Département IF
Cours de quatrième année

INSA de Lyon

Introduction à la statistique

I. Gannaz

Table des matières

Introduction

Les statistiques sont rencontrées dans de très nombreux domaines. Pour n'en citer que quelques uns :

- en sciences humaines, la réalisation et l'étude de sondages permettent par exemple d'analyser le positionnement politique d'une population et de prédire les résultats à de futures élections ; des études statistiques (notamment avec le recensement) offrent aussi une vision des données socio-économiques de la population : niveau de vie, satisfaction globale, etc,
- en économie, les prévisions économétriques (quelle sera la croissance du PIB au prochain semestre ?) sont indispensables,
- en biologie, le diagnostic déduit d'un test sanguin par exemple est souvent issu de développements statistiques ; par exemple une femme enceinte effectuant un test sur la trisomie verra apparaître des *p-valeurs* (cf le cours sur les tests),
- les études des données météorologiques ou sismographiques font appel aux statistiques,
- un ingénieur est confronté de manière très importante aux statistiques : évaluation de la performance d'une méthode développée, contrôle de la qualité de composants, étude de la fiabilité d'un système, analyse de la sensibilité d'un code informatique, etc.

L'idée de la statistique est de fournir des méthodologies rigoureuses d'études des données en présence d'incertitude. Cette incertitude peut avoir diverses sources : les mesures sont entachées d'erreur (contrairement à une pensée communément répandue, ceci est quasi-systématique, même avec des appareils considérés comme *fiables*), elles sont issues de phénomènes inconnus que l'on souhaiterait éventuellement modéliser pour mieux les comprendre et surtout les mesures ne sont en général pas exhaustives : on n'a accès à des données que sur une partie de la population et l'on voudrait déduire de nos observations des résultats généraux pour l'intégralité de la population. Par exemple, si vous mesurez la performance d'un code en terme de temps de calcul :

— le temps de calcul mesuré par votre ordinateur peut être incertain en raison par exemple de tâches subalternes qu'effectue votre ordinateur,
— vous ne pouvez pas lancer indéfiniment votre code et vous allez devoir en déduire sa performance générale pour la comparer à un ancien code (peut-être votre code est-il plus rapide pour certains jeux de paramètres et pas pour d'autres : qu'en est-il globalement ?)
— votre code est complexe et vous ne savez pas quels paramètres sont les plus critiques…

Dans de tels contextes, le but des statistiques est multiple : premièrement décrire les observations, résumer les données à votre disposition ; deuxièmement essayer de comprendre les phénomènes sous-jacents en introduidant des modèles permettant d'analyser ces phénomènes ; cette modélisation permet dans un troisième temps de faire des prédictions sur le comportement du système étudié dans des contextes non encore observés et de mettre en place des procédures d'aide à la décision.

Nous allons dans ce cours développer les méthologies qu'offre la statistique pour répondre à ces différents points. Dans un premier temps, nous allons définir le contexte dans lequel sont développées ces techniques. Le fondement de la statistique étant d'analyser des données en présence d'incertitude, ces incertitudes sont modélisées en faisant appel à la théorie des probabilités. Nous faisons donc dans un premier temps quelques (très) brefs rappels de probabilité, afin de rafraîchir le cours de l'année dernière. Dans un second temps, nous présentons quelques éléments dits de statistique descriptive, permettant de décrire un jeu de données. Nous développerons ensuite des outils de modélisation : estimation ponctuelle et intervalles de confiance, puis des tests permettant de prendre des décisions au vu de cette modélisation et de vérifier la justesse de cette modélisation.

CHAPITRE 1

RAPPELS DE PROBABILITÉ

Ce chapître reprend brièvement quelques éléments de la théorie des probabilités que vous avez vu auparavant. Nous essayons de redonner les principaux outils qui serviront dans ce cours de manière informelle. Nous vous renvoyons à votre cours de probabilité pour des définitions plus rigoureuses.

1.1 Variable aléatoire

Une variable aléatoire X désigne le résultat d'une expérience aléatoire, telle le résultat d'un lancer de dé, la durée de vie d'une clef usb, l'intention de vote aux prochaines élections, etc. Il existe différents types de variables :

Les variables quantitatives. Parmi celles-ci, on peut distinguer les variables discrètes, qui ont un nombre fini de modalités, des variables continues :

- **discrètes :** résultat d'un tirage de dé, nombre de bugs dans un programme, nombre de contrôles dans le tramway sur une année…
- **continues :** âge d'un étudiant, durée de vie d'un système, taux d'un composant chimique dans une solution…

Les variables qualitatives. Par exemple la couleur des yeux d'une personne, le résultat à la question « Aimez-vous les repas au resto U ? » ou « Pour qui allez-vous voter aux prochaines élections ? »

La première démarche à faire face à un problème donné est d'identifier quelle est la variable aléatoire que l'on va considérer et de quel type est cette variable. Ensuite, nous allons essayer de caractériser cette variable.

1.1.1 Loi de probabilité

La loi de probabilité d'une variable aléatoire X permet de déterminer comment se répartit X.

Dans le cas d'une variable discrète, notons $\mathcal{X} = \{e_1, e_2, \ldots\}$ l'ensemble des valeurs prises par X. Alors pour caractériser la loi de X on donne l'ensemble des valeurs de $\mathbb{P}(X = e_i)$, pour $i = 1, 2, \ldots$

Si la variable est quantitative, discrète ou continue, on peut caractériser la loi de X par la fonction de répartition :

$$\begin{aligned} F: \quad \mathbb{R} &\to [0,1] \\ x &\mapsto \mathbb{P}(X \leqslant x) \end{aligned}$$

Rappelons qu'alors pour tout (a, b), $\mathbb{P}(X \in [a, b]) = F(b) - F(a)$.

Dans le cas des variables aléatoires continues, on manipule aussi souvent la fonction de densité f, dérivée de F : pour tout (a, b), $\mathbb{P}(X \in [a, b]) = \int_{[a,b]} f(x)dx$.

Nous verrons dans ce cours comment évaluer ces grandeurs ou ces fonctions caractérisant la loi d'une variable à partir d'observations.

1.1.2 Moments et quantiles

Cette section ne concerne que les variables quantitatives.

Espérance

L'espérance est définie par

$$\begin{cases} \mathbb{E}X = \int x f(x)dx & \text{si on a une variable continue,} \\ \mathbb{E}X = \sum_i x_i \mathbb{P}(X = x_i) & \text{si on a une variable discrète.} \end{cases}$$

C'est la valeur que l'on s'attend à avoir en moyenne lorsqu'on répéte un très grand nombre de fois l'expérience (voir la loi des grands nombres en section suivante).

Prenons un exemple. Nous jouons à un jeu de pile ou face. Introduisons $X = \mathbb{1}_{\{\text{on obtient } face\}}$. X suit une loi de Bernoulli de paramètre $1/2 : \mathbb{P}(X = 0) = \mathbb{P}(X = 1) = 1/2$. Nous avons alors $\mathbb{E}X = 1/2$. Si nous lançons 1000 fois une pièce nous nous attendons à avoir environ $1000 * \mathbb{E}X = 500$ fois *face*.

Plus généralement,

$$\begin{cases} \mathbb{E}\varphi(X) = \int \varphi(x) f(x) dx & \text{si on a une variable continue,} \\ \mathbb{E}\varphi(X) = \sum_i \varphi(x_i) \mathbb{P}(X = x_i) & \text{si on a une variable discrète.} \end{cases}$$

Rappelons aussi qu'on a la linéarité de l'espérance :

$$\begin{aligned} \mathbb{E}[X + Y] &= \mathbb{E}X + \mathbb{E}Y, \\ \mathbb{E}[aX] &= a\mathbb{E}X \quad \text{pour tout } a \in \mathbb{R}. \end{aligned}$$

Variance et écart-type

La variance d'une variable aléatoire X est

$$Var(X) = \mathbb{E}[(X - \mathbb{E}X)^2] = \mathbb{E}[X^2] - (\mathbb{E}X)^2.$$

L'écart-type se définit ensuite comme la racine de la variance :

$$\sigma(X) = \sqrt{Var(X)}.$$

Ces grandeurs sont des indicateurs de dispersion : elles mesurent l'écart entre les valeurs prises par X et son espérance $\mathbb{E}X$.

Rappelons les modalités de manipulation de la variance :

$$\begin{aligned} \text{Si X et Y sont indépendantes } Var(X + Y) &= Var(X) + Var(Y). \\ \text{Pour tout réel } a,\ Var(aX) &= a^2 Var(X). \end{aligned}$$

Remarque : L'espérance est le moment d'ordre 1, la variance est le moment centré d'ordre 2. Il existe des moments (centrés ou non) d'ordre p pour tout $p \in N^*$ mais qui ne sont pas présentés ici.

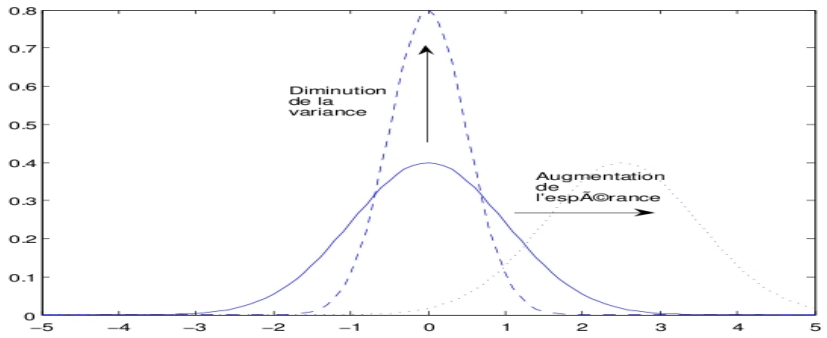

FIGURE 1.1 – Trois densités de loi normale avec différentes valeurs de l'espérance et de la variance.

Quantiles

Le quantile d'ordre α est le réel q_α tel que $\mathbb{P}(X \leqslant q_\alpha) = \alpha$. Par exemple le quantile d'ordre 25%, aussi appelé premier quartile, est le réel $q_{0.25}$ tel que $\mathbb{P}(X \leqslant q_{0.25}) = 25\%$.

Prenons un exemple. Soit X le niveau sonore en décibels d'une voiture. Les autorités ont décidé de taxer les voitures les plus sonores. Elles voudraient taxer 5% des voitures. Alors le seuil qu'elles doivent prendre est le quantile d'ordre 95%, noté $q_{0.95}$: une voiture ayant un niveau sonore supérieur à $q_{0.95}$ décibels sera taxée, mais pas une voiture de niveau inférieur.

La notion de quantile est très utilisée en statistique, comme nous le verrons par la suite.

1.1.3 Convergences

Nous rappelons ici trois notions de convergence ainsi que deux théorèmes fondamentaux de manière très succinte. Nous vous renvoyons à votre cours de probabilité pour plus de détails et pour d'autres types de convergence.

Convergence en loi

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variable aléatoire de fonction de répartition (F_{X_n}) et soit X variable aléatoire de fonction de répartition F. On dit que $(X_n)_n$ converge en loi vers X et on notera $(X_n)_n \xrightarrow[n\to\infty]{loi} X$ si la suite de fonction (F_{X_n}) converge simplement vers F en tout point de continuité de F.

Cela signifie que si n est grand, la variable aléatoire X_n se comporte comme X.

Rappelons le théorème de la limite centrale :

Théorème 1.1. *Soient $X_1, X_2, \dots X_n$ i.i.d. (indépendants et identiquement distribués). Supposons que $m = \mathbb{E}X_i$ et $\sigma^2 = Var(X_i)$ soient finis. Notons $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$. Alors,*

$$\sqrt{n}\frac{\overline{X_n} - m}{\sigma} \xrightarrow[n\to\infty]{loi} \mathcal{N}(0,1).$$

Ce théorème justifie l'importance de la loi normale dans la modélisation statistique : quelque soit la loi des variables aléatoires considérées, leur moyenne se comporte asymptotiquement comme une loi normale. Par exemple, le fait que souvent les résidus dans les modèles de physique soient considérés comme suivant des lois normales est dû au fait qu'on peut les considérer comme la somme de petits phénomènes non pris en compte.

Convergence presque sûre

Soient $(X_n)_{n\in\mathbb{N}}$ et X des variables aléatoires. On dit que $(X_n)_n$ converge presque sûrement vers X et on notera $(X_n)_n \xrightarrow[n\to\infty]{p.s.} X$ si $\mathbb{P}\left(\{w,\, X_n(w) \to X(w)\}\right) = 1$.

La convergence presque sûre implique la convergence en loi.

La loi forte des grands nombres nous dit :

Théorème 1.2. *Soient $X_1, X_2, \dots X_n$ i.i.d. Notons $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$ et $m = \mathbb{E}X_i$. Alors,*

$$\overline{X_n} \xrightarrow[n\to\infty]{p.s.} m.$$

Ce résultat est assez intuitif : plus on réalise une expérience un grand nombre de fois, plus la moyenne observée est proche de la moyenne théorique.

Convergence L^2

Soient $(X_n)_{n\in\mathbb{N}}$ et X des variables aléatoires. On dit que $(X_n)_n$ converge en moyenne quadratique vers X et on notera $(X_n)_n \xrightarrow[n\to\infty]{L^2} X$ si $\mathbb{E}\left((X_n - X)^2\right) \to 0$.

Le principal intérêt de cette convergence est qu'elle est plus facile à étudier en pratique. La convergence en moyenne quadratique implique la convergence en loi, mais aucune implication ne peut être établie avec la convergence presque-sûre.

1.2 Lois usuelles

Nous ne présentons pas ici les lois de manière exhaustive, uniquement quelques unes des lois les plus usuelles.

1.2.1 Variables discrètes

Uniforme discrète $\mathcal{U}(N)$

— Plage des valeurs : $\{1, \ldots, N\}$
— Fonction de masse : $\mathbb{P}(X = k) = 1/N$
— Espérance : $(N+1)/2$ et Variance : $(N^2-1)/12$
— Interprétation : Expérience avec N issues équiprobables possibles.

Bernoulli $\mathcal{B}(p)$

— Plage des valeurs : $\{0, 1\}$
— Fonction de masse : $\mathbb{P}(X = 0) = 1 - p$ et $\mathbb{P}(X = 1) = p$
— Espérance : p et Variance : $p(1-p)$
— Interprétation : Expérience qui n'a que 2 issues possibles (ticket gagant ou perdant)

Binomiale $\mathcal{B}(n, p)$

— Plage des valeurs : $\{0, \ldots, n\}$
— Fonction de masse : $\mathbb{P}(X = k) = C_n^k p^k (1-p)^{n-k}$
— Espérance : np et Variance : $np(1-p)$
— Interprétation : Somme de n Bernouillis indépendantes. Nombre de succès dans n tirages si chaque tirage a une probabilité p d'être gagnant (nombre de tickets gagnats parmi n)

Géométrique $\mathcal{G}(p)$
- — Plage des valeurs : $\{1, \dots \infty\}$
- — Fonction de masse : $\mathbb{P}(X = k) = p(1-p)^{k-1}$
- — Espérance : $1/p$ et Variance : $(1-p)/p^2$
- — Interprétation : Nombre de tirages nécessaires pour obtenir un succès (nombre de tickets à acheter pour en avoir 1 gagant)

Hypergéométrique $\mathcal{H}(N, n, p)$
- — Plage des valeurs : $\{0, \dots, \infty\}$
- — Fonction de masse : $\mathbb{P}(X = k) = \frac{C_{Np}^k C_{N(1-p)}^{n-k}}{C_N^n}$
- — Espérance : np et Variance : $np(1-p)\frac{N-n}{N-1}$
- — Interprétation : Il y a N tickets et chaque ticket a une probabilité p d'être gagnant. On choisit au hasard n tickets. Combien sont gagnants ?

Pascal $\mathcal{P}(r, p)$
- — Plage des valeurs : $\{r, \dots, \infty\}$
- — Fonction de masse : $\mathbb{P}(X = k) = C_{k-1}^{r-1} p^r (1-p)^{k-r}$
- — Espérance : r/p et Variance : $r(1-p)/p^2$
- — Interprétation : On a observé r succès. Sachant que la probabilité d'avoir un succès est p, combien de réalisations y a-t-il eu ?

Poisson $\mathcal{P}(\lambda)$
- — Plage des valeurs : $\{0, \dots \infty\}$
- — Fonction de masse : $\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$
- — Espérance : λ et Variance : λ
- — Interprétation : Loi des évènements rares : nombre de fois où un évenement ayant une faible probabilité de se réaliser va être observé sur un très grand nombre d'expériences.

1.2.2 Variables continues

Uniforme $\mathcal{U}([a, b])$
- — Plage des valeurs : $[a, b]$
- — Densité : $f(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x)$ où $\mathbb{1}$ désigne la fonction indicatrice ($\mathbb{1}_A(x) = 1$ si $x \in A$ et 0 sinon).
- — Espérance : $(a+b)/2$ et Variance : $(b-a)^2/12$

Normale, Gaussienne, Laplace-Gauss $\mathcal{N}(m, \sigma^2)$
- — Plage des valeurs : $\mathbb{R}$
- — Densité : $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-m)^2}{2\sigma^2}\right)$

— Espérance : m et Variance : σ^2
— Interprétation : le loi des moyennes (et des erreurs), avec le théorème de la limite centrale.

Exponentielle $\mathcal{E}(\lambda)$
— Plage des valeurs : $\mathbb{R}^+$
— Densité : $f(x) = \lambda e^{-\lambda x} \mathbb{1}_{\mathbb{R}^+}(x)$
— Espérance : $1/\lambda$ et Variance : $1/\lambda^2$
— Interprétation : la loi des durés de vie ou de réalisations de tâches.

Nous renvoyons à la fiche distribuée pour d'autres lois : Weibull, Gamma, etc.

1.2.3 Quelques rappels spécifiques à la loi de Gauss et aux lois dérivées

La loi de Gauss, ou loi normale, a été rappelée ci-dessus. Cette loi est fondamentale en raison du théorème de la limite centrale, qui justifie qu'elle soit si fréquente en modélisation.

Propriété fondamentale :
Si $X \sim \mathcal{N}(m, \sigma^2)$, alors $aX + b \sim \mathcal{N}(am + b, a^2\sigma^2)$ et $\frac{X-m}{\sigma} \sim \mathcal{N}(0, 1)$.

Cette propriété permet quelque soient les paramètres de la loi normale considérée de se ramener à une loi dite centrée-réduite $\mathcal{N}(0, 1)$. Ainsi, dans les tables statistiques, seule la loi centrée-réduite est donnée, et vous devrez utiliser cette propriété afin de vous y ramener.

Nous donnons maintenant les principales lois obtenues lors de manipulations de la loi normale. Ces lois apparaissent notamment dans le théorème de Fisher donné ci-après, qui est particulièrement utile dans le cadre de ce cours.

Loi du χ^2 :
Si $X \sim \mathcal{N}(0, 1)$, alors $X^2 \sim \chi^2_1$.
Si $X_1, \ldots, X_n$ indépendantes et de même loi χ^2_1, alors $\sum_{i=1}^n X_i^2 \sim \chi^2_n$.

Loi de Student :
Si $U \sim \mathcal{N}(0, 1)$ et $Z \sim \chi^2_n$, et si U et Z sont indépendantes, alors $\frac{U}{Z/\sqrt{n}} \sim St(n)$.

Théorème 1.3. *Théorème de Fisher*
Soient $X_1, \ldots, X_n$ indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$. Posons $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ et $S_n'^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X_n})^2$. Alors :

- $\overline{X_n} \sim \mathcal{N}(m, \sigma^2/n)$,
- $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - m)^2 \sim \chi^2_n$,
- $\frac{(n-1){S'_n}^2}{\sigma^2} \sim \chi^2_{n-1}$,
- $\overline{X_n}$ *et* ${S'_n}^2$ *sont indépendantes,*
- $\frac{\overline{X_n}-m}{S'_n/\sqrt{n}} \sim St(n-1)$.

Ces résultats servent à établir les intervalles de confiance et les procédures de test dans le cas d'un échantillon issu d'une loi normale.

Loi de Fisher-Snedecor :
Si $X \sim \chi^2_n$ et $Y \sim \chi^2_m$, et si X et Y sont indépendantes, alors $\frac{X/n}{Y/m} \sim \mathcal{F}(n, m)$.

Cette loi est utile dans les tests de comparaison de variances.

L'expression explicite des densités de ces lois n'est pas à connaître (sauf pour la loi normale). Des tables statistiques et des logiciels permettent de les manipuler.

CHAPITRE 2

ELÉMENTS DE STATISTIQUE DESCRIPTIVE

Le but est de décrire des observations $x_1, x_2, \dots x_n$ à notre disposition afin d'analyser leur structure et d'en extraire les informations pertinentes.

2.1 Variables discrètes

Nous supposons ici que la variable considérée est à valeurs dans un ensemble dénombrable. En pratique, comme on dispose d'un nombre fini d'observations, on peut considérer en général qu'il y a un nombre fini de modalités. Notons $e_1, e_2, \dots e_k$ les k valeurs observées.

Nous considérons dans un premier temps le cas d'une variable qualitative. Aucun indicateur numérique n'est alors possible, mais des diagrammes permettent une meilleure lisibilité des données. Nous nous intéressons ici aux résultats de l'élection européenne de 2009. Les résultats obtenus sont les suivants :

Partis	LO	NPA	FDG	PS	EE	AEI
Pourcentage des suffrages exprimés	1,2	4,88	6,48	16,48	16,28	3,63
Partis	MoDem		UMP-NC-GM		DLR	Libertas
Pourcentage des suffrages exprimés	8,46		27,88		1,77	4,8

Source : http ://www.france-politique.fr

Seuls les partis ayant obtenu un score supérieur à 1% sont étudiés.

Afin de visualiser comment se répartissent les observations parmi ces valeurs, deux diagrammes sont possibles :

- Le diagramme en bâtons consiste à associer à chaque modalité un bâton de longueur proportionnelle au nombre de fois où la modalité est observée.

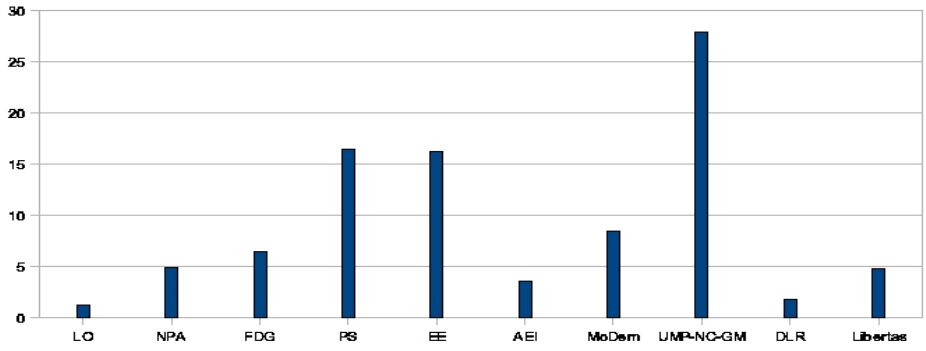

FIGURE 2.1 – Diagramme en bâtons.

L'ordre choisi dans cet exemple pour les modalités respecte les mouvances sur la scène politique (classique gauche-droite). Essayez en général dans un tel diagramme de donner un ordre cohérent.

- Le diagramme sectoriel ou camembert consiste à partitionner un disque en associant à chaque modalité une aire proportionnelle au nombre de fois où la modalité est observée.

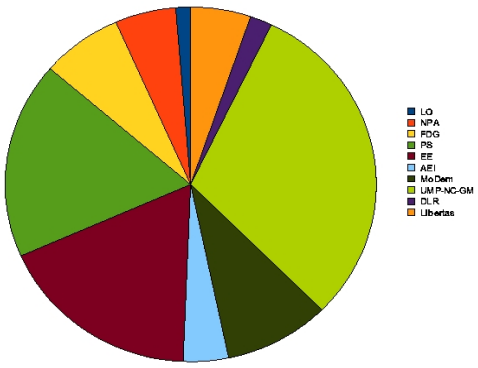

FIGURE 2.2 – Diagramme sectoriel

Dans le cas présent de partis politique, il arrive que l'on considère des demi-disques. Ceci est en général à éviter dans les autres contextes. Notons par ailleurs que le diagramme sectoriel est en général particulièrement bien adapté pour les

répartitions de budget.

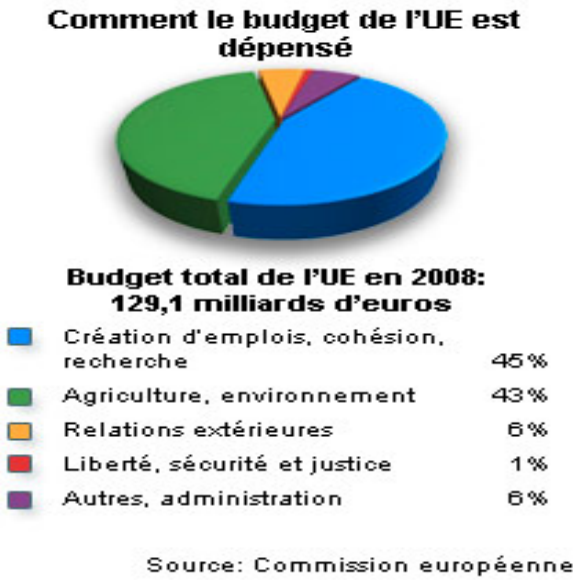

FIGURE 2.3 – Diagramme sectoriel budgétaire

Remarque : Il est préférable d'éviter les représentations en 3 dimensions (comme ci-dessus) qui réduisent la lisibilité d'un graphique.

