Tests statistiques paramétriques : taille d'effet, puissance et taille d'échantillon (package pwr)

Stéphane CHAMPELY
Université Lyon 1, France.

13 décembre 2017

Table des matières

1 Introduction **5**

2 Concepts de base **7**
2.1 Rappels sur les tests paramétriques 7
2.2 La taille d'effet 8
2.3 La puissance d'un test 9
2.4 La taille d'échantillon 10
2.5 Les types d'analyse de puissance 13

3 Tests t de moyennes **15**
3.1 Le test t à un échantillon 15
3.1.1 Rappels sur le test t 15
3.1.2 La taille d'effet 16
3.1.3 La puissance 17
3.1.4 La taille d'échantillon 18
3.2 Le test t à deux échantillons appariés 20
3.2.1 Rappels sur le test 20
3.2.2 La taille d'effet 20
3.2.3 La puissance 20
3.2.4 La taille d'échantillon 21
3.3 Le test t à deux échantillons indépendants 22
3.3.1 Rappels sur le test 22
3.3.2 La taille d'effet 22
3.3.3 La puissance 23
3.3.4 La taille d'échantillon 24

4 Tests de proportions **27**
4.1 Le test de proportion à un échantillon 27
4.1.1 "Rappels" sur le test 27
4.1.2 La taille d'effet 28
4.1.3 La puissance 29
4.1.4 La taille d'échantillon 30
4.2 Le test de proportion à deux échantillons indépendants 30
4.2.1 Rappels sur le test 30

4.2.2 La taille d'effet 30
4.2.3 La puissance 31
4.2.4 La taille d'échantillon 32

5 Test de corrélation **35**
5.1 Rappel du test 35
5.2 La taille d'effet 35
5.3 La puissance 36
5.4 La taille d'échantillon 37

6 Test d'analyse de variance **39**
6.1 Rappel du test 39
6.2 La taille d'effet 39
6.3 La puissance 40
6.4 La taille d'échantillon 41

7 Tests du chi-carré **43**
7.1 Rappels des tests 43
7.1.1 Le test d'ajustement 43
7.1.2 Le test d'indépendance 44
7.2 La taille d'effet 44
7.3 La puissance 45
7.4 La taille d'échantillon 47

8 Tests dans le modèle linéaire général **49**
8.1 Rappels sur les tests 49
8.2 Tailles d'effet 50
8.3 Puissance des tests 50
8.4 Taille d'échantillon 50

Chapitre 1

Introduction

On utilise un test paramétrique afin de démontrer l'*existence d'un effet* (différence entre deux moyennes, présence d'une corrélation). Suite à la collecte de données et aux calculs de statistiques sur l'échantillon recueilli, une décision est prise, grâce aux méthodes de la statistique inférentielle, concernant l'existence ou l'absence de cet effet. Si la conclusion est l'existence de l'effet, le résultat est dit *statistiquement significatif*.

Cette démarche à présent classique passe sous silence un certain nombre de questions essentielles :

- Quel est le rapport entre un effet *statistiquement significatif* et un effet *scientifiquement significatif*, c'est-à-dire qui intéresse concrètement (et pas seulement pour publier) le praticien ? Ces deux notions peuvent-elles ne pas coïncider ?
- Admettons que cet effet existe, quelle est sa taille ?
- On prend grand soin de ne pas conclure à l'existence d'effet lorsqu'il n'existe pas dans la démarche du test statistique (c'est le rôle de l'*erreur de première espèce*), mais prend-on également garde à ne pas conclure à l'absence d'effet lorsqu'il existe ? Comment quantifier ce risque ? C'est le rôle de l'*erreur de deuxième espèce* ou inversement de la *puissance du test*.
- Lorsque l'effet est concrètement important, on imagine bien qu'il faut moins d'observations pour le démontrer que lorsqu'il est petit mais combien au juste ? A-t-on les moyens, en termes de nombre de mesures, de démontrer ce que l'on cherche ? Faut-il s'y prendre autrement et changer le dispositif de son observation/expérimentation ?

Ces questions sont en pratique fondamentales mais l'enseignement initial classique de la statistique ne les aborde pas souvent, par manque de temps probablement mais peut-être aussi parce qu'elles remettent au centre du débat une chose difficile à gérer : le réel et la nécessaire complémentarité du statisticien avec le praticien [1].

1. qui peuvent être la même personne...

L'objectif de ce cours est
- de donner les idées théoriques essentielles des analyses de puissance,
- les aspects pratiques de leur utilisation à l'aide d'exemples souvent issus du monde sportif ainsi que
- les moyens de les réaliser informatiquement.

La construction de ce cours est volontairement inspirée de l'ouvrage classique de Cohen [4] et ses notations ont été en particulier reprises (elles sont classiques dans le domaine de l'analyse de puissance) ; nombre de ses exemples sont donnés en exercices afin de pouvoir comparer avec un ouvrage de référence.

En ce qui concerne la mise en oeuvre informatique, il existe un certain nombre de logiciels dont certains sont d'ailleurs gratuits. Pour ma part, j'ai décidé d'employer le logiciel libre de distribution R [8] qui comprend quelques fonctions spécifiques (power.t.test, power.prop.test et power.anova.test) que j'ai modifiées afin qu'elles deviennent conformes aux notations de Cohen [4]. Ces fonctions sont réunis dans le package pwr 1.2.1 disponible sur le site http ://cran.r-project.org/. L'installation et le chargement de ce package sont indispensables à une lecture active de ce cours.

Chapitre 2

Concepts de base

Un exemple tout à fait artificiel (un test de la loi normale *où l'on connaît l'écart-type*) va permettre de rappeler le principe d'un test paramétrique, d'exposer les notions de taille d'effet, de puissance et de taille d'échantillon dans un contexte simple où les calculs sont réalisables presque "à la main".

2.1 Rappels sur les tests paramétriques

Supposons que l'on s'intéresse à un test de VO2Max (Consommation maximale en oxygène, une mesure de la "caisse" d'un individu) dans une population âgée comme c'est le cas dans [9]. On suppose également que les mesures suivent une loi normale et que l'écart-type est $\sigma = 6$ (ml/kg/min). Pour un groupe de contrôle, il a été montré que l'espérance mathématique est de l'ordre de $\mu = 25.5$.

On pense qu'une population de malades (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 $H_0 : \mu = 25.5$ et l'*hypothèse alternative* notée $H_a : \mu < 25.5$.

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 ?) On va supposer qu'on décide d'employer $n = 15$ sujets dans cette expérience (mais pourquoi pas 20, ou 50 ?). On voit ici que l'on passe sous silence dans les exposés habituels deux questions essentielles.

La statistique que l'on emploie pour prendre une décision est la moyenne de l'échantillon recueilli notée $\overline{Y}$. La position de cette statistique par rapport à 25.5 va nous permettre de choisir entre les deux hypothèses.

Il faut compter avec les variations d'échantillonnages, c'est-à-dire qu'il est tout à fait possible que l'espérance mathématique soit plus grande (ou égale à 25.5) et que la statistique soit plus petite... Inversement, que l'espérance mathé-

matique soit plus petite que 25.5 et qu'on observe une valeur de la statistique plus grande. On va donc fixer un seuil de décision c qui permet de dire qu'en deça de ce seuil on décidera l'alternative et qu'au delà on décidera l'hypothèse nulle. La région définie pour la statistique par $W = \{\overline{Y} < c\}$ est dite *région critique* ou région de rejet.

La loi de la statistique $\overline{Y}$ est bien connue, *à partir du moment où μ l'est* (et σ aussi), c'est une loi normale d'espérance mathématique μ et d'écart-type $\frac{\sigma}{\sqrt{n}}$. On peut donc calculer les deux types d'erreurs : (1) celle qu'on commet en décidant l'alternative alors que la véritable situation est celle de l'hypothèse nulle, on parle d'*erreur de première espèce* et (2) celle où l'on décide l'hypothèse nulle alors que c'est l'alternative qui est vraie, on parle d'*erreur de deuxième espèce*.

Dans le cadre des deux hypothèses postulées ici, si on décide de diminuer la valeur du seuil c on va diminuer l'erreur de première espèce mais on augmentera celle de deuxième. Inversement si on décide d'augmenter c on diminuera l'erreur de deuxième espèce mais on augmentera celle de première. Il faut donc réaliser un arbitrage entre ces deux erreurs.

Depuis les travaux de Neyman et Pearson, on choisit de limiter l'erreur de première espèce à un niveau dit *niveau de significativité* que l'on note généralement α (et qui, le plus souvent, est conventionnellement égal à 0.05). Ceci signifie que sous l'hypothèse nulle, la taille de la région critique W est choisie pour avoir une probabilité de α.

On a donc $P(W|H_0) = P(\overline{Y} < c|H_0) = \alpha$, c'est-à-dire : $P(W|H_0) = P(\frac{\overline{Y}-25.5}{6/\sqrt{15}} < \frac{c-25.5}{6/\sqrt{15}}) = \alpha$. En notant z_α le quantile de la loi normale standard nous obtenons $\frac{c-25.5}{6/\sqrt{15}} = z_\alpha$ c'est-à-dire $c = 25.5 + z_\alpha \times \frac{6}{\sqrt{15}}$. En choisissant le seuil conventionnel de $\alpha = 0.05$, nous avons $z_\alpha = -1.645$ donc $c = 22.95$.

```
25.5+qnorm(0.05)*6/sqrt(15)
```

2.2 La taille d'effet

En résumé, on va calculer la statistique de test $\overline{Y}$. Si elle est plus grande que $c = 22.95$ 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* théorique. Pour l'heure, nous allons éluder le problème de la détermination de cette valeur

et la considérer comme donnée (et exacte!). Nous reviendrons sur ce difficile problème. On suppose donc que la vraie valeur est de 23.5 (ml/kg/min). Ce qui nous intéresse est l'écart à l'hypothèse nulle, soit : $23.5 - 25.5 = -2$.

Remarque 1 *Cela nous ramène à une écriture où l'hypothèse nulle est équivalente à ce que la taille d'effet soit nulle aussi.*

Remarque 2 *Généralement, on standardise également cet effet afin d'obtenir un nombre « sans dimension », c'est-à-dire qui ne dépend pas des unités de mesures.*

La taille d'effet sera donc $d = \frac{23.5-25.5}{6} = -1/3$.

Définition 1 *Une taille d'effet évalue dans quelle mesure on s'écarte de l'hypothèse nulle ($H_0 : \mu = \mu_0$). Dans ce test, elle est égale à*

$$d = \frac{\mu - \mu_0}{\sigma}.$$

Remarque 3 *Même si les calculs sont basés sur une seule quantité, la taille d'effet, elle dépend en fait de plusieurs éléments (espérance et écart-type par exemple).*

2.3 La puissance d'un test

Définition 2 *La puissance d'un test est la probabilité de rejeter l'hypothèse nulle à raison c'est-à-dire lorsqu'on est «en vérité» dans le cadre de l'hypothèse alternative.*

La puissance du test est donc le complément de l'erreur de deuxième espèce β. C'est pourquoi on la note $1 - \beta$.

On va donc rejeter l'hypothèse nulle si $\overline{Y} < 22.95$. Quelle est la probabilité de cet événement lorsque nous sommes dans le cadre de l'hypothèse alternative et, plus précisément, avec une taille d'effet de $d = -\frac{1}{3}$?

$$\begin{aligned} P(W|H_1) &= P(\overline{Y} < 22.95|H_1) \\ &= P(\frac{\overline{Y} - 23.5}{6/\sqrt{15}} < \frac{22.95 - 23.5}{6/\sqrt{15}}) \\ &= P(N(0,1) < -0.355) \\ &= 0.36 \end{aligned}$$

```
(22.95-23.5)/(6/sqrt(15))
pnorm(-.355)
```

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.

Remarque 4 *On considère généralement que la puissance doit au moins être égale à* 0.80 *pour être satisfaisante.*

Exercice 1 *Quelle est la taille d'effet lorsqu'on postule que* $\mu = 21$ *? Quelle est la puissance correspondante ?*

En écrivant de façon plus formelle les calculs précédents et en supposant qu'on s'intéresse à une alternative où l'espérance mathématique est μ lorque l'hypothèse nulle est définie par μ_0, soit une taille d'effet de $d = \frac{\mu - \mu_0}{\sigma}$, la puissance $1 - \beta$ est donnée par la formule

$$1 - \beta = P(N(0,1) < z_\alpha - d \times \sqrt{n})$$

On peut donc facilement calculer la puissance (à l'aide de la fonction pwr.norm.test) pour une séquence de tailles d'effet, ce qui donne la *courbe de puissance du test*[1]. La courbe de puissance est représentée dans la figure 2.1.

```
pwr.norm.test(sig.level=0.05,d=-1/3,n=15,alternative="less")
x<-seq(15,30,l=100)
d<-(x-25.5)/6
plot(x,pwr.norm.test(sig.level=0.05,d=d,n=15,alternative="less")$power,
type="l",xlab="mu",ylab="power")
```

On voit sur le graphique que la taille d'effet correspondant à une puissance "satisfaisante" de 0.80 est située vers $\mu = 21.5$. On peut utiliser la fonction pwr.norm.test pour réaliser le calcul exact : $d = -0.642$ donc $\mu = 21.65$.

```
pwr.norm.test(power=0.80,sig.level=0.05,n=15,alternative="less")
-0.642*6+25.5
```

Remarque 5 *Le graphique 2.2 permet de voir la fonction de puissance d'un test bilatéral avec les spécifications précédentes.*

2.4 La taille d'échantillon

Supposons que l'on souhaite à présent détecter l'effet : $d = -\frac{1}{3}$. Nous avons vu précédemment que la puissance correspondante est de $1 - \beta = 0.36$. On n'a même pas une chance sur deux de montrer ce qui nous intéresse...

