CHAPITRE 2-2
L'INDUCTION STATISTIQUE

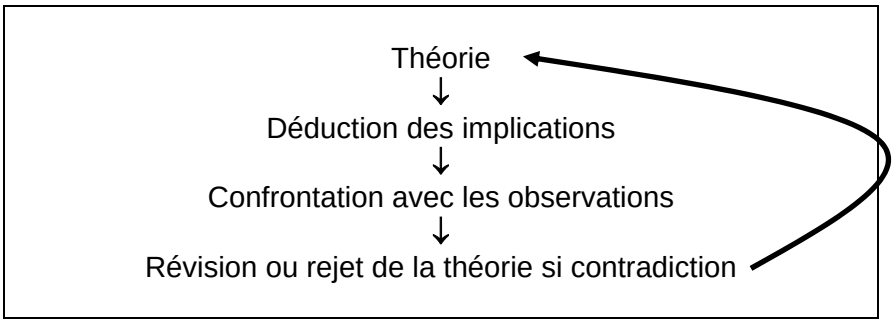
Plan

CHAPITRE 2-2
L'INDUCTION STATISTIQUE

2-2.1 L'induction statistique dans la méthode scientifique : modèles théoriques et modèles aléatoires

Réf. : Malinvaud (1969), chap. 1 et 2 et Blalock (1972), chap. 2 et 8.

Nous examinerons bientôt la logique fondamentale des tests d'hypothèse, qui sont un aboutissement de l'induction statistique. Auparavant, il est utile de situer l'induction statistique dans le cadre de la méthode scientifique hypothético-déductive. Nous partirons d'un schéma simplifié de la méthode [1] :



Les théories, que ce soit en sciences sociales ou en sciences physiques, sont souvent formalisées comme *modèles* :

> « Un modèle consiste en la représentation formelle d'idées ou de connaissances relatives à un phénomène » (Malinvaud, 1969, p. 45).

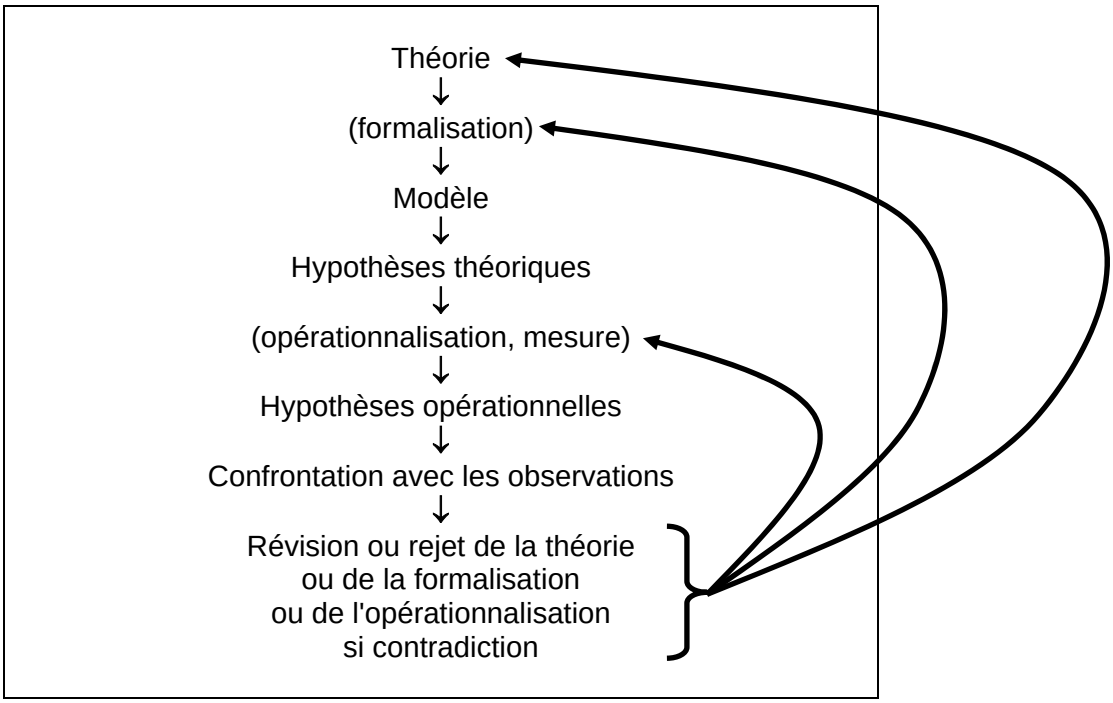
Lorsque, comme c'est souvent le cas, la formalisation choisie est mathématique, on a un *modèle mathématique*.

Une théorie est une construction intellectuelle de caractère *hypothétique*, c'est-à-dire une hypothèse globale à propos d'un phénomène. Il en est de même du modèle qui en donne une représentation formelle, ainsi que des parties du modèle. Mais dans le modèle comme dans la théorie, les concepts ne sont pas définis en termes opérationnels : les hypothèses théoriques ne peuvent pas être confrontées directement aux observations. Pour ce faire, il faut *traduire les*

---

[1] Il faut reconnaître que ce schéma est tronqué. Pour une présentation plus approfondie du processus de développement de la connaissance, voir Robert (1993). Ce qui est représenté ici correspondrait plutôt à ce que Kuhn, dans *La structure des révolutions scientifiques* (1983), appelle la science « normale ».

hypothèses théoriques en hypothèses opérationnelles, ce qui implique de définir des *mesures* (voir le chapitre 1-1). Le schéma complété devient :



Dans ce schéma, il est évident qu'une contradiction entre les observations et les prédictions de la théorie peuvent découler, non seulement de la théorie elle-même, mais aussi de sa formalisation ou de son opérationnalisation. Dans la réalité concrète, le processus de la recherche est moins limpide que le laissait voir le premier schéma...

La statistique intervient au niveau de la confrontation avec les observations : la plupart du temps, les observations sont faites sur un échantillon, alors que le modèle, la théorie, portent sur une population. Or, la plupart des modèles mathématiques théoriques sont déterministes : les variables qui représentent les concepts sont liées les unes aux autres par des relations fonctionnelles (des fonctions mathématiques), sans élément aléatoire.

Pourtant, puisque chaque échantillon n'est que l'un des multiples échantillons possibles, le lien entre un échantillon et la population dont il est tiré est *essentiellement aléatoire*. Pour être confrontés aux observations, les modèles déterministes ont besoin d'être complétés de façon à tenir compte de cet élément aléatoire. Lorsqu'on combine un modèle déterministe et un modèle du lien aléatoire entre l'échantillon et la population, on obtient un *modèle aléatoire* (*probability*

model) [2]. Certains auteurs traduisent cette distinction en désignant par « modèle structurel » (*structural model*) le modèle théorique déterministe et par « modèle d'échantillonnage » (*sampling model*) le modèle du lien entre l'échantillon et la population [3].

Idée-clé No 4 :

Pour être confrontés aux observations, les modèles déterministes ont besoin d'être complétés de façon à tenir compte du lien aléatoire entre la population et l'échantillon. Lorsqu'on combine un modèle déterministe et un modèle du lien aléatoire entre l'échantillon et la population, on obtient un modèle aléatoire (probability model). Dans un modèle aléatoire, le modèle du lien entre l'échantillon et la population s'appelle « modèle d'échantillonnage » (sampling model).

Jusqu'à maintenant, nous n'avons mentionné qu'une seule source d'aléatoire. En fait, il y a trois « portes » par lesquelles l'aléatoire s'introduit dans les modèles [4] :

1. Il y a d'abord la nature aléatoire, déjà mentionnée, du lien entre un échantillon et la population dont il est tiré.
2. Les variables opérationnelles sont des mesures imparfaites des concepts et on peut considérer que l'erreur de mesure est aléatoire (c'est-à-dire déterminée au hasard). On peut donc représenter par un modèle aléatoire l'influence des erreurs de mesure qui interviennent lors de la traduction des hypothèses théoriques en hypothèses opérationnelles (les modèles de la « théorie des erreurs » en sciences physiques ont d'ailleurs été parmi les premiers modèles aléatoires).
3. Enfin, certains phénomènes nous apparaissent comme aléatoires en soi et ils ne peuvent pas être représentés adéquatement par des modèles théoriques non aléatoires. Le hasard dans ces modèles est un concept qui recouvre tantôt une indétermination fondamentale

---

2 « [...] un modèle aléatoire définit, pour tout ensemble de valeurs données aux variables exogènes, la loi de probabilité correspondante des variables endogènes » (Malinvaud, 1969, p. 59).

3 Plus précisément, Upton et Fingleton (1985, p. 264) appellent « structural model » la spécification du lien fonctionnel entre la variable dépendante et les variables indépendantes ; ils appellent « sampling model » l'hypothèse quant à la distribution de probabilité de la variable dépendante (ou, ce qui est équivalent, du terme d'erreur).

