Gaël Millot

Comprendre et réaliser les tests statistiques à l'aide de R

Manuel de biostatistique

3e édition

de boeck

Comprendre et réaliser les tests statistiques à l'aide de R

3ème édition

"Aucun chercheur – sauf s'il est assuré d'avoir du génie et, en outre, beaucoup de chance – ne peut plus ignorer la méthode statistique"
(Schwartz, 1963)

AVANT-PROPOS

Un statisticien et un biologiste sont condamnés à mort. On leur accorde une dernière faveur.
— Je voudrais donner une grande conférence sur la statistique devant tout le monde, dit le statisticien.
— Accordé, répond le juge. Et pour vous ?
Le biologiste n'exprime aucune hésitation :
— Je souhaiterais être exécuté le premier.

Cette histoire ironise avec justesse sur le fait qu'une partie des scientifiques ont un problème avec l'outil statistique. Ceux qui s'adressent à moi le reconnaissent volontiers, qu'ils soient étudiants ou chercheurs, en biologie, médecine, sociologie, ou psychologie. C'est souvent son langage mathématique qui rebute. Pourtant la statistique n'est pas une science abstraite. Jouer aux jeux de hasard, prendre une assurance, écouter la météo, placer son argent, s'intéresser aux sondages : nous côtoyons en permanence, dans notre vie de tous les jours, les probabilités et la statistique. Et, à la différence de la science statistique, l'utilisation de l'outil statistique ne requiert pas de compétences mathématiques élevées. Les chercheurs en science statistique sont comme les développeurs de voitures : ils travaillent sur des domaines ultra-pointus qui visent à améliorer la performance des concepts existants, ou à en créer de nouveaux. Mais pour ceux qui désirent se servir de l'outil, il n'est pas nécessaire d'arriver à leur niveau de maîtrise. Peu de conducteurs savent comment fonctionne une boîte d'embrayage. Mais beaucoup d'entre eux peuvent conduire une voiture en toute sécurité. J'ai donc choisi d'écrire ce livre en m'adressant d'abord et avant tout aux débutants, ainsi qu'à ceux que la statistique rebute, avec deux objectifs :

Le premier est de vous apprendre à manier les tests statistiques sans vous tromper. Si je reprends la parabole sur la voiture, l'objectif est de vous donner le mode d'emploi des tests, et de vous enseigner le code de la route lorsqu'il s'agira d'en conduire un, afin d'éviter les accidents. Difficile de savoir si ces derniers sont fréquents ou non en science (Giles, 2006) mais comme pour la conduite des voitures, les esprits évoluent. Et les accidents sont de moins en moins bien perçus (la revue *Nature* par exemple, a décidé de renforcer la qualité des analyses statistiques de ses publications). Ce livre vous aidera, je l'espère, à devenir des conducteurs consciencieux.

Le deuxième objectif est de vous rendre autonome dans la réalisation d'un test. A mon sens, il ne suffit pas de savoir réaliser des tests, encore faut-il maîtriser un logiciel de statistique. Les deux aspects sont indissociables aujourd'hui. J'ai donc choisi de vous enseigner l'utilisation des tests avec le logiciel R, pour six raisons. (1) Il est le plus accessible à tous puisque gratuit, disponible sur internet et compatible avec les systèmes Windows, MAC OS et Linux. (2) Il est, à mon avis, le logiciel de statistique le plus complet et le plus puissant dans de nombreux domaines. Avec R, peu de chance d'avoir à s'investir dans un autre logiciel parce qu'un type d'analyse statistique n'est pas disponible. (3) A condition de respecter quelques règles, les données à analyser peuvent provenir de tableurs type "Excel". (4) Son utilisation et sa réputation sont grandissantes. (5) Il est très pédagogique puisqu'il permet de comprendre d'où provient cette mystérieuse *p value* qui apparaît lorsqu'on réalise un test statistique. (6) Ses capacités graphiques permettent de réaliser la majorité des graphiques de publications (voir quelques exemples dans l'annexe 13). Si vous pensez que R manque de convivialité, ce livre vous est de fait destiné car l'essentiel du travail est déjà réalisé. Le code qui est associé aux 35 tests décrits n'attend plus que vos données pour vous fournir une analyse statistique détaillée. Et puis n'oubliez pas qu'avec R, l'effort que vous fournirez à sa compréhension n'est pas un

investissement à court terme, puisque vous pourrez installer et utiliser ce logiciel n'importe où, sur n'importe quelle machine, dès que vous en aurez besoin. Ce livre est aussi pratique si vous utilisez un autre logiciel de statistique. En effet, tous les exemples d'application des tests sont réalisés à partir du même tableau de données (tableau 3, paragraphe 2.4.1). Tout ce que vous aurez à faire, c'est de trouver le moyen d'importer ce tableau dans votre logiciel. Ensuite, vous pourrez appliquer le test que vous souhaitez, et vérifier que vous retrouvez bien les résultats indiqués dans ce manuel.

Comme mentionné dans le titre, le fil conducteur du livre est la compréhension et l'utilisation des tests statistiques. C'est pourquoi les autres aspects de la statistique, tels que l'estimation, l'intervalle de confiance, la description ou la modélisation ne sont pas développés (les deux premiers points sont abordés dans l'annexe 2). De même, la notion de probabilité mathématique n'est pas traitée. Ceux qui le souhaitent trouveront ces informations dans la plupart des manuels de statistique, notamment ceux cités en référence. Un bon nombre de statisticiens considèrent, à juste titre, que les tests occupent une place abusive en statistique. En effet, l'ensemble des outils et méthodes statistiques disponibles dépasse très largement le cadre des tests, et ces derniers ne sont adaptés qu'à des situations spécifiques. Le manuel proposé cible donc un domaine restreint de la statistique. Il est dédié à ceux qui recherchent un contenu exhaustif sur les tests.

