

Gaël Millot

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

Manuel de biostatistique

3^e édition



 de boeck

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

3^{ème} 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".

Microsoft Windows 7, Microsoft Excel, Microsoft Word, Mac OS X, Bloc-notes et TextEdit sont des marques commerciales déposées. R, Linux, Tinn-R et ESS sont librement disponibles suivant les termes de la *GNU General Public License*.

SOMMAIRE

ABREVIATIONS ET SYMBOLES	14
CHAPITRE 1 : PRESENTATION DE R	16
1.1. Introduction.....	16
1.2. Comment obtenir et installer R.....	16
1.2.1. Installation de R sous Windows et Mac OS.....	16
1.2.2. Installation de <i>packages</i>	17
1.2.3. Récupérer des manuels d'aide.....	18
1.3. Découverte	18
1.4. Description des principaux attributs de R	20
1.4.1. L'instruction	20
1.4.2. Les objets	21
1.4.2.1. Les objets de données	21
1.4.2.2. Les fonctions.....	22
1.4.3. Les opérateurs	23
1.4.4. Les attributs spéciaux.....	24
1.5. Premiers pas : R est une calculatrice.....	25
1.6. Manipulation des objets de données.....	25
1.6.1. Création d'objets de données.....	25
1.6.1.1 Création par écriture	26
1.6.1.2. Création par importation de fichiers texte	31
1.6.1.3. Création par utilisation du tableur de R	35
1.6.2. Description d'un objet de données	36
1.6.2.1. La fonction <code>length()</code>	36
1.6.2.2. La fonction <code>mode()</code>	37
1.6.2.3. La fonction <code>class()</code>	38
1.6.2.4. La fonction <code>summary()</code>	38
1.6.3. Analyse et modification des données dans un objet de données.....	40
1.7. Fonctions de statistique descriptive	54
1.8. Manipuler les instructions conditionnelles ou répétées en boucles	57
1.8.1. Attributs de condition	57
1.8.2. Attributs de répétition	59
1.9. R est un éditeur de graphiques.....	61
1.9.1. Découverte de la fenêtre graphique	62
1.9.2. Différents types de graphique	62
1.9.2.1. Description des fonctions classiques	62
1.9.2.2. Arguments communs de ces fonctions	66
1.9.3. Ajout d'éléments sur un graphique.....	67
1.9.4. Paramètres graphiques	68
1.9.5. Manipulation des polices de caractères.....	71
1.9.6. Manipulation des couleurs	73
1.9.7. Tracer plusieurs graphiques côte à côte dans une même fenêtre	81
1.9.8. Gérer plusieurs fenêtres graphiques	82
1.9.9. Exporter un graphique.....	83

CHAPITRE 2 : NOTIONS DE BASE DE LA STATISTIQUE	86
2.1. Introduction.....	86
2.2. Définitions de la statistique, de la population et de l'échantillon	87
2.3. L'individu.....	89
2.4. Les variables aléatoires.....	90
2.4.1. Définition	90
2.4.2. Deux types de variables aléatoires	94
2.4.2.1. Variable quantitative.....	94
2.4.2.2. Variable qualitative.....	94
2.4.3. Particularités de certaines variables qualitatives	95
2.4.3.1. Exclusivité des classes	95
2.4.3.2. Classes appariées	96
2.4.3.3. Cas des variables fixées	97
2.4.4. Observer la distribution des valeurs d'une variable quantitative : l'histogramme	99
2.4.4.1. Choix du nombre de classes	99
2.4.4.2. Fixer l'intervalle des classes.....	99
2.4.4.3. Ordonnée en effectif, fréquence ou densité	100
2.4.4.4. La fonction <code>hist()</code> de R.....	101
2.4.5. Observer la distribution des valeurs d'une variable qualitative	103
2.4.6. Limite entre l'aspect quantitatif et qualitatif d'une variable	104
2.5. Les différents types de tableaux de données	107
2.5.1. Cas standard.....	107
2.5.2. Le tableau disjonctif complet pour les variables qualitatives.....	107
2.5.3. Le tableau de contingence pour une ou deux variables qualitatives	108
2.6. Avant d'entreprendre toute analyse statistique : la check-list	109
2.7. Les paramètres de statistique descriptive les plus employés	110
2.7.1. La moyenne et la médiane	110
2.7.2. Les quantiles	111
2.7.3. La variance, l'écart type et le coefficient de variation.....	112
2.7.4. La covariance.....	113
2.7.5. Le coefficient de corrélation linéaire de Pearson	115
2.8. Exercices	117
2.9. Corrections des exercices.....	118
CHAPITRE 3 : DEMARCHE SCIENTIFIQUE ET ERREURS ASSOCIEES	124
3.1. Formulation de la question scientifique	124
3.1.1. Décalage entre la question posée et l'approche envisagée	124
3.1.2. Faits supposés avérés	125
3.2. Organisation de l'étude scientifique	126
3.2.1. Individus non semblables.....	126
3.2.2. Conditions environnementales non semblables	127

3.3. Interprétation du résultat.....	128
3.4. La gestion des individus extrêmes (outliers en anglais).....	128
3.5. Conclusion	130
CHAPITRE 4 : LES ETAPES D'UN TEST STATISTIQUE	131
4.1. Introduction à lire avant de se lancer dans ce chapitre	131
4.2. Les deux hypothèses statistiques.....	133
4.3. La Variable de Test (VT).....	135
4.3.1. Définition.....	135
4.3.2. Différents types de VT.....	135
4.3.2.1. Tests paramétriques et non paramétriques.....	136
4.3.2.2. Estimateur et VT.....	137
4.3.2.3. VT et distribution de probabilité.....	137
4.4. Distributions de probabilité	138
4.4.1. Définition d'une loi de probabilité	138
4.4.2. Paramètres d'une loi de probabilité	138
4.4.2.1. Cas des variables discrètes : quantile, probabilité et fonction de répartition	138
4.4.2.2. Cas des variables continues : quantile, densité de probabilité et fonction de répartition.....	140
4.4.2.3. Calcul de la probabilité de voir apparaître une valeur de variable continue	142
4.4.3. Comment utiliser les lois de probabilité avec R.....	144
4.4.4. Différentes lois de probabilité discrètes.....	146
4.4.4.1. Loi binomiale.....	146
4.4.4.2. Loi multinomiale	151
4.4.4.3. Loi de Pascal et loi binomiale négative	153
4.4.4.4. Loi géométrique.....	156
4.4.4.5. Loi hypergéométrique.....	158
4.4.4.6. Loi de Poisson	160
4.4.5. Différentes lois de probabilité continues.....	162
4.4.5.1. Loi normale ou de Laplace-Gauss	162
4.4.5.2. Loi normale centrée réduite	166
4.4.5.3. Loi exponentielle	169
4.4.5.4. Loi gamma.....	170
4.4.5.5. Loi de χ^2	173
4.4.5.6. Loi de Fisher-Snedecor.....	175
4.4.5.7. Loi de Student.....	178
4.4.6. Distributions de probabilité qui ne suivent pas de loi connue.....	180
4.4.6.1. Distribution de probabilité de Mann-Whitney	180
4.4.6.2. Distribution de probabilité de Wilcoxon.....	185
4.4.6.3. Distribution de probabilité du test des signes de Wilcoxon.....	190
4.4.7. Rapport entre toutes ces distributions de probabilité	195
4.4.8. Remarques importantes.....	196
4.4.8.1. Ne pas confondre la loi de probabilité d'une variable mesurée et celle d'une VT.....	196
4.4.8.2. Simulation avec R de la fluctuation d'une VT due à l'échantillonnage.....	197
4.4.8.3. Importance du tirage aléatoire des individus dans la formation de l'échantillon	199
4.5. Hypothèse H_0, distribution de probabilité de la VT et échantillon : le cocktail magique de l'obtention de la p value.....	200
4.6. Conclusion d'un test statistique et les deux risques d'erreurs associés.....	203
4.6.1. Conclure, c'est deux vérités, deux décisions soit quatre probabilités.....	203
4.6.2. L'hypothèse H_0 et le risque α : définitions.....	205
4.6.3. La correction du seuil de rejet α	207
4.6.3.1. Le problème soulevé.....	207
4.6.3.2. La technique de Bonferroni	209
4.6.3.3. La technique séquentielle (Holm).....	210
4.6.3.4. Quand appliquer la correction ?.....	211

