Probabilités et Statistique

Cours - TD

ENSICA 2ème année

2005/2006

Introduction

La théorie des probabilités procède selon une méthode qui s'apparente à la démarche déductive. Connaissant la loi d'une variable ou d'un vecteur aléatoire, on sait calculer les valeurs exactes des paramètres qui la caractérisent, comme l'espérance ou la variance, et déterminer les lois de nouvelles variables ou vecteurs aléatoires fonction de la variable ou des vecteurs aléatoires donnés ainsi que les limites de suites de variables et de vecteurs aléatoires.

La théorie statistique procède selon une démarche radicalement différente qui s'apparente à l'induction et qui consiste à exploiter des données d'une ou plusieurs variables décrivant plusieurs populations qui ont une existence réelle dans les domaines économiques, industriel, médical ou autre, dans le but de prendre des décisions du type : choix d'une hypothèse parmi plusieurs possibles, comparaison de paramètres, etc.

Par exemple, étant données plusieurs populations décrites par des variables aléatoires numériques dont les paramètres (espérance, variance,...) sont inconnus, il pourra s'agir d'estimer d'abord ces paramètres à l'aide des seules informations contenus dans de petits échantillons extraits de ces populations, puis de tester les hypothèses d'égalité ou d'inégalité de ces paramètres.

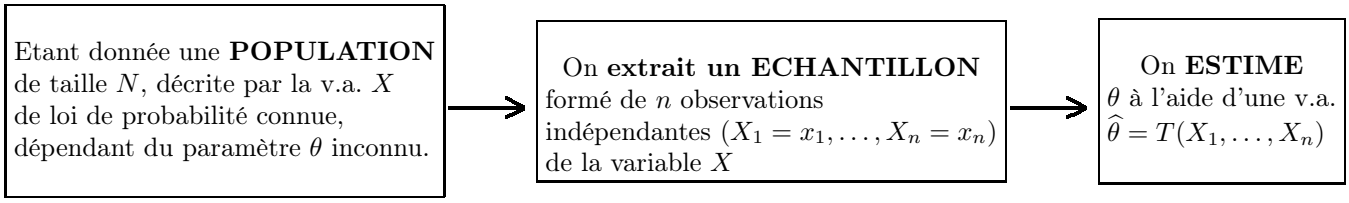
Table des matières

Chapitre 1

Estimation statistique

1 Introduction

La procédure d'estimation s'articule selon le schéma suivant :



L'indépendance n'est rigoureusement acquise que s'il y a tirage avec remise ; toutefois si la population est très grande (plusieurs milliers au moins) on peut faire l'économie de cette hypothèse.
$T(X_1, .., X_n)$ est une v.a. fonction de l'échantillon $(X_1, .., X_n)$, dite **statistique**, construite pour représenter de façon optimale l'information sur un paramètre inconnu, contenue dans l'échantillon.

Exemple Estimation d'une moyenne
Pour estimer la durée de vie moyenne m d'une ampoule électrique, on prélève au hasard un échantillon de 30 ampoules. On réalise une expérience pour observer les durées de vie de ces ampoules. Quelle estimation de m peut-on proposer ?

2 Estimation ponctuelle d'un paramètre

2.1 Modèle statistique inférentiel

Soit $(x_1, \ldots, x_n)$ une observation (ou une réalisation) du vecteur aléatoire $X = (X_1, \ldots, X_n)$; $(x_1, x_2, ..., x_n)$ est dit aussi un **échantillon de données**.

Définition 1 Construire un **modèle statistique** revient à définir sur l'espace probabilisé $\mathbb{R}^n$, muni de la tribu des boréliens $\mathcal{B}(\mathbb{R}^n)$ une probabilité P_θ, où θ est un paramètre ou un vecteur de paramètres inconnu.

La probabilité P sera définie, selon l'un des deux cas, par :

$$\begin{cases} \text{la loi conjointe } \mathrm{P}_\theta\left(X_1 = x_1, \ldots, X_n = x_n\right) \text{ dans le cas de v.a. discrètes,} \\ \text{ou} \\ \text{la densité conjointe } f_{X;\theta}\left(x_1, \ldots, x_n\right) \text{ dans le cas de v.a. continues.} \end{cases}$$

Définition 2 Une **statistique** T est une application de l'espace $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathrm{P}_\theta)^n$ à valeurs dans $\mathbb{R}$.

$$\begin{array}{rcl} T : \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ (x_1, x_2, \ldots, x_n) & \longmapsto & T(x_1, x_2, \ldots, x_n) = t \ . \end{array}$$

La quantité t est une observation de la v.a. $T(X_1, X_2, \ldots, X_n)$.

Par exemple, les statistiques usuelles sont :
- la moyenne empirique $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$,
- la variance empirique $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$,
- ...

Exemple 1 La v.a. $T_n = \sum_{i=1}^n X_i$ (somme des X_i indépendants) a pour loi de probabilité une loi normale $\mathcal{N}(n\mu, n\sigma^2)$ si la structure est $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathcal{N}(\mu, \sigma^2))^n$.

2.2 Qualités d'un estimateur

Définition 3 Un **estimateur** du paramètre inconnu θ est une statistique $\widehat{\theta}$ dont la valeur observée est une approximation de θ.

Notation On note souvent $\widehat{\theta_n} = \widehat{\theta}\left(X_1, X_2, \ldots, X_n\right)$ la v.a. correspondant à l'estimateur T.

La qualité d'un estimateur T du paramètre θ sera évaluée grâce aux propriétés suivantes.

Définition 4 Un **estimateur T est sans biais** si $\mathbb{E}\left(T\right) = \theta$.

Exemple 2 La statistique $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ est un estimateur sans biais de la moyenne théorique $m = \mathbb{E}\left(X_i\right)$.

Solution : Calculons l'espérance de $\overline{X}$.

$$\mathbb{E}\left(\overline{X}\right) \ = \ \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(X_i\right) = m$$

La statistique $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$ est un estimateur sans biais de $\sigma^2 = \mathrm{Var}\left(X_i\right)$.

Solution : Calculons l'espérance de S_{n-1}^2. On a :

$$\begin{aligned}
\mathbb{E}\left(S_{n-1}^2\right) &= \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right) = \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left(\left(X_i-\overline{X}\right)^2\right)\\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left(\left((X_i-m)+\left(m-\overline{X}\right)\right)^2\right)\\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left((X_i-m)^2+\left(m-\overline{X}\right)^2+2\left(X_i-m\right)\left(m-\overline{X}\right)\right)\\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left[\mathbb{E}\left((X_i-m)^2\right)+\mathbb{E}\left(\left(m-\overline{X}\right)^2\right)+2\,\mathbb{E}\left(\left(X_i-m\right)\left(m-\overline{X}\right)\right)\right]\\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left[\mathrm{Var}\left(X_i-m\right)+\mathrm{Var}\left(m-\overline{X}\right)-2\,\mathbb{E}\left(\left(X_i-m\right)\left(\overline{X}-m\right)\right)\right]\\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left[\mathrm{Var}\left(X_i\right)+\mathrm{Var}\left(\overline{X}\right)-2\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left(\left(X_i-m\right)\left(X_j-m\right)\right)\right]
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left(S_{n-1}^2\right) &= \frac{1}{n-1}\sum_{i=1}^{n}\left[\sigma^2+\frac{\sigma^2}{n}-2\frac{1}{n}\,\mathbb{E}\left(\left(X_i-m\right)^2\right)\right]\\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left[\sigma^2+\frac{\sigma^2}{n}-\frac{2}{n}\sigma^2\right] = \frac{1}{n-1}\sum_{i=1}^{n}\left[\frac{n-1}{n}\sigma^2\right] = \sigma^2\ .
\end{aligned}$$

On démontre que : $V(S_{n-1}^2) = \frac{1}{n}(\mu_4 - \frac{n-3}{n-1}\sigma^4)$ où μ_4 est le moment centré d'ordre quatre de X ($\mu_4 = \mathbb{E}((X-\bar{X})^4)$.

Définition 5 Un estimateur T est **biaisé** si le biais $\mathbb{E}(T) \neq \theta$.

Définition 6 Un estimateur T est **asymptotiquement sans biais** si $\lim\limits_{n\to+\infty}\mathbb{E}(T) = \theta$.

Définition 7 Soient deux estimateurs T_1 et T_2 sans biais de θ. On dit que T_1 est **meilleur que** T_2 si on a : $\mathrm{Var}(T_1) < \mathrm{Var}(T_2)$.

Définition 8 L'estimateur T de θ est dit **convergent en probabilité** si :

$$\forall\epsilon > 0,\ \lim_{n\to+\infty}\mathrm{P}\left(|T-\theta| < \epsilon\right) = 1.$$

Théorème 1 Tout estimateur sans biais ou asymptotiquement sans biais tel que $\lim_{n\to+\infty}\mathrm{Var}(T) = 0$ est convergent en probabilité vers θ.

Preuve *Utiliser l'implication* $|T-\mathbb{E}(T)| < \frac{\epsilon}{2} \Rightarrow |T-\theta| < \epsilon$ *et l'inégalité de Bienaymé-Tchebicheff :* $\forall\epsilon > 0,\ \mathrm{P}\left(|T-\mathbb{E}(T)| < \varepsilon\right) \leqslant \frac{\mathrm{Var}(T)}{\epsilon^2}$.

Théorème 2 Sous les hypothèses de régularité suivantes :

H_1 : Le support de la densité $D = \{x \,/\, f(x;\theta) > 0\}$ est indépendant de θ ;

H_2 : θ varie dans un intervalle ouvert I ;

H_3 : $\forall\, \theta \in I$, $\forall\, x \in D$, $\dfrac{\partial}{\partial\theta} f(x;\theta)$ et $\dfrac{\partial^2}{\partial\theta^2} f(x;\theta)$ existent et sont intégrables par rapport à x ;

H_4 : pour tout $\theta \in I$, pour tout $A \in \mathcal{B}(\mathbb{R})$, $\displaystyle\int_A f(x;\theta)\, dx$ est deux fois dérivable par rapport à θ ;

H_5 : $\dfrac{\partial}{\partial\theta} \ln f(X;\theta)$ est une v.a. de carré intégrable d'un estimateur sans biais.

HS : On détermine la borne de Cramer-Rao, $V_n(\theta)$, définie par :

$$V_n(\theta) = \frac{1}{-\,\mathbb{E}\left[\dfrac{\partial^2 \log f\,(X_1,\ldots,X_n;\theta)}{\partial\theta^2}\right]}$$

où $-\,\mathbb{E}\left[\dfrac{\partial^2 \log f\,(X_1,\ldots,X_n;\theta)}{\partial\theta^2}\right]$ est l'**information de Fisher**

et dans le cas particulier d'un échantillon $X_1,\ldots,X_n$ de v.a indépendantes de même loi :

$$V_n(\theta) = \frac{1}{-n\,\mathbb{E}\left[\dfrac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right]}$$

où $f(X;\theta)$ désigne la densité si X est continue ou $\mathrm{P}\,(X = x)$ si X est discrète.

Quel que soit l'estimateur sans biais $\widehat{\theta}$ de θ, la borne $V_n(\theta)$ de **Cramer-Rao** minore $\mathrm{Var}\left(\widehat{\theta}\right)$: $\mathrm{Var}\left(\widehat{\theta}\right) \geqslant V_n(\theta)$.

Définition 9 Un estimateur T_n sans biais de θ est dit **efficace** si sa variance atteint la borne de Cramer-Rao.

Remarque 1 *: De plusieurs estimateurs sans biais, le meilleur est celui qui a la plus petite variance. Attention ! Il est possible qu'il n'existe pas d'estimateur efficace.*

Exercice 1 : Estimation de la borne d'un support borné de variable aléatoire

On considère T la variable aléatoire : «durée d'attente à un feu rouge». La durée du feu rouge est égale à θ, paramètre inconnu strictement positif.
On observe un échantillon $t_1, t_2, \ldots, t_n$ de taille n, où t_i désigne la durée d'attente observée pour le $i^{\text{ème}}$ individu. On fait l'hypothèse que les variables aléatoires $T_1, T_2, \ldots, T_n$ sont indépendantes et de même loi uniforme sur $[0;\theta]$, notée $\mathcal{U}[0;\theta]$.

1) Représenter le graphe de la densité de la loi $\mathcal{U}[0;\theta]$ et préciser ses paramètres de moyenne et de variance.

2) Soit la statistique $\overline{T} = \frac{1}{n}\sum_{i=1}^{n} T_i$. Calculer $\mathbb{E}\left(\overline{T}\right)$ et $\mathrm{Var}\left(\overline{T}\right)$.

Montrer que la statistique $\widehat{\theta_1} = 2\overline{T}$ est un estimateur sans biais de θ et convergent en probabilité.

3) Soit la statistique $Y_n = \sup_i T_i$.

a) En utilisant l'équivalence des événements $(Y_n < y)$ et $(\forall i = 1, \ldots, n\ T_i < y)$, calculer la fonction de répartition de Y_n. En déduire sa densité, calculer $\mathbb{E}(Y_n)$, $\mathrm{Var}(Y_n)$ et tracer le graphe de la densité pour $n = 3$, puis pour $n = 30$. Comparer les graphes et interpréter les.

b) Montrer que la statistique $\widehat{\theta_2} = \frac{n+1}{n} Y_n$ est un estimateur sans biais de θ et convergent en probabilité.

4) Comparer les variances $\mathrm{Var}\left(\widehat{\theta_1}\right)$ et $\mathrm{Var}\left(\widehat{\theta_2}\right)$. Lequel des deux estimateurs $\widehat{\theta_1}$ et $\widehat{\theta_2}$ choisiriez-vous pour estimer θ ? Calculer la borne de Cramer-Rao et conclure.

Application : pour $n = 10$, on a $(t_1, \ldots, t_{10}) = (28; 33; 42; 15; 20; 27; 18; 40; 16; 25)$. Quelle est l'estimation de la durée du feu rouge ?

Exercice 2 : Estimateur sans biais de σ

(Extrait de l'ouvrage de B.Garel. Modélisation probabiliste et statistique).
Soient $X_1, \ldots, X_n$ des v.a. indépendantes de même loi $\mathcal{N}(m; \sigma^2)$. On observe $Y_i = |X_i - m|$, $i = 1, \ldots, n$.

1) Montrer que $\forall i$, Y_i admet la densité g sur $\mathbb{R}^+$ définie par :

$$y \longmapsto g(y) = \frac{2}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}} \mathbb{1}_{\mathrm{R}^+}(y).$$

En déduire que $\mathbb{E}(Y_i) = \sigma\sqrt{\frac{2}{\pi}}$.

Remarquons que $\mathbb{E}(|X_i - m|)$ est un indice de dispersion qui joue un rôle analogue mais non équivalent à la variance.

2) On cherche un estimateur sans biais de σ de la forme $\widehat{\sigma} = \sum_{i=1}^{n} a_i Y_i$. En calculant l'espérance de $\widehat{\sigma}$, trouver une contrainte linéaire sur les a_i.

3) Sous cette contrainte, montrer que $\sum_{i=1}^{n} a_i^2$ est minimale si et seulement si les a_i sont tous égaux.

4) On note alors $\widehat{\sigma_n}$ l'estimateur de σ associé à ce dernier cas : les a_i sont tous égaux. Calculer $\mathrm{V}(\widehat{\sigma_n})$.

5) Calculer la borne inférieure de Cramer-Rao pour un estimateur sans biais de σ. L'estimateur $\widehat{\sigma_n}$ est-il efficace ?

2.3 Méthodes d'estimation

Il existe plusieurs méthodes pour construire un estimateur : la méthode des moindres carrés (théorie de la régression), la méthode des moments et la méthode du maximum de vraisemblance, qui est la méthode de référence présentée ci-dessous.

2.3.1 Résultats généraux pour un échantillon

Théorème 3 Dans le cas d'un échantillon $(X_1, X_2, \ldots, X_n)$ extrait d'une population de moyenne μ et de variance σ^2, inconnues, les propriétés suivantes sont vérifiées :

- La statistique $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur sans biais et convergent en probabilité de la moyenne μ.
- La statistique $S^2_{n-1} = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ est un estimateur sans biais et convergent en probabilité de la variance σ^2.

Preuve *On a démontré (exemples 2) que* $\mathbb{E}\left(\overline{X}\right) = \mu$, *que* $\mathbb{E}\left(S^2_{n-1}\right) = \sigma^2$ *et*

$$\mathrm{Var}\left(\overline{X}\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Par un calcul analogue, mais plus long, on établit que :

$$\mathrm{Var}\left(S^2_{n-1}\right) = \frac{n}{(n-1)^2}\left[\mu_4 - \sigma^4 - 2\frac{(\mu_4 - 2\sigma^4)}{n} + \frac{\mu_4 - 3\sigma^4}{n^2}\right]$$

qui converge vers 0 quand $n \to +\infty$ *où* $\mu_4 = \mathbb{E}\left(X_i - \mu\right)^4$ *(moment centré d'ordre 4 de* X_i*).*

Rappelons le théorème de Fisher, utile à la théorie des tests.

Théorème 4 Si l'échantillon $(X_1, \ldots, X_n)$ est formé de v.a. de même loi normale, les estimateurs $\overline{X}$ et S^2_{n-1} sont indépendantes.

2.3.2 Méthode du maximum de vraisemblance

On se donne une population décrite par une v.a. X de loi connue, fonction d'un paramètre inconnu θ.

Définition 10 On appelle **fonction de vraisemblance** ou **vraisemblance** des v.a. $X_1, \ldots, X_n$,

la fonction l telle que :

$$l : \begin{array}{ccc} \mathbb{R}^n \times \Theta & \longrightarrow & \mathbb{R}^+ \\ (\underline{x}, \theta) & \longmapsto & l(\underline{x};\theta) \end{array} = \begin{cases} f(\underline{x};\theta) \text{ densité conjointe de } X_1, \ldots, X_n, \\ \text{dans le cas à densité} \\ \\ \text{ou} \\ \\ \mathrm{P}_\theta(X_1 = x_1, \ldots, X_n = x_n), \\ \text{dans le cas de v.a. discrètes.} \end{cases}$$

où $\underline{x} = (x_1, x_2, ...x_n)$.

Remarque 2 *Si les v.a. $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées (iid), on a :*

$$l(\underline{x};\theta) = \begin{cases} \displaystyle\prod_{i=1}^{n} f(x_i;\theta) \quad \textit{dans le cas à densité} \\ \\ \textit{ou} \\ \\ \displaystyle\prod_{i=1}^{n} I\!P_\theta[\ X_i = x_i\ ] \quad \textit{dans le cas discret.} \end{cases}$$

Exemple 3 Soit l'échantillon $(X_1, \ldots, X_n)$ où chaque X_i est de loi normale $\mathcal{N}(\mu, \sigma^2)$. La fonction de vraisemblance s'écrit alors :

$$\begin{aligned} \forall\, (x_1, \ldots, x_n)\,,\ l(\underline{x};\theta) &= \prod_{i=1}^{n} f_\theta^{(i)}(x_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\ \exp\left\{-\frac{1}{2}\left(\frac{x_i-\theta}{\sigma}\right)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\prod_{i=1}^{n}\sigma^{-1}\right)\ \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\theta}{\sigma}\right)^2\right\}\ . \end{aligned}$$

Définition 11 On appelle **estimateur du maximum de vraisemblance** (noté e.m.v.), la statistique $\widehat{\theta}$ définie par $\widehat{\theta} = \widehat{\theta}(x_1, x_2, ..., x_n)$ et associée à la valeur de θ qui rend **maximum** la vraisemblance $l(\underline{x};\theta)$, qui équivaut à supposer, sachant que $l(\underline{x};\theta)$ est une probabilité, que l'événement qui s'est produit est le plus vraisemblable.

On a alors :

$$l(x_1, \ldots, x_n; \widehat{\theta}\,(x_1, \ldots, x_n)) = \max_{\theta\in\Theta} l(x_1, \ldots, x_n;\theta).$$

ou bien encore, puisque la fonction log est croissante :

$$\widehat{\theta}\,(x_1, \ldots, x_n) = \arg\ \max_{\theta\in\Theta} l(x_1, \ldots, x_n;\theta) = \arg\ \max_{\theta\in\Theta} \log l(x_1, \ldots, x_n;\theta)\ .$$

En pratique, il sera plus facile de maximiser la fonction $\log l(x_1, \ldots, x_n;\theta)$, dite "log de vraisemblance".

Application Dans l'exemple précédent, déterminons l'estimateur du maximum de vraisemblance du paramètre μ, sous l'hypothèse que σ est connu. Pour cela, calculons $\log l(x_1, \ldots, x_n; \mu)$ et cherchons la solution qui maximise cette quantité, obtenue par la résolution de l'équation :

$$\frac{\partial \log l(x_1, \ldots, x_n; \mu)}{\partial \mu} = 0 \ ,$$

qui définit les valeurs stationnaires de $\log l(x_1, \ldots, x_n; \mu)$. On s'assurera alors que la dérivée seconde en ce point est négative, ce qui garantit que le point critique est bien un maximum.

L'estimateur de vraisemblance n'existe pas toujours, et quand il existe il peut ne pas être unique et peut être biaisé ou encore non efficace. Le théorème suivant modère le caractère relatif de la remarque précédente.

