Les tests d'équivalence et de non-infériorité : *peut-on « démontrer » l'hypothèse nulle ?*

L'enseignement de statistique en Terminale S et ES est notamment orienté vers l'étude post-bac des tests d'hypothèses.

Ceux-ci sont fondés sur une asymétrie de traitement entre les deux hypothèses en présence, H_0 et H_1 [B1, page 36] ; c'est pourquoi un test d'hypothèses ne permet jamais d'affirmer (statistiquement ...) que H_0 est vraie.

Or un besoin nouveau est apparu dans le domaine de la recherche médicale, celui de démontrer statistiquement que telle nouvelle thérapie n'était inférieure que de très peu à telle autre, plus contraignante pour les patients.

Ceci a conduit à définir les tests de non-infériorité et d'équivalence.

On trouvera quelques éléments d'histoire concernant les tests d'équivalence et de non-infériorité en consultant [1].

Plan

1 Les tests d'hypothèses

1.1 Les hypothèses

Dans une population, on a à étudier une caractéristique numérique $\theta \in \Theta$, qui se trouve être un paramètre d'une variable aléatoire X, par exemple une espérance ou une variance.

On définit une hypothèse concernant θ, appelée hypothèse nulle et notée H_0, ainsi qu'une hypothèse contradictoire avec H_0, appelée hypothèse alternative et notée H_1. Souvent, H_0 est « $\theta = \theta_0$ » et H_1 est « $\theta \neq \theta_0$ ». À partir d'un échantillon de n membres de la population tirés au sort, on recueille les données $x_1, x_2, x_3, ..., x_n$, réalisations de la variable aléatoire considérée.

Le but du test est de déterminer si ces données permettent ou non d'affirmer *raisonnablement* que H_0 est fausse et, en conséquence, que H_1 est vraie.

1.2 Les deux erreurs

Pour atteindre ce but, on utilise une statistique
$T = f(X_1, X_2, X_3, ..., X_n)$

- où les variables X_i sont indépendantes et ont la même loi de probabilité que X
- dont la valeur apporte, si possible, des renseignements *optimaux* sur H_0
- et dont la loi de probabilité *correspondant aux cas où H_0 est vraie* est connue, au moins asymptotiquement.

La statistique T nous permet d'instaurer une règle de décision statistique, qui peut conduire à deux types d'erreurs :

- l'erreur de première espèce : rejeter H_0 alors qu'elle est vraie ;
- l'erreur de deuxième espèce : ne pas rejeter H_0 alors qu'elle est fausse.

On appelle risque de première espèce d'un test et on note α la probabilité, calculée en considérant H_0 *comme vraie*, que la règle de décision de ce test conduise à rejeter H_0. D'après les propriétés de T ci-dessus, il est possible de choisir α, souvent avec $\alpha = 0,05$.

Si on sait seulement que le risque de première espèce maximal d'un test est égal à un nombre α, on dit que ce test est de niveau α.

1.3 La fonction risque de deuxième espèce

Le risque de deuxième espèce, noté $\beta(\theta_1)$, où $\theta_1 \in \Theta$ a une valeur en contradiction avec H_0, est la probabilité, calculée en considérant que $\theta = \theta_1$, de ne pas rejeter H_0.

On voit que le risque $\boldsymbol{\beta}$ (risque de deuxième espèce) est une *fonction* et n'est pas aussi simple à maîtriser que le risque α.

On appelle puissance du test la fonction $\theta_1 \mapsto 1-\beta(\theta_1)$. Sa valeur intéressante dépend, évidemment, de la valeur inconnue du *paramètre* θ.

1.4 Premier exemple : le test d'égalité de deux espérances, les écarts-type étant inconnus mais égaux

1.4.1 Statistique de test

Les variables aléatoires X et Y suivent les lois $\mathcal{N}(\mu_X, \sigma^2)$ et $\mathcal{N}(\mu_Y, \sigma^2)$. On se pose la question de l'égalité de μ_X et μ_Y.
On note H_0 : « $\mu_X = \mu_Y$ » (*hypothèse simple*) et H_1 : « $\mu_X \neq \mu_Y$ » (*hypothèse composite*). [1]
(Remarque : sous certaines conditions, H_0 et H_1 peuvent être *toutes les deux composites*).
Les échantillons indépendants étant $X_1, X_2, X_3, ..., X_{n_X}$ et $Y_1, Y_2, Y_3, ..., Y_{n_Y}$, on note $\overline{X} = \frac{\sum_{i=1}^{n_X} X_i}{n_X}$, $\overline{Y} = \frac{\sum_{i=1}^{n_Y} Y_i}{n_Y}$, $S_X = \sqrt{\frac{\sum_{i=1}^{n_X}(X_i-\overline{X})^2}{n_X-1}}$, $S_Y = \sqrt{\frac{\sum_{i=1}^{n_Y}(Y_i-\overline{Y})^2}{n_Y-1}}$ et $S = \sqrt{\frac{(n_X-1).S_X^2+(n_Y-1).S_Y^2}{n_X+n_Y-2}}$, c'est-à-dire $S = \sqrt{\frac{\sum_{i=1}^{n_X}(X_i-\overline{X})^2+\sum_{i=1}^{n_Y}(Y_i-\overline{Y})^2}{n_X+n_Y-2}}$.
Comme statistique de test, on emploie

$$T = \frac{\overline{X} - \overline{Y}}{S.\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}.$$

1.4.2 Loi et règle de décision

On montre que la loi de probabilité de cette statistique est, dans le cas où H_0 est vraie, la loi de Student à $n_X + n_Y - 2$ degrés de liberté.
Pour la règle de décision à prendre, le choix naturel est, par construction de T, de rejeter H_0 au profit de H_1 (« $\mu_X \neq \mu_Y$ ») lorsque la réalisation de T obtenue est trop grande en valeur absolue, c'est-à-dire lorsque cette valeur absolue appartient à la zone de rejet qu'on détermine comme suit :

$$\alpha = \mathbb{P}_{H_0}(\text{Rejet de } H_0) = \mathbb{P}_{H_0}(|T| > c)$$

ce qui, pour une zone de rejet prise symétrique par rapport à 0, équivaut à $c = t_{1-\frac{\alpha}{2}}$ où t_p est le fractile d'ordre p de la loi de Student à $n_X + n_Y - 2$ degrés de liberté, qui est défini par $\mathbb{P}(T \leqslant c) = p$.

