Code

Pratique AMMS : Introduction à la puissance statistique

Bob Verity

August 04, 2022

Dépendances pour Pratique

Veuillez copier et coller le morceau de code ci-dessous dans son intégralité sur votre console pour télécharger les bibliothèques de packages R nécessaires à cette pratique. Si vous rencontrez des difficultés pour installer l'un des packages R, veuillez demander à un instructeur un lecteur flash préchargé.

Hide

```
if (!("tidyverse" %in% installed.packages())) {
  install.packages("tidyverse")
}
```

Chargez maintenant toutes ces bibliothèques dans cette session en utilisant le morceau de code ci-dessous. Veuillez le copier-coller dans son intégralité.

Hide

```
library(tidyverse)
```

Enfin, sourcez les fonctions supplémentaires nécessaires à ce TP en copiant-collant cette fonction:

Hide

```
source("source_functions/power1_utils.R")
```

Introduction aux tests statistiques et à l'analyse de puissance

Très souvent en surveillance moléculaire nous sommes intéressés à répondre à des questions simples et bien définies, telles que :

- La prévalence d'une mutation de résistance aux médicaments a-t-elle augmenté dans ma population au cours des 5 dernières années ?
- L'utilisation des moustiquaires a-t-elle eu un impact sur la diversité génétique des parasites ?
- L'incidence des résultats de TDR faussement négatifs est-elle plus élevée dans une population que dans une autre ?

Nous pourrions simplement mesurer ces quantités et rapporter nos résultats, ce qui est parfois appelé **statistiques descriptives**, mais souvent nous voulons plus que cela. Nous voulons pouvoir "prouver" qu'un effet est réel. Cela signifie tenir compte du rôle du hasard dans nos résultats par le biais de **tests statistiques**, parfois appelés **tests d'hypothèse nulle**. S'il est très peu probable que nos résultats observés se produisent par hasard, cela nous donne plus de confiance qu'un effet est réel. Cela ne "prouve" pas tout à fait que l'effet est réel, car nous aurions toujours pu être très chanceux ou malchanceux, mais cela contrôle au moins la fréquence à laquelle nous arrivons à la mauvaise conclusion.

Une fois que nous avons une bonne maîtrise des tests statistiques, la prochaine idée cruciale à saisir est la **puissance statistique**. Nous pouvons considérer l'**analyse de puissance** comme une sorte de test statistique que nous effectuons avant de voir des données. Au lieu de cela, nous faisons des hypothèses sur la force de l'effet dans le monde réel (par exemple, la différence de prévalence de la résistance aux médicaments) et la taille de notre échantillon, et nous calculons * exactement * la probabilité que nous ayons de détecter cet effet réel avec notre signification choisie test. Si la puissance est faible, nous risquons de manquer des résultats intéressants même s'ils sont là, car notre test statistique ne peut pas exclure la possibilité que nos résultats soient dus au pur hasard. Il est donc extrêmement important que nous effectuions une analyse de puissance avant d'entreprendre tout essai ou enquête sérieux pour nous assurer que nous avons une chance décente de succès.

Aperçu des données

Pour cette pratique, nous travaillerons entièrement avec des ensembles de données inventés ou simulés. Ceux-ci nous permettront de nous familiariser avec les concepts de base, que nous pourrons ensuite appliquer à des ensembles de données du monde réel à un stade ultérieur.

Objectifs pratiques

À la fin de cet exercice pratique, vous devriez être en mesure de :

- Construire un intervalle de confiance à 95%
- Réaliser un t-test pour comparer deux moyennes, et un z-test pour comparer deux proportions
- Calculer la puissance statistique de plusieurs tests
- Interpréter les courbes de puissance et calculer les tailles d'échantillon optimales
- Ajuster la taille des échantillons pour l'abandon attendu
- Commenter les questions d'études sous-alimentées et suralimentées
- Concevoir une étude simple, en tenant compte de considérations statistiques et logistiques

Comparer les COI entre deux populations

Tests statistiques

Il vous a été demandé d'analyser certaines données sur la complexité de l'infection (COI) dans deux populations, dont l'une a connu une mise à l'échelle rapide des moustiquaires et l'autre agissant comme témoin. Le COI a tendance à être plus élevé dans les populations à forte intensité de transmission, donc l'hypothèse ici est que les moustiquaires auront provoqué une baisse du COI en moyenne par rapport à la population témoin.

Chargeons les données COI et examinons rapidement la distribution:

Hide

```
# load COI data
load("data/COI_control.RData")
load("data/COI_nets.RData")

COI_control
COI_nets
```

```
##  [1] 1 2 2 3 1 3 4 2 2 1
##  [1] 1 1 2 1 2 1 2 4 1 2
```

Nous pouvons voir que nous n'avons que 10 échantillons de chaque population, et il est difficile de dire en regardant si le COI est plus élevé dans une population ou dans l'autre.

Q1. Quel est le COI moyen dans chaque population ? Quelle est la variance dans chaque population ? La moyenne est-elle supérieure ou inférieure dans la population disposant de moustiquaires ?

Click For Answer

Le COI moyen dans l'échantillon de contrôle est de 2,1. Il s'agit de notre meilleure *estimation* du COI moyen dans la *population* témoin, mais nous ne serions pas du tout surpris si la valeur de la population différait légèrement de 2,1. Par exemple, il se peut que le COI moyen dans la population soit de 2,0, et il se trouve que nous avons échantillonné par hasard des individus avec des COI légèrement plus élevés.

Nous pouvons représenter notre incertitude dans l'estimation moyenne par l'*erreur standard*. La formule de l'erreur type est :

$$SE = \sqrt{\frac{s^2}{n}}$$

où s^2 est la variance de l'échantillon et n est la taille de l'échantillon.

Q2. Quelle est l'erreur type de la moyenne pour la population témoin ? Quelle est l'erreur type de la moyenne pour la population de moustiquaires ?

