L2 Biochimie
Année 2014 – 2015

Polycopié du cours-TD
Méthodes et outils de calcul

Introduction à la méthodologie statistique

par Sébastien Gerchinovitz

Introduction

Informations pratiques

- Équipe enseignante :
  Sébastien Gerchinovitz sebastien.gerchinovitz@math.univ-toulouse.fr
  Tatiana Labopin-Richard tatiana.labopin@math.univ-toulouse.fr
  Diane Peurichard diane.peurichard@math.univ-toulouse.fr
- Modalités :
  - cours-TD structuré en : cours / exercices / mini-TP
  - évaluation (partiel 30% 1.5h + examen 70% 1.5h)
  - Des informations et documents de cours seront parfois déposés sur Moodle. On peut s'inscrire au cours en suivant le chemin :

    Sciences et technologies $\rightarrow$ $L2$ $\rightarrow$ Biochimie $\rightarrow$ BioCell/Méthodo.

Objectifs du cours-TD

- Prendre conscience de la présence d'incertitudes dans les observations expérimentales. Introduction des outils statistiques nécessaires pour traiter ces incertitudes.

  On s'appuiera régulièrement sur une expérience de biologie cellulaire que la promo réalisera bientôt pendant le "TP endocytose" : on traitera les données expérimentales via des méthodes statistiques.

- On s'intéressera aussi à des questions du type :

  1. Test clinique : au vu d'une base de données clinique, comment conclure sur l'efficacité ou l'absence d'efficacité d'un nouveau médicament par rapport à un médicament standard ?

  2. Agronomie : on dispose de deux fongicides différents pour traiter des plantations de maïs. Sur combien de plantations doit-on comparer ces fongicides avant de recommander un fongicide plutôt qu'un autre ?

  3. Sondage politique : 1000 personnes sont interrogées à la sortie des urnes du second tour des élections présidentielles françaises ; 49% affirment avoir voté pour le candidat A, et 51% pour le candidat B. Si vous étiez journaliste, quel pronostic annonceriez-vous à l'antenne ?

Tout l'enjeu est de répondre aux questions précédentes de façon quantitative, afin de prendre une décision mesurée. Les conséquences de ces décisions sont importantes, que ce soit sur les plans médicaux, économiques ou politiques.

Références bibliographiques

Ce polycopié est partiellement inspiré de notes de cours antérieures de Philippe Monnier et Muriel Casalis. Il puise également quelques explications ou illustrations des ouvrages suivants, que le lecteur pourra consulter afin d'approfondir les notions introduites dans ce cours :

- Richard Weber, *Statistics*, polycopié de cours de deuxième année à Cambridge disponible en anglais à l'adresse http://www.statslab.cam.ac.uk/~rrw1/stats/Sa5.pdf
- Vincent Rivoirard et Gilles Stoltz, *Statistique en action*, Vuibert, seconde édition, 2012.
- Denis D. Wackerly, William Mendenhall III, Richard L. Scheaffer, *Mathematical statistics with applications*, Thomson Brooks/Cole, seventh edition, 2008.
- Peter J. Bickel et Kjell A. Doksum, *Mathematical statistics : basic ideas and selected topics*, Peason Prentice-Hall, second edition, 2006.

Table des matières

Chapitre 1

Statistiques descriptives à l'échelle d'une population

Résumé Ce chapitre a pour objectif de faire prendre conscience de la variabilité présente dans une population, et d'introduire le vocabulaire pour la décrire.

1 Activité introductive : exemple de variabilité dans une population

Considérons une chorale constituée de 40 personnes, dont les tailles (en m) sont données par :

1.84	1.88	1.95	1.95	1.84	1.66	1.75	1.74
1.79	1.89	1.83	1.93	1.85	1.74	1.81	1.74
1.90	1.75	1.89	1.64	1.78	1.84	1.90	1.76
1.92	1.83	1.59	1.79	1.94	1.79	1.73	1.72
1.82	1.74	1.68	1.75	1.90	1.71	1.96	1.76

En langage statistique, on dira que le tableau de données ci-dessus correspond à une *population*. A l'inverse, un *échantillon* correspondrait à un jeu de données restreint, obtenu par sondage, comme par exemple :

1.89 1.74 1.68 1.74 1.84

Dans ce chapitre, nous allons uniquement nous intéresser à la description d'une population. Sur l'exemple de la chorale, comment peut-on procéder pour dégager des informations des 40 données disponibles ?

2 Représentations graphiques des valeurs d'une population

Une première idée de visualisation consisterait à représenter les individus par des points sur un graphique : le numéro de la donnée en abscisses, et sa valeur en ordonnées, cf. figure 1.1.

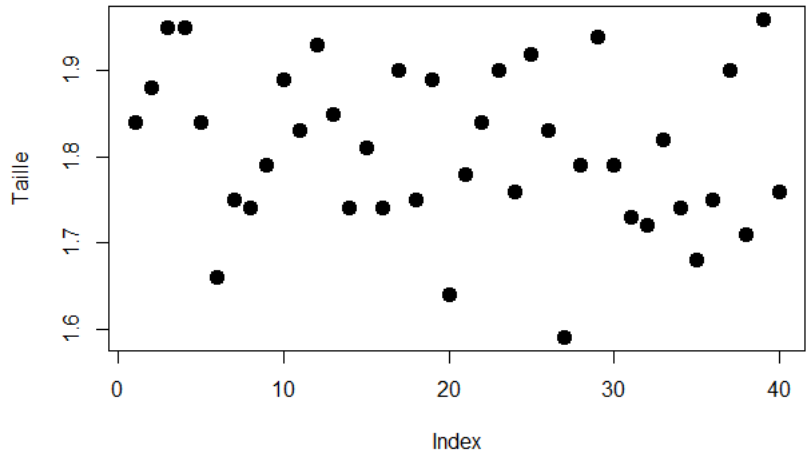


FIGURE 1.1 – Une première représentation des données.

La représentation précédente n'est pas très parlante. Une deuxième idée, plus informative, consiste à tracer un *diagramme en bâtons*, avec la taille en abscisses, et l'effectif (= nombre de personnes) en ordonnées.

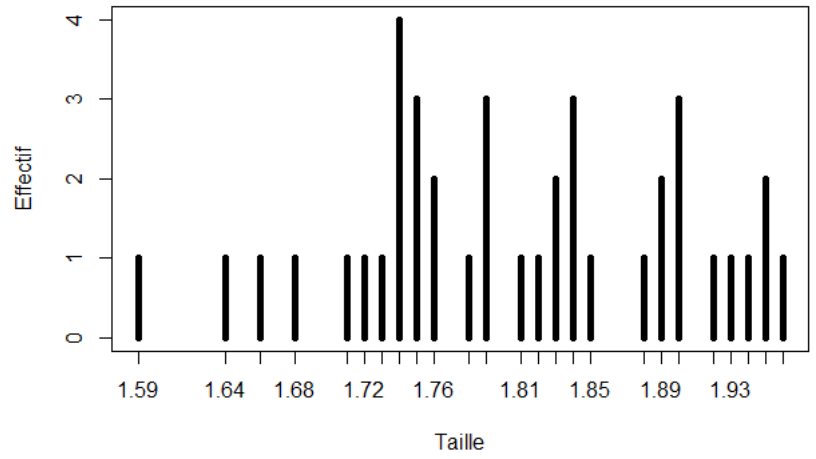


FIGURE 1.2 – Exemple de diagramme en bâtons.

Le diagramme en bâtons précédent n'est pas suffisamment synthétique car de nombreuses tailles ne sont présentes qu'une seule fois dans la population. Un outil graphique plus adapté dans ce cas est l'*histogramme*. Pour l'obtenir, on subdivise les tailles en plusieurs intervalles (qu'on appelle *classes*). Puis, pour chaque intervalle, on trace un rectangle vertical dont l'aire est égale à la proportion d'individus ayant cette taille. Concrètement, la hauteur h d'un rectangle est donnée par :

$$h = \frac{\text{proportion d'individus dans l'intervalle}}{\text{largeur de l'intervalle}} .$$

Par exemple : on obtient $h = (4/40)/L$ si l'intervalle est de largeur L et s'il contient 4 individus parmi les 40 individus de la population. L'exemple de la chorale est illustré en figure 1.3.

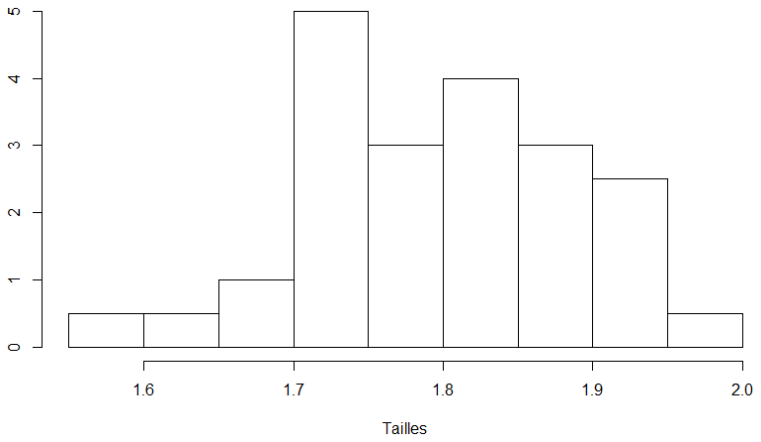


FIGURE 1.3 – Exemple d'histogramme des tailles de la chorale.

L'histogramme peut être aussi utilisé dans le cas où les intervalles ne sont pas de même largeur. Cela peut par exemple être intéressant s'il y a peu d'individus sur une grande plage de valeurs ; dans ce cas, on peut regrouper ces valeurs dans un même intervalle (cf. figure 1.4). Une règle grossière est qu'il ne faut ni trop d'intervalles (car on veut suffisamment de points par intervalle), ni trop peu d'intervalles (car sinon, l'histogramme ne serait pas suffisamment informatif).

Attention dans tous les cas à bien faire en sorte que ce soit **l'aire** et non la hauteur du rectangle qui représente la proportion d'individus dans l'intervalle. Ce choix permet de déterminer, presque *à vue d'oeil*, la proportion d'individus de tailles comprises entre t_1 et t_2 : il suffit d'évaluer l'aire des rectangles entre t_1 et t_2 (à comparer à la somme des aires de tous les rectangles, qui vaut 1, c'est-à-dire 100%). Par exemple, la figure 1.4 indique qu'il y a à peu près 30% des individus dont la taille est comprise entre 1.55 m et 1.75 m.

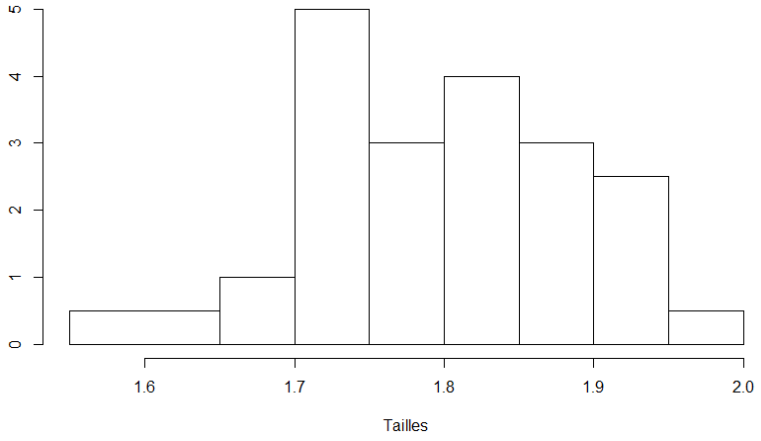


FIGURE 1.4 – Même histogramme, mais avec des intervalles de largeurs différentes.

3 Grandeurs décrivant la répartition des valeurs d'une population

Introduisons maintenant quelques grandeurs quantitatives pour résumer certaines informations sur la population. On distingue les *mesures de position* qui indiquent une tendance centrale (moyenne et médiane), et les *mesures de dispersion* qui reflètent la variabilité de la population autour de cette tendance centrale (écart-type et écart interquartile).

Remarque importante : dans tout ce qui suit, on ne s'intéresse qu'à des valeurs *quantitatives* (taille, âge, nombre d'enfants, etc) et non à des valeurs *qualitatives* (genre, style de musique préféré, prénom, etc). Ces dernières se traitent un peu différemment.

3.1 Mesures de position : moyenne et médiane

3.1.1 Moyenne de population

La *moyenne de population* est la moyenne arithmétique de toutes les valeurs de la population (avec les redondances éventuelles). Ainsi, si la population compte N individus, et que les valeurs sont $x_1, x_2, x_3, \ldots, x_N$, alors la moyenne de population μ_{pop} est définie par

$$\mu_{\text{pop}} = \frac{x_1 + x_2 + x_3 + \ldots + x_N}{N} \overset{\text{notation}}{=} \frac{1}{N} \sum_{i=1}^{N} x_i \, .$$

Dans l'exemple de la chorale, la moyenne de population est égale à $\mu_{\text{pop}} = 1.807$ m.

3.1.2 Médiane de population

La *médiane de population* partage l'ensemble de toutes les valeurs de la population en deux groupes d'effectifs égaux : au moins 50% des valeurs sont inférieures ou égales, et au moins 50% des valeurs sont supérieures ou égales.

Prenons un exemple de population volontairement très simple, avec 5 individus, et dont les valeurs ordonnées sont :

$$1.05 \quad 1.4 \quad \boxed{2.8} \quad 3.1 \quad 6$$

Dans ce cas, la médiane de population est égale à 2.8. Cet exemple est facile car le nombre d'individus dans la population est impair, si bien qu'il existe une valeur "au milieu" des autres. Dans le cas où l'effectif de la population est pair, il y a plutôt deux valeurs "au milieu" des autres, et il est d'usage de retenir la moyenne des deux. Exemple :

$$1.05 \quad 1.4 \quad \boxed{2.8} \quad \boxed{3.1} \quad 6 \quad 7.4$$

Ici, la médiane de population vaut $(2.8 + 3.1)/2 = 2.95$.

Application numérique : dans l'exemple de la chorale, la médiane de population vaut 1.8 m.

3.1.3 Différence entre moyenne et médiane

Fait important : la moyenne est plus sensible aux valeurs extrêmes que la médiane. Prenons l'exemple d'une petite entreprise de 9 personnes, dont les salaires nets mensuels (en euros) sont donnés par :

1160	1160	1160	1160	1160	1500	1500	1700	3500

Dans cet exemple, la médiane de population vaut 1160 euros, alors que la moyenne de population vaut environ 1556 euros. Si le salaire du patron était encore plus élevé, par ex de 8500 euros au lieu de 3500, alors la moyenne passerait à 2111 euros environ, mais la médiane resterait inchangée. La médiane est donc moins sensible aux valeurs extrêmes.

3.2 Mesures de dispersion : écart-type et écart interquartile

3.2.1 Écart-type de population

L'écart-type de population est un indicateur de l'amplitude des variations des valeurs de la population autour de leur moyenne. Ainsi, deux populations qui ont la même moyenne mais des écarts-type différents n'ont pas la même variabilité (un écart-type elévé correspond à une grande variabilité).

Plus précisément, si une population de N individus est formée des valeurs $x_1, x_2, x_3, \ldots, x_N$, on définit d'abord la *variance de population* comme étant la valeur moyenne des carrés des écarts à la moyenne μ_{pop} :

$$\begin{aligned} \text{Var}_{\text{pop}} &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{\text{pop}})^2 \\ &= \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \mu_{\text{pop}}^2 \,. \end{aligned}$$

Les deux formules ci-dessus sont importantes :

- La première ligne est plus facile à interpréter : la variance est d'autant plus grande qu'un grand nombre de valeurs x_i sont éloignées de la moyenne μ_{pop} ; autrement dit, la variance est grande lorsque la population présente beaucoup de variabilité.
- La formule de la seconde ligne, qui s'obtient par un petit calcul, est également utile car souvent plus facile à appliquer.

On définit ensuite *l'écart-type de la population* σ_{pop} comme étant la racine-carrée de la variance :

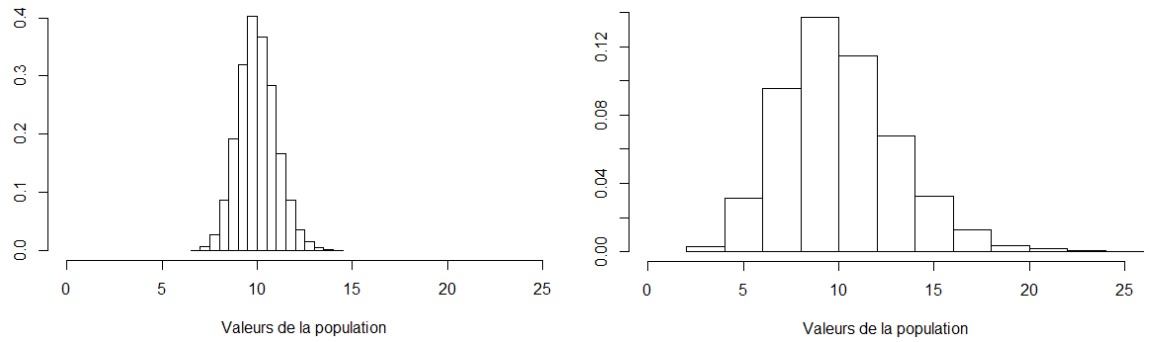
$$\sigma_{\text{pop}} = \sqrt{\text{Var}_{\text{pop}}} \,.$$

Pourquoi la racine carrée ? Dans l'exemple de la chorale, les valeurs sont des tailles (en mètres), donc la variance s'exprime en m^2 (mètres au carré). En prenant la racine carrée, on revient à des mètres (l'écart-type s'exprime en m).

Application numérique : l'écart-type des tailles de la chorale vaut $\sigma_{\text{pop}} \approx 0.09$ m.

Remarque. Attention : lors des applications numériques, il ne faut arrondir que les résultats finaux, en conclusion de question (3 chiffres significatifs par exemple). Ne pas arrondir les résultats de calculs intermédiaires (cela peut entraîner de très grandes erreurs d'arrondis). Ainsi, même si vous avez déjà calculé une moyenne μ_{pop} à la question précédente, vous devez, pour le calcul de la variance, reprendre la valeur de μ_{pop} avec le plus de décimales possibles sur la calculatrice (ne pas utiliser le résultat arrondi de la question précédente).

On décrit en figure 1.5 deux exemples de populations de moyennes égales, mais d'écarts-types différents. Question : laquelle correspond à l'écart-type le plus grand ?



(a) Population n° 1 (b) Population n° 2

FIGURE 1.5 – Exemples de deux populations de moyennes égales ($\mu_{\text{pop}} = 10$), mais d'écarts-types différents ($\sigma_{\text{pop}} = 1$ ou 3).

3.2.2 Écart interquartile de population

L'écart-type n'est pas la seule mesure de dispersion. En effet, deux populations de moyennes et d'écarts-types égaux peuvent malgré tout avoir des histogrammes de formes différentes.

De même que nous avons défini la médiane d'une population, on peut définir le *premier quartile* $Q1$ et le *troisième quartile* $Q3$ de la façon suivante :

$$\overbrace{x_1 \quad x_2 \quad x_3 \quad \dots \quad Q1}^{\text{au moins } 25\% \text{ des individus}} \quad \underbrace{\dots \quad \dots \quad \dots \quad x_{N-2} \quad x_{N-1} \quad x_N}_{\text{au moins } 75\% \text{ des individus}}$$

$$\overbrace{x_1 \quad x_2 \quad x_3 \quad \dots \quad \dots \quad \dots \quad Q3}^{\text{au moins } 75\% \text{ des individus}} \quad \underbrace{\dots \quad x_{N-2} \quad x_{N-1} \quad x_N}_{\text{au moins } 25\% \text{ des individus}}$$

La médiane correspond quant à elle au *second quartile* $Q2$.
Application numérique : dans l'exemple de la chorale, $Q1 = 1.74$ m et $Q3 = 1.89$ m.

On appelle *écart interquartile* d'une population la différence $Q3 - Q1$. Cette différence est la largeur d'un intervalle qui contient au moins 50% de la population. Plus cette différence est élevée, et plus les valeurs de la population sont dispersées.

3.3 Récapitulatif graphique : le boxplot

Le *boxplot*, ou *boîte à moustaches*, est un outil graphique pour représenter simplement la médiane et la dispersion des valeurs d'une population. Typiquement, un boxplot ressemble à la figure 1.6. Il est défini de la façon suivante :

- Le rectangle du milieu s'étend du premier quartile $Q1$ au troisième quartile $Q3$, et le trait vertical du milieu correspond à la médiane.
- Les traits horizontaux, qu'on appelle les *moustaches*, englobent encore davantage de valeurs de la population. Pour simplifier, la convention que nous utiliserons est de placer les extrémités des moustaches au premier décile (10% des valeurs en dessous) et au neuvième décile (10% des valeurs au-dessus). Attention néanmoins : en biologie, la convention est souvent un peu différente (longueur des moustaches égale à 1.5 fois l'écart interquartile).

- Si quelques valeurs de la population ne sont pas comprises entre les extrémités des moustaches, celles-ci sont figurées par des points à part ; on parle de *valeurs extrêmes* ou *outliers*.

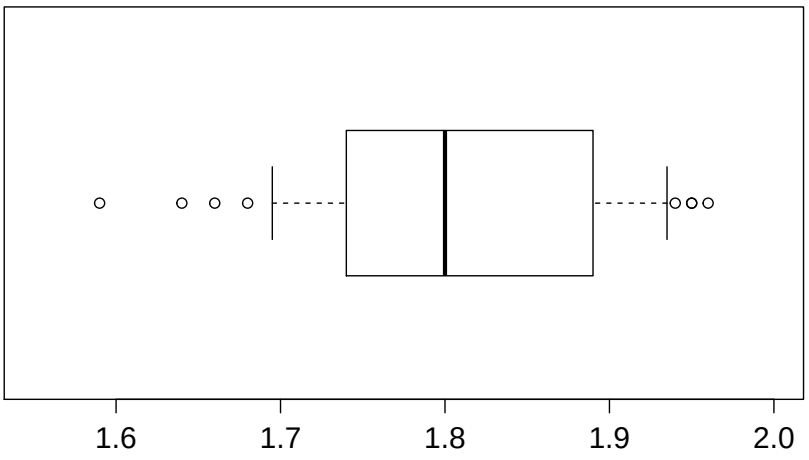


FIGURE 1.6 – Boxplot des tailles des membres de la chorale.

4 Exercices

Exercice 1. La liste des 30 communes françaises les plus peuplées en 2011 est donnée ci-dessous (figure 1.7). Pour plus de lisibilité, on pourra convertir (et arrondir) les effectifs des communes en milliers d'habitants.

1. Tracer un histogramme des effectifs des 30 premières communes françaises.
2. Quel est l'effectif moyen parmi ces 30 communes ? Que vaut la médiane ? Commenter la différence.
3. Que vaut l'écart-type des effectifs des 30 communes ? Est-il élevé ?
4. Déterminer l'écart interquartile des effectifs des 30 communes, puis tracer un boxplot.

Paris	2 249 975	Grenoble	157 424
Marseille	850 636	Dijon	151 672
Lyon	491 268	Angers	148 803
Toulouse	447 340	Saint-Denis	145 347
Nice	344 064	Villeurbanne	145 034
Nantes	287 845	Nîmes	144 940
Strasbourg	272 222	Le Mans	143 240
Montpellier	264 538	Clermont-Ferrand	140 957
Bordeaux	239 399	Aix-en-Provence	140 684
Lille	227 533	Brest	140 547
Rennes	208 033	Limoges	137 758
Reims	180 752	Tours	134 633
Le Havre	174 156	Amiens	133 327
Saint-Étienne	170 049	Metz	119 962
Toulon	163 974	Perpignan	118 238

FIGURE 1.7 – Effectifs des 30 premières communes françaises.

Exercice 2 (Manipulation du symbole Σ)**.** On considère la suite des 7 valeurs $x_1, x_2, \ldots, x_7$ suivantes :

x_1	x_2	x_3	x_4	x_5	x_6	x_7
2	1	5	3	1.5	4	0.5

1. Calculer $\displaystyle\sum_{i=1}^{7} x_i$ et $\displaystyle\sum_{i=2}^{6} x_i$.
2. Calculer $\displaystyle\sum_{i=1}^{7} x_i^2$.
3. Calculer la moyenne et l'écart-type du jeu de données précédent.

Exercice 3. Les tailles en pouces des 105 étudiants d'un cours de biostatistique aux États-Unis sont réparties selon l'histogramme de la figure 1.8 (NB : 1 pouce = 2.54 cm).

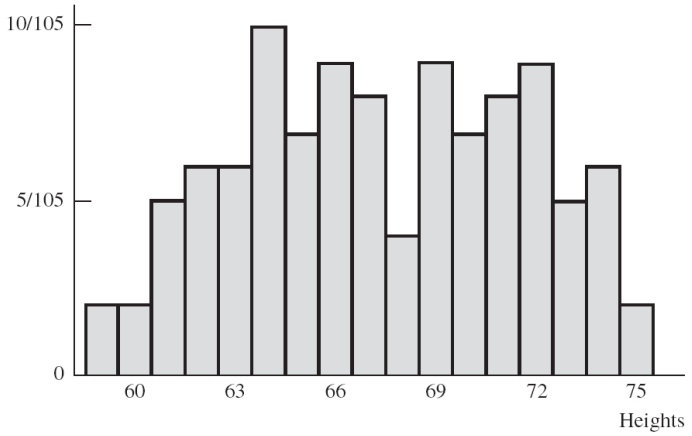


FIGURE 1.8 – Les tailles des 105 étudiants.

1. Décrire la forme de l'histogramme.
2. Cet histogramme présente-t-il une caractéristique particulière ?
3. Voyez-vous une raison particulière expliquant les deux pics de l'histogramme ? Une autre variable que la taille serait-elle utile pour expliquer ces deux pics ?

Exercice 4 (Formule de la variance)**.** $\diamondsuit$
On va expliquer pourquoi la variance de population peut s'exprimer des deux façons équivalentes suivantes :

$$\mathrm{Var}_{\mathrm{pop}} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{\mathrm{pop}})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \mu_{\mathrm{pop}}^2 \,.$$

1. Montrez que $(x_i - \mu_{\mathrm{pop}})^2 = x_i^2 - 2\,\mu_{\mathrm{pop}}\,x_i + \mu_{\mathrm{pop}}^2$ pour tout $i \in \{1, \ldots, N\}$.
2. En déduire que
$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{\mathrm{pop}})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \mu_{\mathrm{pop}}^2 \,.$$

Exercice 5. Cet exercice est extrait de l'ouvrage *Mathematical statistics with Applications* de Wackerly, Mendenhall et Scheaffer (2008, chapitre 1). A vous de traduire...

Are some cities more windy than others? Does Chicago deserve to be nicknamed "The Windy City"? Given below are the average wind speeds (in miles per hour) for 45 selected U.S. cities:

8.9	12.4	8.6	11.3	9.2	8.8	35.1	6.2	7.0
7.1	11.8	10.7	7.6	9.1	9.2	8.2	9.0	8.7
9.1	10.9	10.3	9.6	7.8	11.5	9.3	7.9	8.8
8.8	12.7	8.4	7.8	5.7	10.5	10.5	9.6	8.9
10.2	10.3	7.7	10.6	8.3	8.8	9.5	8.8	9.4

Source: The World Almanac and Book of Facts, 2004.

a Construct a relative frequency histogram for these data. (Choose the class boundaries without including the value 35.1 in the range of values.)

b The value 35.1 was recorded at Mt. Washington, New Hampshire. Does the geography of that city explain the magnitude of its average wind speed?

c The average wind speed for Chicago is 10.3 miles per hour. What percentage of the cities have average wind speeds in excess of Chicago's?

d Do you think that Chicago is unusually windy?

Exercice 6. Quel message peut-on tirer de ce dessin humoristique ?

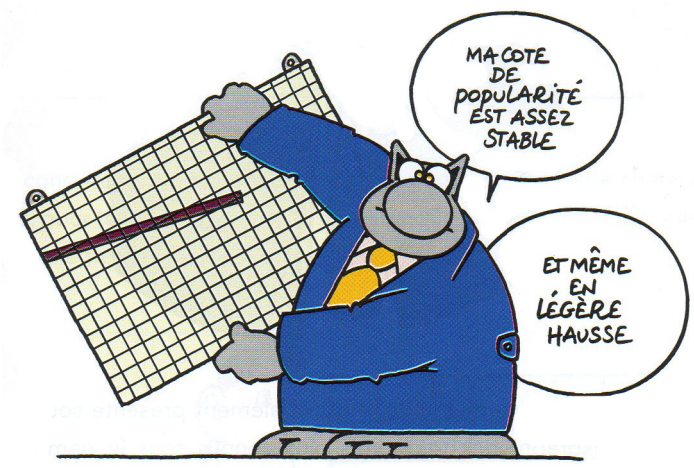


Image extraite de l'ouvrage *La mathématique du chat de Philippe Geluck*, Daniel Justens.

A Introduction au logiciel R

On liste ci-dessous quelques commandes utiles pour démarrer avec le logiciel R. Une liste de commande s'appelle un *script*. Tout le texte qui est placé après un symbole # n'est pas pris en compte par R ; il s'agit de commentaires laissés par l'utilisateur.

