

Table des matières

Chapitre 1 : Rappels sur la statistique et probabilité.....	2
1. La statistique	2
2. Les lois de probabilité.....	3
3. Applications.....	4
Chapitre 2 : Tests d'hypothèses.....	8
1. Estimation.....	8
1. Estimation ponctuelle	8
a. Espérance	8
b. Variance	8
c. Fréquence	9
d. Exemple :.....	9
2. Tests de conformité	10
Comparaison d'une moyenne à une norme	10
a. Principe du test.....	10
b. Variance de la population connue	10
c. Variance de la population inconnue.....	12
d. Comparaison d'une fréquence observée et une fréquence théorique.....	13
3. Tests d'homogénéité.....	15
a. Comparaison de deux moyennes	15
b. Comparaison de deux fréquences	19
Chapitre 3: Analyse de la variance	22
1. Principe.....	22
2. Procédure de résolution.....	22
3. Exemple	23
Chapitre 4 Test d'indépendance de khi deux.....	25
1. Position du problème	25
2. Résolution du problème.....	25
a. Les effectifs marginaux.....	25
b. Le tableau théorique.....	25
c. La distance de Khi deux	25
d. Lecture de la distance limite D.....	25
3. Application	25
Exercice.....	27
Table de Khi deux	28
Chapitre 5 : Regressions	30
Introduction.....	30
Résolution.....	30
a. Calcul de b l'ordonnée à l'origine	30
b. Calcul de la pente a	30
c. Le coefficient de corrélation r	31
d. Interprétation des résultats	31
e. Application	32
Chapitre 6 : Analyse discriminante.....	34
1. Le partitionnement ou Kmeans.....	34
2. Classification Ascendante Hiérarchique(CAH)	38

Chapitre 1 : Rappels sur la statistique et probabilité

1. La statistique

La Médiane	$M_e = a_i + \frac{\frac{1}{2}(n+1) - FACA_{i-1}}{n_i} \alpha_i$
Le Mode	$M_o = a_i + \frac{d_1}{d_1 + d_2} * \alpha_i$
La Moyenne arithmétique	$M = \frac{1}{n} \sum_{i=1}^k n_i x_i = \frac{\sum_{i=1}^k n_i x_i}{\sum_{i=1}^k n_i}$
Moyenne Géométrique	$G = \left[\prod_{i=1}^n x_i \right]^{1/n} = G = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \dots x_k^{n_k}} \Leftrightarrow \ln G = \frac{1}{n} \sum_{i=1}^n n_i \ln x_i$
Moyenne Harmonique	$H = \frac{n}{\sum_{i=1}^n \frac{n_i}{x_i}}$
Moyenne Quadratique	$Q = \sqrt{\frac{1}{n} \sum_{i=1}^n n_i x_i^2}$
La Variance	$V(x) = \sigma^2(x) = \frac{1}{n} \sum n_i (x_i - \bar{x})^2 = q^2 - m^2$
La covariance	$COV(X, Y) = \frac{\sum_i^n (X_i - \bar{X}) * (Y_i - \bar{Y})}{\partial_x * \partial_y}$
La pente	$a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\text{cov}(x, y)}{\text{var}(x)}$
L'ordonnée à l'origine	$b = \bar{y} - a\bar{x}$
Coefficient de corrélation	$r = \frac{\text{cov}(x, y)}{\delta_x * \delta_y} ; r \text{ est compris entre } -1 \text{ et } 1.$

2. Les lois de probabilité

LOIS DISCRETES	Nom	Loi	Notation	E(x)	V(x)
	Uniforme discrete	$P(x=k) = \frac{1}{n}$	U D	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
	Bernoulli	$P(x=1) = p$ $p(x=0) = q$	B(1,p)	p	pq
	Binomiale	$P(x=k) = C_n^k p^k q^{n-k}$	B(n,p)	np	npq
	Hypergéométrique	$P(x=k) = \frac{C_{n1}^k C_{N-n1}^{n-k}}{C_N^n}$	H(N,n,p)	np	$npq \cdot \frac{N-n}{N-1}$
	Poisson	$P(x=k) = e^{-\lambda} \frac{\lambda^k}{k!}$	P(λ)	λ	λ
	Géométrique	$P(x=k) = p q^{k-1}$	G(p)	$\frac{1}{p}$	$\frac{q}{p^2}$
LOIS CONTINUES	Uniforme Continue	$f(x) = \frac{1}{b-a} 1_{[a,b]}(x)$	UC	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$
	Normale	$f_{(\mu, \sigma)}(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	N(m, δ)	m	δ
	Normale Centrée réduite	$f_{(0,1)}(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$	N(0,1)	0	1
	Exponentielle	$a e^{-ax}$ si $x \geq 0$ et $a > 0$	E(a)	$\frac{1}{a}$	$\frac{1}{a^2}$
	Khi deux	$\sum_1^n \chi^2$	χ_n^2	η	2η
	Student Fisher	$\frac{Z}{\sqrt{\chi^2}}$	T _n	0	$\frac{n}{n-2}$
	Fisher Snedecor	$\frac{\frac{X}{n}}{\frac{Y}{m}}$	F _(n,m)	$\frac{m}{m-2}$	$\frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$

3. Applications

Série1

xi	ni	FACA	FACD	nixi	nixi ²	ni ln(xi)	ni/xi	ni xi-m
1	5							
2	5							
3	10							
4	5							
5	5							
Total								

Mo=		G=		M=		Q=		
Me=		H=		V(x)=		σ(x)=		

Série 2

xi	ni	FACA	FACD	nixi	nixi ²	ni ln(xi)	ni/xi	ni xi-m
1	380							
2	455							
3	245							
4	230							
5	100							
6	75							
7	10							
8	5							
Total								

Mo=		G=		M=		Q=		
Me=		H=		V(x)=		σ(x)=		

Série 3

salaire F/h	xi	ni	nixi	nixi ²	ni/xi	ni ln(xi)	FACA	ni xi-m
1200-1250	1 225	20						
1250-1300	1 275	10						
1300-1350	1 325	32						
1350-1400	1 375	25						
1400-1450	1 425	8						
1450-1500	1 475	5						
	Total							

Mo=		G=		M=		Q=		
Me=		H=		V(x)=		σ(x)=		

Série 4

bi	xi	ni	xini	nixi ²	ni/xi	ni ln(xi)	Faca	ni xi-m
36,5-37,5	37	3						
37,5-38,5	38	7						
38,5-39,5	39	17						
39,5-40,5	40	18						
40,5-41,5	41	9						
41,5-42,5	42	4						
42,5-43,5	43	2						
Total								

Mo=		G=		M=		Q=		
Me=		H=		V(x)=		σ(x)=		

Partie probabilité

Loi Binomiale n=4 p=1/6

XI	PI	XIPI	XI ² PI
0			
1			
2			
3			
4			
Total			

E(x)=

V(x)=

Loi Hypergéométrique N=10; n1=6 ; n2=4 et n=5

XI	PI	XIPI	XI ² PI
0			
1			
2			
3			
4			
5			
Total			

E(x)=

V(x)=

Loi de Poisson

$\lambda=2$

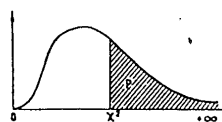
XI	PI	XIPI	XI ² PI
0			
1			
2			
3			
4			
5			
6			
7			
8			
Total			

E(x)=

V(x)=

TABIE 4 Distribution de χ^2 (Loi de K. Pearson).

Valeur de χ^2 ayant la probabilité P d'être dépassée

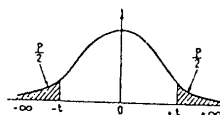


ν	$P = 0,90$	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01
1	0,0158	0,0642	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635
2	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,879	9,210
3	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,345
4	1,064	1,649	2,195	3,357	4,778	5,989	7,779	9,488	11,668	13,277
5	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,086
6	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812
7	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,662	18,475
8	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090
9	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666
10	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209
11	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725
12	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217
13	7,042	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,472	27,688
14	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141
15	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578
16	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	29,633	32,000
17	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	30,995	33,409
18	10,865	12,857	14,430	17,338	20,601	22,760	25,989	28,869	32,346	34,805
19	11,651	13,716	15,332	18,338	21,689	23,900	27,204	30,144	33,687	36,191
20	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,410	35,020	37,566
21	13,240	15,445	17,182	20,337	23,858	26,171	29,615	32,671	36,343	38,932
22	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	37,659	40,289
23	14,848	17,187	19,021	22,337	26,018	28,429	32,007	35,172	38,968	41,638
24	15,659	18,062	19,943	23,337	27,096	29,553	33,196	36,415	40,270	42,980
25	16,473	18,940	20,867	24,337	28,172	30,675	34,382	37,652	41,566	44,314
26	17,292	19,820	21,792	25,336	29,246	31,795	35,563	38,885	42,856	45,642
27	18,114	20,703	22,719	26,336	30,319	32,912	36,741	40,113	44,140	46,963
28	18,939	21,588	23,647	27,336	31,391	34,027	37,916	41,337	45,419	48,278
29	19,768	22,475	24,577	28,336	32,461	35,139	39,087	42,557	46,693	49,588
30	20,599	23,364	25,508	29,336	33,530	36,250	40,256	43,773	47,962	50,892

Nota. — ν est le nombre de degrés de liberté.
 Pour ν compris entre 30 et 100, on admettra que $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ est approximativement distribué suivant la loi normale centrée réduite.
 Pour ν supérieur à 100, on admettra que $(\chi^2 - \nu)/\sqrt{2\nu}$ est approximativement distribué suivant la loi normale centrée réduite.

