

Cours L2 AES S4 : Loi normale, Intervalles de confiance, Tests

Simon ROBY

14 septembre 2022

TABLE DES MATIÈRES

I	Loi normale, Lois associées	2
I.1	Loi normale	2
I.1.1	Définitions	3
I.1.2	Calculs de Probabilités	4
I.2	Loi associées à la loi normale	9
I.2.1	Loi du Khi-deux	9
I.2.2	Loi de Student	11
I.2.3	Loi de Fisher-Snedecor	14
II	Intervalles de confiances	17
II.1	Préliminaires : Convergence en loi et Théorème central limite	17
II.2	Échantillonnage	18
II.3	Estimation par intervalle de confiance	19
II.3.1	Estimateurs	19
II.3.2	Calculs d'intervalle de confiance	21
III	Tests d'hypothèse	23
III.1	Région critique d'un test	24
III.2	Risques d'un test	25
III.3	Prendre une décision	26
III.4	Tests du Khi-deux	26
III.4.1	Test d'adéquation du Khi-deux	26
III.4.2	Test d'indépendance	27

I

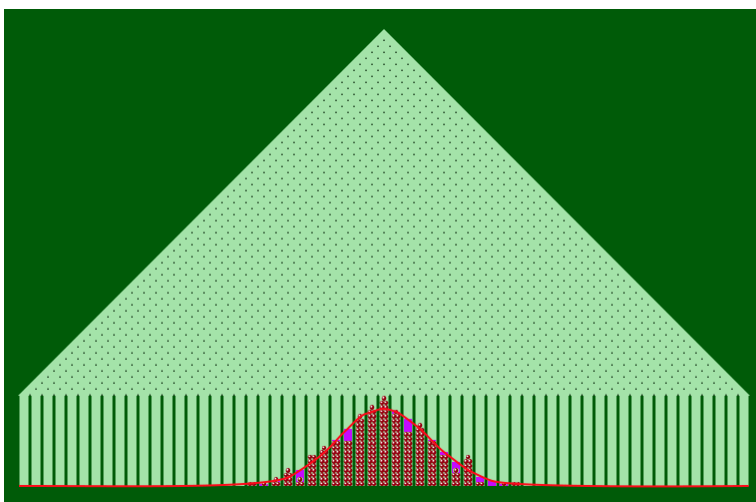
LOI NORMALE, LOIS ASSOCIÉES

I.1 LOI NORMALE

La loi normale (ou loi de Laplace-Gauss) est une loi de probabilité continue définie sur \mathbb{R} , l'ensemble des réels. C'est sans aucun doute la loi de probabilité la plus utilisée et la plus commune dans le réel. C'est, en effet, la loi la plus adaptée à modéliser les phénomènes naturels et physiques issue de plusieurs événements aléatoires.

Exemples :

1. La taille des humains, hommes ou femmes
2. La taille des tiges des pâquerettes dans un champ
3. Mensurations d'une manière générale
4. Notes à une épreuve
5. La planche à clous de Galton (Galton board or bean machine) montre bien ce qui se passe : les billes subissent les chocs sur les chicanes et se retrouvent dans les cases du bas. À la fin, les billes remplissent les cases selon une distribution en cloche. Il est possible de programmer cette expérience. Un amusement intéressant, surtout si on le poursuit en percolant les billes sur les billes déjà en position.



http://sorciersdesalem.math.cnrs.fr/Vulgarisation/Galton/galton_plus.html

I.1.1 DÉFINITIONS

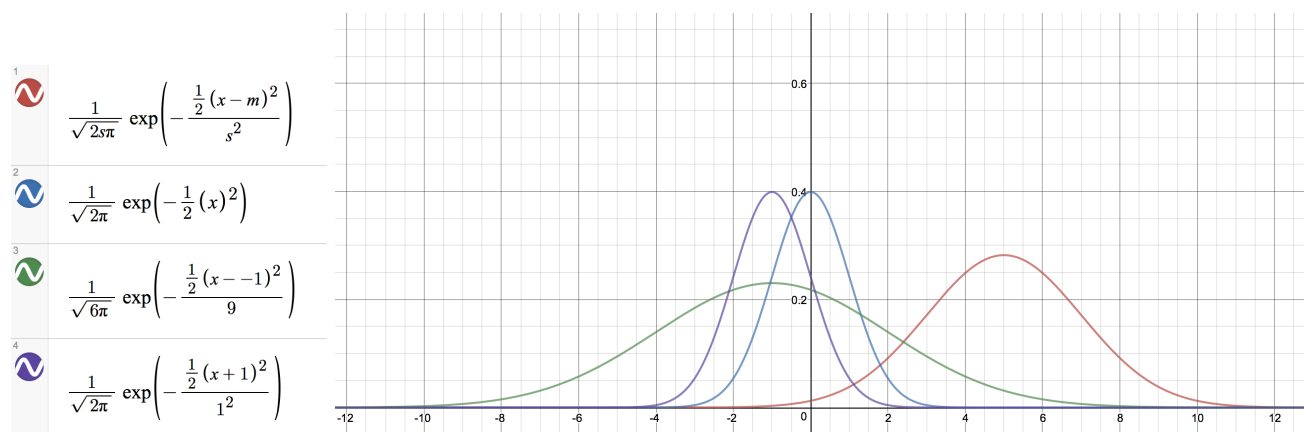
Définition I.1 (Fonctions de densité et de répartition de la loi normale) :

Une variable aléatoire réelle X suit une loi normale d'espérance $\mu \in \mathbb{R}$ et de variance $\sigma^2 \in \mathbb{R}^+$, notée $\mathcal{N}(\mu, \sigma^2)$ (On l'écrit $X \sim \mathcal{N}(\mu, \sigma^2)$, si sa fonction de densité est définie, pour tout $x \in \mathbb{R}$, par :

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

De manière équivalente, sa fonction de répartition est donnée pour tout $x \in \mathbb{R}$ par :

$$F_X(x) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2\right) dz$$



Remarque :

On rappelle que :

1. Pour toute loi de probabilité continue, la fonction de répartition F découle de la fonction de densité f par la relation :

$$F(x) = \int_{-\infty}^x f(z) dz$$

2. La fonction de densité d'une loi de probabilité est toujours de masse 1 : en d'autre terme

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

Cela implique que la fonction de répartition associée vérifie $\lim_{x \rightarrow +\infty} F(x) = 1$

3. La fonction de répartition et/ou la fonction de densité caractérisent la loi de la variable aléatoire. Par exemple si on trouve une variable aléatoire X absolument continue de densité

$$\frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-1}{2}\right)^2\right),$$

alors X suit une loi normale $\mathcal{N}(1, 4)$.

Voyons maintenant comment calculer les probabilités pour des variables aléatoires suivant cette loi.

I.1.2 CALCULS DE PROBABILITÉS

Soit X une variable aléatoire suivant une loi normale $\mathcal{N}(\mu, \sigma^2)$. Si on cherche à calculer la probabilité que $X \leq -0,5$, par exemple, on va se confronter au problème du calcul de la fonction de répartition :

$$P(X \leq -0,5) = F_X(-0,5) = \int_{-\infty}^{-0,5} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2\right) dz$$

Rappel :

Ici la valeur $P(X \leq -0,5)$ représente graphiquement le point sur la courbe de F qui correspond à 0.5. C'est aussi l'aire sous la courbe de f entre $-\infty$ et -0.5 . Le graphe suivant illustre ce calcul pour une loi normale $\mathcal{N}(-1, 3)$.



Un ordinateur fait très bien ce genre de calcul. On peut se donner un peu d'intuition sur ce que l'on fait grâce à l' **animation (créée pour vous)**. Mais en pratique, calculer à chaque fois cette expression serait long et fastidieux. On utilise donc la méthode de réduction de la loi normale.

Proposition I.1

$$X \sim \mathcal{N}(\mu, \sigma^2) \iff \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

On peut alors calculer la probabilité précédente pour toute loi normale, en connaissant seulement la loi $\mathcal{N}(0, 1)$, appelée loi normale centrée réduite.

LA LOI NORMALE CENTRÉE RÉDUITE

Définition I.2 (Fonctions de densité et de répartition de la loi normale centrée réduite) :

Une variable aléatoire réelle X suit une loi normale centrée réduite, notée $\mathcal{N}(0, 1)$, si sa fonction de densité est définie, pour tout $x \in \mathbb{R}$, par :

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

De manière équivalente, sa fonction de répartition est donnée pour tout $x \in \mathbb{R}$ par :

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz$$

Grâce à ces formules on peut établir la table de la loi normale centrée réduite suivante :

z	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	0.7290	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

FIGURE I.1 – Table de la loi normale centrée réduite

De plus, comme la fonction de densité est symétrique par rapport à l'axe des ordonnées, la fonction de répartition est symétrique par rapport au point $(0;0,5)$:

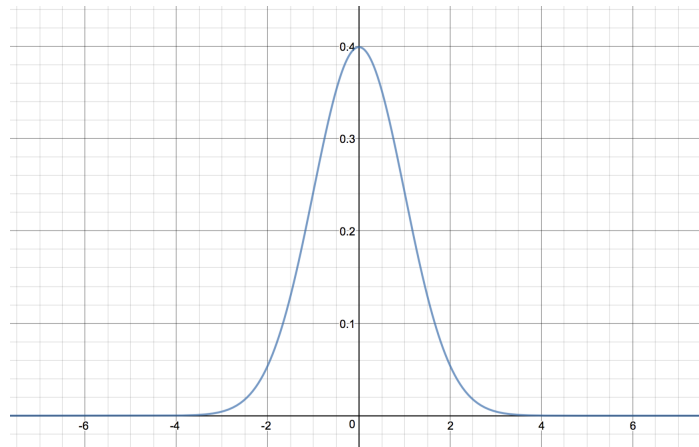


FIGURE I.2 – Fonction de densité de la loi normale centrée réduite

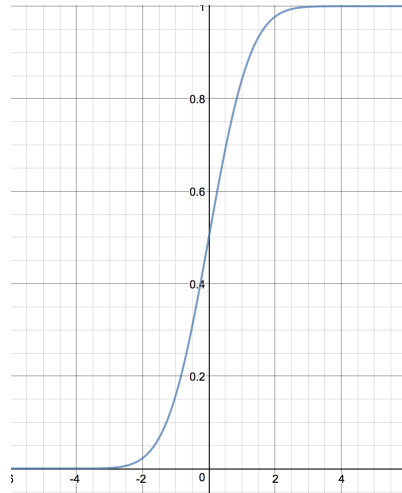


FIGURE I.3 – Fonction de répartition de la loi normale centrée réduite

et on obtient donc la propriété suivante de symétrie de la fonction de répartition :