A.1 Quelques commandes utiles sous R

```
## Opérations usuelles :

1+2
2*5
3^2
sqrt(2) # racine carrée
exp(1)  # exponentielle
log(2)  # logarithme népérien

## Manipulation de vecteurs :

v=c(1,3.1,1,6.8,2,-0.4,-7,6.8,-3.2,6.8) # construction d'un vecteur
v
v[2] # 2ème élément de v
v[1:3] # extraction des 3 premiers éléments de v

sum(v) # somme des valeurs
mean(v) # moyenne des valeurs
mean(v^2) - mean(v)^2 # variance

median(v) # médiane
quantile(v,0.25,type=2) # Q1 (l'option "type = 2" permet de retenir
                              la moyenne de deux valeurs adjacentes)
quantile(v,0.5,type=2) # Q2 = médiane !
quantile(v,0.75,type=2) # Q3

sort(v) # classement par ordre croissant
sort(v,decreasing=TRUE) # par ordre décroissant


## Représentations graphiques

plot(v,type='p',lwd=6,main="Simple graphique",xlab="Index",ylab="Valeur")
plot(table(v),type="h",lwd=5,main="Diagramme en batons", xlab="Taille",
     ylab="Effectif")

hist(v)
hist(v,freq=FALSE) # histogramme avec la règle "aire = proportion"
hist(v,freq=FALSE,main="Titre",xlab="Titre x",ylab="Titre y")

boxplot(v) # convention longueur des moustaches : au max 1.5 (Q3-Q1)
boxplot(v,horizontal=TRUE)
```

A.2 Résolution de l'exercice 4 avec R

```
# On construit le vecteur des vitesses du vent :
wind = c(8.9, 12.4, 8.6, 11.3, 9.2, 8.8, 35.1, 6.2, 7.0, 7.1,
         11.8, 10.7, 7.6, 9.1, 9.2, 8.2, 9.0, 8.7, 9.1, 10.9,
         10.3, 9.6, 7.8, 11.5, 9.3, 7.9, 8.8, 8.8, 12.7, 8.4,
         7.8, 5.7, 10.5, 10.5, 9.6, 8.9, 10.2, 10.3, 7.7, 10.6,
         8.3, 8.8, 9.5, 8.8, 9.4)
length(wind) # nombre total de valeurs

# Question a : histogramme
# L'option "freq=FALSE" correspond à la règle : aire = proportion
hist(wind,freq=FALSE);

# Pour écarter la valeur atypique (35.1), rangeons les valeurs
# par ordre croissant :
sort(wind)
sort(wind)[1:44]

hist(sort(wind)[1:44],freq=FALSE,main="Histogramme des vitesses
     du vent aux US",xlab="Vitesse moyenne du vent",ylab="")

# Question c : combien de villes sont plus venteuses que Chicago ?
sum(wind>10.3)
```

Chapitre 2

Estimation des caractéristiques d'une population par échantillonnage

Résumé Dans ce chapitre, on explique comment faire de l'inférence statistique, c'est-à-dire, comment déduire des propriétés d'une population à partir de l'observation d'un échantillon. On introduira à cette fin les notions importantes d'*expérience aléatoire* et de *variable aléatoire*.

1 Introduction à l'inférence statistique

1.1 Contexte : observation d'un échantillon de la population

Supposons qu'on souhaite étudier certaines propriétés d'une population, comme, par exemple, la taille moyenne des Toulousains, ou la proportion des électeurs français qui vont voter pour le maire sortant de leur ville. Ces questions concernent des populations (les Toulousains, ou les électeurs français). En pratique, on n'a cependant que très rarement accès aux données d'une population dans sa totalité (exception : recensement). On recourt donc à l'étude d'un échantillon de cette population. Par exemple, dans le cas de la chorale, on pourrait observer :

?	?	?	?	1.84	?	?	?
?	1.89	?	?	?	?	?	1.74
?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?
?	1.74	1.68	?	?	?	?	?

Dans cet exemple, l'*échantillon* correspond au jeu de données restreint :

1.89 1.74 1.68 1.74 1.84

L'*inférence statistique* consiste, à partir de l'observation d'un échantillon d'une population, à en déduire certaines propriétés de la population toute entière (avec, néanmoins, des échantillons un peu plus grands que dans l'exemple simpliste précédent).

1.2 Deux estimateurs naturels de la moyenne et de la variance d'une population

Estimateur de la moyenne de population Si on ne connait pas la moyenne d'une population, une approche très naturelle consiste à l'estimer par la moyenne de l'échantillon observé. Dans l'exemple précédent, on est tenté de dire que la moyenne de population est proche de

$$\frac{1.89 + 1.74 + 1.68 + 1.74 + 1.84}{5} = 1.778\ m$$

Estimateur de la variance de population De même, si on ne connait pas la variance d'une population, une approche très naturelle consiste à l'estimer par la variance de l'échantillon observé. Ainsi, dans l'exemple précédent, on est tenté de dire que la variance de population est proche de

$$\frac{1.89^2 + 1.74^2 + 1.68^2 + 1.74^2 + 1.84^2}{5} - 1.778^2 = 0.005776\ m^2$$

Le but de ce chapitre est de donner un cadre mathématique pour comprendre la qualité de ces estimations.

2 L'échantillonnage est une expérience aléatoire

On appelle *échantillonnage* le procédé qui consiste à choisir un échantillon dans une population. Nous allons expliquer pourquoi il s'agit d'une expérience aléatoire, et quelles propriétés cette expérience aléatoire doit vérifier.

2.1 Pourquoi choisir l'échantillon aléatoirement ?

Reprenons l'exemple de la chorale. Imaginons un instant que vous soyez déterminé à l'avance à choisir des individus bien précis dans la chorale, par exemple, les numéros 2, 17, 24, 25 et 38. Il se peut que la moyenne des tailles de ces individus soit proche de la moyenne de toutes les tailles de la chorale, mais il se peut aussi très bien que les individus que vous avez choisis soient très particuliers. Dans ce dernier cas, votre estimation de la taille moyenne de la chorale sera très mauvaise.

A l'inverse, une façon de choisir des individus représentatifs de la population consiste à les choisir complètement au hasard dans la population. En effet, dans ce cas, vous observerez plus souvent des tailles bien représentées et moins souvent des tailles peu représentées. On s'attend donc à ce qu'avec grande probabilité, vous observiez un échantillon représentatif de la population. Peut-être observerez-vous, par malchance, un échantillon très peu représentatif de la population, mais ce cas ne se produira qu'avec une très faible probabilité.

2.2 Pour être vraiment aléatoire, le tirage de l'échantillon doit respecter certaines propriétés

Dans toute la suite du cours, on s'intéressera à l'échantillonnage suivant : les valeurs de l'échantillon sont tirées aléatoirement dans la population et indépendemment les unes des autres. Formalisons un tout petit peu le problème :

1. On choisit un individu complètement aléatoirement dans la population, et on appelle X_1 la valeur observée.
2. Puis, indépendemment du premier tirage, on choisit à nouveau un individu complètement aléatoirement dans la population, et on appelle X_2 la valeur observée.
3. Puis, indépendemment des deux premiers tirages, on choisit à nouveau un individu complètement aléatoirement dans la population, et on appelle X_3 la valeur observée.
   $[\ldots]$

n. Enfin, indépendemment des $n-1$ premiers tirages, on choisit à nouveau un individu complètement aléatoirement dans la population, et on appelle X_n la valeur observée.

Remarque : puisque les tirages sont indépendants les uns des autres, il est possible qu'on choisisse deux fois le même individu dans la population (même si c'est peu probable).

Définition 1 (Variable aléatoire)**.** Les quantités $X_1, X_2, \ldots, X_n$ sont notées en majuscules pour signifier que leurs valeurs sont aléatoires au sens où elles dépendent de l'échantillon choisi. Si deux instituts de sondage réalisaient en parallèle le même sondage, il est fort probable qu'ils observeraient des valeurs différentes. On dit que les quantités $X_1, X_2, \ldots, X_n$ (avant choix effectif de l'échantillon) sont des *variables aléatoires*.

En revanche, une fois l'échantillon réellement choisi et observé, les valeurs de ces quantités deviennent précises, et on note généralement leurs valeurs par des minuscules $x_1, x_2, \ldots, , x_n$. Par exemple :

x_1	x_2	x_3	x_4	x_5
1.89	1.74	1.68	1.74	1.84

On peut alors procéder à une application numérique.

2.3 Estimateurs de la moyenne et de la variance d'une population

Utilisons les notations précédentes, c'est-à-dire, on dispose d'un échantillon $X_1, X_2, \ldots, X_n$ de valeurs tirées aléatoirement et indépendemment les unes des autres dans une population.

Estimateur de la moyenne de population On définit *la moyenne d'échantillon* $\mu_{\acute{e}ch}$ par

$$\mu_{\text{éch}} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \, .$$

C'est la même formule que pour la moyenne de population, mais elle est restreinte aux valeurs de l'échantillon. Remarque importante : la moyenne d'échantillon $\mu_{\text{éch}}$ est encore une variable aléatoire, puisque sa valeur dépend de l'échantillon choisi. En revanche, comme nous l'expliquerons au chapitre suivant, sa valeur est souvent proche de la vraie moyenne μ_{pop} de la population. On dit donc que $\mu_{\text{éch}}$ est un estimateur de μ_{pop}.

Autre remarque : la moyenne d'échantillon est parfois notée $\bar{X}_n$, qui se lit : "moyenne des X_i pour i allant de 1 à n".

Estimateur de l'écart-type de population On définit *l'écart-type d'échantillon* $\sigma_{\acute{e}ch}$ par

$$\sigma_{\text{éch}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\big(X_i - \mu_{\text{éch}}\big)^2} = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \mu_{\text{éch}}^2} \, .$$

Ici encore, il s'agit de la même formule que pour l'écart-type de population, mais celle-ci est restreinte aux valeurs de l'échantillon. Tout comme précédemment, l'écart-type d'échantillon $\sigma_{\text{éch}}$ est une variable aléatoire, puisque sa valeur dépend de l'échantillon choisi. On s'attend néanmoins à ce que sa valeur soit souvent proche de l'écart-type σ_{pop} de la population ; on dit que $\sigma_{\text{éch}}$ est un estimateur de σ_{pop}.

Autre estimateur de l'écart-type de population Pour estimer l'écart-type σ_{pop} d'une population, il est fréquent d'utiliser un estimateur légèrement différent, où on divise par $n-1$ au lieu de diviser par n (pour des raisons qu'on expliquera au chapitre suivant). On définit ainsi *l'écart-type d'échantillon corrigé* $s_{\text{éch}}$ par

$$s_{\text{éch}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\big(X_i - \mu_{\text{éch}}\big)^2} = \sqrt{\frac{n}{n-1}}\,\sigma_{\text{éch}} \, .$$

3 Autres exemples d'expériences aléatoires

Un sondage de n valeurs dans une population est une expérience aléatoire. Il existe en fait de nombreuses autres expériences aléatoires avec des propriétés similaires.

3.1 Des expériences aléatoires "jouets"

3.1.1 Tirage de boules dans une urne

Considérons l'expérience suivante : une urne contient N_0 boules blanches et N_1 boules noires ; on tire avec remise n boules dans l'urne parfaitement au hasard ; à chaque tirage, on note la valeur 0 si la boule est blanche et la valeur 1 si la boule est noire.

L'expérience précédente est une expérience aléatoire très semblable au sondage : les tirages sont parfaitement aléatoires et indépendants les uns des autres (puisque que le tirage est *avec remise*). La population est constituée de N_0 boules blanches et N_1 boules noires, et l'échantillon sera constitué de n valeurs $X_1, X_2, \ldots, X_n$, où chaque valeur X_i vaut 0 ou 1.

Dans ce cadre, faire de l'inférence statistique peut consister à estimer la proportion de boules noires dans l'urne (qui est inconnue de l'expérimentateur) par la proportion de boules noires dans l'échantillon observé.

3.1.2 Lancers de pièces

Un autre exemple classique d'expérience aléatoire consiste à lancer successivement et indépendemment une même pièce équilibrée (c'est-à-dire, qui tombe avec probabilité $1/2$ sur pile et avec probabilité $1/2$ sur face), puis à noter la suite des résultats obtenus (1 pour "pile" et 0 pour "face").

Activité 1 (Arriverez-vous à reproduire le hasard ?)**.**

- Mettez-vous par groupe de deux.
- L'un des deux devra lancer 100 pièces de 1 € sans tricher (par exemple en déposant à chaque lancer la pièce dans un gobelet, en mélangeant, puis en retournant le gobelet). Notez la suite des résultats (1 = pile ou 0 = face) sur une feuille de papier. Par exemple, si les deux premiers lancers sont des "pile" : $1 \quad 1 \quad \ldots$
- L'autre personne devra écrire une suite de 100 résultats (1 ou 0) mais sans lancer réellement de pièces ; il faudra *imaginer* des lancers de pièces successifs. Notez les résultats sur une feuille de papier séparée.

3.2 Expériences en biologie et physique : où est l'aléatoire ?

Dans les sciences expérimentales, les résultats sont souvent accompagnés d'incertitudes statistiques. Quelles formes prennent-elles ?

Echantillonnage dans une population De nombreuses expériences en laboratoire de biologie visent à étudier certains phénomèmes sur une espèce animale ou végétale (ex : longévité, risque de développement de tumeur, etc). Tous les individus de l'espèce ne se comportent peut-être pas exactement de la même façon, et on peut vouloir connaître le comportement moyen dans l'espèce ainsi que la variabilité associée.

Puisqu'on ne peut pas examiner toute l'espèce, seul un échantillon sera observé. On est donc dans le même cas que celui du sondage : les résultats expérimentaux dépendent de l'échantillon prélevé, mais on a espoir qu'ils soient une bonne indication des caractéristiques de l'espèce entière.

Bruits de mesure En physique comme en biologie, il est fréquent de ne mesurer une quantité qu'avec un certain degré de précision. Les capteurs des appareils de mesure sont en effet sujets à des erreurs, faibles, mais existantes. Puisque deux mesures indépendantes peuvent donner deux résultats (proches mais) différents, le protocole de mesure peut être considéré comme une expérience aléatoire, dont le résultat fluctue autour d'une vraie valeur.

Autres formes d'aléa Il existe d'autres raisons de considérer les phénomènes observés comme étant aléatoires. Parfois, les conditions initiales d'un protocole expérimental ne sont pas parfaitement maîtrisées et les résultats de l'expérience, qui en dépendent, fluctueront d'un expérimentateur à l'autre.

Une autre forme d'aléa, plus conceptuelle, survient aussi en modélisation physique ou biologique : il arrive qu'on ne dispose que d'un modèle simplifié de la réalité pour décrire les observations, de sorte que les observations fluctueront autour des valeurs prédites ; l'erreur de prévision est alors modélisée comme un bruit aléatoire.

3.3 Exemples où les n tirages ne sont pas indépendants

La contrainte consistant à effectuer des tirages *indépendants* dans la population n'est pas anodine. Il est des cas où elle n'est pas vérifiée. Prenons un exemple : on souhaite connaître l'âge moyen des Français ; on interroge d'abord quelqu'un complètement au hasard, puis on lui demande les coordonnées d'un de ses amis, qu'on interroge, puis d'un ami de cet ami, qu'on interroge, etc. Dans ce cas, les choix des individus ne sont pas indépendants les uns des autres. Nous n'allons pas observer un échantillon d'âges représentatif de la population française, mais plutôt un échantillon d'âges probablement proches de celui de la première personne interrogée.

4 Caractéristiques importantes d'une variable aléatoire

Rappelons qu'une variable aléatoire est une quantité qui dépend du résultat de l'expérience aléatoire étudiée. On note généralement les variables aléatoires par des majuscules (X, Y, Z, etc). On décrit ci-après trois caractéristiques importantes d'une variable aléatoire : sa loi, son espérance, et sa variance.

4.1 Loi d'une variable aléatoire

Soit X une variable aléatoire. Si X ne prend qu'un nombre fini de valeurs, alors on peut chercher à savoir quelles sont les valeurs les plus probables et lesquelles sont les moins probables. Cela motive la définition suivante : on appelle *loi de la variable aléatoire* X la donnée des probabilités $\mathbb{P}(X = a)$ pour toutes les valeurs possibles a.

La loi d'une variable aléatoire peut-être donnée par une formule du type $\mathbb{P}(X = a) = f(a)$ si cela est possible, mais on peut aussi, dans des cas simples, la décrire complètement à l'aide d'un tableau de la forme :

a	1	2	3	4	5
$\mathbb{P}(X = a)$	0.1	0.3	0.4	0.15	0.05

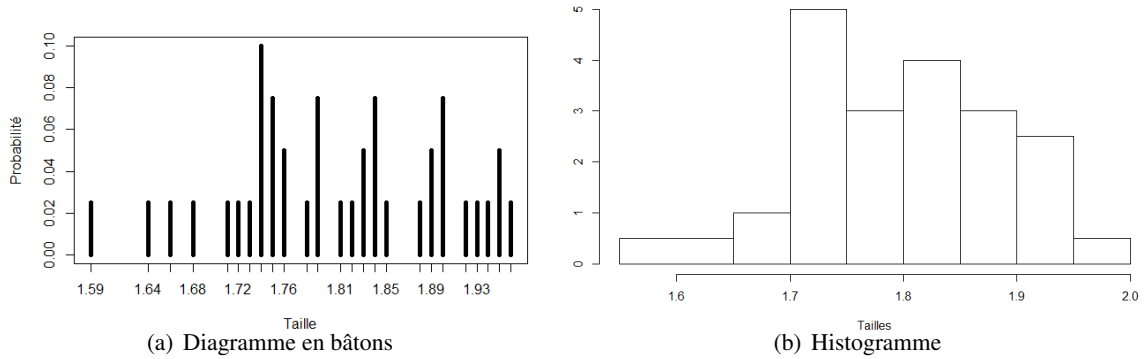
Exemple (Sondage d'un individu dans la chorale)**.** Dans l'exemple de la chorale, si on choisit une personne complètement aléatoirement parmi les 40 chanteurs et qu'on note X_1 sa taille, alors X_1 est une variable aléatoire (sa valeur dépend du tirage). Pour obtenir la loi de X_1, il suffit de remarquer que puisque la personne est choisie complètement aléatoirement, la probabilité que X_1 soit égal à une taille a donnée vaut $\mathbb{P}(X_1 = a) = N_a/40$, où N_a est le nombre d'individus de

taille a dans la chorale ($0 \leqslant N_a \leqslant 40$). En utilisant le diagramme en bâtons de la figure 1.2 du chapitre précédent, on peut calculer tous les nombres N_a et on obtient :

a	1.59	1.64	1.66	1.68	1.71	1.72	1.73	1.74	1.75	1.76	1.78	1.79	1.81
$\mathbb{P}(X_1 = a)$	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.100	0.075	0.050	0.025	0.075	0.025

a	1.82	1.83	1.84	1.85	1.88	1.89	1.9	1.92	1.93	1.94	1.95	1.96
$\mathbb{P}(X_1 = a)$	0.025	0.050	0.075	0.025	0.025	0.050	0.075	0.025	0.025	0.025	0.050	0.025

On peut aussi représenter la loi de X_1 graphiquement comme en figure 2.1 : le diagramme en bâtons est quasiment le même que celui de la figure 1.2, et l'histogramme est identique à celui de la figure 1.4. Cette ressemblance est tout à fait logique puisqu'on tire un individu aléatoirement dans la population de départ !



(a) Diagramme en bâtons (b) Histogramme

FIGURE 2.1 – Loi de la variable aléatoire X_1 dans l'exemple de la chorale. Représentation à l'aide d'un diagramme en bâtons et d'un histogramme.

Exemple (Pile ou face consécutifs)**.** Dans l'activité 1 précédente, si les 100 pièces sont jetées indépendemment et complètement aléatoirement, alors, contrairement à ce que vous auriez peut-être imaginé, il est très probable qu'il y ait au moins 5 pile ou 5 face consécutifs. Plus précisément, si on note Z le nombre maximal de pile ou face consécutifs parmi les 100 lancers, alors Z est une variable aléatoire dont on peut calculer la loi. Nous avons représenté cette loi en figure 2.2. La probabilité $\mathbb{P}(Z = a)$ est maximale en $a = 6$ et on a $\mathbb{P}(Z \geqslant 5) \approx 0.97$.

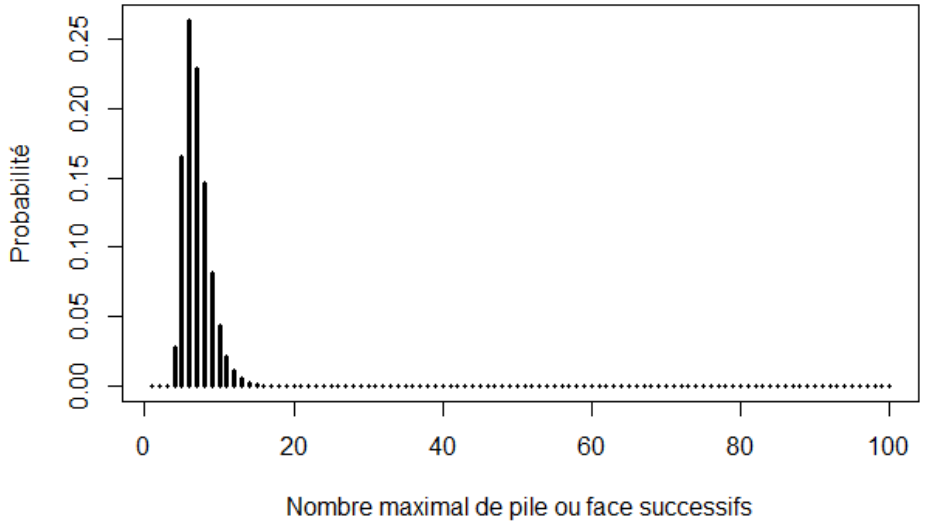


FIGURE 2.2 – Loi du nombre maximal Z de pile ou face consécutifs parmi 100 lancers de pièces.

4.2 Espérance d'une variable aléatoire

Tout comme dans la section précédente, on ne considère que des variables aléatoires X qui ne prennent qu'un nombre fini de valeurs. Dans ce cas, par définition, *l'espérance d'une variable aléatoire* X est la moyenne des valeurs de X pondérées par leurs probabilités d'occurrence. L'espérance est notée $\mathbb{E}(X)$.

Plus formellement, si une variable aléatoire X prend les valeurs $a_1, a_2, \ldots, a_p$, alors l'espérance de X vaut par définition

$$\mathbb{E}(X) = \sum_{i=1}^{p} a_i \, \mathbb{P}(X = a_i) \, .$$

Exemple (Sondage d'un individu dans la chorale)**.** Reprenons l'exemple de la chorale, où la variable aléatoire X_1 désigne la taille d'un individu choisi complètement aléatoirement parmi les 40 chanteurs. Comme vu précédemment, la variable aléatoire X_1 peut prendre 25 valeurs différentes. Son espérance vaut donc :

$$\mathbb{E}(X_1) = 1.59 \times 0.025 + 1.64 \times 0.025 + \ldots + 1.95 \times 0.05 + 1.96 \times 0.025 = 1.807 \text{ m} \, .$$

On remarque que l'espérance de X_1, c'est-à-dire la taille moyenne d'un individu tiré aléatoirement dans la chorale, est égale à la moyenne de population μ_{pop} calculée au chapitre 1. C'est tout à fait logique !

> Aparté mathématique : on peut prouver cette égalité formellement : en notant $a_1, a_2, \ldots, a_{25}$ les 25 tailles différentes dans la chorale, et en notant N_{a_i} le nombre d'individus de taille a_i, on a par définition
>
> $$\mathbb{E}(X_1) = \sum_{i=1}^{25} a_i \, \mathbb{P}(X_1 = a_i) = \sum_{i=1}^{25} a_i \, \frac{N_{a_i}}{40} = \frac{\sum_{i=1}^{25} N_{a_i} \, a_i}{40} = \frac{1}{40} \sum_{i=1}^{40} x_i = \mu_{\text{pop}} \, ,$$
>
> où les quantités $x_1, x_2, x_3, \ldots, x_{40}$ désignent les tailles des 40 chanteurs de la chorale (avec toutes les redondances).

Exemple (Pile ou face consécutifs)**.** Reprenons l'étude de l'activité 1, avec la variable aléatoire Z qui désigne le nombre maximal de pile ou face consécutifs parmi les 100 lancers. Par définition, l'espérance de Z est donnée par

$$\mathbb{E}(Z) = \sum_{a=0}^{100} a \, \mathbb{P}(Z = a) \, .$$

Après calcul, on obtient $\mathbb{E}(Z) \approx 6.98$ pile ou face consécutifs, ce qui est cohérent avec la figure 2.2.

4.3 Variance d'une variable aléatoire

Soit X une variable aléatoire qui ne prend qu'un nombre fini de valeurs $a_1, a_2, \ldots, a_p$. Par définition, *la variance de* X, qu'on note $\text{Var}(X)$, vaut

$$\begin{aligned} \text{Var}(X) &= \sum_{i=1}^{p} \big(a_i - \mathbb{E}(X)\big)^2 \, \mathbb{P}(X = a_i) \\ &= \left(\sum_{i=1}^{p} a_i^2 \, \mathbb{P}(X = a_i) \right) - \big(\mathbb{E}(X)\big)^2 \, . \end{aligned}$$

Quelle formule utiliser pour la variance ?

- La première ligne est facile à interpréter : la variance de X est la moyenne des carrés des écarts des valeurs a_i à l'espérance $\mathbb{E}(X)$, où cette moyenne est calculée en pondérant les valeurs a_i par leurs probabilités d'occurrence. Par conséquent, la variance d'une variable aléatoire X est grande lorsque X fluctue potentiellement beaucoup autour de son espérance $\mathbb{E}(X)$.
- La formule de la deuxième ligne, qui se vérifie à l'aide d'un calcul simple, est également utile car souvent plus pratique pour les applications numériques.

L'*écart-type* $\sigma(X)$ d'une variable aléatoire X est la racine carrée de sa variance : $\sigma(X) = \sqrt{\mathrm{Var}(X)}$.

> Aparté mathématique : il existe une notation pour réécrire les deux formules de la variance de façon plus synthétique. En effet, en remarquant que
>
> $$\mathbb{E}\big(f(X)\big) = \sum_{i=1}^{p} f(a_i)\,\mathbb{P}(X = a_i)$$
>
> pour toute fonction f, la formule de la première ligne se réécrit $\mathbb{E}\big([X - \mathbb{E}(X)]^2\big)$, et celle de la deuxième ligne se réécrit $\mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2$. Il est ainsi fréquent de lire dans les ouvrages de mathématiques :
>
> $$\mathrm{Var}(X) = \mathbb{E}\big([X - \mathbb{E}(X)]^2\big) = \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2 .$$

Exemple (Sondage d'un individu dans la chorale)**.** Reprenons l'exemple de la chorale, où la variable aléatoire X_1 désigne la taille d'un individu choisi complètement aléatoirement parmi les 40 chanteurs. En utilisant la deuxième formule de la variance, on obtient :

$$\begin{aligned}\mathrm{Var}(X_1) &= \big(1.59^2 \times 0.025 + 1.64^2 \times 0.025 + \ldots + 1.95^2 \times 0.05 + 1.96^2 \times 0.025\big) - \big(\mathbb{E}(X_1)\big)^2 \\ &\approx 0.00827\ m^2\ ,\end{aligned}$$

d'où $\sigma(X_1) \approx 0.09$ m. Comme dans le cas de l'espérance, on remarque que l'écart-type de X_1 est égal à l'écart-type de population σ_{pop} calculé au chapitre 1. C'est tout à fait logique puisque X_1 corrrespond à la taille d'un individu tiré aléatoirement dans la population.

Remarque. Attention : lors des applications numériques, il ne faut arrondir que les résultats finaux, en conclusion de question (3 chiffres significatifs par exemple). Ne pas arrondir les résultats de calculs intermédiaires (cela peut entraîner de très grandes erreurs d'arrondis). Ainsi, même si vous avez déjà calculé l'espérance $\mathbb{E}(X_1)$ à la question précédente, vous devez, pour le calcul de la variance, reprendre la valeur de $\mathbb{E}(X_1)$ avec le plus de décimales possibles sur la calculatrice (ne pas utiliser le résultat arrondi de la question précédente).

Exemple (Pile ou face consécutifs)**.** Reprenons l'étude de l'activité 1, avec la variable aléatoire Z qui désigne le nombre maximal de pile ou face consécutifs parmi les 100 lancers. D'après la deuxième formule de la variance, la variance de Z est égale à

$$\mathrm{Var}(Z) = \sum_{a=0}^{100} a^2\,\mathbb{P}(Z = a) - \big(\mathbb{E}(Z)\big)^2 .$$

Après calcul, on obtient $\mathrm{Var}(Z) \approx 3.19$, et donc $\sigma(Z) \approx 1.79$ pile ou face consécutifs. Ce résultat est cohérent avec la figure 2.2.

4.4 Autres caractéristiques d'une variable aléatoire

Tout comme au chapitre 1, il existe de nombreuses autres quantités pour décrire une variable aléatoire : la médiane, les quartiles, l'écart interquartile, etc. Nous ne les définirons pas en détails, mais leurs définitions sont très sembables à celles que nous avons données au chapitre 1 pour décrire la répartition des valeurs dans une population.

5 Exercices

Exercice 1 (Deux exemples d'échantillons)**.** Au cours d'une étude de santé, une équipe de chercheurs d'un hôpital souhaite connaître le rythme cardiaque moyen des 20-50 ans de leur commune. L'un d'entre eux interroge un premier petit groupe de personnes et relève, en nombre de pulsations par minute, les valeurs suivantes :

$$75 \quad 61 \quad 72 \quad 85 \quad 61 \quad 70 \quad 58 \quad 73 \quad 83 \quad 70$$

Parallèlement, un collègue interroge un deuxième groupe de personnes et relève les valeurs suivantes :

$$57 \quad 80 \quad 86 \quad 67 \quad 75 \quad 57 \quad 61 \quad 70 \quad 68 \quad 77$$

1. Décrire l'expérience aléatoire sous-jacente.
2. Calculer les moyennes $\mu_{\text{éch},1}$ et $\mu_{\text{éch},2}$ des échantillons 1 et 2.
3. Calculer les écarts-types $\sigma_{\text{éch},1}$, $\sigma_{\text{éch},2}$ et les écarts-types corrigés $s_{\text{éch},1}$, $s_{\text{éch},2}$ des échantillons 1 et 2. Commenter la différence entre ces deux formes d'écart-types.

