Statistiques
pour
statophobes

Une introduction au monde des tests statistiques
à l'intention des étudiants qui n'y entravent que pouic
et qui détestent les maths par dessus le marché

Denis Poinsot

2004

Pour citer ce document :
D. Poinsot, 2004. *Statistiques pour statophobes*.

Ce petit livre est dédié avec reconnaissance à ***René Merckhoffer****, mon extraordinaire prof de maths de classe de seconde au lycée des sept mares d'Elancourt en 1982, pour son humour pince sans rire, mais surtout pour avoir réussi à m'arracher — même très momentanément — du fond de l'abîme mathématique dans lequel j'avais sombré, sans espoir de revoir la lumière du jour. S'il n'avait pas été là au bon moment, je n'aurais tout simplement pas pu faire d'études scientifiques.*

DP

Avant Propos

Je suis biologiste, et non statisticien. Circonstance aggravante, j'ai collectionné les mauvaises notes en mathématiques sans interruption à partir de la classe de 5ème, litanie interrompue seulement par l'obtention d'une thèse de doctorat en biologie évolutive[1]. Je pense donc être idéalement qualifié pour expliquer les bases des méthodes statistiques aux étudiants en biologie réfractaires aux maths. Si vous voulez bien mettre de côté une incrédulité très naturelle à ce stade de votre lecture, vous réaliserez que cela n'est peut être pas si idiot que ça en a l'air. Bien sûr, les manuels d'introduction aux statistiques pullulent, rédigés par de véritables bio-mathématiciens et statisticiens infiniment plus doués que moi dans leur discipline. Et c'est justement là le problème. Malgré toute leur science, mes chers collègues (dont j'envie sincèrement les compétences) ne pourront jamais se mettre complètement à la place d'un étudiant ne comprenant rien aux maths, parce que, anciens étudiants "matheux", ils n'ont jamais connu cette humiliante expérience eux-mêmes. Moi, si. J'y suis même régulièrement confronté chaque fois que je me heurte durement aux étroites limites de mon savoir dans cette discipline. Je sais tout de la frustration, voire de la rage que l'on peut ressentir face à l'"explication" d'une méthode dont on a besoin pour analyser ses résultats, mais que le manuel décrit uniquement dans un langage mathématique pur et dur. Soyons clairs, je ne blâme évidemment pas les mathématiciens pour l'utilisation d'un langage symbolique précis et rigoureux, il est indispensable à leur discipline. Je souhaiterais cependant qu'ils essayent davantage de comprendre que le pékin moyen ne lit pas cette langue couramment.

Lorsque j'ai eu à enseigner pour la première fois sans bénéficier de la présence rassurante d'un collègue expérimenté, j'étais un étudiant en fin de thèse très heureux de faire de la biologie, ma passion depuis aussi longtemps que je me souvienne d'avoir été à l'école. Bien entendu, je devais utiliser les méthodes d'analyse statistique pour les besoins de ma recherche, mais mon directeur de thèse, chercheur au CNRS, m'apportait alors tout son soutien et sa vaste expérience. J'utilisais en fait à l'époque les techniques statistiques avec la foi enfantine d'un homme des cavernes regardant dans un microscope. Je savais en gros que lorsque mon test révélait que « $P < 0,05$ » il y avait un *effet significatif* dont je pouvais discuter, et que sinon je devais tristement m'abstenir de conclure. Or donc, j'eus la chance d'obtenir un contrat d'enseignement de un an pour finir ma thèse. C'est alors qu'on m'annonça que j'allais y assurer des travaux dirigés de... probabilités et statistiques, à des étudiants de première année. Je me souviens encore de la sensation que tout mon sang venait de se congeler dans mes

[1] Une fois que vous êtes docteur, plus personne n'ose mettre en doute vos compétences en mathématiques en vous obligeant à passer des examens écrits. C'est un des multiples avantages de notre beau métier.

veines. Cependant, les prouesses dont l'être humain est capable lorsqu'il ne peut fuir et que le combat est la seule issue sont véritablement étonnantes. Je parvint en effet à assurer les séances prévues, en les préparant évidemment frénétiquement, physiquement malade de terreur avant chaque TD, et totalement épuisé à la fin. Et à ma grande surprise, je me mis à *comprendre* des choses qui m'étaient pourtant passées des kilomètres au dessus de la tête lorsque j'étais étudiant..

Un an plus tard (c'était vers la fin du XXème siècle), recruté à l'université de Rennes comme maître de conférences (en biologie et non en stats, est il besoin de le préciser ?), j'ai eu a nouveau l'opportunité d'enseigner les biostatistiques de base, cette fois à des étudiants de maîtrise de biologie devant les utiliser pour analyser des données de terrain. J'ai alors pris une folle décision : écrire pour ces étudiants le manuel de stats que j'aurais aimé avoir lorsque j'étais moi même traumatisé par cette matière maudite. Le résultat est entre vos mains. J'espère que ce petit ouvrage vous sera utile et même qu'il vous plaira, parce que je pense honnêtement qu'il est différent de beaucoup d'autres livres de stats. J'en ai tant bavé[2] pour comprendre le peu que je sais dans cette discipline, que j'ai soigneusement évité les "explications" telles que : « *soit (Ω, F, p) un espace probabilisé modélisant une espérance finie* » qui m'ont toujours donné envie de posséder un lance flammes. Ce livre est donc écrit en français normal. Il contient même nombre de remarques plus ou moins saugrenues, parce que je suis viscéralement incapable de résister à l'envie de dire (et d'écrire) des bêtises, juste pour rire. Depuis sa première version, imprimée sous forme de polycopié en octobre 1998, et profondément remaniée cet été, cet ouvrage a été testé par environ 900 étudiants de maîtrise, qui l'ont utilisé pour analyser leurs données de terrain. Quelques uns ont eu la gentillesse de m'en dire du bien. Quasiment tous m'ont fait remarquer que je parlais trop. Ils ont évidemment raison (au moins sur le second point). Je vous invite donc à tourner la page.

Denis Poinsot,

Rennes le 11 octobre 2004

[2] et je suis poli.

Sommaire

1. Pourquoi des stats en biologie?

— Karl-Heinz vonVölapuk vous êtes directeur de production du groupe Bercedes Mens pour toute l'Eurasie.

— Rigoureusement exact.

— Pourriez-vous nous dire quel est le poids total de votre dernier modèle décapotable 450 C ?

— Le poids total de notre modèle 450 C, réservoirs vides, est de 1251 kg exactement.

— Et comment cela se compare t–il avec les caractéristiques de sa principale rivale ?

— Je présume que vous faites référence à la WMB 3.0i?

— Bien entendu.

— Et bien notre voiture pèse précisément 47kg de moins que sa concurrente, qui atteint en effet 1298kg dans les mêmes conditions.

— Peut-on en conclure que la Bercedès-Mens 450 C est plus légère que la WBM 3.0i ?

— C'est l'évidence même.

— Karl-Heinz vonVölapuk, je vous remercie.

— Tout le plaisir a été pour moi.

Que de précision, que de rigueur… « *Ah ces Allemands tout de même !* » direz-vous d'un air admiratif. Peut être, cependant tout constructeur automobile n'aurait eu aucune peine à faire le même genre de réponse ferme et définitive. Avant d'analyser pourquoi, voyons d'abord un dialogue identique au précédent dans sa structure, mais en fait fondamentalement différent…

— Robert Lebouvier vous êtes l'expert mondial incontesté de la race bovine charolaise.

— C'est ce qu'on dit.

— Pourriez-vous nous dire combien pèse un taureau Charolais de trois ans ?

— Eh bien, disons… entre 800 kg et 1,2 tonnes à peu près, mais certains arrivent même à être encore plus gros. Ils sont plus légers bien sûr si la pâture n'a pas été bonne, et puis il faut savoir qui étaient le père et la mère hein, parce que la génétique…

— Heu… oui… bien…, et comment ces résultats peuvent-ils se comparer avec ceux de la race Holstein ?

— Les taureaux Holstein font plutôt 700kg à 1 tonne mais, là encore, ça dépend du type d'élevage et de l'alimentation qui…

— Certes, mais alors peut-on dire que les taureaux Holstein sont plus légers que les taureaux Charolais ?

— Ben… en général peut être… quoique si, par exemple, vous prenez « Lulu le Tarbais », qui a été primé au dernier salon de…

— Je vois. Le temps qui nous était imparti touche hélas à son terme, merci beaucoup pour cette intervention, et maintenant une page de publicité.

Robert Lebouvier est il vraiment l'expert qu'il prétend être, lui qui est visiblement incapable de donner une réponse claire et nette sur un sujet qu'il connaît soi-disant à fond ? Doit-on l'accabler ? Non, évidemment. Contrairement aux voitures, les taureaux ne sont pas construits dans des conditions contrôlées et à partir de pièces qui sont automatiquement rejetées si elles ne satisfont pas le cahier des charges. Résultat : un produit *non* calibré. Il est impossible d'échapper à cette marge d'incertitude, intrinsèque à tout phénomène vivant. Chaque caractéristique d'un organisme (qu'il s'agisse de sa masse ou de son comportement à un moment précis) résulte de l'interaction entre son génome (plusieurs milliers de gènes, donc une infinité de combinaisons possibles) et l'environnement, lui même fort variable. Le résultat final est ce que vous en connaissez : une myriade d'individus tous différents, même s'ils appartiennent à la même espèce, même s'ils ont le même père et la même mère. Toute expérience visant à estimer la différence (éventuelle) entre deux groupes d'individus pour un caractère donné (le poids moyen dans notre exemple, ou le temps de réaction après un stimulus) ne peut donc pas s'appuyer sur *un seul exemplaire de chaque groupe* pris au hasard. Cette approche était pourtant valable pour comparer deux modèles manufacturés (c'est le principe du magazine « Que Choisir »). En biologie, la grande variabilité des individus oblige à se baser sur des échantillons *de plusieurs individus* (et tant mieux s'ils sont nombreux).

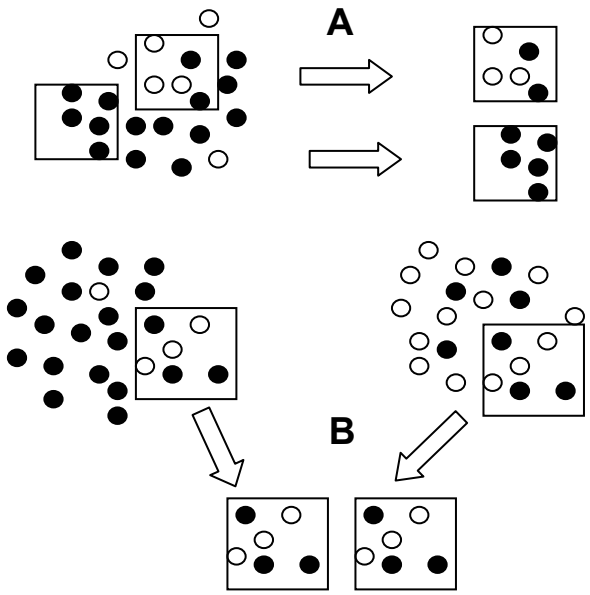


Figure 1.1 Les méfaits des fluctuations d'échantillonnage. A: Deux échantillons, même fort différents, ne proviennent pas nécessairement de deux populations différentes. **B** : Deux échantillons, même fort semblables, ne proviennent pas nécessairement de deux populations semblables.

Le problème qu'il faut bien avoir à l'esprit est que la variabilité du résultat *n'en disparaît pas pour autant*. Puisque tous les individus biologiques sont différents, *il n'y aura jamais deux échantillons semblables* !

Mais assez d'exemples théoriques, passons à de véritables données scientifiques et voyons si une personne raisonnable et compétente a vraiment besoin de tout un attirail mathématique pour les interpréter. La figure 1.2 montre le résultat d'une expérience d'écotoxicologie

(Ishimata & Takahiro, 1967) dont le but était d'établir l'impact potentiel de la cyano-cobalamine[3] (un puissant polluant cyanuré issu de l'industrie minière à ciel ouvert, fréquente à l'époque) sur le rendement du riz.

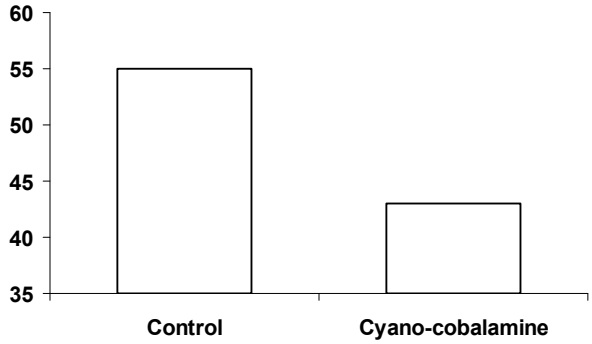


Figure 1.2. Effect of cyano-cobalamine on rice yield in dry farming conditions. Means of three replicates per treatment (kg per plot). D'après Ishimata & Takahiro, 1967, *J. Tropical Rice Res*. **12**:459-463.

Cette figure montre la moyenne du rendement obtenu en comparant deux modalités (sol non pollué *vs* sol pollué par la cyanocobalamine), avec trois répétitions (=trois parcelles expérimentales) par modalité. Ces résultats montrent de manière indiscutable que le rendement moyen obtenu dans les trois parcelles polluées est inférieur au rendement moyen obtenu dans les trois parcelles témoin (non polluées). Ressentez-vous le besoin irrésistible de calculer quelques intégrales ou autres logarithmes avant d'interpréter ces résultats ? Non, bien entendu. Il est clair que ce polluant a un impact négatif sur le rendement du riz, et il n'y a franchement rien d'étonnant là dedans.

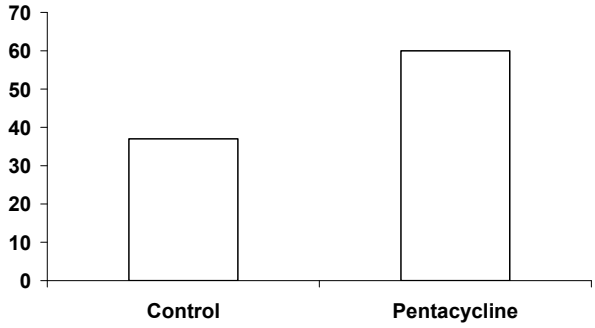


Figure 1.3. Effect of pentacycline (20mg/kg) on survival rate (%) 24 hours post treatment in mice inoculated with *Salmonella sp*.(intra peritoneal route). N=30 per treatment. (D'après Wilkins & Fretwell 1998. *J. Rodent Med. Assoc*. **53**:289-292).

La figure 1.3 montre le résultat d'un essai antibiotique préliminaire (Wilkins & Fretwell, 1998) dans lequel soixante souris ont été inoculées (par injection dans la cavité péritonéale) avec une souche potentiellement mortelle de *Salmonella sp*., bactérie tristement célèbre pour les nombreux cas d'intoxication alimentaire qu'elle provoque chaque année. Après cette injection, 30 souris tirées au hasard (lot témoin) sont laissées tranquilles, tandis que les 30 autres (lot traité) reçoivent une dose de Pentacycline (un antibiotique à large spectre dont on espère qu'il est actif *in vivo* sur cette souche de Salmonelle). On examine les sujets 24 heures

[3] DIMETHYL-5,6 BENZIMIDAZOLYL)-ALPHA CO-CYANO COBAMIDE

après. Comme vous pouvez le voir sur la figure 1.3, seules 37% des souris du lot témoin ont survécu, alors que cette proportion est de 60% dans le lot traité avec l'antibiotique. Il est manifeste que l'antibiotique a eu un effet positif — même s'il est améliorable, il faudrait probablement augmenter la dose — sur le traitement de l'infection. Encore une fois, avez-vous ressenti une envie pressante de faire des mathématiques compliquées avant d'interpréter des résultats aussi limpides ? Bien sûr que non !

Et vous avez eu tort. Précisons tout d'abord que les deux expériences décrites ci-dessus n'ont jamais eu lieu, que la cyanocobalamine n'est pas un puissant polluant cyanuré de l'industrie minière (il s'agit en fait de la vitamine B12, anti-anémique), que les noms des chercheurs cités sont imaginaires, que la pentacycline n'existe pas (la tétracycline oui, et c'est bien un antibiotique), que le Journal de l'Association Médicale des Rongeurs (*J. Rodent Med. Assoc.*) n'existe pas non plus, et précisons enfin qu'aucune adorable petite souris n'a subi de mauvais traitements pour les besoins de la rédaction de cet ouvrage. Donc, j'aurais inventé ces résultats ? Pas exactement. Ils ont été *générés aléatoirement* (tirés au hasard) par ordinateur. Le prétendu "rendement moyen" de la pseudo expérience sur le riz a été obtenu en prenant la moyenne de 30 nombres tirés au hasard entre 0 et 100. La figure 1.2 montre donc en fait les moyennes obtenues lors de deux séries de 30 tirages aléatoires chacune. La différence entre ces deux moyennes (cette différence est *réelle*, et on la voit très bien sur la figure) est évidemment un pur hasard. J'ai également utilisé une astuce bien connue des professionnels du marketing : l'échelle des ordonnées ne part pas de zéro, ce qui permet de rendre la différence de hauteur entre les barres de l'histogramme plus impressionnante que dans la réalité. Si jamais vous voyez une échelle Y qui ne part pas de zéro, méfiez-vous, on essaie de vous vendre quelque chose (par exemple que les chiffres du chômage ont baissé de manière spectaculaire).

De même, le "taux de survie à 24h" de mes deux lots de 30 souris virtuelles a été obtenu en réalisant à chaque fois 30 tirages aléatoires avec comme résultat possible 0 (souris morte) ou 1 (souris vivante), chacun ayant la même probabilité (une chance sur deux), soit l'équivalent de tirer 30 fois à pile ou face. Le "taux de survie" est simplement le nombre de 1 obtenus, divisé par 30 et multiplié par 100 pour permettre d'afficher un pourcentage. Là encore, la différence observée entre les deux pourcentages obtenus (cette différence est *réelle*, on la voit bien sur la figure) est due entièrement au hasard.

Et alors ? A quoi rime ce canular de gamin ? Il démontre mine de rien une chose importante, qui est que *le hasard peut très facilement provoquer des différences notables* (mais évidemment aléatoires) entre des moyennes ou des pourcentages, en particulier lorsque les échantillons sont de taille modeste (ce qui était le cas ici). Ce phénomène sera particulièrement dangereux quand la différence ainsi produite *va dans le sens que l'on attendait* (diminution du rendement par le polluant, amélioration de la survie par l'antibiotique), car il est alors très tentant de conclure que l'expérience "a marché". Conclusion : on ne peut pas se fier aveuglément à une différence observée (entre deux moyennes, entre deux pourcentages). Il est impératif de prendre en compte le fait que le hasard a forcément joué un rôle dans le résultat obtenu. Il y a en fait deux possibilités :

1) la différence observée est due *seulement* au hasard

2) la différence observée est due au hasard et à un *effet réel*

Le problème est évidemment qu'il est *impossible* de déterminer, juste en observant une différence, quelle part (obligatoire) est due au hasard et quelle part (éventuelle) est due à un

véritable effet. Il est *possible* en revanche de calculer *la probabilité d'observer par hasard une différence aussi grande, voire plus grande que celle que vous avez obtenue*. Telle est la tâche la plus utile des tests statistiques. Par ailleurs (et c'est très important), il est possible de calculer autour de chacune de vos valeurs observées (moyenne, pourcentage) une zone dans laquelle se situe vraisemblablement la véritable valeur dans la grande population échantillonnée. Cette zone se nomme "intervalle de confiance". Enfin, il est également possible de déterminer quelle est la *magnitude* possible de l'effet que vous avez peut être observé (la taille de la différence), là encore en calculant un intervalle de confiance pour cette magnitude.

Comme tout ceci est assez flou pour l'instant, voici à quoi cela pourrait ressembler dans les deux cas fictifs décrits ci dessus.

Analyse des résultats concernant l'effet de la Cyanocobalamine (CCA) sur le rendement du riz[4]. On suposera les observations suivantes : Témoin {43, 57, 65 q/ha} soit une moyenne de 55q/ha et Pollué {31, 43, 55qx/ha} soit une moyenne de 43q/ha

- Intervalle de confiance des rendements observés :

  Témoin non pollué : [27—83q/ha] (la véritable valeur a 95% de chances de se situer dans cet intervalle)

  Sol pollué: [13—73q/ha] (idem)

On peut observer que la précision de notre estimation des rendements est catastrophique, ce qui est dû au petit nombre de données et à une grande variabilité d'une parcelle à l'autre.

- Probabilité d'observer un écart aussi grand (voire plus grand) entre les rendements obtenus (ici l'écart observé est de –12q/ha) si seul le hasard entre en jeu :

  $P = 0{,}25$ environ (une chance sur 4)

Si seul le hasard était intervenu, on aurait donc observé *dans près d'un cas sur quatre* un écart au moins aussi important que celui que nous observons ici. De quoi doucher notre enthousiasme si nous étions tentés de conclure à un effet clair de la CCA sur le rendement.

- Magnitude de l'effet (apparent) de la CCA sur le rendement : – 12q/ha, comme déjà dit.

Ce résultat est évidemment à relativiser très fortement vu la facilité à obtenir ce genre d'écart sous l'effet du hasard !

- Intervalle de confiance de la magnitude de l'effet (apparent):[– 49 —— +27q/ha]

En clair, vu la *très grande imprécision* des estimations des rendements, il est certes possible que la CCA ait un effet *très négatif* (– 49q/ha), mais il est *également possible* que le PE ait au contraire un effet... *très positif* (+ 27q/ha), qui aurait été masqué ici par les fluctuations d'échantillonnage !

Avec ces informations en main, vous voyez que nous sommes nettement mieux armés pour ne pas foncer bille en tête sur une conclusion hâtive. Nous pouvons faire dire aux données ce qu'elles peuvent dire, mais pas plus. Ici, la (modeste) conclusion serait grosso modo celle-ci : les valeurs de rendement ayant été estimées de manière extrêmement imprécise (intervalle de confiance très large), il est impossible *d'affirmer* que le PE ait eu un effet négatif sur le rendement dans cette expérience, il est tout aussi impossible *d'exclure* qu'il ait eu un tel effet,

[4] On verra naturellement dans les chapitres suivants comment on peut obtenir les chiffres en question

et on ne peut pas non plus exclure l'hypothèse qu'il ait, en réalité un effet *bénéfique* sur le rendement, qui aurait été masqué ici par les fluctuations d'échantillonnage. Le résultat de cette expérience nous laisse donc *dans une totale incertitude* concernant l'effet du PE sur le rendement du riz. C'est un peu désespérant, mais c'est ainsi. On peut cependant quand même tirer quelques conclusions, grâce au calcul de l'intervalle de confiance de la magnitude de l'effet possible du CCA :

(1) si un effet *négatif* de la CCA existe réellement à la dose employée, il n'est vraisemblablement pas plus sévère que – 49q/ha (ce qui serait déjà catastrophique !),

(2) si au contraire la CCA a un effet *positif* à la dose employée, cet effet n'est vraisemblablement pas plus important que +27q/ha (ce qui en ferait un engrais de rêve).

Ca n'est pas grand chose, mais c'est mieux que rien. La morale de cette histoire est qu'on peut toujours tirer de l'information d'une expérience, même si elle est entachée d'une grande incertitude.

Effet de la Pentacycline sur *Salmonella sp. in vivo* chez la souris *Mus musculus*. (en pratique, 11 souris sur 30 ont survécu dans le lot témoin, soit 37%, alors que 18 souris sur 30 ont survécu dans le lot traité, soit 60%)

- Intervalle de confiance (à 95%) des taux de survie observés :

  Témoin : [20————56%] (la véritable valeur a 95% de chances de se situer dans cet intervalle)

  Pentacycline : [41————77%] (idem)

On note au passage la *très mauvaise précision* de ces estimations (presque du simple au triple pour la première, presque du simple au double pour la seconde !)

- Probabilité d'observer un écart aussi grand (voire plus grand) si seul le hasard entre en jeu (autrement dit, si l'antibiotique n'a en réalité aucun effet sur la survie)

  (chi2=3.27) $P = 0{,}12$ (plus d'une chances sur 10)

Le fait qu'un placebo (un médicament sans effet réel) puisse obtenir le même type de résultat "seulement" une fois sur dix peut sembler encourageant à première vue concernant l'existence d'un effet antibiotique de la pentacycline sur la salmonelle, mais cette possibilité resterait très inquiétante s'il s'agissait de décider de mettre cet antibiotique sur le marché pour sauver des vies !

- Magnitude de l'effet *apparent* sur le taux de survie : + 23% de taux de survie

Effet certes prometteur à première vue, mais à relativiser comme vu plus haut dans la mesure où, quand on teste un produit *n'ayant aucun effet réel* (et que seul le hasard joue) ce type d'écart sera observé tout de même *une fois sur dix*[5].

- Intervalle de confiance de la magnitude de l'effet de l'antibiotique : [–2 —— +48%]

En clair, vu la très mauvaise précision de l'estimation des pourcentages de survie, il est même possible que l'antibiotique ait un – faible – effet *négatif* sur la survie !

[5] je simplifie. En réalité, si les deux traitements sont équivalents (un témoin non traité, un traitement sans effet), on observera 0,5 fois sur 10 un écart de +23% de survie (ou mieux) en faveur du *traitement*, et 0,5 fois sur 10 un écart de +23% de survie (ou mieux) dans le *témoin non traité*. Le "1 chance sur 10" est donc la probabilité d'observer un écart de 23% (ou plus) *quel que soit le sens de l'écart.*

Ces informations étant connues, on peut maintenant tenter de conclure, et le moins que l'on puisse dire est qu'il n'y a pas de quoi pavoiser. Tout au plus peut on dire ceci :

(1) il est impossible d'affirmer ni d'infirmer un effet antibiotique de la pentacycline sur *Salmonella* dans les conditions de l'expérience;

(2) si l'effet antibiotique existe, il ne dépasse probablement pas +48% à la dose utilisée;

(3) un effet *négatif* modéré de l'antibiotique sur la survie (–2%) reste possible.

Il est évidemment hors de question de lancer ce produit sur le marché tout de suite. Ceci dit, si j'étais coincé sur une île déserte et en proie à une grave intoxication à salmonelle, j'utiliserais cet antibiotique sans hésiter, et au moins pour deux raisons évidentes (i) parce qu'il est à large spectre d'action, (ii) parce qu'il a peut être un effet très positif (+48% de taux de survie) et que dans le pire des cas il ne diminuerait mes chances que de 2%. Comme quoi, on peut voir du bon même dans les résultats les plus douteux.

Résumé du chapitre 1.

1. Quand vous comparerez deux moyennes ou deux proportions (pourcentages) issus d'une expérience de labo ou d'observations de terrain, vous observerez *toujours* une différence entre elles.

2. Au moins *une partie* de cette différence (et peut être même *la totalité* !) sera due au *hasard*, à cause d'un phénomène nommé *les fluctuations d'échantillonnage.* Les fluctuations d'échantillonnage sont *totalement inévitables,* aucune méthode, prière, ni juron ne pourra les faire disparaître. Les scientifiques passent, les fluctuations d'échantillonnage restent.

3. Avant de se précipiter vers la mauvaise conclusion, il est donc indispensable de calculer la fiabilité de vos moyennes (ou pourcentage) en calculant leur *intervalle de confiance*, et éventuellement de calculer la probabilité qu'un écart aussi grand puisse être observé simplement sous l'effet du hasard (*test statistique*).

4. L'usage des statistiques (une branche des mathématiques accessible à tous) est le seul moyen connu actuellement d'effectuer ces vérifications de manière objective, et selon une procédure reproductible par les personnes qui auront à examiner vos résultats et vos conclusions.

Voilà pourquoi les biologistes ont impérativement besoin de connaître au moins les bases des statistiques.

2. Présentez vos données

Pour résumer des données, la moyenne arithmétique semble être un choix naturel. Même le pire des cancres sait calculer que les notes 7/20, 8/20 et 9/20 donnent une moyenne de 8/20 en mathématiques. Oui, la bonne vieille moyenne arithmétique est un objet familier, et nous la choisissons spontanément pour résumer une série de données. Les scientifiques résument également leurs données de cette manière lorsqu'ils veulent les présenter. Cependant, ils veulent que le résumé de leurs données chèrement acquises soit le plus fiable possible, c'est pourquoi ils n'utilisent jamais une moyenne *seule*. Voici pourquoi :

2.1 L'île de la tentation

Supposons pour les besoins de la démonstration que vous soyez un étudiant de 22 ans rêvant de rencontrer l'âme sœur lors de vos prochaines vacances d'été. Nous supposerons de surcroît que le coût du voyage et de l'hébergement ne sont pas un problème (ceci est presque un ouvrage de mathématiques après tout alors autant éliminer tout réalisme et y aller carrément). Supposons cependant que, hélas, toutes les destinations soient complètes et qu'il ne vous reste plus que deux possibilités de lieux de vacances:

> Choix 1. *La Datcha du Corbeau Mort*, une paisible pension de famille dans la banlieue industrielle de Verkoïansk (Sibérie).
>
> Choix 2. *Surf Island*, une île paradisiaque baignée par des vagues superbes à quelques miles au large de Hawaii.

Réfléchissez bien. Oh, avant que j'oublie, voici une autre information:

> Moyenne d'âge des 252 hôtes de la Datcha du Corbeau Mort : 64 ans
> Moyenne d'âge des 248 hôtes de Surf Island : 22 ans

C'est un choix difficile. Mais je suppose que vous êtes parvenu à vous décider. Dans l'avion qui vous mène à destination, vous découvrez l'information suivante, imprimée en tout petits caractères en bas de la brochure distribuée par l'agence :

> Variance de l'âge des hôtes de la Datcha du Corbeau Mort : 1225 ans (au carré)
> Variance de l'âge des hôtes de Surf Island : 1209 ans (au carré)

De quoi peut il bien s'agir ? Et que représentent ces unités absurdes (des années *au carré* ?). Nous verrons plus loin comment on calcule une variance, mais pour le moment il suffit de comprendre que la variance mesure la *dispersion* des données autour de leur moyenne. Les valeurs ci-dessus ne vous disent probablement rien, mais vous apprendrez bientôt à reconnaître qu'elles sont *anormalement élevées*. Elles nous informent du fait

que dans les deux lieux de villégiature dont il est question, l'âge d'un hôte pris au hasard sera *très éloigné* de l'âge moyen des hôtes. En d'autres termes, les âges dans ces deux lieux ne sont probablement pas très regroupés à proximité de la moyenne. Serait-il possible (cela semble à peine croyable) que l'agence de voyage ne vous ait pas donné une idée très fiable de la situation réelle ?

L'explication est la suivante. La Datcha du Corbeau Mort est spécialisée dans le quatrième âge. La quasi totalité des pensionnaires a donc dépassé les 90 ans. « Minute ! » ferez vous remarquer « il est donc impossible d'obtenir un âge moyen de 64 ans ! ». Vous avez évidemment raison. La moyenne de 64 ans est atteinte grâce aux nuées d'infirmières (et d'infirmiers) âgés d'une vingtaine d'années qui s'occupent des pensionnaires. Ce personnel dévoué, sympathique, débordant de jeunesse et d'énergie, tue le temps comme il peut pendant les deux heures de sieste quotidienne de ses hôtes (et pendant les longues soirées, car les pensionnaires en question sont au lit vers 20h00).

Vous ne verrez cependant rien de tout ceci puisque – ne mentez pas – vous avez choisi d'aller à Hawaii, et votre petit avion-taxi est justement en train d'atterrir sur le charmant terrain de terre battue de *Surf Island*. Vous découvrez alors de coquets bungalows et une foule paisible de couples âgés de 40 ans environ, et leurs très jeunes enfants. *Surf Island* est en effet spécialisée dans les couples avec jeunes enfants (vous êtes la seule exception) d'où la moyenne d'âge de 21 ans. Souriez. Au moins vous allez échapper à la routine exténuante des vacances en bandes-de-jeunes-fêtards, et pourrez retourner à l'université plein de tonus pour étudier les sciences une année de plus. Et puis, sur *Surf Island*, il y a une véritable fortune à faire en tant que baby-sitter.

Rappelez-vous de vos vacances à *Surf Island* la prochaine fois qu'on vous résumera des données en vous donnant seulement une moyenne. Ce réflexe de méfiance deviendra automatique si vous vous lancez dans une carrière scientifique. Dans la vie de tous les jours, nous baignons dans ce qu'un de mes collègues décrit comme « la culture de la moyenne ». En sciences (tout spécialement en biologie, car les phénomènes biologiques sont si variables) vous rejoindrez la « culture de la variance » et deviendrez très circonspects face aux moyennes « toutes nues ».

Revenons à nos moutons : comment présenter des données de manière synthétique ? Comme nous venons de le voir, il est crucial de ne pas se fier à une moyenne seule, et de prendre aussi en compte la variabilité des données, qui conditionne à quel point leur moyenne est fiable. Il est facile de comprendre que, si les données sont étroitement groupées autour de la moyenne, celle-ci est fiable : elle donne une bonne idée des données. Par exemple, les ongles des mains poussent au rythme approximatif d'un demi millimètre par jour. Il ne s'agit que d'une moyenne, mais elle est fiable car la plupart des cas individuels se situent immédiatement aux alentours de ce chiffre (personne n'a d'ongles poussant de un centimètre par jour).

Au contraire, si les données sont largement dispersées, alors leur moyenne donne une assez mauvaise idée des données. Le cas extrême étant représenté par l'effet *Surf Island*,

dans lequel la plupart des données sont très éloignées de la moyenne ! Par exemple, si on tient compte de toutes les espèces, le poids moyen d'un mammifère adulte doit se situer aux environs de 1kg (et même probablement moins). A l'évidence, cette moyenne ne résume pas les données avec efficacité, car la masse d'un mammifère adulte se situe quelque part entre les 2 grammes de la musaraigne *Suncus etruscus* et les 150 tonnes de la baleine bleue *Balaenoptera musculus,* ce qui laisse tout de même une belle marge d'incertitude. Il se trouve que parmi les quelques milliers d'espèces de mammifères, il y a beaucoup d'espèces de petite taille (des rongeurs en particulier), ce qui fait que le poids *moyen* d'une espèce de mammifère est *faible*. Cette information *moyenne* ne serait cependant pas suffisante à vous rassurer complètement si, marchant en pleine nuit dans une forêt du Bengale, vous appreniez que "un mammifère" allait bientôt vous sauter dessus et vous mordre la nuque. Comment faire parvenir à votre lecteur une information claire sur la fiabilité de la moyenne que vous lui montrez, sans pour autant le noyer sous l'intégralité de votre jeu de données ? En utilisant un paramètre décrivant avec concision la dispersion des données autour de la moyenne. On peut imaginer plusieurs possibilité de le faire.

2.2 L'étendue

J'ai utilisé l'une de ces possibilités plus haut, lorsque j'ai indiqué la plus petite (2 grammes) et la plus grande valeur (150 tonnes) du jeu de données. L'écart qui les sépare se nomme *l'étendue* (*range*, en anglais). Les *étendues* sont utiles car elles donnent une première idée approximative de la situation, et ne nécessitent aucun calcul élaboré. Leur principal défaut est qu'une étendue repose uniquement sur les deux données les plus extrêmes, et reste totalement aveugle à tout ce qui se passe entre les deux. Pour reprendre l'exemple des mammifères, si toutes les espèces animales pesaient 1kg à l'âge adulte *sauf* les musaraignes et les baleines bleues, *l'étendue* des données resterait identique à ce qu'elle est aujourd'hui, alors que la dispersion des données serait devenue quasiment nulle. Pour cette raison, les étendues ne sont guère utilisées autrement que de manière descriptive et ne sont pas utilisées pour les tests statistiques (elles sont trop vulnérables à l'influence d'une seule valeur extrême, en particulier).

Si on n'utilise pas l'étendue, alors quoi ? Si vous deviez inventer à brûle-pourpoint un indice qui rende compte de la dispersion de données autour de leur moyenne, vous mesureriez probablement les écarts entre chacune des données et la moyenne. Pour synthétiser toute cette information, il serait alors naturel de faire tout simplement la moyenne de ces écarts. Une moyenne faible indiquerait sans doute des valeurs groupées et une moyenne élevée des valeurs dispersées ? A cette occasion, vous vous apercevriez cependant que la somme algébrique de ces écarts est... nulle (ce qui rappelle au passage que la moyenne est située en quelque sorte au barycentre des données). Vous contourneriez évidemment cet obstacle en un clin d'œil, en faisant la moyenne des

valeurs absolues des écarts. Vous auriez ainsi réinventé la formule de **l'écart moyen**, qui est bien le paramètre de dispersion le plus intuitif de tous :

$$e_{moyen} = (|x_1 - m| + |x_2 - m| + ... + |x_n - m|) / n$$

m : moyenne des données de l'échantillon
n : effectif de l'échantillon

Cette distance moyenne entre un point de donnée et la moyenne des données utilise la même unité que la variable mesurée. Dans le cas du *Corbeau Mort* par exemple, l'écart moyen aurait été d'environ 35 *ans* (un hôte choisi au hasard aurait en moyenne eu *35 ans de plus ou de moins* que la moyenne d'âge, qui était de 56 ans). Une indication claire que beaucoup d'hôtes étaient *soit* très âgés (56 + 35 = 91 ans) *soit* de jeunes adultes (56 – 35 = 21 ans). Il est fort dommage que l'écart moyen n'ait pas d'application statistique, car il a la même unité que la moyenne (il sera exprimé en kg si la moyenne est en kg) et cette caractéristique (en plus de son calcul très simple) le rend immédiatement compréhensible. Il n'en va pas de même d'un autre paramètre de dispersion moins évident a priori mais bien plus utilisé : la très redoutée **variance.**

2.3 La variance

La pauvre variance ne mérite vraiment pas sa réputation. Regardez sa formule :

$$\sigma^2 = [(x_1 - m)^2 + (x_2 - m)^2 + ... + (x_n - m)^2] / n$$

m : moyenne des données de l'échantillon
n : effectif de l'échantillon

Maintenant, comparez-là à la formule de l'écart moyen. C'est quasiment la même ! (pourquoi croyez vous donc que j'ai consacré du temps à vous présenter l'écart moyen ?). Dans le cas de la variance, le problème du signe des écarts à la moyenne a été éliminé en élevant ces écarts au carré (donc le signe est systématiquement positif, plus besoin de traîner des valeurs absolues). Dans le cas du *Corbeau Mort*, on a donc une variance de l'ordre de $(35 \text{ ans})^2 = 1225$ *années au carré*, ce qui explique à la fois le nombre élevé et l'unité bizarre vue plus haut.

Ce traitement (la mise au carré juste pour se débarrasser des signes négatifs) semble un peu excessif pour une question si banale, mais il sert en réalité à faire apparaître des propriétés mathématiques et géométriques intéressantes pour la suite des événements (traduire : la variance permet d'effectuer des tests statistiques, pas l'écart-moyen).

Pour finir, regardez à nouveau la formule de la variance, en notant "CE" le carré de l'écart entre une donnée *x* et la moyenne *m* des données, la variance est égale à :

$$(CE_1 + CE_2 + ... + CE_n) / n$$

Il semble que... mais oui, ma parole, la variance n'est rien d'autre qu'une banale... **moyenne arithmétique** ! C'est la moyenne des (carrés des) écarts séparant chacune des données x_1, x_2 (etc.) de leur moyenne *m*. Vous le voyez, rien de mystérieux là dedans,

rien de nouveau, juste une bonne vieille moyenne arithmétique mesurant la dispersion des données. Franchement, la *variance* mérite-t-elle vraiment d'inspirer la terreur ?

Notez que pour des raisons expliquées en **Annexe 1**, l'estimation s^2 (basée sur un échantillon) de la variance réelle σ^2 (inconnue) d'une population utilise $(n - 1)$ et non pas n au dénominateur. Vous utiliserez dont en pratique la formule suivante :

$$s^2 = [(x_1 - m)^2 + (x_2 - m)^2 + ... + (x_n - m)^2] / (n - 1)$$

La variance souffre d'un défaut ingrat : son unité ne « parle » absolument pas. En effet, la variance a la dimension d'un carré par rapport à l'unité de la variable mesurée. Par exemple, si vous mesurez la masse de vos individus en grammes, la variance aura la dimension g^2 (grammes au carré…), ce qui n'évoque pas grand chose pour un cerveau humain normal. Pour faire face à cet aspect un peu déroutant de la variance, on fait alors appel à un paramètre de dispersion plus « parlant », qui est *l'écart type* (ne pas confondre avec l'écart moyen).

2.4 L'écart type

C'est simplement ***s***, c'est-à-dire la racine carrée de la variance s^2 (ce qui permet de retomber sur ses pieds en terme de dimensions). L'écart type sera ainsi exprimé dans la même unité que la variable mesurée (des kg, des années etc.), ce qui est quand même plus confortable et facile à interpréter puisqu'on retrouve (à la racine de n près) la notion de l'écart moyen. Vous pouvez donc — *grosso modo* — considérer l'écart type comme l'écart qu'on observera en moyenne entre une donnée prise au hasard dans votre échantillon (et, par extension, dans la population) et la moyenne des données. L'écart type est donc un paramètre bien plus parlant que la variance.

2.5 Ecart type de la moyenne : l'erreur standard

La racine carrée de la variance permet de calculer l'écart-type *des données autour de leur moyenne*. Cependant, la synthèse ultime de vos données consiste à montrer à vos lecteurs l'écart-type *de la moyenne elle même*. Cet écart type reçoit un nom spécial (qui pemet de ne pas le confondre avec l'écart type des données) et devient *l'erreur standard* de la moyenne (abréviation : *e.s.* en Français et *s.e.* en Anglais). Si on appelle s^2 la variance des données de la population, alors l'erreur standard de la moyenne m obtenue à partir d'un échantillon de n individus est :

$$\text{erreur standard} = \text{racine}(s^2/n)$$

D'après mon expérience, il est très difficile de comprendre à première vue comment la moyenne d'un échantillon (valeur *unique,* il y a évidemment *une seule* moyenne par échantillon) peut avoir un écart-type, puisque cette notion est basée sur la moyenne de *plusieurs* écarts (sans parler de la question existentielle « un écart par rapport à *quoi* ? »). En conséquence, beaucoup d'étudiants confondent l'erreur standard (écart-type *de la moyenne*) avec l'écart-type *des données* (défini plus haut). Je comprend tout à fait leurs doutes, car cette notion n'est pas intuitive. Aussi, **l'Annexe 3 « L'erreur standard pour les débutants »** a été rédigée tout spécialement à leur intention. Elle

traite de ce problème lentement et pas-à-pas. N'hésitez pas à en faire usage, ou bien faites moi une confiance aveugle pour l'instant et continuez votre lecture tout de suite.

Les erreurs standards sont très importantes en sciences. Ce sont en effet *les valeurs représentées par les « barres d'erreur » que vous trouverez sur quasiment tous les graphiques scientifiques professionnels*. En voici un exemple.

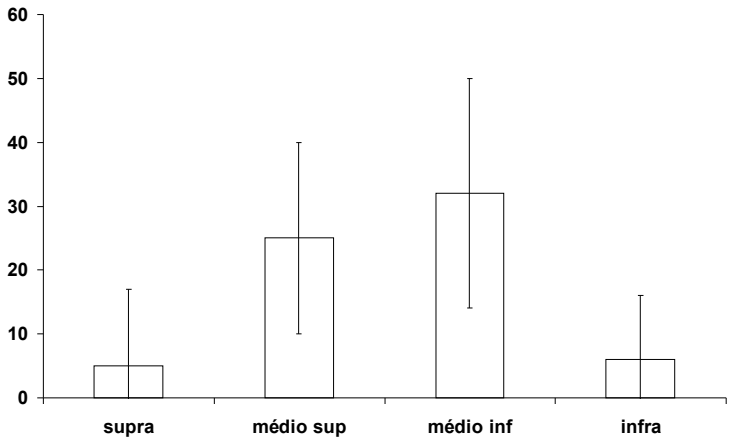


Figure 2.1. Densité (individus/m2) d'une population de Littorina littorea sur l'estran rocheux de Penvins (Morbihan) le long d'un transect en fonction des niveaux cotidaux. Supra : supra littoral. Médio : médio littoral, Infra : intra littoral. Barres : erreur standard. N = 528 individus.

Ce type de représentation donne à vos lecteurs une idée du crédit que l'on peut apporter à la précision des moyennes présentées sur le graphe. Plus la variance des données est importante, plus l'erreur standard (barre d'erreur du graphe) est grande, et moins la moyenne présentée est fiable. D'après la figure ci-dessus par exemple, il serait particulièrement ridicule de prétendre qu'il existe une différence de densité entre le médiolittoral inférieur et le médiolittoral supérieur.

2.6 Ecart-type d'un pourcentage : une autre sorte d'erreur standard

Les pourcentages aussi ont leur écart-type. Son calcul est simple puisqu'il ne nécessite que le pourcentage lui même et l'effectif sur lequel il est calculé :

$$\text{erreur standard} = \text{racine}\ (pq/(n-1))$$

Comme dans le cas de la moyenne, il est a priori difficile de comprendre comment un pourcentage (qui est *unique*) peut présenter un écart-type, qui est basé sur la moyenne de *plusieurs* mesures d'écart (de plus on se pose encore une fois la question : "écart par rapport à *quoi* ?"). Vous avez parfaitement raison de vous poser ces questions, et vous trouverez leur réponse détaillée dans **l'Annexe 3 « L'erreur standard pour les débutants »**.

Quoi qu'il en soit, vous pouvez dès maintenant calculer l'erreur standard de n'importe quel pourcentage, et vous en servir pour représenter les barres d'erreur sur vos graphiques. Comme dans le cas des moyennes, vous pouvez choisir de représenter vos

pourcentagess dans des tableaux ou directement dans le texte. Vous le ferez alors en utilisant le format « pourcentage ± s.e.". Par exemple, si le pourcentage en question est de 10% avec une erreur standard de 2% vous écrirez « 10 ± 2% » ou bien « 10% ± 2% ».

Nous verrons plus tard (chapitre 6) que l'erreur standard, même si elle a l'avantage d'être universellement utilisable et de dissiper l'illusion d'une valeur ponctuelle parfaite, donne encore une impression trompeuse de précision par rapport à la réalité. En effet, la zone (nommée **intervalle de confiance)**, dans laquelle se trouve "presque certainement" la véritable valeur du paramètre moyen (de la population étudiée) est la plupart du temps **environ deux fois plus large** que l'erreur standard.

Exemple 2.1 : Valeurs de l'échantillon : 1, 2, 3, 6.

Effectif : $n = 4$
Moyenne : $m = 3$

somme des carrés des écarts à la moyenne (SCE) :
$(1 – 3)^2 + (2 – 3)^2 + (3 – 3)^2 + (6 – 3)^2 = 14$

Variance *de l'échantillon* : SCE/n = 14/4 = 3,5 (sans intérêt pour nous)
Ecart type *de l'échantillon* : $\sqrt{}$ 3,5 = 1,871 (sans intérêt pour nous)

Variance estimée *de la population* : s^2 = SCE/(n – 1) = 14/3 = 4,667
Ecart type estimé *de la population* : $s = \sqrt{}$ 4,667 = 2,160
Erreur standard de la moyenne : e.s. = $\sqrt{(s^2/n)} = \sqrt{(4{,}667/4)}$ = 1,08

On peut donc écrire dans un tableau : « $m = 3 \pm 1{,}08$ »

Exemple 2 : fréquence observée de p = 0,20 sur 50 individus :

e.s. = $\sqrt{[pq/(n - 1)]} = \sqrt{(0{,}20 \times 0{,}8/49)}$ = 0,057

On peut écrire dans un tableau : « $p = 0{,}20 \pm 0{,}057$ (ou $20 \pm 5{,}7\%$) »

Fonctions à utiliser dans le tableur « Excel ». on suppose dans cet exemple que les 20 données de l'échantillon sont rangées dans les cases C1 à C20)

Paramètre à calculer	**Ecrire dans la cellule du tableur**
m = **Moyenne** *des données de l'échantillon* (c'est la meilleure **estimation de la moyenne** du caractère étudié *chez des individus de la population*	**=MOYENNE(C1:C20)**
Ecart moyen	=ECART.MOYEN(C1 :C20)
s^2 = **estimation de la Variance** du caractère étudié *chez des individus de la population*	**=VAR(C1:C20)**
s = **estimation de l'Ecart type** du caractère étudié *chez des individus de la population*	**=RACINE(VAR(C1:C20))** il existe une formule plus directe mais celle ci vous oblige à retenir ce qu'est l'écart-type
$\sqrt{(s^2/n)}$ =**Erreur standard de la moyenne**	**=RACINE(VAR(C1:C20)/20)** on divise ici par 20 car n = 20 données

Résumé du chapitre 2.

Les moyennes ne donnent aucune information sur la dispersion des données. C'est pourquoi elles doivent être complétées par une valeur rendant compte de cette dispersion, qui conditionne la fiabilité de la moyenne. Cette valeur est basée sur le calcul de la *variance*. La variance d'une série de données est la moyenne du carré des écarts séparant les données de leur moyenne, et elle est notée s^2. La racine carrée de la variance est l'écart type. L'écart type d'une moyenne est nommé *erreur standard* (voir **Annexe 3**) et abrégé « e.s. ». C'est la valeur représentée par les « barres d'erreur » des graphiques scientifiques. Dans les tableaux, ou dans le corps du texte, une moyenne sera toujours accompagnée de son erreur standard, sous la forme « moyenne ± erreur standard». Les pourcentages ne donnent pas non plus la moindre idée de leur degré de fiabilité, car un pourcentage ne vaut que par l'effectif sur lequel il est calculé. Il est donc impératif d'en tenir compte. Dans le cas d'un pourcentage p calculé sur n données, la variance est $p(1–p)/(n – 1)$ (voir **Annexe 3**). La racine carrée de cette variance est l'écart type du pourcentage. Il s'agit encore d'une erreur standard, qui est utilisée pour construire les barres d'erreur sur les graphiques scientifiques représentant des pourcentages. Dans les tableaux, ou dans le corps du texte, un pourcentage sera toujours accompagné de son erreur standard, sous la forme « pourcentage ± erreur standard». L'erreur standard peut toujours être calculée facilement (c'est son avantage). Son inconvénient pour les observateurs non avertis est qu'elle donne encore une impression trompeuse : les intervalles de confiance (cf chapitre 6) sont environ deux fois plus larges que l'erreur standard en général.

3. Observons quelques variables aléatoires sauvages

3.1 Définition d'une variable aléatoire

La définition d'une variable aléatoire dans un manuel d'introduction aux statistiques s'effectue traditionnellement en trois étapes hautement ritualisées. Dans la **première étape**, une définition mathématique rigoureuse est donnée. Comme cette définition est évidemment incompréhensible (sauf pour un mathématicien), une **seconde étape** est consacrée à des exemples très simples cherchant à illustrer cette définition incompréhensible. Ces exemples sont eux même immuables et débutent toujours par l'une des deux options suivantes (1) la pièce de monnaie, (2) le dé à six faces. On passe enfin à une **troisième étape**, dans laquelle on présente des exemples scientifiques réalistes. Comme le présent ouvrage est rédigé par un ancien élève polytraumatisé par les mathématiques, je vais me contenter de définir très vaguement une variable aléatoire comme "*quelque chose dont il est impossible de connaître le résultat à l'avance*". De plus, j'émettrai l'hypothèse selon laquelle vous savez *déjà* qu'une pièce de monnaie bien équilibrée a une chance sur deux de tomber sur pile, et qu'un dé (à six faces) non pipé a une chance sur six de donner le chiffre que vous avez choisi à l'avance. Nous pouvons donc passer directement à la **troisième étape.**

En gros, les variables aléatoires manipulées en sciences appartiennent à l'une des catégories suivantes :

- Toute grandeur physique qui peut être **mesurée** (mensurations et poids d'un organe ou d'un individu, rendement d'une culture, densité d'un matériau, résistance à la flexion, température, intensité lumineuse ou d'un champ magnétique...)
- Tout ce qui peut être **chronométré** (durée de développement, longévité, temps de réaction après un stimulus, temps nécessaire pour accomplir une tâche...)
- Tout ce qui peut être **compté** (nombre de pétales d'une fleur, nombre de bigorneaux dans un cerceau lancé au hasard, nombre de petits dans une portée...)
- Toute **proportion (=pourcentage) résultant d'un comptage d'individus** (proportion de gauchers, de mâles, de juvéniles, de malades, de survivants à un traitement toxique). Ce type de proportion résulte du comptage de n individus d'un type donné parmi un grand ensemble de N d'individus). Cette sorte de proportion est *fondamentalement différente* des proportions découlant d'une **mesure physique.** Par exemple la proportion d'alcool ("degré d'alcool") dans un breuvage, ou le "pourcentage de protéines" d'un aliment sont estimés grâce à une mesure physico-chimique, et non grâce à un comptage d'unités individuelles. Ce second type de "proportion" peut donc être assimilé à une **mesure physique** (voir le premier type de variable aléatoire)
- Tout critère qualitatif qui permet de **hiérarchiser** les individus (exemple : "grand, moyen, petit", " excellent, très bon, bon, moyen, médiocre, mauvais... ", "bon état, état moyen, mauvais état, entièrement détruit", "A, B, C, D, E").

Je dois sans doute en oublier, mais ces catégories permettent de ranger déjà pas mal de choses.

3.2 Examen de quelques variables aléatoires

Une variable aléatoire peut être synthétisée par sa moyenne et sa variance, en tout cas lorsqu'elle est numérique, mais ces indicateurs synthétiques ne remplacent pas les données elles mêmes. L'idéal est d'observer la manière dont se répartissent les valeurs des différents individus de la population étudiée ou, à défaut, comment se répartissent les valeurs obtenues dans l'échantillon sur lequel on travaille. Cette répartition est appelée *distribution* de la variable aléatoire.

