ENSIMAG 2ème année

METHODES STATISTIQUES POUR L'INGENIEUR

Olivier Gaudoin

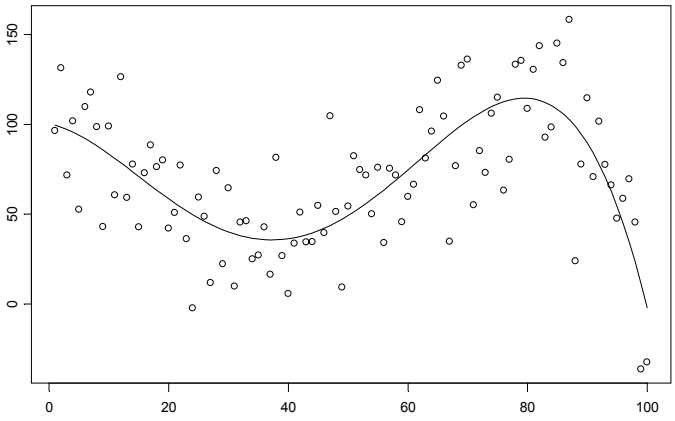


Table des matières

Chapitre 1 : Introduction

1.1. Utilité des méthodes statistiques pour l'ingénieur

La **statistique** est l'ensemble des méthodes et techniques utilisées dans le but d'extraire de l'information de **données** . Ces données peuvent être issues :

- de l'observation de phénomènes naturels (météorologie,…)
- de résultats d'expériences scientifiques (médecine, chimie,…)
- d'enquêtes socio-économiques
- etc…

Dans la plupart des cas, les données sont entachées d'incertitudes et présentent des variations pour plusieurs raisons :

- le résultat des expériences effectuées n'est pas prévisible à l'avance avec certitude
- toute mesure est entachée d'erreur
- une enquête est faite sur quelques individus et on doit extrapoler les conclusions de l'étude à toute une population
- etc…

Il y a donc intervention du hasard et des probabilités. L'objectif essentiel de la statistique est de maîtriser au mieux cette incertitude pour extraire des informations utiles des données, via l'analyse des variations dans les observations.

Les méthodes statistiques se répartissent en deux classes :

- la **statistique descriptive** (ou **statistique exploratoire** ou **analyse des données**) a pour but de **résumer l'information** contenue dans les données de façon synthétique et efficace. Elle utilise pour cela des **représentations de données** sous forme de graphiques, de tableaux et d'indicateurs numériques. Elle permet de dégager les caractéristiques essentielles du phénomène étudié et de suggérer des hypothèses pour une étude ultérieure plus sophistiquée. Les probabilités n'ont ici qu'un rôle mineur.

- la **statistique inférentielle** a pour but de **faire des prévisions** et de **prendre des décisions** au vu des observations. En général, il faut pour cela proposer des **modèles probabilistes** du comportement du phénomène aléatoire étudié et savoir gérer les risques d'erreurs. Les probabilités jouent ici un rôle fondamental.

Les méthodes statistiques sont utilisées dans de très nombreux domaines. Citons quelques exemples :

- ingénierie : contrôle de qualité des procédés de fabrication, sûreté de fonctionnement (fiabilité, sécurité,…)
- médecine : expérimentation de nouveaux traitements ou médicaments
- économie : prévisions économétriques, études quantitatives de marchés
- prévisions de tous ordres : météorologiques, démographiques, sociologiques,…
- politique : sondages d'opinion

- biologie : évolution des espèces, caractérisation des populations naturelles
- physique : théorie cinétique des gaz, mouvements des particules
- agriculture : rendement des cultures, expérimentation de nouvelles espèces,…
- etc…

On s'intéressera ici particulièrement aux applications de la statistique à l'informatique :

- qualité et sûreté de fonctionnement des systèmes informatiques
- évaluation des performances des systèmes informatiques
- évaluation et prévision du trafic sur les réseaux
- débruitage d'images
- etc…

D'autre part, l'informatique est souvent définie comme la science et la technique du traitement des données. L'analogie avec la définition de la statistique est frappante.

Enfin, tout ingénieur est amené à prendre des décisions au vu de certaines informations, dans des contextes où de nombreuses incertitudes demeurent. Il importe donc qu'un ingénieur soit formé aux techniques de gestion du hasard et de traitement de données expérimentales.

1.2. Statistique et probabilités

La statistique et les probabilités sont les deux aspects complémentaires de l'étude des phénomènes aléatoires. Ils sont cependant de natures bien différentes.

Les **probabilités** peuvent être envisagées comme une branche des mathématiques pures, basée sur la théorie de la mesure, abstraite et complètement déconnectée de la réalité.
Les **probabilités appliquées** proposent des **modèles probabilistes** du comportement de phénomènes aléatoires concrets. On peut alors, **préalablement à toute expérience**, faire des prévisions sur ce qui va se produire.

Par exemple, il est usuel de modéliser la durée de bon fonctionnement d'un système par une variable aléatoire X de loi exponentielle de paramètre λ. Ayant adopté ce modèle, on dira que la probabilité que le système ne soit pas encore tombé en panne à la date t est $P(X > t) = e^{-\lambda t}$. On prévoira aussi que si n systèmes identiques et indépendants sont mis en route en même temps, en moyenne $n(1-e^{-\lambda t})$ d'entre eux seront tombés en panne à la date t (car le nombre d'appareils en panne entre 0 et t est alors une variable aléatoire de loi binomiale $B(n, 1-e^{-\lambda t})$, d'espérance $n(1-e^{-\lambda t})$).

Dans la pratique, l'utilisateur d'un tel système est très intéressé par ces résultats. Il souhaite évidemment avoir une évaluation de la durée de bon fonctionnement de ce système, de la probabilité qu'il fonctionne correctement pendant plus d'un mois, un an, etc… Mais si l'on veut utiliser les résultats théoriques énoncés plus haut, il faut d'une part pouvoir s'assurer que la durée de vie de ce système est bien une variable aléatoire de loi exponentielle, et, d'autre part, pouvoir calculer d'une manière ou d'une autre la valeur du paramètre λ. C'est la statistique qui va permettre de résoudre ces problèmes.

Exemple : Dans le but d'étudier la densité du trafic sur internet, on a mesuré les durées de transfert, en millisecondes, d'un même message entre deux sites, à 10 moments différents d'une même journée :

91.6 35.7 251.3 24.3 5.4 67.3 170.9 9.5 118.4 57.1

On souhaite connaître la durée moyenne de transfert, la probabilité qu'un transfert se fasse en moins de 10 ms ou en plus de 200 ms, etc…

Notons $x_1,...,x_n$ $(n=10)$ ces observations. A cause des variations de densité du trafic sur internet, la durée de transfert d'un message n'est pas prévisible avec certitude à l'avance. On va donc considérer que $x_1,...,x_n$ sont les réalisations de variables aléatoires $X_1,...,X_n$. Puisque le message est toujours le même, il est naturel de supposer que les X_i sont de même loi. Si les transferts se sont faits indépendamment les uns des autres, on pourra supposer que les X_i sont des variables aléatoires indépendantes. On peut alors se poser les questions suivantes :

- Au vu de ces observations, est-il raisonnable de supposer que la durée de transfert d'un message est une variable aléatoire de loi exponentielle ?
- Si non, quelle autre loi serait plus appropriée ?
- Comment proposer une valeur (ou un ensemble de valeurs) vraisemblable pour les paramètres de cette loi ?
- Que peut-on garantir aux usagers d'internet sur la durée de transfert des messages ? Sur un paquet de 100 messages, combien seront transférés en moins de 50 ms ?

Notons que, pour répondre à ces questions, on doit prendre des **décisions** : décider si la loi est exponentielle, décider si la valeur du paramètre est dans tel intervalle, décider qu'un objectif de densité de trafic est bien atteint. A chaque fois, il est possible que l'on se trompe en prenant ces décisions. Donc, à toute réponse statistique, il faudra associer le **degré de confiance** que l'on peut accorder à cette réponse.

Pour résumer, la démarche probabiliste suppose que la nature du hasard est connue. Cela signifie que l'on adopte un modèle probabiliste particulier (ici la loi exponentielle), qui permettra d'effectuer des prévisions sur les observations futures. Dans la pratique, la nature du hasard est inconnue. La statistique va, au vu des observations, formuler des hypothèses sur la nature du phénomène aléatoire étudié. Maîtriser au mieux cette incertitude permettra de traiter les données disponibles. Probabilités et statistiques agissent donc en aller-retour dans le traitement mathématique des phénomènes aléatoires.

1.3. Plan du cours

Ce cours a pour but de présenter les principales méthodes statistiques utilisées par les ingénieurs. Ces méthodes seront toujours illustrées par des problèmes concrets, issus de l'informatique, la médecine, le contrôle de qualité, etc… Il ne s'agit pas de donner un catalogue de recettes. Les méthodes statistiques seront la plupart du temps justifiées mathématiquement, ce qui permettra d'éviter un certain nombre d'erreurs d'interprétation des résultats, fréquentes dans la pratique.

Toutes les méthodes décrites ici peuvent être mises en œuvre à l'aide du logiciel S+, qu'elles soient déjà préprogrammées ou pas. En général, on associera à chaque méthode la syntaxe et les sorties (tableaux, graphiques) correspondantes de S+.

Le chapitre 2 présente les techniques de base en statistique descriptive, représentations graphiques et indicateurs statistiques. Le chapitre 3 est consacré aux problèmes d'estimation, ponctuelle et par intervalles de confiance. Le chapitre 4 traite des tests d'hypothèses, tests paramétriques et non paramétriques, sur un ou deux échantillons. Le dernier chapitre est consacré à une des méthodes statistiques les plus utilisées, la régression linéaire. Enfin, des annexes donnent quelques rappels de probabilités utiles en statistique, ainsi que des tables des lois de probabilité usuelles.

Chapitre 2 : Statistique descriptive

La **statistique descriptive** a pour but de **résumer l'information** contenue dans les données de façon à en dégager les caractéristiques essentielles sous une forme simple et intelligible. Les deux principaux outils de la statistique descriptive sont les **représentations graphiques** et les **indicateurs statistiques**.

2.1. Population, individus et variables

Les **données** dont nous disposons sont des mesures faites sur des **individus** (ou unités statistiques) issus d'une **population**. On s'intéresse à une ou plusieurs particularités des individus appelées **variables** ou **caractères**. L'ensemble des individus constitue l'**échantillon** étudié.

Exemple : si l'échantillon est un groupe de TD à l'ENSIMAG,

- un individu est un étudiant
- la population peut être l'ensemble des étudiants de l'ENSIMAG, des écoles d'ingénieur, des habitants de Grenoble, etc…
- la variable étudiée peut être la taille, la filière choisie, la moyenne d'année, la couleur des yeux,…

Si l'échantillon est constitué de tous les individus de la population, on dit que l'on fait un **recensement**. Il est extrêmement rare que l'on se trouve dans cette situation, essentiellement pour des raisons de coût. Quand l'échantillon n'est qu'une partie de la population, on parle de **sondage**. Le principe des sondages est d'étendre à l'ensemble de la population les enseignements tirés de l'étude de l'échantillon. Pour que cela ait un sens, il faut que l'échantillon soit représentatif de la population. Il existe des méthodes pour y parvenir, dont nous ne parlerons pas ici.

Remarque : le mot « variable » désigne à la fois la grandeur que l'on veut étudier (variable statistique) et l'objet mathématique qui la représente (variable aléatoire).

Une variable statistique peut être **discrète** ou **continue**, **qualitative** ou **quantitative**. Les méthodes de représentation des données diffèrent suivant la nature des variables étudiées.

Dans ce chapitre, on ne s'intéresse qu'au cas où on ne mesure qu'une seule variable sur les individus. On dit alors que l'on fait de la **statistique unidimensionnelle**. Dans ce cas, les données sont sous la forme de la série des valeurs prises par la variable pour les n individus, notées $x_1,...,x_n$. On supposera que ces données sont n réalisations indépendantes de la même variable aléatoire X [1], ou, ce qui revient au même, les réalisations de n variables aléatoires $X_1,...,X_n$ indépendantes et de même loi (c'est la même distinction qu'entre la durée de transfert d'un message en général et la durée de transfert du $i^{\text{ème}}$ message). Le terme d'**échantillon** désignera à la fois les séries $x_1,...,x_n$ et $X_1,...,X_n$.

[1] En toute rigueur, il faudrait dire que les données proviennent de la même loi de probabilité et que X est une notation pour une variable aléatoire de cette loi.

Quand on mesure plusieurs variables sur les mêmes individus, on dit que l'on fait de la statistique multidimensionnelle. Des données de ce type seront traitées dans le chapitre consacré aux modèles linéaires.

2.2. Représentations graphiques

2.2.1. Variables discrètes

Une variable discrète est une variable à valeurs dans un ensemble fini ou dénombrable. Mais l'ensemble des valeurs prises par cette variable dans un échantillon de taille n est forcément fini. Les variables qui s'expriment par des nombres réels sont appelées **variables quantitatives** ou numériques (ex : longueur, durée,…). Les variables qui s'expriment par l'appartenance à une catégorie sont appelées **variables qualitatives** (ex : couleur, catégorie socio-professionnelle, …).

2.2.1.1. Variables qualitatives

Si la variable est qualitative, on appelle **modalités** les valeurs possibles de cette variable. L'ensemble des modalités est noté $E = \{e_1,...,e_m\}$.

Par exemple, si la variable est la couleur des yeux d'un individu, l'ensemble des modalités est E = {vert, bleu, brun, gris, noir }. Si on interroge n = 200 personnes, les données brutes se présenteront sous la forme d'une suite du type : brun, vert, vert, bleu, ..., gris, vert. Cette suite n'est pas lisible. La meilleure manière de représenter ces données est d'utiliser les fréquences absolues et relatives.

> *Définition : On appelle **fréquence absolue** de la modalité e_i le nombre total n_i d'individus de l'échantillon pour lesquels la variable a pris la modalité e_i : $n_i = \sum_{j=1}^{n} 1_{\{e_i\}}(x_j)$.*
>
> *On appelle **fréquence relative** de la modalité e_i le pourcentage n_i / n d'individus de l'échantillon pour lesquels la variable a pris la modalité e_i.*

Dans l'exemple, on obtient un tableau de ce type :

couleur des yeux	vert	bleu	brun	gris	noir
fréquences absolues	66	34	80	15	5
fréquences relatives	33%	17%	40%	7.5%	2.5%

Tableau 2.1. : couleur des yeux d'un échantillon de 200 personnes

De même, dans le cas des résultats d'élection en France, les individus sont les n = 20 millions d'électeurs et la variable est la personne ou la liste pour laquelle l'individu a voté. La suite des 20 millions de votes n'a aucun intérêt. Le résultat est exprimé directement sous forme du tableau des fréquences relatives. Par exemple, le tableau 2.2. donne le résultat du premier tour des élections législatives de mai 1997 :

Listes	Blancs + nuls	Ext. Gauche	PC	PS + DvG	Verts	DvD	UDF	RPR	FN
% Voix	2.2	2.2	9.9	26.5	6.2	6.6	14.7	16.8	14.9

Tableau 2.2. : résultat du premier tour des élections législatives de mai 1997

Les représentations graphiques correspondantes sont de deux types :

- **diagrammes en colonnes** ou **en bâtons** : à chaque modalité correspond un rectangle vertical dont la hauteur est proportionnelle à la fréquence relative de cette modalité

- **diagrammes sectoriels** ou **camemberts** : à chaque modalité correspond un secteur de disque dont l'aire (ou l'angle au centre) est proportionnelle à la fréquence relative de cette modalité

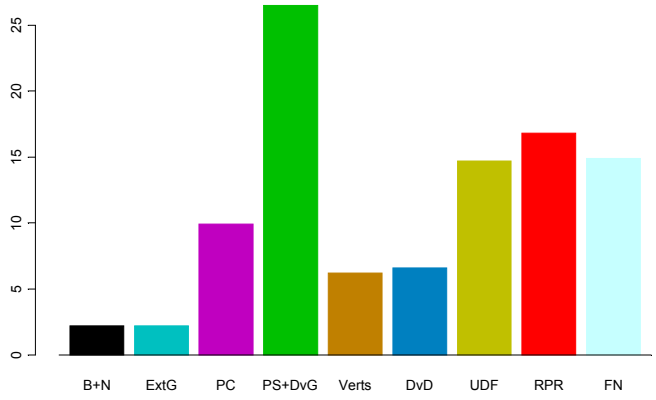


Figure 2.1. : élections législatives, diagramme en colonnes

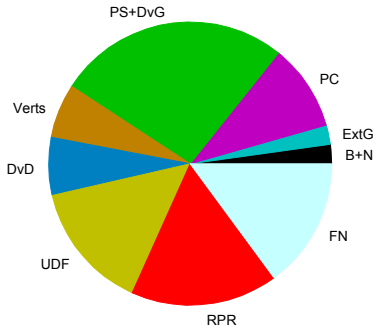


Figure 2.2. : élections législatives, diagramme sectoriel

Les commandes S+ pour les diagrammes en colonnes et sectoriels sont `barplot(x)` et `pie(x)`. Dans l'exemple des élections, les figures 2.1 et 2.2. sont obtenues à l'aide des commandes :

```
> x<-c(2.2,2.2,9.9,26.5,6.2,6.6,14.7,16.8,14.9)
> partis<-c("B+N","ExtG","PC","PS+DvG","Verts","DvD","UDF","RPR","FN")
> barplot(x,names=partis,col=1:10)
> pie(x,names=partis,rotate=F,inner=1.5)
```

2.2.1.2. Variables quantitatives

Quand la variable est quantitative**,** on utilise les mêmes représentations à l'aide des fréquences absolues et relatives. La différence fondamentale entre les représentations pour des variables qualitatives et quantitatives tient au fait qu'il existe un ordre naturel sur les modalités (qui sont des nombres réels) pour les variables quantitatives, alors qu'aucun ordre n'est prédéfini pour les variables qualitatives. C'est pourquoi les diagrammes en bâtons sont toujours utilisés (avec une seule couleur pour les bâtons), mais pas les diagrammes sectoriels.

Par exemple, on a effectué une enquête auprès de 1000 couples en leur demandant notamment leur nombre d'enfants. Le tableau des fréquences et le diagramme en bâtons sont représentés ci-dessous.

Nombre d'enfants	0	1	2	3	4	5	6	> 6
fréquence absolue	235	183	285	139	88	67	3	0
fréquence relative	23.5%	18.3%	28.5%	13.9%	8.8%	6.7%	3%	0

Tableau 2.3. : nombre d'enfants de 1000 couples

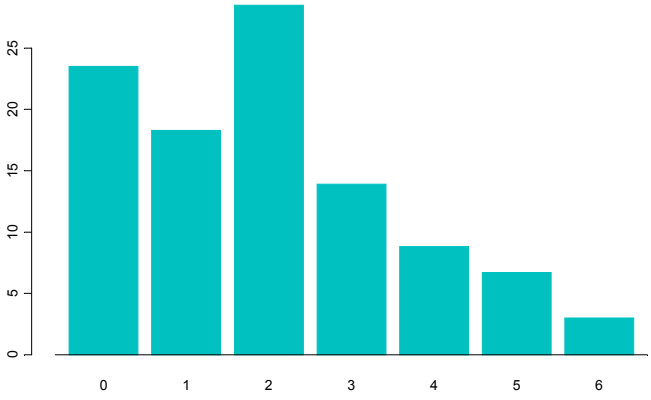


Figure 2.3. : nombre d'enfants de 1000 couples, diagramme en bâtons

2.2.1.3. Choix d'un modèle probabiliste discret

Les représentations graphiques effectuées permettent de guider le statisticien dans le choix d'un modèle probabiliste adapté aux données. En effet, la fréquence relative n_i / n, **pourcentage** d'observation de la modalité e_i dans l'échantillon, est une estimation naturelle de la **probabilité** que

la variable prenne la modalité e_i, $P(X = e_i)$. Une loi de probabilité vraisemblable pour X est une loi telle que le diagramme des $P(X = e_i)$ soit proche, en un certain sens, du diagramme en bâtons.

Par exemple, pour le nombre d'enfants par famille, une loi géométrique est impossible car une variable aléatoire de loi géométrique ne peut pas prendre la valeur 0. Une loi binomiale est envisageable, par exemple la loi $B(6, p)$ ou la loi $B(7, p)$. Le problème est de savoir s'il existe un paramètre p dans [0,1] tel que le diagramme des $P(X = i)$ ait une allure proche de celle de la figure 2.3. Une loi de Poisson est aussi possible a priori.

Pour pouvoir choisir un modèle par cette méthode, il faudrait donc connaître au moins les formes des diagrammes des probabilités élémentaires des lois binomiale et de Poisson. Ce n'est pas simple du fait de la complexité des expressions de ces probabilités. De plus, la forme de ces diagrammes peut changer assez sensiblement suivant la valeur des paramètres. Il est donc difficile de proposer un modèle probabiliste vraisemblable au seul vu d'un diagramme en bâtons. On verra que c'est beaucoup plus facile quand la variable est continue.

Finalement, le diagramme en bâtons sert plus à visualiser l'allure générale de la distribution qu'à véritablement aider à choisir un modèle probabiliste pertinent.

2.2.2. Variables continues

Quand la variable étudiée est continue, les représentations du type diagramme en bâtons sont sans intérêt, car les données sont en général toutes distinctes, donc les fréquences absolues sont toutes égales à 1.

On considèrera ici deux types de représentations graphiques :

- l'**histogramme** et le **polygone des fréquences** qui lui est associé
- la **fonction de répartition empirique**, qui permet notamment de construire des **graphes de probabilités**

Ces deux types de représentations nécessitent d'**ordonner** les données. Si l'échantillon initial est noté $x_1,...,x_n$, l'échantillon ordonné sera noté $x_1^*,...,x_n^*$.

Dans l'exemple du trafic sur internet, l'échantillon initial est :

91.6 35.7 251.3 24.3 5.4 67.3 170.9 9.5 118.4 57.1

et l'échantillon ordonné est :

5.4 9.5 24.3 35.7 57.1 67.3 91.6 118.4 170.9 251.3

On a donc, par exemple,

x_1 = 91.6 = durée de transfert du message n°1

$x_1^* = \min x_i$ = 5.4 = plus petite des durées de transfert des 10 messages

Sous S+, l'échantillon `x` est créé par la commande :

```
x<-c(91.6, 35.7, 251.3, 24.3, 5.4, 67.3, 170.9, 9.5, 118.4, 57.1)
```

La $i^{\text{ème}}$ observation est donnée par `x[i]`.

L'échantillon ordonné est obtenu par la commande `sort(x)`.

2.2.2.1. Histogramme et polygone des fréquences

Le principe de cette représentation est de regrouper les observations « proches ». Pour cela, on se fixe une borne inférieure de l'échantillon $a_o < x_1^*$ et une borne supérieure $a_k > x_n^*$. On partitionne l'intervalle $]a_o, a_k]$ en k intervalles $]a_{i-1}, a_i]$ appelés **classes**. La largeur de la classe i est $\Delta_i = a_i - a_{i-1}$. Les classes ne sont pas forcément toutes de même largeur.

On appelle **effectif** de la classe i le nombre d'observations appartenant à cette classe :

$$n_i = \sum_{j=1}^{n} 1_{]a_{i-1}, a_i]}(x_j).$$

La **fréquence** (ou fréquence relative) de la classe i est n_i / n.

L'**histogramme** est la figure constituée des rectangles dont les *bases* sont les classes et dont les *surfaces* sont égales aux fréquences de ces classes. Autrement dit, la *hauteur* du $i^{\text{ème}}$ rectangle est $\dfrac{n_i}{n\Delta_i}$.

Notons F la fonction de répartition de la variable aléatoire réelle observée X et f sa densité. La proportion n_i / n d'observations dans la classe i est une estimation naturelle de la probabilité qu'une observation appartienne à cette classe : $P(X \in]a_{i-1}, a_i]) = F(a_i) - F(a_{i-1})$. Or, la densité de X au point x est $f(x) = F'(x) = \lim\limits_{dx \to 0} \dfrac{1}{dx}\left[F(x + dx) - F(x)\right]$.

D'où, pour dx petit, $f(x)dx \approx F(x + dx) - F(x)$. En prenant $x = a_{i-1}$ et $dx = \Delta_i$, on obtient $f(a_{i-1}) \approx \dfrac{1}{\Delta_i}\left[F(a_i) - F(a_{i-1})\right] \approx \dfrac{n_i}{n\Delta_i}$.

Par conséquent, l'**histogramme fournit une approximation grossière de la densité des observations.** L'allure de l'histogramme permettra donc de proposer des modèles probabilistes vraisemblables pour la loi de X.

A priori, on a toute liberté pour le choix des classes : bornes inférieure et supérieure, nombre et largeur, ce qui fait que plusieurs histogrammes peuvent être dessinés à partir des mêmes données, et peuvent parfois avoir des allures assez différentes. Il est donc bon de suivre quelques règles :

- Il est recommandé d'avoir entre 5 et 12 classes, jamais moins de 5 ni plus de 20. La règle empirique de Sturges préconise un nombre de classes égal à $k \approx 1 + \log_2 n = 1 + \ln n / \ln 2$, ce qui donne par exemple $k = 5$ pour $n \leq 22$, $k = 6$ pour $23 \leq n \leq 45$, etc…

- Le choix des bornes a_o et a_k n'est pas normalisé. Il doit être fait de sorte que toutes les classes soient homogènes, en largeur ou en effectif. Un choix fréquent est $a_o = x_1^* - 0.025\,(x_n^* - x_1^*)$ et $a_k = x_n^* + 0.025\,(x_n^* - x_1^*)$.

- En ce qui concerne la largeur des classes, le choix le plus fréquent consiste à prendre des **classes de même largeur**. Dans ce cas, la hauteur des rectangles est proportionnelle à l'effectif des classes. Mais il est en général plus intéressant de choisir des **classes de même effectif**.

Dans l'exemple du trafic sur internet, $n = 10$ donc on choisit $k = 5$ classes. Il semble raisonnable de prendre comme bornes a_o = 0 et a_5 = 260. Dessinons dans un premier temps un histogramme à 5 classes de même largeur. Cette largeur est donc $\Delta = 260/5 = 52$. On obtient alors le tableau suivant :

classes $]a_{i-1}, a_i]$	]0, 52]	]52, 104]	]104, 156]	]156, 208]	]208, 260]
effectifs n_i	4	3	1	1	1
fréquences n_i / n	40%	30%	10%	10%	10%
hauteurs $n_i / n\Delta_i$	$7.7\ 10^{-3}$	$5.8\ 10^{-3}$	$1.9\ 10^{-3}$	$1.9\ 10^{-3}$	$1.9\ 10^{-3}$

Tableau 2.4. : trafic sur internet, répartition en classes de même largeur

L'histogramme correspondant est donné par la figure 2.4. Les commandes S+ permettant de construire cette figure sont :

```
> x<-c(91.6,35.7,251.3,24.3,5.4,67.3,170.9,9.5,118.4,57.1)
> abs<-c(0,26,78,130,182,234,275)
> ord<-c(0.0082,0.0077,0.0058,0.0019,0.0019,0.0019,0)
> hist(x,probability=T,breaks=seq(0,260,52),col=0,xlim=c(0,300),
  ylim=c(0,0.009))
> lines(abs,ord,lwd=5)
```

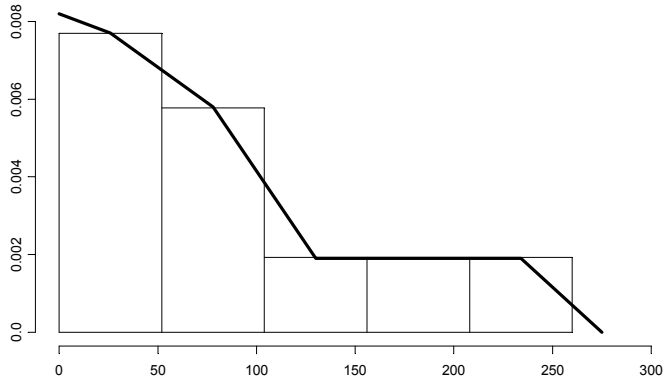


Figure 2.4. : trafic sur internet, histogramme à classes de même largeur et polygone des fréquences

Le **mode** est le milieu de la classe correspondant au rectangle le plus haut (estimation du maximum de la densité). Ici, le mode est 26.

L'histogramme fournit bien une visualisation de la répartition des données. Ici, le phénomène marquant est la concentration des observations sur les petites valeurs et le fait que, plus la durée de transfert grandit, moins il y a d'observations. Autrement dit, la densité de la variable aléatoire représentant la durée de transfert d'un message est une fonction décroissante.

L'histogramme n'est pas une approximation satisfaisante de la densité dans la mesure où c'est une fonction en escalier, alors que la densité est en général une fonction continue. Une meilleure approximation est le **polygone des fréquences**, c'est à dire la ligne brisée reliant les milieux des sommets des rectangles, et prolongée de part et d'autre des bornes de l'histogramme de sorte que l'aire sous le polygone soit égale à 1 (comme une densité). Le polygone des fréquences est représenté en gras dans la figure 2.4.

Avec l'histogramme, on estime qu'il y a 40% de chances que la durée de transfert d'un message soit inférieure à 52 ms, 10% qu'elle soit supérieure à 208 ms, etc… Avec le polygone des fréquences, on peut calculer des valeurs analogues en des points qui ne sont pas forcement des bornes de classes.

Le choix de classes de même largeur fait que certaines classes peuvent être très chargées et d'autres pratiquement vides. Pour connaître la répartition des observations dans les classes chargées, on a envie de scinder celles-ci. De même, on peut regrouper des classes trop peu chargées. A la limite, on peut faire en sorte que toutes les classes aient le même effectif. Dans ce cas, elles ne peuvent pas être de même largeur.

Dans l'exemple du trafic sur internet, on peut faire en sorte d'avoir 2 observations par classe. On détermine par exemple les limites des classes en prenant le milieu de deux observations ordonnées successives. On obtient alors le tableau et l'histogramme 2.5.

classes $]a_{i-1}, a_i]$	]0, 17]	]17, 46]	]46, 79]	]79, 145]	]145, 260]
largeur Δ_i	17	29	33	66	115
effectifs n_i	2	2	2	2	2
fréquences n_i / n	20%	20%	20%	20%	20%
hauteurs $n_i / n\Delta_i$	$11.8\ 10^{-3}$	$6.9\ 10^{-3}$	$6.1\ 10^{-3}$	$3.0\ 10^{-3}$	$1.7\ 10^{-3}$

Tableau 2.5. : trafic sur internet, répartition en classes de même effectif

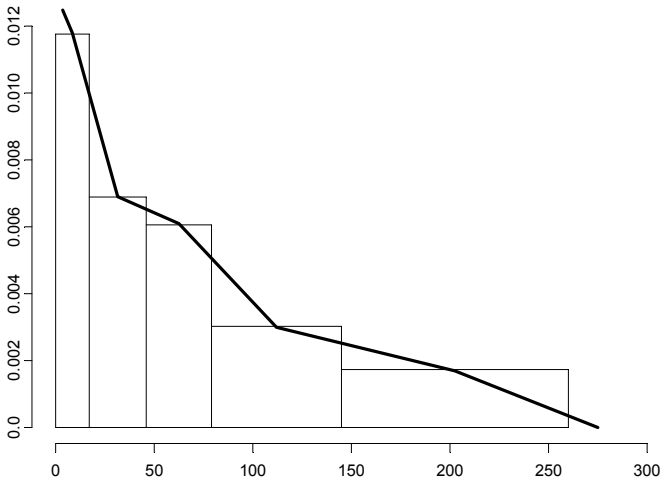


Figure 2.5. : trafic sur internet, histogramme à classes de même effectif et polygone des fréquences

On constate que cet histogramme décrit plus finement la distribution que le précédent. C'est toujours le cas des histogrammes à classes de même effectif. Mais leur usage est moins répandu que celui des histogrammes à classes de même largeur, car ils sont moins faciles à tracer.

On voit que des histogrammes distincts sur les mêmes données peuvent être sensiblement différents. Donc il faudra se méfier des histogrammes si on veut estimer la densité des observations. On se contentera de dire que l'histogramme et, mieux encore, le polygone des fréquences, donnent une allure générale de cette densité.

Par exemple ici, il est clair que la forme des deux histogrammes et polygones n'est pas très éloignée de la densité d'une loi exponentielle ($f(x) = \lambda e^{-\lambda x}$). En revanche, ils ne ressemblent pas du tout à la densité d'une loi normale (en forme de cloche). On en conclura qu'il est très peu probable que la durée de transfert d'un message soit de loi normale, et qu'il est possible, voire vraisemblable, qu'elle soit de loi exponentielle. Ce jugement est pour l'instant purement visuel. Il faudra l'affiner par des techniques quantitatives plus précises.

Remarque 1 : Si au lieu des effectifs n_i, on considère les effectifs cumulés $m_i = \sum_{j=1}^{i} n_j$, on construit un **histogramme** et un **polygone des fréquences cumulées**, qui fournissent une approximation de la fonction de répartition de la variable étudiée.

Remarque 2 : Il est fréquent qu'on ne dispose pas de l'intégralité des données brutes, mais de données déjà groupées. Par exemple, pour mesurer l'influence d'un certain type de grain sur la croissance des poulets, on a mesuré le poids de 1000 poulets nourris avec ce grain. Au lieu d'avoir le détail des 1000 poids, les données sont directement sous forme d'effectifs de classes, dans le tableau 2.6.

poids (en kg)	1.8-2.0	2.0-2.2	2.2-2.4	2.4-2.5	2.5-2.6	2.6-2.8	2.8-3.0	3.0-3.2
nombre de poulets n_i	64	86	140	232	168	160	90	60

Tableau 2.6. : poids de poulets, répartition en classes

L'histogramme peut alors se faire directement de la même manière que précédemment, en remarquant que les classes sont déterminées par les données et qu'elles ne sont pas toutes de même largeur.

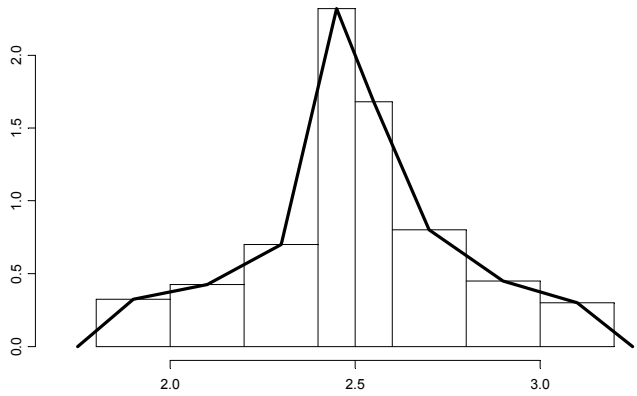


Figure 2.6. : poids de poulets, histogramme

La forme de cet histogramme nous amène à envisager l'hypothèse que le poids des poulets est distribué selon une loi normale.

2.2.2.2. Fonction de répartition empirique

On a vu que le polygone des fréquences cumulées était une approximation de la fonction de répartition des observations. La fonction de répartition empirique en est une autre, de meilleure qualité.

*Définition : La **fonction de répartition empirique** (FdRE) associée à un échantillon* $x_1,...,x_n$ *est la fonction définie par :*

$$F_n : R \to [0,1]$$

$$x \mapsto F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \le x\}} = \begin{cases} 0 & \text{si } x < x_1^* \\ i/n & \text{si } x_i^* \le x < x_{i+1}^* \\ 1 & \text{si } x \ge x_n^* \end{cases}$$

$F_n(x)$ *est le pourcentage d'observations de l'échantillon inférieures ou égales à x.*

La fonction de répartition empirique est une fonction en escalier qui fait des sauts de hauteur $1/n$ en chaque point de l'échantillon. Par exemple, la figure 2.7. représente la fonction de répartition empirique de l'échantillon des durées de transfert. Les commandes S+ permettant de tracer cette fonction pour cet exemple sont :

```
> x<-c(91.6,35.7,251.3,24.3,5.4,67.3,170.9,9.5,118.4,57.1)
> fdr<-seq(0.1,1,1/10)
> plot(sort(x),fdr,xlim=c(0,260),ylim=c(0,1.1),xlab="durees de transfert",
  ylab="")
> abs<-c(0,sort(x),260)
> ord<-c(0,fdr)
> for (i in 1:11) lines(c(abs[i],abs[i+1]),c(ord[i],ord[i]))
```

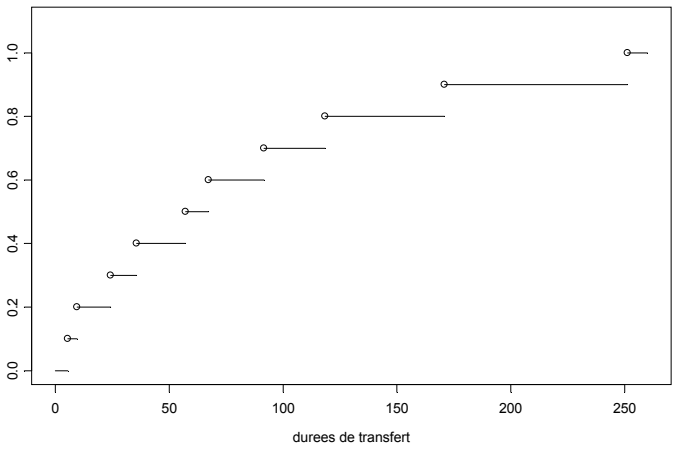


Figure 2.7. : trafic sur internet, fonction de répartition empirique

Il est clair que $F_n(x)$, pourcentage d'observations inférieures ou égales à x, est une estimation de la probabilité qu'une observation soit inférieure à x, c'est à dire $F(x)$. La qualité de cette estimation est donnée par le :

Théorème de Glivenko-Cantelli : $\sup_{x \in R} |F_n(x) - F(x)| \xrightarrow[n \to \infty]{} 0$ *p.s.*

Cela signifie que la fonction de répartition empirique converge uniformément et presque sûrement vers la vraie fonction de répartition de la variable étudiée. C'est donc une approximation d'excellente qualité de celle-ci.

