MTH tc 3 : Probabilités et Statistique

Partie II : Statistique

Christophette BLANCHET & Céline Helbert

MTH TC3: Probabilités et Statistique

Année Universitaire 2023-2024

Table des matières

À la bibliothèque

Pour compléter, on pourra consulter au choix :

- un ouvrage destiné aux ingénieurs avec une base mathématique solide, et une partie supplémentaire sur l'analyse des données :

  Gilbert Saporta : Probabilités, analyse des données et statistique. *Editions Technip 2011*.

- un ouvrage américain très peu mathématique, avec beaucoup d'exemples :

  Douglas C. Montgomery et George C. Runger : Applied Statistics and Probability for Engineers. *Wiley Edition 2013*.

- une très bonne référence très pédagogique et bien illustrée d'exemples :

  Vincent Rivoirard et Gilles Stoltz : Statistique en action. *Vuibert*, *second edition, 2012*.

1 Estimation ponctuelle

La statistique inférentielle a pour but d'estimer un paramètre inconnu $\theta \in \Theta$ d'une population Ω à partir de l'observation d'un échantillon aléatoire $X_1, X_2, \ldots, X_n$. Elle propose de transporter (du latin *fero*) l'information collectée par l'échantillon à la population entière. On dispose de deux classes de méthodes :

- La théorie de l'estimation dont l'objet est d'estimer un ou plusieurs paramètres par un nombre ou un intervalle.
- La théorie des tests dont l'objectif est de confronter une hypothèse concernant les paramètres théoriques d'un modèle statistique aux valeurs mesurées sur un échantillon.

On introduit les notions suivantes que l'on va reprendre au cours de ce chapitre.

✍ **Données et échantillon** Les données sont les valeurs $x_1, x_2, \ldots, x_n$ prises par l'échantillon $X_1, X_2, \ldots, X_n$ prélevé aléatoirement dans la population. Les X_i sont des variables aléatoires supposées indépendantes et identiquement distribuées. Observons que les données sont n valeurs numériques alors que l'échantillon est une variable aléatoire.

✍ **Modèle statistique** C'est la donnée des lois de probabilité suivies par la variables aléatoires $X_1, X_2, \ldots, X_n$. En général, on suppose que les variables aléatoires $X_1, X_2, \ldots, X_n$ sont indépendantes et identiquement distribuées selon la même loi. La loi selon laquelle l'échantillon est distribué dépend d'un paramètre $\theta \in \Theta$ inconnu que l'on cherche à estimer.

✍ **Paramètres** Ce sont ceux de la **loi** de l'échantillon $X_1, \ldots, X_n$. En général, on note θ le paramètre (inconnu) du modèle. On suppose alors que θ appartient à Θ, l'ensemble des paramètres possibles du modèle. Le but est d'estimer ces paramètres à partir de l'observation de $x_1, \ldots, x_n$.

✍ **Estimateur** C'est une fonction f de l'échantillon $X_1, \ldots, X_n, T = f(X_1, \ldots, X_n)$. On l'appelle également statistique et on la note en général T ou $\hat{\theta}$. La statistique $\hat{\theta}$ fournit une estimation du paramètre θ. Observons qu'un estimateur T est une variable aléatoire.

✍ **Estimation** C'est une réalisation $t = f(x_1, \ldots, x_2)$ de la variable aléatoire T appelée estimation du paramètre θ.

1.1 Exemple introductif

🕹 **Contexte :** Plusieurs élèves de 1A à l'ECL habitant Gorge de Loup ont fait le trajet de Vaise vers l'Ecole Centrale via le bus C6/C6E la veille de la rentrée au hasard dans la journée. Ils ont chacun noté leur temps d'attente à l'arrêt de bus avant le départ du bus. A partir de ces différents temps d'attente qu'ils mettent en commun, ils voudraient connaître la périodicité de passage du bus C6/C6E.

Prénom	Mathias	Noé	Élise	Thibaut	Thomas
Temps d'attente (minutes)	2.67	5.5	4.17	1.83	9.33

⍰ Avec ces données peut-on estimer la périodicité θ du bus ?

🎓 **Formalisation probabiliste du problème :** On peut considérer que chaque observation du temps d'attente $x_i, 1 \leq i \leq 5$ est une réalisation d'une variable aléatoire $X_i, 1 \leq i \leq 5$. Ces observations ayant été faites **au hasard** dans la journée, on supposera que $X_1, \ldots, X_5$ sont identiquement distribuées de loi $\mathcal{U}([0, \theta])$. De plus les élèves ne se connaissant pas la veille de la rentrée on supposera que les variables aléatoires sont indépendantes. On dit qu'on est en présence d'un échantillon. Le problème est alors d'**estimer** le paramètre θ.

Remarque : la démarche statistique commence toujours par l'obtention des données. Une fois les observations obtenues, on cherche à caractériser le **modèle probabiliste** associé, c'est-à -dire le modèle dont les observations sont une réalisation. Dans cet exemple la loi de chaque variable aléatoire est uniforme sur $[0, \theta]$, cette loi est entièrement déterminée par un paramètre : θ. On parle d'**estimation paramétrique**. Les données servent à identifier le paramètre de cette loi. Une fois

le paramètre estimé, et donc la loi caractérisée, on peut mener les calculs de probabilité habituels nous permettant par exemple de répondre aux questions suivantes :

- ⑦ Quel est le temps d'attente moyen ?
- ⑦ Quelle est la probabilité d'attendre plus de 7 minutes à l'arrêt de bus ?
- ⑦ etc.

❷ Réponse au problème :
⑦ Que vaut θ ? Que peut-on donner comme estimation de θ ?

- $t_1 = \max(x_1, \ldots, x_5) = 9.33$
- $t_2 = 2 * \frac{x_1+\ldots+x_5}{5} = 9.4$

⑦ Quelle est la meilleure estimation ?
Ici $\theta = 10$, donc t_2 est meilleure (car plus près) que t_1.

⑦ Est-ce toujours le cas ?
On peut regarder le comportement de ces deux estimations sur un grand nombre d'échantillons.

- Sous MATLAB se donner $\theta = 10$ et obtenir une réalisation de (t_1, t_2) à partir d'un échantillon de taille 5 de la loi $\mathcal{U}([0, \theta])$.
- Recommencer 1000 fois et regarder les lois obtenues pour (T_1, T_2).

	T_1	T_2
Moyenne	8.3583	10.0150
Variance	1.8681	7.1578
Ecart-type	1.3668	2.6754

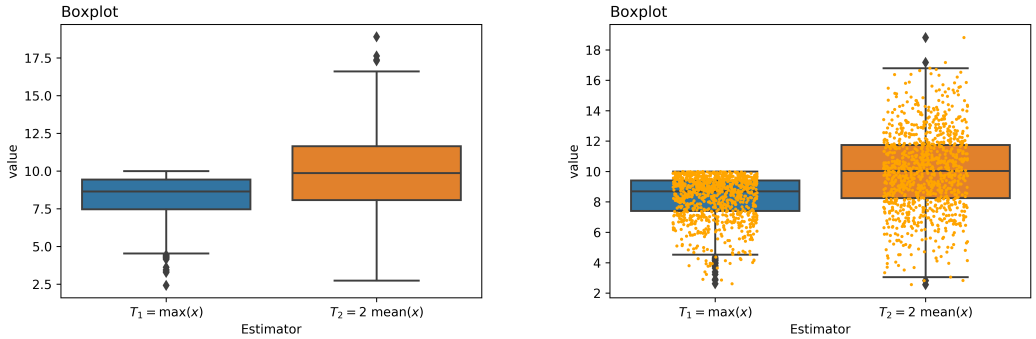

Figure 1 – Comparaison des deux estimations. On a représenté les 1000 points obtenus par MonteCarlo sur la figure de droite.

✍ Définition d'un Boxplot (cf. Figure 1) Il s'agit d'une représentation graphique d'une population sous forme de boite à moustaches comprenant

- Une boite centrale comprenant 50% de la population. Les bornes de la boite correspondent aux quantiles $q_{0.25}$ et $q_{0.75}$ (i.e. premier et troisième quartile). La valeur centrale correspond à la médiane.
- La moustache inférieure s'étend jusqu'à $x_{min} = \min\{x_i, x_i \geq q_{0.25} - 1.5 * (q_{0.75} - q_{0.25})\}$

- La moustache supérieure s'étend jusqu'à $x_{max} = \max\{x_i, x_i \leq q_{0.75} + 1.5 * (q_{0.75} - q_{0.25})\}$
- Toutes les autres observations sont tracées individuellement

Conclusions de l'exemple :

- Certaines estimations sont meilleures que d'autres :
  - la dispersion de t_1 est plus faible celle de t_2 donc T_1 est plus précis que T_2
  - t_1 sous-estime systématiquement la valeur de θ
  - la distribution des valeurs de t_1 est asymétrique alors que celle des valeurs de t_2 est symétrique
- Ce qui paraît pertinent n'est pas tant l'estimation ponctuelle mais plutôt l'incertitude associée à l'estimation.

Dans la suite du cours nous introduisons deux notions très importantes en statistique :

- celle d'estimateur et ses qualités associées
- celle d'estimation par intervalle de confiance en complément à l'estimation ponctuelle.

1.2 Définition et qualité d'un estimateur

1.2.1 Définition

Définition 1.1

> $(X_1, ..., X_n)$ *est un* ***échantillon*** *de la v.a.* X *(ou un* n*-échantillon de la v.a.* X*) si toutes les v.a.* X_i *sont indépendantes et suivent la même loi, celle de la v.a.* X*. On appellera* X *la variable parente.*

Remarque : On dit que l$X_1, ..., X_n$ sont i.i.d. à X (indépendantes, identiquement distribuées à X).

Définition 1.2

> *Soit* $(x_1, ..., x_n)$ *une réalisation d'un échantillon* $(X_1, ..., X_n)$ *de loi parente* $\mathcal{P}_\theta$*,* $\theta \in \mathbb{R}^p$*. On appelle* ***estimateur*** *du paramètre* θ *toute fonction réelle ou vectorielle de* $(X_1, ..., X_n)$*, notée* $T(X_1, ..., X_n)$*.*

Exemples pour $\mathcal{P}_\theta$:

- $\mathcal{P}_\theta = \mathcal{B}(p)$, où $\theta = p \in [0, 1]$
- $\mathcal{P}_\theta = \mathcal{N}(\mu, \sigma^2)$, où $\theta = (\mu, \sigma^2) \in \mathbb{R}^2$

Exemples d'estimateurs :

- $T(X_1, ..., X_n) = \dfrac{X_1 + ... + X_n}{n}$ (moyenne empirique)
- $T(X_1, ..., X_n) = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ (variance empirique)
- $T(X_1, ..., X_n) = (X_{(1)}, ..., X_{(n)})$ où $X_{(1)} < X_{(2)} < ... < X_{(n)}$ sont les statistiques d'ordre de l'échantillon.

Remarque : Un estimateur étant une fonction des v.a. de l'échantillon, c'est lui-même **une variable aléatoire**. En revanche, la réalisation de l'estimateur, appelée estimation, est une constante vectorielle ou scalaire.

1.2.2 Biais, risque et convergence

On a vu que toute fonction de $X_1, \ldots, X_n$ est un estimateur. Ainsi, la fonction constante égale à votre année de naissance est un estimateur. Il est donc utile de définir quelques propriétés "naturelles" que l'on voudrait satisfaites par un "bon" estimateur. Essentiellement, on souhaiterait contrôler l'erreur d'estimation, $T - \theta$. Cette erreur se décompose en une partie aléatoire plus une partie "biais" comme suit :

$$T - \theta = \underbrace{[T - \mathbb{E}_\theta(T)]}_{\text{aléatoire}} + \underbrace{[\mathbb{E}_\theta(T) - \theta]}_{\text{biais}}$$

où le symbole $\mathbb{E}_\theta$ est l'espérance lorsque les $X_1, \ldots, X_n$ suivent la loi de paramètre θ définie par le modèle statistique.

Définition 1.3

Le biais d'un estimateur T est la fonction :

$$\mathcal{B}_T : \Theta \to \mathbb{R}$$
$$\theta \mapsto \mathbb{E}\left[T(X_1, ..., X_n) - \theta\right]$$

où Θ est un ensemble de paramètres.

Remarque : On dit qu'un estimateur n'est pas biaisé, sans biais, ou encore de biais nul, si pour tout $\theta \in \Theta$, $\mathbb{E}\left[T(X_1, ..., X_n)\right] = \theta$.

Définition 1.4

Le risque d'un estimateur T est la fonction :

$$\mathcal{R}_T : \Theta \to \mathbb{R}$$
$$\theta \mapsto \mathbb{E}\left[(T(X_1, ..., X_n) - \theta)^2\right]$$

On peut noter que

$$\mathcal{R}_T(\theta) = \mathbb{V}\mathrm{ar}\left[T\right] + \mathcal{B}_T(\theta)^2$$

En effet,

$$\mathcal{R}_T(\theta) = \mathbb{E}\left[(T - \theta)^2\right]$$

$$= \mathbb{E}\left[(T - \mathbb{E}\left[T\right] + \mathbb{E}\left[T\right] - \theta)^2\right]$$

$$= \mathbb{E}\left[(T - \mathbb{E}\left[T\right])^2\right] + 2\mathbb{E}\left[T - \mathbb{E}\left[T\right]\right](\mathbb{E}\left[T\right] - \theta) + (\mathbb{E}\left[T\right] - \theta)^2$$

Remarque : le risque quantifie la distance moyenne au carré entre l'estimateur et le paramètre. On cherche en général des estimateurs de risque minimal, c'est-à -dire de précision maximale au sens de cette distance.

Définition 1.5

On dira que l'estimateur T (en réalité la suite d'estimateurs $(T_n)_{n\geq 1}$, T_n étant l'estimateur sur l'échantillon de taille n) est, lorsque les moments sont bien définis,

(i) asymptotiquement sans biais pour θ si $\lim_{n\to+\infty} \mathbb{E}\left[T_n\right] = \theta$;

(ii) convergent si $\lim_{n\to+\infty} \mathbb{E}\left[(T_n-\theta)^2\right]=0$.

Remarque : Un estimateur est convergent si et seulement si il est asymptotiquement sans biais et si sa variance converge vers 0 quand n tend vers l'infini.

1.2.3 Retour sur l'exemple

Calculons les biais, les variances et les risques quadratiques de $T_1=\max(X_1,...,X_n)$ et de $T_2=2\frac{\Sigma X_i}{n}$:

	Espérance $\mathbb{E}[T]$	Biais $\mathcal{B}_T(\theta)$	Variance $\mathbb{V}\mathrm{ar}[T]$	Risque $\mathcal{R}_T(\theta)$
T_1	$\frac{n\theta}{n+1}$	$\frac{-\theta}{n+1}$	$\frac{n\theta^2}{(n+1)^2(n+2)}$	$\frac{2\theta^2}{(n+1)(n+2)}$
T_2	θ	0	$\frac{\theta^2}{3n}$	$\frac{\theta^2}{3n}$

L'évolution avec n (taille de l'échantillon) de la variance et du risque de ces estimateurs est présentée sur la FIGURE 2. Ces deux estimateurs étant sans biais pour T_2 et asymptotiquement sans biais pour T_1, de variance tendant vers 0 en $+\infty$, ils sont convergents.

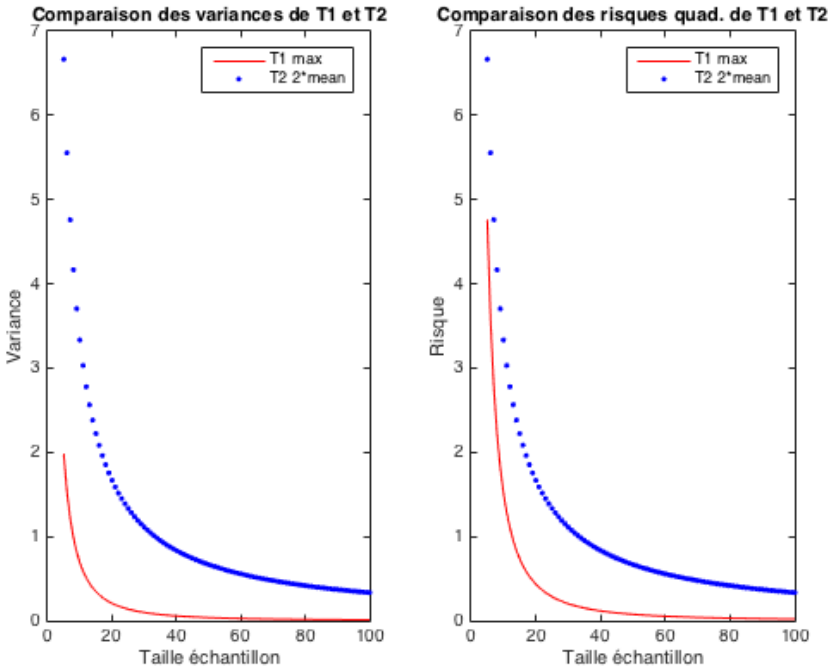

Figure 2 – Comparaison des deux estimateurs

Remarque : On observe que l'estimateur T_1, quoique biaisé, a toujours un risque quadratique plus faible que T_2. On peut avoir l'idée de construire un troisième estimateur non biaisé de θ de risque faible en débiaisant T_1 (cf FIGURE 3).

