Utilisation de tests statistiqueS

Jean-Baptiste DENIS[1]

Hervé MONOD[2]

Rapport technique 2007-1, 22 pp.

Unité Mathématiques et Informatique Appliquées
INRA
Domaine de Vilvert
78352 Jouy-en-Josas Cedex
France

[1] Jean-Baptiste.Denis@Jouy.Inra.Fr
[2] Herve.Monod@Jouy.Inra.Fr

Utilisation de tests statistiqueS

J.-B. Denis et H. Monod*

3 avril 2007

Table des matières

*INRA Jouy-en-Josas, Unité de Mathématiques et Informatique Appliquées, Domaine de Vilvert, 78352 Jouy-en-Josas Cedex

Introduction

Cette note a été écrite pour répondre à la demande de praticiens s'interrogeant sur l'usage intensif de tests qu'ils sont conduits à interpréter sur un même lot de données et/ou sur des lots de données différents pour conclure sur une même question. En la composant, nous n'avons aucune visée théorique, pas plus que nous n'avons la prétention de répondre à toutes les questions pratiques posées. Simplement, nous espérons qu'après l'avoir parcourue les lecteurs auront une meilleure appréhension de la démarche sous-jacente aux tests, et en tireront profit pour une pratique plus raisonnée, mais toujours appuyée sur le sens commun.

Dans la première section, nous redéfinirons par l'exemple ce que sont les tests, puis nous aborderons la question des tests multiples pour terminer par des remarques de bon sens, dont une petite allusion à la posture bayésienne qui propose une démarche non basée sur des tests.

Pour illustrer notre propos, nous nous appuierons sur des exemples tirés du domaine agronomique. Cependant les notions introduites ne seront pas spécifiques de ce domaine. Ainsi, l'exemple à venir de comparaison entre deux variétés de céréales peut être transposé à la comparaison de deux traitements en pharmacologie, de deux produits agro-alimentaires, ou encore de deux procédés de fabrication dans un contexte industriel.

1 Tests Simples

Avant de faire compliqué, il convient de bien cerner la démarche. Dans cet esprit, le mieux est d'aborder les concepts dans la situation la plus simple qui soit, c'est ce qui est proposé dans cette section.

1.1 La posture générale

La présentation classique des tests se fait au travers de la théorie des jeux. Le statisticien[1] est censé deviner un état de la nature qui lui est inconnu. Pour répondre à ce genre de question, il dispose d'informations prodiguées par les données. Le jeu est simplifié par le fait que l'état de la nature est supposé ne comprendre que deux alternatives et que le statisticien doit obligatoirement choisir entre les deux états.

Pour prendre un exemple : *une nouvelle variété B de blé est-elle plus productive dans une région agricole précisée qu'une ancienne variété A* ? Les données disponibles pourraient être les rendements de chacune des variétés mesurés sur une parcelle. Les deux états de la nature envisagés seraient donc :

1. $A < B$, oui c'est intéressant d'introduire la nouvelle variété sur le marché[2],
2. $B \leq A$, non ce n'est pas intéressant.

Cette présentation conduit au Tableau 1 décrivant classiquement les quatre possibilités.

Dans la diagonale du tableau, la décision prise a été correcte, par contre pour les deux autres cases la réponse est fausse ! Comme les deux types d'erreurs n'ont pas forcément les mêmes

[1] Par ce terme, nous désignons celle ou celui qui utilise l'outil *statistique*.

[2] en fait, dans la réalité, on demanderait sans doute que l'écart soit supérieur à un certain niveau pour justifier une nouvelle introduction... cela ne changerait pas le problème posé mais compliquerait seulement les notations.

TAB. 1 – Les quatre possibilités résultant de l'état de la nature et des décisions prises.

	Réalité $A < B$	Réalité $B \leq A$
décision $A < B$	Correct	Erreur de 2de espèce
décision $B \leq A$	Erreur de 1ère espèce	Correct

TAB. 2 – Cadre conceptuel des tests

	H_0 est vrai	H_1 : H_0 est faux
On choisit H_0	Correct	Erreur de 2de espèce
On rejette H_0	Erreur de 1ère espèce	Correct

conséquences, on les distingue en les numérotant[3]. La distinction est particulièrement flagrante s'il s'agit de formuler un diagnostic médical : ne pas détecter un cancer est sans doute plus grave que de se tromper en pensant à tort qu'il existe. C'est pour cela qu'on dissymétrise les deux états de la nature en *hypothèse nulle* (H_0), celle qu'*a priori* on privilégie ; et en *contre hypothèse* ou *hypothèse alternative* (H_1), rassemblant les états complémentaires de H_0 jugés possibles de la nature. De manière générique, on utilise le Tableau 2. En pratique, les quatre situations présentées dans ce tableau peuvent se produire, puisque l'on ne dispose que d'une information incomplète pour trancher entre les deux hypothèses. Les probabilités d'erreur sont appelées *risque de première espèce* (lorsque H_0 est vraie) et *risque de seconde espèce* (lorsque H_1 est vraie). Il est bien entendu nécessaire d'employer des méthodes statistiques permettant de limiter ces deux risques, compte tenu des données disponibles.

Continuant l'exemple de nos deux variétés, nous allons supposer[4] que chacun des deux rendements (R_A et R_B) dont nous disposons est distribué selon une loi normale, avec une espérance propre à chaque variété (notée μ_A et μ_B) et une variance commune connue valant $50\,(q/ha)^2$, plus formellement :

$$\begin{aligned} R_A &\sim N(\mu_A, 50) \\ R_B &\sim N(\mu_B, 50) \end{aligned}$$

Nous admettrons aussi que dans ce cas simple, la différence de rendement observée $D = R_B - R_A$ est la seule information utile. Les hypothèses précédentes conduisent à

$$D \sim N(\delta, 100) \tag{1}$$

Nous avons donc un modèle qui se réduit à un seul paramètre [5] , $\delta = \mu_B - \mu_A$, l'espérance de l'écart entre les deux variétés. La valeur, inconnue, de ce paramètre correspond à l'état de la nature que nous voulons déterminer.

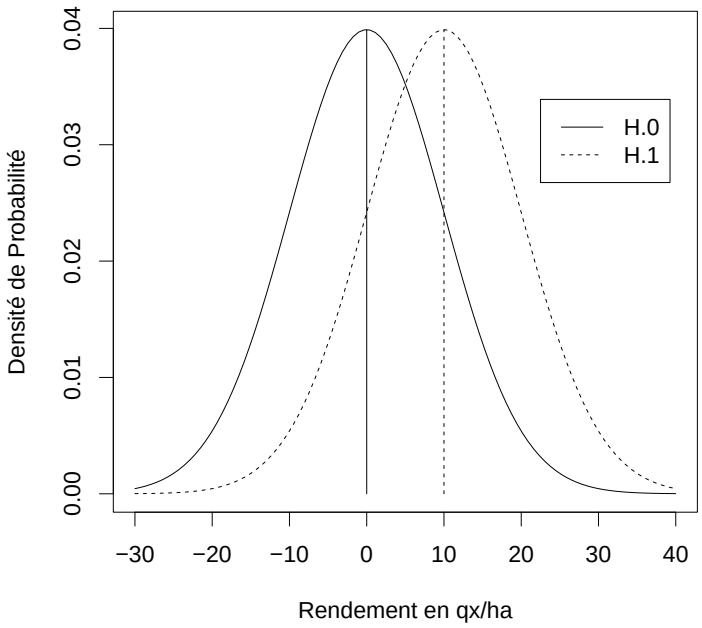
Pour aller progressivement en suivant cet exemple, nous envisagerons trois cas de figure du plus simple au plus compliqué : (1) hypothèse ponctuelle contre hypothèse ponctuelle, (2) hypothèse ponctuelle contre hypothèse composite et (3) hypothèse composite contre hypothèse composite.

[3] Dans cette note nous emploierons systématiquement *de première espèce* et *de seconde espèce* qui se traduisent en anglais par *type 1* et *type 2*.