Maintenant, il reste encore un élément sur lequel on peut "jouer" : n, la taille de l'échantillon. En décidant d'employer une puissance "satisfaisante" de $1 - \beta = 0.80$, on doit résoudre l'équation

$$1 - \beta = P(N(0,1) < z_\alpha - d \times \sqrt{n})$$

1. On parle aussi parfois de façon équivalente de *courbe des caractéristiques opérationnelles* qui est celle de l'erreur de première espèce β

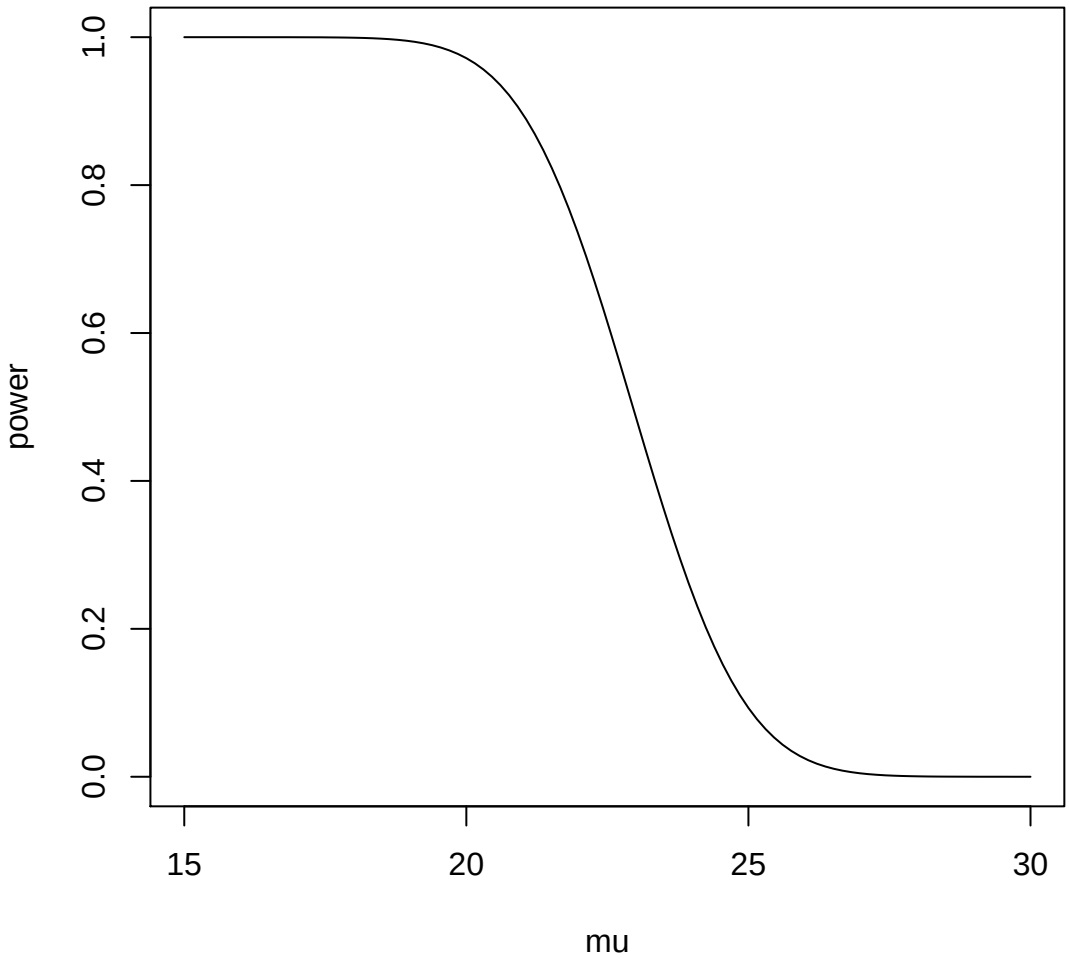

FIGURE 2.1 – Courbe de puissance en fonction de l'espérance mathématique pour un test unilatéral

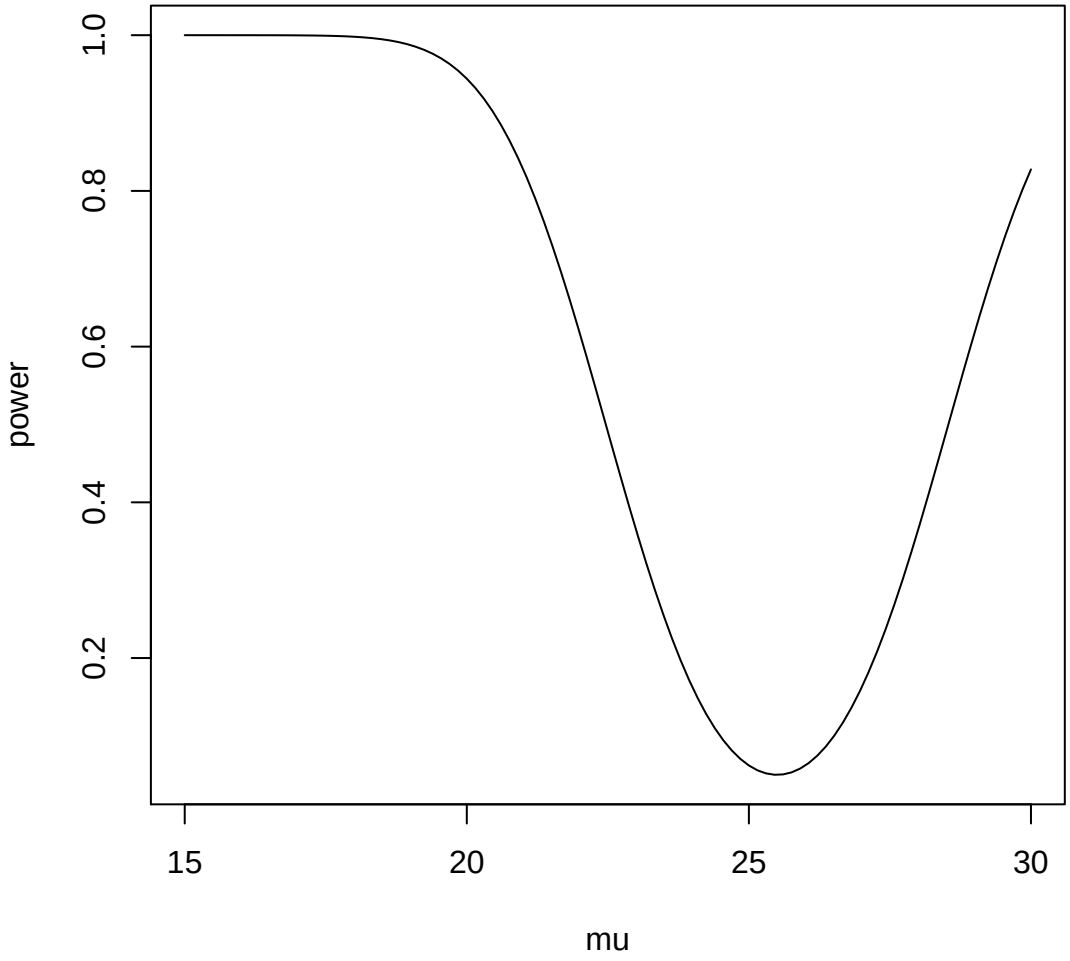

FIGURE 2.2 – Courbe de puissance en fonction de l'espérance mathématique pour un test bilatéral

ce qui est assez simple :

$$n = \left(\frac{z_{1-\beta} + z_{1-\alpha}}{d} \right)^2$$

Il faut ici une taille d'échantillon de $n = 56$ pour parvenir à atteindre une telle puissance. La résolution de l'équation est réalisée de façon complètement numérique par la fonction pwr.norm.test.

```
pwr.norm.test(sig.level=0.05,power=0.8,d=-1/3,alternative="less")
```

Remarque 6 *On peut remarquer que le raisonnement précédent s'applique aussi avec la version bilatérale du test de moyenne. La fonction pwr.norm.test réalise les calculs dans ce cas en utilisant l'option alternative="two.sided".*

Exercice 2 *Dans le cas du test bilatéral, calculer à l'aide de la fonction pwr.norm.test la taille d'échantillon nécessaire pour repérer avec une puissance de $1-\beta = 0.80$ une taille d'effet de $d = 1/3$ (avec le niveau conventionnel de $\alpha = 0.05$) ?*

Pourquoi cette taille d'échantillon est-elle plus grande que dans le cas unilatéral ?

Exercice 3 *Le fichier survey du package MASS contient différentes mesures concernant un groupe d'étudiants. On s'intéressera à leur taille (variable Height) mais uniquement pour les hommes (catégorie "Male" de la variable Sex). On va supposer que la taille suit une loi normale d'écart-type $\sigma = 10$.*

On veut tester si l'espérance mathématique peut-être égale à $\mu_0 = 173$.

- *Réalisez le test au niveau 5% (en unilatéral pour une alternative "greater").*
- *Quelle est la puissance du test correspondant à $\mu = 175$?*
- *Quelle est la taille d'effet 8emphobservée sur cet échantillon ?*
- *Pour détecter une taille d'effet correspondant à $\mu = 174$, quelle taille d'échantillon est nécessaire ?*

2.5 Les types d'analyse de puissance

Une analyse de puissance peut prendre plusieurs formes :

- on connait le niveau du test, la taille d'échantillon et la taille d'effet et on cherche à calculer la puissance. Ceci permet de voir si notre dispositif expérimental est bien calibré.
- on connait la puissance voulue, le niveau du test et la taille d'effet à détecter. On cherche alors à calculer la taille d'échantillon nécessaire pour monter un dispositif expérimental efficace.
- une fois les données collectées (*post-mortem* donc), il est possible également de calculer la puissance pour certains effets, et il est aussi possible de calculer la taille de l'effet mesuré sur l'échantillon. On arrive ainsi parfois à des contradictions entre la significativité statistique et la significativité scientifique.

Remarque 7 *Derrière les questions (techniques) autour de la puissance d'un test, il convient de pas oublier d'autres questions (voir Lenth[7]) :*

- *Est-ce que l'étude est bien aléatoire (randomisation) ?*
- *Devons-nous nous attendre à des non-réponses ?*
- *Que mesure-t-on et comment ? N'y a-t-il pas des méthodes alternatives ?*
- *Quelles sont les sources de variation (sexe, âge...) ? Peut-on les prendre en compte dans le dispositif expérimental (par des blocs, des covariables) ?*
- *Combien de temps pour faire l'expérience ? Quelles sont les contraintes pratiques ?*
- *Les données suivent-elles le modèle probabiliste proposé ?*

Chapitre 3

Tests t de moyennes

3.1 Le test t à un échantillon

3.1.1 Rappels sur le test t

Nous sommes dans le cadre d'un échantillon issu d'une loi $N(\mu, \sigma)$. Cette fois-ci, l'écart-type σ est inconnu. On veut tester les hypothèses (version unilatérale ici) : $H_0 : \mu \leq c$ contre $H_a : \mu > c$.

Dans une étude concernant les souhaits d'équipements des usagers des piscines lyonnaises [11], était demandé sur une échelle de Likert à 5 niveaux (1, 2, 3, 4 et 5) si un sauna était désiré. On veut voir si le niveau de souhait est supérieur à la position neutre $c = 3$. On va supposer que $n = 25$ individus ont accepté de répondre [1].

Le test repose sur le calcul de la moyenne $\overline{Y}$ et de l'écart-type S de l'échantillon puis de la statistique

$$T = \frac{\overline{Y} - c}{S/\sqrt{n}}.$$

Sous l'hypothèse nulle, cette statistique T suit une loi de Student à $n - 1$ degrés de liberté. On peut donc définir la région critique comme étant :

$$W = \{T > q_t(1 - \alpha, n - 1)\}$$

où $q_t(1 - \alpha, n - 1)$ est le quantile correspondant.

On a par exemple pour $n = 25$ et le niveau conventionnel $\alpha = 0.05$ la région critique $W = \{T > 1.711\}$. On rejettera l'hypothèse nulle si la valeur de la statistique T calculée sur l'échantillon appartient à cet ensemble.

```
qt(1-0.05,df=25-1)
```

1. Il est évident que dans cette situation l'hypothèse de la normalité est entendue comme étant une approximation ; une approximation grossière.

