	0793	5675

3754	7829	9473	8264	8502	0364	5146	0609	4708	5229
9278	1828	8171	8788	3821	0923	8249	8431	6516	0911
9152	6396	7516	2959	4988	0943	6070	8342	5643	7476
0306	8452	1326	8892	2571	4860	1907	4843	0248	5283
1775	3205	8496	0201	6864	3375	0599	7516	8592	9823
4448	1897	3406	1429	8153	3408	1136	9173	9582	2866
3406	4332	0083	1214	5107	0912	8257	4015	5933	5520
4869	7491	5786	3633	9450	4572	6046	7844	2536	9502
5042	6524	1138	4001	6957	7220	8715	5082	8909	2384
0371	1656	8756	3369	3347	3534	0519	7230	2516	2674
2969	0056	8199	9383	4840	4135	7713	6317	4188	8073
4680	0551	7807	9470	9460	2253	0146	6082	9037	1862
1979	1845	0247	4813	2052	2758	6032	8288	6840	2677
3463	7252	3753	1178	2766	3207	2332	8262	8499	4501
0698	8601	2945	6077	3785	4647	4226	8959	9006	0964
2709	2447	0580	3375	1775	2038	3797	5163	7845	9397
6014	1671	2362	2315	8297	3930	6686	5835	9464	0916
7219	3355	3933	9312	3808	7879	6254	7075	7818	0295
6900	7276	4131	5402	3263	4026	5185	2862	8450	7749
0652	9020	6533	5737	6390	8723	8240	6442	4775	6040
3559	8683	0358	0118	0825	3360	7913	1403	4016	0202
1133	5094	3564	9818	0188	6367	2887	5038	1039	1658
1066	2065	4018	9132	3343	6165	1351	1312	7876	8452
8099	2678	7288	1970	9523	4070	7258	7276	3138	6818
5599	5836	0212	7112	8857	5894	6647	1660	3518	5780
6204	6540	1791	3190	3727	4500	5370	5231	8629	6291
8288	1891	5014	8442	9712	3435	4570	9493	1563	9165
7590	9691	1601	6615	0848	2885	1863	5682	1666	3398
7162	9599	9286	2819	2867	6533	9931	9217	4987	7722
9948	6283	0839	4175	8654	2005	6128	1306	6879	3152
5187	9791	4301	8481	5699	2522	0394	1538	8492	1812
5330	8112	2323	3056	1282	0543	4135	5819	6172	1017
6454	8783	7254	5267	9809	9964	9835	1111	5988	8017
8771	0872	6538	9975	4349	4106	6047	9630	4211	3234
1804	3896	2518	5665	8766	7161	0755	0886	3256	3198
8109	0020	3347	9221	6511	7593	6133	6123	2128	2735
9371	0132	4794	3110	5357	7242	4790	8002	9268	9733
6062	6416	7311	1167	5131	9955	9738	6038	1119	4832
7072	3929	8902	8062	6898	5499	5278	3407	0544	8772
5867	5384	8700	8017	5235	4094	9441	2381	8478	0981
1390	8293	7525	7188	8218	0131	3543	1679	8610	5737
4974	9904	7964	6038	0910	9364	4842	3873	3495	5511
9086	9898	1529	8544	7800	8523	1353	3312	5255	3096
8786	4498	5476	6266	9636	1897	3924	7298	3764	0906
7215	2019	6780	1005	4812	0787	8463	3784	6072	0940
2701	2584	8904	7799	9877	9015	0310	9330	0037	8215
9830	7090	3878	7553	7460	2845	9183	6429	9249	0246
0008	1130	3811	1862	1670	6389	9179	8571	7621	2169
5338	0351	6437	6148	5015	6174	5761	4690	0799	3291
6508	4163	0794	5801	1272	2814	0989	1130	3918	8596

Example – random students

Problem: We want to randomly select 350 students out of 780 enrolled in the first year. Each student mark with a number.

Rješenje:

1. Squinting, the table is opened to a random page (e.g. by tossing a coin) and a number is chosen at random with the tip of the pencil
2. 3 digits are taken, for example in order:

   7766 7520 1607 6048 2771 4733 8558 8681 5204 3806

3. 827, 855, 886, 815 fall away because the numbers are too big
4. If a number appears again - we do not take it into account
5. We repeat until we gather 350 students

A random number in the computer

- **Physical methods**
  - Random atomic or subatomic physical phenomena (quantum mechanics)
  - Radioactive decay, thermal noise, noise in Zener diodes, clock drift, radio noise...
  - They may contain asymmetries or systematic deviations
- **Computational methods**
  - PRNG (pseudorandom number generator) algorithms - automatically creates a long sequence of numbers with good random properties (but after a while the sequences start repeating themselves or the memory is overloaded)
  - The random sequence is determined by a fixed number "seed"
- **Human-based methods**
  - By collecting different inputs from users and using that as a source of randomness

Types of random samples (1)

1. stratified or layered sampling

- the population is divided into "**subpopulations**" or "**layers**" i.e. "**stratums**" according to some characteristics and a random sample is taken from each group
- strata can be according to age, sex, social composition
- no statistical unit may be located in more than one stratum
- advantage: when it is necessary **to know about each stratum**, and not only about the population as a whole
- **sample size from each stratum**:
  - **proportional to the size** of the group in the entire population