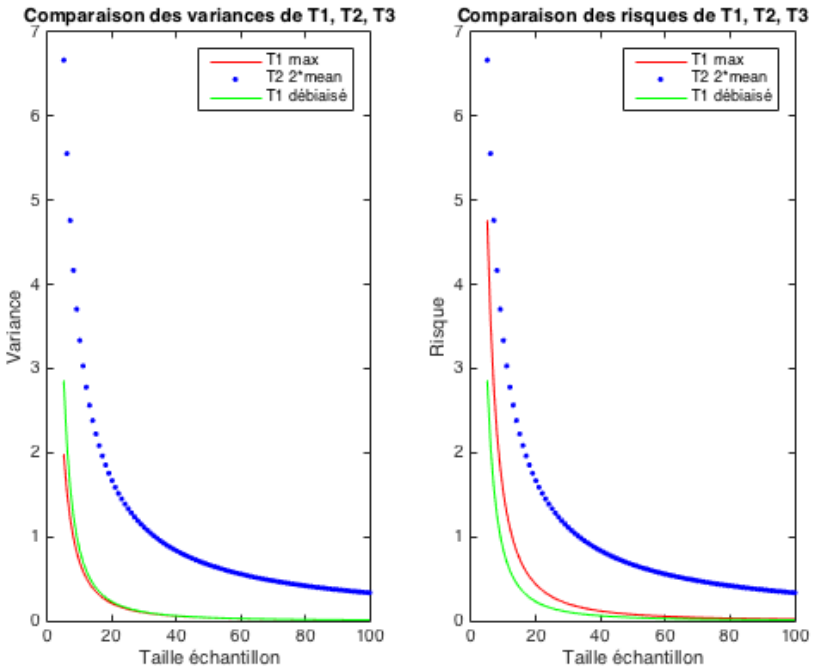

Figure 3 – Comparaison des trois estimateurs

1.3 Construction d'un estimateur

Dans cette section nous présentons plusieurs méthodes de construction d'estimateurs.

1.3.1 Méthode de substitution

On suppose que l'on dispose d'un estimateur $\hat{\theta}$ de θ [1]. Rappelons que θ est un scalaire inconnu alors que $\hat{\theta}$ est une variable aléatoire dont on espère avoir de bonnes propriétés par rapport à θ. On peut alors construire un estimateur de $g(\theta)$ en substituant θ par $\hat{\theta}$.

Exemple : Si on s'intéresse à la quantité $p = P(X > c)$ où X de loi $U([0, \theta])$ et $c < \theta$. Un calcul rapide donne $p = 1 - \frac{c}{\theta}$. On peut donc avoir $\hat{p} = 1 - \frac{c}{\hat{\theta}}$.

1.3.2 Méthode des moments

Exemple : Soit $(X_1, ..., X_n)$ un échantillon d'une v.a. X suivant une loi Gamma de paramètres λ et α que l'on cherche à estimer. On peut remarquer que :

$$\mathbb{E}[X] = \frac{\alpha}{\lambda} \text{ et } \mathbb{V}\text{ar}[X] = \frac{\alpha}{\lambda^2}.$$

Si on inverse le système on obtient que :

$$\lambda = \frac{\mathbb{E}[X]}{\mathbb{V}\text{ar}[X]}, \ \alpha = \frac{\mathbb{E}[X]^2}{\mathbb{V}\text{ar}[X]}.$$

1. **Remarque :** en statistique les estimateurs sont souvent notés par le symbole du paramètre complété d'un chapeau.

La méthode des moments consiste à identifier les paramètres de la loi en fonction des premiers moments de la variable aléatoire parente de l'échantillon. Une fois cette opération effectuée, il suffit de substituer $\mathbb{E}[X]$ et $\mathbb{V}\mathrm{ar}[X]$ par leurs estimateurs. On obtient alors :

Exemple :

$$\hat{\lambda} = \frac{\overline{X}}{S^2},\ \hat{\alpha} = \frac{\overline{X}^2}{S^2}.$$

Exemple : Concernant l'exemple introductif de la loi uniforme sur $[0, \theta]$, l'estimateur T_2 correspond à l'estimateur de la méthode des moments.

$$\mathbb{E}[X] = \frac{\theta}{2} \text{ et donc } \hat{\theta} = T_2 = 2\overline{X}$$

Définition 1.6

On suppose que la variable X est intégrable, de loi de paramètre $\theta \in \mathbb{R}$ et que l'espérance de X s'écrit comme une fonction de θ, soit

$$\mathbb{E}[X] = g(\theta).$$

Un estimateur T de θ obtenu par la méthode des moments est une fonction de l'échantillon $(X_1, ..., X_n)$ telle que

$$\overline{X} = g(T)$$

Définition 1.7

Si X est de carré intégrable (ou d'ordre supérieur) cette méthode peut permettre de déterminer des estimateurs vectoriels. Par exemple lorsque l'on a deux paramètres à estimer : soient θ_1, θ_2 ces paramètres. On suppose

$$\mathbb{E}[X] = g_1(\theta_1, \theta_2) \text{ et } \mathbb{V}\mathrm{ar}[X] = g_2(\theta_1, \theta_2).$$

On cherche alors des estimateurs T_1 et T_2 tels que

$$\overline{X} = g_1(T_1, T_2) \text{ et } S^2 = g_2(T_1, T_2).$$

Exercice 1 : Trouver les estimateurs des paramètres des lois Bernoulli, Exponentielle et Normale par la méthode des moments. |

Remarque : Les propriétés des estimateurs construits avec cette méthode dépendent directement des propriétés des estimateurs de la moyenne et de la variance. Le chapitre suivant a pour objet l'étude de ces estimateurs.

1.3.3 Méthode du maximum de vraisemblance

Cette méthode, quoique très utilisée en pratique, sera présentée en 2ème et/ou 3ème année.

1.4 Moyenne et variance empiriques : définition et premières propriétés.

Définition 1.8

Soit $(X_1, \ldots, X_n)$ un échantillon d'une v.a. X.

1. *On appelle moyenne empirique de l'échantillon la statistique (ou fonction de l'échantillon)*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

2. *On appelle variance empirique de l'échantillon la statistique (ou fonction de l'échantillon)*

$$\Sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Proposition 1.1

Si X est intégrable, $\overline{X}$ est intégrable et

$$\mathbb{E}\left[\overline{X}\right] = \mathbb{E}[X]$$

et si X est de carré intégrable, $\overline{X}$ est de carré intégrable et

$$\mathbb{V}\mathrm{ar}\left[\overline{X}\right] = \frac{\mathbb{V}\mathrm{ar}[X]}{n}.$$

Démonstration : cf TD de Probabilités.

Proposition 1.2

Si X est de carré intégrable,

$$\mathbb{E}\left[\Sigma^2\right] = \frac{n-1}{n}\mathbb{V}\mathrm{ar}\left[X\right].$$

et si $\mu_4 = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^4\right]$ est défini, on a

$$\mathbb{V}\mathrm{ar}\left[\Sigma^2\right] = \frac{n-1}{n^3}\left[(n-1)\mu_4 - (n-3)\mathbb{V}\mathrm{ar}\left[X\right]^2\right].$$

Démonstration : Première partie faite en TD de Probabilités, la seconde partie est un calcul fastidieux et est admise.

La moyenne empirique est un estimateur sans biais de l'espérance de X et que sa variance (et donc sa précision) converge vers 0 quand la taille de l'échantillon tend vers $+\infty$. Par ailleurs la variance empirique a un biais qui disparaît quand n tend vers $+\infty$ (on dit que cet estimateur est asymptotiquement sans biais) et sa précision converge également vers 0. Plutôt que Σ^2, on utilisera dans la suite l'estimateur sans biais de la variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n}{n-1}\Sigma^2.$$

Proposition 1.3

(admise). *Si $|X|^3$ est intégrable, on a :*

$$\mathbb{C}\mathrm{ov}\left[\overline{X}, \Sigma^2\right] = \frac{n-1}{n^2}\mathbb{E}\left[(X - \mathbb{E}\left[X\right])^3\right].$$

En particulier, $\overline{X}$ et Σ^2 sont asymptotiquement décorrélées et si la distribution de X est symétrique alors elles sont décorrélées.

1.5 Fonction de répartition empirique d'un échantillon

On peut définir la fonction de répartition empirique d'un échantillon. C'est un estimateur de la fonction de répartition.

Définition 1.9

Soit $(X_1, \ldots, X_n)$ un échantillon d'une v.a. X. On appelle fonction de répartition empirique de l'échantillon la famille d'estiumateurs définie par

$$F_n^\star(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{]-\infty,x]}(X_i) \text{ pour tout } x \in \mathbb{R}.$$

Théorème 1.1

Soit $(X_i)_{i\geq 1}$ une suite de v.a. i.i.d. à une v.a. X. Alors, pour tout $x \in \mathbb{R}$,

$$\lim_{n\to+\infty} F_n^\star(x) = F_X(x) \text{ presque sûrement}.$$

Démonstration : Soit $x \in \mathbb{R}$ fixé. Pour tout $i \in \{1, ..., n\}$, on pose

$$Y_i = \mathbf{1}_{]-\infty,x]}(X_i).$$

La loi de Y_i est : puisque Y_i prend les valeurs 0 ou 1, $Y_i \rightsquigarrow \mathcal{B}(p)$ avec $p = \mathbb{P}[Y_i = 1] = \mathbb{P}[X_i \leq x] = F_{X_i}(x) = F_X(x)$.
On récrit alors

$$F_n^\star(x) = \frac{Y_1 + ... + Y_n}{n}$$

et, puisque les Y_i sont indépendants, on peut appliquer la loi des grands nombres et il vient, quand $n \to +\infty$,

$$\text{p.s. } F_n^\star(x) \to \mathbb{E}[Y_1] \ = p = F_X(x).$$

2 Estimation par intervalle de confiance

On vient de voir dans le chapitre précédent que pour un seul et même paramètre nous pouvions construire plusieurs estimations, chaque estimation étant la réalisation d'un estimateur. Ces estimateurs ne sont pas tous équivalents :

- certains sont biaisés, d'autres non
- certains sont plus risqués que d'autres, c'est-à -dire qu'ils approchent le paramètre en question avec une moins grande précision.

Faire un choix d'estimateur peut se faire par l'étude du biais et du risque. Cependant une fois l'estimateur choisi, peut-on avoir une idée de la précision du calcul effectué ? Tout naturellement, cette précision va dépendre de la taille de l'échantillon et de la variance de l'estimateur choisi. L'idée de l'estimation par intervalle est de donner à l'utilisateur un intervalle qui a de très grandes chances de contenir le paramètre recherché. Cet intervalle est bien plus informatif sur la quantité d'information que nous apportent les données sur le paramètre que la valeur "ponctuelle" de l'estimation.

2.1 Exemple

Reprenons l'exemple du chapitre section . Peut-on calculer un intervalle qui a de très grandes chances de contenir le vrai paramètre θ ? Rappelons que $(X_1, ..., X_n)$ est un échantillon de loi $U([0, \theta])$ et que les deux estimateurs étudiés sont :

$$T_1 = \max(X_1, ..., X_n) \text{ et } T_2 = 2\overline{X}$$

Etude de T_1 :

- Le support de cette variable aléatoire est : $T_1(\Omega) = [0, \theta]$
- Quelle est la loi de T_1 ? Calculons la fonction de répartition et la densité de T_1 :

$$\text{Soit } t \in [0, \theta],\ P(T_1 \leq t) = P(X_1 \leq t, ..., X_n \leq t) = \left(\frac{t}{\theta}\right)^n$$

$$\text{et donc } f_{T_1}(t) = \frac{nt^{n-1}}{\theta^n}\mathbf{1}_{[0,\theta]}(t)$$

– Où se trouve T_1 avec une forte probabilité, ex. 95% ? Notons q_α tel que $P(q_\alpha \leq T_1) = 0.95$. On remarque que q_α est le quantile d'ordre $\alpha = 0.05$ de la loi de T_1. On a $q_\alpha = 0.05^{1/n}\theta$. $P(0.05^{1/n}\theta \leq T_1 \leq \theta) = 0.95 \Leftrightarrow P([T_1, \frac{T_1}{0.05^{1/n}}] \ni \theta) = 0.95$

Ainsi, avec les observations des élèves Thomas, Elise, Noé etc., on peut construire l'intervalle $[t_1, \frac{t_1}{0.05^{1/5}}] = [9.33, 16.99]$. Il s'agit d'un intervalle de confiance **exact** à 95% pour θ.

Etude de T_2 :

– Le support de cette variable aléatoire est :

$$T_2(\Omega) = [0, 2\theta]$$

– Quelle est la loi de T_2 ? T_2 est une somme de variables aléatoires uniformes indépendantes. Sa loi n'est pas connue, on pourrait la calculer par récurrence avec la formule de la loi de la somme de 2 variables aléatoires indépendantes. Ici on préfère utiliser une loi approchée (TCL) :

$$T_2 \rightsquigarrow N(\theta, \frac{\theta^2}{3n})$$

$$\text{ou encore } \frac{T_2 - \theta}{\frac{\theta}{\sqrt{3n}}} \rightsquigarrow N(0,1)$$

– Où se trouve T_2 avec une forte probabilité, ex. 95% ?

$$P(\theta - 1.96\frac{\theta}{\sqrt{3n}} \leq T_2 \leq \theta + 1.96\frac{\theta}{\sqrt{3n}}) = 0.95$$

$$\Leftrightarrow P([\frac{T_2}{1 + 1.96/\sqrt{3n}}, \frac{T_2}{1 - 1.96/\sqrt{3n}}] \ni \theta) = 0.95$$

Ainsi compte tenu des valeurs des observations concernant les élèves Thomas, Elise, Noé etc., on peut construire l'intervalle $[\frac{t_2}{1+1.96/\sqrt{15}}, \frac{t_2}{1-1.96/\sqrt{15}}] = [6.24, 19.03]$. Il s'agit d'un intervalle de confiance **asymptotique** à 95% pour θ.

À retenir

– Pour construire un intervalle de confiance, on a besoin de la loi de l'estimateur. Ensuite il suffit "d'isoler" θ dans les formules pour obtenir l'intervalle recherché.

– Il est parfois plus simple de faire le calcul d'un intervalle de confiance avec une loi approchée. Attention, dans ce cas, le niveau de confiance de l'intervalle est "approximatif". Sous Matlab, on peut évaluer par Monte Carlo (10^7 simulations) que le niveau réel de l'intervalle annoncé avec T_2 est 95.2%. Ce qui veut dire que l'intervalle construit est légèrement plus large de ce qu'il devrait être réellement.

2.2 Définition d'un intervalle de confiance

Dans la définition qui suit, θ est le paramètre à estimer (la moyenne ou la variance d'une variable aléatoire X dont on dispose d'un échantillon $(X_1,...,X_n)$.

Définition 2.1

Soit $0 < \alpha < 1$. On appelle intervalle de confiance pour le paramètre θ de niveau de confiance $1-\alpha$

(ou au seuil de risque α) un intervalle $[L_1; L_2]$ où L_1 et L_2 sont des variables aléatoires telles que $P(L_1 \leq \theta \leq L_2) = 1 - \alpha$.

Par construction L_1 et L_2 sont des variables aléatoires car elles dépendent des observations donc des X_1,...,X_n. On note l_1 et l_2 la réalisation de ces v.a. sur l'échantillon. On appelle alors aussi intervalle de confiance au seuil de risque α, l'intervale $[l_1, l_2]$.

Il est difficile de proposer une méthode systématique pour déterminer les intervalles de confiance. On détaille ci-dessous les intervalles de confiance pour la moyenne et la variance.

2.3 Construction d'intervalle de confiance pour la moyenne et la variance

Rappel : Nous avons vu que la méthode des moments est une façon d'obtenir des estimateurs. Ceux-ci sont alors des fonctions de la moyenne empirique (estimateur de l'espérance) et de la variance empirique (estimateur de la variance). Pour construire des intervalles de confiance, il convient donc d'étudier la loi de ces deux principaux estimateurs. C'est l'objet des deux sections suivantes.

2.3.1 Lois de $\overline{X}$ et Σ^2 dans le cas gaussien

Proposition 2.1

Soit $(X_1, ..., X_n)$ un échantillon de taille n de la v.a. X. On a les propriétés suivantes :

$$\overline{X} \rightsquigarrow \mathcal{N}\left(m, \frac{\sigma^2}{n}\right), \quad \frac{n\Sigma^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1} \left(\text{ou } \frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}\right),$$

$$\overline{X} \text{ et } \Sigma^2 \text{ (ou } S^2\text{) sont indépendantes et}$$

$$\frac{\overline{X} - m}{\Sigma/\sqrt{n-1}} \rightsquigarrow T_{n-1} \left(\text{ou } \frac{\overline{X} - m}{S/\sqrt{n}} \rightsquigarrow T_{n-1}\right).$$

Démonstration :
$X \rightsquigarrow \mathcal{N}(m, \sigma^2)$. Or

$$\Sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m + m - \overline{X})^2$$

$$\Sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - (\overline{X} - m)^2.$$

En multipliant par n/σ^2, on obtient :

$$\sum_{i=1}^{n}\left(\frac{X_i - m}{\sigma}\right)^2 = n\frac{\Sigma^2}{\sigma^2} + n\frac{(\overline{X} - m)^2}{\sigma^2}$$

soit

$$\sum_{i=1}^{n}\left(\frac{X_i - m}{\sigma}\right)^2 = n\frac{\Sigma^2}{\sigma^2} + \left(\frac{\overline{X} - m}{\sigma/\sqrt{n}}\right)^2.$$

Le théorème dit "de Cochran" (admis — une sorte de réciproque à la construction de la loi du chi-deux) nous permet de conclure ici par : puisque

$$\sum_{i=1}^{n}\left(\frac{X_i - m}{\sigma}\right)^2 \rightsquigarrow \chi^2_n \text{ et } \left(\frac{\overline{X} - m}{\sigma/\sqrt{n}}\right)^2 \rightsquigarrow \chi^2_1,$$

$$n\frac{\Sigma^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1} \quad \textbf{et} \quad \Sigma^2 \text{ et } \overline{X} \text{ indépendants}$$

D'autre part, puisque

$$\frac{\overline{X}-m}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1) \text{ et } n\frac{\Sigma^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$$

et ces variables sont indépendantes, on a, d'après la définition de la loi de Student,

$$\frac{\overline{X}-m}{\sigma/\sqrt{n}} \left(n\frac{\Sigma^2}{\sigma^2}/(n-1) \right)^{-1/2} \rightsquigarrow T_{n-1}.$$

soit

$$\frac{\overline{X}-m}{\Sigma/\sqrt{n-1}} \rightsquigarrow T_{n-1},$$

ce qui conclut la preuve de la proposition.
Remarquer que cette statistique ne dépend pas de σ.

