Une histoire lacunaire

Anne Gégout-Petit, Myriam Maumy-Bertrand, Gilbert Saporta, Christine Thomas-Agnan

Chapitre 1

UNE HISTOIRE LACUNAIRE

Anne Gégout-Petit, Myriam Maumy-Bertrand, Gilbert Saporta et Christine Thomas-Agnan

Tout recueil de données conduit à des données manquantes mais la prise de conscience des problèmes soulevés n'est que très récente puisqu'elle ne date que du début du vingtième siècle. On a coutume de distinguer les non-réponses ou données totalement manquantes, des données manquantes partielles où seules certaines valeurs d'une ou plusieurs variables sont absentes.

La plupart du temps les données manquantes sont subies. Il existe cependant des cas où l'absence de données peut venir d'une volonté délibérée

Ce chapitre ne prétend nullement à l'exhaustivité, est-il besoin de le dire ? Il essaiera juste d'apporter quelques éclairages historiques dans un domaine bien vivant.

1.1 Les recensements dans les anciens empires

Les recensements sont les plus anciennes opérations statistiques connues destinées à compter hommes, bêtes, terres, production agricole, etc. Réalisés par des gouvernements autoritaires, leur caractère obligatoire a fait de la non-réponse un non-problème ! Le lecteur intéressé pourra se reporter à `https://brewminate.com/an-historical-overview-of-the-census-in-the-ancient-and-medieval-world/`.

1.1.1 Babylone, Chine, empire Inca

Le premier recensement connu a été effectué par les Babyloniens en 3800 avant J.C., il y a donc près de 6000 ans. Selon les archives, il était effectué tous les six ou sept ans et recensait le nombre de personnes et le bétail, ainsi que les quantités de beurre, de miel, de lait, de laine et de légumes.

En Chine, le premier recensement aurait eu lieu il y a 4 000 ans. Comme l'écrivent Cartier et Will [1971] « l'ancienne administration impériale ne cherchait pas à obtenir des renseignements démographiques — effectif de la population, statistiques vitales, répartition par sexes ou par âges — mais elle avait en vue l'établissement de rôles d'impôts, de conscription ou de corvées ». Le recensement le plus célèbre est celui de l'an 2 sous la dynastie des Han (206 av. J.C. à 220 apr. J.C) qui dénombra 12 233 062 feux et 59 594 978 bouches.

Dans l'empire inca chaque tribu avait son propre statisticien, appelé le Quipucamayoc représenté Figure 1.1. Cet homme tenait des registres sur, par exemple, le nombre de personnes, le nombre de maisons, le nombre de lamas, le nombre de mariages et le nombre de jeunes hommes qui pouvaient être recrutés pour l'armée. Tous ces faits étaient consignés sur un quipu, un système de noeuds dans des cordes colorées. (Bethlehem *et al.* [2011])

Figure 1.1 : *Le Quipucamayoc*

1.1.2 L'empire romain et la mort civile des non-répondants

Le mot recensement vient du latin *census* et du verbe *censere* , qui signifie estimer. Le recensement romain était le plus développé de tous ceux qui ont été enregistrés dans le monde antique et il a joué un rôle crucial dans l'administration de l'Empire romain.

Selon la tradition romaine, le recensement (voir Figure 1.2) a été établi par le roi Servius Tullius, le sixième roi légendaire qui aurait régné de 575 à 535 av. J.C., pour permettre le recrutement de l'armée romaine. La participation au recensement était obligatoire pour les citoyens. Pendant la royauté étrusque celui qui se soustrayait au recensement pouvait être mis à mort. Pendant la période républicaine, l'État pouvait vendre comme esclave le citoyen qui échappait volontairement à son devoir : c'était la « mort civile », voir Cicéron *ProCaecina*, 99.

Le recensement romain était effectué tous les cinq ans. Cette période était appelée lustre, mot qui désignait soit un sacrifice expiatoire qui avait lieu tous les cinq ans au moment du recensement, soit le recensement lui-même. Depuis 435 av. J.-C., il se déroulait dans la *Villa Publica*, un bâtiment spécialement aménagé au Champ de Mars représenté Figure 1.3.

Le recensement fournissait un registre des citoyens et de leurs biens dressée par des officiers publics nommés censeurs. Les censeurs étaient élus parmi les anciens consuls et disposaient d'un pouvoir absolu. Ils étaient également chargés de mettre à jour l'*album*, c'est-à-dire le registre des personnes admises au Sénat. Leur fonction les amène aussi à surveiller les mœurs : ils détiennent

la *cura morum* qui leur permet de rayer de l'album sénatorial les sénateurs indignes, mais aussi de flétrir publiquement la réputation d'une personne par la *nota censoria*. On retrouvera plus loin le terme de censure dans une acception bien différente.

En plus de servir de base au recrutement de l'armée romaine, le *census* servait à la délimitation des droits politiques, à l'organisation des scrutins, au calcul des impôts, puis à l'élaboration d'un état civil. à partir duquel les devoirs et privilèges pouvaient être répertoriés. Institution fondamentale de la République romaine, le recensement constitue le peuple romain (*populus*) en un corps civique organisé et hiérarchisé. À chaque citoyen est attribué un rang précisant sa dignité, ses droits et ses devoirs envers la cité. Ce rang dépend principalement du patrimoine foncier.

Sources : Nicolet [1985], et `https://fr.vikidia.org/wiki/Recensement_dans_la_Rome_antique`.

Figure 1.2 : *Scène de recensement, dit autel de Domitius Ahenobarbus, le Louvre*

Figure 1.3 : *La Villa Publica*

1.2 Données absentes et pondération

Comme on le verra au chapitre 9 le traitement des non-réponses, ou observations manquantes totalement, s'effectue généralement en pondérant les données observées. La pratique est courante dans les enquêtes par sondage mais on la trouve dans d'autres domaines.