Théorème 5 Quand la taille n de l'échantillon est suffisamment «grande» (dès que $n > 30$), l'e.m.v. $\widehat{\theta}$ possède les trois propriétés suivantes :

1. $\widehat{\theta}$ est asymptotiquement sans biais et asymptotiquement efficace ;
2. $\dfrac{\widehat{\theta} - \theta}{\sqrt{V_n(\theta)}} \xrightarrow[n \to +\infty]{\text{Loi}} \mathcal{N}(0; 1)$, où $V_n(\theta)$ est la borne de Cramer-Rao ;
3. $\widehat{\theta}$ converge vers θ en moyenne quadratique

Le théorème ci-dessus garantit d'une part la bonne qualité de l'estimateur de vraisemblance $\widehat{\theta}$ pour des échantillons suffisamment grands ; d'autre part, il permet grâce à l'approximation gaussienne de $\widehat{\theta}$, de construire des intervalles de confiance pour la valeur exacte du paramètre θ.

Exercice 3 : Estimation d'une proportion

Pour estimer la proportion p de pièces défectueuses à la sortie d'une chaîne de production, on prélève un échantillon de n_1 pièces. A la $i^{\text{ème}}$ pièce tirée, on associe la v.a.

$$X_i = \begin{cases} 1 & \text{si la pièce est défectueuse,} \\ 0 & \text{sinon.} \end{cases}$$

1) Construire l'estimateur du maximum de vraisemblance T_{n_1} de la proportion p et étudier ses propriétés.

2) On tire un deuxième échantillon de n_2 pièces, indépendamment du premier. On note F_1 et F_2 les fréquences relatives des pièces défectueuses.

a) $F = \frac{F_1 + F_2}{2}$ est-il un estimateur sans biais de p ? Quelle est sa variance ?

b) Déterminer les réels a et b de manière que F^* défini par $aF_1 + bF_2$, où $a + b = 1$ et $a, b > 0$, soit un estimateur sans biais de p, de variance minimum.

Exercice 4 : Estimation de la moyenne et de l'écart-type d'une variable gaussienne

Soit $(X_1, X_2, \ldots, X_n)$ un échantillon de loi $\mathcal{N}(m, \sigma^2)$

1) Déterminer l'estimateur du maximum de vraisemblance T de m. Déterminer ses propriétés.

2) La connaissance a priori de σ^2 modifie-t-elle le résultat ?

3) Supposons maintenant m connu. Déterminer l'estimateur du maximum de vraisemblance $\widehat{\sigma_n^2}$ de σ^2 et étudier ses propriétés.

4) En déduire un estimateur $\widehat{\sigma_n}$ de σ. A-t-il un biais ? Calculer sa borne de Cramer-Rao.

5) Supposons que m est inconnu. Déterminer un estimateur S_{n-1}^2 sans biais de σ^2 et calculer sa variance.
On montrera que S_{n-1}^2 est égal à $\frac{n}{n-1}T'$ où $T' = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ et $\frac{nT'_n}{\sigma^2}$ est un chi-deux à $(n-1)$ d.d.l.

2.4 Notion de statistique exhaustive

Tout échantillon $(X_1, \ldots, X_n)$ d'une v.a. X, de loi connue paramétrée par θ inconnu, contient une information sur ce paramètre.

Définition 12 Une statistique $T(X_1, \ldots, X_n)$ destinée à l'estimation du paramètre θ est dite **exhaustive** si elle apporte toute l'information contenue dans l'échantillon $(X_1, \ldots, X_n)$.

Exemple de statistique exhaustive Lors d'un contrôle industriel, on tire avec remise n pièces d'un lot, dans le but d'estimer la proportion p de pièces défectueuses.
A chaque pièce tirée est associée la v.a. de Bernoulli X_i de loi :

$$\mathrm{P}_{X_i} = p\delta_1 + (1-p)\delta_0 \ .$$

L'estimateur classique $\frac{\sum_{i=1}^{n} X_i}{n}$ de p est-il exhaustif ? Il suffit de démontrer que $T = \sum_{i=1}^{n} X_i$ est un estimateur exhaustif de np.

Solution Montrons que la loi de $(X_1, \ldots, X_n)$ conditionnée par T est indépendante du paramètre p à estimer ; on en conclura alors que les valeurs individuelles $(X_1 = x_1, \ldots, X_n = x_n)$ n'apportent pas plus d'information sur p que la seule valeur $T = t$. On a :

$$\begin{aligned}
\mathrm{P}\left(X_1 = x_1, \ldots, X_n = x_n \,/\, T = t\right) &= \frac{\mathrm{P}\left(X_1 = x_1, \ldots, X_n = x_n, T = t\right)}{\mathrm{P}\left(T = t\right)} \\
&= \frac{\mathrm{P}\left(X_1 = x_1, \ldots, X_n = t - \sum_{i=1}^{n} x_i\right)}{\mathrm{P}\left(T = t\right)}
\end{aligned}$$

Rappelons que $\mathrm{P}\left(X_i = x_i\right) = p^{x_i} \cdot (1-p)^{1-x_i}$, où $x_i \in \{0; 1\}$ et que T suit une loi binomiale $\mathcal{B}(n, p)$. L'équation précédente s'écrit :

$$\begin{aligned}
\mathrm{P}\left(X_1 = x_1, \ldots, X_n = x_n \,/\, T = t\right) &= \frac{p^t \cdot (1-p)^{n-t}}{\binom{n}{t} p^t \cdot (1-p)^{n-t}} \\
&= \frac{1}{\binom{n}{t}} \text{ indépendant de } p.
\end{aligned}$$

La statistique $\sum_{i=1}^{n} X_i$ est donc exhaustive.

On pourrait avoir l'intuition de ce résultat, sachant que $\binom{n}{t}$ est le nombre de choix de t indices i vérifiant $x_i = 1$.

Nous ne démontrerons pas le théorème qui caractérise l'exhaustivité :

Théorème 6 (Neyman-Fisher) Soit X une v.a. de densité $f(x;\theta)$. Une statistique $T(X_1, \ldots, X_n)$ est exhaustive s'il existe des applications g et h, positives et mesurables telles que la vraisemblance s'écrit selon une factorisation non nécessairement unique :

$$l(x_1, \ldots, x_n; \theta) = g(t; \theta) \cdot h(x_1, \ldots, x_n)$$

Remarque 3 *De nombreuses lois appartenant à une famille de lois dite exponentielle, que nous ne définirons pas ici, admettent toutes des statistiques exhaustives. Il en est ainsi de la statistique définissant l'estimateur classique de la moyenne* $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$*, pour la loi binomiale, pour la loi de Poisson ou pour la loi gaussienne. Par contre, ni la loi de Cauchy de densité* $\frac{1}{\pi(1+x^2)}$*, ni la loi uniforme sur un segment, n'admettent de statistique exhaustive.*

En pratique, on recherche des statistiques exhaustives conduisant à des estimateurs sans biais et de variance minimale.

3 Estimation par intervalle de confiance

3.1 Cas général

Etant données une estimation ponctuelle $\widehat{\theta}$ du paramètre θ et une probabilité α petite, de l'ordre de 0.01 ou 0.05, on souhaite construire un intervalle dont on soit sûr qu'il contienne la valeur exacte de θ avec la probabilité $(1-\alpha)$. On remarque que pour α fixé, il existe une infinité de tels intervalles, appelés **intervalles de confiance de niveau** $1-\alpha$ de la forme $[\theta_1, \theta_2]$ vérifiant $\mathbb{P}(\theta \in [\theta_1, \theta_2]) = 1-\alpha$, où $\mathbb{P}(\theta < \theta_1) = \alpha_1$, $\mathbb{P}(\theta > \theta_2) = \alpha_2$, $\alpha_1 + \alpha_2 = \alpha$.

En pratique on utilise deux sortes d'intervalles de confiance :

- les intervalles de confiance bilatéraux symétrique $[\theta_1, \theta_2]$ désormais unique puisque $\mathbb{P}(\theta < \theta_1) = \mathbb{P}(\theta > \theta_2) = \alpha/2$; ces intervalles standard conviennent parfaitement au cas où la loi de l'estimateur $\widehat{\theta}$ est symétrique (Normale, Student,...), mais ils conviennent aussi de façon générale.

– les intervalles unilatéraux à gauche $]-\infty, \theta_1]$, ou à droite $[\theta_2, +\infty[$ qui conviennent à des situations spécifiques ; par exemple, s'il s'agit d'estimer la proportion p de pièces défectueuses d'un lot, on cherche à borner supérieurement p, donc $p \leqslant p_1$: dans ce cas $\mathbb{P}(p \leqslant p_1) = 1 - \alpha$.

3.2 Estimation d'une moyenne théorique

On a vu que la moyenne théorique est estimée par la moyenne empirique $\overline{x}$ des valeurs observées $x_1, \ldots, x_n$ de l'échantillon ; on peut donc construire un intervalle de confiance à partir de la loi de probabilité de $\overline{X}$.

3.2.1 Construction de l'intervalle de confiance de la moyenne inconnue m d'une population gaussienne $\mathcal{N}(m, \sigma^2)$ où σ^2 est connue

Etant donnée une population de loi gaussienne, de variance σ^2 *connue* et de moyenne m *inconnue*. On rappelle que pour un échantillon $(X_1, X_2, \ldots, X_n)$ extrait de la population, l'estimateur (classique) de la moyenne est :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \ , \qquad \mathbb{E}\left(\overline{X}\right) = m \qquad \text{et} \qquad \mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$$

En outre $\overline{X} \sim \mathcal{N}(m, \frac{\sigma^2}{n})$ et donc que $\frac{\overline{X}-m}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.

Si U suit une loi $\mathcal{N}(0,1)$, alors pour tout α, $0 < \alpha < 1$, il existe un réel positif, noté $u_{1-\alpha/2}$, tel que :

$$\mathrm{P}\left(-u_{1-\alpha/2} \leqslant U \leqslant u_{1-\alpha/2}\right) = 1 - \alpha \ . \tag{1.1}$$

La v.a. $\frac{\overline{X}-m}{\sigma/\sqrt{n}}$ étant une v.a. $\mathcal{N}(0;1)$, (1.1) s'écrit :

$$\begin{aligned} \mathrm{P}\left(-u_{1-\alpha/2} \leqslant \frac{\overline{X}-m}{\sigma/\sqrt{n}} \leqslant u_{1-\alpha/2}\right) &= \mathrm{P}\left(m - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant \overline{X} \leqslant m + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\ &= \mathrm{P}\left(\overline{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant m \leqslant \overline{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \ . \end{aligned}$$

Définition 13 On appelle **fractile** ou **quantile** d'ordre q (q$\in [0,1]$) de la v.a. X la valeur x_q telle que $\mathbb{P}(X \leqslant x_q) = q$. On désignera par u_q le fractile d'ordre q de la v.a. $\mathcal{N}(0,1)$

Définition 14 On appelle **intervalle de confiance** de la moyenne m d'une population gaussienne $\mathcal{N}(m, \sigma^2)$, où σ est connu, l'intervalle :

$$\boxed{I_\alpha = \left[\overline{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]} \text{ où } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

I_α est l'intervalle de confiance de m, au niveau de confiance $1-\alpha$, qui s'interprète par : «on peut affirmer avec la probabilité $1-\alpha$ que $m \in I_\alpha$». On rappelle que :

- si $\alpha = 5$ (la confiance étant de 95%), alors $u_{1-\alpha/2} = 1{,}96$,
- si $\alpha = 1$ (la confiance étant de 99%), alors $u_{1-\alpha/2} = 2{,}58$.

Illustration numérique : $\sigma^2 = 4$, $n = 25$, $\overline{x} = 8$. On a :

$$\begin{aligned} I_{5\%} &= \left[8 - 1{,}96\frac{2}{5}; 8 + 1{,}96\frac{2}{5}\right] = [7{,}22; 8{,}78] \ , \\ I_{5\%} \subset I_{1\%} &= \left[8 - 2{,}56\frac{2}{5}; 8 + 2{,}56\frac{2}{5}\right] = [6{,}97; 9{,}03] \ .... \end{aligned}$$

Théorème 7 $\alpha_1 > \alpha_2 \Longrightarrow I_{\alpha_1} \subset I_{\alpha_2}$, donc plus la confiance exigée est grande, plus l'intervalle de confiance est grand.

Si l'on veut réduire l'amplitude de l'intervalle de confiance I_α dans un rapport k, il faut multiplier la taille de l'échantillon par k^2.

Illustration numérique : $\sigma^2 = 4$, $\alpha = 5$, $\overline{x} = 8$. On a :

$$\begin{aligned} n = 25 &\longrightarrow I_{5\%} = [7{,}22; 8{,}78] \ , \\ n = 100 &\longrightarrow I_{5\%} = [7{,}61; 8{,}39] \ .... \end{aligned}$$

Rappels :

(A) La loi du chi-deux à p degrés de liberté, noté χ^2_p, est la loi de la somme $\chi^2_p = \sum_{i=1}^{p} U_i^2$, où les U_i sont iid $\mathcal{N}(0;1)$.

$$\mathbb{E}(\chi^2_p) = p, \qquad \text{Var}\left(\chi^2_p\right) = 2p.$$

La densité du χ^2_p est celle de la loi $\Gamma(\frac{p}{2}; \frac{1}{2})$; les v.a. $\frac{\chi^2_p - p}{\sqrt{2p}}$ et $\left(\sqrt{2\chi^2_p} - \sqrt{2p-1}\right)$ convergent en loi vers $\mathcal{N}(0;1)$, quand p tend vers $+\infty$.

Théorème 8 Soit l'échantillon $(X_1, \ldots, X_n)$ de loi $\mathcal{N}(m; \sigma^2)$, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ la moyenne empirique et $S^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ la variance empirique de l'échantillon :

(a) $\overline{X}$ et S^2_{n-1} sont indépendantes.

(b) $\overline{X} \sim \mathcal{N}(m, \frac{\sigma^2}{n})$.

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

Théorème 9 Soient deux échantillons indépendants $(U_i)_{i=1,\dots,p}$ et $(V_j)_{j=1,\dots,q}$ de loi $\mathcal{N}(0;1)$ et $\chi^2_p = \sum_{i=1}^{p} U_i^2$, $\chi^2_q = \sum_{j=1}^{q} V_j^2$.

Alors : $\chi^2_p + \chi^2_q = \chi^2_{p+q}$ (stabilité par somme de χ^2 indépendants).

(B) La loi de Student à p degrés de libertés est la loi de $\frac{X}{\sqrt{Z/p}} \underset{\text{notée}}{=} T_p$, où $X \sim \mathcal{N}(0;1)$ et $Z \sim \chi^2_p$, (X et Z indépendantes)

Théorème 10 Quand p tend vers $+\infty$, T_p converge en loi vers $\mathcal{N}(0;1)$.

(C) La loi de Fisher-Snedecor à p et q degrés de liberté est la loi du quotient $\frac{\chi^2_p/p}{\chi^2_q/q} \underset{\text{notée}}{=} F_{p,q}$, où χ^2_p et χ^2_q sont indépendants.

Exercice 5 : Construction d'un intervalle de confiance de l'espérance m d'une loi normale d'écart-type connu

Une machine produit en série des tiges métalliques dont la longueur X, par suite d'imperfections inhérentes au fonctionnement, peut être considérée comme une v.a. de loi normale $\mathcal{N}(m, \sigma^2)$ lorsqu'elle est réglée à la valeur m. L'écart-type est une caractéristique du procédé de fabrication, de valeur connue $\sigma = 0{,}02$ mm.

1) Pour qu'une tige soit utilisable par un client industriel, sa longueur doit être comprise entre 23,60 mm et 23,70 mm. Quelle valeur m_0 faut-il donner à m pour que la proportion de tiges produites utilisables soit maximale ? Quelle est cette proportion maximale ?

2) L'industriel recevant un lot de 10000 tiges ne connaît pas cette valeur de réglage, lui permettant de décider d'accepter ou de refuser le lot qui lui a été livré ; il ne connaît que $\sigma = 0{,}02$. Il va donc tirer un échantillon de n tiges dont il va mesurer les longueurs $(X_1, \dots, X_n)$ pour se faire une idée de la valeur de m. Construire l'intervalle de confiance de niveau 0,90 pour m. Quelle doit être la valeur minimale de n pour que la longueur de cet intervalle soit au plus égale à 0,01 mm ?

3.2.2 Construction de l'intervalle de confiance d'une moyenne m d'une population gaussienne $\mathcal{N}(m, \sigma^2)$ où σ^2 est inconnue

Théorème 11 Quand la variance σ^2 est inconnue, l'I.C. au niveau de confiance $\gamma = 1 - \alpha$, du paramètre m est défini par :

$$I_\alpha = \left[\overline{x} - t_{\alpha/2} \frac{s_{n-1}}{\sqrt{n}} ; \overline{x} + t_{\alpha/2} \frac{s_{n-1}}{\sqrt{n}} \right]$$

où $\begin{cases} s_{n-1} \text{ est l'écart-type défini par } \sqrt{\dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2}, \\ t_{\alpha/2} \text{ est le fractile d'ordre } \alpha/2 \text{ de la v.a. de Student} \end{cases}$

Preuve *La construction de l'I.C. est analogue au cas précédent : cette fois on utilise la loi de probabilité de la variable aléatoire* $T = \sqrt{n}(\frac{\overline{X}-m}{s_{n-1}})$ *qui suit une loi de Student lorsque les v.a.* X_i *sont indépendantes et de loi normale.*

Exercice 6 : Intervalle de confiance de l'espérance m d'une loi normale d'écart-type inconnu

Un fabricant de piles électriques affirme que la durée de vie moyenne du matériel qu'il produit est de 170 heures. Un organisme de défense des consommateurs prélève au hasard un échantillon de $n = 100$ piles et observe une durée de vie moyenne de 158 heures avec un écart-type empirique s_{n-1} de 30 heures.

1) Déterminer un intervalle de confiance de niveau 0,99 pour la durée de vie moyenne m.

2) Peut-on accuser ce fabricant de publicité mensongère ?

Exercice 7 : Intervalle de confiance de la variance d'une loi normale d'espérance connue

Pour estimer la précision d'un thermomètre, on réalise $n = 15$ mesures indépendantes de la température d'un liquide maintenu à température constante égale à 20°C. Compte tenu des erreurs de mesure, ce sont des réalisations de v.a. Xi, $1 \leqslant i \leqslant n$, de même loi normale $\mathcal{N}(m, \sigma^2)$, où la valeur de m est fixée à 20, σ étant inconnu et caractérisant la précision du thermomètre.

Construire un intervalle de confiance pour σ^2 de niveau 0,99.

On supposera : $\dfrac{1}{15}\displaystyle\sum_{i=1}^{15}(x_i - 20)^2 = 18.$

Exercice 8 : Intervalle de confiance de la variance d'une loi normale d'espérance inconnue

On veut déterminer le poids P d'un objet à l'aide d'une balance à deux plateaux. Le poids marqué à l'équilibre est une v.a. X qui, compte tenu de l'imprécision, suit une loi $\mathcal{N}(P, \sigma^2)$, σ^2 étant inconnu et caractérisant la précision de la balance. On réalise 20 pesées X_i, $1 \leqslant i \leqslant 20$, du même objet. Construire un intervalle de confiance pour σ^2 de niveau 0,95.

On supposera : $\dfrac{1}{20}\displaystyle\sum_{i=1}^{20}(x_i - \overline{x})^2 = 250.$

Exercice 9 : Comparaison d'intervalles

L'épaisseur d'une tôle d'acier sortant d'un laminoir suit une loi $\mathcal{N}(m, \sigma^2)$, les paramètres m et σ^2 étant inconnus. La mesure de l'épaisseur de neuf tôles a permis de calculer $\sum_{i=1}^{9}(x_i - \overline{x})^2 = 32$. Comparer les intervalles de confiance pour σ^2 au niveau 0,95 :

- bilatéral symétrique,
- bilatéral dissymétrique, avec $\alpha_1 = 0{,}04$ et $\alpha_2 = 0{,}01$,
- unilatéral à droite, de la forme $]a; +\infty[$.

3.2.3 Cas de grands échantillons ($n > 30$)

Théorème 12 L'intervalle de confiance symétrique I_α est défini, **quelle que soit la densité de la v.a.** X, par :

$$I_\alpha = \left[\overline{x} - u_{1-\alpha/2}\frac{s_{n-1}}{\sqrt{n}}; \overline{x} + u_{1-\alpha/2}\frac{s_{n-1}}{\sqrt{n}}\right]$$

Preuve On utilise le théorème central limite $\sqrt{n}\frac{\overline{X}-m}{s_{n-1}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0;1)$.

3.3 Estimation d'une proportion p

Soit p la probabilité inconnu d'apparition d'un événement A. Pour estimer p, on réalise n épreuves correspondant au schéma binomial et on observe la valeur r de la v.a. binomiale $\mathcal{B}(n,p)$ R représentant le nombre de réalisations de l'événement A.

On sait que R suit une loi binomiale $\mathcal{B}(n;p)$ et que $\frac{R}{n}$ est un estimateur sans biais du paramètre p.