Exemple : [Intervalle de confiance pour σ^2 dans le cas gaussien]
Pour de bonnes conditions de vieillissement, une cave à vin doit impérativement être bien isolée pour éviter des variations trop importantes de température préjudiciables à la qualité du vin. Il est donc essentiel de contrôler la variabilité de la température. On considère que la température dans une cave est une variable aléatoire sensiblement normale. Afin de contrôler la variabilité de la température, on a relevé 21 fois la température (en degrés Celsius) sur une période de 2 mois. Les résultats étant notés x_1,..., x_{21}, on calcule

$$\overline{x} = \frac{1}{21}\sum_{i=1}^{21} x_i = 11,66 \quad \text{et} \quad \frac{1}{21}\sum_{i=1}^{21} x_i^2 = 139,36.$$

Donner une estimation par intervalle, au seuil de risque 10%, de la variabilité des températures.

Les questions qu'on se pose sont :

1. Quel est le modèle probabiliste ?
2. Quel est le paramètre qu'on cherche à estimer ?
3. Quel est l'estimateur adapté ?
4. Connait-on sa loi ?

Les réponses sont les suivantes :

1. Il s'agit de proposer une formalisation
   Soit $(X_1, ..., X_n)$ un échantillon de taille n d'une v.a. X suivant une loi normale de paramètres m et σ^2 inconnus.
2. On veut construire un intervalle de confiance pour σ^2 au seuil de risque 10%.
3. Ici comme il s'agit du paramètre de variance, on utilise en général l'estimateur non biaisé de la variance : $S^2 = \frac{21}{20}\Sigma^2 = \frac{21}{20}\left(\overline{X^2} - \overline{X}^2\right)$.
4. Connait-on sa loi ? De part la proposition 2.1 on a :

$$(n-1)\frac{S^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$$

En notant $q_{\chi^2_{n-1},0.05}$ et $q_{\chi^2_{n-1},0.95}$ les quantiles d'ordres 5% et 95% de la loi du χ^2_{20} on :

$$P(q_{\chi^2_{n-1},0.05} \leq (n-1)\frac{S^2}{\sigma^2} \leq q_{\chi^2_{n-1},0.95}) = 0.90$$

et donc

$$P(\left[\frac{20S^2}{q_{\chi^2_{20},0.95}}; \frac{20S^2}{q_{\chi^2_{20},0.05}}\right] \text{contienne } \sigma^2) = 0.90$$

L'estimation est $s^2 = \frac{21}{20}(139.36 - 11.66^2) = 3.5746$, $q_{\chi^2_{n-1},0.05} = 10.86$, $q_{\chi^2_{n-1},0.95} = 31.41$, et donc $[2.28; 6.59]$ est un intervalle de confiance pour σ^2 au seuil de risque 10%.

Remarque : On peut noter que :

- Il s'agit d'un intervalle de confiance **exact** de niveau 90%.
- Nous aurions pu construire un intervalle unilatéral en tenant compte du fait que ce sont les grosses variances qui sont les moins souhaitées et les plus risquées.
- Avec ce même raisonnement nous pouvons construire des intervalles de confiance pour m avec variance connue ou inconnue ou pour σ^2 à moyenne connue ou inconnue. La construction de ces intervalles est laissée au lecteur à titre d'exercice (les résultats sont synthétisés dans l'annexe A).

2.3.2 Lois dans le cas général

Soit $(X_i)_{i\geq 1}$ une suite i.i.d. à X, une v.a.r. de carré intégrable d'espérance m et de variance σ^2. Pour tout $n \geq 1$, on note $\overline{X}_n$ la moyenne empirique de l'échantillon $(X_1, \ldots, X_n)$ et Σ_n^2 sa variance empirique.

Proposition 2.2

$(\overline{X}_n)_{n\geq 1}$ converge presque sûrement et en moyenne quadratique vers m et

$$\mathcal{L}\left(\frac{\overline{X}_n - m}{\sigma/\sqrt{n}}\right) \to \mathcal{N}(0,1) \text{ quand } n \to +\infty;$$

De manière équivalente, il existe $(Z_n)_n$ suite de variables aléatoires telles que

$$\overline{X}_n = \underbrace{m}_{\text{moyenne à estimer}} + \underbrace{\frac{\sigma}{\sqrt{n}}}_{\text{ordre de grandeur des fluctuations}} \times \underbrace{Z_n}_{\text{fluctuations standardes à la limite}}$$

$Z_n \to \mathcal{N}(0,1)$ *à la limite $n \to \infty$, Z_n convergence vers une gaussienne standard.*

Proposition 2.3

si X^2 est de carré intégrable alors $(\Sigma_n^2)_{n\geq 1}$ converge presque sûrement et en moyenne quadratique vers σ^2 et

$$\mathcal{L}\left(\frac{\Sigma_n^2 - \mathbb{E}\left[\Sigma_n^2\right]}{\sqrt{\mathbb{V}\text{ar}\left[\Sigma_n^2\right]}}\right) \to \mathcal{N}(0,1) \text{ quand } n \to +\infty.$$

Démonstration : Applications directe de la loi des grands nombres et du théorème limite central.

Exemple : On mesure les diamètres à 1m30 du sol de n arbres au hasard dans un bois. On note $x_1, ..., x_n$ les mesures obtenues. Donner un intervalle de confiance pour le diamètre moyen des arbres du bois.

Par le théorème limite central, un intervalle de confiance pour la moyenne peut-être obtenu si n est assez grand. L'intervalle de confiance au seuil de risque α a la forme suivante :

$$\left]\overline{X} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\right[$$

où $u_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite.

Remarque :

1. Quand la loi est connue et que la variance est une fonction g de l'espérance, on remplace σ par $\sqrt{g(\overline{X})}$. C'est le cas pour la loi de Bernoulli (cf. formulaire A.2), de Poisson, exponentielle etc.
2. Quand la loi n'est connue qu'au travers de réalisations de la variable aléatoire, on remplace σ par l'estimateur habituel S. C'est ce que l'on fait également pour déterminer les intervalles de confiance de la moyenne dans la méthode de Monte Carlo.

3 Théorie des tests

Un test statistique est une procédure qui quantifie l'adéquation entre des données obervées et une hypothèse sur le modèle statistique. A cette fin, un test regarde la valeur de l'écart entre la valeur théorique d'un paramètre (donnée par l'hypothèse faite sur le modèle) et la valeur observée. Un test statistique est plus concluant s'il permet de rejeter l'hypothèse (que l'on cherche à tester) car sa démarche s'apparente à un raisonnement par l'absurde : on suppose une hypothèse et on montre que l'on observe un résultat très peu probable sous cette hypothèse.

On introduit les notions suivantes que l'on va reprendre au cours de ce chapitre.

Hypothèse nulle et hypothèse alternative L'hypothèse nulle est une assertion sur θ que l'on note H_0. Elle est de la forme :

$$H_0 \ : \ \theta \in \Theta_0 \, ,$$

où Θ_0 est un sous ensemble de Θ. L'hypothèse alternative, que l'on note H_1, décrit l'ensemble des situations considérées si l'hypothèse nulle n'est pas satisfaite. Elle est de la forme :

$$H_1 \ : \ \theta \in \Theta_1 \, ,$$

où Θ_1 est un sous ensemble de Θ disjoint de Θ_0.

Règle de décision Soit $T = f(X_1, \ldots, X_n)$ une statistique, appelée "statistique de test", et $\mathcal{R}$ un sous ensemble des valeurs possibles de T, composée de valeurs très peu probables pour T. L'ensemble $\mathcal{R}$ est appelé "région de rejet" du test.

La règle de décision est alors : si $T \in \mathcal{R}$ alors on rejette H_0 et si $T \notin \mathcal{R}$ alors on ne peut pas rejeter H_0. Remarquons que T et $\mathcal{R}$ sont fixés avant d'oberver $x_1, \ldots, x_n$.

Test de l'hypothèse nulle de Fisher : la valeur p Il s'agit d'une autre procédure de test. Elle consiste à fixer au préalable un niveau de confiance α (0.05, 0.02, ou 0.001) et à calculer, à partir des données observées $x_1, \ldots, x_n$, une "valeur p" (encore appelée, le "niveau observé") qui quantifie la confiance que l'on a dans l'hyptohèse nulle. Si α est plus petit que la valeur p alors on ne peut rien conclure et il faut attendre d'autres données. Si α est plus grand que la valeur p alors soit l'hypothèse nulle est fausse, soit l'hypothèse nulle est vraie et il s'est passé quelque chose de très peu probable.

Nous n'aborderons pas le test de l'hypothèse nulle de Fisher dans ce cours. Cependant, la "valeur p" (centrale dans le cadre du test de l'hypothèse nulle de Fisher) est souvent mal utilisée dans le cadre de la théorie de la décision de Neyman-Pearson et il faudra faire très attention à ce qu'elle représente.

3.1 Un exemple : les faiseurs de pluie.

Situation :

Le niveau naturel des précipitations dans la Beauce en mm par an suit une loi normale $\mathcal{N}(600, 100^2)$.

Les "faiseurs de pluie" prétendent augmenter le niveau moyen annuel des précipitations par insémination des nuages avec de l'iodure d'argent.

Les agriculteurs souhaitent que le niveau augmente d'au moins 50mm par an en moyenne avant de financer ce projet.
Après insémination, on obtient les mesures suivantes :

année	1951	1952	1953	1954	1955
mm	510	614	780	512	501

année	1956	1957	1958	1959
mm	534	603	788	650

On a le choix entre deux hypothèses :

H_0 : l'insémination est sans effet,

H_1 : l'insémination augmente de 50 mm le niveau moyen de pluie.

1. *Le point de vue des agriculteurs* : ils adoptent H_0 et n'acceptent de l'abandonner que si la probabilité de le faire à tort est très faible, mettons $\alpha << 1$.
   On chercher un événement A qui se produit avec probabilité α sous l'hypothèse H_0 :

   $$\mathbb{P}[A|H_0] = \alpha \text{ et tel que}$$

   (i) connaissant les résultats des essais, on puisse déterminer s'il a été réalisé ou non ;

   (ii) l'événement A se réalise avec une forte probabilité sous l'hypothèse H_1.

   Remarque : L'idéal serait d'avoir A tel que $\mathbb{P}[A|H_0] = 0$ et $\mathbb{P}[A|H_1] = 1$ mais un tel événement n'existe pas en général.

   Le test consiste ensuite à vérifier si, lors des essais, cet événement s'est réalisé ou non. Deux cas se produisent :

   - Si A s'est réalisé, alors qu'il avait une probabilité α très faible de se réaliser sous H_0, les agriculteurs décideront de rejeter l'hypothèse H_0, et donc d'accepter H_1 ; ce faisant, ils ont une probabilité α de se tromper.
   - Si A ne s'est pas réalisé, les agriculteurs décideront alors de conserver l'hypothèse H_0 faute de raisons suffisantes de la rejeter.

   Remarque : L'événement A^c se réalise avec probabilité $1 - \alpha$ (proche de 1) sous H_0.

2. *Le point de vue des faiseurs de pluie.* Leur risque est mesuré différemment : ils redoutent que l'hypothèse H_1 soit rejetée alors qu'elle est bonne. On vient de voir que cette hypothèse sera rejetée par les agriculteurs si A n'est pas réalisé. Ils calculent donc

   $$\beta = \mathbb{P}[A^c|H_1].$$

   - Si β est petit, cela signifiera donc que, sous l'hypothèse H_1, l'événement qui a conduit à rejeter H_1 a très peu de chances de se réaliser, et l'idée que l'hypothèse H_1 n'est pas la bonne est confirmée.
   - Si β est grand, cela signifiera que l'événement qui a conduit à rejeter H_1 se réalise avec une forte probabilité sous H_1 et donc que le test que l'on a mis en place n'est pas significatif.

Résumé des notations :

- A : région de rejet de H_0.
- α : probabilité de rejeter à tort H_0.
- β : probabilité de conserver H_0 à tort.

Formalisation mathématique. On suppose que, après insémination des nuages :

- X suit une loi $\mathcal{N}(m, 100^2)$.
- Les mesures effectuées sont des réalisations d'un échantillon de taille $n = 9$ de la v.a. X.

On doit choisir entre

$$H_0 \ : \ [m = 600], \quad H_1 \ : \ [m \geq 650].$$

Etape 1
On fixe le seuil de risque accepté par les agriculteurs.

Par exemple $\alpha = 0,05$.

Etape 2
On détermine la région de rejet de H_0, A.

On la choisit de la forme

$$A = [(X_1, ..., X_n) \in W]$$

car A doit dépendre des données. Par ailleurs, puisque l'on teste une moyenne et puisque H_1 " tire vers le haut" la moyenne fixée par H_0, on choisit

$$W = \{(x_1, ..., x_n) \in \mathbb{R}^n / \ \overline{x} > k\}.$$

Reste à déterminer k. On a $A = [\overline{X} > k]$ donc k doit être tel que $\mathbb{P}\left[\overline{X} > k | H_0\right] = \alpha$. Or, sous H_0,

$$\overline{X} \rightsquigarrow \mathcal{N}\left(600, \frac{100^2}{n}\right) \text{ et donc } \frac{\overline{X} - 600}{100/3} \rightsquigarrow \mathcal{N}(0, 1).$$

Il suffit alors d'inverser la fonction de répartition de la loi $\mathcal{N}(0, 1)$: on trouve que

$$\mathbb{P}\left[\frac{\overline{X} - 600}{100/3} \leq 1,645 \, | H_0\right] = 0,95$$

et donc

$$\mathbb{P}\left[\frac{\overline{X} - 600}{100/3} > 1,645 \, | H_0\right] = 0,05.$$

On choisit donc k tel que

$$\frac{k - 600}{100/3} = 1,645 \text{ soit } k = 600 + \frac{100}{3}\text{x}1,645 = 655 \text{ environ.}$$

Finalement $A = [\ \overline{X} > 655]$.

Etape 3
On procède au test : $(x_1, ..., x_n) \in W$?

On trouve $\overline{x} \approx 610,2$. Donc $\overline{x} \leq k$ et $(x_1, ..., x_n) \notin W$, A n'est pas réalisé.

Sans complément d'information, les agriculteurs sont donc conduits à conserver H_0.

Etape 4
Calcul de la probabilité $\beta = \mathbb{P}[A^c|H_1]$ d'avoir conservé H_0 à tort.

Sous H_1, $m \geq 650$ et donc la loi de $\overline{X}$ dépend de m. On calcule le pire des cas, c' est-à -dire celui donnant le β le plus gros, soit

$$\sup_{m \geq 650} \mathbb{P}\left[\overline{X} \leq k | \mathbb{E}\left[\overline{X}\right] = m\right].$$

Ce sup est obtenu pour $m = 650$. En effet, lorsque m croît, $\mathbb{E}\left[\overline{X}\right]$ également et donc $[\overline{X} \leq k]$ a de moins en moins de chance de se réaliser (sa probabilité décroît).
Sous l'hypothèse $m = 650$, $\frac{\overline{X}-650}{100/3} \rightsquigarrow \mathcal{N}(0,1)$ et on trouve

$$\beta = \mathbb{P}\left[\frac{\overline{X}-650}{100/3} \leq \frac{k-650}{100/3} | \mathbb{E}\left[\overline{X}\right] = 650\right]$$

$$= \mathbb{P}\left[\frac{\overline{X}-650}{100/3} \leq 0,15 | \mathbb{E}\left[\overline{X}\right] = 650\right] \approx 0,56.$$

Le risque de conserver H_0 à tort est donc considérable et les agriculteurs ont peut-être eu tort de le faire. On peut aussi interpréter ce résultat avant même de faire le test sur les données en disant que le test n'est pas très bon puisque l'événement A dont la réalisation conduit à rejeter H_0 et accepter H_1, a une probabilité relativement petite $1 - \beta = 0,44$ de se réaliser sous H_1.

3.2 Notions générales.

On dispose d'observations $(x_1, ..., x_n)$ qui sont la réalisation d'un échantillon $(X_1, ..., X_n)$ d'une variable aléatoire X (réelle ou vectorielle). Un test statistique définit une règle de décision pour choisir entre deux hypothèses H_0 et H_1 faites sur la loi de X au vu des données recueillies.

Les hypothèses H_0 et H_1 ne jouent pas le même rôle, l'hypothèse H_0 est celle à laquelle on tient le plus, qu'on ne veut rejeter qu'avec une faible probabilité de le faire à tort. De plus, pour pouvoir procéder à un test il faut impérativement être capable de faire des calculs sous l'hypothèse H_0, elle doit donc être suffisamment précise alors que l'hypothèse H_1 peut être relativement vague (la négation de H_0 par exemple). Bien sûr les hypothèses H_0 et H_1 doivent s'exclure mutuellement.

Construction et utilisation du test :

1. On fixe $\alpha > 0$ petit, le *risque de première espèce* qui est la probabilité de rejeter H_0 à tort.
2. On détermine une *région de rejet de* H_0, $W \in \subset \mathbb{R}^n$, telle que

   $$\mathbb{P}[(X_1, ..., X_n) \in W | H_0] = \alpha.$$

   Cette région dépend fortement des hypothèses que l'on considère. En particulier, elle dépend de H_1 en ce sens que l'on souhaite que la probabilité

   $$1 - \beta = \mathbb{P}[(X_1, ..., X_n) \in W | H_1]$$

   soit la plus grande possible. En effet, $1 - \beta$, qui s'appelle la *puissance du test*, mesure la probabilité que les données soient dans la région de rejet de H_0 lorsque H_1 est vraie.

Cette région de rejet est en règle générale construite à partir d'une *statistique* ou *variable de décision* D (fonction de l'échantillon). Cette statistique est construite de manière à connaître sa loi sous H_0. Dans le cas de l'exemple des faiseurs de pluie, le test portait sur l'espérance de la variable parente de l'échantillon et la statistique était donc liée à l'estimateur de la moyenne (ici dans le cas gaussien à variance connue)

$$D = \frac{\overline{X} - m}{\sigma/\sqrt{n}}.$$

A partir de cette statistique, on a construit une zone de rejet W (ou A) en fonction de H_0 et H_1 et α. La zone de rejet s'écrit $\{(X_1, ..., X_n) \in W\} = \{D \in W_D\}$.