1.2.1 R.A. Fisher et les albinos

C.R. Rao (1920-) dans deux publications ([Rao, 1965a] et [Rao, 1985]) fait remonter l'usage des distributions pondérées à Fisher [1934]. R.A. Fisher (1890-1962) écrivait à peu près ceci : « Supposons que, si la majorité des parents sont incapables d'engendrer des enfants albinos, un certain nombre de couples, qui ne sont pas eux-mêmes albinos, sont tels qu'une proportion fixe p des enfants qu'ils engendrent sont des albinos. Comment estimer p ? Considérons un grand nombre de familles choisies au hasard, le nombre moyen $\bar{R}$ d'enfants albinos divisé par la taille de la fratrie s ne donne pas une estimation sans biais de p car le zéro est ambigü : on ne peut pas distinguer les familles génétiquement capables d'avoir des enfants albinos, mais qui n'en ont pas eu, de celles de parents incapables d'en avoir. »

Fisher montre alors que :

$$\mathbb{E}\left[\frac{\bar{R}}{s}\right] = \frac{p}{1-q^s} \tag{1.1}$$

où $q{=}1-p$. La résolution en p de cette équation d'estimation fournit l'estimateur du maximum de vraisemblance.

On retrouvera plus loin des problèmes similaires quand la valeur zéro n'est pas observable.

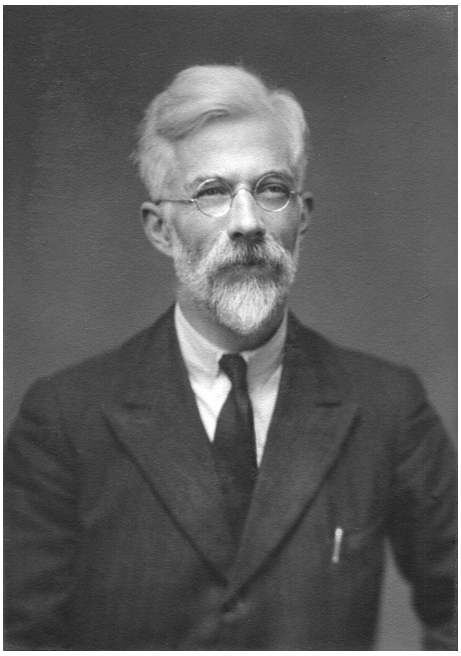

Figure 1.4 : *Ronald Aylmer Fisher vers 1943*

1.2.2 C.R. Rao et les petits crânes

Le doctorat de Rao a commencé par une lettre de J.C. Trevor du musée d'archéologie et anthropologie de l'Université de Cambridge, qui en 1946 a demandé l'aide de l'Institut indien de statistique pour analyser les mesures des squelettes humains obtenus à partir du Jebel Moya au Soudan. Rao était parmi les deux personnes dépêchées par P.C. Mahalanobis pour ce travail. Il y a travaillé de 1946 à 1948, et a fait simultanément son doctorat sous la direction de Ronald Fisher. Sa thèse s'intitulait *Statistical problems of biological classification.*

À l'occasion du centenaire de l'Institut International de Statistique, Rao [1985] est revenu sur les résultats de cette étude archéologique [Mukherjee *et al.*, 1955] portant sur un échantillon de crânes. Certains de ces crânes étaient en bon état et d'autres non. Chaque crâne en bon état était décrit par 4 variables C (capacité), L (longueur), B (largeur), et H (hauteur) alors que sur un crâne fracturé, C ne pouvait pas être mesuré. Un problème était d'estimer la moyenne de C dans la population originale de crânes provenant des échantillons fragmentaires.

Dans un certain nombre d'articles anciens, il était d'usage d'estimer la valeur moyenne en prenant la moyenne des valeurs observées. Une alternative, souvent recommandée, consistait à calculer les estimations du maximum de vraisemblance sur toutes les données disponibles en supposant une distribution normale à quatre variables pour C, L, B et H et en utilisant la distribution marginale pour les mesures incomplètes. Cela reposait sur l'hypothèse que chaque crâne ayant les quatre mesures complètes, ou tout sous-ensemble des quatre, peut être considéré comme un échantillon aléatoire de la population originale des crânes. Cette hypothèse était-elle valable, ce qui renvoie à la question posée dans le titre de l'article *What Population Does a Sample Represent ?* .

Or il y avait des éléments prouvant que les plus petits crânes avaient plus de chances de rester intacts, d'où une sur-représentation des petits crânes parmi les crânes non brisés, conduisant à une sous-estimation de la capacité moyenne de la population d'origine.

Notons $w(c)$ la probabilité qu'un crâne de capacité c soit préservé, et $p(c)$ la densité de probabilité de la variable C dans la population totale. Alors la densité de probabilité de C dans la population observée vaut :

$$\frac{w\left(c\right)p\left(c\right)}{\mathbb{E}\left(w\left(C\right)\right)} \tag{1.2}$$

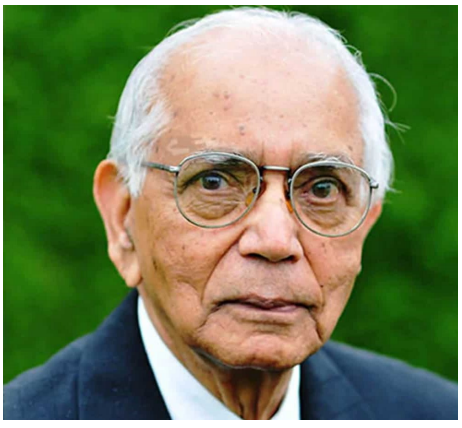

Figure 1.5 : *C.R. Rao en 2020*

1.2.3 A.Wald et la vulnérabilité des avions

Pendant la deuxième guerre mondiale, Abraham Wald (1902-1950) travaillait au *Statistical Research Group* de l'université Columbia. On lui demanda où il fallait renforcer le blindage des bombardiers en se basant sur les impacts de balle constatés au retour de mission. Les plus nombreux étaient sur le fuselage, et plus rarement sur les moteurs.

Figure 1.6 : *Abraham Wald*

Wald recommanda de déterminer avec soin où les avions de retour avaient été touchés et à l'étonnement des militaires, de blinder partout ailleurs et surtout les moteurs ! En effet les avions qui étaient revenus étaient ceux qui n'avaient pas été touchés au moteur, ceux touchés au moteur avaient été abattus et étaient manquants. Les données disponibles souffraient d'un cas particulier de biais de sélection appelé biais du survivant. On trouvera des détails et commentaires dans les références suivantes : Mangel et Samaniego [1984], Wainer [2011] et Ellenberg [2014] dont Ellenberg [2018] est une traduction française.