TABIE 5 Distribution de Student-Fisher.

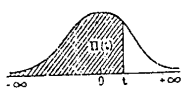
Valeur de t ayant la probabilité P d'être dépassée en module.



ν	$P = 0,90$	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01
1	1,0158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657
2	2,0142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925
3	3,0137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841
4	4,0134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604
5	5,0132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032
6	6,0131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707
7	7,0130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499
8	8,0130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355
9	9,0129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250
10	10,0129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169
11	11,0129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,166
12	12,0128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,035
13	13,0128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012
14	14,0128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977
15	15,0128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947
16	16,0128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921
17	17,0128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,896
18	18,0127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878
19	19,0127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861
20	20,0127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845
21	21,0127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831
22	22,0127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819
23	23,0127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807
24	24,0127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797
25	25,0127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787
26	26,0127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779
27	27,0127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771
28	28,0127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,761
29	29,0127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,751
30	30,0127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750
∞	0,12566	0,25335	0,38532	0,52440	0,67449	0,84162	1,03643	1,28155	1,64485	1,99966	2,32634	2,57582

Nota. — ν est le nombre de degrés de liberté

250



TABIE 2 Fonction de répartition de la loi de Laplace-Gauss.

Probabilité d'une valeur inférieure à t :

$$\Pi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-x^2/2} dx$$

<i>t</i>	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7021	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7290	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9779	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Chapitre 2 : Tests d'hypothèses

1. Estimation

L'**estimation** d'un paramètre quelconque θ est **ponctuelle** si l'on associe **une seule valeur** à l'estimateur $\hat{\theta}$ à partir des données observables sur un échantillon aléatoire.

L'**estimation par intervalle** associe à un échantillon aléatoire, **un intervalle** $[\hat{\theta}_1, \hat{\theta}_2]$ qui recouvre θ avec une certaine probabilité.

1. Estimation ponctuelle

a. Espérance

Soit X une variable aléatoire continue suivant une loi normale $N(m, \sigma)$ dont la valeur des paramètres n'est pas connue et pour laquelle on souhaite estimer **l'espérance m** .

Soient $X_1, X_2, \dots, X_i, \dots, X_n$, n réalisations indépendantes de la variable aléatoire X , un estimateur du paramètre m est une suite de variable aléatoire $\hat{\theta}$ fonctions des X_i :

$$\hat{\theta} = f(X_1, X_2, \dots, X_i, \dots, X_n)$$

La **moyenne arithmétique** constitue le meilleur estimateur de m , espérance de la loi de probabilité de la variable aléatoire X :

$$m = \bar{x} = \frac{\sum n_i x_i}{n}$$

b. Variance

Soit X une variable aléatoire continue suivant une loi normale $N(m, \sigma)$ pour laquelle on souhaite estimer **la variance σ^2** .

Soient $X_1, X_2, \dots, X_i, \dots, X_n$, n réalisations indépendantes de la variable aléatoire X , un estimateur du paramètre σ^2 est une suite de variable aléatoire $\hat{\theta}$ fonctions des X_i :

• Cas où l'espérance m est connue

La **variance observée** constitue le meilleur estimateur de σ^2 , variance de la loi de probabilité de la variable aléatoire X lorsque **m est connue** :

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2$$

- Cas où l'espérance m est inconnue

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

c. Fréquence

Soit le schéma de Bernoulli dans lequel le caractère A correspond au succès. On note p la fréquence des individus de la **population** possédant le caractère A . La valeur de ce paramètre étant inconnu, on cherche à estimer la fréquence p à partir des données observables sur un échantillon.

A chaque échantillon non exhaustif de taille n , on associe l'entier k , nombre d'individus possédant le caractère A .

Soit K une variable aléatoire discrète suivant une loi binomiale $B(n, p)$ et pour laquelle on souhaite estimer **la fréquence p** .

La **fréquence observée** du nombre de succès observé dans un échantillon de taille n constitue le meilleur estimateur de p :

$$P = \frac{k}{n}$$

d. Exemple :

On a prélevé au hasard, dans une population de lapin, 100 individus. Sur ces 100 lapins, 20 sont atteints par la myxomatose. Le pourcentage de lapins atteints par la myxomatose dans la population est donc :

$$P = \frac{k}{n} = \frac{20}{100} = 0,2 \text{ soit } 20\% \text{ de lapins atteints dans la population}$$

Ce résultat n'aura de signification que s'il est associé à **un intervalle de confiance**.

2. Tests de conformité

Les tests de conformité sont destinés à **vérifier si un échantillon peut être considéré comme extrait d'une population donnée ou représentatif de cette population**, vis-à-vis d'un paramètre comme la moyenne, la variance ou la fréquence observée. Ceci implique que la **loi théorique du paramètre est connue au niveau de la population**.

En théorie, si l'on suppose connu la valeur θ_0 d'un paramètre relatif à la population

(par exemple p, μ, σ^2) et $\hat{\theta}$ un **estimateur absolument correct** de θ (par exemple $\frac{k}{n}, (\bar{x}, s^2)$) obtenu à partir d'un **échantillon aléatoire simple de taille n** , on cherche à savoir si l'échantillon est représentatif de la population pour le paramètre considéré.

$$H_0: \theta = \theta_0$$

Comparaison d'une moyenne à une norme

a. Principe du test

Soit X , une variable aléatoire observée sur une population, suivant **une loi normale** et un **échantillon** extrait de cette population.

Le but est de savoir si un **échantillon de moyenne \bar{X} , estimateur** de m , appartient à une **population de référence connue d'espérance μ_0 (H_0 vraie)** et ne diffère de μ_0 que par des fluctuations d'échantillonnage ou bien appartient à une **autre population inconnue d'espérance m (H_1 vraie)**.

Pour tester cette hypothèse, il existe deux statistiques : la variance σ_0^2 de la population de référence est connue ou cette variance est inconnue et il faut l'estimer.

b. Variance de la population connue

Statistique du test

Soit \bar{X} la **distribution d'échantillonnage** de la moyenne dans la population inconnue suit

une loi normale telle que : $\bar{X} \rightarrow N(m, \sqrt{\frac{\sigma^2}{n}})$.

La statistique étudiée est l'écart : $S = \bar{X} - \mu_0$ dont la distribution de probabilité est la suivante

$$S \rightarrow N(0, \sqrt{\frac{\sigma^2}{n}}) \quad \text{avec sous } H_0, E(S) = 0 \text{ et } V(S) = \frac{\sigma^2}{n}$$

Nous pouvons établir grâce au **théorème central limite** la variable Z centrée réduite telle que

$$Z = \frac{S - E(S)}{\sqrt{V(S)}} = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

Sous $H_0: \mu = \mu_0$ avec σ^2 connue

$$Z = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \quad \text{suit une loi normale centrée réduite } N(0,1)$$

Application et Décision

L'hypothèse testée est la suivante :

$$H_0: \mu = \mu_0 \quad \text{contre} \quad H_1: \mu \neq \mu_0$$

Une valeur z de la variable aléatoire Z est calculée :

$$z = \frac{|\bar{x} - \mu_0|}{\sqrt{\frac{\sigma^2}{n}}} \quad \text{notée aussi } T_0 \text{ est comparée avec la valeur } T_\alpha \text{ lue dans la}$$

table de la loi normale centrée réduite pour un risque d'erreur α fixé.

- si $T_0 > T_\alpha$ l'hypothèse H_0 est rejetée au risque d'erreur α : l'échantillon appartient à une population d'espérance μ et n'est pas représentatif de la population de référence d'espérance μ_0 .

- si $T_0 \leq T_\alpha$ l'hypothèse H_0 est acceptée: l'échantillon est représentatif de la population de référence d'espérance μ_0 .

Exemple :

La **glycémie** d'une population suit une loi normale d'espérance $\mu_0 = 1\text{g/l}$ et d'écart-type $\sigma_0 = 0,1\text{ g/l}$.

On relève les glycémies chez 9 patients. On trouve $\bar{x} = 1,12\text{g/l}$. Cet échantillon est-il représentatif de la population ?

c. Variance de la population inconnue

Statistique du test

La démarche est la même que pour le test précédent mais la variance de la population n'étant pas connue, elle est estimée par :

La statistique étudiée est l'écart : $S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$ dont la distribution de probabilité est la suivante

$$S \rightarrow N\left(0, \sqrt{\frac{\hat{\sigma}^2}{n}}\right) \quad \text{avec} \quad E(S) = 0 \quad \text{et} \quad V(S) = \frac{\hat{\sigma}^2}{n}$$

Nous pouvons établir grâce au **théorème central limite** la variable T centrée réduite telle que

$$T_0 = \frac{\bar{X} - m_0}{\sqrt{\frac{S^2}{n}}}$$

Application et Décision

L'hypothèse testée est la suivante :

$$H_0: \mu = \mu_0 \quad \text{contre} \quad H_1: \mu \neq \mu_0$$

Une valeur t de la variable aléatoire T est calculée :

$$T_0 = \frac{\bar{X} - m_0}{\sqrt{\frac{S^2}{n}}}$$

T_0 calculée (t_{obs}) est comparée avec la valeur t_{seuil} lue dans **la table de Student** pour un risque d'erreur α fixé et $(n-1)$ **degrés de liberté**.

