

Mesurer l'effet d'un traitement

Les méta-analyses en sciences de l'éducation


Nathalie ROQUES

Avril 2021

SOMMAIRE

Résumé.....	5
Pour commencer	7
Premier niveau : les études par comparaison de groupes.....	7
Deuxième niveau : les méta-analyses	10
Analyse statistique des données	11
Première partie : les données et le hasard.....	13
Chapitre 1. Description des données	15
Chapitre 2. Variables aléatoires	23
Chapitre 3. La loi normale	29
Chapitre 4. Distribution normale de caractéristiques naturelles.....	37
Deuxième partie : les échantillons	43
Chapitre 5. Échantillonnage et intervalle de fluctuation	45
Chapitre 6. Estimation des paramètres d'une population	51
Chapitre 7. Différence entre deux moyennes statistiquement significative.....	61
Chapitre 8. Taille d'effet du traitement	76
Troisième partie : les études	91
Chapitre 10. Les méta-analyses.....	93
Chapitre 11. Méta-analyse et enseignement des mathématiques	110
Chapitre 12. Interprétation des tailles d'effet.....	117
Pour conclure	123
Références	125
Liste des figures, des tableaux et des exemples	128

Résumé

Les analyses statistiques utilisées pour décrire les résultats des études quantitatives dans le domaine des Sciences de l'éducation sont expliquées. La première partie de ce livre est consacrée à la description d'une série de données (on parle de **statistiques descriptives**), et s'attache particulièrement à analyser les distributions dites *normales*. Les notions de probabilités mises en œuvre par la suite y seront également présentées. La seconde partie traite ensuite de la comparaison entre deux échantillons extraits d'une population quand l'un d'eux a été soumis à un traitement qui intéresse le chercheur (on parle de **statistiques inférentielles**). Dans la troisième partie, les règles qui encadrent la réalisation des méta-analyses et plus particulièrement l'analyse quantitative de leurs données sont détaillées. Des annexes comprenant notamment des formulaires et la liste des fonctions à utiliser avec le tableur Excel (le pictogramme  dans le texte principal signale la présence en annexe 9 d'une information pertinente). Des exemples, dont les données sont soit réelles soit fictives sont proposés tout au long du livre et les fichiers Excel correspondants peuvent être téléchargés sur le site mathadoc.fr

Pour commencer

De nombreuses discussions ont comme sujet l'acquisition par les élèves de connaissances et de compétences dans plusieurs domaines, comme la compréhension de la langue, les mathématiques ou les sciences. Des débats parfois animés agitent la communauté des spécialistes de l'enseignement en France, débats qui sont parfois relayés par les médias auprès d'un grand public friand de ces questions d'actualité. Mais ces discussions laissent systématiquement de côté tout un pan de la recherche scientifique et se privent, se faisant, d'un élément de réflexion et de construction du savoir. Ce pan de la recherche s'appuie sur des études par comparaison de groupes et des méta-analyses qui toutes se fondent sur un corpus mathématique austère, et il est probable que cet aspect technique aride soit au moins en partie à l'origine de cette mise à l'écart. C'est pour remédier à cette situation que ce livre a été écrit.

Les recherches dont il sera question ici fondent leurs conclusions sur la mise en œuvre d'études quantitatives dont le principe de base est de mesurer les influences d'une méthode d'enseignement ou d'une pratique pédagogique sur ce que les élèves apprennent. Deux niveaux de recherche doivent être distingués. Le premier niveau concerne les études qui comparent deux groupes d'élèves, un groupe dans lequel un traitement a été mis en place, comme une méthode pédagogique particulière, et un groupe qui n'est pas soumis à ce traitement, on parle alors de groupe contrôle : ce sont les **études par comparaison de groupes**. Quand plusieurs études indépendantes de ce type ont été menées sur une même question de recherche (on parlera alors d'études primaires), on passe alors au deuxième niveau de recherche qui a comme objectif de synthétiser leurs résultats : le chercheur réalise ce qu'on appelle une **méta-analyse** (on parle d'études secondaires). Ces deux niveaux de recherche ont leurs propres règles et leurs propres spécialistes, mais partagent tout de même un même paradigme scientifique et utilisent les mêmes concepts théoriques basés sur l'analyse statistique de données.

Je m'appuierai souvent dans ce livre sur les travaux menés par le What Works Clearinghouse^a (WWC), initiative lancée en 2002 par l'Institute of Education Sciences (IES) du département de l'Éducation aux États-Unis. Vous trouverez une présentation des procédures mises en œuvre par le WWC dans mon livre *Comment enseigner les maths ? La réponse du What Works Clearinghouse*^b. Enfin les enquêtes PISA mises en œuvre par l'OCDE seront également évoquées, car elles s'imposent comme un élément incontournable dans toutes les discussions portant sur l'enseignement. On devra garder à l'esprit qu'elles n'ont pas comme objectif d'évaluer des méthodes d'enseignement mais de mesurer des corrélations entre certains facteurs et les compétences des élèves.

Premier niveau : les études par comparaison de groupes

Les études quantitatives qui analysent l'effet d'une méthode d'enseignement sur les apprentissages des élèves se basent sur les trois principes importants suivants.

^a <https://ies.ed.gov/ncee/wwc/>

^b www.mathadoc.fr

Le premier principe : mener une expérience scientifique

L'expérience scientifique est un modèle suivi par les chercheurs pour mettre à l'épreuve une théorie. Ces derniers cherchent alors à contrôler au maximum tous les éléments de l'expérience pour en évaluer ou supprimer les influences (ou tout au moins en tenir compte).

Dans le cas qui nous intéresse, l'expérience est menée sur des sujets (des élèves) et consiste à appliquer un traitement (une méthode d'enseignement) puis à récolter des informations (les résultats des tests) une fois le traitement^a terminé. De nombreuses questions peuvent se poser déjà sur

- le choix des sujets : les caractéristiques des élèves sélectionnés doivent être contrôlées (par exemple leur niveau socioéconomique doit être pris en compte).
- l'application du traitement : la qualité de la mise en œuvre de la méthode d'enseignement doit être vérifiée et la méthode d'enseignement utilisé par l'enseignement du groupe contrôle doit être détaillée.
- les informations permettant d'évaluer l'efficacité du traitement : les tests doivent être choisis en fonction des apprentissages ciblés.
- le traitement de ces informations : l'analyse statistique des données doit être rigoureusement menée.

Ces aspects importants mériteraient d'être étudiés et discutés ; mon propos ici ne concernera que le dernier point, c'est-à-dire l'analyse statistique des données obtenues après le traitement et la passation d'un test.

Le deuxième principe : étendre les conclusions à une population entière

Les études et leurs résultats concernent des échantillons prélevés dans une population qui reste la cible des éventuelles préconisations qui seront prononcées. Les échantillons qui regroupent les sujets de l'expérience doivent donc être représentatifs de cette population, et c'est pour cette raison que les élèves sont choisis (dans la mesure du possible) de façon aléatoire dans la population. Quand on élargi les observations et les conclusions faites sur un échantillon à toute une population, on dit alors que l'on infère, ou que l'on procède par inférence.

Travailler sur des échantillons permet de limiter les coûts et d'augmenter la faisabilité d'une expérience (il est parfois impossible de mener des expériences sur l'ensemble de la population) mais offre également l'avantage de permettre la comparaison entre les effets d'un traitement et les effets de l'absence de ce traitement, ce qui nous amène au troisième principe.

Le troisième principe : comparer deux échantillons

Au moins deux échantillons d'élèves sont étudiés dans ces études quantitatives : un groupe traitement qui reçoit le traitement c'est-à-dire la méthode d'enseignement étudiée d'une part, et un groupe contrôle^b qui ne reçoit pas cette méthode d'enseignement (mais une autre qualifiée de « standard ») d'autre part. Pour que les groupes soient équivalents, et que le traitement (ou son absence) soit la seule caractéristique permettant de les différencier, ils sont constitués

^a On utilise également le terme « intervention ».

^b « Échantillon » d'une part et « groupe traitement » (plus simple à utiliser que groupe de traitement) et « groupe contrôle » (pour groupe de contrôle) d'autre part sont employés tout au long de ce texte comme des synonymes.

d'élèves tirés au sort. Une fois l'expérience terminée, le niveau des élèves de chacun des deux groupes est mesuré par la passation d'un test dont les résultats (que l'on nommera **scores** dans toute la suite de ce livre) constituent les données brutes de l'expérience. C'est ensuite la différence entre les mesures récoltées dans chacun des deux groupes qui permet de conclure. On mène alors une analyse statistique de ces données, et c'est de cette analyse dont il sera plus particulièrement question dans la deuxième partie de ce livre.

Quand les premiers et troisièmes principes sont suivis, on dit que l'étude adopte un design^a expérimental. Vous rencontrerez probablement l'acronyme RCT dans les publications anglo-saxonnes (*randomised controled trial*) que l'on traduit par **essai contrôlé randomisé** (ECR) et qui a été schématisé ci-dessous (figure 1). Ce type d'études expérimentales a été plus particulièrement mis au point dans le domaine médical depuis une trentaine d'années et constituent la matière première des méta-analyses déjà évoquées ci-dessus. Les sciences sociales (et les sciences de l'éducation en font partie) doivent souvent adapter leur design à leur objet d'étude : ainsi, il n'est pas toujours possible de distribuer de façon aléatoire les sujets dans les groupes traitement et contrôle. Dans ce cas, on dit que les études suivent un design **quasi-expérimental**. Les essais contrôlés randomisés restent cependant la référence absolue des études qui procèdent par comparaison de groupes et seules les études qui adoptent ce design sont reconnues par le WWC comme conformes à leurs normes sans réserve.

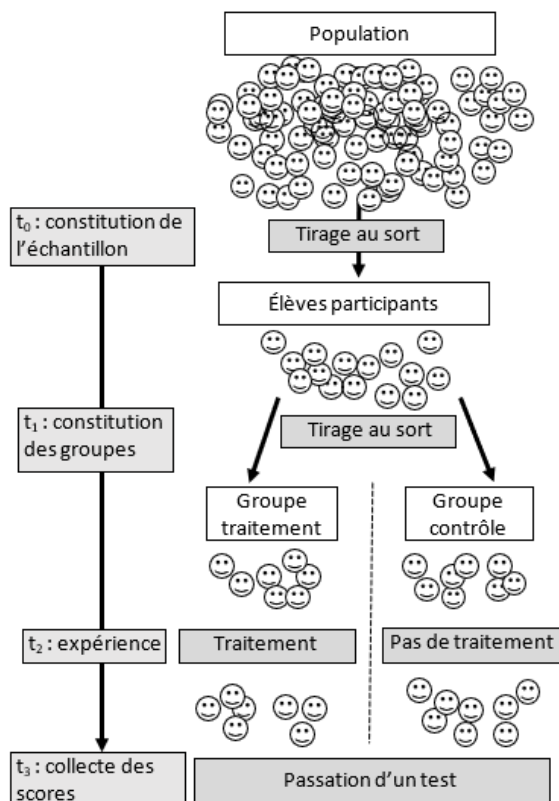


Figure 1 : essai contrôlé randomisé

^a Le mot *design* qui peut être traduit par méthode en français, est également employé dans notre langue.

Le second principe quant à lui n'est que très rarement respecté. C'est le cas par exemple des études PISA. Ces enquêtes transversales (qui ne sont donc pas des études scientifiques telles que présentées ci-dessus) basent leurs conclusions sur des analyse d'échantillons prélevés aléatoirement dans les pays participants et nous auront l'occasion d'évoquer leurs résultats plus loin.

Deuxième niveau : les méta-analyses

Quand plusieurs études par comparaison de groupes ont été menées sur un même traitement, le plus souvent dans des conditions différentes (les élèves sont de niveau ou de nationalités différents par exemple) se pose alors la question des conclusions qui peuvent être tirées de leurs résultats souvent dissemblables et parfois même contradictoires. C'est l'objectif que se fixe le chercheur qui va mener alors une méta-analyse, étude rétrospective^a combinant les résultats de plusieurs études en mettant en œuvre des procédures statistiques^b. Ici aussi un cadre méthodologique est posé que l'on peut broser dans ses grandes lignes en nous appuyant sur trois étapes temporelles majeures. Durant la première, les études traitant d'un sujet commun sont identifiées et sélectionnées en utilisant des critères clairement énoncés ; ce processus transparent se doit d'être systématique et exhaustif, et les motifs d'exclusion explicités. Dans un second temps, les résultats des études sont agrégés et synthétisés ; là aussi, comme pour les études primaires décrites auparavant, c'est l'analyse statistique des données (qui cette fois sont les résultats publiés par les études primaires) qui donne naissance à une poignée d'indicateurs dont la description sera l'objet de la dernière partie de ce livre. La troisième étape de ce processus concerne la publication des résultats qui seront présentés de façon à être compris d'un large public, bien au-delà de la sphère des spécialistes en Sciences de l'éducation (figure 2) ; c'est un point majeur qui distingue les méta-analyses des analyses primaires dont elles sont le fruit, l'audience de ces dernières ne dépassant pas le cercle intime des experts en la matière. Pour une description précise et complète de ces étapes, je vous recommande la lecture du *Handbook of research synthesis and meta-analysis* écrits par les statisticiens de renommée internationale que sont Harris COOPER, Larry V. HEDGES et Jeffrey C. VALENTINE^c.

Les méthodes mises en œuvre dans ces « analyses d'analyses » que sont les méta-analyses ont été initialement développées dans les années 1980 dans le domaine des Sciences de l'éducation : le terme « méta analyse » apparaît pour la première fois en 1976 dans un article écrit par Gene GLASS qui jette les bases de cette nouvelle science. Mais c'est dans le domaine médical que les méta-analyses vont se développer avec le plus de succès au tournant du millénaire, avec notamment le déploiement de la célèbre *Collaboration Cochrane*^d. Cette

^a Contrairement aux études primaires qui sont des études prospectives.

^b « Quantitative procedures that a research synthesist may use to statistically combine the results of studies », dans *Chapter 1. Research synthesis as a scientific process*, H. COOPER, L. HEDGES et J. VALENTINE (2019, troisième édition), p.7.

^c La plupart des recherches dont il sera question ici ont été menées dans les pays anglo-saxons ; il n'existe pas à ma connaissance de document écrit en langue française sur les méta-analyses dans le domaine des Sciences de l'éducation.

^d Réseau international de volontaires qui réalisent, tiennent à jour et disséminent des méta-analyses sur des interventions thérapeutiques et préventives depuis 1993, www.cochrane.org

méthode scientifique à part entière s'est développée dans le cadre plus vaste de l'*evidence-based medicine* ou de l'*evidence-based education* dans le domaine qui nous intéresse, dans le but de traiter de questions médicales, sociales, voire politiques en s'appuyant sur une approche scientifique fondée sur l'exigence de preuves.

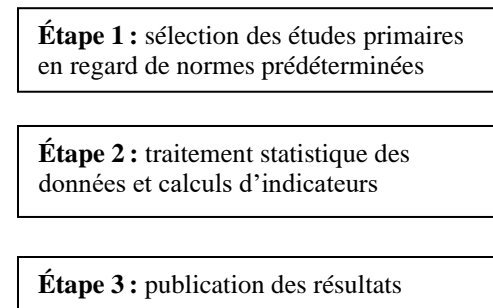



Figure 2 : étapes suivies pour mener une méta-analyse

Analyse statistique des données

L'analyse statistique des données concourt largement à la validité scientifique des conclusions publiées aussi bien par les études par comparaison de groupes (les études primaires) que par les méta-analyses (les études secondaires). Elle s'appuie sur les éléments complexes de la théorie des probabilités, et une façon d'aborder ces questions pratiques (que dire des données une fois l'expérience menée ? une fois les études rassemblées ?) serait donc de comprendre dans un premier temps ces concepts pour ensuite les appliquer à des cas concrets. J'ai choisi ici de suivre une autre voie et, tout en reconnaissant l'importance de cette assise théorique fondamentale, je vous propose de démarrer par des situations concrètes et de faire appel à ces connaissances théoriques au fur et à mesure que le besoin s'en fait sentir. Les connaissances requises pour comprendre ce qui suit n'excèdent pas ce qu'un bon élève de lycée connaît des mathématiques. Ce livre est divisé en trois parties. La première est consacrée à la description d'une série de données (on parle de **statistiques descriptives**), et s'attachera particulièrement à analyser les distributions dites *normales*. Les notions de probabilités mises en œuvre par la suite y seront également présentées. La seconde partie traitera ensuite de la comparaison entre deux échantillons extraits d'une population quand l'un d'eux a été soumis à un traitement qui intéresse le chercheur (on parle de **statistiques inférentielles**). Dans la troisième partie, les règles qui encadrent la réalisation des méta-analyses et plus particulièrement l'analyse quantitative de leurs données seront détaillées. Vous trouverez également à la fin de ce livre des annexes comprenant notamment des formulaires et la liste des fonctions à utiliser avec le tableur Excel^a (le pictogramme  dans le texte principal signalera la présence en annexe 9 d'une information pertinente). Des exemples, dont les données sont soit réelles soit fictives sont proposés tout au long du livre et les fichiers Excel correspondants peuvent être téléchargés sur le site mathadoc.fr

^a Cet outil va vite s'avérer indispensable si vous souhaitez reproduire les calculs proposés dans ce document.

L'analyse statistique des données est un domaine complexe dont les subtilités ne sont connues que d'un nombre limité de spécialistes. Ainsi les équipes de chercheurs mobilisées par le What Works Clearinghouse pour analyser la littérature scientifique sont pluridisciplinaires et comprennent toujours des experts en statistiques. Le but de ce livre n'est pas de former aux techniques pointues utilisées par ces spécialistes (formés après plusieurs années d'études et d'expériences professionnelles), mais bien plus modestement de donner à voir ce que leur expertise peut nous apporter et de permettre à tout un chacun de comprendre quelques-uns des concepts utilisés dans les méta-analyses. Les sources d'information complémentaires qui vous sont proposées tout au long du livre devraient vous permettre d'approfondir dans un second temps ce domaine scientifique dont nous n'aurons fait qu'entrevoir la richesse et la complexité.

Première partie : les données et le hasard

Chapitre 1. Description des données

La seule collecte de centaines, voire de milliers de données, ne dit pas grand-chose au chercheur de la population qu'elles sont censées décrire : c'est leur analyse descriptive qui va finalement leur donner un sens, sens qui dépend bien sûr du contexte et de la question de recherche. Ces données vont alors donner naissance à une poignée d'indicateurs qui vont les « faire parler ». Ils sont nombreux à pouvoir être utilisés et nous nous limiterons ici à ceux qui sont le plus souvent calculés dans les études quantitatives qui nous intéressent. Une seconde façon de décrire une collection de données est d'en proposer une représentation graphique. On verra alors en utilisant des histogrammes qu'un simple coup d'œil peut donner rapidement du sens à d'immenses séries de données.


Pour mieux comprendre les concepts utilisés en statistique descriptive, nous nous appuyerons dans ce chapitre sur deux exemples. Dans le premier exemple, la série de données brutes sera notre point de départ : il s'agit des poids d'un groupe de 252 hommes (**exemple 1**). Pour le second exemple (**exemple 2**), nous analyserons les scores de presque 30 000 élèves âgés de 15 ans habitant l'OCDE. Ces données sont continues : en effet, le poids ou le score d'une personne peuvent prendre n'importe quelle valeur comprise dans un intervalle de nombres réels donné. Dans les faits, on se limite à des valeurs arrondies (on s'arrête ainsi au chiffre des unités pour les scores, au chiffre des dixième de kilogramme pour le poids), ce qui revient à rassembler les données dans des classes dont on a fixé la largeur (voir plus loin). Il existe d'autres types de données, comme des données non numériques dites qualitatives (la couleur des yeux par exemple) ou des données numériques discrètes (comme le nombre lu sur la face d'un dé).

Moyenne et écart-type

Dans notre premier exemple nous disposons des poids d'un groupe de 252 hommes mesurés en kilogramme au dixième près (tableau 1)^a.

Nous allons calculer dans un premier temps la moyenne^b de cette série de données. D'une façon générale, si on considère un ensemble de n données notées $x_1, x_2, x_3, \dots, x_n$, la moyenne μ^c de cette série de données est égale à la somme^d de toutes les données, divisée par le nombre de données. Donc

$$\mu = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{\sum_i x_i}{n}$$


Dans notre exemple, nous calculons une moyenne de 81,2 kg ^e.

^a <http://lib.stat.cmu.edu/datasets/bodyfat>

^b Sans mention particulière dans ce texte, le terme « moyenne » fera référence à la moyenne arithmétique.

^c Vous comprendrez au chapitre 5 pourquoi cette moyenne n'est pas notée m ici.

^d $\sum_i x_i = \sum_{i=1}^n x_i$. C'est la première notation qui sera le plus souvent utilisée dans ce texte, le nombre de termes de la série (ici n) étant systématiquement mentionné par ailleurs.

^e Le pictogramme  indique qu'une information est disponible en annexe 9 concernant l'utilisation d'Excel pour un calcul.

75,6	85,2	76,3	96,5	80,2	78,6	75,7	72,5	85,3	70,8	94,6	93,7	65,2	101,2	69,1	109,7	66,2
101,2	94,7	75,3	88,5	72,8	72,5	63,7	98,1	76,3	88,3	78,4	99,3	67,7	70,1	90,4	70,1	69,5
63,2	62,3	69,3	61,8	89,8	82,3	91,3	91,9	81,5	98,0	81,1	87,7	80,7	93,2	83,2	68,7	70,2
112,2	87,0	91,7	89,2	164,7	92,1	119,2	93,0	98,4	96,2	56,8	74,5	60,6	67,4	61,6	57,8	71,8
70,4	71,1	76,0	66,6	72,9	56,7	64,9	67,2	73,7	80,6	73,1	77,7	74,3	68,2	86,3	77,5	76,2
71,1	90,8	77,8	93,3	82,8	61,9	80,4	68,6	88,9	83,6	63,5	99,2	98,4	75,4	101,9	103,5	78,4
75,7	71,6	72,6	80,2	79,8	80,3	81,5	75,0	87,3	83,6	101,8	85,6	73,7	71,0	89,4	90,0	78,8
95,5	91,7	83,9	69,4	110,8	87,8	101,9	73,8	81,6	70,9	76,2	75,9	77,5	80,9	68,0	90,9	83,5
85,0	93,7	84,0	72,7	68,7	73,0	75,7	80,5	69,1	87,2	75,0	77,9	77,7	89,4	71,2	76,3	84,4
78,4	89,2	80,3	75,1	90,8	92,2	88,0	76,4	77,5	83,1	80,9	73,9	79,5	71,7	80,4	81,2	86,6
70,0	78,6	69,9	83,8	83,6	95,4	82,1	79,8	86,6	89,9	84,5	98,0	81,9	93,1	85,2	73,8	88,8
94,9	83,3	96,0	81,2	90,9	63,6	67,5	68,6	72,2	59,6	67,1	60,4	72,9	82,6	72,7	76,2	99,1
104,3	73,4	64,5	81,5	57,4	76,9	90,0	79,2	76,1	67,0	82,7	79,6	73,4	71,6	76,5	86,9	99,4
69,1	57,0	80,4	79,9	102,9	65,9	68,5	109,4	84,9	106,5	99,4	53,8	66,1	72,2	77,3	76,0	105,6
70,4	86,1	57,8	101,8	106,3	103,3	90,5	70,5	97,7	60,9	91,2	84,7	86,5	94,1			

Tableau 1 : poids de 252 hommes en kg.

Quand plusieurs données ont la même valeur, on dit que cette valeur est une **modalité** de la série. On peut alors calculer la **fréquence** de cette modalité : c'est le quotient du nombre de données égales à cette modalité (ou encore l'**effectif** de la modalité) par le nombre total de données. Par exemple, on remarque que 3 données ont comme modalité 81,5 kg, donc sa fréquence est égale à $3 \div 252 \approx 0,012$ ou encore 1,2%.

Si on désigne par p le nombre de modalités et k_i le nombre de données qui ont comme valeur la modalité x_i (on a donc $\sum_{i=1}^p k_i = n$), la fréquence de la modalité x_i est égale à $f_i = \frac{k_i}{n}$ et la somme de ces fréquences est égale à 1, donc $\sum_i f_i = 1$.

La moyenne peut alors être calculée en utilisant la formule suivante (on verra l'intérêt de ce type de calcul au chapitre suivant) :

$$\mu = \frac{k_1 x_1 + k_2 x_2 + \dots + k_i x_i + \dots + k_p x_p}{n} = \sum_{i=1}^p \frac{k_i}{n} \times x_i = \sum_{i=1}^p f_i \times x_i$$

La moyenne est un indicateur de position central. Pour évaluer la dispersion de la série de données autour de cette moyenne on va calculer sa variance, notée *var*, égale à la moyenne des carrés des écarts à la moyenne. On obtient alors une première formule, puis une deuxième (égalité non démontrée ici) très utilisée pour les calculs

$$var = \frac{\sum_i (x_i - \mu)^2}{n} = \frac{\sum_i x_i^2}{n} - \mu^2$$

Et à nouveau, avec plusieurs données de même valeur, on peut utiliser les fréquences :

$$var = \frac{\sum_{i=1}^p k_i (x_i - \mu)^2}{n} = \sum_{i=1}^p f_i \times (x_i - \mu)^2 = \sum_{i=1}^p f_i \times x_i^2 - \mu^2$$

Dans notre exemple, on calcule une variance de 177 kg² \square .

Plus les données sont dispersées autour de la moyenne, plus la variance est grande ; si toutes les données sont égales à leur moyenne, la variance est égale à zéro. L'unité de la variance est égale à l'unité des mesures élevée au carré. Ainsi, quand les données sont en kg, la variance est exprimée en kg². C'est pour cette raison que l'on utilise très souvent la racine carrée de la variance que l'on nomme l'écart-type σ^a (*standard deviation* en anglais avec comme autre abréviation SD) pour évaluer la dispersion des données autour de la moyenne. En effet, l'écart-type a le mérite d'être dans la même unité que les données. On a donc

$$\sigma = \sqrt{\text{var}} = \sqrt{\frac{\sum_i (x_i - \mu)^2}{n}}$$

Dans de nombreux documents la variance est notée σ^2 .

Dans notre exemple, on calcule un écart-type de 13,3 kg \square .

Moyenne et écart-type sont les véritables vedettes de l'analyse descriptive des séries de données. On donne en général d'abord la moyenne de la série, suivi de l'écart-type entre parenthèses. L'intérêt de la variance réside dans ses propriétés mathématiques dont l'écart-type est privé comme on le verra plus tard.

On va classer les données de notre exemple par ordre croissant et les représenter sous la forme d'un nuage de points : chaque point représente un homme, avec en abscisse son poids et en ordonnée son numéro dans la série (figure 3).

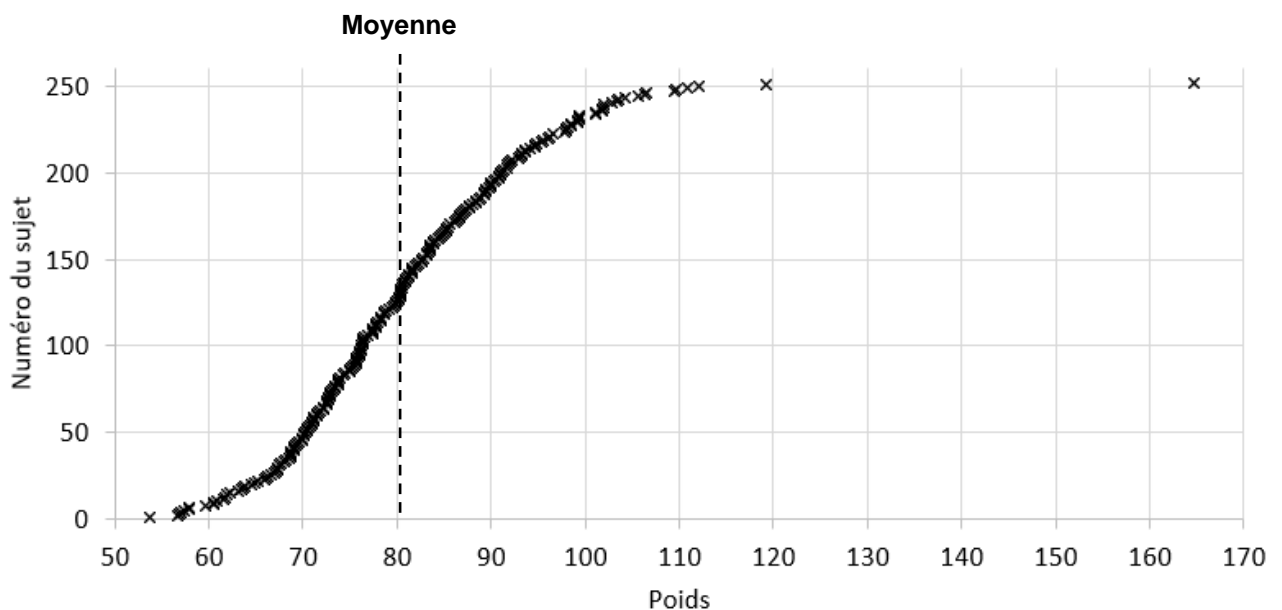


Figure 3 : poids de 252 hommes (nuage de points)

Centrer et réduire une série de données

Il est possible, à partir de n'importe quelle série de données de moyenne et d'écart-type connus, d'appliquer une transformation afin d'obtenir une nouvelle série de données de moyenne et d'écart-type choisis.

^a Vous comprendrez au chapitre 5 pourquoi cet écart-type n'est pas notée s ici.

Prenons comme exemple une série de n données x_i dont on a calculé la moyenne μ et l'écart-type σ . On définit une nouvelle série de n données y_i de la façon suivante :

$$y_i = a + \frac{x_i - \mu}{\sigma} \times b$$

On peut vérifier par le calcul que cette nouvelle série de données a comme moyenne a et comme écart-type b .

On rencontre très souvent la transformation suivante :

$$z_i = \frac{x_i - \mu}{\sigma}$$

Vous pourrez vérifier que la moyenne de cette série de données est égale à zéro et son écart-type égal à un. Dans ce cas, on dit qu'on a **centré et réduit les données**, ou encore qu'on les a **standardisées** et cela revient finalement à s'affranchir des unités et des échelles utilisées pour les mesures. On donne à cette nouvelle donnée le nom de score-z. On a également

$$x_i = \mu + z_i \times \sigma$$

Le nombre z_i est en fait le nombre d'écart-types dont un score x_i est éloigné de la moyenne ; il est positif quand le score est supérieur à la moyenne, et négatif quand le score est inférieur à la moyenne.

Pour notre exemple, cela reviendrait à appliquer la transformation suivante :

$$z_i = \frac{x_i - 81,2}{13,3}$$

On a appliqué cette transformation à 4 poids (tableau 2).

Poids (kg)	scores z
110,8	2,2
101,2	1,5
70	-0,8
60,4	-1,6

Tableau 2 : scores-z de 4 poids

Ces informations ont été signalées sur le nuage de points (en omettant cette fois le point extrême, figure 4). Chaque double flèche sous l'axe des abscisses représente un écart-type.

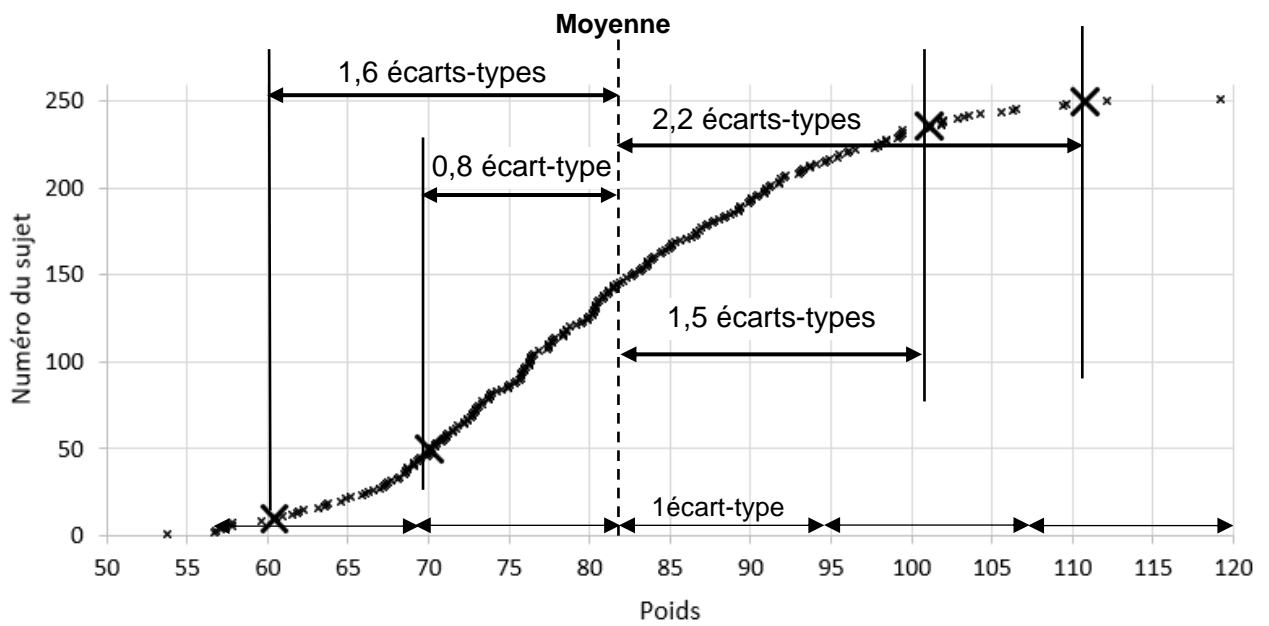


Figure 4 : 4 poids particuliers

Quand on centre et réduit toutes les données, on obtient le nuage de point suivant (les 4 points particuliers sont à nouveau signalés).

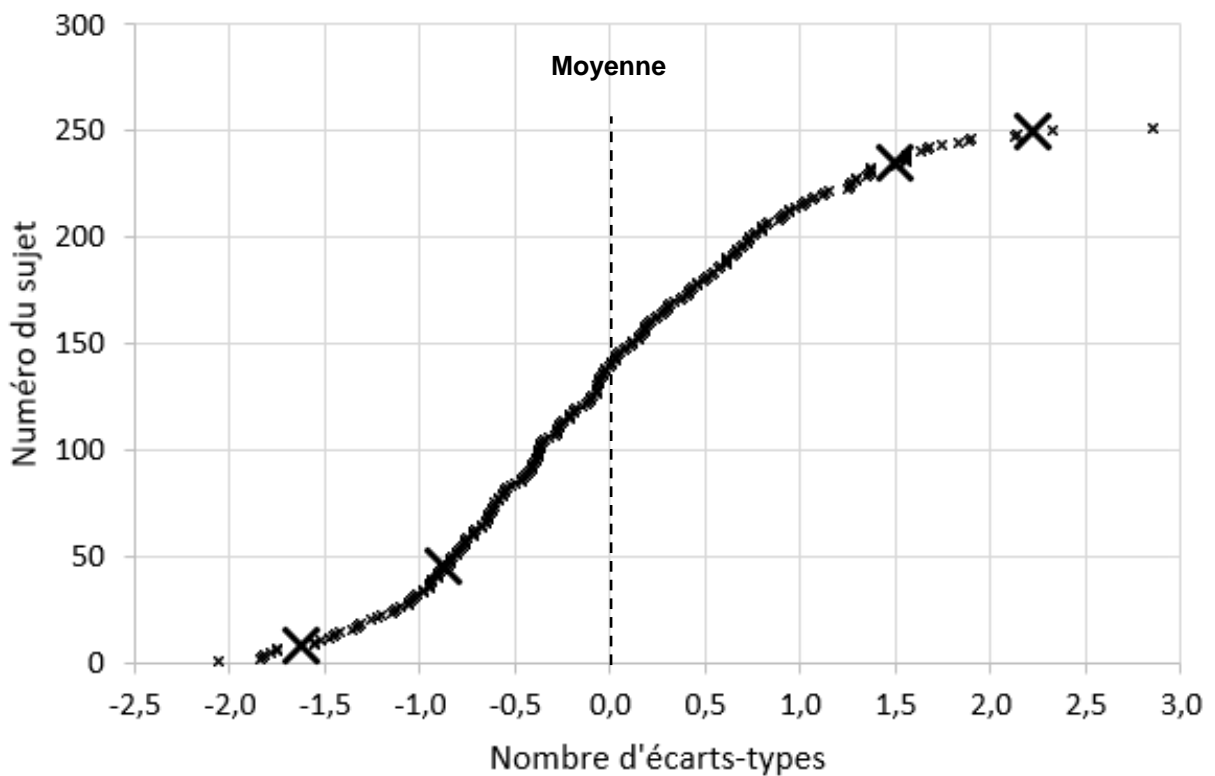


Figure 5 : poids de 252 hommes centrés-réduits

Dans ce cas, l'axe des abscisses est tout simplement gradué en écarts-types.

Histogramme

On va poursuivre la description de notre ensemble de données en visualisant leur répartition. Dans un premier temps on partage la série des données classées par ordre croissant en plusieurs classes de même largeur. Reprenons notre exemple des poids de 252 hommes en partageant la série en classes de largeur 5 kg. Comptons ensuite le nombre de sujets dont le poids appartient à chaque classe (ce sont les effectifs, deuxième ligne du tableau 2 ci-dessous) Σ . Si on divise ces nombres par le nombre total de sujets inclus dans l'étude (252 ici), on obtient les fréquences des classes qui sont des nombres compris entre 0 et 1 que l'on peut aussi donner sous la forme d'un pourcentage. La somme des fréquences de chaque classe est égale à 1 (ou 100 %). On partage également la série en classes de largeur 2,5 kg cette fois et on calcule à nouveau les fréquences des classes (tableau 3). Bien sûr, plus la largeur des classes est importante, plus la fréquence de chaque classe est importante. Pour s'affranchir de cette dépendance, on va diviser les fréquences par la largeur des classes : on obtient alors une **densité de fréquence**, ce sont les dernières lignes des tableaux 3 et 4.

Quand on associe à chaque classe sa densité de fréquence, on définit une **distribution** des données.

Borne supérieure de la classe	50*	55**	60	65	70	75	80	85	90	95	100	105	110	115	120	170	Total
Effectifs	0	1	7	13	26	39	40	39	27	25	16	10	5	2	1	1	252
Fréquences (%)	0	0,4	2,78	5,16	10,3	15,5	15,9	15,5	10,7	9,92	6,35	3,97	1,98	0,79	0,4	0,4	100
Densité fréquence	0,000	0,001	0,006	0,010	0,021	0,031	0,032	0,031	0,021	0,020	0,013	0,008	0,004	0,002	0,001	0,001	0,200

Tableau 3 : distribution des poids dans des classes de largeur 5 kg

* : lire poids « inférieurs ou égaux à 50 kg ».

** : lire poids « strictement supérieurs à 50 kg et inférieurs ou égaux à 55 kg ». Et ainsi de suite.

Borne supérieure de la classe	50,*0	52,5**	55,0	57,5	60,0	62,5	65,0	67,5	70,0	72,5	75,0	77,5	80,0	82,5	85,0
Effectifs	0	0	1	4	3	7	6	10	16	20	19	24	16	22	17
Fréquences (%)	0	0	0,4	1,59	1,19	2,78	2,38	3,97	6,35	7,94	7,54	9,52	6,35	8,73	6,75
Densité fréquence	0,000	0,000	0,002	0,006	0,005	0,011	0,010	0,016	0,025	0,032	0,030	0,038	0,025	0,035	0,027

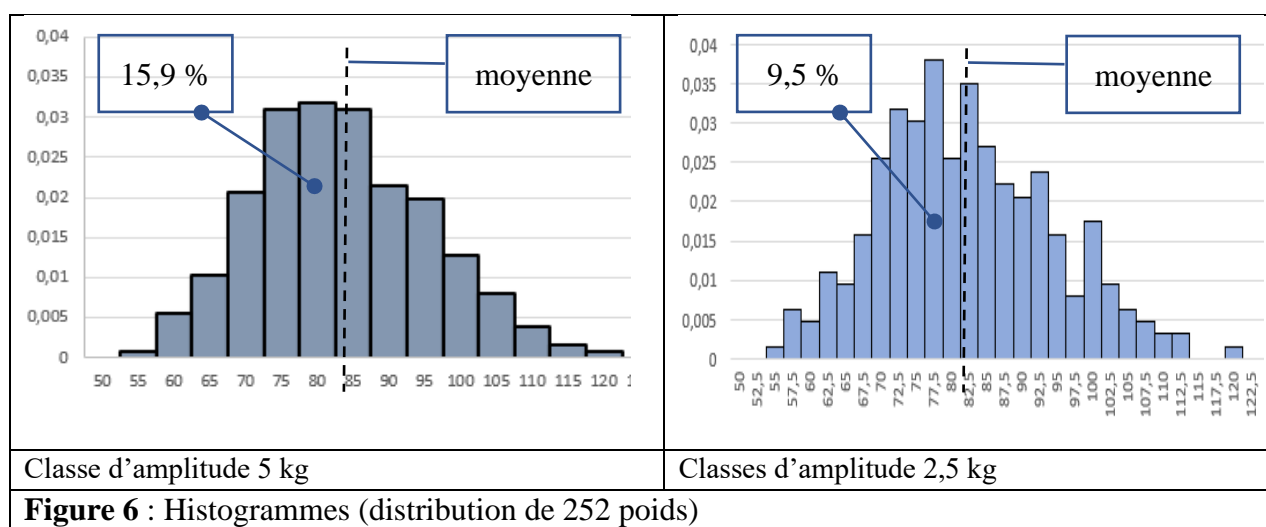
Borne supérieure de la classe	87,5	90	92,5	95	97,5	100	102,5	105	107,5	110	112,5	115	117,5	120	170	Total
Nombre	14	13	15	10	5	11	6	4	3	2	2	0	0	1	1	252
Fréquence (%)	0,056	0,052	0,06	0,04	0,02	0,044	0,024	0,016	0,012	0,008	0,008	0	0	0,004	0,004	100
Densité fréquence	0,022	0,021	0,024	0,016	0,008	0,017	0,010	0,006	0,005	0,003	0,003	0,000	0,000	0,002	0,002	0,4

Tableau 4 : distribution des poids dans des classes de largeur 2,5 kg

* : lire « poids inférieurs ou égaux à 50 kg ».

** : lire « poids strictement supérieurs à 50 kg et inférieurs ou égaux à 52,5 kg ». Et ainsi de suite.

On représente ensuite graphiquement ces distributions des données par une série de rectangles dont la base est à chaque fois la largeur de la classe (les bornes des classes sont placées en abscisse) et la hauteur égale à la densité de fréquence associée à chaque classe. Ce type de graphique est appelé un **histogramme**^a (figure 6, la dernière classe n'est pas représentée). Excel permet bien évidemment de tracer ces histogrammes mais deux défauts sont à noter. En premier lieu, l'utilitaire dénommé *Histogramme* ne permet en fait que de calculer des effectifs de classes (et génère par conséquent un diagramme en bâton et non un histogramme). Le second provient du décalage vers la gauche des bases des rectangles sur l'axe des abscisses de la moitié de la largeur de la classe : ce sont les bornes supérieures des classes qui figurent sous le rectangle dont la hauteur donne la densité de fréquence correspondante, et non le centre de la classe^b.