Exercice 2 (TP d'endocytose : densité optique moyenne)**.** Lors d'un TP de biologie cellulaire en 2013, des étudiants de L2 Biochimie ont étudié le phénomène d'endocytose chez l'amibe sociale. Le ligand utilisé était le HRP, une enzyme que l'amibe ne connait pas et n'arrive pas à digérer, donc facile à mesurer. Nous allons analyser les résultats des 18 binomes ayant travaillé avec une concentration initiale de HRP de 30 $\mu g/mL$. Voici leurs résultats de mesure pour la densité optique / million de cellules, à l'instant $t = 20$ min :

$$\begin{array}{cccccccccc} -0.0003 & 0.0132 & 0.0083 & 0.0154 & 0.0251 & 0.0371 & 0.0032 & 0.0059 & -0.018 \\ 0.0052 & 0.0885 & 0.0196 & 0.0005 & 0.001 & 0.0238 & 0.0138 & 0.0306 & 0.0139 \end{array}$$

1. Où est l'aléatoire dans cette expérience de biologie ?
2. Calculer la moyenne d'échantillon et l'écart-type d'échantillon corrigé.

Exercice 3 (Deux anniversaires le même jour ?)**.** Pendant un match de foot, deux équipes de 11 joueurs sont sur le terrain.

1. Quelle la probabilité que les 22 joueurs sur le terrain soient tous nés un jour de l'année différent ? (Pour simplifier le problème, on oubliera les années bissextiles, et on supposera que tous les jours de l'année sont équiprobables [1].)
2. En déduire la probabilité qu'au moins deux joueurs sur le terrain soient nés le même jour de l'année. Commentaire ?

1. Si certaines dates sont plus représentées que d'autres, on peut en fait montrer (avec un raisonnement plus sophistiqué) que le phénomène est encore plus marqué, au sens où il est encore plus probable que deux joueurs soient nés le même jour de l'année.

Exercice 4. Carine joue au casino sur une machine à sous avec des pièces de 2 euros. A chaque tour, sa machine délivre un gain (en euros) avec les probabilités suivantes :

gain	0	1	2	10
probabilité	0.8	0.1	0.07	0.03

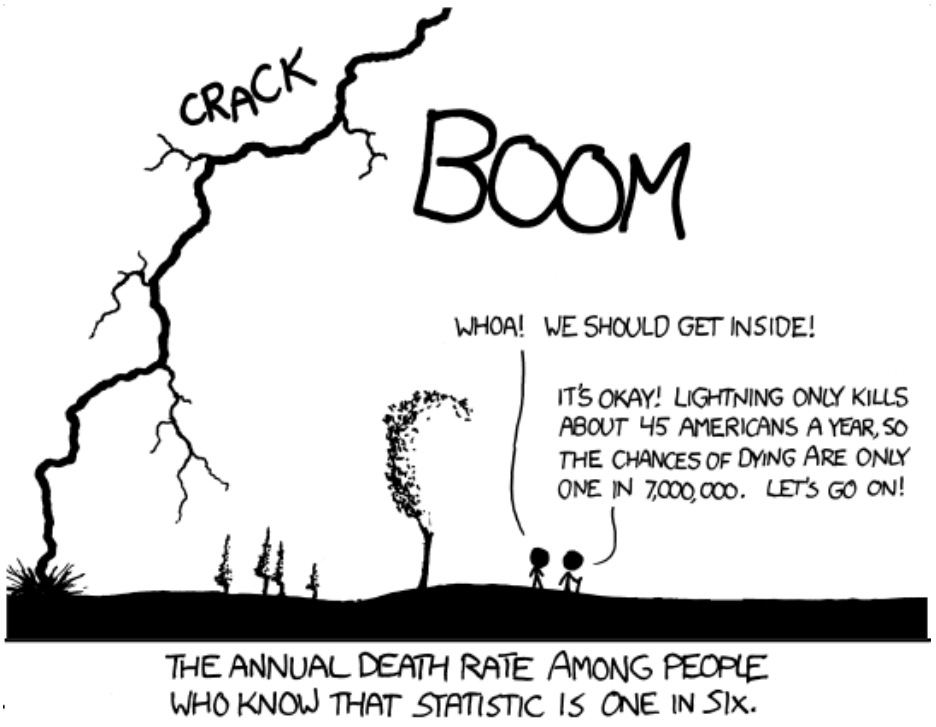
1. Si on appelle X le gain de Carine après un essai, à quoi correspond le tableau précédent en termes mathématiques ?
2. Soit Y le gain net de Carine après un essai. Déterminer la loi de la variable aléatoire Y.
3. Combien Carine peut-elle espérer gagner en moyenne après un essai ? Commenter.

Exercice 5. On lance un dé à quatre faces (tétraèdre régulier dit pyramidal) et un dé à six faces (cubique). Les faces des dés sont numérotées de la manière suivante : le dé pyramidal a trois faces portant le numéro 3 et une face portant le numéro 1, tandis que le dé cubique a trois faces portant le numéro 0, deux faces portant le numéro 2 et une face portant le numéro 4. On appelle X la variable aléatoire égale au numéro de la face cachée du dé pyramidal, et Y celui de la face cachée du dé cubique.

1. Déterminer les lois des variables aléatoires X et Y.
2. Supposons que les lancers des deux dés sont indépendants. Donner la loi de $S = X + Y$.
3. Calculer l'espérance et l'écart-type de la variable aléatoire S.

Exercice 6 (Exercice sur la notion d'indépendance)**.** Bill, Joe et Sam tirent sur une cible au même instant. Pour chacun d'eux, les probabilités d'atteindre la cible sont respectivement $1/2$, $2/3$ et $1/4$. On suppose leurs tirs indépendants. Quelle est la probabilité pour qu'ils atteignent tous les trois la cible ? Qu'aucun des trois n'atteigne la cible ? Que la cible soit atteinte par au moins l'un d'eux ?

Exercice 7. Quel message peut-on tirer de ce dessin humoristique ?



Source : http://xkcd.com/795/ (image légèrement modifiée)

Exercices supplémentaires

Exercice 8 (Exemple de deux événements non indépendants)**.** Camille et Laëtitia sont deux étudiantes qui découvrent le tir à l'arc. Après deux mois, leur entraîneur remarque que Camille atteint sa cible environ trois fois sur quatre, alors que Laëtitia ne l'atteint qu'une fois sur trois. Imaginons l'expérience suivante : l'une de ces deux étudiantes effectue un premier lancer un lundi, puis un deuxième lancer le lendemain (mardi). On ne sait pas qui est l'étudiante à l'origine de ces lancers ; on observe juste les résultats des deux lancers.

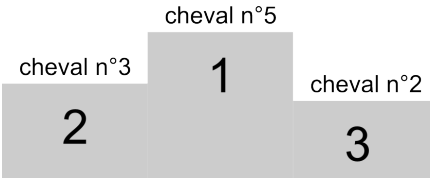
1. Quelle la probabilité que le résultat du premier lancer soit un succès ? que le résultat du second lancer soit un succès ?
2. Les résultats des deux lancers sont-ils indépendants ? (Donner l'intuition puis vérifier votre pressentiment à l'aide d'un calcul.)
3. Donner un exemple biologique où ce type de situation peut se produire.

Exercice 9. Une urne contient 10 boules numérotées : trois boules portent le numéro 1, deux boules portent le numéro 2, cinq boules portent le numéro 3. On tire au hasard deux boules sans remise et on considère la variable aléatoire T représentant le total des nombres marqués sur les deux boules.

1. Déterminer la loi de probabilité de T.
2. Quelle est la probabilité pour que T soit supérieur ou égal à 6 ? pour qu'il soit compris strictement entre 2 et 6 ?
3. Calculer l'espérance de T.

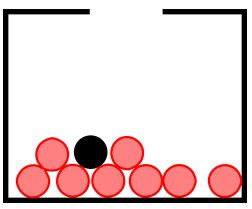
Exercice 10 (partiel 2014)**.** Une course oppose 6 chevaux de force égale. Ces chevaux sont numérotés de 1 à 6 pour mieux les repérer.

1. Quelle est la probabilité que le cheval portant le numéro 1 finisse dans les 3 premiers ?
2. Quelle est la probabilité que le podium soit le suivant ?



Exercice 11. ♢ Un enfant joue au jeu suivant : il place une boule rouge et une boule noire identiques au toucher dans une urne. Après avoir mélangé, il prend au hasard une boule dans l'urne. Si c'est la boule noire, le jeu s'arrête ; si c'est la rouge, il la remet dans l'urne et y ajoute une autre boule rouge. Le jeu recommence alors jusqu'à ce que l'enfant découvre la boule noire.

1. Quelle est la probabilité que l'enfant fasse au moins n tirages ? (pour $n \geqslant 1$)
2. Quelle est la probabilité que l'enfant fasse exactement n tirages ? En déduire la loi de probabilité de la variable aléatoire T correspondant au nombre de tirages.



Exemple de configuration (juste avant le 8ème tirage).

A Résolution de l'exercice 1 avec le logiciel R

On répond aux questions 2 et 3 de l'exercice 1 à l'aide du logiciel R. La fonction mean permet de calculer la moyenne, et la fonction sd permet de calculer l'écart-type corrigé ("sd" est l'abréviation de *standard deviation*). Ainsi, **les logiciels de statistiques donnent plus souvent la valeur de l'écart-type corrigé que celle de l'écart-type**. Pour l'écart-type (tout court), on utilise donc la formule du cours.

```
rythme1=c(75,61,72,85,61,70,58,73,83,70)
rythme2=c(57,80,86,67,75,57,61,70,68,77)

mean(rythme1) # = 70.8
mean(rythme2) # = 69.8

var1 = mean(rythme1^2) - mean(rythme1)^2 # = 73.16 ...
sqrt(var1) # = 8.553362
var2 = mean(rythme2^2) - mean(rythme2)^2 # = 86.16 ...
sqrt(var2) # = 9.282241

sd(rythme1) # = 9.01603
sd(rythme2) # = 9.784341
```

Chapitre 3

Fluctuations d'échantillonnage et intervalles de confiance

Résumé Dans ce chapitre, on étudie l'erreur commise quand on estime la moyenne de population par la moyenne d'échantillon. Cela conduit à construire des intervalles de confiance.

1 Introduction : pourquoi une estimation doit-elle être accompagnée de marges d'erreurs ?

Dans l'exemple de la chorale, supposons qu'on cherche à estimer la moyenne μ_{pop} des tailles des 40 chanteurs à partir d'un échantillon de 5 observations de cette chorale. Si on estime μ_{pop} par la moyenne d'échantillon $\mu_{\text{éch}}$, on ne peut pas raisonablement croire que $\mu_{\text{éch}} = \mu_{\text{pop}}$ exactement ; on fera une petite erreur d'estimation. D'ailleurs, comme l'échantillon est aléatoire, la valeur de $\mu_{\text{éch}}$ que vous auriez obtenue sur un autre échantillon aurait probablement été différente, quoique tout aussi pertinente. Voici quelques exemples d'échantillons ; on remarque effectivement que la valeur de $\mu_{\text{éch}}$ fluctue d'un échantillon à l'autre :

	X_1	X_2	X_3	X_4	X_5	$\mu_{\text{éch}}$
échantillon 1	1.89	1.79	1.74	1.90	1.74	1.812
échantillon 2	1.74	1.95	1.76	1.75	1.71	1.782
échantillon 3	1.84	1.84	1.88	1.85	1.89	1.86
échantillon 4	1.84	1.84	1.75	1.83	1.75	1.802
échantillon 5	1.59	1.68	1.79	1.89	1.79	1.748

Conclusion : pour estimer μ_{pop}, vous ne pouvez pas simplement donner la valeur de $\mu_{\text{éch}}$, mais vous devez l'accompagner de marges d'erreurs. L'objet de ce chapitre est de comprendre comment déterminer ces marges d'erreurs, ou, en termes mathématiques, comment construire un *intervalle de confiance*.

2 Comment quantifier l'erreur associée à l'estimation de la moyenne ?

Formalisons un peu le problème. Dans toute la suite du chapitre, on suppose qu'un statisticien cherche à connaître la moyenne μ_{pop} d'une population de N valeurs $x_1, x_2, \ldots, x_N$. Puisqu'il n'a pas accès à toute la population, le statisticien tire aléatoirement un échantillon de n valeurs

$X_1, X_2, \ldots, X_n$ parmi cette population (on rappelle que les n tirages sont indépendants). Il peut alors calculer la moyenne d'échantillon :

$$\mu_{\text{éch}} = \frac{1}{n} \sum_{i=1}^{n} X_i \ .$$

Une question légitime est alors : quelle est l'erreur commise en proposant $\mu_{\text{éch}}$ comme estimation de μ_{pop} ? Puisque $\mu_{\text{éch}}$ est une variable aléatoire, on peut s'intéresser à son espérance (= sa valeur moyenne) et à son écart-type (= l'ordre de grandeur de l'amplitude de ses fluctuations).

2.1 Espérance de la moyenne d'échantillon

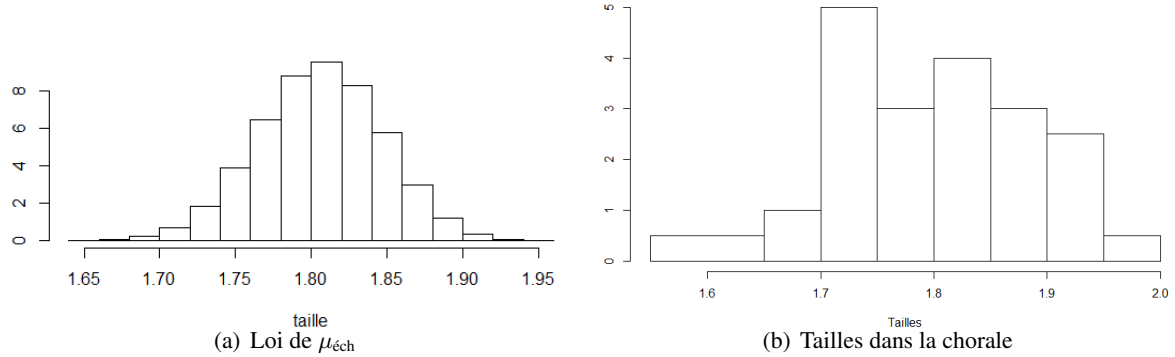
Une première propriété intéressante est que $\mu_{\text{éch}}$ fluctue (d'un échantillon à l'autre) autour de la bonne valeur, c'est-à-dire autour de μ_{pop}. En termes plus mathématiques :

Proposition 1. Soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire de la population initiale (rappel : les n tirages sont indépendants). Alors, la variable aléatoire $\mu_{\text{éch}} = \frac{1}{n} \sum_{i=1}^{n} X_i$ vérifie

$$\mathbb{E}(\mu_{\text{éch}}) = \mu_{\text{pop}} \ .$$

Interprétation : parmi toutes les valeurs possibles que peut prendre la moyenne d'échantillon $\mu_{\text{éch}}$, on obtient en moyenne la bonne valeur μ_{pop}. Ainsi, si on répétait un très grand nombre de fois l'expérience (avec un nouvel échantillon aléatoire à chaque fois), alors la moyenne de toutes les valeurs $\mu_{\text{éch}}$ obtenues serait proche de la vraie valeur μ_{pop}.

Illustration avec l'exemple de la chorale : avec le logiciel R, nous avons tiré aléatoirement et indépendamment 100000 échantillons de 5 observations chacun ($n = 5$), et nous avons relevé à chaque fois la valeur de $\mu_{\text{éch}}$. La répartition de ces valeurs est représentée en figure 3.1(a) ; en termes mathématiques, ce graphique correspond à la loi de la variable aléatoire $\mu_{\text{éch}}$. On remarque que la variable aléatoire $\mu_{\text{éch}}$ fluctue en effet autour de $\mu_{\text{pop}} = 1.807\ m$. Ce graphique ne doit pas être confondu avec la répartition des tailles dans la chorale, qui est représentée en figure 3.1(b).



(a) Loi de $\mu_{\text{éch}}$ (b) Tailles dans la chorale

FIGURE 3.1 – Nous avons représenté : (a) la loi de la variable aléatoire $\mu_{\text{éch}}$ quand $n = 5$; (b) la répartition des tailles dans la chorale.

2.2 Écart-type de la moyenne d'échantillon

Nous avons expliqué au paragraphe précédent que l'estimation $\mu_{\text{éch}}$ fluctue (d'un échantillon à l'autre) autour de la bonne valeur μ_{pop}. Une question importante est maintenant : quelle est l'am-

plitude de ces fluctuations ? Autrement dit, que vaut l'écart-type $\sigma(\mu_{\text{éch}})$ de la variable aléatoire $\mu_{\text{éch}}$?

Proposition 2. Soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire de la population initiale (rappel : les n tirages sont indépendants). Alors, l'écart type de la variable aléatoire $\mu_{\text{éch}} = \frac{1}{n}\sum_{i=1}^{n} X_i$ vérifie

$$\sigma(\mu_{\text{éch}}) = \frac{\sigma_{\text{pop}}}{\sqrt{n}} .$$

(Attention : $\sigma(\mu_{\text{éch}})$ ne doit pas être confondu avec $\sigma_{\text{éch}}$.)

Interprétation : d'un échantillon à l'autre, la moyenne d'échantillon $\mu_{\text{éch}}$ fluctue autour de la bonne valeur μ_{pop} avec une amplitude de l'ordre de $\sigma_{\text{pop}}/\sqrt{n}$. Par conséquent, plus la taille n de l'échantillon est grande, et plus ces fluctuations sont petites, ce qui est logique (un échantillon aléatoire de grande taille donne plus d'informations sur la population). Ainsi, pour que l'estimation $\mu_{\text{éch}}$ soit proche de la valeur inconnue μ_{pop}, il est préférable de disposer d'un échantillon le plus grand possible.

Illustration avec l'exemple de la chorale : sur la figure 3.1, on remarque que les valeurs les plus fréquentes de $\mu_{\text{éch}}$ sont situées entre $1.74\ m$ et $1.88\ m$ environ. Ces valeurs sont donc moins dispersées que les tailles de la population. Ce phénomène est logique car, d'après la proposition 2 ci-dessus, l'écart-type de la variable aléatoire $\mu_{\text{éch}}$ vaut $\sigma_{\text{pop}}/\sqrt{5}$, donc il est environ deux fois plus petit que l'écart-type initial σ_{pop}.

Remarque sur la racine carrée : si on veut augmenter la précision de l'estimation d'un chiffre significatif, il faut prendre un échantillon 100 fois plus grand ; en effet, l'amplitude des fluctuations $\sigma_{\text{pop}}/\sqrt{n}$ est alors $\sqrt{100} = 10$ fois plus petite, d'où un gain d'un chiffre significatif.

3 Contruction d'un intervalle de confiance

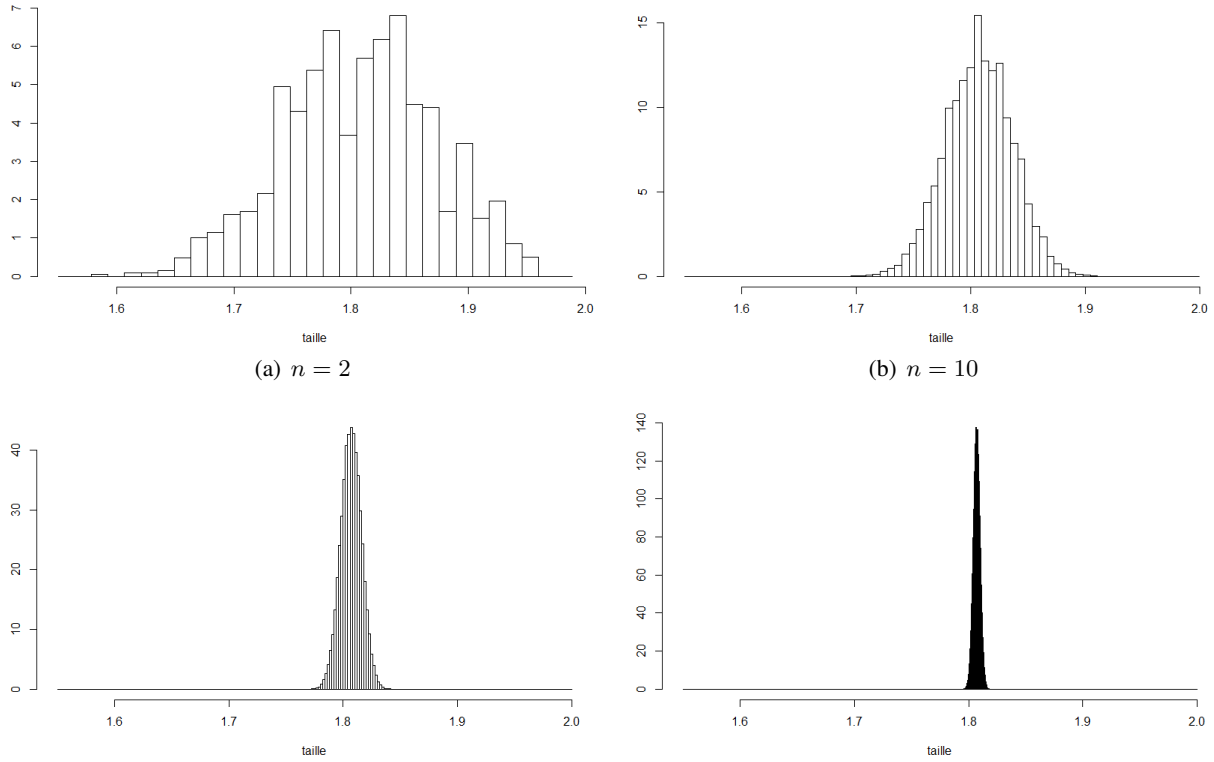
Dans cette section, on explique comment construire des intervalles de confiance. L'idée principale est une simple inversion de rôles : pour connaître l'écart entre la valeur inconnue μ_{pop} et la valeur calculée $\mu_{\text{éch}}$, il suffit de connaître l'écart entre $\mu_{\text{éch}}$ et μ_{pop}. Nous allons donc d'abord construire un intervalle de fluctuation pour $\mu_{\text{éch}}$, puis en déduire un intervalle de confiance pour μ_{pop}.

On reprend les mêmes notations que dans la section précédente : un statisticien dispose d'un échantillon de n valeurs $X_1, X_2, \ldots, X_n$ (tirées aléatoirement et indépendemment les unes des autres) ; il peut alors calculer la moyenne d'échantillon $\mu_{\text{éch}} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

3.1 Intervalle de fluctuation pour la moyenne d'échantillon

Nous avons expliqué dans la section 2 que la moyenne d'échantillon $\mu_{\text{éch}}$ fluctue (d'un échantillon à l'autre) autour de la vraie valeur μ_{pop} avec un écart-type égal à $\sigma_{\text{pop}}/\sqrt{n}$. On peut être encore plus précis en remarquant que, quand n augmente, l'histogramme de la variable aléatoire $\mu_{\text{éch}}$ ressemble de plus en plus à une courbe de Gauss, de plus en plus concentrée autour de μ_{pop}. C'est ce qu'on observe par exemple en figure 3.2, où on a représenté la loi de $\mu_{\text{éch}}$ pour différentes tailles n d'échantillon. [1]

1. Pour que les cas $n = 100$ et $n = 1000$ aient un intérêt, il ne faut plus penser à l'exemple de la chorale (où il y a seulement 40 individus), mais à un exemple beaucoup plus important comme la population d'un pays. On regarde alors la répartition des valeurs possibles de $\mu_{\text{éch}}$ en fonction du nombre n d'observations collectées, avec $n \in \{2, 10, 100, 1000\}$.



(a) $n = 2$ (b) $n = 10$

(c) $n = 100$ (d) $n = 1000$

FIGURE 3.2 – Loi de la variable aléatoire $\mu_{\text{éch}}$ pour différentes tailles n d'échantillon.

3.1.1 Le Théorème de la Limite Centrale

Le fait que la loi de $\mu_{\text{éch}}$ ressemble à une courbe de Gauss de plus en plus resserrée quand n augmente est un phénomène universel : il se produit quelle que soit la répartition des valeurs dans la population initiale. C'est ce qu'énonce un théorème fondamental en statistique : le *Théorème de la Limite Centrale*.

Théorème 1 (Théorème de la Limite Centrale)**.**
Soit $X_1, X_2, \ldots, X_n$ un échantillon de n valeurs tirées aléatoirement et indépendemment dans une population de moyenne μ_{pop} et d'écart-type σ_{pop}. Alors, lorsque $n \to +\infty$, la moyenne d'échantillon $\mu_{\text{éch}} = \frac{1}{n}\sum_{i=1}^{n} X_i$ vérifie :

$$\mathbb{P}\left(a \leqslant \frac{\mu_{\text{éch}} - \mu_{\text{pop}}}{\sigma_{\text{pop}}/\sqrt{n}} \leqslant b\right) \underset{n\to+\infty}{\longrightarrow} \mathbb{P}(a \leqslant Z \leqslant b)\,, \tag{3.1}$$

où Z est une variable aléatoire dont les valeurs se répartissent selon la courbe de Gauss $\mathcal{N}(0,1)$ d'espérance 0 et de variance 1. Cela signifie que la probabilité $\mathbb{P}(a \leqslant Z \leqslant b)$ est égale à l'aire sous la courbe de Gauss $\mathcal{N}(0,1)$ entre les abscisses a et b, comme indiqué sur la figure 3.3. Cette aire peut s'exprimer mathématiquement à l'aide la formule suivante :

$$\mathbb{P}(a \leqslant Z \leqslant b) = \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right) dx\,.$$

En pratique, la valeur de cette intégrale est calculée de façon approchée et est disponible dans un tableau (en fonction de a et b). Tous les logiciels de calcul scientifique peuvent aussi fournir cette valeur.

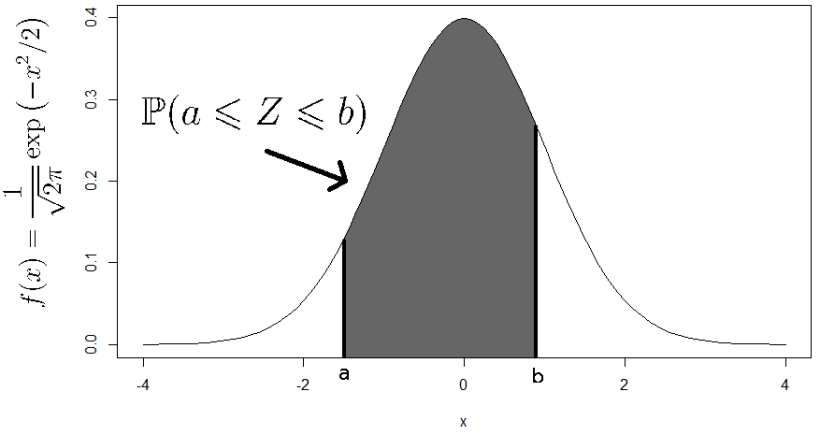


FIGURE 3.3 – La courbe de Gauss $\mathcal{N}(0,1)$. La probabilité $\mathbb{P}(a \leqslant Z \leqslant b)$ est égale à l'aire sous la courbe entre les abscisses a et b.

Remarque. Le théorème signifie que, lorsque n est grand, la loi de la variable aléatoire $\dfrac{\mu_{\text{éch}} - \mu_{\text{pop}}}{\sigma_{\text{pop}}/\sqrt{n}}$ est approximativement gaussienne $\mathcal{N}(0,1)$. En multipliant par $\sigma_{\text{pop}}/\sqrt{n}$ et en ajoutant μ_{pop}, on en déduit que la loi de la variable aléatoire $\mu_{\text{éch}}$ est approximativement gaussienne d'espérance μ_{pop} et d'écart-type $\sigma_{\text{pop}}/\sqrt{n}$, ce qui est cohérent avec la figure 3.2.

Remarque. Le symbole $\mathcal{N}$ est une abréviation de l'adjectif *normale*, qui est un synonyme de *gaussienne*. Dans les livres, on trouve souvent l'expression : "Z suit la loi normale $\mathcal{N}(0,1)$".