Prenons maintenant l'exemple suivant : lors de la campagne de vaccination contre la grippe A en 2009, les autorités cherchent à savoir comment se répartissent les vaccinations. Elles ont observé au cours d'une semaine le nombre de personnes venant se faire vacciner dans un centre donné. Elles ont obtenu les résultats suivants :

Jour	1	2	3	4	5	6	7	Total
Nombre de vaccins	8	19	52	84	141	208	288	800
Pourcentage de vaccins	1	2,38	6,50	10,50	17,63	26	36	100

(Ces données sont fictives.)

Nous considérons ici comme variable aléatoire d'intérêt le nombre de jour avant qu'un individu aille se faire vacciner. Cette variable aléatoire est discrète et quantitative. Contrairement au cas précédent où la variable était qualitative. Concernant la représentation graphique, la repésentation adaptée est un diagramme en bâtons, construit de manière identique à ce qui précède : en abscisse les modalités de la variable et en ordonnée les fréquences observées pour chacune d'elles.

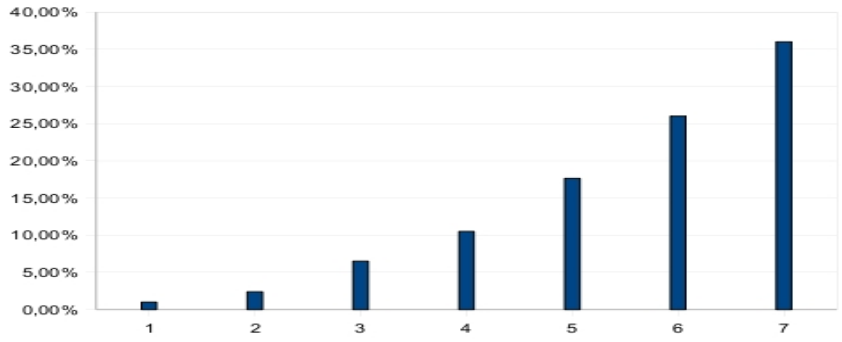

FIGURE 2.4 – Diagramme en bâtons.

Le graphique montre une courbe fortement croissante.

Cette variable étant quantitative, nous avons de plus accès à des indicateurs numériques permettant de résumer les observations. Notons $x_1, \ldots, x_n$ le nombre de jour attendu par chacun des 800 individus observés avant d'aller se faire vacciner.

Indicateurs de tendance

— Moyenne empirique, $\overline{x_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Lorsque les données sont *rangées* : pour chaque modalité $(e_i)_{i=1,\ldots,k}$, il y a n_i individus, alors $\bar{x}_i = \frac{1}{n}\sum_{i=1}^{k} n_i e_i$. C'est un indicateur usuel que vous rencontrez fréquemment. Il donne une *valeur centrale* pour l'échantillon.

— Médiane empirique, $\tilde{q}_{50\%}$. C'est la valeur telle que 50% des individus aient une modalité inférieure à $\tilde{q}_{50\%}$, et 50% aient une modalité supérieure. Pour calculer la médiane empirique, on commence par ordonner l'échantillon : $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$. Si n est impair, nous avons $\tilde{q}_{50\%} = x_{((n+1)/2)}$. Si n est impair, nous savons que $\tilde{q}_{50\%}$ est compris entre $x_{((n)/2)}$ et $x_{((n+1)/2)}$. Par convention, nous choisirons le milieu : $\tilde{q}_{50\%} = \left(x_{((n)/2)} + x_{((n+1)/2)}\right)/2$.

Il est intéressant de comparer les valeurs de la moyenne et de la médiane. Prenons l'exemple des salaires en France d'après l'étude de l'INSEE en 2007. La moyenne des salaires par an par individu est de 33 100 euros. La médiane des salaires par an par individu est de 27 630 euros. On observe que la médiane est significativement inférieure à la moyenne. Cela signifie qu'un petit nombre de français a un

salaire très élevé, comparativement au niveau moyen. En général, on interprète les écarts entre moyenne et médiane comme suit : si les valeurs sont comparables, la distribution est symétrique. Si la moyenne est nettement plus élevée, un petit nombre d'individus admet des valeurs très supérieures à la majorité, et si la moyenne est nettement plus faible, un petit nombre d'individus admet des valeurs très inférieures à la majorité

Nous pourrions généraliser la médiane empirique aux quantiles empiriques, mais ceci n'apporterait pas beaucoup dans cet exemple. Nous introduirons cette notion pour l'exemple des variables continues où elle semble plus adaptée.

Indicateurs de dispersion

- Variance empirique, $s_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x_n}^2$. Ou encore $s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i e_i^2 - \overline{x_n}^2$. Cette grandeur mesure la dispersion des données par rapport à la moyenne. Nous verrons ultérieurement que la variance est très utilisée en théorie car nous connaisons beucoup de propriétés sur son estimation. Mais en pratique, vous ne pouvez pas interpréter une variance : si vous considérer des mètres, la variance est en mètres^2, des secondes, elle est en secondes^2, etc. Mieux vaut regarder, pour l'interprétation, l'écart-type empirique.
- Ecart-type empirique, s_n. Il a même unité que la variable considérée. S'il est grand cela signifie que les données sont très dispersées, s'il est faible que la variabilité entre les individus est relativement faible. Pour déterminer si l'écart-type est grand ou non, on regarde le coefficient de variation empirique, $cv_n = \frac{s_n}{\overline{x_n}}$. Si $cv_n \geqslant 0.25$ on considère que la variabilité est forte et si $cv_n < 0.25$ on considère que la variabilité est raisonnable.

Dans l'exemple des vaccins contre la grippe A, le nombre de jours avant d'aller se faire vacciner vérifie :

moyenne empirique	médiane empirique
5,6338	6

variance empirique	écart-type empirique	coefficient de variation empirique
2,04	1,43	0,25

La moyenne est relativement élevée compte-tenu de la plage de valeurs, ce qui s'explique par cette forte proportion d'individus dans les derniers jours. Par ailleurs, la médiane est

plus élevée que la moyenne, ce qui est dû à une asymétrie de la distribution : les valeurs sont plus *étalées* vers le bas.

On ne peut considérer la variabilité comme très forte. Ceci est sûrement dû au fait que nos données sont peu étalées dans le temps car au vu de la croissance de la courbe, on peut présager d'une forte disparité entre les individus.

2.2 Variables continues

La station de Lyon-Bron a enregistré entre 1921 et 1992 les températures moyennes suivantes :

Mois	janvier	février	mars	avril	mai	juin	juillet	août	septembre
Lyon	2,51	4,01	7.55	10.58	14.55	18.15	20.76	20.13	17.11
Mois	octobre	novembre	décembre						
Lyon	12.14	6.82	3.2						

source : BLANCHET, 1993, Le climat de Lyon et sa région, Bulletin Mensuel de la Société LINEENE de Lyon

2.2.1 Histogrammes

Lorsqu'on observe des données $x_1 \dots x_n$ issues d'une variable aléatoire continue, le moyen usuel de représentation est l'histogramme. L'idée est de regrouper les valeurs dans des classes $([a_{j-1}; a_j[)_{j=1\dots k}$. On peut ensuite associer à chaque classe le nombre n_j d'observations $(x_i)_{i=1,\dots,n}$ appartenant à la classe $[a_{j-1}; a_j[$. On est ramenés à un tableau en effectifs de la forme :

Classe	$[a_0; a_1[$	$[a_1; a_2[$	$\dots$	$[a_{k-1}; a_k[$	total
Effectif	n_1	n_2	$\dots$	n_k	n
Fréquence	f_1	f_2	$\dots$	f_k	1

Les fréquences sont données par $f_i = \frac{n_i}{n}$. Lorsqu'il y a de nombreuses observations, les données sont souvent déjà regroupées en classes.

Une question qui se pose est comment choisir les classes ? Combien en prendre et quelles bornes a_j donner ? En général, le nombre de classes est choisi selon la Règle de Sturges : k est l'entier le plus proche de $1 + \log_2(n)$. Ensuite, les bornes extrêmes a_0 et a_k peuvent être données par

$$a_0 = \min_{i=1...n} x_i - 0.025(\max x_i - \min x_i) \text{ et } a_k = \max_{i=1...n} x_i - 0.025(\max x_i - \min x_i).$$

Enfin, les deux choix les plus usuels de classes sont les classes de même largeur et les classes de même effectif. Nous verrons par la suite comment construire de telles classes.

Afin de faciliter la lecture du tableau obtenu, nous optons pour une représentation graphique : un histogramme.

Définition 2.1. Un histogramme est la juxtaposition de rectangles de bases $([a_{j-1}; a_j[)_{j=1...k}$ et d'**aire** égale à la fréquence f_j associée. La hauteur du $j^{\text{ème}}$ rectangle est donc égale à $\frac{f_j}{a_j - a_{j-1}}$.

Nous voulons faire l'histogramme des températures moyennes de Lyon. On commence par trier les données : 2.51, 3.2, 4.01, 6.82, 7.55, 10.58, 12.14, 14.55, 17.11, 18.15, 20.13, 20.76. La règle de Sturges inciterait à faire 5 classes mais pour des raisons d'ordre pratique, nous choisissons ici d'en faire 6. Les bornes extrêmes des classes sont ensuite $a_0 = 2.55$ et $a_k = 21.45$.

Création d'un histogramme avec 6 classes de même largeur

Nous souhaitons dans un premier temps réaliser un histogramme avec des classes de même largeur. Soit h la largeur de chaque classe : $a_j = a_{j-1} + h$. Alors il est immédiat que $a_k - a_0 = kh$, donc que ici $h = (21.45 - 2.55)/6 = 3.19$.

Nous obtenons donc les classes suivantes : [2.05-5.25[, [5.25-8.44[, [8.44-11.64[, [11.64-14.83[, [14.83-18.02[, [18.02-21.22[. Ces classes constituent les bases des rectangles de l'histogramme. Nous aimerions ensuite déterminer la hauteur des rectangles. Rappelons que l'aire est proportionnelle à la fréquence f_j. Ainsi, la hauteur du $j^{\text{ème}}$ rectangle vaut $f_j/(a_j - a_{j-1})$. Lorsque les classes ont même largeur h, la hauteur vaut f_j/h. Nous obtenons ainsi :

classes	[2.05-5.25[	[5.25-8.44[	[8.44-11.64[	[11.64-14.83[	[14.83-18.02[	[18.02-21.22[
effectifs	3	2	1	2	1	3
fréquences	0.25	0.17	0.08	0.17	0.08	0.25
largeurs	3.19	3.19	3.19	3.19	3.19	3.19
hauteurs	0.078	0.052	0.026	0.052	0.026	0.078

La représentation graphique associée est :

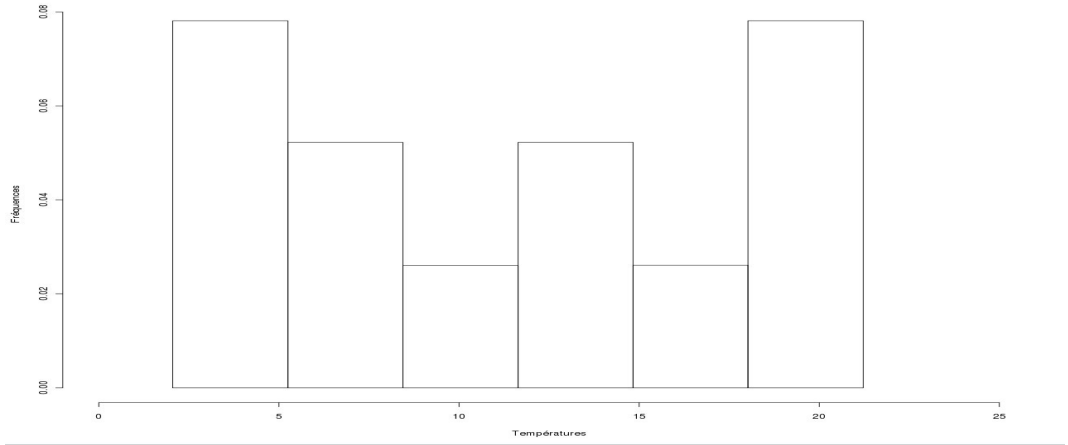

FIGURE 2.5 – Histogramme de la température moyenne à Lyon avec 6 classes de même largeur

Création d'un histogramme avec 6 classes de même fréquence

Nous allons maintenant construire l'histogramme avec 6 classes de même fréquence. Nous manipulons 6 classes et 12 observations donc si les classes ont mêmes fréquences, cela signifie qu'il y a 2 observations par classes. La borne de chaque classe est donnée par le milieu entre les observations que l'on veut séparer (qu'on a bien entendu prélablement triées). On obtient ainsi les classes : [2.05-3.61[, [3.61-7.19[, [7.19-11.36[, [11.36-15.83[, -[15.83-19.14[, [19.14-21.22[.

De même, la hauteur du $j^{\text{ème}}$ rectangle de l'histogramme vaut $f_j/(a_j - a_{j-1})$, mais ici les largeur des classes n'étant pas identiques, il faut calculer cette hauteur pour chacune des classes. Ce qui nous donne :

classes	[2.05-3.61[	[3.61-7.19[	[7.19-11.36[	[11.36-15.83[	[15.83-19.14[	[19.14-21.22[
effectifs	2	2	2	2	2	2
fréquences	0.17	0.17	0.17	0.17	0.17	0.17
largeurs	1.55	3.58	4.18	4.47	3.31	2.08
hauteurs	0.107	0.047	0.040	0.037	0.050	0.080

La représentation graphique associée est :

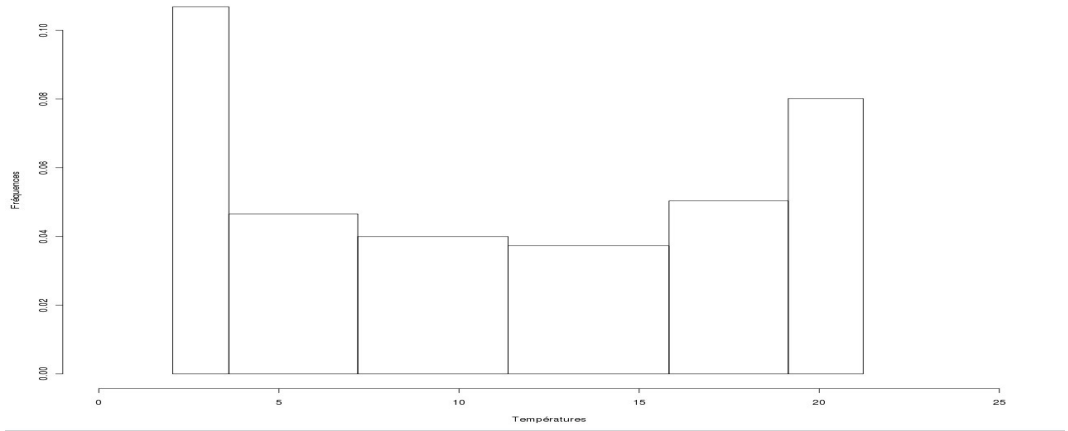

FIGURE 2.6 – Histogramme de la température moyenne à Lyon avec 6 classes de même fréquence

Remarquons que les tableurs usuels (Microsoft Office, Open Office) ne savent pas faire un “vrai” histogramme. En effet ils prennent une hauteur et non une aire proportionnelle à la fréquence. Ceci peut avoir des conséquences importantes quand les bases ne sont pas de même largeur !

Les histogrammes sont en fait des estimateurs de la fonction de densité de probabilité : on approxime la densité à l’aide de fonctions constantes par morceaux. Ce principe se généralise avec des droites par morceaux, des polynômes, etc. En général, un histogramme avec classes de même fréquence rend mieux compte visuellement de la forme de la distribution.

2.2.2 Indicateurs numériques

Afin de mieux comprendre l’intérêt des indicateurs numériques, nous allons les exploiter dans la comparaison de la température de 2 villes. Nous allons comparer les distributions

des températures de Lyon et de Brest :

Mois	J	F	M	A	M	J	J	A	S	O	N	D
Lyon	2.51	4.01	7.55	10.58	14.55	18.15	20.76	20.13	17.11	12.14	6.82	3.2
Brest	7.94	8.01	8.42	9.64	12.08	14.45	16.67	17.72	15.64	12.8	10.29	8.97

Les indicateurs numériques associés sont les suivants :

	Lyon	Brest
Moyenne empirique	11.46	11.89
Médiane empirique	11.36	11.19
Variance empirique	40.81	11.43
Ecart-type empirique	6.39	3.38
Coefficient de variation empirique	0.56	0.28

La moyenne indique que la température est globalement plus élevée à Brest qu'à Lyon, mais la différence semble faible. La médiane étant plus faible pour Lyon, nous avons envie de dire que globalement les températures sont comparables dans les deux villes. Cependant les écarts-types montrent que la variation des températures au cours de l'année est nettement plus importante à Lyon. Pour affiner ce constat, nous regardons les quantiles des distributions. Ces grandeurs donnent en effet une vision plus fine de la façon dont s'étalent les données.

Quantile théorique. Rappelons que le quantile d'ordre α d'une variable aléatoire X est la grandeur q_α telle que $\mathbb{P}(X \leqslant q_\alpha) = \alpha$.

Quantile empirique. Nous noterons $\tilde{q}_\alpha$ la valeur telle qu'une proportion des individus égale (au moins) à α ait une modalité inférieure à $\tilde{q}_\alpha$. Pour les calculer, on commence par ordonner l'échantillon : $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$. Si $n\alpha$ n'est pas entier, nous avons $\tilde{q}_\alpha = x_{([n\alpha]+1)}$ où $[y]$ est la partie entière de y. Si $n\alpha$ est entier, nous prenons : $\tilde{q}_\alpha = \alpha\, x_{(n\alpha)} + (1-\alpha)\, x_{(n\alpha+1)}$.

La médiane est le quantile d'ordre $\alpha = 50\%$, les quartiles sont les quantiles d'ordre $\alpha = 25\%$, 50%, 75% et les déciles d'ordre $\alpha = 10\%$, $20\%, \ldots, 90\%$. Le principe est de découper la population en tranches de mêmes effectifs.

Exemple :

- Calcul du 1^{er} décile. Nous souhaitons que 10% des observations soient inférieures à $\tilde{q}_{10\%}$. Comme nous avons 12 observations, cela signifie que nous voulons $n.10\% = 1.2$ observations sous $\tilde{q}_{10\%}$. Nous prenons donc $\tilde{q}_{10\%} = x_{(2)}$.
- Calcul de la médiane. Nous voulons que 50% des observations soient inférieures à $\tilde{q}_{50\%}$. Comme nous avons 12 observations, cela signifie que nous voulons $n.50\% = 6$ observations sous $\tilde{q}_{50\%}$. Donc $\tilde{q}_{50\%} \leqslant x_{(6)}$. Si nous prenons $\tilde{q}_{50\%} \geqslant x_{(7)}$, nous aurons 7 observations au-dessous, donc nous devons prendre $x_{(6)} \leqslant \tilde{q}_{50\%} < x_{(7)}$. Toute valeur vérifiant ces inégalités conviendrait mais par convention, nous choisissons le milieu : $\tilde{q}_{50\%} = (x_{(6)} + x_{(7)})/2$.

Outre les quantiles, les minima et maxima donnent aussi souvent une information importante sur une distribution.

Pour l'exemple des températures, nous obtenons :

	minimum	1er decile	1er quartile	médiane	3ème quartile	9ème décile	maximum
Lyon	2.51	3.2	6.12	11.36	17.37	20.13	20.76
Brest	7.94	8.01	8.83	11.19	14.75	16.67	17.72

Les extrêmes (minimum et maximum) confirment que l'étalement des températures est nettement plus important à Lyon et indique que ceci est vrai dans les basses comme dans les hautes températures. De même les quantiles tendent à montrer que cette dispersion des températures plus importante est valable sur toute la gamme de température.

Il existe une représentation graphique associée aux quantiles : le *boxplot*, ou *boîte à moustache* dont le principe est décrit par le schéma suivant :

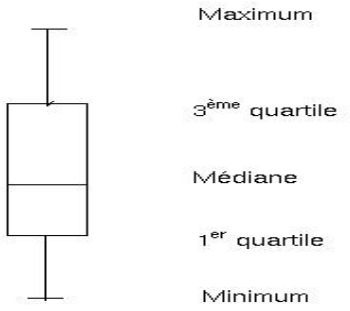

FIGURE 2.7 – Construction d'un boxplot

L'idée est qu'un tel schéma permet de voir si la distribution est asymétrique (étalement plus grand dans les basses ou les hautes valeurs) et de comparer l'étalement de deux

distributions.

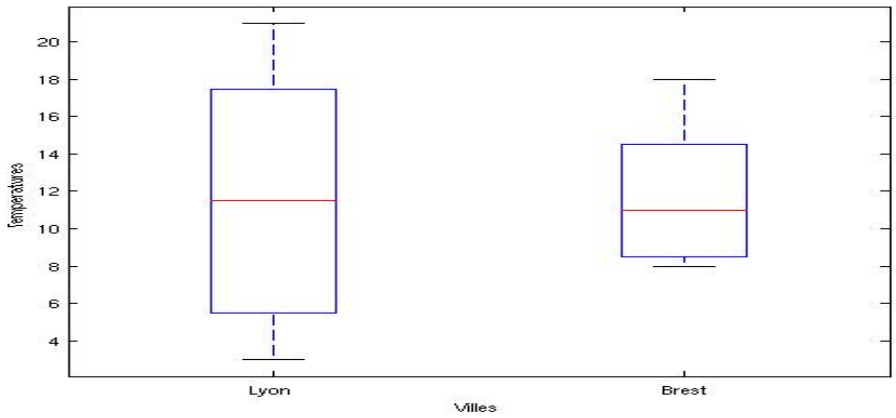

FIGURE 2.8 – Boxplots des distributions de température à Lyon et à Brest

Dans l'exemple, nous observons que la distribution des températures à Lyon est globalement symétrique tandis que celle de Brest présente un étalement plus important dans les valeurs élevées. De plus si l'étalement est plus important pour Lyon que pour Brest dans toutes les gammes de température, cette différence est nettement plus accentuée dans les basses températures. En conclusion, nous pouvons voir que les deux climats sont très différents : Lyon a un climat continental avec des températures très variables, avec équirépartition des périodes chaudes et des périodes froides ; le climat de Brest est océanique, avec des températures moins variées.

Nous n'avons eu ici qu'un petit aperçu des techniques de statistiques descriptives. Un large panel de méthodologies est disponible. La particularité de la statistique descriptive est de décrire les données sans chercher à les modéliser : le but est de donner des outils permettant de représenter et d'analyser les observations sans faire appel aux notions probabilistes. Cependant, les observations peuvent être vues comme des réalisations de variables aléatoires. L'intérêt est alors la caractérisation du comportement d'une variable à l'aide d'une loi de probabilité permet une compréhension plus fine du phénomène, un contrôle de la précision des grandeurs évaluées, une comparaison avec d'autres gran-

deurs données par la théorie ou observées, etc. La modélisation probabiliste des observations est très utile. Néanmoins, il faut bien garder alors en mémoire qu'il ne s'agit que de modèles : nous approchons la réalité à l'aide d'un modèle mathématique, donc nous la déformons. De plus, ceci nécessite que le modèle choisi corresponde au mieux aux donnéees.

Dans les chapîtres suivants, nous considérons que nous avons identifié un modèle pour nos observations. Nous développons les méthodes alors accessibles. Comment le choisir en pratique ? Tout d'abord, les histogrammes ou les diagrammes en bâtons permettent d'identifier une famille de lois possibles. Ensuite, il existe des outils de statistique descriptive, que nous n'avons malheureusement pas présenté ici, permettant de vérifier rapidement notre intuition. Enfin, nous verrons que l'on peut tester si le modèle choisi est adéquat.

CHAPITRE 3

ESTIMATION PONCTUELLE ET INTERVALLES DE CONFIANCE

Soit X la variable aléatoire que nous souhaitons étudier. Nous effectuons pour cela n réalisations de X. Les résultats de ces réalisations sont des variables aléatoires notées $X_1, X_2, \dots X_n$, qui ont même loi que X. Nous considérerons dans l'intégralité de ce cours que les mesures sont effectuées de manières indépendantes. On dit alors que les variables $X_1, X_2, \dots X_n$ sont i.i.d. : indépendantes et identiquement distribuées.

Nous supposons que, par des méthodes de statistique descriptive, nous avons pu déterminer que la loi des variables X_i appartient à une famille de lois paramétrée. Plus précisément, nous considérons que la fonction de répartition de X appartient à $\{F_\theta, \theta \in \mathbb{R}^d\}$.

Le but est ici d'estimer la valeur du paramètre θ, c'est-à-dire de déterminer la valeur de θ la plus vraisemblable pour la loi de $X_1, X_2, \dots X_n$. Et une fois cette grandeur estimée, quelle est la précision de l'approximation réalisée ?

Il faudra bien distinguer dans la suite la théorie : on manipule des variables aléatoires $X_1, X_2, \dots X_n$ dont on ne connaît pas les réalisations, de la pratique : on a observé n valeurs $x_1, x_2, \dots x_n$, réalisations de $X_1, X_2, \dots X_n$ et on veut extraire l'information de ces données. De manière générale, les majuscules désignent des variables aléatoires, donc des manipulations théoriques, tandis que les minuscules correspondent à des observations, donc à des valeurs numériques.

3.1 Construction de l'estimation

La première étape est d'abord de définir plus rigoureusement ce qu'est un estimateur :

Définition 3.1. Une **statistique** t_n est une fonction des observations $x_1, x_2, \dots x_n$.