3.1.2 La taille d'effet

La taille d'effet est définie comme précédemment c'est-à-dire

$$d = \frac{\mu - c}{\sigma}.$$

On voit que pour l'hypothèse nulle la taille d'effet correspondante est nulle, elle mesure donc bien l'éloignement à cette hypothèse. De plus, cette quantité est indépendante des unités de mesure (on exprime les écarts en unités de variabilité).

Comment choisir d ? Il est préférable de le faire en partant de considérations scientifiques basées sur l'expérience du praticien [2]. En réalité, c'est bien difficile, et on se dirige plutôt sur l'emploi de données historiques, tirées de la littérature [3]. Mieux, une étude pilote (de petite taille, ne dilapidant pas notre budget) permet d'en avoir une estimation raisonnable. Il faut bien comprendre, et nous le soulignerons à plusieurs reprises, que la taille d'effet conditionne notre capacité à obtenir des résultats significatifs. Sa détermination, *au moins approximative*, est donc fondamentale.

Dans l'étude de souhaits qui nous intéresse, on va supposer que nous allons prendre un écart-type probablement trop grand, qui correspond à la loi uniforme sur les cinq niveaux soit $\sigma = \sqrt{(2)} = 1.41$. On va considérer comme effet intéressant que le niveau moyen de souhait soit de 3.5.

Remarque 8 *Il s'agit de ce qu'on appelle une taille d'effet prescrite. Dans [6], les dangers de cette approche, pourtant répandue, sont soulignés : Prescrire ce qui est intéressant ne nous dispense pas de la réalité ! Il se peut que la véritable valeur soit toute autre et dans ce cas nos calculs de puissance ou de taille d'échantillon sont tout simplement fantaisistes !*

Ceci conduit donc à une valeur de taille d'effet de $d = \frac{3.5-3}{\sqrt{(2)}}$ soit environ $d = 0.35$.

La dernière méthode, proposée par Cohen [4] est d'utiliser des niveaux *conventionnels* de taille d'effet. Mais il faut bien comprendre, et il le souligne, qu'ils devraient être *conventionnels au domaine d'étude*. Il propose d'employer les trois niveaux suivants :

- $d = 0.2$ faible effet (qui correspond par exemple à la différence de taille entre des filles de 15 et 16 ans)
- $d = 0.5$ effet moyen (différence de taille entre filles de 14 et 18 ans ou différence de QI entre employés et managers...)
- $d = 0.8$ effet fort (différence de taille entre filles de 13 et 18 ans ou différence de QI entre lycéens et titulaires d'un doctorat)

C'est donc un effet de faible à moyen que nous souhaitons étudier.

2. et Cohen [4] montre plusieurs façons de la "faire parler" s'il est mal à l'aise avec la formulation directe de la taille d'effet

3. mais la littérature décrit-elle souvent une expérience comparable ?

Remarque 9 *A nouveau, la taille d'effet est prescrite dans ce cas et sans rapport avec la réalité avec ce que cela implique... Ces métaphores sont intéressantes pour la* taille d'effet observée *. Il s'agit d'une estimation de la taille d'effet théorique, ici par* $\frac{\overline{Y}-c}{S}$*. Elle permet de donner une idée de la taille de l'effet tel qu'il apparaît d'après la meilleure information dont nous disposons : l'échantillon, et de discuter les résultats en termes de* significativité pratique*, de surcroît à la* significativité statistique.

3.1.3 La puissance

On doit donc calculer la probabilité de la région de rejet dans le cadre de l'hypothèse alternative où la taille d'effet est d.

$$1-\beta = P(W/H_a) = P(T > qt(1-\alpha, n-1)/H_a)$$

On montre après quelques étapes que

$$1-\beta = P\left(\sqrt{n-1}\frac{\left(\sqrt{n}\frac{\overline{Y}-\mu}{\sigma}+\sqrt{n}\frac{\mu-c}{\sigma}\right)}{\sqrt{\frac{(n-1)S^2}{\sigma^2}}} > qt(1-\alpha, n-1)\right)$$

que l'on peut ré-écrire

$$1-\beta = P\left(\sqrt{df}\frac{U+ncp}{\sqrt{\chi^2(df)}} > qt(1-\alpha, n-1)\right).$$

Or la loi de $\sqrt{df}\frac{U+ncp}{\sqrt{\chi^2(df)}}$ est une loi qui est connue et tabulée. C'est la loi de Student décentrée avec $df = n-1$ degrés de liberté et comme paramètre de non-centralité $ncp = \sqrt{n}\frac{\mu-c}{\sigma}$. Les fonctions pt ou qt avec les options ncp et df permettent de réaliser ces calculs. Nous allons pour notre part utiliser la fonction pwr.t.test [4] qui permet d'obtenir directement le résultat avec l'option type="one.sample" (puisque nous avons dans cette situation un seul échantillon).

Pour notre exemple d'étude de souhaits basé sur le test unilatéral avec le niveau de signification conventionnel $\alpha = 0.05$, la taille d'échantillon $n = 25$ et la taille d'effet $d = 0.35$, on obtient $1-\beta = 0.52$ ce qui n'est pas satisfaisant.

```
pwr.t.test(n=25,d=0.35,sig.level=0.05,type="one.sample",alternative="greater")
```

Remarque 10 *On généralise facilement le cas de test unilatéral précédent à une situation bilatérale*[5] *en calculant la puissance comme*

$$\begin{aligned} 1-\beta &= P(|T| > qt(1-\alpha/2, n-1)/H_a) \\ &= P(T < qt(\alpha/2, n-1)/H_a) + P(T > qt(1-\alpha/2, n-1)/H_a). \end{aligned}$$

4. la fonction power.t.test le fait aussi avec d'autres arguments
5. il existe une variation dans power.anova.test n'étant pas complètement bilatérale

Ainsi dans l'exercice 2.5 p. 47 de Cohen [4] où l'on cherche à montrer dans un test bilatéral un petit effet $d = 0.2$, avec $n = 60$ et $\alpha = 0.10$, on obtient $1 - \beta = 0.46$ (ce qui là encore ne peut être considéré comme satisfaisant).

```
pwr.t.test(n=60,d=0.2,sig.level=0.10,type="one.sample",alternative="two.sided")
```

3.1.4 La taille d'échantillon

Si l'on souhaite calculer la taille d'échantillon pour une puissance donnée $1 - \beta$, il "suffit" de résoudre l'équation suivante où l'inconnue est n : $1 - \beta = P(W/H_a)$ ce qui se réalise numériquement avec la fonction pwr.t.test.

En ce qui concerne notre étude de souhaits avec une taille d'effet de $d = 0.35$, un niveau de signification conventionnel pour un test unilatéral, si l'on souhaite obtenir une puissance de $1 - \beta = 0.80$, il faut $n = 52$ unités statistiques.

```
pwr.t.test(power=0.8,d=0.35,sig.level=0.05,type="one.sample",alternative="greater")
```

Une possibilité des dernières versions de pwr (ici 1.2.1) est de pouvoir réaliser un graphique montrant la relation entre puissance et taille d'échantillon ainsi que la taille optimale pour la puissance fournie.

```
plot(pwr.t.test(power=0.8,d=0.35,sig.level=0.05,type="one.sample",alternative="greater"))
```

Exercice 4 *L'exercice 2.12 p 61 de Cohen [4] porte sur une échelle d'attitude de 11 points de 0 à 10 avec une réponse neutre à 5 en utilisant cette fois un test bidirectionnel :* $H_0 : \mu = 5$ *contre* $H_a : \mu \neq 5$.

- *En décidant un niveau de significativité* $\alpha = 0.01$, *mais en voulant une puissance de* $1 - \beta = 0.90$ *et voulant montrer un effet de taille* $d = 0.1$, *quelle taille d'échantillon nous faut-il ?*
- *Cette taille vous surprend-elle ?*
- *Changeons* $\alpha = 0.01$ *pour le niveau plus classique de* $\alpha = 0.05$. *Quelle est la conséquence ?*
- *Modifions à présent la puissance en employant le classique* $1 - \beta = 0.80$. *Que se passe-t-il ?*
- *Réalisons plutôt un test unilatéral, que constatons-nous ?*
- *Supposons que la taille d'effet soit plus élevée* ($d = 0.2$), *qu'observe-t-on ?*

Exercice 5 *Afin d'essayer de résumer les relations entre les différents éléments d'un calcul de puissance réaliser un graphique schématique qui représente*

- *la puissance* $1 - \beta$ *en fonction de* n,
- *la puissance* $1 - \beta$ *en fonction de* n,
- *la puissance* $1 - \beta$ *en fonction de* α,
- *la puissance* $1 - \beta$ *en fonction du type d'alternative (unilatérale ou bilatérale).*

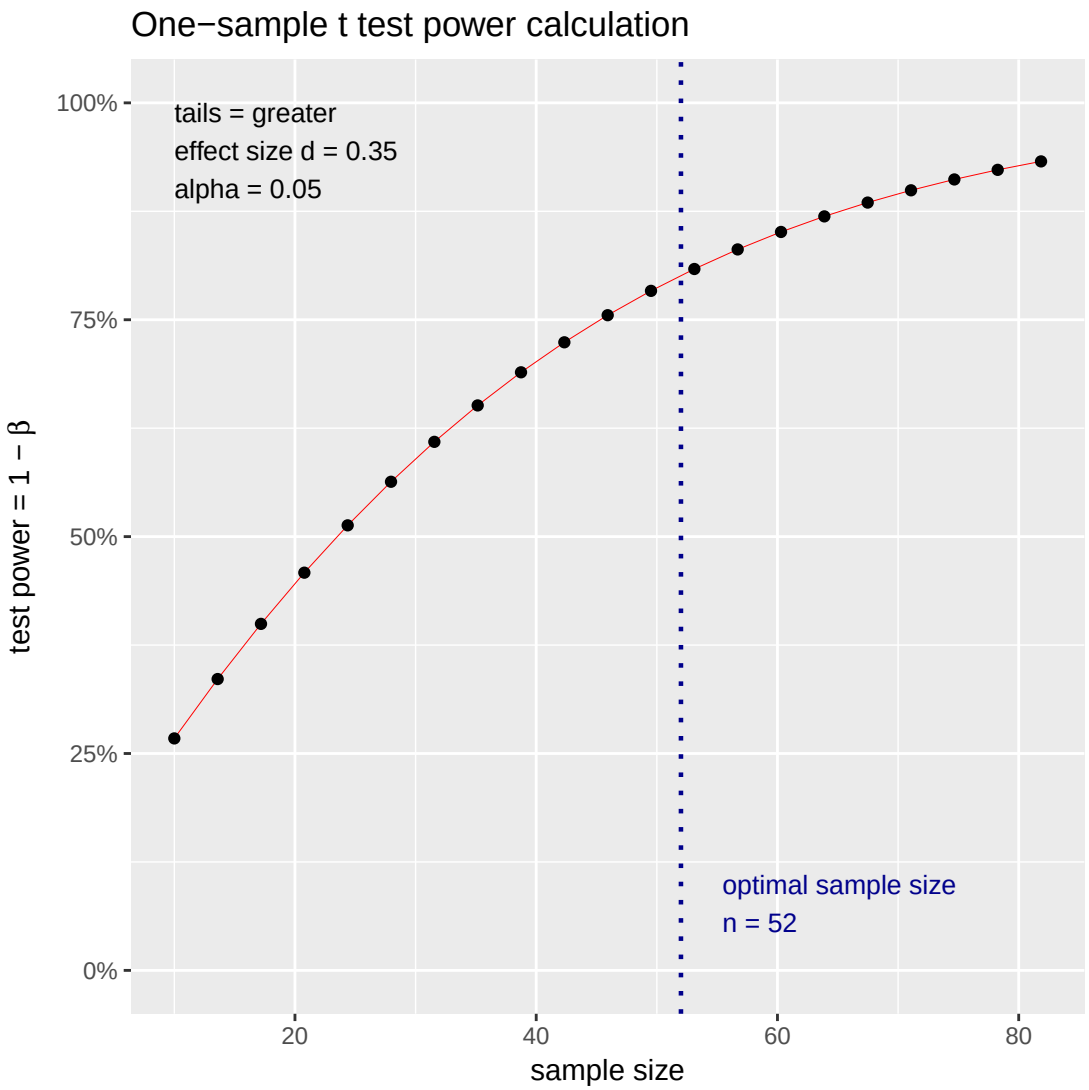

FIGURE 3.1 – Puissance en fonction de la taille d'échantillon pour un test unilatéral. Utilisation de la fonction plot()

3.2 Le test t à deux échantillons appariés

3.2.1 Rappels sur le test

On est en présence de deux mesures X et Y prises sur les mêmes n unités statistiques. On suppose qu'elles suivent une loi normale bivariée $N_2(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$. Ce qui nous intéresse est de tester la différence $\mu_X - \mu_Y$.

