Déterminer la taille des échantillons notion sous-jacente : puissance d'un test

Claire Chabanet, Fabrice Dessaint

26 mai 2015

1 Objectif

Lors de la planification d'une expérience, il est nécessaire de choisir la taille des échantillons, autrement dit le nombre d'unités expérimentales (individus, placettes, parcelles...) sur lesquelles seront appliqués les différents traitements étudiés. Ce choix est bien évidemment contraint par les possibilités financières ou humaines. Cependant il est avant tout lié à l'espérance que l'on a de mettre en évidence un effet d'une taille donnée avec une probabilité donnée. La justification du choix a priori des tailles d'échantillon est de plus en plus demandée dans les publications.

L'objectif de cette note est double : (*i*) donner les explications nécessaires à la compréhension des notions sous-jacentes (puissance d'un test, risque de première et de deuxième espèce), (*ii*) donner les moyens de calculer la taille d'échantillon, avec R, dans les cas les plus courants.

2 La puissance d'un test

2.1 Hypothèse nulle, alternative, et statistique de test

Imaginons le cas d'un test de *Student* dont le but est de savoir si la moyenne de la variable étudiée dans la population traitée est supérieure à la moyenne dans la population témoin. On réalise donc un test unilatéral. L'**hypothèse nulle** H_0 est l'hypothèse d'égalité entre les deux moyennes, alors que l'**hypothèse alternative** H_1 indique que la moyenne dans la population traitée est supérieure à la moyenne dans la population témoin.

$$H_0 : \mu_{\text{traité}} = \mu_{\text{témoin}}$$
$$H_1 : \mu_{\text{traité}} > \mu_{\text{témoin}}$$

On mesure la variable étudiée sur les individus de deux échantillons issus de ces deux populations. On calcule une **statistique de test** qui permettra de décider entre $H0$ et H_1. Cette statistique doit être d'autant plus grande que les moyennes sont différentes. Si on choisit de réaliser un test de *Student*, la statistique calculée est la statistique de *Student*. Elle est égale au rapport entre la différence des deux moyennes et l'écart type de cette différence [1].

$$t = \frac{\bar{x}_{\text{traité}} - \bar{x}_{\text{témoin}}}{s_{\text{différence}}}$$

1. Ici, les moyennes sont notées $\bar{x}_{\text{témoin}}$ et $\bar{x}_{\text{traité}}$. Ce sont des estimations des vraies moyennes inconnues dans les populations témoin et traité que l'on note $\mu_{\text{témoin}}$ et $\mu_{\text{traité}}$.

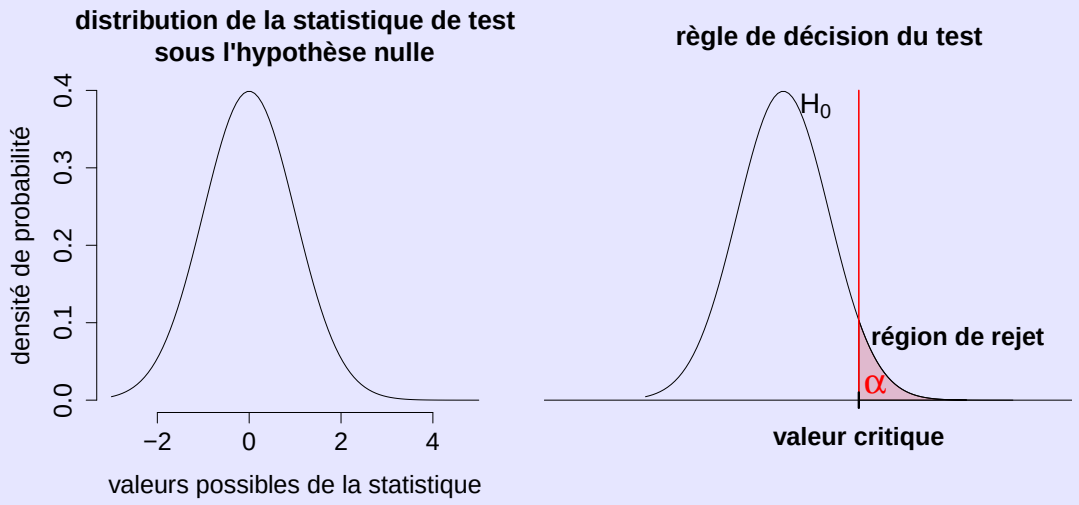

FIGURE 1 – Distribution de la statistique de test sous l'hypothèse nulle, et règle de décision du test

