

Statistiques

par **Alain LAMBOLEY**
*Ancien élève de l'École Polytechnique
Ingénieur Principal de l'Armement*

1. Statistique descriptive. Traitement des données	A 166 - 2
2. Modèle théorique de distribution. Variables aléatoires d'échantillonnage.....	— 6
3. Estimation	— 8
4. Tests d'hypothèse	— 12
5. Tests d'ajustement	— 13
6. Tests paramétriques	— 16
7. Tests non paramétriques	— 21
8. Analyse de la variance.....	— 24
9. Corrélation et régression	— 28
10. Généralisation de l'analyse de la variance	— 32
11. Contrôles statistiques industriels	— 36
Pour en savoir plus.....	Doc. A 166

Les statistiques peuvent être définies comme l'ensemble des techniques qui permettent d'étudier une vaste collection d'éléments, en limitant l'étude à une fraction bien choisie de cette collection.

1. Statistique descriptive. Traitement des données

1.1 Population et échantillon

On appellera *population* la collection d'éléments que l'on envisage, et *échantillon* la fraction d'éléments qui va permettre l'étude d'un caractère de la population, soit qualitatif soit quantitatif. L'opération permettant de constituer l'échantillon s'appellera *prélèvement*. Le nombre d'éléments de l'échantillon est appelé *taille de l'échantillon*.

Pour constituer un échantillon représentatif de la population, le prélèvement doit obéir à certaines règles. La principale est que le prélèvement soit tel qu'il offre à chaque élément de la population une chance égale d'être prélevé. Il est évident, en effet, que, si certaines catégories d'éléments ont des probabilités d'être prélevées supérieures à celles des autres, l'échantillon ne sera représentatif que de ces catégories. Un prélèvement répondant à cette condition est appelé *prélèvement au hasard*.

Effectuer un prélèvement au hasard est en général une opération difficile.

Dans un premier cas, les éléments de la population pourront être rassemblés en un même lieu. Par exemple, des petites pièces mécaniques seront mises dans une caisse, et on prélèvera chaque élément après un brassage soigneux. Toutefois cette méthode, au reste imparfaite, n'est pas toujours réalisable.

Dans un deuxième cas, on pourra numéroté les éléments, reporter ces numéros sur des jetons, et effectuer un tirage au sort de ces jetons. Une autre méthode, plus rigoureuse, consistera à utiliser les tables de nombres au hasard (ces tables présentent des lignes successives de chiffres obtenus au hasard par des procédés analogues à celui qui consisterait à mettre dix boules numérotées de 0 à 9 dans une urne de Bernoulli et à faire des prélèvements successifs). Dans ce cas, chaque élément est numéroté, on prend une suite de nombres dans la table qui permet de désigner les éléments à prélever comme le montre l'exemple ci-dessous.

Exemple

Supposons que l'on veuille prélever au hasard 20 pièces dans un lot de 1 000 pièces. Nous numérotions les pièces de 000 à 999, puis, dans un endroit quelconque de la table, nous prendrons une suite de 60 nombres, soit par exemple : 549727206819403 etc.

On prélèvera alors les pièces 549, puis 727, puis 206, puis 819, etc.

Remarque : pour répondre à la règle énoncée au 3^e alinéa un prélèvement au hasard devra nécessairement être non exhaustif. Toutefois, si l'échantillon ne représente qu'une fraction faible de la population (en pratique inférieure à 0,1), un prélèvement exhaustif pourra, en première approximation, être considéré comme un prélèvement au hasard.

1.2 Distributions statistiques

1.2.1 Mise en ordre des données

En possession d'un échantillon d'éléments, on détermine sur chacun d'eux la valeur de la caractéristique étudiée. On obtiendra ainsi un certain nombre de données qu'il est nécessaire de mettre en ordre. En effet, les résultats ont été obtenus l'un après l'autre, et la série chronologique ainsi constituée n'est pas exploitable.

Dans les cas les plus simples, les résultats seront présentés dans un tableau analogue au tableau 1.

Exemple : contrôle qualitatif de pièces (tableau 1).

Tableau 1 – Contrôle qualitatif de pièces (exemple)	
Pièces bonnes	953
Pièces mauvaises	47
Total	1 000

Dans le cas d'un caractère quantitatif, on représente les résultats sous forme d'un tableau analogue au tableau 2.

Exemple : âge des ingénieurs d'une société (tableau 2).

Tableau 2 – Étude d'un caractère quantitatif d'une population (exemple)					
Âge (ans)	Nombre	Âge (ans)	Nombre	Âge (ans)	Nombre
62	2	48	6	34	4
61	0	47	0	33	12
60	1	46	2	32	9
59	1	45	2	31	9
58	1	44	3	30	12
57	3	43	2	29	9
56	0	42	4	28	6
55	1	41	5	27	8
54	1	40	3	26	15
53	1	39	1	25	9
52	3	38	7	24	3
51	3	37	3	23	5
50	2	36	3	22	1
49	0	35	4		

Si on étudie deux caractères de la même population, on pourra établir un tableau à double entrée.

Exemple : étude de la couleur des yeux et de la couleur des cheveux d'une population (tableau 3).

Tableau 3 – Étude de deux caractères quantitatifs d'une population (exemple)					
Yeux	Cheveux				Totaux
	Blonds	Bruns	Noirs	Roux	
Bleus	1 768	807	189	47	2 811
Gris-vert	946	1 387	746	53	3 132
Noirs	115	438	288	16	857
Totaux	2 829	2 632	1 223	116	6 800

Nous appellerons d'une façon générale *distribution statistique* l'ensemble ordonné des résultats obtenus.

1.2.2 Classes d'une distribution statistique

Quand les résultats obtenus sont nombreux, on les regroupe en *classes* d'intervalles égaux ou variables selon les cas. On affecte ensuite à chaque élément de la classe une valeur du caractère, comprise dans l'intervalle (en général la demi-somme des limites de la classe). Ainsi l'exemple donné dans le tableau 2 peut être résumé par le tableau 4.

Tableau 4 – Étude d'un caractère quantitatif d'une population, tableau simplifié (exemple)

Âge (ans)	Effectif	Effectif cumulé
65 > A ≥ 60	3	166
60 > A ≥ 55	6	163
55 > A ≥ 50	10	157
50 > A ≥ 45	10	147
45 > A ≥ 40	17	137
40 > A ≥ 35	18	120
35 > A ≥ 30	46	102
30 > A ≥ 25	47	56
25 > A ≥ 20	9	9
Total	166	

On appellera *fréquence d'une classe* le rapport de l'effectif de la classe à l'effectif total.

On appellera *effectif cumulé*, ou *fréquence cumulée*, la somme de l'effectif, ou de la fréquence, de la classe considérée et des effectifs, ou fréquences, des classes inférieures.

1.3 Représentations graphiques des distributions statistiques

1.3.1 Diagrammes en barres et diagrammes circulaires

Cette méthode de représentation consiste à diviser la surface d'un rectangle (*diagramme en barres*) ou d'un cercle (*diagramme circulaire*) en secteurs dont la surface est proportionnelle à la fréquence d'observations d'une valeur, ou à l'effectif d'une classe du caractère étudié.

1.3.2 Diagrammes en bâtons et polygones de fréquence

Les diagrammes en bâtons sont obtenus en portant en abscisses sur un graphique les valeurs x_1, x_2, \dots, x_n , prises par le caractère étudié, et en ordonnées les fréquences ou les effectifs f_1, f_2, \dots, f_n correspondants. Ils s'appliquent essentiellement aux distributions de caractère discontinu.

Exemple : nombre de défauts observés sur un lot de pièces de tissus (tableau 5).

Tableau 5 – Étude du nombre de défauts d'une population (exemple)

Nombre de défauts x	Nombre de pièces f
0	3
1	13
2	37
3	26
4	17
5	4
Total	100

Le diagramme en bâtons correspondant au tableau 5 est représenté par la figure 1.

Si, dans le plan x, f , on joint les points représentant les couples $(x_1, f_1) \dots (x_n, f_n)$, on obtient le *polygone des fréquences*.

1.3.3 Histogrammes

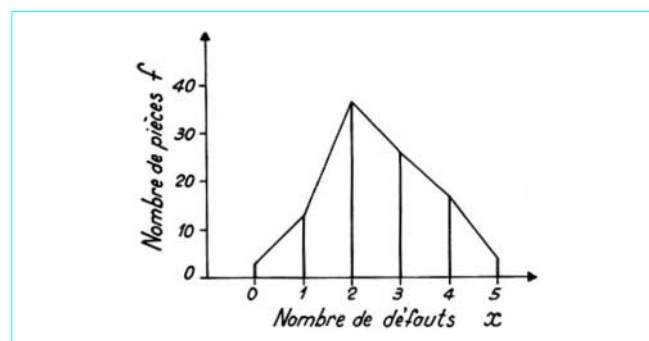
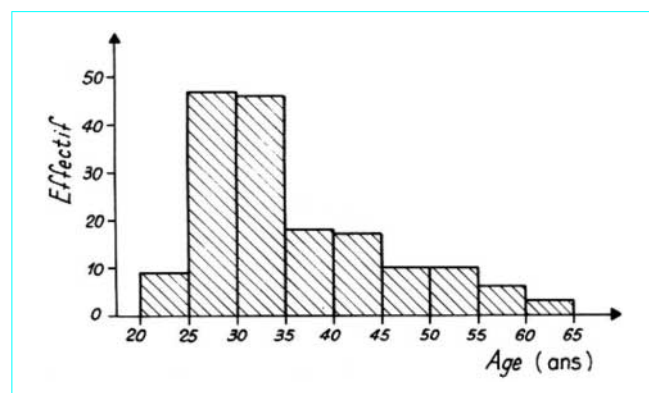
L'histogramme est obtenu en portant sur un graphique en abscisses les intervalles de classes, et en construisant sur ces intervalles de classes des rectangles de surface proportionnelle à l'effectif de la classe considérée. Lorsque les intervalles de classes sont égaux, l'histogramme est une variante du diagramme en bâtons.

L'histogramme correspondant à l'exemple du tableau 2 (âge des ingénieurs d'une société) est représenté par la figure 2.

On utilise également les *histogrammes cumulatifs* (figure 3, d'après l'exemple du tableau 2), obtenus comme les histogrammes simples, mais à partir des effectifs ou des fréquences cumulées.

L'histogramme est extrêmement simple à établir et présente un intérêt considérable ; aussi sera-t-il indispensable de l'établir avant chaque étude statistique.

Il donnera l'allure générale de la distribution statistique, et permettra de déceler certaines anomalies, soit dans la population étudiée, soit dans l'échantillonnage, soit même dans l'obtention des données.

**Figure 1 – Diagramme en bâtons et polygone des fréquences****Figure 2 – Histogramme**

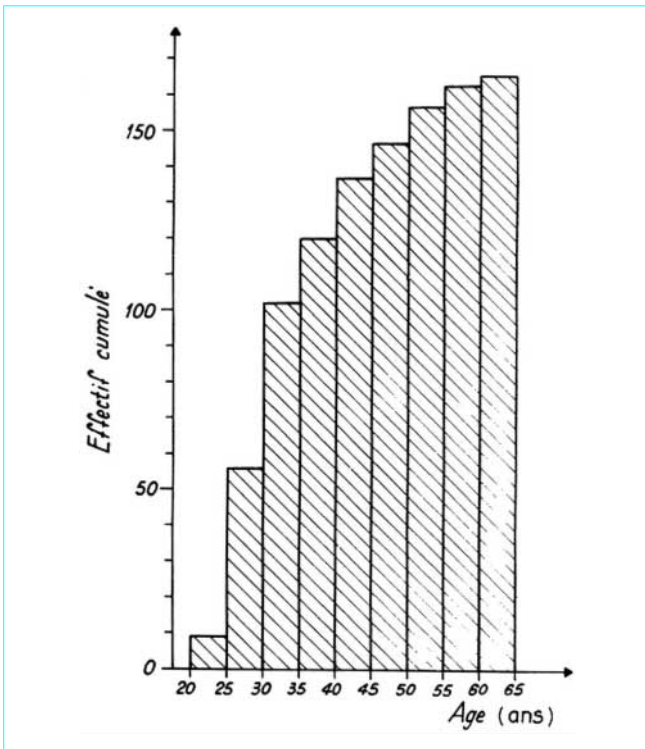


Figure 3 – Histogramme cumulatif

Par exemple, un histogramme présentant plusieurs maximums est très souvent l'indice d'une population hétérogène obtenue par un mélange de plusieurs populations. Dans d'autres cas, on observe des effectifs anormalement élevés dans certaines classes correspondant soit à des valeurs simples de la mesure (par exemple, la classe de mesure 10 est gonflée, au détriment des classes 9,9 et 10,1), soit à des valeurs voisines et intérieures aux tolérances imposées. L'histogramme révèle ainsi la falsification, inconsciente ou volontaire, des résultats par l'opérateur chargé des mesures, qui arrondit à un nombre simple les résultats obtenus ou repousse des mesures *hors-tolérances* à l'intérieur des limites de contrôle.

1.4 Paramètres caractéristiques de la distribution statistique

La distribution statistique rassemble un nombre élevé d'informations élémentaires. De ce fait, elle est d'un maniement assez lourd pour caractériser la population. C'est pourquoi on définit un certain nombre de paramètres qui doivent avoir les propriétés suivantes :

- ils tiennent compte de toutes les données ;
- ils sont peu variables d'un échantillon à un autre ;
- ils sont simples à obtenir et se prêtent au calcul algébrique.