3.1.2 Application du TLC : intervalle de fluctuation

Le Théorème de la Limite Centrale permet d'obtenir des intervalles de fluctuation pour la variable aléatoire $\mu_{\text{éch}}$. Plus précisément, d'après la ligne (3.1) du théorème 1, on a, lorsque n est grand :

$$\mathbb{P}\left(-c \leqslant \frac{\mu_{\text{éch}} - \mu_{\text{pop}}}{\sigma_{\text{pop}}/\sqrt{n}} \leqslant c\right) \approx \mathbb{P}(-c \leqslant Z \leqslant c)\ .$$

On choisit le réel c tel que $\mathbb{P}(-c \leqslant Z \leqslant c)$ soit proche de 100%. Par exemple, pour une probabilité de 95%, une lecture sur un tableau nous donne $c \approx 1.96$, alors que pour une probabilité de 99%, on lit $c \approx 2.576$. Réécrivons maintenant le membre de gauche :

$$\begin{aligned} -c \leqslant \frac{\mu_{\text{éch}} - \mu_{\text{pop}}}{\sigma_{\text{pop}}/\sqrt{n}} \leqslant c &\iff -c\,\frac{\sigma_{\text{pop}}}{\sqrt{n}} \leqslant \mu_{\text{éch}} - \mu_{\text{pop}} \leqslant c\,\frac{\sigma_{\text{pop}}}{\sqrt{n}} \\ &\iff \mu_{\text{pop}} - c\,\frac{\sigma_{\text{pop}}}{\sqrt{n}} \leqslant \mu_{\text{éch}} \leqslant \mu_{\text{pop}} + c\,\frac{\sigma_{\text{pop}}}{\sqrt{n}}\ . \end{aligned}$$

Par conséquent, pour le choix de $c = 1.96$ (et donc $\mathbb{P}(-c \leqslant Z \leqslant c) \approx 0.95$), on a, lorsque n est grand :

$$\mathbb{P}\left(\mu_{\text{pop}} - 1.96\,\frac{\sigma_{\text{pop}}}{\sqrt{n}} \leqslant \mu_{\text{éch}} \leqslant \mu_{\text{pop}} + 1.96\,\frac{\sigma_{\text{pop}}}{\sqrt{n}}\right) \approx 0.95\ . \tag{3.2}$$

Nous avons donc démontré le résultat suivant :

Proposition 3 (Intervalle de fluctuation)**.** Si le nombre n d'observations est grand, alors, pour environ 95% des échantillons de taille n, la moyenne d'échantillon $\mu_{\text{éch}}$ tombe dans l'intervalle

$$\left[\mu_{\text{pop}} - 1.96\,\frac{\sigma_{\text{pop}}}{\sqrt{n}}\,;\ \mu_{\text{pop}} + 1.96\,\frac{\sigma_{\text{pop}}}{\sqrt{n}}\right] .$$

Cet intervalle s'appelle *l'intervalle de fluctuation de $\mu_{éch}$ au niveau de confiance* 95%.

3.2 Intervalle de confiance pour la moyenne d'une population

Nous savons maintenant, lorsque n est grand, comment construire un intervalle de fluctuation pour la moyenne d'échantillon $\mu_{\text{éch}}$. Inversons les rôles de $\mu_{\text{éch}}$ et μ_{pop} dans les deux inégalités apparaissant dans le membre de gauche de (3.2) :

$$\begin{aligned}
\text{d'une part,} \qquad & \mu_{\text{pop}} - 1.96\,\frac{\sigma_{\text{pop}}}{\sqrt{n}} \leqslant \mu_{\text{éch}} \iff \mu_{\text{pop}} \leqslant \mu_{\text{éch}} + 1.96\,\frac{\sigma_{\text{pop}}}{\sqrt{n}} \\
\text{d'autre part,} \qquad & \mu_{\text{éch}} \leqslant \mu_{\text{pop}} + 1.96\,\frac{\sigma_{\text{pop}}}{\sqrt{n}} \iff \mu_{\text{éch}} - 1.96\,\frac{\sigma_{\text{pop}}}{\sqrt{n}} \leqslant \mu_{\text{pop}}
\end{aligned}$$

En combinant les deux nouvelles inégalités obtenues, on peut donc réécrire (3.2) de la façon suivante : lorsque n est grand,

$$\mathbb{P}\left(\mu_{\text{éch}} - 1.96\,\frac{\sigma_{\text{pop}}}{\sqrt{n}} \leqslant \mu_{\text{pop}} \leqslant \mu_{\text{éch}} + 1.96\,\frac{\sigma_{\text{pop}}}{\sqrt{n}}\right) \approx 0.95\ .$$

Cela signifie que dans environ 95% des cas, la valeur de μ_{pop} (que l'on cherche à estimer) est dans l'intervalle :

$$\left[\mu_{\text{éch}} - 1.96\,\frac{\sigma_{\text{pop}}}{\sqrt{n}}\,;\ \mu_{\text{éch}} + 1.96\,\frac{\sigma_{\text{pop}}}{\sqrt{n}}\right] .$$

En remplaçant σ_{pop} (qu'on ne connait généralement pas) par son estimation $\sigma_{\text{éch}}$, on obtient finalement :

Proposition 4 (Intervalle de confiance)**.** Si le nombre n d'observations est grand, alors, pour environ 95% des échantillons de taille n, la valeur de μ_{pop} (que l'on cherche à estimer) est dans l'intervalle :

$$\left[\mu_{\text{éch}} - 1.96\,\frac{\sigma_{\text{éch}}}{\sqrt{n}}\,;\ \mu_{\text{éch}} + 1.96\,\frac{\sigma_{\text{éch}}}{\sqrt{n}}\right] .$$

On dit que cet intervalle est un *intervalle de confiance pour μ_{pop} au niveau de confiance* 95%. Il peut être calculé entièrement à l'aide des données disponibles car le statisticien connait la valeur de $\mu_{\text{éch}}$ et de $\sigma_{\text{éch}}$.

Remarque (Influence du nombre n d'observations)**.**
La largeur de l'intervalle de confiance est proportionnelle à $1/\sqrt{n}$, donc plus le nombre n d'observations est grand, et plus notre intervalle de confiance est précis. C'est logique !

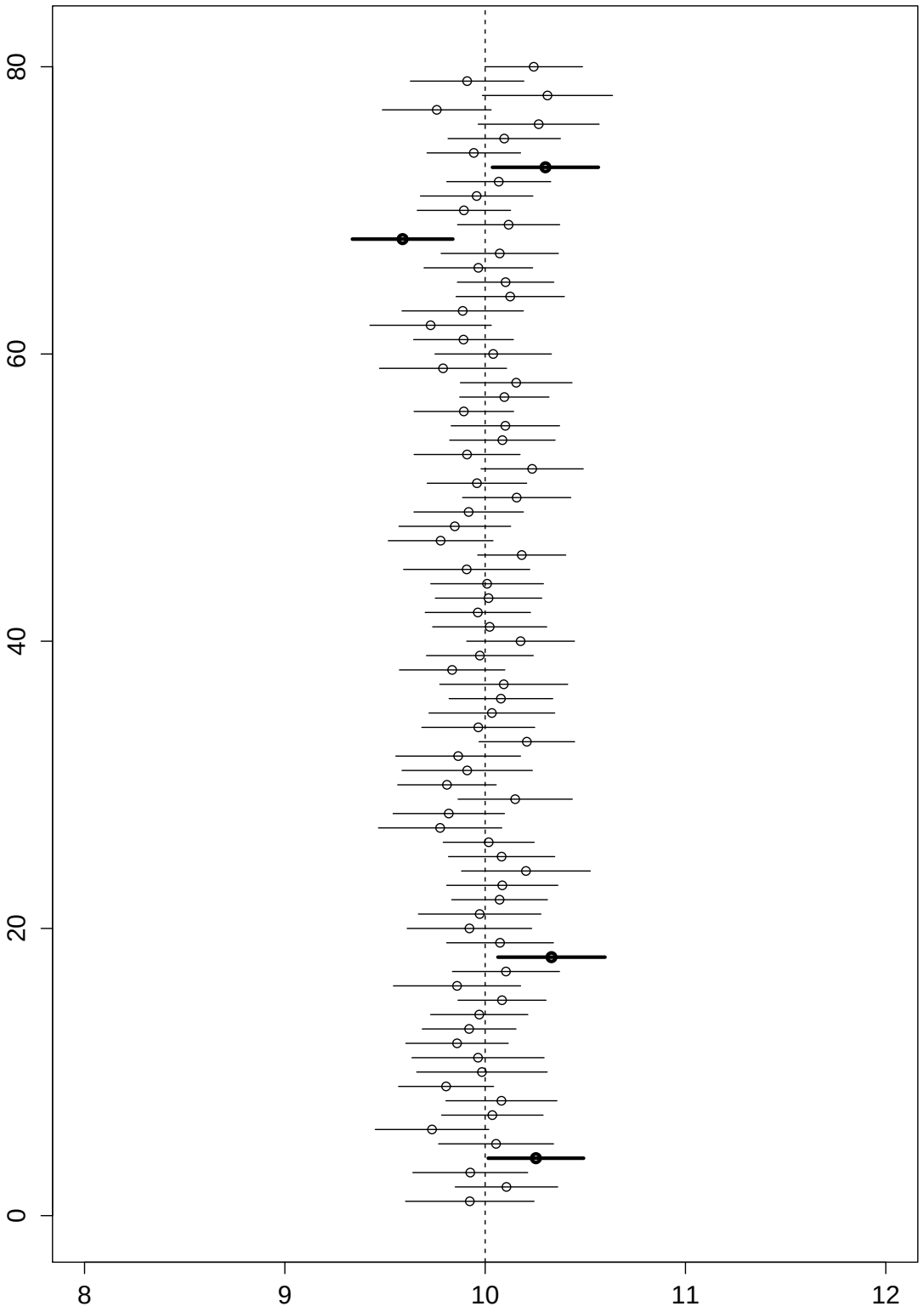


FIGURE 3.4 – Visualisation des intervalles de confiance à 95% pour μ_{pop} obtenus avec 80 échantillons (aléatoires) de taille $n = 50$. La vraie valeur de μ_{pop} est $\mu_{\text{pop}} = 10$. Sur les 80 intervalles de confiance obtenus, 4 ne contiennent pas la valeur de μ_{pop} (en gras). Ainsi, une proportion de $4/80 = 5\%$ des échantillons ont donné un mauvais intervalle de confiance. On s'attendait à une proportion proche de 5% car le niveau de confiance des intervalles est de 95%.

Remarque (Attention à l'interprétation d'un intervalle de confiance !)**.**

1. L'intervalle de confiance est un intervalle aléatoire (à cause de la présence de $\mu_{\text{éch}}$ et de $\sigma_{\text{éch}}$). Cet intervalle aléatoire contient μ_{pop} pour environ 95% des échantillons, mais il ne contient pas μ_{pop} pour les autres d'échantillons (les 5% restants). Ce phénomène est bien illustré sur la figure 3.4.
2. Pour les applications numériques, on disposera de valeurs concrètes pour $\mu_{\text{éch}}$ et $\sigma_{\text{éch}}$. Par exemple, si après $n = 50$ mesures d'une solution dont on cherche à déterminer la concentration en une certaine molécule, on obtient $\mu_{\text{éch}} = 25.2\,\mu g/mL$ et $\sigma_{\text{éch}} = 0.1\,\mu g/mL$, alors on peut calculer explicitement les extrémités de l'intervalle de confiance au niveau 95% pour la concentration inconnue :

$$[25.17\,\mu g/mL\ ;\ 25.23\,\mu g/mL]$$

Cet intervalle n'est plus un intervalle aléatoire (les valeurs 25.17 et 25.23 sont des valeurs précises !). Par conséquent, une phrase du type "la vraie concentration μ_{pop} est comprise entre $25.17\,\mu g/mL$ et $25.23\,\mu g/mL$ avec probabilité 95%" n'aurait pas de sens. En effet, ou bien μ_{pop} est dans cet intervalle, ou bien elle ne l'est pas (il n'y a plus rien d'aléatoire).
$\rightsquigarrow$ L'interprétation des 95% de confiance est en revanche la suivante : parmi tous les échantillons de 50 mesures que j'aurais pu observer, 95% d'entre eux m'auraient donné un intervalle de confiance valide (= qui contient la valeur de μ_{pop}). Dans mon cas précis, j'espère que mon échantillon fait partie de ces 95% ; j'ai seulement pris un risque de 5% de me tromper. Sur la figure 3.4, cela correspond à espérer avoir tiré un échantillon qui ne donne pas l'un des 4 mauvais intervalles de confiance parmi les 80 représentés. On n'a aucun moyen de vérifier si notre intervalle de confiance est correct ; on espère juste ne pas voir avoir eu de malchance.

3.3 Que faire lorsque le nombre n d'observations est petit ?

Les résultats précédents ont été démontrés dans le cas où le nombre n d'observations est grand. Mais que signifie "n est grand" précisément ? Est-ce que $n = 10$ convient ? $n = 50$? $n = 200$? $n = 1000$?

On répondra précisément à cette question dans la section 3.4 pour l'estimation d'une proportion. Dans le cas général, on ne peut malheureusement pas apporter de réponse générale, car la réponse dépend de certains paramètres du problème. [2] Disons néanmoins que $n = 10$ est souvent insuffisant pour utiliser l'intervalle de confiance de la proposition 4, mais que $n = 1000$ sera à l'inverse souvent suffisant. Entre les deux, cela dépend du problème... *Une remarque à propos de la valeur de n est malgré tout indispensable dans tous les exercices.*

Quand n est trop petit, il existe d'autres techniques qui ne reposent pas sur le Théorème de la Limite Centrale, mais sur d'autres théorèmes mathématiques. On verra ce cas plus tard dans le chapitre 4.

3.4 Cas particulier : intervalle de confiance pour une proportion

Dans cette section, on souhaite estimer la proportion p_{pop} des individus d'une population possédant une certaine propriété (ex : la proportion de Français qui voteront pour tel candidat, la proportion de rats qui développent une tumeur, la proportion de machines défectueuses, etc).

Comment estimer p_{pop} ? Comme d'habitude, on se sert des observations disponibles : à partir d'un échantillon de n individus, on calcule la proportion $p_{\text{éch}}$ d'individus dans l'échantillon qui

2. Ce phénomène est décrit par le théorème de Berry-Essen, qui n'est pas au programme de ce cours.

possèdent la propriété en question. La quantité $p_{\text{éch}}$, qu'on appelle la *proportion d'échantillon*, est une estimation de la vraie proportion p_{pop}.

Exemple. On souhaite connaître la proportion de femelles dans une population de lapins. Plusieurs chercheurs ont observé $n = 200$ lapins et ont relevé les résultats suivants (1 pour femelle, 0 pour mâle) :

1	0	1	0	1	0	0	1	1	0	0	1	1	1	1	1	0	0	1	1	1	1	0	1	1	1	0	1	0	1
1	0	1	1	0	1	0	1	1	0	1	1	1	1	1	0	1	0	0	1	0	1	0	0	1	0	0	1	0	0
1	0	1	0	0	0	1	0	0	1	1	1	1	0	1	1	1	0	0	1	0	1	0	0	0	0	0	1	1	1
1	0	1	1	0	0	1	0	0	0	1	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	0	1	0
1	1	0	0	1	1	1	1	1	1	0	1	0	1	1	0	0	0	0	0	1	0	1	1	1	1	0	0	1	1
1	1	0	1	1	1	0	1	1	0	1	1	0	1	1	0	1	1	1	0	1	0	1	1	1	1	0	1	0	1
1	1	1	0	1	0	0	1	0	0	0	0	1	0	1	1	0	1	0	0										

Dans cette exemple, on compte 117 femelles parmi les 200 lapins observés, d'où une proportion d'échantillon $p_{\text{éch}} = 117/200 = 0.585$.

Comme dans le cas de l'estimation d'une moyenne de population, donner une estimation $p_{\text{éch}}$ ne suffit pas. Il faut aussi donner des marges d'erreur, autrement dit, construire un *intervalle de confiance*.

Fort heureusement, tout ce qu'on a fait précédemment pour l'estimation d'une moyenne de population peut être réutilisé dans le cas d'une proportion. En effet, dans l'exemple ci-dessus, la proportion d'échantillon $p_{\text{éch}}$ est égale au nombre de femelles observées divisé par le nombre total de lapins observés, c'est-à-dire :

$$p_{\text{éch}} = \frac{\sum_{i=1}^{n} X_i}{n} ,$$

où $X_i = 1$ si le i-ème lapin observé est une femelle, et $X_i = 0$ si c'est un mâle. De même, la vraie proportion cherchée p_{pop} est égale au nombre total de femelles dans la population divisé par le nombre total N de lapins, d'où

$$p_{\text{pop}} = \frac{\sum_{i=1}^{N} x_i}{N} ,$$

où $x_i = 1$ si le i-ème lapin de la population est une femelle, et $x_i = 0$ si c'est un mâle. Par conséquent, les proportions de population et d'échantillon sont aussi des moyennes ! On peut donc utiliser directement[3] la proposition 4 en remplaçant $\mu_{\text{éch}}$ par $p_{\text{éch}}$ et $\sigma_{\text{éch}}$ par $\sqrt{p_{\text{éch}}(1 - p_{\text{éch}})}$, et on obtient l'intervalle de confiance suivant :

Proposition 5 (Intervalle de confiance pour une proportion)**.** Si les trois conditions

$$n \geqslant 30 \qquad n\, p_{\text{éch}} \geqslant 5 \qquad \text{et} \qquad n\,(1 - p_{\text{éch}}) \geqslant 5$$

sont toutes satisfaites, alors, pour environ 95% des échantillons de taille n, la valeur de p_{pop} (que l'on cherche à estimer) est dans l'intervalle :

$$\left[p_{\text{éch}} - 1.96\, \frac{\sqrt{p_{\text{éch}}(1 - p_{\text{éch}})}}{\sqrt{n}} \,;\; p_{\text{éch}} + 1.96\, \frac{\sqrt{p_{\text{éch}}(1 - p_{\text{éch}})}}{\sqrt{n}} \right] .$$

On dit que cet intervalle est un *intervalle de confiance pour p_{pop} au niveau de confiance* 95%. Il peut être calculé entièrement à l'aide des données disponibles car le statisticien connait la valeur de $p_{\text{éch}}$.

3. Cf. exercice 11 pour l'explication de l'égalité $\sigma_{\text{éch}} = \sqrt{p_{\text{éch}}(1 - p_{\text{éch}})}$.

Remarque. Dans le cas d'une proportion, la condition "lorsque le nombre n d'observations est grand" est remplaçée par un jeu de trois conditions plus précises, à vérifier absolument dans les exercices :

$$n \geqslant 30 \qquad n\,p_{\text{éch}} \geqslant 5 \qquad \text{et} \qquad n\,(1-p_{\text{éch}}) \geqslant 5\,.$$

Exemple. Reprenons l'exemple précédent, où l'on cherche à estimer la proportion p_{pop} de femelles dans une population de lapins. Puisque $n = 200$ et $p_{\text{éch}} = 0.585$, on peut vérifier les trois conditions :

$$\begin{cases} n = 200 \geqslant 30 \\ n\,p_{\text{éch}} = 200 \times 0.585 = 117 \geqslant 5 \\ n\,(1-p_{\text{éch}}) = 200 \times (1-0.585) = 83 \geqslant 5 \end{cases}$$

On peut donc en déduire un intervalle de confiance pour p_{pop} au niveau de confiance 95% :

$$\left[0.585 - 1.96\,\frac{\sqrt{0.585(1-0.585)}}{\sqrt{200}}\,;\; 0.585 + 1.96\,\frac{\sqrt{0.585(1-0.585)}}{\sqrt{200}}\right] \approx \big[0.517\,;\,0.653\big]\,.$$

Ainsi, notre intervalle de confiance au niveau 95% est approximativement $\big[51.7\%\,;\,65.3\%\big]$. Il y a donc vraisemblablement plus de femelles que de mâles dans la population. Le niveau de confiance 95% signifie que notre procédure est valide pour environ 95% des échantillons, donc nous avons seulement pris un risque de 5% de nous tromper.

4 Exercices

Exercice 1. Nous allons considérer des variables aléatoires dont la loi est donnée par :

a	1	2	3
$\mathbb{P}(X_\star = a)$	$1/3$	$1/3$	$1/3$

1. Soit X_1 et X_2 deux variables aléatoires indépendantes et dont la loi est donnée par le tableau ci-dessus (on peut penser à deux tirages indépendants dans la même population). Déterminer la loi de la variable aléatoire $M_2 = (X_1 + X_2)/2$, puis représentez-la graphiquement.
2. Même question avec trois variables aléatoires X_1, X_2 et X_3 indépendantes et de loi donnée ci-dessus. Quelle est la loi de la variable aléatoire $M_3 = (X_1+X_2+X_3)/3$? Représentation graphique ?
3. Calculer et comparer les espérances et les écarts-types de X_1, M_2 et M_3. Commentaires ?

Exercice 2. Un fabricant de boîtes de conserve veut contrôler le poids des boîtes qu'il a produites. On note μ_{pop} le poids moyen de toutes les boîtes de sa production, et σ_{pop} l'écart-type correspondant.

1. Un échantillon de 10 boîtes donne les poids suivants (en grammes) :

   $$490 \quad 492 \quad 497 \quad 502 \quad 505 \quad 490 \quad 495 \quad 492 \quad 490 \quad 497$$

   Proposer des estimations pour μ_{pop} et σ_{pop}.
2. Le fabricant décide de tester davavantage de boîtes (pour limiter les incertitudes d'estimation). Il contrôle donc 33 boîtes supplémentaires, ce qui donne pour les 43 boîtes totales :

poids	490	492	495	497	502	505
effectif	6	11	10	8	5	3

(a) Proposer de nouvelles estimations pour μ_{pop} et σ_{pop}. A votre avis, ces estimations sont-elles meilleures que les précédentes ? Pourquoi ?

(b) Donner un intervalle de confiance pour μ_{pop} au niveau de confiance 95%.

(c) Même question avec un niveau de confiance de 99%. Qu'est-ce qui change ? pourquoi ?

Exercice 3 (TP endocytose)**.** On reprend l'exercice 2 du chapitre 2, correspondant au TP sur le phénomène d'endocytose chez l'amibe sociale. Rappelons les résultats de mesure (densité optique / million de cellules à l'instant $t = 20$ min) des 18 binomes ayant travaillé avec une concentration initiale de HRP de $30\,\mu g/mL$:

-0.0003	0.0132	0.0083	0.0154	0.0251	0.0371	0.0032	0.0059	-0.018
0.0052	0.0885	0.0196	0.0005	0.001	0.0238	0.0138	0.0306	0.0139

1. Où est l'aléatoire dans cette expérience de biologie ?
2. Donner un intervalle de confiance à 95% pour la densité optique / million de cellules à l'instant $t = 20$ min qu'on observerait en moyenne sur toute la population des amibes sociales, dans des conditions expérimentales parfaites, et sans bruit de mesure.

Exercice 4. Sur 12000 individus d'une espèce de lapins, on a dénombré 13 albinos.

1. Estimer la proportion p_{pop} d'albinos dans l'espèce (on donnera un intervalle de confiance pour p_{pop} au niveau de confiance 95%).
2. Même question avec un niveau de confiance de 99%. Quelle la différence avec la réponse à la question précédente ?

Exercice 5 (Etes-vous un-e citoyen-ne éclairé-e ?)**.** Imaginons une situation politique en 2017 : François Hollande se représente aux élections présidentielles, et nous sommes le jour du second tour, en fin d'après-midi. Peu de bulletins ont été dépouillés, mais l'institut Ifop a réalisé un sondage à la sortie des urnes afin d'estimer la proportion des électeurs qui ont voté pour François Hollande.

1. Après avoir interrogé 1225 personnes, l'institut de sondage a compté 637 électeurs ayant voté pour François Hollande. Donner un intervalle de confiance à 95% pour la proportion des électeurs français qui ont voté pour le président sortant.
2. Vous êtes journaliste sur une chaîne de télévision, et vous devez parler des élections. Quel pronostic annoncerez-vous à l'antenne ?

Exercice 6. Quel message peut-on tirer de ce dessin humoristique ?

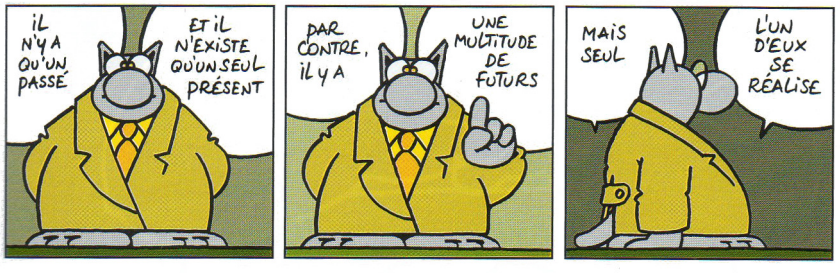


Image extraite de l'ouvrage *La mathématique du chat de Philippe Geluck*, Daniel Justens.

Exercices supplémentaires

Exercice 7. Gaëlle s'apprête à jouer à un jeu de hasard avec un ami : elle dispose 3 cartons numérotés de 1 à 3 sur une table ; les cartons sont retournés pour ne pas que leurs numéros soient visibles.

1. Gaëlle demande à son ami Loïc de choisir un carton, puis de relever son numéro X. Quelle est la loi de X ? Que valent l'espérance et l'écart-type de X ?
2. On considère un protocole un peu plus abouti : Gaëlle demande à Loïc de choisir un premier carton, de relever son numéro X_1, puis de le reposer face cachée. Gaëlle mélange le tas de trois cartons et demande à Loïc de choisir un deuxième carton, puis de noter son numéro X_2. On désigne par S la somme des 2 numéros relevés par Loïc. Quelle est la loi de S ? Déterminer également l'espérance et l'écart-type de S.
3. Protocole encore plus abouti : Loïc ne s'arrête pas au 2-ème carton mais continue jusqu'au 9-ème carton (à chaque fois, Loïc repose son carton face cachée et Gaëlle mélange), et Loïc relève les 9 numéros $X_1, X_2, \dots, X_9$ obtenus. Si M désigne la moyenne des 9 résultats obtenus, que valent l'espérance et l'écart-type de M ?
4. Jeu ultime : Loïc doit relever les numéros de 50 cartons (Gaëlle mélange à chaque fois). Soit M la moyenne des 50 résultats obtenus. Si on traçait l'histogramme des valeurs de M, à quoi ressemblerait-il ? pourquoi ?

Les exercices suivants sont d'un niveau mathématique un peu plus élevé. N'hésitez pas à les chercher pour vous entraîner davantage, mais aussi pour mieux comprendre les théorèmes utilisés dans le cours. Les résultats les plus intéressants sont les exercices 10 et 11.

Exercice 8. $\diamondsuit$ Soit X une variable aléatoire prenant un nombre fini de valeurs $a_1, a_2, \dots, a_p \in \mathbb{R}$. Montrer que pour toute fonction $f : \mathbb{R} \to \mathbb{R}$, on a :

$$\mathbb{E}\big(f(X)\big) = \sum_{i=1}^{p} f(a_i)\mathbb{P}(X = a_i) \, .$$

Exercice 9. $\diamondsuit$ Démontrer les propositions 6 et 7 ci-dessous, qui sont des relations mathématiques utiles sur l'espérance et la variance. On pourra supposer que la variable aléatoire X prend un nombre fini de valeurs $a_1, a_2, \dots, a_p$ et que Y prend un nombre fini de valeurs $b_1, b_2, \dots, b_q$.

Proposition 6 (Relations sur l'espérance)**.** Soit X et Y deux variables aléatoires et $\alpha \in \mathbb{R}$. Alors les deux relations suivantes sont toujours vraies :

$$\mathbb{E}(\alpha\, X) = \alpha\, \mathbb{E}(X)$$
$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

Proposition 7 (Relations sur la variance)**.** Soit X et Y deux variables aléatoires et $\alpha \in \mathbb{R}$. La relation suivante est toujours vraie :

$$\mathrm{Var}(\alpha\, X) = \alpha^2\, \mathrm{Var}(X) \, .$$

Sous une condition supplémentaire, on a aussi :

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \qquad \text{si } X \text{ et } Y \text{ sont indépendantes.}$$

Exercice 10 (Démonstration des propositions 1 et 2 du cours)**.** ♢ Soit $X_1, X_2, \ldots, X_n$ un échantillon tiré aléatoirement dans une population. On rappelle que $\mu_{\text{éch}} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

1. En utilisant la proposition 6 ci-dessus, montrer que $\mathbb{E}(\mu_{\text{éch}}) = \mu_{\text{pop}}$.
2. En utilisant la proposition 7 ci-dessus, montrer que $\sigma(\mu_{\text{éch}}) = \sigma_{\text{pop}}/\sqrt{n}$.

Exercice 11. ♢ Soit $X_1, X_2, \ldots, X_n \in \{0, 1\}$ un échantillon de valeurs égales à 0 ou 1 (interprétation : $X_i = 1$ lorsque le i-ème individu de l'échantillon possède la propriété étudiée). On note $p_{\text{éch}}$ la proportion d'individus tels que $X_i = 1$.

1. Expliquez pourquoi $p_{\text{éch}} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
2. On rappelle que $\sigma^2_{\text{éch}} = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$. Montrer que
$$\sigma_{\text{éch}} = \sqrt{p_{\text{éch}}(1 - p_{\text{éch}})}\,.$$
3. Prouvez la proposition 5 à partir de la proposition 4 et des questions précédentes.

A Résolution de l'exercice 2 avec le logiciel R

Voici un script R avec les calculs nécessaires pour répondre à l'exercice 2.

```
## Exercice 2

# Q1
x=c(490,492,497,502,505,490,495,492,490,497)
mean(x) # = 495
sqrt(mean(x^2)-mean(x)^2) # = 5

# Q2 (a)
poids = c(490,492,495,497,502,505)
effectifs=c(6,11,10,8,5,3)
moy = sum(poids*effectifs)/sum(effectifs) # = 495.4186
variance = sum(poids^2*effectifs)/sum(effectifs)-moy^2
ecarttype = sqrt(variance) # = 4.362806
# Oui, car n plus grand

# Q2 (b)
moy - 1.96*ecarttype/sqrt(43) # = 494.1146
moy + 1.96*ecarttype/sqrt(43) # = 496.7226

# Q2 (c)
moy - 2.576*ecarttype/sqrt(43) # = 493.7047
moy + 2.576*ecarttype/sqrt(43) # = 497.1325
# Commentaire : intervalle plus large, logique...
```

Chapitre 4

Introduction aux tests d'hypothèses

Résumé Nous introduisons la notion de *test d'hypothèses* et décrivons la méthodologie standard pour construire un test. On présentera aussi le vocabulaire usuel : *risques de première et de seconde espèce*, p*-valeur*, etc.

