Université Claude Bernard Lyon 1

NOTES DE COURS DE STATISTIQUES

INFÉRENCE STATISTIQUE, APPLICATION À L'ENTRAÎNEMENT SPORTIF

Formation : M1PPMR

UE : STATISTIQUES/GESTION

2009-2010, Automne

S. CHAMPELY & J. BASTIEN

Document compilé le 19 janvier 2010

Identification Apogée

Matière	Statistiques
Formation	Master 1 P.P.M.R (P)
Formation (code)	SPM103
UE	1MTR2 Statistique/Gestion
UE (code)	SP1002M1

Table des matières

Avant-propos

Ces notes de cours constituent un support **provisoire** de cours, TD et TP de Statistiques pour l'UE Statistiques/Gestion du M1PPMR (2009-2010, Automne).

Le contenu du cours correspond approximativement aux chapitres 5, 6, 8, 10, 11 et 12 de l'ouvrage de Stéphane CHAMPELY [**Cha04**].

Ce polycopié de cours et les fichiers de données sont normalement disponibles à la fois

- en ligne sur `http://utbmjb.chez-alice.fr/UFRSTAPS/index.html` à la rubrique habituelle ;
- en cas de problème internet, sur le réseau de l'université Lyon I : il faut aller sur :
  - 'Poste de travail',
  - puis sur le répertoire '`P:`' (appelé aussi : enseignants sur '`\Univ-lyon1\enseignement\homes`'),
  - puis '`jerome.bastien`',
  - enfin sur '`M1PPMR`'.

Pour l'examen, les données se trouveront aussi, par mesure de précaution à ces deux endroits.

On trouvera quelques références bibliographiques (voir page 157).

Vous trouverez

- au chapitre 10, l'essentiel (et l'exigible aux examens !) des notions, définitions, propriétés, exercices et manipulations avec R et Rcmdr qu'il faut savoir (ou retrouver dans le polycopié de cours) ;
- en annexe A, un petit guide d'installation du logiciel R et du package Rcmdr, pour ceux qui souhaitent l'installer sur leur propre ordinateur ;
- en annexe B, une prise en main à la première séance, pour ceux ceux qui se sentent peu habitués aux opérations de téléchargement de fichiers, de démarrage de logiciels.

Pour ceux qui utilisent les ordinateurs[1] de l'université Lyon I, à cause d'un problème avec le package Rcmdr de la version 2.9 de R, on prendra bien garde à utiliser la version 2.7 de Ret non la version 2.9 ; on trouvera cette version, comme d'habitude, en faisant "démarrer", puis "programmes", puis "R" puis "R 2.7". Si cette version n'est pas installée sur votre ordinateur, il faut le redémarrer !

[1] en date du mois de décembre 2009 ; ce bug a peut-être été corrigé depuis !

CHAPITRE 1

Introduction

Ce cours est une introduction à la statistique inférentielle orientée vers les méthodes d'entraînement sportif. Les chapitres sont d'ailleurs organisés en fonction de questions précises du type "Mesurer la progression d'un groupe". De véritables jeux de données collectés par les étudiants de M2PPMR servent d'illustration aux procédures statistiques les plus opérationnelles dans le champ de l'entraînement.

La statistique inférentielle a pour objet de généraliser les résultats observés sur un échantillon. Toutefois, elle demande d'abord une description précise de cet échantillon par des méthodes graphiques ou des résumés statistiques. C'est pourquoi nous présenterons ces méthodes descriptives pour ceux qui n'ont pas suivi préalablement une formation de statistique en L3 et les verrons en situation pour les autres. Ce sera l'objet, entre autres, des chapitres 2 et 3. Ceux qui n'auront jamais fait de statistiques pourront lire l'annexe C, qui ne sera pas traitée en cours.

L'approche inférentielle choisie ici est particulière car classiquement on présente surtout les tests d'hypothèses dans ce type de cours. Il nous semble que ces techniques de tests bien que largement répandues ne sont pas les meilleures, aussi le lecteur s'apercevra que nous insistons plus largement sur les méthodes descriptives, les notions de taille d'effet et d'intervalle de confiance.

Si les formules mathématiques sont présentées et décortiquées, elles ne sont pas indispensable à l'application concrète de ces méthodes car nous emploierons un logiciel appelé R dans sa version interactive Rcmdr. Ce logiciel est libre de distribution et peut donc être utilisé dans le cadre professionnel sans souci. On trouvera en annexe un guide d'installation de ce logiciel (Voir annexe A).

CHAPITRE 2

Évaluation de la réussite aux lancers francs (basket) (problème à une proportion)

Ceux qui n'auront jamais fait de statistiques pourront lire les sections C.1, C.2 et C.3 de l'annexe C, qui ne sera pas traitée en cours.

Ceux qui se sentent peu habitués aux opérations de téléchargement de fichiers, de démarrage de logiciels et pourront lire l'annexe B en première séance.

2.1. La lecture d'un jeu de données

Les jeux de données utilisés en statistique ressemblent à des feuilles de calcul Excel. Cependant, pour avoir des fichiers plus petits et surtout, non dépendant d'excel, tous les fichiers sont fournis au format texte (avec extension txt).

Les lignes du fichier sont constituées des *unités statistiques*, c'est-à-dire les individus, objets, groupes que l'on mesure et les colonnes des mesures correspondantes.

Ainsi, le jeu de données `BASKET.txt` correspond à 150 tentatives personnels de tirs au panier de basket (en face) suivant différentes distances et à la réussite dans cet exercice.

Manipulation avec Rcmdr 2.1. Afin d'importer ce ficher dans le logiciel R, et une fois que R et Rcmdr sont ouverts (voir annexe B), il faut suivre les étapes suivantes :

(1) Dans le menu déroulant "Données" de Rcmdr, choisir l'option "Importer des données" puis "Depuis un fichier texte ou le presse-papier...". Dans la fenêtre de dialogue qui s'ouvre, donner un nom au jeu de données (à la place de Dataset, choisi par défaut), le nom du fichier texte sans extension, c'est-à-dire : '`BASKET`'. Laisser les autres champs avec les valeurs choisies par défaut.

(2) Employer la fenêtre qui s'ouvre alors pour retrouver le fichier à importer ('`BASKET.txt`').

(3) Cliquer alors éventuellement sur le bouton "Visualiser".

Après cette procédure, le jeu de données est actif pour le logiciel R.

Le jeu de donnée que l'on vient de charger est composé de deux colonnes : '`Result`' et '`Dist`'.

Manipulation avec Rcmdr 2.2. *Attention*, parfois (notamment si vous travaillez sur de vieux sujets d'examens) vous aurez à ouvrir des fichiers xls, format abandonné à cause de problèmes trop nombreux de compatibilité entre R et excel. Néanmoins, la manipulation à faire est presque identique à la précédente :

(1) Dans le menu déroulant "Données" de Rcmdr, choisir l'option "Importer des données" puis "Depuis Excel, Access ou dBase...". Dans la fenêtre de dialogue qui s'ouvre, donner un nom au jeu de données (à la place de Dataset, choisi par défaut), le nom du fichier xls sans extension.

(2) Employer la fenêtre qui s'ouvre alors pour retrouver le fichier xls.

Après cette procédure, le jeu de données est actif pour le logiciel R.

2.2. Les types de mesures

La première chose essentielle dans l'analyse d'un jeu de données est de comprendre la nature des mesures. Il existe deux grands types de mesures : les *mesures catégorielles* et les *mesures numériques*[1]. La variable Result correspond à une mesure catégorielle comme peuvent l'être le sexe, la PCS, le sport pratiqué. La variable Dist, qui mesure en mètres la distance au panier est une mesure numérique comme la taille, le poids, la vitesse, le volume du mollet .... Ce chapitre est consacré à l'analyse des mesures catégorielles.

2.3. Les graphiques statistiques pour les mesures catégorielles

Les mesures catégorielles s'analysent très simplement, il s'agit de dénombrer le nombre d'unités statistiques dans chaque catégorie. Graphiquement, on représente alors ces catégories de façon proportionnelle afin de visualiser leur importance. Le graphique en barres s'obtient en utilisant le menu déroulant "Graphes" puis "Graphe en barres". Le camembert s'obtient avec le menu déroulant "Graphes" et l'option "Graphe en camembert". On voit qu'il y a plus de paniers manqués que réussis.

2.4. Les statistiques pour mesures catégorielles

Ce sont tout simplement les effectifs de chaque catégories, les fréquences et les pourcentages correspondants. Pour les calculer, employer le menu déroulant "Statistiques" puis l'option "'Résumés" puis "Distributions de fréquences". On lit alors que 59 paniers ont été réussis sur 150 soit un pourcentage de 39%.

2.5. Problèmes classiques concernant les données catégorielles

2.5.1. Les problèmes de sondage

Le jeu de données '`DiarrhF.txt`' contient les réponses de 38 femmes marathoniennes à la question : "A l'issue de vos marathons, avez-vous souvent des problèmes digestifs de type diarrhées ?".

EXERCICE 2.3. Analyser ce jeu de données. Quel pourcentage de femmes souffrent de ce type de problèmes ?

Les données manquantes sont un problème récurrents dans les sondages ou les études observationnelles. Le taux de réponse est un bon indicateur de risque de biais.

EXERCICE 2.4. Le jeu de données '`DiarrhH.txt`' comprend les réponses de 363 hommes à la même question. Analyser ce jeu de données. Comparer le, par le calcul seulement, au résultat des femmes. Quel avantage présente ce jeu de donnée sur le précédent ?

Avec les mesures catégorielles, il importe d'avoir une taille d'échantillon relativement grande afin de pouvoir réaliser des calculs précis de pourcentages.

2.6. Extension à l'étude de plus de deux catégories

Nous avons jusqu'à présent étudier des cas ou seules deux catégories existaient. Le fichier '`FFHandi.txt`' contient la nature du handicap des licenciés (1991) de la Fédération Française Handisport. Nous allons voir que l'analyse se généralise sans changement à plus de deux catégories.

EXERCICE 2.5. Utiliser les méthodes précédentes pour décrire la répartition des licenciés dans ces catégories. Que signifie la catégorie "Autres" ? N'y a-t-il pas une catégorie dont la présence est surprenante ? Quel peut être l'intérêt de connaître également la fréquence de ces différents handicaps dans la population française ?

[1] On parle aussi de mesures qualitatives et quantitatives

Voyons à présent dans le cadre d'une étude des pratiques sportives des camerounais (de 15 à 75 ans) la généralisation à un nombre plus important de catégories : le sport pratiqué par les sondés. Le jeu de données '`cameroun1.txt`' comprend les résultats d'un sondage effectué sur 395 individus.

EXERCICE 2.6. Analyser la variable APS qui décrit l'activité sportive déclarée par le sondé. En particulier réaliser le graphe en barres et le graphe en camembert. Quel est le nouveau problème qui se pose ici ?

Dans cet exercice, le grand nombre de catégories rend difficile l'analyse. Les solutions qui s'offrent à nous sont soit de regrouper les catégories en sous-ensemble qui ont du sens (sport de combat, sport collectif...) soit de créer une catégorie "Autre" qui regroupe les catégories ayant les plus petits effectifs. Afin de visualiser l'ensemble, on peut aussi utiliser un graphe particulier, le graphe en points, qui n'est malheureusement ps disponible dans la version interactive de R.

REMARQUE 2.7. Il est possible d'employer R dans une version langage de commandes qui est beaucoup plus puissante. On tape alors des commandes dans la fenêtre de script puis on clique sur le bouton "Soumettre". On peut aussi taper des commandes dans la console (fenêtre "Rgui") et appuyer sur la touche "Enter".

Le résultat s'affiche soit dans la fenêtre de sortie soit dans une fenêtre graphique.

REMARQUE 2.8.

(1) Si vous disposez du document pdf de ce cours, vous pouvez normalement copier-coller une ou plusieurs lignes du document pdf vers la fenêtre de commande "Rgui", ce qui pourra vous éviter des erreurs de frappe !
(2) Dans "Rgui", vous pouvez réutiliser les commandes déjà saisies et les modifier en les rappelant grâce aux flèches "haut" et "bas" du clavier (↑ et ↓) et les modifier en vous aidant des fléches "gauche" et "droite" du clavier (← et →)
(3) Dans ce polycopié,
   - Conformément à la présentation de R, les "entrées" (les instructions à taper dans la fenêtre de commande "Rgui", derrière le "prompt" >) sont présentées en rouge et en police "machine à écrire", comme par exemple ce qui suit :
     ```
     cos(2)
     ```
   - De même, conformément à la présentation de R, les "sorties" (ce qui sera calculé par R) sont présentées en bleu et en police "machine à écrire", comme par exemple ce qui suit :
     ```
     [1] -0.4161468
     ```
   - Enfin, si l'entrée et le résultat sous présentés simultanément, vous verrez
     ```
     cos(2)
     [1] -0.4161468
     ```

EXERCICE 2.9. Soumettez successivement les 4 ordres suivants (attention, ici `cameroun1` est éventuellement à remplacer par le nom de la variable que vous avez choisi lors de l'importation des données, `Dataset` par défaut)

(1) `cameroun1$APS`
(2) `table(cameroun1$APS)`
(3) `dotchart(table(cameroun1$APS))`
(4) `dotchart(sort(table(cameroun1$APS)))`

Que font ces lignes de commandes ? Quelles sont les catégories les plus importantes ? Quelles catégories suggérez-vous de regrouper dans une catégorie "Autres" ?

2.7. Les données ordinales

Le fichier '`sauna.txt`' comprend les réponses de 687 sondés à une enquête de satisfaction concernant les piscines lyonnaises d'hiver. La question portait ici sur l'implantation d'un sauna, était-elle : Trés souhaitée, Souhaité, Indifférent, Pas souhaité, Pas du tout souhaitée ?

EXERCICE 2.10. Analyser ces souhaits en utilisant les méthodes précédemment décrites. Quelle est la nouveauté concernant ces catégories et en quoi les graphes ne sont pas très bons ? Comment regrouper habilement de telles catégories ?

REMARQUE 2.11. Pour modifier l'ordre de catégories, il faut utiliser le menu déroulant "Données", l'option "Gérer les variables dans le jeu de données actif", puis "Réordonner une variable facteur".

2.8. Dangers des mauvaises affectations ! ! (section facultative)

EXERCICE 2.12. *ATTENTION*, comme matlab ou d'autres logiciels du même type, R présente l'inconvénient de pouvoir faire de dangereuses affectations si on utilise les mots-clés (c'est-à-dire, les noms de variables réservées, utilisées par R).

Faire les commandes suivantes et méditer aux dangers mis en évidence, par l'utilisation des mots-clés '`T`' ou '`cos`'.

```
T
```

```
[1] TRUE
```

```
TRUE
```

```
[1] TRUE
```

```
T & F
```

```
[1] FALSE
```

```
T <- c(1, 2, 3)
T & F
```

```
[1] FALSE FALSE FALSE
```

```
rm(T)
T
```

```
[1] TRUE
```

```
sd(c(1, 2))
```

```
[1] 0.7071068
```

```
sd <- cos
cos(c(1, 2))
```

```
[1]  0.5403023 -0.4161468
```

```
sd(c(1, 2))
```

```
[1]  0.5403023 -0.4161468
```

```
sd(1)
```

```
[1] 0.5403023
```

```
rm(sd)
sd(c(1, 2))
```

```
[1] 0.7071068
```

CHAPITRE 3

Évaluation de la performance d'un groupe (problème à une moyenne)

Ceux qui n'auront jamais fait de statistiques pourront lire les sections C.1, C.2 et C.4 de l'annexe C, qui ne sera pas traitée en cours.

3.1. Les données numériques

G. Coquelin lors d'un stage de M2PPMR à l'AS Gien Football est responsable de la catégorie 15 ans a réalisé une série de tests physiques sur ses joueurs afin de réaliser un état des lieux. Le fichier '`coquelin.txt`' comporte donc pour 16 individus la mesure de leur endurance vitesse (test 4*10 en navette), de leur puissance aérobie (VMA, test Vameval), de leur détente verticale et des 5 enjambées.

Nous allons dans un premier temps travailler sur le résultat de détente verticale (jambes semi flêchies, sans élan, aller toucher le plus haut possible, en mètres).

Charger le fichier de données '`COQUELIN.txt`'.

3.2. Les graphes pour données numériques

3.2.1. Le graphe indexé

Le premier graphe disponible dans Rcmdr est le *graphe indexé*. Il indique en abscisses le numéro de la mesure et en ordonnées sa valeur. On obtient ce graphe en utilisant le menu déroulant "Graphes" et l'option "Graphe indexé". Les mesures s'échelonnent généralement de 2.35 m à 2.55 m avec une valeur exceptionnelle à 2.68m. Voir le premier graphique de la figure 3.1 page suivante.

REMARQUE 3.1. L'utilisation du numéro de mesure n'est pas vraiment utile et il est sans doute aussi intéressant de placer toutes les valeurs sur une même ligne. Ceci doit se faire en utilisant la commande suivante :

```
stripchart(coquelin$DétentVert)
```

Attention au problème d'accent de 'DétentVert' qui ne passe par le copier-coller !

Cela doit produire le deuxième graphique de la figure 3.1. Ce graphique est appelé une ligne de point.

Le problème étant que les valeurs égales sont superposées, pour les empiler on peut utiliser :

```
stripchart(coquelin$DétentVert, method = "stack")
```

3.2.2. Le graphe de dispersion

Afin de résumé les aspects essentiels du jeu de données, la *boîte de dispersion* indique la valeur centrale par un trait gras, indique l'étendue des données par (1) une boîte et (2) des moustaches et (3) des points. Le principe en est simple, la médiane est indiquée par un trait central gras, les deux quartiles forment les extrémités de la boîte et, sortant de la boîte, deux moustaches essaient de rejoindre les valeurs minimum et maximum. Si ces deux valeurs sont trop éloignées[1], elles apparaissent comme des points que les moustaches ne rejoignent pas.

[1]Par rapport à un critère que nous ne développerons pas ici

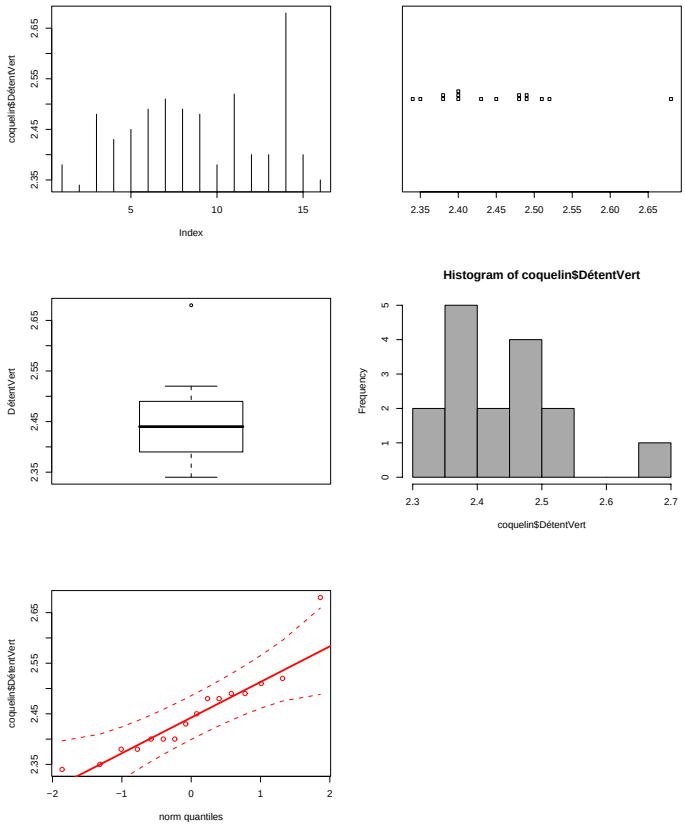


FIG. 3.1. Graphe indexé, Ligne de points, Boîte de dispersion, Histogramme et Graphe quantile-quantile sur les données de détente verticales de Coquelin.

Ce graphique est disponible par le menu déroulant "Graphes" et l'option "Boîte de dispersion". La mesure extrême est fort bien mise en exergue et la valeur centrale est proche de 2,45 m. Voir le troisième graphique de la figure 3.1.

3.2.3. L'histogramme

L'*histogramme* est une technique qui consiste à regroupe les mesures en un certain nombre de catégories et à réaliser un graphe en barres le long d'une échelle numérique indiquant ces catégories.

Le menu déroulant "Graphes" donne accès à l'option "Histogramme". Ce graphique est surtout intéressant lorsque le nombre de mesures est important (plus de 50 données). Il permet alors de décrire la concentration des données. Avec $n = 16$ valeurs comme en l'espèce, il devient très instable en fonction du choix des intervalles définissant les groupes. Voir le quatrième graphique de la figure 3.1.

REMARQUE 3.2. Pour tracer les histogrammes, on a le choix entre
- fréquence (nombre d'unités statistiques par classe) ;
- pourcentage (nombre d'unités statistiques par classe divisée par 100) ;
- densité (fréquence divisée par le produit de la largeur de la classe par le nombre total d'individu).

Ici, il y a peu de différences entre ces trois possibilités. Retenez que les histogrammes en densité sont plus "stables" par rapport aux nombre de classes choisie. De plus, l'histogramme en densité est "normalisé", c'est-à-dire que son aire totale est égale à 1 et il pourra être ainsi comparé à des lois théoriques de probabilité.

3.2.4. Le graphe quantile-quantile

Ce graphe cherche à confronter l'histogramme des données à une forme prototypique, celle de la loi normale[2] dite aussi courbe en cloche. Il est très important de déterminer si les données suivent approximativement cette forme car les procédures statistiques que nous verrons par la suite sont généralement basées sur cette hypothèses. Lorsque les données suivent la loi normale, les points sont situés exactement sur une droite. Toutefois, un écart est inévitable, l'écart normal étant symbolisé par deux courbes en pointillées sur le graphe. On constate sur cet exemple que les données sont globalement compatibles avec l'hypothèse de normalité, sauf en ce qui concerne la valeur extrême. Voir le cinquième graphique de la figure 3.1.

3.3. Les statistiques de centralité

Les graphiques permettent de repérer les aspects essentielles de la distribution (concentration) des données. Il convient ensuite de passer à une description plus précise basée sur des calculs que l'on appelle des *statistiques*.

Les statistiques servent à décrire un aspect précis des mesures. Nous allons commencer par des statistiques qui s'attachent à décrire la mesure typique de l'échantillon.

La statistique la plus classique pour ce faire est la moyenne, il s'agit de la somme des mesures divisée par leur nombre. Le menu déroulant "Statistiques" avec l'option "Résumés" puis "Statistiques descriptives" donne une moyenne[3] de $M = 2.44875$. Rappelons que la moyenne des $(y_i)_{1\leq i\leq N}$ est donnée par

$$M = \frac{1}{N}\left(y_1 + y_2 + \ldots + y_N\right) = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{3.1}$$

La moyenne n'est pas la seule façon de quantifier la centralité, la notion de médiane, notée M_d, est également importante : la médiane est une valeur telle que la moitié des mesures lui sont inférieures (et l'autre moitié supérieures). On trouve ici $M = 2.44$.

La différence est ici faible entre ces deux statistiques de centralité. Ce n'est pas toujours le cas. En particulier, lorsqu'i existe des mesures extrêmes ou bien une *dissymétrie* des données, c'est-à-dire que les mesures s'écartent plus d'un côté de la valeur centrale que de l'autre, moyenne et médiane peuvent être très différentes. Il convient alors de privilégier la médiane qui est plus *résistante*, c'est-à-dire qu'elle est moins influencée par les valeurs exceptionnelles.

Exercice 3.3. Calculer la moyenne et la médiane pour le test des 5 enjambées.

3.4. Les statistiques de variabilité

Au delà de la notion de centralité, la statistique accorde une importance toute particulière à la notion de *variabilité*. Il s'agit de prendre conscience de la variabilité naturelle des mesures et du fait que la relation ne sont pas déterministes. Il existe toujours autour de la mesure typique des écarts, la statistique propose justement de les quantifier.

La façon la plus simple de quantifier la variabilité est l'*étendue*, il s'agit de l'écart entre la valeur maximum et la valeur minimum du jeu de données. On trouve en l'espèce 2.68 et 2.34, ce qui conduit à une étendue de 0.34 m. L'atour majeur de cette statistique est la simplicité mais elle est malheureusement absolument pas résistante. Pour cette raison, on lui préfère l'*écart inter-quartiles*.

Les *quartiles* sont des généralisations de la médiane, il s'agit cette fois de découper les données en 4 sous-ensembles (et non pas 2) de même effectifs : un quart des données est ainsi inférieur au premier quartile, ici $Q_1 = 2.395$, une moitié au deuxième quartile $Q_2 = 2.44$ qui n'est autre que la médiane et trois quart au

[2]Beaucoup de choses plus loin à ce sujet

[3]On note parfois la moyenne $\overline{y}$

troisième quartile ici $Q_3 = 2.49$. Les quartiles sont beaucoup plus résistants que les extrémités, aussi l'étendue inter-quartiles donne une indication stable de la variabilité des données et plus précisément de sa partie centrale.

Enfin, la statistique de variabilité la plus classique est l'écart-type. Elle n'est cependant ni la plus facile à calculer ni la plus résistante. On doit pour l'utiliser à bon escient vérifier sur les graphiques que la distribution est symétrique et qu'il n'y a pas de données extrêmes. L'écart-type est en l'espèce égal à

$$SD = 0.08413283 \tag{3.2}$$

Cette quantité indique la façon dont les données s'écarte usuellement de la valeur moyenne, il s'agit en quelque sorte d'une moyenne des écarts.

REMARQUE 3.4. Si $\overline{y}$ désigne la moyenne des $(y_i)_{1\leq i\leq N}$, l'écart-type des $(y_i)_{1\leq i\leq N}$ est donné

$$\sigma = \sqrt{\frac{1}{N}\left((\overline{y}-y_1)^2+\ldots+(\overline{y}-y_N)^2\right)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\overline{y}-y_i)^2} \tag{3.3}$$

Voir remarque C.11 page 127 de l'annexe C. Attention, l'écart-type calculé par R n'est pas celui-ci. R calcule en fait la déviation standart (notée SD) et définie par

$$\sigma = \sqrt{\frac{1}{N-1}\left((\overline{y}-y_1)^2+\ldots+(\overline{y}-y_N)^2\right)} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\overline{y}-y_i)^2} \tag{3.4}$$

Quand N est grand, il y a peu d'écart entre σ et SD.

Si on veut calculer le "vrai" écart-type σ des $(y_i)_{1\leq i\leq N}$, on pourra rentrer dans Rgui les commandes suivantes :

```
x <- coquelin$DétentVert
sqrt((length(x) - 1)/length(x)) * sd(x)
```

ce qui doit afficher

```
[1] 0.08146126
```

à comparer avec

```
x <- coquelin$DétentVert
sd(x)
```

qui donne SD. Cela affiche le résultat

```
[1] 0.08413283
```

On a donc

$$\sigma = 0.08146126,$$

et

$$SD = 0.08413283,$$

ce qui est bien la valeur donnée par (3.2). Il y a peu de différence entre ces deux valeurs.

Pour toute la suite, on appellera l'écart-type la quantité SD (celle qui est déterminée par R).

EXERCICE 3.5. Calculer les trois statistiques de variabilité (moyenne, écart-type, quartiles) pour le test des 5 enjambées.

REMARQUE 3.6. On présente très souvent les résultats d'un échantillon de mesures numériques sous la forme :

$$M \pm SD.$$

L'intervalle $[M - SD, M + SD]$ est censé contenir la plus grande partie des valeurs y_i.

3.5. Autres aspects des distributions numériques

Afin de savoir quelles statistiques de centralité et de variabilité il est préférable d'utiliser, on doit connaître la forme des données. Trois aspects sont essentiels, la forme, la multi-modalité et les mesures extrêmes.

3.5.1. La notion de symétrie

EXERCICE 3.7. Le jeu de données '`TRANSFERTS.txt`' contient les montants (en Francs 1999) des 93 transferts de la Ligue 1 de football.

- Réaliser un histogramme de ces données
- Recommencer en choisissant 20 comme nombre de classes
- est-ce que cet histogramme à une courbe en cloche ?
- comment décrire la forme de cet histogramme.

L'étude de symétrie essaie de préciser comment s'exerce la variabilité, c'est-à-dire la façon dont les mesures s'écartent de la valeur centrale. Les mesures peuvent en effet s'en écarter de la même façon à gauche et à droite, un peu comme un miroir, on parle alors d'échantillon symétrique ou bien s'écarter de façon plus étendue dans un sens, on parle alors d'échantillon dissymétrique.

Les tests physiques et les mesures anthropométriques sont généralement symétriques. En revanche, les mesures telles que le salaire des individus, la durée de vie d'un équipement sont souvent dissymétriques. La symétrie se détecte sur les graphiques, mais aussi selon la position des deux quartiles par rapport à la médiane.

Si la distribution est dissymétrique, la moyenne est généralement attirée du côté le plus étendue et donne donc une mesure faussée de la valeur typique. De même, l'écart-type, en ne distinguant pas les écarts bien différent de chaque côté constitue un compromis qui n'est pas satisfaisant dans ce type de situation. On leur préférera donc la médiane et les quartiles.

REMARQUE 3.8. Afin de résoudre le problème des distributions dissymétriques, on emploie souvent la technique des transformations. Il s'agit de trouver une formule mathématique qui transforme la variable dissymétrique en une nouvelle variable plus symétrique. L'analyse statistique est alors plus facile. En revanche, on perd l'avantage des unités de mesures originelles souvent mieux connues. Les transformations les plus connues sont le logarithme, la racine carrée et l'inverse.

EXERCICE 3.9. Pour le jeu de données '`TRANSFERTS.txt`', utiliser le menu déroulant "Données" et l'option "Gérer les variables dans le jeu de données actif" puis "Calculer une nouvelle variable". Dans la fenêtre de dialogue, en tant que variable existante indiquer : "Montant", en tant que nom de la nouvelle variable indiquer : "TransLog" et en tant qu'expression à calculer : "log(Montant+1)".

- Réaliser un histogramme sur la nouvelle variable : "Translog". Que constatez-vous ?
- Pourquoi avoir utiliser la formule $\log(x+1)$ plutôt que tout simplement $\log(x)$?

On consulter les élements de correction de cet exercice page 12.

3.5.2. la multi-modalité

EXERCICE 3.10. Le fichier '`ROLLERS.txt`' contient le prix de rollers (en Francs 2000) pour 224 pratiquants de la région lyonnaise.

- Réaliser un histogramme de ces données.
- Recommencer en choisissant comme nombre de classes : 15
- Analyser la forme de cet histogramme. Que constatez-vous ?
- Que pensez-vous de la notion de valeur typique dans ce contexte ?

Il arrive que les données ne soient pas regroupées autour d'une unique valeur centrale mais qu'il existe plusieurs groupes. On parle alors de *multimodalité*. Dans ce cas, chaque groupe devrait faire l'objet d'une

description quantifiée, mais cette analyse (dite analyse de mélange) est complexe et ne sera pas envisagée. On se contentera de descriptions verbales.

3.6. Éléments de correction

Éléments de correction de l'exercice 3.9

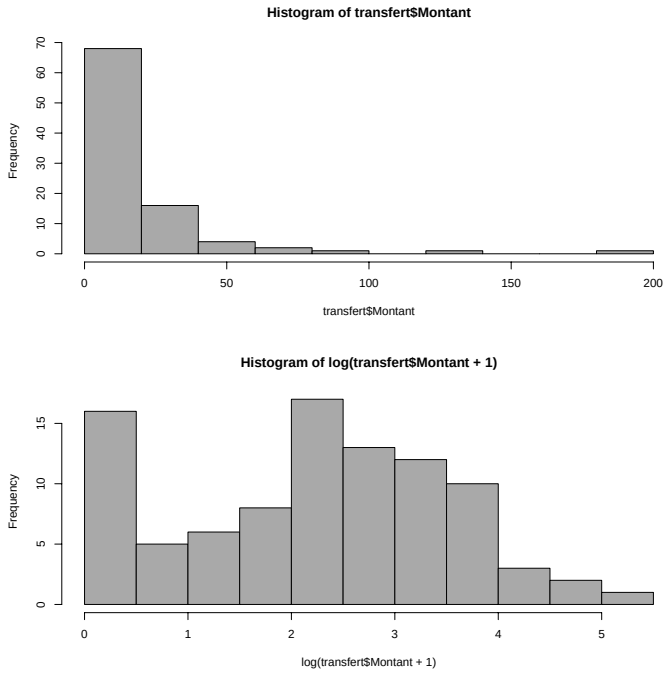


Fig. 3.2. histogrammes des deux variables "Montant" et "TransLog".

Voir sur la figure l'histogtamme de variable "Montant" et celui de la variable "TransLog". On constate que le second histogramme est beaucoup plus proche d'un diagramme en cloche que le premier (mise à part la premiere barre qui correspond à des transfert nuls, pour lequel $\log(x+1)$ est lui aussi nul !).

Cela est confirmé par les deux graphes quantile-quantile de la figure 3.3. Pour tracer ces deux quantiles en enlevant les montants non nuls (voir figure 3.4).

On note x le montant. Puisque $\log(x+1)$ suit une loi normale, cela signie (en prenant l'exponentielle) que x suit une loi en exponentielle d'une loi normale. Autrement dit, ce sont les exposants des montants qui suivent une loi normale, d'ou leur très forte étendue

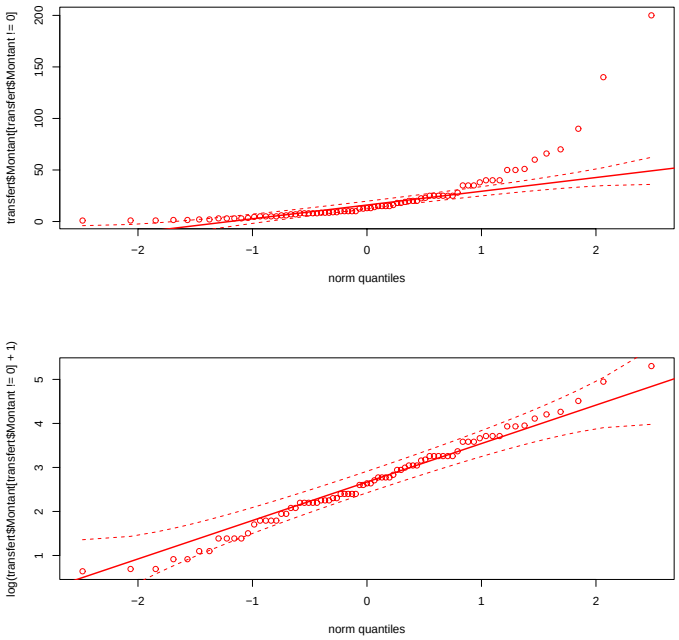


FIG. 3.3. graphes quantile-quantile des deux variables "Montant" et "TransLog" (hors montants nuls).

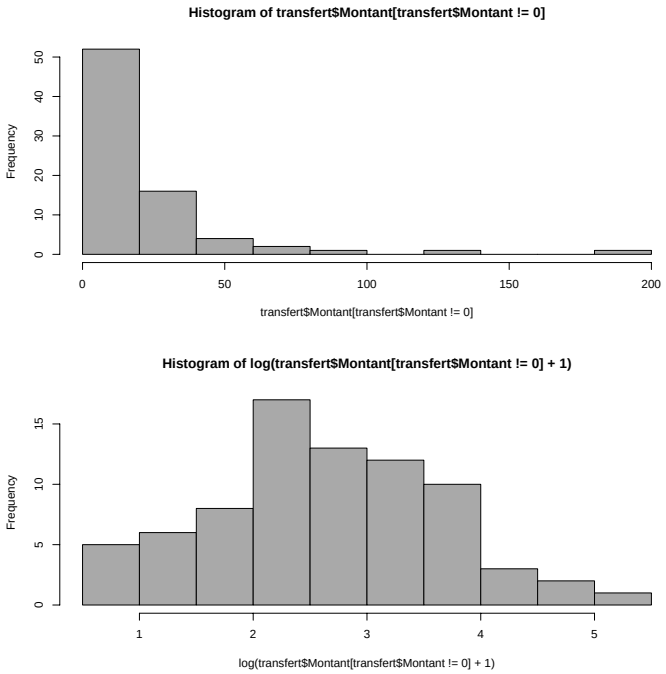


FIG. 3.4. histogrammes des deux variables "Montant" et "TransLog" hors montants nuls..

CHAPITRE 4

Comparaison de la performance d'un groupe à une norme (Comparaison d'une moyenne à une norme)

Les joueurs de football entraînés par Coquelin ont également subi un test physique de VMA. On peut représenter graphiquement les données à l'aide du graphe index (voir le fichier '`COQUELIN.txt`'). On constate que la majorité des résultats vont de 14.5 à 17.0.

Il est très intéressant de posséder une valeur de comparaison afin de pouvoir situer des valeurs. Une première méthode est de disposer d'une "norme", c'est-à-dire d'une valeur moyenne calculée sur une population de référence. Le plus souvent, il s'agit de la population générale correspondante (même âge, même sexe...). Mais en l'espèce, il existe une valeur mesurée sur l'élite de l'INF Clairefontaine pour des joueurs du même âge. La valeur moyenne de l'élite est de 17.35 pour la VMA. Il s'agit donc de voir si le groupe de joueurs est éloigné des caractéristiques de l'élite (pour choisir éventuellement un type spécifique de travail).

4.1. Représentation graphique et comparaison à la norme

Avec $n = 16$ mesures, l'histogramme n'est pas le graphique le plus adapté. On peut alors hésiter entre le graphe indexé qui donne les valeurs individuelles, la boîte de dispersion, le graphe quantile-quantile voire la ligne de points[1].

Il s'agit ensuite d'ajouter au graphique une ligne verticale ou horizontale symbolisant la situation de la norme. Ceci ne peut être que qu'en entrant au clavier une ligne de commandes dans la fenêtre de script et en la soumettant.

Manipulation avec Rcmdr 4.1. Ainsi,

- après avoir tracé un graphe indexé ou une boite de dispersion, il faut dessiner une ligne horizontale en tapant
  ```
  abline(h= 17.35, col = "red", lty = 2)
  ```
- après avoir tracé un histogramme ou une ligne de points, il faut dessiner une ligne verticale en tapant
  ```
  abline(v= 17.35, col = "red", lty = 2)
  ```

Manipulation avec R 4.2. Le graphique le plus adapté est sans doute la ligne de points, elle demande de soumettre deux lignes de commandes dans la fenêtre de script :

```
stripchart(coquelin$VMA, method = "stack", xlab = "VMA")
abline(v= 17.35, col = "red", lty = 2)
```

Quelle que soit le graphe employé (cf figure 4.1 page suivante), la conclusion est la même, il est très rare que les joueurs du groupe aient une puissance équivalente à la norme définie.

Exercice 4.3. Les données du fichier '`DAUCHEZ.txt`' ont été réunies par B. Dauchez (M2PPMR) sur un groupe de joueurs et joueuses de tennis. On s'intéressera en particulier au résultat du test navette de Luc Léger. La norme fournit par la FFT est de 12.75 pour ce test.

(1) réaliser les graphiques

[1] qui malheureusement n'est pas directement disponible dans l'interface graphique

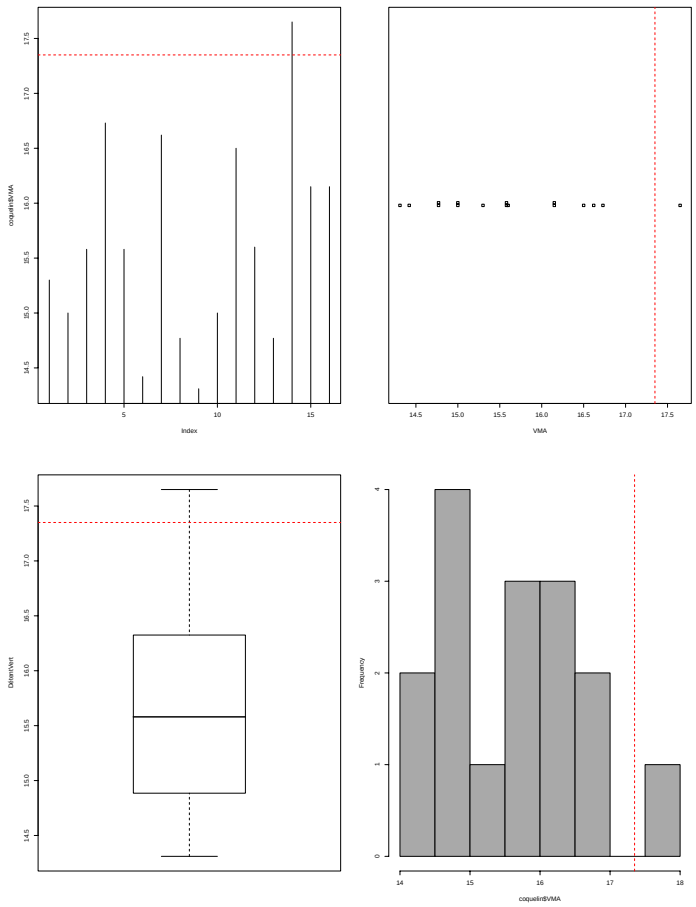


FIG. 4.1. Graphe indexé, Ligne de points, Boîte de dispersion, Histogramme sur les données de VMA de Coquelin. La norme fournie par l'élite de Clairefontaine est indiquée par une ligne rouge en pointillés.

(2) Quel est le graphique le plus adapté à votre sens ?

(3) Reprenez ce graphique et ajoutez la norme de la FFT.

(4) Qu'observez-vous ?

(5) Pouvez-vous trouver dans le fichier de données des explications aux valeurs éloignées de la norme. Pour cela :
   - Introduire une nouvelle variable age, égale à l'année du questionnaire (2007) moins l'année de naissance (stockée dans la variable `Année_2007`) ;
   - Utiliser l'option "graphique", "nuage de points" avec l'age en abscisse et la navette en ordonnée.

Voir les éléments de correction page 18.

4.2. La mesure de la taille de l'effet

Afin de quantifier la comparaison, il s'agit de mesurer l'éloignement de notre échantillon de mesure à la norme. La norme constitue ce qu'on appelle en statistique une *hypothèse*. L'éloignement de l'échantillon est appelé un *effet* qui est en quelque sorte l'effet de l'entraînement et des capacités de ce groupe par rapport à

l'entraînement et aux capacités du groupe de l'élite. Il convient donc de mesurer l'importance de cet effet grâce à une statistique appelée taille d'effet (et notée ES : effect size).

On distingue la taille d'effet absolue :

$$\text{ES} = M - norme, \tag{4.1}$$

et la taille d'effet relative

$$d = \frac{M - norme}{SD}, \tag{4.2}$$

où M est la moyenne et SD l'écart-type. On obtient ici sur les données de VMA de Coquelin [2] comme taille d'effet absolue :

$$\text{ES} = M - norme = 15.63312 - 17.35 = -1.716875$$

et la taille d'effet relative

$$d = \frac{M - norme}{SD} = \frac{15.63312 - 17.35}{0.93674} = -1.83281$$

La taille d'effet absolue peut s'interpréter directement dans les unités originelles tandis que la taille d'effet relative est adimensionnelle (sans dimension). Il s'agit de bien connaître le test physique en question et de décider l'importance pratique d'une différence de VMA de 1.7.

La taille d'effet relative s'émancipe des unités originelles et a l'avantage de permettre de comparer des mesures effectuées dans des unités différentes ou des unités moins connues. Comme guide, il a été proposé par Cohen[**Coh98**] un classement qualitatif de ces tailles d'effet sur la base de la valeur de |d| par rapport à trois valeurs $d_1 = 0.2$, $d_2 = 0.5$ et $d_3 = 0.8$:

$$\text{si } |d| \begin{cases} < d_1, & \text{l'effet est faible,} \\ \in [d_1, d_2[\,, & \text{l'effet est moyenne,} \\ \in [d_2, d_3[\,, & \text{l'effet est fort,} \\ > d_3, & \text{l'effet est très fort} \end{cases} \tag{4.3}$$

Si d est positive ou nulle, la moyenne est supérieure à la norme et sinon, la moyenne est inférieure à la la norme.

Ici, sur les données de VMA de Coquelin, l'effet ainsi mesuré est incontestablement un fort effet ("vers le bas", c'est-à-dire $d \leq 0$ puisque la moyenne est inférieure à la norme).

Exercice 4.4. Pour les données des joueurs de tennis de Dauchez, calculer l'effet par rapport à la norme de la FFT.

Voir les éléments de correction page 19.

4.3. Les tailles d'effet pour les proportions

En ce qui concerne les proportions, il est aussi possible de calculer des tailles d'effet par rapport à une hypothèse. Ainsi en ce qui concerne les joueurs de tennis, il est souvent remarqué qu'il existe beaucoup de gauchers parmi eux. Dans le groupe de Dauchez, on en repère ainsi 3 sur 10. Or dans la population française, il y a environ 10% de gauchers.

Il existe plusieurs méthodes pour calculer une taille d'effet, les deux plus classiques étant simplement de calculer la différence entre la proportion observée p et la norme On distingue la taille d'effet absolue :

$$p - norme \tag{4.4}$$

[2] en tapant directement dans la fenêtre de script et en soumettant la commande

ou le rapport entre ces deux proportions[3]

$$\frac{p}{Norme} \tag{4.5}$$

Nous obtenons dans le cas présent une différence de

$$p - norme = \frac{3}{10} - 0.10 = 0.3 - 0.10 = 0.2 = 20\%$$

ou un rapport de

$$\frac{p}{Norme} = \frac{0.3}{0.10} = 3$$

c'est à dire 20% de gauchers en plus dans l'échantillon ou 3 fois plus de gauchers dans l'échantillon que dans la population française.

Exercice 4.5. Les catégories dans les différents sports sont définies en fonction de l'âge (poussins, benjamins...). Or, à l'intérieur d'une même catégorie, une différence d'un an est considérable en termes de potentiel physique. Il est donc possible que le fait d'être né à la fin de l'année soir un désavantage pour être sélectionné au meilleur niveau et ensuite bénéficier des meilleures conditions et donc finalement avoir une meilleure progression. Ainsi la probabilité d'être dans le quatrième trimestre de l'année doit être normalement de 25%. En revanche, si cet effet d'âge joue la proportion observée dans le haut niveau doit être sensiblement plus basse. Une étude de 1989 sur la LNH canadienne montre que sur 388 joueurs, 69 sont nés durant le dernier trimestre. Calculer le pourcentage correspondant ainsi que les tailles d'effet suivant les deux méthodes.

Voir les éléments de correction page 20.

4.4. La notion de variation d'échantillonnage

Nous savons à présent calculer les performances moyennes d'un groupe et mesurer l'écart à une norme. Toutefois, et c'est l'objet de la statistique inférentielle, on souhaite souvent généraliser les résultats obtenus sur un groupe à un ensemble plus grand pour montrer par exemple l'effet d'une méthode d'entraînement au delà de notre groupe.