2.2 Règle de décision du test et risque de première espèce

La règle de décision entre H_0 et H_1 est la suivante : si la statistique de test est trop grande et si elle dépasse un certain seuil appelé valeur critique, on rejette l'hypothèse nulle H_0 au profit de l'hypothèse alternative H_1 (Figure 1).

Si l'hypothèse nulle H_0 est vraie, les moyennes des populations sont égales et la statistique de *Student* est distribuée autour de 0. La densité de distribution représente les probabilités associées à chacune des valeurs possibles de la statistique de test (Figure 1 gauche). La valeur 0 est la plus probable, mais il y a peu de chance que la statistique calculée soit exactement égale à 0. En revanche, elle a de grandes chances de se trouver dans une zone autour de 0 définie par la densité de probabilité.

Si la statistique de test dépasse un seuil appelé valeur critique, on rejette l'hypothèse nulle. La valeur critique est définie par le **risque de première espèce** α choisi. Si on choisit $\alpha = 0{,}05$ comme risque de première espèce, on choisit d'avoir 5 % de chances de conclure au rejet de l'hypothèse nulle alors que celle ci est vraie (on dit « sous H_0 »). La valeur critique est alors définie de la façon suivante : la probabilité d'observer une valeur de la statistique supérieure à la valeur critique, est égale à 5%, sous l'hypothèse nulle. C'est donc la valeur qui définit la région de rejet matérialisée en rouge (Figure 1 droite).
Ce schéma représente la **règle de décision** du test, c'est à dire la règle qui conduit à rejeter ou non H_0. Cette règle découle du choix de α, risque de première espèce.

2.3 Risque de deuxième espèce et puissance du test

Si la réalité est autre, la statistique de test ne sera pas distribuée comme représenté sous H_0 mais suivra une autre distribution décalée d'une valeur égale à la vraie différence entre les moyennes (Figure 2 gauche). On représente ici une distribution possible en pointillé, il s'agit d'une des multiples réalités possibles. La règle de décision choisie auparavant définit une zone sous la courbe H_1 représentée en bleu. Cette zone correspond à la probabilité

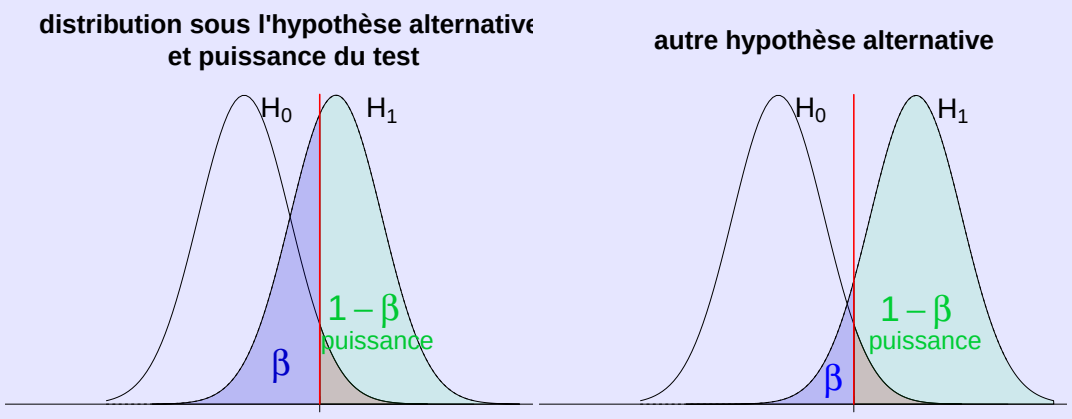

FIGURE 2 – Risque de deuxième espèce β et puissance sous deux hypothèses alternatives différentes

d'accepter à tort l'hypothèse nulle. En effet cette zone est définie par la valeur critique. Pour toute valeur de la statistique plus petite que la valeur critique, on ne peut pas rejeter H_0 alors que pour toute valeur supérieure, on rejette H_0. La zone en bleu correspond bien à un non rejet de H_0, mais cette décision est prise à tort, puisque « l'on est sous H_1 ». On note β la probabilité correspondant à la zone en bleu, c'est le **risque de deuxième espèce**.