3. Règle de décision : si la réalisation $(x_1, ..., x_n)$ de notre échantillon est dans W, on rejette H_0 ; sinon, on conserve H_0. On fait ce test sur la réalisation d de la variable de décision D : si $d \notin W_D$ on ne rejette pas H_0 ; si $d \in W_D$, on rejette H_0.

Finalement, construire un test, c'est se donner les hypothèses H_0 et H_1, le seuil de risque α petit, la variable de décision D, la zone de rejet W_D et, si on peut la calculer, la puissance du test $1 - \beta$.

Remarque : Le paramètre

$$\beta = \mathbb{P}\left[(X_1, ..., X_n) \in W^c | H_1\right]$$

s'appelle le *risque de seconde espèce* ; c'est la probabilité de conserver H_0 alors que H_1 est vraie. Ce risque doit être aussi petit que possible à α fixé.

Remarque : Heuristiquement, il est assez facile de se convaincre que, lorsqu'on diminue α, on diminue la taille de la région de rejet W et donc on diminue également la puissance du test (ou on augmente le risque de seconde espèce). Par conséquent, on ne peut choisir α trop petit (les valeurs usuelles de α sont , 0.1 ou 0.05 ou 0.01).

Qualité d'un test : On a vu que la qualité d'un test est mesurée par sa puissance $1 - \beta$. D'autre part :

- si $1 - \beta > \alpha$, on dit que le test est sans biais.
- si $1 - \beta \to 1$ lorsque la taille de l'échantillon n tend vers l'infini, on dit que le test est convergent.

Risque de première espèce On le note $\alpha(\theta)$, c'est la probabilité de rejeter H_0 alors que celle-ci est vraie

$$\forall \theta \in \Theta_0\,, \quad \alpha(\theta) := \mathbb{P}_\theta(T(X_1,\ldots,X_n) \in \mathcal{R})\,,$$

où $\mathbb{P}_\theta$ indique que l'échantillon $X_1,\ldots,X_n$ suit la loi $\mathcal{P}(\theta)$.

Niveau On la note α, c'est la valeur la plus élevée du risque de première espèce pour $\theta \in \Theta_0$

$$\alpha := \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(T(X_1,\ldots,X_n) \in \mathcal{R})\,.$$

Risque de deuxième espèce On le note $\beta(\theta)$, c'est la probabilité d'accepter H_0 alors que celle-ci est fausse

$$\forall \theta \in \Theta_1\,, \quad \beta(\theta) := \mathbb{P}_\theta(T(X_1,\ldots,X_n) \notin \mathcal{R})\,,$$

où $\mathbb{P}_\theta$ indique que l'échantillon $X_1,\ldots,X_n$ suit la loi $\mathcal{P}(\theta)$.

Puissance C'est une fonction de $\theta \in \Theta_1$ que l'on note $\pi(\theta)$. Elle représente la probabilité de rejet H_0 alors que celle-ci est fausse

$$\forall \theta \in \Theta_1\,, \quad \pi(\theta) = 1-\beta(\theta) = \mathbb{P}_\theta(T(X_1,\ldots,X_n) \in \mathcal{R})\,,$$

où $\mathbb{P}_\theta$ indique que l'échantillon $X_1,\ldots,X_n$ suit la loi $\mathcal{P}(\theta)$.

Classification des tests. On dit que le test est *paramétrique* lorsque les hypothèses portent sur la valeur d'un ou plusieurs paramètres de la loi de X. Si un même test convient pour différentes lois, on dit que le test est *robuste* (comme les tests de moyenne, par exemple). Parmi les tests *non paramétriques* (qui sont robustes), on trouve les *tests d'ajustement* à une loi donnée. Il existe également des tests de *comparaison de plusieurs échantillons* qui permettent de déterminer si des échantillons sont issus d'une même population. Enfin, on verra comment tester si deux variables aléatoires sont indépendantes.

3.3 Tests paramétriques.

On fait une hypothèse sur le paramètre θ (l'espérance ou la variance) de la loi d'une v.a. X. On dispose de la réalisation d'un échantillon $(X_1, ..., X_n)$ de la v.a. X.
Les hypothèses que l'on peut formuler sont de deux types :

- hypothèse simple : $[\theta = \theta_0]$ où $\theta_0 \in \mathbb{R}^d$ est une valeur fixée du paramètre ;
- hypothèse composite : $[\theta \in B]$ où B est une partie de $\mathbb{R}^d$ non réduite à un point.

Noter que, lorsque le paramètre est réel, une hypothèse composite a souvent la forme $[\theta < \theta_0]$, $[\theta > \theta_0]$ ou $[\theta \neq \theta_0]$ pour une valeur fixée θ_0 du paramètre.

Remarque : Pour nous, l'hypothèse H_0 sera toujours une hypothèse simple, pour pouvoir faire tous les calculs.

3.3.1 Test entre deux hypothèses simples.

On suppose $\theta_0 \neq \theta_1$ et

$$H_0 \ : \ [\theta = \theta_0], \quad H_1 \ : \ [\theta = \theta_1].$$

La variable de décision est la variable qui sert à construire un intervalle de confiance pour le paramètre θ, comme cela a été fait chapitre 2. Ces variables de décision sont rappelées en annexe B.
On suppose en toute généralité que D est un estimateur de θ. Alors

- si $\theta_1 > \theta_0$, la zone de rejet a la forme $[D > d]$;
- si $\theta_1 < \theta_0$, la zone de rejet a la forme $[D < d]$.

Remarque : On peut montrer que les tests ainsi construits sont les meilleurs possibles au sens que la zone de rejet construite est celle qui, parmi toutes celles de probabilité α sous H_0, a la plus forte probabilité sous H_1.

3.3.2 Test d'une hypothèse simple contre une hypothèse composite : la fonction puissance.

On suppose

$$H_0 \ : \ [\theta = \theta_0], \quad H_1 \ : \ [\theta \in B]$$

où θ_0 est une valeur fixée du paramètre et B une partie de $\mathbb{R}$ ne contenant pas θ_0.
Même si l'on connaît la loi de la variable parente X, on ne peut calculer la puissance d'un test car H_1 n'est pas assez précise. Par contre, pour tout $\theta_1 \in B$, on peut calculer la puissance d'un test pour les hypothèses

$$H_0 \ : \ [\theta = \theta_0], \quad H_1 \ : \ [\theta = \theta_1].$$

On appelle alors *fonction puissance du test* la fonction, définie sur B par $\theta_1 \in B \longmapsto 1 - \beta(\theta_1)$. On recherche alors le test *uniformément le plus puissant* (UPP en abrégé), c'est-à-dire, s'il existe, celui tel que, pour tout $\theta_1 \in B$, sa puissance en θ_1 est supérieure à celle de tout autre test.

Lorsque H_1 a la forme $[\theta > \theta_1]$ avec $\theta_1 \geq \theta_0$ ou $[\theta < \theta_1]$ avec $\theta_1 \leq \theta_0$, on peut démontrer que les tests utilisés dans la section précédente sont UPP. Ce sont donc ceux-là que l'on utilisera.

Lorsque H_1 est de la forme $[\theta \neq \theta_0]$, on utilise encore ces tests de la façon suivante : on construit la région de rejet de la forme $[D < d_1] \cup [D > d_2]$ de sorte que

$$\mathbb{P}\left[D < d_1\right] = \mathbb{P}\left[D > d_2\right] = \frac{\alpha}{2}.$$

Exemple : *Test bilatéral* : Soit $(X_1, ..., X_n)$ un échantillon de taille n d'une v.a. X suivant une loi normale de paramètres m et σ^2 inconnus. On veut tester les hypothèses :

$$H_0 \ : \ [m = m_0], \quad H_1 \ : \ [m \neq m_0]$$

avec un risque de première espèce α.

Sous l'hypothèse H_0,

$$T = \frac{\overline{X} - m_0}{S/\sqrt{n}} \rightsquigarrow T_{n-1}.$$

On choisit la région critique de la forme

$$A = [|T| > k]$$

où k est choisi de sorte que

$$\mathbb{P}\left[|T| > k | H_0\right] = \alpha.$$

En utilisant la symétrie de la loi de Student et en inversant sa fonction de répartition, on trouve, pour $n = 15$ et $\alpha = 0,1$ (par exemple), $k = 1,761$.
Le test est donc le suivant : on calcule t avec les données observées pour m_0 donné.
Si $|t| > 1.761$, on rejette H_0 (avec probabilité α de se tromper) ; sinon, on accepte H_0.

3.4 Tests d'ajustement.

On dispose de la réalisation d'un échantillon $(X_1, ..., X_n)$ d'une v.a.r. X et on souhaite déterminer la loi de X. La première étape consiste à "deviner" une loi possible pour X, en regardant l'histogramme des fréquences constitué par la réalisation de notre échantillon par exemple. On construit alors un test pour savoir si X suit ou non la loi que l'on a devinée, mettons $\mathcal{L}$; autrement dit, on pose

$$H_0 \ : \ [X \text{ suit la loi } \mathcal{L}],$$

$$H_1 \ : \ [X \text{ ne suit pas la loi } \mathcal{L}].$$

3.4.1 Test d'ajustement du Chi-deux.

Exemple : *Test d'ajustement du Chi-deux à une loi* $\mathcal{G}(1/2)$.

On veut vérifier l'ajustement de la "loi du premier succès" , lorsque l'on répète indéfiniment des tirages d'une pièce de monnaie, à la loi $\mathcal{G}(1/2)$.
On réalise 50 fois l'expérience, c'est-à-dire que 50 fois de suite, on lance la pièce de monnaie jusqu'à obtenir "pile" et on note le rang d'arrivée de ce "pile" .
Les 50 résultats sont : 1 1 3 2 3 1 1 1 4 3 2 7 1 2 1 1 2 4 2 1 1 1 2 2 5 2 1 1 1 1 3 2 2 1 1 1 1 1 6 1 3 1 1 3 2 1 1 2 1 4.
On représente ces résultats par un diagramme en bâtons de hauteurs 1 :26, 2 :12, 3 :6, 4 :3, 5 :1, 6 :1, 7 :1.

On suppose que ces résultats sont les réalisations d'un échantillon de taille 50 d'une v.a. X et on teste les hypothèses

$$H_0 : [X \rightsquigarrow \mathcal{G}(1/2)] \quad H_1 : \overline{H}_0.$$

Pour cela, on commence par faire une partition en classe des valeurs prises par la v.a. X. Ici, X prend ses valeurs dans $\mathbb{N}^\star$. On choisit ces classes de sorte que l'*effectif empirique des classes*, c'est-à-dire le nombre d'observations de chaque classe ne soit pas trop petit. Par exemple, ici, on choisit les classes $C_1 = 1$, $C_2 = 2$, $C_3 = 3$, $C_4 = 4$ et plus.
Les effectifs empiriques de ces classes sont donc les n_j dans le tableau suivant :

j	n_j		
1	26		
2	12		
3	6		
4	6		
totaux	n=50		

Le but du test est d'étudier la différence entre ces effectifs empiriques et les *effectifs théoriques* des classes. En effet, sous l'hypothèse H_0, X suit une loi $\mathcal{G}(1/2)$ et, sous cette hypothèse, l'effectif que devrait avoir une classe est le nombre total d'observations (la taille de l'échantillon) multiplié par la probabilité qu'une observation soit dans la classe.
Calculons, pour chaque classe, la probabilité d'être dans chacune des classes sous H_0 : pour tout $k \in \mathbb{N}^\star$,
$\mathbb{P}[X = k] = (1 - 1/2)^{k-1} 1/2 = (1/2)^k$ donc, si on note $p_j = \mathbb{P}[X \in C_j]$ pour $j = 1$ à 4, on a

$$p_1 = \frac{1}{2}, \ p_2 = \frac{1}{4}, \ p_3 = \frac{1}{8}, \ p_4 = 1 - \frac{1}{2} - \frac{1}{4} - \frac{1}{8} = \frac{1}{8}.$$

Finalement, les effectifs théoriques des classes sont donc les np_j, pour $j = 1$ à 4, que l'on peut mettre dans le tableau.
Avant de passer au test proprement dit, formalisons ce que nous venons de faire.
Soit $X(\Omega)$ l'ensemble des valeurs prises par X sous l'hypothèse H_0. On choisit une partition $C_1, ..., C_J$ de $X(\Omega)$, chaque C_j, pour $j \in \{1, ..., J\}$, étant appelé une *classe* (voir plus loin les commentaires sur le choix des classes). On définit alors les variables aléatoires $N_1, ..., N_J$, *effectifs empiriques des classes*, comme les nombres de v.a. de l'échantillon appartenant aux classes C_1,...,C_J respectivement.
On peut donc calculer ces effectifs par les formules : pour tout $j \in \{1, ..., J\}$

$$N_j = \sum_{i=1}^{n} \mathbf{1}_{C_j}(X_i)$$

ou bien de façon équivalente

$$N_j = \text{Card } \{i \in \{1, ..., n\} / \ X_i \in C_j\},$$

et les n_j du tableau sont les réalisations des N_j.

On note $p_j = \mathbb{P}[X \in C_j]$ pour tout $j \in \{1, ..., J\}$. Alors p_j est la proportion *théorique* de résultat que l'on doit trouver dans la classe j. On appelle alors *effectif théorique de la classe* C_j la quantité np_j.

Remarque : Le J-uplet $(N_1, ..., N_J)$ suit une loi multinomiale de paramètres $(n, p_1, ..., p_J)$. En particulier, chaque N_j suit une loi binomiale de paramètres (n, p_j). L'effectif théorique de la classe j est l'effectif théorique **moyen** de la classe, soit $\mathbb{E}[N_j] = np_j$.

j	n_j	np_j	.
1	26	25	
2	12	12,5	
3	6	6,25	
4	6	6,25	
totaux	n=50	n=50	

Théorème 3.1

Soit D^2 la variable aléatoire définie par

$$D^2 = \sum_{j=1}^{J} \frac{(N_j - np_j)^2}{np_j}.$$

Alors, sous l'hypothèse H_0, lorsque la taille de l'échantillon n tend vers l'infini, D^2 converge en loi vers une variable du chi-deux à $J - 1$ degrés de liberté.

" Preuve" : Par le TLC, D^2 est approximativement quand n grand une somme de J carrés de $\mathcal{N}(0, 1)$ reliées par la relation $\sum_{j=1}^{J} N_j = n$.

Test d'ajustement du chi-deux

- On admet que D^2 suit approximativement une loi du Chi-deux à $J - 1$ degrés de liberté.

— On fixe le risque de première espèce α petit.

— La région de rejet de H_0 est choisie de la forme $[D^2 > c]$ où c est à déterminer en fonction de α dans en inversant la fonction de répartition de la loi du Chi-deux à $J-1$ degrés de liberté.

— Pour appliquer le test, il suffit alors de calculer la réalisation de la variable D^2 que l'on obtient avec nos données et de constater si elle se trouve ou non dans la région de rejet de H_0.

Exemple : *(suite).* $J = 4$. D^2 suit approximativement une loi du chi-deux à 3 degrés de liberté. On se fixe un seuil de risque (par exemple $\alpha = 0.01$).

En inversant la fonction de répartition d'une loi du Chi-deux à 3 degrés de liberté, on trouve que, sous H_0, $\mathbb{P}\left[D^2 > 11,345\right] = 0.01$. On rejettera H_0 si notre réalisation de D^2 est supérieure à ce seuil. On calcule donc la réalisation de D^2,

$$d^2 = \sum_{j=1}^{4} \frac{(n_j - np_j)^2}{np_j}.$$

j	n_j	np_j	$\frac{(n_j-np_j)^2}{np_j}$
1	26	25	$1/25$
2	12	12,5	$(0.5)^2/12,5$
3	6	6,25	$(0,25)^2/6,25$
4	6	6,25	$(0,25)^2/6,26$
totaux	n=50	n=50	$d^2 = 1/12,5$

On fait le calcul à partir du tableau, et on trouve $d^2 = 1/12,5 << 11,345$, donc on conserve H_0 (i.e. la pièce utilisée est équilibrée et les lancers ont été faits indépendamment).

Remarque : L'approximation que l'on fait en supposant que D^2 suit une loi du Chi-deux à $J-1$ degrés de liberté n'est admise que si les effectifs empiriques et théoriques des classes "ne sont pas trop petits". Le seuil fixé dépend largement des auteurs de traités de statistique. Nous demanderons que les réalisations des N_j et que les np_j soient supérieurs à 5. Dans le cas contraire, on modifiera les classes en les regroupant pour obtenir ces conditions.

3.4.2 Test d'ajustement du Chi-deux avec estimation de paramètres.

Très souvent, on veut pouvoir ne spécifier, dans l'hypothèse H_0, que la loi de X et non les paramètres de cette loi, que l'on ignore a priori et que l'on ne peut qu'estimer. Le test vise donc à choisir entre les hypothèses

$$H_0 : \ [X \rightsquigarrow \mathcal{L}(t)], \quad H_1 : \overline{H}_0,$$

où t est la réalisation observée d'un estimateur T du paramètre $\theta \in \mathbb{R}^p$ de la loi $\mathcal{L}$.
On procède exactement commme ci-dessus, en utlisant cette fois-ci le théorème :

Théorème 3.2

Sous l'hypothèse H_0, lorsque la taille de l'échantillon n tend vers l'infini, D^2 converge en loi vers une variable du chi-deux à $J-1-p$ degrés de liberté (p est le nombre de paramètres estimés).

" Preuve" : D^2 est approximativement la somme de $J-1-p$ carrés de $\mathcal{N}(0,1)$ car chaque paramètre fixé donne une relation entre les variables.

3.4.3 Test d'ajustement de Kolmogorov-Smirnov.

Lorsque la loi est entièrement spécifiée, on peut faire des tests d'ajustement qui reposent sur la convergence de la fonction de répartition empirique d'un échantillon vers la fonction de répartition de la variable parente de l'échantillon. Nous décrivons ici le test de Kolmogorov-Smirnov qui s'applique dans le cas de lois dont les fonctions de répartition sont continues.

