shipout/backgroundshipout/foreground

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Centre Universitaire Salhi Ahmed- Naama
Institut des sciences et technologies
Département de Mathématiques et Informatique

Mémoire de fin d'études

En vue de l'obtention du diplôme de Master
En : Mathématiques

Spécialité: Probabilités, Statistique et Applications

Intitulé

Tests d'indépendance et application avec R

Présenté par :
Yakoubi Meryem

Soutenu : Juillet 2022

Devant le jury composé de :

Dr.DAOUDI Hamza	MCA	Univ Iben Kheldoun-Tiaret	Présédent
Dr.LAALA Zeyneb	MCB	C-Univ Naâma	Examinatrice
Dr.BELGUERNA Abderrahmane	MCA	C-Univ Naâma	Encadreur

Année universitaire 2021/2022

Remerciement

Je remercier ALLAH pour ce vers quoi il m'a guidé. J'envoie mes salutations d'appréciation à mes parents et amis pour leur soutien, et je remercie mon encadreur Dr. A. Belguerna pour ces efforts et ces conseils. Puisse Allah, vous préserver et vous accorder santé, longue vie et bonheur.

Contents

List of Figures

List of Tables

Résumé

Dans le test d'indépendance, l'affirmation est que les variables de ligne et de colonne sont indépendantes l'une de l'autre. C'est l'hypothèse nulle.
La règle de multiplication disait que si deux événements étaient indépendants, alors la probabilité qu'ils se produisent tous les deux était le produit des probabilités que chacun se produise. C'est la clé pour travailler le test d'indépendance. Si vous finissez par rejeter l'hypothèse nulle, alors l'hypothèse doit avoir été erronée et la variable de ligne et de colonne est dépendante. N'oubliez pas que tous les tests d'hypothèse sont effectués en supposant que l'hypothèse nulle est vraie.
[1]

Introduction

Lorsqu'on parle sur l'inférence statistique alors on a des techniques permettant de déduire les caractéristiques d'un objet (la population) à partir d'une partie de la population (échantillon), alors on peut dire que l'inférence statistique est un ensemble des méthodes qui permettant de faire des conclusions fiable à l'aide de données d'échantillon statistique, ces méthodes statistique données par Pierre Simon de Laplace et Carl Friedreich Gauss. à la fin du 19^e siècle a été reconnu la première phase de développement des méthodes statistique par E. Pearson, J. Neyman, K. Pearson, R. Ficher qui a donné les concepts fondamentales de vraisemblance, des tests d'hypothèses et d'intervalle de confiance, à partir de la fin des années 1940 jusqu'à aujourd'hui l'outil informatique nous a facilité les calculs.
Grâce à ces calculateurs on peut dépasser les hypothèse habituelles d'indépendance et de normalité.
Dans ce travail on va voir les différents tests d'indépendance avec différents types des variables qualitative où quantitative, ce mémoire se décompose en trois chapitres, dans le premier chapitre on donne les notions de base sur la statistique inférentielle, le deuxième chapitre porte sur l'hypothèse d'indépendance dans les deux cas paramétrique et non paramétrique.
Enfin, le dernier chapitre est réservé a l'applique de quelque tests d'indépendance avec le language R dans plusieurs domaines.

Chapter 1

Généralité et notions de bases

1.1 Quelque définitions

Définition 1
soit le couple (Ω, C) où Ω est l'ensemble des événements et C est une classe de parties de Ω.
On appelle probabilité ou loi de probabilité sur cette espace, l'application $P : C \to [0, 1]$ qui est vérifier:

- $P(\Omega) = 1$
- *Pour tout E événement,* $0 \leqslant P(E) \leqslant 1$
- *Pour $B_1, ..., B_n$ ensemble dénombrable d'événement incompatible tel que:*

$$P(\cup B_i) = \sum P(B_i)$$

1.1.1 Variable aléatoire X

Définition 2
Une variable aléatoire est une fonction définie sur l'ensemble des résultats possibles,d'une expérience aléatoire, on note un événement $(X = x_i)$ où xi sont des valeurs prises par la variable aléatoire X de plus, la probabilité d'obtenir ces valeurs est $P(X = x_i)$. Les variables aléatoires sont utilisées essentiellement pour modéliser les résultats d'une expérience aléatoire non-déterministe ou un phénomène quelconque.

Remarque 1
Soit X $(X : \Omega \to R)$ une variable aléatoire.
Si $X(\Omega)$ est dénombrable, alors X est une variable aléatoire discrète, sinon la variable X est une variable aléatoire continue.

1.1.2 La loi de probabilité

Une variable aléatoire est généralement définie par sa loi de probabilité, cette loi est caractérisée par un domaine de définition (ou encore support) c'est-à-dire l'ensemble

des valeurs qu'elle peut prendre, et les probabilités attribuées pour chaque valeur prise $P(X = x)$

1.1.3 Densité de probabilité

Lorsque on a une variable aléatoire, la fonction de densité est la probabilité ponctuelle $P(X = x) = f(x)$; où $F(x) = P(X < x)$ est la fonction de répartition telle que:

$$F(x) = \int_{-\infty}^{x} f(u)du$$

alors on peut dire que la densité de probabilité d'une variable continue est la dérivée première de la fonction de répartition par rapport à x.

1.1.4 Les moments

cas / moment	discrète	continu
d'ordre1 $E(X)$	$\sum_1 nX_iP(X = x_i$	$\int xf(x)dx$
d'ordre2 $E(X^2)$	$\sum_1^n X^2P(X = x_i)$	$\int x^2f(x)dx$
$var(X)$	$E(X^2) - (E(X))^2$	

1.2 Les lois usuelles discrètes

Distribution	support	loi de probabilité	$E(X)$	$var(X)$
Bernoulli$B(p)$	$[0, 1]$	$P(X = 0) = q; P(X = 1) = p$p+q=1	p	pq
Binomiale$B(n, p)$	$k \in N$	$P(X = K) = C_n^k p^k q^{n-k} q = 1 - p$	np	npq
Poisson$P(\lambda)$	$k \in N$	$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$	λ	λ
Géométrique $G(p)$	$k \in N^*$	$P(X = k) = pq^{k-1} q = 1 - p$	$\frac{1}{p}$	$\frac{q}{p^2}$

Table 1.1: tableau de quelques lois discrètes

1.3 Les lois usuelles continues

1.3.1 Loi Normale ou de Gauss $N(\mu, \sigma)$

Si X une variable aléatoire suit la loi Normale qui dépend deux paramètres l'espérance μ et l'écart type σ (tq $\sigma > 0$ car est une racine carrée de la variance σ^2).
On définie φ densité de probabilité $\forall x \in \mathbb{R}$ par:

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

Une telle variable aléatoire réelle X est dite variable gaussienne.
Lorsque la moyenne vaut 0 et l'écart type vaut 1, $N(0,1)$ appelée loi Normale centrée réduite ou loi Normale standard; sa fonction caractéristique est $e^{-t^2/2}$.
Seule la loi Normale centrée réduite est tabulé par ce que les différentes lois se déduisent à partir du théorème suivant: Si$Y \sim N(\mu,\sigma)$, alors $Z = \frac{Y-\mu}{\sigma} \sim N(0,1)$ et Φ fonction de répartition donnée par:
$P = (Z < x)$ tel que $\Phi(-x) = 1 - \Phi(x)$.

Exemple : 1
$\Phi(0) = 0.5; \Phi(1.96) \simeq 0.97$.

Notation: $Z_{\alpha/2}$ nombre pour lequel:

$$P(Z > z_{\alpha/2}) = \alpha/2$$

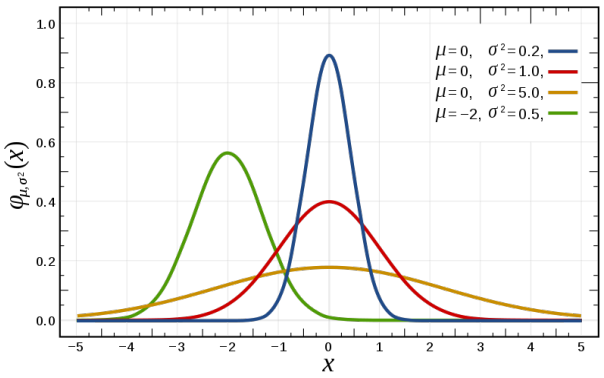


Figure 1.1: La courbe de la fonction de probabilité de loi Normale

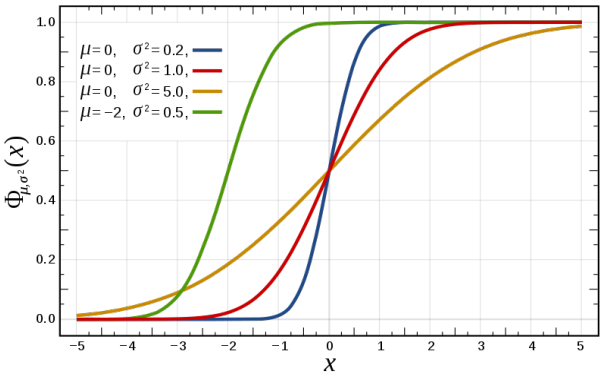


Figure 1.2: La courbe de la fonction de répartition de loi Normale

Propriété 1
X et Y variables aléatoire tels que $X \sim N(\mu_1,\sigma_1)$ *et*$Y \sim N(\mu_2,\sigma_2)$ *donc* $X + Y \sim N(\mu_1+\mu_2, \sqrt{\sigma_1^2+\sigma_2^2})$.

risque	valeur critique $z_{\alpha/2}$	coefficient de sécurité
0.01	2.58	99
0.02	2.33	98
0.05	1.96	95
0.1	1.645	90

Table 1.2: Quelques valeurs courantes pour le seuil de risque α

1.3.2 Loi de Khi-deux $\chi 2$

Soit $X_1, X_2, ..., X_\nu$ une suite des variables aléatoires indépendantes de la loi $N(0,1)$ donc $\sum_{i=1}^{\nu} X_i^2$ variable aléatoire suit la loi du Khi deux et sa densité de probabilité est donnée par:

$$f_\nu(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} \mathrm{e}^{-x/2} pour x > 0. (0 sinon)$$

Ou Γ est la fonction gamma d'Euler $\Gamma = \int_0^\infty x^{\alpha-1}\mathrm{e}^{-x}dx$

Proposition 1

1 *La fonction caractéristique donnée par* $(1-2it)^{-\nu/2}$.

2 *L'espérance et la variance de la loi de Khi-deux sont* $E(x) = \nu$ *et* $Var(x) = 2\nu$.

3 *Soient X et Y deux variables aléatoires tel que* $X \sim \chi^2(\nu_1)$; $Y \sim \chi^2(\nu_2)$ *alors* $X + Y \sim \chi^2(\nu_1 + \nu_2)$
Démonstration

1 *Calculons la fonction caractéristique de* χ^2 *lorsque* $X \sim N(0,1)$

$$\varphi(t) = E(e^{itX^2}) = \int_{-\infty}^{+\infty} e^{itX^2)} \frac{1}{\sqrt{2\phi}} e^{-X^2/2} dx. \tag{1.1}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(1-2it)x^2} dx. \tag{1.2}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{e^{-\frac{1}{2}u^2}}{\sqrt{1-2it}} dt. \tag{1.3}$$

$$\tag{1.4}$$

Prendre $u = (\sqrt{1-2it})X$ *alors* $\varphi(t) = (1-2it)^{-1/2}$

1.3.3 Loi de Student $St(\nu)$

Soient X et Y deux variables aléatoires indépendantes si $X \sim N(0,1)$ et $Y \sim \chi^2(\nu)$ alors on a Z variable aléatoire :
$Z = \frac{X}{\sqrt{Y/\nu}} \sim St(\nu)$ à ν degré de liberté.