3.3.1 Cas d'un échantillon de grande taille ($n > 30$)

Utilisons la convergence en loi de la v.a. binomiale vers la v.a. normale quand $n \to +\infty$:

$$U = \frac{R - np}{\sqrt{np\,(1-p)}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0;1)\ .$$

$$\mathrm{P}\left(-u_{1-\alpha/2} < U < u_{1-\alpha/2}\right) = \gamma = 1-\alpha\ .$$

$$\mathrm{P}\left(-u_{1-\alpha/2} < \frac{R-np}{\sqrt{np\,(1-p)}} < u_{1-\alpha/2}\right) \simeq \gamma = 1-\alpha\ ....$$

On peut alors construire un intervalle de confiance de p, selon deux méthodes :

1. Méthode analytique : il s'agit de résoudre l'inéquation $\left(\frac{r-np}{\sqrt{np\,(1-p)}}\right)^2 < u^2_{1-\alpha/2}$, qui équivaut à : $(r-np)^2 - np(1-p)u^2_{\alpha/2} < 0$, et permet donc de définir l'intervalle de confiance de p au niveau $(1-\alpha)$:

$$p \in \mathrm{IC}_\alpha = \left[\frac{2rn + nu^2_{\alpha/2} - \sqrt{\Delta}}{2\left(n^2 + nu^2_{\alpha/2}\right)}; \frac{2rn + nu^2_{\alpha/2} + \sqrt{\Delta}}{2\left(n^2 + nu^2_{\alpha/2}\right)}\right]$$

où $\Delta = nu^2_{\alpha/2}\,(nz^2 + 4rn - 4r^2)$.

2. Méthode par approximation : on approche $p(1-p)$ par $\frac{r}{n}\left(1-\frac{r}{n}\right)$ et on obtient l'IC_α :

$$\boxed{p \in \mathrm{IC}_\alpha = \left[\frac{r}{n} \pm u_{1-\alpha/2}\sqrt{\frac{r(n-r)}{n^3}}\right]}$$

L'intervalle fourni par cette méthode est évidemment, moins précis que le précédent, mais converge vers celui-ci quand $n \to +\infty$. Cette méthode n'est valide que si : $0.1 < \frac{r}{n} < 0.9$.

Exercice 10 : Construction d'un intervalle de confiance pour une proportion

Un sondage sur la popularité du premier ministre indique que 51% des personnes interrogées sont favorables à sa politique. Construire un estimateur $\widehat{p_n}$, puis un intervalle de confiance au niveau 0,95 de la proportion p de français qui lui sont favorables, sachant que ce sondage a été réalisé auprès de $n = 100$ personnes (utiliser la méthode par approximation). Même question si $n = 1000$. Quelle doit être la valeur minimale de n pour que la longueur de cet intervalle soit au plus égale à 4% ? Conclure sur le peu de confiance que l'on doit accorder aux estimations des intentions de vote données dans la presse à la veille de certains seconds tours très serrés.

On peut construire des intervalles de confiance unilatéraux obtenus par la même d'approximation gaussienne :

- intervalle unilatéral à droite : $[\frac{r}{n} - u_\alpha\sqrt{\frac{r(n-r)}{n^3}}; 1]$

- intervalle unilatéral à gauche : $[0; \frac{r}{n} + u_\alpha\sqrt{\frac{r(n-r)}{n^3}}]$

3.3.2 Cas d'un échantillon de petite taille n

Il existe des tables, en fin de polycopié, qui pour les valeurs n et r permettent de déterminer :

– la valeur p_1 qui est la plus grande valeur du paramètre p telle que :

$$\mathrm{P}(R_1 \geqslant r) = \sum_{k=r}^{n} C_n^k p_1^k (1-p_1)^{n-k} \leqslant \frac{\alpha}{2} \quad \text{où } R_1 \sim \mathcal{B}(n; p_1),$$

– la valeur p_2 qui est la plus petite valeur du paramètre p telle que :

$$\mathrm{P}(R_2 \leqslant r) = \sum_{k=0}^{r} C_n^k p_2^k (1-p_2)^{n-k} \leqslant \frac{\alpha}{2} \quad \text{où } R_2 \sim \mathcal{B}(n; p_2).$$

On pourra utiliser la table de la loi binomiale (en fin d'ouvrage) de paramètre p correspondant.

Le risque de se tromper, en affirmant que l'intervalle $[p_1; p_2]$ contient la vraie valeur p, est alors au plus égal à α. D'où l'intervalle de confiance :

$$\boxed{p \in \mathrm{IC}_\alpha = [p_1; p_2]}$$

Exercice 11 : Intervalle de confiance unilatéral

A la sortie d'une chaîne de montage, 20 véhicules automobiles tirés au sort sont testés de façon approfondie. Sachant que deux d'entre eux présentent des défauts graves et doivent repasser dans la chaîne, construire un intervalle de confiance unilatéral de la forme $p < C$ de niveau 0,95 pour la proportion p de véhicules défectueux, par la méthode par approximation.

Exercice 12

Afin d'établir le profil statistique de certains malades d'un hôpital, on prélève au hasard et avec remise 100 dossiers médicaux. Malheureusement, on constate qu'une proportion p d'entre eux sont incomplets et donc inexploitables. Si on considère qu'il faut pouvoir exploiter au moins 1000 dossiers, combien faudra-t-il en prélever pour que cette condition soit réalisée avec une probabilité égale à 0,95 ?

3.4 Estimation d'un paramètre d'une population quelconque, dans le cas de grands échantillons

3.4.1 Intervalle de confiance obtenu par convergence en loi de l'E.M.V.

Si le paramètre θ de la loi de probabilité définie par $f(x;\theta)$ commune aux v.a. $X_1, \ldots, X_n$ est estimé par la méthode du M.V., on obtient l'E.M.V. $\widehat{\theta_n}$.

Dans le cas où la borne de Cramer-Rao existe, on a :

$$V_n(\theta) = \frac{1}{-n\,\mathbb{E}\left(\dfrac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right)}\ .$$

De plus, on sait que $\dfrac{\widehat{\theta_n} - \theta}{\sqrt{V_n(\theta)}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0;1)$ quand $n \to +\infty$. Ceci nous permet de construire un intervalle de confiance asymptotique de niveau de signification égal à α, à partir de :

$$IC_\alpha = [\widehat{\theta_n} - u_{\alpha/2}\sqrt{V_n(\theta)}; \widehat{\theta_n} + u_{\alpha/2}\sqrt{V_n(\theta)}]$$

3.4.2 Application

Il s'agit d'estimer le paramètre a d'une loi exponentielle décalé par translation de deux unités.

La densité commune aux $T_1, \ldots, T_n$ est :

$$f(t;a) = f_a(t) = \begin{cases} a\ e^{-a(t-2)} & \text{si } t \geqslant 2 \\ 0 & \text{si } t < 2 \end{cases}$$

On trouve que l'E.M.V. est $\widehat{a_n} = \dfrac{1}{\sum_{i=1}^{n} \dfrac{T_i - 2}{n}} = \dfrac{1}{\overline{T} - 2}$ où $\overline{T} = \dfrac{1}{n}\sum_{i=1}^{n} T_i$. Construisons l'IC$_\alpha$ du paramètre a de niveau de confiance $1 - \alpha$.

$$\frac{\widehat{a_n} - a}{\sqrt{V_n(a)}} = \frac{\dfrac{1}{\overline{T} - 2} - a}{\dfrac{a}{\sqrt{n}}} = \frac{\sqrt{n}}{a\left(\overline{T} - 2\right)} - \sqrt{n}\ ,$$

d'où :

$$\text{IC}_\alpha = \left[\frac{\sqrt{n}}{\left(\overline{t} - 2\right)\left(u_{1-\alpha/2} + \sqrt{n}\right)}; \frac{\sqrt{n}}{\left(\overline{t} - 2\right)\left(-u_{1-\alpha/2} + \sqrt{n}\right)}\right].$$

Exercice 13

Un atelier produit des composants électroniques dont la durée de vie est décrite par la variable aléatoire X, de densité de Weibull $f(x) = \frac{2}{\theta} x e^{-\frac{x^2}{\theta}}$ pour tout $x \geqslant 0$, où θ est un paramètre inconnu, strictement positif.

1) Déterminer l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$ de θ, associé à l'échantillon $(X_1, \ldots, X_n)$.

2) Démontrer qu'il est sans biais et convergent.

3) Supposons que n est grand (supérieur à 50), démontrer que la v.a $\widehat{\theta}_n$ est de loi normale, de moyenne θ et de variance $\frac{\theta^2}{n}$.

4) En déduire un intervalle de confiance de θ, au niveau de confiance fixé à 95%. Application numérique : $n = 1000$, $\widehat{\theta}_{1000} = 3$.

Chapitre 2

Tests d'hypothèse

1 Principes généraux

1.1 Exemple introductif : le "test binomial"

Problème posé lors d'un contrôle de fabrication : une machine fabrique des pièces de telle sorte que chaque pièce peut être défectueuse avec la probabilité p, indépendamment de la qualité des autres. Le paramètre p est inconnu et peut varier car la machine peut se dérégler au cours du temps. On souhaite donc intervenir assez rapidement quand la machine se dérègle. On admet que la machine fonctionne de façon satisfaisante tant que $p \leq p_0$ (où p_0 est une valeur fixée par des normes techniques), mais doit être révisée si $p > p_0$.

Problème de décision Choisir entre les deux hypothèses suivantes :

$$\begin{cases} \text{Hypothèse } H_0 : p \leq p_0 \text{ (dite } \textbf{\textit{hypothèse nulle}}) ; \\ \text{Hypothèse } H_1 : p > p_0 \text{ (dite } \textbf{\textit{hypothèse alternative}}) \end{cases}$$

Le paramètre p étant inconnu, au moment où l'on veut contrôler la machine, il faut donc à partir d'un échantillon de taille n définir une *règle de décision* qui nous permette de choisir H_0 ou H_1. On prélève au hasard n pièces et on définit la v.a. X_i par :

$$X_i = \begin{cases} 1 \text{ si la } i\text{ème pièce est défectueuse avec la probabilité } \mathrm{P}(X_i{=}1) = p ; \\ 0 \text{ si la } i\text{ème pièce est bonne avec avec la probabilité } \mathrm{P}(X_i{=}0) = q = 1 - p \end{cases}$$

Sur les n pièces prélevées, on note r le nombre de pièces défectueuses, défini comme l'observation de la variable aléatoire R égale à $\sum_{i=1}^{n} X_i$, et de loi binomiale $B(n, p)$.

Règle de décision Tester l'hypothèse H_0 contre l'hypothèse H_1, c'est adopter une règle de décision qui tienne compte des résultats fournis par l'échantillon. Dans notre cas, on décidera de rejeter H_0 si le nombre observé r de pièces défectueuses est trop grand.

De manière plus précise, "on décide de rejeter H_0 si $R \geq r_0$", où r_0 est un nombre à déterminer.

Le rejet de H_0 ($p \leq p_0$) revient à accepter H_1 et à décider de réviser la machine déréglée.

Dans le cas contraire, si $R < r_0$, on ne rejette pas l'hypothèse H_0 et aucun réglage ne sera fait sur la machine.

Cette règle de décision comporte deux ***risques d'erreurs*** :

- Le risque de rejeter H_0 alors qu'elle est vraie ; il s'agit du risque de trouver dans un échantillon de taille n un nombre r de pièces défectueuses supérieur à r_0 alors que la machine fonctionne correctement (c'est-à-dire : $p \leq p_0$).
Ce risque est appelé **risque de première espèce**.

- Le risque de ne pas rejeter H_0 alors que H_1 est vraie ; il s'agit du risque de trouver dans un échantillon de taille n un nombre r de pièces défectueuses inférieur à r_0 alors que la machine est déréglée ($p > p_0$).
Ce risque est appelé **risque de deuxième espèce** et correspond au risque de ne pas régler la machine alors qu'elle est déréglée, ce qui n'est pas sans gravité.

Le test de H_0 contre H_1 sera totalement déterminé :

- si on connaît le nombre r_0 et qu'on en déduise la probabilité fonction de p de chacun des 2 risques ;

- ou bien, si on se fixe a priori la probabilité du risque de première espèce et qu'on en déduise le nombre r_0 et la probabilité du risque de deuxième espèce ; cette éventualité étant la plus fréquente en pratique.

Calcul des probabilités des deux risques d'erreur pour n et r_0 fixés
Soit $\alpha(p)$ la probabilité de rejeter H_0, pour une valeur p, c'est-à-dire :
$\alpha(p) = \mathrm{P}$(obtenir au moins r_0 pièces défectueuses sur les n pièces)
$= \mathrm{P}(R \geq r_0) = \sum_{k=r_0}^{n} \mathrm{P}(R = k) = \sum_{k=r_0}^{n} C_n^k p^k (1-p)^{n-k}$

L'**erreur de première espèce** est alors :

P(rejeter H_0 alors que H_0 est vraie) $= \alpha(p)$ pour $p \leq p_0$

Erreur de deuxième espèce :

$\forall p > p_0$, P(ne pas rejeter H_0 alors que H_1 est vraie) $= 1 - \alpha(p)$, désigné aussi par $\beta(p)$.

Exemple numérique : pour tester $H_0 : p \leq 4\%$ (donc $p_0 = 0,04$) contre $H_1 : p > 4\%$, on calcule $\alpha(p)$ pour les 3 cas suivants :

- $n = 20$ et $r_0 = 2$
- $n = 20$ et $r_0 = 3$
- $n = 40$ et $r_0 = 2$

Les calculs sont donnés dans le tableau numérique suivant et on peut faire quelques remarques :
- on diminue un risque d'erreur pour augmenter l'autre quand on fait varier r_0 (n égal à 20)
- on augmente la puissance ($\alpha(p)$ pour $p > p_0$) si on augmente n (r_0 étant égal à 3).

1.2 Cas général

Soient $(x_1, ..., x_n)$ les observations de l'échantillon $(X_1, ..., X_n)$, associé au modèle statistique $(R^n, \mathcal{B}(R^n), \mathcal{P})$, où $\mathcal{P}$ est une famille de lois de probabilités.

Définition 15 On appelle ***hypothèse*** tout sous-ensemble de P.

Exemple 4 Soit un échantillon $(X_1, ..., X_n)$ dont on ignore la loi commune ; dans ce cas, $\mathcal{P}$ sera l'ensemble de toutes les lois possibles.

En pratique, on est confronté à des hypothèses moins générales, portant seulement sur les paramètres inconnus de lois connues.

Définition 16 On dit qu'une hypothèse est **paramétrique** si la loi de $(X_1, ..., X_n)$ est connue, mais dépend d'un paramètre θ inconnu, scalaire ou vectoriel.

Exemple 5 Soit un échantillon $(X_1, ..., X_n)$ de loi commune gaussienne, $X_i \sim N(\mu,\ \sigma^2 = 3)$. La famille $\mathcal{P}$ de lois de probabilités de $(X_1, ..., X_n)$ est donc connue au paramètre μ près.

Les hypothèses paramétriques sont décrites par des égalités ou des inégalités.

Définition 17 Une hypothèse paramétrique est dite **simple** si une seule valeur de θ est testée, par exemple $\mu = 1$. Dans le cas contraire, elle est dite **composite** ou **multiple**.

Test d'hypothèses :

Il existe deux hypothèses :

- H_0 (la valeur attendue de θ est θ_0) appelée **hypothèse nulle**
- H_1 (un ensemble de valeurs distinctes de θ_0 ou bien "non H_0"), appelée **hypothèse alternative**.

		Réalité	
		H_0 vraie	H_1 vraie
Décision retenue	H_0 vraie	$1 - \alpha$	$\beta = \text{P(accepter } H_0 \text{ à tort)}$
	H_1 vraie	$\alpha = \text{P(rejeter } H_0 \text{ à tort)}$	$1 - \beta = \eta$

Définition 18 La quantité α représente la probabilité de rejeter H_0 alors qu'elle est vraie, appelée **risque de première espèce**.

Définition 19 La quantité β représente la probabilité de rejeter H_1 alors qu'elle est vraie, appelée **risque de seconde espèce**.

La quantité $1 - \beta$ s'appelle la **puissance du test**.

Lorsqu'on est en présence d'un tel test, on cherche à minimiser les risques de première et seconde espèces. Auparavant, il faut construire une *règle de décision* qui va nous permettre de choisir entre H_0 et H_1. Cette règle de décision est très importante puisqu'elle va induire la forme du calcul de α et de β. Pour minimiser ces deux valeurs, il faut donc jouer sur cette règle de décision. Nous verrons plus loin qu'il n'est pas possible de diminuer simultanément les risques de première espèce et de deuxième espèce, qui varient en sens inverse.

Définition 20 On appelle **statistique du test** une statistique $\phi(X_1, ..., X_n)$ dont la valeur observée $\phi(x_1, ..., x_n)$ permettra de décider ou non le rejet de H_0.

Définition 21 On appelle **région critique du test** l'ensemble W des observations $(x_1, ..., x_n)$ conduisant au rejet de H_0 :

$$W = [(x_1, ..., x_n) \in R^n \ / \ \phi(x_1, ..., x_n) \in D] = \phi^{-1}(D)$$

On rejette H_0 si et seulement si $\phi(x_1, x_2, ..., x_n) \in D$.

Remarque 4 *Si on rejette H_0, c'est qu'au vu des observations il est improbable que H_0 soit vraie ; mais si on décide d'accepter H_0, cela ne signifie pas que H_0 soit vraie. Généralement, H_0 est une hypothèse solidement étayée, qui est vraisemblable. La valeur de α sera prise d'autant plus petite que la gravité conséquente au risque de première espèce est grande.*

1.3 Test d'une hypothèse simple contre une hypothèse simple

Test de Neyman-Pearson

Soit X une v.a de densité $f(.,\theta)$, où $\theta \in R$ est le paramètre inconnu. On note $L(x,\theta)$ (où : $\underline{x} = (x_1, ..., x_n)$) la vraisemblance de l'échantillon $(X_1, ..., X_n)$.

On souhaite tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$.

Dans ce cas, les deux probabilités d'erreur se calculent de la manière suivante :

$\alpha = \mathrm{P}(\text{rejeter } H_0 \ / \ \theta = \theta_0)$;
$\beta = \mathrm{P}(\text{ne pas rejeter } H_0 \ / \ \theta = \theta_1)$.

Théorème 13 (Neyman-Pearson) Pour un risque de première espèce fixé, le test le plus puissant pour tester H_0 contre H_1 est donné par la règle de décision :

$$\text{on rejette } H_0 \text{ si } \frac{L(x,\theta_1)}{L(x,\theta_0)} > \lambda_\alpha$$

où $L(x,\theta_i)$ est la vraisemblance da la variable aléatoire X sous θ_i et λ_α une constante qui dépend du niveau α du test.
Ce test est plus puissant que tout autre test, pour un seuil donné. Il est dit ***uniformément le plus puissant*** (U.M.P).

Remarque 5 *L'interprétation de ce résultat est simple : pour un risque de première espèce α fixé, cette règle de décision parmi toutes les règles possibles, est celle qui permet d'obtenir la plus forte puissance. Mais attention, ce résultat ne veut pas dire que les deux valeurs α et β sont nécessairement proches de 0 : il affirme que pour une valeur de α, la seule règle de décision qui permette d'avoir la plus petite valeur pour β est la règle de décision de Neyman-Pearson.*

Pour mettre en pratique ce théorème, l'idée est de procéder par équivalences successives : on fait passer dans le membre de droite de l'inégalité tout ce qui dépend de θ_0 et de θ_1 ; il ne reste alors plus à gauche qu'une fonction des observations seules : $g(\underline{x})$. La v.a $g(\underline{X})$ est appelée *statistique de test.*

1. L'ensemble des résultats qui vont suivre sont aussi valables dans le cas où la règle de décision nous fournit l'inégalité dans l'autre sens.

2. La loi de $T_n = g(\underline{X})$ dépend du paramètre θ. Soit $h(\underline{x};\theta)$ la densité de T_n. "Sous H_0" veut alors dire que nous prenons comme densité pour T_n, la densité $h(\underline{x};\theta_0)$.

Calcul de la valeur de π_α : trois cas sont possibles.

- Soit nous connaissons la loi de $T_n = g(X_1, ..., X_n)$ sous H_0 et la forme explicite de l'inverse de sa fonction de répartition F_{T_n}. L'équation (2.1) s'écrit alors :
$$F_{T_n}(\pi_\alpha) = \alpha \Rightarrow \pi_\alpha = F_{T_n}^{-1}(\alpha) \ .$$

- Soit nous connaissons la loi de $T_n = g(X_1, ..., X_n)$ sous H_0 et nous possédons la table de cette loi : on se reporte alors à la fiche technique sur les tables de lois pour voir comment calculer la valeur de π_α en fonction de α ;

- Soit nous ne connaissons pas la loi de $T_n = g(X_1, ..., X_n)$: dans ce cas, nous allons utiliser, quand ce sera possible, l'approximation de la loi normale qui découle du théorème de la limite centrale. Pour cela, il faut que la statistique T_n s'écrive sous la forme d'une somme de v.a indépendantes de même loi et de variance finie.

Calcul de la puissance du test

Nous savons d'après le tableau de départ des risques de première et de seconde espèces que :

$$\mathrm{P}(\text{rejeter } H_0 \text{ alors que } H_1 \text{ est vraie}) = \mathrm{P}_{H_1}(\text{rejeter } H_0) = 1 - \beta \ .$$

Supposons que la règle de décision de Neyman et Pearson nous ait fourni l'équivalence suivante :

$$\text{on rejette } H_0 \Leftrightarrow g(\underline{X}) < \pi_\alpha \ .$$

On a l'égalité suivante :

$$\mathrm{P}_{H_1}(g(\underline{X}) < \pi_\alpha) = 1 - \beta \ .$$

La quantité π_α étant donnée, on détermine la puissance $1 - \beta$.