Exemples : $\overline{x_n} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ min(x_i),\ (2x_1 + 3x_2, \log x_6, x_2)\dots$

C'est une réalisation d'une variable aléatoire T_n aussi appelée statistique qui est fonction des variables $X_1, X_2, \dots X_n$.

Exemples : $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ min(X_i),\ (2X_1 + 3X_2, \log X_6, X_2)\dots$

Définition 3.2. Un **estimateur** d'un paramètre θ est une statistique T_n à valeur dans l'ensemble des valeurs possibles de θ.

Une **estimation** de θ est une réalisation d'un estimateur.

Attention car bien souvent on note $\widehat{\theta}_n$ les estimateurs comme les estimations de θ. La distinction entre la théorie et la pratique n'est donc pas visible mais il faut en garder conscience.

Cette définition ouvre un très large champ d'estimateurs possibles : un estimateur est juste une valeur plausible construite à partir des observations. Pour un paramètre donné, nous pouvons alors construire une infinité d'estimateurs. Prenons l'exemple d'une loi uniforme.

Exemple : Estimateurs pour la loi uniforme.
Dans un laboratoire d'astrophysique, un capteur reçoit des particules. Les temps (exprimés en heures) écoulés entre la reception des particules obtenus sont :

75 265 22 402 35 144 346 159 229 62

Le temps entre la $i^{\text{ème}}$ et de la $i+1^{\text{ème}}$ particule est noté X_i. Les variables $X_1, \dots X_n$ sont indépendantes et de même loi. Au vu de l'histogramme, on décide de modéliser par une loi uniforme $\mathcal{U}([0,\theta])$. Les physiciens aimeraient accéder à la valeur du paramètre θ. Il faut donc leur proposer un estimateur : c'est-à-dire une méthode leur permettant de calculer la valeur de θ la plus plausible au vu de leur données. Mais étant donné la définition que nous avons donnée, toute valeur positive convient !

Afin de choisir l'estimateur que nous allons considérer, nous allons essayer de donner des critères mesurant la qualité des estimateurs construits. Nous introduisons alors la notion de biais.

Définition 3.3. Soit T_n un **estimateur** de θ. Le biais de T_n est défini comme

$$Biais(T_n, \theta) = \mathbb{E}T_n - \theta.$$

Si $Biais(T_n, \theta) = 0$, on dit que T_n est un estimateur sans biais de θ.

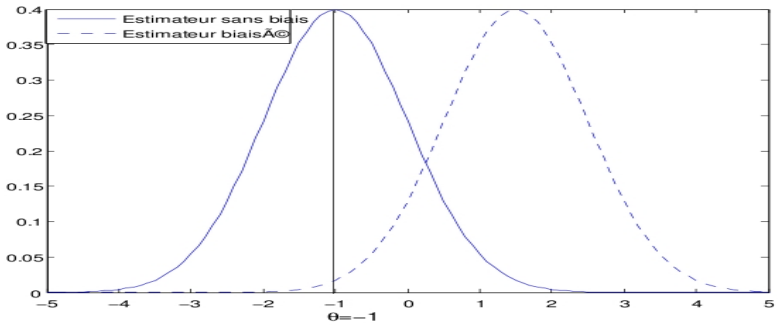

FIGURE 3.1 – Densités d'estimateurs respectivement sans biais et biaisé de la valeur -1.

Le biais mesure si l'estimateur T_n considéré a tendance à sous-estimer (biais négatif) ou sur-estimer (biais positif) la valeur de θ. Nous voyons vite que nous allons chercher en priorité des estimateurs sans biais de θ car cela signifie qu'en moyenne ils retournent bien la vraie valeur de θ. Remarquons qu'un estimateur sans biais ne doit pas avoir un biais nul pour une valeur fixée de θ mais quelque soit la valeur du paramètre que nous souhaitons évaluer.

Exemple : Estimateurs pour la loi uniforme.
Dans l'exemple ci-dessus, considérons $T_1 = 500$ h. Il est alors évident que T_1 est biaisé et ne peut fournir un bon estimateur.

Quel estimateur proposer alors ? L'idée est de regarder quelle opération sur $X_1, \dots X_n$ permet de retrouver θ. Remarquons que si $X_i \sim \mathcal{U}([0, \theta])$, alors en moyenne X_i vaudra $\theta/2$: en effet $\mathbb{E}[X_i] = \theta/2$. La moyenne des X_i est donnée par $\frac{1}{n}\sum_{i=1}^{n} X_i$. Nous pouvons alors proposer $T_2 = \frac{2}{n}\sum_{i=1}^{n} X_i$. Calculons le biais de T_2 :

Par linéarité de l'espérance, $\mathbb{E}T_2 = \frac{2}{n}\sum_{i=1}^{n} EX_i$.
Comme pour tout i $\mathbb{E}X_i = \theta/2$, nous avons $\mathbb{E}T_2 = \theta$.

Ainsi, T_2 est sans biais.

Remarquons par ailleurs que la valeur maximale que peut prendre X_i est θ. Nous pouvons donc proposer un autre estimateur qui est $T_3 = \max_{i=1,\dots,n} X_i$. Etudions la loi de T_3 :

Pour tout x, la fonction de répartition de T_3 vaut $\mathbb{P}(T_3 \leqslant x) = \mathbb{P}(X_1, \dots, X_n \leqslant x) = \mathbb{P}(X_i \leqslant x)^n$ car les variables sont indépendantes et identiquement distribuées. Ainsi $\mathbb{P}(T_3 \leqslant x) = \left(\frac{x}{\theta}\right)^n$ pour $x \in [0, \theta]$.
En dérivant la fonction de répartition, nous obtenons la densité de T_3 :

$$f(x) = nx^{n-1}\theta^{-n}\mathbb{1}_{[0,\theta]}(x).$$

Nous en déduisons que $\mathbb{E}T_3 = \frac{n}{n+1}\theta$. L'estimateur T_3 est donc biaisé : il a tendance à sous-estimer la valeur du paramètre. Mais comme nous connaissons dans quelle mesure il sous-estime θ, nous sommes à-même de corriger cet écart. En effet par linéarité de l'espérance, $\mathbb{E}\left[\frac{n+1}{n}T_3\right] = \theta$. Ainsi, $T_4 = \frac{n+1}{n}T_3$ est un estimateur sans biais de θ.

Les estimations obtenues dans l'exemple valent $t_2 = 347.8h$ et $t_4 = 442.2h$.

Nous voyons ici que le critère que nous avons défini ne suffit pas à choisir entre les deux estimateurs T_2 et T_4 : tous deux sont sans biais, c'est-à-dire d'espérance égale au paramètre que l'on cherche à estimer. Mais il faut encore que la variabililité autour de cette moyenne soit faible :

Définition 3.4. Soit T_n un **estimateur** de θ. L'erreur quadratique de T_n est défini comme

$$EQM(T_n, \theta) = \mathbb{E}[(T_n - \theta)^2] = Biais(T_n, \theta)^2 + Var(T_n).$$

Si $EQM(T_n, \theta) \underset{n\to\infty}{\longrightarrow} 0$, on dit que T_n est un estimateur de θ convergeant en moyenne quadratique.

Pour un estimateur sans biais, la convergence en moyenne quadratique équivaut à la convergence vers 0 de la variance, donc à la concentration des valeurs observées autour de l'espérance pour un grand nombre d'observations. L'idée est alors de chercher à construire de estimateurs sans biais et convergeant en moyenne quadratique.

Remarque : Il existe une borne dite borne de Cramer-Rao donnant la plus petite variance possible pour un estimateur dans un contexte donné. Pour plus de détails, voir la notion d'information de Fisher.

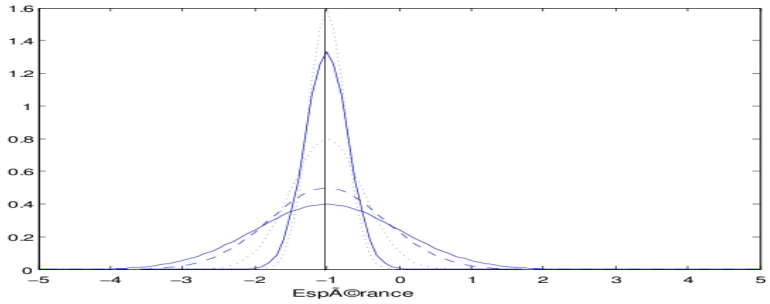

FIGURE 3.2 – Evolution de densités d'estimateurs convergeant en moyenne quadratique.

Exemple : Estimateurs pour la loi uniforme.

Nous avions sélectionné parmi les estimateurs possibles les estimateurs $T_2 = \frac{2}{n}\sum_{i=1}^{n} X_i$ et $T_4 = \frac{n+1}{n}\max_{i=1,\dots,n} X_i$. Nous pouvons montrer que $Var(T_2) = \frac{\theta^2}{3n}$ et que $Var(T_4) = \frac{\theta^2}{n(n+2)}$. Dans les deux cas, nous avons la variance qui tend vers 0 quand n tend vers l'infini, donc ces deux estimateurs sont convergeants. Néanmoins, l'estimateur T_4 admet une variance plus faible, donc est préférable à l'estimateur T_2. En conclusion, vous pouvez proposer aux physiciens d'utilser T_4 pour estimer la valeur de θ. L'estimation retenue est ainsi $t_4 = 442.2$ heures.

Reste maintenant à voir dans un cadre plus général que pour la loi uniforme comment on peut construire des estimateurs avec de bonnes propriétés, c'est-à-dire sans biais et convergeants. Il existe de très nombreuses méthodes pour cela. Les deux plus courantes sont la méthode des moments et la méthode du maximum de vraisemblance. Nous exposons aussi ici le principe de l'estimation bayésienne en raison de son utilisation fréquente, notamment dans le traitement d'images, mais elle n'est pas à retenir.

3.1.1 Méthode des moments

La méthode des moments est relativement intuitive. Il semble naturel d'estimer les moments par leur version empirique :

— L'espérance $\mathbb{E}X$ correspond à une moyenne théorique et peut être estimée par la

moyenne observée $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

— La variance $Var(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2$ peut être estimée par la variance observée $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X_n})^2$.

Lorsque l'on souhaite estimer un paramètre θ, on essaie de l'exprimer en fonction de $\mathbb{E}X$ et de $Var(X)$: $\theta = g(\mathbb{E}X, Var(X))$. On sait ensuite que $\mathbb{E}X$ et de $Var(X)$ peuvent être estimés respectivement par $\overline{X_n}$ et S_n^2. On propose donc d'estimer θ par $\widehat{\theta}_n = g(\overline{X_n}, S_n^2)$.

Ce principe se généralise avec l'ensemble des moments d'ordre p, $p \in \mathbb{N}^*$.

Exemples :

- Si $X_1, \dots X_n$ sont de loi Géométrique de paramètre p, alors, $\mathbb{E}X_i = 1/p$. L'estimateur des moments de p est donc $\widehat{p}_n = 1/\overline{X_n}$.

- Si $X_1, \dots X_n$ sont de loi uniforme sur l'intervalle $[a; b]$ alors, $\mathbb{E}X_i = (a+b)/2$ et $Var(X_i) = (b-a)^2/12$. Par conséquent $a = \mathbb{E}X_i - \sqrt{3Var(X_i)}$ et $b = \mathbb{E}X_i + \sqrt{3Var(X_i)}$. Les estimateurs des moments de a et b sont donc $\widehat{a}_n = \overline{X_n} - \sqrt{3}S_n$ et $\widehat{b}_n = \overline{X_n} + \sqrt{3}S_n$.

3.1.2 Méthode du maximum de vraisemblance

La vraisemblance d'un modèle et d'un échantillon correspond à la probabilité d'avoir obtenu cet échantillon lorsqu'on a ce modèle. Ainsi, si on suppose que le modèle est F_θ, la vraisemblance des observations $x_1, \dots, x_n$ s'écrit sous la forme :

$$\mathcal{L}(\theta, \{x_1, \dots, x_n\}) = \begin{cases} \mathbb{P}_\theta(x_1) \dots \mathbb{P}_\theta(x_n) & \text{si on a une loi discrète,} \\ f_\theta(x_1) \dots f_\theta(x_n) & \text{si on a une loi continue,} \end{cases}$$

avec $\mathbb{P}_\theta$ et f_θ respectivement fonction de masse et densité associées à F_θ. La forme de produit est justifiée par l'hypothèse que les observations sont indépendantes. (Voir le cours de probabilité et les rappels de probabilités en début de polycopié.)

Le principe de l'estimation par maximum de vraisemblance est de se dire que plus la probabilité d'avoir obtenu les observations est forte, plus le modèle est proche de la réalité. Ainsi, on retient le modèle pour lequel la vraisemblance de notre échantillon est la plus élevée :

$$\widehat{\theta}_n = arg\max_\theta \mathcal{L}(\theta, \{x_1, \dots, x_n\}).$$

En pratique, le problème ci-dessus est compliqué à résoudre directement en raison de la présence du produit mais il suffit de prendre le logarithme :

$$\widehat{\theta}_n = arg \max_{\theta} \log \mathcal{L}(\theta, \{x_1, \ldots, x_n\}).$$

Pour trouver le maximum, on résoud l'équation du premier ordre :

$$\left.\frac{\partial \log \mathcal{L}(\theta, \{x_1, \ldots, x_n\})}{\partial \theta}\right|_{\theta=\widehat{\theta}_n} = 0.$$

La théorie nous dit que la solution de cette équation nous donne toujours un maximum. On obtient $\widehat{\theta}_n$ sous la forme $\widehat{\theta}_n = g(x_1, \ldots, x_n)$.

L'estimateur du maximum de vraisemblance est alors $\widehat{\theta}_n = g(X_1, \ldots, X_n)$ et l'estimation du maximum de vraisemblance est obtenue en remplaçant $x_1, \ldots, x_n$ par leurs valeurs numériques dans $\widehat{\theta}_n = g(x_1, \ldots, x_n)$.

Remarque : Il faut toujours raisonner avec $x_1, \ldots, x_n$ sans chercher à remplacer par les valeurs observées. Cette étape ne doit intervenir qu'à la fin, lorsque l'expression de l'estimateur a été établie.

Exemples :

- Si $X_1, \ldots X_n$ sont de loi Géométrique de paramètre p, alors la fonction de vraisemblance du modèle est

  $$\mathcal{L}(p, \{x_1, \ldots, x_n\}) = \prod_{i=1}^{n} p(1-p)^{x_i - 1} = p^n (1-p)^{\sum x_i - n}.$$

  La log-vraisemblance vaut

  $$\log \mathcal{L}(p, \{x_1, \ldots, x_n\}) = n \log(p) + (\sum x_i - n) \log(1-p).$$

  L'équation du premier ordre s'écrit

  $$0 = \frac{\partial \log \mathcal{L}(p, \{x_1, \ldots, x_n\})}{\partial p} = \frac{n}{p} - \frac{\sum x_i - n}{1-p},$$

  ou encore $1/p = (\overline{x_n} - 1)/(1-p)$ dont la solution est $p = 1/\overline{x_n}$. L'estimateur du maximum de vraisemblance de p est donc $\widehat{p}_n = 1/\overline{X_n}$.

- Si $X_1, \dots X_n$ sont de loi uniforme sur l'intervalle $[a; b]$ alors, la fonction de vraisemblance du modèle est

$$\mathcal{L}(a, b, \{x_1, \dots, x_n\}) = \frac{1}{(b-a)^n} \prod_{i=1}^{n} \mathbb{1}_{[a,b]}(x_i).$$

  Nous sommes dans un cas particulier où il n'est pas nécessaire de passer par le logarithme et la dérivée. En effet, il est net que le maximum est atteint pour $\hat{a} = \min X_i$ et $\hat{b} = \max X_i$. Ce sont donc les estimateurs du maximum de vraisemblance de a et b.

- Si $X_1, \dots X_n$ sont de loi normale $\mathcal{N}(m, \sigma^2)$. La fonction de vraisemblance du modèle est

$$\mathcal{L}(m, \sigma^2, \{x_1, \dots, x_n\}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-m)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2).$$

  La log-vraisemblance vaut

$$\log \mathcal{L}(m, \sigma^2, \{x_1, \dots, x_n\}) = \frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2.$$

  Les équations du premier ordre s'écrivent

$$\begin{aligned} 0 = \frac{\partial \log \mathcal{L}(m, \sigma^2, \{x_1, \dots, x_n\})}{\partial m} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - m) \\ 0 = \frac{\partial \log \mathcal{L}(m, \sigma^2, \{x_1, \dots, x_n\})}{\partial (\sigma^2)} &= \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - m)^2 \end{aligned}$$

  La solution de ce système est $m = \overline{x_n}$ et $\sigma^2 = \frac{1}{n} \sum (x_i - m)^2$. Les estimateurs du maximum de vraisemblance sont donc $\widehat{m} = \overline{X_n}$ et $\widehat{\sigma}^2 = S_n^2$.

De manière générale, l'estimation par la méthode des moments est plus rapide que l'estimation par méthode de vraisemblance. Les deux approches donnent parfois le même résultat mais ce n'est pas toujours le cas, comme le montre par exemple le cas de la loi uniforme. De plus la méthode du maximum de vraisemblance présente des avantages théoriques, qui ne seront pas détaillés ici, qui assurent un bon comportement des estimateurs.

3.1.3 Estimation bayésienne

Cette méthode d'estimation est présentée ici dans la mesure où beaucoup d'entre vous sont amenés à manipuler cette estimation, notamment au cours de stages. Le principe général de cette estimation est donné mais les fondements théoriques sont éludés et je m'excuse du manque de rigueur dans la définition des lois et des espaces probabilisés.

Dans l'approche Bayésienne, on suppose que le paramètre θ est lui-aussi issu de la réalisation d'une variable aléatoire M de loi π_0, à valeurs dans Θ. La loi π_0 est appelée la **loi a priori**.

Ainsi, lorsque nous nous intéressons à la probabilité d'avoir observé x dans le modèle F_θ, nous regardons en fait la variable aléatoire $X \mid M = \theta$. Inversement, si nous sommes intéressé par θ, nous allons étudier la variable aléatoire $M|X = x$. La loi de $M|X = x$ est appelée **loi a posteriori**.

Afin de mieux comprendre comment déterminer la loi a posteriori, écrivons-la dans le cas de données continues : $\pi(\theta|X = x) = \frac{f_\theta(x)\pi_0(\theta)}{f_X(x)}$ avec $f_\theta = f_{X|M=\theta}$ densité associé au modèle F_θ et f_X densité marginale de X définie par $f_X(x) = \int_\Theta f_a(x)\pi_0(a)da$.

Généralisons maintenant cette écriture dans le contexte de n variables aléatoires $X_1, \dots X_n$ indépendantes et de même loi. La loi a posteriori s'écrit

$$\pi(\theta|X_1 = x_1, \dots X_n = x_n) = \frac{f_\theta(x_1) \dots f_\theta(x_n)\pi_0(\theta)}{\int_\Theta f_a(x_1) \dots f_a(x_n)\pi_0(a)da}.$$

L'estimateur bayésien de θ est alors défini comme l'espérance conditionnelle de M sachant $X_1 = x_1, \dots, X_n = x_n$,

$$\widehat{\theta}_n = \mathbb{E}[M|X_1 = x_1, \dots, X_n = x_n] \ = \int_\Theta \theta\pi(\theta|X_1 = x_1, \dots X_n = x_n)d\theta.$$

Dans la pratique, la loi π_0 dépend souvent d'un ou de plusieurs paramètres inconnus. La loi marginale de X dépendra donc aussi de ces paramètres. Ceux-ci devront être estimés. Si π_0 dépend de paramètres λ alors on estime λ par maximum de vraisemblance,

$$\mathcal{L}(\lambda, \{x_1, \dots, x_n\}) = \int_\Theta f_a(x_1) \dots f_a(x_n)\pi_0(a)da.$$

On calcule ensuite l'estimateur bayésien en remplaçant λ par son estimation dans la formule de $\mathbb{E}[M|X_1 = x_1, \dots, X_n = x_n]$.

Précisons enfin que l'estimation bayésienne constitue un complément de cours et n'est pas étudié en séance.

3.2 Estimation de la moyenne et de la variance

Supposons que les variables $X_1, \dots X_n$ soient des répliques indépendantes d'une variable aléatoire X vérifiant $\mathbb{E}X = m$ et $Var(X) = \sigma^2$. Nous définissons alors les estimateurs des moments $\widehat{m}_n = \overline{X_n}$ et $\widehat{\sigma}_n^2 = S_n^2$.

Il reste à déterminer si la qualité de ces estimateurs est satisfaisante. Rappelons que nous aimerions qu'ils soient sans biais et convergeants.

- Etude de $\widehat{m}_n$
  - $Biais(\widehat{m}_n, m) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}X_i - m = 0$ Donc $\widehat{m}_n$ est un estimateur sans biais de m.
  - $Var(\widehat{m}_n) = \frac{1}{n^2}\sum_{i=1}^n VarX_i$ car les X_i sont indépendants. D'où $Var(\widehat{m}_n) = \sigma^2/n$. Donc $Var(\widehat{m}_n) \underset{n\to\infty}{\longrightarrow} 0$ et ainsi $\widehat{m}_n$ est un estimateur de m convergeant en moyenne quadratique.

- Etude de $\widehat{\sigma}_n^2$

$$\begin{aligned}
\mathbb{E}(\widehat{\sigma}_n^2) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}X_i^2 - \mathbb{E}(\overline{X_n}^2) \\
&= \mathbb{E}X^2 - E(\widehat{m}_n^2) \\
&= Var(X) + (\mathbb{E}X)^2 - Var(\widehat{m}_n) - (\mathbb{E}\widehat{m}_n)^2 \\
&= \sigma^2 + m^2 - \sigma^2/n - m^2 \\
&= \left(1 - \frac{1}{n}\right)\sigma^2
\end{aligned}$$

  Donc $Biais(\widehat{\sigma}_n^2, \sigma^2) = -\frac{1}{n}\sigma^2$. L'estimateur $\widehat{\sigma}_n^2$ défini n'est pas un estimateur sans biais de σ^2. Nous introduisons alors $S_n'^2 = \frac{n}{n-1}S_n^2$. La statistique $S_n'^2$ est un estimateur sans biais de σ^2. On peut de plus montrer qu'il est convergeant en moyenne quadratique.

BILAN :

Soient $X_1, \ldots, X_n$ indépendantes et de même loi, tels que $\mathbb{E}X_i = m$ et $Var(X_i) = \sigma^2$, Alors un estimateur sans biais convergeant de m est donné par

$$\widehat{m}_n = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

et un estimateur sans biais convergeant de σ^2 est donné par

$$\widehat{\sigma}_n^2 = {S'_n}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^2 - \overline{X_n}^2\right).$$

Remarquons que si nous suggérons ici d'utiliser la méthode des moments (plus facile à appréhender), elle ne donne pas nécessairement un estimateur optimal. Dans le cas d'une loi uniforme considéré auparavant, elle mène par exemple à l'estimateur T_2. Or nous avons construit un estimateur T_4 sans biais de plus petite variance, donc préférable à T_2.

Il faut bien faire attention à distinguer la variance empirique S_n^2 de la variance estimée ${S'_n}^2$. La première de ces grandeurs est la variance observée sur l'échantillon, et la deuxième amène une correction pour annuler le biais. Lorsqu'on parle de la variance des observations, on considère donc S_n^2 et non ${S'_n}^2$.

Les logiciels que vous serez amenés à utiliser font cette distinction. Mais attention, si vous leur demandez la variance d'un échantillon sans préciser ils retournent en général ${s'_n}^2$. Pour obtenir s_n^2, il faut le plus souvent préciser que vous voulez la variance empirique. Ainsi sous (Open ou Microsoft) Office la fonction *var* retourne ${s'_n}^2$ et *var.p* retourne s_n^2.

Concrètement, si vous n'avez pas de logiciels, il est en général plus facile de calculer d'abord S_n^2 puis d'utiliser la relation ${S'_n}^2 = \frac{n}{n-1}S_n^2$.

3.2.1 Application pour la loi normale

Soient $X_1, \ldots, X_n$ indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$. Alors $m = \mathbb{E}X_i$ et $\sigma^2 = Var(X_i)$. D'après ce qui précède, nous pouvons estimer m par $\widehat{m}_n = \overline{X_n}$ et σ^2 par $\widehat{\sigma}_n^2 = {S'_n}^2$.

Considérons un exemple.

Exemple : Estimateurs pour la loi normale.
Nous avons demandé aux élèves de 4IF leur temps de sommeil moyen en heure par nuit en période de projet. Notons X_i les réponses obtenues. Nous avons les indicateurs numériques suivants :

Nombre de réponses	Moyenne	Médiane	Variance	Ecart-type
30	6.36	6	1.80	1.34

Ces résultats sont issus d'un questionnaire anonyme réalisé sur la promotion 2009-2010. Cependant, afin de vous permettre de retrouver ultérieurement les valeurs lues sur les tables, nous avons modifié le nombre de réponses obtenues. Toutes nos excuses pour ce manque de rigueur.

Nous supposons que les X_i sont indépendantes (ceci suppose que les élèves sont sur des projets différents, ne vivent pas ensemble, etc, donc c'est une hypothèse un peu forte) et qu'elles suivent une loi normale $\mathcal{N}(m, \sigma^2)$. Le paramètre m représente ici le temps de sommeil moyen pour la promotion. Et le paramètre σ mesure ici la variabilité entre les individus : si tous les individus disaient dormir autant, on aurait $\sigma = 0$.