4.6.4. L'hypothèse H_1 et son influence sur le risque α	214
4.6.4.1 Le problème de l'hypothèse H_1	214
4.6.4.2. Test bilatéral et unilatéral	215
4.6.4.3. Obtenir la p value en test bilatéral et unilatéral	218
4.6.4.4. Placer les seuils α de rejet en test bilatéral et unilatéral	221
4.6.4.5. Comment choisir entre test bilatéral et unilatéral ?	226
4.6.5. Le risque β et la puissance $1-\beta$ du test	227
4.6.5.1. Retour sur les définitions de β et $1-\beta$	228
4.6.5.2. Variations de β et $1-\beta$ suivant la distribution de probabilité de la VT sous H_1	228
4.6.6. α et β en termes de faux positifs et faux négatifs	234
4.6.7. Propriétés de la puissance $1-\beta$	236
4.6.7.1 A lire avant de se lancer dans ce paragraphe	236
4.6.7.2. La puissance d'un test diminue quand décroît le α_{seuil}	236
4.6.7.3. La puissance d'un test croît quand augmente l'effectif n de l'échantillon	238
4.6.7.4. La puissance d'un test augmente avec l'écart entre les paramètres testés	245
4.6.8. Le danger de considérer la p value comme un indicateur de forte ou faible significativité	246
4.6.9. Alors comment fixer la puissance d'un test ?	248
4.6.9.1. Considérations générales	248
4.6.9.2. Réaliser des abaques	250
4.6.9.3. Les fonctions disponibles sous R	252
4.6.9.4. Le ncp des lois de probabilité de VT sous R	255
4.6.10. Comment conclure finalement ?	258
4.7. Récapitulation	260
4.8. Exercices	261
4.9. Correction des exercices	264
CHAPITRE 5 : LES TESTS STATISTIQUES	280
5.1. A lire absolument avant d'utiliser un test	280
5.2. Quel test appliquer et quelle fonction de R utiliser ?	289
<i>Comparaison d'effectifs et de proportions</i>	
5.3. χ^2 de conformité	293
5.3.1. Méthode	293
5.3.2. Exemple avec R	297
5.3.3. Tests de comparaisons deux à deux	302
5.4. χ^2 d'homogénéité	305
5.4.1. Méthode	305
5.4.2. Exemples avec R	311
5.4.3. Tests de comparaisons deux à deux	318
5.5. Test G	321
5.5.1. Méthode	321
5.5.2. Exemples avec R	324
5.5.3. Tests de comparaisons deux à deux	325
5.6. Test exact de Fisher	328
5.6.1. Tableau de contingence 2×2	328
5.6.1.1. Méthode	328
5.6.1.2. Exemples avec R	334
5.6.2. Tableau de contingence $c \times k$	344
5.6.2.1. Méthode	344
5.6.2.2. Exemple avec R	345
5.6.2.3. Tests de comparaisons deux à deux	347

5.7. Test de Mantel-Haenszel.....	349
5.7.1. Méthode.....	349
5.7.2. Exemples avec R.....	355
5.7.3. Tests de comparaisons deux à deux.....	362
5.8. Comparaison d'une proportion observée à une proportion théorique.....	364
5.8.1. Méthode.....	364
5.8.2. Exemples avec R.....	367
5.9. Comparaison de deux proportions observées.....	377
5.9.1. Méthode.....	377
5.9.2. Exemples avec R.....	381
5.10. Comparaison de deux proportions en séries appariées (test de Mac Nemar).....	386
5.10.1. Méthode.....	386
5.10.2. Exemples avec R.....	391
5.11. Comparaison de plusieurs proportions observées.....	397
5.11.1. Méthode.....	397
5.11.2. Exemple avec R.....	400
5.11.3. Tests de comparaisons deux à deux.....	404
5.12. Comparaison de plusieurs proportions observées à plusieurs proportions théoriques.....	408
5.12.1. Méthode.....	408
5.12.2. Exemple avec R.....	411
5.12.3. Tests de comparaisons deux à deux.....	415

Comparaison de moyennes

5.13. Le test t de Student de comparaison de moyennes.....	417
5.13.1. Comparaison d'une moyenne observée à une valeur théorique.....	417
5.13.1.1. Méthode.....	417
5.13.1.2. Exemples avec R.....	419
5.13.2. Comparaison de deux moyennes observées.....	425
5.13.2.1. Méthode.....	425
5.13.2.2. Exemple avec R.....	428
5.13.3. Comparaison de deux moyennes observées avec variances différentes (test de Welch).....	431
5.13.3.1. Méthode.....	431
5.13.3.2. Exemples avec R.....	432
5.13.4. Comparaison de deux moyennes observées en séries appariées.....	438
5.13.4.1. Méthode.....	438
5.13.4.2. Exemple avec R.....	441
5.14. Comparaison d'au moins deux moyennes observées.....	446
5.14.1. Anova (analyse de variances à un facteur).....	446
5.14.1.1. Méthode.....	446
5.14.1.2. Exemple avec R.....	451
5.14.2. Anova avec variances différentes (correction de Welch).....	454
5.14.2.1. Méthode.....	454
5.14.2.2. Exemple avec R.....	455
5.14.3. Tests de comparaisons deux à deux.....	458

Comparaison de médianes

5.15. Comparaison d'une médiane observée à une valeur théorique (test des signes de Wilcoxon).....	462
5.15.1. Méthode.....	462
5.15.2. Exemples avec R.....	468
5.16. Comparaison de deux médianes observées (test de Mann-Whitney-Wilcoxon).....	476
5.16.1. Méthode.....	476
5.16.2. Exemples avec R.....	485

5.17. Comparaison de deux médianes observées en séries appariées (test des signes de Wilcoxon)	494
5.17.1. Méthode	494
5.17.2. Exemples avec R.....	503
5.18. Comparaison d'au moins deux médianes observées.....	511
5.18.1. Test de Kruskal-Wallis	511
5.18.1.1. Méthode	511
5.18.1.2. Exemple avec R	515
5.18.1.3. Tests de comparaisons deux à deux	520
5.18.2. Test des médianes	523
5.18.2.1. Méthode	523
5.18.2.2. Exemple avec R	527
5.18.2.3. Tests de comparaisons deux à deux	530

Comparaison de variances

5.19. Comparaison de deux variances observées	533
5.19.1. Test de Fisher-Snedecor.....	533
5.19.1.1. Méthode	533
5.19.1.2. Exemple avec R	536
5.19.2. Test d'Ansari-Bradley	540
5.19.2.1. Méthode	540
5.19.2.2. Exemples avec R	547
5.20. Comparaison d'au moins deux variances observées	557
5.20.1. Test de Bartlett.....	557
5.20.1.1. Méthode	557
5.20.1.2. Exemple avec R	560
5.20.2. Test de Fligner- Killeen	564
5.20.2.1. Méthode	564
5.20.2.2. Exemple avec R	568
5.20.3. Tests de comparaisons deux à deux	573