Click For Answer

Nous pouvons utiliser l'erreur standard pour calculer un *intervalle de confiance à 95 %*. L'interprétation d'un intervalle de confiance sous une définition fréquentiste peut prêter à confusion. Il indique que si nous devions prélever des échantillons de la même population plusieurs fois (ou répéter cette "expérience" plusieurs fois), nous nous attendrions à ce que 95 % de nos intervalles de confiance calculés contiennent la moyenne de la population. Une interprétation plus directe est que nous sommes sûrs à *95 % que notre intervalle contient la moyenne de la population*. Les intervalles de confiance (IC) sont un moyen utile de visualiser l'incertitude dans une estimation. La formule pour un IC à 95 % (normal) est la suivante :

$$\bar{x} \pm 1,96 \times SE$$

où $\bar{x}$ est la moyenne de l'échantillon.

Q3. Quel est l'IC à 95 % pour la population témoin ? Quel est l'IC à 95 % pour la population des moustiquaires ?

Click For Answer

Q4. Les IC des deux populations se chevauchent-ils ? Qu'est-ce que cela vous dit sur la confiance que nous accordons aux différences entre les moyennes ?

Click For Answer

Nous comparerons les deux moyennes à l'aide du test t de Student à deux échantillons. Nous avons le même nombre d'échantillons dans les deux groupes, ce qui facilite un peu la vie, et nous supposerons également que les variances sont les mêmes entre les groupes.

La formule de la statistique de test est la suivante :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}_1^2 + \hat{s}_2^2}{n}}}$$

où $\bar{x}_1$ et $\bar{x}_2$ sont les moyennes des deux groupes, $\hat{s}_1^2$ et $\hat{s}_2^2$ sont l'échantillon variances des deux groupes, et n est la taille de l'échantillon (la même dans chaque groupe).

Q5. Complétez la fonction ci-dessous pour calculer cette statistique de test à partir des données d'entrée :

Hide

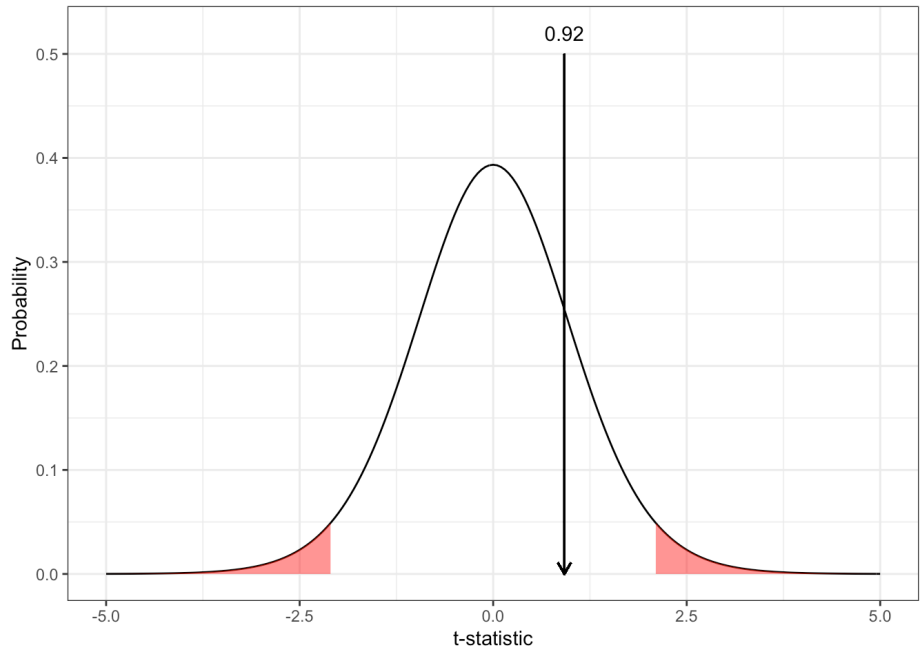
```
get_t_stat <- function(data_series1, data_series2) {
  # calculate both means and variances and the sample size

  # calculate the test statistic

  # return the final value
}
```

Click For Answer

Q6. Utilisez votre fonction complétée pour calculer la statistique du test t sur les données COI. Quelle valeur obtenez-vous ?

Click For Answer

Rappelez-vous de la conférence que chaque statistique de test a une distribution connue sous l'hypothèse nulle. Dans ce cas, la distribution s'appelle la distribution t (ce qui est logique, c'est un test t après tout). Cette distribution a un paramètre "degrés de liberté", qui pour ce test est donné par la formule $2n - 2$. Pour $n = 10$, cela donne une valeur de 18. Le graphique suivant montre la distribution t avec 18 degrés de liberté, et avec notre valeur observée indiquée par une flèche :



Q7. Votre valeur observée de la statistique de test se situe-t-elle dans le corps de la distribution ou dans les queues ? Qu'est-ce que cela vous dit sur la probabilité que cette valeur soit sous l'hypothèse nulle d'aucune différence de COI entre les groupes ?

Click For Answer

Nous pouvons quantifier à quel point cette statistique de test est extrême en utilisant la valeur p.

Q8. Complétez le code ci-dessous pour calculer une valeur de p:

Hide

```
# define sample size and get t statistic
n <- # TO COMPLETE
t_stat <- # TO COMPLETE

# calculate p-value
2*pt(abs(t_stat), df = 2*n - 2, lower.tail = FALSE)
```

Click For Answer

Q9. Quelle est votre valeur p ? Est-ce significatif au niveau $\alpha = 0.05$?

Click For Answer

Il existe un moyen plus simple d'effectuer ce type de test t dans R, nous pouvons utiliser la fonction `t.test()`. Exécutez le code suivant - obtenez-vous les mêmes valeurs que vous avez calculées à la main?