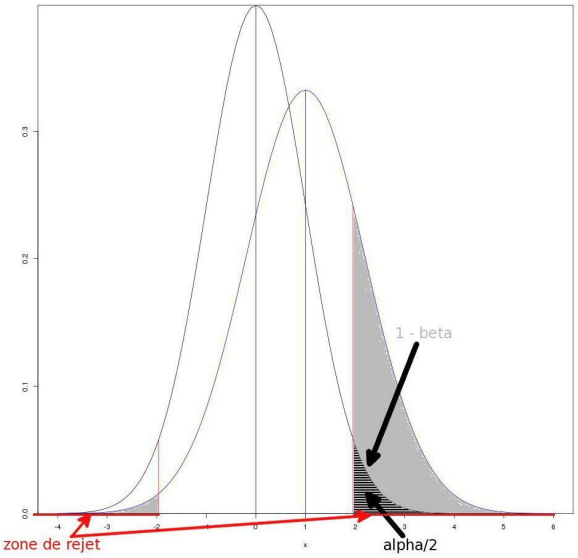
1.4.3 Remarques

— Remarque 1 : une autre écriture de la zone de rejet est $]-\infty\,;\,-c[\,\cup\,]c\,;\,+\infty[$.)

— Remarque 2 : ce test est considéré comme robuste, c'est-à-dire que le résultat obtenu lorsque les variables aléatoires ne suivent pas exactement une loi normale ou que leurs variances ne sont pas franchement égales reste relativement utilisable.

— Remarque 3 : si les effectifs des échantillons sont suffisamment grands, on pourra utiliser, comme loi de probabilité de T dans le cas où H_0 est vraie, la loi $\mathcal{N}(0, 1)$.

1.4.4 Puissance du test



1.4.5 Calcul approximatif de la puissance du test

Pour déterminer la zone de rejet, utilisons $\mathcal{N}(0, 1)$ comme loi approximative pour T puis disons que, en fait,
$\mu_X = \mu_Y + e$ avec $e > 0$ et que S n'est pas aléatoire et a σ comme valeur. Alors,

$$1 - \beta = \mathbb{P}_e(|T| > z_{1-\frac{\alpha}{2}}) \cong \mathbb{P}_e(T > z_{1-\frac{\alpha}{2}})$$
donc $\beta \cong \mathbb{P}_e(T \leqslant z_{1-\frac{\alpha}{2}}) \cong$
$$\mathbb{P}_e\left(\overline{X} - \overline{Y} \leqslant z_{1-\frac{\alpha}{2}}.\sigma.\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}\right) \cong$$
$$\mathbb{P}_e\left(\overline{X} - \overline{Y} - e \leqslant z_{1-\frac{\alpha}{2}}.\sigma.\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} - e\right) \cong$$
$$\mathbb{P}_e\left(\frac{\overline{X} - \overline{Y} - e}{\sigma.\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \leqslant z_{1-\frac{\alpha}{2}} - \frac{e}{\sigma.\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}\right).$$

Or, comme, ici, l'écart entre les espérances inconnues est égal à e, et d'après les décisions prises,
$\frac{\overline{X} - \overline{Y} - e}{\sigma.\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$ suit $\mathcal{N}(0, 1)$. Donc
$z_\beta \cong z_{1-\frac{\alpha}{2}} - \frac{e}{\sigma.\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$ et la puissance est
$1 - \beta \cong 1 - \phi\left(z_{1-\frac{\alpha}{2}} - \frac{e}{\sigma.\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}\right)$ où z_p est le fractile d'ordre p de $\mathcal{N}(0, 1)$ et où ϕ est sa fonction de répartition.

1. Une démonstration de l'impossibilité de choisir comme hypothèse nulle « $\mu_X \neq \mu_Y$ » se trouve dans [2], *page 8*. On trouvera également dans ce document des exemples de fonctions R.

1.4.6 Formule donnant la taille nécessaire pour les échantillons

- Dans le cas où $n_X = n_Y = n$, on a $z_\beta \cong z_{1-\frac{\alpha}{2}} - \frac{e}{\sigma.\sqrt{\frac{2}{n}}}$, soit
- $-z_{1-\beta} \cong z_{1-\frac{\alpha}{2}} - \frac{e.\sqrt{n}}{\sigma.\sqrt{2}}$, c'est-à-dire $n \cong \frac{2.\sigma^2(z_{1-\frac{\alpha}{2}}+z_{1-\beta})^2}{e^2}$.
- Cette formule permet, après avoir *intuité* σ, par exemple à partir d'études précédentes, de déterminer grossièrement l'effectif n des échantillons avec lequel la puissance $1-\beta$ est obtenue pour un choix de l'écart e entre les espérances inconnues.

Une petite application numérique : un choix fait fréquemment est $\alpha = 0,05$ et $\beta = 0,2$. Si nous cherchons à détecter un écart e entre les espérances égal à $\frac{\sigma}{5}$, nous obtenons $n \cong 50.(1,96+0,84)^2 \cong 392$. En fait, réunir deux échantillons indépendants d'environ 400 personnes est rarement accessible : le choix $e = \frac{\sigma}{5}$ n'est pas réaliste.

1.5 Test d'égalité de deux pourcentages

(Dans la suite j'abandonne la notation systématique en majuscule des variables aléatoires $T, \overline{X}, \overline{Y}$, etc.)

Les variables aléatoires X et Y suivent les lois de Bernouilli $\mathcal{B}(1,\pi_X)$ et $\mathcal{B}(1,\pi_Y)$. On se pose la question de l'égalité de π_X et π_Y. On note H_0 : « $\pi_X = \pi_Y$ » et H_1 : « $\pi_X \neq \pi_Y$ ».

Dans la suite j'utilise les notations suivantes :

$p_X := \overline{X}, p_Y := \overline{Y}, p_s := \dfrac{n_X.p_X + n_Y.p_Y}{n_X + n_Y}$.