La surface de chaque rectangle est obtenue en multipliant la hauteur du rectangle, donc la densité de fréquence, par sa largeur ; elle est donc égale à la fréquence de la classe correspondante. Et bien entendu, la somme des surfaces est égale à 100 %. L'allure de ces deux histogrammes est similaire : on observe une certaine symétrie de part et d'autre de la moyenne, et plus on s'éloigne de la moyenne, moins les données sont nombreuses.

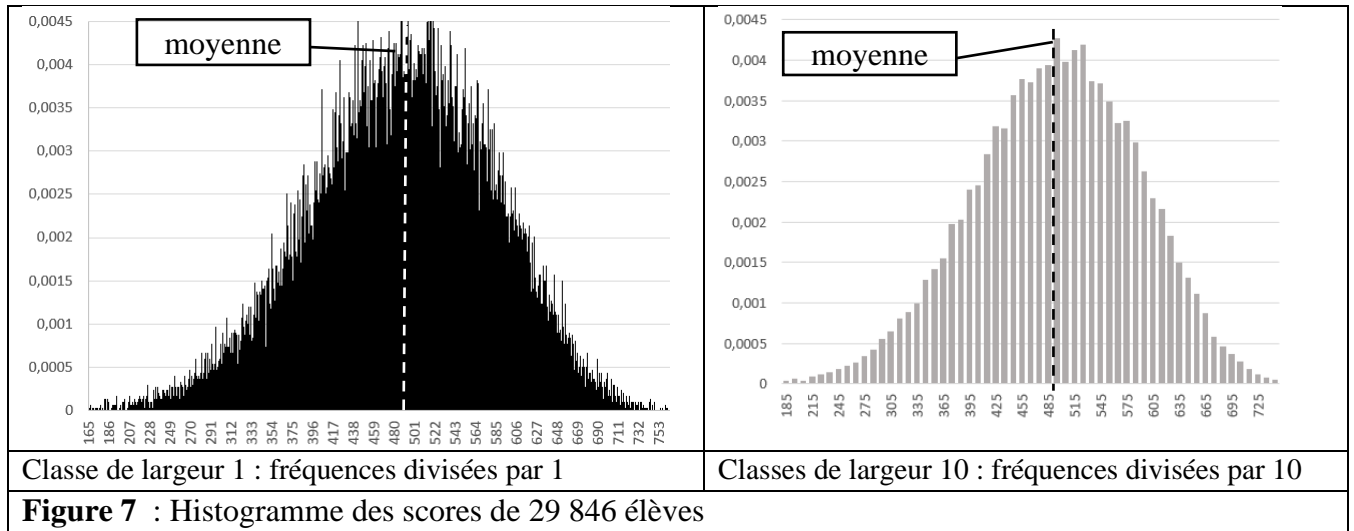
Comme second exemple (**exemple 2**), nous allons utiliser les scores obtenus en mathématiques par 29 846 élèves âgés de 15 ans. Ces scores ont été récoltés et publiés par l'OCDE dans le cadre de l'enquête PISA 2012^c. Le nombre important de données ne permet pas de les présenter individuellement dans un tableau et impose l'utilisation d'indicateurs pour les décrire, comme la moyenne, ici égale à 488,3 et l'écart-type, égal à 95,1. De la même façon que pour notre premier exemple, deux histogrammes ont été tracées (figure 7). Pour le premier, les scores ont

^a On appelle parfois histogramme un diagramme en bâtons dont les hauteurs sont proportionnelles à la fréquence des données (voire de leurs effectifs) ; si les classes sont toutes de même largeur, ce qui est très souvent le cas, la représentation graphique reste globalement la même.

^b Ce qui s'avérera gênant quand on superposera la courbe normale associée à la distribution, voir plus loin.

^c Les données peuvent être téléchargées ici : <http://dx.doi.org/10.1787/888932964813>. Pour en savoir plus sur l'enquête PISA 2012 : www.mathadoc.fr

été arrondi par troncature à l'unité, ce qui revient à les regrouper par classes de largeur 1. Pour le second, les scores ont été regroupés dans des classes de largeur 10. L'allure de ces deux histogrammes est la même que celle des histogrammes de notre premier exemple, épousant encore une fois la forme d'une cloche symétrique de part et d'autre de la moyenne.



Modéliser la distribution des données

On pourrait multiplier les exemples qui aboutissent à des observations similaires. De nombreux caractères / caractéristiques de sujets végétaux ou animaux mais aussi de produits issus de l'industrie ont une distribution dont l'histogramme épouse la forme d'une cloche centrée autour de la moyenne. De là à imaginer une fonction continue qui pourrait modéliser ce type de distribution où la moyenne ainsi que l'écart-type de la série de données joueraient un rôle particulier, il n'y a qu'un pas à franchir. Modéliser la distribution des résultats d'une observation ou d'une expérience par une fonction mathématique connue permet de s'appuyer sur les propriétés de celle-ci. La description en devient plus simple et il devient possible d'émettre des hypothèses qui vont au-delà de la seule observation.

Les distributions de caractères dits « naturels » dans certaines populations mais aussi la répartition d'erreurs (de mesure par exemple) ou de défauts de fabrication sont très souvent modélisés par la loi normale^a. Cette loi joue un rôle important dans la théorie des probabilités (voir plus loin le théorème central limite), et c'est ce rôle qui permet d'expliquer pourquoi elle semble si bien servir de modèle à de nombreuses distributions de données. Mais avant de décrire cette loi normale, nous devons faire une première incursion dans le monde des probabilités.

^a Il existe d'autres modèles (la loi bêta par exemple) qui pourraient également convenir dans certaines situations.

Chapitre 2. Variables aléatoires

Le hasard n'a, a priori, rien à faire dans la description d'une série de données. Mais nous allons découvrir que si de très nombreuses données sont *normalement* distribuées (leur distribution semble suivre une loi normale), c'est bien que le hasard a joué un morceau de la partition. Nous aurons également besoin de nous appuyer sur certains concepts de la théorie des probabilités pour comprendre les procédures mises en œuvre en statistique inférentielle dans la deuxième partie de ce livre. Nous nous contenterons ici des notions qui nous sont indispensables pour la suite. La lecture des 120 pages de *Probabilités et statistiques en S5 IFIPS*^a écrit par J.-P. LENOIR en 2008 devrait satisfaire les plus curieux.

Variabes aléatoires discrètes

On va considérer un exemple fictif simple (**exemple 3**) pour définir quelques notions essentielles à la théorie des probabilités. Dans cet exemple, on va imaginer qu'un groupe de 2000 élèves a répondu à un test et que chaque élève a obtenu un score qui est un nombre entier compris entre 1 et 5. La distribution des scores est donnée dans le tableau ci-dessous, et la moyenne μ , la variance σ^2 et l'écart-type σ de cette série de données ont été calculées. L'histogramme^b correspondant est également présenté (ici, la densité de fréquence est égale à la fréquence car l'amplitude des classes est égale à 1, figure 8).

Modalités des scores	1	2	3	4	5	total
Effectifs	200	400	500	300	600	2000
Densité de fréquences	0,1	0,2	0,25	0,15	0,3	1

Moyenne (μ)	3,35
Variance (σ^2)	1,8275
Écart-type (σ)	1,35185

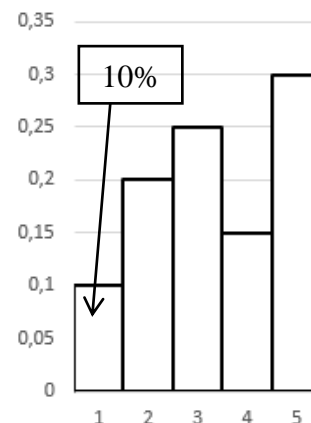


Figure 8 : distribution des scores (exemple 3)

Imaginons maintenant qu'on choisisse une copie au hasard. Chaque copie a une chance sur 2000 d'être choisie ; on dira que la **probabilité** de prendre une copie au hasard est égale à $1/2000=0,0005$. Il y a 200 copies ayant comme score 1 ; on aura donc une probabilité de choisir une copie dont le score est 1 égale à $200 \times 0,0005 = 0,1$. Ce nombre est aussi la fréquence d'apparition de la modalité 1. On peut faire la même remarque pour les autres modalités.

^a Ce texte s'adresse à des personnes familiarisées avec une présentations rigoureuses des mathématiques

^b Les données ne sont pas continues donc le terme « histogramme » est quelque peu usurpé, sauf à considérer que l'enseignant aurait pu choisir de mettre n'importe quel score compris entre 0 et 5 et à arrondir les notes à l'entier supérieur ou égal.

Dans le cadre théorique des probabilités, choisir une copie au hasard et s'intéresser à son score, définit une **variable aléatoire** : on ne connaît pas le résultat de cette action à l'avance mais l'ensemble des résultats possibles est connu (ici les nombres 1, 2, 3, 4, ou 5). Quand on choisit une copie, on obtient un des 5 scores et ce nombre est une **réalisation** de la variable aléatoire. Si à chaque résultat possible on peut associer une probabilité, on définit ce faisant une **loi de probabilité**. Le nombre de réalisations de cette variable aléatoire est fini : on dit que c'est une variable aléatoire **discrète**.

On a comme habitude de désigner les variables aléatoires par des lettres capitales. Par exemple, on va noter ici X la variable aléatoire^a : « choisir une copie au hasard et noter son score ». Les réalisations possibles de cette variable aléatoire seront notées x_i (ici on utilise des minuscules). Dans notre exemple, $i = 1, 2, 3, 4, 5$, $x_i = i$ et on sait, entre autres, que la probabilité que la réalisation de la variable aléatoire X soit égale à 1 est égale à 0,1 ; on écrira $p(X = 1) = 0,1$. On présente ces informations dans un tableau et on en donne une représentation graphique avec en abscisse les réalisations, en ordonnée la densité de probabilité (c'est une probabilité divisée par la largeur des classes ici égale à 1), les probabilités de chaque réalisation sont alors représentées par les surfaces de chacun des rectangles (figure 9).

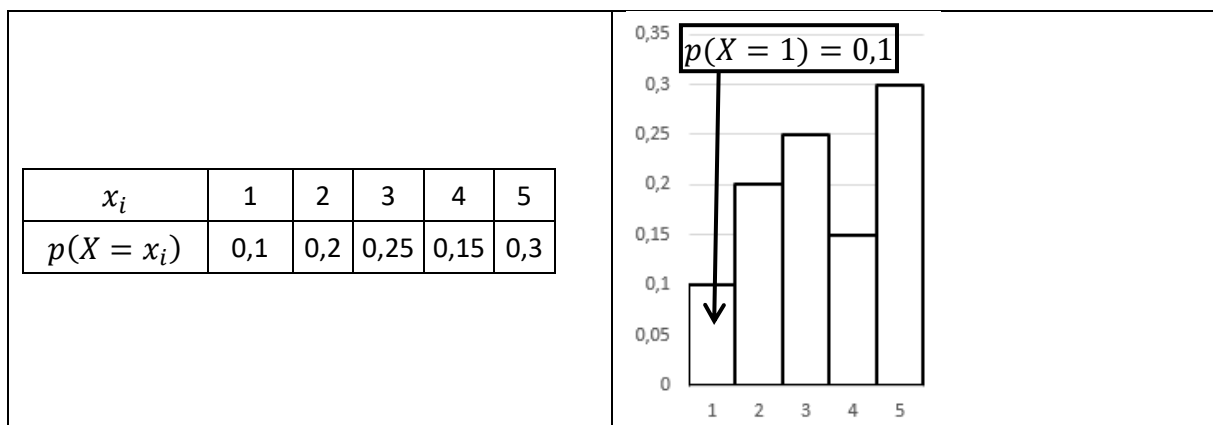


Figure 9 : Probabilités (exemple 3)

On reprend maintenant les informations qui étaient présentées dans le premier tableau, et en notant f_i la fréquence d'apparition du score i , on remarque que

$$p(X = x_i) = f_i$$

Cela signifie tout simplement que la probabilité d'obtenir au hasard un score donné est égal à la fréquence de ce score. Cet exemple pourrait être généralisé à toute situation équivalente, qui consiste à prélever au hasard une donnée dans une série de données dont on connaît la distribution. Et dans tous ces cas, l'égalité précédente sera vérifiée.

On peut également considérer une probabilité comme la limite d'une fréquence calculée à partir d'expériences répétées un nombre infini de fois ; la loi de probabilité, ou encore **distribution**

^a Une variable aléatoire associée à un événement (ici choisir une copie au hasard) un résultat (ici le score écrit sur la copie) ; nous n'aurons pas besoin ici de développer cette notion mathématique et nous nous contenterons donc de cette présentation extrêmement réduite.

de probabilité, est définie par les fréquences limites associées à chaque résultat (ici les scores)^a. Dans ce cas, on dira que la proportion des réalisations de la variable aléatoire X qui prennent la valeur définie x_i est égale à f_i , quand le nombre de ces réalisations tend vers l'infini. La valeur que l'on s'attend à trouver, en moyenne, si l'on répétait un nombre infini de fois une même expérience aléatoire a été baptisée **l'espérance** de cette variable aléatoire, notée $E(X)$. On définit également l'espérance des carrés des écarts de la variable aléatoire X à son espérance comme étant la **variance** de cette variable aléatoire, notée $var(X)$: elle mesure la dispersion de X autour de son espérance^b. On définit enfin **l'écart-type** de la variable aléatoire, égale à la racine carrée de la variance^c. Espérance, variance et écart-type d'une variable aléatoire sont des **paramètres** de cette variable aléatoire. Quand on prélève au hasard une donnée dans une série de données dont on connaît la distribution, on va admettre ici que

$$E(X) = \sum_i p(X = x_i) \times x_i$$

$$var(X) = E \left[(X - E(X))^2 \right] = E(X^2) - (E(X))^2$$

Ce qui donne, d'après la définition de l'espérance,

$$var(X) = \sum_i p(X = x_i) \times (x_i - E(X))^2 = \sum_i p(X = x_i) \times x_i^2 - (E(X))^2$$

Si on remplace $p(X = x_i)$ par f_i , et si on utilise les remarques du chapitre précédent, on a

$$E(X) = \sum_i f_i \times x_i = \mu \quad \text{et} \quad var(X) = \sum_i f_i \times x_i^2 - \mu^2 = \sigma^2$$

Enfin, l'écart-type de la variable aléatoire X est égal à $\sqrt{var(X)} = \sigma$

Et donc finalement, pour toute variable aléatoire du type « prendre au hasard une donnée dans une série de données dont on connaît la distribution », l'espérance, la variance et l'écart-type de cette variable aléatoire sont respectivement la moyenne, la variance et l'écart-type de cette série de données.

Bilan

Pour toute variable aléatoire X consistant à prélever au hasard un élément dans une série de données dont on connaît la moyenne μ et l'écart-type σ , on a

$$E(X) = \mu, \quad var(X) = \sigma^2 \quad \text{et} \quad \sqrt{var(X)} = \sigma$$

On va poursuivre avec notre exemple (**exemple 3**) et terminer l'analyse théorique de la variable aléatoire X en calculant son espérance et sa variance. On utilise les calculs de la moyenne et de la variance de la série de données effectués au début du chapitre, et on a donc

^a Nous verrons plus loin que cette conception des probabilités constitue le socle théorique sur lesquels se fonde la plupart des calculs mis en œuvre dans les méta-analyses.

^b Par analogie avec l'espérance, on peut définir la variance d'une variable aléatoire comme la variance calculée pour des réalisations obtenues après avoir répété l'expérience un nombre infini de fois.

^c Ces définitions sont bien sûr à rapprocher des définitions de la moyenne, la variance et l'écart-type vues auparavant (et l'espérance est aux variables aléatoires ce que la moyenne est à une série de données).

$$E(X) = 3,35, \text{var}(X) = 1,8275 \text{ et } \sqrt{\text{var}(X)} \approx 1,35185$$

Nous allons maintenant étudier cet exemple d'un point de vue expérimental et simuler l'expérience consistant à tirer une copie au hasard, noter son score et la remettre dans le paquet de 2000 copies, tout ceci grâce à Excel ^a. On va mener 4 expériences : pour la première, on effectuera 50 tirages au sort, pour la seconde 200, pour la troisième 1000 et pour la dernière 3000. On calcule pour chaque série de données obtenues leur moyenne, leur variance et leur écart-type ainsi que les fréquences d'apparition des différents scores (tableau 5). On représente également la distribution des scores dans 4 histogrammes (figure 10)^a.

Scores	Les expériences								La théorie
	50 copies		200 copies		1000 copies		3000 copies		Fréquences
	Effectifs	Fréquences	Effectifs	Fréquences	Effectifs	Fréquences	Effectifs	Fréquences	
1	8	0,16	23	0,115	109	0,109	298	0,099333	0,1
2	9	0,18	40	0,2	229	0,229	618	0,206	0,2
3	14	0,28	60	0,3	243	0,243	702	0,234	0,25
4	8	0,16	22	0,11	139	0,139	459	0,153	0,15
5	11	0,22	55	0,275	280	0,28	923	0,307667	0,3
Moyennes	3,1		3,23		3,252		3,363666667		3,35
Variances	1,85		1,8171		1,860496		1,854746556		1,8275
Écart-types	1,360147051		1,347998516		1,364		1,361890802		1,25872

Tableau 5 : 4 tirages au sort (exemple 3)

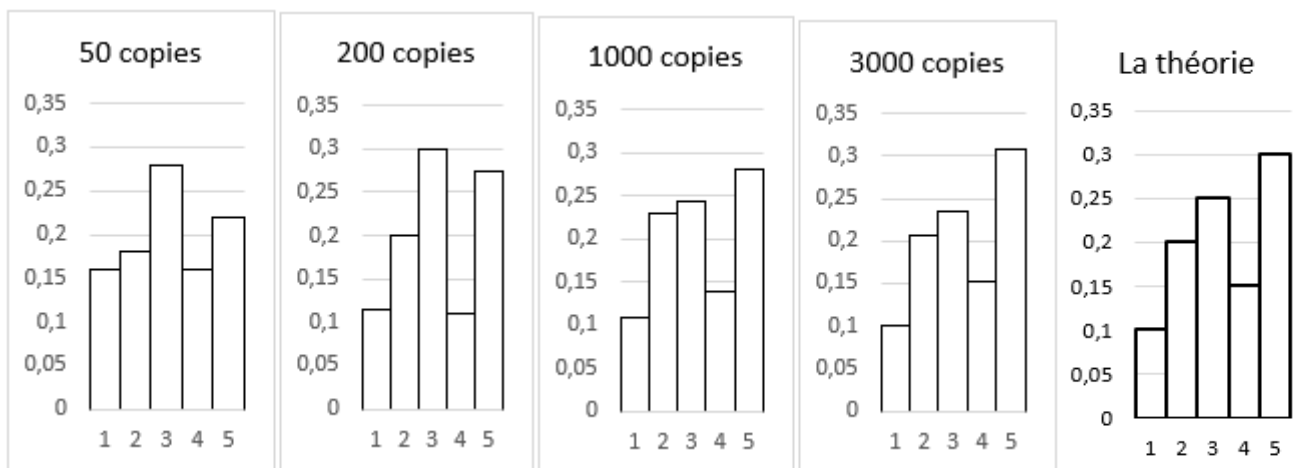


Figure 10 : histogrammes de l'exemple 3

Encore une fois, si on pouvait répéter un nombre infini de fois l'expérience, les fréquences tendraient à être égales aux probabilités théoriques.

^a Fichier Excel téléchargeable sur mathadoc.fr

Variables aléatoires continues

Il existe aussi des variables **aléatoires continues** : leurs réalisations peuvent prendre n'importe quelles valeurs appartenant à un intervalle. Ces variables suivent elles aussi des lois de probabilité ou lois de distribution ou encore **distributions de probabilité**, mais elles sont plus complexes à définir que dans le cas de variables discrètes. Et dans ce cas, c'est la probabilité que les réalisations de cette variable aléatoire soient comprises entre deux valeurs a et b données, qui sera calculée. On admettra ici que cette probabilité est égale à la surface comprise entre l'axe des abscisses, les deux droites verticales passant par a et b et la courbe représentative d'une fonction que l'on appelle la **densité de probabilité de la variable aléatoire** considérée^a (figure 11). Bien entendu, la surface comprise entre la courbe et l'axe des abscisse est égale à 1 (c'est la probabilité de l'événement certain).

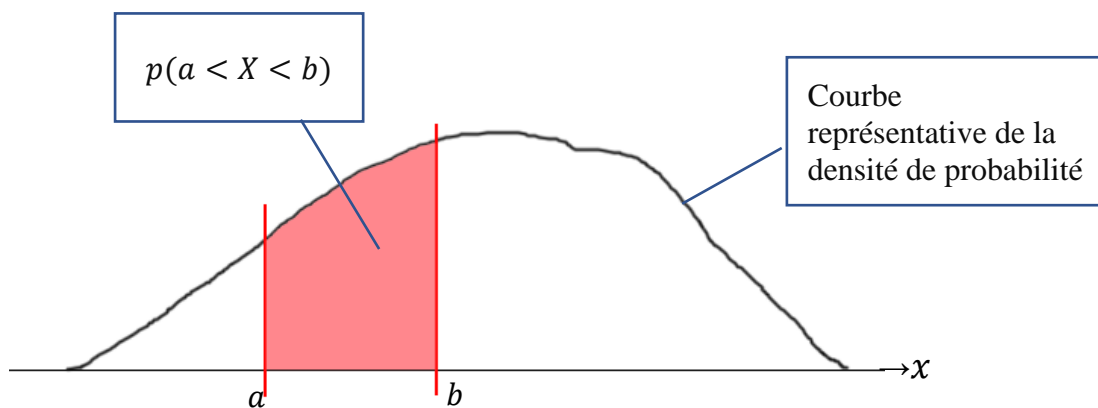


Figure 11 : densité de probabilité

Enfin, de la même façon que pour une variable aléatoire discrète, on définit alors l'espérance de X , notée $E(X)$ ^b sa variance, notée $var(X)$ et son écart-type $\sqrt{var(X)}$, calculés à partir de leur loi de probabilité^c.

Variables aléatoires centrées et réduites

Les variables aléatoires (discrètes ou continues) peuvent s'additionner et subir des transformations mathématiques courantes. Ainsi, on pourra définir, à partir de la variable aléatoire X une autre variable aléatoire notée Z telle que

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

Cette transformation est à rapprocher de celle que nous avons effectuée à partir d'une série de n données x_i (chapitre précédent). On dit ici aussi qu'on a centré et réduit ou encore standardisé

^a Cette surface est l'intégrale de la densité de probabilité sur $[a,b]$. Remarquons que cette définition convient également pour les variables aléatoires discrètes.

^b Dont le calcul utilise l'intégrale de la fonction de densité, qui est aux variables continues ce que la somme est aux variables discrètes.

^c Nous n'aurons pas besoin d'en savoir davantage sur ces notions. Voir LENOIR (2008) pour plus d'informations à ce sujet.

la variable X . On peut montrer que l'espérance de la variable Z est égale à 0 et que sa variance est égale à 1, et on dira que Z est une variable aléatoire centrée et réduite.

On va maintenant s'intéresser au cas d'une variable aléatoire continue dont la loi de probabilité est une loi dite « normale ».

Chapitre 3. La loi normale

La loi de Gauss (pour les allemands et les anglo-saxons) ou de Laplace (pour les français) a été baptisée loi « normale » car elle semble la plus adaptée pour modéliser de nombreuses observations (comme certains phénomènes naturels par exemple). Elle est d'une importance capitale dans le domaine des probabilités et se trouve au fondement des raisonnements développés dans les prochains chapitres.

On dira que la variable aléatoire continue X suit une loi normale (ou distribution normale) de **paramètres** μ et σ , notée $\mathcal{N}(\mu, \sigma)$ quand sa densité de probabilité est de la forme :

$$f_{\mu, \sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

On notera alors $X \sim \mathcal{N}(\mu, \sigma)$.

Nous n'aurons heureusement jamais besoin d'utiliser cette formule, y compris pour des calculs menés sur un tableur (ces derniers ont intégré des fonctions qui permettent d'obtenir des résultats numériques \square). On va admettre ici que les paramètres μ et σ sont l'espérance et l'écart-type de la variable aléatoire X .

$$E(X) = \mu \quad , \quad \text{var}(X) = \sigma^2 \quad \text{et} \quad \sqrt{\text{var}(X)} = \sigma$$

On va tracer les trois courbes représentatives des lois $\mathcal{N}(5,2)$, $\mathcal{N}(4,3)$ et $\mathcal{N}(2,1)$ (figure 12) \square .

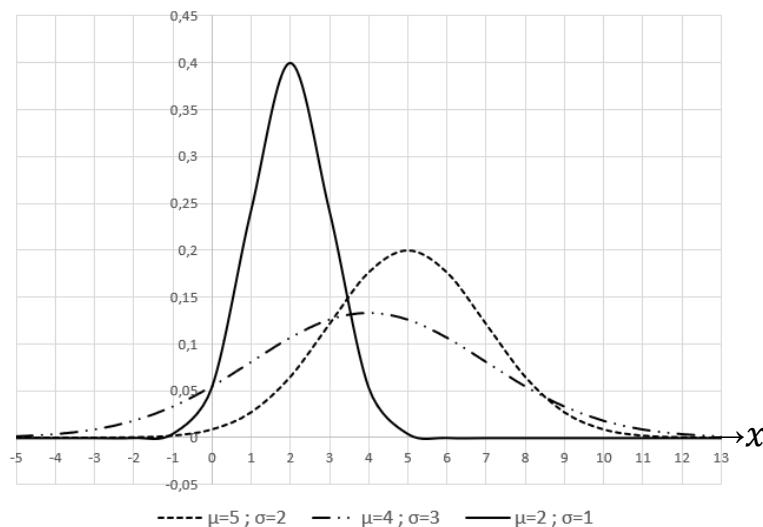


Figure 12 : représentation graphique de trois lois normales

On remarque que les maximums des trois courbes de la figure 12 sont atteints quand x est égal à μ que les courbes sont symétriques par rapport à une droite passant par le point d'abscisse μ et parallèle à l'axe des ordonnées. Également, vous noterez que plus la valeur de σ augmente, plus la courbe est aplatie. Ces courbes ont la forme de cloches parfaites qui font bien entendu

penser aux histogrammes du chapitre précédent. Nous irons plus loin sur ce point après avoir analysé un peu plus en détail ces représentations graphiques.

Les propriétés des courbes normales

On va maintenant graduer l'axe des abscisses en prenant comme point de repère l'espérance μ puis en ajoutant ou retranchant un même nombre d'écart-types σ de part et d'autre de cette espérance. Quelles que soient les valeurs μ et σ , les représentations graphiques auront une allure similaire (figure 13). L'axe des ordonnées a disparu, car on va s'intéresser ici aux surfaces qui vont permettre de déterminer des probabilités.

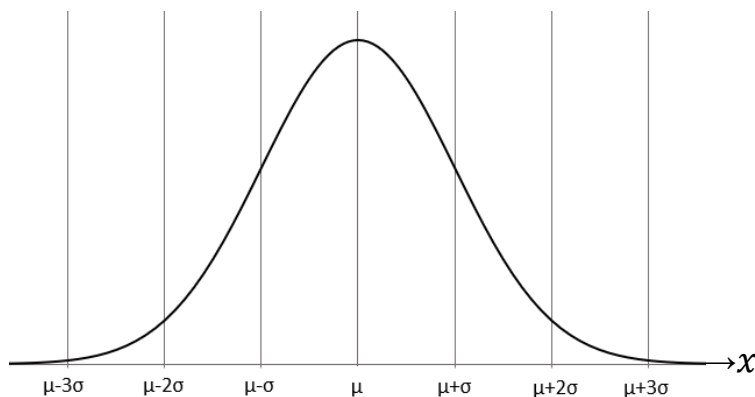


Figure 13 : représentation graphique générale de $\mathcal{N}(\mu, \sigma)$

On va admettre ici l'ensemble des propriétés énumérées ci-dessous.

La surface comprise entre la courbe et l'axe des abscisses est toujours égale à 1 (cela correspond à la probabilité d'un événement certain).

D'un simple coup d'œil, on peut constater que la courbe rejoint (presque) cet axe pour $x = \mu - 3\sigma$ et $x = \mu + 3\sigma$. C'est d'ailleurs pour cette raison qu'on limite souvent la représentation graphique à l'intervalle $[\mu - 3\sigma ; \mu + 3\sigma]$. En d'autres termes, si la variable aléatoire X suit la loi normale de paramètres μ et σ , on a

$$p(\mu - 3\sigma < X < \mu + 3\sigma) \approx 1$$

On peut aussi montrer que la surface comprise entre la courbe, l'axe des abscisses et la gauche d'une droite verticale passant par l'abscisse $\mu - \sigma$ est égale à 16% de la surface totale, et ce pour toutes les lois normales, c'est-à-dire quelles que soient les valeurs de μ et de σ . Le même type de remarque peut être faite quand la droite verticale passe par μ , $\mu + \sigma$, $\mu + 2\sigma$. On va même généraliser cette propriété à toutes valeurs du type $\mu + a \times \sigma$: les surfaces seront toujours égales à la même proportion de la surface totale, quelles que soient les valeurs de μ et de σ pour une valeur de a donnée (voir figure 14 ci-dessous).

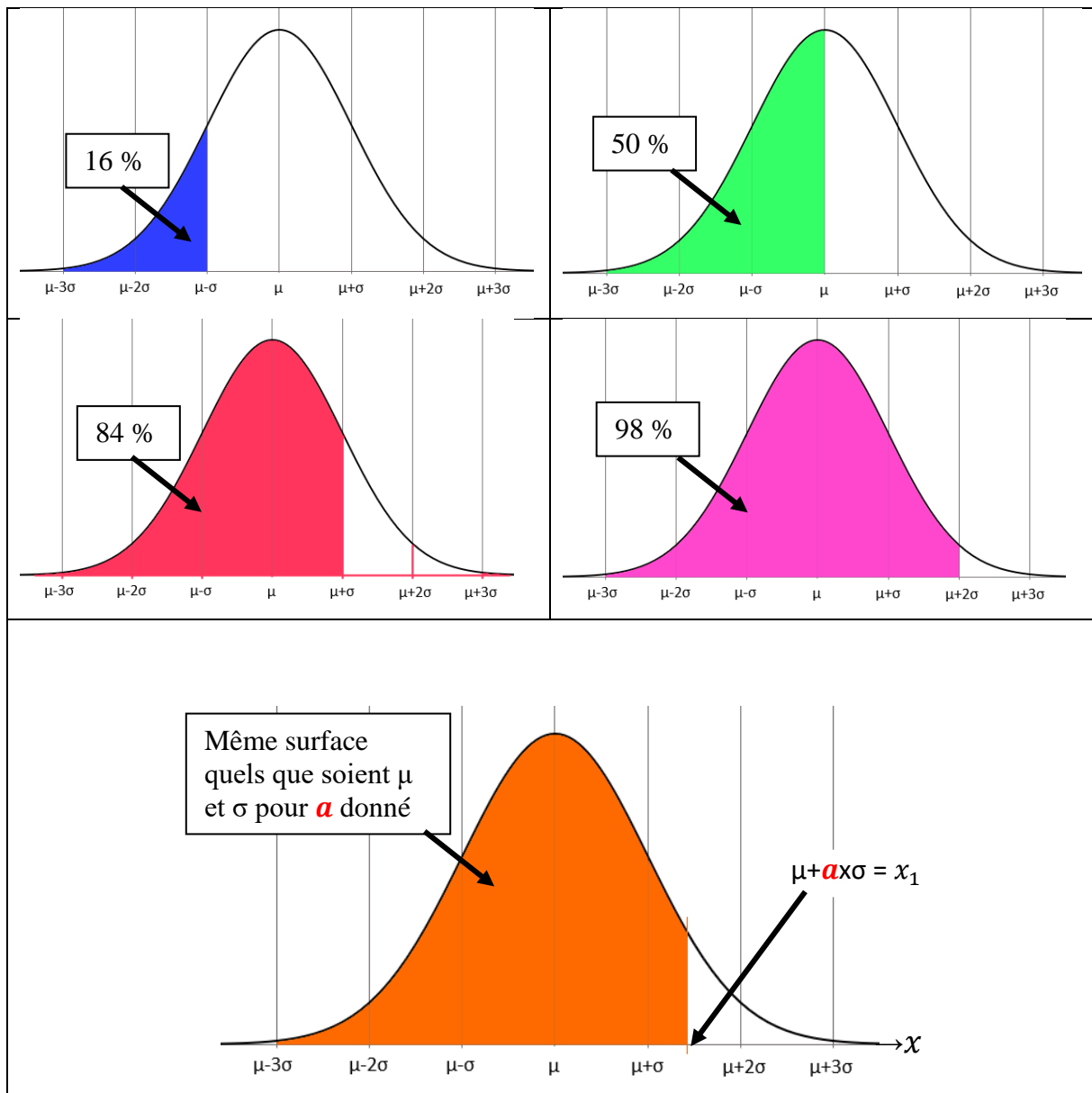


Figure 14 : Proportions de surface délimitée par une courbe normale

Cette dernière propriété est importante car finalement, le calcul de cette proportion de la surface totale va permettre de calculer la probabilité que la variable aléatoire X a d'être inférieure à la réalisation x_1 , ou encore de calculer $p(X < x_1)$. On peut également considérer cette surface comme la proportion des réalisations de la variable aléatoire X inférieures à x_1 quand le nombre de ces réalisations tend vers l'infini. Et d'après ce qui vient d'être dit, c'est le nombre d'écart-types dont la valeur x_1 est éloigné de l'espérance μ qui va permettre de mener à bien ces calculs. C'est ce que nous allons voir maintenant en nous ramenant à l'étude de la loi normale centrée réduite.

Score-z et loi normale centrée réduite

On peut montrer que si X suit une loi normale de paramètres μ et σ , alors la variable aléatoire définie par

$$Z = \frac{X - E(X)}{\sqrt{\text{var}(X)}} = \frac{X - \mu}{\sigma}$$

suit une loi normale de paramètres 0 et 1 : c'est la loi normale centrée réduite ou loi normale standard que l'on note $\mathcal{N}(0,1)$. Vous en trouverez une représentation graphique (figure 15) ci-dessous ; l'axe des abscisses est donc gradué en nombres d'écart-types.

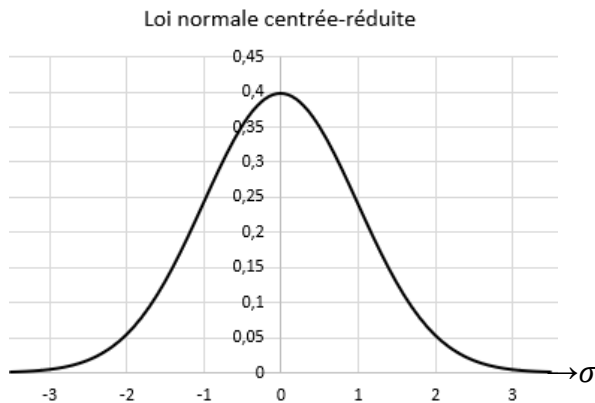


Figure 15 : loi normale centrée-réduite

Cette transformation s'applique aussi aux réalisations de ces variables aléatoires, et quand la variable aléatoire X prend la valeur x_1 alors la variable aléatoire Z précédemment définie prendra la valeur z_1 définie par

$$z_1 = \frac{x_1 - \mu}{\sigma}$$

Remarquons enfin que les deux égalités précédentes donnent

$$X = \mu + \sigma \times Z \quad \text{et} \quad x_1 = \mu + \sigma \times z_1$$

L'étude de la loi normale centrée réduite a permis d'établir des tables de scores-z (on dit que la loi a été tabulée) qui donnent la probabilité qu'une variable aléatoire centrée réduite a, par exemple, d'être inférieure^a à une valeur donnée. Une table est présentée en annexe 5, mais de nos jours ces nombres sont facilement calculés par les tableurs

Admettons que nous cherchions à calculer la proportion de la surface sous la courbe normale de paramètres μ et σ , l'axe des abscisses et à gauche de la droite d'équation $x = x_1$. Cela revient donc à calculer $p(X < x_1)$.

D'après les égalités précédentes, on a finalement

$$p(X < x_1) = p(\mu + \sigma \times Z < \mu + \sigma \times z_1) = p(Z < z_1)$$

Par lecture d'une table de score-z, on peut connaître $p(Z < z_1)$ et donc $p(X < x_1)$.

^a Ou supérieure, les propriétés des probabilités permettant facilement de passer de l'une à l'autre.

Pour fixer les idées, nous allons prendre comme exemple la loi normale d'espérance 488,3 et d'écart-type 95,1^a. Cette loi est notée $\mathcal{N}(488,3 ; 95,1)$. Si on cherche à calculer la probabilité que la variable aléatoire X soit inférieure ou égale à 600, cela revient donc à calculer la proportion de la surface totale comprise entre cette courbe normale, l'axe des abscisses et la gauche de la droite perpendiculaire à l'axe des abscisses passant par $x_1 = 600$. On va dans un premier temps transformer cette valeur en score-z :

$$z_1 = \frac{600 - 488,3}{95,1} \approx 1,17$$

On trouvera dans les tables que la probabilité que Z soit inférieure à 1,17 est égale à 0,879. Cette valeur sera donc la probabilité que X soit inférieure à 600 (figure 16).

$$p(Z < 1,17) = 0,879 = 87,9\% = p(X < 600)$$

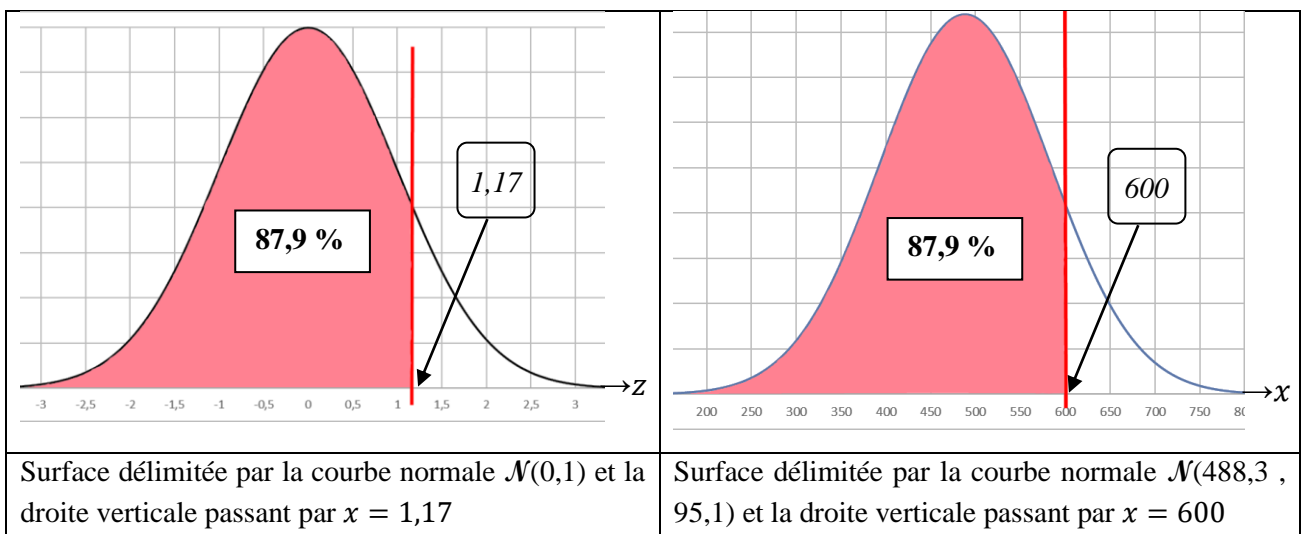


Figure 16 : utilisation des scores-z

On va terminer avec la présentation de notations très souvent utilisées dans les tests d'hypothèses qui utilisent les propriétés de la loi normale centrée réduite.

Dans ces tests on note très souvent α la probabilité que la variable aléatoire Z a d'être supérieure à un score-z donné, lui-même noté z_α . On a donc

$$p(Z > z_\alpha) = \alpha$$

Cette probabilité correspond à la surface comprise entre la courbe représentative de la loi normale centrée réduite, l'axe des abscisses et à droite d'un axe vertical passant par z_α . La représentation graphique de la loi normale est symétrique par rapport à l'axe des ordonnées, ce qui permet de faire les remarques présentées dans la figure 17 ci-dessous^b.

^a Ces valeurs ne sont pas choisies au hasard : il s'agit de la moyenne et de l'écart-type des scores PISA 2012 (voir chapitre 2 et chapitre 4).

^b Nous utiliserons ces remarques au chapitre 7.

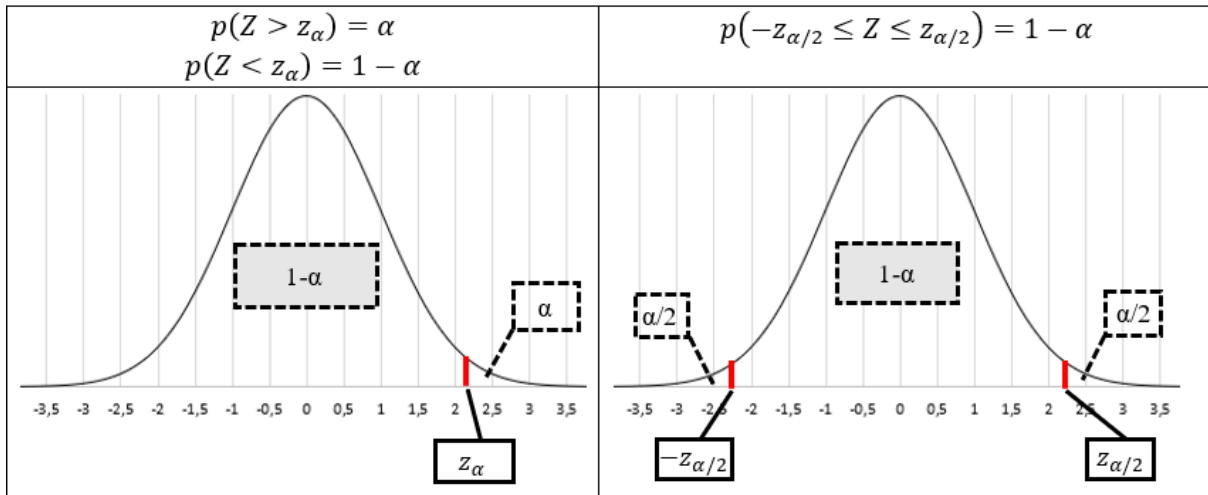


Figure 17 : propriétés de la loi normale centrée réduite

Les chiffres à retenir

On retient souvent les chiffres rassemblés dans le tableau 6 et qui donne les surfaces (qui correspondent à des probabilités comme nous l’avons vu) comprises entre la courbe normale, des droites verticales passant par des valeurs situées de part et d’autre de l’espérance et l’axe des abscisses (figure 18). Ce sont donc les proportions des réalisations supposées infinies d’une variable aléatoire comprises dans des intervalles centrés autour de l’espérance.

