Biostatistique

Marie Laure Delignette-Muller

12 novembre 2024

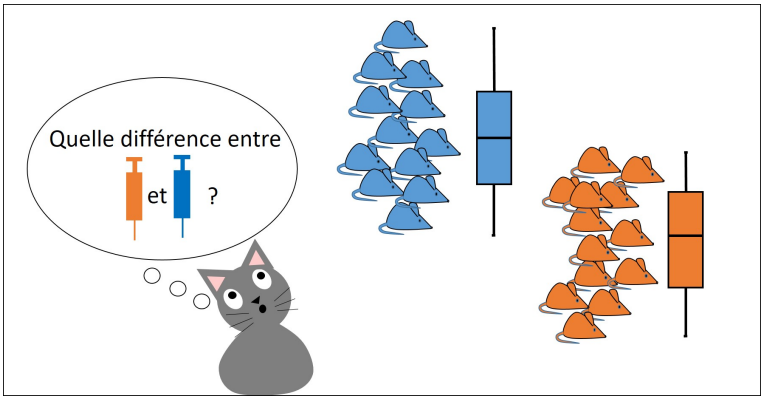


Table des matières

1 Introduction

"La statistique est une méthode et non une science. Elle est une technique. Tantôt elle permet de préciser certains phénomènes, tantôt elle donne des suggestions. Son rôle suggestif est particulièrement important. Mais on doit aussi se méfier d'elle."

Henri Berr (1935)

"Dans notre domaine les méthodes statistiques, qui peuvent être fécondes, peuvent facilement aussi être dangereuses. Le biologiste doit avant tout rester biologiste, c'est-à-dire poser exactement les éléments biologiques d'un problème que l'on veut traiter par la méthode statistique. L'appareil mathématique a ses dangers, car souvent il est artificiel en biologie. Les solutions qui résultent de sa mise en jeu découlent automatiquement des mises en équations initiales. Si ces dernières correspondent vraiment à la réalité, la solution est adéquate à celle-ci : s'il en est autrement les conclusions auxquelles on aboutit ne concernent pas le problème réel et donnent une idée fausse des choses."

Maurice Caullery (1935)

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."

Ronald Aylmer Fisher (1938)

Après ces trois citations et avant de préciser les objectifs pédagogiques de cet enseignement de biostatistique de base, partons de trois constats :

— la méthode statistique est de plus en plus couramment utilisée dans les travaux publiés en médecine vétérinaire et plus largement en biologie,

— rien n'est plus facile que de tirer des conclusions erronées à partir d'une mauvaise utilisation de la méthode statistique ou d'une mauvaise interprétation de résultats statistiques et

— il est indispensable de réfléchir bien en amont d'une collecte de données à leur analyse statistique future et d'en discuter avec un statisticien si besoin, sans quoi les données risquent d'être peu utilisables / informatives.

L'objectif de cet enseignement est triple :

1. vous amener à une compréhension suffisante de la méthode statistique pour vous permettre d'exercer votre esprit critique à la lecture de publications scientifiques,
2. vous donner une maîtrise suffisante des outils de base et de leur cadre d'utilisation pour vous permettre si besoin d'analyser vous-mêmes des données biologiques dans les cas les plus simples et
3. vous donner un recul suffisant par rapport à l'utilisation de la méthode statistique pour savoir à quel moment il devient nécessaire de se former à des méthodes plus sophistiquées et/ou de consulter un statisticien.

? ? ? Dans ce polycopié vous trouverez des blocs encadrés en rouge comme celui-ci. Ils correspondent à des questions auxquelles je vous demande de tenter de répondre au cours de votre lecture ou à des petits exercices d'entrainement à réaliser pour s'assurer que vous avez bien compris un point clef.

Vous trouverez aussi des blocs sur fond jaune comme celui-ci. Ils soulignent des points importants, et visent à vous éviter des erreurs fréquentes.

```
# Enfin les blocs sur fond gris comme celui-ci correspondent à des sorties
# et/ou du code R (code informatique écrit en langage R auquel
# nous vous initierons en S4 pour analyser
# des données sur ordinateur).
```

2 La statistique descriptive

Nous aborderons dans ce chapitre essentiellement les méthodes de réduction et de représentation des données uniquement dans le cas univarié c'est-à-dire lorsqu'on s'intéresse à une seule variable observée. Les méthodes utilisées dans les cas bivariés seront vues au fur et à mesure des cas étudiés dans les chapitres suivants.

2.1 Objectifs pédagogiques

A l'issue de l'étude de ce chapitre vous devriez :

— savoir reconnaître le type d'une variable observée,
— savoir synthétiser et représenter graphiquement des données observées selon le type de la variable,
— être capable d'interpréter les représentations graphiques classiques (dans le cas univarié),
— savoir juger de la normalité d'une distribution à partir des représentations graphiques classiques,
— savoir calculer et interpréter les paramètres statistiques classiques et connaître leurs limites d'utilisation,
— savoir définir et calculer un intervalle de fluctuation (par ex. pour déterminer des valeurs usuelles).

2.2 Quelques définitions

2.2.1 Statistique descriptive et statistique inférentielle

La première étape d'une étude statistique est la collecte des données. Celle-ci implique généralement un échantillonnage (Figure 1) : seul un échantillon de la population étudié fait l'objet de cette collecte de données. L'analyse des données se déroule ensuite le plus souvent en deux étapes.

1. La première étape consiste à décrire les données obtenues sur l'échantillon, c'est-à-dire à les représenter graphiquement et éventuellement à les résumer afin d'en faciliter la prise de connaissance. Il s'agit là de la **statistique descriptive**.

2. La seconde étape consiste à tenter de tirer des conclusions sur la population étudiée à partir des résultats obtenus sur l'échantillon, c'est-à-dire à voir si l'on peut tirer des conclusions généralisables. Il s'agit là de la **statistique inférentielle**. Nous nous intéresserons dans ce premier chapitre uniquement à la première étape de description des données.

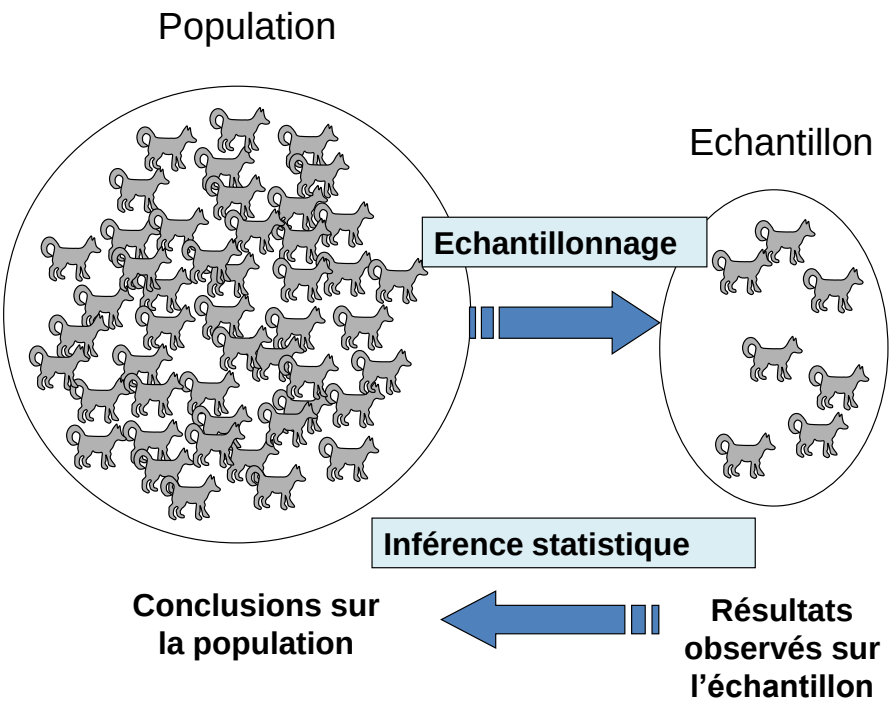


FIGURE 1 – Echantillonnage et statistique inférentielle.

2.2.2 Type de variable aléatoire étudiée

Une variable aléatoire est un caractère qui est susceptible de varier d'un animal à l'autre, d'une unité expérimentale à l'autre, d'un groupe à l'autre .... Ce caractère est parfois quantifiable ou parfois uniquement qualifiable.

Une **variable est dite qualitative** si elle est décrite par une famille de modalités possibles. Lorsque les modalités sont ordonnées, on parle de **variable qualitative ordinale** (ex. : variable « évolution de la maladie » avec comme modalités « aggravation de l'état », « état stationnaire » ou « amélioration de l'état »). Lorsque les modalités ne sont pas naturellement ordonnées, on parle de **variable qualitative nominale** (ex. : variable « couleur du poil de l'animal », variable « sexe de l'animal »).

Lorsque le caractère est quantifiable, on parle de **variable quantitative**. Lorsque ce caractère ne peut prendre qu'un nombre fini de valeurs (généralement entières), on parle de **variable quantitative discrète** (ex. : variable « nombre de chiots par portée », variable « nombre d'animaux domestiques par foyer »). Lorsque ce caractère peut en théorie prendre une infinité de valeurs prises dans une série continue de nombres réels, on parle de **variable quantitative continue** (ex. : variables « taille », « poids », « âge », « taux d'hémoglobine »). Lorsque l'on se demande si une variable quantitative est discrète ou

continue, on considère bien la nature même de cette variable et non la précision de sa mesure. Par exemple une variable âge est une variable par nature continue, même si les âges observés ont été arrondis à l'année.

Certaines variables sont dites **semi-quantitatives** soit parce qu'elles sont quantifiables uniquement sur une partie de leur domaine de définition (par exemple du fait d'une limite de quantification de la méthode de mesure, ce qui induit une censure d'un certains nombre de mesures : on sait qu'elles sont situées en dessous de la limite de quantification, mais on ne sait pas à quelle valeur exactement) soit parce qu'elles sont créées à partir d'observations qualitatives (comme un **score clinique** par exemple, qui sera souvent considéré soit comme une variable **quantitative discrète** soit comme une variable **qualitative ordinale**, en fonction notamment du nombre de valeurs que peut prendre le score).

Toute analyse de données nécessite au préalable de se poser la question de la nature (qualitative ou quantitative ?) de chaque variable étudiée. Or si la différence entre variable qualitative et quatitative semble assez triviale, il n'est pas rare que des erreurs soient commise lors de cette étape préliminaire. Afin d'éviter ces erreurs, la bonne question à se poser est : **quelle est la variable observée sur chaque unité d'observation** ?

Prenons quelques exemples pour illustrer la difficulté parfois rencontrée :

- Lors de l'étude du poids de chiots à la naissance, si l'unité d'observation est le chiot, la variable est le poids qui est une variable quantitative continue.
- Lors de l'étude du taux de mortalité des chiots à la naissance dans divers élevages, si l'unité d'observation est l'élevage, la variable observée est le taux de mortalité qui est une variable quantitative continue.
- Lors de l'étude du taux de mortalité liée à une pathologie donnée sur un groupe de malades, si l'unité d'observation est l'individu, la variable observée le statut de l'individu (mort / vivant) est une variable qualitative nominale.

Au vu des deux derniers exemples on comprend bien qu'il est nécessaire de bien réfléchir à la situation pour savoir si on travaille sur une variable qualitative ou quantitative et que le seul repérage de mots clefs (ex. taux de mortalité) ne permet pas de répondre à la question. Ce commentaire vaut pour bien des étapes d'une analyse statistique. Il est bien plus important de comprendre le cas étudié que de repérer des mots clefs.

2.3 La représentation graphique des données

La représentation graphique des données constitue une étape primordiale dans toute analyse de données, étape trop souvent omise ou négligée.

2.3.1 Quelques représentations classiques de la distribution d'une variable qualitative

Partons d'un exemple d'étude de la reproduction de chiens de race sur 423 élevages, extrait de la thèse vétérinaire de Mathilde Poinssot (Maisons Alfort, 2011). Une des variables étudiées était le **type de fécondation**, **variable qualitative nominale à trois modalités** : 1/ monte naturelle avec un mâle de l'élevage, 2/ monte naturelle avec un autre mâle ou 3/ insémination artificielle (cf. Table 1).

ELEVAGE	FECONDATION
elevage_1	autre_male
elevage_2	insemination
elevage_3	insemination
elevage_4	male_elevage
elevage_5	male_elevage
elevage_6	insemination
elevage_7	autre_male
elevage_8	insemination
elevage_9	insemination
elevage_10	male_elevage

TABLE 1 – Dix premières lignes du fichier des données brutes observées pour la variable "type de fécondation".

Il est assez naturel de résumer ces données brutes par

— les effectifs dans chacune des classes,

```
##
##   autre_male insemination male_elevage
##          197          102          124
```

— ou les fréquences correespondantes obtenues en divisant les effectifs par l'effectif total.

```
##
##   autre_male insemination male_elevage
##        0.466        0.241        0.293
```

Une représentation bien connue de la distribution en fréquences d'une variable qualitative est le diagramme en secteurs appelé communément « camembert » (cf. Figure 2 pour les données précédentes). Néanmoins cette représentation est peu recommandée par les statisticiens. Il convient au moins d'éviter les camemberts en trois dimensions, qui peuvent être très trompeurs (cf. Figure 3).

? ? ? Prenez le temps de bien analyser la figure 3 et demandez-vous pourquoi les scientifiques préfèrent généralement les diagrammes en bâton que les diagrammes en secteurs, et pourquoi le diagramme en secteurs en relief sont à éviter.

Dans les écrits scientifiques on aura donc tendance à privilégier des représentations de type diagrammes en bâtons, notamment pour les variables qualitatives ordinales (cf. Figure 4). Dans un diagramme en bâtons l'axe des ordonnées peut indifféremment être gradué en effectifs ou en fréquences.

Une représentation très proche du diagramme en secteur, mais qui reste sur une échelle linéaire plus simple à lire que l'échelle angulaire, est la représentation dite en barres ou encore en bandes. Cette représentation est souvent utilisée pour visualiser facilement plusieurs distributions sur un même graphe (cf. Figure 5).

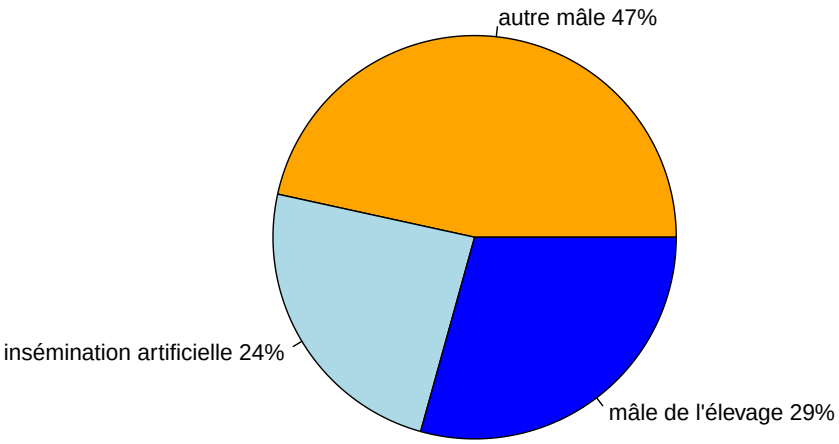


FIGURE 2 – Diagramme en secteurs représentant la ditribution du type de fécondation à partir d'une étude de la reproduction de chiens de race sur 423 élevages.

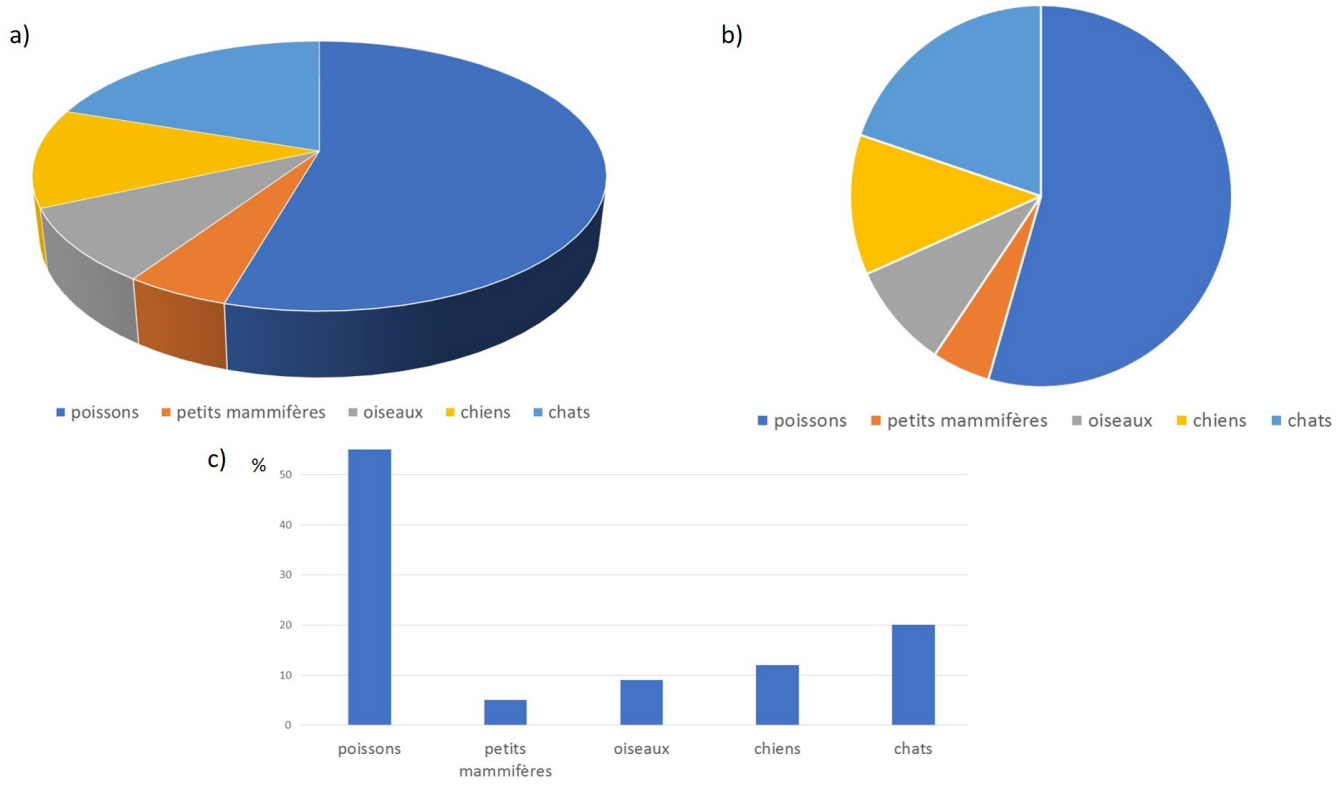


FIGURE 3 – Distribution des animaux domestiques en France en 2014 sous différentes formes, a) diagramme en secteurs en relief (représentation peu recommandable), b) diagramme en secteurs classique, c) diagramme en batons (la plus lisible).

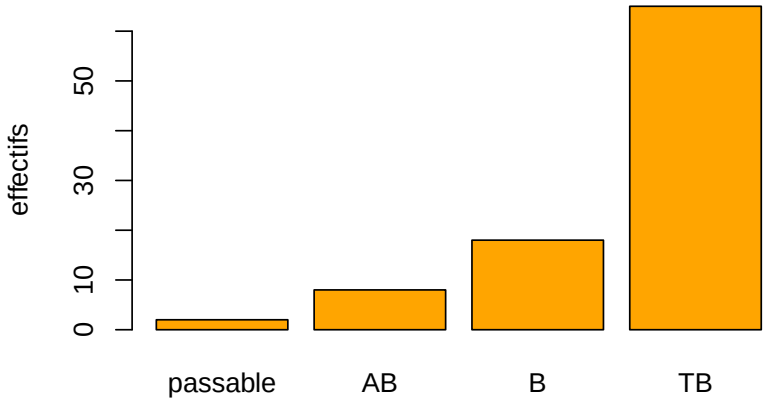


FIGURE 4 – Diagramme en bâtons représentant la distribution des mentions au baccalauréat des étudiants ayant intégré le campus vétérinaire de Lyon en 2016.

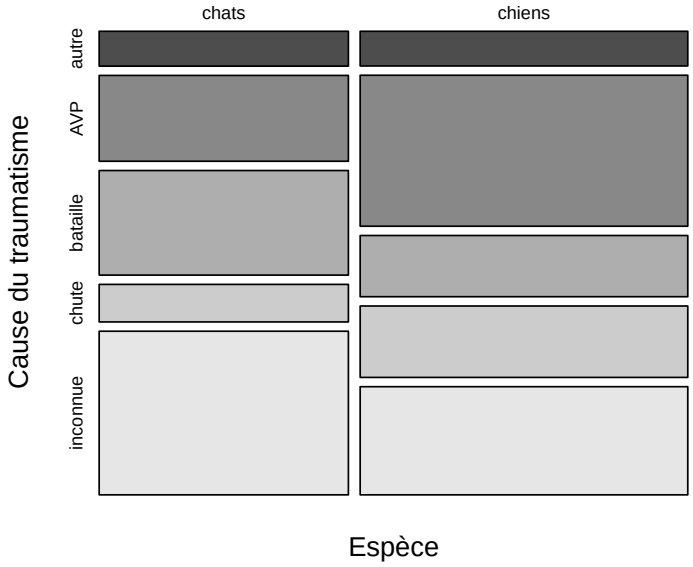


FIGURE 5 – Diagrammes en barres ou en bandes représentant la distribution de la cause du traumatisme (inconnue, chute, bataille, Accident de la Voie Publique, autre) chez les chiens et les chats admis en clinique sur le campus vétérinaire pour traumatisme en 1996.

2.3.2 La représentation classique de la distribution d'une variable quantitative discrète

Toujours dans la thèse vétérinaire (Mathilde Poinssot, Maisons Alfort, 2011) étudiant la reproduction de chiens de race sur divers élevages, prenons l'exemple de la **taille de la portée**, *i.e.* le **nombre de chiots par portée**, qui est une **variable quantitative discrète** (cf. Table 2).

PORTEE	TAILLE
portee_1	6
portee_2	4
portee_3	13
portee_4	9
portee_5	7
portee_6	7
portee_7	7
portee_8	11
portee_9	5
portee_10	10

TABLE 2 – Dix premières lignes du fichier des données brutes observées pour la variable "taille de portée".

Comme pour une variable qualitative il est naturel de résumer ces données brutes par

— les effectifs pour chacune des valeurs observées (qui vont ici de 1 à 17 animaux par portée),

```
##
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
##  39  66  80 119 118 122 131 108  86  52  33  21  17   3   1   0   2
```

— ou les fréquences correspondantes obtenues en divisant les effectifs par l'effectif total.

```
##
##       1       2       3       4       5       6       7       8       9      10
## 0.03908 0.06613 0.08016 0.11924 0.11824 0.12224 0.13126 0.10822 0.08617 0.05210
##      11      12      13      14      15      16      17
## 0.03307 0.02104 0.01703 0.00301 0.00100 0.00000 0.00200
```

Le graphe le plus souvent utilisé pour représenter la distribution d'une variable quantitative discrète est le diagramme en bâtons (cf. Figure 6)

On utilise aussi assez couramment, pour représenter la distribution d'une variable discrète, notamment observée sur plusieurs groupes d'effectifs pas très grands mais avec beaucoup d'ex aequos (plusieurs observations avec la même valeur observée), un graphique en points (dit de type "stripchart" ou "dotplot" en anglais). Cela est notamment conseillé pour représenter des scores cliniques observés pour deux groupes d'animaux ayant subi des traitements différents. Dans la figure 7 vous trouverez un exemple issu d'une enquête réalisée auprès d'étudiants vétérinaires en 2017.

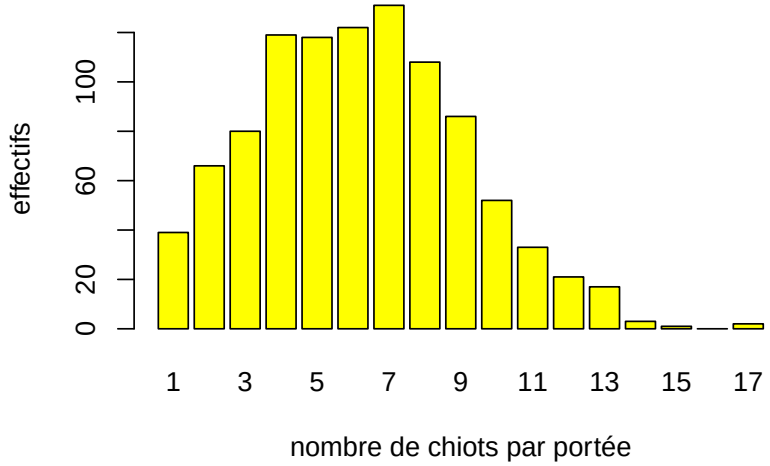


FIGURE 6 – Diagrammes en bâtons de la distribution de la taille des portées de chiots observée sur 998 portées dans le cadre de la thèse de Mathilde Poinssot, Maisons Alfort, 2011

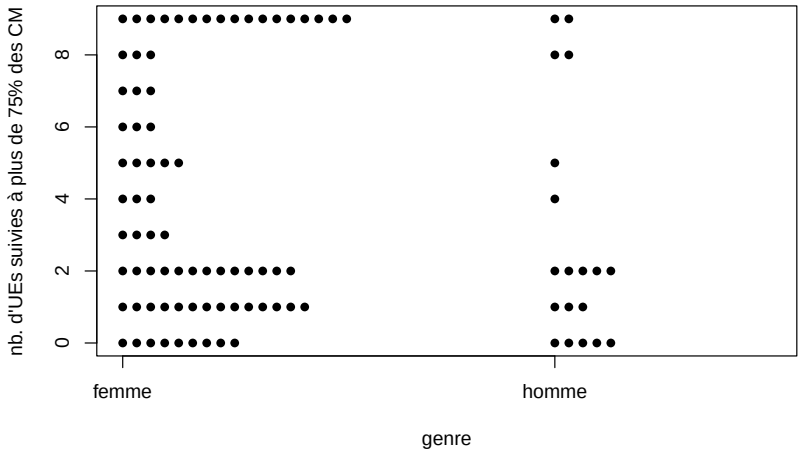


FIGURE 7 – Représentation de la distribution du nombre d'UEs pour lesquelles les étudiants ayant intégré le campus vétérinaire de Lyon en 2016 avaient suivi au moins 75% des cours en amphi au second semestre 2016-2017.

2.3.3 Quelques représentations classiques de la distribution d'une variable quantitative continue

En ce qui concerne la représentation d'une variable aléatoire quantitative continue, quelques rappels théoriques préliminaires s'imposent. Soit X une variable quantitative continue et a une valeur quantitative quelconque, $Pr(X = a) = 0$. Lorsque l'on définit la **fonction de densité de probabilité** f de la variable X, celle-ci ne permet donc pas de quantifier la probabilité d'une valeur unique a, mais la probabilité d'un intervalle $[a; b]$, et cette probabilité correspond à l'aire sous la courbe de densité de probabilité entre a et b ($Pr(a \leq x \leq b) = \int_a^b f(t)dt$, illustré Figure 8).

Toujours dans la thèse vétérinaire (Mathilde Poinssot, Maisons Alfort, 2011) étudiant la reproduction de chiens de race sur divers élevages, prenons l'exemple de la **durée de la gestation**, qui est une **variable quantitative continue** même si elle est exprimée en jours (cf. Table 3).

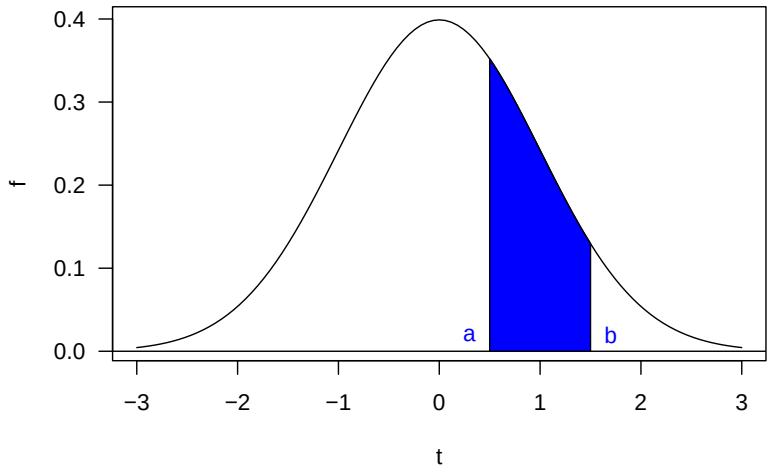


FIGURE 8 – Représentation de la probabilité de l'intervalle [a ; b] (aire sous la courbe coloriée en bleu) sur l'exemple de la fonction de densité de probabilité de la loi normale centrée réduite.

PORTEE	DUREE
portee_1	57
portee_2	62
portee_3	69
portee_4	61
portee_5	61
portee_6	66
portee_7	62
portee_8	63
portee_9	60
portee_10	62

TABLE 3 – Dix premières lignes du fichier des données brutes observées pour la variable "durée de la gestation".

Pour une variable continue, on ne peut résumer les données sous forme d'effectifs et/ou de fréquences qu'en définissant au préalable des classes. Dans cet exemple, sur les classes]45, 50]]50, 55]]55, 60]]60, 65]]65, 70]]70, 75]]75, 80]]80, 85] nous pouvons calculer

— les effectifs pour chacune des classes,

```
## ]45, 50] ]50, 55] ]55, 60] ]60, 65] ]65, 70] ]70, 75] ]75, 80] ]80, 85]
##        2        4      292      577       50        1        1        1
```

— ou les fréquences correspondantes obtenues en divisant les effectifs par l'effectif total.

```
## ]45, 50] ]50, 55] ]55, 60] ]60, 65] ]65, 70] ]70, 75] ]75, 80] ]80, 85]
##  0.00216  0.00431  0.31466  0.62177  0.05388  0.00108  0.00108  0.00108
```

A partir de ces effectifs ou fréquences on réalise un **histogramme de fréquences** (cf. Figure 9) qui donne grossièrement la forme de la **fonction de densité de probabilité** estimée à partir des données.

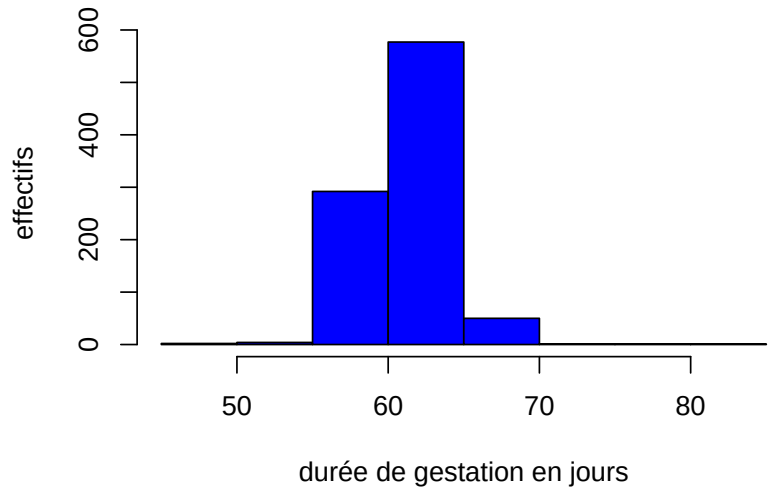


FIGURE 9 – Histogramme de fréquences de la durée de gestation en effectifs, avec des classes de largeur 5 jours, à partir de l'observation de 928 portées dans le cadre de la thèse de Mathilde Poinssot, Maisons Alfort, 2011

Il est très tentant sur ces données de définir des classes plus petites afin de visualiser plus finement la forme de la distribution, ce que l'on peut se permettre de faire ici car l'effectif est grand mais qui poserait plus de problème si l'effectif était petit.

? ? ? Prenez le temps de bien analyser la figure 10 et demandez-vous pourquoi on ne voit pas souvent des histogrammes de fréquences dans les articles publiés dans le domaine des sciences de l'animal.

Dans un histogramme de fréquences, qui est une représentation de la fonction de densité de probabilité, sur chaque classe est représenté un rectangle d'aire proportionnelle à la fréquence ou à l'effectif de la classe. Afin que cette fréquence (ou cet effectif) puisse être lue sur l'axe des ordonnées on choisit généralement des classes de taille identique de sorte que la surface des rectangles soit proportionnelle à leur hauteur. Il est néanmoins possible de tracer un histogramme avec des classes de différentes largeurs, mais celui-ci ne pourra par être présenté classiquement avec l'axe des Y gradué en effectifs ou fréquences, mais sera représenté avec l'axe des Y gradué en densité de probabilité (cf. Figure 11).

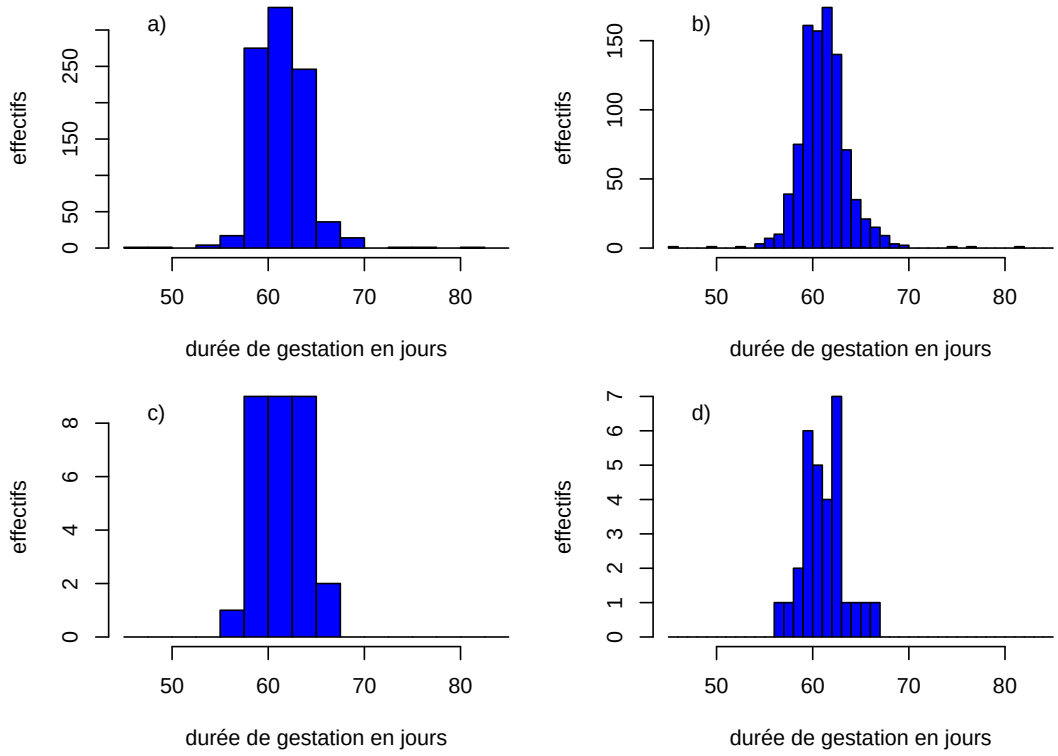


FIGURE 10 – Histogramme de fréquences de la durée de gestation en effectifs, avec des classes de largeur différentes 2.5 jours pour a) et c) et 1 jour pour b) et d), à partir de l'observation des 928 portées pour a) et b) ou d'un sous-échantillon de 30 portées tirées au hasard parmi les 928 pour c) et d)

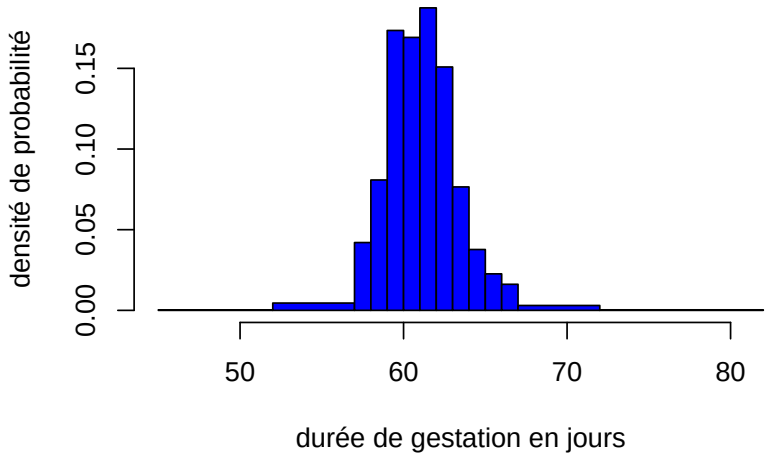


FIGURE 11 – Histogramme de fréquences de la durée de gestation en densité de probabilité à partir de l'observation des 928 portées.

? ? ? Dans une représentation d'une distribution sous forme de fonction de densité de probabilité, comme dans la figure 8 ou l'histogramme 11, on ne peut donner aucune interprétation simple aux ordonnées des points sur la courbe (ou des hauteurs des rectangles pour un histogramme), qui dépendent de la gamme des valeurs de la variable étudiée. Pourquoi les valeurs de densité de probabilité dépendent-elles de la gamme des valeurs de la variable étudiée ?

— Le terme **histogramme** est souvent utilisé à tort pour désigner d'autres graphes, alors qu'il doit être réservé uniquement à ce type de graphe visant à représenter, à partir de données observées d'une variable quantitative continue, la forme de sa fonction de densité de probabilité.