Comme statistique de test, on emploie

$$T = \frac{p_X - p_Y}{\sqrt{p_s.(1-p_s).\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}}$$

et non pas $T = \dfrac{p_X - p_Y}{\sqrt{\frac{p_X.(1-p_X)}{n_X} + \frac{p_Y.(1-p_Y)}{n_Y}}}$.

(Le premier choix est plus puissant et produit un test équivalent au test du χ^2 de K. Pearson).

« T » a, asymptotiquement, comme loi de probabilité correspondant au cas où H_0 est vraie, la loi $\mathcal{N}(0,1)$, p_X et p_Y estimant les proportions observées dans les échantillons et p_s estimant la proportion globale $\frac{p_X.n_X+p_Y.n_Y}{n_X+n_Y}$.

Ce test ne convient que pour des échantillons ayant de grands effectifs.

Ici, la zone de rejet est également déterminée par :

$$\alpha = \mathbb{P}_{H_0}(\text{Rejet de } H_0) = \mathbb{P}_{H_0}(|T| > c)$$

où c est le fractile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$.

1.5.1 Exemple de test d'égalité de deux pourcentages

Essai randomisé [3] : une étude a inclus 200 patients hospitalisés avec une fracture du col du fémur pour comparer la mortalité à 5 ans entre les patients traités chirurgicalement et les patients traités orthopédiquement (traitement non chirurgical).

H_0 : « $\pi_X = \pi_Y$ » contre H_1 : « $\pi_X \neq \pi_Y$ ».

	Décès à 5 ans		
Traitement	Oui	Non	*Total*
Chirurgical	15 (**15%**)	85 (85%)	**100** (50%)
Orthopédique	25 (**25%**)	75 (75%)	**100** (50%)
Total	40 (**20%**)	160 (80%)	200 (100%)

On trouve que T a valu $t = \frac{0,15-0,25}{\sqrt{0,2.0,8.\left(\frac{1}{100}+\frac{1}{100}\right)}} \cong -1,77$; or la zone de rejet de H_0 au risque $\alpha = 0,05$ est bien connue : $]-\infty\,;-1,96[\cup]1,96;+\infty[$ environ. Donc on ne rejette pas H_0, ce qui ne démontre absolument pas H_0. Que faire ?

Par la suite, le cas de deux pourcentages va être principalement envisagé.

1.6 Compléments

1.6.1 Les tests unilatéraux

- On a vu ci-dessus la difficulté principale, qui concerne le non-rejet de H_0.
- Un autre problème apparaît lors de la comparaison d'un traitement expérimental avec un placebo :

il n'est pas intéressant de s'apercevoir que le traitement est « moins ou plus » efficace que le placebo.

Dans ce cas, un *test unilatéral* convient mieux : il est intéressant de s'apercevoir que le traitement est « plus » efficace que le placebo.

1.6.2 Essai randomisé d'un corticoïde, par infiltration, sur les lombosciatiques, test unilatéral [4]

	Succès/échec à J+20		
Traitement	Succès	Échec	*Total*
Corticoïde x	18 (**41,9%**)	25 (58,1%)	**43**(50,6%)
Placebo y	10 (**23,8%**)	32 (76,2%)	**42**(49,4%)
Total	28 (**32,9%**)	57 (67,1%)	85 (100%)

On utilise la même statistique de test,

$$T = \frac{p_X - p_Y}{\sqrt{p_s.(1-p_s).\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}}$$

mais seules les valeurs *trop* positives de T permettent ici de rejeter utilement « $\pi_X = \pi_Y$ ».

Ainsi, on prendra H_1 : « $\pi_X > \pi_Y$ » et, donc, H_0 : « $\pi_X \leqslant \pi_Y$ ».

Les deux hypothèses sont composites ; ce test est seulement *de niveau* α, (mais on ne le dit que rarement ...)
Bien que les deux hypothèses soient composites, un théorème montre que la zone de rejet à considérer est $]c\,;\,+\infty[$ pour le cas unilatéral à droite, où H_1 est $\pi_X - \pi_Y > 0$ comme ici, ou $]-\infty\,;\,-c[$ pour le cas unilatéral à gauche, où H_1 est $\pi_X - \pi_Y < 0$, c étant le fractile d'ordre $1-\boldsymbol{\alpha}$ de la loi $\mathcal{N}(0,1)$.
T a valu $t \cong \frac{0{,}419-0{,}238}{\sqrt{0{,}329.0{,}671.\left(\frac{1}{43}+\frac{1}{42}\right)}} \cong 1,78$ et la zone de rejet au risque $\alpha = 5\%$ est]1,65 ; +∞[. On rejette H_0.
Conclusion : un test de supériorité du *traitement par le corticoïde* avec un *placebo* comme contrôle a été réalisé et le traitement a été effectivement jugé supérieur.

1.6.3 Puissance

On a ici :

$$1-\beta \cong 1-\phi\left(z_{1-\alpha}-\frac{e}{\sqrt{\frac{p_X.(1-p_X)}{n_X}+\frac{p_Y.(1-p_Y)}{n_Y}}}\right)$$

soit $z_{1-\alpha}+z_{1-\beta} \cong \frac{e}{\sqrt{\frac{p_X.(1-p_X)}{n_X}+\frac{p_Y.(1-p_Y)}{n_Y}}}$.
Dans le cas où $n_X = n_Y = n$, on a
$n \cong \frac{(p_X.(1-p_X)+p_Y.(1-p_Y)).(z_{1-\alpha}+z_{1-\beta})^2}{e^2}$.
Ici, pour déterminer l'effectif n des deux échantillons permettant d'obtenir une puissance convenable, on devrait disposer d'une valeur approximative de π_X ou π_Y, pouvant provenir d'études antérieures. Mais, si nous cherchons simplement à majorer n, nous pouvons remarquer que $p_X.(1-p_X)+p_Y.(1-p_Y) \leqslant \frac{1}{2}$
Application numérique : prenons $\alpha = 0,05$, $\beta = 0,2$ et $e = 0,1$. Nous trouvons $n \leqslant \frac{0{,}5.(1{,}65+0{,}84)^2}{0{,}1^2} \cong 310$, ce qui est encore une valeur peu réaliste.
Si nous pensons que $\pi_X = 0,3$, nous remplaçons p_X par 0,3 et p_Y par 0,2 et nous obtenons
$n \cong \frac{(0{,}3.(1-0{,}3)+0{,}2.(1-0{,}2)).(1{,}65+0{,}84)^2}{0{,}1^2} \cong 230$.

