Chapitre 6
ESTIMATIONS ET TESTS D'HYPOTHÈSES

6.1 Introduction

Les statistiques peuvent apporter une aide à la décision. C'est là leur intérêt principal.

Les outils que sont les moyennes, les écarts quadratiques moyens, les coefficients de corrélation, les histogrammes et les nuages de points permettent de décrire de façon synthétique des populations nombreuses et, par conséquent, d'apprécier les situations à partir desquelles des décisions seront prises : décisions qui engagent une action ou qui, plus modestement, attribuent des valeurs précises à des paramètres jusqu'à lors mal connus (c'est l'apport des statistiques à la théorie de la mesure par exemple).

L'une des caractéristiques de l'outil statistique concerne sa capacité à estimer les propriétés d'une population nombreuse à partir d'un échantillon de taille réduite. Ce mécanisme, qui permet de tenir pour vrai certaines propriétés compte tenu de certaines observations qui en seraient la conséquence probable, est connue comme ***"une inférence"*** et les méthodes sont celles des *statistiques inférentielles*. Nous avons déjà rencontré, plusieurs fois, un tel mécanisme, par exemple lors de l'estimation d'une incertitude standard page 78, où des résultats vraisemblablement contradictoires suggèrent de reprendre les expériences. A ce propos, remarquons que décider de l'incompatibilité de deux campagnes de mesures c'est rejeter l'hypothèse que les moyennes des observations sont les mêmes. Ainsi, se donner les moyens décider, c'est, ici, *tester une hypothèse*.

Dans un registre différent, les exemples pages 60 et 75 montrent comment *estimer* la taille d'un échantillon lorsqu'on cherche à connaître une moyenne.

A l'évidence deux sujets importantes émergent : **les estimations et les tests d'hypothèses**. Ce sont ces questions que nous abordons dans ce chapitre.

Les méthodes d'estimation et de comparaison ainsi que les tests d'hypothèses ont été développés par les statisticiens dans une multitude de cas, pour tous les objectifs imaginables. Ce chapitre ne saurait être une présentation complète, ni même approfondie, du domaine, mais seulement une introduction aux méthodes employées et aux façons de penser et de poser les problèmes. Notre intention n'est pas, non plus, de vous assommer de démonstrations*, mais seulement de présenter l'intérêt de ces méthodes, et aussi leurs limites.

Avant de commencer, il est bon de rappeler que les statistiques fournissent une *aide* à la décision, seulement une aide, pas plus : décider n'est pas le fait du statisticien† (voir les commentaires de la figure 4-4, page 62).

---

*Vous devrez donc nous faire confiance dans nos affirmations. Vous pouvez aussi consulter l'un des nombreux ouvrages de statistiques existant et y étudier les démonstrations.

†Si vous devez subir une opération à risque, c'est au médecin, au chirurgien, de vous informer des risques (honnêtement et complètement) mais c'est à vous de prendre la décision et de l'assumer, à vous seul.

6.2 Estimations

Considérons une variable aléatoire, X, dont on a effectué n mesures indépendantes. Les problèmes les plus fréquents portent sur la façon d'obtenir une estimation de l'espérance $E(X)$ et de la variance $\mathcal{V}(X)$ de la loi de probabilité suivie par X, compte tenu des observations. Nous avons déjà étudié ces questions au paragraphe 5.3.2 page 80 pour ce qui concerne l'estimation de l'espérance mathématique d'une variable X et au paragraphe 5.3.4, page 82 pour ce qui concerne la variance de X. Dans ces deux paragraphes, la seule hypothèse posée concernent la répartition normale des erreurs et ne concerne ni la nature, ni la valeur des grandeurs mesurées. Les résultats obtenus sont donc très généraux. Nous les rappelons ici.

La méthode consiste à répéter la même mesure. Si c'est la mesure d'une grandeur physique bien définie, tous les résultats seront voisins les uns des autres. Mais si on cherche à connaître la proportion de ceux qui voteront pour Clovis, il est normal que les résultats diffèrent d'une mesure à l'autre car le résultat de la mesure est 1 si la personne interrogée se propose de voter pour Clovis et 0 dans le cas contraire (voir l'application de l'inégalité de Bienaymé-Tchebychev paragraphe 4.2.2, page 57).

Quelle que soit le cas considéré, la méthode consiste à mesurer la grandeur que l'on souhaite estimer. Cette mesure est une épreuve aléatoire qui définit une variable G. On répète la mesure, construisant ainsi une suite de variables aléatoires qui suivent toutes la même loi de probabilité. Cette suite de variables aléatoires est la suite G_1, G_2, ..., G_n, les résultats obtenus sont g_1, g_2, ..., g_n. L'ensemble de ces n mesures constitue une campagne de mesure.

On introduit la variable aléatoire $Z = \dfrac{1}{n}\left(G_1 + G_2 + ... + G_n\right)$

Le résultat de la campagne de mesures est la série statistique $(g_1, g_2, ..., g_n)$. On note $\overline{g}_n$ la moyenne de cette série, c'est la valeur prise par Z et s_n son écart-quadratique moyen.

Remarquez que $\overline{g}_n$ et s_n sont des quantités numériques connues.

Notre propos est de trouver une estimation de l'espérance mathématique $E[G]$ et de la variance $\mathcal{V}(G)$ de la variable G.

Remarquez que $E[G]$ et $\mathcal{V}(G)$ ne sont pas connus.

Nous notons $\tilde{E}[G]$ et $\tilde{\mathcal{V}}[G]$ des estimations sans biais de $E[G]$ et $\mathcal{V}(G)$. La présence du "~" signifie donc que nous avons affaire à une estimations.

Les résultats déjà obtenus sont les suivants

$$\begin{array}{c} Z = \dfrac{1}{n}\left(G_1 + G_2 + ... + G_n\right) \\ E[Z] = E[G] \qquad \mathcal{V}[Z] = \dfrac{1}{n}\mathcal{V}[G] \\ \boxed{\tilde{E}[G] = \overline{g}_n} \qquad \boxed{\tilde{\mathcal{V}}[Z] = \dfrac{1}{n-1}{s_n}^2} \text{ avec} \\ \overline{g}_n = \dfrac{1}{n}\sum_{k=1}^{n} g_k \text{ et } \qquad {s_n}^2 = \dfrac{1}{n}\sum_{k=1}^{n}\left(g_k - \overline{g}_n\right)^2 \end{array} \tag{6.1}$$

voir les relations 5.3.2 et 5.8.

6.2.1 Echantillon d'effectif élevé ($n \gtrsim 30$)

On définit μ et σ : $\mu = E(G)$ et $\sigma^2 = \mathcal{V}(G)$. Lorsque n est assez grand, on utilise le théorème central limite pour estimer la loi de probabilité suivie par Z.

Celle-ci est une loi normale d'espérance μ et de variance $\dfrac{1}{n}\sigma^2$.

On pose $T = \dfrac{Z-\mu}{\sigma/\sqrt{n}}$. la *variable T suit une loi normale, centrée, réduite.* La probabilité de l'événement $\{Z \in (z, z+dz)\}$ est dP; c'est la probabilité de l'événement $\{T \in (t, t+dt)\}$ avec $t = \dfrac{z-\mu}{\sigma/\sqrt{n}}$.

Les quantités μ et σ ne sont pas connus. On remplace σ par son estimation $\tilde{\sigma}$, $\boxed{\dfrac{\tilde{\sigma}}{\sqrt{n}} = \dfrac{s_n}{\sqrt{n-1}}}$. On admet alors que la variable $\boxed{\tilde{T} = \dfrac{Z-\mu}{\tilde{\sigma}/\sqrt{n-1}}}$ suit une loi normale centrée, réduite, caractérisée par la densité de probabilité $p(t)$:

$$proba\left[\tilde{T} \in [t, t+dt]\right] = dP = p(t)dt = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\ dt$$

Remarquons qu'ici, $E\left[\tilde{T}\right] = \mu$ est inconnu. On peut aussi remplacer μ par son estimation, $\boxed{\tilde{\mu} = \overline{g}_n}$, et admettre que la variable $\boxed{\tilde{\tilde{T}} = \dfrac{Z-\tilde{\mu}}{\tilde{\sigma}/\sqrt{n-1}}}$ suit la loi normale centrée réduite :

$$proba\left[\tilde{\tilde{T}} \in [t, t+dt]\right] = dP = p(t)dt = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\ dt$$

6.2.2 Echantillon d'effectif réduit ($n \lesssim 30$)

Considérons la variable aléatoire $Z = \dfrac{1}{n}(G_1 + G_2 + ... + G_n)$. Les variables aléatoires G_i et G_j étant indépendantes, il vient $\mathcal{V}(Z) = \dfrac{1}{n}\mathcal{V}(G)$. La contribution de la variable G_k à la variance totale est $\dfrac{1}{n^2}\mathcal{V}(G_k) = \dfrac{1}{n^2}\mathcal{V}(G) = \dfrac{1}{n}\mathcal{V}(Z)$.
La condition $\dfrac{1}{n^2}\mathcal{V}(G_k) << \mathcal{V}(Z)$ implique donc $n >> 1$. Pour les petits échantillons, les conditions d'application du théorème centrale limite ne sont donc pas satisfaites. On admet généralement que le problème se pose pour $n \lesssim 30$.