- si $t_{\text{obs}} > t_{\text{seuil}}$ l'hypothèse H_0 est rejetée au risque d'erreur α : l'échantillon appartient à une population d'espérance μ et n'est pas représentatif de la population de référence d'espérance μ_0 .
- si $t_{\text{obs}} \leq t_{\text{seuil}}$ l'hypothèse H_0 est acceptée: l'échantillon est représentatif de la population de référence d'espérance μ_0 .

Remarque : Si $n < 30$, la variable aléatoire X étudiée doit **impérativement** suivre **une loi normale $N(\mu, \sigma)$** . Pour $n \geq 30$, la variable de **student t converge vers** **une loi normale centrée réduite ε** .

Exemple :

Pour étudier un lot de fabrication de comprimés, on prélève au hasard 10 comprimés parmi les 30 000 produits et on les pèse. On observe les valeurs de poids en grammes :

0,81 – 0,84 – 0,83 – 0,80 – 0,85 – 0,86 – 0,85 – 0,83 – 0,84 – 0,80

Le poids moyen observé est-il compatible avec la valeur 0,83g, moyenne de la production au seuil 98% ?

d. Comparaison d'une fréquence observée et une fréquence théorique

1. Principe du test

Soit X une **variable qualitative prenant deux modalités** (succès $X=1$, échec $X=0$) observée sur une population et **un échantillon** extrait de cette population.

Le but est de savoir si un **échantillon** de fréquence observée $\frac{k}{n}$, **estimateur** de p , appartient à une **population de référence connue de fréquence p_0 (H_0 vraie)** ou à une autre **population inconnue de fréquence p (H_1 vraie)**.

Statistique du test

La distribution d'échantillonnage de la fréquence de succès dans la population inconnue,

$\frac{k}{n}$ suit une loi normale telle que : $\frac{k}{n}$ suit $N(p, \sqrt{\frac{p_0 q_0}{n}})$, les variances étant supposées égales dans la population de référence et la population d'où est extrait l'échantillon.

La statistique étudiée est l'écart : $S = \frac{k}{n} - p_0$ dont la distribution de probabilité est la

suivante $S \sim N(0, \sqrt{\frac{p_0 q_0}{n}})$ avec sous H_0 $E(S) = 0$ et $V(S) = \frac{p_0 q_0}{n}$

On établit la variable Z centrée réduite telle que

$$T_0 = \frac{\frac{k}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \quad \text{mais seulement si } np_0 \text{ et } nq_0 \geq 10$$

Sous $H_0: p = p_0$ et $T_0 = \frac{\frac{k}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$ suit une **loi normale centrée réduite** $N(0,1)$

Application et décision

L'hypothèse testée est la suivante :

$H_0: p = p_0$ contre $H_1: p \neq p_0$ Une valeur z de la variable aléatoire Z est

calculée : $z = \frac{\left| \frac{k}{n} - p_0 \right|}{\sqrt{\frac{p_0 q_0}{n}}}$ notée aussi T_0

- si $T_0 > T_{\text{seuil}}$ l'hypothèse H_0 est rejetée au risque d'erreur α : l'échantillon appartient à une population de fréquence p et n'est pas représentatif de la population de référence de fréquence p_0 .

- si $T_0 \leq T_\alpha$ l'hypothèse H_0 est acceptée: l'échantillon est représentatif de la population de référence de fréquence p_0 .

Exemple :

Une anomalie génétique touche au Gabon 1/1000 des individus. On a constaté dans une région donnée : 57 personnes atteintes sur 50 000 naissances.

Cette région est-elle représentative du Gabon entier ?

La **fréquence de l'anomalie** génétique est

$$p_0 = 0,0010 \text{ au Gabon}$$

La **fréquence observée** dans la région étudiée est

$$\hat{p} = \frac{k}{n} = \frac{57}{50000} = 0,00114$$

Cette région est-elle représentative du Gabon entier ?

Hypothèse : $H_0 : p = p_0$ et $H_1 : p \neq p_0$

Conditions : $X \rightarrow B(50\,000 ; 0,001)$ donc np et $nq > 10$

$$\text{Statistique : } \varepsilon_{obs} = \frac{\left| \frac{k}{n} - p_0 \right|}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{|0,00114 - 0,0010|}{\sqrt{\frac{0,001 * 0,999}{50000}}} = 0,99$$

Décision :

Avec un risque d'erreur $\alpha = 0,05$, $T_\alpha = 1,96$, donc

$\varepsilon_{obs} < \varepsilon_{seuil}$ et donc **H_0 ne peut être rejetée. On accepte donc l'hypothèse H_0 .**

La région considérée est **représentative** du Gabon entier en ce qui concerne la fréquence de cette anomalie génétique.

3. Tests d'homogénéité

Les tests d'homogénéité ou d'égalité destinés à comparer deux populations à l'aide d'un nombre équivalent d'échantillons sont les plus couramment utilisés. Dans ce cas **la loi théorique du paramètre étudié** (par exemple p , μ , σ^2) est **inconnue au niveau des populations étudiées**.

a. Comparaison de deux moyennes

1 Principe du test

Soit X un **caractère quantitatif continu** observé sur 2 populations suivant une **loi normale** et **deux échantillons indépendants** extraits de ces deux populations.

On fait l'hypothèse que les deux échantillons proviennent de 2 populations dont les **espérances sont égales**.

Il existe plusieurs statistiques associées à la comparaison de deux moyennes en fonction de la nature des données.

2 Les variances des populations sont connues

Statistique du test

Sous $H_0: \mu_1 = \mu_2$ avec σ_1^2 et σ_2^2 connues

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ suit une loi normale centrée réduite } N(0,1)$$

Application et décision

L'hypothèse testée est la suivante :

$$H_0: \mu_1 = \mu_2 \text{ contre } H_1: \mu_1 \neq \mu_2$$

Une valeur z de la variable aléatoire Z est calculée :

$$T_0 = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- si $T_0 > T_\alpha$, l'hypothèse H_0 est rejetée au risque d'erreur α : les deux échantillons sont extraits de deux populations ayant des espérances respectivement μ_1 et μ_2 .
- si $T_0 \leq T_\alpha$, l'hypothèse H_0 est acceptée: les deux échantillons sont extraits de deux populations ayant même espérance μ .

Remarque : Pour l'application de ce test, il est impératif que $\mathbf{X} \rightarrow \mathbf{N}(\mu, \sigma)$ pour les échantillons de **taille < 30** et que les deux échantillons soient **indépendants**.

Exemple :

On a effectué une étude, en milieu urbain et en milieu rural, sur le rythme cardiaque humain :

Effectif de l'échantillon	300	240
Moyenne de l'échantillon	80	77
Variance de la population	150	120

Peut-on affirmer qu'il existe une différence significative entre les rythmes cardiaques moyens des deux populations ?

3 Les variances des populations sont inconnues et égales

Statistique du test

- Les variances des populations **n'étant pas connues**, on fait l'hypothèse que les deux populations présentent la même variance.

Sous $H_0: \mu_1 = \mu_2$ avec $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$T = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \text{ suit une loi de Student à } (n_1 + n_2 - 2) \text{ degrés de liberté}$$

Application et décision

L'hypothèse testée est la suivante :

$$H_0: \mu_1 = \mu_2 \quad \text{contre} \quad H_1: \mu_1 \neq \mu_2$$

Les variances des populations n'étant pas connues, l'égalité des variances doit être vérifiée

$$H_0: \sigma_1^2 = \sigma_2^2 \quad \text{contre} \quad H_1: \sigma_1^2 \neq \sigma_2^2 \quad \text{test de Fisher-Snedecor.}$$

Une valeur t de la variable aléatoire T est calculée :

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad \text{avec} \quad \hat{\sigma}^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \quad \text{estimation de la variance } \sigma^2 \text{ commune } t$$

calculée (t_{obs}) est comparée avec la valeur t_{seuil} lue dans **la table de Student**

pour un risque d'erreur α fixé et $(n_1 + n_2 - 2)$ degrés de liberté.

- si $t_{\text{obs}} > t_{\text{seuil}}$ l'hypothèse H_0 est rejetée au risque d'erreur α : les deux échantillons sont extraits de deux populations ayant des espérances respectivement μ_1 et μ_2 .
- si $t_{\text{obs}} \leq t_{\text{seuil}}$ l'hypothèse H_0 est acceptée: les deux échantillons sont extraits de deux populations ayant même espérance μ .

Exemple :

Dans le but d'étudier l'influence du type d'atmosphère d'élevage sur la durée de développement des drosophiles femelles, ces dernières ont été élevées à 14°C sous atmosphère normale (N) ou enrichie en CO₂(CO2). Les resultants savants ont été obtenus :

N	864, 768, 912, 804, 924, 984, 888, 816, 840, 936, 792, 876
CO₂	840, 948, 936, 1032, 912, 948, 1020, 936, 1056, 876, 1032, 918

Que peut-on conclure ?

4 Les variances des populations sont inconnues et inégales

Si les variances des populations **ne sont pas connues** et si leurs estimations à partir des échantillons sont **significativement différentes (test de comparaison des variances)**, il faut considérer deux cas de figure selon la taille des échantillons comparés :

les **grands échantillons** avec $n_1 + n_2$ supérieurs à 30.

les **petits échantillons** avec n_1 et/ou n_2 inférieurs à 30.