Exemple : test de comparaison d'une moyenne à chacune de deux valeurs plausibles

Soit une population gaussienne dont on connaît la variance $\sigma^2 = 16$, mais dont la moyenne m est inconnue. Supposons que l'on ait de très bonnes raisons de penser que la moyenne m est égale à 20, mais qu'il n'est pas impossible qu'elle soit égale à 22. On fait alors les deux hypothèses suivantes :

$$\begin{cases} \text{Hypothèse } H_0 : m = m_0 = 20 \, ; \\ \text{Hypothèse } H_1 : m = m_1 = 22 \end{cases}$$

On tire un échantillon de 25 individus de moyenne $\overline{x} = 20{,}7$. La variable aléatoire $\overline{X}$ associée à la moyenne est de loi $N(m; \sigma^2/n) = N(m; 16/25)$. Les fonctions de vraisemblance sous

H_0 et H_1 sont données par :

$$L_0((\underline{x}) \text{ sachant que } m = m_0 = 20) = \prod_{i=1}^{25} f(x_i; 20)$$

$$L_1((\underline{x}) \text{ sachant que } m = m_1 = 22) = \prod_{i=1}^{25} f(x_i; 22)$$

D'après le test de Neyman-Pearson, on a la règle de décision suivante :

$$\begin{aligned}
\text{on rejette } H_0 \quad &\Longleftrightarrow \quad \frac{L_1((x_1, \ldots, x_n) \text{ sachant que } m = m_1 = 22)}{L_0((x_1, \ldots, x_n) \text{ sachant que } m = m_0 = 20)} > \lambda_\alpha \\
&\Longleftrightarrow \quad \frac{\frac{1}{(2\pi)^{n/2}\sigma^n}\ e^{-\sum_{i=1}^n \frac{(x_i-22)^2}{2\sigma^2}}}{\frac{1}{(2\pi)^{n/2}\sigma^n}\ e^{-\sum_{i=1}^n \frac{(x_i-20)^2}{2\sigma^2}}} > \lambda_\alpha \\
&\Longleftrightarrow \quad \overline{x} > \frac{\sigma^2}{n(m_1 - m_0)}\ \log \lambda_\alpha\ + \frac{m_0 + m_1}{2} \equiv \pi_\alpha.
\end{aligned}$$

Tous calculs faits, on obtient : $\pi_\alpha = 21 + 0,32 \log \lambda_\alpha$.

$$\begin{aligned}
\text{Risque de première espèce } : \alpha \quad &\stackrel{déf}{=} \quad \mathrm{P}(\text{choisir} H_1 / H_0 \text{ vraie}) \\
&= \quad \mathrm{P}(\overline{X} > \pi_\alpha / \overline{X} \text{de loi } N(20; 16/25)); \\
\text{Risque de deuxième espèce } : \beta \quad &\stackrel{déf}{=} \quad \mathrm{P}(\text{choisir} H_0 / H_1 \text{ vraie}) \\
&= \quad \mathrm{P}(\overline{X} \leq \pi_\alpha / \overline{X} \text{de loi } N(22; 16/25)).
\end{aligned}$$

On décide d'accepter H_0 si la réalisation $\overline{x}$ de $\overline{X}$ est inférieure ou égale à π_α ; inversement, on rejette H_0 si $\overline{x} > \pi_\alpha$.

Pour déterminer la région critique, donnons à α une valeur, par exemple 5%. On a donc :

$$5\% = P(\overline{X} > \pi_{5\%}/m = 20) = P\left(\frac{\overline{X} - 20}{0,8} > \frac{\pi_{5\%} - 20}{0,8}\right)$$

ce qui entraîne que $\pi_{5\%} = 21{,}32$.

D'où la région critique : $W_{5\%} = \{\underline{x} \in R^n / \overline{x} > 21,32\}$

Nous pouvons maintenant faire le choix de l'hypothèse H_0 ou H_1 avec notre règle de décision. Comme $\bar{x} = 20{,}7$ est inférieur à $\pi_{5\%}$, on décide d'accepter $H_0 : m = 20$.

Pour conclure l'étude de cet exemple, il reste à calculer le risque de deuxième espèce β. Nous avons par définition :

$$\begin{aligned}
\beta &= P(\overline{X} \leq \pi_{5\%}/m = 22) = P\left(\frac{\overline{X} - 22}{0,8} \leq \frac{21,32 - 22}{0,8}\right) \\
&= P(N(0;1) \leq -0,85) = 1 - P(N(0;1) > -0,85) \\
&= 0,197.
\end{aligned}$$

En conclusion, le risque d'avoir choisi H_0 alors que H_1 est vraie est égal à 19,7%. La puissance η du test ($\eta = 1 - \beta$) est donc de l'ordre de 80 %.

Exercice 14 : Test d'une proportion

Avant le second tour d'une élection présidentielle, le candidat D. Magog commande un sondage à une société spécialisée pour savoir s'il a une chance d'être élu. Sachant que la proportion p des électeurs qui lui sont favorables ne peut prendre que deux valeurs, ceci conduit au test :

$$\begin{cases} H_0 : p = p_0 = 0,48; \\ H_1 : p = p_1 = 0,52. \end{cases}$$

1. Quelle est la signification du choix de $p = 0,48$ comme hypothèse nulle ?
2. Déterminer la région critique du test de risque $\alpha = 0{,}10$ si le sondage a été effectué auprès de $n = 100$ électeurs. Que peut-on penser du résultat ? (utiliser le rapport des vraisemblances).
3. Indiquer comment varient la région critique et la puissance $\eta = 1 - \beta$ en fonction de n. Calculer les valeurs de η pour $n = 100$, 500 et 1000.
4. On considère maintenant le test :

$$\begin{cases} H_0 : p = 0,49; \\ H_1 : p = 0,51. \end{cases}$$

Calculer la taille d'échantillon minimum n_0 pour que la puissance soit supérieure à 0,90. Quelle est alors sa région critique ?

Exercice 15

Soit X la variable aléatoire associée à la durée de fonctionnement d'un composant électronique ; X est une variable de Weibull :

- de densité $f(x;\theta,\lambda) = (\lambda/\theta).x^{\lambda-1}\ \exp\{(-1/\theta)\ x^{\lambda}\}, \forall x \geq 0$
- de fonction de répartition $F(x;\theta,\lambda) = 1 - \exp\{(-1/\theta)\ x^{\lambda}\}$

On suppose que λ est connu et θ **inconnu**.

1. Démontrer que $E(X^{\lambda}) = \theta$
2. Déterminer l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de θ, en fonction de l'échantillon $(X_1, \ldots, X_n)$.
3. Est-il sans biais ? Convergent ? (indication : Var $(X^{\lambda}) = \theta^2$).
4. On souhaite tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1 (\theta_1 > \theta_0)$. Déterminer la statistique T_n correspondant à ce test à l'aide du théorème de Neyman-Pearson.
5. Vérifier que $Y = (2/\theta)X^{\lambda}$ est une v.a χ^2 à deux degrés de liberté (on se contentera de démontrer que la fonction de répartition de Y est égale à $1 - \exp\{-y/2\}$ pour y positif).
6. En déduire que sous l'hypothèse H_0, $(2/\theta_0)T_n$ est un χ^2 à $2n$ degrés de liberté. En déduire la borne critique (séparant la région critique de la région d'acceptation) pour un risque de première espèce α fixé.
7. Doit-on rejeter H_0, au risque $\alpha = 5\%$ si $\theta_0 = 0,5$; $\theta_1 = 1$; $n = 10$; $\lambda = 3$ et $\sum_{i=1}^{10} x_i^3 = 18$.

1.4 Test d'une hypothèse simple contre une hypothèse composite

Exemple 6 Sous les mêmes hypothèses de normalité des populations que l'exemple précédent, on teste cette fois :

$$\begin{cases} H_0 : m = 20\ ; \\ H_1 : m > 20. \end{cases}$$

On dispose d'un échantillon de taille $n = 25$ et on fixe le risque de première espèce $\alpha = 0,02$. On a alors :

$$\begin{aligned} \alpha &= 0,02 = P(\text{décider } H_1\ / H_0 \text{ vraie}) = P(\overline{X} \geq \pi_\alpha\ /\ m = 20) \\ &= P\left(\frac{\overline{X} - 20}{0,8} \geq \frac{\pi_\alpha - 20}{0,8}\right) = P\left(N(0;1) \geq \frac{\pi_\alpha - 20}{0,8}\right) \end{aligned}$$

On en déduit par lecture de la table gaussienne centrée réduite :
$(\pi_\alpha - 20)/0,8 = 2,05$; donc, $\pi_\alpha = 21,64$, indépendant de m.

On décide : $m > 20$ si $\overline{x} \geq 21,64$ et $m = 20$ si $\overline{x} < 21,64$

Que peut-on dire de l'erreur de deuxième expèce β ? Elle dépend de la valeur réelle de m :

$$m \to \beta(m) = P(\overline{X} < \pi_\alpha\ /\ m > 20).$$

On conçoit aisément que plus la valeur de m est grande, moins on a de chances de décider H_0 et plus β est petit.

Généralement, soit X une v.a de densité $f(x,\theta)$ où $\theta \in R$ est inconnu. On veut tester :

$$\begin{cases} H_0 : \theta = \theta_0 \ ; \\ H_1 : \theta \in \Theta_1, \text{ sous ensemble de } R, \text{ qui ne contient évidemment pas } \theta_0. \end{cases}$$

Définition 22 On appelle :

- **fonction puissance** $\eta : \theta \in \Theta_1 \to \eta(\theta) = \mathrm{P}(\text{rejet de } H_0 \ / \ \theta \in \Theta_1)$.
- **fonction d'erreur de** $2^{ème}$ **espèce**, ou **fonction d'efficacité** β :
  $\theta \in \Theta_1 \to \beta(\theta) = \mathrm{P}(\text{acceptation de } H_0 \ / \ \theta \in \Theta_1) = 1 - \eta(\theta)$

Recherche du meilleur test

Soit l'hypothèse alternative simple $H_1' : \theta = \theta_1$, avec $\theta_1 \in \Theta_1$.
On cherche le test uniformément le plus puissant (U.M.P.), donné par le théorème de Neyman-Pearson à un niveau α fixé :

$$\begin{cases} H_0 : \theta = \theta_0; \\ H_1' : \theta = \theta_1. \end{cases}$$

Deux cas peuvent se présenter :

- 1^{er} **cas :** le test obtenu ne dépend pas de la valeur θ_1 choisie dans Θ_1 : il est alors U.M.P.
- $2^{ème}$ **cas :** le test dépend du choix de θ_1 : il n'existe pas de test optimal et le choix du test fera une large place à l'intuition.

Exercice 16

(suite du test présenté comme exemple au début du chapitre 2.1.4)

Déterminer les valeurs de β dans le cas des diverses valeurs de m (22, 24, 26, 28), et tracer le graphe de la fonction $m \to \beta(m)$.

Remarque 6

1. *On démontre que la puissance $\eta = 1 - \beta$ d'un test U.M.P. est supérieure à la puissance de tout autre test.*
2. *Un test n'est pas nécessairement U.M.P. : c'est le cas du test $H_0 : m = m_0$ contre $H_1 : m \neq m_0$. Dans ce cas, la région critique est égale à l'union des régions critiques associées aux deux tests :*
   H_0 *contre* H_1' *:* $m > m_0$ *et* H_0 *contre* H_1'' *:* $m < m_0$.

Poursuivons l'étude de l'exemple précédent dans le cas où $H_1 : m \neq 20$, avec un risque $\alpha = 4\%$.

L'hypothèse H_1 est équivalente à $H_1' : m < 20$ ou $H_1'' : m > 20$.
On va tester chaque hypothèse alternative H_1' et H_1'' avec un risque de 2%.
Le cas H_0 contre H_1'' a été traité dans l'exemple précédent où on a déterminé la région critique $[21{,}64\,;\,+\infty[$ au risque de première espèce $\alpha = 2\%$.
Pour des raisons évidentes de symétrie, le test H_0 contre H_1' donne la région critique

$$]-\infty; 18,36]$$

On accepte H_0 si $18,36 < \bar{x} < 21,64$ avec un risque $\alpha = 4\%$. La région d'acceptation de H_0 est donc égale à l'intersection des régions d'acceptation de chacun des deux tests. Le risque de seconde espèce β dépend de la valeur de m : le test n'est donc pas U.M.P. (voir Exercice 16).

Exercice 17 : Test localement optimal

(extrait de l'ouvrage : Modélisation, probabilités et statistique de B. Garel, éd. Cepadues)

Un des problèmes constants du traitement du signal est d'identifier la présence d'un signal dans du bruit. Pour cela, on utilise une statistique de test appelée détecteur. On suppose ici que les observations suivent le modèle :

$$X_j = \theta s_j + Z_j, \quad j = 1, \ldots, n$$

où $\theta \geq 0$, les s_j sont des réels connus et les Z_j sont des v.a indépendantes de même loi centrée de variance σ^2 connue et de densité f positive et dérivable sur R.
On veut tester l'hypothèse $H_0 : \theta = 0$ (absence de signal) contre $H_1 : \theta > 0$ (présence de signal).
On note $L(x_1, \ldots, x_n; \theta)$ la vraisemblance calculée sous l'hypothèse H_1 et
$L_0(x_1, \ldots, x_n)$ celle calculée sous H_0. On appelle détecteur localement optimal le détecteur caractérisé par la règle suivante faisant intervenir la dérivée logarithmique de $\underline{L}(x; \theta)$:

$$\text{on rejette } H_0 \text{ si } \frac{\frac{\partial L(x_1,\ldots,x_n;\theta)}{\partial \theta}|_{\theta=0}}{L_0(x_1, \ldots, x_n)} > \pi_\alpha$$

où C_α dépend de la probabilité de fausse alarme α.

1. Caractériser le détecteur localement optimal en fonction des V.A.R.

   $-\dfrac{f'(X_j)}{f(X_j)}$, $j = 1, \ldots, n$. Pour cela, on remarquera que les X_i sont indépendantes et que leur densité est la fonction $x \to f(x - \theta s_j)$.
2. Préciser ce détecteur dans le cas où les Z_j sont gaussiennes $N(0; \sigma^2)$.
3. A l'aide du théorème de Neyman et Pearson, caractériser le détecteur le plus puissant du test $H_0 : \theta = 0$ contre $H_1' : \theta = \theta_1 > 0$ dans le cas gaussien ci-dessus. En déduire l'expression du test uniformément le plus puissant pour tester H_0 contre H_1. Que constate-t-on ?

Exercice 18

(cf : ouvrage précédemment cité)

En traitement du signal (modèles sismiques, radars, ...), on constate un intérêt croissant pour les modèles multiplicatifs du type

$$X_j = (1 + \theta s_j)Y_j \ , \quad j = 1, \ldots, n$$

où les X_j sont les observations, $(s_j)_{j\geq 1}$ est une suite de réels connus, $\theta \geq 0$ est un paramètre positif inconnu, pour tout j, $1 + \theta s_j > 0$ et Y_j représente un bruit, qualifié ici de multiplicatif. On suppose que les Y_j sont des V.A.R. indépendantes et de même loi de densité f, dérivable sur R.
On veut tester l'hypothèse :

$$H_0 : \theta = 0 \ \text{ contre } \ H_1 : \theta > 0$$

On note $L(x_1, \ldots, x_n; \theta)$ la vraisemblance de $(x_1, \ldots, x_n)$ sous H_1 et $L_0(x_1, \ldots, x_n)$ celle calculée sous H_0. On retient comme statistique celle du test localement optimal caractérisé par la règle de décision :

$$\text{on rejette } H_0 \ \text{ si } \frac{\frac{\partial L(x_1,\ldots,x_n;\theta)}{\partial \theta}|_{\theta=0}}{L_0(x_1, \ldots, x_n)} > \lambda_\alpha$$

où la constante λ_α dépend du risque de première espèce α.

1. Montrer que :
$$L(x_1, \ldots, x_n; \theta) = \prod_{j=1}^{n} \frac{1}{1 + \theta s_j} \ f\left(\frac{x_j}{1 + \theta s_j}\right)$$

2. Montrer que le test localement optimal se ramène à :
$$\text{on rejette } H_0 \ \text{ si } \sum_{j=1}^{n} -s_j X_j \ \frac{f'(X_j)}{f(X_j)} > C_\alpha$$
où C_α dépend de λ_α et des s_j. Penser à utiliser la dérivée logarithmique de la vraisemblance.

3. On appelle *loi normale généralisée* une loi dont la densité f s'écrit :
$$y \longrightarrow f(y) = \frac{\alpha}{2 \ \beta \ \Gamma(\frac{1}{\alpha})} \ \exp\left(-\left|\frac{y - \mu}{\beta}\right|^\alpha\right), \ \text{où } \alpha > 0, \ \beta > 0, \ \text{et } \mu \in R$$
Lorsque $\alpha = 2$, on retrouve une loi normale.
On suppose que les Y_i suivent une loi normale généralisée avec $\alpha > 1, \beta = 1$ et $\mu = 0$.
Montrer que la statistique de test s'écrit alors :
$$T_n = \alpha \sum_{j=1}^{n} s_j |X_j|^\alpha.$$

Exercice 19 : Problème d'extinction d'une population ; capture et recapture
(extrait de l'ouvrage précédemment cité)

On souhaite estimer le nombre N de poissons d'une espèce donnée vivant dans un lac. Pour cela, on effectue une première capture de n poissons que l'on bague et que l'on remet dans l'étang.
Une semaine plus tard, on capture à nouveau n poissons. On note K la V.A.R. qui représente le nombre de poissons bagués recapturés. On suppose que ces recaptures sont effectuées avec remise (autrement dit que la proportion de poissons bagués reste constante pendant la recapture).

1. Quelle est la loi de la v.a K ?
2. en déduire que K/n est une estimateur de n/N. En déduire un estimateur T_n de $1/N$. Calculer son espérance et sa variance. Proposer enfin un estimateur de N. On supposera que N est suffisamment grand pour que K n'ait aucune chance d'être nul.
3. On décide de n'ouvrir la pêche à cet endroit qu'en étant à peu près sûr que le nombre de poissons présents est au moins égal à 1000.
   Pour cela, on souhaite effectuer un test de niveau $\alpha = 5\%$ (risque de première espèce) permettant de tester les hypothèses :

   $$H_0 \ : \ N = 1000 \ \text{ contre } \ H_1 \ : \ N < 1000.$$

   Ecrire un problème équivalent portant sur $1/N$. Montrer que T_n peut être utilisée comme statistique de test avec comme règle de décision :

   $$\text{on rejette } H_0 \text{ si } T_n > u_\alpha$$

   où $\mathrm{P}(T_n > u_\alpha) = 5\%$. Utiliser une approximation normale de la loi binomiale pour calculer la valeur critique u_α.
4. Montrer que :
   $$u_\alpha = \frac{(N-n)^{1/2}}{nN}\ t_{0,95} + \frac{1}{N}\ ,\ N = 1000$$
   où $t_{0,95}$ est le quartile d'ordre 0,95 de la loi normale $N(0;1)$.

Application numérique : on bague 100 poissons. On en retrouve 15 bagués, lors de la recapture. A combien estime-t-on le nombre de poissons dans l'étang ? La pêche sera-t-elle ouverte ?

1.5 Test d'une hypothèse composite contre une hypothèse composite

Soient deux hypothèses :

$$\begin{cases} H_0 : \theta \in \Theta_0 \ ; \\ H_1 : \theta \in \Theta_1 \ . \end{cases}$$

où Θ_0 et Θ_1 sont des sous-ensembles disjoints de R.

Dans ce cas, les deux risques d'erreur α et β dépendent de θ. On a :

$\alpha(\theta) = \mathrm{P}(\text{rejeter } H_0 \ / \ \theta \in \Theta_0)$ erreur de $1^{ère}$ espèce ,
$\beta(\theta) = \mathrm{P}(\text{ne pas rejeter } H_0 \ / \ \theta \in \Theta_1)$ erreur de $2^{ème}$ espèce.

Définition 23 On appelle **niveau ou seuil de signification** du test, la quantité :

$$\alpha = \sup_{\theta \in \Theta_0} \alpha(\theta) = \sup_{\theta \in \Theta_0} \mathrm{P}(\text{rejeter} H_0 \ / \ \theta \in \Theta_0)$$

Il est difficile dans le cas général de déterminer le test optimal. Dans la plupart des cas usuels, on testera une hypothèse simple (H_0) contre une hypothèse simple ou multiple (H_1).

2 Tests pour échantillons gaussiens

Lorsque les études préalables ont permis de considérer que les variables étudiées suivent des lois normales ou approximativement normales, on dispose de tests permettant de tester des hypothèses portant sur les moyennes ou les variances de ces lois normales.

Tous les tests seront de seuil égal à α fixé ou asymptotiquement égal à α.

2.1 Tests de comparaison d'un seul paramètre à une valeur ou un ensemble de valeurs données

Soit $(X_1, \ldots, X_n)$ un échantillon de v.a indépendantes de même loi gaussienne $N(\mu; \sigma^2)$ où μ et σ^2 sont inconnus.

Le lecteur se reportera aux rappels (du paragraphe 3.2.1 du chapitre précédent) sur les propriétés des v.a $\overline{X}, S^2$, omniprésentes dans ce qui suit.

