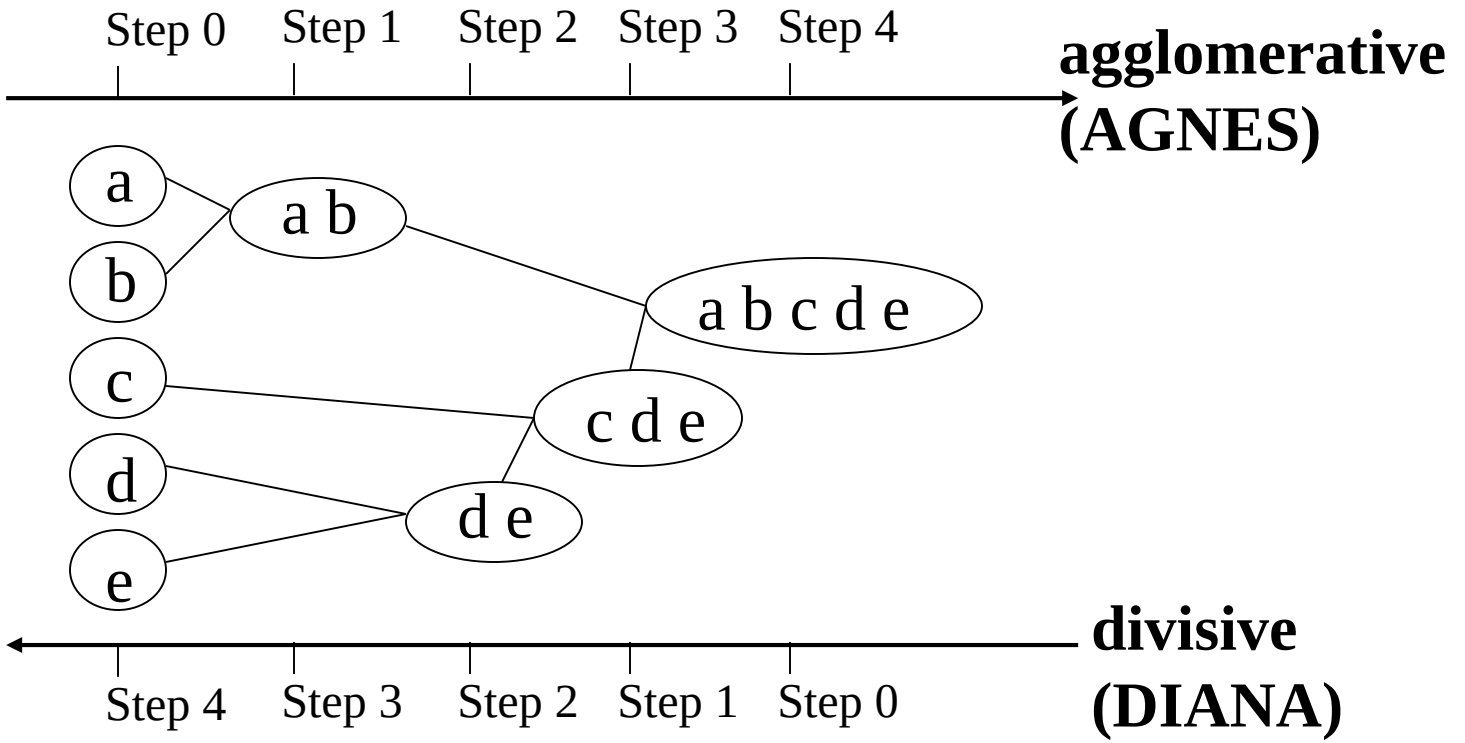
Topic No. 4

1. **Hierarchical Clustering**
2. **Statistical hypothesis testing**
   - **Chi-square test (previous lecture)**
   - **Student's *t*-test**
3. **Bayesian inference**

Hierarchical Clustering

- Use distance matrix as clustering criteria. This method does not require the number of clusters ***k*** as an input, but needs a termination condition



Hierarchical Clustering

- The first step in the hierarchical clustering process is to look for the pair of samples that are the most similar
- The closest in the sense of having the lowest dissimilarity – this is the pair B and F, with dissimilarity equal to 0.2000.
- These two samples are then joined at a level of 0.2000 in the first step of the **dendrogram**, or clustering tree (see the first diagram, and the vertical scale of 0 to 1 which calibrates the level of clustering).
- The point at which they are joined is called a node
- Repeat this process

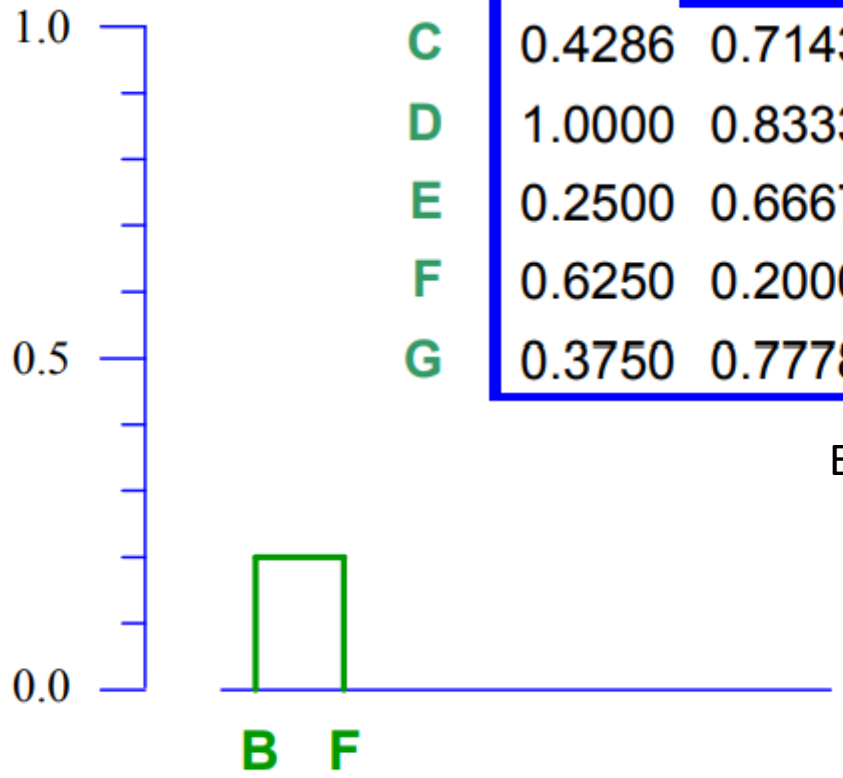
Dendrogram : a tree diagram, especially showing taxonomic relationships

Hierarchical Clustering

Dissimilarity Matrix

samples	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0

Each value is a distance measure



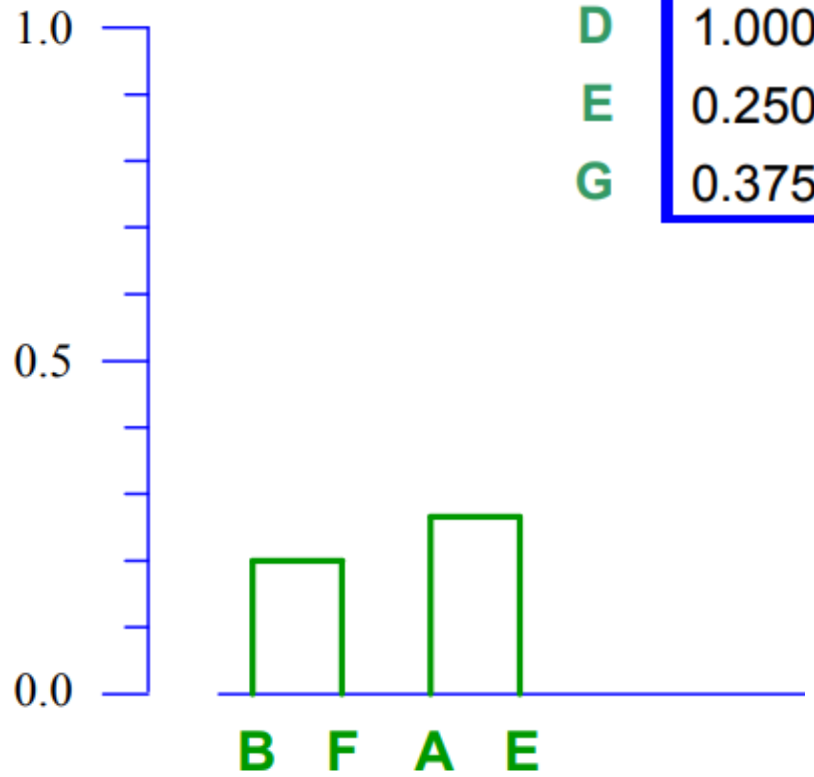
http://www.econ.upf.edu/~michael/stanford/maeb7.pdf