- among 10,000 people there are 60% young, 30% middle-aged, 10% old
- the sample should consist of 60% young, 30% middle-aged and 10% old among 1000 people
- 600 + 300 + 100
- $f_{all} = \frac{1000}{10000} = 0.1, f_{young} = \frac{600}{6000} = 0.1, f_{middle} = \frac{300}{3000} = 0.1, f_{old} = \frac{100}{1000} = 0.1$

Types of random samples (2)

- **disproportionate sample**
  - if some stratum is very small – the sample fraction can be increased
  - ratios of the sizes of individual strata = the same ratio as the products of the standard deviation and the sample size in an individual stratum

- $N_{stratumA} = 1000, \quad SD_{stratumA} = 5$
- $N_{stratumB} = 100, \quad SD_{stratumB} = 20$
- the ratio of the product of the sample size and SD: $\frac{1000 \cdot 5}{100 \cdot 20} = \frac{5}{2}$
- optimal size ratio 5:2
- for n=70, 50 data from the first stratum and 20 data from the second
- (for a proportional sample: ratio 1000:100=10:1 → 64 data from the first stratum and 6 from the second)

Types of random samples (3)

- **arithmetic mean** = weighted arithmetic mean of the stratum
  - if it is a very large random sample from the population - the risk of miscalculation is lower

- $N_{stratumA} = 1000, \; M_{stratumA} = 100$
- $N_{stratumB} = 100, \quad M_{stratumB} = 80$
- $M_{com} = \frac{M_1 N_1 + M_2 N_2}{N_1 + N_2} = \frac{100 \cdot 1000 + 80 \cdot 100}{1000 + 100} = 98.18$
- wrong: $M_{com} = \frac{100+80}{2} = 90$

Types of random samples (3)

2. cluster random sample

- often in economic, political or market research
  - it is necessary to collect the opinions of the inhabitants of the city
  - the city is divided into blocks,
  - a certain number of these blocks is selected by chance
  - interviews with all residents of the selected blocks - one-stage cluster sample
  - two-stage, three-stage, multi-stage samples - choose only some residents by chance

3. systematic random sample

- according to the list of population statistical units
  - a list of people in a factory – which was made randomly
  - chooses one at random,
  - and after that every n-th sample is taken (e.g. alphabetically)

Types of non-random samples (1)

1. convenience sample

- a sample that is "**found at hand**", because we don't have another one: patients currently present, random passers-by on the street, voluntary participants - there is a danger that they are extremely unrepresentative
- samples that are **easiest to come by** – if there is no evidence to the contrary, we can use them as a random sample

2. a deliberate or purposeful sample

- which is taken for a specific goal or purpose
- eg shoppers in a shopping center about price satisfaction

3. modal sample

- variant of purposeful sample - the **most common or typical cases** ("typical" city resident)

4. sample of experts

- the sample includes **experts in a certain field**

Types of non-random samples (2)

5. quota sample

- **proportional quota sample**
  - survey participants are taken according to a predetermined quota (e.g. citizens' opinions on an issue) - to represent the main characteristics of the population by selecting a proportional part of each characteristic
  - the number of people from each individual stratum to be interviewed is selected in advance - these people are randomly selected on the road
- **non-proportional quota sample** – proportional representation in each of the characteristics of the population is not considered

6. sample heterogeneity

- when we want to include **all different opinions or views** when surveying
- when we don't want an average, but to **determine the differences**

7. "snowball" sample

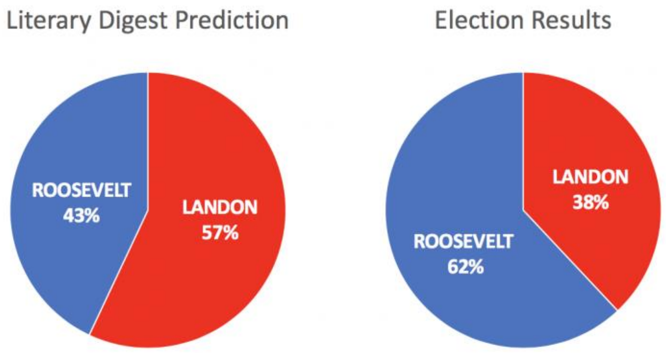
- sample members are collected based on the **recommendation of a previous member** who was included in the sample

Sample size

- the sample must be **representative** - the number is not so important
- it depends on the **variability** of the phenomenon we are measuring (small variability - few samples are needed), the **precision** with which we want to measure the phenomenon (we want less precision - fewer samples)
- **Weber method**: if we can roughly predict in what percentage a certain property is represented in the population
  - sample size = multiply that percentage by the missing percentage up to 100%
  - for example, 5% of the population has the characteristic we are measuring - the sample size should be 5×95=475
  - for 50% of the population – the sample size should be 50x50=2500
  - high variability – a larger number of samples

Example - elections

- **Research**: Literary Digest magazine conducted research on who would win the 1936 election: Landon or Roosevelt
- **2.4 million people** were **surveyed**
- **Survey result**: lost Roosevelt 43%
- **Elections**: Roosevelt won 62%
- **Explanation**: the survey was conducted by telephone. At that time only the rich could have a telephone.
- A biased sample, regardless of the number of respondents
- The Literary Digest soon failed



Variables and scales

- **Variable** = property of what we are studying

- Categorical variables – represent groups:
  - Nominal (and binary)
  - Ordinal

- Quantitative variables – represent values:
  - Continuous
  - Discreet

- **Scale of measurement** = describes the type of information recorded within the value of a variable