3.2.1 L'âge des hôtes de *Datcha du Corbeau Mort* et de *Surf Island*

Voici par exemple (figures 3.1 et 3.2) la *distribution* des âges des hôtes de la Datcha du Corbeau Mort ou vous auriez pu passer vos vacances, et la distribution de l'âge des insulaires de Surf Island, ou vous avez finalement choisi de passer vos vacances (en fondant malheureusement votre décision sur une simple moyenne, une erreur que vous ne commettrez jamais plus).

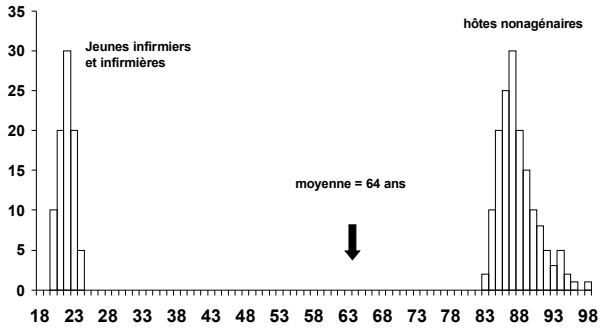


Figure 3.1. Distribution des âges des hôtes de la Datcha du Corbeau Mort

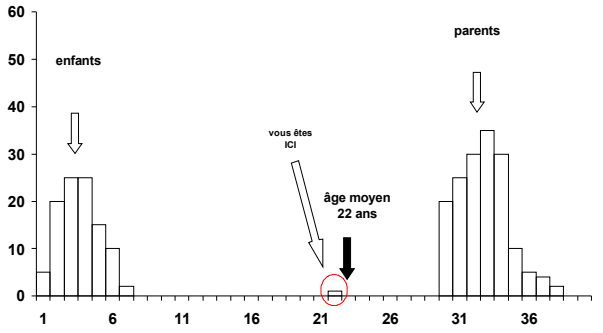


Figure 3.2. Distribution des âges des hôtes de Surf Island

Au vu de ces figures, la répartition très particulière des âges au sein de ces deux sites nous saute littéralement aux yeux. Techniquement parlant, ces distributions sont *bimodales* (c'est à dire qu'elles présentent chacune deux pics, ou *modes*). On peut même

difficilement faire plus bimodal que ça, puisque les pics en question sont carrément séparés par un vaste *no data's land*. D'une manière générale, une distribution bimodale attire notre attention sur le fait qu'il y a probablement deux sous-populations différentes (en ce qui concerne le caractère mesuré – ici, l'âge) au sein de la population dans laquelle nous avons échantillonné. Dans le cas de la Datcha du Corbeau Mort, les jeunes infirmiers/infirmières constituent une sous-population clairement distincte de la sous-population constituée par les pensionnaires très âgés de l'institution. Vous noterez que la moyenne d'âge (64 ans) est dans ce cas *particulièrement peu informative* : aucun des individus échantillonné ne s'en approche, même de loin...

Dans le cas de *Surf Island* également, la répartition des âges est bimodale jusqu'à la caricature. Cette fois, l'une des sous-populations est constituée des enfants, l'autre des parents. On remarque toutefois une donnée très particulière qui semble étrangère à l'une et à l'autre des sous-populations. Il s'agit naturellement de vous-même (si vous êtes un (e) étudiant(e) de 22 ans, comme je l'ai supposé à titre d'exemple).

3.2.2 La taille dans l'espèce humaine

Un exemple extrêmement classique (et plus sérieux) de distribution bimodale est la distribution des tailles des adultes dans l'espèce humaine. Parce que les femmes ont *en moyenne* une taille inférieure à celle des hommes, la distribution des tailles adultes, tous sexes confondus, doit être bimodale (un pic aux alentours de la moyenne des tailles des femmes, un pic aux alentours de la moyenne des tailles des hommes). Cet exemple est souvent employé dans les manuels d'introduction aux statistiques pour introduire la notion de bimodalité, et je l'ai moi même utilisé pendant des années avec bonheur, car il est très pédagogique et son bon sens saute aux yeux. Le seul problème est qu'il est faux. J'aurais pourtant dû m'en méfier, selon le dicton anonyme bien connu : "*Si une chose a été répétée souvent, partout, et à toutes les époques, alors il s'agit probablement d'une erreur*". En effet, la distribution (=répartition) des tailles adultes dans l'espèce humaine *n'est pas bimodale*, comme l'ont démontré récemment trois chercheurs iconoclastes (Schilling *et al*. 2002)[1] qui ont tout simplement pris la peine d'examiner suffisamment de données. Il ressort de leur étude qu'un mélange de deux distributions normales[2] ne peut apparaître bimodale que si l'écart $|m_1 - m_2|$ entre les moyennes (qui sont ici aussi les modes) de chaque distribution dépasse nettement la somme $(s_1 + s_2)$ des écarts-types des deux distributions[3]. Donc, on aura bimodalité seulement si

$$|m_1 - m_2| > (s_1 + s_2)$$

Or, lorsqu'on dispose de suffisamment de données, on constate que cette condition n'est pas remplie dans l'espèce humaine, et on ne peut donc voir qu'un seul pic. Encore un mythe qui s'écroule.

Comment tant de gens ont-ils pu se faire abuser ? C'est encore la faute des fluctuations d'échantillonnage. En effet, avec un échantillon suffisamment petit, les fluctuations d'échantillonnage peuvent facilement faire apparaître deux pics, donnant crédit à la (fausse) notion selon laquelle la distribution des tailles dans la population

[1] Schilling MF, Watkins AE & W Watkins, 2002. Is human height bimodal ? *The American Statistitian* **56**:223-229.

[2] on reviendra sur cette notion

[3] En fait c'est un peu plus compliqué (ça vous étonne ?). Il faut prendre en compte les proportions relatives de garçons et de filles dans l'échantillon, et le ratio entre les écarts-types de chacune des distributions. Passons sur ces détails (les curieux iront lire l'article de Schilling *et al.* 2002).

adulte humaine est bimodale. Voyons ce qu'il en est avec les données dont je dispose, qui m'ont été gracieusement transmises sur la base d'un questionnaire rempli par les étudiants de maîtrise au cours des quelques années pendant lesquelles j'ai eu le plaisir de leur dispenser la bonne parole statistique. Nous retrouverons ces données un peu partout dans cet ouvrage. Les figures 3.3 et 3.4 correspondent respectivement aux tailles auto-déclarées de ces étudiantes et étudiants âgés de 22 ans environ, (donc ayant normalement terminé leur croissance).

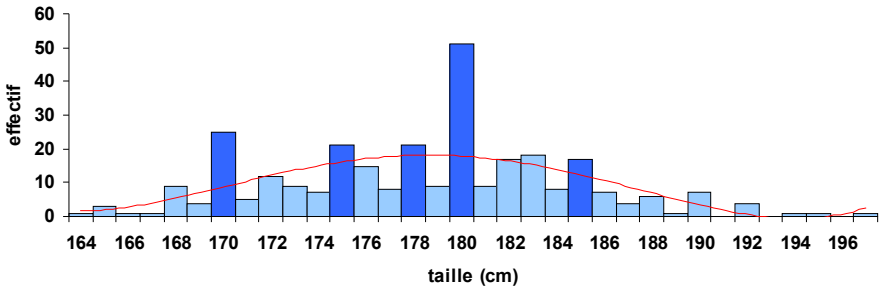


Figure 3.3. tailles auto-déclarées de 303 étudiants de maîtrise (garçons) On observe des "effets de seuils" nets, avec une fréquence anormalement élevée de déclarations pour certaines tailles.

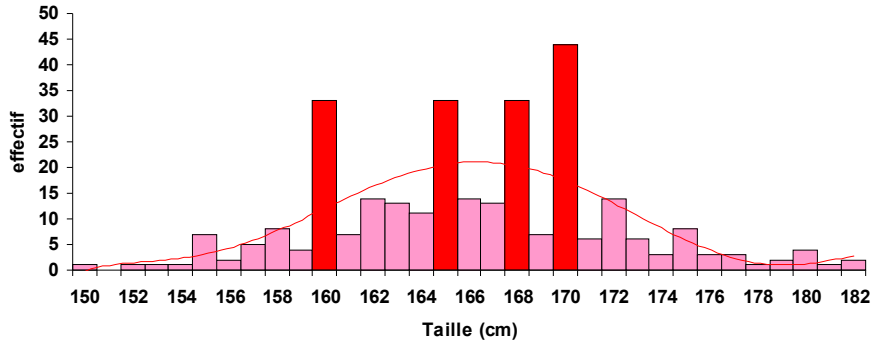


Figure 3.4. tailles auto-déclarées de 305 étudiantes de maîtrise. On observe des "effets de seuils" nets, avec une fréquence anormalement élevées de déclarations pour certaines tailles.

Plusieurs choses sautent aux yeux. La première est que ces étudiants trichent. Il est en effet complètement anormal d'avoir tant de filles déclarant mesurer exactement 1m60 alors que presque aucune ne déclare mesurer 1m59, de même le nombre de garçons mesurant soi-disant 1m80 est stupéfiant quand on considère que aucun ou presque ne déclare mesurer 1m79. L'accusation de tricherie est bien entendu un peu forte. Disons que certains connaissent leur taille approximativement, et ont tendance à donner un chiffre "rond". On constate le même phénomène dans les études anglo-saxonnes (Schilling *et al*. 2002), avec une abondance suspecte de garçons déclarant mesurer *exactement* six pieds (environ 1m82). Si l'on fait abstraction de ces artefacts, on constate que dans chaque sexe les tailles se répartissent à peu près harmonieusement de part et d'autre de la moyenne, qui est (en arrondissant au cm) de 1m66 chez les filles et 1m78 chez les garçons. L'écart entre ces deux moyennes est donc de 12cm. Si on combine ces

deux figures, on obtient la distribution des tailles tous sexes confondus (figure 3.5), qui fait apparaître... damned ! Une distribution bimodale ! (évidemment, il faut les yeux de la foi pour repérer de la bimodalité dans ce fouillis, mais il est tout de même difficile d'ignorer l'énorme pic à 1m70 et le non moins énorme pic à 1m80).

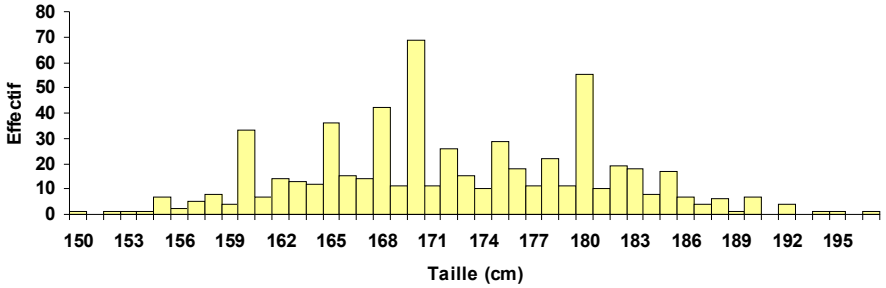


Figure 3.5. tailles auto-déclarées de 608 étudiants MBPE (303 garçons et 305 filles). On observe des "effets de seuils" nets (voir figures 3.1 et 3.2). ,

Vérifions le critère de Schilling. Les écarts-types sont ici de s_1=6,1cm chez les garçons et s_2=5,8cm chez les filles, d'ou $(s_1 + s_2) = 11,9$cm. Or, l'écart entre les moyennes des deux distributions originelles (178 – 166 = 12cm) est *de justesse* plus grand, ce qui suffit effectivement à faire apparaître une bimodalité. Alors, Schilling et ses collaborateurs racontent-ils n'importe quoi en prétendant que la taille des humains n'est pas bimodale ? Bien sûr que non. Nous sommes simplement dans le jeu des fluctuations d'échantillonnage. En effet, nos échantillons sont de taille très modeste comparés à l'enquête d'envergure nationale sur laquelle s'appuient Schilling et al.. Nos estimations de s_1, s_2, m_1 et m_2 sont seulement *approximatives*. Cette fois-ci nous voyons apparaître de la bimodalité, mais, si j'avais utilisé deux autres petits échantillons d'étudiants de maîtrise peut être aurions nous obtenu une courbe unimodale. Comme je sens que vous êtes dubitatifs (après tout, on *voit bien* les deux pics dans les données !), essayons d'avoir une idée de la précision de nos estimations en examinant les intervalles de confiance à 95% des paramètres estimés :

$IC_{95\%}$ de m_1 (en cm): [177,7 — 179,1]

$IC_{95\%}$ de m_2 (en cm): [165,6 — 166,8]

En simplifiant (le calcul correct est malheureusement plus compliqué), l'écart réel entre m_1 et m_2 pourrait être en réalité *aussi petit* que 177,7 – 166,8 = 10,9 cm, mais il pourrait être également *aussi grand* que 179,1–165,6 = 13,5 cm. En clair, nous ne connaissons pas du tout la différence de taille moyenne entre les garçons et les filles au dixième de centimètre près, comme on pouvait le penser, mais avec une incertitude de *plusieurs centimètres*. Notre capacité à affirmer avec force si $|m_1 - m_2| > (s_1 + s_2)$ ou pas est sérieusement compromise.

3.2.3 La longueur des graines d'érable

La taille des individus (ou des organes) est souvent distribuée selon une courbe unimodale bien particulière appelée "loi Normale", dans laquelle les données sont

réparties symétriquement de part et d'autre de la moyenne selon une courbe en cloche bien connue, la moyenne étant elle même le mode (la valeur la plus fréquemment observée). Ce fait est tellement habituel qu'on a tendance à faire cette hypothèse de "normalité" automatiquement chaque fois qu'on manipule une donnée biométrique, et on a souvent raison. Mais pas toujours. La figure 3.6 vous présente par exemple la distribution des tailles de 204 graines ailées d'Erable (ces graines qui tombent comme des hélicoptères), mesurées par mes soins au mm près avec un simple double décimètre.

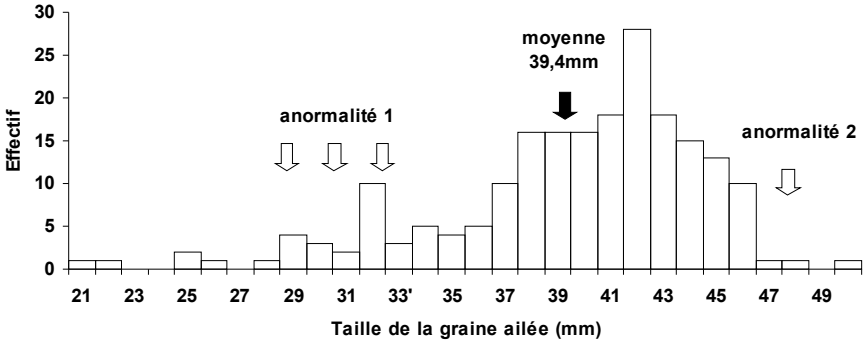


Figure 3.6. Distribution des tailles (mm) de 204 samares d'érable (*Acer sp.*). Il est clair que la distribution ne suit pas la loi normale.

Lorsque j'ai commencé à mesurer ces graines, je comptais en fait utiliser cet exemple pour illustrer une distribution normale, mais comme d'habitude la réalité biologique a été plus subtile que je l'anticipais. On constate en effet que la distribution n'est pas du tout *normale* (au sens statistique), c'est à dire qu'on n'a pas une courbe en cloche symétrique[4]. Cette distribution n'est pas normale pour au moins deux raisons, dont la première est triviale (je m'en suis aperçu très rapidement lors de la mesure des graines), alors que l'autre m'a pris complètement par surprise lorsque j'ai regardé le graphe. La première raison pour laquelle cette distribution n'est pas normale est que certaines graines avaient manifestement subi un gros problème de développement, et restaient rabougries (au point qu'on pouvait se demander si elles étaient viables). Il s'agit du groupe de valeurs entre 20 et 33 mm de long environ. Cependant, si jamais on élimine *arbitrairement* ces graines de la distribution, on ne rétablit pas la normalité de la distribution pour autant. En effet, sur la droite de la distribution vous constatez un deuxième phénomène spectaculaire : les effectifs *s'effondrent brutalement* lorsqu'on dépasse 46 mm de long, alors qu'ils sont encore élevés juste en deçà de cette valeur. Une véritable distribution normale aurait vu une décrue franche certes, mais progressive, ce "coup de hache" est beaucoup trop brutal pour être honnête. Je ne prétend pas avoir l'explication, mais je soupçonne fortement une *contrainte* au delà d'une certaine taille limite de la graine. Je sais en particulier (pour l'avoir lu dans des articles sur le vol) que la forme et la taille de ces graines volantes sont extrêmement optimisées. Je ne serais donc pas surpris de l'existence d'une taille *à ne surtout pas dépasser*. S'il y a des botanistes parmi vous, qu'ils se manifestent. Quoi qu'il en soit, j'ai quand même voulu utiliser cette petite mésaventure, car elle illustre bien l'intérêt de *regarder la distribution des données* pour vérifier si elles se conforment raisonnablement à l'hypothèse de départ, avant de se lancer bille en tête dans des calculs.

[4] et obéissant à une équation bien précise, sa majesté la Loi Normale n'est tout de même pas *n'importe quelle courbe en cloche symétrique*

3.2.4 Temps de développement de *Drosophila simulans*

Cela va peut être constituer un choc pour certains d'entre vous, mais il y a plusieurs dizaines d'espèces de drosophiles en dehors de la célèbre mouche du vinaigre *Drosophila melanogaster*. Parmi elles se trouve *D. simulans*, qui est d'ailleurs l'espèce jumelle de *D. melanogaster,* et c'est à elle que je dois ma thè*se*. La figure 3.7 illustre un des plus hauts faits d'armes de mon année de DEA sur le riant campus de Jussieu (Université Pierre et Marie Curie) vers la fin du siècle dernier.

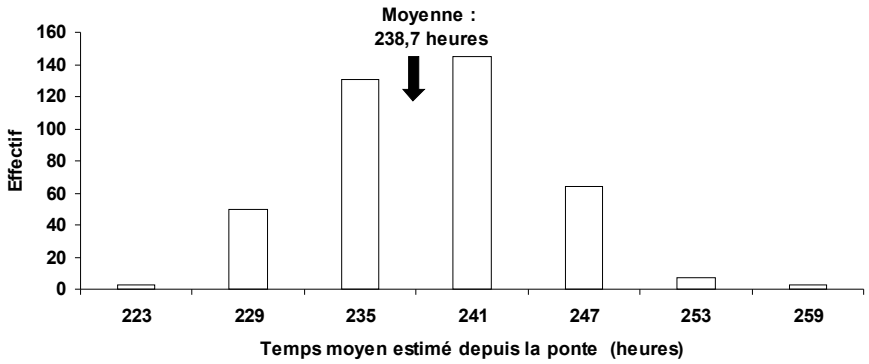


Figure 3.7. Distribution des temps de développement de l'oeuf à l'adulte chez *Drosophila simulans* (N = 403 individus).

J'ai choisi ce jeux de données parce qu'il se comporte (enfin !) à peu près selon une distribution "normaloïde", et aussi pour pouvoir étaler mes prouesses nocturnes[5]. Cette expérience héroïque consistait à mesurer le temps de développement de différentes souches de *D. simulans* selon qu'elle était infestée ou pas par certaines bactéries symbiotiques[6]. D'après ce que j'avais lu dans la littérature, les mesures de temps de développement chez la drosophile se faisaient en routine avec un relevé toutes les 12 heures pendant la période d'émergence (qui dure environ 48 heures à 25°C). Comme il me fallait deux bonnes heures entre l'aller retour domicile/labo et le comptage des émergents, c'était tout ce qu'il me fallait. Mais vous ne connaissez pas mon (adorable) directeur de thèse, et je me retrouvais donc finalement à faire des relevés toutes les 6 heures (24h sur 24, bien entendu) pendant cette fameuse période de 48 heures. Je peux donc témoigner du fait que Jussieu vers trois heures du matin est encore plus sinistre qu'en plein jour[7]. Résultat des courses : tout ça *pour rien* puisque je me suis aperçu que la bactérie en question ne modifiait pas du tout le temps de développement (ou alors de manière infime et sans intérêt). Heureusement pour moi, ce résultat "négatif" était intéressant en lui même, puisqu'on s'attendait plutôt au contraire.

Bref, le principal est que la distribution des temps de développements est à peu près symétrique autour de sa moyenne, avec une forme de courbe en cloche très acceptable. Nous avons enfin mis la main sur une variable qui se comporte (grosso-modo) en suivant une loi normale. Une petite remarque au passage concernant la précision obtenue. L'incertitude de *six heures* sur la mesure peut sembler très médiocre (une

[5] en tout bien tout honneur scientifique, naturellement.

[6] Pour les curieux, il s'agit de la bactérie *Wolbachia*..

[7] Et vous ai-je raconté la fois ou le chien des vigiles du campus m'a attaqué ? Ah... c'était le bon temps.

mouche comptée au temps t à pu émerger jusqu'à 6 heures plus tôt), mais elle représente seulement environ 2,5% du temps de développement. Par ailleurs, voyez vous-mêmes l'intervalle de confiance (à 95%) de la moyenne : [238,3—239,1 heures]. Autrement dit (grâce aux centaines d'individus), la précision sur l'estimation de la moyenne est de moins d'une heure !

3.2.5 Les graines de monnaie-du-pape

La monnaie du pape est ce végétal bien connu formant des siliques plates et translucides en forme de pièce de monnaie et qui ne semble exister que comme fleur séchée. C'est pourtant une plante comme vous et moi, la preuve, il y en a dans mon jardinet. C'est tout ce qu'il fallait pour me fournir facilement un exemple de plus. La monnaie du pape comporte théoriquement six graines par siliques. En réalité, bien entendu, il n'en est rien. Si on élimine du comptage les graines avortées toutes ratatinées et qui ne donneront jamais rien, le nombre de véritables graines dans une silique de monnaie du pape peut être n'importe quel nombre entier entre zéro et six. Voire plus. La figure 3.8 présente la situation à partir d'un échantillon de 210 siliques.

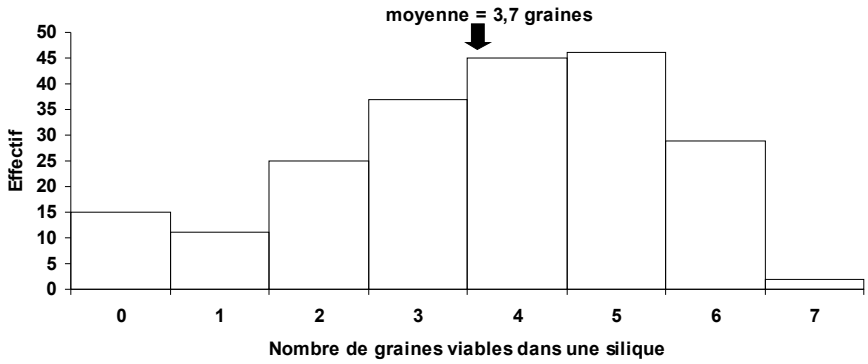


Figure 3.8. Distribution du nombre de graines dans une silique de monnaie-du-pape (N = 210 siliques examinées).

Premier coup d'oeil à la distribution : elle est *nettement dissymétrique*, en particulier à cause des valeurs relativement élevées dans la partie gauche. Ca n'a rien de surprenant : on savait à l'avance que cette distribution était bornée de manière rigide à droite (il ne peut théoriquement y avoir plus de 6 graines), et on s'attendait donc bien à observer une "queue de distribution" traînant sur la gauche. On constate au vu de ces résultats à quel point il peut être difficile en conditions naturelles de former le nombre de graines prévu par le manuel. Dans cet échantillon en tout cas, seules 14% des siliques sont parvenues à mener à maturité le nombre théorique de 6 graines, et une bonne moitié des siliques ne parviennent pas à former plus de trois graines dignes de ce nom. On remarque aussi que deux super-siliques ont réussi à aller là ou la théorie ne les attend pas, en formant 7 graines[8]. Les proportions citées ici ne sont évidemment pas à prendre au pied de la lettre, vu la taille modique de l'échantillon, elles ne représentent que des estimations vagues. Que pouvons nous dire sur la moyenne (qui est de 3,7 graines par silique) ? En premier lieu, bien sûr, on peut dire qu'elle recouvre une réalité bien plus complexe. Cependant, en tant que moyenne, elle représente un paramètre pratique à utiliser. Quelle

[8] ce qui me rappelle le fameux : "*Ces imbéciles ne savaient pas que c'était impossible, alors ils l'ont fait*"

est notre précision dans l'estimation de cette moyenne dans la population dont sont issues ces monnaies du pape ? Son intervalle de confiance (à 95%) est [3,4—3,9 graines], donc une précision d'une demi graine. Pas si mal.

3.2.6 Les pétales des matricaires

Les matricaires, (*Matricaria sp.*) sortes de "marguerites" sauvages, sont traditionnellement effeuillées[9] par les amoureux pour chanter la fameuse comptine "il m'aime, un peu, beaucoup etc...", Mais combien y a t-il de pétales sur une fleur de matricaire ? La question est d'importance, puisque si on connaît le mode de la distribution, il est possible de commencer la comptine de manière à maximiser ses chances de tomber sur "à la folie". Allons-nous rester les bras ballants face à un thème de recherche aussi stratégique ? J'ai donc mené une rapide enquête. Trop rapide, d'ailleurs, on va y revenir. Allons-nous trouver une belle loi normale ? La figure 3.9 vous dit tout.

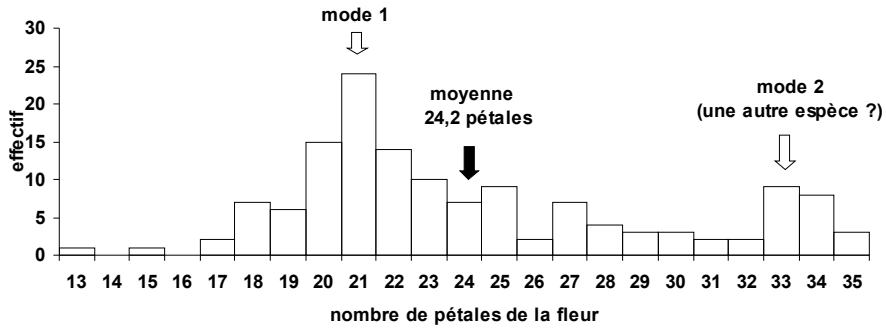


Figure 3.9. Distribution du nombre de pétales chez les fleurs de matricaires (Matricaria sp. (*N* = 139 fleurs examinées). La courbe est bimodale, et donc la moyenne ne veut pas dire grand chose

Caramba ! Encore raté ! Je vous avoue que j'anticipais une belle courbe en cloche qui aurait illustré une loi normale approximative avec des nombres entiers[10]. Nous tombons encore une fois sur une distribution bimodale de la plus belle eau, et je suis prêt à parier que sur le bord de la route de Rennes à Acigné il pousse au moins **deux** espèces de matricaires, l'une (à vue de nez, la plus fréquente) avec des fleurs à une vingtaine de pétales, et l'autre avec des fleurs avec une trentaine de pétales. Quoi qu'il en soit, les amoureux anglo-saxons qui utilisent la comptine très rudimentaire mais directe: "*she/he likes me—she/he likes me not*" ont tout intérêt effectivement à commencer par "she/he-likes me", puisque les deux modes observés dans la distribution (21 et 33) correspondent à des chiffres *impairs*. Je laisse le soin aux francophones de déterminer la stratégie gagnante avec la comptine bien de chez nous si jamais vous effeuillez une fleur de matricaire. Que dire de la moyenne observée dans ces conditions ? Qu'elle ne veut pas dire grand chose. Nous sommes dans une situation assez "Surf Islandesque", et d'ailleurs la variance élevée (26,3 pétales au carré) le montre bien. L'intervalle de confiance de la moyenne est théoriquement [23,4—25,1], mais *je m'empresse d'ajouter* que s'il y a réellement (comme je le soupçonne) deux espèces différentes sous cette distribution bimodale, calculer cette moyenne et son intervalle de confiance n'a

[9] avec les paquerettes, fréquentes sur les pelouses alors que les matricaires affectionnent les talus.

[10] la loi normale s'applique aux variables dites *continues*, c'est à dire qu'on peut découper à l'infini en tranches plus petites qu'une unité ex: 0,845124 grammes.

absolument aucun sens. Tout va fluctuer en effet en fonction des proportions relatives de l'espèce 1 et de l'espèce 2 qui auront été échantillonnées. Un travail plus sérieux serait à l'évidence de retourner sur le terrain avec une flore sous le bras, d'identifier clairement les espèces en présence, et d'étudier la distribution des nombre de pétales espèce par espèce : avis aux amateurs, je reste à l'écoute.

3.2.7 Un nombre choisi au hasard

Chaque année, je demande aux étudiants de maîtrise d'écrire sur un papier un nombre entier choisi au hasard entre zéro et 10 (inclus), pour les convaincre que le cerveau humain est strictement incapable de faire quoi que ce soit de manière aléatoire. Chaque année, les étudiants ont donc le choix entre {0, 1, 2, 3, 4, 5, 6, 7, 8, 9 et 10}, et le jeu consiste à faire tourner une roulette mentalement et à la stopper au hasard sur un chiffre. Les chers petits se livrent chaque année à l'expérience en étant naïvement persuadés d'être plus fort que ceux de l'année précédente (qui ont piteusement échoué).

Prenons un point de référence, pour mieux goûter le sel des résultats que vous allez voir. La figure 3.10 illustre ce qui se passe lorsqu'on demande *à un ordinateur* de s'acquitter de cette tâche en tirant de manière aléatoire 150 fois dans une distribution uniforme comportant tous ces nombres, chacun ayant une chance sur 11 d'être choisi à chaque tirage.

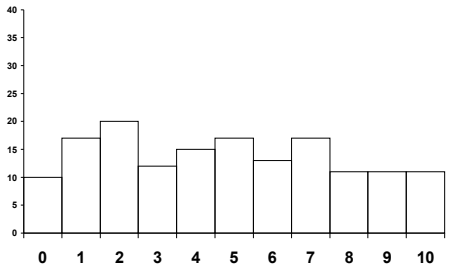


Figure 3.10 Distribution de 150 nombres entiers tirés aléatoirement par ordinateur entre zéro et 10. Les différences d'effectifs observées sont dues uniquement aux fluctuations d'échantillonnage.

Première constatation, certains chiffres (comme 2 en particulier) sont sortis davantage que la moyenne attendue (qui est de 14 fois environ, soit 150/11), et d'autres (comme 0) semblent boudés. Les gens mal informés en concluraient que le tirage aléatoire informatique fonctionne mal. Les gens bien informés (donc, vous) savent en revanche maintenant que ce phénomène est tout à fait normal et se nomme "les fluctuations d'échantillonnage". Il aurait été même *hautement suspect* que tous les chiffres soient tirés 14 fois exactement. Autrement dit, si je recommençais l'expérience avec un nouveau lot de 150 tirages, j'obtiendrais autre chose. Tiens, pendant qu'on en parle, faisons le (figure 3.11)

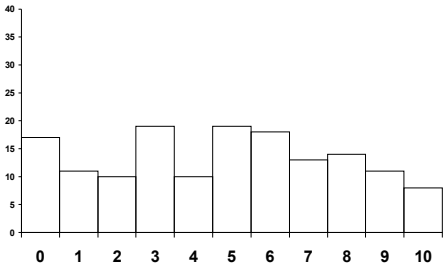


Figure 3.11 Distribution de 150 nombres entiers tirés aléatoirement par ordinateur entre zéro et 10 (même chose que figure 3.10 mais avec 150 nouveaux tirages). On observe toujours des fluctuations d'échantillonnage avec certains chiffres sortant plus que les autres, mais ce ne sont pas les mêmes.

Vous êtes maintenant convaincus que mon ordinateur n'a pas d'inclinaison particulière vers le chiffre 2, et ne boude pas particulièrement le chiffre 0 comme le tirage précédent pouvait le faire soupçonner. Mais en fait vous ne lisez pas ces lignes car vous vous êtes déjà précipités sur la spectaculaire figure 3.12 qui vous montre les tirages "aléatoires" obtenus par les promotions 1999 à 2001 de la maîtrise BPE de Rennes 1.

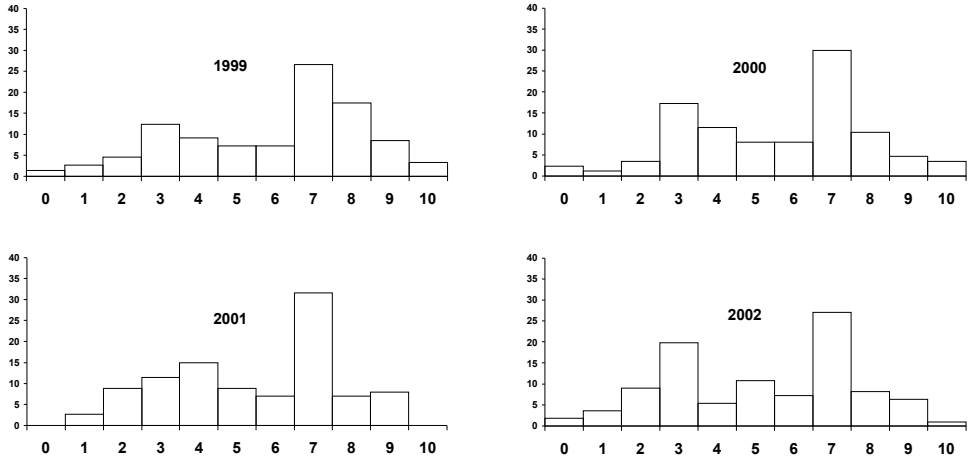


Figure 3.12 Distribution des nombres entiers choisis "au hasard" par des étudiants de maîtrise BPE de Rennes 1 des promotions 1999 à 2002. On constate que la distribution n'est manifestement pas uniforme, parce que les étudiants sont en fait incapables de choisir véritablement "au hasard" (le chiffre 7 est tout particulièrement sur-représenté) alors que les chiffres 0 et 10 sont visiblement évités.

Spectaculaire est le mot. Il saute aux yeux des moins avertis qu'il s'est passé quelque chose de *fondamentalement* différent d'un tirage aléatoire, car cette distribution s'écarte violemment de la distribution uniforme attendue. Sans en avoir conscience, ces étudiants ont fait des *choix non aléatoires*. L'expression "choisir de manière aléatoire" est d'ailleurs une contradiction absolue. On peut maintenant s'amuser à faire de la psychologie à la petite semaine et essayer de comprendre ce qui a pu se passer dans leurs têtes. Clairement, les bords sont boudés (peut être l'idée que "si c'est au bord, c'est à un endroit spécial donc pas choisi au hasard"). On remarque aussi que la valeur 5, qui est "au milieu", est soigneusement évitée, probablement en suivant le même

raisonnement inconscient. Enfin, les valeurs impaires (qui ont un petit coté irrégulier, donc "aléatoire") sont préférées aux valeurs paires "équilibrées". Résultat, les "chiffres de la chance", les "chiffres sacrés", le 3 et le 7 dans la plupart des civilisations, sont plébiscités. Si de véritables psychologues lisent ces lignes, je serais heureux de recevoir leur avis (évidemment plus pertinent) sur la question. Mon propos était simplement ici de vous convaincre que nous ne sommes pas faits pour *générer* de l'aléatoire. Ceci explique peut être que nous ayons parfois du mal à analyser des phénomènes dans lequel l'aléatoire jour un rôle important (l'analyse statistique des résultats, pour prendre un exemple "au hasard"). Dans le même ordre d'idée, mes étudiants ont toujours répondu *majoritairement* "oui" à une question qui était "*répondez par oui ou par non au hasard*.". Ce phénomène est bien connu des hommes politiques, qui l'exploitent à chaque référendum : vous ne verrez **jamais** de référendum dans lequel la réponse "oui" ne soit pas celle qui arrange le gouvernement. En effet, quand nous sommes indécis ou que nous n'avons pas la moindre idée de la réponse à une question, nous avons tendance en moyenne, à y répondre plutôt par *oui* que par *non*...

Résumé du chapitre 3.

Les variables aléatoires existent, et elles sont tout autour de nous. La manière dont les valeurs de ces variables aléatoires sont réparties se nomment distributions, et elles ne prennent pas n'importe quelle forme. Certaines distributions sont assez fréquentes en biologie, et présentent une courbe en cloche. La plus connue est la loi normale, mais il ne faut pas croire que toute variable biologique sera automatiquement distribuée selon une loi normale. En effet, certaines contraintes peuvent tronquer la distribution et provoquer un étalement plus important vers les valeurs faibles ou fortes de la variable aléatoire. La présence de plusieurs pics (ou modes) dans une distribution doit nous faire soupçonner que plus d'une population ont été échantillonnées. Enfin, retenez que l'esprit humain est très malhabile pour générer ou manipuler des événements aléatoires, car il a toujours tendance à rechercher ou à créer des motifs particuliers, ce qui provoque des choix (conscients ou inconscients) et non pas des tirages réellement aléatoires. C'est une des raisons pour lesquelles la randomisation (le tirage au hasard) est très importante en sciences, et c'est aussi la raison pour laquelle on utilise des méthodes statistiques pour analyser les données au lieu de se fier uniquement à nos premières impressions.

4. Tripatouillons les données

On a dit beaucoup de mal du tripatouillage de données. C'est un tort. Utilisé à bon escient, c'est une pratique très recommandable. Je ne connais pas de meilleur moyen de comprendre comment se comportent les variable aléatoires, et les paramètres que nous calculons pour essayer de les cerner. Dans ce chapitre, nous allons donc tripatouiller les données sans honte, et observer ce qui se passe, c'est-à-dire rien de grave si c'est fait dans les règles.

Le tripatouillage de données est un sport, qui comporte des figures imposées. Parmi celles-ci, les principales sont : (1) l'élimination des données qui ne nous plaisent pas, (2) la transformation des données en ajoutant, en retranchant, en multipliant ou en divisant par une constante *C* arbitraire, (3) la transformation des données en utilisant un traitement plus exotique (racine carrée, log, arcsinus racine etc.), qui ne sera pas abordée ici. Une fois qu'on maîtrise ces gammes, on peut passer à la vitesse supérieure, et se mettre à tripatouiller plusieurs variables aléatoires à la fois. Dans l'ivresse de la création, on peut ainsi créer de nouvelles variables aléatoires en en combinant plusieurs autres de différentes manières licites ou illicites.

4.1 Eliminons les données qui nous dérangent

Commençons par la base du tripatouillage amateur : l'élimination des données qui ne nous plaisent pas. Pour comprendre d'où vient cette tentation, il suffit d'observer le comportement de la moyenne et de la variance lorsqu'une donnée extrême (dite "aberrante") entre en jeu. L'échantillon A est puisé au sein de l'exemple des graines ailées d'Erable (il s'agit ici de longueur en mm) :

$$A = \{21 \text{——} 36, 37, 38, 39, 40, 41, 44\} \; ; \; m_A = 37{,}0 \text{ mm} \mid s^2_A = 48{,}0 \text{ mm}^2$$

La moyenne générale est m = 37,0 malgré le fait que la plupart des données *dépassent* cette valeur. La responsable est évidemment la valeur [21], anormalement faible, qui tire la moyenne vers le bas. Ça n'est pas son seul crime. La variance de 48,0 mm^2 implique un écart-type de 6,9 mm, ce qui est anormalement élevé considérant que (mis à part le [21]) la plupart des données sont très proches les unes des autres. Eliminons maintenant [21] du jeu de données.

$$B = \{\sout{21}, 36, 37, 38, 39, 40, 41, 44\}. \; m_B = 39{,}3 \text{ mm} \mid s^2_B = 7{,}2 \text{ mm}^2$$

On obtient une moyenne nettement plus représentative du "coeur" de la distribution. La variance s'effondre brusquement (elle est presque *divisée par neuf !*) soit un écart-type de s = 2,7 mm seulement. Cela montre bien l'influence totalement disproportionnée de l'unique point de données qui s'écartait nettement des autres. Cette influence est due au fait que la variance repose sur la moyenne d'écarts à la moyenne *élevés au carré*. Vis-à-vis de la variance, il faut donc voir la distance séparant une donnée de la moyenne comme une sorte de *bras de levier*, qui démultiplie l'influence de ce point sur la variance globale. C'est pourquoi il suffit d'une donnée vraiment extrême pour non seulement déséquilibrer une moyenne mais surtout pour affoler la variance. En quoi est-ce néfaste ? Examinons avec quelle précision nous avons estimé la moyenne de la

population dont est issu notre échantillon. Selon que [21] est inclus ou exclu des données, les intervalles de confiance de *m* sont, respectivement :

IC_{95} (+21) = [32,2 ——————— 41,8 mm], incertitude de 9,6 mm

IC_{95}(~~21~~) = [37,3—41,3 mm], incertitude de 4 mm

La donnée extrême [21], à elle toute seule, *multiplie par plus de deux* l'incertitude sur la moyenne générale de la population. Voilà pourquoi il est tentant d'éliminer les résultats dits "aberrants" des jeux de données, une pratique totalement condamnable, ou au contraire... complètement justifiée, selon l'origine de l'aberration, et la manière dont l'élimination est pratiquée.

- La *pire* manière de procéder est l'élimination *clandestine, arbitraire, effectuée après l'analyse statistique*. Ce type de tripatouillage malhonnête consiste tout simplement à éliminer une (ou plusieurs) données parce que ça vous permet d'obtenir le résultat que vous espériez, sans aucune autre justification. C'est évidemment inacceptable, et restera toujours une affaire entre vous et votre conscience.

- Une manière *licite* d'éliminer des données extrêmes consiste en revanche à procéder *lors de l'examen préliminaire des données brutes avant analyse*. Cet examen des données est du reste indispensable, et l'idéal est de procéder graphiquement. Ainsi, vous pourrez repérer instantanément les valeurs *véritablement* aberrantes, c'est à dire celles *qui résultent manifestement d'une erreur de mesure ou de transcription* (exemple : une graine d'érable qui mesurerait soi-disant 315 mm au lieu de 31,5mm parce que la virgule n'a pas été saisie). Ces données, résultant d'une erreur *grossière et manifeste* sont les seules que vous pouvez éliminer avec la conscience tranquille à ce stade tardif de l'étude, si vous n'avez plus accès aux individus eux mêmes. Si vous avez encore accès aux individus physiques constituant votre échantillon, rien ne vous interdit bien entendu de *vérifier matériellement* la véracité des mesures qui vous semblent anormalement faibles ou élevées.

- Il est en revanche possible d'éliminer n'importe quelle donnée pour n'importe quelle raison pour faire face à une situation non anticipée apparaissant *au cours de la collecte des données*, mais *à condition* encore une fois de le faire au grand jour et selon un critère clair (exemple : "cette graine a visiblement subi une déformation anormale en cours de développement, je l'élimine (elle et les autres graines du même type) de mon jeu de données"). Ce type d'approche vous obligera éventuellement à revenir en arrière pour ré-examiner les individus précédents. Bien entendu, il faudra *signaler explicitement* lors de la présentation des résultats que les données de ce type ont été éliminées (elles peuvent d'ailleurs être non pas éliminées totalement mais *traitées à part*).

- Enfin, la *meilleure* manière d'éliminer des données consiste à le faire... *avant* la collecte elle même, en décidant de critères d'exclusion *à priori*. Exemple : "toutes les graines rabougries ayant manifestement subi de gros problèmes de développement ou des attaques de parasites ne seront pas prises en considération". Cette approche sera grandement facilitée, si vous avez effectué une étude pilote, ou une petite reconnaissance sur le terrain avant de vous lancer dans la grande manip très importante que vous préparez.

En bref, éliminer des données n'est pas forcément un tripatouillage diabolique démontrant une éthique scientifique douteuse. Cela peut au contraire résulter d'un choix

transparent et justifié, qui prouve que vous prenez soin de collecter des données adaptée à la question précise que vous vous posez. Le tout est d'annoncer clairement la couleur sur ce que vous avez fait.

4.2 Transformons les données en utilisant une constante *C*

Ajouter ou retrancher une constante *C* aux données revient à déplacer la distribution en bloc le long de l'axe des abcisses. Multiplier ou diviser les données par une constante équivaut à changer d'unité de mesure (changement d'échelle). On peut même combiner les deux opérations. Ces changements auront bien sûr une influence sur la moyenne et éventuellement sur la variance. Pour l'illustrer, on va re-utiliser nos 206 mesures de longueur de graines ailées d'érable (variable aléatoire "L"). Premier essai : on ajoute ou on retranche une constante, disons 10mm, à chacune des 206 longueurs mesurées. Que vont devenir la moyenne et la variance ? Réponse :

Variable aléatoire	Moyenne (mm)	Variance (mm^2)
L originale	39,4	25,3
L + 10mm	49,4	25,3
L – 10mm	29,4	25,3

Le contenu de la colonne **Moyenne** n'étonnera personne. En revanche, celui de la colonne **Variance** peut surprendre et mérite d'être commenté. *Ajouter (ou soustraire) aux données une constante C ne modifie pas la variance* parce que cela équivaut simplement à déplacer la distribution « en bloc », de *C* unités le long de l'abscisse. La dispersion des données autour de leur moyenne est inchangée (puisque la moyenne s'est déplacée de *C* unités dans le même sens que les données), donc la variance reste la même.

On peut donc écrire, d'une manière générale, pour une variable aléatoire X de moyenne μ et de variance σ^2

Variable aléatoire	Moyenne	Variance
X	μ	σ^2
$X + C$	$\mu + C$	σ^2
$X - C$	$\mu - C$	σ^2

Supposons maintenant que la norme européenne EN2548-518/B, régissant la taille réglementaire des graines ailées d'érable[1], stipule que cette longueur doive être exprimée non pas en millimètres mais en dixième de millimètres, cependant que la norme américaine US24518-5681 exige que la longueur des graines d'érable soit exprimée en centimètres. Si nous voulons conserver nos crédits de recherche européens, tout en pouvant exporter nos graines d'érable vers les USA, nous devons adapter nos données en multipliant L respectivement par 10 (dans nos documents européens) et en le divisant

[1] Elle n'existe pas encore, mais le temps travaille contre nous.

au contraire par 10 (dans les documents à l'attention des douanes des USA). Que vont devenir la moyenne et la variance ?

Variable aléatoire	Moyenne	Variance
L	39,4	25,3
$10 \times L$	394,0	2530,0
L / 10	3,94	0,253

La moyenne se comporte comme attendu. Encore une fois, la surprise vient de la troisième colonne. La variance semble réagir de manière *exagérée*. Lorsque les données sont multipliées par 10, la variance est multipliée par 100, et lorsque les données sont divisées par 10, la variance est divisée par 100. Il suffit cependant de revenir à la définition de la variance (une moyenne d'écarts *élevés au carré*) pour comprendre facilement pourquoi la constante *C* doit être élevée au carré dans le calcul de la nouvelle variance[2]. Donc, de manière générale on peut écrire :

Variable aléatoire	Moyenne	Variance
X	μ	σ^2
$C \times X$	$C \times \mu$	$C^2 \times \sigma^2$

Ces principes de base étant posés, on peut alors remarquer que pour *une* variable aléatoire X donnée, sa véritable moyenne μ et sa véritable variance σ^2 sont uniques, ce sont donc des *constantes*. On peut donc les manipuler exactement comme la constante *C* du tableau ci-dessus. Ceci permet de passer à quelques cas particuliers très intéressants pour la suite des opérations, en appliquant simplement les principes que nous venons de voir :

Variable aléatoire	Moyenne	Variance
X	μ	σ^2
$X - \mu$	**0**	σ^2
X/σ	μ/σ	**1**
$(X - \mu)/\sigma$	**0**	**1**

L'opération figurant sur la deuxième ligne consiste à retrancher à chaque donnée la moyenne μ et s'appelle un *centrage*. Elle constitue simplement un changement d'origine, qui place la moyenne de la distribution au point 0 de l'axe des abcisses.

L'opération figurant sur la troisième ligne consiste à diviser toutes les valeurs par l'écart-type et s'appelle une *réduction*. Elle représente un simple *changement d'unité* : les valeurs de X ne sont plus exprimées directement dans l'unité d'origine (des mètres, des kg…), mais en écarts-types (σ). L'écart-type étant lui même exprimé en unités

[2] : détail du calcul en annexe 1 pour les incrédules

d'origine (mètres, kg etc...), la division des valeurs par l'écart-type rend le résultat final *sans unité* (kg/kg = sans dimension). Une conséquence intéressante de la réduction est que la variance vaut automatiquement 1 (regardez dans le tableau précédent en remplaçant C par $1/\sigma$ et vous comprendrez rapidement pourquoi).

La combinaison du *centrage* et de la *réduction* (quatrième ligne du tableau) donne une variable *centrée-réduite*. Cette opération est particulièrement intéressante dans le cas d'une loi symétrique par rapport à sa moyenne (cas de la loi normale par exemple), car elle permet de ramener n'importe quelle loi de ce type à une courbe unique calibrée sur laquelle on peut calculer des probabilités par simple lecture dans une table faite une fois pour toutes.

Et maintenant osons ! Osons manipuler plus d'une variable aléatoire à la fois (à propos des mêmes individus), et voyons si le ciel va nous tomber sur la tête.

On pose les deux variables X_A suivant une loi quelconque de moyenne μ_A et de variance σ^2_A, et X_B suivant une autre loi quelconque (pas forcément le même type de loi que X_A) de moyenne μ_B et de variance σ^2_B. On va de plus faire l'hypothèse que les deux variables sont *indépendantes*. Cela signifie concrètement que connaître la valeur de la variable X_A chez un individu ne permet absolument pas de prédire quelle sera sa valeur pour X_B. Dans le cas contraire, on dit que les variables sont *liées*, ou *corrélées*, et les égalités suivantes restent valables pour les moyennes mais doivent être modifiées en ce qui concerne les variances (on doit y ajouter un terme qui est la *covariance*). On verra ça plus en détails dans le chapitre 12 traitant de la corrélation.

Variable	Moyenne	Variance
X_A	μ_A	σ^2_A
X_B	μ_B	σ^2_B
$X_A + X_B$	$\mu_A + \mu_B$	$\sigma^2_A + \sigma^2_B$
$X_A - X_B$	$\mu_A - \mu_B$	$\sigma^2_B + \sigma^2_B$

Tout ceci appelle bien sûr quelques commentaires. Pourquoi *additionner* les variances quand il s'agit d'une *soustraction* de deux variables aléatoires ? Parce que la nouvelle variable aléatoire est certes le résultat d'une soustraction, mais d'une soustraction pour laquelle chacun des deux termes est aléatoire. Le résultat final ne peut donc en être que *plus* variable qu'au départ (il serait proprement miraculeux que retrancher une valeur au hasard d'une autre valeur aléatoire *diminue* le caractère aléatoire de l'ensemble !).

Pour vous en convaincre, réfléchissez à la moyenne et à la variance de la variable aléatoire "$X_1 - X_2$" (autrement dit, la différence entre deux tirages successifs dans la *même* loi, exemple : entre deux lancers du même dé). Certes la *moyenne* de cette variable "$X_1 - X_2$" est bien zéro, car chacun des deux tirages aura la même moyenne en espérance. Cependant, vous savez très bien qu'on obtiendra la plupart du temps deux valeurs de lancer différentes (sinon les casinos feraient faillite) donc $x_1 - x_2$ s'écartera en général de la moyenne attendue qui vaut zéro. Or, qui dit écarts à la moyenne dit *variance*. Ainsi, la variance de "$X_1 - X_2$" n'est clairement pas nulle, alors que cette variable est constituée par une soustraction entre deux variables aléatoires ayant exactement même moyenne et même variance .

Histoire de bien enfoncer le clou, on peut ajouter que soustraire des variances entre elles pourrait vous amener à obtenir parfois des variances *négatives*. Cela serait particulièrement ennuyeux quand on se souvient qu'une variance est formée de la moyenne d'écarts *élevés au carré*, et plus encore si on a compris qu'elle mesure une *dispersion*. A moins d'entrer dans la quatrième dimension, il parait difficile de se disperser *négativement* autour d'une valeur. Retenez donc ceci, ça peut servir : **une variance négative, ça n'existe pas.**

Une fois ces notions d'opérations sur les variables aléatoires claires dans votre esprit, la manipulation de la loi normale à l'endroit, à l'envers et dans le sens des aiguilles d'une montre,-+ ainsi que les tests statistiques en général ne devraient plus vous poser de problèmes insurmontables. Vous avez donc intérêt à bien méditer ces opérations, qui reposent en fait toutes sur la compréhension de deux notions simples : la *moyenne et la variance* (qui est elle même une simple moyenne d'écarts élevés au carré).

Résumé du chapitre 4.

Les tripatouillage des données (suppression de certaines données, troncature en deçà ou au delà d'une certaine valeur) est condamnable lorsqu'il est fait clandestinement. Effectué au grand jour et de manière justifiée, il peut au contraire être utile. Changer d'unités, ajouter ou retrancher une constante ne posent aucun problème et peuvent faciliter l'analyse : le centrage et la réduction sont des opérations non seulement licites mais largement utilisées et très utiles. Les règles de combinaison des variables aléatoires *indépendantes* constituent un cas particulier, d'où on peut retenir ces deux règles : les moyennes s'additionnent et se soustraient, les variances s'additionnent mais ne se soustraient jamais. Les transformations plus complexes (passage au log, arcsinus racine etc.) non abordées ici, ont également leur intérêt pour linéariser des fonctions courbes ou "normaliser" des données ne suivant pas la loi normale, et faciliter ainsi l'analyse.