1 Introduction

1.1 Un problème pratique courant : tester entre deux hypothèses

Dans les chapitres précédents, nous avons expliqué comment estimer la moyenne d'une population à partir de l'observation d'un échantillon, et quelles étaient les marges d'erreurs associées. Dans ce chapitre, on étudie un autre type de problème : les *tests d'hypothèses*. Il s'agit de problèmes où l'on cherche à répondre par "oui" ou "non" à une question sur la population. Par exemple :

1. En Europe, tel médicament est-il plus efficace que tel autre médicament ? $\rightsquigarrow$ oui ou non ;
2. La vitesse de pointe moyenne des centaines de milliers de voitures produites par cette entreprise est-elle bien de $200\ km/h$ comme annoncé par le constructeur ? $\rightsquigarrow$ oui ou non ;
3. Ce gène est-il responsable de la production de vitamine B chez la bactérie *bacillus subtilis* ? $\rightsquigarrow$ oui ou non. (Remarque culturelle : bcp de biostatisticiens s'intéressent actuellement à des questions de ce type, du fait de la collecte de plus en plus massive de données génomiques.)

Ce type de questions survient dans presque tous les domaines de l'ingénierie et des sciences expérimentales au sens large (physique, biologie, chimie, économie, sociologie, etc). L'objectif du statisticien est d'y répondre le plus rigoureusement possible au vu des données expérimentales. Les techniques que nous allons étudier s'appellent des *tests statistiques*.

1.2 Construction d'un test à l'aide d'un intervalle de confiance

Exemple. Un chercheur américain affirme que la taille moyenne des libellules de l'espèce *Aeshna cyanea* est de $69\ mm$. Afin de vérifier la crédibilité de cette hypothèse scientifique, une équipe de chercheurs français prélève 105 libellules de cette espèce et relève leurs tailles, en mm :

taille	67	69	71	72	74	76
effectif	9	18	30	23	18	7

Ces données sont-elles cohérentes avec l'hypothèse du chercheur américain, ou permettent-elles au contraire de l'invalider ?

Un peu de vocabulaire En langage statistique, on dit qu'on cherche à tester

$$\begin{aligned} & H_0 : \mu_{\text{pop}} = 69\ mm \qquad \text{(l'hypothèse nulle/conservative)} \\ \text{contre} \quad & H_1 : \mu_{\text{pop}} \neq 69\ mm \qquad \text{(l'hypothèse alternative)} \end{aligned}$$

La règle de décision qu'on retiendra pour conserver ou rejeter l'hypothèse H_0 s'appelle un *test statistique*. Dans notre cas, une règle de décision naturelle consiste à :

- conserver l'hypothèse H_0 si la valeur $69\ mm$ se trouve dans l'intervalle de confiance

$$\left[\mu_{\text{éch}} - 1.96\,\frac{\sigma_{\text{éch}}}{\sqrt{n}}\,;\ \mu_{\text{éch}} + 1.96\,\frac{\sigma_{\text{éch}}}{\sqrt{n}}\right]\ ;$$

- rejeter l'hypothèse H_0 si la valeur $69\ mm$ n'est pas dans cet intervalle.

Ainsi, on ne rejette l'hypothèse H_0 que si la valeur suggérée 69 mm est *significativement* différente de $\mu_{\text{éch}}$. Ce test est très intuitif ; expliquons pourquoi il est raisonable d'un point de vue mathématique :

1er cas : H_0 est vraie. On a démontré que, lorsque n est assez grand, l'intervalle de confiance ci-dessus contient la vraie valeur μ_{pop} pour environ 95% des échantillons. Donc si l'hypothèse H_0 est vraie, c'est-à-dire si $\mu_{\text{pop}} = 69\ mm$, alors la valeur $69\ mm$ sera dans l'intervalle de confiance pour environ 95% des échantillons, donc on conservera l'hypothèse H_0 pour environ 95% des échantillons. Ainsi, nous ne prenons qu'un risque de 5% de nous tromper si H_0 est vraie.

2ème cas : H_1 est vraie. Si, à l'inverse, H_1 était vraie, alors $\mu_{\text{pop}} \neq 69\ mm$ et notre intervalle de confiance serait concentré autour de μ_{pop} avec une largeur proportionnelle à $1/\sqrt{n}$; donc, *si n est suffisamment grand*, il est peu probable que la valeur $69\ mm$ soit dans l'intervalle de confiance (cf. 2ème cas de la figure 4.1). Par conséquent : si H_1 est vraie et si n est suffisamment grand, il est fort probable qu'on rejette l'hypothèse H_0. (Reste néanmoins un cas problématique : lorsque H_1 est vraie mais que le nombre n d'observations est petit, le test peut conserver H_0 faute de davantage d'informations.)

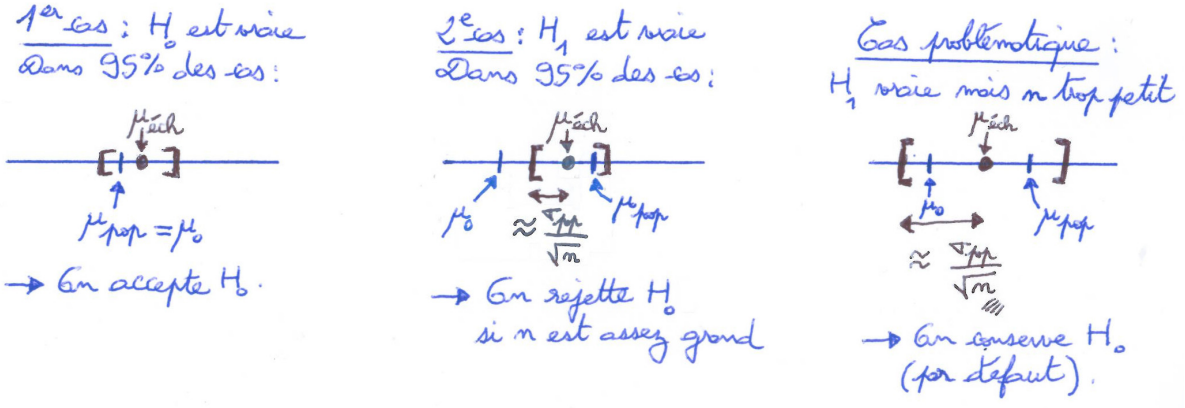


FIGURE 4.1 – (Sur tous les schémas, remplacer μ_0 par $69\ mm$.) A gauche et au milieu : le test détecte bien si l'hypothèse $H_0 : \mu_{\text{pop}} = \mu_0$ est vraie ou fausse. A droite : la situation est plus problématique car l'intervalle de confiance est trop large (à cause d'un petit nombre n d'observations) et ne permet donc pas de détecter que l'hypothèse H_0 est fausse.

Exemple (suite)**.** Appliquons la règle de décision ci-dessus au jeu de données collecté par les chercheurs français. Après calcul, on obtient $\mu_{\text{éch}} \approx 71.38\ mm$ et $\sigma_{\text{éch}} \approx 2.31\ mm$, ce qui

nous donne l'intervalle de confiance à 95% suivant : $[70.9\ mm\,;\ 71.9\ mm]$. D'après notre règle de décision, puisque la valeur $69\ mm$ n'est pas dans l'intervalle, on rejette l'hypothèse H_0. On conclut ainsi que les données collectées rendent peu crédible l'hypothèse du chercheur américain. (Cf. section 2.3 pour des détails sur l'interprétation des résultats.)

2 Méthode classique pour construire un test

Au paragraphe précédent, nous avons expliqué comment construire un test à partir d'un intervalle de confiance. Il existe une autre méthode, très proche, mais plus standard. Elle consiste à étudier ce qui se passerait si l'hypothèse H_0 était vraie, puis à déterminer quelles valeurs expérimentales sont peu probables sous l'hypothèse H_0, mais au contraire plus probables sous l'hypothèse H_1 ; si on observe ces valeurs, on fait alors le pari que c'est l'hypothèse H_1 qui est vraie. Nous détaillons cette démarche de façon rigoureuse dans les prochains paragraphes.

Précisons d'abord un peu les notations. On étudie une certaine population, de moyenne μ_{pop} inconnue. Notre connaissance sur cette population se limite à un échantillon $X_1, X_2, \ldots, X_n$ tiré aléatoirement dans cette population (on rappelle que les n tirages sont indépendants). Le problème est le suivant : on souhaite tester

$$\begin{aligned} & H_0 : \mu_{\text{pop}} = \mu_0 && \text{(l'hypothèse nulle/conservative)} \\ \text{contre} \quad & H_1 : \mu_{\text{pop}} \neq \mu_0 && \text{(l'hypothèse alternative)} \end{aligned}$$

2.1 Choix intuitif de la forme du test

Comment déterminer expérimentalement si $\mu_{\text{pop}} = \mu_0$ ou si $\mu_{\text{pop}} \neq \mu_0$? *Réponse intuitive* : on calcule la moyenne d'échantillon $\mu_{\text{éch}} = \frac{1}{n}\sum_{i=1}^{n} X_i$ puis on la compare à la valeur proposée μ_0 :

- Si $\mu_{\text{éch}}$ et μ_0 sont proches, alors on conserve l'hypothèse H_0.
- Si $\mu_{\text{éch}}$ et μ_0 sont significativement différentes, alors on rejette l'hypothèse H_0 et on opte pour l'hypothèse H_1.

Mais que signifient concrètement "proches" et "significativement différentes" ? Les mathématiques permettent d'apporter une réponse précise. Nous définirons la quantité

$$T = \frac{\mu_{\text{éch}} - \mu_0}{\sigma_{\text{éch}}/\sqrt{n}}$$

et nous retiendrons la règle de décision suivante :

- Si $|T| \leqslant c$ (c-à-d, si $\mu_{\text{éch}}$ et μ_0 sont proches), alors on conserve l'hypothèse H_0.
- Si $|T| > c$ (c-à-d, si $\mu_{\text{éch}}$ et μ_0 sont significativement différentes), alors on rejette l'hypothèse H_0 et on opte pour l'hypothèse H_1.

On explique ci-dessous comment choisir la valeur de c en fonction du risque que l'on souhaite prendre.

2.2 Construction précise et définitions

Encore un peu de vocabulaire Etant donné un test (c'est-à-dire une règle de décision vis-à-vis de H_0 et H_1), on appelle :

- *risque de première espèce* la probabilité de rejeter H_0 alors que H_0 est vraie ;
- *risque de seconde espèce* la probabilité de conserver H_0 alors que H_1 est vraie.

On adopte l'approche dite de *Neyman-Pearson*, qui donne un rôle particulier à l'hypothèse H_0.

Dissymétrie des hypothèses H_0 et H_1 En pratique, les hypothèses H_0 et H_1 ne joueront pas le même rôle :

- L'hypothèse nulle H_0 est une hypothèse par défaut, qu'on ne rejettera que si les preuves expérimentales en faveur de H_1 sont claires. Il peut d'agir d'une hypothèse communément établie, d'une hypothèse de prudence (parce que accepter H_1 serait par exemple coûteux industriellement), ou encore d'une hypothèse plus facile à étudier.
- L'hypothèse alternative H_1, qui est souvent l'hypothèse contraire de H_0, est à l'inverse ce qui doit à tout prix être démontré par l'expérience. Il peut s'agir d'une hypothèse scientifique novatrice, d'une hypothèse industrielle entraînant des coûts importants (ex : H_1 = "nouveau médicament plus efficace que le précédent" $\rightsquigarrow$ coûts de production importants, risques sanitaires), etc.

Comme H_0 est l'hypothèse de prudence (celle qu'on conservera par défaut), alors que H_1 est plus lourde de conséquences, on fera en sorte que tous nos tests aient un *petit risque de première espèce*. Concrètement : on fixera un risque de première espèce maximal α, par exemple $\alpha = 5\%$ ou $\alpha = 1\%$, et on fera en sorte que le risque de première espèce de notre test soit toujours inférieur ou égal à α.

Méthode de construction du test (valable lorsque le nombre n d'observations est assez grand) :

1. On étudie le cas où l'hypothèse H_0 vraie, c'est-à-dire $\mu_{\text{pop}} = \mu_0$. On applique alors le Théorème de la Limite Centrale (valable lorsque n est assez grand) : lorsque $\mu_{\text{pop}} = \mu_0$, on a

$$\mathbb{P}\left(-c \leqslant \frac{\mu_{\text{éch}} - \mu_0}{\sigma_{\text{éch}}/\sqrt{n}} \leqslant c\right) \approx \mathbb{P}(-c \leqslant Z \leqslant c)\,, \qquad \text{où la loi de } Z \text{ est } \mathcal{N}(0,1)\,.$$

2. On se donne un risque de première espèce maximal : $\alpha = 5\%$ par exemple, puis on cherche la valeur de c telle que $\mathbb{P}(-c \leqslant Z \leqslant c) = 1 - \alpha$. Cette valeur s'obtient par lecture d'une table numérique vue en TD (par exemple : $c = 1.96$ pour $\alpha = 5\%$).
3. On en déduit que, sous l'hypothèse H_0, la quantité

$$T = \frac{\mu_{\text{éch}} - \mu_0}{\sigma_{\text{éch}}/\sqrt{n}}$$

est comprise entre $-c$ et c avec probabilité $1 - \alpha$ environ. Avec $\alpha = 5\%$, cela signifie qu'on observerait $|T| \leqslant c$ pour environ 95% des échantillons, si H_0 était vraie.

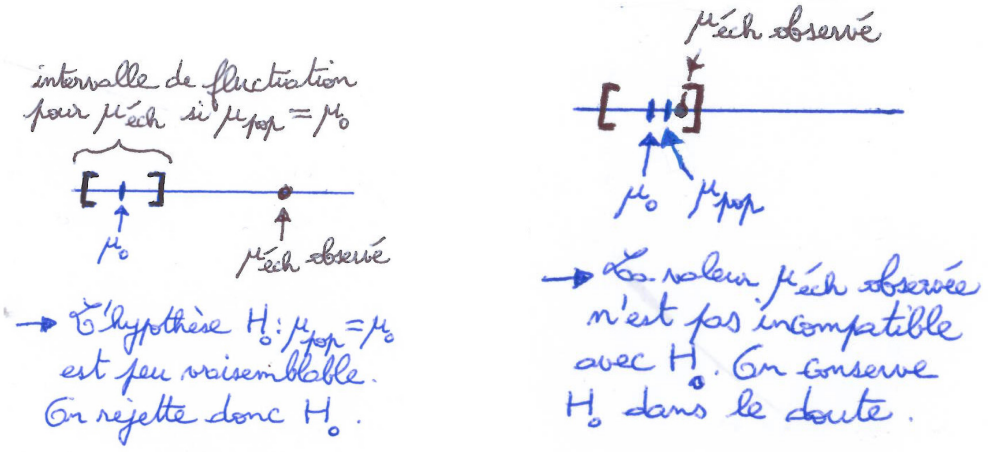
4. Application numérique : si $|T| \leqslant c$, on conserve l'hypothèse H_0 ; dans le cas contraire où $|T| > c$, on rejette H_0 et on opte pour l'hypothèse H_1.

2.3 Interprétation des résultats du test

Les résultats d'un test statistique sont à interpréter avec précaution (illustration en figure 4.2) :

- Si le test rejette l'hypothèse H_0 : alors on a de bonnes raisons de penser que H_0 est fausse (raisonnement par l'absurde : si H_0 était vraie, on aurait observé $|T| \leqslant 1.96$ pour environ 95% des échantillons, et on vient à l'inverse d'observer $|T| > c$.). Ainsi, le test rejette l'hypothèse H_0 lorsqu'on juge que la différence $\mu_{\text{éch}} - \mu_0$ est trop grande pour être simplement le fait de fluctuations aléatoires.
- Si le test conserve l'hypothèse H_0 : la seule chose qu'on peut vraiment dire est que la différence observée $\mu_{\text{éch}} - \mu_0$ n'est pas incompatible avec l'hypothèse H_0. Il se peut que $\mu_{\text{pop}} = \mu_0$, mais il se peut aussi que μ_{pop} soit proche mais distincte de μ_0 (cf. figure 4.2(b)). Dans le

doute, on a conservé H_0. On en conclut donc que ou bien H_0 est vraie, ou bien on n'a pas collecté assez de données pour pouvoir rejeter H_0 avec grande confiance. Attention à ne pas conclure trop rapidement !



(a) Situation où l'on rejette H_0

(b) Situation problématique : on conserve H_0 par défaut

FIGURE 4.2 – Deux situations différentes. A gauche : la valeur de $\mu_{\text{éch}}$ observée rend l'hypothèse H_0 peu vraisemblable, donc le test rejette H_0. A droite : la valeur observée $\mu_{\text{éch}}$ n'étant pas incompatible avec H_0, on conserve H_0 dans le doute (on commet donc une erreur car, en vérité, l'hypothèse H_0 est fausse).

Exemple (libellules)**.** Utilisons le test précédent pour tester l'hypothèse $H_0 : \mu_{\text{pop}} = 69\ mm$ contre l'hypothèse $H_1 : \mu_{\text{pop}} \neq 69\ mm$.

- La justification théorique avec le Théorème de la Limite Centrale, qui amène à comparer $|T|$ à $c = 1.96$, doit être comprise et réexpliquée si cela est demandé. On peut néanmoins sauter cette étape pour une simple application numérique.
- Comme d'habitude en biologie, on fixe le risque de première espèce maximal à $\alpha = 5\%$, donc il faudra comparer $|T|$ à $c = 1.96$.
- En utilisant les données expérimentales, on calcule

$$T = \frac{\mu_{\text{éch}} - 69}{\sigma_{\text{éch}}/\sqrt{n}} \approx 10.58\ .$$

On en déduit que $|T| > 1.96$, donc on rejette l'hypothèse H_0. Par conséquent, les données expérimentales rendent l'hypothèse du chercheur américain très peu crédible. [1]

2.4 Cas particulier : test sur la valeur d'une proportion

Dans cette section, on s'intéresse à la proportion p_{pop} des individus d'une population possédant une certaine propriété (ex : la proportion de Français qui voteront pour tel candidat, la proportion de rats qui développent une tumeur, la proportion de machines défectueuses, etc). Pour mieux connaître p_{pop}, on observe n individus tirés aléatoirement et indépendamment dans la population.

1. Il s'agit de la même conclusion qu'avec l'intervalle de confiance – heureusement !

On peut alors calculer la proportion d'échantillon $p_{\text{éch}}$. On s'intéresse au problème suivant : on souhaite tester

$$\begin{aligned} & H_0 : p_{\text{pop}} = p_0 \qquad \text{(l'hypothèse nulle/conservative)} \\ \text{contre} \quad & H_1 : p_{\text{pop}} \neq p_0 \qquad \text{(l'hypothèse alternative)} \end{aligned}$$

La construction d'un test sur une proportion p_{pop} est presque la même[2] que celle d'un test sur une moyenne de population μ_{pop}. Nous ne détaillerons donc pas la démarche théorique (avec le Théorème de la Limite Centrale) ; on décrit juste ci-dessous la règle de décision obtenue. On notera une différence avec le chapitre sur les intervalles de confiance : ici, on ne remplace pas $\sigma_{\text{éch}}$ par sa valeur $\sqrt{p_{\text{éch}}(1-p_{\text{éch}})}$, mais par une autre[3] valeur $\sqrt{p_0(1-p_0)}$.

Test sur une proportion (valable lorsque le nombre n d'observations est assez grand) :

1. On vérifie les trois hypothèses $n \geqslant 30$, $np_{\text{éch}} \geqslant 5$ et $n(1-p_{\text{éch}}) \geqslant 5$ (nécessaire pour utiliser le TLC).
2. On se donne un risque de première espèce maximal α et on lit sur une table numérique la valeur de c telle que $\mathbb{P}(-c \leqslant Z \leqslant c) = 1-\alpha$ (par exemple : $c = 1.96$ pour $\alpha = 5\%$).
3. On calcule $T = \dfrac{p_{\text{éch}} - p_0}{\sqrt{p_0(1-p_0)/n}}$. Si $|T| \leqslant c$, on conserve l'hypothèse H_0 ; dans le cas contraire où $|T| > c$, on rejette H_0 et on opte pour l'hypothèse H_1.

3 Compléments sur les tests

3.1 La notion de p-valeur

On rencontre très fréquemment le mot *p-valeur*, que ce soit en utilisant un logiciel de statistique ou en lisant un article de recherche en biologie, en physique, etc. Voici sa définition et quelques éléments pour comprendre son utilisation.

Définition (p-valeur) Reprenons le test que nous avons construit en section 2.2 pour tester $H_0 : \mu_{\text{pop}} = \mu_0$ contre $H_1 : \mu_{\text{pop}} \neq \mu_0$. Notre test rejette l'hypothèse H_0 lorsque $|T| > c$, où la quantité T est définie par

$$T = \frac{\mu_{\text{éch}} - \mu_0}{\sigma_{\text{éch}}/\sqrt{n}} \, .$$

En lisant la table de la loi $\mathcal{N}(0,1)$, on s'aperçoit que si on choisit un risque de première espèce α très petit, alors la valeur de c associée est très grande, si bien que $|T| \leqslant c$, donc on conserve H_0 dans ce cas. Si par contre on augmente progressivement la valeur de α, alors la valeur de c diminue jusqu'à ce que $|T| = c$ puis $|T| > c$, auquel cas on commence à rejeter H_0. La valeur de α limite, celle qui correspond à $|T| = c$, s'appelle *la p-valeur du test*.[4] Pour l'obtenir, il suffit de lire la table de la loi $\mathcal{N}(0,1)$ à l'envers.

Notation : on utilise souvent la lettre P pour désigner une p-valeur.

2. Ce n'est pas étonnant, car nous avons déjà expliqué au chapitre 3 qu'une proportion est un cas particulier de moyenne.

3. Utiliser ou non la valeur $\sqrt{p_0(1-p_0)}$ à la place de $\sqrt{p_{\text{éch}}(1-p_{\text{éch}})}$ donne dans les deux cas un test avec de bonnes garanties mathématiques. Cependant, il est plus usuel d'utiliser $\sqrt{p_0(1-p_0)}$. C'est pourquoi on présente cette approche.

4. Ainsi, la p-valeur d'un test est le plus petit risque α pour lequel on rejetterait l'hypothèse H_0 au vu des observations $X_1, X_2, \ldots, X_n$. Cette définition s'étend aisément – avec quelques précautions – à tous types de tests, par exemple tous ceux vus aux chapitre 5.

Utilisation : on rejette H_0 quand la p-valeur est petite Nous venons de voir que la p-valeur P est telle qu'on conserve H_0 pour tout $\alpha < P$ et on rejette H_0 pour tout $\alpha > P$. Cela permet d'en déduire la règle :

$$\text{on rejette } H_0 \Leftrightarrow P < \alpha$$

En biologie, on choisit souvent un risque de première espèce $\alpha = 5\%$, donc la règle ci-dessus devient : on rejette H_0 lorsque $P < 5\%$.

Dangers d'interprétation

- La seule chose que permet de dire une p-valeur est : plus la p-valeur est petite, plus la quantité $|T|$ est grande (d'après la table de loi), donc plus on a envie de croire à l'hypothèse H_1.
- Attention à éviter des phrases incorrectes mais fréquentes du type : "la p-valeur est la probabilité que H_0 soit fausse". Cette phrase n'a pas de sens car l'hypothèse H_0 n'est pas aléatoire : elle est totalement vraie ou totalement fausse, vérité qu'on ignore mais qu'on essaye de deviner à partir de l'échantillon.

3.2 Que faire quand le nombre n d'observations est petit ?

Le test qu'on a présenté jusqu'à maintenant n'est valable que *lorsque le nombre n d'observations est assez grand*. Il s'agit du domaine de validité du Théorème de la Limite Centrale ; on a notamment donné des conditions précises dans le cas du test sur une proportion. Que faire maintenant si n est petit ?

Test de Student (t-test) Le test de Student (ou t-test) est également un classique, qu'on trouve sur tous les logiciels, et qui peut s'appliquer quand n est petit. Il requiert cependant que la répartition des valeurs dans la population soit très proche d'une courbe de Gauss, ce qui n'est pas toujours le cas en pratique, et qui requiert d'avoir déjà des informations sur la population (via une certaine expertise par exemple). A noter que ce test s'applique à des *grandeurs continues* et non à des problèmes d'estimation de proportion. La marche à suivre pour un test de Student est la suivante :

1. On s'assure que la répartition des valeurs dans la population est très proche d'une courbe de Gauss. Cela peut être une information connue par les experts du domaine (si cela a par exemple été démontré de façon statistique sur un échantillon conséquent) ou être simplement une hypothèse de modélisation.
2. On se donne un risque de première espèce maximal α et on lit sur la table numérique de la loi de Student à $n-1$ degrés de libertés la valeur de c telle que $\mathbb{P}(-c \leqslant T \leqslant c) = 1 - \alpha$ (par exemple : $c = 2.365$ pour $\alpha = 5\%$ et $n = 8$).
3. On calcule
$$T = \frac{\mu_{\text{éch}} - \mu_0}{s_{\text{éch}}/\sqrt{n}} \ ,$$
où $s_{\text{éch}}$ est l'écart-type d'échantillon corrigé défini au chapitre 2 (penser à utiliser la formule : $s_{\text{éch}} = \sqrt{n/(n-1)} \cdot \sigma_{\text{éch}}$). Si $|T| \leqslant c$, on conserve l'hypothèse H_0 ; dans le cas contraire où $|T| > c$, on rejette H_0 et on opte pour l'hypothèse H_1.

Test non-asymptotique Dans le cas où la répartition des valeurs dans la population ne ressemble pas à une courbe de Gauss, on ne peut pas utiliser de test de Student. Il existe heureusement d'autres types de tests, qu'on appelle *tests non-asymptotiques*. Les aborder nécessiterait des connaissances mathématiques plus appronfondies. Le lecteur curieux pourra néanmoins en découvrir un exemple dans l'exercice 9.

Ne pas espérer de miracles avec peu d'observations Même s'il est possible de construire des tests quand les conditions d'utilisation du Théorème de la Limite Centrale ne sont pas vérifiées (c'est-à-dire quand n est petit), on ne peut pas espérer déduire beaucoup d'informations sur la population à partir d'un très petit nombre d'observations n. Cela se traduit par des intervalles de confiance très larges (peu précis) et par des tests très conservatifs (qui conservent souvent H_0, dans le doute). Autrement dit : les situations problématiques décrites sur la figure 4.1 (schéma de droite) et sur la figure 4.2(b) se produiront assez souvent.

4 Exercices

Exercice 1. Une étude a montré que l'apport moyen journalier en vitamine D nécessaire au bon développement des bébés est de $\mu_0 = 10\ \mu g$ par jour. On se demande si la quantité de vitamine D apportée par l'alimentation sous forme de lait en poudre produit par une certaine marque est conforme avec cette norme. On note μ_{pop} la quantité moyenne journalière de vitamine D apportée par cette alimentation. On considère un échantillon formé de 50 bébés âgés de trois mois et nourris au biberon. On mesure, lors d'une journée bien identifiée, l'apport en vitamine D du lait en poudre pour chacun de ces bébés. On calcule alors l'apport moyen ($9.1\ \mu g$) et l'écart-type d'échantillon ($1.9\ \mu g$).

1. Où est l'aléatoire dans ce contexte ?
2. Utilisez un test statistique avec le risque de première espèce $\alpha = 0.05$ pour tester l'hypothèse nulle $H_0 : \mu_{\text{pop}} = \mu_0$ contre l'hypothèse alternative $H_1 : \mu_{\text{pop}} \neq \mu_0$. L'alimentation avec cette marque de lait en poudre fournit-elle la quantité voulue de vitamine D aux bébés ?
3. Même question avec le risque de première espèce $\alpha = 0.01$.
4. Quelle grandeur pourrait-on fournir en guise de résultat afin de ne pas reconduire le test pour chaque nouvelle valeur de α ?

Exercice 2. Un fabricant de boîtes de conserve affirme que le poids moyen des boîtes de sa production est de $495\ g$. Une agence de contrôle vient vérifier ses dires et prélève 43 boîtes dont les poids se répartissent de la façon suivante :

poids	490	492	495	497	502	505
effectif	6	11	10	8	5	3

1. Les données prélevées viennent-elles contredire l'affirmation du fabricant ?
2. Une autre société de contrôle vient faire un prélèvement un peu plus conséquent et relève les poids de 149 boîtes :

poids	490	492	495	497	502	505
effectif	12	31	35	38	20	13

   Les données prélevées viennent-elles contredire l'affirmation du fabricant ? Qu'est-ce qui a changé par rapport à la question précédente ?

Exercice 3. Des étudiants de votre promo viennent vous présenter une étude qu'ils ont conduite sur 25 rats d'une même espèce. Sachant que le poids moyen des rats *Sprague-Dawley* est d'environ $350\ g$ et que les poids des 25 rats étudiés étaient (en grammes)

386 389 338 356 378 375 339 376 389 376 366 336 385
349 382 342 335 388 375 331 354 363 371 378 361

pensez-vous qu'ils s'agissait de rats *Sprague-Dawley* ? Quelle hypothèse de modélisation avez-vous faite ?

Exercice 4 (Pile ou face consécutifs)**.** On reprend l'activité 1 du chapitre 2. Vous demandez à un ami de lancer 100 fois une pièce et de noter les résultats obtenus (pile ou face). Vous suggérez à votre ami qu'il peut soit effectuer les 100 lancers consciencieusement, soit tricher en imaginant une suite de 100 lancers.

