Séance 5
Échantillonnage, estimation et décision statistique

Honoka OYAMADA

Introduction

Cette séance vise à comprendre comment produire des conclusions robustes à partir de données partielles. Dans les sciences sociales et en géographie, on observe rarement l'ensemble d'une population : on interroge, on mesure, on extrait des traces. L'enjeu est alors double : estimer des grandeurs d'intérêt (ici, des proportions d'opinions) et quantifier l'incertitude liée au hasard du tirage et aux choix de méthode. La séance articule trois étapes : l'échantillonnage (variabilité des estimateurs), l'estimation (intervalle de confiance) et la décision (tests, risques d'erreur).

1. Questions de cours

1.1 Population, échantillon et représentativité

La population statistique désigne l'ensemble des unités auxquelles on souhaite généraliser une conclusion. L'échantillon est un sous-ensemble observé, supposé informer sur la population. La représentativité est cruciale : un échantillon peut être grand mais biaisé (par exemple, sur-représentation d'un groupe), ce qui produit des estimations systématiquement erronées. En pratique, la représentativité dépend du plan de sondage (tirage aléatoire, stratification, quotas), des non-réponses et de la qualité de mesure. L'inférence n'est donc valide que si l'on contrôle autant que possible les biais de sélection et de mesure.

1.2 Théorie de l'échantillonnage : variabilité et convergence

La théorie de l'échantillonnage décrit la fluctuation naturelle des statistiques d'échantillon (moyenne, fréquence, proportion) lorsque l'on répète des tirages indépendants. Même avec une population fixe, deux échantillons donnent rarement exactement la même proportion observée. Cette variabilité est le cœur de l'incertitude statistique. La loi des grands nombres explique une régularité : quand la taille d'échantillon n augmente, la fréquence \hat{p} se stabilise autour de la proportion vraie p, et l'écart-type de \hat{p} décroît approximativement comme $1/\sqrt{n}$. Ainsi, augmenter n réduit l'ampleur des fluctuations.

1.3 Fréquences, proportions et estimateurs

Dans une enquête à modalités (ici trois catégories), on observe des comptages (effectifs) que l'on transforme en fréquences (effectif/n). La proportion p est un paramètre de population ; \hat{p} est un estimateur, variable d'un échantillon à l'autre. Sur un grand nombre de répétitions, on peut étudier la distribution de \hat{p} et vérifier des propriétés importantes : biais (moyenne de \hat{p} proche de p) et précision (dispersion de \hat{p}).

1.4 Intervalle de fluctuation à 95% : sens et utilité

L'intervalle de fluctuation se construit sous l'hypothèse que p est connue (ou supposée). Il répond à la question : "Si la population a une proportion p, quelle fréquence \hat{p} vais-je observer la plupart du temps avec un échantillon de taille n ?". Avec l'approximation normale : $p \pm 1{,}96 \cdot \sqrt{p(1-p)/n}$. Dans une population finie, on applique parfois une correction (CPF) qui

diminue la variance lorsque n représente une fraction notable de N. L'intervalle de fluctuation est utile pour juger la compatibilité d'un résultat observé avec une hypothèse de référence.

1.5 Intervalle de confiance à 95% : construction et interprétation

L'intervalle de confiance vise à encadrer p inconnu à partir de \hat{p} observé. Il s'écrit généralement : \hat{p} ± 1,96·√(\hat{p}(1−\hat{p})/n) (éventuellement avec CPF). Son interprétation correcte est procédurale : si l'on répétait indéfiniment l'échantillonnage et la construction de l'IC, environ 95% des intervalles contiendraient la vraie valeur p. L'IC quantifie donc l'incertitude d'estimation et permet de comparer des groupes ou des territoires en tenant compte des tailles d'échantillon.

1.6 H0/H1 et région critique

Un test statistique formalise un choix : conserver H0 (conformité) ou rejeter H0 au profit de H1 (différence/effet). On choisit une statistique de test et une région critique déterminée par α. Rejeter H0 signifie que l'observation est jugée suffisamment improbable sous H0. La p value mesure, sous H0, la probabilité d'obtenir un résultat au moins aussi extrême que celui observé.

1.7 Risques α et β : arbitrage

Le risque α (type I) est la probabilité de rejeter H0 alors qu'elle est vraie. Le risque β (type II) est la probabilité de ne pas rejeter H0 alors qu'elle est fausse. Réduire α rend le test plus strict et peut augmenter β. La puissance (1−β) dépend aussi de la taille d'échantillon : à effet égal, augmenter n augmente la puissance. Cet arbitrage doit être discuté selon l'enjeu (coût d'un faux positif vs faux négatif).

1.8 Normalité, approximation et TCL

Beaucoup de méthodes reposent sur une approximation normale. Pour les proportions, la normalité n'est pas une hypothèse sur les données brutes mais sur la distribution de l'estimateur : lorsque n est suffisamment grand, \hat{p} est approximativement normal. Le TCL justifie cette approximation en expliquant que des agrégats (moyennes, sommes) tendent vers une normale sous conditions générales, même si la variable de départ n'est pas normale.