Théorème 3.3

Sous les hypothèses du théorème 1.1, la vitesse de convergence de $F_n^\star$ vers F_X est précisée par, pour tout $y > 0$,

$$\lim_{n\to+\infty} \mathbb{P}\left[\sqrt{n}\sup_{x\in\mathbb{R}}|F_n^\star(x) - F_X(x)| \leq y\right] = K(y) = \sum_{k\in\mathbb{Z}}(-1)^k e^{-2k^2y^2}.$$

Pour le *test d'ajustement de Kolmogorov-Smirnov*, on procède comme suit : on veut tester l'hypothèse selon laquelle X suit une loi $\mathcal{L}$ de fonction de répartition F. On teste toujours

$$H_0 \ : \ [X \text{ suit la loi } \mathcal{L}], \quad H_1 \ : \ \overline{H}_0.$$

On fixe un seuil de risque α petit. Sous l'hypothèse H_0, la variable aléatoire

$$\sqrt{n}D_n = \sqrt{n}\sup_{x\in\mathbb{R}}|F_n^\star(x) - F(x)|$$

admet approximativement comme fonction de répartion la fonction K. On cherche une région de rejet de la forme

$$[\sqrt{n}D_n > y]$$

telle que $\mathbb{P}[\sqrt{n}D_n > y] \approx 1 - K(y) = \alpha$, en déterminant y en fonction de α (la fonction inverse de K n'est pas donnée dans les logiciels de statistique en général, on utilise alors des tables comme celle donnée à la fin du Saporta). On procède ensuite au test en calculant la réalisation de $\sqrt{n}D_n$ obtenue avec nos données.

3.5 Tests de comparaison entre échantillons indépendants.

On dispose de m échantillons indépendants d'une certaine variable X et on désire savoir si ces échantillons proviennent d'une même population. On peut reformuler le problème de la façon suivante : chaque échantillon est une suite finie i.i.d. d'une variable parente X^k et le problème est de savoir si les X^k ont même loi.

3.5.1 Tests paramétriques pour la comparaison de deux échantillons.

Lorsque l'on a deux échantillons indépendants ($m = 2$), on peut chercher à savoir si les variables parentes ont même espérance et même variance.
On suppose que les variables parentes des deux échantillons, X^1 et X^2, admettent des moments d'ordre 2 et on note

$$m_k = \mathbb{E}\left[X^k\right] \text{ et } \sigma_k^2 = \mathbb{V}\text{ar}\left[X^k\right] \text{ pour } k = 1, 2.$$

On note $(X_1^1, ..., X_{n_1}^1)$ l'échantillon de la v.a. X^1 et $(X_1^2, ..., X_{n_2}^2)$ celui de la v.a. X^2.

Cas des échantillons gaussiens. On suppose que X^1 et X^2 suivent des lois gaussiennes. On commence par comparer les variances et, si elles ne sont pas significativement différentes, on comparera les moyennes sous l'hypothèse que les variances sont égales.

Pour comparer les variances, on procède au *test de Fisher-Snedecor* suivant : on pose

$$H_0 \ : \ [\sigma_1 = \sigma_2], \quad H_1 \ : \ [\sigma_1 \neq \sigma_2];$$

on choisit un seuil de risque α petit. Si on note S_1^2 et S_2^2 les variances empiriques (sans biais) des échantillons, alors,

$$(n_k - 1)\frac{S_k^2}{\sigma_k^2} \rightsquigarrow \chi^2_{n_k - 1} \text{ pour } k = 1, 2.$$

et donc, sous l'hypothèse H_0,

$$\frac{(n_1 - 1)S_1^2/\sigma_1^2(n_1 - 1)}{(n_2 - 1)S_2^2/\sigma_2^2(n_2 - 1)} = S_1^2/S_2^2 \rightsquigarrow F_{n_1 - 1, n_2 - 1}$$

On note alors F la variable $F = S_1^2/S_2^2$. On cherche alors la région de rejet sous la forme

$$[F < f_1] \cup [F > f_2]$$

où f_1 (resp. f_2) est le quantile d'ordre $\frac{\alpha}{2}$ (resp. $1 - \frac{\alpha}{2}$) de la loi de Fisher-Snedecor (cf. exemple en TD). Rq : f_1 pourrait se noter $f_{\frac{\alpha}{2}}$ et f_2 pourrait se noter $f_{1 - \frac{\alpha}{2}}$.
Si le test précédent n'a pas conduit à rejeter l'hypothèse $\sigma_1 = \sigma_2$, on procède au test de Student sur les moyennes comme suit : on suppose $\sigma = \sigma_1 = \sigma_2$ (inconnue) et on teste les hypothèses :

$$H_0 \ : \ [m_1 = m_2], \quad H_1 \ : \ [m_1 \neq m_2]$$

en choisissant un seuil de risque α petit.
On note $\overline{X}_k$ la moyenne empirique de l'échantillon k. Sous l'hypothèse H_0,

$$\overline{X}_1 - \overline{X}_2 \rightsquigarrow \mathcal{N}(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}) = \mathcal{N}(0, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2})$$

et donc

$$\frac{\overline{X}_1 - \overline{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \left(\frac{(n_1 - 1)S_1^2/\sigma^2 + (n_2 - 1)S_2^2/\sigma^2}{n_1 + n_2 - 2} \right)^{-1/2} \rightsquigarrow T_{n_1 + n_2 - 2}$$

soit

$$T = \frac{(\overline{X}_1 - \overline{X}_2)\sqrt{n_1 + n_2 - 2}}{\sqrt{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}\sqrt{1/n_1 + 1/n_2}} \rightsquigarrow T_{n_1 + n_2 - 2}.$$

On choisit lors la région de rejet de la forme

$$[|T| > t].$$

Remarque : Lorsque $\sigma_1 \neq \sigma_2$ et que les échantillons sont suffisamment grands (i.e. quelques dizaines d'observations), on peut encore appliquer le test de Student.

Cas des échantillons non gaussiens.
Dans ce cas, le test de Fisher-Snedecor ne peut plus s'appliquer, mais on peut encore appliquer le test de comparaison des moyennes si les échantillons sont assez grands, en remplaçant la loi de Student par la loi normale.

3.5.2 Test non paramétrique de comparaison de deux échantillons ou plus : le test du Chi-deux.

On dispose de m échantillons de v.a. X^1,...,X^m. Comme pour le test d'ajustement du Chi-deux, on partage en J classes l'ensemble des valeurs prises par ces variables aléatoires. Pour tout $k \in \{1, ..., m\}$ et pour tout $j \in \{1, ..., J\}$, on note N_{kj} le nombre de réalisations de l'échantillon k qui sont dans la classe C_j.
On pose

$$N_{.j} = \sum_{k=1}^{m} N_{kj} \text{ l'effectif empirique de la classe } j \in \{1, ..., J\},$$

$$N_{k.} = \sum_{j=1}^{J} N_{kj} = n_k \text{ la taille de l'echantillon } k \in \{1, ..., m\} \text{ et}$$

$$N = \sum_{k=1}^{m} \sum_{j=1}^{J} N_{kj} = n \text{ le nombre total d'observations.}$$

On pose enfin

$$D_0^2 = \sum_{k=1}^{m} \sum_{j=1}^{J} \frac{(N_{kj} - N_{k.}N_{.j}/N)^2}{N_{k.}N_{.j}/N}.$$

On peut montrer que, sous l'hypothèse (H_0) que les échantillons proviennent d'une même population et sont indépendants, D_0^2 suit approximativement une loi du Chi-deux à

$$(m-1)(J-1)$$

degrés de liberté. On procède alors comme dans le test d'ajustement du Chi-deux avec les hypothèses H_0 et $H_1 = \overline{H}_0$.

Exemple : On veut comparer 4 générateurs de nombres aléatoires entre 1 et 13. Pour cela, on fait 100 tirages aléatoires. Les résultats obtenus sont :

gén. 1 : 1 4 2 5 11 4 5 2 4 4 3 13 5 13 4 10 12 12 4 1 3 12 12 9 5 6 12 13 8,
gén. 2 : 7 11 13 5 2 2 4 4 1 12 13 11 6 1 9 13 12 12 12 7 5,
gén. 3 : 13 13 11 5 10 13 13 10 8 12 3 4 5 3 4 4 6 6 10 2 3 5 2 3 13 13 4 9 1,
gén. 4 : 2 1 3 3 11 10 12 8 4 7 2 11 13 1 10 1 12 12 4 2 12.

On choisit de partitionner $\{1, ..., 13\}$ en les classes

$$C_1 = \{1, 2, 3, 4, 5\},\ C_2 = \{6, 7, 8, 9\},\ C_3 = \{10, 11, 12, 13\}.$$

On dispose les résultats dans un tableau :

	C_1	C_2	C_3	Total
G_1	$16 = n_{ij}$	3	10	29$= n_{k.}$
G_2	8	4	9	21
G_3	14	4	11	29
G_4	10	2	9	21
Total	48$=n_{.j}$	13	39	100$=n$

Sous l'hypothèse H_0 selon laquelle les générateurs sont identiques et indépendants, on sait que D_0^2 suit une loi du chi-deux à $3 \times 2 = 6$ degrés de liberté et on applique le test du chi-deux habituel avec un risque de première espèce α à choisir.

3.5.3 Test d'indépendance du chi-deux.

Remarque : Le test non paramétrique de comparaison d'échantillons indépendants s'applique également pour établir l'indépendance de deux variables aléatoires.

En effet, supposons que l'on observe un échantillon

$$((X_1, Y_1), ..., (X_n, Y_n))$$

d'un couple de v.a. (X, Y). L'ensemble des valeurs prises par X est partitionné en les classes $C_i, i = 1$ à I et celui des valeurs prises par Y en les classes C^j, $j = 1$ à J.
On note $C_{ij} = C_i \times C^j$ alors

$$\{C_{ij},\ i = 1..I,\ j = 1..J\}$$

forme une partition de l'ensemble des valeurs prises par (X, Y). On note alors N_{ij} l'effectif de la classe C_{ij}.
On pose

$$H_0 : [X \text{ et } Y \text{ sont indépendantes}], \quad H_1 : \overline{H}_0.$$

Sous l'hypothèse H_0, la variable

$$D_0^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(N_{ij} - N_{i.}N_{.j}/n)^2}{N_{i.}N_{.j}/n} \rightsquigarrow \chi^2_{(I-1)(J-1)}$$

et on procède à un test du chi-deux.

Justification
En effet, cela revient à considérer que l'une des variables est le numéro d'échantillon. La similarité des échantillons correspond à l'indépendance des variables.

4 Régression linéaire

4.1 Un exemple : culture du blé au Burundi (région du Mugamba)

Les agriculteurs de la région du Burundi cultivent le blé depuis des générations. Pour augmenter les rendements, on leur propose de fertiliser la terre avec de l'engrais. Les questions qui se posent sont :

- y-t-il un impact du dépôt d'engrais sur la production ?
- si oui, peut-on prévoir le rendement espéré pour une certaine quantité d'engrais déposée ?

Les sacs d'engrais coûtent cher. Il est donc indispensable de quantifier l'augmentation attendue de rendement pour un investissement donné. Pour répondre à ces questions, une campagne de mesure a été lancée. L'engrais est mesuré en kg/ha, le rendement en t/ha.
Les données sont dans le tableau 1 que l'on représente dans la Figure 4.

Engrais (kg/ha)	Rendement (t/ha)
48	2.1610
48	2.2377
58	2.4087
54	2.3568
60	2.4948
43	2.0117
46	2.1439
43	2.1272
49	2.2224
52	2.2382
54	2.3258
43	2.0617
55	2.3694
54	2.2971
50	2.1957

Table 1 – Données de l'engrais et du rendement

On réalise un ajustement avec l'outil "Basic fitting" de Matlab, on obtient la Figure 5

Remarque : Cette information, est-elle suffisante pour répondre aux questions ?

- On peut répondre visuellement à la question : "y-a-t-il un effet de l'engrais ?". Quantitativement la pente est de 0.024 : est-ce beaucoup ou peu ? L'ajustement est-il bon ?
- On peut effectuer une prédiction pour $engrais = 45$, on obtient $0.024 * 45 + 1.042 = 2.122$. Que peut-on dire sur les incertitudes ?
- Que faire en dimension plus grande quand on ne peut plus visualiser, notamment pour répondre à la question de l'influence d'une variable ?

4.2 Le modèle linéaire et estimation des paramètres

On suppose dans la suite que l'on dispose d'un échantillon de n points $(x_i, y_i) \in \mathbb{R}^2$ et qu'on souhaite analyser la relation entre les x_i (engrais) et les y_i (rendement). Pour cela, nous allons chercher une fonction f telle que :

$$y_i \approx f(x_i)$$

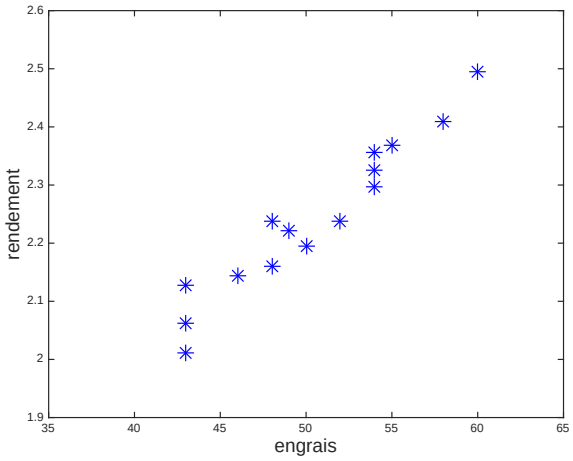

Figure 4 – Représentation du rendement (t :ha) en fonction de la quantité d'engrais (kg/ha)

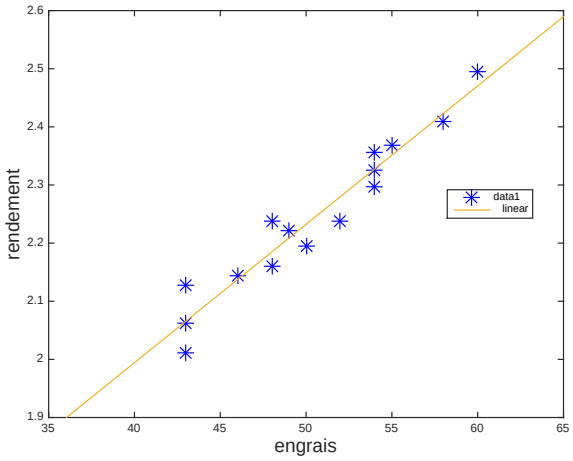

Figure 5 – Ajustement linéaire avec "Basic Fitting" de Matlab

Dans de nombreuses situations, en première approche, une idée naturelle est de supposer que la variable y est une fonction affine de la variable x. C'est le principe de la régression linéaire simple.

4.2.1 Le modèle linéaire

Définition 4.1

Le modèle de régression linéaire simple est :

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

où :

- β_0 *et* β_1 *sont des paramètres inconnus*

- x_i *est la valeur fixée de la variable* x *pour l'individu* i*, ce n'est pas une variable aléatoire (d'où l'utilisation d'une lettre minuscule), mais on l'appelle* **variable explicative**.
- ϵ_i *est une variable aléatoire dite* **erreur (ou bruit)** *pour laquelle on formule les hypothèses suivantes :*
  - *i)* $\epsilon_i \rightsquigarrow N(0, \sigma^2)$ *(*σ^2 *inconnu).*
  - *ii)* $cov(\epsilon_i, \epsilon_j) = 0$ *si* $i \neq j$
  - *iii)* $\epsilon = (\epsilon_1, ..., \epsilon_n)'$ *est un vecteur gaussien.*
- Y_i *est la* **réponse** *pour l'individu* i*. C'est une variable aléatoire, l'aléa venant de* ϵ_i*.* y_i *est une réalisation de* Y_i*.*

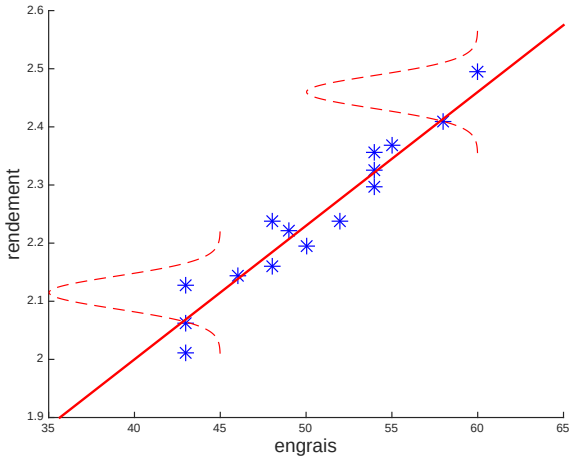

Figure 6 – Distribution de la variable réponse Y si $x = 45$ et $x = 60$.

Interprétation

- β_0 est la valeur attendue pour y en $x = 0$, c'est-à -dire sans ajout d'engrais dans le cas de l'exemple. β_1 est l'augmentation attendue de rendement quand on augmente de 1 kg/ha la quantité d'engrais.
- x_i est déterministe : lorsque l'on décide de cultiver la parcelle avec une quantité d'engrais de 50 kg/ha, 50 n'est pas aléatoire.
- Y_i est aléatoire. Si on pouvait recommencer l'expérience à 50kg/ha d'engrais, le rendement observé sera bien sûr différent de celui dont on dispose dans notre échantillon
- i $\forall i, \mathbb{E}[\epsilon_i] = 0$ signifie que les erreurs sont supposées de moyenne nulles quelque soit x_i, i.e. Y_i est centrée sur $\beta_0 + \beta_1 x_i$ pour tout i.

  ii $\mathbb{V}\text{ar}[\epsilon_i] = \sigma^2$ signifie que la dispersion de Y_i autour de $\beta_0 + \beta_1 x_i$ est supposée constante quelque soit x_i.

  ii $\epsilon_i \rightsquigarrow N(0, \sigma^2)$ signifie que Y_i de loi normale de moyenne $\beta_0 + \beta_1 x_i$ et de variance σ^2 (cf Figures 6 et 7).

  iv $cov(\epsilon_i, \epsilon_j) = 0$ signifie que les erreurs commises en deux points d'observation différents sont décorrélées (ici indépendantes) quels que soient ces points.