Une première catégorie de paramètres rend compte de la valeur dominante ou de la tendance centrale de la population (§ 1.4.1). Une deuxième catégorie caractérise la dispersion ou l'étalement (§ 1.4.2).

1.4.1 Paramètres caractérisant la tendance centrale

Les paramètres les plus utilisés sont le mode, la médiane et les quantiles, le milieu de l'étendue, et la moyenne. Ces paramètres sont confondus quand la distribution est rigoureusement symétrique.

1.4.1.1 Mode

C'est la valeur du caractère mesuré qui est observée le plus souvent. Autrement dit, c'est l'abscisse correspondant au point d'ordonnée maximale du diagramme en bâtons ou de l'histogramme. Ce maximum peut n'être d'ailleurs qu'un maximum relatif et une distribution peut posséder plusieurs modes ; elle est dite alors *plurimodale*.

Nota : nous avons déjà indiqué qu'une population plurimodale était souvent constituée par le mélange de plusieurs populations (§ 1.3.3).

1.4.1.2 Médiane et quantiles

La médiane est la valeur du caractère qui sépare la distribution ordonnée en deux sous-ensembles d'effectifs égaux. Dans l'exemple donné dans le tableau 2, la médiane est égale à 32. Cela signifie que la moitié des ingénieurs sont âgés de moins de 32 ans.

On définit également les *quantiles* qui sont les valeurs du paramètre, qui divisent la distribution en quatre sous-ensembles d'effectifs égaux ; on définit de même les *déciles*, les *centiles*, etc. D'une façon plus générale, on appelle *quantiles d'ordre v*, les $v - 1$ valeurs du caractère qui partagent la distribution en v sous-ensembles d'effectifs égaux. Ces paramètres sont très utilisés dans les statistiques démographiques.

1.4.1.3 Milieu de l'étendue

Il est égal à la demi-somme des valeurs extrêmes prises par la distribution. Il est souvent appelé *midrange*, ayant fait l'objet de nombreux travaux de statisticiens anglais. Il est clair que ce paramètre ne possède aucune des propriétés énoncées plus haut, mise à part la simplicité de son calcul. Toutefois, pour cette raison, il peut être employé dans certains contrôles en cours de fabrication, sous réserve que la taille de l'échantillon soit très faible (en pratique inférieure à 10).

1.4.1.4 Moyenne

Soit x_i la valeur du caractère prise avec l'effectif n_i . La *moyenne* \bar{x} est donnée par l'expression :

$$\bar{x} = \frac{\sum_i n_i x_i}{\sum_i n_i} = \frac{\sum_i n_i x_i}{n} \text{ en posant } n = \sum_i n_i$$

On comprend tout de suite l'importance de la moyenne en notant l'analogie de ce paramètre avec le moment d'ordre 1 d'une loi de probabilité théorique ([A 165] *Probabilités*).

1.4.2 Paramètres caractérisant la dispersion

Les principaux paramètres utilisés sont l'étendue, la variance et l'écart-type, l'écart-moyen, et les interquantiles.

1.4.2.1 Étendue

Elle est égale à la différence des valeurs extrêmes prises par la distribution. On la trouve souvent dans la littérature sous le nom de *range*. Comme pour le milieu de l'étendue, ce paramètre ne rend compte de la dispersion que dans le cas d'échantillons de taille très faible.

1.4.2.2 Variance et écart-type

Avec les notations du paragraphe 1.4.1.4, la *variance* est donnée par l'expression :

$$\text{Var}(x) = \frac{\sum_i n_i (x_i - \bar{x})^2}{n}$$

La racine carrée de la variance s'appelle l'*écart-type* ou encore l'*écart quadratique moyen*.

■ **Remarque** : calcul pratique de la variance et de l'écart-type

L'expression précédente ne doit **jamais** être utilisée pour effectuer le calcul de la variance ; il est plus simple et plus précis d'utiliser la formule suivante :

$$\text{Var}(x) = \frac{\sum_i n_i x_i^2}{n} - \bar{x}^2$$

En effet :

$$\text{Var}(x) = \frac{\sum_i n_i (x_i - \bar{x})^2}{n} = \frac{\sum_i n_i x_i^2}{n} - \frac{2\bar{x} \sum_i n_i x_i}{n} + \frac{\bar{x}^2 \sum_i n_i}{n}$$

or $\frac{\sum_i n_i x_i}{n} = \bar{x}$ et $\sum_i n_i = n$

Comme la moyenne, la variance des données présente une analogie avec le moment centré d'ordre deux d'une loi de probabilité théorique ([A 165] *Probabilités*). Ceci explique l'importance considérable de la variance.

1.4.2.3 Écart-moyen

Il est égal à la moyenne des valeurs absolues des écarts des mesures à la moyenne \bar{x} :

$$e_m = \frac{\sum_i n_i |x_i - \bar{x}|}{n}$$

Ce paramètre est d'un emploi de moins en moins répandu, l'utilisation des valeurs absolues limitant beaucoup son exploitation mathématique.

1.4.2.4 Interquartiles

On peut encore caractériser la dispersion par les interquartiles, qui sont définis par l'intervalle séparant les deux quantiles extrêmes.

Nota : les quantiles ont été définies au paragraphe 1.4.1.2.

1.5 Calcul des paramètres d'une distribution statistique

Nous avons vu (§ 1.2.2) qu'il était commode de regrouper une distribution en classes. Nous allons montrer que l'on peut se servir de cette nouvelle présentation pour calculer plus aisément les paramètres de la distribution, aux dépens d'une très minime perte de précision.

1.5.1 Valeur à affecter aux éléments d'une classe

Elle dépendra de la nature de la distribution statistique envisagée.

Par exemple, dans le cas d'une distribution à caractère continu, l'intervalle de classe est égal à l'unité, et la valeur à affecter sera le milieu de l'intervalle.

Par contre, dans le cas d'une distribution discontinue, il en sera autrement. Supposons, par exemple, qu'on fasse une étude statistique sur le nombre de défauts n de pièces de tissus. La classe $5 \leq n < 10$ renferme des pièces possédant 5, 6, 7, 8, 9 défauts. La valeur à donner à cette classe sera 7 (ce qui est d'ailleurs physiquement beaucoup plus significatif que 7,5).

Quelques difficultés se poseront également au niveau des classes extrêmes qui peuvent regrouper des valeurs très dispersées.

1.5.2 Changement d'échelle et changement d'origine

Même après regroupement en classes de la distribution, il peut être très bénéfique d'effectuer un changement d'origine et un changement d'échelle, consistant par exemple à affecter à la classe centrale une valeur nulle et à l'intervalle de classe une valeur égale à l'unité. Ceci sera précisé dans l'exemple traité au paragraphe 1.5.4.

1.5.3 Présentation des calculs

On ne saurait trop insister sur la nécessité de présenter les calculs sur un tableau bien ordonné. Le maximum d'ordre entraînera le minimum d'erreurs, et, dans ces conditions, même des calculs sur un échantillon de grande taille pourront être effectués *à la main* dans un temps réduit.

1.5.4 Exemple de calcul de la moyenne et de l'écart-type

Reprenons l'exemple du tableau 2. Pour déterminer la moyenne et l'écart-type, nous présenterons les résultats sous la forme du tableau 6.

Tableau 6 – Calcul de la moyenne et de l'écart-type (exemple)

Classe	x_i	u_i	n_i	$n_i u_i$		$n_i u_i^2$
				+	-	
65 > A ≥ 60	62,5	4	3	12	48
60 > A ≥ 55	57,5	3	6	18	54
55 > A ≥ 50	52,5	2	10	20	40
50 > A ≥ 45	47,5	1	10	10	10
45 > A ≥ 40	42,5	0	17
40 > A ≥ 35	37,5	-1	18	- 18	18
35 > A ≥ 30	32,5	-2	46	- 92	184
30 > A ≥ 25	27,5	-3	47	- 141	423
25 > A ≥ 20	22,5	-4	9	- 36	144
Totaux	166	60	- 287	921

Nous avons effectué le changement de variable suivant :

$$u_i = \frac{x_i - 42,5}{5}$$

ou

$$x_i = 5 u_i + 42,5$$

Dans ces conditions, on peut aisément calculer la moyenne :

$$\begin{aligned}\sum_i n_i u_i &= -227 \\ \bar{u} &= \frac{-227}{166} \approx -1,37 \\ \bar{x} &= 5\bar{u} + 42,5 \approx 35,7 \\ \text{Var}(u) &= \frac{\sum_i n_i u_i^2}{n} - \bar{u}^2 \approx 3,68 \\ \text{Var}(x) &= 25 \text{Var } u \approx 92\end{aligned}$$

et l'écart-type s sera donné par :

$$s = \sqrt{\text{Var}(x)} \approx 9,6$$

Les résultats obtenus sont donc :

- âge moyen $\approx 35,7$ ans ;
- variance $\approx 92,25$;
- écart-type $\approx 9,6$.

Nous nous sommes livrés au calcul fastidieux de la moyenne et de l'écart-type sur la distribution brute. Nous avons ainsi trouvé :

- âge moyen ≈ 35 ans ;
- variance $\approx 94,5$.

Compte tenu de la précision avec laquelle les âges ont été donnés, les résultats obtenus plus haut sont largement satisfaisants.

1.6 Distribution statistique empirique et lois de probabilité

Le lecteur sera sans doute frappé par le fait que nous avons utilisé dans ce paragraphe des dénominations telles que moyenne et variance, déjà utilisées en calcul des probabilités ([A 165] *Probabilités*). Cependant, il n'y a pas de confusion. En effet, le calcul des probabilités définit les lois de probabilité de variables aléatoires ; en statistique, quand on mesure un caractère sur des éléments prélevés au hasard, on réalise en fait un événement au sens probabiliste, et la mesure effectuée est une variable aléatoire dont on relève la fréquence. Par le biais des fréquences, et en s'appuyant sur la loi des grands nombres, on voit immédiatement l'usage du calcul des probabilités dans les statistiques.

L'objet des statistiques sera par conséquent de déterminer, à partir des données de l'échantillon, la loi de probabilité de la variable aléatoire représentant le caractère étudié.

Il existe, dans ces conditions, un parallélisme rigoureux entre lois de probabilité et distributions statistiques empiriques, entre probabilités et fréquences, variables aléatoires et mesures, moyenne d'une loi de probabilité et moyenne d'échantillons.

2. Modèle théorique de distribution. Variables aléatoires d'échantillonnage

2.1 Généralités

Dans le paragraphe 1, nous avons décrit les procédés qui permettent de donner une représentation graphique et paramétrique d'une distribution ; nous n'avons toutefois fait aucune hypothèse sur la forme de la distribution. Nous savons déjà que nous ne pourrions, en général, connaître une population qu'au travers d'un échantillon ; par conséquent, il sera nécessaire de faire une hypothèse sur la

forme de la population, c'est-à-dire de choisir un modèle théorique de distribution permettant de remonter de la connaissance de l'échantillon à la connaissance de la population.

Le processus d'étude statistique d'une population sera le suivant :

- prélèvement d'un échantillon ;
- étude descriptive de l'échantillon mis sous forme d'une distribution ordonnée ;
- choix d'un modèle théorique de distribution, l'échantillon étant par hypothèse considéré comme représentatif de la population ; nous avons déjà vu, dans l'article [A 165] *Probabilités*, un certain nombre de lois de probabilité parmi lesquelles nous sélectionnerons le modèle le plus adapté, en fonction des résultats obtenus sur l'échantillon ; nous verrons (§ 5) comment justifier ce choix qui peut être *a priori* assez arbitraire ;
- le choix d'un modèle théorique va permettre alors d'obtenir, à partir des résultats de l'échantillon, des données beaucoup plus élaborées sur la population.

2.2 Choix du modèle de distribution

Le choix d'un modèle de distribution n'obéit pas à des règles bien établies ; en général, il résultera de :

- l'étude d'un échantillon de grande taille fourni par la population : allure de l'histogramme, ou du diagramme en bâtons, juxtaposition et comparaison des fréquences observées avec les fréquences fournies par une loi théorique, etc ;
- l'étude des causes de la dispersion du caractère étudié. Par exemple, si on s'intéresse à la cote d'un assemblage mécanique complexe, il est à peu près certain que la loi de probabilité de cette cote sera normale, car la dispersion est la somme des dispersions sur les cotes des pièces élémentaires (application du théorème central limite).

D'une façon générale, la grande majorité des variables aléatoires continues suit une loi de probabilité normale (signalons cependant que les mesures d'excentricités de pièces circulaires suivent une loi du χ^2). Quant aux variables aléatoires discontinues, on ne rencontre guère que des variables relevant, soit de la loi de Poisson, soit de la loi binomiale.

2.3 Échantillon, variable aléatoire

Supposons que l'on étudie une population dont chaque élément est, par exemple, caractérisé par une valeur variant continûment. Supposons, de plus, que la distribution de ce caractère soit de nature connue, par exemple normale, et que l'on connaisse également les paramètres de cette distribution.

Prélevons un échantillon de n éléments dans cette population ; nous savons qu'il existe dans la population une infinité d'échantillons différents de taille n . L'ensemble de ces échantillons constitue également une population, et on peut considérer chaque échantillon comme une variable aléatoire dans un espace à n dimensions. La loi de probabilité de cette variable aléatoire pourra se déduire de la loi de probabilité correspondant à la population initiale.

