STATISTIQUES POUR LABORATOIRES

Fabien Dumont

Fire testing laboratory
University of Liège

Quality manager, fabien.dumont@uliege.be

TABLE DES MATIERES

1 CONCEPTS STATISTIQUES

1.1 DÉFINITIONS

Dans cette section sont rappelées les définitions utiles à maîtriser dans le cadre de ce document. Ces notions sont exposées en toute généralité, sans rapport à un domaine d'application particulier.

1. Population

Totalité des individus pris en considération.

Exemple :

Si trois villages sont sélectionnés pour une enquête de démographie et de santé, la population est alors constituée des habitants de ces trois villages uniquement. Sinon, si ces trois villages sont sélectionnés aléatoirement parmi tous les villages d'une région donnée, la population est alors constituée de tous les habitants de la région.

2. Echantillon

Sous-ensemble d'une population.

3. Effectif

Nombre *n* d'individus contenus dans une population ou un échantillon considéré.

Notation :

Dans les considérations relatives aux distributions discrètes, le dénombrement de tous les individus sera indicé $i = 1,...,n$*, et le dénombrement des valeurs distinctes prises par les individus sera indicé* $i = 1,...,N$*. Par convention, il sera admis que n et N peuvent être fini ou désigner l'infini (*∞*).*

Remarque :

La notion d'effectif n'est utilisée que pour les distributions discrètes, étant donné que l'effectif est trivialement infini dans le cas des distributions continues.

4. Fréquence

Nombre n_i d'occurrence d'une valeur distincte prise par les individus dans une population ou un échantillon considéré.

Propriété :

$$\sum_{i=1}^{N} n_i = n$$

Remarques :

- *La notion de fréquence n'est utilisée que pour les distributions discrètes.*
- *Voir aussi « fonction de masse » plus bas.*

5. Variable aléatoire

Variable *X* pouvant prendre n'importe quelle valeur *x* d'un ensemble déterminé de valeurs, et à laquelle est associée une loi de probabilité (fonction de densité de probabilité pour une variable continue, fonction de masse pour une variable discrète).

Propriétés :

- *Une variable aléatoire qui ne peut prendre que des valeurs isolées est dite «discrète». Une variable aléatoire qui peut prendre toutes valeurs à l'intérieur d'un intervalle fini ou infini est dite «continue».*
- *Une fonction de variables aléatoires est elle-même une variable aléatoire.*
- *Deux variables aléatoires X et Y sont indépendantes si* $\forall A,B \quad : \quad \mathcal{P}(X \in A, Y \in B) = \mathcal{P}(X \in A).\mathcal{P}(Y \in B)$ *(*$\mathcal{P}$ *désigne la probabilité).*

6. Fonction de densité de probabilité (FDP)

Fonction $p(x)$ donnant la probabilité qu'une variable aléatoire *X* prenne une valeur donnée quelconque *x* ou appartienne à un ensemble donné de valeurs *dx*.

Propriétés :

- $p(x)dx = \mathcal{P}(x \le X < x + dx)$ *est appelée « probabilité élémentaire ». Elle donne la probabilité que la variable aléatoire X prenne une valeur comprise dans l'intervalle* $[x, x + dx[$
- $p(x)dx$ *est aux distributions continues ce que* p_i *est aux distributions discrètes.*

7. Fonction de masse

Fonction donnant, pour chaque valeur x_i d'une variable aléatoire discrète *X*, la probabilité p_i que cette variable aléatoire soit égale à x_i, soit $p_i = \mathcal{P}(X = x_i)$.

Propriétés :

- $p_i = \frac{n_i}{n}$ *: la fonction de masse est la fréquence relative.*
- p_i *est aux distributions discrètes ce que* $p(x)dx$ *est aux distributions continues.*

8. Fonction de répartition

Fonction $F(x)$ donnant pour toute valeur *x* la probabilité cumulée qu'une variable aléatoire *X* prenne une valeur comprise dans l'intervalle $]-\infty, x]$. Elle est notée $F(x)$ et est calculée comme suit :

$$F(x) = \int_{-\infty}^{x} p(x)dx \quad \text{(variable continue)}$$

$$F(x) = \sum_{i=1}^{j} p_i \text{ avec } x_j \le x < x_{j+1} \quad \text{(variable discrète)}$$

Propriétés :

- $F(x) = \mathcal{P}(X \le x)$
- $F(\infty) = \int_{-\infty}^{\infty} p(x)dx = 1$ *et* $F(\infty) = \sum_{i=1}^{N} p_i = 1$

9. Quantile (ou fractile) d'ordre α

Valeur *x* d'une variable aléatoire *X* telle que α soit la probabilité cumulée que *X* prenne une valeur comprise dans l'intervalle $]-\infty, x]$. Il s'agit donc de la valeur de *x* trouvée par la réciproque de $F(x)$ et est calculée en inversant le relation suivante :

$$\alpha = F(x) = \int_{-\infty}^{x} p(x)dx \text{ (variable continue)}$$

$$\alpha = F(x) = \sum_{i=1}^{j} p_i \text{ avec } x_j \le x < x_{j+1} \text{ (variable discrète)}$$

10. Espérance mathématique

Intégrale d'une fonction *g* d'une variable aléatoire *X*, pondérée par la mesure de probabilité de cette variable. Elle est notée $E[g(X)]$ et est calculée comme suit :

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)p(x)dx \text{ (variable continue)}$$

$$E[g(X)] = \sum_{i=1}^{N} g(x_i)p_i \text{ (variable discrète)}$$

Propriété :

En tant que moyenne pondérée des valeurs que peut prendre la fonction g, l'espérance représente intuitivement « la valeur que l'on s'attend à trouver en moyenne pour g lorsqu'on observe un grand nombre de valeurs aléatoires de X ».

11. Moment d'ordre r

Espérance mathématique de la *r*ème puissance d'une variable aléatoire *X*. Il est calculé comme suit :

$$E[X^r] = \int_{-\infty}^{\infty} x^r p(x)dx \text{ (variable continue)}$$

$$E[X^r] = \sum_{i=1}^{N} x_i^r p_i \text{ (variable discrète)}$$

12. Moyenne μ

Moment d'ordre 1 d'une variable aléatoire *X*. Elle est calculée comme suit :

$$\mu = E[X] = \int_{-\infty}^{\infty} x\, p(x)dx \text{ (variable continue)}$$

$$\mu = E[X] = \sum_{i=1}^{N} x_i p_i = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ (variable discrète)}$$

Propriétés :

- *Il s'agit de la moyenne arithmétique.*
- *Il n'existe pas toujours de moyenne pour une variable aléatoire. En particulier, les distributions à queues très étalées, comme la distribution de Cauchy, produisent des intégrales non convergentes et donc des espérances non définies.*
- *La moyenne est un paramètre de position.*
- *Terme anglais : Mean ou Average*

13. Variance σ^2

Moment d'ordre 2 d'une variable aléatoire *X* centrée (dont la moyenne a été soustraite). Elle est calculée comme suit :

$$\sigma^2 = E\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty}(x-\mu)^2\, p(x)dx \quad \text{(variable continue)}$$

$$\sigma^2 = E\left[(X-\mu)^2\right] = \sum_{i=1}^{N}(x_i-\mu)^2\, p_i = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \quad \text{(variable discrète)}$$

Propriétés :

- *Il s'agit du carré de l'écart-type.*
- *On montre que* $\sigma^2 = E\left[(X-\mu)^2\right] = E\left[(X-E(X))^2\right] = E\left[X^2\right] - E[X]^2$
- *La variance est un paramètre d'échelle.*
- *Terme anglais : Variance*

14. Ecart-type σ

Racine carrée de la variance.

Propriété :

Terme anglais : Standard deviation

15. Coefficient d'asymétrie γ_1

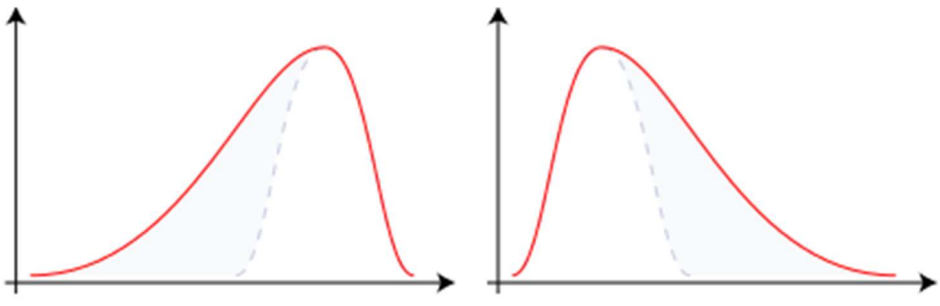
Moment d'ordre 3 d'une variable aléatoire *X* centrée (dont la moyenne a été soustraite) réduite (divisée par son écart-type). Elle est calculée comme suit :

$$\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \int_{-\infty}^{\infty}\left(\frac{x-\mu}{\sigma}\right)^3 p(x)dx \quad \text{(variable continue)}$$

$$\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^3 p_i \quad \text{(variable discrète)}$$

Propriétés :

- *Le coefficient d'asymétrie est un paramètre de forme.*
- *Terme anglais : Skewness*
- *Un coefficient positif indique une distribution décalée à gauche de la médiane, et donc une queue de distribution plus épaisse ou étalée vers la droite.*
- *Un coefficient négatif indique une distribution décalée à droite de la médiane, et donc une queue de distribution plus épaisse ou étalée vers la gauche.*
- *Un coefficient nul indique une distribution symétrique.*



Exemple de coefficient de symétrie négatif (à gauche) et positif (à droite)

16. Coefficient d'aplatissement β_2

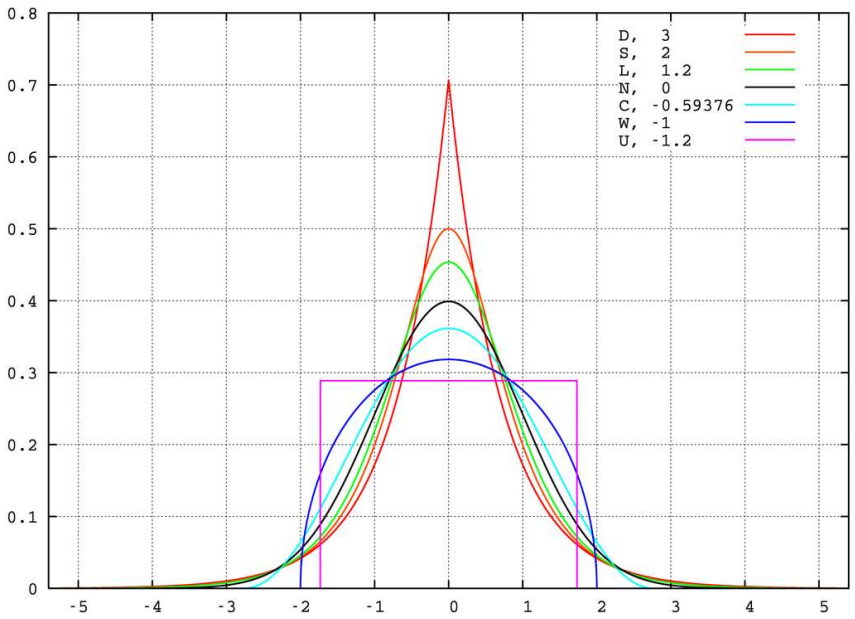
Moment d'ordre 4 d'une variable aléatoire *X* centrée (dont la moyenne a été soustraite) réduite (divisée par son écart-type). Elle est calculée comme suit :

$$\beta_2 = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \int_{-\infty}^{\infty}\left(\frac{x-\mu}{\sigma}\right)^4 p(x)dx \quad \text{(variable continue)}$$

$$\beta_2 = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^4 p_i \quad \text{(variable discrète)}$$

Propriétés :

- *Le coefficient d'aplatissement est un paramètre de forme.*
- *Terme anglais : Kurtosis*
- *Un coefficient élevé indique une distribution amincie en sa moyenne, et dont les queues de distribution sont épaisses.*
- *Un coefficient peu élevé indique une distribution élargie en sa moyenne, et dont les queues de distribution sont mince.*
- *Une distribution normale a un coefficient égal à 3.*



Exemple de coefficient d'aplatissement normalisé γ_2 pour quelques distributions
Le coefficient normalisé est simplement obtenu par $\gamma_2 = \beta_2 - 3$
(N désigne la loi normale, en noir)

17. Covariance

Moyenne du produit de deux variables aléatoires centrées dans leur loi de probabilité combinée. Elle est calculée comme suit :

$$\sigma_{XY} = E\left[(X-\mu_X)(Y-\mu_Y)\right] = E[XY] - E[X]E[Y]$$

18. Coefficient de corrélation

Moyenne du produit de deux variables aléatoires centrées réduites dans leur loi de probabilité combinée. Elle est calculée comme suit :

$$\rho_{XY} = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

1.2 ESTIMATION ET ESTIMATEUR

Pour le praticien, la connaissance exhaustive de la loi de probabilité d'une variable aléatoire est souvent inaccessible. En effet, la fonction de densité de probabilité (ou la fonction de masse) d'une variable aléatoire n'est que rarement connue : il ne s'agit généralement pas d'une donnée de base, mais bien d'une inconnue – soumise à l'observation – qu'on essaie de déterminer (par des mesures par exemple).

La connaissance que l'on peut avoir de la loi de probabilité d'une variable aléatoire sera alors caractérisée par l'observation d'un échantillon de n valeurs $x_1,...,x_i,...,x_n$ prises par cette variable. Les paramètres (moyenne, variance, …) de la loi de probabilité inconnue seront déduits de cette série d'observations et on parlera dans ce cas d'estimateur (en tant qu'estimation du paramètre).

Définitions :

- *Estimateur : valeur calculée sur un échantillon et que l'on espère être une bonne évaluation de la valeur que l'on aurait calculée sur la population totale.*
- *Estimation : opération ayant pour but, à partir des valeurs observées dans un échantillon, d'attribuer des valeurs numériques aux paramètres d'une loi prise comme modèle statistique de la population dont est issu l'échantillon.*
- *Degrés de liberté : la quantité $\upsilon = n-1$ est appelée "nombre de degrés de liberté" de la série d'observations (il s'agit plus précisément du « nombre de termes d'une somme moins le nombre de contraintes sur les termes de la somme »).*

La qualité d'un estimateur s'exprime par sa convergence, son biais et son efficacité, auxquels est parfois ajoutée la robustesse. Cela implique que, s'il est possible de définir plusieurs estimateurs pour un même paramètre (moyenne estimée, variance estimée, …), ces estimateurs seront cependant de qualités différentes : certains seront de meilleurs estimateurs que d'autres. Un bon estimateur devra au-moins être convergent et sans biais.

Propriétés :

- *Un estimateur est convergent s'il tend vers la valeur réelle du paramètre lorsque l'effectif observé augmente.*
- *Un estimateur est sans biais si son espérance est égale à la valeur réelle du paramètre.*
- *Si deux estimateurs sont convergents et sans biais, le plus efficace est celui qui a la variance la plus faible.*
- *La robustesse d'un estimateur est sa capacité à ne pas être modifié par une petite modification dans les données ou dans les paramètres du modèle choisi pour l'estimation.*

En pratique, on ne retiendra que le meilleur estimateur de chaque paramètre, dont les qualités ont été mathématiquement prouvées. Ce sont ceux qui sont présentés ci-dessous.

Dans toute application, les estimateurs doivent être utilisés en lieu et place des valeurs vraies des paramètres (formules vues plus haut) si celles-ci sont inconnues. Ces dernières ne doivent être utilisées que si la loi de probabilité théorique est connue (ou, ce qui revient au même, les valeurs prises par la totalité des individus de la population).

19. Moyenne estimée $\bar{x}$

Les valeurs $x_1,...,x_i,...,x_n$ désignant un échantillon de *n* observations indépendantes de la variable aléatoire *X*, le meilleur estimateur adopté pour la moyenne μ est donné par :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Propriété :

Il s'agit de la moyenne arithmétique.

Excel :

MOYENNE(plage de données)

20. Variance estimée s^2

Les valeurs $x_1,...,x_i,...,x_n$ désignant un échantillon de *n* observations indépendantes de la variable aléatoire *X*, le meilleur estimateur adopté pour la variance σ^2 est donné par :

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \quad \textit{si la moyenne } \mu \textit{ est connue}$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \textit{si la moyenne } \mu \textit{ est inconnue}$$

Remarques :

- *Il est important de souligner que le dénominateur de la variance estimée est n-1 lorsque la moyenne μ est inconnue, alors qu'il est n lorsque la moyenne μ est connue. Seules ces définitions fournissent un estimateur sans biais.*
- *Lorsque la moyenne μ est inconnue, on en déduit notamment que s^2 est toujours plus grand que σ^2. Cela exprime d'une certaine façon qu'il est moins précis. Cela provient du fait que l'estimation de la variance implique l'estimation d'un paramètre en plus, la moyenne estimée de X, ce qui induit une incertitude de plus. C'est cette contrainte qui fait apparaître le nombre de degré de liberté $\upsilon = n-1$ au dénominateur.*

Excel :

VAR.S(plage de données) donne $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ *(S pour « sample »)*

VAR.P.N(plage de données) donne $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ *(P pour « population »)*

21. Ecart-type estimé *s*

Les valeurs $x_1,...,x_i,...,x_n$ désignant un échantillon de *n* observations indépendantes de la variable aléatoire *X*, le meilleur estimateur adopté pour l'écart-type σ est donné par :

$$s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \quad \textit{si la moyenne } \mu \textit{ est connue}$$

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \textit{si la moyenne } \mu \textit{ est inconnue}$$

Excel :

STDEVA(plage de données) donne $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}$

STDEVPA(plage de données) donne $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2}$

22. Covariance estimée s_{XY}

Les valeurs $x_1,...,x_i,...,x_n$ et $y_1,...,y_i,...,y_n$ désignant des échantillons de *n* observations indépendantes des variables aléatoires *X* et *Y*, le meilleur estimateur adopté pour la covariance s_{XY} est donné par :

$$s_{XY}=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu_X\right)\left(y_i-\mu_Y\right) \quad \textit{si les moyennes } \mu \textit{ sont connues}$$

$$s_{XY}=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right) \quad \textit{si les moyennes } \mu \textit{ sont inconnues}$$

Remarques :

Tout comme pour la variance, il est important de souligner que le dénominateur de la covariance estimée est n-1 lorsque les moyennes μ sont inconnues, alors qu'il est n lorsque les moyennes μ sont connues. Seules ces définitions fournissent un estimateur sans biais.

Excel :

COVARIANCE.STANDARD(plage de données X ; plage de données Y) donne $s_{XY}=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)$

COVARIANCE.PEARSON(plage de données X ; plage de données Y) donne $s_{XY}=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu_X\right)\left(y_i-\mu_Y\right)$

23. Coefficient de corrélation estimée r_{XY}

Les valeurs $x_1,...,x_i,...,x_n$ et $y_1,...,y_i,...,y_n$ désignant des échantillons de *n* observations indépendantes des variables aléatoires *X* et *Y*, le meilleur estimateur adopté pour le coefficient de corrélation r_{XY} est donné par :

$$r_{XY}=\frac{\sum_{i=1}^{n}\left(x_i-\mu_X\right)\left(y_i-\mu_Y\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\mu_X\right)^2\sum_{i=1}^{n}\left(y_i-\mu_Y\right)^2}} \quad \textit{si les moyennes } \mu \textit{ sont connues}$$

$$r_{XY}=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}} \quad \textit{si les moyennes } \mu \textit{ sont inconnues}$$

Excel :

COEFFICIENT.CORRELATION(plage de données X ; plage de données Y) donne

$$r_{XY} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$

24. Paramètres de la moyenne estimée $\bar{x}$ d'un échantillon

Lorsqu'on observe des échantillons de *n* observations indépendantes $x_1,\ldots,x_i,\ldots,x_n$ d'une variable aléatoire, la moyenne estimée $\bar{x}$ varie selon l'échantillon observé et est donc elle aussi une variable aléatoire possédant une moyenne, une variance et un écart-type.

Propriétés :

- *L'espérance de la moyenne estimée est* $E\left[\bar{x}\right] = \mu$.
- *Le meilleur estimateur adopté pour la variance de la moyenne estimée est donné par :*

  $$s^2\left(\bar{x}\right) = \frac{s_X^2}{n}$$

- *Le meilleur estimateur adopté pour l'écart-type de la moyenne estimée est donné par :*

  $$s\left(\bar{x}\right) = \frac{s_X}{\sqrt{n}}$$

- *Cela signifie que la moyenne de n variables aléatoires fluctue moins qu'une seule de ces variables aléatoires.*

REFERENCES

[1] ISO 3534-1:2006 — Statistique – Vocabulaire et symboles – Partie 1 : Termes statistiques généraux et termes utilisés en calcul des probabilités

[2] JSGM 100:2008 (GUM) — Évaluation des données de mesure – Guide pour l'expression de l'incertitude de mesure

[3] Wikipedia

2 STATISTIQUES POUR LABORATOIRES

Dans cette section sont exposés les principaux outils statistiques utiles à maîtriser en laboratoire. Ces notions sont exposées en toute généralité, sans rapport à un domaine d'application particulier. L'intérêt sera cependant principalement porté sur les variables aléatoires continues car elles représentent l'essentiel des grandeurs rencontrées en mesurage (des variables aléatoires discrètes pourraient intervenir dans de rares cas mais ne feront pas l'objet de ce document).

2.1 LOIS DE PROBABILITÉ

2.1.1 Loi normale

La loi normale (ou loi gaussienne) de paramètres μ, σ^2 (avec $\sigma > 0$) est la distribution de probabilité définie par la fonction de densité de probabilité :

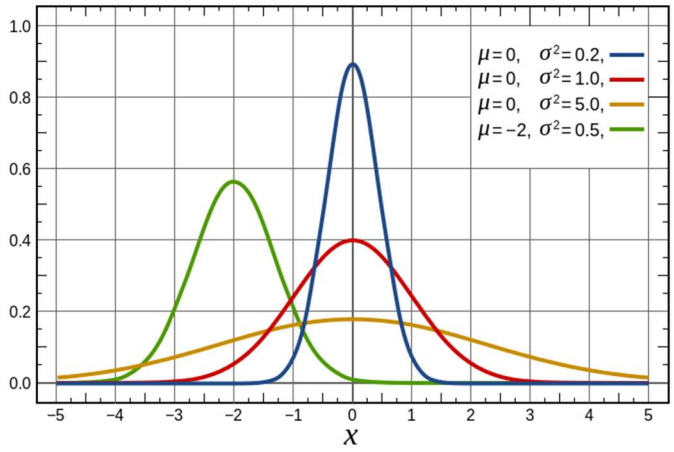
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Propriétés :

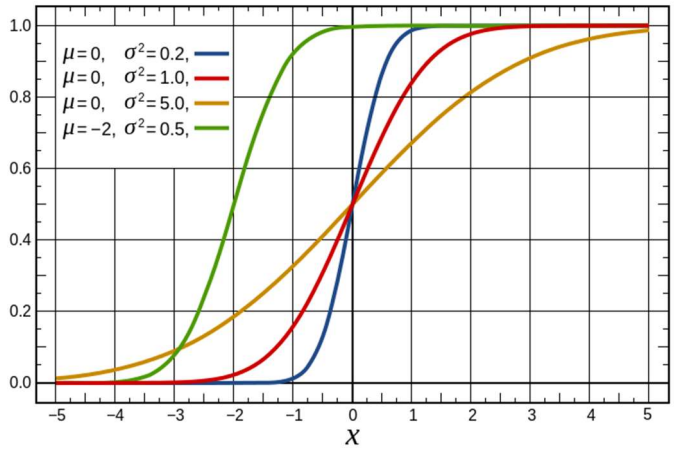
- *Support de définition :* $x \in]-\infty, +\infty[$
- $p(\mu + a) = p(\mu - a)$ *: la loi normale est symétrique autour de* μ
- $Moyenne = E[X] = \mu$
- $Variance = E\left[(X-\mu)^2\right] = \sigma^2$ *et donc* $Ecart-type = \sigma$

Notation :

$$N(\mu, \sigma^2)$$



Fonction de densité de probabilité



Fonction de répartition

Excel :

LOI.NORMALE.N(x, μ, σ, false) donne la FDP $p(x)$ *de la loi normale*

LOI.NORMALE.N(x, μ, σ, true) donne la fonction de répartition $F(x) = \int_{-\infty}^{x} p(x)dx = \mathcal{P}(X \le x)$

LOI.NORMALE.INVERSE.N($F(x)$, μ, σ) donne la valeur de x, appelée quantile

La loi normale est d'une importance primordiale dans la représentation de nombreux phénomènes. Il existe une raison bien établie de cette omniprésence de la loi normale, souvent ignorée par ceux qui l'utilisent, qui découle du théorème central limite.

2.1.2 Loi rectangulaire

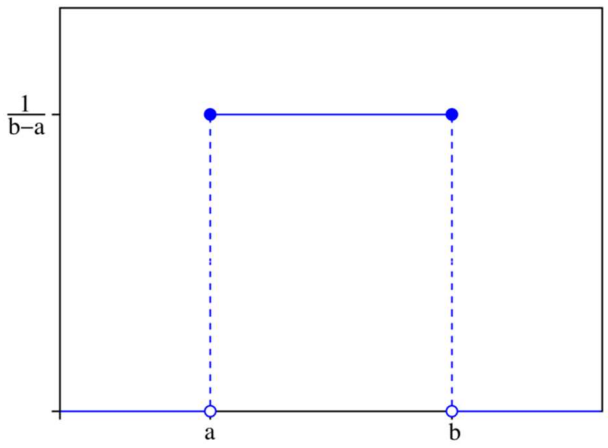
La loi rectangulaire (ou loi uniforme continue) de paramètres *a*, b (avec $a < b$) est la distribution de probabilité définie par la fonction de densité de probabilité :

$$p(x) = \begin{cases} 0 & pour\ x \notin [a,b] \\ \dfrac{1}{b-a} & pour\ x \in [a,b] \end{cases}$$

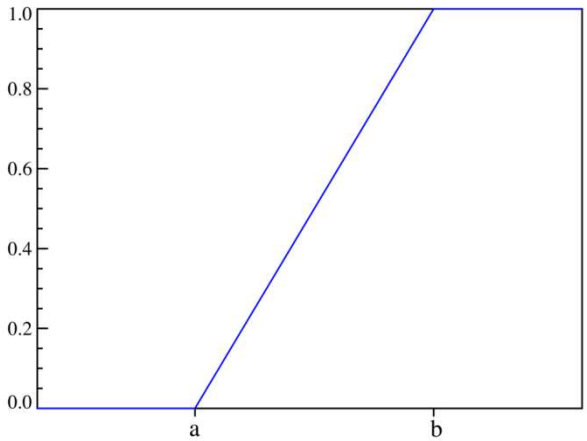
On désignera par *2d* (avec $d > 0$) la différence entre les valeurs limites, de sorte que $d = \dfrac{b-a}{2}$ représente la demi-largeur de la distribution.

Propriétés :

- *Support de définition :* $x \in [a,b]$
- $p\left(\dfrac{a+b}{2}+c\right) = p\left(\dfrac{a+b}{2}-c\right)$ *: la loi rectangulaire est symétrique autour de* $\dfrac{a+b}{2}$
- $Moyenne = E[X] = \dfrac{a+b}{2}$
- $Variance = E\left[(X-\mu)^2\right] = \dfrac{d^2}{3}$ *et donc* $Ecart-type = \dfrac{d}{\sqrt{3}}$



Fonction de densité de probabilité



Fonction de répartition

Cette loi de distribution caractérise une variable aléatoire dont les valeurs sont équiprobables sur l'intervalle $[a,b]$.

