CHAPITRE 2-1
DESCRIPTION ET INDUCTION STATISTIQUES EN SCIENCES SOCIALES

Plan

CHAPITRE 2-1
DESCRIPTION ET INDUCTION STATISTIQUES EN SCIENCES SOCIALES

Réf. : Blalock (1979) ; Wonnacott et Wonnacott (1992)

La statistique est essentiellement une collection de méthodes mathématiques pour traiter les données, soit en vue d'en résumer l'information, soit pour généraliser l'information qu'elles contiennent. La *statistique descriptive* est la partie de la statistique qui a pour but de *résumer l'information*. La démarche de *généralisation* de l'information s'appelle l'*induction statistique*.

2-1.1 Statistique descriptive

Devant un ensemble de données tant soit peu important, le cerveau humain est incapable de saisir d'un seul coup toute l'information détaillée qu'il contient [1]. La solution à ce problème consiste à laisser tomber les détails pour concentrer son attention sur les grands traits. La statistique descriptive permet de traiter méthodiquement les données pour condenser l'information qu'elles contiennent. En calculant des pourcentages, des moyennes, des écarts type ou des coefficients de corrélation, on peut arriver à une vision globale des données. Il ne faut cependant pas perdre de vue qu'en résumant ainsi les données, on laisse de côté une partie de l'information qu'elles contiennent : cela peut induire en erreur, à moins que l'on ne soit prudent dans l'interprétation.

2-1.2 Induction statistique

Il est rare, particulièrement en sciences sociales, de posséder les données pertinentes à la totalité d'un phénomène étudié. Le plus souvent, les observations dont on dispose ne portent que sur une partie du phénomène. D'où la nécessité de généraliser à partir d'une information incomplète. Pour être véritablement scientifique toutefois, une généralisation doit se fonder sur des principes épistémologiques [2].

---

1 Selon Georges Ifrah (1994, Tome 1, p. 33-34, « Les limites de la perception directe des nombres »), le cerveau ne peut pas apréhender concrètement, c'est-à-dire sans les compter (ce qui constitue une abstraction), plus de quatre objets à la fois. Par exemple, on ne peut pas « à l'oeil », sans compter, faire la différence entre cinq et six objets.

2 L'épistémologie est une partie de la philosophie. Elle consiste en l'étude critique des sciences, en vue de déterminer leur origine logique, leur valeur et leur portée.

Les méthodes d'induction statistique sont justement une expression mathématique de principes épistémologiques en vertu desquels, à partir de l'information contenue dans un ensemble de données particulier, on peut arriver à des propositions de portée plus générale. L'induction statistique est donc une façon scientifiquement valide de passer du particulier [3] au général.

Idée-clé No 1 :

Les méthodes d'induction statistique sont une expression mathématique de principes épistémologiques en vertu desquels, à partir de l'information contenue dans un ensemble de données particulier, on peut arriver à des propositions de portée plus générale.

Concrètement, la démarche de l'induction statistique a pour objectif de dégager des conclusions générales quant aux diverses caractéristiques d'une *population*, à partir de faits observés sur un *échantillon* tiré de cette population. La statistique emploie le mot *paramètres* pour désigner les caractéristiques de la population, et le mot *statistiques* [4] pour désigner les caractéristiques de l'échantillon.

Il importe de garder à l'esprit que les paramètres sont normalement considérés comme des valeurs *fixes* se rapportant à la population et qu'ils sont généralement *inconnus* (puisque la population elle-même n'est pas connue dans sa totalité). Au contraire, puisqu'on peut tirer plus d'un échantillon d'une population donnée, les statistiques sont des valeurs qui peuvent *varier* d'un échantillon à l'autre ; mais les valeurs des statistiques pour un échantillon particulier sont *connues* ou peuvent être calculées. On ne sait pas toutefois à quel point un échantillon est représentatif de la population en général, ni dans quelle mesure une statistique calculée sur cet échantillon se rapproche du paramètre correspondant de la population, inconnu.

Exemples d'induction statistique :

1. Sur la base des réponses obtenues par sondage auprès d'un échantillon de la population d'une ville, estimer la proportion des citoyens qui sont favorables à un certain projet d'aménagement urbain.