Ce livre comporte cinq chapitres. Le premier est une introduction à l'utilisation de R. Le deuxième développe les notions minimales de la statistique qu'il faut connaître avant d'utiliser un test. Il est souhaitable que ceux qui débutent en statistique commencent la lecture du livre directement par ce chapitre (vous reviendrez au chapitre 1 lorsque vous le jugerez utile). Le troisième chapitre aborde en peu de lignes quelques règles de méthodologie à respecter afin de ne pas fausser l'analyse statistique. Les chapitres 4 et 5 sont complémentaires. Mon expérience me montre que nous sommes demandeurs de méthodes écrites sous forme de procédures. C'est ce qui est proposé dans le chapitre 5. Si vos données sont simples, vous devriez trouver le test qui convient à votre analyse parmi les 35 décrits, et le moyen de réaliser ce test sous R sans difficulté. L'analyse multivariée n'est pas traitée et l'analyse linéaire, ainsi que la survie, ne sont abordées que dans leur forme la plus simple. Elles constitueront néanmoins une bonne introduction à ceux qui doivent s'engager dans des analyses complexes. Mais appliquer une procédure sans la comprendre, c'est un peu comme conduire de nuit avec des lunettes de soleil : le champ de vision devient extrêmement étroit, et la procédure risquée. D'où l'importance du chapitre 4. Celui-ci explique en effet comment fonctionnent les tests statistiques. Il permet donc de comprendre la procédure proposée dans les tests du chapitre 5. Il fournit également les éléments qui développeront votre sens critique vis-à-vis des résultats de tests, les vôtres ou ceux que l'on rencontre dans la littérature. Ce chapitre 4 aborde les aspects les plus techniques de la statistique des tests mais il a été rédigé avec l'objectif d'être facile d'accès. On peut le subdiviser en deux parties : une première qui explique ce qu'est la *p value* d'un test et comment l'obtenir (paragraphe 4.2 à 4.5), et une deuxième qui traite de l'utilisation de la *p value* et de la conclusion d'un test (paragraphe 4.6). Les cinq chapitres de cet ouvrage devraient donc remplir les deux objectifs cités précédemment, à savoir vous rendre autonome dans la réalisation d'un test à l'aide d'un logiciel de statistique, et ce, sans vous tromper.

Pour terminer, je me dois de souligner que ce livre est le fruit d'une écoute mutuelle : la mienne, devant les étudiants des écoles doctorales de l'Université Pierre et Marie Curie se retrouvant dans le besoin d'analyser leurs résultats de recherche à l'aide de tests statistiques, alors qu'ils n'y comprennent pas toujours grand chose ; la leur, me demandant beaucoup d'efforts pour rendre mon cours utile et comestible. J'espère que nombre de lecteurs néophytes en statistique trouveront dans cet ouvrage solutions à leurs problèmes.

PARTICULARITES DE LA TROISIEME EDITION

La deuxième édition avait bénéficié de profondes modifications. L'élan est poursuivi dans cette nouvelle édition avec les trois objectifs récurrents (1) d'apporter des corrections, notamment celles qui étaient disponibles sur le site http://perso.curie.fr/Gael.Millot/, (2) d'insérer des ajouts afin de produire un ouvrage complet aussi bien sur les tests statistiques que sur l'utilisation de R et (3) de mettre à jour l'ensemble du code décrit dans le livre pour qu'il soit compatible avec une version récente de R. Au total, plus de 15% des pages de la deuxième édition ont été retouchées et 39 pages ont été ajoutées.

Concernant la biostatistique, une nouvelle annexe a été introduite pour expliquer en détail les différentes façons de calculer un quantile (annexe 17). Elle s'avère utile dès que l'on s'intéresse aux médianes de distributions. Elle peut également expliquer les différences de résultats obtenus suivant les logiciels de statistique employés.

Concernant R, l'ensemble du code et des explications relatives aux attributs de R ont été mis à jour pour profiter de la version 3.0.2 du logiciel. Cela s'est traduit par quelques modifications critiques dans le manuscrit (fonctions `mean()`, `sd()` et `median()` ne fonctionnant plus sur les data frames par exemple). Le nombre de fonctions décrites a été augmenté (20 pages au total pour l'annexe 19) et l'index a été renforcé. Le paragraphe de la gestion des couleurs bénéficie maintenant d'une nouvelle section sur l'affichage des couleurs adaptées aux personnes présentant des difficultés à distinguer le rouge du vert. Dans l'annexe 14, un modèle de code est proposé afin d'exporter différents types de résultats (texte, tableaux, etc.) dans un fichier texte unique. Dans le paragraphe sur les attributs de répétition (boucles de type *for, while* et *repeat*), une partie ajoutée présente les avantages de la fonction `get()`, ainsi qu'une astuce pour remplir progressivement un tableau ou une liste lors de l'exécution d'une boucle. D'autres astuces sont proposées dans l'annexe 13, afin de minimiser les modifications d'un code graphique pour l'adapter à de nouvelles données. Par ailleurs, de plus en plus de personnes se familiarisent avec R et, ce faisant, un certain nombre d'entre elles s'intéresse à la création de fonctions. L'ouvrage comporte maintenant une annexe détaillée de 25 pages sur ce point particulier.

Au final, j'espère que vous trouverez dans cette troisième édition un contenu exhaustif sur la compréhension et l'utilisation des tests statistiques, ainsi que toutes les informations nécessaires à l'utilisation de R.

PRECISIONS

Tous les exemples du chapitre 5 sont réalisés à partir du tableau 3 (paragraphe 2.4.1), disponible en ligne. Ce qui fait que vous n'avez plus qu'à l'importer dans R (et si vous désirez le consulter ou le modifier, ouvrez-le dans un tableur type "Excel", la lecture en est plus commode). Vous trouverez ce fichier, appelé *mais.txt*, ainsi que des informations supplémentaires sur le livre, à l'adresse suivante : http://perso.curie.fr/Gael.Millot/.

L'ensemble du code de R fourni dans l'ouvrage s'identifie de la façon suivante : les lignes sont écrites en police de caractères "**`courier style gras`**" et débutent par le symbole ">" avec un renvoi à la ligne décalé si le code dépasse une ligne. Si la ligne de code est suivie d'une explication, notée derrière le symbole "#", la police de caractères est alors "*`courier style italique`*". Enfin, les résultats affichés par R sont reportés immédiatement après le code : ils sont en police de caractères "`courier style normal`".

Ce livre exploite la version 3.0.2 64-bit de R, la version 6.1(SP1) 64-bit de Windows 7 et la version 10.6.8 de MAC OS X. Les graphiques présentés sont obtenus avec une résolution d'écran de 1280 × 1024 pixels.