4 Malinvaud écrit : « On sait que l'emploi du calcul des probabilités pour l'analyse des données statistiques est justifié par l'une ou l'autre des deux considérations suivantes. Ou bien le phénomène étudié est assimilé à un processus comportant une détermination aléatoire de certaines grandeurs ; ces grandeurs sont alors considérées comme aléatoires dans l'univers [NDLR : c'est-à-dire dans la population] comme dans l'échantillon observé. Ou bien la sélection des unités observées résulte d'un tirage aléatoire ; la composition de l'échantillon est alors aléatoire, donc aussi les données obtenues, même si elles portent sur des grandeurs non aléatoires »

(comme en physique des particules), tantôt une multitude de facteurs inobservables (comme c'est plus souvent le cas en sciences sociales [5]), dont les manifestations apparaissent comme régies par des lois de probabilité.

Quoi qu'il en soit, un modèle aléatoire est, comme tout modèle, de caractère hypothétique ; il est constitué entre autres d'*hypothèses sur la structure aléatoire*, sur les lois de probabilité qui régissent le hasard. Dans la confrontation avec les observations, ces hypothèses ne sont pas remises en question (du moins pas toutes). Elles sont pour ainsi dire le péage exigé pour franchir le pont du connu à l'inconnu, puisque l'induction statistique va « au-delà » des données observées. Néanmoins, bien que l'induction repose sur des hypothèses, il y a un gain épistémologique lorsque les hypothèses sur lesquelles se fonde l'induction sont moins restrictives que les résultats obtenus par l'induction.

Idée-clé No 5 :

Un modèle aléatoire est, comme tout modèle, de caractère hypothétique ; il est constitué entre autres d'hypothèses sur la structure aléatoire, sur les lois de probabilité qui régissent le hasard.

Plus généralement, il faut reconnaître que, contrairement à ce que pourrait laisser croire la simplicité du schéma présenté ci-haut, la confrontation de la théorie, des modèles et des hypothèses avec les observations est rarement totale. Chaque exercice de confrontation repose en fait sur un modèle plus général qui n'est pas remis en question. Cela est particulièrement vrai de l'induction statistique et des tests d'hypothèse dont il est question un peu plus loin. La plupart du temps en effet, les tests d'hypothèse portent sur des formes particulières d'un modèle théorique général, qui n'est pas remis en question, et s'appuient sur un modèle aléatoire, qui n'est pas remis en question non plus [6].

(Malinvaud, 1969, p.62). Malinvaud poursuit en disant que le premier type de justification lui semble plus approprié au contexte de l'économétrie.

5 Pensons en particulier aux modèles d'utilité aléatoire (*random utility*) sous-jacents aux modèles de choix discrets (*discrete choice*) logit, probit, etc. Ces modèles sont abordés au chapitre 4-3.

6 Une « confrontation totale » serait le propre d'une révolution scientifique à la Kuhn. Il est douteux cependant que l'induction statistique joue un rôle prédominant dans le processus de changement de paradigme d'une révolution scientifique. Il est néanmoins vrai qu'il existe des tests « de niveau supérieur », pour ainsi dire, qui portent sur certains aspects du modèle aléatoire. Mais ces tests reposent eux-mêmes sur des modèles aléatoires plus généraux qui, à ce niveau, ne sont pas remis en question. On peut imaginer un test du modèle aléatoire du test du modèle aléatoire... Mais peu importe la « hauteur » du niveau auquel on s'élève, il y aura toujours au-dessus un modèle d'échantillonnage qui n'est pas remis en question.

Idée-clé No 6 :

Chaque exercice de confrontation entre une théorie et les observations repose en fait sur un modèle plus général qui n'est pas remis en question.

Idée-clé No 7 :

La plupart du temps, les tests d'hypothèse portent sur des formes particulières d'un modèle théorique général, qui n'est pas remis en question, et s'appuient sur un modèle aléatoire, qui n'est pas remis en question non plus.

2-2.2 Quelques concepts-clés de la théorie des probabilités

Réf. : Wonnacott et Wonnacott (1992, chap. 3 et chap. 4, sections 4.1-4.2)

Avant d'aborder l'induction statistique proprement dite, il faut rappeler sommairement les définitions de quelques concepts clés de la théorie des probabilités.

2-2.2.1 CONCEPTS FONDAMENTAUX

Hasard (du mot arabe *az-zahr*, « le dé ») : « cause fictive de ce qui arrive sans raison apparente ou explicable » (Robert).

Événement aléatoire (alea = dés à jouer en latin) [7]

Événement dont la réalisation ou non dépend du hasard. Par exemple, il est utile de considérer que, dans l'ensemble de tous les échantillons qu'on peut tirer d'une population, chaque possibilité est un événement aléatoire. Lorsqu'on tire un échantillon, l'un de ces événements se réalise, tandis qu'aucun des autres ne se réalise.

Variable aléatoire

Variable dont la valeur est le résultat d'événements aléatoires [8]. Puisque le résultat du tirage d'un échantillon est un événement aléatoire, toutes les mesures que l'on peut faire sur un

---

7 Pensons au fameux « Alea jacta est » (les dés sont jetés) de Jules César franchissant le Rubicon.

8 Dans la plupart des manuels de statistique, la distinction entre la variable aléatoire et ses valeurs possibles ou observées se traduit par une notation où $\boldsymbol{X}$ désigne la variable aléatoire et x ses valeurs possibles ou observées.

échantillon sont des variables aléatoires. Cela s'applique aux données brutes et aux statistiques calculées à partir de ces données.

On distingue les variables aléatoires **discrètes**, qui ne peuvent prendre que certaines valeurs (des nombres entiers la plupart du temps) et les variables aléatoires **continues**, dont la valeur peut être n'importe quel nombre réel sur un intervalle donné (ouvert ou fermé). Les variables aléatoires continues représentent un ensemble de possibilités *infini*, tandis que les variables aléatoires discrètes peuvent représenter un ensemble de possibilités *fini*, lorsque leur domaine de variation est fini [9].

Probabilité d'un événement aléatoire

Nous avons tous une notion intuitive de ce qu'est une probabilité mais il n'est pas facile de donner de ce concept une définition rigoureuse. On peut aborder la notion de probabilité de trois façons [10].

1) On peut concevoir la probabilité dans le contexte d'une suite d'« expériences » ou d'« essais » où le résultat de chaque tentative est un « succès » (l'événement se produit) ou un « échec » (l'événement ne se produit pas) ; c'est la définition « fréquentiste » de la probabilité, en termes de fréquence relative d'un événement aléatoire. Dans une telle suite d'expériences (tirer à pile ou face ou lancer un dé), la probabilité d'un événement aléatoire (comme obtenir « pile » ou un « six ») est défini comme la proportion des expériences où cet événement se réalisera en moyenne.

2) La probabilité d'un événement aléatoire peut être définie comme l'évaluation que l'on se fait, sur une échelle de 0 à 100 %, des chances que cet événement se produise (définition subjectiviste ou bayesienne).

3) On peut enfin considérer le concept de probabilité comme premier et non définissable, pour ensuite énoncer un système d'axiomes auquel doit se conformer toute *mesure* de probabilité.

Ici, nous écrirons « variable aléatoire » au long lorsque ce sera nécessaire ; autrement, nous utiliserons x pour désigner les deux.

9 Une variable aléatoire dont le domaine de variation est l'ensemble des entiers naturels est une variable discrète, mais l'ensemble de ses valeurs possibles est infini.

10 Voir Wonnacott et Wonnacott (1991), p. 110-114.

2-2.2.2 DISTRIBUTIONS DE PROBABILITÉ

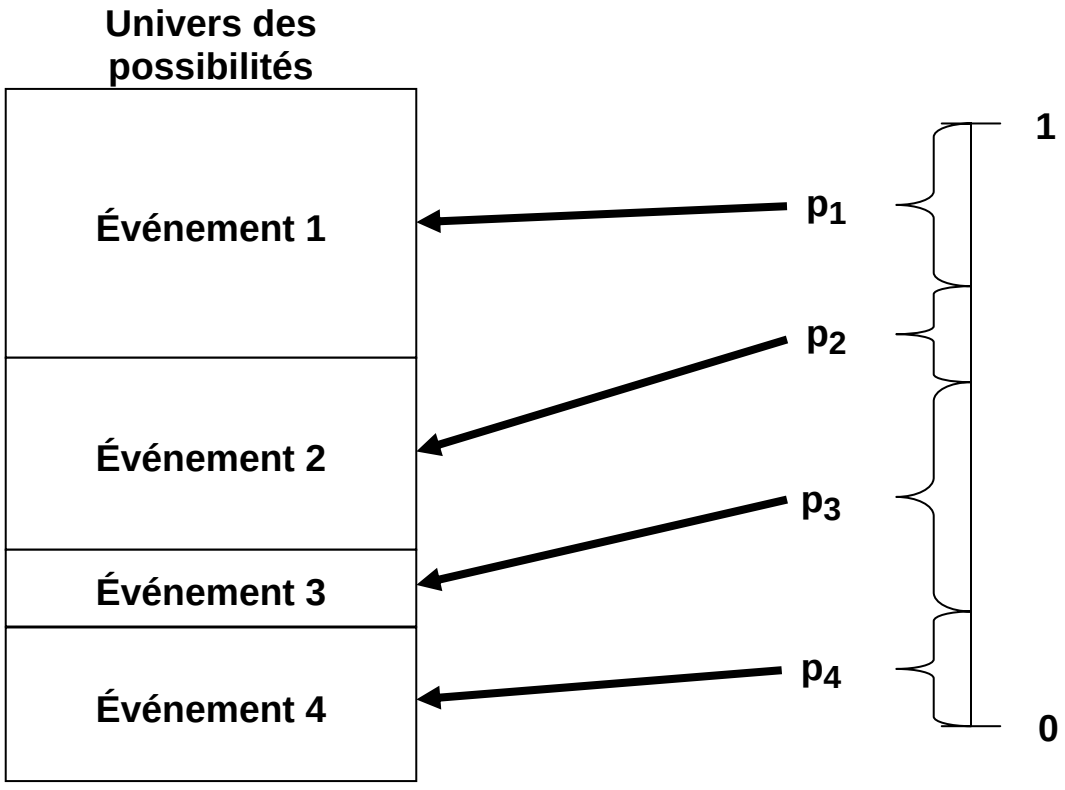
Fonction de distribution de probabilité ou distribution de probabilité