2.2.2.3. Les graphes de probabilités

La fonction de répartition empirique est très utile en statistique. Intéressons-nous ici uniquement à son utilisation pour déterminer un modèle probabiliste acceptable pour les observations.

A priori, la première idée est de tracer le graphe de la fonction de répartition empirique et de déterminer si ce graphe ressemble à celui de la fonction de répartition d'une loi connue. En fait, il est très difficile de procéder ainsi car les fonctions de répartition de toutes les lois de probabilité se ressemblent : à vue d'œil, il n'y a pas de grande différence entre les fonctions de répartition des lois normale et exponentielle.

Une seconde idée est alors d'appliquer une transformation à la fonction de répartition empirique qui permette de reconnaître visuellement une caractéristique d'une loi de probabilité. Un **graphe de probabilités** est un nuage de points tracé à partir de la fonction de répartition empirique, tel que les points doivent être approximativement alignés si les observations proviennent d'une loi de probabilité bien précise.

Construisons les graphes de probabilités pour deux exemples simples, la loi exponentielle et la loi normale.

*** Graphe de probabilités pour la loi exponentielle**

La fonction de répartition de la loi exponentielle de paramètre λ est $F(x) = 1 - e^{-\lambda x}$. On a donc $\ln(1 - F(x)) = -\lambda x$.
On sait que $F_n(x)$ est une excellente approximation de $F(x)$. Donc, si les observations proviennent bien d'une loi exponentielle, on aura pour tout x, $\ln(1 - F_n(x)) \approx -\lambda x$. Par conséquent, si l'échantillon est issu d'une loi exponentielle, le graphe de la fonction $x \mapsto \ln(1 - F_n(x))$ doit être approximativement une droite de pente négative et passant par l'origine. On considère cette fonction aux points $x = x_i^*$, pour lesquels $F_n(x_i^*) = i/n$.

Le graphe de probabilités pour la loi exponentielle est le nuage des points $\left(x_i^*, \ln(1 - i/n)\right)$*, pour* $i = 1..n-1$ *(on ne prend pas en compte le cas* $i = n$ *car* $\ln(1 - n/n) = -\infty$*).*

Si les points de ce nuage sont approximativement alignés sur une droite de pente négative et passant par l'origine, on pourra considérer que la loi exponentielle est un modèle probabiliste vraisemblable pour ces observations. Inversement, si ce n'est pas le cas, il est probable que les observations ne sont pas issues d'une loi exponentielle.

La figure 2.8., construite à partir du tableau 2.7., présente le graphe de probabilités pour la loi exponentielle, pour l'exemple du trafic sur internet.

x_i^*	5.4	9.5	24.3	35.7	57.1	67.3	91.6	118.4	170.9
$\ln(1-i/n)$	- 0.105	- 0.223	- 0.357	- 0.511	- 0.693	- 0.916	- 1.204	- 1.609	- 2.303

Tableau 2.7. : trafic sur internet, tableau du graphe de probabilités pour la loi exponentielle

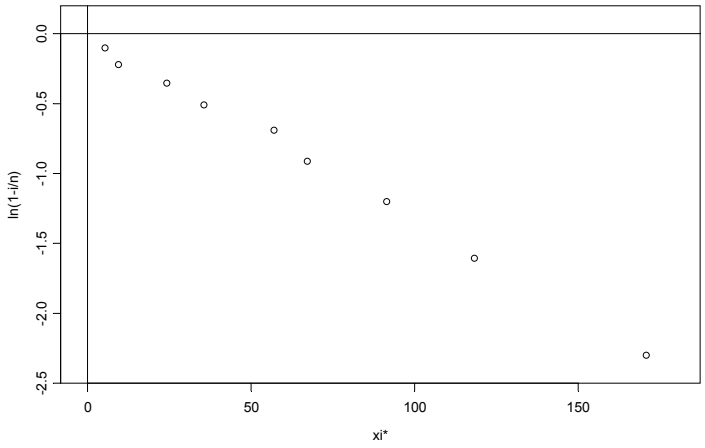


Figure 2.8. : trafic sur internet, graphe de probabilités pour la loi exponentielle

Les points semblent bien alignés sur une droite de pente négative et passant par l'origine. Il est donc vraisemblable que la durée de transfert d'un message soit une variable aléatoire de loi exponentielle. Cette conclusion est cohérente avec celle des histogrammes.

Remarque : la droite en question a pour équation $y = -\lambda x$. Sa pente fournit donc une estimation du paramètre λ. Pour déterminer cette pente, la méthode la plus usuelle est la méthode des moindres carrés, qui sera étudiée dans le chapitre consacré aux modèles linéaires. On obtient ici une pente de l'ordre de 0.013.

*** Graphe de probabilités pour la loi normale**

Si X est de loi normale $N(m,\sigma^2)$, alors $U = \dfrac{X-m}{\sigma}$ est de loi $N(0,1)$. Donc la fonction de répartition de la loi $N(m,\sigma^2)$ peut s'écrire $F(x) = P(X \le x) = P\left(U \le \dfrac{x-m}{\sigma}\right) = \phi\left(\dfrac{x-m}{\sigma}\right)$, où ϕ est la fonction de répartition de la loi normale centrée-réduite.

Etant donné que ϕ est strictement croissante, elle est inversible.

On a alors $\phi^{-1}\big(F(x)\big) = \dfrac{x-m}{\sigma} = \dfrac{1}{\sigma}x - \dfrac{m}{\sigma}$.

Par conséquent, si l'échantillon est issu d'une loi normale, le graphe de la fonction $x \mapsto \phi^{-1}(F_n(x))$ doit être approximativement une droite de pente positive et d'ordonnée à l'origine négative.

Le graphe de probabilités pour la loi normale est le nuage des points $\left(x_i^*, \phi^{-1}(i/n)\right)$*, pour* $i = 1..n-1$ *(on ne prend pas en compte le cas* $i = n$ *car* $\phi^{-1}(1) = +\infty$*).*

Les valeurs $\phi^{-1}(i/n)$ se calculent facilement à l'aide de S+ grâce à la commande `qnorm(x)`, ou sont à lire dans des tables de la loi normale (voir pages 110 et 111).

Si les points sont alignés sur une droite de pente positive et d'ordonnée à l'origine négative, on conclura que la loi normale est une loi vraisemblable pour les observations. La droite en question est alors appelée **droite de Henry**. Son équation permet d'obtenir des estimations de m et σ.
Sous S+, la commande `qqnorm(x)` donne le graphe de probabilités pour la loi normale, moyennant une permutation des abscisses et des ordonnées.

Pour l'exemple du trafic sur internet, on obtient le tableau 2.8 et la figure 2.9.

x_i^*	5.4	9.5	24.3	35.7	57.1	67.3	91.6	118.4	170.9
$\phi^{-1}(i/n)$	- 1.282	- 0.842	- 0.524	- 0.253	0	0.253	0.524	0.842	1.282

Tableau 2.8. : trafic sur internet, tableau du graphe de probabilités pour la loi normale

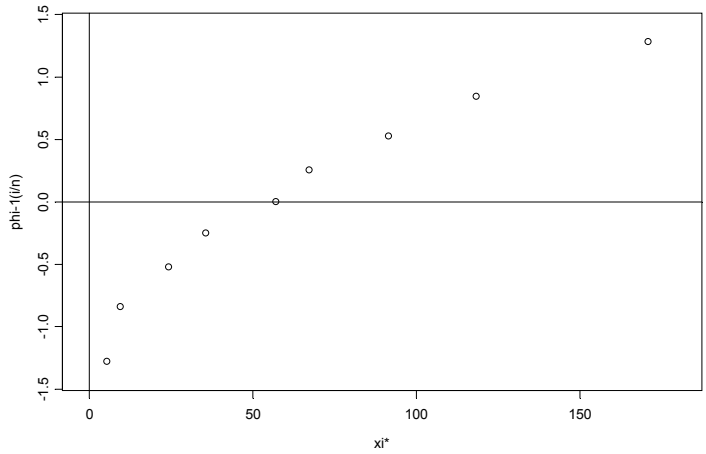


Figure 2.9. : trafic sur internet, graphe de probabilités pour la loi normale

Le graphe de probabilités semble plus proche d'un logarithme que d'une droite. On en conclura donc que la loi normale n'est pas un modèle approprié pour ces données.

On constate ici le principal défaut de la méthode : comment juger visuellement si des points sont plus ou moins alignés ? La réponse est soumise à la subjectivité de l'utilisateur. Il est donc nécessaire d'utiliser des techniques objectives que nous étudierons ultérieurement : les tests d'adéquation.

*** Principe général des graphes de probabilités**

Le principe général des graphes de probabilités est de chercher une transformation de la fonction de répartition de la loi à tester de la forme $h\big(F(x)\big)=\alpha(\theta)g(x)+\beta(\theta)$, où h et g sont des fonctions qui ne dépendent pas du paramètre θ de la loi.

Le graphe de probabilités est alors le nuage des points $\big(g(x_i^*), h(i/n)\big)$ dont on souhaite qu'ils soient alignés. A chaque fois, il s'agit de faire un changement d'échelle en abscisse et en ordonnée à partir du nuage $(x_i^*, i/n)$, qui n'est autre que le graphe de la fonction de répartition empirique.

Il existe des papiers spéciaux, dits papiers d'Alan Plait, pour lesquels ce changement d'échelle est déjà fait, et il ne reste plus qu'à représenter directement les points $(x_i^*, i/n)$. Par exemple, on parle de papier gausso-arithmétique pour la loi normale et de papier Weibull pour la loi de Weibull.

Remarque : Ce principe, appliqué ici à la fonction de répartition, peut s'appliquer aussi à d'autres caractéristiques des lois de probabilité, comme par exemple les probabilités élémentaires $P(X=x)$ pour les lois discrètes.

2.3. Indicateurs statistiques

Les représentations graphiques présentées dans la section précédente ne permettent qu'une analyse visuelle de la répartition des données. Pour des variables quantitatives, il est intéressant de donner des indicateurs numériques permettant de caractériser au mieux ces données. On donne en général deux indicateurs : un indicateur de localisation et un indicateur de dispersion.

2.3.1. Indicateurs de localisation ou de tendance centrale

Le but est de donner un ordre de grandeur général des observations, un nombre unique qui résume au mieux les données. On pense immédiatement à la moyenne des observations.

2.3.1.1. La moyenne empirique

La **moyenne empirique** de l'échantillon est la moyenne arithmétique des observations, notée $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. Son interprétation est évidente. La commande S+ correspondante est `mean(x)` .

Pour l'exemple du trafic sur internet, $\bar{x}_{10} = 83.15$, donc on dira que la durée moyenne de transfert d'un message est de 83.15 ms. Les représentations graphiques nous ont amenés à admettre que la durée de transfert d'un message était une variable aléatoire de loi exponentielle. On rappelle que l'espérance de la loi $\exp(\lambda)$ est $1/\lambda$. D'après la loi des grands nombres, la moyenne empirique converge presque sûrement vers l'espérance de la loi. Il est donc logique de considérer qu'une valeur vraisemblable de λ (ce qu'on appellera plus tard une estimation de λ) est $1/\bar{x}_{10} = 0.012$. Cette valeur est cohérente avec la valeur trouvée à l'aide du graphe de probabilités, 0.013. On retrouvera ce principe d'estimation plus tard, sous le nom de méthode des moments.

2.3.1.2. Les valeurs extrêmes

La plus petite valeur $x_1^* = \min x_i$ et la plus grande valeur $x_n^* = \max x_i$ d'un échantillon sont évidemment des indications intéressantes. Leur moyenne $\frac{1}{2}(x_1^* + x_n^*)$ est un indicateur de localisation.
Sous S+, on peut utiliser les commandes `min(x)` et `max(x)`.

Pour le trafic sur internet, $\frac{1}{2}(x_1^* + x_n^*) = 128.35$.

Problème : Les deux indicateurs que l'on vient de définir sont très sensibles aux valeurs extrêmes. En particulier, il arrive parfois qu'une série statistique présente des **valeurs aberrantes**, c'est à dire des valeurs exagérément grandes ou petites par rapport aux autres valeurs de l'échantillon. Par exemple, ce serait le cas si une durée de transfert était égale à 0.01 ou 10 000. En général, la présence d'une valeur aberrante est due à une erreur de saisie ou une erreur dans l'expérience ayant abouti à cette observation. Il faut alors l'éliminer avant d'effectuer l'analyse statistique. Il existe des méthodes de détection des valeurs aberrantes, mais il est souvent difficile de décider si une valeur est aberrante ou pas. Aussi est-il important de disposer d'indicateurs qui ne soient pas trop sensibles aux valeurs aberrantes. Or la moyenne est très sensible : si une des observations est extrêmement grande, elle va tirer la moyenne vers le haut. La médiane empirique est un indicateur de localisation construit pour être insensible aux valeurs aberrantes.

2.3.1.3. La médiane empirique

La **médiane empirique** de l'échantillon, notée $\tilde{x}_n$ ou $\tilde{x}_{1/2}$, est un réel qui partage l'échantillon ordonné en deux parties de même effectif. La moitié des observations sont inférieures à $\tilde{x}_n$ et l'autre moitié lui sont supérieures. Il y a donc une chance sur deux pour qu'une observation soit inférieure à la médiane, et évidemment une chance sur deux pour qu'une observation soit supérieure à la médiane.

Si n est impair, la médiane empirique est la valeur située au centre de l'échantillon ordonné : $\tilde{x}_n = x^*_{(n+1)/2}$.

Si n est pair, n'importe quel nombre compris entre $x^*_{n/2}$ et $x^*_{(n/2)+1}$ vérifie la définition de la médiane. Par convention, on prend en général le milieu de cet intervalle : $\tilde{x}_n = \frac{1}{2}(x^*_{n/2} + x^*_{(n/2)+1})$.

La commande S+ pour la médiane empirique est `median(x)`.

L'expression de la médiane montre bien que c'est un indicateur qui n'est pas sensible aux valeurs aberrantes. Pour l'illustrer, considérons les deux échantillons suivants :

1 3 5 8 10 et 1 3 5 8 10 000

La médiane empirique est égale à 5 pour les deux échantillons, alors que la moyenne empirique vaut 5.4 pour le premier échantillon et 2 003.4 pour le second. La moyenne est fortement influencée par la valeur aberrante 10 000 du deuxième échantillon, alors que la médiane ne l'est pas du tout.

Dans l'exemple du trafic sur internet, $\tilde{x}_{10} = \frac{1}{2}(57.1 + 67.3) = 62.2$.

On constate que la médiane est ici nettement inférieure à la moyenne : la durée moyenne de transfert est de 83.1 ms, et pourtant un message sur deux sera transféré en moins de 62.2 ms. Cette propriété est caractéristique des distributions non symétriques dites « à queues lourdes » : un petit nombre de messages auront une durée de transfert nettement supérieure à la majeure partie des autres. C'est ce qu'on avait déjà observé sur l'histogramme, et qui peut se remarquer directement sur les données.

Le même phénomène se produit si la variable étudiée est le salaire des français. En 1999, le salaire net mensuel moyen était de 10 930 F, alors que le salaire net mensuel médian était de 8 875 F. Un français sur deux touchait donc moins de 8 875 F par mois, mais un petit nombre de salariés gagnaient beaucoup d'argent, ce qui fait remonter la moyenne.

On voit donc que la connaissance simultanée de la moyenne et de la médiane peut être riche d'enseignements.

Quand la distribution est symétrique, moyenne et médiane empiriques sont proches (pour une variable aléatoire de loi symétrique, l'espérance et la médiane théoriques sont égales).

2.3.1.4. Caractérisation des indicateurs de localisation

Un indicateur de localisation c est fait pour résumer au mieux à lui seul l'ensemble des observations. L'erreur commise en résumant l'observation x_i par c peut être quantifiée par une distance $d(x_i, c)$.

L'erreur moyenne commise sur tout l'échantillon est $e = \frac{1}{n}\sum_{i=1}^{n} d(x_i, c)$. Un bon indicateur de localisation doit minimiser cette erreur globale.

- Si on choisit la distance euclidienne, $e = \frac{1}{n}\sum_{i=1}^{n} (x_i - c)^2$. La valeur de c qui minimise cette erreur est obtenue en annulant la dérivée de e par rapport à c. On obtient $c = \bar{x}_n$. La moyenne empirique est donc la valeur qui résume le mieux l'échantillon au sens dit « des moindres carrés ».

- Si on prend $e = \frac{1}{n}\sum_{i=1}^{n} |x_i - c|$, on trouve $c = \tilde{x}_n$.

- Si on prend $e = \frac{1}{n}\sup_{i=1}^{n} |x_i - c|$, on trouve $c = \frac{1}{2}(x_1^* + x_n^*)$.

Il est donc justifié d'utiliser ces 3 quantités comme indicateurs de localisation.

2.3.2. Indicateurs de dispersion ou de variabilité

Pour exprimer les caractéristiques d'un échantillon, il est nécessaire de compléter les indicateurs de localisation par des indicateurs de dispersion, qui mesureront la variabilité des données.

Par exemple, le tableau 2.9 donne les températures mensuelles moyennes, en degrés Celsius, à New-York et à San Francisco, calculées sur une période de 30 ans.

	J	F	M	A	M	J	J	A	S	O	N	D
New-York	0	1	5	12	17	22	25	24	20	14	8	2
San Francisco	9	11	12	13	14	16	17	17	18	16	13	9

Tableau 2.9. : températures mensuelles moyennes à New-York et à San Francisco

La température annuelle moyenne est de 12.5° à New-York et de 13.7° à San Francisco. En se basant uniquement sur ces moyennes, on pourrait croire que les climats de ces deux villes sont similaires. Or il est clair que la différence de température entre l'hiver et l'été est beaucoup plus forte à New-York qu'à San Francisco. Pour le déceler, il suffit de calculer un indicateur qui exprime la variabilité des observations.

Or, d'après la section 2.3.1.4., l'erreur moyenne commise en résumant l'échantillon par un indicateur de localisation c est $e = \frac{1}{n}\sum_{i=1}^{n} d(x_i, c)$. e exprime bien la variabilité de l'échantillon autour de c. On pourra donc construire des indicateurs de dispersion à partir de e en considérant différentes distances.

2.3.2.1. Variance et écart-type empiriques

Si on choisit la distance euclidienne, on a vu que $c = \bar{x}_n$. L'indicateur de dispersion correspondant est donc $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$. Il est appelé **variance empirique** de l'échantillon, et mesure l'écart quadratique moyen de l'échantillon à sa moyenne.

Il est facile de montrer que la variance empirique peut aussi s'écrire $s_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2$.

L'**écart-type empirique** de l'échantillon est $s_n = \sqrt{s_n^2}$. Il s'exprime dans la même unité que les données, ce qui rend son interprétation plus facile que celle de la variance. Ainsi, l'écart-type des températures annuelles est de 8.8° à New-York et de 3° à San Francisco, ce qui exprime bien la différence de variabilité des températures entre les deux villes.

Cependant, la variabilité doit toujours se comparer à la valeur moyenne. En effet, une variabilité de 10° n'a pas le même sens si la température moyenne de référence est 12° ou 10 000°. Des données présentent une forte variabilité si l'écart-type est fort par rapport à la moyenne.

On est donc amenés à définir le **coefficient de variation empirique** de l'échantillon, comme le rapport entre l'écart-type empirique et la moyenne empirique : $cv_n = \frac{s_n}{\bar{x}_n}$. On considère en général que l'échantillon possède une variabilité significative si $cv_n > 15\%$. Si $cv_n \leq 15\%$, les données présentent peu de variabilité et on considère que la moyenne empirique à elle seule est un bon résumé de tout l'échantillon.

Dans nos exemples, on obtient :

	$\bar{x}_n$	s_n^2	s_n	cv_n
durées de transfert	83.15	5540.2	74.4	89.5 %
t° New-York	12.5	77.7	8.8	70.4 %
t° San Francisco	13.7	8.9	3.0	21.8 %

On remarque donc une très forte variabilité des deux premiers échantillons et une variabilité assez faible du troisième.

Remarque : $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2$ évoque $Var(X) = E\left[(X - E(X))^2\right] = E(X^2) - [E(X)]^2$.

Les similitudes dans les noms et les formules suggèrent que la variance empirique est très liée à la variance de la loi de probabilité de la variable aléatoire sous-jacente. On reviendra sur ce point au chapitre suivant.

Sous S+, la commande `var(x)` donne $\frac{n}{n-1}s_n^2$ au lieu de s_n^2. On en verra l'explication au chapitre suivant. Il n'y a pas de commandes prédéfinies pour l'écart-type et le coefficient de variation empiriques.

2.3.2.2. L'étendue

L'étendue d'un échantillon est $e_n = x_n^* - x_1^*$. Cet indicateur est moins riche que la variance empirique et est évidemment très sensible aux valeurs aberrantes. Il est employé couramment en contrôle de qualité, notamment pour détecter ces valeurs aberrantes.

2.3.2.3. Les quantiles empiriques

Les **quantiles empiriques** sont des valeurs qui partagent l'échantillon ordonné en un certain nombre de parties de même effectif.

- s'il y a 2 parties, on retrouve la médiane empirique $\tilde{x}_n$
- s'il y a 4 parties, on parle de **quartiles**, notés $\tilde{q}_{n,1/4}$, $\tilde{q}_{n,1/2}$ et $\tilde{q}_{n,3/4}$; on a $\tilde{q}_{n,1/2} = \tilde{x}_n$
- s'il y a 10 parties, on parle de **déciles**, notés $\tilde{q}_{n,1/10}, \ldots, \tilde{q}_{n,9/10}$
- s'il y a 100 parties, on parle de **centiles**, notés $\tilde{q}_{n,1/100}, \ldots, \tilde{q}_{n,99/100}$
- etc…

<u>Définition</u> : Le quantile empirique d'ordre p de l'échantillon est défini par :

$$\tilde{q}_{n,p} = \begin{cases} \frac{1}{2}(x_{np}^* + x_{np+1}^*) & \text{si } np \text{ est entier} \\ x_{\lfloor np \rfloor + 1}^* & \text{sinon} \end{cases}$$

Dans l'exemple du trafic sur internet, on n'a que 10 données, donc seuls les quartiles ont un sens. On connaît déjà la médiane empirique $\tilde{q}_{n,1/2} = \tilde{x}_n = 62.2$. On obtient $\tilde{q}_{n,1/4} = x_3^* = 24.3$, et $\tilde{q}_{n,3/4} = x_8^* = 118.4$.

La **distance inter-quartiles** $\tilde{q}_{n,3/4} - \tilde{q}_{n,1/4}$ est un indicateur de dispersion. Son principal intérêt est d'être insensible aux valeurs aberrantes. Dans l'exemple, elle vaut 94.1 ms. On définit de la même manière des distances inter-déciles, inter-centiles, etc…

Les quantiles sont très utiles pour analyser des phénomènes concernant les extrémités des échantillons. Par exemple, une enquête de l'INSEE sur le patrimoine des familles en France en 1997 a obtenu entre autres les résultats suivants :

- le patrimoine moyen des familles était de 900 000 F
- 5% des familles avaient un patrimoine inférieur à 25 000 F
- 5% des familles avaient un patrimoine supérieur à 1 800 000 F ; ces 5% possèdaient 40% du patrimoine total.

Les chiffres fournis ici sont $\bar{x}_n$, $\tilde{q}_{n,5/100}$ et $\tilde{q}_{n,95/100}$.

Sous S+, la commande `quantile(x,p)` donne une version du quantile empirique d'ordre p légèrement différente de celle décrite ici. La commande `summary(x)` donne en une seule fois les minimum, premier quartile, médiane, moyenne, troisième quartile et maximum de l'échantillon.

2.3.3. Indicateurs statistiques pour des données groupées

Quand on ne dispose pas de la totalité des données brutes, mais de données déjà groupées en classes, le calcul exact des indicateurs statistiques est impossible. On peut en proposer une approximation en faisant comme si toutes les données appartenant à une classe étaient égales au centre de la classe.

Pour illustrer cette démarche, reprenons l'exemple des poids de poulets vu en section 2.3.2.1. En conservant les notations utilisées pour l'histogramme, on a k classes $]a_{i-1}, a_i]$. Les centres des classes sont les $c_i = \frac{1}{2}(a_{i-1} + a_i)$. L'effectif de la classe i est n_i.

poids $]a_{i-1}, a_i]$	]1.8, 2.0]	]2.0, 2.2]	]2.2, 2.4]	]2.4, 2.5]	]2.5, 2.6]	]2.6, 2.8]	]2.8, 3.0]	]3.0, 3.2]
centres des classes c_i	1.9	2.1	2.3	2.45	2.55	2.7	2.9	3.1
nombre de poulets n_i	64	86	140	232	168	160	90	60

Tableau 2.10. : poids de poulets, calcul des indicateurs statistiques

Une approximation de la moyenne empirique est $\bar{x}_{n,a} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i$.

Une approximation de la variance empirique est $s^2_{n,a} = \frac{1}{n}\sum_{i=1}^{k} n_i (c_i - \bar{x}_{n,a})^2$.

Ici, on obtient $\bar{x}_{n,a}$ = 2.498 et $s_{n,a}$ = 0.29.

Pour estimer la médiane empirique, on commence par déterminer la classe médiane, c'est à dire celle qui contient la médiane de l'échantillon. Ici, c'est la classe]2.4, 2.5] : 29% des données sont inférieures à 2.4 et 47.8% des données sont supérieures à 2.5.

Dans un premier temps, on peut approcher la médiane empirique par le centre de la classe médiane, ici 2.45. Mais on voit que ceci ne tient pas compte du déséquilibre éventuel entre les effectifs des classes inférieures et supérieures à la classe médiane. Dans l'exemple, il est logique de dire que la médiane est plus proche de 2.5 que de 2.4 car 47.8% est nettement supérieur à 29%. On peut alors procéder par interpolation linéaire.

On obtient : $\tilde{x}_{n,a} = a_{i-1} + \frac{1}{n_i}\left(\frac{n}{2} - \sum_{j=1}^{i-1} n_j\right)(a_i - a_{i-1})$, où i est le numéro de la classe médiane.

Dans l'exemple, $\tilde{x}_{n,a}$ = 2.49. Le fait que la moyenne et la médiane empiriques soient quasiment identiques confirme la symétrie de la distribution, déjà observée sur l'histogramme.

Chapitre 3. Estimation paramétrique

3.1. Introduction

Dans ce chapitre, on suppose que les données $x_1,...,x_n$ sont n réalisations indépendantes d'une même variable aléatoire X, appelée variable parente. Il est équivalent de supposer que $x_1,...,x_n$ sont les réalisations de variables aléatoires $X_1,...,X_n$ indépendantes et de même loi. Nous adopterons ici la seconde formulation, qui est plus pratique à manipuler.

Les techniques de statistique descriptive, comme l'histogramme ou le graphe de probabilités, permettent de faire des hypothèses sur la nature de la loi de probabilité des X_i. Des techniques statistiques plus sophistiquées, appelées tests d'adéquation, permettent de valider ou pas ces hypothèses.
On supposera ici que ces techniques ont permis d'adopter une famille de lois de probabilité bien précises (par exemple, loi normale, loi binomiale, etc ...) pour la loi des X_i, mais que la valeur du ou des paramètres de cette loi est inconnue.

On notera θ le paramètre inconnu. A priori, θ peut-être un paramètre à plusieurs dimensions, mais on supposera ici que θ est un réel. Pour $\theta \in R^p$, $p \geq 2$, toutes les notions de ce chapitre sont généralisables, mais la complexité des résultats augmente notablement.
On notera $F(x;\theta)$ la fonction de répartition des X_i. Pour les variables aléatoires discrètes on notera $P(X = x;\theta)$ les probabilités élémentaires, et pour les variables aléatoires continues on notera $f(x;\theta)$ la densité.

Le problème traité dans ce chapitre est celui de l'**estimation** du paramètre θ. Il s'agit de donner, au vu des observations $x_1,...,x_n$, une approximation de θ que l'on espère la plus proche possible de la vraie valeur inconnue. On pourra proposer une unique valeur vraisemblable pour θ (**estimation ponctuelle**) ou un ensemble de valeurs vraisemblables (**estimation ensembliste** ou **intervalle de confiance**).

3.2. Principes généraux de l'estimation

3.2.1. Définition et qualité d'un estimateur

*Définition : Une **statistique** s est une fonction des observations $x_1,...,x_n$.*

$$\begin{aligned} s: \quad R^n &\to R^m \\ (x_1 \ldots x_n) &\mapsto s(x_1 \ldots x_n) \end{aligned}$$

Par exemple, $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, $x_1^* = \min x_i$, $(x_1, x_3 + x_4, 2\ln x_6)$ sont des statistiques.

Remarque : Puisque $x_1,...,x_n$ sont des réalisations des variables aléatoires $X_1,...,X_n$, $s(x_1,...,x_n)$ est

une réalisation de la variable aléatoire $s(X_1,...,X_n)$. Par exemple, $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ est une réalisation de $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Pour simplifier les écritures, on note souvent $s_n = s(x_1,...,x_n)$ et $S_n = s(X_1,...,X_n)$. Par abus, on donne le même nom de statistique aux deux quantités.

> *Définition : Un* ***estimateur*** *d'une grandeur θ est une statistique S_n à valeurs dans l'ensemble des valeurs possibles de θ. Une* ***estimation*** *de θ est une réalisation s_n de l'estimateur S_n.*

A priori, n'importe quelle fonction des observations à valeurs dans l'ensemble des valeurs possibles de θ est un estimateur de θ. Mais un estimateur S_n de θ ne sera satisfaisant que si, pour n'importe quelle observation $x_1,...,x_n$, s_n est « proche », en un certain sens, de θ.

Pour cela, il faut d'abord que, si on répète plusieurs fois l'expérience, la moyenne des estimations obtenues soit très proche, et dans l'idéal égale à θ. Cela revient à souhaiter que l'espérance de l'estimateur soit égale à θ.

> *Définition : Un* ***estimateur*** S_n de θ *est* ***sans biais*** *si et seulement si $E(S_n) = \theta$. Il est* ***biaisé*** *si et seulement si $E(S_n) \neq \theta$.*

Ensuite, il est souhaitable que, plus on a d'observations, meilleure soit l'estimation. Cela signifie que l'estimateur S_n doit converger vers la valeur à estimer θ.
Il s'agit en fait d'étudier la convergence de la suite de variables aléatoires $\{S_n\}_{n\geq 1}$ vers la constante θ. Dans l'absolu, la convergence la plus forte est la convergence presque sûre. Dans la pratique, on se contente de la **convergence en moyenne quadratique** (ou convergence dans L^2) :

$$S_n \xrightarrow{MQ} \theta \Leftrightarrow \lim_{n\to\infty} E[(S_n - \theta)^2] = 0$$

$E[(S_n - \theta)^2]$ est appelée l'**erreur quadratique moyenne**. Elle mesure l'erreur que l'on fait si on estime θ par S_n, c'est à dire la précision de l'estimateur S_n. Elle doit donc être la plus petite possible.

> *Définition : Un* ***estimateur*** S_n *de θ est* ***convergent*** *si et seulement si S_n converge en moyenne quadratique vers θ quand n tend vers l'infini.*

On remarque que si S_n est sans biais, $E[(S_n - \theta)^2] = E[(S_n - E(S_n))^2] = Var(S_n)$. D'où :

- Un estimateur sans biais est convergent si et seulement si sa variance tend vers 0 quand n tend vers l'infini.
- De deux estimateurs sans biais, le meilleur est celui qui a la plus petite variance.

C'est logique : il faut non seulement que la moyenne des estimations soit proche de θ, mais aussi que *chaque* estimation soit la plus proche possible de θ, donc que la variabilité de l'estimateur S_n soit faible.

Finalement, on considèrera que le meilleur estimateur possible de θ est un **estimateur sans biais et de variance minimum (ESBVM)**. Un tel estimateur n'existe pas forcément.

Il existe des méthodes pour déterminer directement un ESBVM dans certains cas. Elles sont basées sur des techniques sophistiquées (exhaustivité, complétion, espérance conditionnelle), qui ne seront pas abordées dans ce cours. Cependant, on pourra parfois montrer facilement qu'un estimateur est un ESBVM en utilisant la quantité d'information de Fisher, définie dans la section suivante.

Remarque 1 : Un estimateur biaisé peut être intéressant si son erreur quadratique moyenne est inférieure à la variance d'un estimateur sans biais.

Remarque 2 : Ce n'est pas parce que S_n est un bon estimateur de θ que $\varphi(S_n)$ est un bon estimateur de $\varphi(\theta)$. Par exemple, on peut avoir $E(S_n)=\theta$ et $E[\varphi(S_n)]\neq\varphi(\theta)$.

3.2.2. Fonction de vraisemblance, efficacité d'un estimateur

Définition : Quand les observations sont toutes discrètes ou toutes continues, on appelle ***fonction de vraisemblance*** *de l'échantillon $x_1,...,x_n$ pour le paramètre θ la fonction :*

$$\mathcal{L}(\theta\,;x_1,...,x_n)=\begin{cases} P(X_1=x_1,...,X_n=x_n;\theta) & \text{si les } X_i \text{ sont discrètes} \\ f_{(X_1,...,X_n)}(x_1,...,x_n;\theta) & \text{si les } X_i \text{ sont continues} \end{cases}$$

Remarque : La probabilité et la densité utilisées dans cette définition sont des fonctions des observations $x_1,...,x_n$, dépendant du paramètre θ. A l'inverse, la fonction de vraisemblance est considérée comme une fonction de θ dépendant des observations $x_1,...,x_n$, ce qui permet, par exemple, de dériver cette fonction par rapport à θ.

Définition : On appelle ***quantité d'information de Fisher*** *sur θ apportée par l'échantillon $X_1,...,X_n$, la quantité (si elle existe) :*

$$I_n(\theta)=Var\left[\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta;X_1,...,X_n)\right]$$

Propriétés :

- On peut montrer que $E\left[\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta;X_1,...,X_n)\right]=0$. Par conséquent, la quantité d'information peut aussi s'écrire sous la forme $I_n(\theta)=E\left[\left(\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta;X_1,...,X_n)\right)^2\right]$.

- Si le domaine de définition des X_i ne dépend pas de θ, on montre que l'on a également $I_n(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln \mathcal{L}(\theta; X_1,..., X_n)\right]$. Cette écriture peut s'avérer pratique pour les calculs.

L'intérêt de la quantité d'information de Fisher est qu'elle fournit une borne inférieure pour la variance de n'importe quel estimateur de θ. Ce résultat s'exprime sous la forme du théorème suivant :

Inégalité de Fréchet-Darmois-Cramer-Rao (FDCR) : Si le domaine de définition des X_i ne dépend pas de θ, alors pour toute statistique S_n on a :

$$Var(S_n) \geq \frac{\left[\frac{\partial}{\partial\theta} E[S_n]\right]^2}{I_n(\theta)}$$

La quantité $\frac{\left[\frac{\partial}{\partial\theta} E[S_n]\right]^2}{I_n(\theta)}$ est appelée la **borne de Cramer-Rao**. L'inégalité FDCR dit donc que la variance d'un estimateur quelconque de θ est forcément supérieure à cette borne.

*Définition : on appelle **efficacité** d'un estimateur S_n la quantité $Eff(S_n) = \frac{\left[\frac{\partial}{\partial\theta} E[S_n]\right]^2}{I_n(\theta) Var(S_n)}$.*

On a $0 \leq Eff(S_n) \leq 1$.

*S_n est dit un estimateur **efficace** si $Eff(S_n) = 1$.*

*S_n est dit **asymptotiquement efficace** si $\lim_{n\to+\infty} Eff(S_n) = 1$.*

Propriétés :

- Si un estimateur est efficace, sa variance est égale à la borne de Cramer-Rao, donc il est forcément de variance minimum.