[4] Ce genre de postulat doit se fonder sur les connaissances préalables dont on dispose sur les variables aléatoires étudiées.

[5] La variance est 100 car la variance d'une somme de variables aléatoires indépendantes est la somme des variances (ici $50 + 50$).

FIG. 1 – Distribution de D selon les deux hypothèses ponctuelles (trait plein : $\delta = 0$, trait en pointillé $\delta = 10$). Pour chacune des deux courbes, l'aire sous la courbe entre deux valeurs de rendement représente la probabilité d'observer une valeur de D dans cet intervalle.



1.2 Tests d'hypothèse ponctuelle contre hypothèse ponctuelle

Bien que ce ne soit pas tellement réaliste supposons que l'on puisse affirmer que de deux choses l'une : soit nos deux variétés sont équivalentes (hypothèse *a priori* privilégiée), soit la nouvelle variété l'emporte sur l'ancienne de 10 q/ha. Dans le cadre du modèle formulé dans l'équation (1), les deux hypothèses s'expriment comme :

$$\begin{aligned} H_0 &: \delta = 0 \\ H_1 &: \delta = 10 \end{aligned}$$

On les appelle ponctuelles puisqu'elles se traduisent chacune par la restriction du paramètre du modèle, à *un point* sur son domaine de variation initialement possible (ici l'ensemble des valeurs réelles).

Il est alors possible d'exprimer complètement l'état de la nature sous chacune des deux hypothèses.

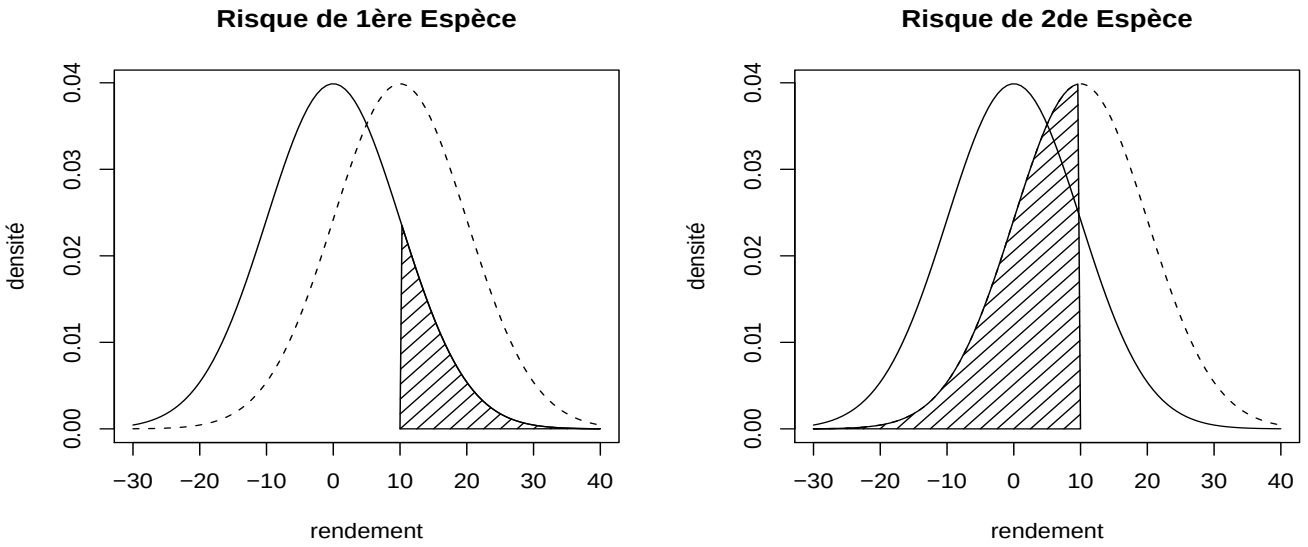
– si H_0 est vraie, alors D suit une loi de probabilité $N(0, 100)$;
– au contraire, si c'est H_1 qui est vraie, alors D suit une loi de probabilité $N(10, 100)$.

La Figure 1 présente dans le même graphe ces deux alternatives.

Logiquement, on va supposer que le statisticien retient la règle de décision basée sur D et définie par

– si D observé est inférieur à k, alors on retient H_0,

FIG. 2 – Figuration des deux risques pour la valeur limite $k = 10$. Les aires sous la courbe associées aux erreurs de décision, en grisé, sont ici égales à 0.16 (risque de 1ère espèce) et 0.50 (risque de 2ème espèce).



TAB. 3 – Probabilités associées à la figure 2, pour la valeur limite $k = 10$

	H_0 est vrai	H_1 : H_0 est faux
On choisit H_0	0.84	0.50
On rejette H_0	0.16	0.50

– dans le cas contraire, on retient H_1,

où k est une valeur limite à déterminer. On peut facilement représenter visuellement les deux risques en reprenant le graphe précédent (Figure 2, pour $k = 10$). Toujours pour $k = 10$, le Tableau de base 2 devient le Tableau 3.

Et sur le même principe, il est tout à fait possible de calculer, pour toute valeur de k, les 4 probabilités de bonnes et mauvaises décisions suivant les deux états précis de la nature que nous avons admis. Voici les valeurs obtenues pour $k = -10, -5, 0, 5, ..., 30$.

```
              <---   H0 vrai --->     <---   H1 vrai --->
              correct      risq.1     correct      risq.2
k = -10         0.16         0.84       0.98         0.02
k =  -5         0.31         0.69       0.93         0.07
k =   0         0.50         0.50       0.84         0.16
k =   5         0.69         0.31       0.69         0.31
k =  10         0.84         0.16       0.50         0.50
k =  15         0.93         0.07       0.31         0.69
k =  20         0.98         0.02       0.16         0.84
k =  25         0.99         0.01       0.07         0.93
```

```
k =  30        1.00        0.00        0.02        0.98
```

Elles montrent bien qu'on ne peut gagner sur tous les tableaux : les deux risques varient en sens inverse en fonction de k et il y a des valeurs de k qui favorisent de manière outrancière une hypothèse ou l'autre. Plus k est petit et plus on a de chance de décider H_1 de manière correcte, mais au détriment d'une décision correcte si c'est H_0 qui est le vrai état de la nature. Il n'est pas possible sans prendre un parti, de déterminer comment équilibrer les deux risques. Une stratégie raisonnable serait certainement d'ajouter un troisième choix possible, celui de ne pas avoir à décider... mais ce n'est pas la coutume ! La coutume est de construire le test pour contrôler le risque de première espèce, c'est à dire de contrôler la probabilité de rejeter à tort H_0[6]. De manière arbitraire, on admet généralement un risque de première espèce de 5% (souvent noté α) ; ce qui conduirait dans notre exemple à prendre comme valeur limite k de la décision un nombre compris entre 15 et 20... et entraînerait un risque de seconde espèce compris entre 69 et 84%. Plus précisément, il faut prendre $k = 16.45$ exactement, et le risque de seconde espèce est alors de 74%. Evidemment, on ne gagne pas sur tous les tableaux !