Corrélations entre variables

5.21. Test du coefficient de corrélation linéaire de Pearson	575
5.21.1. Méthode	575
5.21.2. Exemple avec R	581
5.22. Test du coefficient de corrélation de Spearman	586
5.22.1. Méthode	586
5.22.2. Exemples avec R.....	593
5.23. Test du coefficient de corrélation de Kendall	600
5.23.1. Méthode	600
5.23.2. Exemples avec R.....	605
5.24. Test de χ^2	610
5.24.1. Méthode	610
5.24.2. Exemple avec R	611
5.25. Tests de corrélations multiples.....	612

Comparaison de distributions

5.26. Ajustement d'une distribution observée à une distribution théorique	615
5.26.1. Introduction.....	615
5.26.2. Test de χ^2 de conformité	616
5.26.2.1. Méthode	616
5.26.2.2. Exemple avec R	619
5.26.3. Test de Kolmogorov-Smirnov	624
5.26.3.1. Méthode	624

5.26.3.2. Exemple avec R	630
5.26.4. Test de Shapiro-Wilk	637
5.26.4.1. Méthode	637
5.26.4.2. Exemple avec R	642
5.27. Comparaison de deux distributions observées (test de Kolmogorov-Smirnov).....	646
5.27.1. Méthode	646
5.27.2. Exemple avec R	651
<i>Autres tests</i>	
5.28. Tests autour de la régression.....	657
5.28.1. Introduction.....	657
5.28.2. Principe de la régression linéaire simple.....	658
5.28.3. Comparaison d'une régression observée à une régression nulle.....	661
5.28.3.1. Méthode	661
5.28.3.2. Exemple avec R	670
5.28.4. Comparaison d'une régression observée à une régression théorique.....	680
5.28.4.1. Méthode	680
5.28.4.2. Exemples avec R	683
5.29. Test autour de la survie	688
5.29.1. Introduction.....	688
5.29.2. Comparaison de deux courbes de survie (test du logrank).....	694
5.29.2.1. Méthode	694
5.29.2.2. Exemple avec R	702
ANNEXES	713
1. Formule développée de la variance et de la covariance	713
2. L'estimateur	713
3. Distribution normale de variables mesurées et théorème central limite.....	719
4. Rappel des moyennes et variances des distributions de probabilité.....	720
5. Rappel sur les combinaisons.....	721
6. Passage du χ^2 au Z^2 dans le cas de la comparaison d'une proportion observée à une proportion théorique	722
7. Passage du χ^2 au Z^2 dans le cas de la comparaison de deux proportions observées	723
8. Retrouver la formule de la VT à partir de la formule du χ^2 dans le cas de la comparaison de plusieurs proportions observées	726
9. Estimation de la fluctuation de la VT χ^2 avec correction de continuité de Yates	727
10. Comment se comportent les différents couples de proportions ($p_{G1/F1}$, $p_{G1/F2}$), ($p_{G2/F1}$, $p_{G2/F2}$), ($p_{F1/G1}$, $p_{F1/G2}$) et ($p_{F2/G1}$, $p_{F2/G2}$) lors d'un test exact de Fisher sur tableau de contingence 2×2	729
11. Anova et régression linéaire sont liées.....	731
12. Procédure lorsque la fonction <code>solve()</code> n'est pas utilisable.....	736
13. Exemples graphiques avec R	736
14. Exécution des codes du chapitre 5 depuis un fichier de type "texte" et exportation des résultats.....	741
15. Edition des graphiques du chapitre 5 dans un fichier de type "pdf"	743
16. Différences entre les fonctions <code>sort()</code> , <code>rank()</code> et <code>order()</code>	743
17. Précisions sur la médiane et autres quantiles	744
18. Création de fonctions	746
19. Principaux attributs de R.....	771
REFERENCES.....	792
REMERCIEMENTS	794
INDEX.....	795

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 $m_{\text{théo}}$.

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

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

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

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

Loi de probabilité suivie par la VT sous H_0

La VT t suit une loi de Student à $v = n - 1$ 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 :

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

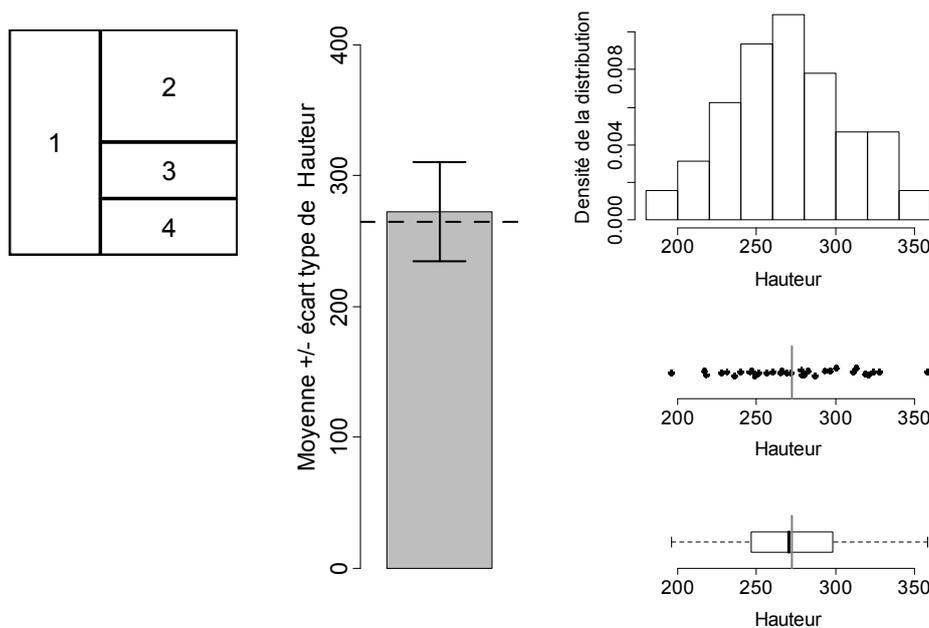
Obtention des graphiques (voir le paragraphe 1.9 pour plus d'informations) :

```
> 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 $v = n - 1$ 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 $v = n - 1$ 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 où l'on ne connaît 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 $N(m_{\text{théo}}, s/\sqrt{n})$ 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 $v = n - 1$ 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.

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

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

$H_{1 \text{ unilat g}}$: $\mu_1 < \mu_2$ 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 $v = n_1 + n_2 - 2$ 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	Individu	Classe 2 de la variable qualitative
1	x_{11}	1	x_{12}
...
j	x_{j1}	g	x_{g2}
...
n_1	x_{n11}	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 :

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

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

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 :

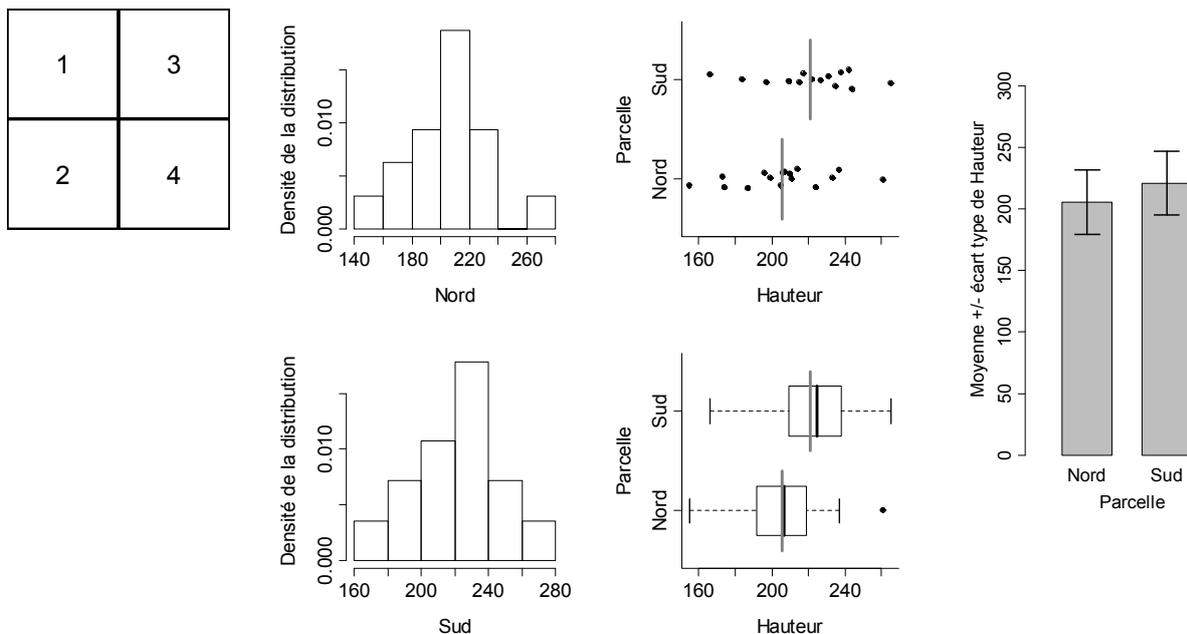
```
> mais<-read.table("C:/Users/Gael/Desktop/mais.txt",header=TRUE) # pour importer
votre fichier de données (voir le paragraphe 1.6.1.2)
> obs1<-mais[which(mais$Parcelle=="Nord" | mais$Parcelle=="Sud"), c(2,11)] # on
sélectionne les individus quand une ligne indique "Nord" ou "Sud" dans la
colonne 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 $v = n_1 + n_2 - 2$ 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.

