Fiches Travaux Dirigés
Statistiques inférentielles

CQLS
cqls@upmf-grenoble.fr
http://cqls.upmf-grenoble.fr/

Remarques autour du cours

- **Points-clés de ce cours** : Voici quelques éléments qui font l'originalité de ce cours qui s'adresse tout particulièrement à des étudiants n'ayant pas nécessairement un bagage mathématique très imposant :
  1. *Le **langage mathématique** (utile pour interpréter les résultats des mathématiciens) avant les **techniques mathématiques** (utiles pour développer de nouveaux résultats mathématiques)*
  2. *Une **A**pproche **E**xpérimentale des **P**robabilités (plus intuitive) introduite pour mieux décoder les résultats de l'**A**pproche **M**athématique des **P**robabilités classiquement enseignée*
  3. *Le **cadre asymptotique** (i.e. nombre de données observées suffisamment grand) bien plus réaliste en pratique avant le **cadre Gaussien** (i.e. l'origine des données est d'un type connu mais particulier)*
  4. Outil d'aide à la décision présenté sous sa forme la plus pratique et universelle à savoir la **p-valeur**.
- **Esprit du cours** : Nous avons conscience qu'avoir fait le choix d'introduire un système de notation plutôt lourd mais avant tout précis, est un point qui peut effrayer l'étudiant lors des premiers cours et séances de T.D.. Cependant, il faut souligner qu'un cours alternatif classique (que nous avons déjà expérimenté avant la mise en place de celui-ci) s'appuie essentiellement sur l'**A**pproche **M**athématique des **P**robabilités nécessitant un niveau mathématique bien supérieur à celui requis dans ce cours. Le cours classique de Probabilités et Statistique était alors beaucoup plus orienté sur les Probabilités et la partie Statistique était principalement évoquée sur un plan méthodologique (avec pour conséquence un très grand risque de mauvaise utilisation). La raison principale est que techniquement parlant les probabilités concernant les variables aléaloires discrètes sont plus accessibles que celles concernant les variables aléatoires continues. Et pourtant, ce sont ces dernières qui sont le plus souvent intéressantes dans un contexte Statistitique. Bien connue des développeurs d'outils statistiques (que nous sommes), l'**A**pproche **E**xpérimentale des **P**robabilités n'est pas souvent enseignée alors qu'elle est accessible (même pour un étudiant "non matheux") puisqu'elle s'appuie sur le sens intuitif des probabilités dont nous semblons tous disposer (dans cette société où les jeux de hasard sont très appréciés). Dans ce contexte, il n'y a notamment aucune différence de traitement dans la façon d'aborder le comportement aléatoire d'une variable aléatoire qu'elle soit discrète ou continue (à la différence de l'approche classique). De plus, les difficultés mathématiques utilisées dans cette approche se limitent tout au plus à celles rencontrées lors du cours de Statistique Descriptive (enseigné généralement l'année précédente). Grâce à ces facilités, nous avons pu atteindre notre d'objectif de présenter la règle de décision d'un test d'hypothèses (nom donné par les matheux à l'outil d'aide à la décision) en fonction de la notion fondamentale sur un plan pratique de **p-valeur** ou valeur-p (traduction de p-value en anglais). Au niveau du langage mathématique, nous avons introduit un système de notation afin de rendre accessible cette approche aux étudiants. Si un étudiant nous fait confiance, il pourra se fixer comme objectif principal de maîtriser dans un premier temps ces nouvelles notations alourdies volontairement dans un but de précision puis dans un second temps de les simplifier (comme le font la plupart des mathématiciens) dès lors que son niveau de compréhension sera satisfaisant. Soulignons que l'**A**pproche **E**xpérimentale des **P**robabilités est spécifiquement adaptée à l'utilisation de l'ordinateur. Nous nous appuierons sur le logiciel `R` (libre, gratuit et accessible sous toutes les plateformes). Pour conclure, nous espérons que ce cours se fera malgré tout avec le *sourire* et *plein de bonne humeur*.
- **Les documents de cours** : Il y a trois documents proposés dans ce cours :
  1. Les supports de cours en amphithéatre (version courte imprimable et version longue à ne consulter qu'en mode présentation).
  2. Le document de T.D. qui s'articule avec les supports de cours.
  3. Le polycopié de cours rassemblant les informations importantes du cours.