Cas où $n_1 + n_2 > 30$

La statistique utilisée est la même que pour le cas où les **variances sont connues**.

Sous $H_0: \mu_1 = \mu_2$

$$T_0 = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ suit une } \mathbf{loi normale centrée réduite} \ N(0,1)$$

Comme les variances sont inconnues et significativement différentes $\sigma_1^2 \neq \sigma_2^2$, on remplace les variances des populations par leurs estimations ponctuelles calculées à partir des

échantillons, $\hat{\sigma}_1^2 = \frac{n_1}{n_1 - 1} s_1^2$ et $\hat{\sigma}_2^2 = \frac{n_2}{n_2 - 1} s_2^2$

L'hypothèse testée est la suivante :

$$H_0: \mu_1 = \mu_2 \text{ contre } H_1: \mu_1 \neq \mu_2$$

Une valeur z de la variable aléatoire Z est calculée :

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

Exemple :

Dans le but d'étudier l'influence éventuelle de la lumière sur la croissance du poisson **Lebistes Reticulus**, on a élevé deux lots de ce poisson dans des conditions d'éclairage différentes. Au 95^{ème} jour, on a mesuré en mm les longueurs x_i des poissons. On a obtenu les résultats suivants :

Lot 1 (180 individus) : éclairage à 400 lux $\sum x_{i1} = 3\,780$ $\sum x_{i1}^2 = 84\,884$

Lot 2 (90 individus) : éclairage à 3 000 lux. $\sum x_{i2} = 2\,043$ $\sum x_{i2}^2 = 46\,586$

Que peut-on conclure ?

Cas où n_1 et/ou $n_2 < 30$

Lorsque les variances sont **inégales** et les échantillons de **petites tailles**, la loi de probabilité suivie par $\bar{X}_1 - \bar{X}_2$ n'est pas connue. On a recours alors aux statistiques non paramétriques.

b. Comparaison de deux fréquences

1 Principe du test

Soit X une variable qualitative prenant deux modalités (succès $X=1$, échec $X=0$) observée sur 2 populations et **deux échantillons indépendants** extraits de ces deux populations. On fait l'hypothèse que les deux échantillons proviennent de 2 populations dont les **probabilités de succès sont identiques**.

Le problème est de savoir si la différence entre les deux fréquences observées est réelle ou explicable par les fluctuations d'échantillonnage. Pour résoudre ce problème, deux tests de comparaison de fréquences sont possibles :

2 Statistique du test

- La distribution d'échantillonnage de la fréquence de succès dans la population 1, $\frac{K_1}{n_1}$ suit une loi normale telle que :

$\frac{K_1}{n_1}$ suit $N(p_1, \sqrt{\frac{p_1 q_1}{n_1}})$ et de même pour $\frac{K_2}{n_2}$ suit $N(p_2, \sqrt{\frac{p_2 q_2}{n_2}})$
si et seulement si $n_1 p_1, n_1 q_1, n_2 p_2, n_2 q_2 > 10$

Sous $H_0: p_1 = p_2$ avec $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

$$Z = \frac{\left(\frac{K_1}{n_1} - \frac{K_2}{n_2} \right)}{\sqrt{pq \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \text{ suit une loi normale centrée réduite } N(0,1)$$

.3 Application et décision

La valeur p , probabilité du succès commune aux deux populations n'est en réalité **pas connue**. On l'estime à partir des résultats observés sur les deux échantillons :

$$\hat{p} = \frac{k_1 + k_2}{n_1 + n_2} \text{ où } k_1 \text{ et } k_2 \text{ représentent le nombre de succès observés respectivement pour l'échantillon 1 et pour l'échantillon 2.}$$

L'hypothèse testée est la suivante :

$$H_0: p_1 = p_2 \text{ contre } H_1: p_1 \neq p_2$$

Une valeur z de la variable aléatoire Z est calculée :

$$z = \frac{\left| \frac{k_1}{n_1} - \frac{k_2}{n_2} \right|}{\sqrt{\hat{p}\hat{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \text{ avec } \hat{p} = \frac{k_1 + k_2}{n_1 + n_2}$$

z ou $|z|$ calculée ($|z|_{\text{obs}}$) est comparée avec la valeur $|z|_{\text{seuil}}$ lue **sur la table**

de la loi normale centrée réduite pour un risque d'erreur α fixé.

- si $|T_0| > T_\alpha$ l'hypothèse H_0 est rejetée au risque d'erreur α : les deux échantillons sont extraits de deux populations ayant des probabilités de succès respectivement p_1 et p_2 .
- si $|T_0| \leq T_\alpha$ l'hypothèse H_0 est acceptée: les deux échantillons sont extraits de deux populations ayant même probabilité de succès p .

Exemple :

On veut tester l'impact des travaux dirigés dans la réussite à l'examen de statistique.

	Groupe 1	Groupe 2
Nbre d'heures de TD	20 h	30 h
Nbre d'étudiants	180	150
Nbre d'étudiants ayant réussi à l'examen	126	129

Reponse

Données :

$$\text{Groupe 1 : 20h de TD avec } n_1 = 180 \quad f_1 = \frac{k_1}{n_1} = \frac{126}{180} = \mathbf{0,70}$$

$$\text{Groupe 2 : 30h de TD avec } n_2 = 150 \quad f_2 = \frac{k_2}{n_2} = \frac{129}{150} = \mathbf{0,86}$$

Conditions : échantillons indépendants, $n_1 p_1, n_2 p_2, n_1 q_1, n_2 q_2 \geq 10$

Hypothèse : $H_0: p_1 = p_2$ contre $H_1: p_1 < p_2$

Test unilatéral : la réussite est meilleure avec plus d'heures de TD.

$$\text{Statistique : } T_0 = \frac{\frac{k_1}{n_1} - \frac{k_2}{n_2}}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \mathbf{-3,45} \quad \text{avec} \quad \hat{p} = \frac{k_1 + k_2}{n_1 + n_2} = \mathbf{0,773}$$

Décision :

$T_0 = -3,45$ correspond à une probabilité critique $\alpha_{\text{obs.}} < \mathbf{0,001}$.

$\alpha_{\text{obs.}} < \mathbf{0,001}$ donc le risque d'erreur de rejeter H_0 alors qu'elle est vraie est très faible. On peut donc rejeter l'hypothèse H_0 avec un risque pratiquement nul de se tromper.

Comme attendu, le taux de réussite est significativement plus grand lorsque le nombre d'heure de TD est plus élevé (plus de pratique).

Chapitre 3: Analyse de la variance

1. Principe

L'analyse de la variance est une technique statistique fondamentale. Elle vise à comparer des moyennes sur plusieurs échantillons. Elle s'applique sur un tableau de contingence. C'est la généralisation de la comparaison des deux moyennes.

L'hypothèse à vérifier (H_0) est que tous les échantillons ont la même moyenne. L'hypothèse alternative H_1 est qu'au moins l'un d'eux joue les trouble-fête avec une moyenne sensiblement différente des autres. Il existe j et i tel que m_j différente de m_i

2. Procédure de résolution

Etant donné un tableau de contingence, on calcule

$n_{.j}$ = Nombre de modalité par groupe

$T_{.j}$ = effectif du groupe

$\overline{X}_{.j}$ = Moyenne du groupe

S_j^2 = Variance du groupe

$n_{..}$ = Nombre total de modalité de la population

$T_{..}$ = Effectif de la population

$\overline{X}_{..}$ = Moyenne de la population

S^2 = Variance de la population

$SST = S^2 * (N_{..} - 1)$

$SSW = \sum (n_{.j} - 1) \cdot S_j^2$

$SSA = SST - SSW$

On remplit le tableau d'analyse comme suit :

	SS	ddl	MS	F
A	ssa	r-1	$Ssa/(r-1)$	MSA/MSW
W	ssw	n-r	$Ssw/(n-r)$	
T	sst	n-1		

On lit dans la table de Fisher snedecor (en ligne la plus petite variance et en colonne la plus grande) la distance limite d à $(r-1)$ et $(n-r)$ ddl en fonction du risque α .

Si $d < D$ on accepte H_0 selon laquelle les moyennes sont sensiblement égales.

