Introduction to the basics of AI

Session 5

Z. TAIA-ALAOUI

Outline

- Definitions
- T-est
- ANOVA Test
- LDA
- Fisher Score
- Mutual Information
- Implementation

Statistical Tools - Dataset

- **Sample**

$$X_{k \in [1,N]} \in R^p$$

- **Set of samples**

$$\mathbf{X} = \{X_k \in \mathrm{R}^p\}_{k \in [1,\mathrm{N}]}$$

IRIS DATASET

p variables

N samples

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	[illegible]	[illegible]	[illegible]	[illegible]
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5.0	3.4	1.5	0.2

Row = Sample = Observation = Measurement

Statistical Tools - Dataset

- **Set of N samples expressed in a space of p Variables**

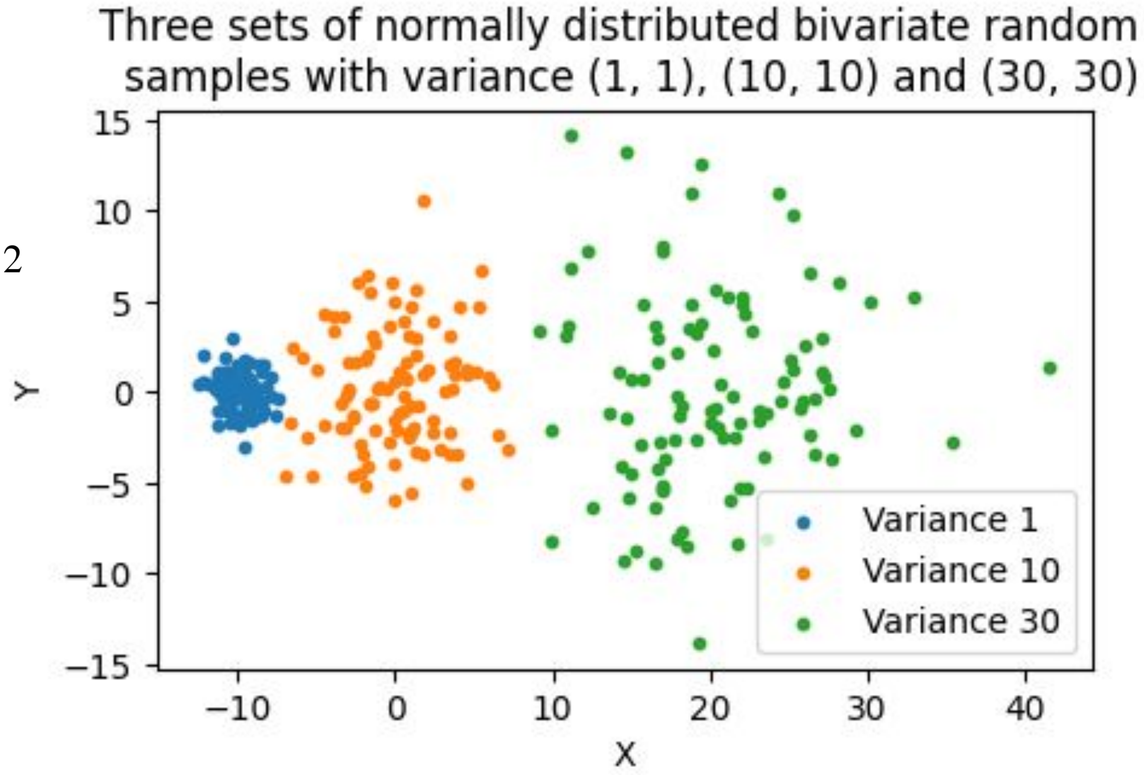
$$
\begin{array}{cc}
 & \begin{array}{ccccc} \phantom{x_{11}} & \phantom{x_{12}} & \phantom{x_{13}} & V_j & \phantom{x_{1p}} \end{array} \\
\mathbf{X} = & \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{N1} & . & . & . & x_{Np} \end{pmatrix} \\
X_i &
\end{array}
= [V_1, V_2, \ldots, V_p] = \begin{bmatrix} X_1^{\mathrm{T}} \\ X_2^{\mathrm{T}} \\ . \\ . \\ . \\ X_N^{\mathrm{T}} \end{bmatrix}
$$

$$
X_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ . \\ . \\ . \\ x_{ip} \end{bmatrix} \qquad V_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ . \\ . \\ . \\ x_{Nj} \end{bmatrix}
$$