Afin de savoir si le sommeil d'un étudiant est significativement différent de celui d'un individu quelconque, les enseignants de IF voudraient comparer le temps de sommeil moyen et la variabilité à ce qui est observé dans le reste de la population. Dans la population française, on observe qu'un individu dort en moyenne 7 heures par nuit et que l'écart-type est de 1.2 heure.

En appliquant les résultats qui précèdent, m peut être estimer par $\widehat{m}_n = 6.36$ heures et σ^2 par $\widehat{\sigma}_n^2 = \frac{30}{29} 1.80 = 1.85 \text{heures}^2$.

Remarque : Il n'est pas question ici de voir $m = 6.36h$! m reste un paramètre inconnu que vous n'avez pas mesuré : vous n'avez qu'une estimation et devez en rester conscients.

Nous observons que l'estimation de l'espérance m obtenue est plus petite que la moyenne de la population, et que l'estimation de la variance est plus élevée. Cependant, nous voyons ici les limites d'une estimation ponctuelle : les différences observées ne sont-elles pas simplement dues au fait que nous n'avons interrogé que 30 étudiants de la promotion ? Notre estimation est-elle assez fiable pour en conclure une différence notable ? Les enseignants de IF ne sont pas convaincus et ils sont demandeurs d'une methodologie plus rigoureuse permettant de répondre si oui ou non la différence observée est significative ou uniquement due aux aléas des sondages. La prochaine section aura

par conséquent pour but de regarder quelle est la précision de notre estimation et de déterminer dans quelle mesure nous pouvons nous y fier, lors d'une comparaison notamment.

3.2.2 Estimation d'une proportion

Soient $X_1, \ldots, X_n$ indépendantes et de même loi $\mathcal{B}(p)$. Alors $m = \mathbb{E}X_i$ et ainsi nous pouvons estimer p par $\widehat{p}_n = \overline{X_n}$. Remarquons que p est la proportion théorique de succès que nous cherchons à déterminer. Nous pouvons interpréter $\widehat{p}_n$ comme la proportion de succès observés dans l'échantillon.

Considérons un exemple.

Lors du deuxième tour de l'élection présidentielle, un sondage est effectué sur un échantillon de n personnes. On fait l'hypothèse que les réponses des sondés sont indépendantes, et que les sondés ne mentent pas. De plus on ne tiendra compte que des suffrages exprimés. Sur $n = 500$ personnes interrogées, 274 ont déclaré qu'elles voteraient pour le candidat A.

Soit $X_i = 1$ si la $i^{\text{ème}}$ personne vote pour le candidat A et 0 sinon. Les variables $X_1, \ldots, X_n$ sont supposées indépendantes et identiquement distribuées selon une loi $\mathcal{B}(p)$. Le paramètre p représente la proportion d'individus votant pour A dans la population totale. Il peut être estimé par la proportion observée dans l'échantillon : $\widehat{p}_n = 274/500 = 54.8\%$.

Remarquons que ce modèle est très imparfait : il ne tient pas compte de l'inhomogénéité de la population et du besoin de représentativité de l'échantillon, ni des changements d'opinion, des indécis, etc.

3.3 Intervalles de confiance

Nous avons vu à la section précédente comment estimer une valeur inconnue, c'est-à-dire comment proposer une valeur plausible pour cette grandeur inconnue. Mais nous commettons nécessairement une erreur : l'aléatoire fait que nous ne donnons pas exactement la valeur théorique, mais une valeur approchée. Le but est donc maintenant de donner cette marge d'erreur. Plus précisément nous allons construire un intervalle (ou une fourchette) dans lequel la grandeur recherchée a une probabilité forte de se trouver.

Définition 3.5. Soit θ un paramètre donné. On appelle intervalle de confiance de niveau β un intervalle aléatoire $[T_1, T_2]$ tel que

$$\mathbb{P}(\theta \in [T_1, T_2]) = \beta.$$

La raison pour laquelle il est précisé que l'intervalle est aléatoire est que les bornes T_1 et T_2 de cet intervalle sont des variables aléatoires.

L'idée d'un intervalle de confiance est donc de donner une plage de valeurs possibles avec un degré de confiance associé. Un intervalle $[T_1, T_2]$ de niveau 95% pour θ, signifie qu'il y a une probabilité de 95% que θ soit bien compris entre T_1 et T_2. Il n'est pas possible en général de donner un intervalle de longueur finie où l'on peut trouver θ avec une probabilité de 100%. On se fixe donc un taux d'erreur acceptable (*i.e.* on admet qu'on peut se tromper avec une probabilité de 5%, 1%, 0.5%...).

Pourquoi l'intervalle de confiance informe-t-il bien sur la précision ? Ceci est lié à sa largeur : plus un intervalle est large, plus les valeurs possibles sont étalées et donc moins on est précis. Nous pouvons remarquer qu'un intervalle de confiance va se rétrécir en fonction du nombre n d'observations réalisées : plus nous avons d'observations, plus nous disposons d'information sur la valeur plausible de θ, donc plus nous sommes précis.

Inversement, quand le niveau de confiance augmente, nous devons vérifier que la largeur d'intervalle augmente. En effet, cela signifie que nous augmentons la probabilité d'être dans l'intervalle.

3.3.1 Paramètres d'une loi normale

Soient $X_1, \ldots, X_n$ indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$. Nous avons vu précédemment comment estimer m et σ^2. A l'aide du théorème de Fisher, nous pouvons déterminer la précision des estimateurs que nous avons construits et nous en déduirons ainsi des intervalles de confiance.

a. Intervalle pour l'espérance

Nous allons essayer de construire un intervalle de confiance pour le paramètre m. Nous avons proposé ci-dessus d'estimer m par $\widehat{m}_n = \overline{X_n}$.

Nous souhaitons construire T_1 et T_2 tels que $\mathbb{P}(T_1 \leqslant m \leqslant T_2) = \beta$. La valeur $\widehat{m}_n$ étant considérée comme la plus vraisemblable pour m, nous allons chercher un intervalle centré en $\widehat{m}_n$. Ceci revient à supposer qu'existe a tel que $T_1 = \overline{X_n} - a$ et $T_2 = \overline{X_n} + a$. Nous voulons ainsi déterminer a tel que $\mathbb{P}\left(\overline{X_n} - a \leqslant m \leqslant \overline{X_n} + a\right) = \beta$.

Rappelons que d'après le Théorème de Fisher donné en section 1.2.3, nous avons

(i) $\frac{\overline{X_n}-m}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$

(ii) $\frac{\overline{X_n}-m}{S'_n/\sqrt{n}} \sim St(n-1)$.

L'inconvénient du résultat (i) est qu'il fait intervenir la variance σ^2 qui est inconnue. Utiliser (i) n'est donc pas cohérent (cela le serait si σ^2 était connu). Nous allons donc exploiter (ii).

Réécrivons l'égalité que doit vérifier a : $\mathbb{P}\left(-a \leqslant \overline{X_n} - m \leqslant a\right) = \beta$. Ceci équivaut à

$$\mathbb{P}\left(-\frac{a}{S'_n/\sqrt{n}} \leqslant \frac{\overline{X_n} - m}{S'_n/\sqrt{n}} \leqslant \frac{a}{S'_n/\sqrt{n}}\right) = \beta.$$

Notons $T_{n-1} = \frac{\overline{X_n}-m}{S'_n/\sqrt{n}}$. Alors $\mathbb{P}\left(-\frac{a}{S'_n/\sqrt{n}} \leqslant T_{n-1} \leqslant \frac{a}{S'_n/\sqrt{n}}\right) = \beta$, avec T_{n-1} suivant une loi de Student de paramètre $n-1$.

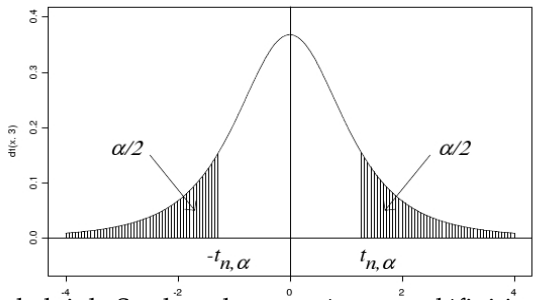

FIGURE 3.3 – Densité de la loi de Student de paramètre n et définition du quantile $t_{n,\alpha}$.

Au vu du schéma ci-dessus, nous pouvons voir que $\frac{a}{S'_n/\sqrt{n}} = t_{n-1,1-\beta}$ avec $t_{n-1,1-\beta}$ donné sur le schéma.

Par conséquent $a = t_{n-1,1-\beta}\frac{S'_n}{\sqrt{n}}$ et ainsi nous avons

$$\mathbb{P}\left(\overline{X_n} - t_{n-1,1-\beta}\frac{S'_n}{\sqrt{n}} \leqslant m \leqslant \overline{X_n} + t_{n-1,1-\beta}\frac{S'_n}{\sqrt{n}}\right) = \beta.$$

L'intervalle de confiance de niveau β recherché est :

$$\left[\overline{X_n} - t_{n-1,1-\beta}\frac{S'_n}{\sqrt{n}}\,;\, \overline{X_n} + t_{n-1,1-\beta}\frac{S'_n}{\sqrt{n}}\right].$$

On constate en effet que la longueur de cet intervalle diminue lorsque n augmente et augmente lorsque β augmente (au vu du schéma ci-dessus, $t_{n-1,1-\beta}$ augmente en fonction de β).

Exemple : Espérance d'une loi normale.
Reprenons l'exemple précédent où 30 étudiants de 4IF nous avaient donné leur temps moyen de sommeil sur une nuit en période de projet, que nous supposions issus d'une loi $\mathcal{N}(m, \sigma^2)$. Nous avions obtenu l'estimation de l'espérance $\widehat{m}_n = \overline{x_n} = 6,36$ heures et l'estimation de la variance $\widehat{\sigma}_n^2 = {s'_n}^2 = 1.85\,\text{heures}^2$. Par conséquent l'intervalle de confiance de niveau 95% $= 1 - \alpha$ de m est

$$\begin{aligned} IC_{95\%}(m) &= \left[\overline{x_n} - t_{n-1,\alpha}\frac{s'_n}{\sqrt{n}}\,;\,\overline{x_n} + t_{n-1,\alpha}\frac{s'_n}{\sqrt{n}}\right] \\ &= \left[6.36 - t_{29;5\%}\sqrt{\frac{1.85}{30}}\,;\,6.36 + t_{29;5\%}\sqrt{\frac{1.85}{30}}\right] \end{aligned}$$

La table nous donne $t_{29;5\%} = 2.045$ et par conséquent

$$IC_{95\%}(m) = [5.85; 6.87].$$

Ceci signifie que la vraie valeur de m appartient à cet intervalle avec une probabilité de 95%.

Les enseignants souhaitaient comparer à la moyenne de la population française qui est de 7 heures de sommeil par nuit. Nous pouvons constater que 7 n'appartient pas à l'intervalle de confiance donné ci-dessus. Nous pouvons en déduire que la moyenne des étudiants de IF est significativement différente de la moyenne du reste de la population, avec une probabilité de se tromper inférieure à 5%.

Peut-on affiner cette probabilité de se tromper ? Considérons l'intervalle de confiance d'ordre 99%. Nous obtenons

$$IC_{99\%}(m) = [5.68; 7.05].$$

Comme 7 appartient à cet intervalle, nous ne pouvons affirmer que les moyennes sont différentes avec un risque de se tromper inférieur à 1%. De même que précédemment, nous aimerions cependant une méthode permettant de déterminer plus précisément quelle est la probabilité de se tromper en affirmant que les deux valeurs sont significativement différentes. Ceci fera l'objet du chapître suivant.

b. Intervalle pour la variance

Nous allons essayer de construire un intervalle de confiance pour le paramètre σ^2. Nous avons proposé ci-dessus d'estimer σ^2 par $\widehat{\sigma}_n^2 = {S'_n}^2$.

Nous souhaitons construire T_1 et T_2 tels que $\mathbb{P}\left(T_1 \leqslant \sigma^2 \leqslant T_2\right) = \beta$. De même que pour l'intervalle de l'espérance, nous allons construire l'intervalle en exploitant le comportement de l'estimateur.

Rappelons que d'après le Théorème de Fisher donné en section 1.2.3, nous avons

$$\frac{(n-1){S'_n}^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Il semble judicieux de faire apparaître la quantité $K_{n-1} = \frac{(n-1){S'_n}^2}{\sigma^2}$ dans l'égalité recherchée. Nous avons

$$\begin{aligned}\mathbb{P}\left(T_1 \leqslant \sigma^2 \leqslant T_2\right) &= \mathbb{P}\left(\frac{1}{T_2} \leqslant \frac{1}{\sigma^2} \leqslant \frac{1}{T_1}\right) \\ &= \mathbb{P}\left(\frac{(n-1){S'_n}^2}{T_2} \leqslant \frac{(n-1){S'_n}^2}{\sigma^2} \leqslant \frac{(n-1){S'_n}^2}{T_1}\right).\end{aligned}$$

Si nous voulons avoir un intervalle centré sur l'estimateur, nous devons prendre T_1 et T_2 tels que

$$\begin{aligned}\mathbb{P}\left(K_{n-1} \leqslant \frac{(n-1){S'_n}^2}{T_2}\right) &= (1-\beta)/2 \\ \mathbb{P}\left(K_{n-1} \geqslant \frac{(n-1){S'_n}^2}{T_1}\right) &= (1-\beta)/2\end{aligned}$$

avec K_{n-1} variable aléatoire de loi χ^2_{n-1}.

Au vu du schéma ci-après, nous pouvons voir que $\frac{(n-1){S'_n}^2}{T_1} = z_{n-1,(1-\beta)/2}$ et $\frac{(n-1){S'_n}^2}{T_2} = z_{n-1,1-(1-\beta)/2}$ avec $z_{n-1,(1-\beta)/2}$ et $z_{n-1,1-(1-\beta)/2}$ définis par le schéma.

Par conséquent $T_1 = \frac{(n-1){S'_n}^2}{z_{n-1,(1-\beta)/2}}$ et $T_2 = \frac{(n-1){S'_n}^2}{z_{n-1,1-(1-\beta)/2}}$. Nous en déduisons que

$$\mathbb{P}\left(\frac{(n-1){S'_n}^2}{z_{n-1,(1-\beta)/2}} \leqslant \sigma^2 \leqslant \frac{(n-1){S'_n}^2}{z_{n-1,1-(1-\beta)/2}}\right) = \beta.$$

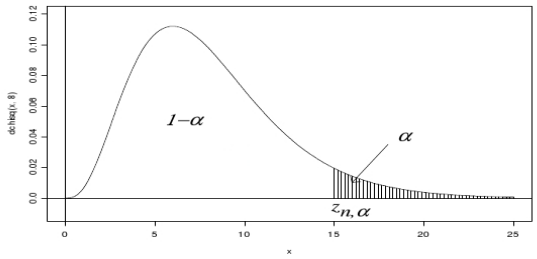

FIGURE 3.4 – Densité de la loi du χ^2 de paramètre n et définition du quantile $z_{n,\alpha}$.

L'intervalle de confiance de niveau β recherché est :

$$\left[\frac{(n-1){S'_n}^2}{z_{n-1,(1-\beta)/2}}\,;\,\frac{(n-1){S'_n}^2}{z_{n-1,1-(1-\beta)/2}}\right].$$

Afin de simplifier l'écriture, on prefère poser $\alpha = 1-\beta$. L'intervalle ci-dessus s'écrit alors

$$\left[\frac{(n-1){S'_n}^2}{z_{n-1,\alpha/2}}\,;\,\frac{(n-1){S'_n}^2}{z_{n-1,1-\alpha/2}}\right].$$

Exemple : Variance d'une la loi normale.
Dans l'exemple sur le temps moyen de sommeil en 4IF, l'intervalle de confiance de niveau $95\% = 1-\alpha$ de la variance σ^2 est

$$\begin{aligned} IC_{95\%}(\sigma^2) &= \left[\frac{(n-1){s'_n}^2}{z_{n-1;\alpha/2}}\,;\,\frac{(n-1){s'_n}^2}{z_{n-1;1-\alpha/2}}\right] \\ &= \left[\frac{29.1,85}{z_{29;2.5\%}}\,;\,\frac{29.1,85}{z_{29;97.5\%}}\right] \end{aligned}$$

La table nous donne $z_{29;2.5\%} = 45,72$ et $z_{29;97.5\%} = 16,05$. D'où

$$IC_{95\%}(\sigma^2) = [1.17; 3.35].$$

Ceci signifie que σ^2 appartient à cet intervalle avec une probabilité de 95%.

Dans la population française la variance est de 1.44 heures^2. La valeur 1.44 appartient à $IC_{95\%}(\sigma^2)$. Nous ne pouvons donc pas affirmer que la variance en 4IF est significativement différente du reste de la population : si nous affirmons le contraire nous nous trompons avec une probabilité supérieure à 5%. La différence observée est donc imputable aux aléas des réalisations, au fait que les données ne sont pas exhaustives, etc. De même qu'auparavant, nous aimerions pouvoir évaluer le risque exact de se tromper en affirmant que les 4IF admettent une variabilité plus forte, ce qui sera fait ultérieurement.

BILAN

Pour les paramètres une loi normale :

Soient $X_1, \dots, X_n$ indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$, alors l'estimateur de m est donné par $\widehat{m}_n = \overline{X_n}$ et l'estimateur de σ^2 est donné par $\widehat{\sigma}_n^2 = {S'_n}^2$. Les intervalles de niveau $1-\alpha$ des paramètres m et σ^2 valent respectivement :

$$\begin{aligned} IC_{1-\alpha}(m) &= \left[\overline{X_n} - t_{n-1,\alpha}\frac{S'_n}{\sqrt{n}} ; \overline{X_n} + t_{n-1,\alpha}\frac{S'_n}{\sqrt{n}}\right], \\ IC_{1-\alpha}(\sigma^2) &= \left[\frac{(n-1){S'_n}^2}{z_{n-1;\alpha/2}} ; \frac{(n-1){S'_n}^2}{z_{n-1;1-\alpha/2}}\right]. \end{aligned}$$

Les notations $t_{k,\alpha}$ et $z_{k;\beta}$ correspondent aux définitions suivantes :
- Si $T \sim St(k)$, $\mathbb{P}(|T| \geqslant t_{k,\alpha}) = \alpha$, soit encore $\mathbb{P}(T \geqslant t_{k,\alpha}) = \alpha/2$.
- Si $Z \sim \chi^2_k$, alors $\mathbb{P}(Z \geqslant z_{k,\beta}) = \beta$.

Remarquons que l'écriture choisie est légérement différente de celle utilisée dans les preuves. La raison est qu'il est en général plus facile de les retenir sous cette forme, notamment car celle-ci est très similaire à ce qui sera vu ultérieurement pour la construction de tests.

3.3.2 Intervalle pour une proportion

Soient $X_1, \dots X_n$ indépendants et de loi de Bernouilli $\mathcal{B}(p)$. Une proportion p est estimée par la proportion observée $\widehat{p}_n = \overline{X_n}$. Remarquons que $n\widehat{p}_n$ suit une loi binomiale $\mathcal{B}(n, p)$. Nous pouvons exploiter ce résultat afin de construire un intervalle de confiance pour le paramètre p. Cependant, la loi binomiale n'étant pas facile à manipuler, nous optons ici pour l'utilisation du théorème de la limite centrale. En effet, celui-ci nous assure que

$$U = \frac{\widehat{p}_n - p}{\sqrt{p(1-p)/n}} \xrightarrow[n\to\infty]{loi} \mathcal{N}(0,1).$$

L'inconvénient est que ce résultat étant asymptotique, il mènera à des intervalles de confiance asymptotiques et non à des intervalles exacts.

Définition 3.6. Soit θ un paramètre donné. On appelle intervalle de confiance asymptotique de niveau β un intervalle aléatoire $[T_{1,n}, T_{2,n}]$ tel que

$$\mathbb{P}\left(\theta \in [T_{1,n}, T_{2,n}]\right) \xrightarrow[n\to\infty]{} \beta.$$

Supposons n suffisamment gand pour que nous ayons l'approximation $U \sim \mathcal{N}(0,1)$. Au vu du schéma ci-dessous,

$$\mathbb{P}(-u_\alpha \leqslant U \leqslant u_\alpha) = 1 - \alpha,$$

avec u_α quantile de la loi normale centrée réduite défini par le graphique.

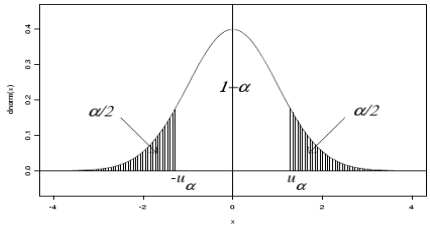

FIGURE 3.5 – Densité de la loi de Gauss de paramètres 0 et 1 et définition du quantile u_α.

Nous avons donc l'ínégalité

$$U^2 = \frac{(\widehat{p}_n - p)^2}{p(1-p)/n} \leqslant u_\alpha^2$$

vérifiée avec une probabilité $1 - \alpha$. Ceci équivaut à résoudre

$$\left(1 + \frac{u_\alpha^2}{n}\right) p^2 - \left(2\widehat{p}_n + \frac{u_\alpha^2}{n}\right) p + \widehat{p}_n^2 \leqslant 0.$$

Nous savons que cette inégalité est vérifiée avec une probabilité $1 - \alpha$. Ainsi le paramètre p a une probabilité $1 - \alpha$ d'être solution de cette inéquation, donc d'appartenir à l'ensemble des valeurs vérifiant l'inégalité. Si nous notons S_α l'ensemble des solutions, nous avons $\mathbb{P}(p \in S_\alpha) = 1 - \alpha$. L'ensemble S_α est alors l'intervalle de confiance recherché.

Reste à résoudre l'inégalité donnée ci-dessus. Le discriminant du polynôme vaut

$$\Delta = u_\alpha^2(u_\alpha^2/n^2 + 4\widehat{p}_n(1-\widehat{p}_n)/n).$$

Les racines du polynôme en p sont

$$\frac{2\widehat{p}_n + u_\alpha^2/n \pm u_\alpha\sqrt{u_\alpha^2/n^2 + 4\widehat{p}_n(1-\widehat{p}_n)/n}}{2(1+u_\alpha^2/n)}.$$

Le polynôme est négatif entre ses racines, donc les solutions de l'inégalité sont données par l'intervalle

$$\left[\frac{2\widehat{p}_n + u_\alpha^2/n - u_\alpha\sqrt{u_\alpha^2/n^2 + 4\widehat{p}_n(1-\widehat{p}_n)/n}}{2(1+u_\alpha^2/n)} ; \frac{2\widehat{p}_n + u_\alpha^2/n + u_\alpha\sqrt{u_\alpha^2/n^2 + 4\widehat{p}_n(1-\widehat{p}_n)/n}}{2(1+u_\alpha^2/n)}\right].$$

Nous pourrions exploiter directement cette conclusion. Cependant, nous raisonnons en asymptotique, c'est-à-dire que nous considérons n grand et que nous acceptons des approximations (par l'application du théorème de la limite centrale). Nous pouvons donc essayer de simplifier l'intervalle obtenu. En remarquant que u_α^2 est négligeable devant n, nous en déduisons l'intervalle

$$\left[\widehat{p}_n - u_\alpha\sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}\,;\,\widehat{p}_n + u_\alpha\sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}\right].$$

Au vu de notre raisonnement ci-dessus, le paramètre p a une probabilité approximativement de $1-\alpha$ d'appartenir à cet intervalle, pour n suffisamment grand. Cet intervalle est donc un intervalle de confiance asymptotique de niveau $1-\alpha$ pour le paramètre p.

BILAN

Pour une proportion :

Soient $X_1, \ldots, X_n$ indépendantes et de même loi $\mathcal{B}(p)$, alors l'estimateur de p est donné par $\widehat{p}_n = \overline{X_n}$. L'intervalle asymptotique de niveau $1-\alpha$ obtenu est :

$$IC_{1-\alpha}(p) = \left[\widehat{p}_n - u_\alpha\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}\,;\,\widehat{p}_n + u_\alpha\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}\right],$$

où u_α est défini comme suit :
Pour tout $U \sim \mathcal{N}(0,1)$, $\mathbb{P}(|U| \geqslant u_\alpha) = \alpha$, soit encore $\mathbb{P}(U \geqslant u_\alpha) = \alpha/2$.

Remarque : Le principe de construction d'un intervalle de confiance est toujours le même : se ramener à une variable aléatoire faisant intervenir le paramètre d'intérêt dont on connaisse la loi de manière exhaustive (*i.e.* aucun paramètre de la loi n'est inconnu). Une telle variable aléatoire est dite *fonction pivotale*.

Exemple : Sur $n = 500$ personnes interrogées, 274 ont déclaré qu'elles voteraient pour le candidat A. Si p représente le score de A aux élections, nous estimions p par $\widehat{p}_n = 274/500 = 54.8\%$. Comme $u_{5\%} = 1.96$, l'intervalle de confiance asymptotique de niveau 95% pour p est $IC_{95\%}(p) = [50.44\%; 59.16\%]$. Au seuil $\alpha = 1\%$, $u_{1\%} = 2.5758$, est ainsi l'intervalle est $IC_{99\%}(p) = [49.07\%; 60.53\%]$.

Nous pouvons constater qu'avec une confiance de 95%, la valeur 50% n'appartient pas à l'intervalle. Cela signifie que dans ce modèle simplifié, nous pouvons affirmer que la candidat A va gagner avec une probabilité supérieure à 95%. Cependant, nous pouvons re-

marquer que la conclusion change avec un degré de confiance de 99% : il y a donc d'après les intervalles de confiance une probabilité supérieure à 1% que A perde les élections.