Ceci peut être fait si (1) notre groupe est représentatif de la population à laquelle nous souhaitons étendre nos conclusions, (2) nous prenons en compte le fait que notre échantillon n'est qu'un exemple parmi d'autres échantillons qui auraient pu survenir, c'est ce qu'on appelle les variations d'échantillonnage et (3) nous savons caractériser les variations d'échantillonnage.

L'exemple de la proportion de gauchers est ainsi très intéressant pour sensibiliser aux notions que nous allons développer longuement dans le chapitre suivant. Le groupe de joueurs de tennis comportait 3 gauchers sur 10, mais un autre groupe aurait pu comprendre 0, 1, 2, 3, 4 gauchers (voir plus). On voit donc que *la statistique calculée varie d'un échantillon à l'autre.*

Toutefois, on peut de façon intuitive sentir que (1) si la proportion de gauchers est importante chez les joueurs de tennis la statistique prendra des valeurs plus élevées donc , à l'inverse, la valeur de la statistique obtenue d'après l'échantillon doit nous donner une bonne estimation de la valeur réelle dans la population de joueurs et (2) si notre échantillon de joueurs de tennis est plus grand, le calcul de la proportion devrait être plus précis.

4.5. Éléments de correction

Éléments de correction de l'exercice 4.3

(1) Voir les graphiques en figure 4.2 page suivante sur lesquels, on a déjà indiqué la norme de la FFT.

(2) les deux derniers graphiques représente bien l'ensemble des données.

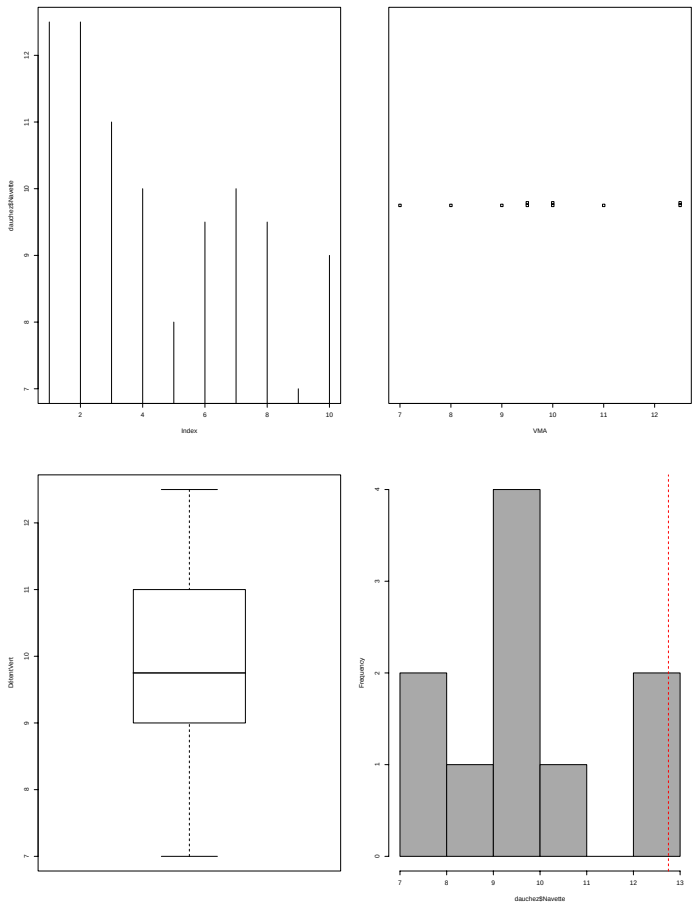


FIG. 4.2. Graphe indexé, Ligne de points, Boîte de dispersion, Histogramme sur les données de Navette de Dauchez. La norme fournie par la FFT est indiquée par une ligne rouge en pointillés.

(3) Voir les graphiques en figure 4.2.

(4) On constate que la norme n'apparaît que sur l'histogramme. En effet, la norme 12.75 est supérieure au maximum des données, ici égale à 12.5. Cette norme apparaît donc uniquement dans la classe $[12, 13]$ de l'histogramme.

(5) Si on réalise le nuage de points (age,navette), on obtient la figure 4.3 page suivante. On a légèrement forcé les limites en y sur ce graphique pour faire apparaître la norme de la FFT. On constate alors que les individus âgés de 19 et 20 ans sont très proche de la norme mais que les autres (excepté celui de 22 ans) sont trop jeunes pour être comparés avec la norme de la FFT.

Sur la figure 4.4 page 21, nous avons tracé les mêmes graphiques par sexe.

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 4.4 On trouve

$$\text{ES} = M - norme = 9.9 - 12.75 = -2.85$$

et la taille d'effet relative

$$d = \frac{M - norme}{SD} = \frac{9.9 - 12.75}{1.76068} = -1.61869$$

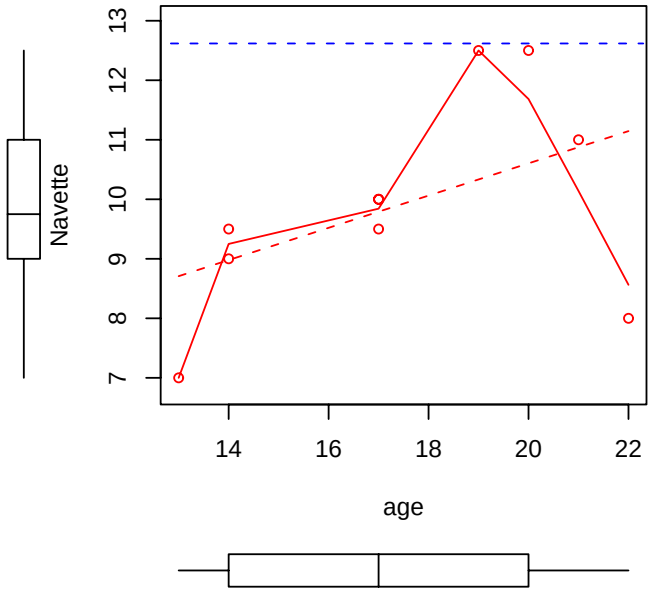


FIG. 4.3. le nuage de points (age,navette) sur les données de Navette de Dauchez. La norme fournie par la FFT est indiquée par une ligne bleue en pointillés.

Ainsi, on a donc un fort effet.

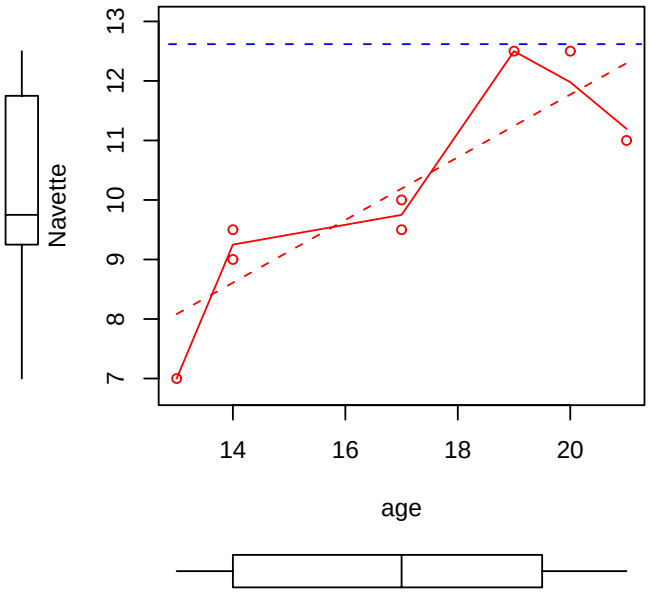
ÉLÉMENTS DE CORRECTION DE L'EXERCICE 4.5 La proportion vaut

$$\frac{69}{388} = 0.1778$$

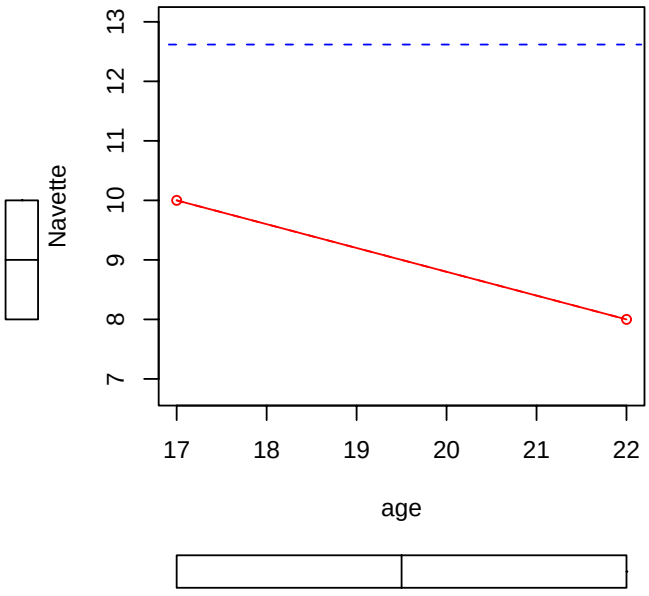
et le rapport vaut

$$\frac{0.1778}{0.25} = 0.7113$$

soit donc 0.7113 "plus" que dans la population totale.



(a) : Hommes



(b) : Femmes

FIG. 4.4. Le nuage de points (age,navette) sur les données de Navette de Dauchez pour les hommes (en haut) et pour les femmes (en bas). La norme fournie par la FFT est indiquée par une ligne bleue en pointillés.

CHAPITRE 5

Généraliser les résultats obtenus avec une proportion

Certains passages seront écrits en petits caractères et finissant par le symbole ♢ (comme la note 37) ; ils pourront être omis ou lus en seconde lecture. Il en est de même de certains exercices facultatif (comme l'exercice 5.4 page 25).

5.1. Qu'est-ce que l'inférence statistique

Le mécanisme de généralisation des résultats est appelé en statistique inférence statistique. Ce mécanisme repose sur une idée très importante, l'échantillon d'unités statistiques que nous observons n'est qu'un échantillon parmi d'autres que nous aurions pu obtenir dans un population plus générale que nous souhaitons décrire.

Or si cet échantillon est aléatoire, c'est-à-dire s'il a été sélectionné sur la base d'un tirage au sort, ou du moins s'il peut être assimilé à un tel mécanisme de sélection, l'échantillon ressemble à la population dont il est issu, il est représentatif. On peut donc estimer une proportion de la population (la proportion de gauchers par exemple) à partir de la proportion que nous observons dans l'échantillon.

Bien plus que cela, la notion de tirage au sort fait qu'on peut définir par des probabilités, les résultats qui peuvent être issus de la population, et en particulier, l'erreur que l'on commet en utilisant la proportion de l'échantillon à la place de la proportion de la population. cela permet de définir ce qu'on appelle un intervalle de confiance.

Une expérience concrète va permettre de comprendre ce qui se passe lorsqu'on utilise un échantillon. Elle servira de base à une courte discussion sur un modèle de probabilité très important : le modèle binomial qui permettra de généraliser les observations faites sur l'expérience concrète à n'importe quel problème concernant des proportions

Ce chapitre est relatif à l'inférence sur la proportion. Nous verrons au cours du chapitre 6, un autre type d'inférence statistique, cette fois-ci relatif à la moyenne.

5.2. L'expérience smarties

Les boîtes de smarties contiennent un nombre aléatoire de bonbons. Nous allons nous intéresser à la proportion de smarties de couleur rouge.

Il s'agit donc dans un premier temps de compter le nombre total de smarties dans la boite, que nous noterons n, par exemple j'observe $n = 32$ smarties dans la mienne puis le nombre de smarties rouges que nous noterons y, j'obtiens personnellement $y = 7$. Il y a donc une proportion de

$$p = \frac{y}{n}$$

smarties rouges, soit pour ma part : $7/32 = 22\%$.

Lorsque nous réalisons une étude, nous ne pouvons disposer que d'un seul échantillon de mesures. Nous devons donc généraliser les résultats à partir de cet unique échantillon. Et la question que nous devons nous poser, c'est : "que se passerait-il si je disposais d'autres échantillons et que je recommençais le même calcul ?"

Ici nous sommes dans une situation particulière car il est possible d'ouvrir une autre boîte de smarties et de recommencer. On peut observer dans le fichier '`SMARTIES.txt`' le résultat de l'ouverture de 15 boîtes par un groupe d'étudiants résumé par les proportions observées.

Avant de faire l'exercice 5.1, bien relire la remarque 2.8 page 5.

EXERCICE 5.1.

(1) Réaliser un histogramme de la variable p du fichier "`SMARTIES.txt`' qui contient les proportions calculées sur 15 boîtes. Décrivez cet histogramme. Conclure sur la proportion de smarties rouges.

(2) Un autre mode de calcul *a posteriori* consiste à prendre tous les résultats ensemble, c'est-à-dire, calculer la proportion obtenue en divisant le nombre total de smarties rouges par le nombre total de smarties. En utilisant la commande,

```
sum(smarties$rouge)/sum(smarties$nombre)
```

déterminer cette proportion et conclure en comparant avec les conclusions de la question 1. Naturellement, quand vous êtes seul face à votre boîte de smarties, vous ne pouvez procéder ainsi !

Voir éléments de correction page 51.

Il reste maintenant à savoir ce qui se passerait si on pouvait multiplier de tels échantillonnages. Les modèles probabilistes nous le permettent.

Nous allons à présent utiliser des probabilités pour expliquer ce qui se passe.

5.3. Notions de probabilités

DÉFINITION 5.2. On parle d'expérience aléatoire, s'il est impossible d'en prévoir l'issue mais qu'en revanche, sur un grand nombre de répétitions, on peut connaître la fréquence avec laquelle les différents résultats apparaitront.

EXERCICE 5.3. Taper les lignes suivantes dans la fenêtre de script et soumettez les successivement.

```
 1:6
[1] 1 2 3 4 5 6
 sample(1:6, size = 1, replace = T)
[1] 1
 sample(1:6, size = 10, replace = T)
 [1] 6 3 1 6 2 3 3 2 1 1
 dede <- sample(1:6, size = 10, replace = T)
```

(Ici `dede` est le nom de la variable dans laquelle on stocke un résultat ; on aurait pu l'appeler autrement.)

```
 dede
 [1] 5 1 5 3 5 4 2 2 5 2
 table(dede)
dede
1 2 3 4 5
1 3 1 1 4
 table(sample(1:6, size = 10000, replace = T))/10000
     1      2      3      4      5      6
0.1714 0.1640 0.1654 0.1689 0.1705 0.1598
 table(sample(1:6, size = 1e+05, replace = T))/1e+05
      1       2       3       4       5       6
0.16582 0.16665 0.16590 0.16827 0.16743 0.16593
```

Attention, puisque des fonctions aléatoires interviennent, elles fournissent donc des valeurs différentes à chaque appel (et donc *a priori* différent de ce qui est écrit en bleu ici !).

Voir éléments de correction page 52.

EXERCICE 5.4 (facultatif).

(1) Quelle ligne de commande vous permet de simuler un tirage d'une grille de loto ?

(2) Comment simuler un nombre entier quelconque $n \geq 1$ grilles de loto ?

(3) Pourriez-vous retirer un enrichissement personnel (au sens propre du terme) de cet exercice ?

Voir éléments de correction page 52.

REMARQUE 5.5. Pour toute la suite de ce cours, nous allons utiliser des fonctions. Lire pour cela l'annexe D page 133.

EXERCICE 5.6. Comme dans l'annexe D page 133, récupérez et sourcez la fonction `graphe_desimple.R`. Pour un entier n donné, cette fonction simule n lancé de dès et calcule les proportions d'apparition de chacun de numéro des faces pour les lancés 1, 2, ..., n. Six courbes sont tracées et permettent de vérifier graphiquement que ces proportions d'apparition des six numéros tendent vers 1/6.

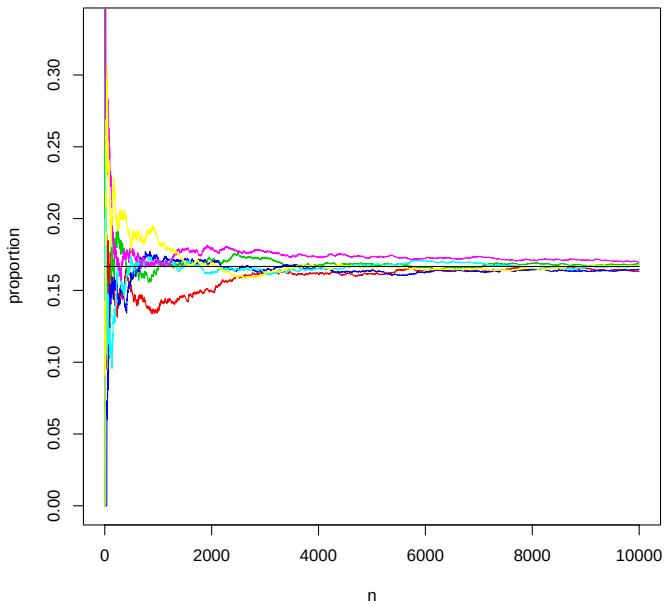


FIG. 5.1. Une simulation de la fonction graphe_desimple pour $n = 10000$.

Tapez pour cela par exemple :

```
graphe_desimple(10000)
```

Vous obtiendrez par exemple la figure 5.1.

EXERCICE 5.7 (facultatif).

Pour ceux qui sont habitués aux logarithmes, consulter la fonction `graphe_desimple.R` et taper par exemple

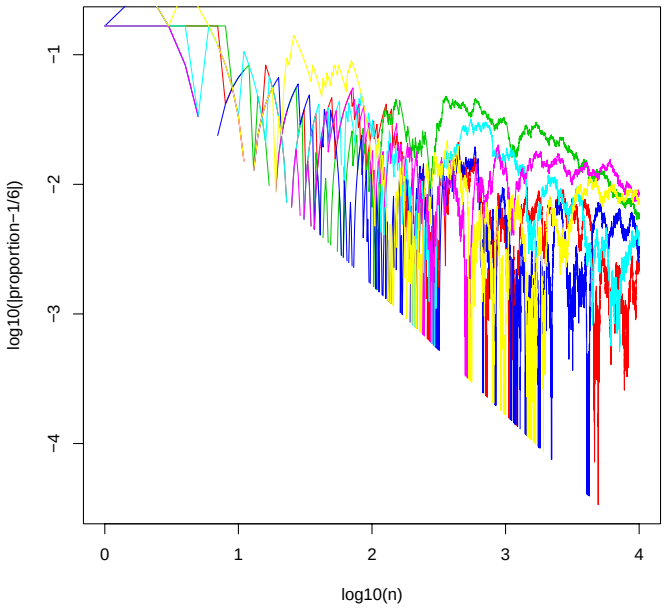


FIG. 5.2. Une simulation de la fonction graphe_desimple pour $n = 10000$ en logarithme.

```
graphe_desimple(10000,translog=TRUE)
```

Vous obtiendrez la figure du type de la figure 5.2.

EXERCICE 5.8 (facultatif). Comme dans l'annexe D page 133, récupérez et sourcez la fonction `desimple.R`, qui permet de calculer la différence entre les proportions d'apparition des 6 numéros et la quantité 1/6 pour un tirage de taille n. pour des valeurs de n de plus en plus grande. Tapez pour cela par exemple :

```
 desimple(100)
```

```
      1       2       3       4       5       6
 0.0033  0.0033 -0.0067 -0.0167 -0.0267  0.0433
```

```
 desimple(1000)
```

```
      1       2       3       4       5       6
-0.0077  0.0053 -0.0057  0.0033  0.0043  0.0003
```

et constatez que les différences obtenus sont de plus en plus faibles.

Attention, puisque ces résultats proviennent de tirages aléatoires, vous aurez tous des résultats différents ! Et même à chaque appel de la fonction !

EXERCICE 5.9 (facultatif). Comme dans l'annexe D page 133, récupérez et sourcez la fonction `deuxdessom` qui renvoie (entre autres) les différentes proportions d'apparitions des événements de l'expérience aléatoire suivante : "jeter deux dès n fois indépendamment et noter la somme des deux numéros sortis". Par exemple :

```
 deuxdessom(1000)
```

```
$prop
res
    2     3     4     5     6     7     8     9    10    11    12
```

```
0.024 0.048 0.083 0.111 0.148 0.155 0.134 0.116 0.086 0.061 0.034

$m
[1] 7.107

$v1
[1] 5.893551

$v2
[1] 5.89945
```

DÉFINITION 5.10. La probabilité d'un événement est la proportion de fois où il se réalise dans une série longue d'une même expérience aléatoire. Une probabilité est toujours comprise entre 0 et 1.

REMARQUE 5.11 (En option). En fait, plus rigoureusement, il faudrait écrire :

DÉFINITION 5.12. Si une expérience est aléatoire, la probabilité d'un événement est la "limite" quand série est "longue" de la proportion de fois où il se réalise.

DÉFINITION 5.13. Dans le cas où l'ensemble des résultats est "discret" (c'est-à-dire avec un nombre fini de valeurs possible), la probabilité d'un événement est égal à

$$p = \frac{\text{Nombre de cas favorables}}{\text{Nombre de cas possible}},$$

où le nombre de cas favorable est le nombre de cas possible où cet événement se réalise et le nombre de cas favorable est le nombre total de possibilités.

Dans l'exemple de l'exercice 5.3 page 24, l'expérience consiste à faire un "grand" nombre de tirages (avec remise) parmis les 6 valeurs posssibles $\{1, 2, 3, 4, 5, 6\}$. La probabilité d'obtenir chacune des 6 valeurs est égale à $1/6$, limite des différentes proportions de fois où chacune des ces valeurs apparaît quand le nombre de tirage est grand.

EXERCICE 5.14. Le livre *l'homme-dé* de Luke Rhinehart décrit l'expérience suivante : Je ramassais le dé en déclarant : "Si c'est un, trois ou cinq, je vais me coucher ; si c'est deux je descends demander à Jake la permission d'essayer de revoire Arlène ; si c'est quatre ou six, je veille pour continuer à réfléchir à tout ça." Quels sont les résultats possibles ? À quels résultats correspond l'événement : veiller un peu pour continuer à réfléchir à tout ça ? Peut-on prévoir sur un jet de dé si cet événement se réalisera ? Que se passe-t-il à votre avis si on jette le dé un millier de fois ?

Voir éléments de correction page 53.

5.4. Notion de variable aléatoire (discrète)

DÉFINITION 5.15. On parle de *variable aléatoire* lorsqu'on associe à des résultats d'expérience aléatoire des valeurs x (numériques ou non). Dans le cas où l'ensemble des valeurs possibles est fini, on parle de *variable aléatoire discrète*. La liste des probabilités associées aux résultats possibles d'une variable aléatoire est appelée *loi de probabilités*. On note $P(X = x)$ la probabilité qu'une variable aléatoire X prenne la valeur x.

EXEMPLE 5.16. Dans le cas de l'exercice 5.3 qui simule un tirage au dès, l'ensemble des valeurs possible pour un jeté de dès est $\{1, 2, 3, 4, 5, 6\}$. X est ici la variable aléatoire égale au numéro de la face obtenue et on a

$$P\big(X = 1\big) = P\big(X = 2\big) = \ldots = P\big(X = 6\big) = \frac{1}{6}. \tag{5.1}$$

EXEMPLE 5.17. Le système de la draft NBA consiste chaque année pour les différentes franchises à recruter les jeunes joueurs les plus talentueux, provenant soit des équipes universitaires américaines soit, ce qui n'est plus rare maintenant, des meilleures équipes européennes. Afin d'équilibrer la compétition, il a été décidé par les instances de la NBA de laisser aux franchises les plus mal classées l'année précédente la priorité du choix.

Pour déterminer l'équipe s'octroyant le premier tour de draft, on affecte le numéro 1 à la mieux classée[1] jusqu'à 11 pour la dernière. Plusieurs systèmes aléatoires se sont succédés, nous n'en considérerons ici que deux :

- Le premier (datant de 1985) repose sur un simple tirage au sort des onze numéros. Il s'agit donc d'un système équiprobable ($\frac{1}{11}$ pour chaque équipe).
- Le second (1990) utilisé pour avantager les plus faibles équipes, consiste à mettre soixante-six balles de ping-pong dans une urne, dont onze portent le numéro 11, dix le numéro 10,..., et 1 le numéro 1 .En ce qui concerne le tirage de la première balle, les probabilités sont donc de $\frac{1}{1+2+...11} = \frac{1}{66}$ pour le numéro 1, $\frac{2}{66}$ pour le numéro 2,..., et $\frac{11}{66}$ pour le numéro 11.

DÉFINITION 5.18. L'*espérance mathématique* d'une variable aléatoire X est la somme des résultats possibles pondérés par leurs probabilités. On la note généralement $\mathbb{E}(X)$.

Si l'ensemble des valeurs prise par la variable aléatoire X est noté $\{n_1, n_2, \ldots, n_q\}$, alors

$$\mathbb{E}(X) = P(X = n_1)n_1 + P(X = n_2)n_2 + \ldots + P(X = n_q)n_q = \sum_{i=1}^{q} P(X = n_i)n_i \tag{5.2}$$

EXEMPLE 5.19. Dans le cas de l'exemple 5.16, nous avons d'après (5.1)

$$\mathbb{E}(X) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) \tag{5.3}$$

On peut calculer cela en se rappelant que

$$1 + 2 + 3 + 4 + 5 + 6 = \frac{6 \times 7}{2} = 21$$

et donc

$$\mathbb{E}(X) = 3.5 \tag{5.4}$$

On peut aussi le calculer avec R en tapant

```
sum(1:6)/6
```

EXEMPLE 5.20. Dans le cas de l'exemple 5.17 on trouve pour le premier système

$$\mathbb{E}(X) = \frac{1}{11}(1 + 2 + \ldots + 11) = 6,$$

et pour le second

$$\mathbb{E}(X) = 1 \times \frac{1}{66} + 2 \times \frac{2}{66} + \ldots + 11 \times \frac{11}{66} = \frac{1}{66}\left(1 + 2^2 + \ldots + 11^2\right) = \frac{11 \times (11 + 1) \times (2 \times 11 + 1)}{6 \times 66} = \frac{3036}{396} \approx 7.666666$$

ce qui peut aussi s'obtenir avec R :

```
sum((1:11)^2)/66
```

[1] De celles ne participant pas aux play-offs, les plus fortes sont donc d'office écartées du premier choix.

Définition 5.21. La *variance* d'une variable aléatoire est la somme des carrés des écarts entre les résultats possibles et l'espérance mathématique, pondérée par les probabilités respectives. On note généralement la variance σ^2. Si l'ensemble des valeurs prise par la variable aléatoire X est noté $\{n_1, n_2, \ldots, n_q\}$, alors

$$\sigma^2 = P(X = n_1)(n_1 - \mathbb{E}(X))^2 + \ldots + P(X = n_q)(n_q - \mathbb{E}(X))^2 = \sum_{i=1}^{q} P(X = n_i)(n_i - \mathbb{E}(X))^2 \tag{5.5}$$

Définition 5.22. L'*écart-type* d'une variable aléatoire est la racine carrée de sa variance. On le note σ.

Exemple 5.23. Dans le cas de l'exemple 5.16, nous avons grâce à (5.4)

$$\sigma = \sqrt{\sum_{i=1}^{6} \frac{1}{6}(i - 3.5)^2}$$

et grâce à R en tapant

```
sqrt((sum((1:6 - 3.5)^2))/6)
```

on obtient

$$\sigma \approx 1.70783 \tag{5.6}$$

Exemple 5.24 (facultatif). Dans le cas de l'exemple 5.17 on trouve grâce à R pour le premier système

```
sqrt((sum((1:11 - 6)^2))/11)
```

```
[1] 3.162278
```

et donc $\sigma \approx 3.162$. Pour le second, on a

```
sqrt((sum((1:11 - 7.666667)^2 * (1:11)))/66)
```

```
[1] 2.687419
```

et donc $\sigma \approx 2.687$.

On peut constater qu'il existe un lien entre la moyenne et l'écart-type d'une variable aléatoire et la moyenne et l'écart-type (au sens déjà vu dans le chapitre 3, voir équations 3.1 page 9 et 3.3 page 10) des valeurs prises par cette variable aléatoire au cours d'une "longue" expérience :

Proposition 5.25. *On réalise l'expérience aléatoire suivante : on effectue N tirages aléatoires d'une variable aléatoire X et on note $(y_i)_{1\leq i\leq N}$ les valeurs obtenues. Alors, si N est grand, la moyenne des valeurs $(y_i)_{1\leq i\leq N}$ (au sens de 3.1 page 9) est approximativement égale à l'espérance $\mathbb{E}(X)$ de la variable aléatoire X et l'écart-type des valeurs $(y_i)_{1\leq i\leq N}$ (au sens de 3.3 page 10) est approximativement égale à l'écart-type de la variable aléatoire σ, soit encore*

$$\frac{1}{N}\sum_{i=1}^{N} y_i \approx \mathbb{E}(X), \tag{5.7a}$$

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N} (\overline{y} - y_i)^2} \approx \sigma \tag{5.7b}$$

Démonstration facultative. Voir l'annexe E. □

Cela justifie la notion d'espérance mathématique, qui correspond la moyenne "espérée" des différentes valeurs aléatoires.

Exercice 5.26 (facultatif). On pourra vérifier cela grâce à la commande suivante (sur l'exemple 5.16)

```
mean(sample(1:6, size = 10000, replace = T))
```

```
[1] 3.501
```

qui doit fournir une valeur proche de (5.4).

EXERCICE 5.27 (facultatif). On pourra vérifier cela grâce à la commande suivante (sur l'exemple 5.16)

```
sd(sample(1:6, size = 10000, replace = T))
```

```
[1] 1.714729
```

qui doit fournir une valeur proche de (5.6).

EXERCICE 5.28. Comme dans l'annexe D page 133, chargez la fonction `desexpetheor.R`, qui dans le cas de l'exemple 5.16 page 27 renvoie les différences entre moyenne et écart-type et moyenne et écart-type mathématiques (donné par (5.4) et (5.6)). Constater en tapant par exemple que ces différences s'amenuisent quand n est "grand" en tapant par exemple

```
desexpetheor(100)
```

```
desexpetheor(1e+06)
```

5.5. Le modèle binomial

5.5.1. Le modèle probabiliste binomial

DÉFINITION 5.29. Le modèle probabiliste *binomial* correspond à des circonstances où
- il n'y a que deux résultats possibles dans une unique expérience aléatoire : pile/face, correct/défectueux, mort/vivant ; l'un est considéré comme un succès et l'autre comme un échec[2],
- les répétitions de cette expérience se réalisent indépendamment les unes des autres et
- la probabilité de succès reste la même à chaque répétition.

Le modèle probabiliste binomial décrit le comportement de la variable aléatoire "nombre de succès apparus" sur l'ensemble des répétitions.

La famille binomiale est engendrée par deux *paramètres* : le nombre de répétitions et la probabilité de succès, notés respectivement n et π.

EXERCICE 5.30. Quels sont les paramètres correspondant au nombre de "pile" dans un lancer de pièce cinq fois de suite ? Quels sont les paramètres correspondant à la roulette russe ?

Voir éléments de correction page 53.

EXERCICE 5.31.

(1) Comme dans l'annexe D page 133, récupérer et sourcer la fonction R `binom.simple.R`.

(2) Utiliser cette fonction en tapant par exemple

```
binom.simple(50, 0.7)
```

```
   E    S
0.36 0.64
```

où le premier paramètre correspond à n et le second à π. Elle renvoie les proportions de succès (S) et d'échec (E) pour un tirage de taille n.

(3) Faire des simulations en prenant par exemple les paramètres de l'exercice 5.30 et constater que quand n est "grand", ces proportions se rapproche des probabilités théoriques $1-\pi$ et π en tapant par exemple

```
binom.simple(1e+05, 5/6)
```

```
      E       S
0.16678 0.83322
```

[2] Sans qu'il y ait pour autant jugement de valeur...

(4) Cette fonction a un argument optionnel `affiche` égal à `FALSE` par défaut et qui affiche l'échantillon obtenu. Attenttion à ne pas l'afficher pour n trop grand ! Exemples :

```
 binom.simple(50, 0.7, affiche = F)
   E    S
0.12 0.88
```

ou

```
 binom.simple(50, 0.7, affiche = T)
[1] "échantillon trouvé"
 [1] S S S S S S E S S S S S E S E S S E S S E S E E S E S S S S E S E S E E
[37] E E S E S E S S S S S S S S
Levels: E S
   E    S
0.32 0.68
```

5.5.2. La loi de probabilités binomiale

PROPOSITION 5.32. *Pour n répétitions avec une probabilité de succès π, la probabilité qu'une variable aléatoire binomiale*[3] *X soit égale à $x \in \{0, \dots, n\}$ est*

$$P(X = x) = C_n^x \pi^x (1 - \pi)^{n-x} \tag{5.8}$$

avec C_n^x qui est appelé coefficient binomial.

REMARQUE 5.33 (facultative). Dans la proposition 5.32, on a

$$C_n^x = \frac{n!}{x!(n-x)!}.$$

PREUVE FACULTATIVE. La proposition 5.32 peut se "montrer" (en simplifiant) de la façon suivante : pour obtenir x succès et $n - x$ échecs dans un ordre donné, la probabilité est de $\pi^x(1 - \pi)^{n-x}$ (on multiplie les diférentes probabilités de chacun des événements, indépendants). On multiplie ce résultas par C_n^x qui correspond au nombre total de combinaisons de x éléments parmi n, puisqu'ici l'ordre des succès ne compte pas. □

MANIPULATION AVEC RCMDR 5.34. On peut calculer cette loi de probabilité en utilisant le menu déroulant "Distributions" , l'option "Distributions discrètes" puis "Distribution binomiale" puis "Probabilités binomiales". Dans la fenêtre de dialogue, il faut alors préciser le nombre d'essais (n) et la probabilité de succès.

Ainsi pour $n = 5$ essais et une probabilité de succès de $\pi = 0.3$, on obtient

```
      Pr
0 0.16807
1 0.36015
2 0.30870
3 0.13230
4 0.02835
5 0.00243
```

[3] ce nom vient de la formule du binôme de Newton :

$$(a + b)^n = \sum_{k=0}^{n} C_n^k a^k b^{n-k}.$$

Si on écrit l'équation proche de l'equation (F.1) page 139, alors on a un moyen "mnémotechnique" de se rappeler (5.8) :

$$g(X) = (X\pi + 1 - \pi)^n = \sum_{k=0}^{n} C_n^k X^k \pi^k (1 - \pi)^{n-k} = \sum_{k=0}^{n} P(X = k) X^k$$

De plus, pour $X = 1$, cela nous montre que la somme des probabilité est bien 1.

et donc en particulier $P(X = 2) = 0.3087$.

MANIPULATION AVEC RCMDR 5.35. On peut tracer un graphe de la loi de probabilité correspondante avec l'option "Graphe de la distribution binomiale". Sur ce graphe, pour chaque x dans $0, \ldots, n$, on voit un segment d'abscisse x et de hauteur $P(X = x)$.

EXERCICE 5.36.

(1) Comme dans l'annexe D page 133, récupérer et sourcer la fonction R `binom.complet.R`.
Cette fonction simule le modèle binomial, pour un nombre de tirage égal à p, avec des paramètres (n, π). Elle affiche le graphe de la loi de probabilité théorique ainsi que celui des proportions observées.

(2) Constater en reprenant les paramètres suivants $\pi = 0.3$ et $n = 5$ et p de plus en plus grand que les deux graphes sont similaires en tapant par exemple :
```
binom.complet(5, 0.3, p = 1000)
```

(3) On pourra aussi afficher les différences entre les probabilités théoriques et les proportions observées en tapant
```
binom.complet(5, 0.3, p = 1000, difference = T)
```

Voir éléments de correction page 53.

EXERCICE 5.37. Observer comment évolue le graphe de la distribution binomiale évolue lorsque vous modifiez la probabilité de succès en $\pi = 0$, $\pi = 0.4$, $\pi = 0.5$, $\pi = 0.7$, $\pi = 0.9$, $\pi = 0.95$ et $\pi = 1$ en conservant le paramètre $n = 5$.

Voir éléments de correction page 54.

EXERCICE 5.38. Représenter graphiquement les lois de probabilités correspondant aux jeux de paramètres suivants : $(n = 10, \pi = 0.1)$, $(n = 10, \pi = 0.25)$, $(n = 10, \pi = 0.5)$, $(n = 10, \pi = 0.85)$.

Voir éléments de correction page 54.

5.5.3. Espérance mathématique et variance binomiales

PROPOSITION 5.39. *L'espérance mathématique d'une variable aléatoire binomiale de paramètres n et π est égale à $n\pi$, sa variance à $n\pi(1-\pi)$ et son écart-type à $\sqrt{n\pi(1-\pi)}$.*

PREUVE FACULTATIVE. Voir la preuve de ce résultats en annexe F. □

PREUVE PARTIELLE "AVEC LES MAINS". On peut remarquer formellement et facilement que l'espérance mathématique d'une variable aléatoire binomiale de paramètres n et π est égale à $n\pi$.

En effet, pour un essai, la proportion de succés est égale à π ; pour n essais, la proportion de succès est égale au nombre moyen de succès (soit $\mathbb{E}(X)$) divisée par le nombre d'essais (n) ; bref, $\pi = \mathbb{E}(X)/n$, d'où le résultat ! □

EXERCICE 5.40. De façon analogue à l'exercice 5.28, chargez la fonction `binom.theorie.exper.R`, qui dans le cas de la proposition 5.39 renvoie les différences entre la moyenne $n\pi$ et l'écart-type $\sqrt{n\pi(1-\pi)}$ et moyenne et écart-type du tirage aléatoire effectué. Constater en tapant par exemple que ces différences s'amenuisent quand n est "grand" en tapant par exemple

```
binom.theorie.exper(4, 0.4, p = 100)
```

```
$ecm
[1] 0.1

$ecec
[1] 0.02726209
```

```
 binom.theorie.exper(4, 0.4, p = 1e+06)
$ecm
[1] 0.001587

$ecec
[1] 0.01979431
```

MANIPULATION AVEC RCMDR 5.41. Afin de calculer l'espérance, la variance et l'écart-type pour une distribution binomiale comportant $n = 5$ essais et une probabilité de succès de $\pi = 0.3$ il faut écrire dans la fenêtre de script les trois commandes suivantes et les soumettre successivement :

```
5*0.3
5*0.3*(1-0.3)
sqrt(5*0.3*(1-0.3))
```

On obtient donc

```
[1] 1.5
[1] 1.05
[1] 1.024695
```

c'est-à-dire

$$\mathbb{E}(X) = 1.5, \quad \sigma = 1.024695 \tag{5.9}$$

EXERCICE 5.42. Trouver l'espérance mathématique et l'écart-type pour les distributions binomiales suivantes :

- $n = 20$ et $\pi = 0{,}50$;
- $n = 40$ et $\pi = 0{,}20$;
- $n = 200$ et $\pi = 0{,}80$.

Voir éléments de correction page 54.

5.5.4. Les probabilités cumulés

Il est souvent utile de calculer des probabilités binomiales cumulées, par exemple la probabilité d'observer 2 succès ou moins ou bien la probabilité d'observer plus de 4 succès.

DÉFINITION 5.43. La *probablité cumulée* jusqu'au *quantile* $i \in \{0, ..., n\}$ correspond à la somme

$$P(X = 0) + P(X = 1) + \ldots + P(X = i)$$

On notera cette somme

$$P(X \leq i)$$

MANIPULATION AVEC RCMDR 5.44. Pour ce faire, il est nécessaire d'employer le menu déroulant "Distributions" , l'option "Distributions discrètes" puis "Distribution binomiale" puis l'option "Probabilités binomiales cumulées" (en prenant par défaut l'aire à gauche)

De façon plus générale,

DÉFINITION 5.45. Si l'ensemble des valeurs prise par la variable aléatoire X est noté $\{n_1, n_2, \ldots, n_q\}$ (*ici, dans un ordre croissant*), alors la *probablité cumulée* jusqu'au *quantile* n_i pour $i \in \{1, ..., q\}$ correspond à la somme

$$P(X = n_i) + P(X = n_{i+1}) + \ldots + P(X = n_q) \tag{5.10}$$

On notera cette somme

$$P(X \leq n_i) \tag{5.11}$$

On note de même

$$P(X > n_i) = P(X \geq n_{i+1}) = P(X = n_{i+1}) + P(X = n_{i+2}) + \ldots + P(X = n_q) \tag{5.12}$$

REMARQUE 5.46. Notons aussi que, de façon plus générale, si on se donne la probabilité p, le quantile est le nombre q tel que

$$P(X \leq q) = p.$$

On choisira parfois la définition (équivalente ici) : si on se donne la probabilité p, le quantile est la plus petite valeur x telle que

$$P(X \leq x) \geq p.$$

REMARQUE 5.47. On a naturellement

$$P(X \leq n_i) + P(X > n_i) = 1 \tag{5.13}$$

MANIPULATION AVEC RCMDR 5.48. Avec R, dans l'option "Probabilités binomiales cumulées", la somme (5.11) est appelée "aire à gauche" tandis que la somme (5.12) est appelée "aire à droite".

Plus précisémment, on utilise le menu déroulant "Distributions" , l'option "Distributions discrètes" puis "Distribution binomiale" puis l'option "Probabilités binomiales cumulées", puis en choisissant "aire à gauche" ou "aire à droite".

On a aussi le résultat suivant

LEMME 5.49. *Si pour $i < j$, on note*

$$P(n_i \leq X \leq n_j) = P(X = n_i) + P(X = n_{i+1}) + \ldots + P(X = n_j), \tag{5.14}$$

alors

$$P(n_i \leq X \leq n_j) = P(X \leq n_j) - P(X \leq n_{i-1}). \tag{5.15}$$

EXERCICE 5.50. Pour une variable aléatoire binomiale X de paramètres $n = 7$ et $\pi = 0.2$

(1) Calculer à l'aide du logiciel R : $P(X = 2)$, $P(X = 0)$, $P(X = 9)$, $P(X \leq 5)$, $P(X \geq 5)$, $P(X > 5)$, et $P(2 \leq X \leq 5)$

(2) Vérifier que la probabilité cumulée $P(X \leq 5)$ est bien égale à

$$P(X \leq 5) = P(x = 0) + \ldots + P(x = 5)$$

(3) Vérifier que la probabilité cumulée $P(X > 5)$ est bien égale à

$$P(X > 5) = P(X = 6) + P(X = 7)$$

(4) Représenter par un graphique en bâtons sa loi de probabilités cumulées.

Voir éléments de correction page 55.

EXERCICE 5.51 (facultatif).

On pourra utiliser les fonctions `dbinom` et `dbinom` qui fournissent dans "Rgui", ce qu'on appelle respectivement les densités et la fonction de distribution pour la loi binomiale ; plus précisement, si $X = (x_1, \ldots, x_n)$ est un vecteur de valeurs (dites quantiles) , alors

- la commande
  ```
  dbinom(X, size = n, prob = Pi)
  ```
  fournit le vecteur des probabilités $\big(P(X = x_1), \ldots, P(X = x_n)\big)$,
- la commande
  ```
  pbinom(X, size = n, prob = Pi)
  ```
  fournit le vecteur des probabilités $\big(P(X \leq x_1), \ldots, P(X \leq x_n)\big)$,
- et la commande

`pbinom(X, size = n, prob = Pi, lower.tail = FALSE)`
fournit le vecteur des probabilités $\big(P(X > x_1), \ldots, P(X > x_n)\big)$.
Voir éléments de correction page 56.

EXERCICE 5.52. La probabilité de contact des sondés au téléphone est généralement estimée à 60 %. Nous décidons de lancer une vague d'appels de $n = 200$ personnes. On considérera que le nombre de personnes que l'on va parvenir à contacter suit une loi binomiale de paramètres $n = 200$ et $\pi = 0.6$. En effet, dans ce cas, on répète $n = 200$ fois l'expérience "appeler quelqu'un au téléphone" qui a deux issues : le succès est la prise de contact de probalité $\pi = 0.6$ et l'échec est la non prise de contact.

(1) Quel nombre moyen de personnes peut-on espérer toucher dans cette première vague d'appels ?
(2) Quelle est la probabilité de contacter au moins 120 personnes ?
(3) Quelle est la probabilité de contacter au moins $n_1 = 150$ personnes ?
(4) Combien de personnes faudrait-il appeler pour espérer, en moyenne, contacter 150 personnes ?

Voir éléments de correction page 57.

5.6. Modèle probabiliste binomial et distribution d'échantillonnage d'une proportion

Il existe deux conceptions majeures d'une population en statistique. Dans la première, la population (bien que grande) est finie et clairement délimitée. Une véritable échantillonnage aléatoire est pratiqué pour connaître la population. C'est le domaine des *sondages*.

Dans la seconde conception, qui est celle exposée dans ce cours, un modèle probabiliste va tenir lieu de population idéale (de taille infinie) et on considérera que les données recueillies ne sont qu'un échantillon aléatoire provenant de ce modèle.

DÉFINITION 5.53. Un *paramètre* est un nombre qui décrit la population modèle. Il est constant (et inconnu en inférence statistique).

Une *statistique* est un calcul mené sur un échantillon. Sa valeur se modifie d'un échantillon à l'autre, elle est donc une variable aléatoire. Elle servira à estimer les paramètres de la population. De façon concrète, on mesurera ou calculera la statistique sur l'échantillon disponible (par exemple calculer une proportion de smarties rouges dans une boîte donnée).

Dans cette section, on connaît le paramètre π et on détermine, grâce à des échantillons aléatoirement produits, des statistiques pr. Nous allons comprendre comment ces statistisques sont en moyenne déterminables par rapport au paramètre connu π. Dans la section 5.7, nous ferons le contraire : nous mesurerons des statistiques pr et, grâce au travail que nous allons faire maintenant, nous allons en déduire des estimations du paramètre π,inconnu.

Une fois le modèle posé, la théorie des probabilités permet d'étudier les propriétés des échantillons qui en sont issus. Voyons ce qui se passe avec un modèle binomial. La véritable valeur du paramètre sera fixée arbitrairement pour les besoins de cette simulation à $\pi = 0.3$. Générons grâce au logiciel R un échantillon de taille $n = 1000$ et calculons la proportion p correspondant au nombre de succès y obtenus.