2.1.3 Loi de Student

La loi de Student (ou loi t) à *k* degrés de liberté (où *k* est un entier positif) est la distribution de probabilité définie par la fonction de densité de probabilité :

$$p(x)=\frac{1}{\sqrt{k\pi}}\frac{\Gamma\left((k+1)/2\right)}{\Gamma\left(k/2\right)}\frac{1}{\left(1+\left(x^2/k\right)\right)^{(k+1)/2}}$$

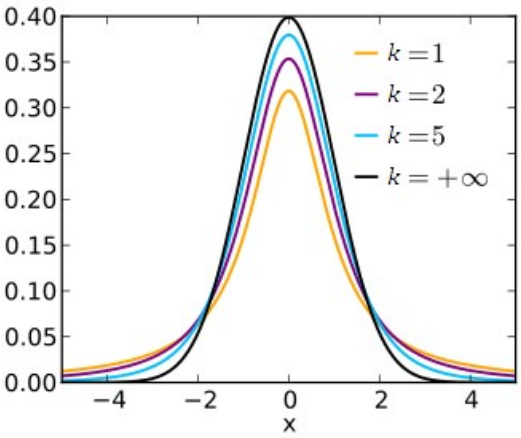
où Γ est la fonction Gamma d'Euler

Propriétés :

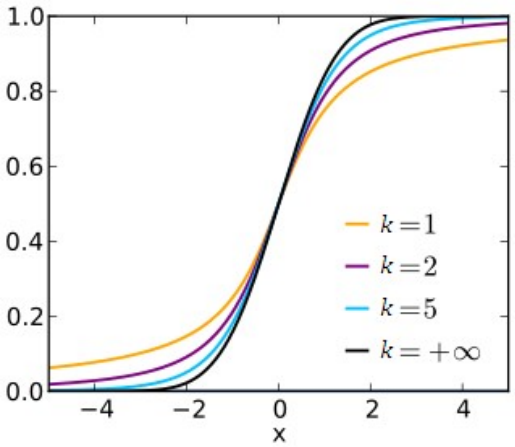
- *Support de définition :* $x\in\left]-\infty,+\infty\right[$
- $p(a)=p(-a)$ *: la loi de Student est symétrique autour de 0*
- $Moyenne=E[X]=0$ *pour* $k>1$ *et indéfinie pour* $k=1$
- $Variance=E\left[(X-\mu)^2\right]=\frac{k}{k-2}$ *pour* $k>2$*, infinie pour* $k=2$ *et indéfinie pour* $k=1$*, et donc* $Ecart-type=\sqrt{\frac{k}{k-2}}$

Notation :

$T(k)$



Fonction de densité de probabilité



Fonction de répartition

Remarques :

- $\lim_{k\to\infty}T(k)=N(0,1)$ *: la loi de Student tend vers la loi normale centrée réduite lorsque* $k\to\infty$
- *La densité de probabilité de la moyenne de la loi de Student est inférieure à la densité de probabilité de la moyenne de la loi normale centrée réduite et tend vers cette dernière lorsque* $k\to\infty$
- *La variance de la loi de Student est supérieure à la variance de la loi normale centrée réduite et tend vers cette dernière lorsque* $k\to\infty$

En d'autres termes, la distribution-t est une version plus écrasée et plus large que la distribution normale centrée réduite.

Excel :

LOI.STUDENT.N(x, k, false) donne la FDP $p(x)$ *de la loi de Student*

LOI.STUDENT.N(x, k, true) donne la fonction de répartition $F(x)=\int_{-\infty}^{x}p(x)dx=\mathcal{P}(X\le x)$

LOI.STUDENT.INVERSE.N($F(x)$*, k) donne la valeur de x, appelée quantile*

La loi de Student a été formulée par William S. Gosset aux alentours de 1910. Gosset travaillait comme statisticien pour la brasserie Guinness en Angleterre. Gosset publia tous ses travaux de statistique sous le pseudonyme de Student. Ainsi on appelle loi de Student la loi de probabilité qui aurait dû être appelée la loi de Gosset.

2.1.4 Loi du khi-deux

La loi du khi-deux (ou loi du khi-carré) à k degrés de liberté (où k est un entier positif) est la distribution de probabilité définie par la densité de probabilité :

$$p(x)=\begin{cases}\dfrac{1}{2^{k/2}\Gamma(k/2)}X^{(k/2)-1}e^{-x/2} & pour\ x\geq 0\\ 0 & pour\ x<0\end{cases}$$

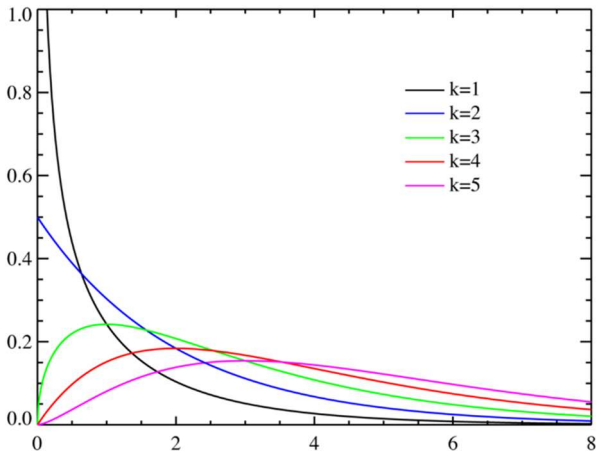
où Γ est la fonction Gamma d'Euler

Propriétés :

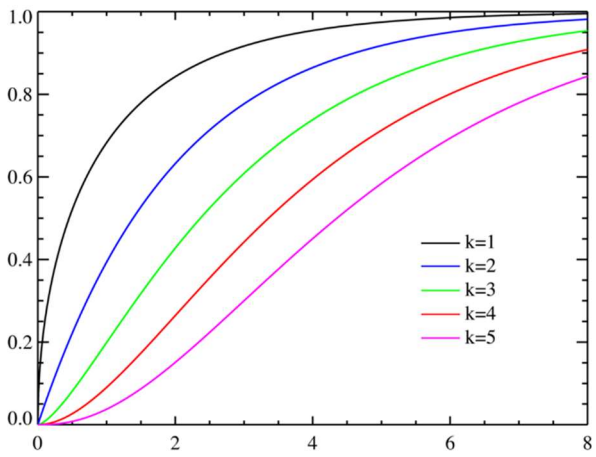
- *Support de définition :* $x\in[0,+\infty[$
- *La loi du khi-deux est asymétrique*
- $Moyenne=E[X]=k$
- $Variance=E\left[(X-\mu)^2\right]=2k$ *et donc* $Ecart-type=\sqrt{2k}$

Notation :

$\chi^2(k)$



Fonction de densité de probabilité



Fonction de répartition

Propriété remarquable :

$\mathcal{P}(x\leq E[X])>\mathcal{P}(E[X]\leq x)$, *en d'autres termes les probabilités cumulées à gauche et à droite de la moyenne ne sont pas égales (à 50%), elle est toujours plus grande à gauche qu'à droite, et on montre que* $\lim_{k\to\infty}\mathcal{P}(x\leq E[X])=\lim_{k\to\infty}\mathcal{P}(E[X]\leq x)=50\%$

Excel :

LOI.KHIDEUX.N(x, k, false) donne la FDP $p(x)$ *de la loi du khi-deux*

LOI.KHIDEUX.N(x, k, true) donne la fonction de répartition $F(x)=\int_0^x p(x)dx=\mathcal{P}(X\leq x)$

LOI.KHIDEUX.INVERSE($F(x)$*, k) donne la valeur de x, appelée quantile*

2.1.5 Loi de Fisher

La loi de Fisher (ou loi de Fisher-Snedecor, ou loi F) à k et l degrés de liberté (où k et l sont des entiers positifs) est la distribution de probabilité définie par la densité de probabilité :

$$p(x)=\begin{cases}\dfrac{\Gamma\left((k+l)/2\right)(k/l)^{k/2}}{\Gamma(k/2)\Gamma(l/2)}\dfrac{x^{(k/2)-1}}{(1+kx/l)^{(k+l)/2}} & pour\ x\geq 0\\ 0 & pour\ x<0\end{cases}$$

où Γ est la fonction Gamma d'Euler

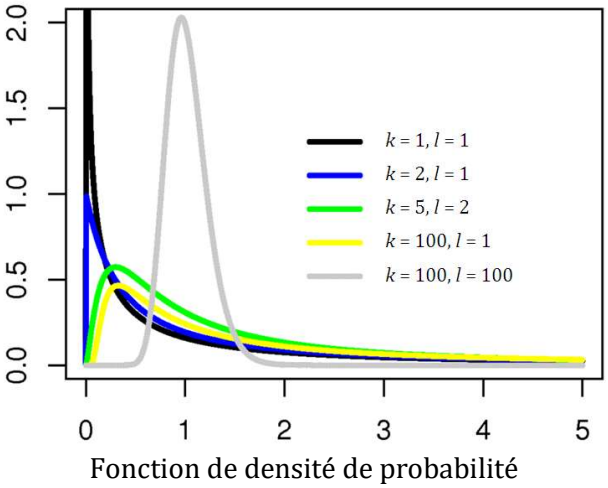
Propriétés :

- *Support de définition :* $x\in[0,+\infty[$
- *La loi de Fisher est asymétrique*
- $Moyenne=E[X]=\dfrac{l}{l-2}$ *pour* $l>2$ *et indéfinie pour* $k\leq 2$
- $Variance=E\left[(X-\mu)^2\right]=\dfrac{2l^2(k+l-2)}{k(l-2)^2(l-4)}$ *pour pour* $l>4$ *et indéfinie pour* $k\leq 4$, *et donc*

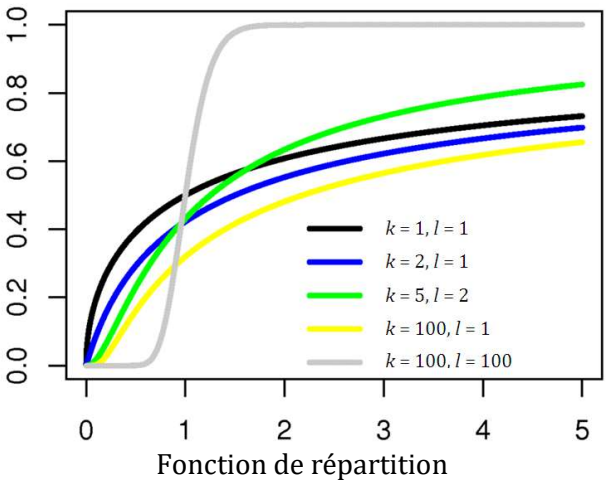
$$Ecart-type=\sqrt{\frac{2l^2(k+l-2)}{k(l-2)^2(l-4)}}$$

Notation :

$F(k,l)$



Fonction de densité de probabilité



Fonction de répartition

Propriétés remarquables :

- *On montre que* $F_\alpha(k,l)=\dfrac{1}{F_{1-\alpha}(l,k)}$ *où* $F_\alpha(k,l)$ *désigne le quantile x d'ordre* α *de la loi de Fisher avec k et l degrés de liberté, et donc* $\alpha=F(x)=\int_0^x p(x)dx=\mathcal{P}(X\leq x)$
- $\mathcal{P}(x\leq E[X])>\mathcal{P}(E[X]\leq x)$, *en d'autres termes les probabilités cumulées à gauche et à droite de la moyenne ne sont pas égales (à 50%), elle est toujours plus grande à gauche qu'à droite, et on montre que* $\lim_{k,l\to\infty}\mathcal{P}(x\leq E[X])=\lim_{k,l\to\infty}\mathcal{P}(E[X]\leq x)=50\%$

Excel :

LOI.F.N(x, k, l, false) donne la FDP $p(x)$ *de la loi de Fisher*

LOI.F.N(x, k, l, true) donne la fonction de répartition $F(x)=\int_0^x p(x)dx=\mathcal{P}(X\le x)$

INVERSE.LOI.F.N($F(x)$*, k, l) donne la valeur de x, appelée quantile*

2.1.6 Loi log-normale

La loi log-normale de paramètres a, b^2 (avec $b>0$) est la distribution de probabilité définie par la fonction de densité de probabilité :

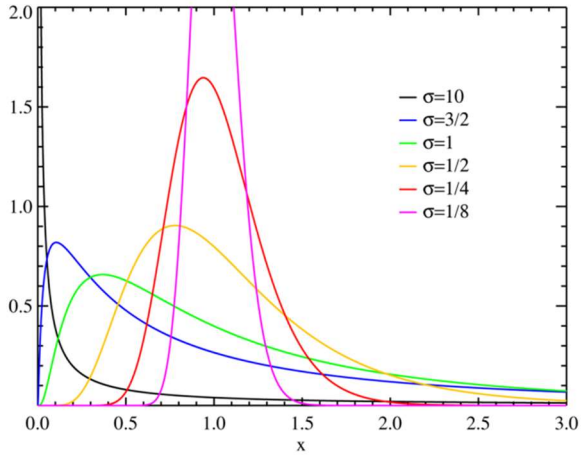
$$p(x)=\frac{1}{xb\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\ln x-a}{b}\right)^2}$$

Propriétés :

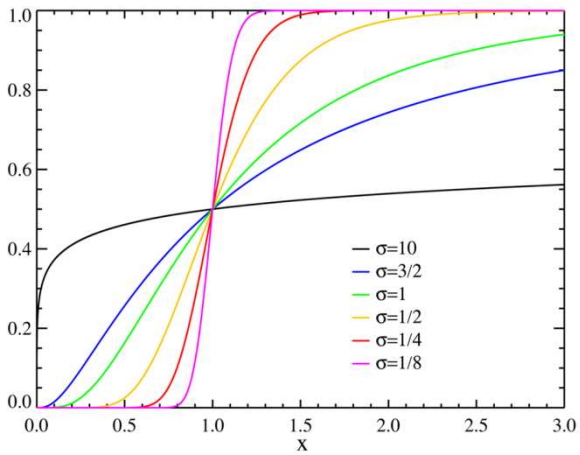
- *Support de définition :* $x\in[0,+\infty[$
- *La loi log-normale est asymétrique*
- $Moyenne=E[X]=e^{a+\frac{b^2}{2}}$
- $Variance=E\left[(X-\mu)^2\right]=\left(e^{b^2}-1\right)e^{2a+b^2}$ *et donc* $Ecart-type=\sqrt{\left(e^{b^2}-1\right)e^{2a+b^2}}$

Notation :

$lnN(a,b^2)$



Fonction de densité de probabilité
(σ désigne b)



Fonction de répartition
(σ désigne b)

Excel :

LOI.LOGNORMALE.N(x, a, b, false) donne la FDP $p(x)$ *de la loi log-normale*

LOI.LOGNORMALE.N(x, a, b, true) donne la fonction de répartition $F(x)=\int_0^x p(x)dx=\mathcal{P}(X\le x)$

LOI.LOGNORMALE.INVERSE.N($F(x)$*, a, b) donne la valeur de x, appelée quantile*

Interprétation :

Une variable aléatoire X est dite suivre une loi log-normale de paramètres a et b^2 si la variable $Y = \ln X$ suit une loi normale de paramètres a et b^2.

La loi log-normale apparaît dans la représentation de certains phénomènes naturels (voir théorème central limite appliqué au produit de variables aléatoires).

2.1.7 Loi Bêta

La loi Bêta de paramètres a, b (avec $a>0$ et $b>0$) est la distribution de probabilité définie par la fonction de densité de probabilité :

$$p(x)=\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$$

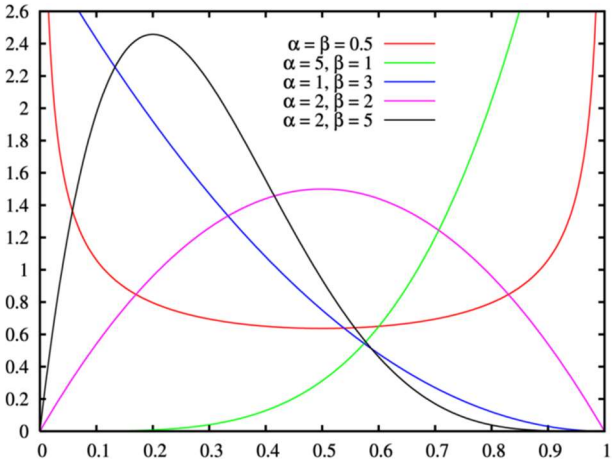
où $B(a,b)$ désigna le fonction Bêta.

Propriétés :

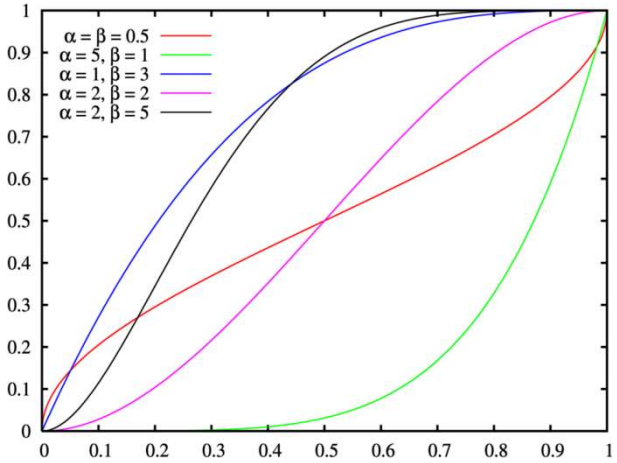
- *Support de définition :* $x \in [0,1]$
- *La loi Bêta est asymétrique*
- $Moyenne = E[X] = \frac{a}{a+b}$
- $Variance = E\left[(X-\mu)^2\right] = \frac{ab}{(a+b)^2(a+b+1)}$ *et donc* $Ecart-type = \sqrt{\frac{ab}{(a+b)^2(a+b+1)}}$

Notation :

$B(a,b)$



Fonction de densité de probabilité
(α et β désignent a et b)



Fonction de répartition
(α et β désignent a et b)

Propriétés remarquables :

- *On montre que :*
  - $p(x,a,b) = p(1-x,b,a)$
  - $E[B(a,b)] = 1 - E[B(b,a)]$
  - $Variance[B(a,b)] = Variance[B(b,a)]$

- *La distribution Bêta peut être exprimée sur un autre support de définition que* $x \in [0,1]$. *En toute généralité, soit alors le support* $x \in [x_{min}, x_{max}]$. *Le changement de variable* $y = \frac{x - x_{min}}{x_{max} - x_{min}}$ *ramène la distribution au cas connu du support* $y \in [0,1]$. *On montre notamment que :*
  - $E[Y] = \frac{E[X] - x_{min}}{x_{max} - x_{min}}$
  - $Variance[Y] = \frac{Variance[X]}{(x_{max} - x_{min})^2}$

Excel :

LOI.BETA.N(x, a, b, false, x_{min}, x_{max}) donne la FDP $p(x)$ *de la loi log-normale*

LOI.BETA.N(x, a, b, true, x_{min}, x_{max}) donne la fonction de répartition $F(x) = \int_{x_{min}}^{x} p(x)dx = \mathcal{P}(X \le x)$

BETA.INVERSE.N($F(x)$*, a, b, x_{min}, x_{max}) donne la valeur de x, appelée quantile*

La loi Bêta admet une grande variété de formes selon les valeurs prises par ses paramètres *a* et *b*. Elle permet ce faisant de modéliser de nombreux phénomènes à support fini. Elle est particulièrement bien adaptée pour la modélisation des comportements aléatoires exprimés en pourcentages ou proportions.

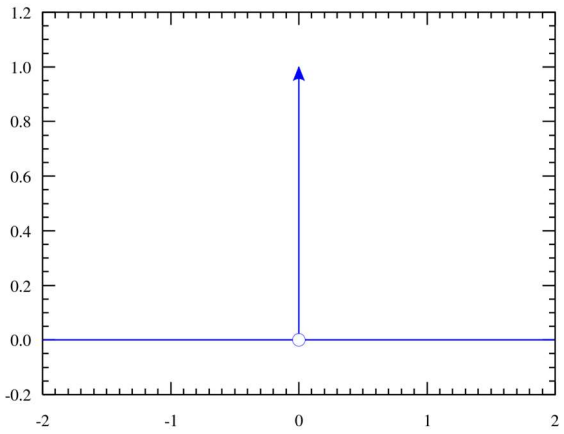
2.1.8 Distribution de Dirac

La distribution de Dirac (ou fonction δ) est la distribution de probabilité définie par la densité de probabilité :

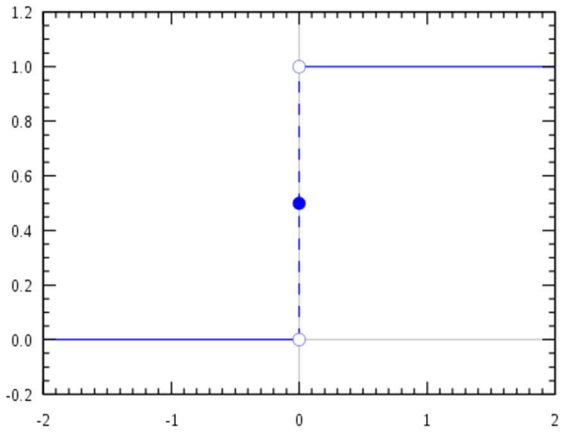
$$\delta(x) = \begin{cases} 0 & pour\ x \neq 0 \\ \infty & pour\ x = 0 \end{cases}$$

Propriétés :

- $p(a) = p(-a)$ *: la distribution de Dirac est symétrique autour de 0*
- $Moyenne = E[X] = 0$
- $Variance = E\left[(X - \mu)^2\right] = 0$ *et donc* $Ecart - type = 0$



Fonction de densité de probabilité
(La flèche indique une valeur infinie)



Fonction de répartition
(il s'agit de la fonction $H(x)$ de Heaviside, aussi appelée fonction échelon ou step function)

On parle souvent par abus de langage de « fonction de Dirac ». La « fonction de Dirac » n'est pas une fonction au sens analytique, mais bien une « fonction généralisée ».

Propriétés remarquables :

- *La distribution de Dirac est la limite de la loi normale lorsque* $\mu = 0$ *et* $\sigma \to \infty$
- $\int_A \delta(x)dx = \begin{cases} 1 & si\ 0 \in A \\ 0 & si\ 0 \notin A \end{cases}$ *et donc* $\int_A \delta(x-a)dx = \begin{cases} 1 & si\ a \in A \\ 0 & si\ a \notin A \end{cases}$
- $\int_A \delta(x)f(x)dx = \begin{cases} f(0) & si\ 0 \in A \\ 0 & si\ 0 \notin A \end{cases}$ *et donc* $\int_A \delta(x-a)f(x)dx = \begin{cases} f(a) & si\ a \in A \\ 0 & si\ a \notin A \end{cases}$
- $(f * \delta)(x) = f(x)$ *: la fonction* δ *est l'élément neutre du produit de convolution*

La FDP de la distribution de Dirac représente un phénomène n'ayant « aucune probabilité de prendre une autre valeur que la valeur nulle », c'est-à-dire un phénomène prenant la valeur nulle à coup sûr. En d'autres termes, la distribution de Dirac représente adéquatement une constante, et non pas une variable aléatoire comme c'est le cas pour les autres distributions vue ci-dessus (en ce sens, une constante pourrait être considérée – du point de vue des distributions – comme la dégénérescence d'une variable aléatoire).
Plus généralement, une constante a pourra être modélisée par la distribution $\delta(x-a)$. Si cette représentation peut sembler atypique, elle offre cependant une formulation mathématique utile pour l'application systématique des lois de propagation.

2.1.9 Autres lois

Il existe d'autres lois de probabilité continues utilisées plus rarement : loi gamma, loi exponentielle, loi de Weibull, … Pour leurs définitions et propriétés, ainsi que pour la généralisation de la loi normale à plusieurs variables, ou encore pour les définitions et propriétés des lois discrètes (loi de Poisson, loi hypergéométrique, …), se reporter à l'ISO 3534-1:2006.

2.1.10 Calage des paramètres

Comme expliqué plus haut, dans la pratique, la fonction de densité de probabilité d'une variable aléatoire n'est que rarement connue à priori : il ne s'agit généralement pas d'une donnée de base, mais bien d'une inconnue. On cherche alors le plus souvent à modéliser le phénomène étudié – dont on dispose d'un échantillon de valeurs observées (obtenues par des mesures par exemple) – par une loi de probabilité de forme usuelle, telles celles présentées ci-dessus.

Ainsi, dans la pratique, il est surtout intéressant d'estimer les paramètres a, b, k, l, … des lois usuelles à partir d'un échantillon de n valeurs $x_1,\ldots,x_i,\ldots,x_n$ prises par la variable observée. Plus précisément, ces paramètres seront déduits par comparaison des moyenne estimée $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et variance estimée $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ aux expressions exactes des moyennes et variances des lois usuelles. L'inversion de ces relations permet enfin de déterminer les valeurs des paramètres.

L'inversion de ces relations est élémentaire pour la plupart des lois de probabilité présentées ci-dessus. Les cas suivant étant moins évident, nous en donnons l'inversion :

- loi log-normale :

$$a = ln\left(\frac{\bar{x}}{\sqrt{1+\frac{s_X^2}{\bar{x}^2}}}\right) \text{ et } b = \sqrt{ln\left(1+\frac{s_X^2}{\bar{x}^2}\right)}$$

- loi Bêta :

$$a = \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{s_X^2}-1\right) \text{ et } b = (1-\bar{x})\left(\frac{\bar{x}(1-\bar{x})}{s_X^2}-1\right)$$

2.2 INTERVALLE DE CONFIANCE, NIVEAU DE CONFIANCE ET FACTEUR D'ÉLARGISSEMENT

La probabilité qu'une variable aléatoire *X*, de FDP $p(x)$ et de fonction de répartition $F(x)$, prenne une valeur dans l'intervalle $[\mu-\delta_-, \mu+\delta_+]$ donné autour de sa moyenne est :

$$\mathcal{P}(\mu-\delta_- \le X \le \mu+\delta_+) = \int_{\mu-\delta_-}^{\mu+\delta_+} p(x)dx = F(\mu+\delta_+) - F(\mu-\delta_-)$$

Très souvent, l'information utile se trouve dans la résolution du problème inverse : on cherche l'intervalle autour de sa moyenne dans lequel la variable aléatoire *X* a une probabilité $\mathcal{P}(\mu-\delta_- \le X \le \mu+\delta_+) = 1-\alpha$ de prendre une valeur, avec $\alpha \in [0,1]$.

Définitions :

- □ *La probabilité $1-\alpha$ exprimée en % est appelée « niveau de confiance »*
- □ *L'intervalle $[\mu-\delta_-, \mu+\delta_+]$ est appelé « intervalle de confiance »*

Les bornes δ_-, δ_+ sont alors déduites par l'inversion de la fonction de répartition :

$$\mathcal{P}(X \le \mu-\delta_-) = F(\mu-\delta_-) = \alpha/2$$
$$\mathcal{P}(X \le \mu+\delta_+) = F(\mu+\delta_+) = 1-\alpha/2$$

Excel :

L'inversion des fonctions de répartition est donnée au § 2.1

Par convention, l'écart-type σ incarnant par excellence la dispersion d'une loi de probabilité, les bornes δ_-, δ_+ de cet intervalle sont exprimées sous la forme de multiples de l'écart-type σ de la distribution de *X*. En outre, généralement, les distributions considérées sont symétriques autour de leur moyenne (lois normale, rectangulaire, de Student, …), de sorte que l'intervalle de confiance s'exprime alors sous la forme suivante :

$$\mathcal{P}(\mu-k\sigma \le X \le \mu+k\sigma) = 1-\alpha$$

Définition :

Le paramètre k est appelé « facteur d'élargissement (au niveau de confiance $1-\alpha$) ».

Valeurs utiles :

- □ *Loi normale*
  - *à k = 1 : $\mathcal{P}(\mu-k\sigma \le X \le \mu+k\sigma) = 68,27\,\%$*
  - *à k = 2 : $\mathcal{P}(\mu-2\sigma \le X \le \mu+2\sigma) = 95,45\,\%$*
  - *à k = 3 : $\mathcal{P}(\mu-3\sigma \le X \le \mu+3\sigma) = 99,73\,\%$*
- □ *Loi rectangulaire :*

à $k=(1-\alpha)\sqrt{3}$ *:* $\mathcal{P}(\mu-k\sigma\le X\le\mu+k\sigma)=1-\alpha$*, et en particulier*
à k = 1,65 : $\mathcal{P}(\mu-1{,}65\sigma\le X\le\mu+1{,}65\sigma)=95{,}26\,\%$

- *Loi de Student :*
  à $k=k(\nu)$ *:* $\mathcal{P}(\mu-k\sigma\le X\le\mu+k\sigma)=95{,}45\,\%$
  où ν *sont les degrés de liberté et* $k=k(\nu)$ *est donné par :*

ν	1	2	3	4	5	6	7	8	10	20	50	∞
k	13,97	4,53	3,31	2,87	2,65	2,52	2,43	2,37	2,28	2,13	2,05	2,00

- *Loi trapézoïdale :*
  Pour une loi trapézoïdale résultant de la convolution (voir la première loi de propagation des FDP) de deux distributions rectangulaires de demi-largeurs respectives d_1 *et* d_2*, on a :*

$$\text{à } k=\frac{1}{\sqrt{\frac{1+\beta^2}{6}}}\begin{cases}\frac{(1-\alpha)(1+\beta)}{2} & \text{pour } \frac{1-\alpha}{1+\alpha}<\beta\\ 1-\sqrt{\alpha(1-\beta^2)} & \text{pour } \beta\le\frac{1-\alpha}{1+\alpha}\end{cases}: \mathcal{P}(\mu-k\sigma\le X\le\mu+k\sigma)=1-\alpha$$

où $\beta=\frac{|d_1-d_2|}{d_1+d_2}$ *est appelé « paramètre de bord ».*

(EA 4/02, S10.13 Mathematical note)

On constate notamment que, pour une loi normale ou rectangulaire, le facteur d'élargissement *k* ne dépend que du niveau de confiance donné.