Hide

```
# Two-sample t-test assuming equal variances between groups
t.test(COI_control, COI_nets, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  COI_control and COI_nets
## t = 0.92036, df = 18, p-value = 0.3696
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5130891  1.3130891
## sample estimates:
## mean of x mean of y
##       2.1       1.7
```

Analyse de puissance et calcul de la taille de l'échantillon

Avant de réaliser une étude comme celle ci-dessus, il est bon d'effectuer une analyse de puissance. Cela peut nous indiquer la chance de trouver quelque chose d'intéressant s'il est vraiment là. Plus exactement, cela nous indique la chance de rejeter correctement l'hypothèse nulle compte tenu de certaines hypothèses sur la taille de l'effet et la taille de l'échantillon.

Ci-dessus, nous avons utilisé cette formule pour la statistique du test t :

$t = $

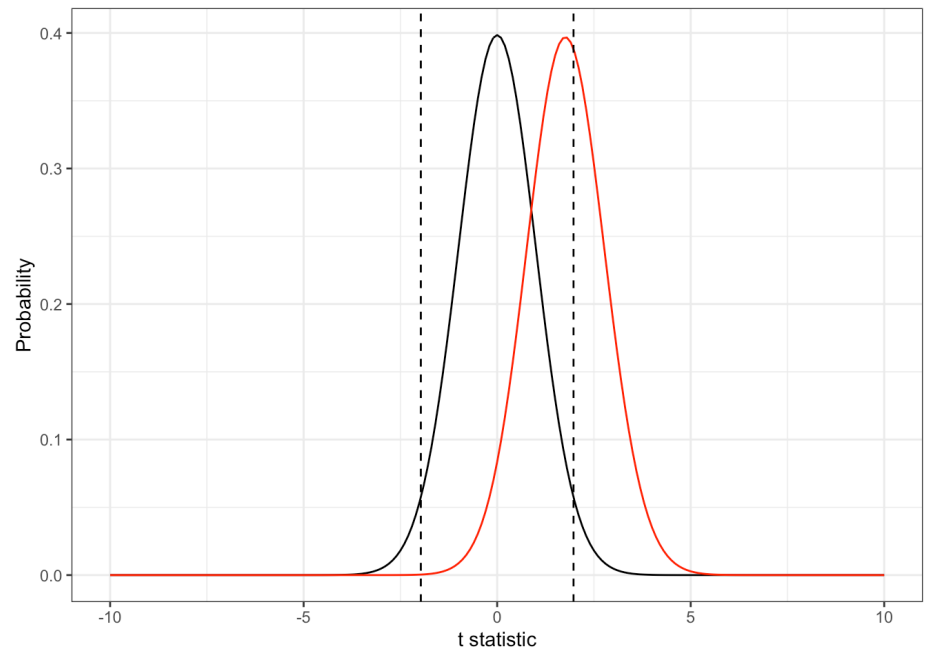
Réécrivons cela légèrement. Premièrement, nous utiliserons *d* pour représenter la différence entre le COI moyen dans les deux populations. C'est cette différence qui nous intéresse - car c'est la différence entre nos deux traitements d'étude, interventions, *etc.*, et donc on lui donne souvent le nom spécial **taille d'effet**. Une taille d'effet plus grande signifie qu'il y a une plus grande différence entre nos populations, et donc nous sommes plus susceptibles de le détecter. Deuxièmement, nous utiliserons s^2 pour représenter la variance du COI dans les deux populations (rappelez-vous, nous avons supposé que la variance est la même dans nos deux populations). La nouvelle version de la formule devient :

$$t = \frac{d}{\sqrt{\frac{2s^2}{n}}}$$

Le bloc de code suivant définit ces paramètres, puis trace la distribution de la statistique de test sous les hypothèses nulle et alternative. Les valeurs critiques, ou points "coupés" pour notre valeur α désignée (niveau de signification), de la distribution nulle sont représentées par des lignes pointillées verticales. Rappelons que la puissance est la proportion de la distribution rouge qui se trouve en dehors de ces lignes.

Hide

```
# input parameters
d <- 0.5
s <- 2
n <- 100
alpha <- 0.05

# produce plot
plot_ttest(d, s, n, alpha)
```



Q10. Expérimentez en modifiant les paramètres d'entrée dans le code ci-dessus

- Que se passe-t-il lorsque vous augmentez la taille de l'effet (`d`) ?
- Que se passe-t-il lorsque vous augmentez l'écart type (`s`) ?
- Que se passe-t-il lorsque vous augmentez la taille de l'échantillon (`n`) ?
- Que se passe-t-il lorsque vous augmentez le seuil de signification (`alpha`) ?

Click For Answer

La fonction suivante renvoie la zone de la courbe rouge qui se trouve en dehors des lignes pointillées (c'est-à-dire la puissance). Copiez cette fonction dans votre console :

Hide

```
# returns the power under the t-test
get_pow_ttest <- function(d, s, n, alpha = 0.05) {
  pt(qt(alpha / 2, df = 2*n - 2), df = 2*n - 2, ncp = d / sqrt( 2*s^2 / n)) +
    pt(qt(1 - alpha / 2, df = 2*n - 2), df = 2*n - 2, ncp = d / sqrt( 2*s^2 / n),
        lower.tail = FALSE)
}
```

Q11. Expérimentez en modifiant les paramètres d'entrée dans le code ci-dessous. Pour `d = 0,5` , `s = 1` , `alpha = 0,05` , pouvez-vous trouver une valeur de `n` qui atteint une puissance de 80 % ?

