Notes de cours Biostatistiques – MIV (L3)
Introduction à l'analyse de puissance

M. Bailly-Bechet – d'après le cours de S. Champely

Université Claude Bernard Lyon 1 – France

Ce cours est une introduction destinée à présenter les concepts de base de l'analyse de puissance. Pour une analyse plus détaillée pour les principaux tests classiques paramétriques, les étudiants sont invités à consulter le polycopié de S. Champely, disponible à l'adresse `http://pbil.univ-lyon1.fr/R/puissance.pdf`.

1 Analyse de puissance : concepts de base

Un test est une règle de décision entre deux hypothèses H_0 et H_1, respectivement nommées hypothèse nulle (choisie par défaut) et hypothèse alternative. La pratique du test consiste à calculer une statistique, puis à estimer la chance d'observer une telle valeur de la statistique (ou une valeur encore plus extrême) sous l'hypothèse H_0. Cette probabilité, la p-value, est ensuite comparée à un seuil de décision fixé à l'avance, α. Si la p-value est inférieure à α, on rejettera H_0 au profit de H_1, en argumentant qu'observer une telle valeur de la statistique calculée est trop peu probable au regard du risque de première espèce que l'on est prêt à prendre. Ce risque – la valeur de α – représente le risque que l'on s'autorise à avoir pour rejeter par erreur H_0 alors que cette hypothèse est vraie. Il est toujours choisi très faible, le raisonnement scientifique étant mu par l'idée de ne pas ajouter de complexité inutile dans les modèles.

Il existe une autre erreur possible : l'erreur de deuxième espèce, notée β. C'est la probabilité de conserver à tort H_0 alors que H_1 est vraie. Cette valeur est souvent plus difficile à calculer, mais est également très importante pour le raisonnement scientifique : si β est très grand, cela revient à dire que le

test pratiqué a de grandes chances de conserver l'hypothèse H_0, qu'elle soit vraie ou non. Dans ce cas, faire un test est relativement inutile, puisque la réponse est "presque" connue à l'avance...

On veut donc minimiser la valeur de β, tout en gardant une valeur de α aussi basse que possible. La minimisation des deux valeurs simultanément n'est pas possible [1], mais il est par contre possible, dans un cadre expérimental donné, de calculer explicitement la valeur de β à α fixé, en fonction des paramètres de l'expérience (taille d'échantillon, etc...). En pratique, on calculera souvent $1-\beta$, que l'on appelle la *puissance* du test, et que l'on veut maximiser.

2 Exemple sur le test de comparaison de moyennes.

2.1 Présentation du problème

Supposons que l'on s'intéresse à un test de VO_2Max (Consommation maximale en oxygène, une mesure de la "caisse" d'un individu) dans une population âgée. On suppose, grâce à de précédentes études populationnelles, que cette variable suit une loi normale de moyenne $\mu_0 = 25.5$ et d'écart-type $\sigma = 6$ (ml/kg/min).

On pense qu'une population de personnes atteintes de la maladie de Parkinson doit avoir, outre les tremblements bien connus, des capacités cardio-respiratoires plus limitées. On souhaite donc tester si dans un tel groupe l'espérance mathématique μ est plus faible. Le principe du test est donc de décider entre deux hypothèses : l'*hypothèse nulle* notée $H_0 : \mu \geq 25.5$ et l'*hypothèse alternative* notée $H_1 : \mu < 25.5$. Il s'agit d'un test unilatéral, comme souvent dans le cadre d'expériences scientifiques.

Remarquons tout de suite qu'on a choisi de poser comme hypothèse nulle l'absence d'effet et comme hypothèse alternative son existence et *qu'on s'est bien gardé de donner une taille quelconque à l'effet* (l'espérance diminue de 1, 2, ou 5 ?).

1. en raison d'arguments théoriques non exposés ici, mais qui peuvent se résumer en disant que quand l'une des deux erreurs diminue, l'autre augmente.

2.2 Statistique de décision

On va supposer que l'on a accès à $n = 15$ sujets dans cette expérience. On note x_i les valeurs des VO_2Max mesurées, et $\bar{x}$ leur moyenne.

On est dans le cadre d'une comparaison de moyennes entre un échantillon et une valeur de référence μ_0. La variance σ^2 est connue. Cet exemple, quoique artificiel, va permettre de présenter la démarche de l'analyse de puissance. La statistique du test est :

$$\epsilon_{obs} = \frac{\|\bar{x} - \mu_0\|}{\sigma/\sqrt{n}} \tag{1}$$

Cette statistique suit une loi normale (voir cours sur les tests paramétriques) sous H_0. Si on veut savoir à partir de quelle valeur observée de $\bar{x}$ on concluera à un effet de la maladie sur la VO_2Max, il faut renverser cette formule. On observera un effet au seuil α si :

$$P\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq \frac{c_\alpha - \mu_0}{\sigma/\sqrt{n}}\right) = \alpha, \tag{2}$$

avec c_α la valeur critique en dessous de laquelle on choisira de rejeter H_0 au risque α. En notant le quantile de la loi normale centrée réduite ϵ_α, on obtient :

$$c_\alpha = \mu_0 + \epsilon_\alpha \frac{\sigma}{\sqrt{n}} \tag{3}$$

Pour les valeurs numériques données plus haut et un seuil à $\alpha = 5\%$, on a ϵ_α=-1.645, et on obtient un effet si pour $c_\alpha = 22.95$, soit un effet si $\bar{x} \leq 22.95$.