Espérance plus ou moins	1 écart-type	2 écarts-types	3 écarts-types
Probabilité (ou surface)	68% (soit 2/3)	95 %	100 %

Tableau 6 : surfaces (ou probabilités) pour certains intervalles de valeurs

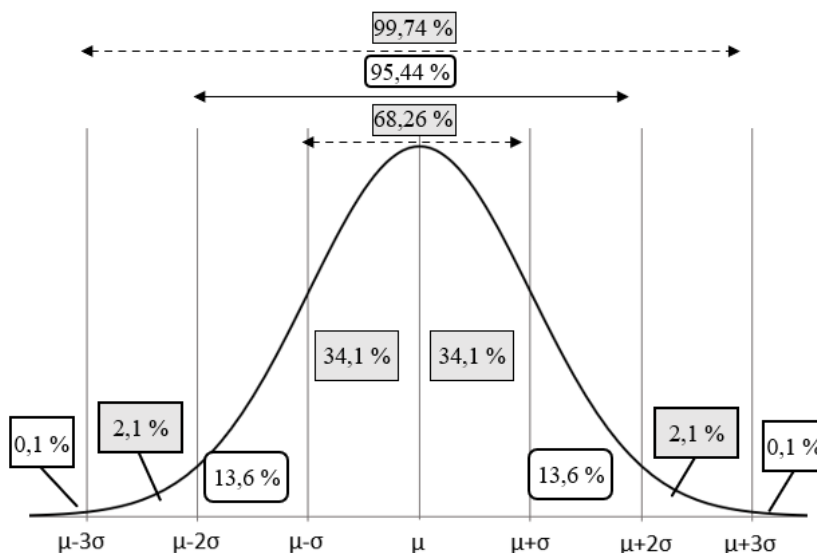


Figure 18 : proportions de la surface totale à retenir

Nous utiliserons également très souvent les deux remarques suivantes (figure 19) :

- 5 % des scores-z sont inférieurs à $-1,96$ et supérieurs à $1,96$, ou encore $p(Z < -1,96) = p(Z > 1,96) = 0,025$
- 5 % des scores-z sont supérieurs à $1,645$, ou encore $p(Z > 1,645) = 0,05$

En reprenant la notation vue ci-dessus, on a donc pour $\alpha = 0,05$, $z_{0,025} = 1,96$ et $z_{0,05} = 1,646$.

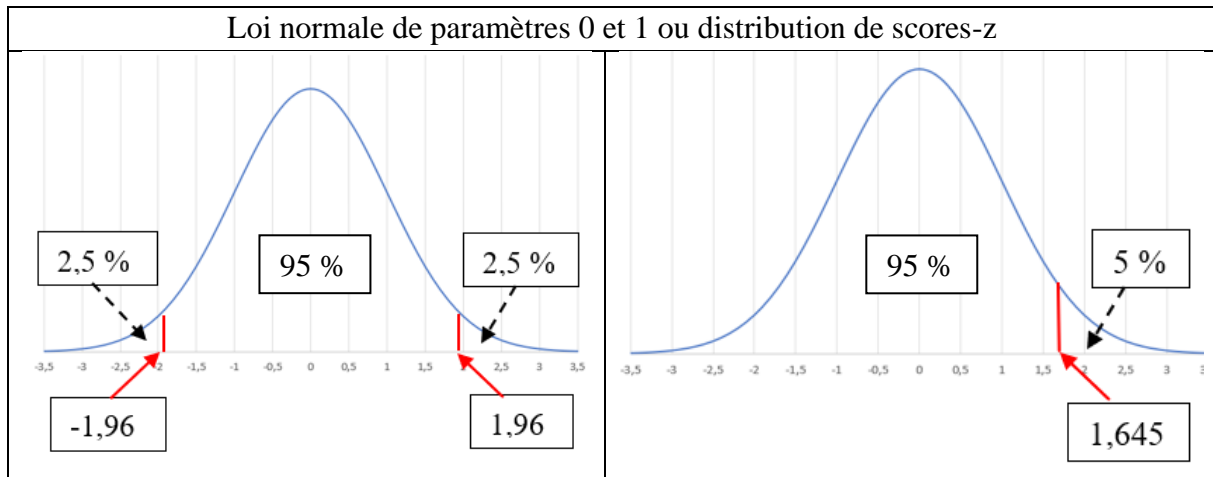


Figure 19 : valeurs particulières d'une loi normale centrée réduite

Quelques propriétés des variables aléatoires suivant des lois normales sont données dans l'annexe 5.

Chapitre 4. Distribution normale de caractéristiques naturelles

Pour comprendre pourquoi de très nombreuses caractéristiques naturelles suivent des distributions normales (ou presque), nous allons nous appuyer sur un théorème fondamental des probabilités : le théorème central limite.

Le théorème central limite

On peut montrer^a que si l'on considère n variables aléatoires indépendantes deux à deux^b qui toutes suivent une même loi de probabilité (et qui ont donc même espérance $E(X)$ et même variance $var(X)$), alors la somme de ces variables est une variable aléatoire d'espérance $n \times E(X)$, de variance $n \times var(X)$ et d'écart-type $\sqrt{n} \times \sqrt{var(X)}$.

Mais ce n'est pas tout. Un théorème fondamental a également été démontré, qui porte le nom de théorème central limite. Ses conséquences sont incommensurables dans le domaine des probabilités, et impose que nous le citions.

Théorème central limite

Dans les conditions énumérées ci-dessus et quand le nombre de variables aléatoires n tend vers l'infini, alors la somme de ces variables aléatoires suit une loi normale de paramètres $n \times E(X)$ et $\sqrt{n} \times \sqrt{var(X)}$

Remarquons tout de suite que les variables aléatoires peuvent ne pas suivre une loi normale^c. Le nombre infini de variables qui est une condition d'application de ce théorème peut apparaître comme une limite contraignante à son utilisation dans des situations concrètes. Mais il a été montré que l'on pourra considérer cette condition comme vérifiée dès que n est supérieur ou égale à 30. On dira alors que la somme des variables aléatoires tend à suivre une loi normale.

Reprenons l'exemple du chapitre précédent (**exemple 3**) qui consiste à prendre une des 2000 copies au hasard puis à noter son score. Dans un premier temps, nous allons poursuivre son étude théorique ; pour dans un second temps, vérifier expérimentalement les résultats obtenus. Il nous faut au moins 30 variables aléatoires qui ont la même loi de probabilité : on va utiliser 40 fois la même variable aléatoire « choisir une copie au hasard et noter son score »^d et noter ces variables X_i , ($1 \leq i \leq 40$).

On sait déjà que $E(X_i) = 3,35$ et $var(X_i) = 1,8275$.

Notons T la somme des 40 variables aléatoires X_i . On dira que T est une **statistique**^e associée aux variables aléatoires X_i . On a

^a Vous trouverez en annexe 5 les règles de calculs qui permettent de démontrer ces affirmations.

^b Deux variables aléatoires sont indépendantes quand les événements aléatoires qui les définissent n'ont aucune influence l'un sur l'autre.

^c Si cela était le cas, on peut montrer que leur somme suit de toutes façons une loi normale, y compris donc pour un petit nombre de variables

^d Une description plus fine de cette situation est proposée dans le chapitre suivant.

^e Une statistique (associée aux variables aléatoires X_i) est toute variable aléatoire calculée à partir des variables aléatoires X_i .

$$T = \sum_i X_i$$

Et d'après ce qui a été dit

$$E(T) = 40 \times 3,35 = 134, \text{ var}(T) = 40 \times 1,8275 = 73,1 \text{ et } \sqrt{\text{var}(T)} = \sqrt{73,1} \approx 8,55.$$

Les variables aléatoires X_i sont indépendantes deux à deux (puisqu'on remet à chaque fois la copie dans le paquet) et de plus 40 est supérieur à 30 : on peut donc appliquer le théorème central limite, et affirmer que la variable aléatoire T tend à suivre une loi normale de paramètres 134 et 8,55, donc $T \sim \mathcal{N}(134, 8,55)$.

Voilà pour la théorie.

Passons à l'expérience.

On commence par un premier essai (on prend une première copie au hasard, on note son score, on la remet dans le paquet, on prend une seconde copie au hasard, on note son score, on la remet dans le paquet, ... et ainsi de suite jusqu'à la 40^{ème} copie). On aura alors la première réalisation des 40 variables aléatoires. On calcule alors leur somme qui est la première réalisation de la variable aléatoire T . Et on recommence la même opération pour un second essai, puis un troisième, jusqu'à faire 500 essais. On obtient donc $40 \times 500 = 20\,000$ scores et 500 sommes. Les résultats expérimentaux ont été rassemblés dans le tableau 7 (et comparés aux résultats théoriques).

L'expérience	
Moyenne de la somme	134,762
Écart-type de la somme	8,352566

La théorie	
Moyenne de la somme	134
Écart-type de la somme	8,549854

Tableau 7 : somme de 40 tirages au sort

On trace enfin un histogramme (figure 20) représentant la distribution des sommes obtenues avec une largeur des classes de 2,5 sur lequel on superpose la courbe normale théorique. Notre expérience semble en accord avec la théorie, au moins par analyse visuelle rapide (il existe des méthodes plus précises pour déterminer dans quelle mesure on peut affirmer qu'une distribution de données suit une loi normale). Comme nous l'avons déjà signalé, Excel décale les rectangles des histogrammes sur la droite, de la moitié de la largeur de la classe ; ce décalage a été rectifié au niveau du calcul des valeurs prises par la loi normale.

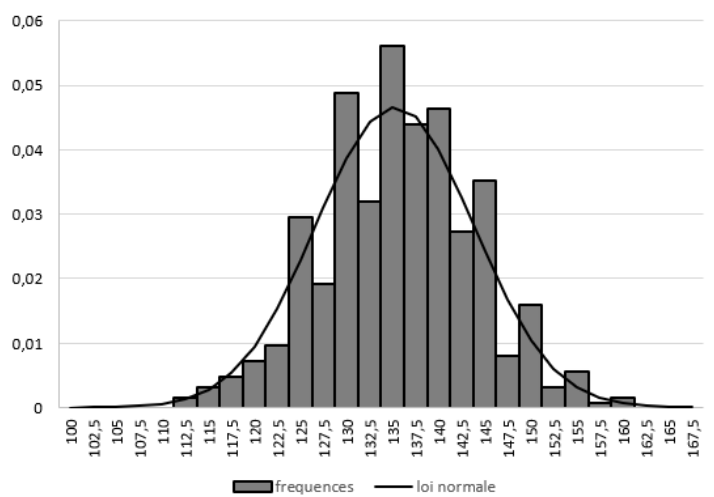


Figure 20 : distribution des sommes (exemple 3)

Des exemples de données expérimentales distribuées normalement

Reprenons maintenant les **exemples 1 et 2** vus au chapitre 2. On va poursuivre dans un premier temps la description de nos données en calculant la proportion de ces données comprises dans des intervalles centrés autour de la moyenne et dont la largeur est un nombre entier d'écart-types. Pour les poids de 252 hommes, cela donne le tableau 8 suivant :

	borne inférieure	borne supérieure	effectif	proportion %
moyenne + ou - 1 écart-type	67,9	94,5	182	72,2
moyenne + ou - 2 écarts-types	54,6	107,8	245	97,2
moyenne + ou - 3 écarts-types	41,2	121,0	251	99,6
Moyenne = 81,2 kg				
Écart-type = 13,3 kg				

Tableau 8 : proportion des poids dans des classes centrées autour de la moyenne

On va faire la même chose pour les scores obtenus par des élèves lors de l'enquête PISA 2012 dans le tableau 9 ci-dessous.

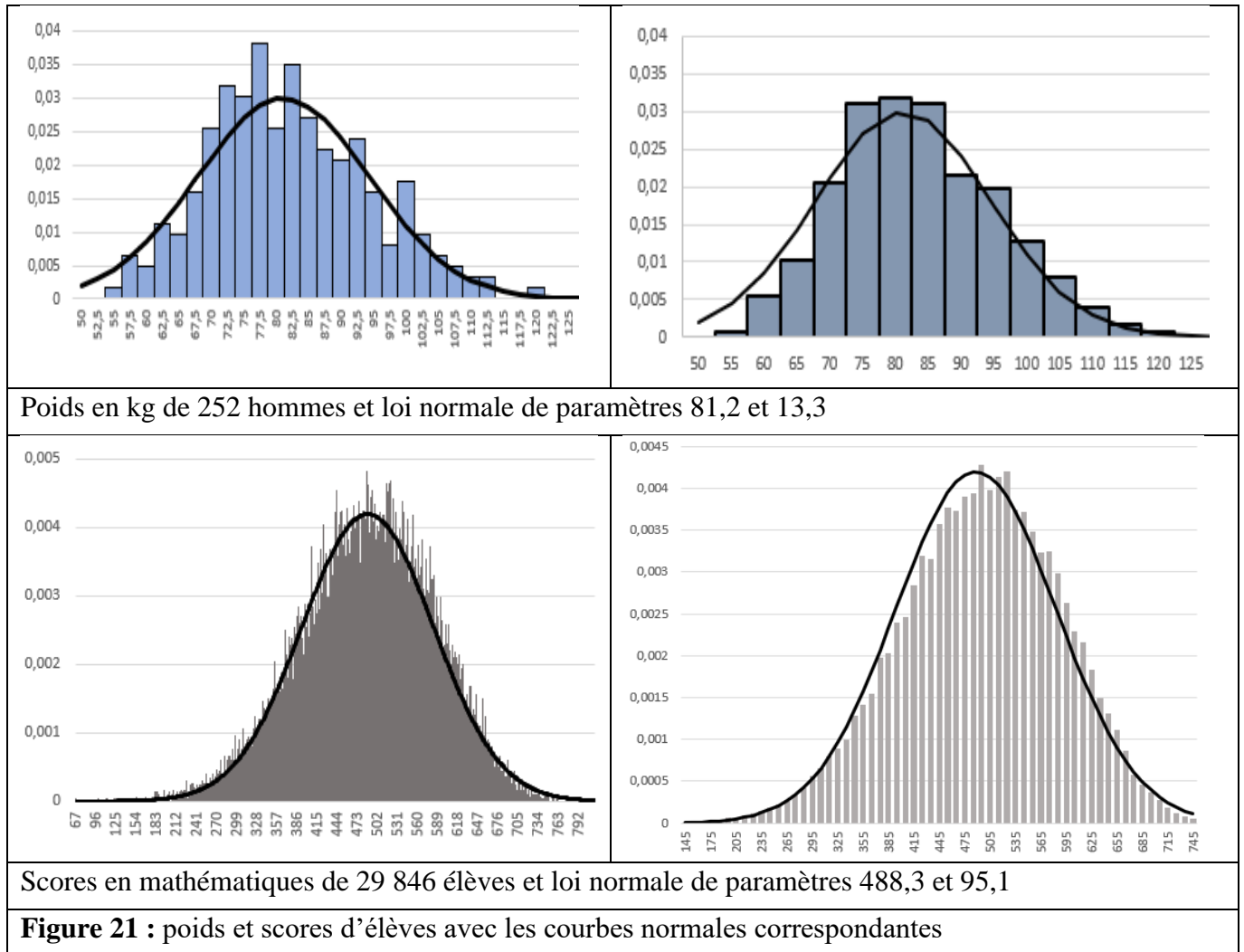
	borne inférieure	borne supérieure	effectif	proportion %
moyenne + ou - 1 écart-type	393,2	583,4	20139	67,5
moyenne + ou - 2 écarts-types	298,1	678,4	28581	95,8
moyenne + ou - 3 écarts-types	203,0	773,5	29767	99,7
Moyenne = 488,3				
Ecart-type = 95,1				

Tableau 9 : proportion des scores dans des classes centrées autour de la moyenne

Dans tous les cas, les proportions calculées à partir des données expérimentales ne sont pas très éloignées de celles calculées théoriquement dans le cas d'une distribution normale (voir le tableau 6).

Nous allons reprendre les quatre histogrammes de nos exemples (figure 8 et 9), en y ajoutant cette fois les courbes normales (de paramètres les moyennes et écarts-types obtenus

expérimentalement) correspondantes (figure 21) : la loi normale semble bien pouvoir servir de modèle à nos distributions statistiques, et on conclura ici que ces dernières tendent à suivre une distribution normale. Encore une fois, les courbes normales semblent légèrement décalées sur la gauche, cela étant dû à la présentation des histogrammes par Excel (ce sont en fait les histogrammes qui sont décalés d'une demi-classe sur la droite).



Revenons maintenant sur le calcul de la probabilité que la variable aléatoire X suivant la loi normale de paramètre 488,3 et 95,1 soit inférieure ou égale à 600 (chapitre 3). Ces paramètres sont en fait la moyenne et l'écart-type des 29 946 scores PISA 2012 de notre **exemple 2**. Nous avons trouvé que cette probabilité est égale à 0,879. Quand on utilise les données fournies par l'OCDE, on calcule une proportion des notes inférieures ou égales à 600 égale à 87,8 %, soit une valeur très proche de celle proposée par le modèle normal.

Finalement, que ce soit par simple analyse visuelle ou en vérifiant certaines de leurs propriétés, on peut considérer que les distributions des poids de 252 sujets masculins et des scores de 29 846 élèves peuvent être modélisées par les lois normales de paramètres la moyenne et

l'écart-type de ces deux populations. Mais comment l'expliquer ? C'est le théorème central limite qui va permettre de répondre à cette question.

Le hasard et les données

On peut en effet penser que certaines caractéristiques (comme le poids ou les compétences mathématiques) sont les résultantes de très nombreux facteurs indépendants distribués de façon aléatoire dans la population : les ressources alimentaires, les caractéristiques de l'écosystème environnant, le patrimoine génétique, le niveau socioéconomique, les caractéristiques culturelles et tant d'autres. Ces très nombreux facteurs en jeux sont les variables aléatoires indépendantes de notre modèle. On pourra cependant rétorquer que ces facteurs que l'on peut supposer être indépendants ne suivent pas la même distribution (et n'ont ni les mêmes moyennes ni les mêmes écarts-types). Mais tant que la proportion de chaque variance de chacune des variables reste modérée par rapport à la variance totale, le théorème central limite reste une explication plausible. Concrètement, cela revient à dire que les mesures des observations d'une caractéristique sont susceptibles de suivre une distribution normale quand tous les facteurs intervenant dans la genèse de cette caractéristique ont des rôles plus ou moins équivalents et qu'aucun ne prenne le pas sur les autres. Ce raisonnement s'applique également à des caractéristiques non naturelles comme des défauts de fabrication ou des erreurs de mesures. C'est donc bien le hasard qui finalement se glisse jusque dans la description de nos données.

Comme nous l'avons déjà remarqué, la loi normale n'est pas toujours le modèle le plus pertinent pour décrire une distribution de données. Il existe d'autres modèles, et parfois aucun d'entre eux ne peut être appliqué.

Jusqu'à présent, nous avons considéré la série de données comme une population entière (par exemple l'ensemble des 252 personnes dont on a mesuré le poids (**exemple 1**), les 29 846 élèves de 15 ans dont on a calculé le score en mathématiques (**exemple 2**) ou les 2000 copies d'élèves (**exemple 3**)). Nous allons maintenant nous intéresser à ce que l'on peut dire et faire quand ce sont les mesures récoltées sur un échantillon extrait d'une population qui sont disponibles. La population sera à partir de maintenant l'ensemble des individus qui intéressent le chercheur et qui sont la cible par exemple de ses préconisations.

Deuxième partie : les échantillons

Chapitre 5. Échantillonnage et intervalle de fluctuation

On va dans ce chapitre considérer une nouvelle situation, puisque nous allons prélever un ou plusieurs échantillons dans une population. Cet échantillon se doit d'être représentatif de la population dont il est issu, et c'est dans ce but qu'il est prélevé de façon aléatoire. On dit encore qu'on procède à un **échantillonnage**.

Imaginons par exemple que nous connaissions les données d'un ensemble de personnes ; il peut s'agir par exemple des poids de sujets masculins ou de scores d'élèves. Ces données constituent ce que nous allons nommer ici la **population**. On désigne leur moyenne par μ et leur écart-type par σ ; on dit également que μ et σ sont des **paramètres** de la population. On prélève maintenant au hasard dans cette population plusieurs échantillons, chacun constitués de n données. On va associer à ce prélèvement la variable aléatoire X : « prélever au hasard une donnée dans la population, en prendre note et replacer cette donnée dans la population ». Et on sait déjà que $E(X) = \mu$ et que $var(X) = \sigma^2$ (voir le chapitre 3). Considérons maintenant n variables aléatoires X_i qui toutes ont la même loi de probabilité que la variable aléatoire X . Ainsi X_1 sera définie comme « prélever au hasard une première donnée », X_2 « choisir au hasard une deuxième donnée », et ainsi de suite jusqu'à X_n « choisir au hasard une $n^{ième}$ donnée ». Quand on procède de la sorte pour le premier échantillon, on obtient finalement n données qui sont les n premières réalisations des n variables aléatoires X_i que l'on note $x_{1,i}$ (pour i compris entre 1 et n). Puis on continue avec un second échantillon, pour lequel on obtient les n données qui sont les n deuxièmes réalisations des n variables aléatoires X_i que l'on note $x_{2,i}$, puis un troisième, et ainsi de suite (voir tableau 10). La somme de ces n variables aléatoires est elle-même une variable aléatoire que l'on a déjà notée T au chapitre précédent. Et il en est encore de même pour la moyenne M de ces variables aléatoires, que l'on pourrait décrire comme « choisir au hasard un échantillon et calculer la moyenne des données ». M et T sont des statistiques associées aux variables aléatoires X_i . On a

$$T = \sum_i X_i \quad M = \frac{\sum_i X_i}{n}$$

La réalisation de la variable aléatoire T pour l'échantillon j seront notée t_j et la réalisation de la variable aléatoire M pour l'échantillon j seront notée m_j . Ainsi

$$t_1 = \sum_i x_{1,i} \quad m_1 = \frac{\sum_i x_{1,i}}{n}$$

	X_1	X_2	...	X_n	T	M
Échantillon n°1	$x_{1,1}$	$x_{1,2}$...	$x_{1,n}$	t_1	m_1
Échantillon n°2	$x_{2,1}$	$x_{2,2}$...	$x_{2,n}$	t_2	m_2
...
Échantillon n°j	$x_{j,1}$	$x_{j,2}$...	$x_{j,n}$	t_j	m_j
...

Tableau 10 : données prélevées dans plusieurs échantillons

Il est temps maintenant de faire un point sur l'emploi du terme « moyenne^a » qui peut finalement désigner trois résultats :

- la moyenne de la population, notée μ ;
- la moyenne des variables aléatoires X_i , variable aléatoire elle-même, notée M ;
- les moyennes des données de chaque échantillon, qui sont les réalisations de la variable aléatoire M , notées m_i .

L'usage de différents alphabets permet en général de savoir un peu mieux à quelle « moyenne » on a affaire : les lettres grecque sont en effet réservées aux paramètres des populations, les majuscules de l'alphabet latin aux variables aléatoires et les minuscules de l'alphabet latin aux réalisations de ces dernières. Vous comprenez maintenant pourquoi dans les chapitres précédents les moyennes et écarts-types des données (qui étaient alors la population considérée) étaient notées μ et σ .

Comme la donnée prélevée est replacée dans la population^b, les variables aléatoires X_i sont indépendantes deux à deux : le résultat obtenu après un tirage ne dépend pas des résultats obtenus aux tirages précédents. Pour des raisons pratiques évidentes, les échantillons prélevés ne suivent pas la plupart du temps cette règle, et les données extraites au hasard ne peuvent pas être extraites une seconde fois. Mais tant que la taille de l'échantillon reste très petite par rapport à la taille de la population, ce qui est très souvent le cas, on pourra considérer que les variables aléatoires X_i sont bien indépendantes deux à deux.

Étude théorique de la moyenne M

On peut montrer^c que si on considère n variables aléatoires indépendantes deux à deux, qui toutes suivent une même loi de probabilité d'espérance μ et d'écart-type σ , alors la moyenne M de ces variables est une variable aléatoire d'espérance $E(M) = \mu$ et d'écart-type $\sqrt{\text{var}(M)} = \frac{\sigma}{\sqrt{n}}$.

De plus, une nouvelle version du théorème central limite peut être énoncée :

Théorème central limite

Dans les conditions énumérées ci-dessus et quand le nombre de variables aléatoires n tend vers l'infini, alors la moyenne M de ces variables aléatoires suit une loi normale de paramètres μ et $\frac{\sigma}{\sqrt{n}}$. Ce qui revient à dire que la variable aléatoire standardisée Z définie ci-dessous suit une loi normale de paramètres 0 et 1 (Z est centrée-réduite).

$$Z = \frac{M - E(M)}{\sqrt{\text{var}(M)}} = \frac{M - \mu}{\frac{\sigma}{\sqrt{n}}}$$

^a mais aussi le terme « somme » ou « écart-type » comme on le verra plus loin.

^b On dit que l'échantillonnage est non exhaustif.

^c Vous trouverez en annexe 5 les règles de calculs qui permettent de démontrer ces affirmations.

Il s'agit en fait du même théorème vu auparavant et qui concernait la somme des n variables. On a vu alors que cette somme suit une loi normale d'espérance $n \times \mu$ et d'écart-type $\sqrt{n} \times \sigma$. La moyenne étant égale à la somme des n variables aléatoires indépendantes divisée par n , alors on conclue assez simplement que cette moyenne suit une distribution normale dont les paramètres sont obtenus à partir des paramètres de la somme en les divisant par n . On parle alors de la **distribution d'échantillonnage** de la moyenne M , en considérant ici la distribution de probabilité de cette variable aléatoire, c'est-à-dire la distribution de ses réalisations obtenues à partir d'expériences que l'on imagine avoir répétées un nombre infini de fois. Et là encore, l'infini commence à partir de 30. Dans ce cas, on peut affirmer, de manière équivalente, que

- la moyenne M tend à suivre une loi normale de paramètres μ et $\frac{\sigma}{\sqrt{n}}$,
- la distribution d'échantillonnage de la moyenne M suit une loi normale de paramètres μ et $\frac{\sigma}{\sqrt{n}}$,
- la moyenne centrée réduite $Z = \frac{M - \mu}{\sqrt{\text{var}(M)}} = \frac{M - \mu}{\frac{\sigma}{\sqrt{n}}}$ suit la loi normale centrée réduite.

On connaît bien les propriétés de cette loi et on dispose de tables (et bien sûr de tableurs) qui permettent de déterminer les valeurs $z_{\alpha/2}$ telles que $p(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$ (voir chapitre 3).

Nous allons souvent considérer le cas où $1 - \alpha = 0,95$, et donc $\alpha = 0,05$. Dans ce cas on a $z_{\alpha/2} = 1,96$ et donc $p(-1,96 \leq Z \leq 1,96) = 0,95$. Et finalement

$$p\left(-1,96 \leq \frac{M - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1,96\right) = 0,95$$

Cela revient à dire que la probabilité que la moyenne centrée réduite Z soit comprise entre $-1,96$ et $1,96$ est égale à $0,95$. Ou encore que 95% de ces réalisations (supposées en nombre infini) sont comprises entre $-1,96$ et $1,96$. On transforme alors cette expression pour aboutir à une dernière égalité :

$$p\left(\mu - 1,96 \times \frac{\sigma}{\sqrt{n}} \leq M \leq \mu + 1,96 \times \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

Intervalle de fluctuation

On peut donc affirmer que la proportion des réalisations de la variable aléatoire M comprises entre $\mu - 1,96 \times \frac{\sigma}{\sqrt{n}}$ et $\mu + 1,96 \times \frac{\sigma}{\sqrt{n}}$ est égale à 95% quand le nombre de ces réalisations tend vers l'infini ; ou encore que dans 95 cas sur 100, la moyenne calculée sur un échantillon aléatoire (qui est une réalisation de la variable aléatoire M) est incluse dans l'intervalle $\left[\mu - 1,96 \frac{\sigma}{\sqrt{n}}; \mu + 1,96 \frac{\sigma}{\sqrt{n}}\right] = \left[\mu - 1,96 \times \sqrt{\text{var}(M)}; \mu + 1,96 \times \sqrt{\text{var}(M)}\right]$. Cet intervalle est appelé **l'intervalle de fluctuation** de la moyenne M **au seuil (ou niveau) de confiance** $0,95$, ou encore au **niveau de signification** $0,5$. L'écart-type de la distribution d'échantillonnage

de cette moyenne, qui est égal à $\frac{\sigma}{\sqrt{n}}$ est également appelé **erreur standard**^a (*standard error* ou SE en anglais) ou bien erreur-type de la moyenne.

α	Seuil ou niveau de signification	Souvent $\alpha = 0,05$ ou $0,01$
$1 - \alpha$	Seuil ou niveau de confiance	Souvent $1 - \alpha = 0,95$ ou $0,99$

Le théorème central limite est fondamental car ses conditions d'application sont très larges. En particulier, quelle que soit la loi de probabilité suivie par les variables aléatoires et donc y compris quand elles ne suivent pas une loi normale, leur moyenne sera distribuée normalement^b. À condition toutefois que la taille des échantillons soit supérieure à 30. Vous verrez plus loin que cela sera une condition d'utilisation de certains tests d'hypothèse employés de nos jours.

Nous poursuivons avec **l'exemple 3** du chapitre précédent. On va à nouveau utiliser 40 variables aléatoires X_i qui sont toutes égales à la même variable aléatoire X « choisir une copie au hasard et noter son score ». On sait déjà que

$$E(X) = \mu = 3,35 \text{ et } var(X) = \sigma^2 = 1,8275.$$

Si on note M la moyenne des 40 variables aléatoires X_i . On a

$$E(M) = \mu = 3,35, \text{ } var(M) = \frac{\sigma^2}{n} = \frac{1,8275}{40} \approx 0,04569 \text{ et } \sqrt{var(M)} \approx 0,214$$

Les variables aléatoires X_i sont indépendantes deux à deux et 40 est supérieur à 30, on peut donc appliquer le théorème central limite et on peut affirmer que la variable aléatoire M tend à suivre une loi normale de paramètres 3,35 et 0,214. L'intervalle de fluctuation de la moyenne au seuil 0,95 est alors

$$\left[3,35 - 1,96 \times \sqrt{0,04569}; 3,35 + 1,96 \times \sqrt{0,04569} \right] = [2,931; 3,769]$$

Cela veut dire que la probabilité que la réalisation de la variable aléatoire M soit comprise entre 2,931 et 3,769 est égale à 0,95.

On reprend maintenant les données obtenues expérimentalement au chapitre précédent à partir de 500 échantillons. On obtient alors 500 moyennes, 2,5 % de ces moyennes sont inférieures à 2,93 et 2,5 % de ces moyennes sont supérieures à 3,8 (et donc 95% de ces moyennes sont comprises entre ces deux valeurs). Ces résultats sont représentés par le nuage de point ci-dessous (figure 22) :

^a L'erreur standard d'une statistique est l'écart-type de sa distribution d'échantillonnage.

^b Si les variables aléatoires suivent une loi normale, alors les conclusions précédentes s'appliquent quel que soit le nombre de variables aléatoires.

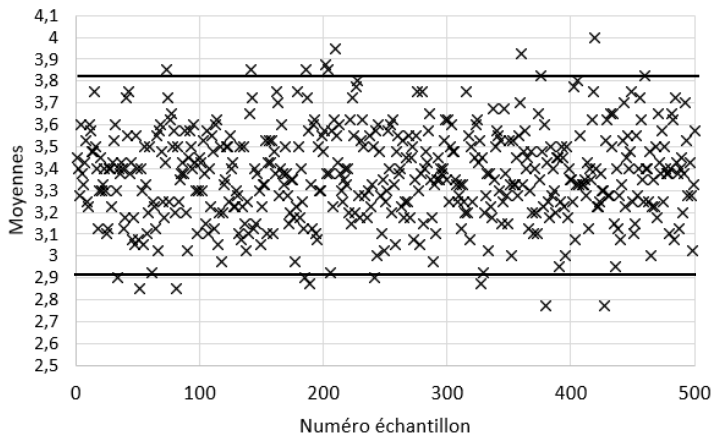


Figure 22 : 500 moyennes expérimentales (exemple 3)

Les moyennes et leurs écarts-types obtenus expérimentalement sont rassemblés dans le tableau 11 ci-dessous (et comparés à nouveau aux résultats théoriques).

L'expérience	
Moyenne	3,202
Écart-type de la moyenne (erreur standard)	0,20881

La théorie	
Moyenne	3,35
Écart-type de la moyenne (erreur standard)	0,21375

Tableau 11 : moyennes de 40 tirages au sort (exemple 3)

On trace enfin un histogramme représentant la distribution des moyennes obtenues avec une largeur des classes de 0,0625 sur lequel on superpose la courbe normale théorique. Une fois de plus, notre expérience semble en accord avec la théorie, au moins par analyse visuelle rapide. On obtient la représentation graphique suivante (figure 23) :

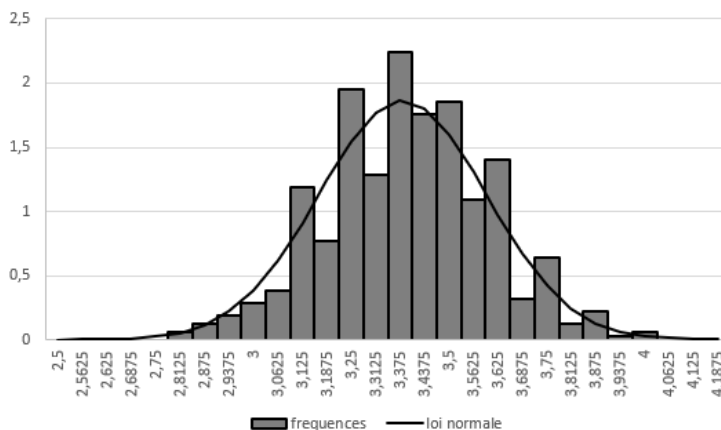


Figure 23 : distribution des moyennes (exemple 3)

Si la taille de l'échantillon est inférieure à 30, alors nous ne sommes plus dans les conditions d'application du théorème central limite. On devra vérifier une condition supplémentaire sur la variable aléatoire X (et donc sur l'ensemble des variables aléatoires X_i qui suivent la même loi de probabilité que la variable aléatoire X) : cette variable aléatoire devra en effet suivre une loi normale de paramètres μ et σ pour que nous puissions conclure que la variable M suit elle aussi

une loi normale (de paramètres μ et $\frac{\sigma}{\sqrt{n}}$) et enfin calculer des intervalles de fluctuation de la même façon que ci-dessus.

Intervalle de prédiction

Revenons aux scores de notre échantillon de taille n . On va admettre ici que les scores de la population dont ils sont issus, sont normalement distribués^a autour de leur moyenne μ avec comme écart-type σ . On a défini les scores de notre échantillon comme les premières réalisations de n variables aléatoires X_i qui toutes suivent une même loi de probabilité ; ici cette loi de probabilité est donc $\mathcal{N}(\mu, \sigma)$. On a donc la variable centrée réduite $Z = \frac{X_i - \mu}{\sigma}$ qui suit la loi normale centrée réduite et à nouveau, on peut trouver facilement les valeurs $z_{\alpha/2}$ telles que $p(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$. Là encore, on utilise souvent $\alpha = 0,05$ et dans ce cas, on a

$$p\left(-1,96 \leq \frac{X_i - \mu}{\sigma} \leq 1,96\right) = 0,95$$

Donc

$$p(\mu - 1,96\sigma \leq X_i \leq \mu + 1,96\sigma) = 0,95$$

Cela revient à dire que dans 95% des cas, le score x_i se trouve dans l'intervalle

$$[\mu - 1,96\sigma ; \mu + 1,96\sigma]$$

Et en particulier, le score x_{n+1} , première réalisation de la variable aléatoire X_{n+1} se trouvera dans 95% des cas dans cet intervalle, que l'on nomme **intervalle de prédiction** au seuil de confiance 0,95 (mais qui aura tout son intérêt au chapitre prochain)^b.

Nous allons maintenant appliquer ces remarques à une situation très courante en statistiques mais aussi dans d'autres domaines comme l'action politique ou même la vie courante : l'utilisation d'un seul et unique échantillon pour analyser une population. Tout le monde a un jour goûté un grain de raisin pour évaluer la qualité de la grappe dont il est issu. Tout le monde a également pris connaissance au moins une fois dans sa vie d'un sondage politique (avant une élection par exemple). L'utilisation des données d'un seul et unique échantillon dans le but d'estimer les paramètres de la population dont il est issu (et dont on ne connaît bien sûr pas les valeurs) est donc une pratique intuitive qui épargne de l'énergie et du temps (inutile de goûter toute la grappe de raisins ou d'interroger tous les électeurs).

Si le simple bon sens semble largement suffisant pour choisir une grappe de raisin, il en va différemment quand il s'agit de sondages ou d'études scientifiques qui vont conduire à des prises de décisions qui peuvent être lourdes de conséquences. Assurer une validité aux conclusions qui seront proposées impose de comprendre les processus mis en jeu pour décrire une population à partir des données d'un seul et unique échantillon. Nous allons quitter le domaine de la statistique descriptive pour pénétrer dans celui de la statistique inférentielle.

^a Leur distribution tend vers une loi normale ; des tests comme le test de Shapiro-Wilk (pour $n > 50$) ou le test de Kolmogorov-Smirnov pour des échantillons de taille plus grande, permettent de déterminer si cette condition est bien vérifiée.

^b Nous ne pouvons pas calculer un intervalle de prédiction pour les données de l'**exemple 3** car les données ne tendent pas à suivre une distribution normale.

Chapitre 6. Estimation des paramètres d'une population

Jusqu'à présent, on connaissait les paramètres μ et σ de la population. On va changer de point de vue maintenant, puisqu'on va se trouver dans la situation très classique où on ne connaît ni μ ni σ et où l'objectif va justement être d'estimer ces deux valeurs à partir d'un seul échantillon de taille n prélevé de façon aléatoire dans la population. On dit dans ce cas qu'on **infère**, c'est-à-dire qu'à partir de l'étude d'un unique échantillon on va conclure au niveau de la population en prenant le risque de commettre une erreur, risque qu'il va falloir expliciter.

On va reprendre le même raisonnement que celui suivi au chapitre précédent. On prélève aléatoirement une donnée de la population (par exemple un score ou un poids). On définit de ce fait une variable aléatoire X du type « prélever au hasard une donnée dans la population, en prendre note et replacer cette donnée dans la population ». On sait déjà que $E(X) = \mu$ et $var(X) = \sigma^2$ (même si ces deux valeurs sont inconnues maintenant, ces égalités restent vraies). Notre échantillon a une taille égale à n , on va donc considérer n variables aléatoires X_i qui suivent toutes la même loi que X . Cette fois, on ne prélève qu'un seul échantillon : les n mesures obtenues à partir de notre unique échantillon sont donc les premières (et uniques) réalisations des n variables aléatoires indépendantes X_i . A partir de ces variables aléatoires, on définit M , la moyenne des n variables X_i , et S' , l'écart-type des n variables X_i , de la manière suivante :

$$M = \frac{\sum_i X_i}{n} \quad S'^2 = \frac{\sum_i (X_i - M)^2}{n}$$

Dans un premier temps on calcule la moyenne m de ces n données (c'est l'unique réalisation de la variable aléatoire M) et l'écart-type s' (unique réalisation de la variable aléatoire S'). On a donc

$$m = \frac{\sum_i x_i}{n} \quad s'^2 = \frac{\sum_i (x_i - m)^2}{n}$$

Notre objectif est d'estimer les paramètres μ et σ de la population, et pour se faire nous utiliserons les propriétés de variables aléatoires qui sont des estimateurs des paramètres dont on cherche une estimation.

Estimateurs

L'**estimateur** d'un paramètre est une variable aléatoire qui doit satisfaire trois conditions :

- Son espérance doit être égale au paramètre. Dans ce cas, on dit que l'estimateur est sans biais. Si ce n'est pas le cas, on dit que l'estimateur est **biaisé**.
- Sa distribution tend à se concentrer autour du paramètre quand la taille d'échantillon augmente. Dans ce cas on dit que l'estimateur est convergent.
- Sa variance est la plus faible possible (si plusieurs estimateurs vérifient les propriétés précédentes, on choisira l'estimateur avec la variance la plus faible).

On sait déjà que $E(M) = \mu$ et $var(M) = \frac{\sigma^2}{n}$. Donc M est un estimateur de la moyenne de la population μ qui vérifie les propriétés citées ci-dessus^a.

On pourrait penser que S'^2 est un estimateur de la variance de la population σ^2 mais l'espérance de cet estimateur n'est pas égale à σ^2 : on peut montrer en effet que $E(S'^2) = \frac{n-1}{n} \times \sigma^2$, donc S'^2 est un estimateur biaisé de la variance de la population σ^2 . Et c'est pour cette raison que l'on choisira $S^2 = S'^2 \times \frac{n}{n-1}$ comme estimateur de σ^2 (on admettra ici qu'il vérifie les propriétés citées ci-dessus). On a donc^b

$$S = \sqrt{\frac{\sum(X_i - M)^2}{n}} \times \sqrt{\frac{n}{n-1}} = \sqrt{\frac{\sum(X_i - M)^2}{n-1}}$$

L'unique réalisation de cette variable aléatoire est donc

$$s = \sqrt{\frac{\sum(x_i - m)^2}{n-1}}$$

Estimations ponctuelles

Les réalisations des estimateurs sont des **estimations ponctuelles** du paramètre associé à la variable aléatoire X . Ce sont donc des valeurs approchées du paramètre de la population qui intéresse le chercheur (souvent la moyenne de la population μ , et son écart-type σ).


D'après ce que nous avons vu, l'estimation ponctuelle de la moyenne μ de la population est la moyenne m de l'échantillon qui est une réalisation de la variable aléatoire M , avec

$$m = \frac{\sum x_i}{n}$$

De la même façon, la meilleure estimation ponctuelle de l'écart-type de la population que l'on notera s , est la réalisation de la variable aléatoire S définie ci-dessus, avec

$$s = \sqrt{\frac{\sum(x_i - m)^2}{n-1}}$$

Par rapport à la formule qui était de mise quand on avait les données pour la population entière, on a donc remplacé μ par son estimation m et on a retranché 1 à la taille de l'échantillon (voir chapitre 1).

Si nous reprenons les données de nos **exemples 1 et 2** concernant les poids de 252 sujets masculins et les scores de 29 846 élèves de 15 ans habitant l'OCDE, on obtient les résultats suivants (tableau 12) 

^a D'autres estimateurs auraient pu prétendre à cette place, comme la médiane, mais qui ne vérifient pas ces trois propriétés.

