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

Exercice 1 (Salaire Juste - Estimation (ponctuelle))

Une équipe de sociologues propose de réunir un comité d'experts pour la création d'un indicateur, appelé ***Salaire Juste****, mesuré pour toute personne active et qui permettra de transformer les ressources individuelles réelles (généralement mesurées par un salaire) en tenant compte de critères aussi importants que les ressources locales, le partage de ces ressources, la pénibilité du travail, le niveau d'expérience, d'expertise et bien d'autres encore... Cet indicateur est conçu de sorte qu'en théorie il devrait être équivalent (en fait égal à une valeur étalon 100) pour toute personne active dans le monde. En conséquence directe, le Salaire Juste moyen dans le monde devrait être égal à 100. Après quelques mois de travail, un premier prototype (très perfectible) du* ***Salaire Juste*** *est élaboré par la fine équipe d'experts. Les sociologues s'accordent à dire qu'un pays peut se dire non civilisé s'il vérifie aux 2 critères de discriminations suivants :*

> ***Discrimination Mondiale*** *: le Salaire Juste moyen dans le pays est très supérieur à la valeur 100 de base. Un Salaire Juste moyen excédant un seuil de 150 est considéré comme intolérable.*
>
> ***Discrimination Intérieure*** *: les Salaires Justes dans le pays sont très dispersés. La variance des Salaires Justes dans le pays supérieur à 30 est considérée comme excessive et donc anormale.*

Par la suite, $\mathcal{Y}_i$ $(i = 1, \cdots, N)$ désigne le Salaire Juste de la $i^{\text{ème}}$ personne actives du pays.

1. *Définir mathématiquement les paramètres (d'intérêt), notés μ^J et σ_J^2, permettant éventuellement d'établir des discriminations mondiale et intérieure. Quelle est la nature de ces paramètres ?*
2. *Soit Y^J la variable aléatoire (v.a.) correspondant au Salaire Juste d'un individu choisi au hasard dans la population des N personnes actives du pays. Etablir la relation entre les paramètres μ^J et σ_J^2 et la v.a. Y^J*
3. *Rappeler alors les estimateurs proposés par les mathématiciens obtenus à partir d'un "futur" échantillon $\boldsymbol{Y^J}$ (en utilisant la Norme CQLS).*
4. *Quelles sont les "bonnes" propriétés de ces estimateurs désirées par les mathématiciens ? Interrogez-vous sur comment les interpréter via l'A.M.P. ?*
5. *Proposer à présent leur interprétation via l'A.E.P. en prenant soin au préalable d'introduire les notations nécessaires (Norme CQLS). Proposer alors une description littérale pour chacune de ces "bonnes" propriétés.*
6. *Une étude est menée par un expérimentateur. Il se fixe l'ensemble des Salaires Justes sur un pays fictif de $N = 1000000$ personnes actives dont il est le seul à en connaître les valeurs. Voici les résultats présentés dans les tableaux ci-dessous :*

$\boldsymbol{Y}$	$\widehat{\mu^J}(\boldsymbol{Y})$	$\widehat{\sigma_J}(\boldsymbol{Y})$	$\widehat{\sigma_{\widehat{\mu^J}}}(\boldsymbol{Y})$	$\widehat{\mu^J}(\boldsymbol{Y}) - \mu^J$	$\delta_{\widehat{\mu}^J,\mu^J}(\boldsymbol{Y})$
$\boldsymbol{y}_{[1]}$	99.91	10.0231	0.317	-0.09	-0.29
$\boldsymbol{y}_{[2]}$	99.65	9.2615	0.2929	-0.35	-1.19
$\boldsymbol{y}_{[3]}$	100.84	10.448	0.3304	0.84	2.54
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{y}_{[9998]}$	99.58	9.3785	0.2966	-0.42	-1.43
$\boldsymbol{y}_{[9999]}$	100.2	9.9372	0.3142	0.2	0.63
$\boldsymbol{y}_{[10000]}$	99.94	9.7991	0.3099	-0.06	-0.21
Moyenne	100.0043	10.034	0.3173	0.0043	-0.0294
Ecart-type	0.3178	0.63	0.0199	0.3178	1.0058

7. *En notant $\boldsymbol{yy}$ un échantillon correspondant à une ligne du tableau ci-dessous (par exemple, la $3^{\text{ème}}$), fournir les instructions `R` qui a permis à l'expérimentateur d'obtenir les valeurs du tableau précédent (Indication : étant ici à la place de l'expérimentateur, n'oubliez pas que vous disposez exceptionnellement les valeurs de μ^J et σ_J^2).*
8. *Proposer les notations mathématiques correspondant aux 2 dernières lignes du tableau qui, nous l'espérons, permet de comprendre à quoi elles correspondent et comment elles ont été obtenues.*