La densité de loi Student est:

$$f(x) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)(1+x^2/\nu)^{\frac{\nu+1}{2}}}.$$

Proposition 2
Espérance de loi Student: Si $\nu = 1$ *alors l'espérance n'est existe pas; Si* $\nu \geqslant 2$, *l'espérance* $E(x) = 0$.
La variance de loi Student: Si $\nu \leqslant 2$ *la variance n'est pas définie, mais si* $\nu \geqslant 3$ *donc la variance* $var(x) = \frac{\nu}{(\nu-2)}$.
La loi de Student converge en loi vers la loi Normale centrée réduite $N(0,1)$

Remarque 2
Si $\nu = 1$ *nous appelons la loi de Student, loi de Cauchy ou loi de Loretnz.*

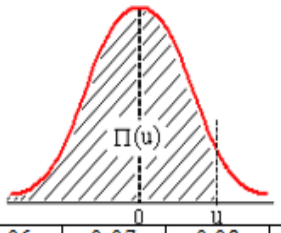
1.3.4 Loi de Ficher-Senedecor $F(\nu_1, \nu_2)$

P et Q deux variables aléatoires indépendantes ou $P \sim \chi^2(\nu_1)$ et $Q \sim \chi^2(\nu_2)$ alors F variable aléatoire donnée par: $F = \frac{P/\nu_1}{Q/\nu_2} \sim F(\nu_1, \nu_2)$ à (ν_1, ν_2) degré de liberté
Sa densité de probabilité est conçue comme suit:

$$f(x) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} (\frac{\nu_1}{\nu_2}) \frac{x^{\nu_1/2-1}}{(1+\frac{\nu_1}{\nu_2 x})^{\frac{\nu_1+\nu_2}{2}}} \quad six > 0 (0 sinon).$$

Proposition 3
L'espérance: si $\nu_2 \geqslant 3$; $E(x) = \frac{\nu_2}{\nu_2-2}$ *et sinon l'espérance n'existe pas. La variance existe si et seulement si* $\nu_2 \geqslant 5$ *tel que* $var(x) = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$
Si $X \sim St(\nu)$ *alors* $X^2 \sim F(1,\nu)$.
Si $Y \sim F(\nu_1,\nu_2)$ *alors* $\frac{1}{Y} \sim F(\nu_2,\nu_1)$.

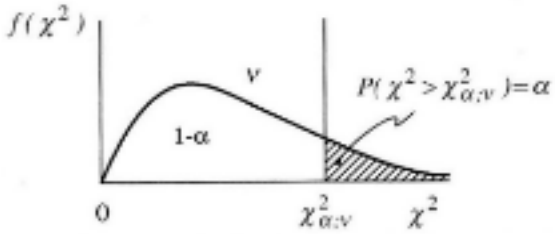
Table de Loi Normale
P(x<u)



	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	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	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	0,8254	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	6,9761	0,9767
2,0	0,9772	0,9778	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
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	.0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Figure 1.3: table de la loi Normale centrée réduite

TABLE DE LA LOI KHI-DEUX (χ^2)

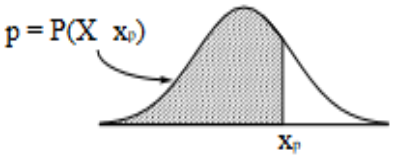


ν \ α	0,995	0,975	0,95	0,9	0,5	0,1	0,05	0,025	0,01	0,005	0,001
1	0,000	0,001	0,004	0,016	0,455	2,706	3,841	5,024	6,635	7,879	10,827
2	0,010	0,050	0,103	0,211	1,386	4,605	5,991	7,378	9,210	10,597	13,815
3	0,072	0,215	0,352	0,584	2,366	6,251	7,815	9,348	11,345	12,838	16,268
4	0,207	0,484	0,711	1,064	3,357	7,779	9,488	11,143	13,277	14,860	18,465
5	0,412	0,831	1,145	1,610	4,351	9,236	11,070	12,832	15,086	16,750	20,517
6	0,676	1,237	1,635	2,204	5,348	10,645	12,592	14,449	16,812	18,548	22,457
7	0,989	1,689	2,167	2,833	6,346	12,017	14,067	16,013	18,475	20,278	24,322
8	1,344	2,179	2,733	3,490	7,344	13,362	15,507	17,535	20,090	21,955	26,125
9	1,735	2,700	3,325	4,168	8,343	14,684	16,919	19,023	21,666	23,589	27,877
10	2,156	3,247	3,940	4,865	9,342	15,987	18,307	20,483	23,209	25,188	29,588
11	2,603	3,815	4,575	5,578	10,341	17,275	19,675	21,920	24,725	26,757	31,264
12	3,074	4,403	5,226	6,304	11,340	18,549	21,026	23,337	26,217	28,300	32,909
13	3,565	5,008	5,892	7,041	12,340	19,812	22,362	24,736	27,688	29,819	34,528
14	4,075	5,628	6,571	7,790	13,339	21,064	23,685	26,119	29,141	31,319	36,123
15	4,601	6,262	7,261	8,547	14,339	22,307	24,996	27,488	30,578	32,801	37,697
16	5,142	6,907	7,962	9,312	15,338	23,542	26,296	28,845	32,000	34,267	39,252
17	5,697	7,564	8,672	10,085	16,338	24,769	27,587	30,191	33,409	35,718	40,790
18	6,265	8,230	9,390	10,865	17,338	25,989	28,869	31,526	34,805	37,156	42,312
19	6,844	8,906	10,117	11,651	18,338	27,204	30,144	32,852	36,191	38,582	43,820
20	7,434	9,590	10,851	12,443	19,337	28,412	31,410	34,170	37,566	39,997	45,315
21	8,034	10,282	11,591	13,240	20,337	29,615	32,671	35,479	38,932	41,401	46,797
22	8,643	10,982	12,338	14,041	21,337	30,813	30,924	36,781	40,289	42,796	48,268
23	9,260	11,688	13,091	14,848	22,337	32,007	35,172	38,076	41,638	44,181	49,728
24	9,886	12,401	13,848	15,659	23,337	33,196	36,415	39,364	42,980	45,558	51,179
25	10,520	13,119	14,611	16,473	24,337	34,382	37,652	40,646	44,314	46,928	52,620
26	11,160	13,843	15,379	17,292	25,336	35,563	38,885	41,923	45,642	48,290	54,052
27	11,808	14,573	16,151	18,114	26,336	36,741	40,113	43,195	46,963	49,645	55,476
28	12,461	15,307	16,928	18,939	27,336	37,916	41,337	44,461	48,278	50,994	56,893
29	13,121	16,047	17,708	19,768	28,336	39,087	42,557	45,722	49,588	52,335	58,302
30	13,787	16,790	18,493	20,599	29,336	40,256	43,773	46,979	50,892	53,672	59,703
40	20,706	24,433	26,051	29,051	39,335	51,805	55,758	59,342	63,691	66,766	73,403
60	35,534	40,481	43,188	46,459	59,335	74,397	79,082	83,298	88,379	91,952	99,608
80	51,171	57,153	60,391	64,278	79,334	96,578	101,879	106,629	112,329	116,321	124,839
100	67,327	74,221	77,929	82,358	99,334	118,498	124,342	129,561	135,807	140,170	149,449

Figure 1.4: table de la loi khi-deux

Table des quantiles de la v.a. de Student

Fournit les quantiles x_p tels que $P(X \ x_p) = p$
pour $X \sim t_{dl}$



p	0.7500	0.9000	0.9500	0.9750	0.9900	0.9950	0.9975	0.9990
dl								
1	1.0000	3.0780	6.3140	12.7060	31.8210	63.6570	127.3213	318.3088
2	0.8160	1.8860	2.9200	4.3030	6.9650	9.9250	14.0891	22.3271
3	0.7650	1.6380	2.3530	3.1820	4.5410	5.8410	7.4533	10.2145
4	0.7410	1.5330	2.1320	2.7760	3.7470	4.6040	5.5976	7.1732
5	0.7270	1.4760	2.0150	2.5710	3.3650	4.0320	4.7733	5.8934
6	0.7180	1.4400	1.9430	2.4470	3.1430	3.7070	4.3168	5.2076
7	0.7110	1.4150	1.8950	2.3650	2.9980	3.4990	4.0293	4.7853
8	0.7060	1.3970	1.8600	2.3060	2.8960	3.3550	3.8325	4.5008
9	0.7030	1.3830	1.8330	2.2620	2.8210	3.2500	3.6897	4.2968
10	0.7000	1.3720	1.8120	2.2280	2.7640	3.1690	3.5814	4.1437
11	0.6970	1.3630	1.7960	2.2010	2.7180	3.1060	3.4966	4.0247
12	0.6950	1.3560	1.7820	2.1790	2.6810	3.0550	3.4284	3.9296
13	0.6940	1.3500	1.7710	2.1600	2.6500	3.0120	3.3725	3.8520
14	0.6920	1.3450	1.7610	2.1450	2.6240	2.9770	3.3257	3.7874
15	0.6910	1.3410	1.7530	2.1310	2.6020	2.9470	3.2860	3.7328
16	0.6900	1.3370	1.7460	2.1200	2.5830	2.9210	3.2520	3.6862
17	0.6890	1.3330	1.7400	2.1100	2.5670	2.8980	3.2225	3.6458
18	0.6880	1.3300	1.7340	2.1010	2.5520	2.8780	3.1966	3.6105
19	0.6880	1.3280	1.7290	2.0930	2.5390	2.8610	3.1737	3.5794
20	0.6870	1.3250	1.7250	2.0860	2.5280	2.8450	3.1534	3.5518
21	0.6860	1.3230	1.7210	2.0800	2.5180	2.8310	3.1352	3.5272
22	0.6860	1.3210	1.7170	2.0740	2.5080	2.8190	3.1188	3.5050
23	0.6850	1.3190	1.7140	2.0690	2.5000	2.8070	3.1040	3.4850
24	0.6850	1.3180	1.7110	2.0640	2.4920	2.7970	3.0905	3.4668
25	0.6840	1.3160	1.7080	2.0600	2.4850	2.7870	3.0782	3.4502
26	0.6840	1.3150	1.7060	2.0560	2.4790	2.7790	3.0669	3.4350
27	0.6840	1.3140	1.7030	2.0520	2.4730	2.7710	3.0565	3.4210
28	0.6830	1.3130	1.7010	2.0480	2.4670	2.7630	3.0469	3.4082
29	0.6830	1.3110	1.6990	2.0450	2.4620	2.7560	3.0380	3.3962
30	0.6830	1.3100	1.6970	2.0420	2.4570	2.7500	3.0298	3.3852

Figure 1.5: table de la loi Student

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	>25
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.08	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
>120	3.84	2.99	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00

Figure 1.6: table de loi de Ficher

1.4 Convergences

1.4.1 Des inégalités utilisables

1 Inégalité de Markov
Si X v.a.réelle, et f fonction croissante et positive (ou null) sur $\mathbb{R}$ vérifier que $f(a) > 0$ alors:

$$P(X \geqslant a) \leqslant \frac{E(f(X))}{f(a)}$$

Démonstration

$$E(f(x)) = \int_{\Omega} f(x)g(x)dx = \int_{X<a} f(x)g(x) + \int_{X\geqslant a} f(x)g(x)dx. \tag{1.5}$$

$$\geqslant \int_{X\geqslant a} f(x)g(x)dx(fpositiveounulle) \tag{1.6}$$

$$\geqslant f(a)\int_{X\geqslant a} g(x)dx.fcroissante \tag{1.7}$$

$$= f(a)P(X \geqslant a). \tag{1.8}$$

Par conséquence $E(f(X)) \geqslant f(a)P(X \geqslant a)$

2 Inégalité de Bienaymé-Chebychev
X v.a accepter une espérance E(X) et de variance finie σ^2 (l'hypothèse de variance finie garantit l'existence de l'espérance conçue comme suit: $\xi > 0$

$$P(|X - E(X)| \geqslant \xi) \leqslant \frac{\sigma^2}{\xi^2}$$

Définition 3
Envisager une suite X_n d'une v.a définie sur Ω et X v.a définie aussi sur Ω
Si $\forall\varepsilon > 0, \lim_{n\to+\infty} P(|X_n - \ell| > \varepsilon) = 0$ alors la suite (X_n) converge en probabilité vers une constante réelle ℓ.
Si $\forall\varepsilon > 0, \lim_{n\to+\infty} P(|X_n - X| > \varepsilon) = 0$. alors la suite X_n est convergé vers X.