Ainsi, si les n mesures obtenues sur l'échantillon sont $x_1, x_2, \dots, x_i, \dots, x_n$ et si la distribution de la population-mère admet une densité de probabilité $f(x)$, la loi de probabilité de cet échantillon, considéré comme une variable aléatoire, aura comme probabilité élémentaire : $f(x_1) f(x_2) \dots f(x_n) dx_1 dx_2 \dots dx_n$.

Cela suppose évidemment que les prélèvements de la population soient considérés comme des événements aléatoires indépendants.

De même, les paramètres de l'échantillon pourront être considérés comme des variables aléatoires, et les lois de probabilité de ces variables seront tirées de la loi de probabilité correspondant à la population initiale, en partant de l'expression précédente. Il va de soi que ces lois dépendront également de la taille de l'échantillon.

Nous allons développer ces considérations, et nous emploierons les notations suivantes, dans la suite de l'article :

— moyenne de la population : m ([A 165] *Probabilités*) ;

— moyenne de l'échantillon : $\bar{x} = \frac{\sum_i x_i}{n}$;

— variance de la population : σ^2 ([A 165] *Probabilités*) ;

— variance de l'échantillon : $s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n}$.

2.4 Moyenne de l'échantillon, variable aléatoire

2.4.1 Espérance mathématique de la moyenne

Quelle que soit la forme de la loi de probabilité de la population-mère, l'espérance mathématique de \bar{x} sera m ; en effet, considérons N échantillons, de taille n , de moyenne \bar{x}_i :

$$E[\bar{x}] = \frac{\bar{x}_1 + \dots + \bar{x}_i + \dots + \bar{x}_N}{N}$$

qui tend vers m quand N tend vers l'infini.

2.4.2 Variance de la moyenne

La variance de \bar{x} peut se mettre sous la forme :

$$\begin{aligned} \text{Var}(\bar{x}) &= E[\bar{x} - m]^2 = E\left[\frac{\sum_i x_i}{n} - m\right]^2 = E\left[\frac{\sum_i (x_i - m)}{n}\right]^2 \\ &= \frac{1}{n} E\left[\frac{\sum_i (x_i - m)^2}{n}\right] = \frac{\sigma^2}{n} \end{aligned}$$

2.5 Variance de l'échantillon, variable aléatoire

On montre que la variance de l'échantillon a pour espérance mathématique :

$$E[s^2] = \frac{n-1}{n} \sigma^2$$

$E[s^2]$ tend vers σ^2 quand la taille de l'échantillon croît indéfiniment.

La variance de s^2 , $\text{Var}(s^2)$, a une forme beaucoup plus complexe.

2.6 Cas d'une population normale

Supposons que la loi de probabilité de la population soit normale, de moyenne m et d'écart-type σ .

Considérons un échantillon de taille n , dont la moyenne et l'écart-type sont donnés par :

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_i x_i \\ s &= \sqrt{\frac{1}{n} \sum_i (x_i - \bar{x})^2} \end{aligned}$$

On montre que :

— les variables aléatoires \bar{x} et s sont indépendantes ;

— la loi de probabilité de \bar{x} est normale, de moyenne m et d'écart-type $\frac{\sigma}{\sqrt{n}}$;

— la quantité $\frac{\sum_i (x_i - m)^2}{\sigma^2}$ suit une loi de χ^2 à n degrés de liberté ;

— la quantité $\frac{ns^2}{\sigma^2}$ suit une loi de χ^2 à $n-1$ degrés de liberté.

Or on sait ([A 165] *Probabilités*) qu'une loi de χ^2 à $n-1$ degrés de liberté a pour espérance mathématique et variance :

$$E[\chi^2] = n-1$$

$$\text{Var}(\chi^2) = 2(n-1)$$

d'où

$$E[s^2] = \frac{n-1}{n} \sigma^2$$

$$\text{Var}(s^2) = \frac{2(n-1)}{n^2} \sigma^4$$

On en déduit également, en ce qui concerne l'écart-type :

$$E[s] = \sqrt{\frac{2}{n}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sigma = b_n \sigma$$

où Γ est la fonction eulérienne de 2^e espèce ; notons que b_n tend vers 1 quand n tend vers l'infini ;

• en première approximation, l'écart-type de s vaut $\frac{\sigma}{\sqrt{2n}}$;

• la variable $t = \sqrt{n-1} \times \frac{\bar{x} - m}{s}$ suit une loi de Student-Fisher à $n-1$ degrés de liberté.

■ Remarque

Le lecteur pourra s'étonner que la quantité $\frac{\sum_i (x_i - m)^2}{\sigma^2}$ suive une loi du χ^2 à n degrés de liberté, alors que $\frac{ns^2}{\sigma^2}$ suit une loi du χ^2 à $n-1$ degrés de liberté.

L'explication est simple : en effet, si on remonte à la définition de la variable χ^2 à n degrés de liberté ([A 165] *Probabilités*), on voit que celle-ci est obtenue comme la somme des carrés de n variables normales réduites indépendantes.

Or si les n variables $\frac{x_i - m}{\sigma}$ sont indépendantes, seuls $n-1$ variables $\frac{x_i - \bar{x}}{\sigma}$ sont indépendantes ; en effet, si on connaît $n-1$

de ces variables, la relation $\bar{x} = \frac{\sum_i x_i}{n}$ permet d'obtenir la $n^{\text{ème}}$.

Par conséquent, la quantité :

$$\frac{\sum_i (x_i - \bar{x})^2}{\sigma^2} = \frac{ns^2}{\sigma^2}$$

suit une loi du χ^2 à $n-1$ degrés de liberté.

2.7 Cas d'un échantillon de taille élevée

Si l'échantillon issu d'une population quelconque est de taille élevée, on peut dire en première approximation :

- la variable \bar{x} suit une loi normale, de moyenne m et d'écart-type $\frac{\sigma}{\sqrt{n}}$;
- la variable s suit une loi normale, de moyenne σ et d'écart-type $\frac{\sigma}{\sqrt{2n}}$.

3. Estimation

3.1 Généralités

Ayant choisi, pour la population étudiée, un modèle de distribution théorique (§ 2), il est nécessaire d'en déterminer les paramètres. Faute de pouvoir prendre en compte la population complète, il sera impossible, en général, de connaître les paramètres réels de la population ; il sera donc nécessaire de les approcher à partir d'un échantillon.

Remarque : si on a pu obtenir d'une même population un grand nombre d'échantillons, les paramètres moyens que l'on en a tirés pourront être considérés comme ceux de la population (§ 11).

La première approche pour déterminer les paramètres de la population consistera à prendre les paramètres de l'échantillon ; il n'est toutefois pas évident qu'on obtienne ainsi la meilleure représentation des paramètres réels de la population. En faisant subir à ceux-ci certaines opérations, il est possible de mieux cerner les paramètres de la population. Ces opérations portent le nom d'*estimation*, et on appelle également *estimation* le résultat obtenu ; l'opérateur est appelé *estimateur*. L'estimation la plus élémentaire consiste évidemment à prendre le paramètre même de l'échantillon.

3.2 Qualités d'un bon estimateur

Un estimateur dépend évidemment de la valeur du paramètre de l'échantillon, et ainsi constitue une variable aléatoire dont on peut obtenir la loi de probabilité à partir de celle qui correspond à la population et de la taille de l'échantillon.

On obtiendra un bon estimateur t d'un paramètre θ de la population s'il remplit les conditions énoncées dans les paragraphes 3.2.1, 3.2.2, 3.2.3, 3.2.4 et 3.2.5.

3.2.1 Convergence

Un estimateur t est dit *convergent* ou *correct* s'il converge en probabilité vers le paramètre θ de la population, quand la taille n de l'échantillon croît indéfiniment, ce qui se traduit par :

η et ε étant quelconques, on peut trouver une valeur N telle que $n > N$ entraîne $P(|t - \theta| < \eta) > 1 - \varepsilon$

Il est utile d'utiliser un estimateur convergent car plus la taille de l'échantillon croît, plus l'estimateur rend compte de la vraie valeur du paramètre.

3.2.2 Estimateur sans biais

Un estimateur est dit *sans biais* ou *sans distorsion* si son espérance mathématique est égale au paramètre à estimer de la population, quelle que soit la taille de l'échantillon.

L'utilité d'un estimateur sans biais est évidente ; en effet, si on a prélevé un grand nombre d'échantillons de la même population, la moyenne des estimations obtenues donnera une valeur assez exacte du paramètre de la population.

Un estimateur convergent et sans biais est dit *absolument correct*.

3.2.3 Estimateur efficace

Un estimateur est dit *efficace* s'il est absolument correct et, de plus, si, parmi tous les estimateurs absolument corrects, il possède la variance la plus faible.

3.2.4 Estimateur exhaustif

On dit qu'un estimateur est *exhaustif* s'il rend compte de toute l'information contenue dans l'échantillon. Ainsi l'étendue (§ 1.4.2.1), qui ne prend en compte que les valeurs extrêmes de l'échantillon, n'est pas un estimateur exhaustif de la dispersion.

3.2.5 Facilités de calcul

Cette qualité évidente n'est pas souvent compatible avec les autres.

3.3 Méthode d'obtention des estimateurs

Soient x_1, x_2, \dots, x_n les valeurs d'une variable aléatoire X obtenues sur un échantillon. On cherche à déterminer une estimation t du paramètre θ de la population. La probabilité de prélever l'échantillon précédent peut se mettre sous la forme :

$$\mathcal{L} = P(x_1) \cdot P(x_2) \dots P(x_n)$$

si on note $P(x_i) = P(X = x_i)$.

$P(x)$ dépend de la loi de probabilité correspondant à la population, en principe connue, et du paramètre réel inconnu θ ; \mathcal{L} s'écrit donc :

$$\mathcal{L} = p(x_1, \theta) \cdot p(x_2, \theta) \dots p(x_n, \theta)$$

La fonction \mathcal{L} s'appelle *fonction de vraisemblance* ; on démontre que l'on peut obtenir un bon estimateur de θ en prenant la fonction $\theta(x_1, \dots, x_n)$ qui rend \mathcal{L} maximale (d'où le nom de *méthode du maximum de vraisemblance* donné à ce procédé). On obtient θ en résolvant l'équation : $\frac{d\mathcal{L}}{d\theta} = 0$.

Nous ne nous étendrons pas plus sur cette méthode dont les développements sont hors de propos avec cet exposé.

3.4 Estimation par intervalle

Nous avons recherché, par les méthodes exposées aux paragraphes 3.2 et 3.3, une estimation d'un paramètre de la population, c'est-à-dire une certaine valeur qui représente le mieux le paramètre de la population, compte tenu des aléas d'échantillonnage. Nous savons déjà (§ 3.2) que cette valeur est elle-même une variable aléatoire, et qu'elle possède une loi de probabilité ; notre intention est maintenant de situer, en terme de probabilités, la vraie valeur du paramètre par rapport à son estimateur, c'est-à-dire de déterminer, autour de l'estimation, un intervalle dans lequel le paramètre de la population a une probabilité $1 - \alpha$ de se trouver (α étant faible).

Cet intervalle portera le nom d'*intervalle de confiance* et constituera ce qu'on appelle ainsi l'*estimation par intervalle*, par opposition à l'*estimation ponctuelle* exposée dans les paragraphes 3.2 et 3.3. L'estimation par intervalle mesure, en quelque sorte, la précision de l'estimation ponctuelle.

3.5 Estimations dans le cas de lois de probabilité classiques

3.5.1 Cas de la loi binomiale et de la loi de Poisson

3.5.1.1 Estimation d'une proportion

Sur un prélèvement de taille n , on observe k éléments identiques. Soit p la proportion inconnue avec laquelle figure cette valeur au sein de la population.

La quantité $\frac{k}{n}$ converge, nous le savons, vers p , quand n augmente indéfiniment (théorème de Bernoulli) et $E\left[\frac{k}{n}\right] = p$. En outre, $\text{Var}\left(\frac{k}{n}\right) = \frac{p(1-p)}{n}$ tend vers zéro, quand n tend vers l'infini.

La quantité k/n est donc un estimateur absolument correct de p (§ 3.2.2).

3.5.1.2 Intervalle de confiance

3.5.1.2.1 Cas d'un échantillon de taille élevée

Nous savons que, dans ce cas, la quantité k/n suit approximativement une loi normale ([A 165] *Probabilités*), de moyenne p et

d'écart-type $\sqrt{\frac{p(1-p)}{n}}$.

Dans ces conditions, on aura :

$$P\left\{p - u_{\alpha} \sqrt{\frac{p(1-p)}{n}} \leq \frac{k}{n} \leq p + u_{\alpha} \sqrt{\frac{p(1-p)}{n}}\right\} = 1 - \alpha$$

u_{α} étant la valeur de la variable normale réduite u telle que :

$$P(|u| > u_{\alpha}) = \alpha$$

$$\text{et } P\left\{\frac{k}{n} + u_{\alpha} \sqrt{\frac{p(1-p)}{n}} \geq p \geq \frac{k}{n} - u_{\alpha} \sqrt{\frac{p(1-p)}{n}}\right\} = 1 - \alpha$$

En substituant k/n à p dans les quantités sous radical, on obtient une première approximation de l'intervalle de confiance, qui sera donnée par l'expression :

$$\frac{k}{n} \pm u_{\alpha} \sqrt{\frac{1}{n} \cdot \frac{k}{n} \left(1 - \frac{k}{n}\right)}$$

3.5.1.2.2 Cas d'un échantillon de taille réduite

Si p est connu, on peut, avec la loi binomiale, trouver les valeurs k_1 et k_2 telles que :

$$P(k \leq k_1) = \frac{\alpha}{2}$$

$$\text{et } P(k \leq k_2) = 1 - \frac{\alpha}{2}$$

Nota : en fait, on ne trouvera pas, en général, de valeurs de k correspondant exactement au seuil α , du fait de la discontinuité de la loi binomiale, mais on utilise les valeurs de k correspondant au seuil α' le plus proche de α .