Correspondance qui associe une probabilité à chacun d'un ensemble exhaustif d'événements mutuellement exclusifs (possibilités). Exemple : lorsqu'on tire à pile ou face une fois avec une pièce qui n'est pas truquée, la fonction de probabilité est donnée par

Prob(*face*) = Prob(*pile*) = 0,5

La distribution de probabilité s'apparente à une *distribution de fréquence relative*, mais elle s'en distingue en ce que la distribution de fréquences rapporte des fréquences *observées*, tandis que la distribution de probabilité assigne à chaque événement la fréquence relative qu'il aurait en moyenne dans le contexte d'une suite infinie d'expériences (voir ci-devant, la définition « fréquentiste » de la probabilité).

Une fonction de distribution de probabilité est une correspondance...



Fonction de distribution cumulative d'une variable aléatoire

La fonction de distribution cumulative d'une variable aléatoire est une fonction $F(x)$ (une correspondance) qui, pour chaque valeur possible x de la variable aléatoire, donne la probabilité que la variable aléatoire prenne une valeur *inférieure ou égale* à x [11].

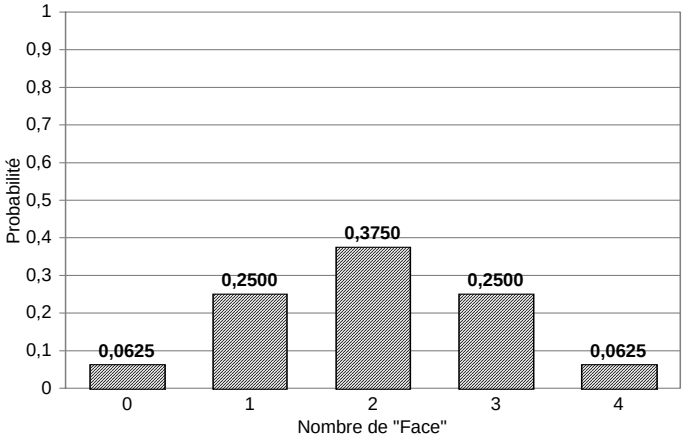
Par exemple, si l'on tire à pile ou face quatre fois, le nombre de fois que l'on obtient face est une variable aléatoire discrète, dont la fonction de probabilité et la fonction de distribution cumulative sont données dans le tableau suivant (cette distribution s'appelle la distribution binomiale). Les figures qui accompagnent le tableau illustrent les notions de distribution de probabilité et de fonction de distribution cumulative.

11 Si les valeurs possibles de la variable aléatoire ne sont pas numériques – comme « pile » et « face » — il faut définir au préalable l'ordre dans lequel on range ces valeurs pour que la relation « inférieure ou égale » ait un sens.

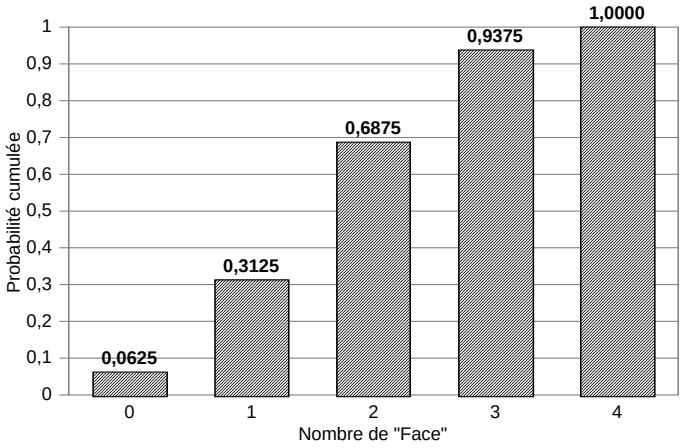
Fonction de probabilité et fonction de distribution cumulative

Nombre de face	Probabilité		Probabilité cumulée
x_i	$f(x_i)$	$F(x_{i-1})$	$F(x_i)$
0	1/16		1/16
1	4/16	+ 1/16	= 5/16
2	6/16	+ 5/16	= 11/16
3	4/16	+ 11/16	= 15/16
4	1/16	+ 15/16	= 16/16

Fonction de probabilité



Fonction de distribution cumulative



2-2.2.3 DISTRIBUTION D'ÉCHANTILLONNAGE

Réf. : Wonnacott et Wonnacott (1992, p. 224-226)

Le concept de **distribution d'échantillonnage** (*sampling distribution*) est central en induction statistique. C'est la forme opérationnelle que prend le modèle d'échantillonnage (le modèle du lien entre un échantillon et la population ; voir 2-2.1).

En effet, une distribution d'échantillonnage est une distribution de probabilité associée à une statistique. Rappelons qu'une statistique est une caractéristique d'un échantillon, alors qu'un paramètre est une caractéristique d'une population.

Or, nous avons vu que tout échantillon n'est qu'un des échantillons de la même taille que l'on pourrait tirer de la population étudiée. Donc, selon l'échantillon tiré, la statistique pourrait prendre différentes valeurs. Et puisque l'échantillon est tiré au hasard, la valeur que prend la statistique est aléatoire et la statistique elle-même est une variable aléatoire. La *distribution d'échantillonnage* de la statistique est sa distribution de probabilité dans la population des échantillons d'une taille donnée qu'on peut tirer au hasard de la population étudiée.

Idée-clé No 8 :

La distribution d'échantillonnage d'une statistique est sa distribution de probabilité dans la population des échantillons d'une taille donnée qu'on peut tirer au hasard de la population étudiée.

En général, la distribution d'échantillonnage d'une statistique dépend des paramètres de la population étudiée. C'est cette dépendance qui permet, à partir de la valeur observée d'une statistique, de formuler des énoncés probabilistes à propos des paramètres. Nous expliciterons cette démarche lorsqu'il sera question des tests d'hypothèse.

Par exemple, supposons que l'on cherche à déterminer si une pièce de monnaie utilisée pour jouer à « pile ou face » est truquée. Si la pièce n'est pas truquée, elle devrait « en moyenne » tomber aussi souvent sur « pile » que sur « face ». Mais pour connaître la vraie moyenne, il faudrait tirer la pièce un nombre littéralement infini de fois, parce que, quel que soit le nombre d'essais, on ne pourra jamais être certain de l'issue des essais supplémentaires qu'on pourrait faire. La population étudiée est donc infinie. Pour décider si l'on doit considérer la pièce comme truquée ou non, une seule possibilité : faire un certain nombre d'essais, calculer la proportion de « pile » et de « face » et accepter la pièce comme honnête si cette proportion (fréquence

relative) est suffisamment proche de 50 %. La distribution d'échantillonnage de cette proportion est la distribution de probabilité de cette statistique. Cette distribution dépend du nombre d'essais et de la vraie valeur de la probabilité.