À la zone sous H_1 représentée en vert, est associée la probabilité $1-\beta$, c'est la probabilité de rejeter H_0 lorsque H_0 est fausse, c'est la **puissance du test**.

Si l'on représente le cas d'une autre hypothèse alternative, où la vraie différence est plus marquée, si de plus la règle de décision est la même, la puissance est supérieure (Figure 2 droite). La puissance dépend donc de l'amplitude réelle de la différence, qui restera inconnue. [2]

2.4 Lien entre puissance $1-\beta$ et α, δ, n, σ

La puissance d'un test dépend du risque α choisi et de la vraie différence entre les moyennes. Elle dépend aussi de la taille des échantillons et de la variabilité du phénomène étudié. En effet, si la variabilité est grande, la distribution des moyennes et donc la distribution de la statistique sont plus étalées que dans le cas contraire. De même, la distribution des moyennes et la distribution de la statistique sont moins étalées lorsque la taille des échantillon augmente.

2. La vraie différence reste inconnue, étant donné que l'on n'a pas accès à l'intégralité de la population, et c'est bien pour cette raison qu'on utilise les tests statistiques. C'est parce que l'hypothèse alternative n'est qu'une hypothèse parmi d'autres qu'elle est représentée en pointillé, et que deux alternatives possibles sont présentées ici (figure gauche et figure droite).

La puissance $1-\beta$ d'un test dépend :
- de la vraie différence δ,
- de la taille des échantillons n,
- de la variabilité du phénomène étudié σ,
- du risque α choisi.

Les 5 paramètres δ, n, σ, α, β sont liés entre eux. De quatre de ces paramètres, on déduit le 5e.

On peut expérimenter cela à l'aide de la fonction `run.power.examp()` du package TeachingDemo, qui illustre graphiquement le concept de puissance et permet à l'utilisateur de modifier interactivement la taille d'échantillon, l'écart-type, la vraie différence, le risque α, et de visualiser les distributions sous l'hypothèse nulle et sous l'hypothèse alternative, ainsi que le risque α et la puissance du test.

```
install.packages("TeachingDemos")
library(TeachingDemos)
run.power.examp()
```

3 Déterminer la taille d'échantillon nécessaire : code R

Quelques unes des fonctions font partie des fonctions R de base [3] , d'autres sont accessibles via le package pwr ou via d'autres packages (Tableau 1).

On trouvera dans cette section des exemples pour les fonctions R de base et pour celles du package pwr. Pour utiliser les fonctions du package pwr, il est nécessaire d'installer le package puis de le charger [4].

```
install.packages("pwr")
library(pwr)
```

Table 1 – Quelques fonctions permettant le calcul de la taille d'échantillons

Fonction	Package	Fonction	Package
power.t.test		pwr.2p.test	pwr
pwr.t.test	pwr	pwr.2p2n.test	pwr
pwr.t2n.test	pwr	pwr.r.test	pwr
power.anova.test		pwr.chisq.test	pwr
pwr.anova.test	pwr	pwr.f2.test	pwr
power.prop.test		lmmpower.lme	longpower
pwr.p.test	pwr	lmmpower.gee	longpower

De nombreux autres packages existent qui permettent de réaliser des calculs de puissance. Parmi eux, on peut citer les packages PoweR, PowerTOST, powerMediation, simsem, MBESS, CP, powerAnalysis, rpsychi, powerpkg, haplo.stats, stats.