Pour chaque valeur de p , on pourra déterminer ainsi deux valeurs de k/n . En portant p en abscisses et k/n en ordonnées, on pourra tracer deux courbes C_1 et C_2 .

Réciproquement, si p est inconnu, et si on a obtenu sur un échantillon l'estimation k/n , en traçant sur le graphique une parallèle à l'axe des abscisses, d'ordonnée k/n , on obtiendra, par intersection avec les courbes C_1 et C_2 , l'intervalle de confiance, comme le montre la figure 4.

Ces courbes ont été établies pour des valeurs de n comprises entre 5 et 100, aux seuils $1 - \alpha = 0,95$ et $0,998$.

Nota : si p est faible et n élevé, il est possible de tracer les courbes C_1 et C_2 à partir de la loi de Poisson.

3.5.2 Cas de la loi normale

3.5.2.1 Estimation et intervalle de confiance de la variance

3.5.2.1.1 Cas où la moyenne m de la population est connue

Nous savons (§ 2.6) que la quantité $\frac{\sum_i (x_i - m)^2}{\sigma^2}$ suit une loi du χ^2 à n degrés de liberté ; par conséquent :

$$E\left[\frac{\sum_i (x_i - m)^2}{\sigma^2}\right] = n$$

$$\text{et } E\left[\frac{\sum_i (x_i - m)^2}{n}\right] = \sigma^2$$

Notons que :

$$\text{Var}\left(\frac{\sum_i (x_i - m)^2}{\sigma^2}\right) = 2n$$

$$\text{et } \text{Var}\left(\frac{\sum_i (x_i - m)^2}{n}\right) = \frac{2\sigma^4}{n}$$

Donc, si m est connue, la quantité $\frac{\sum_i (x_i - m)^2}{n}$ constitue un estimateur sans biais de la variance (§ 3.2.2).

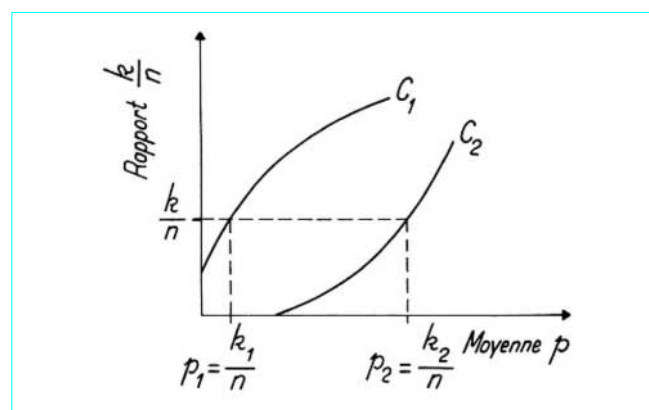


Figure 4 – Intervalle de confiance d'une proportion pour une loi binomiale, dans le cas où la taille de l'échantillon est faible

L'intervalle de confiance est très facilement déterminé par l'expression :

$$\left\{ \chi^2_{\alpha/2}(n) \leq \frac{\sum_i (x_i - m)^2}{\sigma^2} \leq \chi^2_{1-(\alpha/2)}(n) \right\} = 1 - \alpha$$

où $\chi^2_{\alpha/2}(n)$ et $\chi^2_{1-(\alpha/2)}(n)$ sont définis par :

$$P\{\chi^2(n) < \chi^2_{\alpha/2}(n)\} = \frac{\alpha}{2}$$

$$P\{\chi^2(n) < \chi^2_{1-(\alpha/2)}(n)\} = 1 - \frac{\alpha}{2}$$

On en tire :

$$P\left\{ \frac{\sum_i (x_i - m)^2}{\chi^2_{\alpha/2}(n)} \geq \sigma^2 \geq \frac{\sum_i (x_i - m)^2}{\chi^2_{1-(\alpha/2)}(n)} \right\} = 1 - \alpha$$

Ceci détermine l'intervalle de confiance.

Nota : signalons qu'on ne rencontre pas souvent ce cas : il est très rare, en effet, de connaître la moyenne de la population.

3.5.2.1.2 Cas où la moyenne m de la population est inconnue

Nous savons (§ 2.6) que la quantité $\frac{ns^2}{\sigma^2} = \frac{\sum_i (x_i - \bar{x})^2}{\sigma^2}$ suit une loi du χ^2 à $n - 1$ degrés de liberté ; par conséquent :

$$E\left[\frac{ns^2}{\sigma^2}\right] = n - 1$$

et

$$E\left[\frac{ns^2}{n-1}\right] = \sigma^2$$

Notons que

$$\text{Var}\left(\frac{ns^2}{\sigma^2}\right) = 2(n-1)$$

et

$$\text{Var}\left(\frac{ns^2}{n-1}\right) = \frac{2\sigma^4}{n-1}$$

L'estimateur sans biais de la variance (§ 3.2.2) sera donc, dans ce cas, la quantité :

$$s'^2 = \frac{ns^2}{n-1}$$

L'intervalle de confiance sera donné par l'expression :

$$\frac{\sum_i (x_i - \bar{x})^2}{\chi^2_{\alpha/2}(n-1)} \geq \sigma^2 \geq \frac{\sum_i (x_i - \bar{x})^2}{\chi^2_{1-(\alpha/2)}(n-1)}$$

Nota : ce cas est beaucoup plus usuel que le précédent ; il faudra se souvenir qu'il est nécessaire d'utiliser la loi du χ^2 à $n - 1$ degrés de liberté.

Exemple

Sur une série de 10 mesures, on a trouvé les résultats suivants :

$$s'^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1} = 3$$

On recherche l'intervalle de confiance à 95 % de la variance.

Il sera donné par :

$$\frac{\sum_i (x_i - \bar{x})^2}{\chi^2_{0,025}} \geq \sigma^2 \geq \frac{\sum_i (x_i - \bar{x})^2}{\chi^2_{0,975}}$$

Avec un nombre de degrés de liberté égal à 9 on a :

$$\chi^2_{0,025} = 2,70$$

$$\chi^2_{0,975} = 19,0$$

soit

$$1,42 \leq \sigma^2 \leq 10$$

Remarque : cas d'un échantillon de taille élevée.

Les tables de la loi du χ^2 vont, en général, jusqu'à un nombre de degrés de liberté v égal à 30. Si le nombre de degrés de liberté est supérieur à 30, on utilise le résultat suivant : la variable $u = \sqrt{2\chi^2} - \sqrt{2v-1}$ est une variable normale réduite qui permet de calculer les valeurs $\chi^2_{\alpha/2}$ et $\chi^2_{1-(\alpha/2)}$.

3.5.2.2 Estimation et intervalle de confiance de la moyenne

3.5.2.2.1 Cas où la variance σ^2 de la population est connue

Nous savons déjà (§ 2.4.1 et 2.4.2) que $E[\bar{x}] = m$ et $\text{Var}(\bar{x}) = \frac{\sigma^2}{n}$; la moyenne de l'échantillon est donc un estimateur sans biais (§ 3.2.2).

Sachant en outre (§ 2.6), que \bar{x} suit une loi normale, de moyenne m et d'écart-type $\frac{\sigma}{\sqrt{n}}$, on peut écrire :

$$P\left\{ m - u_\alpha \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq m + u_\alpha \frac{\sigma}{\sqrt{n}} \right\} = 1 - \alpha$$

u_α étant la valeur d'une variable normale réduite u définie par :

$$P(|u| > u_\alpha) = \alpha$$

On en déduit :

$$P\left\{ \bar{x} - u_\alpha \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{x} + u_\alpha \frac{\sigma}{\sqrt{n}} \right\} = 1 - \alpha$$

L'intervalle de confiance est par conséquent :

$$\bar{x} \pm u_\alpha \frac{\sigma}{\sqrt{n}}$$

Exemple

Sur un échantillon de taille $n = 25$, on a trouvé $\bar{x} = 117$. L'écart-type de la population est connu, et vaut $\sigma = 7$. L'intervalle de confiance à 95 % est donné par :

$$117 \pm u_{\alpha} \times \frac{7}{5} \text{ avec } u_{\alpha} = 1,96$$

L'intervalle de confiance est approximativement : $117 \pm 2,74$.

3.5.2.2.2 Cas où la variance σ^2 de la population est inconnue

Dans ce cas, l'estimation sans biais (§ 3.2.2) de la moyenne de la population est encore la moyenne de l'échantillon. Par contre, pour calculer l'intervalle de confiance, on utilise la variable aléatoire

$t = \frac{\bar{x} - m}{\frac{s}{\sqrt{n-1}}}$ qui suit une loi de Student-Fisher à $n-1$ degrés de liberté.

On peut alors écrire :

$$P\left\{-t_{\alpha} \leq \frac{\bar{x} - m}{\frac{s}{\sqrt{n-1}}} \leq t_{\alpha}\right\} = 1 - \alpha$$

t_{α} étant définie comme la valeur d'une variable t de Student-Fisher à $n-1$ degrés de liberté, telle que :

$$P\{|t| \leq t_{\alpha}\} = 1 - \alpha$$

Cette expression devient :

$$P\left\{\bar{x} - t_{\alpha} \frac{s}{\sqrt{n-1}} \leq m \leq \bar{x} + t_{\alpha} \frac{s}{\sqrt{n-1}}\right\} = 1 - \alpha$$

L'intervalle de confiance est alors :

$$\bar{x} \pm t_{\alpha} \frac{s}{\sqrt{n-1}}$$

Rappelons que

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}$$

Si on veut utiliser l'estimation $s' = \sqrt{\frac{n}{n-1}} s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$,

l'intervalle de confiance peut alors s'écrire :

$$\bar{x} \pm t_{\alpha} \frac{s'}{\sqrt{n}}$$

Exemple

Reprenons partie des données de l'exercice précédent : $n = 25$, $\bar{x} = 117$ et $s = 7$. L'intervalle de confiance à 95 % est donné par :

$$117 \pm t_{\alpha} \times \frac{7}{\sqrt{24}}$$

Or $t_{\alpha} \approx 2,06$, et l'intervalle de confiance est approximativement : $117 \pm 2,94$.

Remarque : si la taille de l'échantillon croît, t_{α} tend vers u_{α} . Ainsi, pour n élevé, on pourra utiliser la méthode du paragraphe 3.5.2.2.1 en remplaçant σ par s ou s' (n étant grand, $s \approx s'$).

3.5.2.3 Estimation de l'écart-type

Nous avons vu (§ 2.6), que $E[s] = b_n \sigma$. Par conséquent, dans le cas d'une loi normale, on obtiendra une estimation sans biais (§ 3.2.2) de l'écart-type avec l'expression $\frac{1}{b_n} s$.

Nota : signalons, en effet, que si $\frac{n}{n-1} s^2$ est une estimation sans biais de la variance (§ 3.5.2.1.2), sa racine carrée n'est pas une estimation sans biais de l'écart-type car $\sqrt{E[u^2]} \neq E[u]$, u étant une variable aléatoire quelconque.

Rappelons que b_n tend vers 1 assez rapidement quand n croît. Rappelons par ailleurs que, si n est grand, s suit approximativement une loi normale de moyenne σ et d'écart-type $\frac{\sigma}{\sqrt{2n}}$. On pourra, dans ce cas, donner pour expression de l'intervalle de confiance :

$$s \pm u_{\alpha} \frac{s}{\sqrt{2n}}$$

3.5.2.4 Estimation de la variance à partir de deux échantillons

Supposons qu'on ait obtenu deux échantillons respectivement de taille n_1 et n_2 d'une population normale. On peut obtenir une estimation sans biais (§ 3.2.2) de la variance de la population avec l'expression :

$$s' = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

avec

$$s_1^2 = \frac{\sum_i (x_{1i} - \bar{x}_1)^2}{n_1}$$

$$s_2^2 = \frac{\sum_i (x_{2i} - \bar{x}_2)^2}{n_2}$$

En effet, si σ^2 est la variance de la population, la quantité $\frac{n_1 s_1^2}{\sigma^2}$

suit une loi du χ^2 à $n_1 - 1$ degrés de liberté, et la quantité $\frac{n_2 s_2^2}{\sigma^2}$

suit une loi du χ^2 à $n_2 - 1$ degrés de liberté (§ 2.6). Par conséquent ([A 165] *Probabilités*), la somme de ces quantités :

$$\frac{1}{\sigma^2} (n_1 s_1^2 + n_2 s_2^2)$$

suit une loi du χ^2 à $(n_1 + n_2 - 2)$ degrés de liberté, et par conséquent :

$$E\left[\frac{n_1 s_1^2 + n_2 s_2^2}{\sigma^2}\right] = n_1 + n_2 - 2$$

et

$$E\left[\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\right] = \sigma^2$$

3.6 Détermination de la taille de l'échantillon

Signalons, sans entrer dans les détails, que, réciproquement, si on se fixe l'intervalle de confiance, c'est-à-dire la précision avec laquelle on veut connaître le paramètre de la population, on pourra en tirer la taille de l'échantillon permettant d'obtenir cet intervalle de confiance.