- Il est possible qu'il n'existe pas d'estimateur efficace de θ. Alors, s'il existe un ESBVM de θ, sa variance est strictement supérieure à la borne de Cramer-Rao.

- Si S_n est un estimateur sans biais de θ, alors $Var(S_n) \geq \frac{1}{I_n(\theta)}$ et $Eff(S_n) = \frac{1}{I_n(\theta) Var(S_n)}$.

- Si la valeur de la borne de Cramer-Rao est très grande, il est impossible d'estimer correctement θ car tous les estimateurs possibles auront une forte variance.

Remarque : La définition de la quantité d'information ci-dessus est une définition générale, applicable quelle que soit la nature des variables aléatoires observées. Quand celles-ci sont indépendantes et de même loi, il est facile de voir que $I_n(\theta) = nI_1(\theta)$.

Dans cette section, nous avons discuté des propriétés que devrait avoir un estimateur de θ, mais nous n'avons pas encore donné de méthodes pour trouver un estimateur de θ. C'est l'objet de la section suivante.

3.3. Méthodes d'estimation

Il existe de nombreuses méthodes pour estimer un paramètre θ. Par exemple, nous avons déjà vu des estimations graphiques à partir des graphes de probabilité. Nous avons aussi utilisé le principe qu'une probabilité peut s'estimer par une proportion.

Dans cette section, nous ne nous intéressons qu'aux deux méthodes d'estimation les plus usuelles, la méthode des moments et la méthode du maximum de vraisemblance.

3.3.1. La méthode des moments

C'est la méthode la plus naturelle, que nous avons déjà utilisée sans la formaliser.

3.3.1.1. Estimation d'une espérance

Le principe de la méthode des moments est que, si le paramètre à estimer est l'espérance de la loi des X_i, alors on peut l'estimer par la moyenne empirique de l'échantillon. Autrement dit, si $\theta = E(X)$, alors l'**estimateur** de θ **par la méthode des moments** (EMM) est $\tilde{\theta}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

La justification de ce principe est la loi des grands nombres, qui dit que $\overline{X}_n$ converge presque sûrement vers $E(X)$. Donc, si $\theta = E(X)$, $\overline{X}_n$ est un estimateur de θ convergent presque sûrement.

On peut en fait montrer facilement que $\overline{X}_n$ est un bon estimateur de $\theta = E(X)$, sans utiliser la loi des grands nombres.

- $E(\overline{X}_n) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\theta = \theta$, donc $\overline{X}_n$ est un estimateur sans biais de θ.

- $Var(\overline{X}_n) = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right]$

  $= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$ car les X_i sont indépendantes

  $= \frac{1}{n^2} n Var(X) = \frac{1}{n} Var(X)$, qui tend bien vers 0 quand n tend vers l'infini.

Donc $\overline{X}_n$ est un estimateur sans biais et convergent de $E(X)$.

Plus généralement, si $E(X)=\varphi(\theta)$, où φ est une fonction inversible, alors l'estimateur de θ par la méthode des moments est $\tilde{\theta}_n=\varphi^{-1}(\overline{X}_n)$.

3.3.1.2. Estimation d'une variance

De la même manière, on a envie d'estimer la variance de la loi des X_i par la variance empirique de l'échantillon $S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}_n^2$.

Déterminons le biais de cet estimateur.

$$E(S_n^2)=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}_n^2\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i^2)-E(\overline{X}_n^2)=E(X^2)-E(\overline{X}_n^2)$$
$$=Var(X)+[E(X)]^2-\left[Var(\overline{X}_n)+[E(\overline{X}_n)]^2\right]=Var(X)-\frac{1}{n}Var(X)=\frac{n-1}{n}Var(X)$$

On a $E(S_n^2)\neq Var(X)$, donc, contrairement à ce qu'on pourrait croire*, la variance empirique n'est pas un estimateur sans biais de la variance des observations.* Cet estimateur n'est qu'asymptotiquement sans biais.

En revanche, on voit que $E\left[\frac{n}{n-1}S_n^2\right]=Var(X)$. Soit donc $S_n'^2=\frac{n}{n-1}S_n^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2$.

$S_n'^2$ est appelée **variance estimée** de l'échantillon. C'est un estimateur sans biais de $Var(X)$.

On montre que $Var(S_n'^2)=\frac{n-1}{n^3}\left[(n-1)E\left[(X-E(X))^4\right]-(n-3)Var(X)^2\right]$, qui tend bien vers 0 quand n tend vers l'infini.

Donc $S_n'^2$ est un estimateur sans biais et convergent de $Var(X)$.

C'est pour cela que la commande `var(x)` sous S+ donne la variance estimée, et non pas la variance empirique de l'échantillon `x`.

On peut montrer également que $S_n'^2$ et S_n^2 convergent toutes les deux presque sûrement vers $Var(X)$.

Remarque 1 : On n'a pas de résultat général sur la qualité de S_n comme estimateur de l'écart-type de la loi, $\sigma(X)=\sqrt{Var(X)}$.

Remarque 2 : $Cov(\overline{X}_n,S_n^2)=\frac{n-1}{n^2}E\left[(X-E(X))^3\right]$, donc $\overline{X}_n$ et S_n^2 sont asymptotiquement non corrélées. La moyenne et la variance empirique ne sont indépendantes que si les observations sont de loi normale.

Plus généralement, si la loi des X_i a deux paramètres θ_1 et θ_2 tels que $(E(X),Var(X))=\varphi(\theta_1,\theta_2)$, où φ est une fonction inversible, alors les estimateurs de θ_1 et θ_2 par la méthode des moments sont $(\widetilde{\theta}_{1n},\widetilde{\theta}_{2n})=\varphi^{-1}(\overline{X}_n,S_n^2)$. Ce principe peut naturellement se généraliser aux moments de tous ordres, centrés ou non centrés : $E\left[(X-E(X))^k\right]$ et $E(X^k)$, $k\geq 1$.
Le simple exemple de la variance montre qu'un estimateur des moments n'est pas forcément sans biais. On peut montrer qu'il est asymptotiquement sans biais et convergent presque sûrement.

3.3.1.3. Exemples

Exemple 1 : loi normale

Si $X_1,...,X_n$ sont indépendantes et de même loi normale $N(m,\sigma^2)$, les estimateurs de m et σ^2 par la méthode des moments sont évidemment $\widetilde{m}_n=\overline{X}_n$ et $\widetilde{\sigma}_n^2=S_n^2$, et on sait qu'il vaut mieux estimer σ^2 par $S_n'^2$. Il est facile de montrer que $\overline{X}_n$ est un ESBVM de m. $S_n'^2$ est également un ESBVM de σ^2, mais la démonstration est moins immédiate.

Exemple 2 : loi exponentielle

Si $X_1,...,X_n$ sont indépendantes et de même loi exponentielle $\exp(\lambda)$, on sait que $E(X)=1/\lambda$. Donc l'estimateur de λ par la méthode des moments est $\widetilde{\lambda}_n=1/\overline{X}_n$.

> *Exercice* : montrer que $\widetilde{\lambda}_n$ est biaisé, trouver un estimateur $\widetilde{\lambda}_n'$ sans biais, montrer qu'il est convergent, asymptotiquement efficace, mais pas efficace.

En fait, on peut montrer qu'il n'existe pas d'estimateur efficace de λ et que $\widetilde{\lambda}_n'$ est l'ESBVM de λ.

Dans l'exemple du trafic sur internet, on obtient $\widetilde{\lambda}_n$ = 0.012 et $\widetilde{\lambda}_n'$ = 0.0108. Rappelons que l'estimation graphique obtenue à l'aide des graphes de probabilité était 0.013. Ces résultats sont bien cohérents.

Remarque : L'usage veut que la même notation $\widetilde{\theta}_n$ désigne à la fois l'estimateur de θ (variable aléatoire) et l'estimation correspondante (réalisation de cette variable aléatoire sur l'expérience considérée). Par exemple, dans le cas de la loi exponentielle, $\widetilde{\lambda}_n$ désigne aussi bien $1/\overline{X}_n$ que $1/\overline{x}_n$. Il faudra prendre garde à ne pas confondre les deux notions.

3.3.2. La méthode du maximum de vraisemblance

3.3.2.1. Définition

Principe : Si les X_i sont des variables aléatoires discrètes, la fonction de vraisemblance de l'échantillon est $\mathcal{L}(\theta\,;x_1,...,x_n)=P(X_1=x_1,...,X_n=x_n;\theta)$. C'est la probabilité que l'on observe les réalisations $x_1,...,x_n$ quand la vraie valeur du paramètre est θ. Pour certaines valeurs de θ, cette proba-

bilité sera petite : il y a peu de chances d'observer $x_1,...,x_n$. Pour d'autres valeurs de θ, cette probabilité sera forte : il y a de fortes chances d'observer $x_1,...,x_n$. Il est logique de dire qu'une valeur vraisemblable pour θ est la valeur pour laquelle la probabilité d'observer $x_1,...,x_n$ est la plus forte possible. Cela revient à faire comme si c'était l'éventualité la plus probable qui s'était produite au cours de l'expérience.

Mathématiquement, on obtient la définition suivante :

> *Définition : L'**estimateur de maximum de vraisemblance** (EMV) de θ est la valeur $\hat{\theta}_n$ de θ qui rend maximale la vraisemblance $\mathcal{L}(\theta ; X_1,..., X_n)$.*

Dans la plupart des cas, la fonction de vraisemblance s'exprime comme un produit. Il est alors plus commode de remarquer que la valeur qui rend maximale une fonction rend aussi maximal son logarithme. Par conséquent, $\hat{\theta}_n$ sera en général calculé en annulant la dérivée du logarithme de la vraisemblance $\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta; X_1...X_n)$. On remarque que ce calcul est également utile pour déterminer la quantité d'information de Fisher.

Quand $\theta = (\theta_1,...,\theta_p) \in R^p$, $\hat{\theta}_n$ est solution du système d'équations :

$$\frac{\partial}{\partial\theta_i}\ln\mathcal{L}(\theta; X_1...X_n) = 0,\ i = 1..p$$

Un estimateur de maximum de vraisemblance n'est pas forcément unique (la vraisemblance peut avoir plusieurs maxima), ni sans biais, ni de variance minimale, ni efficace. Il n'a pas forcément d'expression explicite (il faut alors résoudre numériquement les équations de vraisemblance).

En revanche, on peut montrer que :

- $\hat{\theta}_n$ converge presque sûrement vers θ
- $\sqrt{I_n(\theta)}(\hat{\theta}_n - \theta) \xrightarrow{L} N(0,1)$, ce qui signifie que, quand n est grand, $\hat{\theta}_n$ est approximativement de loi $N\left(\theta, \frac{1}{I_n(\theta)}\right)$. On en déduit que $\hat{\theta}_n$ est asymptotiquement sans biais et efficace. Cette propriété peut aussi s'écrire $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{L} N\left(0, \frac{1}{I_1(\theta)}\right)$
- si $\hat{\theta}_n$ est l'EMV de θ, alors $\varphi(\hat{\theta}_n)$ est l'EMV de $\varphi(\theta)$; de plus, si φ est dérivable, $\sqrt{n}\left[\varphi(\hat{\theta}_n) - \varphi(\theta)\right] \xrightarrow{L} N\left(0, \frac{\varphi'(\theta)^2}{I_1(\theta)}\right)$
- en général, l'EMV est meilleur que l'EMM au sens où $Var(\hat{\theta}_n) \le Var(\tilde{\theta}_n)$

3.3.2.2. Exemples

Exemple 1 : loi de Poisson

Si les X_i sont de loi $P(\lambda)$, la fonction de vraisemblance est :

$$\mathcal{L}(\lambda ; x_1,...,x_n) = P(X_1 = x_1,..., X_n = x_n ; \lambda) = \prod_{i=1}^{n} P(X_i = x_i ; \lambda)$$

$$= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

D'où $\ln \mathcal{L}(\lambda ; x_1,...,x_n) = -n\lambda + \ln\lambda . \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln x_i!$

Alors $\frac{\partial}{\partial\lambda} \ln \mathcal{L}(\lambda ; x_1,...,x_n) = -n + \frac{1}{\lambda} \sum_{i=1}^{n} x_i$, qui vaut 0 pour $\lambda = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}_n$.

Par conséquent, l'EMV du paramètre de la loi de Poisson est $\hat{\lambda}_n = \overline{X}_n$.
Remarquons que, puisque $E(X) = \lambda$, $\overline{X}_n$ est également l'EMM de λ. On peut montrer que cet estimateur est en fait un ESBVM de λ.

Exemple 2 : loi exponentielle

Si les X_i sont de loi $\exp(\lambda)$, le calcul fait plus haut de l'efficacité de l'EMM a permis d'établir que $\frac{\partial}{\partial\lambda} \ln \mathcal{L}(\lambda ; x_1,...,x_n) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$, ce qui prouve immédiatement que $\hat{\lambda}_n = 1/\overline{X}_n$. Là encore, EMM et EMV sont identiques.

Exemple 3 : loi normale

Si les X_i sont de loi $N(m,\sigma^2)$, la fonction de vraisemblance est :

$$\mathcal{L}(m,\sigma^2 ; x_1,...,x_n) = f_{(X_1...,X_n)}(x_1,...,x_n ; m,\sigma^2) = \prod_{i=1}^{n} f_{X_i}(x_i ; m,\sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-m)^2}{2\sigma^2}} = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2}$$

D'où $\ln \mathcal{L}(m,\sigma^2 ; x_1,...,x_n) = -\frac{n}{2} \ln\sigma^2 - \frac{n}{2} \ln 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2$.

On doit annuler les dérivées partielles de ce logarithme par rapport à m et σ^2. On a :

$\frac{\partial}{\partial m}\ln \mathcal{L}(m,\sigma^2\,;x_1,...,x_n) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} -2(x_i - m) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - nm\right)$, qui s'annule pour $m = \frac{1}{n}\sum_{i=1}^{n} x_i$.

$\frac{\partial}{\partial \sigma^2}\ln \mathcal{L}(m,\sigma^2\,;x_1,...,x_n) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - m)^2$, qui s'annule pour $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$.

$\hat{m}_n$ et $\hat{\sigma}_n^2$ sont les valeurs de m et σ^2 qui vérifient les deux conditions en même temps. On a donc $\hat{m}_n = \overline{X}_n$ et $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = S_n^2$.

Remarque 1 : $S_n'^2$ est un ESBVM de σ^2, mais S_n' n'est pas un ESBVM de σ (ce n'est même pas un estimateur sans biais). On montre qu'en fait, un ESBVM de σ est $\sqrt{\frac{n-1}{2}}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}S_n'$.

Remarque 2 : Dans les trois exemples présentés ici, la méthode des moments et la méthode du maximum de vraisemblance donnent les mêmes résultats. C'est parce que les exemples traités sont élémentaires. En fait, dans la plupart des cas, les deux méthodes fournissent des estimateurs différents (voir par exemple le cas de la loi uniforme sur $[0,\theta]$).

3.4. Intervalles de confiance

3.4.1. Définition

Jusqu'à présent, on a estimé un paramètre θ par une unique valeur $\hat{\theta}_n$ (estimation ponctuelle). Si l'estimateur $\hat{\theta}_n$ est sans biais et de faible variance, on peut s'attendre à ce que chaque réalisation de $\hat{\theta}_n$ soit proche de la vraie valeur de θ. Cependant, $\hat{\theta}_n$ ne sera sûrement pas exactement égal à θ. Donc, plutôt que d'estimer θ par la seule valeur $\hat{\theta}_n$, il semble raisonnable de donner un ensemble de valeurs vraisemblables pour θ, toutes proches de $\hat{\theta}_n$. Comme on supposera ici que $\theta \in R$, on donnera un intervalle (une « fourchette ») ayant une forte probabilité de contenir la vraie valeur de θ.

> *Définition : Un* ***intervalle de confiance*** *de* ***seuil*** *(ou* ***niveau de signification****)* $\alpha \in [0,1]$ *pour un paramètre* θ*, est un intervalle aléatoire* I *tel que* $P(\theta \in I) = 1 - \alpha$.

α est la probabilité que le paramètre θ n'appartienne pas à l'intervalle I, c'est à dire la probabilité que l'on se trompe en affirmant que $\theta \in I$. C'est donc une probabilité d'erreur, qui doit être assez petite. Les valeurs usuelles de α sont 10%, 5%, 1%, etc…

Remarque fondamentale : Les intervalles de confiance suscitent souvent des erreurs d'interprétation et des abus de langage. La raison essentielle est la suivante.
Dans l'écriture $P(\theta \in I)$, θ est une grandeur inconnue mais non aléatoire. Ce sont les bornes de l'intervalle I qui sont aléatoires. Posons $I = [Z_1, Z_2]$. Z_1 et Z_2 sont des variables aléatoires. Soient z_1 et z_2 les réalisations de Z_1 et Z_2 pour une expérience donnée.
A titre indicatif, prenons l'exemple des particules de la fiche d'exercices n°2, pour lequel $\theta = b$. Admettons que $z_1 = 440$ et $z_2 = 460$. Il est correct de dire une phrase du type : « b a 95% de chances d'être compris entre Z_1 et Z_2 », mais il est incorrect de dire : « b a 95% de chances d'être compris entre 440 et 460». En effet, dans cette dernière écriture, il n'y a rien d'aléatoire. b est ou n'est pas dans l'intervalle [440, 460]. La probabilité que b soit compris entre 440 et 460 est donc 0 ou 1, mais pas 95%.
En fait, si on recommence 100 fois l'expérience, on aura 100 réalisations du couple (Z_1, Z_2), et donc 100 intervalles de confiance différents. En moyenne, b sera dans 95 de ces intervalles.
Par conséquent, il vaut mieux dire : « on a une confiance de 95% dans le fait que b soit compris entre 440 et 460 ».
Quand $\theta \in R^p$, $p > 1$, on ne peut plus parler d'intervalle de confiance. L'ensemble des valeurs admissibles pour θ est appelé une **région de confiance**. C'est souvent un ellipsoïde de R^p.

Il semble logique de chercher un intervalle de confiance pour θ de la forme $[\hat{\theta}_n - \varepsilon, \hat{\theta}_n + \varepsilon]$, où $\hat{\theta}_n$ est un estimateur de θ. Il reste alors à déterminer ε de sorte que $P(\hat{\theta}_n - \varepsilon \le \theta \le \hat{\theta}_n + \varepsilon) = 1 - \alpha$. Mais cette démarche ne va pas toujours aboutir car le calcul de ε peut s'avérer très complexe. Le problème est que la loi de probabilité de $\hat{\theta}_n$ dépend de θ, alors que α est un réel fixé à l'avance qui, lui, ne doit pas dépendre de θ. Or $P(\hat{\theta}_n - \varepsilon \le \theta \le \hat{\theta}_n + \varepsilon) = P(-\varepsilon \le \hat{\theta}_n - \theta \le +\varepsilon) = P(|\hat{\theta}_n - \theta| \le \varepsilon)$.
Donc on ne peut déterminer un ε, ne dépendant que des observations et pas de θ, et tel que $P(|\hat{\theta}_n - \theta| \le \varepsilon) = 1 - \alpha$, que si la loi de probabilité de $\hat{\theta}_n - \theta$ ne dépend pas de θ, ce qui n'est pas toujours le cas.

Pour trouver un intervalle de confiance, la méthode la plus efficace consiste à chercher une **fonction pivotale**, c'est à dire une variable aléatoire fonction à la fois du paramètre θ et des observations $X_1,...,X_n$, dont la loi de probabilité ne dépende pas de θ. Les sections suivantes ont pour but d'illustrer cette méthodologie par des exemples.

3.4.2. Intervalles de confiance pour les paramètres de la loi normale

3.4.2.1. Intervalle de confiance pour la moyenne

Si $X_1,...,X_n$ sont indépendantes et de même loi normale $N(m,\sigma^2)$, on sait que l'ESBVM de m est $\overline{X}_n$. La première idée est donc de chercher un intervalle de confiance pour m de la forme $[\overline{X}_n-\varepsilon,\overline{X}_n+\varepsilon]$. Conformément à ce qui précède, le problème revient, pour α fixé, à chercher ε tel que $P(|\overline{X}_n-m|\le\varepsilon)=1-\alpha$.

Les propriétés élémentaires de la loi normale permettent d'établir que $\sum_{i=1}^{n}X_i$ est de loi $N(nm,n\sigma^2)$ et que $\overline{X}_n$ est de loi $N(m,\frac{\sigma^2}{n})$. Par conséquent, $U=\frac{\overline{X}_n-m}{\sqrt{\sigma^2/n}}=\frac{\overline{X}_n-m}{\sigma}\sqrt{n}$ est de loi $N(0,1)$.

Alors $P(|\overline{X}_n-m|\le\varepsilon)=P(|U|\le\frac{\varepsilon\sqrt{n}}{\sigma})=1-P(|U|>\frac{\varepsilon\sqrt{n}}{\sigma})=1-\alpha$. Or la table 2 de la loi normale donne la valeur u_α telle que $P(|U|>u_\alpha)=\alpha$. Par conséquent, $\frac{\varepsilon\sqrt{n}}{\sigma}=u_\alpha$, donc $\varepsilon=\frac{\sigma}{\sqrt{n}}u_\alpha$. D'où le résultat :

> *<u>Propriété</u> : Un intervalle de confiance de seuil α pour le paramètre m de la loi $N(m,\sigma^2)$ est* $[\overline{X}_n-\frac{\sigma}{\sqrt{n}}u_\alpha,\overline{X}_n+\frac{\sigma}{\sqrt{n}}u_\alpha]$

Le problème est que cet intervalle n'est utilisable que si on connaît la valeur de σ. Or, dans la pratique, on ne connaît jamais les vraies valeurs des paramètres.

Une idée naturelle est alors de remplacer σ par un estimateur, par exemple S'_n.

Mais si on fait cela, $P(\overline{X}_n-\frac{S'_n}{\sqrt{n}}u_\alpha\le m\le\overline{X}_n+\frac{S'_n}{\sqrt{n}}u_\alpha)=P(\left|\frac{\overline{X}_n-m}{S'_n}\sqrt{n}\right|\le\varepsilon)$ n'est pas égale à $1-\alpha$, car $\frac{\overline{X}_n-m}{S'_n}\sqrt{n}$ n'est pas de loi $N(0,1)$, donc $[\overline{X}_n-\frac{S'_n}{\sqrt{n}}u_\alpha,\overline{X}_n+\frac{S'_n}{\sqrt{n}}u_\alpha]$ n'est pas un intervalle de confiance de seuil α pour m.
On peut cependant résoudre le problème en utilisant le théorème de Fisher :

Théorème de Fisher : Si $X_1,...,X_n$ *sont* n *variables aléatoires indépendantes et de même loi normale* $N(m,\sigma^2)$*, alors :*

- $\overline{X}_n$ *est de loi* $N(m,\frac{\sigma^2}{n})$
- $\frac{nS_n^2}{\sigma^2}$ *est de loi du khi deux à* $n-1$ *degrés de libertés* χ^2_{n-1}
- $\overline{X}_n$ *et* S_n^2 *sont indépendantes*
- $\frac{\overline{X}_n - m}{S'_n}\sqrt{n} = \frac{\overline{X}_n - m}{S_n}\sqrt{n-1}$ *est de loi de Student* $St(n-1)$

On peut alors écrire $P(|\overline{X}_n - m| \le \varepsilon) = P(|Y| \le \frac{\varepsilon\sqrt{n}}{S'_n}) = 1 - P(|Y| > \frac{\varepsilon\sqrt{n}}{S'_n})$, où Y est une variable aléatoire de loi $St(n-1)$. Or la table de la loi de Student donne la valeur $t_{n-1,\alpha}$ telle que $P(|Y| > t_{n-1,\alpha}) = \alpha$. Par conséquent, $\frac{\varepsilon\sqrt{n}}{S'_n} = t_{n-1,\alpha}$, donc $\varepsilon = \frac{S'_n}{\sqrt{n}} t_{n-1,\alpha}$. D'où le résultat :

Propriété : Un intervalle de confiance de seuil α *pour le paramètre* m *de la loi* $N(m,\sigma^2)$ *est*
$[\overline{X}_n - \frac{S'_n}{\sqrt{n}} t_{n-1,\alpha}, \overline{X}_n + \frac{S'_n}{\sqrt{n}} t_{n-1,\alpha}] = [\overline{X}_n - \frac{S_n}{\sqrt{n-1}} t_{n-1,\alpha}, \overline{X}_n + \frac{S_n}{\sqrt{n-1}} t_{n-1,\alpha}]$

Dans l'exemple des niveaux de bruit de la fiche d'exercices 1, on a $n = 20$, $\bar{x}_n = 64.2$ et $s'_n = 5.02$. Pour $\alpha = 5\%$, la table de la loi de Student donne $t_{19,0.05} = 2.093$. On en déduit qu'un intervalle de confiance de seuil 5% pour le niveau de bruit moyen est [61.8, 66.6].

Interprétation : La meilleure estimation possible du niveau de bruit moyen est 64.2 db. De plus, on a une confiance de 95% dans le fait que ce niveau de bruit moyen est compris entre 61.8 db et 66.6 db.

Sous S+, u_α est obtenu par la commande `qnorm(1-alpha/2)` et $t_{n,\alpha}$ par la commande `qt(1-alpha/2,n)`.

Remarque 1 : Rien n'oblige à prendre un intervalle de confiance du type $[\overline{X}_n - \varepsilon, \overline{X}_n + \varepsilon]$ (intervalle de confiance bilatéral). Tout intervalle I tel que $P(m \in I) = 1 - \alpha$ convient. Par exemple, des intervalles de la forme $[A,+\infty[$ et $]-\infty,B]$ (intervalles de confiance unilatéraux) fournissent des bornes inférieure et supérieure pour l'estimation de m.

Remarque 2 : La largeur de l'intervalle de confiance est $2\frac{S'_n}{\sqrt{n}} t_{n-1,\alpha}$. La table de la loi de Student permet de constater que c'est une fonction décroissante en n comme en α, ce qui est logique. En effet, plus on a d'observations, plus on a d'informations, donc plus l'incertitude sur le paramètre diminue et plus l'intervalle de confiance est étroit. D'autre part, plus α est petit, moins on veut prendre de risques de se tromper en disant que m est dans l'intervalle, donc plus on aura tendance à prendre

des intervalles larges. A la limite, on ne prend aucun risque ($\alpha = 0$) en proposant comme intervalle de confiance R tout entier !
En pratique, un intervalle de confiance trop large n'a aucun intérêt, donc il faut parfois accepter un risque d'erreur relativement fort pour obtenir un intervalle de confiance utilisable.

Remarque 3 : La variable aléatoire $\frac{\overline{X}_n - m}{S'_n}\sqrt{n}$ est une fonction des observations $X_1,...,X_n$ et du paramètre m pour lequel on recherche un intervalle de confiance, dont la loi de probabilité ne dépend pas des paramètres du modèle m et σ^2. C'est ce qu'on a appelé une **fonction pivotale** et c'est ce que nous utiliserons à partir de maintenant pour construire des intervalles de confiance.

3.4.2.2. Intervalle de confiance pour la variance

Conformément à ce qui précède, on recherche une fonction pivotale, c'est à dire une fonction des observations $X_1,...,X_n$ et de σ^2, dont la loi de probabilité ne dépend ni de m ni de σ^2. Une telle fonction est donnée par le théorème de Fisher : $\frac{nS_n^2}{\sigma^2}$ est de loi χ^2_{n-1}.

On a donc, quels que soient les réels a et b, $0 < a < b$:

$$P(a \le \frac{nS_n^2}{\sigma^2} \le b) = P(\frac{nS_n^2}{b} \le \sigma^2 \le \frac{nS_n^2}{a}) = F_{\chi^2_{n-1}}(b) - F_{\chi^2_{n-1}}(a)$$

Il y a une infinité de façons de choisir a et b de sorte que cette probabilité soit égale à $1-\alpha$. On montre que les valeurs pour lesquelles $b-a$ est minimum (on cherche à obtenir l'intervalle de confiance le plus étroit possible) sont telles que $F_{\chi^2_{n-1}}(b) = 1 - \frac{\alpha}{2}$ et $F_{\chi^2_{n-1}}(a) = \frac{\alpha}{2}$.

La table de la loi du χ^2 donne la valeur $z_{n,\alpha}$ telle que, quand Z est une variable aléatoire de loi χ^2_n, alors $P(Z > z_{n,\alpha}) = 1 - F_{\chi^2_n}(z_{n,\alpha}) = \alpha$.

Alors, pour $b = z_{n-1,\alpha/2}$ et $a = z_{n-1,1-\alpha/2}$, on a bien $P(\frac{nS_n^2}{b} \le \sigma^2 \le \frac{nS_n^2}{a}) = 1-\alpha$. D'où le résultat :

> *<u>Propriété</u> : Un intervalle de confiance de seuil α pour le paramètre σ^2 de la loi $N(m,\sigma^2)$ est* $\left[\frac{nS_n^2}{z_{n-1,\alpha/2}}, \frac{nS_n^2}{z_{n-1,1-\alpha/2}}\right] = \left[\frac{(n-1)S'^2_n}{z_{n-1,\alpha/2}}, \frac{(n-1)S'^2_n}{z_{n-1,1-\alpha/2}}\right]$

Dans l'exemple des niveaux de bruit, on a $n = 20$ et $s_n^2 = 25.2$.
Pour $\alpha = 5\%$, on obtient $z_{19,0.025} = 32.85$ et $z_{19,0.975} = 8.91$. On en déduit qu'un intervalle de confiance de seuil 5% pour la variance du niveau de bruit est [15.3, 56.6].
On constate que cet intervalle de confiance est très large : l'estimation de la variance est moins précise que celle de la moyenne.

Sous S+, $z_{n,\alpha}$ est obtenu par la commande `qchisq(1-alpha,n)`.

Remarque 1 : $P(a \le \sigma^2 \le b) = P(\sqrt{a} \le \sigma \le \sqrt{b})$, donc un intervalle de confiance de seuil α pour l'écart-type σ est $\left[\sqrt{\frac{n}{z_{n-1,\alpha/2}}} S_n, \sqrt{\frac{n}{z_{n-1,1-\alpha/2}}} S_n \right]$.

Remarque 2 : L'intervalle de confiance est de la forme $[\varepsilon_1 S_n^2, \varepsilon_1 S_n^2]$ et non pas $[S_n^2 - \varepsilon, S_n^2 + \varepsilon]$. C'est parce que la loi de probabilité de εS_n^2 est plus facile à manipuler que celle de $S_n^2 + \varepsilon$.

Exercice : Montrer qu'un intervalle de confiance de seuil α pour le paramètre λ de la loi exponentielle est $\left[\frac{z_{2n,1-\alpha/2}}{2\sum_{i=1}^{n} X_i}, \frac{z_{2n,\alpha/2}}{2\sum_{i=1}^{n} X_i} \right]$. Qu'obtient-on pour l'exemple des durées de transfert ?

3.4.3. Estimation et intervalle de confiance pour une proportion

On désire évaluer la probabilité p qu'un événement A se produise au cours d'une expérience donnée : $p = P(A)$. Pour cela, on fait n expériences identiques et indépendantes et on compte le nombre x de fois où A s'est produit. x est la réalisation d'une variable aléatoire X qu'on sait être de loi binomiale $B(n, p)$.

Exemple : Une élection oppose deux candidats A et B. Un institut de sondage interroge 800 personnes sur leurs intentions de vote. 420 déclarent voter pour A et 380 pour B. Estimer le résultat de l'élection, c'est estimer le pourcentage p de voix qu'obtiendra le candidat A. En supposant que les réponses des 800 personnes interrogées sont indépendantes, on est bien dans le cas de figure de l'estimation d'une proportion.

3.4.3.1. Estimation ponctuelle

Remarquons que nous n'avons ici qu'une seule réalisation de X, c'est à dire un échantillon de taille 1. Pour une fois, la notation n ne désigne pas la taille de l'échantillon.

Il est naturel d'estimer la probabilité p que A se produise par le pourcentage $\frac{X}{n}$ de fois où A s'est produit au cours des n expériences.

Par la méthode des moments, on a $E(X) = np$, donc l'EMM de p est $\frac{X}{n}$.

Par la méthode du maximum de vraisemblance, on a $\mathcal{L}(p; x) = P(X = x) = C_n^x p^x (1-p)^{n-x}$.

D'où $\ln \mathcal{L}(p; x) = \ln C_n^x + x \ln p + (n-x) \ln(1-p)$.

Alors $\frac{\partial}{\partial p}\ln \mathcal{L}(p;x) = \frac{x}{p} - \frac{n-x}{1-p} = \frac{x-np}{p(1-p)}$, qui s'annule pour $p = \frac{x}{n}$.

Par conséquent, l'EMV, l'EMM et l'estimateur naturel sont tous égaux à $\hat{p} = \frac{X}{n}$. Déterminons les qualités de cet estimateur.

Biais : $E(\hat{p}) = E(\frac{X}{n}) = \frac{1}{n}E(X) = \frac{1}{n}np = p$, donc $\hat{p}$ est sans biais.

Convergence : $Var(\hat{p}) = Var(\frac{X}{n}) = \frac{1}{n^2}Var(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$ donc $\hat{p}$ est convergent.

Efficacité : $Eff(\hat{p}) = \frac{\left[\frac{\partial}{\partial p}E(\hat{p})\right]^2}{I(p)Var(\hat{p})} = \frac{n}{I(p)p(1-p)}$,

avec $I(p) = Var\left[\frac{\partial}{\partial p}\ln \mathcal{L}(p;X)\right] = Var\left[\frac{X-np}{p(1-p)}\right] = \frac{Var(X)}{p^2(1-p)^2} = \frac{np(1-p)}{p^2(1-p)^2} = \frac{n}{p(1-p)}$,

d'où $Eff(\hat{p}) = 1$: $\hat{p}$ est un estimateur efficace. D'où le résultat :

Propriété : $\hat{p} = \frac{X}{n}$ *est l'ESBVM de* p

3.4.3.2. Intervalle de confiance

Une fonction pivotale est une fonction de X et p dont la loi ne dépend pas de p. Il n'en existe pas de simple. On montre le résultat suivant :

Propriété : Un intervalle de confiance exact de seuil α *pour* p *est :*

$$\left[\frac{1}{1+\frac{n-X+1}{X}f_{2(n-X+1),2X,\alpha/2}}, \frac{1}{1+\frac{n-X}{X+1}f_{2(n-X),2(X+1),1-\alpha/2}}\right]$$

où les $f_{\nu_1,\nu_2,\alpha}$ *sont à lire dans des tables de la loi de Fisher-Snedecor*

Sous S+, $f_{\nu_1,\nu_2,\alpha}$ est obtenu par la commande `qf(1-alpha,nu1,nu2)`.

Si on ne dispose pas de logiciel, cet intervalle n'est pas facile à utiliser car il nécessite l'emploi de nombreuses tables. C'est pourquoi on utilise souvent un intervalle de confiance approché, basé sur l'approximation de la loi binomiale par la loi normale.

En effet, si $np \geq 5$ et $n(1-p) \geq 5$, on peut approcher la loi binomiale $B(n,p)$ par la loi normale $N(np, np(1-p))$. Donc $\dfrac{X-np}{\sqrt{np(1-p)}}$ est approximativement de loi $N(0,1)$, ce qui fournit la fonction pivotale cherchée.

On écrit alors $P\left(\left|\dfrac{X-np}{\sqrt{np(1-p)}}\right| \leq u_\alpha\right) = 1-\alpha$. Pour en déduire un intervalle de confiance, il suffit d'écrire $\left|\dfrac{X-np}{\sqrt{np(1-p)}}\right| \leq u_\alpha$ sous la forme $A \leq p \leq B$. On a :

$$\left|\frac{X-np}{\sqrt{np(1-p)}}\right| \leq u_\alpha \Leftrightarrow \frac{(X-np)^2}{np(1-p)} \leq u_\alpha^2 \Leftrightarrow p^2(n+u_\alpha^2) - p(2X+u_\alpha^2) + \frac{X^2}{n} \leq 0$$

Ce trinôme en p est toujours positif sauf entre ses racines. Donc ces deux racines sont les bornes de l'intervalle de confiance cherché. Puisque l'approximation de la loi binomiale par la loi normale n'est valable que quand n est suffisamment grand, cet intervalle porte le nom d'intervalle de confiance asymptotique.