Hierarchical Clustering

- Calculated the dissimilarity between the merged pair (B,F) and the other samples.
- The **maximum**, or **complete linkage**, method
  - The dissimilarity between the merged pair and the others will be the maximum of the pair of dissimilarities in each case.
  - For example, the dissimilarity between B and A is 0.5000, while the dissimilarity between F and A is 0.6250. hence we choose the maximum of the two, 0.6250, to quantify the dissimilarity between (B,F) and A.
- Continuing in this way we obtain a new dissimilarity matrix

Hierarchical Clustering

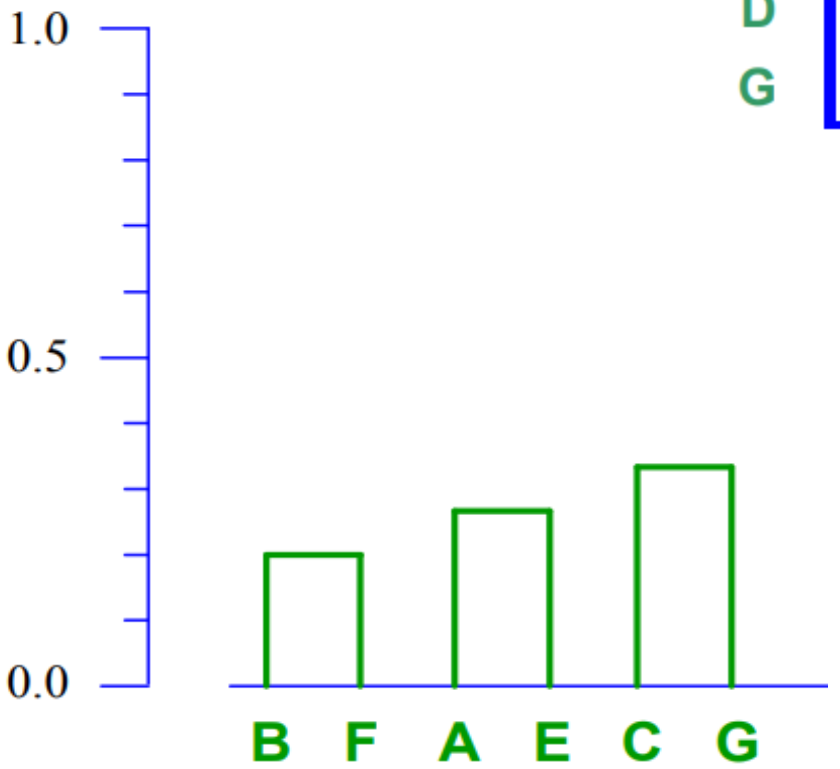
samples	A	(B,F)	C	D	E	G
A	0	0.6250	0.4286	1.0000	0.2500	0.3750
(B,F)	0.6250	0	0.7143	0.8333	0.7778	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8571
E	0.2500	0.7778	0.4286	1.0000	0	0.3750
G	0.3750	0.7778	0.3333	0.8571	0.3750	0



samples	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0

Hierarchical Clustering

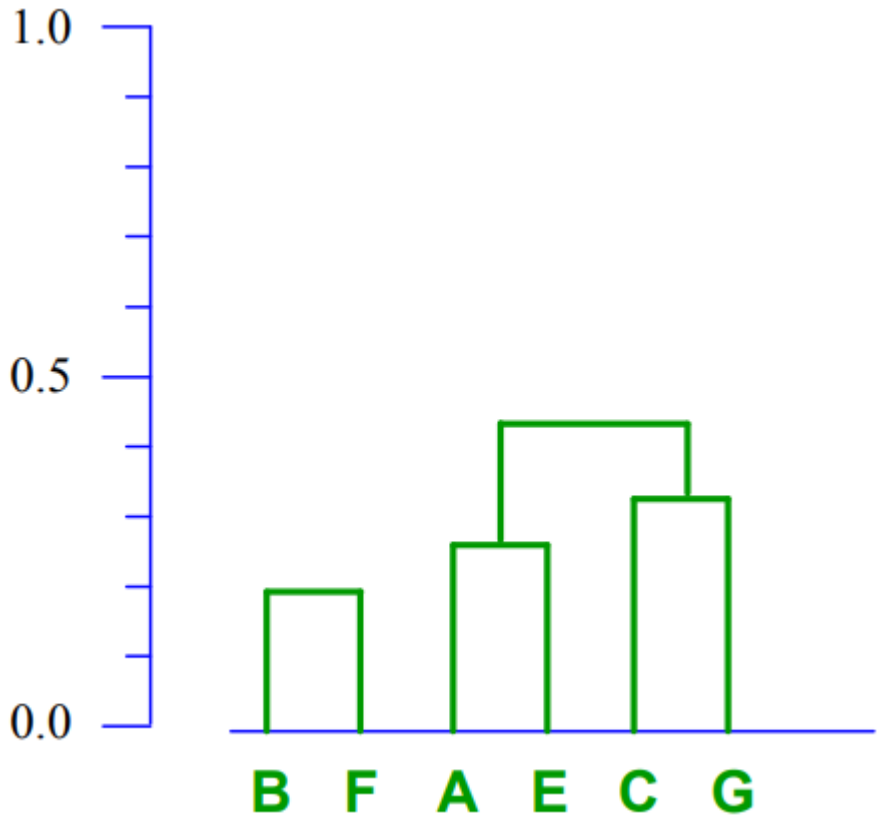
samples	(A,E)	(B,F)	C	D	G
(A,E)	0	0.7778	0.4286	1.0000	0.3750
(B,F)	0.7778	0	0.7143	0.8333	0.7778
C	0.4286	0.7143	0	1.0000	0.3333
D	1.0000	0.8333	1.0000	0	0.8571
G	0.3750	0.7778	0.3333	0.8571	0



samples	A	(B,F)	C	D	E	G
A	0	0.6250	0.4286	1.0000	0.2500	0.3750
(B,F)	0.6250	0	0.7143	0.8333	0.7778	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8571
E	0.2500	0.7778	0.4286	1.0000	0	0.3750
G	0.3750	0.7778	0.3333	0.8571	0.3750	0