1.3 Tests d'hypothèse ponctuelle contre hypothèse composite

La situation précédente n'était pas très réaliste. Il serait sans doute plus raisonnable de prendre comme contre-hypothèse (H_1) que la nouvelle variété ne peut être que supérieure à l'ancienne, car le semencier ne demanderait jamais la mise sur le marché d'une variété très inférieure à celles qui sont déjà disponibles. Ceci se formalise par

$$\begin{aligned} H_0 &: \delta = 0 \\ H_1 &: \delta > 0 \end{aligned}$$

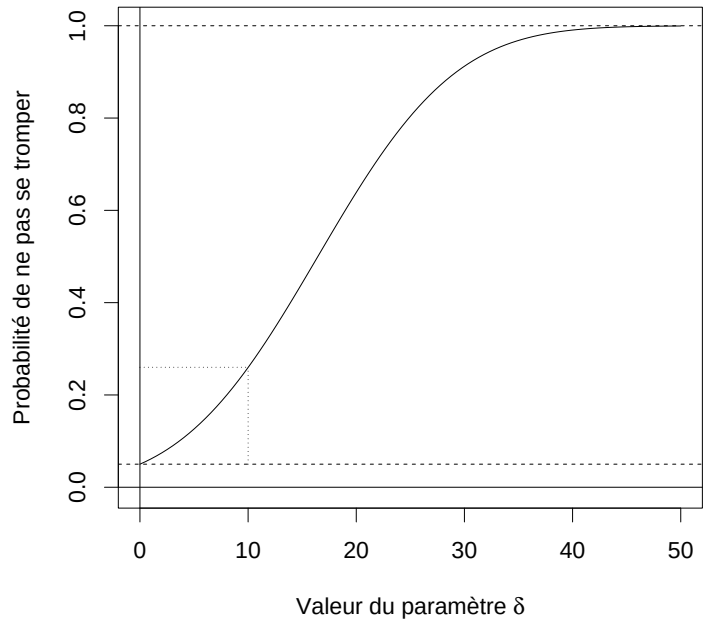
La Figure 1 peut encore être reprise, mais en permettant à la densité sous l'hypothèse alternative (celle tracée en pointillés) de glisser de sa position tout en restant à droite de la densité en trait plein. Elle peut donc se trouver, suivant la valeur de δ, dans sa position actuelle ou encore s'éloigner plus de la première, ou au contraire s'en rapprocher autant qu'on veut (sans pour autant s'y confondre). En fait si on reprend le modèle posé en (1), cela revient à considérer toutes les valeurs strictement positives de δ.

La différence notable par rapport à la situation antérieure, est qu'on se trouve avec une contre-hypothèse H_1 qui peut prendre une infinité de positions, chacune d'entre elles correspondant à un cas hypothèse ponctuelle contre hypothèse ponctuelle, c'est la justification du terme *hypothèse composite*. Pour chacune de ces versions possibles de la contre-hypothèse, on peut mener les calculs précédents sur les risques d'erreur. Si on applique la stratégie coutumière de baser la prise de décision sur le risque de première espèce, alors on aboutira au même choix de valeur limite k et donc aux mêmes propriétés dans le cas de H_0 vrai. Par contre le risque de seconde espèce variera selon la réalité, inconnue, de la contre hypothèse H_1. Le pire des cas pour ce risque sera lorsque les deux densités seront quasi confondues (lorsque δ tend vers 0). A cette situation limite, la probabilité de se tromper (= risque de seconde espèce) est égale à celle d'accepter H_0, soit 95% !

Evidemment, il est un peu tendancieux de se placer dans les circonstances les plus défavorables sous H_1, dans la mesure où elles correspondent à des écarts insignifiants par rapport à l'hypothèse nulle ! Il est par contre important de pouvoir déterminer avec quelle probabilité

[6] qui prend donc le rôle de l'hypothèse de référence.

FIG. 3 – Courbe puissance pour la contre-hypothèse composite, lorsque le test est effectué avec comme valeur limite $k = 16,45$ afin d'assurer un risque de 1ère espèce égal à 5% .
La puissance est la probabilité de ne pas se tromper lorsque la contre-hypothèse H_1 est vraie. Si la contre-hypothèse est composite, c'est une fonction des paramètres qui la spécifient. Ici, il n'y a qu'un paramètre : $\delta > 0$, on peut donc tracer cette puissance comme une fonction de δ. Lorsque $\delta \to 0$, elle tend vers le risque de première espèce (5%=0.05), lorsque $\delta \to +\infty$, elle tend vers 1. Elle vaut 0.26 (26%) lorsque $\delta = 10$.



la décision correcte H_1 sera prise dès lors que l'écart à l'hypothèse nulle atteint des valeurs non négligeables[7]. Cette propriété correspond à ce que l'on appelle la puissance du test, définie comme le complément à un du risque de seconde espèce. La puissance est une fonction du ou des paramètres qui précisent la contre-hypothèse (ici δ). Dans notre exemple, elle est la fonction qui, à chaque valeur possible de $\delta > 0$ sous H_1, associe la probabilité de faire le choix correct H_1. Lorsqu'on pratique un test avec un risque de première espèce égal à 5%, elle prend une valeur égale (5%) lorsque $\delta \to 0$, une valeur de 26% quand $\delta = 10$, et une probabilité qui tend vers 1 au fur et à mesure que le paramètre augmente. La courbe puissance de notre exemple est représentée dans la Figure 3. D'autres considérations sur la puissance sont présentées en §1.5.

Le test que nous venons de discuter est appelé *unilatéral* car la contre-hypothèse H_1 n'envisage des écarts à l'hypothèse nulle que dans une direction. Une variante est le test *bilatéral* défini par

$$\begin{aligned} H_0 &: \delta = 0 \\ H_1 &: \delta \neq 0 \end{aligned}$$

[7] la détermination de ce qui est non négligeable pouvant bien sûr être un problème en soi

Les notions développées ci-dessus s'adaptent facilement à cette variante, la principale différence étant la nécessité de préciser deux valeurs limites k_- et k_+ dans les règles de décision. En pratique, on choisit généralement $k_- = k_+ = k$ et l'on choisit H_0 si $|D| \leq k$, H_1 sinon.

1.4 Tests d'hypothèse composite contre hypothèse composite

L'hypothèse nulle peut aussi ne pas être ponctuelle ! Toujours avec le même exemple de comparaison d'une nouvelle variété par rapport à une ancienne variété, on peut vouloir distinguer : H_0 : la nouvelle variété est de rendement inférieur ou égal à celui de l'ancienne variété, opposée à H_1 : la nouvelle variété est de rendement strictement supérieur à celui de l'ancienne, soit :

$$\begin{aligned} H_0 &: \delta \leq 0 \\ H_1 &: \delta > 0 \end{aligned}$$

Pour une même règle de décision, le risque de première espèce, à son tour, ne prend plus une valeur unique sous H_0 mais dépend de la réalité, inconnue, sous H_0. Sur notre exemple, sa valeur est d'autant plus élevée que δ est proche de 0, quel que soit le choix de la valeur limite k. La pratique des statisticiens dans cette situation consiste à se ramener au cas précédent en se plaçant dans la plus défavorable situation (ici $\delta = 0$) pour évaluer le risque de première espèce et en déduire la règle de décision ...

Un autre cas d'hypothèses doublement composites se produit en renversant le rôle joué par les hypothèses nulle et alternative. L'hypothèse privilégiée *a priori* devient celle qui postule une différence substantielle entre variétés, alors que l'alternative devient celle qui postule une équivalence entre variétés. Dans cette approche, la notion d'égalité stricte est remplacée par celle d'équivalence, c'est-à-dire d'une différence inférieure à une valeur limite admise. Les hypothèses deviennent donc

$$\begin{aligned} H_0 &: \delta \leq -L \text{ ou } L \leq \delta \\ H_1 &: -L < \delta < L, \end{aligned}$$

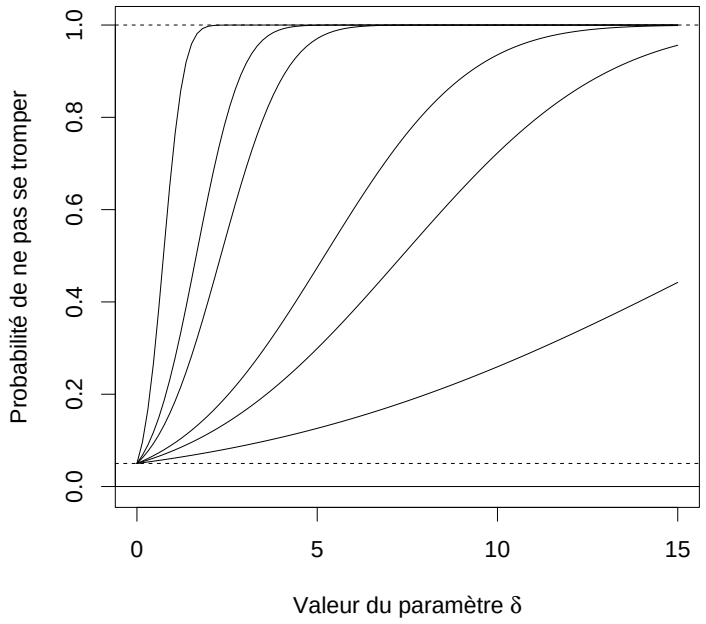
où L représente la valeur limite. Dans le domaine pharmaceutique ou alimentaire, ce type de test est appelé test de bio-équivalence.