2.2.1.1 Comparaison de la moyenne μ à une valeur donnée μ_0

Dans le cas où le paramètre σ^2 est connu, pour tester l'hypothèse $H_0 \ : \ \mu = \mu_0$ contre $H_1 \ : \ \mu = \mu_1$, on a vu que le test de Neyman-Pearson (U.M.P.) nous fournit la règle de décision suivante :

- dans le cas où $\mu_1 > \mu_0$: on rejette H_0 ssi $\overline{X} > \pi_\alpha$, borne critique
- dans le cas où $\mu_1 < \mu_0$: on rejette H_0 ssi $\overline{X} < \pi_\alpha$, borne critique

Dans le cas où le paramètre σ^2 est inconnu, on utilise la statistique :

$$\sqrt{n}\,\frac{\overline{X} - \mu_0}{S_{n-1}} = \frac{\sqrt{n}\,(\frac{1}{n}\sum_{i=1}^n X_i - \mu_0)}{\sqrt{\sum_{i=1}^n \frac{(X_i - \overline{X})^2}{n-1}}} \text{ à la place de } \sqrt{n}\,\frac{\overline{X} - \mu_0}{\sigma}$$

qui suit la loi de Student à $(n-1)$ degrés de liberté quand H_0 est vraie.

On construit donc les tests suivants, appelés *tests de Student*, dans chacun des cas suivants (au seuil α) :

1^{er} **cas :** le test unilatéral pour tester $\begin{cases} H_0 : \mu = \mu_0 \\ H_> : \mu > \mu_0 \end{cases}$ est défini par la règle de décision suivante :

$$\boxed{\text{on rejette H}_0 \text{ ssi } \frac{\sqrt{n}\,(\overline{X} - \mu_0)}{S} > t_{n-1;\,\alpha}}$$

Ce test est également utilisé pour tester $\begin{cases} H_0 : \mu \leqslant \mu_0 \\ H_> : \mu > \mu_0 \end{cases}$

$2^{ème}$ **cas :** le test unilatéral pour tester $\begin{cases} H_0 : \mu = \mu_0 \\ H_< : \mu < \mu_0 \end{cases}$ est défini par la règle de décision suivante :

$$\boxed{\text{on rejette H}_0 \text{ ssi } \frac{\sqrt{n}\,(\overline{X} - \mu_0)}{S} < -t_{n-1;\,\alpha}}$$

Ce test est également utilisé pour tester $\begin{cases} H_0 : \mu \geqslant \mu_0 \\ H_< : \mu < \mu_0 \end{cases}$

$3^{ème}$ **cas :** le test bilatéral pour tester $\begin{cases} H_0 : \mu = \mu_0 \\ H_{\neq} : \mu \neq \mu_0 \end{cases}$ est défini par la règle de décision suivante :

$$\boxed{\text{on rejette H}_0 \text{ ssi } \frac{\sqrt{n}\;\left|\overline{X} - \mu_0\right|}{S} > t_{n-1;\,\frac{\alpha}{2}}}$$

où $t_{n-1;\alpha}$ et $t_{n-1;\alpha/2}$ représentent les fractiles de la loi de Student à $(n-1)$ d.d.l. d'ordre $1-\alpha$ et $1-\alpha/2$.

Remarque 7 ***Les tests de Student restent encore "valables" quand*** n ***est au moins de l'ordre de la trentaine , même si l'hypothèse de normalité n'est pas vérifiée.***

Exercice 20 : Test de l'espérance d'une loi normale d'écart-type connu

Soit X une v.a de loi normale $N(m;\sigma^2)$ d'écart-type connu σ =2. Au vu d'un échantillon $(X_1,\ldots,X_n)$ issu de la loi de X, on veut choisir entre les deux hypothèses :

$$\begin{cases} H_0 : m = 2 \\ H_1 : m = 3 \end{cases}$$

1. On résout ce problème de test par la méthode de Neyman-Pearson. On dispose donc de l'estimateur de la moyenne $\overline{X}$. Déterminer la région critique.
2. Dans le cas où $n = 100$ et $\alpha = 0,05$, calculer la puissance de ce test. Qu'en concluez-vous ?
3. Quelle doit être la taille d'échantillon minimum n_0 pour que la puissance soit égale à 0,95 ?

Exercice 21 : Test de l'espérance d'une loi normale d'écart-type inconnu

Soit X une v.a de loi normale $N(m; \sigma^2)$ d'écart-type inconnu. Au vu d'un échantillon $(X_1, \ldots, X_n)$ issu de la loi de X, on veut choisir entre les deux hypothèses :

$$\begin{cases} H_0 : m = 1 \\ H_1 : m = 2 \end{cases}$$

Déterminer la région critique de ce test dans le cas où $n = 25$ et $\alpha = 0,05$, dans le cas à $s = 0,2$ (préciser quel test utiliser) **relire cette phrase !**

Exercice 22 : Respect d'une norme de fabrication

On étudie la résistance à la rupture X d'un fil fabriqué selon des normes où la résistance moyenne doit être $m_0 = 300$g avec un écart-type $\sigma_0 = 20$g. La v.a X est supposée suivre une loi normale $N(m;\ \sigma^2)$. On désire vérifier le respect des normes.

Tester cette hypothèse sur la base d'un échantillon de 100 bobines de fil fournissant comme résultats une moyenne empirique $\overline{x} = 305$ et un écart-type empirique $s_{100} = 22$.

2.2.1.2 Comparaison de la variance σ^2 à une valeur donnée ${\sigma_0}^2$

a) Dans le cas où le paramètre μ **est connu**, pour tester l'hypothèse $H_0 : \sigma^2 = {\sigma_0}^2$ contre l'hypothèse $H_1 : \sigma^2 = {\sigma_1}^2$, le test de Neyman-Pearson nous donne la règle de décision suivante :

- dans le cas où $\sigma_1 > \sigma_0$: on rejette H_0 ssi $\sum_{i=1}^{n}(X_i - \mu)^2 > K$,
- dans le cas où $\sigma_1 < \sigma_0$: on rejette H_0 ssi $\sum_{i=1}^{n}(X_i - \mu)^2 < K$.

b) Dans le cas où le paramètre μ **est inconnu** et estimé par $\overline{X}$, on considère la statistique :

$$\frac{1}{{\sigma_0}^2} \sum_{i=1}^{n}(X_i - \overline{X})^2$$

qui, sous H_0 est une v.a χ^2_{n-1}.

On construit donc les tests suivants, au seuil α :

- 1^{er} **cas :** le test unilatéral pour tester $\begin{cases} H_0 : \sigma = \sigma_0 \\ H_{\sigma>} : \sigma > \sigma_0 \end{cases}$ est défini par la règle de décision suivante :

$$\boxed{\text{on rejette } H_0 \text{ ssi } \frac{1}{{\sigma_0}^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 > \chi^2_{n-1;\ \alpha}}$$

– $2^{ème}$ **cas :** le test unilatéral pour tester $\left\{\begin{array}{l} H_0 \ : \sigma = \sigma_0 \\ H_{\sigma<} : \sigma < \sigma_0 \end{array}\right.$ est défini par la règle de décision suivante :

$$\boxed{\text{on rejette } H_0 \text{ ssi } \frac{1}{{\sigma_0}^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 < \chi^2_{n-1;\ 1-\alpha}}$$

– $3^{ème}$ **cas :** le test bilatéral pour tester $\left\{\begin{array}{l} H_0 \ : \sigma = \sigma_0 \\ H_{\sigma\neq} : \sigma \neq \sigma_0 \end{array}\right.$ est défini par la règle de décision suivante :

$$\boxed{\text{on rejette } H_0 \text{ ssi } \frac{1}{{\sigma_0}^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 < \chi^2_{n-1;\ 1-\frac{\alpha}{2}} \text{ ou } \frac{1}{{\sigma_0}^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 > \chi^2_{n-1;\ \frac{\alpha}{2}}}$$

où $\chi_{n-1;\alpha}$, $\chi_{n-1;1-\alpha}$, $\chi_{n-1;\alpha/2}$ et $\chi_{n-1;1-\alpha/2}$ représentent les fractiles de la loi du chi-deux à $(n-1)$ d.d.l. d'ordre $1-\alpha$, α, $1-\alpha/2$ et $\alpha/2$.

Exercice 23 : Test de l'écart-type d'une loi normale d'espérance connue

Une entreprise, spécialisée dans la fabrication de pâtes *aux oeufs frais* fonde sa publicité sur l'affirmation suivante : *trois oeufs frais au kilo, soit 14% de la composition.* Le producteur s'engage ainsi à respecter certaines normes de qualité de fabrication portant sur les caractéristiques de la v.a X représentant le pourcentage d'oeufs dans la composition d'un paquet, supposée de loi normale $N(m;\sigma^2)$.

1. On suppose que $\sigma = 1$ et que la machine à fabriquer les pâtes est bien réglée si $m = 14$. Le fabricant considère qu'un paquet est non conforme à sa norme de qualité si $X < 14$. Calculer la probabilité p qu'un paquet ne soit pas conforme lorsque la machine est bien réglée. Quelle doit être la valeur minimale de m pour que la probabilité qu'un paquet ne soit pas conforme ne dépasse pas 5 % ?
2. Une grande surface reçoit un lot de ces pâtes et désire vérifier la qualité du fournisseur. Un échantillon de n paquets est tiré au hasard ; soit $(x_1, \ldots, x_n)$ les observations après analyse. On suppose $m = 6$ connu et σ^2 inconnu.
   Sur un échantillon de taille $n = 76$, on a observé $\sum_{i=1}^{76} x_i = 1140$ et $\sum_{i=1}^{76} {x_i}^2 = 17195$.
   Déterminer la région critique du test de risque $\alpha = 0,05$:
   $$\left\{\begin{array}{l} H_0 : \sigma = 1 \\ H_1 : \sigma = 2 \end{array}\right.$$

Exercice 24

Une boisson gazeuse, mise en vente au public depuis plusieurs mois, a procuré par quinzaine un chiffre d'affaires de loi normale d'espérance 157 000 euros et d'écart-type 19 000 euros. Une campagne publicitaire est alors décidée.
La moyenne des ventes des huit quinzaines suivant la fin de la promotion est 165 000 euros. On admet que l'écart-type reste constant.
La campagne publicitaire a-t-elle permis d'accroître le niveau moyen des ventes de 10 % ?

2.2.1.3 Tests sur les proportions

On se place dans le cadre de l'approximation gaussienne du sous-chapitre 2.3.3.1, auquel on se reportera.

Il s'agit de tester la valeur inconnue de la probabilité p d'un événement A, contre une valeur p_0. Pour ce faire, on dispose d'un échantillon de longueur n ayant donné k réalisations de A : k/n est donc l'estimation de p. Considérons les trois tests asymptotiques :

$$H_0 \ : \ p \leqslant p_0 \text{ contre } H_1 \ : \ p > p_0 \ ; \ W_\alpha = \left\{ \frac{k - np_0}{\sqrt{np_0 \ (1 - p_0)}} > u_\alpha \right\}$$

$$H_0 \ : \ p \geqslant p_0 \text{ contre } H_1 \ : \ p < p_0 \ ; \ W_\alpha = \left\{ \frac{k - np_0}{\sqrt{np_0 \ (1 - p_0)}} < -u_\alpha \right\}$$

$$H_0 \ : \ p = p_0 \text{ contre } H_1 \ : \ p \neq p_0 \ ; \ W_\alpha = \left\{ \left| \frac{k - np_0}{\sqrt{np_0 \ (1 - p_0)}} \right| > u_{\alpha/2} \right\}$$

Les $u_{\alpha/2}$ et u_α sont les fractiles de la loi $\mathcal{N}(0,1)$ d'ordre $1 - \alpha/2$ et $1 - \alpha$.

Exercice 25

Un généticien veut comparer les proportions p de naissances masculines et $1 - p$ de naissances féminines à l'aide d'un échantillon de $n = 900$ naissances où on a observé 470 garçons. Il considère donc le test suivant :

$$\begin{cases} H_0 : p = 0,5 \\ H_1 : p = 0,48 \end{cases}$$

1. Quelle est la conclusion sur cet échantillon et pourquoi le généticien est-il peu satisfait de ce test ? (on choisit $\alpha = 10\%$)
2. Il décide alors d'effectuer le test :
$$\begin{cases} H_0 : p = 0,5 \\ H_1 : p \neq 0,5 \end{cases}$$
Quelle est alors sa conclusion ?

2.2 Tests de comparaison de deux paramètres issus de populations distinctes

On considère deux échantillons extraits de deux populations normales distinctes :

$$\begin{cases} (X_1, \ldots, X_n) \text{ un échantillon de taille } n_1 \text{ de v.a de loi } N(\mu_1; \sigma_1^2) \\ (Y_1, \ldots, Y_n) \text{ un échantillon de taille } n_2 \text{ de v.a de loi } N(\mu_2; \sigma_2^2) \end{cases}$$

Notations
On pose :

$$\overline{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i \; ; \; S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \overline{X})^2 \; , \text{ et}$$

$$\overline{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i \; ; \; S_Y^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \overline{Y})^2$$

Compte tenu des hypothèses précédentes, on sait que la loi de :

$$\frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} \text{ est une loi de Fisher à } (n_1 - 1) \text{ et } (n_2 - 1) \text{ d.d.l.}$$

$$\frac{(\overline{X} - \mu_1) - (\overline{Y} - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ est une loi normale centrée réduite } N(0; 1).$$

On rappelle que : si $U_1 \sim N(\mu_1; \sigma_1^2)$, $U_2 \sim N(\mu_2; \sigma_2^2)$ et U_1 indépendante de U_2, alors $U_1 + U_2 \sim N(\mu_1 + \mu_2; \sigma_1^2 + \sigma_2^2)$.

2.2.2.1. Comparaison des variances σ_1^2 et σ_2^2

On souhaite tester l'hypothèse H_0 : $\sigma_1^2 = \sigma_2^2$ contre une hypothèse alternative H_1, de type inégalité.

Quand H_0 est vraie, la statistique $(S_X^2/\sigma_1^2)/(S_Y^2/\sigma_2^2) = S_X^2/S_Y^2$ suit une loi de Fisher à $(n_1 - 1)$

et $(n_2 - 1)$ d.d.l.

Avec un raisonnement analogue à celui du paragraphe précédent, on construit les tests au seuil α, dits de Fisher-Snedecor :

- 1^{er} **cas :** le test unilatéral pour tester $\left\{ \begin{array}{l} H_0 \ : \sigma_1^2 = \sigma_2^2 \\ H_{1,2} : \sigma_1^2 > \sigma_2^2 \end{array} \right.$ est défini par la règle de décision suivante :

  $$\boxed{\text{on rejette } H_0 \text{ ssi } \frac{S_X^2}{S_Y^2} > f_{n_1-1;n_2-1;\ \alpha}}$$

- $2^{ème}$ **cas :** le test unilatéral pour tester $\left\{ \begin{array}{l} H_0 \ : \sigma_1^2 = \sigma_2^2 \\ H_{2,1} : \sigma_1^2 < \sigma_2^2 \end{array} \right.$ est défini par la règle de décision suivante :

  $$\boxed{\text{on rejette } H_0 \text{ ssi } \frac{S_Y^2}{S_X^2} > f_{n_2-1;n_1-1;\ \alpha}}$$

- $3^{ème}$ **cas :** le test bilatéral pour tester $\left\{ \begin{array}{l} H_0 \ : \sigma_1^2 = \sigma_2^2 \\ H_{\neq} \ : \sigma_1^2 \neq \sigma_2^2 \end{array} \right.$ est défini par la règle de décision suivante :

  $$\boxed{\text{on rejette } H_0 \text{ ssi } \frac{S_X^2}{S_Y^2} > f_{n_1-1;n_2-1;\frac{\alpha}{2}} \text{ ou } \frac{S_Y^2}{S_X^2} > f_{n_2-1;n_1-1;\frac{\alpha}{2}}}$$

où $f_{k;l;\alpha}$, $f_{n_1-1;n_2-1;\alpha/2}$ représentent les fractiles de la loi de Fisher-Snedecor à k et l d.d.l. d'ordre $1-\alpha$ et $1-\alpha/2$.

2.2.2.2. Comparaison des moyennes μ_1 et μ_2 en présence de variances égales

On suppose donc que $\sigma_1 = \sigma_2$ et on notera σ_0 cette valeur commune. En pratique, on fait le test de comparaison des variances (voir paragraphe précédent), ce test devant conclure à l'acceptation de l'égalité des variances.
Le paramètre σ_0^2 est inconnu et est estimé par :

$$S_0^2 = \frac{\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{i=1}^{n_2}(Y_i - \overline{Y})^2}{n_1 + n_2 - 2}$$

On admet que $(n_1 + n_2 - 2)S_0^2/\sigma_0^2$ suit une loi du χ^2 à $(n_1 + n_2 - 2)$ d.d.l.
On considère alors la statistique :

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_0\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

qui suit une loi de Student à (n_1+n_2-2) d.d.l. quand H_0 est vraie (voir rappels du paragraphe 3.2.1)

On construit donc les tests suivants, au seuil α :

– 1^{er} **cas :** le test unilatéral pour tester $\begin{cases} H_0 & : \mu_1 = \mu_2 \\ H_> & : \mu_1 > \mu_2 \end{cases}$
est défini par la règle de décision suivante :

$$\boxed{\text{on rejette } H_0 \text{ ssi } \frac{\overline{X}-\overline{Y}}{S_0\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > t_{n_1+n_2-2;\ \alpha}}$$

– $2^{ème}$ **cas :** le test unilatéral pour tester $\begin{cases} H_0 & : \mu_1 = \mu_2 \\ H_< & : \mu_1 < \mu_2 \end{cases}$
est défini par la règle de décision suivante :

$$\boxed{\text{on rejette } H_0 \text{ ssi } \frac{\overline{X}-\overline{Y}}{S_0\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < -t_{n_1+n_2-2;\ \alpha}}$$

– $3^{ème}$ **cas :** le test bilatéral pour tester $\begin{cases} H_0 & : \mu_1 = \mu_2 \\ H_\neq & : \mu_1 \neq \mu_2 \end{cases}$
est défini par la règle de décision suivante :

$$\boxed{\text{on rejette } H_0 \text{ ssi } \frac{|\overline{X}-\overline{Y}|}{S_0\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > t_{n_1+n_2-2;\ \alpha/2}}$$

où $t_{n_1+n_2-2;\alpha}$ et $t_{n+1+n_2-2;\alpha/2}$ représentent les fractiles de la loi de Student à (n_1+n_2-2) d.d.l. d'ordre $1-\alpha$ et $1-\alpha/2$.

2.2.2.3. Comparaison des moyennes μ_1 et μ_2 en présence de variances inégales

Dans le cas où le test de comparaison de variances conclut à l'inégalité des variances, la loi de probabilité de la statistique de test dépend alors des paramètres σ_1 et σ_2 inconnus, ce qui ne permet pas de déterminer un test de seuil α.
Seul le cas de "grands" échantillons ($n > 30$) nous permet de donner une réponse à ce problème. On construit les tests suivants, de seuil *asymptotiquement égal* à α en utilisant la convergence en loi de :

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2}}} \text{ vers la loi } N(0;1).$$

Par exemple, le test unilatéral pour tester $\begin{cases} H_0 & : \mu_1 = \mu_2 \\ H_1 & : \mu_1 > \mu_2 \end{cases}$
est défini par la règle de décision suivante :

$$\text{on rejette } H_0 \text{ ssi } \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}} > z_\alpha$$

Le test bilatéral pour tester $\begin{cases} H_0 & : \mu_1 = \mu_2 \\ H_1 & : \mu_1 \neq \mu_2 \end{cases}$
est défini par la règle de décision suivante :

$$\text{on rejette } H_0 \text{ ssi } \frac{|\overline{X} - \overline{Y}|}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}} > z_{\alpha/2}$$

où z_α et $z_{\alpha/2}$ représentent les fractiles de la loi de Normale centrée réduite d'ordre $1-\alpha$ et $1-\alpha/2$.

Exercice 26

Le propriétaire d'un magasin d'alimentation constate que son chiffre d'affaires, supposé suivre une loi normale, baisse depuis l'installation à proximité, d'un hypermarché. Pour savoir si cette baisse est significative, il a relevé le montant hebdomadaire de ses ventes x_i durant les vingt semaines précédant l'ouverture de cet hypermarché ($1 \leqslant i \leqslant 20$) et y_i pendant les trente-deux semaines après l'ouverture ($1 \leqslant j \leqslant 32$). Que peut-on conclure à partir des observations $\overline{x} = 33,8$; $\overline{y} = 30,9$; $\sum_{i=1^{20}}(x_i - \overline{x})^2 = 763$ et $\sum_{j=1}^{32}(y_i - \overline{y})^2 = 875$?

Exercice 27

Une usine élabore une pâte de verre dont la température de ramollissement X est supposée suivre une loi normale.