Pour la traduction des termes anglais de statistique, voir le site internet de P. Legendre et L. Legendre : "lexique anglais-français d'écologie numérique et de statistique".

SOMMAIRE

5.13. Le test t de Student de comparaison de moyennes

5.13.1. Comparaison d'une moyenne observée à une valeur théorique

5.13.1.1. Méthode

Application

Comparaison d'une moyenne observée m à une moyenne théorique $m_{théo}$.

Variable mesurée

Une variable quantitative

Conditions d'application

(1) Les individus formant l'échantillon doivent être choisis un par un et aléatoirement dans l'ensemble de la population visée.

(2) La variable quantitative doit suivre une loi normale dans la population visée (voir le point 23 du paragraphe 5.1).

(3) La variable quantitative peut être continue ou discrète (voir le point 14 du paragraphe 5.1).

Voir le test des signes de Wilcoxon (paragraphe 5.15) si les conditions ne sont pas remplies.

Hypothèses de test

H_0 : $\mu = m_{théo}$ La moyenne théorique $m_{théo}$ est la moyenne *réelle* dans la population visée.

$H_{1\ bilat}$: $\mu \neq m_{théo}$ La moyenne théorique $m_{théo}$ n'est pas la moyenne *réelle* dans la population visée.

$H_{1\ unilat\ d}$: $\mu > m_{théo}$ La moyenne *réelle* est strictement supérieure à la moyenne théorique $m_{théo}$ dans la population visée.

$H_{1\ unilat\ g}$: $\mu < m_{théo}$ La moyenne *réelle* est strictement inférieure à la moyenne théorique $m_{théo}$ dans la population visée.

Loi de probabilité suivie par la VT sous H_0

La VT t suit une loi de Student à $\nu = n - 1$ degrés de liberté (ddl).

Calcul de la VT

Le tableau initial est le suivant :

Individu	Variable
1	x_1
...	
i	x_i
...	
n	x_n

i : numéro de l'individu
n : effectif total de l'échantillon
x_i : valeur de la variable quantitative de l'individu i

Formule de la VT : $$t = \frac{m - m_{théo}}{\frac{s}{\sqrt{n}}}$$

t : VT
m : moyenne de la variable quantitative
$m_{théo}$: moyenne théorique
s : écart type non biaisé de la variable quantitative
n : effectif total de l'échantillon

Conclusions du test

Si H_0 ne peut être rejetée :
La différence entre m et $m_{théo}$ n'est pas significative : elle ne semble due qu'à la fluctuation d'échantillonnage. Donc, rien ne permet de réfuter, pour l'instant, que la moyenne $m_{théo}$ est celle existant dans la population visée. Dans certains cas, H_0 peut être *acceptable* (avec le risque β de se tromper en affirmant cela).

Si H_0 peut être rejetée (test bilatéral) :
La différence entre m et $m_{théo}$ est significative : elle n'est pas simplement due à la fluctuation d'échantillonnage. La moyenne $m_{théo}$ n'est donc pas celle existant dans la population visée (avec le risque α de se tromper en affirmant cela).

Vous adapterez sans difficulté cette conclusion à celle d'un test unilatéral (le risque de se tromper en la formulant étant de α).

Remarques

(1) Prendre garde au sens de soustraction du numérateur de t lorsqu'on travaille en unilatéral. Le sens $m - m_{théo}$ doit être respecté pour que le t calculé soit conforme aux hypothèses H_1 unilatérales droite et gauche telles qu'elles sont formulées dans "Hypothèses de test". Sinon il faut procéder à l'opposé. Ainsi en unilatéral droit, si le numérateur est $m_{théo} - m$, la *p value* se détermine à gauche du t calculé et non à droite.

(2) Attention : lorsque l'effectif de l'échantillon est grand, le test reste valable si la distribution de la variable quantitative ne suit pas une loi normale dans la population, à condition qu'elle

ne s'en éloigne pas trop (test robuste). Une distribution en dos de chameau ou exponentielle par exemple, fausserait complètement le résultat du test.

(3) La moyenne est sensible aux individus extrêmes. Un ou deux individus très éloignés des autres déplacent la moyenne qui n'est plus alors au centre du nuage des individus. Dans ce cas, il peut être préférable de travailler avec la médiane.

(4) Lorsque n augmente, la loi de probabilité suivie par la VT t sous H_0 converge vers une loi normale centrée réduite. A partir de n = 30, l'approximation est considérée comme suffisamment bonne pour pouvoir utiliser les VT M ou Z à la place de t, dont les caractéristiques sont les suivantes :

$$M = m$$

$$Z = \frac{m - m_{théo}}{\frac{s}{\sqrt{n}}}$$

M : VT suivant approximativement une loi normale $N(m_{théo}, s/\sqrt{n})$
Z : VT suivant approximativement une loi normale centrée réduite N(0, 1)

La VT M est l'estimateur de moyenne (voir l'annexe 2). La VT Z est l'équivalent de la VT M centrée réduite, dont la formule est identique à celle de la VT t. L'emploi de Z se justifiait à l'époque des calculs manuels, la table de la loi normale centrée réduite étant la plus manipulée. Il n'a plus lieu d'être aujourd'hui.

La VT M conserve un intérêt pédagogique puisqu'elle offre un calcul de *p value* très simple. C'est pour cette raison qu'elle a été choisi dans les paragraphes 4.5 et 4.6 au détriment des autres.

5.13.1.2. Exemples avec R

Description de la fonction `t.test()`

Les données à soumettre doivent correspondre à un vecteur contenant les valeurs de la variable quantitative. Elles peuvent également être dans une colonne de data frame.

Exécuter `t.test(...)[]` pour obtenir la totalité des résultats du test :

`$statistic`	valeur de la VT t_{calc}
`$parameter`	ν (ddl de la loi de Student)
`$p.value`	*p value*
`$conf.int`	intervalle de confiance à 95% par défaut
`$estimate`	valeurs des moyennes observées
`$null.value`	valeur de $m_{théo}$ ou de différence des moyennes observées (0 par défaut)
`$alternative`	type de test : `"two.sided"` pour bilatéral (par défaut), `"greater"` pour unilatéral droit, `"less"` pour unilatéral gauche
`$method`	méthode employée
`$data.name`	nom des données de départ utilisées

La fonction emploie la VT t quel que soit n, jamais la VT Z.