Scales of measurement (1)

1. Nominal scale

- data are categorized but without order between the categories
- a number is used instead of a name (e.g. numbers of players in sports)
- numbers are only for identification, i.e. marking the classes of identical units
- e.g. gender, mobile phone brand, city
- eg either a = b or a != b; if a = b, then b = a; if a = b, b = c then a = c
- possible to use: dominant value (mod, D), calculation of proportions, chi2 test, phi, Cramer's phi, contingency coefficient, visualizations (bar graph, pie)
- **Binary scale** – a special type of nominal scale (yes/no, tail/head, won/lost)

Scales of measurement (2)

2. Ordinal scale

- to indicate the order according to some quantitative or qualitative property according to the order of appearance, according to weight, size, liking... (eg ranking of arrival at the finish line in a race, numbers of houses in a street, Likert scale)
- the difference between individual categories (ranks) does not have to be equal or known, they only determine whether something is greater or lower than another
- eg if a > b then b < a; if a > b, and b > c then a > c
- possible to use: all operations with a constant (squaring, logarithmizing...), all for nominal scales, dominant value (mod), median, range, frequency distribution, correlation coefficient ro, statistical tests (Mood's median test, Mann-Whitney U test (Wilcoxon rank sum test), Wilcoxon matched-pairs signed-rank test, Kruskal–Wallis H test, Spearman's rho or rank correlation coefficient)

Scales of measurement (3)

3. Interval scale

- known order and difference between the numbers on the scale and these differences must correspond to real differences in the measured phenomenon
- quantitative scale but does not have an true zero point (eg a temperature of 100°F is not 2x hotter than 50°F, negative temperatures are also possible, 0°F doesn't mean an absence of temperature)
- equal distances between two values (intervals)
- eg (a - b) + (b - c) = a - c
- possible to use: addition, multiplication, all for nominal and ordinal scales, arithmetic mean, standard deviation, z-value, distributions, almost all tests (T-test, ANOVA, pearson's r, regression)

Scales of measurement (4)

4. Ratio scale

- all the properties of interval scales and also the property that equal numerical ratios also mean equal ratios in the measured phenomenon
- have absolute zero (e.g. length, weight, resistance - a weight of 90kg is 3x greater than a weight of 30kg, temperature in Kelvin - 20K is twice as hot as 10K and nothing can be below 0K)
- eg a : b = 3a : 3b, a : b = 7a : 7b
- possible to use: all for nominal, ordinal and interval scales, geometric mean, coefficient of variability, almost all tests (T-test, ANOVA, pearson's r, regression)

Scales of measurement (5)

scale	categori-zation	rank	equal intervals	true zero
nominal	+			
ordinal	+	+		
interval	+	+	+	
ratio	+	+	+	+

Determine which scale it is

Sample	Type of plant	Added salt (mg/l water)	Initial height (cm)	Current height (cm)	Withering (rank 0-10)	Survived (1 = yes, 0 = no)
1	A	0	12	26	7	1
2	A	100	13	24	8	1
3	A	250	11	25	9	0
4	B	0	25	33	2	1
5	B	100	26	35	4	1
6	B	250	25	34	3	1

Determine which scale it is

Postal code (10000, 10310, 21001)	Nominal
Knowledge of a foreign language (beginner, intermediate, advanced)	Ordinal
IQ test	Interval
Genre (comedy, drama, satire, tragedy)	Nominal
Number of people in the household	Ratio
This lecture is interesting (strongly disagree, disagree, neither agree nor disagree, agree, strongly agree)	Ordinal
Level of pain (painless, mild pain, moderate pain, considerable pain, maximum possible pain)	Ordinal
Pain level on a scale of 0 (no pain) – 10 (greatest possible pain)	Interval
Speed in km/h	Ratio
Number of bicycles in front of FER	Ratio

Descriptive statistics

- Characteristics of the concrete sample

Central tendency

- Measures of central tendency estimate the center (or average) of a data set

- Arithmetic mean M
- Common mean M_{com}
- Central value (median) C
- The dominant value (mode) D

Mean value M

$$M = \frac{sum\ of\ all\ values}{total\ number\ of\ responses} = \frac{1}{N}(X_1 + X_2 + \cdots + X_N)$$

- an indicator of the **true value of the measurement**
- Terms of use:
  1. the results must be true measured values, obtained at least on an interval scale
  2. all results must be obtained under **equal measurement conditions**
  3. a sufficient number of results is required, **at least 30**
  4. the distribution of the results must be normal (which also means symmetrical)

Common mean value

If mean values are **calculated from an equal number of results** (n_M number of mean value):

$$N_1 = N_2 = N_3 = \cdots = N_{n_M}$$

$$M_{com} = \frac{1}{n_M}(M_1 + M_2 + \cdots + M_n)$$

If mean values are **not calculated from an equal number of results:**

$$M_{com} = \frac{M_1 N_1 + M_2 N_2 + \cdots + M_n N_n}{N_1 + N_2 + \cdots + N_n}$$