2.3 Calcul de la puissance

En résumé, on va calculer la statistique de test $\bar{x}$. Si elle est plus grande que $c_\alpha = 22.95$ on décidera de conserver l'hypothèse nulle. Si elle est plus petite, on décidera de rejeter l'hypothèse nulle et on dira que le résultat est *statistiquement significatif au seuil* α.

Si nous sommes effectivement dans le cadre de l'hypothèse nulle, nous savons que nous risquons de nous tromper dans 5% des cas, c'est le risque α que nous avons pris en choisissant le niveau de significativité conventionnel.

Maintenant nous allons poser la question un peu moins conventionnelle : "Mais que se passe-t-il si nous sommes effectivement dans le cadre de l'hypothèse alternative ? Quel risque prenons-nous ?". Il faut choisir dans quelle mesure on s'écarte de l'hypothèse nulle, c'est ce qu'on appelle la *taille d'effet*. C'est une décision qui se prend à partir de considérations scientifiques. Il faut se demander en particulier à partir de quelle taille un effet constitue une différence scientifiquement significative. La consultation avec un expert du domaine est à ce niveau nécessaire... En effet, on imagine bien que si l'hypothèse alternative $H_1 : \mu = 25.499999$ est vraie, on ne pourra pas distinguer par notre test H_0 et H_1 : dans ce cas notre test sera peu puissant, et on acceptera toujours l'hypothèse nulle, la taille d'effet étant trop faible pour être détectée par le test.

On supposera qu'un spécialiste nous répond qu'à partir de 23.5 points l'effet peut être considéré comme important.

Calculons alors la probabilité, si $H_1 : \mu = 23.5$ est vraie – et donc que l'effet est scientifiquement intéressant – que l'on rejette effectivement H_0 :

$$\begin{aligned} 1-\beta &= P(\bar{x} < 22.95) \\ &= P\left(\frac{\bar{x} - 23.5}{6/\sqrt{15}} < \frac{22.95 - 23.5}{6/\sqrt{15}}\right) \\ &= P(\mathcal{N}(0,1) < -0.355) \\ &= 0.36 \end{aligned}$$

On constate sur cet exemple que l'on a une très faible chance de démontrer ce qui nous intéresse. On dit alors que la puissance de ce test n'est pas satisfaisante.

2.4 Taille d'effet

On peut facilement voir dans ce calcul que la puissance, $1 - \beta$, augmente quand l'hypothèse alternative H_1 s'éloigne de H_0 : si on choisit $H_1 : \mu = 22.95$, la puissance $1 - \beta$ devient par construction 0.5, et elle continue à augmenter quand la valeur réelle de μ décroît. Cela implique qu'un test est toujours plus performant pour détecter de grandes différences que de petites différences : on parle de taille d'effet. Plus la différence entre H_0 et H_1 augmente, plus on a de chances de les distinguer avec un test. Ceci est illustré sur la figure 1.

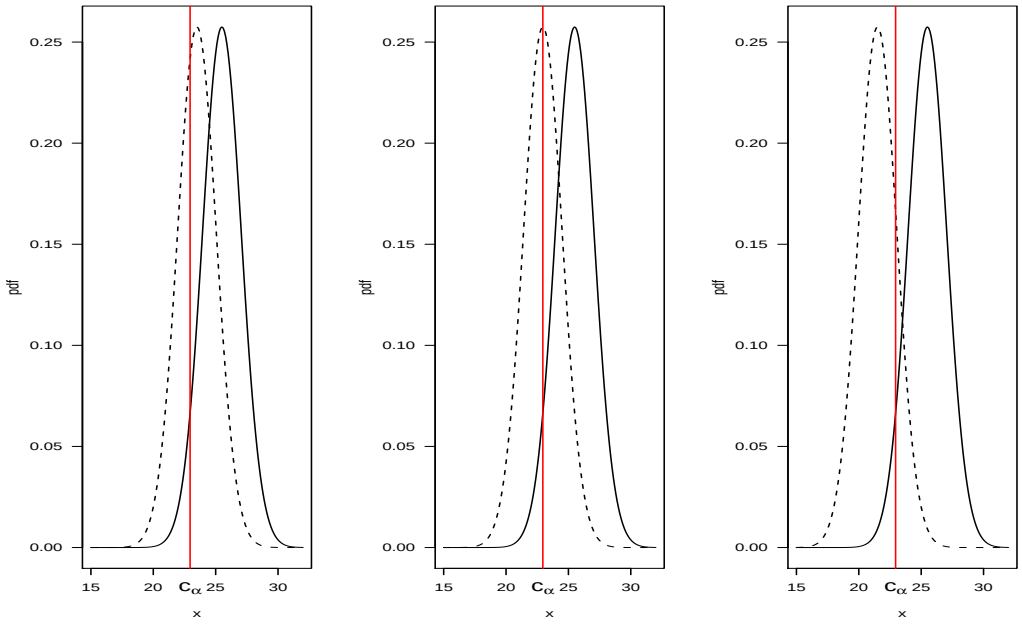

FIGURE 1 – Courbes de densité de probabilité pour $H_0 : \mu_0 = 25.5$ (ligne continue) et pour 3 hypothèses alternatives différentes (ligne pointillée) ; de gauche à droite $\mu = 23.5$, 22.95, 21.5. La ligne rouge représente la valeur critique pour $\alpha = 5\%$.