Schéma 1a

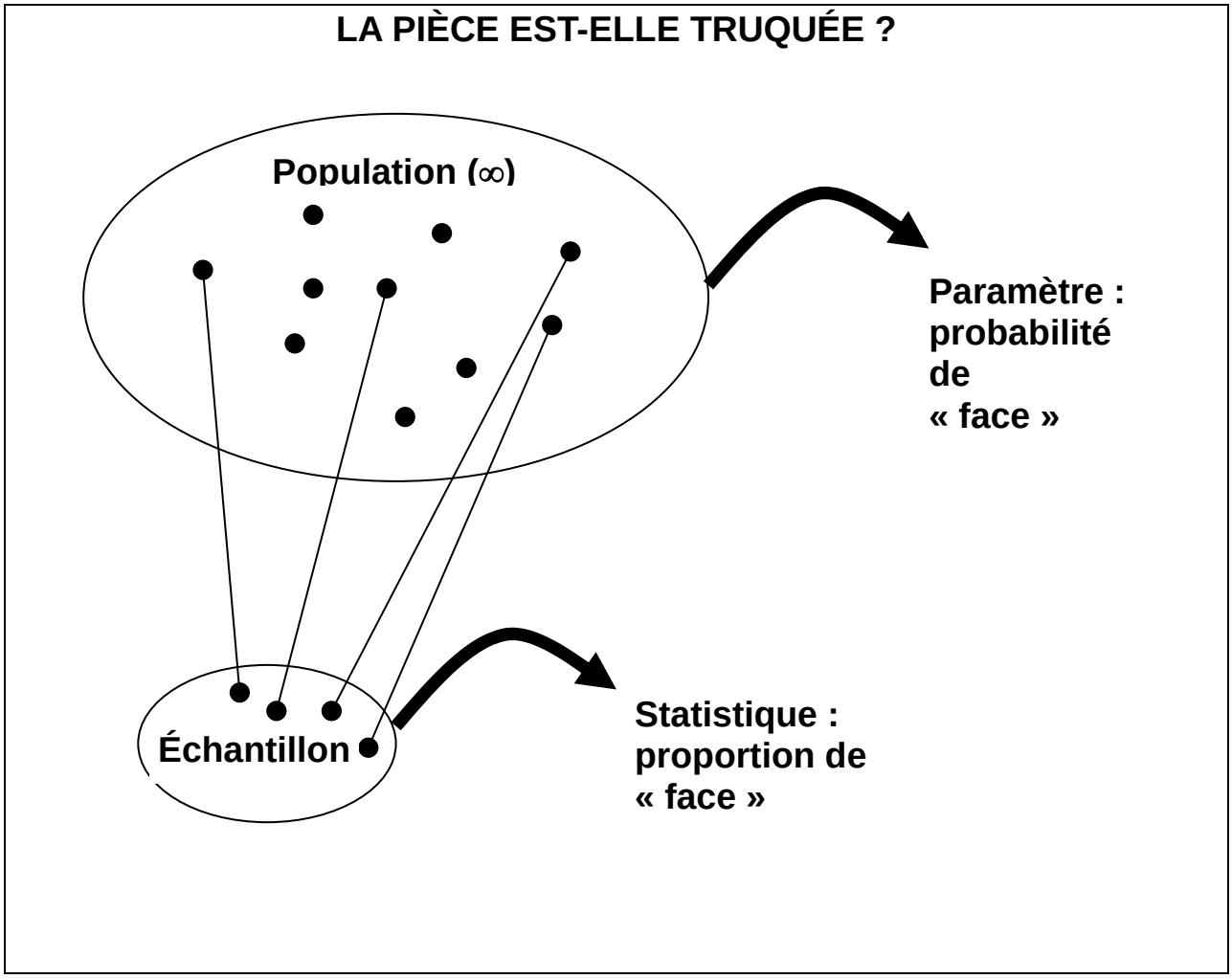
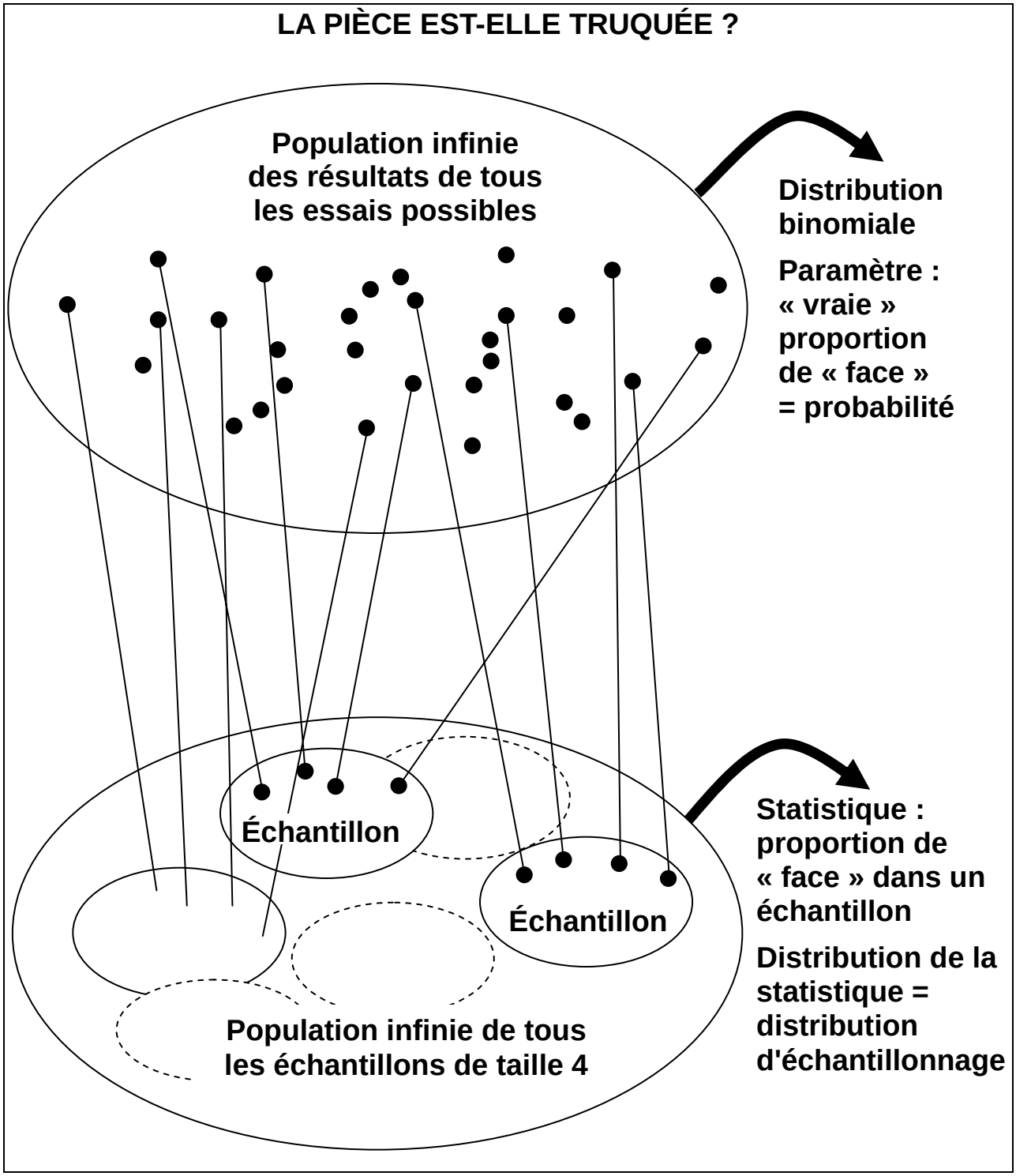


Schéma 1b


LA PIÈCE EST-ELLE TRUQUÉE ?
Population infinie des résultats de tous les essais possibles
Distribution binomiale
Paramètre : « vraie » proportion de « face » = probabilité
Échantillon
Échantillon
Statistique : proportion de « face » dans un échantillon
Distribution de la statistique = distribution d'échantillonnage
Population infinie de tous les échantillons de taille 4

2-2.2.4 VARIABLES ALÉATOIRES CONTINUES : FONCTION DE DENSITÉ DE PROBABILITÉ ET ESPÉRANCE MATHÉMATIQUE

Fonction de densité de probabilité d'une variable aléatoire continue

Réf. : Wonnacott et Wonnacott (1992, p. 138-140) proposent une autre présentation de la fonction de densité de probabilité.

Pour bien comprendre la logique des tests statistiques, il est extrêmement important de saisir la différence entre une fonction de *densité* de probabilité et une fonction de *distribution* de probabilité (même si l'expression plus générale « distribution de probabilité » s'applique indifféremment aux deux).

La plupart des statistiques sont des variables aléatoires continues. Or, avec une variable aléatoire continue, le nombre de valeurs possibles est infini. Il s'ensuit que la probabilité que la variable aléatoire prenne *une valeur en particulier* est normalement *infiniment petite :* on ne peut pas associer une probabilité à chaque valeur possible de la variable aléatoire, de sorte que le concept de fonction de probabilité tel que défini ci-haut ne s'applique pas.

Idée-clé No 9 :

Avec une variable aléatoire continue, le nombre de valeurs possibles est infini. Il s'ensuit que la probabilité que la variable aléatoire prenne une valeur en particulier est normalement infiniment petite.

C'est pourquoi, s'agissant de variables continues, on fait appel à la notion de fonction de densité de probabilité. La fonction de densité de probabilité se définit à partir de la fonction de probabilité cumulative. Car, même si la probabilité qu'une variable aléatoire continue prenne une valeur x en particulier est infiniment petite, il y a une probabilité positive [12] que sa valeur ne dépasse pas cette valeur x : il existe donc une fonction

$$F(x) = \text{Prob (variable aléatoire} \leq x)$$

qui n'est autre que la distribution *cumulative* en fonction des valeurs possibles x.