Proposition I.2

On a les égalités suivantes :

$$\Phi(0) = 0,5 \tag{I.1}$$

$$\Phi(-x) = 1 - \Phi(x) \quad \forall x \in \mathbb{R} \tag{I.2}$$

UN EXEMPLE DE CALCULS DE PROBABILITÉ

Prenons par exemple $X \sim \mathcal{N}(0,6, 4)$ et déterminons les probabilités $P(X \geq 1,86)$ et $P(X \leq -0,22)$:

$$\begin{aligned} P(X \geq 1,86) &= 1 - P(X \leq 1,86) \text{ , par définition de la probabilité } (P(X \in \mathbb{R}) = 1) \\ &= 1 - P\left(\frac{X - 0,6}{\sqrt{4}} \leq \frac{1,86 - 0,6}{\sqrt{4}}\right) \\ &= 1 - P\left(\frac{X - 0,6}{\sqrt{4}} \leq 0,63\right) \\ &= 1 - \Phi(0,63) \text{ , car } X \text{ suit une loi normale centrée réduite d'après la proposition I.1} \\ &= 1 - 0,7357 = 0,2643 \end{aligned}$$

Un peu d'explication sur cette dernière ligne : Pour trouver $\Phi(0,63)$ dans la table de la loi normale centrée réduite, on choisit les chiffres des unités et des dixièmes dans la colonne de gauche (0,6) et le chiffre des centièmes dans la ligne du haut (0,03). L'intersection des ligne et colonne correspondantes donne la valeur recherchée : 0,7357

$$\begin{aligned} P(X \leq -0,22) &= P\left(\frac{X - 0,6}{\sqrt{4}} \leq \frac{-0,22 - 0,6}{\sqrt{4}}\right) \\ &= P\left(\frac{X - 0,6}{\sqrt{4}} \leq -0,41\right) \\ &= \Phi(-0,41) \text{ , car } X \text{ suit une loi normale centrée réduite d'après la proposition I.1} \\ &= 1 - \Phi(0,41) \text{ , d'après l'équation I.14} \\ &= 1 - 0,659097 = 0,340903 \end{aligned}$$

RECHERCHE DE QUANTILES

Dans cette partie on cherche à faire le chemin inverse : On veut inverser la fonction de répartition pour pouvoir trouver à quelle nombre vérifie une probabilité donnée $\alpha \in [0, 1]$. Cela est possible puisque la fonction de répartition F_X est croissante strictement. Cela s'appelle la *recherche de quantile*.

Définition I.3 :

Pour une variable aléatoire réelle X donnée, le quantile $F_X^{-1}(\alpha)$ d'ordre $\alpha \in [0, 1]$, noté Q_α , est le nombre tel que $P(X \leq Q_\alpha) = \alpha$.

Proposition I.3

Si X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$, alors le quantile Q_α de X s'exprime en fonction du quantile de la loi normale centrée réduite $\Phi^{-1}(\alpha)$, pour tout $\alpha \in [0, 1]$:

$$Q_\alpha = \mu + \sigma\Phi^{-1}(\alpha) \tag{I.3}$$

Exemple I.1 :

Supposons que $X \sim \mathcal{N}(1,9)$. Cherchons par exemple le quantile d'ordre $\alpha = 90\%$.

D'après la proposition précédente, il suffit de calculer le quantile d'ordre α pour la loi normale centrée réduite. On cherche alors dans la table de loi les valeurs les plus proches, qui encadrent 0,9. C'est 0,8997 et 0,9015. On regarde alors les deux nombres qui correspondent respectivement 1,28 et 1,29. Donc $\phi^{-1}(0,9)$ est compris entre ces deux valeurs.

z	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	0.7290	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

Si on veut trouver une valeur approchée, on peut faire une interpolation linéaire des valeurs.

D'après la formule :

$$\begin{aligned} \mu + \sigma \times 1,28 &\leq Q_\alpha \leq \mu + \sigma \times 1,29 \\ 1 + \sqrt{9} \times 1,28 &\leq Q_\alpha \leq 1 + \sqrt{9} \times 1,29 \\ 4,84 &\leq Q_\alpha \leq 4,87 \end{aligned}$$

Dans la table de loi, les probabilités présentes vont de 0,5 à 1. Pour pouvoir trouver les quantiles d'ordre compris entre 0 et 0,5 on a besoin de l'analogie de la proposition I.2, pour Φ^{-1} :

Proposition I.4

On a les égalités suivantes :

$$\Phi^{-1}(0,5) = 0 \quad (\text{I.4})$$

$$\Phi^{-1}(1 - \alpha) = -\Phi^{-1}(\alpha) \quad \forall \alpha \in [0, 1] \quad (\text{I.5})$$

Grâce à la seconde formule, on peut retrouver tous les quantiles entre 0 et 0,5. Par exemple, $\Phi(0,1) = -\Phi(0,9)$. Donc, il suffit de trouver le quantile d'ordre 0,9 pour trouver un quantile d'ordre 0,1. Cette formule est donc indispensable à connaître pour pouvoir faire n'importe quel calcul de recherche de quantile.

PROPRIÉTÉS DE LA LOI NORMALE

La loi normale, comme on l'a dit, se retrouve beaucoup dans la nature. Mais ce n'est pas dû au hasard. Elle vérifie beaucoup de propriétés intéressantes qui la conserve :

Proposition I.5 (Linéarité et sommes de lois normales indépendantes)

Soient X une variable aléatoire réelle de loi normale $\mathcal{N}(\mu, \sigma^2)$ et deux constantes $a, b \in \mathbb{R}$, alors la variable aléatoire $aX + b$ suit une loi normale $\mathcal{N}(a + b\mu)$

De plus, si Y_1, Y_2, \dots, Y_n sont des variables aléatoires réelles indépendantes de lois normales respectives $\mathcal{N}(\mu_i, \sigma_i^2)$, pour i allant de 1 à n , alors :

$$\sum_{i=1}^n Y_i \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$$

On verra aussi par la suite que beaucoup de loi peuvent être ramenées à la loi normale. C'est ce qui fait son importance. Le Théorème de la Limite entrée II.1 est un résultat primordial qui ramène presque tout échantillon de variables aléatoires indépendantes et identiquement distribuées à une loi normale.

I.2 LOI ASSOCIÉES À LA LOI NORMALE

Nous allons à présent donner des lois qui sont utiles en théorie des tests. Toutes ces lois sont issues de la loi normale centrée réduite.

I.2.1 LOI DU KHI-DEUX

Nous commençons par la loi dite du Khi-deux, notée $\chi^2(k)$ (ceci n'est pas un x mais bien la lettre grecque Khi), car elle dépend d'un paramètre $k \in \mathbb{N}^*$. l'entier naturel k désigne le nombre de degrés de libertés de la loi.

Définition I.4 (Loi du Khi-deux) :

Si Y_1, \dots, Y_k sont des variables aléatoires indépendantes qui suivent chacune une loi normale $\mathcal{N}(0, 1)$, alors la variable aléatoire $X = \sum_{i=1}^k Y_i^2$ suit une loi du Khi-deux à k degrés de libertés :

$$X \sim \chi^2(k)$$

Remarque :

Si on regarde de plus près la définition on peut remarquer que si X suit une loi normale centrée réduite alors X^2 suit une loi du Khi-deux à 1 degré de liberté.

De plus, une variable aléatoire qui suit une loi du Khi-deux est réalisée par la somme de carrés de variables aléatoires, elle ne peut jamais être négative. C'est pour cela qu'elle est définie sur \mathbb{R}^+ .

Enfin, on peut remarquer que la somme de variables aléatoire suivant une loi du Khi-deux donne une variable aléatoire suivant une loi du Khi-deux. Son degré est la somme des degrés des autres lois.

Proposition I.6 (Fonction de densité et fonction de répartition d'une loi du Khi-deux)

La fonction de densité f_X d'une variable aléatoire X qui suit une loi du Khi-deux est définie, pour tout $x \in \mathbb{R}^+$, par :

$$f_X(x) = \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} \exp\left(-\frac{x}{2}\right) \quad (\text{I.6})$$

où Γ désigne la fonction gamma définie pour tout $y \in \mathbb{R}$ par :

$$\Gamma(y) = \int_0^{+\infty} t^{y-1} \exp(-t) dt \quad (\text{I.7})$$

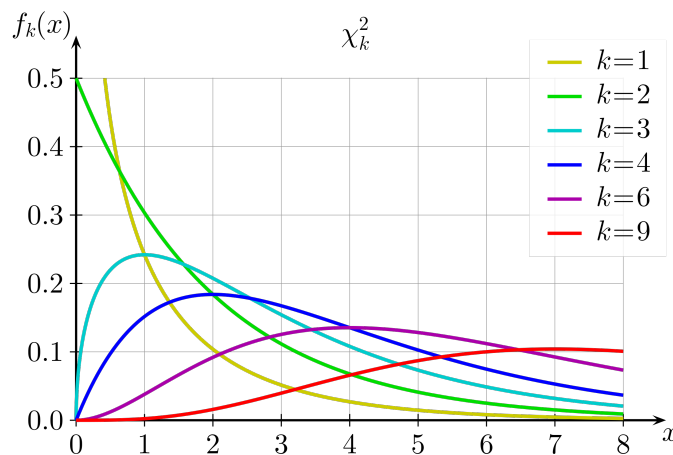


FIGURE I.4 – Fonction de répartition du Khi-deux

De manière équivalente, sa fonction de répartition F_X est donnée par :

$$F_X(x) = \frac{\gamma(\frac{k}{2}, \frac{x}{2})}{\Gamma(\frac{k}{2})} \quad (\text{I.8})$$

où γ désigne la fonction gamma incomplète définie pour tout $y \in \mathbb{R}$ et a un paramètre strictement positif, par :

$$\gamma(a, x) = \int_0^x t^{a-1} \exp(-t) dt \quad (\text{I.9})$$

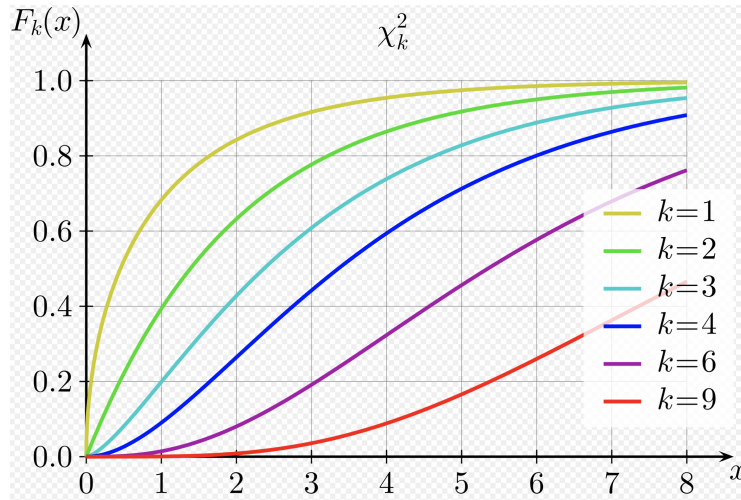


FIGURE I.5 – Fonction de densité du Khi-deux

Remarque :

Les fonctions gamma apparaissent souvent en mathématiques. En effet, elle font parties des *fonctions spéciales* qui sont les fonctions apparues au XIXème siècle dans la résolution d'équations de la physique mathématique, en l'occurrence dans des équations aux dérivées partielles. La fonction Gamma est préprogrammée dans tout logiciel de calcul statistique.