2 Présentation des tests d'équivalence ou de non-infériorité dans le domaine des essais cliniques

2.1 Tests d'équivalence ; pourcentages

- Pour traiter une maladie, on dispose d'un traitement standard et, de plus, on étudie un nouveau traitement, expérimental.
- On juge la qualité de chacun des traitements à l'aide du pourcentage des patients guéris (ou bien soulagés, par exemple), respectivement π_s, s comme *standard*, et π_e, e comme *expérimental*.

Le non-rejet de H_0 : « $\pi_s = \pi_e$ » n'est pas une démonstration de H_0.

On voudrait prendre au maximum un risque de première espèce α, fixé, de conclure à tort « $\pi_s \neq \pi_e$ », ce qui conduirait à tester H_0 : « $\pi_s \neq \pi_e$ » contre H_1 : « $\pi_s = \pi_e$ ˇ.

Mais on a vu qu'il existe une démonstration de l'impossibilité théorique de la réalisation d'un tel test. Donc on se contentera de définir un maximum d'écart δ entre l'efficacité du traitement standard et l'efficacité du traitement expérimental et on prendra :

- H_0 : « $|\pi_s - \pi_e| \geqslant \delta$ »
- H_1 : « $|\pi_s - \pi_e| < \delta$ »,

le choix de la valeur de δ relevant principalement de normes définies par les différentes agences internationales.

Ainsi, le rejet de H_0 constituera une démonstration statistique de l'équivalence des deux traitements.

2.2 Tests de non-infériorité : pourcentages (tests unilatéraux)

En fait, on utilise le plus souvent la version unilatérale de ce test avec

- H_0 : « $\pi_s - \pi_e \geqslant \delta$ »
- H_1 : « $\pi_s - \pi_e < \delta$ ».

Rejeter H_0 signifie ici que

- $\pi_e > \pi_s$ si $\pi_s - \pi_e < 0$
- ou, sinon, que π_e n'est pas trop inférieur à π_s : $\pi_s \geqslant \pi_e > \pi_s - \delta$.

Voici comment la nécessité de l'usage d'un test de non-infériorité est décrit dans un article de 2008, article de référence, sûrement, car directement disponible[5] sur le site des National Institutes for Health : « *Pour ce type de tests, l'erreur de première espèce est la probabilité que nous concluions que les résultats obtenus sur le groupe « traitement nouveau » ne sont pas pires que ceux obtenus avec le groupe « témoins » alors qu'en fait ils sont réellement inférieurs.*
Comme nous contrôlons, en employant un test de non-infériorité, l'erreur de première espèce, nous nous protégeons contre une éventuelle conclusion fausse, qui affirmerait à tort la non-infériorité du traitement nouveau par rapport au traitement standard. »

2.2.1 Justification thérapeutique

- Beaucoup de maladies ont aujourd'hui un traitement qu'on considère comme efficace.
- Mais certains de ces traitements peuvent être, pour les patients, très lourds, voire invasifs, ou bien très chers...

— Il peut donc être utile de rechercher de nouveaux traitements qui n'auraient pas ces inconvénients,
- même s'ils ne disposent que d'une efficacité un peu inférieure,
- la différence pouvant peut-être, d'ailleurs, être de l'ordre de l'incertitude sur les pourcentages obtenus lors de l'essai clinique.

2.2.2 Mise en œuvre d'un test de non-infériorité (pourcentages)

Comme on l'a vu, on définit
— H_0 : « $\pi_s - \pi_e \geqslant \delta$ »
— H_1 : « $\pi_s - \pi_e < \delta$ ».
et, donc, on utilise comme statistique de test

$$T = \frac{p_s - p_e - \delta}{\sqrt{\left(\frac{p_s.(1-p_s)}{n_s} + \frac{p_e.(1-p_e)}{n_e}\right)}}$$

— Ce sont les valeurs *trop* négatives de t qui permettent de rejeter H_0.
— La zone de rejet est donc $]-\infty\,;\,-c[$,
c étant le fractile d'ordre $1-\boldsymbol{\alpha}$ de la loi $\mathcal{N}(0,1)$,
ce qui donne]-∞; $-1,65$[au risque α = 5 %.

Remarque : aucune nouvelle justification théorique n'est nécessaire ici, car ce qui a été dit au paragraphe 1.6.2 s'applique également aux tests de non-infériorité.

2.2.3 Exemple : cellules souches

L'allogreffe de cellules souches hématopoïétiques [5] est utilisée pour le traitement de certaines maladies hématologiques. Une étude de 2007 a eu pour but de comparer l'emploi de cellules sources du sang périphérique (PBSC) avec celui de la moëlle osseuse (BM). Lors d'un prélèvement de cellules souches du sang périphérique, le sang est prélevé par le biais d'un cathéter veineux. Le sang est ensuite dirigé vers un séparateur de cellules, appareil centrifugeur qui en extrait les cellules souches et les stocke. Puis le sang est réinjecté au donneur par le biais d'un second cathéter veineux.

On voit que l'avantage du sang périphérique est que son prélèvement est beaucoup moins invasif pour le donneur ; de plus, des études indiquent qu'il a une meilleure efficacité que BM dans le cas des donneurs jumeaux des receveurs. Cependant, dans le cas des donneurs non apparentés aux receveurs, le risque de réaction du greffon contre l'hôte semble accru.

On prend donc ici BM comme traitement expérimental, PBSC comme traitement de contrôle (de référence) et on cherche à savoir si BM n'est, au pire, que peu inférieur à PBSC pour la survie à 6 mois. Le choix a été fait de $\delta = 0,1 = 10\%$.