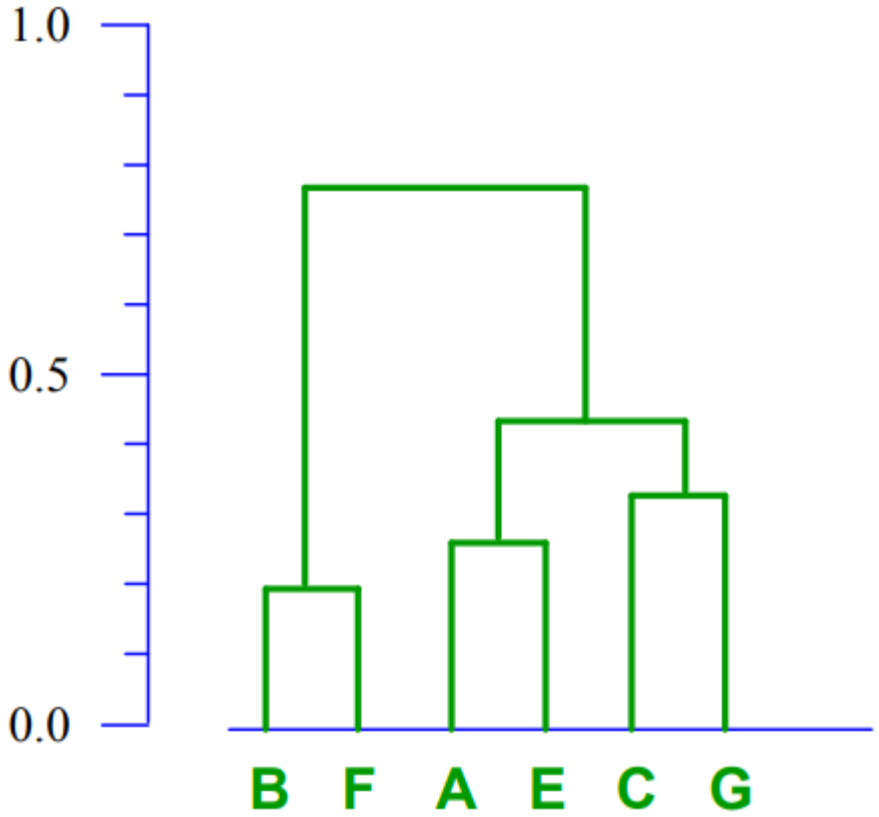
Hierarchical Clustering

samples	(A,E)	(B,F)	(C,G)	D
(A,E)	0	0.7778	0.4286	1.0000
(B,F)	0.7778	0	0.7778	0.8333
(C,G)	0.4286	0.7778	0	1.0000
D	1.0000	0.8333	1.0000	0



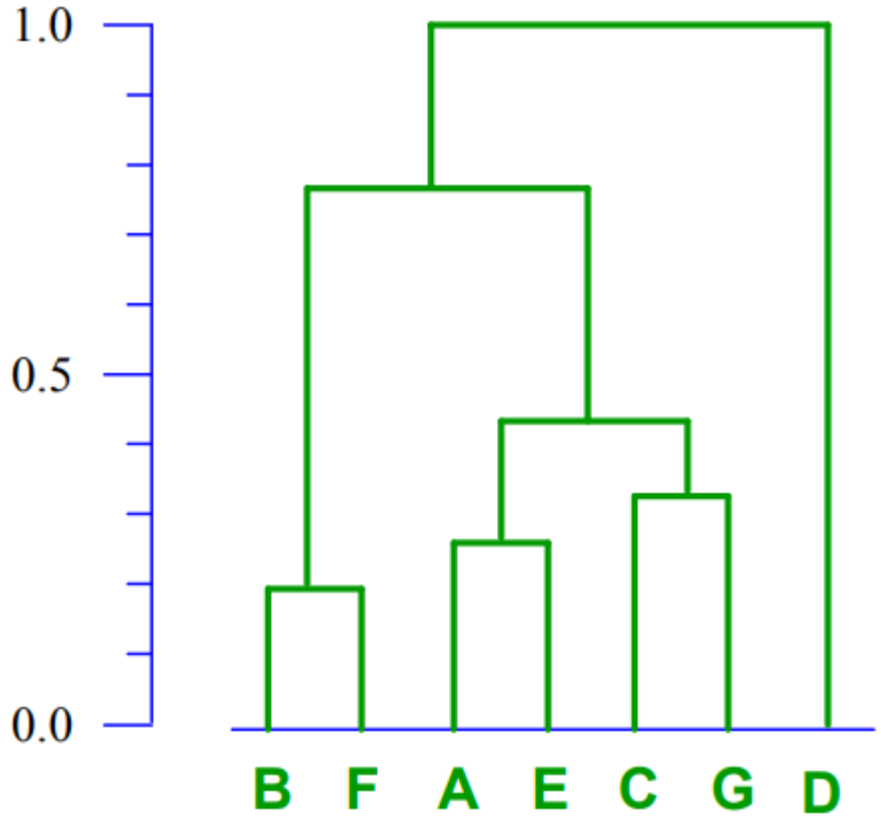
Hierarchical Clustering

samples	(A,E,C,G)	(B,F)	D
(A,E,C,G)	0	0.7778	1.0000
(B,F)	0.7778	0	0.8333
D	1.0000	0.8333	0

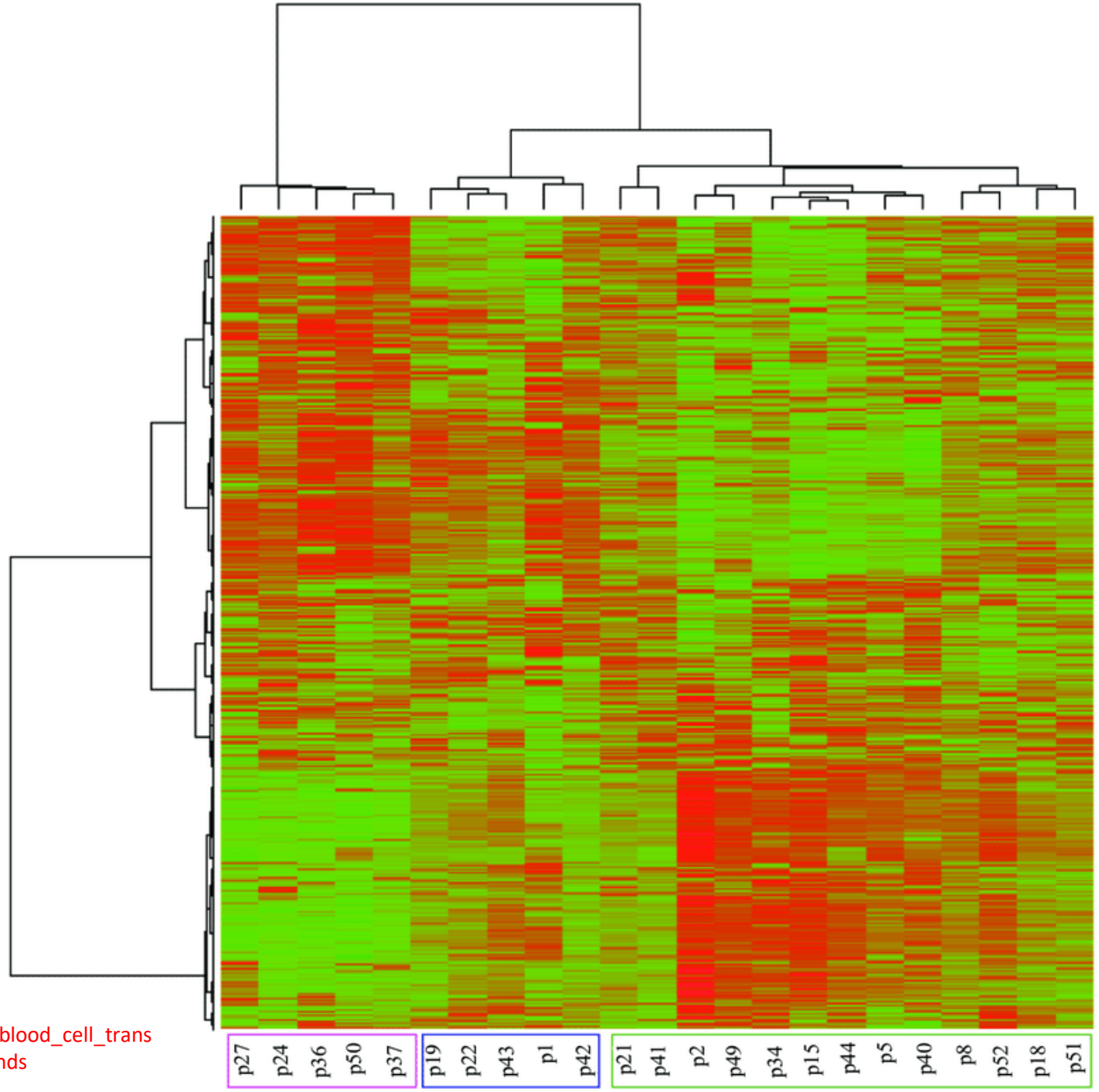


Hierarchical Clustering

samples	(A,E,C,G,B,F)	D
(A,E,C,G,B,F)	0	1.0000
D	1.0000	0



Hierarchical clustering of gene expression data with rows corresponding to genes and columns to samples (patients). Unsupervised classification (Ward method) of 20 000 most variable genes. Up-regulated genes are denoted in green, down-regulated in red



https://www.researchgate.net/publication/312480456_Correlation_between_peripheral_blood_cell_transcriptomic_profile_and_clinical_parameters_of_chronic_mitral_valve_disease_in_Dachshunds

- method='single' assigns

$$d(u, v) = \min(dist(u[i], v[j]))$$

for all points i in cluster u and j in cluster v. This is also known as the Nearest Point Algorithm.