Statistical Tools - Variance

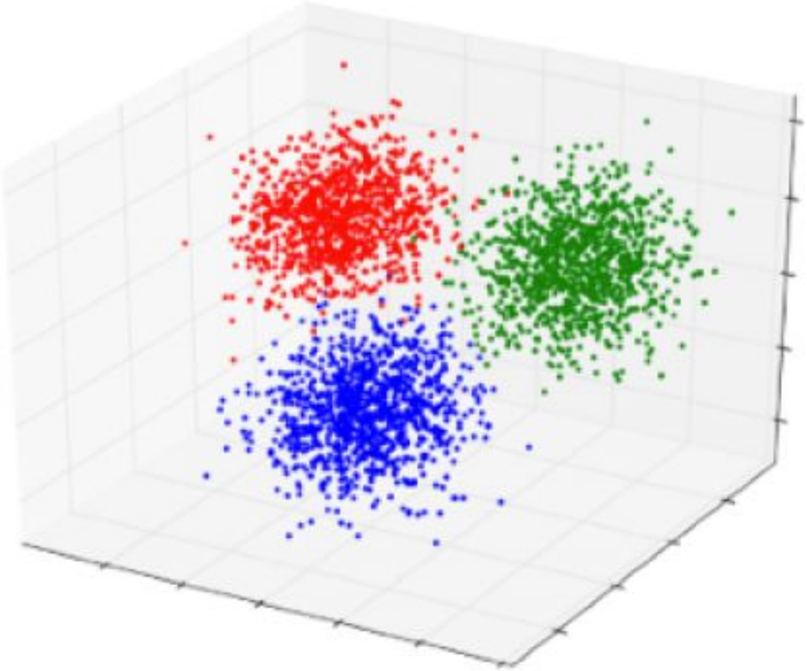
$$Var(V_j) = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_{ij} - \overline{V_j}\right)^2$$

$$\overline{\mathrm{V}}_{\mathrm{j}} = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$$



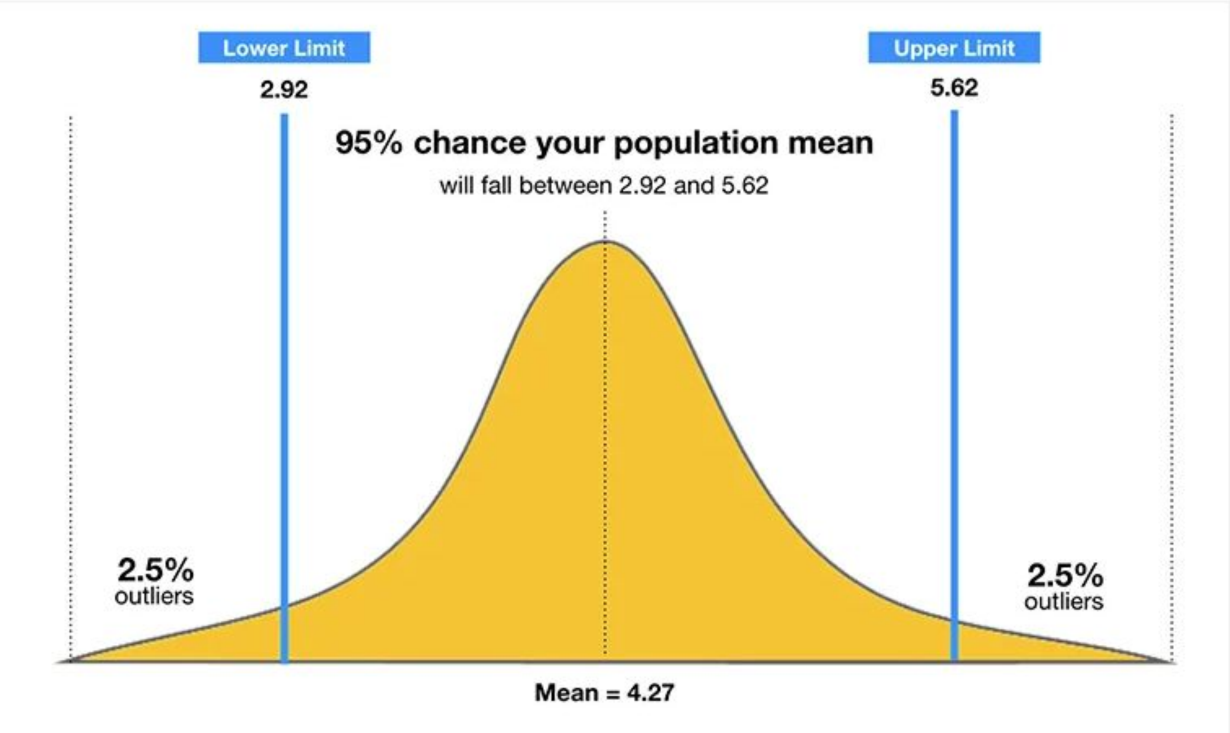
Statistical Tools - Co-Variance

$$Cov\left(\boldsymbol{V}_i,\ \boldsymbol{V}_j \right) = \frac{1}{N-1} \sum_{k=1}^{N} \left(x_{ki} - \overline{V}_i \right)\left(x_{kj} - \overline{V}_j \right)$$



T-Test

- A T-test is used to infer whether two sets of data come from the same distribution.