Les relations 6.1 ci-dessus sont satisfaites. Cependant, dans le cas $n \lesssim 30$, lorsque G est issue d'une population normale, la variable centrée, réduite, $\tilde{T} = \dfrac{Z-\tilde{\mu}}{\tilde{\sigma}}$ suit une ***loi de Student*** (voir page 108) et non pas une loi normale. La loi de Student a été découverte et étudiée par William Gosset (1876-1937), employé à la brasserie Guiness à Dublin. William Gosset avait établi un test de contrôle du brassage de la bière qui permettait de ne considérer que des échantillons peu nombreux. Ce test donnait un avantage certain à son employeur et l'autorisation de publier ses travaux lui fut refusée[‡]. Il les jugea, à juste titre, suffisamment importants pour qu'ils ne restent pas confidentiels. Aussi en assura-t-il la publication sous un pseudonyme.

La probabilité de l'événement $\left\{\dfrac{Z-\tilde{\mu}}{\tilde{\sigma}} \in (t, t+dt)\right\}$ est

$$dP = p(\nu, x)\ dx \quad \text{avec } \nu = n-1$$

La densité de probabilité de la loi de Student, $p(\nu, x)$ dépend du paramètre ν. Le paramètre ν ne figure pas dans la densité de probabilité de la loi normale. En effet la loi normale correspond à $\nu \to \infty$, pratiquement à $\nu \gtrsim 30$.

Le paramètre ν est appelé "nombre de degré de liberté" (ddl).

[‡]A cette époque lointaine (1908) l'employeur, achetait le travail, le temps, le repos dominical, le cerveau de ses employés. Mais c'était autrefois....(?).

6.3 Comparaison de deux moyennes

Avec les méthodes descriptives, et les méthodes d'estimations, les tests d'hypothèses constituent les outils privilégiés des statistiques. Nous allons détailler sur un exemple, la procédure employée pour conduire les tests d'hypothèses les plus courants.

6.3.1 La méthode

Prenons un exemple. Un employeur que nous appellerons Oreste, pour ne désigner personne, se pose la question de l'intérêt du repos dominical[§]. L'effet sur la production en est-il bénéfique ? Ou au contraire néfaste ?

Pour tenter de donner une réponse à cette question, Oreste opère ainsi. Il contrôle $N_L = 300$ pièces issues de la production annuelle du lundi et $N_A = 1000$ pièces issues de la production des autres jours de la même année. La proportion de pièces défectueuses observées le lundi est x_L, tandis que les autres jours elle est x_A. On considère que x_L et x_A sont les valeurs prises par les variables aléatoires X_L et X_A. Les observations donnent $x_L = \dfrac{28}{300} \simeq 0,09$ et $x_A = \dfrac{86}{1000} \simeq 0,086$. Peut-on admettre que la différence est significative ou au contraire, doit on admettre que la différence observée est due au hasard de l'échantillonnage ?

On opère ainsi :

1- On établit clairement le schéma probabiliste auquel on se réfère. En particulier il faut préciser les populations d'où sont tirées, au hasard, les échantillons observés. Ces populations peuvent exister réellement ou exister potentiellement.

• Dans le cas considéré, la population des pièces fabriquées le lundi existe réellement.

• Si on fabrique 300 pièces seulement pour étudier le comportement de la chaîne de production consécutif à une modification, la population a une existence virtuelle : c'est l'ensemble des pièces qu'est susceptible de fabriquer la chaîne étudiée. C'est sur cette population virtuelle que l'on souhaite des informations tandis que les 300 pièces fabriquées en constitue un échantillon.

• Enfin, si on teste toutes les pièces produites le lundi et toutes les pièces produites les autres jours, le schéma probabiliste n'est plus valable. C'est un schéma qui relève des statistiques descriptives car il n'y a pas dans ce cas de tirages au hasard.

Ici nous supposons que les pièces du lundi sont tirées au hasard, de façon non exhaustive, à partir d'une urne où la proportion de pièces défectueuses est p_L et que les pièces fabriquées les autres jours sont tirés au hasard, de la même façon, à partir d'une urne où la proportion de pièces défectueuses est p_A.

2- On établit clairement l'hypothèse que l'on veut tester : par exemple, nous baptisons "hypothèse H" l'hypothèse selon laquelle "la proportion de pièces défectueuse fabriquées le lundi est la même que la proportion de pièces défectueuses fabriquées dans l'année". Cette hypothèse $\{p_L = p_A\}$ est l'*hypothèse nulle.*

Les raisonnements qui suivent présupposent la validité de cette hypothèse ; si les résultats qui s'en déduisent sont "improbables", l'hypothèse est rejetée.

3- On choisit un seuil de confiance π_s. Posons $\pi_s = 0,95 = 95\%$.

[§]Ne croyez pas que cet exemple soit aussi académique qu'il le paraisse. La question se pose réellement dans certains pays de savoir si les voitures assemblées le lundi ne souffriraient pas de l'abus d'alcool pendant la fin de semaine.

4- On construit **un indicateur** R qui présente les propriétés suivantes.

• R est positif ou nul

• R est fonction des observations (ici x_L et x_A, ainsi que les nombres de pièces contrôlée, N_A et N_L) et éventuellement des paramètres caractérisant les hypothèses lorsque ceux-ci sont connus.

On impose comme contrainte de pouvoir calculer numériquement R. Dans le cas considéré, R ne peut pas être une fonction de p car p est inconnu : $R = R(N_A, N_L, x_A, x_L)$.

• On souhaite que R mesure la "distance" entre l'observation et l'hypothèse, ce qui impose deux conditions que l'on satisfait au mieux.

1- La relation $R = 0$ devrait impliquer la validité de H. En fait, le mieux dont on puisse s'assurer c'est que $R = 0$ lorsque H est vérifiée (c'est-à-dire lorsque $p_A = p_L$) et que l'on remplace les paramètres inconnus (comme p_A et p_L) par leur estimation (x_A et x_L).

2- Il faut ajouter la condition que R croît lorsque la vérification de H se dégrade. C'est-à dire, ici, que R est une fonction croissante de $|x_A - x_L|$ par exemple.

Nous avons déjà rencontré la construction d'indicateurs et il n'est pas inutile de se reporter aux pages 14, 17, 22, 26, 27, 34 et 42.

5- La valeur de R est la valeur de la variable aléatoire $R(N_A, N_L, X_A, X_L)$ où X_A et X_L sont des variables aléatoires. Il est nécessaire que sa loi de probabilité soit connue.

De nombreux indicateurs peuvent être construits, par exemple $R = (X_A - X_L)^2$. Nous choisissons ici de poser $R = |X_A - X_L|$ car la loi de probabilité suivie par R est simple (voir ci-dessous).

Remarquons que N_A et N_L sont assez grands (supérieurs à 30) pour que, selon le théorème centrale limite, X_A et X_L suivent (presque certainement) une loi normale de variance $\sigma^2_{X_A} = \dfrac{p(1-p)}{N_A - 1}$ et $\sigma^2_{X_L} = \dfrac{p(1-p)}{N_L - 1}$, sous réserve de l'hypothèse H que nous admettons ici en posant $p_L = p_A = p$.

Comment estimer p?

Suivant l'hypothèse H, on dispose d'un tirage de 1300 individus, issus de la même urne. Dans cet échantillon on dénombre $28 + 86 = 114$ pièces défectueuses, soit une proportion $p = 114/1300 = 0,0877$; c'est cette valeur que l'on prend comme estimation de p. Les estimations sont donc les suivantes

$$\sigma^2_{X_A} = 8 \times 10^{-5} \quad \text{et} \quad \sigma^2_{X_L} = 2,7 \times 10^{-4}$$

La différence $Z = X_A - X_L$ de deux variables qui suivent une loi normale suit une loi normale (cf. paragraphe 5.1 page 71 et l'annexe page 83) : $E(Z) = p - p = 0$ (hypothèse H) et $\sigma^2_Z = \sigma^2_{X_A} + \sigma^2_{X_L} = 3,5 \times 10^{-4}$.

Posons $Z = u\ \sigma_Z$. La variable u présente une distribution normale, centrée, réduite dont la loi de probabilité est connue.

6- Nous calculons la valeur R. Nous trouvons ici $R = |Z| = 7,3 \times 10^{-3}$.

7- A ce stade nous disposons de tous les outils nécessaires. Comme nous connaissons la loi de probabilité suivie par R, nous pouvons déterminer la valeur de R qui a une probabilité $1 - \pi_s$ d'être dépassée. Soit R_s cette valeur. Ici, $1 - \pi_s = 5\%$. Quelle est la valeur u_s, telle que $|u|$ à une probabilité 5% d'être supérieure à u_s? Ou en d'autre termes, connaissant l'aire grise($1 - \pi_s = 0,05$) de la figure 6-1, quelle est la valeur de u_s. La réponse se lit dans les tables de la loi normale : $u_s = 1,96$. Or $|u| > 1,96$ implique $|Z| = R > 1,96 \times \sigma_Z = 0,037 = R_s$.