1.3 Données et distributions tronquées

Les données tronquées (il vaut mieux parler de distributions tronquées) ne doivent pas être confondues avec des données censurées qui seront évoquées dans la partie 1.5. Dans le cas d'une troncature droite en a la variable d'intérêt X n'est pas observable si $X > a$: les données peuvent être non enregistrées comme dans le cas des trotteurs de Galton trop lents ou être non détectables. Une loi tronquée est une loi conditionnelle, celle de $X|X < a$. C'est un cas particulier de distribution pondérée comme dans la partie 1.2.2 avec une fonction de pondération $w(c) = 1$ si $c < a$ et $w(c) = 0$ si $c > a$.

Dans une loi tronquée l'événement $X > a$ est inobservable : on n'en connait parfois même pas la fréquence ou la probabilité. Sachant que l'on n'observe que les données inférieures à a, comment en déduire la loi de X ?

1.3.1 Galton et les trotteurs tronqués

On admet communément que Francis Galton [1898] fut le premier à considérer une distribution tronquée, en l'occurrence une loi normale tronquée à droite, en analysant des données extraites du *Wallace's Year Book*, Vols. 8-12 (1892-1896), une publication de l'*American Trotting Association*. Voir l'introduction historique de Cohen [1991]. Les données consistaient en des temps de parcours, en vue d'une qualification, de trotteurs devant parcourir un mile en 2 minutes et 30 secondes au maximum. Aucune trace n'était gardée des temps des trotteurs les plus lents, donc éliminés, dont le nombre est resté inconnu. Galton se contenta d'une inspection visuelle du polygone de fréquences pour repérer le mode qui lui servit d'estimation de la moyenne. Il repéra de même les quartiles pour estimer la dispersion à l'aide de l'intervalle interquartile. Peu satisfait par cette méthode, Karl Pearson [1902] reprit les données de Galton avec une méthode plus complexe ajustant des paraboles au diagramme des logarithmes des fréquences. Mais les résultats s'écartèrent peu de ceux de Galton. Fisher reprendra le problème en 1931 en utilisant la méthode du maximum de vraisemblance qu'il avait introduite seulement 10 ans auparavant.

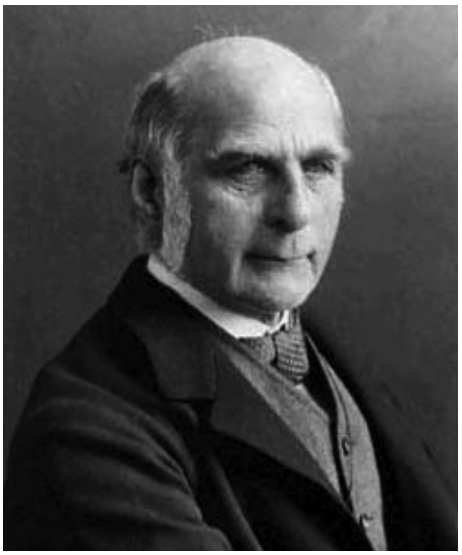

Figure 1.7 : *Francis Galton 1822-1911*

Pour en savoir plus sur la vie de Francis Galton qui fut explorateur, géographe, météorologue entre autres et qui était cousin de Darwin, on se reportera au chapitre historique de Droesbeke et Saporta [2011]. On retrouvera Galton plus loin.

1.3.2 Poisson sans le zéro

On doit à Anderson Gray McKendrick [1926] dans son exemple n°2 la première utilisation d'un modèle de Poisson tronqué en zéro. McKendrick avait été confronté au problème suivant : en étudiant la distribution du nombre de foyers où 0, 1, 2, 3, 4 cas de choléra avaient été observés dans un village indien, il constata un nombre trop élevé de zéros pour une distribution de Poisson.

Il attribua cela au fait que l'on avait dû mélanger des données provenant de foyers proches d'un puits pollué avec des foyers éloignés et donc non contaminés. Le véritable nombre d'observations de foyers touchés n était donc inconnu tout comme n_0.

Table 1.1 : *Distribution des foyers selon le nombre de victimes du choléra*

x	n_x
0	168
1	32
2	16
3	6
4	1

McKendrick décida d'ignorer la première ligne de la Table 1.1 et recourut à une méthode astucieuse pour estimer n puis le paramétre λ de la loi de Poisson. Les moments empiriques de X ne font pas intervenir la valeur zéro, et on sait que le moment factoriel $\mathbb{E}\left[X\left(X-1\right)\right]=\lambda^2$, d'où

$$\mathbb{E}\left[\sum_{x=1}^{\infty} x n_x\right]=n\lambda$$

et

$$\mathbb{E}\left(\sum_{x=1}^{\infty} x\left(x-1\right) n_x\right)=n\lambda^2$$

Le quotient

$$\hat{n}=\frac{\left(\sum\limits_{x=1}^{\infty} x n_x\right)^2}{\sum\limits_{x=1}^{\infty} x\left(x-1\right) n_x}$$

estime donc (mais pas sans biais)

$$\frac{\left(n\lambda\right)^2}{n\lambda^2}=n$$

On déduit ensuite les estimations $\hat{n}_0$ et $\hat{\lambda}$. McKendrick trouve ainsi $\hat{n}=93$, $\hat{\lambda}=0.92$ et $\hat{n}_0=37$.

J. Gani [2001] nous éclaire sur la biographie de McKendrick (1876-1943) : diplômé en médecine de l'université de Glasgow en 1900, il rejoint l'*Indian Medical Service* de l'armée des Indes qu'il quitte en 1920 pour raisons de santé avec le grade de lieutenant-colonel. Il travailla avec les Instituts Pasteur de Coonoor et Jausali dont il fut même directeur. McKendrick fut un des pionniers de l'épidémiologie mathématique et le co-inventeur du célèbre modèle SIR de diffusion des épidémies.

Plus de 30 ans après, dans une courte note Joseph Oscar Irwin [1959] proposa d'itérer la méthode de McKendrick : $\hat{\lambda}$ permet d'obtenir une estimation de n_0, que l'on peut alors utiliser pour réestimer n et λ et ainsi de suite. La méthode d'Irwin converge vers l'estimateur du maximum de vraisemblance pour des lois tronquées que H.O. Hartley [1958] venait de publier. Comme l'indique Xiao-Li Meng [1997] on a ainsi les prémisses de l'algorithme EM qui sera publié près de 20 ans après [Dempster *et al.*, 1977] .