  Tous les documents sont disponibles sur le site :

  `http://cqls.upmf-grenoble.fr`

  à l'onglet `Stat Inf`.
- **Organisation des documents** : Les séances de T.D. et les cours en amphithéatre s'enchaînent dans l'ordre décrit dans le tableau suivant (bien entendu à adapter selon ses préférences) :

Sujet	Cours	T.D.	Nbre séances
Présentation problématiques		Fiche **??**	1
Introduction A.E.P.	Cours 1	Fiche **??**	1
Estimation ponctuelle	Cours 2	Fiche **??**	1
Intervalle de confiance	Cours 3	Fiche **??**	1
Test d'hypothèses (construction)	Cours 4	Fiche **??**	1
A.E.P. en graphique (p-valeur)	Cours 5	Fiche **??**	1
Test d'hypothèses (méthodologie)	Cours 6-8	Fiche **??**	3

Fin

FICHE T.D. 1 Présentation des différents outils

Indications préliminaires

- *Proportion* : Une proportion d'individus ayant une caractéristique (d'intérêt) parmi une population de N individus est le nombre d'individus ayant la caractéristique divisé par la taille N de la population.
- *Moyenne* : La moyenne de l'ensemble des N données $\mathbf{z} := (z_1, z_2, \cdots, z_N)$ correspond à la somme des ces données divisée par le nombre total N de données. Elle est usuellement notée et définie par $\overline{z} := \frac{1}{N}\sum_{i=1}^{N} z_i$.
- *Proportion comme une moyenne* : La proportion d'individus ayant une caractéristique parmi une population de N individus peut être vue comme la moyenne $\overline{z}$ des $\mathbf{z} = (z_1, z_2, \cdots, z_N)$ où z_i vaut 1 lorsque l'individu i a la caractéristique et 0 sinon. Autrement dit, une moyenne de valeurs ne valant que 0 ou 1 est une proportion.
- *Nombre moyen* : Soit z_i un nombre (d'objets) associé à tout individu i de la population. Un nombre (d'objets) moyen (par individu de la population) est alors défini comme la moyenne $\overline{z}$ des nombres (d'objets) $\mathbf{z}$.
- *Echantillon* : Indépendament de son procédé de construction, un échantillon de taille n est un "paquet" de n individus extrait parmi l'ensemble de la population totale des N individus. Lorsqu'en particulier, on n'est intéressé que par une variable $\mathcal{Y}$ relative à la population des N indvidus, de manière un peu abusive mais simplifiée, on appelle population l'ensemble des N valeurs $\boldsymbol{\mathcal{Y}} := (\mathcal{Y}_1, \mathcal{Y}_2, \cdots, \mathcal{Y}_N)$ et un échantillon (de $\mathcal{Y}$) l'ensemble des n valeurs $\boldsymbol{y} := (y_1, \cdots, y_n)$ correspondant aux valeurs de $\mathcal{Y}$ pour les individus extraits de la population.
- *Estimation* : une estimation d'un paramètre inconnu θ sur un échantillon $\boldsymbol{y}$ est noté $\widehat{\theta}(\boldsymbol{y})$ s'exprimant littéralement par "estimation (ou plus généralement remplaçant) du paramètre (inconnu) θ calculée à partir du jeu de données $\boldsymbol{y}$" après avoir appliqué les conventions de notation suivantes :
  1. "$\widehat{\ }$" signifie (usuellement en Statistique) estimation (ou plus généralement, remplaçant) de la quantité sur laquelle il se trouve, ici le paramètre inconnu θ à estimer.
  2. "$(\boldsymbol{y})$" exprime la dépendance fonctionnelle (i.e. symbolisée dans le langage mathématique par les parenthèses qui servent à encadrer une valeur d'entrée à appliquer à une fonction afin de retourner une valeur de sortie) pouvant être traduite littéralement par "calculée à partir de l'échantillon $\boldsymbol{y}$". Le "$\widehat{\theta}$" est alors vu comme une fonction retournant en sortie l'estimation de θ lorsqu'en entrée il lui est donné un échantillon $\boldsymbol{y}$.