— Par ailleurs, une erreur courante, consiste à réaliser un diagramme en batons à partir d'une variable quantitative (discrète ou continue) comme s'il s'agissait d'une variable qualitative, comme dans la figure 12. Cette erreur peut induire des conclusions erronées sur la forme de la distribution.

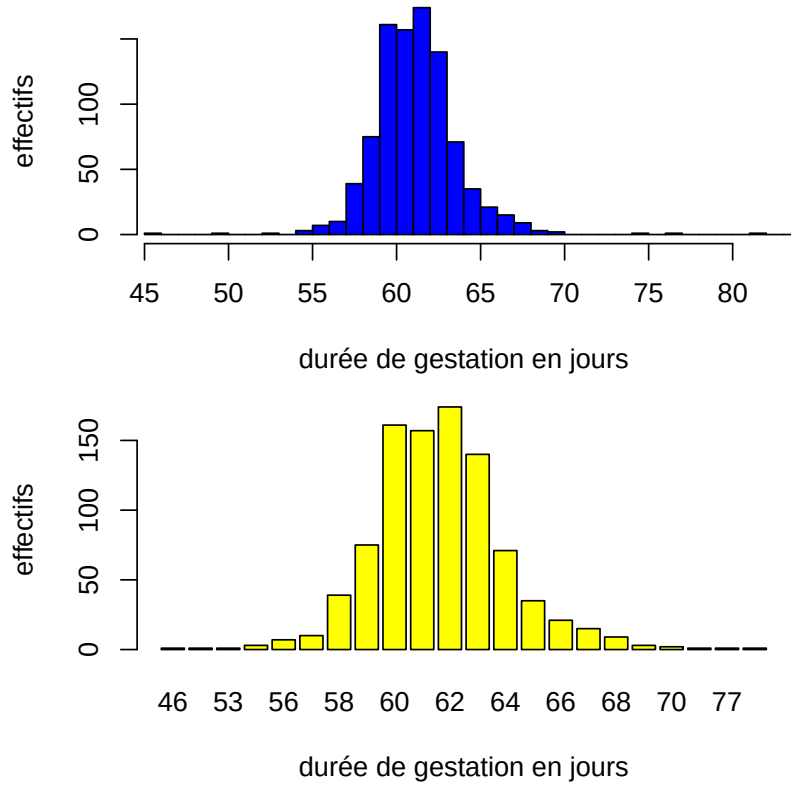


FIGURE 12 – Comparaison de l'histogramme de fréquences de la durée de gestation à partir de l'observation des 928 portées au diagramme en batons réalisé sur les mêmes données sans prendre en compte, à tort, le caractère quatitatif de la variable.

? ? ? Prenez le temps d'analyser la figure 12 pour comprendre en quoi les deux graphes diffèrent et en quoi la non prise en compte de la nature quantitative de la variable peut induire en erreur.

Pour une variable quantitative continue, on définit aussi sa **fonction de répartition** F (cf. Figure 13). La valeur de cette fonction de répartition en un point donné x correspond à la probabilité pour que la variable soit inférieure ou égale à x soit à l'aire sous la courbe de densité de probabilité f à gauche de x : $F(x) = Pr(t \leq x) = \int_{-\infty}^{x} f(t)dt$.

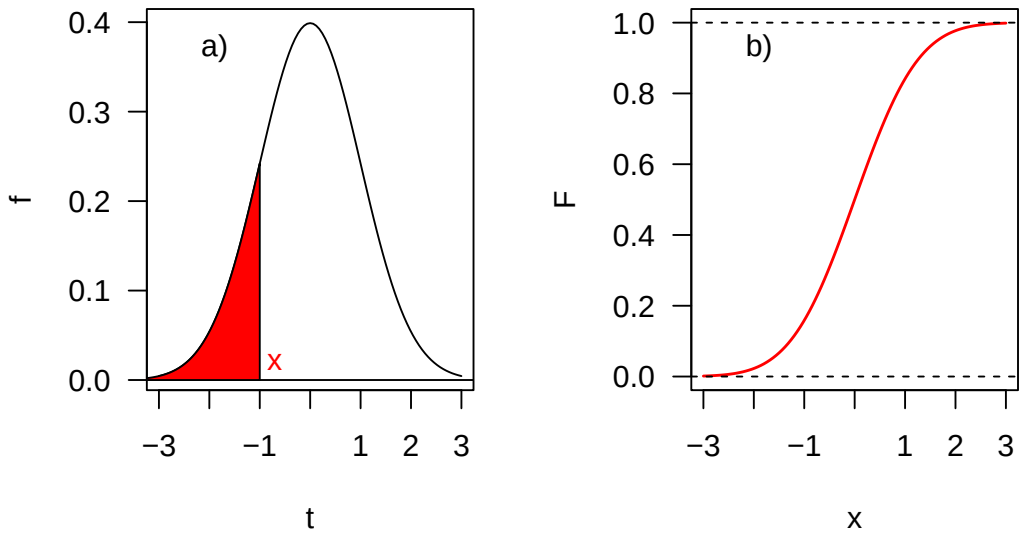


FIGURE 13 – Représentation de la fonction de répartition sur l'exemple de la fonction de densité de probabilité de la loi normale centrée réduite :a) valeur de F pour une valeur de x, aire sous la courbe à gauche de x coloriée en rouge, b) courbe de F pour toutes les valeurs de x .

Le représentation de la fonction de répartition à partir de données, qu'on appelle aussi le **diagramme des fréquences cumulées**, ne nécessite pas de définir des classes (intervalles). On **classe les N observations par ordre croissant**, on attribue **à chaque observation x_i son rang i dans le classement**, et on peut alors dire que $F(x_i) = Pr(t \leq x_i) = \frac{i}{N}$ (proportion d'observations inférieures ou égales à x_i).

Néanmoins la définition de la **fonction de répartition empirique** (c'est-à-dire calculée à partir de données observées) n'est pas unique. Pour la plupart des lois théoriques, dans la définition de la fonction de répartition l'utilisation du signe $\leq$ ou $<$ importe peu, car la probabilité associée à un point est nulle ($F(x) = Pr(t \leq x) = Pr(t < x) = \int_{-\infty}^{x} f(t)dt$). Néanmoins si on utilise le signe $<$ pour définir la fonction de répartition empirique on aura alors comme $F(x_i) = Pr(t < x_i) = \frac{i-1}{N}$

Souvent on utilise encore une troisième définition qui prend la moyenne entre les valeurs de F définies précédemment pour représenter le diagramme des fréquences cumulées. Elle présente l'avantage de donner une graphe qui part au-dessus de 0 et arrive en dessous de 1 : $F(x_i) = \frac{i-0.5}{N}$. Ce choix n'a pas beaucoup d'impact si le nombre total d'observations est grand (cf. Figure 14) mais en a s'il est petit (cf. Figure 15).

? ? ? Attardez-vous sur la figure 15 pour vous assurer que vous avez bien compris et que vous sauriez construire un diagramme des fréquences cumulées à la main à partir de données observées (vous pouvez essayer de refaire le graphe rouge à partir des valeurs observées qui sont en légende de la figure).

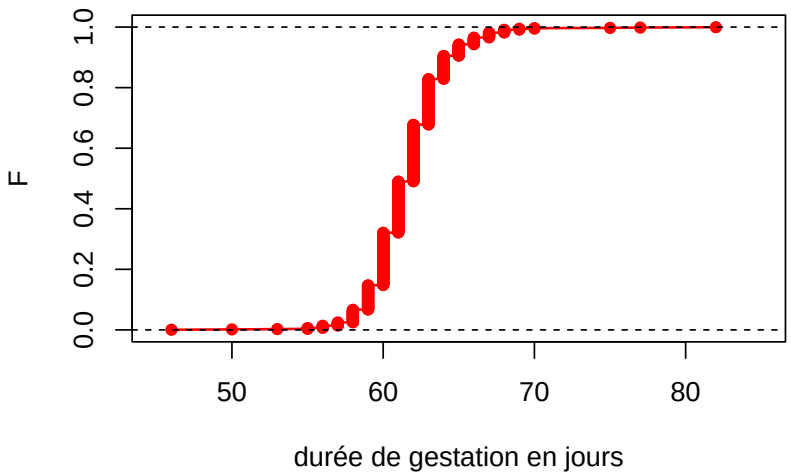


FIGURE 14 – Diagramme des fréquences cumulées (représentation de la fonction de répartition empirique) de la durée de gestation à partir de l'observation de 928 portées dans le cadre de la thèse de Mathilde Poinssot, Maisons Alfort, 2011

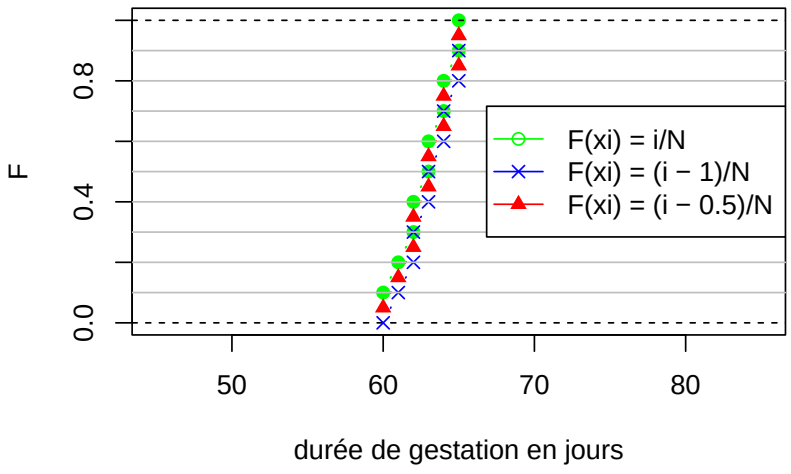


FIGURE 15 – Diagramme des fréquences cumulées pour un petit nombre de valeurs observées égales (valeurs classées par ordre croissant à 60, 61, 62, 62, 63, 63, 64, 64, 65, 65) suivant les trois définitions de la fonction de répartition empirique.

Une autre représentation couramment utilisée pour une variable quantitative continue est le **diagramme en boîte**, ou **diagramme en boîte à moustache**. On y représente en théorie les trois quartiles observés et des valeurs minimale et maximale (cf. Figure 16). Pour calculer les quartiles (quantiles à 25, 50 et 75%) on attribue à chaque observation x_i sa fréquence cumulée comme précédemment (classiquement $F(x_i) = \frac{i-0.5}{N}$) et on définit les valeurs de x correspondant à $F(x) = 0.25$, 0.5 et 0.75.

— Premier quartile : $F(Q_{0.25}) = 0.25$

— Deuxième quartile (médiane) : $F(Q_{0.5}) = 0.50$

— Troisième quartile : $F(Q_{0.75}) = 0.75$

Diverses méthodes sont utilisables pour calculer les quartiles, notamment utilisant ou non une interpolation de la courbe de fonction de répartition empirique. La méthode que vous avez vue au lycée est une méthode simple possible, sans interpolation.

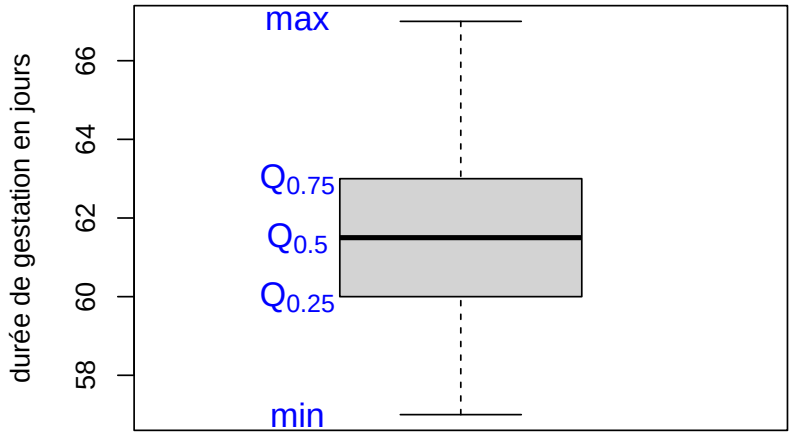


FIGURE 16 – Diagramme en boîte de la durée de gestation sur 30 portées prises au hasard parmi les 928

En pratique, la plupart des logiciels proposent par défaut une variante de cette représentation visant à représenter de façon séparée les valeurs considérées comme extrêmes, en les définissant par exemple comme indiqué en rouge dans la figure 17.

? ? ? Tracez un diagramme en boîte pour chacune des deux séries de valeurs suivantes : série 1) 2, 5, 12, 17, 26, 78, 127, 301 et 500, série 2) 60, 60, 62, 63, puis répondez aux questions suivantes :

— A partir du premier diagramme en boîte, pensez-vous que la distribution dont a été tirée la série 1 soit normale (Gaussienne) ?

— Quel sont d'après vous les avantages du diagramme en boîte par rapport à l'histogramme de fréquences ?

— Pensez-vous qu'il est raisonnable de tracer un diagramme en boîte avec un tout petit nombre d'observations ? Que représenteriez-vous à la place dans un tel cas (cf. série 2) ?

— Pensez-vous que des diagrammes en boîte seraient adaptés pour représenter les deux distributions représentées sur la figure 7 ? Argumentez votre réponse.

Le dernier graphe que nous allons présenter est très souvent utilisé sur des variables quantitatives continues, mais plus rarement montré dans les publications. Il s'agit du **diagramme quantile-quantile**. Il est utilisé lorsque l'on veut vérifier la normalité d'une distribution, ou du moins que la distribution

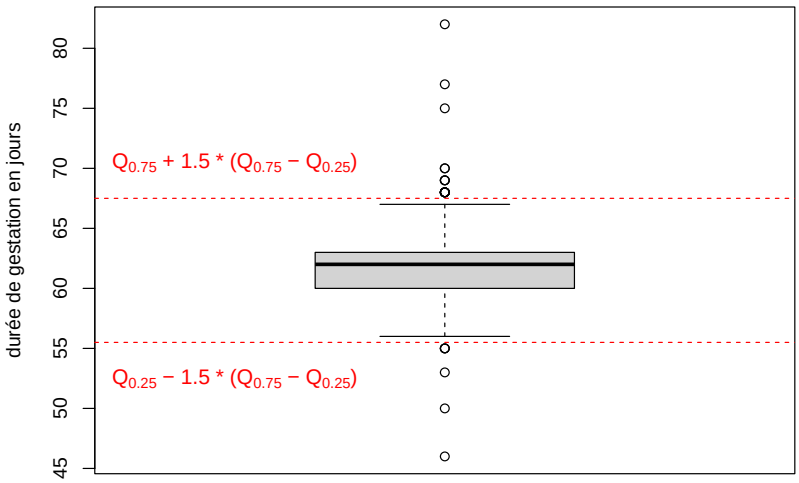


FIGURE 17 – Diagramme en boîte de la durée de gestation sur les 928 portées, avec individualisation des valeurs considérées comme extrêmes, c'est-à-dire ici dépassant les seuils indiqués en rouge.

observée ne s'écarte pas trop d'une distribution normale.

Pour le construire, on attribue à chaque observation x_i de rang i sa fréquence cumulée suivant la définition $F(x_i) = \frac{i-0.5}{N}$, puis pour chaque valeur de i on regarde quelle valeur de u_i dans la loi normale centrée réduite $N(0,1)$ possède la même valeur de F : $F_{N(0,1)}(u_i) = F(x_i)$. Ensuite pour chaque valeur de i on reporte sur le graphe un **point d'abscisse** u_i (quantile de la loi normale) et **d'ordonnée** x_i (quantile observé). **Si la loi observée est normale les points sont à peu près alignés.**

? ? ? Cette description vous paraît peut-être un peu abstraite. Pour bien la comprendre, entraînez-vous à refaire le diagramme quantile-quantile de la figure 18 à partir des valeurs observées données dans sa légende et des quantiles Q de la loi normale pour les valeurs de F utiles donnés ci-dessous :

F	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
Q	-1.64	-1.04	-0.674	-0.385	-0.126	0.126	0.385	0.674	1.04	1.64

Lorsque l'on réalise le diagramme quantile-quantile de la durée de gestation sur l'ensemble des 928 portées (cf. Figure 19) on s'aperçoit que la plupart des points sont alignés, sauf les points les plus extrêmes. On observe donc ici un faible écart à la normalité de la distribution sur ces deux queues de distribution (ici valeurs un peu plus extrêmes qu'attendu avec une loi normale puisque les quantiles observés sont plus éloignés du centre de la distribution qu'attendu).

? ? ? Peut-être que vous vous demandez pourquoi les points d'un diagramme quantile-quantile sont alignés si la distribution observée est normale. Pour vous aider à intuiter cela, demandez-vous sur quelle droite les points du graphe devraient être alignés si la distribution était normale et que les données avaient été centrées (centrer = enlever la moyenne) et réduites (réduire = diviser par l'écart type) avant de faire le graphe.

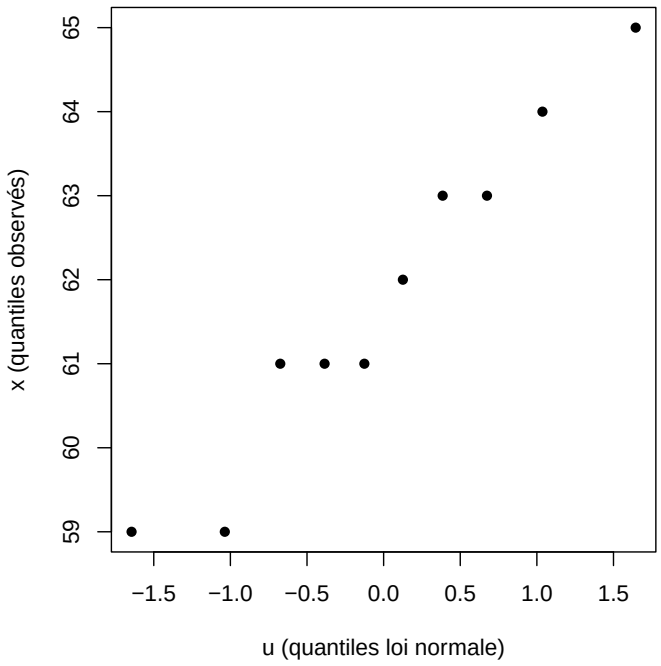


FIGURE 18 – Diagramme quantile-quantile réalisé sur la série de valeurs suivante : 59, 59, 61, 61, 61, 62, 63, 63, 64, 65.

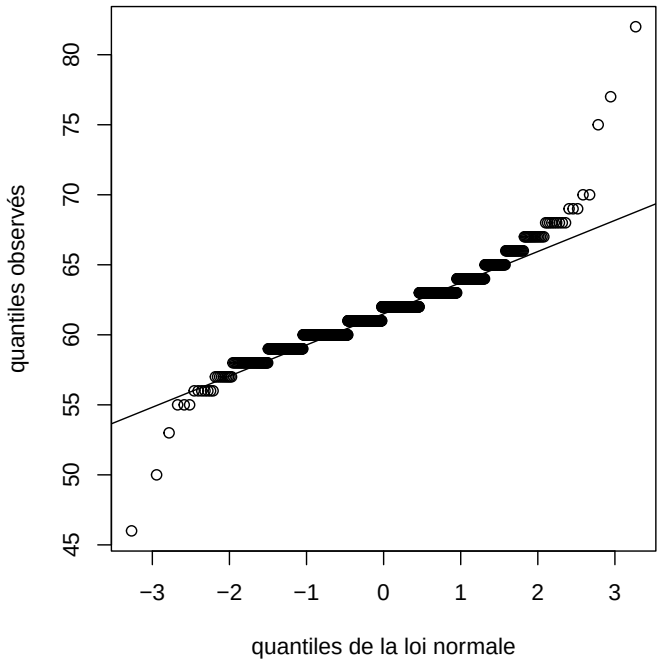


FIGURE 19 – Diagramme Quantile-Quantile de la durée de gestation sur les 928 portées avec une droite attendue ajoutée, obtenue en reliant les points des deux quartiles à 25 et 75%.

Nous avons vu diverses représentations pour le cas d'une variable quantitaive continue. Voici un petit récapitulatif de celles-ci avec quelques éléments à retenir concernant leur utilisation

- **Histogramme de fréquences**
  Vision fine de la densité de probabilité - grand nombre de points et définition appropriée de classes nécessaires.
- **Diagramme des fréquences cumulées ("ECDF plot")**
  Visualisation de la fonction de répartition empirique.
- **Diagramme en boîte ("boxplot")**
  Visualisation synthétique de la densité de probabilité - possible même avec un nombre de points modéré (si nombre trop faible, représentation directe des points)
- **Diagramme Quantile-Quantile ("QQ-plot")**
  Vérification de la normalité d'une distribution.

2.4 La réduction des données pour une variable quantitative

La réduction des données consiste à réduire ou résumer les données par un ou plusieurs paramètres statistiques comme la moyenne. Nous ne traiterons dans ce chapitre que de la réduction des données pour une variable quantitative.

2.4.1 Les paramètres de position

Les paramètres de position visent à localiser le centre de la distribution observée, qu'on appellera parfois la **tendance centrale**. Le plus courant est la **moyenne arithmétique** (appelée plus couramment moyenne) définie pour une variable aléatoire x par : $\overline{x} = \frac{1}{N}\sum_{k=1}^{N} x_i$, N représentant la taille de l'échantillon sur lequel ont été observées les valeurs x_i.

On utilise parfois aussi la **médiane** (deuxième quartile $Q_{0.5}$) our localiser le centre de la distribution. La médiane est un paramètre statistique plus robuste que la moyenne au sens où il est moins sensible aux valeurs extrêmes observées.

Enfin, on parle parfois du **mode** de la distribution qui correspond au pic de la distribution. La valeur numérique de ce dernier paramètre est plus difficile à estimer. Si on la définit par exemple comme le centre de la classe de plus forte fréquence dans un histogramme de fréquences, elle dépend des classes utilisées pour représenter l'histogramme de fréquences. De ce fait, ce dernier paramètre est rarement précisément quantifié mais plutôt cité oralement pour décrire une distribution observée. Par exemple une distribution avec deux pics de densité de probabilité sera appelée bi-modale (c'est le cas par exemple de la distribution observée du nombre d'UEs pour lesquelles les étudiantes vétérinaires avaient suivi au moins 70% des CMs, représentée en figure 7.

2.4.2 Les paramètres de dispersion et l'intervalle de fluctuation

Les paramètres de **dispersion** visent à décrire comment les valeurs observées se dispersent autour de la valeur centrale. Les paramètres de dispersion le plus classiquement utilisés sont :

- **la variance** $V(x) = \frac{1}{N}\sum_{k=1}^{N}(x_i - \overline{x})^2$,

— **l'écart type** (noté généralement SD pour 'Standard Deviation' en anglais) et qui n'est autre que la racine carrée de la variance ($SD = \sqrt{V(x)}$),
— et enfin le **coefficient de variation** noté CV qui représente l'écart type en valeur relative à la moyenne ($CV = \frac{SD}{\overline{x}}$ est souvent exprimé souvent en $\%$).

? ? ? Il est important d'avoir en tête la formule de la variance, pour bien comprendre de quoi il s'agit et comprendre pourquoi il est plus facile d'interpréter l'écart type que la variance. La **variance représente la moyenne des carrés des écarts à la moyenne**. A partir de sa définition essayez d'expliquer :

— pourquoi dans cette définition on a pris la somme des carrés des écarts à la moyenne et non pas simplement la somme des écarts à la moyenne, et
— pourquoi de ce fait l'écart type est plus facile à interpréter que la variance ?

Lorsque l'on souhaite décrire la dispersion d'une distribution à l'aide d'un paramètre plus robuste que les paramètres précédents, on peut utiliser **l'écart interquartile** noté EIQ défini comme l'écart entre le troisième et le premier quartile ($EIQ = Q_{0.75} - Q_{0.25}$). Notons que l'écart inter-quartile n'est autre que la longueur de la boîte dans le diagramme en boîte.

On définit **l'intervalle de fluctuation** à $k\%$ comme l'intervalle dans lequel se trouve $k\%$ de la distribution (plus précisément en supposant qu'on a $\frac{k}{2}\%$ de la distribution au-dessus et $\frac{k}{2}\%$ en-dessous de l'intervalle). On utilise un tel intervalle dans le domaine de la biologie médicale notamment lorsque l'on veut définir un **intervalle de référence pour une variable biologique** dans la population saine (ex. Exemple : valeurs usuelles du taux d'hémoglobine chez le chat sain). L'intervalle de référence est souvent défini comme l'intervalle de fluctuation à 95%. Deux méthodes sont alors classiquement utilisées pour estimer cet intervalle de fluctuation à 95% :

— lorsque le **nombre d'observations est grand** (plus d'une centaine) on calcule cet intervalle de fluctuation à partir des quantiles à 2.5 et 97.5% ($[Q_{0.025}, Q_{0.975}]$),
— et lorsque le nombre d'observations est modéré (quelques dizaines) et que **la distribution est proche d'une loi normale**, on utilise la moyenne et l'écart type et le quantile à 97.5% de la loi normale (égal à 1.96, cf. Table 4) ce qui permet de définir l'intervalle comme $[\overline{x} - 1.96 \times SD, \overline{x} + 1.96 \times SD]$ qui est souvent approché par $[\overline{x} - 2 \times SD, \overline{x} + 2 \times SD]$.

Rappelons que si x suit une loi normale, l'intervalle $[\overline{x} - 1.96 \times SD, \overline{x} + 1.96 \times SD]$ est l'intervalle de fluctuation à 95% et que l'intervalle $[\overline{x} - SD, \overline{x} + SD]$ est l'intervalle de fluctuation à 68%. Ce dernier intervalle est souvent utilisé sans que ces utilisateurs aient en tête qu'il s'agit d'un intervalle qui ne contient que 68% des valeurs (cf. Figure 20).

? ? ? Pour voir si vous avez bien compris, calculez l'intervalle de référence (= intervalle de fluctuation à 95%) de la température corporelle du chat, en supposant que la distribution de cette variable chez les chats sains est normale et qu'on a estimé à partir d'un échantillon de 144 chats sains une moyenne à 39 degrés celsius et un écart type à 0.25.

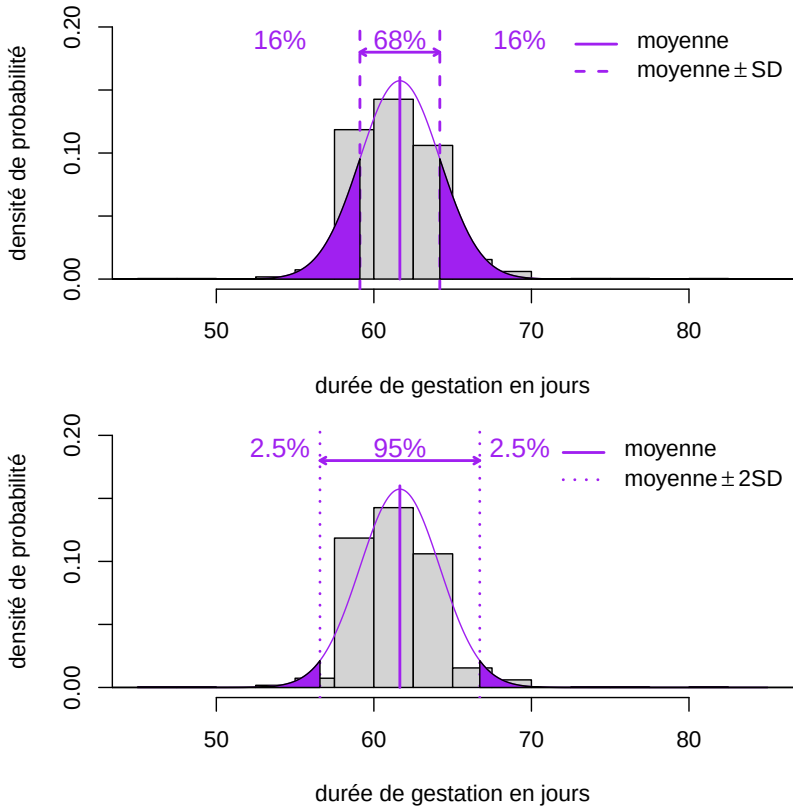


FIGURE 20 – Intervalles de fluctuations à 68 et 95% estimés à partir de la moyenne et de l'écart type des durées de gestation en approchant la distribution par une loi normale

2.4.3 Les limites des paramètres statistiques classiques

Bien que le calcul des paramètres statistiques classiques que sont la moyenne et la variance ou l'écart type soit toujours possible dès que l'on dispose d'au moins deux observations, il est important d'avoir à l'esprit que ces paramètres ne fournissent pas toujours un bon résumé de la distribution observée. Ces paramètres caractérisent très bien les données issues d'une loi proche d'une loi normale, mais dès que la loi observée s'éloigne de cette loi classique, et notamment lorsqu'elle est fortement dissymétrique, ils deviennent beaucoup moins pertinents pour caractériser la loi.

? ? ? Pour vous en convaincre, observez bien les différents graphes de la figure 21 et au cas par cas déterminez s'il serait raisonnable de résumer les données par une moyenne et un écart type, en vous justifiant, et le cas échéant proposez d'autres paramètres résumés plus pertinents.
En imaginant la forme que peut avoir la distribution des salaires des français, pensez-vous qu'il est pertinent de donner une moyenne comme résumé statistique de cette distribution, comme cela est souvent fait dans les médias ? Justifiez votre réponse.

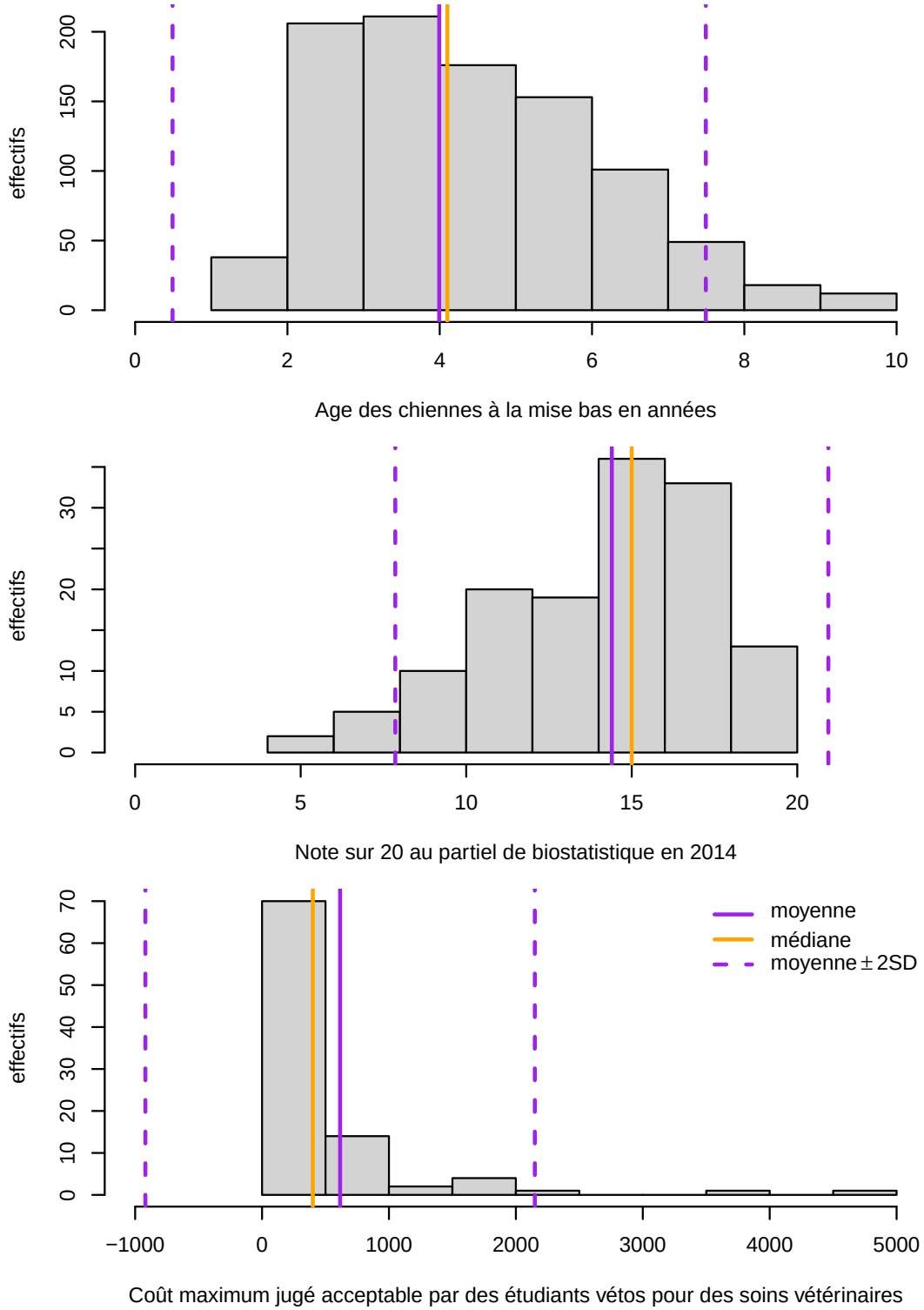


FIGURE 21 – Trois exemples de distributions représentées sous forme d'histogrammes de fréquences, avec ajout d'indications utilisant les moyennes, écarts-types et médianes.

La description des données observées est une étape importante qui doit **IMPERATIVEMENT commencer par une bonne représentation graphique** de la distribution étudiée.
Il convient ensuite de **bien réfléchir avant de calculer les paramètres statistiques classiques** (moyenne, variance ou écart type) : *"Décrivent-ils bien la distribution observée ?"*
Il est parfois plus raisonnable de ne pas résumer les données (cas des très petis effectifs par exemple) ou d'utiliser des paramètres descriptifs alternatifs comme la médiane (cas de distributions avec valeurs extrêmes susceptibles d'avoir une influence importante sur la moyenne par exemple).

3 La statistique inférentielle

Nous aborderons dans ce chapitre les grands principes et concepts de la statistique inférentielle.

3.1 Objectifs pédagogiques

A l'issue de l'étude de ce chapitre et de la réalisation des deux premiers TD de S3, vous devriez :

— Savoir définir les notions suivantes : inférence statistique, échantillonnage aléatoire simple, distribution d'échantillonnage, estimation, estimation sans biais, test de signification, test d'hypothèse, test d'équivalence, différence significative, risques d'erreur de première et deuxième espèces, p-value (valeur de p ou degré de signification), puissance d'un test d'hypothèse.

— Savoir ce que représentent SD et SE (ou SEM).

— Avoir bien compris le théorème de l'approximation normale.

— Savoir juger de l'applicabilité de ce théorème et vérifier les conditions d'utilisation des divers intervalles de confiance.

— Savoir ce que représente un intervalle de confiance et ce qui le différencie d'un intervalle de fluctuation.

— Savoir calculer à la main (avec une calculatrice) un intervalle de confiance sur une moyenne et sur une fréquence.

— Savoir réaliser à la main un test à partir de sa fiche technique.

— Savoir interpréter le résultat d'un test de signification et notamment avoir les idées claires sur les conclusions qu'on peut tirer d'un test.

— Savoir réaliser un test d'équivalence et en interpréter les résultats.

3.2 Echantillonnage et théorème central limite

3.2.1 Principe de l'échantillonnage et définition de la distribution d'échantillonnage

L'échantillonnage concerne la première étape d'une étude statistique qui est la collecte des données (Figure 1). Il s'agit dans cette étape d'obtenir un échantillon représentatif de la population qui nous intéresse appelée souvent population cible. Le principal enjeu de cette étape est d'éviter tout biais d'échantillonnage qui pourrait conduire à des conclusions erronées sur la population cible. La méthode la plus classiquement utilisée pour éviter ces biais est l'**échantillonnage aléatoire simple**. Celle-ci implique que le choix de chaque individu ou unité de l'échantillon fait l'objet d'un tirage au hasard et que les tirages des différents individus ou unités de l'échantillon sont indépendants les uns des autres et que chaque individu ou unité a la même probabilité d'être tiré. Généralement une analyse statistique de données observées sur un échantillon vise non seulement à décrire ce qui a été observé sur l'échantillon mais aussi à en tirer des conclusions sur la population dont a été tiré l'échantillon : c'est ce qu'on appelle l'**inférence statistique** (Figure 1). L'inférence statistique consiste donc à déduire de ce que l'on a observé sur un échantillon des conclusions sur la population dont il a été tiré.

Lorsque l'on veut estimer un paramètre caractérisant la population étudiée à partir d'un échantillon de cette population, la valeur estimée du paramètre dépend néanmoins de l'échantillon. En effet si l'on

pouvait tirer un nouvel échantillon dans la même population on trouverait une valeur estimée du paramètre quelque peu différente de celle obtenue sur le premier échantillon. On parlera des **fluctuations d'échantillonnage** de ce paramètre pour désigner ces fluctuations potentielles d'un paramètre estimé d'un échantillon à l'autre. Si l'on tire un grand nombre d'échantillons, on peut représenter la distribution en fréquences des valeurs estimées du paramètre pour chaque échantillon. On appellera cette distribution la **distribution d'échantillonnage du paramètre** (cf. Figure 22 pour l'exemple de la moyenne d'une variable quantitative continue).