Exemple : 2
pour tout entier $n \geqslant 1$; soit X_n une suite de variables aléatoires telles que X_n admet f_n une densité de probabilité donnée par:

$$f_n(x) = 1_{\mathbb{R}^+}(x)n^2x\exp(-n^2x^2/2)$$

Montrons que la suite X_n converge en probabilité vers une variable aléatoire X que l'on précisera.

Soit $\varepsilon > 0$, on a :

$$\begin{aligned} P(|X_n| \geqslant \varepsilon) &= \int_{\varepsilon}^{+\infty} n^2 t \exp(n^2 t^2/2) dt \\ &= \left[-\exp(-n^2 t^2/2)\right]_{\varepsilon}^{+\infty}. \\ &= \exp(-n^2 \varepsilon^2/2) \end{aligned} \tag{1.9}$$

Temps que $n \to +\infty$; ceci tend vers 0. On en déduit que X_n converge en probabilité vers la variable nulle (X=0).

Théorème 1
Soit (Ω, P) espace de probabilité, X_n suite de variables aléatoires sur cet espace accepter des espérances et des variances telles que: $\lim_{n\to\infty} E(X_n) = \ell$ et $\lim_{n\to\infty} V(X_n) = 0$. Donc les X_n convergent en probabilité vers ℓ

1.4.2 Convergence en moyenne quadratique

Définition 4
On dit que $(X_n)_{n\in N}$ une suite de v.a.r converge en moyenne quadratique vers une v.a X si:

$$\lim_{x\to\infty} E((X_n - X)^2) = 0$$

Propriété 2
la convergence en moyenne quadratique donnée la convergence en probabilité.
Ensuite, pour les variables aléatoires (X_n) d'espérance et de variance finies, on dit que X_n converge en moyenne quadratique vers Y si $E(X_n) \to \mu$ et $var(X_n) \to 0$

Preuve 1
On utilise l'inégalité de Marcov avec $Y = |X_n - X|; a = \varepsilon^2 et f(t) = t^2$. Il suffit d'observer que $P(|X_n - X| > \varepsilon)$.
Ensuite on prendre l'hypothèse où $\lim E((X_n - X)^2) = 0$

$$\begin{aligned} \lim E((X_n - \mu)^2) &= \lim E(X_n^2) - 2\mu E(X) + \mu^2 \\ &= \lim E(X_n^2) - (E(X_n))^2 = \lim V(X_n) = 0 \end{aligned} \tag{1.10}$$

1.4.3 Convergence en loi

Définition 5
Soit (Ω, P) espace de probabilité, les variables aléatoires X_n et X sur cet espace de fonction de répartition F_n et F respectivement, si en tout point x ou F continue. les F_n convergent vers $F(x)$, alors les X_n convergent en loi vers X, et on note $X_n \xrightarrow{\ell} X$.

1.5 Estimation Ponctuelle

L'estimation est essentielle pour construire des valeurs approximatives aux paramètres d'une population à partir d'un échantillon de n observations issues de cette population; Il

est possible de passer à la valeur exacte, mais on utilise la meilleure valeur possible que l'on peut donner.

1.5.1 Définitions

1 Le n-échantillon de X est un n-uplet $(X_1, ..., X_n)$ où les X_k ont la même loi que X tel que ces variables aléatoires sont indépendantes. Donc la réalisation de l'échantillon est un n-uplet $(x_1, ..., x_n)$ de valeurs prises par l'échantillon.

2 La statistique de l'échantillon est la variable aléatoire $f(X_1, ..., X_n)$ tel que f est une application de $\mathbb{R}^n$ dans $\mathbb{R}$.

1.5.2 C'est quoi un estimateur

Soit θ une paramètre d'une population, un estimateur de θ est une statistique T qui est considérée comme une bonne valeur du paramètre θ.

Exemple : 3
la moyenne empirique $\bar{X}$ est une estimateur naturel d'espérance $E(X)$ de la loi X cet estimateur produit une estimation $\bar{x}$ moyenne descriptive de la série des valeurs observées.

1.5.3 L'estimation par la méthode de vraisemblance

Soit X v.a réelle de loi paramétrique discrète ou continue on présenté la fonction f comme suit

$$f(x,\theta) = \begin{cases} P_\theta(X = x) & \textit{si X v.a discréte de probabilité ponctuelle P} \\ f_\theta(x) & \textit{si X v.a continue de densité f} \end{cases}$$

Définition 6
La fonction de vraisemblance de θ pour une réalisation d'un échantillon est la fonction de θ

$$L(x_1, ..., x_n; \theta) = f(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Définition 7
la méthode du maximum de vraisemblance est consistante à estimer θ par la valeur maximise L ou

$$\hat{\theta} = \left\{\theta / L(\hat{\theta}) = sup_\theta L(\theta)\right\}$$

1.5.4 Les étapes d'estimation

Vérifier la condition nécessaire $\frac{\partial L(x_1,..,x_n;\theta}{\partial \theta} = 0$
alors on trouve la valeur θ.

Si la condition suffisante est remplie au point critique alors $\theta = \hat{\theta}$ est un maximum local.
On écrive $\frac{\partial^2 L(x_1,...,x_n;\theta}{\partial\theta^2}(\hat{\theta}) \leqslant 0$
ou $\frac{\partial^2 lnL(x_1,...,x_n;\theta)}{\partial\theta^2}(\hat{\theta}) \leqslant 0$.

Exemple : 4
Avec la loi Bernoulli calculons la fonction de vraisemblance:

$$L(X_1, ..., X_n; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$

L'expression de log-vraisemblance est comme suit :

$$\ln L(X_i, p) = \sum X_i \ln p + (n - \sum X_i \ln(1-p)$$

Dérivons par rapport à p :

$$\frac{\partial}{\partial p} \ln [L(X_i, p)] = \frac{1}{p} \sum X_i - \frac{1}{1-p}(n - \sum X_i)$$

Le maximum de vraisemblance est là où la dérivée s'annule.

$$\frac{\partial}{\partial p} [\ln L(X_i, p)]$$

Soit $\hat{p}$ un estimateur sans biais sur un échantillon, on a

$$\begin{aligned} \frac{1}{\hat{p}} \sum X_i &= \frac{1}{1-\hat{p}}(n - \sum X_i)(1-\hat{p}) \sum X_i \\ &= \hat{p}(n - \sum X_i) \sum X_i - \hat{p} \sum X_i \\ &= \hat{p}n - \hat{p} \sum X_i \hat{p} \\ &= \frac{\sum X_i}{n} \end{aligned} \tag{1.11}$$

En fin $\hat{p} = \bar{X}$.

1.5.5 Des estimateurs classique

$\bar{X}$(moyenne empirique) est un estimateur sans biais de moyenne μ,$\bar{x}$ la moyenne observée est son estimation dans une réalisation de l'échantillon.
$\bar{S}^2$ [1] est un estimateur consistant et biaisé de σ^2
$S^2 = \frac{n}{n-1}\bar{S}^2$ est un estimateur sans biais et consistant de σ^2, tel que $\frac{n}{n-1}\sigma_e^2$ est son estimation, ou σ_e est l'écart type observé dans une réalisation de l'échantillon.
f est l'estimation de p la fréquence d'un caractère, F constitue un estimateur sans biais et (consistant).

[1] $\bar{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ la variance empirique

1.5.6 Qualité d'estimateur

Définition 8

1 T estimateur pour θ le biais de T est donné par

$$b_\theta = E(T) - \theta$$

si $E(T) = \theta$ alors T est sans biais.

*2 Si $E(X) \underset{n\to\infty}{\longrightarrow} \theta$ alors on dit que T est convergent. De plus, T est dit **consistant** s'il est convergent vers θ lorsque n tend vers l'infini.*

Théorème 2

Si T convergent et de variance $var(T) \underset{n\to\infty}{\longrightarrow} 0$, alors T est consistant.

1.5.7 La quantité d'information de Ficher

Soit T un estimateur et n est la taille d'échantillon. L représente la vraisemblance de T en fonction des n ($L(X_1, ..., X_n; T)$. On a la fonction de log-vraisemblance $\ln L$, tel que L'information de Ficher est comme suite:

$$I_n(T) = E\left[\left(\frac{\partial \ln L}{\partial T}\right)^2\right]$$

Le cas ou l'ensemble X ne dépend pas de T on a

$$I_n = -E\left(\frac{\partial^2 \ln L}{\partial T^2}\right)$$

et aussi

$$I_n = V\left(\frac{\partial \ln L}{\partial T}\right)$$

1.5.8 Inégalité de Cramer Rao

Soit T un estimateur sans biais du paramètre θ sur un échantillon, X et θ sont indépendantes, la variance de T est bornée inférieurement par l'inverse d'information de Ficher c'est-à-dire:

$$V(T) \geqslant \frac{1}{I_n(T)}$$

1.5.9 Efficacité d'estimateur

Si T un estimateur sans biais est de variance minimale alors on dit qu'un estimateur est efficace dans lequel

$$V(T) = \frac{1}{I_n}$$

Exemple : 5

soit $\hat{p}$ un estimateur sans biais sur un échantillon de loi Bernoulli tel que $\hat{p} = p$
Calculons la fonction de vraisemblance:

$$L(X_1, ..., X_n; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$

L'expression de log-vraisemblance est comme suite :
$\ln L(X_i, p) = \sum X_i \ln p + (n - \sum X_i \ln(1-p)$ *Dérivons par rapport à p :*
$\frac{\partial}{\partial p} \ln [L(X_i, p)] = \frac{1}{p} \sum X_i - \frac{1}{1-p}(n - \sum X_i)$
maintenant, on va calculer l'information de Ficher $I_n(p) = V(\frac{\partial \ln L}{\partial p})$
$I_n(p) = V\left[\frac{1}{p} \sum X_i - \frac{1}{1-p}(n - \sum X_i)\right]$
$I_n(p) = V\left[\sum X_i(\frac{1}{p} + \frac{1}{1-p}) - \frac{n}{1-p}\right]$
On utilisons les propriétés de la variance :

$$I_n(p) = (\frac{1}{p} + \frac{1}{1-p})^2 V(\sum X_i)$$

$$I_n(p) = \frac{1}{p^2(1-p)^2} V(\sum X_i)$$

lorsque les variables aléatoires sont indépendantes on peut écrire
$I_n(p) = \frac{1}{p^2(1-p)^2} \sum V(X_I)$
Rappelons que la variance de loi Bernoulli est p(1-p)
$I_n(p) = \frac{1}{p^2(1-p)^2} \sum p(1-p)$
On obtient $I_n(p) = \frac{n}{p(1-p)}$ *Si* $\hat{p} = \frac{\sum X_i}{n}$ *alors* $V(\hat{p}) = V\frac{\sum X_i}{n} = \frac{1}{n^2} \sum V(X_i)$
vérifiant l'égalité

$$\frac{1}{n^2} \times np(1-p) = \frac{p(1-p)}{n}$$

Par conséquent $V(\hat{p}) = \frac{1}{I_n(p)}$ *;$\hat{p}$ est bien un estimateur efficace.*

1.6 Intervalle de confiance

Soit $\alpha \in]0, 1[$ est un niveau de risque fixé.