#

-, 774
!, 773
!=, 773
#, 773
 α voir alpha
.Rdata, 16, 18
 β voir bêta
 γ voir gamma
1- β voir puissance
 Δ voir Delta
.GlobalEnv, 755
.Primitive(), 765
.C(), 766
.Call(), 766
.Fortran(), 766
.External(), 766
.Internal(), 770
.GlobalEnv, 772
?, 773
??, 773
<, 773
<=, 773
>, 773
>=, 773
==, 773
&, 773
|, 773
:, 774
e+, 774
+, 774
***, 774
/, 774
^, 774
****, 774
% / %, 774
%%, 774
^, 774
%%*, 775
<-, 775
->, 775
=, 775
<<-, 776
.First, 776
.Last, 776
[], 780
[[]], 780
\$, 780
%in%, 784
~, 791

χ^2

loi de, 136, 173, 293, 306, 321, 397, 408, 511, 523, 557, 564, 694, 789
test du, 290, 293, 305, 616, 650

A

abaque de puissance, 250, 263
abline(), 68, 118, 785
abline(lm()), 672, 791
abs(), 774
addmargins(), 108, 782
aggregate(), 57, 782, 783
agrep(), 784
aide, 22
aire sous la courbe, 141, 168
ajustement, 198, 263, 288, 615
aléatoire
 choix, 88
 échantillon, 88
 variable, 90
all(), 340, 773
alpha
 risque, 203, 205
 seuil, 205, 221, 236, 249
analyse
 de variances, 446, 731
 globale, 212, 282, 283, 284
 intermédiaire, 213
 multiple voir comparaisons deux à deux
Anova, 446, 731
anova(lm()), 290, 451, 734, 790
ansari.test(), 290, 547, 790
ansari_test(), 548
Ansari-Bradley, test d', 540
antislash, 33
any(), 311, 469, 505, 773
appariement, 97
append(), 781
apply(), 781, 783
approche
 descriptive, 124
 expérimentale, 124
 intermédiaire, 128
 par enquête, 124
approx(), 786
approximation d'une loi, 195, 284
apropos(), 771
arborescence, 755
args(), 773
arguments, 22
 d'une fonction, 746
 évaluation des, 750
 graphiques, 66
 supplied, 776

array, 27
array(), 26, 776
arrows(), 68, 739, 785
as.array(), 778
as.character(), 777
as.data.frame(), 778
as.double(), 778, 780
as.environment(), 756, 772
as.expression(), 673
as.factor(), 778
as.list(), 778
as.logical(), 778
as.matrix(), 778
as.numeric(), 778
as.table(), 778
as.vector(), 778
assign(), 60, 760, 764, 775, 776
assignation, 23, 775
astuce, 60, 336, 737, 743
attach(), 759
attributs, 20

- d'aide, 773
- de calcul, 774
- de comparaison, 773
- de condition, 57, 788
- de R, 20
- de répétition, 59, 788
- logiques, 773
- spéciaux, 24, 765

axis(), 68, 785
axTicks(), 785

B

barplot(), 63, 738, 785
barplot2(), 64, 738, 785
barres

- d'écart type, 243, 717, 739
- d'erreurs, 243, 717
- graphique en, 64

Bartlett, test de, 557
bartlett.test(), 290, 560, 790
baseenv(), 755, 772
béta, risque, 203, 228
biais, 199
bilatéral(e)

- hypothèse, 216
- test, 134, 216, 219, 225, 226, 232

binaire, variable, 95
binom.test(), 290, 367, 790
binomiale négative, loi, 153, 789
binomiale, loi, 136, 146, 364, 387, 789
binormalité, 578, 677
bit(s), 74
bitops, 64
bmp(), 83, 788
body(), 766, 776
boîtes, graphique en, 65
Bonferroni, correction de, 209, 407
boucle, 59
boxplot(), 65, 785
break, 61

Breslow-Day, test de, 354
brillance de la couleur, 78
bty, 786
bxp(), 65
by(), 56, 782, 783

C

c(), 26, 49, 775
camembert, 785
caractère, 90
cat(), 741, 784
caTools, 64
cbind(), 43, 48, 782
cbind.data.frame(), 48, 782
ceiling(), 774
censure, 688
centre

- de gravité, 110
- d'inertie, 110

cex, 786
cex.axis, 786
cex.lab, 786
cex.main, 786
chaînes de caractères, 783
chartr(), 783
check-list, 109
chemin absolu, 33
chi2 voir χ^2 en début d'index
chiffre à double précision, 775
chisq.test(), 290, 297, 311, 790
choix

- aléatoire, 88
- d'un test, 292

choose(), 722, 775
citation(), 773
class(), 38, 777, 779
classe(s)

- appariées, 96
- d'un histogramme, 99
- d'un objet, 38
- d'une variable qualitative, 94
- exclusives, 95
- indépendance des, 305
- liaison partielle entre deux, 96

clôture, 747, 758
fonction, 765
cm.colors(), 79, 787
Cochran, règle de, 293, 297, 300
code, 20
code R