Hide

```
# input parameters
d <- 0.5
s <- 1
n <- 30
alpha <- 0.05

# calculate power
get_pow_ttest(d, s, n, alpha)
```

```
## [1] 0.4778965
```

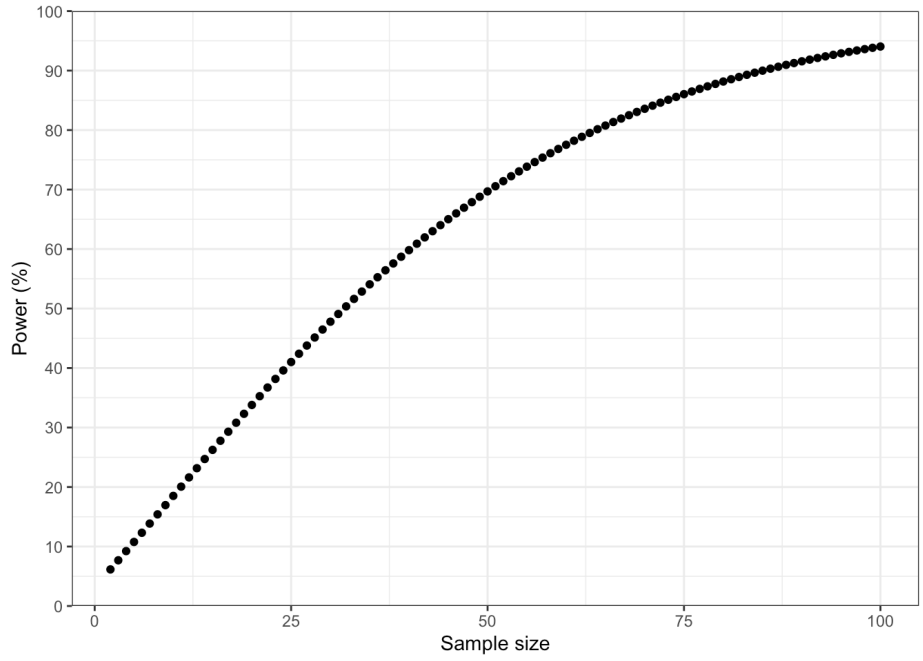
Click For Answer

Parfois, il peut être utile de regarder les **courbes de puissance**. Celles-ci montrent la puissance sur l'axe des y et une autre variable sur l'axe des x, généralement la taille de l'échantillon.

Copiez ce code dans votre console pour produire une courbe de puissance en fonction de 'n' :

Hide

```
# input parameters
d <- 0.5
s <- 1
n <- 2:100
alpha <- 0.05

# plot power curve
qplot(x = n, y = get_pow_ttest(d, s, n, alpha)*100) + theme_bw() +
  scale_y_continuous(breaks = seq(0, 100, 10), limits = c(0, 100), expand = c(0,
        0)) +
  xlab("Sample size") + ylab("Power (%)")
```



Q12. Expérimentez avec différentes valeurs des paramètres d'entrée dans le code ci-dessus et voyez comment ceux-ci modifient la forme de la courbe. Que remarquez-vous sur la forme de cette courbe ? La puissance augmente-t-elle davantage quand on passe de `n = 25` à `n = 50`, ou de `n = 50` à `n = 75` ?

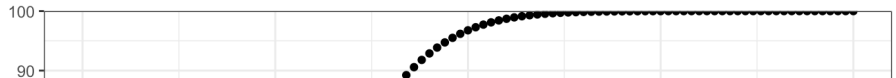
Click For Answer

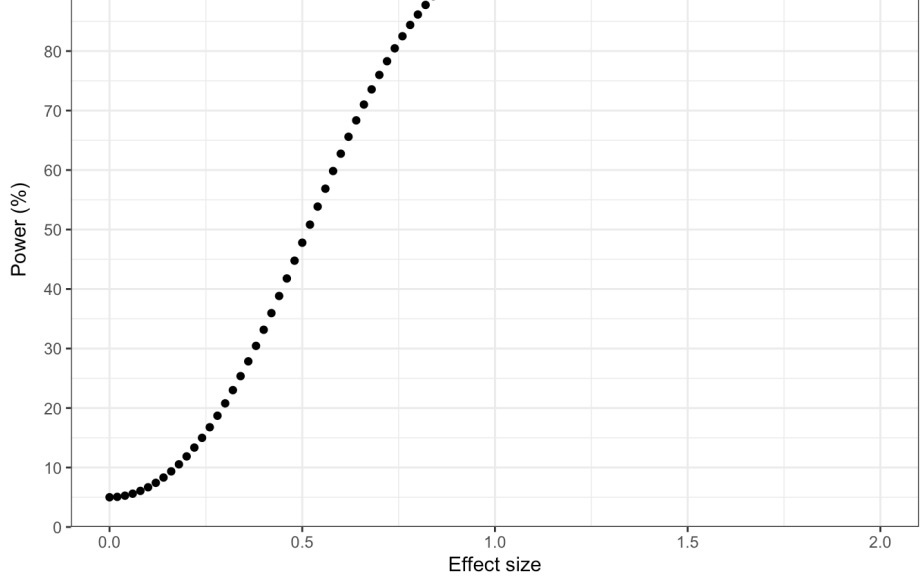
Les courbes de puissance sont idéales pour avoir une idée du nombre d'échantillons que nous pourrions vouloir dans un monde parfait, cependant, en réalité, il existe d'autres contraintes. Il n'est peut-être pas possible d'un point de vue logistique d'obtenir des échantillons de grande taille, ou cela peut être trop coûteux. Cela ne signifie pas que nous devrions abandonner complètement l'analyse de puissance - nous devrions plutôt essayer de travailler dans les limites. Une façon d'y parvenir est de fixer la taille de l'échantillon et de regarder à la place quelle taille d'effet nous sommes capables de détecter.

Le code suivant produit une courbe de puissance pour une taille d'échantillon fixe, et avec la taille d'effet (`d`) sur l'axe des x. Copiez ce code dans votre console et expérimentez différentes valeurs des paramètres d'entrée.

Hide

```
# input parameters
d <- seq(0, 2, l = 101)
s <- 1
n <- 30
alpha <- 0.05

# plot power curve
qplot(x = d, y = get_pow_ttest(d, s, n, alpha)*100) + theme_bw() +
  scale_y_continuous(breaks = seq(0, 100, 10), limits = c(0, 100), expand = c(0,
        0)) +
  xlab("Effect size") + ylab("Power (%)")
```





Q13. Imaginez que la taille de votre échantillon soit fixée par des contraintes logistiques à `n = 17`. Produire une courbe de puissance sous cette limitation. Quelle taille d'effet pouvez-vous détecter avec une puissance de 80 % ? Qu'est-ce que cela signifie pour le COI dans les deux groupes que vous comparez ?