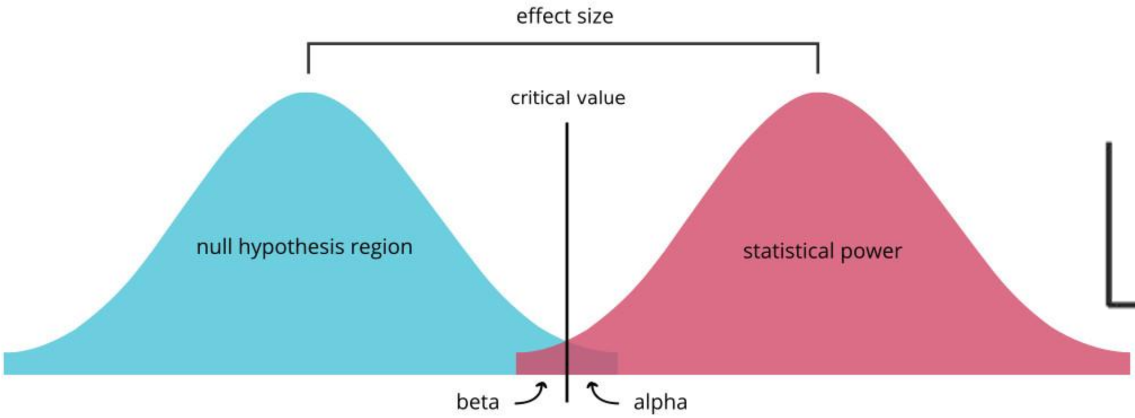


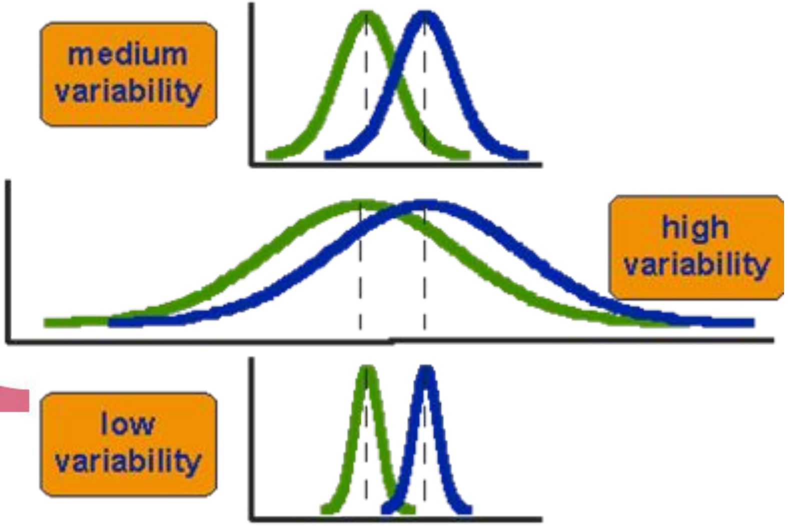
T-Test

- The t-test consists in assessing whether the distance between means of two groups is significant regarding their respective variances.

T-Test

- A high t-value indicates high significance, meaning the two samples come from different distributions.





T-Test

- Calculating a t-test requires three elements:
  - the mean values of the two groups, or the mean difference,
  - the standard deviation of each group,
  - and the number of data samples in each group.

- Results: A value of **t** (degrees of freedom = $n1 + n2 - 2$)

- Interpretation of a t-value is realized using the t-table based on the values of t and df.