Un calcul mathématique ou une simple simulation numérique permet de montrer la chose suivante : si les 100 lancers sont indépendants et si la pièce est bien équilibrée, alors la loi de la variable aléatoire $Z =$ "nombre maximal de pile ou face consécutifs parmi les 100 lancers" est donnée approximativement par

a	1	2	3	4	5	6	7	8	9	10
$\mathbb{P}(Z=a)$	0	0	0.00031	0.02989	0.16911	0.26638	0.22414	0.14388	0.08093	0.04287
a	11	12	13	14	15	16	17	18	...	
$\mathbb{P}(Z=a)$	0.02095	0.01117	0.00546	0.00248	0.00133	0.00056	0.00026	0.00012	...	

Question : comment allez-vous procéder pour déterminer si votre ami a vraiment lancé les pièces ou s'il a triché ? La réponse doit bien sûr être mathématiquement fondée.

Exercice 5. Afin de tester une solution toxique, on fait des injections à un groupe de 80 souris. Il est d'usage de penser que l'injection est mortelle dans 80% des cas. Le fait que 22 souris ne soient pas mortes est-il compatible au seuil 5% avec cette hypothèse ?

Exercice 6. Un journaliste souhaiterait savoir si la parité est respectée parmi les agents secrets français. Ces données n'étant pas publiques, il entreprend une petite étude statistique en interrogeant plusieurs contacts indépendants bien placés dans le renseignement. Ses résultats sont les suivants : sur les 73 agents secrets dont il a vu la fiche descriptive, 28 sont des femmes.

1. En prenant un risque de première espèce de 5%, le journaliste peut-il tirer une conclusion sur la parité parmi les agents secrets français ? Que signifie ce risque de première espèce ?
2. Qu'en est-il avec un risque de première espèce de 2% ?

Exercice 7. Quel message peut-on tirer de ce dessin humoristique ?

Source : http://xkcd.com/1448/

Exercices d'approfondissement

Exercice 8 (Dualité entre test et intervalle de confiance)**.** $\diamondsuit$ On considère une population de N valeurs $x_1, x_2, \ldots, x_N$ dont on souhaite connaître la moyenne μ_{pop}. Pour cela, en tant que statisticien, vous prélevez un échantillon $X_1, X_2, \ldots, X_n$ dans cette population. Vous avez alors accès à la moyenne d'échantillon $\mu_{\text{éch}}$ et à l'écart-type d'échantillon $\sigma_{\text{éch}}$.

1. Rappeler la formule du cours pour un intervalle de confiance $[a, b]$ à 95% pour μ_{pop}. Rappeler ensuite comment effectuer un test de $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$. Remarquez-vous des similarités entre l'intervalle de confiance et le test obtenus ?
2. Montrer que l'intervalle de confiance $[a, b]$ et le test ci-dessus sont liés par la relation

$$\text{On rejette } H_0 \Leftrightarrow \mu_0 \notin [a, b] \, .$$

   Cela vous rappelle-t-il un passage du cours ?
3. $\diamondsuit$ Dans le cadre d'un test sur une proportion (section 2.4 du cours), le test proposé n'est pas exactement celui qu'on pourrait obtenir via l'intervalle de confiance. Quelle est la différence ? Quel test aurait-ton obtenu en utilisant directement l'intervalle de confiance ?

Exercice 9 (Test rigoureux lorsque n est petit)**.** $\diamondsuit\diamondsuit$ On considère une population de N valeurs $x_1, x_2, \ldots, x_N$ dont on souhaite connaître la moyenne μ_{pop}. Pour cela, en tant que statisticien, vous prélevez un échantillon $X_1, X_2, \ldots, X_n$ dans cette population. Vous avez alors accès à la moyenne d'échantillon $\mu_{\text{éch}}$ et à l'écart-type d'échantillon $\sigma_{\text{éch}}$. Nous allons construire un test adapté au cas où le nombre n d'observations est petit (on parle de *test non-asymptotique*). Cet exercice est difficile.

1. (a) Soit Z est une variable aléatoire positive et $x > 0$. Montrer que $\mathbb{E}(Z) \geqslant x\, \mathbb{P}(Z > x)$. En déduire l'inégalité de Markov :

$$\mathbb{P}(Z > x) \leqslant \frac{\mathbb{E}(Z)}{x} \, .$$

   (b) En utilisant la question précédente, prouver l'inégalité de Bienaymé-Tchebychev, valable pour toute variable aléatoire Y (de variance finie) :

$$\mathbb{P}\big(|Y - \mathbb{E}(Y)| > x\big) \leqslant \frac{\mathrm{Var}(Y)}{x^2} \, .$$

   (c) En déduire que, pour tout $x > 0$, on a :

$$\mathbb{P}\big(|\mu_{\text{éch}} - \mu_{\text{pop}}| > x\big) \leqslant \frac{\sigma^2_{\text{pop}}}{n\, x^2} \, .$$

2. Nous allons maintenant construire un test de $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$. On supposera pour simplifier que l'on connait une majoration $\sigma^2_{\text{pop}} \leqslant 1/4$ (bonus : montrer que c'est vrai dès lors que tous les $x_i \in [0, 1]$). En utilisant cette majoration, construire un test de niveau α de H_0 contre de H_1.
3. Dans le même esprit, constuire un intervalle de confiance à 95% sur μ_{pop}.

A Résolution des exercices 2 et 3 avec le logiciel R

Voici un script R permettant d'effectuer les calculs nécessaires à la résolution des exercices 2 et 3.

```
## Exercice 2

poids = c(490,492,495,497,502,505)
effectifs=c(6,11,10,8,5,3)
moy = sum(poids*effectifs)/sum(effectifs) # = 495.4186
variance = sum(poids^2*effectifs)/sum(effectifs)-moy^2
ecarttype = sqrt(variance) # = 4.362806
(moy-495)/(ecarttype/sqrt(43)) # = 0.6291764

poids2 = c(490,492,495,497,502,505)
effectifs2 = c(12,31,35,38,20,13)
moy2 =  sum(poids2*effectifs2)/sum(effectifs2) # = 496.2953
variance2 = sum(poids2^2*effectifs2)/sum(effectifs2)-moy2^2
ecarttype2 = sqrt(variance2) # = 4.304677
sum(effectifs2) # = 149
(moy2-495)/(ecarttype2/sqrt(149)) # = 3.673022

## Exercice 3

poidsrats=c(386,389,338,356,378,375,339,376,389,376,366,336,385,
            349,382,342,335,388,375,331,354,363,371,378,361)
T=(mean(poidsrats)-350)/(sd(poidsrats)/sqrt(25))

# On obtient T = 3.814479 > t(24,5%) --> on rejette H0
```

Chapitre 5

Un aperçu de différents tests d'hypothèses

Résumé Dans ce chapitre, nous passons en revue divers tests d'hypothèses afin de répondre à divers types de questions. La liste n'est bien sûre pas exhaustive. Nous ne détaillons ni la démarche théorique, ni l'interprétation, qui ont été vues en détails au chapitre précédent.

1 Pourquoi un catalogue de tests ?

Dans ce chapitre, nous allons recenser plusieurs tests différents. La raison est que toutes les questions statistiques rencontrées dans les sciences expérimentales ou en ingénierie ne sont pas les mêmes. On peut les regrouper en plusieurs catégories à l'aide des critères suivants :

1. Cherche-t-on à conduire un test sur une seule population, ou plutôt à comparer deux/plusieurs populations entre elles ?
2. S'intéresse-t-on à la valeur d'une moyenne de population, ou à une proportion dans une population ?
3. Dispose-t-on de beaucoup d'observations (n grand) ou de peu d'observations (n petit) ?
4. Quelle est la nature des hypothèses H_0 et H_1 ? Par exemple, cherche-t-on à tester

$$\left.\begin{array}{rl} & H_0 : \mu_{\text{pop}} = \mu_0 \\ \text{contre} & H_1 : \mu_{\text{pop}} \neq \mu_0 \end{array}\right\} \text{test bilatéral}$$ [1]

ou cherche-t-on plutôt à tester

$$\left.\begin{array}{rl} & H_0 : \mu_{\text{pop}} < \mu_0 \\ \text{contre} & H_1 : \mu_{\text{pop}} > \mu_0 \end{array}\right\} \text{test unilatéral}$$

Exemple où le deuxième jeu d'hypothèses intervient : une multinationale qui produit des voitures hésite à investir dans de nouvelles machines (plusieurs milliers). Un responsable en fait donc tester quelques dizaines et note ensuite si un gain de performances a été constaté. Si on note μ_{pop} le gain moyen de performances que permettrait l'achat de milliers de machines, a-t-on plutôt $\mu_{\text{pop}} < 0$ ou $\mu_{\text{pop}} > 0$? La deuxième hypothèse est plus risquée car elle implique un gros investissement financier. On prendra donc comme hypothèse alternative $H_1 : \mu_{\text{pop}} > 0$.

Il existe de nombreux tests statistiques pour tenter de répondre à la multitude de questions possibles. Nous en décrivons quelques-uns dans ce chapitre.

1. En anglais, les termes statistiques *bilatéral* et *unilatéral* se traduisent par *two-sided* et *one-sided*.

2 Tests sur des moyennes ou des proportions de populations

2.1 Etude d'une seule population

2.1.1 Test sur une moyenne μ_{pop}

On s'intéresse à une certaine population, de moyenne μ_{pop} inconnue. Notre connaissance sur cette population se limite à un échantillon $X_1, X_2, \ldots, X_n$ tiré aléatoirement dans cette population (on rappelle que les n tirages sont indépendants). On souhaite effectuer un test bilatéral ou unilatéral :

test bilatéral	test unilatéral
$H_0 : \mu_{\text{pop}} = \mu_0$	$H_0 : \mu_{\text{pop}} \leqslant \mu_0$
$H_1 : \mu_{\text{pop}} \neq \mu_0$	$H_1 : \mu_{\text{pop}} > \mu_0$

On dispose des quantités

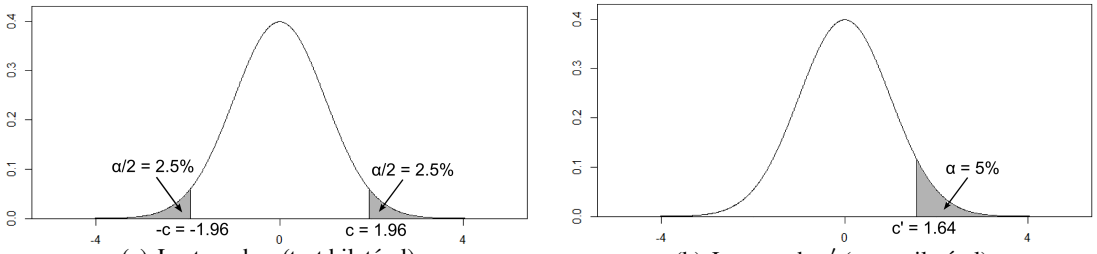
$$\mu_{\text{éch}} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \sigma_{\text{éch}} = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \mu_{\text{éch}}^2} \qquad s_{\text{éch}} = \sqrt{\frac{n}{n-1}}\,\sigma_{\text{éch}}$$

<table>
<tr><th>n grand</th><th>n petit</th></tr>
<tr><td>Nom du test : "z-test"
Condition : n "assez" grand</td><td>Nom du test : "t-test" (ou "test de Student")
Condition : loi des obs. = courbe de Gauss (expertise requise, ou hypothèse de modélisation)</td></tr>
<tr><td>• Choisir le risque de 1ère espèce α.
• Calculer
$$T = \frac{\mu_{\text{éch}} - \mu_0}{\sigma_{\text{éch}}/\sqrt{n}}$$
• Rejeter H_0 lorsque
<table><tr><td>$H_0 : \mu_{\text{pop}} = \mu_0$
$H_1 : \mu_{\text{pop}} \neq \mu_0$</td><td>$H_0 : \mu_{\text{pop}} \leqslant \mu_0$
$H_1 : \mu_{\text{pop}} > \mu_0$</td></tr><tr><td>$\lvert T\rvert > c$</td><td>$T > c'$</td></tr></table>Lecture de c et c' sur la table $\mathcal{N}(0,1)$.</td><td>• Choisir le risque de 1ère espèce α.
• Calculer
$$T = \frac{\mu_{\text{éch}} - \mu_0}{s_{\text{éch}}/\sqrt{n}}$$
• Rejeter H_0 lorsque
<table><tr><td>$H_0 : \mu_{\text{pop}} = \mu_0$
$H_1 : \mu_{\text{pop}} \neq \mu_0$</td><td>$H_0 : \mu_{\text{pop}} \leqslant \mu_0$
$H_1 : \mu_{\text{pop}} > \mu_0$</td></tr><tr><td>$\lvert T\rvert > c$</td><td>$T > c'$</td></tr></table>Lecture de c et c' sur la table de la loi de Student de paramètre $d = n - 1$.</td></tr>
</table>

Remarque (bilatéral versus unilatéral)**.** La règle de rejet de H_0 n'est pas la même pour un test bilatéral ou unilatéral :

- Pour un test bilatéral, on s'attend à ce que $|T|$ soit proche de 0 si H_0 est vraie, et que $|T|$ prenne de grandes valeurs si H_1 est vraie. Donc on rejette H_0 quand $|T| > c$.
- Pour un test unilatéral de $H_0 : \mu_{\text{pop}} \leqslant \mu_0$ contre $H_1 : \mu_{\text{pop}} > \mu_0$, on s'attend à ce que T soit négative (ou peu positive) si H_0 est vraie, et que T soit très positive si H_1 est vraie. Donc on rejette H_0 quand $T > c'$.

Le choix rigoureux des valeurs critiques c et c' est réalisé de sorte que le risque de première espèce du test soit toujours contrôlé par α. On explique ci-après comment lire ces valeurs c et c' en pratique.



(a) Lecture de c (test bilatéral) (b) Lecture de c' (test unilatéral)

FIGURE 5.1 – Lecture de c et c' sur la table de la loi $\mathcal{N}(0,1)$. Les exemples sont donnés pour $\alpha = 5\%$.

Remarque (Comment lire c et c' ?)**.** Dans le cas de grands échantillons, la lecture de c et c' se fait sur les deux tables de la loi $\mathcal{N}(0,1)$ distribuées en TD. Cf. explication en figure 5.1.

Dans le cas de petits échantillons, la lecture de c et c' se fait sur la table de la loi de Student distribuée en TD. Attention néanmoins : une lecture directe de la table ne donne que la valeur de c (test bilatéral). Pour avoir la valeur de c' (test unilatéral), il suffit de lire la même table mais avec une valeur de α deux fois plus grande (car le α indiqué sur la table correspond à l'aire grisée à gauche et à droite). Par exemple, pour $n = 18$ et $\alpha = 5\%$, on lit la table à la ligne $d = 18-1 = 17$ et à la colonne $\alpha = 2 \times 5\% = 10\%$, ce qui donne $c' = 1.74$.

2.1.2 Test sur une proportion p_{pop}

<table>
<tr><th>n grand</th><th>n petit</th></tr>
<tr><td>Nom du test : "z-test"
Condition : $n \geqslant 30$, $np_{\text{éch}} \geqslant 5$
et $n(1-p_{\text{éch}}) \geqslant 5$</td><td>Nom du test : "test binomial"
Condition : aucune</td></tr>
<tr><td>• Choisir le risque de 1ère espèce α.
• Calculer
$$T = \frac{p_{\text{éch}} - p_0}{\sqrt{p_0(1-p_0)/n}}$$
• Rejeter H_0 lorsque
<table><tr><td>$H_0 : p_{\text{pop}} = p_0$
$H_1 : p_{\text{pop}} \neq p_0$</td><td>$H_0 : p_{\text{pop}} \leqslant p_0$
$H_1 : p_{\text{pop}} > p_0$</td></tr><tr><td>$\lvert T\rvert > c$</td><td>$T > c'$</td></tr></table>Lecture de c et c' sur la table $\mathcal{N}(0,1)$.</td><td>Pas au programme
(mais utilisable sur un logiciel de statistiques)</td></tr>
</table>

2.2 Comparaison de deux échantillons/populations

On s'intéresse maintenant au cas où on dispose de deux échantillons $X_1, \ldots, X_{n_1}$ et $Y_1, \ldots, Y_{n_2}$ issus de deux populations de moyennes $\mu_{\text{pop},1}$ et $\mu_{\text{pop},2}$ inconnues. On souhaite comparer ces deux moyennes à l'aide d'un test bilatéral ou unilatéral :

test bilatéral	test unilatéral
$H_0 : \mu_{\text{pop},1} = \mu_{\text{pop},2}$ $H_1 : \mu_{\text{pop},1} \neq \mu_{\text{pop},2}$	$H_1 : \mu_{\text{pop},1} \leqslant \mu_{\text{pop},2}$ $H_1 : \mu_{\text{pop},1} > \mu_{\text{pop},2}$

On dispose des quantités

$$\mu_{\text{éch},1} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \qquad \sigma_{\text{éch},1} = \sqrt{\left(\frac{1}{n_1}\sum_{i=1}^{n_1} X_i^2\right) - \mu_{\text{éch},1}^2} \qquad s_{\text{éch},1} = \sqrt{\frac{n_1}{n_1-1}}\,\sigma_{\text{éch},1}$$

$$\mu_{\text{éch},2} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i \qquad \sigma_{\text{éch},2} = \sqrt{\left(\frac{1}{n_2}\sum_{i=1}^{n_2} Y_i^2\right) - \mu_{\text{éch},2}^2} \qquad s_{\text{éch},2} = \sqrt{\frac{n_2}{n_2-1}}\,\sigma_{\text{éch},2}$$

On distingue deux cas :

- lorsque les échantillons 1 et 2 sont *indépendants* (mesures réalisées sur des individus différents et indépendamment les unes des autres). Dans ce cas, on peut avoir $n_1 \neq n_2$.
- lorsque les échantillons 1 et 2 sont *appariés* (par ex : lorsque les mesures des échantillons 1 et 2 sont prises sur les mêmes individus mais à deux instants différents, ou sur des couples d'individus liés l'un à l'autre). Dans ce cas, on a $n_1 = n_2 = n$.

2.2.1 Échantillons indépendants

• **Comparaison de moyennes $\mu_{\text{pop},1}$ et $\mu_{\text{pop},2}$**

<table>
<tr><th>n_1 et n_2 grands</th><th>n_1 ou n_2 petit</th></tr>
<tr><td>Nom du test : "z-test"
Condition : n_1 et n_2 "assez" grands</td><td>Nom du test : "t-test" (ou "test de Student")
2 conditions : (a) loi des obs. = courbe de Gauss
(b) écarts-types égaux $\sigma_{\text{pop},1} = \sigma_{\text{pop},2}$
(expertise requise, ou hyp de modélisation)</td></tr>
<tr><td>• Choisir le risque de 1ère espèce α.
• Calculer
$$T = \frac{\mu_{\text{éch},1} - \mu_{\text{éch},2}}{\sqrt{\frac{\sigma_{\text{éch},1}^2}{n_1} + \frac{\sigma_{\text{éch},2}^2}{n_2}}}$$
• Rejeter H_0 lorsque
<table>
<tr><td>$H_0 : \mu_{\text{pop},1} = \mu_{\text{pop},2}$
$H_1 : \mu_{\text{pop},1} \neq \mu_{\text{pop},2}$</td><td>$H_0 : \mu_{\text{pop},1} \leqslant \mu_{\text{pop},2}$
$H_1 : \mu_{\text{pop},1} > \mu_{\text{pop},2}$</td></tr>
<tr><td>$|T| > c$</td><td>$T > c'$</td></tr>
</table>
Lecture de c et c' sur la table $\mathcal{N}(0,1)$.</td><td>• Choisir le risque de 1ère espèce α.
• Calculer
$$T = \frac{\mu_{\text{éch},1} - \mu_{\text{éch},2}}{s\sqrt{1/n_1 + 1/n_2}}$$
où $\quad s = \sqrt{\frac{(n_1-1)s_{\text{éch},1}^2 + (n_2-1)s_{\text{éch},2}^2}{n_1+n_2-2}}$
• Rejeter H_0 lorsque
<table>
<tr><td>$H_0 : \mu_{\text{pop},1} = \mu_{\text{pop},2}$
$H_1 : \mu_{\text{pop},1} \neq \mu_{\text{pop},2}$</td><td>$H_0 : \mu_{\text{pop},1} \leqslant \mu_{\text{pop},2}$
$H_1 : \mu_{\text{pop},1} > \mu_{\text{pop},2}$</td></tr>
<tr><td>$|T| > c$</td><td>$T > c'$</td></tr>
</table>
Lecture de c et c' sur la table de la loi de Student de paramètre $d = n_1 + n_2 - 2$.</td></tr>
</table>

Remarque : le test de Student dans le cas "n_1 ou n_2 petit" nécessite 2 conditions, et notamment que les écarts-types $\sigma_{\text{pop},1}$ et $\sigma_{\text{pop},2}$ des deux populations soient égaux. Si on a de bonnes raisons de penser que ce n'est pas le cas, il faut utiliser un autre test (test d'Aspin-Welch).

• **Comparaison de proportions $p_{\text{pop},1}$ et $p_{\text{pop},2}$**

Même problème qu'au paragraphe précédent, si ce n'est qu'on cherche maintenant à comparer des *proportions* $p_{\text{pop},1}$ et $p_{\text{pop},2}$ de deux populations. On suppose toujours que les échantillons 1 et 2 sont *indépendants*.

<table>
<tr><th>n_1 et n_2 grands</th><th>n_1 ou n_2 petit</th></tr>
<tr><td>Nom du test : "z-test"
2 jeux de conditions : $n_1 \geqslant 30$, $n_1 p_{\text{éch},1} \geqslant 5$ et $n_1(1-p_{\text{éch},1}) \geqslant 5$ ainsi que $n_2 \geqslant 30$, $n_2 p_{\text{éch},2} \geqslant 5$ et $n_2(1-p_{\text{éch},2}) \geqslant 5$</td><td>Nom du test : variante du "test binomial"
Condition : aucune</td></tr>
<tr><td>• Choisir le risque de 1ère espèce α.
• Calculer
$$T = \frac{p_{\text{éch},1} - p_{\text{éch},2}}{\sqrt{p(1-p)}\,\sqrt{1/n_1 + 1/n_2}}$$
où $\quad p = \dfrac{n_1 p_{\text{éch},1} + n_2 p_{\text{éch},2}}{n_1 + n_2}$
• Rejeter H_0 lorsque
<table><tr><td>$H_0 : p_{\text{pop},1} = p_{\text{pop},2}$
$H_1 : p_{\text{pop},1} \neq p_{\text{pop},2}$</td><td>$H_0 : p_{\text{pop},1} \leqslant p_{\text{pop},2}$
$H_1 : p_{\text{pop},1} > p_{\text{pop},2}$</td></tr><tr><td>$\lvert T\rvert > c$</td><td>$T > c'$</td></tr></table>Lecture de c et c' sur la table $\mathcal{N}(0,1)$.</td><td>Pas au programme
(mais utilisable sur un logiciel de statistiques)</td></tr>
</table>

2.2.2 Échantillons appariés

On traite maintenant le cas où les échantillons 1 et 2 sont *appariés*, c'est-à-dire lorsque les résultats de mesures ne sont plus indépendants d'un échantillon à l'autre. Plus précisément :

- On dispose du même nombre d'observations dans chaque échantillon ($n_1 = n_2 = n$). On observe ainsi : $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_n$.
- Les i-èmes observations X_i et Y_i des deux échantillons sont appariées (par ex : mesures réalisées sur le même individu mais à deux instants différents, ou sur des couples d'individus liés l'un à l'autre). Les individus (ou couples d'individus) i sont par contre observés indépendamment les uns des autres.

• **Comparaison de moyennes $\mu_{\text{pop},1}$ et $\mu_{\text{pop},2}$**

Attention : les notations $\mu_{\text{pop},1}$ et $\mu_{\text{pop},2}$ sont légèrement trompeuses car les échantillons ne proviennent pas vraiment de deux populations différentes, mais plutôt d'une même population étudiée à 2 instants différents. Cette remarque est d'ailleurs au fondement du test utilisé : nous allons procéder comme si nous ne disposions que d'un seul échantillon composé des n valeurs $X_i - Y_i$, puis regarder si la moyenne des différences

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) = \mu_{\text{éch},1} - \mu_{\text{éch},2}$$

est significativement non nulle (pour le test bilatéral) ou significativement positive (pour le test unilatéral). C'est pourquoi le test suivant ressemble en tous points au cas de l'étude d'une seule population.

<table>
<tr><th>n grand</th><th>n petit</th></tr>
<tr><td>Nom du test : "z-test"
Condition : n "assez" grand</td><td>Nom du test : "t-test" (ou "test de Student")
Condition : (loi de $X-Y$) = courbe de Gauss
(expertise requise, ou hyp de modélisation)</td></tr>
<tr><td>• Choisir le risque de 1ère espèce α.
• Calculer
$$T = \frac{\mu_{\text{éch},1} - \mu_{\text{éch},2}}{\widetilde{\sigma}_{\text{éch}}/\sqrt{n}}$$
$$\widetilde{\sigma}_{\text{éch}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2 - (\mu_{\text{éch},1} - \mu_{\text{éch},2})^2}$$
• Rejeter H_0 lorsque
<table><tr><td>$H_0 : \mu_{\text{pop},1} = \mu_{\text{pop},2}$
$H_1 : \mu_{\text{pop},1} \neq \mu_{\text{pop},2}$</td><td>$H_0 : \mu_{\text{pop},1} \leqslant \mu_{\text{pop},2}$
$H_1 : \mu_{\text{pop},1} > \mu_{\text{pop},2}$</td></tr><tr><td>$\lvert T\rvert > c$</td><td>$T > c'$</td></tr></table>Lecture de c et c' sur la table $\mathcal{N}(0,1)$.</td><td>• Choisir le risque de 1ère espèce α.
• Calculer
$$T = \frac{\mu_{\text{éch},1} - \mu_{\text{éch},2}}{\widetilde{s}_{\text{éch}}/\sqrt{n}}$$
où $\quad \widetilde{s}_{\text{éch}} = \sqrt{\frac{n}{n-1}} \times \widetilde{\sigma}_{\text{éch}}$
• Rejeter H_0 lorsque
<table><tr><td>$H_0 : \mu_{\text{pop},1} = \mu_{\text{pop},2}$
$H_1 : \mu_{\text{pop},1} \neq \mu_{\text{pop},2}$</td><td>$H_0 : \mu_{\text{pop},1} \leqslant \mu_{\text{pop},2}$
$H_1 : \mu_{\text{pop},1} > \mu_{\text{pop},2}$</td></tr><tr><td>$\lvert T\rvert > c$</td><td>$T > c'$</td></tr></table>Lecture de c et c' sur la table de la loi de Student de paramètre $d = n-1$.</td></tr>
</table>

• **Comparaison de proportions $p_{\text{pop},1}$ et $p_{\text{pop},2}$**

Pour simplifier, on ne traite que le cas où le nombre n d'observations est assez grand (sans donner de conditions précises car elles sont plus difficiles à exprimer).

Nous allons décrire la méthode sur un exemple. Une association de consommateurs cherche à comparer l'efficacité de deux lessives de marques différentes (disons A et B). Elle propose donc à un panel de personnes de tester chacune des deux lessives A et B puis de donner sa satisfaction (A_+/A_- : statisfaite/insatisfaite avec la lessive A, même chose avec B_+/B_-). Un statisticien contacté par l'association cherche alors à répondre à la question : au vu de l'échantillon de réponses, la proportion $p_{\text{pop},1}$ de personnes (dans toute la population) satisfaites par la lessive A est-elle égale à la proportion $p_{\text{pop},2}$ de personnes satisfaites par la lessive B ? ou est-elle inférieure, supérieure ? Le problème peut se formaliser à l'aide d'un test bilatéral ou unilatéral :

test bilatéral	test unilatéral
$H_0 : p_{\text{pop},1} = p_{\text{pop},2}$ $H_1 : p_{\text{pop},1} \neq p_{\text{pop},2}$	$H_1 : p_{\text{pop},1} \leqslant p_{\text{pop},2}$ $H_1 : p_{\text{pop},1} > p_{\text{pop},2}$

Attention : les notations $p_{\text{pop},1}$ et $p_{\text{pop},2}$ sont légèrement trompeuses car les résultats de satisfaction pour les lessives A et B ne proviennent pas vraiment de deux populations différentes, mais plutôt d'une même population étudiée à 2 instants différents (essai de la lessive A, puis essai de la lessive B). D'ailleurs, pour un individu i donné, les deux résultats de satisfaction recueillis pour A et B **ne sont pas indépendants** (certaines personnes sont toujours mécontentes, d'autres toujours satisfaites, etc). Les échantillons de réponses pour les lessives A et B sont donc *appariés*.

Le test est intuitif, puisqu'il compare le nombre $n_{(A_+,B_-)}$ de personnes de type (A_+, B_-) au nombre $n_{(A_-,B_+)}$ de personnes de type (A_-, B_+). (Les autres types d'individus n'apportent pas d'information.)

Démarche de test (comparaison de proportions avec deux échantillons appariés) :
- Condition d'utilisation : n "assez" grand.
- Choisir le risque de 1ère espèce α.
- Calculer
$$T = \frac{n_{(A_+,B_-)} - n_{(A_-,B_+)}}{\sqrt{n_{(A_+,B_-)} + n_{(A_-,B_+)}}}$$
où $n_{(A_+,B_-)}$ est le nombre de personnes de type (A_+, B_-), et où $n_{(A_-,B_+)}$ est le nombre de personnes de type (A_-, B_+).
- Rejeter H_0 lorsque

$H_0 : p_{\text{pop},1} = p_{\text{pop},2}$	$H_0 : p_{\text{pop},1} \leqslant p_{\text{pop},2}$
$H_1 : p_{\text{pop},1} \neq p_{\text{pop},2}$	$H_1 : p_{\text{pop},1} > p_{\text{pop},2}$
$\lvert T\rvert > c$	$T > c'$

Lecture de c et c' sur la table $\mathcal{N}(0,1)$.