Click For Answer

D'après l'analyse ci-dessus, nous pouvons voir que l'étude originale, qui utilisait une taille d'échantillon de "n = 10", avait une puissance extrêmement faible. Nous n'étions capables de détecter une différence significative entre les groupes que s'il y avait une différence de COI moyen supérieure à 1,5, ce qui est une différence assez importante. D'autre part, si nous voulions détecter des différences de COI jusqu'à 0,5, nous aurions besoin d'une taille d'échantillon plus proche de `n = 65`. Compte tenu de ces résultats, il n'est pas surprenant que nous ayons obtenu un résultat non significatif dans l'analyse des données réelles. Malheureusement, cette étude a probablement été une perte de temps et d'argent. Il a peut-être généré des statistiques descriptives intéressantes, et les résultats peuvent être utilisés comme données pilotes lors de la conception d'études futures, mais en termes de réponse à la question scientifique clé, il était voué à l'échec dès le départ. ***La dure vérité est que toutes les études ne valent pas la peine d'être faites !***

Comparaison de la prévalence des mutations DR

Tests statistiques

Votre programme national de lutte contre le paludisme (PNLP) vous a demandé d'établir si les mutations au locus de résistance à la chloroquine *pfcrt* ont augmenté de manière significative dans votre zone d'étude entre 2005 et 2020. Deux enquêtes transversales ont été menées, avec un grand nombre d'échantillons obtenus et séquencés avec succès chaque année.

Chargeons les données et regardons :

Hide

```
# load COI data
load("data/pfcrt.RData")

head(pfcrt)
```

```
##   year       ID pfcrt
## 1 2005 ID2005.1     0
## 2 2005 ID2005.2     0
## 3 2005 ID2005.3     0
## 4 2005 ID2005.4     1
## 5 2005 ID2005.5     1
## 6 2005 ID2005.6     0
```

Nous pouvons voir que nos données sont organisées dans un data.frame, et en *format long*, ce qui signifie que tous les facteurs (par exemple, l'année) sont en colonnes. Pour chaque ID individuel, nous avons une valeur binaire 1/0 indiquant si l'échantillon contenait la mutation *pfcrt*. Notre première tâche est de résumer ces données pour nous dire :

- Le nombre d'échantillons obtenus chaque année
- La proportion d'échantillons contenant la mutation *pfcrt* (c'est-à-dire la prévalence de la mutation)

Q14. Complétez le code suivant pour 1) regrouper les données par année, et 2) résumer pour obtenir la prévalence des mutations *pfcrt*. La prévalence de la mutation a-t-elle augmenté ou diminué avec le temps ?

Hide

```
pfcrt_summary <- pfcrt %>%
  group_by( # TO COMPLETE ) %>%
  summarise(n = n(),
            prev = #TO COMPLETE )

pfcrt_summary
```

Click For Answer

Pour notre test statistique, nous voulons comparer deux valeurs comme dans l'exemple COI ci-dessus, mais cette fois nos valeurs sont des proportions, et sont donc contraintes d'être comprises entre 0 et 1. Le test approprié ici n'est pas le test t, mais plutôt le test Z à deux proportions. La statistique de test est calculée comme suit :

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\chapeau{p}_2(1-\chapeau{p}_2)}{n_2}}}$$

où $\hat{p}_1$ et $\hat{p}_2$ sont les prévalences dans les deux groupes et n_1 et n_2 sont les tailles d'échantillon. $\bar{p}$ est la prévalence moyenne, calculée comme suit : $\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}$ Le code suivant calcule cette statistique :

Hide

```
# get values from summary table
n <- pfcrt_summary$n
p <- pfcrt_summary$prev
p_bar <- sum(n*p) / sum(n)

# calculate Z statistic
Z <- (p[1] - p[2]) / sqrt(p_bar*(1 - p_bar) * (1/n[1] + 1/n[2]))
Z
```

```
## [1] -14.2472
```

Pour le test z, la distribution de la statistique de test sous l'hypothèse nulle est la distribution z, également appelée **distribution normale**. Pour la distribution normale, les "queues" à 5 % sont à -1,96 et +1,96.

Q15. Votre valeur Z observée est-elle dans le corps ou les queues de la distribution ? Qu'est-ce que cela vous dit sur la probabilité que nous voyions une valeur aussi extrême par hasard ?

Click For Answer

Nous pouvons calculer une valeur p comme suit :

Hide

```
# calculate p-value
2*pnorm(abs(Z), lower.tail = FALSE)
```

```
## [1] 4.666238e-46
```

Q16. La différence de prévalence *pfcrt* est-elle significative au seuil de 5 % ? Accepteriez-vous ou rejetteriez-vous l'hypothèse nulle ?

Click For Answer

Comme pour le test t, il existe un moyen plus simple d'effectuer un test Z à deux proportions dans R, nous pouvons utiliser la fonction `prop.test()`. Cela effectue la même analyse que nous venons de faire à la main ci-dessus :

Hide

```
prop.test(x = n*p, n = n, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  n * p out of n
## X-squared = 202.98, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.3667931 -0.2832069
## sample estimates:
## prop 1 prop 2
##  0.290  0.615
```

Analyse de puissance et calcul de la taille de l'échantillon

Comme précédemment, nous effectuerons une analyse de puissance pour déterminer si cette étude a été bien conçue. Nous garderons les choses simples en supposant le même nombre d'échantillons au cours des deux années. Nous pouvons réécrire la formule de la statistique Z comme suit :