Définition 9

1 *L'intervalle de confiance de θ de niveau de confiance $1-\alpha$ est un ensemble $C(X) \subset \Theta$ telle que, quelque soit θ on a*

$$P_\theta(\theta \in C(X)) \geq 1 - \alpha.$$

2 *Pour une loi de probabilité P Le quantile d'ordre α est la quantité z_α*

$$z_\alpha = inf\left\{x, P(]-\infty, x]) \geqslant \alpha\right\}.$$

Exemple : 6
pour la loi Normale centré réduite, le quantile d'ordre $97,5\%$ *est 1,96.*

1.6.1 Estimation par intervalle

1.6.1.1 Estimation de la moyenne

1 Quand la variance connue
On a X v.a suit la loi $N(\mu, \sigma^2)$

Théorème 3
Si σ^2 *est connu l'intervalle de confiance au niveau* $1-\alpha$ *de* μ *est donné par*

$$\left[\bar{X}_n - t_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + t_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

$t_{1-\alpha/2}$ *est le quantile d'ordre* $1-\alpha/2$ *de la loi* $N(0,1)$.

2 Quand la variance inconnue

Théorème 4
Si σ^2 *est inconnu alors, l'intervalle de confiance au niveau* $1-\alpha$ *de* μ *est conçu comme suite:*

$$\left[\bar{X}_n - t_{n-1,1-\alpha/2}\frac{\sqrt{\hat{S}_n^2}}{\sqrt{n}}, \bar{X}_n + t_{n-1,1-\alpha/2}\frac{\sqrt{\hat{S}_n^2}}{\sqrt{n}}\right]$$

$t_{n-1,1-\alpha/2}$ *est le quantile d'ordre* $1-\alpha/2$ *de la loi de Student* $n-1$ *ddl.*

1.6.1.2 Estimation de la variance

1 Lorsque la moyenne connu

Théorème 5
Si μ *est connu alors l'intervalle de confiance au niveau* $1-\alpha$ *de* σ^2 *est conçu comme suit:*

$$\left[\frac{1}{q_2}\sum_{i=1}^{n}(X_i-\mu)^2, \frac{1}{q_1}\sum_{i=1}^{n}(X_i-\mu)^2\right]$$

q_1*et* q_2 *sont les quantiles d'ordre* $\alpha/2$ *et* $1-\alpha/2$ *de la loi* χ^2 *à n ddl.*

2 Lorsque la moyenne est inconnu

Théorème 6

Si μ est inconnu alors l'intervalle de confiance au niveau$1-\alpha$ de σ^2 est conçu comme suit:

$$\left[\frac{1}{q_2}\sum_{i=1}^{n}(X_i-\bar{X_n})^2, \frac{1}{q_1}\sum_{i=1}^{n}(X_i-\bar{X_n})^2\right]$$

$q_1$et q_2 sont les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi χ^2 à $n-1$ ddl.

1.7 Test d'hypothèse

1.7.1 Quelque définitions

Hypothèse statistique: est une affirmation concernant les valeurs paramétrique, la forme de la distribution des observations de la population.
Test d'hypothèse: appelé aussi un test statistique est essentiellement pour donner une règle de décision, sur une base de résultat d'échantillon, on peut donc choisir entre deux hypothèses statistiques.
Hypothèse nulle: notée H_0 est quand on fixe le paramètre de la population à une valeur particulier.
Hypothèse alternative: notée H_1 est l'hypothèse qui diffère l'hypothèse nulle H_0.
*Seuil de signification:*Lorsque l'hypothèse nulle est vrai on note α Seuil de signification du test tel que:

$$\alpha = P(rejeterH_0|H_0vraie)$$

Remarque 3

la seuil signification les plus utilisées sont $\alpha = 0.05$ ou $\alpha = 0.01$.

1.7.2 Les étapes pour tester une hypothèse

Premièrement on définir l'hypothèse nulle à étudier; deuxièmement, on chercher un test statistique pour contrôler l'hypothèse nulle et définir le niveau de signification α, puis à l'aide des données représenté par l'échantillon on va calculer la valeur de la statistique; finalement faire une décision à partir l'hypothèse posé et donner un interprétation.

Chapter 2

Tests d'indépendance

Avant de faire n'importe quelle étude d'inférence statistique,doit faire la vérification de l'hypothèse d'indépendance, cette notion est fondamentale en statistique inférentielle.

2.1 Le concept d'indépendance

Si une série chronologique est indépendante, la densité de probabilité jointe de cette sérié, peut être écrit comme un produit de densités marginales pour chacune des n variables aléatoires.
$f_{x_1,x_2,...,x_n}(x_1, x_2, ..., x_n) = f_{x_1}(x_1)f_{x_2}(x_2), ..., f_{x_n}(x_n)$
L'indépendance est essentielle pour plusieurs solutions statistiques. Dans un contexte expérimental, cette hypothèse est souvent garantie, ou presque, soit quand les prélèvements sont faits par un procédé (ou aléatoirement), qui sert à randomiser, où les traitements sont affectés aléatoirement aux unités expérimentales. Dans l'analyse d'une série chronologique, l"hypothèse d'indépendance est cruciale, car les observations de la chronique sont assemblées au cours du temps.Dans une telle situation, on peut dire que les observations ainsi recueillies ne sont pas sans rapport avec le passé. Donc on va construire des différents type de dépendance affectant les observations de la série chronologique. Parmi ces types de dépendance, il y a un effet de persistance avec une valeur n"est pas indépendante de une ou des valeurs précédentes. Un effet de tendance monotone où l'espérance mathématique (la moyenne) de la série croit ou décroît avec le temps de façon continue. Des effets cycliques ou pseudo-cycliques tels que l'espérance mathématique d'une valeur observée est fonction de la chronologie.
Donc, pour vérifier l'indépendance des observations d'une série chronologique, on doit représenter l'hypothèse d'indépendance à des hypothèses alternatives raisonnables. et bien spécifiées de telle manière que celles-ci correspondent à l'un de ces trois types de dépendance qui peut affecter les observations d'une chronique. En pratique, dans un échantillon utilisé pour analyser les caractéristiques d'un phénomène hydrologique, les débits d'une rivière par exemple, il serait préférable de considérer le débit maximum annuel afin d'éviter le plus possible le problème de dépendance. En effet, si l'on considère, par exemple des débits journaliers, il est probable qu'on introduise une dépen-

dance entre les observations de la chronique. Dans le présent contexte, si l'on considère l'hypothèse d'indépendance comme étant l'hypothèse nulle, l'hypothèse alternative associée est l'existence de l'un ou l'autre type de dépendance (persistance, tendance ou cyclicité).

2.2 L'indépendance de deux variables qualitatives

Ici, la condition essentielle est les valeurs de l'échantillon sont indépendantes. Par un test, on va vérifier l'hypothèse d'indépendance, ce test poser un variable aléatoire qui suit la loi de χ^2. On effectuer ce test à l'aide de la statistique de khi deux qui est très complexe à calculer pour cette raison on utilise le tableau de contingence des observations et le tableau des valeurs attendus ou théorique; donc le calcul devient plus facile.
n:la taille de l'échantillon.

Y / X	mod 1 ...	mod j...	
mod 1	n_{11}	n_{1j}	$n_{1.}$
mod i		n_{ij}	$n_{i.}$
	$n_{.1}$	$n_{.j}$	n

Table 2.1: tableau de modalité

$n_{i.}$: la fréquence de modalité i de la v.a X.
$n_{.j}$: la fréquence de la modalité j de la v.a Y.
$n_{1.}/n$: une estimation de la probabilité que la v.a X prenne la modalité 1.
$n_{.1}/n$: une estimation de la probabilité que la v.a Y prenne la modalité 1.
n_{11}/n: une estimation de la probabilité que les v.a X et Y prennent la modalité i et j respectivement.
Si il y a un indépendance on devra vérifier que: $\frac{n_{ij}}{n} \simeq \frac{n_{i.}}{n} \times \frac{n_{.j}}{n}$.
Prenons $T_{ij} = \frac{n_{i.} \times n_{.j}}{n}$ la fréquence attendue pour les modalités i et j s'il y a avait une indépendance.
On calcule la valeur de la variable de test χ^2 d'indépendance:

$$\chi_c^2 = \sum_{i=1}^{l} \sum_{j=1}^{m} \frac{(n_{ij} - T_{ij})^2}{T_{ij}}$$

Où l représente le nombre de modalités de X, et m représente le nombre de modalités de Y; à l'aide de cette statistique on peut mesurer l'indépendance entre les variables X et Y.
Hypothèse testée H_0: Les variables X et Y sont indépendantes.
On cherche la valeur critique χ_ν^2 dans la table de la loi du χ^2à $(l-1)(m-1)$ degrés de liberté
résolution: accepter H_0 si $\chi_c^2 < \chi_\nu^2$. Sinon on la rejette.

Exemple : 7

Pour comparer l'efficacité de deux médicaments agissant sur la même maladie, mais aux prix très différents, la Sécurité Sociale a effectué une enquête sur les guérisons obtenues en suivant chacun des traitements. Les résultats sont consignés dans le tableau suivant:

Les effectifs marginaux sont les suivants:

	Médicament cher	Générique
Guérisons	48	158
Non guérisons	6	44

Table 2.2: tableau de contingence

Les effectifs théorique(fréquences):

	Médicament cher	Générique	totale de ligne
Guérisons	48	158	206
Non guérisons	6	44	50
Totale de colonne	54	202	256

Table 2.3: tableau de totaux de ligne et totaux de colonnes

	Médicament cher	Générique	totale de ligne
Guérisons	$\frac{54\times206}{256}$	$\frac{202\times206}{256}$	206
Non guérisons	$\frac{54\times50}{256}$	$\frac{202\times50}{256}$	50
Totale de colonne	54	202	256

Table 2.4: tableau des fréquence

$\chi^2 = \frac{(48-43,45)^2}{43,45} + \frac{(158-162,55)^2}{162,55} + \frac{(6-10,55)^2}{10,55} + (\frac{(44-39,45)^2}{39,45} \simeq 3,1$ *La variable de test* χ^2 *vaut approximativement, à l'aide de table de* χ^2 *la valeur critique par un niveau de risque de* 5% *est* $\chi^2_1 = 3,84$ *ou* ($\chi^2_c < \chi^2_\nu$)*,alors on accepte l'hypothèse nulle*

On conclure que le taux de guérison ne dépend pas du prix du médicament et se poser des questions sur l'opportunité de continuer à vendre le médicament cher.