^b On montre également que si n variables aléatoires X_i suivent toutes la même loi normale $\mathcal{N}(\mu, \sigma)$ et si $n > 30$, alors $var(S^2) = \sigma^4 \times \frac{2}{n-1}$ (LENOIR, 2008, p. 49, $S^2 \sim \mathcal{N}(\sigma^2, \sigma^2 \sqrt{\frac{2}{n-1}})$)

	Les données sont celles d'une population	Les données sont celles d'un échantillon représentatif d'une population
252 poids	$\sigma = 13,3041531$	$s = 13,330629$
29 846 scores (enquête PISA)	$\sigma = 95,0883273$	$s = 95,0867343$

Tableau 12 : écarts-types d'un échantillon, d'une population

Pour de très grandes tailles d'échantillons, l'utilisation de n ou $n - 1$ n'aura que peu d'impact sur les calculs.

Bilan pour la moyenne

Il faut donc bien faire la différence entre

- la moyenne de la population μ , qui est un paramètre de cette population
- la moyenne M , variable aléatoire, qui est un estimateur associé à un échantillonnage du paramètre μ
- et la moyenne d'un échantillon qui est la réalisation de cette variable aléatoire M , que l'on notera m dans la suite de ce livre et qui est une estimation ponctuelle du paramètre μ . Cette moyenne m est égale à la moyenne des mesures prise sur un échantillon. La plupart du temps, ce sera le seul nombre dont nous aurons connaissance.

Les estimations ponctuelles ne rendent pas compte des fluctuations d'échantillonnage. Pour se faire, nous allons leur associer des intervalles de confiance et affiner de cette manière notre estimation des paramètres de la population. Nous nous intéresserons ici exclusivement à la moyenne, mais le raisonnement que nous allons suivre pourrait s'appliquer à d'autres paramètres, comme par exemple l'écart-type.

Estimation par intervalle de la moyenne

On va devoir distinguer deux cas ici, en fonction de n , la taille de l'échantillon.

Si $n \geq 30$

D'après le chapitre précédent, nous savons déjà que M , moyenne des n variables aléatoires dont on connaît une réalisation m , est une variable aléatoire qui suit une loi normale de moyenne μ (la moyenne de la population, inconnue bien sûr) et d'écart-type $\frac{\sigma}{\sqrt{n}}$ (inconnu également), et ce quelle que soit la loi de probabilité suivie par la variable aléatoire X . Donc la variable centrée réduite Z définie ci-dessous suit une loi normale centrée réduite.

$$Z = \frac{M - E(M)}{\sqrt{\text{var}(M)}} = \frac{M - \mu}{\frac{\sigma}{\sqrt{n}}}$$

On retrouve alors la situation nous ayant permis de calculer un intervalle de fluctuation et on a

$$p\left(\mu - 1,96 \times \frac{\sigma}{\sqrt{n}} \leq M \leq \mu + 1,96 \times \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

Contrairement au chapitre précédent, ici on ne connaît pas σ^a . Mais **comme $n \geq 30$, on peut remplacer σ par son estimation s** (avec la méthode de calcul vue ci-dessus)^b.

$$p\left(\mu - 1,96 \times \frac{s}{\sqrt{n}} \leq M \leq \mu + 1,96 \times \frac{s}{\sqrt{n}}\right) = 0,95$$

Ou encore

$$p\left(M - 1,96 \times \frac{s}{\sqrt{n}} \leq \mu \text{ et } M + 1,96 \times \frac{s}{\sqrt{n}} \geq \mu\right) = 0,95$$

Nous avons noté m la moyenne de l'échantillon, unique réalisation de la variable aléatoire M . On peut donc affirmer que dans 95% des cas, quand on calcule une moyenne à partir des données d'un échantillon, on aura :

$$m - 1,96 \frac{s}{\sqrt{n}} \leq \mu \leq m + 1,96 \frac{s}{\sqrt{n}}$$

C'est cette double inégalité qui va nous permettre de définir un **intervalle de confiance** pour la moyenne μ au seuil (ou niveau) de confiance 0,95.

On peut remarquer que plus la taille de l'échantillon n est grande et plus l'intervalle est petit. De la même façon, plus l'écart-type de l'échantillon (donc la dispersion des données autour de la moyenne) est faible et plus l'intervalle est réduit.

Bilan (si $n \geq 30$)

Soit m la moyenne calculée d'un échantillon prélevé au hasard dans une population dont on ne connaît pas les paramètres, et s son écart-type ; alors on dira que l'intervalle $\left[m - 1,96 \frac{s}{\sqrt{n}}; m + 1,96 \frac{s}{\sqrt{n}}\right]$ contient la moyenne μ de la population avec un niveau de confiance^c de 0,95 : on appelle cet intervalle l'intervalle de confiance à 95% de la moyenne μ . Et si m est une estimation ponctuelle de la moyenne μ , l'intervalle de confiance $\left[m - 1,96 \frac{s}{\sqrt{n}}; m + 1,96 \frac{s}{\sqrt{n}}\right]$ est une estimation par intervalle de cette même moyenne μ .

Finalement, on peut retenir que la moyenne et l'écart-type d'un échantillon prélevé aléatoirement dans une population nous permettent donc de définir les bornes d'un intervalle qui contient la moyenne de la population dans 95 %^d des cas où un tel prélèvement est effectué.

Si nous considérons à nouveau notre **exemple 1** et en considérant maintenant que les poids des 252 sujets masculins sont un échantillon des poids de la population, on obtient l'intervalle de confiance à 95% suivant :

$$\left[81,2 - 1,96 \times \frac{13,3}{\sqrt{252}}; 81,2 + 1,96 \times \frac{13,3}{\sqrt{252}}\right] = [79,6; 82,8]$$

^a Certains auteurs traitent le cas où σ est connu (et dans ce cas on utilise bien entendu la valeur de ce paramètre dans le calcul de l'intervalle de confiance ; mais si μ est inconnu, il est bien peu probable de se trouver dans cette situation).

^b Certains préfèrent dans ce cas utiliser la loi de Student (voir plus loin) et réserve la procédure qui suit aux cas où σ est connu.

^c Attention, il ne s'agit pas d'une probabilité. En particulier, il est faux de dire que « la probabilité que μ appartienne à cet intervalle est égale à 0,95 » (μ n'est pas une variable aléatoire).

^d Les mêmes raisonnements s'appliquent pour des valeurs différentes de 95%

Cela signifie que la moyenne de la population dont nos 252 données sont un échantillon est contenue dans cet intervalle avec un niveau de confiance de 0,95.

Pour les scores PISA de l'**exemple 2** et en suivant le même raisonnement, cela donnerait l'intervalle de confiance à 95% suivant :

$$\left[488,3 - 1,96 \times \frac{95,1}{\sqrt{29\ 846}}; 488,3 + 1,96 \times \frac{95,1}{\sqrt{29\ 846}} \right] = [487,2 ; 489,4]$$

Pour terminer, nous reprenons notre **exemple 3** de 2 000 scores d'élèves. On a pris au hasard 40 copies, et on a obtenu les scores rassemblés dans le tableau 13. On calcule alors leur moyenne et leur écart-type (ici, c'est l'écart-type d'un échantillon qui est calculé ; si on considérait cette série de données comme une population, on aurait calculer un écart-type égal à 1,4824).

3	3	4	3	3	1	3	2	1	5	1	2	4	5	moyenne	3,45
2	5	5	2	5	5	2	5	3	1	2	4	5	5	écart-type	1,5013
5	5	3	5	2	1	5	5	5	2	5	4				

Tableau 13 : 40 scores tirés au sort (exemple 3)

On obtient donc l'intervalle de confiance à 95 % suivant :

$$\left[3,45 - 1,96 \times \frac{1,5013}{\sqrt{40}}; 3,45 + 1,96 \times \frac{1,5013}{\sqrt{40}} \right] = [2,985 ; 3,915]$$

On peut représenter ces résultats par des graphiques avec barres d'erreur (figure 24) .

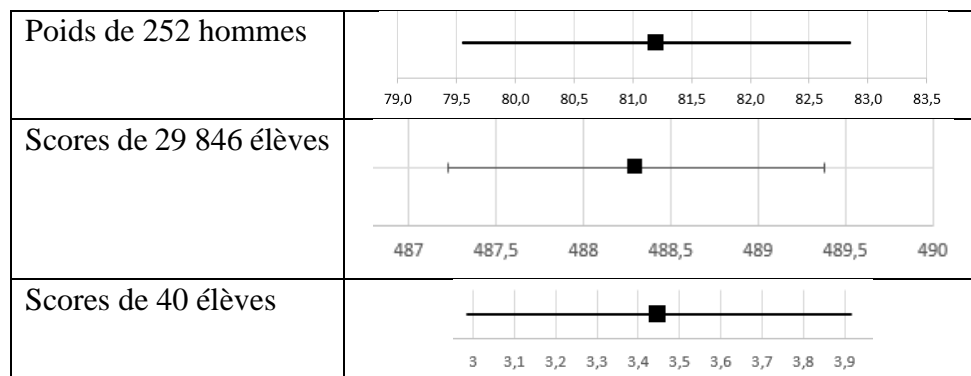


Figure 24 : intervalles de confiance à 95%

Si $n < 30$

Dans ce cas, on ne peut plus estimer l'écart-type de la population par une constante comme nous l'avons fait quand $n \geq 30$, et les écarts-types des échantillons doivent être considérés comme les réalisations de la variable aléatoire S définie auparavant. Pour calculer un intervalle de confiance nous devons renforcer les conditions sur la variable aléatoire X , et pour conclure

il faudra ici que sa loi de probabilité soit normale (de paramètres μ et σ). Ce sera donc la loi suivie par les variables aléatoires X_i .

On va considérer une nouvelle variable aléatoire, la variable aléatoire T^a définie par

$$T = \frac{M - \mu}{\frac{S}{\sqrt{n}}}$$

La modification apportée ici par rapport à la variable Z est l'intervention de la variable aléatoire S en lieu et place du paramètre σ . Cette variable T ne suit donc pas une loi normale centrée réduite mais la loi de Student à $n - 1$ degré de liberté. On notera $T \sim \mathcal{J}(n-1)$

Cette loi, étudiée par Student et qui porte son nom (en fait un pseudonyme), est plus complexe à définir que la loi normale, et nous nous contenterons ici de décrire les résultats nécessaires à sa mise en œuvre. La loi de Student fait intervenir le degré de liberté^b de l'échantillon qui est égal à $n - 1$. Pour chaque valeur de $n - 1$ on obtient une représentation graphique de la loi de Student qui ressemble à celle d'une loi normale centrée réduite mais plus aplatie, et ce d'autant plus que les valeurs de $n - 1$ sont faibles (voir figure 25) \square . On dit aussi que la loi de Student tend vers la loi normale centrée réduite quand les tailles des échantillons tendent vers l'infini.

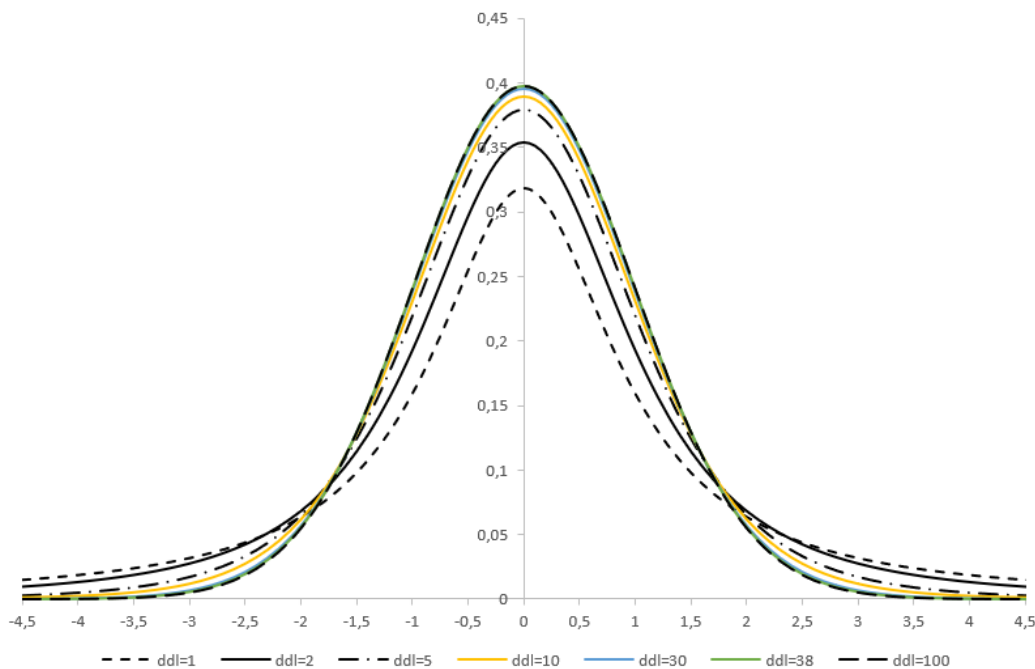


Figure 25 : lois de Student pour 7 degrés de liberté (ddl) (3 courbes superposées)

De la même façon que pour la loi normale centrée réduite, on dispose de tables^c (et bien sûr de tableurs) qui permettent de déterminer les valeurs bornant des intervalles dans lesquels on a une probabilité, définie à l'avance, de trouver les réalisations de la variable aléatoire T et ces valeurs

^a Attention, ce n'est pas la même variable que la variable somme utilisée au chapitre précédent.

^b On utilise dans l'estimation de la variabilité d'une population l'estimation de sa moyenne, on réduit donc de 1 le nombre d'informations indépendantes utilisées pour l'estimation de cette variabilité (voir le calcul de la variance).

^c Un exemple de tables est présenté en annexe 4.

sont notées $\pm t_{\alpha/2;n-1}$ (voir figure 26). Ces notations sont à rapprocher de celles utilisées pour la loi normale centrée réduite (chapitre 3), mais ici le degré de liberté de l'échantillon joue un rôle. Ainsi

$$p(-t_{\alpha/2;n-1} \leq T \leq t_{\alpha/2;n-1}) = 1 - \alpha$$

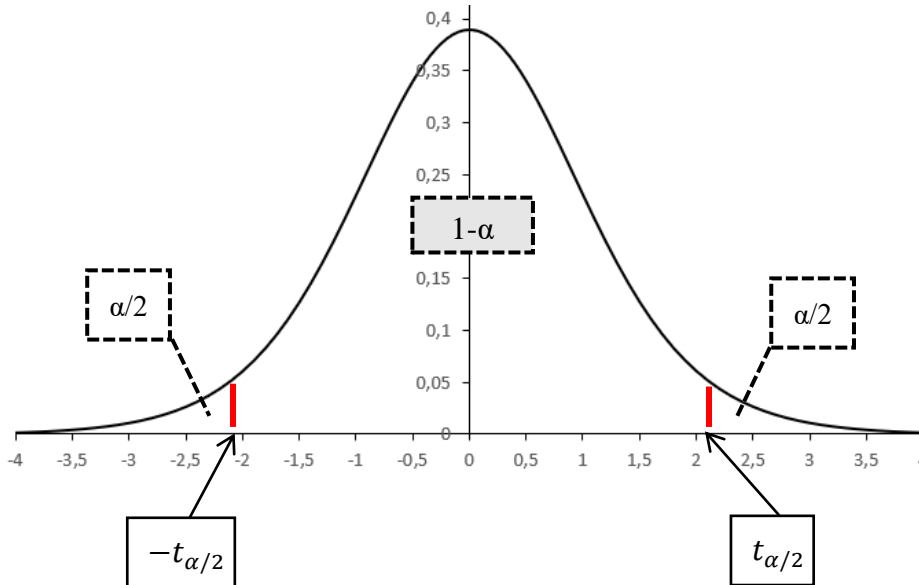


Figure 26 : propriétés des lois de Student (ici le degré de liberté est égal à 5).

Nous avons déjà souvent considéré le cas où $\alpha = 0,05$, et donc $1 - \alpha = 0,95$.

Vous verrez dans les tables de Student que, pour $n - 1 = 60$, on a $t_{0,025;60} = 2,0003$; et pour $n - 1 = 110$ on a $t_{0,025;110} = 1,982$. Plus la taille de l'échantillon est grande, et plus cette valeur se rapproche de 1,96. Donc finalement,

$$p\left(-t_{0,025;n-1} \leq \frac{M - \mu}{\frac{S}{\sqrt{n}}} \leq t_{0,025;n-1}\right) = 0,95$$

$$p\left(M - t_{0,025;n-1} \times \frac{S}{\sqrt{n}} \leq \mu \text{ et } M + t_{0,025;n-1} \times \frac{S}{\sqrt{n}} \geq \mu\right) = 0,95$$

A nouveau, avec m la moyenne de l'échantillon, unique réalisation de la variable aléatoire M et s l'écart-type de l'échantillon, unique réalisation de la variable aléatoire S , on peut affirmer que dans 95% des cas, quand on calcule une moyenne à partir des données d'un échantillon, on aura :

$$m - t_{0,025;n-1} \times \frac{s}{\sqrt{n}} \leq \mu \leq m + t_{0,025;n-1} \times \frac{s}{\sqrt{n}}$$

Et finalement on définit un intervalle de confiance de la moyenne μ au seuil de confiance de 0,95 de la manière suivante :

$$\left[m - t_{0,025;n-1} \times \frac{s}{\sqrt{n}} ; m + t_{0,025;n-1} \times \frac{s}{\sqrt{n}} \right]$$

Bilan (si $n < 30$)

Si $X_i \sim \mathcal{N}(\mu, \sigma)$, soit m la moyenne calculée d'un échantillon prélevé au hasard dans une population dont on ne connaît pas les paramètres, et s son écart-type, alors on dira que l'intervalle $\left[m - t_{0,025; n-1} \times \frac{s}{\sqrt{n}}; m + t_{0,025; n-1} \times \frac{s}{\sqrt{n}} \right]$ contient la moyenne μ de la population avec un niveau de confiance de 0,95 : on appelle cet intervalle l'intervalle de confiance à 95% de la moyenne μ .

Les différences entre les deux procédures (utilisation de la variable aléatoire ou statistique Z qui suit une loi normale centrée réduite ou utilisation de la variable aléatoire ou statistique T qui suit une loi de Student à $n - 1$ degrés de liberté) tient évidemment à leurs conditions d'application. Pour de très grands échantillons, les deux méthodes (il est toujours possible de calculer un intervalle de confiance en utilisant la statistique T d'utilisation plus large que la statistique Z) donnent des résultats similaires. Un bilan vous est proposé en annexe 6.

Les bornes inférieures et supérieures des intervalles de confiance ont été calculées au millième près en utilisant les deux procédures pour nos trois **exemples 1, 2 et 3** (les poids de 252 hommes, les scores de 29 846 élèves de l'OCDE et les 40 copies de notre exemple fictif, tableau 14). On ne remarque aucune différence observable dans les calculs concernant les scores PISA. Et au dixième près, les résultats concernant les deux autres exemples sont égaux.

	n	m	Écart-type	Z		T	
				Borne inférieure	Borne supérieure	Borne inférieure	Borne supérieure
Poids hommes	252	81,2	13,3	79,558	82,842	79,550	82,850
Scores PISA	29846	488,3	95,1	487,221	489,379	487,221	489,379
Scores élèves	40	3,45	1,5013	2,985	3,915	2,970	3,930

Tableau 14 : intervalles de confiance : loi normale ou loi de Student ?

Terminons avec un exemple numérique fictif (**exemple 4**) qui montre au contraire les différences observées dans ces deux cas quand l'effectif est faible. Supposons que nous ayons tiré au sort les 12 scores suivants (toujours des notes entières entre 1 et 5) : 1 ; 2 ; 2 ; 3 ; 3 ; 3 ; 3 ; 4 ; 4 ; 4 ; 5 ; 5. On représente graphiquement (figure 27) la distribution des scores pour vérifier visuellement qu'elle tend à suivre une loi normale.

On calcule alors une moyenne égale à 3,25 et un écart-type égal à environ 1,2154. Les bornes inférieures et supérieures des intervalles de confiance au niveau de confiance 0,95 calculés avec les statistiques Z et T sont données ci-dessous, et montre que dans ce cas, avec un effectif faible, les intervalles de confiance obtenus sont différents. On a $t_{0,025;11} = 2,20098516$

	Borne inférieure	Borne supérieure
Intervalle de confiance avec Z	2,56	3,94
Intervalle de confiance avec T	2,48	4,02

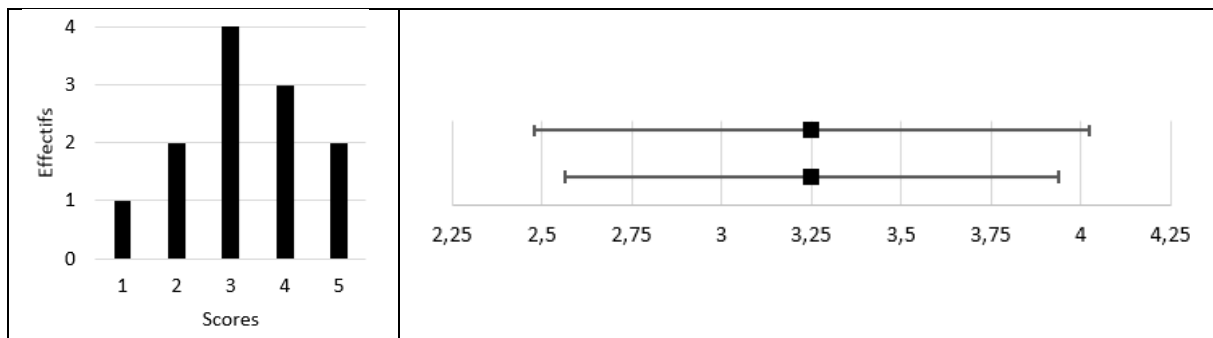


Figure 27 : distribution des scores et intervalles de confiance (statistique Z en bas, statistique T en haut).

Intervalle de prédiction

Nous allons reprendre les mêmes raisonnements qu'au chapitre précédent. On va donc admettre ici que les données de la population dont les mesures effectuées sur les échantillons sont issues, sont normalement distribuées^a autour de leur moyenne μ avec comme écart-type σ . Les données de notre échantillon sont donc les premières réalisations de n variables aléatoires X_i qui toutes suivent la loi $\mathcal{N}(\mu, \sigma)$. Mais contrairement au chapitre précédent, on ne connaît ici ni la moyenne μ , ni l'écart-type σ , mais seulement leur estimations m et s (moyenne et écart-type calculés à partir de l'échantillon). On va considérer la variable aléatoire X_{n+1} (« prendre au hasard une $n + 1^{\text{ème}}$ donnée dans la population »). Dans ce cas, il a été montré (et nous allons admettre ici), que la variable

$$\frac{X_{n+1} - M}{\sqrt{S^2 + \frac{S^2}{n}}}$$

suit une loi de Student à $n - 1$ degré de liberté. On connaît bien entendu les valeurs $t_{\alpha/2; n-1}$ telles que

$$p\left(-t_{\alpha/2; n-1} \leq \frac{X_{n+1} - M}{\sqrt{S^2 + \frac{S^2}{n}}} \leq t_{\alpha/2; n-1}\right) = 1 - \alpha$$

Pour $\alpha = 0,05$, on a

$$p\left(-t_{0,025; n-1} \leq \frac{X_{n+1} - M}{\sqrt{S^2 + \frac{S^2}{n}}} \leq t_{0,025; n-1}\right) = 0,95$$

$$p\left(M - t_{0,025; n-1} \times \sqrt{S^2 + \frac{S^2}{n}} \leq X_{n+1} \leq M + t_{0,025; n-1} \times \sqrt{S^2 + \frac{S^2}{n}}\right) = 0,95$$

^a Leur distribution tend vers une loi normale

Avec m la moyenne de l'échantillon, unique réalisation de la variable aléatoire M et s l'écart-type de l'échantillon, unique réalisation de la variable aléatoire S , on peut affirmer que dans 95% des cas, quand on calcule une moyenne à partir des données d'un échantillon, on aura

$$m - t_{0,025; n-1} \times \sqrt{s^2 + \frac{s^2}{n}} \leq x_{n+1} \leq m + t_{0,025; n-1} \times \sqrt{s^2 + \frac{s^2}{n}}$$

Et finalement on définit l'intervalle de prédiction de la manière suivante :

$$\left[m - t_{0,025; n-1} \times \sqrt{s^2 + \frac{s^2}{n}} ; m + t_{0,025; n-1} \times \sqrt{s^2 + \frac{s^2}{n}} \right]$$

Cet intervalle construit à partir d'un échantillon de taille n contient dans 95 cas sur 100 la $n + 1^{\text{ème}}$ donnée tirée au hasard dans une population dont les données suivent une loi normale de paramètres μ et σ dont m et s (moyenne et écart-type de l'échantillon) sont des estimations ponctuelles.

Forts de ces raisonnements qui nous ont permis d'estimer les paramètres d'une population à partir des résultats obtenus sur un échantillon, nous allons pouvoir maintenant passer à la comparaison de deux échantillons de données, échantillons obtenus après un essai contrôlé randomisé.

Chapitre 7. Différence entre deux moyennes statistiquement significative

Quelle est la question ?

On va considérer l'expérience suivante. Deux groupes expérimentaux d'élèves sont constitués de façon aléatoire en les prélevant dans la population^a. L'un des échantillons est soumis à un traitement particulier, c'est-à-dire une méthode d'enseignement dont on attend des effets positifs ; c'est le groupe traitement. L'autre échantillon n'est pas soumis à ce traitement, mais à une méthode d'enseignement dite « standard » ; c'est le groupe contrôle. Pour évaluer l'impact de ces deux méthodes d'enseignement^b sur les apprentissages des élèves, ces derniers passent le même test en fin de traitement (voir figure 28 ci-dessous). Dans chaque échantillon, les élèves ont donc obtenu un score et on connaît bien entendu les moyennes et écart-types de ces deux séries de données.

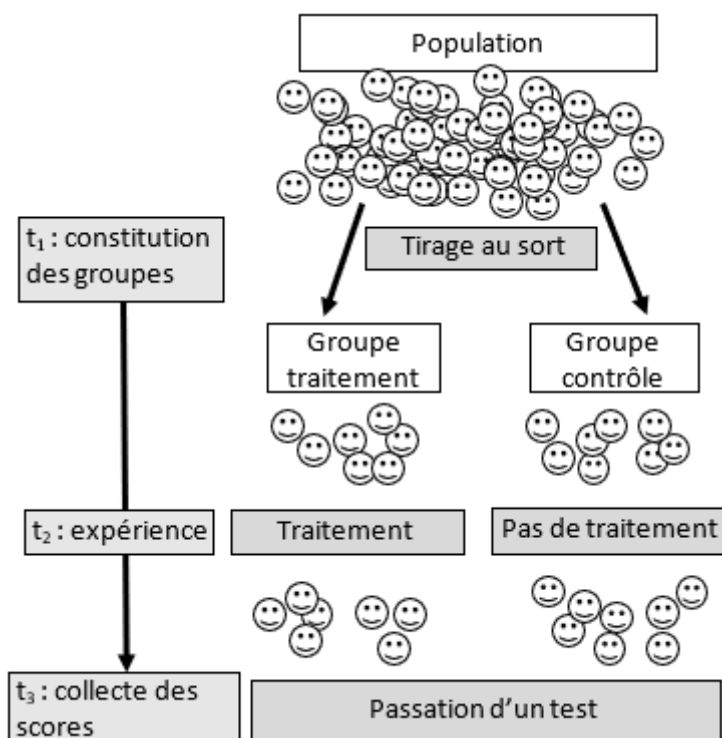


Figure 28 : plan d'expérience dans les études par comparaison de groupes

On souhaite maintenant répondre à la question suivante : « la méthode d'enseignement appliquée au groupe traitement s'est-elle montrée plus efficace que la méthode d'enseignement standard ? »

On sait déjà que si deux échantillons d'élèves sont choisis de façon aléatoire dans une même population, les moyennes des scores de chacun de ces deux groupes se situent dans 95 cas sur 100 dans l'intervalle de fluctuation centré autour de la moyenne des scores de la population (voir le chapitre 4). Donc il ne serait pas surprenant que les moyennes calculées à partir de nos deux échantillons soient différentes : cette différence peut être due à l'échantillonnage (et donc

^a Les deux premières étapes de la figure 1 du chapitre 1 sont réunies en une seule et même étape ici.

^b On peut considérer que ces deux groupes d'élèves sont soumis chacun à un traitement particulier (après tout, le groupe contrôle reçoit bien un enseignement)

au hasard). Il se peut aussi que cette différence soit due à la différence de traitement entre les deux groupes. Si c'est le cas, nous devons nous demander alors si cette différence est suffisante pour justifier une généralisation à toute la population de la méthode d'enseignement étudiée.

Deux questions distinctes attendent donc finalement une réponse.

La première concerne la significativité de la différence des moyennes observée entre les deux échantillons quand on l'analyse d'un point de vue statistique : peut-on prétendre que cette différence ne soit pas uniquement le fruit du hasard qui prévaut dans tout échantillonnage effectué de manière aléatoire (et si oui, dans quelle mesure) ? En voici une autre formulation : la différence observée entre les moyennes des deux groupes (groupe traitement et groupe contrôle) peut-elle être considérée comme due, au moins en partie, à l'influence du traitement appliqué dans un des groupes ?

Si la réponse à cette première question est affirmative, on passera à la seconde question qui concerne l'ampleur de cette différence : comment mesurer et évaluer concrètement cette ampleur dans le contexte de l'étude et quelles préconisations est-on en droit de proposer ? Cette question sera traitée au chapitre suivant.

Revenons donc à la première.

Les tests d'hypothèses

Pour montrer que la différence entre les deux moyennes de deux groupes de données est **statistiquement significative** (ou non), et donc pour montrer qu'elle n'est pas uniquement le résultat du hasard qui prévaut dans tout échantillonnage mais bien la conséquence de la différence entre les deux traitements, les chercheurs ont recours à une méthode : le test d'hypothèse. Un **test d'hypothèse** est une procédure qui met en jeu un raisonnement spécifique avec une terminologie particulière qu'il convient de connaître pour en comprendre les conclusions.

La prudence est à l'origine de l'architecture de cette procédure. En effet, la première étape d'un test d'hypothèse consiste à poser ce qu'il est convenu d'appeler « l'hypothèse nulle » (notée H_0). Cette hypothèse peut s'énoncer ainsi dans notre cas : « la différence observée entre les moyennes des scores des deux échantillons est due au hasard de l'échantillonnage, et non à la différence entre les deux traitements ». L'objectif du test d'hypothèse va être de décider si on rejette cette hypothèse nulle, ou non (c'est-à-dire si on l'accepte). L'hypothèse alternative est la négation de l'hypothèse nulle, notée H_1 . Dans notre cas elle pourrait s'énoncer ainsi : « la différence observée entre les moyennes des scores des deux échantillons n'est pas due au hasard d'échantillonnage mais bien à la différence entre les deux traitements ». Elle est le pendant des intentions du chercheur puisque c'est cette hypothèse que ce dernier souhaite démontrer. Mais ce n'est pas sur elle que porte le test. On acceptera l'hypothèse H_1 dans le cas où l'hypothèse nulle est rejetée^a.

^a Cette conception est sans doute à la source de la difficulté que l'on peut éprouver à comprendre les tests d'hypothèses : l'objectif est finalement de montrer que l'hypothèse nulle est fautive, ce qui va souvent se traduire par des formulations doublement négatives (du type « on ne peut pas rejeter l'hypothèse nulle ») qui sont tout sauf naturelles.

Quand l'hypothèse nulle n'est pas rejetée (et donc quand cette hypothèse est acceptée), on dira que les deux échantillons proviennent de la même population. Quand l'hypothèse nulle est rejetée, on conclura donc en présentant ces deux échantillons comme ne provenant pas de la même population. Ou encore, en affirmant que ces deux groupes proviennent de deux populations différentes : une population qui n'a pas subi le traitement et une population virtuelle qui aurait subi le traitement. Bien sûr, cette deuxième population n'existe pas^a, et ce concept est purement lié à l'utilisation des tests d'hypothèses que nous allons analyser ci-dessous. Par opposition, l'hypothèse nulle devient alors « les moyennes de chacune des deux populations sont égales » (ou bien « la différence de leurs moyennes est égale à 0 »). Les tests d'hypothèses vont donc présenter les échantillons comme provenant de deux populations, et la question revient donc à déterminer s'il s'agit de la même population ou non. Rajoutons tout de même que quand l'hypothèse nulle est rejetée et que l'on conclura en présentant les échantillons comme issus de deux populations différentes, cette différence ne devrait concerner que leurs moyennes, et non leurs écarts-types. Et que la plupart des tests d'hypothèses ne peuvent être utilisés que si l'on admet que ces deux populations ont la même variance (ou, ce qui revient au même, le même écart-type). Nous reviendrons dessus plus loin.

Revenons à nos deux échantillons. Nous avons à notre disposition les moyennes et les écarts-types des scores de deux échantillons : l'échantillon **a** (qui a subi le traitement) et l'échantillon **b** (le groupe contrôle) ainsi bien sûr que la taille de ces échantillons (voir figure 29 ci-dessous). Nous allons passer en revue les trois tests d'hypothèse les plus utilisés dans ce genre de situation : le test Z, le test de Student et l'analyse de la variance (ou ANOVA). Ces tests se fondent sur l'étude de statistiques (telles que définies au chapitre 4) et sont de ce fait souvent dénommés **tests statistiques**. Dans la suite de ce texte, je ferai référence parfois au travail mené par le What Works Clearinghouse (WWC) publié dans son *Procedures Handbook*^b. On va considérer dans tout ce qui suit que les scores sont indépendants^c, les sujets ayant été prélevés aléatoirement dans une population. Le formulaire placé en annexe 4 vous permettra de mieux comprendre le détail des calculs de ce chapitre un peu technique.

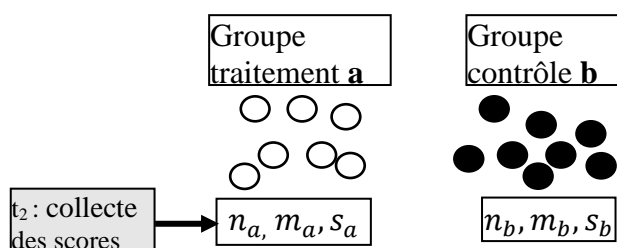


Figure 29 : les données obtenues après le traitement

Le test Z

Les scores sont indépendants (puisque les élèves ont été aléatoirement distribués entre les groupes expérimentaux), et on va se placer dans le cas où les tailles des deux échantillons sont

^a La population réelle étant l'ensemble des sujets à partir duquel deux échantillons sont constitués de façon aléatoire, c'est-à-dire la population expérimentale.

^b WWC (2021)

^c Ce sont les premières réalisations de variables aléatoires indépendantes deux à deux.

supérieures ou égales à 30 : nous sommes donc dans les conditions d'application du théorème central limite. On va considérer que les échantillons proviennent de deux populations dont les paramètres seront notés μ_a et σ_a pour la population dont l'échantillon **a** est extrait ; μ_b et σ_b pour la population dont l'échantillon **b** est extrait. Soient M_a la variable aléatoire « observer la moyenne d'un échantillon pris au hasard dans la population dont est issu l'échantillon **a** » et M_b définie de la même façon pour l'échantillon **b**. Alors m_a et m_b , les moyennes des deux échantillons, sont des réalisations de ces deux variables aléatoires M_a et M_b . On sait que la variable M_a suit une loi normale de paramètres μ_a et $\frac{\sigma_a}{\sqrt{n_a}}$ et que la variable M_b suit une loi normale de paramètre μ_b et $\frac{\sigma_b}{\sqrt{n_b}}$ (voir le chapitre 5). La variable $M_a - M_b$ suit donc une loi normale de paramètres^a $\mu_a - \mu_b$ et $\sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}}$, et donc la variable centrée réduite Z définie ci-dessous suit la loi normale centrée réduite.

$$Z = \frac{M_a - M_b - E(M_a - M_b)}{\sqrt{\text{var}(M_a - M_b)}} = \frac{M_a - M_b - (\mu_a - \mu_b)}{\sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}}}$$

Plaçons-nous maintenant dans le cas où l'hypothèse nulle est vérifiée : les deux populations dont sont issues les deux échantillons ont les mêmes moyennes, donc $\mu_a - \mu_b = 0$ et

$$Z = \frac{M_a - M_b}{\sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}}}$$

En nous fondant sur les propriétés de la loi normale centrée réduite (figure 15), nous pouvons affirmer que si H_0 est vraie, alors $p(-1,96 < Z < 1,96) = 0,95$; ce qui revient à dire que 95% des réalisations de la variable aléatoire Z sont comprises entre $-1,96$ et $1,96$. Et donc que 5% des réalisations de Z sont supérieures à $1,96$ ou inférieures à $-1,96$.

A partir des données de nos deux échantillons, on calcule une réalisation de la statistique Z , la **valeur observée**, que l'on va noter z_{obs} . On ne connaît pas les écarts-types des populations. On va donc les remplacer par leur estimations s_a et s_b puisque les tailles d'échantillon sont supérieures ou égales à 30^b. On obtient finalement

$$z_{obs} = \frac{m_a - m_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

On va maintenant décider si on accepte ou si on rejette l'hypothèse nulle. Ce qui reviendra à rejeter ou à accepter l'hypothèse alternative H_1 . Mais avant, nous devons définir un peu plus précisément cette dernière. En effet, deux cas sont possibles.

^a La moyenne d'une différence de variables aléatoires est égale à la différence des moyennes ; et la variance d'une différence de deux variables aléatoires indépendantes est égale à la somme des variances (voir annexe 5).

^b Mais certains préfèrent calculer un écart-type groupé et utiliser un test de Student y compris quand les tailles des échantillons sont supérieures à 30, suivant en cela la démarche déjà évoquée au chapitre précédent qui concernait les intervalles de confiance.

Premier cas

Si le signe de la différence entre les deux moyennes des populations est inconnu, alors on dit que l'hypothèse H_1 est non directionnelle. C'est le cas par exemple quand il n'est pas impossible que le traitement puisse avoir un effet négatif sur l'échantillon (ce qui, encore une fois, correspond la plupart du temps aux essais contrôlés randomisés mis en œuvre dans le domaine des sciences de l'éducation). On posera $H_1 : \mu_a \neq \mu_b$ ou encore $H_1 : \mu_a > \mu_b$ ou $\mu_a < \mu_b$. On dit que le test est **bilatéral**.

D'après la remarque 1, on sait que dans 95% des cas, les réalisations de Z sont comprises entre $-1,96$ et $1,96$. Cet intervalle est la **zone d'acceptation** de notre test ; la **zone de rejet** (aussi dénommée la zone de risque de première espèce) est alors définie par l'ensemble des valeurs qui n'appartiennent pas à cet intervalle (voir figure 30). Ces zones sont bornées par les **valeurs critiques** de la statistique étudiée (ici $1,96$ et $-1,96$).

Si la statistique observée n'appartient pas à la zone d'acceptation, on rejettera l'hypothèse nulle. On prend bien sûr un risque de se tromper : dans 5% des cas, les réalisations de la variable aléatoire Z sont supérieures à $1,96$ ou inférieures à $-1,96$ quand les deux échantillons proviennent de la même population. Ce risque est le **risque de première espèce** (ou erreur de type I), que l'on note très souvent α : c'est la probabilité de rejeter l'hypothèse nulle alors qu'elle est vraie ; cela revient donc à présenter la différence observée comme étant la conséquence de la différence entre les deux traitements alors que c'est faux. Ce risque a été défini au début du test, quand les zones d'acceptation et de rejets ont été déterminées. Ici $\alpha = 5\%$. On aurait pu choisir un risque plus faible (par exemple $\alpha = 1\%$), et dans ce cas, la zone d'acceptation aurait été bien sûr plus large.

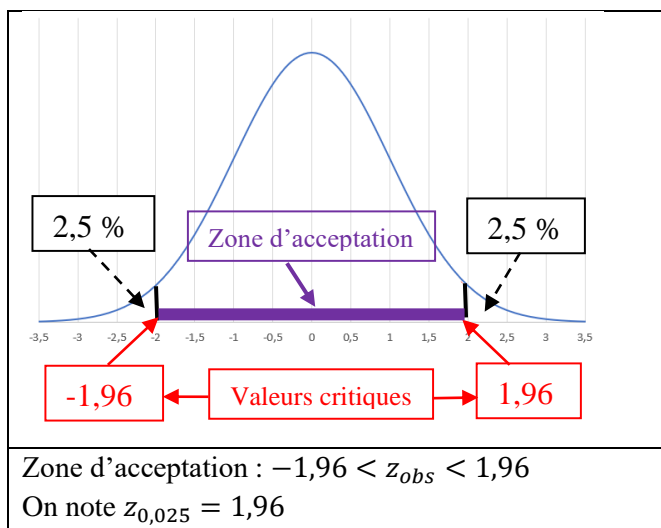


Figure 30 : test bilatéral avec la statistique Z

Si la statistique observée appartient à la zone d'acceptation, on conclura en ne rejetant pas (c'est-à-dire en acceptant) l'hypothèse nulle. Mais là aussi on prend un risque de se tromper : c'est le **risque de seconde espèce** (ou erreur de type II), noté β , qui est le risque d'accepter l'hypothèse nulle alors qu'elle est fautive ; cela revient à présenter la différence observée comme étant le fruit du hasard alors que c'est la différence entre les deux traitements qui en est la cause. On reviendra plus loin sur ces deux risques, le risque α et le risque β .

Deuxième cas

Si le signe de la différence entre les deux moyennes des populations est connu, par exemple dans le cas où il est impossible que le traitement ait un effet négatif (ce qui ne semble pas être le cas des essais contrôlés randomisés mis en œuvre dans le domaine des sciences de l'éducation), alors on dit que l'hypothèse H1 est directionnelle. On posera alors $H1 : \mu_a > \mu_b$ (ou inversement, tout dépend du contexte). On dit que le test est **unilatéral**. Ce cas est traité dans l'annexe 4.

Illustrations concrètes

Six cas fictifs (**exemple 5**) sont présentés ci-dessous pour illustrer ce qui vient d'être dit. Dans chacun de ces six cas, on a deux groupes d'élèves : un groupe traitement, le groupe **a**, et un groupe contrôle, le groupe **b**. On a donc 12 séries de données qui sont les scores sur 20 obtenus par les élèves en fin d'expérience après passation d'un même test. Dans chacun des 12 groupes, les scores sont normalement distribués. Les moyennes des scores des six groupes contrôle sont égales à 10. Pour chacun des 6 cas, les deux groupes traitement et contrôle ont le même écart-type et le même effectif. Enfin, dans chacun des six cas, la moyenne du groupe traitement est telle que son écart avec la moyenne du groupe contrôle (égale à 10) est celui requis pour conclure à une différence statistiquement significative en utilisant le test Z décrit ci-dessus avec $\alpha=0,05$.

Les six cas se différencient par le nombre de données dans les groupes (50, 100 ou 200) et les écarts-types (3 ou 1,5) et la différence des moyennes entre le groupe traitement et le groupe contrôle.

On a rassemblé les résultats dans le tableau 15.