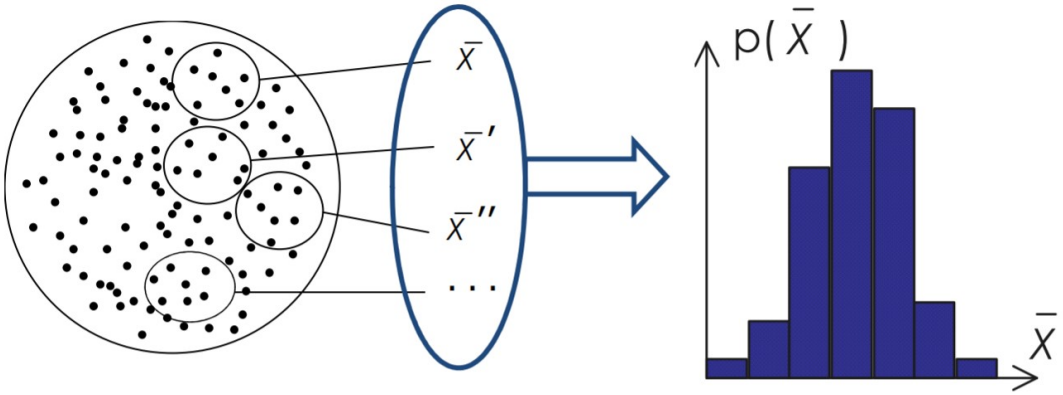


FIGURE 22 – Schéma illustrant la notion de distribution d'échantillonnage d'une moyenne.

Bien entendu, lors d'une étude statistique, il est rare que l'on dispose de plusieurs échantillons. Mais ce raisonnement sur plusieurs échantillons potentiels est à la base de tout ce qui va suivre. Pour bien comprendre la suite, il faut donc s'efforcer d'imaginer ce cadre théorique où l'on pourrait obtenir plusieurs échantillons d'une population.

3.2.2 Théorème central limite (ou théorème de l'approximation normale) pour une moyenne

Le **théorème central limite** que nous appelerons aussi **théorème de l'approximation normale** est à la base de très nombreuses méthodes utilisées couramment en statistique inférentielle, et c'est pourquoi il est nécessaire de bien le comprendre. La première version du théorème décrit la **distribution d'échantillonnage d'une moyenne** :

Théorème central limite pour une moyenne

Pour des échantillons aléatoires simples de taille N, la moyenne $\overline{X}$ de l'échantillon varie autour de la moyenne μ de la population avec une **erreur standard** $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}}$ noté SE ou SEM (**"Standard Error of the Mean"**), σ étant l'écart type de la population.

- Lorsque la distribution de X dans la population est normale, $\overline{X}$ suit la loi $N(\mu, \frac{\sigma}{\sqrt{N}})$.
- Quelle que soit la distribution de X, lorsque l'effectif N est suffisamment grand, la loi de $\overline{X}$ s'approche de la loi normale $N(\mu, \frac{\sigma}{\sqrt{N}})$.

Le point le plus intéressant de ce théorème est bien sûr le dernier qui dit que quelle que soit la distribution de la variable X, lorsque la taille des échantillons N est suffisamment grande la distribution de $\overline{X}$ tend vers une loi normale. Une illustration de ce théorème pour différentes lois de la variable X

est proposée sur les figures 23 à 25. Pour chacune des figures, des échantillons de tailles croissantes ont été tirés aléatoirement dans la distribution initiale et les distributions d'échantillonnage de la moyenne correspondantes ont été représentées.

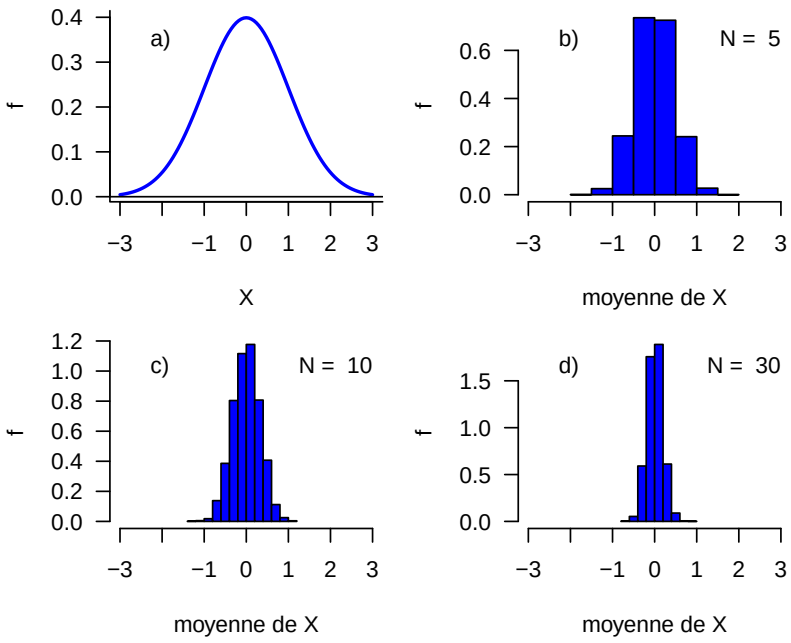


FIGURE 23 – Illustration du théorème central limite pour la moyenne d'une variable distribuée suivant une loi normale. a) loi de départ suivie par X. b) à d) histogrammes de la distribution d'échantillonnage de la moyenne pour différentes tailles N d'échantillons, obtenus à partir de 5000 échantillons de taille N tirés dans la loi de départ.

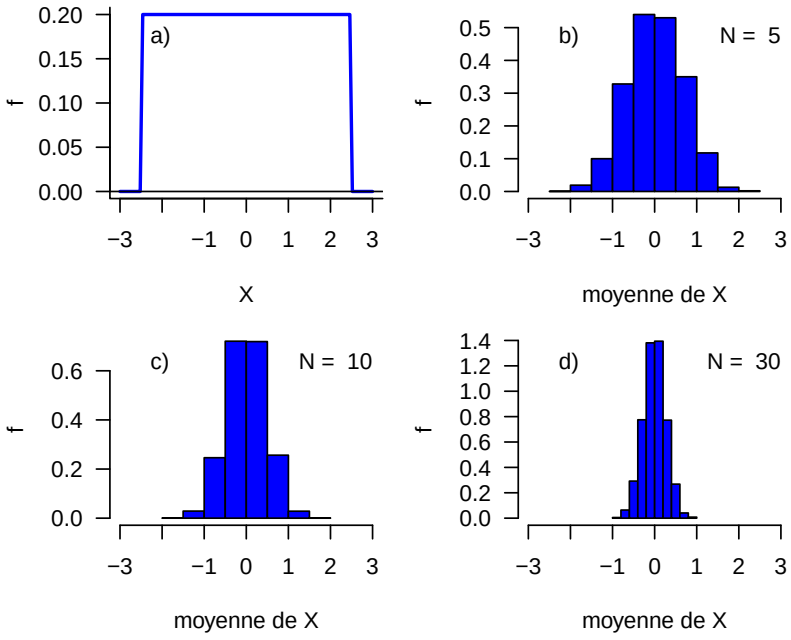


FIGURE 24 – Illustration du théorème central limite pour la moyenne d'une variable distribuée suivant une loi uniforme. a) loi de départ suivie par X. b) à d) histogrammes de la distribution d'échantillonnage de la moyenne pour différentes tailles N d'échantillons, obtenus à partir de 5000 échantillons de taille N tirés dans la loi de départ.

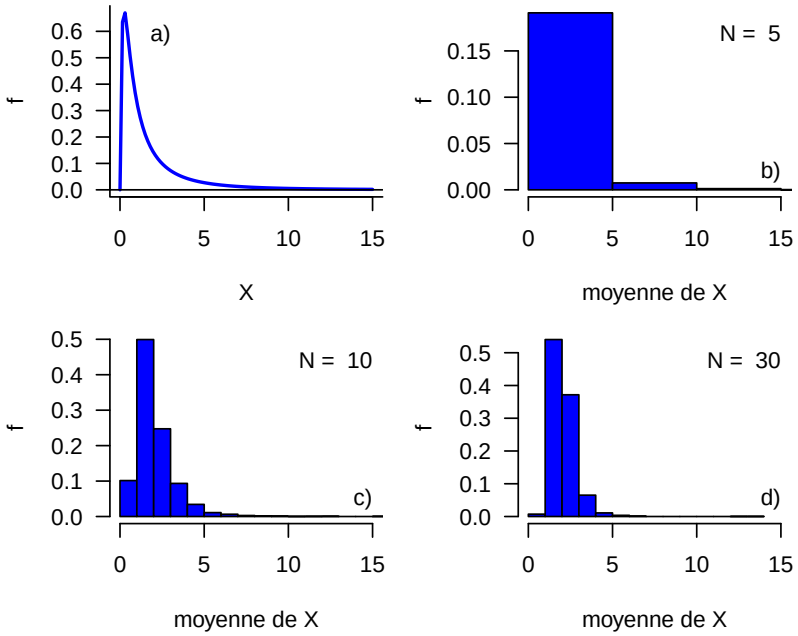


FIGURE 25 – Illustration du théorème central limite pour la moyenne d'une variable distribuée suivant une loi lognormale. a) loi de départ suivie par X. b) à d) histogrammes de la distribution d'échantillonnage de la moyenne pour différentes tailles N d'échantillons, obtenus à partir de 5000 échantillons de taille N tirés dans la loi de départ.

? ? ? Ce théorème indique une convergence asymptotique de la loi de $\overline{X}$ vers la loi normale, c'est-à-dire lorsque la taille N de l'échantillon devient grande, mais en pratique à partir de quelle valeur de N peut-on l'appliquer ? Certains vieux ouvrages indiquent qu'il est applicable dès que $N > 30$. Examinez bien les figures 23 à 25 et posez-vous la question de la pertinence de définir une valeur de N à partir de laquelle ce théorème serait applicable.

Il est impossible de juger de l'applicabilité du théorème de l'approximation normale sans regarder la forme de la distribution de X dans l'échantillon. En effet comme on le voit dans les trois exemples, plus la loi de X s'écarte de la loi normale, plus N doit être grand pour appliquer le dernier point du théorème.

3.2.3 Théorème central limite (ou théorème de l'approximation normale) pour une fréquence

La deuxième version du théorème décrit la distribution d'échantillonnage d'une fréquence :

Théorème central limite pour une fréquence

Pour des échantillons aléatoires simples de taille N, la fréquence F d'un caractère étudié varie autour de la proportion π_0 de ce caractère dans la population, avec une erreur standard $\sigma_F = \sqrt{\frac{\pi_0(1-\pi_0)}{N}}$. Lorsque l'effectif N est suffisamment grand, la loi de F s'approche de la loi normale $N(\pi_0, \sqrt{\frac{\pi_0(1-\pi_0)}{N}})$.

Cette deuxième version du théorème, que vous avez certainement déjà vue dans un cours de probabilités sous la forme de la convergence de la loi binomiale vers la loi normale, peut être présenté comme découlant directement de la première : il suffit dans la version précédente de définir X comme une variable quantitative prenant la valeur 1 si le caractère étudié est observé et 0 s'il ne l'est pas (par exemple dans le cas où l'on veut estimer la proportion d'animaux malades dans une population $X = 1$ si l'animal est malade, $X = 0$ s'il est sain) et de réaliser que la fréquence d'animaux malades n'est autre que la moyenne de X dans ce cas. Ce théorème est illustré par les figures 26 et 27.

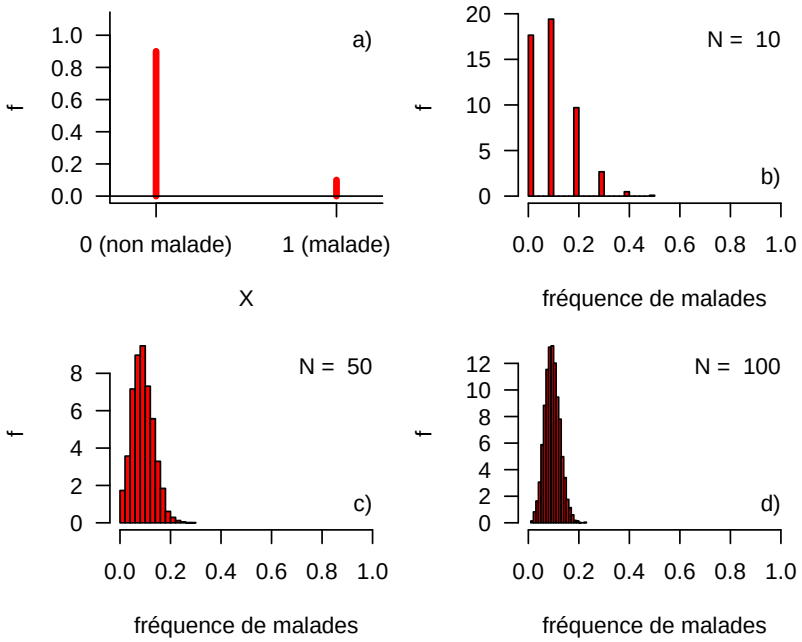


FIGURE 26 – Illustration du théorème central limite pour la fréquence F d'un caractère étudié en supposant la proportion π_0 de ce caractère égale à 10% dans la population. a) loi de départ suivie par X codant pour la présence ou non du caractère. b) à d) histogrammes de la distribution d'échantillonnage de la fréquence observée du caractère pour différentes tailles N d'échantillons, obtenus à partir de 5000 échantillons de taille N tirés dans la loi de départ.

L'effectif requis pour pouvoir appliquer le théorème de l'approximation normale pour l'estimation de la fréquence F d'un caractère étudié dépend de la proportion π_0 de ce caractère dans la population. Plus π_0 est proche de 0 (caractère rare) ou de 1 (caratère très répandu), plus N devra être grand.

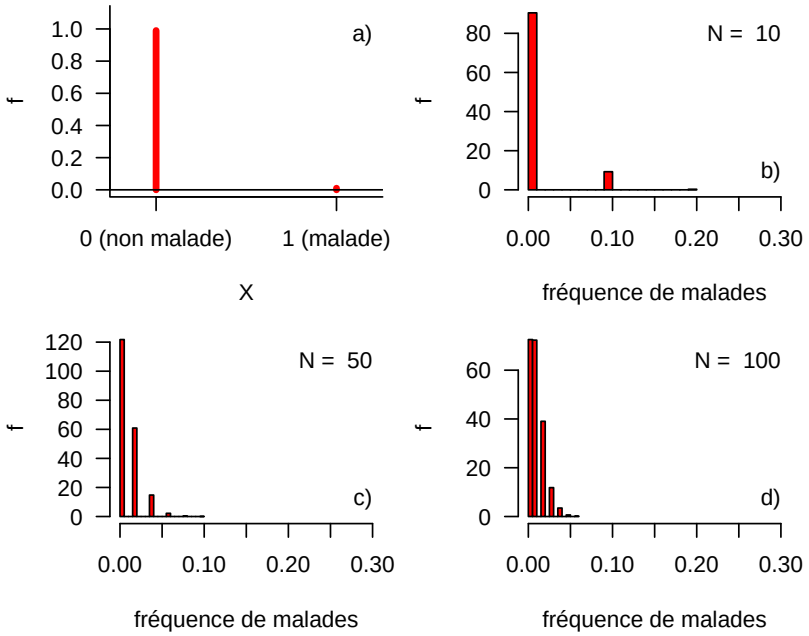


FIGURE 27 – Illustration du théorème central limite pour la fréquence F d'un caractère étudié en supposant la proportion π_0 de ce caractère égale à 1% dans la population. a) loi de départ suivie par X codant pour la présence ou non du caractère. b) à d) histogrammes de la distribution d'échantillonnage de la fréquence observée du caractère pour différentes tailles N d'échantillons, obtenus à partir de 5000 échantillons de taille N tirés dans la loi de départ.

3.2.4 Conclusion

De très nombreux outils statistiques (intervalles de confiance et tests statistiques) sont basés sur le théorème de l'approximation normale. Leur utilisation nécessite donc la vérification au préalable des conditions d'utilisation de ce théorème. Il est donc indispensable, pour une bonne utilisation de ces outils, d'avoir bien compris ce théorème et d'avoir des notions relatives à ses conditions d'utilisation.

3.3 Estimation statistique

Fixons tout d'abord le cadre général de l'estimation statistique.

— On s'intéresse à un **paramètre θ caractérisant la population** cible
(par exemple la moyenne d'une variable biologique quantitative continue, ou sa médiane, ou son écart type, ..., dans la population cible, ou encore si la variable est qualitative binaire, la proportion du caractère étudié dans la population cible).

— **θ est supposé fixe, mais inconnu** du fait qu'on n'a pas accès à la population cible entière mais seulement à un échantillon aléatoire de cette population.

— On souhaite donner la **meilleure estimation possible de** θ (répondre à la question *"Que sait-on sur θ dans la population cible ?"* **à partir** de ce que l'on a observé dans **un échantillon**.

3.3.1 Estimation ponctuelle

Quand on parle d'**estimation ponctuelle**, cela veut dire l'estimation du paramètre θ par une valeur ponctuelle. On notera l'estimation ponctuelle de θ à partir d'un échantillon $\hat{\theta}$. On appelera T l'estimateur

de θ, c'est-à-dire la fonction mathématique qui est appliquée aux données observées sur l'échantillon pour calculer $\hat{\theta}$: $\hat{\theta} =$ T(données observées sur l'échantillon).

On exige souvent d'un estimateur T de θ qu'il soit sans biais, c'est-à-dire qu'en moyenne il ne se trompe pas, autrement dit que la moyenne de la distribution d'échantillonnage de T soit égale à θ : $E(T) = \theta$. Cela revient à dire que si l'on pouvait disposer d'un nombre infini d'échantillons, si on estimait θ sur chacun d'eux à l'aide de l'estimateur T, la moyenne des valeurs obtenues serait égale à θ.

D'après les deux versions du théorème de l'approximation normale que nous avons vues auparavant, nous pouvons dire que **la moyenne $\overline{X}$ est une estimation sans biais la moyenne μ dans la population** dans le cas d'une variable quantitative continue car $E(\overline{X}) = \mu$, et que **la fréquence observée F d'un caractère étudiée est une estimation sans biais de la proportion π_0 du caractère étudié dans la population** car $E(F) = \pi_0$.

Prenons l'exemple de l'estimation de la fréquence de chats FIV positifs à partir d'un échantillon de 50 chats sur lesquels 7 sont détectés FIV positifs. On obtient une fréquence observée de $f = 14\%$ de chats FIV positifs. On estimera la fréquence de chats FIV positifs dans la population correspondante à 14% et on notera $\hat{\pi_0} = f = 0.14$.

Pour obtenir un estimateur sans biais de la variance, une petite correction est nécessaire. En effet, on peut montrer (nous ne ferons pas de démonstration mathématique dans ce cours, pour ne pas le surcharger de formalisme mathématique) que l'espérance de la variance $V(X)$ n'est pas tout à fait égale à la variance de X dans la population σ^2, mais que $E(V(X)) = \frac{(N-1)\times\sigma^2}{N}$. Si l'on estimait sans correction σ^2 par $V(X)$, l'estimation serait donc biaisée. On sous-estimerait σ^2, et ce d'autant plus que N, la taille de l'échantillon, est petit. On utilisera donc comme **estimation sans biais de la variance, la variance corrigée** : $\hat{\sigma}^2 = \frac{N\times V(X)}{N-1} = \frac{1}{N-1}\sum_{k=1}^{N}(X_i - \overline{X})^2$. Cette correction a été calculée de manière à ce que $E(\hat{\sigma}^2) = \sigma^2$.

3.3.2 Estimation par intervalle

Quand on estime un paramètre statistique à partir d'un échantillon, on associe généralement une estimation par intervalle à l'estimation ponctuelle, dans le but de donner une indication quant à la précision de l'estimation ponctuelle et d'indiquer ainsi quelle confiance peut-on accorder à l'estimation à partir du seul échantillon dont on dispose, mais en prenant en compte les fluctuations d'échantillonnage. Pour répondre à cette question sans avoir à répéter l'échantillonnage, on construit classiquement un **intervalle de confiance** autour de l'estimation.

On utilise le plus couramment un intervalle de confiance bilatéral, tel que défini ci-dessous et illustré Figure 28.

Définition d'un intervalle de confiance bilatéral

Intervalle $[t_1; t_2]$ construit de façon à ce qu'**en terme de distribution d'échantillonnage** (sous-entendu si on répétait l'échantillonnage)

$$Pr(t_1 \geq \theta) = Pr(t_2 \leq \theta) = \frac{\alpha}{2}$$
$$\text{donc } Pr(t_1 \leq \theta \leq t_2) = 1 - \alpha$$

t_1 et t_2 sont appelées les limites de confiance et $1 - \alpha$ est appelé le seuil de confiance.

Généralement α est fixé à 5% et l'on parle d'**intervalles de confiance à** 95%.

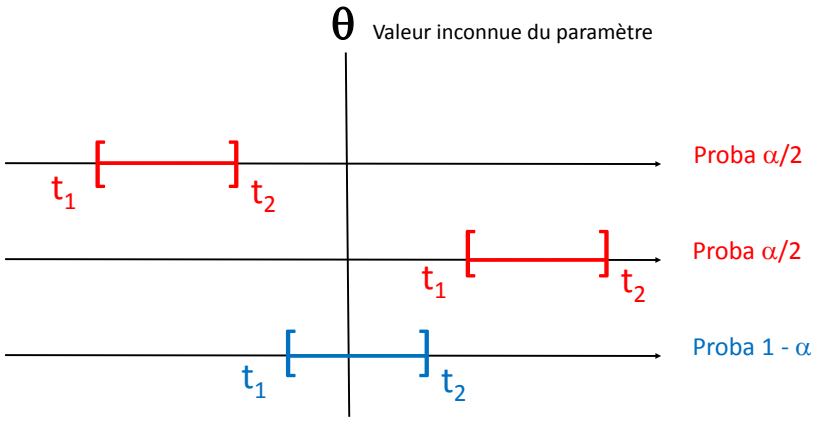


FIGURE 28 – Illustration de la définition d'un intervalle de confiance bilatéral.

Intervalles de confiance sur une moyenne et une fréquence A partir du théorème de l'approximation normale et/ou d'autres résultats de la statistique théorique des intervalles de confiance ont été proposés pour les cas classiques, notamment :

— **l'intervalle de confiance bilatéral autour d'une fréquence** F

$\pi_0 = F \pm u_{1-\frac{\alpha}{2}} \times \sqrt{\frac{F(1-F)}{N}}$

avec $u_{1-\frac{\alpha}{2}}$ le quantile à $1-\frac{\alpha}{2}$ de la distribution normale $N(0,1)$ (pour $\alpha = 0.05$, $u_{1-\frac{\alpha}{2}} = 1.96$, cf. Table 4

utilisable si $NF \geq 20$ et $N(1-F) \geq 20$,

— **l'intervalle de confiance bilatéral autour d'une moyenne** $\overline{X}$

$\mu = \overline{X} \pm t_{N-1;1-\frac{\alpha}{2}} \times \frac{\hat{\sigma}}{\sqrt{N}}$

avec $t_{N-1;1-\frac{\alpha}{2}}$ le quantile à $1 - \frac{\alpha}{2}$ de la distribution de Student de degré de liberté $N-1$ (T_{N-1}),

cf. Table 5

utilisable si le théorème de l'approximation normale est applicable.

La loi normale ainsi que la loi de Student pour quelques valeurs de degrés de liberté vous sont représentées dans la figure 29. Pour rappel la loi de Sudent est plus piquée et avec des queues de distribution plus lourdes que la loi normale lorsque son degré de liberté est faible, et tend vers la loi normale lorsque que son degré de liberté augmente.

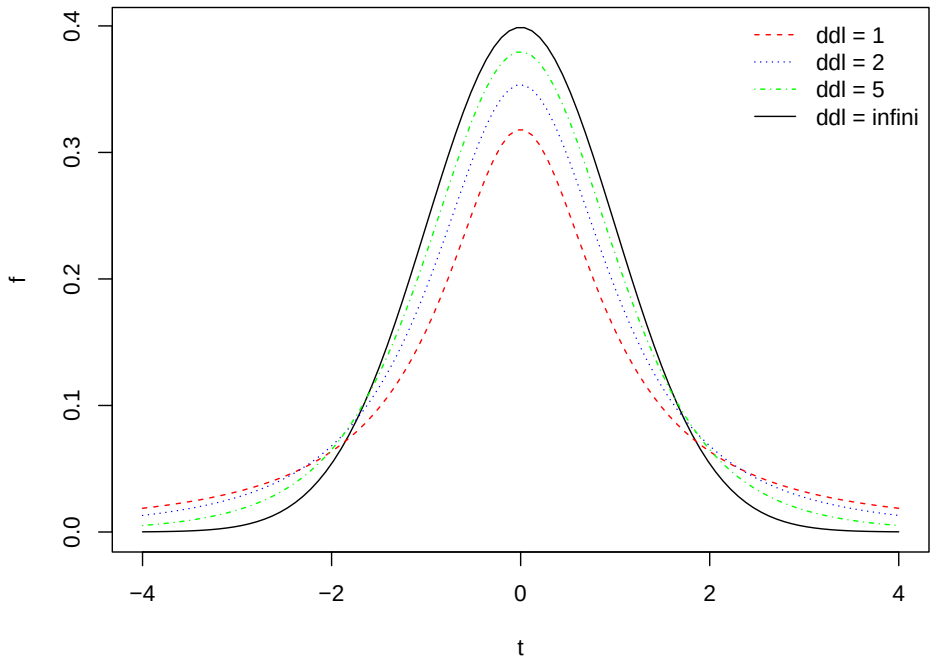


FIGURE 29 – Représentation de la loi de Student pour quelques degrés de liberté et de la loi normale vers laquelle elle tend pour un degré de liberté infini.

alpha	0.1000	0.0500	0.0100	0.0010	0.0001
Q	1.645	1.960	2.576	3.291	3.891

TABLE 4 – Quantiles à $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite N(0,1) pour quelques valeurs de α (cf. schéma associé Figure 30).

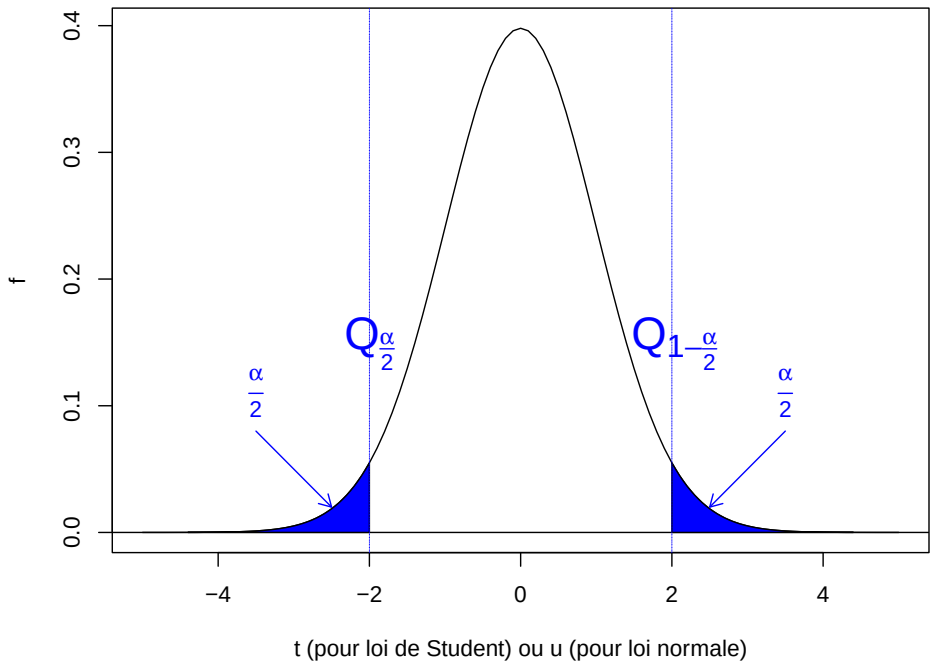


FIGURE 30 – Représentation schématique associée aux tables donnant les quantiles à $1 - \frac{\alpha}{2}$ pour la loi normale et les lois de Student.

alpha	0.1000	0.0500	0.0100	0.0010	0.0001
1	6.314	12.706	63.657	636.619	6366.198
2	2.920	4.303	9.925	31.599	99.992
3	2.353	3.182	5.841	12.924	28.000
4	2.132	2.776	4.604	8.610	15.544
5	2.015	2.571	4.032	6.869	11.178
6	1.943	2.447	3.707	5.959	9.082
7	1.895	2.365	3.499	5.408	7.885
8	1.860	2.306	3.355	5.041	7.120
9	1.833	2.262	3.250	4.781	6.594
10	1.812	2.228	3.169	4.587	6.211
11	1.796	2.201	3.106	4.437	5.921
12	1.782	2.179	3.055	4.318	5.694
13	1.771	2.160	3.012	4.221	5.513
14	1.761	2.145	2.977	4.140	5.363
15	1.753	2.131	2.947	4.073	5.239
16	1.746	2.120	2.921	4.015	5.134
17	1.740	2.110	2.898	3.965	5.044
18	1.734	2.101	2.878	3.922	4.966
19	1.729	2.093	2.861	3.883	4.897
20	1.725	2.086	2.845	3.850	4.837
21	1.721	2.080	2.831	3.819	4.784
22	1.717	2.074	2.819	3.792	4.736
23	1.714	2.069	2.807	3.768	4.693
24	1.711	2.064	2.797	3.745	4.654
25	1.708	2.060	2.787	3.725	4.619
26	1.706	2.056	2.779	3.707	4.587
27	1.703	2.052	2.771	3.690	4.558
28	1.701	2.048	2.763	3.674	4.530
29	1.699	2.045	2.756	3.659	4.506
30	1.697	2.042	2.750	3.646	4.482
infini	1.645	1.960	2.576	3.291	3.891

TABLE 5 – Quantiles à $1-\frac{\alpha}{2}$ de la loi de Student pour quelques valeurs de α (cf. schéma associé Figure 30) et des degrés de liberté croissant de 1 à 30 et leurs valeurs limites pour un degré de liberté infini (la loi tend alors vers loi normale).

Reprenons l'exemple de l'échantillon de 50 chats parmi lesquels 7 sont FIV positifs. Nous avions déduit auparavant de ces chiffres une estimation ponctuelle de 14% *de chats FIV positifs dans la population. Pour pouvoir utiliser la formule précédente de calcul d'un intervalle de confiance sur une fréquence, il faudrait au moins 20 chats FIV positifs. Imaginons qu'on ait eu par exemple 70 chats FIV positifs sur 500, on aurait pu utiliser cette formule, qui nous aurait donné* $[0.14-1.96\times\sqrt{\frac{0.14\times0.86}{500}};0.14+1.96\times\sqrt{\frac{0.14\times0.86}{500}}] = [0.14-0.03;0.14+0.03] = [11\%;17\%]$. *Sur notre exemple avec 7 chats sur 50 FIV positifs, comme les conditions d'utilisation ne sont pas respectées, on utilise un calcul plus complexe (on verra au second semestre comme la réaliser via le logiciel R sur ordinateur) qui utilise la loi binomiale sans approximation par la loi normale et donne l'intervalle de confiance* $[6\%;27\%]$.

Dans le chapitre 3.5 (de lecture optionnelle, pour les plus à l'aise avec le formalisme mathématique) nous verrons sur un exemple comment il est possible d'obtenir ce type de formules de calcul d'intervalle de confiance à partir de résultats de statistique théorique. **A ce stade assurez-vous surtout de comprendre ce que représente un intervalle de confiance** et de savoir le calculer à partir des formules données précédemment pour une fréquence et une moyenne, après avoir vérifié leurs conditions d'utilisation.

En terme d'interprétation d'un intervalle de confiance, étant donné qu'en pratique on calcule un intervalle de confiance sur un seul échantillon, **on n'a aucun moyen de savoir si cet intervalle de confiance contient bien la vraie valeur du paramètre. On peut juste se dire qu'en moyenne, lorsqu'on calcule des intervalles de confiance à** 95%**, on se trompe une fois sur 20 (**5% **des échantillons).**

Avant d'utiliser un intervalle de confiance, il est impératif d'en vérifier ses conditions utilisation, sans quoi on risque de raconter n'importe quoi. Il est très important de savoir juger, à partir d'un échantillon, du respect des conditions d'application du théorème de l'approximation normale. De très nombreux outils statistiques (estimateurs ponctuels et par intervalle, test statistiques) sont basés sur le théorème de l'approximation normale et nécessitent donc la vérification au préalable de ses conditions d'utilisation. **Il est IMPORTANT de se souvenir que la vérification des ces conditions d'utilisation, dans le cas d'une variable quantitative, ne peut pas se faire en regardant uniquement la taille de l'échantillon. L'examen de la distribution sera indispensable (nous nous y entraînerons en travaux dirigés).**

Dans l'exemple précédent (7 chats FIV positifs sur 50), si l'on avait appliqué à tort la formule précédente utilisant la loi normale, on aurait obtenu comme intervalle de confiance $[0.14-1.96\times\sqrt{\frac{0.14\times0.86}{50}};0.14+1.96\times\sqrt{\frac{0.14\times0.86}{50}}] = [0.14-0.096;0.14+0.096] = [4.4\%;23.6\%]$ *au lieu de* $[6\%;27\%]$. *Avec un effectif encore plus petit (ex. 2 chats FIV positifs sur 15, soit* 13.3% *en estimation ponctuelle, on aurait même obtenu l'intervalle* $[-3.9\%;30.5\%]$*, contenant des valeurs négatives, ce qui n'a bien entendu pas de sens pour l'estimation d'une proportion - n'hésitez pas à vérifier le calcul par vous-même, cela vous entraînera).*

Dans certains cas particuliers on définira des **intervalles de confiance unilatéraux** (une seule limite de confiance, cf Figure 31) ayant toujours une probabilité $1-\alpha$ de contenir la vraie valeur du paramètre. Voici deux exemple classiques d'utilisation d'un intervalle de confiance unilatéral :

— calcul du seuil au dessous duquel on veut pouvoir dire avec une confiance de 95% que se trouve une proportion d'animaux malades dans un pays (intervalle du type $[0, t]$),
— calcul du seuil au dessus duquel on veut pouvoir dire avec une confiance de 95% que se trouve la sensibilité d'un test diagnostique (intervalle du type $[t, 1]$).

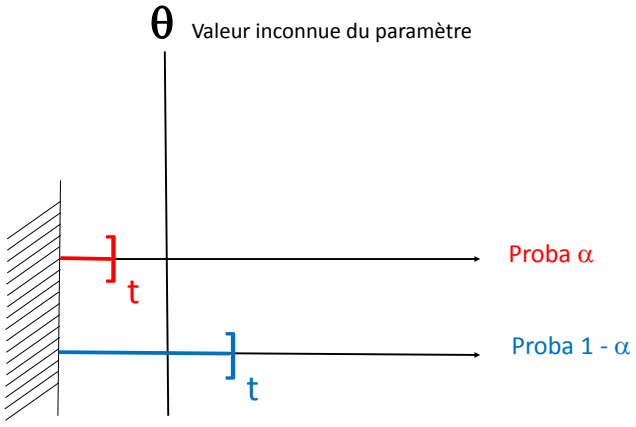


FIGURE 31 – Illustration de la définition d'un intervalle de confiance unilatéral dans le cas d'un intervalle de type $[0, t]$: $Pr(t \geq \theta) = \alpha$ donc $Pr(t \leq \theta) = 1 - \alpha$.

? ? ? A partir de la formule qui vous est donnée dans ce chapitre, calculez l'intervalle de confiance à 95% sur la température corporelle moyenne des chats sains, en supposant que la distribution de cette variable chez les chats sains est normale et qu'on a estimé sur un échantillon de 144 chats sains une moyenne à 39 degrés celsius et un écart type à 0.25 (on prendra les quantiles de la loi normale dans ce calcul en considérant que pour un degré de liberté de plus de 30 on peut approcher la loi de Student par la loi normale).
Refaites le même calcul d'intervalle de confiance à 95% sur la température corporelle moyenne des chats sains, en supposant cette fois que les mêmes moyenne et écart type estimé ont été obtenus sur 25 chats. Comparez cet intervalle de confiance au précédent.
Comparez ces deux intervalles à l'intervalle de fluctuation à 95% que vous aviez obtenu à partir des mêmes données observées dans la section 2.4.2. Ce dernier dépendait-il de l'effectif observé ?

ATTENTION à ne pas confondre la notion d'**intervalle de confiance** visant à quantifier l'**incertitude** associée à l'**estimation d'un paramètre statistique** à partir d'un échantillon (ex. moyenne, proportion) et l'**intervalle de fluctuation** visant à décrire l'intervalle dans lequel fluctuent la plus grande partie (95% si l'intervalle est calculé à 95%) des valeurs observées dans un échantillon (notion vue dans la section 2.4.2).

3.4 Test statistique

Partons d'un petit exemple introductif simple.

? ? ? Imaginez qu'on tire au sort aléatoirement n étudiants vétérinaires sur lesquels on estime la fréquence de filles. **A partir des données observées** (sans utiliser de connaissance *a priori*) **peut-on conclure que la fréquence de filles parmi les étudiants vétérinaires est différente de** 50%**, autrement dit que le ratio filles / garçons n'y est pas équilibré** ?
Imaginons 5 cas :

1. 2 filles sur $n = 2$: 100%
2. 6 filles sur $n = 10$: 60%
3. 15 filles sur $n = 20$: 75%
4. 37 filles sur $n = 50$: 74%
5. 68 filles sur $n = 100$: 68%

Tentez de répondre à la question dans chaque cas en utilisant juste votre intuition et votre bon sens et notez votre réponse pour y revenir plus tard.