T-Test

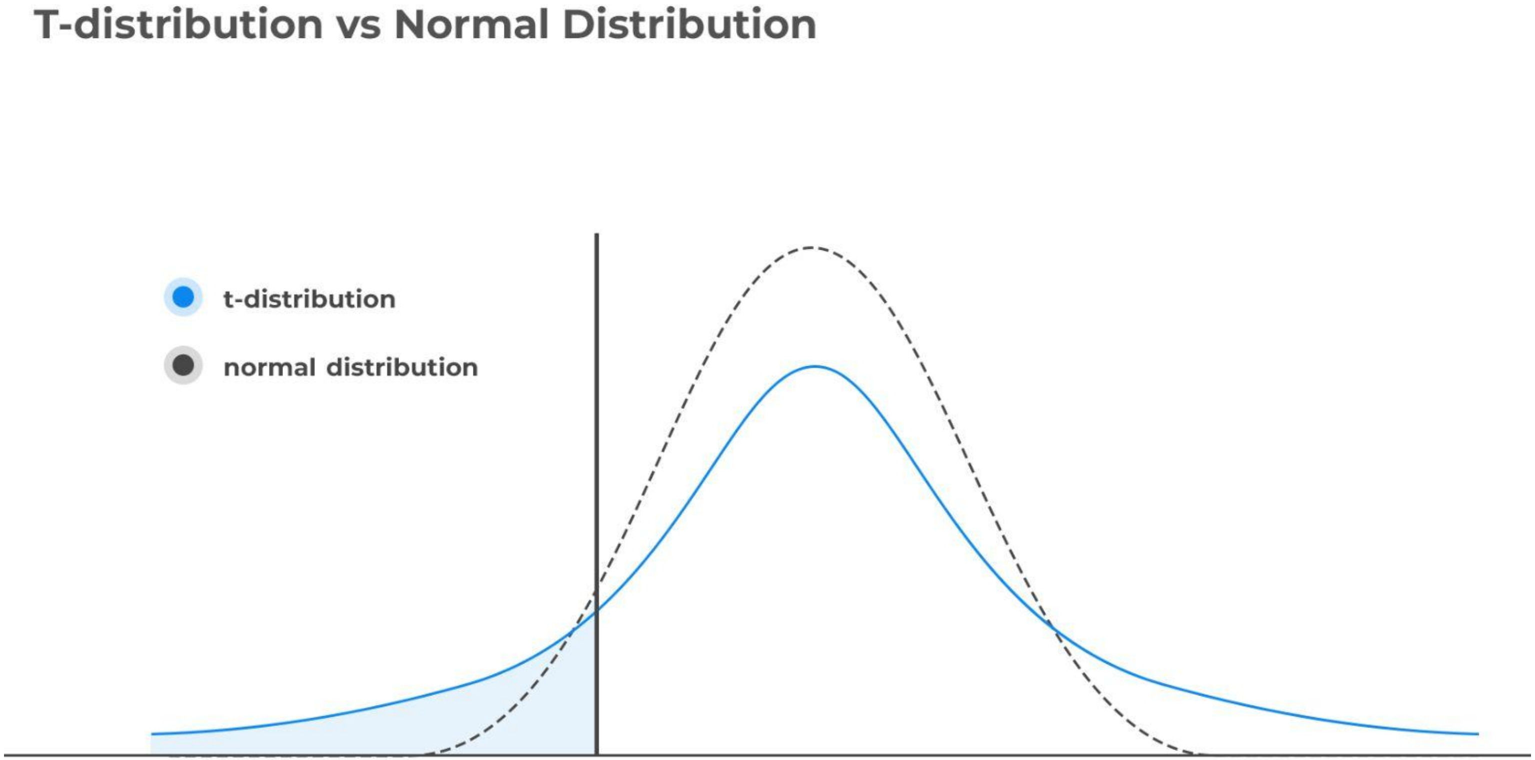
- Equal-Variance Samples/ Pooled T-Test (eq. N° of samples)

$$\text{T-value} = \frac{mean1 - mean2}{\frac{(n1-1)\times var1^2 + (n2-1)\times var2^2}{n1+n2-2} \times \sqrt{\frac{1}{n1} + \frac{1}{n2}}}$$

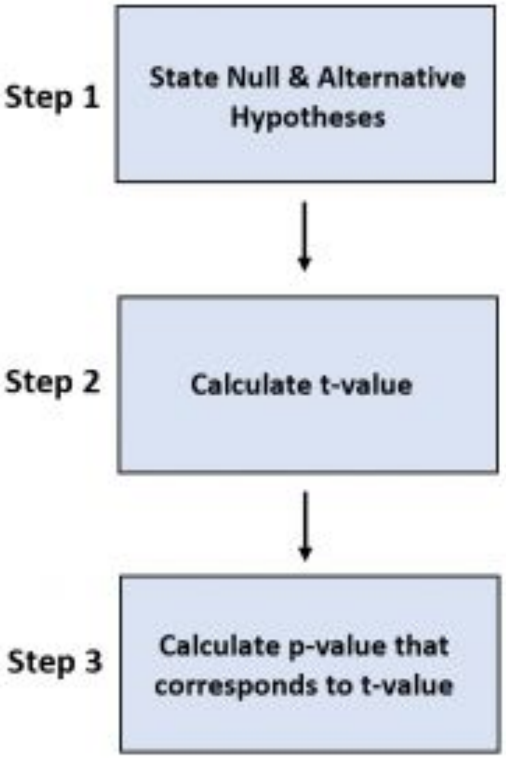
- Welch's t-test (unequal sample size / variances)

$$\text{T-value} = \frac{mean1 - mean2}{\sqrt{\left(\frac{var1}{n1} + \frac{var2}{n2}\right)}}$$