	Cas 1		Cas 2		Cas 3		Cas 4		Cas 5		Cas 6	
	a	b	a	b	a	b	a	b	a	b	a	b
$n_a = n_b$	50				100				200			
$s_a = s_b$	3		4		3		4		3		4	
$m_a - m_b$	1,18		1,76		0,83		1,25		0,59		0,88	
m_a, m_b	11,18	10	11,76	10	10,83	10	11,25	10	10,59	10	10,88	10

Tableau 15 : différences de moyennes statistiquement significatives (exemple 5)

Ainsi, pour le cas 1, les scores des 2 groupes ont comme effectif commun 50, et le même écart-type égal à 3 (ce qui signifie par exemple que pour le groupe contrôle les 2/3 des scores sont compris entre 7 et 13 et que 90% des scores sont compris entre 4 et 16) ; dans ce cas, la moyenne du groupe traitement doit être au moins égale à 11,18 pour que l'on puisse conclure à une différence statistiquement et significativement différente entre les deux moyennes. Pour le cas 2, les scores des groupes ont toujours comme effectif commun 50, mais cette fois leur écart-type est égal à 4 (ce qui signifie par exemple que pour le groupe contrôle les 2/3 des scores sont compris entre 6 et 14 et que 90% des scores sont compris entre 2 et 18) ; cette fois la moyenne du groupe traitement doit être au moins égale à 11,76 pour que l'on puisse conclure à une différence statistiquement significative. Les six représentations graphiques correspondantes à chacun de ces cas sont présentées figure 31.

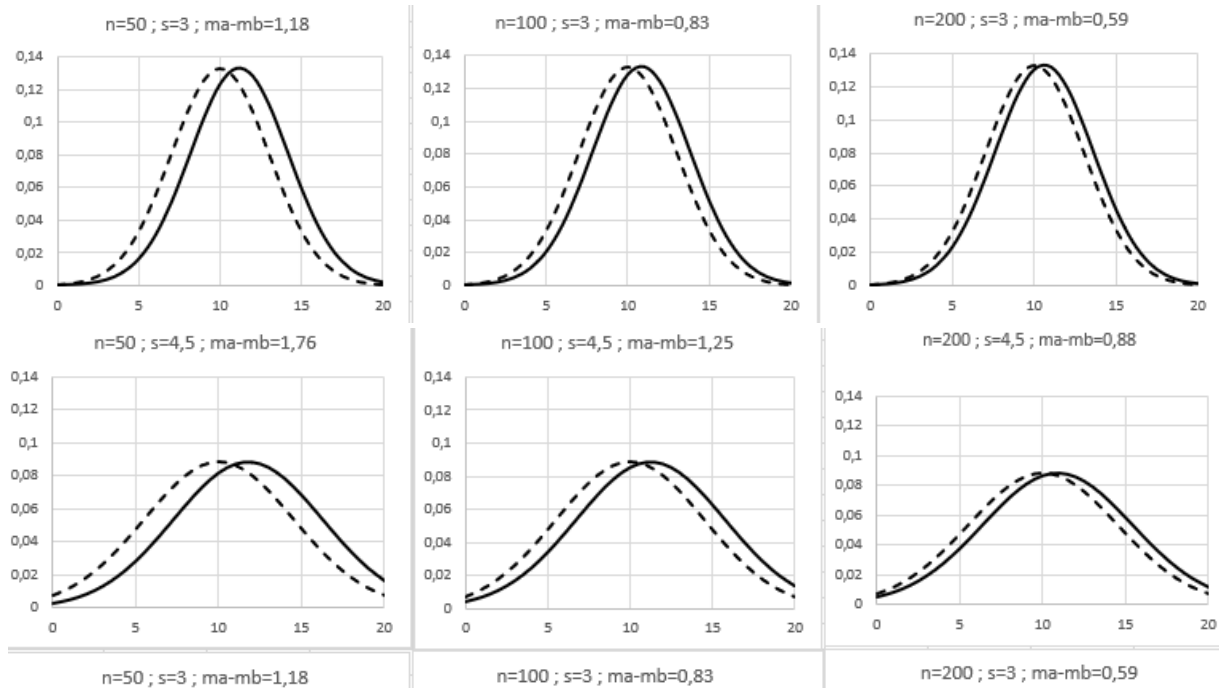


Figure 31 : différences de moyennes statistiquement significatives (exemple 5)

A effectif constant, on remarque donc que la différence des moyennes requise pour conclure à une différence statistiquement significative diminue quand l'écart-type des groupes diminue. Et quand l'écart-type est constant, la différence des moyennes requise pour conclure à une différence statistiquement significative diminue quand l'effectif des groupes augmente. Autrement dit, quand la taille des échantillons augmente (toutes choses égales par ailleurs), l'écart entre la moyenne du groupe traitement et la moyenne du groupe contrôle requis pour conclure à un écart statistiquement significatif diminue (et c'est l'inverse en ce qui concerne l'écart-type des scores).

Nous allons reprendre une nouvelle fois les données de l'**exemple 3** et nous allons prélever expérimentalement deux groupes de 40 copies (tableau 16). Bien sûr, ici nous n'avons pas de groupe traitement, mais juste deux groupes contrôles ; il va donc s'agir d'illustrer le cas où, a priori, on s'attend à ne pas trouver de différence statistiquement significative entre les deux groupes.

Groupe a

3	3	4	3	3	1	3	2	1	5	1	2	4	5
2	5	5	2	5	5	2	5	3	1	2	4	5	5
5	5	3	5	2	1	5	5	5	2	5	4		

moyenne	3,45
écart-type	1,5013

Groupe b

3	3	4	4	1	1	4	2	1	3	5	3	4	3
5	4	5	5	1	3	5	3	2	2	2	5	4	5
5	2	3	2	3	5	5	2	2	5	3	2		

moyenne	3,275
écart-type	1,3585

Tableau 16 : Deux groupes de 40 données (**exemple 3**)

On calcule la valeur observée :

$$z_{obs} = \frac{3,45 - 3,275}{\sqrt{\frac{1,5013^2}{40} + \frac{1,3585^2}{40}}} \approx 0,54665$$

Comme $0,54665 < 1,96$ la différence des moyennes n'est pas statistiquement significative au seuil de 5%.

Risque α , risque β

On a vu que la décision d'accepter ou de rejeter l'hypothèse nulle sera prise en acceptant deux risques. C'est clairement le risque α qui tient la vedette, car on veut avant tout éviter de se tromper en désignant à tort un traitement comme efficace. C'est ce risque que l'on va choisir de prendre (en fixant la valeur α), et qui va conditionner la décision qui sera prise. Mais le risque β doit également être considéré car il est important de détecter les traitements efficaces. Il faudra peut-être faire un choix, car le risque β dépend de α (fixé en premier lieu), et plus α est petit, plus β est grand. Mais il dépend aussi de la taille des échantillons : plus les tailles sont importantes, plus β est petit. Et le risque α étant fixé par le chercheur (encore une fois, on cherche avant tout à limiter le risque de première espèce^a), ce dernier devra jouer sur les tailles d'échantillons pour diminuer le risque β . Ce qui aura comme effet d'augmenter la **puissance** du test, cette puissance étant égale à $1 - \beta^b$.

Le test de Student

Même quand ses conditions d'applications sont réunies, le test Z cède de nos jours la place à un test bien plus répandu dans les études quantitatives qui nous intéresse : le test de Student ou test T . Si j'ai pris le temps de détailler le premier, c'est qu'il permet plus facilement de comprendre les raisonnements et la procédure qui sont au cœur des tests d'hypothèse décrits dans ce livre.

Quand l'un au moins des échantillons à une taille inférieure à 30, on ne peut plus utiliser le test Z . Dans ce cas, pour conclure, nous allons devoir renforcer nos conditions afin de procéder à un test d'hypothèse :

- La loi suivie par la variable aléatoire X doit être normale.
- Les variances des deux populations, σ_a^2 et σ_b^2 , doivent être égales^c.

La première condition pourra être vérifiée par une analyse visuelle de l'histogramme des données ou encore par des tests statistiques non décrits ici. La deuxième condition impose que l'homogénéité des variances des deux échantillons soit vérifiée par un test statistique, également non décrit ici. Remarquons que les variables aléatoires X_i doivent toujours être indépendantes deux à deux.

^a Quand $\alpha < 5\%$ on dit que le test est conservatif.

^b Le statisticien conduit alors des calculs (non traités ici) basés sur certaines hypothèses dans le but de déterminer les tailles d'échantillon minimales requises afin de détecter un effet du traitement.

^c Mais comme nous l'avons déjà signalé, cette égalité est également inhérente au principe même de l'étude par comparaison de groupes qui sont censés ne différer que par leur moyenne.

Si ces conditions sont vérifiées, on pourra utiliser le test de Student, test mis en œuvre (notamment par le WWC) de façon habituelle y compris pour des échantillons de grande taille. Ce test suit les mêmes étapes que celles décrites pour le test Z précédemment étudié. La statistique est cette fois le T de Student, qui, sous l'hypothèse nulle, suit une loi de Student à $n_a + n_b - 2$ degré de liberté. On a

$$T = \frac{M_a - M_b}{\sqrt{\frac{S^2}{n_a} + \frac{S^2}{n_b}}} = \frac{M_a - M_b}{S \times \sqrt{\frac{1}{n_a} + \frac{1}{n_b}}}$$

Ce test se démarque d'un test Z par deux aspects notables :

- l'estimation des écarts-types des populations est ici l'écart-type groupé, ou regroupé, ou commun (*pooled SD*) ; cet écart-type groupé, que l'on notera s , est la moyenne pondérée par leurs degrés de liberté des écarts-types des deux échantillons, on dit aussi que c'est la racine carrée de la variance intra (notée var_{intra}).
- les valeurs critiques dépendent toujours du risque α choisi, mais également des degrés de liberté des échantillons n .

On a

$$s^2 = \frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{n_a + n_b - 2}$$

La statistique observée est calculée de la façon suivante

$$t_{obs} = \frac{m_a - m_b}{\sqrt{\frac{s^2}{n_a} + \frac{s^2}{n_b}}} = \frac{m_a - m_b}{s \times \sqrt{\frac{1}{n_a} + \frac{1}{n_b}}} = \frac{m_a - m_b}{s \times \sqrt{\frac{n_a + n_b}{n_a \times n_b}}}$$

De la même façon que pour le test Z , les valeurs critiques de la statistique T définissent une zone d'acceptation (et donc une zone de rejet) associée à l'hypothèse nulle pour un risque α donné mais aussi pour un certain degré de liberté (calculé à partir des tailles d'échantillon). Ici aussi on pourra choisir entre un test bilatéral et un test unilatéral. Mais dans notre cas, et comme nous l'avons déjà signalé, seuls les tests bilatéraux sont à considérer, et les valeurs critiques présentées ci-dessous ne concernent que ce type de test. D'une façon générale, on les note $\pm t_{\alpha/2; n_a + n_b - 2}$. Par exemple, pour $n_a + n_b - 2 = 60$ et $\alpha=0,05$, on a $\pm t_{0,025; 60} = \pm 2,0003$. Pour $n_a + n_b - 2 = 78$ et $\alpha=0,05$, on a $\pm t_{0,025; 88} = \pm 1,9908$. Ce sont des valeurs peu éloignées de $\pm 1,96$ qui sont les valeurs critiques quand $n_a + n_b - 2$ tend vers l'infini (voir table des valeurs critiques en annexe 4).

Comme vous l'aurez compris, test de Student et test Z ont des conditions d'application différente. Ils peuvent néanmoins être utilisés de façon équivalente dans certaine situation. Une synthèse vous est proposée **figure 32**.

On remarquera que si $n_a = n_b$ ou si $s_a = s_b$, alors $t_{obs} = z_{obs}$.

$n_a \geq 30$ et $n_b \geq 30$?

OUI

NON

($n_a < 30$
ou $n_b < 30$)

Les distributions des données des deux échantillons sont normales (test de normalité ou/et analyse visuelle).

OUI

NON

Les variances sont homogènes (rapport entre 1 et 3 ; test d'homogénéité des variances).

OUI

NON

On utilise d'autres tests

On utilise alors un **test Z** : quand $\mu_a - \mu_b = 0$ (sous hypothèse nulle), alors $M_a - M_b$ (variable aléatoire de la différence des moyennes d'échantillonnage) suit une **loi normale** de paramètres $\mu_a - \mu_b = 0$ et

$\sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}}$. Comme les effectifs sont suffisamment

grands, s_a et s_b sont des estimations de σ_a et σ_b , on

va donc estimer $\sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}}$ par $\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$. On en

déduit alors que la variable aléatoire $z = \frac{M_a - M_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$ suit

une loi normale centrée réduite. Dans 95% des cas les réalisations de cette variable aléatoire sont comprises entre 1,96 et -1,96. On va alors calculer

$$z_{obs} = \frac{m_a - m_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$
 et conclure.

Remarque : on peut également utiliser un test de **Student**.

On utilise alors un **test de Student** en estimant un écart-type groupé s ($s^2 = \frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{n_a + n_b - 2}$) :

quand $\mu_a - \mu_b = 0$ (sous hypothèse nulle), alors la variable aléatoire $T = \frac{M_a - M_b}{\sqrt{\frac{s^2}{n_a} + \frac{s^2}{n_b}}}$ suit une loi de Student

à $n_a + n_b - 2$ degrés de liberté. Dans 95% des cas, les réalisations de cette variable aléatoire sont comprises entre $t_{0,025;n_a+n_b-2}$ et $-t_{0,025;n_a+n_b-2}$. On va alors calculer $t_{obs} = \frac{m_a - m_b}{\sqrt{\frac{s^2}{n_a} + \frac{s^2}{n_b}}}$ et conclure.

Remarques :

1. Si $n_a \geq 30$ et $n_b \geq 30$ alors on peut s'affranchir des conditions de normalité et d'homogénéité des variances
2. Si $n_a \approx n_b$, le test sera plus robuste face à l'inégalité des variances
3. Si n_a et n_b sont grands, le test sera plus robuste face à la non-normalité des données.

Figure 32 : test Z ou test T ? Bilan et arbre de décision


Si nous reprenons les données de notre **exemple 3** (tableau 16), nous sommes dans le cas où $n_a = n_b$, et on peut vérifier par le calcul que $t_{obs} = z_{obs} \approx 0,54665$. Cette fois, on compare cette valeur à la valeur critique $t_{0,025; 88} = 1,9908$: la conclusion est la même qu'après avoir mené un test Z.

Enfin, si la distribution des données ne suit pas une loi normale, nous ne sommes plus dans les conditions d'application du test de Student. Et dans ce cas, des tests dits « non paramétriques » devront être utilisés. Mais dans le domaine des sciences sociales, les études quantitatives font presque toujours l'hypothèse d'une distribution normale de leurs données, et de ce fait ce type de test n'est pratiquement jamais utilisé^a.

La valeur-p

Les résultats publiés après avoir conduit ce type de test d'hypothèse sont souvent accompagnés du calcul de la valeur-p (*p-value*), cette valeur étant espérée la plus petite possible. La valeur-p (notée également p) est la probabilité pour la statistique de se trouver dans la zone de rejet, zone définie par la valeur observée calculée sous l'hypothèse nulle. Pour les tests bilatéraux, c'est donc la probabilité d'observer une différence des moyennes supérieure à la valeur observée, ou inférieure à son opposé, sous le seul effet du hasard. Pour un test Z bilatéral, on a donc

$$p = P(-z_{obs} > Z > z_{obs})$$

Si par exemple nous calculons une valeur observée égale à 1,8, on trouvera une valeur-p égale à 0,072  pour un test Z bilatéral (voir figure 33), et on notera $p = 0,072$ à côté de la valeur de la statistique observée. Pour prendre une décision on peut donc comparer le risque α choisi avant de débiter l'expérience à cette valeur-p calculée à partir de la statistique observée : la valeur-p doit être inférieure au risque α pour rejeter l'hypothèse nulle. Si nous avons choisi $\alpha = 0,05$ (ou 5%), nous aurions donc $p > \alpha$ donc l'hypothèse nulle n'aurait pas été rejetée dans un test bilatéral. En résumé, si $p < \alpha$ alors H_0 est rejetée, et si $p > \alpha$ alors H_0 n'est pas rejetée.

Certains chercheurs vont encore plus loin en se contentant de calculer la valeur-p, sans autre formalité. Il ne s'agit plus ici de prendre une décision en ayant fixé un seuil critique avant de mener l'expérience, mais de donner le résultat de la valeur-p et de conclure en fonction de la valeur trouvée (approche de Fischer). Il est reproché à cette méthode de ne pas définir de règle de décision avant d'obtenir les résultats : on peut alors suspecter que les conclusions sont plus facilement influencées par la conviction du chercheur que dans l'approche précédemment développée (approche de Neyman et Pearson).

^a WWC Procedures Handbook 2020, p.E-7

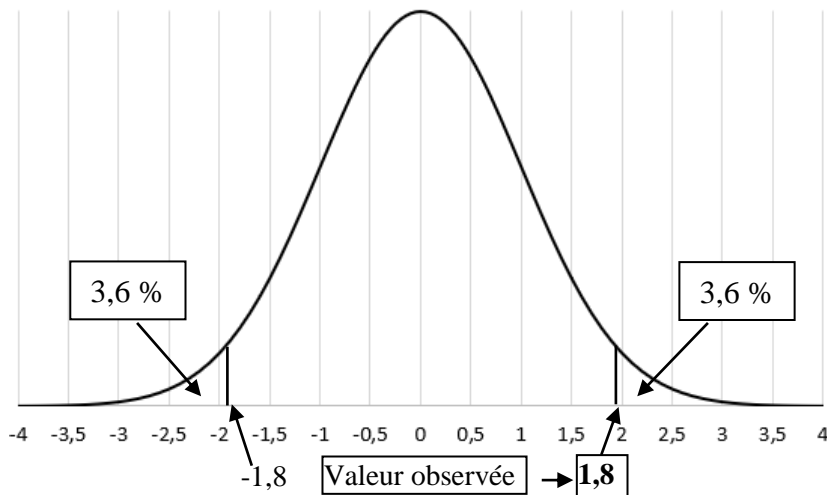


Figure 33 : signification de la valeur- p dans le cas d'un test Z bilatéral

L'analyse de la variance (ANOVA)

On utilise ce test le plus souvent quand on a plus de 2 échantillons^a à étudier, mais il est possible de l'utiliser pour deux échantillons (et on a besoin de comprendre ce test pour utiliser l'ANCOVA, voir plus loin). L'objectif est toujours de conclure sur le sens qu'il convient de donner à la différence entre les moyennes des échantillons. Ici c'est la statistique F qui est utilisée (F de Fisher-Snédecor). Les conditions d'application sont les mêmes que celles du test de Student. Cette statistique est égale au quotient de la variance observée entre les échantillons, variance due aux différents traitements (ou encore variance inter ou expliquée) notée var_{inter} , divisée par la variance interne aux échantillons (ou encore variance intra ou résiduelle encore désignée comme étant l'erreur) notée var_{intra} . La variance inter mesure la dispersion des moyennes de chaque groupe par rapport à la moyenne générale (moyenne pondérée des moyennes), alors que la variance intra mesure, pour l'ensemble des groupes, la dispersion à l'intérieur de chaque groupe des données autour de leur moyenne. Le chercheur souhaite bien sûr une variance inter la plus élevée possible et une variance intra faible. Et c'est pour cette raison que l'on va étudier leur quotient (que l'on souhaite donc le plus grand possible).

$$F = \frac{var_{inter}}{var_{intra}}$$

Mais le degré de liberté des k échantillons, égal à $k - 1$, et le degré de liberté de l'ensemble des N données, égal à $N - k$, interviennent ici dans les calculs de la statistique observée et pour la lecture des tables. C'est pour cette raison que l'on note $F(k - 1, N - k)$ la statistique observée. Dans le cas de deux échantillons $k = 2$, et $N = n_a + n_b$, donc le degré de liberté des échantillons est égal à 1 et le degré de liberté de l'ensemble des données est égal à $n_a + n_b - 2$.

^a De nombreuses études analysent les résultats de plus de deux groupes d'élèves, soumis à des méthodes d'enseignement différentes ainsi qu'à une méthode dite « standard ».

Là encore, les mêmes étapes seront suivies que pour les tests précédents. Et comme toujours, on connaît la distribution de F^a si l'hypothèse nulle est vérifiée. On calcule la statistique de la manière suivante (les calculs détaillés sont proposés en annexe 4) :

$$F(1, n_a + n_b - 2) = \frac{var_{inter}}{var_{intra}}$$

Comme toujours, on compare la valeur observée à une valeur critique donnée par des tables et on calcule une valeur $-p$ \square . Cette valeur critique dépend cette fois de trois nombres : le risque α (souvent 5 ou 1 %), le degré de liberté du nombre des k échantillons et le degré de liberté de l'ensemble des données. Si la valeur observée est inférieure à celle donnée par les tables, on acceptera l'hypothèse nulle. Si la valeur est supérieure à celle donnée par les tables, on rejettera l'hypothèse nulle.

Quand on n'a que deux échantillons, faire un test F ou un test T revient en fait au même, et on peut montrer que $F(1, n_a + n_b - 2) = t_{obs}^2$

La procédure suivie par les tests

D'une façon générale, les tests statistiques mis en œuvre vont toujours suivre les 5 étapes suivantes.

Étape 1 : on énonce l'hypothèse nulle (notée H_0) et l'hypothèse alternative (notée H_1) ; celle-ci peut être directionnelle ou non. L'hypothèse nulle considère que les deux échantillons ont été prélevés dans deux populations de façon aléatoire et que la différence des moyennes de ces deux populations est nulle.

Étape 2 : on choisit une statistique qui suit une loi connue sous l'hypothèse nulle comme un score Z , le T de Student, le F de Fischer-Snédecour par exemple^b, ce choix dépendant en partie des conditions requises pour son application. Par exemple, on a vu que pour utiliser un test Z , les scores devaient être indépendants et les tailles des échantillons supérieures à 30. On choisit dans le même temps un risque α (souvent 5% ou 1%) qui permet de déterminer pour la statistique choisie une zone de rejet et une zone d'acceptation de l'hypothèse nulle.

Étape 3 : on calcule la valeur observée à partir des données des échantillons.

Étape 4 : si la statistique observée appartient à la zone de rejet on rejettera l'hypothèse nulle avec le risque α de se tromper ; si cette valeur est comprise dans la zone d'acceptation, alors on acceptera l'hypothèse nulle avec le risque β de se tromper.

Étape 5 : on conclura sur la significativité de la différence au niveau de la population en assumant un certain risque.

^a Pour plus de simplicité, ici nous ne différencions pas la notation utilisée pour la statistique de celle utilisée pour la valeur observée et donc l'estimation.

^b D'autres lois peuvent être utilisées, comme la loi du Khi2, voir chapitre 10 pour la statistique Q .

Utiliser les intervalles de confiance


Une autre façon de questionner la significativité statistique d'une différence entre deux moyennes passe par le calcul de l'intervalle de confiance à 95 % de cette différence. En reprenant les raisonnements développés au chapitre précédent et en les appliquant à la variable aléatoire $M_a - M_b$ on peut en effet définir un intervalle de confiance estimant la différence des moyennes $\mu_a - \mu_b$ en distinguant deux cas. Dans le premier cas, les tailles d'échantillon sont supérieures ou égales à 30. Cet intervalle est alors égal à :

$\left[m_a - m_b - 1,96 \times \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}; m_a - m_b + 1,96 \times \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}} \right]$ avec $\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$ l'écart-type de la différence des moyennes déjà rencontré auparavant (voir le test Z).

Dans le second cas, si l'une des tailles d'échantillons inférieure à 30, on calculera l'intervalle suivant

$\left[m_a - m_b - t_{\frac{\alpha}{2}; n_a+n_b-2} \times \sqrt{\frac{s^2}{n_a} + \frac{s^2}{n_b}}; m_a - m_b + t_{\frac{\alpha}{2}; n_a+n_b-2} \times \sqrt{\frac{s^2}{n_a} + \frac{s^2}{n_b}} \right]$ avec s l'écart-type groupé déjà rencontré plus haut.

Dans les deux cas, cela se traduit ainsi : l'intervalle de confiance contient la valeur de la différence des moyennes des deux populations avec un niveau de confiance de 0,95. Si l'intervalle inclus la valeur 0 alors il ne sera pas possible d'exclure l'hypothèse d'une différence nulle au niveau de confiance choisi.

On peut également procéder en étudiant les deux intervalles de confiance des deux moyennes : si ces intervalles se recouvrent au moins partiellement, la différence entre les moyennes est caractérisée comme statistiquement non significative. C'est la méthode retenue par l'OCDE pour exploiter les résultats de l'enquête PISA 2012. L'OCDE donne bien sûr les moyennes des scores obtenus par les élèves pour chacun des chaque pays, mais précise bien que ces moyennes doivent être considérées avec précaution et que ce sont leurs intervalles de confiance qui doivent être utilisés pour comparer les pays. Ainsi le score moyen de la France est comparable à celui des pays de l'OCDE suivants : Irlande, Danemark, Nouvelle-Zélande, République Tchèque, Royaume-Uni, Islande, Luxembourg, Norvège, Portugal^a. Les erreurs standards (les écart-type des moyennes d'échantillonnage donc) ont été publiées par l'OCDE mais pas les intervalles de confiance que l'on trouvera dans une publication de la DEPP^b. Les informations concernant la République Tchèque, la France, le Royaume-Uni, l'Italie et l'ensemble des pays de l'OCDE sont rassemblés dans le tableau 17 (**exemple 6**), et les intervalles de confiance correspondants ont été représentés par un **graphique en forêt** (*forest plots*) figure 34 .

^a Table I.2.3a dans OCDE (2014)

^b DEPP (2013)

	Moyenne	Erreur standard
Italie	485	2.0
Royaume-Uni	494	3.3
France	495	2.5
République Tchèque	499	2.9
OCDE	494	0.5

Tableau 17 : moyennes et erreurs standards de certains pays (PISA 2012)

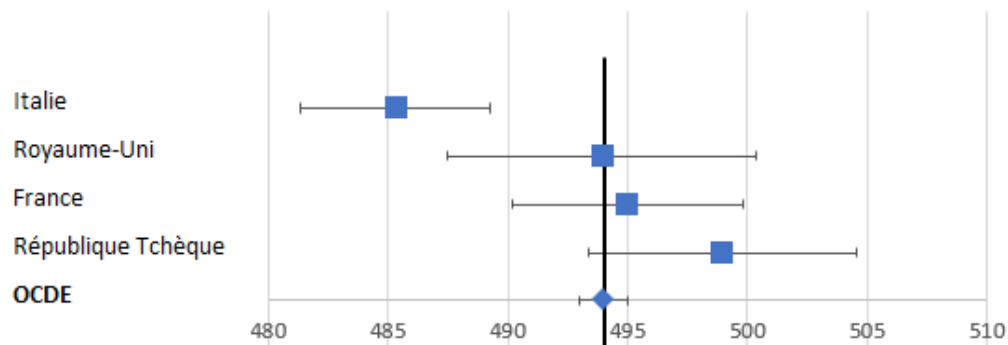


Figure 34 : moyennes et intervalles de confiance à 95% des scores PISA

La question que l'on se pose ici est : la différence observée dans les moyennes est-elle due à des fluctuations d'échantillonnage (les échantillons étant prélevés dans la population des élèves de l'OCDE) ou bien est-elle expliquée par le facteur « pays » ? Si nous nous intéressons plus particulièrement à la France, son intervalle de confiance chevauche celui du Royaume-Uni, de la République Tchèque et de l'ensemble des pays de l'OCDE, donc on acceptera la première hypothèse et on rejettera la seconde. Par contre l'Italie se démarque clairement, et sa moyenne pourra être considérée comme statistiquement significativement différente des moyennes de la France (mais aussi de la République Tchèque et de l'OCDE).

Chapitre 8. Taille d'effet du traitement

Si on cherche à évaluer la différence entre deux groupes expérimentaux, c'est bien pour en tirer des conclusions d'ordre plus général. Faut-il ou non généraliser le traitement à tout un établissement ? À tout un pays ? C'est à cette question que nous allons tenter d'apporter une réponse maintenant.

Ampleur de la différence entre deux moyennes

Dans certains cas, la simple analyse de la différence des moyennes des échantillons peut apporter une réponse satisfaisante à la question précédente, notamment quand les échelles utilisées sont « parlantes^a ». Cependant, utiliser la différence des moyennes des échantillons peut ne pas être satisfaisant pour au moins deux raisons.

La première est que la différence de ces moyennes a bien entendu comme unité celle de l'échelle utilisée pour le calcul des scores et elle ne donne donc pas d'information sur son ampleur quand on n'est pas familier de cette échelle (les mesures peuvent être faites dans des unités dont on n'a pas l'habitude, comme par exemple les scores obtenus dans les enquêtes PISA).

La seconde raison tient au fait que les statisticiens cherchent toujours à tenir compte du « bruit », c'est-à-dire de la variabilité à l'intérieur des groupes expérimentaux (mesurée par une estimation de la variance des données ou encore la variance intra) pour évaluer correctement la variabilité entre deux groupes (mesurée par une estimation de la variance des moyennes ou encore la variance inter)^b. Pour deux moyennes données, les conclusions seront d'autant plus fortes que la variance intra sera faibles.

Enfin l'intérêt porté actuellement aux méta-analyses nous donnera une troisième raison comme nous le verrons dans la dernière partie.

Deux cas fictifs (issus de l'**exemple 5**) sont présentés ci-dessous pour illustrer l'importance à accorder à la variance intra. Dans chaque cas, il s'agit de deux groupes d'élèves, un groupe traitement, le groupe **a**, et un groupe contrôle, le groupe **b**. Tous les élèves ont subi le même test en fin d'expérience et ont obtenu un score sur 20. On considère que les 4 séries de données sont normalement distribuées, et ont le même effectif de 50 élèves. Pour chacun des deux cas, les moyennes du groupe contrôle sont égales à 10 et les moyennes du groupe traitement sont égales à 11,76^c, ce qui revient à dire que les variances inter sont égales. Les deux cas se différencient uniquement par les écarts-types (et donc la variance intra), égale à 4,5 dans le premier cas et à 2 dans le second cas. Une simple analyse visuelle semble indiquer que l'ampleur de la différence entre les deux moyennes des échantillons est plus importante dans le premier cas que dans le second cas (figure 35), le recouvrement des deux courbes étant moins important dans le premier cas que dans le second (les surfaces représentent en fait des

^a Dans certains contextes, les procédures statistiques peuvent céder la place au bon sens.

^b Voir l'annexe 1 pour le calcul de ces variances.

^c Cette différence est statistiquement significativement différente de zéro (test Z avec $\alpha=0,05$, voir chapitre précédent)

proportions d'élèves)^a. C'est une façon simple d'illustrer la nécessité de tenir compte des variances intra dans ce type d'analyse.

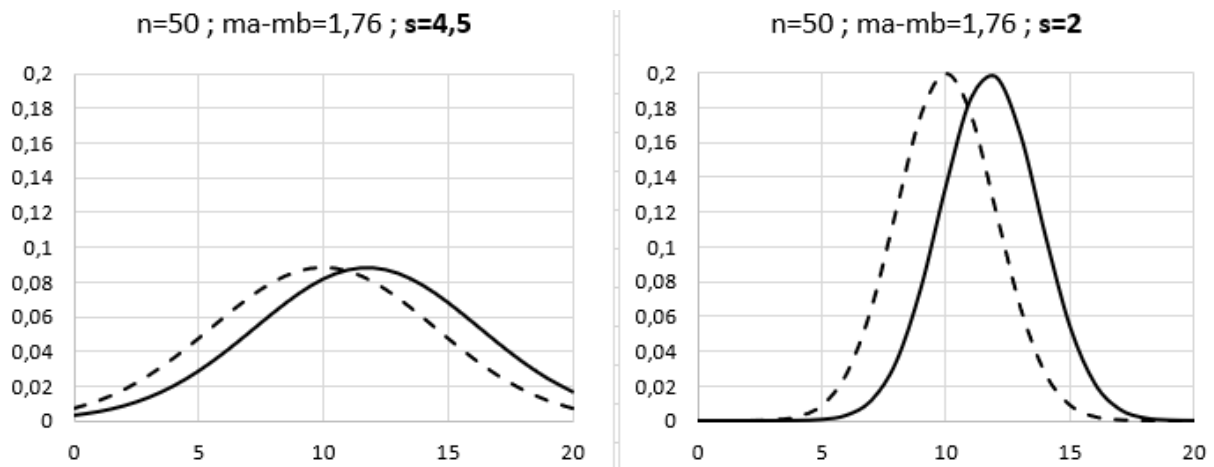


Figure 35 : impact de la valeur de l'écart-type pour des échantillons d'élèves ayant la même différence de moyennes.

Taille de l'effet du traitement

Considérons deux échantillons, l'échantillon **a** qui subit la méthode d'enseignement dont on cherche à évaluer les bénéfices (c'est le groupe traitement) et l'échantillon **b** qui ne subit pas le traitement (c'est le groupe contrôle). Ces échantillons sont représentatifs de deux populations, la population **a** et la population **b**, dont les paramètres sont μ_a et μ_b , leurs moyennes, et σ leur écart-type : ici aussi on considère ici que les variances des populations sont égales et donc que $\sigma_a = \sigma_b = \sigma$. Par définition, la **taille de l'effet (ou taille d'effet) du traitement** au niveau de la population dont sont issus les échantillons est égale à la différence entre les deux moyennes divisée par l'écart-type des populations. On note ce nouveau paramètre δ . On a donc

$$\delta = \frac{\mu_a - \mu_b}{\sigma}$$

On dit aussi que ce quotient est la **différence des moyennes standardisée**^b. Si nous reprenons les données de notre exemple 5 et en considérant que les groupes d'élèves sont des populations, dans le premier cas on aurait $\delta \approx 0,39$ et dans le second cas $\delta \approx 0,88$. On comprendra à la fin du chapitre 12 que la différence entre ces deux tailles d'effet est importante dans le domaine des sciences de l'éducation.

On pourrait, en prélevant plusieurs échantillons, calculer plusieurs estimations de la taille d'effet ; et la distribution de ces valeurs serait alors la distribution d'échantillonnage de l'estimateur de la taille d'effet. Cela veut dire donc que, de la même façon que pour une

^a Le recouvrement des deux courbes est un indicateur permettant de donner du sens à une taille d'effet (voir chapitre 12).

^b « Taille d'effet » est le nom donné à une famille d'indicateurs (seuls quelques uns sont étudiés ici) qui estiment l'ampleur de l'effet d'un traitement ou de l'influence d'un facteur sur une population ; cela peut être aussi, par exemple, un coefficient de corrélation.

moyenne, nous pouvons calculer une estimation ponctuelle de la taille d'effet réelle δ , son erreur standard, c'est-à-dire l'écart-type de la distribution d'échantillonnage de la taille d'effet, et finalement une estimation par intervalle de ce paramètre^a. Mais si paramètres (*parameters*) et estimations (*estimates*) sont clairement identifiés dans tous les articles traitant de la taille d'effet, la notion d'estimateur (*estimator*) quant à elle n'est que rarement rencontrée, et quand c'est le cas il s'agit plutôt d'étudier leurs distributions d'échantillonnage que de les considérer comme des variables aléatoires. Il arrive parfois que l'on soit confrontée à une « variance d'estimation » ou à une « estimation non biaisée », alors que, comme nous l'avons vu dans les chapitres précédents, les termes « variance » ou « non biaisé » ne peuvent que se référer à des estimateurs (qui sont des variables aléatoires, contrairement aux estimations qui sont des nombres réels). C'est Larry HEDGES (1981) qui le premier a fait le point sur les estimateurs de la taille d'effet δ en étudiant leurs distributions dans un article qui reste encore aujourd'hui la référence indiscutable à la source de tous les calculs présentés ci-après.

Dans tout ce qui suit, les calculs ont été menés en assumant que les scores des élèves dans les deux populations sont normalement distribués autour de leur moyenne (μ_a pour le groupe **a**, μ_b pour le groupe **b**) avec le même écart-type σ .

Estimations ponctuelles

Nous allons donc estimer δ , la taille de l'effet au niveau de la population, à partir des données des deux échantillons. On connaît les scores des élèves de ces deux groupes obtenus après passation d'un test en fin d'expérience, ainsi que leurs moyennes (m_a et m_b) et leurs écart-types (s_a et s_b). Les moyennes des populations seront estimées par les moyennes des échantillons m_a et m_b . Mais qu'en est-il de l'estimation de l'écart-type de la population σ ? Deux méthodes de calcul sont très fréquemment utilisées pour estimer ce paramètre, qui ont donné naissance à deux familles d'estimations : le d de Cohen (proposée par Jacob COHEN en 1969) et le g de Hedges (défini par Larry HEDGES en 1981) d'une part, le Δ de Glass (que l'on doit à Gene GLASS en 1977) d'autres part^b. Pour les deux premières, c'est l'utilisation (ou non) d'un facteur correctif pour des échantillons de petite taille qui va faire la différence. La plupart des formules qui vous sont présentées ci-dessous ont été extraites de *Introduction to meta-analysis* (Michael BORENSTEIN, 2009) ; les auteurs de ce livre mentionnent le paramètre δ et ses estimations en éludant la notion d'estimateur. Il peut paraître surprenant que des calculs différents (légèrement, mais différents tout de même) puissent être utilisés par les chercheurs, et ce encore de nos jours. Ces différences sont en partie liées à la difficulté que représente l'estimation de l'écart-type de la population (dont nous n'avons presque jamais la valeur exacte).

^a On applique le même raisonnement que celui conduit au chapitre 5 pour la moyenne M .

^b Des discussions ont agité les statisticiens autour des dénominations mêmes de ces différents calculs (voir par exemple, D. ENZMANN (2015)). On pourra signaler tout de même que L. HEDGES poursuit son travail de recherche en collaboration avec le WWC et ce sont les notations utilisées dans *Introduction to meta-analysis* (2009) dont il est co-auteur qui ont été utilisées ici.

Le d de Cohen

Quand on a de bonnes raisons de penser que les écarts-types des groupes traitement et contrôle sont des estimations de l'écart-type de la population, on estime l'écart-type de la population par l'écart-type groupé déjà rencontré dans le cadre du test de Student au chapitre précédent. On calcule alors d une estimation de la taille d'effet δ de la façon suivante :

$$d = \frac{m_a - m_b}{s}$$

avec

$$s^2 = \frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{n_a + n_b - 2}$$

On peut remarquer que

$$t_{obs} = \frac{d}{\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}}$$

Cette relation implique que la taille de l'effet et la significativité statistique de la différence de deux moyennes varient dans le même sens.

Le g de Hedges

En 1981, HEDGES montre que le d de Cohen^a est biaisé (son espérance n'est pas égale à la taille d'effet de la population, δ) et que ce biais est substantiel pour les petites tailles d'échantillon. L'estimateur non biaisé est obtenu en multipliant le d de Cohen par un facteur multiplicatif correctif ω

$$g = d \times \omega = \frac{m_a - m_b}{s} \times \omega$$

avec

$$\omega = 1 - \frac{3}{4(n_a + n_b) - 9}$$

C'est l'estimation utilisée par le WWC par défaut dans les études par comparaison de groupes utilisant des données continues^b.

Vous constaterez qu'avec des tailles d'échantillons supérieures à 30, le facteur correctif ω est très proche de 1 et se trouve de ce fait ignoré par de nombreux chercheurs. Les tailles d'effets sont la plupart du temps données au dixième près et, dans la majorité des cas, elles sont inférieures à 1,5^c. Dans ce cas, on ne constate une différence entre g de Hedges et d de Cohen au dixième près que pour des tailles d'échantillons totales inférieures à 20 (voir annexe 2).

$$\text{Si } n_a + n_b \geq 60, \text{ alors } \omega \geq 1 - \frac{3}{4 \times 60 - 9} \approx 0,987$$

^a Curieusement, HEDGES ne fait références dans son article qu'au Δ de Glass, tout en utilisant comme point de départ la formule dite du d de Cohen, illustrant parfaitement la difficulté qu'il y a ici à associer de façon claire et déterminée un nom à une formule.

^b Voir plus loin (chapitre 11).

^c Sur 4430 tailles d'effet publiées par le WWC, seules 1,44% sont supérieures à 1,5 (proportions calculées sur les résultats exportés à partir de leur site internet en mai 2020).

Et de la même façon qu'avec le d de Cohen, on peut faire le lien avec le chapitre précédent, en remarquant que

$$t_{obs} = \frac{g}{\omega \times \sqrt{\frac{1}{n_a} + \frac{1}{n_b}}}$$

Cette formule est utilisée par le WWC pour calculer le g de Hedges quand les études sélectionnées dans leurs méta-analyses publient les valeurs de t_{obs} , de n_a et de n_b .

Le Δ de Glass

Quand l'écart-type du groupe contrôle (le groupe **b**) est une bonne estimation de l'écart-type de la population^a ou quand les écart-types des groupes contrôle et traitement sont très différents (et donc que l'écart-type groupé ne semble pas estimer correctement l'écart-type de la population), alors on utilise le calcul suivant pour estimer la taille de l'effet δ ^b :

$$\Delta = \frac{m_a - m_b}{s_b}$$

On justifie aussi ce choix en soulignant que dans les études expérimentales le traitement influe sur la moyenne bien sûr mais également sur l'écart-type, et qu'il serait donc préférable d'utiliser l'écart-type du groupe contrôle comme estimation de l'écart-type de la population.

Variances de d , g et Δ

Pour chacune des trois estimations d , g et Δ , on va calculer une variance de leur distribution d'échantillonnage. Les formules seront admises.

La variance du d de Cohen (notée var_d) est égale à :

$$var_d = \frac{n_a + n_b}{n_a \times n_b} + \frac{d^2}{2(n_a + n_b)}$$

Le premier terme reflète l'incertitude dans l'estimation de la différence des moyennes, le second reflète l'incertitude dans l'estimation de l'écart-type σ^c . Et comme $g = d \times \omega$, on calcule alors la variance de g (notée var_g) de la façon suivante^d :

$$var_g = var_d \times \omega^2 = \left[\frac{n_a + n_b}{n_a \times n_b} + \frac{d^2}{2(n_a + n_b)} \right] \times \omega^2 = \frac{n_a + n_b}{n_a \times n_b} \times \omega^2 + \frac{g^2}{2(n_a + n_b)}$$

Comme $\omega < 1$ alors $\omega^2 < 1$ et $var_g < var_d$. Donc g est l'estimateur de plus faible variance. Comme de plus il est non-biaisé il est utilisé de préférence au d de Cohen de nos jours.

^a C'est l'option retenue par R. SLAVIN (2009) *Effective Programs in Middle and High School Mathematics: A Best-Evidence Synthesis* (voir chapitre 11).

^b Ici tout le monde s'accorde à dénommer cette taille de l'effet le Δ de Glass.

^c Il existe d'autres expressions très légèrement différentes pour cette variance, par exemple le dénominateur du second terme est parfois donné comme égal à $2(n_a + n_b - 2)$, BORENSTEIN (2019, p.213) et HEDGES (1981).