1. A six mois d'intervalle, deux séries d'observations sont réalisées et les moyennes et écart-types empiriques sont les suivants :
   $n_1 = 41$; $\overline{x_1} = 785$; $s_1 = 1,68$; $n_2 = 61$; $\overline{x_2} = 788$; $s_2 = 1,40$
   Les productions sont-elles identiques ? On comparera les variances, puis les moyennes au risque $\alpha = 5\%$.
2. Même question avec :
   $n_1 = 9$; $\overline{x_1} = 2510$; $s_1 = 15,9$; $n_2 = 21$; $\overline{x_2} = 2492$; $s_2 = 24,5$

Exercice 28

Soient deux populations d'individus dont certaines présentent une caractéristique donnée (par exemple : pièce défectueuse) dans les proportions p_1 et p_2 inconnues. On prélève n_1 individus dans la population (1) et n_2 dans la population (2) ; soient $X_i (i = 1, 2)$ les nombres aléatoires d'individus présentant la caractéristique.

1. Déterminer la loi exacte de X_i.
2. Sous l'hypothèse qui garantit l'approximation gaussienne ($i = 1, 2 \ : \ n_i p_i > 5$ et $n_i(1 - p_i) > 5$, démontrer que si $H_0 \ : \ p_1 = p_2 = p$ est vraie, la v.a $(X_1/n_1 - X_2/n_2)$ est $N(0, \sigma^2 = p\ (1-p)(1/n_1 + 1/n_2))$.
3. Approcher p inconnu par une estimation $\hat{p}$ satisfaisante.
4. Construire la région de confiance du test : $H_0 : p_1 = p_2 = p$ contre $H_1 : p_1 \neq p_2$.
5. Construire la région de confiance du test : $H_0 : p_1 \leqslant p_2$ contre $H_1 : p_1 > p_2$.
6. Application : $n_1 = 30,\ n_2 = 40,\ X_1 = 5,\ X_2 = 8$.

2.3 Test du chi-deux

Dans ce paragraphe, nous développons quelques tests classiques d'hypothèse portant sur la nature de la loi de probabilité d'une v.a décrivant une population donnée. Certaines méthodes graphiques permettent de se faire une idée sur la loi de la variable étudiée (histogramme, fonction de répartition empirique, droite de Henry dans le cas d'une loi gaussienne, ...), mais ne permettent pas, contrairement au test du chi-deux, de contrôler les erreurs que l'on risque de faire en acceptant tel ou tel type de loi.

Le test du chi-deux appartient à la classe des tests qui ne concernent pas la valeur d'un paramètre inconnu, appelés tests non-paramétriques. Le test du chi-deux est un test d'ajustement qui permet de déterminer si les observations d'une population donnée vérifient une loi postulée (normale, exponentielle, ...).

Rappels sur la loi multinomiale à k catégories

Dans une urne contenant des boules de k catégories : $C_1,\ C_2, \ldots,\ C_k$ en proportions respectivement $p_1,\ p_2, \ldots,\ p_k$ telles que $\sum_{i=1}^{k} p_i = 1$, on réalise une suite de n *tirages avec remise* (ou indépendants). Soit x_i l'observation de la v.a X_i, représentant le nombre de boules de la catégorie C_i, obtenu parmi les n boules tirées ($i = 1, 2, \ldots, k$ et $\sum_{i=1}^{k} x_i = n$).

La v.a multidimensionnelle $(X_1, \ldots, X_k)$ suit une loi multinomiale de paramètres $(n; p_1, \ldots, p_k)$, c'est-à-dire que la vraisemblance de $(X_1, \ldots, X_k)$ au point $(x_1, \ldots, x_k)$ s'écrit :

$$\mathrm{P}(X_1 = x_1, \ldots, X_k = x_k;\ p_1, \ldots, p_k) = \frac{n!}{x_1! \ldots x_k!}\ p_1^{x_1} \ldots p_k^{x_k}$$

Les n boules tirées reproduisent plus ou moins bien la composition de l'urne et on mesure l'écart entre les observations x_i et les valeurs théoriques np_i espérées, par la valeur de la distance :

$$k_n = \sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i},\ \text{qui est l'observation de la v.a } K_n = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

où chaque v.a X_i est $B(n; p_i)$.

La v.a K_n ne suit pas une loi de probabilité usuelle, par contre on connaît sa loi asymptotique, lorsque n atteint de grandes valeurs.

Théorème 14 Quand $(X_1, \ldots, X_k)$ suit une loi multinomiale $(n;\ p_1, \ldots, p_k)$, la loi de K_n tend vers une loi du χ^2 à $(k-1)$ d.d.l. quand $n \to \infty$.

Remarque 8 *En pratique, il faudra toujours s'assurer que pour tout i, les valeurs de chaque np_i dépassent 5, pour pouvoir approcher la loi de K_n par la loi $\chi^2(k-1)$.*

Hypothèse H_0 : $(X_1, \ldots, X_n)$ suit une loi multinomiale de paramètres $(n;\ p_1, \ldots, p_k)$

On construit un test de niveau asymptotiquement égal à α, pour tester H_0 contre $H_1 =$ non H_0. On a alors la règle de décision suivante :

$$\text{on rejette } H_0 \text{ ssi } K_n = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} > \chi^2_{k-1;\ \alpha}$$

où $\chi_{k-1;\alpha}$ représente le fractile de la loi du chi-deux à $(k-1)$ d.d.l. d'ordre $1-\alpha$.

Intuitivement, cette règle de décision revient donc à rejeter H_0 si l'écart (ou distance) entre la valeur observée x_i et la valeur espérée np_i lorsque H_0 est vraie, est trop "grand".

Remarque 9 *1. Il est important de souligner que, dans les exemples étudiés, on cherchera à ne pas rejeter H_0. Ce raisonnement inhabituel nécessite un test de puissance forte et donc une probabilité du risque de $2^{\text{ème}}$ espèce, P(ne pas rejeter H_0/H_1 vraie), faible. Toutefois, il est difficile de calculer la puissance qui dépend de la loi de K_n, inconnue quand H_1 est vraie.*

2. *Dans un certain nombre de situations, l'hypothèse H_0 sera composée, par exemple :*
$$\begin{cases} H_0 : (X_1, \ldots, X_n) \text{ v.a multinomiale } (n;\ p_1(\theta), \ldots, p_n(\theta)), \\ H_1 :\ \text{non } H_0, \end{cases}$$
où θ est un paramètre inconnu de dimension $l : \theta = (\theta_1, \ldots, \theta_l)$. Dans ce cas, on estime, par exemple, θ à l'aide de l'E.M.V. $\hat{\theta} = (\hat{\theta_1}, \ldots, \hat{\theta_l})$ et on a le résultat suivant :

Théorème 15 Quand H_0 est vraie, la loi de :
$$K_n = \sum_{i=1}^{k} \frac{(X_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})}$$
tend vers une loi de χ^2 à $(k - l - 1)$ d.d.l. quand $n \to \infty$.

On a vu qu'en pratique, il faut vérifier : $\forall i,\ np_i > 5$.

3. *Lorsque le nombre ν de d.d.l. d'une loi de $\chi^2(\nu)$ est grand, on peut approcher la loi de $\chi^2(\nu)$ par une loi normale $N(\nu, 2\nu)$.*

Exemple 7 On jette 100 fois un dé et on observe $y_1, \ldots, y_{100}$ les 100 numéros relevés sur la face du dé, regroupés dans le tableau suivant :

face du dé observée	1	2	3	4	5	6
fréquence observée x_i	15	18	15	17	16	19

Le dé est-il parfait ?

Soit p_i la probabilité d'observer la face i
$$\begin{cases} H_0 : p_1 = 1/6, \ldots, p_6 = 1/6 \ \text{(equiprobabilité des résultats : le dé est parfait).} \\ H_1 :\ \text{non } H_0 \end{cases}$$
On cherche donc l' "écart" k_n entre les fréquences observées x_i et les effectifs théoriques espérés si H_0 est vraie. On a :
$$k_n = \sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i} = \sum_{i=1}^{6} \frac{(x_i - \frac{100}{6})^2}{\frac{100}{6}} = 0,81$$
Si on cherche à conclure au risque $\alpha = 5\%$, on lit sur la table : $\chi^2_{5;\ 5\%} = 11,07$. Comme $k_n < \chi^2_{5;\ 5\%}$, on ne rejette pas l'hypothèse H_0, c'est-à-dire que l'on considère que le dé est parfait.

On ne peut pas calculer la puissance du test du chi-deuxcar l'hypothèse H_1 n'est pas explicitée. Néanmoins, on peut remarquer que la valeur k_n est très petite par rapport à la valeur du $\chi^2_{5;\ 5\%}$. On admet donc que le dé est parfait.

Cas général

Soit un échantillon $(Y_1, \ldots, Y_n)$ et L une loi de probabilité donnée sur R, de fonction de répartition F, *entièrement déterminée* (sans paramètre inconnu). On veut soumettre au test du χ^2 les hypothèses :

$$\begin{cases} H_0 : \text{ la loi } L \text{ est la loi commune aux v.a } Y_i \ , \\ H_1 : \text{ non } H_0. \end{cases}$$

On se ramène aux résultats du paragraphe précédent en découpant R en k classes :

$$]-\infty; a_1], \ ]a_1; a_2], \ldots, \ ]a_{j-1}; a_j], \ldots, \ ]a_{k-1;+\infty}]$$

Quand H_0 est vraie, la probabilité qu'une observation y_i appartienne à la $j^{ème}$ classe $]a_{j-1}; a_j]$ est :

$$p_j = \mathrm{P}(Y_i \in]a_{j-1}; a_j]) = \mathrm{P}(a_{j-1} < Y \leqslant a_j) = F(a_j) - F(a_{j-1}).$$

Si on note X_j la v.a associée au nombre d'observations qui se trouvent dans la $j^{ème}$ classe $(j = 1, \ldots, k)$, la règle de décision est :

$$\text{on rejette } H_0 \text{ ssi } K_n = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} > \chi_{k-1;\ \alpha} \ , \text{ où } p_i = F(a_i) - F(a_{i-1})$$

Rappelons que $F(-\infty) = 0$ et $F(+\infty) = 1$

Remarque 10 *1. Si la loi L dépend de l paramètres inconnus $(\theta_1, \ldots, \theta_l)$ que l'on estimera dans un premier temps à l'aide d'un estimateur $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_l)$, la région de rejet de H_0 sera définie par :*

$$\text{on rejette } H_0 \text{ ssi } \sum_{i=1}^{k} \frac{(x_i - np_i\ (\hat{\theta}))^2}{np_i\ (\hat{\theta})} > \chi_{k-l-1;\ \alpha}$$

2. Il faut noter que la convergence en loi de la v.a K_n vers une loi de χ^2 n'est valable que si les quantités np_i ne sont pas trop petites. En pratique, on constituera donc des classes (et éventuellement on en regroupera certaines) de façon que l'effectif théorique np_i soit supérieur à 5.

Exercice 29 : Test d'adéquation

On étudie la circulation en un point fixe d'une autoroute en comptant, pendant deux heures, le nombre de voitures passant par minute devant un observateur. Le tableau suivant résume les données obtenues :

Nombres x_i de voitures	0	1	2	3	4	5	6	7	8	9	10	11
Fréquences observées n_i	4	9	24	25	22	18	6	5	3	2	1	1

Tester l'adéquation de la loi empirique à une loi théorique simple pour un risque $\alpha = 0,10$, par exemple la loi de Poisson si $\overline{x} \simeq s_x^2$.

Exercice 30

A la sortie d'une chaîne de fabrication, on prélève toutes les trente minutes un lot de 20 pièces mécaniques et on contrôle le nombre de pièces défectueuses du lot. Sur 200 échantillons indépendants, on a obtenu les résultats suivants :

Nombres de pièces défectueuses	0	1	2	3	4	5	6	7
Nombre de lots	26	52	56	40	20	2	0	4

Tester l'adéquation de la loi empirique du nombre de pièces défectueuses par lot de 20 pièces à une loi théorique simple pour un risque $\alpha = 0,05$.

Test d'indépendance de deux variable qualitatives

Table de contingence

Soit deux caractères qualitatifs (appelés *facteurs*) :
- le caractère C ayant c modalités : $C_1,\ C_2, \ldots, C_c$,
- le caractère L ayant l modalités : $L_1,\ L_2, \ldots, L_l$.

Dans la population étudiée, on prélève au hasard n individus et on note x_{ij} le nombre d'observations de la cellule $(C_i;\ L_j)$, c'est-à-dire le nombre d'individus possèdant la $i^{ème}$ modalité de C et la $j^{ème}$ modalité de L, avec $1 \leqslant i \leqslant c$ et $1 \leqslant j \leqslant l$. On dispose alors d'une table de contingence dans laquelle chacun des n individus doit se retrouver dans une seule des $l \times c$ cases.

effectifs observés par couple de facteurs	C_1	$\ldots$	C_j	$\ldots$	C_c	Total
L_1	x_{11}	$\ldots$	x_{1j}	$\ldots$	x_{1c}	$x_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
L_i	x_{i1}	$\ldots$	x_{ij}	$\ldots$	x_{ic}	$x_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
L_l	x_{l1}	$\ldots$	x_{lj}	$\ldots$	x_{lc}	$x_{l\bullet}$
Total	$x_{\bullet 1}$	$\ldots$	$x_{\bullet j}$	$\ldots$	$x_{\bullet c}$	n

$$\text{On calcule les effectifs marginaux par } : x_{i\bullet} = \sum_{j=1}^{c} x_{ij} \ ;\ x_{\bullet j} = \sum_{i=1}^{l} x_{ij} \ .$$

$$\text{De plus, on a } : \sum_{i=1}^{l}\sum_{j=1}^{c} x_{ij} = \sum_{i=1}^{l} x_{i\bullet} = \sum_{j=1}^{c} x_{\bullet j} = n \ .$$

Règle de décision

Les hypothèses du test sont :

$$\begin{cases} H_0 : \text{ les deux caractères sont indépendants,} \\ H_1 : \text{ non } H_0 \end{cases}$$

On se retrouve dans le cas de la loi multinomiale à $l \times c$ catégories et on note p_{ij} la probabilité pour un individu d'appartenir à la cellule $(C_i;\ L_j)$ pour $i = 1, \ldots, l$ et $j = 1, \ldots, c$.
On en déduit les probabilités marginales $p_{1\bullet}, \ldots, p_{i\bullet}, \ldots, p_{c\bullet}$ pour le caractère C et $p_{\bullet 1}, \ldots, p_{\bullet j}, \ldots, p_{\bullet l}$ pour le caractère L. On a alors le tableau ci-dessous :

Probabilités inconnues $\searrow$	C_1	$\ldots$	C_j	$\ldots$	C_c	Probabilité marginale de L
L_1	p_{11}	$\ldots$	p_{1j}	$\ldots$	p_{1c}	$p_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
L_i	p_{i1}	$\ldots$	p_{ij}	$\ldots$	p_{ic}	$p_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
L_l	p_{l1}	$\ldots$	p_{lj}	$\ldots$	p_{lc}	$p_{l\bullet}$
Probabilité marginale de C	$p_{\bullet 1}$	$\ldots$	$p_{\bullet j}$	$\ldots$	$p_{\bullet c}$	1

Tableau des probabilités inconnues

On sait que L et C sont indépendantes en probabilité si $p_{ij} = p_{i\bullet}p_{\bullet j}$ pour tout (i, j). Lorsque H_0 est vraie, on a bien l'indépendance entre L et C.
Il faut alors estimer $(l + c - 2)$ paramètres. En effet, $p_{l\bullet}$ et $p_{\bullet c}$ se déterminent par les relations $\sum p_{i\bullet} = \sum p_{\bullet j} = 1$.

On admet que $p_{i\bullet}$ et $p_{\bullet j}$ sont estimés respectivement par $x_{i\bullet}/n$ et $x_{\bullet j}/n$. Quand H_0 est vraie, on peut donc estimer p_{ij} par $x_{i\bullet}x_{\bullet j}/n^2$.

On utilise la règle de décision suivante :

on rejette H_0 ssi $\displaystyle\sum_{i=1}^{l}\sum_{j=1}^{c} \frac{(x_{ij} - \frac{x_{i\bullet}x_{\bullet j}}{n})^2}{\frac{x_{i\bullet}x_{\bullet j}}{n}} > \chi_{(l-1)(c-1);\ \alpha}$

En effet, le degré de liberté de la loi limite est égal au nombre de paramètres estimés :
$$lc - (l + c - 2) - 1 = (l - 1)(c - 1)$$

Exercice 31 : Test d'indépendance

Un examen est ouvert à des étudiants d'origines différentes : économie, informatique et mathématiques. Le responsable de l'examen désire savoir si la formation initiale d'un étudiant influe sur sa réussite. A cette fin, il construit le tableau ci-dessous à partir des résultats obtenus par les 286 candidats, les origines étant précisées en colonne :

	Economie	Informatique	Mathématiques	Total
Réussite	41	59	54	154
Echec	21	36	75	132
Total	62	95	129	286

Quelle est sa conclusion ?

Exercice 32 : Test de comparaison de proportion

Un homme politique s'interrogeant sur ses chances éventuelles de succès aux élections présidentielles commande un sondage qui révèle que, sur 2000 personnes, 19% ont l'intention de voter pour lui. Il demande alors à une agence de publicité de promouvoir son image.
Un second sondage réalisé après cette campagne publicitaire, auprès de 1000 personnes, montre que 230 d'entre elles ont l'intention de voter pour lui. Peut-on considérer que cette campagne a été efficace (on utilisera la méthode précédente, ainsi que la méthode classique (2.2.1)) ?

3 Tests d'analyse de variance

L'analyse de variance permet d'évaluer et de comparer les effets d'un ou plusieurs facteurs contrôlés sur une population donnée. Sous l'hypothèse de normalité de la population, l'analyse de variance se réduit à un test de comparaison globale des moyennes d'un ensemble de sous-populations associées aux divers niveaux des facteurs.

3.1 Cas d'un facteur

On fait varier un facteur selon k modalités (correspondant en pratique à k sous-populations P_i) : à chaque modalité i du facteur, est associée la variable aléatoire X_i de loi $N(m_i;\ \sigma)$ où m_i est **inconnue** et σ **connue** .

Pour tout i, $1 \leqslant i \leqslant k$, on tire un échantillon de taille n_i :
soit $(X_i^1 = x_i^1,\ X_i^2 = x_i^2, \ldots, X_i^{n_i} = x_i^{n_i})$. On pose : $n = \sum_{i=1}^{k} n_i$.

L'analyse de variance est un test qui permet de tester :
$H_0 : m_1 = m_2 = \ldots = m_k$ contre H_1 (négation de H_0) : $\exists i, j$ tels que $m_i \neq m_j$

La variable X_i^j, associée au $j^{ème}$ tirage de la variable X_i, se décompose ainsi en une somme d'effets :

$$X_i^j = \mu + \alpha_i + \varepsilon_i^j$$

où μ est la valeur moyenne de X, toutes modalités confondues ; α_i est l'effet moyen dû à la modalité i du facteur considéré ; ε_i^j est la v.a résiduelle de loi normale $N(0; \sigma)$.

La quantité $\mu + \alpha_i$ est la valeur moyenne $\bar{X}$ sur la population P_i, correspondant à la modalité i.

Notations :

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_i^j \ , \qquad \overline{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_i^j \ ,$$

$$S_A^2 = \frac{1}{n} \sum_{i=1}^{k} n_i \ (\overline{X_i} - \overline{X})^2 \ , \quad S_R^2 = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_i^j - \overline{X}_i)^2 \ , \quad S^2 = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_i^j - \overline{X})^2 .$$

Commentaires : S_A^2 est la **variance inter-modalités** (moyenne des écarts quadratiques entre $\overline{X}$ et $\overline{X}_i$) ; S^2 est la **variance totale** (moyenne des écarts quadratiques entre $\overline{X}$ et tous les X_i^j) ; S_R^2 est la **variance résiduelle** (moyenne des dispersions autour des $\overline{X}_i$).

Théorème 16 (Formule d'analyse de variance à 1 facteur)
$$S^2 = S_A^2 + S_R^2 \ .$$

Preuve :

$$\forall\ i, j \quad X_i^j - \overline{X} = X_i^j - \overline{X}_i + \overline{X}_i - \overline{X}$$

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_i^j-\overline{X})^2=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_i^j-\overline{X}_i)^2+\frac{1}{n}\sum_{i=1}^{k}n_i\,(\overline{X_i}-\overline{X})^2$$

$$S^2=S_A^2+S_R^2.$$

Remarque 11 *On remarque que la formule précédente s'apparente à la formule de la variance totale (cf : "théorie des probabilités", paragraphe 7.3.4).*

Théorème 17 1. La v.a $n\ S_R^2/\sigma^2$ suit une loi χ^2 à $(n-k)$ d.d.l.

2. Sous l'hypothèse H_0 (égalité des m_i), la v.a $n\ S^2/\sigma^2$ suit une loi χ^2 à $(n-1)$ d.d.l et la v.a $n\ S_A^2/\sigma^2$ suit une loi χ^2 à $(k-1)$ d.d.l.

Preuve :

Pour démontrer le deuxième point du théorème précédent, on rappelle que :

$$\frac{n\ S^2}{\sigma^2}=\sum_{i,j}\left(\frac{X_i^j-\overline{X}}{\sigma}\right)^2$$

et donc : la v.a $n\ S_R^2/\sigma^2$ suit une loi χ^2 à $(n-1)$ d.d.l.

De même :

$$\frac{n\ S_A^2}{\sigma^2}=\sum_{i=1}^{k}\ n_i\left(\frac{\overline{X}_i-\overline{X}}{\sigma}\right)^2=\sum_{i=1}^{k}\ \left(\frac{\overline{X}_i-\overline{X}}{\sigma/\sqrt{n_i}}\right)^2$$

qui est un χ^2 à $(k-1)$ d.d.l.