En tenant compte de l'appariement des deux mesures, on peut travailler sur la variable $Z = X - Y$ qui suit alors une loi normale $N(\mu_Z, \sigma_Z)$ où $\mu_Z = \mu_X - \mu_Y$ et $\sigma_Z = \sqrt{\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y}$

On se retrouve donc avec la nouvelle variable Z dans le cas du test précédent en testant les hypothèses $H_0 : \mu_Z = 0$ contre $H_a : \mu_Z \neq 0$ soit $H_0 : \mu_X = \mu_Y$ contre $H_a : \mu_X \neq \mu_Y$.

3.2.2 La taille d'effet

La taille d'effet est donc $d_Z = \frac{\mu_z - 0}{\sigma_Z} = \frac{\mu_X - \mu_Y}{\sigma_Z}$. Or la valeur σ_Z est dans ce cas plus difficile à définir. On peut s'en sortir (un peu trop facilement ?) en utilisant les tailles d'effet conventionnelles définies par Cohen (0.2, 0.5 et 0.8).

L'autre possibilité repose sur une estimation des différents éléments de σ_Z, c'est-à-dire σ_X, σ_Y et ρ le coefficient de corrélation linéaire entre X et Y (par étude pilote, données historiques, résultats issus de la littérature).

Remarque 11 *Notons que si $\sigma_X = \sigma_Y = \sigma$, on obtient $\sigma_Z = \sigma\sqrt{2(1-\rho)}$, la taille d'effet devient alors $d_Z = \frac{\mu_X - \mu_Y}{\sigma\sqrt{2(1-\rho)}}$.*

Nous retiendrons cette formule car elle permet de comprendre par la suite dans quelle mesure on peut améliorer un dispositif expérimental avec deux populations en choisissant un dispositif apparié.

Remarque 12 *La taille d'effet observée est selon les auteurs calculée pour estimer d_Z ou bien le précédent d.*

3.2.3 La puissance

Les calculs de puissance n'ont rien de nouveau par rapport à ce qui était vu précédemment. Nous allons les mettre en application sur l'exemple p.50 de Cohen [4].

On considère une variable d'aptitude et on souhaite tester l'existence d'une différence entre garçons et filles d'école primaire. On a déjà observé, lors d'études antérieures, que cette variable a chez les deux sexes un écart-type de $\sigma = 16$. On souhaite détecter une différence de 8 points entre les deux populations avec $n = 40$ sujets dans chacune (soit 80 sujets au total). Lorsque l'on peut prendre des paires de frères et soeurs, on a pour cette variable d'aptitude une corrélation de $r = 0.60$. Il est donc possible de calculer la taille d'effet comme $d_Z = \frac{8}{16 \times \sqrt{2(1-0.6)}} = 0.56$.

La puissance correspondante que l'on obtient à l'aide de l'option type="paired" est donc de $1 - \beta = 0.93$, très satisfaisante.

```
dz<-8/(16*sqrt(2*(1-0.6)))
pwr.t.test(n=40,d=dz,sig.level=0.05,type="paired",alternative="two.sided")
```

Exercice 6 *Calculer la puissance si la corrélation entre frères et soeurs est en fait moins élevée : $r = 0.4$? Que doit-on en conclure pratiquement en ce qui concerne les dispositifs appariés ?*

Exercice 7 *(Exercice 2.7, p. 51 de Cohen [4]) Très souvent le dispositif apparié vient du fait que la même personne est mesurée dans deux conditions différentes, par exemple avant et après une manipulation expérimentale. On considère ici l'efficacité d'un programme de régime et d'exercices d'un groupe d'étudiants ($n = 50$) en surpoids. Le chercheur a mesuré le poids avant (X), puis le poids (Y) après 60 jours de programme.*

Il veut connaître la puissance d'un test qui permet de détecter au niveau $\alpha = 0.01$ une perte de poids ($X - Y$) moyenne de 4 kilos, sachant que l'écart-type des poids, avant et après, est estimé à $\sigma = 12$. Il présume que la corrélation entre les poids avant-après sera très forte de l'ordre de $r = 0.80$.

- *Calculer la taille d'effet recherchée. Est-elle petite, moyenne ou grande ?*
- *Expliquer pourquoi le test choisi sera unilatéral.*
- *Calculer la puissance correspondante*
- *Si la taille d'effet est plus faible $d = 0.2$ quelle est la puissance correspondante ? Quels sont les paramètres qui peuvent faire que la taille d'effet soit plus faible que celle proposée précédemment ?*

3.2.4 La taille d'échantillon

Là non plus, rien de nouveau ; on peut directement passer à une application (Exercice 2.15 p. 66 de Cohen [4]) : Une expérimentatrice dans un laboratoire de psychologie s'intéresse chez les rats à la comparaison de deux méthodes d'apprentissage A et B. La méthode A est censée être supérieure.

Les rats vont être répartis aléatoirement entre les deux méthodes. L'expérimentatrice décide de mettre sur pied un dispositif apparié en utilisant des paires de rats provenant de la même portée. Sur la base de travaux précédents, elle sait que la capacité d'apprentissage entre rats de la même portée atteint une corrélation de $r = 0.65$.

En ce qui concerne la taille d'effet, elle souhaite détecter $d = \frac{\mu_A - \mu_B}{\sigma} = 0.5$ (ce qui serait un effet modéré dans le cadre de deux échantillons indépendants voir plus loin...). On obtient donc : $d_Z = \frac{0.5}{\sqrt{2(1-0.65)}} = 0.598$.

Elle choisit un niveau de significativité bas : $\alpha = 0.01$. On obtient alors comme taille d'échantillon nécessaire $n = 47$ [6] dans chaque groupe (soit 94 rats).

6. un résultat légèrement différent de celui de Cohen ($n = 45$)

```
d<-0.5/sqrt(2*(1-0.65))
pwr.t.test(power=0.95,d=d,sig.level=0.01,type="paired",alternative="greater")
```

Exercice 8 *Reprendre l'exemple précédent pour les valeurs du coefficient de corrélation de 0 à 0.8 (r<-seq(0,0.8,by=0.05)), calculer les valeurs de taille d'effet correspondantes et les tailles d'échantillons nécessaires. Représenter graphiquement la relation entre le coefficient de corrélation linéaire et la taille d'échantillon.*

3.3 Le test t à deux échantillons indépendants

3.3.1 Rappels sur le test

On observe deux échantillons X et Y qui suivent la loi normale et qui ont le même écart-type : $X \sim N(\mu_X, \sigma)$ et $Y \sim N(\mu_Y, \sigma)$. On suppose que l'on a la même taille d'échantillon[7] dans chaque groupe : n. On cherche à tester $H_0 : \mu_X = \mu_Y$ contre $H_a : \mu_X \neq \mu_Y$ (en bilatéral).

Á partir des moyennes $\overline{X}$ et $\overline{Y}$ et des écarts-types S_X et S_Y des échantillons, on calcule (1) un écart-type compromis

$$S = \sqrt{\frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}}$$

et (2) une nouvelle statistique T :

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

Sous l'hypothèse nulle, la loi de cette statistique est une loi de Student à $n_X + n_Y - 2$ degrés de liberté. On a donc en bilatéral une région critique de la forme

$$W = \{|T| > qt(1 - \alpha/2, n_X + n_Y - 2)\}.$$

3.3.2 La taille d'effet

La taille d'effet de ce test est

$$d = \frac{\mu_X - \mu_Y}{\sigma}$$

qui est nulle lorsque l'hypothèse nulle est vérifiée et est indépendante des unités de mesures.

7. On pourra relâcher cette hypothèse mais la fonction power.t.test fournie par R la fait. Je la reprends dans ma version pwr.t.test car les dispositifs équilibrés sont fortement conseillés par la théorie. Il existe toutefois une fonction pwr.t2n.test dans le package pwr

On essaie bien sûr de la définir plutôt à partir de considérations pratiques, mais on peut, le cas échéant, employer les tailles d'effet conventionnelles (0.2, 0.5 et 0.8) de Cohen [4].

Remarque 13 *En tenant compte de la relation entre ce test de Student et l'analyse de la variance avec deux groupes (c'est le même test), on peut établir une relation avec le rapport de corrélation dont l'interprétation est parfois plus aisée pour le praticien (bien qu'il ait tendance à la sur-estimer). En effet* $d = 0.2$ *correspond à* $\eta^2 = 0.01$ *un faible pourcentage de variance expliquée,* $d = 0.5$ *à* $\eta^2 = 0.06$ *et* $d = 0.8$ *à* $\eta^2 = 0.138$.

3.3.3 La puissance

La théorie de la puissance repose encore sur la loi de Student décentrée mais elle possède ici $df = n_X + n_Y - 2$ degrés de liberté et son paramètre de non centralité est $ncp = \frac{1}{\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}}d$ qui se simplifie en cas d'effectifs égaux en $ncp = \sqrt{\frac{n}{2}}d$.

Á cause de leur handicap, les aveugles ont souvent une pratique physique moindre. Une étude de Sundberg [10] montre que les adolescents masculins ont une VO2Max dont l'écart-type est supposé proche de $\sigma = 10$. On souhaite les comparer à une population de jeunes filles aveugles du même âge dont on imagine la VO2Max inférieure de 5 points et l'écart-type pas trop différent. La taille d'effet serait donc de $d = -5/10 = -0.5$. Il est proposé de prendre $n = 30$ personnes dans chaque échantillon. On prendra le seuil de signification conventionnel et on pratiquera un test unilatéral. La puissance obtenue est de $1 - \beta = 0.61$, ce qui est trop faible.

```
d<- -5/10
pwr.t.test(n=30,d=d,sig.level=0.05,type="two.sample",alternative="less")
```

Exercice 9 *L'exercice 2.1 p. 40 de Cohen [4] décrit une expérience sur l'apprentissage des souris dans un labyrinthe et l'effet d'une récompense.* $n = 30$ *souris ont été placées dans chaque groupe, l'un étant un groupe avec récompense (X), l'autre étant un groupe de contrôle (Y). L'écart absolu d'intérêt entre les deux groupes est de 2, à rapporter à un écart-type tiré de la littérature de 2.8. La taille d'effet correspondante est donc de* $d = \frac{2}{2.8} = 0.714$*, un effet classé de moyen à fort.*

Il a été choisi de pratiquer un test bilatéral (bien qu'on s'attende à de meilleurs scores dans le groupe expérimental...) avec un niveau conventionnel de $\alpha = 0.05$. *Quelle puissance obtient-on alors ?*

Si on suppose qu'on ne puisse avoir qu'une amélioration, quel changement faut-il faire dans le calcul de puissance ? Et quel résultat obtient-on ?

Et si la taille d'effet est moyenne : $d = 0.5$*, quelle puissance obtient-on ? Que dire alors des expériences, très répandues en pratique, où l'on trouve très souvent cette taille d'échantillon de* $n = 30$ *?*

Exercice 10 *(Exercice 2.2 p. 41 de Cohen [4]) Un psychiatre cherchant des facteurs endocrinologiques impliqués dans la schizophrénie veut réaliser une expérience avec $n = 500$ schizophrènes et $n = 500$ patients normaux. Des échantillons d'urine sont examinés sur lesquels on relève un produit métabolique avec lequel on attend un faible effet $d = 0.2$. Il sélectionne un niveau de significativité conservateur $\alpha = 0.01$ et le test pratiqué sera bidirectionnel.*

Calculer la puissance correspondant à ces conditions. Qu'en pensez-vous ? Que se passe-t-il si on utilise le niveau conventionnel $\alpha = 0.05$?

Remarque 14 *Il arrive parfois pour des raisons pratiques que les deux tailles d'échantillons soient différentes. On peut faire un calcul approché*[8] *avec les outils précédents en employant une taille compromis de $n' = \frac{2n_X n_Y}{n_X + n_Y}$. La puissance sera alors légèrement sous-estimée.*

Exercice 11 *(Exercice 2.3 p.43 de Cohen [4])Dans un service psychiatrique, les patients sont assignés soit à une technique standard (X) soit à une nouvelle technique (Y). Après un certain temps, $n_X = 90$ cas ont été traités de façon standard contre $n_Y = 60$ pour la nouvelle méthode. La variable mesurée est un score d'amélioration du patient. Un taille d'effet de $d = 0.6$ est décidée et on choisit un test unilatéral car on imagine que la nouvelle technique fonctionne mieux. Un niveau de significativité conventionnel est choisi $\alpha = 0.05$.*

- *Calculer la taille compromis.*
- *Calculer la puissance correspondant à ce test.*
- *Utiliser la fonction* pwr.t2n.test*.*

3.3.4 La taille d'échantillon

Dans l'exemple concernant les jeunes aveugles, la puissance de $1 - \beta = 0.61$ étant trop faible avec $n = 30$ individus dans chaque groupe, on se propose d'augmenter la taille des échantillons pour atteindre une puissance de $1 - \beta = 0.80$ avec la même taille d'effet moyenne.