Test bilatéral

Des experts signalent que cette année, la hauteur moyenne des pieds provenant de la parcelle Est devrait être de 265 cm. Monsieur Léon voudrait savoir ce qu'il en est.

Hypothèses de test :
H_0 : La moyenne de hauteur est de 265 cm sur l'ensemble de la parcelle Est.
H_1 : Elle n'est pas de 265 cm sur l'ensemble de la parcelle Est.

Tableaux créés qui aident à la compréhension du test :
obs1 : tableau initial (sert à la fonction `t.test()`)
obs2 : tableau initial sans les lignes contenant au moins une donnée manquante NA (sert à la décomposition du test)

Obtention des tableaux :

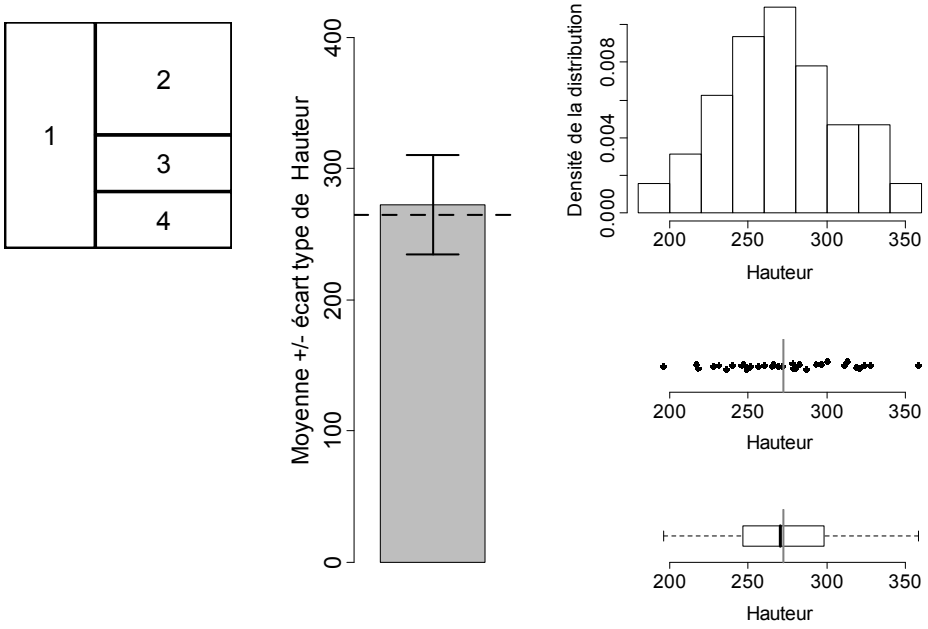
```
> mais<-read.table("C:/Users/Gael/Desktop/mais.txt",header=TRUE) # pour importer
      votre fichier de données (voir le paragraphe 1.6.1.2)
> obs1<-data.frame(mais[which(mais$Parcelle=="Est"), "Hauteur"]) # on sélectionne
      les chiffres de la colonne "Hauteur" quand une ligne indique "Est" dans la
      colonne parcelle
> names(obs1)<-c("Hauteur")
> if(any(is.na(obs1) == TRUE)){nombre.lig.na<- length(table(which(is.na(obs1) ==
      TRUE, arr.ind = TRUE)[,1]))} else{nombre.lig.na<-0} # recherche le nombre de
      lignes de obs1 avec au moins un NA
> cat("Nombre de lignes avec NA :", nombre.lig.na, "\n")
Nombre de lignes avec NA : 1

> obs2<-na.omit(obs1) # supprime les lignes de obs1 avec au moins un NA
> obs2
   Hauteur
2      278
3      260
4      217
...
31     283
32     272
33     246
```

Représentations graphiques :
Fenêtre 2 : emplacement des 4 graphiques dans la fenêtre 3
Fenêtre 3 : représentations de la variable quantitative "Hauteur"
- Graphique 3.1 : moyenne observée +/- écart type, et moyenne théorique de la hauteur (ligne horizontale pointillée)
- Graphique 3.2 : histogramme de la distribution de la hauteur
- Graphique 3.3 : distribution des individus et moyenne observée (ligne verticale grise)
- Graphique 3.4 : distribution en boîte et moyenne observée (ligne verticale grise)

Obtention des graphiques (voir le paragraphe 1.9 pour plus d'informations) :

```
> graphics.off()
> # Fenêtre 2
> zone<-matrix(c(1,1,1:4), ncol=2) ; layout(zone, widths=c(1,1.5), heights=c(2,1,1))
      ; par(cex=5, lwd=10) ; layout.show(max(zone))
> # Fenêtre 3
> windows() # remplacer par quartz() sous Mac OS ou par X11() sous Linux
> layout(zone, widths=c(1,1.5), heights=c(2,1,1))
> par(mar=c(5.1, 6, 4.1, 2.1), mgp=c(3.5,1,0))
> # Graphique 3.1
```

```
> m.theo<-265 # valeur de la moyenne théorique
> library(gplots)
> barplot2(mean(obs2[, 1]), xlim=c(0, 1.5), ylim= range(0, m.theo, obs2[, 1],
     max(m.theo, obs2[, 1])*1.2), names.arg="", plot.ci=TRUE, ci.l= (mean(obs2[,
     1])-sd(obs2[, 1])), ci.u= (mean(obs2[,1])+sd(obs2[,1])), xpd=FALSE,
     ylab=paste("Moyenne +/- écart type de ", names(obs1)[1]), cex.lab=2.2,
     cex.axis=2.2, ci.lwd=2)
> abline(h=m.theo, lty="77", lwd=2) # la moyenne théorique est représentée par une
     ligne horizontale pointillée
> # Graphique 3.2
> hist<-hist(obs2[, 1], freq=FALSE, main=NULL, xlab= names(obs1), ylab="Densité de
     la distribution", cex.lab=1.8, cex.axis=1.8)
> # Graphique 3.3
> par(cex.lab=1.8, cex.axis=1.8, bty="n")
> stripchart(obs2[,1], method="jitter", jitter=1, vertical=FALSE,
     xlab=names(obs1)[1], pch=16, cex=1.2, xlim=range(hist$breaks))
> abline(v=mean(obs2[,1]), lwd=2, col=gray(0.5)) # la moyenne observée est
     représentée par une ligne verticale grise
> # Graphique 3.4
> par(bty="n")
> boxplot(obs2[,1], horizontal=TRUE, xlab=names(obs1)[1], pch=16, cex=1.2,
     cex.lab=1.8, cex.axis=1.8, bty="l", ylim=range(hist$breaks))
> abline(v=mean(obs2[,1]), lwd=2, col=gray(0.5)) # la moyenne observée est
     représentée par une ligne verticale grise
```