Vous vous rendez compte sur cet exemple qu'il n'est pas toujours évident de répondre à ce type de question, et qu'il est certainement utile de prendre en considération à la fois la taille de l'échantillon et la valeur de la différence observée entre la fréquence de filles et la valeur théorique attendue de 50%. C'est l'objet des tests statistiques de fixer un cadre théorique associé à ce type de questionnement et d'y donner une réponse objective. Ce cadre n'étant pas des plus simple à bien comprendre, et faisant régulièrement l'objet d'abus d'interprétation, nous allons faire un petit retour 100 ans auparavant pour bien en comprendre l'historique.

3.4.1 Le test de signification tel que proposé par R.A. Fisher

Dans les années 1920, Ronald Aylmer Fisher a popularisé un concept proposé quelque temps auparavant par Karl Pearson, le concept de "p-value" ou en français "valeur de p", qui est depuis très couramment utilisé. Introduisons le en nous basant sur l'exemple précédent.

Nous définirons tout d'abord l'**hypothèse nulle** H_0 qui est l'**hypothèse de différence nulle**. Dans notre exemple on compare une fréquence observée (f fréquence de filles) à une valeur de référence de 50% et donc H_0 est l'hypothèse selon laquelle la proportion de filles parmi les étudiants vétérinaires est de 50% ($H_0 : \pi_0 = 0.5$). L'objectif du test de signification va être d'**évaluer si les données nous permettent de réfuter cette hypothèse**. Pour cela on va confronter les données à H_0 et se poser la question : "les données sont-elles probables sous H_0 ?" Autrement dit, "est-il probable, sous H_0, d'observer une telle différence, sous-entendu une différence aussi grande (en valeur absolue) ?"

La réponse à cette question est une probabilité qu'on appelle la "p-value" ou "valeur de p" en français. On lui donne aussi parfois le nom de degré de signification. **La p-value est donc la probabilité, si on est sous** H_0**, d'observer une différence au moins aussi grande que celle observée sur les données.** Si on veut formaliser la p-value dans le cas d'un test de signification d'une différence d (dans cet exemple la différence entre la fréquence de filles observée et 50%), elle correspond à $Pr(|d| > |d_{obs}||H_0)$.

Si p **est faible on rejette** H_0 et on en conclut qu'il existe bien une différence, sous entendu que la différence observée n'est pas uniquement due aux fluctuations d'échantillonnage mais est le reflet d'une

différence réelle dans la population. Il est devenu d'usage de fixer un seuil de 5% pour la valeur de p (à noter que R.A. Fisher n'avait pas proposé de seuil particulier) et si $p < 5\%$ de dire que la **différence est significative** mais cet usage est actuellement remis en cause par un grand nombre de statisticiens (cf. Wasserstein et al. (2019)), et vous comprendrez plus tard pourquoi.

A ce stade, je vous demande non pas d'apprendre par coeur une règle de décision sans vraiment la comprendre, mais de bien comprendre ce qu'est la **p-value**. C'est la réponse à la question : "sous l'hypothèse H_0 de différence nulle, quelle est la probabilité d'observer une différence au moins aussi grande (en valeur absolue) que celle que l'on a observé sur cet échantillon ?". Ou encore à retenir en plus court,
"sous l'hypothèse H_0 , quelle est la probabilité d'observer une telle différence ? ! ! !".
On comprend alors bien que plus la p-value est petite, plus on a d'arguments pour rejeter H_0 et donc conclure à une différence.

Gardons en tête que nous sommes dans le même cadre théorique que celui décrit en partie 3.3 où le paramètre d'intérêt (ici π_0 la proportion de filles dans la population d'étudiants vétérinaires) est supposé fixe mais inconnu car on n'a accès qu'à un échantillon de la population. Lorsque l'on parle de probabilité d'observer sous H_0 une différence au moins aussi grande que celle qu'on a observé sur notre unique échantillon, c'est sous-entendu "si on pouvait disposer de plein d'échantillons sous H_0". Le terme probabilité a donc ici le sens de fréquence d'occurences sur un grand nombre d'échantillons tirés au hasard dans la population sous l'hypothèse nulle.

Maintenant que vous avez compris le sens de la p-value, voyons sur ce cas de comparaison d'une fréquence observée f à une fréquence de référence (dite aussi fréquence théorique) π_0, comment il est possible de calculer la p-value à partir de ce qu'on appelle une **variable de décision**. Le **théorème de l'approximation normale**, s'il est applicable (N assez grand) nous dit : $F \sim N(\pi_0, \sqrt{\frac{\pi_0(1-\pi_0)}{N}})$.
Donc la variable centrée réduite $u = \frac{F-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{N}}} \sim N(0,1)$
On va utiliser u comme **variable de décision** :

- on **calcule la variable de décision u sur les données observées** : $u_{obs} = \frac{f-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{N}}}$, puis
- on **confronte u_{obs} à la loi qu'elle est censée suivre sous H_0 pour quantifier la p-value** (cf. Table 4).

Appliquons cette procédure à notre exemple dans le cas 3 (15 filles sur $n = 20$: $75\% \rightarrow u_{obs} = 2.24$). Visualisons la valeur la p-value (aire sous la courbe en bleu) dans la figure 32 qui représente la densité de probabilité de la loi $N(0,1)$ (loi attendue de la variable de décision u sous H_0). $u_{obs} = 2.24$ étant compris entre les quantiles de la loi normales (quantiles qu'on trouve dans la table 4, $u_{1-\frac{0.05}{2}} = 1.96$ et $u_{1-\frac{0.01}{2}} = 2.576$), on en déduit que $0.01 < p < 0.05$ et donc qu'on a assez d'arguments au vu des données pour réfuter H_0 si l'on utilise le seuil classiquement utilisé de 5%.

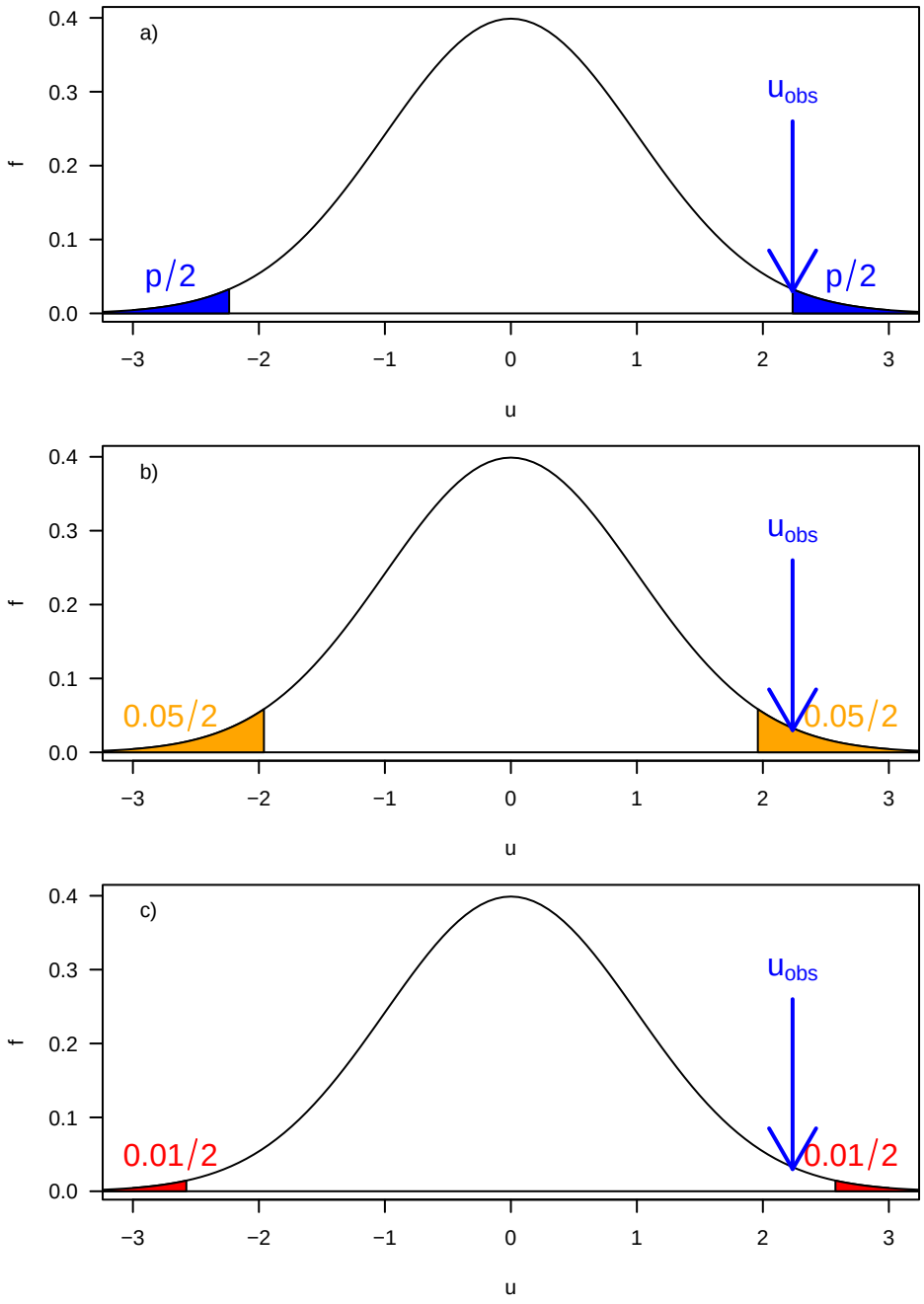


FIGURE 32 – Représentation a) de la p-value (aire sous la courbe en bleu) dans le cas de la comparaison d'une fréquence observée de 0.75 sur un effectif de 20, à une proportion théorique de 0.5 (ce qui donne $u_{obs} = 2.24$), et comparaison cette p-value à b) 5% (sachant que $u_{1-\frac{0.05}{2}} = 1.96$) et à c) 1% (sachant que $u_{1-\frac{0.01}{2}} = 2.576$).

? ? ? Si nous appliquons un test de signification aux autres cas sur le même exemple nous obtenons les résultats suivants.

- **Cas 1 : 2 filles sur** $n = 2$: 100%
  $\rightarrow p > 0.05$ (autre test adapté aux petits effectifs) $\rightarrow$ non rejet de H_0
- **Cas 2 : 6 filles sur** $n = 10$: 60%
  $\rightarrow p > 0.05$ (autre test adapté aux petits effectifs) $\rightarrow$ non rejet de H_0
- **Cas 3 : 15 filles sur** $n = 20$: 75%
  $\rightarrow u_{obs} = 2.24 \rightarrow 0.01 < p < 0.05 \rightarrow$ rejet de H_0
- **Cas 4 : 37 filles sur** $n = 50$: 74%
  $\rightarrow u_{obs} = 3.39 \rightarrow 0.0001 < p < 0.001 \rightarrow$ rejet de H_0
- **Cas 5 : 68 filles sur** $n = 100$: 68%
  $\rightarrow u_{obs} = 3.60 \rightarrow 0.0001 < p < 0.001 \rightarrow$ rejet de H_0

A partir de ces exemples, et avant de lire la suite, je vais vous demander de faire quelques petits exercices de calcul et de réflexion en suivant les points ci-dessous :

- refaites vous même les calculs pour les cas 3, 4 et 5 afin de vérifier que vous avez bien compris le principe et la procédure d'encadrement de la p-value,
- pour chacun des cas, comparez la conclusion que l'on obtient en terme de rejet ou non de H_0 (avec le seuil classique de 5%) avec la décision que vous aviez donnée en utilisant votre intuition en début de chapitre. Avez-vous tendance, dans des cas comme ceux-ci, à rejeter plus facilement ou plus difficilement H_0 que la procédure classique que l'on vient de présenter ?
- en vous basant notamment sur les deux premiers cas, pensez-vous qu'on peut tirer une conclusion d'une p-value élevée, supérieure à 5% ?

Alors peut-on accepter H_0 lorsque p est élevé ? J'espère que vous êtes convaincu, notamment à partir des cas 1 et 2 de l'exemple que non. Voici une citation de R.A. Fisher à ce sujet : *"The null hypothesis is never proved or established, but it is possibly disproved, in the course of experimentation"*, autrement dit **un test de signification peut conduire à rejeter H_0 dans certains cas, mais en aucun cas à l'accepter.**

A RETENIR !

Objectif du test de signification : déterminer si une différence observée est le reflet d'une vraie différence et non le simple reflet des fluctuations d'échantillonnage)

Principe :

- on fait l'hypothèse d'une différence nulle (H_0),
- à l'aide d'une variable de décision on calcule la p-value (p), *i.e.* la probabilité d'observer, sous H_0, une différence au moins aussi grande que celle observée sur l'échantillon,
- si p est faible on rejette H_0 (il est d'usage si $p < 0.05$ de dire que la différence est significative),
- plus p est petit, plus on a d'arguments pour rejeter H_0.
- **On ne peut jamais accepter H_0 seulement à partir de la p-value**

3.4.2 Le test d'hypothèse tel que modifié par E.Pearson et J. Neyman

Une deuxième vision a été proposée par Jerzy Neyman et Egon Pearson en 1928 et présentée comme une amélioration du test de signigication. Ils ont proposé l'utilisation de la p-value pour un **test d'hypothèse**, outil décisionnel permettant de choisir entre deux hypothèses, l'hypothèse nulle H_0 l'hypothèse alternative H_1 de différence non nulle. Le principe est de choisir H_0 si $p > 0.05$ et H_1 si $p < 0.05$.

Dans une telle procédure on a deux risques d'erreur :

— **le risque de 1ère espèce α maîtrisé** ($\alpha = 0.05$), risque de se tromper en rejetant H_0 et

— **le risque de 2ème espèce β non maîtrisé**, risque de se tromper en acceptant H_0

On appelle $1 - \beta$ la **puissance du test**. La figure 33 illustre les concepts de risque α et β et montre comment, pour un α donné, fixé dans la figure classiquement à 5%, β dépend de H_1, donc de la vraie différence. Plus elle grande et plus β est petit, donc plus le test est puissant.

Pour illustrer cette fois l'impact de l'effectif sur la puissance d'un test (en reprenant l'exemple du test de comparaison de la fréquence de filles à 50%), nous avons réalisé des simulations de 1000 échantillons d'étudiants vétérinaires (en faisant varier la taille des échantillons) en supposant que la proportion de filles dans cette population est de 70% et noté le nombre de rejets de H_0.

— sur 1000 échantillons de taille 10 : 161 rejets de H_0

— sur 1000 échantillons de taille 20 : 415 rejets de H_0

— sur 1000 échantillons de taille 50 : 784 rejets de H_0

— sur 1000 échantillons de taille 100 : 977 rejets de H_0

Comme on pouvait aisément l'imaginer, à différence théorique fixée, plus l'effectif est grand, plus la puissance est grande.

Dans le cadre de la réalisation d'un test d'hypothèse, on souhaite avoir un **risque** β faible donc une **puissance** $1 - \beta$ forte, mais β peut être élevé (donc la puissance faible) du fait :

— d'une faible différence théorique (H_1 proche de H_0),

— d'une grande incertitude sur le paramètre estimé (faible effectif, forte variabilité si on travaille sur une variable quantitative continue)

On ne peut raisonnablement utiliser un test d'hypothèse **que si la puissance est maîtrisée** donc si un calcul de puissance *a priori* a été réalisé :

Il s'agit d'un calcul a priori d'effectifs nécessaires pour atteindre une puissance donnée, c'est-à-dire une probabilité donnée de détecter une différence dépassant un seuil d'intérêt prédéfini (nous verrons comment faire ce type de calcul au second semestre lors des travaux dirigés sur ordinateurs).

Néanmoins il est important de savoir ce qu'écrivait R.A. Fisher dans une lettre au magazine Nature en 1935 au sujet du test d'hypothèse et du risque de deuxième espèce :

"Errors of the second kind are committed only by those who misunderstand the nature and the application of tests of significance"

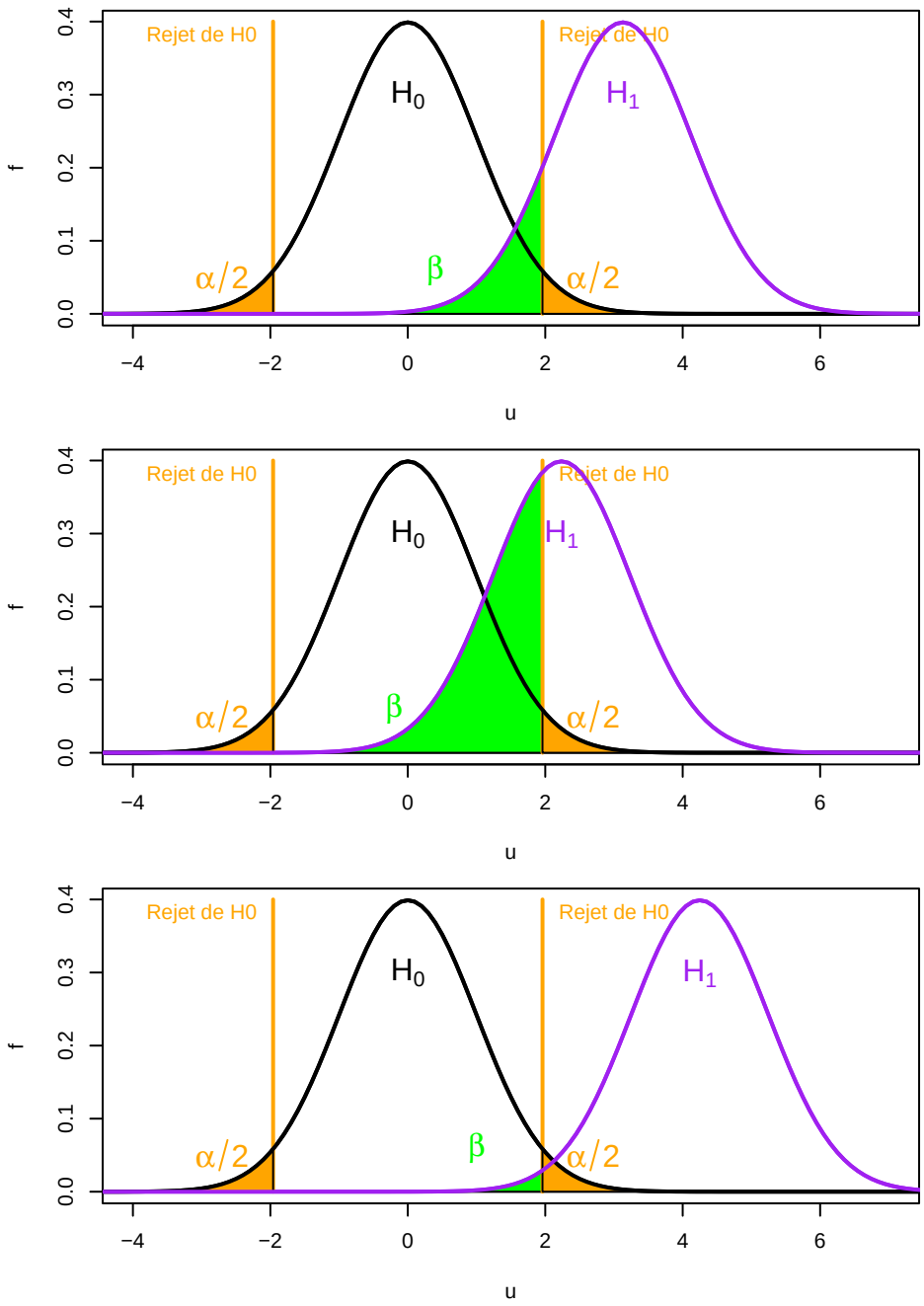


FIGURE 33 – Représentation des risques de première et deuxième espèce pour différentes hypothèses alternatives H_1

3.4.3 L'utilisation raisonnée des tests statistiques recommandée aujourd'hui

On comprend bien à la lecture de la citation précédente de R.A. Fisher, la discorde profonde entre lui d'un côté et J. Neyman et E. Pearson de l'autre. R.A. Fisher craignait sans doute, à raison, que les tests soient utilisés de façon abusive pour valider une hypothèse en l'absence de preuve permettant de la réfuter, sans s'être donné les moyens d'obtenir suffisamment de preuves (via un calcul de puissance *a priori*). Par analogie avec ce qui se passe lors d'un procès, accepter H_0 à partir de peu de données, juste parce que la p-value est supérieure à 5%, c'est comme relaxer un inculpé après un jour d'enquête juste parce qu'on n'a pas trouvé de preuve de sa culpabilité durant cette enquête minimale.

La mauvaise utilisation des tests, venant en partie de l'amalgame entre les tests de signification (version R.A. Fisher) et les tests d'hypothèse (version J.Neyman et E. Pearson), est trop fréquente et a fait couler

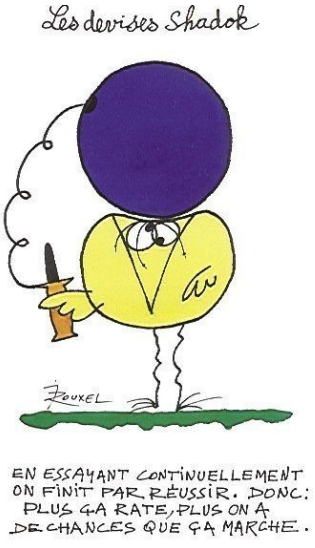
beaucoup d'encre dans tous les domaines d'application des statistiques depuis des décennies, avec des titres comme "Faut-il brûler les tests de signification statistique ?", "What Your Statistician Never Told You about P-Values ?", "A Dirty Dozen : Twelve P-Value Misconceptions". Le débat est même abordé dans les revues de vulgarisation, comme dans l'article publié dans la revue *Pour la science* en avril 2020 et intitulé "La valeur-p : un problème siginificatif. Les méthodes statistiques pour jauger la pertinence d'un résultat expérimental sont attaquées de toute part. Résisteront-elles ?"

Pour un bon usage des tests à l'heure actuelle, on peut s'appuyer sur un article publié en 2016 dans le journal *The American Statistician* (Wasserstein and Lazar (2016)), référence très citée et qui fait consensus auprès des statisticiens. Cette référence décline les six points suivants :

1. **P-values can indicate how compatible the data are with a specified statistical model."** En effet plus la valeur de p est petite et plus l'incompatibilité statistique entre les données et l'hypothèse nulle est grande. On peut voir la **valeur de p comme un indicateur de discordance entre les données et l'hypothèse nulle**.
2. **"P-values do not measure the probability that the studied hypothesis is true."** C'est-à-dire que la valeur de p ne doit surtout pas être interprétée comme la probabilité de l'hypothèse nulle connaissant les données, même si cela est très tentant. On ne peut pas inverser les probabilités aussi facilement !

   Si un jour vous en êtes tenté pensez au cas 1 de notre exemple (avec deux filles sur un échantillon aléatoire de deux étudiants vétérinaires en concluriez-vous que la probabilité pour qu'il y ait autant de filles que de garçons parmi les étudiants vétérinaires est forte ?)
3. **"Scientific conclusions and decisions should not be based only on whether a p-value passes a specific threshold."** Actuellement les scientifiques donnent souvent trop de poids à la valeur de p et au résultat du test en terme de différence significative ou non, parfois sans même regarder la différence estimée. Il convient plutôt de considérer le **test juste comme un garde fou, nous empêchant d'interpréter hâtivement une différence qui ne serait pas significative**.
4. **"Proper inference requires full reporting and transparency."**

   — Les résultats de **tous les tests réalisés doivent être reportés**, et non les seuls résultats significatifs.
   — En moyenne dans tous les cas où H_0 est vraie, une fois sur 20 on a $p < 0.05$. *A force de chercher on finit par trouver !*



5. **"A p-value does not measure the size of an effect or the importance of a result."** Une valeur de p petite n'implique pas forcément la mise en évidence d'une différence d'intérêt

biologique : une différence importante peut ne pas apparaître significative du fait du manque de puissance de l'analyse (par ex. en cas d'effectifs faibles). Et inversement, une différence biologique importante peut ne pas apparaître significative du fait de faibles effectifs. Il est donc capital, dans tous les cas, **d'interpréter in fine l'estimation de la différence (estimation ponctuelle et intervalle de confiance lorsque celui-ci est calculable)**.

6. **"By itself, a p-value dose not provide a good measure of evidence regarding a hypothesis."**
   On ne doit jamais utiliser un test d'hypothèse pour montrer une hypothèse et en particulier pour montrer une équivalence mais privilégier les tests d'équivalence basés sur les intervalles de confiance dans ce cas.

   Le principe des tests d'équivalence (cf. Figure 34) est de définir *a priori* une zone d'équivalence, sur des critères biologiques ("quelle différence maximum sera considérée comme négligeable ?") puis de conclure à l'équivalence si l'intervalle de confiance sur la différence observée est entièrement contenu dans cette zone.

? ? ? Pour voir si vous avez bien compris la différence entre test d'équivalence et test d'hypothèse, essayez de répondre aux questions posées dans l'exemple suivant.
Une étude comparative de deux produits traitant l'otite externe du chien a été réalisée sur 140 chiens atteints par cette pathologie. Les deux produits comparés associent antibiotique et anti-inflammatoire. L'un des deux produits est un traitement de référence considéré comme efficace mais pouvant entraîner des effets secondaires. Le but de l'étude est de montrer que le nouveau produit est aussi efficace que le produit de référence (par un test d'équivalence). Sur ce type de traitement l'équivalence a été définie auparavant par une différence entre les 2 fréquences de guérison inférieure à 0.2 (20%). A l'issue de l'étude randomisée la fréquence de guérison observée était de 58.3% avec le nouveau produit et de 41.2%, avec une différence estimée de 17.1% et un intervalle de confiance à 95% sur cette différence de $[-0.6\%; 34.9\%]$.

- D'après ce résultat peut-on conclure à l'équivalence d'efficacité entre les deux traitements ?
- En admettant que la différence est significative si et seulement si son intervalle de confiance à 95% ne contient pas la valeur 0, conclurait-on ici à une différence significative d'efficacité entre les deux produits ?

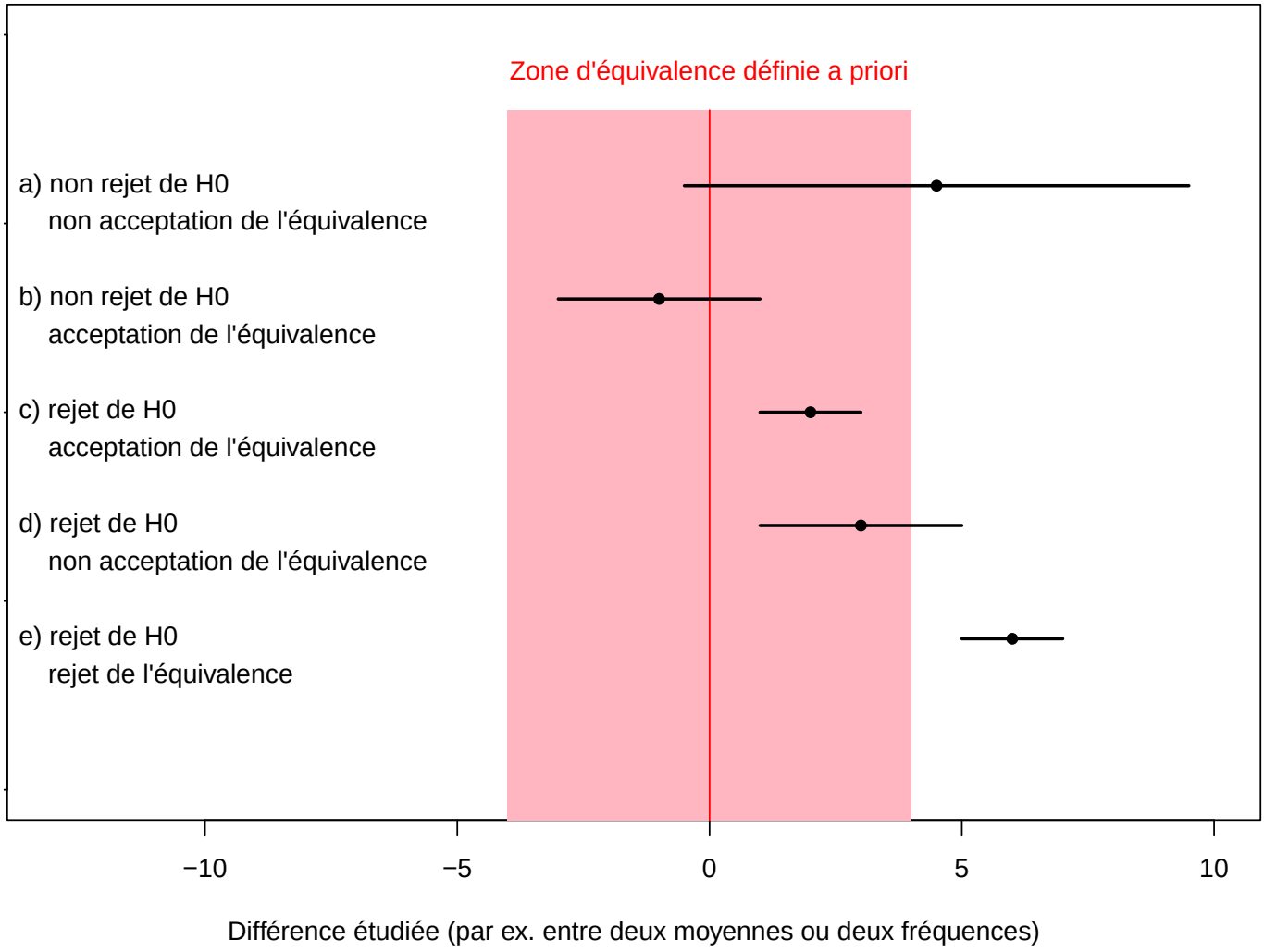


FIGURE 34 – Illustration du principe du test d'équivalence et conclusions comparées des tests de signification et d'équivalence sur 5 exemples. Nous montrerons sur un exemple en travaux dirigés qu'on peut rejeter l'hypothèse de différence si et seulement si l'intervalle de confiance sur la différence ne contient pas la valeur 0.

En 2019, un numéro spécial de la même revue, *The American statistician*, présente les points de vue d'un grand nombre de statisticiens de renom au sujet de l'utilisation et/ou de l'abandon des tests statistiques (Wasserstein et al. (2019)). Sans bannir complètement l'utilisation de la p-value, ils préconisent l'abandon pur et simple des terminologies "différence significative", "statistiquement significatif", ... Il m'est difficile pour le moment de les bannir complètement de mon enseignement car elles sont encore beaucoup utilisées, mais j'espère pouvoir le faire un jour. Nous verrons comment les façons de faire évolueront dans les années à venir.

Ce chapitre est de loin le plus compliqué du cours de biostatistique d'un point de vue conceptuel et il est capital que vous ayez compris toutes les notions qui y sont abordées. En ce qui concerne l'utilisation très controversée des tests de signification ou d'hypothèse, nous nous en tiendrons à la présentation qui en a été faite par R.A. Fisher, en prenant soin de ne jamais conclure à l'acceptation de H_0 à partir de la p-value, et, à chaque fois que cela est possible, nous compléterons le calcul de la p-value par le calcul de l'estimation ponctuelle et par intervalle de l'effet étudié (ex. différence entre deux moyennes ou deux fréquences).

3.5 Pour aller plus loi sur les notions d'intervalle de confiance et de test - chapitre de lecture optionnelle

Cette partie est de lecture optionnelle. Elle est destinée aux étudiants que le formalisme mathématique ne rebute pas, pour les aider à comprendre comment on construit un intervalle de confiance et un test et comment ces deux concepts sont reliés.

3.5.1 Exemple support

Partons d'un exemple qui nous servira de support. Un essai clinique a été réalisé pour évaluer l'effet d'un traitement oral (le telmisartan, vendu sous le nom de Semintra) sur l'hypertension féline. Cet essai a consisté tout d'abord à sélectionner des chats avec une hypertension modérée à sévère mais pas trop (pression artérielle systolique entre 160 et 200 mmHg, mmHg symbolisant l'unité millimètre de mercure). Ces chats ont ensuite été répartis dans deux groupes par randomisation (répartition alétoire entre les deux groupes), le groupe P recevant un placebo et le groupe S recevant le traitement Semintra. Pour chaque chat, la baisse de pression artérielle entre le niveau de base (mesuré à J0 avant traitement) et le niveau mesuré après 28 jours de traitement (placebo ou Semintra) a été calculée. Cette baisse de pression artérielle est considérée comme la variable d'intérêt pour juger de l'efficacité du traitement et sera appelée dans la suite BPA (pour baisse de pression artérielle).

Imaginons que l'essai ait été réalisé sur 10 chats dans chaque groupe, et que l'on ait obtenu dans le groupe S une moyenne à 25 mmHg avec un écart type estimé de 17 mmHg, et sur le groupe P une moyenne à 11 mmHg avec un écart type à 16 mmHg. Imaginons aussi pour la suite que la distribution de notre variable BPA soit normale dans chacun des groupes.

Construction de l'intervalle de confiance autour de la différence entre deux moyennes observées sur deux séries indépendantes

On considère deux échantillons indépendants de tailles n_1 et n_2 sur lesquels ont été mesurées les valeurs d'une variable quantitative continue X. On suppose que X est distribuée normalement dans les deux populations dont ont été tirés les deux échantillons, avec une variance commune $\sigma_1^2 = \sigma_2^2 = \sigma^2$ et des moyennes respectives μ_1 et μ_2. On peut alors montrer que la variable $T = \frac{\overline{X_1}-\overline{X_2}-(\mu_1-\mu_2)}{\widehat{\sigma}\times\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ suit la loi de Student de n_1+n_2-2 degrés de liberté $T(n_1+n_2-2)$ (avec $\widehat{\sigma} = \sqrt{\frac{(n_1-1)\times\widehat{\sigma_1^2}+(n_2-1)\times\widehat{\sigma_2^2}}{n_1+n_2-2}}$ l'estimation de l'écart type commun.).

A partir de ce résultat théorique on peut obtenir une formule de calcul de l'intervalle de confiance bilatéral à 95% autour de la différence entre les 2 moyennes observées. Un intervalle de confiance $[t_1; t_2]$ autour de $\overline{X_1} - \overline{X_2}$ doit avoir une probabilité $1-\alpha$ de contenir la vraie valeur de la différence $\mu_1 - \mu_2$ et plus précisément doit vérifier :

$$Pr(t_1 \geq \mu_1 - \mu_2) = Pr(t_2 \leq \mu_1 - \mu_2) = \frac{\alpha}{2} \tag{1}$$

Pour rappel voir l'illustration associée à cette définition dans le chapitre 3.3.2.

Comme T suit la loi de Student de degré de liberté $\nu = n_1 + n_2 - 2$, pour une valeur de α donnée, la table de la loi de Student (Table 5) nous donne le quantile $t_{\nu,1-\frac{\alpha}{2}}$ tel que :

$$Pr(\frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\widehat{\sigma} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \geq t_{\nu,1-\frac{\alpha}{2}}) = Pr(\frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\widehat{\sigma} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq -t_{\nu,1-\frac{\alpha}{2}}) = \frac{\alpha}{2} \tag{2}$$

que l'on peut aussi écrire

$$Pr(\overline{X_1}-\overline{X_2}-t_{n_1+n_2-2;1-\frac{\alpha}{2}}\times\widehat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \geq \mu_1-\mu_2) = Pr(\overline{X_1}-\overline{X_2}+t_{n_1+n_2-2;1-\frac{\alpha}{2}}\times\widehat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq \mu_1-\mu_2) = \frac{\alpha}{2} \tag{3}$$

Cette dernière écriture permet d'identifier facilement les bornes de l'intervalle de confiance sur $\mu_1 - \mu_2$ et de l'écrire sous la forme classique (forme que vous retrouverez dans le chapitre 5.2.1).

$$\mu_1 - \mu_2 = \overline{X_1} - \overline{X_2} \pm t_{n_1+n_2-2;1-\frac{\alpha}{2}} \times \widehat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{4}$$

Sur notre exemple, à partir de cette formule et en supposant les variances égales dans les 2 groupes comparés (ou écarts types égaux, ce qui revient au même, mais est plus facile à évaluer : ici les écarts types estimés sont proches donc cette hypothèse semble raisonnable), on peut calculer la différence entre les deux moyennes observées de la variable BPA sur les 2 échantillons, et l'intervalle de confiance à 95% associé à cette différence.

$\overline{X_1} - \overline{X_2} = 14$

$\widehat{\sigma} = \sqrt{\frac{9\times(17^2+16^2)}{18}} = 16.51$

ce qui donne pour l'intervalle : $14 \pm 2.101 \times 16.51 \times \sqrt{\frac{2}{10}} = 14 \pm 15.51$ c'est-à-dire [-1.51 ; 29.51].

On peut noter que le calcul de $\widehat{\sigma}$ peut se simplifier dans les cas où $n_1 = n_2$ et qu'il ne dépend alors plus de l'effectif. On a alors $\widehat{\sigma} = \sqrt{\frac{(n_1-1)\times\widehat{\sigma_1^2}+(n_2-1)\times\widehat{\sigma_2^2}}{n_1+n_2-2}} = \sqrt{\frac{\widehat{\sigma_1^2}+\widehat{\sigma_2^2}}{2}}$.

Construction du test de comparaison de deux moyennes observées sur deux séries indépendantes

Pour réaliser maintenant un test de comparaison des 2 moyennes observées, on peut travailler sur variable de décision suivante :

$T = \frac{\overline{X_1} - \overline{X_2}}{\widehat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ qui suit la loi $T(n_1 + n_2 - 2)$ sous H_0 définie par $\mu_1 = \mu_2$.