Exemple

Quand on veut estimer une proportion p (§ 3.5.1.2.1) avec une précision de $\pm 0,02$ au seuil de 95 %, il faudra que :

$$u_{\alpha} \sqrt{\frac{p(1-p)}{n}} = 0,02 \quad \text{et} \quad u_{\alpha} = 1,96 \approx 2$$

Si p est approximativement connue, et de l'ordre de 0,2, on aura :

$$2 \sqrt{\frac{0,2 \times 0,8}{n}} \approx 0,02$$

La taille du prélèvement sera donc de l'ordre de 1 600.

4. Tests d'hypothèse

4.1 Généralités

Les résultats du paragraphe 3 permettent d'obtenir une représentation d'une population à partir d'un échantillon. Cependant, les problèmes d'estimation ne sont pas les seuls que la statistique permet de résoudre. Ainsi, on pourra tenter, par exemple, de répondre à des questions du genre des suivantes :

— *problèmes de comparaison de deux populations* : on considère deux lots de produits fabriqués. Sur chacun d'entre eux, on a effectué un prélèvement, et estimé la moyenne et la variance ; à partir de ces résultats, peut-on conclure que les deux lots sont identiques, ou, plus exactement, peut-on dire que ces deux lots constituent des populations statistiques identiques ? Nous savons que, même si ces deux lots étaient identiques, les échantillons prélevés seraient différents du fait des aléas de l'échantillonnage ;

— *problèmes de comparaison à un standard* : la qualité exigée d'un produit est donnée ; on considère un lot de ce produit dont on prélève un échantillon sur lequel on estime une qualité moyenne ; la qualité moyenne diffère de la qualité exigée ; peut-on dire que le lot répond à la qualité exigée et que l'écart entre celle-ci et la qualité estimée est dû uniquement aux aléas de l'échantillonnage ?

4.2 Définition des tests

En présence de problèmes tels que ceux posés au paragraphe 4.1, on adopte la marche suivante : par exemple, dans le premier cas, on fera l'hypothèse que les deux lots sont identiques, c'est-à-dire que les deux échantillons peuvent être considérés comme extraits d'une même population. Dans le cadre de cette hypothèse, l'écart entre leurs paramètres (par exemple l'écart entre leurs moyennes) suivra une certaine loi de probabilité, qui permettra de déterminer un intervalle dans lequel le paramètre aura une probabilité $1 - \alpha$ de se trouver, si l'hypothèse est exacte.

Si l'écart observé est compris dans cet intervalle, on conviendra d'accepter l'hypothèse d'identité des lots. Dans le cas contraire, on refusera l'hypothèse. Cette technique porte le nom de *test d'hypothèse*.

En fait, dans le cas d'acceptation, remarquons bien que le test d'hypothèse n'apporte pas une réponse affirmative absolue. Il signifie simplement que, compte tenu des résultats observés sur les échantillons, rien ne s'oppose à l'acceptation de l'hypothèse formulée.

De même, si l'écart observé n'est pas dans l'intervalle considéré, on n'a pas une certitude absolue que l'hypothèse d'identité des lots soit fausse. En effet, on sait seulement que, si l'hypothèse est exacte, il y a une probabilité α faible pour que l'écart observé soit hors de l'intervalle d'acceptation. On refuse simplement l'hypothèse parce que la probabilité pour que l'hypothèse soit exacte peut être chiffrée par α , et que l'on court, en l'acceptant, un risque, de probabilité $1 - \alpha$, bien supérieure à α , d'accepter une hypothèse fausse.

4.3 Efficacité des tests

Reprenons le deuxième exemple du paragraphe 4.1 : nous désirons comparer la qualité d'une fabrication, estimée sur un échantillon, à une qualité standard exigée. Le test de comparaison est basé sur la variable aléatoire constituée par l'écart entre l'estimation et la qualité standard. L'hypothèse à tester est que la qualité de la fabrication est identique à la qualité standard. Nous comprenons, tout de suite, que la probabilité d'accepter l'hypothèse est une fonction de la qualité réelle de la fabrication.

On appellera *efficacité du test* la fonction :

$$P = f(Q)$$

avec P probabilité d'acceptation de l'hypothèse,

Q qualité réelle de la fabrication.

D'une façon plus générale, on pourra définir l'efficacité d'un test de comparaison, en déterminant la probabilité d'acceptation de l'hypothèse d'identité des éléments comparés, en fonction de leurs écarts réels. On appellera *courbe d'efficacité* la représentation graphique de l'efficacité.

On pourra ensuite définir un risque, dit *risque de première espèce*, par la probabilité de refuser l'hypothèse d'identité alors qu'elle est exacte. De même, on déterminera un risque, dit *risque de deuxième espèce*, par la probabilité d'accepter l'hypothèse alors qu'elle est fausse.

■ Remarque

On pourra s'étonner que, dans la notion d'efficacité, on fasse intervenir l'écart réel entre les éléments comparés, alors que cet écart est inconnu. En fait, il est essentiel de connaître, en fonction de cet écart, la probabilité d'acceptation de l'hypothèse d'identité, car la notion d'efficacité donne les limites du test ainsi que les risques pris en acceptant ou refusant l'hypothèse.

En effet, si l'hypothèse d'identité est effectivement réalisée, l'écart estimé sur échantillon fluctue autour de zéro, et le test consiste à s'assurer que cette fluctuation est comprise dans des limites compatibles avec les aléas d'échantillonnage. Si l'hypothèse n'est pas réalisée, il existe un écart réel entre les éléments comparés ; l'écart estimé fluctue autour de l'écart réel ; il peut se trouver à l'intérieur des limites précédentes, et ainsi entraîner l'acceptation de l'hypothèse fausse d'identité. Il est donc nécessaire de connaître la probabilité pour que l'écart estimé dans ce cas soit dans les limites fixées pour l'acceptation de l'hypothèse. Plus cette probabilité sera faible, plus le test sera efficace.

Il sera très souvent nécessaire d'accroître l'efficacité d'un test ; on y parviendra, le plus souvent, en augmentant la taille de l'échantillon.

4.4 Diverses catégories de tests

On distingue trois sortes de catégories de tests que l'on étudiera dans les paragraphes suivants :

- les tests d'ajustement (§ 5) ;
- les tests paramétriques (§ 6) ;
- les tests non paramétriques (§ 7).

Les premiers consistent essentiellement à comparer le modèle statistique théorique choisi pour représenter la population, à l'échantillon prélevé.

Les seconds ont pour but de comparer les paramètres (proportions, moyenne, variance) estimés sur deux ou plusieurs échantillons, pour savoir si leurs différences sont significatives, ou, au contraire, si on peut considérer ces paramètres comme issus d'une même population-mère.

Les troisièmes tests ont pour objet les comparaisons d'échantillons dont les populations-mères sont de nature inconnue, ou encore de séries d'observations appariées.

5. Tests d'ajustement

5.1 Généralités

Après étude d'un échantillon prélevé dans une population inconnue, on est amené à faire choix d'un modèle théorique. Tout d'abord, pour diverses considérations, la nature de la distribution est, selon les cas, supposée normale ou binomiale, etc. (§ 2). Le calcul des paramètres de l'échantillon permet ensuite d'obtenir des estimations des paramètres de la population (§ 3). Il est naturel de choisir un modèle théorique construit à partir de ces données.

Il sera nécessaire de tester la légitimité de ce choix. Dans ce but, on fera l'hypothèse que le modèle théorique choisi représente bien la population étudiée. Le principe du test consistera à s'assurer que l'échantillon prélevé est compatible avec ce modèle théorique, autrement dit que la variable aléatoire constituée par l'échantillon fluctue autour du modèle théorique dans des limites compatibles uniquement avec les aléas d'échantillonnage.

5.2 Principe des tests

Les tests ayant pour but de légitimer le choix d'un modèle, on conçoit aisément qu'ils seront basés sur les fréquences des résultats obtenus sur l'échantillon.

Ainsi, ayant opéré un regroupement par classes des données de l'échantillon, on construit un échantillon *idéal* de même taille que l'échantillon prélevé ; on l'ordonne suivant des classes identiques, avec des effectifs obtenus à partir du modèle théorique retenu. Si le modèle théorique choisi est bon, cet échantillon idéal peut ainsi être considéré comme la *moyenne* des échantillons de même taille, prélevés dans la population. Réciproquement, tout échantillon prélevé dans la population est une variable aléatoire qui fluctue autour de l'échantillon idéal.

On construit, avec les données des échantillons observés et des échantillons idéaux, une fonction caractérisant, en quelque sorte, leur écart. Cette fonction sera une variable aléatoire, et, par construction, on connaîtra sa loi de probabilité.

Le principe du test consistera à s'assurer que la valeur de l'écart obtenu reste dans des limites compatibles avec les aléas d'échantillonnage.

On déterminera ainsi un intervalle dans lequel, si le modèle retenu est bon, l'écart a une probabilité $1 - \alpha$ de se trouver (α étant faible). Si l'écart observé est dans cet intervalle, rien ne s'oppose à l'hypothèse que le modèle retenu soit acceptable. Dans le cas contraire, il est préférable de rejeter cette hypothèse et d'attribuer au choix erroné du modèle la valeur improbable observée, plutôt qu'aux aléas de l'échantillonnage. En effet, le risque pour que l'hypothèse soit exacte est au plus égal à α , faible par définition.

5.3 Test du χ^2

5.3.1 Écart entre l'échantillon et le modèle

Supposons que les observations effectuées sur l'échantillon (dont la taille sera notée N) aient été regroupées dans des classes $1, 2, \dots, i, \dots, k$, d'effectifs $n_1, n_2, \dots, n_i, \dots, n_k$, avec :

$$\sum_{i=1}^{i=k} n_i = N$$

Le modèle théorique choisi permettra de construire un échantillon idéal avec des effectifs de classes $n'_1, n'_2, \dots, n'_i, \dots, n'_k$. On aura :

$$\sum_{i=1}^{i=k} n'_i = N$$

On a choisi de donner à l'écart entre l'échantillon obtenu et l'échantillon idéal la forme suivante :

$$E = \sum_{i=1}^{i=k} \frac{(n_i - n'_i)^2}{n'_i}$$

Cette forme est justifiée par les propriétés suivantes :

- la variable E est nulle quand $n_i = n'_i$;
- elle tend vers une variable χ^2 quand N tend vers l'infini.

Supposons en effet que le modèle théorique soit acceptable, et soit p_i la probabilité de prélever un élément de la population appartenant à la classe i . La probabilité de prélever l'échantillon observé est donnée par la loi multinomiale ([A 165] *Probabilités*) et, dans ces conditions, la variable aléatoire u_i donnée par l'expression suivante :

$$u_i = \frac{n_i - N p_i}{\sqrt{N p_i}} = \frac{n_i - n'_i}{\sqrt{n'_i}}$$

tend vers une variable aléatoire normale réduite quand N tend vers l'infini. La variable aléatoire $E = \sum_i u_i^2$ tend alors vers une loi du χ^2

quand N tend vers l'infini (cf. définition de la loi du χ^2 , en [A 165] *Probabilités*). Le nombre de degrés de liberté de cette loi est égal au nombre de variables aléatoires u_i indépendantes.

La procédure du test du χ^2 s'en déduit simplement. On compare E à la valeur $\chi^2_{1-\alpha}$ ayant une probabilité α d'être dépassée. Si $E < \chi^2_{1-\alpha}$, on accepte l'hypothèse. Si $E > \chi^2_{1-\alpha}$, on refuse l'hypothèse.

5.3.2 Détermination du nombre de degrés de liberté de la loi du χ^2

Si les valeurs de p_i ont été déterminées sans utiliser les données de l'échantillon, il existe une seule relation entre les u_i , soit :

$$\sum_{i=1}^{i=k} u_i = 0$$

Il y a donc $k - 1$ variables aléatoires u_i indépendantes, et le nombre de degrés de liberté à employer est $k - 1$.

Si les valeurs de p_i ont été calculées à partir des estimations obtenues sur l'échantillon, on montre alors que le nombre de degrés de liberté à employer est égal au nombre $k - 1$, diminué du nombre d'estimations utilisées.

Il est assez intuitif, en effet, de voir que l'utilisation de la moyenne et de l'écart-type par exemple lie les variables aléatoires u_i par deux relations supplémentaires, ce qui diminue encore de 2 le nombre de variables u_i indépendantes. Par conséquent, le nombre de degrés de liberté à employer en général sera :

$$\lambda = k - 1 - s$$

avec k nombre de classes,

s nombre d'estimations utilisées.

Si les paramètres du modèle ont été pris égaux aux estimations tirées de l'échantillon prélevé, dans le cas d'une loi binomiale ou d'une loi de Poisson, on utilisera la loi du χ^2 à $k - 2$ degrés de liberté et, dans le cas de la loi normale, une loi du χ^2 à $k - 3$ degrés de liberté.

5.3.3 Procédure pratique du test

Remarque préliminaire importante : on peut admettre que E suit une loi du χ^2 quel que soit N , sous réserve que les effectifs des classes de l'échantillon prélevé ne soient pas excessivement faibles. En pratique, on sera souvent obligé d'effectuer, en particulier aux extrémités de la distribution expérimentale, des regroupements de classes, de façon que les effectifs soient au moins égaux à 4 ou 5. Nous allons exposer la procédure du test sur deux exemples (§ 5.3.3.1 et 5.3.3.2).