Comme pour la loi normale centrée réduite, pour faire les calculs de probabilité, on utilise une table de loi :

p	0,999	0,995	0,99	0,98	0,95	0,9	0,8	0,2	0,1	0,05	0,02	0,01	0,005	0,001
ddl														
1	0,0000	0,0000	0,0002	0,0006	0,0039	0,0158	0,0642	1,6424	2,7055	3,8415	5,4119	6,6349	7,8794	10,8276
2	0,0020	0,0100	0,0201	0,0404	0,1026	0,2107	0,4463	3,2189	4,6052	5,9915	7,8240	9,2103	10,5966	13,8155
3	0,0243	0,0717	0,1148	0,1848	0,3518	0,5844	1,0052	4,6416	6,2514	7,8147	9,8374	11,3449	12,8382	16,2662
4	0,0908	0,2070	0,2971	0,4294	0,7107	1,0636	1,6488	5,9886	7,7794	9,4877	11,6678	13,2767	14,8603	18,4668
5	0,2102	0,4117	0,5543	0,7519	1,1455	1,6103	2,3425	7,2893	9,2364	11,0705	13,3882	15,0863	16,7496	20,5150
6	0,3811	0,6757	0,8721	1,1344	1,6354	2,2041	3,0701	8,5581	10,6446	12,5916	15,0332	16,8119	18,5476	22,4577
7	0,5985	0,9893	1,2390	1,5643	2,1673	2,8331	3,8223	9,8032	12,0170	14,0671	16,6224	18,4753	20,2777	24,3219
8	0,8571	1,3444	1,6465	2,0325	2,7326	3,4895	4,5936	11,0301	13,3616	15,5073	18,1682	20,0902	21,9550	26,1245
9	1,1519	1,7349	2,0879	2,5324	3,3251	4,1682	5,3801	12,2421	14,6837	16,9190	19,6790	21,6660	23,5894	27,8772
10	1,4787	2,1559	2,5582	3,0591	3,9403	4,8652	6,1791	13,4420	15,9872	18,3070	21,1608	23,2093	25,1882	29,5883
11	1,8339	2,6032	3,0535	3,6087	4,5748	5,5778	6,9887	14,6314	17,2750	19,6751	22,6179	24,7250	26,7568	31,2641
12	2,2142	3,0738	3,5706	4,1783	5,2260	6,3038	7,8073	15,8120	18,5493	21,0261	24,0540	26,2170	28,2995	32,9095
13	2,6172	3,5650	4,1069	4,7654	5,8919	7,0415	8,6339	16,9848	19,8119	22,3620	25,4715	27,6882	29,8195	34,5282
14	3,0407	4,0747	4,6604	5,3682	6,5706	7,7895	9,4673	18,1508	21,0641	23,6848	26,8728	29,1412	31,3193	36,1233
15	3,4827	4,6009	5,2293	5,9849	7,2609	8,5468	10,3070	19,3107	22,3071	24,9958	28,2529	30,5779	32,8013	37,6973
16	3,9416	5,1422	5,8122	6,6142	7,9616	9,3122	11,1521	20,4651	23,5418	26,2962	29,6332	31,9999	34,2672	39,2524
17	4,4161	5,6972	6,4078	7,2550	8,6718	10,0852	12,0023	21,6146	24,7690	27,5871	30,9950	33,4087	35,7185	40,7902
18	4,9048	6,2648	7,0149	7,9062	9,3905	10,8649	12,8570	22,7595	25,9894	28,8693	32,3462	34,8053	37,1565	42,3124
19	5,4068	6,8440	7,6327	8,5670	10,1170	11,6509	13,7158	23,9004	27,2036	30,1435	33,6874	36,1909	38,5823	43,8202
20	5,9210	7,4338	8,2604	9,2367	10,8508	12,4426	14,5784	25,0375	28,4120	31,4104	35,0196	37,5662	39,9968	45,3147
21	6,4467	8,0337	8,8972	9,9146	11,5913	13,2396	15,4446	26,1711	29,6151	32,6706	36,3434	38,9322	41,4011	46,7970
22	6,9830	8,6427	9,5425	10,6000	12,3380	14,0415	16,3140	27,3105	30,8133	33,9244	37,6595	40,2894	42,7957	48,2679
23	7,5292	9,2604	10,1957	11,2926	13,0905	14,8480	17,1865	28,4288	32,0069	35,1725	38,9683	41,6384	44,1813	49,7282
24	8,0849	9,8862	10,8564	11,9918	13,8484	15,6587	18,0618	29,5533	33,1962	36,4150	40,2704	42,9798	45,5585	51,1786
25	8,6493	10,5197	11,5240	12,6973	14,6114	16,4734	18,9398	30,6752	34,3816	37,6525	41,5661	44,3141	46,9279	52,6197
26	9,2221	11,1602	12,1981	13,4086	15,3792	17,2919	19,8202	31,7946	35,5632	38,8851	42,8558	45,6417	48,2899	54,0520
27	9,8028	11,8076	12,8785	14,1254	16,1514	18,1139	20,7030	32,9117	36,7412	40,1133	44,1400	46,9629	49,6449	55,4760
28	10,3909	12,4613	13,5647	14,8475	16,9279	18,9392	21,5880	34,0266	37,9159	41,3371	45,4188	48,2782	50,9934	56,8923
29	10,9861	13,1211	14,2565	15,5745	17,7084	19,7677	22,4751	35,1394	39,0875	42,5570	46,6927	49,5879	52,3356	58,3012
30	11,5880	13,7867	14,9535	16,3062	18,4927	20,5992	23,3641	36,2502	40,2560	43,7730	47,9618	50,8922	53,6720	59,7031
40	17,9164	20,7065	22,1643	23,8376	26,5093	29,0505	32,3450	47,2685	51,8051	55,7585	60,4361	63,6907	66,7660	73,4020
50	24,6739	27,9907	29,7067	31,6639	34,7643	37,6886	41,4492	58,1638	63,1671	67,5048	72,6133	76,1539	79,4900	86,6608
60	31,7383	35,5345	37,4849	39,6994	43,1880	46,4589	50,6406	68,9721	74,3970	79,0819	84,5799	88,3794	91,9517	99,6072
70	39,0364	43,2752	45,4417	47,8934	51,7393	55,3289	59,8978	79,7146	85,5270	90,5312	96,3875	100,4252	104,2149	112,3169
80	46,5199	51,1719	53,5401	56,2128	60,3915	64,2778	69,2069	90,4053	96,5782	101,8795	108,0693	112,3288	116,3211	124,8392
90	54,1552	59,1963	61,7541	64,6347	69,1260	73,2911	78,5584	101,0537	107,5650	113,1453	119,6485	124,1163	128,2989	137,2084
100	61,9179	67,3276	70,0649	73,1422	77,9295	82,3581	87,9453	111,6667	118,4980	124,3421	131,1417	135,8067	140,1695	149,4493
120	77,7551	83,8516	86,9233	90,3667	95,7046	100,6236	106,8056	132,8063	140,2326	146,5674	153,9182	158,9502	163,6482	173,6174
140	93,9256	100,6548	104,0344	107,8149	113,6593	119,0293	125,7581	153,8537	161,8270	168,6130	176,4709	181,8403	186,8468	197,4508
160	110,3603	117,6793	121,3456	125,4400	131,7561	137,5457	144,7834	174,8283	183,3106	190,5165	198,8464	204,5301	209,8239	221,0190
180	127,0111	134,8844	138,8204	143,2096	149,9688	156,1526	163,8682	195,7434	204,7037	212,3039	221,0772	227,0561	232,6198	244,3705
200	143,8428	152,2410	156,4320	161,1003	168,2786	174,8353	183,0028	216,6088	226,0210	233,9943	243,1869	249,4451	255,2642	267,5405
250	186,5541	196,1606	200,9386	206,2490	214,3916	221,8059	231,0128	268,5986	279,0504	287,8815	298,0388	304,9396	311,3462	324,8324
300	229,9634	240,6634	245,9725	251,8637	260,8781	269,0679	279,2143	320,3971	331,7885	341,3951	352,4246	359,9064	366,8444	381,4252
400	318,2596	330,9028	337,1553	344,0781	354,6410	364,2074	376,0218	423,5895	436,6490	447,6325	460,2108	468,7245	476,6064	493,1318
500	407,9470	422,3034	429,3875	437,2194	449,1468	459,9261	473,2099	526,4014	540,9303	553,1268	567,0698	576,4928	585,2066	603,4460
600	498,6229	514,5289	522,3651	531,0191	544,1801	556,0560	570,6680	628,9433	644,8004	658,0936	673,2703	683,5156	692,9816	712,7712
700	590,0480	607,3795	615,9075	625,3175	639,6130	652,4973	668,3308	731,2805	748,3591	762,6607	778,9721	789,9735	800,1314	821,3468
800	682,0665	700,7250	709,8969	720,0107	735,3623	749,1852	766,1555	833,4557	851,6712	866,9114	884,2789	895,9843	906,7862	929,3289
900	774,5698	794,4750	804,2517	815,0267	831,3702	846,0746	864,1125	935,4987	954,7819	970,9036	989,2631	1001,6296	1013,0364	1036,8260

FIGURE I.6 – Table de la loi du Khi-deux

Exemple I.2 :

Regardons comment calculer une probabilité avec ce tableau :

Supposons que X suit une loi du Khi-deux à 2 degrés de liberté. Pour calculer $P(X \geq 0,01)$, on regarde dans la ligne 2, le nombre le plus proche ou un encadrement de 0,01. Ici 0,01 apparaît dans la ligne. En remontant, la colonne correspondante on obtient la probabilité voulu (ou l'encadrement). Donc on a $P(X \geq 0,01) = 0,995$.

Pour la recherche de quantiles, on fait le travail inverse.