- method='complete' assigns

$$d(u, v) = \max(dist(u[i], v[j]))$$

for all points i in cluster u and j in cluster v. This is also known by the Farthest Point Algorithm or Voor Hees Algorithm.

- method='average' assigns

$$d(u, v) = \sum_{ij} \frac{d(u[i], v[j])}{(|u| * |v|)}$$

for all points i and j where $|u|$ and $|v|$ are the cardinalities of clusters u and v, respectively. This is also called the UPGMA algorithm.

- method='weighted' assigns

$$d(u, v) = (dist(s, v) + dist(t, v))/2$$

where cluster u was formed with cluster s and t and v is a remaining cluster in the forest (also called WPGMA).

- method='centroid' assigns

$$dist(s, t) = ||c_s - c_t||_2$$

where c_s and c_t are the centroids of clusters s and t, respectively. When two clusters s and t are combined into a new cluster u, the new centroid is computed over all the original objects in clusters s and t. The distance then becomes the Euclidean distance between the centroid of u and the centroid of a remaining cluster v in the forest. This is also known as the UPGMC algorithm.

- method='median' assigns $d(s, t)$ like the `centroid` method. When two clusters s and t are combined into a new cluster u, the average of centroids s and t give the new centroid u. This is also known as the WPGMC algorithm.

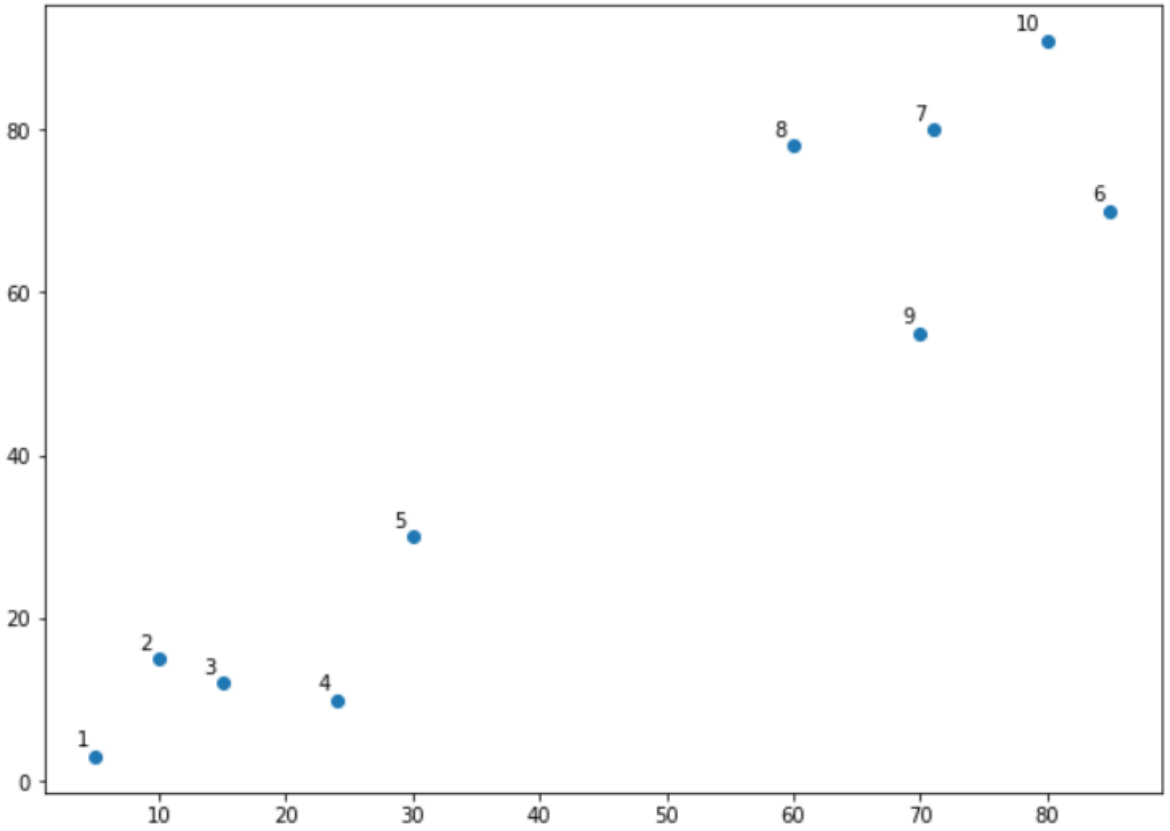
- method='ward' uses the Ward variance minimization algorithm. The new entry $d(u, v)$ is computed as follows,

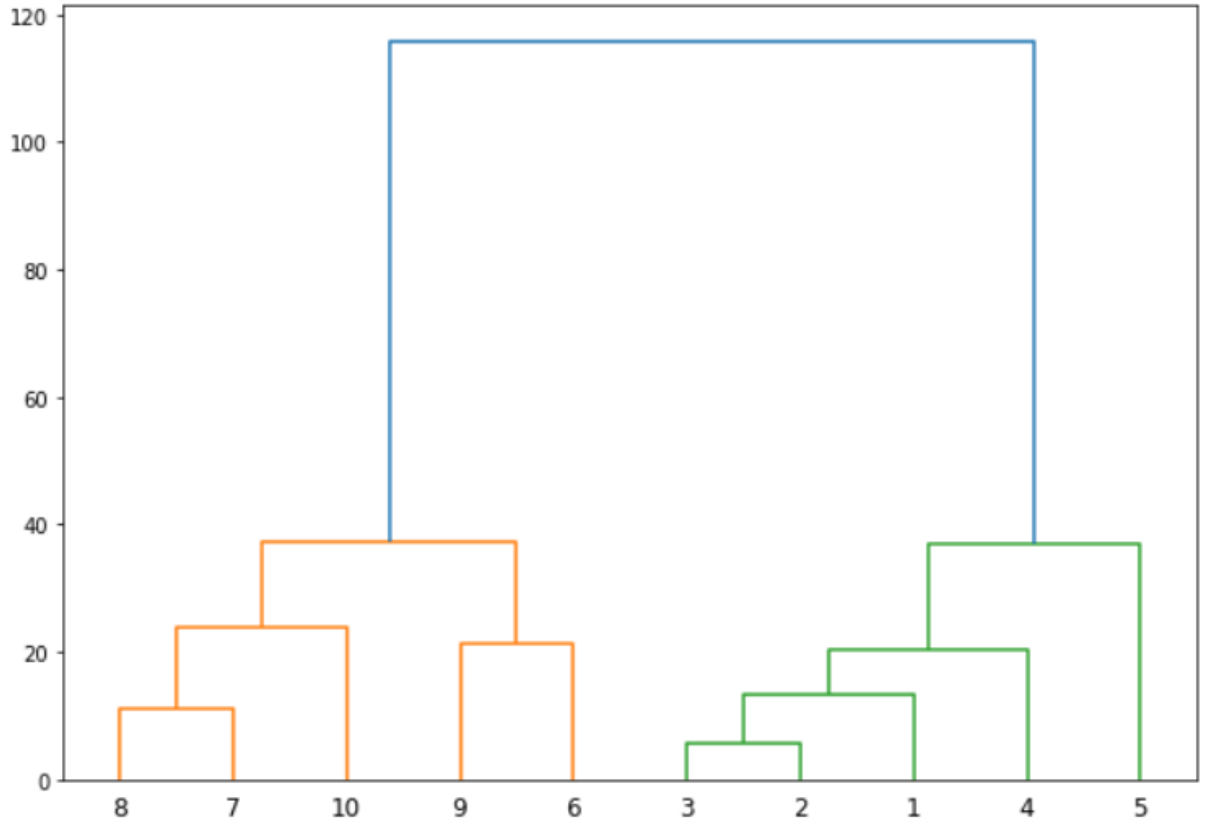
$$d(u,v) = \sqrt{\frac{|v|+|s|}{T}d(v,s)^2 + \frac{|v|+|t|}{T}d(v,t)^2 - \frac{|v|}{T}d(s,t)^2}$$

where u is the newly joined cluster consisting of clusters s and t, v is an unused cluster in the forest, $T = |v| + |s| + |t|$, and $|*|$ is the cardinality of its argument. This is also known as the incremental algorithm.