Par exemple, on peut considérer la durée de vie d'une ampoule électrique incandescente comme une variable aléatoire continue. Quelle est la probabilité qu'une ampoule dure *exactement* 112 heures, 23 minutes, 14 secondes et trois centièmes ? Évidemment, cette

[12] C'est-à-dire non infiniment petite.

probabilité est infiniment petite. Néanmoins, la probabilité que cette ampoule dure 112 heures, 23 minutes, 14 secondes et trois centièmes *ou moins* est assurément positive. Cette dernière probabilité est la probabilité cumulative

$F(x)$ = Prob (durée de vie de l'ampoule $\leq x$)

où x = 112h 23m 14,03s.

En somme, avec une variable aléatoire continue, le concept de fonction de probabilité tel que défini ci-haut ne s'applique pas, mais la fonction de distribution cumulative existe bel et bien en général. Et c'est à partir de la fonction de distribution cumulative $F(x)$ que l'on définit la fonction de *densité* $f(x)$: c'est une fonction qui, pour chaque valeur possible de la variable aléatoire, donne le taux (vitesse, densité) auquel augmente la probabilité cumulative en ce point de la fonction. Techniquement, la fonction de densité de probabilité est la *dérivée* (la pente) de la fonction de distribution cumulative d'une variable continue [13] :

$$f(x) = \frac{d}{dx} F(x)$$

Idée-clé No 10 :

Avec une variable aléatoire continue, le concept de fonction de probabilité tel que défini pour une variable aléatoire discrète ne s'applique pas, mais la fonction de distribution cumulative existe bel et bien en général. Et c'est à partir de la fonction de distribution cumulative *F*(*x*) que l'on définit la fonction de *densité f*(*x*).

Pour bien comprendre les tests statistiques, il est extrêmement important de saisir la différence entre une fonction de *densité* de probabilité et une fonction de *distribution* de probabilité, parce que les tests sont des raisonnements sur des probabilités et que ces probabilités sont calculées à l'aide de fonctions de densité de probabilité. Or, l'ordonnée d'une fonction de densité (sa hauteur) *n'est pas une probabilité* (alors que l'ordonnée d'une fonction de probabilité, elle, est une probabilité). Par contre, la *surface sous la courbe* d'une fonction de densité est une probabilité : techniquement, puisque $f(x)$ est la *dérivée* de $F(x)$, il s'ensuit que $F(x)$ est donnée par l'*intégrale* de $f(x)$.

---

[13] La fonction de densité joue par rapport à la fonction de distribution cumulative le même rôle que la vitesse par rapport à la distance parcourue : sur un graphique de la distance parcourue en fonction du temps écoulé, la pente de la courbe donne la vitesse en cet instant. La dérivée est la vitesse *instantanée* ; celle-ci est différente de la vitesse moyenne sur un intervalle donné, qui correspond dans le graphique de la distance parcourue à la pente moyenne sur cet intervalle. On peut aussi dire que la fonction de densité est à la fonction de distribution cumulative ce qu'est le débit d'un robinet au volume d'eau accumulé dans un réservoir.

Idée-clé No 11 :

L'ordonnée d'une fonction de densité (sa hauteur) *n'est pas une probabilité* (alors que l'ordonnée d'une fonction de probabilité, elle, est une probabilité). Par contre, la *surface sous la courbe* d'une fonction de densité est une probabilité : techniquement, puisque *f*(*x*) est la *dérivée* de *F*(*x*), il s'ensuit que *F*(*x*) est donnée par l'*intégrale* de *f*(*x*).

Ainsi,

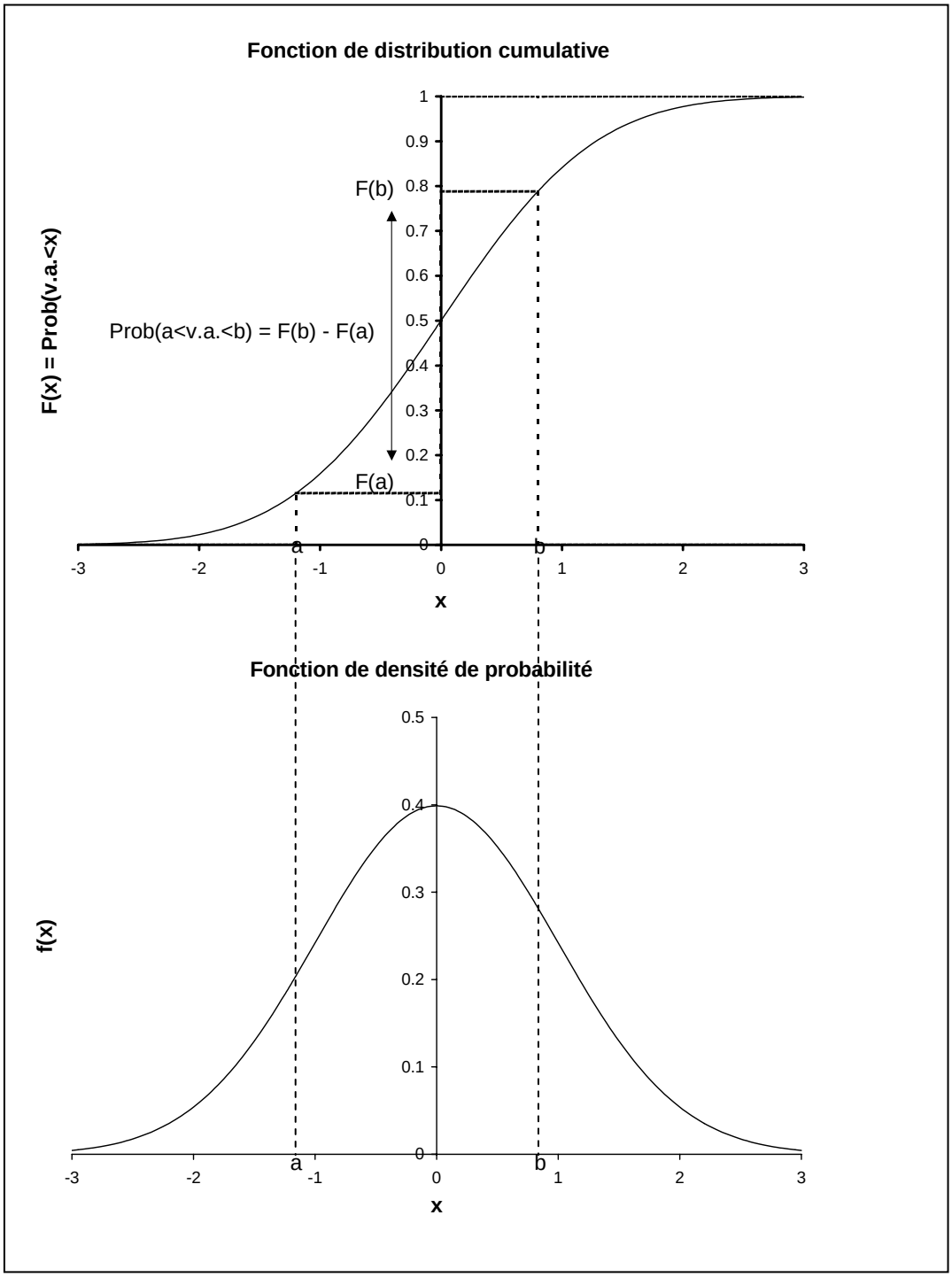
$$\text{Prob}(a \leq \text{variable aléatoire} \leq b) = F(b)-F(a)$$

$$\text{Prob}(a \leq \text{variable aléatoire} \leq b) = \text{surface sous } f(x) \text{ entre } a \text{ et } b = \int_a^b f(x)\,dx$$

Naturellement,

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1 = F(+\infty)$$

FIGURE 1 – FONCTION CUMULATIVE ET FONCTION DE DENSITÉ



Espérance mathématique

Réf. : Wonnacott et Wonnacott (1992, p. 154-155, 184-185)

La moyenne d'une variable aléatoire continue dans une population infinie ne peut pas se calculer à l'aide de la formule bien connue

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$ (où x_i sont les valeurs possibles de la variable aléatoire)

tout simplement parce que le nombre de valeurs possibles, n, est infini. Le concept d'*espérance mathématique* est une généralisation de la moyenne. Pour une variable continue, l'espérance mathématique est donnée par [14]

$$E(x) = \int_{-\infty}^{+\infty} f(x)\, x\, dx$$

Ainsi, quand on parle de la moyenne d'une variable aléatoire continue dans la population, on veut dire

$$\mu_x = E(x)$$

Et quand on parle de la variance d'une variable aléatoire continue dans la population, on veut dire

$$\sigma_x^2 = E\left\{ [x - E(x)]^2 \right\} = \int_{-\infty}^{+\infty} f(x)\, [x - E(x)]^2\, dx$$

Dans le cadre de ce cours, il suffit de se rappeler que les formules de calcul de la moyenne et de la variance peuvent se généraliser au cas d'une variable aléatoire continue dans une population infinie. Pour le reste, on peut se contenter de l'intuition que l'on a du concept de moyenne et de celui de variance à partir des formules de la statistique descriptive.