ATTENTION : Il faut remarquer que la table de loi du Khi-deux ne donne pas la valeur du quantile d'ordre α mais de $1 - \alpha$. Cela vient du fait qu'il donne les valeurs des probabilités pour $X \geq \dots$ contrairement à la valeur de la fonction de répartition : $F_X(y) = P(X \leq y)$.

I.2.2 LOI DE STUDENT

La loi de Student est très utilisée dans la calcul d'intervalles de confiance et en théorie des tests, notamment pour le test t dit aussi, test de Student.

Comme la loi du Khi-deux, elle dépend d'un nombre de degré de liberté, que l'on notera ici $k \in \mathbb{N}^*$. Elle est définie à partir de la loi de normale centrée réduite et de la loi du Khi-deux.

Définition I.5 (Définition de la loi de Student) :

Si Y et Z sont deux variables aléatoires indépendantes suivant respectivement une loi $\mathcal{N}(0, 1)$ et $\chi^2(k)$ alors la variable aléatoire $X = \frac{Y}{\sqrt{Z/k}}$ suit une loi de Student à k degrés de liberté. On le note :

$$X \sim t(k)$$

La loi de Student est définie alors sur \mathbb{R} .

Proposition I.7 (Fonction de densité et fonction de répartition d'une loi de Student)

La fonction de densité f_X d'une variable aléatoire X qui suit une loi de Student, de degré de liberté k , est définie, pour tout $x \in \mathbb{R}$, par :

$$f_X(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}} \quad (\text{I.10})$$

où Γ désigne la fonction gamma définie pour tout $y \in \mathbb{R}$ par :

$$\Gamma(y) = \int_0^{+\infty} t^{y-1} \exp(-t) dt \quad (\text{I.11})$$

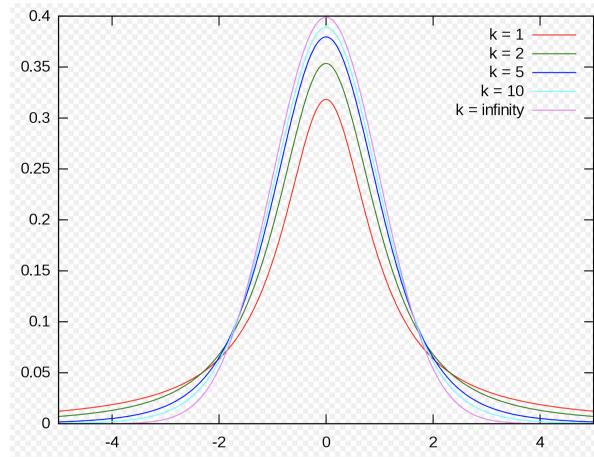


FIGURE I.7 – Fonction de densité de Student

La fonction de répartition de la loi de Student n'a pas de forme explicite utilisable. C'est pourquoi on utilisera, comme pour les autres lois, une table, pour calculer les probabilités.

Pour se donner une idée, on donne une représentation graphique.

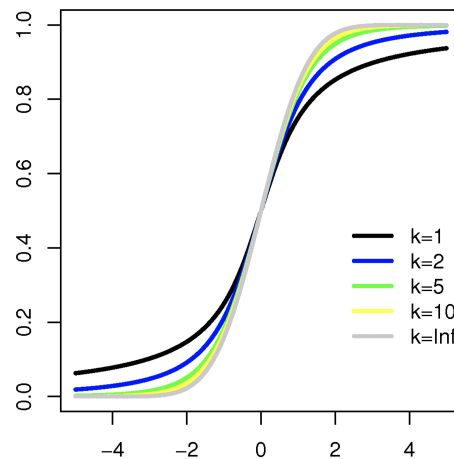


FIGURE I.8 – Fonction de répartition de Student

La parité de la fonction de densité de Student implique la symétrie de la fonction de répartition par rapport au point $(0; 0,5)$. Ainsi, on obtient les mêmes propriétés que pour celle de la loi normale centrée réduite :

Proposition I.8

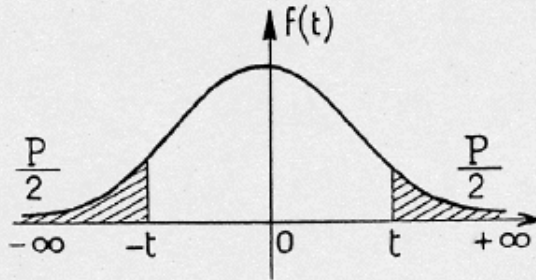
On pose F_k la fonction de répartition de la loi de Student à k degré de liberté. On a alors, pour tout $k \in \mathbb{N}^*$ les égalités suivantes :

$$F_k(0) = 0,5 \tag{I.12}$$

$$F_k(-x) = 1 - F_k(x) \quad \forall x \in \mathbb{R} \tag{I.13}$$

$$F_k^{-1}(1 - \alpha) = -F_k^{-1}(\alpha) \quad \forall \alpha \in [0, 1] \tag{I.14}$$

Voici la table de la loi de Student :



$\frac{P}{v}$	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,929
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
80	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

FIGURE I.9 – Table de la loi de Student

On verra dans le 3ème chapitre qu'on cherche surtout à calculer des probabilités qui ressemblent à $P(|X| \geq \dots)$, si X suit une loi de Student. Par conséquent, la table de loi recense ces probabilités. Par exemple, la première case indique que $P(|X| \geq 0,510) = 0,35$ si X suit une loi de Student à 1 degré de liberté.

On va voir comment retrouver la valeur de la fonction de répartition avec cette table. Calculons, pour $x \geq 0$, la probabilité $P(|X| \geq x)$:

$$\begin{aligned} P(|X| \geq x) &= 1 - P(|X| \leq x) \\ &= 1 - P(-x \leq X \leq x) \\ &= 1 - (P(X \leq x) - P(X \leq -x)) \\ &= 1 - F_X(x) + F_X(-x) \\ &= 1 - F_X(x) + 1 - F_X(x), \text{ d'après la proposition précédente} \\ &= 2 - 2F_X(x) \end{aligned}$$

D'où si $x \geq 0$,

$$F_X(x) = \frac{2 - P(|X| \geq x)}{2}$$

Le même calcul pour $x \leq 0$ donne :

$$F_X(x) = \frac{P(|X| \geq -x)}{2}$$

Le calcul de probabilité avec la table et le calcul des quantiles se fait alors comme précédemment.

Remarque :

La loi de Student a aussi la propriété de "tendre" vers la loi normale centrée réduite quand son degré de liberté tend vers l'infini. C'est une propriété utile que l'on verra plus loin dans ce cours.

$$F_k \xrightarrow[k \rightarrow \infty]{} \Phi$$

I.2.3 LOI DE FISHER-SNEDECOR

La loi de Fisher-Snedecor, ou plus simplement loi de Fisher est souvent utilisée dans le calcul de certaines statistiques de tests, notamment pour le test F dit aussi, test de Fisher.

Elle dépend de deux nombres $n, m \in \mathbb{N}^*$ qui représentent les degrés de liberté. Elle est définie comme le "quotient" de deux lois du Khi-deux.

Définition I.6 (Définition de la loi de Fisher) :

Si Y et Z sont deux variables aléatoires indépendantes suivant respectivement une loi $\chi^2(n)$ et $\chi^2(m)$ alors la variable aléatoire $X = \frac{Y/n}{Z/m}$ suit une loi de Fisher à n et m degrés de liberté. On le note :

$$X \sim \mathcal{F}(n, m)$$

La loi de Fisher est définie alors sur \mathbb{R}^+ .

Proposition I.9 (Fonction de densité et fonction de répartition d'une loi de Fisher)

La fonction de densité f_X d'une variable aléatoire X qui suit une loi de Fisher $\mathcal{F}(n, m)$ est définie, pour tout $x \in \mathbb{R}^+$, par :

$$f_X(x) = \frac{\Gamma\left(\frac{n}{2}\right) \Gamma\left(\frac{m}{2}\right)}{\Gamma\left(\frac{n+m}{2}\right)} n^{\frac{n}{2}} m^{\frac{m}{2}} \frac{x^{\frac{n}{2}-1}}{(m+nx)^{\frac{n+m}{2}}} \quad (\text{I.15})$$

où Γ désigne la fonction gamma définie pour tout $y \in \mathbb{R}$ par :

$$\Gamma(y) = \int_0^{+\infty} t^{y-1} \exp(-t) dt \quad (\text{I.16})$$

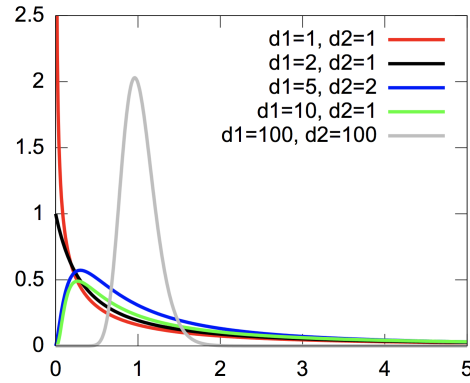


FIGURE I.10 – Fonction de densité de la loi de Fisher

La fonction de répartition de la loi de Fisher n'a pas de forme explicite utilisable. C'est pourquoi on utilisera, comme pour les autres lois, une table, pour calculer les probabilités.

Pour se donner une idée, on donne une représentation graphique de cette fonction :

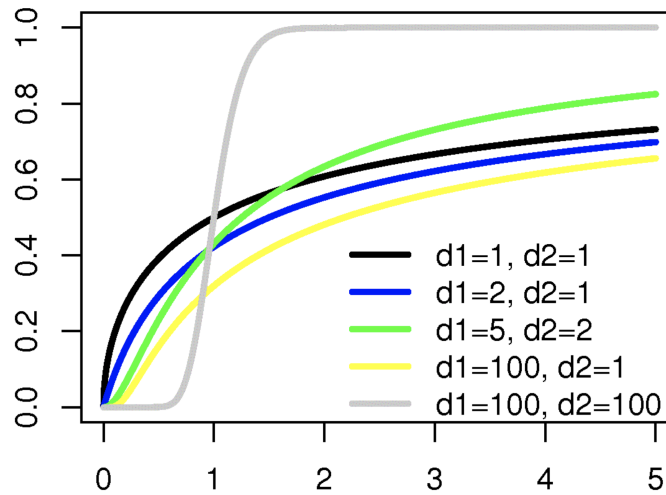


FIGURE I.11 – Fonction de répartition de la loi de Fisher

Voici une table de la loi de Fisher. Sur cette table ν_1 et ν_2 représentent les degrés de liberté de la loi. La différence avec les précédentes c'est qu'elle ne représente que la probabilité 0,95 par manque de place. Il existe donc plusieurs tables avec chaque valeur (en sachant que les plus utilisées sont celles pour 0,95 et 0,99).