MANIPULATION AVEC RCMDR 5.54. Pour effectuer cette simulation, on utilise le menu déroulant "Distributions", l'option "Distributions discrètes" puis "Distribution binomiale", puis "Echantillon d'une distribution binomiale". Dans la fenêtre dialogue, il faut indiquer pour
– "Nombre d'essais" (valeur de n) : 1000,
– "Probabilité de succès" (valeur de π) : 0.3,
– "Nombre d'échantillons" : 1,
– "Nombre d'observations" : 1

Un jeu de données est alors créé qui s'appelle par défaut "EchantillonsBinomiaux". Si on cherche à le visualiser, on obtient (pour une évaluation ; cela changera à chaque fois)

```
        obs1
sample1  305
```

Il se trouve que la valeur est pour ce tirage de $y = 305$ ce qui conduit à une proportion observée (de succès)

$$p = \frac{y}{n} = \frac{305}{1000} = 0.305 \tag{5.16}$$

Un autre tirage conduira (probablement) à une autre valeur, par exemple à

$$p = \frac{y}{n} = \frac{331}{1000} = 0.331 \tag{5.17}$$

ou à

$$p = \frac{y}{n} = \frac{298}{1000} = 0.298$$

Ici, *pr* est la statistique (opposé au paramètre π). Il est essentiel de comprendre que *la valeur de la statistique est différente d'un échantillon à l'autre*, c'est ce qu'on appelle les *variations d'échantillonnage.*

Que se passe-t-il lorsque l'on tire un grand nombre d'échantillons dans les mêmes conditions ? On peut étudier la distribution de la statistique observée, on parle alors de *distribution d'échantillonnage.*

Manipulation avec Rcmdr 5.55. Tirons 15 échantillons binomiaux de taille $n = 1000$ et de paramètres $\pi = 0.3$ en faisant comme ci-dessus avec

- "Nombre d'essais" (valeur de n) : 1000,
- "Probabilité de succès" (valeur de π) : 0.3,
- "Nombre d'échantillons" : 15,
- "Nombre d'observations" : 1

Un nouveau jeu de données est alors créé (qui écrase l'ancien) Si on cherche à le visualiser, on obtient (pour une évaluation ; cela changera à chaque fois)

```
         obs1
sample1   294
sample2   286
sample3   299
sample4   296
sample5   287
sample6   309
sample7   296
sample8   305
sample9   319
sample10  280
sample11  317
sample12  316
sample13  306
sample14  312
sample15  304
```

Manipulation avec Rcmdr 5.56. Prenons maintenant

- "Nombre d'essais" (valeur de n) : 1000,
- "Probabilité de succès" (valeur de π) : 0.3,
- "Nombre d'échantillons" : 10000,
- "Nombre d'observations" : 1

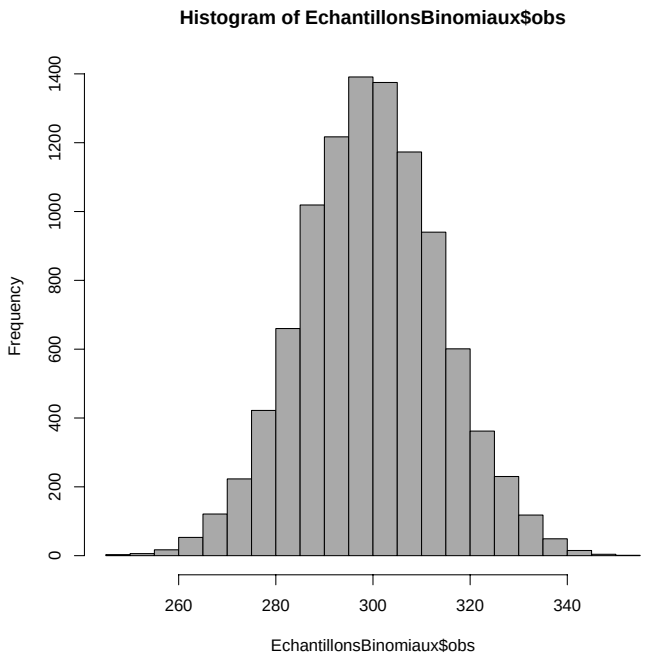


FIG. 5.3. Histogramme de la distribution d'échantillonnage (10000 tirages) des succès observés sur une loi binomiale de paramètre $n = 1000$ et $\pi = 0.3$

On peut ensuite et réaliser un histogramme des succès apparus (cf figure 5.3).

Si on fait un histogramme avec cette fois-ci la densité (voir remaque 3.2 page 8) en prenant 25 classes alors on observe l'histogramme de la figure 5.4 page suivante. Cet histogramme est alors très proche (avec un grossissement) de la figure 5.10 page 55. Voir l'exercice 5.36 page 32.

On pourra créer des échantillons aléatoire binomiaux en utilisant la commande `rbinom`. ♢

Désormais, on s'intéresse non plus à la variable aléatoire X égale au nombre de succès, mais à la variable aléatoire pr égale à la proportion de succès X/n.

Si on refait l'histogramme correspondant à celui de la figure 5.4 page suivante mais avec les proportions observées (donc en introduisant une nouvelle variable obtenue en divisant `EchantillonsBinomiaux` par $n = 1000$: voir pour cela la technique vue dans l'exercice 3.9 page 11) on obtient la figure 5.5. La figure 5.5 montre clairement

(1) que la distribution d'échantillonnage du nombre la proportion observée *pr* est centrée autour de la véritable valeur du paramètre $\pi = 0.3$. En moyenne la proportion d'un échantillon redonne la probabilité de succès.

(2) que l'histogramme a l'allure d'une densité normale (c'est-à-dire, "en cloche").

La théorie des probabilités prouve facilement cette observation. En effet, partant de l'expression donnée dans la proposition 5.39 page 32, si on divise X par n, chacune des valeurs prises par $pr = X/n$ est divisée par n et la loi de probabilité est conservée. On divise donc, par linéarité, l'espérance mathématique par n et l'écart-type par n. On a donc

PROPOSITION 5.57. *L'espérance mathématique de la variable aléatoire pr correspondant à la proportion de succès d'une variable aléatoire binomiale de paramètres n et π est égale à π et son écart-type à $\sqrt{\dfrac{\pi(1-\pi)}{n}}$.*

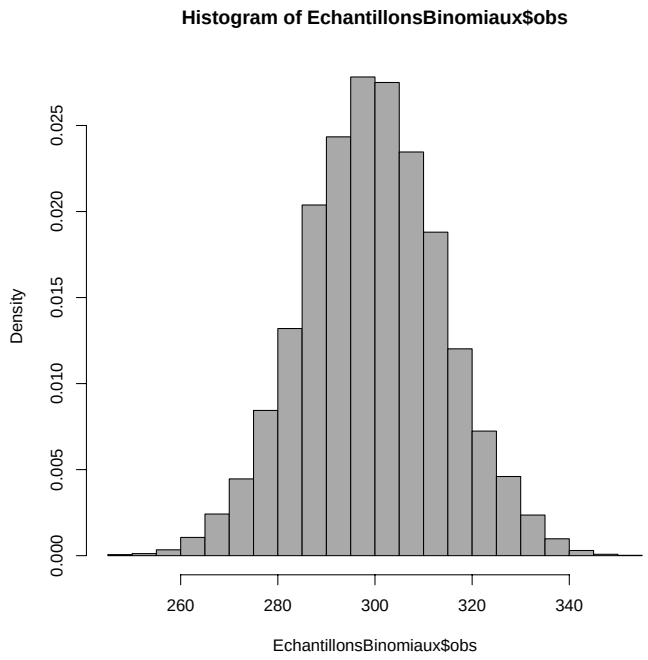


FIG. 5.4. Histogramme de la distribution d'échantillonnage (10000 tirages) des succès observés sur une loi binomiale de paramètre $n = 1000$ et $\pi = 0.3$ en densité avec 25 classes

Si on utilise la proposition 5.25 page 29, on a finalement, en notant $(pr_i)_{1\leq i\leq N}$ les N valeurs des proportions *observées* dans l'échantilonnage de taille n les résultats suivants : Si N est grand, la moyenne des valeurs $(y_i)_{1\leq i\leq N}$ (au sens de 3.1 page 9) est approximativement égale à l'espérance π et et l'écart-type des valeurs $(y_i)_{1\leq i\leq N}$ (au sens de 3.3 page 10) est approximativement égale à l'écart-type $\sqrt{\dfrac{\pi(1-\pi)}{n}}$.

Par exemple, ici pour l'échantillon défini par
- "Nombre d'essais" (valeur de n) : 1000,
- "Probabilité de succès" (valeur de π) : 0.3,
- "Nombre d'échantillons" : 10000,
- "Nombre d'observations" : 1

on obtient des des écart-types et moyennes expérimentaux *pour nos valeurs, différentes des votres !*

$$0.300102,$$
$$0.014326.$$

proches des écart-types et espérances théoriques :

$$\mathbb{E}(pr) = \pi = 0.3, \tag{5.18a}$$

$$\sigma = \sqrt{\frac{\pi(1-\pi)}{n}} = 0.014491, \tag{5.18b}$$

Bien entendu, ces résultats ne sont que des approximations, qui sont de plus en plus précise quand N est grand.

La valeur $\sqrt{\dfrac{\pi(1-\pi)}{n}}$ est appelée *erreur standard de la proportion* :

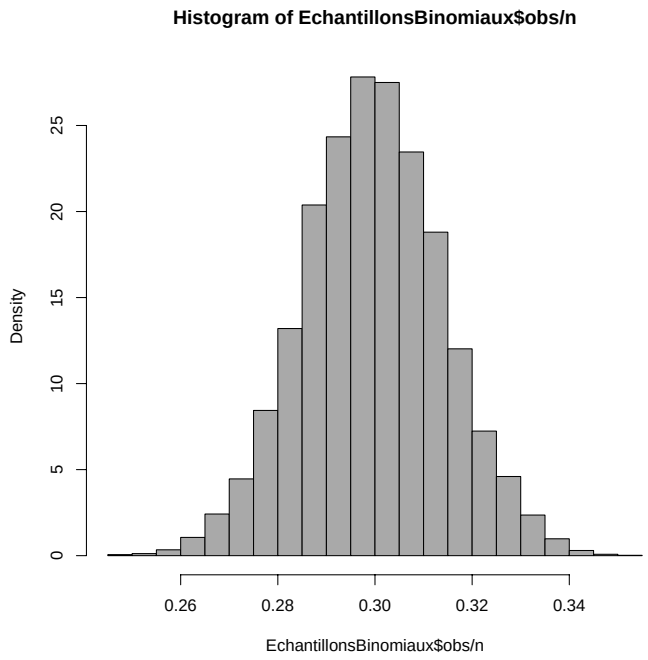


FIG. 5.5. Histogramme de la distribution d'échantillonnage (10000 tirages) d'une proportion observée sur une loi binomiale de paramètre $n = 1000$ et $\pi = 0.3$ en densité avec 25 classes

DÉFINITION 5.58. On appelle erreur standart de poportion la quantité :

$$SEP = \sqrt{\frac{\pi(1-\pi)}{n}} \tag{5.19}$$

On la note SEP à cause de son nom anglais (Standard Error of a Proportion).

On peut remarquer au dénominateur de la formule de cette erreur standard que la taille de l'échantillon n apparaît.

Sur la figure 5.6 page suivante, on a de nouveau tracé l'histogramme des proportion et la loi normale (on reviendra longuement sur cette courbe dans le chapitre 6) correpondant à la moyenne "théorique" $\pi = 0.3$ et l'écart-type théorique $\sigma = 0.014491$. La loi normale est une "courbe en cloche" totalement symétrique dont le sommet est d'abcisse la moyenne π et dont "l'étalement" est caractérisé par σ. On note

$$p_{\min} = \pi - 2SEP, \tag{5.20a}$$

$$p_{\max} = \pi + 2SEP \tag{5.20b}$$

Numériquement, on a ici :

$$p_{\min} = 0.271017, \tag{5.21a}$$

$$p_{\max} = 0.328983 \tag{5.21b}$$

On constate sur la courbe 5.6 page suivante que l'histogramme est proche de la loi normale (cela est justifié *a posteriori* par un calcul difficile de probabilité quand N et n sont "grands"). De plus, le coefficient 2 de l'équation (5.20) est justifié par le fait que 95% de "l'aire totale sous la cloche" se trouve entre les abscisses $p_{\min}$ et $p_{\max}$. Autrement dit, 95% des données de l'histogrammes des proportions se trouvent dans l'intervalle $[p_{\min}, p_{\min}]$. Justifions cela expérimentalement en tapant la séquence suivante dans Rgui qui dénombre le nombre de proportions dans cet intervalle et qui en calcule le pourcentage :

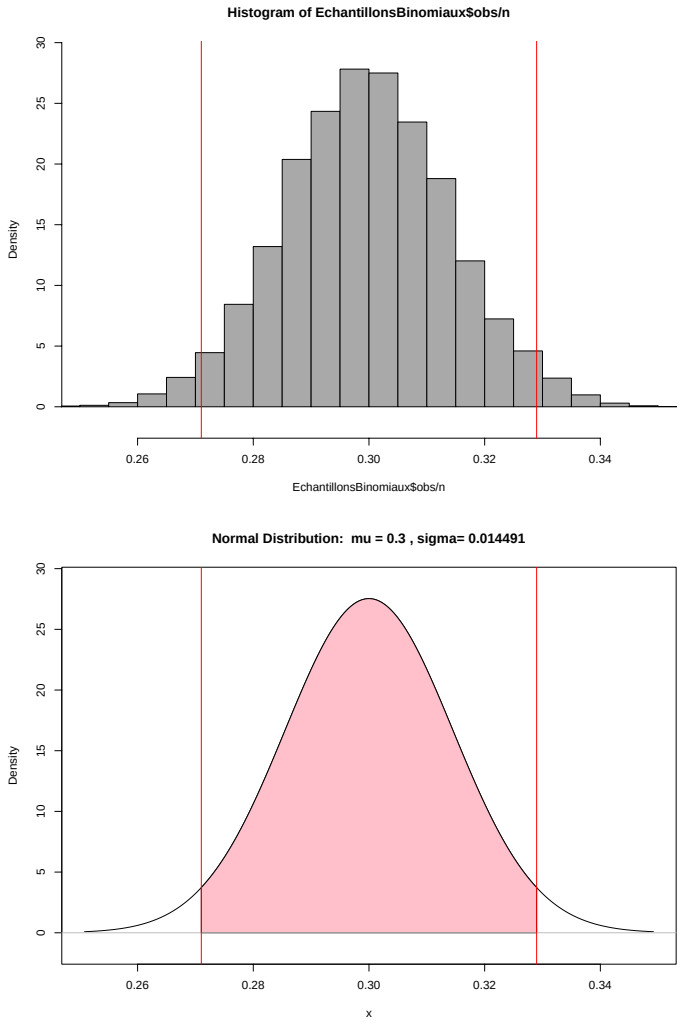


FIG. 5.6. Histogramme de la distribution d'échantillonnage (10000 tirages) d'une proportion et le graphique de la loi normale de moyenne $m = 0.3$ et d'écart-type $\sigma = 0.014491$ avec les deux droites correspondant aux abscisses 0.271017 et 0.328983

```
sum((EchantillonsBinomiaux$obs/1000>=0.2710172)
&(EchantillonsBinomiaux$obs/1000<=0.3289828))
```

puis

```
100*sum((EchantillonsBinomiaux$obs/1000>=0.2710172)
&(EchantillonsBinomiaux$obs/1000<=0.3289828))/10000
```

Cela donne *pour mes valeurs* un pourcentage égal à 95.11%, qui est bien proche de 95 % ! Cela est corroboré sur le graphique du haut de la figure 5.6 : on a représenté les deux droites d'absxisses $p_{\min}$ et $p_{\max}$. L'aire de l'histogramme (en densité) entre ces deux valeurs représente aussi 95% de l'aire totale ! Bref, l'intervalle suivant

(défini à partir de (5.20)) :

$$\left[\mathbb{E}(pr) - 2\sqrt{\frac{\pi(1-\pi)}{n}}, \mathbb{E}(pr) + 2\sqrt{\frac{\pi(1-\pi)}{n}}\right] = \left[\pi - 2\sqrt{\frac{\pi(1-\pi)}{n}}, \pi + 2\sqrt{\frac{\pi(1-\pi)}{n}}\right]$$
$$= [\pi - 2SEP, \pi + 2SEP] = [0.271017, 0.328983]. \quad (5.22)$$

contient environ 95% des proportions tirées au hasard.

Remarque 5.59. Naturellement, le raisonnement fait avec le nombre 95 % souvent utilisé peut se faire avec n'importe quel pourcentage dans $]0, 100[$!

5.7. Intervalle de confiance "d'une proportion"

Dans le contexte probabiliste ci-dessus, nous connaissions π, le paramètre, et nous étudiions comment évoluait la statistique pr (voir définition 5.53 page 35). En réalité, c'est le contraire qui se passe : nous connaissons *une seule valeur* de pr et nous souhaitons *inférer* la valeur de π.

La théorie des probabilités et le raisonement précédent nous a appris qu'il y a 95% de chance que la statistique pr soit située à moins de deux écarts-types du paramètres π. *Ce qui revient à dire qu'il y a aussi* 95% *de chances que le paramètre se trouve à deux écarts-types de la statistique pr !*. Cela s'exprime par le raisonnement suivant :

$$pr \in [\pi - 2\sigma, \pi + 2\sigma] \iff |pr - \pi| \leq 2\sigma \iff \pi \in [pr - 2\sigma, pr + 2\sigma]$$

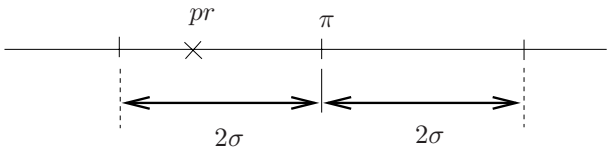


Fig. 5.7. Il y a au plus la distance 2σ entre π et pr.

Voir figure 5.7. Autrement dit, *il y a aussi* 95% *de chances que le paramètre* π *se trouve dans l'intervalle* $[pr - 2\sigma, pr + 2\sigma]$. On a vu que $\sigma = SEP = \sqrt{\pi(1-\pi)/n}$. Ainsi, l'intervalle de confiance donné par (5.22) contiendra pas la valeur recherchée π dans 95 % des cas ! *A contrario*, cet intervalle ne contiendra pas la valeur recherchée π dans 5 % des cas !

Il n'y a qu'un petit problème car on ne connaît pas la valeur de l'écart-type car elle dépend du paramètre (c'est $\sqrt{\frac{\pi(1-\pi)}{n}}$) ! On remplacera dans cette formule le paramètre π par la valeur observée de la statistique pr, et heureusement, on peut prouver que cette approximation marche plutôt bien lorsque l'échantillon atteint une taille raisonnable !

Définition 5.60 (Intervalle de confiance "d'une proportion" ($NC = 95$ %)). L'intervalle au niveau de confiance $NC = 95$ % pour une proportion dans un contexte binomiale de paramètre n et π est donné par les quantités $pr \pm 2\sqrt{\frac{pr(1-pr)}{n}}$ où pr est la proportion observée dans l'échantillon de taille n, qui définissent l'intervalle :

$$\left[pr - 2\sqrt{\frac{pr(1-pr)}{n}}, pr + 2\sqrt{\frac{pr(1-pr)}{n}}\right] \quad (5.23)$$

soit encore en introduisant l'erreur standart de proportion (SEP) (ici "expérimentale", contrairement à (5.19)) la quantité :

$$SEP = \sqrt{\frac{pr(1-pr)}{n}}, \quad (5.24)$$

l'intervalle de confiance est donné par

$$[pr - 2SEP, pr + 2SEP]$$

Avec les valeurs issues de la simulation de ma machine avec
- "Nombre d'essais" (valeur de n) : 1000,
- "Probabilité de succès" (valeur de π) : 0.3,
- "Nombre d'échantillons" : 1,
- "Nombre d'observations" : 1

et les valeurs données par (5.16), on a puisque

$$n = 1000,$$
$$y = 305,$$

$$pr = \frac{y}{n} = 0.305,$$
$$pr \pm 2\sqrt{\frac{pr(1-pr)}{n}} = 0.305 \pm 0.0291187.$$

et donc un intervalle de confiance à 95 %

$$[0.2758813, 0.3341187]. \tag{5.25}$$

La valeur du paramètre ($\pi = 0.3$) est bien dans ce cas compris dans l'intervalle de confiance. Il faut bien comprendre que le contraire ne se produit qu'une fois sur 20.

Parfois d'autres intervalles de confiance plus large, mais plus "sûrs" (car on ne remplace par π par pr) sont donnés, en remarquant que l'on peut majorer $\pi(1-\pi)$ par $1/4$:

$$\left[pr - \sqrt{\frac{1}{n}}, pr + \sqrt{\frac{1}{n}}\right] \tag{5.26}$$

♢

Si on reprend le calcul avec la valeur trouvée définie par (5.17), on a

$$n = 1000,$$
$$y = 331,$$

$$pr = \frac{y}{n} = 0.331,$$
$$pr \pm 2\sqrt{\frac{pr(1-pr)}{n}} = 0.331 \pm \times 0.0297617.$$

et donc un intervalle de confiance à 95 %

$$[0.3012383, 0.3607617].$$

On n'est dans les 5 % des cas non chanceux où l'intervalle de confiance ne contient pas le paramètre π ! Ici, la probabilité de sortie de la valeur 331 vaut 0.0028362 plus faible que la probabilité de sortie de la valeur 305, égale à 0.0258145 et qui a donné un intervalle contenant π. Si on faisait une étude complète des cas où on trouve un intervalle ne contenant pas π, on trouverait une probabilité expérimentale approchant 5% !

Manipulation avec R 5.61. Pour évaluer l'intervalle de confiance au niveau de confiance 0.95 avec R, on pourra taper en ligne de commande :

```
pr<-305/1000
pr-2*sqrt(pr*(1-pr)/1000)
```

puis

```
pr+2*sqrt(pr*(1-pr)/1000)
```

ce qui fournit bien

```
[1] 0.2758813
```

et

```
[1] 0.3341187
```

Mieux, on pourra taper directement

```
pr<-305/1000
pr+2*c(-1,1)*sqrt(pr*(1-pr)/1000)
```

ce qui fournit bien

```
[1] 0.2758813 0.3341187
```

En reprenant la remarque 5.59 page 41, on peut souhaiter modifier le niveau de confiance. La formule de l'intervalle devient alors :

DÉFINITION 5.62 (Intervalle de confiance "d'une proportion"). L'intervalle au niveau de confiance NC pour une proportion dans un contexte binomial est donné par les quantités $p \pm z\sqrt{\frac{p(1-p)}{n}}$ où p est la proportion observée dans l'échantillon de taille n et où z, le coefficient multiplicateur, est donné par la loi normale et pourra être obtenu sous

- soit allant dans le menu "distribution", puis "distribution continue", puis "distribution normale", puis "quantiles normaux" et rentrer à la place de "probabilités", `(1+NC)/2`, où NC est un nombre égal au niveau de confiance ;
- soit grâce à la ligne de commande
  `qnorm((1 + NC)/2)`
  où NC est un nombre égal au niveau de confiance.

À une probabilité continue X est associée une loi de probabilité donnée par $P(X \leq q)$. Ici q est le *quantile* et $P(X \leq q)$ est la *probabilité*. Voir remarque 5.46 page 34. Le coefficient z est défini par

$$P(-z \leq X \leq z) = NC$$

où ici la loi est la loi normale centrée réduite. On peut montrer, puisque la loi normale centrée réduite est paire que

$$P(-z \leq X \leq z) = 2P(X \leq z) - 1 \tag{5.27}$$

En effet, on écrit successivement

$$\begin{aligned} P(-z \leq X \leq 0) + P(0 \leq X \leq z) &= 2P(0 \leq X \leq z), \\ &= 2\left(P(-\infty \leq X \leq z) - P(-\infty \leq X \leq 0)\right), \\ &= 2\left(P(X \leq z) - \frac{1}{2}\right), \\ &= 2P(X \leq z) - 1. \end{aligned}$$

Il nous faut donc résoudre

$$P(X \leq z) = \frac{1 + NC}{2},$$

où z suit la loi normale centrée réduite. Autrement dit, on cherche le quantile associé à la probabilité $(NC + 1)/2$. Pour cela, on tapera donc la ligne de commande `qnorm((1+NC)/2)` qui fournit le quantile correspondant à la probabilité $(1 + NC)/2$ pour la loi normale centrée réduite. Ce point sera revu dans le chapitre suivant 6. ♢

Les différentes valeurs de z en fonction du niveau de confiance NC sont données dans le tableau 5.1 page suivante.

EXERCICE 5.63. Retrouver ces valeurs de z. Voir éléments de correction page 58

REMARQUE 5.64. On constate dans le tableau 5.1 page suivante que le z augmente avec niveau de confiance. En effet, la largeur de l'intervalle de confiance augmente avec le niveau de confiance. Pour avoir un seuil élevé et donc obtenir un intervalle de confiance qui contienne π dans un grand nombre de cas, on "ratisse plus large" !

NC	z	z approché
0	0	0
0.5	0.67449	0.7
0.6	0.841621	0.8
0.7	1.036433	1
0.8	1.281552	1.3
0.9	1.644854	1.6
0.95	1.959964	2
0.99	2.575829	2.6
0.999	3.290527	3.3
1	Inf	Inf

TAB. 5.1. Quelques valeurs de z en fonction du niveau de confiance NC.

Le cas extrême $NC = 0$ correspond à $z = 0$ et donc un intervalle de confiance pas du tout probable et le cas $NC = 1$, correspond au cas où toutes les valeurs sont possibles ! À NC constant, on constate aussi que l'on peut diminuer la largeur en prenant un n plus grand.

REMARQUE 5.65. Attention à ne pas mal interpréter la notion d'intervalle de confiance ! Au niveau de confiance NC, il ne faut pas dire que le paramètre inférer a $100\times NC$ % de chance d'être dans un intervalle de confiance donné ; L'intervalle de confiance doit être plutôt interprété en terme de probabilité ou en fréquence par rapport à un "grand" nombre d'intervalles de confiance calculé sur des échantillons similaires : $100\times NC$ % des intervalles de confiance contiendront la valeur réelle du paramètre inféré. Voir l'exercice 5.71 page 46 par exemple.

REMARQUE 5.66. Grâce aux définitions 5.58 et 5.62 et à la remarque 5.64, on peut mettre l'intervalle de confiance au niveau de confiance NC sous la forme

$$[pr - zSEP, pr + zSEP], \tag{5.28}$$

de telle sorte que la largeur de l'intervalle de confiance est $2zSEP$, quantité :

- qui diminue quand n augmente ;
- dimimue quand NC diminue ;
- diminue quand π se rapproche de 0 ou de 1 (maximum pour 1/2, voir remarque 5.26 page 42).

MANIPULATION AVEC R 5.67. Comme dans l'annexe D page 133, vous pouvez récupérer et sourcer la fonction `int.conf.prop.R` qui donne l'intervalle de confiance en fonction de pr la proportion mesurée, n la taille de l'échantillon et NC le niveau de confiance. Par exemple, on obtiendrait avec les calculs correspondant à (5.16) avec un NC de 0.95 avec

$$n = 1000,$$
$$y = 305,$$

on tape

```
n<-1000
y<-305
NC<-0.95
```

puis

```
int.conf.prop(y/n, n, NC)
```

on obtient

```
[1] 0.2764642 0.3335358
```

et donc un intervalle de confiance donné par

$$[0.2764642, 0.3335358] \tag{5.29}$$

Pour éviter d'avoir à se rappeler l'ordre des arguments, on peut aussi taper, *avec un ordre quelconque des arguements* :

```
int.conf.prop(pr=305/1000,n=1000,NC=0.95)
```

Attention, ici l'intervalle est légèrement différent de (5.25) car on a remplacé la valeur approché de $z = 2$ par $z = 1.959964$. Par exemple, l'intervalle de confiance pour $NC = 0.9$ est donné par

$$[0.281052, 0.328948]$$

et l'intervalle de confiance pour $NC = 0.99$ est donné par

$$[0.2674976, 0.3425024]$$

Manipulation avec R 5.68. On peut aussi utiliser la fonction `prop.test` de R et taper directement :

```
prop.test(y, n, conf.level = NC, correct = FALSE)
```

où

- 'y' est le nombre de succès ;
- 'n', le nombre d'essais ;
- 'NC', le niveau de confiance.

Le résultat obtenu sera très proche du calcul précédent. Par exemple, on obtiendrait avec les calculs correspondant à (5.16) avec un NC de 0.95 avec

$$n = 1000,$$
$$y = 305,$$

on tape

```
prop.test(305,1000,conf.level=.95,correct=FALSE)
```

on obtient

```
        1-sample proportions test without continuity correction

data:  305 out of 1000 null probability 0.5
X-squared = 152.1, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.2772553 0.3342372
sample estimates:
    p
0.305
```

et donc un intervalle de confiance donné par

$$[0.2772553, 0.3342372] \tag{5.30}$$

ce qui proche du résultat donné par (5.29).

EXERCICE 5.69. Quelle représentation de l'enseignant d'éducation physique et sportive (EPS) ont ses collègues d'autres disciplines ? Trois cent trente et un professeurs de collèges et lycées ont accepté de répondre à un questionnaire consacré à ce thème. À part quinze d'entre eux, ils se sont prononcés sur la question : "comparativement aux autres enseignants, l'enseignant d'EPS a une charge de travail :

- moins importante (121 réponses),
- équivalente (185 réponses) ou
- plus importante (10 réponses) ?"

Ici, il n'y a pas véritablement une population clairement définie, sinon une hypothétique population d'enseignants dont une partie pense que... On la "remplace" par un modèle binomial. Nous allons calculer un intervalle de confiance concernant la probabilité qu'un enseignant estime que la charge de travail de l'enseignant d'EPS est moindre. La proportion observée dans l'échantillon est donc de de

$$pr = \frac{121}{316} = 0.3829.$$

(1) Calculer un intervalle de confiance au niveau $NC = 0.95$ de la proportion.

(2) Calculer un intervalle de confiance au niveau $NC = 0.9$ de la proportion.

(3) Supposons que la proportion observée reste la même mais que la taille de l'échantillon soit de $n = 50$. Calculer un intervalle de confiance de la proportion au niveau $NC = 0.95$.

(4) Et si la taille de l'échantillon est de $n = 5000$?

Voir éléments de correction page 59.

EXERCICE 5.70. Un questionnaire envoyé par la Caisse Nationale d'Assurance Maladie à un échantillon représentatif de la population française visait à connaître les risques sportifs. Sur 7408 accidents déclarés, 1037 sont le fait de la pratique sportive.

- Calculer le pourcentage que constituent les accidents sportifs.
- Calculer un intervalle de confiance au niveau $NC = 95$ % de cette proportion ? Que pensez-vous de la précision des résultats ? À quoi est-elle due ?

EXERCICE 5.71. On s'intéresse à l'expérience aléatoire" jeter 30 fois une pièce de monnaie" où l'on note le nombre de face (qui est considéré comme un succès).

(1) Quelle est la valeur du paramètre dans cette expérience ?

(2) J'ai obtenu 18 fois face, soit $pr = 18/30 = 0.6$. Calculer l'intervalle de confiance un niveau de confiance $NC = 0.95$. Contient-il la valeur du paramètre ?

(3) Simuler cette expérience à l'aide du logiciel Rcmdr. Quelle proportion trouvez-vous ? Calculer un intervalle de confiance à $NC = 0.95$. Comprend-il la valeur du paramètre ?

(4) Tapez la commande suivante directement dans la fenêtre de "Rgui"

```
rbinom(1, size=30, prob=0.5)
```

qui simule en fait 1 fois directement la manipulation précédente. Commentez !

(5) *Questions facultatives*

(a) Tapez la commande suivante directement dans la fenêtre de "Rgui"

```
rbinom(20, size=30, prob=0.5)
```

qui simule en fait 20 fois directement la manipulation précédente.
Qu'observez-vous ?

(b) Pour chacun des groupes de commandes donnés, vous les taperez les unes à la suite des autres en tapant sur la touche "enter" après chaque commande.

(i) (A) Récupérez et sourcez la fonction `int.conf.prop.R` et tapez les commandes suivante directement dans la fenêtre de "Rgui"

```
IC<-int.conf.prop(rbinom(1, size=30, prob=0.5)/30,30,0.95)
(0.5>=IC[1])&(IC[2]>=0.5)
```

qui calcule 1 intervalle de confiance et renvoie T (TRUE) s'il contient la valeur du paramètre π et F (FALSE) sinon. Refaites plusieurs fois cette commande (grâce à la flèche $\uparrow$ de votre clavier). Commentez.

(B) *Organiser un sondage sur l'ensemble des étudiants !* Conclure !

(ii) Tapez les commandes suivante directement dans la fenêtre de "Rgui"

```
res<-rbinom(25, size=30, prob=0.5)
IC<-matrix(ncol=2,nrow=25)
for(i in 1:25) IC[i,]<-int.conf.prop(res[i]/30,30,0.95)
(0.5>=IC[,1])&(IC[,2]>=0.5)
```

qui calcule 25 intervalles de confiance et renvoie T (TRUE) si ils contiennent la valeur du paramètre π et F (FALSE) sinon. Commentez.

(iii) Tapez les commandes suivante directement dans la fenêtre de "Rgui"

```
res<-rbinom(1000, size=30, prob=0.5)
IC<-matrix(ncol=2,nrow=1000)
for(i in 1:1000) IC[i,]<-int.conf.prop(res[i]/30,30,0.95)
100*sum((0.5>=IC[,1])&(IC[,2]>=0.5))/1000
```

qui calcule 1000 intervalles de confiance et qui renvoie directement la poportion d'intervalle de confiance construits qui contiennent la valeur du paramètre inféré.

Qu'osbservez-vous ? Commentez !

(iv) Refaite la même suite de commandes, mais en prenant le nombre de jeters $n = 100$, puis $n = 1e + 05$. Commentez !

Voir éléments de correction page 59.

5.8. Le retour des smarties rouges

Reprenons l'expérience initiale de l'exercice 5.1.

Exercice 5.72.

(1) Reprendre le fichier '`SMARTIES.txt`' et calculer les intervalles de confiance au niveau de confiance $NC = 0.95$ correspondant aux échantillons de lignes 6, 3 et 7. Commentez !

(2) On cherche maintenant à calculer automatiquement tous les intervalles de confiance au niveau de confiance $NC = 0.95$. Pour créer des nouvelles variables dans Rcmdr (à partir de données déjà crées), on utilise la technique vue dans l'exercice 3.9 page 11. Dans Rcmdr, créer dans l'ordre trois nouvelles variables
   - de nom `SEP` et égale à `sqrt((p*(1-p))/nombre)`
   - de nom `pmin` et égale à `p-1.959964*SEP`
   - de nom `pmax` et égale à `p+1.959964*SEP`

   Vérifier que les deux dernières colonnes créées contiennent la borne inférieure et supérieure des intervalles de confiance au seuil $NC = 0.95$. Commentez !

(3) On veut maintenant procéder comme dans la question 2 de l'exercice 5.1. Déterminer l'intervalle de confiance au niveau $NC = 0.95$ et concluez.

Voir éléments de correction page 61.

5.9. Intervalles de confiance et test Z d'hypothèse en proportion

Cette section sera étudiée en deuxième lecture, après étude du chapitre 6.

On se contente de donner les résultats essentiels, comme dans la section 6.5.2.1 page 80.

On considère un échantillon avec n essais, une proportion de succès égale à pr. On suppose que cet échantillon provient d'une loi binomiale de paramètres n et π et on se demande si la proportion π est différente, supérieure ou inférieure à une proportion donnée π_0. Comme précédemment, on supposera que n "est grand" de façon à pouvoir utiliser l'approximation, sous l'hypothèse $H_0 : \pi = \pi_0$, la statistique

$$z = \frac{pr - \pi_0}{\sqrt{\frac{\pi - \pi_0}{n}}}$$

suit une loi normale centrée réduite.

De façon plus précise, dans le cadre d'un test sur la moyenne, on a :

Définition 5.73. Pour tester dans une population binomiale de paramètres n et π, l'hypothèse que la proportion π est égale à une norme :

$$H_0 : \quad \pi = \pi_0,$$

on utilise *le test Z d'une proportion*. On calcule la proportion pr observée sur l'échantillons de taille n puis le score normalisé suivant qui servira de statistique de test :

$$z = \frac{pr - \pi_0}{\sqrt{\frac{\pi - \pi_0}{n}}}. \tag{5.31}$$

La probabilité critique de l'hypothèse nulle p_c :

- contre l'hypothèse $H_{a(1)} : \quad \pi > \pi_0$ est égale à la probabilité que la loi normale centrée réduite soit plus élevée que z, c'est-à-dire $P(Z \geq z)$;
- contre l'hypothèse $H_{a(2)} : \quad \pi < \pi_0$ est égale à la probabilité que la même loi soit plus petite que z, soit $P(Z \leq z)$;
- contre l'hypothèse $H_{a(3)} : \quad \pi \neq \pi_0$ est égale à la probabilité que cette loi soit plus éloignée de zéro que z, c'est-à-dire $P(|Z| \geq |z|)$, soit encore $2P(Z \geq |z|)$.

Enfin,

- si $p_c \leq \alpha$, on rejettera H_0 donc on acceptera l'hypothèse alternative.
- si $p_c > \alpha$, on acceptera H_0.

Exercice 5.74.

Cet exercice est issu de [**Cha04**].

Un certain nombre d'études ont montré comment la retransmission télévisée sportive pouvait parfois renforcer les stéréotypes raciaux aux États-Unis. L'une d'elles [**SJT+95**] concerne sept événements sportifs internationaux. Sur 369 apparitions de journalistes sportifs américains, 91.869918699187 % ont été le fait de commentateurs "blancs". Sachant, d'après le recensement, que les blancs constituent 75 % de la population américaine, peut-on parler pour ces retransmissions de sur-représentation ethniques ?

Voir éléments de correction page 62.

Les manipulations à faire dans R sont données dans la correction de cet exercice. On pourra aussi procéder comme suit (voir aussi la manipulation sous R 5.68 page 45) :

Manipulation avec R 5.75. On peut aussi utiliser la fonction `prop.test` de R et taper directement :

```
prop.test(y, n, p, conf.level = NC, correct = FALSE, alternative = alt)
```

où

- '`y`' est le nombre de succès ;
- '`n`', le nombre d'essais ;

- 'p', la proportion π_0 ;
- 'NC', le niveau de confiance ;
- 'alt', une alternative égale à "two.sided", "less" ou "greater".

Par exemple, on obtiendrait avec les calculs correspondant à l'exercice 5.74 page précédente avec un NC de 0.95 avec

$$n = 369,$$
$$y = E(n.pr) = E(369 \times 0.91869918699187) = 339,$$

on tape ici

```
prop.test(339,369,0.75,conf.level=0.95,correct=FALSE,alternative="greater")
```

on obtient

```
        1-sample proportions test without continuity correction

data:  339 out of 369 null probability 0.75
X-squared = 56.0081, df = 1, p-value = 3.609e-14
alternative hypothesis: true p is greater than 0.75
95 percent confidence interval:
 0.8921368 1.0000000
sample estimates:
        p
0.9186992
```

et donc une probabilité critique égale à 3.6086e-14, ce qui est bien la valeur donnée dans la correction page 62.

Exercice 5.76. On pourra voir une application amusante des lois binomiales sur des sourcier : voir annexe G.

5.10. Quelque(s) exercice(s) supplémentaire(s)

Exercice 5.77.

Récupérez et sourcez la fonction `test.sondage.R` qui simule le travail des instituts de sondages en période pré-électorale pour un second tour d'élection avec seulement deux candidats (processus très simplifié, car en réalité, on utilise la méthode des quotats, dont on extrapole certains résultats) : on introduit

- `tpop`, la taille de population des électeurs ;
- `techan`, la taille de l'échantillon choisis (tirés aléatoirement dans la population électorale) ;
- `p0`, la probabilité théorique de succès (ici égale à la proportion de voies du gagnant) ;
- `NC`, le niveau de confiance (argument optionnel, égal par défaut à 0.95)

Cette fonction simule une population de taille `tpop`, détermine aléatoirement un groupe d'électeurs votant pour le gagnant de cardinal `tpop`× `p0`, puis réalise un échantillon aléatoire de taille `techan` et en déduit une proportion et un intervalle de confiance au seuil `NC`.

(1) Faites quelques simulations en prenant par exemple

```
 test.sondage(1000, 100, 0.53)
$prop
[1] 0.43

$intc
[1] 0.3329669 0.5270331
 test.sondage(10000, 1000, 0.53)
```

```
$prop
[1] 0.526

$intc
[1] 0.4950522 0.5569478
 test.sondage(1e+05, 1000, 0.53)
$prop
[1] 0.522

$intc
[1] 0.4910403 0.5529597
 test.sondage(1e+05, 1000, 0.53, 0.99)
$prop
[1] 0.55

$intc
[1] 0.5094767 0.5905233
 test.sondage(1e+05, 1000, 0.53, 0.999)
$prop
[1] 0.53

$intc
[1] 0.4780659 0.5819341
```

Commentez !

(2)

	nombre	pourcentage (/Inscrits)
Inscrits	44 472 733	100
Votants	37 342 004	83,97
	nombre	pourcentage (/Votants)
Blancs ou Nuls	1 568 426	4,20
Exprimés	35 773 578	95,80
	nombre	pourcentage (/Exprimés)
M. Nicolas SARKOZY	18 983 138	53,06
Mme Ségolène ROYAL	16 790 440	46,94

TAB. 5.2. Résultats officiels du second tour de la présidentielle 2007

On s'intéresse au second tour de l'élection présidentielle de 2007, pour laquelle, on donné dans le tableau 5.2 les résultats officiels issus de `http://www.interieur.gouv.fr/sections/a_votre_service/resultats-elections/PR2007/index.html` (pour l'ensemble de la France). Essayez de faire les simulations suivantes (qui risquent de ne pas fonctionner, pour cause de mémoire insuffisante) :

```
test.sondage(35773578, 1000, 18983138/35773578)
```

```
$prop
[1] 0.516

$intc
[1] 0.4850261 0.5469739
 test.sondage(35773578, 1000, 18983138/35773578, 0.9)
$prop
[1] 0.542

$intc
[1] 0.4850261 0.5469739
 test.sondage(35773578, 1e+05, 18983138/35773578)
$prop
[1] 0.52849

$intc
[1] 0.5253961 0.5315839
```

Commentez !

(3) Si votre ordinateur ne vous permet pas de faire des simulations pour la France entière, restreniez-vous par exemple à l'Ile-de-France (`http://www.interieur.gouv.fr/sections/a_votre_service/resultats-elections/PR2007/011/011.html`) voire à Paris (`http://www.interieur.gouv.fr/sections/a_votre_service/resultats-elections/PR2007/011/077/1177.html`)

(4) Refaites les simulations précédentes en imaginant un score proche du résultat du second tour de l'élection présidentielle de 1974 (Giscard : 50,81 % et Mitterand : 49,19 % ; voir `http://fr.wikipedia.org/wiki/%C3%89lection_pr%C3%A9sidentielle_fran%C3%A7aise_de_1974`)

```
 test.sondage(26367807, 1000, 13396203/26367807)
$prop
[1] 0.52

$intc
[1] 0.4890351 0.5509649
```

Commentez !

5.11. Éléments de correction

Éléments de correction de l'exercice 5.1

(1) Voir l'histogramme de la figure 5.8 page suivante. On constate sur cet histogramme que la classe la plus importante est la classe $[0.15, 0.20]$ ce qui signifie que la proportion "la plus fréquente" (donc "la plus probable") est dans l'intervalle $[0.15, 0.20]$.

(2) Une autre façon d'évaluer la proportion "la plus probable" est de calculer le nombre total de smarties rouges divisé par le nombre total de smarties en tapant dans la fenêtre de script :

```
 sum(smarties$rouge)/sum(smarties$nombre)
```

on obtient

$$p \approx 0.173, \tag{5.32}$$

ce qui confirme l'observation faite sur l'histogramme.

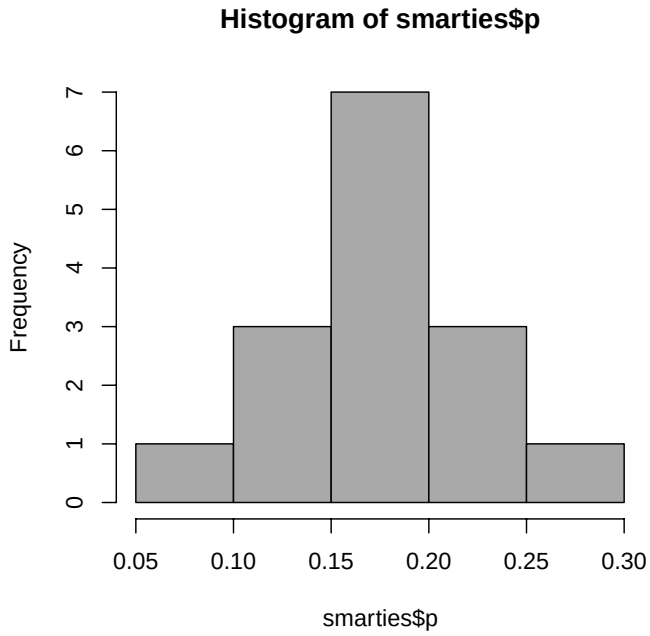


FIG. 5.8. histogrammes de la variable p des données 'SMARTIES.txt'

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 5.3

La fonction `sample` effectue un tirage aléatoire dans l'échantillon {1,2,3,4,5,6}. Au vu de la remarque 5.11 page 27, on observe bien des proportions qui se rapproche de la probabilité d'apparition de chacune des six valeurs, soit 1/6.

Ces instructions permettent donc de simuler numériquement des tirages de dés à six faces.

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 5.4

(1) Tapez dans "Rgui"

```
sample(1:50, size = 6, replace = F)
```

ou mieux

```
sort(sample(1:50, size = 6, replace = F))
```

(2) Faire plusieurs fois la commande précédente ou alors, plus subtilement,

```
n <- 20
dudu <- matrix(nrow = n, ncol = 6)
for (i in 1:n) {
    dudu[i, ] <- sort(sample(1:50, size = 6, replace = F))
}
dudu
```

Ceux qui souhaitent en savoir plus sur le fonctionnement des commandes `for` et `matrix` pourront taper dans R, les commandes `help("for")` et `help("matrix")` ou `?matrix` ♢

Cela donnerait par exemple

```
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    8    9   12   16   26   49
[2,]   22   27   28   32   40   43
```

```
 [3,]   11   25   27   28   37   45
 [4,]    2    7   12   21   28   35
 [5,]    4   17   26   45   46   47
 [6,]    5   13   19   21   36   47
 [7,]   12   15   34   40   41   47
 [8,]   17   30   31   32   41   47
 [9,]    6    9   26   28   40   42
[10,]    7   10   28   31   45   48
[11,]    2   16   21   25   33   39
[12,]    3   17   20   22   34   46
[13,]   18   23   24   44   47   48
[14,]   29   31   33   36   47   49
[15,]    4    5   21   33   42   48
[16,]    3    5   15   24   28   47
[17,]    7   10   14   22   30   33
[18,]    5   31   35   37   42   45
[19,]   11   12   13   15   46   48
[20,]   16   25   27   28   29   35
```

(3) Non, bien sûr ....

Éléments de correction de l'exercice 5.14

Il y a 3 événements possible : la valeur du résultat est

- 1, 3 ou 5, de probabilité 3/6
- 2, de probabilité 1/6
- 4,6 , de probabilité $2/6 = 1/3$

Si on jette le dé un millier de fois, la proportion de l'événement " veiller pour continuer à réfléchir à tout ça" se rapprochera donc de 1/3.