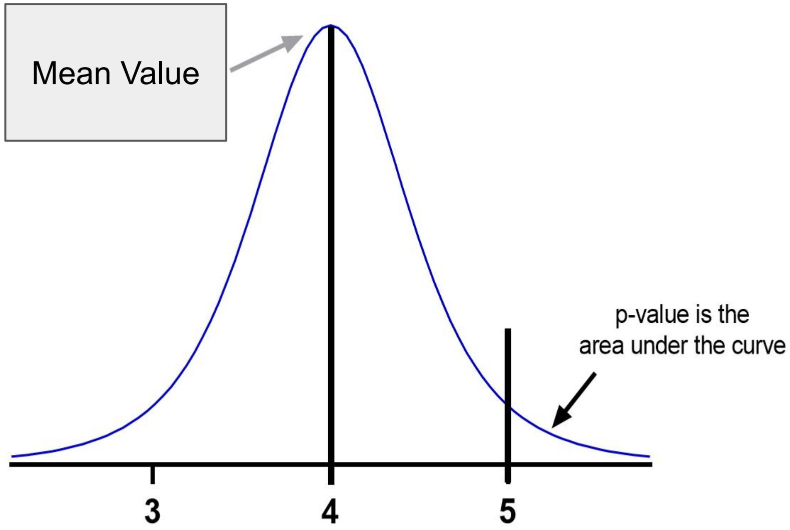
T-Test


T-distribution vs Normal Distribution
t-distribution
normal distribution

T-test



A p-value less than 0.05 is typically considered to be statistically significant, in which case the null hypothesis should be rejected

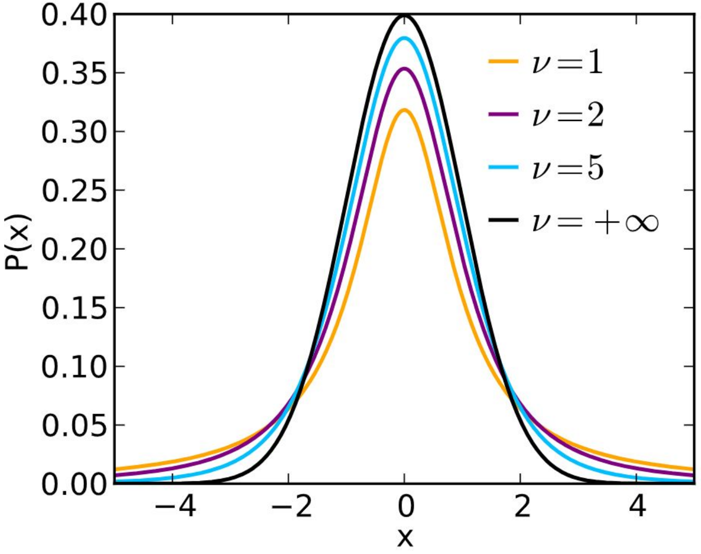


T-test

https://www.statology.org/how-to-calculate-a-p-value-from-a-t-test-by-hand/

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\,\nu}\;\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\,\mathrm{d}t, \qquad \Re(z) > 0$$



t Table

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										

ANOVA

- ANOVA extends the use of the t-test to k>=2 groups

- H0 (Null hypothesis): $\mu 1 = \mu 2 = \mu 3 = \ldots = \mu k$

- H1 (Alternate hypothesis): It states that there will be at least one population mean that differs from the rest

ANOVA

- **Factor**: A variable under consideration that influences an observation (example: The type of flower in regards to sepal length)

- **Level**: A value for this factor (example: 0,1,2 for the type of flower in IRIS dataset)

ANOVA

- **Assumptions**:
  - Normality for each factor level
  - Equal Variances inside each level
  - Variables are drawn independently and randomly from each factor level

MANOVA

$$
X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{N1} & . & . & . & x_{Np} \end{pmatrix}
$$

Number of samples (or levels)	$= k$
Number of observations in ith sample	$= n_i, \quad i = 1, 2, \ldots, k$
Total number of observations	$= n = \sum_i n_i$
Observation j in ith sample	$= x_{ij}, \quad j = 1, 2, \ldots, n_i$
Sum of n_i observations in ith sample	$= T_i = \sum_j x_{ij}$
Sum of all n observations	$= T = \sum_i T_i = \sum_i \sum_j x_{ij}$

MANOVA

$$\mathrm{X} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{N1} & . & . & . & x_{Np} \end{pmatrix}$$

Observation j in ith sample $= x_{ij}, \quad j = 1, 2, \ldots, n_i$

Sum of n_i observations in ith sample $= T_i = \sum_j x_{ij}$

Sum of all n observations $= T = \sum_i T_i = \sum_i \sum_j x_{ij}$

Total sum of squares, $SS_T = \sum_i \sum_j x_{ij}^2 - \frac{T^2}{n}$

Between samples sum of squares, $SS_B = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{n}$

Within samples sum of squares, $SS_W = SS_T - SS_B$

Total Sum of Squares (SST): The SST is the sum of all squared differences between the mean of a sample and the individual values in that sample

MANOVA

Total sum of squares, $$SS_T = \sum_i \sum_j x_{ij}^2 - \frac{T^2}{n}$$

Between samples sum of squares, $$SS_B = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{n}$$