1.5 Influence des répétitions sur la puissance

En général, on ne se contente pas d'un seul rendement par variété pour comparer deux variétés, on mesure des rendements de parcelles expérimentales dans des réseaux d'essais souvent fort complexes pour éliminer ou contrôler les effets des autres facteurs déterminants du rendement. Simplifions à l'extrême cette réalité en supposant que nous soyons capable d'avoir *le même nombre R de vraies répétitions* pour chacun des rendements. Les données dont nous disposons pourraient alors s'écrire :

$$\begin{aligned} R_{Ar} &\sim N(\mu_A, 50) \\ R_{Br} &\sim N(\mu_B, 50) \end{aligned}$$

FIG. 4 – Variation de la fonction puissance en fonction du nombre de répétitions
Il s'agit d'une figure semblable à la figure 3 mais les abscisses ont été limitées pour mieux faire apparaître les formes des six courbes représentées. La courbe la plus basse correspond à une seule répétition (c'est donc la même que celle de la figure 3), les autres correspondent dans l'ordre à $R = 5, 10, 50, 100, 500$ répétitions.



où $r = 1, ..., R$ indique les numéros de répétitions disponibles. La statistique naturelle pour effectuer le test serait bien entendu la différence des rendements moyens : $D_* = \left(\frac{1}{R}\sum_{r=1}^{R} R_{Ar}\right) - \left(\frac{1}{R}\sum_{r=1}^{R} R_{Br}\right)$ et on peut facilement vérifier que

$$D \sim N\left(\delta, \frac{100}{R}\right) \tag{2}$$

on retrouve bien l'équation (1) lorsque le nombre de répétitions est égal à 1. La figure 4 représente les puissances pour différents nombres de répétitions. Pour chacune, on part bien de 5% lorsque δ est nul pour tendre asymptotiquement vers l'unité, c'est à dire la probabilité certaine de ne pas se tromper sous H_1. Le phénomène supplémentaire mis en évidence est que l'on se rapproche plus vite de cette certitude plus le nombre de répétitions est élevé. Avec 50 répétitions, un écart de $\delta = 2,5$ q/ha a la même probabilité (environ 55%) d'être mis en évidence qu'un écart de $\delta = 17,7$ q/ha avec une seule répétition !

Le même type de calcul peut bien sûr être mené lorsque les nombres de répétitions sont inégaux. Si les variétés sont répétées R_1 et R_2 fois respectivement, la variance de D est égale à $\left(\frac{1}{R_1} + \frac{1}{R_2}\right) 50$ (q/ha)^2.

2 Tests Multiples

On dit être en situation de tests multiples

1. quand sur un même jeu de données, on pratique simultanément plus d'un test (c'est à dire qu'on veut statuer sur plusieurs états élémentaires de la nature),
2. ou quand pour répondre à une même question, on dispose de plusieurs jeux de données,
3. ou encore les deux à la fois (les deux situations précédentes ne sont pas forcément très tranchées).

Le nombre de tests en jeu est parfois considérable.

2.1 La problématique

Les données sont le plus souvent chères et donc rares. Lorsqu'on dispose d'un bon jeu de données, il n'est pas d'usage de pratiquer un seul test puis de ne plus les considérer. En dehors d'un premier test, on fait beaucoup d'autres choses, par exemple d'autres tests !

Quelques situations classiques :

- On dispose de mesures indépendantes d'une seule variable pour des individus de différentes zones géographiques, on s'intéresse à l'égalité de la mesure entre couples de zones... on retrouve une situation généralisant celle de la comparaison de deux zones, similaire à la comparaison pratiquée en §1 entre les variétés A et B. Il s'agit du problème des comparaisons multiples (cf. §2.3).
- La même question peut se poser de manière multivariable, c'est à dire que chaque mesure indépendante correspond à un ensemble de variables. Par exemple, on compare des variétés de blé sur la quantité de matière sèche produite par unité de surface, mais aussi sur le taux de protéine du grain et la résistance à un champignon microscopique. Les difficultés supplémentaires sont (1) qu'il n'est pas certain que les mêmes hypothèses de distribution soient vérifiées pour toutes les variables et (2) que les corrélations entre les variables doivent être prises en compte (cf. §2.4).
- Les enquêtes cas témoins (fort utilisées en épidémiologie). Pour chaque malade repéré, on associe un individu sain dont les caractéristiques principales sont les plus proches possibles de celles du malade (âge, sexe, type d'habitat,...) et on regarde pour un certain nombre de facteurs suspectés, si les malades se distinguent significativement des sains (par exemple par l'alimentation, la pratique sportive,...). On est donc amené à pratiquer des tests pour chaque facteur suspecté sur la même population d'individus, alors que les facteurs sont eux mêmes très liés[8] entre eux !
- On s'intéresse à un grand nombre de gènes, chaque individu porte ou pas le gène, ce qui classe la population des individus en deux populations complémentaires pour chaque gène. A côté de cela, on a des variables phénotypiques, comme des caractéristiques agronomiques (rendement, résistance à une maladie...) et on pratique pour chaque gène un test pour décider de ceux qui sont en correspondance avec la variable phénotypique ?
- Sur un ensemble d'individus d'une population définie, on a observé beaucoup de variables, on peut imaginer des mesures morphométriques : longueurs de feuille, largeurs de feuille, hauteur de tige, hauteur de l'inflorescence,... La question est de savoir quelles sont les corrélations réelles qui les lient : la hauteur de l'inflorescence est-elle reliée à la longueur des feuilles mais pas à la longueur de la tige ? Comme on dispose d'un test de la nullité de corrélation entre deux variables, on va l'utiliser pour tous les couples de variables. Mais si les deux couples (U, V) et (V, W) sont déclarés liés, faut-il aussi tester le couple (U, W) ? Sans doute le fera-t-on, ne serait-ce que pour avoir un traitement symétrique entre toutes les variables, mais on s'expose alors à des réponses incohérentes.

[8] soit par nature, soit en conséquence d'un faible nombre d'individus.

– Une situation plus confortable que nous examinerons en détail plus loin (§2.2) est celle de mêmes tests sur des lots de données indépendants. On répugne à mélanger les données parce qu'elles sont d'origines très différentes (pays, année, techniques de mesure) mais la question posée pour chaque lot est bien la même. Dans le domaine médical par exemple, des études indépendantes peuvent avoir été consacrées à la même question et conclues pour chacune par des tests (sur une relation cancer-alimentation par exemple) pas forcément tous cohérents. En déduire des conclusions globales fait l'objet des méthodes statistiques dites de méta-analyse.