L'histogramme semble proche d'une distribution normale. Les graphiques 3.3 et 3.4 ne révèlent pas d'individus extrêmes. Sur le graphique 3.1, la moyenne théorique est proche de celle observée.

Réalisation du test :

```
> m.theo<-265 # valeur de la moyenne théorique
> t.test(obs1,mu=m.theo)

        One Sample t-test

data:  obs1
t = 1.1151, df = 31, p-value = 0.2734
alternative hypothesis: true mean is not equal to 265
95 percent confidence interval:
 258.8607 285.9518
```

```
sample estimates:
mean of x
 272.4062
```

Conclusion :
L'hypothèse H_0 ne peut être rejetée au seuil $\alpha = 5\%$. D'après l'échantillon, rien ne permet de réfuter que la moyenne de hauteur est de 265 cm sur l'ensemble de la parcelle Est.

Calcul des paramètres de notre échantillon (sur obs2 pour éviter les NA) :

```
> n<-length(obs2[, 1]) # effectif
> m<- mean(obs2[, 1]) # moyenne observée
> s<- sd(obs2[, 1]) # écart type
> param <- data.frame(n, m, s)
> names(param) <- c("Effectif", "Moyenne", "Ecart.type")
> param
  Effectif  Moyenne Ecart.type
1       32 272.4062   37.57046
```

Valeur de la VT t sur notre échantillon :

```
> m.theo<-265 # valeur de la moyenne théorique
> t.calc<- (m-m.theo)/(s/n^0.5)
> t.calc
[1] 1.115134
```

La VT t suit une loi de Student à $\nu = n - 1$ ddl et le test est bilatéral. La *p value* correspond donc à la plus petite des deux aires sous la courbe, à droite et à gauche de la valeur t calculée, qu'il faut multiplier par deux. Obtention de la *p value* :

```
> min(pt(t.calc, n-1, lower.tail=FALSE), pt(t.calc, n-1))*2
[1] 0.2733637
```

Pour retrouver les valeurs de l'intervalle de confiance à 95% (voir l'annexe 2) :

```
> m+qt(0.025,n-1)*s/n^0.5
[1] 258.8607
> m+qt(0.025,n-1,lower.tail=FALSE)*s/n^0.5
[1] 285.9518
```

Test unilatéral droit

Monsieur Léon voudrait savoir si la hauteur moyenne des pieds de la parcelle Est est supérieure à 265 cm. Qu'elle soit inférieure ne l'intéresse pas.

Hypothèses de test :
H_0 : La moyenne de hauteur est de 265 cm sur l'ensemble de la parcelle Est.
H_1 : Elle est strictement supérieure à 265 cm sur l'ensemble de la parcelle Est.

Les tableaux, les représentations graphiques, le calcul de la VT et des paramètres, sont identiques à ceux du test bilatéral.

Réalisation du test :

```
> m.theo<-265 # valeur de la moyenne théorique
> t.test(obs1,mu=m.theo, alternative="greater")

        One Sample t-test

data:  obs1
t = 1.1151, df = 31, p-value = 0.1367
alternative hypothesis: true mean is greater than 265
95 percent confidence interval:
 261.1453      Inf
```

```
sample estimates:
mean of x
 272.4062
```

Dans cet exemple, la conclusion n'est pas différente de celle du test bilatéral.

La VT t suit une loi de Student à $\nu = n - 1$ ddl et le test est unilatéral droit. La *p value* correspond donc à l'aire sous la courbe à droite de la valeur t calculée. Obtention de la *p value* :

```
> pt(t.calc, n-1, lower.tail=FALSE)
[1] 0.1366819
```

Pour retrouver la borne inférieure de l'intervalle de confiance à 95% (voir l'annexe 2) :

```
> m+qt(0.05,n-1)*s/n^0.5
[1] 261.1453
```

Notons que l'effectif de l'échantillon est supérieur à 30. La remarque 4 du paragraphe 5.13.1.1 précédent indique que, dans ce cas, les VT M et Z peuvent être utilisées à la place de t car l'approximation est suffisamment bonne. Notre échantillon peut donc servir d'exemple supplémentaire au paragraphe 4.5, dans un cadre ou l'on ne connait rien de la population à l'origine de l'échantillon, soit la totalité de la récolte de Monsieur Léon sur la parcelle Est. Tout ce qui précède la réalisation du test, à savoir la formulation des hypothèses de test, l'analyse graphique et le calcul des paramètres, n'est pas différent de ce qui a été vu ci-dessus. Seuls changent les calculs de la VT et de la *p value* associée.

Valeur de la VT M sur notre échantillon :

```
> Mcalc<-m
> Mcalc
[1] 272.4062
```

La VT M suit une loi normale $N(m_{théo}, s/\sqrt{n})$ et le test est unilatéral droit. La *p value* correspond donc à l'aire sous la courbe à droite de la valeur M calculée. Obtention de la *p value* :

```
> pnorm(Mcalc, m.theo, (s/n^0.5), lower.tail=FALSE)
[1] 0.1323966
```

L'approximation de la *p value* est correcte à deux chiffres après la virgule. On remarque que lorsque l'écart type σ n'est pas connu (il l'était dans le paragraphe 4.5), on exploite celui de l'échantillon s (voir l'annexe 2). La VT Z donne strictement le même résultat que la VT M.