2.5 Taille d'échantillon

Une puissance de 0.36 est insuffisante : dans 64% des cas, si H_1 est vraie, on concluera néanmoins que H_0 est la bonne réponse. On veut donc augmenter cette puissance. Une manière de faire est d'augmenter la taille de l'échantillon. En effet, on a vu que :

$$1 - \beta = P\left(\mathcal{N}(0,1) < \frac{c-\mu}{\sigma/\sqrt{n}}\right). \tag{4}$$

Si on exige une puissance de 80%, avec $H_1 : \mu = 23.5$, on veut trouver n tel que :

$$0.8 = P\left(\mathcal{N}(0,1) < \frac{22.95 - 23.5}{6/\sqrt{n}}\right) \tag{5}$$

$$\frac{22.95 - 23.5}{6/\sqrt{n}} = 0.85 \tag{6}$$

$$\frac{1}{\sqrt{n}} = 0.107 \tag{7}$$

$$n \geq 86. \tag{8}$$

Avec $n = 86$ sujets, on aurait, dans le même cadre expérimental, la puissance nécessaire pour réaliser correctement le test. De grands échantillons permettent donc de mieux distinguer des différences, tout comme de grandes différences. Plus la différence scientifiquement intéressante sera faible, plus il faudra un grand échantillon pour arriver, avec un test, à la faire ressortir. On peut voir cette progression sur la figure 2.

Attention ! Il est classique, en génomique par exemple, de disposer de milliers, voire de millions de points à comparer. Dans ces conditions, la plus infime différence sera détectée par un test, et à la question "Ces deux échantillons proviennent-ils de la même population ?", la réponse donnée par le test sera quasiment toujours négative. Mais dans ce cas, la question intéréssante à poser est : est ce que cette différence entre deux échantillons représente bien une variation intéréssante du phénomène que l'on étudie ? Ce n'est pas toujours le cas...

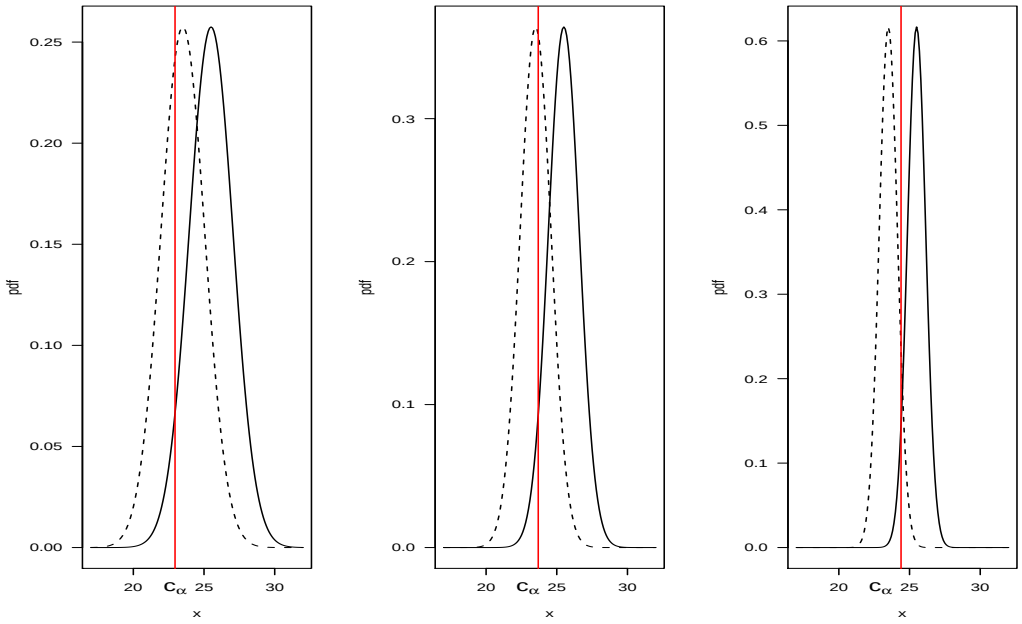

FIGURE 2 – Courbes de densité de probabilité pour $H_0 : \mu_0 = 25.5$ (ligne continue) et $H_1 : \mu = 23.5$ (ligne pointillée) pour 3 tailles d'échantillon différentes ; de gauche à droite $n = 15, 30, 86$. La ligne rouge représente la valeur critique pour $\alpha = 5\%$, qui varie avec n dans ce cas.