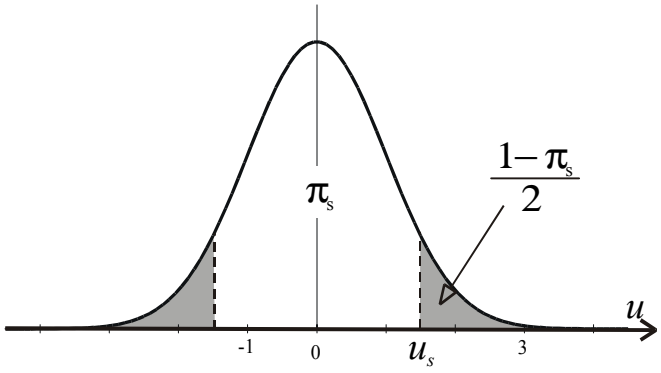


Figure 6-1

La probabilité de $R > 0,037$ est donc de 5%. L'intervalle (0-0,037) est donc l'intervalle de confiance au seuil 0,95 (on dit aussi que **le risque** est $1 - \pi_s = 5\%$ ici).

6.3.2 Premier exemple : l'acceptation

Les données sont résumées dans le tableau ci-dessous appelé *tableau de contingence.*

	nombre de contrôles	nombre de pièces défectueuses
pièces du lundi	300	28
autres jours	1000	86

Ici, nous trouvons $R = 7,3 \times 10^{-3} < R_s = 0,037$. Il n'est donc pas invraisemblable d'observer une valeur de R "*aussi grande*" avec les hypothèses que nous avons posées. Par conséquent rien ne nous permet de rejeter l'hypothèse H que nous acceptons.

Remarquons qu'en acceptant H, nous n'avons aucune certitude. Nous prenons donc le *risque d'accepter une hypothèse fausse.*

6.3.3 Un second exemple : rejet

Reprenons le même exemple. Mais supposons que les observations de pièces défectueuses soient modifiées :28 → 36 le lundi et 86 → 78 les autres jours.

	nombre de contrôles	nombre de pièces défectueuses
pièces du lundi	300	36
autres jours	1000	78

On vérifie que p n'est pas modifié, non plus que σ_Z. L'intervalle de confiance à 95% est inchangé : $R_s = 0,037$; cependant R est modifié : $R = \dfrac{36}{300} - \dfrac{78}{1000} = 0,042 > R_s$. Dans ces conditions, on considère que l'obtention d'une valeur aussi grande de R est improbable (moins de 5%). Il est préférable de rejeter l'hypothèse et de croire que s'est réalisé un événement probable pour lequel $p_L > p_A$.

En rejetant l'hypothèse, on prend cependant le *risque de rejeter une hypothèse exacte.*

6.3.4 Un troisième exemple : rejet

Supposons que les observations du lundi portent sur $\lambda \times 300$ pièces dont $\lambda \times 28$ sont défectueuses, tandis que les autres observations concernent $\lambda \times 1000$ pièces dont $\lambda \times 86$ sont défectueuses.

	nombre de contrôles	nombre de pièces défectueuses
pièces du lundi	$\lambda \times 300$	$\lambda \times 28$
autres jours	$\lambda \times 1000$	$\lambda \times 86$

On vérifie que R conserve la même valeur $R = 7,3 \times 10^{-3}$. De même π_s étant inchangé à 95%, il s'en suit que u_s est inchangé, $u_s = 1,96$. Estimé comme nous l'avons indiqué lors du premier exemple, p conserve sa valeur, $p = 0,0877$. Rien ne change sauf σ_Z qui est divisé par $\sqrt{\lambda}$. On en déduit $R_s = 1,96 \times \sigma_Z = 0,037/\sqrt{\lambda}$. On constate que pour λ assez grand (ici $\lambda > 26$) on obtient $R > R_s$. On est alors conduit à rejeter l'hypothèse. La comparaison avec le premier exemple montre qu'une différence de moyenne entre deux échantillons n'est pas significative si les observations sont peu nombreuses mais que la même différence le devient si les observations sont nombreuses (multipliées par λ, assez grand).

Sur cet exemple, on constate qu'accepter une hypothèse est la conséquence d'une absence d'arguments en faveur du rejet.

6.3.5 Un quatrième et dernier exemple

Considérons l'exemple suivant

	nombre de contrôles	nombre de pièces défectueuses
pièces du lundi	300	33
autres jours	1000	81

On considère deux choix possibles du seuil de confiance. Tous calculs faits on obtient

$$R = \frac{33}{300} - \frac{81}{1000} = 0,029$$

$$p = \frac{33+81}{1300} = 8,77 \times 10^{-2}$$

$$\sigma_Z = \sqrt{p\,(1-p)\left(\frac{1}{299}+\frac{1}{999}\right)} = 1,9 \times 10^{-2}$$

$$\pi_s = 0,95 \Rightarrow R_s = 1,96 \times 1,9 \times 10^{-2} = 0,037$$

$$\pi_s = 0,85 \Rightarrow R_s = 1,44 \times 1,9 \times 10^{-2} = 0,027$$

Toutes choses égales par ailleurs, le choix du seuil de confiance π_s détermine l'acceptation ou le refus. Ainsi pour $\pi_s = 0,95$, il vient $u_s = 1,96$ et par conséquent $R_s = 0,037 > 0,029 = R$. Rien ne s'oppose à ce que nous acceptions l'hypothèse H. Par contre pour $\pi_s = 0,85$, on trouve $u_s = 1,44$ et $R_s = 0,027 < 0,029 = R$. Dans ce cas, il vaut mieux rejeter l'hypothèse et admettre que le lundi le travail est de moindre qualité.

Entre ces deux décisions rien n'a changé sauf des considérations abstraites qui se déroulent dans la tête du décideur.

Si π_s est grand, R_s est grand et on a toutes chances de voir se vérifier la relation $R < R_s$: on accepte donc souvent. Par conséquent il y a peu de risque de rejeter une hypothèse exacte. C'est cette attitude qu'il faut prendre si on teste une hypothèse de travail sur laquelle on souhaite s'appuyer (pour mettre au point un nouveau vaccin par exemple). Dans ce cas $1 - \pi_s$ est petit devant l'unité. C'est le "risque", c'est-à-dire la

probabilité pour avoir $R > R_s$ et donc de rejeter l'hypothèse H alors qu'elle est vérifiée. La condition $R > R_s$ signifie que l'hypothèse est mal vérifiée et qu'il est préférable de ne pas l'accepter (compte tenu des observations). R_s étant grand, cette condition n'est satisfaite que rarement ; on est donc conduit à accepter souvent l'hypothèse H. Enfin, une dernière façon de comprendre le cas étudié consiste à remarquer que $1 - \pi_s$ est le seuil à partir duquel on considère qu'une probabilité est négligeable. Choisir $1 - \pi_s << 1$, c'est décider que seules les situations hautement improbables sont considérées comme improbable. On n'est donc pas surpris que les situations de probabilité inférieures à $1 - \pi_s$ se rencontrent rarement. Or ce sont de telles situations qui conduisent au rejet de l'hypothèse H.

Si π_s est petit R_s est petit par conséquent la relation $R > R_s$ se rencontre souvent : on rejette souvent l'hypothèse H. Il y a donc peu de risque d'accepter une hypothèse fausse. C'est cette attitude qu'il faut prendre pour convaincre un public méfiant que l'hypothèse proposée est pertinente (pour mettre un vaccin sur le marché par exemple).

Un compromis entre les deux risques conduit à la valeur $\pi_s = 0,95$, généralement acceptée.

6.4 Le test de χ^2

Parmi les tests les plus souples d'emploi, le test de χ^2 de Pearson[¶] permet de tester l'hypothèse d'homogénéité de plusieurs populations.

Considérons deux exemples.

6.4.1 Comparaison à un standard

Le tableau ci-dessous donne la répartition moyenne des groupes sanguins parmi les Espagnols et les Français

Groupe	O	A	B	AB	Total
Espagnols	38%	47%	10%	5%	100%
Français	43%	47%	7%	3%	100%

Le pays Basque se répartit entre l'Espagne et la France. La langue y est originale pour la région (ce n'est pas une langue latine) mais en outre le bruit court que la répartition des groupes sanguins est différente de la répartition moyenne rencontrée en Espagne et en France.

Dans un lycée de Biarritz, parmi 96 élèves tirés au hasard, les groupes sanguins se répartissent ainsi

Groupe	O	A	B	AB	total
effectif	48	41	5	2	96

Nous posons la question de savoir si l'on peut considérer la population des élèves comme issue de la population espagnole moyenne ou de la population française moyenne.

Nous supposons que *nous tirons au hasard 96 individus parmi les Espagnols.* Si les proportions étaient exactement celles de la population d'origine nous trouverions la répartition théorique suivante :

Groupe	*O*	*A*	*B*	*AB*	Total
Répartition théorique	36,48	45,12	9,6	4,8	96

[¶] χ est une lettre grecque. χ^2 se prononce et s'écrit khi2 ou encore chi2, lorsque les lettres latines sont les seules disponibles.

Karl Pearson (1857-1936) est l'un des premiers a avoir développé le test auquel est associé son nom.

Par exemple, dans la première colonne, la colonne O, pour obtenir 36,48 nous avons multiplié l'effectif (96 lycéens) par la proportion rencontrée dans la population espagnole (38%) : en effet $96 \times 0,38 = 36,48$.

Bien sûr, on ne saurait prendre une fraction d'élève. La répartition théorique n'est donc pas observable. Toutefois il ne s'agit pas ici de reproduire la répartition théorique mais seulement d'apprécier la "distance" qui sépare les observations et la théorie pour construire un indicateur numérique. La contribution de la colonne A à cet indicateur est prise sous la forme $\dfrac{(41-45,12)^2}{45,12}$, ce que l'on écrit $\dfrac{(n_{obs}-n_{th})^2}{n_{th}}$.