Valeur de la VT Z sur notre échantillon :

```
> m.theo<-265 # valeur de la moyenne théorique
> Zcalc<- (m-m.theo)/(s/n^0.5)
> Zcalc
[1] 1.115134
```

La VT Z suit une loi normale N(0, 1) et le test est unilatéral droit. La *p value* correspond donc à l'aire sous la courbe à droite de la valeur Z calculée. Obtention de la *p value* :

```
> pnorm(Zcalc, lower.tail=FALSE)
[1] 0.1323966
```

Test unilatéral gauche

Monsieur Léon voudrait savoir si la hauteur moyenne des pieds de la parcelle Est est inférieure à 265 cm. Qu'elle soit supérieure ne l'intéresse pas.

Hypothèses de test :
H_0 : La moyenne de hauteur est de 265 cm sur l'ensemble de la parcelle Est.
H_1 : Elle est strictement inférieure à 265 cm sur l'ensemble de la parcelle Est.

Les tableaux, les représentations graphiques, le calcul de la VT et des paramètres, sont identiques à ceux du test bilatéral.

Réalisation du test :

```
> m.theo<-265 # valeur de la moyenne théorique
> t.test(obs1,mu=m.theo, alternative="less")

        One Sample t-test

data:  obs1
t = 1.1151, df = 31, p-value = 0.8633
alternative hypothesis: true mean is less than 265
95 percent confidence interval:
     -Inf 283.6672
sample estimates:
mean of x
 272.4062
```

Dans cet exemple, la conclusion n'est pas différente de celle du test bilatéral.

La VT t suit une loi de Student à $\nu = n - 1$ ddl et le test est unilatéral gauche. La *p value* correspond donc à l'aire sous la courbe à gauche de la valeur t calculée. Obtention de la *p value* :

```
> pt(t.calc, n-1)
[1] 0.8633181
```

Pour retrouver la borne supérieure de l'intervalle de confiance à 95% (voir l'annexe 2) :

```
> m+qt(0.05,n-1,lower.tail=FALSE)*s/n^0.5
[1] 283.6672
```

5.13.2. Comparaison de deux moyennes observées

5.13.2.1. Méthode

Application

Comparaison de deux moyennes observées m_1 et m_2.

Variables mesurées

Une variable quantitative et une variable qualitative à deux classes.

Conditions d'application

(1) Les individus formant l'échantillon doivent être choisis un par un et aléatoirement dans l'ensemble de la population visée.

(2) Les classes de la variable qualitative doivent être exclusives.

(3) Dans la population visée, les distributions de la variable quantitative dans chacune des classes de la variable qualitative doivent suivre une loi normale (voir le point 23 du paragraphe 5.1).

(4) Les variances de la variable quantitative dans chacune des classes de la variable qualitative doivent être identiques dans la population visée ($\sigma^2 = \sigma_1^2 = \sigma_2^2$).

(5) La variable quantitative peut être continue ou discrète (voir le point 14 du paragraphe 5.1).

(6) La variable qualitative est nominale (voir le point 15 du paragraphe 5.1).

(7) La variable qualitative peut être fixée.

Voir la comparaison de deux moyennes observées avec variances différentes (paragraphe 5.13.3), le test de Mann-Whitney-Wilcoxon (paragraphe 5.16) ou de médianes (paragraphe 5.18.2) si les conditions ne sont pas remplies.

Hypothèses de test

H_0 : $\mu_1 = \mu_2$ Les moyennes sont identiques dans la population visée.

$H_{1\ bilat}$: $\mu_1 \neq \mu_2$ Les moyennes sont différentes dans la population visée.

$H_{1\ unilat\ d}$: $\mu_1 > \mu_2$ La moyenne μ_1 est strictement supérieure à la moyenne μ_2 dans la population visée.

$H_{1\ unilat\ g}$: $\mu_1 < \mu_2$ La moyenne μ_1 est strictement inférieure à la moyenne μ_2 dans la population visée.

Loi de probabilité suivie par la VT sous H_0

La VT t suit une loi de Student à $\nu = n_1 + n_2 - 2$ degrés de liberté (ddl).

Calcul de la VT

Le tableau initial est le suivant :

Individu	Variable quantitative	Variable qualitative
1	x_1	c_1
...	...	...
i	x_i	c_i
...	...	...
n	x_n	c_n

i : numéro de l'individu
n : effectif total de l'échantillon
x_i : valeur de la variable quantitative de l'individu i
c_i : valeur de la variable qualitative de l'individu i : soit le nom de la classe 1 soit le nom de la classe 2

Ce tableau peut être décomposé en deux tableaux de test, un pour chaque classe qualitative :

Individu	Classe 1 de la variable qualitative
1	x_{11}
...	...
j	x_{j1}
...	...
n_1	x_{n11}

Individu	Classe 2 de la variable qualitative
1	x_{12}
...	...
g	x_{g2}
...	...
n_2	x_{n22}

j : numéro de l'individu dans la classe 1 de la variable qualitative
g : numéro de l'individu dans la classe 2 de la variable qualitative
n_1 : effectif de la classe 1
n_2 : effectif de la classe 2
n : effectif total de l'échantillon ($n = n_1 + n_2$)
x_{j1} : valeur de la variable quantitative de l'individu j dans la classe 1 de la variable qualitative
x_{g2} : valeur de la variable quantitative de l'individu g dans la classe 2 de la variable qualitative

Formule de la VT :

$$m_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{j1} \qquad \text{et} \qquad m_2 = \frac{1}{n_2}\sum_{g=1}^{n_2} x_{g2}$$

$$t = \frac{m_1 - m_2}{\sqrt{\hat{s}^2 \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ avec } \quad \hat{s}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

t : VT
m_1 : moyenne de la variable quantitative dans la classe 1
m_2 : moyenne de la variable quantitative dans la classe 2
$\hat{s}^2$: estimation de σ^2 (voir la condition d'application 4)
s_1^2 : variance non biaisée de la variable quantitative dans la classe 1
s_2^2 : variance non biaisée de la variable quantitative dans la classe 2
n_1 : effectif de la classe 1
n_2 : effectif de la classe 2

Conclusions du test

Si H_0 ne peut être rejetée :
La différence entre m_1 et m_2 n'est pas significative : elle ne semble due qu'à la fluctuation d'échantillonnage. Donc, rien ne permet de réfuter, pour l'instant, que les moyennes μ_1 et μ_2 sont identiques dans la population visée. Dans certains cas, H_0 peut être *acceptable* (avec le risque β de se tromper en affirmant cela).