Pour les données de l'exemple, nous allons visualiser la valeur observée de cette variable de décision, la valeur du degré de signification p ainsi que la région de rejet de H_0 (cf. Figure 35 et explications associées ci-dessous). Si l'on veut formaliser la p-value sur cet exemple elle correspond à $Pr(|t| > |t_{obs}||H_0)$. Sur les données on obtient $t_{obs} = \frac{14}{16.51 \times \sqrt{\frac{2}{10}}} = 1.896$. Or d'après la table de la loi de Student (Table 5), pour un degré de liberté ν de 18, le quantile $t_{\nu,1-\frac{0.05}{2}} = 2.101$ est plus éloigné du centre de la distribution que t_{obs} (cf. figure 35), donc l'aire sous la courbe correspondant à la p-value est supérieure à 0.05. On ne peut donc pas rejeter H_0 au risque 5%. La différence observée est donc dite non significative.

D'après la table de Student (Table 5), pour un degré de liberté ν de 18, on peut même donner un peu plus d'information sur la p-value. Du fait que 1.896 est entre les quantiles de la loi de Student $t_{\nu,1-\frac{0.05}{2}} = 2.101$ et $t_{\nu,1-\frac{0.1}{2}} = 1.734$, on peut dire que $0.05 < p < 0.1$.

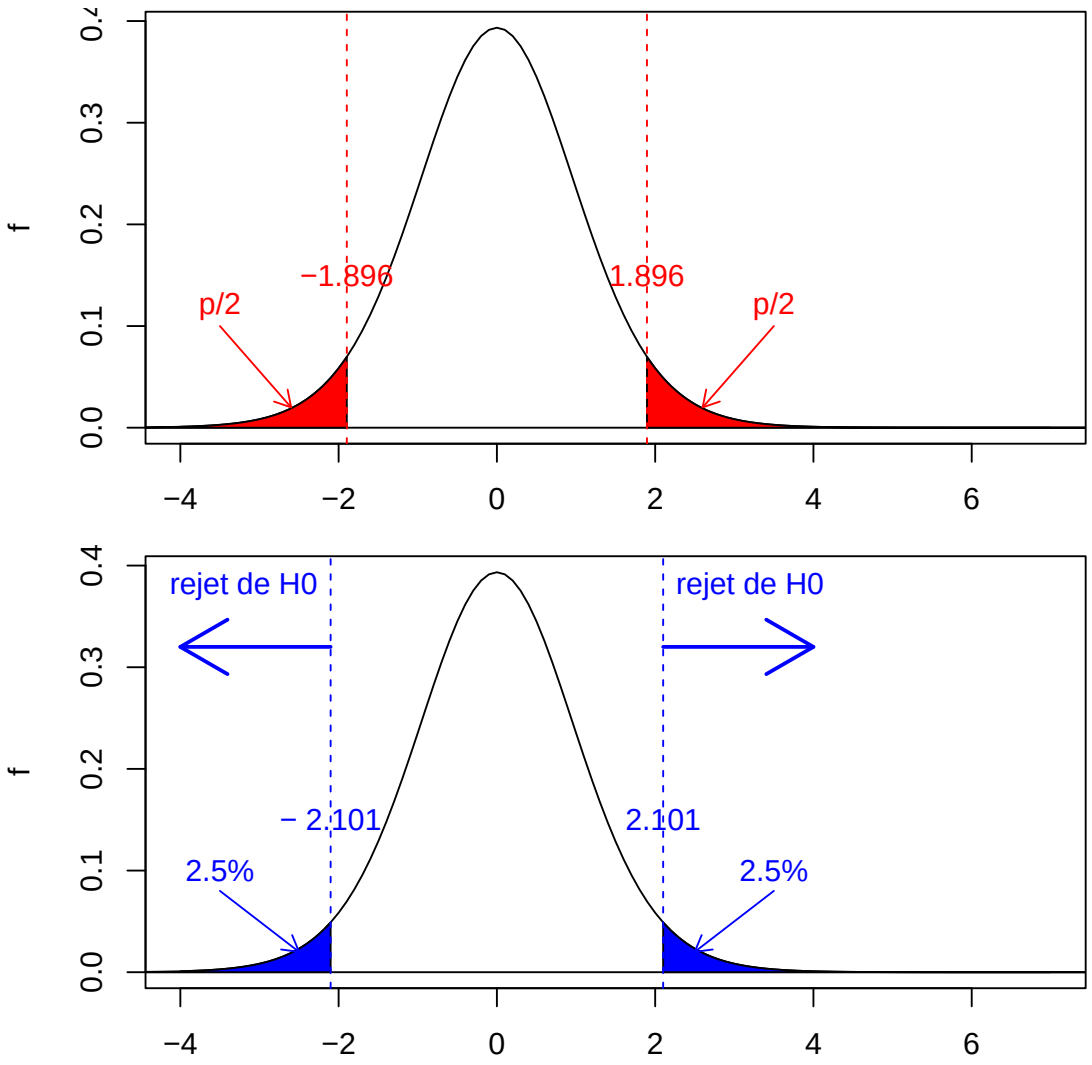


FIGURE 35 – Représentation de la p-value et de la zone de rejet de H_0 sur l'exemple support

Lien entre test de signification et intervalle de confiance

Pouvait-on déjà donner cette conclusion à partir des résultats de la première question ? Oui on peut ici démontrer qu'on pouvait donner cette conclusion à partir de l'intervalle de confiance, du fait qu'il contenait la valeur 0 correspondant à l'hypothèse nulle. Plus généralement il est équivalent de dire que l'intervalle de confiance à 95% ne contient pas la valeur 0 et de dire que la différence est significative au risque de première espèce de 5%.

En voici la démonstration : Rejet de H_0 au risque $\alpha \iff$
$\frac{\overline{X_1}-\overline{X_2}}{\widehat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \geq t_{\nu,1-\frac{\alpha}{2}}$ ou $\frac{\overline{X_1}-\overline{X_2}}{\widehat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leq -t_{\nu,1-\frac{\alpha}{2}} \iff$
$\overline{X_1} - \overline{X_2} - t_{\nu,1-\frac{\alpha}{2}} \times \widehat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \geq 0$ ou $\overline{X_1} - \overline{X_2} + t_{\nu,1-\frac{\alpha}{2}} \times \widehat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq 0 \iff$
$t_1 \geq 0$ ou $t_2 \leq 0$ (cf. Figure 36) $\iff$
l'intervalle de confiance au seuil de confiance $1-\alpha$ ne contient pas 0

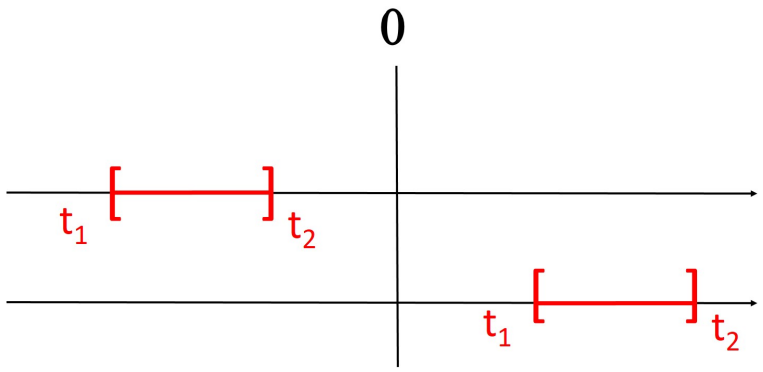


FIGURE 36 – Représentation des conditions sur l'intervalle de confiance dans le cas du rejet de H_0

4 Comparaison de fréquences et de distributions d'une variable qualitative, ou méthodes permettant de corréler deux variables qualitatives

4.1 Objectifs pédagogiques

A l'issue de l'étude de ce chapitre et de la réalisation du deuxième TD de S3, vous devriez :

- A partir de la description de séries d'observations et de leur processus d'acquisition, savoir dire si les séries sont indépendantes ou dépendantes (c'est-à-dire appariées dans le cas de deux séries).
- Savoir repérer dans quels cas on doit utiliser un test du χ^2 d'ajustement, un test du χ^2 d'indépendance, un test de Mc Nemar, et un test de Cochran-Mantel-Haenszel.
- Savoir vérifier les conditions d'utilisation de ces tests et en interpréter les résultats.
- Savoir réaliser à la main les tests du χ^2 (ajustement et indépendance) et le test de McNemar.

4.2 Les tests du χ^2

4.2.1 Le test du χ^2 d'ajustement

Prenons l'exemple d'un échantillon aléatoire de 15 étudiants vétérinaires sur lequel on compte 4 garçons et 11 filles. Peut-on dire à partir de cet échantilllon qu'il y a plus de filles que de garçons dans la population des étudiants vétérinaires ? Autrement dit peut-on rejeter l'hypothèse selon laquelle la proportion de filles est égale à 50% ?

Pour répondre à cette question nous pouvons utiliser la variable de décision présentée précédemment au chapitre 3.4.1 ou la statistique du χ^2 en réalisant la procédure suivante :

Procédure de réalisation d'un test du χ^2 d'ajustement

1. **Calcul des effectifs théoriques sous H_0** : " différence nulle c'est-à-dire $\pi_{filles} = 0.5$"
   - **Effectifs observés** notés O_i : filles 11, garçons 4
   - **Effectifs théoriques** (attendus sous H_0) **notés** C_i : filles 7.5, garçons 7.5

2. **Calcul de la variable de décision suivante de comparaison des O_i et des C_i :**

$$\chi^2 = \Sigma_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i} = (\Sigma_{i=1}^{k} \frac{O_i^2}{C_i}) - N$$

   avec
   - k le nombre de classes de la variable qualitative (ici 2, filles et garçons),
   - N le nombre total d'observations (ici 15).

   Dans notre exemple la valeur observée est $\chi^2_{obs} = 3.267$.

3. **Calcul de la valeur de p**

   Si les conditions d'utilisation du test du χ^2 d'ajustement sont respectées, c'est-à-dire si tous les C_i sont supérieurs à 5 (c'est le cas ici), on peut considérer que la variable de décision suit à peu près la **loi du χ^2 de degré de liberté** $k - 1$ (cf. Table 6 et Figure 37).

   Au vu de la loi du χ^2 à un degré de liberté et de la valeur observée $\chi^2_{obs} = 3.267$ (Figure 38 et Table 6), la p-value est supérieure à 5% (entre 5% et 10%) et on ne peut donc pas conclure à une proportion plus importante de filles à partir de ce seul échantillon.

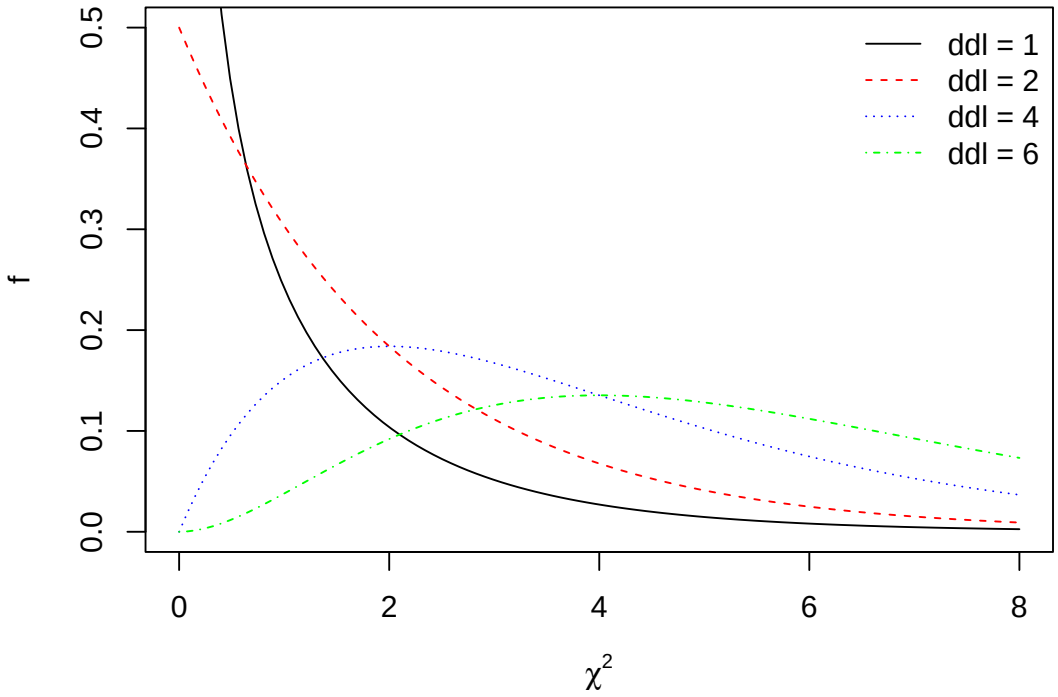


FIGURE 37 – Représentation de la loi du χ^2 pour quelques degrés liberté

alpha	0.1000	0.0500	0.0100	0.0010	0.0001
1	2.706	3.841	6.635	10.828	15.137
2	4.605	5.991	9.210	13.816	18.421
3	6.251	7.815	11.345	16.266	21.108
4	7.779	9.488	13.277	18.467	23.513
5	9.236	11.070	15.086	20.515	25.745
6	10.645	12.592	16.812	22.458	27.856
7	12.017	14.067	18.475	24.322	29.878
8	13.362	15.507	20.090	26.124	31.828
9	14.684	16.919	21.666	27.877	33.720
10	15.987	18.307	23.209	29.588	35.564
11	17.275	19.675	24.725	31.264	37.367
12	18.549	21.026	26.217	32.909	39.134
13	19.812	22.362	27.688	34.528	40.871
14	21.064	23.685	29.141	36.123	42.579
15	22.307	24.996	30.578	37.697	44.263
16	23.542	26.296	32.000	39.252	45.925
17	24.769	27.587	33.409	40.790	47.566
18	25.989	28.869	34.805	42.312	49.189
19	27.204	30.144	36.191	43.820	50.795
20	28.412	31.410	37.566	45.315	52.386
21	29.615	32.671	38.932	46.797	53.962
22	30.813	33.924	40.289	48.268	55.525
23	32.007	35.172	41.638	49.728	57.075
24	33.196	36.415	42.980	51.179	58.613
25	34.382	37.652	44.314	52.620	60.140
26	35.563	38.885	45.642	54.052	61.657
27	36.741	40.113	46.963	55.476	63.164
28	37.916	41.337	48.278	56.892	64.662
29	39.087	42.557	49.588	58.301	66.152
30	40.256	43.773	50.892	59.703	67.633

TABLE 6 – Quantiles à $1-\alpha$ de la loi du χ^2 pour quelques valeurs de α et des degrés de liberté croissant de 1 à 30.

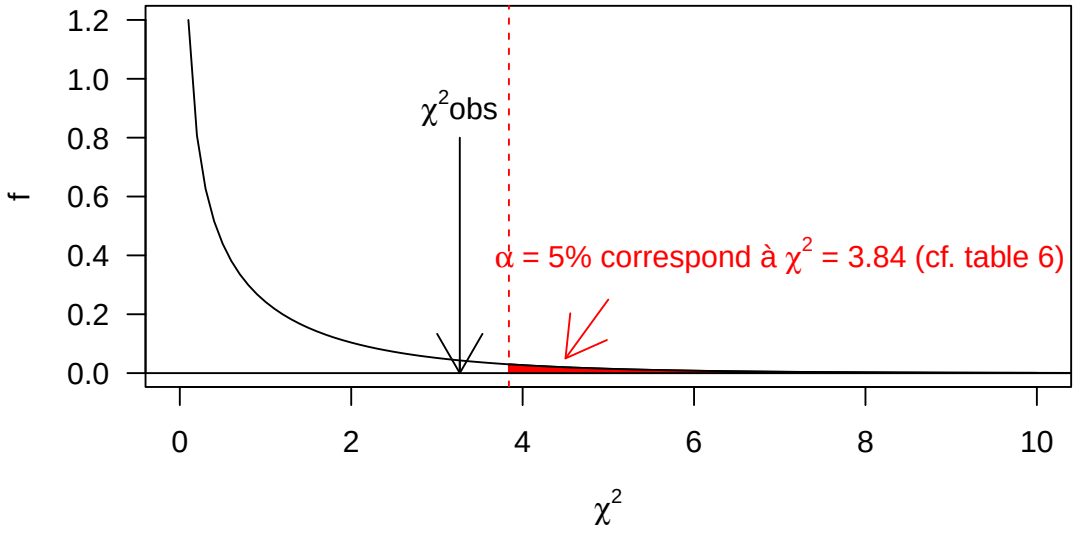


FIGURE 38 – Illustration de l'estimation de la p-value pour un test du χ^2 d'ajustement utilisé dans l'exemple pour la comparaison d'une fréquence observée à une fréquence théorique (à partir de la loi du χ^2 à un degré de liberté).

Ce même test du χ^2 d'ajustement peut être utilisé non seulement pour comparer une fréquence observée à une fréquence théorique (dans ce cas il est équivalent au test utilisant la loi normale présentée au chapitre 3.4.1) mais aussi la distribution observée d'une variable qualitative à k classes (quel que soit la valeur de k), à une distribution théorique.

? ? ? Pour voir si vous avez bien compris le principe du test du χ^2 d'ajustement, essayez de l'appliquer à l'exemple suivant :

On cherche à savoir si une maladie donnée est liée au groupe sanguin, autrement dit si certains groupes sanguins sont plus touchés que d'autres. Sur 200 malades observés, on a dénombré 104 sujets du groupe O, 76 du groupe A, 18 du groupe B et 2 du groupe AB. On admettra que dans la population générale la répartition entre les groupes est : 47% *de O,* 43% *de A,* 7% *de B et* 3% *de AB. Peut-on dire à partir de cette observation que la répartition des groupes sanguins est différente chez les sujets malades et chez les sujets sains. Il s'agit bien de comparer une distribution observée sur un échantillon, ici la distribution des groupes sanguins chez les malades, à une distribution de référence, ici la distribution des groupes sanguins dans la population générale.*

Vérifiez la cohérence des résultats que vous obtenez avec les sorties que vous donnerait le logiciel R à l'issue de ce test (cf. ci-dessous), soit dans l'ordre :

- la valeur du χ^2 observé,
- la valeur du degré de liberté (`df` pour "degree of freedom") de la loi du χ^2 à utiliser et
- la p-value associée.

```
##
##  Chi-squared test for given probabilities
##
## data:  N.groupe.sanguin.obs
## X-squared = 6.036, df = 3, p-value = 0.11
```

4.2.2 Le test du χ^2 d'indépendance

Prenons un exemple de comparaison de plusieurs fréquences observées sur des échantillons indépendants à partir de données tirées de la thèse d'exercice vétérinaire de Mathilde Poinssot (Maisons Alfort, 2011).

A partir d'un échantillon de 999 chiennes d'élevage on voudrait savoir si la fréquence d'intervention de l'éleveur ou du vétérinaire pendant leur mise-bas dépend de la taille des races (variable qualitative à 4 classes qu'on notera ainsi : races géantes (XL), randes races (L), races moyennes (M) et petites races (S)). Autrement dit, on se demande si les fréquences d'intervention sont différentes entre les 4 groupes de taille de race, ou encore si la variable "intervention" est corrélée à la variable "taille de race".

Dans un exemple de ce type les données observées sont présentées sous la forme d'une table de contingence (cf. Table 7), et on calcule généralement les fréquences associées (ici fréquences d'intervention associées à chacun des 4 groupes : 0.681 0.423 0.423 0.497 resp. pour les groupes ou séries XL, L, M,et S). On peut y associer une représentation graphique sous forme de diagramme en barres comme en Figure 39.

Taille de race / **Intervention**	**XL**	**L**	**M**	**S**	Total
NON	29	183	146	170	528
OUI	62	134	107	168	471
Total	91	317	253	338	999

TABLE 7 – Table de contingence décrivant la distribution jointe observée des variables "intervention" et "taille des races" sur les 999 chiennes.

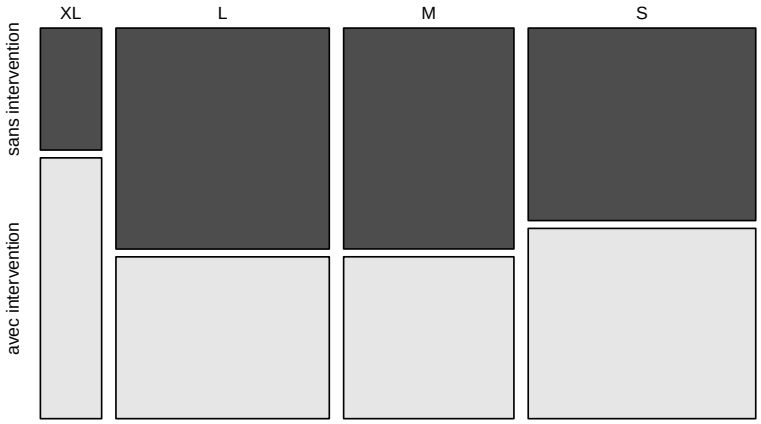


FIGURE 39 – Représentation en barres de la distribution jointe observée des variables "intervention" et "taille des races" sur les 999 chiennes.

Pour **comparer globalement les quatre fréquences**, autrement dit **savoir si l'on peut conclure à une corrélation entre les variables "intervention" et "taille des races"**, on va utiliser le test du χ^2 d'indépendance en suivant la procédure suivante :

Procédure de réalisation d'un test du χ^2 d'indépendance

1. **Calcul des effectifs théoriques sous H_0** : " indépendance entre les deux variables (intervention et taille de races) ce qui équivaut à différence nulle entre les fréquences".

   On calcule les effectifs théoriques C_{ij} à partir des totaux $C_{i.}$ et $C_{.j}$

   Sous H_0, les probabilités marginales et conditionnelles sont les mêmes, c'est-à-dire $\frac{C_{ij}}{C_{i.}} = \frac{C_{.j}}{N}$ donc $C_{ij} = C_{i.} \times \frac{C_{.j}}{N}$

   Dans l'exemple de calcul de l'effectif théorique pour la cellule OUI et L, on utilise le fait que sous H_0 la fréquence d'intervention attendue pour la taille de classe L est la même que celle attendue pour toutes les autres tailles et est donc aussi la fréquence marginale, *i.e.* toutes tailles confondues, égale à $\frac{471}{999}$:

Taille de race Intervention	XL	L	M	S	Total
NON					528
OUI		$\frac{471}{999} \times 317$			471
Total	91	317	253	338	999

   Effectifs théoriques obtenus sur toutes les cellules de la table de contingence :

Taille de race Intervention	XL	L	M	S	Total
NON	48.1	167.5	133.7	178.6	528
OUI	42.9	149.4	119.3	159.4	471
Total	91	317	253	338	999

2. **Calcul de la variable de décision suivante** :

$$\chi^2 = \Sigma_{i=1}^{k} \Sigma_{j=1}^{l} \frac{(O_{ij} - C_{ij})^2}{C_{ij}} = (\Sigma_{i=1}^{k} \Sigma_{j=1}^{l} \frac{O_{ij}^2}{C_{ij}}) - N$$

   avec

   — k le nombre de lignes de la table de contingence (nombre de classes de la variable en ligne),

   — l le nombre de colonnes de la table de contingence (nombre de classes de la variable en colonne),

   — et N le nombre total d'observations.

   Dans notre exemple la valeur observée est $\chi^2_{obs} = 22.38$ (dans le calcul il convient de garder au moins une décimale sur les effectifs théoriques).

3. **Calcul de la valeur de p**

   Si les conditions d'utilisation du test du χ^2 d'indépendance sont respectées, c'est-à-dire si tous les C_{ij} sont supérieurs à 5 (c'est le cas ici), on peut considérer que la variable de décision suit à peu près la **loi du χ^2 de degré de liberté** $(k-1) \times (l-1)$ ce qui nous permet comme précédemment d'encadrer la p-value à partir des valeurs données dans la table 6.

Dans notre exemple, au vu de la loi du χ^2 de degré de liberté 3 (Table 6) et de la valeur observée $\chi^2_{obs} = 22.38$, la p-value est inférieure à 0.0001, et l'on peut donc conclure à une différence globale entre les quatre fréquences, c'est-à-dire à une corrélation entre la taille des races et l'intervention lors de la mise bas. Même si le test ne nous donne qu'une réponse globale (différence globale entre les 4 fréquences), au vu de la figure 39, nous pouvons au moins en conclure que les chiennes de très grande taille ont plus souvent besoin de l'intervention de l'éleveur ou du vétérinaire durant la mise bas que les chiennes de taille inférieure.

Ce test du χ^2 d'indépendance est utilisé pour corréler deux variables qualitatives observées sur les individus d'un échantillon (exemple historique exposé par Karl Pearson : corrélation entre la couleur des cheveux et la couleur des yeux) et pour comparer plusieurs fréquences observées sur des échantillons indépendants. Dans le cas de la comparaison de deux fréquences sur deux échantillons indépendants, il est équivalent au test utilisant la loi normale pour le calcul de la p-value, mais la statistique de la loi normale permet en plus de donner un intervalle de confiance sur la différence entre les deux fréquences (nous verrons des exemples en travaux dirigés).

? ? ? : Pour vous assurer que vous avez bien compris le principe du test du χ^2 d'indépendance, essayez de l'utiliser pour comparer les fréquences d'étudiants vétérinaires vivant avec un animal de compagnie parmi les filles et parmi les garçons, à partir des données obtenues via une enquête réalisée début 2017 auprès d'étudiants ayant intégré le cursus vétérinaire en automne 2016 : 8 étudiants sur 19 (soit 40.5%) vivaient avec un animal de compagnie contre 44 sur 74 chez les étudiantes (soit 57.9%). La différence entre ces deux fréquences observées est-elle significative ?
Vérifiez la cohérence des résultats que vous obtenez avec les sorties que vous donnerait le logiciel R à l'issue de ce test (cf. ci-dessous), soit dans l'ordre :

- la valeur du χ^2 observé,
- la valeur du degré de liberté (`df` pour "degree of freedom") de la loi du χ^2 à utiliser et
- la p-value associée.

Si vous vous souvenez bien du chapitre précédent, peut-on conclure à partir de cette p-value qu'il n'y a pas de corrélation entre le genre des étudiants vétérinaires et le fait qu'ils vivent ou non avec un animal de compagnie ? Justifiez votre réponse.

```
##
##  Pearson's Chi-squared test
##
## data:  t.genre.animal
## X-squared = 1.85, df = 1, p-value = 0.17
```

4.2.3 Quand les conditions d'utilisation des tests du χ^2 ne sont pas respectées

Les tests du d'ajustement et d'indépendance ne peuvent être utilisés que si tous les effectifs calculés sont supérieurs à 5. Que peut-on faire lorsque ce n'est pas le cas ? Considérons tout d'abord les cas de comparaison de deux fréquences. Pour un test d'ajustement (**comparaison d'une fréquence observée**

à une fréquence théorique), on peut faire un **calcul exact** de la p-value à partir de la **loi binomiale** et pour un test d'indépendance de **comparaison de deux fréquences observées**, on peut aussi faire un calcul exact de la p-value à partir du **test de Fisher**. La plupart des logiciels statistiques permettent ces calculs (on les utilisera lors des travaux dirigés sur ordinateur). Une alternative facile à utiliser même sans logiciel statistique est de calculer le critère du χ^2 avec une correction de continuité dite **correction de Yates** ($\chi^2_{cor} = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(|O_{ij}-C_{ij}|-0.5)^2}{C_{ij}}$ pour un test d'indépendance ou $\chi^2_{cor} = \sum_{i=1}^{k}\frac{(|O_i-C_i|-0.5)^2}{C_i}$ pour un test d'ajustement) puis de réaliser le test habituel à partir de cette valeur corrigée (alternative valable si tous les effectifs théoriques restent supérieurs à 2).

En ce qui concerne la comparaison de distributions définies sur plus de deux classes ou de plus de deux fréquences (**tables de contingence de dimensions supérieure à** 2×2), la solution la plus raisonnable, lorsqu'elle a un sens, est de regrouper des classes entre elles afin d'augmenter les effectifs des classes les moins représentées. Mais la meilleure solution est encore de penser à ce problème potentiel en amont et de ne pas travailler avec des classes qui risquent d'être peu représentées. Il existe d'autres alternatives qui font appel aux techniques de tests par permutation qui sont un peu plus délicates à utiliser et que nous ne décrirons pas dans ce cours.

4.3 Comparaison de fréquences sur séries dépendantes

4.3.1 Séries indépendantes ou dépendantes ?

Lorsque l'on compare des moyennes ou fréquences estimées sur plusieurs séries d'observations (issues de plusieurs échantillons / groupes), pour choisir le bon test il est important de savoir si les séries sont indépendantes ou dépendantes. Pour bien comprendre ce point plaçons-nous dans le cas simple de deux séries. **Dire que deux séries sont dépendantes est la même chose que dire qu'elles sont appariées**, *i.e.* qu'on a des paires d'observations. Pour savoir si deux séries sont appariées, imaginez-vous les deux séries d'observations (pour chacun des deux groupes) et demandez-vous si les deux séries les observations sont reliées par paires.

— Dans un exemple de comparaison de deux fréquences, imaginons deux tests diagnostiques utilisés en parallèle sur les mêmes souris. Si on regarde les deux séries d'observations (résultat du test 1 et résultat du test 2 sur chaque souris), les observations y sont reliées par paires, qui correspondent aux réponses obtenues sur un même animal.

— Imaginons un exemple de comparaison de moyennes, où l'on comparerait deux aliments différents sur la prise de poids de jeunes animaux. Imaginons que l'expérimentation consiste à prendre systématiquement deux animaux par portée, sur diverses portées échantillonnées, et à donner à l'un des deux animaux l'aliment A, et à l'autre l'aliment B. Là encore on aurait deux séries d'observations appariées, les paires étant constituées les prises de poids des deux animaux d'une même portée (plus susceptibles de se ressembler que des animaux de portées différentes).

4.3.2 Test de Mc Nemar et test de Cochran pour comparer respectivement deux ou plusieurs fréquences sur des séries dépendantes

Examinons un cas de la comparaison de 2 fréquences observées sur 2 échantillons appariés.

On dispose de 2 tests A et B pour détecter la présence d'une maladie donnée chez des souris. Les 2 tests sont utilisés en parallèle sur 100 souris que l'on sait malades de façon certaine. On souhaite comparer les sensibilités (probabilité de réponse positive chez un malade) des 2 tests.

Il s'agit bien de comparer deux fréquences (les sensibilités) sur 2 séries appariées. Dans un exemple de ce type, on présentera les données sous la forme d'une **table de concordance** (cf. Table 8), à ne pas confondre avec une table de contingence : la table de concordance croise ici "test A (+ ou -)" avec "test B (+ ou -)", alors qu'une table de contingence sur les mêmes données croiserait "Test (A ou B)" avec "résultat du test (+ ou -)".

Résultat du test B / **Résultat du test A**	**positif**	**négatif**
positif	70	6
négatif	18	6

TABLE 8 – Table de concordance décrivant les résultats des deux tests sur les souris testées.

A partir de cette table de concordance on peut calculer les deux sensibilités observées ($Se_A = \frac{70+6}{100} = 0.76$ et $Se_B = \frac{70+18}{100} = 0.88$).

Pour comparer ces deux fréquences, le **test de Mc Nemar se base uniquement sur les nombres de résultats discordants**, et utilise une variable de décision qui **compare les effectifs des deux types de discordances**, c'est-à-dire les nombres de résultats A+B- (ici 6) et A-B+ (ici 18). On comprend bien que les deux fréquences diffèrent si et seulement si ces deux nombres diffèrent.

Principe du test de Mc Nemar

Le test de Mc Nemar est basé sur la comparaison de la proportion de discordances d'un des deux types (indifféremment A+B- ou A-B+ dans cet exemple) à la fréquence théorique de 50% attendue sous l'hypothèse nulle d'égalité des fréquences. En pratique on note f et g les deux **nombres de discordances obtenus dans la table de concordance**, et on utilise comme variable de décision $z = \frac{(|f-g|-1)^2}{f+g}$ qui suit une loi du χ^2 de degré de liberté 1 lorsque $f+g$ le nombre total de discordances est supérieur à 10.

? ? ? : Essayez de faire le test de Mc Nemar sur l'exemple précédent, donc avec $f = 6$ et $g = 18$ et vérifiez la cohérence de ce que vous obtenez avec les sorties que vous donnerait R pour ce test (cf. ci-dessous).