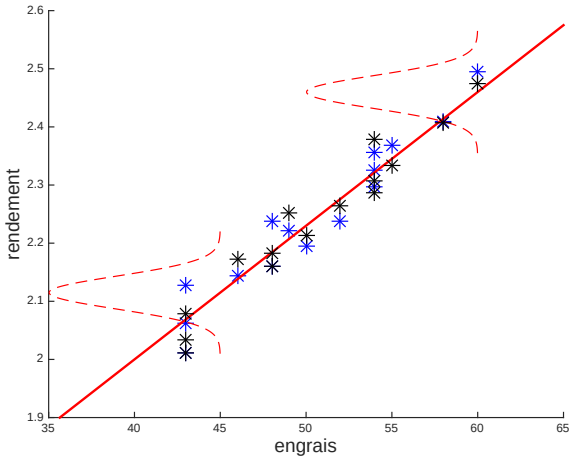

Figure 7 – Distribution de la variable réponse Y si $x = 45$ et $x = 60$. D'autres mesures ont été ajoutées

.

Remarque : Le modèle peut également s'écrire sous la forme matricielle suivante :

$$Y = X\beta + \epsilon$$

$$\text{où } Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$\text{et } \epsilon \rightsquigarrow \mathcal{N}(0, \sigma^2 I_n)$$

Remarque : On remarque que le modèle linéaire s'écrit comme la somme de deux termes. Un terme déterministe et un terme aléatoire. La partie déterministe s'écrit comme une **combinaison linéaire** de fonctions de base (ici 1 et x) d'où le nom "modèle linéaire". L'appellation ne vient pas du fait qu'on soit linéaire en x mais linéaire en les paramètres β_0, β_1. La partie aléatoire (bruit ou erreur) représente un bruit de mesure et des erreurs dues à des aléas incontrôlables : type de sol, conditions climatiques, qualité de la semence. Cette partie aléatoire ne doit pas représenter des erreurs de modélisation de la dépendance entre y_i et x_i . Si tel est le cas, des outils de diagnostic existent et permettent de corriger les hypothèses de modélisation, par exemple en complexifiant la relation $y_i \approx \beta_0 + \beta_1 x_i + \beta_2 x_i^2$. Avec cette dernière expression on reste dans le cadre du modèle linéaire.

4.2.2 Loi des estimateurs des paramètres

Le modèle étant postulé il faut estimer les paramètres β_0 et β_1, ou encore : que peut-on proposer comme estimation pour les paramètres de la droite ?

Notons $B = \begin{pmatrix} B_0 \\ B_1 \end{pmatrix}$ le vecteur des estimateurs des paramètres $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ et $b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$ le vecteur des réalisations.
On souhaite construire B de sorte que pour tout $1 \leq i \leq n$, $B_0 + B_1 x_i$ soit le plus proche possible de Y_i. On a plusieurs possibilités. Par exemple,

- on peut minimiser les $|Y_i - (B_0 + B_1 x_i)|$
- on peut les minimiser $(Y_i - (B_0 + B_1 x_i))^2$

Idée : On cherche b_0 et b_1 qui minimisent

$$S(b_0, b_1) = \Sigma_{i=1}^{n} \left(y_i - b_0 - b_1 x_i\right)^2 = \|y - Xb\|_2^2 = S(b)$$

où $\|\cdot\|_2$ désigne la norme euclidienne sur $\mathbb{R}^n$.

Définition 4.2

On appelle estimation des Moindres Carrés, les valeurs b_0 et b_1 minimisant la quantité :

$$S(b_0, b_1) = \Sigma_{i=1}^{n} \left(y_i - b_0 - b_1 x_i\right)^2$$

Proposition 4.1

La solution du problème précédent est

$$b = (X'X)^{-1} X'y$$

ou encore :

$$b_0 = \overline{y} - b_1 \overline{x} \quad b_1 = \frac{\Sigma_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\Sigma_{i=1}^{n}(x_i - \overline{x})^2}$$

Remarque : On peut noter que

- les expressions pour b_0 et b_1 sont linéaires en les observations y_i (conséquence du modèle linéaire et estimation par Moindres Carrés). Pour chaque nouvel échantillon, les estimations changent (cf Figure 8)
- l'expression de b_0 assure que la droite passe par le centre de gravité $(\overline{x}, \overline{y})$ du nuage de points (x_i, y_i).
- l'expression de b_1 peut aussi s'écrire $\dfrac{cov(x, y)}{var(x)}$

Démonstration : Comme S est une fonction convexe, il suffit de calculer matriciellement sa différentielle et de l'annuler pour trouver son point de minimum. C'est la solution b de $X'Xb = X'y$ encore appelées équations normales.

Définition 4.3

L'estimateur des Moindres Carrés est le vecteur B défini par

$$B = (X'X)^{-1} X'Y$$

Proposition 4.2

$$B \rightsquigarrow \mathcal{N}\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right)$$

Interprétation

- On remarque que la loi de l'estimateur des paramètres obtenus par Moindres Carrés est entièrement connue.
- B est un estimateur non biaisé de β.
- La variance de B ne dépend que de X (c'est à dire des x_i) et est proportionnelle à la variance de l'erreur. Plus le bruit est important, plus l'estimation de β sera imprécise.

Démonstration : B est un vecteur gaussien car Y est un vecteur gaussien ($Y \rightsquigarrow \mathcal{N}(X\beta, \sigma^2 I_n)$). De plus B est un estimateur sans biais de β car :

$$\mathbb{E}\left[B\right] = \mathbb{E}\left[(X'X)^{-1}X'Y\right] = (X'X)^{-1}X'\mathbb{E}\left[Y\right] = (X'X)^{-1}X'X\beta = \beta$$

Enfin

$$\mathbb{V}\mathrm{ar}\left[B\right] = \mathbb{V}\mathrm{ar}\left[(X'X)^{-1}X'Y\right] = (X'X)^{-1}X'\mathbb{V}\mathrm{ar}\left[Y\right]X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

La figure 8) montrent les estimations obtenues pour b_0 et b_1 pour trois échantillons différents.

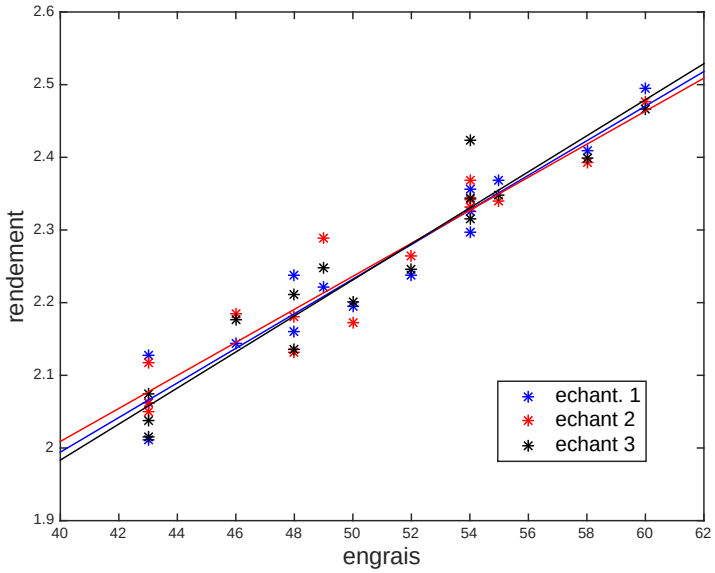

Figure 8 – Ajustement d'un modèle linéaire sur trois échantillons

Définition 4.4

On note $\hat{Y}_i = B_0 + B_1 x_i$ *la prédiction du modèle au point* x_i*. On appelle résidu* E_i *pour l'individu* i

l'écart entre la valeur prédite et la valeur observée :

$$E_i = Y_i - \hat{Y}_i$$

Il s'agit d'une variable aléatoire dont on possède une réalisation notée $e_i = y_i - b_0 + b_1 x_i$ sur notre échantillon.

Remarque : On remarque que le vecteur des prédictions s'écrit sous la forme matricielle $\hat{Y} = XB$.En remplaçant B par son expression on obtient $\hat{Y} = X(X'X)^{-1}X'Y$ qui est linéaire en le vecteur des observations Y.

Définition 4.5

L'estimateur

$$\hat{\Sigma}^2 = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n-2}$$

est un estimateur non biaisé de la variance de l'erreur σ^2. On note $\hat{\sigma}^2$ la réalisation de $\hat{\Sigma}^2$ sur notre échantillon.

En appliquant les formules aux données présentées en introduction on obtient les valeurs suivantes :

- $b_0 = 1.042$: rendement auquel on peut s'attendre sans engrais
- $b_1 = 0.024$: augmentation du rendement par kg/ha supplémentaire d'engrais
- $\hat{\sigma} = 0.0354$: écart-type résiduel

4.3 Qualité de l'ajustement et test de signification d'un coefficient

4.3.1 Le coefficient de détermination $\mathcal{R}^2$

Une fois la droite de régression estimée, on se pose la question de la qualité de l'ajustement. Existe-t-il un **indicateur** permettant de faire la distinction entre les deux situations présentées sur la figure suivante ? On note sur la figure 9 de gauche un très bon alignement des points sur la droite alors que sur la figure 9 de droite l'ajustement est moins bon, la réponse Y semble dépendre moins nettement de la variable explicative X.
De plus, on voudrait pouvoir **tester** la pertinence de la régression ou, lorsqu'il y a plusieurs variables explicatives, la pertinence de chacune des variables explicatives dans la régression.

Proposition 4.3

Dans le contexte du modèle linéaire on a la décomposition de la variance :

$$\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

avec $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

On note cette égalité

$$SST = SSR + SSE$$

où

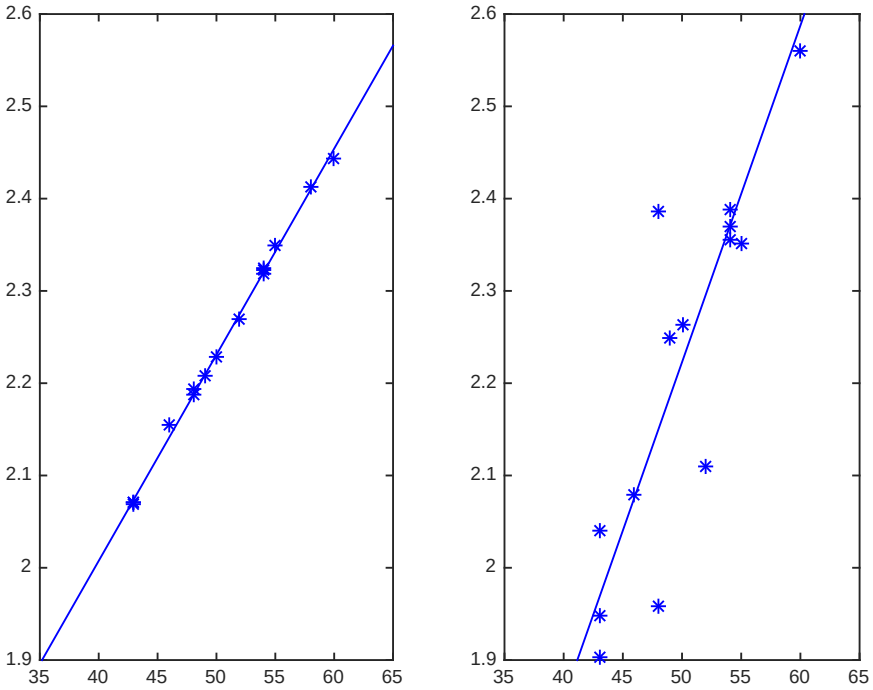

Figure 9 – Ajustements linéaires pour deux niveaux de bruit

- $SST = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$ est la somme des carrés totaux, proportionnelle à la variance des observations
- $SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2$ est la somme des carrés des prédictions de la régression, proportionnelle à la variance des prédictions
- $SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$ est la somme des carrés des résidus (estimations des erreurs).

Interprétation La variance totale se décompose en une partie expliquée par le modèle (ici grâce à l'information apportée par les x_i) et une partie résiduelle que l'on n'a pas réussi à expliquer avec les x_i.

Démonstration : Il suffit d'appliquer Pythagore en ayant observé que $\hat{Y}$ est la projection orthogonale de Y sur le sous-espace engendré par les colonnes de X (cf cours 3A).

Définition 4.6

Le coefficient de détermination $\mathcal{R}^2$ est défini par :

$$\mathcal{R}^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}$$

Interprétation Du fait de la décomposition de la variance, le coefficient $\mathcal{R}^2$ est compris entre 0 et 1. Si $\mathcal{R}^2 = 1$ il existe une relation déterministe (affine ici) entre y_i et x_i. Si au contraire $\mathcal{R}^2 = 0$, les observations y_i ne s'expliquent pas du tout de façon affine en fonction des x_i. Plus précisément, R^2 mesure la qualité de l'ajustement du modèle aux données par la part de variance des données due au modèle. Dans le cas de l'exemple, on trouve $R^2 = 0.934$.

4.3.2 Test de signification d'un coefficient

Ici on s'intéresse à tester la pertinence d'une variable dans le modèle de régression. Si on reprend l'exemple introductif, à partir des données, peut-on considérer que l'augmentation de la dose d'engrais se traduit par une augmentation du rendement ?
Si cela n'était pas le cas, on aurait $\beta_1 = 0$. Sachant que l'on n'observe jamais exactement $b_1 = 0$, on peut se demander si l'écart à 0 est le fruit du hasard ou d'une réelle influence de x sur y. Autrement dit l'estimation de β_1 obtenue est-elle significative ? On répond à cette question à l'aide d'un **test de signification du coefficient** β_1.

Construction du test au seuil de risque α :

. Choix des hypothèses : $H_0 : [\beta_1 = 0]$ contre $H_1 : [\beta_1 \neq 0]$

. Choix de l'estimateur :

$$B_1 \rightsquigarrow \mathcal{N}\left(\beta_1, \sigma^2 (X'X)^{-1}_{22}\right)$$

Attention : cette loi est-elle entièrement spécifiée sous H_0 ? non car elle dépend de σ^2. On va donc utiliser la statistique suivante :

$$T = \frac{B_1}{\sqrt{\hat{\Sigma}^2 (X'X)^{-1}_{22}}} \rightsquigarrow T_{n-2} \text{ sous } H_0$$

. Forme de la région de rejet : $W =]-\infty, -k] \cup [k, +\infty[$

. k est le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $n - 2$ ddl.

. Règle de décision. Si $|\frac{b_1}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{22}}}| > k$ alors on rejette H_0 et on conclut que l'estimation de β_1 est significative et donc que la variable est influente. Sinon, on conserve H_0, i.e. compte-tenu du bruit dans les observations, la pente observée ne peut être différenciée de 0.

Remarque : La plupart des logiciels renvoie une pvaleur, c'est-à -dire $p\_value = P(|T| > |T_{obs}||H_0)$ en notant $T_{obs} = \frac{b_1}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{22}}}$. Si $p\_value < 5\%$ on rejette H_0 et on conclut que la pente estimée est significative.

4.3.3 Application : retour sur l'exemple

```
>> DMobj = readtable('donnees.txt')

DMobj =
    engrais    rendement
    -------    ---------

    48          2.161
```

```
    48         2.2377
    58         2.4087
    54         2.3568
    60         2.4948
    43         2.0117
    46         2.1439
    43         2.1272
    49         2.2224
    52         2.2382
    54         2.3258
    43         2.0617
    55         2.3694
    54         2.2971
    50         2.1957


>> model = 'rendement~engrais'

model =
rendement~engrais


>> mdl = fitlm(DMobj,model)


mdl = Linear regression model: rendement ~ 1 + engrais

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   --------    ---------    ------    ----------

    (Intercept)      1.0419     0.088778    11.737     2.726e-08
    engrais        0.023808    0.0017498    13.607    4.5675e-09

Number of observations: 15, Error degrees of freedom: 13
Root Mean Squared Error: 0.0354
R-squared: 0.934,  Adjusted R-Squared 0.929
F-statistic vs. constant model: 185, p-value = 4.57e-09
```

4.4 Prédiction

L'un des intérêts pratiques des modèles est la prédiction. A quelle valeur de y peut-on s'attendre pour une valeur de x donnée, que nous noterons x_0 ?
Exemples :

- Si un agriculteur peut se permettre un investissement de x_0 kg/ha d'engrais, quel rendement moyen peut-il espérer de ses parcelles cultivées ?
- Si j'étudie le lien entre la température à midi et le pic de pollution d'ozone (en ppm) le lendemain, quel pic de pollution moyen puis-je attendre si j'observe une température de x_0°C ?

Lorsqu'on utilise un modèle de régression, on suppose, pour calculer une prédiction, que la réponse est donnée par

$$Y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0 \quad \text{où} \quad \begin{pmatrix} \epsilon_0 \\ \epsilon \end{pmatrix} \rightsquigarrow \mathcal{N}(0, \sigma^2 I_{n+1}).$$

Remarque : On note que la prédiction attendue est une moyenne : on ne peut empêcher une variabilité individuelle des parcelles de terre, ou des conditions météorologique (vent etc..). Autrement dit il faut prévoir la moyenne (= l'espérance) de Y en x_0 soit :

$$\mathbb{E}[Y_0] = \beta_0 + \beta_1 x_0$$

Le problème est en fait triple :

1. Que peut-on donner comme estimation pour $\beta_0 + \beta_1 x_0$?
2. Peut-on quantifier l'incertitude associée à cette estimation ?
3. Quelle est alors l'erreur de prédiction commise ? Peut-on donner un intervalle de prédiction ?