Propriété : Un intervalle de confiance asymptotique de seuil α pour p est :

$$\left[\frac{\dfrac{X}{n} + \dfrac{u_\alpha^2}{2n} - u_\alpha\sqrt{\dfrac{u_\alpha^2}{4n^2} + \dfrac{X(n-X)}{n^3}}}{1+\dfrac{u_\alpha^2}{n}}, \frac{\dfrac{X}{n} + \dfrac{u_\alpha^2}{2n} + u_\alpha\sqrt{\dfrac{u_\alpha^2}{4n^2} + \dfrac{X(n-X)}{n^3}}}{1+\dfrac{u_\alpha^2}{n}}\right]$$

Souvent on néglige u_α^2 par rapport à n, et on obtient un intervalle de confiance asymptotique approché de seuil α pour p :

$$\left[\frac{X}{n} - u_\alpha\sqrt{\frac{X(n-X)}{n^3}}, \frac{X}{n} + u_\alpha\sqrt{\frac{X(n-X)}{n^3}}\right]$$

Exemple du sondage : l'ESBVM de p est $\hat{p} = \dfrac{X}{n}$. Ici, $\hat{p} = \dfrac{420}{800} = 52.5\%$, donc l'institut de sondage estime que le candidat A va gagner l'élection. Un intervalle de confiance exact de seuil α pour p est :

$$\left[\frac{1}{1+\frac{381}{420}f_{762,840,\alpha/2}}, \frac{1}{1+\frac{380}{421}f_{760,842,1-\alpha/2}}\right].$$

La table de la loi de Fisher-Snedecor permet de voir que, pour les valeurs usuelles de α, $f_{762,840,\alpha/2}$ et $f_{760,842,1-\alpha/2}$ sont de l'ordre de 1. Pour α = 5%, on trouve en fait $f_{762,840,0.025}$ = 1.1486 et $f_{760,842,0.975}$ = 0.8702. On obtient alors comme intervalle de confiance exact [0.4896, 0.5600].

Pour α = 5%, $u_{0.05}$ = 1.96. L'intervalle de confiance asymptotique de seuil 5% est alors [0.49036, 0.55940]. Mais $u_{0.05}^2$ = 3.8 est négligeable par rapport à n = 800. On peut donc utiliser l'intervalle de confiance asymptotique approché [0.49039, 0.55960].

On constate que l'écart entre les trois intervalles est négligeable. C'est souvent le cas, ce qui fait que l'intervalle asymptotique approché est très largement utilisé.

Pour simplifier, on peut dire que l'on a une confiance de 95% dans le fait que le pourcentage de voix obtenu par le candidat A sera compris entre 49% et 56%.

Le problème est que cet intervalle de confiance n'est pas entièrement situé au-dessus de 50%. Il semble donc possible que, malgré l'estimation de 52.5%, le candidat A soit battu. On voit donc que ce qui importe dans cette situation, ce n'est pas vraiment d'estimer p, mais de déterminer si on peut admettre avec une confiance raisonnable que p est supérieur à 50%. C'est, entre autres, l'objet de la théorie des tests d'hypothèses, qui sera abordée au chapitre suivant.

Une autre possibilité pour résoudre le problème est de déterminer à quelle condition l'intervalle de confiance pour p sera entièrement au-dessus des 50%. Il s'agit donc de réduire la taille de l'intervalle de confiance. Si on prend l'intervalle asymptotique approché, sa largeur est $2u_\alpha\sqrt{\frac{X(n-X)}{n^3}}$. Donc, pour diminuer cette largeur, on peut, au choix, diminuer u_α ou augmenter n.

Diminuer u_α, c'est augmenter α, donc augmenter la probabilité de se tromper en affirmant que le candidat est élu. On retrouve ce qui a déjà été dit : pour obtenir des intervalles de confiance exploitables, il faut parfois accepter un risque d'erreur assez élevé.

Augmenter n, c'est augmenter le nombre de personnes interrogées. On peut même, à α fixé, déterminer n de façon à obtenir la largeur que l'on veut pour l'intervalle de confiance.

Soit l une largeur objectif : $l = 2u_\alpha\sqrt{\frac{X(n-X)}{n^3}} = 2u_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Or $\forall p \in [0,1], p(1-p) \le \frac{1}{4}$, donc $2u_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \frac{u_\alpha}{\sqrt{n}}$. Par conséquent, si on détermine n tel que $\frac{u_\alpha}{\sqrt{n}} < l$, c'est à dire $n > \frac{u_\alpha^2}{l^2}$, on est sûr que la largeur de l'intervalle de confiance sera inférieure à l.

Pour α = 5% et n = 800, $\frac{u_\alpha}{\sqrt{n}} = \frac{1.96}{\sqrt{800}} \approx 7\%$. La précision sur l'estimation de p est donc, avec une confiance de 95%, de plus ou moins 3.5%. Si on veut, avec le même niveau de confiance, avoir une précision inférieure à 1%, il faudra interroger au moins $\frac{u_\alpha^2}{l^2} = \frac{1.96^2}{0.01^2} = 38416$ personnes. C'est rarement le cas dans les sondages, pour lesquels le nombre de personnes interrogées est en général de l'ordre de 1000.

En conclusion, il faut toujours tenir compte du nombre de personnes interrogées pour interprêter les résultats d'un sondage. C'est pour cela qu'il est obligatoire de préciser ce nombre quand les résultats du sondage sont publiés.

Chapitre 4 : Tests d'hypothèses

4.1. Introduction : le problème de décision

Dans tous les domaines, de l'expérimentation scientifique à la vie quotidienne, on est amenés à prendre des décisions au vu de résultats d'expériences ou d'observation de phénomènes. Par exemple :

- *contrôle de qualité* : au vu du nombre d'objets défectueux produits par une machine, on doit décider si ce nombre est conforme à une certaine norme, décider si la machine est à remplacer ou pas.
- *essais thérapeutiques* : décider si un nouveau traitement médical est meilleur qu'un ancien au vu du résultat de son expérimentation sur des malades.
- *justice* : décider si l'accusé est innocent ou coupable à partir des informations acquises pendant le procès.

Dans chaque cas, le **problème de décision** consiste à trancher, au vu d'observations, entre une hypothèse appelée **hypothèse nulle**, notée H_0, et une autre hypothèse dite **hypothèse alternative**, notée H_1. En général, on suppose qu'une et une seule de ces deux hypothèses est vraie.

Un **test d'hypothèses** est une procédure qui permet de choisir entre ces deux hypothèses.

Dans un problème de décision, deux types d'erreurs sont possibles :

- **erreur de première espèce :** décider que H_1 est vraie alors que H_0 est vraie.
- **erreur de seconde espèce :** décider que H_0 est vraie alors que H_1 est vraie.

Les conséquences de ces deux erreurs peuvent être d'importances diverses. En général, une des erreurs est plus grave que l'autre :

- *contrôle de qualité* : si on décide à tort que la machine n'est pas aux normes, on engagera des dépenses inutiles de réparation ou de changement de matériel; si on décide à tort qu'elle est aux normes, on risque de produire de mauvaises pièces, ce qui peut aboutir à un mécontentement des clients, voire à des problèmes de sécurité.
- *essais thérapeutiques* : on peut adopter un nouveau traitement moins efficace, voire pire que l'ancien, ou se priver d'un nouveau traitement plus efficace que l'ancien.
- *justice* : on peut condamner un innocent ou acquitter un coupable.

A toute décision correspond une probabilité de décider juste et une probabilité de se tromper :

- la probabilité de l'erreur de première espèce, qui est la probabilité de rejeter à tort H_0, est notée α et est appelée **seuil** ou **niveau de signification** du test. C'est la même terminologie que pour les intervalles de confiance, ce qui n'est pas un hasard, comme nous le verrons plus loin. Dans certains contextes, cette probabilité est appelée **risque fournisseur**.
- la probabilité de l'erreur de deuxième espèce est notée $1-\beta$ et est parfois appelée **risque client**.
- β est la probabilité de décider H_1 ou de rejeter H_0 à raison. Elle est appelée **puissance** du test.
- $1-\alpha$ est parfois appelée **niveau de confiance** du test.

Le tableau 4.1. résume simplement le rôle de ces probabilités.

Vérité / Décision	H_0	H_1
H_0	$1-\alpha$	$1-\beta$
H_1	α	β

Tableau 4.1. : probabilités de bonne et mauvaise décision dans un test d'hypothèses

L'idéal serait de diminuer les deux risques d'erreur en même temps. Malheureusement, on montre qu'ils varient en sens inverse, c'est-à-dire que toute procédure diminuant α va augmenter $1-\beta$ et réciproquement. Dans la pratique, on va donc considèrer que l'une des deux erreurs est plus importante que l'autre, et tacher d'éviter que cette erreur se produise. Il est alors possible que l'autre erreur survienne. Par exemple, dans le cas du procès, on fait en général tout pour éviter de condamner un innocent, quitte à prendre le risque d'acquitter un coupable.

On va choisir H_0 et H_1 de sorte que l'erreur que l'on cherche à éviter soit l'erreur de première espèce. Mathématiquement cela revient à se fixer la valeur du seuil du test α. Plus la conséquence de l'erreur est grave, plus α sera choisi petit. Les valeurs usuelles de α sont 10%, 5%, 1%, …

On appelle **règle de décision** une règle qui permette de choisir entre H_0 et H_1 au vu des observations $x_1,...,x_n$, sous la contrainte que la probabilité de rejeter à tort H_0 est égale à α fixé. Une idée naturelle est de conclure que H_0 est fausse si il est très peu probable d'observer $x_1,...,x_n$ quand H_0 est vraie.

On appelle **région critique** du test, et on note W, l'ensemble des valeurs des observations $x_1,...,x_n$ pour lesquelles on rejettera H_0. La région critique est souvent déterminée à l'aide du bon sens. Sinon, on utilisera une fonction pivotale ou des théorèmes d'optimalité. W dépend du seuil α et est déterminée a priori, indépendamment de la valeur des observations. Ensuite, si les observations appartiennent à W, on rejette H_0, sinon on ne la rejette pas.

Remarque : il vaut mieux dire « *ne pas rejeter* H_0 » que « *accepter* H_0 ». En effet, si on rejette H_0, c'est que les observations sont telles qu'il est très improbable que H_0 soit vraie. Si on ne rejette pas H_0, c'est qu'on ne dispose pas de critères suffisants pour pouvoir dire que H_0 est fausse. Mais cela ne veut pas dire que H_0 est vraie. Un test permet de dire qu'une hypothèse est très probablement fausse ou seulement peut-être vraie.
Par conséquent, dans un problème de test, il faut choisir les hypothèses H_0 et H_1 de façon à ce que ce qui soit vraiment intéressant, c'est de rejeter H_0.

Récapitulons l'ensemble de la démarche à suivre pour effectuer un test d'hypothèses :

1. Choisir H_0 et H_1 de sorte que ce qui importe, c'est le rejet de H_0.
2. Se fixer α selon la gravité des conséquences de l'erreur de première espèce.
3. Déterminer la région critique W.
4. Regarder si les observations se trouvent ou pas dans W.
5. Conclure au rejet ou au non-rejet de H_0.

Pour le même problème de décision, plusieurs tests (c'est-à-dire plusieurs régions critiques) de même seuil sont souvent possibles. Dans ce cas, le meilleur de ces tests est celui qui minimisera la probabilité de l'erreur de seconde espèce, c'est à dire celui qui maximisera la puissance β. Le meilleur des tests possibles de seuil fixé est le **test le plus puissant**. Il arrive, mais pas toujours, que l'on puisse le déterminer.

Dans de nombreux cas, les hypothèses d'un test peuvent se traduire sur la valeur d'un paramètre d'une loi de probabilité. Les tests de ce type sont appelés **tests paramétriques**. Dans l'exemple de l'élection, le problème est de trancher entre les deux hypothèses « $p \leq 1/2$ » et « $p > 1/2$ ».

On s'intéressera ici à des tests paramétriques portant sur un échantillon et à des tests portant sur deux échantillons. Ces derniers tests permettent de comparer deux populations. On pourra par exemple répondre à des questions du type :
« Le nouveau traitement est-il plus efficace que l'ancien ? »
« Les processeurs de la nouvelle génération sont-ils plus rapides que les anciens ? »

Les tests qui ne portent pas sur la valeur d'un paramètre sont appelés **tests non paramétriques**. Il en existe de tous les types. On ne s'intéressera ici qu'aux tests permettant de :

- déterminer si un échantillon provient d'une loi de probabilité donnée : **tests d'adéquation**

- déterminer si deux échantillons proviennent de la même loi de probabilité : **tests de comparaison d'échantillons**.

4.2. Tests paramétriques sur un échantillon

4.2.1. Formalisation du problème

Dans cette section, on supposera que les observations $x_1,...,x_n$ sont les réalisations de variables aléatoires $X_1,...,X_n$ indépendantes et de même loi, dépendant d'un paramètre inconnu θ. On supposera que θ est un réel. Si θ est un paramètre vectoriel, on fera des tests sur chacune de ses composantes. Par exemple, on fera des tests sur la moyenne de la loi normale, puis des tests sur la variance, mais pas sur les deux en même temps.

Une **hypothèse** est **simple** si elle est du type « $\theta = \theta_0$ », où θ_0 est un réel fixé. Une **hypothèse** est **composite** si elle est du type « $\theta \in A$ » où A est une partie de R non réduite à un élément.

4.2.1.1. Tests d'hypothèses simples

Un **test d'hypothèses simples** est un test du type H_0 : « $\theta = \theta_0$ » contre H_1 : « $\theta = \theta_1$ ».

Un tel test est un cas d'école : il permet de dire laquelle des deux valeurs θ_0 et θ_1 est la plus vraisemblable au vu des observations. Mais il ne correspond pas à un problème de décision tel qu'il a été formulé plus haut, dans lequel une des deux hypothèses doit être vraie. Ici, il est possible que θ ne soit égal ni à θ_0 ni à θ_1.

Le seuil du test est la probabilité de rejeter à tort H_0 : $\alpha = P\big((X_1,...,X_n) \in W ; \theta_0\big)$.

La puissance du test est la probabilité de rejeter à raison H_0 : $\beta = P((X_1,...,X_n) \in W ; \theta_1)$.

4.2.1.2. Tests d'hypothèses composites

Un **test d'hypothèses composites** est un test dans lequel l'une au moins des deux hypothèses est composite. Les tests les plus usuels sont du type :

- **test bilatéral** : H_0 : « $\theta = \theta_0$ » contre H_1 : « $\theta \neq \theta_0$ » (seule H_1 est composite).
- **test unilatéral** : H_0 : « $\theta \leq \theta_0$ » contre H_1 : « $\theta > \theta_0$ »
  ou H_0 : « $\theta \geq \theta_0$ » contre H_1 : « $\theta < \theta_0$ » (H_0 et H_1 sont composites).

On pourrait aussi imaginer des tests du type H_0 : « $\theta \in [\theta_1, \theta_2]$ » contre H_1 : « $\theta < \theta_1$ ou $\theta > \theta_2$ ». Toutes les variantes sont envisageables.

Quand une hypothèse est composite, la notion de puissance est à repréciser. En effet, β a été définie comme la probabilité de rejeter H_0 à raison, c'est à dire de rejeter H_0 quand H_1 est vraie. Or, dans les exemples ci-dessus, il y a une infinité de valeurs de θ pour lesquelles H_1 est vraie. Donc la puissance du test doit dépendre de la vraie valeur de θ, ce qui nous amène à redéfinir la puissance et le seuil d'un test :

> *Définition : La **puissance** d'un test portant sur la valeur d'un paramètre réel θ est la fonction de θ définie par :*
>
> $$\beta : R \to [0,1]$$
> $$\theta \mapsto \beta(\theta) = P((X_1,...,X_n) \in W ; \theta)$$
>
> *Le **seuil** du test est* $\alpha = \underset{H_0}{Sup}\, \beta(\theta)$.

$\beta(\theta)$ est la probabilité de rejeter H_0 quand la vraie valeur du paramètre est θ.

$\alpha = \underset{H_0}{Sup}\, \beta(\theta)$ est la probabilité maximale de rejeter H_0 alors que H_0 est vraie, c'est à dire la plus forte probabilité de rejeter à tort H_0. Par exemple, pour un test bilatéral, $\alpha = \beta(\theta_0)$, et pour le premier test unilatéral présenté, $\alpha = \underset{\theta \leq \theta_0}{Sup}\, \beta(\theta)$.

Une fois H_0 et H_1 déterminées et α fixé, il faut construire la région critique W. L'exemple introductif suivant va permettre de comprendre comment on peut déterminer une région critique.

4.2.2. Exemple introductif : tests sur la moyenne d'une loi normale

4.2.2.1. Modélisation

Pour apaiser un certain type de maux de tête, on a l'habitude de traiter les malades avec un médicament A. Une étude statistique a montré que le temps de disparition de la douleur chez les malades

traités avec A était une variable aléatoire de loi normale $N(m_0, \sigma_0^2)$, avec $m_0 = 30$ mn et $\sigma_0 = 5$ mn. Un laboratoire pharmaceutique a conçu un nouveau médicament B et désire tester son efficacité. Pour cela, le nouveau médicament a été administré à n malades cobayes, et on a mesuré le temps de disparition de la douleur pour chacun d'entre eux : $x_1,...,x_n$. Une étude de statistique descriptive sur ces données a amené les bio-pharmaciens à considérer que ce temps était une variable aléatoire de loi normale $N(m,\sigma^2)$.

Remarque : En toute rigueur, on ne devrait pas modéliser une durée (positive) par une variable aléatoire qui, comme pour la loi normale, peut prendre des valeurs négatives. En pratique, on peut le faire quand, pour les lois considérées, la probabilité que la variable soit négative est négligeable.

L'effet du nouveau médicament se traduit facilement sur la valeur de la durée moyenne de disparition de la douleur :

- « $m = m_0$ » : le médicament B a en moyenne le même effet que le médicament A
- « $m < m_0$ » : le médicament B est en moyenne plus efficace que le médicament A
- « $m > m_0$ » : le médicament B est en moyenne moins efficace que le médicament A

Nous reviendrons ultérieurement sur l'interprétation de la valeur de l'écart-type σ en termes d'efficacité du médicament.

Pour savoir s'il faut commercialiser B, il faut trancher entre ces 3 hypothèses. L'important est de ne pas se tromper si on décide de changer de médicament : il est préférable de conserver un médicament moins performant que le nouveau que d'adopter un médicament moins performant que l'ancien. Il faut donc que l'hypothèse « $m < m_0$ » corresponde au rejet de H_0.

Par conséquent, nous allons tester H_0 : « $m \geq m_0$ » contre H_1 : « $m < m_0$ » au vu de n réalisations indépendantes $x_1,...,x_n$ de la loi $N(m,\sigma^2)$.

4.2.2.2. Première idée

Puisque $\overline{X}_n$ est l'ESBVM de m, une première idée est de conclure que $m < m_0$ si et seulement si $\overline{x}_n < m_0$: la durée moyenne de disparition de la douleur sur les malades traités avec B est plus petite que ce qu'elle est sur les malades traités avec A.
Cela revient à proposer comme région critique du test $W = \{(x_1,...,x_n)\,;\,\overline{x}_n < m_0\}$.

Si $\overline{x}_n$ est beaucoup plus petit que m_0, il est en effet très probable que B soit plus efficace que A. Mais si $\overline{x}_n$ est proche de m_0 tout en étant plus petit, on risque de se tromper si on affirme que $m < m_0$. La probabilité de cette erreur, qui n'est autre que le risque de première espèce α, est très facile à calculer :

$$\alpha = \sup_{H_0} \beta(m) = \sup_{m \geq m_0} P(\overline{X}_n < m_0\,;m)$$
$$= \sup_{m \geq m_0} P\left(\frac{\overline{X}_n - m}{\sigma}\sqrt{n} < \frac{m_0 - m}{\sigma}\sqrt{n}\,;m\right) = \sup_{m \geq m_0} \phi\left(\frac{m_0 - m}{\sigma}\sqrt{n}\right)$$

où ϕ est la fonction de répartition de la loi normale centrée-réduite. En effet, si $X_1,...,X_n$ sont indépendantes et de même loi $N(m,\sigma^2)$, alors $\overline{X}_n$ est de loi $N(m,\frac{\sigma^2}{n})$ et $\frac{\overline{X}_n - m}{\sigma}\sqrt{n}$ est de loi $N(0,1)$.

$\phi(u)$ est une fonction croissante de u, donc $\beta(m) = \phi\left(\frac{m_0 - m}{\sigma}\sqrt{n}\right)$ est une fonction décroissante de m.

Par conséquent, $\alpha = \underset{m \geq m_0}{Sup}\ \beta(m) = \beta(m_0) = \phi(0) = \frac{1}{2}$.

Il y a donc une chance sur deux de se tromper si on décide que B est plus efficace que A quand $\bar{x}_n < m_0$. C'est évidemment beaucoup trop.

4.2.2.3. Deuxième idée

On voit qu'il faut en fait rejeter H_0 quand $\bar{x}_n$ est *significativement plus petit* que m_0. Cela revient à prendre une région critique de la forme $W = \{(x_1,...,x_n)\,;\bar{x}_n < l_\alpha\}$, où $l_\alpha < m_0$.

La borne l_α dépend du seuil α que l'on s'est fixé. Moins on veut risquer de rejeter à tort H_0, plus α sera petit, et plus l_α sera petit. Le sens de l'expression *significativement plus petit* est lié à la valeur de α.

Un calcul analogue au précédent montre que :

$$\alpha = \underset{H_0}{Sup}\,\beta(m) = \underset{m \geq m_0}{Sup}\ P(\overline{X}_n < l_\alpha\,;m) = \underset{m \geq m_0}{Sup}\ \phi\left(\frac{l_\alpha - m}{\sigma}\sqrt{n}\right) = \phi\left(\frac{l_\alpha - m_0}{\sigma}\sqrt{n}\right)$$

On obtient donc $\frac{l_\alpha - m_0}{\sigma}\sqrt{n} = \phi^{-1}(\alpha)$, d'où $l_\alpha = m_0 + \frac{\sigma}{\sqrt{n}}\phi^{-1}(\alpha) = m_0 - \frac{\sigma}{\sqrt{n}}u_{2\alpha}$, avec les notations habituelles pour les quantiles de la loi normale.

En conclusion, on a :

> *<u>Propriété</u> : Un test de seuil α de H_0 : « $m \geq m_0$ » contre H_1 : « $m < m_0$ » est déterminé par la région critique* $W = \left\{(x_1,...,x_n)\,;\bar{x}_n < m_0 - \frac{\sigma}{\sqrt{n}}u_{2\alpha}\right\}$

4.2.2.4. Troisième idée

La région critique proposée ci-dessus pose un problème déjà rencontré à propos des intervalles de confiance : ce test est inutilisable si on ne connaît pas la vraie valeur de σ, ce qui est toujours le cas en pratique. Pour pallier cet inconvénient, on utilise la même procédure que pour les intervalles de

confiance : on remplace σ par son estimateur S'_n, ce qui nécessite de remplacer la loi normale par la loi de Student.

Rappelons en effet que si $X_1,...,X_n$ sont indépendantes et de même loi $N(m,\sigma^2)$, $\frac{\overline{X}_n - m}{S'_n}\sqrt{n}$ est de loi $St(n-1)$. Alors, à partir d'une région critique de la forme $W = \{(x_1,...,x_n)\,;\bar{x}_n < l_\alpha\}$, on obtient :

$$\alpha = \sup_{H_0} \beta(m) = \sup_{m \geq m_0} P(\overline{X}_n < l_\alpha\,;m) = \sup_{m \geq m_0} P\left(\frac{\overline{X}_n - m}{S'_n}\sqrt{n} < \frac{l_\alpha - m}{S'_n}\sqrt{n}\,;m\right)$$
$$= \sup_{m \geq m_0} F_{St(n-1)}\left(\frac{l_\alpha - m}{S'_n}\sqrt{n}\right) = F_{St(n-1)}\left(\frac{l_\alpha - m_0}{S'_n}\sqrt{n}\right)$$

D'où $\frac{l_\alpha - m_0}{S'_n}\sqrt{n} = F^{-1}_{St(n-1)}(\alpha) = -t_{n-1,2\alpha}$, avec les notations habituelles pour les quantiles de la loi de Student, et finalement $l_\alpha = m_0 - \frac{S'_n}{\sqrt{n}} t_{n-1,2\alpha}$.

En conclusion, on a :

<u>Propriété</u> : Un test de seuil α de H_0 : « $m \geq m_0$ » contre H_1 : « $m < m_0$ » est déterminé par la région critique $W = \left\{(x_1,...,x_n)\,;\bar{x}_n < m_0 - \frac{s'_n}{\sqrt{n}} t_{n-1,2\alpha}\right\}$

Remarque : La région critique peut aussi s'écrire $W = \left\{(x_1,...,x_n)\,;\frac{\bar{x}_n - m_0}{s'_n}\sqrt{n} < -t_{n-1,2\alpha}\right\}$.

4.2.2.5. Exemple

Avec le médicament A, la durée moyenne de disparition de la douleur était 30 mn. On a administré le médicament B à 12 malades et relevé les durées de disparition de la douleur suivants :

25 28 20 32 17 24 41 28 25 30 27 24

La moyenne empirique de ces données est $\bar{x}_n = 26.75$ et l'écart-type estimé est $s'_n = 6.08$.

On décide de ne commercialiser B que si on est sûr à 95% qu'il est plus efficace que A. Cela revient donc à faire un test de H_0 : « $m \geq 30$ » contre H_1 : « $m < 30$ » au seuil $\alpha = 5\%$.

On voit qu'il s'agit finalement de déterminer si 26.75 est suffisamment inférieur à 30 pour que l'on puisse conclure que le médicament B réduit vraiment la durée de disparition de la douleur.

D'après ce qui précède, on rejettera H_0 si $\frac{\bar{x}_n - m_0}{s'_n}\sqrt{n} < -t_{n-1,2\alpha}$.

Or $\frac{\bar{x}_n - m_0}{s'_n}\sqrt{n} = \frac{26.75 - 30}{6.08}\sqrt{12} = -1.853$ et $t_{n-1,2\alpha} = t_{11,0.1} = 1.796$.

-1.853 < -1.796 , donc les observations sont dans la région critique. On rejette donc H_0, ce qui signifie que l'on conclut que B est plus efficace que A, avec moins de 5% de chances de se tromper. Par conséquent, on peut lancer la commercialisation du médicament B.

4.2.2.6. Remarques

Remarque 1 : On voit ici le rôle fondamental du seuil α. Si on avait pris α = 1%, on aurait eu $t_{11,0.02} = 2.718$. Comme - 1.853 > - 2.718, on n'aurait pas rejeté H_0, donc on n'aurait pas adopté le médicament B.
Ce phénomène est normal : se fixer un seuil α petit revient à éviter au maximum d'adopter à tort le médicament B. Or un bon moyen de ne pas prendre ce risque, c'est de conserver le médicament A. Le test de seuil α = 0 consiste à conserver le médicament A quelles que soient les observations : la probabilité de rejeter à tort H_0 est nulle quand on ne rejette jamais H_0 ! En pratique, plus α est petit, moins on aura tendance à rejeter H_0.
Il est donc fondamental de bien savoir évaluer les risques et de choisir α en connaissance de cause.

Remarque 2 : La remarque précédente met en évidence l'existence d'un seuil critique α_c tel que pour tout seuil α supérieur à α_c, on rejettera H_0, et pour tout seuil α inférieur à α_c, on ne rejettera pas H_0. α_c vérifie $\frac{\bar{x}_n - m_0}{s'_n}\sqrt{n} = -t_{n-1,2\alpha_c}$. Sur l'exemple, la table de la loi de Student permet de constater que $-t_{11,0.05} = -2.201 < -1.853 < -1.796 = -t_{11,0.1}$. On en déduit que 5% < $2\alpha_c$ < 10%, d'où 2.5% < α_c < 5%. Cette valeur est appelée la **p-valeur**. C'est elle qui est calculée par les logiciels de statistique.

Sous S+, la commande permettant d'effectuer un test sur la moyenne d'une loi normale est `t.test`. L'option `alternative` permet de préciser lequel du test bilatéral et des deux tests unilatéraux on choisit. Sur l'exemple, on obtient :

```
> t.test(x,alternative="less",mu=30)

      One-sample t-Test

data:  x
t = -1.8526, df = 11, p-value = 0.0455
alternative hypothesis: true mean is less than 30
95 percent confidence interval:
       NA 29.90056
sample estimates:
 mean of x
     26.75
```

La p-valeur est ici α_c = 4.55%. Cela signifie que, pour tout seuil supérieur à 4.55% (c'est le cas de 5%), on rejettera H_0, donc on conclura que B est plus efficace que A, et pour tout seuil inférieur à 4.55% (c'est le cas de 1%), on ne rejettera pas H_0, donc on conclura que B n'est pas plus efficace que A.

Remarque 3 : Pour des raisons de symétrie, un test de « $m \le m_0$ » contre « $m > m_0$ » aura pour région critique $W = \left\{ (x_1,...,x_n) ; \frac{\bar{x}_n - m_0}{s'_n} \sqrt{n} > t_{n-1,2\alpha} \right\}$.

Remarque 4 : Pour le test bilatéral de H_0 : « $m = m_0$ » contre H_1 : « $m \neq m_0$ », le bon sens veut que l'on rejette H_0 si $\bar{x}_n$ est significativement éloigné de m_0. On prendra donc une région critique du type $W = \left\{ (x_1,...,x_n) ; |\bar{x}_n - m_0| > l_\alpha \right\}$. Alors, comme précédemment on obtient :

$$\alpha = \underset{m=m_0}{Sup}\ P(|\bar{X}_n - m_0| > l_\alpha ; m) = P(|\bar{X}_n - m_0| > l_\alpha ; m_0) = P\left(\frac{|\bar{X}_n - m_0|}{S'_n} \sqrt{n} > \frac{l_\alpha}{S'_n} \sqrt{n} ; m_0 \right)$$

On en déduit que $\frac{l_\alpha}{S'_n} \sqrt{n} = t_{n-1,\alpha}$, d'où $l_\alpha = \frac{S'_n}{\sqrt{n}} t_{n-1,\alpha}$. On obtient donc comme région critique :

$$W = \left\{ (x_1,...,x_n) ; |\bar{x}_n - m_0| > \frac{s'_n}{\sqrt{n}} t_{n-1,\alpha} \right\} = \left\{ (x_1,...,x_n) ; \left| \frac{\bar{x}_n - m_0}{s'_n} \right| \sqrt{n} > t_{n-1,\alpha} \right\}$$

Remarque 5 : Pour éviter d'alourdir les écritures, on écrit souvent une région critique en omettant l'expression $(x_1,...,x_n)$;, ce qui donne par exemple $W = \left\{ \frac{\bar{x}_n - m_0}{s'_n} \sqrt{n} < -t_{n-1,2\alpha} \right\}$. Mais il faut toujours garder à l'esprit que la région critique est l'ensemble des valeurs des observations pour lesquelles on rejettera H_0.

4.2.2.7. Le test de Student

Finalement, on dispose d'une procédure permettant d'effectuer le test bilatéral et les deux tests unilatéraux portant sur la moyenne de la loi normale. Ces trois tests sont connus sous le nom unique de **test de Student**.

Récapitulatif : Test de Student sur la moyenne d'une loi normale :

Test de « $m \le m_0$ » contre « $m > m_0$ » : $W = \left\{ \frac{\bar{x}_n - m_0}{s'_n} \sqrt{n} > t_{n-1,2\alpha} \right\}$

Test de « $m \ge m_0$ » contre « $m < m_0$ » : $W = \left\{ \frac{\bar{x}_n - m_0}{s'_n} \sqrt{n} < -t_{n-1,2\alpha} \right\}$

Test de « $m = m_0$ » contre « $m \neq m_0$ » : $W = \left\{ \left| \frac{\bar{x}_n - m_0}{s'_n} \right| \sqrt{n} > t_{n-1,\alpha} \right\}$

Remarque : Les tests ci-dessus ont été présentés comme des tests portant sur la valeur de la moyenne d'une loi normale. En fait, grâce au théorème central-limite, on sait que, quand n est assez grand, $\bar{X}_n$ est approximativement de loi normale, quelle que soit la loi de probabilité des observations.

Cette propriété permet de montrer qu'en pratique, pour $n \geq 30$, on pourra utiliser le test de Student pour faire un test sur la valeur de la moyenne de n'importe quelle loi de probabilité. On dit que le test de Student est robuste à la non-normalité.

4.2.3. Lien entre tests d'hypothèses et intervalles de confiance

Dans le test bilatéral, on rejette l'hypothèse « $m = m_0$ » à condition que $\left|\frac{\bar{x}_n - m_0}{s'_n}\right|\sqrt{n} > t_{n-1,\alpha}$. Or :

$$\left|\frac{\bar{x}_n - m_0}{s'_n}\right|\sqrt{n} > t_{n-1,\alpha} \quad \Leftrightarrow \quad \bar{x}_n - m_0 < -\frac{s'_n}{\sqrt{n}} t_{n-1,\alpha} \quad \text{ou} \quad \bar{x}_n - m_0 > \frac{s'_n}{\sqrt{n}} t_{n-1,\alpha}$$

$$\Leftrightarrow \quad m_0 < \bar{x}_n - \frac{s'_n}{\sqrt{n}} t_{n-1,\alpha} \quad \text{ou} \quad m_0 > \bar{x}_n + \frac{s'_n}{\sqrt{n}} t_{n-1,\alpha}$$

$$\Leftrightarrow \quad m_0 \notin \left[\bar{x}_n - \frac{s'_n}{\sqrt{n}} t_{n-1,\alpha}, \bar{x}_n + \frac{s'_n}{\sqrt{n}} t_{n-1,\alpha}\right]$$

Cet intervalle n'est autre que l'intervalle de confiance usuel pour la moyenne de la loi normale, vu en 3.4.2.1. Il y a donc un lien étroit entre les tests d'hypothèses et les intervalles de confiance.

C'est logique : on a une confiance $1-\alpha$ dans le fait que m appartient à l'intervalle de confiance. Si m_0 n'appartient pas à cet intervalle, il est vraiment douteux que $m = m_0$. On a même une confiance $1-\alpha$ dans le fait que $m \neq m_0$. On peut donc construire un test d'hypothèses sur la valeur d'un paramètre à partir d'un intervalle de confiance pour ce paramètre.

Or, pour construire un tel intervalle, on a eu besoin d'une fonction pivotale. Par conséquent, pour construire un test paramétrique, il suffit de connaître une fonction pivotale. Dans le cas de la moyenne de la loi normale, la fonction pivotale est $\frac{\bar{X}_n - m}{S'_n}\sqrt{n}$.

La dualité entre intervalles de confiance et tests d'hypothèses fait que, sous S+, la commande `t.test` permet à la fois d'effectuer un test et d'obtenir un intervalle de confiance sur la moyenne de la loi normale. Ainsi, la commande `t.test(x,conf.level=0.95)` effectue par défaut le test de « $m = 0$ » contre « $m \neq 0$ », et donne un intervalle de confiance pour m au seuil 5%.
Dans l'exemple des niveaux de bruit, on obtient :

```
> t.test(x,conf.level=0.95)

      One-sample t-Test

data:  x
t = 55.7889, df = 19, p-value = 0
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 61.82992 66.65008
sample estimates:
 mean of x
     64.24
```

L'intervalle de confiance de seuil 5% pour m est [61.82992, 66.65008], ce qui est bien le résultat déjà énoncé dans la section 3.4.2.1. Etant donné que 0 n'est pas, et de loin, dans cet intervalle, l'hypothèse « $m = 0$ » est très largement rejetée : la p-valeur vaut 0 (en fait un nombre extrêmement proche de 0).

4.2.4. Comment construire un test d'hypothèses

Finalement, le plus simple pour construire un test d'hypothèses portant sur la valeur d'un paramètre θ est de se fier à son bon sens. Si on connaît un estimateur $\hat{\theta}_n$ de θ, on procèdera de la façon suivante :

- Test de « $\theta \le \theta_0$ » contre « $\theta > \theta_0$ » : on rejette H_0 si $\hat{\theta}_n$ est « trop grand ». $W = \{\hat{\theta}_n > l_\alpha\}$.
- Test de « $\theta \ge \theta_0$ » contre « $\theta < \theta_0$ » : on rejette H_0 si $\hat{\theta}_n$ est « trop petit ». $W = \{\hat{\theta}_n < l_\alpha\}$.
- Test de « $\theta = \theta_0$ » contre « $\theta \ne \theta_0$ » : on rejette H_0 si $|\hat{\theta}_n - \theta_0|$ est « trop grand » ou bien si $\hat{\theta}_n$ est « soit trop grand soit trop petit ». $W = \{|\hat{\theta}_n - \theta_0| > l_\alpha\} = \{\hat{\theta}_n < l_{1,\alpha} \text{ ou } \hat{\theta}_n > l_{2,\alpha}\}$.

Pour déterminer $l_\alpha, l_{1,\alpha}, l_{2,\alpha}$, il faut écrire $\alpha = \underset{H_0}{Sup}\, P((X_1,...,X_n) \in W ; \theta)$. Par exemple, dans le premier cas, $\alpha = \underset{\theta \le \theta_0}{Sup}\, P(\hat{\theta}_n > l_\alpha)$. Pour pouvoir calculer $P(\hat{\theta}_n > l_\alpha)$, il faut utiliser une fonction pivotale.