Il faut alors $n = 50$ unités statistiques dans chaque groupe.

```
pwr.t.test(power=0.8,d=0.5,sig.level=0.05,type="two.sample",alternative="greater")
```

Exercice 12 *On va reprendre l'exercice 9 des souris dans le labyrinthe. On souhaitait donc détecter un effet de taille moyenne $d = 0.5$ avec un test bilatéral de niveau $\alpha = 0.05$ pour obtenir une puissance de $1 - \beta = 0.80$. Combien de rats doit-on employer en tout ?*

Calculer la taille d'échantillon correspondant à un effet fort ($d = 0.8$) et un effet faible ($d = 0.2$).

*Expliquer pourquoi Cohen [4] dit qu'*un dispositif expérimental peut difficilement être mis au point en l'absence d'un jugement préalable sur la taille des effets à détecter.

8. Il est possible de faire un calcul exact, voire la fonction pwr.t2n.test

Remarque 15 *On ne peut parfois dépasser une certaine taille pour l'un des deux échantillons. On doit alors construire un dispositif déséquilibré. Dans ce cas, avec n_X fixé, on calcule n comme d'habitude et on a $n_Y = \frac{nn_X}{2n_X - n}$* [9].

Exercice 13 *(Exercice 2.10 p. 59 de Cohen [4]) Dans un test d'un programme informatique d'enseignement à la lecture, on s'attend à un effet unidirectionnel de l'ordre de $d = 0.3$ pour un niveau de significativité conventionnel de $\alpha = 0.05$ et on souhaite atteindre une puissance de $1 - \beta = 0.75$.*

- *Calculer d'abord sans restriction, quelle est la taille d'échantillon nécessaire dans chaque groupe ?*
- *En fait, pour des raisons d'espace, il ne peut y avoir que 80 sujets dans le groupe informatique. Calculer le nombre de sujets nécessaires dans le groupe standard pour atteindre le même objectif de puissance.*
- *Utiliser la fonction* pwr.t2n.test *pour résoudre ce problème.*

Remarque 16 *Nous avons vu que la puissance d'un test dépend du nombre de sujets que l'on peut employer. Il est possible de jouer sur un autre élément : l'organisation de l'expérience. Il est plus efficace d'avoir un appariement - lorsque la corrélation n'est pas nulle - que deux échantillons indépendants. En fait, la théorie des plans d'expériences cherche pour bonne part, c'est la constitution de* blocs, *à augmenter la puissance du test de comparaison en réduisant la variabilité individuelle (le dénominateur de la taille d'effet).*

9. Il n'y a pas toujours de solution à ce problème si $2n_X < n$

Chapitre 4

Tests de proportions

Il existe plusieurs façons de pratiquer un test de proportions (test exact, test Z ou test X^2). Nous utiliserons un test qui est moins connu mais qui possède certains avantages dans le cadre de l'analyse de puissance. En outre, cela permet de rester cohérent avec l'ouvrage de Cohen [4] [1].

4.1 Le test de proportion à un échantillon

4.1.1 "Rappels" sur le test

On suppose que l'on s'intéresse à la probabilité de succès π dans un cadre binomial avec n répétitions. On souhaite en particulier tester $H_0 : \pi \leq c$ contre $H_a : \pi > c$ (version unilatérale).

On utilise comme statistique de test la proportion p observée sur l'échantillon. On se sert [2] alors de l'approximation normale suivante

$$2\arcsin(\sqrt{p}) \sim N\left(2\arcsin(\sqrt{\pi}), \frac{1}{\sqrt{n}}\right).$$

On note

$$\phi(p) = 2\arcsin(\sqrt{p})$$

cette fonction monotone.

La région critique est de la forme $W = \{p > k\}$. Comme d'habitude, cette région critique doit sous l'hypothèse nulle avoir pour probabilité le niveau de significativité α d'où $P(W/H_0) = \alpha$ soit $P(p > k/H_0) = \alpha$ soit en utilisant la monotonie de ϕ on a $P(\phi(p) > \phi(k)/H_0) = \alpha$ puis $P(\sqrt{n}(\phi(p) - \phi(c)) > \sqrt{n}(k - \phi(c))/H_0) = \alpha$. Sous l'hypothèse nulle, la quantité à gauche suit une loi

1. où il dit d'ailleurs que les résultats sont comparables entre les différentes méthodes.
2. la fonction arcsinus est la réciproque de la fonction sinus. A une valeur de sinus, donc une quantité entre -1 et 1, elle associe l'angle correspondant en radians de $-\pi$ à π

normale, on a donc une région critique qui peut s'écrire [3]

$$W = \left\{\sqrt{n}(\phi(P) - \phi(c)) > z_{1-\alpha}\right\}$$

où $z_{1-\alpha}$ est le quantile de la loi normale standard.

Exercice 14 *En utilisant les fonctions R* asin *et* sqrt*, calculer en fonction de* $\phi(p)$ *au niveau* $\alpha = 0.05$ *la région critique correspondant à* $n = 50$ *répétitions pour le test* $H_0 : \pi \leq 0.5$ *contre* $H_a : \pi > 0.5$*. Si on observe* $p = 0.70$*, quelle est notre décision ?*

4.1.2 La taille d'effet

La taille d'effet va être calculée en fonction des proportions transformées. Elle sera

$$h = \phi(\pi) - \phi(c).$$

Il peut être surprenant de choisir une taille d'effet dans des unités transformées mais l'avantage est que cette transformation possède une loi dont l'écart-type ne dépend plus de la probabilité de succès, ce qui rend les calculs de puissance beaucoup plus facile. Pour déterminer cette taille d'effet en pratique, il peut être plus simple de donner un tableau qui décrit la relation entre deux probabilités que le client doit pouvoir supputer, et la valeur de h correspondante (voir tableau 4.1).

TABLE 4.1 – Valeur de la taille d'effet $h = \phi(p_1) - \phi(p_2)$ en fonction de p_1 et p_2

p_1/p_2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.00	-0.28	-0.52	-0.73	-0.93	-1.13	-1.34	-1.57	-1.85
0.2	0.28	0.00	-0.23	-0.44	-0.64	-0.84	-1.06	-1.29	-1.57
0.3	0.52	0.23	0.00	-0.21	-0.41	-0.61	-0.82	-1.06	-1.34
0.4	0.73	0.44	0.21	0.00	-0.20	-0.40	-0.61	-0.84	-1.13
0.5	0.93	0.64	0.41	0.20	0.00	-0.20	-0.41	-0.64	-0.93
0.6	1.13	0.84	0.61	0.40	0.20	0.00	-0.21	-0.44	-0.73
0.7	1.34	1.06	0.82	0.61	0.41	0.21	0.00	-0.23	-0.52
0.8	1.57	1.29	1.06	0.84	0.64	0.44	0.23	0.00	-0.28
0.9	1.85	1.57	1.34	1.13	0.93	0.73	0.52	0.28	0.00

Sinon, Cohen [4] propose conventionnellement les niveaux suivants : (1) effet faible pour $h = 0.2$, (2) effet moyen pour $h = 0.5$ et (3) effet fort pour $h = 0.8$.

Remarque 17 *On peut, là aussi, calculer des tailles d'effet observées en utilisant la proportion* p *observée dans l'échantillon :* $\hat{h} = \phi(p) - \phi(c)$*. En considérant la taille d'échantillon, elle permet de mieux saisir le résultat (significativité statistique) obtenue. Elle permet également de calculer des puissances observées* $1 - \hat{\beta}$*, parfois délicate d'utilisation.*

3. On ne cherchera pas à l'écrire directement en fonction de p bien que cela soit possible

4.1.3 La puissance

Il s'agit comme toujours de la probabilité de la région critique lorsque nous sommes dans le cadre de l'hypothèse alternative en la définissant plus précisément par la taille d'effet h (d'abord en version unilatérale).

$$\begin{aligned} 1-\beta &= P(\sqrt{n}(\phi(p)-\phi(c)) > z_{1-\alpha}/H_a) \\ &= P(\sqrt{n}(\phi(p)-\phi(\pi)+\phi(\pi)-\phi(c)) > z_{1-\alpha}/H_a) \\ &= P(\sqrt{n}(\phi(p)-\phi(\pi)) > z_{1-\alpha}-\sqrt{n}h/H_a) \\ &= P(N(0,1) > z_{1-\alpha}-\sqrt{n}h). \end{aligned}$$

Remarque 18 *Dans le cas bilatéral, on montre que*

$$1-\beta = P(N(0,1) > z_{1-\alpha/2}-\sqrt{n}h) + P(N(0,1) < z_{\alpha/2}-\sqrt{n}h).$$

Chez les jeunes sportifs, l'utilisation de catégories d'âge fait que les individus nés en début d'année sont plus matures et peuvent avoir plus de chances de réussir. On veut tester s'ils sont plus nombreux parmi les joueurs de Hockey professionnels (voir l'article de Grondin [5])) en observant le nombre de ces joueurs nés dans les trois premiers trimestres de l'année.

S'ils sont normalement représentés, on doit avoir une probabilité de présence de $\pi = \frac{3}{4}$. On va donc tester $H_0 : \pi = \frac{3}{4}$ contre $H_a : \pi > \frac{3}{4}$.

On considérera que si la probabilité de présence π est de 80% cela sera "scientifiquement" significatif.

La valeur de la taille d'effet est donc de $h = \phi(0.80) - \phi(0.75) = 0.12$ (valeur obtenue à l'aide de la fonction ES.h), ce qui est classé comme très faible. Si un échantillon de $n = 100$ joueurs est sélectionné pour tester ces hypothèses au niveau $\alpha = 5\%$, la puissance correspondante est de $1-\beta = 0.33$, une catastrophe.

```
ES.h(0.8,0.75)
pwr.p.test(h=ES.h(0.8,0.75),sig.level=0.05,n=100,alternative="greater")
```

Exercice 15 *(Exercice 6.5 p. 203 de Cohen [4]) Un modèle mathématique prédit qu'une certaine réponse devrait apparaître chez $c = 40\%$ des animaux soumis à certaines conditions. Un expérimentateur souhaite tester ce modèle donc cette proportion sur $n = 60$ animaux dans un test bilatéral avec un niveau de significativité de $\alpha = 0.05$. Il présume que le véritable taux est de $\pi = 50\%$.*

- *Quelle est la taille d'effet correspondante ?*
- *Cette taille d'effet est-elle faible, moyenne ou forte ?*
- *Quelle est la puissance correspondante ?*
- *Est-elle satisfaisante ?*

4.1.4 La taille d'échantillon

La taille d'échantillon peut se calculer directement dans le cas unilatéral puisqu'on a $z_{1-\beta} = \sqrt{n}h - z_{1-\alpha}$ d'où

$$n = \left(\frac{z_{1-\alpha} + z_{1-\beta}}{h} \right)^2 .$$

Dans le cas bilatéral, on emploiera une méthode numérique programmée dans la fonction pwr.p.test. Dans le problème de l'âge des joueurs de hockey, la puissance obtenue avec $n = 100$ joueurs n'était pas du tout satisfaisante. Pour une taille d'effet de $h = 0.12$, dans un test unilatéral au niveau $\alpha = 5\%$, il faut en effet $n = 430$ unités statistiques pour atteindre une puissance de 80%!!!

```
pwr.p.test(h=ES.h(0.80,0.75),sig.level=0.05,power=0.80,alternative="greater")
```

Exercice 16 *Dans le problème du modèle mathématique (exercice 15) avec un test bilatéral où $\alpha = 0.05$, la même taille d'effet $h = 0.2$ et une puissance souhaitée de $1 - \beta = 0.95$, quelle taille d'échantillon est-elle nécessaire ?*

Calculer dans les mêmes conditions mais avec des tailles d'effet de $h = 0.5$ et $h = 0.8$ les tailles d'échantillons nécessaires.

4.2 Le test de proportion à deux échantillons indépendants

4.2.1 Rappels sur le test

Cette fois-ci, on a deux échantillons X et Y qui suivent des lois binomiales de probabilités de succès respectives π_X et π_Y. Le nombre de répétitions sera noté n_X et n_Y. On souhaite tester les hypothèses $H_0 : \pi_X = \pi_Y$ contre $H_a : \pi_X \neq \pi_Y$ (ici en bilatéral).

En utilisant les remarques de la section précédente, on en vient à une région critique de la forme :

$$W = \left\{ \sqrt{\frac{n_X n_Y}{n_X + n_Y}} \, |\phi(p_X) - \phi(p_Y)| > z_{1-\alpha/2} \right\}$$

ou plus simplement si $n_X = n_Y = n$

$$W = \left\{ \sqrt{\frac{n}{2}} \, |\phi(p_X) - \phi(p_Y)| > z_{1-\alpha/2} \right\}$$

4.2.2 La taille d'effet

La taille d'effet est dans ce cas

$$h = \phi(\pi_X) - \phi(\pi_Y).$$

On utilise de préférence des proportions suggérées par le praticien et sinon les mêmes tailles conventionnelles c'est-à-dire 0.2 (faible), 0.5 (moyenne) et 0.8 (forte).