On remarque que le théorème 17 est une conséquence du théorème 18, puisque $\chi^2_{n-1}=\chi^2_{n-k}+\chi^2_{k-1}$.

Ainsi, si H_0 est vraie, la v.a $\dfrac{S_A^2/(k-1)}{S_R^2/(n-k)}$ suit une loi de Fisher − Snedecor à $(n-1;n-k)$ d.d.l., que l'on notera $F(k-1;n-k)$.

On vérifie que si la variance inter-modalités S_A^2 est faible par rapport à la variance résiduelle S_R^2, alors la variable de Fisher-Snedecor prend une petite valeur : cette constatation est à la base du test.
A un risque de première espèce α fixé, on conclura au rejet ou à l'acceptation de H_0.

Exemple 8 Trois machines sont normalement réglées pour produire des pièces identiques dont la caractéristique est de loi $N(m;\sigma)$. On veut s'assurer qu'elles ne sont pas déréglées. On prélève donc un échantillon produit par chaque machine :

– machine 1 : $n_1 = 5,\ \overline{x_1} = 8$,
– machine 2 : $n_2 = 5,\ \overline{x_2} = 10$,
– machine 3 : $n_3 = 5,\ \overline{x_3} = 15$.
Les calculs donnent :

$$\sum_{i=1}^{3} n_i\ (\overline{x_i} - \overline{x})^2\ = 130\ ;\ \sum_{i,j} (\overline{x_i^j} - \overline{x_i})^2\ = 104$$

et donc

$$\sum_{i,j} (\overline{x_i^j} - \overline{x})^2\ = 234.$$

On présente les résultats sous forme de tableau :

Origine des variations	Somme des carrés des écarts	Degrés de liberté	
Inter − modalités	$ns_A^2 = 30$	$k - 1 = 2$	$ns_A^2/(k-1) = 65$
Erreur à l′ intérieur des groupes	$ns_R^2 = 104$	$n - k = 12$	$ns_R^2/(n-k) = 8,67$
Total	$ns^2 = 234$	$n - 1 = 14$	

La quantité $ns_R^2/(n-k)$ est une estimation de σ^2 quel que soit l'effet du facteur.
Si H_0 est vraie, alors $ns_A^2/(k-1)$ est une estimation de σ^2, indépendante de la précédente.

$$\text{Donc : } \frac{s_A^2/(k-1)}{s_R^2/(n-k)} \text{ est une v.a de Fisher} - \text{Snedecor, égale à } \frac{65}{8,67} = 7,5 \text{ , que}$$

l'on doit comparer à $F_{0,95}(2;12) = 3,89$, dans la table correspondante.

Comme $7,5 > 3,89$, on rejette H_0 avec un risque de 5%. Dans cet exemple, on peut même être plus exigeant, puisque au risque de 1%, on a $F_{0,99}(2;12) = 6,93$ qui est toujours inférieur à $7,5$ (le rejet de H_0 est donc toujours garanti).

Les estimations des moyennes $m_1,\ m_2,\ m_3$ sont évidemment :
$\overline{x_1} = 8,\ \overline{x_2} = 10,\ \overline{x_3} = 15.$

Attention ! Le rejet de H_0 ne signifie pas que toutes les moyennes m_i sont significativement différentes entre elles, mais que deux d'entr'elles, au moins, le sont. En toute rigueur, il faudrait tester ensuite $(m_i - m_j = 0)$ contre $(m_i - m_j \neq 0)$ pour les couples $(m_i;\ m_j)$ les plus distincts, afin de déterminer lesquels sont significativement différents à un niveau de confiance donné (méthode de Scheffé).

3.2 Cas de deux facteurs

Deux facteurs indépendants A et B varient respectivement selon p et q modalités : au couple (i, j) des modalités respectives des facteurs A et B, correspond un échantillon de taille

$n_{i,j}$ de la v.a descriptive X.
Le modèle statistique correspondant est dit équilibré si : $\forall\ i, j,\ n_{i,j} = r$; c'est l'hypothèse dans laquelle nous nous placerons. Donc, à tout couple de modalités (i, j), on associe l'échantillon $(X_{i,j,1} = x_{i,j,1},\ X_{i,j,2} = x_{i,j,2}, \ldots, X_{i,j,r} = x_{i,j,r})$. La v.a $X_{i,j}$ est supposée de loi $N(m_{i,j};\ \sigma)$. On peut alors décomposer la moyenne $m_{i,j}$ de la sous-population $P(i, j)$ de la façon suivante :

$$\boxed{m_{i,j} = \mu + \alpha_i + \beta_j + \gamma_{i,j}}$$

Notations

$$\overline{X} = \frac{1}{pqr} \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{r} X_{i,j,k} \quad \overline{X_{i,j,\bullet}} = \frac{1}{r} \sum_{k=1}^{r} X_{i,j,k}$$

$$\overline{X_{i,\bullet,\bullet}} = \frac{1}{qr} \sum_{j=1}^{q} \sum_{k=1}^{r} X_{i,j,k} \quad \overline{X_{\bullet,j,\bullet}} = \frac{1}{pr} \sum_{i=1}^{p} \sum_{k=1}^{r} X_{i,j,k}$$

Après calcul, on retrouve une formule d'analyse de variance analogue à la précédente, qui prend en compte les variances des facteurs et de leur interaction.

$$\text{Posons : } S_A^2 = qr \sum_i (\overline{x_{i,\bullet,\bullet}} - \overline{x})^2, \qquad S_B^2 = qr \sum_j (\overline{x_{\bullet,j,\bullet}} - \overline{x})^2,$$

$$S_{AB}^2 = r \sum_i \sum_j (\overline{x_{i,j,\bullet}} - \overline{x_{i,\bullet,\bullet}} - \overline{x_{\bullet,j,\bullet}} + \overline{x})^2, \quad S_R^2 = \sum_i \sum_j \sum_k (x_{i,j,k} - \overline{x_{i,j,\bullet}})^2,$$

$$S^2 = \sum_i \sum_j \sum_k (x_{i,j,k} - \overline{x})^2.$$

Théorème 18 (Formule d'analyse de variance à 2 facteurs)

$$\boxed{S^2 = S_A^2 + S_B^2 + S_{AB}^2 + S_R^2}$$

(la preuve est fastidieuse et de peu d'intérêt)

Tableau d'analyse de variance à 2 facteurs :

Origine des variations	Somme des carrés	d.d.l	Carrés moyens	Variable F
A	S_A^2	$p-1$	$S_A^2/(p-1) = S_{AM}^2$	S_{AM}^2/S_{RM}^2
B	S_B^2	$q-1$	$S_B^2/(q-1) = S_{BM}^2$	S_{BM}^2/S_{RM}^2
Interaction AB	S_{AB}^2	$(p-1)(q-1)$	$\frac{S_{AB}^2}{(p-1)(p-q)} = S_{ABM}^2$	S_{ABM}^2/S_{RM}^2
Résiduelle	S_R^2	$pq(r-1)$	$\frac{S_R^2}{pq(r-1)} = S_{RM}^2$	
Totale	S^2	$pqr-1$		

Exemple 9 Des ampoules électriques sont fabriquées en utilisant :
- 4 types de filament (facteur A à 4 modalités),
- 4 types de gaz (facteur B à 4 modalités).

On sait qu'il n'y a pas d'interaction. Ici, $n_{i,j} = 1,\ \forall i, j$.

Voici le tableau des durées de vie :

		Facteur B			
		B_1	B_2	B_3	B_4
Facteur A	A_1	44	22	36	34
	A_2	47	43	41	53
	A_2	0	9	10	17
	A_2	36	14	1	34

Tableau d'analyse de variance à 2 facteurs :

Origine des variations	Somme des carrés	d.d.l	carrés moyens	Variable F
A	3063	3	1021	$S^2_{AM}/S^2_{RM} = \frac{1021}{90} = 11,3$ $> F_{0,95}(3;9) = 3,86$
B	510	3	170	$S^2_{BM}/S^2_{RM} = \frac{170}{90} = 1,89$ $< F_{0,95}(3;9) = 3,86$
Résiduelle	811	9	90	
Totale	4384			

Conclusion : On admet que seul le facteur A a une influence au risque de première espèce 5%.

Chapitre 3

Régression à une variable

Le modèle de régression permet d' expliquer et d'exprimer une variable aléatoire Y en fonction d'une variable explicative x selon le modèle fonctionnel de la forme

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

où :

- f est une fonction connue dépendante de paramètres $\theta_1, \theta_2, ..., \theta_p$ inconnus ;
- x est soit une variable contrôlée, soit une variable aléatoire variant dans un intervalle $I \subset \mathbb{R}$;
- ε est une variable aléatoire, qui est associée à l'écart aléatoire entre le modèle et la variable expliquée Y ; on l'appelle **résidu** ou **erreur**.

A chaque valeur $x_i, i = 1, 2, ..., n$ est associée une valeur y_i de Y et un résidu ε_i.

Objectif de la théorie :

Etant donné un ensemble de n couples (x_i, y_i) associé à n mesures expérimentales, il s'agira :

1. d'en déduire la meilleure estimation possible des paramètres $(\widehat{\theta_k})_{k=1,...p}$;
2. d'évaluer l'adéquation du modèle ainsi obtenu ;
3. d'effectuer des tests de comparaison sur les paramètres et d'utiliser éventuellement le modèle à des fins prévisionnelles.

1 Régression linéaire

Lorsque la variable explicative est une variable aléatoire X, le modèle explicatif de Y en fonction de X s'écrit sous la forme :

$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x$$

Toutefois, dans le cadre de cet exposé, nous nous placerons dans le cas où la variable x est contrôlée.

1.1 Modèle linéaire standard

A tout x_i est associé $Y(x_i) \underset{\text{notée}}{=} Y_i$ selon le modèle :

$$\forall i, \quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{3.1}$$

Les v.a **résiduelles ou résidus** ε_i sont supposées vérifier les trois hypothèses suivantes :

R$_1$. $\forall i, \mathbb{E}(\varepsilon_i) = 0$,
R$_2$. $\forall i, \mathrm{Var}(\varepsilon_i) = \sigma^2$, inconnu,
R$_3$. $\forall (i,j)$, tel que $i \neq j, \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$.

Les résidus sont centrés, de même variance et décorrélés.

On déduit de **R$_1$**, **R$_2$**, **R$_3$** que : $\mathbb{E}(Y_i) = \beta_0 + \beta_1 \cdot x_i$; $\mathrm{Var}(Y_i) = \sigma^2$ et $\mathrm{Cov}(Y_i, Y_j) = 0$ si $i \neq j$.
Etant données n observations $(x_i, y_i)_{i=1,\ldots,n}$, **ajuster** le modèle linéaire simple, équivaut à estimer les trois paramètres, β_0, β_1 et σ^2.

Remarque 12 *Forme matricielle du modèle.*

Soient $Y = {}^t(Y_1, \ldots, Y_n)$, $\beta = {}^t(\beta_0, \beta_1)$, $\varepsilon = {}^t(\varepsilon_1, \ldots, \varepsilon_n)$ *et* $x = {}^t \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix}$. *Le modèle (3.1) s'écrit alors :*

$$\boxed{Y = x \cdot \beta + \varepsilon} .$$

Pour estimer les paramètres β_0 et β_1, on utilise la **méthode des moindres carrés** qui consiste à minimiser, par rapport à (β_0, β_1), la somme des écarts quadratiques entre les valeurs observées $(y_i)_i$ et les valeurs calculées par la modèle $(\beta_0 + \beta_1 \cdot x_i)_i$:

$$\boxed{\underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\text{Minimiser}} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 \cdot x_i))^2}$$

Notations

$\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$; $\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$; $s_x^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$; $s_y^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$;

$s_{xy} = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ qui est l'estimateur de la covariance de (X, Y) dans le cas où la variable explicative est aléatoire.

On notera $(\widehat{\beta_0}, \widehat{\beta_1})$ la solution du problème de minimisation ; $\widehat{\beta_0}$ est l'estimateur de β_0 et $\widehat{\beta_1}$ celui de β_1.

Définition 24 Le **coefficient de corrélation linéaire** empirique entre x et y est défini par :

$$r = \frac{s_{xy}}{s_x \cdot s_y} .$$

Théorème 19 Propriétés des estimateurs $\widehat{\beta_0}$ et$\widehat{\beta_1}$

(a) Les estimateurs des moindres carrés, notés $\widehat{\beta_0}$ et $\widehat{\beta_1}$, sont définis par :

$$\widehat{\beta_0} = \overline{y} - \widehat{\beta_1} \cdot \overline{x}, \widehat{\beta_1} = \frac{s_{xy}}{s_x^2}.$$

(b) Ces estimateurs sont **sans biais et efficaces** (i.e : de tous les estimateurs linéaires et sans biais de β_0 et de β_1, ils sont ceux qui ont la variance minimum : théorème de Gauss-Markov).

(c) $\widehat{\sigma^2} \underset{\text{notée}}{=} s^2 = \dfrac{\sum_{i=1}^{n}(y_i - (\widehat{\beta_0} + \widehat{\beta_1} \cdot x_i))^2}{n-2}$ est un estimateur sans biais de σ^2, mais non efficace.

(d) $\text{Var}\left(\widehat{\beta_1}\right) = \dfrac{\sigma^2}{(n-1)s_x^2}$ et $\text{Var}\left(\widehat{\beta_0}\right) = \dfrac{\sigma^2}{n} + \overline{x}^2 \,\text{Var}\left(\widehat{\beta_1}\right)$.

Exercice 33

Démonstration des propositions (a), (b), (d).

Supposons que la variable explicative varie dans l'intervalle borné $[a, b]$; pour n fixé, où aura-t-on intérêt à choisir les valeurs $(x_i)_{i=1,\ldots,n}$?

Remarque 13 *Les variances de $\widehat{\beta_0}$ et de $\widehat{\beta_1}$ (voir le théorème 19) qui mesurent la qualité de l'estimation de β_0 et β_1, sont toutes deux fonctions inverses de s_x^2. En conséquence, si l'on est sûr du caractère linéaire du modèle, on aura intérêt à répartir les valeurs x_i de la variable indépendante x, de façon équitable, aux deux extrémités du domaine afin de maximiser $\sum_i (x_i - \overline{x})^2$ et afin de minimiser les variances de $\widehat{\beta_0}$ et $\widehat{\beta_1}$.*

1.2 Modèle linéaire gaussien simple

Si l'on souhaite poursuivre plus avant l'étude statistique de la régression linéaire, par des tests de validation du modèle, de significativité des coefficients, ou même de comparaison de deux modèles linéaires, il est nécessaire de faire l'hypothèse supplémentaire de normalité des résidus ε_i et donc des v.a Y_i. Désormais on supposera donc l'hypothèse supplémentaire :

$\mathbf{R}_4$. $\forall i, \varepsilon_i \sim \mathcal{N}(0; \sigma^2)$ équivalent à $\forall i, Y_i \sim \mathcal{N}(\beta_0 + \beta_1 \cdot x_i; \sigma^2)$.

Théorème 20 Sous les hypothèses R_1, R_2, R_3, R_4, l'estimateur des moindres carrés s'identifie à l'estimateur de vraisemblance.

Preuve

Soit $f_{Y_i}(y) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\dfrac{(y - (\beta_0 + \beta_1 \cdot x_i))^2}{2\sigma^2}\right\}$ la densité des Y_i. La vraisemblance de $(Y_1, \ldots, Y_n)$ est donnée par le produit des densités $f_{Y_i}(y)$, les ε_i étant normaux et décorrélés, donc indépendants.

$$L(\underline{x}; \beta_0, \beta_1) = \sigma^{-n}.(2\pi)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 \cdot x_i))^2\right\}$$

On constate que : Maximiser $\underset{\beta_0,\beta_1}{} L(\underline{x};\beta_0,\beta_1) \Longleftrightarrow$ *Minimiser* $\underset{\beta_0,\beta_1}{} \sum_{i=1}^{n}(y_i-(\beta_0+\beta_1\cdot x_i))^2$. *D'où le résultat attendu.*
L'estimateur du maximum de vraisemblance permet de calculer explicitement un estimateur de σ^2 *:*

$$\begin{aligned}\ln L(\underline{x};\beta_0,\beta_1) &= Cte-\frac{n}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1\cdot x_i))^2\\ \Longrightarrow \frac{\partial}{\partial\sigma^2}\ln L(\underline{x};\beta_0,\beta_1) &= -\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1\cdot x_i))^2\end{aligned}$$

D'où

$$\frac{\partial}{\partial\sigma^2}\ln L(\underline{x};\beta_0,\beta_1)=0\Longrightarrow\widehat{\sigma^2}=\frac{1}{n}\sum_{i=1}^{n}(y_i-(\widehat{\beta_0}+\widehat{\beta_1}\cdot x_i))^2=\frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon_i^2}$$

■

Aux propriétés des estimateurs des moindres carrés de β_0, β_1 et σ^2 rassemblées dans le théorème 19, on adjoint celles provenant du caractère gaussien centré des résidus, résumées dans le théorème suivant.

Théorème 21

(a) $\widehat{\beta_0}\sim\mathcal{N}(\beta_0;\sigma_{\beta_0})$ avec $\sigma^2_{\beta_0}=\sigma^2\left(\dfrac{\overline{x}^2}{\sum_i(x_i-\overline{x})^2}+\dfrac{1}{n}\right)$

(b) $\widehat{\beta_1}\sim\mathcal{N}(\beta_1;\sigma_{\beta_1})$ avec $\sigma^2_{\beta_1}=\dfrac{\sigma^2}{\sum_i(x_i-\overline{x})^2}$

(c) $(n-2)\dfrac{\widehat{\sigma^2}}{\sigma^2}\sim\chi^2$ à $(n-2)$ ddl.

(d) $\widehat{\beta_0}$ et $\widehat{\beta_1}$ sont indépendants des résidus estimés $\widehat{\varepsilon_i}$.

(e) Les estimations de $\sigma^2_{\beta_0}$ et de $\sigma^2_{\beta_1}$ sont :

$$\widehat{\sigma^2_{\beta_0}}\equiv s^2_{\beta_0}=\left(\frac{1}{n}+\frac{\overline{x}^2}{\sum_i(x_i-\overline{x})^2}\right)s^2 \qquad \widehat{\sigma^2_{\beta_1}}\equiv s^2_{\beta_1}=\frac{s^2}{\sum_i(x_i-\overline{x})^2}\ .$$

Remarque 14 *Les résultats précédents permettent de construire des intervalles de confiance des paramètres* β_0 *et* β_1*. En effet, d'après les propositions (a) et (b) du théorème précédent, on a :*

$$\frac{\widehat{\beta_1}-\beta_1}{\sigma_{\beta_1}}\sim\mathcal{N}(0;1) \quad ; \quad \frac{\widehat{\beta_0}-\beta_0}{\sigma_{\beta_0}}\sim\mathcal{N}(0;1)\ .$$

Si σ^2 *est inconnu et estimé par* s^2*, on a :*

$$\frac{\widehat{\beta_1}-\beta_1}{\widehat{\sigma_{\beta_1}}}\sim Student(n-2) \quad ; \quad \frac{\widehat{\beta_0}-\beta_0}{\widehat{\sigma_{\beta_0}}}\sim Student(n-2)\ .$$

On peut donc facilement construire les intervalles de confiance, bilatéraux à un niveau de confiance $(1-\alpha)$ *donné, de chacun des paramètres :*

$$I_\alpha(\beta_0) = \left[\widehat{\beta_0} - t_{1-\frac{\alpha}{2};n-2}\widehat{\sigma_{\beta_0}}; \widehat{\beta_0} + t_{1-\frac{\alpha}{2};n-2}\widehat{\sigma_{\beta_0}}\right]$$
$$I_\alpha(\beta_1) = \left[\widehat{\beta_1} - t_{1-\frac{\alpha}{2};n-2}\widehat{\sigma_{\beta_1}}; \widehat{\beta_1} + t_{1-\frac{\alpha}{2};n-2}\widehat{\sigma_{\beta_1}}\right]$$

où $t_{1-\alpha/2;n-2}$ *est le fractile d'ordre* $(1-\alpha/2)$ *de la loi de Student à* $(n-2)$ *d.d.l.*

Nous allons maintenant construire un critère permettant d'évaluer la qualité d'ajustement des données $(x_i, y_i)_{i=1,\dots,n}$ par le modèle linéaire $y = \widehat{\beta_0} + \widehat{\beta_1}x$. Cherchons tout d'abord à décomposer la variance des y_i (autour de leur moyenne $\overline{y}$) en une somme de deux autres variances.

Théorème 22 (Equation d'analyse de la variance)

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - \overline{y})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

où $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1}x_i$, pour tout $i = 1, \dots, n$.

La quantité $\dfrac{1}{n}\sum_i (y_i - \overline{y})^2$ est la **variance totale** des observations y_i.

La quantité $\dfrac{1}{n}\sum_i (\widehat{y_i} - \overline{y})^2$ est la **variance des valeurs ajustées** par le modèle.

La quantité $\dfrac{1}{n}\sum_i (y_i - \widehat{y_i})^2$ est la **variance des résidus estimés** $\widehat{\varepsilon_i}$ dite aussi **variance résiduelle**.