Irwin (1898-1982) joua un rôle important dans le développement de la statistique médicale au Royaume-Uni. Il rendit un hommage appuyé à McKendrick dans son discours présidentiel à la *Royal Statistical Society* (Irwin [1963]. Dans sa note biographique Peter Armitage [2001] signale qu'Irwin « fut l'un des rares statisticiens à travailler à la fois avec Karl Pearson et Fisher, et il a été capable d'entretenir des relations cordiales avec ces deux personnalités fortes mais disparates ».

1.4 Données manquantes partielles, observations et expériences planifiées

1.4.1 Galton et les *pairwise* corrélations

Dans ses travaux anthropométriques Francis Galton [1889] a rencontré des cas de mesures incomplètes concernant certaines variables ce qui ne le gêna pas particulièrement pour calculer les coefficients de corrélation linéaire entre les paires de variables, utilisant à chaque fois les paires de mesures complètes. Son étude était basée sur les données de 350 hommes, mais comme écrivait Galton le nombre exact de 350 n'est pas conservé tout au long de l'étude, car une blessure à un membre ou à un autre a réduit le nombre disponible de 1, 2 ou 3 dans différents cas. Il calcula donc ainsi des corrélations selon la méthode *pairwise*, sans doute pour la première fois comme le remarque Stigler [1986].

Si la covariance se définit aisément en prenant les paires d'observations disponibles, la manière d'estimer les variances posa problème ultérieurement : fallait-il l'estimer avec les mêmes observations que pour la covariance, ou bien avec toutes les observations disponibles pour chaque variable au risque d'obtenir des corrélations en dehors de l'intervalle $[-1; 1]$?

Wilks [1932] étudia même le cas d'une estimation de variance combinant les deux méthodes pour conclure à la supériorité de l'estimateur *pairwise* du coefficient de corrélation qui a une variance inférieure. Il n'en reste pas moins qu'une matrice de corrélations *pairwise* peut ne pas être définie positive.

1.4.2 Estimer des expériences perdues

Les plans d'expériences et l'analyse de variance furent introduits par Fisher dans les années 1920 pour l'agronomie. L'objectif des plans est de rationaliser et optimiser des expériences longues et coûteuses. Le choix de plans orthogonaux en blocs randomisés ou en carrés latins permettait une simplification du calcul des effets des facteurs ne nécessitant que des différences de moyenne, ce qui était indispensable à une époque où les moyens de calcul étaient inexistants.

Encore fallait-il que toutes les expériences soient réalisées, ce qui n'était pas toujours le cas, en raison par exemple des intempéries. La question des réponses manquantes dans les plans d'expérience se posa donc rapidement. C'est un cas particulier d'une régression linéaire avec une variable réponse partiellement observée (voir la section 2.8.1).

Ce qui suit a été inspiré par l'ouvrage de Yadolah Dodge [1985]. Le chapitre 2 de Little et Rubin [2019] et la note bibliographique du chapitre 1 de Raghunathan [2015] fournissent également des informations utiles.

L'article d'Allan et Wishart [1930] semble être le premier à avoir traité du cas d'une seule observation manquante dans un plan d'expériences. Considérons un plan en p blocs et q traitements tel celui de la figure 1.8 représentant les rendements en pommes de terres de 12 parcelles réparties en trois blocs et soumises à quatre traitements.

	Treatments					
Block	1	2	3	4	Mean	
A.	139·0	219·0	200·5	145·0	175·88	
B.	197·5	205·0	206·0	182·5	197·75	
C.	156·0	229·5	210·0	245·5	210·25	
Mean	164·17	217·83	205·50	191·00	194·63	General mean

Figure 1.8 : *Rendement en livres de pommes de terre*

Si la réponse Y est manquante dans une case repérée par le couple $(i_0; j_0)$, Allan et Wishart utilisent le modèle suivant :

$$y_{ij} = b_i + t_j + k$$

où b_i est l'effet du bloc i, t_j l'effet du traitement j et k un terme constant. En posant $b_{i_0} = t_{j_0} = 0$ pour la case manquante repérée par le couple $(i_0; j_0)$, on voit que k est l'estimation du rendement manquant. En minimisant sur b, t et k la somme des $pq - 1$ carrés suivants :

$$\sum_i \sum_j \left(y_{ij} - b_i - t_j - k\right)^2$$

on trouve que :

$$k = \frac{(p + q - 1)\, S - qS_t - pS_b}{(p - 1)\,(q - 1)}$$

où S est la somme des rendements des $pq - 1$ parcelles non manquantes, S_t la somme des totaux des $q - 1$ traitements en excluant le traitement où l'observation est manquante et S_b la somme des totaux des $p - 1$ blocs en excluant le bloc où l'observation est manquante.

La méthode d'Allan et Wishart est une méthode d'imputation par régression. Dans l'analyse de variance, il faut diminuer le degré de liberté de la variance totale pour tenir compte de la donnée manquante. Allan et Wishart donnent aussi une formule pour les carrés latins.

Yates [1933] proposa le procédé simple suivant qui s'étend à un nombre quelconque (mais pas trop grand !) de valeurs manquantes : on ajuste un modèle linéaire sur les données complètes et on remplace les valeurs manquantes par leurs valeurs estimées. Comme l'écrit Dodge [1985] : « Yates a montré que les valeurs à insérer sont celles pour lesquelles, lorsque les données complétées sont analysées, les résidus dans les parcelles manquantes sont nuls. Le critère de Yates, qui consiste à estimer la ou les observations manquantes en minimisant la somme des carrés résiduels, devint LE critère recommandé dans presque tous les textes importants sur les plans d'expériences ».

Bose et Mahalanobis [1938] et Bose [1938] ont traité le cas où on connait la somme ou plus généralement une combinaison linéaire des valeurs manquantes. Ce cas leur a été inspiré par un incident survenu dans une expérimentation en Inde : « Les récoltes des parcelles individuelles étaient stockées dans de petits sacs étiquetés placés côte à côte. Deux de ces sacs ont été accidentellement endommagés et leur contenu s'est mélangé. La particularité de l'événement dans le cas présent est que, bien que les rendements des parcelles individuelles ne soient pas disponibles séparément, le rendement total des deux parcelles est connu ».