---

3 Il y a une boutade en anglais qui traduit bien le caractère particulier des données : « "Data" is the plural of "anecdote" ».

4 Noter que *les* statistiques désignent des caractéristiques d'un échantillon, alors que *la* statistique désigne l'ensemble des méthodes mathématiques d'analyse de données.

2. Partant de l'hypothèse (du modèle), acceptée *a priori*, que la relation macroéconomique entre le revenu des ménages et les investissements en construction résidentielle est décrite par une équation de la forme

   $I = a + b\,R$

   (où I est le montant des investissements et R le revenu agrégé),

   estimer la valeur des paramètres a et b à partir des données publiées par Statistique Canada pour le Québec de 1974 à 1994 [5].

Il est à noter que les mesures utilisées en statistique descriptive (moyenne, écart type, ...) sont également utilisées dans le contexte de l'induction statistique. En statistique descriptive cependant, la distinction entre paramètres et statistiques n'existe pas, parce que la statistique descriptive ne fait pas de distinction entre population et échantillon.

2-1.3 Les probabilités et l'induction statistique : le lien aléatoire entre un échantillon et la population

Avec l'induction statistique, on quitte le domaine de la certitude. En effet, l'induction statistique a pour point de départ un *échantillon*, qui n'est qu'*un des échantillons possibles* qu'on aurait pu tirer de la population étudiée : si l'on tire d'une population donnée un échantillon d'une taille donnée, en suivant une procédure donnée, et si ensuite on recommence, l'échantillon obtenu au second tirage sera probablement différent du premier. Pour une population donnée, il y a donc un grand nombre d'échantillons possibles. L'ensemble des échantillons possibles forme aussi une population au sens statistique : les « individus » de cette population sont les échantillons.

Par exemple, l'ensemble des abonnés du téléphone dans la ville de Montréal forme une population. On pourrait tirer de cette population un échantillon de 1000 abonnés choisis au hasard dans le bottin téléphonique. On pourrait ensuite recommencer et tirer un second échantillon, puis un troisième, etc. (en fait, on pourrait recommencer à l'infini si l'on suivait une procédure d'échantillonnage où les abonnés qui font partie d'un échantillon redeviennent

---

5 Cet exemple de modèle est évidemment trop simpliste.

éligibles pour le suivant [6]). L'ensemble de tous les échantillons possibles de 1000 abonnés choisis au hasard dans le bottin est la population des échantillons.

Parmi les échantillons possibles, certains sont représentatifs de la population étudiée, tandis que d'autres le sont moins. Puisqu'on ne connaît pas la population autrement qu'à travers l'échantillon, on ne peut jamais savoir avec certitude à quel point l'échantillon particulier que l'on a tiré est représentatif de la population en général. Le lien entre l'échantillon et la population est donc essentiellement *aléatoire* (c'est-à-dire influencé par le hasard).

Idée-clé No 2 :

Un échantillon n'est qu'un des échantillons possibles de même taille qu'on aurait pu tirer de la population étudiée. D'où, le lien aléatoire entre l'échantillon et la population.

Concrètement, il s'ensuit que l'on ne peut pas savoir avec certitude dans quelle mesure une statistique calculée à partir des données d'un échantillon se rapproche du paramètre inconnu correspondant dans la population. Cependant, la théorie des probabilités nous donne des outils pour évaluer la *probabilité* que l'écart (l'erreur d'estimation) entre la statistique et le paramètre se situe à l'intérieur d'une certaine marge. C'est donc sur la théorie des probabilités que se fondent les règles de l'induction statistique.

Idée-clé No 3 :

La théorie des probabilités nous donne des outils pour évaluer la probabilité que l'écart (l'erreur d'estimation) entre la statistique et le paramètre se situe à l'intérieur d'une certaine marge.

6 Ne pas confondre avec l'échantillonnage *avec remplacement*, où les individus qui constituent un échantillon sont tirés de manière séquentielle et où un individu qui est tiré redevient éligible pour le tirage suivant (Freund, 1970, p. 183).