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  t.concordance
## McNemar's chi-squared = 5.04, df = 1, p-value = 0.025
```

? ? ? pour les plus matheux d'entre vous (en option) : On peut assez facilement montrer que réaliser le test de Mc Nemar revient en fait à réaliser un test du χ^2 d'ajustement pour comparer à la fréquence théorique de 50% l'une des proportions de discordances (n'importe laquelle des deux $\frac{f}{f+g}$ ou $\frac{g}{f+g}$), en faisant la correction de continuité de Yates dans la formule du χ^2.
Vous pouvez essayer de le démontrer si le formalisme mathématique ne vous fait pas peur.

Le test de Mc Nemar permet uniquement de comparer 2 fréquences et en aucun cas de juger de la concordance entre les tests. Pour caractériser la **concordance entre deux tests**, la statistique du **Kappa de Cohen** serait utilisable, mais nous ne la verrons pas dans le cadre de ce cours.

? ? ? : Avez-vous bien compris le paragraphe précédent et en êtes-vous convaincu ? Pour vous en convaincre, imaginez un exemple de table de concordance correspondant à un cas où les deux tests seraient très discordants mais caractérisés par deux sensibilités identiques.

Repartons de l'exemple que nous venons de traiter, et supposons que nous n'avons pas seulement deux tests diagnostiques à comparer mais trois. Il s'agit donc de comparer plusieurs fréquences sur des séries dépendantes. Le **test de Cochran-Mantel-Haenszel** permet le calcul de la p-value dans ce cadre. Nous ne décrirons pas la variable de décision utilisée par ce test, ni dans le cours ni en travaux dirigés du premier semestre, mais nous serons amenés à l'utiliser lors de nos travaux dirigés sur ordinateurs au second semestre. L'essentiel pour vous à ce stade est de savoir dans quel cadre il convient de l'utiliser.

Voici un petit récapitulatif qui pourra vous aider pour le choix du bon test lors de la comparaison de fréquences, auquel j'ai ajouté, pour vous aider lors des travaux dirigés du second semestre, le nom des fonctions du langage **R** que nous utiliserons.

- **Un seul échantillon**
  test du χ^2 d'ajustement de comparaison d'une fréquence observée à une fréquence théorique ou d'une distribution observée à une distribution théorique (`chisq.test()`) ou test exact utilisant la loi binomiale (`binom.test()`)
- **Deux ou plusieurs échantillons indépendants**
  test du χ^2 d'indépendance de comparaison de plusieurs fréquences observées ou plusieurs distributions observées (`chisq.test()` et dans le cas de deux fréquences `prop.test()` pour obtenir l'intervalle de confiance sur la différence entre les 2 fréquences et `fisher.test()` pour le test exact)
- **Deux échantillons dépendants (appariés)**
  test de Mc Nemar de comparaison de deux fréquences observées (`mcnemar.test()`)
- **Plusieurs échantillons dépendants**
  test de Cochran-Mantel-Haenszel de comparaison de plusieurs fréquences observées (`mantelhaen.test()`)

5 Comparaison de moyennes ou méthodes permettant de corréler une variable quantitative à une variable qualitative

5.1 Objectifs pédagogiques

A l'issue de l'étude de ce chapitre et de la réalisation du deuxième TD de S3, vous devriez :

— Avoir compris les différences entre un test paramétrique et un test non paramétrique
— Savoir réaliser à la main les deux tests de Student (séries indépendantes ou appariées) et avoir bien compris l principe des tests non paramétriques associés (somme des rangs et rangs signés).
— Connaître le principe de l'analyse de variance
— Connaître le principe des méthodes de comparaisons multiples et leurs limites
— Savoir interpréter les résultats d'un test de normalité et d'un test de comparaison de variances et en connaître les limites.
— Savoir choisir le test adapté pour comparer deux ou plusieurs séries d'une variable quantitative en fonction de la question posée, du plan d'expérience et des données, et en interpréter les résultats.

5.2 Différence entre les deux approches, paramétrique et non paramétrique

Partons d'un exemple.

Un essai randomisé a été réalisé sur 18 chiens, afin d'évaluer l'efficacité d'un supplément alimentaire contre la formation de tartre sur les dents de l'animal. Neuf chiens reçoivent une alimentation supplémentée (groupe supplément) et neuf chiens ne reçoivent aucune supplémentation (groupe témoin). La formation de tartre est quantifiée par un index combinant la proportion de dents atteintes et l'épaisseur de la couche de tartre formée. On se demande si le supplément alimentaire diminue globalement l'index de tartre (ce qui est attendu/souhaité).

Il s'agit d'un exemple de comparaison de deux moyennes (ou plus généralement deux tendances centrales) sur deux séries indépendantes (on n'a pas de paires d'observations). Les données vous ont été représentées sous deux formes en Figure 40.

5.2.1 Approche paramétrique

Une approche paramétrique classique sur ce type d'exemple va supposer que le **théorème de l'approximation normale s'applique** et que les variances sont égales (même dispersion dans les deux groupes). Elle visera à comparer les deux moyennes observées qui sont ici respectivement de 0.747 pour le groupe supplément et de 1.089 pour le groupe témoin.

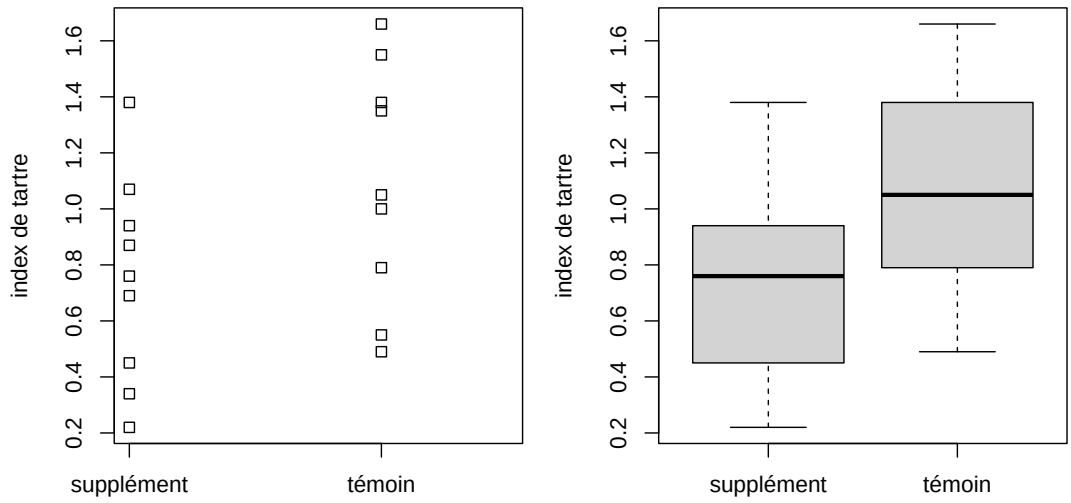


FIGURE 40 – Deux représentations des distributions observées de l'index de tartre sur les deux séries.

Principe du test de Student de comparaison de deux moyennes sur séries indépendantes

Voici la variable de décision utilisée pour réaliser le **test de Student** de comparaison des deux moyennes et la loi approchée sous H_0 si l'on peut supposer le théorème de l'approximation normale applicable chacun des deux groupes et qu'il est raisonnable de supposer les écarts types égaux dans les deux groupes :

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim T(n_1 + n_2 - 2)$$

avec $\hat{\sigma} = \sqrt{\frac{(n_1-1)\hat{\sigma_1}^2+(n_2-1)\hat{\sigma_2}^2}{n_1+n_2-2}}$

où $\hat{\sigma_1}$ et $\hat{\sigma_2}$ représentent les écarts types estimés sur chacun des deux groupes,

et $T(n_1 + n_2 - 2)$ représente la loi de Student de degré de liberté $n_1 + n_2 - 2$ (cf. Table 5 et Figure 29).

A partir de cette variable de décision que l'on peut calculer sur les données (dans notre exemple $t_{obs} = -1.84$), on peut calculer la valeur de la p-value comme une aire sous la courbe, comme nous l'avons vu précédemment dans le chapitre 3.4.1 et comme illustré Figure 41 (ici $0.1 > p > 0.05$).

Intervalle de confiance associé au test de Student de comparaison de deux moyennes sur séries indépendantes

La statistique de Student utilisée pour calculer la p-value peut aussi être utilisée pour calculer un intervalle de confiance autour de la différence entre les deux moyennes :

$$\mu_1 - \mu_2 = \overline{x}_1 - \overline{x}_2 \pm t_{n_1+n_2-2;1-\frac{\alpha}{2}} \times \hat{\sigma} \times \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

avec $t_{n_1+n_2-2;1-\frac{\alpha}{2}}$ le quantile à $1-\frac{\alpha}{2}$ de la loi de Student de degré de liberté n_1+n_2-2 (cf.Table 5).

Dans la partie 3.5 (de lecture optionnelle) il a été montré comment obtenir cet intervalle de confiance à partir du résultat précédent, et démontré **que dire que la p-value est inférieure à** 5% **(donc**

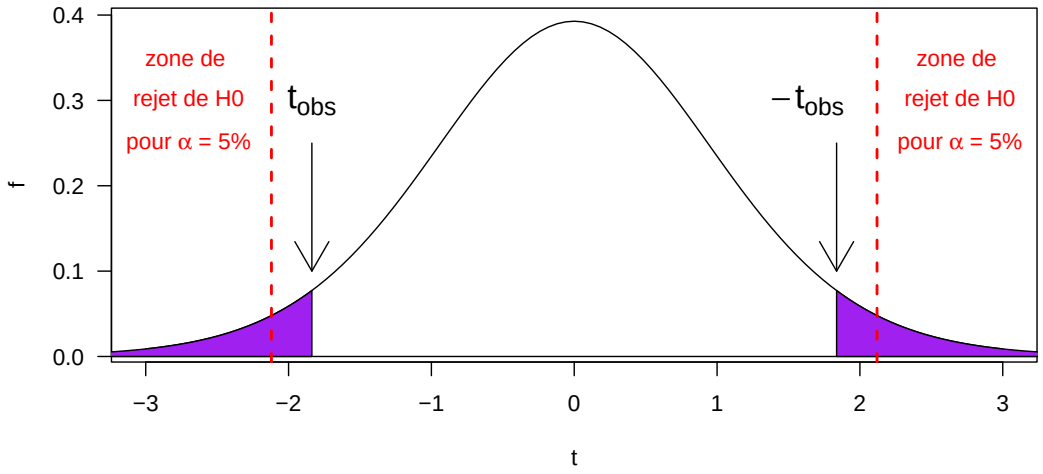


FIGURE 41 – Illustration du calcul de la p-value (aire sous la courbe de densité de probabilité de la loi coloriée en violet) dans le test de Student de comparaison des distributions observées de l'index de tartre sur les deux séries.

qu'on conclut à une différence significative entre les deux moyennes) est équivalent à dire que l'intervalle de confiance à 95% sur la différence entre les 2 moyennes ne contient pas la valeur 0. Mais cet **intervalle de confiance est plus informatif que la p-value** et il est donc important de le donner. Dans cet exemple la différence est estimée à -0.34 avec son intervalle de confiance à 95% de $[-0.74; 0.05]$. Ce résultat ne permet donc pas de mettre en évidence une différence, et il est alors courant dans ce cas de dire que la différence est non significative, mais on gardera en tête qu'il convient de se méfier de cette terminologie (cf. Wasserstein et al. (2019)) et que cela ne permet en aucun cas de conclure qu'il n'y a pas de différence. On n'aurait d'ailleurs pas envie de le faire en regardant ce que nous dit l'intervalle de confiance (cf. Figure 42) qui n'est pas en faveur d'une différence nulle, à moins que l'on considère comme négligeable une différence de -0.74 (peu probable pour un index qui est d'environ 1 dans le groupe témoin).

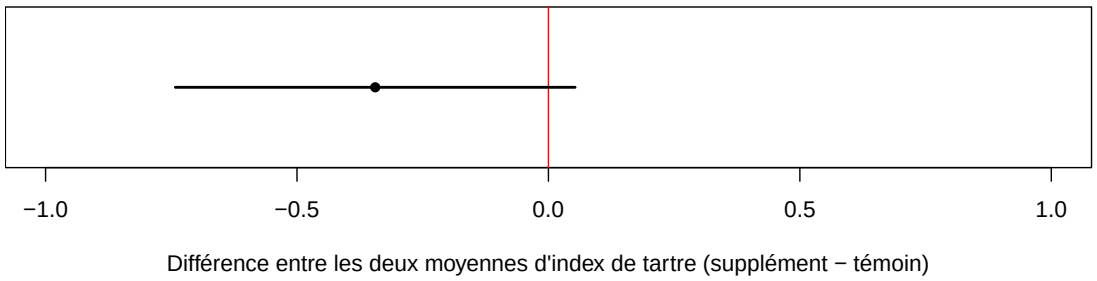


FIGURE 42 – Visualisation de l'estimation de la différence entre les deux moyennes d'index de tartre dans notre exemple

? ? ? Vous trouvez ci-dessous la sortie de R pour ce test. Observez-la bien pour y repérer tous les résultats calculés que l'on a calculé sur cet exemple (valeur observée de la variable de décision, degré de liberté de la loi suivie par cette variable sous H_0, p-value associée, moyennes observées, intervalle de confiance sur la différence entre les deux moyennes).

```
##
##  Two Sample t-test
##
## data:  d$index by d$traitement
## t = -1.84, df = 16, p-value = 0.085
## alternative hypothesis: true difference in means between group supplement and group tem
## 95 percent confidence interval:
##  -0.741977  0.053088
## sample estimates:
## mean in group supplement     mean in group temoin
##                  0.74667                  1.09111
```

5.2.2 Approche non paramétrique

Dans une aproche **paramétrique**, la variable de décision du test est calculée **à partir d'un paramètre statistique caractérisant chaque loi observée** (ici la moyenne) et cette variable de décision n'est de loi connue (ici la loi de Student) que si certaines hypothèses fortes peuvent être faites sur les lois observées. Dans une approche **non paramétrique**, on n'utilise plus directement de paramètre caractérisant chaque loi observée (ici on n'utilise plus la moyenne) et on ne fait plus d'hypothèse forte quant à la forme des lois observées. Les approches non paramétriques les plus courantes sont basées sur des **statistiques de rang**, qui n'utilisent comme information que l'ordonnancement des observations entre elles. Ces statistiques sont dites plus robustes, au sens moins sensibles aux valeurs extrêmes.

Principe du test de la somme des rangs de Mann-Whitney-Wilcoxon

Sur l'exemple précédent, voici le principe du **test de Mann-Whitney-Wilcoxon de la somme des rangs**, qui est aussi illustré en Figure 43.

1. On **classe globalement les observations**
   et on affecte son rang à chaque observation
   en moyennant les rangs des ex aequos
   (dans l'exemple on en a deux à la 15ème position, on leur affecte le rang 15.5 et on continue ensuite au rang 17),
2. **on calcule la somme des rangs de chacun des groupes**
   (dans l'exemple $T_{supplement} = 66.5$ et $T_{temoin} = 104.5$)
3. **on compare les 2 sommes des rangs**
   à l'aide d'une variable de décision adaptée qui n'est pas incluse dans ce polycopié
   (seul le principe du test est à connaître).

Dans l'exemple on obtient une valeur de $p > 0.10$ ce qui ne nous permet pas de conclure à une différence. On utilisera R au second semestre pour réaliser ce test, et il vous suffit à ce stade d'avoir bien compris le principe du test, et de savoir retrouver et interpréter sa p-value dans la sortie de R (cf. ci-dessous pour notre exemple) .

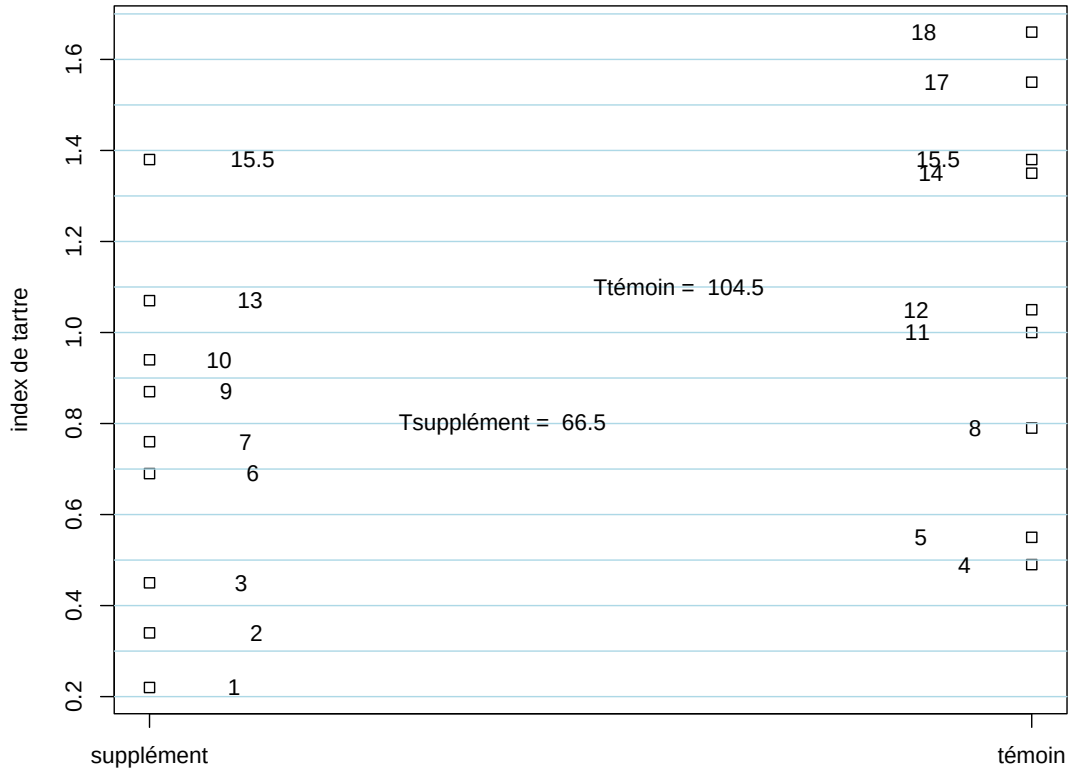


FIGURE 43 – Calcul des rangs des observations et de la somme des rangs dans notre exemple

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  d$index by d$traitement
## W = 21.5, p-value = 0.1
## alternative hypothesis: true location shift is not equal to 0
```

Nous obtenons, dans ce cas (ce n'est pas systématique bien sûr), la même conclusion qu'avec l'approche paramétrique, mais nous ne pouvons pas donner plus d'information qui pourrait nous aider dans l'interprétation du résultat. Effectivement, comme cette approche ne se base pas sur l'estimation d'un paramètre (pas d'estimation des moyennes ni de leur différence), elle ne peut fournir d'indication sur l'estimation (intervalle de confiance).

? ? ? Pour vous assurer que vous avez bien compris le principe du test de la somme des rangs, calculez les sommes des rangs de chacun des deux groupes pour l'exemple suivant. *Dans le cadre d'un essai randomisé on évalue l'effet de deux traitements sur la charge parasitaire (mesurée en nombre d'oeufs par gramme de fèces après 10 jours de traitement).*
Les résultats obtenus sur 13 animaux sont répertoriés dans le tableau suivant :

traitement	A	A	A	A	A	A	B	B	B	B	B	B	B
charge.parasitaire	22	38	42	190	1200	3000	0	0	0	2	15	32	48

Dans un deuxième temps interprétez la sortie de R associé à ce test sur cet exemple (cf. ci-dessous). Notez que le message d'alerte ne nous empêche pas d'interpréter la p-value. Il indique juste qu'en présence d'ex aequos dans les données (ce qui est le cas ici) le calcul de la p-value n'est pas exact mais approché.
A votre avis pourquoi a-t-on choisi une approche non paramétrique dans un tel cas ?

```
## Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): impossible de
calculer la p-value exacte avec des ex-aequos

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dantipara$charge by dantipara$traitement
## W = 38, p-value = 0.02
## alternative hypothesis: true location shift is not equal to 0
```

? ? ? La terminologie "test non paramétrique de comparaison de deux moyennes" pour désigner le test de la somme des rangs que l'on vient de présenter est assez courante et nous l'utiliserons parfois même si elle est impropre. En quoi est-elle impropre ?

5.2.3 Choix entre les deux approches

Pour vous aider à choisir entre les deux approches, sur chaque cas pratique, résumons les avantages et inconvénients de chacune.

— **Approche paramétrique**
Hypothèse forte sur la forme des distributions.
Conditions d'utilisation assez restrictives.
Intervalle de confiance associé pouvant s'avérer très informatif surtout en cas de non rejet de H_0.

— **Approche non paramétrique**
Pas d'hypothèse forte quant à la forme des distributions,
Dégradation de l'information initiale qui peut induire à une perte de puissance.
Pas d'intervalle de confiance associé.

Vous comprendrez bien à partir de ce bilan que l'approche paramétrique à privilégier, si possible, c'est-à-dire si ses conditions d'utilisation sont vérifiées, éventuellement après transformation de la variable pour que ses conditions soient respectées (on verra notamment des exemples de tranformation logarithmique en travaux dirigés).

Revenons à l'exemple. Peut-on utiliser une approche paramétrique sur cet exemple ? Peut-on appliquer le théorème de l'approximation normale ? La variable est un index combinant diverses informations (variable de type score). Rien ne garantit donc à l'avance la normalité de sa distribution. Les effectifs ne sont pas très grands (deux groupes de 9). L'observation des données ne conduit pas à remettre en cause l'hypothèse de normalité des distributions (en plus des diagrammes en boîte de la Figure 40 où l'on voit que les boîtes sont à peu près symétriques, sans valeurs extrêmes, on peut voir sur les diagrammes Quantile-Quantile de la Figure 44 que les points sont à peu près alignés), néanmoins les effectifs ne sont vraiment pas très grands. On est ici dans un **cas un peu limite** ou certains choisiraient une démarche paramétrique et d'autres une démarche non paramétrique. Dans le cas du choix d'une démarche paramétrique, il serait **raisonnable de supposer les variances égales** (écarts types du même ordre de grandeur, 0.37 pour le groupe supplément et 0.42 pour le groupe témoin et dispersions comparables d'après les diagrammes en boîte).

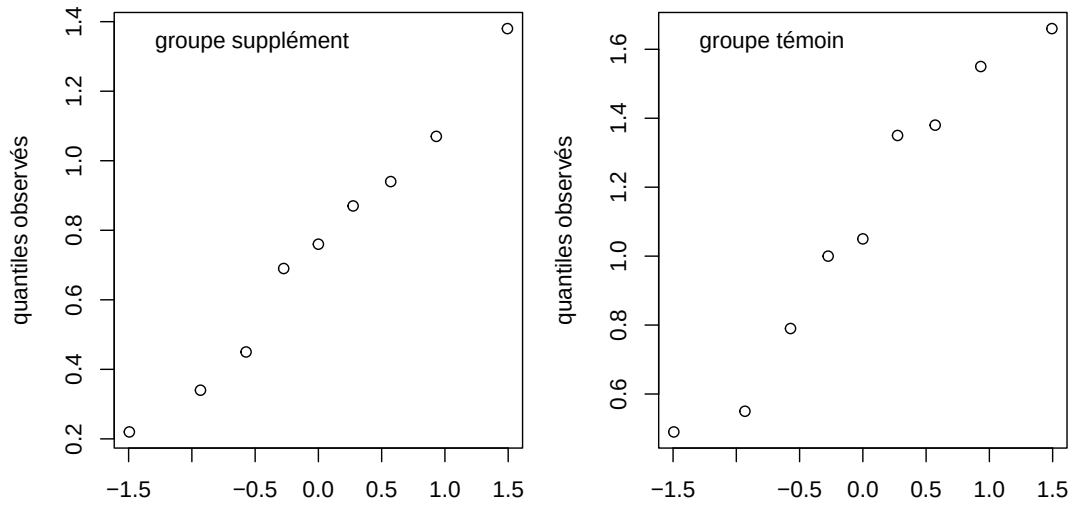


FIGURE 44 – Diagrammes Quantile-Quantile de l'index de tartre sur chacune des deux séries.

Le **choix entre approche paramétrique et non paramétrique** se fera essentiellement à partir de l'**examen visuel des distributions**, visant à évaluer si le théorème de l'approximation normale est applicable. Vous verrez en travaux dirigés que bien souvent le choix s'impose de lui-même. Mais il est pour certains frustrants de ne pas disposer d'une méthode plus "objective" pour choisir entre les deux approches, et il n'est pas rare de voir, dans des articles, l'utilisation de tests de normalité et d'égalité des variances pour vérifier les conditions d'utilisation du test de Student montré précédemment.

— Les **tests de normalité**, dont le test de Shapiro Wilk que nous pourrons être amenés à utiliser au second semestre à l'aide du langage **R**, ont comme hypothèse nulle H_0 : **"la distribution est normale"**.

— Le **test de Fisher de comparaison de variances**, que nous utiliserons au second semestre, a comme hypothèse nulle H_0 : **"les variances sont égales"**. Nous l'utiliserons lorsque notre

objectif sera de mettre en évidence une différence entre deux séries d'observations en terme de dispersion.

Mais l'utilisation de ces deux tests (test de normalité et test d'égalité des variances) ne permet pas à eux seuls de vérifier ni la normalité ni l'égalité des variances, et l'examen visuel des distributions est indispensable pour vérifier les conditions d'utilisation d'une approche paramétrique de type test de Student.

? ? ? Si vous avez bien compris le principe des tests, vous devez être capable de justifier ce que nous venons d'énoncer. Prenez le temps d'y réfléchir et de trouver les bons arguments pour le justifier.

5.3 Méthodes de comparaison de deux moyennes

Dans ce chapitre sont présentées les méthodes de comparaison de deux moyennes. Avant d'analyser ce type de données, il est important de bien caractériser le cas étudié. S'agit-il de comparer une moyenne observée sur un échantillon à une moyenne de référence (ou moyenne théorique), de comparer deux moyennes observées sur deux échantillons et dans ce dernier cas, il conviendra de bien évaluer si les séries sont indépendantes ou appariées (cf. chapitre 4.3.1 pour cette notion).

Dans ce chapitre nous ne détaillerons pas tous les aspects techniques de réalisation des tests (seuls certains d'entre eux seront vus dans le détail dans les travaux dirigés de S3, et tous seront utilisés en travaux dirigés sur ordinateurs en S4, mais le calcul sera alors du ressort du logiciel utilisé), mais nous présenterons les éléments permettant de comprendre le principe de chacun des tests.

5.3.1 Comparaison d'une moyenne observée à une moyenne théorique

Prenons un exemple.

Un laboratoire d'analyse indique comme valeur moyenne de l'urée plasmatique chez les chats sains, une valeur de 8.5 mmol/l. Suite à un remplacement de ses appareils de mesure, le laboratoire dose l'urée sur un échantillon aléatoire de 140 chats en bonne santé. On obtient comme statistiques résumées : m = 9.7 mmol/l et SD = 2.6 mmol/l. Peut-on en conclure que la moyenne de l'urée plasmatique chez les chas sains a bougé suite au remplacement des appareils ?

Il s'agit d'un problème de comparaison d'une moyenne (ou tendance centrale) observée sur un échantillon à une moyenne théorique, ou de référence (non associée à un échantillon mais supposée connue). Voici les approches paramétriques et non paramétriques utilisables sur un tel exemple :

— **Approche paramétrique**
test de **conformité de Student** si le théorème de l'approximation normale s'applique et/ou calcul de l'intervalle de confiance autour de la moyenne observée (à l'aide de la formule donnée au chapitre 3.3.2). **Si l'intervalle de confiance à 95% sur la moyenne observée ne contient pas la moyenne théorique, c'est que la p-value est inférieure à 5%.**

— **Approche non paramétrique**
test de la médiane sinon

Ce test revient à se poser la question : la valeur théorique est-elle au milieu des observations ? Pour y répondre on compte les effectifs observés de part et d'autre de la valeur théorique, et on les compare aux effectifs théoriques 50% - 50%, à l'aide d'un test du χ^2 d'ajustement (cf. comparaison d'une fréquence observée à une théorique au chapitre 4.2.1).

5.3.2 Comparaison de deux moyennes sur des séries indépendantes

Nous avons traité du cas de la comparaison de deux moyennes sur des séries indépendantes dans le chapitre introductif 5.2. Voici en format condensé les approches possibles dans un tel cas :

- **Approches paramétriques**
  si le théorème de l'approximation normale s'applique
  - **test de Student avec "variances égales"** et intervalle de confiance associé s'il est raisonnable de supposer les écarts types égaux
  - **test de Welch** appelé aussi test de Student avec variances inégales, et intervalle de confiance associé, si les écarts types semblent différents et qu'il reste intéressant de comparer les moyennes (nous verrons au second semestre comment le mettre en oeuvre avec R).
- **Test non paramétrique**
  test de la somme des rangs de Mann-Whitney-Wilcoxon

5.3.3 Comparaison de deux moyennes sur des séries appariées

Pour aborder ce dernier cas nous allons prendre un nouvel exemple.

On souhaite comparer une nouvelle méthode de dosage de l'urée urinaire (méthode 2) à la méthode de référence (méthode 1). Pour cela on a dosé l'urée par les 2 méthodes chez 12 animaux (cf. résultats en g/24h dans le tableau ci-dessous).

meth1	20.4	25.4	25.6	25.6	26.6	28.6	28.7	29	29.8	30.5	30.9	31.1
meth2	21.7	26.3	26.8	28.1	26.2	27.3	29.5	32	30.9	32.3	32.3	31.7

Il s'agit bien de séries appariées, puisque sur chaque animal on dispose des résultats des deux méthodes (paire d'observations). Visualisons les données brutes en matérialisant ces paires (cf. Figure 45).

Lorsque l'on a des séries appariées, on peut calculer une différence par paire d'observations (cf. Figure 46 et tableau ci-dessous).

meth1	20.4	25.4	25.6	25.6	26.6	28.6	28.7	29	29.8	30.5	30.9	31.1
meth2	21.7	26.3	26.8	28.1	26.2	27.3	29.5	32	30.9	32.3	32.3	31.7
diff_meth2_meth1	1.3	0.9	1.2	2.5	-0.4	-1.3	0.8	3	1.1	1.8	1.4	0.6

Dans le cadre d'une **approche paramétrique**, on souhaite comparer les moyennes des 2 groupes. Or la différence des moyennes est aussi la moyenne des différences. Comparer les 2 moyennes (donc tester l'égalité de leur différence à 0) revient donc à comparer la moyenne des différences à 0. Dans ce but on peut donc utiliser un test de comparaison d'une moyenne observée sur un échantillon (moyenne des différences) à 0 ce qui nous ramène à une technique vue précédemment.

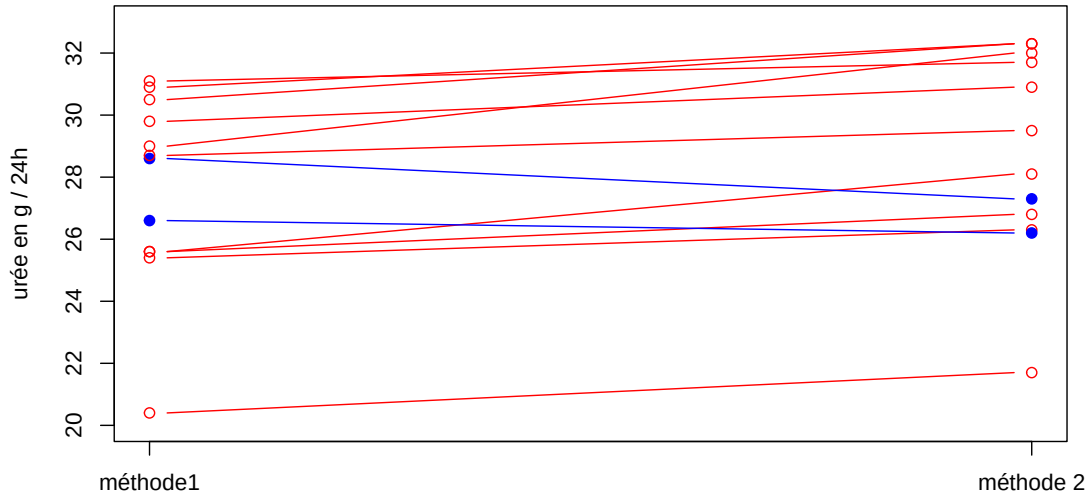


FIGURE 45 – Représentation des paires d'observation de l'urée avec les deux méthodes dans notre exemple, avec les paires associées à une différence méthode2 - méthode1 coloriées en rouge (entre cercles vides) et les autres en bleu (entre cercles pleins).

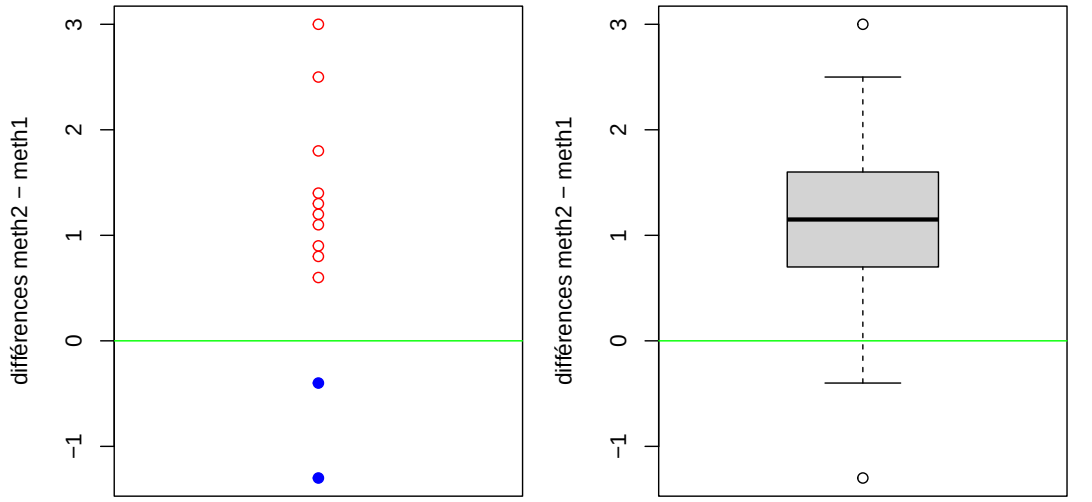


FIGURE 46 – Représentation de la distribution des différences méthode2 - méthode1 en coloriées en rouge si elles sont positives (cercles vides) et en bleu sinon (cercles pleins) (pour la figure de gauche).

Principe du test de Student de comparaison de séries appariées

Dans le cadre d'une approche paramétrique, il suffira de **calculer l'intervalle de confiance sur la moyenne des différence**

en se servant de la formule donnée au chapitre 3.3.2, **et de regarder s'il contient ou non 0**.

Les conditions d'utilisation sont que l'on puisse appliquer le théorème de l'approximation normale sur la série des différences par paire, (cf. Figure 46) et non sur chacune des deux séries.

? ? ? Pour vous assurer que vous avez bien compris le principe de ce test, reprenez les données de l'exemple, calculez l'intervalle de confiance sur la moyenne des différences et interprétez-le, puis retrouvez vos résultats dans les sorties de R ci-dessous correspondant à la mise en oeuvre de l'approche paramétrique sur notre exemple. Pour vous simplifier les calculs on vous donne la moyenne et l'écart type estimé de ces différences (resp. 1.075 pour la moyenne et 1.151 pour l'écart type estimé)

```
##
##  Paired t-test
##
## data:  meth2 and meth1
## t = 3.23, df = 11, p-value = 0.008
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.34345 1.80655
## sample estimates:
## mean difference
##           1.075
```

Dans le cadre d'une **approche non paramétrique**, basée sur les statistiques de rangs, nous utiliserons le **test des rangs signés de Wilcoxon** dont voici le principe.

Principe du test des rangs signés de Wilcoxon
Le principe de ce test consiste à **classer les différences en valeur absolue**
(toujours en associant un rang moyen en cas de différences ex aequos)
puis à **comparer la somme des rangs T+** des différences positives
à la somme des rangs T- des différences négatives
(comme illustré sur la Figure 47)
avec une variable de décision adaptée. Comme pour le test de la somme des rangs, nous ne vous donneront pas la définition de cette variable de décision, mais nous vous demandons uniquement de connaître le principe de ce test.

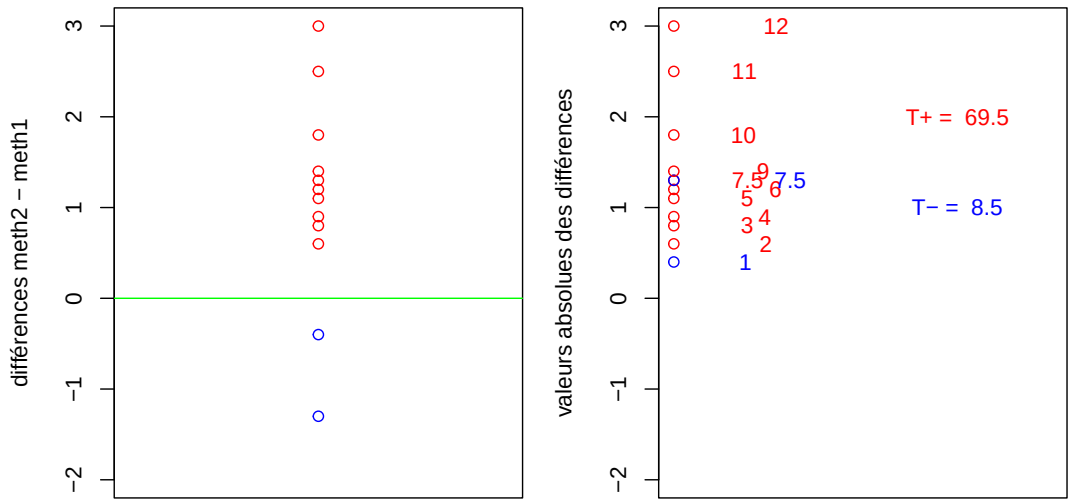


FIGURE 47 – Illustration du principe du test des rangs signés de Wilcoxon.

??? Pour vous assurer que vous avez bien compris le principe de ce test, tentez de recalculer à partir des données les valeurs de T+ et T- qui vous sont reportées dans la figure 47.
Vous trouvez ci-dessous les sorties de R correspondant à la mise en oeuvre de l'approche non paramétrique sur notre exemple. Interprétez l'information intéressante qui apparaît dans cette sortie.