Example – common mean value

- Some measurement was repeated 6 times on different groups of subjects

1. measurement	2. measurement	3. measurement	4. measurement	5. measurement	6. measurement
M_1 = 18.5 N_1 = 5	M_2 = 22.0 N_2 = 17	M_3 = 23.9 N_3 = 40	M_4 = 23.8 N_4 = 48	M_5 = 22.8 N_5 = 19	M_6 = 22.6 N_6 = 25

$$M_{com} = \frac{M_1 N_1 + M_2 N_2 + \cdots + M_6 N_6}{N_1 + N_2 + \cdots + N_6} = 23.1$$

- The same value would be obtained if each of the results were taken individually
- If we ignored the number of samples, we would get the wrong value :

$$M_{com} = \frac{1}{n_M}(M_1 + M_2 + \cdots + M_6) = 22.3$$

Central value (median) C

- the value that is **exactly in the middle** of the sequence of results sorted by size
- if the number of results is even, then the central value is calculated as the sum of the two middle results and divided by 2
- It is used if we have an **extremely large or small value** in the series of results or the **distribution of the results is asymmetrical**

Dominant value (mode) D

- the **most popular** or **most frequent** response value
- data set can have no mode, one mode, or more than one mode
- it is not affected by the number or value of the results, but only by the frequency of individual results

Example

Array: 1, 2, 4, 4, 4, 5, 6

M=3.71

C=4

D=4

Array: 1, 2, 4, 4, 4, 5, 60

M=11.43

C=4

D=4

Measures of variability

- a sense of how spread out the response values are

- Range R
- Mean deviation (dispersion) SO
- Variance *var*
- Standard deviation SD
- Interquartile range Q
- Coefficient of variability V

Range R

- how far apart the most extreme response values are

- **subtract the lowest value from the highest value**
- an outlier significantly increases the range without changing the grouping of results around the mean value
- it is usually higher if we have greater number of measurements - if only a few results are taken into account, the probability that there will be exactly the largest and smallest result among them is reduced
- often the min and max value is displayed instead of the range

Mean deviation (dispersion) MD

- **the average size of deviation of individual results**
- it can be calculated with the mean, central or dominant value
- a rough indicator of distinguishing the results from some environment

$$\boxed{MD = \frac{\sum|X - M|}{N}}$$

Variance *var*

- average sum of squared deviations
- reflects the degree of spread in the data set
- Variance of sample

$$\boxed{var = \frac{\sum(X - M)^2}{N - 1}}$$

- Variance of population

$$\boxed{var = \frac{\sum(X - M)^2}{N}}$$

Standard deviation SD

- average amount of variability in dataset
- the larger the standard deviation → the more variable the data set
- how far each score lies from the mean value → is mean a good or bad representative of the results?
- can be calculated only with the mean and is in the units as the measurements

- Standard deviation of sample

$$\boxed{SD = \sqrt{var} = \sqrt{\frac{\sum(X - M)^2}{N - 1}}}$$

- Standard deviation of population

$$\boxed{SD = \sqrt{var} = \sqrt{\frac{\sum(X - M)^2}{N}}}$$

- count control: range / standard deviation is almost always between 2 and 6.5

Interquartile range Q

- the series of obtained results is ordered by size, from the smallest to the largest - the series has 4 quartiles - in each quartile there is 25% of the results
- **quartile limit values**: Q_1, Q_2 ($Q_2 = C$ median value divides the series into two parts), Q_3, Q_4 (upper limit)
- ordinal place of limit values: $R_{Q_1} = \frac{N}{4} + 0.5, R_{Q_3} = \frac{N}{4} \cdot 3 + 0.5$
- interquartile range Q – **half the difference between the limit values of the third and first quartiles**

$$\boxed{Q = \frac{Q_3 - Q_1}{2}}$$

Coefficient of variability (variation) V

- what percentage of the mean value is the value of the standard deviation

$$\boxed{V = \frac{SD}{M} \cdot 100}$$

- to be able to **compare the variability of different phenomena and properties** (e.g. which case is more favorable $M_1 = 100, SD_1 = 10$ ili $M_2 = 8, SD_2 = 2$ – the first case is more favorable)
- used when we want to know:
  - in which property a group varies more and in which less
  - which of the groups varies more and which less in the same property

Example – weight and height of a ten-year-old

Boys, height	Boys, weight	Girls, height	Girls, weight
$N_1 = 612$ $M_1 = 134.4$ cm $SD_1 = 6.06$ cm $V = \frac{6.06}{134.4} \cdot 100 = 4.51\%$	$N_2 = 612$ $M_2 = 29.2$ kg $SD_2 = 3.89$ kg $V = \frac{3.89}{29.2} \cdot 100 = 13.32\%$	$N_1 = 684$ $M_1 = 134.9$ cm $SD_1 = 6.43$ cm $V = \frac{6.43}{134.9} \cdot 100 = 4.77\%$	$N_2 = 684$ $M_2 = 29.7$ kg $SD_2 = 4.78$ kg $V = \frac{4.78}{29.7} \cdot 100 = 16.09\%$