2.3 L'indépendance de deux variables quantitatives

Un échantillon composé de n paires d'observation extrait de population qui suivent la loi Normale, et r le coefficient de corrélation de l'échantillon. Il s'agit de tester l'hypothèse nulle:
$H_0 : \rho = 0$ (corrélation nulle entre les populations)au risque α
$H_1 : \rho \neq 0$ ($\rho > 0$ liaison positive, $\rho < 0$ liaison négative)

On peut démontrer que la v.a $T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim St(\nu)$ ou $\nu = n - 2$ degré de liberté à l'aide de H_0 avec les étapes suivantes:

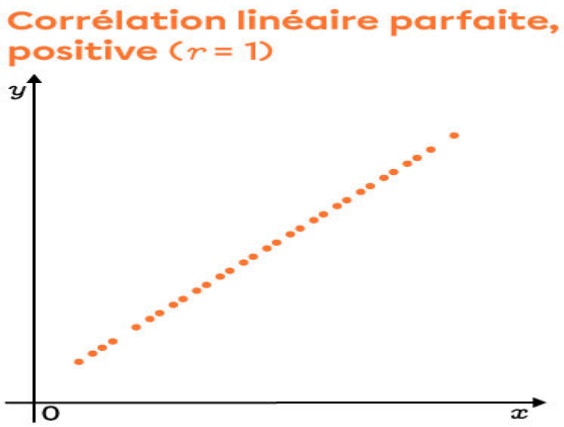
1 On calculera $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.

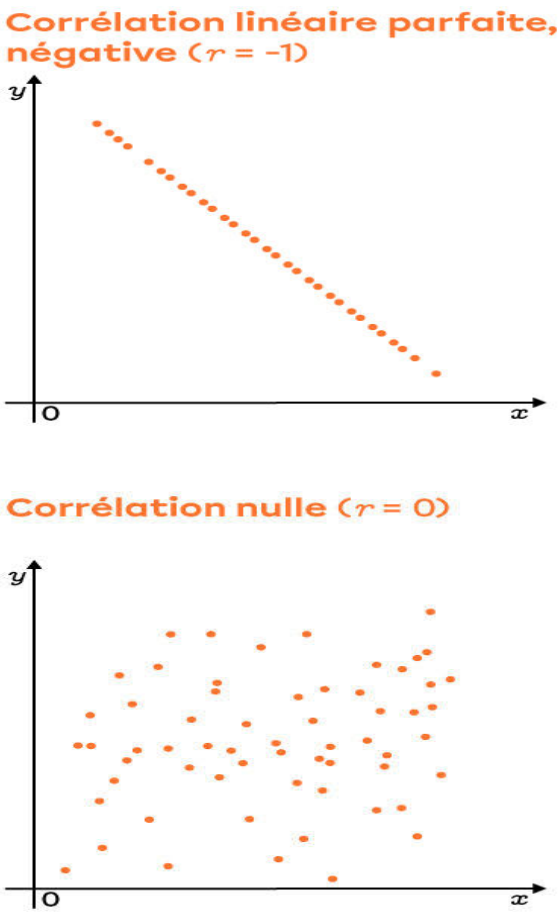
2 On déduire la valeur t_α ou $t_{\alpha/2}$dans la table de loi t de Student à $\nu = n - 2$ degré de liberté,tel que $P(T_{n-2} > t_{\alpha/2}) = \alpha/2$.

3 Analyser la règle de décision comme suite

1. Si l'hypothèse alternative $H_1 : \rho \neq 0$(cas bilatéral):rejet deH_0 au risque α si $t \notin \left]-t_{\alpha/2}; t_{\alpha/2}\right[$.
2. si l'hypothèse alternative $H_1 : \rho > 0$(cas unilatérale):rejet H_0 au risque α si $t > t_\alpha$avec $\nu = n - 2$ degré de liberté.
3. si l'hypothèse alternative $H_1 : \rho < 0$(cas unilatérale):rejet H_0 au risque α si $t < -t_\alpha$ avec $\nu = n - 2$ degré de liberté.

2.3.1 Nuage du point et la corrélation





2.4 Test d'indépendance de variable qualitative et quantitative

Pour étudier la liaison entre une variable qualitative et variable quantitative on utilise le test de Student si on a un variable qualitative à deux modalités, et on fait une ANOVA (analyse de la variance), si il y a une variable qualitative avec plus de deux modalités.

2.4.1 Test de Student

Ce test est l'un des tests paramétriques qui permet de comparer les moyennes de deux groupes, le test de Student suppose que les variables aléatoires suivent une distribution Normal et que les variances sont égales, alors , on tester l'hypothèse nulle suivante :

$$H_0 : m_X = m_Y$$
$$H_1 : m_X \neq m_Y$$

Où X et Y sont deux groupes (échantillons) différents, m_X et m_Y sont leurs moyennes (respectivement).Pour n_X taille du groupe X et n_Y taille du groupe Y, on définie la valeur

t de Student comme suit:

$$\frac{m_X - m_Y}{\sqrt{\frac{S^2}{n_X} + \frac{S^2}{n_Y}}} \sim St(n_X + n_Y - 2)$$

Le calcul de la variance commune S^2 aux deux échantillons est donné par:

$$S^2 = \frac{\sum(x - m_X)^2 + \sum(x - m_Y)^2}{n_X + n_Y - 2}$$

Pour tester l'indépendance tout d'abord il faut donner t_c la valeur critique de table de Student à $n_X + n_Y - 2$ ddl au risque $\alpha = 5\%$
Si $|t| > t_c$ on rejette H_0 sinon on accepte H_0.

2.4.2 Test d'ANOVA

On applique ce test d'ANOVA si les conditions suivantes sont remplies.

1 La population est gaussienne.

2 La variance de chaque sous-population est identique.

3 l'indépendance des sous-échantillons.

On va tester l'hypothèse suivante:

$H_0 : \mu_1 = \mu_2 = ... = \mu_k = \mu$
$H_1 : \exists j, \mu_j \neq \mu$(*il exist une moyenne d'un échantillon qui différente des autres*

$\bar{x_j} = \frac{1}{n_j} \sum_{i=j}^{n_j} x_{ij}$ est la moyenne de chaque échantillon.
$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} n_j \bar{x_j}$ est la moyenne globale.
On obtient l'équation de l'analyse de variance comme suit: SCT=SCE+SCR

$$\sum(\bar{x_{ij}} - \bar{x}) = \sum(\bar{x_j} - \bar{x})^2 + \sum(x_{ij} - \bar{x_j})^2$$

tel que:
SCT: est la somme des carrés totale
SCE: est la somme des carrées expliquées.
SCR: est la somme des carrées résiduelles. La statistique du test est:

$$F = \frac{SCE/k - 1}{SCR/n - k} \sim F_c(k - 1, n - k)$$

Règle de décision: Si $|F| > F_c$ on rejette H_0, sinon on accepte H_0.

2.5 L'indépendance contre l'existence d'une persistance

Le test d'hypothèse d'indépendance s'intéresse à la vérification s'il existe un effet de persistance entre les observations successives de la série chronologique ou non. On vérifier donc si les valeurs faibles (respectivement élevées) de la chronique ont ou non-tendance à suivre des valeurs faibles (respectivement élevées). On montre donc à cette hypothèse. Les tests non-paramétriques d'indépendance, d'application très simple et les tests paramétriques d'indépendance basée en général sur l'auto-corrélation d'ordre 1. Les hypothèses à tester sont :

- H_o : Les observations sont indépendantes
- H_1 : les observations sont auto corrélées.

2.5.1 Test non-paramétrique d'indépendance

On va vérifier s'il y a une auto-corrélation entre les observations successives des $X_t(t = l, ..., n)$. Il est facile de s'attendre à ce que des valeurs élevées (respectivement faibles) suivent fréquemment des valeurs élevées (respectivement faibles). Cette remarque adopte les tests qui vont suivre.

2.5.1.1 Test des groupe

Malinvaud (1978). Pour appliquer ce test, soit la variable R où est le nombre des groupes formés des observations consécutives toutes supérieures (ou toutes inférieures) à la médiane de la chronique. Pour $n \geqslant 50$ la taille d'échantillon (chronique), on peut montrer que la variable R suit approximativement une distribution Normale de moyenne E(R) et de variance Var(R) données par:
$E(R) = \frac{n+1}{2}$.
$Var(R) = \frac{n-1}{4}$.
alors la statistique du test est:
$Z = \frac{R-E(R)}{\sqrt{Var(R)}} \sim$ N(0,1)
on appliquant ce test par les étapes suivantes:

- La vérification de la taille de la chronique $n \geq 50$.
- Calculons le nombre de groupes formés des observations consécutives toutes supérieures (ou inférieures) à la médiane de la chronique x_v donnée par:

$$x_v = \begin{cases} x_{\frac{n+1}{2}} & \textit{si n impaire} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \textit{si n paire} \end{cases}$$

- Tester l'hypothèse d'indépendance par calculer la valeur observée de la statistique du test sous l'hypothèse nulle suivante:
  $Z_{obs} = \frac{R - \frac{n+2}{2}}{\sqrt{\frac{n-1}{4}}}$
- Au niveau de signification α rejeter $H_0 si : |Z_{obs}| > Z_{(1-\frac{\alpha}{2})}$ où $Z_{(1-\frac{\alpha}{2})}$ est le quantile d'ordre $(1 - \frac{\alpha}{2})$ de la loi N(0.1).

2.5.1.2 Test de Walis et Moore

Malinvaud (1978); ce test connu aussi test des points de retournement, considérons P le nombre de points de retournement, c'est-à-dire
P= le nombre des observations X_t telles que:$(x_{t+1} - x_t)(x_t - x_{t-1}) < 0$
$(t = 2, ..., n)$ Pour un échantillon (chronique) de taille $n \geq 50$, on peut démontrer que la variable P suit approximativement une distribution Normale de moyenne E(P) et de variance Var(P) données par :
$E(P) = \frac{2}{3}(n-2)\ V(P) = \frac{16n-29}{90}$
alors la statistique de ce test est:
$Z = \frac{P - E(P)}{\sqrt{Var(P)}}$
on appliquant ce test par les étapes suivantes:

- La vérification de la taille de la chronique $n \geq 50$.
- Calculer la valeur de P.
- Tester l'indépendance par calculer la valeur observée de la statistique du test sous 1 'hypothèse nulle suivante:
  $Z_{obs} = \frac{P - \frac{2}{3}(n-2)}{\sqrt{\frac{16n-29}{90}}}$
- Au niveau de signification α rejeter H_0 si $|Z_{obs}| > Z_{(1-\frac{\alpha}{2})}$ ou $Z_{(1-\frac{\alpha}{2})}$ est le quantile d'ordre$(1 - \frac{\alpha}{2})$ de la loi $N(0, 1)$.