Malheureusement, cette procédure de bon sens ne permet pas toujours de résoudre le problème. C'est le cas par exemple quand la loi de probabilité de $\hat{\theta}_n$ sous H_0 est complexe et qu'on ne peut pas trouver de fonction pivotale. D'autre part, le test obtenu par cette approche n'est pas forcément optimal, au sens où il peut en exister de plus puissants.

Il existe en fait des méthodes statistiques sophistiquées permettant de répondre à ces deux problèmes. Le résultat le plus important est le théorème de Neyman-Pearson. Mais ces procédures débordent du cadre de ce cours et ne seront pas évoquées ici.

4.2.5. Tests sur la variance d'une loi normale

On suppose ici que les observations $x_1,...,x_n$ sont les réalisations de variables aléatoires $X_1,...,X_n$ indépendantes et de même loi normale $N(m,\sigma^2)$. On souhaite tester par exemple H_0 : « $\sigma^2 \le \sigma_0^2$ » contre H_1 : « $\sigma^2 > \sigma_0^2$ ».

Puisque l'ESBVM de σ^2 est S'^2_n, il est naturel de rejeter H_0 si S'^2_n est « trop grand », donc de considérer une région critique de la forme $W = \{s'^2_n > l_\alpha\}$. Pour calculer $\alpha = \underset{H_0}{Sup}\, P(S'^2_n > l_\alpha)$, on utilise la fonction pivotale $\frac{(n-1)S'^2_n}{\sigma^2}$, qui est de loi χ^2_{n-1}. On obtient :

$$\alpha = \sup_{\sigma^2 \le \sigma_0^2} P(S'^2_n > l_\alpha) = \sup_{\sigma^2 \le \sigma_0^2} P\left(\frac{(n-1)S'^2_n}{\sigma^2} > \frac{(n-1)l_\alpha}{\sigma^2} \right)$$
$$= \sup_{\sigma^2 \le \sigma_0^2} \left[1 - F_{\chi^2_{n-1}}\left(\frac{(n-1)l_\alpha}{\sigma^2} \right) \right] = 1 - F_{\chi^2_{n-1}}\left(\frac{(n-1)l_\alpha}{\sigma_0^2} \right)$$

D'où $l_\alpha = \frac{\sigma_0^2}{n-1} F^{-1}_{\chi^2_{n-1}}(1-\alpha) = \frac{\sigma_0^2}{n-1} z_{n-1,\alpha}$, et la région critique du test est $W = \left\{ s'^2_n > \frac{\sigma_0^2}{n-1} z_{n-1,\alpha} \right\}$ ou $W = \left\{ \frac{(n-1)s'^2_n}{\sigma_0^2} > z_{n-1,\alpha} \right\}$.

On aboutirait au même résultat en partant d'un intervalle de confiance de seuil α pour σ^2 du type $[0, a]$.

Exercice : construire le deuxième test unilatéral et le test bilatéral.

Finalement, on obtient :

Propriété : Tests sur la variance d'une loi normale :

Test de « $\sigma^2 \le \sigma_0^2$ » *contre* « $\sigma^2 > \sigma_0^2$ » : $W = \left\{ \frac{(n-1)s'^2_n}{\sigma_0^2} > z_{n-1,\alpha} \right\}$

Test de « $\sigma^2 \ge \sigma_0^2$ » *contre* « $\sigma^2 < \sigma_0^2$ » : $W = \left\{ \frac{(n-1)s'^2_n}{\sigma_0^2} < z_{n-1,1-\alpha} \right\}$

Test de « $\sigma^2 = \sigma_0^2$ » *contre* « $\sigma^2 \ne \sigma_0^2$ » :

$W = \left\{ \frac{(n-1)s'^2_n}{\sigma_0^2} < z_{n-1,1-\alpha/2} \text{ ou } \frac{(n-1)s'^2_n}{\sigma_0^2} > z_{n-1,\alpha/2} \right\}$

Remarque : Contrairement à ce qui se passait pour la moyenne, ces tests ne sont pas généralisables à des tests sur la variance d'une loi non normale, car on n'a pas l'équivalent du théorème central-limite pour S'^2_n.

Dans l'exemple de l'essai thérapeutique, la variance mesure la variabilité de l'effet du médicament. La variabilité est faible si l'effet du médicament est à peu près le même pour tout le monde, et elle est forte si les effets peuvent être très différents d'un individu à un autre. On a évidemment intérêt à avoir une variabilité assez faible pour bien contrôler les effets d'un traitement. Cette variabilité se traduit sur la variance de la loi normale qui modélise le temps de disparition de la douleur chez les malades traités.

Avec le médicament A, l'écart-type était $\sigma_0 = 5$ mn, ce qui signifie que, pour 95% des malades, la douleur disparaît entre $m_0 - 2\sigma_0 = 20$ mn et $m_0 + 2\sigma_0 = 40$ mn. Avec le médicament B, on estime σ par s'_n = 6.08 mn. La variabilité du second médicament est-elle significativement supérieure à celle du premier ?

C'est un test de « $\sigma \leq 5$ » contre « $\sigma > 5$ » , évidemment identique au test de « $\sigma^2 \leq 25$ » contre « $\sigma^2 > 25$ ». La région critique est $W = \left\{ \frac{(n-1)s'^2_n}{\sigma_0^2} > z_{n-1,\alpha} \right\}$.

Au seuil α = 5%, on a $z_{11,5\%} = 19.7$. $\frac{(n-1)s'^2_n}{\sigma_0^2} = \frac{11 \times 6.08^2}{25} = 16.3$.

Comme 16.3 < 19.7, on n'est pas dans la région critique, donc on ne rejette pas H_0 : on n'a pas de preuves suffisantes pour conclure que la variabilité de l'effet de B est supérieure à celle de A. La différence entre 6.08 et 5 n'est pas significative au seuil choisi.

Exercice : Construire les trois tests usuels portant sur le paramètre de la loi exponentielle.

4.2.6. Tests sur une proportion

On désire faire des tests sur la probabilité $p = P(A)$ qu'un événement A se produise au vu du nombre x de fois où A s'est produit au cours d'une série de n expériences identiques et indépendantes. On a déjà vu en 3.4.3. que x est la réalisation d'une variable aléatoire X de loi binomiale $B(n, p)$ et que l'ESBVM de p est $\hat{p} = \frac{X}{n}$.

Pour construire des tests, on peut partir de l'intervalle de confiance exact vu en 3.4.3.2. Mais compte-tenu de sa complexité, on se contentera de l'intervalle de confiance asymptotique, basé sur l'approximation de la loi binomiale $B(n, p)$ par la loi normale $N(np, np(1-p))$. $\frac{X - np}{\sqrt{np(1-p)}}$ est approximativement de loi $N(0,1)$, ce qui fournit la fonction pivotale cherchée et permet de donner directement les tests sur une proportion :

Propriété : Tests asymptotiques sur une proportion :

Test de « $p \leq p_0$ » contre « $p > p_0$ » : $W = \left\{ \frac{x - np_0}{\sqrt{np_0(1-p_0)}} > u_{2\alpha} \right\}$

Test de « $p \geq p_0$ » contre « $p < p_0$ » : $W = \left\{ \frac{x - np_0}{\sqrt{np_0(1-p_0)}} < -u_{2\alpha} \right\}$

$$\textit{Test de « } p = p_0 \textit{ » contre « } p \neq p_0 \textit{ » : } W = \left\{ \left| \frac{x - np_0}{\sqrt{np_0(1-p_0)}} \right| > u_\alpha \right\}$$

Dans l'exemple du sondage de la section 3.4.3., on a interrogé n = 800 personnes et x = 420 d'entre elles ont déclaré vouloir voter pour A. On a donc estimé le pourcentage p de voix qu'obtiendra le candidat A par $\hat{p} = \frac{420}{800} = 52.5\%$. Mais on a vu qu'un intervalle de confiance de seuil 5% pour ce pourcentage est [49%, 56%], dont une partie est située sous les 50%.

En fait, la seule chose qui intéresse le candidat A, c'est de savoir s'il va être élu ou pas. Il s'agit donc de faire un test dans lequel le rejet de H_0 correspond à l'élection de A. Par conséquent, on va tester « $p \leq 1/2$ » contre « $p > 1/2$ ».

$\frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{420 - 800/2}{\sqrt{800/4}} = 1.414$. Au seuil 5%, $u_{0.1} = 1.645$.

1.414 < 1.645, donc on n'est pas dans la région critique, donc on ne rejette pas H_0 : on ne peut pas affirmer que A sera élu avec moins de 5% de chances de se tromper.

La p-valeur du test est la valeur α_c de α telle que $u_{2\alpha_c} = \phi^{-1}(1-\alpha_c) = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = 1.414$. On a donc $\alpha_c = 1 - \phi(1.414) = 7.86\%$.

Sous S+, on peut effectuer le test exact grâce à la commande `binom.test`. On obtient sur l'exemple du sondage :

```
> binom.test(420,800,p=0.5,alternative="greater")

      Exact binomial test

data:  420 out of 800
number of successes = 420, n = 800, p-value = 0.0839
alternative hypothesis: true p is greater than 0.5
```

La p-valeur est 8.39 %, ce qui est bien cohérent avec la valeur donnée par le test asymptotique.

En conclusion, si on décide de conclure, au vu du sondage, que le candidat A sera élu, on a environ 8% de chances de se tromper. Tout ce qui vient d'être dit n'est évidemment valable que si les résultats du sondage sont bien représentatifs de ce qui se passera le jour de l'élection, ce qui est loin d'être certain.

4.3. Tests paramétriques sur deux échantillons

Dans l'exemple de l'essai thérapeutique, nous avons supposé que la durée de disparition de la douleur avec le médicament A était de loi normale $N(m_0, \sigma_0^2)$, où m_0 et σ_0 étaient connus. En réalité, m_0 et σ_0 ne sont pas connus mais estimés à partir d'observations faites sur des malades traités avec le médicament A. Les données sont donc en fait constituées de deux échantillons correspondant aux deux médicaments.
Si les traitements ont été appliqués sur deux groupes de personnes différentes, on peut raisonnablement considérer que les échantillons sont indépendants. Mais il est possible que l'on donne successivement les deux médicaments aux mêmes malades, pour déterminer lequel est le plus efficace. La premier cas correspond à un **test sur deux échantillons indépendants** et le second à un **test sur deux échantillons appariés**.

4.3.1. Comparaison de deux échantillons gaussiens indépendants

Il est très fréquent que l'on ait à comparer deux populations selon un critère quantitatif particulier. Par exemple :

- performances de deux produits concurrents
- occurrences de maladies chez les fumeurs et les non-fumeurs
- résultats scolaires des filles et des garçons

Statistiquement, cela signifie que l'on dispose d'observations de variables aléatoires $X_1, ..., X_{n_1}$ indépendantes et de même loi constituant le premier échantillon, et de variables aléatoires $Y_1, ..., Y_{n_2}$ indépendantes et de même loi constituant le deuxième échantillon, les X_i et les Y_j étant indépendants.

Un problème important est de déterminer si les deux échantillons sont issus de la même loi de probabilité. Ce problème ne peut se traiter que de façon non-paramétrique, ce qui sera fait en 4.4.2.

Dans cette section, on supposera que les deux échantillons sont de loi normale et on comparera leurs moyennes et leur variances.
$X_1, ..., X_{n_1}$ sont supposées de loi $N(m_1, \sigma_1^2)$ et $Y_1, ..., Y_{n_2}$ de loi $N(m_2, \sigma_2^2)$.
Les moyennes empiriques, variances empiriques et variances estimées des deux échantillons sont notées respectivement $\overline{X}_{n_1}$, S_1^2, $S_1'^2$, $\overline{Y}_{n_2}$, S_2^2 et $S_2'^2$.

Exemple : deux groupes d'étudiants de tailles respectives $n_1 = 25$ et $n_2 = 31$ ont suivi le même cours de statistique et passé le même examen. Les moyennes et écarts-types empiriques des notes obtenues dans les deux groupes sont respectivement :

Premier groupe : $\bar{x}_{n_1} = 12.8$, $s_1' = 3.4$.
Deuxième groupe : $\bar{y}_{n_2} = 11.3$, $s_2' = 2.9$.

On suppose que les notes sont réparties dans les deux groupes selon des lois normales et qu'elles sont toutes indépendantes.

Peut-on considérer que le premier groupe est meilleur que le deuxième, c'est-à-dire qu'un point et demi d'écart entre les moyennes est significatif d'une différence de niveau ?

La procédure à suivre consiste à tester d'abord l'égalité des variances, puis l'égalité des moyennes.

4.3.1.1. Test de Fisher de comparaison des variances

Comparer les variances des deux échantillons, c'est tester H_0 : « $\sigma_1^2 = \sigma_2^2$ » contre H_1 : « $\sigma_1^2 \neq \sigma_2^2$ ».

Il est naturel de rejeter l'hypothèse d'égalité des variances si les variances empiriques ou estimées des deux échantillons sont significativement différentes. On peut penser à une région critique de la forme $W = \left\{ \left| s_1'^2 - s_2'^2 \right| > l_\alpha \right\}$, mais la loi de probabilité de $S_1'^2 - S_2'^2$ s'avère complexe.

En revanche, celle de $\dfrac{S_1'^2}{S_2'^2}$ est simple. On utilisera donc plutôt une région critique de la forme $W = \left\{ \dfrac{s_1'^2}{s_2'^2} < l_{1,\alpha} \text{ ou } \dfrac{s_1'^2}{s_2'^2} > l_{2,\alpha} \right\}$, avec $l_{1,\alpha} < l_{2,\alpha}$: on rejettera l'égalité des variances si le rapport des deux variances estimées est soit « trop grand » soit « trop petit ».

D'après le théorème de Fisher, $\dfrac{n_1 S_1^2}{\sigma_1^2} = \dfrac{(n_1 - 1) S_1'^2}{\sigma_1^2}$ est de loi $\chi^2_{n_1 - 1}$ et $\dfrac{n_2 S_2^2}{\sigma_2^2} = \dfrac{(n_2 - 1) S_2'^2}{\sigma_2^2}$ est de loi $\chi^2_{n_2 - 1}$, ces deux variables aléatoires étant indépendantes. Or si X est de loi χ^2_n, Y est de loi χ^2_m, et X et Y sont indépendantes, alors $\dfrac{mX}{nY}$ est de loi de Fisher-Snedecor $F(n, m)$.

Par conséquent, $\dfrac{(n_2 - 1)\dfrac{(n_1 - 1) S_1'^2}{\sigma_1^2}}{(n_1 - 1)\dfrac{(n_2 - 1) S_2'^2}{\sigma_2^2}} = \dfrac{S_1'^2 \sigma_2^2}{S_2'^2 \sigma_1^2}$ est de loi $F(n_1 - 1, n_2 - 1)$.

Sous l'hypothèse H_0, $\sigma_1^2 = \sigma_2^2$ donc $\dfrac{S_1'^2}{S_2'^2}$ est de loi $F(n_1 - 1, n_2 - 1)$.

Le seuil du test est donc :

$$\begin{aligned} \alpha &= P_{H_0}\left(\frac{S_1'^2}{S_2'^2} < l_{1,\alpha} \text{ ou } \frac{S_1'^2}{S_2'^2} > l_{2,\alpha} \right) = P_{H_0}\left(\frac{S_1'^2}{S_2'^2} < l_{1,\alpha} \right) + P_{H_0}\left(\frac{S_1'^2}{S_2'^2} > l_{2,\alpha} \right) \\ &= F_{F(n_1 - 1, n_2 - 1)}(l_{1,\alpha}) + 1 - F_{F(n_1 - 1, n_2 - 1)}(l_{2,\alpha}) \end{aligned}$$

En équilibrant les risques, on choisira $l_{1,\alpha}$ et $l_{2,\alpha}$ de sorte que $F_{F(n_1-1,n_2-1)}(l_{1,\alpha}) = \frac{\alpha}{2}$ et $F_{F(n_1-1,n_2-1)}(l_{2,\alpha}) = 1-\frac{\alpha}{2}$, c'est à dire $l_{1,\alpha} = f_{n_1-1,n_2-1,1-\alpha/2}$ et $l_{2,\alpha} = f_{n_1-1,n_2-1,\alpha/2}$.

La région critique du test s'écrit donc $W = \left\{ \frac{s_1'^2}{s_2'^2} < f_{n_1-1,n_2-1,1-\alpha/2} \text{ ou } \frac{s_1'^2}{s_2'^2} > f_{n_1-1,n_2-1,\alpha/2} \right\}$. On peut simplifier les choses en remarquant que :

1. $f_{n_1-1,n_2-1,1-\alpha/2} = \frac{1}{f_{n_2-1,n_1-1,\alpha/2}}$, donc $W = \left\{ \frac{s_2'^2}{s_1'^2} > f_{n_2-1,n_1-1,\alpha/2} \text{ ou } \frac{s_1'^2}{s_2'^2} > f_{n_1-1,n_2-1,\alpha/2} \right\}$
2. Des deux rapports $\frac{s_1'^2}{s_2'^2}$ et $\frac{s_2'^2}{s_1'^2}$, un seul est plus grand que 1. Or on peut montrer que pour $\alpha < 1/2$, $f_{n,m,\alpha} > 1$. Donc, dans la région critique, il suffit de retenir celui des deux rapports qui est supérieur à 1.

Par conséquent, la région critique du test peut s'écrire simplement sous la forme ci-dessous. Ce test est appelé **test de Fisher**.

Propriété : Test de Fisher d'égalité des variances de deux échantillons gaussiens indépendants:

Test de « $\sigma_1^2 = \sigma_2^2$ » contre « $\sigma_1^2 \neq \sigma_2^2$ » :

- *si* $s_1'^2 > s_2'^2$, $W = \left\{ \frac{s_1'^2}{s_2'^2} > f_{n_1-1,n_2-1,\alpha/2} \right\}$
- *si* $s_1'^2 < s_2'^2$, $W = \left\{ \frac{s_2'^2}{s_1'^2} > f_{n_2-1,n_1-1,\alpha/2} \right\}$

Remarque : Le fait que $\frac{S_1'^2\sigma_2^2}{S_2'^2\sigma_1^2}$ soit de loi $F(n_1-1, n_2-1)$ permet d'obtenir facilement un intervalle de confiance pour le rapport $\frac{\sigma_1^2}{\sigma_2^2}$: $\left[\frac{s_1'^2}{s_2'^2} f_{n_2-1,n_1-1,1-\alpha/2}, \frac{s_1'^2}{s_2'^2} f_{n_2-1,n_1-1,\alpha/2} \right]$.

Dans l'exemple, $s_1'^2 > s_2'^2$ et $\frac{s_1'^2}{s_2'^2} = 1.37$. La table de la loi de Fisher ne fournit des quantiles que pour $\alpha/2$ = 5% ou 1%. On choisit donc de faire le test de Fisher au seuil α = 10%. Alors $f_{24,30,0.05}$ = 1.89.

1.37 < 1.89, donc on n'est pas dans la région critique. On ne peut donc pas conclure que les variances des deux échantillons sont différentes.

Sous S+, la commande permettant d'effectuer un test de Fisher est `var.test`. L'option `conf.level` précise le seuil de l'intervalle de confiance pour le rapport des variances.

```
> var.test(x, y, alternative="two.sided", conf.level=.95)

      F test for variance equality

data:  x and y
F = 1.3746, num df = 24, denom df = 30, p-value = 0.4058
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6435544 3.0363564
sample estimates:
 variance of x variance of y
         11.56          8.41
```

La p-valeur vaut 40.58 %. Cela signifie que, même en prenant un risque d'erreur très grand comme 40%, on ne rejettera pas l'hypothèse d'égalité des variances. Par conséquent, on est très loin de rejeter cette hypothèse. On constate par ailleurs que l'intervalle de confiance de seuil 5% pour le rapport des deux variances est [0.644, 3.036], qui contient bien la valeur 1.

Remarque : Le test de Fisher peut se généraliser à la comparaison des variances de k échantillons gaussiens indépendants, de tailles respectives $n_1, n_2, ..., n_k$.

Soit $n = \sum_{i=1}^{k} n_i$. Le **test de Bartlett** est basé sur le fait que, sous l'hypothèse « $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ », la variable aleatoire $(n-k)\ln\left(\frac{1}{n-k}\sum_{i=1}^{k}(n_i-1)S_i'^2\right) - \sum_{i=1}^{k}(n_i-1)\ln S_i'^2$ est approximativement de loi χ^2_{k-1}.

Exercice : Construire les tests de « $\sigma_1^2 \le \sigma_2^2$ » contre « $\sigma_1^2 > \sigma_2^2$ » et « $\sigma_1^2 \ge \sigma_2^2$ » contre « $\sigma_1^2 < \sigma_2^2$ ».

4.3.1.2. Test de Student de comparaison des moyennes

On veut tester H_0 : « $m_1 = m_2$ » contre H_1 : « $m_1 \neq m_2$ ».

L'idée naturelle est de rejeter « $m_1 = m_2$ » quand la différence entre les moyennes empiriques des deux échantillons est trop grande, d'où une région critique de la forme $W = \left\{ \left| \bar{x}_{n_1} - \bar{y}_{n_2} \right| > l_\alpha \right\}$.
Pour déterminer l_α, on a besoin de la loi de probabilité de $\overline{X}_{n_1} - \overline{Y}_{n_2}$ sous H_0. Or on sait que $\overline{X}_{n_1}$ est de loi $N(m_1, \frac{\sigma_1^2}{n_1})$ et $\overline{Y}_{n_2}$ est de loi $N(m_2, \frac{\sigma_2^2}{n_2})$. Ces deux variables aléatoires étant indépendantes, on en déduit que $\overline{X}_{n_1} - \overline{Y}_{n_2}$ est de loi $N(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$.

Donc finalement, la variable aléatoire $U = \dfrac{\left(\overline{X}_{n_1} - \overline{Y}_{n_2}\right) - (m_1 - m_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ est de loi $N(0,1)$ et, sous H_0, $\dfrac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ est de loi $N(0,1)$.

σ_1^2 et σ_2^2 étant inconnues, on ne peut pas utiliser directement cette variable aléatoire pour construire le test. On va alors construire l'équivalent d'un test de Student.

Pour cela, on pose $Z = \dfrac{(n_1 - 1)S_1'^2}{\sigma_1^2} + \dfrac{(n_2 - 1)S_2'^2}{\sigma_2^2}$. Etant donné que $\dfrac{(n_1 - 1)S_1'^2}{\sigma_1^2}$ est de loi $\chi^2_{n_1 - 1}$, $\dfrac{(n_2 - 1)S_2'^2}{\sigma_2^2}$ est de loi $\chi^2_{n_2 - 1}$ et que ces deux variables aléatoires sont indépendantes, Z est de loi $\chi^2_{n_1 + n_2 - 2}$. Le théorème de Fisher permet d'établir que U et Z sont indépendants.
Par conséquent, par définition de la loi de Student, la variable aléatoire

$$\frac{U}{\sqrt{Z}}\sqrt{n_1 + n_2 - 2} = \frac{\left(\overline{X}_{n_1} - \overline{Y}_{n_2}\right) - (m_1 - m_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}\sqrt{\dfrac{(n_1 - 1)S_1'^2}{\sigma_1^2} + \dfrac{(n_2 - 1)S_2'^2}{\sigma_2^2}}}\sqrt{n_1 + n_2 - 2}$$

est de loi de $St(n_1 + n_2 - 2)$.

Dans cette expression, les paramètres inconnus σ_1^2 et σ_2^2 subsistent. Mais on remarque que, sous l'hypothèse « $\sigma_1^2 = \sigma_2^2$ », ils disparaissent. Pour savoir si cette hypothèse est valide, il suffit d'appliquer le test de Fisher vu précédemment.

Par conséquent, la démarche à suivre consiste à tester d'abord l'égalité des variances. Si le test de Fisher ne rejette pas l'égalité des variances, on considèrera que $\sigma_1^2 = \sigma_2^2$. Alors, la variable aléatoire $\dfrac{\left(\overline{X}_{n_1} - \overline{Y}_{n_2}\right) - (m_1 - m_2)}{\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}\sqrt{(n_1 - 1)S_1'^2 + (n_2 - 1)S_2'^2}}\sqrt{n_1 + n_2 - 2}$ est de loi $St(n_1 + n_2 - 2)$, et, sous l'hypothèse « $m_1 = m_2$ », $T = (\overline{X}_{n_1} - \overline{Y}_{n_2})\sqrt{\dfrac{n_1 n_2 (n_1 + n_2 - 2)}{(n_1 + n_2)[(n_1 - 1)S_1'^2 + (n_2 - 1)S_2'^2]}}$ est de loi $St(n_1 + n_2 - 2)$, ce qui fournit la fonction pivotale cherchée.

<u>Propriété</u> : Test de Student d'égalité des moyennes de deux échantillons gaussiens indépendants de même variance

Test de « $m_1 = m_2$ » contre « $m_1 \neq m_2$ » :

$$W = \left\{ \left| \bar{x}_{n_1} - \bar{y}_{n_2} \right| \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{(n_1 + n_2)[(n_1 - 1)s_1'^2 + (n_2 - 1)s_2'^2]}} > t_{n_1+n_2-2,\alpha} \right\}$$

Remarque 1 : Dans cette approche, on commet une faute de raisonnement. En effet, si le test de Fisher ne rejette pas l'égalité des variances, on peut en conclure qu'on n'a pas de preuves suffisantes pour considérer que les variances sont différentes, mais on ne peut pas pour autant considérer qu'elles sont égales : c'est un exemple de la différence entre *ne pas rejeter* H_0 et *accepter* H_0. Pour bien faire, il faudrait pouvoir tester « $\sigma_1^2 \neq \sigma_2^2$ » contre « $\sigma_1^2 = \sigma_2^2$ ». Mais c'est impossible car l'hypothèse nulle est trop vaste pour que l'on puisse calculer le seuil d'un tel test. On est donc contraints d'adopter la démarche présentée ici. Le résultat ne sera alors qu'approximatif.

Remarque 2 : A partir du test, on peut facilement construire un intervalle de confiance pour la différence des moyennes $m_1 - m_2$:

$$\left[\bar{X}_{n_1} - \bar{Y}_{n_2} - t_{n_1+n_2-2,\alpha} \sqrt{\frac{(n_1 + n_2)[(n_1 - 1)S_1'^2 + (n_2 - 1)S_2'^2]}{n_1 n_2 (n_1 + n_2 - 2)}}, \bar{X}_{n_1} - \bar{Y}_{n_2} + t_{n_1+n_2-2,\alpha} \sqrt{\frac{(n_1 + n_2)[(n_1 - 1)S_1'^2 + (n_2 - 1)S_2'^2]}{n_1 n_2 (n_1 + n_2 - 2)}} \right]$$

Remarque 3 : A priori, si le test de Fisher rejette l'égalité des variances, on ne peut pas appliquer le test. En fait, le théorème central-limite permet de montrer que, si n_1 et n_2 sont suffisamment grands (supérieurs à 30), alors la loi de T est approximativement la loi $N(0,1)$ même si les deux variances sont différentes et en fait même si les deux échantillons ne sont pas de loi normale.

Par conséquent, *si on a beaucoup d'observations, on peut comparer les moyennes d'échantillons issus de n'importe quelle loi de probabilité*. En revanche, si on a peu d'observations, ce test ne fonctionne pas. On utilise alors d'autres tests comme le test de Smith-Satterthwaite ou le test d'Aspin-Welch.

Remarque 4 : La généralisation de ce problème à la comparaison des moyennes de k échantillons gaussiens fait l'objet d'un domaine important de la statistique appelé l'**analyse de variance**.

Exercice : Construire les tests de « $m_1 \leq m_2$ » contre « $m_1 > m_2$ » et « $m_1 \geq m_2$ » contre « $m_1 < m_2$ ».

Dans l'exemple, on n'a pas rejeté l'égalité des variances, donc on peut appliquer le test de Student. Comme il s'agit de déterminer si le premier groupe est meilleur que le deuxième et que cette hypothèse doit correspondre au rejet de H_0, on voit qu'il s'agit ici de tester « $m_1 \leq m_2$ » contre « $m_1 > m_2$ ».

La région critique est $W = \left\{ (\bar{x}_{n_1} - \bar{y}_{n_2}) \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{(n_1 + n_2)[(n_1 - 1)s_1'^2 + (n_2 - 1)s_2'^2]}} > t_{n_1+n_2-2,2\alpha} \right\}$.

Ici, $t = (\bar{x}_{n_1} - \bar{y}_{n_2}) \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{(n_1 + n_2)[(n_1 - 1)s_1'^2 + (n_2 - 1)s_2'^2]}} = 1.78$.

Pour un seuil de 5%, on a $t_{25+31-2,0.1} = t_{54,0.1} \approx 1.68$.

1.78 > 1.68, donc on est dans la région critique, donc on rejette H_0. On conclut que la différence de moyenne entre les deux groupes d'étudiants est significative au seuil 5%.

Sous S+, la commande `t.test` déjà vue pour effectuer des tests sur la moyenne d'un échantillon gaussien, permet également de comparer les moyennes de deux échantillons gaussiens indépendants :

```
> t.test(x,y,alternative="greater",conf.level=0.95)

      Standard Two-Sample t-Test

data:  x and y
t = 1.7816, df = 54, p-value = 0.0402
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.09097004         NA
sample estimates:
 mean of x mean of y
      12.8      11.3
```

On retrouve que $t = 1.7816$. La p-valeur du test est 4.02%. Donc au seuil 5%, on rejettera bien H_0, par contre on ne la rejettera pas au seuil 1%.

4.3.2. Comparaison de deux proportions

Le problème se pose quand on veut comparer deux populations selon un critère qui est une proportion. Par exemple :

- comparer les performances de deux machines au vu de la proportion de pièces défectueuses qu'elles produisent
- comparer les fréquences d'occurrences de cancers selon que l'on habite ou pas à proximité d'une centrale nucléaire

Mathématiquement, on a une première population de taille n_1 et une seconde de taille n_2. On note X_1 et X_2 les nombres d'individus dans chaque population présentant une certaine caractéristique (pièce défectueuse, habitant malade), et p_1 et p_2 les probabilités qu'un individu de chaque population présente cette caractéristique. On souhaite comparer p_1 et p_2, c'est-à-dire effectuer des tests du type « $p_1 \leq p_2$ » contre « $p_1 > p_2$ » ou « $p_1 = p_2$ » contre « $p_1 \neq p_2$ ».

Exemple : La machine 1 a produit 96 pièces dont 12 défectueuses. La machine 2 a produit 55 pièces dont 10 défectueuses. Les pourcentages de pièces défectueuses produites par ces machines sont respectivement $\frac{12}{96} = 12.5\%$ et $\frac{10}{55} = 18.2\%$. Peut-on en conclure que la machine 1 est significativement plus performante que la machine 2 ?

Si les occurrences des évènements qui nous intéressent sur chaque individu sont indépendantes, les variables aléatoires X_1 et X_2 sont de lois binomiales, respectivement $B(n_1, p_1)$ et $B(n_2, p_2)$. On

se contentera ici de supposer que les tailles d'échantillons sont suffisamment grandes pour que l'on puisse faire l'approximation de la loi binomiale par la loi normale : $n_1 p_1 > 5$, $n_1(1-p_1) > 5$, $n_2 p_2 > 5$, et $n_2(1-p_2) > 5$. Alors on peut considérer que X_1 et X_2 sont des variables aléatoires indépendantes et approximativement de lois normales, respectivement $N(n_1 p_1, n_1 p_1(1-p_1))$ et $N(n_2 p_2, n_2 p_2(1-p_2))$.

Les ESBVM de p_1 et p_2 sont $\frac{X_1}{n_1}$ et $\frac{X_2}{n_2}$. Si on veut tester H_0 : « $p_1 = p_2$ » contre H_1 : « $p_1 \neq p_2$ », il est logique de rejeter H_0 si $\left|\frac{X_1}{n_1} - \frac{X_2}{n_2}\right|$ est « trop grand », donc de choisir une région critique de la forme $W = \left\{ \left|\frac{x_1}{n_1} - \frac{x_2}{n_2}\right| > l_\alpha \right\}$.

$\frac{X_1}{n_1}$ et $\frac{X_2}{n_2}$ sont indépendantes et de lois respectives $N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right)$ et $N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$, donc $\frac{X_1}{n_1} - \frac{X_2}{n_2}$ est de loi $N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$. Sous H_0 : « $p_1 = p_2$ », posons $p = p_1 = p_2$. Alors $\frac{X_1}{n_1} - \frac{X_2}{n_2}$ est de loi $N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$ et $\frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ est de loi $N(0,1)$.

Comme p est inconnu, cette variable aléatoire ne peut pas servir de fonction pivotale. Mais, comme les tailles d'échantillon sont grandes, on peut montrer que le résultat reste approximativement vrai quand on remplace p par son estimateur $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$.

Donc finalement, sous H_0, la variable aléatoire $U = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\frac{X_1 + X_2}{n_1 + n_2}\left(1 - \frac{X_1 + X_2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ est approximativement de loi $N(0,1)$, ce qui permet de construire le test.

Propriété : Test de comparaison de deux proportions :

Test de « $p_1 = p_2$ » contre « $p_1 \neq p_2$ » : $W = \{|u| > u_\alpha\}$

Test de « $p_1 \leq p_2$ » contre « $p_1 > p_2$ » : $W = \{u > u_{2\alpha}\}$

Test de « $p_1 \geq p_2$ » contre « $p_1 < p_2$ » : $W = \{u < -u_{2\alpha}\}$

Dans l'exemple, il s'agit de tester « $p_1 \geq p_2$ » contre « $p_1 < p_2$ », avec $\frac{x_1}{n_1} = 12.5\%$ et $\frac{x_2}{n_2} = 18.2\%$. On trouve $\frac{x_1 + x_2}{n_1 + n_2} = \frac{22}{151} = 14.6\%$, d'où $u = \frac{0.125 - 0.182}{\sqrt{0.146(1-0.146)\left(\frac{1}{96} + \frac{1}{55}\right)}} = -0.95$. Au seuil 5% on a $u_{2\alpha} = 1.645$.

-0.95 > -1.645, donc on ne rejette pas H_0 : la différence entre les deux proportions de pièces défectueuses n'est pas significative au seuil 5%.

Sous S+, le test s'effectue à l'aide de la commande `prop.test` et fournit en même temps un intervalle de confiance pour $p_1 - p_2$.

```
> prop.test(c(12,10),c(96,55),alternative="less",conf.level=0.95,correct=F)

2-sample test for equality of proportions without continuity correction

data:  c(12, 10) out of c(96, 55)
X-square = 0.9069, df = 1, p-value = 0.1705
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000  0.04516349
sample estimates:
 prop'n in Group 1 prop'n in Group 2
             0.125         0.1818182
```

La statistique de test calculée (`X-square`) est en fait U^2 qui, sous H_0, est de loi χ^2_1. La p-valeur vaut 17%, donc pour rejeter H_0, il faudrait prendre un risque d'erreur assez grand (supérieur à 17%). On est donc assez confiant dans le fait que la différence des deux proportions n'est pas significative.

4.3.3. Comparaison d'échantillons gaussiens appariés

Deux échantillons sont dits appariés si et seulement si ils sont constitués de deux mesures successives de la même variable sur les mêmes individus.

Exemple : Afin de mesurer les effets d'un nouveau régime amaigrissant, celui-ci a été testé sur 15 individus pris au hasard dans une population. Le tableau 4.2 donne leur poids en kg avant et après le régime. Le régime est-il efficace ?

avant	70	75	80	60	64	66	70	74	78	80	82	90	101	84	77
après	68	76	74	58	65	60	70	70	75	79	78	95	103	80	74

Tableau 4.2. : poids avant et après un régime amaigrissant de 15 individus

Mathématiquement, les observations sont deux échantillons de même taille n, $X_1,...,X_n$ et $Y_1,...,Y_n$. Les X_i sont indépendants entre eux, les Y_j aussi, mais X_i et Y_i ne sont pas indépendants.

On se contentera ici de supposer que les deux échantillons sont gaussiens, les X_i de loi $N(m_1,\sigma_1^2)$ et les Y_j de loi $N(m_2,\sigma_2^2)$. La procédure s'appliquera également à des échantillons de lois quelconques mais de grande taille, en vertu du théorème central-limite.

Pour tout i, posons $Z_i = X_i - Y_i$. Le test se base sur l'hypothèse que les Z_i sont indépendants et de même loi normale d'espérance $E(X_i) - E(Y_i) = m_1 - m_2 = m$. Mais ceci n'est vrai que si le vecteur (X_i, Y_i) est gaussien. Il faut donc rajouter cette hypothèse.

Alors, tester « $m_1 = m_2$ » sur les deux échantillons, c'est tester « $m = 0$ » sur le troisième échantillon. Comme c'est un échantillon gaussien, on peut le faire grâce au test de Student usuel.