ν_2 (dén.)	ν_1 (numérateur)																			
	1	2	3	4	5	6	7	8	9	10	20	30	40	50	60	80	100	200	500	1 000
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	248.02	250.10	251.14	251.77	252.20	252.72	253.04	253.68	254.06	254.19
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.45	19.46	19.47	19.48	19.48	19.48	19.49	19.49	19.49	19.49
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.66	8.62	8.59	8.58	8.57	8.56	8.55	8.54	8.53	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.80	5.75	5.72	5.70	5.69	5.67	5.66	5.65	5.64	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.56	4.50	4.46	4.44	4.43	4.41	4.41	4.39	4.37	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.87	3.81	3.77	3.75	3.74	3.72	3.71	3.69	3.68	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.44	3.38	3.34	3.32	3.30	3.29	3.27	3.25	3.24	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.15	3.08	3.04	3.02	3.01	2.99	2.97	2.95	2.94	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	2.94	2.86	2.83	2.80	2.79	2.77	2.76	2.73	2.72	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.77	2.70	2.66	2.64	2.62	2.60	2.59	2.56	2.55	2.54
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.12	2.04	1.99	1.97	1.95	1.92	1.91	1.88	1.86	1.85
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	1.93	1.84	1.79	1.76	1.74	1.71	1.70	1.66	1.64	1.63
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.84	1.74	1.69	1.66	1.64	1.61	1.59	1.55	1.53	1.52
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.78	1.69	1.63	1.60	1.58	1.54	1.52	1.48	1.46	1.45
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.75	1.65	1.59	1.56	1.53	1.50	1.48	1.44	1.41	1.40
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.72	1.62	1.57	1.53	1.50	1.47	1.45	1.40	1.37	1.36
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.70	1.60	1.54	1.51	1.48	1.45	1.43	1.38	1.35	1.34
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.69	1.59	1.53	1.49	1.46	1.43	1.41	1.36	1.33	1.31
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.68	1.57	1.52	1.48	1.45	1.41	1.39	1.34	1.31	1.30
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.62	1.52	1.46	1.41	1.39	1.35	1.32	1.26	1.22	1.21
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.61	1.50	1.43	1.39	1.36	1.32	1.30	1.23	1.19	1.17
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.59	1.48	1.42	1.38	1.35	1.30	1.28	1.21	1.16	1.14
1 000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.58	1.47	1.41	1.36	1.33	1.29	1.26	1.19	1.13	1.11
2 000	3.85	3.00	2.61	2.38	2.22	2.10	2.01	1.94	1.88	1.84	1.58	1.46	1.40	1.36	1.32	1.28	1.25	1.18	1.12	1.09

FIGURE I.12 – Table de la loi de Fisher pour la quantile $P(X \leq \dots) = 0,95$

Nous allons à présent donner une propriété facile à remarquer sur la loi de Fisher qui peut aider à calculer certaines probabilités.

Proposition I.10

Soit X une variable aléatoire réelle non nulle. Si, de plus, $X \sim \mathcal{F}(n, m)$, alors $X^{-1} \sim \mathcal{F}(m, n)$.

On peut alors en déduire deux relations importantes de la fonction de répartition de la loi de Fisher :

Corollaire I.10.1

Si on note $F_{(n,m)}$ la fonction de répartition d'une loi de Fisher $\mathcal{F}(n, m)$, on a les relations suivantes :

$$F_{(n,m)}(x) = 1 - F_{(m,n)}\left(\frac{1}{x}\right), \quad \forall x \in \mathbb{R}^+ \tag{I.17}$$

$$F_{(n,m)}^{-1}(\alpha) = \frac{1}{F_{(m,n)}(1 - \alpha)}, \quad \forall \alpha \in [0, 1] \tag{I.18}$$

Le calcul de probabilité avec la table et le calcul des quantiles se fait alors comme précédemment. Par exemple, en gardant les notations de la proposition :

$$F_{(1,1)}(161, 45) = 0,95$$

$$F_{(9,300)}(1, 91) = 0,95$$

II

INTERVALLES DE CONFIANCES

II.1 PRÉLIMINAIRES : CONVERGENCE EN LOI ET THÉORÈME CENTRAL LIMITE

On est parfois amené à essayer de comprendre le comportement d'une suite de variables aléatoires. Par exemple, dans une usine de fabrication de boulons, la taille du boulon est une variable aléatoire qui suit une loi normale. Ainsi, comme on la taille de chaque boulon qui sort de la production suit une variable aléatoire, on a une suite de variables aléatoires que l'on peut noter $(X_n)_{n \in \mathbb{N}}$ (comme une suite numérique sauf qu'ici chaque élément de la suite n'est plus un nombre mais une variable aléatoire). Pour cela, les mathématiciens ont développé la théorie asymptotique des variables aléatoires.

On a alors défini des convergences de variables aléatoires plus ou moins fortes. La seule qui va nous intéresser ici (dans ce cours) est celle de la convergence en loi.

Définition II.1 (Convergence en loi) :

Soit une suite de variables aléatoires réelles (X_n) ayant pour fonction de répartition F_n . On dit que la suite de variables aléatoires (X_n) converge en loi vers une variable aléatoire réelle X de fonction de répartition F , si, pour tout x dans l'ensemble de définition de X ,

$$\lim_{n \rightarrow \infty} F_n(x) = F(x)$$

On le note :

$$X_n \xrightarrow{\mathcal{L}} X$$

Remarque :

L'important ici est la loi d'arrivée. Comme la fonction de répartition caractérise la loi, cette définition fait sens. Une définition équivalente pourrait être avec les fonctions de densités (si elles existent pour chaque X_n).

Par exemple, on a déjà vu précédemment que si X_k suit une loi de Student à k degrés de liberté : $X_k \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$.

Supposons que $X_n \xrightarrow{\mathcal{L}} X$ avec $X \sim \mathcal{N}(0, 1)$, alors la variable aléatoire X a peu d'importance puisque toutes les variables aléatoires qui suivent une $\mathcal{N}(0, 1)$ ont la même fonction de répartition. On peut donc écrire :

$$X_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) .$$

Autrement dit, la convergence en loi ne donne que l'information sur la loi d'arrivée de la variable aléatoire, mais rien de plus.

Rappel : La loi d'une variable aléatoire ne définit pas la variable elle-même. Deux variables aléatoires peuvent avoir la même loi tout en étant différentes. Par exemple, on lance une pièce de monnaie équilibrée. Considérons la variable aléatoire X_{pile} qui vaut 1 quand la pièce tombe sur *pile* et 0 sinon et la variable aléatoire X_{face} qui vaut 1 quand la pièce tombe sur *face* et 0 sinon. Ces deux variables ont la même loi. On a une chance sur deux de faire *face* comme de faire *pile*.

$$P(X_{\text{pile}} = 1) = P(X_{\text{face}} = 1) = 1/2 \quad \text{et} \quad P(X_{\text{pile}} = 0) = P(X_{\text{face}} = 0) = 1/2$$

Pourtant ces deux variables aléatoires sont clairement différentes. Par exemple, quand la pièce tombe sur *pile*, $X_{\text{pile}} = 1$ et $X_{\text{face}} = 0$.

Le résultat qui suit donne encore plus d'importance à la loi normale. Il est indispensable à toute la théorie des estimateurs, surtout à la construction de certains intervalles de confiance.

Il donne le comportement asymptotique de la moyenne empirique :

Définition II.2 :

La moyenne empirique (\bar{X}_n) d'une suite de variables aléatoires réelles (X_n) est en fait la moyenne de ses premiers termes :

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Ne voyez pas cela comme une nouvelle formule : C'est juste la moyenne des n premières variables aléatoires de la suite $(X_n)_{n \in \mathbb{N}}$.

Nous pouvons donc énoncer le théorème :

Théorème II.1 (Théorème de la limite centrée (ou Théorème central limite))

Soit (X_n) une suite de variables aléatoires réelles indépendantes et identiquement distribuées (iid) avec une espérance μ et une variance σ^2 finies. Alors la moyenne empirique vérifie :

$$\sqrt{n} \left(\frac{\bar{X}_n - m}{\sigma} \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

Remarque :

Le résultat du théorème peut aussi se noter :

$$\left(\frac{\sum_{i=1}^n X_i - nm}{\sigma \sqrt{n}} \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

II.2 ÉCHANTILLONAGE

On se demande parfois comment un sondage sur 1000 individus d'une population peut donner le résultat d'une élection. Comment en vérifiant que 100 voitures fonctionnent, on a de bonnes chances que toutes les voitures du même modèle fonctionneront de la même manière. En fait, comment en prenant des informations sur une petite partie d'une population, on peut avoir de bonnes idées sur l'ensemble de la population ? Ce n'est pas anodin, car sonder l'ensemble d'une population est souvent difficile voir impossible.

On appelle cette méthode, l'échantillonnage. On prend un échantillon d'une population pour essayer de connaître certaines informations sur la population entière. La première question que l'on peut se poser est comment prendre cet échantillon ?

Il y a les **échantillons aléatoires** : On choisit les individus d'une population au hasard. Un des problèmes est que le hasard n'est jamais parfait. Rien ne peut nous assurer, que nous ne prendrons pas au hasard 90% de moins de 25 ans, dans une population qui n'en a que 30%.

L'autre méthode est donc celle des **échantillons représentatifs** : On choisit alors certains critères qui discriminent les individus d'une population en cherchant à reproduire les catégories de la population. Le problème est qu'il est très difficile voir impossible que considérer tous les critères existants, ainsi on biaise forcément le choix. Par exemple, dans la population française, il n'y a pas un nombre fini de critères à prendre en compte. Les gens diffèrent par leur taille, poids, âges, religions, taille des jambes, taille des pieds, couleurs peaux, couleurs de cheveux, etc... C'est impossible d'avoir un échantillon prenant en compte chaque critère.

De plus, il y a deux points importants à prendre en compte lorsque qu'on choisit un échantillon : La taille de celui-ci et la taille totale de la population. Si la taille de la population est très petite il faut faire attention à la façon de constituer l'échantillon. Celui-ci peut être rapidement biaisé.

Pour tous ces problèmes on restera dans le cadre suivant dans le reste de ce cours : Les populations choisies seront de taille très grande et les échantillons seront aléatoires et de taille importante.

Définition II.3 (échantillon) :

On appelle *n*-échantillon aléatoire une suite de variables aléatoires, notée X_1, \dots, X_n , où X_i désigne la valeur de la caractéristique d'intérêt associée au $i^{\text{ème}}$ individu pris au hasard dans la population.

