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

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

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

Un industriel veut lancer sur le marché deux produits que l'on nommera Produit A et Produit B . Le Produit A est acheté au plus une fois par mois tandis que le Produit B peut être acheté autant de fois que désiré. Après une étude financière, les services comptables indiquent à cet industriel que pour que le lancement de chacun de ces produits soit rentable, il faut qu'il soit vendu à plus de 300000 exemplaires par mois. La population ciblée par l'industriel est une population de taille $N = 2000000$. L'industriel se demande s'il doit ou non lancer le(s) Produit(s) A et/ou B.

Commençons par introduire quelques notations permettant de décrire le choix d'achat des individus de la population totale (ciblée par l'industriel). Les deux études des Produit A et Produit B étant plutôt similaires, nous noterons donc dans un cadre général • aussi bien à la place de A ou B. Ainsi $\mathcal{Y}_i^\bullet$ représente le nombre de produit(s) • acheté(s) par le $i^{ème}$ $(i = 1, \cdots, N)$ individu de la population totale. L'ensemble des choix d'achat des N individus $(\mathcal{Y}_i)_{i=1,\cdots,N}$ sera noté $\boldsymbol{\mathcal{Y}}^\bullet$. $N^\bullet$ désignera le nombre d'exemplaires de Produit • achetés par les N individus de la population.

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.*
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**.*
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 ?*
4. *Est-il alors possible d'évaluer (exactement) les paramètres d'intérêt ? Comment les qualifieriez-vous par la suite ?*
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 • 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 $\widehat{\mu^\bullet}(\boldsymbol{y}^\bullet)$?)
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 $\boxed{p^A = \mu^A}$.*

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

Exercice 4 (Outil pour la problématique des élections)

On se propose d'estimer le paramètre d'intérêt en fournissant un intervalle (ou fourchette, encadrement) obtenu à partir des données. Cet intervalle, appelé intervalle de confiance, est centré en la valeur de l'estimation et sa largeur dépend d'un niveau de confiance que l'on se fixe (généralement plutôt grand, par exemple, 95%).

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) ?*
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) ?*
3. *Compléter les phrases suivantes :*
   (a) *PLUS le niveau de confiance est fort, ... 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 ...%.*
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 ?*
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

Remarque

Cet exemple nous aide à appréhender la notion de niveau de confiance ou plus généralement de probabilité d'un événement en l'exprimant comme la proportion parmi toutes (en théorie, on peut imaginer en faire une infinité) réalisations de l'expérience (a priori supposée aléatoire) qui conduisent à ce que l'événement soit vérifié. Ceci nous conduit naturellement vers la notion d'Approche Expérimentale des Probabilités qui sera présentée dans la fiche de Td suivante en complément de l'Approche Mathématique des Probabilités (qui est classiquement présentée dans les cours de Probabilités et Statistique).