Dans l'exemple, le régime est efficace si le poids moyen après régime est inférieur au poids moyen avant régime. On doit donc faire un test de « $m_1 \le m_2$ » contre « $m_1 > m_2$ », ce qui revient à faire un test de « $m \le 0$ » contre « $m > 0$ » sur l'échantillon des différences de poids avant et après le régime :

2 -1 6 2 -1 6 0 4 3 1 4 -5 -2 4 3

La région critique est $W = \left\{ \frac{\bar{x}_n - 0}{s'_n}\sqrt{n} > t_{n-1,2\alpha} \right\}$, $\bar{x}_n$ et s'_n étant calculées sur le troisième échantillon.

Ici, $n = 15$, $\bar{x}_n = 1.73$ et $s'_n = 3.08$, donc $\frac{\bar{x}_n}{s'_n}\sqrt{n} = 2.18$. Pour $\alpha = 5\%$, $t_{14,0.1} = 1.76$.

2.18 > 1.76, donc on rejette H_0 et on conclut que le régime est bien efficace, avec moins de 5% de chances de se tromper.

Sous S+, on peut soit créer le troisième échantillon et faire un test de Student usuel comme en 4.2.2., soit partir des deux échantillons et préciser dans l'appel du test qu'ils sont appariés. On obtient également un intervalle de confiance pour $m_1 - m_2$.

```
> t.test(x,y,alternative="greater",paired=T,conf.level=0.95)

      Paired t-Test

data:  x and y
t = 2.1786, df = 14, p-value = 0.0235
alternative hypothesis: true mean of differences is greater than 0
95 percent confidence interval:
 0.3319946        NA
sample estimates:
 mean of x - y
      1.733333
```

La p-valeur vaut 2.35%, donc on rejette bien H_0 au seuil 5%, mais on ne la rejetterait pas au seuil 1%.

4.4. Quelques tests non paramétriques

Un test non paramétrique est un test qui ne porte pas sur la valeur d'un paramètre d'une loi de probabilité. Il peut donc y en avoir de toutes sortes. Nous nous contenterons ici de décrire quelques uns des plus usuels de ces tests.

4.4.1. Tests d'adéquation pour un échantillon

Le problème est de déterminer si les observations $x_1,...,x_n$ peuvent être considérées comme des réalisations de variables aléatoires indépendantes de loi donnée (normale, exponentielle, binomiale,…). Nous avons déjà vu que l'histogramme et le graphe de probabilités permettent de répondre à cette question. Mais cette réponse n'est que qualitative et est basée sur un jugement visuel : deux personnes peuvent avoir des conclusions différentes au vu du même histogramme. De plus, on ne sait pas quantifier l'erreur que l'on fait si on refuse telle ou telle loi de probabilité au vu de l'échantillon.

Or il est parfaitement possible de construire un test statistique pour répondre à ce problème. Un tel test est appelé **test d'adéquation** ou **test d'ajustement**. On distinguera deux cas, suivant que l'on veut tester l'adéquation de l'échantillon à une loi de probabilité entièrement spécifiée (par exemple la loi $U[0,1]$ ou $N(2,9)$), ou à une famille de lois de probabilité (par exemple la famille des lois exponentielles).

Soit F la fonction de répartition inconnue des X_i.

Cas 1 : Il s'agit de tester H_0 : « $F = F_0$ » contre H_1 : « $F \neq F_0$ ».

Cas 2 : Il s'agit de tester H_0 : « $F \in \mathcal{F}$ » contre H_1 : « $F \notin \mathcal{F}$ », où $\mathcal{F}$ est une famille de lois de probabilité, dépendant en général d'un paramètre θ : $\mathcal{F} = \{F(.,\theta);\theta \in \Theta\}$.

Remarque : La complexité de l'hypothèse alternative fait qu'il sera impossible de calculer de manière générale la puissance d'un test d'adéquation. On pourra déterminer une puissance contre certaines alternatives spécifiées, par exemple H_1 : « $F = F_1$ ».

L'important dans un test étant de rejeter H_0, on voit que ces tests permettront essentiellement de rejeter des modèles très peu vraisemblables au vu des observations.

4.4.1.1. Le test du χ^2 sur les probabilités d'évènements

Exemple introductif : On jette un dé 204 fois. On obtient les résultats suivants :

1	2	3	4	5	6
40	30	38	34	35	27

Tableau 4.3. : résultat de 204 lancers d'un dé

Peut-on en conclure que le dé est équilibré ?

Une idée naturelle est de dire que, si le dé est équilibré, on devrait avoir à peu près 204/6=34 fois chaque face. Si le résultat s'éloigne trop de 34 sur quelques unes des faces, on peut douter du fait que le dé est équilibré. On peut donc penser à rejeter l'hypothèse que le dé est équilibré si la « distance » entre le vecteur (40, 30, 38, 34, 35, 27) et le vecteur (34, 34, 34, 34, 34, 34) est « trop grande ». Il reste à choisir une distance appropriée.

Plus généralement, on considère une expérience qui a k issues possibles. On sait que, sous une certaine hypothèse H_0, les probabilités d'apparition de ces k issues sont $p_1,...,p_k$, avec $\sum_{i=1}^{k} p_i = 1$. On fait n expériences identiques et indépendantes, et on compte les nombres n_i de fois où chaque issue i s'est produite. On a forcément $\sum_{i=1}^{k} n_i = n$. Le problème est de décider si l'observation de $n_1,...,n_k$ est compatible avec l'hypothèse H_0 que les probabilités des issues sont $p_1,...,p_k$.

Sous H_0, on s'attend à observer en moyenne np_i fois l'issue i. Il s'agit donc de déterminer si les n_i sont significativement proches ou éloignés des np_i. On peut alors penser à une région critique de la forme $W = \left\{ \sum_{i=1}^{k} (n_i - np_i)^2 > l_\alpha \right\}$. Pour déterminer l_α, il faut connaître la loi de probabilité sous H_0 de $\sum_{i=1}^{k} (N_i - np_i)^2$, ou d'une variable aléatoire analogue.

Il est clair que, pour tout i, N_i est de loi binomiale $B(n, p_i)$. Si n est suffisamment grand, on fait l'approximation de la loi binomiale par la loi normale. N_i est donc approximativement de loi normale $N\left(np_i, np_i(1-p_i)\right)$. Alors, $\frac{N_i - np_i}{\sqrt{np_i(1-p_i)}}$ est approximativement de loi $N(0,1)$, et $\frac{(N_i - np_i)^2}{np_i(1-p_i)}$ est approximativement de loi χ_1^2. Si les N_i étaient indépendantes, on en déduirait que $\sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i(1-p_i)}$ est approximativement de loi χ_k^2. Mais elles ne sont pas indépendantes puisque $\sum_{i=1}^{k} N_i = n$. Il y a donc une correction à faire, qui est donnée par le théorème de Pearson :

<u>Théorème de Pearson</u> : Sous H_0 : « les probabilités des k issues sont $p_1,...,p_k$ », la variable aléatoire $D_n^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$ converge en loi vers la loi χ_{k-1}^2.

On en déduit alors facilement un test , appelé **test du khi-deux** :

<u>Test du χ^2</u> : Test de H_0 : « les probabilités des k issues sont $p_1,...,p_k$ » contre $H_1 = \overline{H}_0$:

$$W = \left\{ \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} > z_{k-1,\alpha} \right\}$$

Sur l'exemple du dé, $d_n^2 = \frac{(40-34)^2}{34} + ... + \frac{(27-34)^2}{34} = 3.47$. Au seuil 5%, $z_{5,0.05} = 11.07$.

3.47 << 11.07, donc on ne rejette pas H_0 : rien n'indique que le dé n'est pas équilibré.

Remarque 1 : La table de la loi du χ^2 indique que la p-valeur est comprise entre 50% et 70%. Donc pour rejeter H_0, il faudrait tolérer une probabilité d'erreur exagérément grande. Il y a donc toutes les raisons de penser que le dé est équilibré.

Remarque 2 : Le test repose sur l'approximation de la loi binomiale par la loi normale. Pour l'appliquer, on doit donc avoir pour tout i, $np_i \geq 5$ et $n(1-p_i) \geq 5$. En pratique, on considère que l'on peut effectuer un test du χ^2 si, pour tout i, $n_i \geq 5$.

4.4.1.2. Le test du χ^2 d'adéquation à une famille de lois de probabilité

La démarche précédente peut être utilisée pour effectuer un test d'adéquation pour un échantillon.

Quand les variables observées $x_1,...,x_n$ sont *discrètes*, on peut se demander si elles sont issues d'une loi de probabilité discrète comme la loi binomiale, loi de Poisson, etc… Les n observations $x_1,...,x_n$ ont pris k valeurs différentes $e_1,...,e_k$. Soit $n_i = \sum_{j=1}^{n} 1_{\{e_i\}}(x_j)$ le nombre d'observations égales à e_i. Sous l'hypothèse H_0 que les X_j sont indépendantes et de loi donnée, on connaît les $p_i = P(X = e_i)$. Alors le théorème de Pearson s'applique et on peut utiliser le test du χ^2 pour tester l'adéquation de l'échantillon à cette loi.

Quand les variables observées $x_1,...,x_n$ sont *continues*, on peut se demander si elles sont issues d'une loi de probabilité continue comme la loi exponentielle, loi normale, etc… Revenons alors à la construction de l'histogramme vue en 2.2.2.1. On a choisi k intervalles $]a_{i-1}, a_i]$ et compté les nombres $n_i = \sum_{j=1}^{n} 1_{]a_{i-1},a_i]}(x_j)$ d'observations appartenant à chaque intervalle. Sous l'hypothèse H_0 que les X_j sont indépendantes et de loi donnée, la probabilité qu'une observation appartienne à la classe i est $p_i = P(X \in]a_{i-1}, a_i]) = F(a_i) - F(a_{i-1})$. Le théorème de Pearson permet alors d'utiliser le test du χ^2 pour tester l'adéquation de l'échantillon à cette loi.

Remarque : Le nombre de classes conseillé pour effectuer un test du χ^2 n'est pas le même que le nombre de classes conseillé pour dessiner un histogramme. On prendra en général $k \approx 2n^{2/5}$. D'autre part, il est conseillé de n'effectuer le test que si on a au moins 5 observations par classe.

Rappelons que, dans les tests d'adéquation, deux cas sont à considérer selon que l'on connaît parfaitement ou pas la loi à tester.
Dans le cas 1, sous H_0 : « $F = F_0$ », $p_i = F_0(a_i) - F_0(a_{i-1})$ pour tout i. Les p_i sont parfaitement connus et on peut appliquer le test tel quel.

Dans le cas 2, sous H_0 : « $F \in \mathcal{F} = \{F(.,\theta);\theta \in \Theta\}$», $p_i = F(a_i,\theta) - F(a_{i-1},\theta)$ pour tout i. Les p_i dépendent d'un paramètre θ inconnu, donc on ne peut pas utiliser directement le test. L'idée naturelle est de remplacer θ par un estimateur $\hat{\theta}_n$ et de remplacer p_i par le $\hat{p}_i$ correspondant.
On montre alors que, si $\hat{\theta}_n$ est l'estimateur de maximum de vraisemblance de θ, la variable aléatoire $\hat{D}_n^2 = \sum_{i=1}^{k} \frac{(N_i - n\hat{p}_i)^2}{n\hat{p}_i}$ converge approximativement en loi vers la loi χ^2_{k-1-p}, où p est la dimension de θ. D'où le résultat final :

Test du χ^2 d'adéquation à une famille de lois : Test de H_0 : « $F \in \mathcal{F} = \{F(.,\theta);\theta \in \Theta\}$» contre $H_1 = \overline{H}_0$: $W = \left\{ \sum_{i=1}^{k} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} > z_{k-1-p,\alpha} \right\}$

Exemple : Reprenons l'exemple des données sur les niveaux de bruit à Montréal. On souhaite tester H_0 : « Les observations proviennent d'une loi normale » contre H_1 : « Les observations ne proviennent pas d'une loi normale ». La loi normale a $p = 2$ paramètres. Les estimations de maximum de vraisemblance de m et σ^2 sont respectivement $\bar{x}_n = 64.2$ et $s_n^2 = 25.2$. Notons qu'il faut bien utiliser l'estimateur de maximum de vraisemblance de σ^2, même s'il n'est pas optimal.

Nous avons construit un histogramme à $k = 5$ classes de même effectif. Nous ne sommes pas tout à fait dans les conditions d'application du test puisqu'il faudrait en théorie au moins 5 observations par classe et que nous n'en avons que 4. Appliquons néanmoins le test pour comprendre son fonctionnement.

On a $p_i = P(a_{i-1} < X \le a_i) = P\left(\frac{a_{i-1} - m}{\sigma} < \frac{X - m}{\sigma} \le \frac{a_i - m}{\sigma} \right) = \phi\left(\frac{a_i - m}{\sigma} \right) - \phi\left(\frac{a_{i-1} - m}{\sigma} \right)$,
d'où $\hat{p}_i = \phi\left(\frac{a_i - \bar{x}_n}{s_n} \right) - \phi\left(\frac{a_{i-1} - \bar{x}_n}{s_n} \right)$.
Le test d'adéquation à la loi normale aura donc pour région critique $W = \left\{ \sum_{i=1}^{5} \frac{(4 - 20\hat{p}_i)^2}{20\hat{p}_i} > z_{2,\alpha} \right\}$.

Le vecteur des $\hat{p}_i$ est (0.170, 0.232, 0.181, 0.211, 0.155), d'où $\hat{d}_n^2 = \sum_{i=1}^{5} \frac{(4 - 20\hat{p}_i)^2}{20\hat{p}_i} = 0.514$.
Au seuil 5%, $z_{2,0.05} = 5.99$. 0.514 << 5.99, donc on ne rejette pas H_0, et de loin. On peut donc avoir une bonne confiance dans la normalité des observations.

Sous S+, la commande permettant d'effectuer un test du χ^2 est `chisq.gof(x)`. Il faut préciser le nombre et les bornes des classes, la loi à tester, le nombre de paramètres à estimer et comment on estime ces paramètres. Pour l'exemple, on obtient :

```
> chisq.gof(x,n.classes=5,cut.points=c(54.3,59.9,63.3,65.6,68.8,73.9),
  distribution="normal",n.param.est=2,mean=mean(x),
  sd=sqrt(var(x,unbiased=F)))
```

```
      Chi-square Goodness of Fit Test

data:  x
Chi-square = 0.5141, df = 2, p-value = 0.7733
alternative hypothesis: True cdf does not equal the normal Distn. for at
least one sample point.

Warning messages:
  Expected counts < 5. Chi-squared approximation may not be appropriate
```

On retrouve que $\hat{d}_n^2$ vaut 0.514. La p-valeur est 77.33%, qui est très élevée. On est donc en effet bien loin de rejeter H_0. S+ signale qu'il faudrait pour bien faire avoir au moins 5 observations par classe.

4.4.1.3. Les tests basés sur la fonction de répartition empirique

On a vu que la fonction de répartition empirique F_n était un excellent estimateur de la vraie fonction de répartition inconnue des observations. Il est donc naturel de rejeter l'hypothèse H_0 : « $F = F_0$ » si les fonctions F_n et F_0 sont significativement éloignées. Il y a plusieurs façons de mesurer cet écart :

- statistique de Kolmogorov-Smirnov : $K_n = \sqrt{n} \underset{x \in R}{Sup} \left| F_n(x) - F_0(x) \right|$
- statistique de Cramer-von Mises : $W_n^2 = n \int_{-\infty}^{+\infty} \left(F_n(x) - F_0(x) \right)^2 dF_0(x)$
- statistique d'Anderson-Darling : $A_n^2 = n \int_{-\infty}^{+\infty} \frac{\left(F_n(x) - F_0(x) \right)^2}{F_0(x)(1 - F_0(x))} dF_0(x)$

On montre que, sous H_0, K_n, W_n^2 et A_n^2 convergent en loi vers des lois de probabilité indépendantes de F_0, ce qui permet de réaliser des tests d'adéquation, avec des régions critiques du type $W = \{K_n > l_\alpha\}$. Mais les lois limites ont des expressions complexes ou même pas d'expressions explicites. On est donc obligés de se référer à des tables ou à des logiciels de statistique.

D'autre part, si on teste H_0 : « $F \in \mathcal{F} = \{F(.,\theta) ; \theta \in \Theta\}$», il faut remplacer θ par un estimateur $\hat{\theta}_n$. Les lois limites des statistiques correspondantes $\hat{K}_n$, $\hat{W}_n^2$ et $\hat{A}_n^2$ ne sont alors plus les mêmes que précédemment, et sont en plus différentes suivant le type de loi testée.

On voit donc que ces tests peuvent très difficilement être effectués « à la main ». Heureusement, ils sont implémentés dans certains logiciels. Sous S+, le test de Kolmogorov-Smirnov peut s'effectuer à l'aide de la commande `ks.gof`. Pour l'exemple des niveaux de bruit, on obtient :

```
> ks.gof(x,distribution="normal",mean=mean(x),
  sd= sqrt(var(x,unbiased=F)))

      One-sample Kolmogorov-Smirnov Test
      Hypothesized distribution = normal

data:  x
ks = 0.0758, p-value = 0.9993
```

```
alternative hypothesis: True cdf is not the normal distn. with the speci-
fied parameters
```

La p-valeur est de 99.93%, ce qui signifie qu'on est très loin de rejeter H_0. Donc on conclut à la normalité des observations, ce qui est cohérent avec le résultat obtenu par le test du χ^2.

Les tests basés sur la fonction de répartition empirique sont nettement plus complexes à mettre en œuvre que le test du χ^2, mais ils sont plus puissants, car ils évitent la perte d'information due au regroupement en classes dans le test du χ^2. Il est donc conseillé de les utiliser.

Notons pour terminer qu'il existe des tests d'adéquation spécifiques à certaines familles de lois. Par exemple, le meilleur des tests d'adéquation à la loi normale est le test de Shapiro-Wilk. Mais ce test n'est pas implémenté dans S+.

4.4.2. Tests non paramétriques de comparaison de deux échantillons

Dans cette section, on suppose que l'on dispose de deux échantillons indépendants $X_1,...,X_{n_1}$ et $Y_1,...,Y_{n_2}$. On désire savoir si les X_i et les Y_j ont même loi, sans faire d'hypothèses sur cette loi. Le problème est donc de tester H_0 : « $F_X = F_Y$ » contre H_1 : « $F_X \neq F_Y$ ». On a vu que l'on sait répondre à cette question si on suppose que les deux échantillons sont gaussiens.

4.4.2.1. Test de Kolmogorov-Smirnov

Si les deux échantillons proviennent de la même loi, ils ont même fonction de répartition, donc leurs fonctions de répartition empiriques doivent être très proches. Le test de Kolmogorov-Smirnov consiste à rejeter l'hypothèse H_0 : « $F_X = F_Y$ » si et seulement si $\underset{x\in R}{Sup}\left|F_{X,n_1}(x) - F_{Y,n_2}(x)\right|$ est « trop grand ». On utilise pour cela le fait que, sous H_0, la variable aléatoire $\sqrt{\frac{n_1 n_2}{n_1+n_2}}\underset{x\in R}{Sup}\left|F_{X,n_1}(x) - F_{Y,n_2}(x)\right|$ converge en loi vers la même loi limite que la statistique K_n du test de Kolmogorov-Smirnov sur un échantillon.

Sous S+, on peut comparer visuellement les deux fonctions de répartition empiriques à l'aide de la commande `cdf.compare`. Le test de Kolmogorov-Smirnov s'effectue à l'aide de la commande `ks.gof`, comme dans le cas d'un seul échantillon. Mais cette fois, il est inutile de préciser une loi de probabilité que l'on désire tester.

Exemple : Un même logiciel a été vendu à deux sociétés, 8 exemplaires à la société A et 10 exemplaires à la société B. On a relevé le nombre d'utilisations de chaque exemplaire, sur la même période de temps :

Société A : 110 82 121 47 103 78 97 143
Société B : 92 101 38 71 52 108 65 64 88 111

Peut-on en conclure que le logiciel est utilisé de façon similaire dans les deux sociétés ?

Des histogrammes et des tests d'adéquation montrent que ces deux échantillons ne sont pas gaussiens. Il est donc nécessaire d'adopter une démarche non paramétrique.

Commençons par comparer visuellement les deux fonctions de répartition empiriques :

```
> a<-c(110,82,121,47,103,78,97,143)
> b<-c(92,101,38,71,52,108,65,64,88,111)
> cdf.compare(a,b)
```

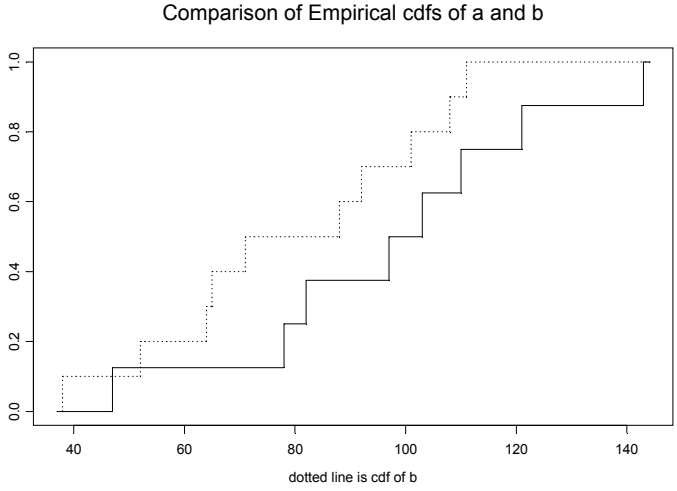


Figure 4.1. : Comparaison des fonctions de répartition empiriques des deux échantillons a et b

On constate qu'il est assez difficile de déduire quoi que ce soit de cette figure. On effectue alors un test de Kolmogorov-Smirnov qui va, en fait, déterminer si la distance verticale maximum entre ces deux fonctions est significative d'une différence entre les deux lois.

```
> ks.gof(a,b, alternative="two.sided")

      Two-Sample Kolmogorov-Smirnov Test

data:  a and b
ks = 0.375, p-value = 0.4853
alternative hypothesis:
  cdf of a does not equal the
             cdf of b for at least one sample point.
```

La p-valeur est 48.53%. Elle est élevée, donc on ne va pas rejeter H_0 : rien ne prouve que le logiciel est utilisé différemment dans les deux sociétés.

4.4.2.2. Test de Wilcoxon-Mann-Whitney

Le principe de ce test est que, si les X_i et les Y_j ont même loi, alors si on mélange les deux séries de valeurs, on doit obtenir un mélange homogène. Plus précisément, soit U le nombre de couples (i, j) pour lesquels $X_i \leq Y_j$. Il y a en tout $n_1 n_2$ couples (i, j). Comme les X_i et les Y_j sont indépendan-

tes, si elles ont même loi on aura $P(X_i \le Y_j) = \frac{1}{2}$, donc U devrait être de l'ordre de $\frac{n_1 n_2}{2}$. Le **test de Mann-Whitney** consiste alors à rejeter H_0 : « $F_X = F_Y$ » si et seulement si $\left| U - \frac{n_1 n_2}{2} \right|$ est « trop grand ».

Pour déterminer la région critique, on doit connaître la loi de U sous H_0. Pour de petits échantillons, on utilise des tables de cette loi. Pour de grands échantillons ($n_1 \ge 8$ et $n_2 \ge 8$), on utilise une approximation normale : sous H_0, U est approximativement de loi $N\left(\frac{n_1 n_2}{2}, \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \right)$. On en déduit le test :

Test de Mann-Whitney : Test de H_0 : « $F_X = F_Y$ » contre H_1 : « $F_X \ne F_Y$ » :

$$W = \left\{ \left| \frac{u - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \right| > u_\alpha \right\}$$

Dans l'exemple, on obtient u= 1+5+0+9+2+5+3+0=25. Comme on a plus de 8 observations par échantillon, on peut utiliser l'approximation gaussienne.

$\frac{u - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = -1.33$. Au seuil 5%, $u_{0.05} = 1.96$.

$|-1.33| < 1.96$, donc on ne rejette pas H_0 : comme avec le test de Kolmogorov-Smirnov, on conclut que rien ne prouve que le logiciel est utilisé différemment dans les deux sociétés.

Une autre façon de tester l'égalité des lois dans les deux échantillons est de compter la somme W des rangs des observations de l'échantillon $X_1, ..., X_{n_1}$ dans la série résultant du mélange des deux échantillons initiaux. Le test correspondant est appelé **test de Wilcoxon**. En fait, on montre que $U = \frac{n_1 (n_1 + 2n_2 + 1)}{2} - W$, ce qui prouve que les tests de Mann-Whitney et de Wilcoxon sont équivalents. On emploie souvent le terme de **test de Wilcoxon-Mann-Whitney**.

Sous S+, c'est le test de Wilcoxon qui a été retenu, avec la commande `wilcox.test`. Dans notre exemple, on obtient :

```
> wilcox.test(a,b,alternative="two.sided")

      Exact Wilcoxon rank-sum test

data:  a and b
```

```
rank-sum statistic W = 91, n = 8, m = 10, p-value = 0.2031
alternative hypothesis: true mu is not equal to 0
```

On retrouve bien que $\frac{n_1(n_1+2n_2+1)}{2} - w = 116 - 91 = 25 = u$. La p-valeur du test est 20.31%, donc on ne rejette pas H_0, ce qui est bien le résultat déjà trouvé.

Remarque 1 : On montre que le test de Wilcoxon-Mann-Whitney est plus puissant que le test de Kolmogorov-Smirnov, ce qui signifie qu'il détectera plus facilement si les deux lois ne sont pas les mêmes. Il est donc recommandé d'employer le test de Wilcoxon-Mann-Whitney.

Remarque 2 : Les deux tests présentés ici sont des tests de « $F_X = F_Y$ » contre « $F_X \neq F_Y$ ». Il est très facile d'en déduire des tests de « $F_X \leq F_Y$ » contre « $F_X > F_Y$ » et de « $F_X \geq F_Y$ » contre « $F_X < F_Y$ ».

Chapitre 5 : La régression linéaire

5.1. Introduction

Dans la quasi-totalité de ce cours, on a considéré que les observations étaient unidimensionnelles, c'est-à-dire que les variables aléatoires étudiées étaient à valeurs dans R ou un sous-ensemble de R. On a commencé à aborder le cas de données bidimensionnelles quand il a été question de comparaison d'échantillons appariés : $X_1,...,X_n$ et $Y_1,...,Y_n$. En effet, on peut considérer que l'on dispose en fait de l'observation d'un seul échantillon de n couples aléatoires : $(X_1,Y_1),...,(X_n,Y_n)$. Si les X_i et les Y_i sont indépendants, le traitement se ramène à celui de données unidimensionnelles. Quand les X_i et les Y_i ne sont pas indépendants, il faut utiliser des méthodes spécifiques.

Le problème principal est l'étude de la dépendance entre X_i et Y_i. Un problème de **régression** consiste à chercher une fonction f telle que pour tout i, Y_i est approximativement égal à $f(X_i)$. Le cas le plus simple est celui de la **régression linéaire**, où on cherche f de la forme $f(x)=ax+b$. Dans ce cadre, les problèmes usuels sont l'estimation de a et b, ponctuelle et par intervalle de confiance, et la construction de tests d'hypothèses portant sur a et b. La méthode d'estimation est bien connue sous le nom de **méthode des moindres carrés**.

Exemple : Pour tester la performance du système de freinage d'une voiture, on la fait rouler jusqu'à atteindre une vitesse x, à laquelle on freine. On mesure alors la distance de freinage y. On fait l'expérience pour n vitesses différentes $x_1,...,x_n$ et on mesure les n distances de freinage correspondantes $y_1,...,y_n$. On obtient le tableau 5.1. :

vitesse (m/s)	5	10	15	20	25	30	35	40
distance de freinage (m)	3.42	5.96	31.14	41.76	74.54	94.92	133.78	169.16

Tableau 5.1. : vitesse et distance de freinage d'une voiture

Quel modèle de dépendance entre la distance de freinage et la vitesse peut-on proposer ? Peut-on estimer la distance de freinage d'une voiture lancée à 50 m/s ? Avec quelle précision ?

5.2. Le modèle de régression linéaire

On dispose de données bidimensionnelles, qui sont n couples (x_i,y_i). C'est le cas de l'exemple. On souhaite modéliser la dépendance entre la vitesse x et la distance de freinage y. Il est clair que y dépend de x, mais pas seulement : l'état de la route, la météo, la nervosité du conducteur, peuvent influer sur la distance de freinage. En tous cas, même quand on connaît x, y n'est pas prévisible avec certitude à l'avance. Par conséquent, on considèrera que la distance de freinage y est la réalisation d'une variable aléatoire Y. Dans l'exemple, il est clair que la vitesse à laquelle on freine est contrôlée par le conducteur, donc n'est pas aléatoire. Aussi on supposera que x est une constante

connue. Mais tout ce qui est dit dans ce chapitre se généralisera au cas où x est la réalisation d'une variable aléatoire X.

Il faut donc exprimer le fait que la variable aléatoire Y dépend de la grandeur mesurée x et d'un certain nombre d'autres facteurs imprévisibles et non mesurés. Le modèle de régression suppose que l'effet de tous les facteurs autres que x est aléatoire et s'ajoute à l'effet de x :

*Définition : Le **modèle de régression** de Y sur x est défini par*

$$Y = f(x) + \varepsilon$$

où :

- *Y est la **variable expliquée** ou **variable observée***
- *x est la **variable explicative** ou **prédicteur***
- *ε est l'**erreur** de prédiction de Y par x ou **résidu***

Les données consistent en plusieurs observations de Y, obtenues pour différentes valeurs de x. Le modèle de régression s'écrit alors $Y_i = f(x_i) + \varepsilon_i$, $\forall i \in \{1,...,n\}$. On suppose en général que les variables aléatoires Y_i sont indépendantes.
Pour signifier que les facteurs autres que le prédicteur x ont des effets qui se compensent, on considère en général que les résidus ε_i sont centrés : $\forall i, E(\varepsilon_i) = 0$. Pour signifier que les expériences ont toutes été faites dans les mêmes conditions, on suppose en général que les résidus sont de même loi, et on note σ^2 leur variance.

Enfin, on définit le modèle de régression linéaire quand, en plus de ces hypothèses, on suppose que f est linéaire.

*Définition : Le **modèle de régression linéaire simple** est défini par :*

$$\forall i \in \{1,...,n\},\ Y_i = ax_i + b + \varepsilon_i$$

où les résidus ε_i sont indépendants, de même loi, centrés et de variance σ^2.

On a alors :

- $\forall i \in \{1,...,n\},\ E(Y_i) = E(ax_i + b + \varepsilon_i) = ax_i + b + E(\varepsilon_i) = ax_i + b$
- $\forall i \in \{1,...,n\},\ Var(Y_i) = Var(ax_i + b + \varepsilon_i) = Var(\varepsilon_i) = \sigma^2$

σ^2 mesure le bruit, ou le poids des facteurs autres que le prédicteur. Plus σ^2 est élevé, plus Y_i fluctue autour de $ax_i + b$.

On verra que l'on peut estimer a, b et σ^2 sans connaître plus précisément la loi des résidus. Cependant, dans de nombreux cas, il est raisonnable de supposer que les ε_i sont de loi normale. Dans notre exemple, on peut considérer que les facteurs autres que la vitesse sont très nombreux et s'ajoutent, ce qui aboutit à une hypothèse de loi normale grâce au théorème central-limite. On obtient alors le modèle linéaire gaussien :

Définition : Le ***modèle linéaire simple gaussien*** *est défini par :* $\forall i \in \{1,...,n\}$, *les variables aléatoires* Y_i *sont indépendantes et de lois de probabilité respectives* $N(ax_i + b, \sigma^2)$

Remarque fondamentale : ce qui compte en fait, c'est que la fonction de régression f soit linéaire par rapport aux paramètres a et b, pas par rapport au prédicteur x. Ainsi, à l'aide d'une généralisation simple, on pourra considérer que les modèles suivants sont, contrairement aux apparences, des modèles linéaires :

- $Y_i = ax_i^2 + bx_i + c + \varepsilon_i$
- $Y_i = a \ln x_i + b + \varepsilon_i$
- $Y_i = ax_i^b \varepsilon_i$, car $\ln Y_i = \ln a + b \ln x_i + \ln \varepsilon_i$

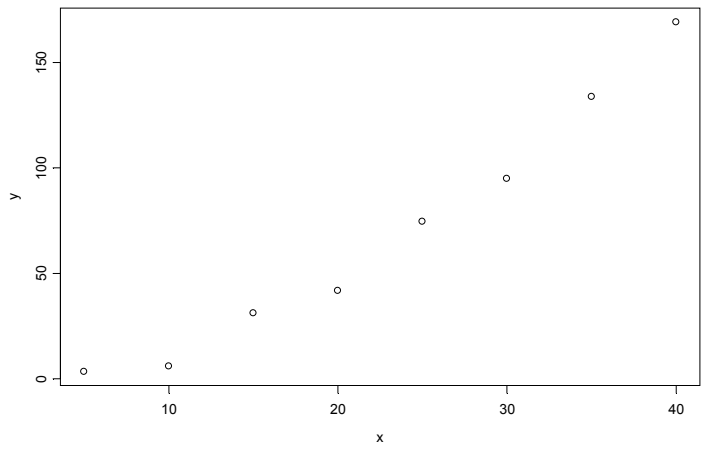
En revanche, le modèle $Y_i = e^{ax_i} + b + \varepsilon_i$ n'est pas un modèle linéaire.

5.3. Estimation des paramètres : la méthode des moindres carrés

Considérons le modèle de régression linéaire simple $Y_i = ax_i + b + \varepsilon_i$.
La première chose à faire est de dessiner le nuage des points (x_i, y_i) $\forall i \in \{1,...,n\}$, de manière à s'assurer visuellement qu'une hypothèse de dépendance linéaire entre x et y n'est pas absurde.

Sous S+, la fonction `plot(x,y)` permet de dessiner ce nuage. Sur l'exemple, on obtient :

```
> x<-c(5,10,15,20,25,30,35,40)
> y<-c(3.42,5.96,31.14,41.76,74.54,94.92,133.78,169.16)
> plot(x,y)
```



A première vue, l'hypothèse de dépendance linéaire peut être retenue pour ces données. En fait, il existe des méthodes statistiques permettant de juger de la pertinence de cette hypothèse plus précisément que par une simple impression visuelle.

Le problème maintenant est de déterminer la droite « la plus proche » de ce nuage de points, en un certain sens. La méthode la plus couramment utilisée est la **méthode des moindres carrés**, due à Gauss. Elle consiste à retenir la droite pour laquelle la somme des distances verticales des points à la droite est minimum.

Autrement dit, il faut trouver a et b tels que $\sum_{i=1}^{n}(y_i - ax_i - b)^2$ soit minimum. C'est ce qui justifie le nom de « moindres carrés » pour cette méthode. On préfère en fait, ce qui revient au même, minimiser **l'erreur quadratique moyenne** $\delta^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i - b)^2$. Pour cela, on annulle les dérivées partielles de δ^2 par rapport à a et b :

$$\frac{\partial \delta^2}{\partial a} = -\frac{2}{n}\sum_{i=1}^{n} x_i(y_i - ax_i - b) = -2\left[\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{a}{n}\sum_{i=1}^{n} x_i^2 - \frac{b}{n}\sum_{i=1}^{n} x_i\right]$$

$$\frac{\partial \delta^2}{\partial b} = -\frac{2}{n}\sum_{i=1}^{n}(y_i - ax_i - b) = -2\left[\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{a}{n}\sum_{i=1}^{n} x_i - \frac{nb}{n}\right] = -2\left[\bar{y}_n - a\bar{x}_n - b\right]$$

$\frac{\partial \delta^2}{\partial b} = 0 \Rightarrow \bar{y}_n = a\bar{x}_n + b$. Par conséquent, la droite des moindres carrés passe par le centre de gravité du nuage, le point $(\bar{x}_n, \bar{y}_n)$.

Outre les notations habituelles $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$, il faut faire intervenir de nouvelles notations :

- $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2$ est la variance empirique des x_i
- $s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}_n^2$ est la variance empirique des y_j
- $c_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}_n\bar{y}_n$ est la **covariance empirique** entre les x_i et les y_j
- $r_{xy} = \frac{c_{xy}}{s_x s_y}$ est le **coefficient de corrélation linéaire empirique** entre les x_i et les y_j

c_{xy} et r_{xy} sont les versions empiriques de la covariance $Cov(X,Y) = E(XY) - E(X)E(Y)$ et du coefficient de corrélation linéaire $\rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$ (voir annexe de probabilités). On peut montrer en particulier que r_{xy} vérifie des propriétés analogues à celles de $\rho(X,Y)$:

- $r_{xy} \in [-1,+1]$

- $r_{xy} = +1 \Leftrightarrow$ les points (x_i, y_i) sont alignés sur une droite de pente positive
- $r_{xy} = -1 \Leftrightarrow$ les points (x_i, y_i) sont alignés sur une droite de pente négative
- si y ne dépend pas de x, r_{xy} doit être proche de 0. Réciproquement, si r_{xy} est proche de 0, alors il n'y a pas de dépendance linéaire entre x et y, mais il est possible qu'il existe une dépendance non linéaire.