Remarque :

Il ne faut pas confondre, un échantillon qui est composé d'une suite de **variables aléatoires** et la réalisation de cet échantillon après expérience, que l'on notera x_1, \dots, x_n . Le premier est probabiliste, la seconde est déterministe. Par exemple, considérons un échantillon aléatoire X_1, \dots, X_n qui donne la taille de n étudiants en L2 AES tirés au hasard. Ici X_1, \dots, X_n sont des variables aléatoires (toutes de même loi) et non des données numériques. Par contre, on note x_1, \dots, x_n les résultats d'une expérience, qui sont n tailles de n étudiants. Ces tailles x_1, \dots, x_n sont donc des nombres.

II.3 ESTIMATION PAR INTERVALLE DE CONFIANCE

Supposons que nous cherchons à connaître un des paramètres de la loi qui régit un caractère d'une population (par exemple, les paramètres μ et σ^2 qui déterminent la loi normale qui régit la taille des individus). Nous essayons alors, à partir d'un échantillon aléatoire de la population, d'estimer les valeurs de ces paramètres. C'est ce que l'on appelle l'**estimation**.

Il existe plusieurs manières d'estimer les paramètres des lois en jeu. Dans ce cours on se concentrera sur l'estimation par intervalle de confiance. L'avantage : Donner l'appartenance d'une variable aléatoire à un intervalle selon une probabilité donnée.

Définition II.4 (Intervalle de confiance) :

Soit X_1, \dots, X_n un n -échantillon. Un intervalle de confiance sur le paramètre θ pour un niveau de risque $\alpha \in]0, 1[$ est un encadrement donné par deux variables aléatoires A, B fonctions de l'échantillon X_1, \dots, X_n :

$$P(A \leq \theta \leq B) = 1 - \alpha$$

Par exemple, si on veut connaître un intervalle de "tailles d'individus" dans lequel si on tire un français au hasard, on a 95% de chance qu'il ait une taille qui soit dans cet intervalle.

Pour pouvoir construire les intervalles de confiance, on a besoin d'un peu de théorie des estimateurs.

II.3.1 ESTIMATEURS

Définition II.5 (Estimateur) :

Un estimateur d'un paramètre θ , noté $\hat{\theta}$, est une fonction d'un n -échantillon X_1, \dots, X_n :

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

La notion d'estimateur est donc très vaste. On a besoin de poser quelques définitions supplémentaires pour savoir si un estimateur est "bon" ou "mauvais" pour θ .

Définition II.6 (Estimateur non-biaisé) :

On dit qu'un estimateur $\hat{\theta}$ est non-biaisé si son espérance est égale à θ :

$$\mathbb{E}(\hat{\theta}) = \theta$$

Le biais d'un estimateur désigne donc son écart avec θ . Ci dessous la courbe rouge représente la fonction de densité d'un estimateur non-biaisé pour $\theta = 2$, tandis que les courbes bleues représentent les fonctions de densité d'estimateurs biaisés.

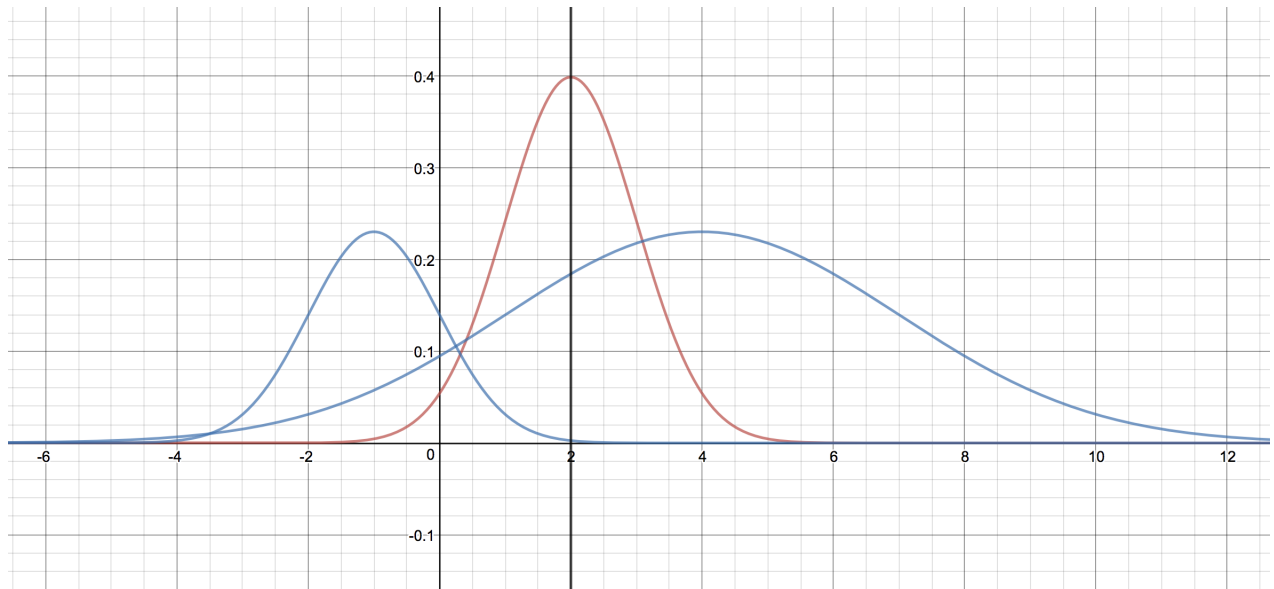


FIGURE II.1 – Estimateurs biaisés / non-biaisés

Il est clair que l'on préfère donc les estimateurs non biaisés.

Exemple II.1 :

Montrons que la moyenne empirique d'un échantillon est un estimateur non-biaisé de l'espérance. Considérons donc une population dont la taille en mètres suit une loi normale $\mathcal{N}(m, 1)$. Montrons que

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

est un estimateur sans biais de m :

$$\begin{aligned} \mathbb{E}(\bar{X}) &= \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) \quad , \text{ par linéarité de l'espérance} \\ &= \frac{1}{n} \sum_{i=1}^n m \quad , \text{ car les } X_i \text{ sont dans l'échantillon et suivent une loi } \mathcal{N}(m, 1) \\ &= m \end{aligned}$$

Un autre exemple pourrait être de montrer que la variance empirique

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

est un estimateur biaisé de la variance. C'est pour cela que l'on a défini la variance empirique corrigée

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

qui est un estimateur non biaisé de la variance.

II.3.2 CALCULS D'INTERVALLE DE CONFIANCE

Nous allons établir une méthode générale pour établir un intervalle de confiance. Attention, cette méthode ne fonctionne pas dans tous les cas, mais dans tous ceux que l'on aura à traiter dans ce cours.

Imaginons que nous souhaitons estimer un paramètre θ inconnu. On veut donc l'encadrer par deux variables aléatoires A et B qui sont fonctions de l'échantillon X_1, \dots, X_n .

1. Soit $\hat{\theta}$ un estimateur sans biais et convergent du paramètre θ . On essaie alors de changer cette variable aléatoire de manière à ce que sa loi soit connue (on utilise en général le théorème central limite II.1).
Notons la nouvelle variable aléatoire $g(\theta, \hat{\theta})$. Cette variable aléatoire doit être vide de tout paramètre inconnu différent de θ .
2. Comme on connaît la loi de $g(\theta, \hat{\theta})$ on peut trouver deux constantes a et b telles que :

$$P(a \leq g(\theta, \hat{\theta}) \leq b) = 1 - \alpha$$

Si $g(\theta, \hat{\theta})$ suit une loi normale, on choisit a et b de tel manière que :

$$P(a \leq g(\theta, \hat{\theta})) = 1 - \frac{\alpha}{2} \quad P(g(\theta, \hat{\theta}) \leq b) = 1 - \frac{\alpha}{2} .$$

De cette manière, on a un intervalle symétrique.

3. Ensuite, on isole θ au centre pour avoir un encadrement de θ , et on a notre intervalle de confiance :

$$P(A(\hat{\theta}) \leq \theta \leq B(\hat{\theta})) = 1 - \alpha$$

4. Enfin, si on veut connaître l'intervalle de notre paramètre θ , il suffit de nous donner une réalisation $x = (x_1, \dots, x_n)$ de notre échantillon, puis de remplacer cette réalisation dans notre encadrement. On obtient alors notre intervalle de confiance :

$$IC_{1-\alpha} = [A(\hat{\theta}(x)), B(\hat{\theta})] = [a, b]$$

Cela veut dire, que l'on est sûr avec une probabilité de $(1 - \alpha)$ que θ est dans l'intervalle $[a, b]$.

Exemple II.2 :

Dans cet exemple, on veut estimer l'espérance μ d'une variable aléatoire X qui suit une loi normale $\mathcal{N}(\mu, 3, 25)$. On a à notre disposition un échantillon de taille $n = 200$ de variables aléatoires i.i.d. de même loi que X . On sait d'autre part que la moyenne de la réalisation cet échantillon est $\bar{x}_n = 5,25$. On veut que l'intervalle ait niveau de risque $\alpha = 5\%$. Autrement dit, on veut que la probabilité que μ soit dans cet intervalle de confiance soit de 0,95.

1. On veut estimer l'espérance μ (qui est la moyenne de la variable aléatoire) donc on choisit ... la moyenne empirique comme estimateur. D'après la linéarité de la loi normale (Proposition I.5) on a (*On choisit ici un intervalle centré sur la moyenne. C'est souvent le plus utilisé quand on a une loi normale*) :

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

D'où par réduction :

$$\frac{\bar{X}_n - \mathbb{E}(\bar{X}_n)}{\sqrt{\mathbb{V}(\bar{X}_n)}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

On a réussi à ramener notre estimateur \bar{X}_n à une loi qui ne dépend pas de θ . Ici

$$g(\bar{X}_n, \mu) = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

2. Les quantiles de la loi normale centrée réduite $\mathcal{N}(0, 1)$ sont donnés par sa fonction de répartition Φ . Ainsi, on a :

$$P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \frac{\alpha}{2}$$

et

$$P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \geq \Phi^{-1}\left(\frac{\alpha}{2}\right)\right) = 1 - \frac{\alpha}{2}$$

D'où :

$$P\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha \iff$$

3. On isole μ :

$$P\left(\bar{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{\alpha}{2}\right) \geq \mu \geq \bar{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$

Or, d'après la propriété de symétrie de la loi normale, on a $\Phi^{-1}\left(\frac{\alpha}{2}\right) = -\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$. Donc on peut conclure sur notre intervalle de confiance :

$$P\left(\bar{X}_n + \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \geq \mu \geq \bar{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$

On peut donc appliquer numériquement au risque $\alpha = 5\%$. Notre intervalle de confiance est :

$$\begin{aligned} IC &= \left[\bar{x}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \bar{x}_n + \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right] \\ &= \left[5,25 - \frac{\sqrt{3,25}}{\sqrt{200}}\Phi^{-1}(0,975), 5,25 + \frac{\sqrt{3,25}}{\sqrt{200}}\Phi^{-1}(0,975)\right] \\ &= [5, 5,5] \end{aligned}$$

Supposons maintenant que l'on garde l'exemple précédent mais que l'on ne connaisse pas la loi de X (mais que l'on connaisse quand même la variance). On doit alors se remettre au Théorème de la limite centrée II.1. D'après ce théorème (si $n > 30$) :

$$\sqrt{n}\left(\frac{\bar{X}_n - m}{\sigma}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

On a donc un intervalle de confiance quand n est assez grand. On peut considérer ici qu'il l'est. On appelle cela un intervalle de confiance asymptotique. Le reste se calcule de la même manière que l'exemple précédent.