3. Exemple

Effectuez une analyse de variance sur ce tableau avec un risque de 5%

22	20	8	12	7
30	18	9	15	8
20	25	21	23	42
10	10	28	22	7
8	9	7	8	10

Résolution




	22	20	8	12	7		
	30	18	9	15	8		
	20	25	21	23	42		
	10	10	28	22	7		
	8	9	7	8	10		
$N_{.j}$	5	5	5	5	5	25,00	N..
$T_{.j}$	90	82	73	80	74	399,00	T..
$M_{.j}$	18	16,4	14,6	16	14,8	15,96	M..
$S^2_{.j}$	82	46,3	88,3	41,5	232,7	83,37	S²

Tableau d'analyse

TABLEAU D'ANALYSE DE VARIANCE

	<i>SS</i>	<i>ddl</i>	<i>MC</i>	<i>F</i>	<i>Probabilité</i>	<i>D</i>
A	37,76	4	9,44	0,0962	0,9825	2,8661
W	1963,2	20	98,16			
T	2000,96	24				

INVERSE.LOI.F

Probabilité  = 0,05
Degrés_liberté1  = 4
Degrés_liberté2  = 20
 = 2,866081402

Code Python

```
import pandas as pd
import statsmodels.formula.api
import statsmodels.api
from scipy.stats.distributions import f
```

```
av= pd.read_excel("données/application.xlsx", sheet_name="AV1")
av.head()
```

```
av=av.melt()
av
```

```
fit = statsmodels.formula.api.ols('value ~ variable', data = av).fit()
avr = statsmodels.api.stats.anova_lm(fit)
avr
```

```
fo=avr['F'][0]
ddla=avr['df'][0]
ddlw=avr['df'][1]
pvalue=avr['PR(>F)'][0]
print("dda=",ddla,"\nddw=",ddlw,"\nFo=",round(fo,3),'\nPvalue=',round(pvalue,3))
```

```
alpha=float(input("Donnez le risque Alpha:"))
D=f.ppf(1-alpha, ddla,ddlw)
print("\nD=",round(D,3))
if fo<= D:
    print("H0 est acceptée c'est à dire les moyennes sont égales")
else:
    print("H0 est rejetée c'est à dire les moyennes ne sont pas égales")
```


Chapitre 4 Test d'indépendance de khi deux

1. Position du problème

Le test d'indépendance de khi deux a pour but de savoir si deux ou plusieurs variables sont indépendantes. Il utilise un tableau de contingence qui est un tableau à deux entrées dont toutes les entrées sont des variables. L'intersection n_{ij} désigne le nombre d'individu présentant le caractère i de la variable V_1 et le caractère j de la variable V_2 . L'hypothèse de base H_0 suppose que les variables sont indépendantes contre H_1 qui suppose que les variables sont liées.

2. Résolution du problème

Étant donné un tableau de contingence à r lignes et s colonnes, on calcule :

a. Les effectifs marginaux

$N_{i.}$ = Total ligne
 $N_{.j}$ = Total colonne
 $N_{..}$ = Total général

b. Le tableau théorique

$$\alpha_{ij} = \frac{n_{i.} \cdot n_{.j}}{n_{..}}$$

c. La distance de Khi deux

$$d = \sum \sum \frac{(n_{ij} - \alpha_{ij})^2}{\alpha_{ij}} = \sum \sum \frac{n_{ij}^2}{\alpha_{ij}} - n_{..}$$

d. Lecture de la distance limite D

Le degré de liberté ; $ddl = (r-1) \cdot (s-1)$

Le risque α étant donnée on lit dans la table de Khi deux la distance limite D

e. Conclusion

Si $d \leq D$ on accepte l'hypothèse H_0 c'est à dire les variables sont indépendantes. Dans le cas contraire c'est H_1 qui est acceptée c'est-à-dire les variables sont liées.

3. Application

Étant donné le tableau observé suivant effectuez le test d'indépendance de Khi deux pour un risque de 5%

Tableau observé

	P1	P2	P3	P4	P5	P6
A1	9	35	44	24	8	13
A2	66	72	171	122	48	71
A3	77	139	380	195	69	233
A4	50	78	155	152	57	85
A6	52	86	274	43	26	48
A6	55	103	191	40	25	46

Les effectifs marginaux ;

	P1	P2	P3	P4	P5	P6	Ni.
A1	9	35	44	24	8	13	133
A2	66	72	171	122	48	71	550
A3	77	139	380	195	69	233	1 093
A4	50	78	155	152	57	85	577
A6	52	86	274	43	26	48	529
A6	55	103	191	40	25	46	460
N.j	309	513	1 215	576	233	496	3 342

Tableau théorique

Exemple 12.297= $\frac{309 \cdot 133}{3342}$

$$\alpha_{ij} = \frac{n_{i.} \cdot n_{.j}}{n_{..}}$$

12,297	20,416	48,353	22,923	9,273	19,739	133
50,853	84,425	199,955	94,794	38,345	81,628	550
101,058	167,776	397,365	188,381	76,203	162,217	93
53,349	88,57	209,771	99,447	40,228	85,635	577
48,911	81,202	192,32	91,174	36,881	78,511	529
42,531	70,61	167,235	79,282	32,071	68,27	460
309	513	1 215	576	233	496	3 342

La distance de khi deux

$$d = \sum \sum \frac{n_{ij}^2}{\alpha_{ij}} - n_{..} \quad d = 3598.007 - 3342 = 256.007$$

- ddl=(6-1)(6-1)=25 et
- La distance limite D=37.652

$d \gg D$ alors H_0 rejetée c'est-à-dire H_1 acceptée les variables sont liées

Arguments de la fonction

KHIDEUX.INVERSE

Probabilité 0,05 = 0,05

Degrés_liberté 25 = 25

= 37,65248413

Renvoie, pour une probabilité unilatérale donnée, la valeur d'une variable aléatoire suivant une loi du Khi-deux.

Degrés_liberté représente le nombre de degrés de liberté, un nombre entre 1 et 10^{10} , 10^{10} exclus.

Résultat = 37,65248413

[Aide sur cette fonction](#) OK Annuler

Exercice

Un tableau de contingence indiquant des quantités de CD vendus sur quatre points de vente en fonction de leur style musical :

Point de Vente	Classique	Variété	Rock	Electro	Jazz& Blues
Libreville	21	340	46	210	9
Port gentil	15	150	20	110	5
FranceVille	17	180	19	99	6
Oyem	22	175	22	187	6

On souhaite savoir si, compte tenu de leur emplacement, ces points de vente attirent ou non des clientèles différentes pour un risque de 2%.

Table de Khi deux

Khi deux	0,1	0,09	0,08	0,07	0,06	0,05	0,04	0,03	0,02	0,01
1	2,7055	2,8744	3,0649	3,2830	3,5374	3,8415	4,2179	4,7093	5,4119	6,6349
2	4,6052	4,8159	5,0515	5,3185	5,6268	5,9915	6,4378	7,0131	7,8240	9,2103
3	6,2514	6,4915	6,7587	7,0603	7,4069	7,8147	8,3112	8,9473	9,8374	11,3449
4	7,7794	8,0434	8,3365	8,6664	9,0444	9,4877	10,0255	10,7119	11,6678	13,2767
5	9,2364	9,5211	9,8366	10,1910	10,5962	11,0705	11,6443	12,3746	13,3882	15,0863
6	10,6446	10,9479	11,2835	11,6599	12,0896	12,5916	13,1978	13,9676	15,0332	16,8119
7	12,0170	12,3372	12,6912	13,0877	13,5397	14,0671	14,7030	15,5091	16,6224	18,4753
8	13,3616	13,6975	14,0684	14,4836	14,9563	15,5073	16,1708	17,0105	18,1682	20,0902
9	14,6837	15,0342	15,4211	15,8537	16,3459	16,9190	17,6083	18,4796	19,6790	21,6660
10	15,9872	16,3516	16,7535	17,2026	17,7131	18,3070	19,0207	19,9219	21,1608	23,2093
11	17,2750	17,6526	18,0687	18,5334	19,0614	19,6751	20,4120	21,3416	22,6179	24,7250
12	18,5493	18,9395	19,3692	19,8488	20,3934	21,0261	21,7851	22,7418	24,0540	26,2170
13	19,8119	20,2140	20,6568	21,1507	21,7113	22,3620	23,1423	24,1249	25,4715	27,6882
14	21,0641	21,4778	21,9331	22,4408	23,0166	23,6848	24,4855	25,4931	26,8728	29,1412
15	22,3071	22,7319	23,1993	23,7202	24,3108	24,9958	25,8162	26,8479	28,2595	30,5779
16	23,5418	23,9774	24,4564	24,9901	25,5950	26,2962	27,1356	28,1907	29,6332	31,9999
17	24,7690	25,2150	25,7053	26,2514	26,8701	27,5871	28,4450	29,5227	30,9950	33,4087
18	25,9894	26,4455	26,9467	27,5049	28,1370	28,8693	29,7451	30,8447	32,3462	34,8053
19	27,2036	27,6694	28,1814	28,7512	29,3964	30,1435	31,0367	32,1577	33,6874	36,1909
20	28,4120	28,8874	29,4097	29,9910	30,6489	31,4104	32,3206	33,4624	35,0196	37,5662
21	29,6151	30,0998	30,6322	31,2246	31,8949	32,6706	33,5972	34,7593	36,3434	38,9322
22	30,8133	31,3071	31,8494	32,4526	33,1350	33,9244	34,8673	36,0492	37,6595	40,2894
23	32,0069	32,5096	33,0616	33,6754	34,3696	35,1725	36,1311	37,3323	38,9683	41,6384
24	33,1962	33,7077	34,2690	34,8932	35,5990	36,4150	37,3891	38,6093	40,2704	42,9798
25	34,3816	34,9015	35,4721	36,1065	36,8235	37,6525	38,6416	39,8804	41,5661	44,3141
26	35,5632	36,0915	36,6711	37,3154	38,0435	38,8851	39,8891	41,1460	42,8558	45,6417
27	36,7412	37,2777	37,8662	38,5202	39,2593	40,1133	41,1318	42,4066	44,1400	46,9629
28	37,9159	38,4604	39,0577	39,7213	40,4710	41,3371	42,3699	43,6622	45,4188	48,2782
29	39,0875	39,6398	40,2456	40,9187	41,6789	42,5570	43,6038	44,9132	46,6927	49,5879
29	39,0875	39,6398	40,2456	40,9187	41,6789	42,5570	43,6038	44,9132	46,6927	49,5879
30	40,2560	40,8161	41,4304	42,1126	42,8831	43,7730	44,8336	46,1599	47,9618	50,8922

```
# Code en Python
import numpy as np
from scipy.stats import chi2_contingency

# Données d'exemple pour le test
data = np.array([[10, 20, 30], [15, 25, 35]])

# Effectuer le test d'indépendance du khi-deux
chi2, p, dof, expected = chi2_contingency(data)

# Afficher les résultats
print("Résultats du test d'indépendance du khi-deux :")
print("Statistique du chi-carré :", chi2)
print("Valeur p :", p)
print("Degrés de liberté :", dof)
print("Fréquences attendues :", expected)
```

```
#Code en langage R
# Données d'exemple pour le test
observed_data <- matrix(c(10, 20, 30, 15, 25, 35), nrow = 2, byrow = TRUE)
# Effectuer le test d'indépendance du khi-deux
result <- chisq.test(observed_data)
# Afficher les résultats
print("Résultats du test d'indépendance du khi-deux :")
print("Statistique du chi-carré :", result$statistic)
print("Valeur p :", result$p.value)
print("Degrés de liberté :", result$parameter)
print("Fréquences attendues :", result$expected)
```