- accès, 767
- écriture du, 18
- zone grisée, 281

codispersion, 110
coef(lm()), 791
coefficient

- d'association, 610
- de corrélation
 - de Kendall, 603
 - de Spearman, 588

- linéaire de Pearson, 115, 575
- test d'un, 290, 575, 586, 600
- de détermination, 640, 663, 682
- de variation, 112
- multinomial, 152
- coin*, 548
- col*, 786
- col.axis*, 786
- col.lab*, 786
- col.main*, 786
- col2rgb()*, 76, 787
- colMeans()*, 789
- colnames()*, 45, 782
- colors()*, 73, 787
- colSums()*, 782
- combinaison, 721
- combinations()*, 301, 314, 341, 775
- combinatoire, 180
- comparaison(s)
 - attributs de, 773
 - de courbes de survie, 290, 694
 - de distributions, 290, 646
 - de médianes, 290, 462, 476, 494, 511, 523
 - de moyennes, 290, 417, 425, 438, 446, 454
 - de proportions, 290, 364, 377, 386, 397, 408
 - de quantiles, 290, 526
 - de variances, 290, 533, 540, 557, 564
 - deux à deux, 212, 283
 - multiples, 283
- compiler, 769
- complete.cases()*, 785
- conclusion d'un test, 204, 258, 281
- concordance, 441, 503, 602
- condition(s)
 - attributs de, 57, 788
 - d'application d'un test, 281, 287
 - environnementales, 127
- confirmatoire, valeur, 212
- conformité, test de, 288
- confusion, facteur de, 356
- conservateur, test, 615
- console, 18
- contingence, 108
- continuité, correction de, 285
- contraste(s)
 - technique des, 212
- convergence
 - d'un estimateur, 718
 - d'une loi, 195
- Cook, distance de, 669, 678
- cooks.distance(lm())*, 791
- cor()*, 54, 55, 789
- cor.test()*, 290, 581, 790
- corps d'une fonction, 746
- correction
 - de Bonferroni, 209, 407
 - de continuité, 285
 - de Holm, 210, 283, 405, 406
 - de la *p value*, 211
 - du seuil α , 209
- corrélation voir coefficient de corrélation
- cos()*, 774
- couleurs, gestion des, 73, 787
- courbe voir graphique(s)
 - de Kaplan-Meier, 689, 703
 - de survie, 290, 694
 - en cloche, 162
- cov()*, 789
- covariance, 113
- Cramer, coefficient d'association de, 610
- création
 - de fonctions, 746
 - d'objets de données, 25
 - d'opérateurs, 754
- cumprod()*, 775, 781
- cumsum()*, 775, 781
- curseur, 18
- curve()*, 785
- cut()*, 781

D

- daltonisme, 80
- data frame, 28
- data.frame()*, 26, 776
- data.matrix()*, 778
- dbinom()*, 145, 148, 789
- dchisq()*, 145, 174, 789
- décomposition du test, 282, 283
- Delta, 245
 - réel, 246
 - seuil, 246, 249, 259
- démarche
 - scientifique, 124
 - statistique, 124
- densité, 100
 - de probabilité, 141
- détermination, coefficient de, 640, 663, 682
- deutéranopie, 80
- dev.cur()*, 82, 788
- dev.list()*, 82, 788
- dev.off()*, 82, 788
- dev.set()*, 82, 788
- deviance(lm())*, 791
- device region*, 62
- dexp()*, 145, 169, 789
- df()*, 145, 176, 789
- dffits(lm())*, 791
- dfitts*, 669, 678
- dgamma()*, 145, 172, 789
- dgeom()*, 145, 157, 789
- dhyper()*, 145, 159, 789
- diag()*, 783
- dichromat*, 80
- dichromat()*, 80
- diff()*, 775
- différence significative, 204, 284
- dim()*, 782
- dimnames()*, 782
- dir()*, 777
- dir.create()*, 742, 777
- discordance, 602
- discrète, variable, 94

dispersion, 110
dissymétrique, distribution, 111, 283
distance de Cook, 669, 678
distribution(s)
 comparaison de, 290, 646
 de probabilité, 138, 139, 141, *voir aussi loi*
 ajustement d'une, 263, 615
 de Mann-Whitney, 180, 789
 de Wilcoxon, 185, 190, 790
 de valeurs mesurées, 54, 99
 dissymétrique, 111, 283
 normale à deux dimensions, 577
 symétrique, 65, 111, 462, 494
dmultinom(), 145, 152, 789
dmbinom(), 145, 154, 789
dnorm(), 145, 164, 789
do.call(), 770, 777
données manquantes, 32, 784
double aveugle, 127
double précision, 775
double(), 775
dpois(), 145, 161, 789
dput(), 742, 777
droite de régression linéaire, 659
dsignrank(), 145, 193, 790
dt(), 145, 179, 789
dump(), 742, 777
duplicated(), 779
dwilcox(), 145, 184, 188, 789

E

écart entre deux paramètres, 245
écart type, 112
échantillon, 87
 aléatoire, 88
 biaisé, 199
 effectif de l', 88, 213, 238, 249
 paramètre d', 714
 représentatif, 88
échantillonnage
 biaisé, 129, 200
 fluctuation d', 197
 unité d', 89
EDI, 19
edit(), 35, 777
éditeur, 18
effectif
 de l'échantillon, 88, 213, 238, 249
 d'une classe, 100, 108
 d'une population, 87
efficacité d'un test, 292
ellipse, 116, 613
emptyenv(), 755, 772
enquête, approche par, 124
ensemble de tests, 211
entité de mesure, 89
environment(), 755, 762, 772
environmentName(), 755, 772
environnement, 772
 de développement intégré, 19
 de travail, 18, 754

global, 755
 notion d', 754
 supérieur, 756
 vide, 755
épidémiologie, 128
équivariance, 667
erreur(s)
 barres d', 243, 717
 écologique, 128
 géographique, 128
 standard *voir SEM*
espace de travail, 18, 755
estimateur, 137, 713
étendue lexicale, 747, 758
étude de terrain, 124
évaluation des arguments, 750
exclusives
 classes, 95
 hypothèses, 133
exclu-vivant, 689, 700
exp(), 774
expand.grid(), 194, 775
exploratoire, valeur, 212
exponentielle, loi, 169, 789
exporter vers un fichier, 83, 742, 777
expression, 20
expression(), 738, 784

F

facteur
 de confusion, 356
 objet, 29
 variable, 90
factor(), 26, 776
factorial(), 722, 775
factorielle, 722
FALSE, 24, 774
faux positif, 234
fenêtre(s) graphique(s)
 active, 62, 82
 description d'une, 62
 diviser une, 81
 gestion des, 82, 788
fichier maïs.txt, 93
figure region, 62
find(), 757, 771
findMethods(), 769, 777
Fisher, test exact de, 328, 344
fisher.test(), 290, 334, 790
Fisher-Snedecor
 loi de, 136, 175, 446, 534, 662, 789
 test de, 290, 533
fitted(lm()), 791
fix(), 35, 777
fligner.test(), 290, 568, 790
Fligner-Killeen, test de, 564
floor(), 774
fluctuation
 d'échantillonnage, 197
 d'une variable mesurée, 196
 d'une VT, 196, 197

follow up, 688
fonction(s)
 accès au code, 767
 arguments d'une, 746
 clôture, 765
 corps d'une, 746
 création de, 746
 de masse, 139
 de répartition, 140, 141
 graphique, 62, 785
 graphique de bas niveau, 68
 graphique de haut niveau, 63
 imbriquées, 758, 762, 764
 implémentée, 746
 intégrée, 746
 objet, 22
 primitive, 765
 système, 767, 771
 trigonométrie, 774
font, 71, 786
font.axis, 786
font.lab, 786
font.main, 786
 fonte, 71
for(){}, 59, 788
foreign, 33
formals(), 766, 776
format(), 784
formatC(), 784
formula, 767
 Freedman-Diaconis, règle de, 100
 fréquence, 100, 109
ftable(), 776, 782
function(){}, 24, 577, 746, 776