4.2.3 La puissance

On démontre que la puissance est, dans le cas unilatéral, de

$$1-\beta = P\left(N(0,1) > z_{1-\alpha} - \sqrt{\frac{n_X n_Y}{n_X+n_Y}}h\right)$$

et, dans le cas bilatéral, de

$$1-\beta = P\left(N(0,1) > z_{1-\alpha/2} - \sqrt{\frac{n_X n_Y}{n_X+n_Y}}h\right) + P\left(N(0,1) < z_{\alpha/2} - \sqrt{\frac{n_X n_Y}{n_X+n_Y}}h\right).$$

Ces calculs sont accomplis par les fonctions suivantes :

— pwr.2p2n.test pour un test avec tailles différentes d'échantillons

— pwr.2p.test pour un test avec mêmes tailles d'échantillons.

En football, le fait de jouer à domicile peut constituer, même pour l'arbitrage, un avantage. L'étude de Avanzini [1] s'intéresse à la probabilité de voir ses fautes sanctionnées pour l'équipe qui est reçue (π_X) par rapport à l'équipe qui reçoit (π_Y). On teste donc les hypothèses $H_0 : \pi_X = \pi_Y$ contre $H_a : \pi_X > \pi_Y$.

On compte visionner suffisamment de matchs afin de repérer $n_X = n_Y = 1000$ fautes commises par chacune des deux équipes. On considérera que si on observe une différence du type $\pi_X = 40\%$ contre $\pi_Y = 35\%$ cela sera scientifiquement significatif. La taille d'effet correspondante est de $h = \phi(0.40) - \phi(0.35) = 0.10$.

En utilisant un test unilatéral au niveau $\alpha = 5\%$, la puissance sera de $1-\beta = 75\%$.

```
ES.h(0.40,0.35)
pwr.2p.test(h=ES.h(0.40,0.35),n=1000,sig.level=0.05,alternative="greater")
```

Exercice 17 *(Exercice 6.1 p.198 de Cohen [4]) Un socio-psychologue cherche à savoir si les enfants uniques (X) préfèrent attendre plus fréquemment avec d'autres qu'attendre seuls (anticipant une anxiété) par rapport à des enfants (Y) nés dans une fratrie (mais pas au premier rang).*

Dans une culture non-occidentale, il obtient la coopération de $n_X = n_Y = 80$ personnes. Un travail précédent aux E.U. suggère que les 2/3 des X préfèrent attendre "ensemble" ($\pi_X = 2/3$) et que seulement la moitié des Y est ainsi ($\pi_Y = 1/2$).

— *En s'attendant sur cette nouvelle population à une différence de ce type, quelle est la taille d'effet correspondante ?*

— *Il prévoit un test unilatéral avec le niveau de test conventionnel de $\alpha = 5\%$. Quelle est la puissance correspondante ?*

Exercice 18 *(Exercice 6.2 p. 199 de Cohen [4]) Une psychologue clinicienne planifie une recherche où les patients, lors de leur admission à l'hôpital, sont assignés aléatoirement dans deux* athmosphères *: (X) "Autoritaire" et (Y) "Démocratique". Elle va mesurer la proportion d'entre eux qui au bout de six mois seront guéris. L'hôpital admettant 50 patients par mois, elle attendra quatre mois afin de réunir deux groupes de $n_X = n_Y = 100$ sujets.*

- *En s'appuyant sur son expérience, elle s'attend à des différences de proportions de l'ordre de 20% autour de 50% ce qui la conduit à choisir $h = 0.4$. Expliquer son raisonnement.*
- *Le test sera bilatéral au niveau de significativité conventionnel. Calculer la puissance correspondante.*
- *Quelle est la puissance atteinte si elle attend un mois de plus (50 patients supplémentaires) ?*

Exercice 19 *Dans l'expérience avec les premiers nés (exercice 17), il est probable que, dans une culture non occidentale, l'on trouvera plus de personnes provenant d'une fratrie (et non premier né). On souhaite détecter une taille d'effet d'environ $h = 0.30$ pour un test unilatéral au niveau conventionnel, mais avec $n_X = 80$ et $n_Y = 245$. Quelle est la puissance de ce test ?*

4.2.4 La taille d'échantillon

En unilatéral avec la même taille d'échantillon le calcul est simple :

$$n = 2\left(\frac{z_{1-\alpha} + z_{1-\beta}}{h}\right)^2.$$

Pour les autres cas, on emploie les solutions obtenues numériquement par les fonctions **R**.

Dans l'exemple des fautes d'arbitrages, la puissance obtenue n'était pas tout à fait satisfaisante. Nous allons être très exigeant en demandant qu'elle s'élève à $1-\beta = 95\%$ pour détecter un très faible effet de $h = 0.10$, toujours en unilatéral avec le niveau de signification conventionnel. Le nombre de fautes à repérer pour chaque équipe est alors de $n = 2027$ (dans l'article correspondant, la taille des échantillons est effectivement proche de 2000).

```
pwr.2p.test(h=ES.h(0.40,0.35),power=0.95,sig.level=0.05,alternative="greater")
```

Exercice 20 *Dans l'exemple des premiers nés (cf exercice 17) pour un test unilatéral avec niveau de signification conventionnel, un effet de taille $h = 0.3$ était détecté avec une puissance de $1 - \beta = 0.60$ pour $n_X = n_Y = 80$ cas dans chaque groupe. Quelle taille d'échantillon faut-il prendre pour atteindre une puissance de $1 - \beta = 0.80$?*

Exercice 21 *Toujours avec l'exemple des premiers nés, en supposant que le nombre de personnes enfants uniques est limité à $n_X = 80$ combien de personnes*

de l'autre groupe (n_Y) faut-il prendre pour détecter la taille d'effet $h = 0.3$ avec une puissance "satisfaisante" de $1 - \beta = 0.80$?

Combien cela fait-il au total d'individus ? Expliquer la différence du nombre total d'individus obtenu dans le calcul avec échantillons équilibrés.

Chapitre 5

Test de corrélation

En présence de deux mesures prises sur les mêmes unités statistiques, on peut s'intéresser à la relation entre ces variables. Lorsqu'elles sont toutes deux quantitatives, la méthode la plus classique repose sur le coefficient de corrélation linéaire.

5.1 Rappel du test

Nous avons donc deux variables X et Y qui suivent une loi bi-normale de coefficient de corrélation ρ. On cherche à tester $H_0 : \rho = 0$ contre $H_a : \rho \neq 0$ (en version bilatérale).

Le test statistique est basé sur le coefficient de corrélation linéaire r calculé sur l'échantillon. On définit alors la statistique de Student

$$T = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}.$$

Sous l'hypothèse nulle, cette statistique suit une loi de Student à $n-2$ degrés de liberté. On obtient donc une région critique de la forme :

$$W = \{|T| > qt(1-\alpha/2, n-2)\}$$

où $qt(1-\alpha/2, n-2)$ est le quantile de la loi de Student.

5.2 La taille d'effet

La taille d'effet est ici très simple, elle s'exprime directement en fonction de r [1]. En effet, cette quantité traduit l'éloignement à l'hypothèse nulle et elle est en même temps indépendante des unités de mesure.

1. En fait de ρ mais c'est la notation (mauvaise) choisie par Cohen [4] que je conserve.

Comme toujours, on essaie d'appuyer le choix de r sur des considérations scientifiques, sur des données historiques, sur la littérature du champ d'investigation. Á défaut, Cohen [4] propose d'utiliser les tailles d'effet conventionnelles suivantes : $r = 0.1$ (faible), $r = 0.3$ (moyenne) et $r = 0.5$ (forte).

5.3 La puissance

Le calcul de la puissance repose sur une transformation du coefficient de corrélation linéaire (qui a pour objectif de rendre son écart-type indépendant du paramètre d'intérêt) :

$$Z(r) = arctangh(r) + \frac{r}{2(n-1)}.$$

Cette transformation suit approximativement une loi normale d'espérance $Z(\rho)$ et d'écart-type $\frac{1}{\sqrt{n-3}}$. Dès lors, en supposant que nous traitions du cas unilatéral $W = \{T > qt(1-\alpha, n-2)\}$, on peut ré-écrire cette région $W = \{r > r_c\}$ où

$$r_c = \sqrt{\frac{q_t^2}{q_t^2 + n - 2}}$$

avec $q_t = qt(1-\alpha, n-2)$.

La puissance est donc

$$\begin{aligned} 1-\beta &= P(W/H_a) \\ &= P(r > r_c/H_a) \\ &= P(Z(r) > Z(r_c)/H_a) \\ &= P(\sqrt{n-3}(Z(r) - Z(\rho)) > \sqrt{n-3}(Z(r_c) - Z(\rho))/H_a) \\ &= P(N(0,1) > \sqrt{n-3}(Z(r_c) - Z(\rho))) \\ &= P(N(0,1) < \sqrt{n-3}(Z(\rho) - Z(r_c))). \end{aligned}$$

Remarque 19 *De façon similaire, on montre que, dans le cas bilatéral, mais en utilisant $q_t = qt(1-\alpha/2, n-2)$ pour le calcul de r_c, la puissance est alors*

$$1-\beta = P(N(0,1) < \sqrt{n-3}(Z(\rho) - Z(r_c))) + P(N(0,1) < \sqrt{n-3}(-Z(\rho) - Z(r_c)))$$

La concentration en lactates est-elle reliée à la vitesse de sprint ? On supposera que cette corrélation est moyenne $r = 0.30$, qu'elle est positive (donc le test sera unilatéral), que le niveau conventionnel de significativité convient et que nous disposons d'un échantillon de $n = 36$ individus comme dans l'article de Bret [2]. La fonction pwr.r.test permet de réaliser le calcul de puissance qui est ici de $1 - \beta = 55\%$ ce qui est trop faible.

```
pwr.r.test(r=0.3,n=36,sig.level=0.05,alternative="greater")
```

Exercice 22 *(Exercice 3.1 p. 96 de Cohen [4]) Une expérience est réalisée par un psychologue de la personnalité qui a mesuré $n = 50$ sujets. Une des variables (X) est un score d'extraversion et l'autre (Y) une mesure neuro-physiologique qui, en théorie, est reliée à la première. Le test est bilatéral avec le niveau de significativité conventionnel. Un effet moyen est attendu $r = 0.30$. Quelle est la puissance de ce test ?*

Si on considère la version unilatérale de ce test (il doit quand même connaître la direction de l'association entre les deux mesures), quelle est alors la puissance ?

Exercice 23 *(Exercice n. 3.2 p.97 de Cohen [4]) Un psychologue de l'éducation est consulté pour établir un nouveau critère d'admission dans un lycée en utilisant un test de personnalité (X). On cherche à savoir si cette mesure est corrélée à la moyenne obtenue par l'élève en fin d'année (Y).*

Pour l'instant, il n'y a pas de critère de sélection donc même un faible effet $r = 0.1$ serait considéré comme une amélioration. Le nombre annuel d'entrants sur lequel on projette de faire cet essai est de $n = 500$.

Calculer dans un contexte bilatéral les puissances correspondant aux niveaux de significativité $\alpha = 0.01$ et $\alpha = 0.05$. Qu'en pensez-vous ?

5.4 La taille d'échantillon

En ce qui concerne la relation de la production de lactates et de la vitesse de sprint, avec $n = 36$ sujets, la puissance était beaucoup trop faible. Pour atteindre avec un test unilatéral de niveau $\alpha = 5\%$, une puissance satisfaisante de $1 - \beta = 80\%$ lorsque la taille d'effet recherchée est de $r = 0.3$, il faut $n = 67$ sujets soit près du double.

Remarquons que, dans les mêmes conditions, si on attend un fort effet $r = 0.5$ on a $n = 23$ et si, au contraire, on attend $r = 0.1$ on doit avoir $n = 616$! Á nouveau, on voit clairement que si on prend une certaine taille d'échantillon, c'est bien que l'on attend un certain type d'effet !

```
pwr.r.test(r=0.3,power=0.8,sig.level=0.05,alternative="greater")
pwr.r.test(r=0.5,power=0.8,sig.level=0.05,alternative="greater")
pwr.r.test(r=0.1,power=0.8,sig.level=0.05,alternative="greater")
```

Exercice 24 *Reprenons l'exercice 22 avec le psychologue de la personnalité, où il souhaitait repérer une taille d'effet de $r = 0.3$ pour un test bilatéral avec le niveau conventionnel. Avec $n = 50$ sujets, la puissance obtenue est bien trop faible. Quelle taille d'échantillon doit-il prendre pour obtenir une puissance plus satisfaisante de $1 - \beta = 0.90$?*

Exercice 25 *Reprenons l'exercice 23, et essayons de déterminer la taille d'échantillon nécessaire pour repérer un faible effet ($r = 0.1$) dans un test bilatéral avec le niveau conventionnel pour atteindre une puissance de $1 - \beta = 0.90$.*

Chapitre 6

Test d'analyse de variance

6.1 Rappel du test

On suppose que l'on affaire à k populations qui suivent indépendamment des lois normales d'espérances μ_j et de même écart-type σ. On cherche à savoir si ces espérances sont égales : $H_0 : \mu_1 = ...\mu_k$ contre $H_a : \exists \mu_i \neq \mu_j$.