Fin

FICHE T.D. 2 Introduction aux probabilités

Indications préliminaires

- *Objectif* : L'originalité de ce cours réside essentiellement dans l'axe qui a été choisi pour présenter les probabilités. Dans un cours classique, les développements mathématiques (de nature plutôt technique) sont proposés en priorité en laissant peu de place à l'interprétation des concepts théoriques véhiculés. Cette approche pour introduire les concepts de probabilités sera par la suite appelée **A.M.P.** pour désigner **A**pproche **M**athématique des **P**robabilités. La Statistique (Inférentielle ou Inductive, celle présentée dans ce cours) repose largement sur la théorie des Probabilités, mais de part sa vocation à être largement utilisée par les praticiens sous une forme plutôt méthodologique, il s'ensuit souvent une difficulté pour ces utilisateurs à appréhender les conditions d'applicabilité et les points-clés des outils statistiques qui bien souvent s'expriment en fonction des concepts probabilistes pas toujours faciles à assimiler (compte tenu de leurs aspects mathématiques). Afin de remédier à cet incovénient, nous avons choisi de proposer une approche complémentaire, appelée **A.E.P.** pour désigner **A**pproche **E**xpérimentale des **P**robabilités, qui nous semble plus intuitive car basée sur l'expérimentation et dont la difficulté technique se limite aux outils de Statistique Desciptive présentés en première année (faciles à appréhender par les praticiens motivés surtout lorsqu'ils en ont l'utilité). L'objectif de cette fiche T.D. est essentiellement de faire le lien entre les deux approches **A.M.P.** et **A.E.P.**. Notamment, il sera essentiel de comprendre comment les praticiens pourrons être éclairés via l'**A.E.P.** sur les résultats techniques obtenus grâce à l'**A.M.P.** par les mathématiciens.
- *L'**A.E.P.** en complément de l'**A.M.P.*** : Soit Y une variable aléatoire réelle dont on suppose disposer (via l'**A.E.P.**) d'un vecteur $\mathbf{y}_{[m]} := (y_{[\cdot]})_m := (y_{[1]}, y_{[2]}, \cdots, y_{[m]})$ de m (a priori très grand) réalisations indépendantes entre elles. En théorie, on pourra aussi imaginer disposer du vecteur $\mathbf{y}_{[+\infty]} := (y_{[\cdot]})_{+\infty}$ qui est l'analogue de $\mathbf{y}_{[m]}$ avec $m \to +\infty$. Supposons aussi que $m = 10000$ expériences aient été réalisées et les m composantes de $\mathbf{y}_{[m]}$ aient été stockées dans `R` sous le vecteur nommé `yy`.

Quantité	**A.M.P.**		**A.E.P.** $(+\infty)$		**A.E.P.**		**Traitement** `R`
Probabilité	$\mathbb{P}(Y = a)$	$=$	$\overline{(y_{[\cdot]} = a)}_{+\infty}$	$\simeq$	$\overline{(y_{[\cdot]} = a)}_{m}$	$\overset{R}{=}$	`mean(yy==a)`
Probabilité	$\mathbb{P}(Y \in]a,b])$	$=$	$\overline{(y_{[\cdot]} \in]a,b])}_{+\infty}$	$\simeq$	$\overline{(y_{[\cdot]} \in]a,b])}_{m}$	$\overset{R}{=}$	`mean(a<yy & yy<=b)`
Moyenne	$\mathbb{E}(Y)$	$=$	$\overline{(y_{[\cdot]})}_{+\infty}$	$\simeq$	$\overline{(y_{[\cdot]})}_{m}$	$\overset{R}{=}$	`mean(yy)`
Variance	$\mathbb{V}ar(Y)$	$=$	$\overline{(y_{[\cdot]})^2}_{+\infty}$	$\simeq$	$\overline{(y_{[\cdot]})^2}_{m}$	$\overset{R}{=}$	`var(yy)=sd(yy)^2`
Quantile	$q_Y(\alpha)$	$=$	$q_\alpha\left((y_{[\cdot]})_{+\infty}\right)$	$\simeq$	$q_\alpha\left((y_{[\cdot]})_{m}\right)$	$\overset{R}{=}$	`quantile(yy,alpha)`