9. *Quelles valeurs du tableau sont sensées mesurer (approximativement) les qualités de l'estimateur* $\widehat{\mu^J}\left(\boldsymbol{Y^J}\right)$ *? Comment les noter dans l'A.M.P. ? Sont-elles accessibles le jour J ? Mêmes questions pour l'estimateur* $\widehat{\sigma_J^2}\left(\boldsymbol{Y^J}\right)$.
10. *Comment obtient-on les estimations des qualités mesurées par les écarts-type des estimateurs* $\widehat{\mu^J}\left(\boldsymbol{Y^J}\right)$ *et* $\widehat{\sigma_J^2}\left(\boldsymbol{Y^J}\right)$. *Comment sont-elles appelées ?*
11. *A partir de maintenant, on s'imagine être le jour J. Pour cela, on suppose ne disposer que du* $3^{\text{ème}}$ *échantillon dans le tableau ci-dessus. Comment doit-on noter ce jeu de données. Proposer à partir du tableau toutes les estimations intéressantes relativement aux problèmes de discriminations mondiale et intérieure. N'en manque-t-il pas une ou plusieurs ? Retrouvez-les ou complétez-les à partir de la sortie* `R` *suivante :*

```
> length(yy)
[1] 1000
> mean(yy)
[1] 100.8388
> sd(yy)
[1] 10.44798
> var(yy)
[1] 109.1603
> seMean(yy)
[1] 0.3303941
> sd(yy)/sqrt(length(yy))
[1] 0.3303941
> seVar(yy)
[1] 9.475496
```

12. *Voici les sorties* `R`*, correspondant aux mêmes informations mais sur l'échantillon des n=100 premiers individus :*

```
> mean(yy[1:100])
[1] 101.7301
> sd(yy[1:100])
[1] 12.13053
> var(yy[1:100])
[1] 147.1498
> seMean(yy[1:100])
[1] 1.213053
> sd(yy)/sqrt(100)
[1] 1.044798
> seVar(yy[1:100])
[1] 32.54073
```

Comparer ces résultats à ceux obtenus à partir de l'échantillon initial de taille n=1000. Quelle type d'estimation vaut-il mieux préconiser lorsqu'on désire intégrer l'erreur standard ?

Exercice 2 (Salaire Juste - Estimation par intervalle de confiance)

1. *A partir de votre formulaire, rappeler les expressions des "futurs" intervalles de confiance à* 95% *(généralement noté* $1-\alpha$*) de niveau de confiance pour les paramètres* μ^J *et* σ_J^2. *Rappeler à partir de quel résultat mathématique (probabiliste) ont-ils été construits ? Evaluer la probabilité* $\mathbb{P}\left(|\delta_{\widehat{\theta},\theta}(\boldsymbol{Y^J})| \leq 1.96\right) = \mathbb{P}\left(-1.96 \leq \delta_{\widehat{\theta},\theta}(\boldsymbol{Y^J}) \leq 1.96\right)$ *où* θ *désigne indifféremment* μ^J *et* σ_J^2. *L'interpréter via l'A.E.P. notamment avec le tableau précédent.*
2. *Question optionnelle (pour ceux qui ne sont pas rebutés par de simples calculs mathématiques) : Construire mathématiquement les futurs intervalles de confiance ci-dessus.*
3. *Fournir l'instruction* `R` *permettant de les obtenir le jour J (Indication : en* `R`*,* `qnorm(.975)` $\simeq 1.96$*) et le calculer éventuellement en utilisant votre machine à calculer. Déduire un intervalle de confiance à* 95% *pour* σ_J.
4. *Voici sur les résultats expérimentaux pour les intervalles de confiance* $IC_{\mu^J}(\boldsymbol{Y^J})$ *et* $IC_{\sigma_J^2}(\boldsymbol{Y^J})$ *de* μ^J *et* σ_J^2. *Interpréter via l'approche expérimentale*

$\boldsymbol{Y}$	$IC_{\mu^J}(\boldsymbol{Y^J})$	$\mu^J \in IC_{\mu^J}(\boldsymbol{Y^J})$	$IC_{\sigma_J^2}(\boldsymbol{Y^J})$	$\sigma_J^2 \in IC_{\sigma_J^2}(\boldsymbol{Y^J})$
$\boldsymbol{y_{[1]}}$	$[99.29, 100.53]$	1	$[61.61, 139.31]$	1
$\boldsymbol{y_{[2]}}$	$[99.08, 100.23]$	1	$[71.6, 99.95]$	0
$\boldsymbol{y_{[3]}}$	$[100.19, 101.49]$	0	$[90.59, 127.73]$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{y_{[9998]}}$	$[99, 100.16]$	1	$[72.18, 103.74]$	1
$\boldsymbol{y_{[9999]}}$	$[99.58, 100.81]$	1	$[82.57, 114.93]$	1
$\boldsymbol{y_{[10000]}}$	$[99.33, 100.54]$	1	$[77.91, 114.13]$	1
Moyenne		94.86%		92.02%

