FICHE T.D. 1 Présentation des différents outils

Exercice 1 *Entre les deux tours d'une élection présidentielle, un candidat, Max, souhaiterait "rapidement" avoir un a priori sur la proportion d'intentions de vote en sa faveur. On notera* $\boldsymbol{\mathcal{Y}}^{Max} = \left(\mathcal{Y}_1^{Max}, \ldots, \mathcal{Y}_N^{Max}\right)$ *l'ensemble des réponses des* N *électeurs (où* $\mathcal{Y}_i^{Max}$ *vaut 1 si l'individu* i *a l'intention de voter pour Max et 0 sinon).*

1. *Déterminez en fonction de* $\boldsymbol{\mathcal{Y}}^{Max}$*, le nombre puis la proportion d'intentions de vote en faveur de Max, notée respectivement* N^{Max} *et* p^{Max}*.*

   Réponse

   $p^{Max} := \frac{1}{N} \sum_{i=1}^{N} \mathcal{Y}_i^{Max}$.

   Fin

2. N *étant très grand, quel serait une solution réalisable permettant d'obtenir un remplaçant (i.e. estimation) de* p^{Max}*. Proposez les notations adéquates.*

   Réponse

   Une solution consiste à interroger un nombre $n << N$ *d'individus. Ce jeu de données pourrait être noté* $\boldsymbol{y^{Max}}$ *et le remplaçant de* p^{Max} *pourrait alors être calculé par* $\boldsymbol{E}p^{Max}y^{Max} := \frac{1}{n} \sum_{i=1}^{n} y_i^{Max}$.

   Fin

3. *Deux personnes se proposent d'interroger chacun* $n = 1000$ *électeurs. On notera* $\boldsymbol{y}_{[1]}$ *et* $\boldsymbol{y}_{[2]}$ *ces deux jeux de données recueillis. Les estimations correspondantes sont respectivement de 47% et 52%. Comment interpréter la différence des résultats qui, si on leur fait une confiance aveugle, conduit à deux conclusions différentes ?*

   Réponse

   La différence des résultats peut s'expliquer par le fait qu'un échantillon de taille n *ne constitue qu'une sous-information de la population de taille* N*.*

   Fin

4. *Connaissez-vous d'autres applications nécessitant une estimation d'un paramètre inconnu ?*

   Réponse

   normes de production, normes écologiques,...

   Fin

Exercice 2 (Présentation des problématiques des produits A et B)

1. *Exprimez N^A (resp. N^B) en fonction des $\boldsymbol{\mathcal{Y}}^A$ (resp. $\boldsymbol{\mathcal{Y}}^B$). Exprimez la rentabilité du Produit A (resp. Produit B) en fonction du nombre total N^A (resp. N^B) d'exemplaires du Produit A (resp. Produit B) vendus.*

   Réponse

   Le produit $\bullet$ est rentable si $N^\bullet := \sum_{i=1}^N \mathcal{Y}_i^\bullet > 300000$.

   Fin

2. *Même question mais en fonction du nombre moyen (par individu de la population) μ^A (resp. μ^B) d'exemplaires du Produit A (resp. Produit B) en ayant au préalable établi la relation entre μ^A et N^A (resp. μ^B et N^B) et ainsi entre μ^A et $\boldsymbol{\mathcal{Y}}^A$ (resp. μ^B et $\boldsymbol{\mathcal{Y}}^B$). Quelle relation y a-t-il donc entre μ^A et $\overline{\mathcal{Y}^A}$ (resp. entre μ^B et $\overline{\mathcal{Y}^B}$) ?*

   *Les quantités μ^A et μ^B seront appelées **paramètres d'intérêt**.*

   Réponse

   Le produit $\bullet$ est rentable si $\mu^\bullet := \frac{1}{N}\sum_{i=1}^N \mathcal{Y}_i^\bullet = \overline{\mathcal{Y}^\bullet} > \frac{300000}{2000000} = 0.15$.

   Fin

3. *Est-il possible pour l'industriel de ne pas se tromper dans sa décision quant au lancement de chaque produit ? Si oui, comment doit-il procéder ? Cette solution est-elle réalisable ?*

   Réponse

   Pour ne pas se tromper, il lui faut recueillir les intentions des N individus ce qui paraît peu réalisable.

   Fin

4. *Est-il alors possible d'évaluer (exactement) les paramètres d'intérêt ? Comment les qualifieriez-vous par la suite ?*

   Réponse

   Les paramètres d'intérêt ne peuvent donc pas être évalués et sont donc considérés comme INCONNUS.

   Fin

5. *Une solution réalisable est alors de n'interroger qu'une sous-population de taille raisonnable $n << N$ (ex $n = 1000$). On notera alors $\boldsymbol{y}^\bullet$ le jeu de données (appelé aussi échantillon), i.e. le vecteur des n nombres d'achat $(y_i^\bullet)_{i=1,\cdots,n}$ du produit $\bullet$ des n ($n << N$) individus interrogés.*

   Comment l'industriel pourra-t-il évaluer un remplaçant de $\mu^\bullet$ à partir de son échantillon $\boldsymbol{y}^\bullet$?

   (quelle est la relation entre $\overline{y^\bullet}$, représentant la moyenne empirique des $(y_i^\bullet)_{i=1,\ldots,n}$, et l'estimation $\widehat{\mu^\bullet}(\boldsymbol{y}^\bullet)$?)

   Réponse

   En évaluant la moyenne sur l'échantillon observé, i.e. en calculant $\widehat{\mu^\bullet}(\boldsymbol{y}^\bullet) = \frac{1}{n}\sum_{i=1}^n y_i^\bullet = \overline{y^\bullet}$.