Example

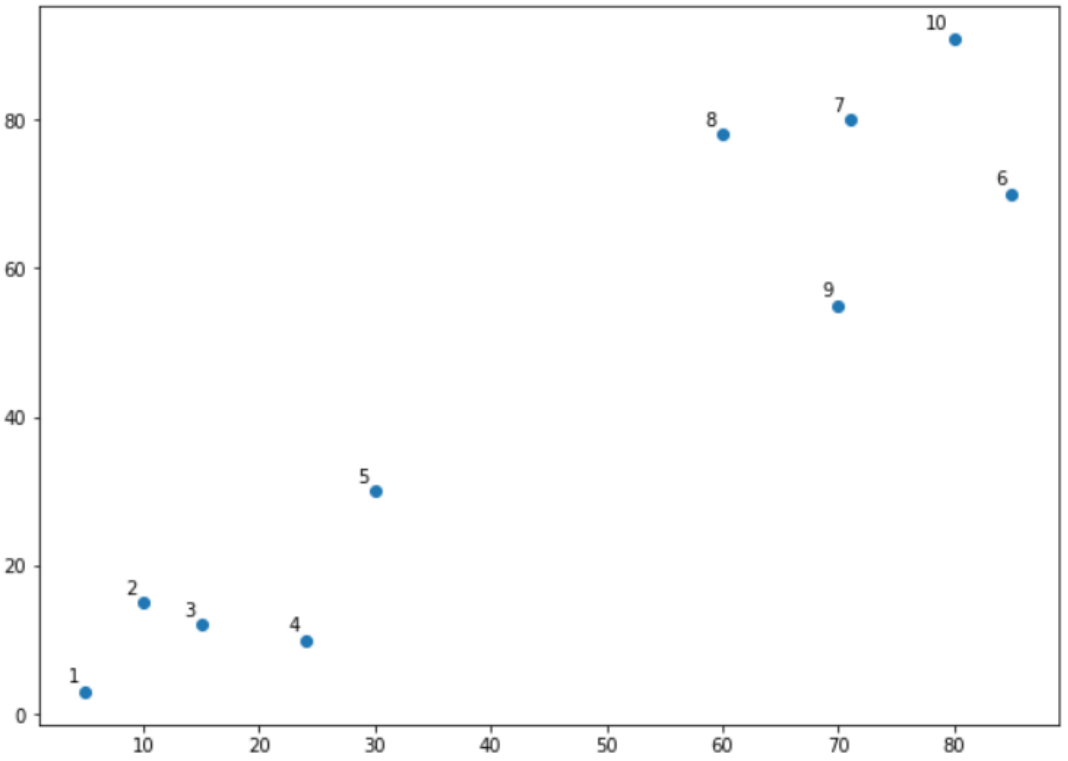
[5,3], [10,15], [15,12], [24,10], [30,30], [85,70], [71,80], [60,78], [70,55], [80,91]