Chapitre 5 : Regressions

Introduction

L'ajustement linéaire simple, également connu sous le nom de régression linéaire simple, est une méthode statistique utilisée pour modéliser la relation entre une variable indépendante (X) et une variable dépendante (Y) à l'aide d'une équation linéaire. L'objectif est de trouver la meilleure ligne droite qui représente au mieux la relation entre les deux variables.

Résolution

L'équation générale d'un ajustement linéaire simple est donnée par :

$$Y = aX + b$$

où Y représente la variable dépendante, X représente la variable indépendante, **b** est l'ordonnée à l'origine (intercept) et **a** est la pente de la ligne (coefficient directeur).

L'objet est de minimiser la somme des écarts des y_i . $S = \sum (y_i - a x_i - b)^2$

En cherchant les coefficients **a** (la pente) et **b** (l'ordonnée à l'origine)

a. Calcul de b l'ordonnée à l'origine

$$\begin{aligned} \frac{\partial S}{\partial b} &= -2 \sum_1^n (y_i - a x_i - b) = 0 \\ \Rightarrow \sum_1^n (y_i - a x_i - b) &= 0 \\ \Rightarrow \sum_1^n y_i - a \sum_1^n x_i - \sum_1^n b &= 0 \\ \Rightarrow \bar{y} - a \bar{x} - b &= 0 \Leftrightarrow b = \bar{y} - a \bar{x} \end{aligned}$$

b. Calcul de la pente a

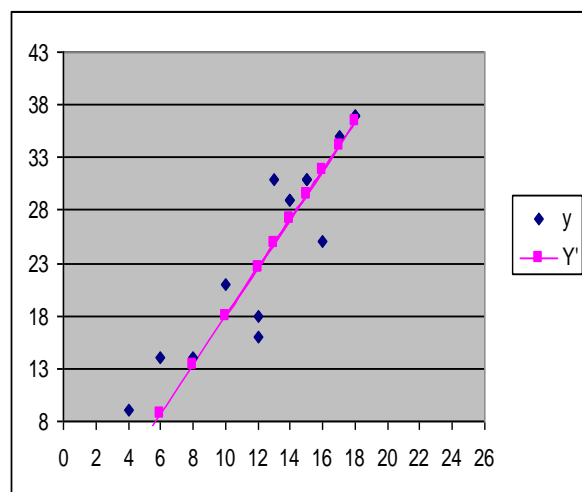
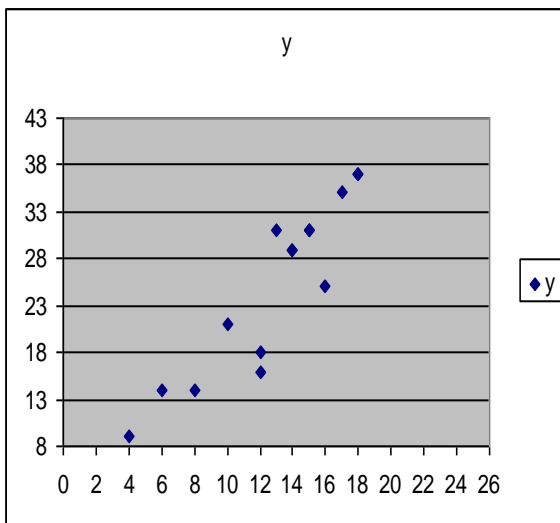
$$\begin{aligned} S &= \sum (y_i - a x_i - \bar{y} + a \bar{x})^2 \\ S &= \sum (y_i - \bar{y} - a(x_i - \bar{x}))^2 \\ \frac{\partial S}{\partial a} &= -2 \sum (x_i - \bar{x})(y_i - \bar{y} - a(x_i - \bar{x})) \\ \Rightarrow \sum (x_i - \bar{x})(y_i - \bar{y}) - a \sum_1^n (x_i - \bar{x})(x_i - \bar{x}) &= 0 \end{aligned}$$

$$\Rightarrow a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \Leftrightarrow a = \frac{\text{cov}(x, y)}{\text{var}(x)}$$

c. Le coefficient de corrélation r

$$r = \sqrt{a * a'} \text{ avec } a' = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2} \text{ d'où } r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (y_i - \bar{y})^2 * \sum (x_i - \bar{x})^2}}$$

; r est compris entre -1 et 1



- Cas du coefficient de corrélation (r) : $r = \frac{\text{cov}(x, y)}{\sigma_x * \sigma_y}$

Le coefficient de corrélation mesure la force et la direction de la relation linéaire entre les variables X et Y. Sa valeur se situe entre -1 et 1. Voici comment interpréter le coefficient de corrélation :

Si r est proche de 1, cela indique une corrélation positive forte, ce qui signifie que lorsque les valeurs de X augmentent, les valeurs de Y ont tendance à augmenter également.

Si r est proche de -1, cela indique une corrélation négative forte, ce qui signifie que lorsque les valeurs de X augmentent, les valeurs de Y ont tendance à diminuer.

Si r est proche de 0, cela indique une corrélation faible, ce qui signifie qu'il y a peu ou pas de relation linéaire entre les variables X et Y.

- Cas du coefficient de détermination (R^2) :
$$\frac{\sum_1^n (Y'_i - \bar{Y})^2}{\sum_1^n (Y_i - \bar{Y})^2}$$

Le coefficient de détermination mesure la proportion de la variance totale de la variable dépendante (Y) qui peut être expliquée par la variable indépendante (X). Il est également compris entre 0 et 1. Voici comment interpréter le coefficient de détermination :

Plus R^2 est proche de 1, plus le modèle d'ajustement linéaire explique une grande partie de la variance de la variable dépendante. Cela indique une bonne adéquation du modèle aux données.

Si R^2 est proche de 0, cela signifie que le modèle ne parvient pas à expliquer la variance de la variable dépendante. Il est possible que d'autres facteurs non inclus dans le modèle aient une influence sur la variable dépendante.

- Cas de la Racine carrée de l'erreur quadratique moyenne (RMSE) :

$$RMSE = \sqrt{\frac{\sum_1^n (Y'_i - Y)^2}{n}}$$

La RMSE mesure l'écart moyen entre les valeurs prédites par le modèle d'ajustement linéaire et les valeurs réelles de la variable dépendante. Elle est utile pour évaluer la précision de prédiction du modèle. Voici comment interpréter la RMSE :

Une RMSE plus proche de 0 indique une meilleure adéquation du modèle et une meilleure précision de prédiction.

Une RMSE plus élevée indique une plus grande dispersion des valeurs prédites par rapport aux valeurs réelles, ce qui suggère que le modèle ne s'ajuste pas bien aux données.

e. Application

X	Y	X ²	Y ²	XY	Y'=ax+b	Y'-Y	(Y'-Y) ²	(y'-ybar) ²	(y-ybar) ²
2	7								
5	13								
6	15								
1	5								
4	11								
8	19								
6	15								
5	13								

Total

	X	Y	a	
Moyenne			b	RMSE
Variances			r	
Covariance			MSE	R ² =

Code Python

```
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Données d'exemple
X = np.array([1, 2, 3, 4, 5]) # Variable indépendante
Y = np.array([2, 4, 5, 4, 6]) # Variable dépendante

# Créer le modèle d'ajustement linéaire
model = LinearRegression()

# Entraîner le modèle
model.fit(X.reshape(-1, 1), Y)

# Effectuer les prédictions
Y_pred = model.predict(X.reshape(-1, 1))

# Calculer la RMSE
rmse = np.sqrt(mean_squared_error(Y, Y_pred))

# Calculer le coefficient de détermination ( $R^2$ )
r2 = r2_score(Y, Y_pred)

# Obtenir les coefficients de l'ajustement linéaire
intercept = model.intercept_
slope = model.coef_[0]
# le coefficient de corrélation r
r = np.corrcoef(X, Y)

# Afficher les résultats
print("Intercept :", intercept)
print("Slope :", slope)
print("RMSE :", rmse)
print("R2 :", r2)
print("r= :", r)
```

Chapitre 6 : Analyse discriminante

1. Le partitionnement ou Kmeans

Définition

Le partitionnement est une méthode d'analyse des données non supervisée qui vise à diviser un ensemble d'individus ou d'objets en plusieurs groupes distincts (clusters) en fonction de leurs similarités. Contrairement à la classification ascendante hiérarchique (CAH), le partitionnement ne vise pas à créer une hiérarchie de groupes, mais plutôt à diviser les individus en groupes non hiérarchiques.