   Fin

6. *Quelle est la nature du paramètre d'intérêt μ^A dans le cas où les données ne sont que des 0 et 1 ? Désormais cette moyenne, puisqu'elle bénéficiera d'un traitement particulier, sera notée $\boxed{p^A = \mu^A}$.*

   Réponse

   Une moyenne de 0 et de 1 correspond à une proportion.

   Fin

Exercice 3 *Dans le but d'estimer un paramètre d'intérêt inconnu, on dispose d'un échantillon. Nous nous proposons maintenant de préciser plus en détail son procédé de construction.*

1. *Proposez des critères de qualité d'un tel échantillon.*
2. *A quoi correspond la notion de représentativité ?*

   Réponse

   à essayer de faire "ressembler" l'échantillon à la population totale.

   Fin

3. *Est-il possible de construire un échantillon représentatif d'une (ou plusieurs) caractéristique(s) donnée(s) ?*

   Réponse

   oui par exemple en tentant de respecter la proportion de femmes dans la population totale avec la proportion de femmes présentes dans l'échantillon.

   Fin

4. *Même question sans aucun a priori (i.e. aucune caractéristique fixée).*
5. *Proposez un critère de qualité qui permettra de construire un échantillon le plus représentatif sans aucun a priori.*

   Réponse

   voir réponse ci-après.

   Fin

6. *Fournissez un (ou plusieurs) procédé(s) d'échantillonnage satisfaisant au critère suivant de représentativité (maximale) sans a priori (RSAP) :*

   Tous les individus de la population totale ont la même chance d'être choisi dans l'échantillon.

   Réponse

   Selon ce critère, on pourrait choisir n individus au hasard au sein de la population avec remise et sans remise. Notons qu'étant donné les ordres de grandeurs, $n = 1000$ et $N = 2000000$ ces deux procédés sont quasiment équivalents.

   Fin

7. *Si on répète le procédé d'échantillonnage suivant le critère RSAP et que pour chaque échantillon on évalue l'estimation du paramètre d'intérêt, pensez-vous que les résultats seront toujours les mêmes ? Comment qualifie-t-on alors la nature du procédé d'échantillonnage ?*

   Réponse

   L'échantillonnage est dit aléatoire.

   Fin

Exercice 4 (Outil pour la problématique des élections)

1. *Pensez-vous qu'il soit possible qu'une estimation $\widehat{p}(\boldsymbol{y})$ soit égale au paramètre estimé ? Pouvez-vous savoir l'ordre de grandeur de l'écart entre l'estimation et le paramètre inconnu ? Quel niveau de confiance accordez-vous à la valeur d'une estimation (dans notre exemple, 47% et 52% sur deux échantillons) ?*

   Réponse

   Excepté dans de très rares contextes, une estimation ne peut pas correspondre à la vraie valeur du paramètre inconnu. Il n'y a aucun moyen de mesurer avec certitude l'écart entre l'estimation et le paramètre. Cependant, on peut seulement espérer qu'ils ne sont pas très éloignés. Compte tenu ce ces réponses, il est alors difficile de répondre à la dernière question autrement que de proposer un avis personnel plutôt arbitraire.

   Fin

2. *Si on vous annonce qu'un statisticien sait généralement fournir en plus de l'estimation du paramètre, l'estimation de sa fiabilité mesurée en terme de variabilité attendue, quel est la*

mission principale d'un intervalle de confiance ? Quelles sont les qualités souhaitées d'un intervalle de bonne confiance (95% par exemple) du paramètre d'intérêt (inconnu) ?

Réponse

L'objectif est d'intégrer dans le procédé d'estimation du paramètre sa fiabilité (voir énoncé) afin de fournir un intervalle plus ou moins large selon le niveau de confiance fixé. La qualité attendue est que cet intervalle ait de bonnes chances (traduites par le niveau de confiance) de contenir le paramètre d'intérêt inconnu. En outre, on peut espérer obtenir un intervalle de longueur raisonnablement faible pour que l'estimation soit suffisamment informative (bien que l'on ne peut en être assuré en général).

Fin

3. *Compléter les phrases suivantes :*
   (a) *PLUS le niveau de confiance est fort, PLUS l'intervalle de confiance est petit.*
   (b) *Vue comme un intervalle de confiance de largeur 0, une estimation peut donc être associé à un niveau de confiance 0%.*
4. *Un statisticien construit les intervalles à* 95% *de confiance (via une formule d'obtention étudiée plus tard dans le cours ne faisant pas l'objet) et informe le candidat que les intervalles associés aux estimations 47% et 52% sont respectivement* $[43.90655\%, 50.09345\%]$ *et* $[48.90345\%, 55.09655\%]$. *Les élections effectuées, on évalue* $p^{Max} = 51.69\%$, *qu'en pensez-vous ?*

   Réponse

   Il semble qu'il soit difficile d'affirmer que le candidat sera élu.

   Fin

5. *Si vous avez des difficultés à traduire ce que signifie le niveau de confiance d'un intervalle, comparez-le avec celui que vous accorderiez à une personne qui serait censée dire la vérité avec un niveau de confiance fixé à 95%. Dans le cas de cette personne, comment traduiriez-vous (ou expliqueriez-vous) le concept de niveau de confiance ?*

   Réponse

   Parmi toutes les assertions énoncées par cette personne (dont on peut vérifier la véracité ou fiabilité), 95% (en moyenne) seraient censées être justes ou fiables.

   Fin