Remarques :

- *Un intervalle de confiance associé à un niveau de confiance est une mesure de la fiabilité que l'on peut accorder à une estimation. Le risque d'erreur est donné par* α *et le degré de certitude par* $1-\alpha$*.*
- *Dans de nombreuses applications pratiques (tests statistiques, estimation d'incertitudes, …), le niveau de confiance choisi sera* $1-\alpha=95\,\%$ *et donc* $\alpha=5\,\%$*. Cette valeur de référence fait l'objet d'un consensus à l'échelle internationale.*

2.3 LOIS DE PROPAGATION

Soit *n* variables aléatoires $X_1,\ldots,X_n$, entièrement définies par leur FDP, et la grandeur de sortie *Y* donnée – en toute généralité – par une relation fonctionnelle :

$$Y=f(X_1,\ldots,X_n)$$

Une fonction de variables aléatoires étant elle-même une variable aléatoire, la grandeur de sortie *Y* est donc une variable aléatoire, dont on cherche la FDP.

La particularité est que ces grandeurs sont chacune définies non pas par une valeur algébrique unique mais bien par une fonction (leur FDP). En d'autres termes, la résolution d'un tel problème exige de travailler dans un espace fonctionnel et non pas algébrique. Les réflexes du calcul algébrique traditionnel – où on réalise des opérations sur des valeurs – doivent être oubliés et remplacés par les méthodes du calcul fonctionnel – où on réalise des opérations sur des fonctions (les FDP en l'occurrence).

Remarque importante :

Il faut bien comprendre ce que cela implique :

- *la FDP d'une fonction f de variables aléatoires n'est pas la fonction f des FDP de ces variables aléatoires ;*

- *la moyenne d'une fonction f de variables aléatoires n'est pas la fonction f des moyennes de ces variables aléatoires (sans dans le cas particulier ou f est linéaire) ;*
- *la variance d'une fonction f de variables aléatoires n'est pas la fonction f des variances de ces variables aléatoires ;*
- *…*

Par exemple, la somme de deux variables aléatoires $Y = X_1 + X_2$, de FDP respectives $f_1(X)$ et $f_2(X)$, donnera une variable aléatoire Y dont la FDP ne sera pas simplement $f_1(X) + f_2(X)$.

2.3.1 Lois de propagation des FDP

Définition :

Le produit de convolution $h(s)$ de deux fonctions $f(x)$ et $g(x)$ est défini par :

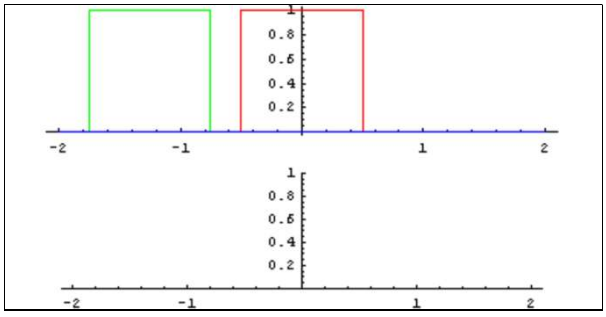
$$h(s) = (f * g)(s) = \int_{-\infty}^{\infty} f(s-x)g(x)dx$$

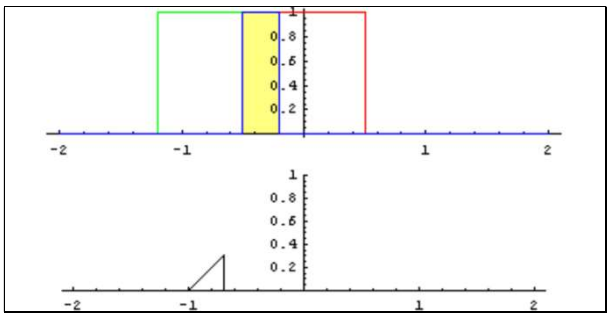
Propriétés :

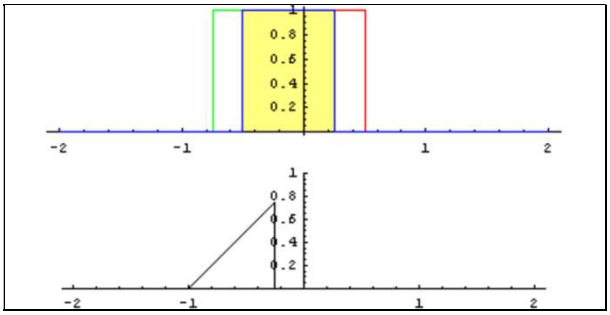
- *On montre que le produit de convolution est associatif et commutatif.*
- *L'étalement (intervalle de définition non-nulle) de la loi de probabilité de la variable aléatoire $h(s) = (f_1 * \ldots * f_n)(s)$ est la somme des étalements des lois de probabilité des variables aléatoires f_i*

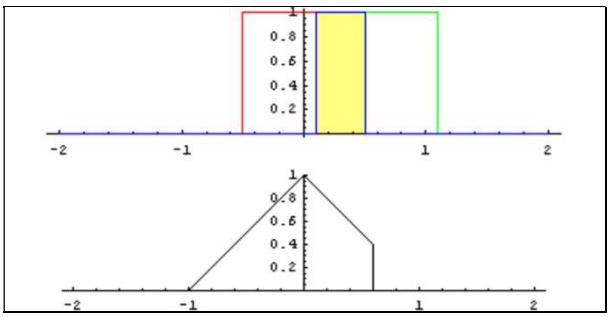
Exemple :

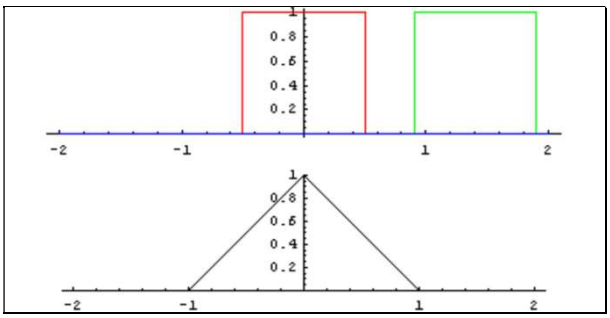
L'exemple suivant illustre comment interpréter graphiquement cette opération. On considère ici le produit de convolution de deux lois rectangulaires. La courbe rouge représente $g(x)$. La courbe verte représente $f(s-x)$, où s agit sur $f(x)$ comme un opérateur de translation. Les figures successives correspondent à différentes valeurs croissantes de s. Sur chaque figure, le graphique supérieur montre $f(s-x)$ et $g(x)$ (avec x en abscisse), et le graphique inférieur montre $h(s)$ (avec s en abscisse). En tout s, $h(s)$ est égal à la surface de recouvrement de $f(s-x)$ et $g(x)$ (en jaune).











Somme de deux variables aléatoires

Soit X_1 et X_2 deux variables aléatoires indépendantes continues dont les distributions de probabilité sont $f_{X1}(X)$ et $f_{X2}(X)$. Alors, la distribution de probabilité de la variable aléatoire $Y = X_1 + X_2$ est donnée par le produit de convolution de leur FDP respective :

$$f_Y(Y) = (f_{X1} * f_{X2})(Y) = \int_{-\infty}^{\infty} f_{X1}(Y - X) f_{X2}(X) dX$$

Cas particulier :

Si $Y = a + X$, où a est une constante, alors $f_Y(Y) = f_X(Y - a)$ (voir distribution de Dirac). En appliquant les définitions des moyenne et variance, on trouve $\mu_Y = \mu_X + a$ et $\sigma_Y^2 = \sigma_X^2$.

L'addition d'une constante à une variable aléatoire agit donc comme un opérateur de translation sur la loi de probabilité de cette variable. Cela laisse notamment inchangée sa « largeur » (écart-type) et sa « forme » (normale, rectangulaire, …).

Exemple :

Appliquons cette loi de propagation à l'exemple illustrant plus haut l'interprétation graphique du produit de convolution. On considère donc deux variables aléatoires indépendantes X_1 et X_2 dont les FDP sont les lois rectangulaires :

$$f_{X1}(X) = f_{X2}(X) = \begin{cases} 0 & pour\ X \notin [-0,5,+0,5] \\ 1 & pour\ X \in [-0,5,+0,5] \end{cases}$$

Alors, la variable aléatoire $Y = X_1 + X_2$ *sera caractérisée par la FDP :*

$f_Y(Y) = (f_{X1} * f_{X2})(Y)$

soit la fonction triangulaire :

$$f_Y(Y) = \begin{cases} 0 & pour\ Y \notin [-1,+1] \\ Y+1 & pour\ Y \in [-1,0] \\ -Y+1 & pour\ Y \in [0,+1] \end{cases}$$

On retrouve bien mathématiquement l'interprétation graphique.

Cette loi de propagation se généralise quel que soit le nombre de variables dans la somme.

Exemples :

- *Reprenons l'exemple ci-dessus. On considère à présent trois variables aléatoires indépendantes* X_1, X_2 *et* X_3 *ayant chacune pour FDP la loi rectangulaire considérée ci-dessus. Leur somme* $Y = X_1 + X_2 + X_3$ *aura pour FDP le produit de convolution de ces trois lois, soit une loi définie sur* $[-1,5,+1,5]$ *par 3 morceaux de polynômes du 2ème degré et nulle en dehors de cet intervalle.*
- *Le même exercice peut être extrapolé à n variables aléatoires indépendantes* $X_1,...,X_n$, *ayant chacune pour FDP la même loi rectangulaire. Leur somme* $Y = \sum_{i=1}^{n} X_i$ *aura pour FDP une loi symétrique, définie sur* $[-n.0,5,+n.0,5]$ *par n morceaux de polynômes du n-1ème degré et nulle en dehors de cet intervalle.*

Ce dernier exemple est important. Il permettra d'illustrer le théorème central limite.

Produit de deux variables aléatoires

Soit X_1 et X_2 deux variables aléatoires indépendantes continues dont les distributions de probabilité sont $f_{X1}(X)$ et $f_{X2}(X)$. Alors, la distribution de probabilité de la variable aléatoire $Y = X_1.X_2$ est donnée par :

$$f_Y(Y) = \int_{-\infty}^{\infty} f_{X1}(X) f_{X2}\left(\frac{Y}{X}\right) \frac{1}{|X|} dX$$

Cas particulier :

Si $Y = aX$, *où a est une constante, alors* $f_Y(Y) = \frac{1}{|a|} f_X\left(\frac{Y}{a}\right)$ *(voir distribution de Dirac). En appliquant les définitions des moyenne et variance, on trouve* $\mu_Y = a\mu_X$ *et* $\sigma_Y^2 = a^2 \sigma_X^2$.

La multiplication d'une variable aléatoire par une constante agit donc comme un opérateur d'échelle sur la loi de probabilité de cette variable. Cela laisse notamment inchangée sa « forme » (normale, rectangulaire, …).

Fonction d'une variable aléatoire (transformation d'une loi continue)

Soit *X* une variable aléatoire continue dont la distribution de probabilité est $f_X(X)$ et $\Psi(z)$ une fonction monotone et dérivable. Alors, la distribution de probabilité de la variable aléatoire $Y = \Psi(X)$ est donnée par :

$$f_Y(Y)=\left|\frac{d\Psi^{-1}}{dz}(Y)\right| f_X\left(\Psi^{-1}(Y)\right)$$

où Ψ^{-1} désigne la fonction réciproque de Ψ.

Cas particulier :

Si $Y=aX$, où a est une constante, alors $f_Y(Y)=\left|\frac{1}{a}\right| f_X\left(\frac{Y}{a}\right)$ (voir distribution de Dirac). On retrouve donc bien le résultat établi ci-dessus.

Exemples :

Le cas des lois dites « lois en cosinus » – avec argument distribué symétriquement autour de 0° – apparaît fréquemment dans la pratique. Aussi est-t-il intéressant de se pencher sur les deux exemples suivants.

Attention ! La fonction $Y=\cos(\theta)$ n'est pas monotone dans l'intervalle envisagé, mais sa symétrie sur cet intervalle et sa monotonie sur un demi-intervalle permet de ne considérer le calcul que sur un demi-intervalle de valeurs de θ, et ensuite de doubler la densité de probabilité obtenue.

- *On considère une variable aléatoire θ dont la FDP est la loi rectangulaire :*

$$f_\theta(\theta)=\begin{cases}0 & pour\ \theta\notin[-5\pi/180,+5\pi/180]\\ \frac{18}{\pi} & pour\ \theta\in[-5\pi/180,+5\pi/180]\end{cases}$$

Cette variable représente un angle pouvant prendre des valeurs distribuées équiprobablement dans l'intervalle $[-5\pi/180,+5\pi/180]$rad, soit dans $[-5°,+5°]$.

Alors, la variable aléatoire $Y=\cos(\theta)$ sera caractérisée par la FDP :

$$f_Y(Y)=\begin{cases}0 & pour\ Y\notin[\cos(5\pi/180),1]\\ 2\frac{1}{\sqrt{1-Y^2}}\frac{18}{\pi} & pour\ Y\in[\cos(5\pi/180),1]\end{cases}$$

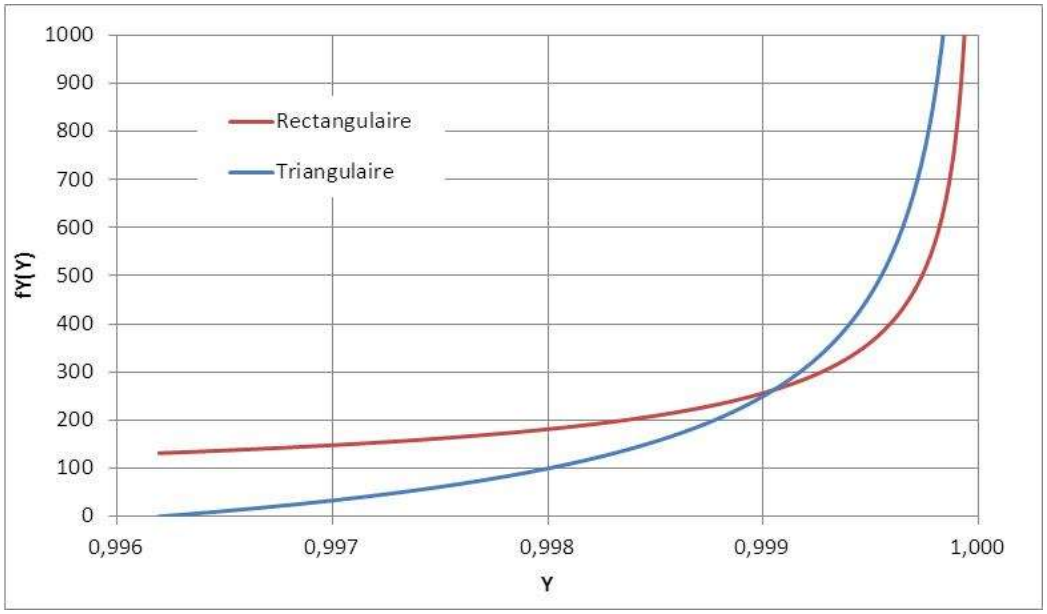
- *On considère une variable aléatoire θ dont la FDP est la loi triangulaire :*

$$f_\theta(\theta)=\begin{cases}0 & pour\ \theta\notin[-5\pi/180,+5\pi/180]\\ \frac{180}{5\pi}+\left(\frac{180}{5\pi}\right)^2\theta & pour\ \theta\in[-5\pi/180,0]\\ \frac{180}{5\pi}-\left(\frac{180}{5\pi}\right)^2\theta & pour\ \theta\in[0,+5\pi/180]\end{cases}$$

Cette variable représente un angle pouvant prendre des valeurs distribuées triangulairement dans l'intervalle $[-5\pi/180,+5\pi/180]$rad, soit dans $[-5°,+5°]$, avec un maximum en 0°.

Alors, la variable aléatoire $Y=\cos(\theta)$ sera caractérisée par la FDP :

$$f_Y(Y)=\begin{cases}0 & pour\ Y\notin[\cos(5\pi/180),1]\\ 2\frac{1}{\sqrt{1-Y^2}}\left(\frac{180}{5\pi}+\left(\frac{180}{5\pi}\right)^2 ar\cos(Y)\right) & pour\ Y\in[\cos(5\pi/180),1]\end{cases}$$



□ *Il est tentant de vouloir explorer le cas d'une variable aléatoire θ dont la FDP est la loi normale :*

$$f_\theta(\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\theta}{\sigma}\right)^2}$$

Cette variable représente un angle pouvant prendre des valeurs distribuées normalement autours de 0° avec un écart-type σ.
Cependant, la loi normale possède deux queues de distribution étalées jusqu'à l'infini. En conséquence, si la fonction $Y = cos(\theta)$ est symétrique sur l'intervalle envisagé (de $-\infty$ à ∞), elle n'est pas monotone sur un demi-intervalle et aucun calcul n'est envisageable.
A titre d'information, on trouvera dans le GUM (annexe F, § F.2.4.4) le cas particulier d'une variable aléatoire θ dont la FDP est la loi normale lorsque $\theta <<$.

Ces trois propriétés générales constituent les « lois de propagation des FDP » et permettent de résoudre l'équation $Y = f(X_1,...,X_n)$ lorsque f prend une forme algébrique (non-différentielle) et explicite, ce qui représente la plupart des cas rencontrés en pratique. On met alors ces propriétés en œuvre pour réduire la fonction *f* de proche en proche jusqu'à la formulation finale de la distribution de probabilité de *Y*.

Cependant, la pratique montre que même dans les cas non-linéaires les plus simples (fonction *f* élémentaire et ne comportant que peu de grandeurs d'entrée) les développements mathématiques peuvent devenir rapidement intraitables par la voie analytique. Deux solutions sont alors envisageables :

- la résolution par voie numérique,
- la linéarisation de la relation $Y = f(X_1,...,X_n)$.

La première ne sera pas développée ici. Pour plus d'information, se reporter par exemple à la publication « Évaluation des données de mesure — Supplément 1 du "Guide pour l'expression de l'incertitude de mesure" – Propagation de distributions par une méthode de Monte Carlo » (JCGM 101:2008).
La seconde est développée ci-dessous. La plupart des publications dans ce domaine n'envisagent d'ailleurs que le cas de relations linéaires tout en passant sous silence les lois générales de propagation des FDP. Il est vrai que bon nombre de cas pratiques se résument de toute façon à des relations linéaires, et que les autres cas sont généralement linéarisés faute d'autre possibilité de résolution.

2.3.2 Formulation linéaire

Un développement de $Y = f(X_1,...,X_n)$ en série de Taylor autour des moyennes $(\mu_1,...,\mu_n)$ donne :

$$Y = f(\mu_1,...,\mu_n) + \sum_{i=1}^{n} \left.\frac{\partial f(X_1,...,X_n)}{\partial X_i}\right|_{\mu_1,...,\mu_n} (X_i - \mu_i) + O^2$$

que nous conviendrons de noter plus simplement :

$$Y = f(\mu) + \sum_{i=1}^{n} \left.\frac{\partial f(X)}{\partial X_i}\right|_{\mu} (X_i - \mu_i) + O^2$$

Un système linéaire (ou linéarisé) est caractérisé par une contribution nulle (ou négligeable) des termes de degrés supérieurs O^2, menant à la relation qui nous intéresse :

$$Y = \sum_{i=1}^{n} c_i X_i + b$$

où les $c_i = \left.\frac{\partial f(X)}{\partial X_i}\right|_{\mu}$ et $b = f(\mu) - \sum_{i=1}^{n} \left.\frac{\partial f(X)}{\partial X_i}\right|_{\mu} \mu_i$ sont des constantes.

Dans la pratique, la détermination analytique exacte de la distribution de probabilité de *Y* n'est que très rarement utile. Nous verrons qu'il est jugé plus commode de n'en retenir qu'un nombre limité de paramètres représentatifs : la moyenne, la variance (ou l'écart-type), et la « forme » générale de la distribution.

Dans le cas de telles relations linéaires, les lois de propagation des FDP dégénèrent et font place aux lois de propagation simplifiées exposées ci-dessous. On bénéficie alors de simplifications surprenantes qui permettent notamment de systématiser substantiellement le traitement mathématique.

Loi de propagation des moyennes

Soit *n* variables aléatoires $X_1,...,X_n$ de moyennes $\mu_1,...,\mu_n$. Alors la moyenne de la variable aléatoire $Y = \sum_{i=1}^{n} c_i X_i + b$ où les c_i et *b* sont des constantes est donnée par (appliquer la définition de la moyenne) :

$$\mu_Y = \sum_{i=1}^{n} c_i \mu_i + b$$

Loi de propagation des variances

Soit *n* variables aléatoires $X_1,...,X_n$ de variances $\sigma_1^2,...,\sigma_n^2$. Alors la variance de la variable aléatoire $Y = \sum_{i=1}^{n} c_i X_i + b$ où les c_i et *b* sont des constantes est donnée par (appliquer la définition de la variance) :

$$\sigma_Y^2 = \underbrace{\sum_{i=1}^{n} c_i^2 \sigma_i^2}_{\substack{\text{contribution de chaque } X_i \\ \text{indépendante}}} + \underbrace{2\sum_{i=1}^{n-1}\sum_{l=i+1}^{n} c_i c_l \sigma_{X_i X_l}}_{\substack{\text{contribution des } X_i \\ \text{de la corrélation}}}$$

où $\sigma_{X_i X_l}$ désigne la covariance de X_i et X_l.

La plupart du temps, les variables $X_1,...,X_n$ sont non-corrélées, c'est-à-dire mutuellement indépendantes. On montre alors que les coefficients de corrélation sont nuls, et la variance de *Y* est égale à la somme des variances des X_i pondérée par les coefficients c_i^2 :

$$\sigma_Y^2 = \sum_{i=1}^{n} c_i^2 \sigma_i^2$$

Exemples :

On considère la variable aléatoire Y donnée par la relation $Y = X_1 + X_2 + X_2 + X_2$. La dernière formule fournit $\sigma_Y^2 = \sigma_1^2 + \sigma_2^2 + \sigma_2^2 + \sigma_2^2$, soit $\sigma_Y^2 = \sigma_1^2 + 3\sigma_2^2$. Bien entendu, la relation $Y = f(X_1, X_2)$ peut être réexprimée sous la forme $Y = X_1 + 3X_2$ pour laquelle la même formule fournit $\sigma_Y^2 = \sigma_1^2 + 9\sigma_2^2$. Or ce résultat n'est pas celui trouvé ci-avant ! C'est parce que la première formulation de la relation $Y = f(X_1, X_2)$ comporte 3 termes égaux, donc dépendants, donc corrélés. Dans ce cas, c'est la formule complète (avec les termes de covariance) qui aurait dû être appliquée. Celle-ci aurait ajouté trois fois les doubles produits $2\sigma_2^2$ conduisant à la variance correcte.

On remarque directement l'intérêt de ces deux lois de propagation : elles sont extrêmement simples d'application (elles dispensent de mettre en œuvre les lois de propagation des FDP vues plus haut) et, surtout, elles ne nécessitent pas de connaître l'expression analytique exacte des FDP des variables aléatoires $X_1,...,X_n$, mais seulement leurs moyennes et variances, pour établir la moyenne et la variance de la variable aléatoire *Y*.

Forme de la distribution résultante

La moyenne et la variance ne caractérisent que partiellement la variable aléatoire *Y*. Dans de nombreuses applications pratiques (voir les niveaux de confiance par exemple), il est nécessaire de compléter la caractérisation de *Y* par une information supplémentaire : la « forme » de sa distribution de probabilité (pas sa formulation analytique exacte, mais simplement sa forme : normale, rectangulaire, …).

Puisque nous considérons ici le cas d'une relation linéaire $Y = \sum_{i=1}^{n} c_i X_i + b$, et que nous avons montré que la multiplication d'une variable aléatoire par une constante (termes « $c_i X_i$ ») et que l'addition d'une constante à une variable aléatoire (terme « *b* ») laissent inchangée sa « forme » (normale, rectangulaire, …), il suffit d'appliquer la première loi de propagation des FDP (somme de variables aléatoires). On montre ainsi aisément que :

- le produit de convolution de deux lois rectangulaires est une loi trapézoïdale,
- le produit de convolution d'un nombre quelconque de lois normales reste une loi normale.

Mais qu'en est-il des autres cas ? Existe-t-il une règle permettant de généraliser cette démarche ? Le théorème central limite répond à cette question.

2.4 THÉORÈME CENTRAL LIMITE

2.4.1 Enoncé

La version originale du théorème central limite est la suivante.

Soit *n* variables aléatoires $X_1,...,X_n$, indépendantes et de même loi de probabilité (quelle que soit cette loi), de moyenne μ et de variance σ^2, et $\Sigma_n = \sum_{i=1}^{n} X_i$ la somme de ces variables. Alors, la loi de probabilité de la variable aléatoire $Z_n = \frac{\Sigma_n - n\mu}{\sigma\sqrt{n}}$ tend vers la loi normale centrée réduite $N(1,0)$ lorsque $n \to \infty$.

Interprétation :

Cela revient à dire – de façon équivalente – que la loi de probabilité de la variable aléatoire $\Sigma_n = \sum_{i=1}^{n} X_i$ tend vers la loi normale $N(n\mu, n\sigma^2)$ lorsque $n \to \infty$. Les paramètres $n\mu$ et $n\sigma^2$ se déduisent par ailleurs directement des lois de propagation des moyennes et des variances vues plus haut.

Illustration intuitive :

Nous avons montré, en exemple de la première loi de propagation des FDP (somme de variables aléatoires), que la somme de n variables aléatoires indépendantes $X_1,...,X_n$, ayant chacune pour FDP la même loi rectangulaire, a pour FDP une loi symétrique, définie sur $[-n.0,5,+n.0,5]$ par n morceaux de polynômes du n-1ème degré et nulle en dehors de cet intervalle. Ceci illustre, lorsque n augmente, le lissage progressif de la loi obtenue et l'étalement croissant de ses queues.

Application particulière :

La loi de probabilité de la variable aléatoire $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ tend vers la loi normale $N(\mu, \sigma^2/n)$ lorsque $n \to \infty$. En d'autres termes, pour les grands échantillons, le théorème central limite permet d'affirmer que la moyenne empirique d'échantillons suit – au moins à peu près – une loi normale, et ce quelle que soit la loi de distribution sous-jacente.

2.4.2 Généralisation

Cette version originale du théorème central limite a par la suite admis plusieurs généralisations qui assurent la même conclusion sous des hypothèses beaucoup plus faibles. Principalement, la condition de Lindeberg (« la variance de chacun des termes individuels est négligeable vis-à-vis de la variance de la somme ») généralise ce théorème lorsque les lois de probabilité individuelles sont différentes. D'autres généralisations autorisent même une dépendance « faible » entre les variables.

Signification remarquable :

Toute somme de n variables aléatoires, indépendantes ou faiblement dépendantes, de distributions de probabilité respectives quelconques (et donc même différentes), converge vers une variable aléatoire de distribution de probabilité normale, à condition qu'aucune des variables n'exerce une influence (en terme de largeur) significativement plus importante que les autres.

Remarque :

On a montré en cas particulier de la seconde loi de propagation des FDP (produit de deux variables aléatoires) que si $Y = aX$, où a est une constante, alors $\sigma_Y^2 = a^2\sigma_X^2$. En conséquence, si $\Sigma_n = \sum_{i=1}^{n} c_i X_i$ alors la condition de Lindeberg doit être appliquée aux $c_i^2\sigma_i^2$ et non plus aux σ_i^2.

Le théorème central limite donne par ailleurs lieu au corollaire suivant.