Le tableau de contingence pour la comparaison entre l'échantillon d'élèves de Biarritz et la population espagnole est le suivant

Groupe	O	A	$B\&AB$	Total
Théorie	36,48	45,12	14,4	96
Observation	48	41	7	96

Nous avons regroupé en une seule catégorie les deux catégories les moins nombreuses (B et AB). La raison est la suivante. L'indicateur, R, que nous utilisons (et dont nous donnons l'expression ci-dessous) suit une loi de probabilité dont le comportement asymptotique est connu, c'est pratiquement celui d'une loi de χ^2, lorsque **l'effectif de chacune des classes est supérieure où égale à 5**.

L'indicateur R est construit de la façon suivante.

Pour chacune des classes de la population observée, on forme $\dfrac{(n_{obs}-n_{th})^2}{n_{th}}$. L'indicateur R, appelé " indicateur de χ^2 " est obtenu en additionnant toutes les valeurs obtenues :

$$\boxed{R=\sum_k \frac{(n_{obs}-n_{th})_k^{\,2}}{(n_{th})_k}} \tag{6.2}$$

Dans l'exemple considéré il vient

$$R=\frac{(48-36,48)^2}{36,48}+\frac{(41-45,12)^2}{45,12}+\frac{(7-14,4)^2}{14,4}=7,8$$

- On peut vérifier que R a les propriétés qui en font un indicateur convenable pour apprécier combien sont différents l'échantillon et la population de référence.
- Pour aller plus loin, il faut choisir un seuil de confiance π_s.

Prenons $\pi_s = 95\%$. Le "risque" est donc 5%.

- Lorsque le tirage de l'échantillon s'effectue au hasard dans la population de référence (les Espagnols), R est une variable aléatoire qui suit une loi de χ^2 à ν degrés de liberté (ddl). Le nombre de ddl est obtenu à partir du tableau de contingence :

$$\boxed{\nu=(L-1)\,(C-1)} \tag{6.3}$$

où L est le nombre de lignes et C le nombre de colonnes de ce tableau.

Ici $L = 2$ et $C = 3$, on en déduit $\nu = 2$.

Cette hypothèse que le tirage a été effectué dans la population des Espagnols est l'hypothèse H.

- Sous réserve de la validité de H, la valeur de R_s au seuil de confiance de 95% est donné par les tables de la loi (voir page 106) : $R_s = 5,99$.
- La relation $R > R_s$ (ici $7,8 > 5,99$) suggère que nous rejetions l'hypothèse H. En effet, si H est satisfaite, l'observation d'une valeur si grande de R est improbable.

Il vaut mieux attribuer une si grande valeur de R au fait que l'échantillon est issu d'une population différente de la population de référence.

Une étude analogue peut être faite avec la population française comme population de référence. Le tableau de contingence est le suivant

Groupe	O	A	$B\&AB$	Total
Théorie	41,28	45,12	9,6	96
Observation	48	41	7	96

L'indicateur R prend la valeur $R = 2,2$.

Au seuil de confiance de 95%, il vient $R_s = 5,99 > R = 2,2$. Rien ne nous permet de rejeter l'hypothèse H. Nous l'acceptons donc.

En réalité une étude de la population basque donne les résultats suivants

Groupe	O	A	B	AB	Total
Basques	51%	44%	4%	1%	100%

On constate que la population basque diffère notablement de la population française moyenne (cf. classes O et $B\&AB$). Un calcul semblable aux précédents donne $R = 1,1$. Au seuil de confiance de 95%, nous acceptons l'hypothèse que l'échantillon de lycéens est issu de la population Basque

Que cet échantillon puisse être issu de la population française n'est pas incompatible avec la possibilité qu'il soit issu de la population basque. Cependant, la valeur de R étant ici plus petite que la précédente on peut affirmer que la population des lycéens étudiée ressemble plus à la population basque qu'à la population française moyenne.

.

Dans les exemples précédents la variable est une variable non numérique dont les valeurs possibles sont *O, A, B, AB.* Le test de χ^2 peut aussi s'appliquer lorsque la variable est numérique et que la population de référence est une une population normale ou poissonnienne par exemple. Dans ce cas, on découpe l'intervalle $(-\infty, \infty)$ en classes. On compare les effectifs observés dans chaque classe avec les effectifs théoriques des mêmes classes, obtenus à partir de la loi de référence. On peut ainsi tester la normalité d'une variable, par exemple.

6.4.2 Test d'homogénéité

Le test de χ^2 permet aussi la comparaison de deux ou plusieurs échantillons afin de décider si ces échantillons sont issus d'une même population, inconnue.

Considérons l'exemple suivant. Une même maladie est traitée de 4 façons différentes dans 4 services hospitaliers différents noté $S1$, $S2$, $S3$, $S4$. Les résultats obtenus sont classés en trois catégories Guérison, Rémission, Décès. Le tableau de contingence qui contient l'ensemble des informations à prendre en compte est le suivant

	$S1$	$S2$	$S3$	$S4$
Guérison	123	95	152	132
Rémission	15	18	22	20
Décès	28	19	63	53

De ce tableau, nous tirons le nombre de ddl suivant la formule 6.3 : $\boxed{\nu = 6}$

Nous formons ensuite la population théorique la plus vraisemblable, par addition des populations de chaque classe

	S1	S2	S3	S4	*total*	*proportion*
Guérison	123	95	152	132	*502*	$502/740 = 68\%$
Rémission	28	19	63	53	*163*	$163/740 = 22\%$
Décès	15	18	22	20	*75*	$75/750 = 10\%$
Total	*166*	*132*	*237*	*205*	***740***	100%

Nous pouvons donc admettre que la variable aléatoire associée à chaque patient est une variable non numérique dont les valeurs sont G, R ou D. La population de base contient 68% de G, et 22% de R ainsi que 10% de D. On considère que $S1, S2, S3$ et $S4$ sont des échantillons tirés au hasard de la population théorique, dont les cardinaux sont respectivement 166 pour $S1$, 132 pour $S2$, 237 pour $S3$ et 205 pour $S4$.

On peut obtenir pour chaque échantillon, l'effectif théorique des trois classes G, R et D.

Par exemple pour $S2$:

$$\begin{array}{ll} \text{G} & 132\times 0,68 = 89,76 \\ \text{R} & 132\times 0,22 = 29,04 \\ \text{D} & 132\times 0,1 = 13,2 \end{array}$$

On construit un tableau qui contient pour chaque échantillon et chaque classe les deux valeurs n_{obs} et n_{th} où n_{osb} est tiré du tableau de contingence tandis que n_{th} est calculé comme nous l'avons indiqué[‖].

	$S1$		$S2$		$S3$		$S4$	
Guérison	123	112,88	95	89,76	152	161,16	132	139,4
Rémission	28	36,52	19	29,04	63	52,14	53	45,1
Décès	15	16,6	18	13,2	22	23,7	20	20,5

Dans chacune des 12 cases on lit $\boxed{n_{obs} \quad n_{th}}$ et on calcule $(n_{obs} - n_{th})^2 / n_{th}$

	$S1$	$S2$	$S3$	$S4$	*total*
Guérison	0,91	0,31	0,52	0,39	*2,13*
Rémission	1,99	3,47	2,26	1,38	*9,10*
Décès	0,15	1,28	0,12	0.012	*1,56*
total	*3,05*	*5,06*	*2,90*	*1,78*	***12,79***

Tableau des $(n_{obs} - n_{th})^2 / n_{th}$ (6.4)

On additionne ensuite le contenu de chaque case. On obtient $\boxed{R = 12,79}$.

On fait l'hypothèse que les quatre échantillons sont issus de la population de référence (hypothèse H). Dans ces conditions, R est une variable aléatoire qui suit une loi de χ^2 à $\nu = 6$ ddl. Au niveau de risque $1 - \pi_s = 0,025$, soit au niveau de confiance $\pi_s = 0,975$, la table donne $R_s = 14$. A ce niveau, on peut accepter l'hypothèse H (hypothèse d'homogénéité). Dans ces conditions, on ne peut pas affirmer que l'un des traitements est plus efficace que les autres. Cependant, au seuil de confiance de 95%, il vient $R_s = 12,59$; on est conduit à rejeter l'hypothèse d'homogénéité et à accepter une conclusion opposée.

Ce dernier résultat indique qu'il faut, en termes de recherche, rejeter H et tenter de comprendre l'origine de l'hétérogénéité présumée (aspects positifs et négatifs des traitements dans les divers services). Il y a là une piste de recherche à approfondir. Cependant, il n'est pas scandaleux de rester sceptique.

Le tableau de contingence peut recevoir une interprétation différente. Nous pouvons considérer que celui-ci représente trois échantillons de patients : ceux qui sont guéris

[‖] Si un effectif n'est pas un entier, c'est vraisemblablement un effectif théorique.

(échantillon G), ceux qui sont en rémission (échantillon R) et ceux qui sont décédés (échantillon D). A chaque patient est associée la valeur d'une variable S1, S2, S3 ou S4 selon le service qui l'a soigné. Ainsi le tableau représente la décomposition des 3 échantillons en 4 classes. Cette nouvelle interprétation du tableau de contingence consiste donc à échanger les échantillons et les classes. Le calcul de R donne cependant le même résultat que précédemment.