Mise en œuvre du test

Données	n	mortalité liée au traitement
BM (*exp*)	583	187
PBSC (*std*)	328	95

— La statistique de test vaut donc

$$t = \frac{\frac{233}{328} - \frac{396}{583} - 0,1}{\sqrt{\left(\frac{\frac{233}{328}.(1-\frac{233}{328})}{328} + \frac{\frac{396}{583}.(1-\frac{396}{583})}{583}\right)}} \cong -2,18$$

— Or la zone de rejet pour un « z-test » unilatéral à gauche au risque $\alpha = 5\%$ est environ $]-\infty\,;\,-1,65[$.
— Donc, à ce risque, on rejette l'hypothèse H_0 : « $\pi_s - \pi_e \geqslant \delta$ »
et on déclare que BM est non-inférieur à PBSC du point de vue de la survie liée au traitement,
— en plus de sa plus faible tendance à la réaction du greffon contre l'hôte.

2.2.4 Exemple : trypanosomiase

Objectif de l'étude NECT [6] : comparaison de la combinaison thérapeutique d'eflornithine en administration I.V. (deux fois par 24 h, sept jours) et de nifurtimox par voie orale (10 jours) au traitement de référence d'elfornithine I.V. (quatre fois par 24 h, 14 jours) en termes d'efficacité thérapeutique pour le traitement des patients atteints de trypanosomiase TBG en phase méningo-encéphalique.

Données	n	nombre d'échecs
EN (*exp*)	47	3
E (*std*)	51	5

La statistique de test vaut donc ici

$$t = \frac{\frac{46}{51} - \frac{44}{47} - 0,1}{\sqrt{\left(\frac{\frac{46}{51}.(1-\frac{46}{51})}{51} + \frac{\frac{44}{47}.(1-\frac{44}{47})}{47}\right)}} \cong -2,45$$

Même conclusion : au risque $\alpha = 5\%$, on rejette l'hypothèse H_0 : « $\pi_s - \pi_e \geqslant \delta$ » et on déclare que EN est non-inférieur à E du point de vue de l'efficacité, et plus simple à mettre en œuvre.
Défaut : les données sont telles que l'approximation normale n'est pas justifiée ici ...

2.2.5 Test de non-infériorité et intervalles de confiance (pourcentages)

En admettant que l'approximation normale soit justifiée, un intervalle de confiance de niveau $1-\alpha$ de la différence $\pi_X - \pi_Y$ est $\left[p_X - p_Y - z_{1-\frac{\alpha}{2}}.\sqrt{v}\,;\,p_X - p_Y + z_{1-\frac{\alpha}{2}}.\sqrt{v}\right]$ où $v = \frac{p_X.(1-p_X)}{n_X} + \frac{p_Y.(1-p_Y)}{n_Y}$.
Il a aussi été défini « intervalle de confiance à droite » : $]-\infty\,;\,p_X - p_Y + z_{1-\alpha}.\sqrt{v}]$; de même à gauche.
Or, dans un test de non-infériorité (à gauche, comme plus haut), H_0 : « $\pi_s - \pi_e \geqslant \delta$ » est rejetée
si et seulement si

$$\frac{p_s - p_e - \delta}{\sqrt{\left(\frac{p_s.(1-p_s)}{n_s} + \frac{p_e.(1-p_e)}{n_e}\right)}} < -z_{1-\alpha}$$

c'est-à-dire :

$$p_s - p_e + z_{1-\alpha}.\sqrt{\left(\frac{p_s.(1-p_s)}{n_s} + \frac{p_e.(1-p_e)}{n_e}\right)} < \delta.$$

On voit donc que H_0 est rejetée si et seulement si l'intervalle de confiance à droite de $\pi_s - \pi_e$ de niveau $1-\alpha$ est inclus dans $]-\infty\,;\,\delta[$.

(Pour une justification théorique de la formule employée ici pour obtenir un intervalle de confiance, voir [B3], page 241.)

2.2.6 Intervalle de confiance symétrique (pourcentages)

Si on ne veut pas utiliser d'intervalle de confiance non symétrique, on peut aussi dire que H_0 est rejetée si et seulement si :
la borne droite de l'intervalle de confiance bilatéral de $\pi_s - \pi_e$ de niveau $1-2.\alpha$ est inférieure à la borne droite de l'intervalle $]-\delta\ ;\ \delta[$.
Cette conclusion est compliquée, voire même tortueuse, mais elle s'avèrera utile au paragraphe 3.1.

2.3 Test de non-infériorité pour deux espérances

2.3.1 Exemple des antihypertenseurs

On compare un antihypertenseur expérimental, e, avec un antihypertenseur de référence, s.

- Le critère de jugement est la variation de tension artérielle avant/après traitement (mmHg).
- On considère que ce caractère suit une loi normale, de même variance pour les deux traitements. La statistique de test est

$$T = \frac{\overline{X_s} - \overline{X_e} - \delta}{S.\sqrt{\frac{1}{n_s}+\frac{1}{n_e}}} \text{ avec } s = \sqrt{\frac{(n_s-1).S_s^2+(n_e-1).S_e^2}{n_s+n_e-2}}.$$

- La borne d'équivalence δ choisie est 2 mmHg.

Données :

n_e	$\overline{x_e}$	s_e	n_s	$\overline{x_s}$	s_s
140	13,1	7,8	138	12,0	8,0

T a donc valu $t \cong \frac{12-13,1-2}{\sqrt{\frac{137.8^2+139.7,8^2}{276}}.\sqrt{\frac{1}{138}+\frac{1}{140}}} \cong -3,27.$

- La zone de rejet au risque $\alpha = 5\%$ étant environ $]-\infty\ ;\ -1,65[$, on rejette H_0 et on affirme que le nouvel antihypertenseur n'est pas inférieur au contrôle employé.

2.3.2 Harpagophyton

Voici un extrait d'article [7] *que nous allons décoder et dont nous allons vérifier les calculs* :

« L'hypothèse principale à tester était la suivante : essai d'équivalence unilatérale ou de non-infériorité de l'harpagophyton par rapport à la diacerhéine sur la douleur spontanée mesurée à l'aide d'une Échelle Visuelle Analogique (100 mm). Ce test d'équivalence devait être mené en situation unilatérale avec un risque à 0,05.