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  meth1 and meth2
## V = 8.5, p-value = 0.019
## alternative hypothesis: true location shift is not equal to 0
```

5.4 Comparaison de plusieurs moyennes sur des séries indépendantes

5.4.1 Analyse de variance à un facteur (ANOVA 1) et méthode non paramétrique associée

Dans le cadre de ce cours, **nous n'aborderons le cas de la comparaison de plus de deux moyennes uniquement sur des séries indépendantes**. Le cas des séries dépendantes ne sera abordée que dans l'enseigement personnalisé d'introduction à la modélisation, comme un cas particulier de la prise en compte de facteurs aléatoires (chapitre modèle linéaire mixte).

Partons d'un exemple à partir de données tirées de la thèse d'exercice vétérinaire de Mathilde Poinssot (Maisons Alfort, 2011). A partir d'un échantillon de 928 chiennes d'élevage, on voudrait savoir si la durée de gestation (variable quantitative) dépend de la taille des races (variable qualitative à 4 modalités - XL, L, M et S - que nous avons déjà manipulée). Il s'agit de **corréler une variable quantitative à une variable qualitative, ce qui revient à comparer les moyennes (ou plus généralement les tendances centrales) obtenues pour les quatre groupes** de taille de races. Les données observées on été représentées sous forme de diagrammes en boîte sur la figure 48.

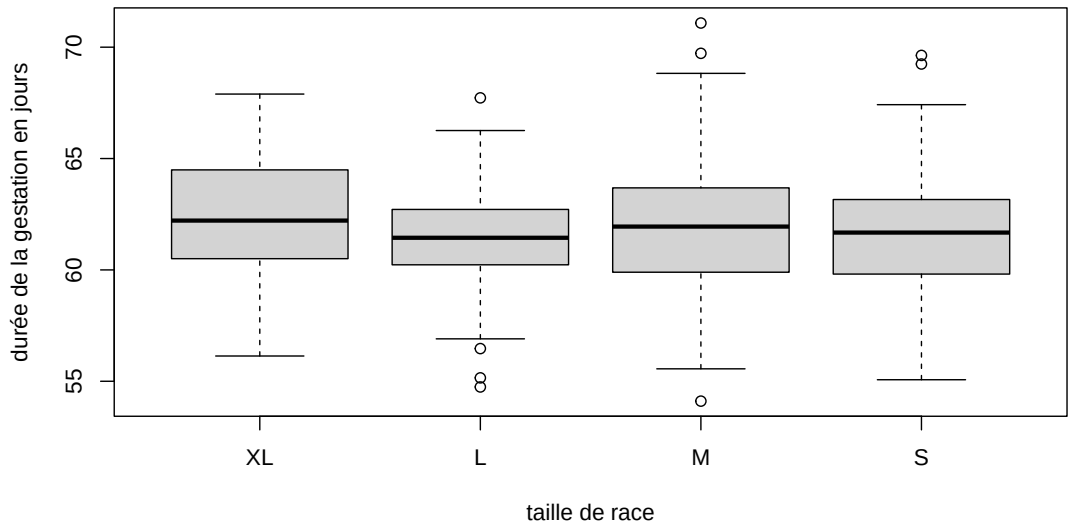


FIGURE 48 – Diagrammes en boîte des durées de gestation pour les quatre tailles de races

Si le théorème de l'approximation normale s'applique et que les variances peuvent être supposées égales, on peut réaliser une **analyse de variance à un facteur (ANOVA 1)**, généralisation du test de Student avec variances égales. Une **variante de l'ANOVA 1 ne supposant pas les variances égales** existe et nous l'utiliserons au cours des travaux dirigés sur ordinateurs sans donner plus de détails techniques. Si le théorème de l'approximation normale ne s'applique pas, nous pourrons utiliser le **test de la somme des rangs de Kruskal-Wallis**, qui est une généralisation du test de Mann-Whitney-Wilcoxon basé exactement sur le même principe (ce pourquoi nous le détaillerons pas ici). Nous allons donc dans la suite présenter uniquement le principe de l'ANOVA 1.

On appelle **facteur la variable qualitative définissant les groupes** (ici la taille de race). L'ANOVA 1 est basé sur le modèle suivant pour décrire les observations x_{ij} de la variable quantitative, i indiquant le numéro du groupe, donc de la modalité du facteur étudié et j indiquant le numéro de l'observation dans le groupe :

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij} \text{ avec } \epsilon_{ij} \sim N(0, \sigma)$$

L'hypothèse nulle de l'ANOVA 1 est :

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0.$$

Sous H_0 on a tous les α_i nuls donc une seule distribution $N(\mu, \sigma)$, alors que sous H_1 : au moins un des α_i est non nul donc on a plusieurs distributions $N(\mu + \alpha_i, \sigma)$ (cf. illustration des deux hypothèses sur la figure 49).

L'ANOVA, qui est, comme son nom ne l'indique pas, une méthode de comparaison globale de plusieurs moyennes, se base sur une **décomposition de la variance** totale en une **variance intra-groupe (résiduelle)** et une **variance inter-groupe (factorielle)** et sur la comparaison de ces 2 composantes de la variance.

Décomposition de la variation totale (ou somme des carrés des écarts totale) :

$$SCE_T = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2 =$$

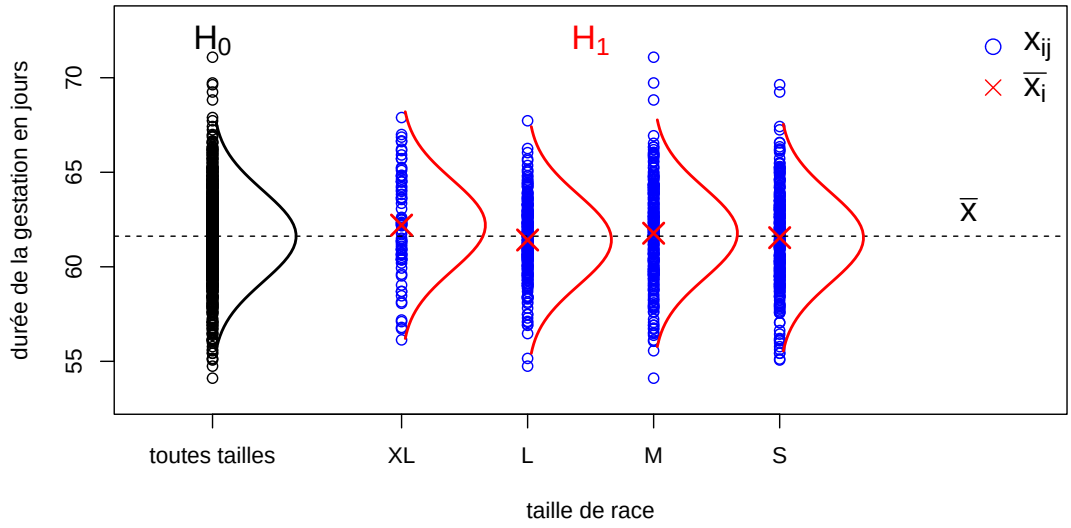


FIGURE 49 – Illustration de l'hypothèse nulle de l'ANOVA 1

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2 + \sum_{i=1}^{p}\sum_{j=1}^{n_i}(\overline{x}_i-\overline{x})^2 = SCE_R + SCE_A$$

avec p le nombre de modalités du facteur A (nombre de groupes) et n_i l'effectif du groupe i.

A partir de ces deux sommes des carrés des écarts, on va estimer les variances intra-groupe et inter-groupe appelés aussi carrés moyens

$$CM_R = \frac{SCE_R}{\sum_{i=1}^{p}(n_i-1)} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2}{\sum_{i=1}^{p}(n_i-1)} = \frac{\sum_{i=1}^{p}(n_i-1)\hat{\sigma}_i^2}{N-p}$$
$$CM_A = \frac{SCE_A}{p-1} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(\overline{x}_i-\overline{x})^2}{p-1} = \frac{\sum_{i=1}^{p}n_i(\overline{x}_i-\overline{x})^2}{p-1}$$

Le principe de l'ANOVA est ensuite de comparer ces variances intra-groupe et inter-groupe en prenant comme variable de décision leur rapport F, qu'on appelle la statistique de Fisher. Sous H_0, $F = \frac{CM_A}{CM_R}$ suit la **loi** $F(p-1, N-p)$ **de Fisher et Snédécor** de degrés de liberté $p-1$ et $N-p$. On rejettera donc H_0 quand $CM_A \gg CM_R$ ce qui sera associé à une p-value faible.

Nous ne vous demanderons à aucun moment de réaliser à la main les calculs présentés ci-dessus (calculs un peu lourds et chronophages). Nous vous demandons juste d'essayer de comprendre le principe de cette méthode, afin de savoir interpréter les résultats que nous obtiendrons lors des travaux dirigés sur ordinateurs.

A titre d'exemple, voici la sortie que nous obtenons avec le langage **R** lors de l'analyse des données de notre exemple (les variables sont codées `duree` et `taille` dans cette sortie) avec en sus le calcul des moyennes et écarts types pour les quatre groupes.

```
## Analysis of Variance Table
##
## Response: d$duree
##            Df Sum Sq Mean Sq F value Pr(>F)
## d$taille    3     46   15.23    2.65  0.048
## Residuals 924   5317    5.75
```

Moyennes estimées sur les quatre groupes

```
##   XL    L    M    S
## 62.2 61.4 61.8 61.5
```

Ecarts types estimés sur les quatre groupes

```
##   XL    L    M    S
## 2.74 1.96 2.62 2.48
```

? ? ? Pour voir si vous avez bien compris, essayez d'interpréter la sortie informatique ci-dessus en reliant tous les résultats qui y sont donnés aux notions théoriques que nous venons de développer. Voici enfin la sortie de R pour le test de la somme des rangs de Kruskal-Wallis appliqué à cet exemple. Interprétez le résultat de ce test non paramétrique.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  d$duree by d$taille
## Kruskal-Wallis chi-squared = 7, df = 3, p-value = 0.06
```

5.4.2 Problématique des comparaisons multiples

Suite à la mise en évidence d'une différence globale significative entre plusieurs moyennes (par ANOVA ou test de Kruskal-Wallis), on souhaite parfois comparer les moyennes 2 à 2. La **méthode basique dite PLSD de Fisher** (protected least significant difference) utilise la statistique de Student pour chaque test mais avec comme σ la valeur commune estimée à partir de l'ensemble des groupes (soit la racine carrée du CM_R de l'ANOVA). Le **problème majeur** associé à ce type de comparaisons multipes est la **répétition des tests qui induit une inflation du risque α global** (n'oublions pas, en prenant un risque α pour chaque test à 5%, quand on est sous H_0 on s'autorise dans un cas sur 20 à rejeter H_0).
Si l'on veut maîtriser le risque α global (risque α global = probabilité de détecter au moins une différence significative parmi toutes celles testées si on est sous H_0), il apparaît donc nécessaire de **corriger le risque α (ou de façon équivalente les valeurs de p)**.

Deux méthodes classiques de correction du risque α sont les suivantes :

— **La méthode de Bonferroni** :

elle est utilisable après la mise en évidence d'une différence globale significative entre plusieurs moyennes (par ANOVA ou test de Kruskal-Wallis). Le principe est, pour chaque test, de corriger α en le divisant par k le nombre de tests réalisés ($\alpha_{cor} = \frac{0.05}{k}$) ou de façon équivalente de corriger chaque p-value en la multipliant par le nombre de tests ($p_{cor} = p \times k$). Cela permet d'être sûr que $\alpha_{global} < 5\%$. Le problème de cette méthode est qu'elle est très conservative lorsque le nombre de groupes augmente, c'est-à-dire qu'elle rejette peu de différences. Il arrive alors souvent qu'une différence globale soit significative sans qu'aucune différence 2 à 2 n'apparaisse significative.

— **La méthode de Bonferroni-Holm** :

Une amélioration de cette méthode a été proposée, qui est aujourd'hui souvent préférée. Elle moins conservative, car la correction est moins drastique, tout en garantissant un risque $\alpha_{global} < 5\%$. Le principe est de :

— classer les valeurs de p par ordre croissant (p_1, p_2, ..., p_k) et

— corriger chaque p_i en le multipliant par $k + 1 - i$ ($p_{i.cor} = p_i \times (k + 1 - i)$)

Les logiciels donnent en général comme sortie des comparaisons multiples les valeurs de p-value corrigées, qui ne peuvent plus vraiment être interprétées comme des probabilités, mais sont comparées à 5% pour savoir si l'on peut rejeter l'hypothèse nulle pour chaque différence.

Voici à titre d'exemple les sorties de **R** correspondant aux comparaisons multiples sur notre exemple, réalisées sans correction, puis avec les corrections de Bonferroni et Bonferroni-Holm.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  d$duree and d$taille
##
##   XL   L    M
## L 0.01 -    -
## M 0.17 0.09 -
## S 0.03 0.53 0.25
##
## P value adjustment method: none
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  d$duree and d$taille
##
##   XL   L    M
## L 0.07 -    -
## M 1.00 0.54 -
## S 0.17 1.00 1.00
```

```
##
## P value adjustment method: bonferroni
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  d$duree and d$taille
##
##   XL   L    M
## L 0.07 -    -
## M 0.52 0.36 -
## S 0.15 0.53 0.52
##
## P value adjustment method: holm
```

? ? ? Essayer d'interpréter les sorties obtenues avec ces trois méthodes, et demandez-vous à quoi correspondent en réalité les valeurs à 1.00.

Il existe d'autres méthodes de correction de p-value, dont voici un petit récapitulatif dans un cadre un peu plus général de tests répétés (pas forcément uniquement de comparaisons multiples suite à une comparaison globale de moyennes) :

- **Comparaisons 2 à 2** : $k = \frac{p(p-1)}{2}$ comparaisons
  de très nombreuses autres méthodes disponibles (Tukey, Duncan, Rodger, Scheffé, Dunn-Sidak, ...), avec prédominance actuelle de la méthode **Bonferroni-Holm**.
- **Comparaisons à un groupe témoin** : $k = p - 1$ comparaisons
  méthode paramétrique de **Dunnett** couramment employée : statistique de Student avec estimation d'un σ global et correction des valeurs de p adaptée à ce cas particulier.
- **Tests répétés avec maîtrise du taux de fausses découvertes** : méthode de **Benjamini-Hochberg** (méthode couramment utilisée par exemple en transcriptomique - analyse de l'expression d'un très grand nombre de gènes, avec un test statistique par gène).

Certaines de ces méthodes peuvent être utilisées pour la comparaison d'autres paramètres statistiques que des moyennes, notamment à l'issue d'un test de comparaison de plusieurs fréquences sur séries indépendantes.

Les comparaisons multiples ne sont pas du tout une obligation à l'issue d'une comparaison globale de moyennes, et souvent elles n'apportent pas grand chose de plus qu'une bonne représentation graphique. Retenons les points suivants quant à leur utilisation :

- Il est indispensable de **vérifier que la différence globale est significative avant** de faire des comparaisons multiples (origine du terme "protected" dans PLSD de Fisher).
- Il faut **corriger le risque** α lors de la réalisation de comparaisons multiples **si l'on souhaite limiter le nombre de faux positifs** (rejets à tort de H_0).
- Les comparaisons multiples suite à une comparaison globale ne sont **pas à préconiser systématiquement** : elles n'apportent souvent pas grand chose à l'analyse globale et sont souvent difficiles à interpréter.

— **NE JAMAIS OUBLIER** qu'une différence non significative ne permet pas de conclure à une non différence.

— **TOUJOURS PENSER à interpréter les effets (différences entre groupes)** : ne pas rester au niveau des valeurs de p.

Voici un petit récapitulatif qui pourra vous aider pour le choix du bon test lors de la comparaison de moyennes, auquel j'ai ajouté, pour vous aider lors des travaux dirigés du second semestre, le nom des fonctions du langage **R** que nous utiliserons.

— **Un seul échantillon**
  test de conformité de Student (`t.test()`) ou test de la médiane

— **Deux échantillons indépendants**
  test de Student avec variances égales ou non (test de Welch - `t.test()`) ou test de la somme des rangs de Mann-Whitney-Wilcoxon (`wilcox.test()`)

— **Deux échantillons dépendants (appariés)**
  test de Student des séries appariés (`t.test()`) ou test des rangs signés de Wilcoxon (`wilcox.test()`)

— **Plusieurs échantillons indépendants**
  ANOVA (`oneway.test()`)ou test de la somme des rangs de Kruskal-Wallis (`kruskal.test()`)

Et souvenez-vous, il est capital de bien examiner visuellement la ou les distributions observées afin de choisir entre une approche paramétrique et une approche non paramétrique.

6 Corrélation linéaire et régression linéaire simple

6.1 Objectifs pédagogiques

A l'issue de l'étude de ce chapitre, vous devriez :

— Connaître la définition et les conditions d'utilisation du coefficient de corrélation linéaire (de Pearson).

— Connaître le principe et les conditions d'utilisation des tests paramétriques et non paramétriques de corrélation ainsi que les limites de ces tests.

— Savoir identifier dans la pratique les cas où l'utilisation d'un test de corrélation n'est pas approprié et dans les autres cas savoir faire le choix entre le test paramétrique et le test non paramétrique, et réaliser ces tests.

— Savoir interpréter les conclusions de ces tests.

— Connaître le modèle utilisé en régression linéaire simple et la méthode d'estimation de ses paramètres à partir de données.

— Savoir expliquer ce que représente la valeur de r^2 dans le cadre de la régression linéaire.

— Savoir identifier les cas sur lesquels il convient d'utiliser une régression linéaire et dans ces cas identifier la variable indépendante (explicative) et la variable dépendante (à expliquer).

— Savoir interpréter les résultats d'une régression linéaire issus d'un logiciel et vérifier ses conditions d'utilisation.

— Savoir utiliser un modèle de régression linéaire en prédiction (avec distinction entre les deux intervalles de confiance associés).

— Savoir ce qui distingue régression et corrélation linéaire.

6.2 La corrélation linéaire

Dans ce chapitre nous allons aborder les méthodes visant à mettre en évidence une **corrélation entre deux variables quantitatives**. Partons d'un exemple tiré de la littérature (Figure 50).

Plus classiquement, nous représentons ce type données par un **nuage de points**, représentation appelée aussi **diagramme de dispersion** (cf. Figure 51).

6.2.1 Le test de corrélation linéaire de Pearson

Le test de corrélation linéaire de Pearson vise à montrer une corrélation linéaire entre deux variables quantitatives continues. Supposons x et y 2 variables aléatoires observées sur un échantillon aléatoire simple et distribuées suivant une loi normale bivariée (cf. Figure 52), et définissons le coefficient de corrélation linéaire de la façon suivante :

$$r = \frac{Cov(x,y)}{\sqrt{V(x)V(y)}} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 \times \frac{1}{N}\sum_{i=1}^{N}(y_i-\overline{y})^2}}$$

r peut être considéré comme un "indicateur unidirectionnel de l'allongement du nuage de points" : $-1 \leq r \leq 1$ et plus les points sont alignés et plus $|r|$ est proche de 1.

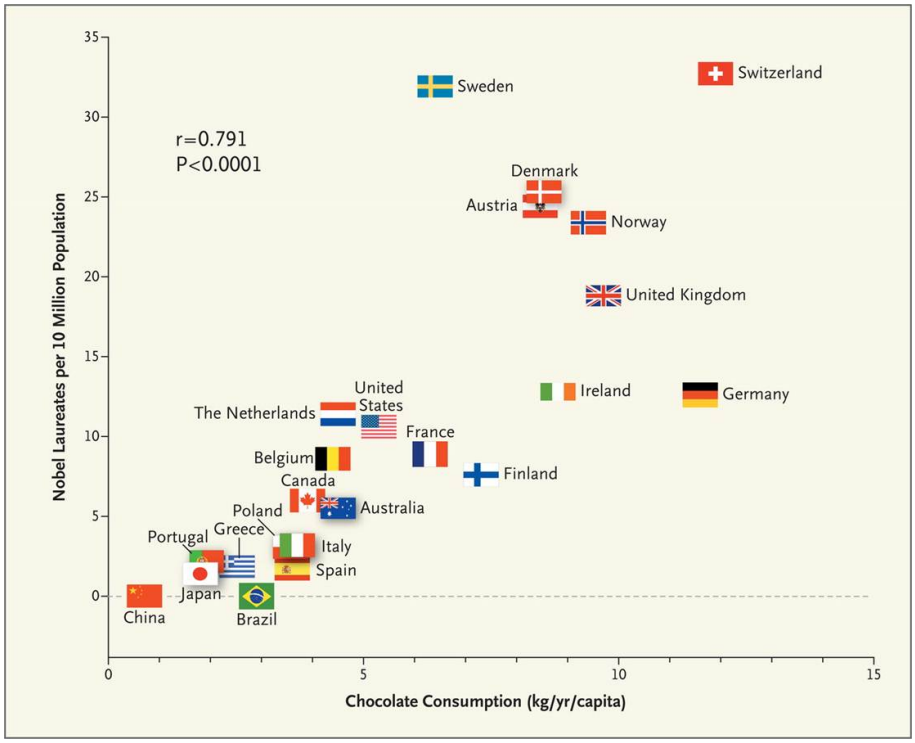


FIGURE 50 – Figure extraite de Messerli (2012), Chocolate Consumption, Cognitive Function, and Nobel Laureates, *the New England Journal of Medicine*

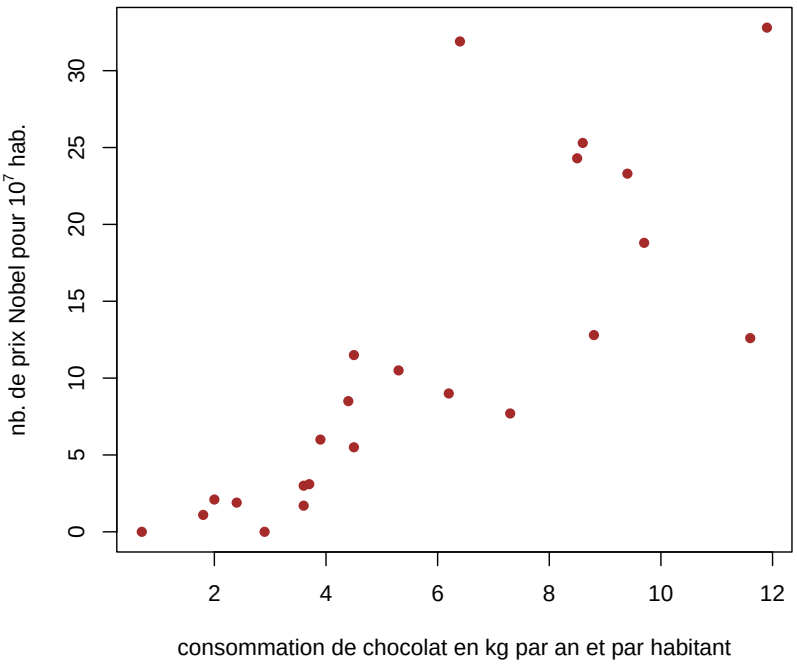


FIGURE 51 – Représentation classique sous forme de nuage de points (ou diagramme de dispersion) des données de l'exemple

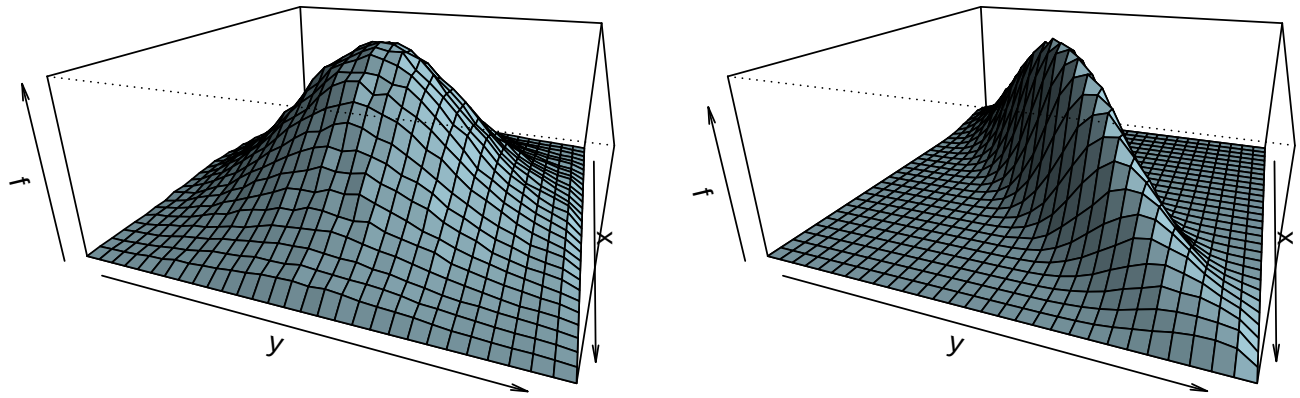


FIGURE 52 – Représentation de la densité de probabilité de deux distributions normales bivariées, à gauche sans corrélation entre x et y ($r = 0$) et à droite avec une forte corrélation entre x et y ($r = 0.9$)

Sous l'hypothèse nulle (H_0 : "absence de corrélation entre x et y"),

$$t = \sqrt{\frac{r^2(N-2)}{1-r^2}} \sim T(N-2)$$

avec N le nombre de points observés.

Nous pourrons donc utiliser t ainsi défini comme variable de décision pour tester la corrélation linéaire entre x et y. Ce test étant très peu robuste, en particulier très sensible aux valeurs extrêmes (un seul point éloigné des autres pourra modifier de façon très importante la valeur du coefficient de corrélation), il est indispensable de bien vérifier ses conditions d'utilisation avant de l'utiliser. Il est généralement très difficile de représenter la distribution conjointe de x et y comme dans la Figure 52 (cela nécessiterait un très grand nombre de points observés), mais si la distribution est normale bivariée le nuage de points est de type elliptique (cf. Figure 53 pour un exemple) et c'est ce que nous vérifierons avant d'appliquer le test de corrélation linéaire dans ce contexte.

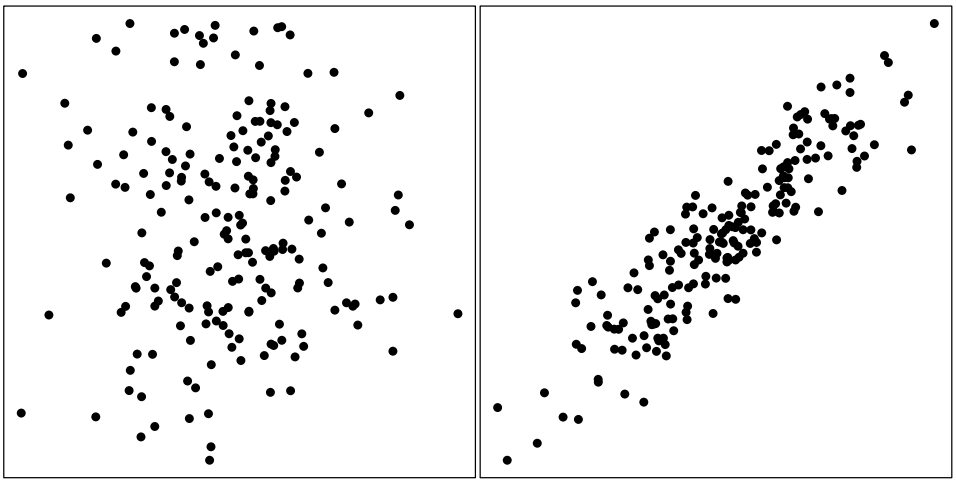


FIGURE 53 – Exemples de nuages de points issus de lois normales bivariées à gauche sans corrélation ($r = 0$), à droite avec une forte corrélation ($r = 0.9$)

? ? ? Interprétez les résultats du test de corrélation de Pearson réalisé sur les données de notre exemple avec R (cf. ci-dessous).
Pensez-vous que ce test est applicable sur ces données ?

```
##
##  Pearson's product-moment correlation
##
## data:  d$Chocolate and d$Nobels
## t = 6, df = 21, p-value = 6e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.567 0.909
## sample estimates:
##   cor
## 0.794
```

6.2.2 Le test de corrélation de rangs de Spearman

Revenons à notre exemple. Au vu de la figure 51, il semble difficile de considérer le nuage de points comme elliptique, et il serait imprudent d'utiliser le test de corrélation linéaire de Pearson. Dans un tel cas, lorsque la relation entre les deux variables observées semble monotone mais que le nuage de points n'est pas elliptique, le **test de corrélation de rangs de Spearman** peut être adapté. Son principe est très simple. On classe les valeurs de x d'un côté, et celle de y de l'autre et on associe à chaque point du nuage le rang de x et le rang de y, puis on calcule le coefficient de corrélation linéaire sur les rangs des x et les rangs des y. On appelle ce coefficient le coefficient de rangs de Spearman et on pourra faire un test à partir de celui-ci (cf. illustration Figure 54). Nous n'aborderons pas les détails techniques de ce test dans ce cours, la compréhension de son principe nous suffisant. On comprend bien, notamment, que le fait de

calculer un coefficient sur les rangs des observations rend le test beaucoup plus robuste par rapport aux éventuelles valeurs extrêmes.

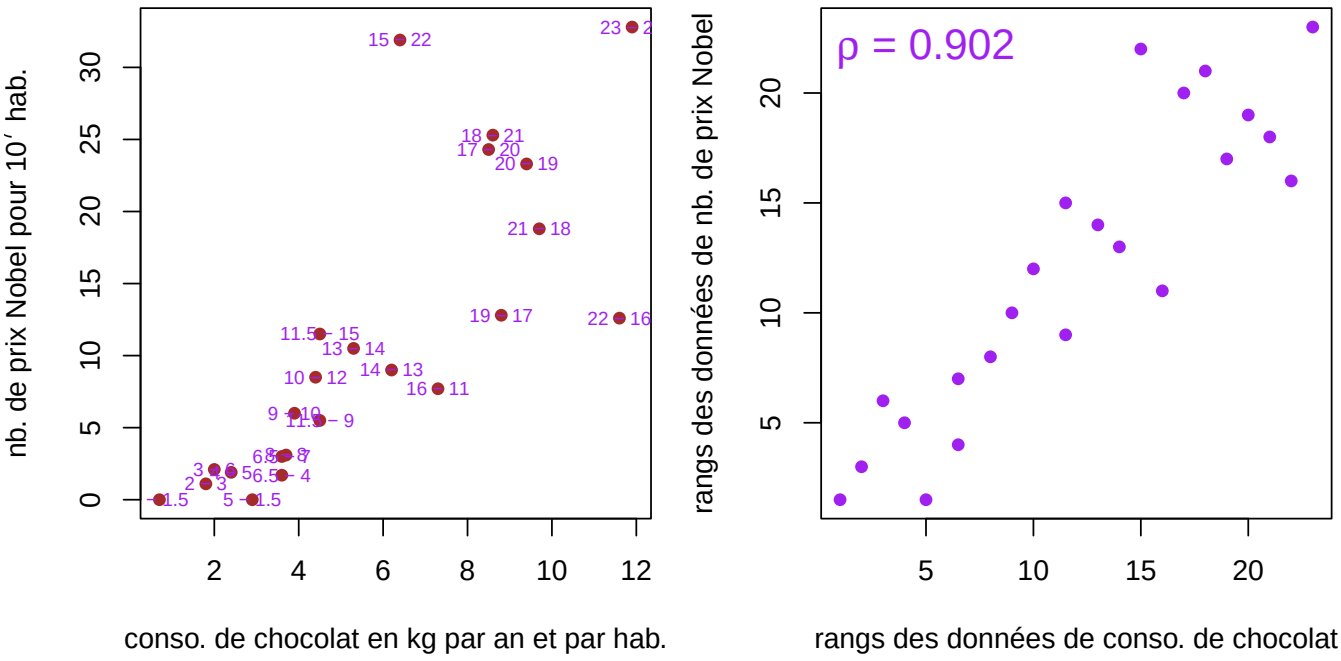


FIGURE 54 – Illustration du principe du calcul du coefficient de corrélation de rangs de Spearman

? ? ? Interprétez les résultats du test de corrélation de rangs de Spearman réalisé sur les données de notre exemple avec R (cf. ci-dessous).

```
##
##  Spearman's rank correlation rho
##
## data:  d$Chocolate and d$Nobels
## S = 198, p-value = 4e-09
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##   rho
## 0.902
```

A partir de ces résultats on conclut donc à une corrélation significative entre la consommation individuelle de chocolat dans les états et le nombre de prix Nobel pour 10 millions d'habitants. **MAIS ATTENTION !** On n'en déduira bien entendu pas de lien de causalité. **Une corrélation entre 2 variables observées n'implique pas forcément un lien de causalité**. Une corrélation peut être :

— due à un facteur de causalité commun (ex. de la corrélation entre vente de glaces et noyades, cause commune : la chaleur)

— due à une causalité dans le sens opposé à celui présenté (ex. : myopie des enfants et veilleuse, effet cigogne, ... cf https://www.chosesasavoir.com/quest-leffet-cigogne/ et https://cortecs.org/la-zetetique/effets-cigogne-correlation-vs-causalite/),

— complètement fortuite (cf. site qui montre des corrélations fortuites entre deux variables suivies au cours du temps (http://www.tylervigen.com/spurious-correlations).

En réponse à l'étude de Messerli (chocolat / prix Nobel), des chercheurs ont d'ailleurs montré une corrélation entre la consommation de chocolat et le nombre de tueurs en séries que le pays engendre).

6.2.3 Les limites des tests de corrélation

Une grande prudence s'impose dans le cadre de l'utilisation et de l'interprétation des coefficients de corrélation. Retenons ces différents points.

— Le test de corrélation de Pearson n'est pas adapté en cas de corrélation non linéaire (ex. Figure 55 a) sur lequel un test de corrélationde rangs serait plus adapté).

— Le test de corrélation de Pearson n'est pas du tout robuste (très influencé notamment par les valeurs extrêmes) et ne doit donc surtout pas être utilisé sur un nuage de points comme représenté Figure 55 b) sur lequel un test de corrélation de rangs serait plus adapté)

— Les tests de corrélation (Pearson et Spearman) ne sont pas adaptés en cas de corrélation non monotone (ex. Figure 55 c) ou les coefficients de corrélation seraient proches de 0, bien qu'il y ait une relation non monotone entre les deux variables) et plus généralement de nuage de points non elliptique (ex. Figure 55 d) pour lequel la relation ne peut pas être décrite sous forme de corrélation simple : on a plus l'impression que plus x est grand et plus y est variable)

— Les tests de corrélation (Pearson et Spearman) ne sont pas adaptés en cas de nuage de points formé de sous-nuages (sous-groupes se distinguant comme en Figure 55 e) et e bis) avec couleurs et remplissage différentiel des points - l'ex. fictif représente le poids du bagage en y et le poids du voyageur en x, avec les hommes en cercles vides bleu et les femmes en cercles pleins rouge).

On ne devrait donc jamais reporter une valeur de r (coefficient de corrélation linéaire de Pearson) ou de ρ (coefficient de corrélation de rangs de Spearman) non assortie du nuage de points, ni interpréter une valeur de r ou de ρ tant qu'on n'a pas vu le nuage de points associé !

? ? ? Examinez de plus près le dernier exemple de la figure 55. Si l'on effectuait "bêtement" les tests de corrélation sur ce nuage de points on obtiendrait $r = -0.47$ ($p < 0.0001$) et $\rho = -0.52$ ($p < 0.0001$). Qu'en conclurait-on quand à la corrélation entre le poids du bagage et le poids du voyageur ?
Maintenant si l'on effectue ces tests sur chacun des deux sous-groupes, on obtient $r = 0.48$ ($p < 0.0001$) pour les femmes et $r = 0.41$ ($p < 0.0001$) pour les hommes. Qu'en conclut-on ?

Nous nous sommes placés dans ce chapitre dans le cas où les deux variables quantitatives à corréler sont toutes les deux observées, c'est-à-dire qu'aucune des deux n'est contrôlée (valeurs fixées par l'expérimentateur). Les tests de corrélation peuvent aussi s'appliquer dans le cas où l'une des variables est contrôlée, et nous aborderons ce cas dans le chapitre suivant.

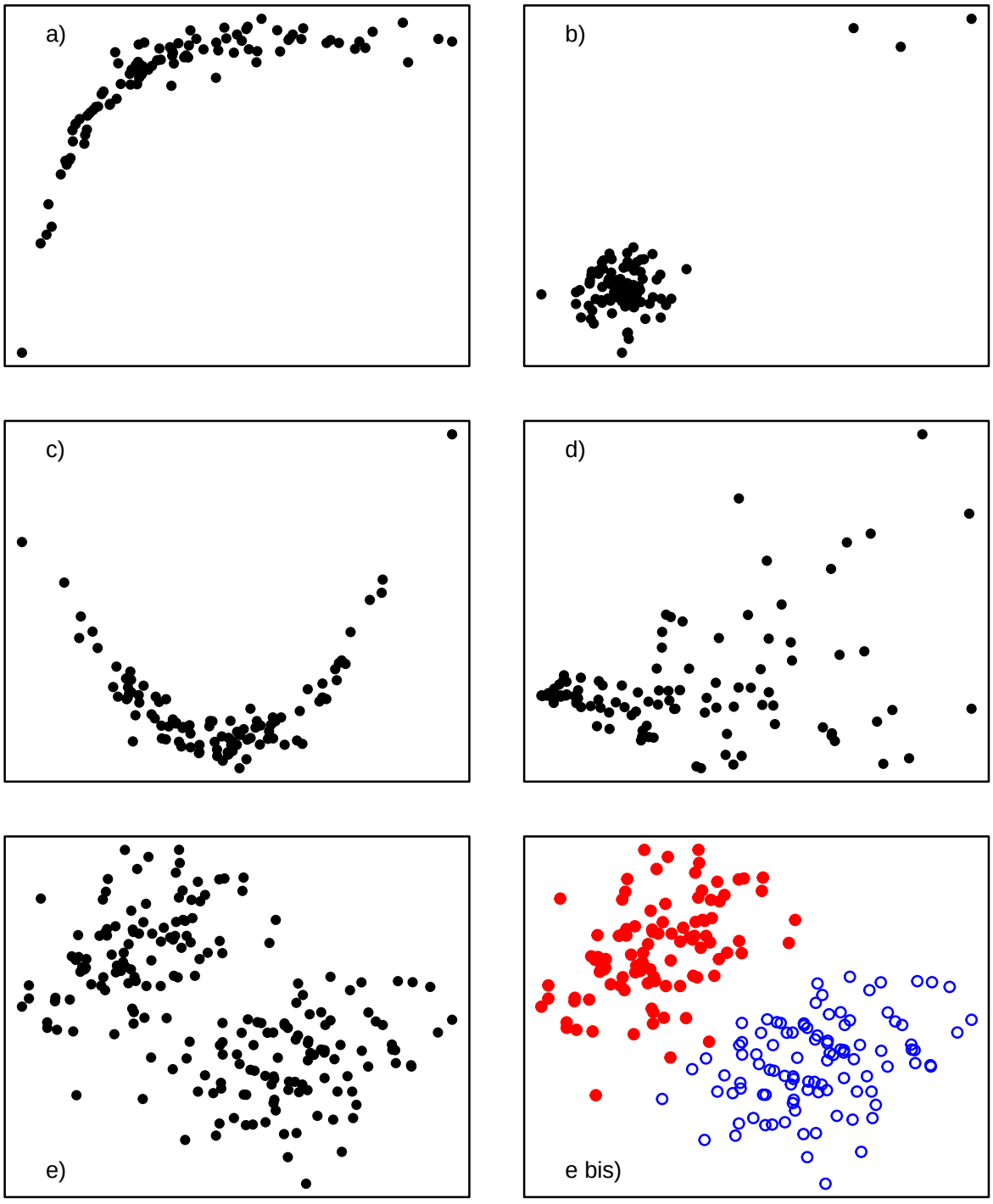


FIGURE 55 – Illustration de quelques nuages de points qui devraient nous imposer la prudence !

6.3 Bilan, mise en garde et transition

Nous avons vu dans le cours jusque là toutes les méthodes qui permettent de **mettre en corrélation deux variables** :

— **deux variables qualitatives** : chapitre 4
— **une variable qualitative, une variable quantitative** : chapitre 5
— **deux variables quantitatives** : chapitre 6

Etant donné que la majorité des test vus précédemment visent à démontrer l'existence d'une corrélation entre deux variables, il serait bien imprudent de penser que vous allez trouver le test adapté dans le chapitre "corrélation linéaire" dès que vous voyez le terme "corrélation" dans la question posée !
Encore une fois, en statistique l'important est la compréhension des méthodes, et le simple repérage de mots clefs risquerait fort de vous induire en erreur.

Nous allons maintenant aborder, sur un exemple encore très simple à deux variables, un thème plus général qui est à la base de très nombreuses approches qui permettent d'aborder des cas plus complexes : la **modélisation**.

6.4 La régression linéaire simple

La régression linéaire simple est une méthode de **modélisation**, par une relation linéaire, de la relation entre une **variable quantitative observée et une variable quantitative contrôlée**. Comme toute démarche de modélisation, elle peut être utilisée pour mieux caractériser / comprendre cette relation et/ou pour l'utiliser en prédiction. Elle est parfois utilisée à la marge de ce cadre, avec des variables quantitatives toutes deux observées, mais qui ont des statuts différents (contrairement au cadre du chapitre précédent) : on veut expliquer / prédire l'une d'entre elle en fonction de l'autre.

Partons d'un exemple inspiré de la littérature (Roomi *et al.* 2011, Nutrient mixture inhibits *in vitro* and *in vivo* growth of human acute promyelocytic leukemia HL-60 cells, *Experimental Oncology*).

Cette publication étudie l'impact, in vitro, d'un mélange de nutriments (acide ascorbique, extrait de thé vert, lysine, proline, ...) sur la prolifération de cellules tumorales. La variable contrôlée est la dose de nutriments en concentration dans le milieu ($\mu g.ml^{-1}$) et la variable observée est la prolifération cellulaire quantifiée en pourcentage de celle observée sans nutriments dans le milieu de culture. Les données sont représentées dans la figure 56, à gauche en données brutes, et à droite avec la dose en logarithme décimal.

6.4.1 Modèle de la régression linéaire simple

En régression linéaire simple on utilise donc un **modèle linéaire** pour expliquer **une variable à expliquer** Y appelée aussi variable **dépendante** (toujours une variable observée) en fonction **d'une variable explicative notée** X (souvent contrôlée mais pas toujours), appelée aussi variable **variable indépendante**.

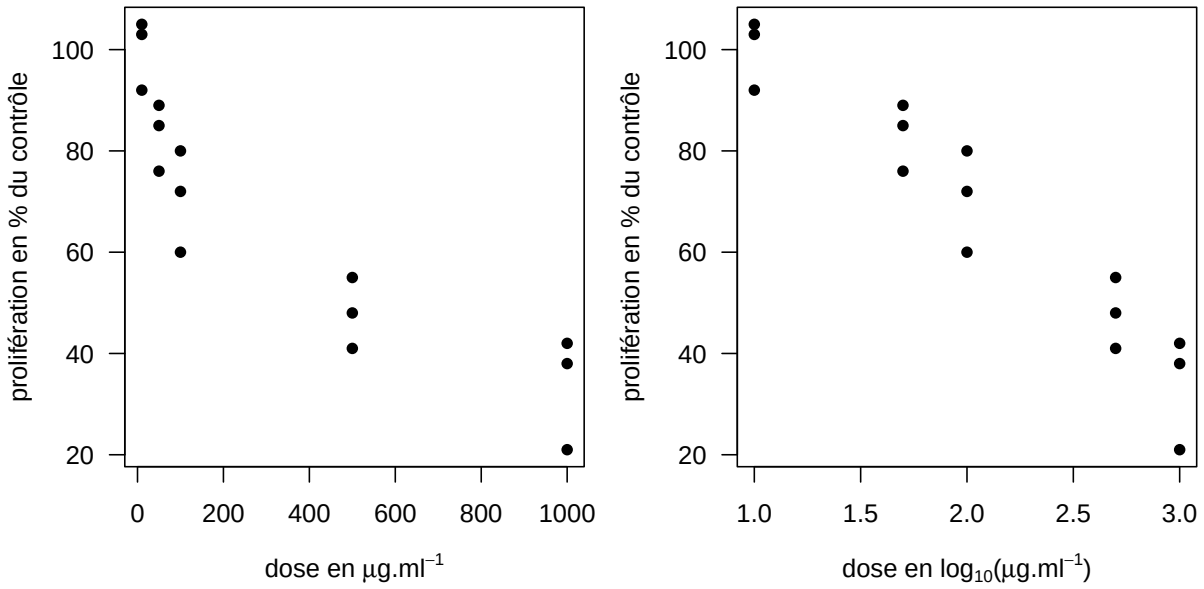


FIGURE 56 – Diagrammes de dispersion présentant les données de l'exemple, à gauche en données brutes, à droite avec la variable explicative transformée en logarithme décimal, en vue de linéariser la relation dans ce cas particulier.

Le modèle théorique de la régression linéaire simple, illustré Figure 57 est le suivant :

$$Y_i = \alpha + \beta X_i + \epsilon_i \text{ avec } \epsilon_i \sim N(0, \sigma)$$

On a dans ce modèle

— une partie déterministe sous forme de relation linéaire et

— une partie stochastique sous forme de modèle Gaussien avec des ϵ_i aléatoires, indépendants, suivant une loi normale (loi de Gauss) de variance résiduelle σ^2 constante.

Pour estimer les paramètres de ce modèle à partir des données observées, on utilise la méthode de maximisation de la vraisemblance (maximisation de $Pr(Y \mid \alpha, \beta, \sigma)$, les données Y étant connues), qui revient, dans le cadre du modèle Gaussien, à minimiser la Somme des Carrés des Ecarts (SCE, cf. illustration Figure 57) :

$$SCE = \sum_{i=1}^{n} e_i^2 \text{ avec } e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\alpha} + \hat{\beta} X_i).$$

Les estimations ponctuelles des paramètres du modèle linéaire Gaussien, obtenues par moindres carrés, sont les suivantes :

— Pente (ou coefficient de régression) : $\hat{\beta} = \frac{cov(X,Y)}{V(X)}$

— Ordonnée à l'origine ("intercept" en anglais) : $\hat{\alpha} = \overline{Y} - \hat{\beta} \times \overline{X}$

— Ecart type résiduel ("residual standard error" en anglais) : $\hat{\sigma} = \sqrt{\frac{SCE}{n-2}}$

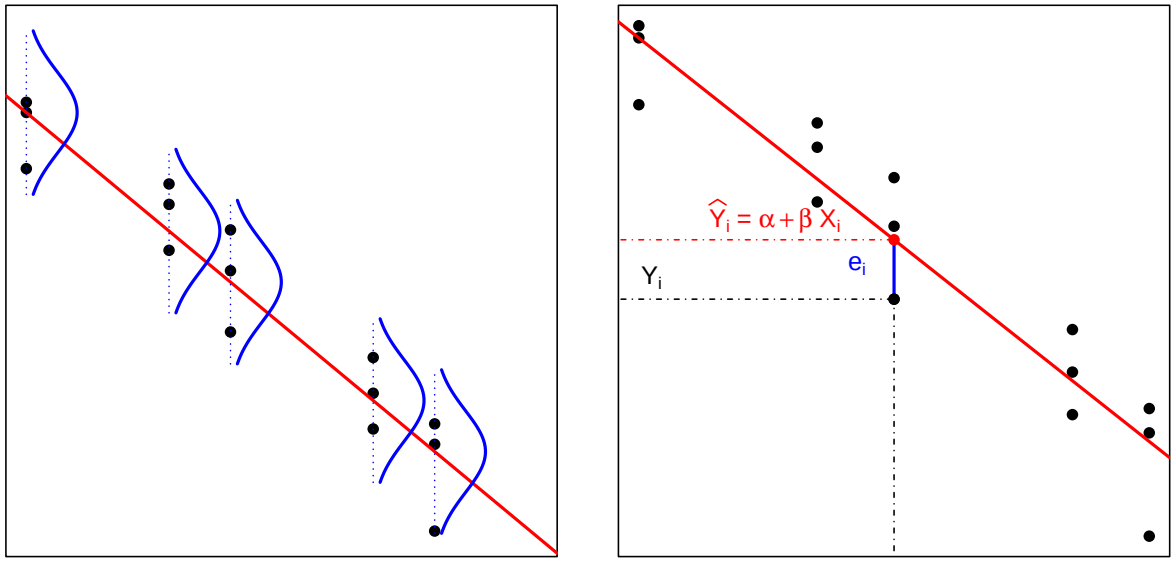


FIGURE 57 – Illustration des hypothèses du modèle linéaire Gaussien (à gauche) et de la méthode des moindres carrés associée (à droite)

Si l'on utilise **R** pour estimer ces paramètres sur notre exemple, on obtient la sortie suivante dans laquelle il vous faudra savoir repérer les différents résultats importants décrits en-dessous de la sortie R (n'hésitez pas à annoter cette sortie R sur votre polycopié pour bien savoir vous repérer à l'avenir) :

