Séance 5 : Les statistiques inférentielles

L'échantillonnage sert à réaliser une étude statistique puisqu'on ne peut pas étudier l'ensemble de la population, cette dernière étant trop grande donc impossible à traiter entièrement. Un échantillon correspond donc à un groupe restreint de la population mère. De cet échantillon pris dans la population, on réalise ensuite la démarche de l'inférence, c'est-à-dire qu'on part du petit groupe étudié pour déduire informations sur le plus grand groupe, la population mère. Cela permet d'éviter d'utiliser la population en entier, ce qui serait impossible tant par la masse de données à traiter dans certains cas (pour la population d'un pays par exemple). Les échantillons rentrent ensuite dans deux catégories : les échantillons non biaisés, c'est-à-dire pris complètement au hasard où tous les individus ont la même possibilité de se retrouver dans l'échantillons (ils sont alors équiprobables). L'échantillons biaisé consiste au contraire sélectionner des individus sans intervention du hasard. L'échantillon doit pouvoir être représentatif, c'est-à-dire qu'on doit pouvoir généraliser les informations déduites à l'ensemble de la population mère. Les échantillons sont indépendants quand ils sont constitués par des individus différents. Les échantillons sont appariés quand ils sont constitués des mêmes individus tirés au sort. Il existe deux méthodes d'échantillonnage : les méthodes aléatoires (avec tirage au sort) et les méthodes non aléatoires. Les méthodes aléatoires s'appliquent par exemple pour un SAS (sondage aléatoire simple). C'est un tirage au sort dans la population mère. Le tirage peut se faire avec remise (possibilité qu'un numéro soit tiré deux fois) ou sans remise (le numéro ne peut être tiré qu'une seule fois). Les méthodes non aléatoires utilisent aussi un tirage au sort mais d'une autre façon. L'échantillon systématique suppose une base de sondage où les individus sont numérotés. On choisit ensuite dans ces numéros selon un intervalle aléatoirement défini (tous les cinq individus par exemple). La méthode des quotas consiste à prendre en compte des variables de contrôle à partir de caractéristiques connues de la population parente (la CSP, le sexe, etc.). Le choix de ces différentes méthodes dépend de ce qu'on souhaite analyser et aussi de notre sélection de l'échantillon. Si on veut une représentation précise de N agriculteurs et N ouvriers, alors on fera un échantillonnage par quotas sur critères de la CSP. Si on mène des sondages dans la rue, il est préférable d'utiliser la méthode d'échantillonnage systématique on l'on interroge une personne sur cinq selon un décompte effectué à chaque piéton.

Un estimateur est une fonction de donnée qui s'organise de manière à être le plus proche possible de la vraie valeur du paramètre. C'est une fonction des données qui a pour but de trouver le meilleur estimateur, qui donne une estimation ponctuelle la plus proche du paramètre. C'est le processus de rechercher le meilleur estimateur possible qui s'appelle une estimation. L'estimation n'est jamais certaine, elle est seulement plus probable que d'autres estimateurs. Pour obtenir une estimation fiable, on extrait un échantillon de la population statistique qu'on étudie et on applique les estimateurs sur cet échantillon. Les estimations comportent nécessairement des erreurs, même minimes mais qui ne peuvent être ignorées donc on ne peut pas en faire une loi absolue. Une erreur est appelée un biais qui correspond à la différence entre l'espérance de l'estimateur et la valeur à estimer dans la population. On dit que l'estimateur est consistant si sa distribution se concentre dans une zone de plus en plus étroite alors que la taille de l'échantillon tend vers l'infini. On dit d'un estimateur qu'il est robuste quand il est peu sensible aux données aberrantes. Les données aberrantes sont des erreurs aléatoires , qui peuvent venir de fausses valeurs, de données mal transmises ou mal copiées, etc. on considère que les données de bonnes qualités contiennent entre 10% à 20% de valeurs aberrantes.