Éléments de correction de l'exercice 5.30

On a

$$\pi = 0.5, \quad n = 5,$$
$$\pi = 5/6, \quad n = 1.$$

Éléments de correction de l'exercice 5.36

(1)

(2)

Voir sur la figure 5.9 les graphes obtenus en tapant

```
binom.complet(5, 0.3, p = 1e+06)
```

Voir sur la figure 5.10 les graphes obtenus en tapant

```
binom.complet(1000, 0.3, p = 10000)
```

Voir sur la figure 5.11 les graphes obtenus en tapant

```
binom.complet(1000, 0.3, p = 1e+06)
```

(3) Les différences entre les probabilités théoriques et les proportions observées s'amenuisent comme le montre

```
binom.complet(5, 0.3, p = 1000, difference = T)
[1]  0.00907 -0.02085  0.02170 -0.00970 -0.00065  0.00043
```

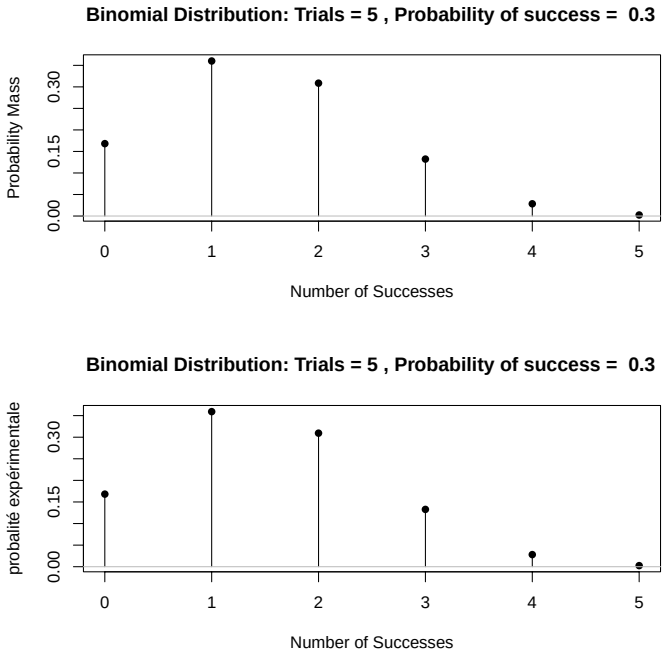


FIG. 5.9. le résultat de la fonction binom.complet pour $n = 5, \pi = 0.3, p = 1e6$

```
 binom.complet(5, 0.3, p = 10000, difference = T)
[1] -0.01083  0.00905 -0.00560  0.00480  0.00325 -0.00067
 binom.complet(5, 0.3, p = 1e+05, difference = T)
[1]  0.00080 -0.00258  0.00106  0.00067 -0.00004  0.00009
 binom.complet(5, 0.3, p = 1e+06, d = T)
[1]  0.000170  0.000310 -0.000609  0.000277 -0.000173  0.000025
```

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 5.37

Voir en figure 5.12 page 57 les cinq courbes obtenues. On constate que quand le paramètre π augmente et se rapproche de 1, les courbes de décalent vers la droite : cela signifie que quand la probabilité de succès se rapproche de 1, on a plus de chance d'obtenir un grand nombre de succès ! Autrement dit, les résultats les plus probables sont à droite quand π se rapproche de 1, à gauche quand π proche de 0, et proche du milieu quand π proche de 0.5.

Pour les cas extrèmes $\pi = 0$ et $\pi = 1$, les graphes de probabilités sont très simples !

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 5.38

Voir en figure 5.13 les quatres courbes obtenues.

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 5.42

On obtient successivement en écrivant par exemple

```
20*0.5
20*0.5*(1-0.5)
sqrt(20*0.5*(1-0.5))
```

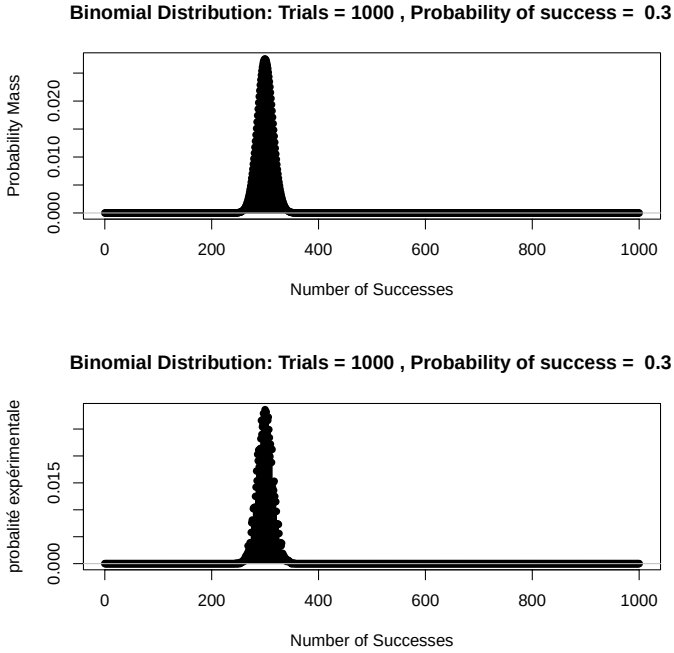


FIG. 5.10. le résultat de la fonction binom.complet pour $n = 1000, \pi = 0.3, p = 10000$

$$\mathbb{E}(X) = 10, \quad \sigma = 2.236068,$$
$$\mathbb{E}(X) = 8, \quad \sigma = 2.529822,$$
$$\mathbb{E}(X) = 160, \quad \sigma = 5.656854.$$

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 5.50

(1) On obient successivement pour une variable aléatoire binomiale X de paramètres $n = 7$ et $\pi = 0.2$:

$$P(X = 2) = 0.275251,$$
$$P(X = 0) = 0.209715,$$
$$P(X = 9) = 0,$$
$$P(X \leq 5) = 0.999629,$$
$$P(X \geq 5) = 0.000371,$$
$$P(X > 5) = 1 - P(X \leq 5) = 1 - 0.999629 = 0.000371$$

pour cette dernière, on peut aussi passer par l'aire à droite :

$$P(X > 5) = 0.000371,$$
$$P(2 \leq X \leq 5) = P(X \leq 5) - P(X \leq 1) = 0.999629 - 0.576717 = 0.422912.$$

(2) On a

$$\begin{aligned} P(X = 0) + \ldots + P(X = 5) &= 0.209715 + 0.367002 + 0.275251 + 0.114688 + 0.028672 + 0.004301 \\ &= 0.999629 \end{aligned}$$

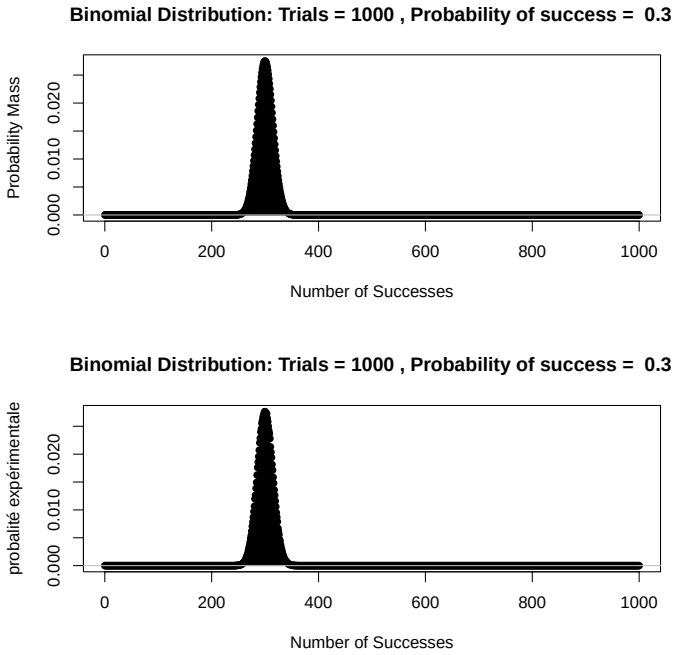


FIG. 5.11. le résultat de la fonction binom.complet pour $n = 1000, \pi = 0.3, p = 1e6$

et

$$P(X \leq 5) = 0.999629$$

(3) On a

$$\begin{aligned} P(X = 6) + P(X = 7) &= 0.000358 + 1.3e - 05 \\ &= 0.000371 \end{aligned}$$

et

$$P(X > 5) = 0.000371$$

Voir le graphique 5.14 page 59.

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 5.51 On retrouve beaucoup plus rapidement les résultats de l'exercice 5.50 en tapant

```
Pi <- 0.2
n <- 7
```

puis

```
dbinom(c(2, 0, 9), size = n, prob = Pi)
```

```
[1] 0.2752512 0.2097152 0.0000000
```

```
pbinom(5, size = n, prob = Pi)
```

```
[1] 0.9996288
```

puis

```
pbinom(5, size = n, prob = Pi, lower.tail = FALSE)
```

```
[1] 0.0003712
```

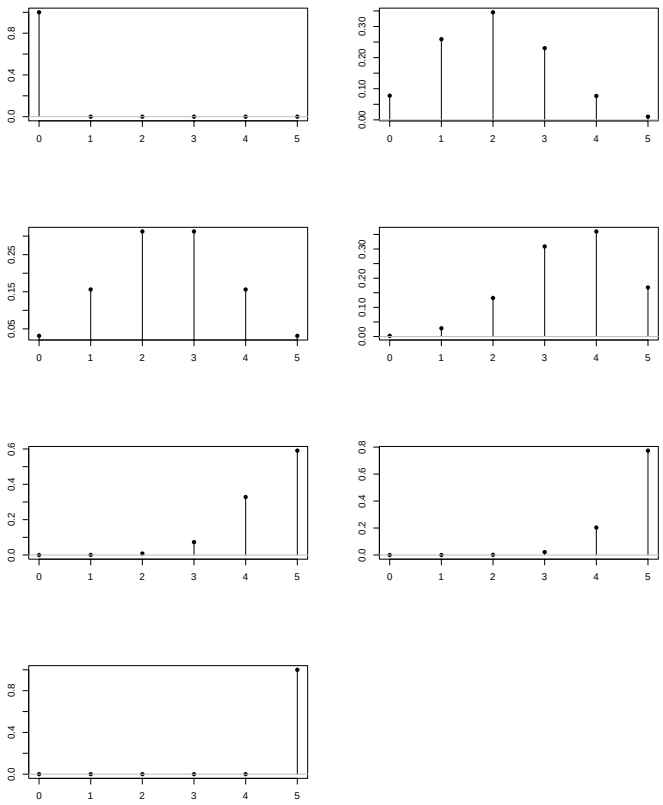


FIG. 5.12. Les cinq graphes correspondant aux graphes de la loi binomiale avec la probabilité de succès en $\pi = 0$, $\pi = 0.4$, $\pi = 0.5$, $\pi = 0.7$, $\pi = 0.9$, $\pi = 0.95$, $\pi = 1$ et $n = 5$.

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 5.52

(1) Le nombre moyen de personnes que l'on peut espérer toucher dans cette première vague d'appels est par définition la moyenne des valeurs de succès, soit encore l'espérance de la loi binomiale de paramètres n et π, c'est-à-dire $\mathbb{E}(X) = n\pi$, soit 120 personnes.

(2) La probabilité de contacter au moins $n_3 = 120$ personnes est égale à $P(X \geq 120)$. On peut la calculer de deux façon :
   - On écrit et on calcule avec les probabilités cumulées
   $$P(X \geq 120) = 1 - P(X < 120) = 1 - P(X \leq 119) = 0.5306621577.$$
   - On peut aussi passer par "l'aire à droite" :
   $$P(X \geq 120) = P(X > 119) = 0.5306621577.$$

(3) La probabilité de contacter au moins $n_1 = 150$ personnes est égale à $P(X \geq 150)$. On peut la calculer de même de deux façon :
   - On écrit et on calcule avec les probabilités cumulées
   $$P(X \geq 150) = 1 - P(X < 150) = 1 - P(X \leq 149) = 5.8989e - 06.$$
   - On peut aussi passer par "l'aire à droite" :
   $$P(X \geq 150) = P(X > 149) = 5.8989e - 06.$$

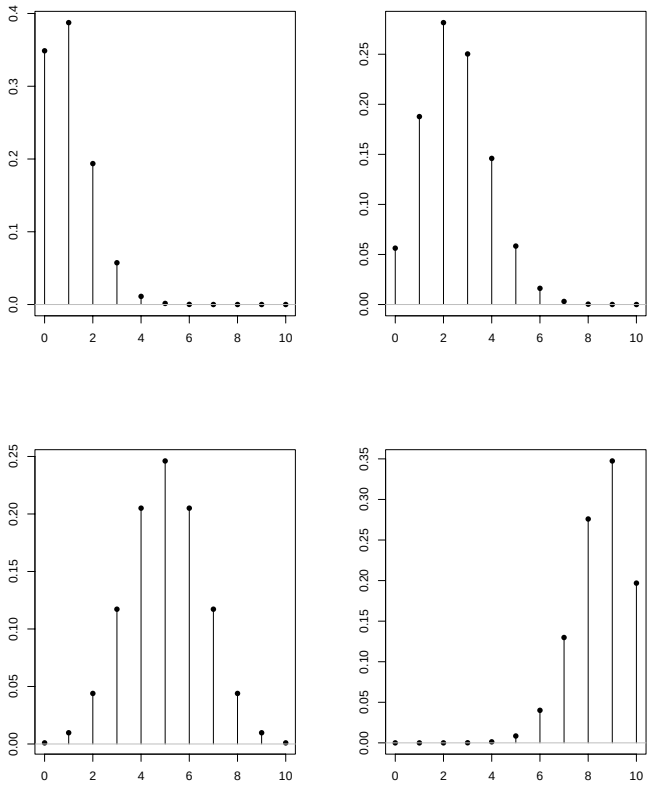


FIG. 5.13. Les quatre graphes correspondant aux graphes de la loi binomiale avec la probabilité de succès en $\pi = 0.1$, $\pi = 0.25$, $\pi = 0.5$, $\pi = 0.85$ et $n = 10$.

(4) Si on veut appeler n_2 personnes pour espérer, en moyenne, obtenir 150 succès, il faut utiliser l'argument de la question 1 à "l'envers" :

$$\mathbb{E}(X(n_2, 0.6)) = n_2\pi = 150$$

On a donc

$$n_2 = \frac{150}{\pi} = \frac{150}{0.6} = 250.$$

Il faut donc contacter $n_2 = 250$ personnes.

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 5.63

En procédant comme dans la définition 5.62, on obtient par exemple en ligne de commande

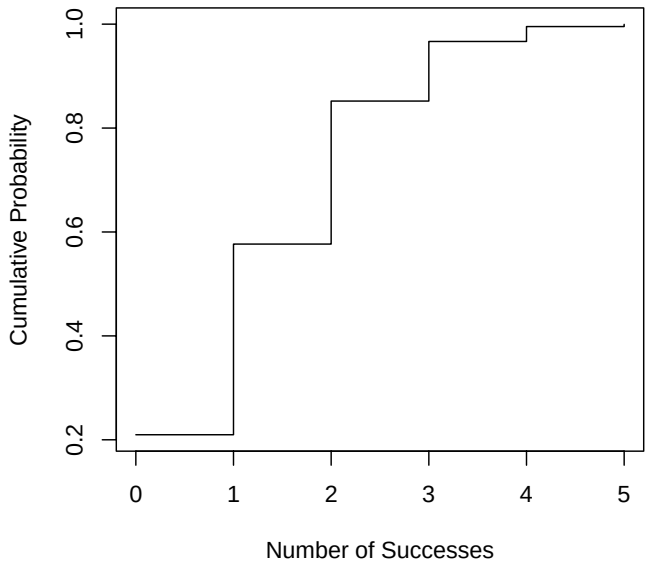
```
qnorm((1 + 0.6)/2)
```

```
[1] 0.8416212
```

et donc le z associé au Niveau de confiance $NC = 0.60$ vaut $z = 0.841621$. Pour les plus avertis, on pourra taper

```
NC <- c(0, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, 0.999, 1)
qnorm((1 + NC)/2)
```

```
[1] 0.0000000 0.6744898 0.8416212 1.0364334 1.2815516 1.6448536 1.9599640
[8] 2.5758293 3.2905267       Inf
```



(4)

FIG. 5.14. Le graphes des probabilités cumulées pour $n = 7$ et $\pi = 0.2$.

voire

```
NC <- c(0, seq(50, 90, by = 10)/100, 0.95, 0.99, 0.999, 1)
qnorm((1 + NC)/2)
```

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 5.69

En tapant par exemple

```
pr<-121/(121+185+10)
int.conf.prop(pr,316,0.95)
int.conf.prop(pr,316,0.9)
int.conf.prop(pr,50,0.95)
int.conf.prop(pr,5000,0.95)
```

on obtient successivement les intervalles

$$[0.329316, 0.4365068],$$
$$[0.3379327, 0.4278901],$$
$$[0.2481747, 0.5176481],$$
$$[0.3694377, 0.3963851],$$

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 5.71

On s'intéresse à l'expérience aléatoire "jeter 30 fois une pièce de monnaie" où l'on note le nombre de face (qui est considéré comme un succès).

(1) Si la pièce est bien équilibrée, on a $\pi = 0.5$ et $n = 30$.

(2) On obtient l'intervalle de confiance suivant :

$$[0.4246955, 0.7753045],$$

qui contient la valeur de $\pi = 0.5$.

(3) On procède comme page 35 : on utilise le menu déroulant "Distributions", l'option "Distributions discrètes" puis "Distribution binomiale", puis "Echantillon d'une distribution binomiale". Dans la fenêtre dialogue, il faut indiquer
- "Nombre d'essais" (valeur de n) : 30,
- "Probabilité de succès" (valeur de π) : 0.5,
- "Nombre d'échantillons" : 1,
- "Nombre d'observations" : 1

Un jeu de données est alors créé qui s'appelle par défaut "EchantillonsBinomiaux". Sa visualisation donne

```
        obs1
sample1   12
```

On obtient donc ici 12 succès. L'intervalle de confiance à $NC = 0.95$ est

$$[0.2246955, 0.5753045],$$

qui contient *pour mes valeurs* la valeur de $\pi = 0.5$.

Refaisons un autre tirage. Sa visualisation donne

```
        obs1
sample1    9
```

On obtient donc ici 9 succès. L'intervalle de confiance à $NC = 0.95$ est

$$[0.1360176, 0.4639824],$$

qui ne contient *pour ces valeurs* pas la valeur de $\pi = 0.5$. Ici, on sera dans le cas défavorable où l'intervalle de confiance ne contient pas le paramètre inféré !

(4) Si on tape la commande suivante directement dans la fenêtre de "Rgui"

```
rbinom(1, size=30, prob=0.5)
```

on observe un tirage binomial de paramètre $n = 30$ et $\pi = 0.5$, directement dans "Rgui" sans passer par Rcmdr !

(5) *Questions facultatives*

(a) De même, la commande suivante directement dans la fenêtre de "Rgui"

```
rbinom(20, size=30, prob=0.5)
```

qui simule en fait 20 tirages binomiaux de paramètre $n = 30$ et $\pi = 0.5$, directement dans "Rgui" sans passer par Rcmdr !

(b) (i) (A)

(B) *Prévoir organisation d'un sondage sur l'ensemble des étudiants !*

(ii) Les commandes suivantes

```
res<-rbinom(25, size=30, prob=0.5)
IC<-matrix(ncol=2,nrow=25)
for(i in 1:25) IC[i,]<-int.conf.prop(res[i]/30,30,0.95)
(0.5>=IC[,1])&(IC[,2]>=0.5)
```

donnent par exemple

```
 [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

soit la plupart des intervalle qui contiennent la valeur du paramètre inféré et quelques autres malheureux qui ne le contiennent pas !

(iii) Les commandes suivantes

```
res<-rbinom(1000, size=30, prob=0.5)
IC<-matrix(ncol=2,nrow=1000)
for(i in 1:1000) IC[i,]<-int.conf.prop(res[i]/30,30,0.95)
100*sum((0.5>=IC[,1])&(IC[,2]>=0.5))/1000
```

donnent par exemple

```
[1] 95.8
```

c'est-à-dire une proportio proche de 0.95 !

(iv) Avec $n = 100$, puis $n = 1e + 05$, on obtient successivement

```
[1] 93
```

puis

```
[1] 94.9
```

Éléments de correction de l'exercice 5.72

(1) Les intervalles de confiance au niveau de confiance $NC = 0.95$ correspondant aux échantillons de lignes 7, 3 et 6 sont données successivement par

$$[-0.0213685, 0.1463685],$$
$$[0.1040313, 0.4120978],$$
$$[0.0522661, 0.3227339].$$

On remarque que le premier intervalle contient des valeurs négatives, ce qui n'est pas choquant *a fortiori* puisque la loi normale qui a permis de construire la largeur de l'intervalle de confiance sous la forme (voir remarque 5.66) est en fait définie sur $\mathbb{R}$ et rien n'interdit mathématiquement qu'elle donne des valeurs négatives, même si celles-ci correspondent à des grandeurs physiquement toujours positives (comme des proportions !)

(2)

Vous devriez obtenir le tableau 5.3 page suivante. On constate sur ce tableau que la SEP la plus faible est 0.04279 correspondant à l'intervalle de confiance le moins large (pour l'échantillon de numéro 7 et d'intervalle associé $[-0.021368, 0.146368]$). La SEP la plus grande est 0.07859 correspondant à l'intervalle de confiance le plus large (pour l'échantillon de numéro 3 et d'intervalle associé $[0.104031, 0.412098]$).

(3) On calcule la proportion totale en écrivant de nouveau :

```
sum(smarties$nombre)
sum(smarties$rouge)/sum(smarties$nombre)
```

Grâce à la valeur de p fournie par (5.32), c'est-à-dire $p = 0.1729167$ et le nombre total de smarties rouge $N = 480$, on détermine alors l'intervalle de confiance au niveau NC :

$$[0.1390852, 0.2067481].$$

Ici, cet intervalle de confiance est de largeur plus faible que les précédents puisque le nombre total de smarties est ici 480 et donc la SEP, ici égale à 0.017261 est plus faible (voir remarque 5.66).

	rouge	nombre	p	SEP	pmin	pmax
1	5	31	0.1612903	0.0660586	0.0318179	0.2907627
2	5	31	0.1612903	0.0660586	0.0318179	0.2907627
3	8	31	0.2580645	0.0785898	0.1040313	0.4120978
4	4	31	0.1290323	0.0602101	0.0110226	0.2470419
5	5	31	0.1612903	0.0660586	0.0318179	0.2907627
6	6	32	0.1875000	0.0689981	0.0522661	0.3227339
7	2	32	0.0625000	0.0427908	-0.0213685	0.1463685
8	8	32	0.2500000	0.0765466	0.0999715	0.4000285
9	7	32	0.2187500	0.0730792	0.0755173	0.3619827
10	6	32	0.1875000	0.0689981	0.0522661	0.3227339
11	4	33	0.1212121	0.0568144	0.0098580	0.2325662
12	6	33	0.1818182	0.0671408	0.0502246	0.3134118
13	6	33	0.1818182	0.0671408	0.0502246	0.3134118
14	7	33	0.2121212	0.0711647	0.0726409	0.3516015
15	4	33	0.1212121	0.0568144	0.0098580	0.2325662

TAB. 5.3. Intervalles de confiance [pmin,pmax] pour les données 'smarties.txt'

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 5.74

On cherche à montrer que la proportion observée $pr = 0.91869918699187$ est supérieure à la proportion $\pi_0 = 0.75$.

On procède au *test Z d'une proportion.*

On fait l'hypothèse nulle H_0 : $\pi = \pi_0$. avec $\pi_0 = 0.75$. On cherche à montrer que le paramètre π de la loi binomiale, dont proviendraient les données de l'échantillon étudié, est plus grande que π_0. On fait donc l'hypothèse alternative suivante : H_1 : $\pi > \pi_0$.

Puisque $n = 369$, est "grand", on remplacera la loi binomiale de paramètre n et π_0 par la loi normale de moyenne $\mu = \pi_0$ et d'écart-type $\sqrt{\pi_0(1-\pi_0)/n}$. Grâce à R, on trouve la valeur suivante de la statistique

$$z = \frac{pr - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = 7.483858$$

La probabilité critique $P(Z \geq z)$ (pour la loi normale centrée réduite) est égale à

$$p_c = 3.6086e - 14.$$

Puisque p_c est inférieure au égal au niveau de signification $\alpha = 0.05$, on rejette l'hypothèse nulle H_0. Ainsi, H_1 est vraie et *la proportion π est plus grande que* $\pi_0 = 0.75$, au risque 0.05.

Les commentateurs blancs sont donc bien sur-représentés dans ces retransmissions, au seuil usuel $\alpha = 0.05$.

CHAPITRE 6

Généraliser les résultats obtenus avec une moyenne

6.1. Le modèle statistique normal

6.1.1. Les données normales

Comme dans le chapitre, 5, on peut charger le fichier '`STUDENTH.txt`' et tracer l'histogramme, en densité (voir remaque 3.2 page 8) et le graphe quantile-quantile des données '`Taille`'. On obtient les deux graphes de la figure 6.1. Le graphe quantile-quantile nous montre un bon accord avec la loi normale.

Sur la figure 6.2, nous avons rajouté en rouge, la loi normale idéale, loi théorique en forme de cloche. Cette loi théorique est centrée sur la moyenne de l'échantillon (tracée en bleu en pointillé sur la figure) et présente, dans un grand nombre de cas, un phénomène idéal de répartition de données.

En fait, les données sont dites normales, mais on verra que ce sont leurs moyennes sur un "grand nombre" de valeurs qui le sont réellement. Nous reviendrons longuement sur ces moyennes au cours du chapitre. ♢

6.1.2. Les intervalles : des événements particuliers

Définition 6.1. Lorsqu'une variable aléatoire peut prendre toutes les valeurs au sein d'un intervalle, on dit qu'elle est *continue*. Pour les variables aléatoires continues, les probabilités sont définies sur des intervalles par l'intermédiaire d'une fonction de densité. La probabilité d'un intervalle s'identifie en effet à la surface qui lui correspond sous la fonction de densité.

On consultera pour plus de détails l'annexe H qui montre sur des exemples simples comment les propriétés des variables aléatoires dsicrètes "passent" à la limite pour devenir continue. On retiendra en particulier l'équation (H.6) pour la "courbe en cloche idéale" qui se généralise pour toute fonction de densité p :

$$P(a \leq X \leq b) = \int_a^b p(x)dx \tag{6.1}$$

La figure 6.3 montre une densité de probabilité où les valeurs sont concentrées autour de 0, rarement supérieures à 2 en valeur absolue et symétrique : on a autant de chance de voir des valeurs positives que négatives. Par exemple, la probabilité qu'une valeur issue de cette loi soit située dans l'intervalle $[0.5, 2]$ soit

$$P(0.5 \leq X \leq 2) = \int_{0.5}^{2} p(x)dx = 0.285787, \tag{6.2}$$

est donnée par la surface en rouge (l'aire totale sous la courbe est de 1).

Les propriétés habituelles des probabilités des variables aléatoires s'étendent aux variables continues (inférieure à 1, somme unitaire, addition des événements disjoints).

Remarque 6.2. De façon analogue à la remarque 5.46 page 34 (dans le cas discret), notons aussi que, si on se donne la probabilité p, le quantile est le nombre q tel que

$$P(X \leq q) = p.$$

c'est-à-dire

$$\int_{-\infty}^{q} p(x)dx = p.$$

On choisira parfois la définition (équivalente ici, quand la densité est "intégrable" et de primitive continue) : si on se donne la probabilité p, le quantile est la plus petite valeur x telle que

$$P(X \leq x) \geq p.$$

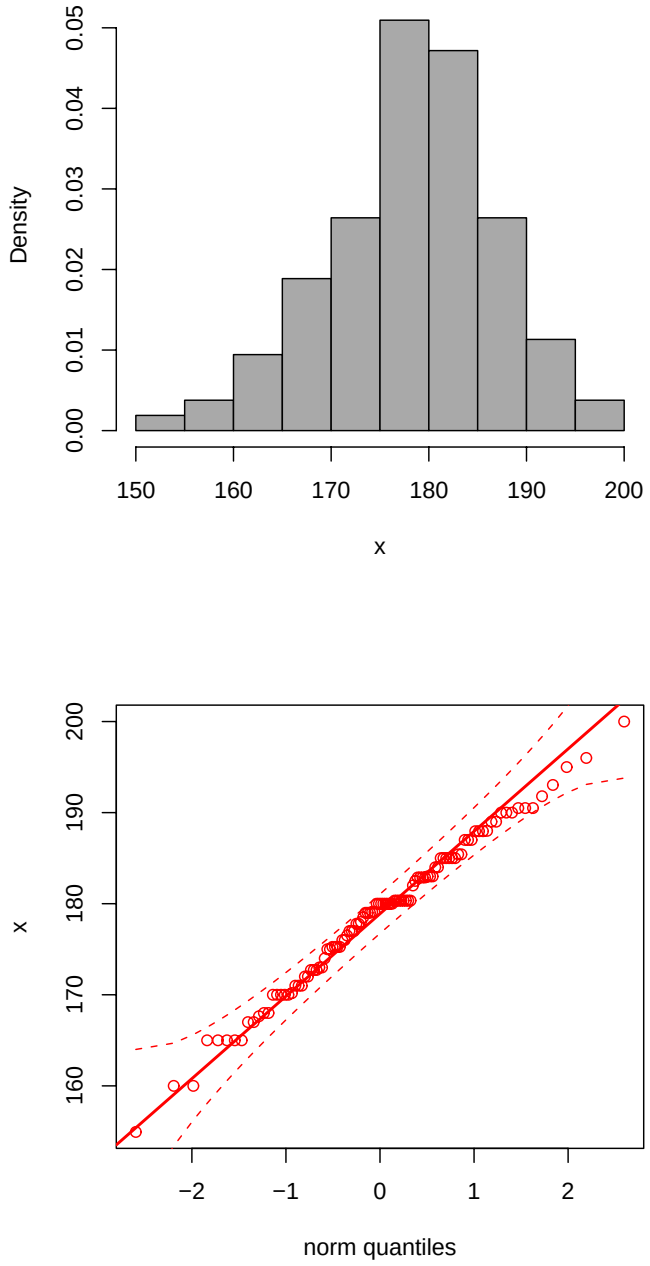


FIG. 6.1. Histogramme (en densité) et graphe quantile-quantile sur les données de 'Taille' de 'STUDENTH.txt'.

♢

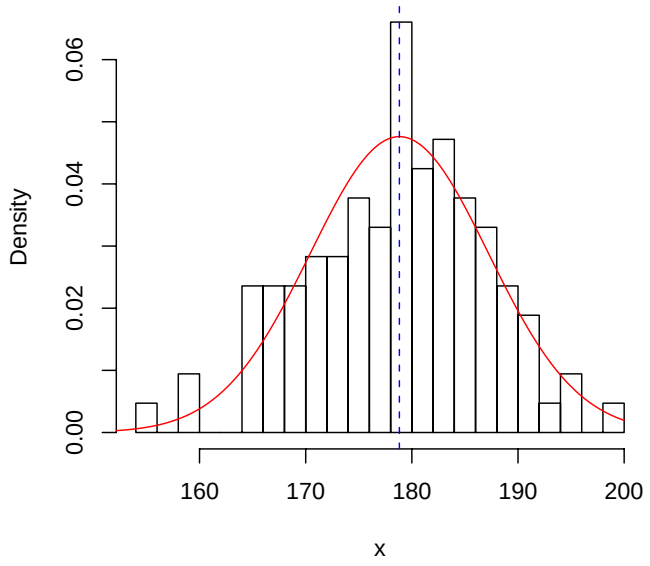


FIG. 6.2. Histogramme (en densité et avec 18 classes) et la loi normale associée sur les données de 'Taille' de 'STUDENTH.txt'.

6.1.3. La loi normale

La "courbe en cloche idéale" correspond à une fonction de densité p, donnée par

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{6.3}$$

Attention, le π dont on parle maintenant est le "vrai" π (celui qui vaut le rapport de la circonférence d'un cercle sur son diamètre) à ne pas confondre avec la proportion π du chapitre 5 !

Cette fonction est donnée sur la courbe 6.4 page 67.

Cette fonction de densité est symétrique, elle atteint son maximum en zéro et est pratiquement nulle au delà de trois (en valeur absolue).

MANIPULATION AVEC RCMDR 6.3. Pour obtenir l'allure de cette fonction de densité, choisir le menu déroulant "Distributions", puis "Distributions continues" puis "Distribution normale" puis "Graphe de la distribution normale". Il faut laisser dans la fenêtre de dialogue les valeurs par défaut pour $mu = 0$ et $sigma = 1$.

Pour avoir une famille de modèles relativement souples et utiles en pratique, il faut que la loi de probabilité puisse avoir n'importe quelle espérance et variance. On peut généraliser en ce sens cette loi normale.

DÉFINITION 6.4. En probabilité, une variable aléatoire suit une loi normale (ou loi normale gaussienne ou loi de Laplace-Gauss) d'espérance $\mu \in \mathbb{R}$ et d'écart-type $\sigma > 0$ si elle admet une densité de probabilité p telle que :

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{6.4}$$

Nous dirons que la variable aléatoire X soit une loi normale et on notera

$$X \sim \mathcal{N}(\mu, \sigma) \tag{6.5}$$

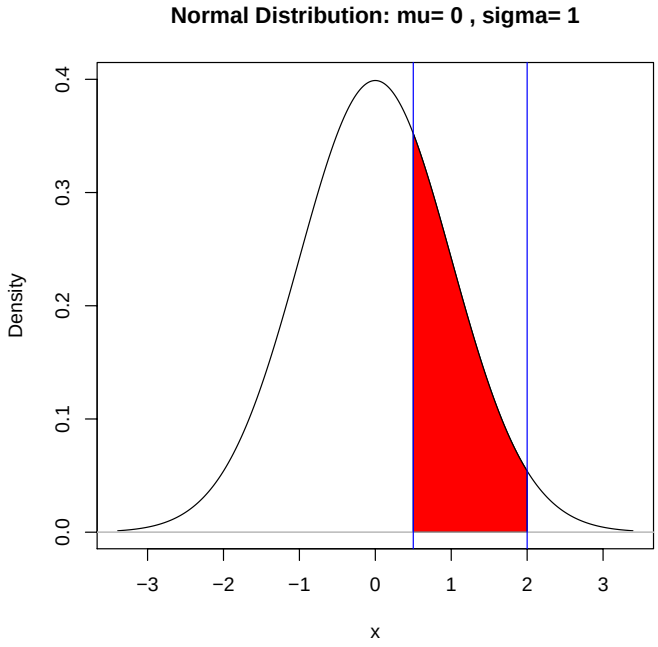


FIG. 6.3. Une fonction de densité p et l'aire en rouge sous la courbe représentant la valeur de la probabilité qu'une valeur issue de cette loi soit située dans l'intervalle $[0.5, 2]$

Attention, on voit aussi la notation, non utilisée dans ce cours,

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

♢

La "courbe en cloche idéale" est le cas particulier suivant :

DÉFINITION 6.5. La loi normale centrée réduite correspond au cas où $\sigma = 1$ et $\mu = 0$.

La dénomination d'espérance $\mu \in \mathbb{R}$ et d'écart-type $\sigma > 0$ provient du fait que l'espérance et l'écart-type de la loi normale sont égaux (au sens de (H.7) page 151) justement à μ et σ :

$$\mu = \mathbb{E}(X) = \int_{-\infty}^{\infty} xp(x)dx,$$
$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx}.$$

Voir par exemple sur la figure 6.5, la densité de la loi normale de moyenne $\mu = 2$ et d'écart-type $\sigma = 0.5$.

MANIPULATION AVEC RCMDR 6.6. Pour obtenir l'allure de cette fonction de densité, choisir le menu déroulant "Distributions", puis "Distributions continues" puis "Distribution normale" puis "Graphe de la distribution normale". Il faut choisir dans la fenêtre de dialogue les valeurs $mu = 2$ et $sigma = 0.5$.

Attention à ne pas confondre densité et probabilité, on obtient les probabilités en intégrant la fonction de densité (d'après (6.1)).

MANIPULATION AVEC RCMDR 6.7. Les probabilités se calculent à l'aide du menu déroulant "Distributions", puis "Distributions continues" puis "Distribution normale" puis "Probabilités normales". Il faut choisir dans la

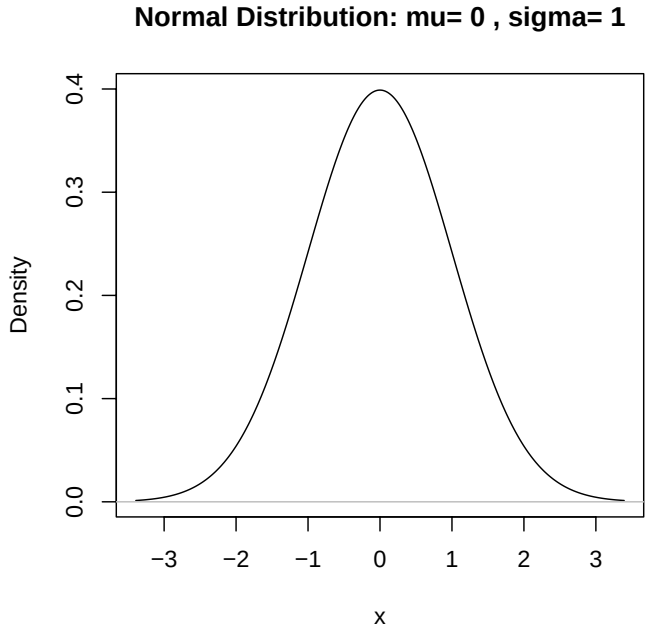


FIG. 6.4. La "courbe en cloche idéale"

fenêtre de dialogue les valeurs de *mu* et *sigma*. On peut alors calculer l'aire, à gauche ou à droite, sous la courbe de densité pour n'importe quelle valeur (appelée *quantile*, voir remarque 6.2 page 63).

On rappelle que de façon analogue au cas discret, l'aire à gauche désigne :

$$P(X \leq a) = \int_{-\infty}^{a} p(x)dx \tag{6.6}$$

et que l'aire à droite désigne

$$P(X \geq a) = \int_{a}^{\infty} p(x)dx \tag{6.7}$$

et que la somme de ces deux aires vaut 1. Pour calculer d'autres types de probabilité, on pourra utiliser la formule (6.1) qui s'écrit aussi

$$P(a \leq X \leq b) = \int_{a}^{b} p(x)dx = \int_{-\infty}^{b} p(x)dx - \int_{-\infty}^{a} p(x)dx. \tag{6.8}$$

soit encore

$$P(a \leq X \leq b) = P(x \leq b) - P(x \leq a) \tag{6.9}$$

Si on passe par les "aires à droites", on a

$$P(a \leq X \leq b) = P(x \geq a) - P(x \geq b) \tag{6.10}$$

♢ On rappelle aussi que

$$P(X = a) = 0 \tag{6.11}$$

et donc

$$P(X \leq a) = P(X < a). \tag{6.12}$$

EXERCICE 6.8. On étudie la loi normale centrée réduite (c'est-à-dire de moyenne nulle et d'écart-type égal à 1)

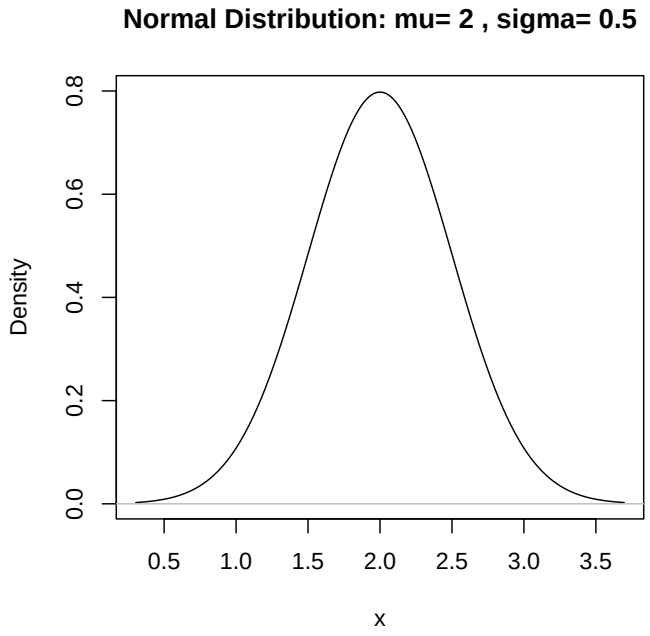


FIG. 6.5. La densité de la loi normale de moyenne $\mu = 2$ et d'écart-type $\sigma = 0.5$.

(1) calculer la probabilité correspondant aux intervalles suivants : $P(X \leq -0.5)$, $P(X \leq 4.5)$, $P(X \geq 1.25)$ et $P(X \geq -2)$.

(2) Puis, en procédant en deux temps, calculer $P(1.25 \leq X \leq 1.5)$ et $P(-0.65 \leq X \leq 1.4)$.

Voir éléments de correction page 85.

EXERCICE 6.9. Reprendre les questions de l'exercice 6.8 page précédente en remplaçant la loi normale centrée réduite par la loi $X \sim \mathcal{N}(\mu = 1.5, \sigma = 2)$.

Voir éléments de correction page 85.

EXERCICE 6.10 (facultatif).

Comme dans l'exercice 5.51 page 34, on peut utiliser la fonction `pnorm` qui fournit directement dans "Rgui", la fonction de distribution (et donc le calcul des probabilités) pour la loi normale normale ; plus précisement, si $X = (x_1, \ldots, x_n)$ est un vecteur de valeurs (dites quantiles) , alors

- la commande
  ```
  pnorm(X, mean = mu, sd = sigma)
  ```
  fournit le vecteur des probabilités $\big(P(X \leq x_1), \ldots, P(X \leq x_n)\big)$, associées à la lois normales de moyenne `mu` et d'écart-type `sigma`,
- et la commande
  ```
  pnorm(X, mean = mu, sd = sigma, lower.tail = FALSE)
  ```
  fournit le vecteur des probabilités $\big(P(X \geq x_1), \ldots, P(X \geq x_n)\big)$, associées à la même loi normale. Dans ce dernier cas, notons que $P(X \geq x_i) = P(X > x_i)$!

Reprendre les questions de l'exercice 6.9, avec ces fonctions.

REMARQUE 6.11. On peut montrer que

$$X \sim \mathcal{N}(\mu, \sigma) \Longleftrightarrow \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

Autrement dit on passe la densité de la loi normale centrée réduite à celle de la loi normale générale par une "dilatation-translation". Par exemple, les deux graphes des figures 6.4 et 6.5 sont rigoureusement superposables ; seuls différent les échelles des axes.

EXERCICE 6.12. Sundberg étudie la VO_2max d'une population d'enfants (8-15 ans). L'histogramme de ces VO_2max a une allure de courbe en cloche. On considérera donc que la loi normale d'espérance $\mu = 55.4$ et d'écart-type $\sigma = 7.7$ modélise correctement cette mesure.

(1) Tracer la densité de la loi normale associée.

(2) Quelle est la probabilité qu'un enfant de cet âge ait une VO_2max supérieure à 60 ? Supérieure à 70 ? Inférieure à 50 ?

(3) Douze enfants aveugles ont également été évalués. La moyenne observée n'est alors que de 45.3 . Quelle est la probabilité qu'un enfant voyant présente une VO_2max inférieure à cette valeur ?

Voir éléments de correction page 86.

EXERCICE 6.13. Comment établir des limites concernant le taux d'hématocrites (volume total des globules rouges par rapport au sang) ? O'toole *et al.* se sont intéressés à une population de triathlètes et ont établi que les taux observés avant une course chez ces hommes suivent assez fidèlement une loi normale d'espérance $\mu = 43{,}2$ (%) et d'écart-type $\sigma = 2{,}9$.

(1) Quelle est la probabilité d'observer dans ce contexte un triathlète dont le taux dépasse 45 % ? Dont le taux dépasse 50 % ?

(2) Les auteurs suggèrent d'employer une valeur de détection de 52 % car elle est située à trois écarts-types au dessus de la moyenne. Qu'en pensez-vous ?

REMARQUE 6.14. Lorsque le nombre de répétitions est assez grand, on peut employer la loi normale pour approximer certaines probabilités binomiales.

Lorsqu'on s'intéresse à des mesures quantitatives, on calcule le plus souvent la moyenne de l'échantillon, et on aimerait avoir une idée de sa variation d'un échantillon à l'autre.

6.2. Une remarque sur le lien entre la densité de probabilité et les histogrammes en densité

REMARQUE 6.15. *Attention*, quand on parle d'écart-type (expérimentaux) déterminés par  , par la fonction `sd`, il s'agit en fait de la déviation standart (défini par (3.4) page 10). En toute rigueur, l'écart-type est défini par (3.3) page 10.

On pourra consulter l'annexe I, qui vous montre sur un exemple simple la propriété suivante : si on fait un tirage aléatoire d'une variable aléatoire définie par la densité p de sa loi continue et que l'on trace son histogramme en densité (voir remarque 3.2 page 8), ce dernier se rapproche du graphique de la densité de probabilité : la loi expérimentale se rapproche donc de la loi théorique (exactement comme dans le cas discret, voir définition 5.12 page 27). De plus, les moyennes et écart-type expérimentaux se rapprochent des moyennes et écart-type théoriques, définis par (H.7), page 151 (exactement comme dans le cas discret, voir propriété 5.25 page 29).

6.3. La distribution d'échantillonnage d'une moyenne

Afin de définir un intervalle de confiance dans ce contexte, on considère à nouveau que les données ont été produites par un modèle probabiliste, la loi normale s'imposant alors *naturellement*. L'espérance mathématique μ de cette loi étant inconnue[1], l'objectif sera de la situer par un encadrement.

[1]Ainsi que l'écart-type σ

Nous allons procéder comme dans la page 35 (section 5.6 page 35) du chapitre 5. Nous allons ensuite développer la notion d'intervalle de confiance (comme dans la section 5.7 page 41).

MANIPULATION AVEC RCMDR 6.16. Nous allons supposer que les données sont générées par un modèle normal d'espérance $\mu = 15$ et d'écart-type $\sigma = 3$. Commençons par générer un échantillon de n=16 valeurs issues de cette loi grâce à au logiciel Rcmdr. Il faut choisir le menu déroulant "Distributions", puis "Distributions continues" puis "Distribution normale" puis "Echantillon d'une distribution normale". Dans la fenêtre de dialogue qui s'ouvre il faut indiquer pour

- "mu" : 15,
- "sigma" :3,
- "Nombre d'échantillons" : 1,
- "Nombre d'observations" : 16

Il faut également cocher dans la rubrique "Ajouter au jeu de données" les cases moyennes et écarts-types (et ne pas cocher la rubrique "somme"). Il se crée alors un tableau à 1 ligne et 18 colonnes, les deux dernières indiquant *pour ma machine* une moyenne de 16.146579 (et incidemment un écart-type de 2.543566) :