La statistique de test est

$$F = \frac{\sum_{j=1}^{k} n_j(\overline{Y_j} - \overline{Y})^2}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2} \times \frac{N-k}{k-1}$$

Sous l'hypothèse nulle, cette statistique suit une loi dite de *Fisher-Snedecor* à $df_1 = k-1$ et $df_2 = N-k$ degrés de liberté. On obtient une région critique de la forme $W = \{F > q_f\}$ où $q_f = qf(1-\alpha, k-1, N-k)$ est le quantile correspondant de la loi de Fisher-Snedecor.

6.2 La taille d'effet

La taille d'effet doit être comme d'habitude une quantité qui traduit l'éloignement à l'hypothèse nulle et ne pas dépendre des unités de mesures. Il s'agit ici de

$$f = \frac{\sigma_m}{\sigma}$$

où $\sigma_m = \sqrt{\sum_{j=1}^{k} \frac{n_j}{N}(\mu_j - \mu)^2}$ avec $\mu = \sum_{j=1}^{k} \frac{n_j}{N}\mu_j$.

Si toutes les espérances sont égales, on retrouve bien $f = 0$.

Il existe plusieurs méthodes pour définir la taille de f :

1. en postulant directement les valeurs des espérances et de l'écart-type sur la base de l'expérience ou de la littérature,
2. en utilisant une relation entre l'étendue des espérances $d = \frac{\mu_{max} - \mu_{min}}{\sigma}$ et f lorsque l'on postule certaines dispositions de ces espérances (voir tableau 8.2.1 p 278 dans [4]),

3. en employant la relation $f = \sqrt{\frac{\eta^2}{1-\eta^2}}$ où η^2 est le rapport de corrélation que l'on peut estimer parfois en tant que pourcentage de variabilité expliquée [1],
4. en utilisant des valeurs conventionnelles suggérées par Cohen [4] : $f = 0.10$ (faible), $f = 0.25$ (moyenne) et $f = 0.40$ (forte).

Nous allons montrer sur un exemple concernant la comparaison de détentes horizontales d'enfants pratiquant des activités différentes comment on peut estimer cette taille d'effet. On va chercher à comparer des enfants pratiquant l'athlétisme, le cyclisme et la natation. On peut prédire que les athlètes auront des performances supérieures (200 cm) en moyenne par rapport aux cyclistes (185 cm) et encore plus par rapport aux nageurs (175 cm). L'écart-type commun aux trois groupes est évalué à 20 cm.

On a alors puisqu'on va prendre des groupes de même taille : $\mu = \frac{200+185+175}{3} = 186.67$, $\sigma_m = \sqrt{\frac{(200-186.67)^2+(185-186.67)^2+(175-186.67)^2}{3}} = 10.3$ et donc $f = \frac{10.3}{20} = 0.51$.

On choisit dans ce cas une taille d'effet de $f = 0.5$ qui est donc très forte.

Remarque 20 *Afin de calculer la taille d'effet observée, on peut se servir d'une estimation R^2 de η^2, ce qui permettra de calculer $\hat{f}$.*

6.3 La puissance

On traitera d'abord le cas où la taille des différents groupes est la même n (on a donc $N = nk$). Dans le cadre de l'hypothèse alternative définie par la taille d'effet f, la statistique F suit une loi de Fisher-Snedecor (que l'on notera FS) à $df_1 = k - 1$ et $df_2 = N - k$ degrés de liberté mais qui est décentrée. Son paramètre de non-centralité [2] est $\lambda = nkf^2$.

La puissance est alors de

$$1-\beta = P\left(FS(df_1 = k-1, df_2 = N-k, \lambda) > qf(1-\alpha, df1 = k-1, df2 = N-k)\right)$$

En reprenant le problème des détentes horizontales à comparer pour trois groupes d'enfants, nous avions défini une très forte taille d'effet $f = 0.5$. Si on souhaite réaliser un test d'ANOVA avec trois groupes de $n = 20$ enfants pour un niveau $\alpha = 5\%$, la puissance sera de $1 - \beta = 93\%$ d'après la fonction pwr.anova.test.

```
pwr.anova.test(f=0.5,k=3,n=20,sig.level=0.05)
```

1. Notons toutefois, qu'il semble difficile d'estimer correctement cette quantité. D'après mon expérience, nous avons tendance à la surévaluer, comme c'était également le cas avec le coefficient de corrélation linéaire ρ
2. Lorsque les tailles de groupe ne sont pas égales, le paramètre est le même : $\lambda = Nf^2$

Exercice 26 *(Exercice 8.1 p. 357 de Cohen [4]) Un psychologue de l'éducation veut réaliser une expérience pour comparer $k = 4$[3] méthodes d'enseignement. Un total de $N = 80$ enfants seront répartis par groupes de $n = 20$. Les enfants sont évalués sur un critère de performance. Une analyse de variance permettra donc de savoir, au niveau $\alpha = 0.05$ si les résultats moyens des groupes sont différents.*

En choisissant $f = 0.28$ (lire l'ouvrage pour comprendre ce choix), quelle puissance obtient-on ?

Exercice 27 *(Exercice 8.2 p. 358 de Cohen [4]) Une recherche de grande ampleur dans un hôpital psychiatrique concernant des schizophrènes est envisagée. On dispose de 600 patients qui seront répartis en trois groupes égaux. Un indice d'amélioration du comportement sera mesuré suite à trois traitements différents de ces troubles. Le niveau de significativité du test d'analyse de variance est fixé à 0.01. L'équipe de recherche est divisée sur la taille de l'effet. Certains pensent que la variabilité expliquée pourrait être de l'ordre de $\eta^2 = 5\%$ alors que d'autres la voient plutôt vers $\eta^2 = 10\%$.*

- *D'après la relation donnée entre f et η^2, quelles sont les deux tailles d'effets correspondantes ?*
- *Calculer les puissances correspondantes.*

Remarque 21 *Si les tailles des groupes ne sont pas égales, on utilise le calcul précédent avec leur moyenne : $n' = \frac{\sum_{j=1}^{k} n_k}{k} = \frac{N}{k}$*

Exercice 28 *(Exercice 8.3 p. 362 de Cohen [4]) Une classe universitaire de sciences politiques veut entreprendre un sondage des étudiants au sujet de leur l'opinion sur la répartition des responsabilités et droits au niveaux local, fédéral et gouvernemental. Un indice de centralité sera dérivé de ce sondage. Il y a six départements dans cette université (sciences, lettres...). La classe souhaite interroger 300 étudiants tirés au hasard (ce qui signifie que les effectifs des départements ne seront pas les mêmes). Le test sera pratiqué au niveau conventionnel en espérant une taille d'effet de 0.15.*

Calculer le nombre d'étudiants moyen (n') puis la puissance correspondante.

6.4 La taille d'échantillon

En ce qui concerne la taille d'échantillon, on supposera qu'elle est la même n dans chacun des k groupes. Si ce n'est pas le cas (dommage, car le dispositif équilibré est le plus puissant), on multiplie la valeur obtenue par le nombre de groupe ce qui donne la taille totale de l'expérience : $N = nk$ que l'on peut alors répartir suivant nos contraintes[4].

En ce qui concerne l'exemple des détentes horizontales, la puissance avec $n = 20$ enfants par groupe s'élevait à $1 - \beta = 93\%$ ce qui est élevé. On va se

3. Notons que Cohen utilise la notation $u = k - 1$ pour indexer le nombre de groupes, je préfère *pour une fois* m'en écarter et directement utiliser k

4. Si elles ne sont pas trop fortes...

demander si en tablant sur une taille d'effet plus faible de $f = 0.4$, en prenant donc moins de risques, et en se "contentant" de la puissance classique de 80%, la taille d'échantillon nécessaire est fortement modifiée? Il semble que non car la fonction pwr.anova.test nous indique $n = 21$.

```
pwr.anova.test(f=0.4,k=3,power=0.80,sig.level=0.05)
```

Exercice 29 *Reconsidérons l'expérience avec les 4 méthodes d'enseignement de l'exercice 26. Le niveau du test était* $\alpha = 0.05$ *pour détecter un effet de taille* $f = 0.28$. *En prenant* $n = 20$ *enfants par groupe, la puissance est trop faible. Quelle taille d'échantillon faut-il pour atteindre une puissance de* $1 - \beta = 0.80$ *?*

Exercice 30 *En reprenant l'expérience de l'exercice 27, où le niveau du test d'anova est de* $\alpha = 0.01$ *pour* $k = 3$ *groupes, on souhaite calculer les tailles d'échantillons correspondants aux deux tailles d'effets postulées* $f = 0.229$ *et* $f = 0.333$ *afin d'atteindre une puissance de* $1 - \beta = 0.90$.

Chapitre 7

Tests du chi-carré

Il existe deux tests du chi-carré qui sont très utilisés : le test d'ajustement à un ensemble de probabilités et le test d'indépendance entre deux variables qualitatives. Ils partagent la même taille d'effet, le même calcul de puissance (à un coefficient près : le nombre de degrés de liberté du test) et la même taille d'échantillon. Comme dans Cohen [4], les deux tests seront traités simultanément.

7.1 Rappels des tests

7.1.1 Le test d'ajustement

On a N mesures d'une variable qualitative à m catégories qui suit une loi multinomiale de probabilités π_1,..., π_m avec $\sum_{i=1}^{m} \pi_i = 1$. On cherche à tester si $H_0 : \pi_i = P_{0i}$ contre $H_a : \exists \pi_i \neq P_{0i}$.

La statistique employée est celle dite du chi-carré (d'ajustement) :

$$X^2 = \sum_{i=1}^{m} \frac{(n_i - NP_{0i})^2}{NP_{0i}}$$

où n_i est le nombre de réalisations observées pour la catégorie i dans l'échantillon.

Sous l'hypothèse nulle, cette statistique suit une loi du chi-carré χ^2 avec $df = m - 1$ degrés de liberté. On définit donc la région critique comme

$$W = \left\{ X^2 > k(m-1, 1-\alpha) \right\}$$

où $k(m-1, 1-\alpha)$ est le quantile de la loi du χ^2 avec $df = m - 1$ degrés de liberté.

7.1.2 Le test d'indépendance

Dans ce cas, deux variables qualitatives à respectivement r et h catégories sont mesurées simultanément sur N unités statistiques et on s'intéresse au croisement de ces deux variables qui est décrit par un tableau de probabilités π_{ij}.

L'hypothèse nulle correspond à l'hypothèse d'indépendance [1] entre les deux variables et l'hypothèse alternative à l'existence d'une "relation".

La statistique employée est dite du chi-carré d'indépendance de Pearson :

$$X^2 = \sum_{i=i,j=1}^{r,h} \frac{(n_{ij} - n_{i+}n_{+j}/N)^2}{n_{i+}n_{+j}/N}$$

où n_{ij} est le nombre de réalisations correspondant au croisement des catégories i et j, n_{i+} est le nombre de réalisations de la catégorie i et n_{+j} celui de la catégorie j.

Sous l'hypothèse nulle, cette statistique suit aussi une loi du chi-carré χ^2 mais avec $df = (r-1)(h-1)$ degrés de liberté. La forme de la région critique est donc la même (attention aux degrés de liberté toutefois [2]).

7.2 La taille d'effet

L'indice de taille d'effet s'appelle w. On va supposer qu'il y a m cases dans le tableau, que les proportions sont notées P_{0i} dans le cadre de l'hypothèse nulle et P_{1i} pour l'alternative.

On a

$$w = \sqrt{\sum_{i=1}^{m} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}.$$

On voit la grande similarité avec la statistique X^2. Cet indice mesure bien l'écart à l'hypothèse nulle et est indépendant du nombre de mesures N.

On peut *relativement facilement* dans le test d'ajustement définir des tailles d'effet en fonction des proportions attendues dans l'alternative.

(Exercice 7.1 p. 249 de Cohen [4]) un chercheur en marketing souhaite déterminer la préférence des consommateurs entre quatre packagings. Il veut réunir un panel de $N = 100$ personnes. Le niveau du test sera $\alpha = 0.05$ et concernera une hypothèse nulle "d'indifférence" : $H_0 : \pi_i = \frac{1}{4}$.

On suppose que dans la population la vraie distribution des préférences est : $\pi_1 = 0.3750$, $\pi_2 = 0.2083$, $\pi_3 = 0.2083$ et $\pi_4 = 0.2083$, c'est-à-dire que la préférence va au premier packaging, les autres étant "équivalents". La fonction ES.w1 indique que la taille d'effet est $w = 0.289$.