Pour toutes ces situations, la première question qui se pose est de savoir comment on définit H_0 et le risque de première espèce associés, puis quels risques de première espèce on utilise pour chacun des tests individuels. Deux pratiques extrêmes peuvent se rencontrer dans une telle situation :

1. On pratique et interprète chaque test comme s'il était unique. C'est à dire qu'on applique la théorie précédente sans se poser de question. Le risque d'une telle pratique est de trouver - par simple coïncidence - beaucoup d'effets significatifs qui ne le sont pas. En agro-physiologie, il est classique de faire remarquer que si on teste sur un ensemble de parcelles, la corrélation entre le rendement et la pluviométrie de chaque jour de la phase de végétation, ce serait bien le diable si on ne trouvait pas que la pluviométrie d'un certain jour (par exemple le 17 mars) est en forte relation avec le rendement...
2. On définit comme une hypothèse nulle globale que *toutes les hypothèses nulles élémentaires sont vraies*, et on cherche à contrôler le risque de première espèce associé, c'est-à-dire la probabilité de rejeter au moins une des hypothèses nulles élémentaires alors qu'elles sont toutes vraies. Il faut alors adapter le risque de première espèce utilisé pour les tests élémentaires. Pour cela on utilise une inégalité[9] d'autant plus grossière que le nombre de tests est élevé. La conséquence est inverse que pour la stratégie précédente, à savoir que la puissance diminue extrêmement au point d'aboutir le plus souvent à ne rien mettre en évidence.

Il est sans doute recommandable d'utiliser la technique intermédiaire, détaillée par exemple dans Benjamin et Hochberg (1995). Cette technique permet de contrôler la proportion des tests que l'on déclare significatifs à tort (car l'hypothèse nulle associée est en fait vraie) parmi l'ensemble des tests que l'on déclare significatifs (en anglais cette proportion est appelé "False Discovery Rate"). En fait, on ne peut pas contrôler ce taux de façon exacte, mais on contrôle son espérance. Le niveau de rejet appliqué à chaque test est alors fonction du rang de sa *p-value*. Graphiquement si on place les *p-values* en y et les rangs en x, la significativité correspond au plus grand sous-ensemble des tests de plus petite *p-value* en dessous d'une certaine droite croissante. Cf. la référence pour les détails techniques.

2.2 Cas de tests indépendants

Lorsque les tests sont pratiqués sur le même jeu de données, l'appréciation de la situation n'est pas simple. En effet, basés sur les mêmes observations les tests pratiqués ne sont pas indépendants. Au contraire, si chaque test élémentaire est pratiqué sur un jeu de données indépendant des autres, les tests élémentaires sont également indépendants et il est plus facile de se livrer à un certain nombre de calculs caractérisantles propriétés. C'est la raison pour laquelle nous allons d'abord examiner cette situation en nous livrant à un certain nombre de calculs des probabilités de ce qui peut arriver quand nous connaissons l'état de la nature.

[9] due à Bonferroni.

TAB. 4 – Répartition des résultats de N tests indépendants, lorsque H_0 est vraie dans n cas sur N.

	Réalité H_0	Réalité H_1
décision H_0	$n-x$	y
décision H_1	x	$N-n-y$

2.2.1 Comportement global d'une série de N tests indépendantset semblables

Supposons donc, le cas de N tests semblables[10] indépendants. Pour simplifier encore, admettons qu'il s'agisse de tests de la catégorie *hypothèse ponctuelle contre hypothèse ponctuelle* comme nous l'avons vu en §1.2. Alors le problème est simplement spécifié par trois quantités :

1. le risque de 1ère espèce *individuel*, c'est-à-dire celui utilisé pour chaque test, nous prendrons 5% ;
2. le risque de seconde espèce individuel ; il est unique dans notre cas simplifié et nous prendrons les 74% déjà obtenus ;
3. le nombre inconnu de tests, n, pour lesquels l'état de la nature est H_0, ce qui implique que les $N-n$ autres sont dans le cas de la contre-hypothèse.

Le résultat de ces N tests est donné par leur répartition dans les quatre cases du Tableau 2. Par exemple, on pourrait obtenir le Tableau 4, avec $x \in 0...n$ et $y \in 0...N-n$.

Il est quasi immédiat de vérifier que[11] $x \sim Bino\,(n, 0.05)$ et $y \sim Bino\,(N-n, 0.74)$ de manière indépendante.

A titre d'illustration, voici quelques indications sur les valeurs que peut prendre x pour des valeurs croissantes de N, lorsque $n = N$, c'est à dire lorsque jamais la contre-hypothèse n'est vraie.

Examinons d'abord graphiquement (Figure 5) le cas de $N = 100$. Le nombre de tests significatifs que l'on trouve n'est pas nul (la probabilité qu'il soit zéro est même très faible : $0.95^{100} = 0.006$), ce qui est assez logique et la conséquence du risque de première espèce.

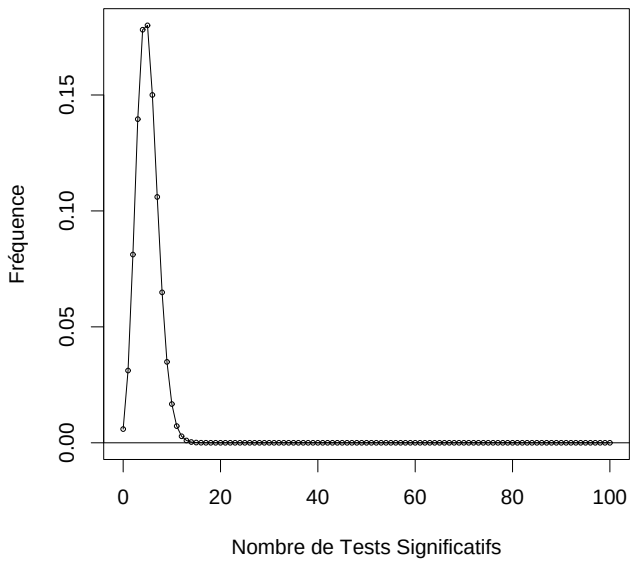
Si nous caractérisons une telle distribution par le mode, le percentile 75%, ainsi que la probabilité d'être égal à 0, on trouve les valeurs suivantes pour une gamme de tests s'étendant de $N = 10$ à $N = 10000$.

```
N =     10  mode =    0 Q75% =    1    P(x=0) = 0.599
N =     30  mode =    1 Q75% =    2    P(x=0) = 0.215
N =    100  mode =    5 Q75% =    6    P(x=0) = 0.006
N =    300  mode =   15 Q75% =   17    P(x=0) < 10^-6
N =   1000  mode =   50 Q75% =   55    P(x=0) < 10^-22
N =   3000  mode =  150 Q75% =  158    P(x=0) < 10^-66
N =  10000  mode =  500 Q75% =  515    P(x=0) < 10^-222
```

[10] par semblables, nous pensons "qui sont pratiqués au même niveau pour la même hypothèse nulle (H_0) et qui sont sujets aux mêmes contre-hypothèses (H_1)".

[11] Nous notons $Bino\,(N, p)$ la distribution binomiale de taille N et de paramètre de probabilité p.

FIG. 5 – Distribution du nombre de tests significatifs lorsqu'on fait $N = 100$ tests simultanéments et que tous correspondent à l'hypothèse nulle, c-a-d que $n = N$ dans le tableau 4.



Le mode est la valeur la plus probable de x, et le percentile 75% est la valeur de x qui est dépassée avec une probabilité de 25%. L'interprétation de ce tableau est donc que si on réalise 1000 tests indépendants au risque de première espèce de 5% et que l'état de la nature est tel que c'est toujours l'hypothèse nulle qui est vraie, alors le nombre de tests significatifs (c'est-à-dire rejetant H_0) tourne autour de 50. Ces 50 faux résultats ne sont que le pur fruit du hasard. Ils sont la conséquence de la définition du risque de première espèce : en espérance, on rejettera à tort l'hypothèse nulle dans 5% des cas.

Lorsque les tests ne sont pas indépendants, pour obtenir des résultats similaires des taux d'erreurs que l'on risque d'obtenir, il faut modéliser leur dépendance, en fonction de nombreuses caractéristiques en général inconnues. Ce que l'on peut imaginer, c'est que les tests iront le plus souvent dans le même sens (si un est significatif, alors les autres aurant tendance à être plus significatifs). Dans ce cas, on conserve toujours les 5% en espérance mais avec une variabilité beaucoup plus grande autour de cette valeur centrale.