G

G, test, 321
gamma
 loi, 170, 789
 risque, 227
gamma(), 171, 775
gdata, 64
 générique, fonction, 767
 géométrique, loi, 156, 789
get(), 60, 761, 789
getAnywhere(), 768
getMethod(), 769, 777
gl(), 26, 30, 776
glm(), 791
global environment, 755
globalenv(), 755, 772
gplots, 64
 graduation secondaire, 736
graphics.off(), 82, 788
graphique(s)
 3D, 785
 ajout d'éléments sur un, 67, 785
 barres, 64
 barres avec intervalles, 64
 boîtes, 65
 camembert, 785

création de, 62, 785
 de fonction mathématique, 785
 fenêtre, 62
 fonctions, 62, 785
 histogramme, 99
 mosaïque, 66
 nuages de points, 63, 65
 paramètres, 68, 786
 polygone de fréquences, 139
quantile-quantile plot, 633
gray(), 73, 787
grep(), 784
grepl(), 784
grey(), 73, 787
grid, 64
 grisée, 281
 groupes non comparables, 126
gsub(), 784
gtools, 64
 Guateverde, 132
 guillemets, 748

H

H_0 , 133, 200, 215
 H_1 , 133, 216
hat(), 791
heat.colors(), 44, 79, 787
help(), 22, 773
help.search(), 773
help.start(), 773
 hexadécimal, système, 76
high-level plotting commands, 63
hist(), 101, 785
 histogramme, 99
history(), 771
 Holm, correction de, 210, 283, 405, 406
 homogénéité, test d', 288
 homoscedasticité, 667
hsv(), 73, 787
 HSV, représentation, 78
hue, 73
 hypergéométrique, loi, 136, 158, 328, 789
hypothèse(s)
 acceptée, 203, 258
 alternative, 133
 bilatérale, 216
 composite, 134
 de test, 281
 H_0 , 133, 200, 215
 H_1 , 133, 216
 nulle, 133
 principale, 133
 rejetée, 203
 simple, 134
 unilatérale, 216
 droite, 217
 gauche, 217

I

I(), 28, 778

IDE, 19
identify(), 786
if(){else}, 57, 788
ifelse(), 57, 788
image(), 44, 785
implémentée, fonction, 746
importation de fichiers, 31, 741, 777
incidence instantanée, 701
indépendance
 des classes, 305
 des résidus, 667
 test d', 288
individu(s), 89
 extrêmes, 128
 non semblables, 126
inertie, 110
Inf, 24, 774
installation
 de packages, 17
 de R, 16
instruction, 20
 conditionnelle, 57, 788
 répétée en boucles, 57, 788
integrated development environment, 19
intégrée, fonction, 746
interaction(), 52, 774, 781
interpolation, 744
interquartile range, 65
intervalle
 de confiance, 287, 717
 de temps, 689
 d'une classe, 99
 mesure d'un, 104
is.array(), 778
is.character(), 777
is.data.frame(), 778
is.double(), 778
is.environment(), 755, 772
is.factor(), 778
is.finite(), 779
is.function(), 765, 778
is.infinite(), 779
is.list(), 778
is.logical(), 778
is.matrix(), 778
is.na(), 779, 785
is.null(), 773
is.numeric(), 778
is.primitive(), 778
is.Surv(), 778
is.table(), 778
is.vector(), 778
isGeneric(), 768, 778
isS4(), 768, 778

J

jackknife, 668
jpeg(), 83, 788

K

Kaplan-Meyer, courbe de, 689, 703
Kendall, coefficient de corrélation de, 603
KernSmooth, 64
Kolmogorov-Smirnov, test de, 624, 646
kruskal.test(), 290, 515, 790
Kruskal-Wallis, test de, 511
ks.test(), 290, 630, 790

L

Laplace-Gauss, loi de, 162
lapply(), 782, 783
las, 786
layout(), 81, 788
layout.show(), 81, 788
legend(), 68, 737, 785
length(), 36, 779
letters, 25, 784
LETTERS, 784
Levels, 30
levels(), 779, 780
leverage, 668, 678
lexical scope, 747
library(), 772
likelihood ratio, 322, 559
lillie.test(), 629
Lilliefors, test de, 629
linéaire
 modèle, 660, 790
 régression, 657, 731
 relation, 578
lines(), 68, 739, 785
list(), 26, 776
list.dirs(), 777
list.files(), 742, 777
liste, 29
lm(), 791
lm.influence(lm()), 671, 791
load(), 771
loadedNamespaces(), 755, 772
locator(), 251, 786
log(), 766, 774
log10(), 774
log2(), 774
loglm(), 290, 324, 790
logrank, test du, 694
loi(s) de probabilité, 138
 ajustement d'une, 198, 263, 615
 approximation d'une, 195, 284
 avec R, 144, 789
 binomiale, 136, 146, 364, 387, 789
 binomiale négative, 153, 789
 continue, 141, 162
 de χ^2 , 136, 173, 293, 306, 321, 397, 408, 511, 523, 557, 564, 694, 789
 de Fisher-Snedecor, 136, 175, 446, 534, 662, 789
 de Laplace-Gauss, 162
 de Pascal, 153

- de Poisson**, 160, 789
- de Student**, 136, 178, 417, 426, 431, 438, 576, 587, 680, 789
- discrète**, 139, 146
- d'une variable mesurée**, 196
- d'une VT**, 196
- exponentielle**, 169, 789
- gamma**, 170, 789
- géométrique**, 156, 789
- hypergéométrique**, 136, 158, 328, 789
- multinomiale**, 151, 789
- non connue**, 180
- normale**, 162, 789
- normale centrée réduite**, 166, 350, 364, 377, 462, 477, 495, 541, 601
- relation entre les**, 195
- low-level plotting commands**, 68
- ls()**, 24, 771
- ls.diag(lsfitt())**, 671, 791
- ls.str()**, 746, 771
- lsfitt()**, 791
- lty**, 786
- lwd**, 786

M

- Mac Nemar, test de**, 386
- main**, 66
- maïs**, 93
- Mann-Whitney**
 - distribution de probabilité de**, 180, 789
 - test de**, 476
- mantelhaen.test()**, 290, 355, 790
- Mantel-Haenszel, test de**, 349
- mar**, 786
- margin.table()**, 52, 782
- MASS**, 64, 324
- match()**, 779, 784
- match.arg()**, 751, 776, 784
- matrice(s)**, 27
 - diagonalisation d'une**, 783
 - inversion d'une**, 783
 - produit de**, 733
 - transposition d'une**, 783
- matrix()**, 26, 776
- max()**, 54, 789
- mcnemar.test()**, 290, 391, 790
- mean()**, 54, 789
- median()**, 54, 745, 789
- médiane(s)**, 54, 111, 744
 - comparaison de**, 290, 462, 476, 494, 511, 523
- menu**, 18
- merge()**, 783
- meta.MH()**, 355
- methods()**, 767, 773, 777
- mfcoll**, 81, 787
- mfrow**, 81, 786
- mgp**, 786
- min()**, 54, 789
- modalité**, 94
- mode d'un objet**, 37
- mode()**, 37, 779, 780

- modèle linéaire**, 660, 790
- modeltools**, 548
- modulo**, 774
- monotone, relation**, 591
- month.name**, 61
- mosaicplot()**, 65, 785
- mosaïque**, 66
- moustaches** voir *whiskers*
- moyenne(s)**, 54, 110
 - comparaison de**, 290, 417, 425, 438, 446, 454
- mtext()**, 68, 740, 785
- multinomiale, loi**, 151, 789
- mvtnorm**, 548

N

- NA**, 24, 32, 774, 784
- na.fail()**, 784
- na.omit()**, 785
- names()**, 780
- NaN**, 774
- nchar()**, 783
- ncol()**, 782
- ncp**, 255
- new.env()**, 757, 772
- nlevels()**, 779
- nominale, variable**, 95
- non central parameter**, 255
- normale centrée réduite, loi**, 166, 350, 364, 377, 462, 477, 495, 541, 601
- normale, loi**, 162, 789
- nortest**, 629
- nrow()**, 782
- nuage de points**, 63, 65, 110
- NULL**, 24
- numeric()**, 775