Pour quel degré de confiance a-t-on la borne inférieure exactement égale à 50% ? Ceci est vérifié lorsque

$$\widehat{p}_n - u_\alpha \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}} = 0.5.$$

Nous pouvons écrire cette équation sous la forme :

$$u_\alpha = \frac{\widehat{p}_n - 0,5}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}} = 2.16.$$

Par définition de u_α cela signifie que pour toute variable aléatoire U de loi $\mathcal{N}(0,1)$, nous avons $\mathbb{P}(U \leqslant 2.16) = 1 - \alpha/2$. Ainsi $\alpha = 2(1 - \mathbb{P}(U \leqslant 2.16))$. La lecture de la table de la fonction de répartition d'une loi normale centrée réduite nous donne $\alpha = 2(1 - 0.9846) = 3.08\%$. Si nous acceptons de nous tromper avec une probabilité supérieure à 3.08%, nous pouvons affirmer au vu des intervalles de confiance que A va gagner. Nous ne pouvons pas conclure quant au résultat si nous ne voulons pas nous tromper avec une probabilité inférieure à 3.08%.

Nous venons ici de chercher à évaluer la probabilité avec laquelle nous pouvions affirmer que $p > 50\%$ sans nous tromper. Le but du chapître suivant est de définir une approche équivalente, généralisable à d'autres hypothèses, qui soit un peu plus rigoureuse.

CHAPITRE 4

TESTS D'HYPOTHÈSES PARAMÉTRIQUES

Notre but est maintenant de vérifier si une hypothèse est ou non valide. L'idée est que la prise de décision qui s'ensuit dépend de cette hypothèse. Le premier problème sera donc de voir quelle hypothèse importe dans la décision à prendre, puis de la formuler en fonction de paramètres. Par exemple, dans l'exemple du temps de sommeil en IF, les enseignants veulent savoir si un aménagement de l'emploi du temps est nécessaire et souhaitent donc savoir si les étudiants dorment assez. En médecine, les médecins sont à la recherche d'aide au diagnostic : la prise de sang permet-elle de dire si le patient va développer une maladie ? En industrie pharmaceutique, le nouveau médicament développé est-il plus performant ? Une entreprise cherche à savoir si elle a intérêt à renouveler son parc informatique, des experts à savoir si un produit vérifie les normes en vigueur, etc.

Une fois formulée l'hypothèse qui nous intéresse, nous voulons étudier sa vraisemblance. Nous allons pour cela nous intéresser à la notion de test statistique. Un test est une procédure qui permet de décider si à partir des observations obtenues nous devons accepter ou rejeter l'hypothèse concernée. En raison des aléas, un tel test ne peut être catégorique : il faut accepter de se tromper dans la conclusion, mais en sachant avec quelle probabilité nous risquons de nous tromper.

Que signifie « se tromper » ? Etudions ceci sur un exemple. Dans une exploitation, un expert mesure sur différents légumes le taux d'un pesticide donné. Il obtient un indicateur de 16, sachant que la norme à ne pas dépasser est 15. Doit-il interdire la mise en vente des légumes ? L'hypothèse que veut tester l'expert est « la norme est dépassée ». Il y a deux façons de se tromper dans sa conclusion :

— S'il conclue que la norme est dépassée alors qu'en réalité ce n'est pas le cas mais que l'écart dans les observations est uniquement dû aux aléas des mesures, alors l'exploitant agricole aura une sanction financière non justifiée.
— S'il conclue que la norme n'est pas dépassée alors qu'en réalité c'est le cas, alors il

y a un risque de santé publique pour les consommateurs.

Notons (H_0) l'hypothèse que nous souhaitons tester. En fait, lorsque nous testons une hypothèse (H_0), nous testons en réalité si cette hypothèse est plus vraisemblable qu'une hypothèse alternative (H_1). On appelle (H_0) l'hypothèse nulle et (H_1) l'hypothèse alternative. Dans le cadre de ce cours, nous considérerons toujours (H_0) et (H_1) complémentaires. Au vu de l'exemple ci-dessus, on distingue deux types de risque :

— On appelle risque de $1^{\text{ère}}$ espèce le risque de rejeter (H_0) à tort.

— On appelle risque de $2^{\text{ème}}$ espèce le risque de ne pas rejeter (H_0) à tort.

Idéalement, nous voudrions minimiser ces deux risques simultanément dans notre procédure de test. Mais ceci n'est pas faisable. Par convention, nous décidons de contrôler le rique de $1^{\text{ère}}$ espèce.

Remarque : Les conclusions d'un test s'expriment toujours comme "on rejette (H_0)" ou "on ne rejette pas (H_0)". La nuance avec "on accepte (H_0)" est subtile, mais ceci est dû au fait que la procédure que nous développerons ne permet pas d'affirmer que (H_0) est réalisée, uniquement de conclure si (H_0) est plausible ou non.

Un test de seuil α est un test dont le risque de $1^{\text{ère}}$ espèce vaut α. Autrement dit la probabilité de conclure que (H_0) est faux lorsque (H_0) est vérifiée vaut α.

Le choix des hypothèses se fait ensuite à partir de la formulation du risque qui nous intéresse. La personne qui réalise le test veut minimiser le risque de première espèce. Dans l'exemple ci-dessus, l'expert indépendant veut minimiser le risque d'affirmer que la norme de pesticides est respectée à tort, en raison du risque pour les consommateurs. Son hypothèse (H_0) est donc « la norme est dépassée ». Un expert mandaté par l'agriculteur choisira au contraire (H_0) « la norme est respectée ».

Prenons un autre exemple. En finance, comment déterminer si une opération financière doit ou non être lancée ? Si (H_0) est « l'opération peut être lancée », alors le risque contrôlé est celui de ne pas lancer l'opération alors qu'elle est rentable : on ne veut pas se priver de bénéfices et on préfère tenter l'opération, quitte à perdre des sous. (H_0) est « l'opération ne peut pas être lancée », alors le risque contrôlé est celui de lancer l'opération alors qu'elle n'est pas rentable : on préfère rester prudent et ne pas perdre de sous, quitte à ne pas en gagner non plus.

Détaillons en quoi consiste plus précisément un test. Nous avons une prise de décision et un risque associé que nous voudrions contrôler. Comment au vu des données allons-nous procéder ? Les étapes d'un test sont les suivantes :

1. Formaliser le problème et la décision à prendre.

2. Expliciter le risque que l'on cherche à minimiser. En déduire les hypothèses (H_0) et (H_1).
3. Choisir le seuil du risque α selon la gravité des conséquences : plus α est petit plus le risque associé est petit. On prend en général α inférieur à 5%.
4. Construire une règle de décision, c'est-à-dire une procédure qui permette de dire si on accepte ou non (H_0) au vu des données $x_1, \dots x_n$. Cette procédure consiste à trouver une **région critique** RC_α telle que
   — si $\{x_1, \dots x_n\} \in RC_\alpha$ on rejette (H_0),
   — si $\{x_1, \dots x_n\} \notin RC_\alpha$ on ne rejette pas (H_0),
5. Les observations $x_1, \dots x_n$ appartiennent-elles à RC_α ? Conclure quant au rejet ou non-rejet de l'hypothèse (H_0).
6. Répondre au problème posé.

Reste à construire la région critique. La région critique est en fait une condition telle que si nos observations la vérifie, on rejette (H_0). Comment déterminer ces conditions ? Nous allons distinguer deux types de tests pour la construction de cette région :

— Les tests paramétriques : les données que nous observons sont modélisées. Notre hypothèse peut se formuler à l'aide d'un paramètre θ, que notre modèle permet d'estimer. Nous n'avons pas accès à la vraie valeur de θ mais nous allons prendre notre décision au vu de son estimation et de sa précision. (Exemples : le degré de pesticides dans un légume suit une loi normale, d'espérance θ et le test se formulera sur θ ; la réponse à un sondage est une loi $\mathcal{B}(p)$ avec p le score aux élections : le candidat gagne si $p \geqslant 50\%\dots$)

— Les tests non paramétriques : nous voulons tester une hypothèse indépendamment de toute modélisation préalable de nos données. Par exemple, nous voulons tester si les données suivent bien une loi normale ou une loi de Poisson, nous voulons tester si deux variables sont corrélées, etc. Ceci fera l'objet du chapître suivant.

4.1 Test sur un paramètre

Dans cette section, nous nous intéressons aux tests ne faisant intervenir qu'une seule grandeur estimée. Le but est de comparer un paramètre θ inconnu avec une grandeur θ_0 donnée, soit car elle correspond à une grandeur physique connue que vous souhaitez vérifier, soit parce que la comparaison permet d'aider dans la prise de décision qui vous intéresse. Afin de réaliser ces tests, nous allons donc devoir estimer le paramètre θ. La conclusion de notre test dépendra alors de la précision de notre estimation. Plus

nous sommes précis plus nous pouvons rejetter l'hypothèse (H_0) formulée de manière affirmative.

4.1.1 Test sur l'espérance d'une loi normale

Soit $X_1, \dots X_n$ n variables aléatoires indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$. Les paramètres m et σ^2 sont inconnus. Nous allons construire le test de (H_0) $m \leqslant m_0$ contre (H_1) $m > m_0$ de risque α.

La première étape consiste à estimer m : d'après ce qui précède, nous proposons d'estimer m par $\overline{X_n}$. La région critique RC_α est la zone où l'on considère que l'hypothèse (H_1) est plus vraisemblable que (H_0). Il semble logique de chercher RC_α sous la forme $RC_\alpha = \{\overline{X_n} > c_\alpha\}$ avec $c_\alpha > m_0$. En effet, étant donné l'erreur commise par l'approximation de m par $\overline{X_n}$, nous n'accepterons $m > m_0$ que si $\overline{X_n}$ est significativement plus grand. Le seuil choisi c_α sera d'autant plus proche de m_0 que notre estimation sera précise.

Par définition du risque, nous avons $\sup_{(H_0)} \mathbb{P}(\{X_1, \dots X_n\} \in RC_\alpha) = \alpha$. Donc

$$\sup_{m \leqslant m_0} \mathbb{P}(\overline{X_n} > c_\alpha) = \alpha.$$

Afin de calculer ces probabilités, nous avons besoin de la loi de $\overline{X_n}$. De même que pour la construction des intervalles de confiance, nous utilisons le Théorème de Fisher :

$$T_n = \frac{\overline{X_n} - m}{S'_n/\sqrt{n}} \sim St(n-1).$$

Alors

$$\mathbb{P}(\overline{X_n} > c_\alpha) = \mathbb{P}\left(T_n > \frac{c_\alpha - m}{s'_n/\sqrt{n}}\right) = 1 - F_{St(n-1)}\left(\frac{c_\alpha - m}{s'_n/\sqrt{n}}\right),$$

avec $F_{St(n-1)}$ fonction de répartition de la loi de Student $St(n-1)$. Une fonction de répartition est une fonction croissante. Par conséquent $1 - F_{St(n-1)}\left(\frac{c_\alpha - m}{s'_n/\sqrt{n}}\right)$ est une fonction croissante de m. Alors

$$\sup_{m \leqslant m_0} 1 - F_{St(n-1)}\left(\frac{c_\alpha - m}{s'_n/\sqrt{n}}\right) = 1 - F_{St(n-1)}\left(\frac{c_\alpha - m_0}{s'_n/\sqrt{n}}\right).$$

Ainsi, nous avons

$$1 - F_{St(n-1)}\left(\frac{c_\alpha - m_0}{s'_n/\sqrt{n}}\right) = \mathbb{P}\left(T_n > \frac{c_\alpha - m_0}{s'_n/\sqrt{n}}\right) = \alpha.$$

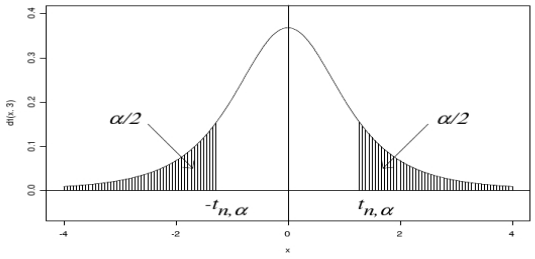

FIGURE 4.1 – Densité de la loi de Student de paramètre n et définition du quantile $t_{n,\alpha}$.

Au vu du dessin ci-dessus, $\frac{c_\alpha - m_0}{s'_n/\sqrt{n}} = t_{n-1;2\alpha}$ avec $t_{n-1;2\alpha}$ défini par le schéma. D'où $c_\alpha = m_0 + \frac{s'_n}{\sqrt{n}} t_{n-1;2\alpha}$. En conclusion la région critique obtenue est

$$RC_\alpha = \{\overline{X_n} > m_0 + \frac{s'_n}{\sqrt{n}} t_{n-1;2\alpha}\}.$$

L'intérêt de cette formulation est sa similitude avec celle des intervalles de confiance. Cependant, nous pouvons remarquer que la région critique s'écrit aussi :

$$RC_\alpha = \{T_n > t_{n-1;2\alpha}\}, \text{ avec } T_n = \frac{\overline{X_n} - m_0}{S'_n/\sqrt{n}}.$$

Cette écriture présente le grand intérêt que si vous voulez réaliser un test pour différents seuils, vous calculez la valeur de T_n une seule fois, ce qui allège considérablement le calcul.

Nous ne redémontrons pas les autres régions critiques ici, car leur construction est identique à ce qui précède. Au final,

BILAN

Test sur l'espérance d'une loi normale :

Soient $X_1, \dots, X_n$ indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$, alors l'estimateur de m est donné par $\widehat{m}_n = \overline{X}_n$.
Si $m = m_0$, on a

$$T = \frac{\overline{X_n} - m_0}{S'_n/\sqrt{n}} \sim \mathcal{S}t(n-1).$$

(H_0)	(H_1)	**Région critique** RC_α
$m = m_0$	$m \neq m_0$	$\{\lvert T\rvert > t_{n-1;\alpha}\}$
$m \leqslant m_0$	$m > m_0$	$\{\ T > t_{n-1;2\alpha}\}$
$m \geqslant m_0$	$m < m_0$	$\{\ T < -t_{n-1;2\alpha}\}$

Exemple : Reprenons l'exemple du temps de sommeil des 4IF : le temps de sommeil par nuit en période de projet est supposé suivre une loi $\mathcal{N}(m, \sigma^2)$. Après interrogation de 30 étudiants de 4IF, nous avions obtenu $\overline{x_n} = 6,36$ heures et $s'^2_n = 1.85$ heures^2. Les enseignants de 4IF souhaitent savoir si l'espérance de sommeil est significativement inférieure au temps moyen de sommeil des autres individus de la poulation, qui est de 7 heures.

Les enseignants veulent minimiser le risque de déclarer à tort que les élèves ne dorment pas assez. Ils posent donc les hypothèses (H_0) $m \geqslant 7$ et (H_1) $m < 7$. La région critique de ce test est $RC_\alpha = \{T_n < -t_{n-1;2\alpha}\}$, avec $T_n = \frac{\overline{X_n} - m}{S'_n/\sqrt{n}}$. La réalisation de T_n vaut $t_n = -2.57$. Au seuil $\alpha = 5\%$, nous comparons cette grandeur à $t_{29;10\%} = 1.699$. Nous avons $t_n < -t_{29;10\%}$. Nous sommes dans la région critique $RC_{5\%}$. Alors nous rejettons (H_0). Nous pouvons conclure que les étudiants de 4IF en période de projet dorment significativement moins que 7 heures.

Avec les mêmes observations, quel risque prendre pour rejeter $m < 7$? Observons la conclusion du test lorsque α diminue.

α	5%	2.5%	1%	0.5%
$t_{29;2\alpha}$	1.699	2.045	2.462	2.756
observations dans RC_α	oui	oui	oui	non

Nous observons que la conclusion n'est plus la même pour $\alpha = 0.5\%$: nous rejetons (H_0) pour $\alpha \geqslant 1\%$ et nous ne la rejetons pas pour $\alpha \leqslant 0.5\%$. Ceci signifie que si nous affirmons que (H_1) est valide, la probabilité de se tromper est comprise entre 0.5% et 1%.

On peut montrer que la valeur critique à partir de laquelle la conclusion du test change est $\alpha = 0.78\%$.

La notion mise en évidence ici est fondamentale : il s'agit de la **p-valeur** d'un test. La p-valeur est la valeur α_c du risque telle que, lorsqu'on réalise le test avec un seuil α :

$$\begin{cases} \text{si } \alpha \leqslant \alpha_c \text{ on ne rejette pas } (H_0), \\ \text{si } \alpha > \alpha_c \text{ on rejette } (H_0). \end{cases}$$

Si la p-valeur est petite cela signifie donc que l'on peut rejeter (H_0) avec un faible risque de se tromper. En général, on cherchera donc à avoir une p-valeur petite, de sorte à pouvoir valider l'hypothèse (H_1) avec un faible risque de se tromper.

Remarque : Souvent les logiciels ne donnent que la p-valeur : elle résume à elle seule le résultat du test pour tous les seuils et donne une information supplémentaire qu'est le risque maximal que l'on prend en rejetant (H_0).

Lien avec les intervalles de confiance

Etudions le cas du test de (H_0) $m = m_0$ et (H_1) $m \neq m_0$. La région critique est

$$RC_\alpha = \{\overline{X_n} < m_0 - \frac{s'_n}{\sqrt{n}} t_{n-1;\alpha} \text{ ou } \overline{X_n} > m_0 + \frac{s'_n}{\sqrt{n}} t_{n-1;\alpha}\}.$$

Nous avons

$$\begin{aligned} \{X_1, \ldots, X_n\} \notin RC_\alpha &\Leftrightarrow m_0 - \frac{s'_n}{\sqrt{n}} t_{n-1;\alpha} \leqslant \overline{X_n} \leqslant m_0 + \frac{s'_n}{\sqrt{n}} t_{n-1;\alpha} \\ &\Leftrightarrow \overline{X_n} - \frac{s'_n}{\sqrt{n}} t_{n-1;\alpha} \leqslant m_0 \leqslant \overline{X_n} + \frac{s'_n}{\sqrt{n}} t_{n-1;\alpha} \\ &\Leftrightarrow m_0 \in IC_{1-\alpha}(m), \end{aligned}$$

où $IC_{1-\alpha}(m)$ est l'intervalle de confiance de niveau $1-\alpha$ construit dans le chapître précédent. Ainsi, ne pas rejeter (H_0), c'est-à-dire considérer que m_0 est une valeur plausible pour m équivaut à vérifier que m_0 appartient à l'intervalle de confiance.

Dans le cas d'un test de (H_0) $m \leqslant m_0$ et (H_1) $m > m_0$, on peut montrer que ceci équivaut à regarder si m_0 appartient à un intervalle de confiance de la forme $[m_{inf}; +\infty[$. Et tester (H_0) $m \geqslant m_0$ contre (H_1) $m > m_0$, équivaut à regarder si m_0 appartient à un intervalle de confiance de la forme $]-\infty; m_{sup}]$.

Attention : il n'est pas équivalent de tester (H_0) $m = m_0$ contre (H_1) $m \neq m_0$ ou de tester (H_0) $m \leqslant m_0$ contre (H_1) $m > m_0$. Si nous avons observé une estimation $\widehat{m}_n < m_0$, le second test validera systématiquement (H_0), pas le premier. Si par contre nous avons observé $\widehat{m}_n > m_0$, alors il est équivalent de réaliser le premier test avec un risque α et de réaliser le deuxième test avec un risque $\alpha/2$.

4.1.2 Test sur la variance d'une loi normale

Soient $X_1, \dots X_n$ n variables aléatoires indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$. Les paramètres m et σ^2 sont inconnus. Nous allons construire le test de (H_0) $\sigma^2 \geqslant \sigma_0^2$ contre (H_1) $\sigma^2 < \sigma_0^2$ de risque α.

Nous estimons σ^2 par ${S'_n}^2$. Nous savons que $\frac{(n-1){S'_n}^2}{\sigma^2}$ suit une loi $\chi^2(n-1)$. La région critique RC_α étant la zone où l'on considère que $\sigma^2 < \sigma_0^2$ est plus plausible, la forme de RC_α est : $RC_\alpha = \{{S'_n}^2 < c_\alpha\}$ avec $c_\alpha < \sigma_0^2$. Par définition du risque, nous avons :

$$\begin{aligned}
\alpha &= \sup_{\text{sous } (H_0)} \mathbb{P}(RC_\alpha) \\
&= \sup_{\sigma^2 \geqslant \sigma_0^2} \mathbb{P}({S'_n}^2 < c_\alpha) \\
&= \sup_{\sigma^2 \geqslant \sigma_0^2} \mathbb{P}\left(K_n < \frac{(n-1)c_\alpha}{\sigma^2}\right) \text{ avec } K_n \text{ de loi } \chi^2(n-1).
\end{aligned}$$

Par croissance des fonction de répartition, nous en déduisons que $\alpha = \mathbb{P}(K_n < \frac{(n-1)c_\alpha}{\sigma_0^2})$. Au vu du schéma ci-dessous, nous avons donc $\frac{(n-1)c_\alpha}{\sigma_0^2} = z_{n-1;1-\alpha}$, avec $z_{n-1;1-\alpha}$ défini sur le schéma. Il vient : $c_\alpha = \frac{\sigma_0^2}{n-1} z_{n-1;1-\alpha}$.

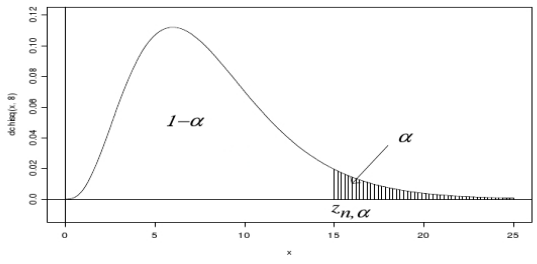

FIGURE 4.2 – Densité de la loi du χ^2 de paramètre n et définition du quantile $z_{n,\alpha}$.

La région critique est ainsi donnée par $RC_\alpha = \{S_n'^2 < \frac{\sigma_0^2}{n-1} z_{n-1;1-\alpha}\}$, ce qui peut aussi s'écrire sous la forme $RC_\alpha = \{K_n < z_{n-1;1-\alpha}\}$, avec $K_n = \frac{(n-1)S_n'^2}{\sigma_0^2}$.

En raisonnant de même pour les autres tests, nous obtenons :

BILAN

Test sur la variance d'une loi normale :

Soient $X_1, \dots, X_n$ indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$. L'estimateur de σ^2 est donné par $\widehat{\sigma}_n^2 = S_n'^2$.
Si $\sigma^2 = \sigma_0^2$, on a

$$K = \frac{(n-1)S_n'^2}{\sigma_0^2} \sim \chi^2(n-1).$$

(H_0)	(H_1)	**Région critique** RC_α
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\{K < z_{n-1;1-\alpha/2}$ ou $K > z_{n-1;\alpha/2}\}$
$\sigma^2 \leqslant \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\{K > z_{n-1;\alpha}\}$
$\sigma^2 \geqslant \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\{K < z_{n-1;1-\alpha}\}$

Remarquons que l'équivalence avec les intervalles de confiance établie pour l'espérance est toujours valable.

Exemple : Rappelons que 30 étudiants de 4IF avaient donné leur temps moyen de sommeil sur une nuit en période de projet et que nous avions obtenu $\overline{x_n} = 6.36$ heures et $s_n'^2 = 1.85$ heures^2. Dans la population, l'écart-type vaut 1,2 heure. Les enseignants de IF voudraient savoir si la variabilité est plus forte au sein de la promotion de IF que dans le reste de la population. Ils posent les hypothèses (H_0) $\sigma^2 \leqslant 1,2^2$ et (H_1) $\sigma^2 > 1,2^2$. La région critique de ce test est $RC_\alpha = \{K_n > z_{n-1;\alpha}\}$, avec $K_n = \frac{(n-1)S_n'^2}{1.2^2}$. La réalisation de K_n vaut $k_n = 37.26$. Au seuil $\alpha = 5\%$, nous comparons cette grandeur à $z_{29;5\%} = 42.56$. Nous avons $k_n < z_{29;5\%}$. Nous ne sommes pas dans la région critique $RC_{5\%}$. Nous ne pouvons pas conclure que la variabilité du sommeil au sein de la promotion de 4IF est significativement plus importante que dans le reste de la population.

(La p-valeur de ce test vaut environ 14%.)

4.1.3 Test sur une proportion

Soient $X_1, \dots X_n$ indépendants et de même loi de Bernouilli $\mathcal{B}(p)$. Le paramètre p est estimé par la proportion observée $\widehat{p}_n$. De même que pour les intervalles de confiance, nous choisissons ici une approche asymptotique. (Remarquons que certains logiciels proposent le test non-asymptotique.) Le théorème de la limite centrale nous assure que

$$U = \frac{\widehat{p}_n - p}{\sqrt{p(1-p)/n}} \xrightarrow[n\to\infty]{loi} \mathcal{N}(0,1).$$

Construisons la région critique RC_α pour le test de (H_0) $p = p_0$ contre (H_1) $p \neq p_0$ de risque α. Nous recherchons la région critique sous la forme

$$RC_\alpha = \left\{ \widehat{p}_n < p_0 - c_\alpha \text{ ou } \widehat{p}_n > p_0 + c_\alpha \right\}.$$

Nous avons $\alpha = \mathbb{P}(RC_\alpha \text{ sous } (H_0))$. Ainsi, il vient :

$$\alpha = \mathbb{P}\left(U < -\frac{c_\alpha}{\sqrt{p_0(1-p_0)/n}} \text{ ou } U > \frac{c_\alpha}{\sqrt{p_0(1-p_0)/n}} \right),$$

avec $U = \frac{\widehat{p}_n - p_0}{\sqrt{p_0(1-p_0)/n}} \xrightarrow[n\to\infty]{loi} \mathcal{N}(0,1)$.