L'estimation par intervalle de confiance doit avoir le plus de possibilité de contenir la vraie valeur du paramètre. Il y a en revanche toujours nécessairement un risque d'erreur. L'intervalle de confiance sert à estimer un paramètre inconnu de la population statistique à partir d'un échantillon. Souvent la

proportion de la population est inconnue, on observe un échantillon et on calcule cette proportion et on cherche un intervalle plausible pour la vraie valeur du paramètre. Un intervalle de confiance à 95% est un intervalle qui a 95% de chances de contenir la vraie valeur du paramètre inconnu. L'estimation par intervalle de fluctuation sert en revanche à annoncer ce qu'on s'attend à observer dans un échantillon si une hypothèse sur la population est vraie. On connaît donc la proportion de population, on prélève un échantillon d'une certaine taille et on s'intéresse à la proportion observée. Le but est de tester la cohérence d'un échantillon avec une hypothèse. Pour récapituler, l'intervalle de fluctuation s'applique avec le paramètre connu, donc la statistique fluctue alors que pour l'intervalle de confiance la statistique est connue mais le paramètre est inconnu.

Comme expliqué plus haut, dans la théorie de l'estimation, un biais mesure le décalage systématique entre un estimateur et la vraie valeur du paramètre que l'on cherche à estimer.

Une statistique traitant la population totale est appelée paramètre (de population). La quantité définie correspond à la population entière. En contexte de statistique classique, les frontières sont floues entre paramètre et statistique (sur un échantillon) puisque les populations sont trop grande donc on travaille souvent à partir d'un échantillon. Seulement avec les *Big data*, le traitement des informations change : on peut souvent travailler sur la quasi-totalité de la population, comme c'est le cas par exemple si on travaille sur les utilisateurs d'une plateforme en ligne. Les quantités calculés uniquement dans ces cas là sont alors des paramètres empiriques et plus des estimations. L'estimation n'intervient que si on se sert d'un échantillon et que ensuite on en tire un paramètre général par déduction. Néanmoins il ne faut pas non plus prendre un paramètre (de population) comme une science absolue puisque avec les *Big data* l'évolution temporelle de la base de données est souvent rapide, on se sert alors davantage du paramètre pour faire des prédictions statistiques ou probabilistes.

Les enjeux autour du choix d'un estimateur sont multiples puisque l'estimateur modifie la qualité, la fiabilité et l'utilité pratique des conclusions tirées des données. Les enjeux sont donc théoriques et opérationnels. Comme vu plus haut, pour que l'estimateur soit efficace il doit être consistant, notamment quand les données sont nombreuses pour réduire le champ des erreurs et converger au mieux vers la vraie valeur. Il doit être performant, donc être comparé par l'erreur quadratique moyenne. L'estimateur doit aussi faire face aux données aberrantes, donc il doit être robuste.

Il existe plusieurs méthodes d'estimation d'un paramètre. On peut d'abord citer la méthode des moindres carrés qui fonctionne notamment quand plusieurs valeurs aléatoires sont étudiées. Cette méthode sert lorsque les quantités à estimer sont des espérances. Ensuite il y a la méthode du maximum de vraisemblance (MV). C'est une approche générale de l'estimation de paramètres inconnus en s'appuyant sur des données. L'objectif est de trier les différentes valeurs du paramètre selon leur probabilité. La sélection d'une méthode d'estimation d'un paramètre dépend du type de données que nous avons à étudier et de l'objectif de l'étude. Par exemple si on souhaite faire des probabilités, il faut plutôt utiliser la méthode MV.

Les tests statistiques permettent d'établir un jugement sur échantillon. On peut ensuite savoir si un événement a ou a eu un impact sur les observations considérées. Le test sert à comprendre s'il existe une relation de cause à effet entre les différents échantillons observés. Il y a deux types de tests : les tests paramétriques (moyenne, écart type…) et les tests non paramétriques (effectif, médiane, …). Le test *t* de Student (paramétrique) par exemple permet d'évaluer l'impact d'une variable qualitative sur une variable quantitative. Le test du Chi2 (non-paramétrique) évalue les relations entre deux variables qualitatives et sert à vérifier l'indépendance de ces deux variables étudiées. Un test statistique sert donc à vérifier si une série statistique d'observations est compatible avec une loi de probabilité entièrement spécifique. Pour créer un test il faut définir à quelle condition l'une ou l'autre des