III

TESTS D'HYPOTHÈSE

Dans ce chapitre, on introduit la **théorie des tests**. La notion de "test" en statistiques nous sert à avoir des raisons objectives de prendre des décisions. Par exemple, pour savoir si le modèle que l'on a choisit est bon, si deux évènements sont indépendant, ou encore si un évènement a bouleversé ou pas le chiffre d'affaire d'une entreprise. Cette théorie étudie la construction et les propriétés de ces tests. C'est la base de la "*business intelligence*". En pratique, les tests ne sont pas seulement là pour nous dire quelle décision prendre, ils mesurent aussi les risques de prendre une décision ou une autre.

Par exemple, quand une banque veut faire un prêt à un client, il peuvent mesurer le risque que le client ne rembourse pas. Pour cela, il fait une hypothèse appelée **hypothèse nulle** selon laquelle le client est solvable. Ensuite, il décide d'une **règle de décision** qui entraînera le **rejet** ou le **non-rejet** de cette hypothèse.

Autre exemple : Une entreprise qui veut savoir si une manifestation a eu un impact sur ses vente. Elle établit alors l'hypothèse nulle selon laquelle la manifestation n'a pas eu d'impact et construit une règle de décision à partir d'un échantillon quant au rejet ou non de l'hypothèse.

Nous verrons deux tests basés sur la loi du **Khi-deux** en fin de cours et d'autres si le temps le permet.

Définition III.1 (Test statistique) :

Un **test statistique** est une règle de décision relative à une hypothèse sur la loi d'une variable, qui se fonde sur les **observations** d'un échantillon de la population.

Dans un test statistique, on prend toujours une décision en confrontant deux hypothèses :

- **L'hypothèse nulle**, notée H_0 qui est l'hypothèse de référence que l'on veut tester.
- **L'hypothèse alternative**, notée H_1 contre laquelle on veut tester l'hypothèse H_0 .

On peut se demander comment choisir l'hypothèse nulle. Par exemple, imaginons que l'on veut savoir si une action marketing a de l'influence sur les ventes d'un magasin. On peut choisir de tester H_0 : "L'action a de l'influence" contre H_1 : "L'action n'a pas d'influence". Mais on peut aussi tester H_0 : "L'action n'a pas d'influence" contre H_1 : "L'action a de l'influence".

En général, on choisit H_0 comme l'hypothèse dont le coût d'une erreur la plus élevée. Par exemple, ici, dans le premier cas on risque de refinancer l'action alors que dans l'autre on risque seulement d'avoir un manque à gagner. On peut donc choisir le premier cas. Ce choix influe sur la règle de décision.

On distingue deux types de tests :

- Les **tests paramétriques** : Ils portent sur un ou plusieurs paramètres de la loi d'une caractéristique d'une population.
- Les **tests non-paramétriques** : Ceux-ci concernent une loi ou certaines caractéristiques de variables aléatoires.

Exemple III.1 :

Un exemple de test paramétrique est si on veut tester la paramètre d'une loi. Par exemple, pour une variable aléatoire X qui suit une loi de Poisson de paramètre θ , on peut tester :

$$H_0 : \theta = 3 \quad \text{contre} \quad \theta = 4$$

Un exemple de test non paramétrique a été donné ci dessus (celui du test de l'action marketing). On verra aussi les tests du Khi-deux.

Parmi les tests paramétriques (on prendra θ comme paramètre quelconque), on distingue les **hypothèse simples** qui caractérise complètement la loi en cas d'acceptation (Si X suit une loi du Khi-deux de paramètre k , " $H_0 : k = 5$ " caractérise la loi en cas d'acceptation puisque on aura $X \sim \chi^2(5)$) aux **hypothèses composites** qui laisse plusieurs possibilités à la loi (dans le même exemple, " $H_0 : k \neq 5$ " laisse beaucoup de possibilités à k , en cas d'acceptation). Dans le deuxième cas, sous l'hypothèse H_0 , il est clair que l'on ne peut pas calculer avec exactitude $P(X \leq c)$, où $c \in \mathbb{R}$. Généralement on construit souvent les tests avec H_0 une hypothèse simple.

Dans les tests "hypothèse simple contre hypothèse composite", il existe deux familles :

- Les **tests unilatéraux** : Ils sont de la forme $H_0 : \theta = \theta_0$ contre $H_1 : \theta < \theta_0$ (ou $H_1 : \theta > \theta_0$).
- Les **tests bilatéraux** : Ils sont de la forme $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$

III.1 RÉGION CRITIQUE D'UN TEST

La région critique d'un test est la zone où on rejette l'hypothèse nulle H_0 . Pour cela, on doit définir une **statistique de test** et une **valeur critique**.

Définition III.2 :

Une **statistique de test**, notée T_n , est une variable aléatoire définie comme fonction des variables de l'échantillon X_1, \dots, X_n :

$$T_n(X_1, \dots, X_n)$$

La **région critique** d'un test, notée W , est un ensemble de réalisations de la statistique de test pour lesquelles l'hypothèse nulle est rejetée. Elle est de la forme :

$$W = \{x_1, \dots, x_n | T_n(x_1, \dots, x_n) \in R(c)\}$$

où $R(c)$ est un ensemble délimité par des valeurs notées c ici.

La **région de non rejet** de l'hypothèse nulle H_0 est donc le complémentaire de W noté \overline{W} .

Remarque :

Ces définitions sont ici à titre informatif, pour fixer les bases mathématiques. Le but de ce cours n'est pas de faire comprendre l'ensemble des fondements mathématiques des tests statistiques. C'est pour cela que l'on se rappellera surtout les points suivants : **Généralement (et sûrement à tout moment dans ce cours) la statistique du test n'est rien d'autre qu'un estimateur du paramètre θ à tester. De plus, la région critique est seulement la zone où l'on rejette l'hypothèse H_0 . Si la réalisation de la statistique de test, pour les données concrètes de l'échantillon, appartient à la région critique, on rejette H_0 ;**

Exemple III.2 :

Lors des dernières élections, un parti a réalisé 21% des suffrages. On réalise, 1 an plus tard, un sondage sur 1000 personnes et 18% des sondés affirment vouloir voter pour ce parti. On s'interroge alors sur cette baisse de pourcentage. On pose alors l'hypothèse nulle $H_0 : p = 21\%$ et l'hypothèse alternative $p < 21\%$.

Sous l'hypothèse H_0 , le nombre de vote pour le parti, N suit une loi binomiale de paramètre $n = 1000$ et $\pi = 0,21$. Ainsi, ce nombre admet pour espérance $n\pi$ et pour variance $n\pi(1 - \pi)$. Donc la proportion $\pi = \frac{N}{n}$ a pour espérance π et pour variance $\frac{\pi(1-\pi)}{n}$. Or, comme n est assez grand, d'après le théorème de la limite centrée, $\frac{p-\pi}{\sqrt{\pi(1-\pi)/n}} \sim \mathcal{N}(0, 1)$. Une région critique liée à l'observation est :

$$W = \{x|p(x) < 0,18\}$$

III.2 RISQUES D'UN TEST

Quand on fait un test statistique, on considère 2 risques : les risques de première et de deuxième espèce. Ce sont seulement des définitions !

Définition III.3 :

Le risque de première espèce (ou risque d'erreur de type I) est le risque que H_0 soit vraie alors qu'on l'a rejeté. Le risque de deuxième espèce (ou risque d'erreur de type II) est le risque que H_1 soit vraie alors qu'on n'a pas rejeté H_0 .

		Décision	
		Non-rejet de H_0	Rejet de H_0
Population	H_0 Vrai	Bonne décision	Erreur de type I
	H_1 Vrai	Erreur de type II	Bonne décision

TABLE III.1 – Risques de type I et II

Les probabilités de ces risques donnent des informations importantes pour la prise de décision. Elles sont souvent fixées par ceux qui commandent les tests statistiques pour leur entreprise (par exemple).

Définition III.4 (Niveau d'un test) :

Le niveau d'un test, noté α , est la probabilité du risque d'erreur de type I. Autrement dit, c'est la probabilité d'être dans la région critique W sachant que H_0 est vrai, donc la probabilité que H_0 soit fausse alors qu'on l'a accepté :

$$\alpha = P(W|H_0)$$

Il est clair, que le but d'un test est de se tromper le moins possible. C'est pourquoi plus le niveau d'un test est faible, mieux c'est ! On ne veut pas que H_0 soit fausse alors qu'on l'a accepté !

Définition III.5 (Puissance d'un test) :

On appelle puissance d'un test, la probabilité de rejeter l'hypothèse nulle H_0 alors que l'hypothèse H_1 est vraie dans la population. C'est donc la probabilité du complémentaire du risque de seconde espèce :

$$\text{Puiss} = P(W|H_1) = 1 - \beta$$

où β est la probabilité du risque d'erreur de type II, $\beta = P(\bar{W}|H_1)$.

Comme précédemment, on aimerait que le risque de seconde espèce soit le plus petit possible. Donc plus le test est puissant (= plus la puissance du test est grande), mieux c'est !