Produits de variables aléatoires

Le logarithme d'un produit (à facteurs strictement positifs) est la somme des logarithmes des facteurs, de sorte que le logarithme d'un produit de variables aléatoires (à valeurs strictement positives) tend vers une loi normale en vertu du théorème central limite. En conséquence, tout produit de variables aléatoires (à valeurs strictement positives) converge vers une loi log-normale.

Bon nombre de grandeurs physiques ne peuvent prendre que des valeurs positives – comme la masse ou la longueur – et s'avèrent être le produit de différents facteurs aléatoires, de sorte qu'elles suivent une loi log-normale.
La loi log-normale apparaît régulièrement dans la description de phénomènes naturels caractérisés par un processus de croissance obéissant à la loi de Gibrat, selon laquelle le taux de croissance de l'entité est indépendant de la taille de l'entité. On peut montrer que les phénomènes qui suivent la loi de Gibrat sont généralement caractérisés par une loi log-normale. On en trouve de nombreux exemples en biologie, médecine, hydrologie, économie, finance, démographie, …

2.4.3 Convergence

Propriétés :

- *Le théorème de Berry–Esseen quantifie la vitesse de convergence vers la loi normale. Soit* $F_n(x)$ *la fonction de répartition de* $\Sigma_n = \sum_{i=1}^{n} X_i$ *et* $\Phi(x)$ *la fonction de répartition de* $N(n\mu, n\sigma^2)$*. Alors, pour tout x et pour tout n, on a* $|F_n(x) - \Phi(x)| \leq \frac{C}{\sqrt{n}}$ *où C est une constante.*
- *La convergence est d'autant plus rapide que les distributions de probabilité des variables sont symétriques (c'est le cas des lois normales et rectangulaires), et que ces variables sont indépendantes.*
- *L'approximation normale est particulièrement efficace au voisinage des valeurs centrales. La précision se dégrade à mesure qu'on s'éloigne de ces valeurs centrales.*

En elle-même, la convergence vers la loi normale d'une somme de variables aléatoires lorsque leur nombre tend vers l'infini n'intéresse que le mathématicien. Pour le praticien, il est intéressant de s'arrêter avant la limite : la somme d'un grand nombre de ces variables est presque normale, ce qui fournit une approximation plus facilement utilisable que la loi exacte. Le problème est de savoir à partir de quelle valeur *n* est « assez grand » pour obtenir la précision souhaitée.
On trouve à ce sujet de nombreux avis peu argumentés. On citera l'interprétation suivante pour exemple.

« L'approximation donnée par le théorème central limite est bonne pour :

- $n \geq 4$ si les lois de probabilité des X_i sont proches d'une loi normale,
- $n \geq 12$ si les lois de probabilité des X_i sont moyennement proches d'une loi normale (par exemple de loi rectangulaire),
- $n \geq 100$ si les lois de probabilité des X_i ne sont pas proches d'une loi normale (par exemple de loi très asymétrique). »

Exemples :

Il est utile d'illustrer par quelques exemples concrets la vitesse de convergence vers la loi normale d'une somme de n variables aléatoires de même loi de probabilité :

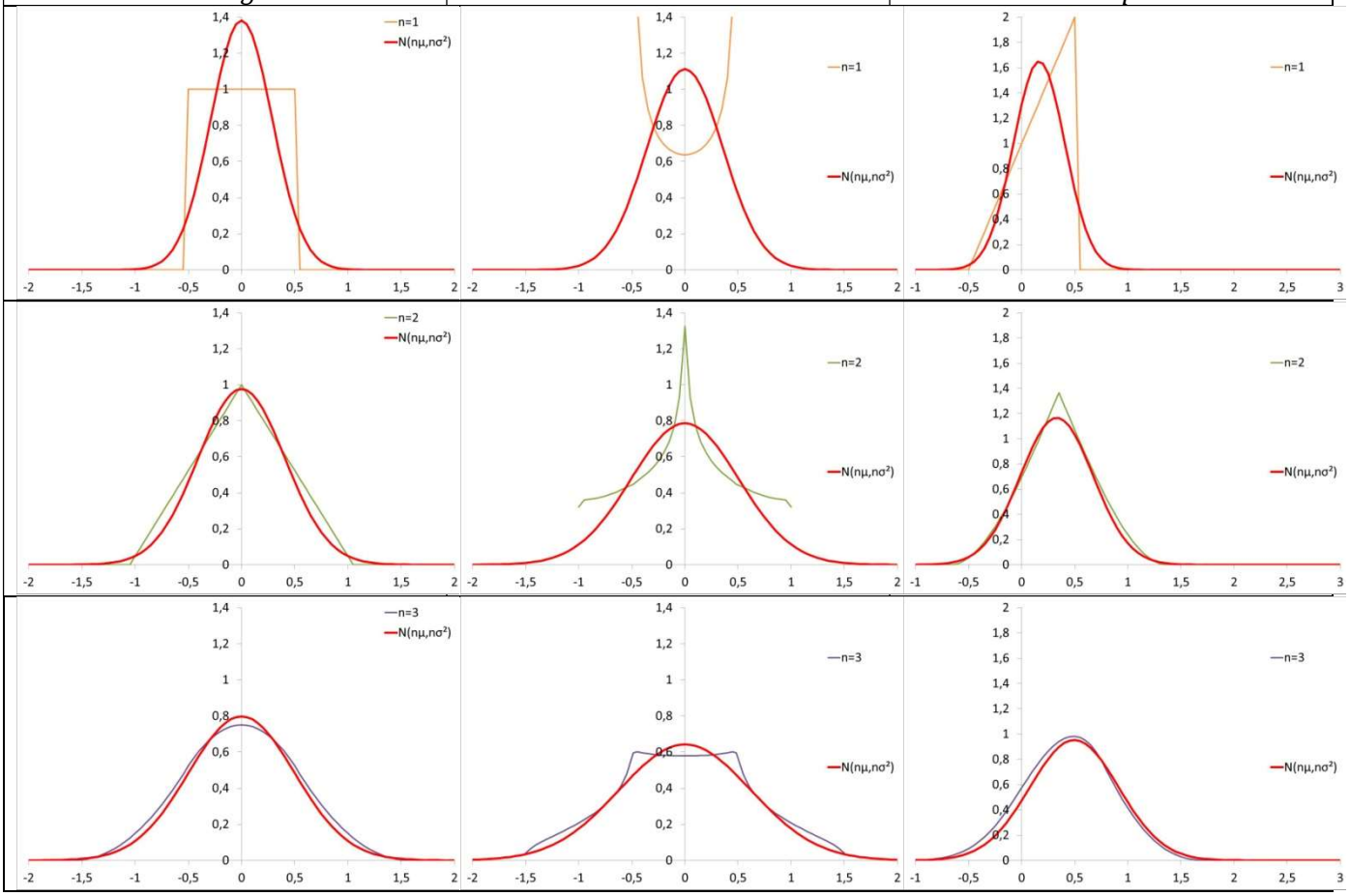
□ *Exemple 1 (à gauche) : lois rectangulaires* $f(X)=\begin{cases}0 & pour\ X\notin[-0{,}5,+0{,}5]\\ 1 & pour\ X\in[-0{,}5,+0{,}5]\end{cases}$, *il s'agit donc de lois symétriques, de moyenne* $\mu=0$ *, avec densité de probabilité uniforme autour de la moyenne*

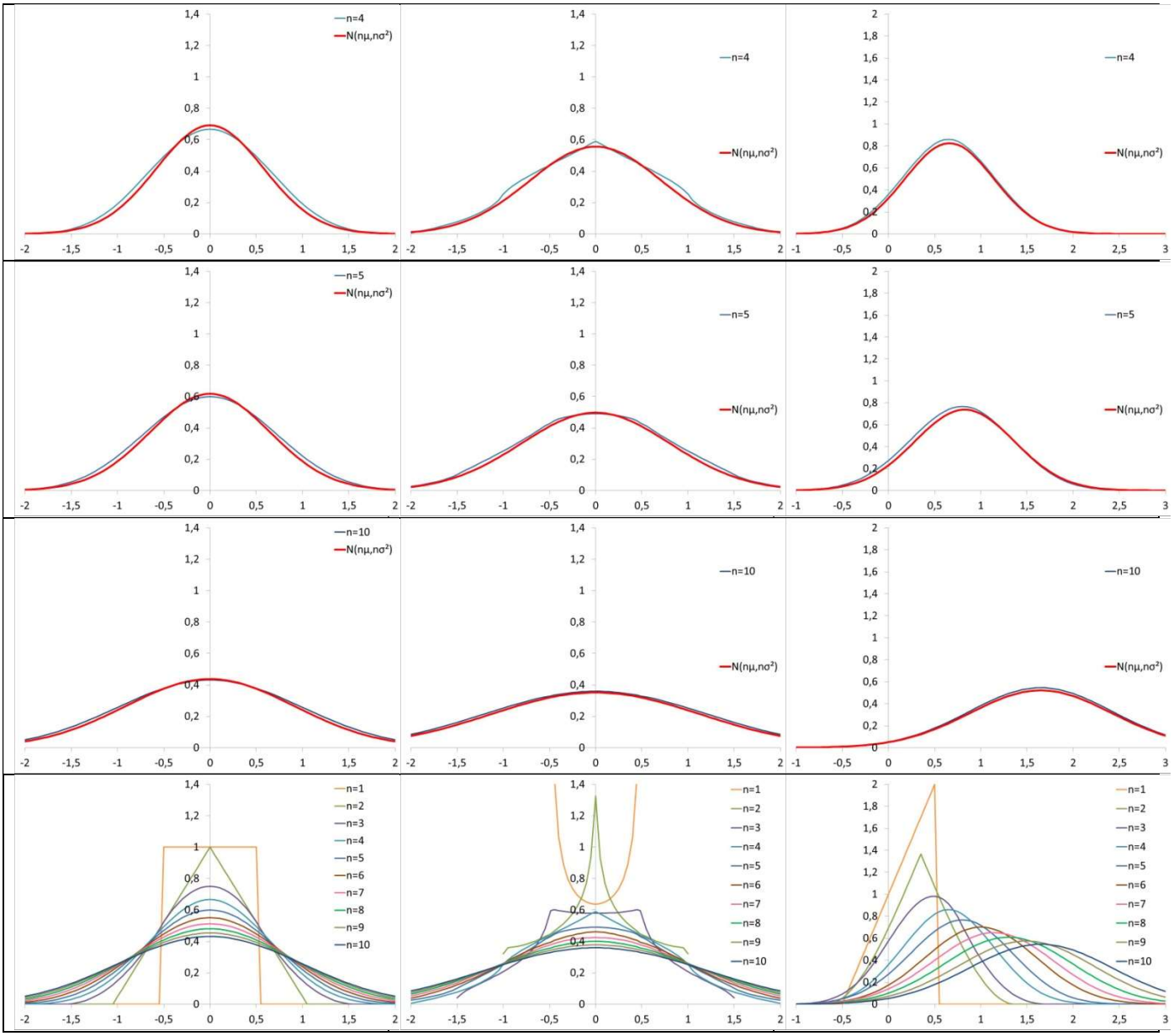
□ *Exemple 2 (au milieu) : lois en U* $f(X)=\begin{cases}0 & pour\ X\notin[-0{,}5,+0{,}5]\\ \dfrac{2}{\pi\sqrt{1-4X^2}} & pour\ X\in[-0{,}5,+0{,}5]\end{cases}$, *il s'agit donc de lois symétriques, de moyenne* $\mu=0$ *, avec densité de probabilité excentrée autour de la moyenne*

□ *Exemple 3 (à droite) : lois en pente* $f(X)=\begin{cases}0 & pour\ X\notin[-0{,}5,+0{,}5]\\ 2X+1 & pour\ X\in[-0{,}5,+0{,}5]\end{cases}$, *il s'agit donc de lois asymétriques, de moyenne* $\mu=1/6$ *, avec densité de probabilité excentrée à droite de la moyenne*

Ces lois sont celles représentées en orange sur les graphiques de la première ligne. Le tableau ci-dessous illustre l'évolution pour $n=1,2,3,4,5\ et\ 10$ *de la somme (produit de convolution) de n lois en comparaison avec la loi normale (en rouge) prédite par le théorème central limite.*

Lois rectangulaires	*Lois en U*	*Lois en pente*





On peut y observer l'évolution des paramètres de moyenne $n\mu$ (nulle pour les exemples 1 et 2, déplacement croissant vers la droite de la distribution résultante pour l'exemple 3) et de variance $n\sigma^2$ (étalement croissant de la distribution résultante).

On déduit de cet exemple que l'approximation donnée par le théorème central limite est bonne pour $n \geq 4$ dans tous les cas, voire même pour $n \geq 3$ pour autant que la densité de probabilité ne soit pas trop excentrée autour de la moyenne. Il apparaît en tout cas qu'il n'est pas nécessaire d'exiger $n \geq 10$ pour garantir une bonne convergence…

2.4.4 Conséquence

On peut parfois lire dans la littérature que la « courbe en cloche » de la loi normale représente bien la « loi du hasard », ou peu importe comment elle est appelée, ce qui n'a pas grande signification en soi. En réalité le succès sans égal de la loi normale est la conséquence directe du théorème central limite. Car en pratique,

bon nombre de phénomènes naturels sont dus à la superposition de causes nombreuses, aléatoires, et plus ou moins indépendantes. Il en résulte que la loi normale les représente de manière raisonnablement fidèle.

Le théorème central limite explique donc l'omniprésence de la loi normale dans la représentation de nombreux phénomènes.

Limitation :

On a vu que l'approximation normale se dégrade à mesure qu'on s'éloigne des valeurs centrales. La raison principale en est la suivante.
Une conséquence directe de la première loi de propagation des FDP (somme de variables aléatoires) est que l'étalement (intervalle de définition non-nulle) de la loi de probabilité de la variable aléatoire $\Sigma_n = \sum_{i=1}^{n} X_i$ *est la somme des étalements des lois de probabilité des variables aléatoires* X_i *(propriété du produit de convolution). En pratique, toute grandeur physique est nécessairement bornée et – à fortiori – la somme de grandeurs physiques le sera aussi. Or la loi normale possède deux queues de distribution étalées jusqu'à l'infini : la loi normale fait toujours apparaître des valeurs positives et négatives étalées jusqu'à l'infini. L'identification à la loi normale loin des valeurs centrales peut apparaître encore moins appropriée pour une somme de variables positives par nature.*
On peut relativiser en se rappelant que les valeurs étalées loin des valeurs centrales ne le sont qu'avec des probabilités certes non nulles mais faibles, et en étant conscient que la loi normale n'est donc qu'une approximation utile.

2.5 ESTIMATION D'ÉCHANTILLONS ISSUS DE LOIS NORMALES

En raison de son omniprésence, la loi normale (voir théorème central limite) a naturellement fait l'objet de plus d'attentions que les autres lois. Ainsi dispose-t-on notamment des propriétés remarquables exposées ci-dessous et relatives à la loi normale.

ATTENTION ! Les propriétés de ce chapitre ne s'appliquent que pour des échantillons issus de lois normales.

2.5.1 Objet

Nous avons expliqué que, dans toute application, les estimateurs doivent être utilisés en lieu et place des valeurs vraies des paramètres si celles-ci sont inconnues. Nous avons alors donné les formules des meilleurs estimateurs pour la moyenne, la variance et l'écart-type. Ces formules fournissent ce qu'on appelle des estimations ponctuelles, en raison du fait que l'estimation produite est un nombre. La faiblesse de l'estimation ponctuelle vient de ce qu'elle est fournie sans aucune information sur sa fiabilité.

Il est parfois possible de mesurer cette fiabilité. C'est l'objectif que se fixe une autre forme d'estimation, l'**estimation par intervalle**. Nous avons déjà vu qu'un intervalle de confiance associé à un niveau de confiance est une mesure de la fiabilité que l'on peut accorder à une estimation. L'objet de ce chapitre est d'appliquer ces notions au cas des estimateurs d'échantillons issus de lois normales.

En deux mots, étant donné un échantillon, l'estimation par intervalle détermine l'intervalle (intervalle de confiance) tel qu'il soit possible de calculer la probabilité (niveau de confiance) pour que cet intervalle contienne la valeur vraie du paramètre estimé. Pour un niveau de confiance donné, plus court est l'intervalle de confiance, meilleure est la précision avec laquelle la valeur vraie du paramètre a été localisée.

Ces propriétés sont particulièrement importantes car elles constituent les bases sur lesquelles sont construits les tests statistiques vus plus bas (tests paramétriques de conformité et d'homogénéité).

Ci-dessous :

- $N_\alpha(0,1)$ ($=-N_{1-\alpha}(0,1)$ par symétrie de la loi normale) désigne le quantile d'ordre α de la loi normale centrée réduite,
- $T_\alpha(k)$ ($=-T_{1-\alpha}(k)$ par symétrie de la loi de Student) désigne le quantile d'ordre α de la loi de Student à *k* degrés de liberté,
- $\chi^2_\alpha(k)$ désigne le quantile d'ordre α de la loi du khi-deux à *k* degrés de liberté,
- $F_\alpha(k,l)$ désigne le quantile d'ordre α de la loi de Fisher avec k et l degrés de liberté.

Excel :

LOI.NORMALE.INVERSE.N(α, 0, 1) donne $N_\alpha(0,1)$
LOI.STUDENT.INVERSE.N(α, k) donne $T_\alpha(k)$
LOI.KHIDEUX.INVERSE(α, k) donne $\chi^2_\alpha(k)$
INVERSE.LOI.F.N(α, k, l) donne $F_\alpha(k,l)$

2.5.2 Estimations d'une moyenne

Variance σ^2 connue

Soit une variable aléatoire *X* de loi normale $N(\mu,\sigma^2)$ où σ^2 est connu. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,\ldots,x_i,\ldots,x_n$ de cette variable *X*, de moyenne estimée $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$. Alors on montre que $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ suit la loi normale centrée réduite $N(0,1)$.

Conséquence :

La propriété ci-dessus nous permet de faire de l'inférence sur la moyenne μ d'une loi normale avec variance σ^2 connue. Cette propriété donne en effet l'intervalle de confiance bilatéral

$$\bar{x}-N_{1-\alpha/2}(0,1)\frac{\sigma}{\sqrt{n}}\le\mu\le\bar{x}+N_{1-\alpha/2}(0,1)\frac{\sigma}{\sqrt{n}}$$

au niveau de confiance $1-\alpha$ pour la moyenne μ d'une loi normale $N(\mu,\sigma^2)$ lorsque la variance σ^2 est connue.

Variance σ^2 inconnue

Soit une variable aléatoire *X* de loi normale $N(\mu,\sigma^2)$ où σ^2 est inconnu. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,\ldots,x_i,\ldots,x_n$ de cette variable *X*, de moyenne estimée $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ et de variance estimée $s_X^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$. Alors on montre que $\frac{\bar{x}-\mu}{s_X/\sqrt{n}}$ suit la loi de Student $T(n-1)$.

Conséquence :

La propriété ci-dessus nous permet de faire de l'inférence sur la moyenne μ d'une loi normale avec variance σ^2 inconnue. Cette propriété donne en effet l'intervalle de confiance bilatéral

$$\bar{x} - T_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + T_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

au niveau de confiance $1-\alpha$ pour la moyenne μ d'une loi normale $N(\mu,\sigma^2)$ lorsque la variance σ^2 est inconnue.

2.5.3 Estimations d'une variance

Moyenne μ connue

Soit une variable aléatoire *X* de loi normale $N(\mu,\sigma^2)$ où μ est connu. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,...,x_i,...,x_n$ de cette variable *X*, de variance estimée $s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$. Alors on montre que $\frac{ns_X^2}{\sigma^2}$ suit la loi du khi-deux $\chi^2(n)$.

Conséquence :

La propriété ci-dessus nous permet de faire de l'inférence sur la variance σ^2 d'une loi normale avec moyenne μ connue. Cette propriété donne en effet l'intervalle de confiance bilatéral

$$\frac{ns^2}{\chi^2_{1-\alpha/2}(n)} \le \sigma^2 \le \frac{ns^2}{\chi^2_{\alpha/2}(n)}$$

au niveau de confiance $1-\alpha$ pour la variance σ^2 d'une loi normale $N(\mu,\sigma^2)$ lorsque la moyenne μ est connue.

Les racines carrées des bornes donnent un intervalle de confiance pour l'écart-type au risque α.

Moyenne μ inconnue

Soit une variable aléatoire *X* de loi normale $N(\mu,\sigma^2)$ où μ est inconnu. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,...,x_i,...,x_n$ de cette variable *X*, de moyenne estimée $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$ et de variance estimée $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$. Alors on montre que $\frac{(n-1)s_X^2}{\sigma^2}$ suit la loi du khi-deux $\chi^2(n-1)$.

Conséquence :

La propriété ci-dessus nous permet de faire de l'inférence sur la variance σ^2 d'une loi normale avec moyenne μ inconnue. Cette propriété donne en effet l'intervalle de confiance bilatéral

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}$$

au niveau de confiance $1-\alpha$ pour la variance σ^2 d'une loi normale $N(\mu,\sigma^2)$ lorsque la moyenne μ est inconnue.

Les racines carrées des bornes donnent un intervalle de confiance pour l'écart-type au risque α.

2.5.4 Estimations du rapport de deux variances

Propriété :

Soit deux variables aléatoires indépendantes X et Y de lois du khi-deux $\chi^2(k)$ *et* $\chi^2(l)$*. Alors on montre que* $\frac{X/k}{Y/l}$ *suit la loi de Fisher* $F(k,l)$.

Moyennes μ_X et μ_Y connues

Soit deux variables aléatoires indépendantes *X* et *Y* de lois normales $N(\mu_X,\sigma_X^2)$ et $N(\mu_Y,\sigma_Y^2)$ où μ_X et μ_Y sont connus. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,...,x_i,...,x_n$ de cette variable *X*, de variance estimée $s_X^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_X)^2$. Soit tout échantillon de *m* valeurs indépendantes $(m>1)$ $y_1,...,y_i,...,y_n$ de cette variable *Y*, de variance estimée $s_Y^2=\frac{1}{m}\sum_{i=1}^{m}(y_i-\mu_Y)^2$. Alors on montre que $\frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2}$ suit la loi de Fisher $F(n,m)$.

Conséquence :

La propriété ci-dessus nous permet de faire de l'inférence sur le rapport des variances $\frac{\sigma_x^2}{\sigma_y^2}$ *de lois normales avec moyennes* μ_X *et* μ_Y *connues. Cette propriété donne en effet l'intervalle de confiance bilatéral*

$$\frac{1}{F_{1-\alpha/2}(n,m)}\frac{S_x^2}{S_y^2}\le\frac{\sigma_x^2}{\sigma_y^2}\le\frac{1}{F_{\alpha/2}(n,m)}\frac{S_x^2}{S_y^2}$$

au niveau de confiance $1-\alpha$ *pour le rapport des variances* $\frac{\sigma_x^2}{\sigma_y^2}$ *de lois normales* $N(\mu_X,\sigma_X^2)$ *et* $N(\mu_Y,\sigma_Y^2)$ *lorsque les moyennes* μ_X *et* μ_Y *sont connues.*

Les racines carrées des bornes donnent un intervalle de confiance pour le rapport des écarts-types au risque α.

Moyennes μ_X et μ_Y inconnues

Soit deux variables aléatoires indépendantes *X* et *Y* de lois normales $N(\mu_X,\sigma_X^2)$ et $N(\mu_Y,\sigma_Y^2)$ où μ_X et μ_Y sont inconnus. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,...,x_i,...,x_n$ de cette variable *X*, de moyenne estimée $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ et de variance estimée $s_X^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$. Soit tout échantillon de *m* valeurs indépendantes $(m>1)$ $y_1,...,y_i,...,y_n$ de cette variable *Y*, de moyenne estimée $\bar{y}=\frac{1}{m}\sum_{i=1}^{m}y_i$ et de variance estimée $s_Y^2=\frac{1}{m-1}\sum_{i=1}^{m}(y_i-\bar{y})^2$. Alors on montre que $\frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2}$ suit la loi de Fisher $F(n-1,m-1)$.

Conséquence :

La propriété ci-dessus nous permet de faire de l'inférence sur le rapport des variances $\frac{\sigma_x^2}{\sigma_y^2}$ *de lois normales avec moyennes* μ_X *et* μ_Y *inconnues. Cette propriété donne en effet l'intervalle de confiance bilatéral*

$$\frac{1}{F_{1-\alpha/2}(n-1,m-1)}\frac{S_x^2}{S_y^2} \le \frac{\sigma_x^2}{\sigma_y^2} \le \frac{1}{F_{\alpha/2}(n-1,m-1)}\frac{S_x^2}{S_y^2}$$

au niveau de confiance $1-\alpha$ *pour le rapport des variances* $\frac{\sigma_x^2}{\sigma_y^2}$ *de lois normales* $N(\mu_X, \sigma_X^2)$ *et* $N(\mu_Y, \sigma_Y^2)$ *lorsque les moyennes* μ_X *et* μ_Y *sont inconnues.*

Les racines carrées des bornes donnent un intervalle de confiance pour le rapport des écarts-types au risque α.

2.6 TESTS STATISTIQUES

2.6.1 Généralités sur les tests

Un ***test d'hypothèse*** est une démarche consistant à accepter ou rejeter une hypothèse statistique, appelée hypothèse nulle, émise sur un ou plusieurs échantillons de données. Il s'agit donc, à partir de calculs réalisés sur un ou plusieurs échantillons de données observées, d'émettre des conclusions sur la ou les populations totales dont sont issus ces échantillons. Ces conclusions sont accompagnées d'une déclaration concernant le risque d'erreur.

En effet, les méthodes de l'inférence statistique vont nous permettre de déterminer, avec une probabilité donnée, si les différences constatées entre le ou les échantillons et l'hypothèse émise peuvent être imputables au hasard (auquel cas l'hypothèse sera acceptée) ou si elles sont suffisamment importantes pour signifier que les échantillons ne satisfont pas à cette hypothèse (auquel cas l'hypothèse est rejetée). Les hypothèses émises concernent les paramètres (moyenne, variance, …) ou la nature (loi normalité, …) des populations dont proviennent le ou les échantillons étudiés.

Catégories de tests

Les tests peuvent être classés selon différents critères. Citons notamment : leur finalité, le nombre de variables d'intérêt, l'existence d'hypothèses à priori sur les distributions des données, et le mode de constitution des échantillons.

Les tests selon leur finalité

La finalité définit l'objectif du test, en d'autres termes : l'information que l'on souhaite extraire des données.

- Le ***test de conformité*** consiste à confronter un paramètre calculé sur un échantillon (moyenne, variance, …) à une valeur pré-établie. On parle alors de test de conformité à un standard. Le plus connu est certainement le test portant sur la moyenne.
- Le ***test d'homogénéité*** consiste à confronter les paramètres calculés pour différents échantillons (moyenne, variance, …) entre eux. Ce test peut être utilisé pour évaluer si différents échantillons proviennent de la même population.
- Le ***test d'adéquation*** consiste à vérifier la compatibilité des données avec une distribution choisie a priori. Le test le plus utilisé dans cette optique est le test d'adéquation à la loi normale.
- Le ***test d'association*** consiste à éprouver l'existence d'une liaison entre 2 variables.

Les tests selon le nombre de variable

Principalement concernant les tests de conformité et d'homogénéité, on dit que le test est ***univarié*** s'il ne porte que sur une seule variable d'intérêt (par exemple, comparer des échantillons de résultats de mesure réalisées à différentes températures ambiantes), il est ***multivarié*** s'il met en jeu simultanément plusieurs variables (par exemple la comparaison porte sur la température ambiante, l'équipement utilisé, la méthode de mesure mise en oeuvre, l'opérateur, …).

Tests paramétriques et non paramétriques

On parle de ***tests paramétriques*** lorsque l'on stipule que les données sont issues d'une distribution paramétrée. Dans ce cas, les caractéristiques des données peuvent être résumées à l'aide de paramètres estimés sur l'échantillon, la procédure de test subséquente ne porte alors que sur ces paramètres.

Exemple :

L'hypothèse de normalité sous-jacente des données est le plus souvent utilisée. Dans ce cas, la moyenne et la variance suffisent pour caractériser complètement la distribution. Dans le cas d'un test d'homogénéité par exemple, pour éprouver l'égalité des distributions, il suffira donc de comparer les moyennes et les variances.