5. Lois statistiques à connaître en biologie

Ce chapitre est probablement le plus soporifique de tout cet ouvrage (c'est vous dire...). Vous pouvez évidemment le contourner et aller voir plus loin des choses plus concrètes et importantes, par exemple comment on calcule les intervalles de confiance. Vous réaliserez cependant tôt ou tard que vous avez besoin des informations qu'il contient.

5.1 La loi binomiale

Il y a en fait deux lois binomiales: la positive et la négative. Lorsqu'on ne précise pas de laquelle on parle, il s'agit toujours de la loi binomiale *positive*. C'est la loi suivie par les résultats de tirages aléatoires, lorsqu'il n'y a que *deux possibilités mutuellement exclusives* de résultats (ex: mâle ou femelle, vivant ou mort, fromage ou dessert[1]) et que la probabilité d'obtenir chaque possibilité est *constante* au cours de l'expérience (ce qui ne veut pas dire *égale* entre l'une et l'autre). Cela sera le cas dans deux sortes de situation : soit quand la population est de taille infinie, soit quand on effectue le tirage avec remise. En effet, s'il n'y a pas remise, le tirage d'un individu modifie la probabilité de tirer un individu de ce type la fois suivante (puisqu'il y en a un de moins disponible). Dans le cas de l'échantillonnage en situation réelle sur le terrain, on peut presque toujours faire l'hypothèse d'une population infinie. La population n'est pas vraiment infinie mais elle est tellement grande que l'approximation est suffisante.

A chaque tirage, la probabilité d'obtenir l'événement "A" qui nous intéresse (par exemple "l'individu est fumeur" sera p, et celle d'obtenir l'événement complémentaire (ici, "l'individu est non-fumeur") sera $(1 - p) = q$. Si on effectue n tirages aléatoires, la probabilité notée $P(X = k)$ d'obtenir au total k individus ou *événements* de type "A", se calcule au moyen d'une formule mathématique qu'on assène trop souvent comme un coup de massue alors qu'elle se déduit d'un raisonnement. Ce raisonnement vous est donc gratuitement fourni en ANNEXE II et vous êtes cordialement invités à l'examiner à l'occasion. Il en ressort que :

$$P(X = k) = C_n^k \times p^k \times q^{n-k}$$

avec :

- $C_n^k = n! / [\, k!\,(n-k)!\,]$
- $x! = x \times (x-1) \times (x-2) \times \ldots \times 3 \times 2 \times 1$

C'est la formule bien connue[2] de la loi binomiale. Le calcul des probabilités pour les différentes valeurs de k (k peut varier de zéro à n) permet d'établir le graphe de répartition des fréquences (ou *distribution*) des événements suivant cette loi binomiale positive **B**, dont les paramètres caractéristiques sont n et p. Ces deux paramètres suffisent à eux seuls à caractériser totalement la loi, puisqu'ils permettent de calculer chacune des probabilités associées aux valeurs k possibles entre zéro et n. On note une

[1] dans le cas ou le menu vous impose l'un OU l'autre, naturellement

[2] De ceux qui adorent perdre leur temps à apprendre les formules par cœur

loi binomiale positive de la façon suivante : **B** (*n* : *p*), où *n* est le nombre de tirages et *p* est la probabilité de l'événement qu'on cherche à étudier. La figure 5.1 représente la loi binomiale **B** (4 : 0,5), associée au nombre de filles attendue dans un échantillon de 4 individus tirés au hasard au sein d'une population dont le sex ratio est parfaitement équilibré (50% de filles).

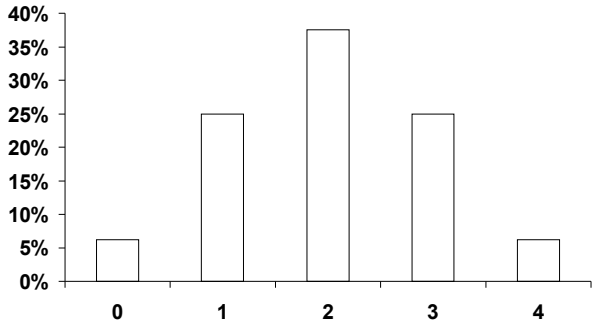


Figure 5.1. Distribution des nombres de filles possibles dans un échantillon aléatoire de 4 personnes prélevé au sein d'une population dont le sex-ratio est parfaitement équilibré (50% de filles) On remarque que le résultat attendu en théorie (deux filles et 2 garçons) est atteint dans moins de la moitié des cas.

Dans ce cas *particulier*, la distribution est symétrique. Cependant si *p* s'écarte de 0,5 la courbe est dissymétrique. La figure 5.2 représente cette situation avec la loi binomiale **B** (10 : 0,1), suivie par le nombre de gauchers attendus dans un groupe de 10 personnes s'il y a exactement 10% de gauchers dans la population.

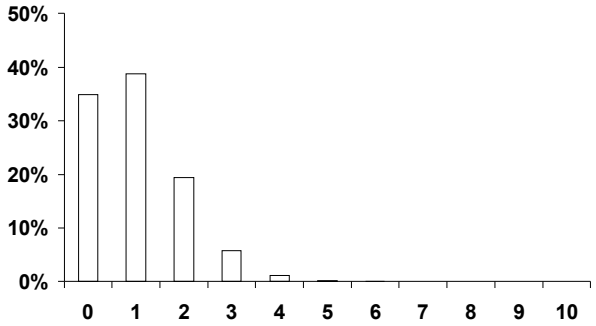


Figure 5.2. Distribution des nombres de gauchers attendus dans un échantillon aléatoire de 10 personnes prélevé au sein d'une population comportant exactement 10% de gauchers. On remarque que le résultat attendu en théorie (un gauchers en 10 tirages) est atteint dans moins de la moitié des cas

Le dernier exemples (figure 5.3) représente les notes attendues lors d'un QCM de 20 questions (avec 4 réponses dont une bonne à chaque fois) lorsque l'étudiant répond entièrement au hasard (une situation absurde, jamais un étudiant ne répondrait au hasard à une question, n'est-ce pas ?). On constate deux choses à la lecture de ce graphique : la note attendue en théorie (5/20 soit le quart des points) est obtenue dans moins de 20% des cas. On constate également (car il y a une morale) que la probabilité d'obtenir la moyenne avec ce genre d'approche est heureusement très faible.

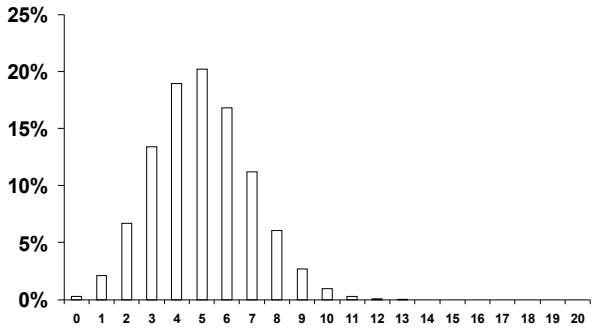


Figure 5.3. Distribution des notes attendues chez des étudiants répondant au hasard à un QCM de vingt questions (1 point par bonne réponse) sachant qu'il y a quatre réponses possibles dont une seule correcte à chaque questions (probabilité de répondre correctement à une question données : 0,25).

La moyenne μ d'une loi binomiale B (n : p), est *en espérance*, c'est à dire sur une infinité de répétitions de n tirages :

$$\mu = np$$

Exemple : en tirant au hasard dix individus dans une population comportant 10% de gauchers, on attend *en moyenne* 1 gaucher, ce qui est bien égal à $np = 10 \times 0{,}1$ et correspond au fait que le nombre de gauchers après 10 tirages suit une loi **B** (10 : 0,1). Mais cela sera seulement vrai *en moyenne*. Dans la réalité, sur plusieurs expériences de n tirages, on observera le phénomène décrit sur la figure 5.2 : les nombres k obtenus vont *fluctuer* autour de la moyenne théorique μ. Cette valeur théorique sera la valeur individuelle *la plus fréquemment observée* (c'est elle qui correspond au *mode*, c'est à dire au pic de la distribution), mais ce ne sera pas la plus fréquemment observée *au total.* Dans l'exemple des gauchers, on s'aperçoit que la valeur théorique de 1 gaucher en 10 tirages sera observée seulement avec la probabilité $C_{10}^1 \times (0{,}1)^1 \times (0{,}9)^9 = 0{,}387$ soit dans *moins de 40% des cas*.

La variance σ^2 de la loi binomiale est :

$$\sigma^2 = npq$$

Ceci dans le cas où la variable X étudiée est le **nombre** x_A d'événements "A" obtenus en n tirages. On peut aussi choisir d'utiliser comme variable la **proportion** observée p_A des événements "A" obtenus en n tirages (=fréquence =pourcentage des événements "A"). Cela revient à effectuer un changement de variable où la variable étudiée n'est plus directement X mais X/n (la variable X est multipliée par une constante C valant 1/n). Or, pour des raisons expliquées plus haut (TRIPATOUILLONS LES DONNÉES) si X est une variable aléatoire et « C » est une constante,

$$\text{Variance de } (CX) = C^2 \times \text{Variance de (X)}$$

Donc, puisque ici $\text{var}(X) = \sigma^2_x = npq$ pour la variable X "nombre de tirages de type A", on peut calculer la variance de la nouvelle variable $p = (1/n)\,X$ ainsi :

$$\sigma^2_{pObs} = \left(\frac{1}{n}\right)^2 npq = \frac{pq}{n}$$

La loi binomiale positive étant la loi qui régit le comportement des **pourcentages,** on comprend qu'elle ait une importance toute particulière, et voilà pourquoi elle occupe une place centrale en sciences.

5.2 la loi de Poisson[1] :

La loi de Poisson est (entre autres) la loi vers laquelle tend la loi binomiale positive lorsque *p* tend vers zéro et *n* tend vers l'infini. Elle régit donc les cas où on doit effectuer un grand nombre de tirages (ou de prélèvements) pour pouvoir observer les événements qu'on étudie (exemple: l'observation d'individus mutants très rares). Dans ce cas, les calculs du C_n^k de la binomiale deviennent extrêmement lourds (essayez un peu de calculer C_{2000}^{50} avec votre si jolie calculatrice...). L'approximation de la loi binomiale par la loi de Poisson est alors la bienvenue car sa formule est moins lourde (une seule factorielle). Cette formule *approchée* est (mêmes notations pour *n*, *p*, et k):

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

Le produit *np* n'est autre que la moyenne (en espérance) de la loi, puisqu'on peut se considérer ici comme dans un cas limite d'une loi binomiale. D'autre part, la variance de la loi de poisson *est elle aussi égale à np*. Cette caractéristique frappante se comprend en remarquant que la variance de la loi binomiale (qui vaut *npq*) tend forcément vers *np* (sa moyenne) lorsque *q* tend vers 1 (et donc lorsque *p* tend vers zéro*)*. Une loi de Poisson est donc entièrement caractérisée en donnant simplement le produit *np*. Pour cette raison, et aussi sans soute parce que ce produit apparaît deux fois dans la formule, on a donné le symbole particulier λ au produit *np* dans la cas de cette loi. Ce paramètre λ est donc la moyenne et la variance de la loi, et la formule de la probabilité que la variable soit égale à une valeur entière k (la loi de poisson comme la binomiale correspond uniquement à des valeurs entières puisqu'elle compte le nombre de fois ou un événement est réalisé) s'écrit habituellement :

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

La notation d'une la loi de Poisson elle même étant simplement : P(λ). Du fait qu'elle représente les cas où *p* est faible (en reprenant l'analogie avec la binomiale), la loi de Poisson est aussi appelée "loi des événements rares". Les cas typiques en biologie sont le nombre d'individus atteint d'une mutation dans une bouteille de 100 drosophiles soumises à une expérience de mutagenèse par rayonnement ou encore le nombre de colonies bactérienne par boîte de Pétri après avoir ensemencé avec une culture très (trop !) diluée.

---

[1] Formalisée par Siméon Denis Poisson. Etait-il lassé de calculer des factorielles pour la binomiale ? La science avance-t-elle grâce aux paresseux (la roue étant l'archétype de l'invention d'un flemmard) ? Voire…Poisson était quand même sorti major de polytechnique, ce qui relativise un peu les choses…

Retenez que pour avoir la possibilité d'utiliser la formule de Poisson pour calculer des probabilités binomiales, il faut avoir *simultanément* un effectif *n* important et une valeur de *p* faible. Pour fixer les esprits, disons $p < 0{,}1$ et $n > 30$. Au fur et à mesure qu'on s'éloigne de ces conditions, l'approximation devient de plus en plus mauvaise et il faut revenir à la formule générale de la loi binomiale. Vous êtes cordialement invités à le vérifier par vous même en calculant *plusieurs* probabilité P (X = k) en utilisant la formule de la loi binomiale et celle de la loi de Poisson, dans un cas ou on *peut* faire l'approximation (*p* petit, *n* grand) et dans un cas ou il ne *faut pas* la faire (*p* grand et/ou *n* petit). La mémorisation du phénomène sera bien plus efficace si vous constatez les choses suite à votre propre calcul.

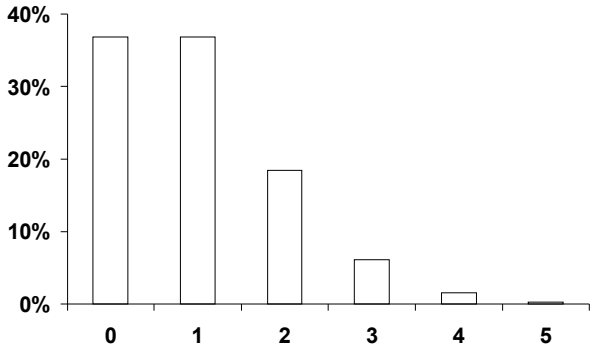


Figure 5.4. Distribution des nombres possibles d'individus albinos obtenus dans un échantillon aléatoire de 10 000 individus si la probabilité dêtre albinos est de 1/10000. La dissymétrie de la loi de Poisson est très nette de part et d'autre de la valeur attendue en théorie (un seul albinos).

5.3 La loi binomiale négative, ou loi de Pascal[(1)].

La situation de départ est celle de la loi binomiale positive : seulement deux possibilités A et B, *mutuellement exclusives,* et de probabilité respectives *p* et *q* constantes au cours de l'expérience. La loi binomiale négative est la loi suivie quand la variable étudiée est *le nombre de tirages successifs nécessaires pour obtenir r événements de type A* (et non pas le nombre d'événements de type A obtenus en *n* tirages, comme dans le cas de la loi binomiale *positive*). NB: On utilise la notation *r* au lieu de *n* pour ne pas introduire de confusion éventuelle avec *n*, le nombre de tirages. La formule qui suit s'explique tout aussi bien que celle de la loi binomiale positive (explication disponible en ANNEXE III). En bref, la probabilité que le nombre de tirages nécessaire pour obtenir le dernier des r événements souhaités soit égal à k est :

$$P(X=k)=C_{k-1}^{r-1}\times p^{r}\times q^{k-r}$$

(1) Car étudiée par Blaise Pascal , Celui là même qui avait peur du silence des espaces infinis (Pascal a vécu bien avant l'invention du périphérique).

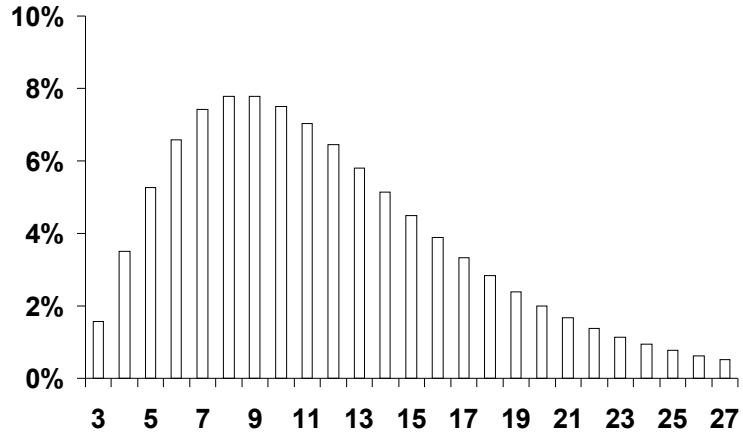


Figure 5.5. Distribution des nombres de tentatives avant d'avoir attrapé trois proies chez un prédateur réussissant ses attaques en moyenne une fois sur quatre (P attraper proie = 0,25 à chaque attaque). Le nombre théorique attendu est de 12 tentatives (3/0,25 soit "0,25 proie attrapée par tentative") mais on constate que le prédateur sera parfois très chanceux... ou très malchanceux. La longue "queue de distribution" vers la droite et les valeurs extrêmes est le trait typique d'une binomiale négative.

Ce type de loi est illustrée ici avec le nombre d'attaques que doit réaliser un prédateur avant d'avoir capturé trois proies, s'il réussit ses attaques une fois sur quatre (ce qui serait un excellent rendement, le tigre est largement en dessous de ces performances). On constate que si, les jours de chances, trois attaques suffiront, il y aura aussi des journées épuisantes où il faudra sonner la charge une bonne trentaine de fois, voire plus pour se caler l'estomac (la distribution est tronquée ici à 27 mais contine théoriquement à l'infini).

5.4 Sa Majesté la Loi Normale.

5.4.1 présentation sans ménagements

La loi normale est la loi **continue** dont la **densité de probabilité** est :

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

Voilà qui est envoyé. En pratique, heureusement, vous n'aurez jamais à toucher à cette formule. Sachez simplement que si on l'applique, on obtient la fameuse courbe en cloche, point à la ligne.

Il est beaucoup plus important d'insister sur la notion de loi ***continue*** d'où découle celle de *densité de probabilité*. Les lois présentées plus haut (Binomiales, Poisson) s'appliquaient à des valeurs entières (comptages), ou bien à des pourcentages *issus de comptages*. Il était donc toujours possible (et souvent intéressant) de calculer la probabilité exacte d'un résultat particulier. La valeur obtenue pouvait être éventuellement faible, mais on pouvait la connaître. On a vu que la probabilité d'obtenir exactement 1 gaucher en 10 tirages dans une population comportant 10% de gauchers peut être calculée, et elle vaut exactement 0,38742049 soit environ 39% des cas.

En revanche, pour une variable X *continue* (temps, taille, poids, température...), la probabilité que X soit égal à une valeur k *précise* est *impossible* à calculer, parce qu'elle est *infiniment faible*. Il en est ainsi parce qu'il y a une *infinité* de valeurs possibles dans n'importe quel intervalle choisi : il y a une infinité de températures possibles entre 10 et 11°C, il y a une infinité de tailles possibles entre 50 et 51 millimètres. Calculer la probabilité qu'il fasse *exactement* 20°C en ce moment dans la pièce où je me trouve n'a aucun sens parce que cette probabilité est infinitésimale[3].

Cependant (et c'est l'intérêt de la formule compliquée vue plus haut), on peut calculer la probabilité que X soit compris *entre* deux valeurs *a* et *b*. Ce calcul est possible à condition de calculer une intégrale, et plus précisément en intégrant la *fonction de densité de probabilité* de la variable aléatoire, notée *f(x),* entre ces deux bornes *a* et *b*. C'est le sens de la notion de *densité de probabilité* : *f(x)* est la fonction telle que :

$$P(X \in [a.b]) = \int_a^b f(x)dx$$

Pas de panique, vous n'aurez jamais à faire ce calcul à la main, les tables et les logiciels sont là pour ça. Retenez plutôt la notion qu'il représente : il n'y a probablement aucune graine d'érable dans l'univers mesurant *exactement* 39 millimètres, au milliardième de micron près (« *aucune* » = probabilité infiniment faible) **mais** la probabilité de trouver une graine d'érable de taille comprise *entre* 38 et 40 millimètres est en revanche loin d'être nulle (les modestes données de la figure 3.6 permettent même de donner un ordre de grandeur pour cette probabilité, qui pourrait être grosso modo 20%).

5.4.2 Le miracle du théorème central limite

La loi normale est sans contestation possible la reine des statistiques. Elle le doit à un théorème extrêmement important, le **théorème de la limite centrale** (*central limit theorem*), nommé selon la tradition Française "Théorème Central–Limite" (par la grâce d'une traduction assez rudimentaire). C'est le seul théorème mathématique de tout ce livre, promis juré. Que dit-il ?

[3] et comme je réécris ce chapitre un premier août dans le Vaucluse, je peux même vous dire que la probabilité qu'il fasse 20°C dans cette pièce n'est même pas infinitésimale, elle est nulle.

"Théorème Central Limite"

Soient *n* variables aléatoires X_1, X_2... Xn

- *indépendantes* deux à deux
- distribuées selon la *même densité de probabilité*,
- ayant *même moyenne* μ et *même variance* σ^2.

On pose :

$$Y = X_1 + X_2 + \ldots + X_n$$

$$Z = \frac{(Y - n\mu)}{\sqrt{n\sigma^2}}$$

Alors, Z $\rightarrow$ N (0 : 1) quand *n* tend vers l'infini.

Il est fort possible (doux euphémisme) que la beauté de la chose ne vous apparaisse pas du premier coup d'œil. Analysons donc la situation posément. La variable Y est simplement la somme de *n* variables X qui ont toutes la même moyenne μ et la même variance σ^2. Selon les opérations sur les variables aléatoires vues précédemment (Tripatouillons les données), et *si les variables sont indépendantes les unes des autres* (la condition clairement posée par le théorème), alors :

- la moyenne μ_Y de cette somme sera *la somme des n moyennes*,

$$\mu_Y = \mu + \mu + \mu \ldots \text{(etc. } n \text{ fois)} = n\mu$$

- La variance σ^2_Y de cette somme sera *la somme des n variances*,

$$\sigma^2_Y = \sigma^2 + \sigma^2 + \sigma^2 \ldots \text{(etc. } n \text{ fois)} = n\sigma^2$$

Donc, à ce stade, on peut dire que Y suit une loi de type inconnu mais ayant les caractéristiques ($n\mu$: $n\sigma^2$). Pour l'instant, rien de bien fabuleux, car on est toujours sans la moindre information sur la loi suivie par les *n* variables indépendantes X_1 à X_n. La très grande nouvelle annoncée par le théorème central-limite est que *cela n'a pas la moindre importance* : quelle que soit la loi suivie par les *n* variables X_i, leur somme suit **une loi normale** quand *n* tend vers l'infini.

En effet, la seconde variable introduite, Z, consiste simplement à retrancher à chaque valeur de Y sa moyenne $n\mu$ (c'est un *centrage*) et à diviser par son écart-type $\sqrt{n\sigma^2}$ (c'est une réduction). Z est donc une variable centrée-réduite (sa moyenne vaut zéro, et sa variance 1), mais l'opération de centrage-réduction ne change pas la nature profonde de la loi. Puisqu'on nous dit que Z est une loi normale centré réduite, c'est donc que Y était une loi normale non centrée et non réduite (de moyenne $n\mu$ et de variance $n\sigma^2$). A ce stade, la question qui tourne comme un écureuil dans votre cerveau perplexe est bien entendu "**ET ALORS** ????". En effet, vous êtes en droit de vous demander sur quelle planète tordue on peut espérer tomber par hasard sur une suite de *n* variables aléatoires *ayant miraculeusement la même moyenne et la même variance.* Vous pouvez aussi vous demander quel intérêt appliqué peut bien présenter un théorème mathématico-

philosophique qui nécessite de disposer d'*une infinité de variables aléatoires* avant de pouvoir en tirer quelque chose (une loi normale)?

Revenons donc sur terre, et examinons un humble objet beaucoup plus concret, le pain noir quotidien du chercheur, une modeste *moyenne* calculée sur un échantillon de taille non infinie de *n* individus. Comment calcule-t–on une moyenne, déjà ? Ah oui, on prend les *n* valeurs observées, et on les *additionne* (avant de diviser le tout par une simple constante, qui est *n*). Et maintenant, posons quelques questions simples. Chacune des *n* valeurs observées est-elle connue à l'avance ? Non. Il s'agit donc de *n variables aléatoires*, que l'on peut appeler par exemple X_1, X_2... X_n. Connaître X_1 vous permet-il de connaître X_2 ? Absolument pas. Donc ces *n* variables aléatoires sont *indépendantes* les unes des autres. Sont-elles tirées dans des distributions différentes ? A l'évidence non, puisqu'elles sont tirées *dans la même population,* elles sont donc issues de la même distribution statistique et auront donc toutes *la même densité de probabilité*. On peut en déduire également que X_1 ou X_2 ou X_n auront la *même moyenne*, que je peux appeler μ. De même, X_1, X_2 ou X_n auront la *même variance*, que je peux appeler σ^2. Conclusion (je pense que vous commencez à voir vaguement où je veux en venir ?), lorsque vous calculez une *moyenne*, vous additionnez *n variables aléatoires indépendantes distribuées selon la même densité de probabilité, de même moyenne et de même variance*. **Vous êtes donc dans les conditions du théorème central limite**. En fait, vous avez manipulé le théorème central limite des centaines de fois depuis le CM2 (chaque fois que vous avez calculé une moyenne), et sans même l'avoir appris. Vous êtes très forts.

Bref, si la moyenne de votre variable aléatoire est μ et sa variance σ^2, en appliquant la règle selon laquelle $\mathrm{var}(C\mathrm{X}) = C^2\mathrm{var}(\mathrm{X})$ avec $C = 1/n$, vous pouvez en déduire que la moyenne *m* que vous calculez sur votre échantillon de taille *n* suit une loi normale ayant les caractéristiques suivantes :

$$m \rightarrow N\left(\mu : \frac{\sigma^2}{n}\right)$$

Ceux d'entre vous qui ont suivi depuis le début devraient cependant émettre une objection sérieuse : le théorème central limite *n'est valable que si n tend vers l'infini*. Voilà quand même une sacrée contrainte du point de vue pratique. Mais ce n'est point un hasard si le titre de cette partie est "le miracle du théorème central limite". En effet, en statistiques, *l'infini commence à 30* (les statisticiens sont finalement des gens très raisonnables). En clair, même si le théorème central limite n'est vrai dans sa parfaite pureté que lorsque *n* atteint l'infini, il est *approximativement* vrai largement avant, et la valeur 30 suffit pour pouvoir l'utiliser dans la vie scientifique de tous les jours.

Une moyenne suit une loi approximativement normale dès lors qu'elle est établie à partir d'un échantillons d'une trentaine d'individus ou plus.

Cerise sur le gâteau, il a été démontré par la suite que ce théorème reste valable dans une large palette de conditions car :

- la condition théorique « *n* variables de même moyenne et de même variance » n'est pas vitale tant que la variance de chaque variable *est faible par rapport à la variance du total;*
- la condition d'indépendance 2 à 2 des variables aléatoires (rarement absolue en biologie) peut supporter quelques entorses.

Il en résulte que beaucoup de variables biologiques suivent des lois approximativement normales *avant même de faire la moindre moyenne*, car elles résultent *intrinsèquement* de l'addition (approximative) de *nombreux* et *petits* effets aléatoires, génétiques et environnementaux, dont chacun apporte une contribution à la variance qui est *faible* par rapport à la variance totale. On se retrouve ainsi presque automatiquement dans les conditions du théorème centrale limite. Toutefois, méfiance méfiance, souvenez vous des exemples biologiques que nous avons vus ensembles (dont l'incontournable taille des graines d'érable) dont la distribution n'est pas du tout normale !

Quid du deuxième ingrédient formant le pain quotidien du biologiste, c'est-à-dire les fréquences (ou pourcentages) ? Bonne nouvelle : **les fréquences suivent aussi très rapidement des lois normales quand *n* grandit** (rappelons le « lien de parenté » entre la loi normale et la loi binomiale, dont elle est une limite à l'infini). En effet, pour calculer une fréquence *p*, on additionne une suite de zéro (l'événement "A" n'est pas réalisé) et de 1 (l'événement "A" est réalisé) qui sont autant de variables binomiales indépendantes de même moyenne et de même variance (le tout étant divisée par une constante, *n* le nombre de tirages). **On retrouve les conditions du théorème central limite**, et il en ressort que *si np > 5 et nq>5*, on a avec une bonne approximation :

$$p_{Obs} \rightarrow N\left(p : \frac{pq}{n}\right)$$

Vous avez immédiatement remarqué que la condition sur la taille de l'échantillon n'est plus ici directement $n > 30$, mais fait intervenir les produits *np* et *nq*. Cela est dû au fait que si *p* est proche de 0,5 la loi binomiale est presque symétrique et déjà très proche de la forme d'une loi normale. Il suffit donc de peu de répétitions pour aboutir à une forme normale. A la limite, si $p = 0{,}5$ vous pouvez constater qu'il suffit théoriquement de 10 individus pour pouvoir appliquer le théorème central limite, car la loi **B**(*n* : 0,5) est parfaitement symétrique. En revanche, si *p* est plus éloigné de 0,5 la loi binomiale devient nettement dissymétrique, et il faudra éventuellement plus de 30 tirages pour parvenir à normaliser la somme des Xi. Les produits *np* et *nq* sont donc une façon commode de prendre en compte ce phénomène. Il ne faut quand même pas oublier que le tirage binomial de base n'accepte que deux valeurs : 0 ou 1. C'est une situation vraiment extrême et il est appréciable que l'on puisse aboutir rapidement à une loi normale quand même.

6. La confiance règne (par intervalles)

Ce chapitre est probablement le plus important de tous[1]. Lisez-le, relisez-le, faites le lire à vos amis, dissimulez–le aux yeux de vos ennemis, enseignez–le à vos enfants (s'ils s'intéressent aux statistiques). Bref, il contient des choses importantes qui peuvent vous éviter de dire et de faire bien des bêtises. Grâce à lui, vous allez pouvoir calculer à quel point, malgré vos efforts d'échantillonnage et vos exploits expérimentaux, vous connaissez mal la réalité. Cette expérience est douloureuse, mais elle est nécessaire. Les biologistes connaissant la valeur exacte de la moyenne et de la variance du caractère qu'ils étudient dans une population d'organismes vivants sont en effet comme les orangers sur le sol irlandais : on n'en verra jamais. En pratique, vous devrez donc toujours vous contenter d'*estimations* de ces valeurs, basées sur une partie (généralement minuscule) de la population étudiée. Remarque importante : "l'échantillon" désigne ici *l'ensemble des individus échantillonnés*, et non pas un échantillon parmi d'autres s'il y en a plusieurs. Rappelez-vous qu'il faut chercher à établir son estimation des paramètres de la population sur le maximum de données possible.

La question cruciale qui se pose est maintenant la suivante : *jusqu'à quel point peut-on se fier aux valeurs estimées à partir d'un simple échantillon* ? La réponse s'obtient par le calcul des intervalles de confiance. Précisons tout de suite ce qu'un intervalle de confiance *n'est pas*. Il *n'est pas* l'intervalle dans lequel la véritable valeur du paramètre se trouve *avec certitude*. En effet, la variable aléatoire peut théoriquement prendre toutes les valeurs possibles dans les limites des lois de la physique, ce qui fait quand même beaucoup. L'intervalle de confiance représente en fait la zone dans laquelle se trouve « très probablement », et *avec une probabilité qu'on choisit*, la véritable valeur (à jamais inconnue) du paramètre que l'on étudie dans la population. On utilise en pratique les probabilité 0,95 ou 0,99 (respectivement : seuil de confiance à 95% et 99%).

Sur quel paramètre peut-on calculer un intervalle de confiance ? Sur littéralement *n'importe quoi de chiffré*. On calcule le plus couramment les intervalles de confiance des moyennes et des pourcentages observés, mais on peut calculer l'intervalle de confiance d'une variance, l'intervalle de confiance d'une différence entre deux moyennes ou pourcentages (magnitude de l'effet ou *effect size*), l'intervalle de confiance d'un coefficient de corrélation, de la pente d'une droite de régression, de la valeur d'un indice quelconque, bref, les possibilités sont infinies. L'idée de base est toujours la même : "quelle est la précision de mon estimation ?". Mais voyons d'abord ce dont vous aurez besoin à coup sûr:

[1] mais vous ne le comprendrez que si vous avez lu et assimilé les précédents. On n'a rien sans rien.

6.1 Intervalle de confiance d'une moyenne.

6.1.1 Grand échantillon ($n > 30$), loi quelconque

Selon le Théorème Central Limite[2], si votre variable aléatoire X suit une loi de distribution quelconque, avec pour moyenne μ et pour variance σ^2, alors, pour un grand échantillon ($n > 30$), la moyenne m calculée sur cet échantillon suivra une loi approximativement **normale,** de moyenne μ et de variance σ^2/n :

$$m \rightarrow \mathbf{N}\ (\mu : \sigma^2/n)$$

Du fait que l'échantillon est grand, on peut sans dommages remplacer σ^2 (inconnu) par son estimation s^2 calculée sur l'échantillon, l'équation ci–dessus restera valable, donc:

$$m \rightarrow \mathbf{N}\ (\mu : s^2/n)$$

Or, une loi normale est ainsi faite que 95% des valeurs sont situées dans un intervalle de $\pm$ 1,96 écarts-types autour de sa moyenne. L'écart-type de μ est ici estimé par *l'erreur standard* (racine carrée de l'estimation de la variance de la moyenne $s^2_m = s^2/n$), donc :

$$s_m = \text{erreur standard} = \sqrt{(s^2/n)}$$

On en déduit l'intervalle dans lequel la véritable valeur μ (à jamais inconnue) a 95% de chances de se trouver :

$$\mu = m \pm 1{,}96\sqrt{\frac{s^2}{n}}$$

Ainsi, μ se trouve très probablement *quelque part* dans un rayon de 1,96 erreur-standard autour de notre valeur calculée m. Cet intervalle est l'intervalle de confiance à 95% de m, et vous donne une idée sur *la précision de votre estimation*.

Exemple 6.1 : Taille de 40 étudiants (garçons) de la maîtrise BPE.

Moyenne observée : m = 178,025 cm. Variance estimée : s^2 = 50,384. Quel est l'intervalle de confiance à 95% de la moyenne ?

Erreur standard : e.s. = $\sqrt{(s^2/n)}$ = $\sqrt{(50{,}384/40)}$ = 1,122 cm
Borne inférieure au seuil 95% : 178,025 – 1,96 × 1,122 = 175,82 = 175,8 cm
Borne supérieur au seuil 95% : 178,025 + 1,96 × 1,122 = 180,225 = 180,2 cm

[2] voir chapitre 5 "lois à connaître en biologie"

IC $_{95\%}$ = [175,8 —180,2 cm]
Amplitude : 4 centimètres

Exemple 6.2 : Taille de 228 étudiantes de la maîtrise BPE

Moyenne observée : m = 166,5 cm. Variance estimée : s^2 = 33,1. Quel est l'intervalle de confiance à 95% de la moyenne ?

Erreur standard : e.s. = $\sqrt{(s^2/n)}$ = $\sqrt{}$ (33,1/228) = 0,4 cm
Borne inférieure au seuil 95% : 166,5 – 1,96 × 0,4 = 165,7 cm
Borne supérieur au seuil 95% : 166,5 + 1,96 × 0,4 = 167,2 cm

IC $_{95\%}$ = [165,7 —167,2 cm]
Amplitude : 1,5 cm.

La meilleure précision (par rapport à l'exemple 4.1) est due à la taille plus élevée de l'échantillon, qui réduit la taille de l'erreur standard. Vous noterez cependant que le gain de précision n'est hélas pas proportionnel à l'augmentation de la taille de l'échantillon: En effet, la précision s'améliore proportionnellement à **la racine carrée** de n, et non pas proportionnellement à n...

Exemple 6.3 : Longueur de 204 graines ailées d'Erable

Moyenne observée : m = 39,4 mm. Variance estimée : s^2 = 25,3. Quel est l'intervalle de confiance à 95% de la moyenne ?

Erreur standard : e.s. = $\sqrt{(s^2/n)}$ = $\sqrt{}$ (39,4/204) = 0,4 mm
Borne inférieure au seuil 95% : 39,4 – 1,96 × 0,4 = 38,7 mm
Borne supérieur au seuil 95% : 39,4 + 1,96 × 0,4 = 40,1 mm

IC$_{95\%}$ = [38,7 — 40,1 mm]
Amplitude : 1,4mm

6.1.2 Petit échantillon (n < 30), loi normale

Si la variable suit une loi normale, nul besoin d'invoquer le Théorème Central Limite, toute moyenne de variables normales est une variable normale, donc la moyenne observée m (variable aléatoire de moyenne μ et de variance σ^2/n par la seule application des règles des opérations sur les variables aléatoires) est automatiquement **normale**,:

$$m \rightarrow \mathbf{N}\ (\mu : \sigma^2/n)$$

Le hic survient au moment de remplacer σ^2 (inconnu) par son estimation s^2 basée sur l'échantillon. En effet, l'approximation est trop grossière si l'échantillon est petit : la sous-estimation de σ^2 (probable quand on utilise un échantillon) risque d'être ici trop importante. Si on applique la formule habituelle : « $\mu = m \pm$ **1,96** $\sqrt{(s^2/n)}$ » On risque de *sous-estimer* la taille réelle de l'intervalle de confiance (c'est à dire que l'estimation de μ va apparaître *plus précise qu'elle ne l'est en réalité*). Heureusement, la loi suivie par la variable centrée-réduite :

$$t = \frac{m - \mu}{\sqrt{\frac{s^2}{n}}}$$

…a été étudiée par un Anglais nommé William GOSSET qui a publié ses travaux en 1908 sous le pseudonyme de STUDENT[3]. Comme Sir R. A. FISHER (le père des statistiques modernes) a mis son nez dans cette loi par la suite, elle porte le nom de **S**TUDENT-**F**ISHER et est désignée (admirez la logique) par la lettre… **t**. Le test statistique qui s'y rattache (et qu'on verra au chapitre 9) s'appelle le *test t de Student*. Deux détails de pure forme: la variable t est désignée par une *minuscule* mais il faut une *majuscule* à **S**tudent, car il s'agit d'un nom propre, tout pseudonyme qu'il est (donc, écrivez "t de Student" et non pas "T de student").

Les valeurs critiques de la distribution du t de Student sont consignées dans une table, mais la lecture de cette table diffère de celle de la loi normale car à chaque effectif n correspond une distributions du t de Student spécifique. Plus exactement, la table se lit en fonction du nombre $(n - 1)$ qui désigne le nombre de variables aléatoires *indépendantes* dans l'échantillon. Il n'y a en réalité que $n - 1$ variables aléatoires parmi les n individus car elles sont toutes liées par leur total. Il suffit en effet de connaître $n - 1$ valeurs pour déduire la dernière à partir du total. Ce nombre de variables aléatoires indépendantes est le nombre de *degrés de liberté (d.d.l.)* de la variable aléatoire t. La notion de degré de liberté[4] en statistiques est un épouvantail à étudiants notoire, et elle est en effet pleine de pièges diaboliques. La lumière viendra de la pratique. On verra d'autres exemples où le nombre de d.d.l. interviendra.

Au final, la formule permettant le calcul de l'intervalle de confiance est la même que précédemment, sauf qu'il faut remplacer la valeur $|\varepsilon| = 1{,}96$ de la loi normale par la valeur figurant dans la table du t de Student en fonction du risque α considéré et du nombre de degrés le liberté. Pour un intervalle de confiance à 95% on choisit $\alpha = 0{,}05$:

$$\mu = m \pm t_{(n-1)\,ddl}\sqrt{\frac{s^2}{n}}$$

Voici donc réglé le cas où le caractère étudié *suit une loi normale*.

Exemple 6.4 : Taille de 10 étudiants (garçons) fictifs (mais je reprend volontairement les mêmes valeurs de moyenne et de variance que dans l'exemple 6.1).

Moyenne : $m = 178{,}025$ cm. Variance estimée : $s^2 = 50{,}384$ Quel est l'intervalle de confiance à 95% de la moyenne ?

L'échantillon est trop petit ($n = 10$) pour pouvoir utiliser le théorème central limite. Cependant, la taille dans l'espèce humaine est un caractère approximativement distribué selon une loi normale. La moyenne m va donc être distribuée approximativement selon la distribution du t de Student avec $(n - 1) = 9$ degrés de liberté.

Erreur standard : e.s. $= \sqrt{(s^2/n)} = \sqrt{(50{,}384/10)} = 2{,}24$ cm

Valeur seuil de la table du t de Student pour $\alpha = 0{,}05$ et 9 degrés de liberté : $t_{(\alpha=0{,}05;9ddl)} = \mathbf{2{,}262}$

[3] William Gosset n'a pas publié sous un pseudonyme parce qu'il avait honte de faire des statistiques, mais parce que son employeur (les célèbres bières Guiness), lui avait interdit de publier sous son vrai nom, pour des raisons qui m'échappent.

[4] *degrees of freedom*, que vous trouverez dans les articles scientifiques abrégé en "d.f."

Borne inférieure au seuil 95% : 178,025 – **2,262** × 2,24 = 172,958 = 173,0 cm
Borne supérieur au seuil 95% : 178,025 + **2,262** × 2,24 = 183,092 = 183,1 cm

$IC_{95\%}$: [173,0 ——————— 183,1 cm]
Amplitude : 10 centimètres

A comparer avec l'amplitude de 4 centimètres de l'exemple 4.1 : [175,8 — 180,2 cm]. Bien que la variance estimée pour le caractère soit artificiellement identique avec l'exemple **4.1**, l'intervalle de confiance de la moyenne est ici plus large (donc l'estimation est *beaucoup moins précise*). Deux raisons à cela : (i) l'erreur standard est plus grande (car *n* est plus petit), et (ii) la loi du t de Student a une plus grande variance (= est plus « étalée ») que la loi normale, d'où la valeur critique **2,262** au lieu du **1,96** utilisable pour les grands échantillons.

6.1.3 Petit échantillon ($n < 30$), loi quelconque.

Le calcul d'un intervalle de confiance en utilisant la loi du t de Student reste approximativement valable même si la loi suivie par la variable aléatoire n'est pas exactement une loi normale. L'important est (entre autres) que la distribution du caractère ne soit pas *trop* dissymétrique. En pratique, ces conditions approchées sont souvent vérifiées (regardez donc vos données), et vous pourrez alors utiliser le t de Student même sans avoir des courbes en cloche impeccables. Faites-le cependant en ayant conscience de l'approximation commise, et du fait que vous êtes en train de pousser une méthode dans ses limites.

En revanche, vous pouvez être face à une distribution qui *s'écarte fortement* de la loi normale : *Loi de Poisson*, *binomiale négative* (voir chapitre 5 lois stat), *distribution "en J"* (grande partie des valeurs massées à droite vers une valeur maximum) *distribution "en L"* (toutes les valeurs massées à gauche vers une valeur minimum) voire, l'horreur absolue, distribution *bimodale* (Surf Island !)[5] et même, encore pire, *distribution "en U"*. Il est alors hors de question d'utiliser la distribution du t de Student comme référence. La solution consiste à utiliser la technique de re-échantillonnage dite du **bootstrap**. Cette technique vous est présentée plus loin, dans la section 6.5 Intervalle de confiance de *tout ce que vous voulez*.

6.2 Intervalle de confiance d'un pourcentage.

6.2.1 Grand échantillon (np et $nq > 5$)

Comme dit plus haut, une fréquence p est l'addition de n variables correspondant chacune a *un* tirage de type « oui – non » (1 ou 0) dans une loi binomiale, le résultat de l'addition étant divisé par n (et multiplié par 100 dans le cas d'un pourcentage). Dans le cas d'un grand échantillon, on peut alors appliquer le Théorème Central Limite. On sait que la loi binomiale **B** (n : p) a pour variance pq/n si on considère non pas l'effectif X mais la fréquence X/n. La loi normale qu'on lui substitue aura donc les mêmes paramètres : moyenne p et variance pq/n (avec $q = 1 - p$).

Notre fréquence observée suivra donc :

$$p_{obs} \rightarrow \mathbf{N}\ (p\ :\ pq/n)$$

[5] voir le chapitre 2 "Présentez vos données"

L'échantillon étant "grand", cette relation reste approximativement valable en remplaçant la variance exacte pq/n par son estimation $p_{obs}q_{obs}/(n-1)$, donc

$$p_{obs} \rightarrow \mathbf{N}\,(p\ :\ p_{obs}q_{obs}/(n-1))$$

Rappel : la division par $n-1$ élimine le biais de sous-estimation de la variance à partir d'un échantillon

En présence d'une loi normale de variance connue, nous sommes tirés d'affaire et la suite des opérations est *exactement* la même que dans le cas du calcul de l'intervalle de confiance d'une moyenne. Les mêmes causes produisant les mêmes effets, on en arrive à la même formule pour l'intervalle de confiance avec $\alpha = 0{,}05$:

$$p = p_{obs} \pm 1{,}96 \sqrt{\frac{p_{obs}q_{obs}}{n-1}}$$

Note : le fait d'utiliser la fréquence elle même (varie de 0 à 1) ou le pourcentage (de 0 à 100) n'a aucune importance **à condition évidemment de ne pas faire de mélanges audacieux dans la formule**. Donc *tout en pourcentages* ou *tout en fréquences* mais restez homogènes dans vos calculs !

Exemple 6.5 : Sur 146 étudiants de maîtrise BPE ayant fourni l'information, 20 étaient gauchers ou ambidextres (soit p_{obs} = 13,698% et q_{obs}= 86,302% droitiers). Quel est l'intervalle de confiance du pourcentage de la catégorie [gauchers & ambidextres] ?

Petite vérification préliminaire : $np_{obs} = 146 \times 0{,}13698 = 19{,}99 >> 5$; $nq_{obs} = 146 \times 0{,}86302 = 126 >> 5$. On a bien $np_{obs} > 5$ et $nq_{obs} > 5$, on peut donc utiliser le Théorème Central Limite.

Erreur standard : e.s. = $\sqrt{[p_{obs}\, q_{obs}/(n-1)]} = \sqrt{[(0{,}13698 \times 0{,}86302)/(146-1)]} = 0{,}02845 = 2{,}845\ \%$
Borne inférieure au seuil 95% : $0{,}13698 - \mathbf{1{,}96} \times 0{,}02845 = 0{,}081218$ = 8,1%
Borne supérieur au seuil 95% : $0{,}13698 + \mathbf{1{,}96} \times 0{,}02845 = 0{,}192742$ = 19,3%

$IC_{95\%}$: [8,1% ——————— 19,3%]

Cet intervalle est *très large*, une parfaite illustration de la difficulté à estimer les fréquences avec précision, même avec des échantillons de taille respectable.

6.2.2 Petit échantillon (np et $nq < 5$)

Dans ce cas, l'approximation par la loi normale n'est plus possible. Pour tout arranger, l'estimation de p par p_{obs} n'est pas suffisamment précise non plus pour ne pas fausser l'estimation de la variance pq/n en lui substituant $p_{obs}q_{obs}/(n-1)$, cette estimation étant indispensable au calcul de l'intervalle de confiance. Si de grands anciens n'étaient pas passés avant nous, il faudrait « tout simplement » revenir à la base (c'est à dire au niveau de la binomiale) et calculer *une par une* les probabilités $P(X/n = p_i)$ grâce à la formule de la loi binomiale. Il faudrait ensuite « éliminer » les 2,5% les plus extrêmes de la distribution de chaque côté et déterminer au bout du compte l'intervalle de confiance à 95%. Heureusement, d'autres ont déjà fait le sale boulot, et ils nous ont légué une table qui vous donnera directement l'intervalle de confiance d'un pourcentage dans le cas des petits effectifs (voir TABLES). Comme quoi venir trop tard dans un monde trop vieux n'a pas que des inconvénients.

Exemple 6.6 : sur **35** étudiants on observe 2 gauchers (soit p_{obs} = 0,05714 soit 5,71%). Quel est l'intervalle de confiance de ce pourcentage basé sur un soi-disant "grand" échantillon (puisque $n > 30$) ?

Petite vérification (mais on connaît le résultat à l'avance...)
$np_{obs} = 35 \times 0{,}05714 = 2 < 5$; on ne peut **pas** utiliser l'approximation par la loi normale.

La table de l'intervalle de confiance des pourcentages donne directement les bornes (ici par interpolation approximative entre les valeurs concernant p_{obs} = 5% et celles pour p_{obs} = 10%) : Borne inférieure au seuil 95% : 0,5% environ. Borne supérieur au seuil 95% : 19% environ

$IC_{95\%}$: environ [0,5 ——————————— 19%]

Selon une expression anglo-saxonne très imagée, cet intervalle de confiance est *suffisamment grand pour qu'un vaisseau de guerre puisse y faire demi-tour*. **Abandonnez** une bonne fois pour toutes l'idée selon laquelle on peut estimer des pourcentages de façon fiable sans avoir *beaucoup* de données.

6.3 Intervalle de confiance d'une différence entre deux moyennes.

6.3.1 Grands échantillons (n_A et $n_B > 30$)

Si deux variables aléatoires X_A et X_B suivent des lois de distribution quelconque de moyennes μ_A et μ_B et de variances σ^2_A et σ^2_B (dont on possède les estimations s^2_A et s^2_B, basées sur deux *grands* échantillons A et B avec n_A>30 et n_B>30), alors, par la grâce du Théorème Central Limite, les moyennes m_A et m_B calculées sur ces deux échantillons suivront des lois approximativement *Normales* ayant les caractéristiques suivantes :

$$m_A \rightarrow \mathbf{N}\ (\mu_A :\ s^2_A/n_A)$$
$$m_B \rightarrow \mathbf{N}\ (\mu_B :\ s^2_B/n_B)$$

En supposant que X_A et X_B sont indépendantes, les règles d'opération sur les variables aléatoires nous permettent de déduire que la différence $D = m_A - m_B$ suivra elle aussi une loi normale, ayant pour moyenne $\Delta = \mu_A - \mu_B$ la *véritable différence* des moyennes et comme variance la *somme* des variances (souvenez vous que les variances ne se soustraient **jamais**), donc :

$$D = (m_A - m_B) \rightarrow \mathbf{N}\ (\mu_A - \mu_B : s^2_A/n_A + s^2_B/n_B)$$

Or, on sait que 95% des valeurs d'une loi normale sont situées dans un intervalle de ± 1,96 erreurs standards de la moyenne. L'erreur standard étant l'écart-type de la moyenne, donc la racine carrée de la variance, on a ici

$$s_D = \sqrt{(s^2_A/n_A + s^2_B/n_B)}$$

On en déduit l'intervalle de confiance de **D**:

$$\mathbf{\Delta = D \pm 1{,}96 \sqrt{(s^2_A/n_A + s^2_B/n_B)}}$$

Exemple 6.7 : IC95% de la différence de taille entre les étudiants et les étudiantes de MBPE

Les données réelles sont les suivantes.
Etudiants : $n_A = 232$; $m_A = 178{,}6$ cm ; $s^2_A = 36{,}9$
Etudiantes: $n_B = 228$; $m_B = 166{,}5$ cm ; $s^2_B = 33{,}1$
différence observée : $D = m_A - m_B = 178{,}6 - 166{,}5 =$ **12,1cm**. Quel est son $IC_{95\%}$?

Les effectifs sont de grande taille, m_A et m_B suivent approximativement des lois normales de variances respectives :

$$s^2_{mA} = s^2_A / n_A = 36{,}9/232 = 0{,}159$$
$$s^2_{mB} = s^2_B / n_B = 33{,}1/228 = 0{,}145$$

La variance de D est donc $s^2_D = s^2_{mA} + s^2_{mB} = 0{,}159 + 0{,}145 = 0{,}304$
Son écart-type $s_D = \sqrt{s^2_D} = \sqrt{0{,}304} = 0{,}552$

Borne inférieure de l'IC_{95} : $12{,}1 - 1{,}96 \times 0{,}552 = 11{,}061 = 11{,}1$ cm
Borne supérieure de l'IC_{95} : $12{,}1 + 1{,}96 \times 0{,}552 = 13{,}224 = 13{,}2$ cm

$IC_{95\%}$: environ [11,1—- 13,2cm]
Amplitude : 2,1 cm
.

6.3.2 Petits échantillons (n_A et $n_B < 30$)

Si la distributions suivie par la variable aléatoire qui vous intéresse *s'écarte nettement* de la loi normale (Poisson, Binomiale négative, distribution en J, en L, bimodale ou en U), dirigez-vous tout de suite vers la section 6.5 intervalle de confiance de *tout ce que vous voulez*. En revanche, si vous travaillez avec une variable aléatoire raisonnablement proche d'une distribution normale, ce qui suit vous concerne. Le raisonnement est le même que dans le cas des grands échantillons, sauf qu'on va utiliser les valeurs de la table du t de Student au lieu du 1,96 de la loi normale.