2.2.2 Conséquences pratiques

La notion d'erreur en situation de tests multiples devient plus complexe et peut donner lieu à plusieurs attitudes différentes, comme évoqué plus haut. En particulier, il faut distinguer :

1. le risque d'erreur de première espèce *individuel*, c'est-à-dire celui associé à chacun des tests (5% dans notre exemple) ;
2. le risque d'erreur de première espèce dit *global* ou *simultané*[12], qui est la probabilité de rejeter au moins l'une des hypothèses H_0 alors qu'elles sont toutes vraies (égal par exemple à $1 - 0.599 = 0.401$, lorsque $N = 10$).

[12]le terme anglais est *family wise error rate*

Lorsque l'on veut contrôler le risque global, on privilégie en fait le test de l'hypothèse nulle globale $\mathbf{H_0}$: "toutes les N hypothèses H_0 sont vraies" contre la contre-hypothèse $\mathbf{H_1}$: "au moins une hypothèse H_0 est fausse" [13]. Les calculs ci-dessus montrent qu'en prenant un risque de 1ère espèce individuel de 5%, le risque global n'est pas contrôlé et augmente rapidement avec N. Une première recommandation lorsque N est grand consiste à évaluer si le pourcentage de tests significatifs reste compatible avec la fréquence attendue αN si l'hypothèse H_0 est vraie pour tous les tests. Une autre solution est d'ajuster le niveau des tests individuels. L'ajustement le plus courant et le plus général est celui dit de Bonferroni[14], qui consiste à choisir comme risque de 1ère espèce individuel $\alpha = \alpha_G/N$, garantissant que le risque global est inférieur ou égal à α_G. Ce n'est cependant pas toujours le plus efficace et il existe d'assez nombreuses alternatives.

En contrôlant mieux le risque de 1ère espèce global, on diminue les risques de 1ère espèce individuels mais *on diminue parallèlement la puissance de ces tests*. Globalement cependant, le fait d'utiliser l'information apportée par plusieurs tests indépendants augmente la probabilité de mettre en évidence des différences s'il en existe. Dans l'exemple de comparaison entre deux variétés, nous avions obtenu, lorsque $\delta = 10$, une puissance égale à 0.26 pour un test sans répétition. Supposons que l'on réalise dix tests similaires indépendamment et que la vraie valeur de δ est égale à 10 dans tous les cas. Sans ajustement du niveau individuel des tests, la puissance globale serait égale à 0.95[15], mais le risque de première espèce serait bien au dessus de 0.05 (0.401 comme vu précédemment). Avec ajustement de Bonferroni, la puissance est égale à 0.447. Elle est égale à 0.452 avec un ajustement alternatif nécessitant l'hypothèse d'indépendance.

Dans certaines situations, il est préférable de privilégier le risque individuel des tests, qui s'interprète aussi comme l'espérance du taux de faux positifs. C'est par exemple le cas en bio-informatique, lorsque l'on cherche à identifier parmi un très grand nombre de gènes ceux qui sont exprimés de façon différente entre deux traitements.[16].

2.3 Cas des comparaisons multiples

Un cas très classique de tests liés est celui des comparaisons multiples. Reprenant notre exemple initial des variétés, imaginons que nous expérimentions six variétés. Pour chacune desquelles nous associons les paramètres de production moyenne $\mu_A, \mu_B, \mu_C, \mu_D, \mu_E, \mu_F$. Dans le cadre de l'analyse de variance, on peut tester l'hypothèse nulle :

$$\mu_A = \mu_B = \mu_C = \mu_D = \mu_E = \mu_F$$

contre l'hypothèse que ces cinq égalités ne sont pas vérifiées. Cette contre-hypothèse peut s'exprimer par le fait qu'au moins un couple de variété n'a pas même espérance.

Mais ces deux alternatives sont un peu courtes pour les expérimentateurs qui n'accepteront pas de savoir seulement si les six variétés sont équivalentes ou pas. Si elles ne le sont pas, ils sont bien entendu désireux de savoir, quelle est la meilleure (ou quelles sont les meilleures), quelle est la pire (ou quelles sont les pires), de manière plus générale si on peut établir un classement plus ou moins complet.

Pour répondre à cette question, il est naturel de pratiquer les tests pour tous les couples de variétés à l'image de ce qui a été fait en §1.4 pour comparer les variétés A et B. Mais alors

[13] c'est une situation du type : on n'accepte l'équivalence entre deux variétés que si elles sont non significativement différentes pour tous les tests

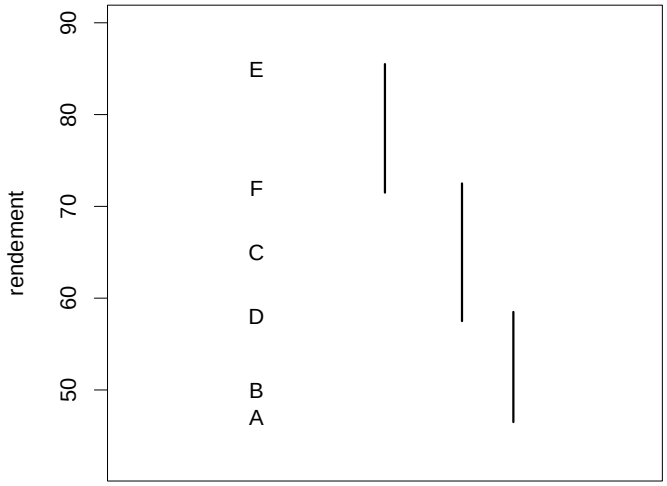
[14] en particulier il s'applique aussi si les tests ne sont pas indépendants

[15] cette valeur est une pure coïncidence et n'a rien à voir avec $1-\alpha$. Elle se calcule par $1-(1-0.26)^{10}$

[16] c'est d'ailleurs un sujet de recherche très actif, et plusieurs variantes ont été proposées dans la bibliographie de ce domaine

FIG. 6 – Affichage d'un test de comparaison multiple.

Les deux barres verticales indiquent :
- les variétés **F,C,D** ne peuvent pas être déclarées différentes,
- les variétés **D,B,A** ne peuvent pas être déclarées différentes,
- toutes les autres comparaisons ont été trouvées significatives. Par exemple la variété E est considérée différente de toutes les autres (et supérieure de plus).



on doit en pratiquer $\frac{6(6-1)}{2} = 15$. Tests qui sont très liés puisque chaque variété participe à 5 d'entre eux. C'est en général ce qui se fait. Soit avec un risque de première espèce égal à 5%, soit avec une valeur plus faible déterminée selon une règle plus ou moins empirique telle que celle de Bonferroni. Et généralement, on présente les résultats comme une série de sous-ensembles de variétés (le plus souvent se recoupant) pour lesquels on ne trouve pas de différences significatives. La Figure 6 en propose un exemple.

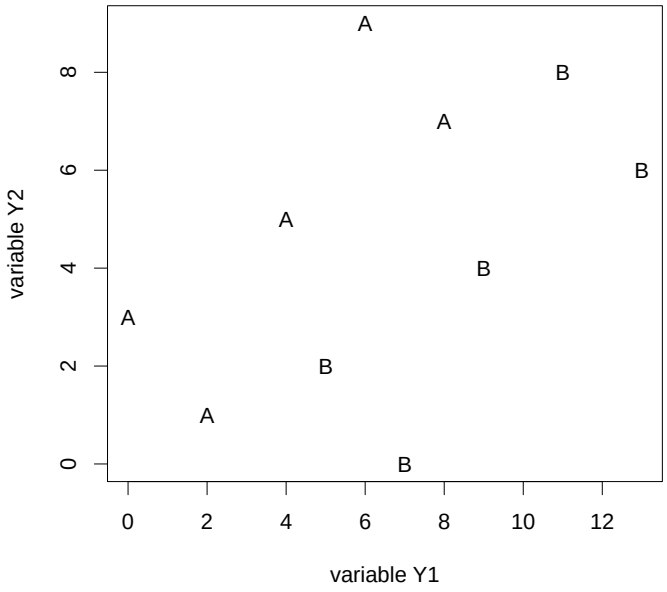
Une incohérence logique (mais pas statistique) existe puisque $\mu_F = \mu_C = \mu_D$ et $\mu_D = \mu_B = \mu_A$ ce qui devrait entraîner l'égalité $\mu_F = \mu_A$, or ces deux variétés sont déclarées significativement différentes ! Mais, est-ce si grave que la réponse fournie par le statisticien ne soit pas plus précise ? Puisque c'est à l'expert d'intégrer toutes les informations disponibles[17] ?