2.5.1.3 Test de Von Neumann

Soit la variable η appelée aussi rapport de Von Neumann qui est le rapport de la moyenne des carrés des différences successives à la variance de l'échantillon sa valeur donnée par

$$\eta = \frac{n \sum_{t=1}^{n-1} (x_{t+1} x_t)^2}{n - 1 \sum_{t=1}^{n} (x_t - \bar{x})^2} \tag{2.1}$$

avec $\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t$
Pour une taille d'échantillon $n \geqslant 30$, η variable suit approximativement une distribution Normale de moyenne et de variance données par :
$E(\eta) = 2n/n - 1, Var(\eta) = 4\frac{n-2}{(n-1)^2}$
donc la statistique du test donnée est : $Z = \frac{\eta - E(\eta)}{\sqrt{Var(\eta)}} \sim N(0, 1)$
on appliquant ce test par les étapes suivantes :

- La vérification de la taille d'échantillon $n \geqslant 30$.
- Calculer η (2.1).
- Tester l'hypothèse nulle d'indépendance par calculer la valeur observée de la statistique du test sous l'hypothèse nulle donnée par:

$$Z_{obs} = \frac{\eta - 2n/n - 1}{\sqrt{4\frac{n-2}{(n-1)^2}}}.$$

- Au niveau de signification α rejeterH_o si$|Z_{obz}| > z_{\frac{1-\alpha}{2}}$
  $z_{\frac{1-\alpha}{2}}$ est le quantile d'ordre $\frac{1-\alpha}{2}$de la loi N(0,1)

2.5.1.4 Test de Wald-Wolfowitz (1978)

Pour appliquer ce test, soit la variable R donnée par :

$$R = \sum_{t=1}^{n-1} = x_t x_{t+1} + x_1 x_n \tag{2.2}$$

Pour $n > 40$; sous l'hypothèse nulle d'indépendance, R suit approximativement une distribution Normale de moyenne E(R)
$E(R) = \frac{(\sum_{t=1}^{n} x_t)^2 - x_t^2}{n-1}$
et de variance Var(R) donnée par:
$\frac{(\sum_{t=1}^{n} x_t^2)^2 - \sum_{t=1}^{n} x_t^4}{n-1} + \frac{(\sum_{t=1}^{n} x_t)^4 - 4(\sum_{t=1}^{n} x_t)^2 \sum_{t=1}^{n} x_t^2 + 4\sum_{t=1}^{n} x_t \sum_{t=1}^{n} x_t^3 + (\sum_{t=1}^{n} x_t^2)^2 - 2\sum_{t=1}^{n} x_t^4}{(n-1)(n-2)} - (E(R))^2$
donc la statistique du test est: $Z = \frac{R - E(R)}{\sqrt{Var(R)}} \sim N(0,1)$

on appliquant ce test par les étapes suivantes :

- La vérification de la taille de la chronique $n > 40$.
- Calculer R (2.2).
- Tester 1'hypothèse nulle d'indépendance par calculer la valeur observée de la statistique du test sous l'hypothèse nulle donnée par:

$$Z_{obs} = \frac{R - \frac{(\sum_{t=1}^{n} x_t)^2 - \sum_{t=1}^{n} x_t^2}{n-1}}{\sqrt{\frac{(\sum_{t=1}^{n} x_t^2)^2 - \sum_{t=1}^{n} x_t^4}{n-1} + \frac{(\sum_{t=1}^{n} x_t)^4 - 4(t=1^n x_t)^2 \sum_{t=1}^{n} x_t^2 + 4\sum_{t=1}^{n} x_t \sum_{t=1}^{n} x_t^3 + (\sum_{t=1}^{n} x_t^2)^2 - 2\sum_{t=1}^{n} x_t^4}{(n-1)(n-2)} - (E(R))^2}}$$

- Au niveau de signification α rejeterH_o si$|Z_{obz}| > z_{\frac{1-\alpha}{2}}$
  $z_{\frac{1-\alpha}{2}}$ est le quantile d'ordre $\frac{1-\alpha}{2}$de la loi N(0,1)

2.5.2 Test paramétrique d'indépendance

Quand un coefficient d'auto-corrélation d'ordre $(B \neq 0)$, les observations de la chronique sont auto-corrélées. Donc, naturellement, on applique un test d'indépendance sur ces coefficients d'auto-corrélation ρ_θ conçu comme suit :

$$\rho_\theta = \frac{E(x_t - \bar{x})(x_{t+\theta} - \bar{x})}{Var(x)} \tag{2.3}$$

Ce test sera basé sur une estimation du coefficient d'auto-corrélation d'ordre θ noté $r_\theta$et est définie par:
$\rho_\theta = \frac{\sum_t^n x_t x_{t+\theta} - \frac{1}{n}(\sum_t^n x_t)^2}{\sum_t^n x_t^2 - \frac{1}{n}(\sum_t^n x_t)^2}$;où $x_{t+\theta} = x_{t+\theta+n}$
pour toutes valeurs$t + \theta > n$.
r_θ est un estimateur naturel de ρ_θ de plus est une fonction symétrique des n valeurs observées de la chronique.
La loi de distribution d'un ensemble de plusieurs ρ_θ est connue, mais son utilisation en pratique est rare, les tests étant appliqué habituellement sur un seul coefficient, le plus souvent celui du premier ordre. Donc on estime juste pour l'analyse des débits de crue annuels, il n'est pas essentiel de considérer des coefficients d'auto-corrélation d'ordre supérieur à 1. Les tests paramétriques qui vont suivre, sont basés sur ρ_θ et dans le cadre des observations de la chronique proviennent d'une population gaussienne.

2.5.2.1 Test de Ljung-box(1994)

Il est le unique test entre les autre qui appliquer un coefficient d'auto-corrélation d'ordre supérieur à 1. ensuite il a été mentionné, si les coefficients d'auto-corrélation sont différents de zéro, alors les observations d'une chronique sont indépendantes Donc ,étudier l'indépendance des observations contre l'existence d'une auto corrélation revient à faire le test définie par:

$$\begin{cases} H_0 : \rho_1 = \rho_2 = ... = \rho_\theta = 0 \\ H_1 : \textit{au moins un} \rho_i \neq 0 \end{cases}$$

La statistique du test Q_θ est calculée en fonction de n (taille de l'échantillon). Pour $n > 40$, Ljung et Box (1978) ont supposé que
$Q_\theta = n(n+2)\sum_{t=1}^{\theta} \frac{\rho_t^2}{n-t} \sim \chi_\theta^2$($Q_\theta$ suit une loi de Khi-deux à θdegrés de liberté)
Pour une chronique composée d'observations sur une base annuelle, les auteurs suggèrent de prendre$\theta = 5$.
On appliquant ce test par les étapes suivantes:

- Vérifier que les observation de la chronique sont gaussiennes.
- tester 1 'hypothèse nulle d'indépendance par calculer la valeur observée de la statis-

tique du test sous l'hypothèse nulle suivante :

$$Q_{\theta_{obs}} = n(n+2)\sum_{t=1}^{\theta}\frac{r_\theta^2}{(n-t)} \tag{2.4}$$

- Au niveau de signification α rejeter H_0 si $Q_{\theta_{obs}} \notin \left[\chi^2_{(\frac{\alpha}{2};\theta)}, \chi^2_{(1-\frac{\alpha}{2};\theta)}\right]$ ou $\chi^2_{(\frac{\alpha}{2};\theta)}$ et $\chi^2_{(1-\frac{\alpha}{2};\theta)}$ sont respectivement les quantiles d'ordre $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ d'une loi χ^2_θ

2.5.2.2 Test de box-Pierce (1970)

on applique les mêmes hypothèses du test précédent, pour un échantillon $n < 40$, Box et Pierce ont supposé que:

$$Q_\theta = n\sum_{t=1}^{\theta}\rho_t^2 \sim \chi^2_\theta \tag{2.5}$$

Pour une chronique composée d'observations sur une base annuelle, $\theta = 5$ (une hypothèse avancée par les auteurs).
On appliquant ce test par les étapes suivantes:

- Vérifier que les observations de la chronique sont gaussienne.
- Tester 1 'hypothèse nulle d'indépendance par calculer la valeur observée de la statistique du test sous l'hypothèse nulle suivante:

$$Q_{\theta_{obs}} = n\sum_{t=1}^{\theta} r_t^\theta.$$

- Au niveau de signification α rejeter H_0si $Q_{\theta_{obs}} \notin \left[\chi^2_{(\frac{\alpha}{2};\theta)}, \chi^2_{(1-\frac{\alpha}{2};\theta)}\right]$ou $\chi^2_{(\frac{\alpha}{2};\theta)}et\chi^2_{(1-\frac{\alpha}{2};\theta)}$sont respectivement les quantiles d'ordre $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ d'une loi χ^2_θ

2.5.2.3 Test de Bartllet (1993)

Pour appliquer ce test , on envisager les séries A et B suivantes:
Série P : $x_1, x2, ..., x_{n-1}$
Série Q : $x_2, x_3, ..., x_n$
On pose ρ_1 le coefficient de corrélation entre ces deux séries (c'est l'équivalent du premier coefficient d'auto-corrélation de la série $x_i tq(i = 1, 2, ..., n)$.Tester l'hypothèse d'indépendance, par le test définie par :

$$\begin{cases} H_0 : \rho_1 = 0 \\ H_1 : \rho_1 \neq 0 \end{cases}$$

$T = \rho_1\frac{\sqrt{n-3}}{\sqrt{1-\rho_1^2}}$;est la statistique du test, elle suit une distribution de Student $n-3$ degrés de liberté.

Bartleu (1935) suggère afin d'accroître l'efficacité de ce test, de prendre le nombreυ de degrés de liberté tels que :

$$\upsilon = (n-3)\frac{1 - r_P r_Q}{1 + r_P r_Q}$$

$\upsilon = (n-3)\frac{(1-r_1^2)}{(1+r_1^2)}$ comme $r_P \simeq r_Q \simeq r_1$
Finalement,la statistique du test est:
$T = \frac{\rho_1\sqrt{r}}{\sqrt{1-\rho_1^2}} \sim St_{(\upsilon)}$(T est une loi de Student υ de degrés de liberté)
On appliquant ce test par les étapes suivantes:

- Former la chronique initiale pour obtenir les séries P et Q.
- déduire le coefficient de corrélation r_1 entre ces deux séries.
- Tester l'hypothèse nulle par calculer la valeur observée de la statistique du test suivante :

$$T_{obs} = \frac{r_1\sqrt{\upsilon}}{\sqrt{1 - r_1^2}}.$$

- Au niveau de signification α rejeter H_0 si $|T_{obs}| > t_{(1-\alpha/2;\upsilon)}$ ou$t_{(1-\alpha/2;\upsilon)}$est le quantile d'ordre $1 - \alpha/2$ d'une loi St_υ

2.5.2.4 Test d'Anderson(1941)

Bobée et al. (1978). Il est essentiel pour des séries circulaires, c'est-à-dire lorsque, la dernière valeur de la série chronologique est suivie par la première. Pour une chronique de taille k, le coefficient d'auto-corrélation d'ordre 1; $r_1$est définie par:

$$r_1 = \frac{\frac{1}{k}\sum_i^k a_i a_{i+1} - \frac{1}{k^2}(\sum_{i=1}^k a_i)^2}{\frac{1}{k}\sum_{i=1}^k a_i^2 - \frac{1}{k^2}(\sum_{i=1}^k a_i)^2}.$$

D'après ce test, on peut montrer lorsque la population est Normale de taille k que sous l'hypothèse nulle suit approximativement une loi Normale de variance $Var(r_1)$et de moyenne$E(r_1)$ définie par:

$$var(r_1) = \frac{k-2}{(k-1)^2}.$$

$$E(r_1) = \frac{-1}{k-1}$$

Donc,sous l'hypothèse nulle la statistique du test est:
$z = \frac{r_1 - E(r_1)}{\sqrt{Var(r_1)}} \sim N(0,1)$
On appliquant ce test par les étapes suivantes :

- vérifier que les observations de la chronique suivant la loi Normale.
- Tester l'hypothèse nulle d'indépendance à partir de la statistique du test z et calculé les valeurs observées données par :

$$z_{obs} = \frac{r_1 + \frac{1}{k-1}}{\sqrt{\frac{k-2}{(k-1)^2}}}$$

- Au niveau de signification α rejeter H_0, si $|Z_{obs}| > z_{(1-\frac{1}{\alpha})}$ tq $z_{(1-\frac{1}{\alpha})}$ est le quantile d'ordre$1 - \frac{1}{\alpha}$ de la loi Normale centrée réduite.