Dans le tableau 6.4 on constate que l'essentiel de R est apporté par la contribution des "*Rémissions*". Il est donc intéressant d'éliminer cette catégorie et de tester l'homogénéité des divers échantillons à partir du nouveau tableau de contingence. Cette méthode permet de localiser l'origine des hétérogénéités qu'il reste alors à interpréter et à comprendre. On peut aussi se pencher sur l'homogénéité des critères qui permettent le classement en "*Rémission*" des patients concernés.

6.5 Conclusion

Pour terminer ce chapitre, nous présentons ce qui pourrait en être un prolongement aux applications multiples.

Les tests d'hypothèses se rapportent souvent à la question de savoir à quel point diverses populations se ressemblent. Tournée en termes de probabilités, cette question peut s'énoncer ainsi "Quelle est la probabilité pour que je puisse considérer que ces échantillons sont issus de la même population ?". On ne peut pas répondre à une telle question mais on peut considérer un indicateur, R, et répondre à la question "Etant donnée la valeur observée de R que nous notons R_0, quelle est la probabilité de l'événement $R < R_0$ si les échantillons sont issues d'une même population ?" Soit π_0 cette probabilité. L'indice de ressemblance sera pris égal à $\alpha_0 = 1 - \pi_0$ (voir la figure 6-2).

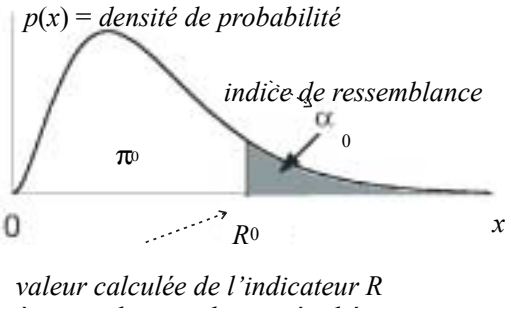


Figure 6-2.

Pour interpréter l'indice de ressemblance, considérons deux populations P_1 et P_2, pour lesquelles l'indice de ressemblance est α_0. Les propriétés suivantes sont satisfaites :

- L'indice de ressemblance est compris entre 0 et 1.

- Pour $\alpha_0 \sim 1$ (c'est-à-dire $\pi_0 = 1 - \alpha_0 << 1$), l'indicateur R_0 est "petit" par conséquent les populations se ressemblent fortement. Ou encore, posant $\pi_s = \pi_0$, et tirant au hasard deux échantillons de la population P, on est amené à rejeter le plus souvent l'hypothèse d'homogénéité. Cela signifie que ces populations se ressemblent moins que P_1 et P_2. Ou encore que P_1 et P_2 se ressemblent tellement que la probabilité d'une ressemblance plus forte est très faible.

$\alpha_0 \sim 1$ **signifie une forte ressemblance.**

- A l'opposé, $\alpha_0 << 1$ **signifie une forte dissemblance.**

Remarquons que α_0 est le "risque" : risque que l'on prend en admettant $P_1 \neq P_2$ (voir page 94).

Pour illustrer l'usage de α_0 considérons les exemples que nous avons traités pour présenter le test de χ^2.

Dans le tableau ci-dessous, R est la valeur de l'indicateur de χ^2, tel que nous l'avons calculé, ν est le nombre de degrés de liberté et α_0 est la valeur du risque, α, qui est lue sur la table $\chi^2(\nu, \alpha)$ telle que $\chi^2(\nu, \alpha) = R$.

Comparaison :	$R = R_0$	ν	$\alpha = \alpha_0$
Lycéens/Espagnols	$7,8$	2	0,02
Lycéens/Français	$2,2$	2	0,33
Lycéens/Basques	$1,1$	2	0,58
$\{S1, S2, S3, S4\}$	$12,79$	6	0,047

Au seuil de confiance 90%, le risque est $10\% = 0,1 = a$. Cette valeur apparaît comme une toise qui permet de qualifier de petit ou grand un individu selon qu'il passe ou non sous la toise. Ici si $\alpha_0 < a$ on qualifie les échantillons comparés de "hétérogènes" (rejet de l'hypothèse H), dans le cas contraire on les qualifie de "homogènes" (acceptation de H).

Si la "toise" est mise à la "hauteur" $a = 0,1$ on rejettera l'hypothèse H dans le premier et le dernier cas. Si la toise est mise à la hauteur $a = 0,05$ on rejettera H dans le premier cas (ressemblance insuffisante), on acceptera H dans les autres cas.

L'indice de ressemblance est un outil plus subtile que la dichotomie *oui* ou *non*. Il peut en particulier être utilisé pour décrire la ressemblance entre deux populations que l'on connaît complètement, ce qu'exclut le modèle probabiliste, faute de sens.

L'usage de l'indice de ressemblance permet (en principe) de laisser la décision à prendre au décisionnaire plutôt qu'à l'expert. Le décisionnaire est en effet libre de fixer la toise où bon lui semble et c'est de sa responsabilité.

Enfin, il donne un sens à une proposition comme "Pour les caractères considérés, les quatre services hospitaliers $S1, S2, S3, S4$ se ressemblent plus que ne ressemblent les lycéens de Biarritz à la population moyenne espagnole".

Chapitre 7
TABLES, LOIS ET FORMULES

N'attendez aucune explication nouvelle de ce chapitre. C'est seulement un résumé, un rappel des formules utiles et la présentation de quelques tables avec leur mode d'emploi.

7.1 Notations et formules

• La **moyenne** statistique de X est notée $\overline{X}$, l'*espérance* mathématique d'une variable aléatoire X, est notée $E(X)$. De façon générale, l'une ou l'autre de ces deux quantités est notée $\langle X \rangle$:

$$\langle X \rangle = \sum_k P_k \; X_k$$

où P_k est la **proportion (*probabilité*)** de valeurs X_k (ce qui implique $\sum_k P_k = 1$)

• λ est une constante :

$$\sum_k P_k = 1 \Rightarrow \qquad \langle \lambda \rangle = \lambda \times \sum_k P_k = 1$$

• L'**écart quadratique moyen** (***écart-type*** dans le langage des probabilités) est la racine carrée, σ, de la *variance*, $\mathcal{V}$:

$$\sigma^2 = \left\langle \;\; (X - \langle X \rangle)^2 \;\; \right\rangle = \mathcal{V}(X) = \left\langle X^2 \right\rangle - \langle X \rangle^2$$

• La ***covariance*** est

$$cov\,[X,Y] = \left\langle \;\; (X - \langle X \rangle)\;\; (Y - \langle Y \rangle) \;\; \right\rangle = \langle XY \rangle - \langle X \rangle \langle Y \rangle$$

• Le coefficient de ***corrélation*** linéaire est

$$\rho\,[X,Y] = \frac{cov\,[X,Y]}{\sigma_X \; \sigma_Y}$$

• Pour λ et μ constants, on démontre

$$\begin{array}{ccc} \langle X_1 + X_2 + ... \rangle & = & \langle X_1 \rangle + \langle X_2 \rangle + ... \\ \langle \lambda X \rangle & = & \lambda \langle X \rangle \end{array}$$

$$\begin{array}{ccc} \mathcal{V}(\lambda X) = \lambda^2 \mathcal{V}(X) & \Leftrightarrow & \sigma_{\lambda X} = |\lambda| \;\; \sigma_X \\ \mathcal{V}(X+Y) = \mathcal{V}(X) + \mathcal{V}(Y) + 2\; cov\,[X,Y] & \Leftrightarrow & \sigma^2_{(X+Y)} = \sigma^2_X + \sigma^2_Y + 2\sigma_X \sigma_Y \; \rho\,[X,Y] \\ cov\,[\lambda X, \mu Y] = \lambda\mu \; cov\,[X,Y] & \Leftrightarrow & \rho\,[\lambda X, \mu Y] = \rho\,[X,Y] \times (\text{signe de } \lambda\mu) \end{array}$$

• **Probabilité conditionnelle et formule de Bayes** :

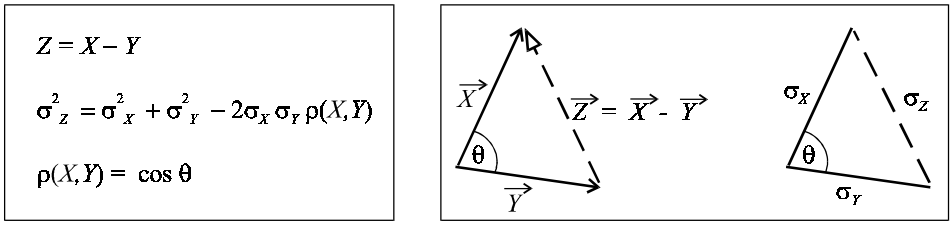
probabilité de X conditionnée par $Y = P_{X/Y} = proba\,(X \text{ et } Y) \;\; / \; proba(Y)$

$$proba(X \text{ et } Y) = P_Y \times P_{X/Y} = P_X \times P_{Y/X}$$

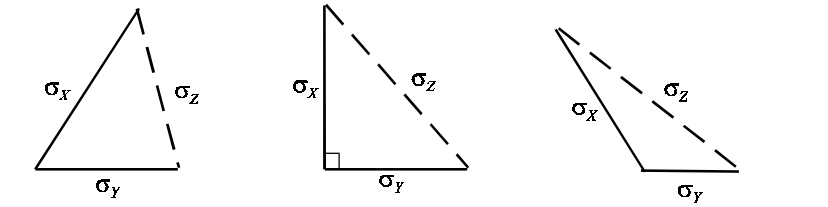
• Variables, X et Y, ***indépendantes*** :

$$proba\ _{(X=x \text{ et } Y=y)} = proba\ _{(X=x)} \times proba\ _{(Y=y)}$$

$$\rho\,[X,Y] = 0 \text{ et } \ \sigma^2_{(X+Y)} = \sigma^2_X + \sigma^2_Y$$



Soustraction de variables aléatoires ..... analogie géométrique



Corrélation positive *Corrélation nulle* *Corrélation négative*

Variables orthogonales.