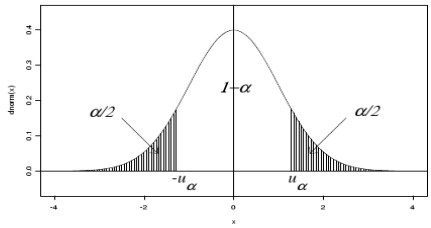

FIGURE 4.3 – Densité de la loi de Gauss de paramètres 0 et 1 et définition du quantile u_α.

Asymptotiquement, nous avons U approximativement de loi gaussienne et nous en déduisons ainsi que $\frac{c_\alpha}{\sqrt{p_0(1-p_0)/n}} = u_\alpha$ avec u_α quantile de la loi normale centrée-réduite défini sur le graphique ci-dessus. Ainsi, la région critique obtenue est :

$$RC_\alpha = \left\{ \widehat{p}_n < p_0 - u_\alpha \sqrt{p_0(1-p_0)/n} \text{ ou } \widehat{p}_n > p_0 + u_\alpha \sqrt{p_0(1-p_0)/n} \right\}.$$

Celle-ci s'écrit aussi

$$RC_\alpha = \{|U| > u_\alpha\} \text{ avec } U = \frac{\widehat{p}_n - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

Cette région est asymptotique dans la mesure où nous avons utilisé le théorème de la limite centrale pour l'obtenir.

BILAN

Test sur une proportion :

Soient $X_1, \ldots, X_n$ indépendantes et de même loi $\mathcal{B}(p)$, alors l'estimateur de p est donné par $\widehat{p}_n = \overline{X}_n$.
Si $p = p_0$, on a

$$U = \frac{\widehat{p}_n - p_0}{\sqrt{p_0(1-p_0)/n}} \xrightarrow{loi} \mathcal{N}(0,1).$$

(H_0)	(H_1)	**Région critique** RC_α
$p = p_0$	$p \neq p_0$	$\{\|U\| > u_\alpha\}$
$p \leqslant p_0$	$p > p_0$	$\{U > u_{2\alpha}\}$
$p \geqslant p_0$	$p < p_0$	$\{U < -u_{2\alpha}\}$

Ces régions sont asymptotiques.

Remarque : Contrairement à la loi normale, il n'y a pas ici équivalence entre les intervalles de confiance et les tests ! Dans les deux cas, nous avons utilisé que

$$\frac{\widehat{p}_n - p}{\sqrt{p(1-p)/n}} \xrightarrow{loi} \mathcal{N}(0,1).$$

Lors de la construction d'un test, le fait de comparer p avec une valeur de référence p_0 a permis d'appliquer ce résultat avec $p = p_0$. Dans le cas des intervalles, nous avons dû réaliser des approximations supplémentaires liées au fait que la variance $p(1-p)$ était inconnue. Dans un test, l'approximation est donc moindre, et la source d'erreur plus faible.

Exemple : Sur $n = 500$ personnes interrogées lors d'un sondage, nous avions obtenu un score pour le candidat A de $\widehat{p}_n = 274/500 = 54.8\%$. Nous aimerions savoir si A va

gagner les élections. Soit p le score de A sur la population. Nous souhaitons donc tester (H_0) $p \leqslant 50\%$ contre (H_1) $p > 50\%$. La région critique de ce test est

$$RC_\alpha = \{U > u_{2\alpha}\} \text{ avec } U = \frac{\widehat{p}_n - 0.5}{\sqrt{0.5(1-0.5)/n}}.$$

Plutôt que de réaliser le test pour différentes valeurs de risque, nous cherchons ici directement la p-valeur. Celle-ci est atteinte lorsque

$$u_{2\alpha} = \frac{\widehat{p}_n - 0.5}{\sqrt{0.5(1-0.5)/n}} = 2.15.$$

Par définition de $u_{2\alpha}$, si Φ désigne la fonction de répartition de la loi $\mathcal{N}(0,1)$, nous avons alors $\alpha = 1 - \Phi(2.15) = 1 - 0.9842 = 1.6\%$. La p-valeur de ce test est donc 1.6%. Nous constatons en effet que la valeur obtenue est différente de celle que nous avions calculé pour les intervalles de confiance. Celle donnée ici est plus précise. Cela signifie donc que si nous acceptons de nous tromper avec une probabilité supérieure à 1.6%, nous pouvons affirmer que A va gagner les élections. Notons que cette conclusion peut se faire sous réserve que l'échantillon soit représentatif de la poulation, qu'aucun individu ne change d'avis ni ne soit indécis, etc.

4.2 Tests de comparaison d'échantillons

Nous souhaitons maintenant comparer un paramètre non pas à une valeur donnée mais à une valeur inconnue que nous avons estimée. Par exemple, nous souhaitons comparer l'effet de deux traitements médicaux sur des patients, comparer le nombre de pièces défectueuses produites par deux machines, comparer des résultats de code après modification d'un paramètre, etc. La difficulté supplémentaire par rapport à la section précédente est que l'imprécision due à l'estimation porte sur les deux grandeurs comparées, et non plus sur une seule. Nous n'allons pas dans ce cours présenter de tests de comparaison dans un contexte général, mais nous nous focaliserons sur les cas d'échantillons suivants des lois gaussiennes et de Bernouilli.

4.2.1 Comparaison d'échantillons gaussiens indépendants

Soient $X_1, \dots X_{n_X}$ indépendants et identiquement distribués selon une loi $\mathcal{N}(m_X, \sigma_X^2)$ et $Y_1, \dots Y_{n_Y}$ indépendants et identiquement distribués selon une loi $\mathcal{N}(m_Y, \sigma_Y^2)$. Nous sup-

posons que les (X_i), (Y_j) sont mutuellement indépendants. Nous souhaiterions comparer m_X et m_Y.

Lorsque les échantillons ne sont pas de grande taille, nous sommes à-même de réaliser un test lorsque les variances σ_X^2 et σ_Y^2 sont égales, mais nous ne pouvons pas effectuer ce test s'il y a une trop forte disparité des variances. Nous allons donc construire une procédure en deux étapes, afin de vérifier que le cadre d'étude convient :

1. Test d'égalité des variances.
   Nous allons tester $(H_0)\ \sigma_X^2 = \sigma_Y^2$ contre $(H_1)\ \sigma_X^2 \neq \sigma_Y^2$. Idéalement, nous voudrions inverser ces deux hypothèses car le risque qui nous intéresse ici est le risque de deuxième espèce alors que les tests permettent de contrôler le risque de première espèce. Le problème est que nous ne pouvons pas échanger les hypothèses : aucune procédure de test n'est alors accessible.
2. Test sur les moyennes.
   Si les échantillons sont de taille supérieure à 100, on peut passer directement à cette étape sans faire de test sur les variances. Sinon, il faut au préalable vérifier qu'il est cohérent de supposer les variances égales. Si, au vu de l'étape précédente, cette hypothèse ne peut être formulée, alors vous répondez que vous ne pouvez pas conclure.

a. Comparaison des variances

Nous voulons tester $(H_0)\ \sigma_X^2 = \sigma_Y^2$ contre $(H_1)\ \sigma_X^2 \neq \sigma_Y^2$. Nous disposons des estimateurs respectifs de σ_X^2 et σ_Y^2 : ${S'_X}^2$ et ${S'_Y}^2$. Nous avons vu que les constructions de test précédentes se faisaient à l'aide d'une fonction pivotale, c'est-à-dire d'une variable aléatoire basée sur l'estimateur concerné, dont la loi était connue de manière exhaustive. Peut-on construire une telle fonction pivotale ? C'est-à-dire trouver une variable aléatoire faisant intervenir ${S'_X}^2$ et ${S'_Y}^2$ dont nous connaissons la loi ?

Rappelons que d'après le théorème de Fisher, $\frac{(n_X-1){S'_X}^2}{\sigma_X^2}$ suit une loi $\chi^2(n_X-1)$ et $\frac{(n_Y-1){S'_Y}^2}{\sigma_Y^2}$ suit une loi $\chi^2(n_Y-1)$. Au vu du rappel sur la loi de Fisher (voir la section 1.2.3 sur la loi gaussienne et ses dérivées), nous pouvons en déduire que $F = \frac{{S'_X}^2}{{S'_Y}^2}\frac{\sigma_Y^2}{\sigma_X^2}$ suit une loi de Fisher de paramètres n_X-1 et n_Y-1, $\mathcal{F}(n_X-1; n_Y-1)$. En particulier, lorsque $\sigma_X^2 = \sigma_Y^2$, nous avons $F = \frac{{S'_X}^2}{{S'_Y}^2} \sim \mathcal{F}(n_X-1; n_Y-1)$.

Construisons maintenant la région critique de notre test. La région critique RC_α est la

zone où l'on considère que ${S'_X}^2$ et ${S'_Y}^2$ sont significativement différentes. La fonction pivotale que nous avons construite faisant intervenir le rapport $\frac{{S'_X}^2}{{S'_Y}^2}$, nous allons chercher la région critique sous la forme $RC_\alpha = \{\frac{{S'_X}^2}{{S'_Y}^2} < c_{1,\alpha}$ ou $\frac{{S'_X}^2}{{S'_Y}^2} > c_{2,\alpha}\}$ avec $c_{1,\alpha} < 1$ et $c_{2,\alpha} > 1$.

Le risque de première espèce vaut

$$\alpha = \mathbb{P}_{\text{sous } (H_0)}(RC_\alpha) = \mathbb{P}_{\text{sous } \sigma_X^2=\sigma_Y^2}\left(\frac{{S'_X}^2}{{S'_Y}^2} < c_{1,\alpha} \text{ ou } \frac{{S'_X}^2}{{S'_Y}^2} > c_{2,\alpha}\right).$$

Ainsi $\alpha = \mathbb{P}\left(F < c_{1,\alpha} \text{ ou } F > c_{2,\alpha}\right)$ avec $F \sim \mathcal{F}(n_X - 1; n_Y - 1)$. Alors, nous posons

$$\begin{aligned} c_{1,\alpha} &= f_{n_X-1;n_Y-1;1-\alpha/2} \\ c_{2,\alpha} &= f_{n_X-1;n_Y-1;\alpha/2} \end{aligned}$$

avec $f_{n_1;n_2;\beta}$ quantile d'ordre $1-\beta$ de la loi $\mathcal{F}(n_1; n_2)$, *i.e.* tel que si $F \sim \mathcal{F}(n_1; n_2)$, alors $\mathbb{P}(F > f_{n_1;n_2;\beta}) = \beta$. Ces valeurs vous sont données soit par un logiciel soit par vos tables statistiques. Ainsi

$$RC_\alpha = \left\{\frac{{S'_X}^2}{{S'_Y}^2} < f_{n_X-1;n_Y-1;1-\alpha/2} \text{ ou } \frac{{S'_X}^2}{{S'_Y}^2} > f_{n_X-1;n_Y-1;\alpha/2}\right\}.$$

BILAN

Test de comparaison des variances de deux lois normales.

Soient $X_1, \ldots, X_{n_X}$ et $Y_1, \ldots, Y_{n_Y}$ deux échantillons indépendants de lois respectives $\mathcal{N}(m_X, \sigma_X^2)$ et $\mathcal{N}(m_Y, \sigma_Y^2)$.
Si $\sigma_X^2 = \sigma_Y^2$, alors

$$F = \frac{{S'_X}^2}{{S'_Y}^2} \sim \mathcal{F}(n_X - 1, n_Y - 1).$$

(H_0)	(H_1)	**Région critique** RC_α
$\sigma_X^2 = \sigma_Y^2$	$\sigma_X^2 \neq \sigma_Y^2$	$\{F < f_{n_X-1;n_Y-1;1-\alpha/2}$ ou $F > f_{n_X-1;n_Y-1;\alpha/2}\}$

On peut construire les autres tests en utilisant la variable aléatoire F ci-dessus.

Si vous regardez vos tables, vous n'arriverez pas à trouver les valeurs que vous cherchez. En effet, en raison de la relation $f_{n_1;n_2;\beta} = 1/f_{n_2;n_1;1-\beta}$ vous pouvez retrouver les autres valeurs. (Cette relation découle de la définition de la loi de Fisher comme quotient : inversez le quotient pour la retrouver.) En pratique, afin de ne regarder qu'une fois sur une table, procédez comme suit :

— Si ${s'_X}^2 > {s'_Y}^2$ alors, comme $f_{n_X-1;n_Y-1;1-\alpha/2} < 1$ quelque soit α, il suffit de comparer $\frac{{s'_X}^2}{{s'_Y}^2}$ à $f_{n_X-1;n_Y-1;\alpha/2}$. Ainsi $RC_\alpha = \left\{ \frac{{s'_X}^2}{{s'_Y}^2} > f_{n_X-1;n_Y-1;\alpha/2} \right\}$.

— Si ${s'_X}^2 < {s'_Y}^2$, de même $RC_\alpha = \left\{ \frac{{s'_Y}^2}{{s'_X}^2} > f_{n_Y-1;n_X-1;\alpha/2} \right\}$.

b. Comparaison des espérances

Nous souhaitons maitenant construire un test de comparaison des espérances m_X et m_Y. Ces deux paramètres sont estimés respectivement par $\overline{X_{n_X}}$ et par $\overline{Y_{n_Y}}$. Nous avons besoin pour déterminer la région critique de trouver une variable aléatoire faisant intervenir $\overline{X_{n_X}}$ et $\overline{Y_{n_Y}}$ dont nous connaisons la loi. Nous savons que $\overline{X_{n_X}}$ suit une loi gaussienne, comme somme de gaussiennes indépendantes, et $\overline{X_{n_X}} \sim \mathcal{N}(m_X, \frac{\sigma_X^2}{n_X})$. De même $\overline{Y_{n_Y}} \sim \mathcal{N}(m_Y, \frac{\sigma_Y^2}{n_Y})$. Les variables aléatoires $(X_i)_{i=1,\dots,n_X}$ et $(Y_j)_{j=1,\dots,n_Y}$ étant indépendantes, $\overline{X_{n_X}}$ et $\overline{Y_{n_Y}}$ sont indépendantes. Nous en déduisons que $\overline{X_{n_X}} - \overline{Y_{n_Y}}$ est de loi $\mathcal{N}\left(m_X - m_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)$. Ainsi

$$U = \frac{\overline{X_{n_X}} - \overline{Y_{n_Y}} - (m_X - m_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim \mathcal{N}(0,1).$$

Lorsque les variances σ_X^2 et σ_Y^2 sont connues, nous pouvons exploiter ce résultat, mais nous devons introduire une estimation des variances dans un cadre général. Soit $Z = \frac{(n_X-1){S'_X}^2}{\sigma_X^2} + \frac{(n_Y-1){S'_Y}^2}{\sigma_Y^2}$. La statistique Z est un estimateur de $\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$. Les variables aléatoires $\frac{(n_X-1){S'_X}^2}{\sigma_X^2}$ et $\frac{(n_Y-1){S'_Y}^2}{\sigma_Y^2}$ sont indépendantes et de loi respectives $\chi^2_{n_X-1}$ et $\chi^2_{n_Y-1}$, d'après le Théorème de Fisher (section 1.2.3). Nous en déduisons que Z suit une loi $\chi^2_{n_X+n_Y-2}$.

Par définition de la loi de Student, la variable aléatoire $T = \sqrt{n_X + n_Y - 2}\frac{U}{\sqrt{Z}}$ suit une loi de Student de paramètre $n_X + n_Y - 2$. Nous avons

$$T = \sqrt{n_X + n_Y - 2}\frac{\overline{X_{n_X}} - \overline{Y_{n_Y}} - (m_X - m_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}\sqrt{\frac{(n_X-1){S'_X}^2}{\sigma_X^2} + \frac{(n_Y-1){S'_Y}^2}{\sigma_Y^2}}}.$$

Lorsque nous avons $\sigma_X^2 = \sigma_Y^2$, nous pouvons constater que T s'éscrit :

$$T = \sqrt{n_X + n_Y - 2}\frac{\overline{X_{n_X}} - \overline{Y_{n_Y}} - (m_X - m_Y)}{\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}\sqrt{(n_X-1){S'_X}^2 + (n_Y-1){S'_Y}^2}}.$$

Les variances inconnues n'apparaissent donc plus dans l'expression de T. Ceci justifie la procédure en deux étapes suggérée auparavant : nous commençons par tester l'hypothèse $\sigma_X^2 = \sigma_Y^2$ est si celle-ci est validée alors nous pouvons utiliser la variable T.

La variable T suivant une loi de Student, la construction de la région critique est parfaitement similaire à ce qui a été vu précédemmment pour le test sur une espérance de la loi normale. Il suffit de traduire les hypothèse de la forme $m_X < m_Y$ par des hypothèses sur la différence entre les espérances $m_X - m_Y < 0$. Les régions critiques obtenues sont les suivantes :

BILAN

Test de comparaison des espérances de deux lois normales.

Soient $X_1, \dots, X_{n_X}$ et $Y_1, \dots, Y_{n_Y}$ deux échantillons indépendants de lois respectives $\mathcal{N}(m_X, \sigma_X^2)$ et $\mathcal{N}(m_Y, \sigma_Y^2)$.
Si $\sigma_X^2 = \sigma_Y^2$ et si $m_X = m_Y$, alors

$$T = \frac{\overline{X_{n_X}} - \overline{Y_{n_Y}}}{\gamma\sqrt{\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \sim \mathcal{S}t(n_X + n_Y - 2) \text{ avec } \gamma^2 = \frac{(n_X - 1){S'_X}^2 + (n_Y - 1){S'_Y}^2}{n_X + n_Y - 2}.$$

(H_0)	(H_1)	**Région critique** RC_α
$m_X = m_Y$	$m_X \neq m_Y$	$\{\lvert T\rvert > t_{n_X+n_Y-2;\alpha}\}$
$m_X \leqslant m_Y$	$m_X > m_Y$	$\{\ T > t_{n_X+n_Y-2;2\alpha}\}$
$m_X \geqslant m_Y$	$m_X < m_Y$	$\{\ T < -t_{n_X+n_Y-2;2\alpha}\}$

Remarque sur le test d'égalité des variances. Nous voulons dans le test d'égalité des variances vérifier que $\sigma_X^2 = \sigma_Y^2$. Nous voulons donc contrôler le risque d'accepter à tort que ces deux valeurs sont égales. Les hypothèses que nous devrions poser dans le test sont donc (H_0) $\sigma_X^2 \neq \sigma_Y^2$ et (H_1) $\sigma_X^2 = \sigma_Y^2$. Cependant nous ne pouvons pas réaliser un tel test. En effet, le risque de première espèce de ce test est $\sup_{\text{sous } (H_0)} \mathbb{P}(\text{refuser } (H_0))$; la borne supérieure sur l'ensemble des variances vérifiant (H_0) est une borne supérieure sur l'ensemble des variances, étant donné qu'un nombre *négligeable* de variances ne vérifie pas l'hypothèse. Nous réalisons donc le test de (H_0) $\sigma_X^2 = \sigma_Y^2$ et (H_1) $\sigma_X^2 \neq \sigma_Y^2$, dans la mesure où nous savons réaliser ce test et où il s'approche de celui que nous désirerions faire. Remarquons simplement que dans la majorité des tests que nous réalisons, nous

cherchons une p-valeur faible permettant de rejeter avec un faible risque l'hypothèse (H_0). Dans le cas présent, à l'inverse, nous cherchons une p-valeur élevée pour ne pas rejeter (H_0).

Remarque. Ce test a été développé pour la comparaison d'espérances de lois normales, mais il est applicable pour des échantillons qui ne sont pas issus de lois normales. Lorsque les échantillons sont grands (plus de 100 observations), on peut montrer que la variable $T = \sqrt{n_X + n_Y - 2}\frac{\overline{X_{n_X}} - \overline{Y_{n_Y}} - (m_X - m_Y)}{\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}\sqrt{(n_X - 1){S'_X}^2 + (n_Y - 1){S'_Y}^2}}$ suit toujours une loi de Student à l'aide du théorème de la limite centrale. Lorsque les échantillons sont de petites tailles, des tests spécifiques doivent être appliqués.

c. Exemple

Nous voudrions comparer le temps de sommeil des étudiants de 4IF en période de projet avec celui d'étudiants de BioSciences. Onze étudiants de ce département ont accepté de répondre. Notons $(X_i)_{i=1,\dots,n_X}$ les temps de sommeil en IF et $(Y_j)_{j=1,\dots,n_Y}$ les temps de sommeil en BioSciences. Nous avons :

$$\begin{array}{llll} n_X = 30 & \bar{x} = 6.36h & {s'_X}^2 = 1.85h^2 \\ n_Y = 11 & \bar{y} = 6.68h & {s'_Y}^2 = 1.35h^2 \end{array}$$

Les données de BioSciences sont malheureusement fictives.

Nous souhaitons déterminer si les étudiants de IF dorment significativement moins en moyenne que les étudiants de BioSciences.

Nous supposons les observations indépendantes et nous supposons que les variables $X_1, \dots X_{n_X}$ sont identiquement distribuées de loi $\mathcal{N}(m_X, \sigma_X^2)$ et les variables $Y_1, \dots Y_{n_Y}$ sont identiquement distribuées de loi $\mathcal{N}(m_Y, \sigma_Y^2)$. Nous souhaitons réaliser le test de (H_0) $m_X \geqslant m_Y$ contre (H_1) $m_X < m_y$. Nous procédons en deux étapes :

1. Test d'égalité des variances.
   Nous réalisons ce test avec la plus grande valeur de α possible. Vos tables permettent de prendre $\alpha = 10\%$. Comme ${s'_Y}^2 > {s'_X}^2$, nous comparons $\frac{{s'_Y}^2}{{s'_X}^2} = 1.37$ à $f_{n_Y-1;n_X-1;\alpha/2} = f_{10;29;5\%} \in [2,16; 2,24]$. Comme $\frac{{s'_Y}^2}{{s'_X}^2} < f_{n_Y-1;n_X-1;\alpha/2}$, nous ne pouvons pas rejeter (H_0) au seuil $\alpha = 10\%$.
2. Test de comparaison des espérances.
   Au vu de ce qui précède, nous pouvons supposer que $\sigma_X^2 = \sigma_Y^2$. La région critique pour le test de (H_0) $m_X \geqslant m_Y$ contre (H_1) $m_X < m_y$ est donc donnée par

$$RC_\alpha = \{\, T < -t_{n_X+n_Y-2;2\alpha} \}$$

$$\text{avec } T = \frac{\overline{X_{n_X}} - \overline{Y_{n_Y}}}{\gamma\sqrt{\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \text{ où } \gamma^2 = \frac{(n_X-1){S'_X}^2 + (n_Y-1){S'_Y}^2}{n_X+n_Y-2}.$$

   La réalisation de T vaut ici $t = -0.392$. La lecture de table nous donne $t_{n_X+n_Y-2;\alpha} = t_{39;5\%} \in [2.021; 2.042]$. Nous ne sommes donc pas dans la région critique.
   Nous ne pouvons donc pas conclure que les étudiants de IF dorment significativement moins en moyenne que les étudiants de BioSciences.

4.2.2 Comparaison d'échantillons gaussiens appariés

Soient $X_1, \dots X_n$ indépendants et identiquement distribués selon une loi $\mathcal{N}(m_X, \sigma_X^2)$ et $Y_1, \dots Y_n$ indépendants et identiquement distribués selon une loi $\mathcal{N}(m_Y, \sigma_Y^2)$. Nous supposons que pour tout $i = 1, \dots, n$, le couple (X_i, Y_i) est mesuré sur le même individu. Les échantillons (X_i), et (Y_j) sont alors dits **appariés** et ne peuvent être considérés comme indépendants. Nous souhaiterions comparer m_X et m_Y.

Introduisons $Z = X - Y$. Alors les variables aléatoires $Z_1, \dots, Z_n$ sont indépendantes. Nous supposons que Z suit une loi normale $\mathcal{N}(m, \sigma^2)$. (Ceci est une hypothèse car si la différence de deux lois normales indépendantes est bien une gaussienne, ceci n'est pas assuré dans un contexte de dépendance.)

Faire un test de comparaison de m_X et de m_Y est équivalent dans ce cadre à comparer la valeur m à 0. En effet, nous avons $m = m_X - m_Y$. Nous sommes donc ramenés à un test sur l'espérance d'une loi normale.

Exemple : Nous voudrions comparer le temps de sommeil des étudiants de 4IF en période de projet et en temps normal. Pour cela nous avons demandé aux 30 étudiants ayant participé à l'enquête de donner leur temps moyen de sommeil en dehors des période de projet. Nous ne donnons ici que les temps donnés par les 10 premiers étudiants pour illustrer la méthodologie :

Projet	4.5	7	6	6	6	7	6	6	7	5
Normal	6	8	7	7	6	7	8	7	7	7
Projet-Normal	-1.5	-1	-1	-1	0	0	-2	-1	0	-2

Notons Z la différence de temps de sommeil entre une période de projet et en temps normal. Z_i correspond à cette différence pour l'étudiant i. Sur les 30 étudiants, nous obtenons une estimation de l'espérance de Z qui vaut $\bar{z_n} = -1.06\,h$ et une estimation de la variance ${s'_n}^2 = 1.07\,h^2$. (Les résultats furent de même obtenus sur la promotion 2009-2010, avec cependant une modification du nombre de réponses données.)