```
            obs1     obs2     obs3     obs4     obs5     obs6     obs7     obs8
sample1 13.45538 13.85967 19.08061 18.03557 16.37041 17.97913 16.65436 21.09333
            obs9    obs10    obs11    obs12    obs13    obs14    obs15    obs16
sample1 15.19221 16.01543 16.25203 19.42371 15.44985 11.83550 12.65738 14.99071
            mean       sd
sample1 16.14658 2.543566
```

Comme précédemment nous allons considérer que cet échantillon est un parmi d'autres.

Attention, dans le chapitre 5, un échantillon consistait en la donnée d'une proportion, crée aléatoirement. Ici, dans ce chapitre, un échantillon consiste en la donnée de 16 nombres. Pour considérer que cet échantillon est un parmi d'autres, nous allons donc maintenant créer un tableau rectangulaire de nombres. Chaque ligne de ce tableau sera un échantillon.

EXERCICE 6.17.

(1) En choisissant :
   - "mu" : 15,
   - "sigma" :3,
   - "Nombre d'échantillons" : 10000,
   - "Nombre d'observations" : 16

   générer $p = 10000$ échantillons semblables au précédent.
(2) Vérifier qu'il se crée alors un tableau à 10000 lignes et 18 colonnes.
(3) Représenter par un histogramme la distribution d'échantillonnage de la moyenne. Tracer aussi un graphe quantile-quantile (voir chapitre 3). Calculer la moyenne de ces moyennes et l'écart-type de ces moyennes.

   Vous devriez obtenir pour l'histogramme une figure analogue à celle de la figure du haut du graphique 6.6 page ci-contre et pour le graphe quantile-quantile, une figure analogue à celle de la figure 6.7.

   Cette figure est tout à l'analogue de la figure 5.6 (où les tirages étaient binomiaux) et la démarche suivie est l'analogue exact de la section 5.6 page 35 (sauf qu'ici la loi de probabilité est continue !).
(4) La moyenne et l'écart-type de toutes les moyennes (c'est-à-dire de l'avant dernière colonne de mon échantillon) que j'obtiens sur *mon* échantillon sont

$$m = 14.984508, \tag{6.13a}$$

$$sd = 0.74611 \tag{6.13b}$$

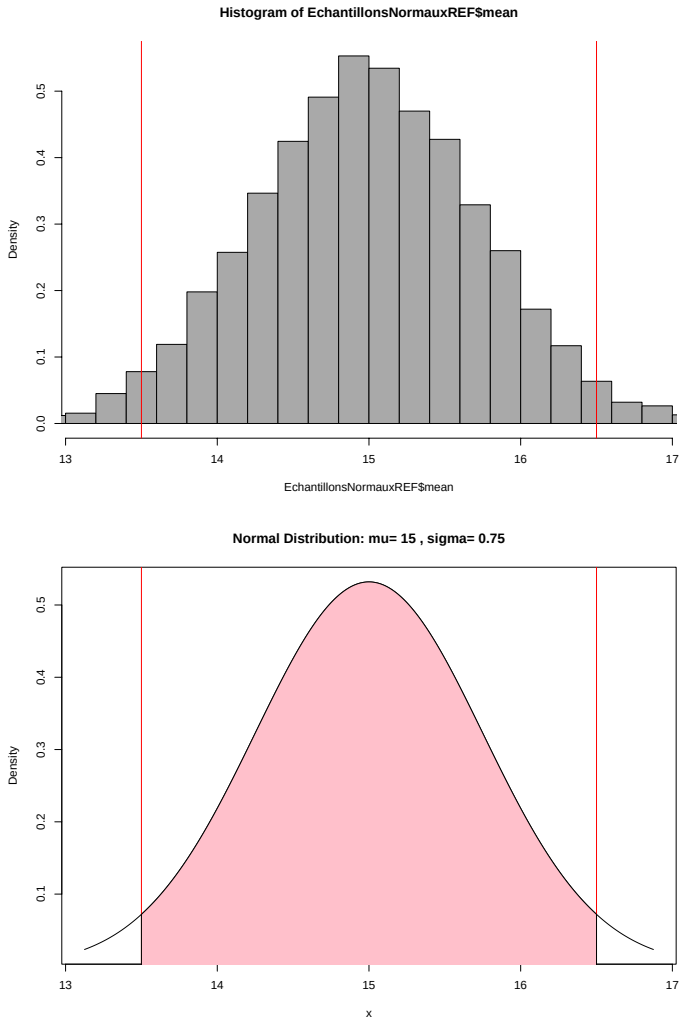


FIG. 6.6. Histogramme de la distribution des moyennes (échantillon de taille n=16, 10000 tirages) observées sur une loi normale de moyenne $\mu = 15$ et d'écart-type $\sigma = 3$ en densité avec 30 classes représenté uniquement sur l'intervalle $[13.125, 16.875]$. En dessous se trouve la loi normale de moyenne $\mu = 15$ et d'écart-type $\sigma/\sqrt{n} = 0.75$. On a aussi représenté les deux droites verticales d'abscisses 13.5 et 16.5.

Comparez avec les votres (accessibles avec Rcmdr). Comparez avec les nombres suivants (où μ et σ sont les moyenne et écart-type de la loi normale initiale) :

$$\mu = 15, \tag{6.14a}$$

$$\frac{\sigma}{\sqrt{n}} = \frac{3}{4} = 0.75 \tag{6.14b}$$

(5) En faisant vous même la démarche de la section 5.6 page 35, essayez de construire *vous-même*, un intervalle qui contiennent 95 % des données.

Refaisons maintenant "proprement" le raisonnement de la section 5.6 page 35 :

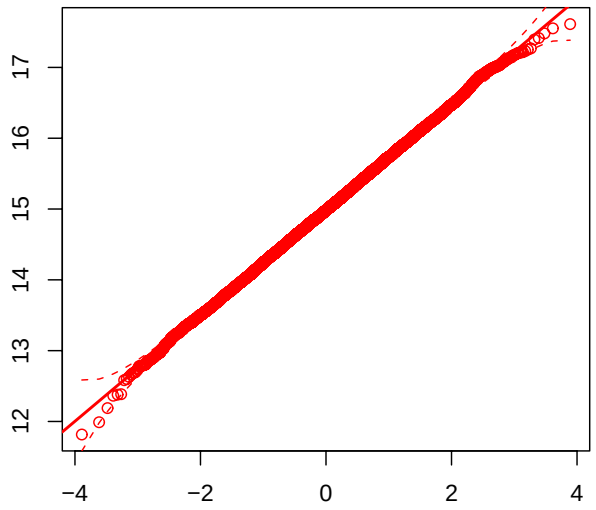


FIG. 6.7. Histogramme graphe quantile-quantile (comparaison avec la loi normale).

- D'après les observations faites sur la figure 6.6, il semblerait que la distribution d'échantillonnage (des moyennes) obtenue présente les caractéristiques suivantes :
  - elle semble suivre la loi normale,
  - elle est centrée sur le paramètre μ , l'espérance mathématique (ici 15). Cela peut-être montré par la théorie des probabilités. Cela est confirmé par le très beau graphe quantile quantile 6.7.
  - son écart-type est en relation avec l'écart-type σ de la distribution et la taille de l'échantillon n. Plus précisément on démontre grâce à la théorie des probabilités qu'il est proche de

$$\widetilde{\sigma} = \frac{\sigma}{\sqrt{n}} \tag{6.15}$$

    soit à peu près 0.75.
- Cela est confirmé par la similarités des résultats donnés par (6.13) et (6.14).
- On suppose donc que les moyennes obtenues m suivent une loi normale de moyenne μ et d'écart-type $\widetilde{\sigma}$ défini par (6.15), soit

$$m \sim \mathcal{N}(\mu, \widetilde{\sigma})$$

  D'après la remarque 6.11 page 68, cela est équivalent à

$$\frac{m-\mu}{\widetilde{\sigma}} \sim \mathcal{N}(0,1) \tag{6.16}$$

  Cherchons la probabilité que "la moyenne m soit à moins de deux écart-types de μ" (comme dans la section 5.6), c'est-à-dire

$$P\left(-2 \leq \frac{m-\mu}{\widetilde{\sigma}} \leq 2\right)$$

  Selon (6.16), on cherche donc à "résoudre"

$$P(-z \leq X \leq z) = NC \tag{6.17}$$

où NC est le niveau de confiance (c'est-à-dire la proportion de cas que l'on veut couvrir) et X suit une loi de probabilité normale centrée réduite. Comme dans le raisonnement de la note en petit caractère 5.7 page 43, on utilise l'équation (5.27) : ainsi, (6.17) est équivalent à

$$P(X \leq z) = \frac{1+NC}{2} \tag{6.18}$$

– Si on connaît $z = 2$, on trouve alors grâce à Rcmdr

$$\frac{1+NC}{2} = 0.97725$$

et donc

$$NC = 2 \times 0.97725 - 1 = 0.9545 \approx 0.95.$$

On retrouve donc le fameux $NC = 0.95$ introduit en section 5.6 !

– Si réciproquement, on connait $NC = 0.95$ est que l'on cherche z, on procède comme en définition 5.62 page 43 : on calcule

$$\frac{1+NC}{2} = 0.975.$$

et on obtient

$$z = 1.959964 \approx 2.$$

De plus, on justifie *a posteriori* la note en petits caractères page 43 ainsi que la construction du z en fonction du NC !

- Si on utilise la remarque de l'annexe I, on vient donc de montrer que les moyennes sont à moins de deux écart-types dans 95 % des cas !
- Justifions cela expérimentalement en tapant la séquence suivante dans Rgui qui dénombre le nombre de moyennes dans l'intervalle

$$[\mu - 2\widetilde{\sigma}, \mu + 2\widetilde{\sigma}] = [13.5, 16.5]$$

et qui en calcule le pourcentage :

```
100*sum((EchantillonsNormaux$mean>=13.5)
&(EchantillonsNormaux$mean<=16.5))/10000
```

Cela donne *pour mes valeurs* un pourcentage égal à 95.7 qui est bien proche de 95 % !

L'écart-type de la distribution d'échantillonnage s'appelle l'*erreur standard de la moyenne* (SEM). L'écart-type initial σ est pour l'heure inconnu. On l'estimera par l'écart-type issu de l'échantillon *sd*. Donnons l'analogue de la définition 5.58 page 38

Définition 6.18. On estime l'*erreur standard de la moyenne* par

$$SEM = \frac{sd}{\sqrt{n}}. \tag{6.19}$$

On la note SEP à cause de son nom anglais (Standard error of the mean).

Exercice 6.19.

Du fait de leur handicap, les aveugles n'ont-ils pas une pratique physique moindre ? Quelles sont leurs capacités physiques par rapport à leurs pairs voyants ? Connaître ces limites peut permettre de programmer des interventions plus efficaces. L'étude de Sundberg concern$e n = 12$ sujets masculins âgés de huit à quatorze ans qui montrent une VO_2max moyenne de $\overline{y} = 45.3$ ml/kg/min avec un écart-type de $\widehat{\sigma} = 8.9$ (estimation faite sur ergo-cycle).

Estimer l'erreur standard de la moyenne (SEM).

Voir éléments de correction page 87.

6.4. L'intervalle de confiance "d'une moyenne"

Sachant que la distribution d'échantillonnage suit la loi normale et donc que la valeur observée de la statistique (la moyenne) à 95% de chance d'être située à moins de deux erreurs standards du paramètre (l'espérance mathématique), on peut comme dans la section 5.7 page 41, "retourner" l'argument et donner la définition suivante.

DÉFINITION 6.20. L'intervalle de confiance au niveau 95 % d'une espérance mathématique est donné par les quantités

$$m \pm 2SEM = m \pm 2\frac{sd}{\sqrt{n}}$$

où n est la taille de l'échantillon, m est sa moyenne et sd son écart-type.

Les quantités correspondantes pour l'exercice 6.19 sont donc :

$$m - 2\frac{sd}{\sqrt{n}} = 45.3 - 2\frac{8.9}{\sqrt{12}} = 40.162,$$

et

$$m + 2\frac{sd}{\sqrt{n}} = 45.3 + 2\frac{8.9}{\sqrt{12}} = 50.438.$$

L'intervalle au niveau de confiance 95 % de la VO_2max moyenne d'un handicapé visuel est de donc de $[40.162, 50.438]$ Dans le même article, la VO_2max d'un groupe d'enfants du même âge est évaluée en moyenne à 55,4 ml/kg/min. On peut donc *voir* qu'il existe une différence notable.

6.4.1. La loi de Student

La formule de base est une approximation commode d'une formule plus précise où le coefficient multiplicateur 2 sera remplacé par une valeur dite t de Student, et ce, pour deux raisons.

D'une part, on peut souhaiter varier le niveau de confiance. D'autre part, à cause de l'approximation de σ par sd, la loi qui fait référence pour la distribution d'échantillonnage de la moyenne n'est finalement pas la loi normale, mais une loi un peu plus étalée qui s'appelle *loi de Student*. Cette loi dépend d'un paramètre qui est le *degré de liberté* et correspond ici à la taille de l'échantillon moins un $(n-1)$. Pour le problème de l'échantillon normal 'EchantillonsNormaux', on utilisera donc $ddl = n - 1 = 16 - 1 = 15$.

Si on reprend l'échantillon normal 'EchantillonsNormaux' crée précédemment, et que l'on trace l'histograme non plus des moyennes mais de la variable suivante

$$t = \frac{m - \mu}{sd/\sqrt{n}} = \frac{m - 15}{sd/4} \tag{6.20}$$

où n (ici égal à 16) est la taille de l'échantillon, m est sa moyenne mesurée sd son écart-type *mesuré*, on obtient l'histogramme de la figure 6.8. Sur cette figure, se trouvent aussi la loi de student correspondant à ddl=15 (en rouge, trait plein) et la loi normale (en bleu, trait pointillé). On constate un léger écart entre la loi de Student et la loi normale. La loi de Student est en théorie plus proche de l'histogramme que la loi normale.

On peut confirmer cela de façon précise : sur la figure 6.9 page 76 se trouvent les deux graphes quantile quantile de l'échantillon des t en comparaison avec la loi de Student et la loi normale. On constate que la loi de Student "colle légèrement mieux à l'histogramme " que la loi de Student !

EXERCICE 6.21. On considére l'échantillon normal 'EchantillonsNormaux' créee précédemment.

(1) En procédant comme dans l'exercice 3.9 page 11, introduire une nouvelle variable appelée t et définie par (6.20). Elle sera définie par l'expression suivante
```
(mean-15)/(sd/sqrt(16))
```
(2) Tracer avec Rcmdr, l'histogramme de la figure 6.8 page ci-contre.

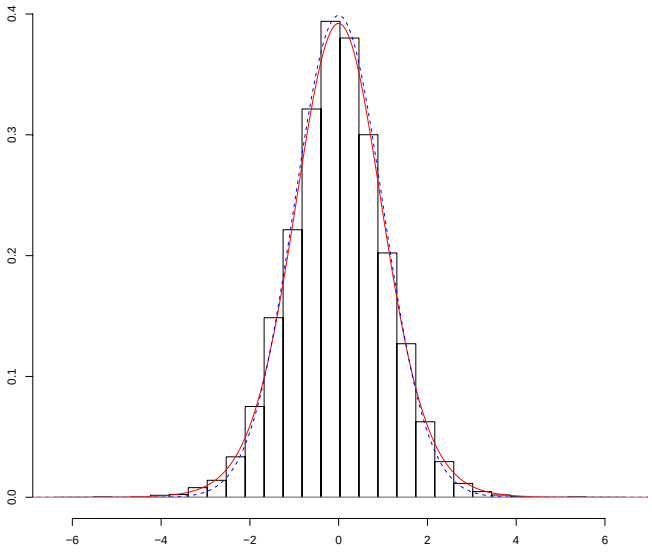


FIG. 6.8. Histogramme de la distribution des t (définie par (6.20) défini à partir de l'échantillon 'EchantillonsNormaux' crée précédemment en densité avec 30 classes. En dessous se trouvent la loi de student correspondant à ddl=15 (en rouge, trait plein) et la loi normale (en bleu, trait pointillé).

(3) Tracer avec Rcmdr, les graphes quantile-quantile de la figure 6.9 page suivante.

PROPOSITION 6.22. *L'intervalle de confiance au niveau NC d'une moyenne est donné par*

$$\left[m - t\frac{sd}{\sqrt{n}}, m + t\frac{sd}{\sqrt{n}}\right], \tag{6.21}$$

où t est le quantile[2] d'une loi de Student à $n-1$ degrés de liberté correspondant à la probabilité $q = (1+NC)/2$, n est la taille de l'échantillon, m est sa moyenne et sd son écart-type. Le coefficient multiplicateur t et pourra être obtenu[3] sous

- *soit allant dans le menu "distribution", puis "distribution continue", puis "distribution t", puis "quantiles t" et rentrer à la place de "probabilités",* `(1+NC)/2`*, où NC est un nombre égal au niveau de confiance et à la place de "degrés de liberté" $n-1$ (en laissant par défaut "aire à gauche") ;*
- *soit grâce à la ligne de commande*

  ```
  qt((1 + NC)/2, df = n - 1)
  ```

  où NC est un nombre égal au niveau de confiance et n est la taille de l'échantillon.

PREUVE FACULTATIVE. Voir par exemple `http://fr.wikipedia.org/wiki/Loi_de_Student`. □

[2]Voir remarque 6.2 page 63 ; on a donc

$$P(T \leq t) = \frac{1+NC}{2} \tag{6.22}$$

où T suit la loi de Student à $n-1$ degrés de liberté.

[3]Manipulation proche de celle de la definition 5.62 page 43 !

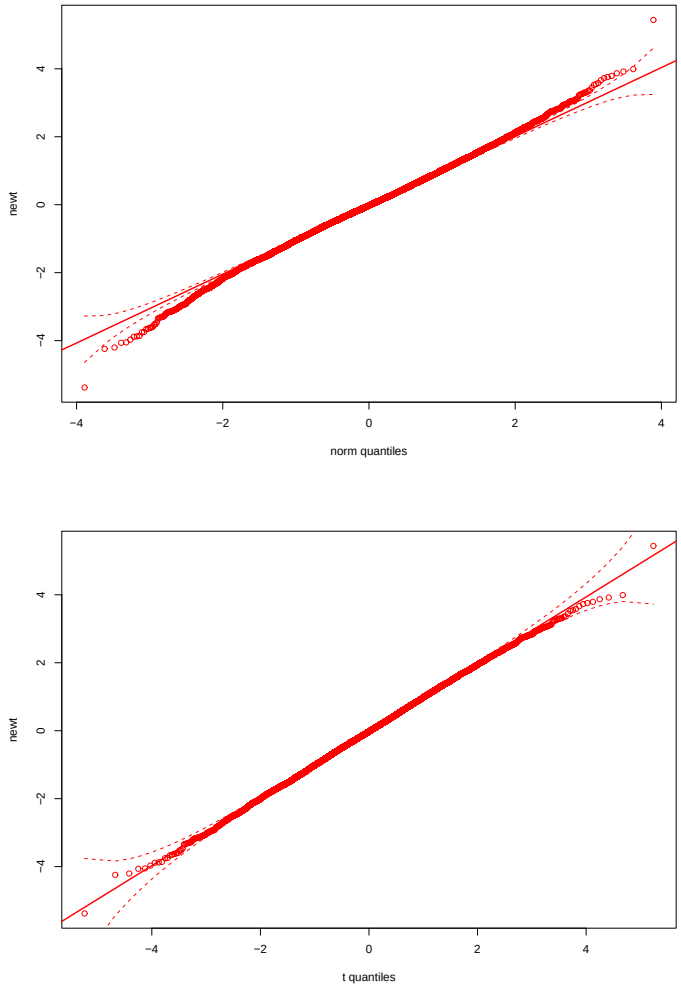


FIG. 6.9. Les deux graphes quantile quantile de l'échantillon des t en comparaison avec la loi normale (en haut) et la loi de Student (en bas) (avec $ddl = 16 - 1$).

REMARQUE 6.23. En fait, la loi de Student "tend" vers la loi normale centrée réduite quand *ddl* tend vers l'infini, comme le montre la figure 6.10. Les densités de probabilité de la loi de Student se tracent avec Rcmdr de façon analogue à celle de la loi normale.

EXERCICE 6.24.
Calculer l'intervalle de confiance à 90 %, puis à 95 %, puis à 99 % pour les données de l'exercice 6.19.
Voir éléments de correction page 87.

REMARQUE 6.25. Si on dispose des données, et pas seulement des statistiques, c'est beaucoup plus simple, il suffit de choisir le menu déroulant "Statistiques", puis les options "Moyennes" et "t-test univarié". Il faut laisser les champs relatifs aux hypothèses avec les valeurs définies par défaut.

REMARQUE 6.26. Comme dans l'annexe D page 133, vous pouvez aussi télécharger et sourcer la fonction `int.conf.moy.R` qui détermine l'intervalle de confiance en fonction de

- `mu` : moyenne mesurée

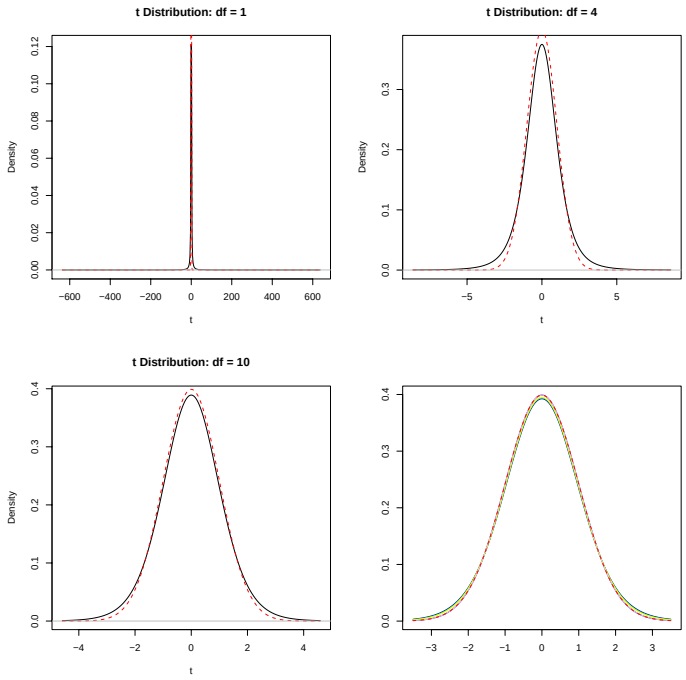


FIG. 6.10. Tracés des loi de student pour quelques valeurs de *ddl* : *ddl*=1, voir graphique numéro 1 ; *ddl*=4, voir graphique numéro 2 ; *ddl*=10, voir graphique numéro 3. Pour $ddl \in \{15, 20, 25, 1000\}$, voir dernière figure ; sur l'ensemble des figures, la loi normale centrée réduite est tracé en rouge pointillé.

- `sd` : écart-type (déviation standart) mesuré ;
- `n` : la taille de l'échantillon ;
- `NC` : seuil de confiance.

EXERCICE 6.27. R. ROLLIER (M2PPMR) a pris en charge l'entraînement de 39 joueurs de moins de 19 ans de l'ASVEL Rugby. Un test de vitesse (30 mètres avec deux changements de direction) a été passé par tous ces joueurs. Le temps (sec.) a été relevé. Voir le fichier de données `ROLLIER.txt`.

(1) Représenter graphiquement les données de temps
(2) Ces valeurs semblent-elles approximativement suivre une loi normale ?
(3) Calculer un intervalle de confiance au niveau $NC = 0.90$ de l'espérance mathématique de cette variable. Vous essayerez d'utiliser la méthode de la proposition 6.22 page 75, ainsi que celle suggérée dans la remarque 6.25 page ci-contre (c'est-à-dire avec Rcmdr) ou dans la remarque 6.26 page précédente (c'est-à-dire avec la fonction `int.conf.moy.R`.

Voir éléments de correction page 87.

EXERCICE 6.28. Pour les tailles d'échantillons (n) et niveaux de confiance (NC) suivants, trouver les valeurs du coefficient multiplicateur t à utiliser pour construire l'intervalle de confiance de la moyenne

(1) $n = 25$, NC =95 % ;
(2) $n = 25$, NC =90 % ;
(3) $n = 15$, NC =99 % ;
(4) $n = 15$, NC =98 % ;

Voir éléments de correction page 89.

EXERCICE 6.29. La ménarche (apparition des règles) est une date importante dans la vie de la jeune fille. Elle semble varier en fonction de la pratique sportive, une activité physique intense ayant tendance à la retarder. Dans quelle mesure est-ce vrai ? Une population de femmes africaines a été étudiée. La date de leur ménarche est obtenue, pour ces femmes âgées de 20 à 30 ans, par une technique de rappel (recoupement d'événements), généralement au mois près.

Cinquante handballeuses de haut niveau, ayant commencé l'entraînement avant la ménarche ont été interrogées. Leurs déclarations présentent une moyenne de $\overline{y} = 15{,}6$ années et un écart-type de $\widehat{\sigma} = 1{,}57$.

- Donner un intervalle de confiance à $NC = 95$ % de l'espérance mathématique de la ménarche pour cette population.
- En outre, un groupe témoin de 214 femmes sédentaires a été étudié, aboutissant à un âge moyen de 13,78 ans pour la ménarche. En rapprochant l'intervalle de confiance précédent de cette"norme" , que peut-on suspecter de l'influence de l'exercice physique intense ? Quelles précautions faut-il toutefois prendre dans l'énoncé de cette conclusion ?

6.5. Les tests d'hypothèses

La méthode inférentielle la plus utilisée est la technique des tests d'hypothèses. Elle est cependant décriée par nombre de méthodologistes pour sa complexité et le fait qu'elle apporte au final moins d'information que les intervalles de confiance. Cependant, malgré ses défauts, comme elle est massivement utilisée dans la publication d'études, il convient de la connaître.

On étudie des données en nombre n dont on suppose qu'elle proviennent d'une loi normale de moyenne μ et d'écart-type σ inconnus. On peut en calculer la moyenne et l'écart-type m et sd.

Soit μ_0 une norme fixée à l'avance.

On se donne un niveau de condiance NC et on pose

$$\alpha = 1 - NC. \tag{6.23}$$

Le nombre α est appelé seuil ou niveau de signification. Souvent, $NC = 0.95$ et donc $\alpha = 0.05$. De façon générale, NC est "proche de 1" et donc α "petit".

6.5.1. Introduction informelle

On se demande si $\mu = \mu_0$ ou si au contraire $\mu \neq \mu_0$.

On sait d'après la section 6.4.1 page 74 que l'intervalle de confiance au niveau NC est donné par (6.21) où t est donné par (6.22). Remarquons que si $\mu = \mu_0$, alors la probabilité que μ_0 appartiennent à cet intervalle de confiance est égale à NC. Ainsi, la probabilité que μ_0 n'appartienne pas à cet intervalle de confiance est égale à α. Ainsi, si μ_0 n'appartient pas à cet intervalle de confiance, nous chosissons de décider que $\mu \neq \mu_0$. Si au contraire μ_0 appartient à cet intervalle de confiance, nous chosissons de décider que $\mu = \mu_0$. Le "risque" de décider que $\mu \neq \mu_0$ alors que $\mu = \mu_0$ a donc une probabilité égale à α, faible.

On donne souvent l'image suivante des enchères. On veut acheter une voiture qui a chance faible α d'être "en mauvais état". Ici, l'analogue de l'hypothèse $\mu = \mu_0$ est "la voiture est en bon état" et l'analogue du contraire ($\mu \neq \mu_0$) est "la voiture est en mauvais état". On a quatre possibilités :

- On décide d'acheter la voiture alors qu'elle est en bon état.
- On décide de ne pas d'acheter la voiture alors qu'elle est en bon état.
- On décide d'acheter la voiture alors qu'elle est en mauvais état.
- On décide de ne pas d'acheter la voiture alors qu'elle est mauvais état.

• Remarquons que selon (6.21), on décide que $\mu \neq \mu_0$ si

$$\mu_0 \notin \left]m - t\frac{sd}{\sqrt{n}}, m + t\frac{sd}{\sqrt{n}}\right[,$$

c'est-à-dire si

$$|\mu_0 - m| \geq t\frac{sd}{\sqrt{n}}$$

D'après (6.23) et (6.22), on vient donc de montrer que l'on décide $\mu \neq \mu_0$ si

$$t_\alpha \leq \frac{|\mu_0 - m|}{sd/\sqrt{n}} \tag{6.24a}$$

où le nombre (positif) t_α est défini par

$$P(T \leq t_\alpha) = \frac{2-\alpha}{2} \tag{6.24b}$$

où T suit la loi de Student à $n-1$ degrés de liberté. Si au contraire (6.24a) n'est pas vraie, on décidera que $\mu = \mu_0$.

- Remarquons que (6.24a) est équivalent à

$$\frac{m-\mu_0}{sd/\sqrt{n}} \geq t_\alpha, \text{ si } m - \mu_0 \geq 0, \tag{6.25a}$$

$$\frac{m-\mu_0}{sd/\sqrt{n}} \leq -t_\alpha, \text{ si } m - \mu_0 \leq 0. \tag{6.25b}$$

La statistique $(m-\mu_0)/sd/\sqrt{n}$ appartient à une région de probabilité $P(|T| \geq t_\alpha)$, égale à α.

En effet, cette quantité vaut, d'après (5.27) et (6.24b)

$$\begin{aligned} P(|T| \geq t_\alpha) &= 1 - P(|T| \leq t_\alpha), \\ &= 1 - P(-t_\alpha \leq T \leq t_\alpha), \\ &= 2 - 2P(T \leq t_\alpha), \\ &= 2 - 2\frac{2-\alpha}{2}, \\ &= 2 - 2 + \alpha, \\ &= \alpha \end{aligned}$$

♢

Bref, on décide $\mu \neq \mu_0$

si la statistique $(m-\mu_0)/sd/\sqrt{n}$ appartient à la région $]-\infty, -t_\alpha] \cup [t_\alpha, \infty[$

qui a une probabilité $P(|T| \geq t_\alpha) = \alpha$ (si $\mu = \mu_0$). (6.26)

- On peut encore donner une autre forme de cela : on calcule la statistique $(m-\mu_0)/sd/\sqrt{n}$. On décidera $\mu \neq \mu_0$ si

$$P\left(|T| \geq \frac{|m-\mu_0|}{sd/\sqrt{n}}\right) \leq \alpha. \tag{6.27}$$

En effet, cela est équivalent à

$$1 - P\left(|T| \leq \frac{|m-\mu_0|}{sd/\sqrt{n}}\right) \leq \alpha.$$

soit

$$1 - 2P\left(T \leq \frac{|m-\mu_0|}{sd/\sqrt{n}}\right) \leq \alpha$$

et donc à

$$P\left(T \leq \frac{|m-\mu_0|}{sd/\sqrt{n}}\right) \geq \frac{2-\alpha}{2}$$

soit d'après (6.24b)

$$P\left(T \leq \frac{|m-\mu_0|}{sd/\sqrt{n}}\right) \geq P(T \leq t_\alpha)$$

ce qui est bien équivalent à

$$\frac{|m-\mu_0|}{sd/\sqrt{n}} \geq t_\alpha.$$

♢

6.5.2. Le modèle et les hypothèses

Formalisons maintenant cela. On envisagera d'autre type de décision analogue à $\mu \neq \mu_0$ comme $\mu < \mu_0$ ou $\mu > \mu_0$.

La question de référence est la comparaison à la norme, qu'on va appeler *hypothèse nulle* :

$$H_0 : \quad \mu = \mu_0.$$

On va confronter cette hypothèse à une *hypothèse alternative*, il y a ici trois choix possibles[4] :

$$H_{a(1)} : \quad \mu > \mu_0,$$
$$H_{a(2)} : \quad \mu < \mu_0,$$
$$H_{a(3)} : \quad \mu \neq \mu_0.$$

DÉFINITION 6.30. Pour réaliser un test sur un paramètre du modèle, il faut définir deux hypothèses : l'hypothèse alternative (souvent notée H_1) correspond à ce que l'on souhaite démontrer et l'hypothèse nulle à une hypothèse de référence (souvent notée H_0). Ces deux alternatives sont exhaustives ou non et exclusives (si on accepte l'une, on rejette l'autre et réciproquement).

REMARQUE 6.31. Ce qui est particulier dans la démarche de test c'est qu'il s'agit d'un raisonnement par l'absurde : on va se placer dans la situation de référence (l'hypothèse nulle) afin d'essayer de démontrer qu'elle est fausse et donc que l'alternative (qui nous intéresse) est vraie.

En reprenant l'image de la section 6.5.1, H_0 correspond à l'hypothèse "la voiture est en bon état" et H_1 correspond à l'hypothèse "la voiture est en mauvais état". Rejetter ou accepter H_0 correspond au rejet ou l'acceptation de l'achat. On a quatre possibilités :

- On accepte H_0 alors qu'elle est juste.
- On refuse H_0 alors qu'elle est juste ; ici la probabilité est donc α.
- On accepte H_0 alors qu'elle est fausse.
- On refuse H_0 alors qu'elle est fausse.

Nous avons déjà étudié de façon formelle le cas où l'alternative $H_{a(3)}$ était considérée dans la section 6.5.1. Nous reprenons et étendons les deux façons de conclure ((6.25) ou (6.27)). Les deux méthodes des sections 6.5.2.1 et 6.5.2.2 page ci-contre sont rigoureusement équivalentes. Les statisticiens préconisent la seconde, mais la première est encore presque partout utilisée !

6.5.2.1. Par probabilité critique.

DÉFINITION 6.32. Pour réaliser un test statistique, il faut, de façon générale,

- identifier le modèle statistique qui correspond à la situation
- définir les deux hypothèses nulle et alternative portant sur les paramètres du modèle
- calculer une statistique sur l'échantillon dont on connaît le comportement si l'hypothèse nulle est vraie
- Utiliser cette statistique pour calculer une probabilité critique qui mesure la compatibilité entre les données (résumée par la statistique calculée) et l'hypothèse nulle
- Conclure en comparant cette probabilité critique au seuil (ou niveau de signification) α, souvent pris conventionnellement égal à 5 %.

De façon plus précise, dans le cadre d'un test sur la moyenne, on a :

DÉFINITION 6.33. Pour tester dans une population normale d'espérance mathématique μ et d'écart-type σ, l'hypothèse que la moyenne est égale à une norme :

$$H_0 : \quad \mu = \mu_0,$$

[4] Les puristes diront qu'il y a *un seul choix possible* parmi trois possiblités !

on utilise *le test de Student pour un échantillon*. On calcule la moyenne m et l'écart-type sd observés sur l'échantillons de taille n puis le score normalisé suivant qui servira de statistique de test :

$$t = \frac{m - \mu_0}{sd/\sqrt{n}}. \tag{6.28}$$

La probabilité critique de l'hypothèse nulle p_c :
- contre l'hypothèse $H_{a(1)}$: $\mu > \mu_0$ est égale à la probabilité que la loi de Student à $n-1$ degrés de liberté soit plus élevée que t, c'est-à-dire $P(T \geq t)$;
- contre l'hypothèse $H_{a(2)}$: $\mu < \mu_0$ est égale à la probabilité que la même loi soit plus petite que t, soit $P(T \leq t)$;
- contre l'hypothèse $H_{a(3)}$: $\mu \neq \mu_0$ est égale à la probabilité que cette loi soit plus éloignée de zéro que t, c'est-à-dire $P(|T| \geq |t|)$, soit encore $2P(T \geq |t|)$.

Enfin,
- si $p_c \leq \alpha$, on rejettera H_0 donc on acceptera l'hypothèse alternative.
- si $p_c > \alpha$, on acceptera H_0.

REMARQUE 6.34. Pour le choix de l'hypothèse alternative $H_{a(3)}$: $\mu \neq \mu_0$, on parle de test bilatéral ; pour les deux autres hypothèses, on parle de test unilatéral.

Le niveau de signification α correspond, quand H_0 est vrai, à la probabilité de rejeter H_0. On parle aussi de risque de premier espèce[5]

Comprenons bien que lorsqu'on décide de rejetter H_0, c'est pour deux raisons :
- soit on le fait "à raison" ; dans ce cas l'hypothèse alternative est exacte et dans ce cas, la moyenne est bien inférieure à la norme ;
- soit on le fait à tort ; dans ce cas, H_0 est vraie ; la moyenne est bien égale à la norme, mais un événement qui se produit très rarement (ce qui arrive avec une probabilité inférieure à α) s'est produit.

Et le pire, c'est que nous ne saurons jamais si nous avons pris la bonne décision ....

REMARQUE 6.35. Il est important de se rappeler que si H_0 est vraie, *la statistique* $t = (m - \mu_0)/(sd/\sqrt{n})$ *suit une loi de Student à* $n-1$ *degrés de liberté*.

6.5.2.2. Par région critique (section facultative).

La problématique est de donner une *région critique* qui correspondra à l'ensemble des valeurs pour lesquels la probabilité critique devient inférieure au niveau de signification. On décide de rejeter l'hypothèse nulle si la statistique observée sur l'échantillon appartient à cette région critique.

DÉFINITION 6.36. Pour tester dans une population normale d'espérance mathématique μ et d'écart-type σ, l'hypothèse que la moyenne est égale à une norme :

$$H_0 : \quad \mu = \mu_0,$$

on utilise *le test de Student pour un échantillon*. On définit la région critique
- correspondant à l'hypothèse alternative $H_{a(1)}$: $\mu > \mu_0$ par

$$R_c = [t_\alpha, +\infty[, \tag{6.29}$$

où t_α est défini par

$$P(T \geq t_\alpha) = \alpha, \tag{6.30}$$

pour la loi de student à $n-1$ degrés de liberté ;
- correspondant à l'hypothèse alternative $H_{a(2)}$: $\mu < \mu_0$ par

$$R_c =]-\infty, t_\alpha], \tag{6.31}$$

[5]opposé au risque de seconde espèce, non évoqué ici ; il correspond à la probabilité d'accepter H_0, alors qu'elle est fausse. Ce risque est en général plus difficile à calculer. Il est en fait "plus grave" : dans l'image des enchères, il correspond au risque que l'on prend d'acheter la voiture alors qu'elle est en mauvais état, ce qui est plus grave de refuser l'achat si elle est en bon état !

où t_α est défini par

$$P(T \leq t_\alpha) = \alpha, \tag{6.32}$$

pour la loi de student à $n-1$ degrés de liberté ;

– correspondant à l'hypothèse alternative $H_{a(3)}$: $\mu \neq \mu_0$ par

$$R_c =]-\infty, -t_\alpha] \cup [t_\alpha, +\infty[\,, \tag{6.33}$$

où le réel positif t_α est défini par

$$P(|T| \geq t_\alpha) = \alpha, \tag{6.34}$$

soit encore

$$P(T \geq t_\alpha) = \frac{\alpha}{2}, \tag{6.35}$$

pour la loi de student à $n-1$ degrés de liberté.

Comme dans le début de la définition 6.33 page 80, on déterminera ensuite la statistique t définie par (6.28).

Enfin,

– si t appartient à R_c, on rejettera H_0 donc on acceptera l'hypothèse alternative.

– si t n'appartient à R_c, on acceptera H_0.

♢

6.5.3. Un exemple

Nous allons reprendre le problème de Coquelin (données `COQUELIN.txt`) qui souhaite comparer la VMA de son groupe de joueurs de football à celle de l'élite de Clairefontaine (μ_0=17.35). Voir début du chapitre 4.

On souhaite démontrer que la moyenne est inférieure à la norme et on privilégie donc l'hypothèse alternative $H_{a(2)}$.

- On peut vérifier tout d'abord la normalité des données en procédant comme au à la section 6.1.1 page 63 :

  On trace l'histogramme, en densité, des données et le graphe quantile-quantile . Sur l'histogramme en densité, on peut ajouter en rouge la loi normale (ce qu'on n'exige pas de vous !). On obtient les deux graphes de la figure 6.1. Le graphe quantile-quantile nous montre un bon accord avec la loi normale.
- On utilise la méthode de la définition 6.33 page 80 : Sur cet échantillon échantillon, on observe (retrouvez vous-même ces grandeurs !)

  $$\begin{aligned} n &= 16, \\ m &= 15.633125, \\ sd &= 0.936744 \end{aligned}$$

  Ceci conduit à la valeurs suivante de t donnée par (6.28) :

  $$t = \frac{m - \mu_0}{sd/\sqrt{n}} = \frac{15.633125 - 17.35}{0.936744/\sqrt{16}}$$

  soit

  $$t = -7.33124416. \tag{6.36}$$

  Il faut donc calculer la probabilité pour une loi T de Student à ddl=15 degrés de liberté que $T < -7.33124416$. On cherche donc $P(T \leq t) = P(T \leq -7.33124416)$. On procédant comme d'habitude avec RCmdr (avec l'aire à gauche), on trouve une probabilité égale à

  $$p_c = 1.2401284e-06. \tag{6.37}$$

  Cette quantité est inférieure au seuil conventionnel de $\alpha = 0.05$. Ainsi, on rejettera H_0 et donc on décidera que l'hypothèse alternative est préférable (ici $H_{a(2)}$) et donc que les joueurs du groupe ont une moyenne une VMA plus faible que le groupe de Clairefontaine.

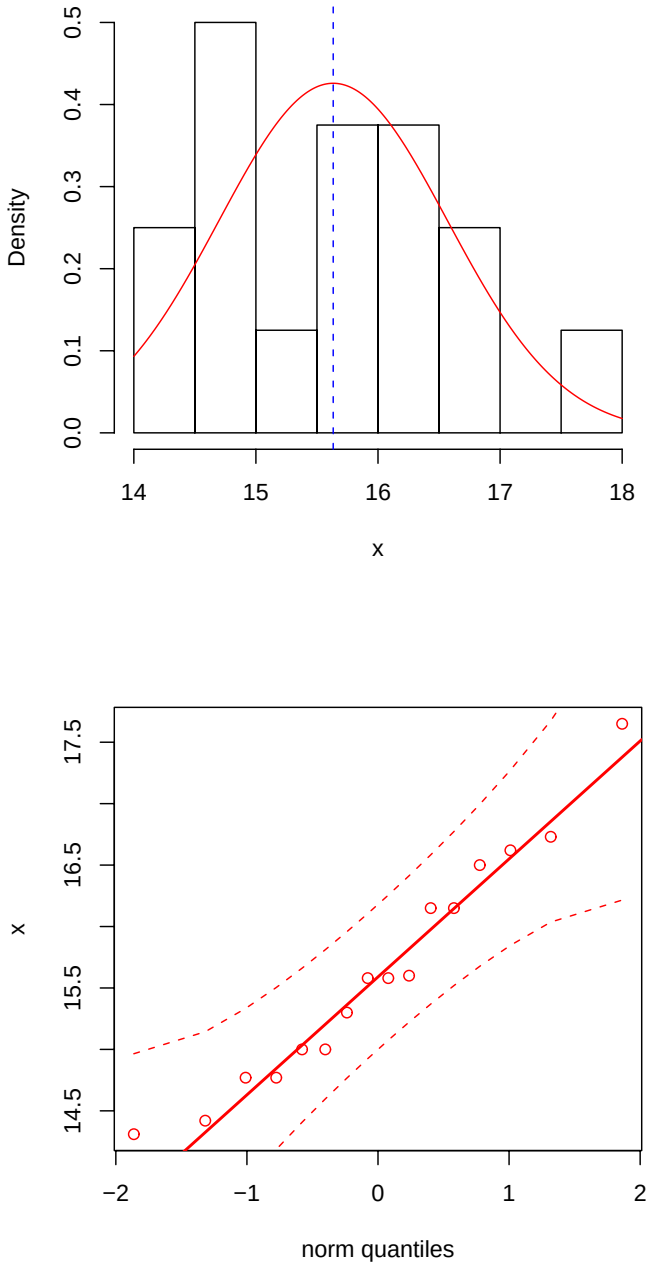


FIG. 6.11. Histogramme (en densité) et graphe quantile-quantile sur les données VMA de 'Coquelin.txt'.

REMARQUE 6.37. Les calculs précédents se réalisent automatiquement en employant le menu déroulant "Statistiques", et les options "Moyennes" et "t-test univarié". il faut alors indiquer la valeur de la moyenne μ_0 et l'une des trois hypothèses alternatives déjà citées. Attention, il y a un petit bug d'affichage dans la fenêtre de dialogue ; il faut lire pour la première hypothèse alternative "moyenne de la population" $\neq mu_0$. Cela revient

aussi à procéder comme dans la remarque 6.25 page 76 à condition de choisir l'hypothèse "la moyenne = μ_0". Le niveau de confiance NC correspond naturellemment à $1-\alpha$, égal classiquement à 0.95.

EXEMPLE 6.38. Reprenons le calcul précédent sur la VMA de Coquelin avec la méthode de la définition 6.37 page précédente. En choisissant comme hypothèse alternative $H_{a(2)}$ ($\mu < \mu_0$), la manipulation avec Rcdmr fournit :