Loi normale

Réf. : Wonnacott et Wonnacott (1992, p. 142-148)

La loi normale est un exemple de distribution de probabilité d'une variable aléatoire continue. C'est une distribution dont la fonction de densité a la forme d'une cloche symétrique. Cette distribution est une bonne approximation de plusieurs distributions de probabilité observées

[14] Si l'on veut faire un parallèle entre les deux formules, on peut dire que $\int$ joue le rôle de Σ et $f(x)$, celui de $(1/n)$.

empiriquement. C'est aussi la distribution asymptotique vers laquelle tendent plusieurs autres distributions (à propos de distribution asymptotique, voir 3.2).

L'une des caractéristiques les plus importantes de la distribution normale est qu'elle ne comporte que deux paramètres : la moyenne et l'écart type. Cela signifie que si l'on sait qu'une variable a une distribution normale et qu'on connaît sa moyenne et son écart type, on connaît parfaitement sa fonction de densité de probabilité.

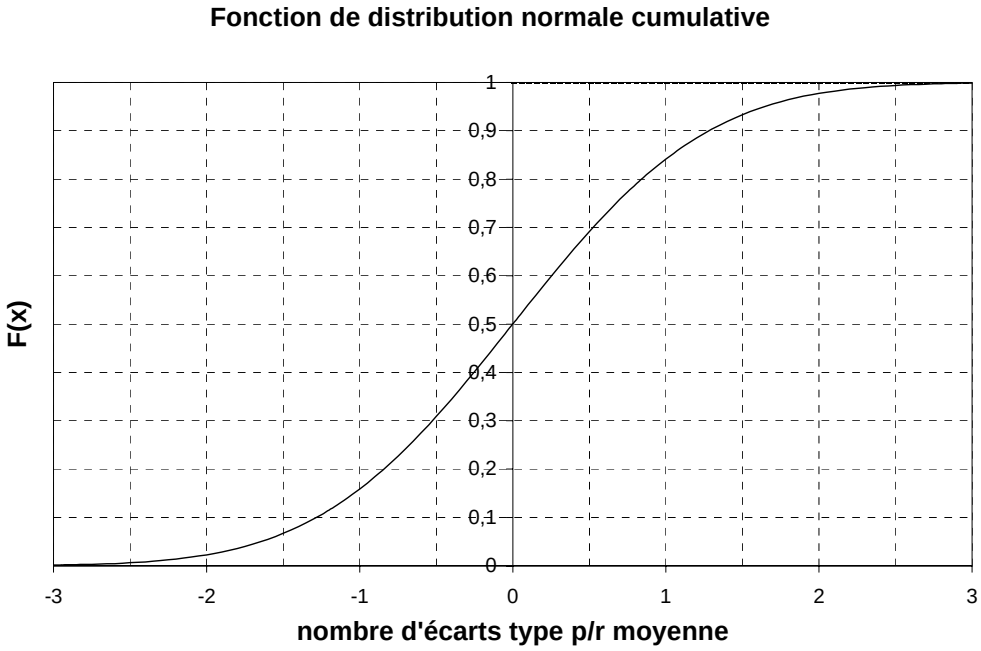
De plus, si la variable aléatoire *x* a une distribution normale, avec une moyenne μ_x et un écart type σ_x, alors la variable « standardisée »

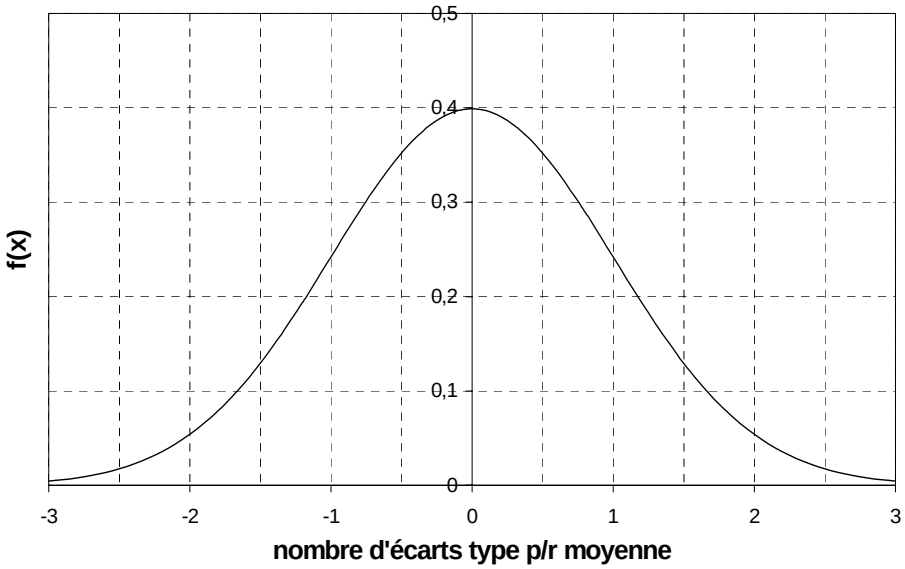
$$z = \frac{x - \mu_x}{\sigma_x}$$

a une distribution normale de moyenne 0 et d'écart type 1 [15].

[15] Les tables statistiques donnent la valeur de la densité de probabilité (ordonnée de la fonction de densité normale) et de la probabilité cumulée (ordonnée de la fonction de distribution cumulative normale) pour différentes valeurs de la variable normale standard. Elles donnent aussi, à l'inverse, la valeur de la variable normale standard qui correspond à différentes probabilités cumulées. Le logiciel Excel, fournit l'équivalent par le truchement des fonctions `NORMSDIST` et `NORMSINV` ; les fonctions correspondantes pour les variables normales non standard sont `NORMDIST`, `NORMINV`.

FIGURE 2 – DISTRIBUTION NORMALE





2-2.3 Échantillonnage, estimation et tests d'hypothèses

Grosso modo, la démarche de l'induction consiste à utiliser les statistiques relatives à un échantillon, pour apprendre quelque chose à propos des paramètres de la population.

Cette démarche pose trois ordres de questions :

- Quelles propriétés est-il souhaitable que l'*échantillon* possède ? Sous quelles conditions est-il raisonnable de considérer celui-ci comme « représentatif » ?
- Quelles sont les statistiques de l'échantillon qui peuvent être utilisées pour *estimer* la valeur des paramètres de la population ? Quelles sont les propriétés de ces estimateurs ?
- Comment peut-on évaluer la fiabilité des estimations obtenues ? Plus généralement, que peut-on affirmer au sujet de la population, et avec quel degré de confiance ?

Par rapport à ce troisième ordre de questions, les principes épistémologiques sous-jacents à l'induction statistique conduisent aux *tests d'hypothèses*.

On peut donc diviser l'induction statistique en trois parties :

- l'échantillonnage
- l'estimation
- les tests d'hypothèses

2-2.3.1 ÉCHANTILLONNAGE

Réf. : Wonnacott et Wonnacott (1992, chap. 23)

Un plan d'échantillonnage doit répondre aux questions suivantes :

1. Comment sélectionner l'échantillon de façon à respecter les conditions requises par le modèle d'échantillonnage (*sampling model*, modèle du lien entre l'échantillon et la population) ? Ou inversement, quel modèle d'échantillonnage reflète correctement le mode de sélection de l'échantillon ?
2. Quelle taille d'échantillon est nécessaire pour obtenir la précision et le niveau de confiance désirés ?

Un préalable à tout plan d'échantillonnage est le choix de l'unité d'observation. Par exemple, dans une enquête auprès des consommateurs, l'unité d'observation peut être la personne ou le ménage. L'ensemble des unités d'observation constitue la population dont on veut tirer un échantillon. Là encore, il faut que la population soit clairement définie. Par exemple dans une

enquête auprès des ménages, il faut évidemment circonscrire l'univers d'échantillonnage au moyen de limites géographiques ou autres. Mais il faut aussi délimiter l'univers conceptuellement : par exemple, dans une enquête auprès des ménages, veut-on prendre en considération les ménages d'une seule personne ? les ménages collectifs (prisons, communautés religieuses, casernes militaires) ?

(1) Sélection

La première distinction à faire dans les méthodes de sélection est entre celles qui conduisent à un échantillon *aléatoire* et les autres. Un échantillon *aléatoire* est un échantillon constitué selon une méthode de sélection qui permet de savoir, pour chacun des échantillons possibles, quelle est la probabilité qu'il soit sélectionné ; la plupart du temps, cela revient à connaître pour chaque individu la probabilité qu'il soit sélectionné, les probabilités des individus étant indépendantes les unes des autres (la probabilité qu'un individu soit sélectionné n'est pas affectée par le fait que tel ou tel autre individu le soit).

Mentionnons les trois grands types d'échantillons aléatoires.

Échantillon ***aléatoire simple*** (obtenu par tirage au sort) : chaque individu de la population a une chance égale d'être choisi. Cette méthode d'échantillonnage exige généralement un inventaire préalable de la population. Un inventaire incomplet comporte des risques de biais; par exemple, une sélection aléatoire dans le bottin téléphonique écarte *a priori* ceux qui n'ont pas le téléphone ou dont le numéro est confidentiel. Il y aura biais si ceux qui sont exclus sont différents des autres. Il y a d'autres risques de biais lors de la collecte. Par exemple, ceux que les politicologues ont pris l'habitude d'appeler les « électeurs discrets » (qui refusent de répondre, ou « ne savent pas ») ont peut-être en moyenne une opinion différente de ceux qui s'expriment plus volontiers.