^d Voir les propriétés des variances dans l'annexe 5.

Et enfin, pour le Δ de Glass, on calcule^a la variance de Δ notée var_{Δ} :

$$var_{\Delta} = \frac{n_a + n_b}{n_a \times n_b} + \frac{\Delta^2}{2(n_b - 1)}$$

Les racines carrées de ses variances permettent alors de calculer les écarts-types de ces estimateurs, et on parle là aussi (comme pour la moyenne), d'erreurs standards. On a

$$s_d = \sqrt{var_d} \quad ; \quad s_g = \sqrt{var_g} \quad \text{et} \quad s_{\Delta} = \sqrt{var_{\Delta}}$$

Remarquons enfin que pour toutes ces variances, plus les tailles d'échantillon sont importantes, et plus les variances sont faibles^b.

Estimation par intervalle

La distribution du g de Hedges est une loi de Student non centrée asymétrique que l'on peut approcher par une loi normale d'espérance δ pour des degrés de liberté suffisamment grands (HEDGES, 1981). On va donc estimer δ en calculant un intervalle de confiance à 95 % de la façon suivante :

$$[g - 1,96 \times s_g \ ; \ g + 1,96 \times s_g]$$

Si cet intervalle ne contient pas la valeur zéro, on pourra conclure à un effet statistiquement significatif au niveau de confiance 0,95. Comme toujours, plus les tailles des échantillons sont importantes et plus l'étendue de l'intervalle est faible.

Tests statistiques

On peut également suivre la procédure des tests d'hypothèse pour estimer la signification statistique à associer au g de Hedges. Il va s'agir ici de décider si on peut écarter l'hypothèse nulle H_0 qui est « la taille d'effet au niveau de la population est égale à zéro », ou encore $\delta = 0$. Comme la distribution du g de Hedges pour des tailles d'échantillons suffisamment grandes tend à suivre une loi normale, sous H_0 la variable centrée réduite^c suit une loi normale centrée-réduite.

On calcule alors (voir annexe 4) :

$$z_{obs} = \frac{g}{s_g}$$

On compare cette valeur à la valeur critique au niveau de confiance choisi (par exemple pour $\alpha = 0,05$ on compare z_{obs} à 1,96 et à -1,96). La valeur- p correspondante sera également calculée, et comparée à α .

^a MARCO, 2019.

^b Estimations ponctuelles, variances et écarts-types sont rassemblés dans un tableau en annexe 2.

^c En reprenant les notations des chapitres 6 et 7, on a $Z = \frac{G - E(G)}{\sqrt{var(G)}} = \frac{G}{\sqrt{var(G)}}$ sous H_0 .

Un exemple numérique

L'exemple numérique ci-dessous^a (**exemple 7**) concerne deux études repérées par les lettres A et B. Les données sont fictives et permettent de s'appropriier les règles de calcul vues ci-dessus. Cet exemple sera repris et développé au chapitre 10.

Les données de départ

Nom de l'étude	n_a	n_b	m_a	m_b	s_a	s_b
A	65	65	98	92	21	22
B	200	200	94	82	19	17

Calcul du d de Cohen, du g de Hedges, du Δ de Glass pour l'étude A

On calcule pour l'étude A les tailles d'effet selon les trois méthodes vues ci-dessus. On a besoin de l'écart-type groupé s pour calculer le d de Cohen et le g de Hedges :

$$s^2 = \frac{(65 - 1) \times 21^2 + (65 - 1) \times 22^2}{65 + 65 - 2} \approx 462,5$$

$$d = \frac{98 - 92}{\sqrt{462,5}} \approx 0,279$$

On calcule ω le terme correctif qui permet de calculer le g de Hedges :

$$\omega = 1 - \frac{3}{4(65 + 65) - 9} \approx 0,994$$

Et donc,

$$g = 0,994 \times 0,279 \approx 0,277$$

On remarque qu'au centième près, les deux calculs donnent le même résultat 0,28.

Pour le Δ de Glass, on divise la différence des moyennes par l'écart-type du groupe contrôle :

$$\Delta = \frac{98 - 92}{22} \approx 0,273$$

Intervalle de confiance, test Z (études A et B)

On va d'abord s'intéresser à l'étude A et calculer les bornes de l'intervalle de confiance et la valeur- p associée au test Z. On va donc avoir besoin de la variance de g :

$$var_g = \frac{65 + 65}{65 \times 65} \times 0,994^2 + \frac{0,277^2}{2(65 + 65)} \approx 0,0307$$

Et donc

$$\frac{g}{s_g} = \frac{0,277}{\sqrt{0,0307}} \approx 1,5829$$

^a Cet exemple est inspiré de BORENSTEIN (2009, p.88) qui accompagne ses exemples de fichiers Excel à télécharger sur www.meta-analysis.com (puis cliquer sur « Books on meta-analysis »).

On calcule alors une valeur- p égale à 0,1134. La taille d'effet calculée n'est pas statistiquement significativement différente de 0 (test Z avec $\alpha=0,05$). L'intervalle de confiance au seuil de 0,95 est $[-0,0661 ; 0,6208]$ qui inclue la valeur 0.

Pour l'étude B, on calcule de la même façon le g de Hedges et sa variance. On a alors

$$g = 0,6644 \quad var_g \approx 0,0105$$

Et donc

$$\frac{g}{s_g} = \frac{0,6644}{\sqrt{0,0105}} \approx 6,48$$

La taille d'effet est ici statistiquement significativement différente de 0 (valeur- $p = 0,0000$) et l'intervalle de confiance au seuil de 0,95 qui est $[0,4634 ; 0,8653]$ n'inclue pas la valeur 0.

Cela montre ici l'impact des tailles d'échantillons dans ce type de calcul au niveau de l'estimation par intervalle (et donc du sens que l'on peut donner au résultat).

Les deux intervalles de confiance sont représentés ci-dessous (figure 36).

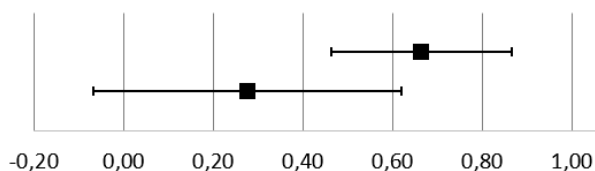


Figure 36 : intervalles de confiance du g de Hedges pour l'étude A (en bas) et l'étude B (en haut)

Conclusions

Nous n'avons toujours pas répondu à la question posée au début de ce chapitre. Il nous manque encore une grille de lecture pour interpréter les tailles d'effet, et ce que nous ferons au chapitre 12. Avant cela, il nous reste à évoquer quelques unes des difficultés rencontrées par les chercheurs qui sont évoquées au prochain chapitre. Mais surtout nous avons encore à comprendre comment les résultats de plusieurs études peuvent être agrégés. Les tailles d'effet globales qui seront alors calculées (vous en saurez plus après la lecture de la troisième partie de ce livre) devront elles aussi être interprétées et là aussi, des grilles de lectures nous seront indispensables.

Chapitre 9. La complexité du terrain

Jusqu'à présent, nous avons considéré que les deux échantillons d'élèves tirés au sort étaient tous les deux des échantillons représentatifs d'une même population et que ces deux groupes étaient équivalents avant le traitement. Cela reviendrait, si nous nous intéressons par exemple aux élèves scolarisés en classe de troisième en France, à prélever des noms au hasard sur une liste recensant tous ces élèves, puis à constituer deux échantillons pour lesquels deux traitements distincts seraient mis en place. Mais dans les faits, les expériences seront le plus souvent menées sur des groupes qui n'auront pas été constitués par tirage au sort dans la population et les élèves n'auront pas été distribués aléatoirement dans les groupes traitement et contrôle. Ces groupes sont par exemple des classes d'élèves constituées avant l'expérience, et l'étude suivra alors un design quasi-expérimental. Dans ce cas, on ne pourra pas considérer que les caractéristiques des élèves (comme leur niveau socioéconomique, leur niveau de compétence avant l'expérience) sont distribuées de manière équivalente dans les groupes. Or, si par exemple l'un des groupes s'avère plus performant que l'autre dans un domaine ciblé par le traitement, on ne pourra pas considérer ces élèves (et leurs scores) comme indépendants les uns des autres, puisque l'appartenance de l'élève à tel ou tel groupe influe sur son score ; la même remarque peut être faite concernant le niveau socioéconomique. Cette indépendance est pourtant une condition d'utilisation des analyses statistiques habituellement menées. Il sera donc indispensable de prendre en compte les connaissances acquises (et/ou d'autres facteurs) par les élèves avant l'expérience.

Quand les deux échantillons sont différents avant le traitement

Quel que soit le plan d'expérience utilisé pour constituer les deux groupes d'élèves (essai contrôlé randomisé ou étude quasi-expérimentale), il est possible que l'un de ces échantillons soit plus performant que l'autre avant l'expérience, et cette différence peut avoir une influence sur la significativité statistique qu'on pourra associer à la différence des scores après le traitement, mais aussi sur celle de la taille de l'effet du traitement.

Pour tenir compte de cette différence initiale, les deux échantillons sont testés deux fois : une première fois avant le traitement, puis une seconde fois après le traitement. On obtient donc 4 jeux de données : les scores prétest des groupes contrôle et traitement, et les scores posttest de ces mêmes groupes (qui nous intéressent plus particulièrement). Dans chaque échantillon, ces derniers sont susceptibles d'être dépendants des premiers : on dit que les variables sont **corrélées**. On va tenir compte de cette dépendance en évaluant la **corrélation** entre les deux variables, la variable score posttest (qui est qualifiée ici de variable dépendante), et la variable score prétest (qui est qualifiée ici de variable indépendante). Le plan d'expérience est schématisé par la figure 37.

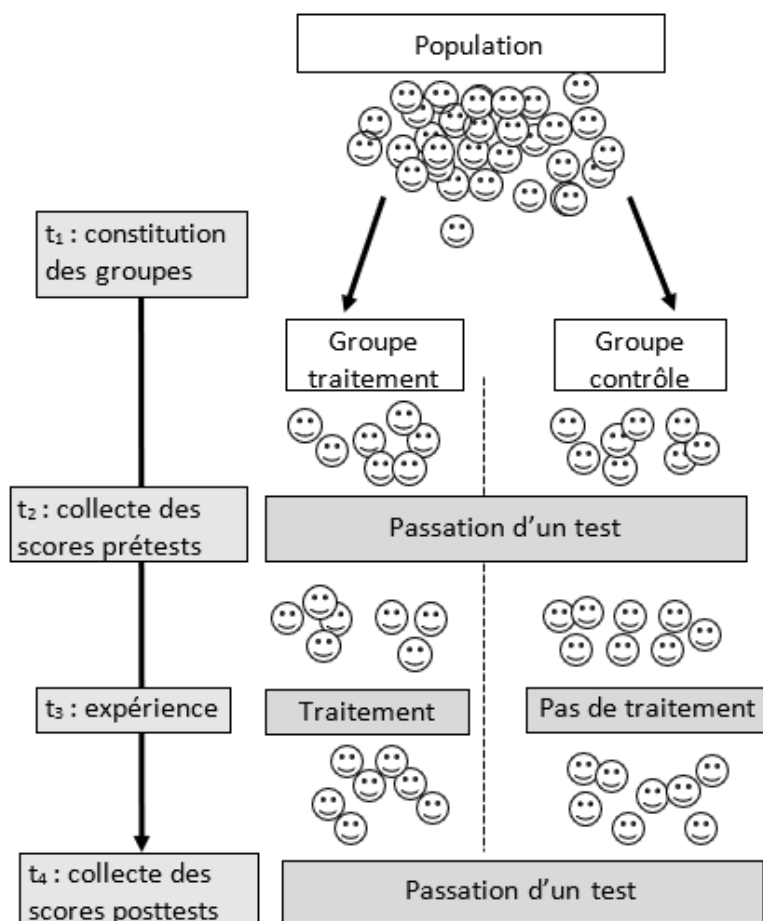


Figure 37 : plan d'expérience tenant compte du niveau des élèves avant le traitement

Corrélation entre scores posttest et scores prétest dans un groupe

Considérons dans un premier temps un seul groupe de n élèves. Dans ce groupe, chaque élève i a obtenu un score prétest (noté x_i) et un score posttest (noté y_i). Nous avons donc une série de n couples de valeurs (x_i, y_i) qui constituent notre ensemble de données. On représente alors graphiquement cet ensemble de données par un nuage de points : chaque point représente un élève avec comme abscisses son score prétest et comme ordonnée son score posttest (plusieurs points peuvent se trouver les uns sur les autres \square). Ce nuage de points prend bien souvent la forme d'un nuage étiré semblant suivre plus ou moins une droite dont la pente est positive. À titre d'exemple, le nuage de points des notes sur 10 obtenues par 56 élèves^a est présenté ci-dessous (figure 38).

^a Ce sont les données fictives du groupe traitement de l'exemple 8.

Scores posttest

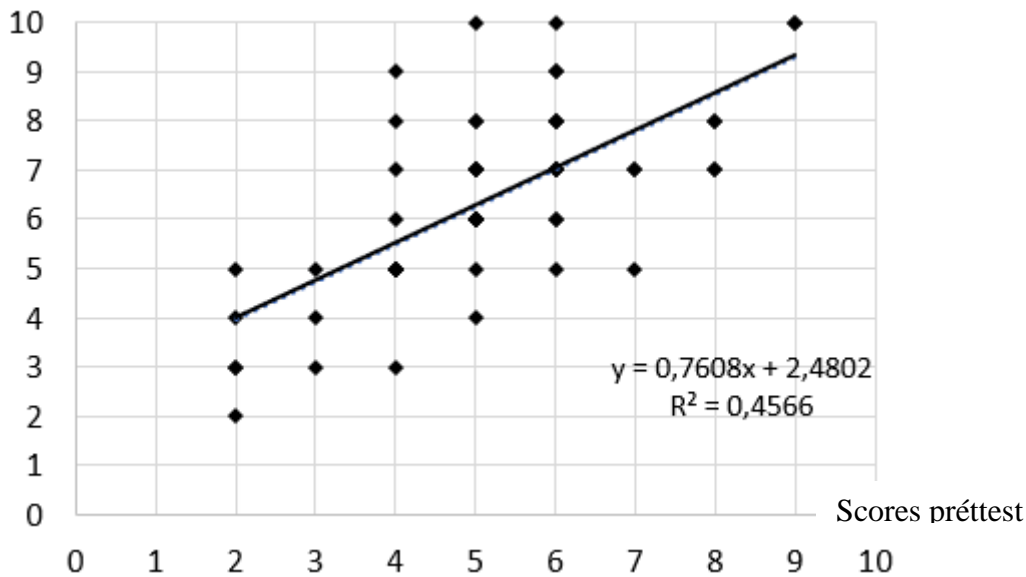


Figure 38 : scores posttests en fonction des scores prétests

Cela reflète le fait bien connu que le score d'un élève à un moment donné dépend de son score antérieur. La droite tracée sur le nuage peut prétendre modéliser notre nuage de point, c'est une des **droites de régression**. On utilise souvent la méthode des moindres carrés^a. La pente de cette droite donne une information sur la variation de la variable dépendante (ici le scores posttest) quand la variable indépendante (ici le scores prétest) augmente d'une unité.

La qualité de ce modèle est évaluée par deux nombres, le coefficient de corrélation r et le coefficient de détermination r^2 . Ce **coefficient de détermination** est un nombre compris entre 0 et 1 (ou 0% et 100%) ; il représente le pourcentage de variation de la variable dépendante (donc le score posttest) en fonction des variations de la variable indépendante (donc le score prétest). Plus les points sont proches de la droite de régression et plus le coefficient de détermination est important (il est égal à 100 % quand tous les points sont alignés sur la droite). Les calculs de ces deux indicateurs (pente de la droite et coefficient de détermination) vous sont expliqués en annexe 1.

Dans notre exemple, le coefficient de détermination est de 45,7 % et la droite a comme équation $y = 0,76x + 2,48$. Donc 45,7 % de la variation des scores posttests peut être expliqué par la variation des scores prétests d'après ce modèle linéaire. Et quand le score prétest augmente de 1 point, le score posttest augmente de 0,76 point toujours selon ce modèle linéaire. Les résultats de cette expérience sont considérés dans le champ des sciences sociales comme significatifs^b.

Il nous reste maintenant à considérer deux échantillons d'élèves qui ont passé deux tests : un test avant l'expérience et l'autre après l'expérience. Nous allons reprendre la procédure des tests d'hypothèses en tenant compte cette fois de la corrélation entre les scores prétest et posttest.

^a On minimise le carré des distances des points expérimentaux à la droite.

^b Et là aussi des tests statistiques peuvent être menés pour déterminer si ces valeurs sont significativement différentes de zéro.

ANCOVA ou ANOVA de la différence^a

On a déjà vu précédemment que pour analyser une différence de deux moyennes, on pouvait tester statistiquement cette différence par une analyse de la variance ou encore ANOVA (chapitre 7), en calculant la statistique F . Maintenant que nous avons remarqué que les scores posttest des élèves étaient dépendants de leur score prétest, on va associer à cette ANOVA les informations obtenues par l'analyse de la corrélation. C'est toujours le rapport entre la variance inter des scores posttest et la variance intra de ces mêmes scores que nous calculons, mais ces variances vont d'abord être ajustées et prendre en compte la corrélation mentionnée ci-dessus. On va en fait ôter à ces variances la partie qui correspond à la corrélation entre scores prétest et scores posttests. Il s'agit finalement de répondre à la question de recherche suivante : si les deux groupes (contrôle et traitement) avaient eu le même niveau au début de l'expérience, la différence entre les deux groupes en fin d'expérience est-elle (au moins en partie) expliquée par la différence des traitements ?

Comme pour les tests précédemment analysés, des conditions devront être respectées pour appliquer ces procédures (normalité des données ou taille des échantillons supérieure à 30, homogénéité des variances), auxquelles on devra ajouter l'homogénéité des pentes des droites de régression.

On a toujours les mêmes étapes à franchir.

L'hypothèse nulle H_0 est encore : « les moyennes des scores posttests des deux populations dont sont issues les deux groupes sont égales ». L'hypothèse H_1 est la négation de H_0 , donc : « les moyennes sont différentes (au moins deux moyennes sont statistiquement différentes entre elles s'il y a plus de deux groupes) ». Il n'y a pas de test unilatéral possible ici.

On connaît la distribution du F de Fischer-Snédecour sous l'hypothèse nulle. Par rapport au calcul effectué dans une ANOVA, deux choses ont tout de même changé : les variances ont été ajustées aux scores prétests et le degré de liberté de la variance ajustée intra a été réduit de 1^b.

On a donc

$$F(k - 1, N - k - 1) = \frac{var_{inter}ajustée}{var_{intra}ajustée}$$

Avec deux échantillons ($k = 2$) cela donne :

$$F(1, N - 3) = \frac{var_{inter}ajustée}{var_{intra}ajustée}$$

On compare comme toujours la statistique observée à une valeur critique donnée par des tables (cette valeur dépend bien entendu du risque α choisi ainsi que des deux degrés de liberté). Si la valeur observée est inférieure à celle donnée par les tables, on ne rejettera pas l'hypothèse nulle. Si la valeur est supérieure à celle donnée par les tables, on rejettera l'hypothèse nulle. L'**exemple 8** vous permettra de mieux assimiler les techniques de calcul.

^a Les détails de ces calculs sont proposés en annexe 4. Voir également <http://vassarstats.net/textbook/>

^b la valeur critique de F est plus élevée quand on perd un degré de liberté.

Une autre façon de tenir compte de ces scores prétest, est de considérer non plus les scores posttests comme la variable intéressante, mais bien la différence des scores (score posttest – scores prétest) de chaque élève. Dans ce cas, on va se ramener à une analyse de la variance de la différence des scores (*gain scores*). La question de recherche n'est plus exactement la même, puisqu'on s'intéresse à la différence moyenne des progrès des élèves selon qu'ils soient dans le groupe traitement ou le groupe contrôle. L'hypothèse nulle pose donc que les moyennes des différences de scores des deux populations sont égales. On procède alors à une ANOVA des différences des scores (voir chapitre 7) dont on sait que, quand seul deux groupes sont étudiés, elle est équivalente à un test de Student. Là aussi les calculs détaillés sont fournis en annexe 4.

Vous trouverez certainement des auteurs préconisant une ANOVA à mesures répétées qui prend en considération un facteur inter-sujet (le traitement), un facteur intra-sujet (le temps) et l'interaction entre ces deux facteurs. J'ai laissé ce type d'analyse de côté car ANOVA à mesure répétée et ANOVA de la différence des scores conduisent à des résultats similaires^a.

Taille de l'effet ajustée

Nous avons déjà vu comment calculer une taille d'effet quand on dispose des scores posttest de deux échantillons. Dans les études au design quasi expérimental il est indispensable d'intégrer à ces calculs la corrélation entre les scores prétest et les scores posttest, toujours dans le but de tenir compte de l'influence des connaissances initiales des élèves sur leur performances en fin d'expérience. Quand on tient compte des scores prétests, on se trouvera dans l'un de ces deux cas.

Premier cas : les moyennes des scores prétest des deux échantillons sont très différentes. On considèrera alors que ces échantillons sont trop différents au départ et on exclura la possibilité de pouvoir comparer raisonnablement le traitement à l'absence de traitement à partir de ces deux échantillons. Ainsi le WWC écarte les études pour lesquelles les scores prétests des deux échantillons ont plus d'un quart d'écart-type de différence^b.

Second cas : les moyennes des scores prétests sont peu différentes (par exemple leur différence est inférieure à 0,25 écart-type). On a alors deux solutions.

Dans le cadre d'une analyse de type ANCOVA, on pourra calculer des moyennes posttests ajustées aux moyennes prétest, puis calculer la taille de l'effet avec ces moyennes ajustées. On doit dans ce cas diviser la différence de ces moyennes posttests ajustées par s , une estimation de l'écart-type de la population calculé à partir des écarts-types non ajustés des scores posttests^c. C'est cette solution qui est privilégiée par le WWC.

$$g = \frac{m_a \text{ajustée} - m_b \text{ajustée}}{s}$$

On peut également calculer une taille d'effet à partir de la différence des différences des moyennes dans le groupe traitement d'une part et dans le groupe contrôle d'autre part (dans le

^a BECKER L. (1999) et KNAPP R. (2009). Ce dernier propose également une équation permettant de passer du t de Student calculé dans une analyse de la différence des scores au F calculée dans une ANCOVA.

^b *Standards Handbook*, WWC, p.14

^c Voir Chapitre 7 pour le calcul de s .

cas d'une ANOVA de la différence des scores par exemple). Cette méthode est réservée aux cas où les moyennes ajustées ne sont pas disponibles et les écarts-types des scores (et non des différences de scores) non ajustés publiés. Tous ses calculs sont détaillés en annexe 4.

Les analyses multiples

Si vous jetez un œil sur les études publiées dans le domaine scientifique qui nous intéresse, vous remarquerez vite que bien peu d'auteurs se contentent de ne mesurer qu'une seule compétence ne concernant qu'un seul groupe traitement à un unique point final d'une seule expérience. Au contraire, vous constaterez qu'une multitude de mesures sont rapportées, qui peuvent concerner plusieurs sous-domaine (par exemple la maîtrise des calculs algébriques et la résolution de problèmes) ou bien être faites à différents moments après l'intervention (immédiatement après puis 6 mois après par exemple). Dans ce cas, plusieurs tests d'hypothèse seront menés simultanément par les auteurs et la significativité statistique des résultats est susceptible d'être surévaluée^a. Le WWC applique alors la méthode de Benjamini-Hochberg^b qui vise à corriger la significativité des différentes p -valeurs calculées et par là-même la significativité statistique de certaines différences. Quand des sous-échantillons disjoints d'élèves sont étudiés (par exemple quand l'échantillon est partagé en groupes de niveaux scolaires différents), des calculs de tailles d'effet plus complexes sont alors menés^c. Enfin si une étude présente des résultats conformes aux normes WWC pour plus d'une mesure de résultats dans un domaine, les tailles d'effet de tous les résultats de cette étude sont combinées en une taille d'effet moyenne de l'étude en calculant la moyenne arithmétique des tailles d'effets.

Quand l'unité d'analyse n'est pas l'unité d'affectation

Une autre source de difficulté vient du fait que bien souvent ce ne sont pas des élèves qui sont attribués aléatoirement à un groupe traitement ou contrôle, mais des classes entières (on les désigne alors comme des clusters ou des grappes). Et cela change un peu la donne. Car même si c'est le hasard qui intervient quand on distribue les classes soit dans le groupe traitement soit dans le groupe contrôle, ce sont bien les compétences des élèves (pris individuellement) qui seront évaluées et mesurées. On se trouve donc dans la situation où l'unité d'affectation (une classe) n'est pas l'unité d'analyse (un élève), et cela va compliquer l'évaluation de la significativité statistique d'une différence de moyennes. Si on se contente d'appliquer les calculs vus précédemment, l'erreur standard de l'effet estimé (l'écart-type de la taille d'effet par exemple) sera sous-estimée, et donc sa significativité statistique sera quant à elle surestimée. On va donc devoir tenir compte de la corrélation entre les réponses des élèves d'une même classe. Le WWC apporte dans ce cas une correction dans les calculs de la statistique observée t en calculant dans un premier temps cette statistique de façon classique (voir chapitre 8) puis dans un second temps en calculant une valeur corrigée ainsi qu'un degré de liberté corrigé également (qui interviendra dans la détermination de la valeur- p). Une taille d'effet

^a Plus on effectue de tests sur un même échantillon, plus la probabilité d'obtenir un résultat significatif augmente ; par exemple, quand un test est négatif, on peut avoir tendance à poursuivre les recherches jusqu'à trouver un résultat convenable (*look-elsewhere effect*).

^b WWC (2020), annexe F du *Procedures Handbook*

^c Voir WWC 2021 p.23

corrigée peut également être calculée, mais dans ce cas l'effet est négligeable. Les formules corrigées sont proposées en annexes 2 et 7 et l'**exemple 9** applique ces méthodes de calculs^a.

Nous venons de faire le point sur les principaux résultats publiés par les études quantitatives par comparaison de groupes. Comme annoncé en introduction, il s'agit ici du premier niveau de recherche : des études individuelles sont menées par des chercheurs dans le but d'évaluer l'efficacité d'un traitement. Ces études, dites primaires, constituent la matière première des méta-analyses, les études secondaires, et c'est de ce deuxième niveau de recherche dont il va être question maintenant.

^a H. COOPER, *The Handbook of research synthesis and meta-analysis* (2019), p.235

Troisième partie : les études

Chapitre 10. Les méta-analyses

De très nombreuses études quantitatives ont été réalisées dans le but d'évaluer l'efficacité d'interventions menées auprès d'élèves, comme une approche pédagogique particulière ou l'utilisation de supports innovants. Il arrive souvent que la même intervention soit explorée par plusieurs études. Et même si ces études partagent une méthode commune basée notamment sur les calculs statistiques décrits dans les chapitres précédents, elles se différencient souvent les unes des autres par plusieurs aspects : l'âge des élèves, leur nationalité, la date de l'expérience, les équipes de recherches mais aussi leurs résultats. Se pose alors la question de l'exploitation de ces derniers : comment peut-on les synthétiser afin d'éclairer les responsables administratifs et politiques en charge de l'organisation du système éducatif d'un pays ? Quelle réponse peut-on apporter à un chef d'établissement, un formateur, un enseignant qui s'interroge sur l'efficacité d'une méthode d'enseignement ?

Nous sommes ici au deuxième niveau de recherche décrit en introduction de ce livre, et c'est pour répondre à ces questions que le chercheur va réaliser une **méta-analyse**. Les méta-analyses, tout comme les études par comparaison de groupes qui en sont la matière première, suivent une démarche scientifique rigoureuse. L'ensemble des concepts et des méthodes mises en œuvre pour la réalisation de ces études secondaires dans le domaine des Sciences de l'éducation est explicité de manière complète et détaillée dans *The Handbook of research synthesis and meta-analysis* (Harris COOPER, Larry HEDGES et Jeffrey VALENTINE, 2019). Pour aller à l'essentiel, la lecture de *Introduction to Meta-analysis* (Michael BORENSTEIN, Larry HEDGES, Julian HIGGINS et Hannah ROTHSTEIN, 2009) permet de saisir le cadre méthodologique qui soutient la réalisation de ces méta-analyses et plus particulièrement de comprendre les procédures statistiques déployées (dont la description constitue le squelette de ce livre), mais aussi le sens que l'on peut donner aux résultats. Les nombreux exemples résolus (pour lesquels des fichiers Excel sont disponibles sur le site internet compagna) font de cet ouvrage un outil pédagogique particulièrement utile aux néophytes. Enfin, pour entrer dans le détail des protocoles mis en place par certaines institutions financées par des gouvernements qui attendent des solutions pratiques et efficaces, je vous invite à consulter les *What Works Clearinghouse : Procedures Handbook* et *What Works Clearinghouse : Standards Handbook* du WWC.

Nous ne parlerons pas de la première étape suivie par les méta-analyses ici (figure 2 p.9). On va donc supposer qu'un protocole de recherche a été défini, qu'une recherche bibliographique exhaustive a été menée, et que des études ont été sélectionnées en s'appuyant sur des normes explicitées (figure 39). C'est la deuxième étape qui va nous intéresser, car c'est elle qui va exiger un traitement statistique des données qui, à partir de maintenant, seront les résultats des études primaires sélectionnées.

^a www.meta-analysis.com

^b Ces documents ont subi depuis 2002 six mises à jour qui témoignent de l'évolution constante de la recherche dans ce domaine.

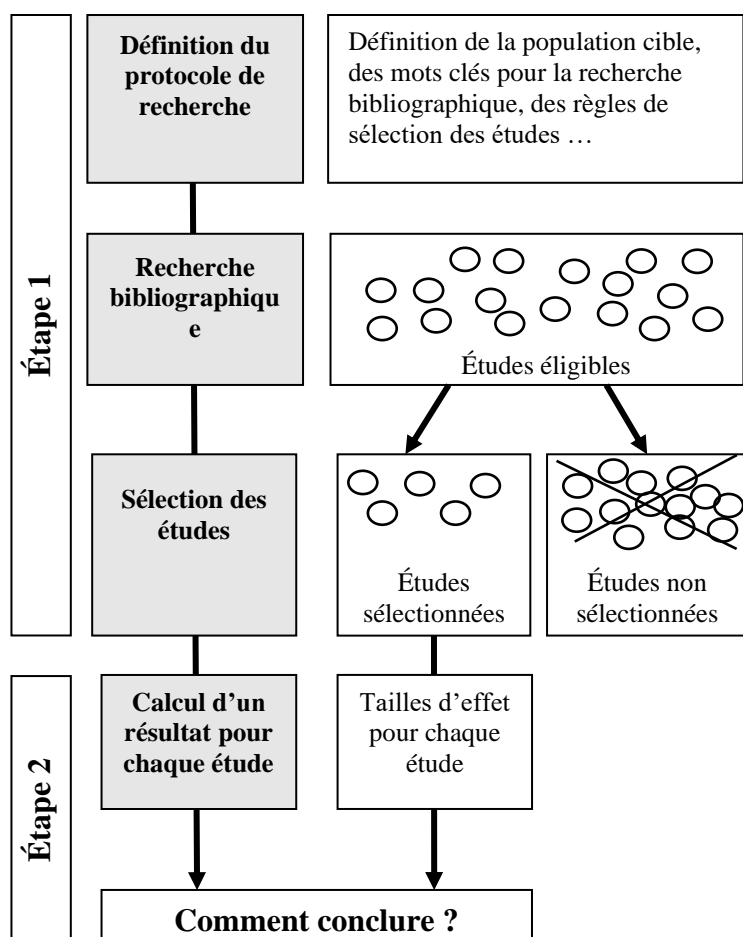


Figure 39 : étapes suivies pour mener une méta-analyse

Dans ce qui suit, les études sélectionnées sont des études par comparaison de groupes^a qui s'appuient toutes sur des concepts communs aux statistiques inférentielles mais qui ont pu appliquer des méthodes statistiques différentes pour répondre à leur question de recherche : test Z , test T , ANOVA, calcul d'une valeur- p , d'une différence des moyennes brutes (avec des échelles différentes) ou d'une taille d'effet (d de Cohen, g de Hedges, Δ de Glass) en sont quelques unes. Pour comparer puis synthétiser tous ces résultats, il est indispensable dans un premier temps de produire pour chacune des études un indicateur commun qui sera à chaque fois calculé exactement de la même façon : c'est la taille d'effet qui va jouer ce rôle. Et c'est sans doute la demande pressante et récente des autorités publiques dans certains pays (aux USA et au Royaume-Uni^b notamment) pour disposer de résultats concrets, de bonne qualité mais surtout comparables et cumulables, qui a donné au travail des statisticiens une importance qui impose que nous soyons capables de mieux en comprendre les principes. Les procédures et les formules statistiques générales mises en œuvre pour réaliser une méta-analyse seront l'objet de

^a D'autres design sont parfois considérés par les méta-analystes, comme les études de cas (WWC 2021).

^b Voir le projet *Education Endowment Foundation* <https://educationendowmentfoundation.org.uk/>

ce chapitre ; quelques unes de leurs utilisations pratiques seront commentées au chapitre suivant.

La première étape du statisticien qui réalise une méta-analyse sera donc de calculer pour chacune des études sélectionnées une taille d'effet à partir des données publiées dans un article scientifique^a, en utilisant la même méthode de calcul. Si une taille d'effet a déjà été calculée par les auteurs, elle sera recalculée par le méta-analyste. Il arrive parfois que les auteurs des études primaires soient invités à fournir des éléments indispensables à ce calcul et qui n'étaient pas publiés, ou qu'une étude soit rejetée par manque d'information. Ces tailles d'effet sont des estimations de tailles d'effet réelles (qui seraient celles calculées sur les populations entières). Mais pour alléger le texte, on parlera ici de taille d'effet en omettant volontairement « estimation de ». Dans le même ordre d'idée, quand la variance d'une taille d'effet est calculée, on considère alors cette taille d'effet comme un estimateur de la taille d'effet de la population (ici aussi on parlera plus volontiers d'une distribution des estimations que d'une variable aléatoire).

L'objectif poursuivi par le méta-analyste est de même nature que l'objectif poursuivi par l'auteur d'études primaires : calculer un indicateur de position centrale qui est une estimation d'un paramètre de la population (c'était par exemple la moyenne des scores pour les études primaires) et évaluer la dispersion des données autour de cet indicateur (comme l'écart-type des scores ou l'intervalle de confiance de leur moyenne ou l'intervalle de prédiction pour le premier type d'études). Certains des éléments caractéristiques des recherches primaires sélectionnées vont ensuite orienter le méta-analyste vers l'un des deux modèles suivants : le modèle de l'effet fixe ou le modèle des effets aléatoires. Ces deux modèles sont aujourd'hui utilisés dans presque toutes les méta-analyses, même si d'autres peuvent encore prévaloir comme une analyse de type narratif ou la méthode par comptage sur lesquelles nous reviendrons plus loin.

Dans ce qui suit, on va se placer dans le cas où k études partageant des caractéristiques communes (comme le traitement étudié et les compétences testées) ont été menées qui ont abouties pour chaque étude i (compris entre 1 et k) au calcul d'une taille de l'effet qui sera ici le g de Hedges (noté g_i pour l'étude i) de leur variance notée var_{g_i} et de leur écart-type s_{g_i} . On note n_{ai} et n_{bi} les tailles d'échantillons des groupes traitement et contrôle et on définit N_i la taille totale de l'échantillon de l'étude i ; on a donc $N_i = n_{ai} + n_{bi}$.

Le modèle de l'effet fixe

Ce modèle sera suivi quand les études dont dispose le méta-analyste ont toutes été menées dans des conditions similaires, sur des populations identiques ayant subi un même traitement. Ces études peuvent être considérées comme des répétitions d'un même protocole qui permet d'estimer un seul et unique effet du traitement réel (l'effet du traitement sur toute la population dont sont issus les échantillons). Les différences observées entre leurs résultats sont donc le fruit du hasard qui prévaut dans tout échantillonnage. Les tailles d'effet calculées par ces études sont alors des estimations d'une seule et unique taille d'effet réelle δ , et on va considérer

^a La véritable matière première du méta-analyste est une pile d'articles dans lesquels il lui faudra trouver des données numériques pertinentes.

qu'elles sont normalement distribuées autour de cette unique taille d'effet réelle. On calcule alors une taille d'effet globale^a, notée M , qui est une estimation de la taille d'effet réelle unique δ associée à une population définie par les études primaires (c'est cette même population qui est analysée dans chacune des études). Ce modèle ne semble pas correspondre à notre situation (et on verra plus loin que c'est le modèle des effets aléatoires qui sera le plus adapté), mais nous devons dans un premier temps comprendre comment mettre en œuvre le modèle de l'effet fixe avant de pouvoir utiliser le second modèle. De la même façon que pour une étude primaire, nous déterminerons enfin la variance de cette taille d'effet globale (notée var_M) pour définir un intervalle de confiance ou procéder à un test d'hypothèse.

La taille d'effet globale M calculée dans le modèle de l'effet fixe est égale à la moyenne des tailles d'effets calculées pour chacune des études, moyenne pondérée par l'inverse de leurs variances. On admettra ici que ce calcul permet d'obtenir un estimateur de variance minimale. Le poids affecté à chaque étude sera noté P_i . On a donc

$$M = \frac{\sum_i P_i \times g_i}{\sum_i P_i} \quad P_i = \frac{1}{var_{g_i}} = \frac{1}{s_i^2} \quad var_{g_i} = \frac{n_{ai} + n_{bi}}{n_{ai} \times n_{bi}} \times \omega^2 + \frac{g_i^2}{2(n_{ai} + n_{bi})}$$

Plus la variance de la taille d'effet de l'étude i est faible, plus son poids dans le calcul de la taille d'effet globale est important.

On peut montrer que pour une taille totale N_i fixée, le poids de la taille d'effet calculé dans une étude est maximal quand les tailles des deux échantillons sont égales. Et que le poids de la taille d'effet de l'étude augmente avec la taille totale (voir figure 40 ci-dessous).

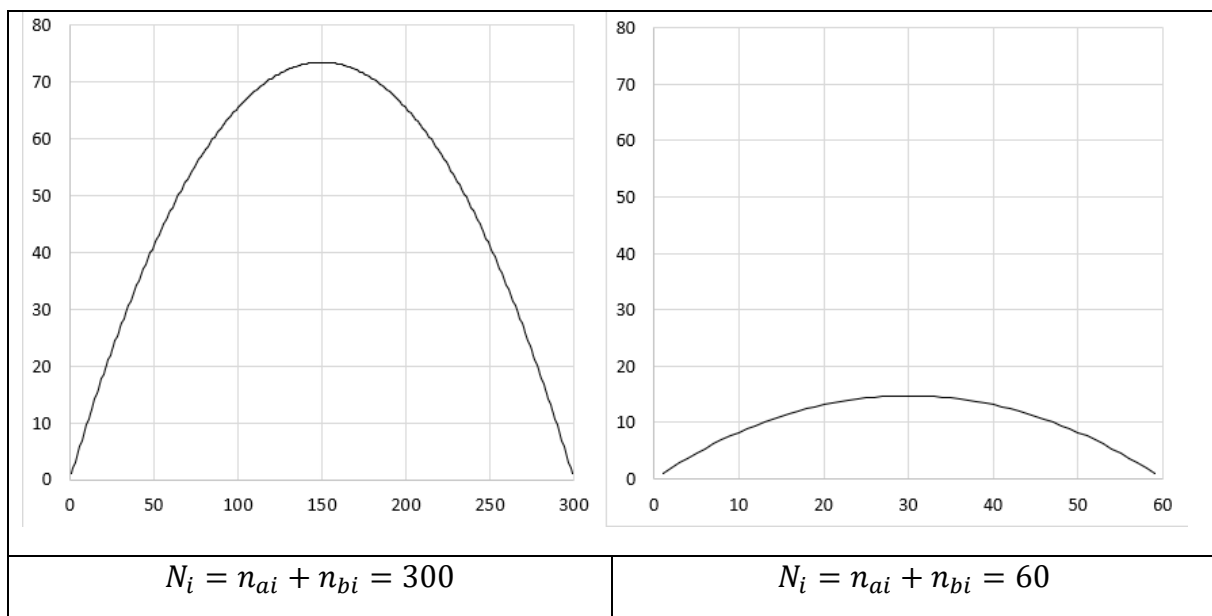


Figure 40 : variation de P_i en fonction de n_{ai} avec N_i fixé (ici $g = 0,4$)

On calcule^b alors la variance de M (notée var_M) et son écart-type (noté s_M) :

^a Ce n'est pas une moyenne arithmétique (voir plus loin), et c'est pour cette raison que le terme « globale » a été préféré ici.

^b Les propriétés de calculs des variances sont rappelées en annexe 5.

$$var_M = \frac{1}{\sum_i P_i} \quad ; \quad s_M = \sqrt{var_M} = \frac{1}{\sqrt{\sum_i P_i}}$$

On va admettre que la statistique^a M est normalement distribuée. On pourra donc mener un test Z (voir annexe 4) et déterminer si M est statistiquement et significativement différente de zéro en ayant fixé un risque α . L'hypothèse nulle est ici que « la taille d'effet réelle est nulle » ou encore $\delta = 0$. On calcule la valeur observée z_{obs} à laquelle on associe une valeur $-p$.

$$z_{obs} = \frac{M}{s_M} = \frac{\sum_i P_i \times g_i}{\sqrt{\sum_i P_i}}$$

On peut également définir un intervalle de confiance à 95 % de la manière suivante :

$$[M - 1,96 \times s_M; M + 1,96 \times s_M]$$

Si cet intervalle comprend la valeur zéro, on conclura à une absence de signification statistique au niveau de confiance 0,95.

Le modèle des effets aléatoires

Ce modèle sera choisi quand les études primaires sélectionnées ont été menées sur des populations différentes (par exemple d'âges différents ou de pays différents) en appliquant des protocoles expérimentaux divers (par exemple les durées du traitement ne sont pas les mêmes, les designs sont différents). La plupart du temps, les études primaires rassemblées lors de méta-analyses réalisées dans le domaine des sciences de l'éducation correspondent bien à cette description, et c'est donc ce modèle qui sera le plus souvent choisi. Dans ce cas, les différences entre les résultats expérimentaux obtenus d'une étude à l'autre sont dues comme toujours à l'échantillonnage mais aussi aux caractéristiques diverses évoquées ci-dessus. On va donc considérer que la taille d'effet calculée pour chaque étude i estime une taille d'effet réelle unique (que l'on notera δ_i) associée aux caractéristiques particulières de cette étude et que ces tailles d'effet réelles δ_i sont normalement distribuées autour d'une taille d'effet moyenne réelle δ , avec un écart-type que l'on notera τ . C'est cette taille d'effet moyenne réelle δ qui est le paramètre que l'on cherche à estimer.