Les formules d'obtention des quantités ci-dessus pour les colonnes **A.M.P.** et **A.E.P.** n'ont pas été fournies. Celles concernant l'**A.M.P.** requiert un niveau plutôt avancé en mathématiques et diffèrent selon la nature (discrète ou continue) de Y. Un point fort de l'**A.E.P.** est que les formules d'obtentions ne dépendent pas de la nature de Y et sont normalement déjà connues en 1ère année dans le cours de Statistique Descriptive (pour rappel, voir polycopié de notre cours).
IMPORTANT : L'objectif principal de la fiche T.D. est l'assimilation des concepts décrits dans le tableau ci-dessus.

- *Quelques résultats sur **A.M.P.*** : Soient Y, Y_1 et Y_2 trois variables aléatoires réelles (v.a.r.) et λ un réel.
  Fonction de répartition $F_Y(y) := \mathbb{P}(Y \leq y)$: Dans l'**A.M.P.**, elle permet de calculer, pour tout $a \leq b$:
  $$\mathbb{P}(a < Y \leq b) = \mathbb{P}(Y \leq b) - \mathbb{P}(Y \leq a) = F_Y(b) - F_Y(a).$$
  Moyenne (théorique) : $\mathbb{E}(\lambda \times Y) = \lambda \times \mathbb{E}(Y)$ et $\mathbb{E}(Y_1 + Y_2) = \mathbb{E}(Y_1) + \mathbb{E}(Y_2)$
  Variance : $\mathbb{V}ar(\lambda \times Y) = \lambda^2 \times \mathbb{V}ar(Y)$ et $\mathbb{V}ar(Y_1 + Y_2) = \mathbb{V}ar(Y_1) + \mathbb{V}ar(Y_2)$
  où Y_1 et Y_2 sont en plus supposées indépendantes.

Fin

Quelques commentaires

- Un étudiant suivant ce cours n'est pas censé comprendre comment les résultats de l'**A.M.P.** ont été mathématiquement obtenus. Ils sont généralement proposés sans démonstration. Sa mission est en revanche de savoir comment les vérifier via l'**A.E.P.** en prenant soin de bien les interpréter. Autrement dit, l'**A.E.P.** permet à un praticien de mieux comprendre les tenants et les aboutissants des outils statistiques (qu'il utilise) développés dans le contexte de l'**A.M.P.**.
- Afin d'éviter de surcharger l'étude de l'**A.E.P.**, il a été décidé dans ce cours d'étaler son introduction en deux étapes. La première qui vous a été présentée dans cette fiche est naturellement complétée par une deuxième étape qui s'appuie sur la représentation graphique des répartitions de $\boldsymbol{y}_{[m]} := \left(y_{[\cdot]}\right)_m$ (avec m généralement très grand). Cette étape est présentée en Annexe. Un étudiant motivé pourra à sa guise choisir de compléter sa connaissance sur l'**A.E.P.** en lisant dès à présent la fiche Annexe **??** en Annexe consacrée à l'**A.E.P.** dans sa version "graphique". Il est toutefois important de rappeler que les 2 fiches T.D. **??** et **??** suivantes ne s'appuient que sur les outils présentées dans la fiche T.D. présentée ici.
- Dans la suite du cours (nous en avons déjà eu un aperçu dans la fiche introductive précédente), la plupart des variables aléatoires d'intérêt, appelées statistiques, seront de la forme $T := t(\boldsymbol{Y})$ où t est une fonction s'appliquant à $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ qui représente le "futur" échantillon, seule source d'aléatoire dans la variable aléatoire $t(\boldsymbol{Y})$. C'est en effet le cas pour l'estimation d'un paramètre inconnu θ qui s'écrit $\widehat{\theta}(\boldsymbol{y})$ lorsqu'il est évalué à partir de l'échantillon que l'on obtient le **Jour J** (i.e. jour d'obtention des données) et qui est la réalisation de $\widehat{\theta}(\boldsymbol{Y})$ représentant le procédé d'obtention de l'estimation à partir du "futur" échantillon $\boldsymbol{Y}$. L'étude **A.E.P.** consistera alors à construire m échantillons $\left(\boldsymbol{y}_{[\cdot]}\right)_m$ où $\boldsymbol{y}_{[k]} := (y_{1,[k]}, \cdots, y_{n,[k]})$ représente le $k^{ème}$ échantillon de taille n construit parmi les m. Le comportement aléatoire d'une statistique $T := t(\boldsymbol{Y})$ sera donc appréhendé via l'**A.E.P.** en proposant m réalisations indépendantes $\left(t_{[\cdot]}\right)_m := \left(t(\boldsymbol{y}_{[\cdot]})\right)_m$ avec $t_{[k]} := t(\boldsymbol{y}_{[k]})$ la $k^{ème}$ réalisation de T parmi les m.