4.4.1 Estimation de $\mathbb{E}[Y_0]$

Rappel : $\mathbb{E}[Y_0] = \beta_0 + \beta_1 x_0$. L'estimation naturelle est donc : $\hat{y}_0 = b_0 + b_1 x_0$? Quelle est la précision de cette estimation ? Nous cherchons donc à construire un intervalle de confiance pour $\mathbb{E}[Y_0]$. L'estimateur associé est le suivant : $B_0 + B_1 x_0$ que nous noterons vectoriellement $\begin{pmatrix} 1 & x_0 \end{pmatrix} \begin{pmatrix} B_0 \\ B_1 \end{pmatrix}$.

Quelle est la loi de cet estimateur ?

B étant un vecteur gaussien, $\begin{pmatrix} 1 & x_0 \end{pmatrix} \begin{pmatrix} B_0 \\ B_1 \end{pmatrix} \rightsquigarrow N\left(\begin{pmatrix} 1 & x_0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & x_0 \end{pmatrix} (X'X)^{-1} \begin{pmatrix} 1 \\ x_0 \end{pmatrix} \right)$.

σ^2 étant inconnu on utilise la loi de student et on obtient :

$$\frac{\begin{pmatrix} 1 & x_0 \end{pmatrix} \begin{pmatrix} B_0 \\ B_1 \end{pmatrix} - \begin{pmatrix} 1 & x_0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}{\sqrt{\hat{\Sigma}^2 \begin{pmatrix} 1 & x_0 \end{pmatrix} (X'X)^{-1} \begin{pmatrix} 1 \\ x_0 \end{pmatrix}}} \rightsquigarrow T_{n-2}$$

Proposition 4.4

Un intervalle de confiance de niveau de confiance $1-\alpha$ pour $\mathbb{E}[Y_0]$ est donné par :

$$(1\ x_0) \begin{pmatrix} B_0 \\ B_1 \end{pmatrix} \pm t_{1-\alpha/2} \sqrt{\hat{\Sigma}^2 (1\ x_0) (X'X)^{-1} \begin{pmatrix} 1 \\ x_0 \end{pmatrix}}$$

où $t_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-2$ degrés de liberté.

4.4.2 Intervalle de prédiction pour Y_0

Attention, l'intervalle de la proposition 4.4 représente l'incertitude sur la moyenne et n'est pas un encadrement des valeurs individuelles. Ce qui est intéressant pour l'utilisateur, c'est de donner une plage de variation pour **une** observation au point x_0. On souhaite donc donner un intervalle de prédiction pour Y_0.

Définition 4.7

Soit X une variable aléatoire de loi μ. Soit $\alpha \in]0,1[$. On appelle **intervalle de prédiction** *pour X de probabilité $1-\alpha$ l'intervalle $[q_{\alpha/2}; q_{1-\alpha/2}]$ où $q_{\alpha/2}$ et $q_{1-\alpha/2}$ sont les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de μ.*

Proposition 4.5

On a

$$\mathbb{P}\left[q_{\alpha/2} \leq X \leq q_{1-\alpha/2}\right] \geq 1-\alpha.$$

Démonstration : En effet, pour des variables discrètes, on définit l'inverse généralisé de la fonction de répartition par

$$F^{-1}(\alpha) = \min\{x \in \mathbb{R} / \ F(x) \geq \alpha\}.$$

On vérifie qu'avec cette définition on a bien l'inégalité ci-dessus. □

Proposition 4.6

$$Y_0 - \hat{Y}_0 \rightsquigarrow N\left(0, \sigma^2\left(1 + (1\ x_0)\,(X'X)^{-1}\begin{pmatrix}1\\ x_0\end{pmatrix}\right)\right),$$

$$\frac{Y_0 - \hat{Y}_0}{\sqrt{\hat{\Sigma}^2\left(1 + (1\ x_0)\,(X'X)^{-1}\begin{pmatrix}1\\ x_0\end{pmatrix}\right)}} \rightsquigarrow T_{n-2}$$

et $[-t_{1-\alpha/2}; t_{1-\alpha/2}]$ *est un intervalle de prédiction pour*

$$\frac{Y_0 - \hat{Y}_0}{\sqrt{\hat{\Sigma}^2\left(1 + (1\ x_0)\,(X'X)^{-1}\begin{pmatrix}1\\ x_0\end{pmatrix}\right)}}$$

de probabilité $1-\alpha$ *où* $t_{1-\alpha/2}$ *est le quantile d'ordre* $1-\alpha/2$ *de la loi de Student à* $n-2$ *degrés de liberté.*

Proposition 4.7

Un intervalle de prediction de probabilité $1-\alpha$ *pour* Y_0 *est donné par les bornes*

$$(1\ x_0)\begin{pmatrix}b_0\\ b_1\end{pmatrix} \pm t_{1-\alpha/2}\sqrt{\hat{\sigma}^2\left(1 + (1\ x_0)\,(X'X)^{-1}\begin{pmatrix}1\\ x_0\end{pmatrix}\right)}$$

où $t_{1-\alpha/2}$ *est le quantile d'ordre* $1-\alpha/2$ *de la loi de Student à* $n-2$ *degrés de liberté.*

Remarque : On remarque que à l'incertitude sur $\beta_0+\beta_1 x_0$ s'ajoute la variabilité sur ϵ_0 (rappel : $Y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0$).

Code Matlab Le code Matlab suivant permet de tracer sur la Figure 10 les données, l'ajustement linéaire, les intervalles de confiance à 95% pour $\mathbb{E}[Y_0]$ et les intervalles de prédiction à 95% pour Y_0.

```
>> DMobj = readtable('donnees.txt')
>> plot(DMobj.engrais,DMobj.rendement,'b*','MarkerSize',12);
>> xlabel('engrais','fontsize',16);
>> ylabel('rendement','fontsize',16);axis([40,65,1.9,2.6])
>> model = 'rendement~engrais';
>> mdl = fitlm(DMobj,model);
```

```
>> xnew = table((35:65)','VariableNames',{'engrais'});
>> [ypred,yci1] = predict(mdl,xnew,'prediction','curve');
>> hold on;plot((35:65),ypred,'b');
>> plot((35:65),yci1(:,1),'--b',(35:65),yci1(:,2),'--b');
>> [ypred,yci2] = predict(mdl,xnew,'prediction','observation');
>> plot((35:65),yci2(:,1),'.b',(35:65),yci2(:,2),'.b');
```

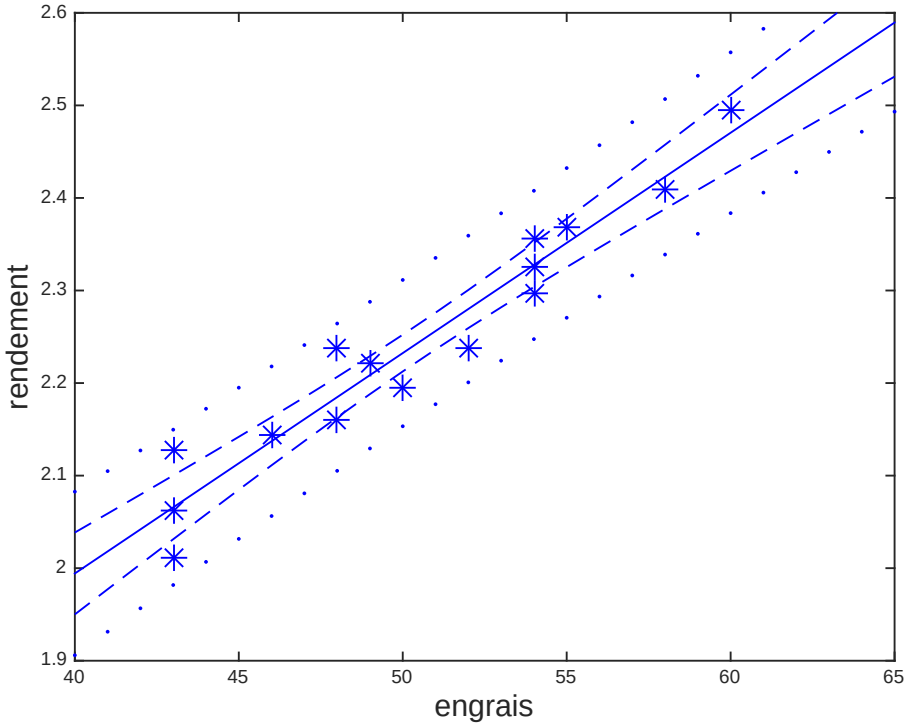

Figure 10 – Données brutes (*)-Ajustement linéaires (-)-Intervalle de confiance (–) à 95% et intervalles de prédiction (..) à 95%

4.5 Extensions

Nous avons présenté quelques éléments incontournables de la régression linéaire. Cependant plusieurs points n'ont pas été abordés :

- Les hypothèses probabilistes portant sur les erreurs sont à valider : la loi normale, la moyenne et la variance constante, l'indépendance des erreurs. Des outils existent et permettent de valider l'ensemble de ces hypothèses.
- La table d'analyse de la variance, en sortie des logiciels, comporte plusieurs informations numériques que nous n'avons pas introduites.
- Nous avons présenté la régression linéaire simple, c'est-à -dire avec une seule variable explicative. Cependant les expressions matricielles introduites permettent facilement de généraliser à une matrice X à plus de 2 colonnes et un vecteur β à plus de 2 paramètres. Pour

l'exemple en introduction, la variable densité du semis pourrait être étudiée, voire des propriétés de qualité du sol, l'exposition des parcelles etc. En dimension supérieure, le problème est la sélection des variables pertinentes pour maintenir de bonnes qualités prédictives du modèle. Le test de Student présenté en section 4.3.2 reste pertinent en l'absence de colinéarité des variables d'entrée. Dans le cas contraire d'autres outils doivent être introduits.

Appendice

A Formulaire : Intervalles de Confiance

Nous donnons ici les intervalles de confiance pour la moyenne et la variance.
On suppose que $(X_1, ..., X_n)$ est un échantillon d'une v.a. X.

A.1 IC sur la moyenne et la variance d'un échantillon gaussien.

On suppose que X suit une loi normale $\mathcal{N}(m, \sigma^2)$.

IC sur la moyenne avec variance connue. On utilise l'estimateur $\overline{X}$. On sait que

$$X^\star = \frac{\overline{X} - m}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0, 1).$$

Pour tout $\beta \in]0, 1[$, on note u_β le quantile d'ordre β de la loi normale centrée réduite. On a alors

$$\mathbb{P}\left[u_{\alpha/2} < \frac{\overline{X} - m}{\sigma/\sqrt{n}} < u_{1-\alpha/2}\right] = \phi(u_{1-\alpha/2}) - \phi(u_{\alpha/2}) = 1 - \alpha,$$

on choisira l'intervalle de confiance

$$\left]\overline{X} - \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}\right[$$

puisque $u_{1-\alpha/2} = -u_{\alpha/2}$.

IC sur la moyenne avec variance inconnue. On sait que la variable

$$\frac{\overline{X} - m}{S/\sqrt{n}} \rightsquigarrow T_{n-1}.$$

On note à nouveau, pour tout $\beta \in]0, 1[$, t_β le quantile d'ordre β de la loi de Student à $n-1$ degrés de liberté. Alors, comme la densité d'une loi de Student est symétrique, $t_{1-\alpha/2} = -t_{\alpha/2}$ et donc

$$\mathbb{P}\left[-t_{1-\alpha/2} < \frac{\overline{X} - m}{S/\sqrt{n}} < t_{1-\alpha/2}\right] = 1 - \alpha.$$

L'intervalle

$$\left]\overline{X} - \frac{S}{\sqrt{n}}t_{1-\alpha/2}, \overline{X} + \frac{S}{\sqrt{n}}t_{1-\alpha/2}\right[$$

est un intervalle de confiance pour m au seuil de risque α.

IC sur la variance avec moyenne connue. On utilise l'estimateur

$$T = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2.$$

On sait que $nT/\sigma^2 \rightsquigarrow \chi^2_n$. Si on note x_β le quantile d'ordre β de la loi du chi-deux à n degrés de liberté pour tout $\beta \in]0,1[$, alors

$$\mathbb{P}\left[x_{\alpha/2} < n\frac{T}{\sigma^2} < x_{1-\alpha/2}\right] = 1 - \alpha$$

et donc un intervalle de confiance pour σ^2 au seuil de risque α est

$$\left]\frac{nT}{x_{1-\alpha/2}}, \frac{nT}{x_{\alpha/2}}\right[.$$

IC sur la variance avec moyenne inconnue. On utilise l'estimateur S^2. On a

$$(n-1)\frac{S^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$$

et on procède comme ci-dessus : on obtient un intervalle de confiance au seuil de risque α

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

A.2 IC pour le paramètre d'un échantillon d'une loi de Bernoulli.

On suppose que X suit une loi de Bernoulli $\mathcal{B}(p)$ et que l'échantillon est grand.

$$\text{On sait que , } \frac{\overline{X} - p}{\sqrt{p(1-p)}/\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1) \text{ approximativement.}$$

Pour tout $\beta \in]0,1[$, on note u_β le quantile d'ordre β pour la loi normale centrée réduite. Un intervalle de confiance au seuil de risque α pour le paramètre p est

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

A.3 IC pour la moyenne pour d'un échantillon non gaussien de carré intégrable.

Par le théorème limite central, un intervalle de confiance pour la moyenne peut-être obtenu si n est assez grand. L'intervalle de confiance au seuil de risque α a la forme suivante :

$$\left]\overline{X} - \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}\right[$$

où $u_{1-\alpha}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite. Dans la pratique si la variance est inconnue, σ est remplacé par S dans l'expression de l'IC.
On n'utilisera pas d'intervalle de confiance pour la variance dans le cadre non gaussien.

B Formulaire : Tests statistiques

Nous donnons ici les *variables de décision* utilisées dans les tests statistiques usuels. Si la variable de décision est D et d sa réalisation, la région de rejet de H_0 est à choisir sous la forme $W_D = \{d > c\}$, $W_D = \{d \leq c\}$ ou $W_D = \{|d| > c\}$ en fonction de la forme de H_1, la constante c étant déterminée par le risque de première espèce α.

B.1 Tests paramétriques pour un échantillon.

On suppose que $(X_1, ..., X_n)$ est un échantillon d'une v.a. X.

B.1.1 Tests sur la moyenne et la variance d'un échantillon gaussien.

On suppose que X suit une loi normale $\mathcal{N}(m, \sigma^2)$.

Tests sur la moyenne avec variance connue.

$$H_0 \ : \ [m = m_0]$$

$$H_1 \ : \ [m = m_1] \text{ ou } [m < m_0] \text{ ou } [m > m_0] \text{ ou } [m \neq m_0].$$

$$\text{Sous } H_0, \ \frac{\overline{X} - m_0}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0, 1).$$

Tests sur la moyenne avec variance inconnue. Sous les mêmes hypothèses que ci-dessus,

$$\text{Sous } H_0, \ \frac{\overline{X} - m_0}{S/\sqrt{n}} \rightsquigarrow T_{n-1}.$$

Tests sur la variance avec moyenne connue.

$$H_0 \ : \ [\sigma = \sigma_0]$$

$$H_1 \ : \ [\sigma = \sigma_1] \text{ ou } [\sigma < \sigma_0] \text{ ou } [\sigma > \sigma_0] \text{ ou } [\sigma \neq \sigma_0].$$

$$\text{Sous } H_0, \ \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (X_i - m)^2 \rightsquigarrow \chi^2_n.$$

Tests sur la variance avec moyenne inconnue. Sous les mêmes hypothèses que ci-dessus,

$$\text{Sous } H_0, \ (n-1)\frac{S^2}{\sigma_0^2} \rightsquigarrow \chi^2_{n-1}.$$

B.1.2 Test sur le paramètre d'un échantillon d'une loi de Bernoulli.

On suppose que X suit une loi de Bernoulli $\mathcal{B}(p)$ et que l'échantillon est grand.

$$H_0 \ : \ [p = p_0]$$

$$H_1 \ : \ [p = p_1] \text{ ou } [p < p_0] \text{ ou } [p > p_0] \text{ ou } [p \neq p_0].$$

$$\text{Sous } H_0, \ \frac{\overline{X} - p_0}{\sqrt{p_0(1-p_0)}/\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1) \text{ approximativement.}$$

B.1.3 Tests sur la moyenne d'un échantillon quelconque.

On suppose que X est une variable de carré intégrable, que n est grand. Alors on peut utiliser les tests de moyenne, en utilisant la loi normale à la place de la loi de Student.

B.2 Tests paramétriques de comparaison d'échantillons indépendants.

On suppose que $(X_1^1, ..., X_{n_1}^1)$ et $(X_1^2, ..., X_{n_2}^2)$ sont des échantillons indépendants de v.a. X^1 et X^2 respectivement.

B.2.1 Cas des échantillons gaussiens.

On suppose que $X^1 \rightsquigarrow \mathcal{N}(m_1, \sigma_1^2)$ et $X^2 \rightsquigarrow \mathcal{N}(m_2, \sigma_2^2)$.

Test d'égalité des variances.

$$H_0 \ : \ [\sigma_1 = \sigma_2], \quad H_1 \ : [\sigma_1 \neq \sigma_2].$$

$$\text{Sous } H_0, \ \frac{S_1^2}{S_2^2} \rightsquigarrow F_{n_1-1, n_2-1}.$$

Test d'égalité des moyennes si les variances sont connues.

$$H_0 \ : \ [m_1 = m_2], \quad H_1 \ : \ [m_1 \neq m_2].$$

$$\text{Sous } H_0, \ \frac{\overline{X}^1 - \overline{X}^2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \rightsquigarrow \mathcal{N}(0,1).$$

Test d'égalité des moyennes si les variances sont inconnues mais testées égales. Sous les mêmes hypothèses que ci-dessus,

$$\text{Sous } H_0,\ T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{(n_1-1)S_1^2 + (n_2-1)S_2^2}} \frac{\sqrt{n_1+n_2-2}}{\sqrt{1/n_1+1/n_2}} \rightsquigarrow T_{n_1+n_2-2}.$$

Nota Bene : On choisit dans ces derniers tests une région de rejet bilatérale.

B.2.2 Cas des échantillons de Bernoulli.