Exemple III.3 (Comment calculer la région critique à partir d'un risque de première espèce α donné) :

Reprenons l'exemple du sondage précédent. Supposons que l'on accepte un risque de première espèce de $\alpha = 5\%$. Alors, par définition, on cherche c tel que :

$$P(p < c) = 0,05$$

Ramenons nous alors à une loi normale centrée réduite grâce au théorème de la limite centrée :

$$P(p < c) = P\left(\frac{p - \pi}{\sqrt{\pi(1 - \pi)/n}} < \frac{c - \pi}{\sqrt{\pi(1 - \pi)/n}}\right) = 0,05$$

On cherche alors la quantile d'ordre α dans la table et on a : $\Phi^{-1}(\alpha) = -1,645$. Donc :

$$\frac{c - \pi}{\sqrt{\pi(1 - \pi)/n}} = -1,645$$

D'où :

$$c = 18,4\%$$

Ce résultat indique que l'on doit rejeter l'hypothèse nulle, au risque de 5% si la proportion est inférieure à 18,4%. Ce qui est le cas ici. Donc on peut conclure à une baisse réelle de la popularité du parti.

III.3 PRENDRE UNE DÉCISION

Maintenant que nous avons défini toutes les notions qu'il faut pour comprendre ce qu'est un test statistique. On peut alors se poser la question : Comment prendre une décision lors d'un test. On va établir des règles pour cela. Essayons de donner une manière pour procéder à un test statistique.

1. On pose l'hypothèse nulle H_0 et l'hypothèse alternative H_1 . Il ne faut pas oublier de choisir H_0 comme l'hypothèse pour laquelle l'erreur a un "coût" moins élevé.
2. On définit la statistique de test T_n . En général, celle ci est un estimateur du paramètre θ , pour un test paramétrique (Pour les tests du Khi-deux, $T_n \sim \chi^2(k)$).
3. Calculer la loi de T_n sachant H_0 .
4. Maintenant, il y a deux possibilités. Les variables β , α et c (la constante qui apparaît dans la définition de la région critique) sont dépendantes les unes des autres. On a vu que la connaissance de la région critique permet de définir α et β (voir les définitions). Il arrive aussi que l'utilisateur impose la valeur de α (1, 5 ou 10% par exemple). Dans ce cas, on peut trouver la région critique (et donc la valeur de β).
5. A partir des observations faites sur l'échantillon, calculer la réalisation de la statistique de test T_n .
6. Pour finir, on peut comparer la réalisation à la région critique du test. Alors, si la réalisation appartient à la région critique du test, on peut rejeter l'hypothèse nulle pour un niveau de risque α . Sinon, on accepte l'hypothèse nulle.

III.4 TESTS DU KHI-DEUX

III.4.1 TEST D'ADÉQUATION DU KHI-DEUX

Parfois, un ensemble de mesures nous donne un tableau de valeurs qui provient de l'observation d'un phénomène. On cherche alors une loi probabiliste qui modélise ce phénomène. Le test d'adéquation du Khi-deux sert à savoir si notre loi correspond bien aux observations. Ce test suit alors les hypothèses suivantes :

$$H_0 : X \sim L(\theta) \quad \text{contre} \quad H_1 : X \text{ ne suit pas la loi } L(\theta)$$

où $L(\theta)$ est une loi de paramètre θ **connu**.

L'idée est de comparer les valeurs données par la loi, aux observations données. On définit alors la statistique du Khi-deux qui définit le test :

Définition III.6 :

Considérons le tableau d'observations suivant :

Variable X	$X = O_1$	\dots	$X = O_j$	\dots	$X = O_k$	Total
Effectifs empiriques (basés sur les observations)	n_1	\dots	n_j	\dots	n_k	n
Effectifs théoriques (calculés à partir de la loi)	N_1	\dots	N_j	\dots	N_k	n

La statistique de test d'adéquation du Khi-deux, notée K_n , est définie par :

$$K_n = \sum_{j=1}^k \frac{(n_j - N_j)^2}{N_j}$$

La variable aléatoire K_n suit alors, sous l'hypothèse H_0 , une loi du Khi-deux à $k - 1$ degrés de liberté.

Exemple III.4 :

On réalise un sondage sur 200 personnes qui doivent donner le nombre d'ordinateurs chez eux :

Nombre d'ordinateurs	0	1	2	3	4	5	6 et plus
Effectifs	60	73	43	17	5	2	2

On se propose de modéliser la cette observation par une loi de Poisson de paramètre 1,2. On rappelle que la loi de Poisson est donnée par la formule :

$$P(X = k) = e^{-1,2} \times \frac{1,2^k}{k!}$$

En faisant les calculs on trouve le nombre de personnes théoriques dans chaque cas.

Nombre d'ordinateurs	0	1	2	3	4 et plus
Effectifs	60	73	43	17	7
Effectifs théoriques	60,25	72,28	43,38	17,34	6,76

Le "6,76" a été obtenu de manière à avoir un effectif théorique total à 200. On a regroupé les effectifs trop faibles pour ne pas fausser le test. Les écarts sont faibles. Mais sans un test, on ne peut que prendre des décisions subjectives. Le résultats du test du khi-deux ici donne :

La réalisation de K_n est alors : $K_n(x) = 0,02664$.

Ce résultat doit être comparé avec la loi du Khi-deux à $5 - 1 = 4$ degrés de liberté. La région critique du test du Khi-deux est donnée par $W = \{x | K_n(x) > F_4^{-1}(1 - \alpha)\}$, où F_5 est la fonction de répartition de la loi du Khi-deux à 4 degrés de liberté.

Imaginons ici que l'on souhaite faire teste de niveau $\alpha = 1\%$. Sur la table du Khi-deux on peut trouver $F_4^{-1}(1 - \alpha) = 13,2767$ ce qui est bien supérieur à la réalisation de la statistique de test. Ainsi, on peut accepter sans problème H_0 et donc on peut conclure que la loi de Poisson modélise bien le phénomène.

III.4.2 TEST D'INDÉPENDANCE

Le test d'indépendance du Khi-deux, un un système de prise de décision quant à l'indépendance de deux variables aléatoires. Les hypothèses mise en jeu sont donc :

$$H_0 : X \text{ et } Y \text{ sont indépendants} \quad \text{contre} \quad H_1 : X \text{ et } Y \text{ ne sont pas indépendants}$$

L'idée est de comparer les valeurs d'un tableau indépendant, calculé à partir des effectifs totaux, et les valeurs observées.

Définition III.7 :

Considérons le tableau d'observations suivant :

$X \setminus Y$	$X = a_1$	\dots	$X = a_j$	\dots	$X = a_k$	Total
$Y = b_1$	$n_{1,1}$	\dots	$n_{j,1}$	\dots	$n_{k,1}$	$n_{Y=1}$
\dots	\dots	\dots	\dots	\dots	\dots	\dots
$Y = b_i$	$n_{1,i}$	\dots	$n_{j,i}$	\dots	$n_{k,i}$	$n_{Y=i}$
\dots	\dots	\dots	\dots	\dots	\dots	\dots
$Y = b_l$	$n_{1,l}$	\dots	$n_{j,l}$	\dots	$n_{k,l}$	$n_{Y=l}$
Total	$n_{X=1}$	\dots	$n_{X=j}$	\dots	$n_{X=k}$	n

On mets en comparaison le tableau de contingence théorique, calculé, sous l'hypothèse H_0 , grâce à la définition d'indépendance entre X et Y :

$X \setminus Y$	$X = a_1$	\dots	$X = a_j$	\dots	$X = a_k$	Total
$Y = b_1$	$\frac{n_{Y=1} \times n_{X=1}}{n}$	\dots	$\frac{n_{Y=1} \times n_{X=j}}{n}$	\dots	$\frac{n_{Y=1} \times n_{X=k}}{n}$	$n_{Y=1}$
\dots	\dots	\dots	\dots	\dots	\dots	\dots
$Y = b_i$	$\frac{n_{Y=i} \times n_{X=1}}{n}$	\dots	$\frac{n_{Y=i} \times n_{X=j}}{n}$	\dots	$\frac{n_{Y=i} \times n_{X=k}}{n}$	$n_{Y=i}$
\dots	\dots	\dots	\dots	\dots	\dots	\dots
$Y = b_l$	$\frac{n_{Y=l} \times n_{X=1}}{n}$	\dots	$\frac{n_{Y=l} \times n_{X=j}}{n}$	\dots	$\frac{n_{Y=l} \times n_{X=k}}{n}$	$n_{Y=l}$
Total	$n_{X=1}$	\dots	$n_{X=j}$	\dots	$n_{X=k}$	n

La statistique du test d'adéquation du Khi-deux, notée K_n , est définie par :

$$C_n = \sum_{j=1}^k \sum_{i=1}^l \frac{(n_{j,i} - \frac{n_{Y=i} \times n_{X=j}}{n})^2}{\frac{n_{Y=i} \times n_{X=j}}{n}}$$

La variable aléatoire C_n suit alors, sous l'hypothèse H_0 , une loi du Khi-deux à $(k-1)(l-1)$ degrés de libertés.

Exemple III.5 :

Une entreprise essaie de savoir s'il y a un lien entre le salaire et le fait d'avoir plus ou moins de 30 ans. On a les données suivantes :

Age \ Salaire	< 1500	≥ 1500	Total
moins de 30 ans	68	35	103
Plus de 30 ans	70	50	120
Total	138	85	223

En faisant les calculs on a :

Age \ Salaire	< 1500	≥ 1500	Total
moins de 30 ans	63,74	39,26	103
Plus de 30 ans	74,26	45,74	120
Total	138	85	223

Par exemple, pour la première case on fait :

$$\frac{103 \times 138}{223} = 63,74$$

ou pour la dernière case :

$$\frac{85 \times 120}{223} = 45,74$$

On peut ensuite calculer la réalisation de la statistique du test :

$$K_n(x) = 1,39$$

Ce résultat doit être comparé avec la loi du Khi-deux à $(2-1)(2-1) = 1$ degré de liberté. La région critique du test du Khi-deux est donnée par $W = \{x | K_n(x) > F_1^{-1}(1-\alpha)\}$, où F_1 est la fonction de répartition de la loi du Khi-deux à 1 degrés de libertés.

Imaginons ici que l'on souhaite faire teste de niveau $\alpha = 1\%$. Sur la table du Khi-deux on peut trouver $F_1^{-1}(1 - \alpha) = 6,64$ ce qui est bien supérieur à la réalisation de la statistique de test. Ainsi, on peut accepter sans problème H_0 et donc on peut conclure que le salaire est indépendant du fait d'avoir plus ou moins de 30 ans dans cette entreprise.

BIBLIOGRAPHIE

- Statistiques et probabilités en économie-gestion, B. Legros, Edition DUNOD
- Statistique, J.-P. Lecoutre, S. Levait-Maille et P. Tassi, Edition DUNOD
- Statistiques et probabilités en économie-gestion, C. Hurlin et V. Mignon, Edition DUNOD
- L'essentiel de la statistique pour l'économie et la gestion, J.-L. Boursin, Gualino éditeur
- QCM La statistique pour l'économie et la gestion, J.-L. Boursin, Gualino éditeur