```
##
## Call:
## lm(formula = proliferation ~ log10(dose), data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15.57  -4.67   1.43   5.33  10.22
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   136.18       6.37    21.4  1.6e-11
## log10(dose)   -33.20       2.90   -11.5  3.6e-08
##
## Residual standard error: 8.01 on 13 degrees of freedom
## Multiple R-squared:  0.91, Adjusted R-squared:  0.903
## F-statistic:  131 on 1 and 13 DF,  p-value: 3.64e-08
```

— la pente (slope or regression coefficient) qui est présentée dans la sortie de R avec comme indication le nom de la variable explicative associée : $\hat{\beta} = -33.2$

— l'ordonnée à l'origine (intercept) : $\hat{\alpha} = 136$

— écart type résiduel (residual standard error) : $\hat{\sigma} = 8.01$

— et le coefficient de détermination (carré du coefficient de corrélation linéaire, dont on expliquera le sens plus loin) : $r^2 = 0.91$

Avant de tirer des conclusions des résultats du modèle, il conviendra de vérifier *a posteriori* les conditions d'utilisation du modèle. On utilise pour cela les résidus e_i, résidus qu'on ne peut calculer d'une fois le modèle ajusté, d'où la vérification *a posteriori*. Ces résidus seront représentés sous deux formes pour vérifier les hypothèses quant à leur distribution, sous forme d'un **graphe de résidus** représentant les résidus e_i en fonction des valeurs prédites $\hat{Y}_i$ (cf. Figure 58) et d'un diagramme Quantile-Quantile des résidus (cf. Figure 59)). Ces deux graphes sont complémentaires et permettent de s'assurer visuellement que les **résidus** sont bien **indépendants les uns des autres** et distribués suivant une **loi normale centrée en 0** et d'**écart type σ constant**.

On peut remarquer sur la sortie R précédente, que les premiers résultats donnés sont des statistiques descriptives des résidus, qui permettent aussi de se faire une idée rapide de leur distribution globale.

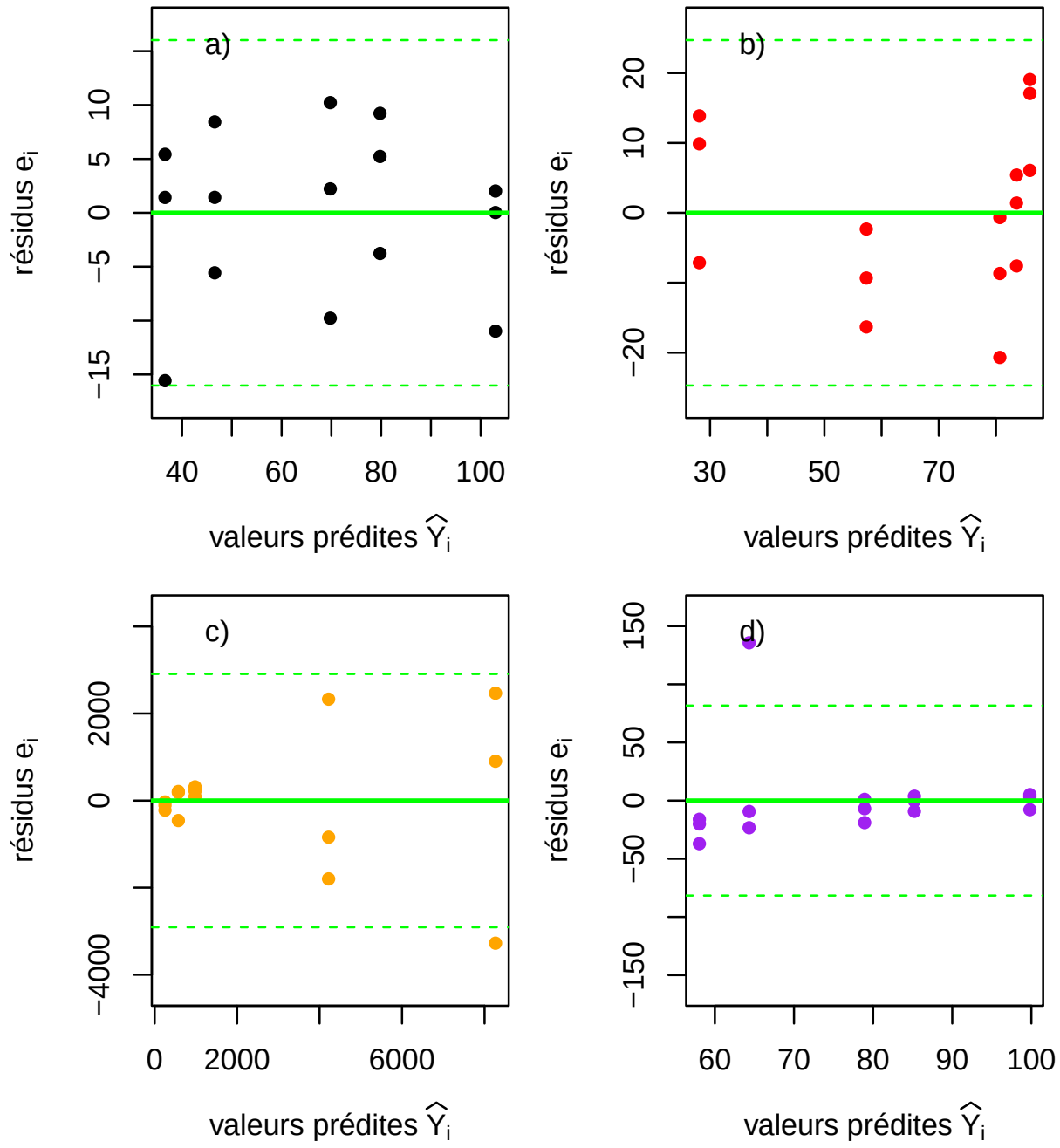


FIGURE 58 – Graphes des résidus a) dans le cadre de notre exemple où les conditions d'utilisation du modèle ne sont pas remises en cause et dans trois autres cas où ces conditions d'utilisation sont clairement non respectées, en b) car les résidus ne sont pas aléatoires, en c) à cause d'un écart type des résidus non constant (hétéroscédasticité des résidus caractérisés par un effet "entonnoir") et en d) à cause d'un résidu extrême. Les lignes vertes représentent en trait plein la valeur 0 attendue pour la moyenne des résidus, et en pointillés $-2\hat{\sigma}$ et $-2\hat{\sigma}$, la bande censée contenir à peu près 95% des résidus.

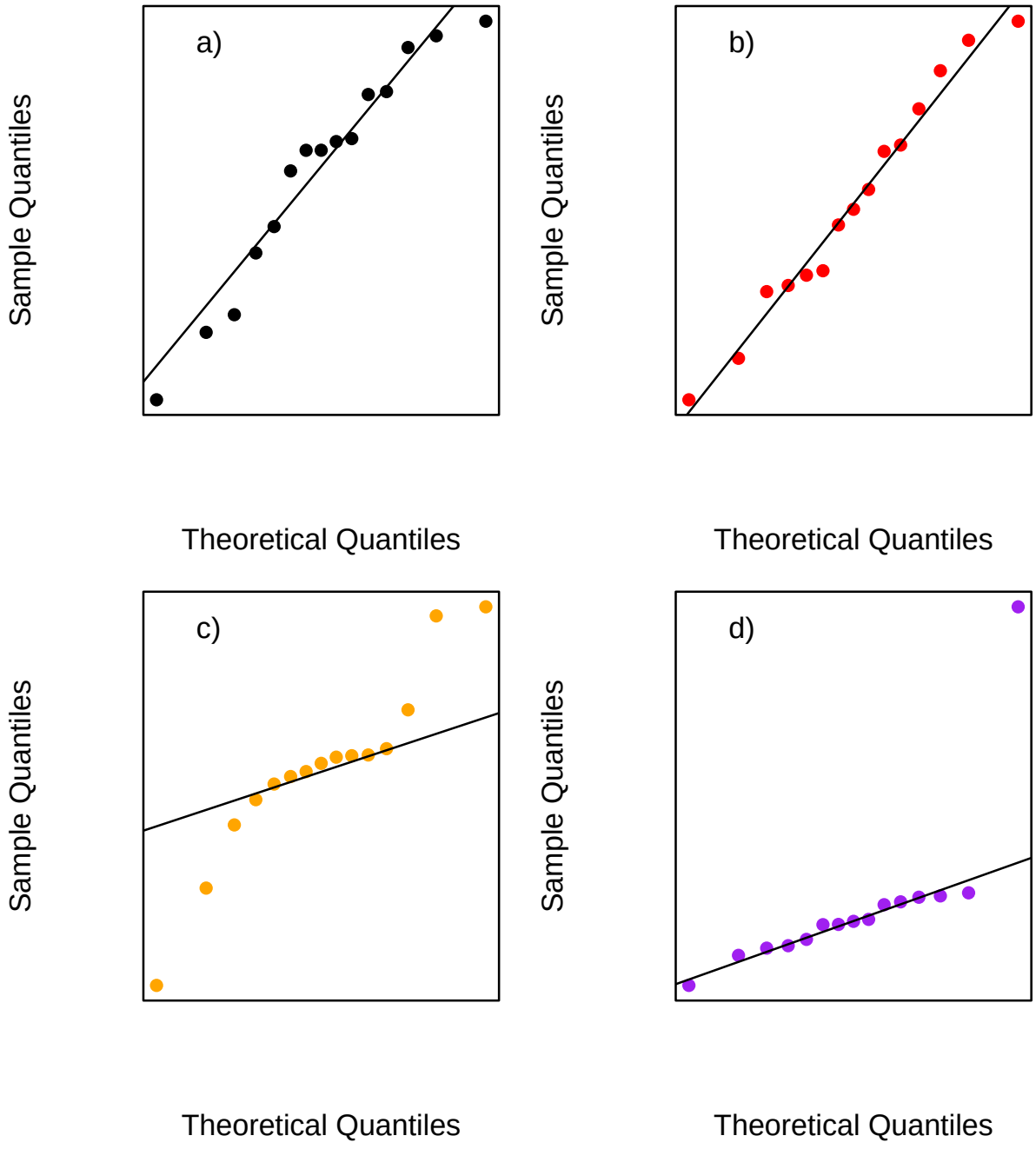


FIGURE 59 – Diagramme Quantile-Quantile des résidus a) dans le cadre de notre exemple où les conditions d'utilisation du modèle ne sont pas remises en cause et dans les trois autres cas où ces conditions d'utilisation sont non respectées (cf. Figure 58), ce qui n'apparaît clairement sur ce graphe que sur les cas c) ou la distribution des résidus n'apparaît pas normale et d) à cause d'un résidu extrême.

6.4.2 Prédiction et intervalles de confiance

Une fois les conditions d'utilisation du modèle linéaire Gaussien vérifiées, on utilise parfois directement les résultats de la régression linéaire par exemple pour estimer un paramètre biologique correspondant à l'un des paramètres du modèle (par exemple un taux de croissance microbien estimé comme la pente d'une courbe de croissance microbienne exprimée en logarithme de la concentration microbienne en fonction du temps). Il est alors important d'associer à leur estimation ponctuelle un intervalle de confiance, afin de quantifier leur incertitude. On utilisera une fonction R très simple pour obtenir les **intervalles de confiance l'ordonnée à l'origine et la pente** (cf. sortie ci-dessous sur notre exemple).

```
##                2.5 % 97.5 %
## (Intercept)  122.4   149.9
## log10(dose)  -39.5   -26.9
```

Il est aussi possible d'utiliser le modèle ajusté aux données pour prédire une valeur de Y_0 pour une valeur donnée $X = X_0$ prise dans le domaine étudié (couvert par le jeu de données). On prédit la valeur de Y pour $X = X_0$ dans le domaine étudié, par $\hat{Y_0} = \hat{\alpha} + \hat{\beta}X_0$ et on peut associer à une telle prédiction deux intervalles de confiance (cf. Figure 60 :

- l'intervalle de confiance sur la moyenne des Y pour $X = X_0$ qui quantifie l'incertitude sur la droite estimée,
- et l'intervalle de prédiction qui quantifie l'incertitude sur une observation individuelle prédite de Y pour $X = X_0$ et prend en compte non seulement l'incertitude sur la droite estimée, mais aussi la variabilité des observations autour de la droite suivant la loi $N(Y_0, \sigma)$. L'intervalle de prédiction à 95% est souvent approché par défaut par $\hat{Y_0} \pm 2 \times \hat{\sigma}$ (en pointillés sur la Figure 60), ce qui revient à négliger l'incertitude sur la droite estimée, et qui sera d'autant plus raisonnable qu'on a beaucoup de points observés et que X_0 se trouve proche du centre du domaine des X étudié.

A titre d'exemple, la sortie R ci-dessous vous indique dans l'ordre

- l'intervalle de confiance à 95% sur la moyenne prédite de prolifération cellulaire pour une dose de 200 $\mu g.ml^{-1}$,
- puis l'intervalle de prédiction à 95% pour cette même dose.

Pour chacun des intervalles, il faut lire en premier l'estimation ponctuelle de la prédiction, puis les bornes inférieure ("lower") et supérieure ("upper") de l'intervalle.

```
##    fit  lwr  upr
## 1 59.8 55.1 64.5
##    fit  lwr  upr
## 1 59.8 41.8 77.7
```

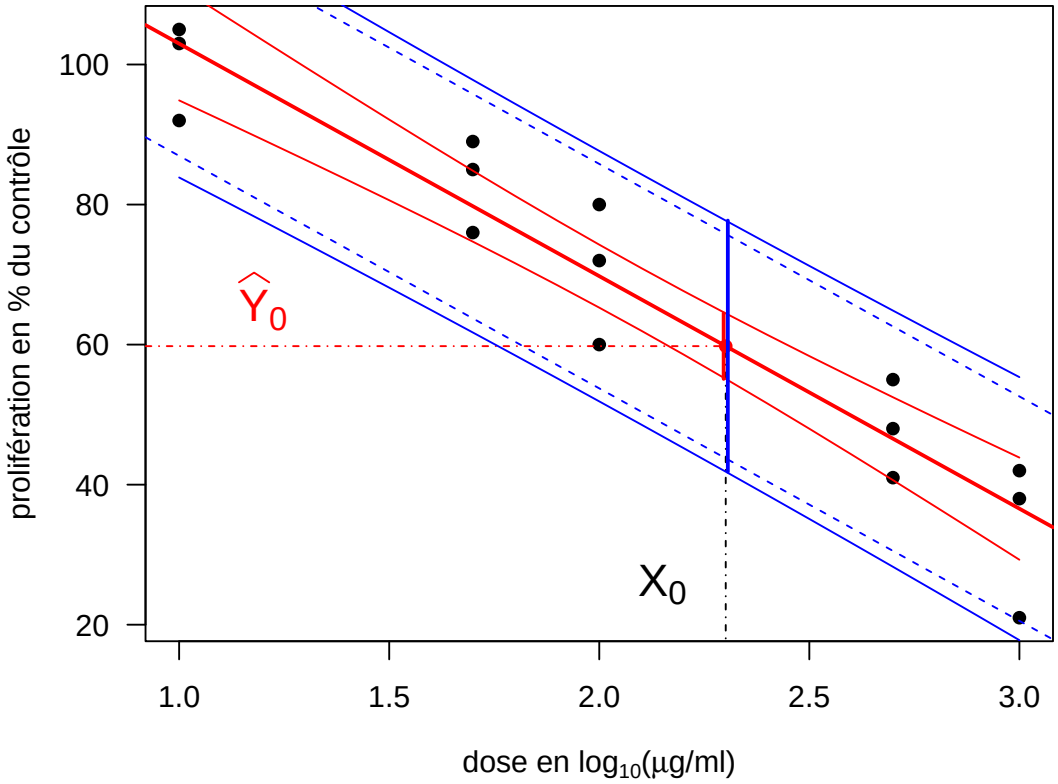


FIGURE 60 – Illustration des deux intervalles de confiance associés à une prédiction en régression linéaire, l'intervalle de prédiction sur une moyenne prédite (en rouge) et l'intervalle de prédiction, c'est-à-dire l'intervalle de confiance sur une prédiction individuelle (en bleu) ainsi que son approximation par défaut (en bleu pointillé).

Lorsque l'on vise l'**utilisation d'un modèle linéaire Gaussien en prédiction**, on calcule souvent le **coefficient de détermination** r^2 qui n'est autre que le **carré du coefficient de corrélation** défini au chapitre 6.2.1. Pourquoi donne-t-on son carré plutôt que le r lui-même ? Tout simplement parce qu'on peut en donner une interprétation en terme de part de variation de Y expliquée par le modèle. On peut en effet montrer mathématiquement que :

$$r^2 = \frac{cov(X,Y)^2}{V(X)V(Y)} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \text{variation de } Y \text{ expliquée / variation de } Y \text{ totale}$$

Interprétation du r^2
Le **coefficient de détermination** (carré du coefficient de corrélation linéaire de Pearson) est souvent **exprimé en pourcentage** et donc **interprété comme le pourcentage de la variation de Y qu'on pourra prédire à partir du modèle connaissant la variation de X** (dite couramment la part de variation expliquée par le modèle, la part restante étant la variation résiduelle).

Comme en corrélation, la valeur de r^2 ne permettra pas à elle seule de vérifier que le modèle est de bonne qualité, et il sera impératif de visualiser le graphe d'ajustement (points ajustés à la droite) et les graphes des résidus (on peut avoir un r^2 très proche de 1 avec des données qui ne respectent pas les conditions d'utilisation de la régression linéaire).

6.4.3 Régression et corrélation

Revenons maintenant au chapitre sur la corrélation et posons-nous deux questions.

— **Peut-on réaliser un test de corrélation linéaire dans le cadre de la régression linéaire ? OUI**,

Celui-ci correspond au test d'égalité à 0 de la pente, affiché classiquement dans le résumé de la régression et appelé test de signification de la pente. Il répond à la question : "y a-t-il un effet significatif de X sur Y ?"

— **Peut-on tracer une droite de régression dans le cadre de la corrélation linéaire, si X et Y sont deux variables observées qui ont des rôles symétriques ? NON**.

Le choix de la variable de contrôle (X) a un impact sur la droite de régression (cf. Figure 61), donc si X et Y ont des rôles symétriques, aucune des 2 droites n'a de justification.

Cette erreur est pourtant très fréquente !

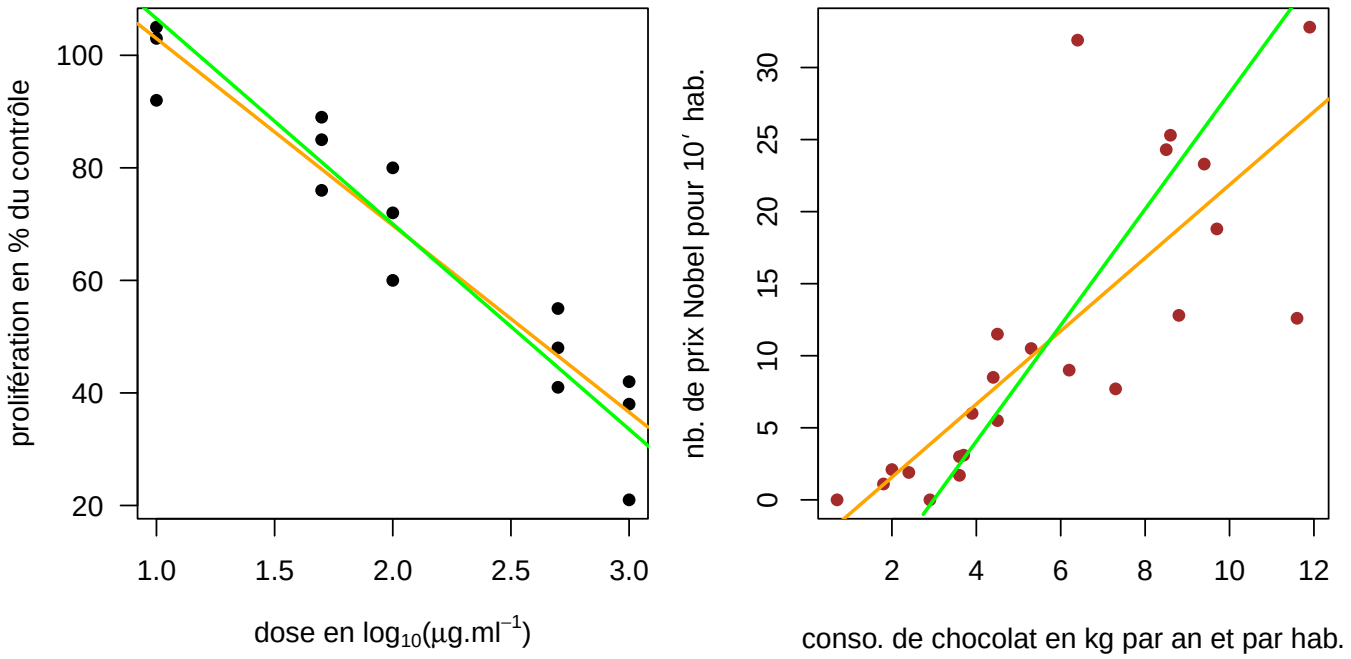


FIGURE 61 – Illustration de l'impact du choix de X et Y sur la droite de régression (en orange la droite de régression de Y en fonction de X et en vert la droite de régression de X en fonction de Y). Comparaison des 2 droites sur notre exemple à gauche et sur l'exemple que nous avions pris dans le chapitre corrélation à droite.

? ? ? En reprenant la définition de la méthode des moindres carrés utilisée pour estimer les paramètres du modèle en régression linéaire, expliquez pourquoi le choix des variables X et Y a une influence sur la droite de régression.

6.5 Le modèle linéaire simple, une brique de base pour construire d'autres modèles plus complexes

Dans ce chapitre nous avons abordé le modèle linéaire simple qui est une brique de base pour construire d'autres modèles plus complexes. Les notions que nous avons vues seront donc capitales pour quelqu'un qui voudrait aborder des modèles plus complexes. Parmi les complexifications du modèle dont nous aurons souvent besoin en médecine vétérinaire ou plus largement en biologie, citons-en quelques-uns :

— prise en compte de **plus d'une variable explicative** (*régression linéaire multiple*)
par exemple si l'on veut décrire l'effet à la fois de la température (variable explicative quantitative) et du pH (seconde variable explicative quantitative) sur un taux de croissance microbien (variable à expliquer quantitative),

— prise en compte de **variables explicatives quantitatives et qualitatives** (*modèle linéaire, ANOVA à plusieurs facteurs*)
par exemple si l'on veut comparer la croissance d'animaux (sur la base d'une variable à expliquer quantitative) nourris avec des aliments différents (variable explicative qualitative) en prenant en compte l'âge des animaux (variable explicative quantitative),

— modélisation d'une **variable à expliquer** non pas quantitative mais **qualitative binaire** (*régression logistique*)
par exemple si l'on veut évaluer l'impact sur le risque de maladie (variable binaire observée sur chaque animal : malade / non malade) de diverses variables explicatives qualitatives et/ou quantitatives,

— prise en compte de **facteurs aléatoires** (*modèles mixtes*)
par exemple si l'on doit prendre en compte le fait que les observations ne sont pas indépendantes les unes des autres parce que plusieurs observations sont réalisées sur chaque animal (mesures répétées, généralisation des séries appariées).

Les modèles associés (et d'autres encore) ne sont pas au programme de ce cours mais vous pourrez vous y former au sein du campus vétérinaire de Lyon dans le cadre de l'enseignement optionnel de A6 (https://biostatistique.vetagro-sup.fr/formcont.html).

7 Traduction anglaise des termes clefs qui ne sont pas évidents à traduire

Terme français	Terme anglais
comparaison deux à deux	pairwise comparisons
coefficient de corrélation	correlation coefficient
coefficient de détermination	coefficient of determination (often simply called R-squared)
diagramme de dispersion ou nuage de point	scatter plot
diagramme en bâtons	bar chart of barplot
diagramme en boîte à moustaches	boxplot
diagramme en secteurs	pie chart
diagramme quantile-quantile	QQ plot
écart type	standard deviation (SD)
écart type résiduel	residual standard error
effectifs	frequency counts (ATTENTION, faux ami !)
erreur standard de la moyenne	standard error of the mean (SEM)
fonction de densité de probabilité	density function
fonction de répartition empirique	empirical cumulative distribution fonction (ECDF)
intervalle de confiance	confidence interval
intervalle de fluctuation	fluctuation interval
ordonnée à l'origine	intercept
pente ou coefficient de régression	slope or regression coefficient
test de signification	significance test OR Null Hypothesis Significance Testing (NHST)
test d'hypothèse	hypothesis test
test d'équivalence	equivalence test
test de la somme des rangs	rank sum test
test des rangs signés	signed rank test
test de Student	T test
valeurs extrêmes	outliers

TABLE 9 – Table de traduction des termes clef en anglais.

8 Récapitulatif des principales fonctions R pour mettre en oeuvre les tests et la régression linéaire

Voici la liste des fonctions utilisées pour réaliser les différents tests au programme et ajuster un modèle linéaire, dans leur ordre d'apparition dans le guide R qui vous sera fourni au second semestre, avec la spécification de leurs arguments, et leur nom en anglais donné dans la sortie de R.

Test de normalité de Shapiro-Wilk (Shapiro-Wilk normality test)

```
shapiro.test(variable_quanti)
```

Test de Student de conformité à une moyenne théorique (One Sample t-test)

```
t.test(variable_quanti, mu = moyenne_theorique)
```

Test du χ^2 d'ajustement (Chi-squared test for given probabilities)

```
chisq.test(vecteur_effectifs_obs, p = vecteur_proportions_theo)
```

Test exact (utilisant la loi binomiale) de comparaison d'une fréquence observée à une fréquence théorique (Exact binomial test)

```
binom.test(nb_realisations_evenement, nb_tirages_total, p = proportion_theo)
```

Test de Student de comparaison de 2 moyennes sur séries indépendantes avec variances égales (Two Sample t-test)

```
t.test(variable_quanti ~ groupe_variable_quali, var.equal = TRUE)
```

Test de Welch de comparaison de 2 moyennes sur séries indépendantes avec variances inégales (Welch Two Sample t-test)

```
t.test(variable_quanti ~ groupe_variable_quali, var.equal = FALSE)
```

Test non paramétrique de la somme des rangs de Mann-Whitney-Wilcoxon de comparaison de séries indépendantes (Wilcoxon rank sum exact test)

```
wilcox.test(variable_quanti ~ groupe_variable_quali)
```

Test de Fisher de comparaison de 2 variances (F test to compare two variances)

```
var.test(variable_quanti ~ groupe_variable_quali)
```

Test du χ^2 d'indépendance réalisé à partir de la table de contingence (Pearson's Chi-squared test)

```r
chisq.test(table_contingence)
# Fonction aussi utilisable pour la comparaison de 2 fréquences sur séries
# indépendantes qui donne en plus l'intervalle de confiance sur la différence
# entre les 2 fréquences comparées - ATTENTION de bien vérifier
# que les fréquences affichées en sorties (prop 1 et prop 2)
# sont bien celles sur lesquelles vous voulez calculer cette différence
prop.test(table_contingence)
```

Test exact de comparaison de 2 fréquences sur séries indépendantes (Fisher's Exact Test for Count Data)

```r
fisher.test(table_contingence)
```

Test de Student de comparaison de 2 moyennes sur séries appariées (Paired t-test)

```r
t.test(variable_quanti_1, variable_quanti_2, paired = TRUE)
```

Test non paramétrique de Wilcoxon des rangs signés de comparaison de séries appariées (Wilcoxon signed rank test)

```r
wilcox.test(variable_quanti_1, variable_quanti_2, paired = TRUE)
```

Test de Mc Nemar de comparaison de 2 fréquences sur séries appariées, à partir de la table de concordance (McNemar's Chi-squared test)

```r
mcnemar.test(table_concordance)
```

Test de comparaison de plusieurs moyennes sur séries indépendantes en supposant les variances égales - analyse de variance classique avec variances égales (One-way analysis of means)

```r
oneway.test(variable_quanti ~ groupe_variable_quali, var.equal = TRUE)
```

Test de comparaison de plusieurs moyennes sur séries indépendantes sans supposer les variances égales extension du test de Welch - analyse de variance classique avec variances inégales (One-way analysis of means (not assuming equal variances))

```r
oneway.test(variable_quanti ~ groupe_variable_quali, var.equal = FALSE)
```

Test non paramétrique de la somme des rangs de Kruskal-Wallis de comparaison de séries indépendantes (Kruskal-Wallis rank sum exact test)

```
kruskal.test(variable_quanti ~ groupe_variable_quali)
```

Test de comparaison de plusieurs variances (Bartlett test of homogeneity of variances)

```
bartlett.test.test(variable_quanti ~ groupe_variable_quali)
```

Test de Cochran-Mantel-Haenszel de comparaison de plusieurs fréquences sur séries dépendantes (Cochran-Mantel-Haenszel test)

```
mantelhaen.test(variable_quali_A, variable_quali_B, identifiant)
```

Test de corrélation linéaire de Pearson (Pearson's product-moment correlation)

```
cor.test(variable_quanti_A, variable_quanti_B, method = "pearson")
```

Test non paramétrique de corrélation de rangs de Spearman (Spearman's rank correlation rho)

```
cor.test(variable_quanti_A, variable_quanti_B, method = "spearman")
```

Régression linéaire simple (simple linear model)

```
# Spécification du modèle
modele <- lm(variable_quanti_Y_a_expliquer ~ variable_quanti_X_explicative,
             data = jeu_de_donnes)
# Graphes des résidus
plot(predict(modele), residuals(modele))
qqnorm(residuals(modele))
# Paramètres estimés et diverses statistiques résumées
summary(modele)
# Intervalle de confiance sur les paramètres estimés
confint(modele)
# Intervalles de confiance sur les moyennes prédites pour des valeurs de X données
predict(modele, data.frame(variable_quanti_A_expliquer = vecteur_valeurs_X),
        interval = "confidence")
# Intervalles de confiance sur des prédictions individuelles pour
#   des valeurs de X données
predict(modele, data.frame(variable_quanti_A_expliquer = vecteur_valeurs_X),
        interval = "prediction")
```

Références

Wasserstein, R. L. and Lazar, N. A. (2016). The asa's statement on p-values : context, process, and purpose. *The American Statistician*, 70(2) :129–133.

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a world beyond "p< 0.05". *The American Statistician*, 73(sup1) :1–19.