1.9 Tests de normalité (Shapiro–Wilk) : lecture critique

Le test de Shapiro–Wilk évalue la compatibilité d'un échantillon avec une loi normale. Une p value faible conduit à rejeter la normalité, mais l'interprétation doit être nuancée : avec de très grands échantillons, de petites déviations deviennent significatives ; avec de petits échantillons, le test peut manquer de puissance. On combine donc tests, diagnostics graphiques (histogrammes, QQ plots) et compréhension du processus générateur.

2. Mise en œuvre avec Python

(Cette partie reprend exactement les calculs réalisés dans le code fourni : moyennes des échantillons, fréquences, intervalles de fluctuation, intervalles de confiance et tests de normalité.)

2. Mise en œuvre avec Python

2.1 Théorie de l'échantillonnage

Le fichier Echantillonnage-100-Echantillons.csv contient cent échantillons indépendants issus d'une population mère de 2185 individus, répartis en trois catégories d'opinion.

Le calcul des moyennes par catégorie sur l'ensemble des échantillons montre que la taille moyenne d'un échantillon est proche de 100 individus. Les fréquences moyennes observées pour chaque opinion sont très proches des fréquences réelles de la population mère.

Cette convergence empirique illustre concrètement la loi des grands nombres : lorsque le nombre d'échantillons augmente, les fréquences moyennes se stabilisent autour des proportions réelles de la population.

Des intervalles de fluctuation à 95 % sont ensuite calculés pour chaque fréquence moyenne, en utilisant l'approximation normale et une correction de population finie. Les fréquences réelles de la population appartiennent à ces intervalles, ce qui confirme la cohérence du processus d'échantillonnage.

2.2 Théorie de l'estimation

Dans un second temps, l'analyse porte sur un échantillon unique, correspondant à la première ligne du fichier. Les fréquences observées dans cet échantillon constituent des estimateurs ponctuels des proportions de la population.

À partir de ces fréquences, des intervalles de confiance à 95 % sont construits. Contrairement aux intervalles de fluctuation, ces intervalles encadrent un paramètre inconnu de la population à partir d'une réalisation unique.

Cette étape met en évidence l'incertitude inhérente à toute estimation fondée sur un seul échantillon, et montre l'importance de l'intervalle de confiance pour interpréter correctement les résultats.

2.3 Théorie de la décision

La théorie de la décision est abordée à travers des tests de normalité appliqués à deux séries de données distinctes. Le test de Shapiro-Wilk est utilisé afin d'évaluer la compatibilité des données avec une distribution normale.

Pour la première série, la p-value est supérieure au seuil de 5 %, ce qui conduit à ne pas rejeter l'hypothèse de normalité. Pour la seconde série, la p-value est inférieure au seuil, ce qui conduit à rejeter l'hypothèse de normalité.

Dans le cas de la série non normale, une procédure complémentaire permet d'identifier la loi statistique la mieux ajustée parmi plusieurs distributions candidates, à l'aide du test de Kolmogorov–Smirnov.

3. Analyse et interprétation

Les résultats obtenus montrent que l'échantillonnage aléatoire permet d'approcher efficacement les caractéristiques d'une population, à condition de disposer d'un nombre

suffisant d'échantillons. Les intervalles de fluctuation confirment la stabilité des fréquences observées, tandis que les intervalles de confiance traduisent l'incertitude liée à une estimation ponctuelle.

Les tests statistiques illustrent quant à eux le rôle central de la décision en statistique inférentielle. Le rejet ou l'acceptation d'une hypothèse dépend à la fois des données observées et du seuil de risque choisi.

4. Difficultés rencontrées et apprentissages

La principale difficulté a consisté à distinguer clairement les rôles respectifs des intervalles de fluctuation et des intervalles de confiance. Cette distinction a été clarifiée par la mise en pratique sur des données simulées.

L'interprétation des tests statistiques a également nécessité une attention particulière, notamment pour éviter une lecture trop mécanique des p-values.

5. Réflexion personnelle

Cette séance met en évidence l'apport fondamental de la statistique inférentielle dans l'analyse des données géographiques. Elle montre que toute conclusion repose sur des hypothèses et sur une gestion explicite de l'incertitude.

Les outils numériques permettent d'automatiser les calculs et de multiplier les simulations, mais ils ne dispensent pas d'une réflexion critique sur les méthodes utilisées et les décisions prises.

Conclusion

La séance 5 a permis de comprendre les principes de l'échantillonnage, de l'estimation et de la décision statistique à travers des exemples concrets. Elle constitue une étape essentielle dans l'apprentissage de l'analyse de données, en reliant les fondements théoriques de la statistique inférentielle à leur mise en œuvre pratique.