Ce sont donc deux distributions que nous devons considérer :

1. la distribution normale des tailles d'effet calculées^b pour une étude donnée autour de la taille d'effet réelle de cette étude avec une erreur standard qui est aussi la racine carrée de la variance intra-étude et que nous savons estimer (voir chapitre 8) ;
2. la distribution normale des tailles d'effet réelles de toutes les études autour de leur moyenne δ avec un écart-type τ , moyenne et écart-type qui est aussi la racine carrée de la variance inter-étude τ^2 que nous cherchons à estimer. Nous noterons ici M^* et T ces estimations.

Les tailles d'effets réelles des études seront estimées par le g de Hedges. On peut déjà dresser le bilan suivant (tableau 18) concernant les notations et le vocabulaire utilisés :

^a On considère ici la distribution des estimations de la taille d'effet réelle si un nombre infini d'expériences identiques étaient menées.

^b Bien sûr une seule taille d'effet a été calculée ; le pluriel ici fait allusion à une distribution des tailles d'effets qui seraient calculées à partir de plusieurs échantillons.

Pour chaque étude i	
Taille d'effet réelle	δ_i
Estimation de la taille d'effet réelle	g_i
Estimation de la variance intra-étude	var_{g_i}

Pour l'ensemble des études	
Moyenne des tailles d'effet réelles	δ
Estimation de la taille d'effet réelle moyenne	M^*
Variance inter-étude	τ^2
Estimation de la variance inter étude	T^2

Tableau 18 : notations et vocabulaires (modèles des effets aléatoire et de l'effet fixe)

On peut retenir qu'il n'existe dans ce modèle pas une seule et unique vraie valeur de la taille d'effet, contrairement au modèle de l'effet fixe (en reprenant les notations ci-dessus, dans le modèle de l'effet fixe on a $\delta_1 = \delta_2 \dots = \delta$ et $\tau^2 = 0$), et que M^* la taille d'effet globale calculée sera donc l'estimation de la moyenne des tailles d'effets réelles. Comme les populations analysées par les études primaires sont différentes, et que ces différences sont prises en compte dans ce modèle, alors il sera possible d'inférer au-delà de ces populations que l'on considère ici comme des échantillons d'un ensemble plus large (ce qui n'était pas possible avec le modèle de l'effet fixe précédemment étudié).

Au niveau des calculs, les principes sont les mêmes que ceux mis en œuvre dans le modèle de l'effet fixe et l'estimation de la taille d'effet réelle moyenne est toujours égale la moyenne des tailles d'effet expérimentales pondérée par l'inverse de leurs variances. Il faudra en fait rajouter un terme à la variance intra-étude des tailles d'effet calculées dans chaque étude, ce terme étant T^2 , l'estimation de la variance inter-études des tailles d'effets réelles τ^2 . On va noter $var_{g_i}^*$ la variance de la taille d'effet calculée pour l'étude i dans ce modèle ; cette variance est donc la somme de la variance intra-étude (que nous avons notée var_{g_i} et qui a été calculée auparavant au chapitre 8 et utilisée dans le modèle de l'effet fixe), et de l'estimation de la variance intra-étude T^2 . La taille d'effet globale calculée dans ce modèle est notée M^* , et on conservera la notation M pour la taille d'effet globale calculée dans le modèle de l'effet fixe. Enfin, nous noterons^a var_{M^*} et s_{M^*} la variance et l'écart-type de M^* . Les tailles d'effets calculées pour chaque étude sont les g de Hedges^b. On a donc

$$var_{g_i}^* = var_{g_i} + T^2 \quad ; \quad P_i^* = \frac{1}{var_{g_i}^*} \quad ; \quad M^* = \frac{\sum_i P_i^* \times g_i}{\sum_i P_i^*}$$

$$var_{M^*} = \frac{1}{\sum_i P_i^*} \quad ; \quad s_{M^*} = \frac{1}{\sqrt{\sum_i P_i^*}}$$

^a Ce sont les notations utilisées par BORENSTEIN (2009) qui seront employées ici, l'étoile (*) indiquant que les calculs concernent le modèle des effets aléatoires.

^b Le calcul de la variance des tailles d'effet $var_{g_i}^*$ n'est valable que pour le g de Hedges ; dans le cas où un estimateur biaisé des tailles d'effet est utilisé, la covariance des variances intra et inter-étude n'est pas nulle (HEDGES, 1983).

De la même façon que pour le modèle précédent, en considérant que M^* est normalement distribuée, on procède à un test Z avec comme hypothèse nulle que « la taille d'effet moyenne de la population est égale à zéro » et en associant à ce test une valeur- p avec

$$Z_{obs} = \frac{M^*}{s_{M^*}}$$

On calcule également un intervalle de confiance au niveau 0,95
 $[M^* - 1,96 \times s_{M^*} ; M^* + 1,96 \times s_{M^*}]$

Il va donc falloir calculer T^2 , l'estimation de τ^2 . Ce calcul va se faire en plusieurs étapes. On utilisera des résultats calculés précédemment dans le modèle de l'effet fixe (donc sans étoile). On calcule d'abord la somme des carrés des écarts des tailles d'effets de chaque étude à la taille d'effet globale calculée dans le modèle de l'effet fixe divisés par l'écart-type intra-étude. On va noter Q ce résultat.

$$Q = \sum_i \left[\frac{g_i - M}{s_{g_i}} \right]^2 = \sum_i P_i \times g_i^2 - \frac{[\sum_i P_i \times g_i]^2}{\sum_i P_i}$$

BORENSTEIN (2009) affirme que si les tailles d'effets réelles de chaque étude étaient les mêmes, cette quantité serait égale au degré de liberté du nombre d'étude (donc égale à $k - 1$). La quantité $Q - (k - 1)$ reflète donc la quantité de variation ajoutée à la variation attendue quand toutes les tailles d'effets réelles sont les mêmes ; c'est en fait la variation supplémentaire que l'on peut attribuer aux variations entre les tailles d'effet réelles. Pour obtenir une estimation de la variance inter-étude T^2 , on divise cette différence par un nombre, noté C , afin d'obtenir une moyenne dans la même unité (élevée au carré) que les tailles d'effet^a avec

$$C = \sum_i P_i - \frac{\sum_i P_i^2}{\sum_i P_i}$$

On a donc finalement

$$T^2 = \frac{Q - (k - 1)}{C}$$

Quand ce calcul aboutit à un résultat négatif, on attribuera à T^2 la valeur nulle (τ^2 ne peut pas être négative).

Pour pouvoir estimer T^2 avec assez de précision un nombre suffisant^b d'études est requis. Si ce nombre est trop faible, il ne sera pas possible de conclure à partir de l'intervalle de confiance ; le modèle de l'effet fixe pourra alors être utilisé, mais ne permettra pas d'inférer au-delà des populations analysées dans les études primaires sélectionnées.

^a On peut remarquer en effet que Q est une somme dépendante du nombre d'étude et n'a pas d'unité.

^b Cette remarque de BORENSTEIN n'est accompagnée d'aucune valeur numérique ; il paraît raisonnable de fixer ce seuil à au moins 5 études. Une analyse de la puissance peut alors être menée, voir par exemple PIGOTT (p.53).

Modèle de l'effet fixe et modèle des effets aléatoires : le bilan

Le tableau ci-dessous (tableau 19) et le schéma qui l'accompagne (figure 41) résument les aspects principaux de chacun de ces deux modèles pour lesquels un bilan est également dressé dans l'annexe 3.

	Modèle de l'effet fixe	Modèle des effets aléatoires
Point de départ	Les études sont semblables (répétition d'un même protocole sur une même population)	Les études sont différentes (traitements, procédures, populations différents)
Taille d'effet	On a une seule et unique taille d'effet réelle commune à toutes les études : $\delta_1 = \delta_2 \dots = \delta$	Les tailles d'effet réelles δ_i associées à chaque étude sont normalement distribuées autour de leur moyenne δ .
Variances des tailles d'effet expérimentales	La variance utilisée pour calculer la taille d'effet globale M est la variance intra-étude, var_{g_i}	La variance utilisée pour calculer la taille d'effet globale M^* est la somme de la variance intra-étude et de la variance inter-études, $var_{g_i}^* = var_{g_i} + T^2$
Conclusion	On ne peut inférer au-delà de la population étudiée	On peut inférer au-delà des populations étudiées.

Tableau 19 : bilan sur les modèles des effets aléatoires et de l'effet fixe.

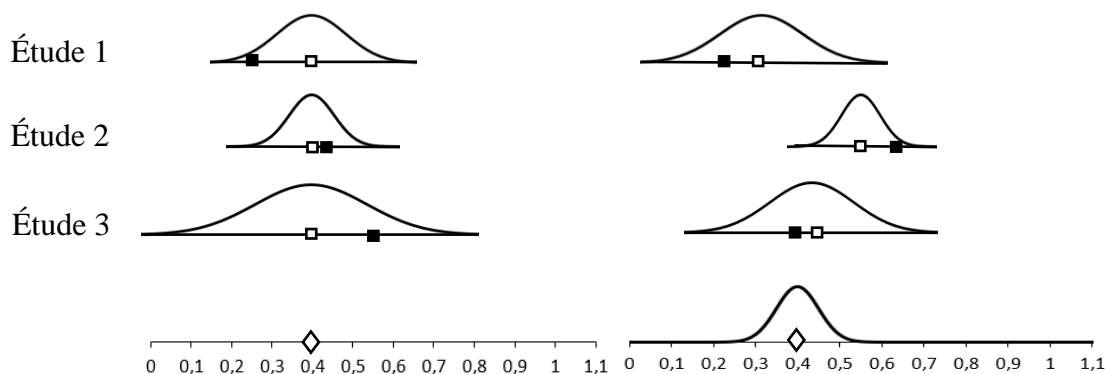


Figure 41 : à gauche le modèle de l'effet fixe, à droite le modèle des effets aléatoires.

◇: taille d'effet réelle moyenne ; □: taille d'effet réelle de l'étude ; ■: taille d'effet expérimentale

Un exemple numérique

Nous poursuivons avec l'**exemple 7**. Cette fois, ce sont les données fictives de 6 études qui vont être utilisées pour calculer les résultats obtenus en suivant le modèle de l'effet fixe et le modèle des effets aléatoires. Les tailles d'effet calculées sont les g de Hedges (voir chapitre 8 pour le calcul de g pour les études A et B et le calcul de la variance de g notée var_g).

Les données de départ

Nom de l'étude	n_a	n_b	g	var_g	s_g
A	65	65	0,277	0,031	0,175

B	200	200	0,664	0,011	0,103
C	60	60	0,095	0,033	0,182
D	40	40	0,367	0,050	0,223
E	50	45	0,462	0,043	0,207
F	85	85	0,185	0,023	0,153

Modèle de l'effet fixe

Nom de l'étude	Poids ($P = \frac{1}{var_g}$)
A	32,572
B	95,129
C	30,352
D	20,055
E	23,449
F	42,700

$$M = \frac{32,572 \times 0,277 + 95,129 \times 0,664 + \dots + 42,700 \times 0,185}{32,572 + 95,129 + \dots + 42,700} \approx 0,414$$

$$var_M = \frac{1}{32,572 + 95,129 + \dots + 42,700} \approx 0,004$$

$$\frac{M}{s_M} = \frac{0,414}{\sqrt{0,004}} \approx 6,475$$

On calcule alors une valeur-p égale à 0,0000. La taille d'effet globale calculée est statistiquement significativement différente de 0 (test Z avec $\alpha=0,05$). L'intervalle de confiance au seuil de 0,95 est [0,289 ; 0,540] qui n'inclue pas la valeur 0.

Modèle des effets aléatoires

On a besoin de calculer la variance intra-étude T^2 . On a $k = 6$, donc $k - 1 = 5$

$$Q = \left[\frac{0,277 - 0,414}{0,175} \right]^2 + \left[\frac{0,664 - 0,414}{0,103} \right]^2 + \dots + \left[\frac{0,185 - 0,414}{0,153} \right]^2 \approx 12,005$$

$$C = 244,256 - \frac{32,572^2 + 95,129^2 + \dots + 42,700^2}{244,256} \approx 187,729$$

Et donc

$$T^2 = \frac{12,005 - 5}{187,729} \approx 0,037$$

On calcule alors les variances des tailles d'effet en rajoutant à la variance intra-étude la variance inter-étude T^2 . Par exemple pour l'étude A on a

$$var_g^* = 0,031 + 0,037 = 0,068$$

On calcule également les nouveaux poids.

Nom de l'étude	var_g^*	poids*($P^* = \frac{1}{var_g^*}$)
A	0,068	14,703
B	0,048	20,910
C	0,070	14,233
D	0,087	11,471
E	0,080	12,506
F	0,061	16,466

On procède de la même façon que pour le modèle de l'effet fixe pour calculer ensuite les autres résultats.

$$M^* = \frac{14,703 \times 0,277 + 20,910 \times 0,664 + \dots + 16,466 \times 0,185}{14,703 + 20,910 + \dots + 16,466} \approx 0,358$$

$$var_M^* = \frac{1}{14,703 + 20,910 + \dots + 16,466} \approx 0,011$$

$$\frac{M^*}{s_M^*} = \frac{0,358}{\sqrt{0,011}} \approx 3,404$$

On calcule alors une valeur-p égale à 0,0007. La taille d'effet globale calculée est statistiquement significativement différente de 0 (test Z avec $\alpha=0,05$). L'intervalle de confiance au seuil de 0,95 est [0,152 ; 0,565] qui n'inclue pas la valeur 0.

Les intervalles de confiance^a des tailles d'effet de chaque étude et des tailles d'effets globales calculées pour les deux modèles (effet fixe et effets aléatoires) sont représentés ci-dessous (figure 42) :

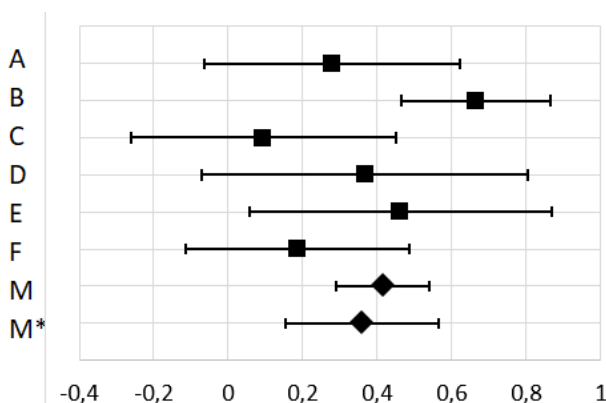


Figure 42 : intervalles de confiance des tailles d'effet des études et des tailles d'effets globales (effet fixe en haut, effets aléatoires en bas)

^a Ce sont les variances var_g qui doivent être utilisées ici, et non var_g^* , quel que soit le modèle utilisé.

Intervalle de prédiction

Nous avons déjà calculé les bornes d'un intervalle de prédiction dans le cas d'une étude primaire. La même procédure peut être suivie pour les méta-analyses qui suivent le modèle des effets aléatoires. Il s'agit alors de proposer un intervalle dans lequel on a une certaine probabilité (définie à l'avance) de trouver une taille d'effet si une expérience supplémentaire était menée. Et si on a noté k le nombre d'études utilisées pour calculer la taille d'effet globale M^* , on définira cet intervalle de prédiction comme devant contenir, par exemple dans 95 cas sur 100, la taille d'effet réelle pour une $k + 1^{\text{ème}}$ étude. Le degré de liberté est ici égal à $k - 2$.

Par analogie avec l'intervalle de prédiction calculé au chapitre 6, on a ici l'intervalle de prédiction suivant ($\alpha=0,05$)

$$\left[M^* - t_{0,025; k-2} \times \sqrt{T^2 + var_{M^*}}; M^* + t_{0,025; k-2} \times \sqrt{T^2 + var_{M^*}} \right]$$

Si l'intervalle de confiance construit autour de la taille d'effet globale M^* quantifie la précision avec laquelle celle-ci estime la taille d'effet réelle δ (dans 95 cas sur 100, elle est incluse dans cet intervalle), l'intervalle de prédiction quantifie lui la distribution des tailles d'effets réelles autour de la taille d'effet réelle moyenne (dans 95 cas sur 100, la taille d'effet réelle d'une nouvelle étude est incluse dans cet intervalle).

On reprend notre **exemple 7**, pour calculer l'intervalle de prédiction suivant (avec $t_{0,025; 4} = 2,776$) :

$$\left[0,358 - 2,776 \times \sqrt{0,037^2 + 0,011}; 0,358 + 2,776 \times \sqrt{0,037^2 + 0,011} \right] = [-0,253; 0,969]$$

Les intervalles de confiance et de prédiction sont représentés figure 43.

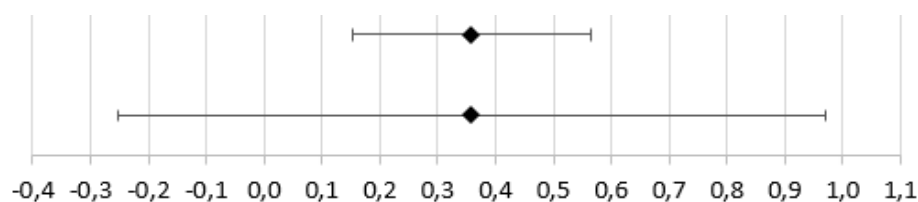


Figure 43 : intervalles de confiance (en haut) et de prédiction (en bas) pour l'**exemple 7**.

Hétérogénéité

Au-delà de cette première description menée dans le cadre du modèle des effets aléatoires, l'objectif du chercheur va être de quantifier l'hétérogénéité des tailles d'effets réelles. Les variations observées expérimentalement sont dues à l'échantillonnage et aux variations entre les tailles d'effets réelles des études. Ce sont ces dernières qui nous intéressent, et pour en prendre la mesure, plusieurs choix sont possibles.

L'une des options consiste à mener un test d'hypothèse sur la statistique Q . Sous l'hypothèse nulle « les tailles d'effets réelles sont toutes égales », cette statistique suit une loi du Khi-2 avec

un degré de liberté égal à $k - 1$. On calcule alors une valeur- p associée à un risque α déterminé. En fonction des résultats, on conclura que l'hétérogénéité est (ou non) statistiquement significative. Pour l'**exemple 7**, on calcule une valeur- p égale à 0,03472, qui est inférieure à 0,05, l'hypothèse nulle est donc rejetée et l'hypothèse alternative, « les tailles d'effet réelles sont différentes » est acceptée. La valeur de Q (comme la valeur- p) sont dépendants du nombre d'études et ne dépendent pas de l'unité des scores. Ce test d'hypothèse ne devrait pas être utilisé pour choisir entre le modèle à effet fixe et le modèle des effets aléatoires : ce choix dépend des caractéristiques des études incluses dans la méta-analyse.

Il est également possible d'utiliser l'estimation de la variation inter-étude T^2 , ou l'écart-type T . Pour notre exemple 7, on a $T \approx 0,192$. On peut aussi calculer l'indicateur I^2 qui évalue la variation inter-étude comme une proportion de la variation totale

$$I^2 = \frac{Q - (k - 1)}{Q}$$

Si I^2 est proche de 0%, on dira que la variance calculée est due à l'échantillonnage ; plus I^2 se rapproche de 100%, plus les variations des tailles d'effets réelles prennent de l'importance. On considère que 25% représente une petite part, 50% une part modérée et 75 % une grande part. Pour notre **exemple 7**, on a

$$I^2 = \frac{12,005 - (6 - 1)}{12,005} \approx 58,3 \%$$

On dira donc qu'une part modérée de la variance est due à la variation des tailles d'effets réelles.

Admettons maintenant que l'hétérogénéité évaluée semble importante. La prochaine étape sera alors d'identifier les causes de cette hétérogénéité, et une des façons de procéder consiste à comparer des sous-groupes d'études entre eux, ce qui va consister à inclure dans l'analyse une troisième variable, ou encore un modérateur.

Analyse de sous-groupes d'études

Les études incluses dans la méta-analyse vont être partagées en plusieurs sous-groupes. Pour constituer ces sous-groupes, le chercheur va prendre en compte des facteurs qui semblent avoir une influence sur l'hétérogénéité des résultats. Ces facteurs peuvent être liés à l'intervention, comme la population ciblée (le niveau d'étude ou l'âge par exemple) ou la durée du traitement, mais ils peuvent aussi être liés à la méthodologie mise en œuvre, comme la qualité du design des études (essais contrôlés randomisés, études quasi-expérimentales, ...) ou la taille de l'échantillon. Cela revient donc à considérer une troisième variable (qualitative ou quantitative) qui joue un rôle dans l'influence de la variable indépendante (comme l'application d'une méthode d'enseignement) sur la variable dépendante (comme le niveau des élèves dans une certaine discipline). Nous laisserons de côté ici la discussion sur la définition de ces variables modératrices (on parle aussi de modérateur) pour nous concentrer, comme d'habitude, sur l'analyse statistique des données. Cette analyse de sous-groupes peut être menée dans le cadre des deux modèles que nous venons d'étudier, le modèle de l'effet fixe et le modèle des effets aléatoires. Comme c'est ce dernier qui semble convenir au domaine des Sciences de l'éducation, nous laisserons de côté ici le modèle de l'effet fixe. Pour fixer les idées, on va

considérer deux groupes d'études, le groupe A et le groupe B (les calculs pour plus de deux groupes peuvent être déduits de ceux que nous allons proposer assez simplement).

Dans le modèle des effets aléatoires, on a calculé pour l'ensemble des études une estimation de la variance inter-étude (variance entre les tailles d'effets réelles δ_i). De la même façon, dans chacun des groupes, on va estimer une variance inter-études, que l'on notera T_A^2 et T_B^2 . Ces variances seront dénommées maintenant variances *intra-groupe* (elles sont internes à chaque groupe). Deux possibilités s'offrent alors au chercheur : soit il considère que chaque variance intra-groupe est propre à son groupe, et c'est cette variance qui sera utilisée dans les calculs ; ou bien il considère que ces variances sont des estimations d'une seule et unique variance, et dans ce cas on les regroupe en une seule variance. C'est cette option qui a été retenue ici (ce choix, comme celui qui doit être fait quand on sélectionne un modèle dépend du contexte de la recherche^a), et on notera T^2 cette variance groupée. On a :

$$T^2 = \frac{Q_A + Q_B - (k_A - 1 + k_B - 1)}{C_A + C_B}$$

Trois procédures peuvent ensuite être suivies, qui sont en fait équivalentes (et le choix pour l'une ou l'autre de ces procédures sera fait en fonction des éléments dont on dispose).

Première méthode : test-Z

On procède de la même manière que dans une étude primaire, en comparant des moyennes (ces moyennes ici ne concernent pas des scores mais des tailles d'effet). Sous l'hypothèse nulle « la valeur réelle de la moyenne des tailles d'effet est la même dans les deux groupes », la statistique Z_{diff}^* (voir ci-dessous) suit une loi normale centrée réduite. On calcule alors une valeur $-p$ que l'on compare au risque α choisi.

$$Z_{diff}^* = \frac{M_B^* - M_A^*}{\sqrt{var_{M_B^*} + var_{M_A^*}}}$$

Deuxième méthode : analyse de la variance

On procède de la même manière que dans une étude primaire, en partageant la variance totale (variation de toutes les études autour de la moyenne générale) en une variance intra (variation des études autour des moyennes de leur groupe respectif) et une variance inter (variation de la moyenne des groupes autour de la moyenne générale). Ici, la variance totale va être quantifiée par la statistique Q^* : ce sont les poids calculés en suivant le modèle des effets aléatoires (après avoir ajouté la variance inter-études aux variances intra) qui sont utilisés dans la formule. La variance intra notée Q_{intra}^* est égale à la somme des variances de chaque groupe, donc $Q_{intra}^* = Q_A^* + Q_B^*$ (là aussi, ce sont les poids calculés en suivant le modèle des effets aléatoires qui sont utilisés). La variance inter est notée Q_{inter}^* . C'est elle qui intéresse le chercheur puisque son hypothèse est que les tailles d'effet diffèrent selon que les études appartiennent à tel ou tel groupe. On a

^a Vous trouverez dans *Introduction to Meta-analysis* les procédures à suivre dans les autres cas.

$$Q_{inter}^* = Q^* - Q_{intra}^* = Q^* - (Q_A^* + Q_B^*)$$

Sous l'hypothèse nulle « les tailles d'effets moyennes réelles des deux groupes sont égales », la statistique Q_{inter}^* suit une loi du Khi 2 avec un degré de liberté égal à $2 - 1 = 1$ (le nombre de groupes diminué de 1). On calcule ensuite une valeur $-p$ que l'on compare au risque α choisi.

Troisième méthode : test Q

On considère les tailles d'effet globales calculées pour chaque groupe comme si elles provenaient d'études individuelles. On procède alors à un test d'hétérogénéité en calculant la statistique Q , comme on le ferait pour une méta-analyse ne comportant que deux études primaires. Sous l'hypothèse nulle « les tailles d'effets moyennes réelles des deux études sont égales », Q suit une loi du Khi 2 avec comme degré de liberté $2 - 1 = 1$ (le nombre de groupes diminué de 1). On calcule ensuite une valeur $-p$, que l'on compare au risque α choisi.

On peut remarquer que

- les trois méthodes sont équivalentes.
- les deux dernières peuvent s'appliquer à plus de deux sous-groupes d'études.
- cette analyse de sous-groupes requiert un nombre suffisant d'études dans les groupes (au moins 5 selon BORENSTEIN).

Un exemple numérique

On termine ici l'exploitation des données fictives de l'**exemple 7**. On va considérer que les études A à F (dont nous avons déjà exploité les données, voir ci-dessus) constituent le groupe A ; et le groupe B sera constitué de cinq nouvelles études G à K.

Les données de départ

Groupe A		
Nom de l'étude	g	var_g
A	0,277	0,031
B	0,664	0,011
C	0,095	0,033
D	0,367	0,050
E	0,462	0,043
F	0,185	0,023

Groupe B		
Nom de l'étude	g	var_g
G	0,440	0,015
H	0,492	0,020
I	0,651	0,015
J	0,710	0,025
K	0,740	0,012

Premiers calculs

On procède de la même façon pour le groupe B que pour le groupe A (voir ci-dessus), et on a finalement :

Groupe A			Groupe B		
k_A	Q_A	C_A	k_B	Q_B	C_B
6	12,005	187,729	5	4,543	241,667

La variance groupée T^2 est calculée de la façon suivante :

$$T^2 = \frac{12,005 + 4,543 - (6 - 1 + 5 - 1)}{187,729 + 241,667} \approx 0,0176$$

Elle permet alors de calculer les variances des tailles d'effet en prenant en compte les variations entre les études, puis de calculer les tailles d'effet globales M_A^* et M_B^* selon le modèle des effets aléatoires comme nous l'avons vu précédemment. Enfin les statistiques Q^* , Q_A^* et Q_B^* sont calculées. Les résultats sont rassemblés dans le tableau ci-dessous.

Groupe A			Groupe B		
Q_A^*	M_A^*	$var_{M_A^*}$	Q_B^*	M_B^*	$var_{M_B^*}$
5,738	0,371	0,008	2,115	0,608	0,007

Pour l'ensemble des 10 études, on calcule $Q^* = 11,737$

Première méthode

$$Z_{diff}^* = \frac{0,608 - 0,371}{\sqrt{0,005 + 0,005}} \approx 1,971$$

On calcule une valeur $-p$ égale à 0,049, la différence est statistiquement significative pour $\alpha = 0,05$.

Deuxième méthode

$$Q_{intra}^* = 5,738 + 2,115 = 7,853 \quad \text{et donc} \quad Q_{inter}^* = 11,737 - 7,853 = 3,884$$

On calcule alors une valeur $-p$ égale à 0,049, les tailles d'effets sont donc statistiquement significativement différentes entre les deux groupes A et B pour $\alpha = 0,05$.

Troisième méthode

$$Q = \frac{0,371^2}{0,008} + \frac{0,608^2}{0,007} - \frac{\frac{1}{0,008} \times 0,371 + \frac{1}{0,007} \times 0,608}{\frac{1}{0,008} + \frac{1}{0,007}} \approx 3,884$$

On remarque que ce résultat est le même que celui précédemment calculé, les conclusions sont donc les mêmes.

On vérifie également que $Z_{diff}^{*2} = Q = Q_{inter}^*$

Les autres méthodes

Nous verrons dans le chapitre suivant que d'autres méthodes ont été parfois utilisées dans les méta-analyses pour conclure quant à l'effet d'un traitement.

Entre autres, des tailles d'effet globales sont parfois calculées différemment (elles peuvent être des moyennes arithmétiques ou des moyennes pondérées par la taille de l'échantillon par exemple), et dans ces cas, la signification statistique de ces résultats ne peut pas être évaluée. Pour conclure, les auteurs procèdent alors à un décompte en comparant le nombre des études dont la taille d'effet est significativement supérieure à zéro et le nombre des études qui ont

trouvé un effet positif mais non significatif, voire même négatif (*vote counting*). Cette procédure interprète les résultats non significatifs comme concluant à une absence d'effet du traitement ; il est cependant possible que cette absence de significativité des résultats soit la conséquence d'un manque de puissance (et donc d'une taille d'échantillon trop faible) et cette procédure n'est pas recommandée.

Des revues de la littérature ou des synthèses bibliographiques s'appuient parfois sur des résultats calculés pour chaque étude analysée sans qu'une taille d'effet globale ne soit proposée et les résultats publiés pour chaque étude ne sont pas toujours accompagnés d'une estimation de leur significativité statistique. Mais quand le nombre des études devient important, ou que des facteurs particuliers associés à ces études (comme leur design, la durée du traitement, la population étudiée, ...) sont pris en compte, des calculs sont alors inévitablement menés, qui peuvent être de simple comparaison entre deux nombres, et des conclusions sont tirées. Quand ces dernières ne s'appuient que sur le jugement du chercheur en l'absence de toute règle de décision clairement énoncée, leur interprétation ne peut qu'être sujette à discussion.

Chapitre 11. Méta-analyse et enseignement des mathématiques

Dans ce chapitre, les méta-analyses conduites par le What Works Clearinghouse d'une part et Robert SLAVIN d'autre part, et qui concernent l'enseignement des mathématiques, vont nous servir d'exemples d'application des règles de calcul vu au chapitre précédent. Leurs évolutions dans le temps et leurs différences permettent de mieux rendre compte des différents paramètres que les chercheurs vont considérer afin de répondre à un même objectif : identifier les méthodes ou les facteurs qui semblent favoriser l'apprentissage des mathématiques en milieu scolaire ordinaire. Elles seront l'occasion de discuter aussi de la légitimité de certains calculs.

Les réponses du What Works Clearinghouse

Le What Works Clearinghouse (WWC) conduit des méta-analyses depuis le début des années 2000. Une description plus détaillée de leurs procédures est proposée dans *Comment enseigner les maths ? La réponse du What Works Clearinghouse* (Nathalie ROQUES, 2021). Les documents publiés par le WWC et qui décrivent très précisément leurs procédures^a constituent une mine d'informations à la fois théoriques et concrètes inestimable.

Les méta-analyses du WWC peuvent être classées dans deux catégories.

Dans la première, ce sont des interventions comme par exemple l'utilisation d'un nouveau logiciel informatique ou de manuels présentés comme innovants, qui sont évaluées. Chaque taille d'effet globale calculée par le WWC associe un domaine d'apprentissage (l'algèbre par exemple), un public particulier (les élèves scolarisés au lycée par exemple) et une intervention ciblée (les données des interventions Odyssey Math et Knowledge Is Power Program (KIPP) sont reprises dans l'**exemple 10**). Un *Rapport d'intervention* est alors publié qui rassemble les résultats calculés.

Les méta-analyses de la seconde catégorie ont un objectif différent : il s'agit ici d'identifier des éléments pédagogiques efficaces quand certaines acquisitions particulières sont attendues, pour aboutir à un ensemble de recommandations. Ces dernières sont alors publiées dans des *Guides des Pratiques*, dont les titres explicitent les objectifs poursuivis (par exemple *Improving mathematical problem solving in grades 4 through 8*).

Des différences notables entre ces deux catégories concernent pratiquement toutes les étapes d'une méta-analyse, mais comme d'habitude, nous nous intéresserons ici à l'étape 2, c'est-à-dire l'étape des calculs. Le point commun à toutes les méta-analyses du WWC sont l'utilisation du g de Hedges^b pour calculer la taille d'effet du traitement pour chacune des études primaires sélectionnées. C'est ensuite que les procédures vont diverger.

Pour les méta-analyses de la première catégorie et jusqu'en 2020, pour que l'effet de l'intervention soit caractérisé comme positif, au moins deux études devaient montrer des effets statistiquement significativement positifs (dont une devait être un essai contrôlé randomisé), et aucune autre étude ne devait avoir montré d'effet statistiquement significativement négatif (ou important, c'est-à-dire avec une taille d'effet inférieur à $-0,25$)^c. L'effet était considéré comme

^a Notamment dans leur *WWC Procedures Handbook 4.1* et *WWC Standards handbook 4.1* (2021) déjà plusieurs fois cités ici et disponibles sur leur site internet *Find What Works*.

^b Larry HEDGES est un expert qui participe de près à l'élaboration des procédures statistiques mises en œuvre par le WWC.

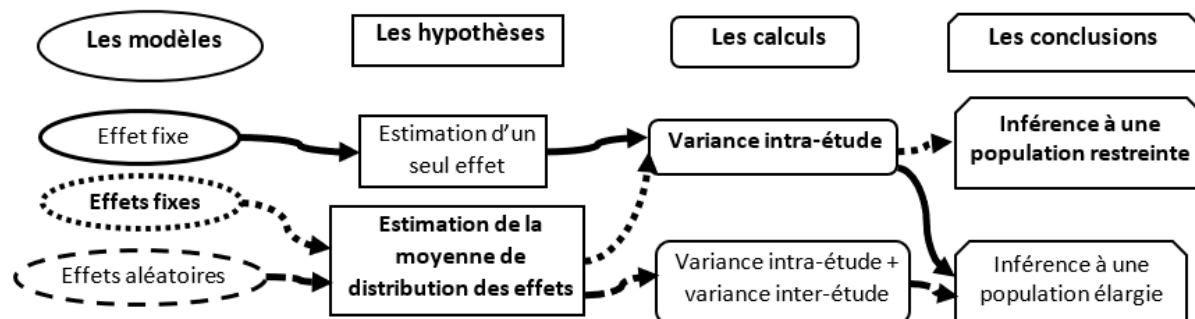
^c Figure IV.3 p.25 du *What Works Clearinghouse : Procedures Handbook* (version 4.0)

potentiellement positif quand au moins pour une étude la taille d'effet calculée était statistiquement significativement positive ou importante (taille d'effet supérieure ou égale à 0,25), que le nombre d'études montrant des effets indéterminés était inférieur ou égal au nombre des études montrant un effet positif et que pour aucune étude les tailles d'effets calculées n'étaient statistiquement significativement négatives ou importantes. Cette méthode qui s'apparente à un comptage (*vote counting*) a été critiquée car toutes les études avaient le même poids, et ce quelle que soit leur taille : une étude conduite sur un échantillon de 20 élèves avait autant d'importance qu'une étude conduite sur un échantillon de 2000 élèves (voir plus loin la discussion en fin de chapitre). Notons enfin qu'à cette efficacité était associé un niveau de preuve : pour obtenir un niveau de preuve considéré comme modéré à important, plusieurs terrains (des établissements dans des milieux différents) devaient être étudiés et la taille totale des échantillons devaient être supérieure à 350.

Ce classement était accompagné de la taille d'effet globale (notée *ES* pour *Effect Size* dans leur publication), égale dans ce cas à la moyenne arithmétique des tailles d'effet. Aucune variance n'est publiée. En reprenant les notations utilisées au chapitre précédent, on a donc pour *i* variant de 1 à *k* (*k* est le nombre des études incluses dans la méta-analyse) :

$$ES = \frac{\sum_i g_i}{k}$$

Depuis janvier 2020, le What Works Clearinghouse a changé ses procédures pour utiliser un modèle qu'ils décrivent comme le modèle des effets fixes^a. Dans ce modèle, on considère qu'il n'existe pas une seule et unique taille d'effet réelle au niveau de la population, mais plusieurs qui dépendent des populations analysées dans chaque étude, reprenant en cela l'hypothèse de départ du modèle des effets aléatoires. Mais pour des raisons techniques^b sans doute liées entre autres au nombre réduit d'études sélectionnées dans chacune de leur méta-analyse (les critères de sélection très restrictifs et le champ étroit de leurs investigations pouvant expliquer ce faible nombre), les calculs ne tiennent pas compte de la variance inter-étude (voir figure 44). Il ne sera donc possible de généraliser les conclusions qu'à une population similaire aux populations analysées par les études incluses dans la méta-analyse (et par-là même impossible d'élargir les conclusions au-delà de celles-ci).



^a Attention, ce n'est pas le modèle de l'effet fixe, le pluriel a ici son importance. Au moment de la rédaction de ce texte, aucune méta-analyse suivant ces nouvelles procédures n'avait été publiée.

^b Communication personnelle

Figure 44 : le modèle des effets fixes du WWC

Les résultats de deux méta-analyses menées par le WWC sont proposés dans l'**exemple 10**, avec une présentation du calcul de la taille d'effet d'une étude et de sa significativité statistique, la prise en compte de l'échantillonnage en classes et le calcul de la taille d'effet globale. La quantité de documents publiés par le WWC sur son site internet est une particularité qu'il convient de souligner ; et au-delà, leur volonté de mettre leurs procédures à jour et de communiquer avec les chercheurs et toute personne intéressée est à mettre à leur crédit. Ils utilisent des formules légèrement différentes de celles qui ont été proposées tout au long de ce livre, qui sont proposées en annexe 7.

Dans les *Guides des Pratiques*, le WWC se limite à donner pour chaque recommandation la liste des études sélectionnées accompagnées du calcul de leur taille d'effet et de leur significativité statistique (une valeur-*p* est calculée) : aucune taille d'effet globale n'est calculée et les auteurs se contentent de dénombrer les études qui montrent des effets statistiquement significativement positifs.

Les réponses de Robert SLAVIN

Deux méta-analyses menées par Robert SLAVIN concernent l'enseignement des mathématiques, la première publiée en deux parties en 2008 (la première partie concerne l'enseignement au primaire, la seconde partie porte sur l'enseignement au secondaire), la seconde publiée en 2011. La première méta-analyse avait comme objectif de déterminer les méthodes d'enseignement les plus efficaces. La seconde s'intéresse plus particulièrement à l'utilisation de logiciels informatiques par les élèves et analyse les facteurs susceptibles d'expliquer l'hétérogénéité des tailles d'effet réelles. Ces deux méta-analyses et d'autres documents les accompagnants sont publiés sur le site internet^a de Robert SLAVIN. Plus de détails concernant ces méta-analyses peuvent également être trouvés dans *Comment enseigner les maths ? La réponse de Robert SLAVIN* (Nathalie ROQUES, en cours de publication)

Dans la méta-analyse de 2008, la taille d'effet calculée pour chacune des études est le Δ de Glass et une estimation de leur significativité statistique est proposée. Les études qui partagent un élément commun (par exemple l'utilisation de logiciels informatiques par les élèves ou la mise en œuvre en classe d'un travail coopératif) sont ensuite regroupées et une taille d'effet globale est calculée. Dans la publication concernant les élèves scolarisés en primaire, cette taille d'effet globale (notée *ES* ici aussi) est la médiane des tailles d'effet du groupe d'études considéré. Dans la publication concernant les élèves du secondaire, c'est la moyenne des tailles d'effet de chaque étude pondérée par N_i , les tailles des échantillons (si $N_i > 2500$, le poids de l'étude était fixé à 2500) qui a été calculée. On a donc

$$ES = \frac{\sum_i N_i \times \Delta_i}{\sum_i N_i}$$

Aucune variance n'est calculée pour ces tailles d'effets globales. Robert SLAVIN considère qu'un traitement est efficace et que cette efficacité est soutenue par de fortes preuves quand la taille d'effet globale est supérieure ou égale à 0,20, qu'au moins deux études ont étudié le

^a www.bestevidence.org

traitement dont une est un essai contrôlé randomisé et que la taille totale des échantillons est au moins égale à 500.

Dans la méta-analyse publiée en 2011, pour chacune des 74 études sélectionnées, c'est une différence des moyennes ajustée aux scores prétest (ou autres covariables) divisée par un écart-type groupé non ajusté qui est calculée (donc soit le d de Cohen, soit le g de Hedges tels qu'ils ont été définis ici). Après avoir identifié une hétérogénéité importante avec $Q = 345,80$ (le degré de liberté est égal à 73 et la valeur- p calculée est inférieure à 0,00, voir chapitre 10), c'est le modèle des effets aléatoires qui a été choisi^a. Là aussi, les études sont réparties dans des groupes constitués à partir de critères très variés (comme par exemple la durée du traitement, la population ciblée, le design de l'étude, le type de logiciel étudiée), et des tailles d'effet globales sont calculées pour chacun des groupes. Cette fois, contrairement à la méta-analyse de 2008, leurs variations ont été étudiées et interprétées (calculs de variances, d'intervalles de confiance et de valeurs- p sont proposés). L'analyse des variances (voir chapitre précédent) est menée afin d'évaluer l'influence du critère ayant permis de départager les études. Quelques données numériques ont été reprises ici et constituent l'**exemple 11** dont quelques résultats sont repris ici. Ainsi, la taille des échantillons a montré avoir un impact sur les tailles d'effet calculées : un premier sous-groupe a été constitué avec des études de petite taille (effectif total inférieur à 250), un second avec des études de grandes tailles (effectif total supérieur à 250). Les résultats obtenus sont présentés dans le tableau 20^b :

	k	M^*	var_{M^*}	s_{M^*}	Z	Valeur-p	Borne inférieure	Borne supérieure
$N > 250$	44	0.12	0.001	0.02	0.08	0.000	0.17	5.15
$N < 250$	30	0.26	0.003	0.05	5.19	0.000	0.16	0.36

Tableau 20 : résultats pour deux sous-groupes d'études (SLAVIN, 2011, p.38)

Robert SLAVIN calcule $Q_{inter}^* = 6.13$, et une valeur- p égale à 0,01 avec un degré de liberté égal à 1 : les tailles d'effets globales pour chacun des deux sous-groupes sont statistiquement significativement différentes de 0 et la différence entre les études de grande taille et les études de petite taille est statistiquement significative.

Quand il s'intéresse au niveau d'études des élèves en regroupant les études dans la catégorie Secondaire ou Primaire, les résultats obtenus sont les suivants (tableau 21) :

	k	M^*	var_{M^*}	s_{M^*}	Z	Valeur-p	Borne inférieure	Borne supérieure
Primaire	45	0.17	0.001	0.03	6.00	0.00	0.11	0.22
Secondaire	29	0.14	0.001	0.04	3.92	0.00	0.07	0.21

Tableau 21 : résultats pour deux sous-groupes d'études (SLAVIN, 2011, p.39)

Robert SLAVIN calcule $Q_{inter}^* = 0.43$ et une valeur- p égale à 0,51 avec un degré de liberté égal à 1. Les tailles d'effets globales pour chacun des deux sous-groupes sont statistiquement

^a Démarche qui n'est pas recommandée par BORENSTEIN (2009), voir chapitre précédent.