Si H_0 peut être rejetée (test bilatéral) :
La différence entre m_1 et m_2 est significative : elle n'est pas simplement due à la fluctuation d'échantillonnage. Les moyennes μ_1 et μ_2 sont donc différentes dans la population visée (avec le risque α de se tromper en affirmant cela).

Vous adapterez sans difficulté cette conclusion à celle d'un test unilatéral (le risque de se tromper en la formulant étant de α).

Remarques

(1) Prendre garde au sens de soustraction du numérateur de t lorsqu'on travaille en unilatéral. Le sens $m_1 - m_2$ doit être respecté pour que le t calculé soit conforme aux hypothèses H_1 unilatérales droite et gauche telles qu'elles sont formulées dans "Hypothèses de test". Sinon il faut procéder à l'opposé. Ainsi en unilatéral droit, si le numérateur est $m_2 - m_1$, la *p value* se détermine à gauche du t calculé et non à droite.

(2) Attention : lorsque l'effectif de l'échantillon est grand, le test reste valable si, au niveau de la population, les distributions de la variable quantitative ne suivent pas une loi normale dans chacune des classes de la variable qualitative, à condition qu'elles ne s'en éloignent pas trop (test robuste). Une distribution en dos de chameau ou exponentielle par exemple, fausserait complètement le résultat du test.

(3) Par contre, le test est très sensible au non-respect de l'égalité des variances. Mais vérifier cette égalité avec le test de Fisher-Snedecor n'est pas souvent opportun car ce test est très sensible au non-respect de la distribution normale dans les deux classes. B. Scherrer cite G.E.P. Box à ce propos : "vérifier préalablement l'égalité des variances revient à mettre à l'eau une barque à rames afin de se rendre compte si les conditions sont suffisamment calmes pour

qu'un paquebot puisse quitter le port" (Scherrer, 1984). Si vous êtes dubitatifs vis-à-vis de la condition d'application 4, reportez-vous au test de Welch (paragraphe 5.13.3).

(4) La moyenne est sensible aux individus extrêmes. Un ou deux individus très éloignés des autres déplacent la moyenne qui n'est plus alors au centre du nuage des individus. Dans ce cas, il peut être préférable de travailler avec la médiane.

(5) Lorsque n augmente, la loi de probabilité suivie par la VT t sous H_0 converge vers une loi normale centrée réduite. Avec $n_1 \geq 30$ et $n_2 \geq 30$, l'approximation est considérée comme suffisamment bonne pour pouvoir utiliser la VT Z à la place de t, dont les caractéristiques sont les suivantes :

$$Z = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Z : VT suivant approximativement une loi normale centrée réduite N(0, 1)

L'emploi de Z se justifiait à l'époque des calculs manuels, la table de la loi normale centrée réduite étant la plus manipulée. Il n'a plus lieu d'être aujourd'hui.

5.13.2.2. Exemple avec R

Description de la fonction `t.test()`

Se reporter à la comparaison d'une moyenne observée à une valeur théorique (paragraphe 5.13.1.2). Les données à soumettre doivent être deux vecteurs contenant les valeurs de la variable quantitative, le premier vecteur correspondant à la classe 1 et le deuxième correspondant à la classe 2 de la variable qualitative. Les données soumises peuvent aussi correspondre à deux vecteurs séparés par le symbole "~" qui signifie *expliqué par* : le premier contenant les valeurs de variable quantitative et le deuxième les valeurs de la variable qualitative de chaque individu de l'échantillon.

Utiliser les arguments `paired = FALSE` et `var.equal = TRUE`.

Attention : en unilatéral, l'ordre des deux vecteurs à soumettre est important. Pour que le test soit conforme aux hypothèses H_1 unilatérales droite et gauche telles qu'elles sont formulées dans "Hypothèses de test" du paragraphe 5.13.2.1 précédent, le premier vecteur doit correspondre à m_1 et le deuxième à m_2.

Test bilatéral

Monsieur Léon voudrait savoir si les moyennes de hauteur des pieds de maïs sont similaires entre les parcelles Nord et Sud.

Hypothèses de test :
H_0 : Dans la récolte, les moyennes de hauteur sont identiques entre les parcelles Nord et Sud.
H_1 : Dans la récolte, elles sont différentes.

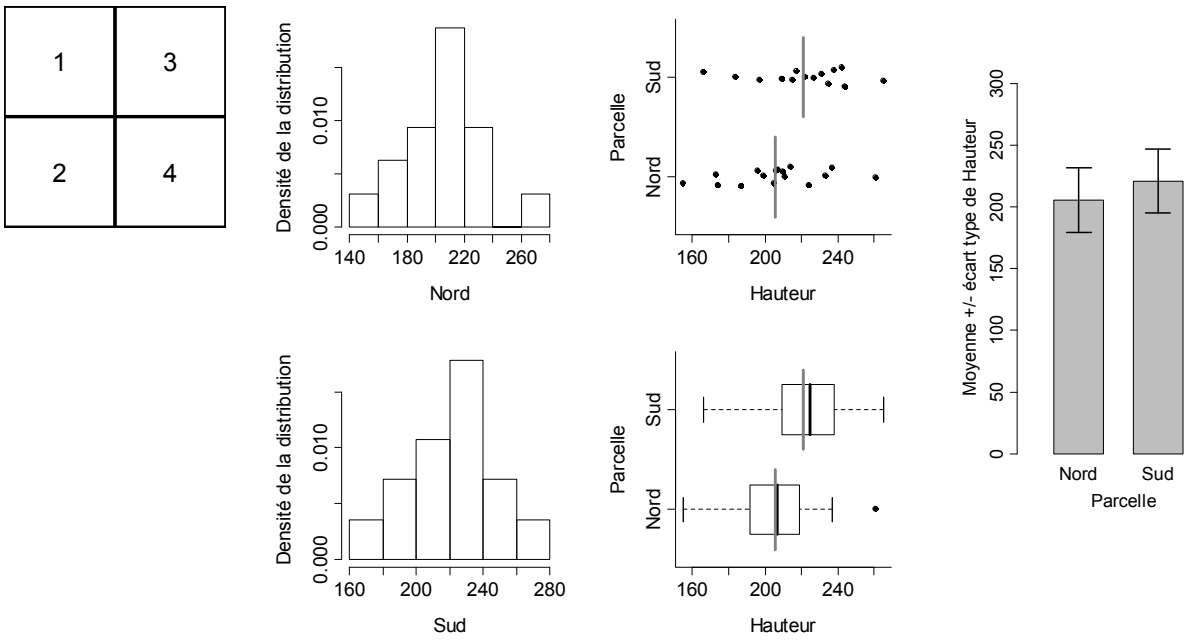
Voir le test de Welch (paragraphe 5.13.3.2) pour l'obtention des tableaux et des représentations graphiques. La seule différence concerne le début du code :

```
> mais<-read.table("C:/Users/Gael/Desktop/mais.txt",header=TRUE) # pour importer
      votre fichier de données (voir le paragraphe 1.6.1.2)
> obs1<-mais[which(mais$Parcelle=="Nord" | mais$Parcelle=="Sud"), c(2,11)] # on
      sélectionne les individus quand une ligne indique "Nord" ou "Sud" dans la
      colonne parcelle. Mettre la variable quantitative dans la 1ère colonne
```