$$V = \frac{SD}{M} \cdot 100$$

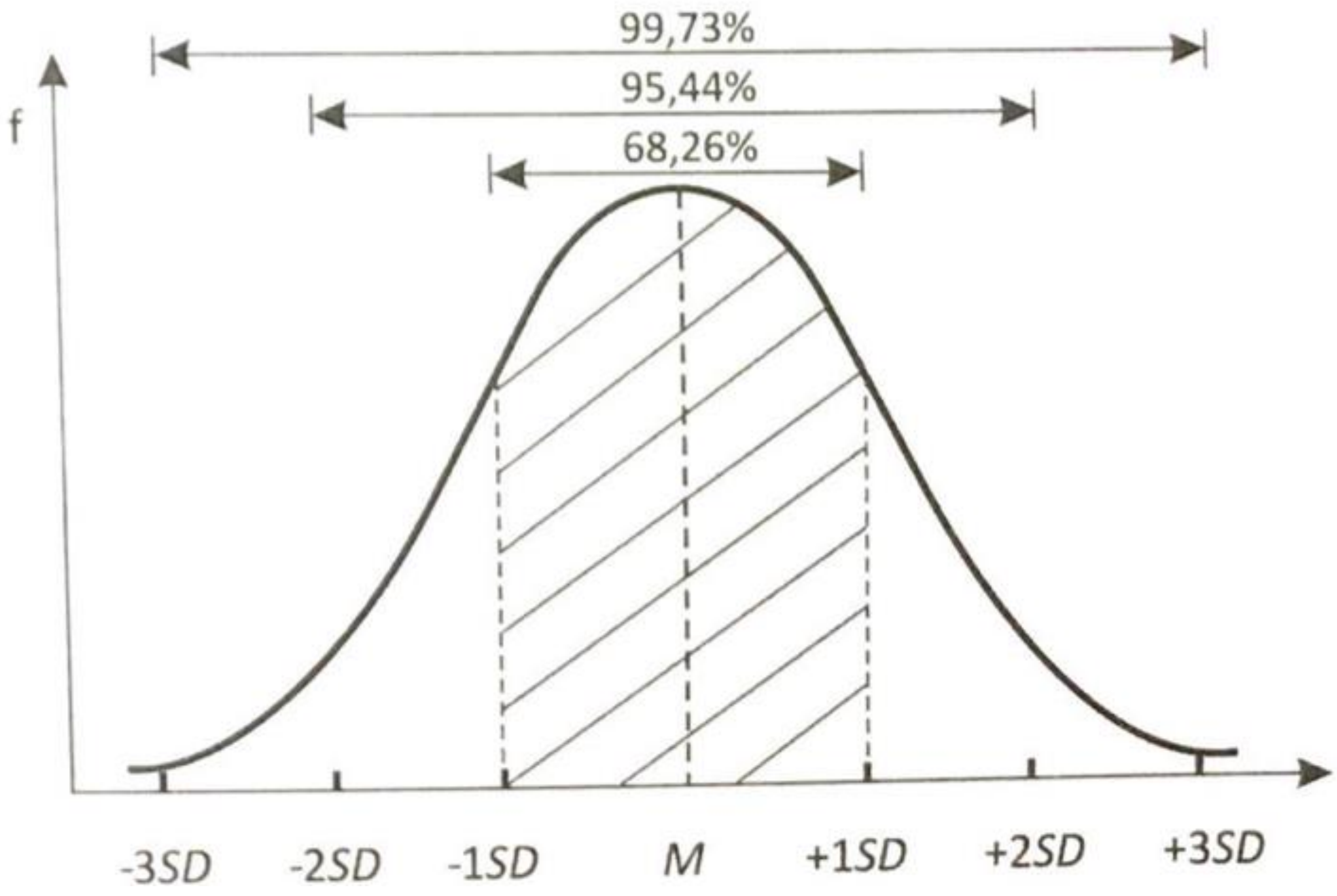
Do boys vary more in height or weight? In weight.

Do girls or boys vary in height more? Girls.

Do girls or boys vary in weight more? Girls.

Normal distribution

- Conditions:
  - if it can be assumed that there is a **true measurement value** that is relatively stable over time and that, apart from itself, only non-systematic variable factors act during its measurement
  - that we have a **large number of measurements**
  - that all measurements were carried out **using the same m**ethod and **in as similar external circumstances** as possible (e.g. the experimental and control groups must be equal in all other factors, except for the one we are currently investigating)
  - the group in which we perform measurements must be **homogeneous in all other properties, and heterogeneous in the property we measure**



99,73%
95,44%
68,26%
f
-3SD
-2SD
-1SD
M
+1SD
+2SD
+3SD

Asymmetric distribution

- when the majority of results are grouped more towards the left or right side of the range of results obtained
- when an asymmetric distribution is obtained during the application of questionnaires and tests - this is most likely due to some errors or omissions during the implementation of measurements (influence of undesirable factors on the results) - a sign for the researcher to check the data collection
- e.g. if different people worked on the questionnaire in different situations of distraction (in silence or noise)

Asymmetry indeks - skewness (1)