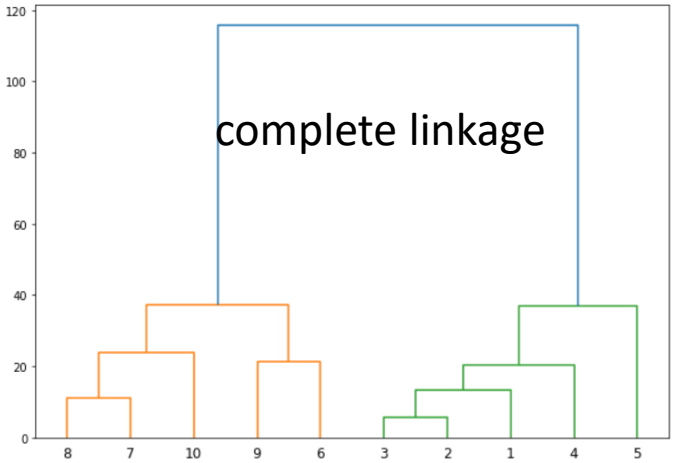


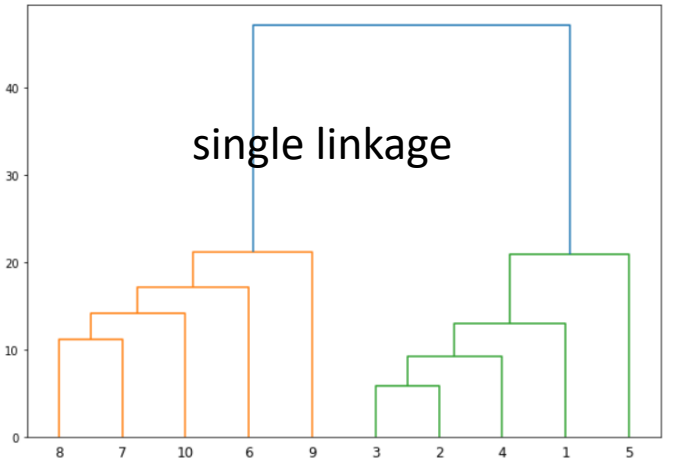


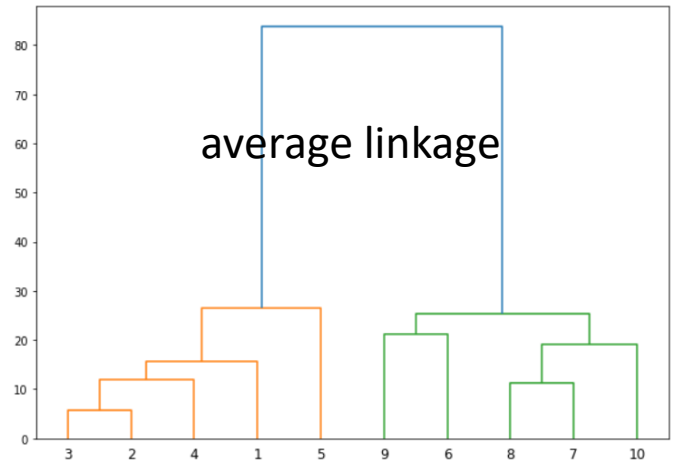
complete linkage

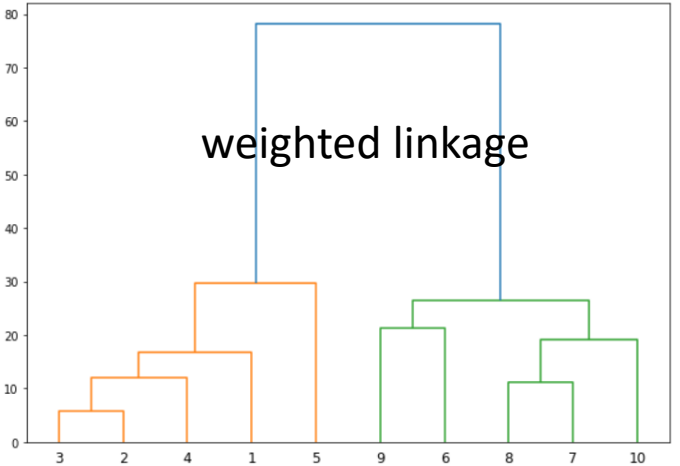
Example

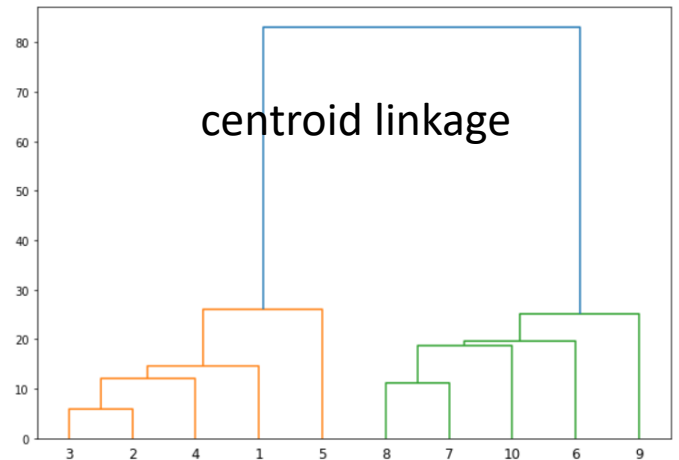

10
8
7
6
9
5
2
3
4
1
0
20
40
60
80
10
20
30
40
50
60
70
80

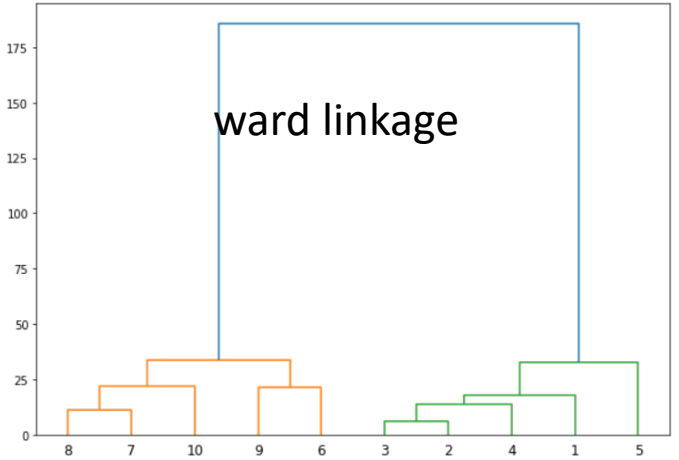

complete linkage
120
100
80
60
40
20
0
8 7 10 9 6 3 2 1 4 5


single linkage
40
30
20
10
0
8 7 10 6 9 3 2 4 1 5


average linkage
80
70
60
50
40
30
20
10
0
3 2 4 1 5 9 6 8 7 10


weighted linkage
80
70
60
50
40
30
20
10
0
3 2 4 1 5 9 6 8 7 10


centroid linkage
80
70
60
50
40
30
20
10
0
3 2 4 1 5 8 7 10 6 9


ward linkage
175
150
125
100
75
50
25
0
8 7 10 9 6 3 2 4 1 5

Statistical Hypothesis Testing

- Most beans in this bag are white. (population)
- Few beans of this handful are white. (sample)
- Therefore: Probably, these beans were taken from another bag. (reject null hypothesis)

To be a real statistical hypothesis test, it requires

1) the formalities of a probability calculation
2) and a comparison of that probability to a standard.

A simple generalization : a mixed bag of beans and a handful that contain either **very few** or **very many** white beans.

1) If the composition of the handful is greatly different from that of the bag, then the sample probably originated from another bag.
2) The original example is termed a one-sided or a **one-tailed test** while the generalization is termed a two-sided or **two-tailed test**.

The statement also relies on the inference that the **sampling was random**.

1) If someone had been picking through the bag to find white beans, then it would explain why the handful had so many white beans

https://en.wikipedia.org/wiki/Statistical_hypothesis_testing

Courtroom trial

In the start of the procedure, there are two hypotheses

"the defendant is not guilty“ → *null hypothesis* (H_0)

"the defendant is guilty". → *alternative hypothesis* (H_1)

Type I error (false positive)

1) The hypothesis of innocence is rejected only when an error is very unlikely, because one doesn't want to convict an innocent defendant.
2) The conviction of an innocent person.
3) Therefore, the occurrence of this error is controlled to be rare.

Type II error (false negative)

1) As a consequence of the asymmetric behavior.
2) acquitting a person who committed the crime
3) more common.

A hypothesis test can be regarded as either a judgment of a hypothesis or as a judgment of evidence.

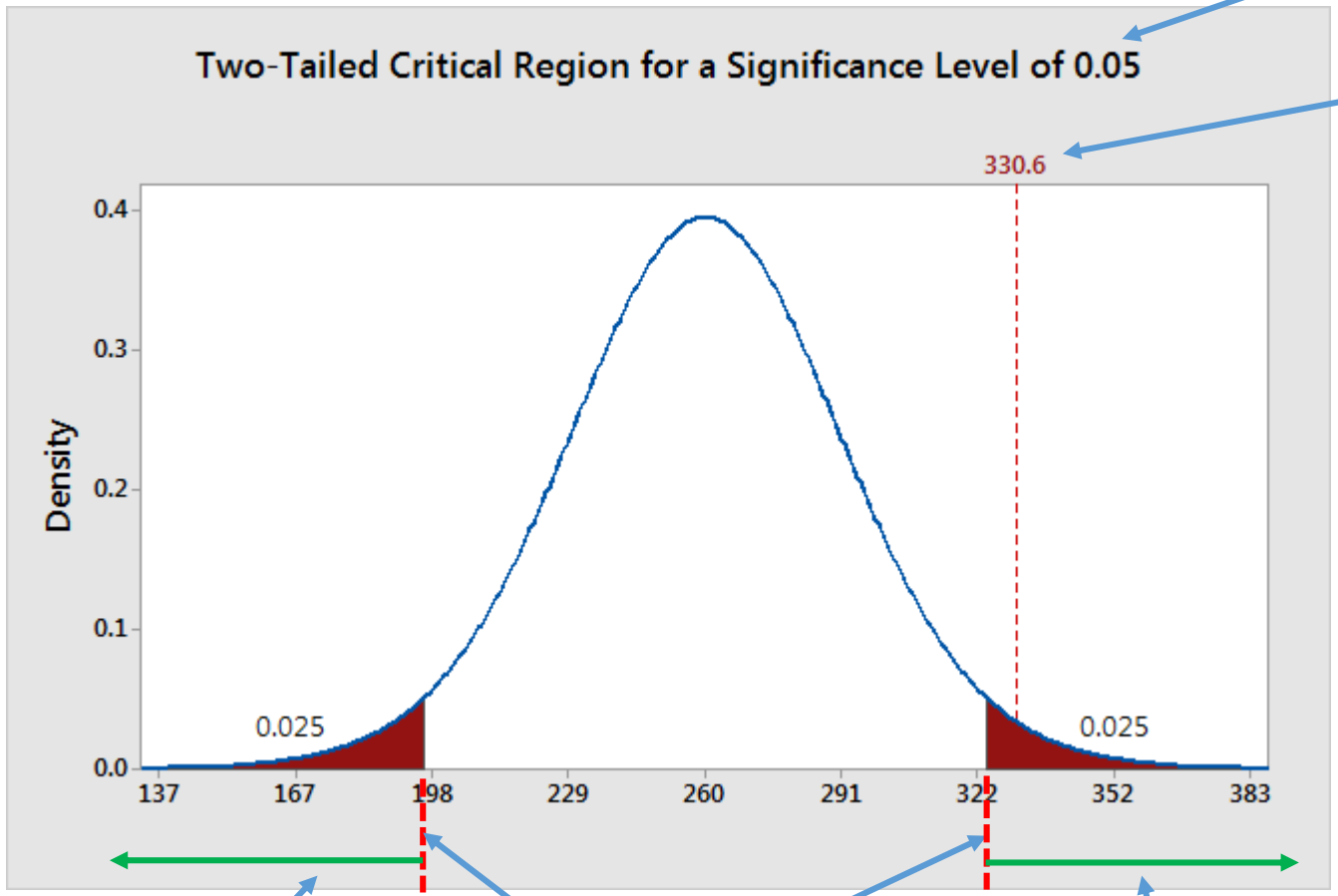
	Truly not guilty	Truly guilty
Do not reject the null hypothesis (Acquittal)	Right decision	Wrong decision Type II Error
Reject null hypothesis (Conviction)	Wrong decision Type I Error	Right decision