5.3.3.1 Premier exemple : ajustement à une loi normale

Un échantillon, prélevé dans une population inconnue, a donné les résultats regroupés dans le tableau 7.

Tableau 7 – Résultats expérimentaux (exemple du § 5.3.3.1)	
Classe	Effectif n_i
$x \leq 110$	7
$110 < x \leq 120$	14
$120 < x \leq 130$	16
$130 < x \leq 140$	9
$140 < x$	4
Total N	50

On fait l'hypothèse que la loi de probabilité de la population est une loi normale dont les paramètres sont des estimations obtenues sur l'échantillon soit, d'après le paragraphe 3.5.2 :

$$\begin{aligned}\bar{x} &\approx 122,8 \\ s' &\approx 11,5\end{aligned}$$

Nota : on a affecté à chaque élément de la première classe la valeur 105, et à chaque élément de la dernière classe la valeur 145.

Nous allons donc tester l'ajustement de la distribution empirique avec loi normale, de moyenne $m = 122,8$, et d'écart-type $\sigma = 11,5$.

Il est nécessaire de connaître les valeurs de p_i pour obtenir les valeurs n'_i des effectifs théoriques. On considère pour cela la variable réduite :

$$u_i = \frac{x_i - 122,8}{11,5}$$

On calculera les intervalles de classes en variable réduite ; on obtiendra ainsi, avec les tables de la loi normale réduite, les valeurs de p_i . Les résultats sont regroupés dans le tableau 8.

Tableau 8 – Calcul de l'écart E (exemple du § 5.3.3.1)				
Classe	n_i	p_i	$n'_i = Np_i$	$\frac{(n_i - n'_i)^2}{n'_i}$
$u_i \leq -1,113$	7	0,133	6,65	0,018 4
$-1,113 < u_i \leq -0,243$	14	0,271	13,55	0,015 0
$-0,243 < u_i \leq 0,626$	16	0,320	16,00	0,000 0
$0,626 < u_i \leq 1,496$	9	0,199	9,95	0,090 7
$1,496 < u_i$	4	0,077	3,85	0,005 8
Totaux	$N = 50$	1,000	$N = 50,00$	$E = 0,129 9$

Le nombre de degrés de liberté à utiliser est $5 - 3 = 2$. Prenons un seuil $\alpha = 5 \%$. La valeur du χ^2 , à deux degrés de liberté, est alors égale à 5,99.

Par conséquent ($E \ll 5,99$), rien ne s'oppose à l'hypothèse avancée, et on peut choisir pour modèle théorique la loi normale, de moyenne $m = 122,8$, et d'écart-type $\sigma = 11,5$.

5.3.3.2 Deuxième exemple : ajustement à une loi binomiale

On effectue sur une fabrication 100 prélèvements de 20 pièces. On classe chaque prélèvement en pièces bonnes et en pièces défectueuses. Le nombre de prélèvements contenant 0, 1, 2, ... pièces défectueuses est donné dans le tableau 9.

Tableau 9 – Résultats expérimentaux (exemple du § 5.3.3.2)		
Nombre de pièces défectueuses par prélèvement k	Nombre de prélèvements n_i	Nombre total de pièces défectueuses kn_i
0	13	0
1	18	18
2	39	78
3	15	45
4	15	60
Totaux	$N = 100$	201

Sur les 2 000 pièces prélevées, on a ainsi observé 201 pièces défectueuses, soit une proportion de 10 %.

Nous allons tester l'hypothèse que la distribution des pièces mauvaises dans la fabrication obéit à une loi binomiale avec $p = 10 \%$. Les résultats sont condensés dans le tableau 10.

Tableau 10 – Calcul de l'écart E (exemple du § 5.3.3.2)			
Nombre de pièces défectueuses par prélèvement k	n_i	$n'_i = Np_i$ (1)	$\frac{(n_i - n'_i)^2}{n'_i}$
0	13	12,16	0,06
1	18	27,02	3,01
2	39	28,52	3,86
3	15	19,00	0,84
4	15	13,3	0,22
Totaux	$N = 100$	$N = 100,00$	$E = 7,99$
(1) Les valeurs p_i sont directement extraites des tables de loi binomiale.			

La loi du χ^2 à employer est à $5 - 2 = 3$ degrés de liberté. La valeur du χ^2 ayant une probabilité de 5 % d'être dépassée vaut 7,82 ; nous voyons que $E > 7,82$.

Il est, par conséquent, préférable de repousser l'hypothèse, la probabilité pour que l'hypothèse soit exacte étant de l'ordre de 5 %.

5.4 Droite de Henry

Nous allons exposer ci-dessous une méthode permettant de justifier le choix d'une loi morale. Cependant, il faut bien insister sur le fait que cette méthode ne constitue, en aucune manière, un test

statistique. En effet, en cas de rejet de l'hypothèse, il n'est pas possible de chiffrer la probabilité pour que l'hypothèse repoussée soit en fait exacte.

Supposons que l'on ait classé les résultats et que, pour chaque valeur x_i , on connaisse la fréquence empirique f_i correspondante. On peut ainsi établir les fréquences cumulées empiriques F_i .

En utilisant la table de la loi de répartition d'une variable normale réduite, on pourra déterminer la valeur u_i de la variable réduite correspondant à une probabilité cumulée égale à F_i .

Portons sur un graphique, en abscisses, les valeurs x_i observées, et, en ordonnées, les valeurs u_i obtenues comme indiqué plus haut. Si la distribution étudiée est voisine d'une distribution normale, les points de coordonnées (x_i, u_i) sont approximativement alignés, en raison de la définition même de la variable réduite :

$$u_i = \frac{x_i - m}{\sigma}$$

La droite sur laquelle ces points sont alignés est appelée *droite de Henry*. Sa pente et son ordonnée à l'origine permettent d'obtenir aisément une approximation de la moyenne et de l'écart-type.

Exemple

Nous allons tracer la droite de Henry pour l'échantillon explicité dans le tableau 11.

On établit le tableau 12 en calculant les fréquences cumulées F_i , et en tirant les valeurs de u_i des tables de la loi normale.

Nota : signalons qu'il existe du papier quadrillé gradué en F_i et donnant directement les valeurs de u_i . On évite ainsi la recherche des u_i dans les tables de la loi normale. Ce papier est appelé *papier gauss-arithmétique*.

On obtient la droite de Henry représentée par la figure 5.

En posant $u_i = 0$, on a $x_i = m$. On obtient ainsi sur le graphique la valeur de la moyenne qui vaut approximativement $m = 13,7$. Le calcul donne par contre 14,1.

En posant $u_i = 1$, on a $\sigma = x_i - m$. On obtient, graphiquement, $\sigma = 1,7$, valeur proche de celle obtenue par le calcul.

Remarques

- La droite de Henry est couramment utilisée ; cependant, certaines précautions sont à conseiller dans son emploi. Il est, en effet, bien difficile, selon l'échelle du graphique, d'estimer si des points sont bien alignés ou non.

- Le graphique de Henry se présente parfois aux extrémités sous forme de deux familles distinctes de points, alignés suivant deux droites parallèles ; ceci indique en général qu'on a affaire à un mélange de deux populations de même écart-type mais de moyennes légèrement différentes.

- On peut également appliquer la méthode de la droite de Henry pour reconnaître si une distribution est gauss-logarithmique, mais en portant en abscisses les logarithmes de x_i .

Nota : il existe de même du papier quadrillé pour cet usage ; il est dit *gausso-logarithmique*.

Tableau 11 – Résultats expérimentaux
(exemple du § 5.4)

x_i	11	12	13	14	15	16	17	18
n_i	6	13	18	23	19	12	6	3

Tableau 12 – Détermination de la droite de Henry
(exemple du § 5.4)

x_i	n_i	f_i	F_i	u_i
11	6	0,06	0,06	-1,56
12	13	0,13	0,19	-0,88
13	18	0,18	0,37	-0,33
14	23	0,23	0,60	+0,26
15	19	0,19	0,79	+0,81
16	12	0,12	0,91	+1,34
17	6	0,06	0,97	+1,88
18	3	0,03	1,00	∞

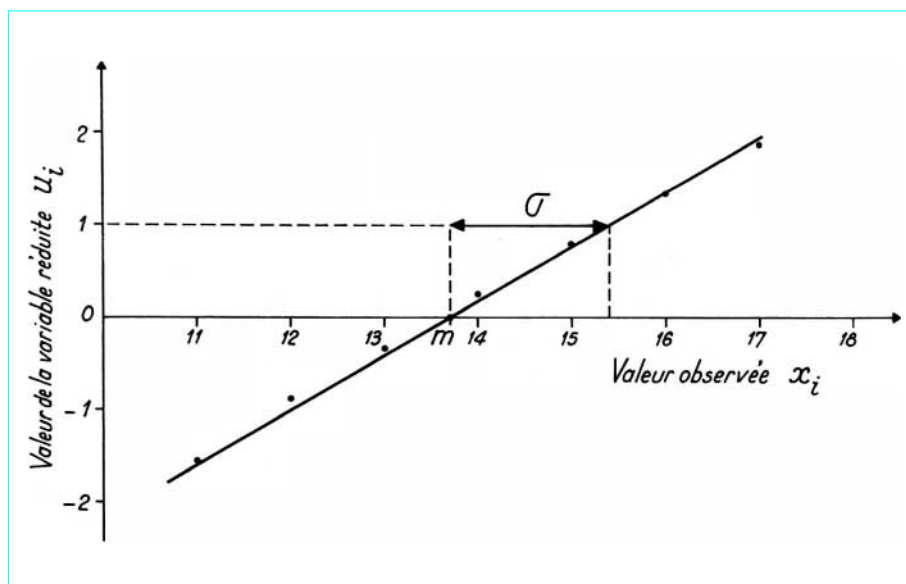


Figure 5 – Droite de Henry

6. Tests paramétriques

On distingue deux catégories de tests paramétriques.

■ D'une part, on rencontre les *tests de comparaison à un standard* (§ 6.1), qui consistent à comparer la valeur d'une estimation d'un paramètre à une valeur quelconque, dite standard, afin de savoir si cette estimation, compte tenu des fluctuations d'échantillonnage, peut être considérée comme égale au standard.

■ D'autre part, on rencontre les *tests de comparaison de paramètres de populations* (§ 6.2) qui ont pour objet de tester la différence entre deux échantillons avec les réponses suivantes : les deux échantillons sont extraits d'une même population, et leur écart peut être mis au compte des aléas d'échantillonnage ; ou les deux échantillons proviennent de deux populations différentes.

D'une manière générale, signalons qu'il est possible de déterminer presque toutes les courbes d'efficacité de ces tests, mais que ces notions, dont l'exposé sortirait du cadre de cet article, ne recevront qu'un développement très succinct dans quelques cas particuliers.

6.1 Tests de comparaison à un standard

Soit λ un paramètre quelconque, estimé à partir d'un échantillon, à comparer à un paramètre λ_0 . La comparaison de λ à λ_0 pourra se faire suivant trois hypothèses différentes :

- première hypothèse : $\lambda = \lambda_0$;
- deuxième hypothèse : $\lambda \geq \lambda_0$;
- troisième hypothèse : $\lambda \leq \lambda_0$.

Ces trois hypothèses se résument en fait dans la première.

6.1.1 Efficacité des tests

La notion d'efficacité permet d'enrichir considérablement le test. En effet, faisons l'hypothèse $\lambda = \lambda_0$; l'efficacité se définit comme la probabilité d'accepter l'hypothèse, en fonction de la valeur vraie, du paramètre λ . Or, si on considère un échantillon donné, et si on se donne le risque de première espèce α (probabilité de refuser l'hypothèse alors qu'elle est exacte), l'efficacité du test est, en général, entièrement déterminée. Par contre, on pourra, si l'échantillon n'est pas fixé, définir le test de telle sorte qu'en plus du risque α , on puisse se fixer à l'avance le risque β d'accepter l'hypothèse $\lambda = \lambda_0$ alors qu'elle est fautive et que la valeur de λ est égale à λ_1 . Autrement dit, on pourra déterminer le test à partir de deux points de la courbe d'efficacité (*déterminer le test* signifie, en fait, s'assigner une taille de l'échantillon, et définir les conditions d'acceptation et de refus).

6.1.2 Test de comparaison d'une proportion à un standard

6.1.2.1 Test

Plaçons-nous, tout d'abord, dans le cas où l'approximation de la loi binomiale par une loi normale est valable (§ 3.5.1.2.1). Dans ce cas, soit $f = k/n$ la fréquence observée sur l'échantillon, et soit p_0 le standard.

Faisons l'hypothèse suivante : la proportion p de la population est égale à p_0 . Dans ce cas, la variable u définie par l'expression :

$$u = \frac{f - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

est une variable aléatoire normale réduite. Soit u_α la valeur de la variable normale réduite, telle que :

$$P(|u| > u_\alpha) = \alpha$$

Dans ces conditions, on aura une probabilité $1 - \alpha$ pour que la fréquence f soit comprise dans l'intervalle :

$$p_0 \pm u_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}$$

Si f est effectivement comprise dans cet intervalle, rien ne s'oppose, en fait, à accepter l'hypothèse $p = p_0$. Sinon, avec un risque α , on pourra rejeter l'hypothèse.

On pourra transformer l'intervalle d'acceptation de la fréquence en intervalle d'acceptation de la valeur k .