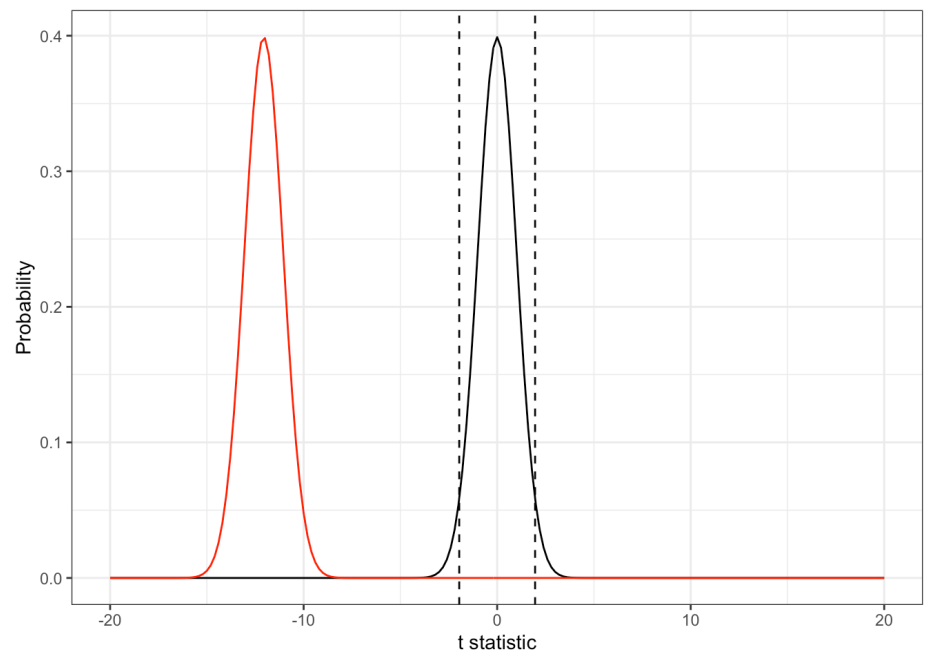
$$Z = \frac{p_1 - p_2}{\sqrt{\frac{\bar{2p}(1-\bar{p})}{n}}}$$

où n est maintenant la taille de l'échantillon pour les deux années (remplaçant n_1 et n_2 dans la version précédente).

Le code suivant trace la distribution de la statistique Z sous les hypothèses nulle et alternative, similaire à ce que nous avons fait pour l'exemple de test t ci-dessus. Expérimentez avec différentes valeurs et voyez comment elles affectent les distributions.

Hide

```
# variable parameters
p1 <- 0.3
p2 <- 0.6
n <- 800
alpha <- 0.05

# produce plot
plot_ztest(p1, p2, n, alpha)
```



Nous pouvons également écrire une fonction pour calculer exactement la puissance (la zone de la distribution rouge qui est au-delà des lignes pointillées) :

Hide

```
# function that returns the power given these parameters
get_pow_ztest <- function(p1, p2, n, alpha = 0.05) {
  p_bar <- mean(c(p1, p2))
  alt_mean <- (p1 - p2) / sqrt(2*p_bar*(1 - p_bar) / n)
  pnorm(qnorm(alpha / 2), mean = alt_mean) + pnorm(qnorm(1 - alpha / 2), mean = al
        t_mean, lower.tail = FALSE)
}
```

Q17. Lors du calcul de la puissance, nous devons faire des hypothèses sur la taille de l'effet (dans ce cas, la prévalence réelle au cours des deux années) et la taille de l'échantillon. Le NMCP vous a demandé de calculer la puissance selon les hypothèses suivantes :

- La prévalence double de 40% à 80%
- La prévalence double de 30% à 60%
- La prévalence double de 20% à 40%
- La prévalence passe de 30% à 45%

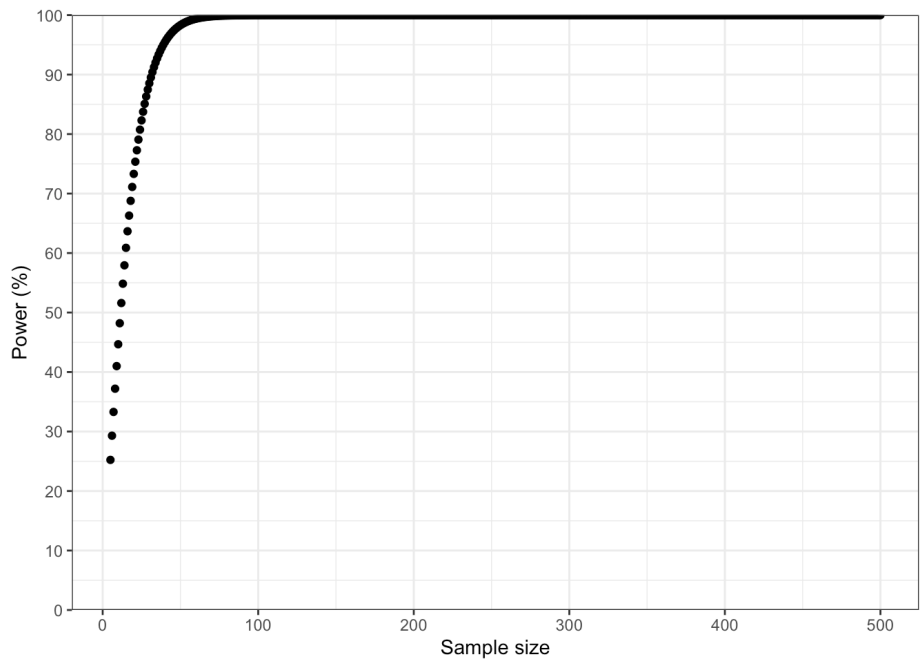
Supposons une taille d'échantillon de `n = 1000` tout au long.