O

- objects()**, 771
- objet**, 21
 - nommer un**, 21
- objet fonction**, 22, 776
- objet(s) de données**, 21
 - création d'**, 25, 775
 - description d'**, 36, 777, 779
 - modification d'**, 40, 777, 780, 782
- odds ratio**, 334, 353, 354
- omd**, 71
- one-sided test**, 216
- one-tailed test**, 216
- oneway.test()**, 290, 455, 734, 790
- opérateur(s)**, 23, 765, 770
 - création d'**, 754
 - de super-assignation**, 748, 763, 776
- options()**, 748, 772
- order()**, 743, 780
- ordered()**, 51, 104, 780
- ordinaire, variable**, 95
- ordonnée à l'origine de la droite**, 658, 680
- outer()**, 776, 781
- outliers**, 128

P

- p value*, 200, 205, 218
 - calcul de la, 282
 - correction de la, 211
 - significative, 205
 - significativité de la, 246
- p.adjust()*, 303
- packages*
 - création de, 777
 - gestion des, 772
 - installation de, 17
- pairs()*, 785
- pairwise.prop.test()*, 404
- pairwise.t.test()*, 459
- pairwise.wilcox.test()*, 520
- palette()*, 73, 787
- par()*, 68, 786
- paramètre(s)
 - centraux, 482
 - de codispersion, 110
 - de dispersion, 110
 - de population, 245, 714
 - de position, 110
 - d'échantillon, 714
 - formel, 746, 750
 - graphiques, 68, 786
 - homogénéité entre, 284
- parent environment*, 756
- parent.env()*, 756, 772
- Pascal, loi de, 153
- paste()*, 784
- pbinom()*, 145, 149, 789
- pch*, 786
- pchisq()*, 145, 174, 789
- pdf()*, 83, 743, 788
- pdf.options()*, 788
- Pearson, coefficient de corrélation de, 115, 575
- penché de la droite, 658, 680
- permutation, 722
- permutations()*, 190, 519, 598, 655, 775
- persp()*, 577, 785
- pexp()*, 145, 169, 789
- pf()*, 145, 177, 789
- pgamma()*, 145, 172, 789
- pgeom()*, 145, 157, 789
- phyper()*, 145, 159, 342, 789
- pi*, 25, 774
- pie()*, 785
- plot region*, 62
- plot()*, 63, 785
- plot(lm())*, 672, 791
- plot(survfit())*, 703
- plotcorr()*, 613
- plotmath*, 784
- plt*, 737, 787
- pmax()*, 779
- pmin()*, 779
- pbinom()*, 145, 155, 789
- png()*, 83, 788
- pnorm()*, 145, 164, 789
- points()*, 66, 68, 209, 785
- Poisson, loi de, 160, 789
- polices, 71
- polygon()*, 68, 168, 740, 785
- polygone de fréquences, 139, 339
- population, 87
 - paramètre de, 245, 714
 - statistique, 87
 - visée, 87
- portée lexicale, 747, 758
- position, 110
- power.anova.test()*, 254
- power.prop.test()*, 253
- power.t.test()*, 252
- ppoints()*, 635
- ppois()*, 145, 161, 789
- pré-test, 214, 284
- pretty()*, 102
- primitive
 - fonction, 765
- print()*, 741, 749, 784
- probabilité, 139, 142
- prod()*, 775, 781
- produit matriciel, 733
- programmation, 776
- prompt, 18
- prop.table()*, 782
- prop.test()*, 290, 368, 790
- proportion(s), 100, 109
 - comparaison de, 290, 364, 377, 386, 397, 408
- protanopie, 80
- psignrank()*, 145, 193, 472, 790
- pt()*, 145, 179, 789
- puissance $1-\beta$, 228
 - abaque de, 250, 263
 - et test de χ^2 , 263
 - et test de Mann-Whitney-Wilcoxon, 263
 - et test t, 262
 - fonctions de R, 252
 - propriétés de la, 236
 - seuil de la, 248
- pwilcox()*, 145, 184, 189, 490, 789
- pwr*, 252

Q

- q()*, 771
- qbinom()*, 145, 149, 789
- qchisq()*, 145, 174, 789
- qexp()*, 145, 170, 789
- qf()*, 145, 177, 789
- qgamma()*, 145, 172, 789
- qgeom()*, 145, 789
- qhyper()*, 145, 789
- qbinom()*, 145, 789
- qnorm()*, 145, 165, 566, 789
- qpois()*, 145, 789
- qqline()*, 635, 786
- qqnorm()*, 635, 785
- qqplot*, 633
- qqplot()*, 635, 785

qsignrank(), 145, 790
qt(), 145, 179, 789
quantile(), 54, 744, 789
quantile(s), 54, 111, 139, 141, 744
 comparaison de, 290, 526
quantile-quantile plot, 633
quartile, 111
quartz(), 82, 788
quartz.options(), 83, 788
Quotes, 67, 784
qwilcox(), 145, 789

R

R

accès au code, 767
 astuce, 60, 336, 737, 743
 attributs de, 20
 découverte de, 18
 écriture du code, 18
 installation de, 16
R Commander, 19
 raccourcis clavier, 18
rainbow(), 79, 787
 randomisation, 126, 262
rang(s), 288
 test de, 182, 288, 480
range(), 779, 789
rank(), 743, 779
 rapport de vraisemblance, 322, 559
 rayure, 738
rbind(), 43, 48, 782
rbind.data.frame(), 48, 782
rbinom(), 145, 727, 789
rchisq(), 145, 789
Rcmdr, 19
read.table(), 34, 777
 recensement, 89
 réciproque, 774
rect(), 68, 71, 785
 région
 de la fenêtre graphique, 62
 de la figure, 62
 du tracé, 62
 régression, 657
 droite de, 659
 linéaire, 657, 731
 logistique, 658
 test d'une droite de, 290, 661, 680
 rejet, seuil de, 205, 221
 relation
 causale, 128, 286
 linéaire, 578
 monotone, 591
relevel(), 51, 780
rep(), 766, 780
repeat{}, 59, 788
 répertoire, 33, 755
 courant / de travail, 33
replace(), 781
require(), 772
resid(lm()), 791

résidu, 659, voir aussi valeur résiduelle
 résiduelle, variance, 448, 669
return(), 749, 776
rev(), 780
rexp(), 145, 789
rf(), 145, 789
rgamma(), 145, 789
rgb(), 73, 787
RGB, représentation, 75
rgb2hsv(), 79, 787
rgeom(), 145, 789
rhyper(), 145, 789
 risque(s)
 alpha, 203, 205
 béta, 203, 228
 d'acquérir le statut final, 692
 de 1^{ère} espèce, 203
 de 2^{ème} espèce, 203
 de 3^{ème} espèce, 227
 gamma, 227
 proportionnels, 701
 relatif, 702
 relatif instantané, 701
rle(), 779
rm(), 772
rmeta, 355
rmultinom(), 145, 272, 789
rbinom(), 145, 789
rnorm(), 145, 197, 789
round(), 766, 774
rowMeans(), 789
rownames(), 782
rowSums(), 782
rpois(), 145, 789
RR, 702
rsignrank(), 145, 790
rstandard(lm()), 791
rstudent(lm()), 791
RStudio, 19
rt(), 145, 789
rug(), 68, 251, 737, 785
runif(), 268, 775
rwilcox(), 145, 789

S

S3, 768
S4, 768
sample(), 126, 269, 775, 780
sapply(), 782, 783
 saturation de la couleur, 78
 sauvegarder des objets, 21
save(), 771
save.image(), 771
savehistory(), 771
scale(), 167, 789
scan(), 35, 777
scatterplot3d, 785
scatterplot3d(), 152, 785
SCE, 448, 662, 733
SCI, 448, 734
Scott, règle de, 100