Les ***tests non paramétriques*** ne font aucune hypothèse sur la distribution sous-jacente des données. On les qualifie souvent de tests « distribution free ». L'étape préalable consistant à estimer les paramètres des distributions avant de procéder au test d'hypothèse proprement dit n'est plus nécessaire.

Les tests selon la constitution des échantillons

Cet aspect est surtout associé aux tests d'homogénéité.
On parle d'***échantillons indépendants*** lorsque les observations sont indépendantes à l'intérieur des échantillons et d'un échantillon à l'autre. C'est le cas lorsque l'échantillon provient d'un échantillonnage simple dans la population totale.

Exemple :

Un industriel affirme que son additif pour essence permet de réduire la consommation des automobiles. Pour vérifier cette assertion, nous choisissons au hasard n_1 véhicules, nous leur faisons emprunter un parcours routier donné avec un carburant ordinaire, et nous notons la consommation de chaque véhicule. Puis nous choisissons au hasard n_2 autres véhicules, nous rajoutons l'additif dans le réservoir, nous les faisons emprunter le même parcours routier, et nous mesurons les consommations. Pour tester la réduction de la consommation, nous confrontons les deux moyennes observées $\overline{x_1}$ et $\overline{x_2}$. Il s'agit dans ce cas d'un schéma de test sur échantillons indépendants.

On parle d'***échantillons appariés*** lorsque les individus sont liés d'un échantillon à l'autre. C'est le cas lorsque nous procédons à des mesures répétées sur les mêmes individus, de sorte qu'une correspondance 2 à 2 des observations puisse être établie, par exemple :

- mesures « avant-après » des mêmes individus,
- mesures de deux caractères sur les mêmes individus.

Exemple :

Reprenons l'exemple ci-dessus. Avec un peu de (mal)chance, il se peut que les petites berlines soient majoritaires dans le premier échantillon, les grosses berlines dans le second. Cela faussera

totalement les résultats, laissant à penser que l'additif a un effet néfaste sur les consommations. Le principe de l'appariement est d'écarter ce risque en créant des paires d'observations. Dans notre exemple, nous choisirons n véhicules au hasard dans la population : nous leur faisons faire le trajet avec un carburant ordinaire une première fois, puis nous rajoutons l'additif dans le réservoir, et nous leur refaisons parcourir le même chemin. L'écart entre les consommations sera à présent un bon indicateur des prétendus bénéfices introduits par l'additif.

Ce schéma « avant-après » est la forme la plus populaire de l'appariement. Il permet de réduire le risque de seconde espèce du test, c'est-à-dire d'augmenter la puissance du test.
L'appariement est en réalité plus large que le seul schéma « avant-après ». Il est efficace à partir du moment où nous réunissons les deux conditions suivantes :

- les individus dans chaque paire se ressemblent le plus possible, ou appartiennent à une même entité statistique ;
- les paires d'observations sont très différentes les unes des autres.

Ainsi, les individus ne sont pas forcément les mêmes dans les différents échantillons. Ils sont cependant liés d'une façon ou d'une autre.

Exemple :

Reprenons l'exemple ci-dessus. Nous souhaitons à présent comparer les mérites respectifs de 2 additifs concurrents. On ne peut pas mettre le premier additif, faire faire le trajet, puis ajouter le second additif. Quand bien même nous aurions vidangé le réservoir entre temps, nous ne savons pas si les effets du premier additif sur le moteur se sont estompés. Pour dépasser cet écueil, il serait plus judicieux d'échantillonner des paires de modèles identiques (marque, modèle, kilométrage), et de comparer leurs consommations deux à deux. Nous y gagnerons si les paires sont différentes les unes des autres, c'est-à-dire couvrant aussi largement que possible le spectre des véhicules existants (petites citadines, familiales, grosses berlines, …).

L'appariement (paired samples ou matched pairs samples en anglais) est une procédure très populaire en statistique. Elle permet une analyse fine des différences entre les populations.

Hypothèses soumises au test

L'énoncé du problème se résume en une alternative constituée de deux hypothèses H_0 et H_1, qui s'excluent mutuellement (H_1 est la négation de H_0) et qui sont appelées respectivement l'***hypothèse nulle***, ou fondamentale, et l'***hypothèse alternative***, ou contraire.

On choisira de privilégier une hypothèse (H_0) par rapport à l'autre (H_1). Ainsi, on choisira pour hypothèse nulle H_0 l'hypothèse à laquelle « on croit », celle dont le rejet est lourd de conséquences. C'est l'hypothèse nulle qui sera soumise au test et toute la démarche du test s'effectuera en considérant cette hypothèse comme vraie.

Tests

Les hypothèses à confronter H_0 et H_1 étant identifiées, leur validité est soumise à l'épreuve à l'aide d'un test d'hypothèses. Un test d'hypothèses est une règle de décision qui permet, sur la base des données observées et avec des risques d'erreur déterminés, d'accepter ou de rejeter l'hypothèse nulle H_0.

Il est important de comprendre que le but d'un test d'hypothèse est de déterminer si l'hypothèse nulle H_0 peut être rejetée, et non pas de déterminer si elle peut être acceptée !

Attention !

Le terme « accepter H_0 » est donc essentiellement un abus de langage, on devrait plutôt parler de « ne pas rejeter H_0 ». Il existe une dissymétrie importante dans les conclusions des tests. En effet, la décision d' « accepter H_0 » n'est pas équivalente à « H_0 est vraie et H_1 est fausse ». La conclusion d'un test sera surtout que « il n'y a pas d'évidence nette pour que H_0 soit fausse » (acceptation de H_0) ou que « il n'y a pas d'évidence nette pour que H_0 soit vraie » (rejet de H_0).

Fonction discriminante

Un test basé sur un échantillon de taille *n* est caractérisé par une région *R* de R^n appelée ***région critique***, ou ***région de rejet*** de l'hypothèse H_0. Le complémentaire *A* de *R* est appelé la ***région d'acceptation*** de H_0. La règle de décision d'un test est la suivante : si $\underline{x} = (x_1,...,x_n)$ est le vecteur des données observées, on décide de rejeter H_0 (et d'accepter H_1) si $\underline{x} \in R$, et on décide d'accepter H_0 si $\underline{x} \notin R$.

Dans la pratique, pour chaque test, on cherche à définir une fonction *D* de variables aléatoires (elle-même variable aléatoire donc), qui constituera la variable de décision, qu'on appelle ***statistique du test***, ou ***fonction discriminante***, et dont la loi de probabilité est connue sous l'hypothèse H_0.

On calcule alors la ***valeur observée*** *d* prise par la fonction discriminante *D* dans le cas particulier du ou des échantillons de valeurs en présence. Sous l'hypothèse H_0, les valeurs de *D* - et donc sa valeur observée *d* – « ne devraient pas trop s'éloigner » de sa moyenne μ_D.

La règle de décision d'un test devient alors la suivante :

On acceptera H_0 si la valeur observée *d* appartient à l'intervalle de confiance $[\mu_D - \delta_-, \mu_D + \delta_+]$ défini autour de la moyenne μ_D de la statistique de test *D* avec un niveau de confiance $1-\alpha$, et on rejettera H_0 sinon.

Dans la plupart des applications pratiques, le niveau de confiance choisi sera $1-\alpha = 95\,\%$ et donc $\alpha = 5\,\%$. Plus α est choisi grand, plus l'intervalle de confiance se resserre autour de μ_D, moins il est probable que *d* appartienne à cet intervalle, plus le test est exigeant envers l'acceptation de H_0. De façon équivalente, plus α est choisi grand, plus la frontière entre la région d'acceptation *A* et la région de rejet *R* se resserre sur *A* et réduit *A*, moins il est probable que $\underline{x} = (x_1,...,x_n)$ appartienne à *A* (pour $\alpha = 0\,\%$, *A* s'étend à l'infini et *R* s'annule ; pour $\alpha = 100\,\%$, *A* se réduit à un point et *R* s'étend à l'infini). En d'autres termes :

Plus α est choisi petit, plus le test aura une propension élevée à conclure à l'acceptation de l'hypothèse nulle H_0, et vice-versa.

Dans la cadre présent des tests statistiques, α est appelé le ***seuil de signification du test***.

Probabilité critique

En pratique, plutôt que de fixer un niveau α et d'en déduire la région d'acceptation de H_0 associée, il est plus intéressant de calculer la probabilité critique P_{crit} du test associée à la valeur observée *d* de la fonction

discriminante *D*. Cette probabilité critique donne une vision plus complète de la situation et a l'avantage de donner une mesure de crédibilité de l'hypothèse H_0.

La ***probabilité critique*** – très généralement dénommée *p-value* – est la valeur limite maximale de α pour laquelle H_0 est acceptée. C'est la situation limite dans laquelle l'échantillon de valeurs observées $\underline{x} = (x_1,...,x_n)$ se retrouve sur la frontière entre *A* et *R*.

Ainsi, la règle de décision finale est donnée quel que soit le niveau α choisi par :
- **si $\alpha > P_{crit}$: rejet de H_0 ,**
- **si $\alpha \leq P_{crit}$: acceptation de H_0 .**

La probabilité critique *P* fournit une ***mesure de la crédibilité de l'hypothèse nulle*** H_0 :
- une valeur très faible de la probabilité critique signifie que H_0 n'est pas valable,
- une valeur trop élevée permet de mettre en doute le caractère aléatoire de l'expérience et la fiabilité des données et des calculs.

Erreur, risque, puissance, robustesse

La règle de décision d'un test étant basée sur l'observation d'un échantillon et non sur la base d'une population totale, on n'est jamais sûr de l'exactitude de la conclusion : il y a donc toujours un risque d'erreur.
On a dit plus haut que, sous l'hypothèse H_0, les valeurs de D - et donc sa valeur observée d – « ne devraient pas trop s'éloigner » de sa moyenne μ_D. Or, il arrivera que le hasard de l'échantillonnage produise de façon fortuite des valeurs observées *d* qui seront situées « loin de la moyenne μ_D », *de facto* dans la zone de rejet de H_0, bien que H_0 soit vraie. En regard de l'intervalle de confiance adopté plus haut, on déduit que cette situation se produira avec une probabilité α.

L'erreur de première espèce consiste à rejeter H_0 à tort : le ***risque d'erreur de première espèce*** est précisément α, c'est le risque de rejeter H_0 alors que H_0 est vraie. On l'appelle aussi le ***niveau du test***.
L'erreur de deuxième espèce consiste à ne pas rejeter H_0 (et donc à rejeter H_1) à tort : le ***risque d'erreur de deuxième espèce*** est noté β, c'est le risque de ne pas rejeter H_0 (et donc de rejeter H_1) alors que H_0 est fausse (et donc que H_1 est vraie).
L'aptitude d'un test à rejeter H_0 alors que H_0 est fausse est donc $\eta = 1 - \beta$ et est appelée la ***puissance du test***.

On a vu que plus α est choisi petit, plus la région de rejet *R* diminue. Ainsi, tout sera une question de compromis : à vouloir commettre moins d'erreurs, on conclura plus rarement au rejet de H_0 (on a déjà insisté sur ce fait et il est important de le rappeler : le but du test d'hypothèse est de déterminer si l'hypothèse nulle H_0 peut être rejetée, et non pas de déterminer si elle peut être acceptée).

On s'efforce de construire des tests qui limitent les risques à des niveaux jugés acceptables. En règle générale, le seuil de signification α est choisi à priori (5% par défaut) et, compte tenu de cette contrainte, on cherche à construire les tests ayant la plus grande puissance η possible.

On montre notamment que :
- β augmente lorsque α diminue et vice-versa : les risques de première et de deuxième espèce α et β ne peuvent pas être minimisés en même temps,

- lorsque l'effectif des échantillons augmente, les risques diminuent.

On choisit donc un risque de première espèce α, et on réalise le test. Si la conclusion est qu'on rejette H_0, alors l'erreur de première espèce α sera le risque associée au rejet de H_0. Si la conclusion est qu'on accepte H_0, alors on calcule éventuellement l'erreur de deuxième espèce β qui donnera le risque associée à l'acceptation de H_0.

Remarque :

Pour quantifier le risque β, il faut connaître la loi de probabilité de la statistique D sous l'hypothèse H_1 et spécifier des valeurs particulières du paramètre soumis au test dans l'hypothèse H_1 que l'on suppose à présent vraie.

La ***robustesse*** d'une technique statistique représente sa sensibilité à des écarts aux hypothèses faites (hypothèse de normalité par exemple).
Certains tests sont robustes, c'est-à-dire qu'ils restent valables même si l'on s'écarte légèrement des hypothèses sous-jacentes initiales. Dans ce cas, il faut que la violation soit patente (par exemple, distributions très dissymétriques ou bimodales lorsque l'hypothèse de normalité est requise) pour que la procédure ne soit plus opératoire. D'autres tests en revanche ne sont pas robustes du tout.
En particulier, le théorème central limite nous a montré que la somme de variables aléatoires de même moyenne et variance tend vers la loi normale lorsque l'effectif augmente. De ce fait, les statistiques composées à partir d'une somme de variables aléatoires (la moyenne mais aussi la proportion) tendent vers la loi normale dès que l'effectif est suffisamment élevé, quelle que soit la distribution initiale sous-jacente. Ce résultat élargit considérablement le champ d'action des tests basés sur de telles statistiques de test et qui reposent sur l'hypothèse de normalité. Ce qui explique d'ailleurs leur popularité dans la pratique.

Test bilatéral et test unilatéral

Avant d'appliquer tout test statistique, il s'agit de bien définir le problème posé. En effet, selon les hypothèses formulées, on appliquera soit un test bilatéral, soit un test unilatéral.

Un ***test bilatéral*** s'applique quand on cherche une différence entre deux paramètres, ou entre un paramètre et une valeur donnée, sans se préoccuper du signe ou du sens de la différence. Dans ce cas, la zone de rejet de l'hypothèse nulle se fait de part et d'autre de la moyenne μ_D de la fonction discriminante.

Un ***test unilatéral*** s'applique quand on cherche à savoir si un paramètre est supérieur (ou inférieur) à un autre, ou à une valeur donnée. La zone de rejet de l'hypothèse nulle est située d'un seul côté de la moyenne μ_D de la fonction discriminante.

Pour chaque test H_0, on est donc amené à considérer trois versions : un test bilatéral H_{10} et deux tests unilatéraux H_{1+} et H_{1-}. Le choix de la version qui sera testée dépend de la question posée et doit être arrêté avant d'inspecter les données et de réaliser un quelconque test. Les conclusions varient en effet selon la version du test.

Remarque :

La puissance d'un test est fonction de la nature de H_1 : un test unilatéral est plus puissant qu'un test bilatéral.

L'hypothèse nulle sera exprimée par « $H_0 : d = \mu_D$ » et les hypothèses alternatives à tester comme suit :

- H_{10} : $d \neq \mu_D$, test bilatéral (« sous l'hypothèse H_0, la valeur observée *d* ne doit pas être trop différente de la moyenne μ_D de la statistique de test , sinon on est plutôt en faveur de l'hypothèse H_{10} »)
- H_{1+} : $d > \mu_D$, test unilatéral à droite (« sous l'hypothèse H_0, la valeur observée *d* ne doit pas être trop grande par rapport la moyenne μ_D de la statistique de test , sinon on est plutôt en faveur de l'hypothèse H_{1+} »)
- H_{1-} : $d < \mu_D$, test unilatéral à gauche (« sous l'hypothèse H_0, la valeur observée *d* ne doit pas être trop petite par rapport la moyenne μ_D de la statistique de test , sinon on est plutôt en faveur de l'hypothèse H_{1+} »)

On peut à présent formaliser précisément la définition de la probabilité critique comme suit :

- H_{10} (test bilatéral) : $P_{crit} = 2.min(\mathcal{P}(D \leq d), \mathcal{P}(d \leq D))$
  Dans le cas où *D* suit une distribution symétrique centrée sur *0* (lois normale centrée réduite, de Student, … mais pas du khi-deux, de Fisher, …), on peut écrire plus simplement $P_{crit} = \mathcal{P}(|d| \leq |D|)$
- H_{1-} (test unilatéral à gauche) : $P_{crit} = \mathcal{P}(D \leq d)$
- H_{1+} (test unilatéral à droite) : $P_{crit} = \mathcal{P}(d \leq D)$

Tests présentés

Il existe des dizaines de tests statistiques, d'objectifs extrêmement variés. Nous exposerons les plus pertinents dans le cadre du présent document.

Nous nous intéresserons tout d'abord aux principaux tests paramétriques sur un ou plusieurs échantillons indépendants issus de lois normales (tests de conformité et d'homogénéité). Nous présenterons ensuite trois tests de cohérence pour la détection de valeurs incohérentes (tests d'homogénéité). Nous aborderons enfin les tests de normalité (tests d'adéquation).

Hormis les tests de normalité, les tests non paramétriques ne seront pas développés. On trouvera dans la littérature les principaux tests non paramétriques (test du khi-deux de Pearson, test d'ajustement de Kolmogorov-Smirnov, tests de Wilcoxon, …).

Ci-dessous :

- $N_\alpha(0,1)$ ($= -N_{1-\alpha}(0,1)$ par symétrie de la loi normale) désigne le quantile d'ordre α de la loi normale centrée réduite.
- $T_\alpha(k)$ ($= -T_{1-\alpha}(k)$ par symétrie de la loi de Student) désigne le quantile d'ordre α de la loi de Student à *k* degrés de liberté.
- $\chi^2_\alpha(k)$ désigne le quantile d'ordre α de la loi du khi-deux à *k* degrés de liberté.
- $F_\alpha(k,l)$ désigne le quantile d'ordre α de la loi de Fisher avec k et l degrés de liberté.

Excel :

LOI.NORMALE.N(x, 0, 1, true) donne la fonction de répartition $F(x) = \int_{-\infty}^{x} p(x)dx = \mathcal{P}(X \leq x)$

LOI.STUDENT.N(x, k, true) donne la fonction de répartition $F(x) = \int_{-\infty}^{x} p(x)dx = \mathcal{P}(X \leq x)$

LOI.KHIDEUX.N(x, k, true) donne la fonction de répartition $F(x) = \int_{-\infty}^{x} p(x)dx = \mathcal{P}(X \leq x)$

LOI.F.N(x, k, l, true) donne la fonction de répartition $F(x) = \int_{-\infty}^{x} p(x)dx = \mathcal{P}(X \leq x)$

2.6.2 Tests paramétriques

En raison de son omniprésence, la loi normale (voir théorème central limite) a naturellement fait l'objet de plus d'attentions que les autres lois. Ainsi dispose-t-on notamment des tests remarquables exposées ci-dessous et relatifs à la loi normale.

Les tests présentés dans ce chapitre sont la conséquence directe des propriétés établies plus haut concernant l'estimation d'échantillons issus de lois normales.

Les tests dont la statistique suit une loi de Student sous l'hypothèse H_0 sont dénommés « tests de Student ». Les tests dont la statistique suit une loi de Fisher sous l'hypothèse H_0 sont dénommés « tests de Fisher ».

ATTENTION ! Les propriétés de ce chapitre ne s'appliquent que pour des échantillons issus de lois normales. Cependant, l'hypothèse sous-jacente de normalité n'est pas aussi restrictive qu'on peut le penser, certaines conditions favorisent la robustesse des tests comme on l'a déjà laissé entendre plus haut. Lorsque des informations concernant la robustesse des tests seront disponibles, nous les mentionnerons.

2.6.2.1 Tests de comparaison d'une moyenne à une valeur de référence μ_0

Hypothèses à tester

H_0 : $\mu = \mu_0$
H_{10} : $\mu \neq \mu_0$, test bilatéral V_0.
H_{1+} : $\mu > \mu_0$, test unilatéral à droite V_+.
H_{1-} : $\mu < \mu_0$, test unilatéral à gauche V_-.

Variance σ^2 connue

Soit une variable aléatoire *X* de loi normale $N(\mu,\sigma^2)$ où σ^2 est connu. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,...,x_i,...,x_n$ de cette variable *X*, de moyenne estimée $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Alors on montre que sous l'hypothèse H_0 la fonction discriminante $D = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ suit la loi normale centrée réduite $N(0,1)$.

V_0 (test bilatéral)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas trop s'éloigner de 0. La région d'acceptation est un intervalle de la forme $\left[-N_{1-\alpha/2}(0,1),+N_{1-\alpha/2}(0,1)\right]$.

V_+ (test unilatéral à droite)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop grandes, sinon on est plutôt en faveur de l'hypothèse H_{1+} : $\mu > \mu_0$. La région d'acceptation est un intervalle de la forme $\left]-\infty,+N_{1-\alpha}(0,1)\right]$.

V_- (test unilatéral à gauche)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop petites, sinon on est plutôt en faveur de l'hypothèse H_{1-} : $\mu < \mu_0$. La région d'acceptation est un intervalle de la forme $\left[-N_{1-\alpha}(0,1),+\infty\right[$.

Variance σ^2 inconnue

Soit une variable aléatoire *X* de loi normale $N(\mu,\sigma^2)$ où σ^2 est inconnu. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,...,x_i,...,x_n$ de cette variable *X*, de moyenne estimée $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et de variance estimée $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$. Alors on montre que sous l'hypothèse H_0 la fonction discriminante $D = \frac{\bar{x} - \mu_0}{s_X/\sqrt{n}}$ suit la loi de Student $T(n-1)$.

V_0 (test bilatéral)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas trop s'éloigner de 0. La région d'acceptation est un intervalle de la forme $\left[-T_{1-\alpha/2}(n-1),+T_{1-\alpha/2}(n-1)\right]$.

V_+ (test unilatéral à droite)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop grandes, sinon on est plutôt en faveur de l'hypothèse H_{1+} : $\mu > \mu_0$. La région d'acceptation est un intervalle de la forme $\left]-\infty,+T_{1-\alpha}(n-1)\right]$.

V_- (test unilatéral à gauche)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop petites, sinon on est plutôt en faveur de l'hypothèse H_{1-} : $\mu < \mu_0$. La région d'acceptation est un intervalle de la forme $\left[-T_{1-\alpha}(n-1),+\infty\right[$.

2.6.2.2 Tests de comparaison d'une variance à une valeur de référence σ_0^2

Hypothèses à tester

$H_0 : \sigma^2 = \sigma_0^2$
$H_{10} : \sigma^2 \neq \sigma_0^2$, test bilatéral V_0.
$H_{1+} : \sigma^2 > \sigma_0^2$, test unilatéral à droite V_+.
$H_{1-} : \sigma^2 < \sigma_0^2$, test unilatéral à gauche V_-.

Moyenne μ connue

Soit une variable aléatoire *X* de loi normale $N(\mu, \sigma^2)$ où μ est connu. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,...,x_i,...,x_n$ de cette variable *X*, de variance estimée $s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$. Alors on montre que sous l'hypothèse H_0 la fonction discriminante $D = \frac{n s_X^2}{\sigma_0^2}$ suit la loi du khi-deux $\chi^2(n)$.

V_0 (test bilatéral)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas trop s'éloigner de son espérance *n*. La région d'acceptation est un intervalle de la forme $\left[\chi^2_{\alpha/2}(n), \chi^2_{1-\alpha/2}(n)\right]$.

V_+ (test unilatéral à droite)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop grandes, sinon on est plutôt en faveur de l'hypothèse $H_{1+} : \sigma^2 > \sigma_0^2$. La région d'acceptation est un intervalle de la forme $\left[0, \chi^2_{1-\alpha}(n)\right]$.

V_- (test unilatéral à gauche)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop petites, sinon on est plutôt en faveur de l'hypothèse $H_{1-} : \sigma^2 < \sigma_0^2$. La région d'acceptation est un intervalle de la forme $\left[\chi^2_{\alpha}(n), +\infty\right[$.

Moyenne μ inconnue

Soit une variable aléatoire *X* de loi normale $N(\mu, \sigma^2)$ où μ est inconnu. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,...,x_i,...,x_n$ de cette variable *X*, de variance estimée $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Alors on montre que sous l'hypothèse H_0 la fonction discriminante $D = \frac{(n-1)s_X^2}{\sigma_0^2}$ suit la loi du khi-deux $\chi^2(n-1)$.

V_0 (test bilatéral)

Sous l'hypothèse H_0, les valeurs de D ne doivent pas trop s'éloigner de son espérance $n-1$. La région d'acceptation est un intervalle de la forme $\left[\chi^2_{\alpha/2}(n-1), \chi^2_{1-\alpha/2}(n-1)\right]$.

V_+ (test unilatéral à droite)

Sous l'hypothèse H_0, les valeurs de D ne doivent pas être trop grandes, sinon on est plutôt en faveur de l'hypothèse H_{1+} : $\sigma^2 > \sigma_0^2$. La région d'acceptation est un intervalle de la forme $\left[0, \chi^2_{1-\alpha}(n-1)\right]$.

V_- (test unilatéral à gauche)

Sous l'hypothèse H_0, les valeurs de D ne doivent pas être trop petites, sinon on est plutôt en faveur de l'hypothèse H_{1-} : $\sigma^2 < \sigma_0^2$. La région d'acceptation est un intervalle de la forme $\left[\chi^2_{\alpha}(n-1), +\infty\right[$.

2.6.2.3 Tests de comparaison de deux moyennes μ_X et μ_Y

Hypothèses à tester

H_0 : $\mu_X = \mu_Y$
H_{10} : $\mu_X \neq \mu_Y$, test bilatéral V_0.
H_{1+} : $\mu_X > \mu_Y$, test unilatéral à droite V_+.
H_{1-} : $\mu_X < \mu_Y$, test unilatéral à gauche V_-.

Robustesse

Un écart modéré par rapport à la normalité des distributions ne perturbe pas (trop) le test de comparaison de moyennes, surtout si les distributions restent symétriques. Si les distributions sont dissymétriques, mais qu'elles le sont de la même manière pour chaque échantillon, le test s'applique quand même. Lorsque l'effectif est élevé, le théorème central limite balaye toutes les hésitations.
Dans la section « Echantillons indépendants – Variances σ_X^2 et σ_Y^2 inconnues mais égales », en toute rigueur le test de comparaison de moyennes devrait être précédé par un test de comparaison de variances. En effet, nous émettons l'hypothèse que les variances sont identiques, et elle doit être vérifiée au préalable. En pratique, il semble que ce ne soit pas nécessaire dans la grande majorité des cas, sauf violation flagrante visible dans les variances estimées des échantillons. Un écart modéré par rapport à cette hypothèse n'est pas problématique, ceci d'autant plus que les effectifs sont équilibrés ($n_X \approx n_Y$).
En revanche, lorsque les effectifs sont déséquilibrés, on privilégiera le test de la section « Echantillons indépendants – Variances σ_X^2 et σ_Y^2 inconnues et différentes ». Certains auteurs précisent à ce sujet que l'on devrait toujours utiliser la variante pour variances inégales dès que n_X et n_Y sont très différents, quand bien même le ratio entre la plus grande et la plus petite variance n'excéderait pas *1,5*.

Echantillons indépendants – Variances σ_X^2 et σ_Y^2 connues

Soit deux variables aléatoires indépendantes *X* et *Y* de lois normales $N(\mu_X, \sigma_X^2)$ et $N(\mu_Y, \sigma_Y^2)$ où σ_X^2 et σ_Y^2 sont connus. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,...,x_i,...,x_n$ de cette variable *X*, de moyenne estimée $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Soit tout échantillon de *m* valeurs indépendantes $(m>1)$ $y_1,...,y_i,...,y_m$ de cette variable *Y*, de moyenne estimée $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$. Alors on montre que sous l'hypothèse H_0 la fonction discriminante $D = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$ suit la loi normale centrée réduite $N(0,1)$.

V_0 (test bilatéral)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas trop s'éloigner de 0. La région d'acceptation est un intervalle de la forme $\left]-N_{1-\alpha/2}(0,1), +N_{1-\alpha/2}(0,1)\right[$.

V_+ (test unilatéral à droite)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop grandes, sinon on est plutôt en faveur de l'hypothèse H_{1+} : $\mu_X > \mu_Y$. La région d'acceptation est un intervalle de la forme $\left]-\infty,+N_{1-\alpha}(0,1)\right]$.

V_- (test unilatéral à gauche)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop petites, sinon on est plutôt en faveur de l'hypothèse H_{1-} : $\mu_X < \mu_Y$. La région d'acceptation est un intervalle de la forme $\left[-N_{1-\alpha}(0,1),+\infty\right[$.