Il n'était pas simple à l'époque de minimiser le critère de Yates puisque l'orthogonalité était perdue. Il fallut attendre plus de 20 ans pour que l'algorithme itératif de Healy et Westmacott [1956] dont Yates prouva la convergence, permette une solution pratique. L'algorithme est ainsi décrit par Little et Rubin [2019] : Avec cette méthode, (i) des valeurs d'essai sont substituées à toutes les valeurs manquantes, (ii) l'analyse complète des données est effectuée, (iii) des valeurs prédites sont obtenues pour les valeurs manquantes, (iv) ces valeurs prédites sont substituées aux valeurs manquantes, (v) une nouvelle analyse complète des données est effectuée, et ainsi de suite, jusqu'à ce que les valeurs manquantes ne changent pas de manière appréciable ou, de manière équivalente, jusqu'à ce que la somme des carrés résiduels cesse essentiellement de diminuer. On retrouve ici encore un ancêtre de la méthode EM.

Frank Yates, longtemps resté dans l'ombre de Fisher, n'a pas eu droit à une notice dans des livres d'histoire de la statistique. Il mérite pourtant d'être mieux connu. Sa nécrologie de quelques 18 pages écrite par [Healy, 1995] cité plus haut fournit bien des informations. Recruté en 1931 par Fisher à Rothamsted comme assistant pour y remplacer J. Wishart, il devint en 1933 directeur du département de statistique quand Fisher fut nommé sur la chaire de génétique de l'*University College* de Londres. Yates y resta jusqu'à sa retraite en 1968. Il publia avec Fisher un

fameux recueil de tables ([Fisher et Yates, 1938]). En plus de ses importantes contributions à la théorie des plans d'expériences pour lesquels il est le plus connu, et de la correction de continuité pour les tables de contingence 2×2 qui porte son nom, Yates s'intéressa après la seconde guerre mondiale aux méthodes de sondage couvrant donc ainsi les deux principales méthodes statistiques d'obtention de données. Il fut un des 5 membres de la sous-commission des Nations Unies sur les méthodes d'échantillonnage avec Darmois, Deming, Fisher et Mahalanobis et rédigea à la suite de cette expérience un manuel à succès de plus de 300 pages plusieurs fois réédité ([Yates, 1949]). Les spécialistes des sondages connaissent bien la formule de Yates-Grundy pour estimer la variance de l'estimateur de Horwitz-Thompson.

Figure 1.9 : *Frank Yates 1902-1994*

1.4.3 L'essor du maximum de vraisemblance dans les années 50

L'article de Wilks [1932] déjà cité utilisait la méthode du maximum de vraisemblance pour estimer une matrice de variance covariance quand des données sont manquantes. Mais c'était pour deux variables seulement. Il fallut plus de 20 ans pour que Frederic Lord [1955] la généralise à 3 variables !

Rao [1956] établit des formules générales pour le Λ de Wilks dans le cadre de l'analyse de variance multidimensionnelle à k groupes et p variables dont une comporte des valeurs manquantes.

Hartley [1958] donne les estimateurs du maximum de vraisemblance et leurs variances pour des distributions discrètes tronquées et censurées. Il annonce *a second paper in which we shall also discuss the fit of a truncated Gamma distribution to rainfall data* qui semble ne jamais avoir été publié.

La portée pratique de ces avancées restait cependant limitée en raison de la faiblesse des moyens de calcul à l'époque. On remarque qu'aucun de ces auteurs ne se pose la question du mécanisme qui conduit à des données manquantes.

Si on ne présente pas C.R. Rao (il faudrait un livre entier), les deux autres protagonistes cités dans cette partie méritent d'être mieux connus.

Frederic M. Lord (1912-2000) docteur en psychologie de Princeton (1951) fit toute sa carrière à l'*Educational Testing Service* qui est la plus grande organisation à but non lucratif privée de mesure et d'évaluation éducative au monde et emploie 2700 personnes. Ledyard Tucker, Karl Jöreskog, Donald Rubin, Paul Holland, Howard Wainer y ont travaillé. Frederic Lord est considéré comme le père des tests modernes ; il est à l'origine de la plupart de ces célèbres tests : SAT, GRE, GMAT, LSAT et TOEFL. Le paradoxe de Lord, extension en continu du paradoxe de Simpson, est bien connu dans les études sur la causalité et a fait l'objet d'un article récent de Judea Pearl [2016].

Né Herman Otto Hirschfeld à Berlin, H.O. Hartley (1912-1980) immigra en Angleterre en 1934 pour fuir le régime nazi. Son père étant né en Angleterre, il eut la double nationalité allemande et britannique. Il anglicisa alors son nom. Avec Egon Pearson il publia les fameuses *Biometrika Tables for Statisticians*. Il s'installa aux USA en 1953 et eut une brillante carrière. Il fut président de l'*American Statistical Association* en 1979. Il a laissé son nom à un test d'égalité des variances et à la méthode d'échantillonnage à probabilités inégales RHC (Rao-Hartley-Cochran). Sous le nom de Hirschfeld il publia un court article ([Hirschfeld, 1935]) souvent cité comme l'une des sources de l'analyse des correspondances.

1.4.4 Imputation, *hot deck* etc.

Les méthodes précédentes s'appliquaient difficilement au cas des enquêtes et recensements où les données manquantes sont bien plus nombreuses. Ces méthodes mathématiques étaient d'ailleurs moins connues des praticiens qui ont développé des méthodes d'*imputation* sans modèle explicite, consistant à remplacer les données manquantes par des valeurs plausibles issues de données semblables. Eric Rancourt donne l'étymologie du mot imputation dans le premier numéro du *Bulletin d'imputation* publié par Statistique Canada, Rancourt [2001b] :

« Imputation et imputer viennent du verbe *imputare* qui possède plusieurs significations : la première est celle de porter au compte. Cette signification a été utilisée la première par Columelle, agronome sous les empereurs romains Tibère et Claude. Ensuite le mot a été utilisé par Tacite et Pline le jeune pour signifier attribuer quelque chose à quelqu'un ».