1. soit $\pi_{ij} = \pi_i \times \pi_j$
2. Notons que Cohen [4] choisit d'appeler le nombre de degrés de liberté u. Exceptionnellement je ne suis pas ses notations car *df* s'est, il me semble, imposé dans la littérature statistique

```
P0<-rep(1/4,4)
P1<-c(0.3750,0.2083,0.2083,0.2083)
ES.w1(P0,P1)
```

En revanche, il faut *imaginer* la forme des probabilités attendues dans l'alternative pour le test d'indépendance, c'est-à-dire un tableau de probabilités pour ensuite calculer "l'équivalent indépendant" ce qui permet d'obtenir w (voir fonction ES.w2).

Ainsi, imaginons [3] que le chercheur en marketing souhaite déterminer si les hommes et les femmes expriment la même préférence quant aux quatre packagings ou si, comme il le pense, les femmes ont une préférence plus grande pour les deux premiers packagings, du type $\pi_1 = 0.40$, $\pi_2 = 0.40$, $\pi_3 = 0.10$ et $\pi_4 = 0.10$. Les femmes représentant 40% des consommateurs de ce produit, on en déduit le tableau de probabilités ci-dessous.

Packaging	1	2	3	4
Hommes	0.225	0.125	0.125	0.125
Femmes	0.16	0.16	0.04	0.04

La fonction ES.w2 permet de calculer l'écart à l'indépendance de ce tableau de probabilité qui sert de taille d'effet. On obtient ici $w = 0.256$.

```
Prob<-matrix(c(0.225,0.125,0.125,0.125,0.16,0.16,0.04,0.04),nrow=2,byrow=T)
print(Prob)
ES.w2(Prob)
```

Une autre solution repose sur la relation qu'entretient w avec le plus connu V de Cramer qui est $w = V\sqrt{min(r,h)-1}$.

Sinon Cohen [4] propose les tailles d'effets conventionnelles suivantes : $w = 0.1$ (petite), $w = 0.3$ (moyenne) et $w = 0.5$ (grande).

7.3 La puissance

La puissance se calcule à nouveau à l'aide d'une loi décentrée, celle du chi-carré cette fois. On montre que

$$1 - \beta = P\left(\chi^2(df, \lambda) > k(df, 1-\alpha)\right)$$

où

3. En fait, c'est assez difficile à imaginer. On ne voit pas comment il est possible de donner des indications aussi précises sur ce qui possible. Et quant à trouver des données correspondant exactement à la situation décrite par une telle tableau croisée, c'est peu probable.

— $k(df, 1-\alpha)$ est le quantile de la loi du χ^2 centrée avec df degrés de liberté. On a généralement $df = m - 1$ pour le test d'ajustement [4] et $df = (r-1)(h-1)$ pour le test d'indépendance.
— et $\chi^2(df, \lambda)$ qui est une loi du χ^2 décentrée avec df degrés de liberté et $\lambda = nw^2$ comme paramètre de non centralité.

Dans l'exemple de comparaison des quatre packagings où la taille d'effet est de $w = 0.289$, avec $df = 4-1 = 3$, $\alpha = 0.05$ et $N = 100$ on obtient une puissance de $1 - \beta = 67\%$ (un peu faible).

```
pwr.chisq.test(sig.level=0.05,N=100,df=(4-1),w=ES.w1(P0,P1))
```

Exercice 31 *(Exercice 7.2 p. 250 de Cohen [4]) Une autre application classique du test d'ajustement consiste à vérifier si la distribution d'une variable numérique suit bien une loi spécifiée (ici la loi normale bien sûr...). On détermine alors des valeurs de la loi normale (des quantiles) pour construire des intervalles et on compte le nombre d'occurences de la variable étudiée dans chacun de ces intervalles. On cherche alors à les comparer aux valeurs exactes de probabilités des intervalles que fournit la loi spécifiée.*

En choisissant neuf valeurs de la loi normale d'après la méthode de Hays, on obtient le jeu de probabilités suivant : $H_0 : \pi_1 = 0.020, \pi_2 = 0.051, \pi_3 = 0.118, \pi_4 = 0.195, \pi_5 = 0.232, \pi_6 = 0.195, \pi_7 = 0.118, \pi_8 = 0.05, \pi_9 = 0.020$.

Après plusieurs essais, le psychométricien choisit une taille d'effet de $w = 0.20$. *Le niveau du test est fixé à* $\alpha = 0.10$. *Attention, lors de l'ajustement de la loi normale, deux paramètres ont dû être estimés, ce qui fait "tomber" le nombre de degrés de liberté à* $df = 9 - 1 - 2 = 6$. *Avec* $N = 200$ *observations, quelle est la puissance de ce test de normalité ?*

Si on passe au test d'indépendance avec la relation sexe et choix de packaging, pour la taille d'effet précédemment calculée $w = 0.256$, au niveau $\alpha = 5\%$, le nombre de degrés de liberté est cette fois de $df = (2-1) \times (4-1) = 3$, ce qui donne une puissance de $1 - \beta = 87\%$, ce qui est bien, si on emploie $N = 200$ consommateurs des deux sexes.

```
pwr.chisq.test(sig.level=0.05,N=200,df=(2-1)*(4-1),w=ES.w2(Prob))
```

Exercice 32 *(Exercice 7.3 p. 251 de Cohen [4]) Un chercheur en sciences politiques s'intéresse à la relation entre le sexe (H,F) et la préférence politique (Démocrate, Républicain ou Indépendant). Elle a à sa disposition un échantillon de* $N = 140$ *votants et souhaite réaliser un test au niveau* $\alpha = 0.01$.

La taille d'effet (voir l'ouvrage de Cohen) qui l'intéresse est de $w = 0.346$. *Quelle est (1) le nombre de degré de liberté du test d'indépendance et (2) la puissance de ce test*[5] *?*

4. et un peu moins si des paramètres sont estimés voir exercice 31
5. Le résultat est légérement différent de celui de Cohen : $1 - \beta = 85\%$

7.4 La taille d'échantillon

L'exemple du chercheur en marketing avec pour $N = 100$ consommateurs une puissance de $1 - \beta = 67\%$ montre que l'échantillon n'était pas assez grand. En reprenant les mêmes caractéristiques, c'est-à-dire $\alpha = 0.05$, $df = 3$, $w = 0.289$ mais cette fois-ci en désirant une puissance de $1-\beta = 80\%$, il faut $N = 131$ consommateurs.

```
pwr.chisq.test(sig.level=0.05,power=0.80,df=3,w=0.289)
```

Exercice 33 *(Exercice 7.8 p. 270 de Cohen [4]) Un psychiatre étudie la relation entre des groupes ethniques ($r = 5$) et un diagnostic dans une clinique ($h = 6$). Pour détecter un faible effet $w = 0.1$ avec un niveau de signification de $\alpha = 0.05$ et avec une puissance de $1 - \beta = 0.80$, (1) quel nombre de degrés de liberté doit-on employer dans le test d'indépendance et (2) quelle taille d'échantillon est-elle nécessaire ?*

Chapitre 8

Tests dans le modèle linéaire général

Le modèle linéaire général permet de réunir l'analyse de variance et la régression simple et de les généraliser en analyse de variance à plusieurs facteurs, régression multiple, régression polynomiale et analyse de covariance.

Il s'agit donc d'une situation très importante qui englobe la plupart des chapitres précédents.

8.1 Rappels sur les tests

Deux modèles seront considérés :

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \tag{8.1}$$

qui est le modèle le plus simple où les variables prédictrices peuvent être numériques ou catégorielles grâce à un système de codage (contrastes). On s'intéresse au test $H_0 : \beta_1 = \cdots = \beta_p = 0$.

Le test se fait sur la base d'une décomposition de la somme des carrés en deux parties, l'une correspondant à la prédiction par le modèle et l'autre à l'erreur : $SS_T = SS_X + SS_E$. Le test repose alors sur la statistique $F_1 = \frac{SS_X}{SS_E} \times \frac{n-p-1}{p}$ qui suit sous l'hypothèse nulle une loi de F($u = p$,$v = n - p - 1$).

Le second modèle est :

$$y = \alpha_0 + \alpha_1 z_1 + \cdots + \alpha_k z_k + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \tag{8.2}$$

où il existe dans le modèle des variables prédictrices supplémentaires (les z_j) et on teste également l'hypothèse $H_0 : \beta_1 = \cdots = \beta_p = 0$ mais une fois prises en compte les variables supplémentaires.

Cette fois le test repose sur la décomposition $SS_T = SS_{X+Z} + SS_E = SS_Z + (SS_{X+Z} - SS_Z) + SS_E$ et sur la statistique : $F_2 = \frac{SS_{X+Z} - SS_Z}{SS_E} \times \frac{n-p-k-1}{p}$ qui suit sous l'hypothèse nulle une loi de F($u = p$,$v = n - p - k - 1$).

8.2 Tailles d'effet

Les tailles d'effet définies vont être "inspirées" de la statistique F en éliminant la partie correspondant aux degrés de liberté et en écrivant la première partie *en version probabiliste* en fonction des carrés des coefficients de corrélation multiple théoriques (respectivement sur les sev engendrés par les variables X puis par les variables X et Z) :

$$f_1^2 = \frac{R^2_{Y|X}}{1 - R^2_{Y|X}}$$

et

$$f_2^2 = \frac{R^2_{Y|(X+Z)} - R^2_{Y|Z}}{1 - R^2_{Y|(X+Z)}} = \frac{R^2_{YX|Z}}{1 - R^2_{YX|Z}}$$

où $R^2_{YX|Z}$ est le carré du coefficient de la corrélation multiple partielle entre Y et X par rapport à Z [1].

8.3 Puissance des tests

Pour le calcul de la puissance on va utiliser une loi F décentrée de degrés de liberté u et v (voir ci-dessus suivant le test) et de paramètre de non centralité : $\lambda = (u + v + 1)f^2$.

En reprenant l'exemple 9.1 p424 de Cohen [4] il s'agit d'une situation de régression multiple avec 5 prédicteurs : âge, éducation, expérience, score sur un test d'aptitude verbale et un autre d'extraversion (toutes ces variables sont numériques) pour expliquer les ventes de 95 individus. On va supposer que le pourcentage théorique d'explication des ventes par les 5 variables est de $R^2_{Y|X} = 10\%$.

On en déduit que $f_1^2 = \frac{0.10}{1-0.10}$, que $u = 5$, $v = 95 - 5 - 1 = 89$ et on prend conventionnellement $\alpha = 0.05$. On obtient $1 - \beta = 0.673$.

```
pwr.f2.test(f2=0.1/(1-0.1),u=5,v=89,sig.level=0.05)
```

8.4 Taille d'échantillon

La puissance dans l'exemple précédent n'était pas complètement satisfaisante. On peut donc se demander pour $1 - \beta = 0.80$ quelle est la taille d'échantillon pour un effet $f_1^2 = \frac{0.10}{1-0.10}$. Il s'avère qu'elle est de $v = 115.1043 = n-u-1$ soit $n = 115.1 + 5 + 1 = 121.1$.

1. autrement dit $R^2_{YX|Z} = \frac{R^2_{Y|(X+Z)} - R^2_{Y|Z}}{1 - R^2_{Y|Z}}$

```
pwr.f2.test(f2=0.1/(1-0.1),u=5,power=0.80,sig.level=0.05)
```

Bibliographie

[1] Avanzini G. et Pfister R. (1994) Le phénomène de l'arbitrage à domicile en football : mythe ou réalité ? Science et motricité, 21, pp. 48-52.

[2] Bret C., Rahmani A., Messonnier L., Bedu E. et Lacour J.R. (2001) Relation entre la concentration sanguine de lactate mesurée en fin de course et la performance sur 100m. Science et motricité, 42, pp.24-28.

[3] Champely S. (2003) Statistique vraiment appliquée au sport. Éditions de Boeck.

[4] Cohen J. (1988) Statistical power analysis for the behavioral sciences. Lawrence Erlbaum associates, publishers.

[5] Grondin S. et Trudeau F. (1991) Date de naissance et Ligue Nationale de Hockey : analyse en fonction de divers paramètres. STAPS, 26, pp. 37-45.

[6] Laurencelle L. (2007) Inventer ou estimer la puissance statistique ? Quelques considérations utiles pour les chercheur. Tutorial for quantitative methods in psychology, 3(2), pp.35-42.

[7] Lenth R.V. (2001) Some pratical guidelines for effective sample-size determination.

[8] R : Copyright 2005, The R Foundation for Statistical Computing Version 2.2.1 (2005-12-20 r36812) ISBN 3-900051-07-0

[9] Stanley R.K., Protas, E.J. et Jankovic J. (1999) Exercise performance in those having Parkinson's disease and healthy normals. Medecine and Science in Sports and Exercises, 31, 6, pp.761-766.

[10] Sundberg S. (1982) Maximal oxygen uptake in relation to age in blind and normal boys and girls. Acta Poeditrica Scandinavica, 71, pp. 603-608.

[11] Vignal B., Champely B. et Terret Th. (2000) Piscines d'hier, loisirs d'aujourd'hui ? Espaces, 66, pp. 143-156.