Click For Answer

Nous pouvons produire une courbe de puissance en passant une plage de valeurs de "n" dans notre fonction. Expérimentez avec différents paramètres dans le code ci-dessous pour voir comment ils modifient la forme de la courbe de puissance :

Hide

```
# variable parameters
p1 <- 0.4
p2 <- 0.8
n <- 5:500

# plot power curve
qplot(x = n, y = get_pow_ztest(p1, p2, n, alpha)*100) + theme_bw() +
  scale_y_continuous(breaks = seq(0, 100, 10), limits = c(0, 100), expand = c(0,
      0)) +
xlab("Sample size") + ylab("Power (%)")
```



Q18. Produisez des courbes de puissance pour les quatre scénarios que le PNLP vous a demandé d'explorer. Quelle est la taille approximative de l'échantillon nécessaire dans chaque cas pour atteindre une puissance de 80 % ?

Click For Answer

Parfois, il est possible de trouver une formule pour la taille de l'échantillon. Dans ce cas, la formule peut s'écrire :

$$n = (z_{\alpha/2} + z_{\beta})^2 \frac{2\bar{p}(1-\bar{p})}{(p_1-p_2)^2}$$

où $z_{\alpha/2}$ est la valeur critique à un niveau de signification α (bilatéral), dont nous avons déjà noté qu'il est d'environ 1,96. z_{β} est une valeur similaire, cette fois calculée à partir de β qui est défini comme 1 moins la puissance souhaitée (dans notre cas $\beta = 0,2$ pour une puissance de 80%). La fonction suivante implémente cette formule pour donner la taille d'échantillon nécessaire pour une puissance donnée :

Hide

```
get_n_ztest <- function(p1, p2, alpha = 0.05, power = 0.8) {
  p_bar <- mean(c(p1, p2))
  (qnorm(1 - alpha / 2) + qnorm(power))^2 * 2*p_bar*(1 - p_bar) / (p1 - p2)^2
}
```

Cette formule donne parfois des valeurs non entières, auquel cas elles doivent être arrondies au nombre entier le plus proche.

Q19. Utilisez cette fonction exacte pour calculer la taille d'échantillon requise pour chacun des quatre

scénarios demandés par le PNLP. Quelles valeurs obtenez-vous ? Laquelle de ces valeurs devriez-vous adopter comme recommandation finale ?

Click For Answer

Q20. Le coût de l'étude, prenant en compte le temps clinique, le temps de laboratoire et le coût du séquençage, est estimé à 50 USD par échantillon. Quel est le coût estimé de l'étude originale qui a utilisé 800 échantillons en 2005 et 1200 échantillons en 2020 ? Quel est le coût de votre nouvelle conception d'étude sur la base de votre analyse de puissance (rappelez-vous que le "n" que vous avez calculé est pour *chacune* des deux années) ?

Click For Answer

Ce que nous avons vu ici est un exemple d'une étude surpuissante. Bien que nous puissions penser qu'il n'y a pas de mal à collecter plus d'échantillons, nous devons garder à l'esprit que chaque étude a des coûts et que les fonds pourraient être mieux dépensés ailleurs. Dans ce cas, nous avons constaté qu'une conception correctement alimentée était plus de 5 fois moins chère que l'étude originale. Le PNLP aurait pu mener la même étude dans 5 régions différentes du pays, ou sur plusieurs années, et aurait quand même réalisé une économie de coûts. Ainsi, bien qu'il soit généralement judicieux d'être prudent lors du calcul de la taille des échantillons, en optant pour des valeurs plus élevées en cas de doute, cette approche a des limites et il est possible de collecter trop d'échantillons. Quelle que soit la situation, réaliser une analyse de puissance formelle **avant** la réalisation de l'étude est un excellent moyen d'explorer ces questions.

Une dernière chose à noter est que les formules utilisées dans ce calcul de puissance ne sont que des approximations. En règle générale, les méthodes de simulation ou de calcul fourniront des résultats plus précis car elles ne font pas les mêmes approximations. La fonction `get_pow_ztest_exact()` ci-dessous utilise cette approche pour calculer exactement la puissance, qui peut être comparée à la fonction `get_pow_ztest()` utilisée ci-dessus. Pouvez-vous trouver des combinaisons de paramètres où ils sont d'accord/pas d'accord ?

Hide

```
get_pow_ztest(p1 = 0.2, p2 = 0.5, n = 20)
get_pow_ztest_exact(p1 = 0.2, p2 = 0.5, n = 20)
```

```
## [1] 0.5116136
## [1] 0.5308455
```

Gérer le décrochage

Une chose dont nous devons être conscients lors du calcul de la taille des échantillons est l'abandon. Cela fait référence à tout ce qui fait que la taille de notre échantillon final est inférieure à ce que nous avions initialement prévu. Le décrochage peut survenir pour de nombreuses raisons, notamment :

- Personnes retirant leur consentement à l'étude
- Personnes mourant ou migrant hors de la zone d'étude
- Échantillons ne répondant pas aux critères requis pour l'analyse (par exemple, être non-* vivax * dans une étude * falciparum *)
- Échantillons perdus ou contaminés
- Échantillons en échec de séquençage, par exemple en raison d'une faible parasitémie

Nous devons tenir compte de l'abandon dans nos calculs de taille d'échantillon pour nous assurer qu'il nous reste suffisamment d'échantillons pour notre analyse finale. La formule pour ajuster le décrochage est assez simple :

$$n_{\text{ajusté}} = \frac{n_{\text{original}}}{1 - p_{\text{abandon}}}$$

où n_{original} est la taille de l'échantillon brut que nous obtenons de notre analyse de puissance, et p_{dropout} est la proportion d'abandons attendus pour une raison particulière. Par exemple, si la taille de notre échantillon d'origine est `n = 100` et que nous prévoyons un abandon de 20 %, nous faisons $n_{} = 100 / (1 - 0,2) = 100 / 0,8 = 125 $. Nous avons donc besoin de 125 personnes pour tenir compte de ce nombre d'abandons. Nous pouvons facilement vérifier ceci : si nous avons 125 personnes et que 20 % d'entre elles abandonnent, nous perdons 25 personnes, ce qui nous ramène à 100.