On distingue l'échantillonnage aléatoire simple avec, et sans remplacement : dans le premier cas, les individus membres de l'échantillon sont tirés successivement et après chaque tirage, l'individu choisi redevient éligible aux tirages suivants (ainsi, un même individu peut être sélectionné plus d'une fois dans le même échantillon; il y figure alors en plusieurs exemplaires) ; dans le second cas, un individu qui a été choisi est retiré de la population d'où l'on tirera les autres membres de l'échantillon.

L'échantillonnage ***systématique*** est une méthode qui se rapproche de l'échantillonnage aléatoire simple [16]. Il consiste à sélectionner un individu à tous les *n* cas (ce qui suppose que les cas soient déjà rangés dans un certain ordre : par exemple, l'ordre alphabétique dans le bottin téléphonique ou l'ordre des numéros civiques le long d'une rue). Pour ce faire, on divise la taille de la population par la taille désirée de l'échantillon : c'est l'intervalle d'échantillonnage *n*. Il suffit alors de tirer au hasard le premier individu et le choix des autres s'ensuit.

Échantillon ***aléatoire stratifié*** : quand la population se divise en plusieurs sous-populations (« strates ») dont les paramètres peuvent être différents, on voudra que l'échantillon soit représentatif, non seulement de la population en général, mais aussi de chaque sous-population.

Cette représentativité n'implique pas nécessairement que chaque strate de l'échantillon soit proportionnelle à la sous-population qu'elle représente. En fait, si l'on recherche une égale précision pour toutes les strates de la population, il faut que les sous-populations moins nombreuses soient sur-représentées. Cela est dû au fait que la précision dans l'estimation des paramètres n'est pas proportionnelle à la taille de l'échantillon. Nous y reviendrons (voir chapitre 2-3, la fin de la section 2-3.5).

Si l'échantillon aléatoire simple exige un inventaire préalable de la population, l'échantillon aléatroire stratifié exige un inventaire par strate. Cela n'est pas toujours disponible. On tente souvent d'approximer un échantillon aléatoire stratifié au moyen de l'échantillonnage ***par quotas*** [17]. Cette méthode consiste à classer les individus au moment de leur sélection, jusqu'à ce que l'on ait atteint le nombre désiré (quota) dans chaque strate. Dans une enquête par questionnaire basée sur l'échantillonnage par quotas, la première partie du questionnaire sert à classer les individus ; on ne complète pas le questionnaire avec les surnuméraires.

Échantillon par ***grappes*** (« cluster ») : cette méthode consiste à diviser la population en groupes (grappes), puis à tirer au sort un certain nombre de grappes ; les membres des grappes choisies constituent l'échantillon. On recourt souvent à cette méthode faute d'inventaire préalable de la population.

Par exemple, pour faire une enquête auprès des ménages d'un quartier d'habitat informel (où les données du recensement ne sont pas fiables), on peut subdiviser le quartier en pâtés de

---

16 May (1993), p. 70.
17 May (1993), p. 71.

maisons, choisir une certaine proportion des pâtés de maison, et interviewer tous les ménages qui habitent à l'intérieur des pâtés sélectionnés. L'induction statistique est plus difficile à pratiquer avec un échantillon par grappes, parce que les distributions d'échantillonnage des statistiques ont plus complexes.

Il y a des méthodes d'échantillonnage **non** aléatoire qui sont néanmoins appropriées dans des contextes non statistiques. Ainsi, les enquêtes de type qualitatif reposent parfois sur la méthode « boule de neige » ou sur la méthode de saturation. Ces méthodes d'échantillonnage ne sont toutefois pas pertinentes ici.

(2) Taille

En général, plus la taille de l'échantillon est grande, plus celui-ci a de bonnes chances d'être représentatif, et plus la précision de l'estimation est grande pour un même degré de confiance. Mais le degré de précision n'est pas directement proportionnel à la taille de l'échantillon (nous verrons cela plus clairement lorsque nous examinerons un test d'hypothèse sur la moyenne ; voir 2-3.5). Selon les analyses qu'on prévoit vouloir faire, il y a des règles qui permettent de déterminer la taille d'échantillon requise pour atteindre la précision et le niveau de confiance désirés. Mais les coûts de cueillette croissent avec la taille de l'échantillon...

2-2.3.2 ESTIMATION

- Quelles sont les statistiques de l'échantillon qui peuvent être utilisées pour *estimer* la valeur des paramètres de la population ? Quelles sont les propriétés de ces estimateurs ?

Nous faisons ici une distinction entre un *estimateur*, qui est une formule, une méthode de calcul, et une *estimation* ou *valeur estimée*, qui est une valeur obtenue comme résultat de l'application de cette formule. Un estimateur est une variable aléatoire : un même estimateur, appliqué à des données issues d'échantillons différents, conduit généralement à des valeurs estimées différentes.

(1) Méthodes

Réf. : Wonnacott et Wonnacott (1992, chap. 18)

Trois approches :

1. Analogique
2. Moindres carrés

3. Maximum de vraisemblance (Theil, 1971, p.89 ; Freund, 1962, p. 223)

1. Estimation selon l'approche analogique

Le principe de l'estimation analogique (aussi appelé méthode des moments) est simple : pour estimer un paramètre, on applique à l'échantillon la même formule mathématique qu'à la population.

Exemple :

Pour estimer la valeur moyenne μ_X d'une variable x dans une population, on prend la moyenne de la même variable dans l'échantillon :

$$m_X = \frac{1}{n}\sum_i x_i$$

Ce procédé est purement mécanique. En général cependant, un estimateur est l'expression mathématique d'un principe de choix de la « meilleure » valeur comme estimation du paramètre. Différents principes de choix conduisent à différents estimateurs : les principes les plus répandus sont le principe des moindres carrés et celui du maximum de vraisemblance.

On pourrait comparer l'estimation à l'action de syntoniser un poste de radio : on essaie différentes fréquences jusqu'à ce que le signal reçu soit le meilleur possible ; à la fin, la fréquence sélectionnée sur le récepteur est une valeur estimée du paramètre recherché, qui est la fréquence d'émission. La fréquence sélectionnée dépendra du critère de choix utilisé (supposons aux fins de la comparaison que l'on n'applique qu'un seul critère à la fois) : force du signal, absence de chuintement et de distorsion, absence de parasites... [18].

2. Principe des moindres carrés

Le principe des moindres carrés peut s'appliquer sans modèle aléatoire. Il consiste à « syntoniser » les valeurs estimées des paramètres du modèle de façon à ce que, lorsque ce modèle est appliqué à l'échantillon, ses erreurs de prédiction soient aussi petites que possible. L'expression « moindres carrés » se réfère à la mesure d'erreur utilisée : c'est la *somme des carrés des erreurs* de prédiction, chaque erreur étant mesurée par l'écart entre une valeur

---

[18] Il ne faudrait pas pousser trop loin cette analogie : alors que la syntonisation d'un poste de radio dont on ne connaît pas la fréquence se fait le plus souvent par tâtonnement, l'application de l'un ou l'autre des principes d'estimation conduit la plupart du temps à une formule qui permet de calculer directement la valeur estimée correspondante.

observée et la valeur prédite correspondante. Cette mesure d'erreur est donc le carré de la distance euclidienne généralisée entre la série des observations et la série des prédictions.

3. Principe du maximum de vraisemblance

L'application du principe du maximum de vraisemblance fait directement appel au modèle aléatoire choisi pour représenter le lien aléatoire entre l'échantillon et la population ou pour représenter le caractère aléatoire du phénomène étudié. Donc, contrairement au principe des moindres carrés, celui du maximum de vraisemblance ne peut pas s'appliquer sans modèle aléatoire. Le principe du maximum de vraisemblance consiste à « syntoniser » les valeurs estimées des paramètres du modèle de façon à ce que, dans l'hypothèse où ces valeurs seraient les bonnes, l'échantillon soit le plus « vraisemblable » possible. On mesure la vraisemblance au moyen de la *fonction de vraisemblance*, qui est la fonction de densité de probabilité de l'échantillon étant donné les valeurs des paramètres.

Lorsqu'on maximise la fonction de vraisemblance, le rôle des valeurs observées de l'échantillon et celui des paramètres sont inversés par rapport à ce qu'ils sont dans la fonction densité de probabilité : au lieu que les valeurs observées soient considérées comme des variables aléatoires dont la fonction de densité de probabilité dépend de la valeur des paramètres, ce sont les valeurs observées qui sont considérées comme fixes et l'on fait varier les valeurs estimées des paramètres de façon à ce que la vraisemblance atteigne son maximum. Les valeurs choisies comme valeurs estimées des paramètres sont donc celles pour lesquelles la densité de probabilité de l'échantillon est la plus grande (le mode de la distribution) ; par conséquent, les intervalles autour de ce point sont ceux qui ont la plus grande probabilité.