En posant l'hypothèse que la différence vraie entre les traitements était nulle avec un δ de 10 mm, un test d'équivalence unilatéral avec un risque α à 0,05, un risque β à 0,10 et un écart type de 18 mm, nécessitait l'inclusion de 56 patients par groupe. »

Tests d'équivalence basés sur les intervalles de confiance de la différence entre les groupes (douleur à l'EVA)

	Moyenne des différences par rapport à la valeur de base		Différence entre les groupes (H-D)	Intervalle de confiance à 90 % de la différence (H-D)	Hypothèse : Harpagophyton est au moins aussi bon que la diacerhéine
	Harpagophyton	Diacerhéine			
Douleur à j120	-30,6 ± 3,3	-25,5 ± 3,6	-5,1	-13,1 ; 3,0	Acceptée

Décodage :

- « Équivalence unilatérale » est une périphrase représentant, comme il est indiqué, un test de non-infériorité. L'échelle visuelle analogique ayant un support borné, l'usage d'un test de Student ne peut être justifié que pour de grands effectifs, ce qui n'est pas encore assuré. Le risque égal à 0,05 est, en fait, le risque α, et c'est indiqué trois lignes plus bas.
- Adaptons la formule vue au paragraphe 1.4.6 : $n \cong \frac{2.\sigma^2(z_{1-\alpha}+z_{1-\beta})^2}{e^2} \cong \frac{2.18^2(1,65+1,28)^2}{10^2} \cong 56.$
- « $-30,6 \pm 3,3$ » indique que la moyenne, calculée sur le groupe de patients recevant de l'harpagophyton, des diminutions des douleurs mesurées par EVA a été égale à 30,6 mm (signe - car il s'agit de diminutions) et que l'estimation de l'écart-type de cette moyenne (ou *erreur-type*) a été égale à 3.3. De même pour« $-25,5 \pm 3,6$ ».
- Un intervalle de confiance de niveau $1 - 2.0.05$ de $\mu_H - \mu_D$ est donc $[m_H - m_D - z_{0,95}.\sqrt{v}\ ;\ m_H - m_D + z_{0,95}.\sqrt{v}]$ où $v = 3.3^2 + 3.6^2$, ce qui donne $[-13,16\ ;\ 2,96]$.
- Nous avons dit que le rejet de H_0 était obtenu lorsque la borne droite de l'intervalle de confiance ci-dessus était inférieure à celle de l'intervalle $[-\delta\ ;\ \delta]$. Ceci s'applique également ici, bien que la soustraction faite, $m_H - m_D$ corresponde à $m_e - m_s$ et non pas à $m_s - m_e$ comme indiqué. En effet, ici, un bon traitement est un traitement qui fait *diminuer* la douleur, alors que, précédemment, il s'agissait d'*augmenter* le nombre de personnes guéries, ce qui compense le changement de signe. H_0 est donc rejetée, ce qui revient à affirmer que l'harpagophyton n'est pas inférieur (« est au moins aussi bon ») que la diacerhéine.

Discussion figurant dans la suite de l'article :

«

- Les anti-inflammatoires non stéroïdiens sont utilisés depuis longtemps pour lutter contre la douleur des patients arthrosiques.
- Leur efficacité n'est pas contestable, attestée par une multitude d'essais thérapeutiques...
- Plusieurs médicaments, les anti-arthrosiques d'effet différé et prolongé, ont récemment montré leur efficacité symptomatique dans l'arthrose sans démontrer pour le moment un effet structural. Ils sont utiles comme traitement de fond des patients arthrosiques souffrant de façon régulière et chez lesquels

les mesures non médicamenteuses ne suffisent plus à maîtriser les symptômes. Leur intérêt principal est de réduire la consommation d'antalgiques ou d'anti-inflammatoires...

— Il n'a pas été inclus de groupe placebo dans cette étude car il est souvent difficile de recruter des patients lorsqu'ils sont informés du risque (égal à 50 %) d'être dans un groupe placebo, et ce d'autant plus que la durée de traitement est particulièrement longue (quatre mois).

— Par ailleurs, l'appréciation de l'intensité de l'effet placebo dans l'arthrose a récemment été étudiée.

— Une revue portant sur 457 patients traités par placebo dans le cadre de six essais contrôlés d'anti-arthrosiques symptomatiques d'action lente a montré que sous placebo, la baisse moyenne de la douleur à l'EVA était de l'ordre de 10 à 16 mm.

— Par conséquent, la baisse observée dans le groupe harpagophyton (30,6 mm) est nettement supérieure à celle habituellement observée sous placebo. »

Commentaire :

— L'auteur rappelle ici les raisons pour lesquelles on utilise un test de non-infériorité, en particulier le caractère parfois non éthique de l'emploi d'un groupe témoin ne recevant qu'un placebo. Cependant, le test de non-infériorité tolère, malgré son nom, une légère déficience du traitement expérimental par rapport au traitement standard. Si ce dernier n'est supérieur que de peu à l'effet d'un placebo, est-ce que le traitement en cours d'expérimentation ne risque pas de recevoir son brevet de non-infériorité alors qu'il serait, en réalité, inférieur à un placebo ? L'angoisse du chercheur est ici dissipée par la connaissance d'expérimentations antérieures !
(Une question reste sans réponse : pourquoi les placebos ont-ils une certaine efficacité ?)

3 Tests d'équivalence

3.1 Usage d'un intervalle de confiance

Il a été indiqué au paragraphe 2.1 que, pour un test d'équivalence, on prend

— H_0 : « $\pi_s - \pi_e \geqslant \delta$ ou $\pi_e - \pi_s \geqslant \delta$ »

— H_1 : « $\pi_s - \pi_e < \delta$ et $\pi_e - \pi_s < \delta$ ».

Donc [8], en s'inspirant des idées des paragraphes 2.2.5 et 2.2.6, on obtient un test d'hypothèse de H_0 contre H_1 en rejetant H_0 lorsque l'intervalle de confiance de $\pi_s - \pi_e$ de niveau $1 - 2.\alpha$ est entièrement inclus dans $]-\delta\ ;\ \delta[$.