2.4 Situation multivariable

La situation multivariable est celle dans laquelle plusieurs variables sont mesurées sur les mêmes échantillons ou les mêmes individus. Il faut alors considérer que des corrélations peuvent exister entre ces différentes variables. Une illustration en est donnée dans la Figure 7. Il y a 10 échantillons, cinq pour la variété A et cinq pour la variété B, et deux variables mesurées. Le

[17] Certaines données n'ont peut-être pas été prises en considération, certains ressentis d'experts ne sont pas forcément quantifiés, le contexte social et économique peut justifier une décision contradictoire avec des résultats expérimentaux peu marqués...

FIG. 7 – Exemple de situation bivariée
Les deux axes représentent deux variables mesurées sur cinq échantillons de la population A et cinq échantillons de la population B.



graphique indique une corrélation positive entre ces deux variables : pour une même variété, les échantillons ont tendance à s'éloigner de la moyenne dans la même direction pour les deux variables. Sur cet exemple, on perçoit également que suivant la direction selon laquelle on compare les observations, les deux variétés semblent quasiment équivalentes ou au contraire différentes. La démarche des tests multivariables a pour objectif d'exploiter ce type d'information tout en contrôlant les risques d'erreur.

Globalement, on peut distinguer deux démarches possibles :

1. on effectue des tests individuels sur chacune des variables, en ajustant si nécessaire le risque de 1ère espèce afin de contrôler au mieux le risque global ;
2. on effectue des tests dits multivariés, qui prennent en compte simultanément l'ensemble des variables et leurs corrélations.

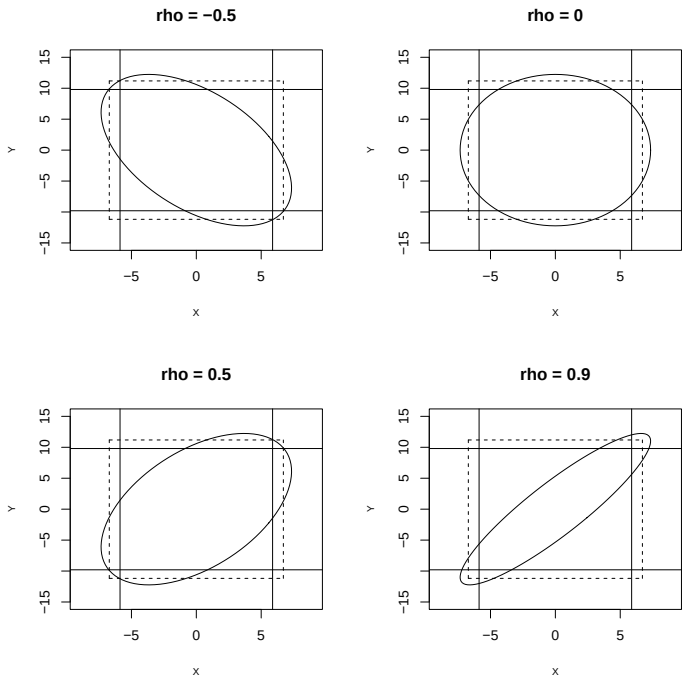
La construction d'un test multivariable est cependant une opération compliquée pour les raisons suivantes :

– la spécification des hypothèses nulle et alternative n'est pas si simple, par exemple la notion d'inférieur/supérieur n'a plus de sens,
– il n'existe pas de procédure de test globalement optimale, sauf sous des conditions très restreintes,
– les approximations par des distributions connues sont beaucoup plus difficiles à assurer, même la multinormalité de quelques variables dont chacune individuellement peut être admise comme normale n'est pas acquise facilement...

Le cadre le plus confortable pour conduire des tests multivariés est celui où l'on peut supposer la normalité de l'ensemble des variables. Le test multivarié le plus classique dans ce cas est le T^2 de

FIG. 8 – Deux tests univariables versus un test bivariable
Sous hypothèse de binormalité et pour des valeurs de corrélation $-0.5, 0, 0.5$ et 0.9 :

1. l'intérieur de l'ellipse correspond à une zone de probabilité 0.95 et peut donc servir à la construction d'un test bivariable (X, Y) de risque de première espèce 0.05.
2. la zone à l'extérieur des deux lignes verticales pleines (-1.96,1.96) correspond au test bilatéral sur la seule variable X.
3. la zone à l'extérieur des deux lignes horizontales pleines (-1.96,1.96) correspond au test bilatéral sur la seule variable Y.
4. l'intérieur de la zone délimitée par les lignes en pointillé forme une zone de probabilité 0.95 *si* X et Y sont indépendantes. Sous cette hypothèse, il correspond à un test bilatéral simultané sur les deux variables X et Y, de risque de première espèce global 0.05.



Hotelling. Dans le cas d'une comparaison entre traitements, il peut être interprété comme une moyenne des carrés des écarts entre traitements calculés variable par variable, pondérée par les variances (estimées) de chacune des variables et prenant en compte leurs corrélations (estimées). Le test s'effectue en comparant la valeur observée de T^2 à des valeurs limites calculées sous l'hypothèse nulle.

La Figure 8 illustre le cas de la binormalité. Les deux axes correspondent, par exemple, aux différences δ_X et δ_Y entre les variétés A et B pour deux variables X et Y. L'hypothèse nulle et la contre-hypothèse du test bivariable sont :

$$\begin{aligned} H_0 &: \delta_X = 0 \text{ et } \delta_Y = 0 \\ H_1 &: \delta_X \neq 0 \text{ ou (inclusif) } \delta_Y \neq 0 \end{aligned}$$

Les régions de rejet des hypothèses nulles d'un test bivariable et des tests univariables ont été tracées. On y constate que les deux régions ne se recouvrent pas et que celle du test bivariable

dépend fortement de la corrélation entre les deux variables. Dans le cas corrélé, les deux types de tests donnent des réponses différentes surtout lorsque les valeurs relatives des différences moyennes sont peu cohérentes avec la corrélation entre les deux variables. Le test bilatéral simultané est alors conservatif, c'est-à-dire que le risque de première espèce est strictement inférieur à 0.05.

Les données répétées sont un cas particulier de la situation multivariable, lorsque les différentes variables sont en fait une même variable réponse mesurée à plusieurs reprises au cours du temps sur les mêmes individus. Il est parfois suffisant et justifié d'effectuer des tests individuels limités à certaines dates clés. Mais il est également possible d'utiliser l'ensemble de l'information disponible pour comparer les dynamiques dans leur ensemble. Cela peut se faire soit par des tests multivariables généraux, soit en modélisant explicitement la relation entre la variable réponse et le temps, et en testant l'égalité de certains paramètres de ce modèle.

3 Remarques (de bon sens)

3.1 Les conclusions consécutives à un test

La présentation de la démarche des tests (§1.1) a mis en évidence la dissymétrie entre l'hypothèse nulle (H_0) et la contre-hypothèse (H_1). Par construction, on se protège contre l'erreur de première espèce (rejeter à tort l'hypothèse nulle). Si aucune étude de puissance (§1.3) n'est menée, les conclusions qu'on peut en retirer sont aussi dissymétriques.

Si l'hypothèse nulle est rejetée, alors on peut affirmer qu'un effet a (très probablement) été mis en évidence.