Supposons que Z suive une loi $\mathcal{N}(m, \sigma^2)$. Nous souhaitons tester si le sommeil en période de projet est significativement plus faible qu'en temps normal. Pour cela, nous allons réaliser le test de (H_0) $m \geqslant 0$ contre (H_1) $m < 0$. La région critique de ce test est $RC_\alpha = \{T_n < -t_{n-1;2\alpha}\}$, avec $T_n = \frac{\overline{X_n}}{S'_n/\sqrt{n}}$. La réalisation de T_n est ici $t_n = -5.63$. Comme $t_{n-1;0,1\%} = 3.659$, nous pouvons rejeter (H_0) au seuil de 0.05%. Ainsi nous pouvons conclure que les étudiants de 4IF dorment significativement moins en période de projet qu'en temps normal ! Sous réserve bien entendu que leurs réponses soient justes et le modèle de loi gaussienne adéquat.

4.2.3 Comparaison de 2 proportions

Soient $X_1, \dots X_{n_X}$ indépendants et identiquement distribués selon une loi $\mathcal{B}(p_X)$ et $Y_1, \dots Y_{n_Y}$ indépendants et identiquement distribués selon une loi $\mathcal{B}(p_Y)$. Nous supposons que les $(X_i), (Y_j)$ sont mutuellement indépendants. Nous souhaiterions comparer p_X et p_Y : peut-on construire des test de $p_X = p_Y$ contre $p_X \neq p_Y$ ou de $p_X \leqslant p_Y$ contre $p_X > p_Y$?

Notons respectivement $\widehat{p}_X$ et $\widehat{p}_Y$ les estimateurs de p_X et de p_Y dans les échantillons. En raison de la difficulté de manipulation des lois binomiales, nous utilisons le théorème de la limite centrale :

$$\widehat{p}_X \xrightarrow[n_X \to \infty]{loi} \mathcal{N}\left(p_X;\, p_X(1-p_X)/n_X\right) \text{ et } \widehat{p}_Y \xrightarrow[n_Y \to \infty]{loi} \mathcal{N}\left(p_Y;\, p_Y(1-p_Y)/n_Y\right).$$

Par indépendance, lorsque n_X et n_Y sont grands, nous avons approximativement

$$\widehat{p}_X - \widehat{p}_Y \sim \mathcal{N}\left(p_X - p_Y;\ p_X(1-p_X)/n_X + p_Y(1-p_Y)/n_Y\right).$$

Dans les constructions de régions critiques, nous sommes toujours ramenés à considérer le cas $p_X = p_Y$ où sera atteinte la borne supérieure du risque. Lorsque $p_X = p_Y = p$, nous avons $\frac{\widehat{p}_X - \widehat{p}_Y}{\sqrt{p(1-p)(1/n_X+1/n_Y)}} \sim \mathcal{N}(0;1)$. Le paramètre p peut être estimé à partir des deux échantillons par $\overline{p} = \frac{n_X\widehat{p}_X + n_Y\widehat{p}_Y}{n_X+n_Y}$. Nous pouvons montrer (ceci est admis ici) qu'asymptotiquement, pour n_X et n_Y tendant vers l'infini, remplacer p par son estimation dans la variable aléatoire ci-dessus ne change pas sa distribution. Ainsi, on admet que

$$U = \frac{\widehat{p}_X - \widehat{p}_Y}{\sqrt{\overline{p}(1-\overline{p})(1/n_X+1/n_Y)}} \xrightarrow[n_X,n_Y\to\infty]{loi} \mathcal{N}(0;1) \text{ lorsque } p_X = p_Y.$$

A partir de ce résultat, les régions critiques peuvent aisèment être établies, en procédant de manière similaire à ce qui a été fait dans le cas du test sur une proportion, en ramenant les hypothèses à une comparaison de $p_X - p_Y$ avec 0. Les régions ainsi obtenues sont données ci-après.

BILAN

Pour deux proportions :
Soient $X_1, \ldots, X_{n_1}$ et $Y_1, \ldots, Y_{n_2}$ deux échantillons indépendants de lois respectives $\mathcal{B}(p_1)$ et $\mathcal{B}(p_2)$. Les estimateurs de p_1 et p_2 sont donnés par $\widehat{p}_1 = \overline{X}_{n_1}$ et $\widehat{p}_2 = \overline{Y}_{n_2}$.
Si $p_1 = p_2$, alors

$$U = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \xrightarrow{loi} \mathcal{N}(0,1) \text{ avec } \overline{p} = \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1+n_2}.$$

(H_0)	(H_1)	**Région critique** W_α
$p_1 = p_2$	$p_1 \neq p_2$	$\{\lvert U\rvert > u_\alpha\}$
$p_1 \leqslant p_2$	$p_1 > p_2$	$\{U > u_{2\alpha}\}$
$p_1 \geqslant p_2$	$p_1 < p_2$	$\{U < -u_{2\alpha}\}$

Ces régions sont asymptotiques.

Exemple. Sur $n = 500$ personnes interrogées lors d'un sondage, nous avions obtenu un score pour le candidat A de $\widehat{p}_n = 274/500 = 54.8\%$. Nous souhaiterions distinguer deux catégories d'individus dans ce sondage (par exemple les moins de 30 ans et les plus de 30 ans, les catégories socio-professionnelles qualifiés et les non qualifiées, les hommes et les femmes, etc). Soit p_1 le score de A sur la catégorie C1 dans la population, et p_2 son score sur la catégorie C2. Nous voudrions savoir si le choix dans l'élection est significativement différent selon les catégories. Nous nous fixons une acceptation de 5% d'erreur dans notre conclusion.

$n_1 = 150$ personnes interrogées dans le sondage appartiennent à la catégorie C1 et les $n_2 = 350$ autres appartiennent à la catégorie C2. Dans la catégorie C1, le score du candidat A dans le sondage vaut $\widehat{p}_1 = 46.67\%$ et dans la catégorie C2, on a observé une proportion $\widehat{p}_2 = 58.29\%$ de personnes votant pour A.

Nous allons donc effectuer le test de (H_0) $p_1 = p_2$ contre (H_1) $p_1 \neq p_2$. La région critique de ce test au seuil α est $RC_\alpha = \{|U| > u_\alpha\}$ avec $U = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ où $\overline{p} = \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1+n_2}$.
La réalisation de U vaut $u = -2.39$. Comme $u_{5\%} = 1.96$, nous concluons que les deux catégories votent significativement de manière différente.

CHAPITRE 5

TESTS DU χ^2

Les tests du χ^2 se placent dans un cadre où les hypothèses ne peuvent être formulées à l'aide de paramètres. Nous allons nous intéresser en fait à deux types de question :

— Le modèle choisi est-il valide ?

— A-t-on indépendance entre deux jeux de donnnées ?

L'intérêt de la première question est évident : si votre modèle est faux, toutes vos conclusions seront potentiellement fausses ! La deuxième question permet de répondre à des interrogations telles que : un médicament a-t-il une influence sur la guérison ? Le poids et la taille d'un individu sont-il liés ? etc.

5.1 Test d'adéquation

Le but d'un test d'adéquation est de vérifier que le modèle choisi est cohérent avec les données, plus exactement nous allons construire une procédure de test afin de vérifier si l'hypothèse que les observations sont issues d'une loi donnée est valide. Nous allons dans un premier temps étudier les tests d'adéquation pour des données discrètes puis pour des données continues.

5.1.1 Variables dicrètes

Soient $X_1, \ldots, X_n$ n répétitions indépendantes d'une variables aléatoire X, à valeurs dans $\{e_1, \ldots, e_k\}$. Notons N_j la variable aléatoire correspondant au nombre de variable X_i ayant pour réalisation e_j.

Nous voulons construire une procédure permettant de tester (H_0) "La variable X admet pour fonction de répartition la fonction F^*" contre (H_1) "La variable X n'admet pas pour

fonction de répartition la fonction F^*".

Remarque : Dans la majorité des cas, nous voudrions conclure que la fonction F^* est bien la fonction de répartition de X. Idéalement, nous aimerions alors échanger les hypothèses (H_0) et (H_1) dans la procédure de test mais nous ne savons pas alors comment contrôler le risque de première espèce. Nous serons donc intéressés par des p-valeurs élevées.

Comment reformuler les hypothèses testées à l'aide des données ? Soit p_j^* la probabilité qu'une variable aléatoire ayant pour fonction de répartition F^* prenne la modalité e_j, pour $j = 1, \ldots, k$. Nous voulons alors tester (H_0) "Pour tout $j = 1, \ldots, k$, nous avons $N_j = np_j^*$" contre (H_1) "Il existe $j \in \{1, \ldots, k\}$ tel que $N_j \neq np_j^*$.".

Exemple. Nous voudrions tester si un dé est truqué ou équilibré. Pour cela, nous lançons le dé 300 fois et notons les résultats obtenus :

e_j	1	2	3	4	5	6
n_j	42	43	56	55	43	61

Nous voulons tester (H_0) "Le dé est équilibré" contre (H_1) "Le dé est truqué". Si le dé est équilibré, la probabilité d'avoir chacune des faces vaut $p_j^* = 1/6$. La théorie nous dit ainsi que sur 300 lancers, il devrait y avoir 50 lancers pour chaque face. D'où le tableau :

e_j	1	2	3	4	5	6
n_j	42	43	56	55	43	61
np_j^*	50	50	50	50	50	50

Nous voulons savoir si les observations (deuxième ligne) sont significativement éloignées de la théorie (troisième ligne).

Revenons sur la construction du test. L'idée est de construire ensuite une notion de distance à la situation (H_0) testée. introduisons

$$\Delta^2 = \sum_{j=1}^{k} \frac{(N_j - np_j^*)^2}{np_j^*}.$$

La variable Δ^2 est appelée distance du χ^2. Elle mesure l'écart entre les effectifs observés $\{N_1, \ldots, N_k\}$ et les effectifs théoriques sous (H_0) : $\{np_1^*, \ldots, np_k^*\}$. Nous sommes alors ramenés à tester si Δ^2 est significativement supérieur à 0, ce qui correspondrait à un

écart significatif entre ce qui a été observé et la théorie. Nous cherchons donc une région critique de la forme $RC_\alpha = \{\Delta^2 > c_\alpha\}$ avec $c_\alpha > 0$.

Afin de construire ce test, nous devons étudier la loi de probabilités de Δ^2 sous (H_0). Lorsque l'hypothèse (H_0) est vérifiée, la probabilité que $X_i = e_j$ vaut p_j^*. Alors la loi du vecteur $(N_1, \dots N_k)$ est donnée par :

$$\text{Pour } n_1, \dots, n_k \text{ entiers tels que } \sum n_j = n,$$

$$\mathbb{P}((N_1, \dots N_k) = (n_1, \dots, n_k)) = C_n^{n_1} p_1^{*n_1} C_{n-n_1}^{n_2} p_2^{*n_2} \dots C_{n-\sum_{j<k} n_j}^{n_k} p_k^{*n_k}.$$

Le vecteur $(N_1, \dots N_k)$ suit une loi dite multinomiale de paramètres $(n, p_1^*, \dots p_k^*)$. Il en résulte que Δ^2 suit asymptotiquement une loi du χ^2 de paramètre $k-1$ (ce résultat est admis).

Nous en déduisons que la région critique asymptotique est

$$RC_\alpha = \{\Delta^2 > z_{k-1,\alpha}\}.$$

Cette région étant asymptotique, elle n'est valable qu'à partir d'un nombre suffisamment élevé d'observations. En pratique, il ne faut pas appliquer ce test si plus de 20% des réalisations n_j des variables N_j sont inférieures à 5 (critère de Cochran).

Exemple. Dans l'exemple précédent :

e_j	1	2	3	4	5	6
n_j	42	43	56	55	43	61
np_j^*	50	50	50	50	50	50

Alors la distance du χ^2 vaut

$$\delta^2 = \frac{(42-50)^2}{50} + \frac{(43-50)^2}{50} + \dots + \frac{(61-50)^2}{50} = 6.48.$$

Nous comparons δ^2 à $z_{5;5\%} = 11.7$. Comme $\delta^2 < z_{5;5\%}$, nous ne sommes pas dans la région critique. Ainsi on ne rejette pas (H_0). On ne peut pas conclure que le dé est truqué avec un risque de 5%.

Généralisation : Et si l'on souhaitait tester l'appartenance de X à une famille de loi de fonction de répartition appartenant à $\mathcal{F} = \{F_\theta, \theta \in \mathbb{R}^p\}$? Dans ce cas, il faut procéder en deux étapes :

1. On estime la valeur de θ par $\widehat{\theta}_n$.
2. On teste (H_0) "X suit la loi $F_{\widehat{\theta}_n}$" contre (H_1) "X ne suit pas la loi $F_{\widehat{\theta}_n}$".

La région critique asymptotique du test est alors $RC_\alpha = \{\Delta^2 > z_{k-1-p;\alpha}\}$ avec Δ^2 calculé de manière similaire à ce qui précède en considérant $p_j^* = F_{\widehat{\theta}_n}(e_j)$. Remarquons que l'on perd autant de degrés de liberté que l'on a de paramètres à estimer.

5.1.2 Variables continues

Nous observons $X_1, \ldots, X_n$ variables indépendantes de même fonction de répartition F. Nous voudrions construire un test de (H_0) $F = F^*$ contre (H_1) $F \neq F^*$. L'idée est de se ramener à un cadre similaire à celui des variables discrètes pour appliquer ce qui précède.

Partitionnons l'ensemble des modalités en k intervalles, de manière similaire à ce qui a été fait lors de la construction des histogrammes. Nous obtenons les classes $c_j =]a_{j-1}; a_j]$, pour $j = 1, \ldots, k$. Notons N_j les effectifs associés : pour tout $j = 1, \ldots, k$, la variable aléatoire N_j correspond au nombre d'observations dans la classe c_j.

Si la fonction de répartition des variables observées est bien F^*, alors la probabilité pour chacune des variables de se trouver dans la classe c_j vaut

$$p_j^* = \mathbb{P}(X_i \in c_j) = F^*(a_j) - F^*(a_{j-1}).$$

L'effectif théorique associé à chacune des classes est ainsi $n_j^* = np_j^*$.

Le raisonnement est alors en tout point similaire à ce qui précède. L'hypothèse (H_0) est rejetée au seuil α si $\delta^2 > z_{k-1,\alpha}$, avec $\delta^2 = \sum_{j=1}^k \frac{(n_j - n_j^*)^2}{n_j^*}$, où n_j est la réalisation de N_j. Ce test est asymptotique.

La généralisation à une famille de loi est similaire à ce qui est décrit dans le cas discret.

5.2 Test d'indépendance

Le but est de déterminer si deux variables aléatoires X et Y sont indépendantes. Par exemple ici nous nous intéresserons à l'influence de la présence en cours sur la note. Nous souhaitons savoir si la note obtenue en examen est significativement corrélée avec

le taux de présence aux cours magistraux. Avant de décrire plus avant le test effectué, nous rappelons les principales notions de vocabulaire liées à l'observation d'un cuople de variables.

5.2.1 Tableaux de contingence

Nous avons demandé à la promotion de 2009-2010 de donner de manière anonyme son taux de présence en cours magistral ainsi que sa note à l'examen final. Remarquons qu'en raison d'un grand nombre d'auto-censure, nous ne disposons malheureusement pas de suffisament de données à notre goût. Les résultats obtenus sont résumés dans le tableau ci-après :

Note | Taux de présence	$< 50\%$	$\geqslant 50\%$	total
0 à 5	0	2	2
5 à 10	5	3	8
10 à 15	3	11	14
15 à 20	0	6	6
total	8	22	30

Un tel tableau est appelé un tableau de distribution conjointe. Les *marges* donnant les totaux sont les distributions marginales.

Donnons les notations utilisées dans un cadre plus général. Considérons deux variables aléatoires X et Y, de modalités respectives $\{e_1, \ldots, e_p\}$ et $\{\tilde{e}_1, \ldots, \tilde{e}_k\}$. Dans le cas de variables continues, ces modalilités représentent des classes. Le tableau de contingence des observations de X et Y sur n individus est le suivant :

X | Y	$\tilde{e}_1$	$\ldots$	$\tilde{e}_j$	$\ldots$	$\tilde{e}_k$	total
e_1	n_{11}		n_{1j}		n_{1k}	$n_{1.}$
$\vdots$			$\vdots$			
e_i	n_{i1}	$\ldots$	n_{ij}	$\ldots$	n_{ik}	$n_{i.}$
$\vdots$			$\vdots$			
e_p	n_{p1}		n_{pj}		n_{pk}	$n_{p.}$
total	$n_{.1}$		$n_{.j}$		$n_{.k}$	n

La grandeur n_{ij} correspond au nombre d'individus tels que X a pour modalité e_i et Y a pour modalité $\tilde{e}_j$. La donnée des $\{n_{ij},\, i = 1, \ldots, p,\, j = 1, \ldots, k\}$ correspond à la distribution conjointe de (X, Y) en effectif.

La grandeur $n_{i.} = \sum_j n_{ij}$ correspond au nombre d'individus tels que X a pour modalité e_i. La donnée des $\{n_{i.},\, i = 1, \ldots, p\}$ correspond à la distribution marginale de X en effectif.

La grandeur $n_{.j} = \sum_i n_{ij}$ correspond au nombre d'individus tels que Y a pour modalité $\tilde{e}_j$. La donnée des $\{n_{.j},\, j = 1, \ldots, k\}$ correspond à la distribution marginale de Y en effectif.

Comment à partir de tableaux similaires obtenir une première vision de la dépendance entre les variables ? Nous pouvons étudier les distributions dites conditionnelles afin de mieux visualiser les liens éventuels.

La distribution conditionnelle en fréquence de X selon Y est donnée par le tableau suivant :

X | Y	$\tilde{e}_1$	…	$\tilde{e}_j$	…	$\tilde{e}_k$
e_1	$f_{1\|1}$		$f_{1\|j}$		$f_{1\|k}$
⋮	⋮		⋮		⋮
e_i	$f_{i\|1}$		$f_{i\|j}$		$f_{i\|k}$
⋮	⋮		⋮		⋮
e_p	$f_{p\|1}$		$f_{p\|j}$		$f_{p\|k}$
total	1		1		1

avec $f_{i|j} = \frac{n_{ij}}{n.j}$. La grandeur $f_{i|j}$ est une estimation de la probabilité $\mathbb{P}(X = e_i \,|\, Y = \tilde{e}_j)$. Ainsipour j donné $\{f_{i|j},\, i = 1, \ldots, p\}$ est la distribution empirique de X sachant $Y = \tilde{e}_j$ (voir votre cours de probabilité pour des rappels sur les distributions conditionnelles).

Lorsqu'il y a indépendance entre les variables X et Y, le conditionnement de X par une condition sur Y ne doit pas modifier la distribution. Ainsi dans le tableau ci-dessus, les distributions, données en colonnes, doivent être similaires.

Prenons l'exemple de la note et du taux de présence en cours. Le tableau de distribution conjointe en fréquence de la note selon la présence en cours est :

Note | Taux de présence	$< 50\%$	$\geqslant 50\%$
0 à 5	0	0,09
5 à 10	0,63	0,14
10 à 15	0,38	0,50
15 à 20	0	0,27
total	1	1

Au vu de ce tableau, les distributions conditionnelles étant très différentes, nous avons envie de conclure que les variables sont liées, c'est-à-dire qu'il y a une corrélation en les notes obtenues et les taux de présence aux cours magistraux. Nous allons construire une procédure de test afin de le vérifier.

5.2.2 Construction du test d'indépendance

Tableau des effectifs théoriques sous l'hypothèse d'indépendance

Dans le cas de l'exemple de l'étude du lien entre la note à l'examen et le taux de pésence en cours magistral, si les variables étaient indépendantes, nous aurions les effectifs théoriques suivants :

Note | Taux de présence	$< 50\%$	$\geqslant 50\%$	total
0 à 5	0.53	1.47	2
5 à 10	2.13	5.87	8
10 à 15	3.73	10.27	14
15 à 20	1.6	4.4	6
total	8	22	30

Le but est alors de tester si la différence entre le tableau de distribution conjointe observé et le tableau de distribution conjointe théorique obtenu sous l'hypothèse d'indépendance est significative. de manière similaire au test du χ^2 sur l'adéquation, nous introduisons une distance du Chi-deux, mesurant l'écart entre les deux tableaux. La distance du χ^2 s'écrit :

$$\delta^2 = \sum_{i,j} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}.$$

De manière similaire à ce qui a été développé pour le test d'adéquation, nous pouvons

montrer que si $\delta^2 > z_{(k-1)\times(p-1);\alpha}$ alors nous pouvons rejeter l'hypothèse d'indépendance avec un risque α. La quantité $z_{(k-1)\times(p-1);\alpha}$ est ici le quantile d'ordre α de la loi du χ^2 de paramètre $(k-1)\times(p-1)$.

Remarque : Pour retenir le nombre de degré de liberté de la loi du χ^2, notez que celui-ci est égal à (nombre de lignes-1)×(nombre de colonnes-1).

De manière similaire au test d'adéquation, ce test est asymptotique. Ainsi, afin de l'appliquer vous devez vous assurer que moins de 20% des effectifs théoriques sont inférieurs à 5.

Dans le cas de l'exemple considéré ici, nous ne pouvons pas appliquer le test car la condition ci-dessus n'est pas vérifiée. Dans un but pédagogique, nous extrapolons les données recueillies ici pour toute la promotion. Cette démarche ne peut bien sûr être validée et a uniquement pour but de vous montrer comment s'applique ce test.

Supposons que sur 120 etudiants, nous ayons observé la répartition suivante :

Note | Taux de présence	$< 50\%$	$\geqslant 50\%$	total
0 à 5	0	8	8
5 à 10	20	12	32
10 à 15	12	44	56
15 à 20	0	24	24
total	32	88	120

Le tableau des effectifs théoriques sous l'hypothèse d'indépendance est :

Note | Taux de présence	$< 50\%$	$\geqslant 50\%$	total
0 à 5	2.13	5.87	8
5 à 10	8.53	23.47	32
10 à 15	14.93	41.07	56
15 à 20	6.40	17.60	24
total	32	88	120

La distance du χ^2 vaut

$$\delta^2 = \frac{(0-2.13)^2}{2.13} + \frac{(8-5.87)^2}{5.87} + \ldots + \frac{(24-17.6)^2}{17.6} = 33.43.$$

Nous comparons cette grandeur à $z_{3\times 1;5\%} = 7.81$. Nous avons $\delta^2 > z_{3;5\%}$. Par conséquent nous rejetons l'hypothèse d'indépendance. Au seuil de 5% la présence en cours magistral et la note à l'examen sont liées. Remarquons de plus qu'au vu des données, plus le taux de présence est élevé, plus la note est élevée ... mais rappelons que nous avons étendu les résultats observés sur 30 étudiants donc que ces conclusions ne sont pas valides. Cependant je me permets de vous inciter à suivre les cours d'amphi car même si cela est probablement moins significatif, la présence en cours semble influencer les notes...

CHAPITRE 6

INTRODUCTION À LA REGRESSION LINÉAIRE

Dans ce qui précède, nous nous sommes majoritairement intéressés à des modèles sur une seule variable aléatoire. Même lors de la comparaison d'échantillons, nous avons modélisé séparément les variables dans les deux échantillons, ou bien nous nous sommes ramenés à une unique variable. Le but de ce chapître est d'introduire les modèles faisant intervenir simultanément plusieurs variables, et plus précisément les modèles cherchant à établir un lien entre deux variables.

Par exemple, vous souhaitez établir un lien entre le poids et la taille des individus, entre la distance de freinage d'une voiture et la vitesse à laquelle elle roule, entre le prix des biens immobiliers et leur surface, etc. Pour formaliser de tels liens, nous allons introduire un modèle linéaire simple. Ce n'est qu'un aperçu du principe de ces modèles.

Remarque : En informatique, on rencontre parfois la notion de *surface de réponse* : l'idée est de comprendre comment les sorties obtenues par votre code dépendent des paramètres d'entrée à l'aide d'un modèle de régression.

6.1 Le modèle de regression linéaire simple

Définissons tout d'abord le modèle de régression de Y sur x :

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

où

- Y est la variable expliquée,
- x est la variable explicative,
- ε représente l'erreur de modélisation, ou résidu du modèle.

Pour un échantillon $(x_i, y_i)_{i=1\ldots,n}$ d'observations, le modèle s'écrit :

$$\forall i = 1, \ldots, n, \quad Y_i = f(x_i) + \varepsilon_i.$$

Nous considérerons dans le cadre de ce cours que les variables x_i sont déterministes, c'est-à-dire que ce sont des constantes et qu'elles ne seront pas vues comme des réalisations de variables aléatoires. Les résidus ε_i sont des variables aléatoires centrées (pour que le modèle ait une signification). Les variables Y_i sont alors également des variables aléatoires.

Il existe des méthodes, dites de régression non paramétriques, permettant d'estimer la fonction f sans lui supposer une forme donnée. Cependant, de telles méthodes dépassent le cadre de ce cours. Nous nous restreindrons ici à un cas simple, où la fonction f est affine.

6.1.1 Définition

Le modèle de régression linéaire simple s'écrit :

$$\forall i = 1, \ldots, n, \quad Y_i = a\, x_i + b + \varepsilon_i,$$

avec a et b paramètres inconnus et des résidus ε_i indépendants, de même loi et tels que $\mathbb{E}\varepsilon_i = 0$ et $Var(\varepsilon_i) = \sigma^2$.