Si deux variables aléatoires X_A et X_B suivent des lois de distribution proches de la loi normale, de moyennes μ_A et μ_B et de variances σ^2_A et σ^2_B, alors, les moyennes m_A et m_B calculées sur des échantillons de tailles n_A et n_B suivront des lois du t de Student ayant approximativement les caractéristiques suivantes :

$$m_A \rightarrow \mathbf{t}\,(\mu_A : s^2_A/n_A) \text{ avec } n_A - 1 \text{ degrés de liberté}$$
$$m_B \rightarrow \mathbf{t}\,(\mu_B : s^2_B/n_B) \text{ avec } n_B - 1 \text{ degrés de liberté}$$

En supposant que les deux échantillons sont indépendants, les règles d'opération sur les variables aléatoires nous permettent de déduire que la différence $D = m_A - m_B$ suivra elle aussi une loi du t de Student, ayant pour moyenne $\Delta = \mu_A - \mu_B$ la *véritable différence* des moyennes et comme variance s^2_D la *somme* des variances (car les variances ne se soustraient **jamais**), donc :

$$D = (m_A - m_B) \rightarrow \mathbf{t}\,(\mu_A - \mu_B : s^2_A/n_A + s^2_B/n_B)$$

Cette loi aura pour nombre de degrés de liberté la somme des degrés de liberté des deux moyennes m_A et m_B :

$$(n_A - 1) + (n_B - 1) = n_A + n_B - 2$$

Or, par définition 95% des valeurs d'une loi du t de Student sont situées dans un intervalle de $\pm$ $\mathbf{t_{(\alpha=0,05;\ n)}}$ écarts-types autour de sa moyenne, avec $t_{(\alpha=0,05;\ n)}$ la valeur lue dans la table du t de Student pour un risque α=0,05 et n degrés de liberté. L'écart-type de la moyenne est comme d'habitude la racine carrée de la variance, on a ici :

$$s_D = \sqrt{(s^2_A/n_A + s^2_B/n_B)}$$

On en déduit l'intervalle de confiance de la différence $m_A - m_B$ en déterminant la valeur du t dans la table du t de Student pour le seuil α choisi et $n_A + n_B - 2$ ddl. (pour un intervalle de confince à 95%, on choisit $\alpha = 0,05$):

$$\Delta = \mathbf{D \pm t_{(\alpha,\ nA+nB-2\ ddl)}}\ \sqrt{(s^2_A/n_A + s^2_B/n_B)}$$

Exemple 6.8 : IC95% de la différence de taille entre des étudiants fictifs

Les données sont les suivantes (*seuls les effectifs changent* par rapport à l'exemple 6.7).

Etudiants : $n_A = 7$; $m_A = 178,6$ cm ; $s^2_A = 36,9$
Etudiantes: $n_B = 8$; $m_B = 166,5$ cm ; $s^2_B = 33,1$
D = $m_A - m_B$ = 178,6 – 166,5 = **12,1cm**. Quel est l'intervalle de confiance de D à 95% ?

Les effectifs sont très petits, mais la taille dans l'espèce humaine est distribuée approximativement de manière **normale**, donc, m_A et m_B suivent approximativement des lois du t de Student avec respectivement $n_A - 1 = 6$ ddl et $n_B - 1 = 7$ ddl

$s^2_{mA} = s^2_A/n_A = 36,9/7 = 5,271$
$s^2_{mB} = s^2_B/n_B = 33,1/8 = 4,138$

La variance de D est donc $s^2_D = s^2_{mA} + s^2_{mB} = 5,271 + 4,138 = 9,409$
Son écart-type $s_D = \sqrt{s^2_D} = \sqrt{9,409} = 3,067$ cm

La loi suivie par D est une loi du t de Student avec $n_A + n_B - 2 = 7 + 8 - 2 = 13$ ddl
La valeur seuil de la table est $t_{(\alpha=0,05,\ 13ddl)}$ = **2,16**

Borne inférieure de l'IC95 : 12,1 – **2,16** × 3,067 = 5,475 = 5,5 cm
Borne supérieure de l'IC95 : 12,1 + **2,16** × 3,067 = 18,724 = 18,7 cm

$IC_{95\%}$: environ [5,5—- 18,7cm]
Amplitude : 13,2 cm

La différence de taille réelle dans la population dont ces échantillons sont issus est donc connue cette fois avec une **très mauvaise précision**, qui reflète la petite taille des échantillons.

6.4 Intervalle de confiance d'une différence entre deux pourcentages.

6.4.1 Grands échantillons (np et $nq > 5$)

Le raisonnement est strictement le même que pour le calcul de l'intervalle de confiance entre deux moyennes calculées sur de grands échantillons. Dans le cas des pourcentages, cependant, souvenez-vous que la notion de *grand* échantillon est différente : il faut que :

$$n_A\, p_A > 5 \quad \text{et} \quad n_A\,(1-p_A) > 5$$
$$n_B\, p_B > 5 \quad \text{et} \quad n_B\,(1-p_B) > 5$$

avec les proportions p exprimées en fréquences, c'est à dire entre zéro et 1.

Deux proportions observées p_{obsA} et p_{obsB} calculées sur des tels "grands"échantillons de taille n_A et n_B suivent des lois de distribution approximativement **normales** de moyennes p_A et p_B (les véritables valeurs des proportions dans les populations A et B) et de variances $p_A \times q_A / (n_A - 1)$ et $p_B \times q_B / (n_B - 1)$,

$$p_{obsA} \rightarrow \mathbf{N}\,(\,p_A : p_A \times q_A/(n_A - 1))$$
$$p_{obsB} \rightarrow \mathbf{N}\,(p_B : p_B \times q_B/(n_B - 1))$$

Les échantillons étant grands, ceci reste approximativement valable en remplaçant les valeurs inconnues p_A et p_B par les valeurs observées p_{obsA} et p_{obsB} dans le calcul des variances :

$$p_{obsA} \rightarrow \mathbf{N}\,(\,p_A : p_{obsA} \times q_{obsA} / (n_A - 1))$$
$$p_{obsB} \rightarrow \mathbf{N}\,(p_B : p_{obsB} \times q_{obsB} / (n_B - 1))$$

avec q = (1 – p)

En supposant que p_A et p_B sont indépendantes, les règles d'opération sur les variables aléatoires nous permettent de déduire que la différence observée D = $p_{obsA} - p_{obsB}$ entre ces deux lois normales suivra *elle aussi* une loi normale,
(i) ayant pour moyenne $\Delta = p_A - p_B$ la *véritable différence* des proportions:

$$D = p_A - p_B$$

(ii) ayant pour variance la *somme* des variances (souvenez vous que les variances ne se soustraient **jamais**) :

$$s^2_D = p_{obsA} \times q_{obsA} / (n_A - 1) + p_{obsB} \times q_{obsB} / (n_B - 1)$$

En résumé :

$$D \rightarrow \mathbf{N}\,(p_A - p_B : p_{obsA} \times q_{obsA} / (n_A - 1) + p_{obsB} \times q_{obsB} / (n_B - 1))$$

Or, 95% des valeurs d'une loi normale sont situées à moins de 1,96 écarts-types de sa moyenne. L'écart-type étant la racine carrée de la variance, on a ici

$$s_D = \sqrt{(p_{obsA} \times q_{obsA} / (n_A - 1) + p_{obsB} \times q_{obsB} / (n_B - 1))}$$

On en déduit l'intervalle de confiance de la différence $D = p_{obsA} - p_{obsB}$:

$$\Delta = \mathbf{D} \pm \mathbf{1{,}96} \sqrt{(p_{obsA} \times q_{obsA} / (n_A - 1) + p_{obsB} \times q_{obsB} / (n_B - 1))}$$

6.4.2 Petits échantillons (*np* ou *nq* <5)

Il y a plusieurs solutions. La plus sophistiquée est d'utiliser le Bootstrap (voir section 6.5 intervalle de confiance de *tout ce que vous voulez*), mais on peut employer des moyens bien plus rudimentaires. Par exemple la méthode suivante, qui a l'inconvénient majeur de vous fournir en fait un intervalle de confiance à environ *un pour mille*.

En utilisant la table des intervalles de confiance pour les pourcentages estimés sur les petits échantillons, déterminez les bornes A_{inf}, A^{sup}, B_{inf} et B^{sup} des intervalles de confiance à 95% des deux proportions. Ces $IC_{95\%}$ seront sous la forme suivante :

$IC_{95\%}$ de A :$[A_{inf} — A^{sup}]$
$IC_{95\%}$ de B :$[B_{inf} — B^{sup}]$

Le but est de construire l'IC de la différence D = A – B, soit $[D_{inf} — D^{sup}]$

Commençons par l'écart le plus grand, et supposons que A > B. Cet écart maximum se produira si on a simultanément $A > A^{sup}$ et $B < B_{inf}$. Or, par définition de l'intervalle de confiance à 95%, il n'y a que 2,5% de chances pour que $A > A^{sup}$, de même il n'y a que 2,5% de chances pour que $B < B_{inf}$. La probabilité combinée d'avoir les deux simultanément est de 0,025 × 0,025 = 0,000625 soit 0,0625%. Voilà déjà décrite la taille de l'écart D qui n'a que 0,0625% de chances d'être dépassé sous l'effet du hasard.

Reste l'écart le plus petit (voire une écart *dans l'autre sens*, avec B > A si les $IC_{95\%}$ se chevauchent largement !). Cet écart se produira si on a simultanément $A < A_{inf}$ et $B > B^{sup}$. Le même raisonnement que précédemment nous indique que cette probabilité combinée est de 0,0625%. Nous avons ainsi défini l'écart minimum (voire inverse) que peut avoir la différence entre A et B.

Le tout nous permet de tracer un intervalle de confiance à plus de 99% pour D. L'inconvénient est qu'il est très large. Pour tracer par cette méthode un véritable $IC_{95\%}$ de D, il nous faudrait en fait partir d'$IC_{70\%}$ des pourcentages observés, mais les tables ne donnent pas ces IC, il faut donc les calculer à partir de la loi binomiale, ce qui est assez long. Quelques exemples seraient les bienvenus. Les voici.

Exemple 6.9
A : 5 individus sur n = 10, donc A = 50% et $IC_{95\%}$ [19% —— 81%] (table des IC pour petits effectifs)
B : 5 individus sur n = 50, donc B = 10% et $IC_{95\%}$ = [1% —— 15%] (table des IC pour petits effectifs)

Ecart maximum pour D : si A > 81% et B < 1%. Alors D > 80%. Proba : 0,0625%
Ecart opposé pour D : si A < 19% et B > 15%. Alors D < 4%. Proba : 0,0625%

L'$IC_{>99\%}$ de D est donc [+ 4% —— + 80%]

Exemple 6.10
A : 2 individus sur n = 10, donc A = 20% et $IC_{95\%}$ = [3% —— 56%] (table des IC pour petits effectifs)
B : 5 individus sur n = 50, donc B = 10% et $IC_{95\%}$ = [1% —— 15%] (table des IC pour petits effectifs)

Ecart maximum pour D : si A > 56% et B < 1%. Alors D > 55%. Proba : 0,0625%
Ecart opposé pour D : si A < 3% et B > 15%. Alors D < – 12%. Proba : 0,0625%

L'$IC_{>99\%}$ de D est donc [– 12% —— + 55%]

6.5 Intervalle de confiance de *tout ce que vous voulez.*

6.5.1 Présentation générale des méthodes de re-échantillonnage

Les méthodes de re-échantillonnage sont surprenantes. En fait, il n'existe probablement pas d'autres procédures statistiques qui donnent autant l'impression de *se moquer du monde*. Ne tentez jamais de les expliquer à vos étudiants un 1er avril, ils croiraient forcément que vous leur faites une blague. Vous pensez que j'exagère ? Alors imaginez la situation suivante. Vous voulez acheter une voiture, et le modèle que le vendeur est en train de vous présenter ne vous convient pas. Vous demandez donc à voir une autre voiture. Le vendeur vous répond alors "mais naturellement, aucun problème", puis il sort de sa poche une espèce d'énorme couteau multi-fonctions, démonte le rétroviseur de la voiture, puis vous la montre à nouveau fièrement en vous disant : "En voilà une autre, qu'en pensez vous" ?

Vous répondriez probablement quelque chose comme : "Vous me prenez vraiment pour un abruti ? Ca n'est *pas* une nouvelle voiture, c'est la *même*, vous avez juste enlevé le rétroviseur !". Et pourtant, les méthodes de rééchantillonnage fonctionnent exactement selon ce principe, et tout le monde trouve ça normal. Lisez plutôt :

6.5.2 La technique du Jackknife

Le *Jackknife* est un très robuste couteau multifonctions américain avec lequel on peut par exemple scier une branche ou couper un fil de fer barbelé (voire démonter un rétroviseur). C'est l'outil à tout faire du bricoleur[6]. On a baptisé de cette manière une technique de réechantillonnage particulière parce qu'elle est, comme un Jackknife, très rudimentaire, mais efficace quand même.

Le principe du Jackknife consiste à créer de soi-disant *nouveaux* échantillons en se servant... *de votre propre échantillon* dont on aura exclu à chaque fois un[7] élément

[6] le surnom du bricoleur est *Jack-of-all-trades*, d'où le nom du couteau

[7] on peut exclure théoriquement n'importe quel nombre d'éléments mais cela présente peu d'intérêt, la technique habituellement employée consiste à exclure un élément à la fois.

différent ! Vous voyez maintenant l'analogie avec la voiture et son rétroviseur ? Donc en gros la procédure est la suivante :

1. Prenez une voiture (=votre échantillon de *n* données)
2. Calculez la variable ***V*** qui vous intéresse
3. Démontez le rétroviseur (= enlevez *une* donnée de votre échantillon)
4. Re-calculez ***V***, mais en utilisant ce "nouvel" échantillon
5. Remontez le rétroviseur (= remettez l'élément enlevé dans l'échantillon)
6. Démontez une roue (= enlevez une *autre* donnée, ne touchez plus au rétroviseur)
7. Re-re-calculez ***V***, en utilisant ce "nouvel" échantillon
8. Remontez la roue (replacez dans l'échantillon la donnée enlevée)
9. Persévérez jusqu'à avoir démonté et remonté entièrement la voiture pièce par pièce
(= jusqu'à avoir passé en revue les *n* "nouveaux" échantillons résultant de l'élimination temporaire d'une des *n* données à chaque fois).

Complètement fou n'est ce pas ? Et pourtant ce procédé est *correct*, et il est décrit dans les ouvrages de statistiques les plus sérieux (ou alors une vieille blague de premier avril traîne dans les manuels depuis plus de 30 ans). A la fin de cette procédure, vous allez avoir sur les bras (1 + ***n***) estimations différentes de ***V***, à savoir :

• **une** estimation initiale ***V*** calculée sur l'échantillon complet des *n* individus

• ***n*** estimations (notées v_1, v_2 ... v_n) calculées sur les *n* "nouveaux" échantillons de (*n*–1) individus chacun, créés par re-échantillonnage au sein de votre propre échantillon.

Les *n* "nouvelles" estimations vont vous permettre de calculer autant de *pseudo-valeurs* Φ selon la formule suivante :

$$\Phi_1 = n\,V - (n-1)\,v_1$$
$$\Phi_2 = n\,V - (n-1)\,v_2$$
$$\ldots$$
$$\Phi_n = n\,V - (n-1)\,v_n$$

La distribution de ces *n* pseudo valeurs vous donnera une idée, imparfaite mais exploitable, de *la distribution du paramètre **V** dans la population dont votre échantillon est issu*, et ce, même si apparemment votre échantillon ne pouvait vous fournir qu'une seule valeur de ***V***. La technique du Jackknife, comme toutes les techniques de reéchantillonnage, permet en fait de *simuler* ce qui s'est produit lorsque vous avez échantillonné dans la grande population, mais en utilisant votre propre échantillon pour modèle. Ainsi, par un tour de passe-passe (qui ressemble à première vue à une escroquerie intellectuelle), on parvient à accéder à une distribution qui semblait hors d'atteinte. Si vous avez peu d'individus, vous pourrez pratiquement considérer (de manière prudente) que la gamme de pseudovaleurs trouvées représente l'intervalle de confiance. Si vous avez beaucoup de données, vous pouvez exclure les 10% les plus hautes et les 10% les plus basses et créer ainsi un très approximatif $IC_{80\%}$

Il faut noter que le Jackknife n'est d'aucune utilité pour calculer l'intervalle de confiance d'une moyenne, car dans le cas particulier où c'est *m* qui est calculée, les pseudo valeurs du Jackknife ne font que reproduire... les valeurs de l'échantillon. Le Jackknife sera donc surtout utile pour déterminer les intervalles de confiance de variances ou de choses plus complexes (telles un indice de Shannon par exemple).

Exemple 6.11 : Jackknife d'une variance avec 15 quadrats

A = {0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 2, 5, 10, 25, 45}; m_A = 6 ind/quadrat, s^2_A = 160,1

les 15 pseudovaleurs de la variance obtenues par Jacknife sont :

28,5 (7 fois) /15,8 (3 fois) / 5,5 / 0 (car valeur négative) / 403 / 1742 (moyenne 160,3: es 115,9):

$IC_{"95\%"}$ très approximatif de la variance réelle : ($s^2_A \pm 2$ es) = [0 —— 390]

6.5.3 La technique du Bootstrap

Les *bootstraps* sont les boucles de cuir souvent cousues en haut des bottes, et qui aident à les enfiler. L'expression anglo-saxonne "*se soulever du sol en tirant sur ses bootstraps*" signifie "se tirer d'affaire tout seul avec les moyens du bord", mais fait référence volontairement à un acte impossible (se soulever du sol en tirant sur ses bootstraps est aussi illusoire que faire avancer un bateau à voiles avec une soufflerie installée *sur le bateau lui même* : ça ne marche pas[8] !). Le bootstrap statistique, lui, consiste effectivement à se débrouiller avec les moyens du bord, c'est-à-dire l'échantillon lui-même, et ça marche. Le bootstrap est devenu LA méthode de re-échantillonnage (on peut démontrer que le Jackknife n'est jamais qu'une simplification extrême et très grossière du bootstrap). Le bootstrap peut être utilisé pour calculer l'intervalle de confiance d'absolument n'importe quoi avec une précision (apparente !) au pour-cent près. En réalité, naturellement, la véritable précision du Bootstrap dépend de la taille de l'échantillon. Une valeur de bootstrap calculée sur un petit échantillon n'est donc pas du tout aussi précise qu'elle en a l'air.

Le bootstrap est né de la capacité des ordinateurs à effectuer des calculs répétitifs très rapidement. C'est là son seul inconvénient : il nécessite impérativement un ordinateur, et si possible avec un logiciel approprié. On peut certes envisager d'effectuer un bootstrap artisanal avec un tableur du type Excel et un peu de patience, mais cela restera fastidieux à mettre en place. Quant à espérer faire un bootstrap avec une simple calculatrice, c'est possible en théorie, mais pas du tout en pratique.

La méthode est la suivante, (je garde initialement l'analogie de la voiture contenant *n* pièces détachées).

[8] Dans le cas du bateau, la poussée vers l'avant exercé par la soufflerie sur la voile est contrebalancée par la poussée vers l'arrière que reçoit la soufflerie du fait de la réaction de l'air mis en mouvement. En fait, comme la voile ne captera pas 100% de l'air qui est soufflé, le bateau aurait même théoriquement plutôt tendance à *reculer*.

1. Prenez votre voiture (=votre échantillon de *n* données)

2. Démontez une pièce *choisie au hasard* (= tirez au hasard *une* donnée de votre échantillon)

3. Obtenez une pièce identique chez le fournisseur et mettez la de côté(=notez la valeur de votre donnée)

4. Remontez la pièce d'origine sur la voiture (=les tirages se font avec remise)

5. Démontez à nouveau *une pièce choisie au hasard.* Il peut donc s'agir éventuellement de la pièce que vous veniez juste de remonter, si le hasard est d'humeur taquine (c'est le principe même du tirage avec remise)

6. Poursuivez la procédure de tirage *avec remise* jusqu'à obtenir autant de pièces détachées qu'il y en a dans votre voiture (=continuez jusqu'à avoir tiré *n* données, *avec remise*, au sein de votre échantillon). Comme les tirages se font *avec remise*, ne vous étonnez pas d'avoir trois volants mais une seule roue, c'est le hasard qui décide (le même individu peut être tiré plusieurs fois, certains ne le seront pas du tout).

7. Calculez alors la valeur V de la variable qui vous intéresse.

Attention, voici le moment pénible

8. Recommencez ces 7 premières étapes... ***mill e fois*** (c'est le nombre consacré, pour un bootstrap). Vous comprenez maintenant pourquoi il vaut mieux avoir un logiciel prévu pour.

Une fois que vous avez rassemblé vos 1000 estimations différentes de *V*, chacune étant établie sur un échantillon de *n* données tirées avec remise dans l'échantillon original, dressez le graphe de la distribution de *V* obtenue. Eliminez les 2,5% de valeurs les plus élevées de la distribution, et les 2,5% de valeurs les plus basses (ce besoin de précision explique de devoir effectuer tant de répétitions). Les valeurs qui restent sont situées dans l'intervalle de confiance à 95%, il est donc enfantin de déterminer ses bornes. Enfin, donnez une petite tape amicale à votre ordinateur, il vient de vous économiser des mois de calcul à périr d'ennui.

En l'absence de logiciel approprié, vous pouvez faire un bootstrap rudimentaire en recommençant la procédure 100 fois seulement au moyen d'un simple tableur. La seule "difficulté" consiste à faire un tirage avec remise. Pour cela :

1) copiez *n* fois dans une colonne chacune de vos *n* valeurs observées (si vous avez 10 individus votre colonne contiendra donc 100 chiffres, chaque valeur étant répétée 10 fois).

2) copiez la formule = alea() , c'est à dire <signe égal>alea<suivi d'une parenthèse ouverte et fermée immédiatement> dans une colonne adjacente de même hauteur. Cette formule effectue dans chaque case un tirage aléatoire dans une loi uniforme entre zéro et un.

3) sélectionnez vos deux colonnes et triez en fonction de la colonne contenant les nombres aléatoires.

Les 10 premiers chiffres de votre colonne initiale représentent autant de tirages avec remise dans votre échantillon initial, et permettent de calculer le paramètre (moyenne, variance etc.) dont vous voulez estimer l'intervalle de confiance. Avec un peu d'astuce, vous pourrez répéter cette manoeuvre un grand nombre de fois sans trop d'efforts. En éliminant les 5 valeurs les plus élevées et les 5 valeurs les moins élevées parmi 100 répétitions, vous obtiendrez un IC90% rudimentaire.

Dans le cas où vous voulez calculer l'IC de la différence D entre deux moyennes (ou pourcentages) dans le cas de petits échantillons, les choses se corsent un peu puisqu'il vous faudra simuler des tirages aléatoires avec remise dans les deux échantillons, en calculant à chaque fois D, ce qui vous permettra d'obtenir une distribution d'une centaine de D vous fournissant, selon le raisonnement ci dessus, un IC90% rudimentaire de D.

Résumé du chapitre 6.

Vos moyennes, vos pourcentages et autres paramètres estimés dans la population ne sont, justement, que des *estimations*. Il est vital d'avoir une idée de la *précision* avec laquelle ces paramètres ont été estimés. Le calcul des intervalles de confiance permet de répondre à ce besoin. Les IC sont très faciles à calculer dans certains cas (grands échantillons, lois normales) mais plus compliqués dans d'autres (petits échantillons de loi inconnue). Leur connaissance est cependant impérative dans tous les cas, sous peine d'avoir une fausse impression de précision. Les intervalles de confiance permettent dans un deuxième temps de connaître la *magnitude* de l'effet observé (dans les cas où on compare plusieurs populations), c'est-à-dire la taille vraisemblable de la différence réelle entre les deux moyennes ou pourcentages observés. Ici encore, la simulation informatique permet de traiter les cas les plus complexes, pour lesquels il n'existe pas de solution analytique toute faite.

7. Les tests statistiques : une saga faite de risques, d'erreurs et de rêves de puissance

Le *principe de base* des tests statistiques est assez simple à *expliquer*, et les tests eux-mêmes sont devenu dangereusement simples à *appliquer* grâce aux logiciels. Il est cependant *difficile* de les comprendre *vraiment*. Quant à en saisir toutes les finesses, il suffit de voir comment les spécialistes se corrigent les uns les autres... Si vous êtes comme moi, vous allez vraisemblablement traverser des cycles de lumière et d'obscurité (dont je ne vois toujours pas la fin) sous la forme "J'ai compris ! c'était donc ça ! — quoique finalement non — si, cette fois c'est bon ! — mouais, sauf que dans ce cas particulier, heu... — eureka ! — damned..." etc.

Je vais dans un premier temps présenter la méthode standard utilisée dans les manuels d'introduction aux statistiques. Nous irons ensuite plonger dans l'origine historique des tests, ou nous découvrirons que la réalité est plus... complexe. Le décor sera alors planté pour le coup de théâtre spectaculaire du chapitre suivant[1].

7.1 Principes de base.

En très bref, un test statistique consiste à :

(1) poser une hypothèse, nommée Ho (H zéro), ou "hypothèse nulle" .

(2) Calculer, dans la gamme des résultats expérimentaux possibles, ceux qui sont tellement *éloignés* du résultat moyen attendu selon Ho, que ces résultats n'ont *presque aucune chance de se produire si Ho est vraie.*

(3) Comparer ces résultats avec celui qui a été réellement obtenu.

(4) Conclure que Ho est *peu crédible* (et donc la rejeter) si le résultat obtenu appartient aux résultats qui n'avaient *presque aucune chance de se produire si Ho était vraie.* Donner le résultat du test en indiquant la probabilité P d'observer des résultats encore plus éloignés de Ho que celui qui a été observé dans l'expérience. Ex : $P = 0{,}001$

(5) Conclure que Ho *reste crédible* (et donc ne pas la rejeter) si le résultat obtenu appartient aux résultats qui avaient une chance, même *relativement* modeste, de se produire *si Ho était vraie.*

En sciences, le "presque aucune chance" se traduit par "dans *moins* de 5% des cas où Ho est vraie". Ou encore "avec une probabilité $P < 0{,}05$ sous Ho".

[1] *Voilà* qui s'appelle faire monter le suspense.

Exemple 7.1 Où sont les hommes ?

(1) Ho : "Le pourcentage de garçons chez les étudiants de la licence de Rennes 1 préparant au métier de professeur des écoles (instituteur/trice) est de 50%".

(2) Si cette hypothèse est correcte (donc si $p_{garçon}$ = 50%) alors sur une promotion de 50 étudiants on observera un pourcentage de garçons inférieur à 36% ou supérieur à 64% dans moins de 5% des cas (car si p = 50% alors l'$IC_{95\%}$ d'un p_{obs} sur 50 individus est [36,0 — 64,0%]).

(3) Or, je constate qu'il y a seulement 5 petits veinards — pardon, 5 garçons — et *45 filles* dans la promotion de cette licence, soit p_{obs}= 10%.

(4) J'en conclus qu'il est *peu vraisemblable* que le pourcentage de garçons soit vraiment de 50% dans la population dont sont représentatifs les étudiant(e)s de cette licence (c'est à dire la population des étudiant(e)s choisissant de devenir instituteur/trice). Je peux même affirmer (soyons audacieux) que ce pourcentage est vraisemblablement *inférieur* à 50%.

Remarque 1. On sait évidemment depuis des lustres qu'il y a plus de filles que de garçons dans ce métier, ce test statistique ne nous apprend donc rien d'intéressant à lui tout seul (ce qui est le cas de beaucoup de tests statistiques, comme nous le verrons au chapitre suivant).

Remarque 2. *Quel peut bien être l'intérêt* de ce test par rapport au calcul de l'intervalle de confiance du pourcentage de garçons observé (qui est : [1,6 — 18,4%]) qui non seulement permet la même conclusion mais fournit une idée de la *précision de notre estimation* et permet par la même occasion de calculer les bornes de l'intervalle de confiance de *l'écart concret* entre la valeur observée et la valeur théorique de 50% [31,6—48,4%]? C'est une bonne question, débattue également dans le chapitre suivant. Patience.

7.2 Détail des étapes d'un test statistique.

7.2.1 Choix de Ho

Pour des raisons expliquées plus loin, l'hypothèse Ho sera habituellement *soit* du type "rien à signaler" (Ho : les moyennes µA et µB sont *égales*), *soit* une valeur *ponctuelle* (Ho : p = 10%). Cette hypothèse est choisie de manière à pouvoir connaître la distribution d'une certaine variable aléatoire de test (que je nommerais T) *si Ho est vraie*. On dira qu'on connaît la distribution de T ***sous Ho***. Cette variable aléatoire T sera basée sur une moyenne, un pourcentage, une différence entre deux moyennes ou entre deux pourcentages, peu importe, l'essentiel est qu'on connaisse sa distribution *sous Ho*, donc *si Ho est vraie*. Pensez-vous que si je répète encore une demi douzaine de fois "*si Ho est vraie*", vous comprendrez tous dès maintenant que la probabilité P que fournit un test statistique est la probabilité d'observer un certain type de résultat *si Ho est vraie* et non pas la probabilité *que Ho soit vraie* ? J'en doute un peu. Ceci dit, vous finirez par le comprendre, même si c'est probablement une des notions les plus retorses en stats (et ça n'est pas la compétition qui manque !).

Le choix de Ho est déroutant pour les débutants, car il donne l'impression de fonctionner à l'envers. En effet, les expériences sont conçues pour détecter (et quantifier) des *effets*, ou des *phénomènes*. On s'attend donc spontanément à des hypothèses du type "le résultat dans le groupe traité va être *différent* de celui du groupe témoin". Or, la mécanique des tests statistiques utilise habituellement des hypothèses Ho du type "le traitement n'a *aucun* effet", que le test est capable de *rejeter*. Dans les tests statistiques tels qu'ils sont enseignés de manière classique, l'hypothèse Ho sera donc du type "rien à signaler". Exemples :

Ho : "la moyenne de la population est *égale* à la moyenne théorique"
Ho : "il y a autant de mâles que de femelles (i.e. 50%) dans la population"
Ho : "le pourcentage de gauchers est *identique* chez les hommes et les femmes"
Ho : "le coefficient de corrélation entre la taille et le poids est *nul*"
Ho : "l'azote n'a *aucun effet* sur le rendement du gougnafier à fleurs bleues"

Il semblait pourtant raisonnable, si on soupçonnait un effet particulier, de poser les hypothèses de travail en supposant une *direction* de l'effet, par exemple :

Ho : "la moyenne de la population est *supérieure* à la moyenne théorique"
Ho : "il y a *davantage* de femelles que de mâles dans cette population
Ho : "le pourcentage de gauchers *est plus* élevé chez les hommes"
Ho : "le coefficient de corrélation entre la taille et le poids est *supérieur à 0,8*"
Ho : "l'azote *diminue* (!) le rendement du gougnafier à fleurs bleues"

Car bien entendu, ce sont ces possibilités là qui nous intéressent, ce sont elles qui nous disent qu'il "se passe quelque chose", ce sont pour elles que les expériences sont réalisées. Cependant, il est techniquement beaucoup plus facile de *rejeter* une hypothèse basée sur un *point* (Ho : "50% de garçons") parce qu'elle est *largement* contredite par les résultats (p_{obs} = 10% de garçons), que de *valider* une hypothèse, surtout si elle est constituée d'une multitude de possibilités. Une hypothèse comme "il y a *davantage* de femelles que de mâles" est une hypothèse composée d'une infinité de possibilités : elle sera vraie si le pourcentage de femelles est en réalité de 51%, de 60%, ou même de 99%. Comment calculer une gamme de résultats qui n'auraient *presque aucune chance de se produire si Ho était vraie*, s'il y a une infinité d'hypothèses Ho? Que conclure par exemple si p_{obs}=49% de garçons ? Clairement, vu les fluctuations d'échantillonnage, cette observation serait *très compatible* avec l'hypothèse "51% de filles", mais serait *presque totalement incompatible* avec l'hypothèse "90% de filles".

En posant "Ho : le pourcentage de mâles *est de 50%*" et que si le test rejette Ho, on dispose de trois informations : (i) on sait que le sex-ratio est déséquilibré, (ii) on sait dans quelle direction il est déséquilibré, (iii) la probabilité associée au test nous permet de savoir à quel point.

7. 2. 2 Calcul de la zone de rejet de Ho

Connaissant la distribution de T, on peut déterminer quelles sont ses valeurs les plus extrêmes, les plus éloignées de la valeur moyenne attendue, autrement dit les valeurs *qui n'ont presque aucune chance d'être observées si Ho est vraie.* Notre connaissance de la distribution de T sous Ho, nous permet en particulier de calculer avec précision les gammes de valeurs extrêmes qui seront observées avec une probabilité α que nous pouvons choisir librement. L'idée est bien entendu de choisir α petit, car souvenez-vous que nous souhaitons connaître les valeurs qui n'ont *presque aucune chance* d'être observées si Ho est vraie. En sciences, le risque *maximum* que l'on consente est α = 0,05 soit *pas plus de 5% des cas*. Supposons ici que $\alpha = 0,05$

Nous voilà nantis de deux[3] zones de rejet de Ho, contenant $\alpha/2$=2,5% de la distribution de T sous Ho. L'une de ces zones de rejet concerne les 2,5% des valeurs extrêmes de T "incroyablement élevées" par rapport à la valeur moyenne attendue sous Ho. L'autre

[3] il peut n'y en avoir qu'une groupant 5% des valeurs à elle seule, exemple : la distribution du Chi2

zone de rejet concerne les 2,5% des valeurs extrêmes de T "incroyablement basses" par rapport à la valeur moyenne attendue sous Ho.

7.2.3 Calcul de la valeur de la variable de test

Avec les données observées, nous calculons la valeur de T. Soit T_{obs} cette valeur. Il est évidemment possible de calculer T_{obs} puisque T est choisie à la base selon deux critères (1) pouvoir être calculé avec les données, (2) avoir une distribution connue *sous Ho*.

Maintenant, le moment de vérité[4].

7.2.4 Verdict du test

Si T_{obs} appartient à une des deux zones de rejet de Ho, on... rejette Ho (logique) selon l'argument que si Ho était vraie, on n'aurait (presque) jamais pu observer une telle valeur de T. Il est donc plus parcimonieux d'accepter l'hypothèse selon laquelle Ho est probablement fausse (ce qui expliquerait très facilement le résultat obtenu).

Si T_{obs} n'appartient pas à une des deux zones de rejet de Ho, on ne rejette pas Ho (toujours aussi logique), ce qui signifie qu'on considère ne pas avoir d'éléments suffisamment solides pour la déclarer peu crédible. Il est de la plus **extrême** importance que vous compreniez *que cela ne signifie pas qu'on a démontré que* Ho est vraie. J'insiste, cela signifie simplement qu'on n'a pas suffisamment de raison de soupçonner que Ho soit fausse.

7.3 Les risques du métier.

Les décisions que nous avons prises ne vont pas sans risque, et on peut en distinguer deux : le risque d'avoir rejeté Ho alors qu'elle était vraie, le risque de ne pas avoir rejeté Ho alors qu'elle était fausse. Restez bien assis : malgré tout ce que vous pourrez lire, *il n'y a (généralement) aucun moyen de connaître ces risques*. En revanche, on peut très précisément en calculer d'autres. Le tableau suivant présente la situation. La rangée du haut présente la réalité : peut être que Ho est vraie (la probabilité en est p(Ho), inconnue), mais peut être qu'elle est fausse (et 1-p(Ho) est tout aussi inconnue). Les lignes correspondent aux différentes décisions possibles, qui sont bonnes ou mauvaises selon la situation. Chaque case comporte une probabilité, qui n'est pas vraie dans l'absolu, mais conditionnée à la probabilité de la colonne dans laquelle elle se trouve.

Tableau 7.1 Les règles de décision lors d'un *test d'hypothèse*

	Réalité (à jamais inconnue, habituellement)	
Décision	Ho est vraie : P(Ho)	Ho est fausse : [1 – P(Ho)]
Ho rejetée	mauvaise décision, **risque** α	**bonne décision**, **(1 – β) = puissance**
Ho non rejetée	**bonne décision**	mauvaise décision, **risque** β

On conclut de ce tableau que :
la véritable probabilité de *rejeter Ho par erreur* est P(Ho) × α **et non pas** α
la véritable probabilité de ne pas rejeter Ho *à tort* est [1– P(Ho)] × β **et non pas** β

[4] vérité " probable" seulement, hélas. En statistiques, vous ne serez jamais SUR de quoi que ce soit.

7.3.1 Le risque de première espèce, α

Lorsque nous rejetons Ho parce que la valeur de T_{obs} se situe dans la zone de rejet, nous savons que ce résultat pouvait se produire dans (au maximum) 5% des cas *si Ho était vraie*. Nous maîtrisons donc totalement ce risque là, puisque c'est nous qui le choisissons. Ce risque est nommé α, risque de première espèce[5]. Sa définition *correcte* est "le risque de rejeter Ho *si elle est vraie*". En revanche, il est presque systématiquement compris (*incorrectement*) comme "le risque de rejeter Ho *alors qu'elle était vraie*" ou "le risque de rejeter Ho *par erreur*". Ca n'est *pas* la même chose. Je vous avais prévenu, c'est subtil (et je me suis certainement fait piéger moult fois à dire, voire écrire, une chose pour l'autre). Le *véritable* risque de rejeter Ho par erreur est *inconnu,* tout simplement parce qu'il dépend de la probabilité que... Ho soit vraie, qui est inconnue (en général). La bonne nouvelle est que cette erreur d'interprétation est sans conséquences : le véritable risque de rejeter Ho par erreur est forcément... inférieur à α puisque la probabilité que Ho soit vraie est inférieure ou égale à 1.

7.3.2 Le risque de seconde espèce, β

Nous voulons ne rejeter Ho qu'en étant sûrs de nous. Cela signifie que nous ne le ferons que pour des valeurs de T_{obs} extrêmes, des valeurs vraiment *très peu crédibles si Ho est vraie*. Si jamais une valeur de T_{obs} est assez peu crédible sous Ho mais pas *très* peu crédible, nous conservons Ho faute de preuves, par prudence. Vous voyez quel est le problème : nous allons passer à côté d'effets qui n'auront pas été suffisamment forts pour provoquer une valeur de T_{obs} extrême. Ce risque, le risque de ne pas rejeter Ho *si elle est fausse* est β, le risque de seconde espèce. Ce risque ne peut pas être calculé sans apporter d'autres éléments que ceux qui figurent dans le tableau. En effet, Ho "$\mu_A = \mu_B$" peut être fausse de milliers de manières différentes (exemple : "$\mu_A - \mu_B = 2$ " mais aussi "$\mu_A - \mu_B = 2000$ "...), et pour chacune d'elle, il existe un risque β spécifique. Le risque β global est donc la combinaison de *l'infinité* des possibilités d'écarts à Ho avec *chacune* son risque β spécifique. Inutile donc d'essayer de le calculer globalement, d'autant plus que lui aussi dépend de la probabilité (inconnue) que Ho soit vraie.

7.3.3 La puissance du test, (1 – β)

Sachant que β est totalement hors d'atteinte, s'intéresser à **(1 – β)** semble un bon moyen de complètement perdre son temps. Comme quoi les apparences sont trompeuses. Ce paramètre est d'une importance capitale, si on ne veut pas passer sa vie à faire des manips sans jamais pouvoir rejeter Ho. Il se nomme **puissance** (tout un programme), et représente la probabilité de rejeter Ho *si elle est fausse*. Autrement dit, c'est la probabilité de voir quelque chose, *quand il y a quelque chose à voir*. On conçoit qu'un chercheur ne puisse rester indifférent à un tel paramètre, et cherche à en augmenter la valeur par tous les moyens. Ces moyens, au nombre de trois, sont bien connus, même s'il n'est pas toujours possible de les utiliser tous les trois :

[5] *Type I error* en anglais

Augmenter l'effectif — Ce moyen est d'une efficacité évidente puisque la taille de l'échantillon augmente la précision de vos estimations (il diminue les intervalles de confiance). Malheureusement, on a vu que cette augmentation de précision est proportionnelle à la *racine carrée* de l'effectif : si vous voulez une précision *dix fois* meilleure, il vous faudra... *cent fois plus* d'individus. Votre capacité d'augmenter la puissance ainsi n'est donc pas illimitée pour des raisons de coûts, de temps, de matériel (sans même parler des problèmes *éthiques et réglementaires* si ce sont des organismes supérieurs, voire des patients, qui constituent l'échantillon).

Limiter la variance — Le *second* moyen consiste à planifier son expérience de manière à réduire au maximum les variations des facteurs autres que celui qui est testé : utiliser quand c'est possible des individus du même sexe, du même âge, génétiquement proches, ayant grandi dans les mêmes conditions, standardiser au maximum toutes les procédures expérimentales, effectuer les observations dans la même plage horaire, avec le même expérimentateur etc...

Favoriser les effets spectaculaires — Ce moyen consiste, lorsqu'on manipule le traitement librement, à utiliser une "cause" de grande intensité, de manière à maximiser les chances de déceler un effet. Exemple : si vous soupçonnez que {la substance X/le facteur F} a un effet, testez la/le *d'abord* à forte dose/avec une forte intensité. Si un effet est effectivement décelé, vous pourrez ensuite réduire la dose/l'intensité et établir la relation dose/intensité—effet.

La puissance est un thème si important qu'il fera l'objet d'un chapitre à lui tout seul. Pour que le débat puisse être plus concret, ce chapitre est placé en fin d'ouvrage, c'est dire après vous avoir présenté l'utilisation des tests.

7.4 Les sources historiques d'un problème actuel.

Sans vouloir déflorer le sujet du chapitre suivant, tout n'est pas rose et consensuel dans le monde merveilleux des tests statistiques. Et ça ne date pas d'hier. Mais au fait, d'où viennent tous ces tests ?

7.4.1 King Fisher et les tests de *significativité* [6]

Le temple originel des tests statistique existe. Il s'agit de la station agronomique anglaise de Rothamsted, où travaillait un certain Ronald Aylmer Fisher, devenu par la suite *Sir* R. A. Fisher par la grâce de *Her Majesty The Queen*, pour services rendus à la science. Surmontant un lourd handicap de départ (il venait des mathématiques dites pures), RA Fisher a su se mettre à la portée des utilisateurs (de modestes ingénieurs agronomes, ou leur équivalent anglais) et a développé toute une série d'outils mathématiques extrêmement impurs, c'est-à-dire utilisables par le chercheur moyen, qui sont universellement en usage aujourd'hui. On lui doit le concept de l'hypothèse nulle Ho, le seuil *complètement anecdotique* $\alpha = 0{,}05$ (même s'il ne l'appelait pas α ni "risque de

[6]: CM Casado, *Fisher: La statistique, entre mathématique et expérience*, Ed RBA, juillet 2018

première espèce"), seuil de 5% auquel Fisher n'accordait qu'une importance pratique et non sacrée. On lui doit aussi les fameuses valeur *P* fournies par les tests statistiques.

Dans l'esprit de Fisher, le but du test est de rejeter des hypothèses nulles du type "pas d'effet" de manière à trier rapidement parmi des tas de traitements (agronomiques) lesquels ont probablement un effet.

La procédure utilisée par Fisher pour procéder à un test statistique est la suivante :

(1) fixer une hypothèse nulle du type "pas d'effet";

(2) calculer la probabilité *P* d'observer des données encore plus éloignées de Ho que ce qui est effectivement observé lors de l'expérience;

(3) considérer que plus *P* est faible, moins Ho est crédible.

Un point, c'est tout. En particulier, *ne prendre aucune décision* sans avoir fait des répétitions de l'expérience. Fisher utilise donc la probabilité *P* du test comme une *mesure des présomptions qui pèsent à l'encontre de l'hypothèse nulle*. Il nomme cette présomption la *significativité* : plus *P* est faible, moins l'hypothèse nulle est crédible, et plus la *significativité* est élevée. Ce mot significativité renvoie à la notion que l'effet observé *signifie* quelque chose, il a un *sens*, il n'est probablement *pas dû au seul hasard*.

Fisher travaille dans un contexte d'expérimentation appliquée. Il réalise des *séries* d'expériences, et cette notion est fondamentale. Le seuil de $\alpha = 0{,}05$ est un simple *premier crible*. Lorsque $P > 0{,}05$ (test "non significatif"), Fisher préfère passer à autre chose, car il effectue un travail de défrichage, il y a beaucoup d'effets à découvrir et la vie est courte. Mais si $P < 0{,}05$, Fisher ne crie pas sur les toits "Hourra ! Les amis, j'ai découvert un effet !". Il se contente de conclure que ce traitement *vaut la peine qu'on s'y intéresse*, et lance une série d'expériences pour essayer de *répliquer* l'effet qu'il a — peut être — découvert. C'est seulement lorsqu'on connaît un protocole tel qu'une probabilité de $P < 0{,}05$ est presque systématiquement obtenue, répétition après répétition de l'expérience, qu'il s'estime satisfait. Cette attitude est à des années lumière de la manière moderne d'utiliser les tests. Vous noterez que Fisher se moque apparemment comme de son premier quintal de blé du risque béta et de la puissance. Non seulement ces notions ne font pas partie de son vocabulaire, mais il va même combattre avec acharnement les théoriciens (on y vient tout de suite) qui vont les introduire. Fisher ne néglige pas ces concepts parce qu'il se moque des chances de ne pas rejeter Ho si elle est fausse, mais parce qu'il travaille dans des conditions appliquées. A ce titre, il suit son feeling d'expérimentateur et n'applique pas lui même le critère $\alpha = 0{,}05$ avec rigidité. N'oubliez pas que Fisher compte surtout sur la *répétition* des expériences pour avoir confiance dans la réalité d'un effet réel. Malgré ses défauts (qui n'en a pas ?) Sir R. A. Fisher reste à ce jour le plus grand statisticiens de tous les temps.

7.4.2 Neyman et Pearson (le fils de son père)

Jerzy Neyman et Egon Pearson sont avant tout des mathématiciens (Dieu les bénisse) et non des expérimentateurs[7]. Ils trouvent probablement que la notion de test de *significativité* mis sur pied par Fisher (dont ils respectent tout à fait les grandes compétences, qu'ils ne mettront jamais en doute) est un peu trop empirique à leur goût. Surtout, ils ont repéré une faille dans son système, qui est la négligence du risque béta,

[7] Je caricature outrageusement. Neyman a touché un peu à tout en réalité.

le risque de ne pas rejeter Ho si elle est fausse. Ils remarquent également au passage que Fisher ne donne pas de nom à l'hypothèse selon laquelle Ho est fausse, et qu'il ne donne pas de règle de décision fixe pour accepter ou rejeter Ho. Ils vont remédier à tout ça et créer un cadre théorique solide, dans lequel on puisse prendre des décisions en utilisant une règle fixe, qui permet de maîtriser alpha et béta à long terme. On doit donc à Neyman et Pearson le tableau 7.1 et la notion de *test d'hypothèse* (et non pas de test de significativité). Ce sont eux qui vont choisir les symboles alpha et béta et donc les notions de risque de première et de seconde espèce. Enfin, contribution essentielle au débat, Neyman et Pearson introduisent la notion de puissance.

Le raisonnement de Neyman et Pearson est celui-ci : en se fixant une règle claire, et tout en sachant très bien qu'un test statistique isolé ne peut pas dire le vrai et le faux, on va pouvoir fixer notre risque alpha et béta à long terme, c'est-à-dire sur une infinité d'expériences. La règle du jeu devient donc la suivante :

(1) choisir une hypothèse Ho et une hypothèse H_1 explicite (H_1 peut être par exemple que $mu_B > mu_A$). Cette hypothèse H_1 est nommée alternative hypothesis, ce qui signifie en anglais "l'autre hypothèse" mais qui a été mal traduit en Français "hypothèse alternative";

(2) *fixer à l'avance* un risque alpha *ET un risque béta lorsque c'est possible* (ce qui est le cas en particulier si l'hypothèse H_1 est une valeur fixe), et par là même, une puissance pour le test;

(3) déterminer la zone de rejet de Ho;

(4) calculer la variable de test T;

(5) si T tombe dans la zone de rejet, on rejette Ho au risque alpha = 0,05 et on accepte donc H_1 ;

(6) si T ne tombe pas dans la zone de rejet, on choisit Ho au risque béta généralement inconnu (sauf exceptions).

Si ce cadre est respecté sur une grande série d'expériences, on rejettera Ho à tort dans une proportion alpha * p(Ho) et on conservera Ho à tort dans une proportion béta (1-P (Ho)) où béta est en fait la moyenne des différents β généralement inconnus.

Comme vous le voyez, selon le cadre théorique de Neyman et Pearson, *chaque* test statistique amène à une décision : soit on *choisit* H_1, soit on *choisit Ho.* Bien entendu, choisir Ho ne signifie pas qu'on a démontré qu'elle est vraie, mais signifie qu'on considère qu'elle est vraie, sachant qu'on a un risque de se tromper qui est béta * (1-p (Ho)). Par ailleurs, vous avez peut être remarqué l'absence remarquable de *P* dans cette histoire. C'est normal, Neyman et Pearson ne s'intéressent à *P* que très indirectement : pour déterminer si oui ou non T tombe dans la zone de rejet. Si T tombe dans la zone de rejet de Ho avec *P* = 0,04 ou bien *P* = 0,000004 la décision *est la même* : on rejette Ho au seuil alpha = 0,05. Je dis bien au seuil alpha = 0,05 et non au seuil alpha = 0,000004. Selon la logique à *long terme* de Neyman et Pearson, on ne s'amuse pas à changer d'objectif au gré des circonstances : si on s'est fixé comme objectif à long terme un risque alpha de 5%, on colle à ce seuil quoi qu'il arrive.

Les êtres humains étant ce qu'ils sont, Fisher est *immédiatement* rentré dans le lard de Neyman et Pearson. Dès la publication de leur premier article, il s'est même mis à tirer à boulets rouges sur leur système, qui était à ses yeux totalement technocratique et digne

de la planification communiste, bref totalement déconnecté de la manière dont la science était faite au jour le jour. Il y a eu par la suite une longue série d'échanges, toujours sur le mode de l'agression, entre les deux équipes, et chacun a campé sur ses positions, jusqu'à la mort de tous les protagonistes. Se rajoute à cela une histoire de famille et de gros sous, cocasse avec le recul historique, mais qui a eu des conséquences internationales inattendues. Le père de Egon Pearson était le grand Karl Pearson (l'inventeur du coefficient de corrélation, dont le nom complet est "r *de Pearson*"). Karl Pearson était l'éditeur en chef de la revue de biostatistiques *Biometrika*, et avait refusé à Fisher de reproduire dans son ouvrage certaines tables statistiques qui y figuraient, car il vendait lui même ces tables (il n'y avait ni photocopieuses, ni évidemment de logiciels statistiques à l'époque) et en tirait un petit revenu. En rétorsion, Fisher avait donc publié ses propres tables dans lesquelles, pour des raisons pratiques, il utilisait les seuils de alpha = 0,05 et alpha = 0,01. Egon Pearson devait noter par la suite avec humour que la publication de ses tables avait fait beaucoup pour l'établissement de standards internationaux du risque de première espèce qu'ils défendaient. Et voilà pourquoi encore à ce jour, les tables statistiques utilisent les seuils alpha = 0,05 et alpha = 0,01. J'ai moi même utilisé ces seuils (parce qu'il n'y a pas d'autres moyens quand on imprime des tables), mais ils constituent une survivance du passé. Dans les articles scientifiques modernes, on tend de plus en plus à indiquer la valeur exacte de P puisque c'est ainsi que les logiciels procèdent. Donner la valeur exacte de P est une très bonne chose, car cela tend à désacraliser le seuil de 0,05. Ceci dit, cette tradition change lentement, ce qui est le propre des traditions.

7.5 L'approche moderne : le beurre, l'argent du beurre, et une belle béchamel...

Si vous relisez la section 7.1 qui définit les étapes d'un test statistique "moderne" vous constaterez qu'il s'agit en fait d'une hybridation sauvage entre l'approche de Fisher et celle de Neyman et Pearson. De Fisher, on a conservé Ho et P, tout en présentant les décisions sous l'angle de Neyman et Pearson (risques alpha et béta, puissance), mais en faisant valser alpha selon ce qu'on trouve, puisqu'on accorde une grande importance à la valeur *P*, qui est la première — voire la seule — chose que beaucoup de gens regardent lorsqu'ils font un test statistique, et qui est fortement mise en valeur lorsqu'on présente triomphalement les résultats d'un test, résultats qui portent habituellement sur UNE expérience, sur laquelle on tire quand même des conclusions plus ou moins définitives.

En bref, tous les tests statistiques modernes sont effectués selon une méthode qui aurait mis Fisher dans une colère noire et qui auraient désolé Neyman et Pearson par leur aspect décousu. Le gag est que personne n'est capable d'expliquer comment cette synthèse hybride s'est mise en place, puisque les deux approches se combattaient farouchement. Toujours est-il que nous faisons actuellement nos tests statistiques d'une manière étrange, qui n'a été validée par aucun des deux courants de pensée ayant fondé le principe des tests statistiques. Cette situation est paradoxale, voire inquiétante, ce qui nous amène à découvrir... *le côté obscur de la force*, dans le chapitre suivant.

Deuxième partie

Sachez utiliser les tests statistiques

Avec Parsimoni et Abonessian

Les tests statistiques sont un sujet difficile, mais également – comme vous le découvrirez peut-être avec surprise – hautement polémique. J'ai donc fait appel à deux fins experts, qui interviendront tout au long de cette seconde partie, chaque fois qu'ils en ressentiront le besoin. Ils se connaissent depuis longtemps et se chamaillent à tout propos (comme il sied à des experts), mais savent toujours tomber d'accord sur l'essentiel.

Giuseppe Parsimoni occupe depuis de très nombreuses années la chaire d'économie en statistiques de l'université méditerranéenne de Chevapiano. Farouche partisan des intervalles de confiance, sa vision des tests statistiques actuels est extrêmement critique. Il soutient que dans neuf cas sur dix, le calcul d'un intervalle de confiance autour des valeurs estimées, ainsi que le calcul de la magnitude de l'effet observé (avec son intervalle de confiance également) sont largement suffisants (et supérieurs à un test) pour répondre concrètement et intelligemment à la question posée. Ses ouvrages majeurs incluent *Statistica al' economia*, *Testi i tutti quanti ma non tropo,* ainsi que le lapidaire *Data e basta !*. Son chef d'oeuvre est bien entendu *Il principio di Parsimoni*, traduit en 25 langues et largement utilisé en biologie évolutive.