Sous certaines conditions, le principe des moindres carrés et celui du maximum de vraisemblance conduisent au même estimateur. Dans certains cas (comme l'estimation de la moyenne), cet estimateur est en même temps celui de l'approche analogique.

(2) Propriétés désirables

Réf. : Wonnacott et Wonnacott (1992, p. 262-266, 275-276) ; Freund (1962, p. 215-220).

Les estimateurs sont des variables aléatoires. Par conséquent, leurs propriétés sont les propriétés de leur distribution d'échantillonnage.

1. Absence de biais

Parmi les propriétés désirables d'un estimateur, l'absence de biais est particulièrement importante. Un estimateur *non biaisé* est un estimateur dont la valeur sera en moyenne égale à la valeur du paramètre estimé. L'expression « en moyenne » renvoie ici à la distribution d'échantillonnage de l'estimateur.

Par exemple, si l'on veut estimer la variance d'une variable dans la population à l'aide des données d'un échantillon, et si on applique la formule de la méthode analogique

$$\frac{1}{n}\sum_i (x_i - m_x)^2$$

on peut démontrer que l'on obtient un estimateur biaisé : si on répétait le calcul avec un très grand nombre d'échantillons (une infinité), le résultat serait en moyenne différent de la vraie variance. C'est pourquoi on utilise plutôt un estimateur corrigé de façon à éliminer le biais ; cet estimateur non biaisé est donné par

$$s_x^2 = \frac{1}{n-1}\sum_i (x_i - m_x)^2$$

De même, l'estimation de la covariance au moyen de la formule de la méthode analogique

$$\frac{1}{n}\sum_i (x_i - m_x)(y_i - m_y)$$

donne un estimateur biaisé de la covariance entre *x* et *y* dans la population, tandis que

$$s_{xy} = \frac{1}{n-1}\sum_i (x_i - m_x)(y_i - m_y)$$

est un estimateur non biaisé.

2. Efficacité relative : les estimateurs « best unbiased »

Dans l'univers des échantillons possibles, tout estimateur non biaisé donne des résultats qui sont en moyenne centrés sur la « cible » que constitue la valeur du paramètre que l'on cherche à estimer. Comment choisir alors entre deux estimateurs non biaisés ? On choisira évidemment un estimateur dont le « tir » est groupé, de préférence à un estimateur qui donne des résultats largement dispersés autour de la cible.

C'est la variance qui mesure la dispersion d'une variable aléatoire autour de sa moyenne. On appelle « variance d'échantillonnage » la variance d'un estimateur dans la population des

échantillons possibles (c'est la variance de la distribution d'échantillonnage) ; la racine carrée de la variance d'échantillonnage est l'« erreur d'échantillonnage » (*sampling error*). On dit qu'un estimateur non biaisé est plus *efficace* qu'un autre si sa variance d'échantillonnage est inférieure à celle de l'autre.

On appelle « best unbiased » un estimateur non biaisé dont l'efficacité relative est supérieure à celle de tout autre estimateur non biaisé. On utilise aussi cette appellation dans un sens plus restrictif pour une classe donnée d'estimateurs. Par exemple, dans la classe des estimateurs dont la valeur est une fonction linéaire des données, celui qui a la meilleure efficacité relative est qualifié de « Best Linear Unbiased Estimate », ou « BLUE ».

3. Convergence

Une autre propriété désirable d'un estimateur est que sa précision soit supérieure lorsque l'on a un échantillon de plus grande taille ou, autrement dit, que sa variance d'échantillonnage soit plus petite quand l'échantillon est plus grand. On dit qu'un estimateur est *convergent* si sa variance d'échantillonnage tend vers zéro lorsque la taille de l'échantillon tend vers l'infini (la distribution d'échantillonnage tend à se concentrer sur un seul point).

4. Suffisance

Enfin, un estimateur est *suffisant* s'il incorpore toute l'information contenue dans l'échantillon à propos du paramètre à estimer : une fois qu'on a calculé la valeur de l'estimateur (à partir des données de l'échantillon), on n'apprendra rien de plus sur la valeur du paramètre en examinant de nouveau les données de l'échantillon.

Techniquement, cette propriété se traduit de la façon suivante : si un estimateur est suffisant, alors la probabilité de l'échantillon (sa vraisemblance) étant donné la valeur estimée est indépendante de la valeur du paramètre [19].

2-2.3.3 LA LOGIQUE FONDAMENTALE DES TESTS D'HYPOTHÈSE

Réf. : Blalock (1972), chap. 8, « The fallacy of affirming the consequent ».

Revenons un instant au schéma de la méthode scientifique discuté en 2-2.1. La logique fondamentale de cette démarche est la suivante :

---

19 Freund (1962, p. 19-20).

- Si une théorie (ou un modèle, ou une hypothèse) est vraie, alors ses implications doivent être vraies aussi.
- Donc, si les observations contredisent les implications d'une théorie, cette théorie ne peut pas être vraie : elle est fausse.

Ce qu'il est crucial de comprendre dans ce raisonnement, c'est que si les observations ne contredisent pas les implications d'une théorie, on n'a pas le droit de conclure que cette théorie est vraie ! Plus exactement, pour que l'on puisse conclure que cette théorie est vraie, il faudrait qu'il n'y ait pas d'autre théorie possible qui soit compatible avec les observations. En pratique, cette condition est tellement exigeante qu'elle n'est jamais remplie.

Idée-clé No 12 :

Logique de rejet/non-rejet : « si les observations ne contredisent pas les implications d'une théorie, on n'a pas le droit de conclure que cette théorie est vraie ! » (p. 2-2.27); elle demeure « acceptable », mais elle n'est pas « acceptée ».

En somme, lorsque les implications d'une théorie sont confrontées aux observations, la théorie est rejetée si les observations en contredisent les implications ; si les observations n'en contredisent pas les implications, la théorie n'est pas pour autant confirmée, elle est seulement « non rejetée » [20].

C'est cette même logique du rejet/non rejet qui prévaut dans les tests d'hypothèse. Mais, il y a une différence capitale : dans les tests d'hypothèse, le lien entre les hypothèses et l'échantillon observé est aléatoire. Il s'ensuit que le raisonnement ne peut plus être déterministe, il doit être probabiliste.

Dans une logique déterministe, une observation est compatible avec l'hypothèse ou elle ne l'est pas : il n'y a pas d'entre-deux. Dans une logique probabiliste, une observation est *plus ou moins compatible* avec l'hypothèse : plus une observation est improbable lorsqu'on suppose que l'hypothèse est vraie, moins elle est compatible avec cette hypothèse. Un exemple quelque peu caricatural illustrera ce qui précède :

---

[20] Personnellement, je préfère l'expression « non rejetée » au mot « acceptable » employé par Wonnacott et Wonnacott (1992), à cause du risque de glisser d'« acceptable » à « accepté », qui n'est pas la même chose ! On aura aussi reconnu ici une parenté avec le falsificationnisme popperien : pour Popper, une hypothèse qu'il est logiquement ou empiriquement impossible de rejeter n'est pas « scientifique ».

> Considérons l'hypothèse que le dromadaire ne fait pas partie de la faune sauvage du continent australien. Supposons qu'un voyageur, à sa grande surprise, croise un dromadaire sans maître dans le désert australien. Cette observation est contraire à son hypothèse. Mais ce dromadaire pourrait s'être échappé d'un cirque ou d'un zoo, ou même ce pourrait être un mirage. L'observation d'un dromadaire n'est pas impossible, elle est seulement improbable : l'observation d'un seul dromadaire, ou même de quelques-uns, ne serait sans doute pas considérée comme incompatible avec l'hypothèse. Mais supposons que le même voyageur repère des dromadaires à plusieurs occasions. Si l'hypothèse était vraie, ces observations répétées seraient extraordinairement improbables. À la longue, l'observateur finira par se dire que ses observations ne sont pas compatibles avec son hypothèse [21].

Dans l'exemple qui précède, notre voyageur se contentera sans doute d'une approche intuitive. S'agissant de tests d'hypothèse, il va sans dire que la démarche est davantage formalisée : en particulier,

1. les probabilités sur lesquelles portent le raisonnement doivent être quantifiées (S'il est vrai que le dromadaire ne fait pas partie de la faune australienne, quelle est la probabilité exacte de néanmoins rencontrer un dromadaire ? Deux ? Trois ? Etc.) ;
2. la décision de rejeter l'hypothèse doit se prendre en fonction d'un critère précis, défini d'avance (Quelle est la probabilité en-deçà de laquelle on décidera que les observations sont incompatibles avec l'hypothèse ?)

C'est la première de ces deux exigences qui, de loin, pose les plus grandes difficultés, tant conceptuelles que pratiques. Nous verrons que la seconde n'est en fin de compte rien d'autre qu'une exigence de transparence.

Idée-clé No 13 :

En logique probabiliste, une observation est *plus ou moins compatible* avec l'hypothèse.

---

21 Les dromadaires sauvages font partie de la faune des déserts australiens depuis qu'ils ont été abandonnés par les caravaniers afghans qui les avaient importés pour assurer les liaisons trans-continentales avant la construction du chemin de fer.