hypothèses sera considérées comme vraisemblable avec une hypothèse nulle *H0* et une hypothèse alternative *H1*. Il existe plusieurs tests. Le test de conformité permet de comparer un échantillon à une référence théorique. Le test d'homogénéité compare plusieurs échantillons entre eux (comparaisons de moyennes, d'écarts types, de variances…). Le test d'adéquation à une loi de probabilité montre qu'une distribution étudiée suit vraisemblablement une loi de probabilité donnée : en d'autres termes il s'agit de vérifier si la distribution de l'échantillon est compatible avec celle de la population mère. Le test d'indépendance de deux caractères compare deux caractères en supposant qu'ils sont indépendants. On notera que ce test appliqué à des valeurs qualitatives s'appelle le test du Chi2. Pour les tests paramétriques, on suppose normalement la forme des distributions testées puisqu'elle est connue *a priori* tandis que pour les tests non paramétriques, la forme des distributions n'est pas prise en compte. Les tests robustes fonctionnent peu importe la forme de la loi de la variable aléatoire (test libre). Le test unilatéral répond aux affirmations « plus que », « moins que », « pire que »… alors que le test bilatéral s'utiliser plutôt pour répondre à « différent de », « non égal à », … Le test d'hypothèse sert à vérifier si les données de l'échantillon collectées sont compatibles avec une hypothèse effectuée sur la population mère. L'hypothèse sera ensuite réfutée ou acceptée. Les tests paramétriques s'appliquent sur la moyenne, l'écart type, la variance, sur les paramètres d'une série statistique. Le test d'ajustement permet de juger l'adéquation entre une situation réelle et un modèle théorique. Le Chi2 est un exemple d'un test d'ajustement qui utilise la loi multinomiale.

Certaines critiques des statistiques inférentielles sont justes, notamment celle qui mentionne bien qu'avec un échantillon trop petit on ne peut pas obtenir un test significatif. Effectivement il faut un échantillon suffisamment raisonnable pour pouvoir appliquer des déductions à l'ensemble de la population, sinon l'échantillon est trop petit et pas représentatif. Ensuite les hypothèses restent des hypothèses, il ne faut pas les surinterpréter mais on ne peut pas non plus les réfuter dès le début sans les avoir testées avant. A l'état d'hypothèse, on ne peut pas tirer de conclusion actives sur l'ensemble de la population statistique, ce qu'on observe est valable seulement pour l'échantillon étudié au début, et ne peut pas être appliqué tout de suite à la population. Attention en revanche, ce n'est pas parce qu'un échantillon est grand qu'on doit et qu'on peut imposer les paramètres à l'ensemble de la population : un échantillon restera un échantillon et ne sera jamais égal à la population mère. Enfin les tests statistiques restent des tests. Les scientifiques peuvent manipuler leurs données comme ils le veulent, nous ne pouvons pas et ne devons pas être aliénés par les tests comme une science absolue.

Commentaires code Python

L'intervalle de fluctuation s'avère comprendre les fréquences dans les trois cas ce qui nous permet donc de préciser que le résultat est compatible avec l'hypothèse dans le cas des POUR, CONTRE et SANS OPINION. La confiance en l'hypothèse est donc de 95% pour la population mère. L'échantillon est bien issu de cette population. Les quelques écarts éventuelles peuvent s'expliquer par le fait du hasard de l'échantillonnage.

Pour l'intervalle de confiance, les fréquences sont aussi proches de l'intervalle réelle de la population mère. On constate sur les cinq premiers échantillons les fréquences varient légèrement d'un échantillon à l'autre. Mais ces variations restent contenues dans l'intervalle de confiance à 95%. Donc environ 95% des intervalles de confiances contiennent la vraie proportion. Ici l'intervalle de confiance contient la proportion, les intervalles de fluctuation confirment la cohérence.

La loi normale est celle du fichier Test 1 puisque sa valeur est supérieure à 0,05 ce qui n'est pas le cas pour la loi du Test 2 qui est inférieure à 0,05 donc qui ne suit pas une loi normale.