Tigran Abonessian dirige d'une main de fer le *Black Sea Institute for the Wise Use of Modern Statistics* de Testanova, au bord de la Mer Noire. Tout en admettant volontiers le bien fondé de nombreuses critiques de son collègue Parsimoni, il soutient que les tests statistiques gardent leur mot à dire dans le monde scientifique moderne, à condition de les utiliser avec pertinence, et seulement pour ce qu'ils sont, et non comme des oracles miraculeux. Il est l'auteur de plusieurs ouvrages sur l'usage inadapté des tests, dont *ANOVA is Not a dying star, Kurtosity killed the cat*, ainsi que *A test in need is a friend indeed*. Son oeuvre majeure est cependant *On The Origin of Slopiness by Means of Statistical Confusion.*

8. La fin des tests statistiques ?

Il y a quelque chose de pourri au royaume du Danemark

SHAKESPEARE (Hamlet)

Viens, viens découvrir le côté sombre de la force, Luke.

DARTH VADOR (Star Wars)

Attention : si vous ne voulez pas perdre définitivement votre foi dans l'infaillibilité de la Science, ne lisez surtout pas ce chapitre ! Fuyez pendant qu'il en est encore temps !

Bon, vous l'aurez voulu.

Une information alarmante est habituellement passée sous silence, dans l'introduction des dizaines de manuels d'introduction aux tests statistiques qui garnissent les étagères des bibliothèques universitaires à l'attention des débutants (et des moins débutants). Il s'agit du fait que l'utilisation des omniprésents tests d'hypothèses (Z, t, χ^2, ANOVA etc...) telle qu'elle est pratiquée dans les revues de recherche scientifiques du monde entier (autrement dit l'approche « H_0 contre H_1, si $P < 0,05$, je rejette Ho ») *est vigoureusement remise en cause depuis plus d'un demi siècle*. Plus perturbant encore, cette critique radicale (et de plus en plus pressante) n'est pas issue d'un collectif anarchiste ou d'un ramassis de feignants incultes, allergiques aux mathématiques et n'ayant jamais analysé des données de leur vie. Bien au contraire, la charge contre l'utilisation traditionnelle des tests d'hypothèse est menée depuis 1935 environ par des statisticiens chevronnés et des chercheurs très expérimentés, qui utilisent les statistiques dans leur travail de recherche. William Thomson, un chercheur de la Colorado State University, a recensé dans la littérature scientifique du plus haut niveau, plus de 400 articles et de chapitres d'ouvrages (voire d'ouvrages entiers) sur ce thème ! Les plus anciennes de ces protestations remontent aux années 1930 et sont apparues dès la mise au point des tests (autrement dit, on ne vient pas de s'apercevoir du problème !). La vague de critiques a cru et embelli dans les années 50, 60, 70, 80, 90 (avec la publication par un groupe de psychologues d'un ouvrage (contesté) intitulé sobrement "**What if there were no significance tests ?**"[1]) et elle n'a rien perdu de sa vigueur, bien au contraire. Un symposium portant entièrement sur la question a rassemblé des statisticiens à Buffalo en 1998, et une *Task Force* spéciale comme seuls les américains en ont le secret a

[1] LL Harlow, SA Mulaik, JH Steiger (2016). *What if there were no significance tests?* Routledge Classic Editions, NewYork.

été formée par la *American Psychological Association* pour édicter des recommandations aux auteurs publiant dans les plus prestigieuses revues scientifiques de cette discipline. Le rapport de cette *task force* inclut un bon nombre des critiques acerbes pleuvant sur les tests statistiques, et enjoint à tout auteur désireux de publier dans les revues de l'APA d'en tenir compte.

J'ai enseigné les bases des statistiques à des étudiants de maîtrise de biologie pendant plusieurs années, sans ***jamais*** – honte à moi – avoir entendu parler de cette polémique pourtant mondiale et ancienne, preuve qu'elle ne fait pas franchement partie des enseignements traditionnels que j'ai moi même reçus, et je n'ai encore rien lu à ce sujet dans un *manuel* de statistiques. De là à penser qu'il y a un complot mondial pour nous cacher certaines choses, il n'y a qu'un pas. Il serait évidemment grotesque de le franchir, et 400 articles scientifiques en accès libre dans de prestigieuses revues à comité de lecture sont là pour témoigner du fait que ces problèmes sont en fait identifiés depuis longtemps, même s'ils semblent avoir du mal à diffuser en dehors de la sphère des spécialistes. Je remercie donc chaleureusement mon collègue, le Pr. Jean-Sébastien Pierre (éthologue et *véritable* biomathématicien, lui) de m'avoir fait découvrir le côté sombre de la force, c'est-à-dire cette polémique troublante sur la légitimité des tests statistiques, qui m'a amené à faire ma petite enquête sur la question. Imaginez un instant mon désespoir initial ("quoi, après tout ce que j'ai souffert pour apprendre à utiliser ces satanés tests, maintenant on vient me dire qu'il ne faut plus s'en servir ???")

Mais quelles sont ces critiques au juste ? Ecoutons le vénérable mais toujours énergique Giuseppe Parcimoni vous les décrire avec flamme :

> **G. Parcimoni** — Mes enfants, ne tombez pas dans le piège comme tous les moutons de Panurge qui vous ont précédés. Méfiez-vous des tests, leurs *P* qui veulent simplement dire "*P*oudre aux yeux", et leurs "étoiles" qui vous cachent le ciel et — pire encore — vos propres données ! Remplacez-les chaque fois que vous pourrez — c'est-à-dire presque tout le temps ! — par de bon vieux **intervalles de confiance**. On n'a jamais rien fait de mieux. Apprenez à utiliser les intervalles de confiance, car ils sont la base de tout et vous obligent à *regarder vos données*. Un bon intervalle de confiance vaut tous les tests du monde. Et si vous ne me croyez pas, méditez ceci :
>
> (1) Dans l'écrasante majorité des cas, les hypothèses Ho utilisées par les tests sont du type "*aucun* effet". Or, elles sont presque forcément *fausses* (et parfois, on le sait même parfaitement dès le départ !) car tout ou presque a un effet sur tout ou presque, même si l'effet en question est minusculissime. Le fait qu'un test soit statistiquement significatif revient donc la plupart du temps à *enfoncer une porte ouverte*. On se doute bien que la substance X va avoir un effet sur le taux de division cellulaire. La vraie question qu'on se pose est en fait : **quelle est la magnitude** de cet effet. Ce calcul nécessite de se concentrer sur les *valeurs* obtenues avec ou sans la substance X, et sur leur *fiabilité*, donc, sur leurs *intervalles de confiance*.
>
> (2) Puisqu'en général, il y a *toujours* un effet (aussi minuscule soit-il), il suffit d'un échantillon suffisamment grand pour montrer que presque *n'importe quoi* est statistiquement significatif. La belle affaire ! En revanche, le fait qu'un effet soit statistiquement significatif *n'apporte aucune information concrète* sur la **magnitude**

de l'effet en question (on ne peut donc pas mesurer son intérêt scientifique), ni sur la **précision** avec laquelle il a été estimé (on ne peut donc pas connaître la fiabilité de la magnitude de l'effet observé). En revanche, calculer les **intervalles de confiance** permet encore une fois de répondre, de manière naturelle, à ces questions fondamentales.

(3) Avec un échantillons suffisamment petit, on peut obtenir au contraire un résultat *non significatif* sur n'importe quoi, par simple manque de puissance du test. Le fait qu'un résultat ne soit pas statistiquement significatif, n'apporte donc aucune information non plus, si on s'en tient là (et 9 fois sur 10, on s'en tient justement là). Or, en général, la puissance des tests est faible, car l'habitude n'est pas encore prise d'estimer la puissance du test avant de lancer la manip, particulièrement parce que cela oblige à réfléchir en profondeur sur les objectifs de l'expérience et à obtenir des réponses précises à des questions difficiles. Le calcul d'un **intervalle de confiance** autour d'une valeur estimée, vous protège automatiquement contre ce risque, car il vous **montre** littéralement la gamme de valeurs dans lesquelles se trouve probablement la réalité. Observer que l'intervalle de confiance du pourcentage de mâles d'une population est [10 – 90%] vous empêche de conclure comme un automate "*J'ai testé l'écart à la valeur théorique 50% par un chi2, mon test est non significatif avec P = 0,85, donc le sex-ratio est sûrement équilibré*", ce qui est une aberration. Encore une fois, l'intervalle de confiance est bien plus pertinent et informatif que le test.

(4) Le fait que Ho ne soit pas rejetée (test non significatif) est trop souvent abusivement interprété comme une confirmation (au moins implicite) de Ho (pas d'effet), alors que dans la plupart des cas, les résultats seraient *également compatibles* avec des hypothèses *très distinctes* de Ho (justement à cause du manque de puissance du test). On passe ainsi à côté d'effets intéressants. La réponse : regardez donc la taille de votre intervalle de confiance, et la sagesse reviendra immédiatement.

(5) Lorsque Ho est rejetée (test significatif), beaucoup de chercheurs confondent la probabilité *P* du test avec la probabilité que Ho soit vraie (par exemple, *P* =0,001 est supposé signifier que la probabilité que Ho soit vraie est 0,001). Or, la probabilité *P* est la probabilité d'observer les données (ou des données encore plus éloignées de Ho) **si Ho est vraie.** Il n'existe en fait **aucun** moyen de connaître la probabilité que Ho soit vraie, que ce soit avant le test ou après, sauf cas **très** particuliers. J'ajoute que la probabilité *P* du test ne change pas de nature quand le test est significatif, puisque c'est **nous** qui décidons **arbitrairement** que le test devient "significatif" lorsque $P < 0,05$! Dans l'exemple du sex-ratio cité au point 4, la proba $P = 0,85$ ne signifie en rien que le sex-ratio à 85 chances sur 100 d'être équilibré !

(6) La manière dont les résultats sont présentés dans les articles scientifiques rend les méta-analyses (la synthèse entre plusieurs études, en combinant leurs résultats pour gagner en précision) très difficiles. En particulier, un effet peut être jugé non significatif dans dix études de suite utilisant de trop petits effectifs, *alors même que les dix résultats vont dans le même sens*, sans qu'il soit possible de synthétiser les résultats (ce qui permettrait de mettre en évidence qu'il y a bien un effet, et de le quantifier). Les intervalles de confiance sont largement supérieurs à cette approche, car ils concentrent l'attention sur *la valeur estimée elle même*, qui est tout de même la base de tout ! Ils nous poussent à comparer les valeurs *concrètes, réellement observées* dans les différentes études, et pas le résultat des "tests", qui sont désincarnés.

(7) En bref, les scientifiques modernes s'hypnotisent beaucoup trop sur la "significativité statistique" de leurs tests, au lieu de se consacrer aux questions importantes qui sont (i) quelle est *l'importance scientifique* de l'effet mesuré, qui est fonction de la *magnitude* de l'effet (*effect size*) et du *domaine scientifique* considéré

(en physique des particules, même un effet infinitésimal peut avoir une importance théorique très grande), et (ii) avec quelle *précision* (intervalle de confiance) cet effet a t-il été mesuré ?

Alors, faut-il du passé faire table rase et envoyer les tests d'hypothèses traditionnels aux poubelles de l'histoire ? Certains chercheurs (en particulier dans le domaine de la psychologie expérimentale) pensent que **oui** et l'expriment avec force. D'autres ont protesté **contre** cette idée, avec tout autant de force. En revanche, il existe entre les deux camps un vaste consensus (y compris dans le monde des statisticiens purs et durs) sur le fait que les tests statistiques sont souvent **mal utilisés**, qu'on attend d'eux des réponses pour lesquels ils ne sont **pas conçus**, et qu'on accorde à leurs résultats une importance si exagérée qu'elle va jusqu'à **éclipser les données elles mêmes**. C'est pourquoi, d'autres spécialistes plus modérés prêchent plutôt (ouf !) pour une approche moins mécanique (voire "religieuse") des tests. Vous pouvez télécharger sur internet des articles récents qui font le point très clairement sur la question, et je vais juste ci-dessous faire un petit tour d'horizon des critiques listées plus haut, en m'efforçant de faire la synthèse de tout ce que j'ai pu lire à ce sujet. C'est le moment de laisser la parole à Tigran Abonessian.

T. Abonessian— Sacré Giuseppe ! Il tape dur, mais je dois admettre qu'il tape juste. Disons simplement que, comme tous les gens du Sud, il exagère parfois un chouïa. Voyons ses critiques d'un peu plus près.

(1) Les hypothèses Ho du type "aucun effet" sont fausses dès le départ.

Giuseppe est dans le vrai... la plupart du temps. Il a raison de souligner qu'il est difficile d'imaginer un facteur qui réussisse à n'avoir *aucun* effet sur une variable aléatoire, en tout cas dans le cadre d'expériences réelles, dans lesquelles on étudie des facteurs qui peuvent être *raisonnablement soupçonnés* d'avoir un effet sur la variable aléatoire examinée (personne à ma connaissance ne cherche à mettre en évidence, par exemple, un effet de la production annuelle de betterave française sur la fréquence des taches solaires, ou un effet du signe astrologique sur la pression sanguine— quoique...). On peut se faire l'avocat du diable et "critiquer cette critique" en faisant remarquer que l'hypothèse Ho "l'effet de la transmission de pensée est strictement nul" est correcte jusqu'à plus ample informé, et que le scientifique qui serait capable de démontrer un effet de ce type, aussi faible soit-il, aurait fait grandement avancer nos connaissances ! Il existe donc bien, certes exceptionnellement, des hypothèses Ho qui soient à la fois *scientifiquement intéressantes à rejeter* et du type "pas d'effet *du tout*". Elles sont cependant une minorité. La plupart des hypothèses Ho que nous rejetons, sont des hypothèses Ho dont on sait dès le départ qu'elles sont fausses (on se doute qu'il y a un effet de A sur B), on veut surtout savoir s'il est **négligeable** (car très petit) ou **digne d'intérêt** (car suffisamment important). C'est pourquoi vous ne devez surtout pas voir un test significatif comme une fin, mais plutôt comme une *étape très préliminaire*. Un test significatif n'est qu'un *argument* (et non une démonstration absolue) en faveur de l'existence (et surtout de la **direction**) de l'effet que l'on soupçonnait. Il ne dispense pas d'estimer la magnitude de l'effet apparent (effect size) et son intervalle de confiance, étapes qu'il faut de toute manière effectuer, que l'effet soit significatif ou non, pour permettre de futures méta-analyses. Le seul moyen de démontrer un effet de manière relativement certaine, est d'être capable de *le répliquer un bon nombre de fois*. Aucune étude ponctuelle, (même si $P<0{,}0001$!) ne possédera jamais ce pouvoir, et il n'y a aucune raison de le lui donner. Les tests statistiques ont

été mis au point dans les années 20 par R.A. Fisher dans le contexte de *séries* d'expériences, pour faire le tri rapidement entre des effets significatifs (qu'il fallait impérativement *répliquer*) et des effets non significatifs (qu'il fallait se contenter de ranger dans la catégorie "reste à déterminer"). Cette philosophie a été pervertie par la suite, car le résultat des tests a été utilisé pour couler dans le bronze le résultat d'études ponctuelles. L'attitude raisonnable consiste donc, non pas à déboulonner ou brûler les idoles (les tests), mais à les faire descendre de leur piédestal. Ce ne sont pas les tests qui décident du vrai ou du faux. Ils ne sont qu'un indicateur utile.

(2) Il suffit d'un échantillon suffisamment grand pour montrer que n'importe quoi est statistiquement significatif.

Soulignée de multiples fois par le passé, cette vérité est une évidence pour les statisticiens professionnels, et elle ne peut piéger en théorie que les amateurs. Le problème est que nous sommes tous des amateurs. Même les scientifiques, qui sont évidemment très compétents dans leur domaine d'expertise (la biologie, l'écologie, la médecine...), ont bénéficié au cours de leurs études d'une formation en statistiques finalement assez modeste, voire sommaire, sauf exceptions. Bien entendu, chacun d'entre nous peut se bercer de l'illusion de faire partie de ces exceptions, mais il ne faut pas se voiler la face : les statistiques sont une discipline scientifique à part entière, une branche des mathématiques avec ses propres revues de recherche et ses congrès, et ça n'est pas un hasard s'il existe des statisticiens professionnels ayant passé une thèse dans ce domaine.

Ironiquement, la meilleure protection des chercheurs contre le danger de déceler des effets minuscules et sans intérêt (bien que statistiquement significatifs) est probablement... leur manque chronique de temps et de moyens. Je n'ai encore jamais entendu de chercheur se plaindre du fait qu'il avait trop de données. En général, lorsque nous décelons un effet, ça n'est pas un effet minuscule, tout simplement parce que la taille de nos échantillons ne rend pas nos tests suffisamment puissants pour cela ! En revanche, dans les rares cas où l'on peut manipuler de grands, voire de *très grands* jeux de données, il faut être conscient du fait que nous devenons capables de déceler des effets microscopiques, qui ne présentent pas forcément d'intérêt scientifique ou pratique. C'est tout l'intérêt du calcul de la magnitude d'un effet mesuré (effect size), avec son intervalle de confiance pour pouvoir discuter de cet aspect, et je suis pleinement en accord avec Giuseppe sur ce point.

(3) Avec un échantillons suffisamment petit, on peut obtenir au contraire un résultat *non significatif* sur n'importe quoi, par simple manque de puissance du test.

Cette remarque constitue le pendant évident de la remarque (2). Si votre échantillon est très (trop) petit, la puissance du test est faible, donc seulement les écarts quantitativement importants pourront être décelés. Dans la pratique, ce risque est largement plus élevé que le précédent, car le temps et les bras disponibles (sans parler du budget) vous obligeront souvent à travailler avec des tailles d'échantillons qui ne sont pas aussi grandes que vous le souhaiteriez. Donc, si votre expérience montre un écart dans la direction attendue, mais sans atteindre le seuil magique de 0,05 c'est peut être qu'il n'y avait rien à voir, mais c'est peut être aussi parce que votre test manquait de puissance. Il ne faut donc pas enterrer trop vite H_1. Si vous croyez toujours à votre idée, il n'y a aucune raison de ne pas persévérer, cette fois avec un échantillon plus grand. En revanche, sachez que calculer la puissance d'un test *a posteriori* (c'est-à-dire après le test) est complètement stérile. Si le test est non significatif (et qu'il y avait quand même un effet), alors par définition sa puissance était *trop faible*, et si le test est significatif alors sa puissance était *suffisante*. Un calcul de puissance, pour avoir un sens, nécessite que vous définissiez *a priori* (donc à l'avance) quelle est la magnitude de l'effet que vous jugez scientifiquement intéressant. C'est parfois

difficile, mais c'est seulement par rapport à cet effet là que le calcul de puissance a un sens.

Malheureusement, le manque de puissance attaché aux petits échantillons peut générer des effets pervers. En effet, il est parfaitement possible de choisir *volontairement* une taille d'échantillon si faible qu'on a pratiquement aucune chance de rejeter Ho. Mais franchement, qui ferait une chose pareille ? Et bien la vérité n'est peut être pas si reluisante. Dans une méta-analyse d'expériences en psychologie, il a été noté que la puissance moyenne des tests (la probabilité de rejeter Ho si elle était fausse) était plus faible lorsque l'hypothèse Ho était l'hypothèse privilégiée par les chercheurs, que lorsque Ho était une théorie qu'ils souhaitaient rejeter. S'il est volontaire, ce comportement est évidemment condamnable puisqu'il est malhonnête. Il est (souhaitons-le) inconscient. En clair, même si vous pensez que Ho est vraie, donnez sa chance à H_1 ! Pour cela, il faut une puissance raisonnable. Cette puissance étant fonction de la magnitude de l'effet à déceler, lui même dépendant de la discipline considérée, il est cependant impossible de donner une règle générale. Voir la vaste littérature sur la puissance statistique pour une discussion approfondie de la "bonne" taille des échantillons.

(4) Le fait que Ho ne soit pas rejetée est trop souvent abusivement interprété comme une confirmation (au moins implicite) de Ho.

Cette faute (conclure que Ho est vraie puisque le test est non significatif) ne devrait jamais être commise tant les mises en garde contre cette notion sont répétées toutes les deux pages dans les manuels d'introduction aux statistiques même les plus élémentaires. Et pourtant, c'est une des choses les plus difficiles à faire comprendre aux étudiants, quand on les initie à la pratique des tests d'hypothèse. On pourrait être tenté sournoisement d'en déduire que les étudiants sont intellectuellement limités. Il est plus raisonnable de conclure que la notion selon laquelle on ne peut *jamais* accepter Ho est tout simplement difficile à avaler (et n'oublions pas que les chercheurs les plus brillants sont souvent d'anciens étudiants). Que penser alors lorsque des "pros" se laissent glisser sur cette pente dangereuse en affirmant dans la conclusion de leur article que "A n'a pas d'effet sur B" simplement parce que le test était non significatif ? Une explication raisonnable est qu'il faut lire ce genre de déclaration entre les lignes, comme le raccourci plein de sous-entendus d'une phrase beaucoup plus lourde qui serait "*Je suis un chercheur compétent qui a réalisé une expérience en m'assurant que la taille de mon échantillon donnait suffisamment de puissance à mon test statistique pour déceler, avec une grande probabilité, un effet dont la magnitude serait digne d'intérêt (et je sais évidemment quelle magnitude est digne d'intérêt dans mon propre domaine de recherche), or aucun effet significatif n'a été décelé, donc, si effet il y a, il n'est pas d'une magnitude digne d'intérêt, en conclusion, je propose de dire qu'il n'y a "pas d'effet" parce que nous sommes entre chercheurs donc vous voyez ce que je veux dire, chers collègues*". Evidemment, cette phrase est un peu plus longue. Sans tomber dans ce genre de formulation ridicule, il est cependant utile de ne jamais donner l'impression qu'on a démontré Ho, ne serait-ce que parce que certains étudiants font l'effort de lire de véritables articles scientifiques et que, ne sachant pas (encore) lire entre les lignes, ils risqueraient de prendre de mauvaises habitudes.

(5) Lorsque Ho est rejetée, beaucoup de chercheurs confondent la probabilité *P* du test avec la probabilité que Ho soit vraie.

La tentation est effectivement irrésistible, lorsque votre test est significatif par exemple à $P = 0{,}001$, de conclure que la probabilité que Ho soit vraie est de une chance sur mille. Malheureusement, ça n'est pas ce que dit le test. Comme il est écrit dans tous les manuels, le test fournit en fait la probabilité d'avoir observé vos données

(ou des données *encore plus éloignées* de Ho que les votres) *si Ho est vraie*. Donc, si $P = 0{,}001$, tout ce qu'on peut déduire est que *si Ho était vraie*, alors on observerait vos résultats (ou des résultats *encore plus éloignés* de Ho) une fois sur mille. Il n'existe cependant aucun moyen au monde pour déduire de ces seules données la probabilité *que Ho elle-même* soit vraie, aussi étrange que cela puisse paraître. Il suffit pour s'en convaincre d'étudier le résultat d'un test pour lequel la différence observée entre le témoin et le traité est si petite que le résultat est non significatif, avec $P = 0{,}99$. Qui oserait en déduire que Ho (**aucun** effet) a 99 chances sur 100 d'être vraie, alors qu'il suffit que l'effet existe mais soit très faible, pour obtenir facilement le même petit écart entre le témoin et le traitement ?

Pire encore, il existe des situations dans lesquelles, malheureusement, la probabilité que Ho soit vraie est très élevée *même si le test est significatif*. Cohen (1994)[1] décrit un exemple édifiant de situation dans laquelle un patient diagnostiqué comme Schizophrène par un test clinique fiable à 95% (95% des schizophrènes testés sont détectés) et spécifique à 97%(97% des gens normaux testés sont jugés normaux) a une probabilité de plus de 60% de ne pas être schizophrène (alors que le test clinique le diagnostique comme schizophrène !). Voir Cohen (1994) pour les détails du calculs.

Prenons un autre exemple. Nous sommes en 1940. Monsieur Robert W. est un citoyen américain honnête de sexe mâle, qui paye ses impôts. On peut donc émettre fermement l'hypothèse Ho: "Robert W. est un homme". Il semble difficile de la rejeter, mais nous allons quand même essayer, au moyen d'un test statistique, basé sur une variable de test T qui sera tout simplement sa taille, car on connaît la distribution des tailles dans l'espèce humaine donc, en langage statistique, on connaît la distribution de T si Ho est vraie. En particulier, si Ho est vraie, la probabilité que $T >$ 2,70 mètres est inférieure à 10^{-9} (il y a moins d'une chance sur un milliard qu'un être humain mesure plus de 2,70 mètres), ce qui nous permet de définir une zone de rejet pour notre test au seuil très sévère $\alpha = 10^{-9}$ (car exclure à tort un être humain de notre espèce est un acte grave, nous voulons être sacrément sûrs de notre décision !). Notre règle de décision sera :

- si $T > 2{,}70$, on rejette Ho au risque $\alpha = 10^{-9}$ (un risque vraiment *infinitésimal*)
- si $T < 2{,}70$, on ne rejette pas Ho, autrement dit, on accorde à Robert W. le bénéfice du doute, et on refuse jusqu'à preuve du contraire de l'exclure de l'espèce humaine.

On effectue alors l'expérience (c'est-à-dire la mesure) : stupeur, Robert W. mesure 2,72 mètres! La réaction bête et méchante serait : "Monsieur Robert W., j'ai le regret de vous dire que vous n'êtes pas un homme, un vrai. ($P < 10^{-9}$)", mais évidemment personne de sensé ne dirait une chose pareille (en tout cas je ne m'y risquerais certainement pas face à un type qui fait deux mètres soixante douze !). L'intérêt de cet exemple, est de montrer que *la probabilité P* associée au test n'est pas du tout *la probabilité que Ho soit vraie*. Ici, on connaissait parfaitement la probabilité de Ho avant même d'effectuer le test. En effet, Robert W. étant un citoyen américain de sexe mâle payant ses impôts, la probabilité qu'il soit un homme était *certaine*, elle était de 100%, elle valait 1, et pas du tout 10^{-9}. Ca n'a pas empêché le test de donner un $P < 10^{-9}$. A l'évidence, il peut donc exister un très grand écart entre le *P* donné par le test et la probabilité réelle de Ho.

Rappelez-vous bien une chose, c'est que le test nous dit **ceci :**

"*Si Ho est vraie (donc, si le citoyen américain de sexe mâle Robert W. est un homme), alors la probabilité que sa taille dépasse 2,72 m est* $P < 10^{-9}$" (c'est **exact**).

et non pas **cela :**

"*Sachant que le citoyen américain de sexe mâle Robert W. a une taille de 2,72 m, la probabilité qu'il soit un homme est* $P < 10^{-9}$" (c'est **faux**).

[1] Jacob Cohen (December 1994), *The Earth is round (p < .05)*, American Psychologist, 49 (12): 997–1003

Au passage, vous aviez peut être reconnu Robert Wadlow [1918-1940], citoyen américain qui reste, à ce jour, l'homme le plus grand de tous les temps.

En conclusion sur ce point, il est exact que *plus* la valeur *P* du test est faible, *moins* l'hypothèse Ho est vraisemblable, mais *on ne peut pas aller plus loin que cette relation "qualitative"*, et en particulier la valeur *P* du test ne permet en rien de connaître la probabilité exacte que Ho soit vraie.

(6) La manière dont les résultats sont présentés dans les articles scientifiques rend souvent les méta-analyses (la synthèse entre plusieurs études, en combinant leurs résultats pour gagner en précision) très difficiles.

C'est probablement moins vrai de nos jours, mais cette critique mérite qu'on s'y attarde. Elle fait référence aux articles qui se contentent de citer les résultats de significativité des tests (les valeurs de *P*) sans donner les estimations des moyennes et des effets observés eux-mêmes. Ce cas extrême réduisant toute la substance de l'article au résultat des tests effectués était (d'après ses détracteurs), semble-t–il encore courant il y a quelques années. Exemple caricatural (et fictif): "L'apport de 50kg de potasse/ha augmente le rendement du haricot par rapport à un témoin sans potasse ($P < 0,01$)". Si c'est toute l'information disponible dans la partie "résultats", on ne peut effectivement pas en tirer grand chose. La question qui vient immédiatement à l'esprit est bien entendu "*de combien* le rendement est-il augmenté ?" (magnitude de l'effet). Si la réponse est "de 10 quintaux à l'hectare", la question qui vient ensuite est alors "quelle est la *précision de l'estimation* de cet effet ?" (l'intervalle de confiance de l'effet est-il [9,5—10,5 q/ha] ou bien [2—18 q/ha] ?). On voudra aussi savoir par ailleurs quel était le rendement du témoin (10q/ha ou bien 100q/ha ?), qui donne une idée du niveau d'intensité de la culture.

Il est cependant difficile de croire que l'article (fictif) contenant cette phrase sur le haricot, ne mentionnerait pas les rendements obtenus concrètement dans le témoin et le traitement, avec leur erreur standard et les effectifs (nombre de répétitions). Ce sont ces informations dont les méta-analyses ont besoin. Il ne serait pas étonnant en revanche que cette phrase soit citée dans le résumé, et que les informations plus concrètes (valeurs obtenues, magnitude) n'y figurent pas. C'est contre cette tendance qui met trop en valeur les tests, par rapport aux résultats concrets, qu'on peut essayer de lutter, car les résumés sont très utiles aux auteurs de méta-analyses, qui ont besoin de pouvoir passer en revue un très grand nombre d'articles, le plus rapidement possible. Pour garder notre exemple agricole, on pourrait vouloir effectuer une méta-analyse pour savoir, par exemple, quel est *en général* l'effet moyen d'une dose de 50kg de potasse à l'hectare sur le haricot (sachant que de nombreux paramètres entrent en jeu : variété utilisée, climat, le type de sol, façon de mener la culture etc...). L'auteur de cette méta-analyse va donc chercher à passer en revue toutes les études dans lesquelles on a testé l'impact de l'engrais potassique sur le haricot. Si les seules informations qu'il y trouve sont du type "l'engrais potassique à 50kg/ha a un effet positif, $P < 0,01$" on comprend tout de suite qu'il n'y a rien à en tirer en dehors d'un comptage élémentaire du type "S*ur les 500 études réalisées, l'apport de potasse à la dose de 50kg/ha avait un effet positif significatif sur le rendement du haricot dans 495 études et un effet non significatif dans 5 études*". La seule "conclusion" qu'on pourrait dégager de cette débauche d'énergie serait alors "*Cette fois mes petits gars, c'est certain, l'engrais potassique à 50kg/ha, c'est bon pour les haricots*", ce dont on pouvait vaguement se douter avant même de se lancer dans la méta-analyse ! Il est impossible d'accumuler un savoir utile, donc chiffré, dans ces conditions. Cohen (1994) cite Tukey (1991), qui faisait remarquer avec humour à propos de la notion d'élasticité en physique :

> *Si, par exemple, la notion d'élasticité avait été restreinte à "quand on tire dessus, ça s'allonge !", alors la loi de Hooke, la limite d'élasticité, la plasticité et beaucoup d'autres thèmes importants n'auraient pas pu*

apparaître. Mesurer les bonnes choses sur une échelle communicable nous permet de stocker de l'information à propos des quantités.

(7) En bref, les scientifiques donnent l'impression de s'hypnotiser sur la significativité statistique de leurs tests.

C'est souvent vrai, mais ils y sont un peu forcés aussi par la tyrannie des journaux scientifiques qui ne souhaitent publier que des résultats "significatifs". La publication d'un article scientifique moderne est donc une frénétique "chasse aux astérisques" (les symboles d'un test significatif) là où nos nobles anciens pouvaient prendre le temps de solidement asseoir leurs théories, en répétant de nombreuses fois leurs expériences, sans subir l'obligation de publier rapidement, le nombre de publications étant synonyme de vie (des crédits de recherche) ou de mort (pas de crédits de recherche). Le résultat est ce que nous en connaissons tous : un des plus prestigieux journaux scientifiques du monde, qui a bâti sa réputation sur la publication des grands scoops scientifiques du siècle (le plus célèbre étant la description de la structure de l'ADN par Watson et Crick) est obligé de publier presque dans chaque numéro des démentis, "corrigendum" et autres "erratum", parce que certaines découvertes annoncées dans le numéro précédent avaient été faites un peu... précipitamment. Les choses changeront de toute manière lentement, mais il n'est pas interdit d'espérer que l'on revienne à plus de modération dans ce domaine, en mettant plus les données — et non les tests — en valeur, comme Giuseppe le souhaite si ardemment.

Résumé du chapitre 8.

Tout n'est pas rose et consensuel dans le monde des statistiques. En un siècle, les sciences biologiques sont passées de l'absence totale d'analyse statistique (Pasteur, Darwin, et plus près de nous, Konrad Lorenz), à l'omniprésence obsédante des tests, y compris dans des domaines éloignés du laboratoire et du champ expérimental en petites parcelles rigoureusement répliquées qui ont vu la naissance des tests, et pour lesquels ils avaient été conçus. La formation statistique des biologistes ayant très imparfaitement suivi le formidable développement des méthodes statistiques (et c'est bien normal, les biologistes sont avant tout des biologistes), le risque est grand pour le praticien moyen d'utiliser les méthodes d'analyse de manière inadaptée, et de très nombreuses voix se sont élevées, depuis les années 20, pour dénoncer cette situation. Il ne faut donc jamais hésiter à aller consulter un véritable statisticien, si possible avant de réaliser l'expérience. Lui (ou elle) saura vous dire si votre protocole est suffisamment simple, si votre puissance de détection correspond à ce que vous espérez être capable de déceler et si vos résultats seront... analysables ! C'est son métier, qui a demandé des années de formations très spécifiques, et il (elle) effectuera forcément cette tâche bien mieux que nous, simples biologistes.

Pour ma part, sauf dans les cas d'analyse les plus simples que je sais traiter, je consulte **systématiquement** des gens plus forts que moi (ce qui n'est pas difficile à trouver !) plutôt que de faire des bêtises, et je pense que c'est la seule attitude possible.

9. Comparaison de moyennes

9.1 Comparaison entre une moyenne observée et une moyenne théorique.

Comme dans le cas des intervalles de confiance, les calculs sont différents selon qu'on a un grand ($n > 30$) ou un petit échantillon, mais l'approche générale est très similaire. Si vous ressentez, à la lecture de ce chapitre, une soporifique impression de déjà-vu, c'est plutôt bon signe, puisqu'elle signifie que vous commencez à bien connaître les raisonnements de base, qui sont toujours les mêmes.

9.1.1 L' échantillon est "grand" ($n > 30$)

C'est le cas idéal. En effet, notre ami le théorème de la limite centrale nous dit que, si une variable aléatoire X suit une loi quelconque de moyenne μ et de variance σ^2, alors la moyenne m calculée sur un grand échantillon de taille n, suivra une loi approximativement **normale** ayant la même moyenne μ, mais une variance n fois plus petite et valant donc σ^2/n. Comme déjà vu, dans le cas d'un grand échantillon, on commet une erreur tout à fait négligeable en remplaçant la valeur σ^2 (inconnue en général) par son estimation s^2 calculée sur l'échantillon. On a donc, avec une très bonne approximation :

$$m \to N\left(\mu : \frac{s^2}{n}\right)$$

Cette loi normale peut être ramenée à la loi normale centrée réduite N (0 , 1) par centrage-réduction comme d'habitude d'où, si on appelle Z la variable centrée-réduite :

$$Z = \frac{m - \mu}{\sqrt{\frac{s^2}{n}}} \to N(0:1)$$

Or, on sait que 95% des valeurs d'une loi normale quelconque sont situées dans un intervalle de 1,96 écarts-types autour de sa moyenne. Comme Z suit N (0 : 1), la valeur de Z va donc être comprise dans 95% des cas entre –1,96 et +1,96 puisque la moyenne vaut zéro et l'écart type vaut racine carrée de 1, c'est-à-dire 1. Si l'hypothèse Ho "*la moyenne du caractère dans la population échantillonnée est bien de* μ" est vraie, la valeur absolue du Z observé devrait donc 95 fois sur 100 être inférieure à 1,96. Le principe du test va donc être de rejeter l'hypothèse Ho chaque fois que l'écart Z observé est supérieur (en valeur absolue) à 1,96 :

Si $|Z| > 1{,}96$ on rejette Ho au risque $\alpha = 0{,}05$
Si $|Z| < 1{,}96$ on ne rejette pas Ho, au risque β inconnu, mais d'autant plus grand que la valeur réelle de la population est proche de μ.

La lecture dans la table de la loi normale N (0 : 1) permettra même de déterminer si on peut toujours rejeter Ho au risque $\alpha = 0{,}01$ (si $|Z| > 2{,}576$) voire $\alpha = 0{,}001$ (si $|Z| > 3{,}29$) etc. C'est cette valeur α que vous nommerez "*P*" en donnant le résultat de votre test.

Exemple 9.1. La taille moyenne des étudiants Français de maîtrise de sexe mâle étant supposée être de 1,76 m, peut-on dire que la taille moyenne des étudiants de maîtrise BPE s'en écarte significativement ? On dispose pour cela d'un échantillon de $n = 70$ garçons de maîtrise BPE (1998) ayant les caractéristiques suivantes : $m = 177{,}65$ cm, $s^2 = 40{,}6$.

$$Z = \frac{177{,}65 - 176}{\sqrt{\frac{40{,}6}{70}}} = 2{,}17 > 1{,}96$$

Au risque $\alpha = 0{,}05$ on rejette l'hypothèse Ho de l'absence de différence. Dans un rapport, on écrira : « *La taille des garçons de maîtrise BPE 1998 était significativement supérieure à celle de la population française des étudiants de maîtrise (Z = 2,17 ; P < 0,05)* ». Notez cependant que (i) la valeur théorique de 1,76 est purement inventée, (ii) la valeur 1,77 résulte d'un simple interrogatoire et non d'une mesure effectuée à la toise. Il existe ici un biais expérimental important car de nombreuses personnes connaissent leur taille très approximativement.

L'opinion de Parsimoni & Abonessian

Parsimoni — Voilà bien le type même du test complètement superflu et sans intérêt ! Tout ce qu'il conclut est "la moyenne observée est différente de la valeur théorique, $P < 0{,}05$". Il est beaucoup plus efficace de calculer directement l'intervalle de confiance de la valeur observée. Non seulement voit-on immédiatement si la valeur théorique est dans cet intervalle ou pas, mais on détermine en un coup d'oeil la gamme des valeurs plausibles de la valeur observée. Il est alors enfantin de calculer l'intervalle de confiance de la différence entre l'observé et le théorique. On peut ensuite écrire : l'intervalle de confiance de la différence D entre OBS et THEO est [borne inf — borne sup]. Faire un test statistique de comparaison de moyennes dans ces conditions est une totale perte de temps et apporte une information squelettique. De telles pratiques devrait être interdites !

Abonessian — Giuseppe exagère, c'est son côté latin. D'abord, quand on rend compte des résultats d'un test statistique, il ne faut pas écrire simplement "machin est significativement différent de truc" il faut préciser si "machin est significativement plus grand (ou plus petit) que truc" : le test statistique s'intéresse en premier lieu au sens de la différence observée. Deuxièmement, l'utilisation des seuils du type $P < 0{,}05$ est une survivance du passé (elle est liée à l'utilisation de tables, qui datent des années 1920). Dans la pratique moderne des tests statistiques, on préfère maintenant indiquer la valeur P exacte qui est donnée par le logiciel d'analyse statistique. Plus cette valeur est faible, plus on peut avoir confiance dans l'existence d'une différence entre OBS et THEO. Il est donc intéressant de connaître P. En revanche ce que dit Giuseppe sur l'intérêt de calculer l'intervalle de confiance de la moyenne est exact (cela permet de le représenter sur une figure), et je suis également d'accord avec le fait qu'il faille ensuite quantifier l'écart observé entre OBS et THEO et avoir une idée de la taille maximum et minimum qu'il peut avoir, car c'est cet écart qui nous intéresse scientifiquement parlant.

9.1.2 L'Echantillon est trop petit ($n < 30$) mais X suit une loi proche de la loi normale.

Vous pouvez alors tranquillement appliquer le même principe que ci-dessus *avec la différence que* la valeur seuil ne va plus être exactement de 1,96 pour un risque $\alpha = 0{,}05$. En effet la variable centrée-réduite dans laquelle on remplace σ^2 par s^2 va (parce que l'estimation s^2 est ici peu précise) suivre une loi du t de Student, dont la variance est plus grande que celle de la loi normale centrée-réduite (les distributions de Student sont plus aplaties, plus "étalées" que la loi normale centrée-réduite). Si on appelle cette variable t et non plus Z, on va observer des valeurs absolues de t en moyenne plus grandes que dans le cas précédent. Pour ne pas

augmenter artificiellement ainsi le risque α, c'est-à-dire le taux de rejet erroné de Ho, il faut donc augmenter la taille de l'intervalle autour de μ, pour lequel on ne va pas rejeter Ho. Il est ici impossible de donner une valeur fixe comme 1,96 qui serait valable quel que soit $n < 30$, car la distribution de la loi du t de Student varie en fonction de la taille de l'échantillon (très exactement en fonction du nombre de degré de liberté = $n - 1$). Il faut donc consulter la table du t de Student et lire la valeur t figurant à l'intersection de la ligne des ddl = $n - 1$ et de la colonne du risque α choisi (en général 0,05) cette valeur est la *valeur critique* du t. Le test s'écrit d'une façon générale :

$$t = \frac{m - \mu}{\sqrt{\frac{s^2}{n}}} \rightarrow t_{(n-1)ddl}$$

Si $|t| > t_{critique}$, on rejette Ho, au risque α choisi.
Si $|t| < t_{critique}$ on ne rejette pas Ho, au risque β inconnu, mais d'autant plus grand que la valeur réelle de la moyenne est proche du μ théorique.

Comme dans le cas du test Z, la table du t vous permettra de préciser α et donc la valeur "*P*" de votre test.

Exemple 9.2. Même problème que l'exemple 9.1, mais on ne dispose plus que de 10 étudiants. On va supposer (artificiellement) qu'on obtient la même variance et la même moyenne que précédemment, d'où : m = 177,65 cm et s^2 = 40,6. La valeur théorique reste 176 cm.

$$t = \frac{177{,}65 - 176}{\sqrt{\frac{40{,}6}{10}}} = 0{,}82 < 2{,}262$$

La valeur 2,262 est la valeur seuil de la table du t pour 9 ddl et α = 0,05. Cette valeur n'étant pas dépassée (et il s'en faut de beaucoup), on n'a pas de raison de rejeter « Ho : pas de différence » *sur la base de ces données*, **mais on n'a pas démontré pour autant l'absence de différence**. La preuve: on a vu dans l'exemple 5.1 qu'en réalité, il y a bel et bien une différence. La seule raison pour laquelle on n'arrive pas à la mettre en évidence ici est le manque d'individus. On écrira dans un rapport : «*Sur la base de nos données, la taille des étudiants de maîtrise BPE 1998 ne diffère pas significativement de la valeur nationale de 1,76 m (t = 0,82 ; 9 d.d.l. ; NS[1])* ». Toute personne sensée aura cependant soin de **relativiser** la portée de ce jugement dans la phrase suivante, en mettant en avant le faible pouvoir de détection du test (= faible puissance), dû à la très petite taille de l'échantillon.

L'opinion de Parsimoni & Abonessian

Parsimoni — tout ce que j'ai dit précédemment concernant la supériorité du calcul d'un intervalle de confiance (utilisant ici la valeur du t de Student) au lieu de se lancer dans un test statistique, reste valable. Je maintiens que ce genre de test est superflu et que le calcul d'un intervalle de confiance est plus riche d'informations.

Abonessian — et je maintiens de même que connaître *P* garde son intérêt, même si nous sommes bien d'accord sur le besoin de calculer un intervalle de confiance et la magnitude de l'effet éventuellement observé.

[1] NS = Non significatif (équivaut à écrire « *P > 0,05* »). Utilisé avec des majuscules ou des minuscules, selon les goûts.

9.1.2. L'échantillon est petit et la loi inconnue (ou connue pour être éloignée de la loi normale).

Il vous est alors impossible d'utiliser un test Z ou un test t de Student, car ces tests sont basés sur l'hypothèse de la normalité de la variable ou au moins de la moyenne calculée sur l'échantillon, le second cas nécessitant impérativement un grand échantillon. Cependant, tout n'est pas perdu: au prix d'une (modeste) perte de puissance, vous pouvez quand même utiliser une méthode *non paramétrique*, telle que le Boostrap, qui vous permettra de calculer l'intervalle de confiance de la valeur observée. Si la valeur théorique se trouve en dehors de cet intervalle, le test est significatif pour $\alpha = 0,05$.

L'opinion de Parsimoni & Abonessian

Parsimoni — Où l'on voit bien l'intérêt des intervalles de confiance. Ceci dit, pour ne pas se mettre dans ce genre de situation, il faut surtout travailler davantage de manière à avoir de grands échantillons !

Abonessian — Avoir de grands échantillons n'est pas toujours possible Giuseppe, vous le savez bien. Certains chercheurs travaillent sur des mutations rares, ou sur des explosions de supernovae qui ne se produisent qu'une fois par million d'années. On ne peut tout de même pas attendre la suivante pour augmenter l'effectif !

Parsimoni — Les jeunes n'ont plus aucune patience.

Abonessian — Face à un petit échantillon d'une loi fortement éloignée de la loi normale, une méthode non paramétrique telle que le Bootstrap peut toujours être appliquée, même si la fiabilité du résultat sera évidemment d'autant plus faible que l'échantillon est petit. Ceci dit, dans les autres cas, les moyennes sont rapidement distribuées selon un t de Student, et si vous avez plus d'une dizaine de données, vous commettrez une erreur modérée en utilisant la méthode classique du t de Student pour utiliser un IC à 95%.

Parsimoni — il est toujours piquant de constater que les grands statisticiens sont les premiers à tordre le cou de leurs propres règles.

Abonessian — En l'occurrence, je me fonde sur le résultat de simulations informatiques qui montrent que, dans une vaste gamme de situations réalistes, les moyennes convergent rapidement vers une distribution proche du t de Student.

Parsimoni — Amen. Si c'est un modèle informatique qui le dit, nous voilà pleinement rassurés.

Abonessian — Pourquoi alors cet air sarcastique ?

9.2 Comparaison de deux moyennes observées.

9.2.1. Les deux échantillons sont "grands" ($n > 30$)

Dans le cas d'échantillons A et B de taille n_A et $n_B > 30$, on va appliquer encore une fois le théorème de la limite centrale et en conclure que les moyennes m_A et m_B suivent des lois approximativement normales N (μ_A, σ^2/n_A) et N(μ_B, σ^2B/n_B) avec le bonus supplémentaire qu'on pourra (vu la taille des échantillons) remplacer sans problèmes les valeurs σ^2_A et σ^2_B inconnues, par leurs estimations s^2_A et s^2_B basées sur les échantillons. Notre hypothèse Ho (qui sera éventuellement rejetée) sera qu'il n'y a aucune différence significative entre m_A et m_B, autrement dit, que les eux échantillons proviennent d'une seule et unique population de

moyenne $\mu_A = \mu_B = \mu$. L'approche va consister à utiliser la **différence** observée $m_A - m_B$. Rappelez-vous à cette occasion, les règles d'opérations sur les variables aléatoires, on est ici dans le cas $Y = X_A - X_B$, avec, dans l'hypothèse nulle, deux tirages *dans la même loi*. Donc, si l'hypothèse Ho est correcte, la nouvelle variable devrait avoir une moyenne nulle (en espérance) et une variance égale à la *somme* des deux variances (rappel : les variances ne se soustraient *jamais*). En pratique, bien sûr, la différence $m_A - m_B$ observée ne sera jamais nulle, à cause des fluctuations d'échantillonnage. On s'attend donc en réalité à trouver une différence mais, et c'est ça qui est fondamental, *on connaît la loi de distribution de cette différence si Ho est vraie*. NB : il n'y a pas de covariance ici car les deux moyennes sont des variables indépendantes (ce qu'on trouve dans un échantillon n'a aucune influence sur ce qu'on trouve dans l'autre). Bref,

$$Y \to N\left(m_A - m_B = 0 : \frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)$$

Cette loi normale étant déjà centrée « automatiquement » (si Ho est vraie), on va donc se contenter de la réduire en divisant par son écart type égal à $\sqrt{(s^2_A/n_A + s^2_B/n_B)}$. Cette nouvelle variable va être :

$$Z = \frac{m_A - m_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \to N(0:1)$$

Revoici notre grande amie la loi normale centrée réduite. La suite devrait maintenant vous faire bailler distraitement tellement vous avez l'habitude :

Si |Z| > 1,96 on rejette Ho, au risque α = 0,05 (on conclut que les données ne sont pas compatibles avec l'hypothèse que les deux échantillons proviennent d'une population de même moyenne).

Si |Z| < 1,96 on ne rejette pas Ho, au risque β inconnu, mais d'autant plus grand que les valeurs réelles des deux populations sont proches, au cas où elles seraient différentes.

Ce qui a été dit plus haut sur le moyen d'utiliser la table pour rechercher *P* reste valable.

Exemple 9.3 n_A =71 garçons et n_B = 74 filles de la maîtrise BPE 1998 ont mesuré leur main droite avec un double décimètre. Les résultats sont les suivants : m_A = 18,98 cm, m_B = 17,59 cm s^2_A = 1,26, s^2_B = 0,69. Les tailles moyennes sont-elles significativement différentes ?

$$Z = \frac{18{,}98 - 17{,}59}{\sqrt{\frac{1{,}26}{71} + \frac{0{,}69}{74}}} = 8{,}45 >> 1{,}96$$

D'après la table de la loi normale, on peut rejeter *très* confortablement l'hypothèse que les garçons et les filles aient la même longueur de main moyenne : la valeur obtenue ici dépasse largement la valeur seuil pour α = 0,000 000 001 (qui est de 6,10). Autrement dit, si Ho était vraie, il y aurait *moins d'une chance sur cent millions* d'observer une telle différence, avec cette taille d'échantillon. Dans un rapport, on écrira : « *Les garçons de la maîtrise BPE 1998 ont une taille de main plus élevée que celle des filles, cette différence étant très hautement significative (Z = 8,45 ; P < 0,001)* ». Remarque : inutile de mentionner la valeur 0,000 000 001 ; on considère en sciences que le seuil 0,001 est suffisamment « convainquant » pour ne pas en rajouter.

L'opinion de Parsimoni & Abonessian

Parsimoni — Le test Z présente le défaut majeur de **tous** les tests : il met l'accent sur le *fait* qu'il y ait une différence (il y en a *toujours* une, aussi petite soit-elle, car tout à un effet sur tout), au lieu de parler du véritable sujet : quelle est *la taille* vraisemblable (magnitude) de cette différence. Seul le calcul de la magnitude de l'effet observé répondra à cette question. Le $P < 0{,}001$ indiqué ci-dessus ne dit rien de l'importance pratique de l'écart décelé.

Abonessian — Certes, mais le test Z a également les qualités de **tous** les tests : il indique si vous *arrivez à déceler* la différence en question, dans *quel sens* elle se trouve, et vous donne avec la probabilité P une mesure quantitative (approximative) de la *fiabilité* de votre conclusion. Il est donc tout à fait *complémentaire* du calcul des intervalles de confiance et de la magnitude de l'effet observé.

Exemple 9.4 De combien papa est-il plus grand que maman ?
Me basant sur les déclarations de n_A =207 garçons et n_B = 203 filles de la maîtrise BPE ayant répondu à un questionnaire sur (entre autres) la taille de leurs parents, j'ai calculé la différence de taille entre leur père et leur mère. D'après les 207 garçons, leur père est en moyenne plus grand que leur mère de 12,0 cm (s^2 = 45). Cependant, d'après les 203 filles, cet écart est en fait de 14,7cm (s^2 = 45,6). Y a t-il un effet significatif du sexe sur la perception de la différence de taille entre les parents ?

Le test donne Z = 2,7 (même formule que dans l'exemple **9.3**)

D'après la table de la loi normale, on a largement $P < 0{,}01$. Dans un article, on écrirait : « *Les filles de la maîtrise BPE perçoivent un écart de taille entre leur père et leur mère significativement plus important que les garçons (Z = 2,7 ; P < 0,01)* ». Pourquoi cet écart, je l'ignore. Un examen des données suggère que les fils semblent voir leur père plus petit que ne le voient les filles, alors que les filles voient leur mère plus petite que ne la voient les fils (je vois d'ici les interprétations freudiennes que l'on pourrait faire de ces résultats...). Ces écarts ne sont pas significatifs (de justesse) lorsqu'ils sont considérés séparément, mais la différence globale père – mère l'est largement. S'il y a des psychologues dans la salle, leur avis m'intéresse.

9.2.2 Au moins un de vos deux échantillons est trop petit ($n < 30$), mais la loi suivie par X est proche de la loi normale.

Dans ce cas, aucun problème pour conclure que les moyennes m_A et m_B suivent une loi approximativement normale (c'est même garanti) et le seul souci concerne la mauvaise approximation faite lorsqu'on remplace les σ^2 inconnus par leurs estimations s^2_A et s^2_B. Pour améliorer la situation, et sachant qu'on part de l'hypothèse Ho que les deux échantillons proviennent *de la même loi*, on va estimer la variance du caractère dans la population sur l'ensemble des deux échantillons en un seul bloc. Cette estimation est :

$$s^2 = \frac{s_A^2 \times (n_A - 1) + s_B^2 \times (n_B - 1)}{n_A + n_B - 2}$$

Il s'agit d'une simple moyenne pondérée, chaque estimation de variance (s^2_A ou s^2_B) étant pondérée par le nombre de ddl ($n_i - 1$) de son échantillon. C'est cette estimation *pondérée* de s^2 qu'on utilise alors dans la formule centrée réduite équivalente à celle du Z des grands échantillons :

$$t = \frac{m_A - m_B}{\sqrt{\frac{s^2}{n_A} + \frac{s^2}{n_B}}}$$

(notez bien l'usage aux numérateurs de
s^2 l'estimation pondérée, et non pas s^2_A et s^2_B)

Tout est bien qui finit bien, car il se trouve que la loi suivie par la différence centrée réduite suivra alors quand même une loi connue : la loi du t de Student. La seule chose à retenir est alors que le nombre de degrés de liberté à utiliser pour la lecture dans la table est la somme des d.d.l. de chaque échantillon, soit :

$(n_A - 1) + (n_B - 1) = n_A + n_B - 2.$ Donc, $t \to t_{(n_A+n_B-2)ddl}$

Comme dans le cas vu précédemment, on compare la valeur du t calculée à partir des données avec la valeur seuil de la table pour le α choisi et $n_A + n_B - 2$ degrés de liberté.