Echantillons indépendants – Variances σ_X^2 et σ_Y^2 inconnues mais égales

Soit deux variables aléatoires indépendantes *X* et *Y* de lois normales $N(\mu_X,\sigma_X^2)$ et $N(\mu_Y,\sigma_Y^2)$ où σ_X^2 et σ_Y^2 sont inconnus mais égaux. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,...,x_i,...,x_n$ de cette variable *X*, de moyenne estimée $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ et de variance estimée $s_X^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$. Soit tout échantillon de *m* valeurs indépendantes $(m>1)$ $y_1,...,y_i,...,y_m$ de cette variable *Y*, de moyenne estimée $\bar{y}=\frac{1}{m}\sum_{i=1}^{m}y_i$ et de variance estimée $s_Y^2=\frac{1}{m-1}\sum_{i=1}^{m}(y_i-\bar{y})^2$. Soit $\nu=n+m-2$ et l'estimateur de la variance commune $s^2=\frac{1}{\nu}\left((n-1)s_X^2+(m-1)s_Y^2\right)$. Alors on montre que sous l'hypothèse H_0 la fonction discriminante $D=\frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n}+\frac{1}{m}}}$ suit la loi de Student $T(\nu)$.

V_0 (test bilatéral)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas trop s'éloigner de 0. La région d'acceptation est un intervalle de la forme $\left[-T_{1-\alpha/2}(\nu),+T_{1-\alpha/2}(\nu)\right]$.

V_+ (test unilatéral à droite)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop grandes, sinon on est plutôt en faveur de l'hypothèse H_{1+} : $\mu_X > \mu_Y$. La région d'acceptation est un intervalle de la forme $\left]-\infty,+T_{1-\alpha}(\nu)\right]$.

V_- (test unilatéral à gauche)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop petites, sinon on est plutôt en faveur de l'hypothèse H_{1-} : $\mu_X < \mu_Y$. La région d'acceptation est un intervalle de la forme $\left[-T_{1-\alpha}(\nu),+\infty\right[$.

Echantillons indépendants – Variances σ_X^2 et σ_Y^2 inconnues et différentes

Soit deux variables aléatoires indépendantes *X* et *Y* de lois normales $N(\mu_X,\sigma_X^2)$ et $N(\mu_Y,\sigma_Y^2)$ où σ_X^2 et σ_Y^2 sont inconnus et différents. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,...,x_i,...,x_n$ de cette

variable *X*, de moyenne estimée $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ et de variance estimée $s_X^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$. Soit tout échantillon de *m* valeurs indépendantes $(m>1)$ $y_1,...,y_i,...,y_m$ de cette variable *Y*, de moyenne estimée $\bar{y}=\frac{1}{m}\sum_{i=1}^{m}y_i$ et de variance estimée $s_Y^2=\frac{1}{m-1}\sum_{i=1}^{m}\left(y_i-\bar{y}\right)^2$. Soit $\nu=\frac{\left(\frac{s_X^2}{n}+\frac{s_Y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{s_X^2}{n}\right)^2+\frac{1}{m-1}\left(\frac{s_Y^2}{m}\right)^2}$ (formule de Welsch-Satterthwaite). Alors on montre que sous l'hypothèse H_0 la fonction discriminante $D=\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_X^2}{n}+\frac{s_Y^2}{m}}}$ suit la loi de Student $T(\nu)$.

Remarque :

La formule de Welsch-Statterthwaite générera pour ν un nombre réel positif, dont il sera question en pratique de ne garder que la partie entière. Sous l'environnement Excel, la fonction LOI.STUDENT.INVERSE.N accepte comme argument des degrés de liberté non-entiers mais n'en retient systématiquement que la partie entière.

V_0 (test bilatéral)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas trop s'éloigner de 0. La région d'acceptation est un intervalle de la forme $\left[-T_{1-\alpha/2}(\nu),+T_{1-\alpha/2}(\nu)\right]$.

V_+ (test unilatéral à droite)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop grandes, sinon on est plutôt en faveur de l'hypothèse H_{1+} : $\mu_X>\mu_Y$. La région d'acceptation est un intervalle de la forme $\left]-\infty,+T_{1-\alpha}(\nu)\right]$.

V_- (test unilatéral à gauche)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop petites, sinon on est plutôt en faveur de l'hypothèse H_{1-} : $\mu_X<\mu_Y$. La région d'acceptation est un intervalle de la forme $\left[-T_{1-\alpha}(\nu),+\infty\right[$.

Echantillons appariés – Variances σ_X^2 et σ_Y^2 et covariance σ_{XY} connues

Dans le cas d'échantillons appariés, $n=m$. Le test se ramène à un test de conformité à une moyenne nulle de l'échantillon $z_1,...,z_i,...,z_n$, avec $z_i=x_i-y_i$.

Soit deux variables aléatoires appariées *X* et *Y* de lois normales $N\left(\mu_X,\sigma_X^2\right)$ et $N\left(\mu_Y,\sigma_Y^2\right)$ où σ_X^2, σ_Y^2 et σ_{XY} sont connus. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,...,x_i,...,x_n$ de cette variable *X*, de moyenne estimée $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $y_1,...,y_i,...,y_n$ de

cette variable *Y*, de moyenne estimée $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Alors on montre que sous l'hypothèse H_0 la fonction discriminante $D = \frac{\bar{z}}{\sigma_Z / \sqrt{n}}$ suit la loi normale centrée réduite $N(0,1)$, avec $\bar{z} = \bar{x} - \bar{y}$ et $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}$ où σ_{XY} désigne la covariance de *X* et *Y*.

Attention !

Les variables ne sont pas indépendantes ici ! La variance ne peut donc pas se résumer à la somme des variances des variables individuelles. Il faut prendre en compte la covariance.

V_0 (test bilatéral)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas trop s'éloigner de 0. La région d'acceptation est un intervalle de la forme $\left]-N_{1-\alpha/2}(0,1), +N_{1-\alpha/2}(0,1)\right[$.

V_+ (test unilatéral à droite)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop grandes, sinon on est plutôt en faveur de l'hypothèse H_{1+} : $\mu_X > \mu_Y$. La région d'acceptation est un intervalle de la forme $\left]-\infty, +N_{1-\alpha}(0,1)\right[$.

V_- (test unilatéral à gauche)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop petites, sinon on est plutôt en faveur de l'hypothèse H_{1-} : $\mu_X < \mu_Y$. La région d'acceptation est un intervalle de la forme $\left]-N_{1-\alpha}(0,1), +\infty\right[$.

Echantillons appariés – Variances σ_X^2 et σ_Y^2 inconnues

Dans le cas d'échantillons appariés, $n = m$. Le test se ramène à un test de conformité à une moyenne nulle de l'échantillon $z_1, \ldots, z_i, \ldots, z_n$, avec $z_i = x_i - y_i$.

Soit deux variables aléatoires appariées *X* et *Y* de lois normales $N(\mu_X, \sigma_X^2)$ et $N(\mu_Y, \sigma_Y^2)$ où σ_X^2 et σ_Y^2 sont inconnus. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1, \ldots, x_i, \ldots, x_n$ de cette variable *X*, de moyenne estimée $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et de variance estimée $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $y_1, \ldots, y_i, \ldots, y_n$ de cette variable *Y*, de moyenne estimée $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ et de variance estimée $s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$. Alors on montre que sous l'hypothèse H_0 la fonction discriminante $D = \frac{\bar{z}}{s_Z / \sqrt{n}}$ suit la loi de Student $T(n-1)$, avec $\bar{z} = \bar{x} - \bar{y}$ et $s_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2$.

Attention !

Les variables ne sont pas indépendantes ici ! La variance ne peut donc pas se résumer à la somme des variances des variables individuelles. Il faudrait prendre en compte la covariance, soit $s_Z^2 = s_X^2 + s_Y^2 - 2s_{XY}$. *L'estimation directe de* s_Z^2 *par* $s_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(z_i - \bar{z}\right)^2$ *permet cependant de s'affranchir de cette considération.*

V_0 (test bilatéral)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas trop s'éloigner de 0. La région d'acceptation est un intervalle de la forme $\left]-T_{1-\alpha/2}(n-1),+T_{1-\alpha/2}(n-1)\right[$.

V_+ (test unilatéral à droite)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop grandes, sinon on est plutôt en faveur de l'hypothèse H_{1+} : $\mu_X > \mu_Y$. La région d'acceptation est un intervalle de la forme $\left]-\infty,+T_{1-\alpha}(n-1)\right[$.

V_- (test unilatéral à gauche)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop petites, sinon on est plutôt en faveur de l'hypothèse H_{1-} : $\mu_X < \mu_Y$. La région d'acceptation est un intervalle de la forme $\left]-T_{1-\alpha}(n-1),+\infty\right[$.

2.6.2.4 Tests de comparaison de deux variances σ_X^2 et σ_Y^2

Hypothèses à tester

H_0 : $\sigma_X^2 = \sigma_Y^2$
H_{10} : $\sigma_X^2 \neq \sigma_Y^2$, test bilatéral V_0.
H_{1+} : $\sigma_X^2 > \sigma_Y^2$, test unilatéral à droite V_+.
H_{1-} : $\sigma_X^2 < \sigma_Y^2$, test unilatéral à gauche V_-.

Robustesse

Certains tests sont très sensibles à la normalité sous-jacente, d'autres en revanche sont plus robustes.

Tests pour échantillons indépendants dont la statistique suit la loi de Fisher sous l'hypothèse H_0

Les tests de Fisher ne sont pas robustes du tout. Un écart, même minime, par rapport à la distribution normale fausse les résultats. Il faut absolument s'assurer du caractère gaussien des échantillons avant de les utiliser. Ce qui en limite considérablement la portée. Dans la pratique, on se tournera avantageusement vers le test de Levene ou de Brown-Forsythe (tests de comparaison de *k* variances, également applicables dans le cas présent où $k = 2$, voir plus bas).

Tests pour échantillons appariés dont la statistique suit la loi de Student sous l'hypothèse H_0

Le test de Pitman est relativement peu robuste par rapport à l'hypothèse sous-jacente de normalité bivariée du couple (U,V). Cela limite quelque peu sa portée.

Echantillons indépendants – Moyennes μ_X et μ_Y connues

Soit deux variables aléatoires indépendantes *X* et *Y* de lois normales $N(\mu_X, \sigma_X^2)$ et $N(\mu_Y, \sigma_Y^2)$ où μ_X et μ_Y sont connus. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,...,x_i,...,x_n$ de cette variable *X*, de variance estimée $s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^2$. Soit tout échantillon de *m* valeurs indépendantes $(m>1)$ $y_1,...,y_i,...,y_m$ de cette variable *Y*, de variance estimée $s_Y^2 = \frac{1}{m}\sum_{i=1}^{m}(y_i - \mu_Y)^2$. Alors on montre que sous l'hypothèse H_0 la fonction discriminante $D = \frac{s_x^2}{s_y^2}$ suit la loi de Fisher $F(n,m)$.

V_0 (test bilatéral)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas trop s'éloigner de 1. La région d'acceptation est un intervalle de la forme $\left[F_{\alpha/2}(n,m), F_{1-\alpha/2}(n,m)\right]$.

V_+ (test unilatéral à droite)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop grandes, sinon on est plutôt en faveur de l'hypothèse H_{1+} : $\sigma_X^2 > \sigma_Y^2$. La région d'acceptation est un intervalle de la forme $\left[0, F_{1-\alpha}(n,m)\right]$.

V_- (test unilatéral à gauche)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop petites, sinon on est plutôt en faveur de l'hypothèse H_{1-} : $\sigma_X^2 < \sigma_Y^2$. La région d'acceptation est un intervalle de la forme $\left[F_\alpha(n,m), +\infty\right[$.

Echantillons indépendants – Moyennes μ_X et μ_Y inconnues

Soit deux variables aléatoires indépendantes *X* et *Y* de lois normales $N\left(\mu_X, \sigma_X^2\right)$ et $N\left(\mu_Y, \sigma_Y^2\right)$ où μ_X et μ_Y sont inconnus. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,...,x_i,...,x_n$ de cette variable *X*, de moyenne estimée $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et de variance estimée $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$. Soit tout échantillon de *m* valeurs indépendantes $(m>1)$ $y_1,...,y_i,...,y_m$ de cette variable *Y*, de moyenne estimée $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$ et de variance estimée $s_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(y_i - \bar{y}\right)^2$. Alors on montre que sous l'hypothèse H_0 la fonction discriminante $D = \frac{s_x^2}{s_y^2}$ suit la loi de Fisher $F(n-1, m-1)$.

V_0 (test bilatéral)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas trop s'éloigner de 1. La région d'acceptation est un intervalle de la forme $\left[F_{\alpha/2}(n-1,m-1), F_{1-\alpha/2}(n-1,m-1)\right]$.

V_+ (test unilatéral à droite)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop grandes, sinon on est plutôt en faveur de l'hypothèse H_{1+} : $\sigma_X^2 > \sigma_Y^2$. La région d'acceptation est un intervalle de la forme $\left[0, F_{1-\alpha}(n-1,m-1)\right]$.

V_- (test unilatéral à gauche)

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop petites, sinon on est plutôt en faveur de l'hypothèse H_{1-} : $\sigma_X^2 < \sigma_Y^2$. La région d'acceptation est un intervalle de la forme $\left[F_\alpha(n-1,m-1), +\infty\right[$.

Echantillons appariés – Moyennes μ_X et μ_Y inconnues

Dans le cas d'échantillons appariés, $n = m$. Le test le plus utilisé est le test de Pitman.

Soit deux variables aléatoires appariées *X* et *Y* de lois normales $N\left(\mu_X, \sigma_X^2\right)$ et $N\left(\mu_Y, \sigma_Y^2\right)$ où μ_X et μ_Y sont inconnus. Soit tout échantillon de *n* valeurs indépendantes $(n>1)$ $x_1,...,x_i,...,x_n$ de cette variable *X*, de

moyenne estimée $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ et de variance estimée $s_X^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$. Soit tout échantillon de n valeurs indépendantes $(n>1)$ $y_1,...,y_i,...,y_n$ de cette variable Y, de moyenne estimée $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$ et de variance estimée $s_Y^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$. Alors on montre que sous l'hypothèse H_0 la fonction discriminante $D=\frac{r_{UV}}{\sqrt{\frac{1-r_{UV}^2}{n-2}}}$ suit la loi de Student $T(n-2)$, avec $r_{UV}=\frac{s_{UV}}{s_U^2 s_V^2}$ le coefficient de corrélation estimé des variables aléatoires $U=X+Y$ et $V=X-Y$ calculé comme suit : $r_{UV}=\frac{\sum_{i=1}^{n}\left(u_i-\bar{u}\right)\left(v_i-\bar{v}\right)}{\sqrt{\sum_{i=1}^{n}\left(u_i-\bar{u}\right)^2\sum_{i=1}^{n}\left(v_i-\bar{v}\right)^2}}$.

Remarques :

Ce test repose sur le fait que la transformation $U=X+Y$ et $V=X-Y$ a pour conséquences

$$\begin{cases} r_{UV}<0 \Leftrightarrow \frac{\sigma_X^2}{\sigma_Y^2}<1 \\ r_{UV}=0 \Leftrightarrow \frac{\sigma_X^2}{\sigma_Y^2}=1 \\ r_{UV}>0 \Leftrightarrow \frac{\sigma_X^2}{\sigma_Y^2}>1 \end{cases}$$

V_0 (test bilatéral)

Sous l'hypothèse H_0, les valeurs de D ne doivent pas trop s'éloigner de 0. La région d'acceptation est un intervalle de la forme $\left[-T_{1-\alpha/2}(n-2),+T_{1-\alpha/2}(n-2)\right]$.

V_+ (test unilatéral à droite)

Sous l'hypothèse H_0, les valeurs de D ne doivent pas être trop grandes, sinon on est plutôt en faveur de l'hypothèse H_{1+} : $\sigma_X^2>\sigma_Y^2$. La région d'acceptation est un intervalle de la forme $\left]-\infty,+T_{1-\alpha}(n-2)\right]$.

V_- (test unilatéral à gauche)

Sous l'hypothèse H_0, les valeurs de D ne doivent pas être trop petites, sinon on est plutôt en faveur de l'hypothèse H_{1-} : $\sigma_X^2<\sigma_Y^2$. La région d'acceptation est un intervalle de la forme $\left[-T_{1-\alpha}(n-2),+\infty\right[$.

2.6.2.5 Test de comparaison de *k* moyennes μ_i (ANOVA)

Objet

Le test de comparaison de *k* moyennes est appelé « **analyse de variance** » et est connu sous l'acronyme « **ANOVA** ». Il est utilisable pour tout $k \geq 2$, le cas particulier $k = 2$ dégénérant en le test de comparaison de deux moyennes vu plus haut (cas des variances égales).

Nous présenterons ici le cas le plus fréquent de l'analyse de variance à un facteur contrôlé.
On étudie dans ce cas l'influence d'un facteur *A* sur une variable *X*. Le facteur *A* peut prendre *k* valeurs – appelées modalités ou niveaux ou traitement – et on évalue l'effet de ces modalités sur la moyenne de *X*.
Les applications sont nombreuses. Il peut s'agir d'une véritable comparaison (comparer la teneur en sel (*X*) réelle de différentes marques (*A*) de sandwich, comparer le salaire moyen (*X*) des étudiants selon les filières (*A*) à la sortie de l'université, …). Il peut aussi s'agir d'analyser l'effet d'un facteur représenté par une variable catégorielle sur une variable continue (évaluer l'impact des différentes méthodes (*A*) d'enseignement sur les notes (*X*) des étudiants, comparer les émissions polluantes (*X*) des véhicules selon le type de filtre (*A*) incorporé dans les pots d'échappement, …).

Hypothèses à tester

H_0 : $\mu_i = \mu_j$ pour tout $i \in (1,..,k)$ et tout $j \in (1,..,k)$, $i \neq j$

H_1 : les moyennes μ_i ne sont pas toutes égales, i.e. au-moins deux moyennes sont différentes.

On peut interpréter ces hypothèses comme suit :
« Aucune modalités du facteur d'influence *A* n'a un effet différent des autres » (ou « Toutes les modalités du facteur d'influence *A* produisent le même résultat ») vs. « Une des modalités du facteur d'influence *A* au-moins se démarque d'une autre ».

Remarque :

Parler de version de test (bilatéral, unilatéral à droite ou unilatéral à gauche) n'a plus de sens dans les tests de comparaison de $k > 2$ échantillons (seule une comparaison de deux valeurs peut être envisagée selon ces trois cas).

Robustesse

L'ANOVA est très robuste par rapport à l'hypothèse de normalité. Il suffit que les distributions des échantillons aient des formes similaires, même asymétriques.

De manière générale, on gagne toujours à équilibrer les échantillons, c'est à dire faire en sorte que $n_1 = ... = n_k$. Cela permet aussi de réduire le risque de deuxième espèce du test. En pratique, on évitera également d'utiliser le test lorsque les échantillons sont trop petits.

En toute rigueur l'analyse de variance devrait être précédée par un test de comparaison de variances. En effet, nous émettons l'hypothèse que les variances sont identiques, et elle doit être vérifiée au préalable.
Dans les faits, cette hypothèse semble relativement secondaire lorsque les effectifs des échantillons ne sont pas trop différents entre eux. Le test est d'autant plus robuste que les échantillons sont équilibrés. Dans ce cas, la variance conditionnelle la plus élevée peut être jusqu'à 4 fois supérieure à la plus petite variance.

Lorsque les deux obstacles sont cumulés – les variances sont manifestement hétérogènes et les effectifs sont déséquilibrés – on notera qu'il existe une adaptation de l'ANOVA, dite de « Welsch » (voir à ce sujet le

cours sur les tests paramétriques de Ricco Rakotomalala, Université Lumière Lyon 2). Cette adaptation nécessite en revanche que les distributions sous-jacentes soient gaussiennes.

Enfin, lorsque les conditions d'application du test ne sont pas satisfaites, il existe des techniques de transformation qui permettent de normaliser les distributions et de stabiliser les variances. Des tests non paramétriques sont également disponibles.

Echantillons indépendants

Exemples :

- *Reprenons l'exemple des additifs pour carburants. Nous souhaitons maintenant comparer $k=5$ marques différentes. Nous choisissons au hasard $n_1,...,n_5$ véhicules, nous rajoutons l'additif dans le réservoir, nous les faisons emprunter le même parcours routier, et nous mesurons les consommations. Pour tester la réduction la consommation, nous confrontons les cinq moyennes observées $\overline{x_1}$ à $\overline{x_5}$. Il s'agit dans ce cas d'un schéma de test sur 5 échantillons indépendants.*
- *On veut savoir si la quantité de nitrates varie d'une station à l'autre le long d'une rivière. Pour cela, on prélève dans 3 stations différentes ($k=3$), en 10 points de prélèvement aléatoires pour chacune d'elles, une certaine quantité d'eau. Le nombre de points de prélèvement pourrait même être différent d'une station à l'autre. Il s'agit dans ce cas d'un schéma de test sur 3 échantillons indépendants.*

Analyse de variance

Soit *k* variables aléatoires indépendantes X_i de lois normales $N(\mu_i,\sigma^2)$ (leur variance est donc supposée identique), avec $i=1,...,k$. Soit tout ensemble de *k* échantillons de n_i valeurs indépendantes $(n_i>1)$ $x_{i1},...,x_{in_i}$ de ces variables X_i, de moyennes intra-échantillon estimées $\overline{x_i}=\frac{1}{n_i}\sum_{j=1}^{n_i}x_{ij}$, avec $i=1,...,k$, d'effectif total $n=\sum_{i=1}^{k}n_i$ et de moyenne totale estimée $\overline{x}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}=\frac{1}{n}\sum_{i=1}^{k}n_i\overline{x_i}$.

L'équation de l'analyse de variance s'écrit :

$$SCT=SCE+SCR$$

où

- $SCT=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{x}\right)^2$ est la somme des carrés des écarts totaux,
- $SCE=\sum_{i=1}^{k}n_i\left(\overline{x_i}-\overline{x}\right)^2$ est la somme des carrés des écarts expliqués,
- $SCR=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{x_i}\right)^2$ est la somme des carrés des écarts résiduels.

Alors on montre que sous l'hypothèse H_0 la fonction discriminante $D=\frac{\frac{SCE}{k-1}}{\frac{SCR}{n-k}}$ suit la loi de Fisher $F(k-1,n-k)$.

Test

Lorsque H_1 est vraie, certaines moyennes μ_i ne sont pas égales. On montre alors que *SCE* (la somme des carrés des écarts expliqués) a tendance à prendre de grandes valeurs, quels que soient les signes des inégalités entre moyennes, et donc que la statistique de test *D* ne peut qu'augmenter.

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop grandes, sinon on est plutôt en faveur de l'hypothèse H_1 : les moyennes μ_i ne sont pas toutes égales, i.e. au-moins deux moyennes sont différentes. La région d'acceptation est un intervalle de la forme $\left[0, F_{1-\alpha}(k-1, n-k)\right]$.

Echantillons appariés

Dans le cas d'échantillons appariés, $n_i = n$ pour tout $i \in (1,..,k)$.

Le test est basé sur les plans d'expériences en **blocs aléatoires complets** (en anglais : randomized blocks)

Exemple :

Reprenons l'exemple des additifs pour carburants. Nous souhaitons maintenant comparer $k=5$ marques différentes. De la même manière que précédemment, nous constituons n unités statistiques (n blocs), chaque unité étant composée de 5 véhicules. Nous attribuons totalement au hasard la population à l'intérieur de chaque bloc. Plus les individus à l'intérieur d'un bloc se ressemblent, plus nous réduisons la variabilité intra-blocs, en revanche nous avons tout intérêt à élaborer des blocs aussi différents que possible les uns des autres. Il s'agit dans ce cas d'un schéma de test sur 5 échantillons appariés.

L'appariement peut également faire référence aux mesures répétées (en anglais : repeated measures). Il s'agit en quelque sorte d'une généralisation du canevas « avant-après » introduit en exemple du cas de deux populations.

Exemple :

Reprenons l'exemple des nitrates en rivière. On veut savoir si la quantité de nitrates varie d'une saison à l'autre dans une rivière. Pour cela, on prélève dans une même station le long d'une rivière, à chaque fois aux 10 mêmes points de prélèvement ($n=10$), une certaine quantité d'eau au cours de 3 saisons ($k=3$). Il s'agit dans ce cas d'un schéma de test sur 3 échantillons appariés.

Analyse de variance

Soit *k* variables aléatoires appariées X_i de lois normales $N(\mu_i, \sigma^2)$ (leur variance est donc supposée identique), avec $i = 1,...,k$. Soit tout ensemble de *k* échantillons de n valeurs indépendantes $(n>1)$ $x_{i1},...,x_{in}$ de ces variables X_i, de moyennes intra-échantillon estimées $\overline{x_{i.}} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$, avec $i = 1,...,k$, de moyennes intra-bloc estimées $\overline{x_{.j}} = \frac{1}{k}\sum_{i=1}^{k} x_{ij}$, avec $j = 1,...,n$, d'effectif total $N = nk$ et de moyenne totale estimée $\overline{x_{..}} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij} = \frac{1}{k}\sum_{i=1}^{k}\overline{x_{i.}} = \frac{1}{n}\sum_{j=1}^{n}\overline{x_{.j}}$.

L'équation de l'analyse de variance s'écrit :

$$SCT = SCE + SCB + SCR'$$

où

- $SCT = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \overline{x_{..}}\right)^2$ est la somme des carrés des écarts totaux,
- $SCE = n\sum_{i=1}^{k}\left(\overline{x_{i.}} - \overline{x_{..}}\right)^2$ est la somme des carrés des écarts expliqués par les échantillons,
- $SCB = k\sum_{j=1}^{n_i}\left(\overline{x_{.j}} - \overline{x_{..}}\right)^2$ est la somme des carrés des écarts expliqués par les blocs,
- $SCR' = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x_{i.}} - \overline{x_{.j}} + \overline{x_{..}}\right)^2$ est la somme des carrés des écarts résiduels.

Alors on montre que sous l'hypothèse H_0 la fonction discriminante $D = \dfrac{\dfrac{SCE}{k-1}}{\dfrac{SCR'}{(n-1)(k-1)}}$ suit la loi de Fisher $F\left(k-1,(n-1)(k-1)\right)$.

Test

Lorsque H_1 est vraie, certaines moyennes μ_i ne sont pas égales. On montre alors que *SCE* (la somme des carrés des écarts expliqués par les échantillons) a tendance à prendre de grandes valeurs, quels que soient les signes des inégalités entre moyennes, et donc que la statistique de test *D* ne peut qu'augmenter.

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop grandes, sinon on est plutôt en faveur de l'hypothèse H_1 : les moyennes μ_i ne sont pas toutes égales, i.e. au-moins deux moyennes sont différentes. La région d'acceptation est un intervalle de la forme $\left[0, F_{1-\alpha}\left(k-1,(n-1)(k-1)\right)\right]$.

Comparaisons multiples

Lorsque l'hypothèse nulle d'égalité des moyennes est rejetée, nous savons qu'au-moins deux des moyennes μ_i sont différentes. La question se pose alors légitimement de savoir quelles moyennes sont différentes, quels écarts sont les plus importants, et quelles moyennes s'écartent significativement d'une moyenne de référence.
Si une simple observation des valeurs peut apporter une première réponse, seul un test formel permettra de conclure sans détour.

Dans le cas où tous les n_i ont la même valeur *h* ($n = kh$), le test de Newmann-Keuls donne des précisions supplémentaires sur les contrastes $\mu_i - \mu_j$ et permet de regrouper les modalités du facteur d'influence *A* en groupes homogènes ne présentant pas de différence significative de moyennes (à un seuil de signification donné). Notamment, si l'on conclut à un seul groupe homogène, c'est que toutes les moyennes sont homogènes.
(voir à ce sujet le cours sur les tests paramétriques de Ricco Rakotomalala, Université Lumière Lyon 2)

2.6.2.6 Test de comparaison de *k* variances σ_i^2

Objet

La littérature recense plusieurs tests de comparaison de *k* variances. Ils sont utilisables pour tout $k \geq 2$.

Certains tests usuels cumulent les désavantages mais sont malgré cela très largement répandus, utilisés dans les études et disponibles dans les logiciels statistiques. Il ne faut probablement y voir que des raisons purement historiques. Citons à ce sujet :

- le test de Bartlett : ce test est très sensible à la non-normalité des échantillons, il n'est absolument pas robuste ;
- les tests de Cochran et de Hartley : ces tests ne sont satisfaisant que si les effectifs des échantillons sont proches les uns des autres, voir même parfaitement équilibrés c'est-à-dire $n_1 = ... = n_k$, et si les distributions sous-jacentes des échantillons sont normales, ces tests sont peu robustes par rapport à ces hypothèses.