En effet, la probabilité, dans le cas où H_0 est vraie, de rejeter H_0 par erreur par cette méthode, se calcule de deux façons différentes, selon que la vraie valeur de $\pi_s - \pi_e$ est supérieure à δ ou bien inférieure à $-\delta$, ces deux cas étant les seuls possibles pour que H_0 soit vraie.

Dans les deux cas, le rejet ne peut avoir lieu, par définition, que si l'intervalle de confiance est inclus dans $]-\delta\ ;\ \delta[$.

Rappelons aussi que c'est l'intervalle de confiance qui est aléatoire, alors que $\pi_s - \pi_e$ est une valeur de la nature, non aléatoire

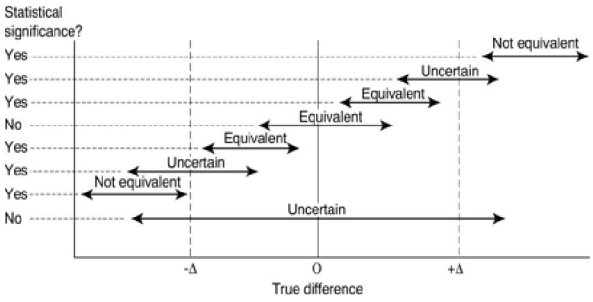
Dans le premier cas, cette probabilité est donc inférieure à la probabilité que tous les éléments de l'intervalle de confiance soient inférieurs à $\pi_s - \pi_e$, cette dernière probabilité étant par définition α pour un intervalle de confiance de niveau $1 - 2.\alpha$.

Dans le deuxième cas, cette probabilité est cette fois inférieure à la probabilité que tous les éléments de l'intervalle de confiance soient supérieurs à $\pi_s - \pi_e$, cette dernière probabilité étant aussi, par définition, α pour un intervalle de confiance de niveau $1 - 2.\alpha$.i

Par disjonction des cas, nous avons démontré qu'il s'agit bien d'un test de niveau **inférieur ou égal à** α de H_0 contre H_1 ; cependant, rien ne prouve qu'il ait une puissance optimale[2].

3.2 Des graphiques

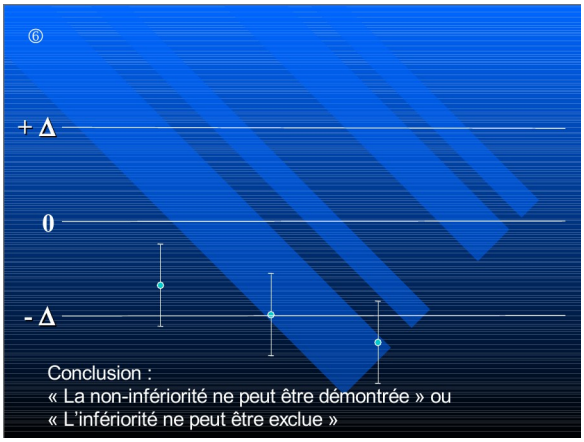
3.2.1 Université du Nord Texas



Ce graphique figure parmi les documents destinés aux étudiants de psychologie de l'université du Nord Texas. Il donne la vision que ces étudiants sont censés avoir concernant les tests d'équivalence.

3.2.2 Université de Rouen

Le graphique ci-dessous est extrait, parmi une dizaine d'autres du même genre, d'un polycopié de médecine de l'Université de Rouen [9]. Le but est le même.



3.3 Un exemple de test d'équivalence

Vérifions les calculs de l'article ci-dessous :

« Staszewski et al. [10] ont rapporté les résultats d'un essai clinique portant sur 562 patients atteints du VIH, conçu pour démontrer l'équivalence entre un traitement par abacavir, lamivudine et zidovudine d'une part et un traitement par indinavir, lamivudine et zidovudine d'autre part. Le critère principal était la proportion de patients ayant un niveau d'ARN du VIH de 400 copies / ml ou moins à la semaine 48. Sur la base de discussions avec des chercheurs, des cliniciens, et la Food and Drug Administration, la marge d'équivalence pour la différence des proportions avait été fixée à $\delta = 12$ %. Les taux de réponses positives à la maladie furent respectivement de 50,8% et 51,3% pour l'abacavir et l'indinavir. L'intervalle de confiance de niveau 95% pour la différence des les taux de réponses positives était $[-9, 8]$ et, comme cet intervalle est inclus dans $[-12, 12]$, les deux thérapies ont pu être déclarées équivalentes au risque $\alpha = 0,025$. »

— L'article indique également que 35 patients avaient été soumis à une intention de traitement mais n'avaient finalement pas pu suivre le protocole. Donc le nombre de patients traités a été $562 - 35 = 527$, dont 262 dans le groupe abacavir et 265 dans le groupe indinavir.

— Dans le groupe abacavir, le nombre de succès a été $262.0,508 \cong 133$; dans le groupe indinavir, il y a eu $265.0,513 \cong 134$ succès.

— La figure ci-dessous illustre la réalisation des calculs indiqués dans cet article à l'aide d'un tableur, les formules contenues dans certaines cellules étant affichées dans les cellules voisines. À droite figure la copie d'écran obtenue avec le logiciel R, qui se trouve reprise sous la figure.

	A	B	C	D	E	F	G
1		Effectif	Echec	Succès	%		
2	abacavir	262	129	133	0.508		
3	indinavir	265	129	136	0.513		
4	+ ou -	527	258	269	-0.00557		
5	Pas pris part	35					
6		562					
7	Z	1.960	LOI.NORMALE.STANDARD.INVERSE(0.975)				
8	Borne g :	-0.0909	E2-E3-B7*RACINE(E3*(1-E3)/B3+E2*(1-E2)/B2)				
9	Borne d :	0.0798	E2-E3+B7*RACINE(E3*(1-E3)/B3+E2*(1-E2)/B2)				

```
> x<-c(133,136)
> n<-c(262,265)
> prop.test(x,n,conf.level=0.95)

2-sample test for equality of proportions
with continuity correction
data:  x out of n
X-squared = 0.0017, df = 1, p-value = 0.9674
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.09472809  0.08358017
sample estimates:
   prop 1    prop 2
0.5076336 0.5132075
```

— On constate bien que l'intervalle de confiance, obtenu comme précédemment, est inclus dans $[-0, 12; 0, 12]$, autre écriture de $[-12\%; 12\%]$.