Supposons que nous voulions tester l'hypothèse $p > p_0$ toujours avec le même risque u_α . Il suffira alors de vérifier que :

$$f > p_0 + u_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}$$

u_α étant la valeur de la loi normale réduite, telle que :

$$P(|u| > u_\alpha) = \alpha$$

6.1.2.2 Efficacité

L'efficacité du test sera égale à la probabilité $P(\varpi)$ d'accepter l'hypothèse d'égalité, quand p est en réalité égal à ϖ .

$$P(\varpi) = P(\text{accepter } p = p_0 \text{ si } p = \varpi)$$

$$P(\varpi) = F \left\{ \frac{p_0 + u_\alpha \sqrt{\frac{p_0(1-p_0)}{n}} - \varpi}{\sqrt{\frac{\varpi(1-\varpi)}{n}}} \right\} - F \left\{ \frac{p_0 - u_\alpha \sqrt{\frac{p_0(1-p_0)}{n}} - \varpi}{\sqrt{\frac{\varpi(1-\varpi)}{n}}} \right\}$$

où F est la fonction de répartition de la loi normale réduite. En effet, la probabilité d'acceptation de l'hypothèse est égale à la probabilité d'obtenir une valeur de f comprise dans l'intervalle d'acceptation, quand cette valeur expérimentale f fluctue autour de la valeur ϖ . La figure 6 permet, mieux que tous commentaires, de comprendre l'expression de P .

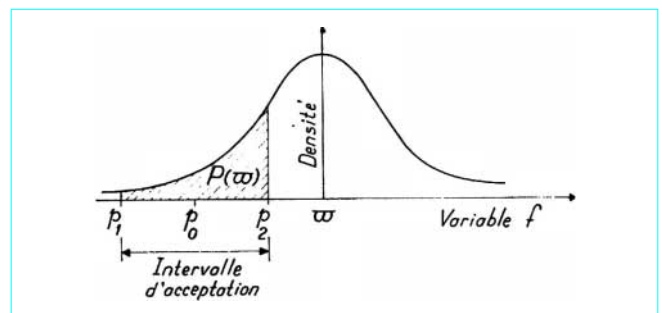


Figure 6 – Courbe de densité de probabilité de f

Ainsi, il est possible de résoudre le problème suivant : déterminer la taille de l'échantillon de façon qu'on ne puisse refuser l'hypothèse $p = p_0$ qu'avec un risque α si cette hypothèse est exacte, et qu'on n'accepte l'hypothèse $p = p_0$ qu'avec une probabilité β si en réalité $p = p_1$. On déterminera en même temps l'intervalle d'acceptation.

6.1.2.3 Remarque

Dans le cas où l'approximation de la loi binomiale par la loi normale n'est pas valable, il est toujours possible d'effectuer des tests de comparaison d'une proportion à un standard. Le lecteur pourra trouver dans les ouvrages spécialisés – dont certains sont cités dans la bibliographie [Doc. A 166] – les abaques permettant d'effectuer ces tests.

6.1.3 Test de comparaison de la variance à un standard

L'échantillon est extrait d'une population normale (§ 3.5.2.1.2). Soit s^2 sa variance.

Faisons l'hypothèse suivante : la variance σ^2 de la population est égale à σ_0^2 . Dans ces conditions, la variable aléatoire ns^2/σ_0^2 suit une loi de probabilité du χ^2 à $n-1$ degrés de liberté. Soient $\chi_{\alpha/2}^2$ et $\chi_{1-(\alpha/2)}^2$ les valeurs de la variable χ^2 telles que :

$$P\{\chi^2 < \chi_{\alpha/2}^2\} = \alpha/2$$

$$P\{\chi^2 < \chi_{1-(\alpha/2)}^2\} = 1 - (\alpha/2)$$

Si $\frac{ns^2}{\sigma_0^2}$ est compris dans l'intervalle $[\chi_{\alpha/2}^2, \chi_{1-(\alpha/2)}^2]$ on accep-

tera l'hypothèse que la variance de la population est égale à σ_0^2 . Dans le cas contraire, on rejettera l'hypothèse, avec un risque α pour que l'hypothèse soit en fait exacte.

Exemple

Tester l'hypothèse que la variance est égale à 25, dans les conditions suivantes :

$$n = 10, \quad s^2 = 15, \quad \frac{ns^2}{\sigma_0^2} = 6$$

L'intervalle d'acceptation de l'hypothèse, avec $\alpha = 5\%$, est :

$$2,7 < \frac{ns^2}{\sigma_0^2} < 19$$

L'hypothèse est acceptée.

6.1.4 Test de comparaison de la moyenne à un standard : premier cas

L'échantillon est extrait d'une population normale, de variance σ^2 connue (§ 3.5.2.2.1).

Faisons l'hypothèse suivante : la moyenne m de la population est égale à m_0 . Dans ces conditions, la variable aléatoire donnée par l'expression :

$$u = \frac{\bar{x} - m_0}{\sigma/\sqrt{n}}$$

suit une loi normale réduite. Soit u_α la valeur de la variable aléatoire normale réduite u , telle que :

$$P(|u| > u_\alpha) = \alpha$$

Si la valeur \bar{x} est comprise dans l'intervalle $m_0 \pm u_\alpha \sigma/\sqrt{n}$, rien ne s'opposera à l'acceptation de l'hypothèse $m = m_0$. Dans le cas contraire, on rejettera l'hypothèse, le risque pour que celle-ci soit exacte étant au plus égal à α .

Exemple

Tester l'hypothèse que la moyenne de la population est égale à 70, dans les conditions suivantes :

$$n = 25, \quad \bar{x} = 58, \quad \sigma^2 = 36$$

Dans ce cas, avec $\alpha = 5\%$, $u_\alpha = 1,96$ et $u_\alpha \sigma/\sqrt{n} \approx 2,35$, l'intervalle d'acceptation de l'hypothèse est alors :

$$67,65 < \bar{x} < 72,35$$

L'hypothèse est rejetée.

6.1.5 Test de comparaison de la moyenne à un standard : deuxième cas

L'échantillon est extrait d'une population normale, de variance σ^2 inconnue (§ 3.5.2.2.2).

Faisons l'hypothèse suivante : la moyenne m de la population est égale à m_0 . Dans ces conditions, la variable aléatoire, donnée par l'expression :

$$t = \frac{\bar{x} - m_0}{s'/\sqrt{n}} \quad \text{avec} \quad s' = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

suit une loi de Student-Fisher à $n-1$ degrés de liberté. Soit t_α la valeur de la variable de Student-Fisher, telle que :

$$P(|t| \leq t_\alpha) = 1 - \alpha$$

Si la valeur \bar{x} est comprise dans un intervalle $m_0 \pm t_\alpha s'/\sqrt{n}$, rien ne s'opposera à l'acceptation de l'hypothèse. Dans le cas contraire, on rejettera l'hypothèse, le risque pour que celle-ci soit exacte étant au plus égal à α .

Exemple

Tester l'hypothèse que la moyenne de la population est égale à 120, dans les conditions suivantes :

$$n = 12, \quad \bar{x} = 117, \quad s'^2 = 48$$

L'intervalle d'acceptation, avec $\alpha = 5\%$ ($t_\alpha = 2,20$), est :

$$124,4 > \bar{x} > 115,6$$

Par conséquent, l'hypothèse peut être acceptée.

6.2 Tests de comparaison de populations

6.2.1 Test de comparaison de deux proportions (loi binomiale)

6.2.1.1 Cas d'échantillons de taille élevée

On a extrait de deux populations deux échantillons, l'un de taille n_1 et de proportion f_1 , l'autre de taille n_2 et de proportion f_2 . La taille des échantillons est élevée ; l'approximation par la loi normale est donc valable (n_1 et $n_2 \geq 100$).

Faisons l'hypothèse suivante : la proportion p_1 de la première population est égale à la proportion p_2 de la deuxième population : $p_1 = p_2 = p$.

Nous savons ([A 165] *Probabilités*) que f_1 est une variable normale, de moyenne p_1 , et de variance $p_1 q_1 / n_1$ avec $q_1 = 1 - p_1$. De même, f_2 est une variable normale, de moyenne p_2 , et de variance $p_2 q_2 / n_2$ avec $q_2 = 1 - p_2$.

Dans ces conditions, $f_1 - f_2$ est une variable normale, de moyenne $p_1 - p_2$, et de variance $p_1 q_1 / n_1 + p_2 q_2 / n_2$. Dans le cadre de l'hypothèse, $f_1 - f_2$ sera une variable normale, de moyenne nulle, et de variance $p q (1/n_1 + 1/n_2)$.

Considérons la variable réduite :

$$u = \frac{f_1 - f_2}{\sqrt{pq \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

On aura une probabilité $1 - \alpha$ pour que la variable u soit comprise dans l'intervalle $[-u_\alpha, +u_\alpha]$.

Si u est effectivement comprise dans cet intervalle, rien ne s'oppose à l'hypothèse $p_1 = p_2 = p$. Dans le cas contraire, on pourra repousser l'hypothèse, avec un risque α .

Pratiquement, pour exécuter le test, il nous manque la valeur p .

On utilise pour cela une estimation basée sur les deux échantillons :

$$p = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

Exemple

On prélève 100 pièces dans un lot et on trouve 5 pièces défectueuses. Dans un second lot, on prélève 200 pièces et on trouve 7 pièces défectueuses. Peut-on considérer que ces deux lots sont identiques du point de vue qualité ?

$$p = \frac{5 + 7}{100 + 200} = 0,04$$

$$f_1 = 0,05 \quad \text{et} \quad f_2 = 0,035$$

On trouve alors :

$$u = \frac{f_1 - f_2}{\sqrt{pq \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \approx 0,63$$

au risque $\alpha = 5\%$, $u_\alpha = 1,96$. Dans ces conditions, on pourra accepter l'hypothèse ; la différence entre les deux prélèvements n'est pas statistiquement significative.

6.2.1.2 Cas d'échantillons de taille réduite

Soient deux échantillons (n_1, k_1) , (n_2, k_2) . On peut établir le tableau 13 résumant les résultats en supposant que les éléments de la population sont classés en bons et mauvais.

Tableau 13 – Comparaison de deux populations : données expérimentales			
Échantillon	Taille	Éléments mauvais	Éléments bons
n° 1	n_1	k_1	l_1
n° 2	n_2	k_2	l_2
Totaux	$n_1 + n_2$	$k = k_1 + k_2$	$l = l_1 + l_2$

On fait l'hypothèse identique à celle du paragraphe 6.2.1.1 : $p_1 = p_2 = p$. Dans ces conditions, une estimation de p sera donnée par :

$$p = \frac{k_1 + k_2}{n_1 + n_2} = \frac{k}{n_1 + n_2}$$

Connaissant p , on pourra construire des échantillons idéaux correspondant au modèle théorique résumé dans le tableau 14.

Tableau 14 – Comparaison de deux populations : construction d'échantillons idéaux			
Échantillon	Taille	Éléments mauvais	Éléments bons
n° 1	n_1	$n_1 p$	$n_1 (1 - p)$
n° 2	n_2	$n_2 p$	$n_2 (1 - p)$
Totaux	$n_1 + n_2$	$(n_1 + n_2) p$	$(n_1 + n_2) (1 - p)$

La technique du test est alors très proche de celle utilisée dans le paragraphe 5. On teste l'écart entre les échantillons théoriques et les échantillons pratiques, dans l'hypothèse $p_1 = p_2 = p$. Si l'écart est compatible avec les aléas d'échantillonnage, on accepte l'hypothèse ; dans le cas contraire elle est repoussée.

On calcule la variable E qui suit, dans ce cas particulier, une loi du χ^2 à 1 degré de liberté :

$$E = \frac{(k_1 - n_1 p)^2}{n_1 p} + \frac{(k_2 - n_2 p)^2}{n_2 p} + \frac{[l_1 - n_1 (1 - p)]^2}{n_1 (1 - p)} + \frac{[l_2 - n_2 (1 - p)]^2}{n_2 (1 - p)}$$

On compare alors E à la valeur de $\chi^2_{1-\alpha}$ à 1 degré de liberté ; si

$E < \chi^2_{1-\alpha}$, l'hypothèse est acceptée.

Nota : les quatre variables $(k_1 - n_1 p)$, $(k_2 - n_2 p)$, $[l_1 - n_1 (1 - p)]$ et $[l_2 - n_2 (1 - p)]$ ne sont pas indépendantes ; elles sont liées par trois relations, ce qui explique le degré de liberté qui est égal à 1.

■ **Remarque :** les effectifs k_1, k_2, l_1, l_2 ne doivent pas être trop faibles ; en pratique ils doivent être supérieurs à 4 ou 5.

Exemple

Tester l'hypothèse que les proportions de deux populations sont identiques, les résultats sur deux prélèvements étant ceux du tableau 15.

L'estimation de p est 0,1.

Les résultats théoriques sont résumés dans le tableau 16.

La variable E est alors $E \approx 1,85$ et la valeur du χ^2 ayant une probabilité de 5 % d'être dépassée vaut 3,84.

L'hypothèse d'égalité des proportions est acceptée.

Tableau 15 – Comparaison de deux populations : données expérimentales (exemple § 6.2.1.2)			
Échantillon	Taille	Éléments mauvais	Éléments bons
n° 1	40	6	34
n° 2	60	4	56
Totaux	100	10	90

Tableau 16 – Comparaison de deux populations : construction d'échantillons idéaux (exemple du § 6.2.1.2)

Échantillon	Taille	Éléments mauvais	Éléments bons
n° 1	40	4	36
n° 2	60	6	54
Totaux	100	10	90

6.2.1.3 Cas d'échantillons de taille réduite : généralisation du test

Le test se généralise aisément à plusieurs échantillons classés, non plus suivant deux caractères, mais suivant plusieurs caractères. Le nombre de degrés de liberté de la variable χ^2 est alors $(\mu - 1)(\lambda - 1)$, μ étant le nombre d'échantillons, λ le nombre de caractères.