2.6 L'indépendance contre l'existence d'une tendance

Dans cette section, on présente des différents tests d'indépendance contre l'existence d'une tendance monotone, les hypothèses à tester sont :

$$\begin{cases} H_0 : \textit{les observations sont indépendantes} \\ H_1 : \textit{il existe une tendance dans les observations} \end{cases}$$

Les tests qui vont suivre sont essentiellement non paramétriques

2.6.1 Test de Foster et Stuart (1954)

Soit la variable D telle que $D = w_r + v_r$ pour $(t < t')$
w_r : le nombre d'observation tq $x_t > x_{t'}$
v_r : le nombre d'observation tq $x_t < x_{t'}$. On peut montrer que la valeur $D \sim N(\mu, \sigma)$ de moyenne $E(D)$ et variance$var(D)$; pour une chronique de taille $n \geqslant 40$ où $E(D) = 0$ et $var(D) = 2log(n - 0,8756)$ on obtient la statistique du test comme suit

$$Z = \begin{cases} \frac{D-E(D)-1/2}{\sqrt{var(D)}}, & siD > E(D) \\ \frac{-D+E(D)+1/2}{\sqrt{var(D)}}, & siD < E(D) \end{cases}$$

de plus $z \sim N(0, 1)$
On applique ce test par les étapes suivantes :

- La vérification de la taille de la chronique $n \geqslant 40$.
- Calculer D.
- Tester l'hypothèse nulle d'indépendance par calculer la valeur observée de la statistique sous l'hypothèse nulle suivante :

$$Z_{obs} = \begin{cases} \frac{D-1/2}{\sqrt{2log(n-0,8756)}}, & siD > E(D) \\ \frac{-D+1/2}{\sqrt{2log(n-0,8756)}}, & siD < E(D) \end{cases}$$

- Au niveau de signification on rejeter H_0 si $|Z_{obs}| > z_{1-\alpha/2}$ tel que $z_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi Normale centré réduite

2.6.2 Test de Cox Stuart

Soit la variable R définie par $R = \sum_{t=1}^{n/2}$ où :

$$R = \begin{cases} 1 & si(x_t - x_{n/2+t} > 0 \\ 0 & sinon \end{cases} \tag{2.6}$$

pour $n \geqslant 20$ la taille de l'échantillon, sous l'hypothèse nulle H_0, $R \sim N(\mu, \sigma)$ ou la moyenne est $E(R)$ et la variance est $var(R)$ données par:
$E(R) = n/2$ et $var(R) = n/4$; alors on obtient la statistique du test comme suit :

$$Z = \frac{R - E(R)}{\sqrt{var(R)}}$$

qui suit la loi Normale centrée réduite.
On appliquant ce test par les étapes suivantes:

- la vérification de la taille de la chronique $n \geqslant 20$
- Calculer R (2.6).
- Tester 1 'hypothèse nulle d'indépendance par calculer la valeur statistique observée donnée par :

$$Z_{obs} = \frac{R - n/2}{\sqrt{n/4}}$$

- au niveau de signification α on rejeter l'hypothèse nulle si: $|Z_{obs}| > z_{1-\alpha/2}$ tel que $z_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi Normale centré réduite.

2.6.3 Test de différences positives

Soit la variable N donnée par $N =$ *le nombre des différences qui sont positive de*$(x_t - x_{t-1})$.
avec une taille de la chronique $n \geqslant 12$, on peut montrer que la variable N est gaussienne de moyenne $E(N)$ et de variance $var(N)$ tel que $E(N) = n-1/2$ et $var(N) = n+1/12$.
On obtient la statistique du test comme suit

$$Z = \begin{cases} \frac{N-E(N)-1/2}{\sqrt{var(N)}} & si \quad N > E(N) \\ \frac{-N+E(N)+1/2}{\sqrt{var(N)}} & si \quad N < E(N) \end{cases}$$

On appliquant ce test par les étapes suivantes :

- la Vérification de la taille de la chronique $n \geqslant 12$
- Calculer N
- Tester l 'hypothèse nulle d'indépendance par calculer la valeur observée de la statistique tel que:

$$Z_{obs} = \begin{cases} \frac{N-\frac{n-1}{2}-1/2}{\sqrt{n+1/12}} & si \quad N > E(N) \\ \frac{-N+\frac{n-1}{2}+1/2}{\sqrt{n+1/12}} & si \quad N < E(N) \end{cases}$$

ou Z_{obs}suit la loi Normale centrée réduite.

- Au niveau de signification α on rejeter H_0 si $|Z_{obs}| > z_{1-\alpha/2}$ tel que $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi Normale centré réduite.

2.7 L'indépendance contre l'existence d'un effet cyclique

Dans une base de données chronologique, on peut déduire que les observations possèdent un caractère cyclique comme le cycle des saisons. Dans cette section, on va tester l'hypothèse d'indépendance à partir de la vérification de l'absence de cycle entre les variables successives de la série chronologique, pour ce type des effets, le test appliqué est non paramétrique de plus les hypothèses à tester sont :

H_0 : *les observations sont indépendante*
H_1 : *il y a un effet cyclique entre les observations*

2.7.1 Test de Mann-Whitney

Soit s saisons dans la série d'observation, on divise en deux parties les groupes des observations de chaque saison, ensuite on faire les séparations soient consistantes d'une saison à l'autre. Alors on obtient la variable suivante:

$$US = \sum_{t=1}^{s} U_t$$

U_t : représente la statistique de Mann-Whitney. Pour une taille d'échantillon $n \geqslant 40$, US suit la loi Normale de moyenne $E(US)$ et $var(US)$ tel que: $E(US) = \sum_t^s E(U_t)$ et $var(US) = \sum_t^s var(U_t)$,on obtient la statistique du test comme suit :
$Z = \frac{US - E(US)}{\sqrt{var(US)}}$ suit la loi Normale centrée réduite.
On appliquant ce test par les étapes suivantes:

- la vérification de la taille de l'échantillon $n \geqslant 40$.
- calculer US.
- Tester l'hypothèse nulle par calculer la variable statistique observée donnée par $Z_{obs} = \frac{US - \sum_t^s E(U_t)}{\sqrt{\sum_t^s var(U_t)}}$
- au niveau de signification α on rejeter l'hypothèse nulle si : $|Z_{obs}| > z_{1-\alpha/2}$ tel que $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi Normale centré réduite.

Chapter 3

Application avec R

3.1 Test de Khi deux d'indépendance

Exemple : 8

Un atelier de reprographie voudrait faire l'économie des frais d'entretien de ses photocopieurs .
En 2015 , sur 200 photocopieurs utilisés dans les mêmes conditions ,48 seulement ont été régulièrement entretenu. Fin 2015 , on a constaté que pendant l'année écoulée 28 photocopieurs ont du subir au moins une réparation au cours de l'année, dont 9 avaient été entretenus. Tester au seuil de 5% ,l'indépendance entre le bon fonctionnement des photocopieurs et le fait qu'ils subissent ou non un entretien
Le tableau des résultats est le suivantes :

	Au moins une panne	Pas de panne	Total
Entretien	9	39	48
Pas d'entretien	19	133	152
Total	28	172	200

Dans ce tableau de contingence on notera les effectifs observés O_i on déterminera les effectifs théorique T_i que l'on aurait si était en situation d'indépendance et s'il n'y avait pas de fluctuations d'échantillonnage,en gardant inchangés les effectifs marginaux. Puis on va calculer l'indicateur d'écart D les entre données observées et les données théoriques . l'indicateur d'écart D donné par : $D = \sum \frac{(O_i - T_i)^2}{T_i} \sim \chi^2$ à $(n-1)(p-1)$ degrés de liberté
n:le nombre de lignes
p:le nombre de colonnes On calcule la valeur de la variable de test :

	Au moins une panne	pas de panne	Total
Entretien	$\frac{28\times48}{200} = 6.72$	$\frac{172\times48}{200} = 41.28$	48
pas d'entretien	$\frac{28\times152}{200} = 21.28$	$152 - 21.28 = 130.72$	152
Total	28	172	200

$\chi^2 = \frac{(9-6.72)^2}{6.72} + \frac{(39-41.28)^2}{41.28} + \frac{(19-21.28)^2}{21.28} + \frac{(133-130.72)^2}{130.72}$ *alors on obtient* $\chi^2_c = 1.18$ *et pour seuil de risque* $\alpha = 0.05$ *à 1 degré de liberté* $\chi^2_{th} = 3.84$

Or $\chi^2_c < \chi^2_{th}$*. Donc au seuil de risque de* 5% *on accepte l'hypothèse null, on dit que les pannes constatées sont indépendantes de l'entretien des photocopieurs.*

Le test par R

```
# faire le test d'indépendance avec le calcul
#Declarer la matrice des données

donne<-matrix(c(9,39,19,133)

        ,byrow = TRUE,ncol = 2)
 donne
     [,1] [,2]
[1,]    9   39
[2,]   19  133
#Nommer les lignes rt les colonnes

colnames(donne)<-
    c("Au moins une panne","Pas de panne")
rownames(donne)<-
          c("Entretien","Pas d'entretien")
 donne
                 Au moins une panne Pas de panne
Entretien                         9           39
Pas d'entretien                  19          133
#Calculer les totaux des lignes et des colonnes

donne<-cbind(donne,c(0,0))
donne<-rbind(donne,c(0,0,0))
donne
                 Au moins une panne Pas de panne
```

```
Entretien                    9          39  0
Pas d'entretien             19         133  0
                            0           0   0
for(i in 1:nrow(donne))
   {donne[i,3]<-sum(donne[i,])}
for(i in 1:ncol(donne))
   {donne[3,i]<-sum(donne[,i])}
 donne
             Au moins une panne Pas de panne
Entretien                    9          39  48
Pas d'entretien             19         133 152
                            28         172 200
#Calculer les effectifs théoriques

 eff<-matrix(nrow =2,ncol=2)
 for(i in 1:2) for (j in 1:2)
{eff[i,j]<-donne[3,j]*donne[i,3]/donne[3,3]}
 eff
      [,1]   [,2]
[1,]  6.72  41.28
[2,] 21.28 130.72
 colnames(eff)<-
c("Au moins une panne","Pas de panne")

 rownames(eff)<-
 c("Entretien","Pas d'entretien")
 eff
            Au moins une panne Pas de panne
```

```
Entretien                    6.72         41.28
Pas d'entretien              21.28        130.72
 eff<-cbind(eff,c(0,0))
 eff<-rbind(eff,c(0,0,0))
 eff
             Au moins une panne Pas de panne
Entretien                    6.72         41.28  0
Pas d'entretien              21.28        130.72 0
                             0.00          0.00  0
 for(i in 1:nrow(eff))
    {eff[i,3]<-sum(eff[i,])}
 for(i in 1:ncol(eff))
    {eff[3,i]<-sum(eff[,i])}

 eff
              Au moins une panne Pas de panne
Entretien                  6.72         41.28  48
Pas d'entretien            21.28        130.72 152
                           28.00     172.00    200
#Calculer la valeur de khi deux
chideux<-matrix(nrow =2,ncol=2)
 for(i in 1:2) for(j in 1:2)
{chideux[i,j]<-(donne[i,j]-eff[i,j])^2/eff[i,j]}
 chideux

 0.7735714 0.12593023
 0.2442857 0.03976744
 for (i in 1:2)
```

```
 {valeurchideux<-sum(chideux[,])}

 valeurchideux
[1] 1.183555
#Calculer la valeur de degré de liberté

 ddl<-(ncol(chideux)-1)*(nrow(chideux)-1)
 ddl
[1] 1
#Tester l'independance avec la fonction

chisq.test(donne)
Pearson's Chi-squared test
  with Yates' continuity correction
data:  donne
X-squared = 0.72137, df = 1, p-value = 0.3957
```