Le processus de partitionnement commence par sélectionner un nombre donné de clusters et en attribuant les individus ou les objets au hasard à l'un des clusters. Ensuite, les centres de chaque cluster sont calculés et les individus ou les objets sont réattribués aux clusters en fonction de leur distance au centre de chaque cluster. Cette étape est répétée plusieurs fois jusqu'à ce que les centres des clusters ne se déplacent plus ou que la qualité de la partition soit considérée comme suffisante.

Il existe plusieurs algorithmes de partitionnement, tels que l'algorithme de k-means, l'algorithme de k-medoids ou encore l'algorithme de clustering spectral. En fonction de l'algorithme choisi, la mesure de la similarité entre les individus ou les objets peut varier, ainsi que la manière dont les centres des clusters sont calculés et les individus ou les objets sont réattribués aux clusters.

Le partitionnement est largement utilisé dans différents domaines, tels que la biologie, la géologie, la sociologie, la finance ou le marketing. Il peut être utilisé pour regrouper les clients en fonction de leurs caractéristiques démographiques ou comportementales, pour détecter des anomalies dans les données financières, ou encore pour classer les images médicales en fonction de leurs caractéristiques.

Soient 8 points suivants, effectuez un Kmeans avec $U1=A1$; $U2=A4$ et $U3= A7$

Point	X	Y
A1	2	10
A2	2	5
A3	8	4
A4	5	8
A5	7	5
A6	6	4
A7	1	2
A8	4	9

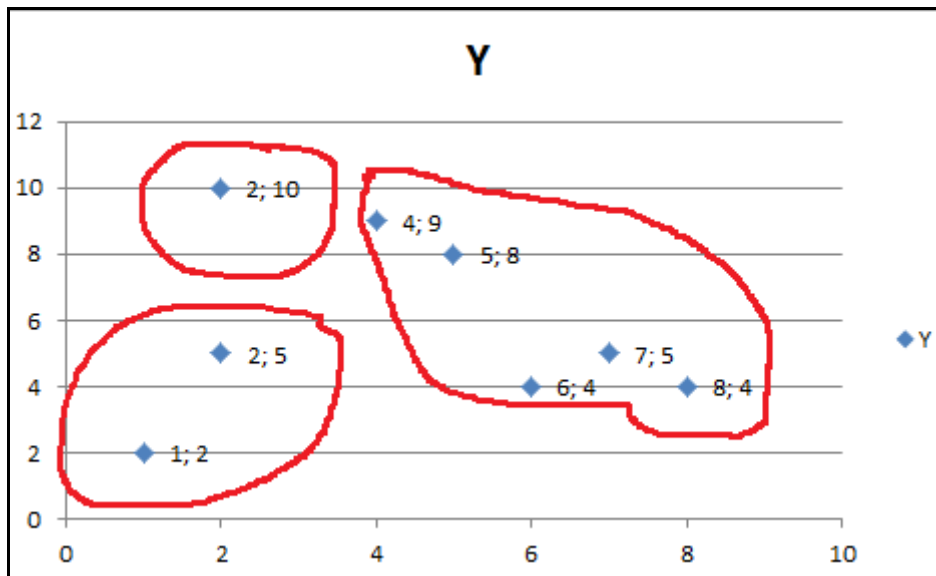
Etape 1

		2 10 A1	2 5 A2	8 4 A3	5 8 A4	7 5 A5	6 4 A6	1 2 A7	4 9 A8
U1=A1	2	-	5,00	8,49	3,61	7,07	7,21	8,06	2,24
	10								
U2=A4	5	3,61	4,24	5,00	-	3,61	4,12	7,21	1,41
	8								
U3=A7	1	8,06	3,16	7,28	7,21	6,71	5,39	-	7,62
	2								

C1={A1}

C2={A3,A4,A5,A6,A8}

C3={A2,A7}



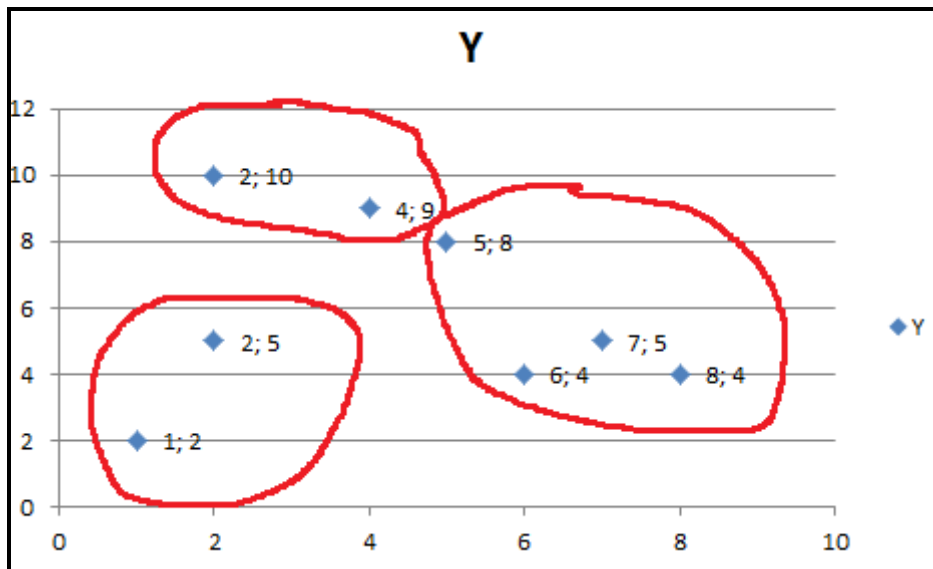
Etape2

		2 10 A1	2 5 A2	8 4 A3	5 8 A4	7 5 A5	6 4 A6	1 2 A7	4 9 A8
U1	2	-	5,00	8,49	3,61	7,07	7,21	8,06	2,24
	10								
U2	6	5,66	4,12	2,83	2,24	1,41	2,00	6,40	3,61
	6								
U3	1,5	6,52	1,58	6,52	5,70	5,70	4,53	1,58	6,04
	3,5								

C1={A1,A8}

C2={A3,A4,A5,A6}

C3={A2,A7}



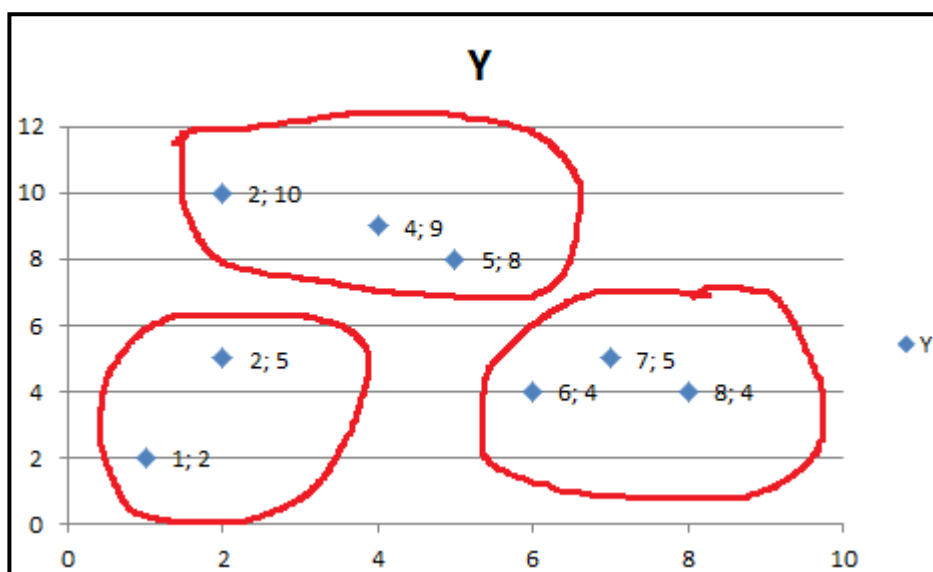
Étape 3

		2 10 A1	2 5 A2	8 4 A3	5 8 A4	7 5 A5	6 4 A6	1 2 A7	4 9 A8
U1	3	1,12	4,61	7,43	2,50	6,02	6,26	7,76	1,12
	9,5								
U2	6,5	6,54	4,51	1,95	3,13	0,56	1,35	6,39	4,51
	5,25								
U3	1,5	6,52	1,58	6,52	5,70	5,70	4,53	1,58	6,04
	3,5								