Within samples sum of squares, $$SS_W = SS_T - SS_B$$

Total mean square, $$MS_T = \frac{SS_T}{n-1}$$

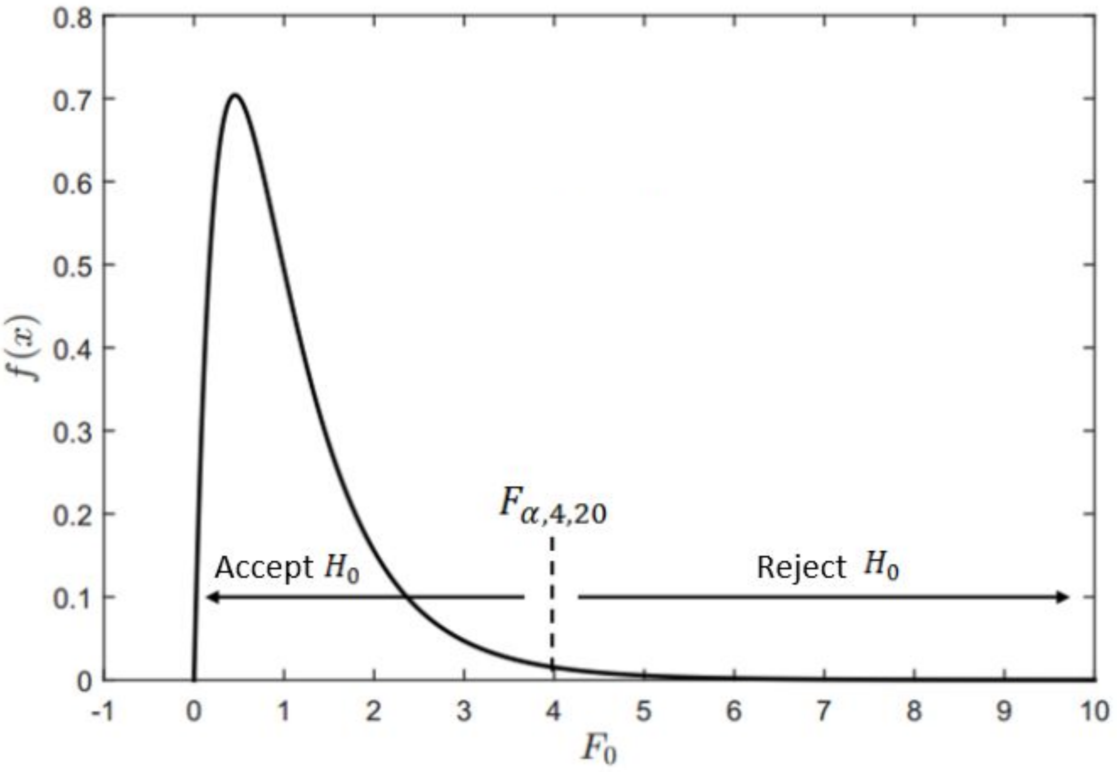
Between samples mean square, $$MS_B = \frac{SS_B}{k-1}$$

Within samples mean square, $$MS_W = \frac{SS_W}{n-k}$$

MANOVA

Source of variation	Sum of squares	Degrees of freedom	Mean square	F ratio
Between samples	SS_B	$k-1$	MS_B	$\frac{MS_B}{MS_W}$
Within samples	SS_W	$n-k$	MS_W	
Total	SS_T	$n-1$		

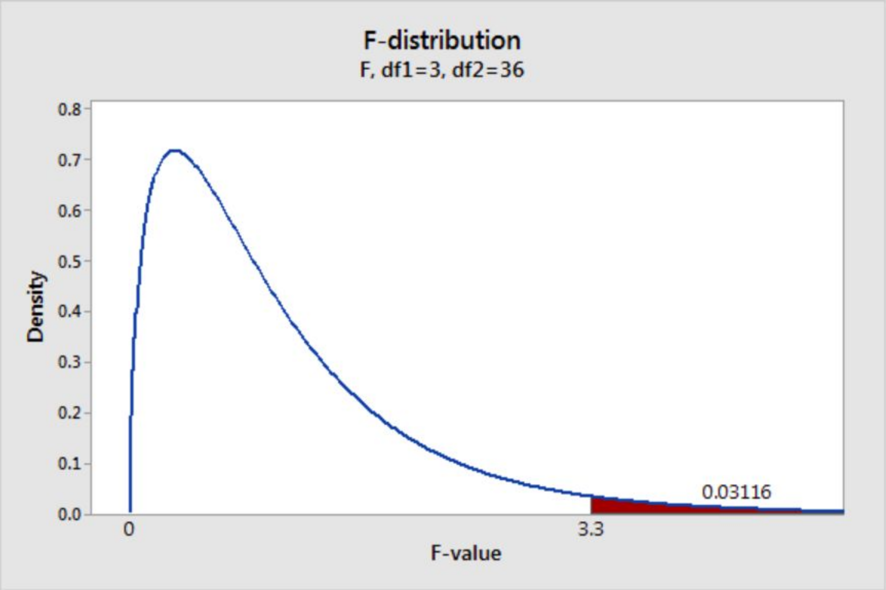
MANOVA



0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
$f(x)$
-1
0
1
2
3
4
5
6
7
8
9
10
F_0
$F_{\alpha,4,20}$
Accept H_0
Reject H_0