5. *Evaluer les probabilités suivantes :*
$\mathbb{P}\left(\mu^J \in IC_{\mu^J}(\boldsymbol{y^J})\right) = \mathbb{P}\left(\mu^J \in [100.19, 101.49]\right)$ *et* $\mathbb{P}\left(\sigma_J^2 \in IC_{\sigma_J^2}(\boldsymbol{y^J})\right) = \mathbb{P}\left(\sigma_J^2 \in [90.59, 127.73]\right)$

Exercice 3 *(taille étudiants)*

*Pour mettre en pratique ce qu'il a appris dans son cours de Statistique Inférentielle, un étudiant souhaite utiliser l'**Approche Expérimentale** pour comprendre la notion d'intervalle de confiance. Son but est d'estimer par intervalle de confiance **la taille moyenne**, notée μ, des $N = 300$ étudiants de sa promotion.*

1) Il construit un premier échantillon (avec remise) de taille $n = 30$ (i.e. pour se placer dans le cadre asymptotique), qu'il note $\boldsymbol{y_{[1]}}$, dans la population des $N = 300$ étudiants de sa promotion :

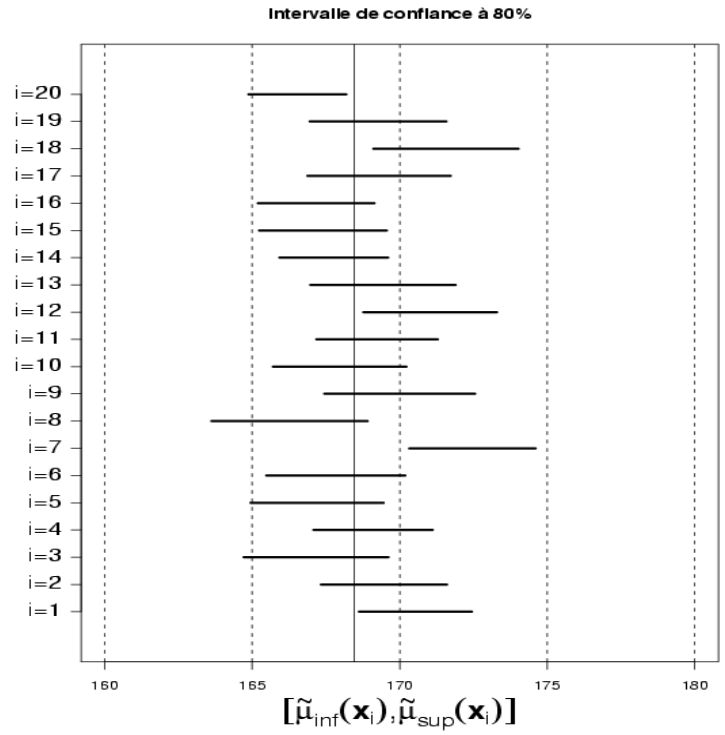
```
> y1
 [1] 165 179 171 178 171 168 166 171 182 178 177 165 174 164 175 178 167 168 185
[20] 166 162 180 167 174 159 159 184 154 172 157
```

Proposez l'instruction `R` ayant permis d'obtenir le résultat ci-dessous correspondant à un intervalle de confiance au niveau de confiance de 80% de μ :

Indication(s) R :

```
> # IC <- (instruction R à fournir dans la rédaction)
> IC
[1] 168.6308 172.4359
```

2) Ne sachant pas comment interpréter ce résultat, il construit 19 autres échantillons de taille $n = 30$ dans la population des étudiants de sa promotion que l'on notera respectivement $\boldsymbol{y_{[2]}}, \ldots, \boldsymbol{y_{[20]}}$. Il représente alors sur un même graphique ces 20 différents intervalles de confiance de μ à 80% de niveau de confiance :



Afin de confronter ses résultats expérimentaux avec la réalité, l'étudiant décide d'interroger tous les étudiants de sa promotion (notez que ceci est possible car $N = 300$). Il peut alors calculer la valeur de μ, à savoir 168.45. *Elle est représentée par le trait vertical (en trait plein) sur le graphique précédent. Sur les 20 intervalles de confiance calculés, combien contiennent μ ? Est-ce surprenant ?*

3) Que se passerait-il si l'étudiant construisait une infinité d'intervalles de confiance de μ à 80% *de niveau de confiance sur des échantillons de taille $n = 30$?*

Exercice 4 *(élection présidentielle) Entre les deux tours d'une élection présidentielle, on souhaite estimer par intervalle de confiance à 95% les proportions d'intentions de vote des deux candidats finalistes. Avant même d'effectuer un sondage sur une sous-population de taille $n = 1000$, quelle serait la plus grande longueur des deux intervalles de confiance (en utilisant la formule approchée) ?*

Indication(s) R :

```
> 2*qnorm(0.975)*sqrt(0.5*0.5/1000)
[1] 0.0619795
```

Quelle doit être la taille de l'échantillon n pour être certain que la longueur de l'intervalle de confiance au niveau 95% *n'excède pas* 0.04% *?*

Indication(s) R :

```
> (2*qnorm(.975)*sqrt(.5^2)/.0004)^2
[1] 24009118
> (qnorm(.975)/.0004)^2
[1] 24009118
```