Tableau obs3 obtenu :

```
Nombre de lignes avec NA : 1

> obs3
   Nord Sud
1   199 215
2   205 242
3   173 197
...
14  210 209
15  187  NA
16  211  NA
```

Représentations graphiques obtenues :



Les histogrammes semblent proches d'une distribution normale. Le graphique 3.4 révèle un individu éloigné mais il ne l'est pas suffisamment pour le considérer comme extrême (graphique 3.3). Dans la fenêtre 4, les moyennes semblent similaires.

Réalisation du test :

```
> t.test(obs3[,1], obs3[,2], var.equal=TRUE)# peut également s'écrire :
      t.test(obs1[,1]~obs1[,2], var.equal=TRUE)

        Two Sample t-test

data:  obs3[, 1] and obs3[, 2]
t = -1.5821, df = 28, p-value = 0.1249
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -34.66703   4.45274
```

```
sample estimates:
mean of x mean of y
 205.7500  220.8571
```

Conclusion :
L'hypothèse H_0 ne peut être rejetée au seuil $\alpha = 5\%$. D'après l'échantillon, rien ne permet de réfuter que les moyennes de hauteur des pieds de la parcelle Nord et Sud sont identiques dans la récolte.

Le calcul des paramètres de notre échantillon est identique à celui du test de Welch (paragraphe 5.13.3.2). Résultats obtenus :

```
> param
     Effectif  Moyenne Ecart.type
Nord       16 205.7500   26.29195
Sud        14 220.8571   25.86015
```

Valeur de la VT t sur notre échantillon :

```
> s2<-((n1-1)*s.1^2+(n2-1)*s.2^2)/(n1+n2-2)
> t.calc<-(m1-m2)/(s2*(1/n1+1/n2))^0.5
> t.calc
     Nord
-1.582094
```

La VT t suit une loi de Student à $\nu = n_1 + n_2 - 2$ ddl et le test est bilatéral. La *p value* correspond donc à la plus petite des deux aires sous la courbe, à droite et à gauche de la valeur t calculée, qu'il faut multiplier par deux. Obtention de la *p value* :

```
> nu<-n1+n2-2
> nu # valeur de ddl
Nord
  28
> min(pt(t.calc, nu, lower.tail=FALSE), pt(t.calc, nu))*2
[1] 0.1248596
```

INDEX

Les attributs de R et les termes anglo-saxons sont en italique.

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z

Gaël Millot

Comprendre et réaliser les tests statistiques à l'aide de R

Manuel de biostatistique

Ce livre s'adresse aux étudiants, médecins et chercheurs désirant réaliser des tests alors qu'ils débutent en statistique.

Une approche simple et détaillée

Illustré par 76 figures et accompagné d'exercices avec correction, l'ouvrage aborde la statistique de la manière la plus simple qui soit, sans démonstration mathématique, mais en insistant sur les détails, afin de bien maîtriser toutes les subtilités des tests.

Des notions essentielles traitées en profondeur

L'ouvrage explore des points fondamentaux en statistique : la check-list à effectuer avant de réaliser un test, la gestion des individus extrêmes, l'origine de la p value, la puissance ou la conclusion d'un test. Il explique comment choisir un test à partir de ses propres données. Il décrit 35 tests statistiques sous forme de fiches, dont 24 non paramétriques, ce qui couvre la plupart des tests à une ou deux variables observées. Il traite de toutes les subtilités des tests, comme les corrections de continuité, les corrections de Welch pour le test t et l'anova, ou les corrections de p value lors des comparaisons multiples. Il propose un exemple d'application de chaque test à l'aide de R, en incluant toutes les étapes du test, et notamment l'analyse graphique des données.

R, le logiciel de référence

L'originalité de ce manuel est de proposer non seulement une explication très détaillée sur l'utilisation des tests les plus classiques, mais aussi la possibilité de réaliser ces tests à l'aide de R, logiciel de référence en statistique, gratuit, disponible sur Internet et compatible avec Windows, Mac OS et Linux. Ce livre parlera également à ceux qui ne souhaitent pas utiliser R, car tous les exemples d'application des tests sont réalisés à partir d'un même fichier de données, qui peut facilement être adapté à un autre logiciel de statistique.

- La 3e édition de la référence dans le domaine des tests statistiques et de R
- Version 3 de R
- Accessible aux débutants : aucun pré-requis nécessaire en mathématique ou en informatique
- Nombreux exemples d'application et exercices corrigés

Gaël MILLOT, Docteur en Génétique Humaine, Maître de Conférences en Génétique et Biostatistique à l'Université Pierre et Marie Curie (Paris VI) / Institut Curie.

L'auteur reverse la moitié de ses droits d'auteur à différents organismes de lutte contre le cancer

Conception graphique : Primo&Primo®
illu : ©Tiero

ISBN : 978-2-8041-6365-5

9 782804 163655

RETESTA

de boeck

www.deboeck.com