MANOVA

- Fisher Test - Critical Value



F - Distribution (α = 0.01 in the Right Tail)

df_2 \ df_1 (Numerator Degrees of Freedom →; Denominator Degrees of Freedom ↓)	1	2	3	4	5	6	7	8	9
1	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158
6	13.745	10.925	9.7795	9.1483	8.7459	8.4661	8.2600	8.1017	7.9761
7	12.246	9.5466	8.4513	7.8466	7.4604	7.1914	6.9928	6.8400	6.7188
8	11.259	8.6491	7.5910	7.0061	6.6318	6.3707	6.1776	6.0289	5.9106
9	10.561	8.0215	6.9919	6.4221	6.0569	5.8018	5.6129	5.4671	5.3511
10	10.044	7.5594	6.5523	5.9943	5.6363	5.3858	5.2001	5.0567	4.9424
11	9.6460	7.2057	6.2167	5.6683	5.3160	5.0692	4.8861	4.7445	4.6315
12	9.3302	6.9266	5.9525	5.4120	5.0643	4.8206	4.6395	4.4994	4.3875
13	9.0738	6.7010	5.7394	5.2053	4.8616	4.6204	4.4410	4.3021	4.1911
14	8.8616	6.5149	5.5639	5.0354	4.6950	4.4558	4.2779	4.1399	4.0297
15	8.6831	6.3589	5.4170	4.8932	4.5556	4.3183	4.1415	4.0045	3.8948
16	8.5310	6.2262	5.2922	4.7726	4.4374	4.2016	4.0259	3.8896	3.7804
17	8.3997	6.1121	5.1850	4.6690	4.3359	4.1015	3.9267	3.7910	3.6822
18	8.2854	6.0129	5.0919	4.5790	4.2479	4.0146	3.8406	3.7054	3.5971
19	8.1849	5.9259	5.0103	4.5003	4.1708	3.9386	3.7653	3.6305	3.5225
20	8.0960	5.8489	4.9382	4.4307	4.1027	3.8714	3.6987	3.5644	3.4567
21	8.0166	5.7804	4.8740	4.3688	4.0421	3.8117	3.6396	3.5056	3.3981
22	7.9454	5.7190	4.8166	4.3134	3.9880	3.7583	3.5867	3.4530	3.3458
23	7.8811	5.6637	4.7649	4.2636	3.9392	3.7102	3.5390	3.4057	3.2986
24	7.8229	5.6136	4.7181	4.2184	3.8951	3.6667	3.4959	3.3629	3.2560
25	7.7698	5.5680	4.6755	4.1774	3.8550	3.6272	3.4568	3.3239	3.2172
26	7.7213	5.5263	4.6366	4.1400	3.8183	3.5911	3.4210	3.2884	3.1818
27	7.6767	5.4881	4.6009	4.1056	3.7848	3.5580	3.3882	3.2558	3.1494
28	7.6356	5.4529	4.5681	4.0740	3.7539	3.5276	3.3581	3.2259	3.1195
29	7.5977	5.4204	4.5378	4.0449	3.7254	3.4995	3.3303	3.1982	3.0920
30	7.5625	5.3903	4.5097	4.0179	3.6990	3.4735	3.3045	3.1726	3.0665
40	7.3141	5.1785	4.3126	3.8283	3.5138	3.2910	3.1238	2.9930	2.8876
60	7.0771	4.9774	4.1259	3.6490	3.3389	3.1187	2.9530	2.8233	2.7185
120	6.8509	4.7865	3.9491	3.4795	3.1735	2.9559	2.7918	2.6629	2.5586
∞	6.6349	4.6052	3.7816	3.3192	3.0173	2.8020	2.6393	2.5113	2.4073

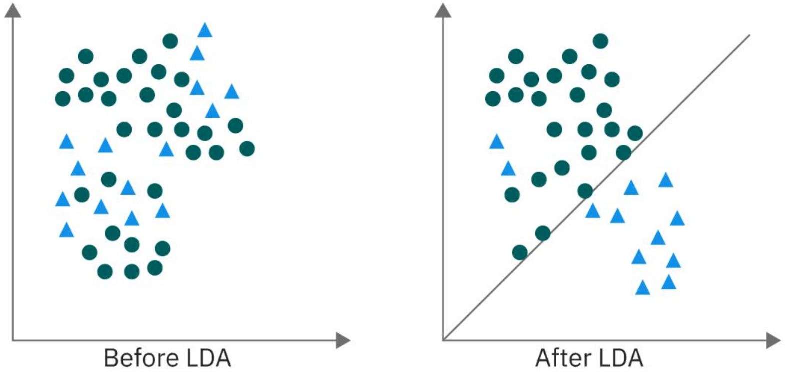
Linear Discriminant Analysis

- LDA is an equivalent for PCA that is well suited for **supervised** classification problems.