NB. Deux variables indépendantes sont orthogonales

• Passage des variables discrètes aux **variables continues** (ci-après F et G sont des fonctions quelconques) :

$$\begin{array}{lcl}
X_k & \to & x \\
proba\,\{X = X_k\} = P_k & \to & proba\{X \in [x, x+dx]\} = p(x)\ dx \\
F(X_k) = F_k & \to & F(X) \\
\sum_k F(X_k) & \to & \displaystyle\int_{-\infty}^{+\infty} F(x)\ dx \\
\langle G(X)\rangle = \sum_k G(X_k)\ P_k & \to & \displaystyle\int_{-\infty}^{+\infty} G(x)\ p(x)\ dx = \langle G\rangle
\end{array}$$

• **Deux formules utiles** :

Une intégrale $\displaystyle\int_{-\infty}^{+\infty} e^{-\alpha^2 x^2 + \beta x}\ dx = \sqrt{\frac{\pi}{|\alpha|}}\ e^{\frac{\beta^2}{4\alpha^2}}$

et la formule de Stirling $\displaystyle n! = \left(\frac{n}{e}\right)^n \sqrt{2n\pi} \times \left(1 + \frac{1}{12n} + \mathcal{O}\left(\tfrac{1}{n^2}\right)\right)$

dont on déduit $\ln n! = n \ln n\ (1 + \varepsilon_n)$ avec $\varepsilon_n = -\dfrac{1}{\ln n} + \mathcal{O}\left(\tfrac{1}{n}\right)$

7.2 Lois usuelles

7.2.1 *Loi normale*

Variable continue $X \in (-\infty, \infty)$. La probabilité de l'intervalle $(x, x+dx)$ est

$$p(x)\ dx = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\ \sigma^2}\right] \times dx$$

$$\begin{array}{ll} \text{moyenne :} & \langle X \rangle = \mu \\ \text{écart-type :} & \sigma \end{array}$$

La variable aléatoire, somme de plusieurs variables normales indépendantes, suit une loi normale.

7.2.2 *Loi de Poisson*

La variable X est discrète : $k \in \mathbb{N}$ (entiers non négatifs). La probabilité de k est P_k :

$$proba\,[X = k] = P_k = e^{-m}\ \frac{m^k}{k!}$$

$$\begin{array}{lccc} \text{moyenne :} & \langle k \rangle & = & m \\ \text{écart-type} & \sigma & = & \sqrt{m} \end{array}$$

7.2.3 *Loi binomiale*

Dans une population contenant la proportion p d'individus remarquables, on prélève au hasard un échantillon de cardinal n. La probabilité d'avoir k individus remarquables dans cet échantillon est P_k :

$$P_k = C_n^k p^k\,(1-p)^{n-k} \qquad \text{avec} \qquad C_n^k = \frac{n!}{k!\,(n-k)!}$$

$$\begin{array}{ll} \text{moyenne :} & \langle k \rangle = n\ p \\ \text{écart-type} & \sqrt{n\ p\ (1-p)} \end{array}$$

• La loi binomiale peut être assimilée à **une loi de Poisson lorsque** $p \to 0$ (par exemple $p < 0,10$), avec $m = np$

$$\begin{array}{lll} \text{variable poissonnienne :} & k & \text{de moyenne } m = np \\ \text{probabilité :} & P_k = & e^{-m}\ \dfrac{m^k}{k!} \end{array}$$

• La loi binomiale peut être assimilée à **une loi normale lorsque** $n\ p\ (1-p) \to \infty$ (par exemple $n\ p\ (1-p) \geq 20$), avec $\mu = np$ et $\sigma = \sqrt{n\ p\ (1-p)}$, ce qui suppose que l'on assimile la variable k à une variable continue :

$$\begin{array}{lll} \text{variable continue, normale réduite :} & u = & \dfrac{k-np}{\sqrt{np\,(1-p)}} \\ \text{densité de probabilité :} & p(u) = & \dfrac{1}{\sqrt{2\pi}} \times e^{-\frac{u^2}{2}} \end{array}$$

Voir la remarque ci-dessus (Figure C-1) concernant l'assimilation d'une variable entière à une variable continue.

• De façon générale, l'assimilation d'une variable entière à une variable continue* demande certaines précautions. Nous considérons une variable discrète X susceptible de prendre les valeurs entières k avec la probabilité P_k. Sur la figure C-1, nous avons représenté la *probabilité* P_k sous la forme d'un bâtonnet. La courbe représentant la *densité de probabilité*, $\tilde{P}_k$, de la variable continue k, est également représentée. Pour des raisons graphiques, nous avons supposé que la différence $\tilde{P}_k - P_k$ n'excédait pas l'épaisseur du trait pour k entier.

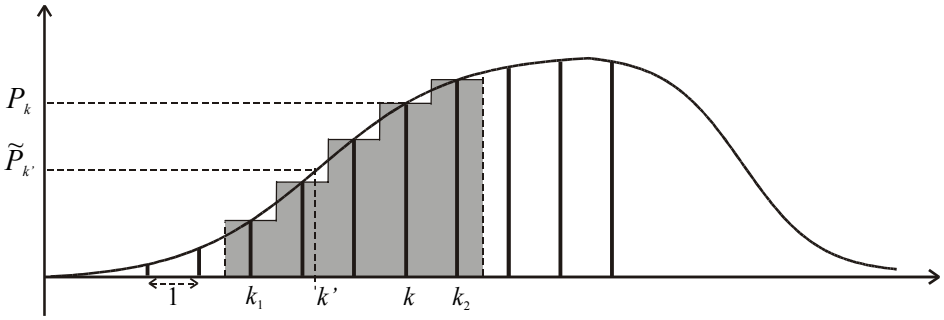


Figure C-1

Posons $\sum_{k_1}^{k_2} F(k)P_k = \sum_{k_1}^{k_2} F(k)\ P_k \times 1$. Assimiler k à une variable continue revient à utiliser $\tilde{P}_k\ dk$ au lieu de $P_k \times 1$. Une somme entre k_1 et k_2 deviendra une intégration entre $k_1 - \dfrac{1}{2}$ et $k_2 + \dfrac{1}{2}$. Ainsi, $\sum_{k_1}^{k_2} P_k = \sum_{k_1}^{k_2} (P_k \times 1) = \mathcal{A}$ où $\mathcal{A}$ est l'aire en gris sur la figure C-1, soit pratiquement $\mathcal{A} = \displaystyle\int_{k_1-0,5}^{k_2+0,5} \tilde{P}_k\ dk$.

7.2.4 <u>*Autres lois*</u>

Il existe de nombreuses autres lois : la loi de χ^2, celle de Student (voir ci-dessous), la loi de Snedecor destinée à l'étude des variances, celle de Fisher pour les corrélations et bien d'autres, parmi lesquelles cette étonnante loi de Benford selon laquelle dans une liste quelconque de grandeurs mesurées la probabilité pour que d soit le premier chiffre significatif est $P_d = \log_{10}\left(\dfrac{d+1}{d}\right)$. Cette loi est expérimentalement bien vérifiée que les grandeurs impliquées soient des masses (en kg), des charges (en C), des épaisseurs (en cm), des cours de bourses (en €), etc... Cette loi est invariante sous un changement d'unités ; dans une base quelconque, a, elle s'écrit $P_d = \log_a\left(\dfrac{d+1}{d}\right)$. La loi de Benford est la seule loi présentant ces propriétés.

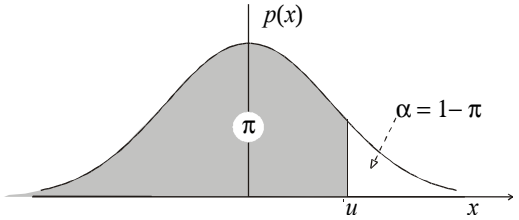
*Voir page 7.

7.3 Tables

Nous donnons ici une présentation succincte des tables de quelques lois d'usage courant. Il existe des outils informatiques et des calculatrices spécialisées qui rangent les tables dans les greniers aux souvenirs. Elles restent cependant bien utiles pour étudier un cours ou suivre quelques exemples.

7.3.1 Tables de la loi normale, centrée ($\mu = 0$) et réduite ($\sigma = 1$)

La densité de probabilité est $p(x) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, La fonction de répartition est $f(u) = \pi = \displaystyle\int_{-\infty}^{u} p(x)dx$. La table ci-dessous donne $f(u) = \pi$ en fonction de u.