Les hypothèses de ce modèle, notamment celles sur les résidus, peuvent bien sûr être relachées, mais nous préférons étudier le cas simple.

Remarquons que l'hypothèse affine n'est pas aussi restrictive qu'il n'y paraît, dans la mesure où il est souvent possible de s'y ramener à l'aide de changement de variables. Par exemple :

— $Y_i = a\, \ln(x_i) + b + \varepsilon_i$,
— $Y_i^2 = a\, e^{x_i} + b + \varepsilon_i$,
— $\ln(Y_i/(1 - Y_i)) = a\, x_i + b + \varepsilon_i$, (modèle logistique)

sont aussi des modèles linéaires.

Supposons par exemple que vous disposiez d'un appartement à Lyon que vous souhaitez vendre. Vous faîtes réaliser une estimation de votre bien mais vous ne trouvez pas l'évaluation obtenue cohérente avec les prix du marché. Vous souhaitez donc établir une

estimation du prix au vu de ce que vous avez observé. Sur un site d'annonces immobilières vous avez relevé les prix et les surfaces de 12 biens immobiliers sur Lyon. Les données datent du 16/11/2010.

Surface (en m^2)	146	70	22	105	76	95	120	32	141	98	114	46
Prix (en milliers d'euros)	465	225	105	321	229	305	369	109	424	314	369	155

Le prix dépend-il de manière linéaire de la surface ? Si oui, quels sont les coefficients de la droite de régression ?

La première étape consiste à faire une représentation graphique des données afin de bien s'assurer qu'il n'est pas absurde d'envisager un modèle linéaire.

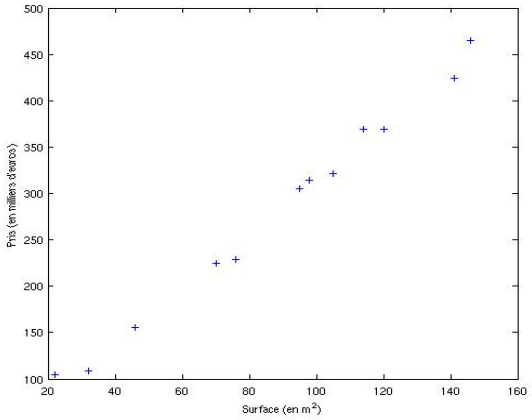

FIGURE 6.1 – Prix des biens immobiliers sur Lyon en fonction de leur surface, le 16/11/2010.

Dans le cas présent il semble tout à fait adapté de modéliser le prix en fonction de la surface par une régression linéaire simple.

6.1.2 Coefficient de corrélation linéaire empirique

Le but est de déterminer une grandeur permettant de confirmer ou d'infirmer que la relation entre les variables est linéaire. Nous rappelons la définition du coefficient de corrélation linéaire qui a été vue dans le cours de probabilité.

Si X et Y sont des variables aléatoires quantitatives, alors la covariance est définie par

$$Cov(X,Y) = \mathbb{E}\left[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Si la covariance est nulle ont dit que les variables ne sont pas corrélées. Deux variables indépendantes ne sont pas corrélées mais la réciproque est fausse.

La covariance donne une mesure de dépendance (on dit plutôt de corrélation) entre les variables X et Y. L'inconvénient est qu'elle n'est pas facilement interprétable dans la mesure où une covariance aura un ordre de grandeur très variable selon le type de données étudiées. C'est la raison pour laquelle on lui préfère en général le coefficient de corrélation linéaire.

Si X et Y sont des variables aléatoires quantitatives, alors le coefficient de corrélation linéaire est défini par

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

Nous avons $-1 \leqslant \rho(X,Y) \leqslant 1$. Lorsque $\rho(X,Y)^2 = 1$, cela signifie que nous avons une relation du type $Y = aX + b$ entre les variables X et Y.

Donnons maintenant les versions empiriques de ces grandeurs. Pour cela, nous introduisons les notations suivantes :

$\overline{x_n} = \frac{1}{n}\sum_{i=1}^n x_i$	moyenne empirique des x_i
$\overline{Y_n} = \frac{1}{n}\sum_{i=1}^n Y_i$	moyenne empirique des Y_i
$s_x^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \overline{x_n}^2$	variance empirique des x_i
$S_Y^2 = \frac{1}{n}\sum_{i=1}^n Y_i^2 - \overline{Y_n}^2$	variance empirique des Y_i

La covariance empirique et le coefficient de corrélation linéaire empirique sont alors donnés respectivement par :

$C_{xY} = \frac{1}{n}\sum_{i=1}^n x_i Y_i - \overline{x_n}\,\overline{Y_n}$	covariance empirique entre les x_i et les Y_i
$R_{xY} = \frac{C_{xY}}{s_x S_Y}$	coefficient de corrélation linéaire empirique entre les x_i et les Y_i

De même $-1 \leqslant R_{xY} \leqslant 1$.

Afin de mieux visualiser ce que représente le coefficient R_{xY}, voici quelques exemples :

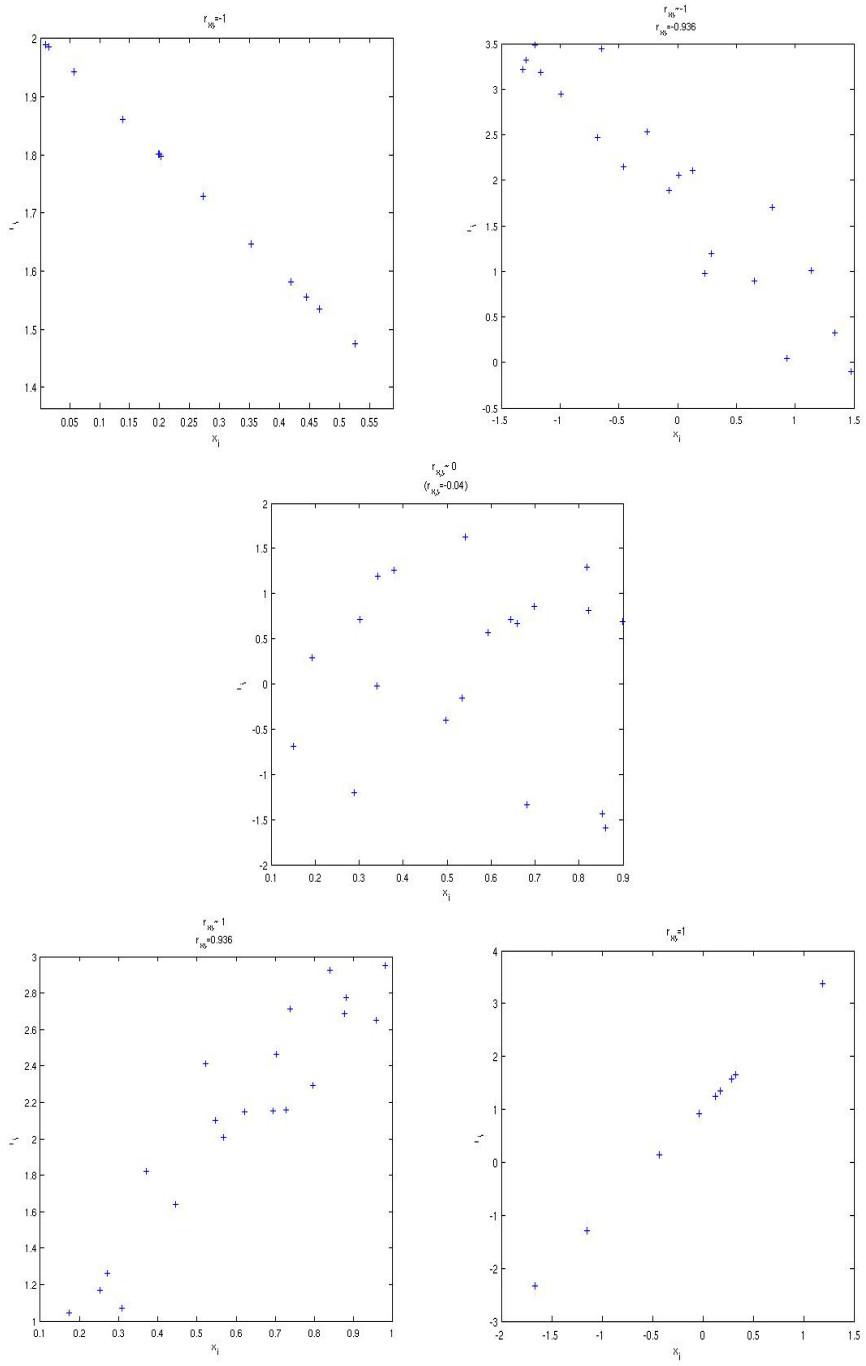

FIGURE 6.2 – Représentation des Y_i en fonction des x_i pour différentes valeurs de R_{xY}.

Nous voyons tout de suite qu'un modèle linéaire simple sera valide pour une valeur de R^2_{xY} proche de 1 et ne sera pas cohérent lorsque R_{xY} sera proche de 0. Sous une hypothèse de lois sur les résidus, il sera alors envisagé de construire un test de pertinence de la régression en testant si R^2_{xY} est suffisament proche de 1. Dans le cas présent, en l'absence de lois, il n'est pas possible d'envisager un tel test.

Dans l'exemple des prix des biens immobiliers sur Lyon en fonction de la surface, nous avons un coefficient de corrélation linéaire empirique égal à $r_{xY} = 0,996$. Le modèle linéaire semble donc bien adapté.

6.1.3 Estimation de la droite de régression par moindres carrés

Une fois que la représentation graphique et le calcul du coefficient de corrélation linéaire empirique incitent à essayer d'ajuster un modèle de régression linéaire simple, reste à déterminer comment réaliser cet ajustement. Rappelons que le modèle de régression linéaire simple s'écrit :

$$\forall i = 1, \dots, n, \quad Y_i = a_0\, x_i + b_0 + \varepsilon_i,$$

avec a_0 et b_0 paramètres inconnus et des résidus ε_i indépendants, de même loi et tels que $\mathbb{E}\varepsilon_i = 0$ et $Var(\varepsilon_i) = \sigma^2$.

Le but est alors d'estimer a et b puis ultérieurement σ^2. Nous optons pour une approche dite *des moindres carrés*. L'idée est de minimiser un critère mesurant l'erreur commise lorsqu'on résume le nuage de points par une droite.

Supposons que nous considérons que les points sont sur la droite d'équation $Y = a\, x + b$. Alors l'erreur commise pour le $i^{\text{ème}}$ point vaut $\delta_i = Y_i - (a\, x_i + b)$. L'erreur totale peut alors être évaluée par $\delta^2 = \sum_{i=1}^n \delta_i^2$, chaque erreur étant élevé au carré afin de bien prendre en compte l'éloignement des points à la droite de manière similaire. Cette grandeur est appelé *critère des moindres carrés*, elle s'écrit donc

$$\delta^2 = \sum_{i=1}^{n} (Y_i - a\, x_i - b)^2.$$

Le but est alors de trouver l'équation de droite minimisant ce terme d'erreur, c'est-à-dire de déterminer a et b minimisant δ^2.

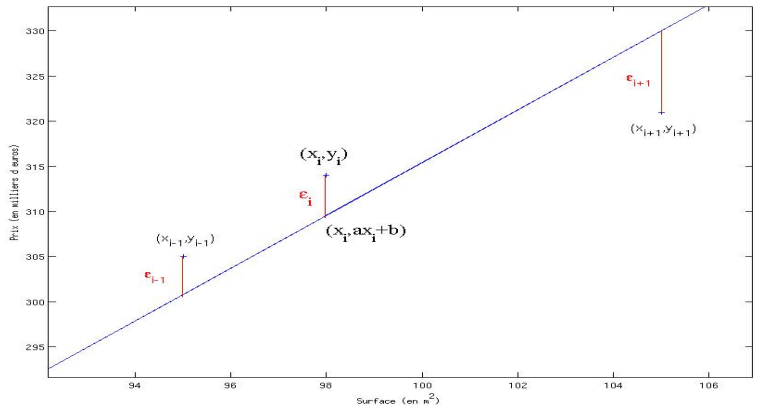

FIGURE 6.3 – Construction du critère des moindres carrés.

Pour ce faire, nous dérivons δ^2 et nous résolvons le système :

$$\begin{aligned} \frac{\partial \delta}{\partial a} &= 0, \\ \frac{\partial \delta}{\partial b} &= 0. \end{aligned}$$

Nous obtenons :

$$\begin{aligned} \frac{1}{n}\sum x_i(Y_i - a\,x_i - b) &= 0, \\ \frac{1}{n}\sum (Y_i - a\,x_i - b) &= 0. \end{aligned}$$

La deuxième équation donne $b = \overline{Y_n} - a\,\overline{x_n}$. En remplaçant dans la première, nous en déduisons que $a = C_{xY}/s_x^2$. Ainsi nous avons le résultat suivant :

BILAN :

Les estimateurs des moindres carrés de a_0 et b_0 sont

$$\widehat{a}_n = \frac{C_{xY}}{s_x^2} \text{ et } \widehat{b}_n = \overline{Y_n} - \widehat{a}_n\,\overline{x_n}.$$

En appliquant ces formules dans l'exemple des biens immobiliers sur Lyon, il vient : $\widehat{a}_n = 2,92$ et $\widehat{b}_n = 23,11$. Le paramètre b s'interprète comme un coût fixe commun à tout les biens vendus, et le paramètre a comme le prix d'un m^2, en milliers d'euros.

Ainsi, le coût du mètre carré à Lyon est d'environ 2923 euros au vu de nos données. (Ce résultat est cohérent avec les chiffres donnés par les agences notariales.) A titre d'information, voici les prix approximatifs du mètre carré dans différentes villes de France à cette date : Brest 1450 euros/m^2, Toulouse 2350 euros/m^2, Marseille 2500 euros/m^2, Bordeaux 2700 euros/m^2 et Paris 7500 euros/m^2.

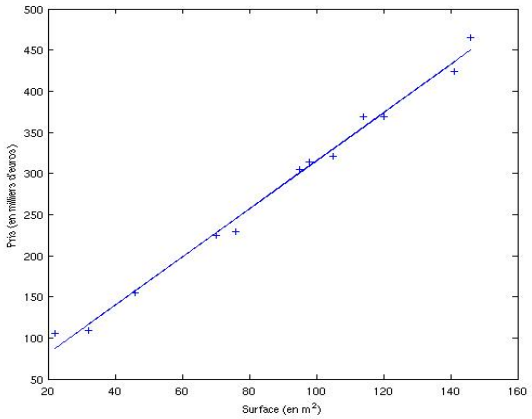

FIGURE 6.4 – Prix des biens immobiliers sur Lyon en fonction de leur surface et droite des moindres carrés.

Les résidus résultant de notre modélisation peuvent ensuite être estimés par $\widehat{\varepsilon}_i = Y_i - \widehat{a}_n\, x_i$. Nous admettons le résultat suivant :

BILAN :

La variance des résidus peut être estimée par

$$\widehat{\sigma}_n^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{a}_n x_i - \widehat{b}_n)^2 = \frac{n}{n-2}S_Y^2(1 - R_{xY}^2).$$

Dans l'exemple des prix des biens immobiliers sur Lyon en fonction de la surface, nous avons $\widehat{\sigma}_n^2 = 118,6$.

Remarquons enfin que l'intérêt d'avoir élaboré un tel modèle, outre qu'il aide à mieux comprendre le lien entre les deux variables x et Y, est qu'il permet de faire de la prévision. En effet, pour un x_0 donné, nous pouvons estimer que la réalisation de Y correspondante sera $\widehat{y}_0 = \widehat{a}_n\, x_0 + \widehat{b}_n$.

Dans l'exemple des biens immobiliers, supposons que vous ayez un bien d'une surface

de 35 m^2 situé proche des biens relevés ici. Alors il semble raisonnable de le mettre en vente à $\widehat{y} = 2,92 \times 35 + 23,11 = 125,31$ millers d'euros.

Cependant, si ce modèle permet déjà le calcul de plusieurs grandeurs pertinentes et la compréhension des liens, nous pouvons constater qu'il est assez limité : aucnu outil statistique puissant n'est disponible, pour calculer des intervalles de confiance ou réaliser des tests. Pour cela, il est nécessaire d'introduire une loi de probabilité sur les résidus. C'est l'objet de la section suivante.

6.2 Le modèle de regression linéaire simple gaussien

Nous supposons ici que les résidus suivent une loi normale. Autrement dit, nous avons :

$$\forall i = 1, \dots, n, \quad Y_i = a_0\, x_i + b_0 + \varepsilon_i,$$

avec a_0 et b_0 paramètres inconnus et des résidus ε_i indépendants, de même loi $\mathcal{N}(0, \sigma^2)$. Cette hypothèse est de loin la plus usuelle, mais remarquez qu'en pratique, vous pouvez rencontrer d'autres lois : lois de Poisson en astronomie, lois binomiales lorsque Y correspond à l'appartenance à une classe, etc.

Lorsque les résidus sont gaussiens, il est immédiat que nous avons alors les lois suivantes :

BILAN :

- $\widehat{a}_n \sim \mathcal{N}(a_0, \frac{\sigma^2}{n s_x^2})$,
- $\widehat{b}_n \sim \mathcal{N}(b_0, \frac{\sigma^2}{n}(1 + \frac{\overline{x_n}^2}{s_x^2}))$,
- $\widehat{a}_n\, x + \widehat{b}_n \sim \mathcal{N}(a_0\, x + b_0, \frac{\sigma^2}{n}(1 + \frac{(x-\overline{x_n})^2}{s_x^2}))$,
- $\frac{(n-2)\widehat{\sigma}_n^2}{\sigma^2} \sim \chi^2_{n-2}$,
- $\widehat{\sigma}_n^2$ est indépendante de $\widehat{a}_n$ et de $\widehat{b}_n$.

Nous pouvons en déduire des intervalles de confiance et des tests sur les paramètres a, b et σ^2, ainsi que pour une prévision, $\widehat{Y} = \widehat{a}_n x + \widehat{b}_n$.

Considérons par exemple le cas où nous voulions estimer le prix d'un bien de 35 m^2. Nous avions estimé un prix de $\widehat{y} = 125,31$ milliers d'euros. Pouvons-nous donner un intervalle de confiance ? Ceci permettrait notamment de vérifier que le prix proposé par une agence est cohérent avec le marché.

D'après ce qui précède, en appliquant le théorème de Fisher,

$$\frac{\widehat{Y} - y}{\sqrt{\frac{\widehat{\sigma}^2}{n}(1 + \frac{(x-\overline{x_n})^2}{s_x^2})}} \sim \mathcal{S}t(n-2).$$

Ainsi, l'intervalle de confiance d'ordre $1-\alpha$ pour y est :

$$\left[\widehat{Y} \pm t_{n-2,\alpha}\sqrt{\frac{\widehat{\sigma}^2}{n}(1 + \frac{(x-\overline{x_n})^2}{s_x^2}}\right].$$

L'application pour un bien de 35 m^2 donne $IC95\%(y) = [113,67; 136,96]$. Si l'agent immobilier a proposé 110 milliers d'euros, vous pouvez demander à monter le prix.

6.2.1 Test de pertinence

Afin de vérifier que le choix d'un modèle linéaire était bien justifié nous souhaitons tester (H_0) $a = 0$ contre (H_1) $a \neq 0$. C'est ce qu'on appelle le test de pertinence de la régression.

Les résultats qui précèdent permette de déduire que

$$\frac{\widehat{a}_n - a_0}{\widehat{\sigma}_n/(\sqrt{n}s_x)} \sim \mathcal{S}t(n-2).$$

Par conséquent la région critique de ce test pour le risque α est

$$RC_\alpha = \{|T| > t_{n-2;\alpha}\} \text{ avec } T = \frac{\widehat{a}_n}{\widehat{\sigma}_n/(\sqrt{n}s_x)}.$$

Dans l'exemple sur les biens immobiliers de Lyon, la réalisation de T vaut $t = 37.6$. Comme $t_{10;1\%} = 3.169$ nous pouvons en déduire qu'au seuil de 1%, la régression est bien pertinente.

6.2.2 Test sur la constante

La présence d'un prix fixe peut sembler surprenante. Il paraît naturel de considérer que le prix des biens immobiliers Y_i varie en fonction de la surface selon $Y_i = a\,x_i + \varepsilon_i$. Nous souhaitons donc réaliser le test de (H_0) $b = 0$ contre (H_1) $b \neq 0$.

Nous pouvons montrer que

$$\frac{\widehat{b}_n - b}{\frac{\widehat{\sigma}}{\sqrt{n}}\sqrt{1 + \frac{\overline{x_n}^2}{s_x^2}}} \sim \mathcal{S}t(n-2).$$

La région critique de seuil α est donc

$$RC_\alpha = \{|T| > t_{n-2;\alpha}\} \text{ avec } T = \frac{\widehat{b}_n}{\frac{\widehat{\sigma}}{\sqrt{n}}\sqrt{1 + \frac{\overline{x_n}^2}{s_x^2}}}.$$

La réalisation de T vaut $t = 3.05$ et est inférieure à $t_{10;1\%}$. Au seuil de 1%, nous ne pouvons pas affirmer que la présence de la constante b dans la régression est pertinente.

6.2.3 Etude des résidus

Afin de valider le modèle il faut aussi vérifier que les hypothèses formulées sur les résidus sont bien valables. Plutôt que de donner une approche numérique, nous optons ici pour une approche graphique.

Dans un premier temps, nous pouvons vérifier que les résidus sont indépendants. Si nous représentons le nuage de points des résidus estimés, nous ne devons pas voir apparaître de structure : si la valeur d'un résidu semble dépendre de la valeur du résidu précédent, cela signifie qu'il y a dépendance. Ceci est souvent le cas par exemple lorsque l'on étudie une régression au cours du temps (i.e. où x_i est le temps d'observation de Y_i).

Dans l'exemple de l'immobilier, nous observons bien un nuage de points sans structure visible, ce qui conforte donc l'hypothèse d'indépendance.

Ensuite, nous voulons vérifier qu'ils suivent bien une loi normale et que leur variance est identique. En utilisant les lois des estimateurs ci-dessus et le théorème de Fisher, cela implique que $\widehat{\varepsilon}_i/\widehat{\sigma}$ suit une loi de student à $n-2$ degré de liberté. Si nous représentons ces variables, dits résidus studentisés, nous devons observer que tous (excepté éventuellement une exception) doivent être compris dans l'intervalle inter-déciles $[t_{n-2;1\%};\ t_{n-2;99\%}]$.

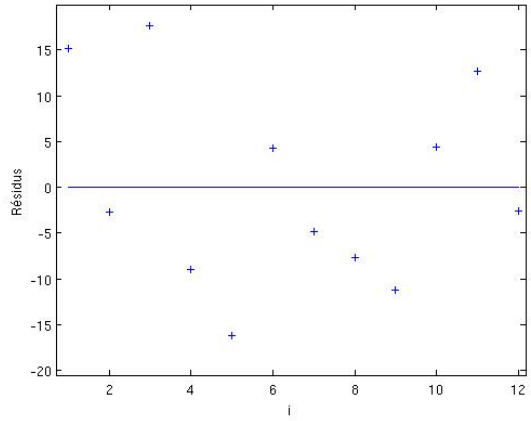

FIGURE 6.5 – Résidus estimés dans la régression des prix des biens immobiliers sur Lyon en fonction de leur surface.

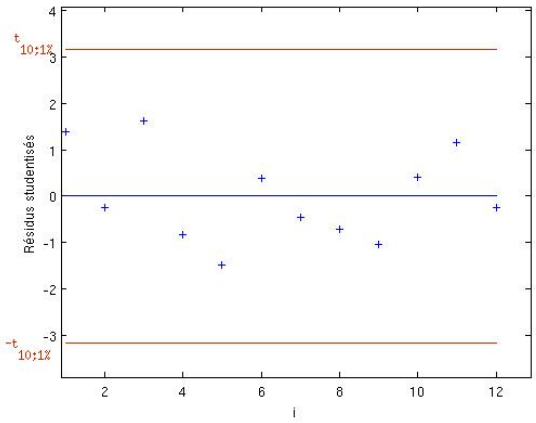

FIGURE 6.6 – Résidus studentisés dans la régression des prix des biens immobiliers sur Lyon en fonction de leur surface.

Enfin, remarquons que le caractère linéaire de la régression n'est validé que si les résidus ne dépendent pas des x_i. En effet, cela signifierait que nous pouvons trouver une fonction f telle que $\varepsilon_i = f(x_i) + \xi_i$ et donc que le modèle $Y_i = a\,x_i + b + f(x_i) + \xi_i$ serait plus adapté que le modèle choisi.

Dans notre exemple, le modèle linéaire est validé : aucune structure n'apparaît dans la représentation des ε_i en fonction des x_i.

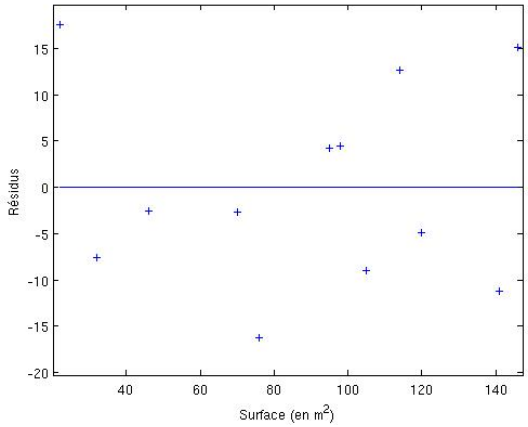

FIGURE 6.7 – Représentations des résidus an fonction de la surface dans la régression des prix des biens immobiliers sur Lyon.