Voir également Rancourt [2001a] qui signale la première utilisation au sens actuel du mot imputation dans Hansen *et al.* [1953] page 546 pour les corrections faites dans l'enquête américaine *Survey of retail shares* de 1948.

On emploie souvent l'expression *hot deck* pour désigner la pratique de remplacer les données manquantes d'un receveur par celles d'un répondant appelé donneur. Cette procédure a été développée en 1947 par le *US Bureau of Census*.

Lavrakas [2008], pages 315-317, et Andridge et Little [2010] en donnent l'explication suivante : le terme *hot deck* vient de l'utilisation des cartes perforées pour le stockage des données, et fait référence à la pile de cartes des donneurs disponibles. Les donneurs et les receveurs provenant du même ensemble de données, la pile de cartes était « chaude »parce qu'elle était en cours de traitement : le passage rapide dans le lecteur chauffait les cartes perforées. L'imputation « *cold deck* », en revanche, sélectionne les donneurs à partir d'ensembles de données externes.

1.5 Données censurées

Il semble que la première occurrence du terme *censuré* pour désigner ce type particulier de données manquantes remonte à Anders Hald [1949] qui en attribue la paternité à J.E. Kerrich. Hald (1913-2007) était professeur à l'université de Copenhague et est surtout connu pour ses ouvrages sur l'histoire de la statistique. John Edmund Kerrich (1903–1985) était un mathématicien et actuaire anglais qui fit sa carrière en Afrique du Sud (université de Witwaterstrand). Il fut fait

prisonnier par les nazis en 1940 alors qu'il était en visite à Copenhague. Il occupa son temps de détention en effectuant 10 000 tirages de pile ou face pour illustrer la loi des grands nombres, parmi d'autres activités similaires.

Les données censurées sont fréquentes dans les analyses de durée de vie, en biostatistique et en fiabilité notamment. On parle de censure droite si pour une observation on sait seulement que la valeur de la variable d'intérêt X est supérieure à une valeur t, mais que l'on ne connaît pas sa valeur exacte. On sait par exemple que la durée de vie d'un individu encore présent dans une étude est supérieure à t contrairement au cas de la troncature où l'observation n'existe pas.

On parle de censure de Type I si t est le même pour toutes les observations, c'est le cas fréquent des études de fiabilité où les essais sont arrêtés au bout d'un temps t. La censure de type II intervient si les essais sont menés jusqu'au $r^{\text{ème}}$ « décès ». La censure aléatoire, la plus commune en biostatistique, correspond au cas où chaque observation i peut être ou non censurée par une valeur t_i. On se reportera utilement à Droesbeke *et al.* [1989] ou Cohen [1991] pour une description des différents types de censure et de troncature.

1.5.1 Daniel Bernoulli et la vaccination contre la variole

Le mémoire présenté à l'Académie royale des sciences en 1760 par Daniel Bernoulli (1700-1782), intitulé « Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, & des avantages de l'inoculation pour la prévenir »[Bernoulli, 1760] est sans doute le premier modèle mathématique en épidémiologie [Gabriel et de la Harpe, 2010], une discipline qui ne se développera que bien plus tard.

On attribue abusivement à cette étude la première utilisation de données censurées puisque les durées de vie exactes des individus vaccinés n'étaient pas intégralement connues. En réalité Daniel Bernoulli avait construit un modèle mathématique à l'aide d'équations différentielles pour calculer ce que serait l'espérance de vie à la naissance d'individus vaccinés et non vaccinés en se fondant sur la table de mortalité établie par Halley pour la ville de Breslau. Le modèle montrait un gain d'espérance de vie à la naissance de 3 ans et lui permettait de conclure : « l'intérêt public demandera toujours, non seulement que l'on emploie l'Inoculation, mais encore qu'on se hâte de l'employer ».

1.5.2 L'estimateur de Kaplan-Meier

Edward Kaplan a relaté lui-même en 1985 dans le bulletin hebdomadaire des *Current Contents* de l'*Institute for Scientific Information* (incorporé dans Clarivate Analytics) la genèse de leur fameux estimateur non-paramétrique de la fonction de survie qui prend en compte des durées censurées à droite ([Kaplan et Meier, 1958]) :

« Cet article a débuté en 1952 lorsque Paul Meier, de l'Université Johns Hopkins a découvert le travail de Greenwood sur la durée du cancer. Un an plus tard, alors aux Bell Labs, je me suis intéressé à la durée de vie des tubes à vide dans les répéteurs de câbles téléphoniques sous-marins. Quand j'ai montré mon manuscrit à John W. Tukey, il m'a informé du travail de Meier, qui circulait déjà parmi certains de nos collègues. Les deux manuscrits étaient soumis au *Journal of the American Statistical Association* qui recommanda d'en faire un article commun. Une correspondance abondante pendant quatre ans fut nécessaire pour concilier nos approches divergentes, et nous étions inquiets qu'entre-temps, quelqu'un d'autre pourrait publier l'idée ».

Il est intéressant de voir qu'en les personnes d'Edward Kaplan et Paul Meier se rencontraient les deux principaux domaines où les durées de vie sont utilisées : la fiabilité des matériels et la santé. Edward Kaplan (1920-2006) était un mathématicien, condisciple à Princeton de John Nash. Le sujet de sa thèse soutenue en 1950 était *Infinite permutations of stationary random*

sequences. Il travailla ensuite aux Bell Labs jusqu'en 1957 avant de rejoindre l'université d'état de l'Oregon. Paul Meier (1924-2011) fit sa thèse à Princeton sous la direction de John Tukey. Il fut également un ardent promoteur des essais randomisés en médecine. On lira avec intérêt son interview (Marks [2004]).