C1={A1,A4,A8}

C2={A3,A5,A6}

C3={A2,A7}



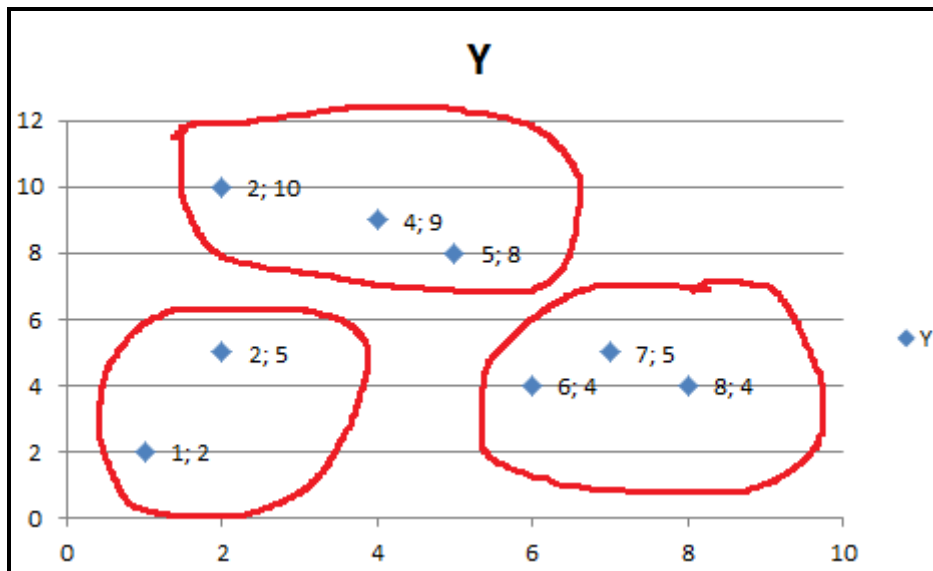
Étape 4

		2 10 A1	2 5 A2	8 4 A3	5 8 A4	7 5 A5	6 4 A6	1 2 A7	4 9 A8
U1	3,67	1,94	4,33	6,62	1,67	5,21	5,52	7,49	0,33
	9,00								
U2	6,50	6,54	4,51	1,95	3,13	0,56	1,35	6,39	4,51
	4,33								
U3	1,50	6,52	1,58	6,52	5,70	5,70	4,53	1,58	6,04
	3,50								

$C1=\{A1,A4,A8\}$

$C2=\{A3,A5,A6\}$

$C3=\{A2,A7\}$



Aucun individu n'a changé de classe dont l'algorithme est stable

2. Classification Ascendante Hiérarchique(CAH)

a. Définition

La classification ascendante hiérarchique (CAH) est une méthode d'analyse des données non supervisée qui vise à regrouper un ensemble d'individus ou d'objets en plusieurs clusters (groupes) en fonction de leurs similarités.

La méthode de CAH commence par considérer chaque individu ou objet comme un cluster distinct, puis elle combine progressivement les clusters en groupes plus grands en fonction de leurs similarités. Le processus de combinaison se poursuit jusqu'à ce que tous les individus soient regroupés dans un même cluster.

Pour mesurer les similarités entre les individus ou les objets, différentes métriques peuvent être utilisées, telles que la distance euclidienne, la distance de Manhattan ou la corrélation. En fonction de la métrique choisie, la CAH peut être utilisée pour regrouper les individus ou les objets en fonction de leurs caractéristiques ou de leurs comportements similaires.

La CAH peut être utilisée dans différents domaines, tels que la biologie, la géologie, la sociologie ou le marketing. Par exemple, en marketing, elle peut être utilisée pour regrouper les clients en fonction de leurs comportements d'achat, afin de créer des segments de marché et d'adapter les offres en fonction de chaque groupe.

b. Exemple

Soient 8 points suivants, effectuez la Classification Ascendante Hiérarchique

Point	X	Y
A1	2	10
A2	2	5
A3	8	4
A4	5	8
A5	7	5
A6	6	4
A7	1	2
A8	4	9

Etape 1

		2 10 A1	2 5 A2	8 4 A3	5 8 A4	7 5 A5	6 4 A6	1 2 A7	4 9 A8
A1	2 10	-	5,00	8,49	3,61	7,07	7,21	8,06	2,24
A2	2 5	5,00	-	6,08	4,24	5,00	4,12	3,16	4,47
A3	8 4	8,49	6,08	-	5,00	1,41	2,00	7,28	6,40
A4	5 8	3,61	4,24	5,00	-	3,61	4,12	7,21	1,41
A5	7 5	7,07	5,00	1,41	3,61	-	1,41	6,71	5,00
A6	6 4	7,21	4,12	2,00	4,12	1,41	-	5,39	5,39
A7	1 2	8,06	3,16	7,28	7,21	6,71	5,39	-	7,62
A8	4 9	2,24	4,47	6,40	1,41	5,00	5,39	7,62	-

Min=1,41
A3-A5

G1={7,5;4,5}

Etape2 : On va remplacer le couple {A3 ; A5 } par son centre de gravité G1(7.5 ;4.5) et recalculer les distances.

		2 10 A1	2 5 A2	7,5 4,5 C1	5 8 A4	6 4 A6	1 2 A7	4 9 A8
A1	2 10	-	5,00	7,78	3,61	7,21	8,06	2,24
A2	2 5	5,00	-	5,52	4,24	4,12	3,16	4,47
C1	7,5 4,5	7,78	5,52	-	4,30	1,58	6,96	5,70
A4	5 8	3,61	4,24	4,30	-	4,12	7,21	1,41
A6	6 4	7,21	4,12	1,58	4,12	-	5,39	5,39
A7	1 2	8,06	3,16	6,96	7,21	5,39	-	7,62
A8	4 9	2,24	4,47	5,70	1,41	5,39	7,62	-

Min=1,41
A4-A8

G2={4,5;8,5}

Étape 3 : On va remplacer le couple {A4 ; A8 } par son centre de gravité G2(4,5 ;8.5) et recalculer les distances

		2 10 A1	2 5 A2	7,5 4,5 G1	4,5 8,5 G2	6 4 A6	1 2 A7
A1	2 10	-	5,00	7,78	2,92	7,21	8,06
A2	2 5	5,00	-	5,52	4,30	4,12	3,16
G1	7,5 4,5	7,78	5,52	-	5,00	1,58	6,96
G2	4,5 8,5	2,92	4,30	5,00	-	4,74	7,38
A6	6 4	7,21	4,12	1,58	4,74	-	5,39
A7	1 2	8,06	3,16	6,96	7,38	5,39	-

Min=1,58
G1 et A6

G3(6,75;4,25)

Étape 4 : On va remplacer le couple {G1 ; A6} par son centre de gravité G3 (6,75 ;4,25 } et recalculer les distances.

		2 10 A1	2 5 A2	7,5 4,5 G1	4,5 8,5 G2	6 4 A6	1 2 A7
A1	2 10	-	5,00	7,78	2,92	7,21	8,06
A2	2 5	5,00	-	5,52	4,30	4,12	3,16
G1	7,5 4,5	7,78	5,52	-	5,00	1,58	6,96
G2	4,5 8,5	2,92	4,30	5,00	-	4,74	7,38
A6	6 4	7,21	4,12	1,58	4,74	-	5,39
A7	1 2	8,06	3,16	6,96	7,38	5,39	-

Min=1,58
G1 et A6

G3(6,75;4,25)

Étape 5 : On va remplacer le couple {G1 ; A6} par son centre de gravité G4(1,5 ;3,5} et recalculer les distances.

		2 10 A1	2 5 A2	6,75 4,25 G3	5 8 G2	1 2 A7
A1	2 10	-	5,00	7,46	3,61	8,06
A2	2 5	5,00	-	4,81	4,24	3,16
G3	6,75 4,25	7,46	4,81	-	4,14	6,17
G2	5 8	3,61	4,24	4,14	-	7,21
A7	1 2	8,06	3,16	6,17	7,21	-

Min=3,16

A2 et A7

G4(1,5;3,5)

Étape 6 : On va remplacer le couple {A2 ; A6} par son centre de gravité G4(1,5 ;3,5} et recalculer les distances.

		2 10 A1	1,5 3,5 G4	6,75 4,25 G3	5 8 G2
A1	2 10	-	6,52	7,46	3,61
G4	1,5 3,5	6,52	-	5,30	5,70
G3	6,75 4,25	7,46	5,30	-	4,14
G2	5 8	3,61	5,70	4,14	-

Min=3,61
A1 et G2
G5(3,5;9)

Étape 7 : On va remplacer le couple {A1 ; G2} par son centre de gravité G6(3,5 ;9} et recalculer les distances.

		3,5 9 G5	1,5 3,5 G4	6,75 4,25 G3
G5	3,5 9	-	5,85	5,76
G4	1,5 3,5	5,85	-	5,30
G3	6,75 4,25	5,76	5,30	-

Min=5,30
G4 et G3
G6(4,125;3,875)

Étape 8 : On va remplacer le couple {G4 ; G3} par son centre de gravité G4(4,125 ;3.875} et recalculer les distances.

		3,5 9 G5	4,125 3,875 G6
G5	3,5 9	-	5,16
G6	4,125 3,875	5,16	-

Min=5,16
G5 et G6
Fin
d'algorithme

Code Python

```
from matplotlib import pyplot as plt
import pandas as pd
from scipy.cluster.hierarchy import dendrogram, linkage
import scipy
%matplotlib inline

df = pd.read_excel("./données/CAH.xlsx",index_col=0)
df

Z = linkage(df,method='ward',metric='euclidean')

plt.scatter(df.X,df.Y)

#affichage du dendrogramme
fig = plt.figure(figsize=(30, 30))
plt.title("CAH")
dendrogram(Z,labels=df.index,orientation='top',color_threshold=4)
plt.show()
```