Densité de probabilité.

La somme des valeurs de tête d'une ligne et d'une colonne donne une valeur de u; la valeur de $f(u) = \pi$ est lue à l'intersection de cette ligne et de cette colonne.

u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7290	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9779	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Table pour les grandes valeurs de u

u	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
π=F(u)	0,99865	0,99904	0,99931	0,99952	0,99966	0,99976	0,999841	0,999928	0,999968	0,999997

7.3.2 Loi de χ^2 de Pearson

Soient $X_1, X_2, ...X_\nu$ des variables aléatoires suivant une loi normale, centrée, réduite. On pose $R = X_1^2 + X_2^2 + ...X_\nu^2$. La variable R suit une loi de χ^2 à ν degrés de liberté dont les propriétés sont données dans le tableau ci-dessous

Espérance	$E(R)$	=	ν
Variance	$\mathcal{V}(R)$	=	2ν
Ecart-type	σ	=	$\sqrt{2\nu}$

On utilise couramment le test de χ^2 comme test d'homogénéité.

Les données sont constituées par des populations, $P1$, $P2$, .., Pn, divisées en classes, Cl_1, Cl_2, ..., Cl_k. Parmi les données peut figurer, éventuellement, la population de référence qui constitue le standard. Le cas échéant, on note cette population $P0$.

Chaque classe constitue une colonne et chaque ligne une population. L'ensemble des données forme donc un tableau de k colonnes et de n où $n+1$ lignes suivant que $P0$ est donné *a priori* ou non. Ce tableau est le tableau de contingence.

1- Le tableau de contingence présente L lignes et C colonnes ($C = k$ et $L = n$ ou $n+1$ selon le cas). Le nombre de degrés de liberté (ddl) est $\nu = (L-1)(C-1)$.

2- Si la population théorique n'est pas donnée, on la construit en additionnant dans chaque classe les effectifs des diverses populations. Chaque classe de la population $P0$ est alors définie en pourcentages : le pourcentage qui apparaît dans la classe Cl_j représente la proportion p_j de la population $P0$.

Remarquons que si $P0$ n'est pas donnée au départ, la ligne correspondante n'intervient pas dans la détermination de ν.

3- Pour chaque classe de chaque population, on calcule l'effectif théorique : $(n_{th})_{j/m}$ est l'effectif théorique de la classe Cl_j de la population Pm; elle vaut $(n_{th})_{j/m} = p_j \times N_m$ où N_m est l'effectif de la population Pm.

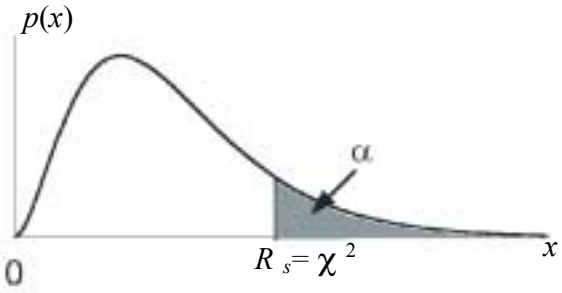
4- Pour que les hypothèses soient vraisemblablement vérifiées on regroupe les classes de façon à ce que les effectifs observés et les effectifs théoriques soient supérieurs (ou égaux) à 5 pour chaque classe de chacune des populations $P1$ à Pn.

5- Pour chaque classe Cl_j de chaque population, Pm, on forme $\left(\frac{(n_{th}-n_{obs})^2}{n_{th}}\right)_{j/m}$.

On additionne alors toutes ces quantités. On obtient $R = \sum_{j,m}\left(\frac{(n_{obs}-n_{th})^2}{n_{th}}\right)_{j/m}$.

R suit une loi de χ^2 à ν ddl.

Distribution du χ^2 (khi2 ou chi2) de K. Pearson



Densité de probabilité d'une loi de χ^2 à 6 ddl.

α / ν	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,001
1	0,0002	0,0010	0,0039	0,0158	2,71	3,84	5,02	6,63	10,83
2	0,02	0,05	0,10	0,21	4,61	5,99	7,38	9,21	13,82
3	0,12	0,22	0,35	0,58	6,25	7,81	9,35	11,34	16,27
4	0,30	0,48	0,71	1,06	7,78	9,49	11,14	13,28	18,47
5	0,55	0,83	1,15	1,61	9,24	11,07	12,83	15,09	20,52
6	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81	22,46
7	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,47	24,32
8	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	26,13
9	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	27,88
10	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	29,59
11	3,05	3,82	4,57	5,58	17,27	19,67	21,92	24,72	31,26
12	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	32,91
13	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	34,53
14	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14	36,12
15	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58	37,70
16	5,81	6,91	7,96	9,31	23,54	26,30	28,84	32,00	39,25
17	6,41	7,56	8,67	10,08	24,77	27,59	30,19	33,41	40,79
18	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,80	42,31
19	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19	43,82
20	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57	45,32
21	8,90	10,28	11,59	13,24	29,61	32,67	35,48	38,93	46,80
22	9,54	10,98	12,34	14,04	30,81	33,92	36,78	40,29	48,27
23	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64	49,73
24	10,86	12,40	13,85	15,66	33,20	36,41	39,37	42,98	51,18
25	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31	52,62
26	12,20	13,84	15,38	17,29	35,56	38,88	41,92	45,64	54,05
27	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96	55,48
28	13,57	15,31	16,93	18,94	37,92	41,34	44,46	48,28	56,89
29	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59	58,30
30	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89	59,70

Lorsque le nb de ddl, ν, est supérieur à 30, on peut utiliser la formule $\chi^2 = \frac{1}{2}\left(u + \sqrt{2\nu - 1}\right)^2$ où, α étant donné, u est obtenu en utilisant la loi normale, centrée, réduite : $\alpha = 2\int_u^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\, du$

7.3.3 Loi de Student

Soit X une variable normale centrée, réduite et Y une variable qui suit une loi de χ^2 à ν degrés de liberté (ddl). La variable $\dfrac{X}{\sqrt{Y/\nu}}$ suit une loi de Student à ν ddl.

Un échantillon de n valeurs, $X_1, X_2, ..., X_n$, prélevées dans une population normale de moyenne μ donne une moyenne observée $\overline{x}_n$ et un écart type observé s_n :

$$\overline{x}_n = \frac{1}{n}\sum_k X_k \ , \quad s_n^2 = \sum_k \frac{(X_k - \overline{x})^2}{n}$$

On pose $s = s_n/\sqrt{n-1}$. La variable $t = \dfrac{\overline{x} - \mu}{s}$ suit une loi de Student à $\nu = n-1$ ddl.

La loi de Student est couramment utilisée pour $n \lesssim 30$. Dans le cas $n \gtrsim 30$, la loi de Student peut être assimilée à une loi normale, centrée, réduite.

1- Avec la loi de Student, on peut tester l'hypothèse que la moyenne a une valeur donnée, m, ou encore calculer un intervalle de confiance. Le résultat s'écrit sous la forme

$$\mu = m \pm t \times s \ \text{ au risque } \alpha$$

où α est donné en fonction de t par la table ci-contre.

2- Deux échantillons d'effectifs n_1 et n_2 donnent les <u>observations</u> suivantes :

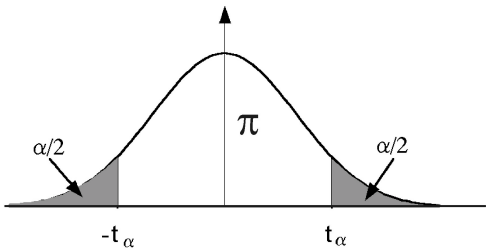
	moyenne	écart quadrtique moyen
$P1$	$\tilde{\mu}_1$	$\tilde{s}_1$
$P2$	$\tilde{\mu}_2$	$\tilde{s}_2$

On pose $s^2 = \dfrac{\left(\tilde{s}_1^2 n_1 + \tilde{s}_1^2 n_2\right) \times (n_1 + n_2)^2}{(n_1 + n_2 - 2) \times n_1 \ n_2}$.

La variable $t = \dfrac{\tilde{\mu}_1 - \tilde{\mu}_2}{s}$ suit une loi de Student à $\nu = n_1 + n_2 - 2$ ddl si les deux échantillons proviennent de la même population. Cette propriété permet de tester cette hypothèse.

3- Si ρ est le coefficient de corrélation calculé à partir de l'observation de n paires de variables suivant une loi normale à deux dimensions, la variable $t = \dfrac{\rho}{\sqrt{1-\rho^2}}\sqrt{n-2}$ suit une loi de Student à $\nu = n-2$ ddl si le "vrai" coefficient de corrélation est nul.

Cette propriété permet de tester la dépendance stochastique de deux variables (d'autres méthodes sont néanmoins préférables).



Densité de probabilité d'une loi de Student.

ν \ α	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,929
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Pour $\nu \gtrsim 30$, la loi de Student est pratiquement assimilable à une loi normale, centrée, réduite.

7.3.4 Nombres au hasard

Pour obtenir une liste de nombres au hasard, on part de n'importe où dans la table et on se déplace dans une direction et un sens fixé à l'avance (de droite à gauche et de haut en bas le plus souvent). On prélève alors les nombres dans l'ordre où on les rencontre : par exemple pour les nombres au hasard entre 27 et 103, on prélève la succession de paquets de 3 chiffres que l'on rencontre. Chaque paquet de 3 chiffres est lu comme un nombre compris entre 000 et 999. On retient alors dans l'ordre où ils apparaissent les nombres compris entre 27 et 103.