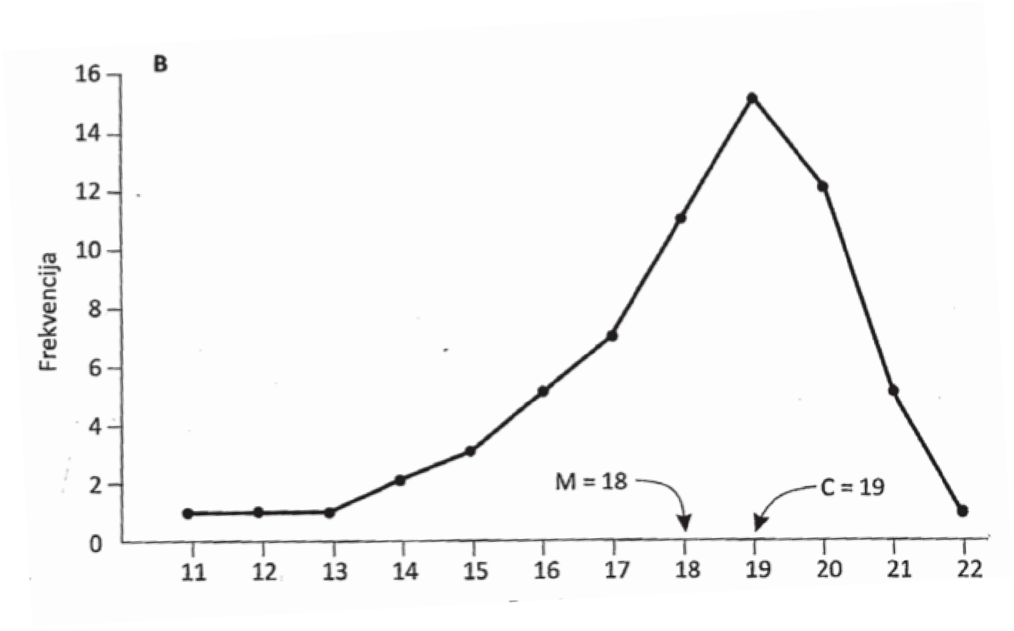
$$\alpha = \frac{3(M - C)}{SD}$$

$$\alpha_D = M - D$$

- **asymmetric distribution**: M = C, $\alpha = 0$
- **positive asymmetric distribution**: M>C>D, $\alpha > 0$
- **negative asymmetric distribution**: M<C<D, $\alpha < 0$
- it is rarely used because it is not clear how much asymmetry its numerical value shows, so it does not allow comparisons - to determine whether the obtained distribution differs from the normal one, the Kolomogorov-Smirnov test is used

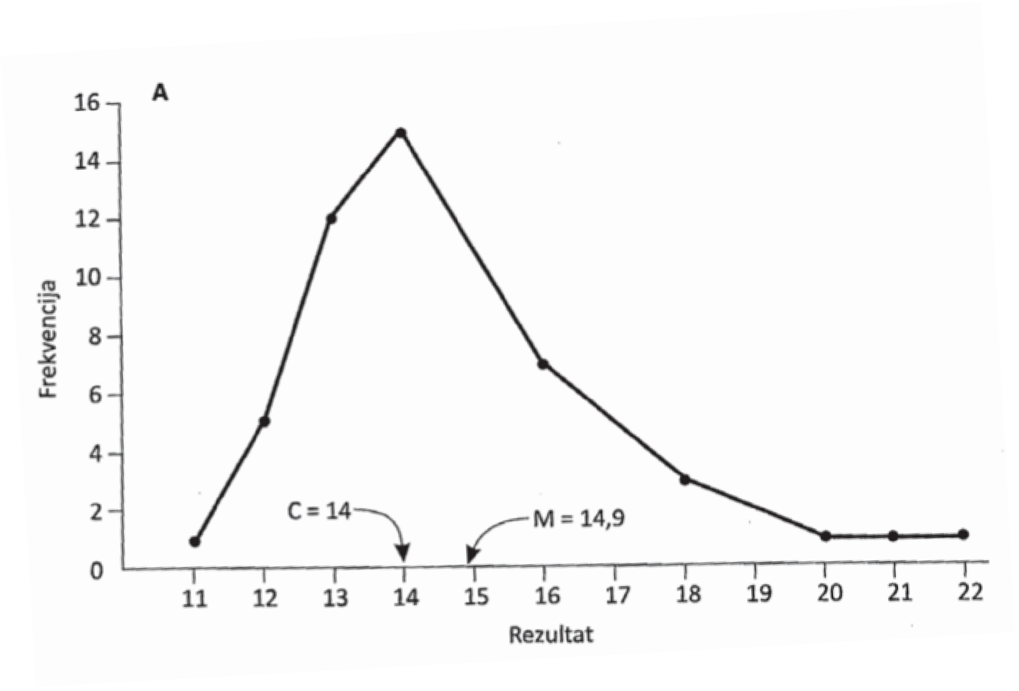
- Kurtosis - bulge, curvature, convexity - a type of deviation from a normal distribution

Asymmetry index (2)



negative asymmetric distribution

$$\alpha = \frac{3(M-C)}{SD} = \frac{3(18-19)}{SD} = \frac{-3}{SD} < 0$$

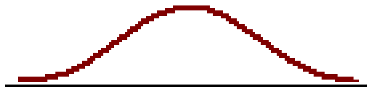


positive asymmetric distribution

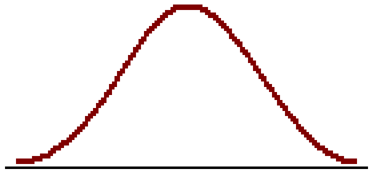
$$\alpha = \frac{3(M-C)}{SD} = \frac{3(14.9-14)}{SD} = \frac{2.7}{SD} > 0$$

Kurtosis

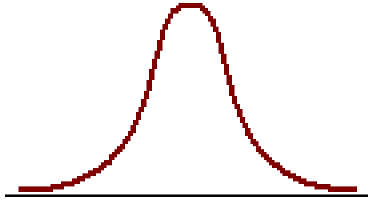
- bulge, curvature, convexity – a type of deviation from a normal distribution
- convexity of the normal distribution curve - how much the data is "peaked" or more flat compared to the normal distribution
- high kurtosis – a peak around the mean value M, falls quickly and has a long tail
- low kurtosis – almost flat distribution around the mean value M