2.3 Comparaison d'au moins trois populations : l'ANOVA

Les tests de la section 2.2 sont utilisables lorsqu'on cherche à comparer deux populations. Si on souhaite étudier au moins trois populations simultanément, on peut recourir à l'ANOVA (*analysis of variance*). Ce test statistique permet de déterminer l'influence (ou pas) d'un ou plusieurs facteurs catégoriels. On parle alors d'ANOVA à un facteur, deux facteurs, etc.

Exemple d'ANOVA à un facteur : vous êtes membre d'une association d'agriculteurs et vous souhaitez comparer l'efficacité de quatre fongicides différents pour traiter des plantations de maïs. Le *facteur* étudié est le type de fongicide ; il est catégoriel et possède quatre modalités. Vous effectuez alors une étude statistique en étudiant un échantillon de parcelles de maïs réparties en quatre groupes (selon le fongicide utilisé). Vous relevez ensuite le rendement $X_{k,i}$ de chaque parcelle de maïs, où k est le type du fongicide utilisé, et où i est le numéro de la parcelle dans le groupe k. Le principe de l'ANOVA à un facteur consiste à comparer les moyennes de chacun des sous-échantillons
$$\mu_{\text{éch},k} = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{k,i}$$
à la moyenne $\mu_{\text{éch}} = \frac{1}{n} \sum_{k=1}^{4} \sum_{i=1}^{n_k} X_{k,i}$ sur les quatre groupes (avec $n = n_1 + \ldots + n_4$). Si la quantité
$$T = \frac{\sum_{k=1}^{4} n_k (\mu_{\text{éch},k} - \mu_{\text{éch}})^2}{\sum_{k=1}^{4} \sum_{i=1}^{n_k} (X_{k,i} - \mu_{\text{éch},k})^2}$$
est significativement grande, la conclusion du test est que le facteur étudié est influent (ici, le type de fongicide a une influence sur le rendement).

En fait, le problème précédent est peut-être plus compliqué qu'il n'y paraît si le rendement du maïs est affecté non pas par un, mais par deux facteurs, par exemple : le type de fongicide et le degré de précipitation. Dans ce cas, il faudra utiliser une ANOVA à deux facteurs.

Le test ANOVA est présent sur tous les logiciels de statistique ; il suffit alors de saisir les données, de relever la p-value et de la comparer au risque de première espèce α. Attention néanmoins

à ne pas utiliser ce test hors de ses conditions d'application (les observations $X_{k,i}$ doivent être indépendantes et gaussiennes).

3 Tests du χ^2

Les tests du χ^2 (chi-deux) sont aussi un standard en statistique. Ils permettent de tirer des informations sur la loi de probabilité de variables aléatoires *catégorielles* (= possédant plusieurs modalités). Il s'agit en quelque sorte d'une extension de ce que nous avons vu précédemment avec des proportions (deux modalités) à un nombre arbitraire k de modalités.

3.1 Test du χ^2 d'adéquation à une loi discrète

On considère une expérience aléatoire où k résultats $a_1, a_2, \ldots, a_k$ différents peuvent se produire ; on note p_i la probabilité d'observer le résultat a_i. Dans ce paragraphe, on souhaite tester si les probabilités p_i sont toutes égales à des probabilités suggérées $p_i^{\text{réf}}$ ou pas :

$$H_0 : p_1 = p_1^{\text{réf}} \quad \text{et} \quad p_2 = p_2^{\text{réf}} \quad \text{et} \quad \ldots \quad \text{et} \quad p_k = p_k^{\text{réf}}$$
$$H_1 : \text{ce n'est pas le cas}$$

Deux exemples :

- On place une souris dans une boîte dont chacun des 4 coins correspond à un habitat donné. La souris se dirige alors vers l'un des coins (donc 4 résultats possibles). La question est : la souris est-elle indifférente au type d'habitat ($H_0 : p_1 = p_2 = p_3 = p_4 = 1/4$) ou aura-t-elle des préférences d'habitat (H_1 : les probabilités p_i ne valent pas toutes $1/4$) ?
- Une population est constituée de k groupes distincts, de proportions respectives $p_1, p_2, \ldots, p_k$. On peut alors se demander si les proportions p_i (inconnues) sont égales à des proportions suggérées $p_i^{\text{réf}}$. (Remarque : ici, l'expérience aléatoire consiste à tirer aléatoirement un individu dans la population ; la probabilité qu'il appartienne au i-ème groupe vaut p_i.)

Comme d'habitude en statistique, on s'appuie sur un échantillon : on répète *indépendamment* n fois la même expérience, et on relève le nombre n_i de fois où chaque modalité i a été observée (d'où $n_1 + n_2 + \ldots + n_k = n$).

Test du χ^2 d'adéquation à une loi discrète :

- Condition d'utilisation : $n \geqslant 30$ et $np_i^{\text{réf}} \geqslant 5$ pour tout $i = 1, \ldots, k$.
- On fixe le risque de première espèce α.
- On calcule les effectifs théoriques $np_i^{\text{réf}}$ de chaque modalité a_i, qu'on compare aux effectifs observés n_i :
$$T = \sum_{i=1}^{k} \frac{(\text{observé} - \text{théorique})^2}{\text{théorique}} = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{\text{réf}}\right)^2}{np_i^{\text{réf}}}$$
- On rejette H_0 lorsque $T > c$, où la valeur critique c est lue sur la table de la loi du χ^2 à $d = k - 1$ degrés de libertés.

3.2 Test du χ^2 d'indépendance entre deux variables catégorielles

On considère maintenant une expérience aléatoire où on relève simultanément les modalités de deux variables catégorielles. La première variable possède k modalités $a_1, \dots, a_k$, et la seconde variable possède ℓ modalités $b_1, \dots, b_\ell$. Exemple : on tire aléatoirement un individu dans une population et on relève son genre (deux modalités possibles) et son style de musique préféré parmi pop-rock, classique, jazz, électro, rap et variété française (6 modalités possibles).

La question est ici de déterminer si les deux variables catégorielles sont indépendantes ou non :

H_0 : les deux variables sont indépendantes

H_1 : les deux variables ne sont pas indépendantes

Comme d'habitude en statistique, on répète *indépendamment* n fois la même expérience, et on relève dans un tableau le nombre de fois $n_{i,j}$ où on a observé la modalité a_i pour la première variable et la modalité b_j pour la deuxième variable. Ce tableau s'appelle une *table de contigence*.

	b_1	b_2	$\dots$	b_ℓ	
a_1	$n_{1,1}$	$n_{1,2}$	$\dots$	$n_{1,\ell}$	$n_{1,\bullet}$
a_2	$n_{2,1}$	$n_{2,2}$	$\dots$	$n_{2,\ell}$	$n_{2,\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
a_k	$n_{k,1}$	$n_{k,2}$	$\dots$	$n_{k,\ell}$	$n_{k,\bullet}$
	$n_{\bullet,1}$	$n_{\bullet,2}$	$\dots$	$n_{\bullet,\ell}$	n

On a aussi calculé les sous-totaux des lignes et des colonnes : $n_{i,\bullet}$ désigne le nombre de fois où on a observé la modalité a_i pour la première variable (somme i-ème ligne), et $n_{\bullet,j}$ désigne le nombre de fois où on a observé la modalité b_j pour la deuxième variable (somme j-ème colonne). On a ainsi les relations suivantes :

$$n_{i,\bullet} = \sum_{j=1}^{\ell} n_{i,j} \qquad n_{\bullet,j} = \sum_{i=1}^{k} n_{i,j} \qquad n = \sum_{i=1}^{k} n_{i,\bullet} = \sum_{j=1}^{\ell} n_{\bullet,j} = \sum_{i=1}^{k}\sum_{j=1}^{\ell} n_{i,j}$$

Puisqu'on cherche à déterminer si les deux variables catégorielles sont indépendantes, on va comparer les effectifs observés $n_{i,j}$ aux effectifs théoriques $n\widehat{p}_i\widehat{q}_j$, où

$$\widehat{p}_i = \frac{n_{i,\bullet}}{n} \qquad \text{et} \qquad \widehat{q}_j = \frac{n_{\bullet,j}}{n}$$

sont les probabilités estimées pour la i-ème modalité de la première variable et la j-ème modalité de la seconde variable.

Test du χ^2 d'indépendence entre deux variables catégorielles :

- Condition d'utilisation : $n \geqslant 30$ et $n_{i,j} \geqslant 5$ pour tout i, j.
- On fixe le risque de première espèce α.
- On calcule les effectifs théoriques $n\widehat{p}_i\widehat{q}_j$ à l'aide des définitions ci-dessus, puis on les compare aux effectifs observés $n_{i,j}$:

$$T = \sum_{i=1}^{k}\sum_{j=1}^{\ell} \frac{(\text{observé} - \text{théorique})^2}{\text{théorique}} = \sum_{i=1}^{k}\sum_{j=1}^{\ell} \frac{\left(n_{i,j} - n\widehat{p}_i\widehat{q}_j\right)^2}{n\widehat{p}_i\widehat{q}_j}$$

- On rejette H_0 lorsque $T > c$, où la valeur critique c est lue sur la table de la loi du χ^2 à $d = (k-1)(\ell-1)$ degrés de libertés.

4 Autres exemples de tests

Nous n'avons recensé que les tests statistiques les plus simples, mais il en existe beaucoup d'autres. Par exemple, dans des articles de recherche en biologie, il est fréquent de rencontrer des tests non-paramétriques d'adéquation à une loi ou à un ensemble de lois pour des variables continues. [2] En voici quelques-uns :

- **Test de Kolmogorov-Smirnov**. Il vise à répondre à la question : est-ce que la répartition des valeurs dans ma population est égale (H_0) ou différente (H_1) d'une loi donnée ?
- **Test de normalité de Shapiro-Wilk**. Il vise à répondre à la question : est-ce que la répartition des valeurs dans ma population est (H_0) ou n'est pas (H_1) une courbe de Gauss ?
- **Test de Wilcoxon**. Il vise à répondre à la question : est-ce que la répartition des valeurs dans ma population est symétrique autour de zéro (H_0) ou pas (H_1) ?
- **Test de Mann-Whitney**. Il vise à répondre à la question : étant donnés deux échantillons indépendants issus de deux populations différentes, est-ce que les deux populations ont la même répartition de valeurs (H_0) ou est-ce que les valeurs de l'une sont globalement plus grandes que celles de l'autre (H_1) ?

5 Exercices

Exercice 1. Une étude médicale vise à connaître les risques de vertiges liés à la prise d'un certain médicament. Au total, 1163 patients soignés avec ce médicament ont été suivis et on a relevé les résultats suivants :

vertiges	pas de vertiges
108	1055

Ces résultats mettent-ils en doute l'affirmation du laboratoire pharmaceutique selon laquelle la proportion de gens subissant des vertiges après prise du médicament est inférieure à 8% ?

Exercice 2. Dans l'article *Sexual activity and the lifespan of male fruitflies*, Nature, 1981, Partridge et Farquhar rapportent des résultats d'expérience visant à étudier le coût de reproduction en termes de longévité pour les drosophiles mâles. Parmi les expériences menées, deux groupes de 25 mâles ont été comparés : dans le premier groupe, chacun des 25 mâles a été isolé avec une femelle vierge réceptive ; dans le second groupe, chacun des 25 mâles a été isolé avec une femelle non réceptive (rôle de contrôle). La durée de vie des mâles a été mesurée dans chaque groupe (en jours) :

groupe de 25 mâles avec	durée de vie moyenne	écart-type corrigé
1 femelle non-réceptive	64.80	15.6525
1 femelle réceptive	56.76	14.9284

Ces données permettent-elles d'affirmer qu'une activité sexuelle plus importante est un facteur de réduction de longévité chez la drosophile mâle ? Par ailleurs, sur quelle hypothèse repose votre test ?

2. A l'inverse, les tests du χ^2 sont d'abord conçus pour des variables catégorielles. (Même s'ils peuvent aussi s'utiliser pour des variables continues préalablement discrétisées.)

Exercice 3. Un chercheur laisse tomber (volontairement) un stylo dans un ascenceur et note si l'autre occupant l'aide à le ramasser. Il relève les résultats suivants, en fonction du genre de l'occupant :

	a aidé	n'a pas aidé	total
hommes	370	950	1320
femmes	300	1003	1303

Peut-on en déduire que, dans un ascenceur, les hommes et les femmes sont autant prêts à aider un autre occupant venant de laisser tomber son stylo ?

Exercice 4. Un médecin cherche à savoir si la pratique régulière du jogging entraîne une réduction de la fréquence cardiaque. Huit volontaires, qui ne pratiquaient pas le jogging auparavant, acceptent de suivre un programme d'entraînement de jogging pendant un mois. Leurs fréquences cardiaques moyennes ont été mesurées avant et après le programme d'entraînement (en nombre de battements par minute) :

fréquence cardiaque avant	74	86	98	102	78	84	79	70
fréquence cardiaque après	70	85	90	110	71	80	69	74

1. Ces données mettent-elles en évidence une réduction du rythme cardiaque après un mois d'entraînement ?
2. Quelle hypothèse de modélisation avez-vous faite à la question précédente ? Comment pourrait-on vérifier que cette hypothèse est valide ?
3. Parmi les huit volontaires, combien ont diminué leur rythme cardiaque ? Est-ce en contradiction avec votre réponse à la première question ?

Exercice 5. On souhaite savoir si les chimpanzés préfèrent le rouge, le bleu, le vert ou le jaune. Un chercheur réalise donc l'expérience suivante sur 83 chimpanzés : il leur montre 4 cartons (un par couleur, les 4 cartons sont de forme identique et sont présentés dans une disposition aléatoire) ; il attend ensuite que le chimpanzé choisisse l'un des cartons et note la couleur choisie. Les résultats obtenus sont les suivants :

rouge	bleu	vert	jaune
30	22	15	16

Cette expérience est-elle une preuve expérimentale du fait que les chimpanzés ont une préférence de couleur dans le choix des cartons ?

Exercice 6. Imaginons le scénario suivant : un sondage citoyen effectué auprès de 937 couples hétérosexuels en France vise à savoir si chacune des deux personnes du couple serait prête à partager la voiture de son conjoint pour aller au travail le matin. On obtient les résultats suivants :

	pour le partage de voiture	contre le partage de voiture	total
femmes	106	831	937
hommes	87	850	937

1. Les résultats obtenus sur ces deux échantillons d'hommes et de femmes mettent-ils en évidence un comportement différent des hommes et des femmes face au covoiturage ?
2. On dispose en fait des données plus précises suivantes : les 937 couples interrogés se répartissent selon

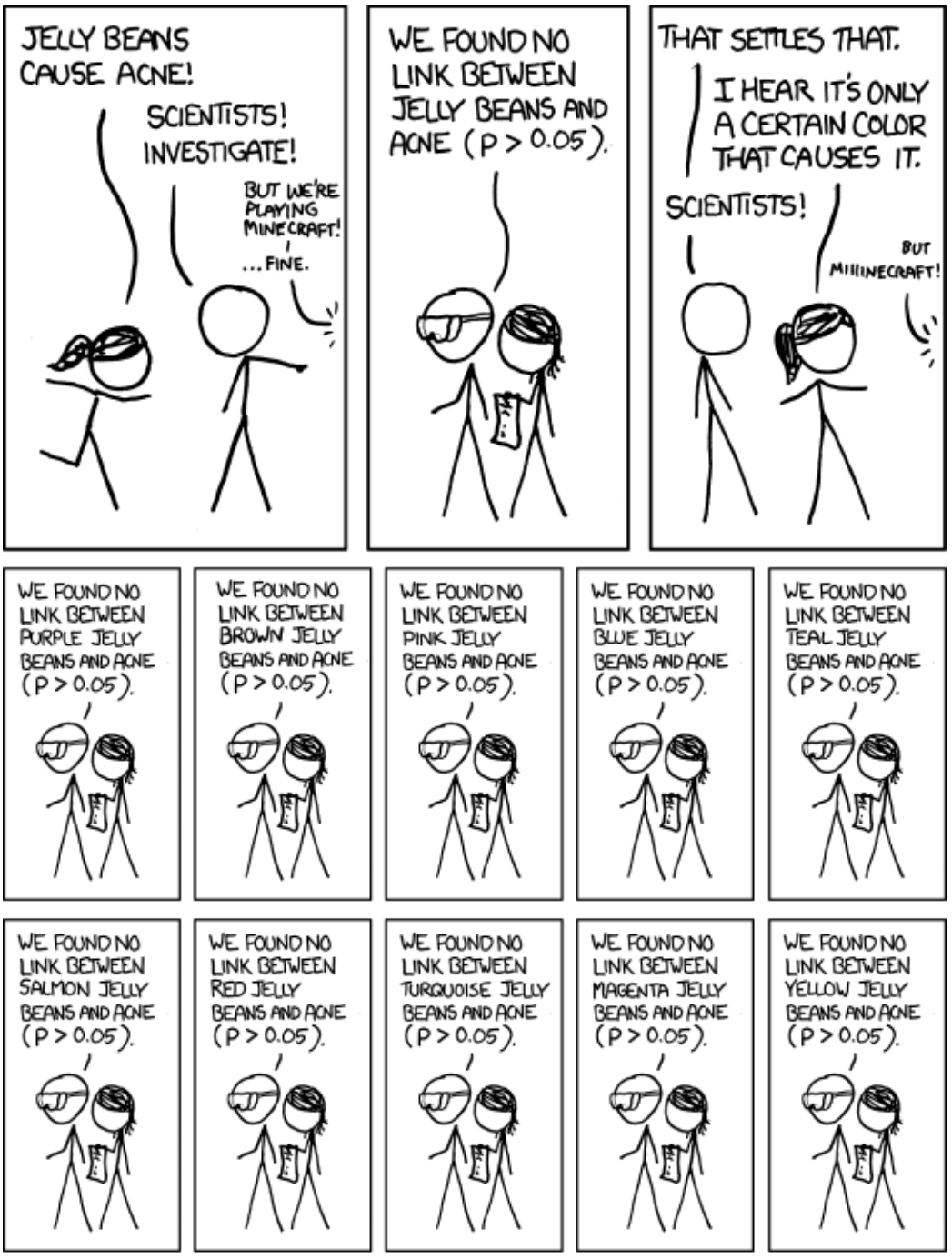
	hommes pour	hommes contre
femmes pour	73	33
femmes contre	14	817

(a) Que remarquez-vous dans ces données ?

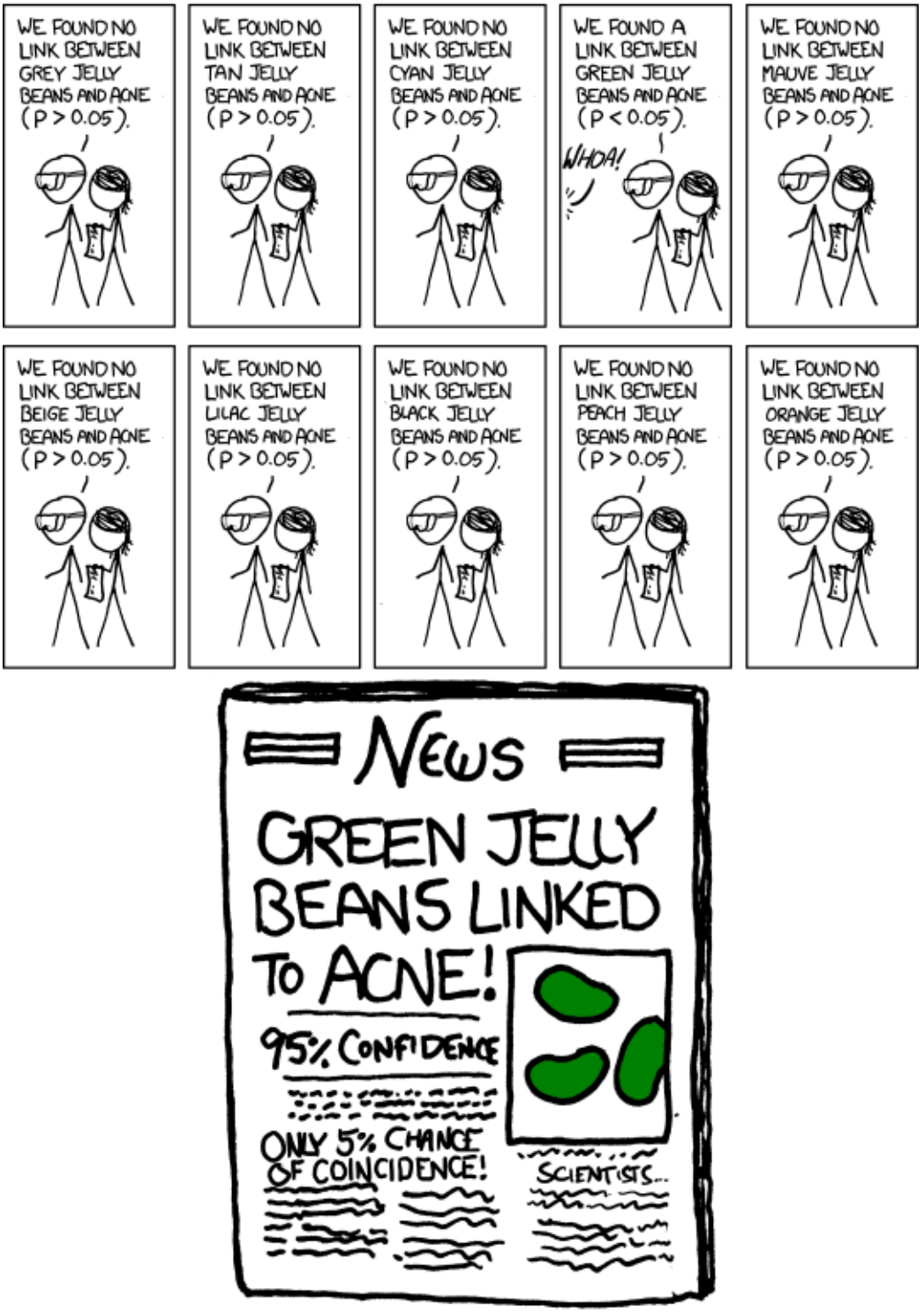
(b) Ces données permettent-elles de mettre en évidence un comportement différent des hommes et des femmes face au covoiturage au sein d'un couple ?

(c) Que vaut la p-valeur du test ?

Exercice 7. Quel message peut-on tirer de ce dessin humoristique ?



Première partie (source : http://xkcd.com/882/)



Deuxième partie (source : http://xkcd.com/882/)

Exercices supplémentaires

Exercice 8.

1. La mortalité infantile est naturellement élevée chez les blaireaux. On note $p_{\text{pop},1}$ la proportion de nouveaux-nés qui meurent avant d'avoir atteint un an dans une région où le blaireau n'est pas spécialement chassé. Sur un échantillon de 120 nouveaux-nés, on en a dénombré

54 qui sont morts dans leur première année. Donner un intervalle de confiance pour $p_{\text{pop},1}$ au niveau de confiance 95%.

2. Les principaux ennemis du blaireau dans certaines régions sont les chasseurs et agriculteurs qui détruisent leurs terriers. On note $p_{\text{pop},2}$ la proportion de nouveaux-nés qui meurent avant d'avoir atteint un an dans une région où la destruction des terriers de blaireaux est fréquente. Sur un échantillon de 100 nouveaux-nés, on en a dénombré 65 qui sont morts dans leur première année.
   (a) En utilisant un test statistique, dites si ces résultats ont mis en évidence l'influence néfaste de la destruction de terriers sur la mortalité infantile des blaireaux.
   (b) Que vaut la p-valeur de votre test ?

Exercice 9 (examen 2014)**.** Un fournisseur d'accès à Internet indique dans son contrat d'abonnement que 50% des problèmes techniques sont résolus en moins d'une demie-journée, que 35% sont résolus au bout de 24h, et que seulement 15% nécessitent au moins 2 jours d'attente.

Une enquête est menée par une association de consommateurs. Parmi les 863 clients interrogés ayant déjà rencontré un problème technique, 401 clients ont vu leur problème résolu en moins d'une demie-journée, 326 au bout de 24h, et 136 ont du attendre au moins 2 jours.

1. Les résultats de l'enquête sont-il en accord avec les affirmations du fournisseur d'accès à Internet ? (Précisez bien la démarche et les éventuelles conditions d'application.)
2. Que vaut la p-valeur de votre test ?

Exercice 10 (examen 2014)**.** Une équipe de biologistes cherche à comprendre l'effet de la malnutrition sur les rats *Sprague-Dawley*. Une expérience est alors menée sur deux groupes de rats : dans le premier groupe (10 rats), chaque rat est isolé et nourri normalement ; dans le deuxième groupe (15 rats), chaque rat est isolé et nourri un jour sur deux seulement. Les poids des rats (en grammes) sont mesurés après une semaine :

groupe de rats	poids moyen	écart-type corrigé
nourris normalement	347.2	27.279
nourris un jour sur deux	328.5	31.911

1. Quelles sont les sources d'aléatoire dans cette expérience ?
2. Les données expérimentales permettent-elles d'affirmer que la malnutrition a un effet négatif sur le poids des rats ? (Précisez bien la démarche et les éventuelles conditions d'application.)
3. Si les données ci-dessus avaient été obtenues avec des échantillons de 39 rats (nourris normalement) et 46 rats (nourris un jour sur deux), un calcul montre qu'on obtiendrait des conclusions différentes. Pourquoi est-ce logique ?

Exercice 11 (examen 2014)**.** Une enquête de consommation alimentaire est effectuée en France auprès de 1048 adultes âgés de 20 à 50 ans. Voici quelques résultats obtenus :

	aime le thé	n'aime pas le thé
aime le café	580	270
n'aime pas le café	140	58

1. Une société commercialisant du thé en sachets affirme qu'en France, les proportions d'amateurs de café et de thé sont identiques. Les résultats de l'enquête permettent-ils de contredire l'affirmation de cette société ? (Précisez bien la démarche.)
2. Que pouvez-vous dire sur la p-valeur de votre test ?

A Résolution de l'exercice 4 avec le logiciel R

Voici un script R permettant d'effectuer les calculs nécessaires à la résolution de l'exercice 4.

```
## Exercice 4

# Q1 : Test de comparaison de deux moyennes, échantillons appariés,
       n petit, cas unilatéral (t-test)
avant=c(74,86,98,102,78,84,79,70)
apres=c(70,85,90,110,71,80,69,74)
muech1 = mean(avant) # = 83.875
muech2 = mean(apres) # = 81.125
avant-apres
# 4  1  8 -8  7  4 10 -4
sigmaech12 = sqrt(mean((avant-apres)^2)-(muech1-muech2)^2) # = 5.760859
sech12 = sqrt(8/7)*sigmaech12 # 6.158618
T = (muech1-muech2)/(sech12/sqrt(8)) # = 1.262974
qt(p=0.95,df=7) # = 1.894579

# On obtient T < 1.895 donc on conserve H0 : pas réduction significative
  pour affirmer qu'un mois de jogging entraîne une réduction
  de rythme cardiaque en moyenne sur la population.

# Q2 : loi(X-Y) = courbe de Gauss -> vérification graphique (histogramme)
       ou quantitative (test de normalité de Shapiro-Wilk, qui est
       malheureusement très conservatif avec n=8)

# Q3 : 6 diminutions, pas de contradiction car ces diminutions ne sont
       pas assez grandes pour affirmer qu'elles ne sont pas simplement
       dues au hasard résultant de l'échantillonnage.
```

Conclusion

L'objet de ce cours était de présenter un des enjeux principaux de l'inférence statistique : comment, à partir de l'observation d'un échantillon d'une population, en déduire des propriétés de la population toute entière ? En sciences expérimentales, où les observations sont très souvent entâchées d'incertitudes, les techniques statistiques vues en cours (intervalle de confiance, tests) permettent de prendre en compte ces incertitudes de façon rigoureuse, et donc de conclure avec précaution ou au contraire avec grande confiance, selon les données statistiques observées.