NOMBRES AU HASARD

5	10	15	20	25	30	35	40	45	50
43645	89232	00384	10858	21789	14093	06268	46460	97660	23490
61618	19275	40744	22482	12424	98601	19089	53166	41836	28205
68136	06546	04029	47946	19526	27122	42515	55048	23912	81105
74005	34558	93779	96120	01695	47720	88646	73520	40050	90553
54437	88825	07943	81795	31709	13358	04626	64838	92133	44221
01990	94762	89926	84764	19159	95355	98213	17704	47400	30837
02404	42408	67981	43684	55467	47030	42545	43920	11199	36521
59253	71535	26149	35629	87127	45581	00185	01041	46662	98897
20471	13914	99330	37938	69649	57964	97149	41628	78664	80727
65946	60766	74084	22484	49514	89820	41310	19722	07045	28808
00939	47818	75949	44707	49105	06777	31998	79942	98351	10265
49952	29123	45950	67578	13524	03023	18046	75287	74989	58152
17328	70732	46319	26950	19037	02831	36558	82712	05590	64941
19420	70215	90476	76400	51553	12158	14668	15656	37895	94559
19121	41190	49145	05373	00755	17817	22757	76116	76977	94570
44300	56179	71202	49238	83682	21989	63268	74644	53625	10791
99403	96757	34512	06475	89028	00290	93766	70812	98331	09611
78578	51589	83195	56332	75076	58202	58038	38817	63835	13486
89830	60177	94550	10119	09083	33398	29974	67721	75037	70444
89502	83947	99940	60969	79452	91472	12611	41681	95285	44153
11187	95098	50369	94874	19853	06933	69767	88842	35676	49766
47886	49549	64465	14508	28215	47766	03076	25940	47239	93425
21325	89726	96964	66106	68517	67954	16570	72433	91514	79333
59927	79213	96072	64540	59002	26619	02930	83677	26442	97346
44232	30754	59691	34893	92531	70313	24969	14458	91409	79369
15956	31379	21224	20366	74348	66239	32704	41018	31937	84761
58597	14598	23589	50700	96194	15831	08968	45321	04207	34438
99185	70628	95475	94156	39588	57825	36521	85188	64339	27460
20986	57081	53928	47768	18313	82950	12335	32298	08662	54552
75371	04678	96443	72965	68012	52485	55139	73430	74306	85960
75775	60178	51110	30735	29761	39565	45332	13671	69405	11186
91592	54102	25242	00063	42467	23339	55311	81275	08602	03508
16106	87812	92476	07849	65510	77763	33684	77092	32490	40345
30764	57054	12611	21455	01332	33101	64795	56555	84390	12982
63826	14146	40993	93849	49799	41080	48621	29555	83653	07742

Conclusion

En écrivant les pages qui s'achèvent ici, notre ambition n'était pas de vous donner un outil spécialisé, ni pour la physique, la sociologie, la pharmacie ou l'économie. Mais un outil pour aborder, sans en préjuger, celui de ces domaines dont vous aurez besoin demain, un outil pour la vie.

Au long des sept chapitres qui précèdent nous avons introduit les notions essentielles communes aux diverses branches des statistiques mais nous n'avons pas développé les méthodes qui leur sont spécifiques. Ne soyez donc pas surpris si l'ensemble vous paraît incomplet ; il l'est. Mais c'est sans difficulté que vous trouverez sur la toile des cours bien faits et les références de livres très complets.

Notre intention était de vous offrir ce que vous trouverez ailleurs avec une certaine difficulté :

- une présentation minimale des outils, de leur nature et des motivations qui présidèrent à leur construction,
- une présentation d'ensemble des questions et des méthodes spécifiques aux statistiques, avec un minimum de théorie, de démonstrations et de calculs,
- un regard aussi critique que le permet la bienséance.

Maintenant que tout est fini, c'est avec un peu de nostalgie que vient le temps de la conclusion : nostalgie d'un safari au coeur d'une jungle de concepts emmêlés, nostalgie de ces semaines passées avec vous, sans vous.

Avec la conclusion vient le doute. Avons-nous assez insisté sur la diversité et la complémentarité des moyens disponibles pour un même but ? Sur les possibilités d'évaluer les qualités et les faiblesses de ces moyens ? Sur les contradictions des résultats qu'ils fournissent parfois ? Sur les possibilités de manipulation qu'offrent le jeu des définitions[†], le choix des descripteurs et des indicateurs ?

Un pont est un ouvrage d'art, la résistance des matériaux est une science. La physique et la biologie sont des sciences, la médecine est un art. De même, par la subtilité des interprétations, les statistiques sont un art*, même si les moyens employés sont de nature scientifique. Mais c'est un art qui présente cette propriété étonnante d'être *probablement* une science, probablement une science exacte parfois, probablement une illusion aussi.

Jouez à *pile ou face*. Comptez le nombre de *piles* après 10 000 lancers, calculez la proportion, p, de *pile*. Une science exacte vous prédirait $p = 0,5 \times (1 \pm 0,02)$. Les théories que nous vous avons présentées prédisent la même chose, *probablement...* Il ne vous a pas

[†]Pour bien des gens, le nombre de chômeurs décrit, plus que l'état de l'emploi, le niveau de difficulté à vivre décemment. C'est, un indicateur de misère assez grossier, mais c'est ainsi. Partant de ce fait on comprend tout l'intérêt qu'il y a, pour ceux qui en sont tenus responsables, à une décroissance du nombre de chômeurs. C'est ce que l'on constate à chaque changement de définition de ce qu'est un chômeur.

*Ce qui ne présuppose aucun jugement de valeur, aucun jugement moral : la théorie des probabilités est une science, le jeu est un art ; certains y voient un gagne-pain ou un amusement, d'autres une passion ou une maladie.

échappé que "*probablement*" signifie qu'il y a doute ; le risque d'erreur est peut-être faible (inférieur à 2,5% ici), mais il n'est pas nul. Le doute est encore accentué dès que des comportements humains sont en cause.

Les sondages d'opinion auraient un pouvoir prédictif si l'opinion existait de façon stable. Dans l'argot des statisticiens, on dirait "si la population étudiée n'évoluait pas". C'est bien sûr une hypothèse forte que l'on signale honnêtement en rappelant qu'un sondage est une "photo de l'opinion à un moment donné" et que les résultats du sondage ne sont que *probablement* exacts. Après toutes ces précautions on croit que tout est dit. Non ! Il reste un silence : un mensonge par omission, peut-être pire, selon Disraeli, que ce fameux fieffé mensonge !

Evoquer l'opinion c'est admettre son existence. Pour le statisticien, c'est admettre que l'on étudie une population qui existe. En abandonnant toute idée préconçue, en ne tenant compte que des observations[†], il faut admettre que cette existence n'est que probable, vraisemblable dirons-nous pour n'être désobligeant envers personne. Il reste donc une certaine probabilité pour que cette population n'existe pas[‡].

Vous n'oseriez pas évoquer la "couleur" de l'année 1900, ni sa "masse". Un médecin n'oserait plus évoquer l'humidité des humeurs de tel malade, même si on lui en fournit la valeur, obtenue par un moyen qui en présupposerait l'existence. Ce sont pourtant dans les mêmes conditions que l'on évoque une "opinion", c'est-à-dire une"population" dont l'existence n'est vérifiée que par des moyens qui la présupposent.

Dans sa période médicale, Gil Blas de Santillane[§], sur les instructions de son maître Sangrado, pratique des saignées à répétition sur les malades. L'effet est assuré : ils meurent. Le diagnostique est donc confirmé : ils étaient malades. Ainsi, une explication fausse peut se trouver légitimée par l'expérience, du seul fait qu'elle ait été acceptée ("*Ils auront raison à force d'avoir tort*" disait Jean-Paul Sartre en évoquant le risque d'une guerre nucléaire). Il est très difficile de mettre en défaut de telles théories dont l'acceptation comme hypothèse de travail implique *ipso facto* la vérification [¶]. C'est ici d'autant plus difficile que la théorie prédit, avec une probabilité non rigoureusement nulle, des résultats compatibles avec les observations, quelles qu'elles soient.

Ainsi, lorsque l'existence d'une "population" statistique est supposée sans qu'elle ne puisse être observée directement, le doute s'impose quant à la justification scientifique de l'emploi des statistiques.

Apprenez à pratiquer le doute scientifique sans douter de la science : ce sera notre dernière leçon.

[†]Comme, par exemple, les résultats obtenus par divers organismes de sondages (souvent homogènes mais pas toujours).

[‡]La situation est similaire en économie où pour éviter le risque que ce ne soit pas une science selon les livres néoclassiques, l'homme doit se comporter comme un *homo œconomicus*.

[§]Roman de Alain-René Lesage (1668-1747).

[¶]C'est important car les seules certitudes que l'on puisse acquérir ne concernent pas la validité d'une théorie mais, le cas échéant, la nécessité de son rejet. Pour des raisons philosophiques, il est donc nécessaire de construire des théories que l'on peut mettre en défaut expérimentalement : théories "*falsifiables*" au sens de Popper (Karl Popper 1902-1994).