Fin

FICHE T.D. 3 Estimation ponctuelle et par intervalle de confiance

Indications préliminaires

- *Objectif* : Dans la fiche d'introduction, le cadre de ce cours de Statistique Inférentielle a été posé. En question préliminaire, nous aurons, pour chaque problématique considérée, à identifier le paramètre d'intérêt (noté θ en général lorsque la problématique n'est pas encore précisée) et à bien prendre conscience que ce dernier est **inconnu**. A partir d'un échantillon $\boldsymbol{y}$ récolté le **jour J** (cette appellation sera utilisée tout au long de ce cours), nous aurons alors comme objectif de proposer une estimation, notée $\widehat{\theta}(\boldsymbol{y})$ (pour bien exprimer la dépendance en l'échantillon $\boldsymbol{y}$), afin d'avoir une idée sur l'ordre de grandeur de θ (inconnu). Dans un deuxième temps, nous réaliserons que ce type d'estimation ponctuelle (i.e. un paramètre inconnu estimé par une unique valeur estimée) n'est pas satisfaisant en termes de confiance que l'on peut apporter à l'estimation. Le statisticien se doit alors de proposer à partir du même échantillon $\boldsymbol{y}$, un niveau de qualité de l'estimation $\widehat{\theta}(\boldsymbol{y})$. L'**erreur standard** ("standard error" en anglais) est alors introduite s'exprimant comme une estimation de l'écart-type (i.e. indicateur de variabilité) de la "future" estimation $\widehat{\theta}(\boldsymbol{Y})$ (à partir du futur échantillon $\boldsymbol{Y}$) ayant pour réalisation $\widehat{\theta}(\boldsymbol{y})$ le **jour J**. En appliquant le système notation Norme CQLS (voir polycopié de cours), cette erreur standard se note $\widehat{\sigma_{\widehat{\theta}}}(\boldsymbol{y})$. La meilleure façon de proposer une estimation tenant compte du couple d'informations $\left(\widehat{\theta}(\boldsymbol{y}), \widehat{\sigma_{\widehat{\theta}}}(\boldsymbol{y})\right)$ disponible le **jour J** est de construire un intervalle de confiance $IC_{\theta,1-\alpha}(\boldsymbol{y}) := [\widetilde{\theta}_{\inf}(\boldsymbol{y}), \widetilde{\theta}_{\sup}(\boldsymbol{y})]$ à $1-\alpha$ de niveau de confiance. Grâce à la l'**A.E.P.**, nous aurons comme mission prioritaire de bien interpréter la notion de niveau de confiance.
- *Loi de probabilité de l'écart standardisé* : Les paramètres d'intérêt considérés dans ce cours sont de manière plus ou moins directe tous reliés à la moyenne. Ainsi, dans un cadre asymptotique où nous supposerons disposer d'un nombre suffisant de données, nous pourrons hériter pleinement de la puissance du Théorème de la limite centrale que nous avons étudié précédemment (notamment dans la fiche T.D. **??** mais aussi dans la fiche Annexe **??** consacrée aux représentations graphiques des lois de probabilité). Dans le contexte de l'estimation d'un paramètre θ traité dans ce cours, il s'exprime par (n supposé suffisamment grand) :