Platykurtic distribution
Low degree of peakedness
Kurtosis <0



Normal distribution
Mesokurtic distribution
Kurtosis = 0



Leptokurtic distribution
High degree of peakedness
Kurtosis > 0

Kurtosis

- Standardized fourth moment of the distribution
- Kurtosis of the sample → estimation of the kurtosis of the population (the ratio of the central fourth moment and the standard deviation to the fourth power)

$$\boxed{k = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \frac{\sum(x_i - M)^4}{(\sum(x_i - M)^2)^2} - 3\frac{(n-1)^2}{(n-2)(n-3)}}$$

Inferential statistics

- to create a conclusion about the population from the sample

Hypothesis testing

- Objective: to **compare populations** or **estimate relationships** between variables **using samples**
- Hypotheses or predictions are tested using statistical tests
- Statistical tests estimate sampling errors in order to draw valid conclusions

- Statistical tests:
  - **parametric**
  - **non-parametric**
- Assumptions of parametric tests:
  - the population from which the sample comes follows a normal distribution
  - the sample size is large enough to represent the population
  - the variances (a measure of variability) of each group being compared are similar
- When the data violates any of these assumptions → **non-parametric tests** → "distribution-free tests"

- Types of tests: **comparison**, **correlation** or **regression tests**

Null hypothesis

1. **null-hypothesis = there is NO difference** – there is no difference between the phenomena we measure

   "there is no statistically significant difference between them"

2. **null-hypothesis = any hypothesis we want to test** → it can probably also be written in the form of the first definition

Decision	Actual situation in the population	
We reject the null hypothesis	Type I error	There is no error **the difference IS statistically significant** (eg $p<0.05$)
We accept the null hypothesis	There is no error **the difference is NOT statistically significant** (eg $p>0.05$)	Type II error

Statistical test

- A statistical test summarizes observed data into a single number using means, variances, sample size, and number of variables
- Statistical test for hypothesis testing:
    - determining whether the **independent variable has a statistically significant relationship with the dependent variable**
    - estimation of the **difference between two or more groups**
- **Null hypothesis** – there is no relationship or no difference between the groups
- A statistical test determines whether the observed data fall outside the range of values predicted by the null hypothesis
- It is generally calculated as the ratio of eg the correlation between variables or the difference between groups and variance in the data
- **p-value:**
    - estimates how likely it is to see the difference described by the test statistic if the null hypothesis is true
    - helps in making a decision whether to reject the null hypothesis

Hypothesis testing

Defining null hypothesis

- H0: Men, on average, are not taller than women.

Data collection

- The sample must have an equal number of men and women, from different socioeconomic and other groups, in one country.

Implementation of statistical test

- E.g. t-test with which we check that men are actually taller than women:
  - estimation of the difference in the average heights of the 2 groups
  - p-value shows the probability of observing this difference if H0 is true

The decision to reject the null hypothesis

- $p=0.002 \rightarrow p<0.05 \rightarrow$ we reject the null hypothesis of no difference

Publication of results

- We reject the null hypothesis that men are NOT on average taller than women and conclude that there is a possible difference in the height of men and women

What significance level to take?