Soyons clair : l'utilisation de ces tests n'est pas vraiment conseillée…

Le test de Levene offre une alternative crédible au test de Bartlett (et de Fisher). Il est robuste face à la non-normalité des distributions sous-jacentes, mais ne sera réellement performant que si ces distributions sont symétriques, avec une queue de distribution modérée.

Un test ressort enfin parmi les autres : la variante du test de Brown-Forsythe basée sur la médiane conditionnelle (il existe d'autres variantes du test de Brown-Forsythe), qui est une généralisation du test de Levene. Il en précise les conditions de robustesse et le rend performant également lorsque les distributions sous-jacentes sont à queue lourde (loi de Cauchy par exemple) ou asymétriques.
Lorsque nous n'avons pas de connaissances précises sur les distributions, ce test est conseillé. Il réalise le meilleur compromis quelles que soient les distributions sous-jacentes. C'est la procédure à utiliser en priorité pour tester l'homogénéité des variances dans un contexte générique. C'est donc le test qui sera présenté et conseillé.

Hypothèses à tester

H_0 : $\sigma_i^2 = \sigma_j^2$ pour tout $i \in (1,..,k)$ et tout $j \in (1,..,k)$, $i \neq j$

H_1 : les variances σ_i^2 ne sont pas toutes égales, i.e. au-moins deux variances sont différentes.

Remarque :

Parler de version de test (bilatéral, unilatéral à droite ou unilatéral à gauche) n'a plus de sens dans les tests de comparaison de $k > 2$ échantillons (seule une comparaison de deux valeurs peut être envisagée selon ces trois cas).

Robustesse

Comme déjà dit ci-dessus, la variante du test de Brown-Forsythe basée sur la médiane conditionnelle est le meilleur compromis quelles que soient les distributions sous-jacentes. Il est robuste face à la non-normalité de ces distributions.

Test de Brown-Forsythe, variante basée sur la médiane conditionnelle

Soit *k* variables aléatoires indépendantes X_i de lois normales $N(\mu_i, \sigma_i^2)$, avec $i = 1,...,k$. Soit tout ensemble de *k* échantillons de n_i valeurs indépendantes $(n_i > 1)$ $x_{i1},...,x_{in_i}$ de ces variables X_i.

Définition :

La médiane d'un ensemble de valeurs est la valeur centrale de cet ensemble ordonné en ordre croissant. Pour une liste ordonnée de $2n+1$ éléments, la valeur du $n+1^{ème}$ élément est la médiane. Pour une liste ordonnée de $2n$ éléments, toute valeur comprise entre l'élément n et l'élément $n+1$ est une médiane et en pratique on utilisera la moyenne arithmétique de ces deux valeurs centrales.

Excel :

MEDIANE(plage de données)

Soit la transformation de variables $z_{ij} = \left| x_{ij} - \tilde{x}_i \right|$ où $\tilde{x}_i$ désigne la médiane des valeurs de l'échantillon *i*. On définit les moyennes intra-échantillon estimées $\overline{z_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} z_{ij}$, avec $i = 1,...,k$, l'effectif total $n = \sum_{i=1}^{k} n_i$ et la moyenne totale estimée $\overline{z} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} z_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i \overline{z_i}$.

Le test de Brown-Forsythe n'est rien d'autre qu'une analyse de variance sur la variable transformée z_{ij}. Définissant donc :

- $SCE = \sum_{i=1}^{k} n_i \left(\overline{z_i} - \overline{z}\right)^2$, somme des carrés des écarts expliqués,
- $SCR = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(z_{ij} - \overline{z_i}\right)^2$, somme des carrés des écarts résiduels,

on montre alors que sous l'hypothèse H_0 la fonction discriminante $D = \frac{\frac{SCE}{k-1}}{\frac{SCR}{n-k}}$ suit la loi de Fisher $F(k-1, n-k)$.

Test

Lorsque H_1 est vraie, certaines variances σ_i^2 ne sont pas égales. On montre alors que *SCE* (la somme des carrés des écarts expliqués) a tendance à prendre de grandes valeurs, quels que soient les signes des inégalités entre variances, et donc que la statistique de test *D* ne peut qu'augmenter.

Sous l'hypothèse H_0, les valeurs de *D* ne doivent pas être trop grandes, sinon on est plutôt en faveur de l'hypothèse H_1 : les variances σ_i^2 ne sont pas toutes égales, i.e. au-moins deux variances sont différentes. La région d'acceptation est un intervalle de la forme $\left[0, F_{1-\alpha}(k-1, n-k)\right]$.

2.6.2.7 Autres tests paramétriques

On trouvera dans la littérature les principaux autres tests de comparaison :

- comparaison de proportions,
- analyse de variance multivariée (i.e. à plusieurs facteurs contrôlés) appelée « MANOVA »,
- …

2.6.3 Tests de cohérence

En raison de son omniprésence, la loi normale (voir théorème central limite) a naturellement fait l'objet de plus d'attentions que les autres lois. Ainsi dispose-t-on notamment des tests remarquables exposées ci-dessous et relatifs à la loi normale.

ATTENTION ! Les propriétés de ce chapitre ne s'appliquent que pour des échantillons issus de lois normales.

2.6.3.1 Objet

On peut définir les observations incohérentes comme étant des observations qui s'écartent de façon anormale de l'ensemble des autres observations du groupe auquel elles appartiennent, et cela par référence à un modèle théorique donné, par exemple une distribution normale.

Il arrive fréquemment que les données observées contiennent une proportion de valeurs extrêmes résultant d'erreurs de procédure, d'observation ou – parfois – de phénomènes inhabituels. Il est cependant souvent difficile de distinguer les valeurs erronées des valeurs résultant de variations fortuites, qui peuvent elles aussi donner lieu à des valeurs extrêmes occasionnelles. La détection des valeurs aberrantes par des tests de cohérence permet de faire la distinction entre une valeur fortuite issue possiblement de la population normale des données, et les valeurs qui ne peuvent raisonnablement résulter d'une variabilité aléatoire. Ainsi, il sera statistiquement peu probable qu'une valeur aberrante soit présente par hasard.

De nombreuses techniques statistiques sont sensibles à la présence de valeurs aberrantes. Par exemple, le simple calcul de la moyenne et de l'écart type peuvent être biaisés par la présence d'une seule donnée grossièrement imprécise. La détection de valeurs aberrantes devrait être une étape routinière préalable à toute analyse de données (voir à ce sujet le paragraphe « Interprétation des données incohérentes » plus bas).

Les méthodes graphiques (voir le test de Mandel ci-dessous) sont généralement suffisantes lorsque le but principal est d'identifier les données qui nécessitent un examen plus approfondi, ou pour identifier les éventuelles erreurs de procédure ou d'observation. Toutefois, s'il est question de prendre des décisions critiques (y compris la décision de rejeter ces données), ou de garantir la fourniture d'un lot de données utilisable pour tout traitement ultérieur, l'inspection graphique doit toujours être soutenue par des tests numériques (voir tests de Cochran et de Grubbs ci-dessous).

Il est autorisé de répéter l'application des tests de cohérence sur des données. En effet, il n'est pas rare de découvrir qu'une valeur extrême exceptionnelle est accompagnée d'une autre valeur un peu moins extrême. Cela implique simplement de re-tester l'ensemble des données après en avoir extrait la valeur aberrante qui y aurait été découverte.

Ci-dessous :

- $T_\alpha(k)$ ($=-T_{1-\alpha}(k)$ par symétrie de la loi de Student) désigne le quantile d'ordre α de la loi de Student à *k* degrés de liberté.
- $F_\alpha(k,l)$ désigne le quantile d'ordre α de la loi de Fisher avec k et l degrés de liberté.

Excel :

LOI.STUDENT.INVERSE.N(α, k) donne $T_\alpha(k)$

INVERSE.LOI.F.N(α, k, l) donne $F_\alpha(k,l)$

2.6.3.2 Tests de Mandel

Les tests de Mandel consistent en des approches graphiques.

Test *k* de Mandel

Le test *k* de Mandel est un « test de cohérence intra-échantillon » qui indique si la variance d'un échantillon de données est significativement plus grande que les variances d'un groupe d'autres échantillons avec lesquels il est supposé comparable. Il est nécessaire de disposer d'au-moins deux données par échantillon pour calculer leur variance, et donc pour évaluer la statistique *k* de Mandel.

Soit *p* variables aléatoires indépendantes X_i de lois normales $N(\mu_i, \sigma^2)$ (leur variance est donc supposée identique), avec $i = 1,...,p$. Soit tout ensemble de *p* échantillons de n_i valeurs indépendantes $(n_i > 1)$ $x_{i1},...,x_{in_i}$ de ces variables X_i, de moyennes intra-échantillon estimées $\overline{x_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$ et de variance intra-échantillon estimée $s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x_i}\right)^2$, avec $i = 1,...,p$. Le test *k* de Mandel suppose que $n_i = n$ pour $i = 1,...,p$. Soit $s_{max} = \underset{i=1,...,p}{Max}(s_i)$. Alors la statistique *k* de Mandel pour l'échantillon *i* est $k_i = \frac{s_i\sqrt{p}}{\sqrt{\sum_{i=1}^{p} s_i^2}}$, $i = 1,...,p$.

L'observation *i* sera considérée incohérente au niveau de confiance α si $k_i > k_{p,n,1-\alpha}^{crit}$ avec $k_{p,n,1-\alpha}^{crit} = \sqrt{\frac{p}{1 + (p-1)F_\alpha((p-1)(n-1), n-1)}}$.

Remarque :

Le choix de $1-\alpha$ dénote un test unilatéral à droite H_{1+} et se justifie par le fait que les statistiques k_i ne peuvent qu'être positives, et ne peuvent donc s'éloigner de 0 que dans le sens des valeurs positives.

Le test *k* de Mandel ne s'applique strictement que lorsque toutes les variances sont calculées à partir du même nombre $n_i = n$ de valeurs. Sinon voir la généralisation présentée plus bas.

Test *h* de Mandel

Le test *h* de Mandel est un « test de cohérence inter-données » qui indique si la valeur d'une donnée est significativement plus grande ou plus petite que les valeurs d'un groupe d'autres données issues du même échantillon, avec lesquelles elle est supposée comparable.

Soit une variable aléatoire *X* de loi normale $N(\mu, \sigma^2)$. Soit tout échantillon de *p* valeurs indépendantes $(p > 1)$ $x_1,...,x_i,...,x_p$ de cette variable *X*, de moyenne estimée $\overline{x} = \frac{1}{p}\sum_{i=1}^{p} x_i$ et de variance estimée $s_X^2 = \frac{1}{p-1}\sum_{i=1}^{p}\left(x_i - \overline{x}\right)^2$. Alors la statistique *h* de Mandel pour l'observation *i* est $h_i = \frac{x_i - \overline{x}}{s_X}$, $i = 1,...,p$.

L'observation i sera considérée incohérente au niveau de confiance α si $|h_i| > h^{crit}_{p,1-\alpha/2}$ avec $h^{crit}_{p,1-\alpha/2} = \dfrac{(p-1)T_{1-\alpha/2}(p-2)}{\sqrt{p\left(p-2+T^2_{1-\alpha/2}(p-2)\right)}}$.

Remarque :

Le choix de $1-\alpha/2$ dénote un test bilatéral H_{10} et se justifie par le fait que les statistiques h_i peuvent être positives ou négatives, et peuvent donc s'éloigner de 0 dans un sens comme dans l'autre.

Dans le cas (essais interlaboratoires par exemple) où l'étude porte sur p échantillons de n_i valeurs indépendantes $x_{i1},...,x_{in_i}$, avec $i = 1,...,p$, issus de p variables aléatoires, il est possible d'appliquer :

- un test *h* de Mandel aux n_i données x_{ij} d'un échantillon *i* (test de cohérence inter-données), et
- un test *h* de Mandel aux *p* moyennes $\overline{x_i}$ des échantillons (test de cohérence inter-échantillons).

Généralisation des tests de Mandel lorsque les n_i sont différents

Soit p variables aléatoires indépendantes X_i de lois normales $N(\mu_i, \sigma^2)$, avec $i = 1,...,p$. Soit tout ensemble de p échantillons de n_i valeurs indépendantes $(n_i > 1)$ $x_{i1},...,x_{in_i}$ de ces variables X_i, de moyennes intra-échantillon estimées $\overline{x_i} = \dfrac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$ et de variance intra-échantillon estimée $s_i^2 = \dfrac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x_i}\right)^2$, avec $i = 1,...,p$.

On trouve (Nordtest project No. 1483-99) la généralisation suivante lorsque les n_i sont différents :

- la statistique *k* de Mandel pour l'échantillon *i* est $k_i = \dfrac{s_i}{s_r}$, $i = 1,...,p$,
- la statistique *h* de Mandel pour l'observation *i* est $h_i = \dfrac{\overline{x_i} - \overline{\overline{x}}}{\sqrt{\dfrac{s_d^2}{\overline{\overline{n}}}}}$, $i = 1,...,p$,

avec $s_r = \dfrac{\sum_{i=1}^{p}(n_i - 1)s_i^2}{\sum_{i=1}^{p}(n_i - 1)}$, $\overline{\overline{x}} = \dfrac{\sum_{i=1}^{p} n_i \overline{x_i}}{\sum_{i=1}^{p} n_i}$, $s_d^2 = \dfrac{1}{p-1}\sum_{i=1}^{p} n_i\left(\overline{x_i} - \overline{\overline{x}}\right)^2$ et $\overline{\overline{n}} = \dfrac{1}{p-1}\left(\sum_{i=1}^{p} n_i - \dfrac{\sum_{i=1}^{p} n_i^2}{\sum_{i=1}^{p} n_i}\right)$

Application des tests de Mandel

Au premier abord, les tests *k* et *h* de Mandel peuvent paraître fort semblables aux tests de Cochran et de Grubbs (test simple) respectivement (la statistique *C* de Cochran est le maximum des carrés des statistiques *k* de Mandel divisé par *p*, la statistique simple *G* de Grubbs est le maximum des statistiques *h* de Mandel). Il n'en est cependant rien (comme on peut le percevoir dans l'excellent article « Critical values of Mandel's h and k, the Grubbs and the Cochran test statistic » de Peter-T. Wilrich par exemple). C'est pourquoi l'application de ces tests est recommandée en première approche, et non pas pour prendre des décisions critiques. Dans cette logique, bien qu'ils adoptent également des niveaux de confiance critiques à

1% et 5% (comme pour Cochran et Grubbs), les tests de Mandel ne qualifient cependant pas les individus testés de correct, isolés ou aberrant.

L'approche graphique des tests de Mandel consiste à tracer sur un graphique les valeurs de h_i pour chaque donnée avec des lignes correspondant aux indicateurs $\pm h_{\alpha=1\%}^{crit}$ et $\pm h_{\alpha=5\%}^{crit}$, et sur un autre graphique les valeurs de k_i pour chaque donnée avec des lignes correspondant aux indicateurs $k_{\alpha=1\%}^{crit}$ et $k_{\alpha=5\%}^{crit}$. Ces lignes indicatrices servent de guides lors de l'examen des formes des données.

L'examen des graphiques des valeurs *h* et *k* peut montrer que des échantillons spécifiques fournissent des ensembles de valeurs qui sont de façon marquée différents des autres.
Par exemple, si un échantillon apparaît sur le tracé de *k* comme ayant de fortes valeurs, il faut alors en rechercher la raison : cela indique qu'il a une plus faible variabilité. A l'opposé, un échantillon pourrait susciter des valeurs conséquemment petites de *k* en raison d'un arrondi excessif de ses données ou d'une échelle de mesure non adaptée.

Les tracés des valeurs *h* et *k* de Mandel doivent ainsi être examinés pour la cohérence des données. Ces graphiques peuvent indiquer la pertinence des données pour la suite des analyses, la présence de valeurs aberrantes possibles et d'échantillons aberrants possibles. Cependant, aucune décision définitive n'est prise à ce stade, mais est repoussée jusqu'à l'interprétation des tests de Cochran et de Grubbs.

2.6.3.3 Test de Cochran

Objet

Le test *C* de Cochran est un test numérique d'homogénéité des variances.

Le test *C* de Cochran est un « test de cohérence intra-échantillon » qui indique si la variance d'un échantillon de données est significativement plus grande que les variances d'un groupe d'autres échantillons avec lesquels il est supposé comparable.
Le test *C* de Cochran n'évalue que la plus forte valeur d'un ensemble de variances. Il s'agit donc d'un test pour valeur unique. Il est nécessaire de disposer d'au-moins deux données par échantillon pour calculer leur variance, et donc pour évaluer la statistique *C* de Cochran.

Hypothèses à tester

H_0 : toutes les variances sont égales,
H_{1+} : au-moins une variance est significativement plus grande que les autres.

Robustesse

Le test de Cochran est assez robuste par rapport à l'hypothèse de normalité. Il reste applicable tant que les distributions sous-jacentes sont unimodales.

Le test de Cochran ne s'applique strictement que lorsque toutes les variances sont calculées à partir du même nombre $n_i = n$ de valeurs. En pratique, on rencontrera de nombreuses applications où n_i sera légèrement différent de *n* pour certains échantillons. On peut cependant raisonnablement considérer que si de telles variations dans le nombre de valeurs par échantillon sont limitées, elles pourront être ignorées et le test de Cochran s'appliquera en utilisant pour *n* le nombre de valeurs n_i présentent dans la majorité des échantillons.

Test

Soit *p* variables aléatoires indépendantes X_i de lois normales $N(\mu_i, \sigma_i^2)$, avec $i = 1,...,p$. Soit tout ensemble de *p* échantillons de n_i valeurs indépendantes $(n_i > 1)$ $x_{i1},...,x_{in_i}$ de ces variables X_i, de moyennes intra-échantillon estimées $\overline{x_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$ et de variance intra-échantillon estimée $s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x_i}\right)^2$, avec $i = 1,...,p$. Le test de Cochran suppose que $n_i = n$ pour $i = 1,...,p$. Soit $s_{max} = \underset{i=1,...,p}{Max}(s_i)$. Alors la statistique de Cochran est $C = \frac{s_{max}^2}{\sum_{i=1}^{p} s_i^2}$.

La variance la plus élevée s_{max}^2 sera considérée incohérente au niveau de confiance α si $C > C_{p,n,1-\alpha}^{crit}$ avec $C_{p,n,1-\alpha}^{crit} \approx \frac{1}{1 + (p-1)F_{\alpha/p}((p-1)(n-1), n-1)}$.

Remarque :

Le choix de $1-\alpha$ dénote un test unilatéral à droite H_{1+} et se justifie par le fait que la statistique C ne peut qu'être positive, et ne peut donc s'éloigner de 0 que dans le sens des valeurs positives.

L'erreur absolue maximale sur $C_{p,n,1-\alpha}^{crit}$ est 0,008 pour $p = 2,...,120$ et $n = 2,...,145$ par rapport aux données originales tabulées.

Les valeurs testées sont qualifiées comme suit :

- $C \le C_{\alpha=5\%}^{crit}$: si la statistique du test est inférieure ou égale à sa valeur critique à 5 %, l'individu testé est accepté comme correct,
- $C_{\alpha=5\%}^{crit} < C \le C_{\alpha=1\%}^{crit}$: si la statistique du test est supérieure à sa valeur critique à 5 % et inférieure ou égale à sa valeur critique à 1 %, l'individu testé est appelé une valeur isolée,
- $C_{\alpha=1\%}^{crit} < C$: si la statistique du test est supérieure à sa valeur critique à 1 %, l'individu est appelé une valeur aberrante.

2.6.3.4 Test de Grubbs

Objet

Le test *G* de Grubbs est un test numérique d'homogénéité des données.

Le test *G* de Grubbs est un « test de cohérence inter-données » qui indique si la valeur d'une donnée est significativement plus grande ou plus petite que les valeurs d'un groupe d'autres données issues du même échantillon, avec lesquelles elle est supposée comparable.
Il existe un test simple de Grubbs qui permet d'évaluer la plus petite ou la plus grande valeur d'un ensemble de données, et un test double de Grubbs qui permet d'évaluer les deux plus petites ou les deux plus grandes valeurs d'un ensemble de données. Il s'agit donc d'un test pour valeur unique et valeurs doubles. Il est nécessaire de disposer d'au-moins trois données pour réaliser le test simple de Grubbs (comparer seulement deux valeurs ne peut pas permettre d'en suspecter l'une par rapport à l'autre) et d'au-moins quatre données pour réaliser le test double de Grubbs.

Hypothèses à tester

H_0 : tous les x_i ($i = 1,...,p$) sont des réalisations indépendantes d'une même distribution normale $N(\mu,\sigma^2)$,
H_1 : tous les x_i sont des réalisations indépendantes d'une même distribution normale à l'exception des valeurs extrêmes de l'échantillon qui suivent une distribution de moyenne plus grande (ou plus petite).

Robustesse

On ne trouve dans la littérature à peu près aucune indication à ce sujet. On peut simplement lire que le test de Grubbs reste raisonnablement applicable aux distributions « approximativement » normales, ce qui semble devoir être compris comme étant « aux distributions unimodales et symétriques ».

Test

Soit une variable aléatoire *X* de loi normale $N(\mu,\sigma^2)$. Soit tout échantillon de *p* valeurs indépendantes $(p>1)$ triées par ordre croissant $x_1,...,x_i,...,x_p$ ($x_1 \le ... \le x_i \le ... \le x_p$) de cette variable *X*, de moyenne estimée $\bar{x} = \frac{1}{p}\sum_{i=1}^{p} x_i$ et de variance estimée $s_X^2 = \frac{1}{p-1}\sum_{i=1}^{p}\left(x_i - \bar{x}\right)^2$.

Une observation incohérente

La statistique de Grubbs est :

- $G_p = \frac{x_p - \bar{x}}{s_X}$ pour la plus grande valeur,
- $G_1 = \frac{\bar{x} - x_1}{s_X}$ pour la plus petite valeur.

La donnée considérée sera considérée incohérente au niveau de confiance α si $G > G_{p,1-\alpha}^{crit,simple}$ avec $G_{p,1-\alpha}^{crit,simple} \approx \frac{(p-1)T_{1-\alpha/2p}(p-2)}{\sqrt{p\left(p-2+T_{1-\alpha/2p}^2(p-2)\right)}}$.

Remarque :
Le choix de $1-\alpha$ dénote un test unilatéral à droite H_{1+} et se justifie par le fait que la statistique G ne peut qu'être positive, et ne peut donc s'éloigner de 0 que dans le sens des valeurs positives.

L'erreur absolue maximale sur $G_{p,1-\alpha}^{crit,simple}$ est 0,009 pour $p = 4,...,149$ par rapport aux données originales tabulées.

Les valeurs testées sont qualifiées comme suit :

- $G \le G_{\alpha=5\%}^{crit,simple}$: si la statistique du test est inférieure ou égale à sa valeur critique à 5 %, l'individu testé est accepté comme correct,
- $G_{\alpha=5\%}^{crit,simple} < G \le G_{\alpha=1\%}^{crit,simple}$: si la statistique du test est supérieure à sa valeur critique à 5 % et inférieure ou égale à sa valeur critique à 1 %, l'individu testé est appelé une valeur isolée,

- $G_{\alpha=1\%}^{crit,simple} < G$: si la statistique du test est supérieure à sa valeur critique à 1 %, l'individu est appelé une valeur aberrante.

<u>Deux observations incohérentes</u>

La statistique de Grubbs est :

- $G_{p-1,p} = \dfrac{s_{p-1,p}^2}{s_0^2}$ pour les deux plus grandes valeurs,

  avec $s_o^2 = \sum_{i=1}^{p}\left(x_i - \bar{x}\right)^2$, $s_{p-1,p}^2 = \sum_{i=1}^{p-2}\left(x_i - \bar{x}_{p-1,p}\right)^2$ et $\bar{x}_{p-1,p} = \dfrac{1}{p-2}\sum_{i=1}^{p-2} x_i$,

- $G_{1,2} = \dfrac{s_{1,2}^2}{s_0^2}$ pour les deux plus petites valeurs,

  avec $s_o^2 = \sum_{i=1}^{p}\left(x_i - \bar{x}\right)^2$, $s_{1,2}^2 = \sum_{i=3}^{p}\left(x_i - \bar{x}_{1,2}\right)^2$ et $\bar{x}_{1,2} = \dfrac{1}{p-2}\sum_{i=3}^{p} x_i$.

Les deux données considérées seront considérées incohérentes au niveau de confiance α si $G < G_{p,1-\alpha}^{crit,double}$ avec $G_{p,\alpha}^{crit,double} \approx \dfrac{1}{1+\dfrac{2}{p-3}F_{(1-\alpha/2)^{1/f(n)}}(2,p-3)}$, où $f(n) = ap^2 + bp + c$ et

α	a	b	c	*Erreur absolue maximale sur* $G_{crit,double}$
0,2%	0,0443	1,0012	-4,2493	0,0007
1%	0,0388	0,9558	-3,6613	0,0012
2%	0,0362	0,9250	-3,3101	0,0014
5%	0,0322	0,8833	-2,8580	0,0017
10%	0,0289	0,8501	-2,5075	0,0023
20%	0,0251	0,8169	-2,1615	0,0029

La dernière colonne donne l'erreur absolue maximale sur $G_{p,1-\alpha}^{crit,double}$ pour $p = 4,...,149$ par rapport aux données originales tabulées.

Remarque :

Le choix de α dénote un test unilatéral à gauche H_{1-} et se justifie par le fait que les statistiques G ne peuvent qu'être inférieures à 1 (tout en restant positives), et ne peuvent donc s'éloigner de 1 que dans le sens des plus petites valeurs.

Les valeurs testées sont qualifiées comme suit :

- $G_{\alpha=5\%}^{crit,double} \le G$: si la statistique du test est supérieure ou égale à sa valeur critique à 5 %, l'individu testé est accepté comme correct,
- $G_{\alpha=1\%}^{crit,double} \le G < G_{\alpha=5\%}^{crit,double}$: si la statistique du test est inférieure à sa valeur critique à 5 % et supérieure ou égale à sa valeur critique à 1 %, l'individu testé est appelé une valeur isolée,
- $G < G_{\alpha=1\%}^{crit,double}$: si la statistique du test est inférieure à sa valeur critique à 1 %, l'individu est appelé une valeur aberrante.

Cas de *p* échantillons de n_i valeurs

Dans le cas (essais interlaboratoires par exemple) où l'étude porte sur *p* échantillons de n_i valeurs indépendantes $x_{i1},...,x_{in_i}$, avec $i = 1,...,p$, issus de *p* variables aléatoires, il est possible d'appliquer :

- un test *G* de Grubbs aux n_i données x_{ij} d'un échantillon *i* (test de cohérence inter-données), et
- un test *G* de Grubbs aux *p* moyennes $\overline{x_i}$ des échantillons (test de cohérence inter-échantillons).

2.6.3.5 Application des tests de Cochran et de Grubbs

Interprétation des données incohérentes

L'identification des observations incohérentes ne doit en aucun cas se confondre avec leur élimination. D'une manière générale, la décision d'éliminer ou de ne pas éliminer de telles observations ne doit pas découler de l'application aveugle de l'une ou l'autre procédure, aussi élaborée soit-elle, mais doit rester de la compétence du statisticien qui est responsable de la collecte et du traitement des données.

A cet égard, il faut savoir que l'élimination inconsidérée de valeurs extrêmes peut conduire à des erreurs systématiques importantes. Dans toute la mesure du possible, il y a donc lieu de limiter l'élimination aux seuls cas pour lesquels des raisons objectives, autres que les observations elles-mêmes, le justifient. L'ISO 5725-2 suggère de procéder comme décrit ci-dessous.

Lorsque des valeurs incohérentes (isolées ou aberrantes) sont détectées, il convient de rechercher si ces valeurs peuvent être expliquées par des erreurs techniques. Par exemple :

- un dérapage dans l'exécution de la mesure,
- une erreur de calcul,
- une simple erreur d'écriture lors de la transcription d'un résultat, ou
- l'analyse du mauvais échantillon.