$$\widehat{\Theta}_n := \widehat{\theta}(\boldsymbol{Y}) \overset{approx.}{\rightsquigarrow} \mathcal{N}\left(\theta, \sigma_{\widehat{\theta}}\right) \Leftrightarrow \Delta_n := \frac{\widehat{\theta}(\boldsymbol{Y}) - \theta}{\sigma_{\widehat{\theta}}} \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1).$$

où $\sigma_{\widehat{\theta}} := \sigma\left(\widehat{\theta}(\boldsymbol{Y})\right) = \sqrt{\mathbb{V}ar\left(\widehat{\theta}(\boldsymbol{Y})\right)}$ est l'écart-type de la "future" estimation $\widehat{\theta}(\boldsymbol{Y})$. Cependant, en général, le paramètre $\sigma_{\widehat{\theta}}$ est lui-même inconnu et doit être estimé par $\widehat{\sigma_{\widehat{\theta}}}(\boldsymbol{y})$) correspondant à l'erreur standard. Un résultat applicable dans le cas où $\sigma_{\widehat{\theta}}$ est inconnu, est le suivant :

$$\Delta_{\widehat{\theta},\theta} := \delta_{\widehat{\theta},\theta}(\boldsymbol{Y}) := \frac{\widehat{\theta}(\boldsymbol{Y}) - \theta}{\widehat{\sigma_{\widehat{\theta}}}(\boldsymbol{Y})} \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1).$$

- *La probabilité comme une extension de la logique* : Nous insistons sur le fait qu'une probabilité d'un <u>événement</u> égale à **0** ou **1** signifie respectivement de manière équivalente que l'<u>événement</u> (dit **certain**) est **Faux** ou **Vrai**. C'est en ce sens que la "probabilité" est une extension de la "logique" (en tant que théorie mathématique). Un événement **incertain** a donc une probabilité strictement comprise entre 0 et 1 et exprime donc qu'il est peut-être Vrai ou peut-être Faux, la probabilité de l'événement étant d'autant plus grande (resp. petite) que l'événement a de plus en plus de chance d'être Vrai (resp. Faux). Dans le contexte statistique, un événement s'exprime à partir d'une statistique $T := t(\boldsymbol{Y})$ sous la forme $(T \in E) \Leftrightarrow (t(\boldsymbol{Y}) \in E)$ où E est un sous-ensemble de modalités de $T := t(\boldsymbol{Y})$. Ainsi, connaissant la loi de probabilité de $T := t(\boldsymbol{Y})$, nous serons en mesure d'évaluer $\mathbb{P}(T \in E) = \mathbb{P}(t(\boldsymbol{Y}) \in E)$ comprise strictement entre 0 et 1 puisque $\boldsymbol{Y}$ est intrinsèquement aléatoire. Une erreur très courante est de confondre, le **Jour J**, $\mathbb{P}(t(\boldsymbol{y}) \in E)$ avec $\mathbb{P}(t(\boldsymbol{Y}) \in E)$ alors que $\boxed{\mathbb{P}(t(\boldsymbol{y}) \in E) \in \{0,1\}} \neq \boxed{\mathbb{P}(t(\boldsymbol{Y}) \in E) \in]0,1[}$ puisque $\boldsymbol{y}$ est déterministe (i.e. strictement non aléatoire) en tant que réalisation de $\boldsymbol{Y}$.

Fin

FICHE T.D. 4

Traitement des problématiques des produits A et B

Indications préliminaires

- *Objectif* : En pratique, on peut être spécialement intéressé par une prise de décision qui dépend de la comparaison du **paramètre d'intérêt** θ inconnu par rapport à une **valeur de référence** θ_0 (fixée selon la problématique). Cette comparaison sera par la suite appelée **assertion d'intérêt**. Ne disposant que d'une estimation $\widehat{\theta}(\boldsymbol{y})$ la décision conduisant à conclure que l'assertion d'intérêt est vraie à partir de l'échantillon $\boldsymbol{y}$ du **jour J** ne peut pas être complètement fiable. L'objectif est de construire un outil d'aide à la décision nous garantissant un risque d'erreur de se tromper dans notre décision de valider l'assertion d'intérêt n'excédant pas une valeur que nous nous sommes fixée (généralement autour des 5%).
- *Paramètre d'écart standardisé* : Comparer la paramètre d'intérêt θ à une valeur de référence θ_0 est strictement équivalent à comparer leur différence ou leur rapport à 0 ou 1. Dans le cadre asymptotique, l'assertion d'intérêt pourra toujours se réécrire en fonction d'un paramètre d'écart standardisé $\delta_{\theta,\theta_0} := \frac{\boldsymbol{\theta - \theta_0}}{\sigma_{\widehat{\theta}}}$. Il est important d'apprendre à lire cette expression où le numérateur $\theta - \theta_0$ a été mis en **gras** pour souligner son rôle plus important (en termes d'information pour l'utilisateur) par rapport au dénominateur $\sigma_{\widehat{\theta}}$ ayant été introduit principalement pour des raisons techniques (mais toutefois indispensables dans la construction de l'outil d'aide à la décision). Il est alors direct de voir que :