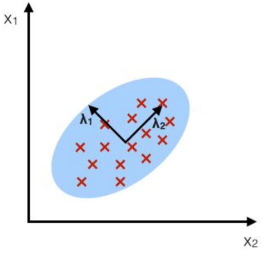
- Unlike ANOVA, LDA has continuous independent variable (measurements) and categorical dependent variables (class labels).

- LDA is a feature **extraction** method that makes linear combination of input features in order to optimize class separability.

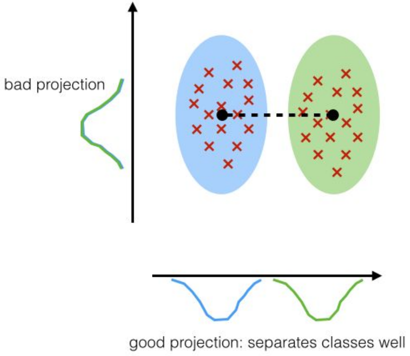
Linear Discriminant Analysis



PCA:
component axes that
maximize the variance



LDA:
maximizing the component
axes for class-separation



Linear Discriminant Analysis

- LDA projects data into a new space in which between-class separation is maximized.
- Separation means maximizing the distance between the projected means and minimizing the projected variance within classes. (Fisher method again !).
- Assumptions:
  - Normal Distribution
  - Covariance Homogeneity

Linear Discriminant Analysis

1. Compute the d-dimensional mean vectors for the different classes from the dataset.
2. Compute the scatter matrices (in-between-class and within-class scatter matrix).
3. Compute the eigenvectors ($\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d$) and corresponding eigenvalues ($\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_d$) for the scatter matrices.
4. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a $d \times k$ dimensional matrix $\boldsymbol{W}$ (where every column represents an eigenvector).
5. Use this $d \times k$ eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the matrix multiplication: $\boldsymbol{Y} = \boldsymbol{X} \times \boldsymbol{W}$ (where $\boldsymbol{X}$ is a $n \times d$-dimensional matrix representing the n samples, and $\boldsymbol{y}$ are the transformed $n \times k$-dimensional samples in the new subspace).

Linear Discriminant Analysis

- Within class scatter matrix

$$S_W = \sum_{i=1}^{c} S_i \qquad S_i = \sum_{\boldsymbol{x} \in D_i}^{n} (\boldsymbol{x} - \boldsymbol{m}_i)\,(\boldsymbol{x} - \boldsymbol{m}_i)^T$$

- Between class scatter matrix

$$S_B = \sum_{i=1}^{c} N_i (\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^T$$

Linear Discriminant Analysis

- Eigenvalue and Eigenvectors computation

$$\boldsymbol{A} = S_W^{-1} S_B$$
$$\boldsymbol{v} = \text{ Eigenvector}$$
$$\lambda = \text{ Eigenvalue}$$

$$\boldsymbol{Av} = \lambda \boldsymbol{v}$$

- Final Projection of data

$$\boldsymbol{Y} = \boldsymbol{X} \times \boldsymbol{W}$$

Feature Selection Based on score thresholds

- **Fisher Score**

$$F(\mathbf{x}^j) = \frac{\sum_{k=1}^{c} n_k(\mu_k^j - \mu^j)^2}{(\sigma^j)^2}$$

μ^j, σ^j are mean and variance of j-th feature

μ_k^j is the mean of the j-th feature for group k

Feature Selection Based on score thresholds

- **Mutual Information**

  - The Mutual Information is a measure of the similarity between two labels of the same data, so the input feature must be first categorized.

  - Qualitatively, entropy is a measure of uncertainty – the higher the entropy, the more uncertain one is about a random variable.

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

$$H(X) = -\sum_{i} p(x_i) \log_2 p(x_i)$$

Sources

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

https://en.wikipedia.org/wiki/Student%27s_t-distribution

https://www.investopedia.com/terms/t/t-test.asp#:~:text=A%20t%2Dtest%20is%20an,flipping%20a%20coin%20100%20times

https://dmn92m25mtw4z.cloudfront.net/img_set/stat-6-6-x-6-article/v1/stat-6-6-x-6-article-1253w.png

https://www.kaggle.com/code/bhagyashree12/anova-test-on-iris-dataset

https://www.automacaodedados.com.br/en/stories/estatistica-em-testes-para-nao-matematicos-parte-5/images/thumbnail.jpg7

https://arxiv.org/html/2404.13664v1/extracted/5549913/TrueClusterPlot.png

https://analystprep.com/cfa-level-1-exam/quantitative-methods/t-distribution-and-degrees-of-freedom/

https://en.wikipedia.org/wiki/F-test

https://www.cimt.org.uk/projects/mepres/alevel/fstats_ch7.pdf