Notation

$$\begin{aligned}
SSE \text{ (sum of squared errors)} &\equiv \text{somme des carrés des résidus} \\
&\equiv \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \\
SST \text{ (total sum of squared)} &\equiv \text{somme totale des carrés} \\
&\equiv \sum_{i=1}^{n}(y_i - \overline{y})^2 \\
SSR \text{ (regression sum of squares)} &\equiv \text{somme des carrés des écarts entre la moyenne et les valeurs ajustées} \\
&\equiv \sum_{i=1}^{n}(\widehat{y_i} - \overline{y})^2
\end{aligned}$$

Remarque 15 *Avec les notations précédentes, l'équation d'analyse de la variance s'écrit :*

$$SST = SSR + SSE\ .$$

Exercice 34

Un chimiste relève la concentration d'un produit en fonction de la quantité d'eau qu'il apporte au mélange. Il s'agit pour lui de dire si oui ou non la concentration dépend de la quantité d'eau, connaissant les relevés suivants :

x (quantité d'eau)	0	2	4	6	8	10	12	14	16	18	20	22	24
y (concentration)	0	0.1	0.2	0.4	0.1	0.6	0.7	0.1	1.1	0.8	0.6	1.6	0.7
x (quantité d'eau)	26	28	30	32	34	36	38	40					
y (concentration)	1.2	1.9	0.3	1.9	2	0.3	2.6	1.9					

1) Déterminer $\widehat{\beta_0}$, $\widehat{\beta_1}$ et $\widehat{\sigma^2}$ et les intervalles de confiance au niveau de confiance 95% de β_0 et β_1. Représenter le nuage de points et la droite de régression. Quelles hypothèses peut-on envisager ? Calculer le coefficient de détermination R^2 (cf Définition ci-dessous).

2) Le chimiste modifie la dose et la composition de la concentration X sur la quantité d'eau et refait les mêmes expériences. Il obtient le tableau suivant :

x (quantité d'eau)	0	2	4	6	8	10	12	14	16	18	20	22	24
y (concentration)	5	5.9	4.3	4.8	6.2	4.7	4.7	6.3	5	4.6	6.3	5.4	4.6
x (quantité d'eau)	26	28	30	32	34	36	38	40					
y (concentration)	6.3	5.8	4.6	6.2	6.2	4.7	6	6.6					

Confirmer ou infirmer les hypothèses formulées en 1).

A-t-on une idée de la dépendance des deux séries ?

Définition 25 Le **coefficient de détermination** R^2 du modèle linéaire est égal au rapport de la variance des valeurs ajustées sur la variance totale des observations :

$$R^2 = \frac{\sum_i (\widehat{y_i} - \overline{y})^2}{\sum_i (y_i - \overline{y})^2} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\text{Variance résiduelle}}{\text{Variance totale}}.$$

Interprétation :

1. Plus la variance résiduelle est proche de 0, plus R^2 est proche de 1 : R^2 est donc un indicateur de qualité de l'ajustement.
2. Il est facile d'établir que R^2 est égal au carré du coefficient de corrélation $\rho = \frac{s_{xy}}{s_x s_y}$.

1.3 Tests sur les paramètres β_0 et β_1

(A) Test de proximité de β_1 avec 0, appelé test de significativité de β_1

Il répond à la question : «La variable explicative x a-t-elle une effet significatif sur Y ?» autrement dit : «Peut-on considérer β_1 significativement différent de 0 ?»

Définissons les hypothèses du test : $\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$.

Il existe quatre méthodes équivalentes qui permettent de décider d'accepter ou de rejeter H_0, à un niveau de confiance $(1-\alpha)$ fixé à l'avance. Nous présentons brièvement les deux plus classiques.

(A1) Si l'intervalle de confiance de β_1 au niveau de confiance $(1-\alpha)$ ne contient pas 0, on rejette H_0 $(\beta_1 = 0)$ au risque α.

(A2) Si $|\frac{\widehat{\beta_1}}{s_{\beta_1}}| \leqslant t_{\frac{\alpha}{2}}(n-2)$ (fractile d'ordre $\frac{\alpha}{2}$ de la loi de Student à $(n-2)$ ddl), on accepte H_0.

(B) Test de comparaison de β_1 avec une valeur donnée b_1

Posons : $\begin{cases} H_0 : \beta_1 = b_1 \\ H_1 : \beta_1 \neq b_1 \end{cases}$.

A un niveau de confiance $(1-\alpha)$, on accepte H_0 si $\frac{|b_1 - \widehat{\beta_1}|}{s_{\beta_1}} \leqslant t_{1-\frac{\alpha}{2}}(n-2)$; sinon on rejette H_0.

(C) Test de comparaison des pentes de deux régressions distinctes

Ces régressions sont par exemple, effectuées sur des échantillons issus de deux populations distinctes : désignons par $\widehat{\beta_1'}$ et $\widehat{\beta_1''}$, les estimations respectives des pentes β_1' et β_1''. D'après le théorème 23, on teste la significativité de la différence des deux v.a normales $\widehat{\beta_1'}$ et $\widehat{\beta_1''}$ par rapport à 0.

Deux cas doivent être envisagés :

1er cas : Les variances $s'^2_{\beta_1}$ et $s''^2_{\beta_2}$ sont considérées comme égales, après application du test de Fisher-Snedecor : dans ce cas, les variances seront considérées comme égales à leur estimation commune $s^2 = \dfrac{\sum_i (y_i' - \widehat{y_i'})^2 + \sum_j (y_j' - \widehat{y_j''})^2}{n_1 + n_2 - 4}$ où n_1 et n_2 sont les tailles respectives des échantillons $(y_i')_i$ et $(y_j'')_j$.

Posons : $\begin{cases} H_0 : \beta'_1 = \beta''_1 \\ H_1 : \beta'_1 \neq \beta''_1 \end{cases}$

Si $\dfrac{\left|\widehat{\beta'_1} - \widehat{\beta''_1}\right|}{\sqrt{s^2\left(\dfrac{1}{\sum_i (x'_i - \bar{x'})^2} + \dfrac{1}{\sum_j (x''_j - \bar{x''})^2}\right)}} \leqslant t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 4)$, alors on accepte H_0, au niveau de confiance $(1-\alpha)$.

2ème cas : Le test dee comparaison des variances conclut à leur inégalité : on utilise alors les tests d'Aspin-Welch (qu'il n'est pas nécessaire de développer ici).

Les tests de significativité ou de comparaison portant sur l'ordonnée à l'origine β_0, sont strictement analogues aux précédents.

Exercice 35

1. Définir le test de comparaison de β_0 à une valeur fixée à l'avance b_0.
2. Suite de l'exercice 34 : tester les hypothèses $\beta_0 = 0$ et $\beta_1 = 0$. Déterminer R^2 dans chacun des cas.

1.4 Prédiction

Etant donnée une valeur $x = x^*$ différente des x_i, $\widehat{y(x^*)} = \widehat{\beta_0} + \widehat{\beta_1} \cdot x^*$ est l'estimation de $y(x^*)$ noté y^*. Quelle est la précision de cette estimation ? A une confiance $(1-\alpha)$ donnée, on définit l'intervalle de confiance de y^* :

$$[\widehat{\beta_0} + \widehat{\beta_1} \cdot x^* - t_{1-\frac{\alpha}{2}}(n-2)s^*; \widehat{\beta_0} + \widehat{\beta_1} \cdot x^* + t_{1-\frac{\alpha}{2}}(n-2)s^*]$$

où $s^* = s\sqrt{\dfrac{1}{n} + \dfrac{(x^* - \overline{x})^2}{\sum_i (x_i^* - \overline{x})^2} + 1}$ et $s^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \widehat{y_i})^2$.

Si l'on fait varier x^*, alors y^* varie à l'intérieur d'une région de confiance délimitée par deux branches d'hyperboles (par rapport à la droite de régression).

Remarque 16 *La précision de l'estimation mesurée par s^*, est fonction de $(x^* - \overline{x})^2$: c'est donc au voisinage de $\overline{x}$ que les prédictions sont les meilleures, ce qui est visible sur le graphe ci-dessus. En pratique, les prédictions ne sont valides qu'à l'intérieur du domaine d'observation délimité par $[\min_i(x_i); \max_i(x_i)]$: on évitera donc les extrapolations hors de ce domaine.*

Exercice 36

Soit le tableau de données

x	0	10	20	30	40	50	60	70	80
y	1,76	1,69	1,61	1,54	1,46	1,4	1,32	1,25	1,2

1. Déterminer $\bar{X}$; $\bar{Y}$; s_X^2 ; s_Y^2 ; $s_{X,Y}$; R.
2. Estimer β_0 et β_1. Déterminer leurs intervalles de confiance.
3. Calculer les variances totales des valeurs ajustées et des résidus.
4. Prédire les valeurs de y^* pour x^* égal à 35 et 150. Déterminer, comparer et interpréter leurs intervalles de prédiction.

2 Régression non-linéaire

2.1 Modèle général

La variable aléatoire réelle Y est cette fois dépendante de la variable contrôlée x selon le modèle général :

$$Y = f(x;\theta) + \varepsilon, \qquad x \in D \subseteq \mathbb{R}^k$$

où : θ désigne le vecteur des p paramètres inconnus $(\theta_1, \ldots, \theta_p)$; A tout $x_i \in D$, on associe la

f une fonction connue non linéaire par rapport à θ ;

ε la v.a résiduelle.

v.a. $Y_i = f(x_i;\theta) + \varepsilon_i$.

Hypothèses : Les variables résiduelles ε_i sont décorrélées de loi normale $\mathcal{N}(0; \sigma^2 \cdot v(x_i;\theta))$ où v est une fonction connue, différentiable par rapport à θ.

Ainsi $\forall i = 1, 2, \ldots, n$, $Y_i \sim \mathcal{N}(f(x_i;\theta); \sigma^2 \cdot v(x_i;\theta))$.

Remarque 17 *La fonction $f(x;\theta)$ est intrinséquement non linéaire par rapport à θ, si elle résiste à toute transformation de linéarisation.*

Par exemple, une fonction de la forme $f(x;\theta) = \sum_{j=1}^{p} \theta_j f_j(x)$ où les $f_j(\bullet)$ sont des fonctions connues ne dépendant pas des θ_j, est une fonction linéaire par rapport à $\theta = (\theta_1, ..., \theta_p)$.

Les fonctions :

$$f_1(x;\theta) = \theta_1 e^{\theta_2 x} \text{ et } f_2(x;\theta) = \frac{\theta_1 x}{1 + \theta_2 e^{\theta_3 x}}$$

ne sont pas intrinséquement non linéaires par rapport à θ, car elles sont linéarisables par les transformations respectives :

$$g_1 = \log\left(f_1(x;\theta)\right) \text{ et } g_2 = \log\left(\frac{f_2(x;\theta)}{1 - f_2(x;\theta)}\right) .$$

Par contre, la fonction $f_3(x;\theta) = \theta_1 + \theta_2 e^{\theta_3 x}$ est intrinséquement non linéaire. On pourra résoudre un problème de régression intrinséquement non linéaire en le transformant en un modèle linéaire, mais il est préférable de le traiter dans le cadre de la théorie de la régression non linéaire.

Notation θ^* désignera la valeur exacte inconnue du paramètre θ.

2.2 Estimation des paramètres

2.2.1 Modèle à variance constante (pour tout i, $\mathrm{Var}(\varepsilon_i) = \sigma^2$)

Sous l'hypothèse de normalité, l'estimateur du maximum de vraisemblance est identique à celui des moindres carrés. On cherche donc le vecteur $\widehat{\theta}$ qui minimise :

$$\sum_{i=1}^{n} (y_i - f(x_i;\theta))^2 \underset{\text{notée}}{=} SCR(\theta) \qquad \text{(Somme des Carrés des Résidus)}$$

Théorème 23 L'estimateur $\widehat{\sigma^2}$ de la variance σ^2 est défini par : $\dfrac{SCR(\theta)}{n-p}$.

Distinguons deux sous-cas :

(a) Expériences répétées : pour toute valeur x_i de x, on fait m mesures $y_{i,j}$ de la réponse Y_i ; dans ce cas, $n = m \cdot k$. Les erreurs ε_{i_j} sont supposées indépendantes, de même loi d'espérance nulle et de variance égale à σ^2.

Théorème 24 Sous les hypothèses précédentes (a), les propositions suivantes sont vraies :

1. Si $\theta \mapsto f(x;\theta)$ est continue et injective, alors $\widehat{\theta_n}$ converge presque sûrement vers θ^* et $\widehat{\theta_n}$ est un estimateur asymptotiquement sans biais, c'est-à-dire que $\lim_{n\to+\infty} \mathbb{E}\left(\widehat{\theta_n} - \theta^*\right) = 0$.
2. Si, de plus, $f(\bullet;\theta)$ est C^2, de hessien convergent uniformément au voisinage de θ^* ainsi que ${}^tDf(\bullet;\theta) \cdot Df(\bullet;\theta)$, la limite étant une matrice définie positive $k\Gamma(\theta^*)$; alors $\widehat{\theta_n}$ est asymptotiquement gaussien et $\sqrt{n}(\widehat{\theta_n} - \theta^*)$ converge en loi vers une loi gaussienne $\mathcal{N}(0;\sigma^2\Gamma^{-1}(\theta^*))$.

Remarque 18 *La convexité de $SCR(\theta)$ n'étant plus acquise, l'unicité du minimum ne l'est pas non plus. D'où la nécessité de pratiquer avec précaution les méthodes numériques de minimisation, pouvant converger vers des minima locaux. Se pose alors le problème de l'initialisation de la procédure d'optimisation : il s'agira de choisir une valeur θ_0 d'initialisation proche de θ^*. Si on a pas (ou peu) d'information sur θ^*, on pourra calculer $SCR(\theta)$ sur les noeuds d'un maillage discret rectangle inclus dans $\mathbb{R}^p$ et dont on est sûr qu'il contient θ^* ; ensuite, il suffira de prendre comme initialisation θ_0, le noeud du maillage qui minimise $SCR(\theta)$.*

Les méthodes classiques de Gauss-Newton ou de Newton-Raphson donnent la plupart du temps de bons résultats (cette dernière est toutefois sensible au choix initial).

Une méthode de minimisation, particulièrement adaptée au cas où existent plusieurs minima locaux, est la méthode du recuit simulé.

Exemple 10 Soit à estimer les coefficients β_1 et β_2 du modèle $Y = \beta_1 \cdot e^{-\beta_2 x}$. On dispose du fichier,

x	0	1	2	3	4	5	6	7	8	9	10
y	1,85	2,3	2,61	2,65	3,25	3,35	3,44	4,2	4,7	5	5,2

Méthode de la linéarisation du problème : soit $z = \log(y)$ alors $z = \beta_1' - \beta_2 x := \log(\beta_1) - \beta_2 x$.

Après traitement par régression linéaire, on obtient :

$$\widehat{\beta_1'} = 0.702, \quad \widehat{\beta_2} = 0.1 \text{ d'où } \widehat{\beta_1} = e^{0.702} \text{ et } R^2 = 0.975$$

(b) Cas d'expériences non répétées : Sous des hypothèses dont certaines sont analogues à celles du théorème précédent, on obtient les résultats :

$\widehat{\theta_n}$ est fortement consistant, asymptotiquement gaussien et sans biais. De plus, $\sqrt{n}(\widehat{\theta_n} - \theta^*) \sim \mathcal{N}(0; \sigma^2 \Gamma^{-1}(\theta^*))$, où $\Gamma(\theta^*)$ est la limite uniforme de $\frac{1}{n}\, {}^t Df(\bullet;\theta) \cdot Df(\bullet;\theta)$.

2.2.2 Modèle à variance : $\sigma_i^2 = \omega_i \cdot \sigma^2$

Les ω_i sont connus et positifs. On utilise la méthode des moindres carrés pondérés qui consiste à minimiser :

$$SCR(\theta) = \sum_{i=1}^{n} \frac{1}{\omega_i} (y_i - f(x_i;\theta))^2$$

Là encore, sous l'hypothèse de normalité, l'estimateur des moindres carrés pondérés est équivalent à l'estimateur du maximum de vraisemblance.

Il est souvent possible de transformer la réponse Y en une réponse $Z = g(Y)$ de variance approximativement constante, ce qui nous ramène au cas précédent.

Par exemple, si $\mathrm{Var}(Y_i) = \sigma^2 \mathbb{E}(Y_i)$ et si $\mathbb{E}(Y_i)$ est grand, on démontre que la transformation «racine carré» convient ; ainsi, $Z = \sqrt{Y}$ et $V(\sqrt{Y}) \sim$ constante.

Exercice 37

Démontrer ces résultats.

2.3 Détermination des intervalles de confiance sous les hypothèses de normalité et d'équivariance résiduelle (dite aussi homoscédasticité)

2.3.1 Méthode d'approximation linéaire

Cette méthode convient d'autant plus que l'échantillon est grand et le modèle $f(\bullet;\theta)$ quasi-linéaire en θ, au voisinage de l'estimation $\widehat{\theta}$.

Elle est basée sur l'approximation de $f(\bullet,\theta)$ au premier ordre du développement de Taylor :

$$f(\bullet;\theta) \simeq f(\bullet;\widehat{\theta}) + Df(\bullet;\widehat{\theta}) \cdot \left(\theta - \widehat{\theta}\right) \ .$$

On détermine la région de confiance, de forme ellipsoïdale, grâce au théorème suivant :

Théorème 25 A un niveau de confiance $(1-\alpha)$ donné, la région de confiance de θ^* est formé de l'ensemble des θ défini par :

1. si σ^2 connue : ${}^t\left(\theta - \widehat{\theta}\right) \cdot \widehat{V^{-1}} \cdot \left(\theta - \widehat{\theta}\right) \leqslant \sigma^2 \left(\chi^2_p\right)^{-1}(1-\alpha)$.
2. si σ^2 inconnue : ${}^t\left(\theta - \widehat{\theta}\right) \cdot \widehat{V^{-1}} \cdot \left(\theta - \widehat{\theta}\right) \leqslant \widehat{\sigma^2} \left(\chi^2_p\right)^{-1}(1-\alpha)$.

On a désigné ${}^tDf(\bullet;\widehat{\theta}) \cdot \left(Df(\bullet;\widehat{\theta})\right)^{-1}$ par $\widehat{V^{-1}}$.

2.3.2 Méthode des isocontours

On traitera le cas avec mesures non répétées.

Si σ^2 est inconnue et estimée par $\frac{1}{n}\sum_{i=1}^n \left(y_i - f(x_i;\theta)\right)^2 \equiv \widehat{\sigma^2}$.

On considère la statistique de déviance définie par :

$$D = 2\left(\mathcal{L}\left(\widehat{\theta};\widehat{\sigma^2}\right) - \mathcal{L}\left(\theta;\widehat{\sigma^2}\right)\right)$$

où $\mathcal{L}\left(\theta;\widehat{\sigma^2}\right)$ est le logarithme de la vraisemblance.

Théorème 26 La statistique D est asymptotiquement décrite par une v.a χ^2 à p d.d.l.

On en déduit la région de confiance de θ^*, au niveau $1-\alpha$:

$$\left\{\theta \ \middle|\ 2\left(\mathcal{L}\left(\widehat{\theta};\widehat{\sigma^2}\right) - \mathcal{L}\left(\theta;\widehat{\sigma^2}\right)\right) \leqslant \left(\chi^2_p\right)^{-1}(1-\alpha)\right\} \ .$$

2.4 Tests d'hypothèses

2.4.1 Test du rapport de vraisemblance

Applicable à des cas plus généraux ($\mathrm{Var}(\varepsilon_i) = v_i(\theta) \cdot \sigma^2$).

Soit à trancher pour l'une des deux hypothèses : $\begin{cases} H_0 : \theta_i = a \\ H_1 : \theta_i \neq a \end{cases}$ où a est fixé.

Notation $\widehat{\theta_1}$ l'estimateur de vraisemblance sous l'hypothèse H_1. $\widehat{\theta_0}$ l'estimateur de vraisemblance sous l'hypothèse H_0.

Théorème 27 Avec les notations précédentes, on a :

$$-2\left(\mathcal{L}\left(\widehat{\theta_0}\right) - \mathcal{L}\left(\widehat{\theta_1}\right)\right) \sim \chi_1^2 \ .$$

Si $-2\left(\mathcal{L}\left(\widehat{\theta_0}\right) - \mathcal{L}\left(\widehat{\theta_1}\right)\right) > \left(\chi_1^2\right)^{-1}(\alpha)$, alors on rejette H_0 avec un risque α.

Ce test se généralise à des tests, où H_0 et H_1 concernent un sous-ensemble des θ_i ; par exemple :

$$\begin{cases} H_0 : \theta_2 = a_2, \theta_5 = a_5, \theta_p = a_p \\ H_1 : \theta_2 \neq a_2, \theta_5 \neq a_5, \theta_p \neq a_p \ . \end{cases}$$

Remarque 19 *L'optimisation des vraisemblances, qui sont des fonctions en général fortement non linéaires, nécessitent l'utilisation d'algorithmes, tels que celui de Marquardt, qui peuvent converger difficilement ou même ne pas converger du tout. En conséquence, dans les cas (fréquents) où la variance résiduelle est indépendante des paramètres θ_i, on pourra avoir recours au test de Fisher-Snedecor, en n'oubliant pas que sa validité sera d'autant plus contestable que l'approximation linéaire du modèle l'est.*