Q21. Un programme de contrôle souhaite déterminer si la fréquence des mutations *dhps* K540E dans son pays est supérieure à 10 %. Si tel est le cas, ils prévoient de remplacer les médicaments de première intention par la sulfadoxine-pyriméthamine. Ils prévoient de reproduire cette étude dans 5 régions distinctes à travers

le pays. Un calcul statistique de la taille de l'échantillon a révélé que 220 échantillons seront nécessaires pour obtenir la puissance souhaitée. Cependant, chacun des 5 laboratoires impliqués dans le traitement des échantillons a des niveaux d'expérience différents, ce qui entraîne des taux différents d'échantillons perdus. La proportion estimée d'échantillons perdus dans chacun des laboratoires est la suivante :

Labo1 : 10 % Labo2 : 3 % Labo3 : 14 % lab4 : 25 % lab5 : 9 %

Quelle taille d'échantillon ajustée est nécessaire pour chacune des 5 régions ? N'oubliez pas d'arrondir les valeurs au nombre entier le plus proche.

Click For Answer

Parfois, nous devons effectuer l'ajustement ci-dessus plusieurs fois. Par exemple, nous pourrions nous attendre à perdre 5 % des échantillons en raison du retrait du consentement, et sur les échantillons restants, nous prévoyons de perdre 10 % en raison d'un échec du séquençage. Dans ce cas, nous devons d'abord ajuster pour la perte de 5 %, puis * en utilisant la nouvelle valeur ajustée * nous devons ajuster pour la perte de 10 %. Notez que cela ne donne pas exactement la même chose que si nous comptabilisions la totalité des 15 % en une seule fois.

Q22. Un programme de contrôle mène une étude dans laquelle ils suivent des personnes sur une période de 6 mois et mesurent l'incidence du paludisme. Les parasites seront génotypés périodiquement pour déterminer si des génotypes identiques ou différents sont présents. Le calcul statistique de la taille de l'échantillon a indiqué qu'ils avaient besoin de 400 échantillons au total sur la période de 6 mois. Ils s'attendent à perdre 15 % des échantillons en raison de la perte de suivi (par exemple, les personnes qui migrent ou abandonnent l'étude). Parmi ceux qui sont séquencés, ils s'attendent à ce que 10 % des échantillons échouent. Quelle est la taille finale ajustée de l'échantillon dont ils ont besoin ?

Click For Answer

Mise en pratique

J'espère qu'à ce stade, vous vous sentez à l'aise avec les bases de l'analyse de puissance et du calcul de la taille de l'échantillon. Les exemples ci-dessus ont été conçus pour illustrer les principaux points d'apprentissage, mais les analyses du monde réel ont tendance à être un peu plus compliquées et impliquent une réflexion créative. Essayez d'aborder le problème plus réaliste suivant.

Q23. (exercice long) Vous avez été recruté par le NMCP de Zambie pour mener une étude sur l'évolution de la prévalence des mutations *dhps* K540E. Ils ont une série d'échantillons qui ont été recueillis dans une étude pilote en 2001 à partir de 5 emplacements d'échantillonnage différents. Ces échantillons ont été séquencés et donnent des estimations de base de la prévalence des mutations K540E dans chacun des emplacements. Ils prévoient de mener une étude actuelle dans les mêmes endroits pour déterminer si la prévalence a changé de manière significative au cours de cette période.

Voici les données pilotes :

Hide

```
load("data/Zambia_pilot.RData")
Zambia_pilot
```

```
##   location total_samples K540E
## 1    Kabwe            80    11
## 2    Ndola            24     6
## 3  Chipata           110    13
## 4    Mansa            90    14
## 5   Lusaka            70    12
```

Vous avez également quelques informations sur les contraintes logistiques. Les échantillons seront séquencés dans plusieurs laboratoires différents, et vous disposez d'estimations de la fraction susceptible d'échouer à chaque emplacement :

Hide

```
load("data/Zambia_logistics.RData")
Zambia_logistics
```

```
##   location fail_fraction
## 1    Kabwe          0.10
## 2    Ndola          0.06
## 3  Chipata          0.30
## 4    Mansa          0.02
## 5   Lusaka          0.04
```

Le budget de l'étude permet de séquencer 1000 échantillons au total sur les 5 sites.

Dans cet exemple, nous connaissons la taille de l'échantillon dans le premier groupe (n_1) et nous essayons de déterminer la taille de l'échantillon requise dans le deuxième groupe (n_2). Cela conduit à la formule suivante :

$$n_2 = \frac{p_2(1-p_2)}{\frac{(p_1-p_2)^2}{(z_{\alpha/2}+z_\beta)^2} - \frac{p_1(1-p_1)}{n_1}}$$

Cette formule n'est valable que tant que $n_1 > (z_{\alpha/2} + z_\beta)^2 \frac{p_1(1-p_1)}{(p_1-p_2)^2}$. Si n_1 est inférieur à cette valeur, il est impossible d'atteindre la puissance souhaitée avec n'importe quelle taille d'échantillon.

Il vous a été demandé de :

1. Estimer la prévalence des mutations K540E en 2001 à partir des données pilotes.
2. Écrivez une nouvelle fonction pour implémenter la formule de taille d'échantillon ci-dessus. Vous pourriez trouver utile de regarder la fonction `get_n_ztest()` définie dans le précédent comme guide.
3. Utilisez votre nouvelle fonction pour effectuer le calcul de la taille de l'échantillon dans chaque emplacement, en supposant que la prévalence a doublé depuis l'étude pilote. Visez 80% de puissance.
4. Ajuster la taille des échantillons pour tenir compte de l'abandon.
5. Calculez votre nombre total d'échantillons pour l'étude. Est-ce dans le budget ?
6. Si ce n'est pas le cas, y a-t-il un emplacement que vous pourriez supprimer pour respecter votre budget ? Justifiez votre choix d'emplacement.
7. Rédigez un paragraphe de synthèse à envoyer au PNLP avec votre recommandation. Cela devrait décrire vos hypothèses ainsi que vos conclusions. Il doit contenir une valeur claire pour la taille d'échantillon requise dans chaque emplacement.

Click For Answer