```
        One Sample t-test

data:  coquelin$VMA
t = -7.3312, df = 15, p-value = 1.240e-06
alternative hypothesis: true mean is less than 17.35
95 percent confidence interval:
     -Inf 16.04366
sample estimates:
mean of x
 15.63312
```

Ici, on voit apparaître la valeur de la statistique t donnée par (6.28) qui vaut ici

$$t = -7.3312,$$

qui correspond bien à la valeur (6.36). On voit aussi la valeur de la probabilité critique p_c qui vaut ici

$$p_c = 1.24e-06,$$

qui correspond bien à la valeur (6.37). On retrouve aussi $ddl = 15$ et la moyenne estimée $m = 15.63312$. Un intervalle de confiance est aussi fourni par :

$$I_c =]-\infty, 16.04366,]$$

Il est en fait rarement utilisé.

EXEMPLE 6.39 (facultatif). Reprenons le calcul de l'exemple 6.38 avec la méthode de la définition 6.36 page 81. Pour α=0.05, On définit le réel t_α par (6.32), en passant par les quartiles à droite de la loi de Student avec $ddl = 15$ et on obtient

$$t_\alpha = -1.75305$$

La région critique est donc donnée par (6.31) :

$$R_c =]-\infty, t_\alpha] =]-\infty, -1.75305] .$$

On définit la statistique t de nouveau grâce à (6.28) :

$$t = -7.3312,$$

Puisque t appartient à la région critique, on rejette H_0 et donc, on retrouve comme précédemment, le fait que les joueurs du groupe ont une moyenne une VMA plus faible que le groupe de Clairefontaine.

EXERCICE 6.40. Reprendre les données du fichier `DAUCHEZ.txt` sur un groupe de joueurs et joueuses de tennis et leur résultat au test navette. Tester ce groupe par rapport à la norme fournie par la FFT de 12.75 pour ce test.

Comme dans l'exercice 4.3 page 15, on cherchera à montrer que la moyenne est inférieure à celle de la norme de la FFT.

Voir éléments de correction en page 89.

REMARQUE 6.41. On pourra consulter la page de Wikipédia `http://fr.wikipedia.org/wiki/Test_d%27hypoth%C3%A8se` (cherchez dans Wikipédia la rubrique "test d'hypothèse") pour avoir plus d'information sur les tests d'hypothèse.

6.5.4. Remarque sur le choix de l'hypothèse alternative

Le choix de l'hypothèse alternative doit de faire *a priori* et indépendamment des données. Dans l'exemple de la VMA, on suppose *a priori* qu'on "arrivera pas à depasser" la norme ; on veut donc démontrer que la moyenne est inférieure à la norme et on choisit donc l'hypothèse alternative $H_{a(2)}$. Dans l'exemple de l'exercice 6.40 page précédente, on raisonne de la même façon. Si on a aucun autre renseignement, le test le plus usuel sera le test bilatéral.

Attention, pour choisir l'hypothèse alternative, on peut être tenté, au vu des données de calculer la moyenne, de la comparer à la norme et de choisir si elle est plus petite l'hypothèse alternative $H_{a(2)}$: $\mu < \mu_0$ (et réciproquement). Il ne faut pas procéder ainsi, le test est en effet "biaisé", car fondé sur une connaissance *a fortiori* des données : on en forcera le résultat.

6.6. Retour sur les intervalles de confiance et test d'hypothèse en proportion

Vous pourrez maintenant lire la section 5.9 page 48 du chapitre 5.

6.7. Éléments de correction

Éléments de correction de l'exercice 6.8

(1) On utilise le menu déroulant "Distributions", puis "Distributions continues" puis "Distribution normale" puis "Probabilités normales". Il faut choisir dans la fenêtre de dialogue les valeurs par défaut de $mu = 0$ et $sigma = 1$. On obtient en utilisant "l'aire à gauche" :

$$P(X \leq -0.5) = 0.308538,$$
$$P(X \leq 4.5) = 0.999997.$$

En passant par "l'aire à droite", on obtient

$$P(X \geq 1.25) = 0.10565,$$
$$P(X \geq -2) = 0.97725.$$

Ces dernières peuvent aussi être obtenue par "l'aire à gauche" :

$$P(X \geq 1.25) = 1 - 0.89435 = 0.10565,$$
$$P(X \geq -2) = 1 - 0.02275 = 0.97725.$$

(2) De même, on obtient en utilisant (6.8) (en passant par "l'aire à gauche") :

$$P(1.25 \leq X \leq 1.5) = P(X \leq 1.5) - P(X \leq 1.25) = 0.933193 - 0.89435 = 0.038843,$$
$$P(-0.65 \leq X \leq 1.4) = P(X \leq 1.4) - P(X \leq -0.65) = 0.919243 - 0.257846 = 0.661397,$$

Éléments de correction de l'exercice 6.9

(1) On procède comme dans l'exercice 6.8 : on utilise le menu déroulant "Distributions", puis "Distributions continues" puis "Distribution normale" puis "Probabilités normales". Il faut alors choisir dans la fenêtre de dialogue les valeurs de $mu = 1.5$ et $sigma = 2$. On obtient en utilisant "l'aire à gauche" :

$$P(X \leq -0.5) = 0.158655254,$$
$$P(X \leq 4.5) = 0.933192799.$$

En passant par "l'aire à droite", on obtient

$$P(X \geq 1.25) = 0.549738225,$$
$$P(X \geq -2) = 0.959940843.$$

Ces dernières peuvent aussi être obtenue par "l'aire à gauche" :

$$P(X \geq 1.25) = 1 - 0.450261775 = 0.549738225,$$
$$P(X \geq -2) = 1 - 0.040059157 = 0.959940843.$$

(2) De même, on obtient en utilisant (6.8) (en passant par "l'aire à gauche") :

$$P(1.25 \leq X \leq 1.5) = P(X \leq 1.5) - P(X \leq 1.25) = 0.5 - 0.450261775 = 0.049738225,$$
$$P(-0.65 \leq X \leq 1.4) = P(X \leq 1.4) - P(X \leq -0.65) = 0.480061194 - 0.141187364 = 0.33887383.$$

Éléments de correction de l'exercice 6.12

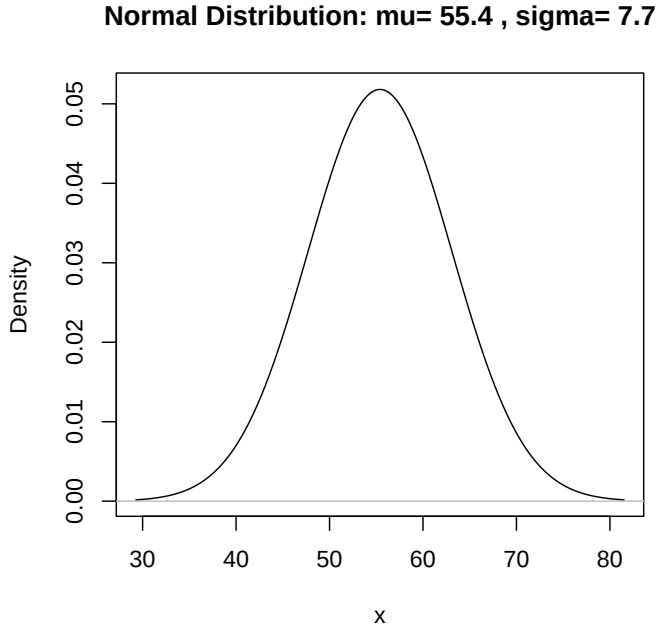


Fig. 6.12. La densité de la loi normale de moyenne $\mu = 55.4$ et d'écart-type $\sigma = 0.936744139737918$.

(1) Comme au début de la section 6.1.3 page 65, on obtient le graphe de la figure 6.12.

(2) On suppose que $\mathrm{VO_2max} \sim \mathcal{N}(55.4, 7.7)$. Comme dans l'exercice 6.9 page 68, on obtient

$$P(\mathrm{VO_2max} \geq 60) = 0.275119,$$
$$P(\mathrm{VO_2max} \geq 70) = 0.028973,$$
$$P(\mathrm{VO_2max} \leq 50) = 0.241558.$$

(3) De même, on obtient

$$P(\mathrm{VO_2max} \leq 45.3) = 0.094813.$$

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 6.19

D'après (6.19), on a

$$SEM = \frac{sd}{\sqrt{n}} = \frac{\widehat{\sigma}}{\sqrt{n}} = \frac{8.9}{\sqrt{12}} = 2.569209.$$

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 6.24

En faisant comme indiqué dans la proposition (6.22), en tapant par exemple

```
qt((1+0.9)/2, df=12-1)
```

on obtient pour le niveau de confiance NC=90%, $t = 1.7958848$. On obtient donc l'intervalle de confiance défini par

$$m - t\frac{sd}{\sqrt{n}} = 45.3 - 1.795885\frac{8.9}{\sqrt{12}} = 40.686,$$

et

$$m + t\frac{sd}{\sqrt{n}} = 45.3 + 1.795885\frac{8.9}{\sqrt{12}} = 49.914.$$

De même, pour le niveau de confiance NC=95%, on a $t = 2.2009852$ et donc l'intervalle de confiance défini par

$$m - t\frac{sd}{\sqrt{n}} = 45.3 - 2.200985\frac{8.9}{\sqrt{12}} = 39.645,$$

et

$$m + t\frac{sd}{\sqrt{n}} = 45.3 + 2.200985\frac{8.9}{\sqrt{12}} = 50.955.$$

Enfin, pour le niveau de confiance NC=99%, on a $t = 3.1058065$ et donc l'intervalle de confiance défini par

$$m - t\frac{sd}{\sqrt{n}} = 45.3 - 3.105807\frac{8.9}{\sqrt{12}} = 37.321,$$

et

$$m + t\frac{sd}{\sqrt{n}} = 45.3 + 3.105807\frac{8.9}{\sqrt{12}} = 53.279.$$

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 6.27

(1)

(2) Voir la figure 6.13 qui représente à la fois l'histogramme et le graphe quantile-quantile (voir chapitre 3) : la courbe est à peu près en "cloche" (sauf peut-être quelques valeurs extrémales).

(3) (a) Grâce à Rcmdr, on obtient la moyenne et l'écart-type

$$m = 11.0674359, \quad sd = 0.6517294 \tag{6.38}$$

On fait comme indiqué dans la proposition (6.22) en tapant par exemple

```
qt((1+0.9)/2, df=39-1)
```

on obtient pour le niveau de confiance NC=90%, $t = 1.6859545$. On obtient donc l'intervalle de confiance défini par

$$m - t\frac{sd}{\sqrt{n}} = 11.067436 - 1.685954\frac{0.651729}{\sqrt{39}} = 10.891, \tag{6.39a}$$

et

$$m + t\frac{sd}{\sqrt{n}} = 11.067436 + 1.685954\frac{0.651729}{\sqrt{39}} = 11.243. \tag{6.39b}$$

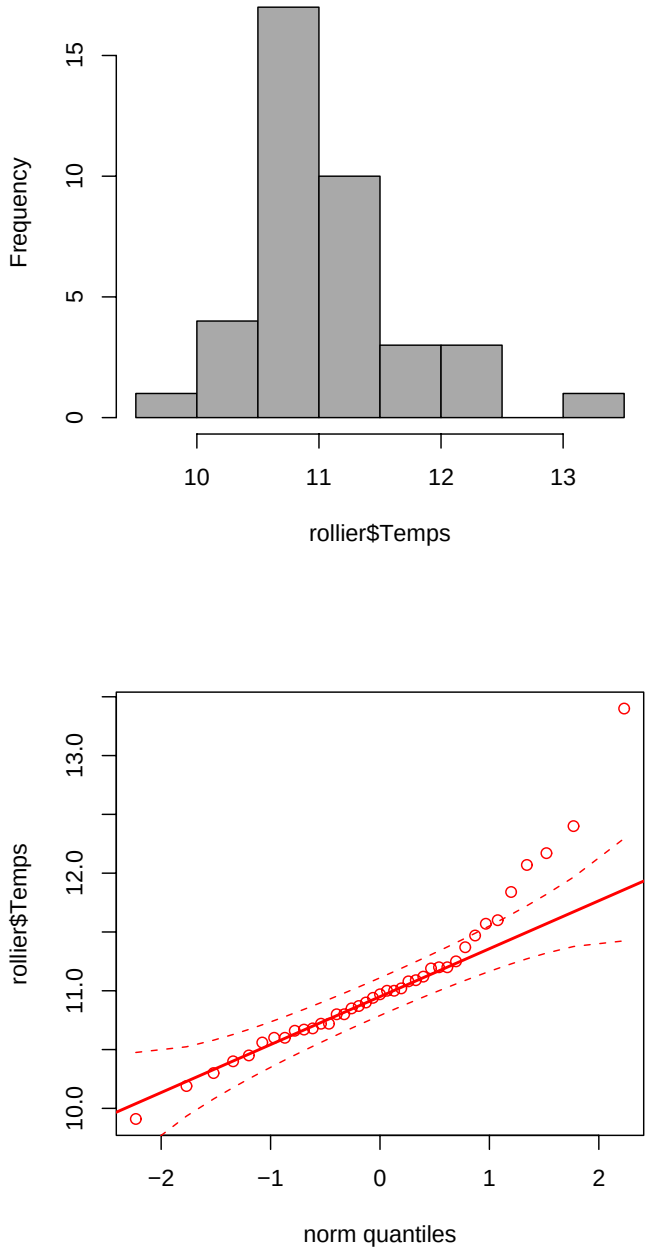


FIG. 6.13. Histogramme et graphe quantile-quantile sur les données de Temps de ROLLIER.txt.

(b) Utilisons la méthode suggérée dans la remarque 6.25 page 76. On obtient alors

```
        One Sample t-test

data:  rollier$Temps
```

```
t = 106.0503, df = 38, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 10.89149 11.24338
sample estimates:
mean of x
 11.06744
```

c'est-à-dire un intervalle de confiance donné par $[10.89149, 11.24338]$ qui est bien celui donné par (6.39) et une moyenne de 11.067436 qui est bien celle donnée par (6.38).

(c) Si on utilise la fonction la méthode de la remarque 6.26 page 76, c'est-à-dire l'utilisation de la fonction `int.conf.moy.R` on tape

```
int.conf.moy(11.067436,0.651729,11.067436,39,0.9)
```

ou directement

```
mu <- mean(rollier$Temps)
sigma <- sd(rollier$Temps)
n <- length(rollier$Temps)
int.conf.moy(mu, sigma, n, 0.9)
```

et on obtient un intervalle de confiance donné par

```
[1] 10.89149 11.24338
```

qui est bien celui donné par (6.39).

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 6.28
On trouve successivement

(1) pour $n = 25$, NC =95 % : t=2.0638986 ;

(2) pour $n = 25$, NC =90 % : t=1.7108821 ;

(3) pour $n = 15$, NC =99 % : t=2.9768427 ;

(4) pour $n = 15$, NC =98 % : t=2.6244941.

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 6.40

- On trace l'histogramme, en densité, des données et le graphe quantile-quantile . Sur l'histogramme en densité, on peut ajouter en rouge la loi normale. On obtient les deux graphes de la figure 6.14. Le graphe quantile-quantile nous montre un bon accord avec la loi normale.
- En procédant comme dans la remarque 6.37 page 83, (avec un niveau de signification α=0.05 donc un NC égal à 0.95) on trouve les résultats suivants avec Rcmdr :

```
        One Sample t-test

data:  data
t = -5.1188, df = 9, p-value = 0.0003145
alternative hypothesis: true mean is less than 12.75
95 percent confidence interval:
     -Inf 10.92063
sample estimates:
mean of x
      9.9
```

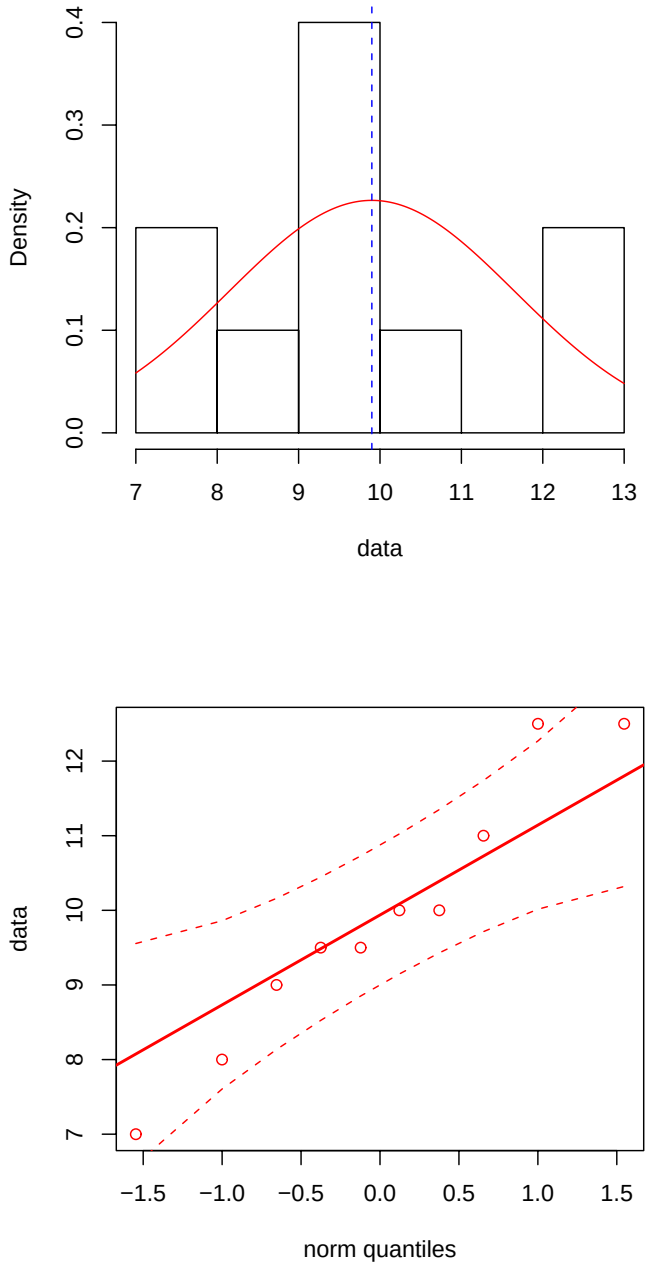


FIG. 6.14. Histogramme (en densité) et graphe quantile-quantile sur les données Navette de 'DAUCHEZ.txt'.

- Si on utilise la méthode usuelle de la section 6.5.2.1 page 80, on écrira : "
  On procède au *test de Student pour un échantillon* (à la moyenne).

On fait l'hypothèse nulle H_0 : $\mu = \mu_0$. avec $\mu_0 = 12.75$. On cherche à montrer que la moyenne de la loi normale, dont proviendraient les données de l'échantillon étudié, est plus petite que μ_0. On fait donc l'hypothèse alternative suivante : H_1 : $\mu < \mu_0$.

Grâce à R, on trouve la valeur suivante de la statistique

$$t = \frac{m - \mu_0}{sd/\sqrt{n}} = -5.118751$$

La probabilité critique $P(T \leq t)$ (pour la loi de Student à $ddl = 9$ degrés de libertés) est égale à

$$p_c = 0.000314518$$

Puisque p_c est inférieure au égal au niveau de signification $\alpha = 0.05$, on rejette l'hypothèse nulle H_0. Ainsi, H_1 est vraie et *la moyenne est plus petite que* $\mu_0 = 12.75$, au risque 0.05. "

- Si on utilise la méthode de la section 6.5.2.2 page 81 (facultative), on écrira : "

  On procède au *test de Student pour un échantillon* (à la moyenne).

  On fait l'hypothèse nulle H_0 : $\mu = \mu_0$. avec $\mu_0 = 12.75$. On cherche à montrer que la moyenne de la loi normale, dont proviendraient les données de l'échantillon étudié, est plus petite que μ_0. On fait donc l'hypothèse alternative suivante : H_1 : $\mu < \mu_0$.

  Grâce à R, on trouve la valeur suivante de la statistique

$$t = \frac{m - \mu_0}{sd/\sqrt{n}} = -5.118751$$

  Le réel t_α défini par $P(T \leq t_\alpha) = \alpha$ (pour la loi de Student à $ddl = 9$ degrés de libertés) est égal à

$$t_\alpha = -1.833113$$

  La région critique est donc égale à

$$R_c =]-\infty, t_\alpha] =]-\infty, -1.833113].$$

  Puisque cette région critique contient la statistique $t = -5.118751$, on rejette l'hypothèse nulle H_0. Ainsi, H_1 est vraie et *la moyenne est plus petite que* $\mu_0 = 12.75$, au risque 0.05. "

CHAPITRE 7

Mesurer la progression d'un groupe (données numériques appariées)

Avec un groupe de 12 nageuses de natation synchronisée (14 à 17 ans), C. Tolleron (M2PPMR) a proposé un cycle de développement de la souplesse. Cette caractéristique est testée par évaluation du grand écart droit au sol (c'est-à-dire la distance entre le pubis et le sol). Afin de mesurer l'évolution, une première mesure a été prise le 23 septembre avant le cycle de souplesse et une autre le 28 octobre à l'issue du cycle.

7.1. Compréhension de la différence entre échantillons appariés (mesures répétées ou blocs) et échantillons indépendants

De telles mesures sont dites *mesures répétées*. Ceci signifie que sur les mêmes individus plusieurs mesures sont effectuées soit à des dates différentes (mesures longitudinales) soit dans des conditions différentes. Il se peut aussi que des individus différents soient regroupés deux par deux sur des caractéristiques telles que le sexe, l'âge, les performances passées et que l'on souhaite comparer leurs performances dans différentes conditions. On parle alors de dispositifs en blocs.

Les deux situations (deux mesures sur les mêmes individus ou même mesure sur deux individus différents) sont similaires, on parle de *mesures appariées*. On va prendre en compte une proximité attendue entre les résultats de chaque paire afin d'obtenir des résultats plus précis.

7.2. Graphiques pour échantillons appariés

7.2.1. Le graphe parallèle

Le graphique le plus efficace est sans doute le *graphe parallèle*. Ce graphe n'est malheureusement pas inclus dans Rcmdr et demande une programmation. En revanche, Stéphane Champely a créé une fonction R qui se nomme `parallelplot.R`.

Téléchargez et sourcez cette fonction comme indiqué dans l'annexe D page 133. Pour obtenir le graphe de la figure 7.1 page suivante, on peut taper la commande (une fois que l'on a chargé le fichier `TOLLERON.txt`) :

```
parallelplot(tolleron$Début, tolleron$Fin, glab = c("Début", "Fin"),
    ylab = "Souplesse")
```

Il est très clair que pour toutes les nageuses sauf une (ligne en rouge), l'écart a diminué, c'est-à-dire que la souplesse a progressé.

7.2.2. Le graphe des différences

Une autre approche graphique va être utilisable qui va nous conduire directement aux techniques statistiques inférentielles que nous utiliserons. Il s'agit de s'intéresser aux différences entre les deux mesures, si ces différences sont globalement positives, il y a augmentation de la mesure d'une date à l'autre, si elles sont négatives il y a bien sûr diminution.

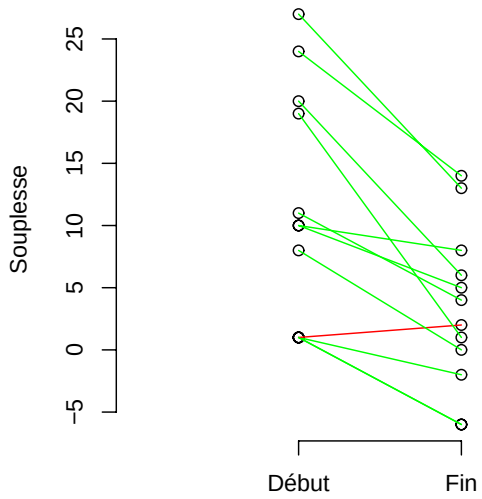


FIG. 7.1. Graphe parallèle des souplesses en début de cycle et fin de cycle d'entraînement pour l'étude de Tolleron.

Pour réaliser cette étude, il faut employer dans Rcmdr le menu déroulant "Données", l'option "Gérer les variables dans le jeu de données actif", "Calculer une nouvelle variable". Dans la fenêtre de dialogue, employer en l'espèce comme nom de la nouvelle variable : Diff et comme "Expression à calculer" :

```
Début-Fin
```

Tous les graphiques adaptés à l'étude d'une variable numérique sont alors envisageable, la norme à considérer étant la valeur zéro qui forme la limite entre augmentation et diminution. On pourra en particulier trace une ligne de point (voir page 7) et rajouter sur le graphe une ligne verticale (voir page 15 dans la section 4.1 du chapitre 4).

On pourra tracer le graphique de la figure 7.2 page ci-contre, en tapant

```
stripchart(tolleron$Diff, method = "stack")
abline(v = 0, col = "red", lty = 2)
```

voire mieux en tapant directement

```
stripchart(tolleron$Début - tolleron$Fin, method = "stack")
abline(v = 0, col = "red", lty = 2)
```

Sur la figure 7.2 page suivante, on constate que la différence entre la valeur en début de cycle et la valeur en fin est généralement positive, c'est-à-dire que la mesure de souplesse a diminuée, ce qui indique une augmentation de la qualité de souplesse !

EXERCICE 7.1. G. Compassi (M2PPMR) a entraîné un groupe de 12 judokas inscrits dans le pôle espoir de Lyon. La mesure de performance est celle du tirage planche (kg) (c'est-à-dire la mesure d'une charge qui peut être portée). La première évaluation a eu lieu le 12 septembre, un entraînement en hypertrophie a été suivi, puis une évaluation finale a eu lieu le 7 novembre. Les données sont disponibles dans le fichier `COMPASSI.txt`.

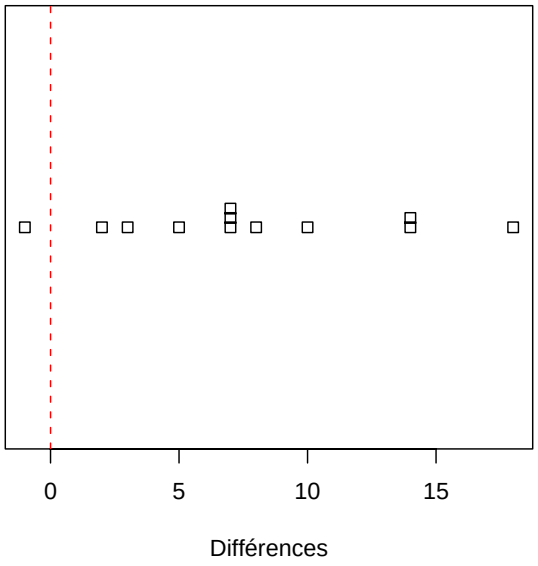


FIG. 7.2. Ligne de points des différences de souplesse entre le début de cycle et la fin de cycle d'entraînement pour l'étude de Tolleron

Réaliser un graphe adapté à l'étude de la progression. Qu'en concluez-vous sur l'efficacité de l'entraînement ? Comment préciser les résultats ?

7.3. taille d'effet

L'hypothèse de référence est qu'il n'y a pas de différence entre les deux mesures (x=début et y=fin d'entraînement). L'effet recherché est l'existence d'une différence.

La taille d'effet va donc être soit absolue en mesurant la différence entre les moyennes des deux mesures

$$m_d = m_x - m_y \tag{7.1}$$

soit relative en rapportant cette différence à un écart-type. Il y a ici deux possibilités, l'écart-type peut être soit l'écart-type des différences sd_d ($d = x - y$) soit l'écart-type de la mesure de référence, ici la mesure initiale c'est-à-dire sd_x.

Les deux tailles d'effet standardisées envisageables sont donc :

$$d_d = \frac{m_x - m_y}{sd_d} = \frac{m_d}{sd_d} \tag{7.2}$$

et

$$d_x = \frac{m_x - m_y}{sd_x} = \frac{m_d}{sd_x}. \tag{7.3}$$

Nous préférons employer la quantité d_d proposée par Cohen [**Coh98**] : on utilise un classement qualitatif de ces tailles d'effet sur la base de la valeur de $|d_d|$ par rapport à trois valeurs $d_1 = 0.2$, $d_2 = 0.5$ et $d_3 = 0.8$:

$$\text{si } |d_d| \begin{cases} < d_1, & \text{l'effet est faible,} \\ \in [d_1, d_2[, & \text{l'effet est moyenne,} \\ \in [d_2, d_3[, & \text{l'effet est fort,} \\ > d_3, & \text{l'effet est très fort} \end{cases} \tag{7.4}$$

On obtient les statistiques nécessaires à ces calculs à l'aide du menu déroulant "Statistiques", l'option "Résumés", puis "Statistiques descriptives", en sélectionnant les trois variables (x=Début, y=Fin et d=Diff). On peut lire alors entre autres les renseignements suivants sur les données de Tolleron (à faire vous-même !) :

$$\begin{aligned} m_x &= 11.083333, \\ m_y &= 3.25, \\ m_d &= 7.833333, \\ sd_x &= 9.443211, \\ sd_d &= 5.474459. \end{aligned}$$

On obtient donc comme taille d'effet absolu

$$m_d = 7.833333,$$

et comme tailles d'effet relatives

$$\begin{aligned} d_d &= \frac{m_d}{sd_d} = \frac{7.833333}{5.474459} = 1.430887, \\ d_x &= \frac{m_d}{sd_x} = \frac{7.833333}{9.443211} = 0.82952. \end{aligned}$$

Nous avons incontestablement affaire à un effet de grande taille.

EXERCICE 7.2. Quelles tailles d'effet peut-on associer à l'entraînement en hypertrophie de Compassi sur les judokas (fichier `COMPASSI.txt`) ?

7.4. Intervalles de confiance

On fera l'hypothèse que les différences suivent une loi normale de moyenne et d'écart-type inconnus. Comme dans le chapitre 6, on cherche d'une part à trouver un intervalle de confiance de la moyenne et un test d'hypothèse par rapport à la moyenne. Il est important de comprendre que l'on applique les thechniques vues au chapitre 6 *aux différences* et qu'on s'intéresse ici *au signe de la moyenne des différences* ; en effet, si la moyenne des différences est positive, la moyenne du premier groupe est supérieure à celle du second.

L'anlogue de la proposition 6.22 page 75 est donc :

PROPOSITION 7.3. *L'intervalle de confiance au niveau NC de la différence entre les espérances mathématiques de deux populations appariées est donné par*

$$\left[m_d - t\frac{sd_d}{\sqrt{n}}, m_d + t\frac{sd_d}{\sqrt{n}} \right], \tag{7.5}$$

où t est le quantile d'une loi de Student à $n-1$ degrés de liberté correspondant à la probabilité $q = (1+NC)/2$ (il est obtenu exactement comme dans la proposition 6.22), n est la taille de l'échantillon, m_d est la moyenne des différences et sd_d l'écart-type des différences.

REMARQUE 7.4. Pour réaliser ce calcul, on procède d'une façon analogue à celle de la remarque 6.25 page 76 : utiliser le menu déroulant "Statistiques", l'option"Moyennes" puis "t-test apparié". Dans la fenêtre de dialogue, déclarer comme première variable X=Début et seconde variable Y=Fin. Il faut laisser les champs relatifs aux hypothèses avec les valeurs définies par défaut.

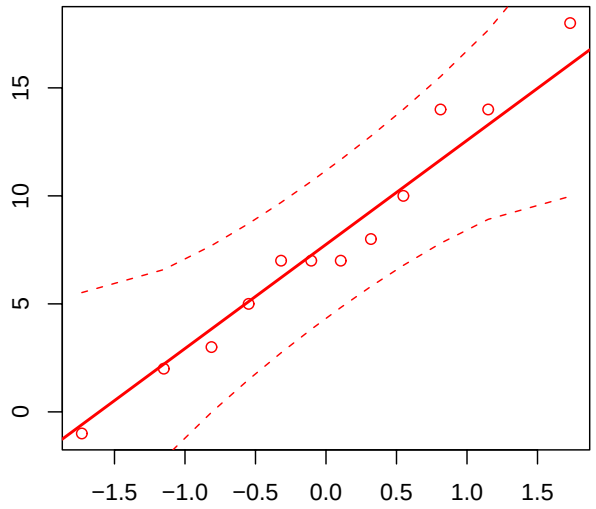


FIG. 7.3. graphe quantile-quantile (comparaison avec la loi normale) pour l'étude de Tolleron.

EXEMPLE 7.5. Pour les donnés de Tolleron, on peut, pour vérifier la normalités des différences, tracer le graphe quantile-quantile de la différence, pour obtenir le graphe de la figure 7.3. On procédant comme dans la remarque 6.25 page 76, on obtient avec $NC = 0.95$,

```
        Paired t-test

data:  tolleron$Début and tolleron$Fin
t = 4.9567, df = 11, p-value = 0.000431
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.355028 11.311638
sample estimates:
mean of the differences
               7.833333
```

Nous obtenons donc ici un intervalle au niveau $NC = 0.95$ s'étendant de 4.355028 à 11.311638. Ici, c'est un intervalle de confiance pour la différence des moyennes, qui est donc positive ; ainsi la souplesse a, en moyenne, diminué : l'entrainement semble donc être bénéfique !

EXEMPLE 7.6. On peut aussi procéder à un test univarié sur la différence des données (variable 'Diff' ou directement sur la différence calculée) ; on obtient

```
        One Sample t-test

data:  tolleron$Début - tolleron$Fin
t = 4.9567, df = 11, p-value = 0.000431
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  4.355028 11.311638
sample estimates:
mean of x
 7.833333
```

ce qui est bien identique à l'exemple 7.5 page précédente !

EXERCICE 7.7. Calculer un intervalle de confiance au niveau $NC = 90\%$ de la progression au tirage plache suite à l'entraînement en hypertrophie de Compassi sur les judokas (fichier `COMPASSI.txt`).

Voir éléments de correction page 100.

7.5. Test d'hypothèses

Nous adopterons maintenant l'approche par les probabilités critiques (section 6.5.2.1 page 80) au détriment de celle par les régions critiques (section 6.5.2.2 page 81). Notre démarche est donc tout à fait identique à celle de la section 6.5.2.1, *adpatée au test sur la différence*.

Donnons maintenant l'équivalent de la définition 6.33 page 80, *adpatée au test sur la différence* avec $\mu_0 = 0$.

On se donne un seuil de signification α, souvent égal à 0.05.

DÉFINITION 7.8. Deux populations sont étudiées de façon appariées. On suppose que la différence entre les deux mesures $d = x - y$ suit un modèle probabiliste normal. Pour tester l'hypothèse que la différence des moyennes est nulle :

$$H_0 : \quad \mu_x - \mu_y = 0,$$

on utilise *le test de Student apparié*. On calcule la moyenne et l'écart-type m_d et sd des différences $x - y$ sur l'échantillons de taille n puis le score normalisé suivant qui servira de statistique de test :

$$t = \frac{m_d}{sd_d/\sqrt{n}}. \tag{7.6}$$

La probabilité critique de l'hypothèse nulle p_c :

- contre l'hypothèse $H_{a(1)}$: $\mu_x - \mu_y > 0$ est égale à la probabilité que la loi de Student à $n-1$ degrés de liberté soit plus élevée que t, c'est-à-dire $P(T \geq t)$;
- contre l'hypothèse $H_{a(2)}$: $\mu_x - \mu_y < 0$ est égale à la probabilité que la même loi soit plus petite que t, soit $P(T \leq t)$;
- contre l'hypothèse $H_{a(3)}$: $\mu_x - \mu_y \neq 0$ est égale à la probabilité que cette loi soit plus éloignée de zéro que t, c'est-à-dire $P(|T| \geq |t|)$, soit encore $2P(T \geq |t|)$.

Enfin,

- si $p_c \leq \alpha$, on rejettera H_0 donc on acceptera l'hypothèse alternative.
- si $p_c > \alpha$, on acceptera H_0.

REMARQUE 7.9. De la même façon que pour la remarque 6.35 page 81, il est important de se rappeler que si H_0 est vraie, *la statistique* $t = (m_d)/(sd_d/\sqrt{n})$ *suit une loi de Student à* $n-1$ *degrés de liberté*.

REMARQUE 7.10. Pour réaliser ce calcul, on procède d'une façon analogue à celle de la remarque 6.37 page 83 utiliser le menu déroulant "Statistiques", l'option"Moyennes" puis "t-test apparié". Dans la fenêtre de dialogue, choisir la première variable x et la seconde y. Le niveau de confiance NC correspond naturellemment

à $1-\alpha$, égal classiquement à 0.95. L'hypothèse alternative "Bilatéral" correspond à l'hypothèse alternative $H_{a(3)}$.

EXEMPLE 7.11. Reprenons les données de Tolleron et l'exemple 7.5 page 97.

•

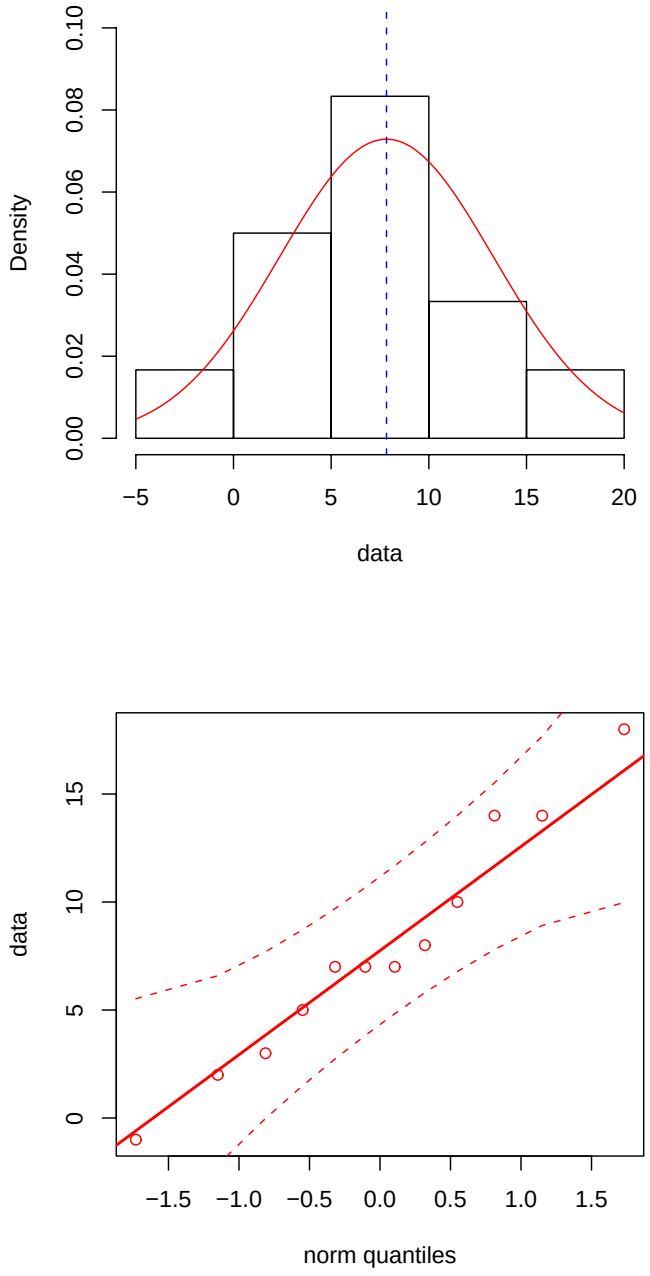


FIG. 7.4. Histogramme (en densité) et graphe quantile-quantile sur les différences de 'TOLLERON.txt'.

Comme dans le chapitre 6, on peut tracer l'histogramme, en densité, des différences et le graphe quantile-quantile . Sur l'histogramme en densité, on peut ajouter en rouge la loi normale. On obtient les deux graphes de la figure 7.4. Le graphe quantile-quantile nous montre un bon accord avec la loi normale.

- On procédant comme dans la remarque 6.37 page 83, on choisit $NC = 0.95$. En l'espèce avec la souplesse en natation synchronisée, il est logique de choisir l'hypothèse alternative $H_{a(1)} : \mu_x - \mu_y > 0$ car les mesures de la date finales doivent être inférieurs aux mesures de la date initiale (si l'entraînement a fonctionné). On obtient

```
        Paired t-test

data:  tolleron$Début and tolleron$Fin
t = 4.9567, df = 11, p-value = 0.0002155
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 4.995225      Inf
sample estimates:
mean of the differences
               7.833333
```

On obtient donc $t = 4.95674, 11$ degrés de liberté et $p_c = 0.0002155$ comme probabilité critique, inférieure à 0.05. Ainsi, on rejettera l'hypothèse nulle et donc la différence est positive. Les effets de l'entraînement sont donc statistiquement significatifs au seuil classique de 5%.

Exercice 7.12. Tester la progression au tirage planche suite à l'entraînement en hypertrophie de Compassi sur les judokas au seuil de 0.1. Attention à définir correctement l'hypothèse alternative.

Voir éléments de correction page 100.

7.6. Éléments de correction

Éléments de correction de l'exercice 7.7

On procède exactement comme dans dans la remarque 6.25 page 76 et l'exemple 7.5 page 97 : on obtient avec $NC = 0.9$,

```
        Paired t-test

data:  compassi$Septembre and compassi$Novembre
t = -6.3635, df = 11, p-value = 5.344e-05
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -10.791970  -6.041363
sample estimates:
mean of the differences
              -8.416667
```

Nous obtenons donc ici un intervalle au niveau $NC = 0.9$ s'étendant de -10.79197 à -6.041363. Ici, c'est un intervalle de confiance pour la différence des moyennes, qui est donc négative ; ainsi le tirage planche a, en moyenne, augmenté. L'entrainement semble donc être bénéfique !

Éléments de correction de l'exercice 7.12

Ici, contrairement à l'exercice 7.7, l'entraînement fonctionne sur la charge portée augmente. On cherchera donc à démontrer que la moyenne augmente.

•

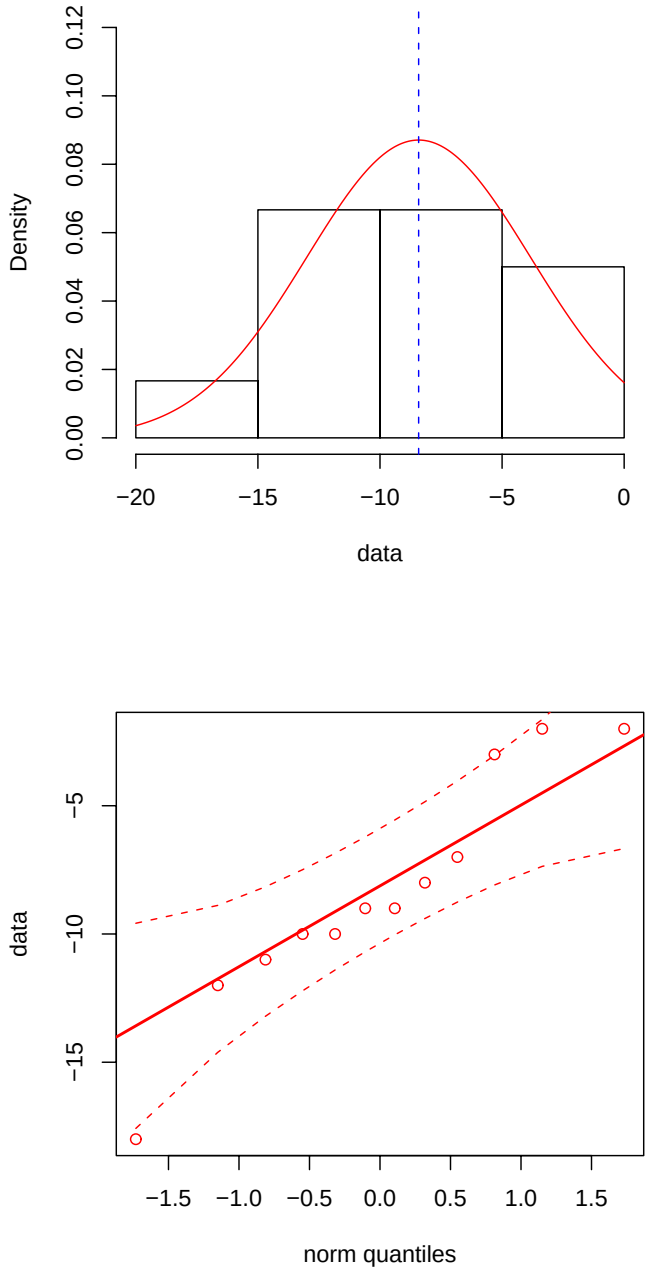


FIG. 7.5. Histogramme (en densité) et graphe quantile-quantile sur les différences de 'COMPASSI.txt'.

On peut tracer l'histogramme, en densité, des différences et le graphe quantile-quantile . Sur l'histogramme en densité, on peut ajouter en rouge la loi normale. On obtient les deux graphes de la figure 7.5. Le graphe quantile-quantile nous montre un bon accord avec la loi normale.

- On obtient