Lorsque l'erreur est une erreur de calcul ou d'écriture, il convient de remplacer le résultat suspect par la valeur correcte; lorsque l'erreur résulte de l'analyse d'un mauvais échantillon, il convient de réassocier le résultat à l'échantillon approprié. Après avoir effectué ces corrections, il convient de répéter l'examen des valeurs isolées ou aberrantes. Si l'explication de l'erreur technique est telle que cela devient impossible de remplacer le résultat suspect, il convient de l'écarter en tant que valeur aberrante « authentique » et n'incarnant pas en propre le phénomène représenté par les données.
Lorsqu'il reste des valeurs incohérentes qui n'ont pas pu être expliquées ni rejetées comme résultant d'une erreur identifiée, les valeurs isolées sont conservées comme individus corrects et les valeurs aberrantes sont écartées à moins que le statisticien ait de bonnes raisons pour les conserver.

Ordre d'application des tests

On a donc vu que dans le cas (essais interlaboratoires par exemple) où l'étude porte sur *p* échantillons de n_i valeurs indépendantes $x_{i1},...,x_{in_i}$, avec $i = 1,...,p$, issus de *p* variables aléatoires, il est possible de réaliser un test d'homogénéité des *p* variances s_i^2 des échantillons, un test d'homogénéité des n_i données x_{ij} d'un échantillon *i* et un test d'homogénéité des *p* moyennes $\overline{x_i}$ des échantillons.

Dans ce cas, les tests doivent être appliqués selon la séquence suivante.

1. Appliquer le test de Cochran aux variances intra-échantillon s_i^2.

2. Si le test de Cochran met en évidence une variance intra-échantillon aberrante, déterminer si cela pourrait être attribuable à une seule donnée x_{ij} de cet échantillon. Appliquer pour cela une seule fois le test de Grubbs pour valeur unique aux données x_{ij} de cet échantillon.
3. Si le test de Grubbs mené en 2 met en évidence une donnée aberrante au sein de l'échantillon, exclure cette valeur. Sinon exclure cet échantillon suite à 1.
4. Répéter 1, 2 et 3 aux données restantes jusqu'à ce que plus aucune variance intra-échantillon aberrante ne soit trouvée.
5. Appliquer le test simple de Grubbs aux moyennes intra-échantillon $\overline{x_i}$. Le test portera sur celle des deux valeurs extrêmes ($\overline{x_{min}}$ ou $\overline{x_{max}}$) la plus éloignée de $\overline{\overline{x}}$ en valeur absolue.
6. Si le test simple de Grubbs mené en 5 met en évidence une moyenne aberrante parmi les échantillons, exclure cette valeur. Sinon, passer directement à 9.
7. Appliquer le test simple de Grubbs à l'autre moyenne intra-échantillon extrême (par exemple, si la moyenne aberrante trouvée en 5 était la plus élevée, appliquer à présent le test à la moyenne la moins élevée).
8. Si le test simple de Grubbs mené en 7 met en évidence une moyenne aberrante parmi les échantillons, exclure cette valeur. Dans tous le cas, passer ensuite directement à 11.
9. Appliquer le test double de Grubbs.
10. Si le test double de Grubbs mené en 9 met en évidence deux moyennes aberrantes parmi les échantillons, exclure ces valeurs.
11. Fin

Remarque :

Le test de Cochran peut être répété mais il peut conduire à des rejets excessifs, comme c'est parfois le cas lorsque l'hypothèse sous-jacente de normalité n'est pas suffisamment bien respectée. L'application répétée du test de Cochran est proposée ici uniquement en tant qu'outil d'aide en vue de combler l'absence d'un test statistique permettant de tester plusieurs valeurs aberrantes simultanément. Le test de Cochran n'est pas fait dans ce but, et il est recommandé de prendre de grandes précautions lorsque l'on tire des conclusions. Ainsi, lorsque deux ou trois échantillons donnent des variances élevées, il convient d'examiner soigneusement les conclusions du test de Cochran.

2.6.3.6 Autres tests

Le test de Dixon est très populaire. Il présente cependant de nombreux inconvénients si l'effectif de l'échantillon est petit, ou si l'on cherche à détecter simultanément deux valeurs incohérentes. On lui préfère alors le test de Grubbs.

Le test de Tietjen-Moore est idéal pour détecter des données incohérentes multiples. Il s'agit d'une généralisation du test double de Grubbs au cas des valeurs aberrantes multiples.

Il existe également des tests adaptés à la détection de valeurs aberrantes dans des échantillons multivariés.

2.6.4 Tests de normalité

Objet

De nombreuses applications (estimation des valeurs vraies des paramètres par intervalles de confiance, tests statistiques vus plus haut, …) supposent que les données qui leur sont soumises sont issues de lois normales. En toute rigueur, il est donc indispensable de vérifier cette normalité avant d'utiliser ces applications.

Remarque :

S'assurer au préalable de la compatibilité des distributions avec l'hypothèse de normalité avant d'utiliser les applications concernées devrait être incontournable, surtout pour les petits effectifs. Fort heureusement, dans le cas des tests statistiques, ce n'est pas une contrainte forte en pratique. En effet, ces tests sont souvent suffisamment robustes pour rester applicables même si l'on s'écarte légèrement des conditions de normalité.

Les tests de normalité permettent d'examiner la compatibilité d'un échantillon de données avec la loi normale. On parle également de test d'adéquation à la loi normale.

La littérature se révèle extrêmement prolifique en matière de tests de normalité. On y trouve des tests graphiques et empiriques (histogramme, boîte à moustache, graphe quantile-quantile dit « qq-plot », droite de Henry, coefficients d'asymétrie et d'aplatissement, …), des tests statistiques (Kolmogorov-Smirnov, Lilliefors, Anderson-Darling, Cramer-Von Mises, Jarque Bera, adéquation du khi-deux, Shapiro-Wilk, Shapiro-Francia, D'Agostino, …), des tests de symétrie (Wilcoxon, Van der Waerden, …), des tests bayesiens, …

Devant l'étendue du choix, on ne peut que se montrer perplexe. Certains auteurs ont ainsi mené des comparaisons entre ces tests. La principale conclusion est que la plupart des méthodes proposées ont une efficacité influencée par la taille des échantillons (ils sont peu efficaces sur les petits effectifs), ou se montrent substantiellement sensibles à certains paramètres (parties centrales des distributions, queues des distributions, …).
On peut lire que les tests graphiques et empiriques ne donnent aucun critère objectif, ou ne donnent qu'une information très parcellaire (ils vérifient certaines conditions nécessaires à la normalité, mais aucune de ces conditions n'est suffisante). Les avis sont partagés quant à la puissance du test de Lilliefors (sa performance est réduite lorsque le désaccord porte sur les queues de distribution), au point qu'il soit déconseillé par certains. Le test d'Anderson-Darling doit comporter une correction pour les très petits effectifs (certains logiciels tels que Statistica ne valident l'utilisation du test d'Anderson-Darling que pour $10 \le n \le 40$). La formulation du test de Jarque-Bera est très simple par rapport au test de D'Agostino, le prix en étant une puissance moindre (ce test est toujours moins puissant que le test de D'Agostino, c'est à dire qu'il a une propension plus élevée à conclure à la compatibilité avec la loi normale ; les écarts s'amenuisant à mesure que les effectifs augmentent, il ne devient réellement intéressant que lorsque les effectifs sont élevés). Les tests de symétrie ne sont que très peu restrictifs puisqu'ils ne portent que sur un seul aspect de la forme de la distribution. Quant au test de normalité historiquement le plus populaire, Kolmogorov-Smirnov, il a une puissance si faible que l'unanimité est qu'il ne devrait plus être considéré pour tester la normalité (le logiciel Graphpad cite : « The Kolmogorov-Smirnov test is only a historical curiosity. It should never be used. »).

Deux tests ressortent enfin parmi les autres. Le test *W* de Shapiro-Wilk (Shapiro and Wilk 1965) et le test *K2* de D'Agostino-Pearson (D'Agostino and Pearson 1973) émergent comme étant d'excellents tests. Ils partagent la propriété intéressante d'être des tests dits « omnibus », en cela qu'ils ont de bonnes propriétés de puissance portant sur un large spectre de distributions non-normales. Ils se complètent par ailleurs parfaitement quant aux étendues d'effectifs sur lesquelles ils sont efficaces. Ce sont donc les deux tests qui seront présentés et conseillés.

Hypothèses à tester

$H_0 : F(x) = F_N(x),$
$H_1 : F(x) \neq F_N(x),$

où $F(x)$ désigne la fonction de répartition de l'échantillon testé et $F_N(x)$ la fonction de répartition de la loi normale.

2.6.4.1 Test *W* de Shapiro-Wilk

Robustesse et puissance

Ce test est particulièrement puissant pour les petits effectifs. Il est recommandé comme le meilleur choix lorsque $n \leq 50$ et à condition que l'échantillon de données testé ne comporte que des valeurs uniques. Le test *W* ne donne en effet pas de bons résultats pour $n > 50$ ou lorsque l'échantillon comporte des doublons.

Test

Soit tout échantillon de *n* valeurs triées par ordre croissant $x_1,...,x_i,...,x_n$ ($x_1 \leq ... \leq x_i \leq ... \leq x_n$), de moyenne estimée $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. La statistique du test *W* est $W = \frac{\left(\sum_{i=1}^{n/2} a_i \left(x_{n-i+1} - x_i\right)\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$ où $n/2$ désigne la partie entière de $n/2$ et les a_i sont des constantes générées à partir de la moyenne et de la matrice de variance covariance des quantiles d'un échantillon de taille *n* suivant la loi normale. Ces constantes sont tabulées en annexe.

Plus *W* est élevé, plus la compatibilité avec la loi normale est crédible. L'échantillon de données sera considéré non-normal au niveau de confiance α si $W < W_{crit}$. Les valeurs seuils W_{crit} sont tabulées en annexe.

2.6.4.2 Test *K2* de D'Agostino-Pearson

Robustesse et puissance

Sa puissance est considérée comme très bonne, et il est apprécié pour l'information qu'il fournit sur la nature de la non-normalité. Il est utilisable dès $n \geq 9$ et présente une puissance similaire à celle de Shapiro-Wilk à mesure que les effectifs augmentent. Il devient particulièrement efficace à partir de $n \geq 20$. Par rapport au test de Shapiro-Wilk, il est de surcroît peu sensible à l'existence des ex-aequo dans l'échantillon. Il est recommandé comme choix pour $n > 50$, où le test de Shapiro-Wilk n'est plus utilisable.

Test

Soit tout échantillon de *n* valeurs $x_1,...,x_i,...,x_n$, de moyenne estimée $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et de variance estimée $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$.

Si l'idée du test de D'Agostino est simple, les formules sont relativement complexes. Le principe est de centrer et réduire les coefficients d'asymétrie et d'aplatissement (moments d'ordres 3 et 4) de manière à générer des variables z_1 et z_2 distribuées asymptotiquement selon une loi normale centrée réduite $N(0,1)$. Cette transformation intègre des corrections supplémentaires de manière à rendre l'approximation normale plus efficace.

Transformation du coefficient d'asymétrie

On calcule successivement :

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^{3/2}}$$

$$A = g_1\sqrt{\frac{(n+1)(n+3)}{6(n-2)}}$$

$$B = \frac{3\left(n^2+27n-70\right)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$

$$C = \sqrt{2(B-1)} - 1$$

$$D = \sqrt{C}$$

$$E = \frac{1}{\sqrt{ln(D)}}$$

$$F = \frac{A}{\sqrt{\frac{2}{C-1}}}$$

$$z_1 = E\,ln\left(F + \sqrt{F^2+1}\right)$$

Transformation du coefficient d'aplatissement

On calcule successivement :

$$g_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_X}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

$$G = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

$$H = \frac{(n-2)(n-3)g_2}{(n+1)(n-1)\sqrt{G}}$$

$$J = \frac{6\left(n^2-5n+2\right)}{(n+7)(n+9)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}}$$

$$K = 6 + \frac{8}{J}\left(\frac{2}{J} + \sqrt{1+\frac{4}{J^2}}\right)$$

$$L = \frac{1-\frac{2}{K}}{1+H\sqrt{\frac{2}{K-4}}}$$

$$z_2 = \frac{\left(1 - \frac{2}{9K}\right) - L^{1/3}}{\sqrt{\frac{2}{9K}}}$$

La statistique du test *K2* est $K2 = z_1^2 + z_2^2$. Elle suit asymptotiquement une loi du khi-deux à deux degrés de liberté.

Plus *K2* est faible, plus la compatibilité avec la loi normale est crédible. L'échantillon de données sera considéré non-normal au niveau de confiance α si $K2 > \chi^2_{1-\alpha}(2)$.

REFERENCES

[] ISO 3534-1:2006 Statistique – Vocabulaire et symboles – Partie 1 : Termes statistiques généraux et termes utilisés en calcul des probabilités

[] Claude Bélisle, Université Laval Quelques rappels sur la loi normale

[] Claude Bélisle, Université Laval La loi du khi-deux, la loi de Student, la loi de Fisher

[] Wikipedia

[] EA 4/02 M rev 01:2013 Evaluation of the Uncertainty of Measurement In Calibration

[] Jean-Michel JOLION, mai 2006, INSA Lyon Probabilités et Statistique

[] Henry Immediato, février 2010, Université Claude-Bernard Lyon 1 Probabilités

[] JSGM 100:2008 (GUM) Évaluation des données de mesure – Guide pour l'expression de l'incertitude de mesure

[] James N Miller, Jane C Miller Statistics and Chemometrics for Analytical Chemistry, Sixth edition, 2010

[] D. Mouchiroud, février 2003, Université Claude-Bernard Lyon 1 Mathématiques : Outils pour la Biologie – Deug SV

[] JP Lenoir, Université Paris Sud Les tests d'hypothèse

[] Henry Immediato, Université Claude-Bernard Lyon 1 Cours de statistiques (2ème partie)

[] Matthieu Kowalski, 2008, Aix-Marseille Université Tests de Student-Fisher – Compléments

[] Ricco Rakotomalala, 2013, Université Lumière Lyon 2 Comparaison de populations - Tests paramétriques (Version 1.2)

[] Ramousse R., Le Berre M. & Le Guelte L., 1996 Introduction aux statistiques

[] Henry Immediato, février 2010, Université Claude-Bernard Lyon 1 Statistiques

[] Pierre Dagnelie, 2006, De Boeck Statistique théorique et appliquée – 2. Inférence statistique à une et à deux dimensions (2ème édition)

[] ISO 5725-1 et -2 :1996 Exactitude (justesse et fidélité) des résultats et méthodes de mesure

[] Tang Luping and Björn Schouenborg, SP REPORT 2000:35 Methodology of Inter-comparison Tests and Statistical Analysis of Test Results - Nordtest project No. 1483-99

[] Académie Nancy-Metz, Génie mécanique-productique Test de Cochran

[] Pierre Jost, Université de Strasbourg — Statistiques à l'usage des ingénieurs et des techniciens

[] Peter-T. Wilrich, 2011, Springer-Verlag — Critical values of Mandel's h and k, the Grubbs and the Cochran test statistic

[] Ricco Rakotomalala, 2011, Université Lumière Lyon 2 — Tests de normalité - Techniques empiriques et tests statistiques (Version 2.0)

[] Ralph B. D' Agostino; Albert Belanger; Ralph B. D' Agostino, Jr., The American Statistician, Vol. 44, No.4. (Nov., 1990), pp. 316-321 — A Suggestion for Using Powerful and Informative Tests of Normality

3 ANNEXE – COEFFICIENTS a_i DU TEST DE SHAPIRO-WILK

$a_i \setminus n$	2	3	4	5	6	7	8	9	10
a1	0,7071	0,7071	0,6872	0,6646	0,6431	0,6233	0,6052	0,5888	0,5739
a2			0,1677	0,2413	0,2806	0,3031	0,3164	0,3244	0,3291
a3					0,0875	0,1401	0,1743	0,1976	0,2141
a4							0,0561	0,0947	0,1224
a5									0,0399

$a_i \setminus n$	11	12	13	14	15	16	17	18	19	20
a1	0,5601	0,5475	0,5359	0,5251	0,5150	0,5056	0,4968	0,4886	0,4808	0,4734
a2	0,3315	0,3325	0,3325	0,3318	0,3306	0,3290	0,3273	0,3253	0,3232	0,3211
a3	0,2260	0,2347	0,2412	0,2460	0,2495	0,2521	0,2540	0,2553	0,2561	0,2565
a4	0,1429	0,1586	0,1707	0,1802	0,1878	0,1939	0,1988	0,2027	0,2059	0,2085
a5	0,0695	0,0922	0,1099	0,1240	0,1353	0,1447	0,1524	0,1587	0,1641	0,1686
a6		0,0303	0,0539	0,0727	0,0880	0,1005	0,1109	0,1197	0,1271	0,1334
a7				0,0240	0,0433	0,0593	0,0725	0,0837	0,0932	0,1013
a8						0,0196	0,0359	0,0496	0,0612	0,0711
a9								0,0163	0,0303	0,0422
a10										0,0140

$a_i \setminus n$	21	22	23	24	25	26	27	28	29	30
a1	0,4643	0,4590	0,4542	0,4493	0,4450	0,4407	0,4366	0,4328	0,4291	0,4254
a2	0,3185	0,3156	0,3126	0,3098	0,3069	0,3043	0,3018	0,2992	0,2968	0,2944
a3	0,2578	0,2571	0,2563	0,2554	0,2543	0,2533	0,2522	0,2510	0,2499	0,2487
a4	0,2119	0,2131	0,2139	0,2145	0,2148	0,2151	0,2152	0,2151	0,2150	0,2148
a5	0,1736	0,1764	0,1787	0,1807	0,1822	0,1836	0,1848	0,1857	0,1864	0,1870
a6	0,1399	0,1443	0,1480	0,1512	0,1539	0,1563	0,1584	0,1601	0,1616	0,1630
a7	0,1092	0,1150	0,1201	0,1245	0,1283	0,1316	0,1346	0,1372	0,1395	0,1415
a8	0,0804	0,0878	0,0941	0,0997	0,1046	0,1089	0,1128	0,1162	0,1192	0,1219
a9	0,0530	0,0618	0,0696	0,0764	0,0823	0,0876	0,0923	0,0965	0,1002	0,1036
a10	0,0263	0,0368	0,0459	0,0539	0,0610	0,0672	0,0728	0,0778	0,0822	0,0862
a11		0,0122	0,0228	0,0321	0,0403	0,0476	0,0540	0,0598	0,0650	0,0697
a12			0,0000	0,0107	0,0200	0,0284	0,0358	0,0424	0,0483	0,0537
a13					0,0000	0,0094	0,0178	0,0253	0,0320	0,0381
a14							0,0000	0,0084	0,0159	0,0227
a15									0,0000	0,0076

$a_i \setminus n$	31	32	33	34	35	36	37	38	39	40
a1	0,4220	0,4188	0,4156	0,4127	0,4096	0,4068	0,4040	0,4015	0,3989	0,3964
a2	0,2921	0,2898	0,2876	0,2854	0,2834	0,2813	0,2794	0,2774	0,2755	0,2737
a3	0,2475	0,2463	0,2451	0,2439	0,2427	0,2415	0,2403	0,2391	0,2380	0,2368
a4	0,2145	0,2141	0,2137	0,2132	0,2127	0,2121	0,2116	0,2110	0,2104	0,2098
a5	0,1874	0,1878	0,1880	0,1882	0,1883	0,1883	0,1883	0,1881	0,1880	0,1878
a6	0,1641	0,1651	0,1660	0,1667	0,1673	0,1678	0,1683	0,1686	0,1689	0,1691
a7	0,1433	0,1449	0,1463	0,1475	0,1487	0,1496	0,1505	0,1513	0,1520	0,1526
a8	0,1243	0,1265	0,1284	0,1301	0,1317	0,1331	0,1344	0,1356	0,1366	0,1376
a9	0,1066	0,1093	0,1118	0,1140	0,1160	0,1179	0,1196	0,1211	0,1225	0,1237
a10	0,0899	0,0931	0,0961	0,0988	0,1013	0,1036	0,1056	0,1075	0,1092	0,1108
a11	0,0739	0,0777	0,0812	0,0844	0,0873	0,0900	0,0924	0,0947	0,0967	0,0986
a12	0,0585	0,0629	0,0669	0,0706	0,0739	0,0770	0,0798	0,0824	0,0848	0,0870
a13	0,0435	0,0485	0,0530	0,0572	0,0610	0,0645	0,0677	0,0706	0,0733	0,0759
a14	0,0289	0,0344	0,0395	0,0441	0,0484	0,0523	0,0559	0,0592	0,0622	0,0651
a15	0,0144	0,0206	0,0262	0,0314	0,0361	0,0404	0,0444	0,0481	0,0515	0,0546
a16	0,0000	0,0068	0,0131	0,0187	0,0239	0,0287	0,0331	0,0372	0,0409	0,0444
a17			0,0000	0,0062	0,0119	0,0172	0,0220	0,0264	0,0305	0,0343
a18					0,0000	0,0057	0,0110	0,0158	0,0203	0,0244
a19							0,0000	0,0053	0,0101	0,0146
a20									0,0000	0,0049

$a_i \setminus n$	41	42	43	44	45	46	47	48	49	50

a1	0,3940	0,3917	0,3894	0,3872	0,3850	0,3830	0,3808	0,3789	0,3770	0,3751
a2	0,2719	0,2701	0,2684	0,2667	0,2651	0,2635	0,2620	0,2604	0,2589	0,2574
a3	0,2357	0,2345	0,2334	0,2323	0,2313	0,2302	0,2291	0,2281	0,2271	0,2260
a4	0,2091	0,2085	0,2078	0,2072	0,2065	0,2058	0,2052	0,2045	0,2038	0,2032
a5	0,1876	0,1874	0,1871	0,1868	0,1865	0,1862	0,1859	0,1855	0,1851	0,1847
a6	0,1693	0,1694	0,1695	0,1695	0,1695	0,1695	0,1695	0,1693	0,1692	0,1691
a7	0,1531	0,1535	0,1539	0,1542	0,1545	0,1548	0,1550	0,1551	0,1553	0,1554
a8	0,1384	0,1392	0,1398	0,1405	0,1410	0,1415	0,1420	0,1423	0,1427	0,1430
a9	0,1249	0,1259	0,1269	0,1278	0,1286	0,1293	0,1300	0,1306	0,1312	0,1317
a10	0,1123	0,1136	0,1149	0,1160	0,1170	0,1180	0,1189	0,1197	0,1205	0,1212
a11	0,1004	0,1020	0,1035	0,1049	0,1062	0,1073	0,1085	0,1095	0,1105	0,1113
a12	0,0891	0,0909	0,0927	0,0943	0,0959	0,0972	0,0986	0,0998	0,1010	0,1020
a13	0,0782	0,0804	0,0824	0,0842	0,0860	0,0876	0,0892	0,0906	0,9190	0,0932
a14	0,0677	0,0701	0,0724	0,0745	0,0765	0,0783	0,0801	0,0817	0,0832	0,0846
a15	0,0575	0,0602	0,0628	0,0651	0,0673	0,0694	0,0713	0,0731	0,0748	0,0764
a16	0,0476	0,0506	0,0534	0,0560	0,0584	0,0607	0,0628	0,0648	0,0667	0,0685
a17	0,0379	0,0411	0,0442	0,0471	0,0497	0,0522	0,0546	0,0568	0,0588	0,0608
a18	0,0283	0,0318	0,0352	0,0383	0,0412	0,0439	0,0465	0,0489	0,0511	0,0532
a19	0,0188	0,0227	0,0263	0,0296	0,0328	0,0357	0,0385	0,0411	0,0436	0,0459
a20	0,0094	0,0136	0,0175	0,0211	0,0245	0,0277	0,0307	0,0335	0,0361	0,0386
a21	0,0000	0,0045	0,0087	0,0126	0,0163	0,0197	0,0229	0,0259	0,0288	0,0314
a22			0,0000	0,0042	0,0081	0,0118	0,0153	0,0185	0,0215	0,0244
a23					0,0000	0,0039	0,0076	0,0111	0,0143	0,0174
a24							0,0000	0,0037	0,0071	0,0104
a25									0,0000	0,0035

4 ANNEXE – VALEURS CRITIQUES W_{crit} DU TEST DE SHAPIRO-WILK

n \ α	0,01	0,02	0,05	0,1	0,5	0,9	0,95	0,98	0,99
3	0,753	0,756	0,767	0,789	0,959	0,998	0,999	1,000	1,000
4	0,687	0,707	0,748	0,792	0,935	0,987	0,992	0,996	0,997
5	0,686	0,715	0,762	0,806	0,927	0,979	0,986	0,991	0,993
6	0,713	0,743	0,788	0,826	0,927	0,974	0,981	0,986	0,989
7	0,730	0,760	0,803	0,838	0,928	0,972	0,979	0,985	0,988
8	0,749	0,778	0,818	0,851	0,932	0,972	0,978	0,984	0,987
9	0,764	0,791	0,829	0,859	0,935	0,972	0,978	0,984	0,986
10	0,781	0,806	0,842	0,869	0,938	0,972	0,978	0,983	0,986
11	0,792	0,817	0,850	0,876	0,940	0,973	0,979	0,984	0,986
12	0,805	0,828	0,859	0,883	0,943	0,973	0,979	0,984	0,986
13	0,814	0,837	0,866	0,889	0,945	0,974	0,979	0,984	0,986
14	0,825	0,846	0,874	0,895	0,947	0,975	0,980	0,984	0,986
15	0,835	0,855	0,881	0,901	0,950	0,975	0,980	0,984	0,987
16	0,844	0,863	0,887	0,906	0,952	0,976	0,981	0,985	0,987
17	0,851	0,869	0,892	0,910	0,954	0,977	0,981	0,985	0,987
18	0,858	0,874	0,897	0,914	0,956	0,978	0,982	0,986	0,988
19	0,863	0,879	0,901	0,917	0,957	0,978	0,982	0,986	0,988
20	0,868	0,884	0,905	0,920	0,959	0,979	0,983	0,986	0,988
21	0,873	0,888	0,908	0,923	0,960	0,980	0,983	0,987	0,989
22	0,878	0,892	0,911	0,926	0,961	0,980	0,984	0,987	0,989
23	0,881	0,895	0,914	0,928	0,962	0,981	0,984	0,987	0,989
24	0,884	0,898	0,916	0,930	0,963	0,981	0,984	0,987	0,989
25	0,888	0,901	0,918	0,931	0,964	0,981	0,985	0,988	0,989
26	0,891	0,904	0,920	0,933	0,965	0,982	0,985	0,988	0,989
27	0,894	0,906	0,923	0,935	0,965	0,982	0,985	0,988	0,990
28	0,896	0,908	0,924	0,936	0,966	0,982	0,985	0,988	0,990
29	0,898	0,910	0,926	0,937	0,966	0,982	0,985	0,988	0,990
30	0,900	0,912	0,927	0,939	0,967	0,983	0,985	0,988	0,990
31	0,902	0,914	0,929	0,940	0,967	0,983	0,986	0,988	0,990
32	0,904	0,915	0,930	0,941	0,968	0,983	0,986	0,988	0,990
33	0,906	0,917	0,931	0,942	0,968	0,983	0,986	0,989	0,990
34	0,908	0,919	0,933	0,943	0,969	0,983	0,986	0,989	0,990
35	0,910	0,920	0,934	0,944	0,969	0,984	0,986	0,989	0,990
36	0,912	0,922	0,935	0,945	0,970	0,984	0,986	0,989	0,990
37	0,914	0,924	0,936	0,946	0,970	0,984	0,987	0,989	0,990
38	0,916	0,925	0,938	0,947	0,971	0,984	0,987	0,989	0,990
39	0,917	0,927	0,939	0,948	0,971	0,984	0,987	0,989	0,991
40	0,919	0,928	0,940	0,949	0,972	0,985	0,987	0,989	0,991
41	0,920	0,929	0,941	0,950	0,972	0,985	0,987	0,989	0,991
42	0,922	0,930	0,942	0,951	0,972	0,985	0,987	0,989	0,991
43	0,923	0,932	0,943	0,951	0,973	0,985	0,987	0,990	0,991
44	0,924	0,933	0,944	0,952	0,973	0,985	0,987	0,990	0,991
45	0,926	0,934	0,945	0,953	0,973	0,985	0,988	0,990	0,991
46	0,927	0,935	0,945	0,953	0,974	0,985	0,988	0,990	0,991
47	0,928	0,936	0,946	0,954	0,974	0,985	0,988	0,990	0,991
48	0,929	0,937	0,947	0,954	0,974	0,985	0,988	0,990	0,991
49	0,929	0,939	0,947	0,955	0,974	0,985	0,988	0,990	0,991
50	0,930	0,938	0,947	0,955	0,974	0,985	0,988	0,990	0,991