On suppose que $X^1 \rightsquigarrow \mathcal{B}(p_1)$ et $X^2 \rightsquigarrow \mathcal{B}(p_2)$. On suppose également que les échantillons sont de grande taille et on pose

$$\hat{p} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1+n_2}.$$

$$H_0 \ : \ [p_1 = p_2], \quad H_1 \ : \ [p_1 \neq p_2].$$

$$\text{Sous } H_0,\ \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{1/n_1+1/n_2}} \rightsquigarrow \mathcal{N}(0,1) \text{ approximativement.}$$

B.2.3 Cas des échantillons quelconques.

On suppose que X^1 et X^2 sont des variables de carré intégrable. Si les échantillons sont assez grands, on peut encore utiliser le test 2.1.2, en remplaçant la loi de Student par la loi normale.

B.3 Tests non paramétriques.

B.3.1 Tests d'ajustement d'un échantillon à une loi donnée.

On suppose que $(X_1, ..., X_n)$ est un échantillon d'une v.a. X.

Loi entièrement déterminée.

$$H_0 \ : \ [X \rightsquigarrow \mathcal{L}], \quad H_1 \ : \ \overline{H}_0.$$

Test d'ajustement du Chi-deux. Soit Y une v.a. de loi $\mathcal{L}$. On choisit une partition finie C_1,...,C_J de l'ensemble des valeurs prises par Y. Pour tout $j \in \{1, ..., J\}$, on note

$$p_j = \mathbb{P}\left[Y \in C_j\right] \text{ et } N_j = \text{Card}\left\{i \in \{1, ..., n\} / \ X_i \in C_j\right\}$$

$$\text{Sous } H_0,\ D^2 = \sum_{j=1}^{J} \frac{(N_j - np_j)^2}{np_j} \rightsquigarrow \chi^2_{J-1} \text{ approximativement.}$$

Test de Kolmogorov-Smirnov. Soit F la fonction de répartition d'une v.a. Y suivant la loi $\mathcal{L}$. On suppose que F est continue. Pour tout $x \in \mathbb{R}$, on note

$$F_n^\star(x) = \frac{1}{n}\text{Card}\{i \in \{1,...,n\}/\ X_i \leq x\}$$

la fonction de répartition empirique de l'échantillon. On pose

$$D_n = \sup_{x\in\mathbb{R}} |F_n^\star(x) - F(x)|.$$

$$\text{Sous } H_0,\ \mathbb{P}\left[\sqrt{n}D_n \leq y\right] \approx K(y) \text{ pour tout } y > 0$$

où la fonction K est tabulée. On choisit la région de rejet sous la forme $W = \{d_n > c\}$ où d_n réalisation de D_n.

Ajustement avec paramètres à estimer : test d'ajustement du chi-deux avec paramètres. On suppose que $\mathcal{L}$ dépend de paramètres θ_1,...,θ_p. On note T_1,...,T_p des estimateurs de ces paramètres et t_1,...,t_p les réalisations observées de ces paramètres sur l'échantillon.

$$H_0\ :\ [X \rightsquigarrow \mathcal{L}(t_1,...,t_p)], \quad H_1\ :\ \overline{H}_0.$$

Avec les mêmes notations que pour le test d'ajustement du chi-deux sans paramètre à estimer,

$$\text{Sous } H_0,\ D^2 \rightsquigarrow \chi^2_{J-1-p} \text{ approximativement.}$$

B.3.2 Test de provenance d'échantillons d'une même population : le test du chi-deux.

Soient $(X_1^1,...,X_{n_1}^1)$,...,$(X_1^m,...,X_{n_m}^m)$ m échantillons indépendants de v.a. X^1,...,X^m.

$$H_0\ :\ [\text{les échantillons proviennent de la même population}], \quad H_1\ :\ \overline{H}_0.$$

On note

$$N_{kj} = \text{Card}\{i \in \{1,...,n_k\}/\ X_i^k \in C_j\},$$

$$N_{k\cdot} = \sum_{j=1}^{J} N_{kj}, \quad N_{\cdot j} = \sum_{k=1}^{m} N_{kj}, \quad N = \sum_{k=1}^{m}\sum_{j=1}^{J} N_{kj}.$$

$$\text{Sous } H_0,\ D_0^2 = \sum_{k=1}^{m}\sum_{j=1}^{J} \frac{(N_{kj} - N_{k\cdot}N_{\cdot j}/N)^2}{N_{k\cdot}N_{\cdot j}/N} \rightsquigarrow \chi^2_{(J-1)(m-1)}.$$

B.3.3 Test d'indépendance : le test du chi-deux.

On applique le test précédent pour tester l'indépendance de deux v.a. X et Y à partir d'un échantillon $((X_1,Y_1),...,(X_n,Y_n))$ de (X,Y). On note alors $\{C_i,\ i \in \{1,...,I\}\}$ et $\{C^j,\ j \in \{1,...,J\}\}$ des partitions des valeurs prises par X et Y respectivement, et

$$N_{ij} = \text{Card}\{l \in \{1,...,n\}/\ X_l \in C_i \text{ et } Y_l \in C^j\}.$$

$$H_0\ :\ [X \text{ et } Y \text{ indépendantes}], \quad H_1\ :\ \overline{H}_0.$$

On procède comme ci-dessus.

C Enoncés des TD

Statistique - Séance de TD 1

Dans cette séance, nous allons étudier les propriétés de quelques estimateurs et procéder à des quantification de l'incertitude par intervalle de confiance.

Exercices à préparer : 1 à 4

Exercice 1 : [*Estimateur*] On considère un échantillon $(X_1, ..., X_n)$ de loi de Exponentielle de paramètre λ. On pose pour tout $1 \leq i \leq n$

$$Y_i = 1 \text{ si } X_i > 1 \text{ et } Y_i = 0 \text{ sinon.}$$

1. Montrer que $\overline{Y}$ est un estimateur sans biais de $e^{-\lambda}$.
2. Calculer son risque.

Exercice 2 : [*Estimateur*] On considère un échantillon $(X_1, ..., X_n)$ de la variable aléatoire parente X de loi $\mathcal{N}(0, \theta^2)$.

1. Calculer le moment d'ordre 1 de $|X|$.
2. En déduire un estimateur de θ.
3. Calculer son risque.

Exercice 3 : [*Intervalle de confiance au niveau α du paramètre μ d'une loi Normale*]
On a pesé 15 poulpes mâles adultes pêchés au large des côtes Mauritaniennes. On suppose que, pour cette espèce de poulpe, les poids sont répartis selon une loi normale d'espérance μ et de variance σ^2. Le tableau suivant donne l'échantillon des 15 valeurs obtenues.

1150	1500	1700	1800	1800
1850	2200	2700	2900	3000
3100	3500	3900	4000	5400

1. Donner une estimation de μ et de σ.
2. Construire un intervalle de confiance de μ de niveau de confiance 95%. Donner l'amplitude de cet intervalle de confiance.
3. Si n désigne la taille d'un échantillon, donner l'amplitude de d'intervalle de confiance de μ à 95% en fonction de n.
4. Quelle doit être la taille de l'échantillon pour que cette amplitude soit inférieur à 500g ? Faire numériquement sous MATLAB.

Exercice 4 : [*Intervalle de confiance au niveau α du paramètre σ^2 d'une loi de Normale*]
Un groupement de citoyen.ne.s du Finistère veux évaluer le taux d'azote dans les eaux de leurs villages. Dans cette étude, nous nous intéressons à la variabilité de ce taux exprimé en unités internationales. Ce dernier est évalué à partir de $n = 23$ prélèvements choisis indépendamment et de manière aléatoire. Les résultats observés dans cet échantillon sont indiqués en pourcentage dans la feuille MATLAB "azote.txt". La distribution du taux est considérée comme sensiblement gaussienne.

1. Donner un estimateur de σ^2 et la valeur de l'estimation sur l'échantillon disponible.
2. Construire un intervalle de confiance de σ^2 de niveau de confiance 90%. Le niveau de confiance est-il exact ou approché ?

Exercice 5 : Une société de service de nettoyage envisage d'ajouter à ses prestations habituelles le nettoyage des rideaux et tentures. La société veut évaluer en pourcentage le nombre de clients intéressés par un tel service.
Un sondage est réalisé auprès de 300 personnes choisies aléatoirement parmi les clients. Dans cet échantillon, on observe que 23 % des clients sont intéressés par ce nouveau service.

1. Estimer la proportion p de clients prêts à utiliser ce nouveau service.
2. Déterminer un intervalle de confiance de cette proportion au niveau de confiance 95%. Le niveau de confiance annoncé est-il exact ou approché ?

Exercice 6 : [*Estimateur*] On considère un échantillon $(X_1, ..., X_n)$ de loi de Poisson de paramètre λ.

1. En utilisant la méthode des moments, proposer un estimateur T du paramètre λ.
2. Calculer son risque.
3. Construire un intervalle de confiance de λ au seuil de risque 1%.

Statistique - Séance de TD 2

Dans cette séance, on met en oeuvre les différents tests statistiques vus en cours.

📢 **Exercices à préparer : 1 et 3.**

📢 **Fichier de données donneesTD2 disponible sur la plateforme pédagogique.**

Exercice 1 : On suppose que la résistance à la traction d'un fil est une variable aléatoire qui suit la loi $\mathcal{N}(m, \sigma^2)$ et que l'on est en présence de deux procédés de fabrication différents avec les paramètres respectifs $m_0 = 100$ et $m_1 = 120$, $\sigma^2 = 100$ étant connue dans les deux cas.
On dispose d'un échantillon de fils fabriqués à l'aide d'un des procédés et on souhaite déterminer de quel procédé il s'agit. L'échantillon dont on dispose est de taille $n = 4$ et, après calculs, on a $\overline{x} = 110$.

1. Faire un test avec un risque $\alpha = 0,05$ en renseignant les rubriques suivantes :
   - Variable de décision D :
   - Forme de la zone de rejet A ou W :
   - Zone de rejet explicite numériquement A ou W :
   - Résultat du test :
   - Conclusion
   - Risque de seconde espèce β :
2. Faire un ajustement numérique sous Matlab pour déterminer quel devrait être l'effectif de l'échantillon pour que l'on ait $\alpha = \beta \leq 0,01$? Quelle est alors la région de rejet ?

Exercice 2 : On considère la réalisation d'un échantillon gaussien $\{5, 7, 9, 10, 6, 8, 6, 5, 9, 4, 13\}$. Calculer la variance empirique s^2 de l'échantillon. Vérifier qu'au niveau de risque 5%, on ne peut rejeter l'hypothèse $\sigma^2 = 4$. Pour cela, on procédera à un test en renseignant les rubriques suivantes :

- Variable de décision D :
- Forme de la zone de rejet A ou W :
- Zone de rejet explicite numériquement A ou W :
- Résultat du test :
- Conclusion
- Le risque de seconde espèce β s'il est calculable :

Exercice 3 : [*Ajustement à loi une loi normale avec paramètres à estimer*] La répartition des durées de 670 vols Paris-Alger (en heure) en Caravelle est donnée dans le fichier vol.txt.

1. Tracer l'histogramme avec Matlab.
2. Calculer l'estimation de la moyenne $\bar{x}$ et l'écart-type s.
3. On veut tester l'ajustement de ces données à une loi normale au seuil de risque 10%.
   (a) Ecrire $\mathbb{H}_0$ et $\mathbb{H}_1$
   (b) Donner la variable de décision D^2, sa loi sous $\mathbb{H}_0$ ainsi que la zone de rejet.

(c) Calculer la valeur de la réalisation de D^2 et conclure. Pour ce faire, on utilisera les commandes Matlab suivantes pour créer les bornes des classes et calculer les effectifs dans chaque classe : $edges = 1.9 : 0.05 : 2.55$ et $effectifs = histcounts(T.vols, edges)$ où $T.vols$ contient les données.

(d) Pour quels seuils de risque a-t-on la même conclusion ?

Exercice 4 : Le contrôle effectué sur deux lots de bobines électriques en provenance de deux fournisseurs différents donne le résultat suivant :

Fournisseur 1 : bobines livrées 1200, nombre de pièces défectueuses 42 ;

Fournisseur 2 : bobines livrées 1500, nombre de pièces défectueuses 30.

On veut savoir s'il y a une différence significative de la qualité de fabrication entre les deux fournisseurs avec un risque de première espèce $\alpha = 0,05$.

1. On suppose que les données des fournisseurs sont des réalisations d'échantillons indépendants de v.a. X^i suivant des lois de Bernoulli de paramètres p_i, pour $i = 1, 2$. Proposer des lois approchant celles de $\overline{X^1}$ et $\overline{X^2}$.
2. Proposer une loi approchant celle de $\overline{X^1} - \overline{X^2}$ si $p_1 = p_2 = p$.
3. Réaliser un test au seuil de risque α de l'égalité des moyennes.

Indication : *on pourra estimer la valeur commune de p_1 et p_2 sous l'hypothèse $p_1 = p_2$ par*

$$\hat{p} = \frac{n_1\overline{X^1} + n_2\overline{X^2}}{n_1 + n_2}$$

en justifiant ce choix.

Statistique - Séance de TD 3

Exercices à préparer : 1 et 2

Exercice 1 : Justifier les affirmations suivantes

1. En régression linéaire simple, quand $\mathcal{R}^2 = 1$ les points sont alignés.
   Indication : Écrire la formule de la décomposition de la variance.
2. L'expression de b_0 obtenue par moindres carrés est $\overline{y} - b_1\overline{x}$ et donc la droite de régression passe par le point $(\overline{x}, \overline{y})$
3. La variance d'estimation de $\mathbb{E}[Y_0]$ est minimale pour $x_0 = \overline{x}$.
   Indication :
   $\mathbb{V}\mathrm{ar}[\hat{y}_0] = \sigma^2(1\, x_0)\,(X'X)^{-1}\,(1\, x_0)'$ où $(X'X)^{-1}$ est proportionnelle à $\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$

Exercice 2 : Le fichier "temperatures.txt contient les températures moyennes annuelles de 34 villes européennes ainsi que leur latitude et longitude. On réalise une régression linéaire de la température en fonction de la latitude.

1. Visualiser les données.
2. Donner l'équation du modèle ainsi que les hypothèses probabilistes.
3. Quelles sont les valeurs des paramètres estimés ?
4. Ajouter la droite de régression sur la figure précédente (fonction *refline* pour le tracer d'une droite)
5. Quel est le pourcentage de variance expliquée par le modèle ?
6. L'effet de la latitude sur la température est-il significatif ? On donnera l'hypothèse H_0, la statistique du test, sa loi sous H_0 et la conclusion du test.
7. La ville de Zurich se positionne à la latitude 47.2. Que prévoit le modèle pour cette ville ? avec quelle incertitude (donner un intervalle de confiance à 95% sur l'estimation de cette quantité et un intervalle de prédiction à 95%) ?
8. La température moyenne à Zurich est de $8.7°C$. La prédiction était-elle bonne ?
9. On décide d'améliorer le modèle en ajoutant la longitude. Conclure sur l'effet de cette variable sur la température.

Exercice 3 :
On suppose que l'on observe des réalisations $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, pour $i \in \{1, \dots, n\}$, associées au modèle linéaire

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_i^{(k)} + \varepsilon_i\,,$$

où $x_i = (x_i^{(1)}, \dots, x_i^{(p)})$, $(\varepsilon_i)_i$ sont des variables aléatoires telles que

$$\mathbb{E}(\varepsilon_i) = 0$$
$$\mathbb{E}(\varepsilon_i \varepsilon_j) = \mathbb{1}_{i=j}$$

Un estimateur $\tilde{\beta}$ de β est dit *linéaire* s'il existe une matrice $\mathbf{A}_n \in \mathbb{R}^{(p+1)\times n}$ telle que

$$\tilde{\beta} = \mathbf{A}_n \mathbf{y}_n \quad \text{où} \quad \mathbf{y}_n = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \tag{1}$$

Si $\tilde{\beta}$ est un estimateur, la matrice $\mathbf{A}_n$ peut dépendre de la matrice $\mathbf{X}_n := [\mathbb{1}_n \, x_1 \dots x_p] \in \mathbb{R}^{n\times(p+1)}$ (dont les colonnes sont x_i sauf la première remplie de 1) mais pas de β.
On suppose que $\mathbf{X}_n$ est de rang $p+1$. Le but de cet exercice est de montrer que l'estimateur des moindres carrés

$$\widehat{\beta}_n := \underbrace{(\mathbf{X}_n^\top \mathbf{X}_n)^{-1}\mathbf{X}_n^\top}_{\mathbf{A}_n^\star} \mathbf{Y}_n$$

est le *Best Linear Unbiased Estimator* (BLUE), résultat connu sous le nom de **théorème de Gauss-Markov**. Ici, "Best" signifie de plus petite matrice de covariance au sense de l'ordre partiel sur les matrices symétriques

$$A \preccurlyeq B \quad \Leftrightarrow \quad B - A \text{ est une matrice semi-définie positive, i.e., } \langle u, (B-A)u\rangle \geq 0 \, , \forall u \in \mathbb{R}^{p+1} \, .$$

1. Si $\tilde{\beta}$ est non biaisé et défini par (1) alors $\mathbf{A}_n\mathbf{X}_n = \mathbb{I}_{p+1}$ où $\mathbf{X}_n := [\mathbb{1}_n \, x_1 \dots x_p] \in \mathbb{R}^{n\times(p+1)}$ et $\mathbb{I}_{p+1}$ est la matrice identité de taille $p+1$.
   Indication : Utiliser la définition d'estimateur non-biaisé.
2. Donner la covariance de $\tilde{\beta}$ en fonction de $\mathbf{A}_n$.
3. Montrer que cette covariance est plus grande, au sens des matrices symétriques, que la covariance de $\widehat{\beta}_n$.
   Indication : Poser $\mathbf{D}_n := \mathbf{A}_n - \mathbf{A}_n^\star$, calculer $\mathbf{D}_n\mathbf{D}_n^\top$, et montrer que

$$\mathbf{A}_n\mathbf{A}_n^\top = (\mathbf{X}_n^\top\mathbf{X}_n)^{-1} + \mathbf{D}_n\mathbf{D}_n^\top \, .$$