```
        Paired t-test

data:  compassi$Septembre and compassi$Novembre
t = -6.3635, df = 11, p-value = 2.672e-05
alternative hypothesis: true difference in means is less than 0
90 percent confidence interval:
      -Inf -6.613343
sample estimates:
mean of the differences
              -8.416667
```

On procède au *test de Student apparié* (à la différence des moyennes).

On fait l'hypothèse nulle H_0 : $\mu_x - \mu_y = 0$. On cherche à montrer que la moyenne de la loi normale, dont proviendraient les différences entre le premier et le second échantillon (qui sont appariés) est strictement négative. On fait donc l'hypothèse alternative suivante : $H_{a(2)}$: $\mu_x - \mu_y < 0$

Grâce à R, on trouve la valeur suivante de la statistique

$$t = \frac{m_d}{sd_d/\sqrt{n}} = -6.36355$$

La probabilité critique $P(T \leq t)$ (pour la loi de Student à $ddl = 11$ degrés de libertés) est égale à

$$p_c = 2.67176e - 05$$

Puisque p_c est inférieure au égal au niveau de signification $\alpha = 0.1$, on rejette l'hypothèse nulle H_0. Ainsi, H_1 est vraie et *la moyenne μ_x du premier échantillon est strictement inférieure à celle du second échantillon μ_y*, au risque 0.1. Les effets de l'entraînement sont donc statistiquement significatifs au seuil de 0.1.

CHAPITRE 8

Comparer les performances de deux groupes (deux échantillons numériques indépendants)

Chapitre non traité pour la saison : Automne 2009-2010.

CHAPITRE 9

Relation entre les caractéristiques anthropométriques et la performance (étude de régression linéaire)

Chapitre non traité pour la saison : Automne 2009-2010.

CHAPITRE 10

Récapitulatifs des notions essentielles

Vous trouverez dans ce chapitre l'essentiel (et l'exigible aux examens !) des notions, définitions, propriétés, exercices et manipulations avec R et Rcmdr qu'il faut savoir (ou retrouver dans le polycopié de cours). Ces notions sont présentées sous forme de listes, chapitre par chapitre, avec renvois aux points importants.

Compte tenu des modifications mineurs faites en cours de semestre, les numéros de pages indiquées peuvent avoir changé par rapport à la version papier distrinuée : il faut donc se référer au dernier document électronique de ce cours en pdf, disponible sur le web et sur le réseaux de l'université.

Chapitre 2

- La lecture d'un jeu de données : voir manipulations avec Rcmdr 2.1 page 3 et 2.2 page 3 ;
- Les graphiques statistiques pour les mesures catégorielles : voir section 2.3 page 4 ;
- Les statistiques pour mesures catégorielles : voir section 2.4 page 4.

Chapitre 3

- Les graphes pour données numériques : voir section 3.2 page 7 ;
- Les statistiques de centralité (moyenne et médiane) : voir section 3.3 page 9 ;
- Les statistiques de variabilité (quartiles et écart-type) : voir section 3.4 page 9 ;

Chapitre 4

- Manipulation avec Rcmdr 4.1 page 15 ;
- Manipulation avec R 4.2 page 15 ;
- La mesure de la taille de l'effet : voir section 4.2 page 16.

Chapitre 5

- Quelques simulations de dés : voir exercice 5.3 page 24 ;
- Notions de probabilités : voir définitions 5.10 page 27, 5.12 page 27 et éventuellement 5.13 page 27 ;
- Notion de variable aléatoire (discrète) : définition 5.15 page 27 et exemple 5.16 page 27 ;
- Espérance mathématique d'une variable aléatoire (discrète) : voir définition 5.18 page 28 et exemple 5.19 page 28 ;
- Variance et écart-type d'une variable aléatoire (discrète) : voir définitions 5.21 page 29 et 5.22 page 29 et exemple 5.23 page 29 ;
- Proposition 5.25 page 29 ;
- Le modèle binomial : voir définition 5.29 page 30 et exercice 5.30 page 30.
- La loi de probabilité binomiale : voir Proposition 5.32 page 31 et manipulation avec Rcmdr 5.34 page 31 et 5.35 page 32 ;
- Espérance mathématique, variance et écart-type binomiaux : voir Proposition 5.39 page 32, manipulation avec Rcmdr 5.41 page 33 et exercice 5.42 page 33 ;

- Les probabilités cumulés : Voir définitions 5.43 page 33, 5.45 page 33 et manipulations avec Rcmdr 5.44 page 33 et 5.48 page 34 ;
- Exercices 5.50 page 34 et 5.52 page 35 ;
- Paramètre et statistique : voir définition 5.53 page 35 ;
- Simulations d'échantillons binomiaux : manipulations avec Rcmdr 5.54 page 35 5.55 page 36 et 5.56 page 36 ;
- Proposition 5.57 page 37 ;
- Erreur standart de proportion (SEM) : voir définition 5.58 page 38 ;
- Intervalle de confiance "d'une proportion" au niveau de confiance $NC = 95\ \%$: voir définition 5.60 page 41 et manipulation avec 5.61 page 42 ;
- Intervalle de confiance "d'une proportion" à un niveau de confiance quelconque : voir définition 5.62 page 43, manipulation avec 5.67 page 44, et exercice 5.69 page 46.
- Test en Z d'une proportion : définition 5.73 page 48, exercice corrigé 5.74 page 48, et manipulation avec 5.75 page 48.

Chapitre 6

- Les données normales : voir section 6.1.1 page 63 ;
- Définition de probabilité continue : voir définition 6.1 page 63, équations 6.1 page 63, figure 6.3 page 66 et équation 6.2 page 63 ;
- Tracé de la loi normale centrée réduite (de moyenne nulle et d'écart-type 1) : voir manipulation avec Rcmdr 6.3 page 65 ;
- Tracé d'une loi normale quelconque : voir manipulation avec Rcmdr 6.6 page 66 ;
- Calculs de probabilités normales : voir manipulation avec Rcmdr 6.7 page 66 et exercices 6.8 page 67, 6.9 page 68 et 6.12 page 69 ;
- Simulations d'échantillons normaux : manipulations avec Rcmdr 6.16 page 70 et exercice 6.17 page 70 ;
- Erreur standart de la moyenne SEM : voir définition 6.18 page 73 et exercice 6.19 page 73 ;
- L'intervalle de confiance "d'une moyenne" (approché) : définition 6.20 page 74 ;
- L'intervalle de confiance "d'une moyenne" (exact avec la loi de Student) : définition 6.22 page 75 et exercice 6.24 page 76 ;
- Autre calcul de l'intervalle de confiance "d'une moyenne" : voir remarque 6.25 page 76 et 6.26 page 76 ;
- Exercice 6.27 page 77 ;
- Les tests d'hypothèses, introduction informelle : voir section 6.5.1 page 78 ;
- Les tests d'hypothèses : voir définitions 6.30 page 80, 6.32 page 80 et 6.33 page 80 ;
- Calculs pratiques : voir exemple de la section 6.5.3 page 82, remarque 6.37 page 83 et exercices 6.38 page 84 et 6.40 page 84.

Chapitre 7

- Graphique : voir section 7.2.1 page 93 ;
- Taille d'effet : voir section 7.3 page 95 ;
- Intervalle de confiance : définition 7.3 page 96, calcul avec la remarque 7.4 page 97 et exercice 7.7 page 98 ;
- Test d'hypothèse : voir définition 7.8 page 98, calcul de la remarque 7.10 page 98 et exercice 7.12 page 100.

Exercices de révision : chapitre 11

Tous les exercices !

CHAPITRE 11

Exercices de révisions

11.1. Énoncés

Exercice 11.1.
On étudie le fichier '`L3APA06.txt`'.
Analyser la variable '`sport`'.

Exercice 11.2.
On étudie le fichier '`L3APA06.txt`'.
Analyser la variable '`rythmcard`'.

Exercice 11.3.
On désire estimer la proportion de personnes qui se déclarent favorables a un certain projet dans une population de taille importante. On interroge finalement 1000 personnes et on trouve une proportion de 53 % dans l'échantillon.

(1) Proposez un intervalle de confiance de la proportion de personnes favorables au niveau de confiance de 99 %.

(2) Cet intervalle de confiance est-il exploitable ?

(3) Comment le rendre exploitable ?

Exercice 11.4.
Cet exercice a été donné au CT de statistique du M1PPMR (Janvier 2010).
D'après une expérience décrite dans *Newsweek* (7 mars 1994), 164 femmes séropositives ont été choisies au hasard pour recevoir un traitement par AZT lors de leur grossesse alors que 160 autres ont reçu un placebo. Sur les 164 mères traitées avec l'AZT, 13 ont eu un bébé séropositif à la naissance alors que dans le groupe placebo, 40 enfants étaient séropositifs sur 160.
Cette différence de proportion est-elle statistiquement significative ?

Exercice 11.5.
Le prix d'un certain nombre de cagettes est donné dans le fichier '`CAGETTE.txt`'.

(1) Proposer un intervalle de confiance du prix moyen au niveau de confiance 0.99.

(2) Peut-on en déduire sans aucun calcul que le prix moyen est différent de 10 ?

(3) Est-ce que, au seuil 0.01, le prix de ces cagettes est supérieur à 10 euros ?

Exercice 11.6.
Charles affirme que son score moyen au golf est 75. Charles a la réputation d'être un fieffé menteur. Vous l'observez pendant 9 parcours et vous calculez que, pour ces parcours, la moyenne est de 80 coups avec un écart type de 4 coups. On admet que son score au golf se distribue à peu près normalement. Au seuil de 0.01, qu'allez-vous conclure ?

Exercice 11.7.

Un ensemble de 15 personnes a fait un régime. Chargez le fichier de données `'REGIME.txt'`. Les poids avant le régime sont notés dans la colonne '`avant`' et ceux à la suite du régime sont notés dans la colonne '`après`'.

Au seuil 0.05, le régime a-t-il fonctionné ?

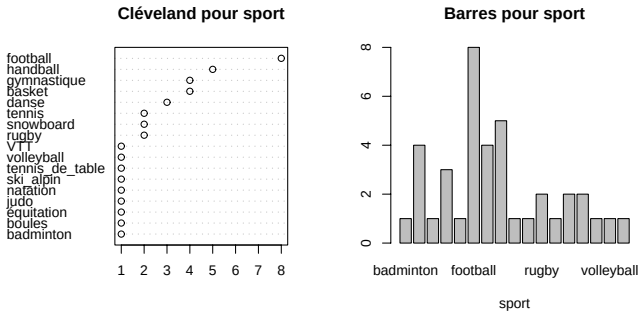
11.2. Corrigés

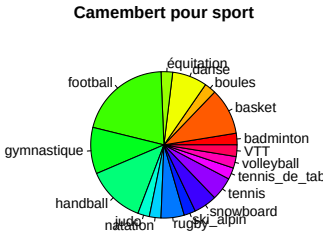
Éléments de correction de l'exercice 11.1

- On étudie la variable qualitative (ou catégorielle) '`sport`'. Pour les manipulations avec R, on renvoie donc aux sections 2.3 et 2.4 du document de cours.
- Les effectifs et les pourcentages déterminés par R sont donnés dans le tableau suivant

	effectifs	pourcentages
badminton	1	2.564
boules	1	2.564
équitation	1	2.564
judo	1	2.564
natation	1	2.564
ski_alpin	1	2.564
tennis_de_table	1	2.564
volleyball	1	2.564
VTT	1	2.564
rugby	2	5.128
snowboard	2	5.128
tennis	2	5.128
danse	3	7.692
basket	4	10.256
gymnastique	4	10.256
handball	5	12.821
football	8	20.513

-





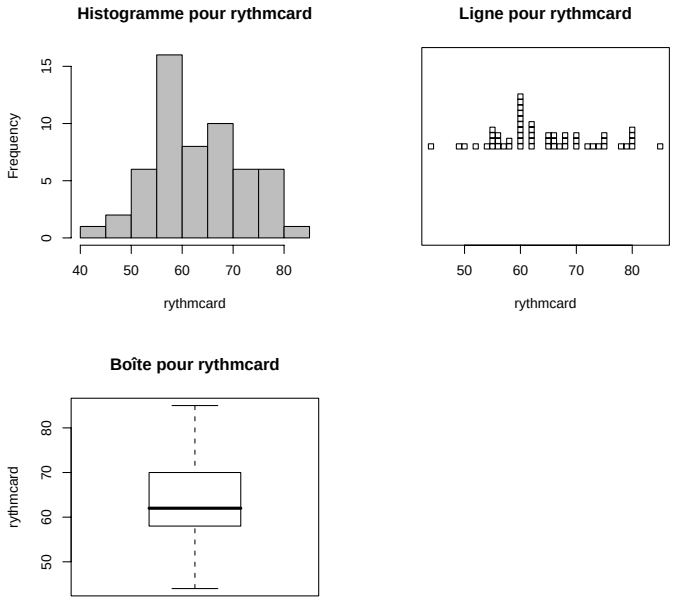
Voir les trois graphiques ci-dessus pour la variable '`sport`'.

Éléments de correction de l'exercice 11.2

- On étudie la variable quantitative (ou numérique) '`rythmcard`'. Pour les manipulations avec R, on renvoie donc aux sections 3.2, 3.3 et 3.4 du document de cours.
- Les différents résultats déterminés par R sont donnés dans le tableau suivant

noms	valeurs
moyenne	64.285714
sd	9.096881
Q_1 (quartile à 25 %)	58
médiane	62
Q_3 (quartile à 75 %)	70
minimum	44
maximum	85
nombre	58

-



Voir les trois graphiques ci-dessus pour la variable '`rythmcard`'.

Éléments de correction de l'exercice 11.3

(1) En utilisant, par exemple, la fonction `int.conf.prop.R`, on trouve l'intervalle de confiance

$$[0.48935, 0.57065]$$

(2) Ici, ce n'est guère exploitable, car cet intervalle contient la valeur 50 % !

(3) Pour le rendre exploitable, il faut le rendre plus petit ; pour cela, deux solutions : augmenter le nombre de personnes n ou diminuer le civeau de confiance NC. Par exemple, avec les données de l'énoncé et $NC = 0.9$, on obtient

$$[0.50404, 0.55596]$$

qui ne contient pas la valeur 50 % ! Avec les données de l'énoncé et $n = 10000$, on obtient

$$[0.51714, 0.54286]$$

qui ne contient pas la valeur 50 % !

Éléments de correction de l'exercice 11.4

- On renvoie à la section 5.9 page 48.
- On procède au *test Z d'une proportion.*

  On fait l'hypothèse nulle H_0 : $\pi = \pi_0$. avec $\pi_0 = 0.25$. On cherche à montrer que le paramètre π de la loi binomiale, dont proviendraient les données de l'échantillon étudié, est différente de π_0. On fait donc l'hypothèse alternative suivante : H_1 : $\pi \neq \pi_0$.

  Puisque $n = 164$, est "grand", on remplacera la loi binomiale de paramètre n et π_0 par la loi normale de moyenne $\mu = \pi_0$ et d'écart-type $\sqrt{\pi_0(1-\pi_0)/n}$. Grâce à R, on trouve la valeur suivante de la statistique

$$z = \frac{pr - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = -5.04935$$

  La probabilité critique $P(|Z| \geq |z|) = 2P(T \geq |z|)$ (pour la loi normale centrée réduite) est égale à

$$p_c = 4.43316e - 07.$$

  Puisque p_c est inférieure au égal au niveau de signification $\alpha = 0.05$, on rejette l'hypothèse nulle H_0. Ainsi, H_1 est vraie et *la moyenne est différente de* $\mu_0 = 0.25$, au risque 0.05.
- Le traitement à l'AZT semble donc efficace !

Éléments de correction de l'exercice 11.5

(1) Si on utilise par exemple la fonction '`t.test`' de R, la commande

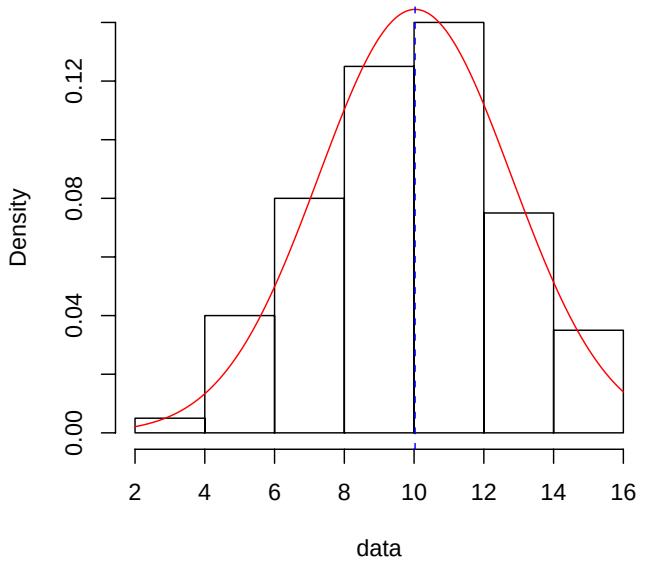
```
t.test(CAGETTE$cagette, conf.level=0.99)
```

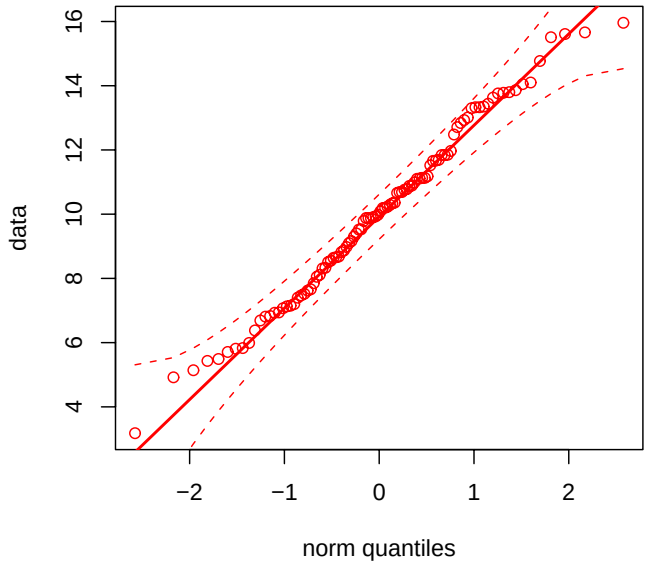
donne l'intervalle de confiance

$$NA$$

(2) L'intervalle de confiance contient la valeur 10 ; en reprenant le raisonnement de la section 6.5.1, on peut en déduire sans aucun calcul que le prix moyen est égal à 10 au seuil 0.01.

(3)
- Pour les manipulations sous R, on renvoie à la section 6.5.2.1 page 80.
-





On trace l'histogramme, en densité, des données et le graphe quantile-quantile . Sur l'histogramme en densité, on peut ajouter en rouge la loi normale. On obtient les deux graphes de la figure ci-dessus. Le graphe quantile-quantile nous montre un bon accord avec la loi normale.

- On procède au *test de Student pour un échantillon* (à la moyenne).
  On fait l'hypothèse nulle H_0 : $\mu = \mu_0$. avec $\mu_0 = 10$. On cherche à montrer que la moyenne de la loi normale, dont proviendraient les données de l'échantillon étudié, est plus grande que μ_0. On fait donc l'hypothèse alternative suivante : H_1 : $\mu > \mu_0$.

Grâce à ®, on trouve la valeur suivante de la statistique

$$t = \frac{m - \mu_0}{sd/\sqrt{n}} = 0.103551$$

La probabilité critique $P(T \geq t)$ (pour la loi de Student à $ddl = 99$ degrés de libertés) est égale à

$$p_c = 0.458867$$

Puisque p_c est strictement supérieure au niveau de signification $\alpha = 0.01$, on accepte l'hypothèse nulle H_0. Ainsi, H_0 est vraie et donc *la moyenne est égale à* $\mu_0 = 10$, au risque 0.01.

Éléments de correction de l'exercice 11.6

- Dans cet exercice, on va tester la valeur de la moyenne μ par rapport à la norme μ_0. Ici, on pense *a priori* que Charles ne dit pas la vérité (puisque il est estimé être un fieffé menteur). En tant que menteur, il à intérêt a se sur-estimer donc annoncer une moyenne $\mu_0 = 75$ "meilleure" que sa moyenne dont on suppose que provient son score $\mu = 80$ sur les $n = 9$ coups. Au golf, cela signifie qu'il annonce en fait que $\mu < \mu_0$. On veut montrer qu'il ment donc montrer que $\mu > \mu_0$.
- Pour les manipulations sous ®, on renvoie à la section 6.5.2.1 page 80.
- On procède au *test de Student pour un échantillon* (à la moyenne).

  On fait l'hypothèse nulle $H_0 : \quad \mu = \mu_0$. avec $\mu_0 = 75$. On cherche à montrer que la moyenne de la loi normale, dont proviendraient les données de l'échantillon étudié, est plus grande que μ_0. On fait donc l'hypothèse alternative suivante : $H_1 : \quad \mu > \mu_0$.

  Grâce à ®, on trouve la valeur suivante de la statistique

$$t = \frac{m - \mu_0}{sd/\sqrt{n}} = 3.75$$

La probabilité critique $P(T \geq t)$ (pour la loi de Student à $ddl = 8$ degrés de libertés) est égale à

$$p_c = 0.00281212$$

Puisque p_c est inférieure au égal au niveau de signification $\alpha = 0.01$, on rejette l'hypothèse nulle H_0. Ainsi, H_1 est vraie et *la moyenne est plus grande que* $\mu_0 = 75$, au risque 0.01.

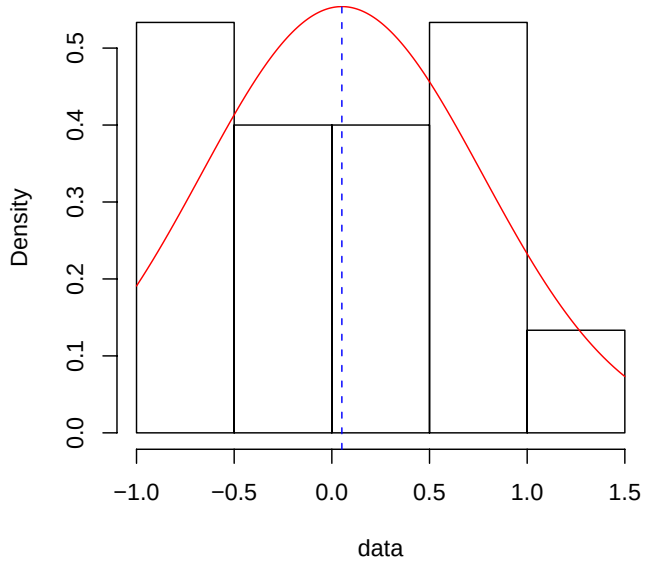
- Bref, Charles est bien un menteur (au seuil 0.01) !

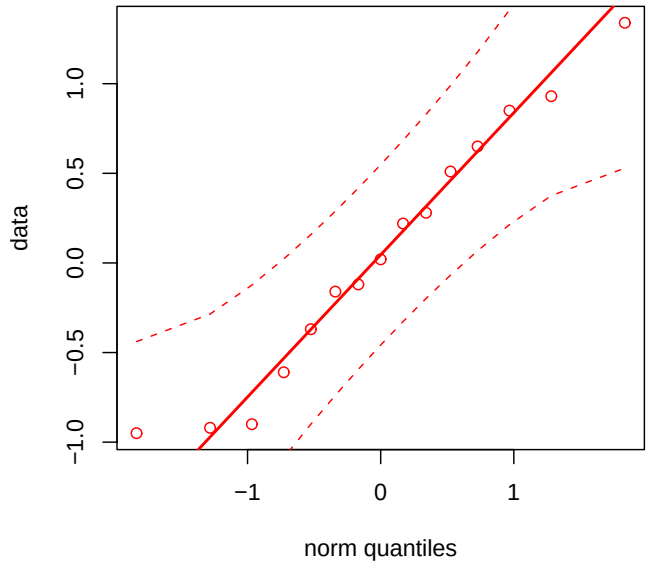
  Attention, il se peut aussi que Charles ne soit pas un menteur et que le jour où on l'observe se passe un événement de probabilité inférieure au seuil 0.01 qui a fait que l'on a rejetté à tort ses dires !

Éléments de correction de l'exercice 11.7

On veut montrer que le régime a fonctionné, c'est-à-dire que la moyenne du second groupe est inférieure à celle du premier groupe.

- Pour les manipulations sous ®, on renvoie à la remarque 7.4 page 97.
-





On peut tracer l'histogramme, en densité, des différences et le graphe quantile-quantile . Sur l'histogramme en densité, on peut ajouter en rouge la loi normale. On obtient les deux graphes de la figure ci-dessus.

Le graphe quantile-quantile nous montre un bon accord avec la loi normale.

- On procède au *test de Student apparié* (à la différence des moyennes).

On fait l'hypothèse nulle H_0 : $\mu_x - \mu_y = 0$. On cherche à montrer que la moyenne de la loi normale, dont proviendraient les différences entre le premier et le second échantillon (qui sont appariés) est strictement positive. On fait donc l'hypothèse alternative suivante : $H_{a(1)}$: $\mu_x - \mu_y > 0$.

Grâce à R, on trouve la valeur suivante de la statistique

$$t = \frac{m_d}{sd_d/\sqrt{n}} = 0.276061$$

La probabilité critique $P(T \geq t)$ (pour la loi de Student à $ddl = 14$ degrés de libertés) est égale à

$$p_c = 0.393267$$

Puisque p_c est strictement supérieure au niveau de signification $\alpha = 0.05$, on accepte l'hypothèse nulle H_0. Ainsi, H_0 est vraie et donc *la moyenne μ_x du premier échantillon est égale à celle du second échantillon μ_y*, au risque 0.05.

Les effets du régime ne sont donc pas statistiquement significatifs au seuil de 0.05.

ANNEXE A

Installation du logiciel R et du package Rcmd

A.1. Installation de R pour Windows

(1) Aller sur le site `http://www.r-project.org/`

(2) Cliquer sur "Download, Packages, CRAN", puis pour limiter le temps de téléchargement, choisir "France", `http://cran.univ-lyon1.fr/`.

(3) Dans la rubrique "Download and Install R", choisir (pour Windows ; bien entendu, sont aussi distribuées des versions pour Mac et Linux) Windows.

(4) Puis, cliquer sur "base Binaries for base distribution".

(5) Cliquer enfin sur "Download R-2.10.1 for Windows (32 megabytes)"

(6) Télécharger alors le logiciel d'installation `R-2.10.1-win32.exe` *Attention, le numéro de version de R est souvent réactualisé depuis la compilation de ce document !*

(7) Double-cliquer sur le logiciel `R-2.10.1-win32.exe` (ou la dernière version en date) afin de procéder à l'installation de R. Choisir les options par défaut afin de l'installer sur le disque `C:\`.

Remarque A.1. Dans la rubrique "R-2.10.1 for Windows" (ou la dernière en date), vous pouvez aussi télécharger les anciennes versions de R, parfois utiles quand les plus récentes peuvent être instables ou présenter un bug non encore résolus ! Voir "Previous releases". Dans cette même rubrique, vous pouvez aussi récupérer des packages das anciennes versions, non distribuées dans la plus récente.

Remarque A.2. Dans la rubrique "The Comprehensive R Archive Network", sous-rubrique "Source Code for all Platforms", puis "Contributed extension packages", vous pouvez aussi récupérer des packages des anciennes versions (sous forme de zip) , non distribuées dans la plus récente.

A.2. Utilisation de R

Utiliser le menu démarrer, puis Tous les programmes, puis R, puis R-2.10.1 (ou la dernière version en date). Le logiciel s'ouvre alors et une fenêtre "Rconsole" apparaît.

Il faut indiquer au logiciel R dans quel répertoire windows il doit aller chercher les fichiers (en particulier les jeux de données) dont nous avons besoin et où les sauvegarder également ; ce répertoire est dit *répertoire de travail*. Dans le menu déroulant "Fichier", existe une option "Changer de répertoire courant" qui par l'intermédiaire d'une arborescence permet de choisir le répertoire qui nous convient (par défaut c'est `C:\R\R-2.10.1`).

En quittant R, ne pas sauvegarder la session !

A.3. Installation et chargement du package Rcmdr

Le problème du logiciel R est qu'il s'agit d'un logiciel à langage de commandes, c'est à dire que pour l'utiliser, il faut taper des commandes dans la console, les valider pour obtenir des calculs ou des graphiques. Toutefois, il existe une version interactive employant des menus déroulants qui s'appelle Rcmdr que nous allons employer.

A.3.1. Installation de Rcmdr

Avant de commencer à utiliser ce package, il faut toutefois l'installer. La démarche suivante n'est donc à réaliser qu'une fois :

(1) L'un des menus déroulants de R s'appelle "Packages". Choisir dans ce menu l'option "Installer le(s) package(s)".

(2) Une fenêtre de dialogue s'ouvre qui vous propose un ensemble de site où vous pouvez chercher ce package. Choisir un site situé en France (Lyon devrait figurer sur la liste !).

(3) Une autre fenêtre s'ouvre qui propose une (longue) série de package, il faut choisir Rcmdr. Le téléchargement se produit alors automatiquement sur votre ordinateur.

A.3.2. Utilisation de Rcmdr

Pour utiliser le package Rcmdr (Une fois qu'il est installé...), il suffit d'aller dans le menu déroulant "Package" et de choisir l'option "Charger le package". Il faut choisir dans la liste des packages déjà installés sur votre ordinateur Rcmdr. *Attention*, à la première utilisation, R vous prévient qu'il manque des packages dont Rcmdr a besoin ; il faut répondre "OK" et laisser les champs par défaut ; un grand nombre de packages sont alors téléchargés et installés automatiquement. Pour les fois suivantes, cela ne produit plus ! Une fenêtre s'ouvre alors qui s'appelle "R commander". On utilise alors les menus déroulants pour choisir différentes options (charger un fichier, calculer des statistiques, réaliser des graphiques, ...). Il faut noter que le résultat des actions s'inscrit dans la fenêtre du bas du R commander qui s'appelle "Fenêtre de sortie". En revanche, les graphiques apparaissent dans la fenêtre habituelle de R dite "RGui". Il faut donc jongler entre "RGui" et "Rcommander" (on sy habitue).

Enfin on peut noter que lorsqu'une action est choisie par un menu déroulant, des lignes s'inscrivent dans la fenêtre dite "Fenêtre de script". Il s'agit des commandes réelles que R exécute (et dont on voit le résultat dans la fenêtre de sortie). Il se peut que dans certains cas (très rares), les menus déroulants ne soient pas suffisants, nous entrerons alors directement des commandes soit dans cette fenêtre de script soit dans la fenêtre de "Rgui" pour les exécuter.

On pourra consulter la doc en pdf `Getting-Started-with-the-Rcmdr.pdf` disponible normalement dans votre ordinateur (si vous y avez installé R), à l'adresse habituelle du site de ce cours ou dans l'ordinateur de l'université (dans le répertoire où R est installé, en général `C:\Program Files`) dans le répertoire `\R\R-2.10.1\library\Rcmdr\doc`

ANNEXE B

Prise en main à la première séance

Cette annexe est destinée à ceux qui se sentent peu habitués aux opérations de téléchargement de fichiers, de démarrage de logociels et pourra être lue en première séance.

B.1. Création d'un dossier de travail (ou répertoire courant)

Il est nécessaire de créer un dossier de travail pour stocker le polycopié de cours et les fichiers de données. Pour cela,

(1) Ouvrez un "Explorateur Windows" ou dans "poste de travail", allez dans le répertoire `W:`. Ce répertoire vous est propre et vous y aurez accès à chaque ouverture de session (avec vos propres identifiants).

(2) Créez-y un dossier (ou répertoire), par exemple appelé "statistiques".

Ce dossier constitue votre répertoire de travail ou répertoire courant.

B.2. Téléchargement du cours et des fichiers de données

Ce polycopié de cours et les fichiers de données sont normalement disponibles à la fois

- en ligne sur `http://utbmjb.chez-alice.fr/UFRSTAPS/index.html` à la rubrique habituelle ;
- en cas de problème internet, sur le réseau de l'université Lyon I : il faut aller sur :
  - 'Poste de travail',
  - puis sur le répertoire '`P:`' (appelé aussi : enseignants sur '`\Univ-lyon1\enseignement\homes`'),
  - puis '`jerome.bastien`',
  - enfin sur '`M1PPMR`'.

Pour l'examen, les données se trouveront aussi, par mesure de précaution à ces deux endroits.

(1) Rendez-vous sur donc soit sur internet soit (en cas de problème de connexion) sur le réseau et
   - ou bien sur internet, téléchargez dans votre répertoire de travail le polycopié de cours (rubrique "Version provisoire du cours" ou "Version définitive du cours"), grâce au clic droit "enregistrer sous"
   - ou bien sur le réseau, copiez-collez le polycopié de cours vers votre répertoire de travail.

(2) Faites de même pour les fichiers de données (disponibles soit sous la forme de fichiers txt ou xls, soit la forme d'un fichier "zipé").

(3) Dans votre répertoire courant, cliquez sur la version pdf du cours.

(4) Dans votre répertoire courant, dézipez éventuellement (clic droit, "extraire ici") les fichiers de données.

B.3. Installation du logiciel R et du package Rcmd

(1) Bouton "démarrer", puis "tous les programmes", puis "R".

(2) Déclarer le répertoire courant avec le menu déroulant "Fichier" puis l'option "Changer le répertoire courant", et indiquer le répertoire créé en section B.1.

(3) Il faut charger l'interface interactive de R en utilisant le menu déroulant "Packages" puis l'option "Charger le package", choisir alors le package Rcmdr (une interface graphique doit alors s'ouvrir).

ANNEXE C

Une toute petite introduction à la statistique descriptive (sans R)

C.1. Introduction

Cette annexe a pour objectifs de donner les notions de bases relatives à différents types de données. Il est conseillé de la lire sans utiliser d'ordinateurs (une petite calculatrice suffira).

C.2. Les données, les variables et le principe de la statistique descriptive

Taille	Poids	Sexe	Sport pratiqué
183	80	H	Basket-ball
182	75	H	Escalade
173	66	F	Basket-ball
178	78	H	Gymnastique
192	77	H	Basket-ball
158	57	F	Natation
163	50	F	Judo
172	53	F	Tennis

TAB. C.1. Tailles, poids, sexes et sports pratiqués pour $N = 8$ individus.

Nous allons dans toute cette séance nous intéresser aux données que l'on pourra trouver dans le tableau C.1 ; elles ont été collectée à partir d'un échantillon de 8 personnes (réalisé pour l'année 2008 dans un groupe de M1APA).

Ces données sont des informations, de deux types : numériques (on parle aussi de données quantitatives) ou catégorielles (on parle aussi de données qualitatives). Elles ont un sens dans un contexte précis.

On pourra consulter l'article de Wikipédia intitulé Statistique descriptive (voir `http://fr.wikipedia.org/wiki/Statistique_descriptive`).

On cherche à décrire, c'est-à-dire résumer ou représenter, par des statistiques, les données disponibles quand elles sont nombreuses[1]. Il est important de résumer les observations sans détruire l'informations qu'elles contiennent.

Ces données varient (dans le temps, chez les individus) et prennent des valeurs différentes. Cette variabilité est si importante que l'on va donner aux mesures le nom de *variables*. Ainsi, on évoquera pour la population des $N = 8$ individus déjà évoqués, les variables taille, poids, sexe et sport pratiqué.

Les valeurs de la variable poids sont successivement 80, 75, 66, 78, 77, 57, 50, 53.

Les valeurs de la variable sexe sont successivement H, H, F, H, H, F, F, F.

[1] ce qui n'est guère pertinent dans notre cas ici !

Nous commencerons par le cas simple où il n'y a qu'une seule variable. On parle de phénomène mono-varié. À la fin du semestre, nous étudierons des phénomèmes multivariés (en fait, seul le cas de deux variables sera étudié).

On parle de variable quantitative (ou numérique) ou variable qualitative (ou catégorielle).

C.3. Étude de donnée qualitatives

On s'intéressera au sexe des $N = 8$ étudiants de M1APA (voir le tableau C.1).

C.3.1. Statistiques

On détermine tout d'abord le nombre de catégorie, puis pour chacune d'elles, le nombre d'effectifs (c'est-à-dire le nombre d'individu pour lesquels la variable associée est dans cette catégorie).

On divisant ces effectifs par le nombre total d'individu, on obtient les fréquences. En multipliant ces fréquences par 100, on obtient les pourcentages.

Exercice C.1. Déterminer les valeurs de ces statistiques pour l'échantillon des huit étudiants étudié.

Voir éléments de correction page 129.

C.3.2. Graphiques

On peut produire des graphiques du type graphe en barres : on trace autant de barres que de catégories, chacune d'elle étant de même largeur, et de hauteur proportionnelle à la fréquence.

On peut aussi tracer un camembert, où chaque catégorie est représentée par un secteur angulaire proportionnel à la fréquence ; l'ensemble des secteurs angulaire est le disque total.

Exercice C.2.

(1) Déterminer le graphe en barres et le camembert pour les sexes des huit étudiants.
(2) Ces graphes sont-ils pertinents ?

Voir éléments de correction page 129.

Exercice C.3. On s'intéresse maintenant au sport pratiqué par les huit étudiants déjà étudiés. Il faudra prende garde au fait qu'ici, il existe des cas de non réponse possibles (si aucun sport n'est pratiqués).

(1) Reprendre l'analyse précédente de cette variable.
(2) Représenter graphiquement ces données
(3) Quel ordre choisir pour les catégories ? Peut-on regrouper les catégories ou utiliser une catégorie "Autre" ?

C.4. Étude de données quantitatives

On s'intéressera au poids des $N = 8$ étudiants de M1APA (voir le tableau C.1 page précédente).

C.4.1. Statistiques

Ces données constitue un ensemble de nombres relatifs à une population de $N = 8$ individus. On les notera n_1, n_2, ..., n_8. De façon générale, ils seront notés $(n_i)_{1\leq i\leq N}$.

On a donc successivement

- $n_1 = 80$,
- $n_2 = 75$,
- $n_3 = 66$,
- $n_4 = 78$,

- $n_5 = 77$,
- $n_6 = 57$,
- $n_7 = 50$,
- $n_8 = 53$

C.4.1.1. La centralité.

On cherche tout d'abord à définir la centralité, c'est-à-dire, la valeur autour de laquelle s'organisent les différentes données.

La notion la plus connue est *la moyenne*[2].

Si l'on dispose de deux nombre, la moyenne est tout simplement le milieu, c'est-à-dire la demi-somme. De façon plus générale, la moyenne est le nombre, souvent noté m, qui se trouve à égale distance de tous les nombres $(n_i)_{1\leq i\leq N}$, soit encore le nombre m tel que

$$(m - n_1) + (m - n_2) + \ldots + (m - n_N) = 0$$

Cela revient à donner la définition suivante :

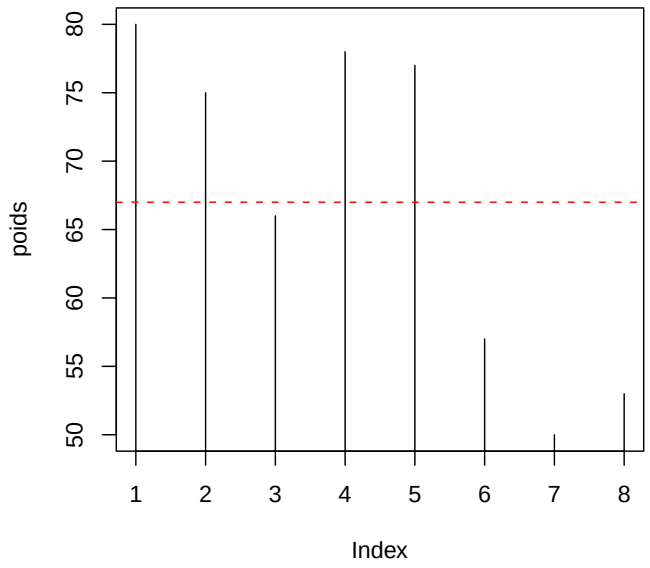
DÉFINITION C.4. La moyenne m des N nombres $(n_i)_{1\leq i\leq N}$ est définie par

$$m = \frac{1}{N}(n_1 + n_2 + \ldots + n_N) = \frac{1}{N}\sum_{i=1}^{N} n_i \tag{C.1}$$

EXEMPLE C.5. La moyenne m des $N = 8$ poids du tableau C.1 page 123 se calcule de la façon suivante : Le détail du calcul est le suivant :

$$\begin{aligned} m &= \frac{1}{N}\sum_{i=1}^{N} n_i, \\ &= \frac{1}{8}(80 + 75 + 66 + 78 + 77 + 57 + 50 + 53), \\ &= \frac{536}{8}, \\ &= 67 \end{aligned}$$

[2]l'adjectif "moyen" provient du latin *medianus*, qui signifie "du milieu" ; cet adjectif a été substantivé au féminin dans "moyenne" qui a perdu son sens initial, pour exprimer ce qui est également distant des deux extrèmes et correspond au type le plus répandu [**Rey98**]



Si on trace le *graphe indexé* ci-dessus avec en pointillé la moyenne, on constate que la somme des distances (algèbriques) de chacun des poids à la moyenne est nulle. Un autre image peut-être donnée : on place sur une régle graduée (infiniment légère) différentes masses égales à des abscisses correspondant aux différentes données. Cette régle, posée horizontale sur une pointe, sera en équilibre si cette pointe correspond à la moyenne.

Une notions moins connue est la *la médiane*[3]. La médiane est un nombre qui divise en deux parties la population. On la note Q_2.

DÉFINITION C.6. La médiane Q_2 des N nombres $(n_i)_{1\leq i\leq N}$ est une valeur choisie pour que la moitié des données lui soit inférieure et l'autre moitié supérieure. On la notera Q_2.

De façon plus précise, pour définir la médiane, on classe les données dans l'ordre croissant. S'il y a un nombre pair de valeurs, la moyenne des deux valeurs centrales est prise. S'il y a un nombre impair de valeurs, la valeur centrale est choisie. Contrairement à la moyenne, la valeur médiane permet d'atténuer l'influence perturbatrice des valeurs extrêmes enregistrées lors de circonstances exceptionnelles. On dit que la médiane est moins sensible aux extrèmes que la moyenne.

EXEMPLE C.7. La médiane Q_2 des $N = 8$ poids du tableau C.1 se calcule de la façon suivante. Le détail du calcul est le suivant :

- les données dans l'ordre croissant sont : 50, 53, 57, 66, 75, 77, 78, 80 ;
- le nombre de données est pair.
- on calcule donc la moyenne des deux valeurs centrales : 66 et de 75.
- la médiane vaut donc 70.5.

EXEMPLE C.8. Si on avait voulu calculer la médiane des $N - 1 = 7$ premiers poids du tableau C.1, on aurait procédé ainsi. Le détail du calcul est le suivant :

- les données dans l'ordre croissant sont : 50, 57, 66, 75, 77, 78, 80 ;
- le nombre de données est impair.
- on prend la valeur centrale 75.
- la médiane vaut donc 75.

[3]qui provient du latin *medianus*, qui signifie "du milieu" [**Rey98**]

C.4.1.2. La dispersion ou l'hétérogénéité.

On cherche maintenant à définir si les données sont rassemblées ou non autour de la moyenne ou de la médiane.

Les extréma (minimum et maximum), notés $\min\left(n_{i1\leq i\leq N}\right)$ et $\max\left(n_{i1\leq i\leq N}\right)$.

Les deux notions les plus importantes pour mesurer la dispersion des données autour de la moyenne sont *la variance* et *l'écart-type*.

On s'intéresse à la somme des écarts entre les données et la moyenne. Par définition la somme

$$(m-n_1)+(m-n_2)+\ldots+(m-n_N)=0$$

est nulle. On considérera une autre somme, où chaque quantité est toujours positive, en prenant par exemple le carré de chacun de ces termes :

$$(m-n_1)^2+(m-n_2)^2+\ldots+(m-n_N)^2$$

On divise cela par N pour donner autant d'importance à chaque terme. On obtient donc la variance

$$\frac{1}{N}\left((m-n_1)^2+(m-n_2)^2+\ldots+(m-n_N)^2\right)$$

Pour obtenir une quantité homogène à chacune des données, on prend la racine carrée de la variance. On obtient donc l'écart-type :

$$\sqrt{\frac{1}{N}\left((m-n_1)^2+(m-n_2)^2+\ldots+(m-n_N)^2\right)}$$

Définition C.9. La *variance*, notée σ^2, des N nombres $(n_i)_{1\leq i\leq N}$ est définie par

$$\sigma^2=\frac{1}{N}\left((m-n_1)^2+(m-n_2)^2+\ldots+(m-n_N)^2\right)=\frac{1}{N}\sum_{i=1}^{N}(m-n_i)^2 \tag{C.2}$$

Définition C.10. L'*écart-type*, noté σ, des N nombres $(n_i)_{1\leq i\leq N}$ est défini par

$$\sigma=\sqrt{\frac{1}{N}\left((m-n_1)^2+(m-n_2)^2+\ldots+(m-n_N)^2\right)}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(m-n_i)^2} \tag{C.3}$$

Remarque C.11. Attention, on parle quelque fois de l'écart-type estimé :

$$\sqrt{\frac{1}{N-1}\left((m-n_1)^2+(m-n_2)^2+\ldots+(m-n_N)^2\right)}$$

On note aussi l'écart-type par son nom anglais, *sd*, comme standart deviation. Attention, ®détermine la déviation standart et non l'écart-type !

Exemple C.12. La variance σ^2 des $N=8$ poids du tableau C.1 se calcule de la façon suivante. Le détail du calcul est le suivant :

$$\begin{aligned}
\sigma^2&=\frac{1}{N}\sum_{i=1}^{N}(m-n_i)^2,\\
&=\frac{1}{8}\left((-13)^2+(-8)^2+1^2+(-11)^2+(-10)^2+10^2+17^2+14^2\right),\\
&=\frac{1}{8}\left(169+64+1+121+100+100+289+196\right),\\
&=\frac{1040}{8},\\
&=130
\end{aligned}$$

EXEMPLE C.13. L'écart-type σ des $N = 8$ poids du tableau C.1 se calcule de la façon suivante. Le détail du calcul est le suivant :

$$\begin{aligned}
\sigma &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(m-n_i)^2}, \\
&= \sqrt{\frac{1}{8}\left((-13)^2+(-8)^2+1^2+(-11)^2+(-10)^2+10^2+17^2+14^2\right)}, \\
&= \sqrt{\frac{1}{8}\left(169+64+1+121+100+100+289+196\right)}, \\
&= \sqrt{\frac{1040}{8}}, \\
&= \sqrt{130}, \\
&= 16.25
\end{aligned}$$

La notion de *quartile* permet de mesurer la dissymétrie et d'appréhender de façon différente la question de la dispersion.

DÉFINITION C.14. De la même façon que la médiane partageait le jeu de données en deux groupes de même effectif, les quartiles vont le partager en quatre groupes d'effectifs égaux. Ainsi 25% des données seront inférieures au premier quartile (Q_1), 50% au deuxième quartile qui n'est autre que la médiane (Q_2) et 75% au troisième quartile (Q_3).

Autrement dit,

- 25% des données sont inférieures à Q_1,
- 25% des données sont comprises entre Q_1 et Q_2,
- 25% des données sont comprises entre Q_2 et Q_3,
- 25% des données sont supérieures à Q_3.

Parfois le minimum est noté Q_0 et le maximum noté Q_4.

Les autres quartiles Q_1 et Q_3 sont donc définis comme la médiane de l'ensemble des valeurs inférieures à la médiane et la médiane de l'ensemble des valeurs supérieures à la médiane.

EXEMPLE C.15. les trois quartiles des $N = 8$ poids du tableau C.1 valent respectivement 56, 70.5 et 77.25.

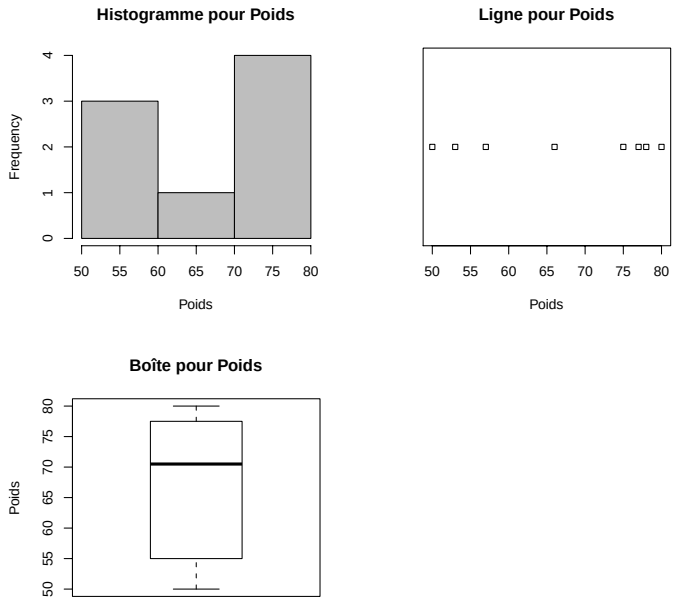
C.4.1.3. Une remarique sur la moyenne et l'écart type.

Souvent, les étudiants retiennent des statistiques descriptives (quand ils retiennent quelque chose) les notions de moyenne et d'écart-type. C'est très souvent, en effet, les statistiques toujours présentées. Si ces deux nombres ont autant d'importance, c'est parce que, dans un grand nombre de cas, les données étudiées suivent une loi idéale, dite en cloche ou "normale". Cette loi à la forme d'une cloche et est décrite par deux paramètres, notés moyenne et écart-type. Ces deux nombres sont proches de la moyenne et de l'écart-type des données considérées. Autrement dit, si les données suivent bien la loi normale, les deux nombres en question reflétent totalement les données considérés et les caractérisent donc.

C.4.2. Les graphiques

Pour représenter graphiquement des données quantitatives, on peut représenter les valeurs individuelles le long d'une échelle (ligne de points, avec empilement des points égaux). On peut aussi les regrouper par tranche et tracer un histogramme : pour chaque tranche choisie, on détermine le nombre d'individus pour lesquels la variable qualitative appartient à cette tranche. On trace ensuite des colonnes dont la base correspond à la tranche et la hauteur est proportionnelle au nombre d'individu.

On peut aussi tracer des boîtes de dispersion mettant en évidence les extrêmes et les quartiles.



Voir les trois graphiques ci-dessus pour la variable poids. Le nombre de données étant faible (8), l'histogramme et la boîtes à moustache ne sont pas très pertinents ici.

Ces trois graphiques ne sont pas toujours pertinents.

Exercice C.16. Déterminer les statistiques et faire les graphiques à main levée de la variable taille. Pour l'histogramme, on prendra des classes de largeur 10 à partir de 150.

Voir éléments de correction page 130.

C.5. Élements de correction

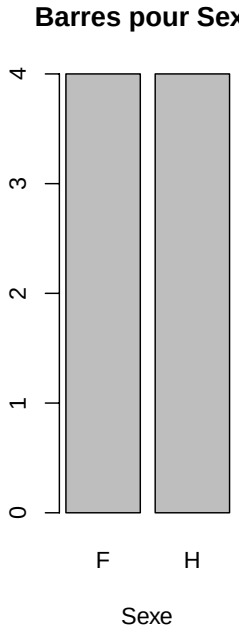
Éléments de correction de l'exercice C.1

Les effectifs et les pourcentages déterminés par R sont donnés dans le tableau suivant

	effectifs	pourcentages
F	4	50.000
H	4	50.000

Éléments de correction de l'exercice C.2

(1)

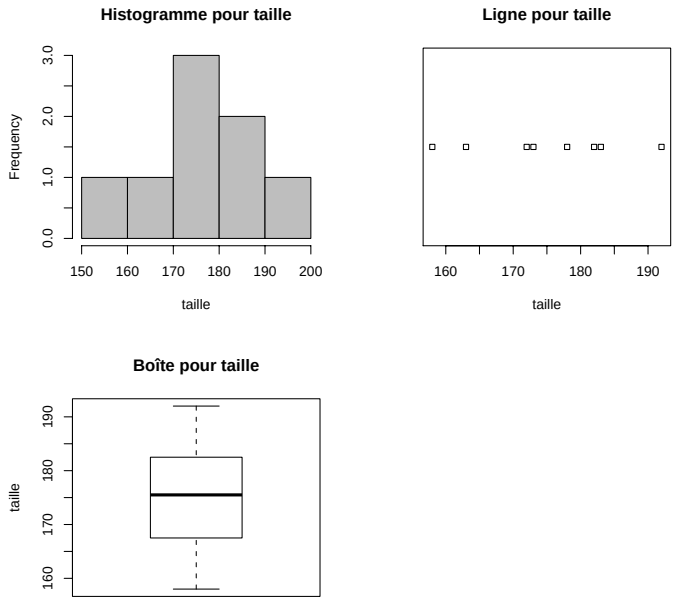


Voir les trois graphiques ci-dessus pour la variable sexe.

(2) À vous de voir ....

Éléments de correction de l'exercice C.16
Les différents résultats déterminés par R sont donnés dans le tableau suivant

noms	valeurs
moyenne	175.125
sd	11.063938
Q_1 (quartile à 25 %)	169.75
médiane	175.5
Q_3 (quartile à 75 %)	182.25
minimum	158
maximum	192
nombre	8



Voir les trois graphiques ci-dessus.

ANNEXE D

Utilisation de fonctions avec R

Ici, la notion de fonction est introduite pour vous simpflifier vos démarches avec R, mais l'usage de ces fonctions n'est nullement imposé (sauf éventuellement pour faire tourner des démonstrations) !

D.1. Une fonction "simple"

Que vous soyez un adepte de Rcmdr ou non, on peut maintenant évoquer la notion de fonction. Quand vous tapez par exemple