SCR, 662
SCT, 448, 662, 734
sd(), 54, 789
sdv(), 736
search(), 755, 772
searchpaths(), 755, 772
segments(), 68, 737, 785
SEM, 716
semi-qualitative, variable, 105
semi-quantitative, variable, 95
seq(), 774
séquence, 181, 288
séries appariées, 96, 386, 438, 494
seuil
 1-β, 248
 alpha, 205, 221, 236, 249
 correction du, 209
 de rejet, 205, 221
 de signification, 205
 Delta, 246, 249, 259
shapiro.test(), 290, 642, 790
Shapiro-Wilk, test de, 637
show(), 768
showMethods(), 768, 777
sign(), 775, 779
signif(), 766, 775
significatif, seuil, 205
significative
 différence, 204, 284
 p value, 205
significativité de la p value, 246
sin(), 774
slash, 33
solve(), 736, 783
sondage, 89, 124
sort(), 743, 780
source(), 741, 777
Spearman, coefficient de corrélation de, 588
splines, 548
split(), 781
sqrt(), 774
standard error of the mean, 716
standardized residual value, 666
statistique(s), 87
 de test, 135
 décisionnelle, 89
 démarche, 124
 descriptive, 89, 789
 étapes d'un test, 131
 hypothèses, 133
 inférentielle, 89
 population, 87
stats4, 548
statut
 final, 688
 initial, 688
strate, 349, 700
stripchart(), 64, 785
strsplit(), 784
Student
 loi de, 136, 178, 417, 426, 431, 438, 576, 587, 680, 789
 test t de, 290, 417
 studentized residual value, 667
Sturges, règle de, 99
sub, 66
sub(), 783
subset(), 781
substitute(), 673
substr(), 783
substring(), 783
sum(), 765, 775, 781
summary(), 38, 54, 779, 789
summary(aov(lm())), 791
summary(lm()), 290, 670, 790, 791
summary(survfit()), 703
super-assignation, 748, 763, 776
suppressWarnings(), 319, 788
Surv(), 702
survdiff(Surv()), 290, 706, 790
survfit(Surv()), 702
survie, 688
survival, 548, 702, 703, 706
symétrique, distribution, 65, 111, 462, 494
Syntax, 24

T

t voir Student
t(), 44, 783
t.test(), 290, 419, 767, 790
table, 30
table(), 26, 776, 779
tableau
 de contingence, 30, 108
 de données standard, 107
 disjonctif complet, 107
tableur de R, 35, 777
tan(), 774
tapply(), 782
tcltk, 80
télécharger
 le fichier maïs.txt, 93
 R, 16
 un package, 17
terrain, étude de, 124
terrain.colors(), 79, 787
test(s)
 bilatéral, 134, 216, 219, 225, 226, 232
 choix d'un, 292
 conclusion d'un, 204, 258, 281
 conditions d'application d'un, 281, 287
 confirmatoire, 212
 conservateur, 615
 d'ajustement, 288, 615
 d'analyse globale, 212, 282, 283, 284
 d'Ansari-Bradley, 540
 de Bartlett, 557
 de Breslow-Day, 354
 de conformité, 288
 de Fisher-Snedecor, 290, 533
 de Fligner-Killeen, 564

- de Kolmogorov-Smirnov, 624, 646
- de Kruskal-Wallis, 511
- de Lilliefors, 629
- de Mac Nemar, 386
- de Mann-Whitney-Wilcoxon, 476
- de Mantel-Haenszel, 349
- de rangs, 182, 288, 480
- de Shapiro-Wilk, 637
- de signification, 288
- de Welch, 431
- de Wilcoxon-Mann-Whitney, 476
- décomposition du, 282, 283
- des signes de Wilcoxon, 462, 494
- d'homogénéité, 288
- d'indépendance, 288
- du χ^2 , 290, 293, 305, 616, 650
- du logrank, 694
- d'un coefficient de corrélation, 290, 575, 586, 600
- d'une droite de régression, 290, 661, 680
- efficacité d'un, 292
- ensemble de, 211
- exact de Fisher, 328, 344
- exploratoire, 212
- G, 321
- hypothèses de, 281
- initial, 284
- multiples *voir* comparaisons deux à deux
- non paramétrique, 137
- paramétrique, 136
- pré-, 214, 284
- robuste, 287
- t de Student, 290, 417
- unilatéral, 134, 216, 226, 282
 - droit, 218, 221, 229, 282
 - gauche, 219, 223, 230, 282
- text()*, 68, 785
- théorème central limite, 719
- tiff()*, 83, 788
- tirage, 126, 146, 199
- title()*, 68, 785
- tolower()*, 783
- topo.colors()*, 79, 787
- toupper()*, 783
- transform()*, 48, 781
- travail
 - environnement de, 18, 754
 - espace de, 18, 755
 - répertoire de, 33
- trigonométrique, fonction, 774
- TRUE*, 24, 774
- trunc()*, 774
- try()*, 277, 788
- Tschuprow, coefficient d'association de, 610
- two-sided test*, 216
- two-tailed test*, 216
- typeface*, 71
- typeof()*, 755, 772, 778

U

- unilatéral(e)
 - hypothèse, 216
 - test, 134, 216, 218, 221, 226, 229, 282
- unique()*, 781
- unité d'échantillonnage, 89
- unlist()*, 778
- UseMethod()*, 767
- user's workspace*, 755
- usr*, 787

V

- valeur résiduelle, 659
 - standardisée, 666, 678
 - studentisée, 667, 678
- var()*, 54, 789
- var.test()*, 290, 536, 790
- variable
 - aléatoire, 90
 - binaire, 95
 - continue, 94
 - de durées, 688
 - de suivi, 688
 - de survie, 688
 - de test (VT), 135
 - discontinue, 94
 - discrète, 94
 - explicative, 657
 - expliquée, 657
 - fixée, 97
 - fluctuation d'une, 196
 - libre, 746, 747
 - locale, 746, 747
 - nominale, 95
 - ordinaire, 95
 - prédictive, 657
 - prédite, 657
 - qualitative, 94
 - quantitative, 94
 - semi-qualitative, 105
 - semi-quantitative, 95
- variance(s), 112
 - comparaison de, 290, 533, 540, 557, 564
 - interclasse, 448
 - intraclasse, 448
 - résiduelle, 448, 669
 - totale, 448
- vecteur, 26
- vector()*, 26, 29, 776
- virgule flottante, 775
- VT, 135
 - calcul de la, 281
 - discrète, 283
 - distribution réelle de la, 215, 243
 - fluctuation d'une, 196, 197
 - orientée, 282

W

- warnings()*, 788

Welch, 431, 454
which(), 41, 779
which.max(), 779
which.min(), 779
while(){}, 59, 788
whiskers, 65
wilcox.test(), 290, 468, 485, 503, 790

Wilcoxon

distribution de probabilité de, 185, 190, 790
test de, 476

test des signes de, 462, 494

Williams, correction de, 323

windows(), 82, 788

windows.options(), 83, 788

write(), 742, 777

write.fwf(), 742

write.table(), 742, 777

X

X11(), 82, 788

X11.options(), 83, 788

xaxp, 787

xaxs, 736, 787

xaxt, 787

xlab, 66

xlim, 66

xor(), 773

xpd, 787

xtabs(), 31, 776

Y

Yates, 727, voir aussi correction de continuité

yaxp, 737, 787

yaxs, 736, 787

yaxt, 787

ylab, 66

ylim, 66

Z

zone grisée, 281

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 3^e 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