The usual reasoning procedure:

1. State the relevant **null** and **alternative hypotheses**.
2. The statistical assumptions being made about the sample; for example, assumptions about the statistical independence or about the form of the distributions of the observations. This is equally important as invalid assumptions will mean that the results of the test are invalid.
3. Decide which test is appropriate, and state the relevant **test statistic** T.
4. Derive the distribution of the test statistic under the null hypothesis from the assumptions. In standard cases this will be a well-known result. For example, the test statistic might follow a Student's t distribution with known degrees of freedom, or a normal distribution with known mean and variance.
5. Select a significance level (α), a probability threshold below which the null hypothesis will be rejected. Common values are 5% (5% risk of rejecting null hypothesis) and 1%.
6. The distribution of the test statistic under the null hypothesis partitions the possible values of T into those for which the null hypothesis is rejected—the so-called *critical region*—and those for which it is not. The probability of the critical region is α.
7. Compute from the observations the observed value t_{obs} of the test statistic T.
8. Decide to either reject the null hypothesis in favor of the alternative or not reject it. The decision rule is to reject the null hypothesis H_0 if the observed value t_{obs} is in the critical region, and not to reject the null hypothesis otherwise.

α : shaded area

t_{obs} value : calculated t value



critical region

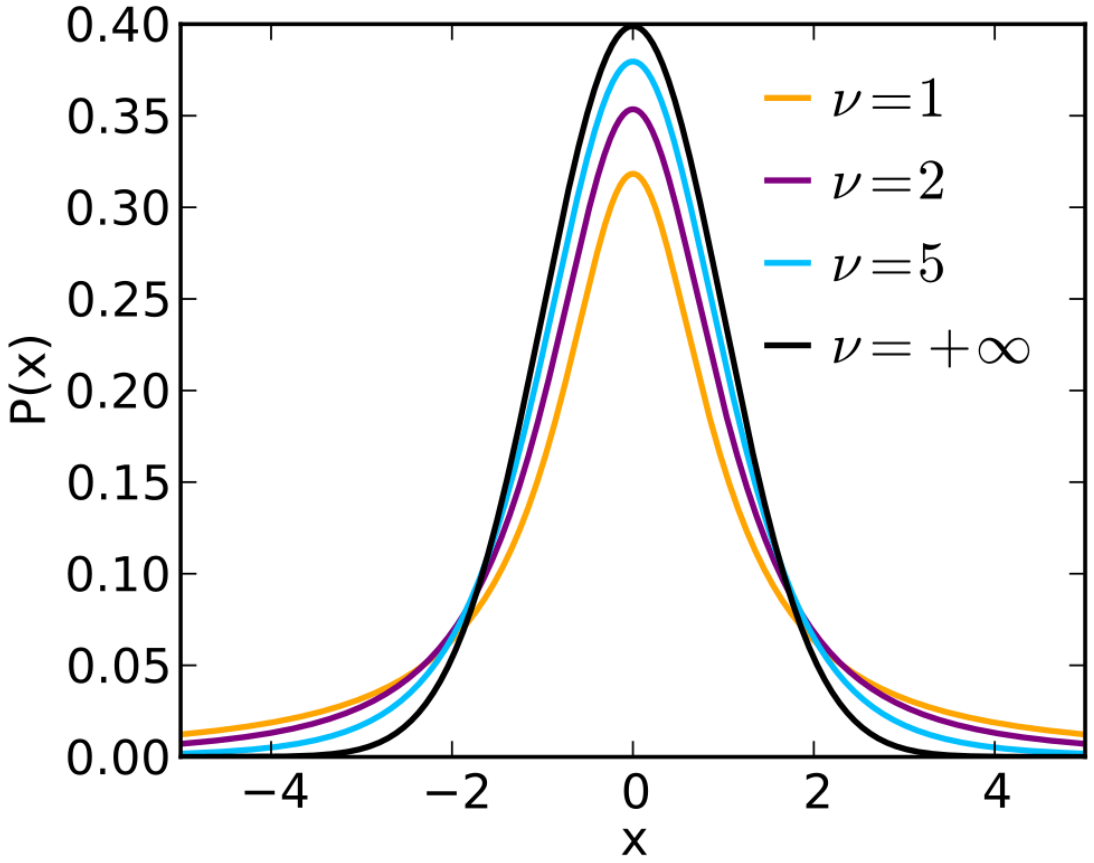
t_{crt} value : from the table or from pdf(cdf)

critical region

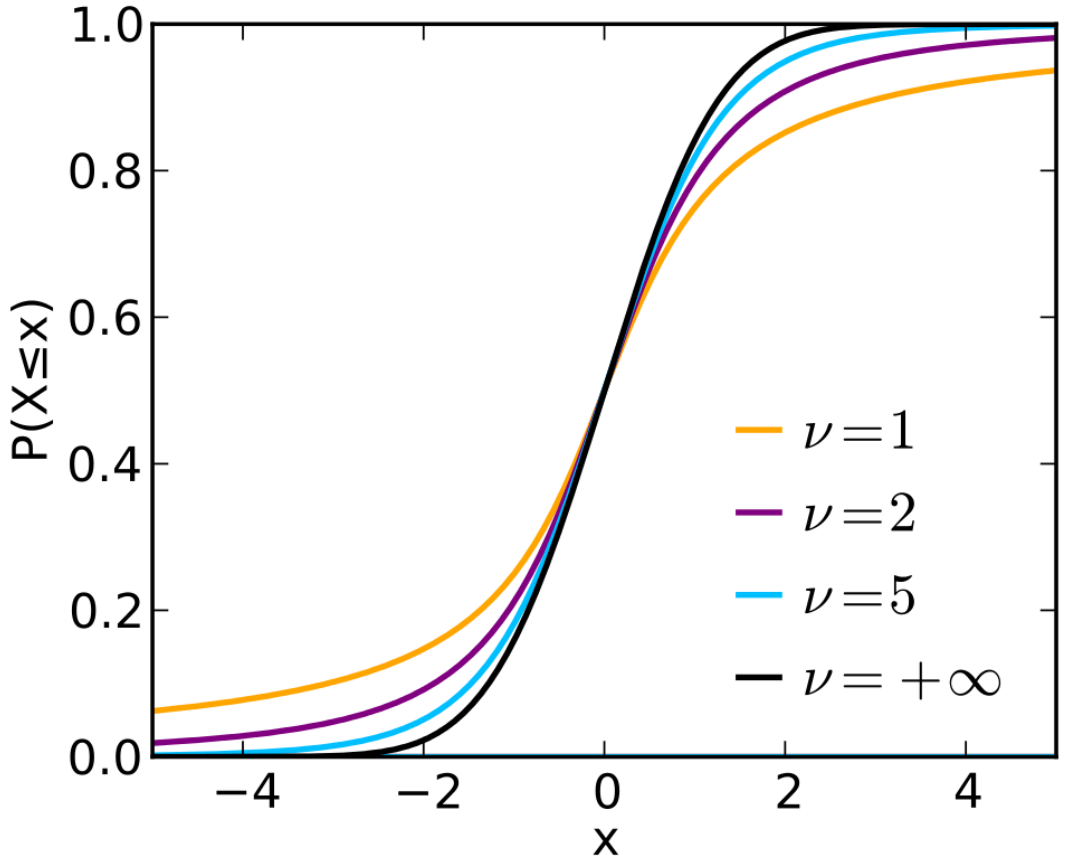
A common alternative formulation :

1. Compute from the observations the observed value t_{obs} of the test statistic T.
2. Calculate the *p*-value. This is the probability, under the null hypothesis, of sampling a test statistic at least as extreme as that which was observed.
3. Reject the null hypothesis, in favor of the alternative hypothesis, if and only if the *p*-value is less than (or equal to) the significance level (the selected probability) threshold.