Sous S+, s_x^2 est donnée par `var(x,unbiased=F)` (rappelons en effet que `var(x)` donne la variance empirique débiaisée), c_{xy} par `var(x,y,unbiased=F)` et r_{xy} par `cor(x,y)`.

Grâce à ces notations, on peut écrire :

$$\frac{\partial \delta^2}{\partial a} = 0 \;\Rightarrow\; \frac{1}{n}\sum_{i=1}^{n} x_i y_i - a\frac{1}{n}\sum_{i=1}^{n} x_i^2 - b\frac{1}{n}\sum_{i=1}^{n} x_i = 0$$
$$\Rightarrow\; c_{xy} + \bar{x}_n \bar{y}_n - a(s_x^2 + \bar{x}_n^2) - b\bar{x}_n = 0$$

et en prenant en compte le fait que $\bar{y}_n = a\bar{x}_n + b$, on obtient :

$$c_{xy} + a\bar{x}_n^2 + b\bar{x}_n - as_x^2 - a\bar{x}_n^2 - b\bar{x}_n = c_{xy} - as_x^2 = 0$$

d'où $a = \dfrac{c_{xy}}{s_x^2}$ et $b = \bar{y}_n - \dfrac{c_{xy}}{s_x^2}\bar{x}_n$. Le problème est résolu.

> <u>Définition</u> : *La* ***droite des moindres carrés*** *est la droite d'équation* $y = \hat{a}_n x + \hat{b}_n$, *où* $\hat{a}_n = \dfrac{c_{xy}}{s_x^2}$ *et* $\hat{b}_n = \bar{y}_n - \dfrac{c_{xy}}{s_x^2}\bar{x}_n$. *Elle peut aussi s'écrire* $y = \dfrac{c_{xy}}{s_x^2}(x - \bar{x}_n) + \bar{y}_n$.

L'erreur quadratique moyenne minimum est alors :

$$\delta^2_{\min} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a}_n x_i - \hat{b}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{c_{xy}}{s_x^2}(x_i - \bar{x}_n) - \bar{y}_n\right)^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 + \frac{c_{xy}^2}{s_x^4}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 - 2\frac{c_{xy}}{s_x^2}\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n)$$
$$= s_y^2 + \frac{c_{xy}^2}{s_x^2} - 2\frac{c_{xy}^2}{s_x^2} = s_y^2 - \frac{c_{xy}^2}{s_x^2} = s_y^2 - r_{xy}^2 s_y^2 = s_y^2(1 - r_{xy}^2)$$

On retrouve le fait que l'erreur quadratique moyenne est nulle si et seulement si $r_{xy}^2 = 1$, c'est-à-dire si et seulement si les points sont alignés.

Comme d'habitude, les quantités que l'on vient de manipuler sont des réalisations des variables aléatoires correspondantes. Notons ici que les y_i sont des réalisations des variables Y_i, tandis que les x_i sont des constantes connues. On pose alors $c_{xY} = \frac{1}{n}\sum_{i=1}^{n} x_i Y_i - \bar{x}_n \bar{Y}_n$ et $s_Y^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}_n^2$ et on obtient :

> *Définition : Dans le modèle de régression linéaire simple* $Y_i = ax_i + b + \varepsilon_i$, $\forall i \in \{1,...,n\}$, *les* ***estimateurs des moindres carrés*** *de* a *et* b *sont* $\hat{A}_n = \frac{c_{xY}}{s_x^2}$ *et* $\hat{B}_n = \bar{Y}_n - \frac{c_{xY}}{s_x^2}\bar{x}_n$.

Il reste maintenant à déterminer si ces estimateurs sont de bonne qualité. Etudions leur biais et leur variance.

$$E(\hat{A}_n) = E\left(\frac{c_{xY}}{s_x^2}\right) = \frac{1}{s_x^2} E\left(\frac{1}{n}\sum_{i=1}^{n} x_i Y_i - \bar{x}_n \bar{Y}_n\right) = \frac{1}{s_x^2}\left[\frac{1}{n}\sum_{i=1}^{n} x_i E(Y_i) - \bar{x}_n E(\bar{Y}_n)\right].$$

Or $E(Y_i) = ax_i + b$ et $E(\bar{Y}_n) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = a\bar{x}_n + b$, d'où :

$$E(\hat{A}_n) = \frac{1}{s_x^2}\left[\frac{1}{n}\sum_{i=1}^{n}(ax_i^2 + bx_i) - a\bar{x}_n^2 - b\bar{x}_n\right] = \frac{1}{s_x^2}\left[a\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2\right) + b\bar{x}_n - b\bar{x}_n\right] = \frac{1}{s_x^2} a s_x^2 = a .$$

D'autre part, $E(\hat{B}_n) = E(\bar{Y}_n) - E(\hat{A}_n)\bar{x}_n = a\bar{x}_n + b - a\bar{x}_n = b$.

Par conséquent, $\hat{A}_n$ et $\hat{B}_n$ sont des estimateurs sans biais de a et b.

De la même façon, on montre que $Var(\hat{A}_n) = \frac{\sigma^2}{n s_x^2}$ et $Var(\hat{B}_n) = \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}_n}{s_x^2}\right)$, ce qui prouve que ces estimateurs sont convergents.

En fait, on a un résultat beaucoup plus fort :

> *Théorème de Gauss-Markov : $\hat{A}_n$ et $\hat{B}_n$ sont les estimateurs sans biais et de variance minimum de a et b parmi tous les estimateurs sans biais qui s'écrivent comme des combinaisons linéaires des Y_i.*

Nous avons estimé a et b, il reste maintenant à estimer la variance σ^2. On sait que, pour tout i, $Var(\varepsilon_i) = Var(Y_i - ax_i - b) = \sigma^2$. Les résidus $\varepsilon_i = Y_i - ax_i - b$ sont naturellement estimés par les **résidus empiriques** $\hat{\varepsilon}_i = Y_i - \hat{A}_n x_i - \hat{B}_n$. Une idée naturelle pour estimer σ^2 est de prendre la variance empirique des résidus empiriques. Cette variance est :

$$s_{\hat{\varepsilon}}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 - \bar{\hat{\varepsilon}}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{A}_n x_i - \hat{B}_n)^2 - (\bar{Y}_n - \hat{A}_n \bar{x}_n - \hat{B}_n)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{A}_n x_i - \hat{B}_n)^2 \quad \text{car } \bar{Y}_n = \hat{A}_n \bar{x}_n + \hat{B}_n$$

$$= \delta_{\min}^2 = s_Y^2(1 - r_{xY}^2)$$

Dans le cas d'un échantillon, la variance empirique est un estimateur biaisé de la variance de l'échantillon. Pour la débiaiser, on la multiplie par $\frac{n}{n-1}$. Ici, on a deux échantillons et deux paramètres à estimer. On peut montrer qu'alors la variance empirique ci-dessus est un estimateur biaisé de σ^2, et que, pour la débiaiser, il faut la multiplier par $\frac{n}{n-2}$. D'où finalement :

Propriété : $\hat{\sigma}_n^2 = \frac{n}{n-2}s_Y^2(1 - r_{xY}^2) = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{A}_n x_i - \hat{B}_n)^2$ *est un estimateur sans biais de* σ^2.

On n'a pas de résultat particulier sur la variance de cet estimateur dans le cas général.

Remarque : Il est important de noter que toutes les propriétés énoncées dans cette section sont valables *quelle que soit la loi des résidus* ε_i. Quand on rajoute une hypothèse sur cette loi, on peut donner des précisions sur la loi des estimateurs, leur qualité (efficacité), et construire des intervalles de confiance et des tests d'hypothèses sur les paramètres du modèle.

Revenons maintenant à l'exemple sur la liaison entre vitesse et distance de freinage. Les indicateurs statistiques sont :

$\bar{x}_n = 22.5$ $\quad \bar{y}_n = 69.33$ $\quad s_x^2 = 131.25$ $\quad s_y^2 = 3172.54$ $\quad c_{xy} = 632.31$ $\quad r_{xy} = 0.9799$

Le fait que r_{xy} soit très proche de 1 indique une forte corrélation linéaire positive, ce qui se voit clairement sur le nuage de points.

Les estimations des paramètres du modèle de régression linéaire simple sont données par :

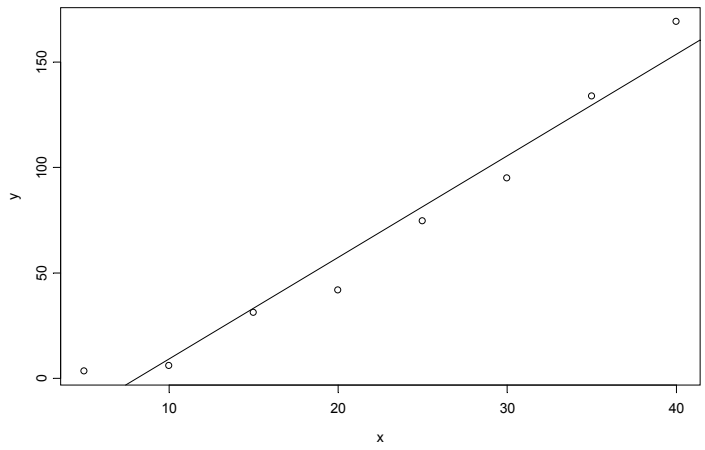
$$\hat{a}_n = \frac{c_{xy}}{s_x^2} = 4.82 \qquad \hat{b}_n = \bar{y}_n - \frac{c_{xy}}{s_x^2}\bar{x}_n = -39.06 \qquad \hat{\sigma}_n^2 = \frac{n}{n-2}s_y^2(1 - r_{xy}^2) = 168.4$$

La droite des moindres carrés a donc pour équation $y = 4.82\,x - 39.06$. On peut la superposer au nuage des points grâce à la commande S+ `abline(a,b)` :

```
> achapeau <- var(x,y,unbiased=F)/var(x,unbiased=F)
> achapeau
[1] 4.817619
> bchapeau <- mean(y)-achapeau*mean(x)
> bchapeau
[1] -39.06143
> sigma2chapeau <- n/(n-2) * var(y,unbiased=F)*(1-cor(x,y)^2)
```

```
> sigma2chapeau
[1] 168.3939

> abline(achapeau,bchapeau)
```



On peut alors facilement prévoir la distance de freinage d'une voiture lancée à 50 m/s :
4.82 x 50 – 39.06 = 201.9 m

5.4. Intervalles de confiance et tests d'hypothèses dans le modèle linéaire gaussien

On supposera dans cette section que le modèle linéaire est gaussien, c'est-à-dire que les variables aléatoires Y_i sont indépendantes et de lois de probabilité respectives $N(ax_i + b, \sigma^2)$. Les résidus ε_i sont indépendants et de même loi $N(0, \sigma^2)$.

Propriétés :

- $\hat{A}_n$ *est de loi* $N\left(a, \dfrac{\sigma^2}{ns_x^2}\right)$
- $\hat{B}_n$ *est de loi et* $N\left(b, \dfrac{\sigma^2}{n}\left(1 + \dfrac{\bar{x}_n}{s_x^2}\right)\right)$
- $Cov(\hat{A}_n, \hat{B}_n) = -\dfrac{\sigma^2 \bar{x}_n}{ns_x^2}$, *ce qui entraîne que* $\hat{A}_n$ *et* $\hat{B}_n$ *ne sont pas indépendants*
- $\dfrac{n-2}{\sigma^2}\hat{\sigma}_n^2$ *est de loi* χ^2_{n-2}
- $\hat{\sigma}_n^2$ *est indépendant de* $\hat{A}_n$ *et* $\hat{B}_n$
- $\hat{A}_n$, $\hat{B}_n$ *et* $\hat{\sigma}_n^2$ *sont les ESBVM de* a, b *et* σ^2

Les résultats pour $\hat{A}_n$ et $\hat{B}_n$ se démontrent facilement en utilisant le fait que toute combinaison linéaire de variables aléatoires indépendantes et de lois normales (les Y_i) est une variable aléatoire de loi normale. Les résultats sur $\hat{\sigma}_n^2$ sont plus complexes à démontrer et peuvent se comprendre comme une généralisation du théorème de Fisher.

Propriété : $\hat{A}_n$, $\hat{B}_n$ et $\frac{n-2}{n}\hat{\sigma}_n^2$ sont les estimateurs de maximum de vraisemblance de a, b et σ^2.

Démonstration : La fonction de vraisemblance est :

$$\mathcal{L}(a,b,\sigma^2;y_1,...,y_n)=\prod_{i=1}^{n} f_{Y_i}(y_i)=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y_i-ax_i-b)^2}{2\sigma^2}}=\frac{1}{\sigma^n(\sqrt{2\pi})^n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-ax_i-b)^2}$$

D'où $\ln\mathcal{L}(a,b,\sigma^2;y_1,...,y_n)=-\frac{n}{2}\ln\sigma^2-\frac{n}{2}\ln 2\pi-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-ax_i-b)^2$.

Maximiser $\ln\mathcal{L}(a,b,\sigma^2;y_1,...,y_n)$ en a et b revient à minimiser $\sum_{i=1}^{n}(y_i-ax_i-b)^2$ en a et b. On voit que l'on retrouve bien les estimateurs des moindres carrés.

Quant à σ^2, on a : $\frac{\partial}{\partial\sigma^2}\ln\mathcal{L}(a,b,\sigma^2;y_1,...,y_n)=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-ax_i-b)^2$, qui vaut 0 pour $\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-ax_i-b)^2$.

Donc l'estimateur de maximum de vraisemblance de σ^2 est $\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{A}_n x_i-\hat{B}_n)^2=\frac{n-2}{n}\hat{\sigma}_n^2$.

CQFD

Le fait de connaître la loi de probabilité de $\hat{A}_n$, $\hat{B}_n$ et $\frac{n-2}{\sigma^2}\hat{\sigma}_n^2$ permet d'obtenir facilement des intervalles de confiance pour les paramètres.

En effet, la définition de la loi de Student permet d'établir directement que $\frac{\hat{A}_n-a}{\hat{\sigma}_n}s_x\sqrt{n}$ est de loi $St(n-2)$ et $\frac{\hat{B}_n-b}{\hat{\sigma}_n\sqrt{s_x^2+\bar{x}_n^2}}s_x\sqrt{n}$ de loi $St(n-2)$, d'où on en déduit les intervalles de confiance suivants :

Propriétés :

Un intervalle de confiance de seuil α pour a est :

$$\left[\hat{A}_n - \frac{t_{n-2,\alpha}\hat{\sigma}_n}{s_x\sqrt{n}}, \hat{A}_n + \frac{t_{n-2,\alpha}\hat{\sigma}_n}{s_x\sqrt{n}}\right]$$

Un intervalle de confiance de seuil α pour b est :

$$\left[\hat{B}_n - \frac{t_{n-2,\alpha}\hat{\sigma}_n\sqrt{s_x^2+\bar{x}_n^2}}{s_x\sqrt{n}}, \hat{B}_n + \frac{t_{n-2,\alpha}\hat{\sigma}_n\sqrt{s_x^2+\bar{x}_n^2}}{s_x\sqrt{n}}\right]$$

Un intervalle de confiance de seuil α pour σ^2 est :

$$\left[\frac{(n-2)\hat{\sigma}_n^2}{z_{n-2,\alpha/2}}, \frac{(n-2)\hat{\sigma}_n^2}{z_{n-2,1-\alpha/2}}\right]$$

Dans l'exemple, choisissons pour seuil α = 10%. On a $t_{6,0.1} = 1.943$, $z_{6,0.05} = 12.6$ et $z_{6,0.95} = 1.64$. On obtient donc : $IC(a) = [4.04, 5.60]$, $IC(b) = [-58.71, -19.41]$, $IC(\sigma^2) = [80.2, 617.8]$.

Les intervalles de confiance pour b et σ^2 sont larges, ce qui traduit le fait que ces paramètres sont plutôt mal estimés, essentiellement à cause du faible nombre de données. En revanche, a semble assez bien estimé.

Compte-tenu de la dualité entre intervalles de confiance et tests d'hypothèses, on peut de la même manière construire des tests d'hypothèses sur la valeur des paramètres a, b et σ^2.

Par exemple, admettons que l'on veuille tester H_0 : « $a = a_0$ » contre H_1 : « $a \neq a_0$ ». Le bon sens dit que l'on rejettera H_0 si et seulement si $|\hat{a}_n - a_0|$ est « trop grand », donc on propose une région critique de la forme $W = \{|\hat{a}_n - a_0| > l_\alpha\}$.

Or, sous H_0, $\frac{\hat{A}_n - a_0}{\hat{\sigma}_n} s_x\sqrt{n}$ est de loi $St(n-2)$. On obtient donc :

$$\alpha = P_{H_0}\left(\left|\hat{A}_n - a_0\right| > l_\alpha\right) = P_{H_0}\left(\left|\frac{\hat{A}_n - a_0}{\hat{\sigma}_n} s_x\sqrt{n}\right| > \frac{l_\alpha}{\hat{\sigma}_n} s_x\sqrt{n}\right),$$

d'où $\frac{l_\alpha}{\hat{\sigma}_n} s_x\sqrt{n} = t_{n-2,\alpha}$ et $l_\alpha = t_{n-2,\alpha}\frac{\hat{\sigma}_n}{s_x\sqrt{n}}$.

On constate qu'il ne s'agit que d'une variante du test de Student. On peut donc facilement construire des tests d'hypothèses sur les paramètres du modèle. La propriété suivante donne les tests bilatéraux sur a, b et σ^2.

Propriété :

Test de seuil α de H_0 : « $a = a_0$ » contre H_1 : « $a \neq a_0$ » :

$$W = \left\{ \left| \frac{\hat{a}_n - a_0}{\hat{\sigma}_n} s_x \sqrt{n} \right| > t_{n-2,\alpha} \right\}$$

Test de seuil α *de* H_0 *: «* $b = b_0$ *» contre* H_1 *: «* $b \neq b_0$ *» :*

$$W = \left\{ \left| \frac{\hat{b}_n - b_0}{\hat{\sigma}_n} \frac{s_x \sqrt{n}}{\sqrt{s_x^2 + \bar{x}_n^2}} \right| > t_{n-2,\alpha} \right\}$$

Test de seuil α *de* H_0 *: «* $\sigma^2 = \sigma_0^2$ *» contre* H_1 *: «* $\sigma^2 \neq \sigma_0^2$ *» :*

$$W = \left\{ \frac{n-2}{\sigma_0^2} \hat{\sigma}_n^2 < z_{n-2,1-\alpha/2} \text{ ou } \frac{n-2}{\sigma_0^2} \hat{\sigma}_n^2 > z_{n-2,\alpha/2} \right\}$$

Parmi les autres hypothèses intéressantes à tester figure évidemment celle qui fonde le modèle : y a-t-il vraiment une dépendance linéaire entre y et x ? On a vu que, si c'est le cas, le coefficient de corrélation linéaire empirique r_{xy} doit être proche de 1. Inversement, si r_{xy} est proche de 0, l'hypothèse de dépendance linéaire doit être rejetée. Il est donc naturel de construire un test, dit **test de pertinence de la régression**, qui consiste à considérer que l'hypothèse de dépendance linéaire est pertinente si et seulement si r_{xy} est significativement proche de 1 ou significativement éloigné de 0. En pratique, cela revient à se demander pour quelle valeur de r_{xy} on peut considérer que des points sont approximativement alignés.

Pour cela, on remarque que $\hat{A}_n = \frac{c_{xY}}{s_x^2} = r_{xY} \frac{s_Y}{s_x}$. Donc quand r_{xy} est proche de 0, $\hat{A}_n$ est aussi proche de 0. Ou bien, quand r_{xy} est significativement éloigné de 0, $\hat{A}_n$ est aussi significativement éloigné de 0. D'où l'idée que la région critique du test de pertinence de la régression pourrait être de la forme $W = \left\{ |\hat{a}_n| > l_\alpha \right\}$. On constate qu'il s'agit simplement d'effectuer le test de H_0 : «$a = 0$» contre H_1 : «$a \neq 0$», qui est décrit ci-dessus.

On a alors $W = \left\{ \left| \frac{\hat{a}_n}{\hat{\sigma}_n} s_x \sqrt{n} \right| > t_{n-2,\alpha} \right\} = \left\{ \left| \frac{r_{xy}}{\hat{\sigma}_n} \frac{s_y}{s_x} s_x \sqrt{n} \right| > t_{n-2,\alpha} \right\} = \left\{ \left| \frac{r_{xy}}{\hat{\sigma}_n} s_y \sqrt{n} \right| > t_{n-2,\alpha} \right\}$.

Mais on a aussi $\hat{\sigma}_n^2 = \frac{n}{n-2} s_y^2 (1 - r_{xy}^2)$,

d'où $W = \left\{ \left| \frac{r_{xy}}{\sqrt{\frac{n}{n-2} s_y^2 (1 - r_{xy}^2)}} s_y \sqrt{n} \right| > t_{n-2,\alpha} \right\} = \left\{ \left| \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \sqrt{n-2} \right| > t_{n-2,\alpha} \right\}$.

Enfin, $\left|\frac{r_{xy}}{\sqrt{1-r_{xy}^2}}\sqrt{n-2}\right| > t_{n-2,\alpha} \Leftrightarrow (n-2)r_{xy}^2 > (1-r_{xy}^2)t_{n-2,\alpha}^2 \Leftrightarrow r_{xy}^2 > \frac{t_{n-2,\alpha}^2}{t_{n-2,\alpha}^2+n-2}$.

Or on sait que si T est de loi $St(n-2)$, alors T^2 est de loi $F(1,n-2)$, d'où $t_{n-2,\alpha}^2 = f_{1,n-2,\alpha}$.

La région critique du test peut donc finalement s'écrire $W = \left\{ \frac{(n-2)r_{xy}^2}{1-r_{xy}^2} > f_{1,n-2,\alpha} \right\}$ ou $W = \left\{ r_{xy}^2 > \frac{f_{1,n-2,\alpha}}{f_{1,n-2,\alpha}+n-2} \right\}$.

<u>Propriété</u> : Le test de pertinence de la régression est le test de H_0 : « $a=0$ » contre H_1 : « $a \neq 0$ ». Sa région critique peut s'écrire sous les formes suivantes :

$$W = \left\{ \left| \frac{\hat{a}_n}{\hat{\sigma}_n} s_x \sqrt{n} \right| > t_{n-2,\alpha} \right\} = \left\{ \frac{(n-2)r_{xy}^2}{1-r_{xy}^2} > f_{1,n-2,\alpha} \right\} = \left\{ r_{xy}^2 > \frac{f_{1,n-2,\alpha}}{f_{1,n-2,\alpha}+n-2} \right\}$$

Dans l'exemple, $\frac{(n-2)r_{xy}^2}{1-r_{xy}^2} = 144.7$. La table de la loi de Fisher-Snedecor donne $f_{1,6,0.05} = 5.99$ et $f_{1,6,0.01} = 13.8$. Même au seuil 1%, on est très largement dans la région critique, donc on conclut que la régression linéaire est ici très pertinente.

Sous S+, la commande permettant d'effectuer une régression linéaire de y sur x est `lm(y~x)`. Le résultat d'une régression est donné grâce à la commande `summary`. Sur l'exemple, on obtient :

```
> reg <- lm(y~x)
> summary(reg)

Call: lm(formula = y ~ x)
Residuals:
    Min     1Q Median    3Q   Max
 -15.53 -7.766 -2.609 7.048 18.39

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept) -39.0614  10.1113    -3.8631   0.0083
          x   4.8176   0.4005    12.0300   0.0000

Residual standard error: 12.98 on 6 degrees of freedom
Multiple R-Squared: 0.9602
F-statistic: 144.7 on 1 and 6 degrees of freedom, the p-value is 0.00002002

Correlation of Coefficients:
  (Intercept)
x -0.8911
```

La colonne `Value` donne les estimations des moindres carrés de b et a, $\hat{b}_n$ = -39.06 et $\hat{a}_n$ = 4.82.

La colonne `Std.error` donne les valeurs de $\frac{\hat{\sigma}_n \sqrt{s_x^2 + \bar{x}_n^2}}{s_x \sqrt{n}}$ et $\frac{\hat{\sigma}_n}{s_x \sqrt{n}}$, ce qui permet de déterminer des intervalles de confiance pour b et a.

La colonne `t value` donne les valeurs de $\left| \frac{\hat{b}_n}{\hat{\sigma}_n} \frac{s_x \sqrt{n}}{\sqrt{s_x^2 + \bar{x}_n^2}} \right|$ et $\left| \frac{\hat{a}_n}{\hat{\sigma}_n} s_x \sqrt{n} \right|$, ce qui permet d'effectuer les tests de « $b = 0$ » contre « $b \neq 0$ » et « $a = 0$ » contre « $a \neq 0$ ».

La colonne `Pr(>|t|)` donne les p-valeurs de ces tests. Dans l'exemple, ces p-valeurs sont très faibles, donc les hypothèses « $b = 0$ » et « $a = 0$ » sont largement rejetées. C'est logique puisque 0 n'appartient pas aux intervalles de confiance déterminés pour b et a.

La `Residual standard error` est $\hat{\sigma}_n$, ce qui permet de retrouver $\hat{\sigma}_n^2 = 12.98^2 = 168.4$.

Le `Multiple R-Squared` est r_{xy}^2, ce qui permet de faire le test de pertinence de la régression. La `F-statistic` est la statistique de ce test, $\frac{(n-2) r_{xy}^2}{1 - r_{xy}^2}$. On retrouve qu'elle vaut 144.7. La `p-value` fournie est la p-valeur de ce test. Elle est très faible, donc on conclut bien que la régression linéaire est pertinente sur notre exemple.

Les commandes `plot(x,y)` puis `lines(x,fitted.values(reg))` permettent de retrouver la figure de la section 5.3. représentant le nuage de points et la droite des moindres carrés.

Le modèle de régression linéaire simple gaussien semble donc satisfaisant pour l'exemple. Cependant, on s'aperçoit que ce modèle prévoit une distance de freinage négative pour toute vitesse inférieure à 8.1 m/s ! D'autre part, la forme du nuage peut évoquer plus un polynôme qu'une droite, et des raisons physiques incitent à penser que la distance de freinage est plutôt une fonction quadratique de la vitesse. Enfin, il est obligatoire que la distance de freinage correspondant à une vitesse nulle soit zéro.

Tous ces arguments amènent à penser que le modèle $Y_i = ax_i + b + \varepsilon_i$ pourrait être avantageusement remplacé par le modèle $Y_i = ax_i^2 + bx_i + \varepsilon_i$. On peut montrer que c'est encore un modèle linéaire, qui se traite de façon similaire au précédent. Nous n'avons pas le temps d'étudier théoriquement ce modèle, mais il est facile de le mettre en oeuvre grâce à S+. On obtient sur l'exemple :

```
> reg2<-lm(y~x^2+x-1)
> summary(reg2)

Call: lm(formula = y ~ x^2 + x - 1)
Residuals:
    Min     1Q  Median     3Q    Max
 -6.557 -3.04 -0.9151 2.734 5.561

Coefficients:
         Value Std. Error t value Pr(>|t|)
I(x^2)  0.1005  0.0078    12.8417  0.0000
     x  0.2467  0.2566     0.9615  0.3734
```

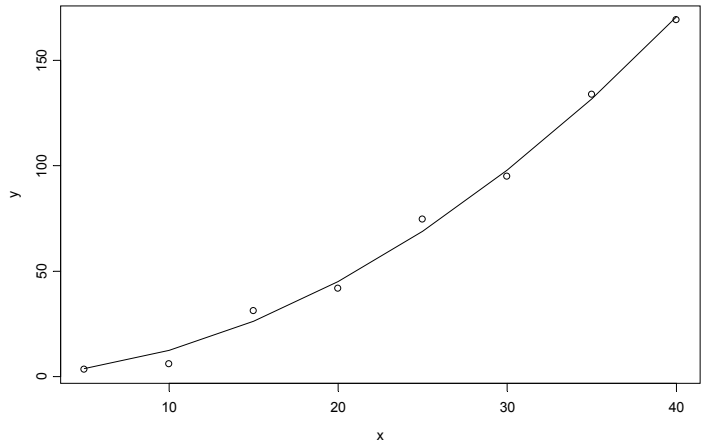
```
Residual standard error: 4.54 on 6 degrees of freedom
Multiple R-Squared: 0.9981
F-statistic: 1545 on 2 and 6 degrees of freedom, the p-value is 7.275e-009

Correlation of Coefficients:
   I(x^2)
x -0.9688
```

On a donc $\hat{a}_n = 0.1005$, $\hat{b}_n = 0.2467$ et $\hat{\sigma}_n^2 = 4.54^2 = 20.51$.

Graphiquement, on obtient :

```
> plot(x,y)
> lines(x,fitted.values(reg2))
```



Le coefficient de corrélation linéaire empirique est $r_{xy} = \sqrt{0.9981} = 0.99905$. Il est nettement plus proche de 1 que celui du modèle précédent, qui valait 0.9799. De la même façon, la p-valeur du test de pertinence de la régression vaut $7.3\ 10^{-9}$, qui est nettement plus petite que celle que l'on avait obtenue dans le modèle précédent, $2\ 10^{-5}$. Ces deux arguments montrent que le nouveau modèle est meilleur que le précédent.

La prévision de distance de freinage à la vitesse de 50 m/s est maintenant de $0.100 \times 50^2 + 0.247 \times 50 = 263.6$ m, alors qu'elle était de 201.9 m pour le modèle précédent. Cette importante différence peut avoir de grandes conséquences pratiques et met en évidence l'importance du choix d'un bon modèle de régression.

Annexe A : Rappels de probabilités pour la statistique

Cette annexe rappelle quelques résultats de base du calcul des probabilités utiles pour la statistique. Les notions sont présentées sans aucune démonstration. Les détails sont à aller chercher dans le cours de Probabilités Appliquées de première année.

A.1. Variables aléatoires réelles

A.1.1. Loi de probabilité d'une variable aléatoire

Mathématiquement, une variable aléatoire est définie comme une application mesurable. On se contentera ici de la conception intuitive suivante :

Une **variable aléatoire** est une grandeur dépendant du résultat d'une expérience aléatoire, c'est-à-dire non prévisible à l'avance avec certitude. Par exemple, on peut dire que la durée de bon fonctionnement d'une ampoule électrique ou le résultat du lancer d'un dé sont des variables aléatoires. Pour une expérience donnée, ces grandeurs prendront une valeur donnée, appelée **réalisation** de la variable aléatoire. Si on recommence l'expérience, on obtiendra une réalisation différente de la même variable aléatoire.
On ne s'intéresse ici qu'aux **variables aléatoires réelles**, c'est-à-dire à valeurs dans R ou un sous-ensemble de R.

On note traditionnellement une variable aléatoire par une lettre majuscule (X) et sa réalisation par une lettre minuscule (x).

Le calcul des probabilités va permettre de calculer des grandeurs comme la durée de vie moyenne d'une ampoule ou la probabilité d'obtenir un 6 en lançant le dé. Ces grandeurs sont déterminées par la **loi de probabilité** de ces variables aléatoires.

Il y a plusieurs moyens de caractériser la loi de probabilité d'une variable aléatoire. La plus simple est la fonction de répartition :

On appelle **fonction de répartition** de la variable aléatoire X la fonction

$$\begin{aligned} F_X \ : \ & R \to [0,1] \\ & x \to F_X(x) = P(X \le x) \end{aligned}$$

F_X est une fonction croissante, continue à droite, telle que $\lim_{x \to -\infty} F_X(x) = 0$ et $\lim_{x \to +\infty} F_X(x) = 1$. Elle permet de calculer la probabilité que X appartienne à n'importe quel intervalle de R :

$$\forall (a,b) \in R^2, a < b, P(a < X \le b) = F_X(b) - F_x(a)$$

Les variables aléatoires peuvent être classées selon le type d'ensemble dans lequel elles prennent leurs valeurs. Dans la pratique, on ne s'intéressera qu'à deux catégories : les variables aléatoires discrètes et les variables aléatoires continues (ou à densité).

A.1.2. Variables aléatoires discrètes et continues

Une variable aléatoire X est dite **discrète** (v.a.d.) si et seulement si elle est à valeurs dans un ensemble E fini ou dénombrable. On peut noter $E = \{x_1, x_2, ...\}$.

Exemples :

- Face obtenue lors du lancer d'un dé : $E = \{1,2,3,4,5,6\}$
- Nombre de bugs dans un programme : $E = N$

La loi de probabilité d'une v.a.d. X est entièrement déterminée par les probabilités élémentaires $P(X = x_i)$, $\forall x_i \in E$.

La fonction de répartition de X est alors $F_X(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i)$.

Une variable aléatoire X est dite **continue** (v.a.c.) si et seulement si sa fonction de répartition F_X est partout dérivable. Sa dérivée f_X est alors appelée **densité de probabilité** de X, ou plus simplement densité de X. Une v.a.c. est forcément à valeurs dans un ensemble non dénombrable.

Exemples :

- Appel de la fonction Random d'une calculatrice : $E = [0,1]$
- Durée de bon fonctionnement d'un système : $E = R^+$

On a alors $\forall (a,b) \in R^2, a < b, P(a < X \leq b) = F_X(b) - F_x(a) = \int_a^b f_X(x)dx$.

Plus généralement, $\forall D \subset R, P(\mathrm{X} \in D) = \int_D f_X(x)dx$. Donc la densité détermine entièrement la loi de probabilité de X.

f_X est une fonction positive telle que $\int_{-\infty}^{+\infty} f_X(x)dx = P(X \in R) = 1$

Connaissant la loi de X, on est souvent amenés à déterminer celle de $Y = \varphi(X)$. Quand X est discrète, il suffit d'écrire $P(Y = y) = P(\varphi(X) = y)$. Quand X est continue, on commence par déterminer la fonction de répartition de Y en écrivant $F_Y(y) = P(Y \leq y) = P(\varphi(X) \leq y)$, puis on en déduit sa densité par dérivation.

Remarque : Il existe des lois de probabilité de variables aléatoires réelles qui ne sont ni discrètes ni continues. Par exemple, si X est la durée de bon fonctionnement d'un système qui a une probabilité non nulle p d'être en panne à l'instant initial, on a :

$\lim_{x \to 0^-} F_X(x) = 0$ (une durée ne peut pas être négative) et $F_X(0) = P(X \leq 0) = P(X = 0) = p$.

Par conséquent F_X n'est pas continue en 0, donc pas dérivable en 0. La loi de X ne peut donc pas être continue, et elle n'est pas non plus discrète. Ce type de variable aléatoire ne sera pas étudié ici.

A.1.3. Moments d'une variable aléatoire réelle

Si X est une variable aléatoire discrète, son **espérance mathématique** est définie par

$$E(X) = \sum_{x_i \in E} x_i P(X = x_i)$$

Si X est une variable aléatoire continue, son espérance mathématique est définie par

$$E(X) = \int_{-\infty}^{+\infty} x f_X(x) dx$$

Concrètement, $E(X)$ est ce qu'on s'attend à trouver comme moyenne des résultats obtenus si on répète l'expérience un grand nombre de fois. Par exemple, si on lance une pièce de monnaie 10 fois, on s'attend à trouver en moyenne 5 piles.

Plus généralement, on peut s'intéresser à l'espérance mathématique d'une fonction de X :

Si X est une v.a.d., $E[\varphi(X)] = \sum_{x_i \in E} \varphi(x_i) P(X = x_i)$

Si X est une v.a.c., $E[\varphi(X)] = \int_{-\infty}^{+\infty} \varphi(x) f_X(x) dx$

Ce résultat permet de calculer l'espérance de $\varphi(X)$ sans avoir à déterminer entièrement sa loi.

Soit k un entier naturel quelconque. Le **moment d'ordre *k*** de X est $E(X^k)$ et le **moment centré d'ordre *k*** est $E\left[(X - E(X))^k\right]$.

De tous les moments, le plus important est le moment centré d'ordre 2, appelé aussi variance :

La **variance** de X est $Var(X) = E\left[(X - E(X))^2\right]$, qui se calcule plus facilement sous la forme $Var(X) = E(X^2) - [E(X)]^2$.

L'**écart-type** de X est $\sigma(X) = \sqrt{Var(X)}$.

La variance et l'écart-type sont des indicateurs de la dispersion de X : plus la variance de X est petite, plus les réalisations de X seront concentrées autour de son espérance.