Une méthode pour trouver un grand nombre de ces fonctions :

3. R3.1.1
4. L'installation ne se fait qu'une seule fois (à moins d'avoir installé une nouvelle version de R entre temps), alors que le chargement du package est nécessaire à chaque session R.

```
install.packages("sos")
library(sos)
findFn("power")
```

3.1 Test t de Student

Notations : s, s_1, s_2 sont des écarts-type observés, δ est la différence à mettre en évidence (ou vraie différence), d est la taille de l'effet.

$$\delta = \bar{x}_1 - \bar{x}_2$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}}$$

La fonction `power.t.test()` nous indique qu'il faut au moins 4 observations par échantillon pour avoir 80 % de chances de conclure à une différence significative au seuil de $\alpha = 0{,}05$ si la vraie différence est $\delta = 1{,}2$ et si les écarts-type s_1 et s_2 sont égaux à 0,5. Il faudra donc 4 répétitions pour espérer mettre en évidence une différence de 1,2.

```
power.t.test(power=0.80,delta=1.2,sd=.5, sig.level=0.05)      # n = 4
```

On peut aussi utiliser les fonctions `pwr.t.test()` et `pwr.t2n.test()` du package pwr. **Attention**, pour ces deux fonctions l'argument d= permet de spécifier la taille d de l'effet que l'on souhaite mettre en évidence.

$$d = \frac{\delta}{s}$$

```
pwr.t.test(power=0.80,d=1.2/0.5, sig.level=0.05)      # n = 4
```

La fonction `pwr.t2n.test()` permet de gérer le cas où les tailles des deux échantillons ne sont pas identiques.

3.2 ANOVA à un facteur

Utiliser la fonction `power.anova.test()` (voir les exemples de l'aide en ligne).

```
?power.anova.test
power.anova.test(groups=4,between.var=1,within.var=3,power=0.80)  # n=12
```

Si une publication fait état de moyennes données, et si l'on souhaite mettre en évidence des écarts entre les moyennes au moins égales aux écarts préalablement reportés :

```
groupmeans=c(120,140,130,150)
power.anova.test(groups=length(groupmeans),between.var=var(groupmeans),
                                          within.var=500,power=0.90)  # n=15
```

3.3 Comparaison de deux proportions

Utiliser la fonction `power.prop.test()`.

```
power.prop.test(p1=0.50,p2=0.75,power=0.90,sig.level=0.05)  # n=77
```

3.4 Test du chi2

Utiliser la fonction `pwr.chisq.test()` du package pwr.

```
pwr.chisq.test(w,N,df,sig.level=.05,power)
```

3.5 Test de nullité d'une corrélation

Utiliser la fonction `pwr.r.test()` du package pwr.

```
pwr.r.test(r=.5,power=.8,sig.level=.05)  # n=29
```

4 Compléments

Quelques références bibliographiques, trouvées au cours de recherches, sans prétention à l'exhaustivité et sans sélection :

- `http://www.statmethods.net/stats/power.html`
- Hoenig, J.M., Heisey, D.M. *The Abuse of Power : The Pervasive Fallacy of Power Calculations for Data Analysis* (2001). The American Statistician 55, 19-24. (`http://www.vims.edu/people/hoenig_jm/pubs/hoenig2.pdf`)
  Les auteurs insistent sur le fait que le calcul de puissance se fait a priori et non a posteriori. Ce calcul se fait a priori, pour une différence minimale δ que l'on souhaite mettre en évidence, il ne peut et ne doit pas être utilisé a posteriori en utilisant l'effet observé dans l'expérimentation comme indicateur descriptif des résultats.
- Castelloe, J.M. *Sample Size Computations and Power Analysis with the SAS System* (`http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.128.5923&rep=rep1&type=pdf`)
- Stroup, W.W. *Mixed model procedures to assess power, precision, and sample size in the design of experiments* (`http://www.stat.ncsu.edu/people/arellano/courses/st524/Fall08/Homeworks/homewotk2/Stroup_MixedModelspower_sas.pdf`)
  cas du modèle mixte
- Donohue M.C., Edland, S.D., Gamst, A.C. *Power for linear models of longitudinal data with applications to Alzheimer's Disease Phase II study design* (2015) `http://127.0.0.1:20494/library/longpower/doc/longpower.pdf`
  cas d'un modèle GEE (generalized estimating equation)
- Mayr, S., Erdfelder, E., Buchner, A., Faul, F. *A short tutorial of GPower* (2007). Tutorials in quantitative methods for psychology 3(2), 51-59. (`http://www.stat.ncsu.edu/people/arellano/courses/st524/Fall08/Homeworks/homewotk2/Stroup_MixedModelspow`
  `sas.pdf`)