Exemple 9.5 On reprend l'exemple 9.3 des longueurs de main, mais avec beaucoup moins de données : $n_A = 8$ garçons et $n_B = 7$ filles de la maîtrise BPE 1998 ont mesuré leur main droite avec un double décimètre. les résultats sont les suivants : $m_A = 18{,}98$ cm $m_B = 17{,}59$ cm $s^2_A = 1{,}26$ $s^2_B = 0{,}69$ (j'ai gardé les mêmes valeurs volontairement, pour que seule la taille des échantillons ait varié par rapport à l'exemple 5.4). Les tailles moyennes sont-elles significativement différentes ?

La variance pondérée est : $s^2 = \frac{1{,}26 \times (8-1) + 0{,}69 \times (7-1)}{8+7-2} = 0{,}99$ $\quad t = \frac{18{,}98 - 17{,}59}{\sqrt{\frac{0{,}99}{8} + \frac{0{,}99}{7}}} = 2{,}70 > 2{,}16$

La valeur obtenue dépasse la valeur seuil de la table du t de Student pour $\alpha = 0{,}05$ et 13 ddl (qui est de 2,16). On rejette donc encore — mais sans tambour ni trompette cette fois — l'hypothèse que les garçons et les filles aient la même longueur de main en moyenne. Notez que la différence est ici suffisante pour pouvoir être décelée même avec deux échantillons de taille très faible. Qui a dit qu'il fallait de grands nombres pour faire des stats ? On écrira dans un rapport « *Sur la base de nos résultats, les garçons de la maîtrise BPE 1998 ont des mains significativement plus longues que celles des filles (t = 2,70 ; 13 d.d.l., P < 0,05)* ».

L'opinion de Parsimoni & Abonessian

Parsimoni — Le test t présente exactement les mêmes défauts que le test Z

Abonessian — et exactement les mêmes qualités.

9.2.3 Au moins un de vos deux échantillons est trop petit ($n < 30$), mais la loi suivie par X est inconnue, ou connue pour être éloignée de la loi normale.

Dans ce cas, l'approximation normale n'est pas possible (air connu) et le remplacement de σ^2_A (ou σ^2_B) par son estimation s^2_A (ou s^2_B) n'est pas satisfaisant non plus. Il devient nécessaire d'utiliser un test qui ne nécessite pas que X suive une loi normale, c'est à dire un test non paramétrique comme le test U de Mann et Whitney ou le test W de Wilcoxon (qui sont équivalents). Voir : **chapitre 10. Tests non paramétriques**.

9.3 Comment comparer plus de deux moyennes ?

Dans ce genre de situation, la tentation est grande de se lancer dans une série de comparaisons 2 à 2 utilisant les tests décrits plus haut, puis de classer les échantillons les uns par rapport aux autres. Cette méthode souffre de deux inconvénients, surtout s'il y a beaucoup d'échantillons (outre qu'elle est fastidieuse).

D'abord, multiplier les tests signifie que vous allez augmenter artificiellement la probabilité de tomber sur un cas ou vous allez rejeter Ho à tort. Selon la définition du risque de première espèce α, si Ho « *tous les échantillons proviennent d'une population de même moyenne* » est vraie, cette hypothèse sera *pour un test donné*, rejetée à tort avec une probabilité de α (avec en général $\alpha = 0,05$). Multiplier les tests augmente donc artificiellement les chances de « découvrir » une (fausse) différence significative.

Un moyen de contourner l'obstacle est de se montrer plus exigeant à chaque test, autrement dit, d'abaisser α en prenant $\alpha = 0,05/k$ avec k le nombre de comparaisons 2 à 2 que vous faites (cette précaution s'appelle la *correction de Bonferoni*). Vous aurez alors une probabilité globale de 0,05 de rejeter Ho à tort, même si vous effectuez plusieurs tests. Le revers de la médaille est que, en vous montrant plus exigeant, vous risquez au contraire de passer à coté de différences réelles (qui auraient peut être été décelées en prenant $\alpha = 0,05$ à chaque test). La vie est dure.

Le second désavantage de l'approche « multi-test » est qu'elle utilise des estimations de la variance de la population totale, basées sur seulement deux échantillons par test, en ignorant totalement les autres données. Cette perte d'information a pour conséquence une moins bonne estimation de la variance réelle au sein de la population. Plus précisément, la variance estimée sur un échantillon ayant tendance *à sous* estimer la variance de la population, chaque test va pouvoir être amené à signaler comme « anormalement élevés » (donc significatifs) des écarts entre moyennes qui auraient été jugés « dans le domaine de variation attendu sous Ho » par un test global ayant une meilleure estimation de la variance totale (qui est probablement plus élevée). On peut donc améliorer la méthode décrite précédemment en utilisant comme estimation de la variance une estimation portant sur tous les échantillons disponibles.

Les méthodes utilisées en réalité pour comparer les moyennes de plusieurs échantillons reposent sur un test englobant *en une seule fois* toutes les données. Le plus classique historiquement est l'ANOVA (analyse de la variance), qui suppose, pour être applicable, que les distributions sont proches de la loi normale et que les variances des populations comparées ne sont pas significativement différentes. Si vous n'avez jamais utilisé l'ANOVA, laissez tomber cette solution pour l'instant. L'autre option consiste à réaliser un test non paramétrique de comparaison multiple, appelé test H de Kruskal-Wallis. Son principe est très facile à comprendre (voir **chapitre 10. Tests non paramétriques**).

Ceci dit, si vous avez peu d'échantillons à comparer, l'approche la plus simple (même si elle a des défauts) reste cependant la comparaison « deux par deux » décrite plus haut, en vous rappelant qu'il faut, pour l'utiliser au mieux, prendre deux précautions :

(i) utiliser un seuil α égal au plus à « **0,05/nombre de comparaisons** »,

(ii) utiliser une estimation de la variance utilisant **toutes les données disponibles**, si vous utilisez des tests t ou Z.

Exemple 9.6 On a trois échantillons d'individus de sexe mâle provenant respectivement des maîtrises de BEP (Biomathématiques Extrêmement Pures), BPE (Biologie des Populations et des Ecosystèmes) et enfin SEB (Sport Etudes Basket). Les données concernant la taille moyenne sont les suivantes : BEP ($n_1 = 15$) : $m_1 =$ 175 cm, $s^2_1 = 39{,}5$; BPE ($n_2 = 12$) $m_2 = 177$ cm, $s^2_2 = 40{,}3$; SEB ($n_3 = 13$) $m_3 = 198$ cm, $s^2_3 = 45{,}3$. Ces trois échantillons proviennent-ils de la même population ? Si non, lesquels diffèrent significativement entre eux ?

La taille étant une variable distribuée normalement, on va effectuer des tests t. Notre estimation de la variance s'appuiera sur la totalité des individus (nb : en toute rigueur, il faudrait d'abord s'assurer que les variances des échantillons ne sont pas *elles mêmes* significativement différentes, ne compliquons pas les choses…). Il suffit de calculer la variance pondérée $S^2 = (14\ s^2_1 + 11\ s^2_2 + 12\ s^2_3)/(14 + 11 + 12)$. Rappel : on pondère par les d.d.l. de chaque échantillon, soit $n - 1$ à chaque fois.

Application numérique : $S^2 = 41{,}62$

Les trois tests t à effectuer seront alors (formule habituelle du t de Student, en utilisant la variance pondérée S^2 calculée ci dessus):

BEP vs BPE : t = – 0,335 ; (15 + 12 – 2 =) 25 d.d.l., NS ($P > 0{,}9$!)
BEP vs SEB : t = – 9,4 ; (15 + 13 – 2 =) 26 d.d.l. ; $P < 0{,}001$
BPE vs SEB : t = – 8,13 ; (12 + 13 – 2 =) 23 d.d.l. ; $P < 0{,}001$

Rappel : les signes des valeurs de t **n'ont aucune importance**, seule la valeur absolue compte. On obtiendrait par exemple t = + 0,335 en faisant « BPE vs BEP » au lieu de « BEP vs BPE **».**

La valeur seuil à dépasser ($|t| = 2{,}5$ environ, par interpolation) est la valeur de la table du t de Student associée au risque de première espèce à utiliser ici : $\alpha = 0{,}05/3$ comparaisons soit $\alpha = 0{,}016$ au maximum. Cette valeur de t = 2,5 est largement dépassée (nb : en valeur absolue) dans les deux cas impliquant les joueurs de basket, mais n'est même pas approchée dans la première comparaison. On conclurait dans un rapport « *Les étudiants de la maîtrise Basket sont significativement plus grands que les étudiants de Biomathématiques (t = – 9,4 ; 26 d.d.l., P < 0,001) et sont également significativement plus grands que les étudiants de BPE (t = –- 8,13 ; 23 d.d.l. ; P < 0,001). En revanche, il n'apparaît, sur la base de nos données, aucune différence significative de taille entre les étudiants de BEP et de BPE (t = – 0,335 ; 25 d.d.l. ; NS).* »

L'opinion de Parsimoni & Abonessian

Parsimoni — encore une fois, la supériorité des intervalles de confiance est éclatante. Que vous ayez deux moyennes ou cinquante, l'intervalle de confiance ne change pas, son calcul est toujours aussi simple et naturel, la lecture de la figure obtenue est toujours aussi intuitive. Quant à la correction de Bonferoni dans l'approche multi-tests (l'utilisation de α/nb comparaisons) elle est à la fois une nécessité, une absurdité, et un boulet. S'il y a beaucoup de moyennes à comparer, α devient si faible que vous n'avez plus aucune chance d'avoir quoi que ce soit de "significatif", alors que dans la réalité chaque moyenne est forcément différente des autres (même de manière infinitésimale). Encore une fois, la véritable question n'est pas tant "*qui est différent de qui*" mais "*quelle est la taille de la différence vraisemblable entre A et B*".

Abonessian — La comparaison simultanée de nombreuses moyenne est un vaste problème qui n'a jamais été complètement résolu. On est dans une situation que les anglais appellent "catch 22" (pile je gagne, face tu perds), puisque, à chaque comparaison, le risque α intervient, ce qui nous pousse à faire le moins de comparaisons possibles, alors que si on fait peu de comparaisons, on peut passer à côté des "bonnes". En réalité, ni l'ANOVA, ni le H de Kruskall-Wallis ne résolvent ce problème. Leur grand mérite est

cependant de signaler, en un seul test, qu'il y a des différences "quelque part" au sein d'un ensemble de moyennes.

Parsimoni — A ceci près que les conditions d'application de l'ANOVA sont drastiques en théorie : normalité des distributions, égalité des variances !

Abonessian — Mais assez souples en pratique. Giuseppe, vous savez bien que cette histoire d'égalité des variances est presque un conte pour enfant. C'est même vous qui me l'avez appris.

Parsimoni — Exact, et c'est bien la preuve que les manuels sont trop dogmatiques.

Abonessian — Et quand des données ne sont pas normales, certaines transformations peuvent les normaliser : la transformation des pourcentages *p* en utilisant arcsinus racine de *p* est un classique.

Parsimoni — C'est un classique tellement classique que tout le monde l'utilise de manière cabbalistique, sans se donner le moins du monde la peine de vérifier si les données ont bien été normalisées par la transformation.

Abonessian — Là encore, Giuseppe, vous savez mieux que personne que les tests de normalité sont presque inutiles : ils n'ont aucune puissance là où on en a besoin (petits effectifs) et sont très puissants lorsqu'une normalité parfaite est totalement superflue (grands effectifs).

Parsimoni — Je suis heureux de constater que tu as de bonnes lectures.

Abonessian — L'approche actuelle est en fait de coller à la distribution observée, et donc d'utiliser les *modèles linéaires généralisés*. Mais ils ne règlent pas le problème de la comparaison multi-moyennes.

Parsimoni — Je ne te le fais pas dire.

Résumé du chapitre 9.

Pour **comparer une moyenne à une valeur théorique**, on peut utiliser un test Z si l'échantillon est grand ($n > 30$). Si l'échantillon est petit, mais que la variable étudiée suit une loi proche de la loi normale, on peut utiliser un test t de Sudent. Ces deux tests sont en fait équivalents à calculer un intervalle de confiance autour de la valeur observée et examiner si la valeur théorique se trouve dans cet intervalle. Si l'échantillon est petit et que la loi est inconnue (ou connue pour être éloignée de la loi normale), on peut se reposer là encore sur le calcul d'un intervalle de confiance, mais en utilisant une approche non paramétrique (le bootstrap). Pour **comparer deux moyennes entre elles**, on retrouve les mêmes tests (Z et t de Student) et les méthodes non paramétriques (U de Mann et Whitney, W de Wilcoxon décrits dans le chapitre 10). Enfin, pour **comparer plus de deux moyennes simultanément**, on peut utiliser l'ANOVA si les distributions sont normales et si les variances ne diffèrent pas significativement. Dans le cas contraire, on peut utiliser un test non paramétrique de comparaison multiple, le test H de Kruskall-Wallis. La méthode la plus moderne repose sur une approche nommée *modèle linéaire généralisée.* Remarque : l'ANOVA et le modèle linéaire généralisé ne sont pas traités dans cet ouvrage d'introduction. La plus ou moins grande robustesse de l'ANOVA face à des situations où ses conditions d'application ne sont pas respectées est affaire de spécialiste, et ils ne sont pas forcément d'accords entre eux.

10. Les tests non paramétriques

10.1 *De naturae testii non parametricii.*

Un test *paramétrique* est un test pour lequel on suppose que le caractère étudié (ou sa moyenne) *suit une loi dont la distribution est connue*, et dont on estime les *paramètres* (moyenne, variance) au moyen des données de l'échantillon. Le test Z et le test t de Student vus dans les chapitres précédemment, sont tous deux des tests paramétriques. En effet :

(i) le test Z s'appuie sur le Théorème de la limite centrale selon lequel, pour un grand échantillon ($n > 30$), *la moyenne du caractère suit approximativement une loi normale*, quelle que soit la loi suivie par le caractère étudié lui même.
(ii) Le test t de Student, utilisé dans le cas des petits échantillons (pour lesquels on ne peut pas invoquer le TCL), nécessite que *le caractère étudié lui même suive une loi normale* (on doit au minimum avoir des arguments pour supposer que la loi suivie est proche de la loi normale). Quand on compare deux petits échantillons observés, il apparaît d'ailleurs une condition supplémentaire (dont le test Z peut se passer) : que les variances des deux échantillons ne soient pas significativement différentes.

Ce sont ces hypothèses qui permettent de faire le test : si on ne connaît pas la loi de distribution de la moyenne, *comment* calculer la probabilité que les écarts observés soient dus au hasard ? Tout simplement en s'affranchissant complètement du besoin de connaître la loi de distribution de la moyenne. C'est le principe des tests *non paramétriques*, qui sont, pour cette raison, qualifiés de *distribution-free* en Anglais (quoique *nonparametric tests* soit le terme habituel). Comme rien n'est gratuit en ce bas monde, l'abandon de toute connaissance sur la loi de distribution de la moyenne, s'accompagne d'une perte de **puissance** du test (voir le chapitre 7), assimilable à de la myopie (c'est-à-dire qu'il faudra que l'écart entre deux moyennes soit plus grand pour pouvoir déceler qu'il est significatif). Heureusement, cette perte de puissance est modérée. Selon la théorie, dans les conditions ou le test paramétrique resterait applicable, un test non paramétrique conserve *au moins* 80% de la puissance du test paramétrique (Schwartz 1993)[1]. Dans le cas où un test non paramétrique est *seul* applicable, la question ne se pose d'ailleurs plus, par manque de compétiteur !

10.2 Comparaison de deux moyennes : le test U de Mann et Whitney (et Wilcoxon).

Il existe *deux* tests non-paramétriques, utilisables indifféremment, que l'échantillon soit grand ou petit (ils remplacent donc chacun *à la fois* le test Z et le t de Student). L'un s'appelle le *test U de Mann et Whitney* et l'autre le *test W de Wilcoxon*. Cette dualité n'est qu'apparente et ces deux tests sont en fait rigoureusement équivalents, puisqu'il suffit de connaître le résultat chiffré de l'un, pour déduire automatiquement le résultat donné par l'autre, par une simple formule de conversion. Une tendance récente est d'ailleurs de mettre en avant leur gémellité en parlant du test de Mann-Whitney-Wilcoxon.

[1] Schwartz, D. 1993. *Méthodes statistiques à l'usage des médecins et des biologistes* (4ème édition). Medecines Sciences, Flammarion. 314 pages.

En conséquence, je présenterai une seule de ces deux approches strictement équivalentes : le **test U de Mann et Whitney**.

Voici vos deux échantillons A et B, d'effectifs n_A et n_B :

$x_1, x_2 \ldots\ldots\ldots\ldots x_{nA}$
$y_1, y_2 \ldots\ldots x_{nB}$

Pour traiter le cas général où la taille des deux échantillons est différente, je vais supposer que $n_A > n_B$.

Le test est bâti sur le principe que, si les individus proviennent en fait de la même population (hypothèse Ho), alors la probabilité qu'un x pris au hasard soit supérieur à un y pris au hasard est de 0,5 (une chance sur deux). Si en revanche la moyenne est plus élevée dans A que dans B, on aura plus souvent x > y que l'inverse (et, si on réalise un classement on trouvera les x préférentiellement dans le haut du classement). Si la moyenne est supérieure dans B, on aura l'inverse (quelle révélation bouleversante), et ce sont les y qui occuperont plutôt le haut du classement.

Donc, si on examine méthodiquement *toutes* les comparaisons possibles entre les x et les y (il y a $n_A \times n_B$ comparaisons différentes possibles), la *proportion* des comparaisons pour lesquelles on aura x > y doit être (en espérance) de 0,5 (soit 50%). Le *nombre* de cas où on va avoir x > y sera donc, toujours en espérance :

$$U_0 = (n_A \times n_B) / 2$$

U_0 est la valeur attendue de U, si on répétait l'expérience une infinité de fois avec des effectifs n_A et n_B, dans deux populations de même moyenne pour le caractère étudié.

Un moyen simple d'effectuer rapidement toutes les comparaisons possibles est de réaliser un classement de ce type (du plus grand au plus petit):

1. x_1	**2.** x_2 et x_3 (ex æquo)	**3.** x_4	**4.** x_5 et y_1 (ex æquo)	**5.** y_2	**6.** y_3	**7.** x_6

Il peut sembler ridicule de faire un test statistique sur des échantillons de taille aussi réduite qu'ici ($n_A = 6$ et $n_B = 3$). Ma réponse sera double :(i) la limite basse d'utilisation de ce test est *encore plus faible* : on peut déceler un écart significatif ($\alpha = 0,05$) entre deux échantillons ayant chacun seulement 4 individus (!), il faut cependant pour cela que les 4 individus de A aient tous une valeur plus élevée que le meilleur individu de B, et (ii) le principe du test est évidemment plus facile à présenter avec peu d'individus.

Dans un premier temps, on va compter pour chaque x le nombre de y qui lui sont inférieurs et on somme les résultats obtenus pour tous les x (cela revient bel et bien à passer en revue toutes les comparaisons possibles entre les x et les y et à noter *le nombre de cas où x > y*). Le nombre obtenu est une variable U de Mann et Whitney, notée dans ce cas précis U*xy*. Dans un deuxième temps, on fait de même pour obtenir son alter ego U*yx*, c'est-à-dire qu'on va

compter, pour chaque y, le nombre de x qui lui sont inférieurs, et on somme le résultat pour tous les y, obtenant ainsi la variable U*yx* (cela revient à faire une deuxième fois toutes les comparaisons possibles, mais à noter cette fois-ci *le nombre de cas où y > x)*.

Vous aurez cependant remarqué une difficulté : nous avons parmi nos données des valeurs x et y qui sont ex-aequo. Que faire ? Notre exemple illustre plus précisément *deux cas distincts, qui n'appellent pas la même réponse*. Premier cas : x_2 et x_3 sont ex-aequo et appartiennent au même échantillon. Cela ne perturbe en rien le comptage : ces deux individus comptabilisent *chacun* 3 individus y qui leurs sont inférieurs (y_1, y_2, y_3). Deuxième cas : x_5 et y_1 sont ex-aequo et sont « adversaires ». Dans le comptage concernant x_5, l'individu y_1 va compter seulement pour 0,5 (une demi-part en quelque sorte). Le score obtenu par x_5 sera donc : 0,5 (à cause de y_1) + 2 (y_2 et y_3 sont strictement inférieurs à x_5) = 2,5. Côté y_1, même raisonnement et son score sera donc 0,5 (à cause de x_5) + 1 (pour x_6, strictement inférieur à y_1) = 1,5. NB : si un individu a *plusieurs* « adversaires » avec lesquels il est ex aequo, le principe restera le même : il engrange 0,5 point par « adversaire ex-aequo» et fournit 0,5 point à chacun de ces « adversaires ex aequo». Compliqué ? Voyez le tableau de résultats :

	A	B	
3pts	x_1		
3pts et 3pts	x_2, x_3		
3pts	x_4		
2,5pts	x_5	y_1	1,5pts
		y_2	1pt
		y_3	1pt
0	x_6		
U*xy* = **14,5**			U*yx* = **3,5**

Vous remarquerez qu'on a bien comptabilisé *toutes* les comparaisons possibles : U*xy* + U*yx* = 14,5 + 3,5 = 18, et on a bien $n_A \times n_B = 6 \times 3 = 18$ (vous n'oublierez pas de faire cette vérification en pratique, l'expérience montre qu'il est très facile d'oublier un demi-point lors du comptage). Le bout du tunnel se profile, car la distribution des variables U est connue et tabulée (c'est la moindre des choses). Il ne reste plus qu'à aller lire dans la table du U de Mann et Whitney, exercice quelque peu… déroutant au début.

En effet, pour une raison historique, on utilise pour le test la plus *petite* des valeurs calculée entre U*xy* et U*yx* (ici, U*yx* = **3,5)** tout simplement parce que son calcul était plus rapide du temps où l'addition des scores s'effectuait avec un papier et un crayon. La table traditionnelle du U de Mann et Whitney est donc habituellement présentée « à l'envers » par rapport à la logique des autres tables (loi normale, t de student, chi 2, etc…). Le test est en effet significatif si *le plus petit* (entre U*xy* et U*yx*) est **inférieur** (et non pas **supérieur)** à la valeur de la table (le test n'est donc **pas** significatif quand on **dépasse** la valeur critique, ce qui est la conclusion inverse de tous les autres tests que nous avons abordés). Dans notre exemple, la plus petite valeur calculée est donc U*yx* = 3,5 et la valeur critique U de la table pour $n_A - n_B = 6 - 3 = 3$ est : $U_{table} = 1$. Ici, U*yx* n'est pas inférieur à 1, donc on ne peut pas rejeter Ho « *les populations dont ces échantillons sont tirés, ont la même moyenne* » et on conclut à une différence *non* significative.

En conclusion, en notant *Min* (Uxy, Uyx) la plus petite des deux valeurs U calculées, le principe général de la lecture dans la table du U est le suivant :

Si *Min* (Uxy, Uyx) < U_{table}, on rejette Ho, au risque α choisi.
Si *Min* (Uxy, Uyx) > U_{table}, on ne rejette pas Ho, au risque β inconnu, mais d'autant plus grand que les moyennes réelles sont proches, si elles sont différentes.

Remarque : le fait de classer vos valeurs dans l'ordre décroissant et de compter les « inférieurs à » ou dans l'ordre croissant et de compter les « supérieur à » est évidemment strictement équivalent. Choisissez donc l'ordre qu'il vous plaira !

Exemple 10.1.
On doit comparer deux séries de longévités d'animaux rares, nés en captivité dans deux types de zoos qui utilisent des méthodes d'élevage différentes. La question est de savoir si un type de méthode d'élevage est meilleur que l'autre, en comparant 10 zoos utilisant la première méthode et 8 zoos utilisant la deuxième méthode. Ce type de naissance étant très rare, on n'a pu observer qu'un seul cas par zoo. La distribution des longévités est une variable notoirement distincte de la loi normale (en particulier à cause de la mortalité en bas âge, qui peut être particulièrement élevée). Les longévités (en années) obtenues dans les deux types de zoos sont les suivantes:

Zoo de type A (n_A = 10 naissances, dans 10 zoos) : 1, 1, 1, 1, 5, 6, 6, 8, 9, 9
Zoo de type B (n_B = 8 naissances, dans 8 zoos) : 1, 7, 7, 8, 10, 11, 13, 15

Le calcul donne U*min* = 16,5 (voir détail du calcul dans le tableau ci-dessous).

La valeur seuil pour α = 0,05 et une taille d'échantillon A = 8 avec une différence $n_A - n_B = 2$ est de **17**, et elle est de **11** pour α = 0,01. Du fait que 16,5 est **inférieur** à 17, le test est significatif au risque α = 0,05. On peut donc rejeter au risque α = 0,05 l'hypothèse Ho selon laquelle la longévité de cette espèce d'animal est la même dans les deux types de zoos comparés. On conclurait éventuellement dans un rapport :" la longévité moyenne de l'espèce X est significativement plus faible dans les zoos de type A, (U de Mann et Whitney, $\boldsymbol{n_A = 10}$, n_B = 8, U = 16,5; $P < 0,05$). En réalité, dans ce genre d'études, il est particulièrement délicat de conclure sur la cause réelle de la différence observée, tant il est impossible de standardiser des "objets" tels que des zoos, qui ont forcément des localisations géographiques, des équipes soignantes, des directeurs différents etc... Par ailleurs, rien ne dit que les mères soient de la même origine pour chaque type de zoo. Bref, cet exemple totalement artificiel visait simplement à montrer un cas pour lequel il est difficile d'avoir un grand échantillon (naissances rares) et dans lequel la variable étudiée (longévité), est largement éloignée de la loi normale (test t de Student peu pertinent).

score	A	B	score
		15	*10*
		13	*10*
		11	*10*
		10	*10*
4 -4	**9, 9**		
3,5	**8**	**8**	*7,5*
		7	*7*
		7	*7*
1	**6**		
1	**6**		
1	**5**		
0,5	**1**	**1**	*4 × 0,5 = 2*
0,5	**1**		
0,5	**1**		
0,5	**1**		
16,5			***63,5***

Bonne nouvelle, pour des effectifs n_A et n_B supérieurs chacun à la dizaine, la variable U suit une loi approximativement... normale (on n'y échappe décidément pas !). Comme d'habitude, on peut alors centrer-réduire la variable en question pour obtenir la loi normale centrée-réduite N (0 : 1). La variance de la loi suivie par U vaut :

$$\sigma^2_U = n_A \times n_B \times (n_A + n_B) / 12$$

(NB : ne cherchez aucun rapport entre le "**12**" du dénominateur et l'effectif de A et de B, cette valeur est fixe).

Ainsi, quand n_A et n_B sont trop grands pour utiliser la table du U de Mann et Whitney (la table fournie en annexe déclare forfait au dessus de $n = 20$), on retombe sur… le test paramétrique Z, après l'opération habituelle de centrage-réduction :

$$Z = (\mathbf{U} - U_{o)} / \sqrt{(\sigma^2_U)} \rightarrow N\ (0 : 1)$$

avec **U** = Umin ou Umax, peu importe, seule compte la valeur absolue de Z

Si |Z| > 1,96 on rejette Ho, au risque $\alpha = 0{,}05$.
Si |Z| < 1,96 on ne rejette pas Ho, au risque β inconnu, mais d'autant plus grand que les moyennes, si elles sont différentes, sont proches.

Exemple 10.2.
Même exemple des zoos, mais on a réussi à obtenir les données en provenance de quelques zoos supplémentaires. Les longévités (en années) obtenues dans les deux types de zoos deviennent les suivantes

Zoo de type A (n_A = 22 naissances, dans 22 zoos) : 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 5, 5, 6, 6, 7, 8, 9, 12, 12
Zoo de type B (n_B = 12 naissances, dans 12 zoos) : 1, 7, 7, 8, 9, 10, 11, 12, 13, 14, 15, 15

Uo = (22 × 12) / 2 = 132
$\sigma^2_U = 22 \times 12 \times (22 + 12) / 12 = 748$
$U_{min} = 38{,}5$ (et $U_{max} = 225{,}5$). Remarquez que $|U_{min} - Uo| = |U_{max} - Uo| = 93{,}5$ — il est donc indifférent d'utiliser U_{max} ou U_{min} dans le calcul de la variable Z. Juste pour éviter un signe négatif, j'utiliserai U_{max}

$$Z = (225{,}5 - 132)/\sqrt{748} = \mathbf{3{,}41}$$

D'après la table de la loi normale, on est au delà de la valeur seuil pour $\alpha = 0{,}001$ (qui est de 3,29). On conclurait éventuellement dans un rapport : "La longévité moyenne de l'espèce X est significativement plus faible dans les zoos de type A ($Z = 3{,}41$; $P < 0{,}001$)". Tout ce qui a été dit sur les problèmes méthodologiques de ce type d'étude (comment garantir que les deux groupes de zoos diffèrent *uniquement* par le facteur "type d'élevage de l'espèce X" ?) reste valable, et la plus grande prudence reste à l'ordre du jour.

10.3 Comparaison de plus de deux moyennes par le test H de Kruskall et Wallis.

On a vu plus haut que, multiplier les tests sur de petits échantillons, au lieu de faire un test global, augmentait le risque α et obligeait à utiliser de mauvaises estimations de la variance globale. La solution, s'il y a beaucoup d'échantillons (ce qui obligerait à multiplier les tests) repose donc sur un test global, mais *non paramétrique*. Notez au passage qu'on peut parfaitement utiliser un test non paramétrique **alors qu'on serait dans les conditions de l'utilisation d'une ANOVA.** C'est même fortement préférable à se lancer dans une ANOVA sans comprendre ce que l'on fait ! ! ! Un équivalent *non paramétrique* de l'ANOVA est le *test H de Kruskal et Wallis*.

Test H de Kruskall et Wallis.

On a k séries de valeurs (et non plus deux comme dans un U de Mann-Whitney). Si on appelle A, B, C… les échantillons, d'effectifs n_A, n_B, n_C… on aura donc pour A les valeurs a_1, a_2, a_3… pour B les valeurs b_1, b_2, b_3…etc.

Comme pour le test U de Mann et Whitney, on va d'abord classer toutes ces valeurs par ordre croissant sans tenir compte de leur échantillon d'origine. Cependant, on va ensuite attribuer à chaque individu un **rang**, et c'est ce rang qui sera utilisé dans les calculs. Une fois le classement effectué, on s'occupe des ex aequo. S'il y a des ex aequo occupant par exemple les rangs 10, 11 et 12, on leur donnera à tous le rang moyen 11 (s'ils occupent les rangs 10 et 11 on leur donnera à chacun le rang 10,5). Notez que ceci est valable *que les ex aequo appartiennent à des échantillons différents ou pas*. On calcule ensuite le rang moyen théorique R, qui vaut (N + 1)/2 avec N le nombre total d'individus, puis on calcule les rangs moyens r_A, r_B, r_C à l'intérieur de chaque classe. Concrètement, le rang moyen r_A de l'échantillon A sera la moyenne des rangs occupés dans le classement *général* par les individus a_1, a_2 a_3 etc…

Si l'hypothèse Ho « *tous les échantillons proviennent de la même populati*on » est vraie, alors le rang moyen de chaque échantillon devrait être très proche du rang moyen théorique R. Si elle est fausse, au moins un échantillon va s'écarter de la norme et le test doit le détecter. On calcule alors l'indice H :

$$\mathbf{H = [\ n_A\ (r_A - R)^2 + n_B\ (r_B - R)^2 + n_C\ (r_C - R)^2\ etc\ldots]\ /\ [\ N(N + 1)/12\ ]}$$

Où l'on retrouve au dénominateur le mystérieux « 12 » du U de Mann et Whitney et du test de Wilcoxon, le test de Wilcoxon étant d'ailleurs un cas particulier du H de Kruskall-Wallis avec deux classes seulement.

Nous sommes au bout de nos peines, car il se trouve que H suit approximativement une loi du χ^2 avec (k – 1) degrés de liberté (rappel : k est le nombre d'échantillons à comparer). La lecture du test se fait donc exactement comme s'il s'agissait d'un χ^2.

Si le test H est non significatif, il est inutile (et absurde) de faire *ensuite* des comparaisons deux par deux. En revanche si le test H est significatif, on est tenté de savoir ce qui en est la cause. On retombe alors dans le problème des tests multiples. Si vous *voulez* faire ces comparaisons deux à deux, utilisez alors un seuil α égal au maximum à 0,05/nombre de tests.

Résumé du chapitre 10

Face à de petits échantillons et une variable aléatoire de loi inconnue (ou connue pour être éloignée de la loi normale), on peut quand même comparer deux moyennes observées en utilisant le **test U de Mann et Whitney** (=**test W de Wilcoxon**), équivalent non paramétrique du test Z et du test t de Student. Dans le cas de plusieurs moyennes à comparer simultanément, on peut utiliser le **test H de Kruskall et Wallis**, équivalent non paramétrique de l'ANOVA. L'utilisation d'un test non paramétrique s'accompagnant d'une légère perte de puissance, ces tests ne sont utilisés que dans les cas où on ne peut utiliser les tests paramétriques, plus puissants.

11. Comparaisons de pourcentages.

11.1 Comparaison entre un pourcentage observé et un pourcentage théorique : le (test du) χ^2 de conformité.

Ce sujet n'est qu'un cas particulier d'un problème plus général, qui est la comparaison d'une répartition observée de *n* objets répartis en k classes, avec une répartition théorique. Notez qu'il est nécessaire que les classes soient mutuellement exclusives (un individu ne peut pas appartenir à deux classes à la fois). Par exemple, le pourcentage de gauchers dans un groupe de *n* personnes est basé sur la répartition des *n* observations en deux catégories (k = 2), à savoir obs_1 gauchers et de obs_2 droitiers, sachant qu'on ne peut pas être gaucher et droitier à la fois[1]. Du fait que la résolution du problème est aussi simple avec un nombre de catégories k quelconque (c'est-à-dire supérieur ou égal à 2), c'est ce cas général qu'on va examiner. L'application au cas d'un seul pourcentage sera ensuite immédiate.

La répartition *observée* de nos *n* individus au sein des k classes aura la forme suivante :

$$obs_1, obs_2, obs_3 \ldots obs_k$$

obs_i l'effectif observé de la classe i

$obs_1 + obs_2 + \ldots + obs_k = n$

Notre but est de comparer cette répartition observée avec une répartition théorique (qui correspondra à notre hypothèse Ho). Il y a plusieurs façons d'obtenir cette répartition théorique, et on y reviendra, car cela a de l'importance dans l'interprétation du test. Pour l'instant, on va négliger les détails et considérer simplement qu'on *connaît* cette répartition a priori (par exemple, on sait que la célébrissime pièce-de-monnaie-équilibrée doit tomber sur pile une fois sur deux en moyenne). La répartition *théorique* de nos *n* individus en k classes (= la répartition si Ho est vraie) sera de la forme :

$$théo_1, théo_2, théo_3 \ldots théo_k,$$

$théo_i$ l'effectif observé de la classe i

$théo_1 + théo_2 + \ldots + théo_k = n$

Il serait évidemment absurde de rejeter Ho chaque fois qu'il n'y a pas accord *parfait* entre nos observations et la théorie (qui considérerait la pièce de monnaie comme faussée si elle tombait deux fois sur pile en deux lancers ?). Il faut tenir compte du fait que nos observations, *basées sur un échantillon*, sont soumises aux incontournables fluctuations d'échantillonnage. Il est donc parfaitement normal d'observer des différences avec la théorie. Notre problème, comme dans la comparaison de deux moyennes, consiste à mesurer l'écart entre l'observation et la répartition attendue sous Ho, puis à déterminer si cet écart est trop grand pour être expliqué par la seule erreur d'échantillonnage. On va donc, encore une fois, se construire une variable de test, dont il faudra connaître la distribution sous Ho. On pourra alors calculer la probabilité qu'on aurait d'observer l'écart constaté avec nos données, si Ho était vraie. Si cette probabilité

[1] en fait si, on peut. Dans un cas pareil on devrait créer une *troisième* catégorie pour les ambidextres ou bien séparer les individus en "purement droitier" et "non purement droitier" si on voulait seulement deux catégories mutuellement exclusives.

est trop faible, on rejettera Ho. Le premier mouvement spontané pour mesurer *globalement* un écart entre deux distributions est simplement de mesurer les écarts entre effectifs observés et théoriques classe par classe, puis de faire la somme de ces écarts. Autrement dit de calculer :

$$(obs_1 - théo_1) + (obs_2 - théo_2) + \ldots + (obs_k - théo_k)$$

Ce calcul donne systématiquement un résultat égal à… zéro. A la réflexion, on comprend que, l'effectif total étant fixé, les individus se trouvant en excès dans une classe feront forcément défaut dans une autre, d'où la somme nulle des écarts. La solution semble s'imposer d'elle même : utiliser les valeurs absolues. Acceptons cette solution pour l'instant. Notre mesure de l'écart global entre la répartition observée et la répartition théorique devient :

$$|obs_1 - théo_1| + |obs_2 - théo_2| + \ldots + |obs_k - théo_k|$$

Cette mesure prendra indubitablement des valeurs d'autant plus grandes que les répartitions diffèrent, ce qui va dans le bon sens. Elle a cependant un défaut gênant, qui est de traiter les écarts sur le même pied d'égalité, sans considération pour la taille relative des classes. Or, il est évident que, par exemple, un excès observé de 10 individus dans une classe qui en comporte théoriquement 1000 (soit + 1%), méritera beaucoup moins d'attention que le même écart pour une classe qui en comporte théoriquement 5 (soit + 200%). Il faut donc pouvoir *relativiser* les écarts observés, et le moyen le plus simple est de les diviser par l'effectif théorique de leur classe. On obtient une mesure de l'écart global plus pertinente, qui est :

$$\frac{|obs_1 - théo_1|}{théo_1} + \frac{|obs_2 - théo_2|}{théo_2} + \ldots + \frac{|obs_k - théo_k|}{théo_k}$$

Il ne reste plus qu'à déterminer si cet écart est « trop grand » pour pouvoir être expliqué par l'erreur d'échantillonnage. Il faut pour cela connaître la loi de distribution de notre mesure si Ho est vraie. Si l'écart global que nous venons de calculer a une probabilité moindre que 0,05 d'être aussi important par hasard, nous rejetterons Ho, en concluant que la distribution théorique n'est pas respectée. Dans le cas contraire, nous conclurons que Ho ne peut pas être rejetée avec nos données.

Hélas, *on ne connaît pas* la loi de probabilité de la mesure présentée ici. De quoi ? Tout ça pour rien ? ! Non, car de brillants esprits se sont attelés au problème et ont évidemment trouvé la solution. Il suffit de considérer non pas *la valeur absolue* des écarts mais *leur carré*. Et pourquoi donc ? Constatons d'abord que cette modification n'altère pas le comportement général de notre indice de distance : l'indice utilisant les carrés des écarts sera toujours d'autant plus grand que l'écart entre la répartition observée et la répartition théorique est grand. Réjouissons-nous ensuite de ne plus avoir à traîner comme un boulet, toutes ces valeurs absolues, diaboliques sources d'erreurs, comme vous l'avez appris douloureusement en classe de 3ème. Enfin, et c'est évidemment l'intérêt de la manœuvre, *on connaît la loi de distribution sous Ho de ce nouvel indice utilisant les carrés des écarts.* Des considérations mathématiques remplissant plusieurs pages permettent en effet de démontrer que la variable :

$$\frac{(obs_1 - théo_1)^2}{théo_1} + \frac{(obs_2 - théo_2)^2}{théo_2} + \ldots + \frac{(obs_k - théo_k)^2}{théo_k}$$

suit, avec une excellente approximation, une loi connue, qui est la loi du χ^2. Certes, la loi en question est définie *strictement parlant* comme une *somme de variables normales centrées réduites élevées au carré*, mais heureusement, sous certaines conditions souvent remplies (voir 11.3 conditions d'application du chi2), notre somme des (obs – théo)^2/théo se comporte de la même manière qu'une "*somme de variables normales etc.*". Il y a des jours où les dieux sont avec nous. Toujours est-il que la distribution du χ^2 étant disponible sous forme de tables, il suffit de lire dans la table, si la valeur calculée dépasse la valeur de la table au seuil α choisi. Si le χ^2 calculé dépasse la valeur lue dans la table, on rejette Ho au seuil α choisi et on conclut donc que la répartition observée est significativement différente de la répartition théorique. Dans le cas contraire, on ne peut pas rejeter Ho. Reste maintenant un détail à régler : sur quelle ligne de la table doit-on lire ? Ici, les choses se corsent, car il existe une infinité de distributions du χ^2 , de même qu'il existe une infinité de distributions du t de Student. Il y a en effet une distribution pour chaque nombre de degrés de liberté. Comment calculer ce nombre ?

Premier cas (le plus simple) : les valeurs théoriques sont totalement indépendantes des données observées. Ce sera le cas chaque fois qu'elles ont été calculées sur des données antérieures, ou lorsqu'elles reposent sur un modèle abstrait sans rapport avec les données observées (exemple : la pièce de monnaie qui doit donner *a priori* 50% de pile). Dans ce cas, le nombre de ddl est égal au nombre de classes – 1, c'est-à-dire à k – 1. On enlève 1 ddl comme vu précédemment, car le nombre de ddl correspond au nombre de variables aléatoires *indépendantes*. Or, pour un total donné, la connaissance de k – 1 valeurs va *fixer* immédiatement la dernière par différence au total. Les k effectifs des k classes ne représentent donc pas k, mais k – 1 variables *indépendantes*.

Exemple 11.1: Sur 80 individus répartis en 4 classes, la répartition observée est 40 ; 30 ; 6 ; 4 alors que la répartition théorique (connue a priori, donc **indépendamment** des données) donnerait avec cet effectif de 80 individus la répartition : 35 ; 25 ; 10 ; 10. Nos données s'écartent-elles significativement de la théorie ?

Le calcul du χ^2 donne : $\chi^2 = \frac{(40-35)^2}{35} + \frac{(30-25)^2}{25} + \frac{(6-10)^2}{10} + \frac{(4-10)^2}{10} = 6{,}91$

Il y a k = 4 classes, donc 4 – 1 = 3 ddl. La table du χ^2 donne pour 3 ddl, la valeur seuil de 7,815 pour le risque α = 0,05. La valeur seuil n'étant pas dépassée, on ne peut pas rejeter Ho et on écrira : « *Sur la base de nos données, on ne constate pas d'écart significatif par rapport aux proportions attendues* (χ^2 *= 6,91 ; 3 d.d.l., NS).*

Deuxième cas : les valeurs théoriques ont été au moins en partie déterminées à partir des données. Exemple classique entre tous : en génétique des populations, on vous fournit les effectifs d'individus de génotype AA, Aa et aa et on vous demande de déterminer s'ils vérifient les proportions de Hardy-Weinberg. Si c'est la seule information qu'on vous donne, vous êtes obligés d'utiliser vos données *observées* pour estimer la fréquence *p* de l'allèle « A »

(la fréquence q de l'allèle « a » se déduisant automatiquement du fait que $p + q = 1$). Ce faisant, vos valeurs théoriques pour les effectifs AA, Aa et aa, basées sur la prédiction théorique des proportions mendélienne "p^2 individus AA : $2pq$ individus Aa : q^2 individus aa" ne sont plus indépendantes des données. Elles en sont en fait artificiellement rapprochées, ce qui va diminuer la valeur du χ^2 calculée et rendre plus difficile le rejet de Ho. Pour tenir compte de ce fait, il faut diminuer le nombre de ddl de 1. Ainsi, au lieu de lire dans la table pour un nombre de 3 – 1 = 2 ddl, il faudra utiliser la valeur correspondant à 3 – 1 – 1 = 1 ddl. Le premier ddl est enlevé parce qu'il suffit de connaître le nombre de AA et de aa (par exemple) pour connaître celui des Aa par différence au total. Le second ddl est enlevé parce qu'on a estimé la fréquence du gène A en utilisant les données observées.

Exemple 11.2. Dans un échantillon de 50 individus, on observe pour un locus diallèlique les effectifs des génotypes suivants : AA = 25 ; Aa = 20 ; aa = 5. Ces effectifs sont-ils en accord avec l'hypothèse de Hardy-Weinberg ?

On estime tout d'abord, les fréquences alléliques *d'après les données* : P(A) = p = P(AA)+1/2 P(Aa) =(25 + 20/2)/50 = 0,7 d'où on déduit *sans avoir à se servir une nouvelle fois des données* que P(a) = 1 – p = 0,3. On en déduit les proportions, puis les effectifs théoriques des trois génotypes, en se basant sur la relation de Hardy-Weinberg (hypothèse Ho):

P(AA) = p^2 = $(0,7)^2$ = 0,49 — effectif théo (AA) = 0,49 × 50 = 24,5
P(Aa) = $2pq$ = 2 × 0,7 × 0,3 = 0,42 — effectif théo (Aa) = 0,42 × 50 = 21
P(aa) = q^2 = $(0,3)^2$ = 0,09 — effectif théo (aa) = 0,09 × 50 = 4,5

On compare maintenant les effectifs observés avec les effectifs théoriques par un χ^2 :
Le nombre de ddl est 3 – 1 (total) – 1(estimation de p à partir des données) = 1 ddl

$$\chi^2 = \frac{(25-24{,}5)^2}{24{,}5} + \frac{(20-21)^2}{21} + \frac{(5-4{,}5)^2}{4{,}5} = 0{,}113$$

La valeur seuil du χ^2 pour α = 0,05 et 1 d.d.l. est de 3,84. Cette valeur n'étant pas dépassée, on ne peut pas rejeter Ho. On écrira « *Sur la base de nos données, on ne constate pas d'écart significatif par rapport aux proportions de Hardy-Weinberg* (χ^2 *= 0,113 ; 1 d.d.l., NS).* ».

En conclusion sur ce type de χ^2 (appellé « chi2 de conformité » car il vérifie la conformité de données observées avec un modèle théorique), on va maintenant traiter très facilement le cas particulier d'un seul pourcentage (donc le cas où k = 2 classes):

Exemple 11.3. Sur 50 individus adultes capturés au hasard dans une population, il y a 32 mâles (soit 64%). Cette proportion de mâles est-elle trop élevée pour accepter, au risque α = 0,05, l'hypothèse Ho que le sex-ratio de la population est équilibré (ce qui supposerait 50% de mâles) ?

L'erreur à ne pas commettre ici, est de négliger les femelles et faire le test du χ^2 sur la moitié des données. Il s'agit, rappelons-le, de tester l'écart entre *deux répartitions en k classes*, et non pas simplement entre deux effectifs d'*une* classe considérée isolément de l'ensemble. Les valeurs observées sont donc ici 32 mâles **et** 18 femelles, les valeurs théoriques (totalement **indépendantes** des données) étant 25 mâles **et** 25 femelles. Le nombre de ddl est ici 2 (classes) – 1 (total) = 1 degré de liberté.

$$\chi^2 = \frac{(32-25)^2}{25} + \frac{(18-25)^2}{25} = 3{,}92$$

La valeur seuil du χ^2 pour 1 ddl et $\alpha = 0,05$ est de 3,84. Cette valeur étant dépassée (de justesse), on rejette Ho, et on écrira dans un rapport : « *Le sex-ratio de cette population est déséquilibré en faveur des mâles (χ^2 = 3,92 ; 1 d.d.l., P < 0,05).* »

L'opinion de Parsimoni & Abonessian

Parsimoni — Dans le cas d'un pourcentage unique, voici encore un bel exemple de test inutile. Pourquoi ne pas calculer directement l'intervalle de confiance du pourcentage observé ? Pour le même prix, vous avez le test (si la valeur théorique est dans cet intervalle, vous savez d'avance que le test sera NS) et la gamme de valeur vraisemblable pour le pourcentage de la population réelle.

Abonessian — Dans ce cas précis, effectivement, le test est équivalent au calcul d'un intervalle de confiance autour du pourcentage observé. Mais encore une fois, il vous donne une probabilité *P* que le simple calcul de l'intervalle de confiance ne donne pas. Il y a donc complémentarité entre les deux approches.

11.2 Comparaison entre plusieurs distributions observées : le χ^2 d'homogénéité.

Le χ^2 de conformité compare une distribution observée avec une distribution théorique, en utilisant, dans chaque classe, les (carrés des) écarts entre l'effectif observé et l'effectif théorique. Que faire maintenant si l'on doit comparer *entre elles* plusieurs distributions *observées* ? Bien que mesurer simultanément l'homogénéité de *N* distributions ne soit en fait pas plus compliqué qu'en comparer seulement deux, ce dernier exemple est plus simple à suivre. On va donc commencer par essayer de comparer *deux* distributions observées comportant k classes chacune :

$$\begin{array}{ll} & obs_1,\ obs_2\ \ldots\ obs_k \\ \text{et} & \\ & obs'_1, obs'_2\ \ldots\ obs'_k \end{array}$$

Notez qu'il faut impérativement abandonner l'idée qui vient éventuellement à l'esprit d'utiliser une des deux distribution observée comme la distribution « théorique » et d'effectuer le même calcul que pour un χ^2 de conformité :

$$\frac{(obs_1 - obs'_1)^2}{obs'_1} + \frac{(obs_2 - obs'_2)^2}{obs'_2} + \ldots + \frac{(obs_k - obs'_k)^2}{obs'_k}$$

Outre qu'un tel calcul obligerait à avoir des effectifs égaux dans chaque échantillon (une contrainte importante dont on aime à se passer), il constitue une erreur de raisonnement à la base. En effet, on a bien ici en présence *deux* répartitions *observées*, qui sont *chacune* soumise aux fluctuations d'échantillonnage. Il est donc fondamentalement erroné d'en considérer une comme fixe et parfaite.

La méthode correcte consiste à comparer les effectifs de chaque classe des deux distributions à ce que chacun d'eux *devrait être* sous « Ho : les deux populations ne diffèrent pas ». Pour éviter un débat trop abstrait, voici un exemple choisi totalement au hasard.

On doit comparer la répartition entre mâles adultes, femelles adultes et immatures « non sexables » à deux endroits différents de l'estran de la pointe de *P******, dans une espèce de mollusque gastéropode qui a également choisi de rester anonyme. Nos deux échantillons comportent 20 mâles, 20 femelles et 60 immatures pour le premier (n_A = 100) ; 5 mâles, 30 femelles et 90 immatures pour le second (n_B = 125).

> **Parenthèse utile :** comme dans le cas de *tous* les autres tests décrits dans ce document, le fait que les deux échantillons n'aient pas le même effectif *n'a absolument aucune importance et n'empêche en rien de les comparer*. Je lance donc un appel à tous les étudiants de bonne volonté : par pitié arrêtez une bonne fois pour toutes d'inventer cette « règle ». Les stats sont suffisamment compliquées comme ça, non ?

Il nous faut maintenant répondre à la question suivante : sous l'hypothèse Ho « les échantillons proviennent en fait de la même population et les différences de proportions observées sont dues simplement à l'erreur d'échantillonnage » quels devraient être les effectifs théoriques des classes de ces deux distributions ? Il nous faut pour répondre à cette question déterminer au mieux quelles sont les proportions relatives entre mâles, femelles et immatures dans la population unique de notre hypothèse Ho. Il suffira ensuite d'appliquer ces proportions aux effectifs totaux de chaque échantillon pour déterminer les effectifs de chaque classe.

Du fait même que notre hypothèse Ho stipule que les échantillons proviennent d'une même population, l'estimation des proportions dans la population se base évidemment sur les deux échantillons *réunis*, donc sur les 225 individus disponibles. On obtient les proportions suivantes :

mâles : 25/225 = 0,111
femelles : 50/225 = 0,222
immatures : (par différence au total de 1) = 0,667.

En appliquant ces proportions à chacun des deux échantillons, on obtient les effectifs *théoriques* classe par classe, soit 11,1 mâles ; 22,2 femelles ; 66,7 immatures pour l'échantillon A et 13,875 mâles ; 27,75 femelles et 83,375 immatures pour l'échantillon B. Le calcul du χ^2 est ensuite identique à celui du χ^2 de conformité, en appliquant pour chaque classe la célèbre formule (obs – théo)^2/théo:

$$\chi^2 = \frac{(20-11{,}1)^2}{11{,}1} + \frac{(20-22{,}2)^2}{22{,}2} + \frac{(60-66{,}7)^2}{66{,}7} + \frac{(5-13{,}875)^2}{13{,}875} + \frac{(30-27{,}75)^2}{27{,}75} + \frac{(90-83{,}375)^2}{83{,}375} = 14{,}41$$

Arrive maintenant le moment que vous adorez tous : il faut déterminer le nombre de degrés de libertés, pour pouvoir lire dans la table du χ^2. Je vous propose deux méthodes de raisonnement, vous verrez par la suite celle qui vous convient le mieux (ceci est une pure clause de style, tout individu normalement constitué préfère la méthode rapide, qui évite d'avoir à réfléchir).