t statistics



PDF



CDF

t-test

1. The *t*-test is one of hypothesis testing in statistics.
2. A *t*-test is used to determine if there is a **significant difference between the means of two groups**, which may be related in certain features.
3. That is, a *t*-test allows us to compare the average values of the two data sets and determine if they came from the same population
4. Take a sample from each of the two sets and establishes a **null hypothesis** that the *two means are equal*.
5. There are several different types of *t*-test that can be performed depending on the data and type of analysis required.
6. Assumptions
    1) Randomly sampled
    2) Sampled data is Gaussian
    3) Homogeneity of variance (standard deviations are approximately equal)

Types of *t-test*

1. Number of populations
    1) One-sample : mean of a population has the values specified by a null hypothesis
    2) Two-sample : means of 2 populations are equal
2. Dependency
    1) Unpaired (independent) samples : two separate sets of independent and identically distributed samples are obtained, one from each of the two populations being compared (test group and control group)
    2) Paired samples : sample of matched pairs of similar units, or one group of units that has been tested twice (subjects are tested prior to a treatment)

1. One-sample t-test

the null hypothesis that the population mean is equal to a specified value μ_0

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

t_{obs} value

$\bar{x}$ is the sample mean
s is the sample standard deviation
n is the sample size.
The degree of freedom is $n - 1$.

1. One-sample t-test : example

Collect random sample of 31 screws (same kind) from Home Depot. The labels claim that each screw weighs 20 grams. Is the label statistically correct?

Null hypothesis : the screws are 20 grams

20.70	27.46	22.15	19.85	21.29	24.75	
20.75	22.91	25.34	20.33	21.54	21.08	
22.14	19.56	21.10	18.04	24.12	19.95	
19.72	18.28	16.26	17.46	20.53	22.12	
25.06	22.44	19.08	19.88	21.39	22.33	25.79

$\bar{x} = 21.40$

$s = 2.54$

$n = 31.$

d.f. = $n - 1 = 30.$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = 3.07$$

α = 0.05

3.07 > 2.042

➡ Reject H_0 (the screw is heavier than 20grams)

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

2. Independent two-sample *t*-test

1) Equal sample sizes and variance

- the two sample sizes (that is, the number *n* of participants of each group) are equal;
- it can be assumed that the two distributions have the same variance

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{n}}}$$

- $S_{X_1}^2$ and $S_{X_2}^2$ are the unbiased estimators of the variances of the two samples.
- The denominator of *t* is the standard error of the difference between two means.
- The degree of freedom is $2n - 2$ where *n* is the number of participants in each group

2. Independent two-sample *t*-test

2) Equal (or unequal) sample sizes and unequal variances (Welch's *t*-test)

- the two population variances are not assumed to be equal and hence must be estimated separately
- the two sample sizes may or may not be equal

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

- S_i^2 is the unbiased estimators of the variance sample i.

$$\mathrm{d.f.} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

3. Dependent *t*-test for paired-sample

- the samples are dependent
  - when there is only one sample that has been **tested twice** (repeated measures) or
  - when there are two samples that have been **matched** or “paired”

$$t = \frac{\bar{X}_D}{s_D / \sqrt{n}}$$

- $\bar{X}_D$ and S_D are the average and standard deviation of the differences between all pairs
- The degree of freedom is *n* − 1, where *n* represents the number of pairs

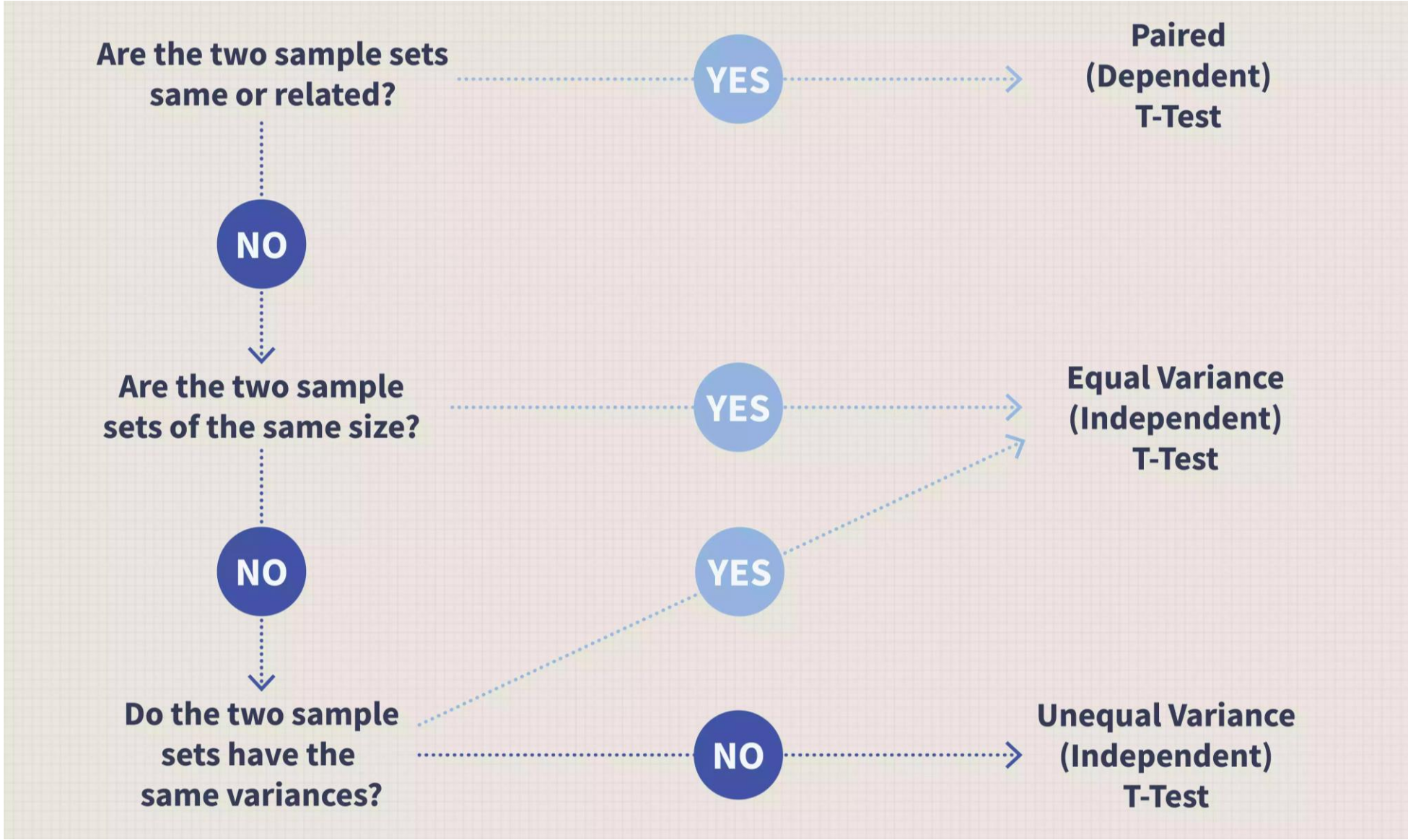
Subject #	Score 1	Score 2	X-Y
1	3	20	-17
2	3	13	-10
3	3	13	-10
4	12	20	-8
5	15	29	-14
6	16	32	-16
7	17	23	-6
8	19	20	-1
9	23	25	-2
10	24	15	9
11	32	30	2
		SUM:	**-73**

Null hypothesis : the difference between the 2 scores is 0

• $\bar{X}_D$ and S_D are the average and standard deviation of the differences between all pairs
• The degree of freedom is *n* - 1, where *n* represents the number of pairs

t = -2.74 α = 0.05

2.74 > 2.228 ➡ Reject H_0 (score 1 and score 2 are different)


Are the two sample sets same or related?
YES
Paired (Dependent) T-Test
NO
Are the two sample sets of the same size?
YES
Equal Variance (Independent) T-Test
NO
YES
Do the two sample sets have the same variances?
NO
Unequal Variance (Independent) T-Test