Exemple : 9

Pour étudier les risque du tabagisme, nous avons examiné des personnes atteintes de différents types des cancer et étudié la relation entre le tabagisme et le cancer.

test	poumon	estomac	colon	nez	gorge	bouche
fumeur	122	150	153	132	200	122
nonfumeur	132	144	120	167	321	165

Table 3.1: Risque du tabagisme

Le test par R

```
 TAB<-
matrix(c(122,150,153,132,200,122,132,144,120
     ,167,321,165),byrow=TRUE,ncol=6)

 TAB

  122  150  153  132  200  122
  132  144  120  167  321  165
 #Nommer les lignes et les colonnes

 colnames(TAB)<-c("poumon","estomac","colon",
           "nez","gorge","bouche")

rownames(TAB)<-c("fumeur","nfumeur")
#calculer les totaux des lignes et des colonnes
 addmargins(TAB)
      poumon estomac colon nez gorge bouche  Sum
fumeur  122     150   153   132   200    122  879
nfumeur 132     144   120   167   321    165 1049
Sum     254     294   273   299   521    287 1928
```

```
 TAB<-cbind(TAB,rep(0,2))
 TAB<-rbind(TAB,rep(0,7))
 TAB
      poumon estomac colon nez gorge bouche
fumeur  122     150   153   132   200    122 0
nfumeur 132     144   120   167   321    165 0
          0       0     0     0     0      0 0

 for(i in 1:nrow(TAB)){TAB[i,7]<- sum(TAB[i,])}
 for(i in 1:ncol(TAB)){TAB[3,i]<- sum(TAB[,i])}
TAB
        poumon estomac colon nez gorge bouche
fumeur  122     150   153   132   200    122  879
nfumeur 132     144   120   167   321    165 1049
        254     294   273   299   521    287 1928

 #calculer les frequences theorique
khideux<-matrix(nrow = 2,ncol = 6)
 for(i in 1:2) for(j in 1:6)
 {khideux[i,j]<-(TAB[i,j]-frec[i,j])^2/frec[i,j]}
 khideux

0.3317464 1.900749 6.542372 0.1367733 5.929956 0.5
0.2779839 1.592715 5.482121 0.1146080 4.968953 0.5

 valeurkhideux<-sum(khideux[,])
 valeurkhideux
[1] 28.37739
```

```
#calculer la valeur de degré de liberté

 ddl<-(nrow(frec)-1)*(ncol(frec)-1)
ddl
[1] 5

 #tester avec la fonction chisq

 TAB<-tab<-matrix(c(122,150,153,132,200,122,132,
     144,120,167,321,165),byrow=TRUE,ncol=6)
colnames(TAB)<-
 c("pomon","estomac","colon","nez","gorge","bouche

rownames(TAB)<-c("fumeur","nfumeur")
 TAB
         pomon estomac colon nez gorge bouche
fumeur     122     150   153  132   200    122
nfumeur  132     144   120  167   321    165

 #faire le test avec la fonction chisq
 chisq.test(TAB)

Pearson's Chi-squared test

data:  TAB
X-squared = 28.377, df = 5, p-value = 3.071e-05
```

Au seuil de risque de 5%*trouvons le quantile d'ordre 0.975 de*

loi de Khi deux avec $\nu = 5$ ddl $\chi^2_5 = 11.07$, donc $\chi_c > \chi_{th}$ alors rejette H_0, On dit qu'il y a une relation entre le tabagisme et le cancer.

3.2 Test de corrélation nulle

Exemple : 10

on va proposer un tableau des variables quantitative et donner les calcules de la coefficient de corrélation puis appliquer le test de Student pour vérifier si la corrélation entre les variables est nulle (indépendante) ou non.
à partir de table de Student à 1 degré de liberté on a $(n = 3, ddl = n - 2)$ $t_1(0,975) = 12.71)$ *Pour tester l'indépendance des variables quantitative on proposer le tableau suivante:*

x	y
-1	-1
0	-1
1	1

Lorsque la corrélation est nulle on accepte H_0 et dit que les variables x et y sont indépendantes sinon on rejette H_0 et dit que les variables x et y sont corrélés (dépendantes)
L'étude est basé sur calculer la valeur $t = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$ pour faire une décision.

Le test par R

```
 # inportation des données
library(readxl)

Donne<-
read_excel("C:/Users/KR/Desktop/quantitative.xlsx"

 Donne
     [,1] [,2]
[1,]   -1   -1
[2,]    0   -1
[3,]    1    1
#calculer la moyenne des colonnes
 X<-sum(Donne[,1])/nrow(Donne)
 X
[1] 0
 Y<-sum(Donne[,2])/nrow(Donne)
 Y
[1] -0.3333333
#clculer la covariance

 cova<- 1/3 *sum((Donne[,1]-X)*(Donne[,2]-Y))
 cova
[1] 0.6666667
 sigmax<-sqrt(1/3*sum((Donne[,1]-X)^2))
 sigmax
[1] 0.8164966
```

```
#calculer l'ecartype
 sigmay<-sqrt(1/3*sum((Donne[,2]-Y)^2))
sigmay
[1] 0.942809
#calculer la coefficient de corrélation
 r<-cova/sigmax *sigmay
 r
[1] 0.8660254
#calculer l'indicateur t
 t=r/(sqrt(1-r^2))
 t
[1] 1.732051
 #tester avec la fonction de test de Student

t.test(Donne)
One Sample t-test data:  Donne
t = -0.41523, df = 5, p-value = 0.6952
alternative hypothesis: true mean

       is not equal to 0
95 percent confidence interval:
 -1.1984635  0.8651301
sample estimates:
mean of x
-0.1666667

plot(Donne)# tracer le nuage du point
```

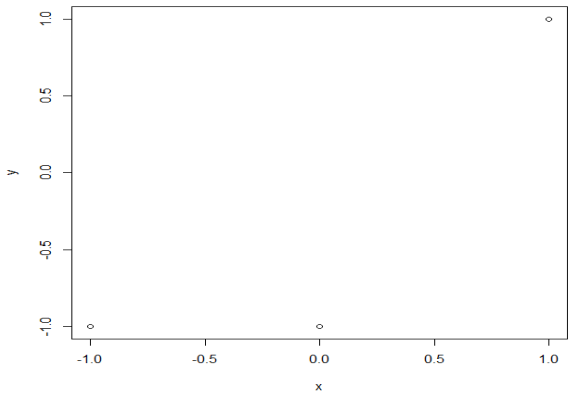


Figure 3.1: nuage du point de corrélation positive

Conclusion

La vérification d'indépendance dépendue cela nécessite une bonne identification d'hypothèse alternative H_1 d'indépendance, c'est pour ça, il est essentiel identifier le type de dépendance dans l'échantillon des données ; de plus, le test que l'on appliquera doit être efficace. D'une part, si les variables sont qualitatives, on utilise le test Khi-deux d'indépendance par contre si on a des variables quantitatives, on passe au test statistique de Student.
D'autre part, on a trois effets pour la dépendance la première effet est la persistance le test le plus indiqué est test Anderson où les variables d'échantillon sont gaussiennes sinon on applique le test de Wald-Wolf est le plus connu dans les tests non-paramétriques avec la vérification de la taille d'échantillon, le deuxième effet est de tendance où les tests non-paramétriques les plus utilisables sont le test de Spearman et le test de Kendall le dernier effet est de cyclique dans cette situation, on a un seul test qui le candidat est le test de Mann-Whitney.

Bibliography

[1] Cleophas.O, Taha.B.M.G.Oranda,Bernard.B, 1997, Revue Bibliographique des test d'homogénéité et d'indépendance.

[2] Louis Houde, PAF-1010 Analyse quantitative de problème de gestion, Département de Mathématique et Informatique, Université de Québec à Trois-Rivières.

[3] Pierre DUSART,2018 ,Licence 2-S4 SI-MASS Cours de Statistique inférentielle.

[4] Pierre DUSART, 2013, Licence 2-S3 SI-MASS, Cour de probabilité.

[5] http://www.jybaudot.fr/Inferentielle/efficacite.html.

[6] U.F.R SPSE Master 1, PMP STA 21 Méthodes statistiques pour l'analyse des données en psychologie.

[7] Jean-Jacques Ruch, 2012-2013, Statistique, Estimation,Préparation à l'Agrégation Bordeaux 1.

[8] FIFO 3, PROBABILITÉS-STATISTIQUES, LES TESTS D'HYPOTHÈSE.

Annexe

La vérification du type d'effet

Pour vérifier l'existence d'un effet de persistance ou cyclique dans l'échantillon, Bouvier suggéré une mesure pour la détection des effets à partir d'un corrélogramme; on définie le coefficient d'auto-corrélation r_θ d'ordre θ, conçu comme suit:
$r_\theta = \frac{\sum_{t=1}^{n} x_t - x_{t+\therefore} - 1/2(sum_{t=1}^{n} x_t)^2}{\sum_{t=1}^{n} x_t^2 - 1/2(sum_{t=1}^{n} x_t)^2}$ tel que pour toute valeur $k + \theta > n$ on a $x_{t+\theta} = x_{t+\theta-n}$

Définition 10
le corrélogramme est une représentation de suit des r_θ en fonction des θ à l'aide de ce représentation on va faire une interprétation et détecter la présence de persistance ou de cyclicité entre l'échantillon.

Si le corrélogramme présente le paterne d'une fonction périodique de période T alors on dit que le type d'effet est cyclique.
Si le corrélogramme présente le paterne d'une fonction décroissante,on peut dit aussi les coefficients d'auto-corrélation décroissante en fonction des θ (leurs ordre), alors on a un effet persistance.

La corrélation

Pour estimer la relation entre deux variables (ou plus), il existe des coefficients de corrélations pour traiter des différent types des donnée tel que le coefficient de corrélation est compris dans l'intervalle -1 à $+1$, si la valeur est égale à -1 on dit qu'il existe une parfaite corrélation négative ensuite si la valeur est égale à $+1$ il existe une parfaite corrélation positive, et lorsque la valeur vaut 0 on a une absence de corrélation, c'est-à-dire les variables sont indépendantes.

Coefficient de corrélation

Le coefficient de corrélation théorique ρ donné par: $\rho = \frac{cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$
ou $cov(X, Y) = E[(X - \mu x)(Y - \mu x)]$ (la covariance théorique).
Le Coefficient de corrélation échantillonnal r donné par:
$r = \frac{cov(\hat{X},Y)}{S_X S_Y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\frac{1}{n-1}(Y_i - \bar{Y})^2}}$.