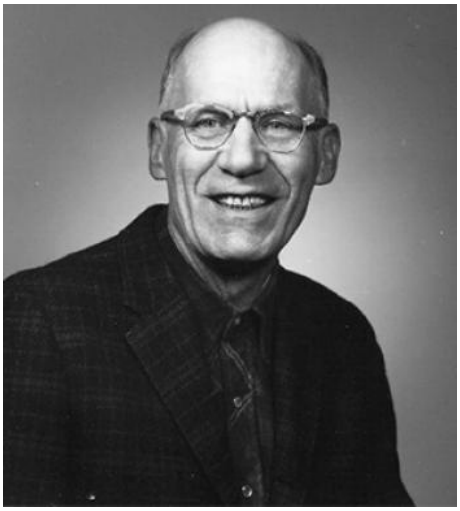

Figure 1.10 : *Edward Kaplan*
1920-2006

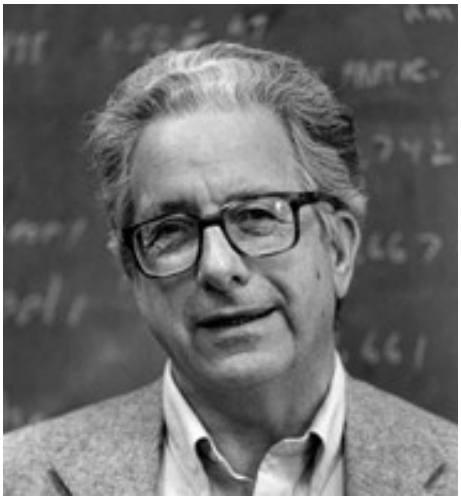

Figure 1.11 : *Paul Meier*
1924-2011

1.5.3 Créer des données manquantes : *winsorization*, *trimming*

Il peut sembler paradoxal d'éliminer des données vu le mal que l'on se donne en général pour les obtenir ! C'est pourtant une pratique courante dès que des valeurs sont suspectes ou quand on veut obtenir des estimateurs robustes. On se reportera avec profit à l'ouvrage Droesbeke *et al.* [2015].

Les moyennes tronquées ou *trimmed means* sont utilisées depuis longtemps et dans différents domaines allant de l'astronomie au sport. Elles consistent à enlever une fraction souvent de 5 à 25 % des plus grandes et plus petites valeurs avant de calculer la moyenne arithmétique. La *moyenne olympique* consiste à éliminer la plus petite et la plus grande valeur et est utilisée dans les disciplines comportant un jury mais aussi pour estimer des rendements agricoles. Le cas limite où on élimine 50% des plus grandes et des plus petites valeurs correspond à la médiane.

La *winsorisation* consiste à remplacer les valeurs extrêmes par des valeurs identiques aux sous-extrêmes correspondant aux x % des plus petites et plus grandes valeurs initiales. Il s'agit donc d'une forme de double censure gauche et droite, suivie d'une imputation par la plus grande ou la plus petite valeur disponible. Le nom de cette technique a été donné en hommage au biostatisticien britannique Charles P. Winsor (1895–1951) par Wilfrid Dixon [1960] et non par J.W. Tukey comme souvent indiqué de manière inexacte car Tukey et Winsor avaient travaillé ensemble.

1.6 Variables inobservables : de l'analyse factorielle aux classes latentes

Les modèles à variables latentes font référence à un cas très particulier de données manquantes, celui où toutes les valeurs d'une ou plusieurs variables sont manquantes, car ces variables sont inobservables. Les variables que l'on observe sont souvent appelées *variables manifestes*. L'existence même de variables latentes fait l'objet de débats philosophiques dans certaines sciences sociales mais pas dans toutes comme le remarque Bartholomew *et al.* [2011] : « en économie, une variable latente peut être réelle dans le sens où elle peut, en principe du moins, être mesurée. Ainsi la richesse personnelle est un concept raisonnablement bien défini qui pourrait être exprimé en termes monétaires, mais en pratique, nous ne sommes pas toujours capable de le mesurer ou désireux de le faire. Néanmoins, nous pouvons souhaiter l'inclure comme variable explicative dans les modèles économiques ».

1.6.1 Les quatre modèles et l'hypothèse d'indépendance conditionnelle

La figure 1.12 reprise de Bartholomew *et al.* [2011] classe les quatre principaux modèles selon la nature qualitative-discrète ou bien continue, des variables latentes et des variables manifestes.

	Variables latentes	
Variables manifestes	**qualitatives**	**quantitatives**
qualitatives	classes latentes	traits latents
quantitatives	profils latents	analyse factorielle

Figure 1.12 : *Les quatre modèles de structures latentes*

Ces modèles ont été développés indépendemment les uns des autres dans diverses sciences sociales : psychologie pour l'analyse factorielle, sociologie pour les classes latentes, sciences de l'éducation pour les traits latents (les modèles de Rasch en sont une sous-classe). Leur hypothèse fondamentale commune est la suivante : les corrélations entre variables manifestes s'expliquent par le fait qu'elles sont causées par un petit nombre de variables communes appelés *variables latentes*. Si ces variables latentes sont fixées, alors les variables manifestes deviennent indépendantes conditionnellement aux variables latentes : il y a factorisation de leurs lois conditionnelles.

1.6.2 L'analyse factorielle : Spearman et Thurstone

Le premier modèle à variable latente, l'analyse factorielle, appelée également *analyse en facteurs communs et spécifiques* pour la distinguer de l'analyse en composantes principales est attribué au psychologue anglais Charles Edward Spearman [1904] qui selon Bartholomew *et al.* [2011] avait remarqué que les personnes, surtout les enfants, qui réussissaient bien à un test d'aptitude mentale avaient aussi tendance à bien réussir dans d'autres. Cela le conduisit à l'idée que les scores de tous les individus étaient des manifestations d'une capacité générale sous-jacente notée g qu'il identifia à l'intelligence. Le modèle unifactoriel de Spearman s'écrit ainsi :

$$x_i = \lambda_i g + u_i \tag{1.3}$$

Les x_i sont les variables manifestes au nombre de p, les u_i les facteurs spécifiques et λ_i est appelée *saturation* ou *factor loading*. On doit aussi à Spearman le coefficient de corrélation des rangs qu'il proposa également en 1904 et qui porte son nom.

Le modèle de Spearman céda la place au modèle plurifactoriel de Thurstone [1931] qui s'écrit ainsi :

$$x_i = \sum_{j=1}^{k} \lambda_{ij} f_j + u_i \tag{1.4}$$

Il s'agit donc d'un modèle de régression multiple multilinéaire sur variables inobservables. Les p variables manifestes x_i sont supposées centrées et réduites, les k facteurs communs f_j centrés réduits et indépendants entre eux, les p facteurs spécifiques u_i centrés indépendants entre eux et indépendants des facteurs communs.