$$\textbf{Assertion d'intérêt} \Longleftrightarrow \left\{ \begin{array}{lllll} \theta < \theta_0 & \Longleftrightarrow & \theta - \theta_0 < 0 & \Longleftrightarrow & \delta_{\theta,\theta_0} < 0 \\ \theta > \theta_0 & \Longleftrightarrow & \theta - \theta_0 > 0 & \Longleftrightarrow & \delta_{\theta,\theta_0} > 0 \\ \theta \neq \theta_0 & \Longleftrightarrow & \theta - \theta_0 \neq 0 & \Longleftrightarrow & \delta_{\theta,\theta_0} \neq 0 \end{array} \right\}$$

  En commentaire non prioritaire, on peut tout de même remarquer que l'interprétation du $\sigma_{\widehat{\theta}}$ dans l'expression de δ_{θ,θ_0} est assez naturelle : plus l'estimation de θ est fiable, se traduisant par un $\sigma_{\widehat{\theta}}$ d'autant plus faible, plus le paramètre d'écart standardisé δ_{θ,θ_0} est grand et ainsi plus facile à comparer à 0.
- *Estimation du paramètre d'écart standardisé* : Dépendant du paramètre d'intérêt θ inconnu, le paramètre d'écart standardisé δ_{θ,θ_0} est lui-même inconnu (en fait doublement inconnu puisque dépendant aussi de $\sigma_{\widehat{\theta}}$ inconnu). Il est facilement estimable à partir de l'échantillon $\boldsymbol{y}$ du **jour J**. Nous l'exprimons ci-dessous à partir du "futur" échantillon $\boldsymbol{Y}$:

$$\widehat{\delta_{\theta,\theta_0}}(\boldsymbol{Y}) := \frac{\widehat{\theta}(\boldsymbol{Y}) - \theta_0}{\widehat{\sigma_{\widehat{\theta}}}(\boldsymbol{Y})}$$

  Pour mesurer les risques d'erreur de décision, nous serons tout particulièrement intéressés par la loi de probabilité de $\widehat{\delta_{\theta,\theta_0}}(\boldsymbol{Y})$ lorsque $\theta = \theta_0$. Dans ce cas très particulier, nous remarquons que lorsque n est suffisamment grand :

$$\widehat{\delta_{\theta,\theta_0}}(\boldsymbol{Y}) := \frac{\widehat{\theta}(\boldsymbol{Y}) - \boxed{\theta_0}}{\widehat{\sigma_{\widehat{\theta}}}(\boldsymbol{Y})} = \frac{\widehat{\theta}(\boldsymbol{Y}) - \boxed{\theta}}{\widehat{\sigma_{\widehat{\theta}}}(\boldsymbol{Y})} =: \delta_{\widehat{\theta},\theta}(\boldsymbol{Y}) \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1)$$

  où $\delta_{\widehat{\theta},\theta}(\boldsymbol{Y})$ a été introduit au début de la fiche T.D. **??**.

Fin

ANNEXE

A Représentations graphiques dans l'A.E.P.