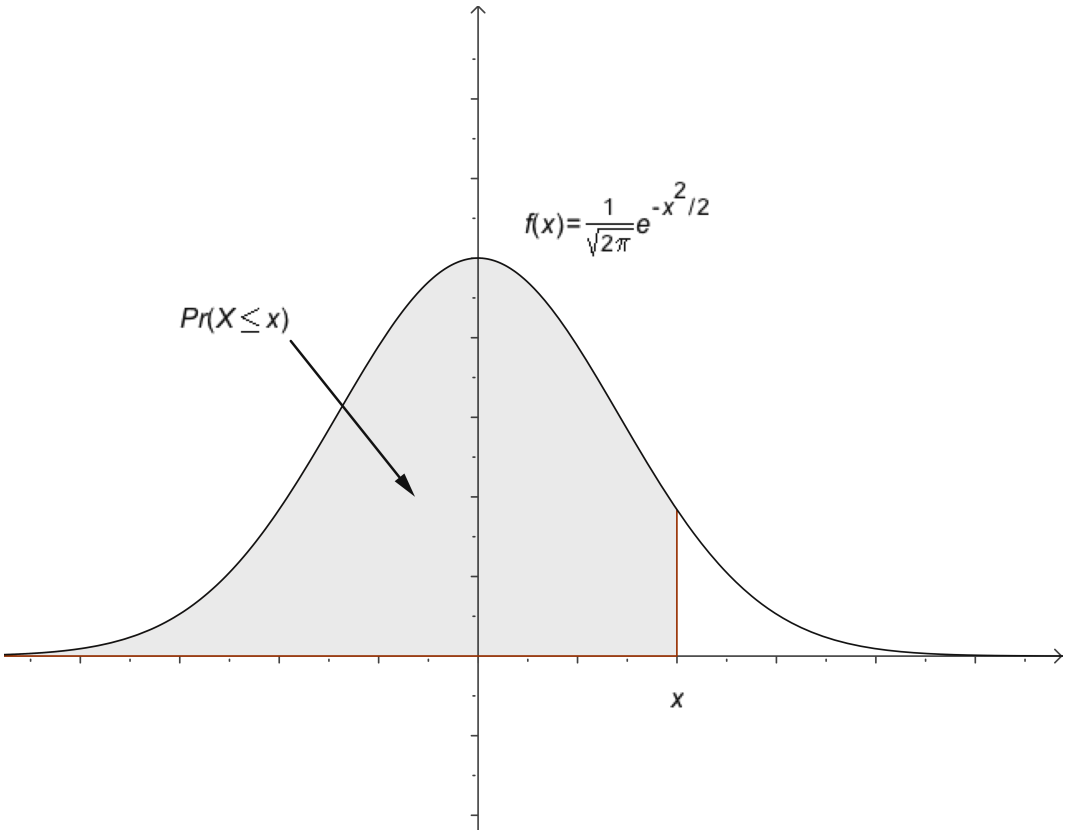
Ce cours avait pour vocation de vous *sensibiliser* à l'importance de traiter les incertitudes via des techniques statistiques, pour vous aider par exemple, à l'avenir, à mener des recherches scientifiques rigoureuses, à conduire une étude pharmaceutique sérieuse, ou tout simplement à interpréter les sondages politiques avec précaution. Ce cours introductif n'a nullement la prétention d'être exhaustif. Nul ne vous empêche cependant d'aller demander plus tard de l'aide à un statisticien ou un biostatisticien si vous rencontrez un problème statistique plus complexe !

Fonction de répartition de la loi normale centrée réduite

On suppose que X suit une loi normale centrée réduite $\mathcal{N}(0;1)$.
La fonction de répartition de X est la fonction $F : \mathbb{R} \longrightarrow \mathbb{R}$ donnée par

$$F(x) = Pr(X \leq x) = \int_{-\infty}^{x} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt$$

Pour tout réel x, le nombre $F(x)$ est l'aire de la partie représentée sur le graphique :



Remarque : Noter que pour des raisons de symétrie, on a la relation :

$$\begin{aligned} Pr(X \leq x) &= 1 - Pr(X \geq x) \\ &= 1 - Pr(X \leq -x)\,. \end{aligned}$$

Le tableau suivant donne des valeurs approximatives de la fonction de répartition de X pour x entre 0 et 2.99.

x	$0,00$	$0,01$	$0,02$	$0,03$	$0,04$	$0,05$	$0,06$	$0,07$	$0,08$	$0,09$
$0,0$	$0,5000$	$0,5040$	$0,5080$	$0,5120$	$0,5160$	$0,5199$	$0,5239$	$0,5279$	$0,5319$	$0,5359$
$0,1$	$0,5398$	$0,5438$	$0,5478$	$0,5717$	$0,5557$	$0,5596$	$0,5636$	$0,5675$	$0,5714$	$0,5753$
$0,2$	$0,5793$	$0,5832$	$0,5871$	$0,5910$	$0,5948$	$0,5987$	$0,6026$	$0,6064$	$0,6103$	$0,6141$
$0,3$	$0,6179$	$0,6217$	$0,6255$	$0,6293$	$0,6331$	$0,6368$	$0,6406$	$0,6443$	$0,6480$	$0,6517$
$0,4$	$0,6554$	$0,6591$	$0,6628$	$0,6664$	$0,6700$	$0,6736$	$0,6772$	$0,6808$	$0,6844$	$0,6879$
$0,5$	$0,6915$	$0,6950$	$0,6985$	$0,7019$	$0,7054$	$0,7088$	$0,7123$	$0,7157$	$0,7190$	$0,7224$
$0,6$	$0,7257$	$0,7291$	$0,7324$	$0,7357$	$0,7389$	$0,7422$	$0,7454$	$0,7486$	$0,7517$	$0,7549$
$0,7$	$0,7580$	$0,7611$	$0,7642$	$0,7673$	$0,7704$	$0,7734$	$0,7764$	$0,7794$	$0,7823$	$0,7852$
$0,8$	$0,7881$	$0,7910$	$0,7939$	$0,7967$	$0,7995$	$0,8023$	$0,8051$	$0,8078$	$0,8106$	$0,8133$
$0,9$	$0,8159$	$0,8186$	$0,8212$	$0,8238$	$0,8264$	$0,8289$	$0,8315$	$0,8340$	$0,8365$	$0,8389$
$1,0$	$0,8413$	$0,8438$	$0,8461$	$0,8485$	$0,8508$	$0,8531$	$0,8554$	$0,8577$	$0,8599$	$0,8621$
$1,1$	$0,8643$	$0,8665$	$0,8686$	$0,8708$	$0,8729$	$0,8749$	$0,8770$	$0,8790$	$0,8810$	$0,8830$
$1,2$	$0,8849$	$0,8869$	$0,8888$	$0,8907$	$0,8925$	$0,8944$	$0,8962$	$0,8980$	$0,8997$	$0,9015$
$1,3$	$0,9032$	$0,9049$	$0,9066$	$0,9082$	$0,9099$	$0,9115$	$0,9131$	$0,9147$	$0,9162$	$0,9177$
$1,4$	$0,9192$	$0,9207$	$0,9222$	$0,9236$	$0,9251$	$0,9265$	$0,9279$	$0,9292$	$0,9306$	$0,9319$
$1,5$	$0,9332$	$0,9345$	$0,9357$	$0,9370$	$0,9382$	$0,9394$	$0,9406$	$0,9418$	$0,9429$	$0,9441$
$1,6$	$0,9452$	$0,9463$	$0,9474$	$0,9484$	$0,9495$	$0,9505$	$0,9515$	$0,9525$	$0,9535$	$0,9545$
$1,7$	$0,9554$	$0,9564$	$0,9573$	$0,9582$	$0,9591$	$0,9599$	$0,9608$	$0,9616$	$0,9625$	$0,9633$
$1,8$	$0,9641$	$0,9649$	$0,9656$	$0,9664$	$0,9671$	$0,9678$	$0,9686$	$0,9693$	$0,9699$	$0,9706$
$1,9$	$0,9713$	$0,9719$	$0,9726$	$0,9732$	$0,9738$	$0,9744$	$0,9750$	$0,9756$	$0,9761$	$0,9767$
$2,0$	$0,9772$	$0,9778$	$0,9783$	$0,9788$	$0,9793$	$0,9798$	$0,9803$	$0,9808$	$0,9812$	$0,9817$
$2,1$	$0,9821$	$0,9826$	$0,9830$	$0,9834$	$0,9838$	$0,9842$	$0,9846$	$0,9850$	$0,9854$	$0,9857$
$2,2$	$0,9861$	$0,9864$	$0,9868$	$0,9871$	$0,9875$	$0,9878$	$0,9881$	$0,9884$	$0,9887$	$0,9890$
$2,3$	$0,9893$	$0,9896$	$0,9898$	$0,9901$	$0,9904$	$0,9906$	$0,9909$	$0,9911$	$0,9913$	$0,9916$
$2,4$	$0,9918$	$0,9920$	$0,9922$	$0,9925$	$0,9927$	$0,9929$	$0,9931$	$0,9932$	$0,9934$	$0,9936$
$2,5$	$0,9938$	$0,9940$	$0,9941$	$0,9943$	$0,9945$	$0,9946$	$0,9948$	$0,9949$	$0,9951$	$0,9952$
$2,6$	$0,9953$	$0,9955$	$0,9956$	$0,9957$	$0,9959$	$0,9960$	$0,9961$	$0,9962$	$0,9963$	$0,9964$
$2,7$	$0,9965$	$0,9966$	$0,9967$	$0,9968$	$0,9969$	$0,9970$	$0,9971$	$0,9972$	$0,9973$	$0,9974$
$2,8$	$0,9974$	$0,9975$	$0,9976$	$0,9977$	$0,9977$	$0,9978$	$0,9979$	$0,9979$	$0,9980$	$0,9981$
$2,9$	$0,9981$	$0,9982$	$0,9982$	$0,9983$	$0,9984$	$0,9984$	$0,9985$	$0,9985$	$0,9986$	$0,9986$

Comment lire cette table numérique ?

On écrit le développement décimal de x avec 2 chiffres après la virgule : $x = a,bc$ (où a, b et c sont des entiers entre 0 et 9). Le début du développement a,b se lit sur la bordure verticale du tableau et la fin $0,0c$ se lit sur la bordure horizontale. Le nombre qui se trouve à l'intersection de la ligne de a,b et de la colonne de $0,0c$ est approximativement $Pr(X \leq a,bc)$.

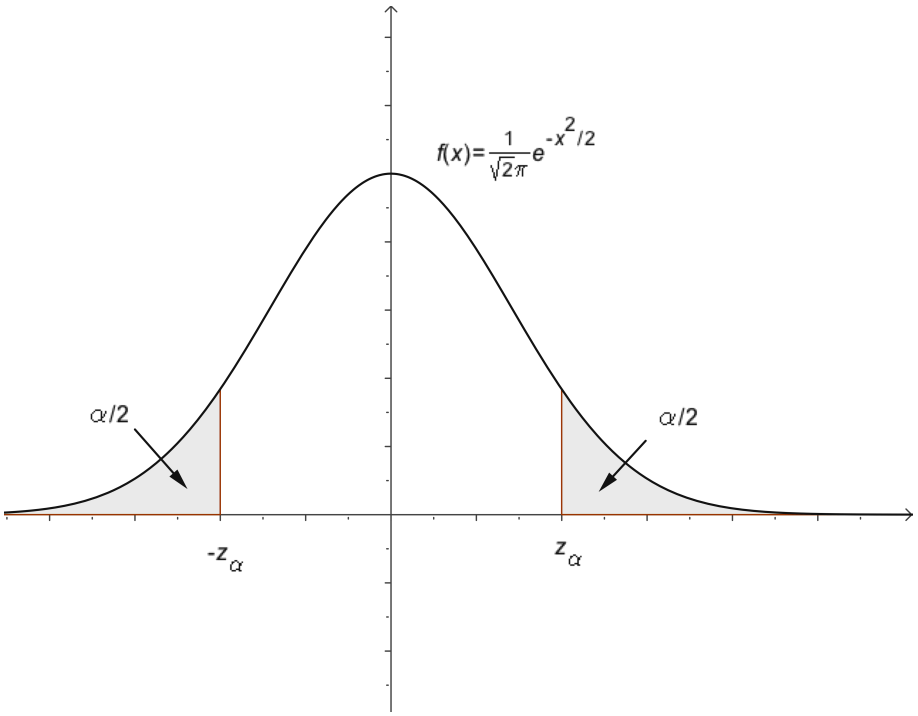
Exemple : $Pr(X \leq 1,24) \approx 0,8925$.

Valeurs extrêmes de la loi normale centrée réduite

On suppose que X suit une loi normale centrée réduite $\mathcal{N}(0;1)$.
Dans les tableaux numériques suivants, la correspondance entre α et z_α est

$$Pr(|X| > z_\alpha) \approx \alpha$$

Cette correspondance peut se visualiser sur le graphe :



La symétrie de la fonction densité de X permet d'écrire

$$Pr(X > z) = \frac{1}{2} Pr(|X| > z)$$

α	z_α
$0,000001$	$5,066$
$0,00001$	$4,414$
$0,0001$	$3,891$
$0,001$	$3,290$
$0,01$	$2,576$
$0,02$	$2,326$
$0,03$	$2,170$
$0,04$	$2,054$
$0,05$	$1,960$
$0,06$	$1,881$

α	z_α
$0,07$	$1,812$
$0,08$	$1,751$
$0,09$	$1,695$
$0,10$	$1,645$
$0,11$	$1,598$
$0,12$	$1,555$
$0,13$	$1,514$
$0,14$	$1,476$
$0,15$	$1,440$
$0,20$	$1,282$

α	z_α
$0,25$	$1,150$
$0,30$	$1,036$
$0,35$	$0,935$
$0,40$	$0,842$
$0,45$	$0,755$
$0,50$	$0,674$
$0,55$	$0,598$
$0,60$	$0,524$
$0,65$	$0,454$
$0,70$	$0,385$

α	z_α
$0,75$	$0,319$
$0,80$	$0,253$
$0,85$	$0,189$
$0,90$	$0,126$
$0,95$	$0,063$
$0,96$	$0,050$
$0,97$	$0,038$
$0,98$	$0,025$
$0,99$	$0,013$
$0,999$	$0,001$

Exemple : on a $Pr(|X| > 1,645) \approx 0,10$.

On peut éventuellement utiliser cette table pour donner des valeurs de la fonction de répartition de X :

$$\begin{aligned} Pr(X \leq 2,054) &= 1 - Pr(X > 2,054) \\ &= 1 - \frac{1}{2} Pr(|X| > 2,054) \quad \text{par symétrie} \\ &= 1 - 0,04/2 \\ &= 0,98 \end{aligned}$$

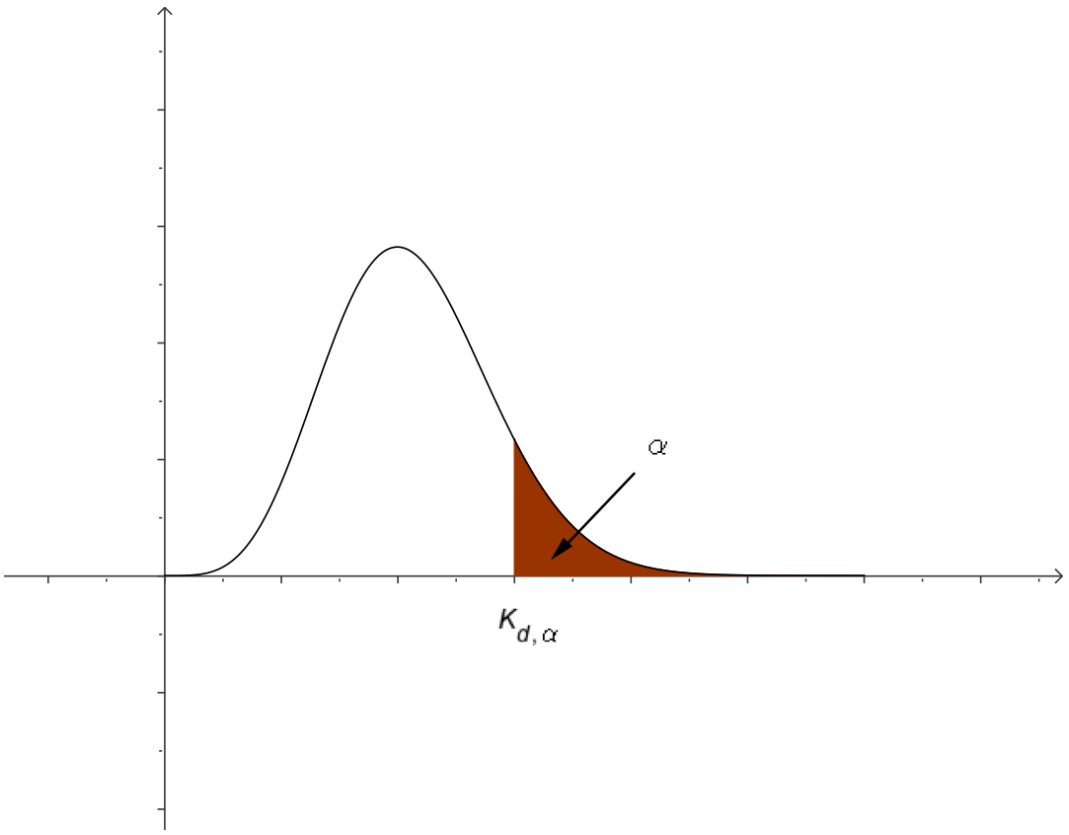
Valeurs extrêmes de la loi du χ^2

On suppose que X suit une loi du χ^2 à d degrés de liberté.

Dans le tableau numérique suivant, le degré de liberté se lit dans la colonne de gauche, la probabilité α se lit sur la première ligne et la valeur $K_{d,\alpha}$ se lit au milieu du tableau. Le lien entre ces trois nombres est

$$Pr(X > K_{d,\alpha}) \approx \alpha$$

Cette correspondance peut se visualiser sur le graphe suivant, où on a représenté la fonction densité de X :



α / d	0,975	0,90	0,50	0,05	0,04	0,03	0,025	0,02	0,01	0,001	0,0001
1	0,001	0,0158	0,455	3,841	4,218	4,709	5,024	5,412	6,635	10,827	15,137
2	0,051	0,211	1,386	5,991	6,438	7,013	7,378	7,824	9,210	13,815	18,421
3	0,216	0,584	2,366	7,815	8,311	8,947	9,348	9,837	11,345	16,266	21,108
4	0,484	1,064	3,357	9,488	10,026	10,712	11,143	11,668	13,277	18,467	23,513
5	0,831	1,610	4,351	11,070	11,644	12,375	12,833	13,388	15,086	20,515	25,745
6	1,237	2,204	5,348	12,592	13,198	13,968	14,449	15,033	16,812	22,457	27,856
7	1,690	2,833	6,346	14,067	14,703	15,509	16,013	16,622	18,475	24,322	29,878
8	2,180	3,490	7,344	15,507	16,171	17,010	17,535	18,168	20,090	26,125	31,828
9	2,700	4,168	8,343	16,919	17,608	18,480	19,028	19,679	21,666	27,877	33,720
10	3,247	4,865	9,342	18,307	19,021	19,922	20,483	21,161	23,209	29,588	35,564
11	3,816	5,578	10,341	19,675	20,412	21,342	21,920	22,618	24,725	31,264	37,367
12	4,404	6,304	11,340	21,026	21,785	22,742	23,337	24,054	26,217	32,909	39,134
13	5,009	7,042	12,340	22,362	23,142	24,125	24,736	25,472	27,688	34,528	40,871
14	5,629	7,790	13,339	23,685	24,485	25,493	26,119	26,873	29,141	36,123	42,579
15	6,262	8,547	14,339	24,996	25,816	26,848	27,488	28,259	30,578	37,697	44,263
16	6,908	9,312	15,338	26,296	27,136	28,191	28,845	29,633	32,000	39,252	45,925
17	7,564	10,085	16,338	27,587	28,445	29,523	30,191	30,995	33,409	40,790	47,566
18	8,231	10,865	17,338	28,869	29,745	30,845	31,526	32,346	34,805	42,312	49,189
19	8,907	11,651	18,338	30,144	31,037	32,158	32,852	33,687	36,191	43,820	50,795
20	9,591	12,443	19,337	31,410	32,321	33,462	34,170	35,020	37,566	45,315	52,386
21	10,283	13,240	20,337	32,671	33,597	34,759	35,479	36,343	38,932	46,797	53,962
22	10,982	14,041	21,337	33,924	34,867	36,049	36,781	37,659	40,289	48,268	55,525
23	11,689	14,848	22,337	35,172	36,131	37,332	38,076	38,968	41,638	49,728	57,075
24	12,401	15,659	23,337	36,415	37,389	38,609	39,364	40,270	42,980	51,179	58,613
25	13,120	16,473	24,337	37,652	38,642	39,880	40,646	41,566	44,314	52,620	60,140
26	13,844	17,292	25,336	38,885	39,889	41,146	41,923	42,856	45,642	54,052	61,657
27	14,573	18,114	26,336	40,113	41,132	42,407	43,194	44,140	46,963	55,476	63,164
28	15,308	18,939	27,336	41,337	42,370	43,662	44,461	45,419	48,278	56,893	64,662
29	16,047	19,768	28,336	42,557	43,604	44,913	45,722	46,693	49,588	58,302	66,152
30	16,791	20,599	29,336	43,773	44,834	46,160	46,979	47,962	50,892	59,703	67,633

Exemples :
Pour $d = 9$ et $\alpha = 0,05$ on a $K_{9;\,0,05} = 16,919$.
Pour $d = 23$ et $\alpha = 0,001$ on a $K_{23;\,0,001} = 49,728$.

On peut également utiliser cette table pour donner des valeurs de la fonction de répartition de X :
Si $d = 13$, on a

$$\begin{aligned} Pr(X \leq 7,042) &= 1 - Pr(X > 7,042) \\ &= 1 - 0,90 \\ &= 0,10 \end{aligned}$$

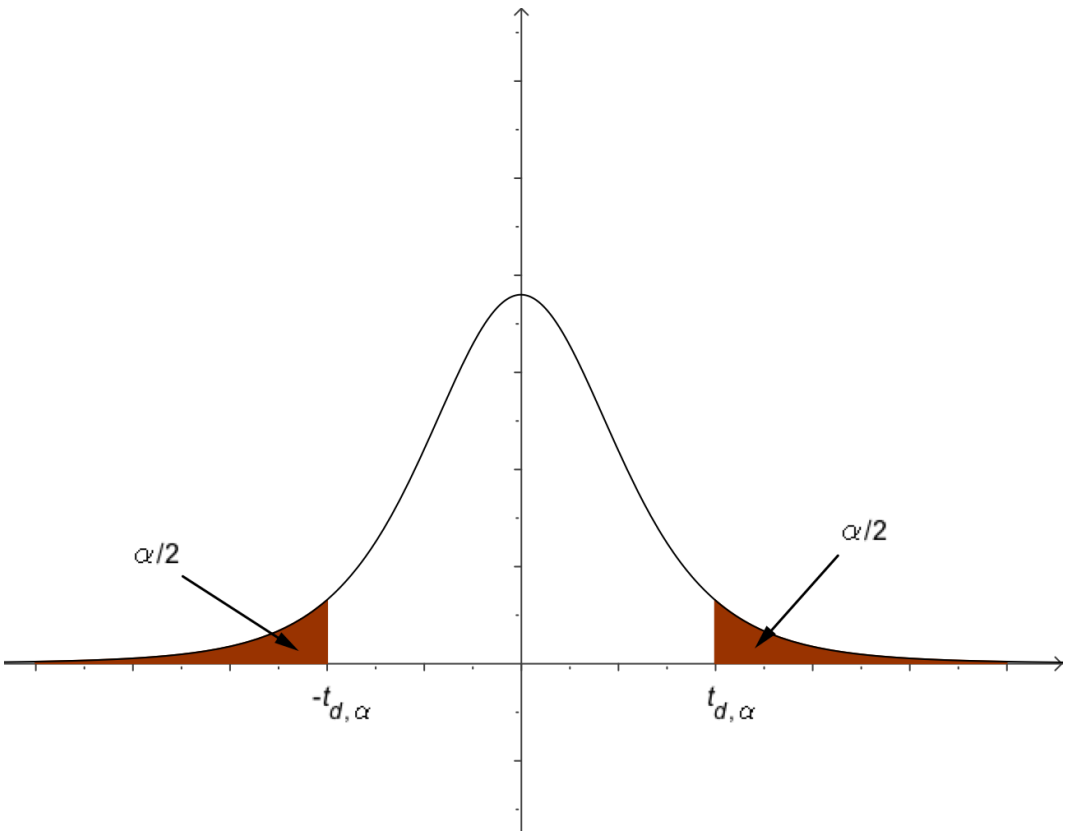
Valeurs extrêmes de la loi du Student

On suppose que X suit une loi du Student à d degrés de liberté.

Dans le tableau numérique suivant, le degré de liberté se lit dans la colonne de gauche, la probabilité α se lit sur la première ligne et la valeur $t_{d,\alpha}$ se lit au milieu du tableau. Le lien entre ces trois nombres est

$$Pr(|X| > t_{d,\alpha}) \approx \alpha$$

Cette correspondance peut se visualiser sur le graphe suivant, où on a représenté la fonction densité de X :



α / d	$0,50$	$0,30$	$0,20$	$0,10$	$0,05$	$0,04$	$0,03$	$0,02$	$0,01$	$0,001$	$0,0001$
1	$1,000$	$1,963$	$3,078$	$6,314$	$12,706$	$15,895$	$21,205$	$31,821$	$63,657$	$636,619$	6366.198
2	$0,816$	$1,386$	$1,886$	$2,920$	$4,303$	$4,849$	$5,643$	$6,965$	$9,925$	$31,598$	99.993
3	$0,765$	$1,250$	$1,638$	$2,353$	$3,182$	$3,482$	$3,896$	$4,541$	$5,841$	$12,924$	28.000
4	$0,741$	$1,190$	$1,533$	$2,132$	$2,776$	$2,999$	$3,298$	$3,747$	$4,604$	$8,610$	15.544
5	$0,727$	$1,156$	$1,476$	$2,015$	$2,571$	$2,757$	$3,003$	$3,365$	$4,032$	$6,869$	11.178
6	$0,718$	$1,134$	$1,440$	$1,943$	$2,447$	$2,612$	$2,829$	$3,143$	$3,707$	$5,959$	9.082
7	$0,711$	$1,119$	$1,415$	$1,895$	$2,365$	$2,517$	$2,715$	$2,998$	$3,499$	$5,408$	7.885
8	$0,706$	$1,108$	$1,397$	$1,860$	$2,306$	$2,449$	$2,634$	$2,896$	$3,355$	$5,041$	7.120
9	$0,703$	$1,100$	$1,383$	$1,833$	$2,262$	$2,398$	$2,574$	$2,821$	$3,250$	$4,781$	6.594
10	$0,700$	$1,093$	$1,372$	$1,812$	$2,228$	$2,359$	$2,527$	$2,764$	$3,169$	$4,587$	6.211
11	$0,697$	$1,088$	$1,363$	$1,796$	$2,201$	$2,328$	$2,491$	$2,718$	$3,106$	$4,437$	5.921
12	$0,695$	$1,083$	$1,356$	$1,782$	$2,179$	$2,303$	$2,461$	$2,681$	$3,055$	$4,318$	5.694
13	$0,694$	$1,079$	$1,350$	$1,771$	$2,160$	$2,282$	$2,436$	$2,650$	$3,012$	$4,221$	5.513
14	$0,692$	$1,076$	$1,345$	$1,761$	$2,145$	$2,264$	$2,415$	$2,624$	$2,977$	$4,140$	5.363
15	$0,691$	$1,074$	$1,341$	$1,753$	$2,131$	$2,249$	$2,397$	$2,602$	$2,947$	$4,073$	5.239
16	$0,690$	$1,071$	$1,337$	$1,746$	$2,120$	$2,235$	$2,382$	$2,583$	$2,921$	$4,015$	5.134
17	$0,689$	$1,069$	$1,333$	$1,740$	$2,110$	$2,224$	$2,368$	$2,567$	$2,898$	$3,965$	5.044
18	$0,688$	$1,067$	$1,330$	$1,734$	$2,101$	$2,214$	$2,356$	$2,552$	$2,878$	$3,922$	4.966
19	$0,688$	$1,066$	$1,328$	$1,729$	$2,093$	$2,205$	$2,346$	$2,539$	$2,861$	$3,883$	4.897
20	$0,687$	$1,064$	$1,325$	$1,725$	$2,086$	$2,197$	$2,336$	$2,528$	$2,845$	$3,850$	4.837
21	$0,686$	$1,063$	$1,323$	$1,721$	$2,080$	$2,189$	$2,328$	$2,518$	$2,831$	$3,819$	4.784
22	$0,686$	$1,061$	$1,321$	$1,717$	$2,074$	$2,183$	$2,320$	$2,508$	$2,819$	$3,792$	4.736
23	$0,685$	$1,060$	$1,319$	$1,714$	$2,069$	$2,177$	$2,313$	$2,500$	$2,807$	$3,767$	4.693
24	$0,685$	$1,059$	$1,318$	$1,711$	$2,064$	$2,172$	$2,307$	$2,492$	$2,797$	$3,745$	4.654
25	$0,684$	$1,058$	$1,316$	$1,708$	$2,060$	$2,167$	$2,301$	$2,485$	$2,787$	$3,725$	4.619
26	$0,684$	$1,058$	$1,315$	$1,706$	$2,056$	$2,162$	$2,296$	$2,479$	$2,779$	$3,707$	4.587
27	$0,684$	$1,057$	$1,314$	$1,703$	$2,052$	$2,158$	$2,291$	$2,473$	$2,771$	$3,690$	4.558
28	$0,683$	$1,056$	$1,313$	$1,701$	$2,048$	$2,154$	$2,286$	$2,467$	$2,763$	$3,674$	4.530
29	$0,683$	$1,055$	$1,311$	$1,699$	$2,045$	$2,150$	$2,282$	$2,462$	$2,756$	$3,659$	4.506
30	$0,683$	$1,055$	$1,310$	$1,697$	$2,042$	$2,147$	$2,278$	$2,457$	$2,750$	$3,646$	4.482

Exemples :
Pour $d = 6$ et $\alpha = 0,05$, on a $t_{6;\,0,05} = 2,447$.
Pour $d = 19$ et $\alpha = 0,01$, on a $t_{19;\,0,01} = 2,861$.

On peut également utiliser cette table pour donner des valeurs de la fonction de répartition de X :
Si $d = 11$, on a

$$\begin{aligned} Pr(X \leq 1,796) &= 1 - Pr(X > 1,796) \\ &= 1 - \frac{1}{2} Pr(|X| > 1,796) \quad \text{par symétrie} \\ &= 1 - 0,10/2 \\ &= 0,95 \end{aligned}$$