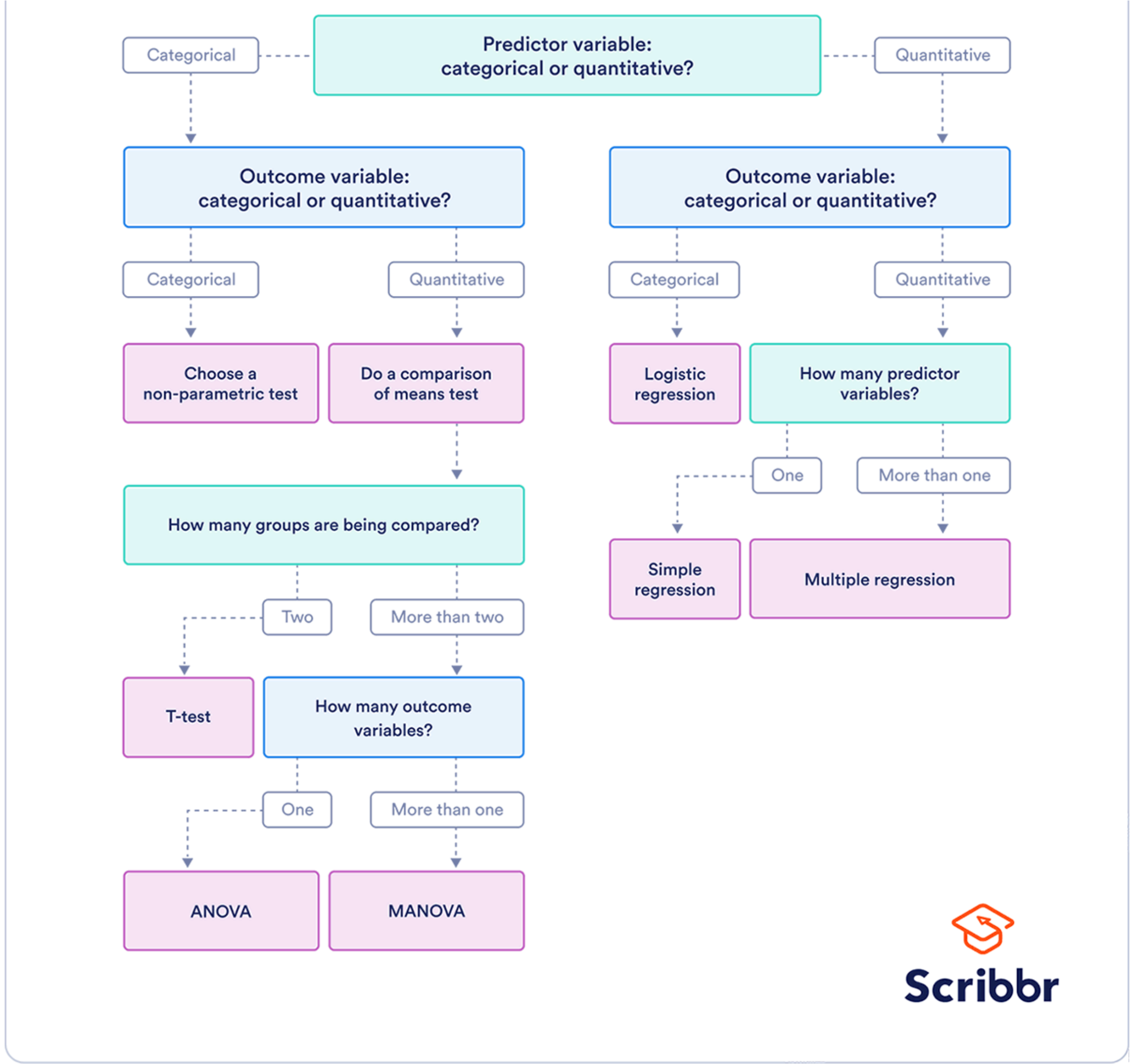
- **it depends on the importance of the consequences if the conclusion is wrong - it should be careful**
- the effect of 2 drugs: one very dangerous and excluded from use → we want the second one to be definitely different from the first → we need a higher degree of safety: 1% or 0.1%
- investigation of side effects of a sedative → better to admit that a side effect exists, than to ignore it → declare it significant → lower degree of safety: 5% or 10%

- if we tend to accept problematic evidence about the defendant's guilt - we risk punishing many innocent people → if we do not take everything into account, but only the strongest evidence about someone's guilt - many people who are guilty will remain unpunished

- whether the new organization of work has an effect, because otherwise the procedure is too expensive or not worth it → stricter level <1%
- whether the new organization of work has an effect, and the procedure is neither more expensive, nor more complex, nor more dangerous → milder level >5%

- a new tumor surgery technique (if it certainly cannot harm the patient and has a higher percentage of recovery) → we can agree to that method even with 20% or 30%
- when saving lives, we agree to a new method even with a 1% probability that it is better than the old method

Choice of statistical test



Parametric tests

- Regression
- Comparison
- Correlation

Regression tests

	Independent variable	Dependent variable
Simple linear regression	• Continuous • 1 independent variable	• Continuous • 1 dependent variable
Multiple linear regression	• Continuous • 2 or more independent variables	• Continuous • 1 dependent variable
Logistric regression	• Continuous	• Binary

What is the effect of salary on longevity? **SLR**

What is the effect of income and minutes of exercise per week on longevity? **MLR**

What is the effect of the drug dose on the survival of a test animal? **LR**

Comparison tests

	Independent variable	Dependent variable
Paired t-test	• Categorical • 1 independent variable	• Quantitative • groups from the same population
Independent t-test	• Categorical • 1 independent variable	• Quantitative • groups from different populations
ANOVA	• Categorical • 1 or more independent variables	• Quantitative • 1 dependent variable
MANOVA	• Categorical • 1 or more independent variables	• Quantitative • 2 or more dependent variables

What is the difference in the average grades at the graduation exam for students from two different schools?

ITT

What is the difference in mean pain levels among postoperative patients who received three different pain medications?

ANOVA

What is the effect of two different preparation programs for graduation exam on the average test scores of students from the same class?

PTT

What is the effect of flower type on petal length, petal width and stem length?

MANOVA

Correlation tests

	Variable
Pearson’s r	• 2 continuous variables

How are latitude and temperature are related?

Non-parametric tests

	Nezavisna varijabla	Zavisna varijabla	Koristi se umjesto
Spearman's r	• Quantitative	• Quantitative	Pearson's r
Chi square test of independence	• Categorical	• Categorical	Pearson's r
Sign test	• Categorical	• Quantitative	One-sample T-test
Kruskal-Wallis H	• Categorical • 3 or more groups	• Quantitative	ANOVA
ANOSIM	• Categorical • 3 or more groups	• Quantitative • 2 or more dependent variables	MANOVA
Wilcoxon Rank-Sum test	• Categorical • 2 groups	• Quantitative • groups from different populations	Independent t-test
Wilcoxon Signed-rank test	• Categorical • 2 groups	• Quantitative • groups from the same population	Paired t-test

References

Boris Petz, Vladimir Kolesarić, Dragutin Ivanec. Petzova statistika. Osnovne statističke metode za nematematičare. Naklada Slap.

https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/

https://www.scribbr.com/statistics/statistical-tests/