```
cos(5)
```

cela calcule le cosinus de 5. On peut aussi sois-même écrire des fonctions. Voir la fonction `somme.R`, disponible à l'URL habituelle (dans la rubrique "fonctionsR"). Vous pouvez visualiser ce fichier et constatez qu'il comporte :

- un entête :
  ```
  somme <- function(x, y) {
  }
  ```
- éventuellement des commentaires (lignes commençant par `#` et qui expliquent le fonctionnement de la fonction) :
  ```
  # exemple de fonction : somme de deux nombres
  # ****************************************
  # somme(x,y) :
  #  * Variables d'entrées :
  #    *   x,y : les deux nombres dont on veut la somme
  #  * Variable de sortie :
  #    *  la somme de x et de y
  ```
- et un corps de fonction
  ```
  s <- x + y
  return(s)
  ```
  qui retourne la somme des deux nombre. Ce corps de fonction comporte en fait les diférentes étapes (ici très simple) nécessaires au calcul exigé.

Pour utiliser cette fonction, il faut

(1) d'abord "sourcer" cette fonction. Vous avez quatre possibilités :

(a) Soit récupérer le fichier `somme.R` dans le répertoire de travail et faire "fichier", puis "Sourcer du code R" et choisir `somme.R`.

(b) Soit récupérer le fichier `somme.R` dans le répertoire de travail et taper la ligne de commande (dans Rgui) :

```
source("somme.R")
```

(c) Soit s'affranchir de la récupération du fichier `somme.R` en tapant directement

```
source("http://utbmjb.chez-alice.fr/UFRSTAPS/M1PPMR/fonctionsR/somme.R")
```

ce qui ne marche que si la connexion internet est correcte !

(d) Si vous avez accès au texte de la fonction, ici

```
somme<-function(x,y){

# exemple de fonction : somme de deux nombres
# ***************************************
# somme(x,y) :
#  * Variables d'entrées :
#    *   x,y : les deux nombres dont on veut la somme
#  * Variable de sortie :
#    *  la somme de x et de y

s<-x+y
return(s)
}
```

il faut en faire un copier-coller et à partir d'un éditeur simple (type bloc-note) l'enregister dans un fichier de nom `somme.R` dans votre répertoire de travail. Vous pouvez aussi utiliser l'éditeur *ad hoc* de R, en allant dans "fichier", "Nouveau script", puis une fois le texte collé, faite "sauver".

REMARQUE D.1. Cette éditeur vous permet aussi de voir des fichiers R déjà écrits en allant dans "fichier", puis "ouvrir un script".

Comme précédemment, il faudra alors le "sourcer".

(2) Cette fonction est donc chargée et, ensuite, vous pourrez la faire tourner en tapant par exemple (ici, les deux arguments sont '`x`' et '`y`')

```
 somme(2, 3)
```

```
[1] 5
```

à comparer à

```
 2 + 3
```

```
[1] 5
```

D.2. Une fonction à deux valeurs de sortie

Considérons maintenant la fonction `somme_rapport.R`, disponible à l'URL habituelle. Comme indiqué dans la setion D.1, récupérez et sourcez-la. Cette fonction renvoie deux expressions, la somme et le rapport de deux nombres. Tapez par exemple

```
 somme_rapport(2, 3)
```

```
$s
[1] 5

$r
[1] 0.6666667
```

Cette fonction renvoie en fait une liste avec deux éléments (ce qui permet d'avoir plusieurs valeurs de sortie). On pourra pour comprendre comment fonctionne une liste en tapant par exemple

```
 res <- somme_rapport(2, 3)
 class(res)
```

```
[1] "list"
```

```
 names(res)
[1] "s" "r"
 res$s
[1] 5
 res$r
[1] 0.6666667
```

On peut aussi définir les valeurs des arguments "dans le désordre" à condition de spécifier quel argument est x et quel argument est y. Comparez ce que donne

```
 somme_rapport(2, 3)
$s
[1] 5

$r
[1] 0.6666667
 somme_rapport(3, 2)
$s
[1] 5

$r
[1] 1.5
 somme_rapport(x = 2, y = 3)
$s
[1] 5

$r
[1] 0.6666667
 somme_rapport(y = 3, x = 2)
$s
[1] 5

$r
[1] 0.6666667
```

Une fonction peut aussi avoir un argument optionnel. Quand il n'est pas indiqué, il prend la valeur imposée par défaut par la fonction. Par exemple, si y n'est pas indiqué, il vaut 1. Comparez ce que donne

```
 somme_rapport(2, 1)
$s
[1] 3

$r
[1] 2
 somme_rapport(y = 1, x = 2)
```

```
$s
[1] 3

$r
[1] 2
```

```
somme_rapport(2)
```

```
$s
[1] 3

$r
[1] 2
```

D.3. D'autres fonctions

Nous utiliserons dans ce cours un certain nombre de fonctions, déjà écrites et disponibles à l'URL habituelle. Elles permettent de faire des calculs déjà programmés et fréquemment utilisés.

Bien entendu, vous pourrez vous même écrire vos propres fonctions. Consultez la section 6.3 page 72 de l'excellente introduction à R, [**Par05**] disponible sur internet.

ANNEXE E

Lien entre la moyenne et l'écart-type d'une variable aléatoire et la moyenne et l'écart-type des valeurs prises par cette variable aléatoire au cours expérience (preuve de la proposition 5.7)

Montrons l'égalite (5.7a). On rappelle que l'on effectue N tirages aléatoires d'une variable aléatoire X et on note $(y_i)_{1\leq i\leq N}$ les valeurs obtenues. Notons $(n_j)_{1\leq j\leq q}$ l'ensemble des valeurs prises par la variable aléatoire X. Notons, pour chaque $j \in \{1,\ldots,q\}$, α_j, le nombre de fois où est apparue la valeur n_j parmi les valeurs $(y_i)_{1\leq i\leq N}$. Ainsi, la moyenne des valeurs $(y_i)_{1\leq i\leq N}$ (au sens de 3.1 page 9) est égale à

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} y_i &= \frac{1}{N}\left(\alpha_1 n_1 + \alpha_2 n_2 + \ldots + \alpha_q n_q\right),\\ &= \frac{\alpha_1}{N} n_1 + \ldots + \frac{\alpha_q}{N} n_q\end{aligned}$$

En vertu, de la définition 5.10 page 27, si N est "grand", alors,

$$\begin{aligned}\frac{\alpha_1}{N} &\approx p(X = n_1),\\ \frac{\alpha_2}{N} &\approx p(X = n_2),\\ &\vdots\\ \frac{\alpha_q}{N} &\approx p(X = n_q),\end{aligned}$$

et donc

$$\frac{1}{N}\sum_{i=1}^{N} y_i \approx p(X = n_1)n_1 + \ldots + p(X = n_q)n_q$$

ce qui est exactement la valeur de l'espérance $\mathbb{E}(X)$ donnée par (5.2).

On ferait de même pour (5.7b).

ANNEXE F

Preuve de la proposition 5.39

Calculons tout d'abord l'espérance de la variable aléatoire binomiale de paramètres n et π égale par définition et d'après la proposition 5.32 et la formule du binome

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x=0}^{n} xP(X=x)\\ &= \sum_{x=0}^{n} xC_n^x \pi^x (1-\pi)^{n-x}\end{aligned}$$

Pour calculer cela, on introduit la fonction

$$f(X) = (X+1-\pi)^n = \sum_{k=0}^{n} C_n^k X^k (1-\pi)^{n-k}. \tag{F.1}$$

Dérivons ces égalités termes à termes (par rapport à X)

$$f'(X) = ((X+1-\pi)^n)' = \left(\sum_{k=0}^{n} C_n^k X^k (1-\pi)^{n-k}\right)',$$

et donc

$$n(X+1-\pi)^{n-1} = \left(\sum_{k=0}^{n} C_n^k X^k (1-\pi)^{n-k}\right)'$$

soit encore

$$\begin{aligned}n(X+1-\pi)^{n-1} &= \left(C_n^0 (1-\pi)^n\right)' + \left(\sum_{k=1}^{n} C_n^k X^k (1-\pi)^{n-k}\right)'\\ &= \sum_{k=1}^{n} C_n^k k X^{k-1} (1-\pi)^{n-k}.\end{aligned}$$

et donc

$$\begin{aligned}nX(X+1-\pi)^{n-1} &= X\left(\sum_{k=0}^{n} C_n^k k X^{k-1} (1-\pi)^{n-k}\right)\\ &= \sum_{k=1}^{n} C_n^k k X^k (1-\pi)^{n-k}\\ &= \sum_{k=0}^{n} C_n^k k X^k (1-\pi)^{n-k}\end{aligned}$$

En remplaçant X par π, on a donc finalement

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=0}^{n} kC_n^k \pi^k (1-\pi)^{n-k}\\ &= n\pi(\pi+1-\pi)^{n-1}\\ &= n\pi\end{aligned}$$

ce qui est bien la formule annoncée.

Pour calculer l'écart-type et la variance, on ferait de même. ♢

ANNEXE G

Un sourcier et la loi binomiale (sous forme d'exercice corrigé)

Cet exercice a déjà été donné en examen de M1 PPMR (CCF2 Automne 2009).

Énoncé

Exercice G.1.

Un sourcier affirme être capable de ressentir la présence d'eau à l'aide d'une "baguette". Une expérience a été mise en place pour tester ses aptitudes. Un nombre $n = 20$ de containers identiques ont été utilisés dont certains sont remplis avec de l'eau. On a demandé au sourcier lesquels étaient pleins.

(1) Le nombre de réponses justes données par le sourcier est de 12. Ce niveau de succès prouve-t-il les compétences du sourcier ?

(2) On suppose maintenant que le sourcier n'a pas de compétences et que sa probabilité de succès est donc de $\pi = 0.5$. Un nombre $n = 20$ d'essais sont tentés et on fait l'hypothèse que le nombre de réponses justes suit une loi binomiale.

   (a) Calculer la probabilité pour le sourcier d'avoir exactement 10 réponses justes.

   (b) Calculer également sa probabilité d'avoir plus de 13 réponses justes.

(3) (a) Supposons maintenant que la sourcier ait effectivement des capacités et que sa probabilité de succès soit en vérité de π=0.75. Quelle est alors sa probabilité d'avoir plus de 13 réponses justes ? Prendre comme critère de décision que le sourcier ait des compétences si on enregistre plus de 13 succès vous semble-t-il raisonnable (expliquer pourquoi oui ou non) ?

   Indication On pourra utiliser l'une des commandes suivante :

```
prop.test(x=13,n=20,p=0.75,alternative="greater",correct=F,conf.level=0.95)
```

   ou

```
prop.test(x=13,n=20,p=0.75,alternative="less",correct=F,conf.level=0.95)
```

   ou

```
prop.test(x=13,n=20,p=0.75,alternative="two.sided",correct=F,conf.level=0.95)
```

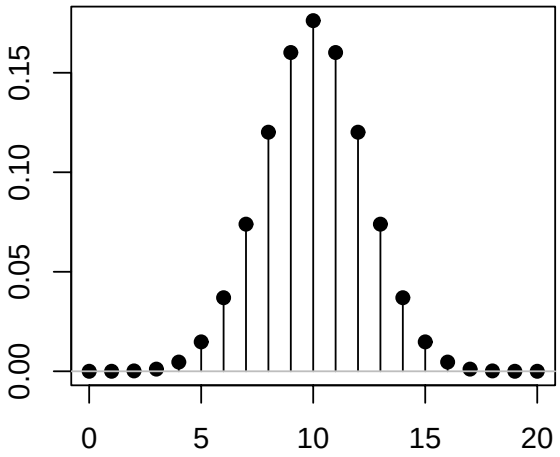
   qui permettent respectivement de faire un test d'hypothèse sur la proportion observée $\pi = 13/20$ avec $n = 20$ essais avec l'hypothèse nulle $\pi = 0.75$ contre l'hypothèse alternative :
   - $\pi > 0.75$
   - $\pi < 0.75$
   - $\pi = 0.75$

   (b) Toujours dans ce cas où π=0.75, à combien de succès peut-on s'attendre pour le sourcier ?

Corrigé

Éléments de correction de l'exercice G.1

(1)
- Le nombre de réponses justes données par le sourcier est de 12 ce qui est proche de la moitié du nombre de containers (10).
  Ce résultat est à comparer avec celui qu'il aurait obtenu "par hasard". Si le sourcier n'a pas de compétences, on peut faire l'hypothèse que la probabilité de succès sur chaque container est de $\pi = 0.5$. Un nombre $n = 20$ d'essais sont tentés et on fait donc l'hypothèse que le nombre de réponses justes suit une loi binomiale.



  On peut tracer la loi de probabilité comme sur la figure ci-dessus. On renvoie aux manipulations avec Rcmdr 5.35 page 32 du cours.
  La question que l'on se pose est "est-ce que le sorcier fait mieux que le hasard". Déterminons la probabilité que le score du sorcier soit inférieur celui du hasard, c'est-à-dire $P(X \geq 12)$ où X suit une loi binomiale de paramètres $\pi = 0.5$ et $n = 20$. On renvoie aux manipulations avec Rcmdr 5.48 page 34 du cours. On écrit $P(X \geq 12) = P(X > 11)$, pour passer par l'aire à droite avec Rcmdr, ce qui fournit 0.25172. On obtient donc

$$p = 0.25172. \tag{G.1}$$

  Cette probabilité est supérieure au seuil usuel de $\alpha = 0.05$. Si elle avait été inférieure, on aurait pu la considérer comme suffisamment proche de zéro pour considérer le score du sourcier comme très peu probable, et donc non dû au hasard ! Donc, ici, on rejette au contraire le fait que le score du sourcier est non dû au hasard et on conclue donc que *le sourcier n'a pas de compétences !*.
- On peut aussi, en utilisant les manipulations avec Rcmdr 5.48 page 34 du cours, calculer les probabilités cumulées $P(X \geq k)$ pour chaque valeur de $k \in \{0, \ldots, 20\}$:
  On obtient le tableau G.1 page suivante.
  Dans ce tableau, on voit deux sous-ensembles :
  - l'ensemble $\{0, ..., 14\}$, où la probabilité cumulée $P(X \geq k)$ est strictement supérieure à 0.05.
  - l'ensemble $\{15, \ldots, 20\}$, où la probabilité cumulée $P(X \geq k)$ est inférieure ou égale à 0.05.

  Comme on vient de faire, dans la première région, appelée R_c, on accepte le fait que le résultat du sourcier est dû au hasard et dans la seconde, on accepte qu'il ait des compétences.

k	probabilités cumulées
0	1.00000000
1	0.99999905
2	0.99997997
3	0.99979877
4	0.99871159
5	0.99409103
6	0.97930527
7	0.94234085
8	0.86841202
9	0.74827766
10	0.58809853
11	0.41190147
12	0.25172234
13	0.13158798
14	0.05765915
15	0.02069473
16	0.00590897
17	0.00128841
18	0.00020123
19	0.00002003
20	0.00000095

TAB. G.1. Les différentes probabilités cumulées

Bref, *au seuil* 0.05*, on accepte les compétences du sourcier pour un score supérieur ou égal à à 15.*
Graphiquement, on peut tracer le graphe des probabilités cumulées (avec l'aire à droite) comme dans la figure G.1 page suivante à gauche (voir manipulations avec Rcdmr 5.44 page 33 du cours). Sur ce graphe, on a rajouté en rouge la droite d'ordonnée 0.05 et en bleu pointillé la droite d'abscisse 15.
Voir aussi la figure G.2 page suivante où on a représenté le graphe des probabilités simples avec la région du "vrai don du sourcier" en rouge. Sur ce graphe, on a rajouté en rouge la droite la droite d'abscisse 15.

REMARQUE G.2. On pourra consulter la partie statistique de l'observatoire de zététique `http://www.zetetique.fr/stats/` plus particulièrement faire tourner le cas correspondant à celui que l'on vient de traiter grâce à `http://www.zetetique.fr/stats/stats.php?cas=11`.

- Une autre façon de procéder est la suivante : On fait un test Z d'hypothèse en proportion, comme indiqué dans la section 5.9 page 48.
  On procède au *test Z d'une proportion.*
  On fait l'hypothèse nulle $H_0 : \quad \pi = \pi_0$. avec $\pi_0 = 0.5$. On cherche à montrer que le paramètre π de la loi binomiale, dont proviendraient les données de l'échantillon étudié, est plus grande que π_0. On fait donc l'hypothèse alternative suivante : $H_1 : \quad \pi > \pi_0$.
  Puisque $n = 20$, est "grand", on remplacera la loi binomiale de paramètre n et π_0 par la loi normale de moyenne $\mu = \pi_0$ et d'écart-type $\sqrt{\pi_0(1-\pi_0)/n}$. Grâce à ℝ, on trouve la valeur suivante de la statistique

$$z = \frac{pr - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = 0.894427$$

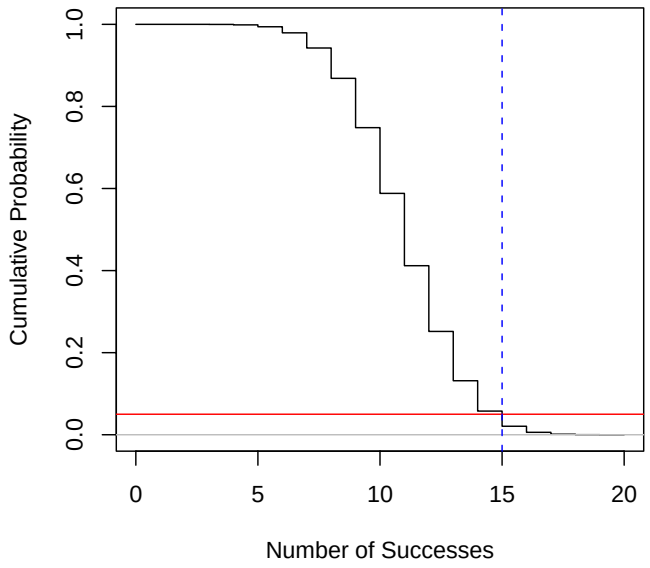


FIG. G.1. Le graphes des probabilités cumulées $P(X \geq k)$ pour $n = 20$ et $\pi = 0.5$.

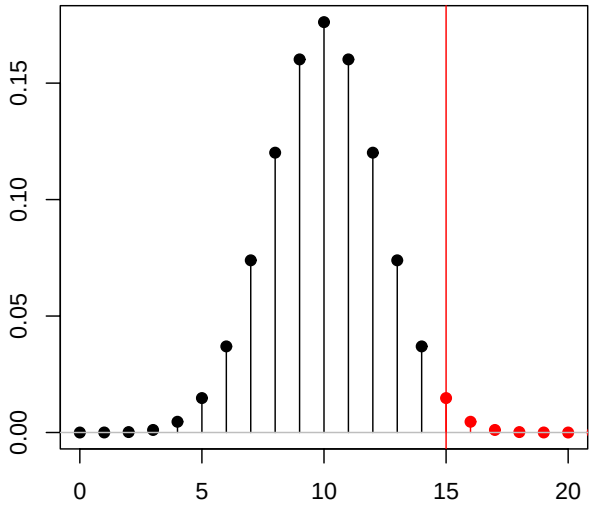


FIG. G.2. Le graphes des probabilités simples avec la région du "vrai don du sourcier" en rouge.

La probabilité critique $P(Z \geq z)$ (pour la loi normale centrée réduite) est égale à

$$p_c = 0.185547.$$

Puisque p_c est strictement supérieure au niveau de signification $\alpha = 0.05$, on accepte l'hypothèse nulle H_0. Ainsi, H_0 est vraie et donc *la proportion π est égale à $\pi_0 = 0.5$,* au risque 0.05.
Le sourcier n'a donc pas de réelles compétences !
On pourra utiliser directement la commande donnée en examen :

```
prop.test(x=12,n=20,p=0.5,alternative="greater",correct=F,conf.level=0.95)
```

adaptée à l'hypothèse alternative $H_1 : \pi > \pi_0$, qui redonne

```
[1] 0.1855467
```

REMARQUE G.3. Si on tape

```
prop.test(x=12,n=20,p=0.5,alternative="greater",correct=T,conf.level=0.95)
```

ou directement

```
prop.test(x=12,n=20,p=0.75,alternative="greater",conf.level=0.95)
```

on obtient une valeur légèrement différente de la probabilité critique

$$p_c = 0.251167.$$

Cette valeur est presque identique à la valeur donnée par (G.1). Pour ce calcul, puisque l'on a choisit une valeur de '`correct`' égale à `T`, le calcul est exact, l'approximation par la loi normale (valable pour les grandes valeurs de n) non utilisée.

(2) (a) Pour les manipulations avec R, on renvoie à la manipulation avec Rcmdr 5.34 dans document de cours. Les paramètres de la loi binomiale sont $n = 20$ et $\pi = 0.5$. La probabilité pour le sourcier d'avoir exactement 10 réponses justes est égale à 0.1762.

(b) Pour calculer la probabilité d'avoir plus de 13 réponses justes, on écrit que cette probabilité vaut $P(X \geq 13) = P(X > 12)$, pour passer par l'aire à droite avec Rcmdr, ce qui fournit 0.13159.

(3) (a)
- Les paramètres de la loi binomiale sont $n = 20$ et $\pi = 0.75$. La probabilité d'avoir plus de 13 réponses justes est maintenant égale à 0.89819 (supérieure à 0.13159).
- Si la probabilité de succés est effectivement de $\pi = 0.75$, la probabilité d'avoir plus de 13 réponses justes est égale à $P(X \geq 13) = P(X > 12)$, soit 0.89819. Ainsi, si $\pi = 0.75$, dans 89.8 % des cas, le sourcier obtiendra plus de 13 réponses justes ! Cependant, le nombre 13 ne prouve pas *a posteriori* que $\pi = 0.75$.
- On fait de nouveau un test d'hypothèse Z en proportion.
  Deux façon de procéder :

  (i) On procède au *test Z d'une proportion.*

  On fait l'hypothèse nulle $H_0 : \quad \pi = \pi_0$. avec $\pi_0 = 0.75$. On cherche à montrer que le paramètre π de la loi binomiale, dont proviendraient les données de l'échantillon étudié, est plus petite que π_0. On fait donc l'hypothèse alternative suivante : $H_1 : \quad \pi < \pi_0$.
  Puisque $n = 20$, est "grand", on remplacera la loi binomiale de paramètre n et π_0 par la loi normale de moyenne $\mu = \pi_0$ et d'écart-type $\sqrt{\pi_0(1-\pi_0)/n}$. Grâce à R, on trouve la valeur suivante de la statistique

$$z = \frac{pr - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = -1.032796$$

  La probabilité critique $P(Z \leq z)$ (pour la loi normale centrée réduite) est égale à

$$p_c = 0.15085.$$

  Puisque p_c est strictement supérieure au niveau de signification $\alpha = 0.05$, on accepte l'hypothèse nulle H_0. Ainsi, H_0 est vraie et donc *la proportion π est égale à $\pi_0 = 0.75$,* au risque 0.05.

  Le sourcier a donc de réelles compétences ! !

(ii) "avec R" On pourra utiliser directement la commande donnée en examen :

```
prop.test(x=13,n=20,p=0.75,alternative="less",correct=F,conf.level=0.95)
```

adaptée à l'hypothèse alternative $H_1 : \pi < \pi_0$, qui redonne

```
[1] 0.1508498
```

REMARQUE G.4. Si on tape

```
prop.test(x=13,n=20,p=0.75,alternative="greater",correct=T,conf.level=0.95)
```

ou directement

```
prop.test(x=13,n=20,p=0.75,alternative="greater",conf.level=0.95)
```

on obtient une valeur légèrement différente de la probabilité critique

$$p_c = 0.219289.$$

Pour ce calcul, puisque l'on a choisit une valeur de '`correct`' égale à `T`, le calcul est exact, l'approximation par la loi normale (valable pour les grandes valeurs de n) non utilisée.

REMARQUE G.5. On vient de montrer avec la théorie des tests d'hypthèse en proportion que la valeur de $n = 20$ permettait d'accepter que $\pi = 0.75$. Mais la question était plus précise : "Prendre comme critère de décision que le sourcier ait des compétences si on enregistre plus de 13 succès vous semble-t-il raisonnable". Autrement dit, on cherche la valeur minimale de n qui permette d'accepter que $\pi = 0.75$!
Ici, on raisonne en fait en terme de région critique (voir section 6.5.2.2).

k	probabilités critiques
0	0.00000000
1	0.00000000
2	0.00000000
3	0.00000000
4	0.00000001
5	0.00000012
6	0.00000168
7	0.00001805
8	0.00015030
9	0.00097289
10	0.00491164
11	0.01943355
12	0.06066763
13	0.15084979
14	0.30278831
15	0.50000000
16	0.69721169
17	0.84915021
18	0.93933237
19	0.98056645
20	0.99508836

TAB. G.2. Les différentes probabilités critiques en fonction du nombre de réponses justes

Voir le tableau G.2. Dans ce tableau, on voit deux sous-ensembles :

- l'ensemble $\{0, ..., 11\}$, où la probabilité critique est inférieure ou égale à 0.05.
- l'ensemble $\{12, \ldots, 20\}$, où la probabilité critique est strictement supérieure à 0.05.

Dans la première région, appelée R_c, on rejette H_0 et donc $\pi < \pi_0$. Dans la seconde, on acceptera H_0 et donc $\pi = \pi_0$.
On constate que 13 appartient à la seconde région et de plus, que jusqu'à 11, on peut considérer le sourcier comme capable ! Autrement dit, au seuil 0.05, la bonne décision est "Prendre comme critère de décision que le sourcier ait des compétences si on enregistre plus de 11 succès " et non "si on enregistre plus de 13 succès" !

REMARQUE G.6. *Attention*, on vient de montrer dans cette question, que, au delà de 11 succès, le sourcier a de réelles compétences, au sens où, au seuil 0.05, la proportion annoncée $\pi = 0.75$ est conforme au nombre de succès. En revanche dans la question 1, on a montré que le sourcier obtient des résultat non dûs au hasard pour un score supérieur à 15, ce qui n'est pas identique !

(b) Si π=0.75, on peut s'attendre, en moyenne à $n\pi = 0.75 \times 20 = 15$ succès.

ANNEXE H

Passage d'une loi de probabilité discrète à une loi de probabilité continue

H.1. Une manipulation sur la loi binomiale

Exercice H.1. Observer comment évolue le graphe de la distribution binomiale évolue lorsque vous conservez $\pi = 0.3$ et que vous choisissez $n \in \{5, 20, 100, 150, 200, 400\}$.

Voir éléments de correction page 152

Reprenons maintenant le très bon exemple issu de la page 7 de [**DC08**]. Reprenons loi binomiale de paramètres n et π et modifions-la pour que son espérance ($\mu = n\pi$) soit égale à 0 et son écart-type ($\sigma = \sqrt{n\pi(1-\pi)}$) égal à 1. On peut montrer qu'il suffit de garder la même loi (définie par (5.8) page 31) et de remplacer chacune des valeurs de succès $k \in \{0, \ldots, n\}$ par $(k-\mu)/\sigma$ pour $k \in \{0, \ldots, n\}$. Autrement dit, on a

$$\forall x \in \{0, \ldots, n\}, \quad P\left(X = \frac{x-\mu}{\sigma}\right) = C_n^x \pi^x (1-\pi)^{n-x} \tag{H.1}$$

On dit que l'on obtient la loi binomiale normalisée (d'espérance nulle et d'écart-type égal à 1).

Sur le graphique H.1 page suivante, on a indiqué le tracé de la loi binomiale normalisée sur l'intervalle $[-3, 3]$ $\pi = 0.3$ et $n \in \{5, 20, 150, 200, 1000, 10000, 1e+06\}$, ainsi que (pour la dernière courbe) la "loi normale de moyenne $\mu = 0$ et d'écart-type $\sigma = 1$", qui est la "courbe en cloche idéale" et qui sera définie dans le chapitre 6.

Sur la figure suivante H.2 page 151, on a tracé les mêmes éléments en "normant" les probablilités de telle sorte que la valeur maximale atteinte soit égale à la valeur maximale de la "courbe en cloche" idéale (en 0 : $1/\sqrt{2\pi}$, ici le "vrai" π).

On constate alors que "la courbe des probabilités discrètes" ont l'air de "se rapprocher" d'une "courbe continue" quand n augmente. Cette courbe est la "courbe en cloche idéale".

Le nombre de valeurs possibles de la loi de probabilité discrète tend vers l'infini, chacune des probabilités tend vers 0, mais ce qui devient constant c'est la probabilité d'être dans un intervalle :

$$P(\alpha \leq X \leq \beta) = \sum_{\alpha \leq x \leq \beta} P(X = x)$$

Il y a dans un intervalle, de plus en plus d'événements mais la somme de leur probabilité tend vers une quantitée fixée

$$P(\alpha \leq X \leq \beta) = \text{Aire "sous la cloche idéale" entre les abscisses } \alpha \text{ et } \beta \tag{H.2}$$

La forme de la distribution se stabilise. Pour rendre compte de ce phénomène il faut utiliser la fonction de répartition, c'est-à-dire les *probabilités cumulées de la loi binomiale* (voir section 5.5.4 page 33).

En procédant comme dans l'exercice 5.50 page 34, on peut tracer le graphique des probabilités cumulées de la loi binomiale ou ceux de la loi binomiale normalisée. Voir graphique H.3 page 152. La dernière courbe est celle de la loi normale : elle représente "l'aire sous la cloche idéale"entre $-\infty$ et x.

On appelle ce phénomène la convergence en loi de la loi binomiale normalisée vers la loi normale ("loi de la cloche idéale") quand n tend vers l'infini (avec π constant). La fonction qui est représentée sur le dernier graphe de la figure H.3 est la fonction de répartition de la loi normale".

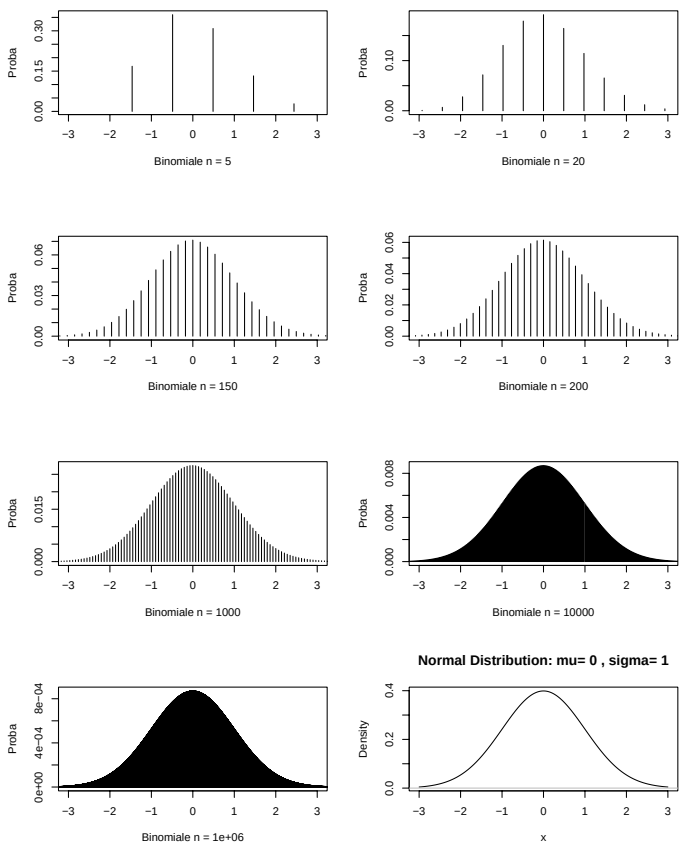


FIG. H.1. Les 7 graphes correspondant aux graphes de la loi binomiale normalisée avec $\pi = 0.3$ et $n \in \{5, 20, 150, 200, 1000, 10000, 1e+06\}$, ainsi que (pour la dernière courbe) la loi normale de moyenne $\mu = 0$ et d'écart-type $\sigma = 1$.

H.2. Passage du discret au continu

Idées lors du passage du discret au continu
- Nombre de valeurs possible tend vers l'infini.
- La loi de probabilité cumulée discrète :

$$P(X \leq n_i) = \sum_{j:n_j \leq n_i} P(X = x_j)$$

remplacée par cette même somme multipliée par un coefficient "petit" et qui représente une aire approchée (formule des rectangles) qui est l'aire "sous la courbe" pour $x \leq X$, soit encore, on connaît

$$P(X \leq x) = \int_{-\infty}^{x} p(y)dy \tag{H.3}$$

p est appelée la *densité de la loi de probabilité (continue)*. Dans ce cas, on a pour a et b :

$$P(a \leq X \leq b) = \int_{-\infty}^{b} p(y)dy - \int_{-\infty}^{a} p(y)dy = \int_{a}^{b} p(y)dy \tag{H.4}$$

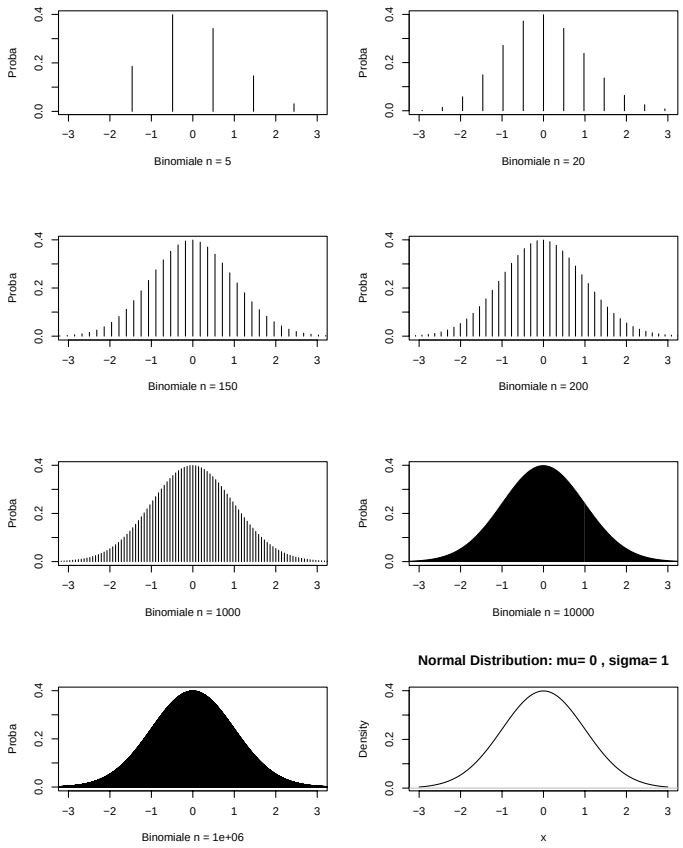


FIG. H.2. Les 7 graphes correspondant aux graphes de la loi binomiale normalisée avec $\pi = 0.3$ et $n \in \{5, 20, 150, 200, 1000, 10000, 1e+06\}$, ainsi que (pour la dernière courbe) la loi normale de moyenne $\mu = 0$ et d'écart-type $\sigma = 1$ avec une valeur maximale imposée.

Si $a = x$ et $b = x + dx$, où dx est une "petite variation",

$$P(x \leq X \leq x + dx) = \int_x^b x + dx p(y)dy \approx p(x)dx \tag{H.5}$$

Ainsi, la formule (H.2) s'écrit rigoureusement

$$P(a \leq X \leq b) = \int_a^b p(x)dx \tag{H.6}$$

où p est "la densité de la cloche idéale". Elle sera définie dans le chapitre 6.

– Enfin, les notions d'espérance et d'écart-type des variables discrète (voir (5.2) et définitions 5.21 et 5.22) "passe" à la limite, en remplaçant les sommes par des intégrales :

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xp(x)dx, \tag{H.7a}$$

$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 p(x)dx}. \tag{H.7b}$$

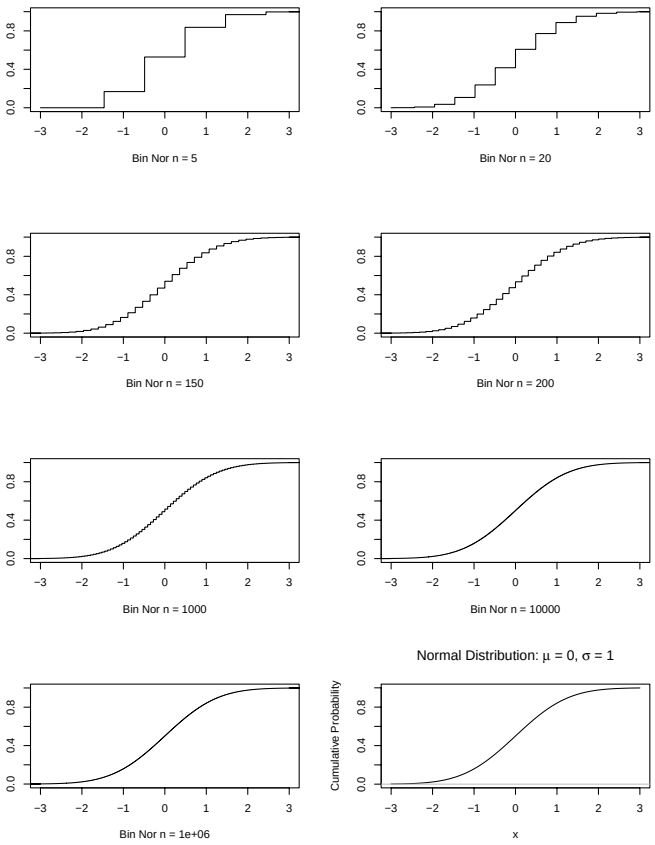


FIG. H.3. Les 7 graphes correspondant aux graphes des probabilités cumulées de la loi binomiale normalisée avec $\pi = 0.3$ et $n \in \{5, 20, 150, 200, 1000, 10000, 1e+06\}$, ainsi que (pour la dernière courbe) l'aire de la "cloche idéale".

H.3. Éléments de correction

ÉLÉMENTS DE CORRECTION DE L'EXERCICE H.1

Voir en figure H.4 les 6 loi de probabilité obtenues, qui ont l'air de "se rapprocher" d'une "courbe continue" quand n augmente.

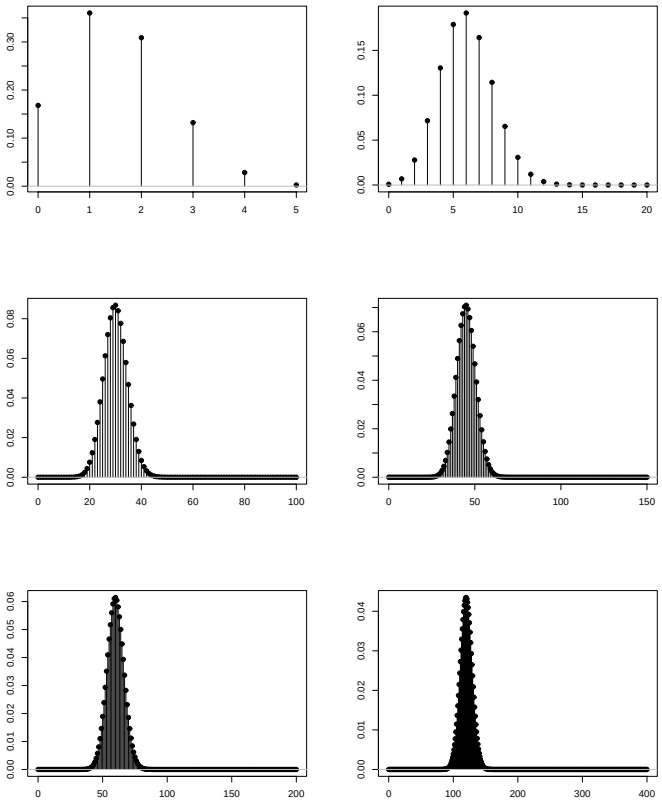


FIG. H.4. Les 6 graphes correspondant aux graphes de la loi binomiale avec $\pi = 0.3$ et $n \in \{5, 20, 100, 150, 200, 400\}$.

ANNEXE I

Lien entre lois de probabilité continue et histogramme en densité

Le principe de cette annexe est d'observer de façon expérimentale que si on fait un tirage aléatoire d'une variable aléatoire définie par sa loi continu et que l'on trace son histogramme en densité (voir remarque 3.2 page 8), ce dernier se rapproche du graphique de la densité de probabilité : la loi expérimentale se rapproche donc de la loi théorique (exactement comme dans le cas discret, voir définition 5.12 page 27). De plus, les moyennes et écart-type expérimentaux se rapprochent des moyennes et écart-type théoriques (exactement comme dans le cas discret, voir propriété 5.25 page 29).

Exercice I.1.

Comme dans l'exercice 5.36 page 32 (analogue discret), récupérer et sourcer la fonction `simule.norm.R`.

Cette fonction simule le modèle d'une loi normale de moyenne μ et d'écart-type σ, affiche l'histogramme en densité obtenu et le graphique de la densité de la loi de probabilité continue ainsi que les écarts entre les moyennes théoriques et observées et entre écarts-types théoriques et observés.

On choisit par exemple $\mu = 2$ et $\sigma = 0.3$.

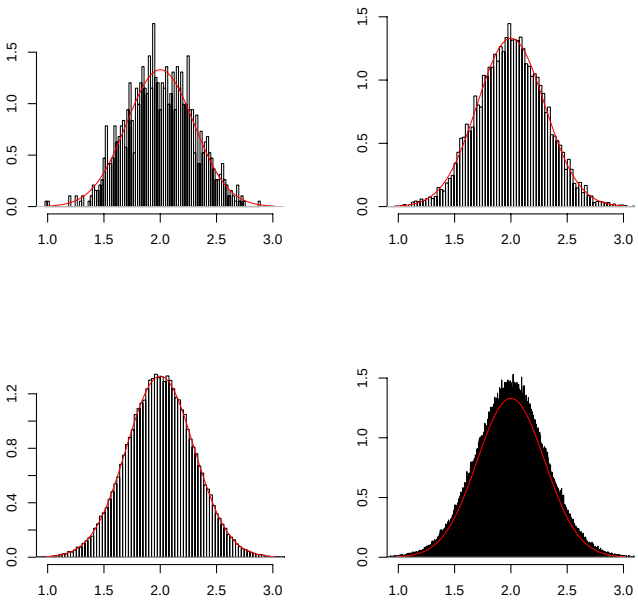


Fig. I.1. Quelques simulations de lois normales

Par exemple, en tapant respectivement

```
simule.norm(1000, 100, 2, 0.3)
simule.norm(10000, 100, 2, 0.3)
```

```
 simule.norm(1e+05, 100, 2, 0.3)
 simule.norm(1e+06, 10000, 2, 0.3)
```

vous devriez obtenir les écarts suivants

```
$ecm
[1] 0.01409247

$ecec
[1] 0.002404708
$ecm
[1] 0.003225252

$ecec
[1] 0.001254524
$ecm
[1] 0.0008862828

$ecec
[1] 0.001210139
$ecm
[1] 0.0005993887

$ecec
[1] 4.111430e-05
```

et les graphiques proche de ceux de la figure I.1 page précédente.

Bibliographie

[Cha04] Stéphane Champely. *Statistique vraiment appliquée au sport*. de Boeck, 2004. disponible à la BU de Lyon I sous la cote 519.5 CHA.

[Cha07] Stéphane Champely. Introduction à la statistique descriptive (sous R). Note de cours de l'UE de statistique L3MOS, disponible sous spiral, 2007.

[Coh98] J Cohen. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, publishers, 1998.

[DC08] AB Dufour et D. Chessel. Fiche de cours bs1 (cours de biostatistique, illustrations dans r) : Vraisemblance d'une hypothèse. Disponible sur `http://pbil.univ-lyon1.fr/R/enseignement` rubrique cours, puis test d'hypothèse, 2008.

[Par05] Émmanuel Paradis. R pour les débutants. disponible sur internet : `http://www.r-project.org/`, puis rubrique `Manuals`, puis `contributed documentation`, puis `Non-English Documents`, puis `French`, puis "R pour les débutants" by Emmanuel Paradis, the French version of "R for Beginners" (PDF)", ou alors directement sur `http://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf`, 2005.

[Rey98] Alain Rey, editor. *Le robert, dictionnaire historique de la langue française*. Dictionnaires le Robert, Paris, 1998.

[SJT+95] D Sabo, S.C Jansen, D Tate, M.C Duncan et D Leggett. *The portrayal of race, ethnicity and nationality in televised international athletic events*. AAF publications, 1995.