Indications préliminaires

- *Objectif* : Comme nous l'avons vu à la fiche T.D. **??**, l'**A.E.P.** s'appuie sur l'étude descriptive de m (grand) réalisations indépendantes $\boldsymbol{y}_{[m]} := \left(y_{[\cdot]}\right)_m$ d'une variable aléatoire d'intérêt Y. Afin d'alléger l'introduction de l'**A.E.P.**, cette étude descriptive a été volontairement limitée à une étude quantitative n'utilisant aucune représentation graphique issue de la Statistique Descriptive. Les graphiques étant d'une grande aide pour représenter les répartitions de séries de données, ils vont donc nous aider à mieux appréhender le comportement aléatoire des variables aléatoires d'intérêt.
- *Représentations graphiques usuelles* : Les représentations graphiques des répartitions diffèrent selon la nature des variables. Ainsi, lorsque la variable est de nature discrète (i.e. les modalités ou valeurs possibles sont dénombrables), on utilise un diagramme en bâton, et lorsqu'elle est de nature continue (i.e. à valeurs dans un continuum qui est non dénombrable), un histogramme est utilisé. Ce choix pose problème lorsqu'une étude expérimentale nous amène à comparer sur un même graphique des répartitions de plusieurs variables n'ayant pas la même nature. Il n'est pas possible de représenter une variable continue par un diagramme en bâton mais il en est tout autrement pour une variable discrète qui peut se représenter via un histogramme discret que nous allons introduire très prochainement.
- *Histogramme (continu)* : Rappelons les règles générales pour construire un histogramme représentant la répartition de la série $z_1, z_2, \cdots, z_m$:
  1. L'ensemble des modalités est découpé en une partition d'intervalles pas forcément de même largeur.
  2. Chaque intervalle de la partition est représenté par un rectangle ayant pour base l'intervalle et de surface égale à la proportion des $z_1, z_2, \cdots, z_m$ appartenant au dit intervalle.
  3. La somme de tous les rectangles est donc égale à 100%=1.

  Pour construire pratiquement un histogramme, il est conseillé au préalable de trier les $z_1, z_2, \cdots, z_m$ afin de les regrouper et ainsi de les affecter plus facilement à leurs intervalles d'appartenance. Au lieu de directement construire les rectangles associés aux intervalles de la partition, nous avons choisi d'associer à chaque donnée z_i un i$^{\text{e}}$ rectangle, que l'on appellera brique (l'histogramme étant un vu de manière imagée comme un "mur") :
  1. Toute brique associée à z_i a pour base l'intervalle d'appartenance de z_i et a une surface égale à $\frac{1}{m}$.
  2. La somme de toutes les briques associées aux $z_1, z_2, \cdots, z_m$ est donc égale à 1.
  3. Un rectangle associé à un intervalle est l'empilement de toutes les briques associés aux z_i de l'intervalle.

  Cette version de l'histogramme est équivalente à la version originale et propose en plus une représentation individualisée de toutes les données $z_1, z_2, \cdots, z_m$. Ce point particulier sera un atout pour comprendre la représentation graphique usuelle de loi de probabilités de variable aléatoire continue.
- *Histogramme discret* : Dans le même esprit, nous allons maintenant présenter la notion d'histogramme discret en l'adaptant à des données $z_1, z_2, \cdots, z_m$ à valeurs dans un espace dénombrable (donc pas dans un continuum) :
  1. Une brique associée à z_i a pour surface $\frac{1}{m}$ et sa base est centrée en z_i (à la différence d'une brique d'un histogramme continu où z_i doit appartenir à l'intervalle représentant la base de la brique).
  2. Les bases des briques sont fixées de sorte que le mur (i.e. l'histogramme) constitué de l'ensemble des briques ait le moins d'espace (trou) possible. Les briques voisines doivent se toucher dès que possible.
  3. La somme des briques est toujours égale à 100%=1 et l'empilement des briques associées aux données ayant même modalité forme un rectangle dont la surface est égale à la proportion des données égales à la modalité associée.

  IMPORTANT : Nous remarquons que dans la représentation d'histogramme (discret ou continu) les valeurs en ordonnée n'ont pas d'unité réellement interprétable. En revanche, lorsque les bases des rectangles sont toutes les mêmes, les hauteurs pourront donc être comparées entre elles pour nous éclairer sur la répartition des données puisque les aires des surfaces des rectangles (vues comme des empilements de briques) représentent naturellement des proportions de données.

Fin