^b Nous reprenons ici les notations utilisées aux chapitres précédents, ce ne sont pas les notations de SLAVIN.

significativement différentes de 0 et la différence entre le primaire et le secondaire n'est pas statistiquement significative.

Remarques sur les tailles d'effet globales

Commençons par remarquer que si toutes les tailles d'effet des études incluses dans une méta-analyse sont statistiquement et significativement positives, alors la taille d'effet globale ES est statistiquement et significativement positive (voir annexe 3 pour le détail des calculs).

La deuxième remarque appelle quant à elle une discussion sur la puissance des résultats obtenus dans les études secondaires. Un avantage des méta-analyses souvent mis en avant par leurs auteurs tient à l'augmentation de la puissance statistique des résultats quand on compare cette dernière à la puissance statistique des résultats des études primaires. En d'autres termes, quand les études primaires rassemblées par le chercheur dans une méta-analyse ne présentent que peu de résultats statistiquement significativement positifs, la méta-analyse quant à elle peut souvent présenter un profil plus avantageux. Cela peut s'expliquer assez simplement si on considère la méthode de calcul mise en œuvre dans les test-Z. En effet, pour déterminer la significativité statistique qu'il convient d'attribuer à une taille d'effet globale, on calcule le score-z suivant (présenté ici en utilisant les notations du modèle de l'effet fixe) :

$$\frac{M}{S_M} = \frac{\frac{\sum_i P_i \times g_i}{\sum_i P_i}}{\frac{1}{\sqrt{\sum_i P_i}}}$$

Plus on ajoute de tailles d'effet (plus le nombre d'études incluses dans la méta-analyse augmente), plus S_M l'erreur standard de M diminue (puisque on ajoute des termes positifs à une somme qui est un dénominateur). De plus M est toujours comprise entre la valeur minimale et la valeur maximale des tailles d'effet g_i (propriété d'une moyenne), et sera toujours du même ordre de grandeur que les tailles d'effet g_i . Dans le cas où les tailles d'effet sont toutes positives, au fur et à mesure que l'on ajoute des études à la méta-analyse, le rapport $\frac{M}{S_M}$ va augmenter et deviendra inévitablement supérieur à 1,96 à un moment donné. Une simulation (**exemple 12**) vous est proposée qui illustre l'effet de l'augmentation du nombre d'études.

La discussion annoncée quelques lignes plus haut vient du fait que la taille d'effet d'une étude est vraisemblablement dépendante de la taille de l'échantillon : il semblerait en effet que plus cette dernière est petite, plus la taille d'effet est importante. Le nuage de points obtenu à partir des données du WWC montre par exemple que les études de taille modeste concentrent les tailles d'effet les plus importantes (figure 45).

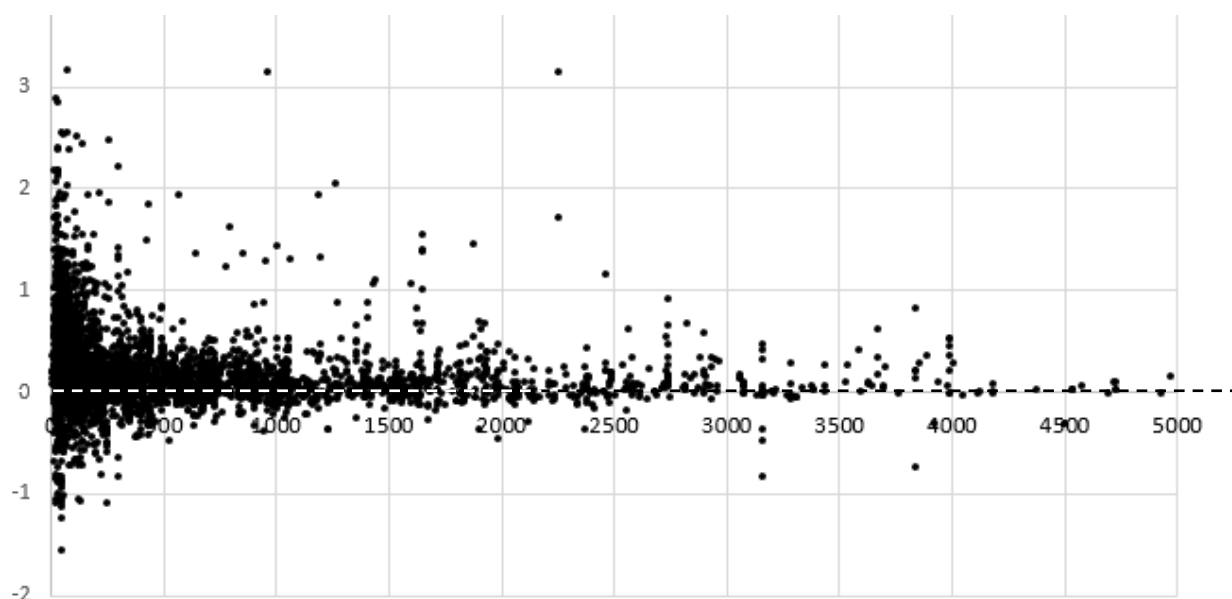


Figure 45 : tailles d'effet (ordonnée) et tailles d'échantillon (abscisses) du WWC

Robert SLAVIN aboutit lui aussi à la même conclusion (voire plus haut). Cela tiendrait probablement à la qualité moindre des études de ce type qui souffrent d'un certain nombre de biais dont s'affranchissent plus facilement les études de grandes tailles menées de façon plus rigoureuse (ZHANG, 2013). Les auteurs de l'étude sont souvent les enseignants ou leurs formateurs, les interventions sont plus faciles à mettre en place dans une ou deux écoles que dans un grand nombre d'établissements (et donc plus efficaces) et les mesures sont souvent issues de tests « maison » plus avantageux pour l'intervention que des tests standards. Il est probable aussi que les petites études qui aboutissent à des tailles d'effet négatives souffrent davantage du biais de publication que de grandes études aboutissant à des conclusions négatives^a. Et dans le même temps, plus la taille totale des échantillons d'une étude par comparaison de groupes est faible, plus sa puissance statistique est faible également : de petites études primaires, avec des tailles d'effet probablement surestimées car de qualité discutable et non statistiquement significatives vont donc pouvoir donner naissance à des tailles d'effet globale statistiquement et significativement positive, ce qui bien évidemment pose la question de la légitimité de ces conclusions.

De nombreux autres points de discussion n'ont pas manqué d'agiter la communauté scientifique, et continuent de le faire encore aujourd'hui. Pour en savoir plus, je vous invite à consulter les deux ouvrages cités au début du chapitre 10.

^a Quels que soient les résultats obtenus (et y compris en cas de résultat décevant), les études de grande taille courent moins le risque de ne pas être publiées.

Chapitre 12. Interprétation des tailles d'effet

Si la rigueur mise en œuvre dans les protocoles et les calculs statistiques évoqués tout au long des chapitres précédents donne aux conclusions des études une légitimité scientifique indiscutable, il nous faut tout de même revenir à la fin dans le monde concret et répondre à cette question à laquelle tout le monde attend une réponse : « la méthode d'enseignement mise en œuvre dans le groupe traitement doit-elle être généralisée et proposée à l'ensemble de la population ? »^a. Si les chercheurs calculent pour une intervention une taille d'effet globale (notée *ES* dans tout ce chapitre) par exemple égale à 0,4 et qu'ils ont montré qu'elle est statistiquement significativement différente de zéro, que doivent en penser les enseignants ou les personnels des administrations concernées ? La communication des résultats et leur présentation sous une forme compréhensible par le plus grand nombre est un objectif important des méta-analyses. Et c'est le soin apporté à cet aspect qui explique en grande partie le succès d'initiatives comme le What Works Clearinghouse (dont nous avons abondamment utilisé les ressources ici), ou encore de Education Endowment Foundation au Royaume-Uni.

Faisons déjà une remarque triviale : une taille d'effet peut être négative. On conclurait alors que la méthode d'enseignement expérimentée n'est pas meilleure (voire plus mauvaise) que la méthode d'enseignement « standard ». Nous ne nous intéresserons ici qu'à des tailles d'effet positives.

On utilise encore de nos jours la grille d'interprétation de Cohen (COHEN, 1988, p.26) :

- Quand $ES = 0,2$ (ou moins), l'effet est considéré comme faible.
- Quand $ES \approx 0,5$, l'effet est considéré comme moyen (visible « à l'œil nu »).
- Quand $ES = 0,8$ (ou plus), l'effet est considéré comme important.

Ces valeurs ont été proposées par COHEN en 1969^b et correspondent aux tailles d'effet calculées pour les comparaisons suivantes : différences de tailles entre des jeunes filles de 15 ans et de 16 ans ($ES = 0,2$) ; différences de taille entre des jeunes filles de 14 ans et de 18 ans ($ES = 0,5$) ; différences de taille entre des jeunes filles de 13 ans et de 18 ans ($ES = 0,8$). Mais de nombreux chercheurs pensent aujourd'hui que les tailles d'effet peuvent être considérées comme importantes dans le domaine de l'éducation même à des niveaux inférieurs à 0,5. Par exemple, jusqu'en janvier 2020 le WWC considérait qu'une taille d'effet de 0,25 était déjà importante, dans le sens où même si elle n'est pas significativement statistiquement différente de zéro, le traitement sera considéré comme ayant un effet positif^c.

^a Encore une fois, nous avons laissé de côté un grand nombre de questions qui se posent dans ce genre d'études en nous concentrant plutôt sur le sens à donner à des résultats mathématiques.

^b Date de la publication de la première édition de son livre *Statistical Power Analysis for the Behavioral Sciences*

^c C'était en tous cas la position adoptée avant la publication des nouvelles procédures 4.1 mises en place depuis 2020 (*The WWC Procedures Handbook*, 4.0, 2014).

Pour répondre à l'attente des décideurs, des indicateurs calculés à partir des tailles d'effets sont proposés qui sont censés traduire en langage clair les résultats des méta-analyses décrits dans les chapitres précédents^a.

Dans tout ce qui suit, on va pour fixer les idées admettre que l'estimation de la taille de l'effet (ES) est égale à 0,4 et que les scores des 2 populations (la population traitée et la population contrôle) suivent une courbe normale de même écart-type^b. Cela signifie alors, en utilisant les notations des chapitres précédents, que

$$ES^c \approx \frac{\mu_a - \mu_b}{\sigma} = 0,4$$

C'est-à-dire que $\mu_a - \mu_b = 0,4 \sigma$ ou encore que $\mu_a = \mu_b + 0,4 \sigma$ (rappelons que μ et σ sont des paramètres de la population et que les groupes contrôles et traitement sont identifiés respectivement par les lettres a et b). Cela revient donc à dire que la différence des moyennes des deux populations est égale à 0,4 écart-type.

Une représentation graphique des distributions des scores des deux populations permet de visualiser cette première analyse. Les données sont centrées-réduites^d : pour le groupe contrôle la moyenne des scores est égale à 0 et leur écart-type est égal à 1 ; pour le groupe traitement, la moyenne des scores est égale à la taille d'effet (donc 0,4 dans notre exemple) et leur écart-type est aussi égal à 1. L'axe des abscisses est donc gradué en nombre d'écart-type ou scores-z (figure 46).

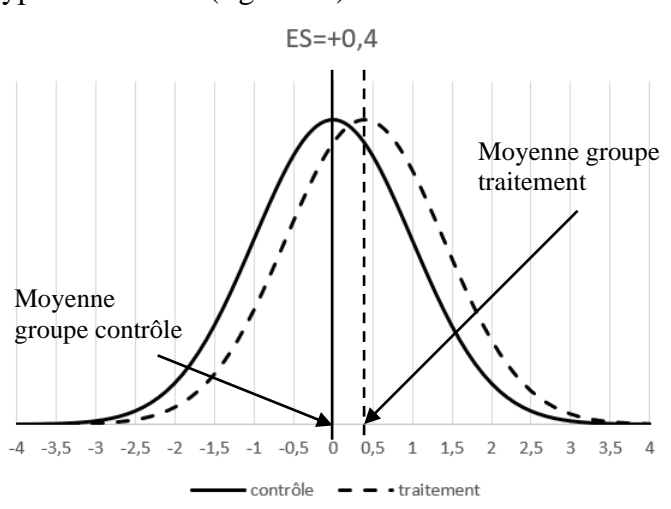


Figure 46 : distribution des scores du groupe contrôle et du groupe traitement

^a Pour vous aider à mieux visualiser ces indicateurs, je vous recommande le site internet

<https://rpsychologist.com/d3/cohend/>

^b Vous trouverez des explications plus détaillées sur les calculs suivants dans l'annexe 8.

^c Encore une fois, ES est l'estimation de la vraie valeur de la taille de l'effet notée δ précédemment. Pour simplifier, cette différence est ici escamotée.

^d On retranche la moyenne des scores du groupe contrôle et on divise par l'écart-type des scores

L'indice d'amélioration du What Works Clearinghouse

Le What Works Clearinghouse associe à chaque taille d'effet calculée un indice d'amélioration qui traduit cette dernière en termes de progression attendue du rang centile d'un élève « moyen » du groupe contrôle si ce dernier avait subi le traitement. Ce rang centile est la position du score de l'élève quand la série des scores a été ordonnée et divisée en 100 classes représentant chacune 1% de la population. Ainsi, un élève « moyen » (son score est le score moyen de son groupe) se trouve au 50^{ème} centile.

Cet indice d'amélioration est calculé à partir de l'indice U_3 de Cohen qui est le pourcentage d'élèves de la population contrôle dont le score est inférieur à la moyenne de la population traitement (qui est en fait la taille d'effet). Il s'agit donc de la probabilité que la variable aléatoire centrée et réduite « prendre au hasard un score du groupe contrôle » soit inférieure à ES . On peut utiliser une table de score- z^a pour déterminer cette probabilité : l'indice U_3 correspondant à une taille d'effet de 0,4 est égal à 66 % (à comparer aux 50% qui correspondent à une absence d'effet du traitement). On a représenté graphiquement ces conclusions avec les deux courbes normales réduites ci-dessous (figure 47).

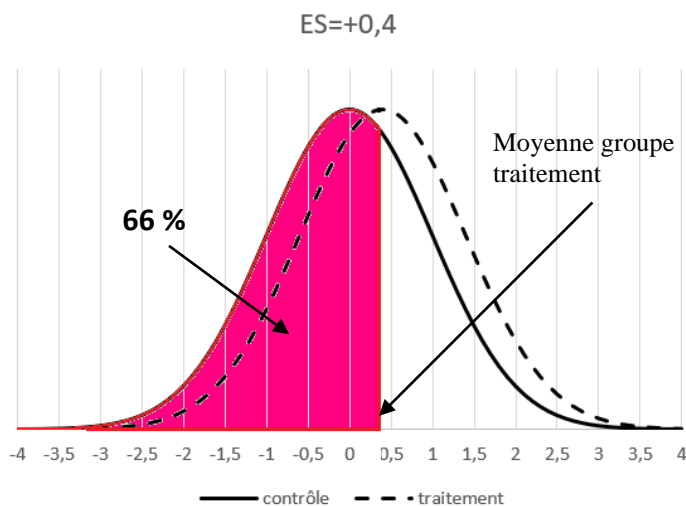


Figure 47 : indice U_3 de Cohen

Toujours avec notre exemple, si les élèves qui ont la moyenne dans la population contrôle avaient subi le traitement, ils auraient progressé de 0,4 écarts-types, et auraient le même niveau que les élèves du 66^{ème} centile de la population contrôle. Autrement dit, on peut considérer que les élèves qui ont subi le traitement et qui étaient dans la moyenne au départ, ont progressé de 16 centiles dans la cohorte. C'est de cette manière que le WWC définit l'indice d'amélioration^b (IA), qui dans cet exemple serait donc égal à +16 (figure 48).

D'une façon générale, on a $IA = U_3 - 50 \%$

^a Annexe 5

^b Procedures Handbook, p.14.

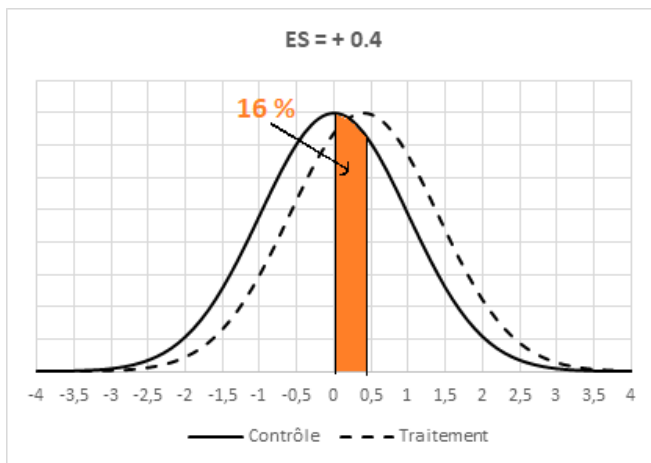


Figure 48 : indice d'amélioration du WWC

Recouvrement des deux courbes

Nous avons déjà remarqué au début du chapitre 8 (figure 36) que plus l'écart-type de la population diminuait, moins les courbes des groupes contrôle et traitement se recouvraient. Et la même remarque peut être faite quand c'est la différence des moyennes qui augmente. La proportion de l'aire commune aux deux courbes par rapport à la surface totale de la courbe des scores du groupe contrôle est un autre indicateur permettant d'évaluer l'efficacité du traitement ; là encore elle est calculée grâce aux propriétés de la loi normale χ^2 . On trouve ici que ce pourcentage est égal à 84 % (à comparer aux 100% quand les deux courbes se superposent, voir figure 49). Cela signifie donc que 84% des scores de chaque groupe correspond à un score de l'autre groupe. Donc 16% des scores de chaque groupe ne correspond à aucun des scores de l'autre groupe (pour le groupe contrôle, il s'agit des scores les plus faibles, pour le groupe traitement il s'agit des scores les plus hauts). Cet indice est dérivé de l'indice U_1 de Cohen^a.

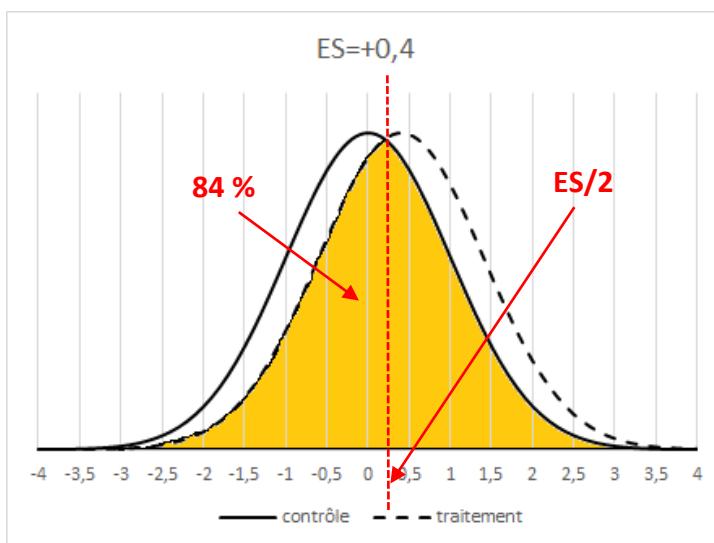


Figure 49 : recouvrement des courbes traitement et contrôle

^a Plus d'informations sur les indices U_1, U_2, U_3 de Cohen en annexe 8.

En langage courant

Je n'ai pas trouvé de traduction française pour ce que les anglo-saxons nomment le *Common Language Effect Size (CLE)*^a qui sera dénommé par la suite « effet en langage courant ». Il s'agit en fait de calculer la probabilité qu'un élève prélevé au hasard dans le groupe traitement ait un score plus élevé qu'un élève prélevé au hasard dans le groupe contrôle. Et ce sont toujours les propriétés de la loi normale qui vont permettre de répondre à cette question \square . Ici on trouve que cette probabilité est égale à 0,61. Si le traitement n'avait pas d'effet sur les élèves, alors la probabilité qu'un élève du groupe traitement ait un score supérieur à un élève du groupe contrôle serait de 0,5.

Nombre de mois « gagnés »

Au Royaume-Uni, les chercheurs britanniques travaillant dans le cadre de l'Education Endowment Foundation (EEF)^b utilisent une autre façon de visualiser les tailles d'effet. Ils reprennent à leur compte une remarque de GLASS qui souligne qu'une augmentation d'un écart-type dans les échelles utilisées pour évaluer les élèves à l'école primaire correspond à une année scolaire (soit à 11 mois d'étude) et donc 1 mois d'étude correspond à 0,09 écart-type^c. Dans ce cas, il suffit de transformer la taille d'effet en nombre de mois d'études pour évaluer l'importance à accorder à cette dernière et donc à l'efficacité du traitement. On a donc :
Nombre de mois « gagnés » = $ES \div 0,09$, et quand $ES = 0,4$ le gain est d'environ 4,4 mois.

Là aussi une évaluation de l'importance de la taille d'effet accompagne ces calculs, un peu à la façon de Cohen (tableau 22).

Taille d'effet	Nombre de mois « gagnés »	Description de l'effet
-0,1 à 0,18	0 ou 1 mois	Faible voire pas d'effet
0,19 à 0,44	2 à 5 mois	Modéré
0,45 à 0,69	6 à 8 mois	Élevé
0,70 à 0,96	9 à 12 mois	Très élevé

Tableau 22 : taille d'effet et nombre de mois « gagnés »^d

En résumé

Finalement, nous avons déterminé l'importance à accorder à une taille d'effet calculée^e de sept manières différentes. Et pour clore ce chapitre et ce livre, nous allons reprendre les données fictives de l'exemple 5 du chapitre 8 : nous avons imaginé deux populations et calculé pour l'une une taille d'effet égale à +0,4, et pour l'autre une taille d'effet de 0,9. Voilà comment nous pourrions interpréter ces deux résultats (tableau 23).

^a WUENSCK K. (2015)

^b <https://educationendowmentfoundation.org.uk>

^c Toolkit Manual https://educationendowmentfoundation.org.uk/public/files/Toolkit/Toolkit_Manual_2018.pdf

^d D'après le tableau 11.1 p.28 du Toolkit Manual

^e Il en existe d'autres, voir par exemple COE R. (2002).

Méthodes d'interprétation	Conclusions avec $ES = 0,4$	Conclusions avec $ES = 0,9$
Grilles de lectures	<ul style="list-style-type: none"> D'après la grille de Cohen, la taille d'effet est faible voire moyenne. D'après le WWC, l'effet est important. D'après les chercheurs de l'EEF, l'effet est modéré. 	<ul style="list-style-type: none"> D'après la grille de Cohen, la taille d'effet est importante. D'après le WWC, l'effet est important. D'après les chercheurs de l'EEF, l'effet est très élevé.
Nombre d'écart-types à rajouter au groupe contrôle	Subir le traitement augmente de 0,4 écart-types le score des élèves moyens de la population.	Subir le traitement augmente de 0,9 écart-types le score des élèves moyens de la population.
Indice U3 de Cohen	66 % des élèves de la population contrôle ont un score inférieur à la moyenne des élèves qui ont subi le traitement.	82 % des élèves de la population contrôle ont un score inférieur à la moyenne des élèves qui ont subi le traitement.
Indice d'amélioration (IA) du WWC	Les élèves qui ont la moyenne dans la population contrôle auraient progressés de 16 centiles dans la cohorte s'ils avaient subi le traitement.	Les élèves qui ont la moyenne dans la population contrôle auraient progressés de 32 centiles dans la cohorte s'ils avaient subi le traitement.
Recouvrement des deux courbes	84 % des populations se recouvrent	66 % des populations se recouvrent
Langage courant	La probabilité qu'un élève de la population traitée tiré au sort ait un score supérieur à un élève de la population contrôle tiré au sort également est égale à 0,61.	La probabilité qu'un élève de la population traitée tiré au sort ait un score supérieur à un élève de la population contrôle tiré au sort également est égale à 0,73.
Nombre de mois gagnés	Un élève qui subi le traitement « gagne » 4,4 mois dans sa scolarité	Un élève qui subi le traitement « gagne » 9,8 mois dans sa scolarité

Tableau 23 : interprétations d'une taille d'effet pour $ES = 0,4$ et $ES = 0,9$

Pour conclure

De nombreuses méta-analyses sont conduites de nos jours qui interrogent l'efficacité des méthodes d'enseignement proposées aux élèves. Ignorer leurs conclusions par manque de connaissances revient à rejeter un pan entier de la recherche et s'apparente à un gaspillage inacceptable pour la communauté scientifique. Si les concepts statistiques vous paraissent plus compréhensibles après la lecture de ce livre, alors mon objectif aura été atteint et vous devriez maintenant donner du sens aux informations publiées, par exemple, dans les *Rapports d'interventions* du What Works Clearinghouse.

Comme cela a été déjà souligné à plusieurs reprises, la conduite de méta-analyses ne se conçoit que dans le cadre d'un protocole précisément défini et nous n'avons abordé qu'une seule des multiples étapes qui permettent aux chercheurs de conclure quant à l'effet d'un traitement. Je vous invite à consulter les documents fréquemment cités tout au long de ce livre pour en savoir plus sur ces autres étapes tout autant décisives (la définition du protocole, la recherche bibliographique, la sélection des études et la publication des résultats notamment). En ce qui concerne les calculs statistiques abordés dans ce livre, il nous faut rester modeste. Car pour appliquer des procédures complexes à une expérience particulière dans un contexte donné avec des questions de recherche inédites afin d'identifier et d'appliquer les modèles statistiques les plus pertinents, l'intervention d'un statisticien expérimenté, et ce dès la définition du protocole d'étude, s'avère indispensable.

Mais au bout du compte, derrière tous ces calculs statistiques parfois ardues et souvent austères se cachent deux idées finalement assez simples.

La première concerne la taille des échantillons. Plus cette taille est importante, et plus fortes seront les conclusions des études. Cette remarque de simple bon sens peut être justifiée par des éléments issus du calcul statistique : en effet, une taille d'échantillon importante est en particulier associée à une meilleure estimation de l'écart-type de la population, une utilisation plus confortable du théorème central limite sur lequel sont basés les tests d'hypothèse, une plus grande puissance de ces tests, une variance plus faible des tailles d'effets calculées et une meilleure qualité des études intégrées aux méta-analyses. Il faut aussi garder à l'esprit que bien souvent ces calculs sont des simplifications d'expressions complexes, ce qui amène parfois des auteurs à proposer des méthodes de calculs (légèrement) différentes pour un même résultat. Là encore, plus les tailles d'échantillon augmentent et plus ces (légères) différences sont gommées. La seconde concerne les caractéristiques des élèves qui sont répartis dans les deux échantillons que l'on souhaite comparer (le groupe traitement et le groupe contrôle) avant le traitement. En effet, plus ces échantillons seront similaires au début de l'expérience et plus la différence observée après cette dernière sera liée à la différence des traitements. Là encore, la remarque de bon sens trouve sa traduction dans le domaine des calculs statistiques : ainsi, l'homogénéité des variances et des coefficients de détermination est une condition de l'utilisation de certains tests d'hypothèse, et des tailles d'échantillons proches augmentent l'impact des résultats des études (voir le calcul de la variance de la taille de l'effet par exemple).

Nous avons terminé ce livre par un chapitre consacré à l'interprétation des tailles d'effet. Cette vulgarisation des résultats a bien sûr comme objectif de permettre à des personnes peu familiarisées avec l'analyse statistique de prendre des décisions et d'agir en se basant sur des conclusions scientifiques. En effet, une des finalités des méta-analyses financées par les pouvoirs publics est de mettre les conclusions d'études complexes à la portée des décideurs politiques mais également des enseignants, des chefs d'établissement et des parents. L'exercice est difficile : cela implique de sélectionner un ou deux indicateurs censés résumer des recherches qui fourmillent de chiffres. Et c'est peut-être cet effort pour une vulgarisation honnête des résultats d'analyses scientifiques menées sur des interventions en milieu scolaire qui explique le succès des méta-analyses réalisées dans les pays anglo-saxons de nos jours, comme le montre l'exemple du site internet *Find What Works* (« trouver ce qui marche ») du What Works Clearinghouse dont il a si souvent été question dans ce livre.

Références

BECKER Lee A.	1999	Analysis of Pretest and Posttest Scores with Gain Scores and Repeated Measures https://www.uccs.edu/lbecker/gainscore
BORENSTEIN Michael, HEDGES Larry, HIGGINS Julian, ROTHSTEIN Hannah	2009	Introduction to Meta-Analysis , Éditions WILEY
COE Robert	2002	It's the Effect Size, Stupid. What effect size is and why it is important Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England
COHEN, Jacob	1988 (1969)	Statistical power analysis for the behavioral sciences - 2nd ed. NY: Academic Press ISBN 0-8058-0283-5
COOPER, H., HEDGES, L., & VALENTINE, J.	2019	The Handbook of Research Synthesis and Meta-Analysis. NEW YORK: Russell Sage Foundation.
DEPP	2013	Note 13,31
ENZMANN D	2015	Notes on Effect Size Measures for the Difference of Means from two Independent Groups : the Case of Cohen's d and Hedges 's g.
GLASS, Gene	1976	Primary, Secondary, and Meta-Analysis of Research , Educational Researcher, volume 5, p.3-8
GLASS Gene, SMITH Mary L.	1977 (octobre)	Meta-Analysis of Psychotherapy Outcome Studies , American Psychologist, 32(9), 752–760
HEDGES, Larry	1981	Distribution Theory for Glass's Estimator of Effect Size and Related Estimators , Journal of Educational Statistics, 1981, vol 6, n°2, p.107-128.
HEDGES, Larry	1983	A Random Effects Model for Effect Sizes , Psychological Bulletin Vol. 93, No.2, 388-395.
HIEBERT, J. GROUWS, D.	2007	The effect of classroom mathematics teaching on student's learning Second Handbook of Research on Mathematics Teaching and Learning, Chapter 9, p.371–404, 2007. Traduction sur mathadoc.fr
KNAPP Thomas R. SCHAFFER, William D.	2009	From Gain Score t to ANCOVA F (and vice versa) Practical Assessment, Research & Evaluation, 14(6). Available online: http://pareonline.net/getvn.asp?v=14&n=6 .
MARFO P., OKYERE G.A.	2019	The accuracy of effect size estimates under normals and contaminated normal in meta-analysis ; https://doi.org/10.1016/j.heliyon.2019.e01838
OCDE	2014	PISA 2012 Results: What Students Know and Can Do Student Performance in Mathematics, Reading and Science, Volume I
PIGOTT, Terri D.	2012	Advances in Meta-Analysis , Springer, New York DOI 10.1007/978-1-4614-2278-5

ROQUES, Nathalie	2019	Inégalités sociales et mathématiques dans l'OCDE. Volume 1 : comprendre l'enquête PISA. Volume 2 : l'enquête PISA revue et commentée. www.mathadoc.fr
ROQUES, Nathalie	2019	Comment enseigner les maths ? La réponse du What Works Clearinghouse. www.mathadoc.fr
SLAVIN, Robert et al.	2009	Effective programs in Middle and High School Mathematics : a best-evidence Synthesis, DOI 10,3102/0034654308330968
What Works Clearinghouse	2014	Procedures Handbook, 4.0 https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_handbook_v4.pdf
What Works Clearinghouse	2014	Standards Handbook, 4.0
What Works Clearinghouse	2020	Procedures Handbook, 4.1
What Works Clearinghouse	2017	Primary mathematics intervention report : Odyssey Math
What Works Clearinghouse	2018	Charter Schools intervention report : Knowledge is Power Program (KIPP)
What Works Clearinghouse	2021	Responses to comments from the public on updated version 4.1 of the WWC Procedures Handbook and WWC Standards Handbook
WIJEKUMAR Kay et al.	2009	A Multisite Cluster Randomized Trial of the Effects of CompassLearning Odyssey® Math on the Math Achievement of Selected Grade 4 Students in the Mid-Atlantic Region Final Report, NCEE
ZHONGHENG Zhang, XIAO Xu et HONGYING Ni	2013	Small studies may overestimate the effect sizes in critical care meta-analyses: a meta-epidemiological study http://ccforum.com/content/17/1/R2 <i>Critical Care</i> 17:R2

Sur internet uniquement

https://www.math.u-psud.fr/~pansu/web_ifips/proba_S5-IFIPS.pdf

LENOIR, 2008, PROBABILITÉS ET STATISTIQUE EN S5 IFIPS (en français)

<http://vassarstats.net/textbook/> de Richard Lowry

Nombreux exemples numériques permettant de faire les calculs à la main. Des simulateurs sont également proposés (en anglais).

[https://perso.univ-rennes1.fr/denis.poinsot/Statistiques %20pour statophobes /STATISTIQUES%20POUR%20STATOPHOBES.pdf](https://perso.univ-rennes1.fr/denis.poinsot/Statistiques_%20pour_statophobes/STATISTIQUES%20POUR%20STATOPHOBES.pdf)

Denis POINSOT, 2004, *Statistiques pour statophobes*, (en français)

http://www.txrating.org/spc/bookma_web/frame.htm

Manuel pratique de méta-analyse des essais thérapeutiques

Michel Cucherat, Jean Pierre Boissel, Alain Leizorovicz, (en français).

<http://core.ecu.edu/psyc/wuenschk/docs30/CL.pdf>

WUENSCH Karl, 2015a, *The Common Langage Effect Size Statistics*, (en anglais).

<http://core.ecu.edu/psyc/wuenschk/MV/LSANOVA/Pretest-Posttest-ANCOV.pdf>

WUENSCH Karl, 2015b, *The Pretest-Posttest x Groups Design: How to Analyze the Data*, (en anglais).

D'autres leçons sur ce site : <http://core.ecu.edu/psyc/wuenschk/StatsLessons.htm>

<https://stats.stackexchange.com/questions/ask>

Forum d'entraide (en anglais).

<http://www.les-mathematiques.net/>

Forum d'entraide (en français).

[www.meta-analysis](http://www.meta-analysis.com), puis chercher le livre de BORENSTIN (2009) pour télécharger les fichiers Excel des chapitres 14, 18 et 19.

Liste des figures, des tableaux et des exemples

Figures

Figure	1 :	essai contrôlé randomisé
Figure	2 :	étapes suivies par une méta-analyse
Figure	3 :	poids de 252 hommes (nuage de points)
Figure	4 :	4 poids particuliers
Figure	5 :	poids de 252 hommes centrés réduits
Figure	6 :	histogramme des poids de 252 hommes (en kg)
Figure	7 :	histogramme des scores de 29 846 élèves
Figure	8 :	distribution des scores (exemple 3)
Figure	9 :	probabilités (exemple 3)
Figure	10 :	histogrammes (exemple 3)
Figure	11 :	densité de probabilité
Figure	12 :	représentation graphique de 3 lois normales
Figure	13 :	représentation graphique générale de $\mathcal{N}(\mu, \sigma)$
Figure	14 :	proportions de surface délimitée par une courbe normale
Figure	15 :	loi normale centrée réduite
Figure	16 :	utilisation des scores-z
Figure	17 :	propriétés de la loi normale centrée réduite
Figure	18 :	proportions de la surface totale à retenir
Figure	19 :	valeurs particulières d'une loi normale centrée et réduite
Figure	20 :	distribution des sommes (exemple 3)
Figure	21 :	poids et scores d'élèves avec les courbes normales correspondantes
Figure	22 :	distribution des moyennes (exemple 3)
Figure	23 :	intervalle de confiance à 95%
Figure	24 :	lois de Student pour 7 degrés de liberté
Figure	25 :	propriétés des lois de Student
Figure	26 :	Plan d'expérience dans les études par comparaison de groupes
Figure	27 :	les données utilisées après le traitement
Figure	28 :	test bilatéral avec la statistique Z
Figure	29 :	test unilatéral avec la statistique Z
Figure	30 :	différences de moyennes statistiquement significatives (exemple 4)
Figure	31 :	signification statistique de la valeur-p
Figure	32 :	test-z ou test t ? Arbre de décision
Figure	33 :	moyennes et intervalles de confiance à 95% des scores PISA
Figure	34 :	impact de la valeur de l'écart-type pour des groupes d'élèves ayant la même différence de moyennes.
Figure	35 :	plan d'expérience tenant compte du niveau des élèves avant le traitement
Figure	36 :	scores posttests en fonction des scores prétests
Figure	37 :	étapes suivies pour mener une méta-analyse
Figure	38 :	trois modèles pour évaluer l'effet du traitement

Figure	39 :	variation de W_i en fonction de n_{ai} avec N_i fixé (ici $g = 0,4$)
Figure	40 :	tailles d'effet et tailles d'échantillon (WWC)
Figure	41 :	scores du groupe contrôle et du groupe traitement
Figure	42 :	indice U3 de Cohen
Figure	43 :	indice d'amélioration du WWC
Figure	44 :	recouvrement des courbes traitement et contrôle

Tableaux

Tableau	1 :	poids de 252 hommes en kg
Tableau	2 :	scores-z de 4 poids
Tableau	3 :	effectifs et fréquences des poids dans des classes d'amplitude 5 kg
Tableau	4 :	effectifs et fréquences des poids dans des classes d'amplitude 2,5 kg
Tableau	5 :	4 tirages au sort (exemple 3)
Tableau	6 :	proportions de la surface normale pour certains intervalles de valeurs
Tableau	7 :	sommes de 40 tirages au sort (exemple 3)
Tableau	8 :	proportions des poids dans des classes centrées autour de la moyenne
Tableau	9 :	proportions des scores dans des classes centrées autour de la moyenne
Tableau	10 :	mesures dans plusieurs échantillons
Tableau	11 :	moyennes de 40 tirages au sort (exemple 3)
Tableau	12 :	écart-type d'un échantillon d'une population
Tableau	13 :	40 scores tirés au sort (exemple 3)
Tableau	14 :	intervalles de confiance : test Z ou test T ?
Tableau	15 :	différences de moyennes statistiquement significatives (exemple 4)
Tableau	16 :	deux groupes de 40 données (exemple 3)
Tableau	17 :	moyennes et erreurs standards de certains pays (PISA 2012)
Tableau	18 :	taille d'effet et nombre de mois « gagnés »
Tableau	19 :	interprétations d'une taille d'effet pour $ES = 0,4$ et $ES = 0$

Exemples

Les équations permettant de calculer certains des résultats présentés dans ces exemples sont identifiées par des numéros permettant de les retrouver dans les annexes publiées séparément.

Exemple 1 : les poids de 252 hommes (chapitres 1, 4 et 6, annexe 6)

Exemple 2 : les scores PISA de 29 846 élèves de l'OCDE (chapitres 1 à 7)

Exemple 3 : les scores de 2000 élèves (chapitres 2, 4, 5, 6 et 7)

Exemple 4 : intervalles de confiance : loi normale ou loi de Student ? (chapitre 6)

Exemple 5 : différences de moyennes significatives (chapitres 7, 8 et 12)

Exemple 6 : intervalles de confiance de 4 pays de l'OCDE (chapitre 7)

Exemple 7 : calculs de tailles d'effets, méta-analyse, sous-groupes d'études (chapitre 8 et 10)

Exemple 8 : étude de deux groupes d'élèves (chapitre 1, 7, 8 et 9)

Exemple 9 : correction pour cluster (chapitre 9, annexe 2)

Exemple 10 : deux méta-analyses du WWC (chapitre 8, 9 et 10)

Exemple 11 : 2 sous-groupes d'études A et B (chapitre 10)
Exemple 12 : deux méta-analyses, exemple fictif (chapitre 10)

Table des matières

Pour commencer	5
Premier niveau : les études par comparaison de groupes	7
Le premier principe : mener une expérience scientifique.....	8
Le deuxième principe : étendre les conclusions à une population entière.....	8
Le troisième principe : comparer deux échantillons	8
Deuxième niveau : les méta-analyses	10
Analyse statistique des données	11
Première partie : les données et le hasard	13
Chapitre 1. Description des données	15
Moyenne et écart-type	15
Centrer et réduire une série de données	17
Histogramme	20
Modéliser la distribution des données	22
Chapitre 2. Variables aléatoires	23
Variables aléatoires discrètes	23
Variables aléatoires continues.....	27
Variables aléatoires centrées et réduites	27
Chapitre 3. La loi normale	29
Les propriétés des courbes normales.....	30
Score-z et loi normale centrée réduite.....	31
Les chiffres à retenir	34
Chapitre 4. Distribution normale de caractéristiques naturelles.....	37
Le théorème central limite	37
Des exemples de données expérimentales distribuées normalement	39
Le hasard et les données	41
Deuxième partie : les échantillons	43
Chapitre 5. Échantillonnage et intervalle de fluctuation	45
Étude théorique de la moyenne M	46
Intervalle de fluctuation	47
Chapitre 6. Estimation des paramètres d'une population	51
Estimateurs.....	51
Estimations ponctuelles.....	52
Estimation par intervalle de la moyenne.....	53
Chapitre 7. Différence entre deux moyennes statistiquement significative.....	61
Quelle est la question ?	61
Les tests d'hypothèses.....	62

Le test Z	63
Illustrations concrètes	66
Risque α , risque β	68
Le test de Student.....	68
La valeur- p	71
L'analyse de la variance (ANOVA).....	72
La procédure suivie par les tests	73
Utiliser les intervalles de confiance	74
Chapitre 8. Taille d'effet du traitement	76
Ampleur de la différence entre deux moyennes.....	76
Taille de l'effet du traitement.....	77
Estimations ponctuelles.....	78
Le d de Cohen.....	79
Le g de Hedges	79
Le Δ de Glass.....	80
Variances de d , g et Δ Pour chacune des trois estimations d , g et Δ , on va calculer une variance de leur distribution d'échantillonnage. Les formules seront admises.....	80
Estimation par intervalle	81
Tests statistiques	81
Un exemple numérique	82
Conclusions.....	83
Chapitre 9. La complexité du terrain	84
Quand les deux échantillons sont différents avant le traitement.....	84
Corrélation entre scores posttest et scores prétest dans un groupe.....	85
ANCOVA ou ANOVA de la différence.....	87
Taille de l'effet ajustée	88
Les analyses multiples	89
Quand l'unité d'analyse n'est pas l'unité d'affectation.....	89
Troisième partie : les études	91
Chapitre 10. Les méta-analyses.....	93
Le modèle de l'effet fixe	95
Le modèle des effets aléatoires.....	97
Modèle de l'effet fixe et modèle des effets aléatoires : le bilan.....	100
Un exemple numérique	100
Intervalle de prédiction	103
Hétérogénéité.....	103
Analyse de sous-groupes d'études	104
Un exemple numérique	106

Les autres méthodes.....	107
Chapitre 11. Méta-analyse et enseignement des mathématiques	110
Les réponses du What Works Clearinghouse	110
Les réponses de Robert SLAVIN	112
Remarques sur les tailles d'effet globales	114
Chapitre 12. Interprétation des tailles d'effet.....	117
En résumé	121
Pour conclure.....	123
Références	125
Liste des figures, des tableaux et des exemples	128
Figures.....	128
Tableaux.....	129
Exemples.....	129

Annexes