Première méthode : le raisonnement. Il y a, dans le tableau de données observées, deux échantillons comportant chacun 3 effectifs qui sont des variables aléatoires, mais il n'y a pas 6 ddl pour autant dans notre χ^2 car ces variables ne sont pas toutes *indépendantes*. Dans chaque échantillon, il suffit de connaître deux effectifs, pour déduire le troisième par rapport au total. Ce troisième effectif n'apporte donc aucune variabilité à l'ensemble et ne constitue pas un degré de liberté. Il n'y a ainsi que 2 ddl par échantillon d'où un total provisoire de 4 d.d.l. dans le calcul de notre χ^2. Provisoire, parce qu'il faut encore enlever des ddl. En effet, on a estimé des proportions théoriques *à partir des données*. Selon le principe exposé précédemment (**11.1 χ^2 de conformité**), il faut enlever 1 d.d.l. supplémentaire pour chaque paramètre estimé à partir des données. On perd donc 1 ddl pour avoir estimé la proportion théorique des mâles adultes à partir des données et 1 ddl pour avoir estimé celle des femelles adultes de la même manière. En revanche, **notez bien** qu'on n'enlève pas de d.d.l. pour la proportion des immatures, car cette proportion se déduit des deux autres par différence à 1, c'est-à-dire *sans avoir à utiliser les données une troisième fois*. Résumons-nous : 6 variables au départ – 2 d.d.l. pour les totaux – 2 d.d.l. pour avoir utilisé deux fois les données = **2 ddl** pour notre χ^2.

Deuxième méthode (rapide). Le tableau ayant 2 colonnes et 3 lignes, le nombre de d.d.l. est $(2 - 1) \times (3 - 1) =$ **2 d.d.l.**

Je vous avais bien dit que vous préféreriez cette méthode là.

> **Justification de la méthode rapide** : notre tableau de données observées est un tableau dit *de contingence* de C = 2 colonnes et L = 3 lignes. Les totaux de ces lignes et de ces colonnes étant connus, combien de cases peut-on faire varier librement ? Réponse : sur chaque ligne on peut faire varier L – 1 effectifs, ce qui fixe le dernier par différence au total de la ligne. De même, on peut faire varier librement C – 1 effectifs sur une colonne. Le nombre de « cases » pouvant *varier librement* (= nombre de d.d.l.) est donc calculé rapidement par le produit $(L - 1) \times (C - 1)$, d'où ici $(3 - 1) \times (2 - 1) = 2$ ddl.

Si d'aventure (on voit de tout de nos jours) un(e) inconscient(e) vous dit avoir maîtrisé du premier coup la notion de degré de liberté en statistiques, ouvrez de grands yeux admiratifs (ça lui fera plaisir), mais ne vous privez pas de ricaner intérieurement. Pour votre part, abordez ce sujet épineux avec la plus grande attention et prenez votre temps pour réfléchir.

Le χ^2 d'homogénéité est également appelé **χ^2 d'indépendance**, car il équivaut à tester (hypothèse Ho) **l'absence de liaison** entre les lignes et les colonnes. Dans l'exemple ci dessus, un χ^2 significatif nous indique que la probabilité qu'un individu soit d'un certain sexe n'est pas indépendante de l'endroit où on le prélève sur l'estran. *Ceci n'implique pas* un rapport de cause à effet (= que le niveau sur l'estran influence directement le sexe d'un individu). Il y a bien des possibilités pouvant expliquer un tel lien apparent, et il appartiendra à l'expérimentateur de les examiner *si* (et seulement *si*) le χ^2 d'indépendance est significatif.

Exemple 11.4. La latéralité est-elle indépendante du sexe ?
D'après un sondage effectué sur 616 étudiants et étudiantes de la maîtrise BPE des promotions 1998 à 2001 (308 garçons et 308 filles), j'ai observé 57 gauchers et ambidextres déclarés parmi les garçons (soit 18,5%) et seulement 31 gauchères et ambidextres déclarées chez les filles (soit 10,1%). Peut-on accepter l'hypothèse Ho d'indépendance entre la latéralité et le sexe au sein de cette population ?

Les proportions théoriques des catégories [gauchers & ambidextres] et [droitiers] se calculent sur la population globale, et on obtient (57 + 31)/616 = 14,3% et 85,7% respectivement. Ces proportions nous permettent de calculer les effectifs théoriques utilisés par le test du χ^2 d'indépendance, soit 44 G&A et 264 D parmi les filles et les garçons. Il n'y aura ici que 1 degré de liberté : (2 – 1) lignes x (2 – 1) colonnes.

Le calcul du chi2 donne : $\chi^2 = \mathbf{8,96}$

La valeur seuil du χ^2 pour 1 ddl et $\alpha = 0,05$ est de 3,84. Cette valeur est largement dépassée, et on lit dans la table du chi2 que la valeur seuil pour $\alpha = 0,01$ (qui est de 6,63) est elle aussi largement dépassée. On peut donc rejeter Ho, au risque $\alpha = 0,01$ et on écrira dans un rapport : « *Le pourcentage de gauchers et ambidextres est significativement supérieur chez les garçons au sein de la population des étudiants de MBPE ($\chi^2 = 8,96$; 1 d.d.l., $P < 0,01$).* ». Ce résultat suggère que le facteur "latéralité" et "sexe" ne sont pas totalement indépendants au sein de cette population. Peut-être est-ce valable dans la population mondiale en général, peut être aussi que plus de garçons gauchers que de filles gauchères sont attirées par les études biologiques, il y a toujours plusieurs manières d'interpréter ce genre de résultat.

11.3 Conditions d'application du χ^2

Le test du χ^2 (qu'il soit de conformité ou d'indépendance) *n'est pas applicable dans n'importe quelles conditions*. Il repose en effet sur une approximation (une de plus), puisque les effectifs des classes suivent des lois *binomiales* (un individu appartient à la classe avec la probabilité p et n'y appartient pas avec la probabilité $q = 1 - p$), alors que la loi du χ^2 est *stricto sensu* la distribution d'une somme de carrés de lois *normales centrées réduites*. Pour que l'approximation puisse être faite, il est nécessaire que toutes les binomiales en présence soient suffisamment proches d'une loi normale. On a vu (INTERVALLES DE CONFIANCE D'UN POURCENTAGE) que cette condition était satisfaite pour une binomiale si np et $nq > 5$. En clair et sans décodeur, les effectifs *théoriques* de chaque classe doivent être au moins de 5 individus. Je dis bien les effectifs **théoriques**, pas les effectifs **observés.** Les effectifs **observés** peuvent prendre n'importe quelle valeur y compris la valeur zéro sans aucun problème. Confondre ces deux aspects est une erreur fréquemment commise, donc je me permets d'insister sur ce point. Il existe par ailleurs une certaine tolérance vis-à-vis de cette condition idéale concernant les effectifs théoriques, car le test du χ^2 est relativement *robuste*. Cette tolérance est énoncée dans la **règle de Cochran**, que l'on peut résumer ainsi :

> « **On peut effectuer un χ^2 si au moins 80% des valeurs théoriques sont au moins égales à 5, et que toutes les valeurs théoriques sont supérieures à 1** »

Traduit en détails cela signifie que, pour utiliser la formule classique du chi2:

(i) *dans le cas de deux effectifs théoriques*, il faut impérativement des effectifs théoriques de 5 ou plus;

(ii) *à partir de 5 effectifs théoriques*, on peut admettre **un** effectif théorique « faible » (entre 1 et 5);

(iii) *à partir de 10 effectifs théoriques,* on peut en admettre **deux** « faibles », etc…

Soyez cependant conscient du fait qu'on touche là les extrêmes limites des possibilités du test. Que faire alors si vous vous trouvez coincés ? Il y a trois méthodes.

Méthode 1. Regroupez des classes entre elles. Vous aurez ainsi des effectifs théoriques plus grands. N'oubliez pas de diminuer le nombre de d.d.l. en conséquence (chaque regroupement de deux classe fait disparaître un d.d.l.). Attention également à ce que votre « regroupement » soit logique : ne regroupez pas entre eux « pour faire masse » les rares individus très clairs avec les rares individus très sombres ! Inconvénient de cette méthode : vous perdez évidemment une partie de votre information.

Méthode 2. Utilisez une formule du chi2 adaptée aux petits effectifs.
Achtung ! Cette méthode n'est valable que pour le cas ou k = 2 classes, et si les effectifs *théoriques* ne sont quand même pas microscopiques (disons supérieurs à 2). La correction à apporter à la formule du chi2, due à Yates, consiste à diminuer la valeur absolue de chaque différence obs – théo de 0,5 avant d'élever au carré. La formule du chi2 devient :

$$\chi^2_{Yates} = \sum \frac{\left(\left|obs - théo\right| - 0{,}5\right)^2}{théo}$$

La lecture dans la table du chi2 se fait sans modification (le nombre de d.d.l. ne change pas).

Méthode 3. Utilisez un test indifférent aux petits effectifs : le test exact de Fisher.
Ce test peut être effectué (de manière fastidieuse) à la calculatrice dans le cas où k = 2 classes (tableau de 2 × 2 cases), en revanche il demande **impérativement** un logiciel dans les autres cas (rares sont d'ailleurs les logiciels qui fassent ce calcul, on peut citer SAS en particulier). En effet, le test exact de Fisher consiste à générer les *milliers* voire les *millions* de tableaux possibles ayant les mêmes totaux de lignes et de colonnes que le tableau de données observées, puis de calculer la proportion *exacte* de ceux qui sont encore plus éloignés de l'hypothèse Ho que le vôtre. Si cette proportion est inférieure à 5%, on peut déduire que votre résultat appartient à une catégorie de résultats très improbables si Ho était vraie, et on rejette Ho en prenant un risque α de 5%. La procédure à utiliser dans le cas d'un tableau de 2 × 2 cases est décrite partout, et très clairement dans Schwartz (1993)[1], je vous renvoie donc à ces bonnes lectures, si jamais vous devez utiliser un test exact de Fisher.

1: Daniel Schwartz (1993), Méthodes statistiques à l'usage des médecins et des biologistes. Médecine-Sciences, Flammarion

Résumé du chapitre 11

La comparaison entre une distribution observée et une distribution théorique s'effectue au moyen du test du **chi2 de conformité**. Dans le cas particulier d'un seul pourcentage, cette approche est en fait équivalente au calcul d'un intervalle de confiance autour du pourcentage observé, en déterminant si la valeur théorique appartient ou non à cet intervalle. La comparaison mutuelle de deux ou N distributions observées s'effectue au moyen d'un test du **chi2 d'homogénéité** également appelé **chi2 d'indépendance** (car il revient à tester s'il existe un lien entre les lignes et les colonnes du tableau de données). Les tests du chi2 ne peuvent pas s'effectuer si certains effectifs théoriques sont inférieurs à 5 individus, mais cette obligation présente en fait une certaine flexibilité, délimitée par la **règle de Cochran**. Lorsqu'on ne peut pas la respecter avec le tableau de données initial, il est possible soit de modifier ce tableau en fusionnant des lignes ou des colonnes, soit d'utiliser la correction de Yates (seulement dans les tableaux de 2×2 cases) soit en dernier lieu de faire appel au **test exact de Fisher**, (ce qui nécessite impérativement un logiciel si le tableau dépasse 2×2 cases).

12. Corrélation n'est pas raison.

12.1 Corrélation ou régression ?

Dans ce domaine plus que dans tout autre, une grande confusion règne dans l'esprit des débutants, qui emploient indifféremment un mot pour l'autre, et seraient bien en peine d'expliquer la différence entre les deux. C'est parfaitement normal, puisque la corrélation et la régression poursuivent le même but (caractériser la liaison statistique entre deux variables quantitatives), peuvent s'appliquer aux même données en fournissant la même conclusion, et sont souvent utilisées... conjointement. Cependant, ce dernier fait illustre bien qu'elles posent des questions un peu différentes. Pour faire court :

La **corrélation** cherche à mesurer la *force*, la *rigidité* de la liaison statistique entre X et Y. Si cette liaison est rigide, il sera en particulier possible d'avoir une bonne idée de Y en connaissant seulement X, et vice versa. Exemple : s'il existe une bonne corrélation entre la taille d'une dent et la taille de son propriétaire chez les tyranosaures (*Tyranosaurus rex*), alors il est possible de déduire de manière approximative la taille d'un spécimen fossile dont on a juste retrouvé une dent. Réciproquement, la découverte éventuelle d'un squelette de *T. rex* sans tête (donc sans dents) permettrait quand même d'estimer quelle était la taille de celle-ci.

La **régression (linéaire)** cherche à caractériser la *pente* de la droite pouvant résumer au mieux la relation entre X et Y, une fois choisies des unités de mesure pour X et Y. Exemple, si la dose efficace d'un anesthésique est de 5mg/kg de poids de corps (pente de 5 pour 1 *avec ces unités là*), un gain de poids de 10kg chez un patient obligera pour le même effet anesthésique à augmenter la dose de $10 \times 5 = 50$mg. Cependant (et c'est là où corrélation et régression marchent main dans la main), la pente en question n'a d'intérêt que si la relation entre la dose efficace et l'effet est suffisamment rigide. Si cette relation est en réalité très floue, le risque de sous-doser ou de sur-doser l'anesthésique devient inquiétant. D'où l'intérêt de connaître la force de la liaison en plus de sa pente.

En résumé, **corrélation** et **régression** permettent toutes deux de répondre à la question "y a-t-il un *lien statistique* entre X et Y", quelle est la *force* de ce lien éventuel (corrélation) et quelle est la *pente* de la relation éventuelle pour un jeu d'unités donné (régression):

Régression	***Corrélation***
Question 1. Y a-t-il un lien *statistique* entre X et Y **Question 2.** Quelle est la *relation numérique* entre X et Y pour un jeu d'unités donné ? (Si X vaut « telle valeur », dans les unités choisies pour exprimer X, quelle sera *en moyenne* la valeur de Y , dans les unités choisies pour exprimer Y ?)	**Question 1.** Y a-t-il un lien *statistique* entre X et Y ? **Question 2.** S'il existe, quel est la *rigidité* de ce lien ? En particulier, quelle est la fraction de la variance de Y qui subsiste si je *fixe* X ?
Test de la pente	Test du coefficient de corrélation
Si test de la pente significatif, Il existe **un lien, mais** un test significatif est à lui seul **insuffisant** pour démontrer que c'est X qui agit sur Y, sauf dans une situation expérimentale dans laquelle tous les facteurs sont strictement contrôlés et que l'expérimentateur fait varier X. Le *calcul de l'équation de la droite de régression* est justifié et permet de répondre à la deuxième question. Cependant, le calcul du coefficient de corrélation est utile, car il est intéressant de connaître la force du lien entre X et Y.	*Si coefficient de corrélation significatif*, Il existe **un lien**. Là encore, **impossible** de déterminer la nature du lien sur la seule base du test, sauf dans le cas expérimental décrit ci-contre, où l'on conclurait à un lien causal. La valeur du coefficient de corrélation R et du coefficient de détermination R^2 ont un sens, et permettent de répondre à la question 2.
Si test de la pente non significatif, Deux explications possibles: (i) le lien existe mais il n'y avait pas suffisamment de données pour le mettre en évidence; (ii) il n'y a pas de lien.	*Si coefficient de corrélation non significatif* Deux explications possibles: (i) le lien existe mais il n'y avait pas suffisamment de données pour le mettre en évidence; (ii) il n'y a pas de lien.

Le présent chapitre traitera de la **corrélation** (pour la régression, voir Chapitre 13).

12.2 Corrélation n'est pas raison.

La corrélation entre deux variables quantitatives X et Y est l'existence d'une liaison *statistique* entre elles, quelle qu'en soit la raison. "Liaison statistique" signifie ici que les deux variables ne semblent pas varier indépendamment l'une de l'autre : connaître la valeur de la variable X (ou Y) pour un individu vous fournit une information sur sa valeur pour la variable Y (ou X) . L'existence d'une liaison *statistique* ne signifie pas nécessairement l'existence d'une liaison *causale* (c'est-à-dire un lien *de cause à effet*, une action de X sur Y ou de Y sur X). Si une liaison statistique est décelée entre deux variables, l'existence d'une liaison causale directe entre elles sera seulement *une* des possibilités à explorer. Nous reviendrons en profondeur sur cette notion, mais il était important de la signaler tout de suite, car la confusion entre "*X et Y sont corrélées significativement*" et "*X a un effet sur Y (ou vice versa)*" est probablement une des erreurs les plus communes de toute l'analyse des données.

12.3 La notion de covariance (co-variance : "variance ensemble").

La covariance aurait pu également être nommée "coévolution" (mais ce terme est déjà utilisé, et avec un tout autre sens, en biologie évolutive). En statistiques, lorsque deux variables évoluent dans le même sens, on dit qu'elles *covarient* de manière *positive*. Par exemple, la taille et la longévité des mammifères varient globalement dans le même sens : les espèces de grande taille vivent en général plus longtemps que les espèces de petite taille. Il n'y a cependant aucun lien de cause à effet entre la *taille* et la *longévité*, comme l'examen des longévités à l'intérieur de n'importe quelle espèce (la nôtre y compris) permet de s'en apercevoir : être de grande taille ne vous dit rien sur vos chances de vivre vieux, on n'a remarqué aucune tendance au gigantisme parmi les centenaires !

Lorsque deux variables évoluent de manière *opposée* l'une par rapport à l'autre, on dit qu'elles *covarient* de manière *négative*. Par exemple, plus une espèce d'oiseau a une vitesse de croisière *élevée*, plus son rythme cardiaque au repos est *lent*. Cette relation paradoxale en apparence, s'explique tout simplement parce que les espèces qui volent vite sont (sauf exception) de grande taille, alors que les petits passereaux volent assez lentement. Or les grands animaux ont des rythmes cardiaques bien plus lents que les petits. D'où une corrélation négative très nette, lorsqu'on raisonne au niveau de l'espèce. Comme d'habitude en biologie, cette règle aura des exceptions remarquables (les oiseaux les plus rapides de tous sont... les martinets, qui sont de petite taille et ont donc un rythme cardiaque très élevé). Il est donc possible, connaissant la vitesse de vol typique d'une espèce, de déduire quel sera approximativement son rythme cardiaque au repos (et réciproquement). Là encore, malgré la corrélation négative très nettement observable, il n'y a aucun lien direct de cause à effet. Il est même parfaitement évident que la vitesse *en vol* ne peut influencer le rythme cardiaque *au repos* et réciproquement !

Enfin, lorsque deux variables sont (*a priori*) totalement *indépendantes* l'une de l'autre, on dit que leur covariance est *nulle*. On peut par exemple supposer que les fluctuations de température à –11000 m au fond de la fosse des Mariannes et les fluctuations de température

au fond de vos fosses nasales sont indépendantes[1]. Si on avait l'idée saugrenue de mesurer chaque jour ces deux données et de le représenter sous forme d'un nuage de points, on obtiendrait probablement un graphe démontrant (mais était-ce nécessaire ?) que votre température nasale est superbement indépendante de ce qui se passe à 11 000 mètres de fond dans le Pacifique (et vice versa). Ainsi, la rare faune adaptée à ces profondeurs abyssales et à la température très stable qui y règne ne court aucun danger si vous piquez une bonne fièvre.

La covariance entre deux séries de données X et Y (par exemple deux mesures quelconques effectuées sur *n* individus) peut évidemment être calculée de manière précise, et sa formule ressemble très logiquement à celle de la variance :

$$\mathrm{Cov}(XY) = [(x_1 - \boldsymbol{m}_x)(y_1 - \boldsymbol{m}_y) + (x_2 - \boldsymbol{m}_x)(y_2 - \boldsymbol{m}_y) + \ldots + (x_n - \boldsymbol{m}_x)(y_n - \boldsymbol{m}_y)] / (n - 1)$$

Avec :
n le nombre d'individus dans l'échantillon
$x_1, x_2 \ldots x_n$ les valeurs de X pour les individus 1, 2…n
$y_1, y_2 \ldots y_n$ les valeurs de Y pour les individus 1, 2…n
$\boldsymbol{m}_x$ la moyenne observée des valeurs X sur l'échantillon
$\boldsymbol{m}_y$ la moyenne observée des valeurs Y sur l'échantillon

La covariance revient donc à remplacer les carrés $(x - \mathbf{m})$ $(x - \mathbf{m})$ de la variance classique, par les produits $(x - \boldsymbol{m}_x)(y - \boldsymbol{m}_y)$, réalisant ainsi une sorte de variance à deux dimensions.

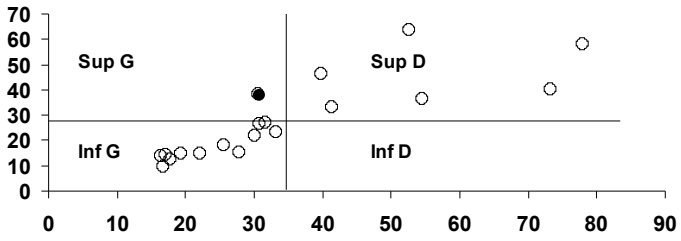


Fig 12.1 longueur de l'œuf (mm) en fonction de la longueur de l'oiseau (cm) chez 19 espèces d'oiseaux d'Europe choisies au hasard. La croix est centrée sur le barycentre du nuage de points, de coordonnées (m_X,m_Y)

On peut comprendre de quelle manière la covariance va se comporter en observant le nuage de points de la **figure 12.1**, dans lequel on peut soupçonner une covariance *positive,* puisque X et Y semblent varier de concert (les plus gros oiseaux – quelle surprise – pondent les plus gros œufs). Ce nuage de points a été découpé en 4 secteurs centrés sur son barycentre de coordonnées G = ($\boldsymbol{m}_x$, $\boldsymbol{m}_y$), les moyennes de X et de Y. Il est facile de deviner quel sera le signe de la contribution apportée à la covariance globale par un point de coordonnées (x,y) selon le secteur dans lequel il se trouve :

[1] Je pars évidemment de l'hypothèse raisonnable que vous ne vous trouvez pas en ce moment au fond de la fosse des Mariannes, sous une pression de plus de mille tonnes au cm².

Secteur du point (x,y)	Signe de $(x - \boldsymbol{m}_x)$	Signe de $(y - \boldsymbol{m}_y)$	Signe du produit $(x - \boldsymbol{m}_x)(y - \boldsymbol{m}_y)$
Supérieur droit	+	+	+
Inférieur gauche	–	–	+
Inférieur droit	+	–	–
Supérieur gauche	–	+	–

Lorsque, comme ici, X et Y varient de concert, le nuage de points s'incline le long d'un axe imaginaire partant du secteur inférieur gauche et se dirigeant vers le secteur supérieur droit, en passant par le barycentre G des données. En conséquence, la majorité des points sont situés dans ces deux secteurs. Leur contribution (positive) à la covariance est supérieure à la contribution (négative) des points situés dans les deux autres secteurs (ici, il n'y en a qu'un, coloré en noir). On obtient donc une covariance de signe positif, qui sera d'autant plus forte que le nuage est allongé (et donc que la liaisons entre X et Y est franche). Si jamais les variables X et Y avaient varié plutôt en opposition l'une par rapport à l'autre, le nuage de points aurait été incliné au contraire vers le bas, et ce sont les points des secteurs supérieur gauche et inférieur droit dont la contribution (négative) à la covariance auraient dominé. Comme vous le voyez, (1) le signe de la covariance permet bien de déceler si X et Y sont liées de manière positive ou négative, (2) la valeur absolue de la covariance mesure semble-t-il la force de la liaison.

Naturellement, il serait trop simple de pouvoir calculer la covariance à partir de vos données et de conclure directement. En effet, comme d'habitude, les fluctuations d'échantillonnage vont entrer en jeu, et vous n'obtiendrez jamais une covariance nulle, même s'il n'y a pas la moindre liaison entre vos données. Il faudrait être capable de déterminer si la valeur de la covariance s'éloigne *trop* de zéro, pour que l'hypothèse Ho "absence de liaison entre X et Y" soit crédible. Cependant, il est impossible de savoir directement si la covariance est "trop positive" ou "trop négative" pour une raison simple : la valeur de la covariance dépend entièrement des unités choisies pour graduer les axes X et Y ! La valeur de la covariance sera évidemment complètement différente selon que la taille de l'œuf est mesurée en millimètres ou en années-lumière. Pour contourner ce problème, il y a deux solutions. La première sera décrite dans le chapitre 13, car elle concerne la notion de droite de régression. La seconde consiste à utiliser le fait que la valeur maximale que peut prendre la covariance entre X et Y est le produit des écarts-types $[s_X\, s_Y]$. On peut donc créer un indice *sans dimension*, en divisant la valeur de la covariance observée (dimension : "unité-de-X × unité-de-Y") par ce produit (qui a les mêmes unités, donc la même dimension). Ce ratio, par construction, est donc sans unité, varie de – 1 à + 1, et n'est autre que le ***coefficient de corrélation*** (ou R de Pearson).

12.4 Mon nom est Pearson.

Comme vu ci-dessus, le coefficient de corrélation, inventé par Karl Pearson et désigné par la lettre R, se calcule de la façon suivante :

$$R = \text{cov}(X,Y) / (s_X\, s_Y)$$

avec
cov(X,Y) la covariance estimée à partir des données
s_X et s_Y les écarts-types des variables X et Y estimés à partir des données

La lettre R normalement utilisée est le r *minuscule*, mais, comme on le voit moins bien dans un texte imprimé, j'utiliserai le R majuscule pour des raisons de lisibilité.

Totalement *indépendant des unités choisies,* R varie entre –1 (liaison statistique négative totalement rigide) à +1 (liaison statistique positive totalement rigide). La "rigidité" dont il est question ici signifie que si R vaut 1 (ou –1), il est possible de connaître *exactement* la valeur de X en connaissant celle de Y (et vice versa bien entendu). En revanche, si R est proche de zéro, connaître X ne donne qu'une très vague indication sur la valeur de Y (et vice versa).

Cependant, là encore, pas question d'utiliser le R calculé à partir des données directement, comme s'il représentait la véritable liaison existant dans la réalité. A cause des inévitables fluctuations d'échantillonnage, la valeur de R calculée à partir de vos données, n'est jamais qu'une *estimation* de la véritable valeur ρ (la lettre grecque **rho**), qui vaut peut-être zéro, reliant (ou pas) les variables X et Y dans la réalité. Il nous faudrait donc connaître la distribution du paramètre R, sous l'hypothèse Ho de l'absence de liaison entre X et Y. Heureusement, on la connaît.

12.5 Test du coefficient de corrélation.

La loi de distribution de R, sous l'hypothèse Ho "*Aucune liaison statistique entre X et Y*", est connue, et ses valeurs seuils sont consignées dans une table qui se lit en fonction du nombre de degrés de liberté permettant le calcul de R à partir de vos *n* couples de données. Ce nombre est $n - 2$ degrés de liberté. On perd deux d.d.l. car on a au départ *n* points de données, mais il nous a fallu utiliser nos propres données **deux fois** pour estimer les moyennes m_x et $\boldsymbol{m_Y}$ (on en avait en effet besoin pour estimer la covariance et des écarts-types s_X et s_Y). Comme d'habitude, cette manière de procéder va avoir tendance à rapprocher artificiellement notre modèle théorique de nos propres données. Cela revient à dire qu'il y a, en réalité, dans notre système, moins de variables aléatoires indépendantes que les *n* données observées. Deux ddl sont ainsi perdus, ce qui nous en laisse $n - 2$.

Par exemple, si $n = 20$ (donc 18 d.d.l.), on lit dans la table du R de Pearson, la valeur seuil 0,4438 pour un risque $\alpha = 0{,}05$. Cela signifie concrètement, qu'il y a seulement 5% de chances que R sorte de l'intervalle [–0,4438 : +0,4438] sous le seul effet du hasard dans un échantillon de 20 individus. Donc, après avoir calculé R sur les données, on conclura à partir de sa valeur absolue | R | :

Si | R | > R_{seuil} on rejette Ho. On conclut donc que les données indiquent un lien statistique, *mais pas forcément un lien de cause à effet*, entre les variables X et Y. Cette décision est associée au risque α choisi pour le test.

Si | R | < R_{seuil} on ne peut pas rejeter Ho sur la base de ces données. On conclut donc qu'on n'a pas de preuves suffisantes pour affirmer l'existence d'une liaison statistique entre X et Y. Cette décision est associée à un risque β, inconnu, mais d'autant plus grand que l'échantillon est petit et que le lien entre X et Y, si il existe en réalité, est faible.

La table du R de Pearson est limitée à $n = 100$, mais il existe une relation entre la loi de r et la loi du t de Student (**valable uniquement sous notre hypothèse Ho que R = 0**) qui permet d'utiliser la table du t si l'effectif dépasse ce nombre. La voici :

$$\frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$ *suit une loi du t de Student à (n – 2) degrés de liberté.*

On pourra donc utiliser la table du t de Student quel que soit n (en lisant sur la ligne "infini" pour les $n > 100$, sinon lire dans la table du R).

L'avis de Parsimoni et Abonessian.

Parsimoni — Et voilà, comme d'habitude, on se précipite tête baissée sur le test statistique !

Abonessian — C'est tout de même comme ça depuis Pearson, Giuseppe. Vous n'allez tout de même pas remettre en cause aussi le R de Pearson ?

Parsimoni — Ai-je dit une chose pareille ? J'ai tout de même le droit de me demander à haute voix pourquoi on utilise un test, au lieu de calculer un intervalle de confiance autour du R observé !

Abonessian — Vous savez bien qu'il faut pour cela passer par une transformation de Fisher, alors que la table du R donne un résultat immédiat, même s'il est plus pauvre d'information. Vous devez aussi reconnaître que la conclusion finale est la même : si la valeur zéro est contenue dans votre intervalle de confiance, vous concluez que la corrélation n'est pas significative.

Parsimoni — Toi et les autres testomanes compulsifs n'avez donc que le mot "significatif" dans votre vocabulaire ? Je t'ai répété mille fois que je n'accordais aucun statut magique à la valeur zéro ! Je veux savoir quelle est la *gamme de valeurs plausibles* pour la véritable valeur de **rho** reliant X et Y dans la réalité. Toute étude raisonnable entre deux variables X et Y va forcément concerner deux variables qui ont une *certaine* liaison, même faible, l'une avec l'autre. Je veux savoir quelle est la ***force*** vraisemblable de la liaison. Je n'ai que faire de votre significativité !

Abonessian — Je n'idolâtre pas le seuil $\alpha = 0{,}05$ Giuseppe, c'est juste une norme pratique, c'est tout. Il n'a jamais été question d'éradiquer les intervalles de confiance.

Parsimoni — Encore heureux! Je note également qu'on n'a toujours pas abordé les choses sérieuses : où est R^2, l'indispensable coefficient de détermination dans tout ça ? Comment mesurer de manière très concrète la force réelle du lien entre X et Y avec un simple R ? Et pourquoi ne pas avoir dit tout de suite que le coefficient de corrélation de Pearson se fait piéger chaque fois que la relation entre X et Y n'est pas une droite ou que les variables ne sont pas distribuées selon une loi normale ?

Abonessian — Mais Giuseppe, le chapitre n'est pas encore fini !

12.6 Ce qu'un coefficient de corrélation de Pearson ne sait pas voir.

Le Professeur Parsimoni soulève, avec sa fougue habituelle, deux thèmes importants que je n'avais pas encore abordés. Je vais commencer par le second. En effet, le coefficient de corrélation de Pearson est conçu pour mesurer précisément la liaison statistique entre deux variables qui évoluent *proportionnellement* l'une par rapport à l'autre, et qui le font de manière *constante* (on dit "monotone"). Si on s'éloigne de cette condition, le coefficient de corrélation sera artificiellement faible. Expliquons cela plus en détail.

Relation proportionnelle — Que la liaison entre X et Y soit rigide ou floue, le nuage de point doit tendre à s'aligner le long d'une *droite* imaginaire. S'il tend à s'aligner le long d'une *courbe* croissante (relation logarithmique, exponentielle) ou décroissante (exponentielle négative par exemple) le R calculé sera artificiellement faible par rapport à la véritable force de la liaison entre X et Y. Dans les cas où la relation entre X et Y est mathématiquement connue, on peut toutefois appliquer certaines transformations pour se ramener à un cas linéaire, et on en verra un exemple plus loin avec les *relations d'allométrie*.

Constance (monotonie) de la relation proportionnelle — Le coefficient R se fera totalement piéger si la relation entre X et Y change de sens à un moment donné (forme parabolique par exemple). Dans ce cas, le R calculé pourra être très proche de zéro, alors que les points expérimentaux suivent impeccablement une courbe parabolique. En termes techniques, on dit que la relation entre X et Y doit être *monotone* (constamment croissante ou constamment décroissante) pour que le calcul de R soit correct. Conclusion : si une inspection visuelle de vos données vous fait soupçonner une relation nettement non linéaire, voire non monotone entre X et Y, le R de Pearson n'est pas l'outil approprié, en tout cas pas sur les données brutes.

12.7 R^2 dit tout[2].

Passons maintenant à la deuxième remarque de G. Parsimoni, concernant l'aspect trompeur de R. Il est en effet très facile de se laisser bluffer par une valeur de R élevée, et d'en conclure avec enthousiasme que la liaison entre X et Y doit être vraiment forte. C'est pourquoi, il est utile de se familiariser avec le grand frère de R, j'ai nommé R^2 (R au carré), le coefficient de détermination. Ce coefficient représente (attention, accrochez-vous) *la proportion de la variance de Y qui disparaît, si on fixe X (ou vice versa)*. Si vous avez compris du premier coup, vous êtes très forts. Reprenons. Supposons que X et Y soient liés de manière absolue (connaître X permet de déduire Y exactement). Cela signifie que, même si les valeurs de Y sont différentes *lorsque X varie* (la variance globale de Y n'est pas nulle), elles ne varient pas *pour un X donné*. Dans cette situation extrême, quelle est la variance de Y si on fixe X ? Elle vaut évidemment zéro : si on fixe X, alors Y est fixé aussi. Or, que vaut R dans ce cas ? Il vaut 1, donc R^2 aussi. Un R^2 de 1 (soit 100%) signifie que si on fixe X, alors 100% de la variance de Y disparaît. Vous voyez, ça marche. Prenons l'exemple opposé : l'absence totale de liaison entre X et Y. Dans ce cas, R vaut zéro donc R^2 aussi (donc 0%). Si on fixe X, quelle est la fraction de la variance de Y qui est éliminée ? Réponse 0% : la variance de Y n'est pas

[2] il était impossible de résister à l'envie de rendre hommage au sympathique petit robot de Star Wars.

diminuée d'un iota si on fixe X, puisque Y se moque éperdument de ce que fait X : il y a indépendance entre X et Y. Passons maintenant à des situations plus intéressantes.

Supposons un coefficient de corrélation R = 0,7. En morphologie, c'est courant. En écologie, ce type de coefficient de corrélation permettrait de sabrer le champagne tant il est rare. Mais que veut-il dire au juste concernant la force de la liaison entre X et Y ? On peut en avoir une idée avec le coefficient de détermination R^2, qui vaut donc 0,7 × 0,7 = 0,49 soit 49%. En clair, si on fixe X, alors 49% de la variance de Y disparaît. C'est déjà bien, mais cela signifie quand même que *la moitié de la variance de Y subsiste* même si on fixe la valeur de X. Connaître X ne donne donc qu'une idée finalement *très vague* de la valeur de Y. Donc, ne vous laissez pas hypnotiser par les valeurs de R et ayez le réflexe de toujours calculer R^2, c'est plus parlant. Voilà pourquoi Giuseppe Parsimoni accordait tant d'importance à cette notion. Il nous faut maintenant aborder sa critique sur test "Significatif" et ce qu'il implique concrètement.

Le fait que le test de R soit "significatif" veut dire simplement qu'il y a (vraisemblablement) une liaison statistique entre X et Y. Ne tombez surtout pas dans l'erreur consistant à croire que le niveau de significativité du test ($P < 0,05$ ou $P < 0,001$) vous indique la force de la liaison. Observez la table du R et vous constaterez en particulier, qu'il suffit d'un effectif relativement modeste (une quarantaine d'individus), pour que le test soit "significatif" dès que R dépasse 0,3. Vous savez maintenant comment avoir une idée un peu plus concrète de la force de la liaison que cela traduit : il suffit de calculer R^2. Celui-ci vaut 0,3 × 0,3 = 0,09 soit 9%. En clair, un coefficient de corrélation de 0,3 veut dire que pour un X fixé, la variance de Y a seulement été diminuée de 9%. En d'autres termes, connaître X ne réduit quasiment pas l'incertitude sur Y. Un coefficient de corrélation de 0,3 traduit une liaison *très faible*, très floue, et ça n'a pas empêché le test d'être significatif. Voyons plus loin. Supposons que vous ayez beaucoup d'individus (disons mille). Alors, un coefficient de corrélation de R = 0,1 sera "*hautement significatif*" et vous écrirez triomphalement "*X et Y sont corrélées de manière hautement significative, P < 0,001*". Mais concrètement, quelle est la force de la liaison mise à jour ? Elle est tout simplement risible, puisque $R^2 = 0,1 \times 0,1 = 0,01$ soit 1%. En clair, pour un X fixé, il restera encore 99% de la variabilité totale de Y. Dans ces conditions, connaître X ou rien, c'est la même chose. Conclusion : avec beaucoup d'individus, on est capable de déceler une liaison statistique *très faible* entre deux variables. C'est plutôt une bonne nouvelle, encore faut-il être conscient qu'un test peut être "significatif" même si la liaison statistique est si faible qu'elle a un intérêt pratique nul. Hélas, la capacité d'un coefficient de corrélation "significatif" à vous induire en erreur ne s'arrête pas là. Lisez plutôt ce qui suit.

12.8 Interprétation prudente d'un coefficient de corrélation significatif.

"***Corrélation*** *n'est pas* ***raison***" est un dicton statistique important. Il rappelle que le fait de trouver une corrélation (même "élevée" et "hautement significative") entre la variable X et la variable Y ne démontre pas du tout qu'il y a *un lien de cause à effet* entre X et Y (ou entre Y et X). Dans une situation d'observation (c'est-à-dire en dehors d'un dispositif expérimental dans lequel tous les facteurs sont strictement contrôlés), si on constate une forte corrélation entre deux variables, il faut donc impérativement résister à l'envie de proclamer tout de suite un lien

de cause à effet. L'établissement d'un tel lien nécessite une expérimentation en conditions contrôlées ou une accumulation d'autres observations dans différentes situations, qui permette d'exclure les autres possibilités non causales. Un grand classique est la corrélation observée systématiquement entre deux variables X et Y, lorsqu'elles sont toutes les deux influencées par la même *troisième* variable Z. Le nombre Y de coups de soleil attrapés sur une plage est fortement corrélé à la température moyenne X de l'air ce jour là. Naturellement, la température X n'influence pas Y[3]. En réalité, X et Y sont sous la dépendance de Z, la véritable variable causale, c'est à dire la quantité de rayonnement solaire.

On peut également observer une très forte corrélation négative entre la vitesse des ordinateurs et leur prix entre 1945 et 2004. Ces données démontreraient, avec une clarté aveuglante, que *plus un ordinateur est rapide, moins il coûte cher*, s'il ne s'agissait d'une simple corrélation. La variable explicative pertinente ici est bien évidemment le progrès technique qui a simultanément augmenté la vitesse et baissé les coûts de fabrication des ordinateurs depuis 1945, avec la très spectaculaire corrélation négative qui en résulte. N'ayons pas peur de le rabâcher : **corrélation n'est pas raison**.

12.9 Calcul de l'intervalle de confiance d'un coefficients de corrélation de Pearson.

La distribution de R est complexe, et elle n'est tabulée que sous l'hypothèse que R = 0. De même, la relation entre la loi du R et la loi du t de Student n'est valable que dans cette condition (R = 0). Calculer un intervalle de confiance autour d'un R observé serait donc pénible, si Sir R. A. Fisher (cet homme ne dormait-il donc jamais ?), n'était pas passé par là. Il nous a mitonné une transformation dite "de Fisher", qui permet d'obtenir une variable auxiliaire Z, suivant une loi normale de variance connue à partir de la loi du R. Cette transformation se présente sous l'avenante forme suivante

$$Z = 0{,}5 \times [\ln (1 + R) - \ln (1 - R)]$$

avec :
R le coefficient de corrélation calculé sur vos *n* données
ln le logarithme népérien

Cette variable Z suit donc, comme dit plus haut, une loi approximativement normale. Sa variance s^2_Z est particulièrement simple à calculer :

$$s^2_Z = 1/(n - 3)$$

avec *n* le nombre de couples de données (*x*,*y*)

d'où un écart type valant :

$$s_Z = \sqrt{[1/(n - 3)]}$$

[3] les petits malins feront remarquer qu'en réalité X agit bien sur Y dans ce cas, même si c'est indirectement : une température élevée pousse à découvrir son épiderme. Rien n'est jamais simple en biologie.

Puisque Z suit une loi normale, on sait que 95% de ses valeurs sont situées dans un intervalle de 1,96 écarts types autour de sa moyenne. On peut alors estimer l'intervalle de confiance de Z comme d'habitude :

$$IC_{95}Z = [Z \pm 1{,}96\ s_Z]$$

En clair, les bornes inférieures et supérieures de l'intervalle de confiance de Z sont :

$$Z_{inf} = Z - 1{,}96\ s_Z$$
$$Z_{sup} = Z + 1{,}96\ s_Z$$

Il suffit maintenant de faire la manœuvre inverse de la transformation de Fisher pour trouver R_{inf} et R_{sup} les bornes de l'intervalle de confiance de R. On pose :

$$Z_{inf} = 0{,}5 \times [\ln (1 + R_{inf}) - \ln (1 - R_{inf})]$$
$$Z_{sup} = 0{,}5 \times [\ln (1 + R_{sup}) - \ln(1 - R_{sup})]$$

Quelques lignes de calcul (passionnantes, comme d'habitude) plus tard, et en se souvenant que $e^{\ln(x)} = x$, on obtient la transformation inverse permettant de trouver le R à partir d'une valeur de Z :

$$R = (e^{2Z} - 1)/(e^{2Z} + 1)$$

Formule qui nous permet de retrouver R_{inf} et R_{sup}, les bornes de l'intervalle de confiance de R que nous cherchions:

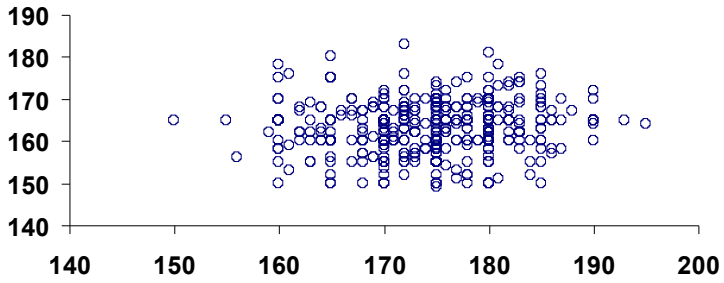
$$IC_{95}R = [(e^{2Z_{inf}} - 1)/(e^{2Z_{inf}} + 1) \text{ —— } (e^{2Z_{sup}} - 1)/(e^{2Z_{sup}} + 1)]$$

Ceux d'entre vous qui ont définitivement condamné les touches ln et e^x de leurs calculatrices en leur enfonçant un épieu d'argent dans le cœur, pourront se baser sur une table de conversion Z⇔R (voir la fin de ce chapitre), ou bien ils utiliseront un logiciel statistique.

Il est grand temps de passer à un exemple concret, en étudiant la relation possible entre la taille du père et la taille de la mère de 460 étudiants de la maîtrise BPE, d'après les tailles fournies par les étudiants eux-mêmes. Si nous décelons une liaison statistique entre ces deux valeurs, nous pourrons (par exemple) émettre l'hypothèse que les individus tiennent compte de la taille de leur partenaire pour former leur couple. Cet exemple nous permettra de passer en revue l'utilisation de R, son test, en passant par la loi du t de Student, le calcul de son intervalle de confiance grâce à la transformation de Fisher, l'utilisation de R^2 et le calcul de son intervalle de confiance. Tout un programme.

Voici d'abord le graphe obtenu en reportant la taille de la mère en fonction de la taille du père.

Fig 12.2 Taille de la mère en fonction de la taille du père chez 460 étudiants de la maîtrise MBPE.



Exemple 12.1 Corrélation taille de la mère/taille du père
Les données de base sont les suivantes : n = 460
Pères : $m_X = 174,5$ cm et $s^2_X = 48,7$ d'où $s_X = 7,0$
Mères : $m_Y = 163,2$ cm et $s^2_Y = 35,4$ d'où $s_Y = 5,9$
Covariance (X,Y) = 3,938
$R = cov(XY)/(s_X \times s_Y) = 0,095$

Faisons une petite pause à ce stade : le graphe ne révèle aucune tendance qui saute aux yeux (quelque soit la taille des pères, celle des mères fluctue dans la même gamme, et vice versa) et le coefficient de corrélation que nous venons de trouver est vraiment très faible. La cause semble donc entendue : il n'y a aucune liaison significative entre X et Y. Pas si vite. Faisons maintenant le test. Du fait que n >> 100, nous ne pouvons utiliser la table du R de Pearson (qui va jusqu'à 100), mais il suffit de se ramener à une loi du t de Student avec $n - 2 = 460 - 2 = 458$ degrés de liberté. En pratique, ce nombre de ddl se confond avec l'infini, et nous allons en réalité comparer la valeur obtenue avec les valeurs de la loi normale. Pour un seuil de significativité de $P < 0,05$ en particulier, cette valeur est le fameux **1,96**. Si le t obtenu dépasse 1,96 alors il existera une liaison statistique "significative" entre X et Y. Rappel, la liaison entre la loi du R et du t de Student sous l'hypothèse Ho "R = 0" est :

$$t = [R/\sqrt{(1 - R^2)}]\times\sqrt{(n - 2)}$$

Ici on obtient **t = 2,0447...** qui est *supérieur* à 1,96 !

Enfer et putréfaction ! Le test est bel et bien *significatif* (même si c'est de justesse) au risque $\alpha = 0,05$! Dans un article scientifique on écrirait donc "*La corrélation entre la taille des époux est significative, (R = 0,0951, n = 460, P < 0,05)*". Comment un tel miracle est il possible ? Tout simplement parce que, avec beaucoup d'individus, il devient possible de déceler des liaisons même ténues entre X et Y. Cela ne veut pas dire que la liaison en question a un intérêt biologique quelconque. Pour s'en convaincre, il suffit de calculer R^2, le coefficient de détermination. ici, $R^2 = 0,009$ soit moins de 1%. Traduction, si on fixe la taille du père, la variance de la taille des mères possibles diminue seulement de 1% par rapport à la variance observée dans la population totale. Autrement dit, n'essayez pas de deviner la taille d'une femme à partir de la taille de son mari (ou vice versa), cette tentative est vouée à l'échec !

Précisons maintenant les choses en abordant, justement, la *précision* avec laquelle nous avons estimé la véritable valeur (inconnue à jamais) de Rho, le véritable coefficient de corrélation entre X et Y dans la population. Nous pourrons évidemment en déduire immédiatement la gamme de valeurs plausibles pour R^2. Il nous faut pour cela utiliser la transformation de Fisher décrite dans la section **12.9.** On commence par déterminer la variance de Z qui vaudra $1/(n - 3)$ soit $s^2_Z = 1/457 = 0,00219$ d'où l'écart type $s_Z = 0,0468$. Finalement :

$Z_{inf} = 0,0037138$ d'où (par la manoeuvre inverse vue plus haut) $R_{inf} = 0,0037137$
$Z_{sup} = 0,1874$ d'où $R_{sup} = 0,1849$

l'$IC_{95\%}$ de R est ainsi environ [0,004 — 0,185]
l'$IC_{95\%}$ de R^2 est donc environ [0,000014 — 0,034]

Décodons. Première constatation, la corrélation entre X et Y est *peut-être* nulle, mais elle est peut-être aussi substantiellement plus élevée que ce que nous avions cru : l'intervalle de confiance de R "monte" tout de même jusqu'à 0,185 ce qui ne serait pas ridicule en biologie. Quoi qu'il en soit, même cette valeur représenterait une liaison *très faible* entre X et Y, puisque le R^2 correspondant serait seulement de 0,034 (soit 3,4%). Ainsi, même si la liaison était "aussi forte" que R = 0,185, connaître la taille d'un des époux permettrait seulement de restreindre de 3,4% la variance de la gamme des valeurs vraisemblables pour l'autre. Il serait toujours aussi illusoire de prétendre deviner la taille d'un mari à partir de celle de sa femme, comme le nuage de données nous l'indiquait d'ailleurs fort clairement dès le départ ! Ce résultat témoignerait du fait que la taille du partenaire entre au mieux pour une *très petite* partie dans le complexe processus d'appariement qui permet de former les couples dans notre espèce. Comme le disent si bien les anglais "*Size isn't everything*".

12.10 Comparaison de deux coefficients de corrélation R_A et R_B

La situation est la suivante : deux échantillons A et B d'effectifs n_A et n_B, provenant de deux populations soupçonnées d'être différentes dans la force de la relation entre X et Y. Sur chacun de ces échantillons, vous avez calculé le coefficient de corrélation de Pearson, et avez donc obtenu deux valeurs qui sont R_A et R_B. Les valeurs obtenues sont évidemment différentes, mais la question est, comme d'habitude, sont elles *suffisamment* différentes pour que le hasard puisse difficilement expliquer cette différence ? Il vous faut donc comparer R_A et R_B pour déterminer si la différence que vous observez est statistiquement significative.

Cette comparaison est très facile grâce à la transformation du Fisher décrite ci-dessus. En effet, sous l'hypothèse Ho habituelle qu'il n'y a, en réalité, aucune différence entre les deux populations dont on veut comparer les coefficients de corrélation, on va avoir $Z_A = Z_B$, donc leur différence D est une loi normale de moyenne nulle : $D = Z_A - Z_B = 0$.

Par ailleurs, les variances ne se soustrayant jamais, la variance de D notée s^2_D est simplement l'addition des deux variances de Z_A et Z_B notées s^2_{ZA} et s^2_{ZB}, valant respectivement :

$$s^2_{ZA} = 1/(n_A - 3)$$
$$s^2_{ZB} = 1/(n_B - 3)$$

donc : $s^2_D = s^2_{ZA} + s^2_{ZB}$ d'où on calcule l'écart-type $s_D = \sqrt{\ s^2_D}$

La variable aléatoire D étant automatiquement "centrée" (elle vaut zéro) si Ho est vraie, il ne reste plus qu'à la réduire en la divisant par son écart-type pour obtenir une variable normale centrée-réduite Z_0:

$$Z_0 = (Z_A - Z_B) / s_D$$

Nantis d'une loi normale centrée réduite, nous sommes sauvés et il ne reste plus qu'à suivre une procédure qui vous est maintenant familière :

> **Si $|Z_0| > 1{,}96$ on rejette Ho. On conclut donc que les variables Z_A et Z_B, et donc les coefficients de corrélation R_A et R_B, sont significativement différents au risque α = 0,05 (la table de la loi normale vous permettra d'affiner la probabilité en fonction de la valeur de Z_0 observée, comme d'habitude**).

Si $|Z_0| \leq 1{,}96$ on ne rejette pas Ho. On conclut donc qu'on n'a pas suffisamment de preuves pour conclure à une différence entre R_A et R_B. Cette conclusion ne démontre en rien que Ho est vraie, notre conclusion étant associée à un risque β d'autant plus grand que l'écart entre R_A et R_B (s'il existe) est faible.

L'exemple qui suit va permettre de comparer la force de la liaison entre la taille d'un individu et la longueur de sa main, chez les garçons et les filles de la promotion 1998 de la maîtrise BPE. Cet exemple m'a été inspiré au cours de ma thèse, par une conversation entre étudiants comme il en existe dans tous les laboratoires pour se détendre entre deux manips, et où le sujet était tombé sur un grand classique : le sexe. Une des participantes à notre débat, qui avait apparemment une certaine expérience, nous avait alors affirmé mordicus qu'il était facile de deviner chez un homme quelle serait la longueur de ce fameux organe que les mâles ont et que les femmes n'ont pas, en regardant la longueur des mains du monsieur en question. Je lui avais alors répondu que ça me semblait assez évident dans la mesure ou *tous* nos organes (que ce soit la main, l'oreille, le foie ou cet organe si particulier) sont *à peu près* proportionnés à la taille du bonhomme, et que donc, elle aurait tout aussi bien pu prendre comme estimateur la hauteur du crâne ou beaucoup plus simplement la *taille* de l'homme "visé". Elle avait rétorqué que non, pas du tout, que les mains étaient un estimateur bien plus fiable. J'étais resté dubitatif. Quelques années plus tard, j'ai donc inclus innocemment dans le questionnaire proposé à mes étudiants une question sur la longueur de leur... main. Nous allons donc voir déjà avec quelle précision la taille d'un individu permet de déduire la longueur de sa main, organe osseux qui fait partie du squelette. En effet, la *taille* d'un individu est grosso modo un caractère osseux (c'est une combinaison de la longueur des os de nos jambes, de notre colonne vertébrale et celle de notre crâne). On peut donc émettre l'hypothèse (que je n'ai évidemment pas pu vérifier) que la liaison qui existe entre la taille de la main et l'organe sexuel mâle sera plutôt *moins bonne* que la relation qui existe entre la taille et la main, puisque le caractère osseux est absent chez l'organe en question (sauf chez les ours, comme chacun le sait). Les figures ci-dessous présentent le lien taille totale/taille de la main chez les garçons et les filles de la promotion 1998.