Une abondante littérature est consacrée à l'analyse factorielle portant en particulier sur l'existence et l'unicité de la solution. Si l'existence de facteurs communs relève du modèle lui même, il n'y a pas unicité de la solution car le modèle est sur-paramétré. On se réferera aux articles historiques de Mulaik [1986] et de Steiger [1979] sur l'indétermination des facteurs.

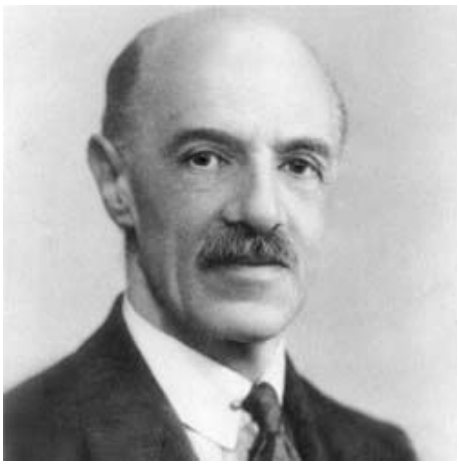

Figure 1.13 : *Charles Spearman 1863-1945*

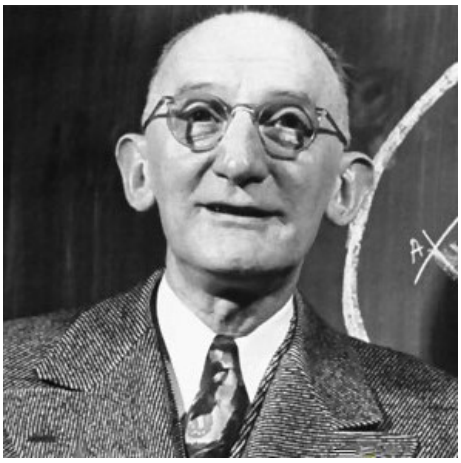

Figure 1.14 : *Louis Leon Thurstone 1887-1955*

1.6.3 Lazarsfeld et les classes latentes

Le modèle des classes latentes a été développé à partir des années 1940 jusqu'aux années 60 par le sociologue et ex-mathématicien Paul Lazarsfeld (1901-1976). Le contexte initial était celui des études psycho-sociales menées par le département de la guerre américain sur le personnel militaire pendant la deuxième guerre mondiale. L'ouvrage de référence est Lazarsfeld et Henry [1968]. On se reportera utilement à l'article de synthèse de Henry [2014].

Ce modèle est l'équivalent de l'analyse factorielle dans le cas entièrement qualitatif : les p variables observées sont qualitatives (souvent dichotomiques) et on postule l'existence d'une variable latente également qualitative à k modalités (les classes latentes).

Notons π_j la probabilité de la classe latente j et p_{ij} la probabilité que la variable dichotomique x_i prenne la valeur 1 dans la classe j. Le modèle de classes latentes s'écrit :

$$\mathbb{P}[\mathbf{x}] = \sum_{j=1}^{k} \pi_j \prod_{i=1}^{p} p_{ij}^{x_i} (1 - p_{ij})^{1-x_i} \tag{1.5}$$

On reconnaît un modèle particulier de mélange de distributions indépendantes dans chaque classe. On peut le voir comme une méthode de classification des observations dans des classes telles que les variables observées soient localement indépendantes.

En 1950, on ne disposait ni de test d'ajustement, ni de méthode convenable d'estimation. En 1974, Leo Goodman identifia le modèle de classes latentes comme un modèle log-linéaire. Considérer l'appartenance aux classes latentes comme des données manquantes conduisit ensuite à utiliser des algorithmes d'estimation de type EM.

La biographie de Paul Lazarsfeld est retracée dans le chapitre historique de Droesbeke *et al.* [2013].

Figure 1.15 : *Paul F. Lazarsfeld 1901-1976*

Ajoutons ici quelques éléments sur les relations complexes entre Pierre Bourdieu (1930-2002) et Paul Lazarsfeld. Bourdieu y revient à deux fois dans Bourdieu [2004] son dernier ouvrage posthume issu de son cours final au Collège de France. Il rappelle qu'au début des années soixante il avait « obstinément refusé d'assister à l'enseignement que Paul Lazarsfeld avait donné à la Sorbonne, devant la sociologie française toute entière réunie car cela lui était « apparu comme une cérémonie collective de soumission ». Bourdieu relate ensuite l'histoire de sa « confrontation à première vue désespérée avec Paul Lazarsfeld qui trouva un dénouement heureux » : à la fin des années soixante Bourdieu fut « littéralement convoqué »par Lazarsfeld dans l'hôtel où il séjournait à Paris, car Lazarsfeld voulait lui faire part de ses critiques sur son livre *L'Amour de l'art* publié avec Alain Darbel. À l'issue de l'entretien, Lazarsfeld déclara avec solennité qu'ils n'avaient jamais fait aussi bien aux Etats-Unis mais comme le note Bourdieu « il se garda bien de l'écrire » !

On notera pour finir que Pierre Bourdieu écrivit en 1981 la préface de l'édition française des *Chômeurs de Marienthal*, travail fondateur de Lazarsfeld en sociologie publié en 1933.

1.7 En quête des mécanismes

L'article de Fisher [1934] commençait par ces mots *It is a statistical commonplace that the interpretation of a body of data requires a knowledge of how it was obtained.*

Ce principe vaut également pour les données manquantes mais la prise en compte des mécanismes conduisant à des données manquantes a été longtemps ignorée jusqu'à ce que ce concept soit formalisé dans l'article fondateur de Rubin [1976]. L'ouvrage de Little et Rubin [2019] attribue le mérite de la formalisation des mécanismes de données manquantes à un *simple device of treating the missingness indicators as random variables and assigning them a distribution.* On note cependant une utilisation antérieure des variables indicatrices en régression linéaire dans Glasser [1964] mais insuffisamment exploitée.

Pour en savoir plus sur Donald Rubin, on se reportera avec profit à l'excellente entrevue qu'il a accordée à Li et Mealli [2014].

La description des mécanismes *MAR, MCAR, MNAR* est traitée dans le chapitre suivant ce qui nous permet de conclure cet historique : place au présent !

Figure 1.16 : *Donald Rubin 1943-*