— Le logiciel R est un logiciel de statistique, agréable à utiliser avec un peu d'habitude ; voici comment nous lui faisons déterminer l'intervalle de confiance de cet exemple :

> x<-c(133,136)

> n<-c(262,265)

> prop.test(x,n,conf.level=0.95)

2-sample test for equality of proportions with continuity correction

data : x out of n
X-squared = 0.0017, df = 1, p-value = 0.9674
alternative hypothesis : two.sided

95 percent confidence interval :
-0.09472809 0.08358017
sample estimates :
prop 1 prop 2
0.5076336 0.5132075

— R ne trouve pas le même intervalle de confiance ! C'est qu'il a utilisé la formule conseillée en 1934 par le statisticien Frank Yates (1902, 1994).
Il est possible de ramener R à plus de naïveté :

> prop.test(x,n,conf.level=0.95,correct= FALSE)

2-sample test for equality of proportions without continuity correction

data : x out of n
X-squared = 0.0164, df = 1, p-value = 0.8982
alternative hypothesis : two.sided

95 percent confidence interval :
-0.09093290 0.07978498
sample estimates :
prop 1 prop 2
0.5076336 0.5132075

Et voilà !

3.4 Autre exemple de test d'équivalence

« Comparaison d'anesthésiques [11]. Helmy (1999) a comparé l'efficacité et la sécurité des effets anti-émétiques de l'ondansétron par voie intraveineuse avec ceux de la meto-clopramide, lors de leur administration préchirurgicale pour les patients qui a subi une cholécystectomie laparoscopique sous anesthésie générale intraveineuse (TIVA).
Dans cette étude, 80 patients ont été répartis au hasard en deux groupes recevant respectivement de l'ondansétron (4 mg) ou du dropéridol (1,25 mg), donnés en une seule dose intraveineuse immédiatement avant la mise en oeuvre d'une anesthésie générale standard ...
Un (autre) indice important pour la comparaison est le score relatif au "bien-être" du patient. Celui-ci est évalué sur une échelle comportant trois valeurs : "pauvre", "modéré", et

"confortable". Des études antérieures ont généralement suggéré l'absence de différence entre les scores de bien-être de l'ondansétron et du dropéridol.

Par conséquent, il est intéressant de déterminer sii, pour les *deux groupes*, "pauvre ou modéré" et "confortable" [2] Droperidol et Ondansetron sont équivalents en termes de bien-être. »

Score de bien-être du droperidol et de l'ondansetron		
	Score de bien-être	
	Pauvre ou modéré	Comfortable
Droperidol	12	28
Ondansetron	9	31

Man-Lai Tang, , Wai-Yin Poon, Hong Kong, 2006.

Appliquons la méthode que nous avons vue.

Borne gauche de l'intervalle de confiance de $\pi_X - \pi_Y$:

$t = \frac{28}{40} - \frac{31}{40} - z_{1-\alpha} * \sqrt{\left(\frac{\frac{28}{40}.(1-\frac{28}{40})}{40} + \frac{\frac{31}{40}.(1-\frac{31}{40})}{40}\right)}$

$\cong -0,24$

Borne droite :

$t = \frac{28}{40} - \frac{31}{40} + z_{1-\alpha} * \sqrt{\left(\frac{\frac{28}{40}.(1-\frac{28}{40})}{40} + \frac{\frac{31}{40}.(1-\frac{31}{40})}{40}\right)}$

$\cong 0,09$

Ici, si l'écart maximal $\delta = 0,1$ est utilisé, on ne peut rien conclure car l'intervalle de confiance, approximativement [-0,24 ; 0,09], n'est pas inclus dans [-0,1 ; 0,1].

Références

1 http://jacques.faisant.pagesperso-orange.fr/JN2014/
2 http://jacques.faisant.pagesperso-orange.fr/JN2014/Atelier_2014/PrEsentation2014_texte.pdf
3 http://unf3s.cerimes.fr/media/paces/Grenoble_1112/labarere_jose/labarere_jose_p06/labarere_jose_p06.pdf
4 Test construit à partir de http://www.sante.univ-nantes.fr/med/cidmef/diapo/TestStatistique_Principe.ppt
5 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2701110/ *ou bien*
da Silva GT, Klein JP. Methods for equivalence and noninferiority testing. Biol Blood Marrow Transplant. 2008 ;15 :120-7.
6 http://dumas.ccsd.cnrs.fr/dumas-00623076/document
7 L'harpagophyton dans le traitement de la gonarthrose et de la coxarthrose. Rev Rhum [Ed Fr] 2000 ;67 :634-40
8 Wilfred J. Westlake. Symmetrical Confidence Intervals for Bioequivalence Trials. Biometrics. Déc 1976 ;32 :741-744
9 http://medecine-pharmacie.univ-rouen.fr/servlet/com.univ.collaboratif.utils.LectureFichiergw?ID_FICHIER=9935
10 Abacavir-Lamivudine-Zidovudine vs Indinavir-Lamivudine-Zidovudine in Antiretroviral-Naive HIV-Infected Adults. JAMA. 2001 ;285(9) :1155-1163. http://jama.jamanetwork.com/article.aspx?articleid=193616
11 Statistical inference for equivalence trials with ordinal responses. Computational Statistics & Data Analysis 51 (2007) 5918-5926

Bibliographie

B1 Fisher, Neyman, and the Creation of Classical Statistics, Erich L. Lehmann, 2011, Springer
— (en ligne : http ://bookzz.org/md5/FC1B6E96A1AEE1538A6EA73D65288E9E)
B2 Testing Statistical Hypotheses, Erich Leo Lehmann, 1959, John Wiley & Sons
B3 Méthodes statistiques, Philippe Tassi, 1985, Economica

2. Modification du texte original de l'article : deux modalités ont été regroupées.