Si l'hypothèse nulle n'est pas rejetée, alors on ne peut certainement pas affirmer avoir mis en évidence qu'elle était valide : simplement avec les moyens expérimentaux dont on s'est doté, il n'est pas possible de la contredire. Si les expériences étaient plus précises, ou simplement plus nombreuses, cela n'est pas exclu. Il est donc indispensable pour pouvoir conclure quelque chose de se livrer à une étude de la puissance du test réalisé qui permettra de statuer sur une affirmation du genre : *avec les moyens déployés, nous avions une probabilité de β de mettre en évidence un écart à l'hypothèse nulle de δ...* La difficulté étant qu'hormis pour les cas d'école, comme celui que nous avons développé, la formulation de δ n'est pas immédiate et peut être sujet à caution. Un autre moyen de quantifier l'écart possible à l'hypothèse nulle consiste à donner des intervalles de confiance aux paramètres ou combinaisons de paramètres dont on veut tester la nullité.

Un autre élément important pour l'interprétation de tests statistiques est que la notion de variabilité recouvre généralement plusieurs origines[18]. Il est important pour juger de la significativité *biologique* de procédures statistiques, d'identifier les niveaux de variabilité par rapport auxquels on doit se situer. Par ailleurs, lorsque l'on effectue des tests statistiques pour comparer des variétés ou des traitements sur des données expérimentales, il est nécessaire que la variabilité entre répétitions d'un même traitement corresponde bien à la variabilité entre traitements[19].

[18] en particulier on a coutume de distinguer la variabilité biologique et l'incertitude due aux mesures, notions à préciser dans chaque situation

[19] c'est en particulier le rôle de la randomisation en planification expérimentale

3.2 Interprétation et validation comme renfort

Si les résultats d'une batterie de tests sont cohérents avec une interprétation plausible, basée par exemple sur une théorie explicative, la situation est beaucoup plus confortable que si aucun appui raisonnable ne peut leur être trouvé.

Si on prend un ensemble de chevaux et qu'on les fait courir, il y aura toujours un premier et un dernier. Si la distance qui les sépare est courte, ce n'est pas pour cela que le premier est meilleur que le second. Si on mène beaucoup de tests, on en trouvera très probablement une proportion proche au risque de première espèce qui sera significative, comment distinguer les hypothèses rejetées par hasard, de celles rejetées par suite d'un effet ?

Si on en a la possibilité, le mieux est sans doute de recueillir de nouvelles données, orientées pour contrôler un sous-ensemble d'hypothèses nulles sur lesquelles il est difficile de trancher.

3.3 Tests de permutation

Lorsque les hypothèses distributionnelles nécessaires pour pratiquer des tests classiques sont difficiles à assumer, une très bonne alternative consiste à pratiquer des tests de permutation. Le principe est de générer à partir des données dont on dispose, un ensemble de données équivalentes sous l'hypothèse nulle, par permutation des valeurs observées. On compare ensuite une statistique de test calculée sur les données observées à la distribution de cette statistique sur les données permutées un grand nombre de fois.

Supposons par exemple que l'on a observé, de façon indépendante, les huit rendements suivants sur chacune des variétés A et B :

```
A : 51.58 63.88 59.56 61.88 58.34 52.95 50.11 58.40
```

```
B : 72.05 73.21 84.53 76.44 67.28 78.78 71.40 63.20
```

La différence observée entre les médianes[20] des deux variétés est égale à 14,26 q/ha. En effectuant 10000 permutations aléatoires des 16 données et en réattribuant les 8 premières valeurs à A et les 8 dernières à B, on obtient la distribution de la Figure 9 pour la différence entre médianes, avec une valeur limite égale à 11 q/ha pour un test unilatéral avec un risque de première espèce de 5%[21]. La valeur observée dépasse la valeur limite, on rejette donc l'hypothèse H_0 d'absence de différence entre les variétés.

Le test de permutations fait partie des tests dits non paramétriques, qui reposent sur un minimum d'hypothèses sur les lois de distribution des observations. Un autre exemple qui pourrait être appliqué sur les données ci-dessus serait le test Wilcoxon-Mann-Whitney, basé sur les rangs des observations plutôt que sur la médiane des différences.

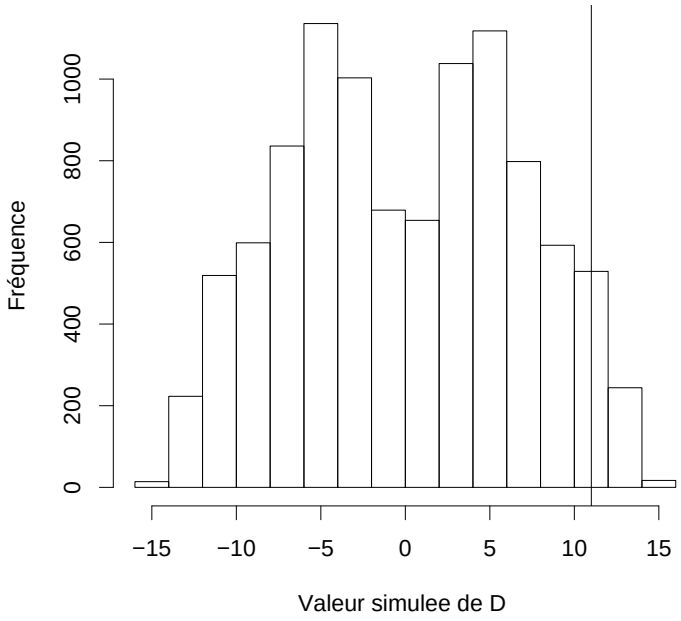
3.4 Optique bayésienne

Il ne serait pas raisonnable de développer des éléments de statistique bayésienne dans ce petit document. Néanmoins, il peut être judicieux de faire remarquer que toute la difficulté de la théorie qui vient d'être esquissée repose sur l'état non accessible de la nature, sur le fait qu'on suppose l'existence d'une valeur vraie inconnue des paramètres.

[20] dans le cadre de ces tests, on utilise plus fréquemment la médiane que la moyenne, car elle est moins sensible à la distribution des données

[21] cette valeur limite est celle qui est juste supérieure à 95% des valeurs obtenues par permutations aléatoires

FIG. 9 – Test de permutation : distribution de la différence des médianes entre les variétés A et B, pour 10000 permutations aléatoires des données. Le trait vertical correspond à la valeur limite pour un risque de 1ère espèce de 5%



La posture bayésienne est différente. On intègre dans la formalisation le degré de connaissance que suppose avoir l'interprétateur vis à vis du paramètre au moyen d'une distribution de probabilité. Si la connaissance est très bonne cette distribution sera très piquée, si elle est très mauvaise elle sera très étalée (peu informative). Il n'est donc pas supposé qu'il existe une valeur vraie du paramètre, c'est une variable aléatoire dont la distribution associée traduit notre degré de connaissance à son égard. L'équivalent du test ne sera pas une réponse en tout ou rien, mais l'expression d'un degré de certitude.

Une optique bayésienne ne donne donc pas prise à décider si un paramètre (ou une fonction des paramètres) est supérieur à 0 ou pas. La réponse apportée par le statisticien bayésien est donc du style *le paramètre est supérieur à* 0 *avec une probabilité de* 0.85... au décideur ensuite de conclure !

A Quelques références

Benjamin Y. et Hochberg Y. (1995). Controlling the False Discovery Rate : a Practical and Powerful Approach to Multiple Testing. *J. R. Statist. Soc. B,* **57**, 289-300.

Daudin J.-J., Robin S. et Vuillet C. (1999). *Statistique Inférentielle. Idées, démarches, exemples.* Société Française de Statistique, Presses Universitaires de Rennes.

Jouhan-Flahault C., Casset-Semanaz F. et Minimi P. (2004). Du bon usage des tests dans les essais cliniques. *Médecine/Sciences,* **20**, 231-5.

Miller R. G. Jr (1981). *Simultaneous Statistical Inference.* Springer-Verlag, Berlin.

Saporta G. (1990). *Probabilités, Analyse des Données et Statistique.* Editions Technip, Paris.