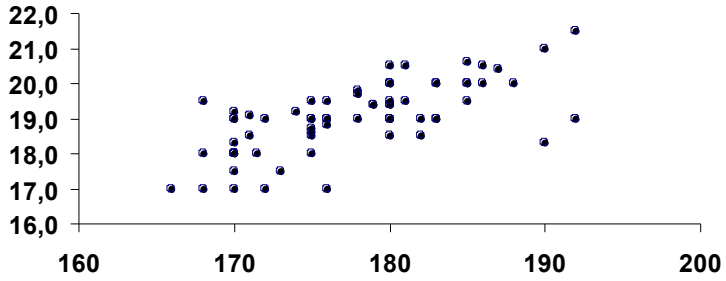


Fig 12.3 longueur de la main (cm) en fonction de la taille (cm) chez les 70 garçons de MBPE 1998

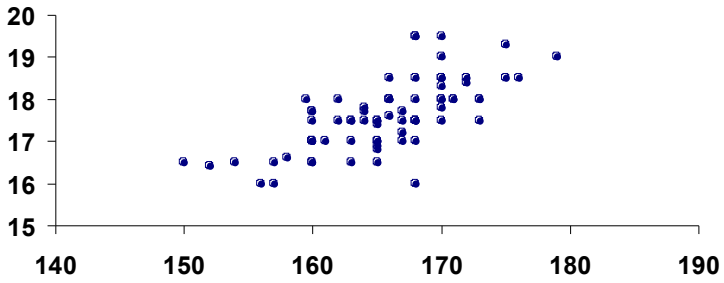


Fig 12.4 longueur de la main (cm) en fonction de la taille (cm) chez les 73 filles de MBPE 1998

Et maintenant, analysons ces données.

Exemple 12.2 comparaison des Corrélation taille (X)/longueur de main(Y)
Les données de base sont les suivantes :

Garçons :
$n_A = 70$ $m_X = 177{,}09$ cm et $s^2_X = 65{,}33$ d'où $s_X = 8{,}08$
$m_Y = 18{,}98$ cm et $s^2_Y = 1{,}27$ d'où $s_Y = 1{,}13$
Covariance (X,Y) = 5,57 $R_A = 0{,}61$ d'où $R^2_A = 0{,}37$

Filles :
$n_B = 73$ $m_X = 165{,}51$ cm et $s^2_X = 31{,}5$ d'où $s_X = 5{,}61$
$m_Y = 17{,}59$cm et $s^2_Y = 0{,}696$ d'où $s_Y = 0{,}834$
Covariance (X,Y) = 3,18 $R_B = 0{,}679$ = d'où $R^2_B = 0{,}461$

Commençons par le commencement : avant de comparer deux coefficients de corrélation, encore faut-il s'assurer qu'ils existent (autrement dit qu'ils sont significatifs). Aucun souci dans le cas présent, la table du R vous indique que pour $n = 70$ ddl un R est significatif pour $P < 0{,}01$ dès qu'il dépasse 0,30. Or, on est *largement* au dessus. La transformation vers une variable du t de Student vous indiquerait ici respectivement $t_A = 6{,}33$ et $t_B = 7{,}79$ soit, dans les deux cas, une probabilité inférieure à *une chance sur un milliard* d'observer un R si éloigné de zéro, simplement sous l'effet du hasard. Il est donc bien clair qu'il existe une liaison entre la taille d'un individu et celle de sa main, quel que soit le sexe, et ça n'a vraiment rien d'un scoop. Comparons maintenant les deux coefficients R_A et R_B obtenus. Il nous faut pour cela utiliser la transformation de Fisher :

$Z_A = 0{,}71$et $s^2_{ZA} = 1/(70 - 3) = 0{,}0149$ d'où $s_{ZA} = 0{,}122$
$Z_B = 0{,}83$ et $s^2_{ZB} = 1/(73 - 3) = 0{,}0143$ d'où $s_{ZB} = 0{,}119$

finalement $|Z_0| = |0{,}71 - 0{,}83|/\sqrt{(0{,}0149+0{,}0143)} = 0{,}70$

soit $0{,}48 < P < 0{,}49$ d'après la table de la loi normale. Donc, si les coefficients R_A et R_B étaient identiques dans la réalité, on observerait sous l'effet du hasard un écart plus grand que celui que nous voyons 5 *fois sur 10* en utilisant des échantillons de cette taille. Il n'y a vraiment *aucune raison* de soupçonner que la force de la liaison statistique entre la taille et la longueur de la main soit différente chez les garçons et les filles.

Ayant calculé Z_A, Z_B, s_{ZA} et s_{ZB} il est maintenant facile de calculer les intervalles de confiance autour de R_A, R_B, R^2_A et R_{2B} comme déjà décrit dans l'exemple **12.1**. Résultat :

IC_{95}de R_A [0,44 — 0,74] IC_{95} de R^2_A [0,19 — 0,55]
IC_{95}de R_B [0,53 — 0,79] IC_{95} de R^2_B [0,28 — 0,62]

Nous constatons à présent que notre estimation des R n'est pas d'une très grande précision (avec des valeurs vraisemblables qui varient presque du simple au double). Quant aux R^2, ils nous indiquent que la fraction de la variance de Y qui disparaît si on fixe X pourrait être aussi faible que 19% et ne dépasse très probablement pas 62%. Autrement dit, prétendre deviner précisément

la longueur de la main d'un individu en connaissant sa taille reste assez aventureux, comme la dispersion des données sur les figures **12.3** et **12.4** nous le montrait déjà. Je vous laisse en tirer les conclusions qui s'imposent concernant la "fiabilité" des estimations que vous seriez tentés de faire en partant de la longueur de la main d'un individu pour deviner la taille d'un autre de ses organes n'ayant aucun rapport avec le squelette...

12.11 Un cas particulier utilisant le R de Pearson: la droite d'allométrie ou droite de Tessier.

12.11.1 Calcul de la droite de Tessier

L'allométrie est le rapport des proportions des mesures biométriques que l'on peut effectuer sur un organisme vivant (par exemple le rapport longueur/masse d'un animal, ou bien le ratio longueur/largeur d'une feuille d'arbre). Ce genre d'étude peut être utilisé aussi bien entre espèces qu'à l'intérieur d'une même espèce, en particulier pour étudier la variation des proportions de différentes parties du corps au cours de la croissance. Il est bien connu dans l'espèce humaine que le rapport taille de la tête/taille du corps varie de manière spectaculaire au cours de la croissance : la tête représente pratiquement la moitié de la longueur du corps d'un foetus à un certain stade, mais seulement environ 1/7ème à 1/8ème de la longueur du corps à l'âge adulte.

Les relations d'allométrie peuvent se modéliser au moyen de courbes d'équation générale :

$$Y = B\,X^{a}$$

avec a et B des constantes propres à l'espèce
et aux dimensions Y et X considérées.

Comme il est beaucoup plus pratique de manipuler des droites, on utilise les logarithmes pour transformer cette fonction en droite, obtenant ainsi :

$$\log(Y) = a\,\log(X) + \log(B)$$

Qui est bien une équation de droite, comme on le fait apparaître encore plus clairement en changeant quelques noms : $\log(Y) = y$, $\log(X) = x$ et $\log(B) = b$, d'où finalement :

$$y = a\,x + b$$

Cette droite se nomme **droite d'allométrie** ou droite de Tessier.

Si on trace une droite D quelconque, passant par le centre de gravité du nuage de points, on peut mesurer la distance d_y entre un point de données et cette droite D, parallèlement à l'axe des Y. C'est ce qu'on fait dans le cas d'une étude de régression. Cependant, si les deux variables X et Y sont considérées de manière égale, il est tout aussi logique de mesurer la distance d_x séparant le point et la droite D, mais cette fois parallèlement à l'axe des X. Pour tenir compte de ces deux distances, on va simplement les multiplier l'une par l'autre (sans tenir

compte de leur signe : on considère les valeurs absolues) et obtenir une distance combinée $d_x d_y$ pour chaque point.

La droite d'allométrie est la droite qui minimise la somme de ces distances combinées $d_x d_y$ pour l'ensemble des points du nuage.

Elle a pour coefficient directeur *a*, vu plus haut, qui se trouve être égal au ratio entre les écarts types concernant *x* et *y* (donc calculé sur les données transformées par le passage au log et non pas sur les données originelles)

$$a = s_y / s_x$$

Cependant, le *a* calculé sur nos modestes données n'est qu'une *estimation* du véritable coefficient α reliant *x* et *y* dans la réalité. Ce n'est donc qu'une variable aléatoire comme une autre, possédant une variance, que l'on sait heureusement calculer. La variance de ce coefficient directeur est :

$$s^2_a = a^2 \, [(1 - R^2) / (n - 2)]$$

avec

R: le coefficient de corrélation entre *x* et *y* les données transformées
n: le nombre de couple de données (*x*,*y*)

Vous aurez noté que le R de Pearson (sous la forme du coefficient de détermination R^2) intervient dans cette équation, et il est important de le calculer sur les données transformées (le nuage de points de coordonnées log(X), log(Y)) car si la relation entre X et Y n'est pas une droite (ce qui sera le cas si jamais le coefficient a n'est pas égal à 1) le calcul de R sur les données originelles serait faussé par la courbure du nuage.

12.11.2 Calcul de l'intervalle de confiance de *a*

La variance de *a* (notée s^2_a) et donc son écart-type s_a étant connus, et sachant que *a* converge rapidement vers une loi du t de Student, puis une loi normale, l'intervalle de confiance à 95% de *a* se calcule à partir des valeurs seuil du **t** de Student pour $\alpha = 0{,}05$ et en délimitant un intervalle de $\pm$ **t** s_a autour de lui. Par exemple, pour 20 valeurs, donc 18 ddl, la valeur seuil du t de Student pour $\alpha = 0{,}05$ à utiliser est **2,101.**

$$\text{l'IC}_{95} \text{ serait : } [a - 2{,}101 \; s_a \text{ —— } a + 2{,}101 \; s_a]$$

Dès que vous avez plus d'une trentaine de valeurs, vous pourrez utiliser la valeur 1,96 de la loi normale, au lieu de la valeur seuil de la table du t de Student.

12.11.3 Comparaison de deux droites d'allométries

Comment comparer les pentes a_1 *et* a_2 mesurées sur deux "populations" (par exemple, les mâles et les femelles capturés sur un estran) ?

On utilise pour cela la sempiternelle manœuvre de centrage réduction, avec comme hypothèse Ho : "les deux populations comparées ont en réalité la même pente α".

Variable aléatoire	sa moyenne sous Ho	sa variance
a_1	α	s^2_{a1}
a_2	α	s^2_{a2}
$a_1 - a_2$	0	$s^2_{a1} + s^2_{a2}$
$(a_1 - a_2)$ /racine $(s^2_{a1} + s^2_{a2})$	0	1

La variable centrée réduite suit théoriquement une loi du t de Student à $n_1 + n_2 - 4$ ddl. Cependant, cette loi converge très rapidement vers la loi normale centrée réduite donc, à moins que vous n'ayez vraiment un petit échantillon, le tableau ci-dessus revient en fait à faire un test Z et à juger par rapport à la valeur 1,96 pour un risque α de 5% :

Si | Z | > 1,96, on conclut que la relation entre X et Y diffère entre vos deux populations, puisque les pentes sont significativement différentes au risque α = 0,05 (donc, *P* < 0,05). La table de la loi normale vous permettra cependant de préciser α, et vous serez alors peut être en mesure d'écrire que *P* < 0,01 voire *P* < 0,001.

Si | Z | < 1,96, on n'a pas suffisamment d'éléments pour rejeter avec confiance l'hypothèse Ho selon laquelle les pentes seraient identiques (mais il n'y a aucun moyen d'aller plus loin, vous ne connaissez pas le risque β), et on conclut donc que la différence entre les pentes observées *n'est pas significative* (ce qui ne veut *pas* dire "je suis certain qu'il n'y a absolument aucune différence dans la réalité").

12.12 Comment étudier une corrélation quand on n'est pas dans les conditions d'utilisation du coefficient de corrélation R de Pearson ?

Le coefficient de corrélation de Pearson est utilisable dans des conditions qui varient de manière surprenante d'un ouvrage à l'autre, mais le minimum requis semble être que la distribution d'une des deux variables soit normale, quand on fixe l'autre (pour un X fixé, les Y doivent être distribués normalement). Si cette condition n'est manifestement pas remplie, on doit utiliser un autre coefficient de corrélation non paramétrique basé, non sur les valeurs, mais encore sur les rangs, le coefficient de corrélation de Spearman.

Coefficient de corrélation de Spearman.

On a deux variables X et Y avec chacune *n* individus ind_1 à ind_n, caractérisé par un couple de valeurs observées (x,y). On classe chaque individu selon sa valeur pour X (donc on obtient un rang r_x) et sa valeur Y (on obtient un autre rang r_y). Bref, chaque individu ind_i sera donc caractérisé par un couple de rangs (r_x, r_y). La suite consiste à calculer le coefficient de corrélation paramétrique habituel (selon la formule de Pearson) **mais** sur ces nouvelles

variables aléatoires « rangs ». Dans ces conditions où seules les valeurs entières sont possibles, la formule se « simplifie » (si on peut dire) en :

$$r' = 1 - 6 \times [\ (r_{x1} - r_{y1})^2 + (r_{x2} - r_{y2})^2 + \ldots + (r_{xn} - r_{yn})^2] \ / \ n(n^2 - 1)$$

soit sous une forme compacte en appelant d_i la différence $(rx_i - ry_i)$

$$\mathbf{r' = 1 - 6\Sigma d_i^2 \ / \ n(n^2 - 1)}$$

Pour les tout petits effectifs (jusqu'à $n = 10$), il existe une table du **r'** et c'est celle là qu'il faut utiliser. Au delà de $n = 10$, la table du r « paramétrique » donne une valeur suffisamment approchée de r' et on lit directement dans la table du r « paramétrique ». Toute la discussion concernant la corrélation et la causalité reste évidemment valable.

ICI

(mais seulement dans un certain temps)
se dressera fièrement le 13ème (et probablement dernier) chapitre
de *Statistiques pour Statophobes*.

Il se nommera "Faites des progrès en régression" et traitera, bien entendu, de la régression linéaire.

Vu la très faible probabilité qu'il apparaisse dans un délai raisonnable (puisque ça fait quatre ans que je me dis "tiens, il faudrait vraiment que je l'écrive"), je vous conseille cependant de ne surtout pas l'attendre, et de consulter les innombrables ouvrages expliquant cette technique, par exemple Schwartz 1994 (4ème ed), *Méthodes statistiques à l'usage des médecins et des biologistes.*

Denis POINSOT, 29 février 2008

Epilogue

Le Professeur Giuseppe Parsimoni me reçoit dans son antique villa, perchée sur les hauteurs de Chevapiano. Il fait très doux en cette fin d'été et les cigales vont bientôt se taire pour la nuit. Je trouve sur la terrasse un petit bonhomme au menton napoléonien et aux yeux noirs incroyablement vifs, tranquillement installé devant une table garnie de trois énormes assiettes fumantes de penne all' arrabiata. Le soleil couchant fait flamboyer ses cheveux blancs un peu en désordre. La vue sur la baie de Valontano est superbe.

DP — *Professore*, vous me faites vraiment un grand honneur de me recevoir chez vous ! Il y a encore des tas de sujets sur lesquels je rêve de vous interroger ! (*un très long silence s'installe*) heu... vous ne dites rien ?
Parsimoni — *Santa Madonna* ! Discutailler alors que *la pasta* est servie et commence à refroidir ? Il n'y a donc rien de sacré pour vous ?
DP — Je suis absolument confus !
Parsimoni — Allez plutôt ouvrir la porte, je vois ce chenapan de Tigran qui monte le chemin. Ce gamin n'a jamais été capable d'arriver à l'heure. Je parie qu'il va encore essayer de me tuer à coup de cholestérol avec ses oeufs de poisson.

Quelques instants plus tard, la formidable silhouette de Tigran Abonessian, précédée d'une non moins formidable barbe noire, se découpe dans l'embrasure de la porte. Il tient négligemment en équilibre trois énormes boites de caviar Beluga.

Parcimoni — Tiens, qu'est ce que je vous disais !

La voix de Tigran éclate dans la pièce comme un coup de canon

Abonessian — *CIAO PEPE* ! Ah, bonsoir *dottore* Poinsot. Bon voyage ? Tiens... je vous imaginais beaucoup plus grand !
Parsimoni — Ne faites pas attention, Tigran a toujours été un ours.
Abonessian *(hilare)* **—** *Pepe* sait de quoi il parle, c'est lui qui m'a élevé.
DP — Elevé ?
Abonessian — J'ai fait une grande partie de mes études à Chevapiano. Giuseppe était mon directeur de thèse. Vous l'ignoriez ?
DP — C'est-à-dire... vous ne semblez pas toujours sur la même longueur d'onde en matière de statistiques.
Abonessian — Seulement sur certains détails. Giuseppe m'a tout appris. Disons simplement que j'ai continué à apprendre par la suite.
Parsimoni — Mais rien de bon, je le crains. Allez, pose ces boîtes, tu vas assommer quelqu'un**.**
Abonessian — Ah, mais cette fois, il n'y en a que deux pour vous *Maestro*. La troisième est pour le *dottore francese* (*il me tend une des boîtes de caviar. Elle fait au moins un kilo*) Vous aimez ?
DP — Je vous le dirai quand je saurais le goût que ça a. Merci beaucoup, en tout cas !
Abonessian (*abasourdi*) **—** Vous ne mangez jamais de caviar Beluga ? mais c'est pourtant très bon !

DP — Nous discuterons du salaire des universitaires français une autre fois, si vous le voulez bien. Maintenant que vous êtes là, je voudrais votre avis à tous les deux sur la meilleure manière de clore ce petit livre.
Parsimoni — Alors c'est très simple. Il vous suffit de rappeler avec force qu'on n'a rien inventé de valable depuis les intervalles de confiance.
Abonessian — Qui sont aux statistiques modernes ce que la massue est au rayon laser.
Parsimoni (*levant les yeux au ciel*) — Seigneur, pardonnez-lui, ce garçon a besoin d'un ordinateur même pour écrire son courrier.
DP — Messieurs, s'il vous plaît, de jeunes étudiants désorientés nous lisent en ce moment même. Ils ont peut-être découvert à la lecture de ce petit opuscule — et grâce à vos interventions pugnaces — que les tests statistiques n'étaient pas forcément ce monument de marbre qu'ils imaginaient, que certains aspects étaient même assez polémiques. Ils savent pourtant qu'ils auront impérativement à utiliser les statistiques eux-mêmes. Que leur dire pour les aider à aller plus loin ? Pour calmer leurs angoisses légitimes face à toute cette incertitude ? Comment faire en sorte que des mots tels que ANOVA, analyse multivariée, modèle linéaire généralisé, ne les fassent pas fuir à toutes jambes ?
Parsimoni — D'abord, rappelez-leur bien que les statistiques ne sont qu'une petite partie de la science, et que l'écrasante majorité des découvertes majeures, de la circulation du sang, à la gravitation, en passant par les ondes radio, la théorie de l'évolution des espèces ou la structure de l'ADN s'en sont très bien passées. De nos jours encore, les mathématiciens, les physiciens et les chimistes vivent très bien sans elles et font de l'excellente science.
Abonessian — Je m'excuse de rappeler que la physique quantique repose sur des équations probabilistes et qu'elle représente une formidable avancée conceptuelle.
Parsimoni (sarcastique) **—** Expliquer qu'un chat peut être à la fois vivant et mort est effectivement une formidable avancée conceptuelle. Ce Schrödinger était vraiment un sacré farceur.
Abonessian — Vous ne pouvez tout de même pas nier l'apport majeur des statistiques en agronomie, en biologie en général et en médecine en particulier. Peut-on concevoir aujourd'hui un essai clinique sans analyse statistique ?
Parsimoni — Non, mais on peut tout à fait concevoir un médecin qui dit à son malade qu'il n'a pas mal au genou parce qu'on ne voit rien sur le scanner. Les scientifiques modernes s'éloignent du concret d'une manière qui m'effraie parfois.
DP — Soyons concrets, alors. Faut-il vraiment faire des statistiques en biologie, comme je le prétends dans le premier chapitre ?
Parsimoni — Bien entendu, mais il faut bien reconnaître que des pans entiers de la biologie échappent sans problèmes aux statistiques. Un biologiste moléculaire ne vous dira pas que l'enzyme QbumIII coupe tel ou tel plasmide "significativement". Elle le coupe ou elle ne le coupe pas. De même la description d'une nouvelle espèce, élément vital de l'étude de la biodiversité, s'effectue sans le moindre petit test. Il ne faut pas confondre "science" et "analyse statistique".
Abonessian — D'un autre côté, des domaines biologiques sortis du néant, comme la génomique (le décryptage des génomes et leur étude à grande échelle) ont amené une masse tellement colossale d'information dans les bases de données, qu'elles ont stimulé la création de méthodes d'analyses statistiques entièrement nouvelles. Biologie moléculaire ne signifie pas non plus "pas de stats".

DP — J'aimerais avoir votre avis sur l'utilisation de l'ANOVA pour comparer simultanément plusieurs moyennes. C'est un sujet assez austère en général pour les débutants.
Abonessian (*malicieux*) — Je note d'ailleurs que vous l'avez soigneusement évité dans cet ouvrage.
DP — J'avoue en bloc. Je ne maîtrise pas suffisamment les finesses de l'ANOVA pour me risquer sans crainte sur son territoire.
Parsimoni — Tigran vous taquine, mais votre prudence vous honore. L'ANOVA fait bien suffisamment de ravages comme ça, quand elle est expliquée par des gens qui la "maîtrisent".
DP — C'est tout de même une méthode extrêmement utilisée.
Parsimoni — Bien entendu. L'inertie du monde scientifique moderne est sans équivalent. Quatre-vingt ans après avoir pointé du doigt les travers des tests d'hypothèse nulle, on continue d'en faire des milliers chaque année, dont la plupart sont superflus et peu informatifs.
Abonessian — Dont *certains* sont superflus et peu informatifs. L'ANOVA est une méthode éprouvée, basée sur des fondements théoriques solides et sur laquelle on a accumulé une très grande expérience. Elle a rendu de fiers services et continuera à le faire, Giuseppe, reconnaissez-le.
Parsimoni — Ah oui, vraiment, quel fier service de nous dire "Quelque part parmi ces quinze moyennes, il y en a au moins une qui n'est pas comme les autres, $P < 0,05$". La belle affaire que voilà ! Je sais bien qu'au moins une de mes quinze moyennes ne va pas être comme les autres. C'est le contraire qui serait miraculeux ! Mais si c'est seulement au niveau du deuxième chiffre après la virgule, je m'en moque complètement.
DP — Que faudrait-il faire alors ?
Abonessian (*plissant les yeux, comme frappé par une vision subite*) — Etant un peu médium, je sens confusément que les mots "intervalle de confiance" ne vont pas tarder à être prononcés.
Parsimoni — Mais parfaitement ! Nous voulons savoir *à quel point* ces moyennes sont différentes les unes des autres, nous voulons savoir *lesquelles* se détachent du lot *et de combien*, c'est la seule question scientifiquement valide, et cela implique — efface ce sourire de nigaud, Tigran — de calculer leurs intervalles de confiance.
Abonessian — Il existe pour cela des tests a posteriori, tel le test de Tukey, que l'on effectue après l'ANOVA, lorsqu'elle est significative.
Parsimoni — Et on ne fait rien lorsqu'elle ne l'est pas, ce qui équivaut à jeter à la poubelle le peu d'information contenue dans des données parfois durement acquises. Quel gâchis !
DP — Mais... si l'ANOVA est non significative, je ne vois pas comment...
Parsimoni — Cela indique simplement qu'on connaît les moyennes très approximativement, et voilà tout. Ca n'a jamais empêché personne de calculer un intervalle de confiance. Pensez- vous rationnel de conduire en fermant les yeux, simplement parce que la visibilité est mauvaise ?
DP — Vu sous cet angle, évidemment... je suppose que l'avantage de ce système est qu'il s'affranchit des conditions d'application assez strictes de l'ANOVA ?
Abonessian — Beaucoup moins strictes qu'on veut bien le dire. L'ANOVA est assez robuste. De plus, traditionnellement, lorsque les conditions d'application de l'ANOVA ne sont pas vraiment remplies, on effectue une transformation de variable, ou, en dernier

recours, on utilise un équivalent non paramétrique comme le H de Kruskal-Wallis dont vous parlez au chapitre 10. Cependant, une approche plus *moderne*...
Parsimoni — Préparons-nous à un déluge de jargon prétentieux.
Abonessian — ... consiste à utiliser le modèle linéaire généralisé (GLM), qui repose sur la théorie de l'information et donc sur le concept de maximum de vraisemblance.
Parsimoni — J'ai connu personnellement plusieurs calmars qui lançaient des nuages d'encre beaucoup moins opaques.
Abonessian — Giuseppe, allons, cela consiste simplement à coller à la véritable distribution des données au lieu de vouloir à toute force la faire rentrer au chausse-pied dans le moule de la loi normale. Ce genre d'approche devrait pourtant vous plaire !
Parsimoni — Malgré ce bel habillage, cela reste un test d'hypothèse nulle de type "pas d'effet".
Abonessian — Exact, mais il est très puissant.
Parsimoni — Et donc particulièrement dangereux, car il permet de monter en épingle des différences minuscules, sans aucun intérêt pratique.
Abonessian — Dans ce cas, il ne faut pas accuser le test, mais éduquer celui qui l'utilise. Un test ne fait que vous donner une probabilité.
Parsimoni — Je ne te le fais pas dire. Alors qu'un intervalle de confiance te donne du concret : la zone d'ombre dans laquelle se cache vraisemblablement la vérité que tu cherches, même si nous devons nous contenter de cette ombre, tels les hommes de la caverne dont parle Platon.
Abonessian — Changez vite le sujet, *dottore*, sinon *pepe* va nous faire une citation en grec ancien .
DP — Le *professore* Parsimoni est donc un fin connaisseur de Platon ?
Abonessian (*angélique*) — Bien entendu : ils sont de la même génération.
Parsimoni — C'est ça, fais le malin. Tu veux que je raconte la fois où tu m'as écrit une pleine page de discussion pour commenter un pourcentage calculé sur douze individus ?
Abonessian — *pepe*, j'avais vingt ans ! Il y a prescription !
Parsimoni (le visage de marbre) — Certains crimes sont imprescriptibles.

(S'ensuit un dîner délicieux)

DP — Messieurs, le mot de la fin ?
Parsimoni — Eh bien... surtout n'y voyez aucune offense, *dottore*... mais je pense vraiment que les statistiques ne s'apprennent pas dans les livres.
DP — Nulle offense *professore*, vous prêchez un convaincu !
Parsimoni — A la bonne heure. Dites donc bien à vos étudiants qu'ils devront encore se colleter avec des milliers et des milliers de données réelles avant de commencer à se sentir vraiment à l'aise. Je leur dirai ensuite ceci : aimez vos données, chérissez les. Traitez les comme elles le méritent : avec douceur et lenteur. Comment dites-vous en français ? "*Plus fait patience et longueur de temps*..."
DP — Je vois ce que vous voulez dire. Et vous, professeur Abonessian, quel dernier conseil donneriez-vous à un étudiant qui veut apprivoiser les stats ?
Abonessian — Je lui dirais la même chose, et je lui dirais surtout qu'il a l'immense chance d'être né à l'époque des ordinateurs...
Parsimoni — J'aurai vraiment tout entendu dans ma longue vie.
Abonessian — ... parce que grâce à l'informatique il n'a jamais été aussi facile et ludique de s'entraîner par simulation pour vraiment saisir de quoi l'aléatoire est capable.

On trouve sur internet des dizaines de sites consacrés aux stats, des encyclopédies et des cours en ligne. Enfin je dirai à vos étudiants : allez donc sur http://the-r-project.org, téléchargez le programme professionnel gratuit R, une bonne fois pour toutes, et apprenez tranquillement à vous en servir, à votre rythme. Vous ne devriez plus jamais avoir besoin d'utiliser grand chose d'autre.
DP — Et bien il me reste à vous remercier tous les deux, j'ai passé une excellente soirée.
Parsimoni — Tout le plaisir est pour moi. Sinon, j'aurais dû rester en tête à tête avec Tigran.
DP — C'est donc si terrible ?
Parsimoni — C'est bien simple : il ne parle que de statistiques !
Abonessian — C'est un mensonge éhonté. Mais la soirée n'est pas encore finie pour vous *pepe*, je vous ai apporté des diapos sur ma dernière partie de pêche à l'esturgeon.
Parsimoni — Surtout pensez bien à emporter vos oeufs de poisson *dottore*, j'en ai plein mes placards.
Abonessian (*moqueur*) — Tiens donc. Il m'a pourtant semblé tout à l'heure qu'il ne restait pas grand chose de la cargaison apportée lors de mon dernier voyage.
Parsimoni (*sérieux comme un pape*) — Je les donne à mes chats. Les chats aiment les oeufs de poisson.

Je m'éclipse sans faire de bruit. Le chemin du retour est obscur et sent bon le romarin et la mer. Soudain, je me fige : deux yeux blancs me fixent dans la nuit. Une fraction de seconde plus tard, il n'en reste que l'impression sur ma rétine. Je viens sûrement de vivre un moment rare, la rencontre avec un chat quantique. Dans le lointain on entend comme un roulement de tonnerre, mais je sais que c'est seulement le rire de Tigran.

F I N

ANNEXE 1 : Estimation s^2

Pourquoi faut il utiliser :

$$s^2 = [\sum (x_i - m)^2] / (n -1)$$

et non pas

$$s^2 = [\sum (x_i - m)^2] / n$$

dans l'estimation de σ^2 à partir d'un échantillon ?

Le problème vient du fait que *m* n'est pas la véritable valeur de μ, mais seulement son estimation basée sur les données de l'échantillon. Or, par définition, la moyenne d'une série de données est la valeur qui *minimise* les écarts entre les données et cette valeur. Ceci reste vrai pour les *carrés* des écarts. Il s'ensuit que le terme $\sum (x_i - m)^2$ est *le plus petit possible* avec les valeurs x_i de l'échantillon. Si on avait pu disposer de la *véritable* valeur de μ, le terme $\sum (x_i - \mu)^2$ aurait forcément été plus grand (puisque μ est différent de m). Bref, procéder en utilisant *m* (la seule valeur dont on dispose en pratique !) amène à un s^2 qui va *systématiquement sous estimer* σ^2.

Pour contrebalancer cet effet, il faut donc augmenter le numérateur (ou diminuer le dénominateur). Il a été démontré (brillamment je suppose) que la meilleure façon possible était de remplacer *n* par (*n - 1*) au dénominateur. Cette façon de procéder a pour avantage que l'effet est sensible pour les petites valeurs de *n* (pour lesquelles l'écart entre *m* et μ est probablement grand, donc la sous estimation importante) et il devient négligeable lorsque *n* grandit (la correction étant alors moins nécéssaire car l'estimation de μ par *m* devient de plus en plus fiable).

ANNEXE

Pourquoi Var(aX) = a^2 Var(X) et non pas a Var (X)

Si X est une variable aléatoire suivant une loi quelconque de moyenne μ_X et de variance σ^2_X. Pour une taille de population N, on a par définition :

$$\mu_X = \sum x/N \qquad \text{et} \qquad \sigma^2_X = (\sum(x - \mu_X)^2)/N$$

Si on pose Y = aX (ce qui revient à remplacer chaque résultat x par ax), on aura donc pour la nouvelle variable aléatoire Y :

$$\mu_Y = \sum ax/N \qquad \text{et} \qquad \sigma^2_Y = (\sum(ax - \mu_Y)^2)/N$$

En remplaçant μ_Y par son expression complète, on obtient :

$$\sigma^2_Y = [\sum(ax - \sum ax/N)^2]/N$$

On peut alors remarquer que l'expression élevée au carré contient des termes qui sont tous multipliés par a. Si on met la constante a en facteur, on obtient successivement (les étapes sont détaillées au maximum volontairement pour qu'on puisse bien suivre le mouvement):

$$\sigma^2_Y = \sum[a (x - \sum x/N)]^2/N$$

$$\sigma^2_Y = \sum a^2 (x - \sum x/N)^2/N$$

Tous les termes du premier $\sum$ sont multipliés par a^2, donc on peut mettre a^2 en facteur :

$$\sigma^2_Y = a^2 \sum(x - \sum x/N)^2/N$$

Or, $\sum x/N$ n'est autre que μ_X, donc :

$$\sigma^2_Y = a^2 \ \sum(x - \mu_X)^2/N = a^2 \ \sigma^2_X$$

On a bien Var(Y) = a^2 Var(X), ce qu'il fallait démontrer. Ce résutat provient, comme on le voit, de la simple application des définitions de la moyenne et de la variance (et le contraire aurait été plutôt inquiétant !). Notez (maintenant que je vous ai asséné cette belle démonstration avec plein de lettres grecques) qu'on aurait pu prévoir le résultat intuitivement sans se fatiguer: il suffisait pour cela de se souvenir que la variance utilise une unité (peu parlante) qui a la dimension du *carré* de celle utilisée pour la variable (des mm^2 pour une longueur en millimètres etc...). Si votre variable prend des valeurs *a* fois plus grandes, il est donc parfaitement normal que sa variance soit a^2 fois plus grande...

ANNEXE 2 : D'OÙ VIENNENT LES FORMULES DES LOIS BINOMIALES ?

Formule de la loi binomiale positive.

Soient deux types d'individus, type A (fréquence p), type B (fréquence $q = 1 - p$) et n tirages aléatoires indépendants. Les n tirages peuvent être conçus comme deux ensembles de cases, l'un de k cases où un individu A a été obtenu, et de (n - k) cases où un individu B a été obtenu. Cependant, ces n cases sont susceptibles d'être obtenues de bien des façons différentes. par exemple, si n = 3 et k = 2 on aura AAB ou bien ABA ou encore BAA = 3 possibilités. Ce genre de situation correspond à une combinaison (au sens probabiliste du terme) dont le nombre est de C_n^k (dans l'exemple choisi on a bien $C_3^2 = 3$). Il faut garder ce facteur multiplicatif en mémoire, il réapparaît plus loin. Reste maintenant à calculer la probabilité que k cases contiennent un A *et* les (n - k) autres cases un B.

(i) La probabilité que k cases prises au hasard contiennent un A est égale à : p^k
(ii) La probabilité que n-k cases prises au hasard contiennent chacune un B est égale à : q^{n-k}

Ces deux probabilités sont totalement indépendantes (le fait d'avoir un A dans une case donnée est sans aucune influence sur le contenu de la case d'à coté). La probabilité d'avoir (i) **ET** (ii) s'obtient donc en multipliant leurs probabilités, soit

P(k cases avec A ***et*** (n - k) cases avec B) = $p^k q^{n-k}$

L'important est ici de se souvenir que, lorsqu'on effectue n tirages, *il y a* C_n^k *combinaisons ayant cette probabilité*, puisqu'il y a C_n^k façons de ranger k éléments équivalents parmi n places (c'est ce qui a été calculé au début du raisonnement). On en déduit que la probabilité d'obtenir k évènements de probabilité constante p au cours de n tirages est :

$$P(X = k) = C_n^k \; p^k \; q^{n-k}$$

Formule de la loi binomiale négative.

Rappel : on veut r individus se trouvant en fréquence p dans la population où on effectue les tirages et on s'intéresse à la variable X « nombre de tirages nécessaires pour obtenir le $r^{ième}$ individu désiré».

Au cours des k – 1 premiers tirages sont apparus les r – 1 premiers individus qui nous intéressent (ils ont pu apparaître très tôt ou très tard dans cette série, mais maintenant ils sont là). Or, la probabilité d'obtenir Z = r – 1 individus de fréquence p au cours d'une expérience de n = k – 1 tirages suit une loi binomiale *positive* B (n : p). On sait que cette probabilité vaut

$P(X = Z) = C_n^Z \, p^Z \, q^{n-Z}$ (voir raisonnement de la loi binomiale positive)

Soit en remplaçant n et Z par leur valeur ici, une probabilité de $C_{k-1}^{r-1} \, p^{r-1} \, q^{k-r}$

Il nous faut *d'autre part* que le $k^{ième}$ tirage nous amène un individu qui nous intéresse pour en avoir r, et cette probabilité est p.

On a donc bien la probabilité totale : $P(X = k) = C_{k-1}^{r-1} \, p^{r-1} \, q^{k-r} \times p = \mathbf{C_{k-1}^{r-1} \, p^r \, q^{k-r}}$

Annexe 3 : L'erreur standard pour les débutants.

L'erreur standard (abréviation : e.s. en français et s.e. en anglais) est simplement le nom spécial donné à **l'écart-type d'un paramètre (moyenne, pourcentage, indice de Shannon etc.) calculé à partir de vos données.** Prenons l'exemple le plus courant, l'écart-type **de la moyenne**, et voyons en quoi cette "erreur standard" est différente du simple "écart-type des données" (racine carrée de la variance) qui est calculé automatiquement par la touche σ des calculatrices statistiques.

Erreur standard d'une moyenne

L'erreur standard de la moyenne **dépend du nombre *n* de données dans l'échantillon**. Plus l'échantillon est **grand**, plus l'erreur standard est **petite.** Elle traduit donc la *précision* dans l'estimation de la moyenne réelle dans la population. C'est pour cela qu'on utilise habituellement l'erreur-standard pour tracer les "barres d'erreur" sur les graphes scientifiques. Plus spécifiquement, la taille de l'erreur standard est *inversement proportionnelle à la racine carrée du nombre n de données*. En pratique, si on multiplie courageusement la taille de son échantillon par 10 au moyen d'un effort expérimental exténuant, l'erreur standard diminuera "seulement" dans la proportion √10 (donc l'estimation sera environ *trois* fois plus précise, et non pas — hélas — *dix* fois plus précise). Bref, dans le cas de la moyenne, le nom "erreur standard" pourrait avantageusement être remplacé par "écart-type-**de-la moyenne**". Alors pourquoi maintenir l'appellation "erreur standard" ?

Probablement parce que le terme "erreur standard" permet d'éviter la confusion avec l'écart-type *des données* (égal comme vous le savez à la racine carrée de la variance des données). Cet écart-type estime la dispersion des données autour de la moyenne dans la population échantillonnée. Or, le fait que les données soient *peu* ou *beaucou*p dispersées autour de leur moyenne, dans la population échantillonnée, **ne dépend évidemment pas du nombre *n* de données dans votre échantillon.** Au contraire, on a vu que l'erreur standard de la moyenne calculée à partit de votre échantillon diminue mécaniquement lorsque la taille de l'échantillon augmente. Il s'agit donc bien de deux notions différentes.

Pour résumer : lorsque la taille de votre échantillon grandit, *l'erreur standard* de la moyenne diminue (un effectif plus grand permet une estimation plus précise de la moyenne), mais la *dispersion* des données autour de leur moyenne dans la population d'origine (estimée par la variance et *l'écart-type de vos données*) reste immuable.

On peut exprimer tout ceci en deux formules très simples:

Ecart-type (d'une variable aléatoire au sein d'une population d'individus) :
σ
(immuable)

Erreur standard (d'une moyenne calculée sur *n* individus) :
$\sigma/\sqrt{n} = \sqrt{(\sigma^2/n)}$
(diminue lorsque *n* augmente)

Et maintenant, revenons à l'objet de cette Annexe 3, qui est de comprendre *comment* la moyenne d'un échantillon, qui est à première vue une valeur *unique*, peut avoir un écart-type (nommé erreur standard), puisque pour calculer un écart-type il faudrait *plusieurs* valeurs de moyennes. Où sont donc les *autres* valeurs qui permettraient de calculer cet écart-type? L'explication est tout simplement que la moyenne dont on calcule l'écart-type n'est pas la moyenne de *votre* modeste échantillon de *n* individus. Il s'agit de l'écart-type d'une variable aléatoire que l'on pourrait nommer : "*moyenne d'un échantillon de n individus tirés au hasard dans la population étudiée*". En effet, rappelons que l'échantillon sert seulement à accéder à une meilleure connaissance de notre véritable objet d'étude : la population dont il est issu. Donc, la moyenne calculée sur *votre* échantillon de *n* individus n'est jamais qu'un tirage aléatoire parmi l'infinité de moyennes d'échantillons de *n* individus que l'on peut réaliser dans la population étudiée. Vous savez très bien, qu'en tirant deux échantillons de *n* individus dans la *même* population, vous obtiendrez deux moyennes *différentes* (à cause des inévitables fluctuations d'échantillonnage). Donc, la "*moyenne d'un échantillon de n individus dans la population*" est bien une variable aléatoire. Qui dit variable aléatoire, dit variance, et donc écart-type. Tout ceci étant encore un peu trop abstrait, il serait plus parlant de le vérifier en pratique. C'est ce que nous allons faire par simulation.

Grâce au logiciel statistique R (très puissant et complet, gratuit, à télécharger sur www.r-project.org), on peut effectuer facilement des tirages aléatoires dans une loi quelconque. J'ai choisi ici la loi statistique régissant le fameux QI (Quotient Intellectuel) dans l'espèce humaine, car elle est connue : il s'agit d'une distribution approximativement normale, sa moyenne vaut 100 (par convention) et son écart-type vaut environ 15 (on le sait pour avoir fait passer des centaines de milliers de tests de QI sur tous les continents). Voici comment demander à R de créer un premier échantillon A de 10 individus en piochant dans cette distribution :

```
>A=rnorm(10, m=100, sd=15)
```

Ces instructions se décryptent ainsi :

"Créer un objet nommé **A**, auquel il faut attribuer (**=**) des tirages aléatoires (**r**andom) dans une loi **norm**ale, je veux **10** tirages, la **m**oyenne de cette loi vaut **100** et son écart-type (**s**tandard **d**eviation) vaut **15**".

On obtient alors (par exemple) un échantillon comme celui-ci :

```
119, 95, 102,  99, 118, 113,  97, 107, 105,  89.
```
(données arrondies pour des raisons de lisibilité)

Moyenne : $m_A = 104{,}46$ (calculée sur les données non arrondies)

On constate que, évidemment, on n'obtient pas *exactement* la moyenne théorique de 100 points de QI, à cause des fluctuations d'échantillonnage.

Et maintenant "recommençons l'expérience" en réclamant un échantillon B:

```
>B= rnorm(10, m=100, sd=15)
```

```
74, 125,  86, 123,  71,  97,  89,  97, 101, 102
```

Moyenne : $m_B = 96{,}55$

On constate (toujours sans aucune surprise) que la seconde moyenne est différente de la première. Comme le second échantillon a été tiré dans la même population statistique (même loi de distribution normale N(100:15)), cette différence s'explique bien entendu uniquement à cause des fluctuations d'échantillonnage.

Et maintenant, le grand jeu. On va obtenir par simulation **l'erreur standard**, (autrement dit l'écart type) de la variable aléatoire "*moyenne du QI de 10 individus choisis au hasard"*. Et avant de le simuler, on va d'abord essayer de le prédire.

Selon ce qui a été vu plus haut, on a ici :

écart-type des données = $\sigma = 15$ (c'est l'écart type de la "loi du QI")

Puisque erreur standard. $= \sigma/\sqrt{n}$ on peut prédire que ici, l'erreur standard (i.e. l'écart type du QI moyen de 10 individus) sera "racine de dix fois" plus petite que 15. On s'attend donc à la valeur :

erreur standard $= 15/\sqrt{10} = 4{,}743416$**.** Soit environ **4,7**.

Obtenir la valeur *exacte* 4,743416 par simulation est impossible, car il faudrait effectuer une *infinité* de tirages. On va se contenter ici de vérifier si on obtient bien une valeur *proche* de **4,7** en collectant mille échantillons de 10 individus, et en calculant à chaque fois la moyenne de l'échantillon. On se retrouvera donc avec mille moyennes, et il ne restera plus qu'à calculer l'écart-type de cette série de mille données. Si toute cette belle théorie de l'erreur standard n'est pas de la fumisterie, on devrait tomber sur une valeur proche de **4,7**.

Dans R, voici la manœuvre (le caractère # signale des commentaires):

```
> mille.valeurs=numeric(1000)        #création d'un tableau nommé mille.valeurs qui servira à stocker les 1000 valeurs numériques obtenues par la simulation
> for(i in 1:1000)                   #On va faire 1000 répétitions en faisant varier un compteur nommé i de i=1 à i=1000
{                                    #début de la boucle
mille.valeurs[i]=mean(rnorm(10,m=100,sd=15))   #pour chacune des 1000 répétitions, ranger dans le tableau mille.valeurs, dans la case de rang [i], la moyenne (mean) obtenue à partir d'un échantillon aléatoire (random) de 10 individus tirés dans une loi normale de moyenne 100 et d'écart-type (standard deviation) 15
}                                    #fin de la boucle
```

Voyons déjà la moyenne générale du QI obtenue sur cette foule de gens :

```
> mean(mille.valeurs)
[1] 99.86966
```

A un pouillème près, c'est exactement la valeur théorique de 100. Notre estimation est très précise, car on a tout de même 10 000 individus au total (mille échantillons de 10 personnes).

Et maintenant le moment de vérité, réclamons à grand cris **l'écart-type** de ces mille moyennes d'échantillons (donc, la fameuse **erreur standard**) et voyons si on est proche de **4,7.** (Rappel : `écart-type = standard deviation en anglais`)

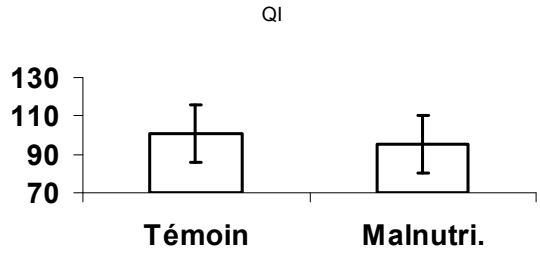
```
> sd(mille.valeurs)
[1] 4.714994
```

C'est pas beau ça ?

On constate donc, par une expérience concrète (bien que *in silico*) que la relation erreur standard = $\sigma/\sqrt{n}$ tient la route remarquablement bien.

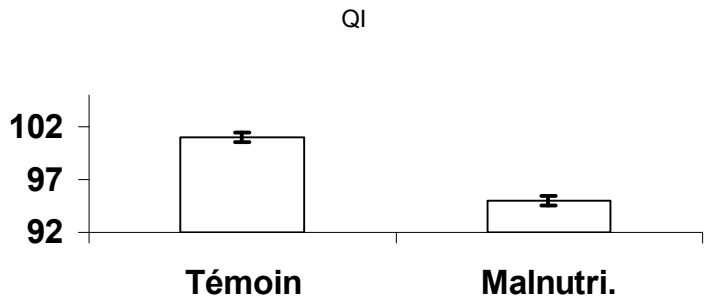
Reparlons maintenant des **barres d'erreur** des graphes scientifiques. Vous comprenez peut-être maintenant pourquoi il est **important** d'apprendre à faire la

distinction entre l'écart-type des données individuelles (ici, σ = 15) et celle de la moyenne que vous présenteriez sur le graphe, si vous aviez utilisé un échantillon de 10 individus (ici, e.s. = 4,71). Supposons en effet que votre expérience teste l'effet de la grave malnutrition d'une femme enceinte sur le QI d'un enfant à naître. Je ne connais rien à ce thème de recherche, mais on va supposer, sans prendre de grands risques, que la grave malnutrition d'une femme enceinte a peu de chances d'être positive pour la maturation du cerveau de son enfant. Supposons donc, qu'une malnutrition sévère fasse perdre 5 points de QI en moyenne chez le futur enfant, mais que personne n'en sache rien. Vous en avez pourtant l'intuition, et souhaitez le démontrer. Supposons encore que, grâce à une étude à très grande échelle, vous soyez parvenu à rassembler deux échantillons de 1000 naissances correctement constitués, de manière à ce que la malnutrition de la mère soit bien le *seul* critère les séparant (ce qui me semble difficilement réalisable, soit dit en passant). Vous calculez les deux moyennes de QI chez les enfants qui en sont issus, et le résultat est : 101 chez le groupe témoin, 95 dans le groupe "malnutrition". Si vous utilisez pour les barres d'erreur de votre graphe le simple écart-type des données (la racine carrée de la variance, donc ici σ=15), cela va donner ceci :



Un scientifique jetant un coup d'oeil à ce graphe conclurait immanquablement (en pensant avoir affaire à de véritables barres **d'erreur standard**) que la différence constatée est **non significative**, et donc qu'on ne peut rien conclure de particulier sur l'effet de la malnutrition des mères sur le QI de leurs enfants. Comme ce thème est important, et que le groupe "malnutrition" obtient cependant un score inférieur, le scientifique en question vous encouragerait peut-être à continuer vos recherches mais avec une échantillon beaucoup plus grand, pour réduire l'incertitude de vos mesures. Bref, de telles "barres d'erreur" trompent complètement le lecteur et donnent, par dessus le marché, une image de grande imprécision à vos estimations, même lorsqu'elles sont très précises. Un comble !

En revanche, si vous avez compris que mille individus permettent normalement une estimation très fiable, vous réaliserez que la valeur correcte à utiliser pour vos barres d'erreur n'est pas le simple écart-type σ =15 mais bien l'erreur standard d'une moyenne de mille individus, qui est *racine de 1000* (soit 31,6) *fois plus petit* = c'est-à-dire 15/31,6 = 0,47. Et sur le graphe, ça change tout :



Cette fois, vous apercevez à peine les barres d'erreur tant elles sont minuscules, et pourtant j'ai donné un coup de zoom (regardez l'échelle des ordonnées). Ce graphique indiquerait, sans même faire de test statistique, qu'il existe une diminution significative de plusieurs points de QI en cas de malnutrition. En effet, les *intervalles de confiance* des moyennes sont environ deux fois plus larges que ces barres d'erreur minuscules. On voit bien (en multipliant mentalement la taille des barres par deux) qu'aucune des deux moyennes n'est située dans l'intervalle de confiance de l'autre. Les deux moyennes sont bien distinctes : la malnutrition diminue significativement le QI (selon cette expérience fictive).

Erreur standard d'un pourcentage.

L'erreur standard d'un pourcentage est simplement son écart-type. Si on appelle p_o le pourcentage observé (exprimé de 0 à 100) calculé sur n individus et $q_o = 1 - p_o$ son complémentaire à 100, alors on peut représenter l'erreur standard. de p_o sur un graphique en calculant :

$$es = \sqrt{(p_o \times q_o / n)}$$

Comme pour le cas des moyennes, on peut se poser la question suivante : "Mais comment diable peut on prétendre calculer l'écart-type d'un pourcentage *unique*, calculé sur mon seul échantillon de n individus ?". La réponse est similaire à la situation de la moyenne : vous calculez en fait l'écart-type d'une variable aléatoire que l'on pourrait nommer "*pourcentage obtenu à partir d'un échantillon de n individus tirés au hasard dans cette population*". Vous savez bien que deux échantillons de n individus donneront deux pourcentages différents à cause des fluctuations d'échantillonnage. Le pourcentage que vous obtenez à partir de *votre* échantillon n'est donc qu'un tirage au sein de l'infinité des pourcentages que l'on pourrait obtenir dans cette population en prélevant des échantillons de n individus. Ces pourcentages fluctueraient autour du véritable pourcentage p, qui restera à jamais inconnu. Or, une variable aléatoire possède une variance et un écart-type, et c'est lui que vous calculez ici. Démonstration par simulation.

Si la proportion réelle de gauchers dans une population est de $p = 0,1$ (soit 10%) d'où $q = 0,9 = 1 - p$, alors la variance de la fréquence des gauchers *dans la population réelle* (qui mesure la dispersion autour de la valeur moyenne $p = 0,1$ gaucher par tirage) est invariable et vaut pq soit $0,1 \times 0,9 = 0,09$. L'écart-type est donc $\sqrt{0,09} = 0,3$. Ceci découle des propriétés de la loi binomiale qui régit les pourcentages, et reste immuable, quel que soit le nombre n d'individus dans *votre* échantillon. Il n'en est pas de même de *l'erreur standard* de votre pourcentage observé p_o, qui traduit la *précision* avec laquelle ce pourcentage estime le véritable pourcentage p. En effet, cette erreur standard sera d'autant plus petite que n sera grand. Il suffit pour s'en convaincre d'étudier par simulation l'écart-type d'une grande série de pourcentages estimés p_o, obtenus sur des échantillons de n individus. On s'attend à ce que cet écart-type (qui est une erreur standard) soit $\sqrt{n}$ fois plus petit que l'écart type réel dans la population, qui vaut 0,3. Si on prend des échantillons de 9 individus par exemple, l'erreur standard devrait être (aux fluctuations d'échantillonnage près) de $0,3/\sqrt{9} = 0,3/3 = 0,1$. Vérifions si ça marche.

```
> mille.valeurs=numeric(1000)        #création d'un tableau nommé mille.valeurs qui servira à stocker les 1000 valeurs numériques obtenues par la simulation
> for(i in 1:1000)                   #On va faire 1000 répétitions en faisant varier un compteur nommé i de i=1 à i=1000
{                                    #début de la boucle
mille.valeurs[i]=mean(rbinom(9,      #pour chacune des 1000 répétitions, ranger dans le tableau mille.valeurs, dans la case de rang [i], la moyenne (mean) obtenue à partir d'un seul (size=1) échantillon aléatoire (random) de 9 individus tirés dans une loi binomiale de moyenne p=0,1.
size=1, p=0.1))
}                                    #fin de la boucle
```

Il n'y a plus qu'à calculer l'écart-type et voir si on est proche de $0,3/3 = 0,1$.

```
> sd(mille.valeurs)
[1] 0.1009985
```

Suffisamment proche de 0,1 pour votre goût ? Avec cet effectif, on peut donc représenter le pourcentage observé dans notre échantillon avec une barre d'erreur de 0,1 unités (ou 10%) de part et d'autre de la valeur observée.