Le **coefficient de variation** de X est $CV(X) = \dfrac{\sigma(X)}{E(X)}$. C'est également un indicateur de dispersion, dont l'avantage est d'être sans dimension. Il permet de comparer les dispersions de variables aléatoires d'ordres de grandeur différents ou exprimées dans des unités différentes. En pratique, on considère que, quand $CV(X)$ est inférieur à 15%, l'espérance peut être considérée comme un bon résumé de la loi.

Soit $p \in]0,1[$. Si F_X est inversible, le **quantile d'ordre *p*** de X est $q_p = F_X^{-1}(p)$.

A.2. Vecteurs aléatoires réels

On ne s'intéressera ici qu'aux vecteurs aléatoires $(X_1,...,X_n)$ constitués de n variables aléatoires réelles toutes discrètes ou toutes continues.

A.2.1. Loi de probabilité d'un vecteur aléatoire

La loi d'un vecteur aléatoire $(X_1,...,X_n)$ est déterminée par sa fonction de répartition :

$$\begin{aligned} F_{(X_1,...,X_n)} : \quad R^n &\to [0,1] \\ (x_1,...,x_n) &\mapsto F_{(X_1,...,X_n)}(x_1,...,x_n) = P([X_1 \le x_1] \cap ... \cap [X_n \le x_n]) \end{aligned}$$

Si les X_i sont discrètes, la loi de $(X_1,...,X_n)$ est aussi déterminée par les probabilités élémentaires $P([X_1 = x_1] \cap ... \cap [X_n = x_n])$.

Si les X_i sont continues, la densité de $(X_1,...,X_n)$ est définie, si elle existe, par :

$$f_{(X_1,...,X_n)}(x_1,...,x_n) = \frac{\partial^n}{\partial x_1 ... \partial x_n} F_{(X_1...X_n)}(x_1,...,x_n)$$

On a alors $\forall D \subset R^n$, $P((X_1,...,X_n) \in D) = \int ... \int_D f_{(X_1,...,X_n)}(x_1,...,x_n) dx_1 ... dx_n$.

Les variables aléatoires $X_1,...,X_n$ sont (mutuellement) **indépendantes** si et seulement si :

$$F_{(X_1,...,X_n)}(x_1,...,x_n) = \prod_{i=1}^{n} P(X_i \le x_i)$$

Pour des variables discrètes cela donne $P([X_1 = x_1] \cap ... \cap [X_n = x_n]) = \prod_{i=1}^{n} P(X_i = x_i)$.

Et pour des variables continues, $f_{(X_1,...,X_n)}(x_1,...,x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$.

Concrètement, l'indépendance signifie que la valeur prise par l'une des variables n'a aucune influence sur la valeur prise par les autres.

A.2.2. Espérance et matrice de covariance d'un vecteur aléatoire

L'**espérance mathématique** d'un vecteur aléatoire $(X_1,...,X_n)$ est le vecteur des espérances mathématiques de ses composantes : $E((X_1,...,X_n)) = (E(X_1),...,E(X_n))$.

L'équivalent de la variance en dimension n est la **matrice de covariance** du vecteur $(X_1,...,X_n)$, notée $K_{(X_1,...,X_n)}$ ou K, dont l'élément $i \times j$ est $k_{ij} = Cov(X_i, X_j)$, $\forall (i,j) \in \{1,...,n\}^2$.

$Cov(X_i, X_j)$ est la **covariance** des variables aléatoires X_i et X_j et est définie par :

$$Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$$

Pour $i = j$, $Cov(X_i, X_i) = E(X_i^2) - E(X_i)^2 = Var(X_i)$.

Pour $i \neq j$, la covariance de X_i et X_j traduit le degré de corrélation entre ces deux variables. En particulier, si X_i et X_j sont indépendantes, $Cov(X_i, X_j) = 0$ (mais la réciproque est fausse). Par conséquent, si $X_1,...,X_n$ sont indépendantes, leur matrice de covariance K est diagonale.

Le **coefficient de corrélation linéaire** entre X_i et X_j est $\rho(X_i, X_j) = \dfrac{Cov(X_i, X_j)}{\sigma(X_i)\sigma(X_j)}$. On montre que :

- $\rho(X_i, X_j) \in [-1,+1]$
- $\rho(X_i, X_j) = +1 \Leftrightarrow X_i = aX_j + b,\ a > 0,\ b \in R$
- $\rho(X_i, X_j) = -1 \Leftrightarrow X_i = -aX_j + b,\ a > 0,\ b \in R$
- si $\rho(X_i, X_j) > 0$, X_i et X_j sont corrélées positivement, ce qui signifie qu'elles varient dans le même sens. Par exemple, X_i et X_j peuvent être la taille et le poids d'individus pris au hasard
- si $\rho(X_i, X_j) < 0$, X_i et X_j sont corrélées négativement, ce qui signifie qu'elles varient en sens contraire. Par exemple, X_i et X_j peuvent être l'âge et la résistance d'un matériau
- si $\rho(X_i, X_j) = 0$, il n'y a pas de corrélation linéaire entre X_i et X_j. Cela ne signifie pas que X_i et X_j sont indépendantes. Il peut éventuellement y avoir une corrélation non linéaire

L'espérance mathématique est linéaire : si X et Y sont des variables aléatoires et a, b et c des réels, alors $E(aX + bY + c) = aE(X) + bE(Y) + c$.

En revanche, la variance n'est pas linéaire : si X et Y sont des variables aléatoires et a, b et c des réels, alors $Var(aX + bY + c) = a^2 Var(X) + 2abCov(X, Y) + b^2 Var(Y)$.

Si X et Y sont indépendantes, $Cov(X_i, X_j) = 0$, donc $Var(aX + bY + c) = a^2 Var(X) + b^2 Var(Y)$. En particulier, la variance de la somme de variables aléatoires indépendantes est égale à la somme des variances de ces variables. Mais ce résultat est faux si les variables ne sont pas indépendantes.

A.3. Convergences et applications

Deux des résultats les plus importants des probabilités sont le théorème central-limite et la loi des grands nombres. Ces résultats nécessitent d'utiliser la notion de convergence d'une suite de variables aléatoires.

Une suite de variables aléatoires $(X_n)_{n \geq 1}$ **converge en loi** vers la loi de probabilité de fonction de répartition F si et seulement si $\lim_{n \to \infty} F_{X_n}(x) = F(x)$ en tout point x où F est continue. Cela signifie

que, quand n est grand, la loi de probabilité de X_n est approximativement la loi de fonction de répartition F.

Théorème central-limite : Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes et de même loi d'espérance m et de variance σ^2 finies. Alors la suite de variables aléatoires $\dfrac{\sum_{i=1}^{n} X_i - nm}{\sigma\sqrt{n}}$ converge en loi vers la loi normale centrée réduite $N(0,1)$.

Concrètement, cela signifie que la loi de toute variable aléatoire égale à la somme d'un nombre « suffisamment grand » de variables aléatoires indépendantes et de même loi est approximativement une loi normale. Plus précisément, pour n grand, $\sum_{i=1}^{n} X_i$ est approximativement de loi $N(nm, n\sigma^2)$.
Ce qui est remarquable, c'est que ce résultat est vrai quelle que soit la loi des X_i.

De très nombreux phénomènes naturels sont la résultante d'un grand nombre de phénomènes élémentaires identiques, indépendants et additifs ce qui justifie l'importance (et le nom) de la loi normale.

La plus forte des convergences de suites de variables aléatoires est la convergence presque sûre. Ce concept nécessite d'avoir défini une variable aléatoire comme une application mesurable d'un espace probabilisé dans un autre. Une suite de variables aléatoires $(X_n)_{n\geq 1}$ **converge presque sûrement** vers la variable aléatoire X si et seulement si $P\left(\left\{\omega;\ \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1$.

Une suite de variables aléatoires $(X_n)_{n\geq 1}$ **converge en probabilité** vers la variable aléatoire X si et seulement si $\forall \varepsilon > 0,\ \lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$.

On montre que la convergence presque sûre entraîne la convergence en probabilité, qui elle-même entraîne la convergence en loi.

Loi des grands nombres : Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes et de même loi d'espérance m. Alors la suite des variables aléatoires $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ converge presque sûrement vers m.

Concrètement, cela signifie que quand on fait un très grand nombre d'expériences identiques et indépendantes, la moyenne des réalisations de la variable aléatoire à laquelle on s'intéresse tend vers l'espérance de sa loi.
Ce résultat permet de justifier l'idée naturelle d'estimer une espérance par une moyenne et une probabilité par une proportion.

En fait, la convergence la plus utile en statistique est la convergence en moyenne quadratique ou dans L^2. L^2 est l'ensemble des variables aléatoires réelles X telles que $E(X^2) < \infty$. Une suite de variables aléatoires $(X_n)_{n\geq 1}$ de L^2 **converge en moyenne quadratique** vers la variable aléatoire X si et seulement si $\lim_{n\to\infty} E(|X_n - X|^2) = 0$.

On montre que la convergence en moyenne quadratique entraîne la convergence en probabilité, qui elle-même entraîne la convergence en loi. Mais il n'y a pas de lien entre la convergence en moyenne quadratique et la convergence presque sûre.

A.4. Quelques résultats sur quelques lois de probabilité usuelles

Les tables de lois de probabilité fournies donnent notamment, pour les lois les plus usuelles, les probabilités élémentaires ou la densité, l'espérance et la variance. On présente dans cette section quelques propriétés supplémentaires de quelques unes de ces lois.

A.4.1. Loi binomiale

Une variable aléatoire K est de loi binomiale $B(n, p)$ si et seulement si elle est à valeurs dans $\{0,1,...,n\}$ et $P(K = k) = C_n^k p^k (1-p)^{n-k}$.

Le nombre de fois où, en n expériences identiques et indépendantes, un évènement de probabilité p s'est produit, est une variable aléatoire de loi $B(n, p)$.

Si $X_1,...,X_n$ sont indépendantes et de même loi $B(m, p)$, alors $\sum_{i=1}^{n} X_i$ est de loi $B(nm, p)$.

A.4.2. Loi géométrique

Une variable aléatoire K est de loi géométrique $G(p)$ si et seulement si elle est à valeurs dans N^* et $P(K = k) = p(1-p)^{k-1}$.

Dans une suite d'expériences identiques et indépendantes, le nombre d'expériences nécessaires pour que se produise pour la première fois un évènement de probabilité p, est une variable aléatoire de loi $G(p)$.

A.4.3. Loi de Poisson

Une variable aléatoire K est de loi de Poisson $P(\lambda)$ si et seulement si elle est à valeurs dans N et $P(K = k) = e^{-\lambda} \frac{\lambda^k}{k!}$.

Pour $n \geq 50$ et $p \leq 0.1$, la loi binomiale $B(n, p)$ peut être approchée par la loi de Poisson $P(np)$. On dit que la loi de Poisson est la loi des évènements rares : loi du nombre de fois où un événement de

probabilité très faible se produit au cours d'un très grand nombre d'expériences identiques et indépendantes.

Si $X_1,...,X_n$ sont indépendantes et de même loi $P(\lambda)$, alors $\sum_{i=1}^{n} X_i$ est de loi $P(n\lambda)$.

A.4.4. Loi exponentielle

Une variable aléatoire X est de loi exponentielle $\exp(\lambda)$ si et seulement si elle est à valeurs dans R^+ et $f(x) = \lambda e^{-\lambda x}$.

La loi exponentielle est sans mémoire : $\forall (t,x) \in R^{+2}, P(X > t + x | X > t) = P(X > x)$.

Si $X_1,...,X_n$ sont indépendantes et de même loi $\exp(\lambda)$, alors $\sum_{i=1}^{n} X_i$ est de loi gamma $G(n,\lambda)$.

Si $X_1,...,X_n$ sont indépendantes et de même loi $\exp(\lambda)$, et représentent les durées entre occurrences successives d'un même événement, alors le nombre d'évènements survenus sur une période de longueur t est une variable aléatoire de loi de Poisson $P(\lambda t)$.

A.4.5. Loi gamma et loi du khi-2

Une variable aléatoire X est de loi gamma $G(\alpha,\lambda)$ si et seulement si elle est à valeurs dans R^+ et $f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1}$. Les propriétés de la fonction gamma sont rappelées sur les tables.

La loi $G(\frac{n}{2},\frac{1}{2})$ est appelée loi du khi-2 à n degrés de liberté, notée χ^2_n.

Si X est de loi $G(\alpha,\lambda)$ et a est un réel strictement positif, alors aX est de loi $G(\alpha,\frac{\lambda}{a})$.

Si X et Y sont des variables aléatoires indépendantes de lois respectives $G(\alpha,\lambda)$ et $G(\beta,\lambda)$, alors $X+Y$ est de loi $G(\alpha+\beta,\lambda)$. En particulier, si X et Y sont indépendantes et de lois respectives χ^2_n et χ^2_m, alors $X+Y$ est de loi χ^2_{n+m}.

A.4.6. Loi normale

Une variable aléatoire X est de loi normale $N(m,\sigma^2)$ si et seulement si elle est à valeurs dans R et $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-m)^2}{2\sigma^2}}$.

Si X est de loi $N(m,\sigma^2)$, alors $aX+b$ est de loi $N(am+b,a^2\sigma^2)$. En particulier, $\frac{X-m}{\sigma}$ est de loi $N(0,1)$.

Si X est de loi $N(0,1)$, alors X^2 est de loi χ_1^2.

Si $X_1,...,X_n$ sont indépendantes et de même loi $N(m,\sigma^2)$, alors :

- $\sum_{i=1}^{n} X_i$ est de loi $N(nm, n\sigma^2)$
- $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ est de loi $N(m, \frac{\sigma^2}{n})$
- $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - m)^2$ est de loi χ_n^2
- $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ est de loi χ_{n-1}^2

Si (X_1, X_2) est un vecteur gaussien tel que X_1 est de loi $N(m_1, \sigma_1^2)$ et X_2 est de loi $N(m_2, \sigma_2^2)$, alors $aX_1 + bX_2$ est de loi $N\left(am_1 + bm_2, a^2\sigma_1^2 + b^2\sigma_2^2 + 2abCov(X_1, X_2)\right)$.

Enfin, les lois de probabilité de Student et de Fisher-Snedecor sont très utilisées en statistique. Elles sont liées à la loi normale à travers les résultats suivants.

Soit U une variable aléatoire de loi $N(0,1)$ et X une variable aléatoire de loi χ_n^2. Si U et X sont indépendantes, alors $\frac{U}{\sqrt{X}}\sqrt{n}$ est de **loi de Student** à n degrés de liberté $St(n)$.

Soit X une variable aléatoire de loi χ_n^2 et Y une variable aléatoire de loi χ_m^2. Si X et Y sont indépendantes, alors $\frac{mX}{nY}$ est de **loi de Fisher-Snedecor** $F(n,m)$.

Ces deux définitions entrainent que si T est de loi $St(n)$, alors T^2 est de loi $F(1,n)$.

Les lois de Student et de Fisher-Snedecor sont toujours utilisées par l'intermédiaire de tables ou à l'aide d'un logiciel de statistique. Il n'est donc pas nécessaire de donner l'expression de leur densité.

Annexe B : Tables de lois de probabilités usuelles

Ces tableaux présentent les lois de probabilité les plus usuelles pour une variable aléatoire réelle X. Pour chaque loi de probabilité, on donne son nom usuel, son symbole, son support, sa définition à l'aide de probabilités élémentaires pour les lois discrètes ou de densité pour les lois continues, son espérance et sa variance.

Les fonctions spéciales suivantes sont utilisées :

- la fonction Gamma est définie pour $a > 0$ par $\Gamma(a) = \int_0^{+\infty} e^{-x} x^{a-1} dx$.
  Propriétés : $\forall n \in \mathbb{N}^*,\ \Gamma(n) = (n-1)!,\ \Gamma(1) = 1,\ \Gamma(1/2) = \sqrt{\pi},\ \forall a > 1,\ \Gamma(a) = (a-1)\Gamma(a-1)$.
- la fonction Béta est définie pour $a > 0$ et $b > 0$ par $\beta(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1}(1-x)^{b-1} dx$.

Tableau 1 : Variables aléatoires discrètes

Nom	Symbole	Support	Probabilités élémentaires	Espérance	Variance
Loi de Bernoulli $p \in]0,1[$	$B(p)$	$\{0,1\}$	$P(X=0) = 1-p$ $P(X=1) = p$	p	$p(1-p)$
Loi binomiale $p \in]0,1[,\ n \in \mathbb{N}^*$	$B(n,p)$	$\{0,1,\ldots,n\}$	$P(X=k) = C_n^k p^k (1-p)^{n-k}$	np	$np(1-p)$
Loi binomiale négative $p \in]0,1[,\ n \in \mathbb{N}^*$	$BN(n,p)$	$\{n, n+1, \ldots\}$	$P(X=k) = C_{k-1}^{n-1} p^n (1-p)^{k-n}$	$\dfrac{n}{p}$	$\dfrac{n(1-p)}{p^2}$
Loi de Poisson $\lambda \in \mathbb{R}^{+*}$	$P(\lambda)$	$\mathbb{N}$	$P(X=k) = e^{-\lambda} \dfrac{\lambda^k}{k!}$	λ	λ
Loi géométrique $p \in]0,1[$	$G(p)$	$\mathbb{N}^*$	$P(X=k) = p(1-p)^{k-1}$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$
Loi hypergéométrique $N \in \mathbb{N}^*,\ (m,n) \in \{1,\ldots,N\}^2$	$H(N,m,n)$	$\{0,\ldots,\min(m,n)\}$	$P(X=k) = \dfrac{C_m^k C_{N-m}^{n-k}}{C_N^n}$	$\dfrac{nm}{N}$	$\dfrac{nm(N-n)(N-m)}{N^2(N-1)}$

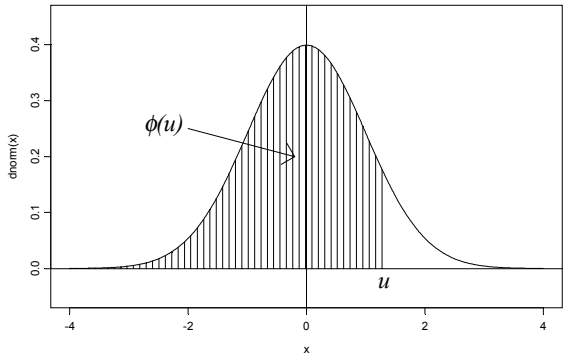
Tableau 2 : Variables aléatoires continues

Nom	Symbole	Support	Densité	Espérance	Variance
Loi uniforme $[a,b]\subset\mathbb{R}$	$U[a,b]$	$[a,b]$	$f_X(x)=\frac{1}{b-a}\ 1_{[a,b]}(x)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Loi normale ou de Gauss $m\in\mathbb{R},\ \sigma\in\mathbb{R}^{+*}$	$N(m,\sigma^2)$	$\mathbb{R}$	$f_X(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}$	m	σ^2
Loi gamma $\alpha\in\mathbb{R}^{+*},\ \lambda\in\mathbb{R}^{+*}$	$G(\alpha,\lambda)$	$\mathbb{R}^+$	$f_X(x)=\frac{\lambda^\alpha}{\Gamma(\alpha)}e^{-\lambda x}x^{\alpha-1}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Loi exponentielle $\lambda\in\mathbb{R}^{+*}$	$\exp(\lambda)$ $=G(1,\lambda)$	$\mathbb{R}^+$	$f_X(x)=\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Loi du chi-deux $n\in\mathbb{N}^*$	χ^2_n $=G\left(\frac{n}{2},\frac{1}{2}\right)$	$\mathbb{R}^+$	$f_X(x)=\frac{1}{\Gamma\left(\frac{n}{2}\right)2^{\frac{n}{2}}}e^{-\frac{x}{2}}x^{\frac{n}{2}-1}$	n	$2n$
Loi béta de 1ère espèce $a\in\mathbb{R}^{+*},\ b\in\mathbb{R}^{+*}$	$\beta_1(a,b)$	$[0,1]$	$f_X(x)=\frac{1}{\beta(a,b)}x^{a-1}(1-x)^{b-1}\ 1_{[0,1]}(x)$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$
Loi béta de 2ème espèce $a\in\mathbb{R}^{+*},\ b\in\mathbb{R}^{+*}$	$\beta_2(a,b)$	$\mathbb{R}^+$	$f_X(x)=\frac{1}{\beta(a,b)}\frac{x^{a-1}}{(1+x)^{a+b}}$	$\frac{a}{b-1}$ si $b>1$	$\frac{a(a+b-1)}{(b-1)^2(b-2)}$ si $b>2$
Loi de Weibull $\eta\in\mathbb{R}^{+*},\ \beta\in\mathbb{R}^{+*}$	$W(\eta,\beta)$	$\mathbb{R}^+$	$f_X(x)=\frac{\beta}{\eta^\beta}x^{\beta-1}e^{-\left(\frac{x}{\eta}\right)^\beta}$	$\eta\Gamma\left(1+\frac{1}{\beta}\right)$	$\eta^2\left[\Gamma\left(1+\frac{2}{\beta}\right)-\Gamma\left(1+\frac{1}{\beta}\right)^2\right]$

TABLE 1 DE LA LOI NORMALE CENTREE REDUITE

U étant une variable aléatoire de loi $N(0,1)$, la table donne la valeur de $\Phi(u) = P(U \leq u)$.

Sous S+, la commande correspondante est `pnorm(u)`.



u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

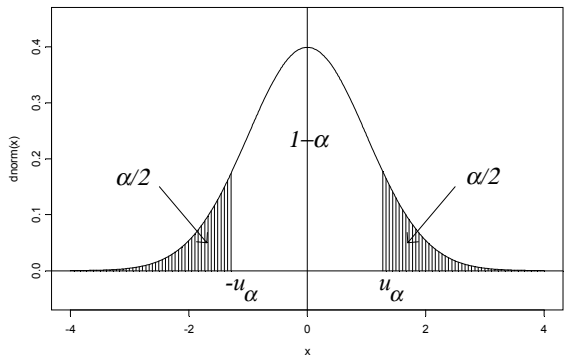
Grandes valeurs de u

u	3.0	3.5	4.0	4.5
$\Phi(u)$	0.9987	0.99977	0.999968	0.999997

TABLE 2 DE LA LOI NORMALE CENTREE REDUITE

U étant une variable aléatoire de loi $N(0,1)$ et α un réel de [0,1], la table donne la valeur $u_\alpha = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$, telle que $P(|U| > u_\alpha) = \alpha$.

Sous S+, la commande correspondante est `qnorm(1-alpha/2)`.



α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	$+\infty$	2.5758	2.3263	2.1701	2.0537	1.9600	1.8808	1.8119	1.7507	1.6954
0.1	1.6449	1.5982	1.5548	1.5141	1.4758	1.4395	1.4051	1.3722	1.3408	1.3106
0.2	1.2816	1.2536	1.2265	1.2004	1.1750	1.1503	1.1264	1.1031	1.0803	1.0581
0.3	1.0364	1.0152	0.9945	0.9741	0.9542	0.9346	0.9154	0.8965	0.8779	0.8596
0.4	0.8416	0.8239	0.8064	0.7892	0.7722	0.7554	0.7388	0.7225	0.7063	0.6903
0.5	0.6745	0.6588	0.6433	0.6280	0.6128	0.5978	0.5828	0.5681	0.5534	0.5388
0.6	0.5244	0.5101	0.4959	0.4817	0.4677	0.4538	0.4399	0.4261	0.4125	0.3989
0.7	0.3853	0.3719	0.3585	0.3451	0.3319	0.3186	0.3055	0.2924	0.2793	0.2663
0.8	0.2533	0.2404	0.2275	0.2147	0.2019	0.1891	0.1764	0.1637	0.1510	0.1383
0.9	0.1257	0.1130	0.1004	0.0878	0.0753	0.0627	0.0502	0.0376	0.0251	0.0125

Petites valeurs de α

α	0.002	0.001	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}
u_α	3.0902	3.2905	3.8906	4.4171	4.8916	5.3267	5.7307	6.1094

Pour $p < \frac{1}{2}$, $\Phi^{-1}(p) = -u_{2p}$

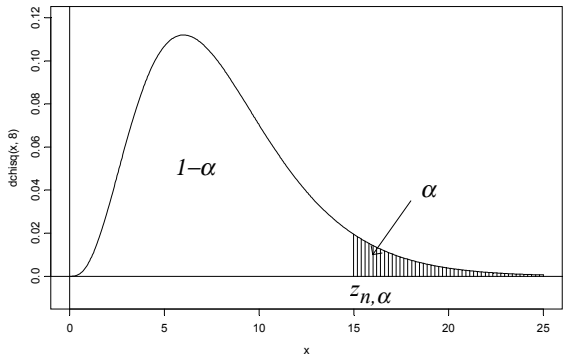
Pour $p \geq \frac{1}{2}$, $\Phi^{-1}(p) = u_{2(1-p)}$

TABLE DE LA LOI DU χ^2

X étant une variable aléatoire de loi du χ^2 à n degrés de liberté, et α un réel de [0,1],

la table donne la valeur $z_{n,\alpha} = F^{-1}_{\chi^2_n}(1-\alpha)$, telle que $P(X > z_{n,\alpha}) = \alpha$.

Sous S+, la commande correspondante est `qchisq(1-alpha,n)`.



n \ α	0.995	0.990	0.975	0.95	0.9	0.8	0.7	0.5	0.3	0.2	0.1	0.05	0.025	0.01	0.005	0.001
1	0.00004	0.0002	0.001	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	5.02	6.63	7.88	10.80
2	0.01	0.02	0.05	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.61	5.99	7.38	9.21	10.60	13.82
3	0.07	0.11	0.22	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.81	9.35	11.34	12.84	16.27
4	0.21	0.30	0.48	0.71	1.06	1.65	2.19	3.36	4.88	5.99	7.78	9.49	11.14	13.28	14.86	18.47
5	0.41	0.55	0.83	1.15	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	12.83	15.09	16.75	20.52
6	0.68	0.87	1.24	1.64	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	14.45	16.81	18.55	22.46
7	0.99	1.24	1.69	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	17.53	20.09	21.95	26.12
9	1.73	2.09	2.70	3.33	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	6.99	8.15	10.34	12.90	14.63	17.28	19.68	21.92	24.72	26.76	31.26
12	3.07	3.57	4.40	5.23	6.30	7.81	9.03	11.34	14.01	15.81	18.55	21.03	23.34	26.22	28.30	32.91
13	3.57	4.11	5.01	5.89	7.04	8.63	9.93	12.34	15.12	16.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.07	4.66	5.63	6.57	7.79	9.47	10.82	13.34	16.22	18.15	21.06	23.68	26.12	29.14	31.32	36.12
15	4.60	5.23	6.26	7.26	8.55	10.31	11.72	14.34	17.32	19.31	22.31	25.00	27.49	30.58	32.80	37.70
16	5.14	5.81	6.91	7.96	9.31	11.15	12.62	15.34	18.42	20.47	23.54	26.30	28.85	32.00	34.27	39.25
17	5.70	6.41	7.56	8.67	10.09	12.00	13.53	16.34	19.51	21.61	24.77	27.59	30.19	33.41	35.72	40.79
18	6.26	7.01	8.23	9.39	10.86	12.86	14.44	17.34	20.60	22.76	25.99	28.87	31.53	34.81	37.16	42.31
19	6.84	7.63	8.91	10.12	11.65	13.72	15.35	18.34	21.69	23.90	27.20	30.14	32.85	36.19	38.58	43.82
20	7.43	8.26	9.59	10.85	12.44	14.58	16.27	19.34	22.77	25.04	28.41	31.41	34.17	37.57	40.00	45.31
21	8.03	8.90	10.28	11.59	13.24	15.44	17.18	20.34	23.86	26.17	29.62	32.67	35.48	38.93	41.40	46.80
22	8.64	9.54	10.98	12.34	14.04	16.31	18.10	21.34	24.94	27.30	30.81	33.92	36.78	40.29	42.80	48.27
23	9.26	10.20	11.69	13.09	14.85	17.19	19.02	22.34	26.02	28.43	32.01	35.17	38.08	41.64	44.18	49.73
24	9.89	10.86	12.40	13.85	15.66	18.06	19.94	23.34	27.10	29.55	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	18.94	20.87	24.34	28.17	30.68	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	19.82	21.79	25.34	29.25	31.79	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	20.70	22.72	26.34	30.32	32.91	36.74	40.11	43.19	46.96	49.64	55.48
28	12.46	13.56	15.31	16.93	18.94	21.59	23.65	27.34	31.39	34.03	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	22.48	24.58	28.34	32.46	35.14	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	23.36	25.51	29.34	33.53	36.25	40.26	43.77	46.98	50.89	53.67	59.70

Pour $n > 30$, on admet que : $z_{n,\alpha} \approx \frac{1}{2}\left(u_{2\alpha} + \sqrt{2n-1}\right)^2$ si $\alpha < \frac{1}{2}$

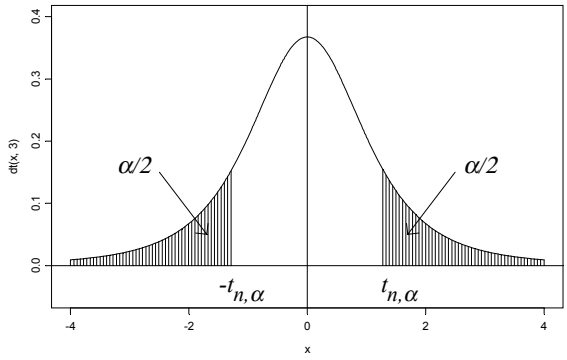
$$z_{n,\alpha} \approx \frac{1}{2}\left(\sqrt{2n-1} - u_{2(1-\alpha)}\right)^2 \text{ si } \alpha \geq \frac{1}{2}$$

TABLE DE LA LOI DE STUDENT

X étant une variable aléatoire de loi $St(n)$ et α un réel de [0,1],

la table donne la valeur $t_{n,\alpha} = F^{-1}_{St(n)}\left(1-\frac{\alpha}{2}\right)$ telle que $P(|X| > t_{n,\alpha}) = \alpha$.

Sous S+, la commande correspondante est `qt(1-alpha/2,n)`. $t_{+\infty,\alpha} = u_\alpha$



α / n	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.62
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
80	0.126	0.254	0.387	0.527	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.416
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
$+\infty$	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

TABLES DE LA LOI DE FISHER-SNEDECOR

X étant une variable aléatoire de loi $F(\nu_1,\nu_2)$, les tables donnent les valeurs $f_{\nu_1,\nu_2,\alpha} = F^{-1}_{F(\nu_1,\nu_2)}(1-\alpha)$ telles que $P(X > f_{\nu_1,\nu_2,\alpha}) = \alpha$ pour $\alpha = 5\%$ et $\alpha = 1\%$.

Sous S+, la commande correspondante est `qf(1-alpha,nu1,nu2)`. $f_{\nu_2,\nu_1,\alpha} = \dfrac{1}{f_{\nu_1,\nu_2,1-\alpha}}$

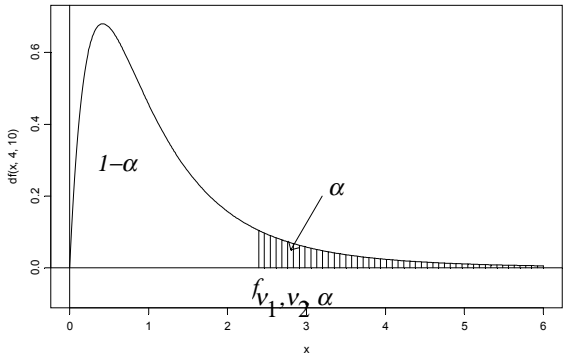


Table 1 : $\alpha = 5\%$

ν_1 / ν_2	1	2	3	4	5	6	7	8	10	12	16	20	24	40	60	100	$+\infty$
1	161.5	199.5	215.7	224.6	230.2	234.0	236.8	238.9	241.9	243.9	246.5	248.0	249.1	251.1	252.2	253.0	254.2
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.40	19.41	19.43	19.45	19.45	19.47	19.48	19.49	19.49
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.79	8.74	8.69	8.66	8.64	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.84	5.80	5.77	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.60	4.56	4.53	4.46	4.43	4.41	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.06	4.00	3.92	3.87	3.84	3.77	3.74	3.71	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.64	3.57	3.49	3.44	3.41	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.35	3.28	3.20	3.15	3.12	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.14	3.07	2.99	2.94	2.90	2.83	2.79	2.76	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.98	2.91	2.83	2.77	2.74	2.66	2.62	2.59	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.85	2.79	2.70	2.65	2.61	2.53	2.49	2.46	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.75	2.69	2.60	2.54	2.51	2.43	2.38	2.35	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.67	2.60	2.51	2.46	2.42	2.34	2.30	2.26	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.60	2.53	2.44	2.39	2.35	2.27	2.22	2.19	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.54	2.48	2.38	2.33	2.29	2.20	2.16	2.12	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.49	2.42	2.33	2.28	2.24	2.15	2.11	2.07	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.45	2.38	2.29	2.23	2.19	2.10	2.06	2.02	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.41	2.34	2.25	2.19	2.15	2.06	2.02	1.98	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.38	2.31	2.21	2.16	2.11	2.03	1.98	1.94	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.35	2.28	2.18	2.12	2.08	1.99	1.95	1.91	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.32	2.25	2.16	2.10	2.05	1.96	1.92	1.88	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.30	2.23	2.13	2.07	2.03	1.94	1.89	1.85	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.27	2.20	2.11	2.05	2.01	1.91	1.86	1.82	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.25	2.18	2.09	2.03	1.98	1.89	1.84	1.80	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.24	2.16	2.07	2.01	1.96	1.87	1.82	1.78	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.16	2.09	1.99	1.93	1.89	1.79	1.74	1.70	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.08	2.00	1.90	1.84	1.79	1.69	1.64	1.59	1.51
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.03	1.95	1.85	1.78	1.74	1.63	1.58	1.52	1.44
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	1.99	1.92	1.82	1.75	1.70	1.59	1.53	1.48	1.39
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	1.95	1.88	1.77	1.70	1.65	1.54	1.48	1.43	1.32
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.93	1.85	1.75	1.68	1.63	1.52	1.45	1.39	1.28
$+\infty$	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.83	1.75	1.64	1.57	1.52	1.39	1.32	1.24	1.00

Table 2 : $\alpha = 1\%$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	10	12	16	20	24	40	60	100	$+\infty$
1	4052	4999	5403	5624	5764	5859	5928	5981	6056	6106	6170	6209	6235	6287	6313	6334	6368
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5
3	34.1	30.9	29.5	28.7	28.2	27.9	27.7	27.5	27.2	27.1	26.8	26.7	26.6	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.6	14.4	14.2	14.0	13.9	13.8	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.0	9.89	9.68	9.55	9.47	9.29	9.20	9.13	9.02
6	13.8	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.87	7.72	7.52	7.40	7.31	7.14	7.06	6.99	6.88
7	12.3	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.62	6.47	6.28	6.16	6.07	5.91	5.82	5.75	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.81	5.67	5.48	5.36	5.28	5.12	5.03	4.96	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.26	5.11	4.92	4.81	4.73	4.57	4.48	4.41	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.85	4.71	4.52	4.41	4.33	4.17	4.08	4.01	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.54	4.40	4.21	4.10	4.02	3.86	3.78	3.71	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.30	4.16	3.97	3.86	3.78	3.62	3.54	3.47	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.10	3.96	3.78	3.66	3.59	3.43	3.34	3.27	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	3.94	3.80	3.62	3.51	3.43	3.27	3.18	3.11	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.80	3.67	3.49	3.37	3.29	3.13	3.05	2.98	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.69	3.55	3.37	3.26	3.18	3.02	2.93	2.86	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.59	3.46	3.27	3.16	3.08	2.92	2.83	2.76	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.51	3.37	3.19	3.08	3.00	2.84	2.75	2.68	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.43	3.30	3.12	3.00	2.92	2.76	2.67	2.60	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.37	3.23	3.05	2.94	2.86	2.69	2.61	2.54	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.31	3.17	2.99	2.88	2.80	2.64	2.55	2.48	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.26	3.12	2.94	2.83	2.75	2.58	2.50	2.42	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.21	3.07	2.89	2.78	2.70	2.54	2.45	2.37	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.17	3.03	2.85	2.74	2.66	2.49	2.40	2.33	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.13	2.99	2.81	2.70	2.62	2.45	2.36	2.29	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	2.98	2.84	2.66	2.55	2.47	2.30	2.21	2.13	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.80	2.66	2.48	2.37	2.29	2.11	2.02	1.94	1.80
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.70	2.56	2.38	2.27	2.18	2.01	1.91	1.82	1.68
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.63	2.50	2.31	2.20	2.12	1.94	1.84	1.75	1.60
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.55	2.42	2.23	2.12	2.03	1.85	1.75	1.65	1.49
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.50	2.37	2.19	2.07	1.98	1.80	1.69	1.60	1.43
$+\infty$	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.32	2.18	2.00	1.88	1.79	1.59	1.47	1.36	1.00