Exemple

Peut-on considérer que les trois lots de fabrications classés en produits de 1^{er}, 2^e, 3^e choix, sont identiques du point de vue qualité, les résultats sur prélèvements étant ceux du tableau 17.

On a alors $p_1 = 0,74$, $p_2 = 0,15$, $p_3 = 0,11$.

Les résultats théoriques sont donnés dans le tableau 18.

$$E = \frac{(0,4)^2}{44,4} + \frac{(1)^2}{9} + \frac{(0,6)^2}{66,6} + \frac{(1,6)^2}{66,6} + \frac{(0,5)^2}{13,5} + \frac{(1,1)^2}{9,9} + \frac{(2)^2}{37} + \frac{(1,5)^2}{7,5} + \frac{(0,5)^2}{5,5} \approx 0,8$$

Le nombre de degrés de liberté est : $(3 - 1)(3 - 1) = 4$; donc la valeur du χ^2 ayant une probabilité de 5 % d'être dépassée vaut 9,49.

On peut, par conséquent, accepter l'hypothèse.

Tableau 17 – Comparaison de trois populations : données expérimentales (exemple du § 6.2.1.3)

Lot	Taille	Premier choix	Deuxième choix	Troisième choix
n° 1	60	44	10	6
n° 2	90	65	14	11
n° 3	50	39	6	5
Totaux	200	148	30	22

Tableau 18 – Comparaison de trois populations : construction d'échantillons idéaux (exemple du § 6.2.1.3)

Lot	Taille	Premier choix	Deuxième choix	Troisième choix
n° 1	60	44,4	9,0	6,6
n° 2	90	66,6	13,5	9,9
n° 3	50	37,0	7,5	5,5
Totaux	200	148,0	30,0	22,0

6.2.2 Test de comparaison des variances de populations normales

6.2.2.1 Cas d'échantillons de taille élevée

Quand n est élevé, l'écart-type s d'un échantillon suit approximativement une loi normale, de moyenne σ , et d'écart-type $\sigma/\sqrt{2n}$ (§ 2.7). Considérons deux échantillons, respectivement de taille n_1 et n_2 , et d'écart-type s_1 et s_2 .

Faisons l'hypothèse suivante : les variances des deux populations sont égales :

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

Dans ces conditions, s_1 et s_2 sont des variables normales, de moyenne σ , et d'écart-type $\frac{\sigma}{\sqrt{2n_1}}$ et $\frac{\sigma}{\sqrt{2n_2}}$. De même, $s_1 - s_2$ est une variable normale, de moyenne nulle, et d'écart-type :

$$\sigma \sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}$$

Considérons la variable réduite :

$$u = \frac{s_1 - s_2}{\sigma \sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}}$$

On aura une probabilité $1 - \alpha$ pour que la variable u soit comprise dans l'intervalle $[-u_\alpha, +u_\alpha]$.

Si u est effectivement comprise dans cet intervalle, rien ne s'oppose à l'hypothèse $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Dans le cas contraire, on pourra repousser l'hypothèse, avec un risque α .

Pratiquement, pour effectuer ce test, il nous manque la valeur σ ; on utilise, pour cela, l'estimation basée sur les deux échantillons (§ 3.5.2.4) :

$$s'^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

Exemple

Comparer les variances de deux populations, les résultats obtenus sur deux échantillons étant les suivants : pour le 1^{er} échantillon, $n_1 = 100$ et $s_1 = 1$; pour le 2^e échantillon, $n_2 = 300$ et $s_2 = 4$.

$$\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{100 + 4 \cdot 800}{398} \approx 12,31$$

$$u = \frac{s_1 - s_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}} \approx \frac{-3}{3,5 \sqrt{\frac{1}{200} + \frac{1}{600}}}$$

$$|u| \approx 10,5$$

Or $u_\alpha = 1,96$, avec $\alpha = 0,05$; par conséquent, on peut rejeter l'hypothèse d'égalité des variances.

6.2.2.2 Cas d'échantillons de taille réduite

Soient n_1 et s_1^2 la taille et la variance de l'échantillon extrait de la première population, n_2 et s_2^2 celles du deuxième échantillon.

Faisons l'hypothèse suivante : les variances des deux populations sont égales. Les estimations des variances des deux populations sont (§ 3.5.2.1.2) :

$$s_1'^2 = \frac{n_1}{n_1 - 1} s_1^2 \quad \text{et} \quad s_2'^2 = \frac{n_2}{n_2 - 1} s_2^2$$

Dans le cadre de l'hypothèse d'égalité des variances, on montre que la variable aléatoire $F = s_1'^2 / s_2'^2$ suit une loi de Snédécov à $v_1 = n_1 - 1$, $v_2 = n_2 - 1$ degrés de liberté, dont la forme est donnée en [A 165] *Probabilités*.

On détermine l'intervalle de probabilité dans lequel la variable F a une probabilité $1 - \alpha$ de se trouver. Si F se trouve dans cet intervalle, l'hypothèse est acceptée.

Les degrés de liberté du numérateur dans les tables de Snédécov se lisent en tête de colonne, ceux du dénominateur en tête de ligne. En effet, les degrés de liberté de la loi de Snédécov n'interviennent pas d'une façon symétrique.

On forme alors le rapport $s_1'^2 / s_2'^2$ en mettant au numérateur la variance la plus élevée, et on recherche la valeur $F_{\alpha/2}(v_1, v_2)$ ayant une probabilité $\alpha/2$ d'être dépassée, qui donne la borne supérieure de l'intervalle.

Si on considère le rapport inverse $s_2'^2 / s_1'^2$, les tables de Snédécov donnent une valeur $F_{\alpha/2}(v_2, v_1)$ ayant une probabilité $\alpha/2$ d'être dépassée. Si on revient au rapport $s_1'^2 / s_2'^2$, $1/F_{\alpha/2}(v_2, v_1)$ constituera la limite inférieure $F_{1-(\alpha/2)}(v_1, v_2)$ ayant une probabilité $1 - (\alpha/2)$ d'être dépassée. C'est la borne inférieure de l'intervalle.

Exemple

Tester l'hypothèse d'égalité des variances de deux populations dans les conditions suivantes : $n_1 = 3$ et $s_1'^2 = 7$; $n_2 = 5$ et $s_2'^2 = 2$; $\alpha = 0,05$.

$$F = \frac{s_1'^2}{s_2'^2} = 3,5$$

$F_{\alpha/2}(2, 4) = 10,6$ avec 2 degrés de liberté au numérateur et 4 degrés de liberté au dénominateur.

$F_{\alpha/2}(4, 2) = 39,2$ avec 4 degrés de liberté au numérateur et 2 degrés de liberté au dénominateur.

$$F_{1-(\alpha/2)}(2, 4) = \frac{1}{F_{\alpha/2}(4, 2)} = \frac{1}{39,2}$$

On a donc :

$$F_{1-(\alpha/2)}(2, 4) < F < F_{\alpha/2}(2, 4)$$

et l'hypothèse d'égalité des variances est acceptée.

■ **Remarque :** en pratique, on se contente de vérifier que $F < F_{\alpha/2}$ en mettant en numérateur la variance la plus élevée.

6.2.2.3 Généralisation : tests de comparaison de plusieurs variances

Il existe de nombreux tests permettant de contrôler l'hypothèse d'égalité des variances de plusieurs populations. Ces tests sont, en

général, peu efficaces et assez approximatifs. En principe, il sera nécessaire d'en effectuer plusieurs en recomposant leurs résultats. Parmi ces tests, citons celui de Harthy et celui de Cochran.

Le test de Harthy donne les valeurs maximales, aux seuils $\alpha = 5\%$ et 10% , de la variable aléatoire :

$$\frac{s_{\max}'^2}{s_{\min}'^2}$$

avec $s_{\max}'^2$ maximum des estimations obtenues,

$s_{\min}'^2$ minimum des estimations obtenues.

Le test de Cochran est basé sur la variable aléatoire :

$$g = \frac{s_{\max}'^2}{\sum_i s_i'^2}$$

qui donne les valeurs maximales g_α de cette variable au seuil α .

6.2.3 Test de comparaison de moyennes de populations

Nota : les tests qui vont être exposés dans ce paragraphe supposent qu'au préalable on ait testé avec succès l'égalité des variances.

6.2.3.1 Cas d'échantillons de taille élevée

Rappelons que la moyenne \bar{x} d'un échantillon de taille n , extrait d'une population normale de moyenne m et d'écart-type σ , suit une loi normale de moyenne m et d'écart-type σ/\sqrt{n} (§ 2.7). Soient deux échantillons, de taille n_1 et n_2 , de moyenne \bar{x}_1 et \bar{x}_2 , et de variance $s_1'^2$ et $s_2'^2$.

Faisons l'hypothèse suivante : les moyennes des populations sont égales : $m_1 = m_2 = m$. Nous savons déjà que $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Nous savons alors que \bar{x}_1 et \bar{x}_2 sont des variables normales, de moyenne m , et d'écart-type $\sigma/\sqrt{n_1}$ et $\sigma/\sqrt{n_2}$. La variable $\bar{x}_1 - \bar{x}_2$ suit une loi normale, de moyenne nulle, et d'écart-type $\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

Rappelons (§ 3.5.2.4) que l'estimation s' de σ est donnée par :

$$s'^2 = \frac{n_1 s_1'^2 + n_2 s_2'^2}{n_1 + n_2 - 2}$$

On considère alors la variable aléatoire normale réduite :

$$u = \frac{\bar{x}_1 - \bar{x}_2}{s' \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

On aura une probabilité $1 - \alpha$ pour que cette variable soit comprise dans l'intervalle $[-u_\alpha, u_\alpha]$. Si u est compris dans cet intervalle, rien ne s'oppose à l'hypothèse $m_1 = m_2 = m$. Dans le cas contraire, on peut repousser l'hypothèse, avec un risque au plus égal à α .

Exemple

Comparer les moyennes de deux populations, les résultats suivants ayant été obtenus sur échantillons : $n_1 = 150$; $\bar{x}_1 = 80,5$; $s_1^2 = 6,0$; $s_1 = 2,45$ et $n_2 = 250$; $\bar{x}_2 = 81,7$; $s_2^2 = 6,4$; $s_2 = 2,53$.

Testons d'abord l'égalité des variances au seuil 5 % (§ 6.2.2.1) :

$$s'^2 = 6,25$$

$$s' = 2,5$$

$$\frac{s_2 - s_1}{s' \sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}} \approx 0,5 < u_\alpha \text{ puisque } u_\alpha = 1,96$$

On peut considérer les deux variances comme identiques.

Revenons à la comparaison des moyennes :

$$\frac{\bar{x}_1 - \bar{x}_2}{s' \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \approx 4,8 > u_\alpha$$

L'hypothèse d'égalité des moyennes est à rejeter.

6.2.3.2 Cas d'échantillons de taille réduite

Le cas est identique au précédent, mais la variable :

$$\frac{\bar{x}_1 - \bar{x}_2}{s' \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

suit une loi de Student-Fisher à $n_1 + n_2 - 2$ degrés de liberté.

Le test est conduit comme précédemment mais en utilisant les valeurs t_α de la loi de Student-Fisher.

Exemple

Tester l'hypothèse d'égalité des moyennes de deux populations, les résultats suivants ayant été obtenus sur deux échantillons : $n_1 = 8$; $\bar{x}_1 = 14,2$; $s_1 = 0,1$ et $n_2 = 5$; $\bar{x}_2 = 14,5$; $s_2 = 0,2$.

Testons l'égalité des variances (§ 6.2.2.2) :

$$F = \frac{s_2^2}{s_1^2} \frac{n_2(n_1 - 1)}{n_1(n_2 - 1)} \approx 4,37$$

$$F_{0,025}(4,7) = 5,52 > F$$

On peut accepter l'hypothèse d'égalité des variances.

Revenons à la comparaison des moyennes :

$$s'^2 \approx 2,55 \times 10^{-2}$$

$$s' \approx 0,16$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s' \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ |t| \approx 3,3$$

Or t_α , avec 11 degrés de liberté, vaut 2,201 pour $\alpha = 0,05$; on peut refuser l'hypothèse d'égalité des moyennes.

6.2.3.3 Généralisation : tests de comparaison de plusieurs moyennes

On peut, d'une façon plus générale, tester l'égalité des moyennes de plusieurs populations par une technique appelée *analyse de la variance* dont l'importance est telle qu'il lui sera consacré un paragraphe entier (§ 8).

7. Tests non paramétriques**7.1 Généralités**

Lorsque les populations que l'on désire comparer ont des lois de probabilité connues, comme nous l'avons vu (§ 6), le problème se réduit à la comparaison des paramètres des lois de probabilité. Quand les variables aléatoires à comparer ont des lois de probabilité inconnues, les tests ne sont plus formulés de la même façon ; ils sont dits alors *non paramétriques*.

Nous allons, dans la suite de ce paragraphe, exposer quelques-uns de ces tests. On remarquera que ces tests sont, en général, assez simples à effectuer, mais il faudra se souvenir que leur efficacité est assez médiocre.

Après avoir étudié les tests de comparaison de populations (§ 7.2, 7.3 et 7.4), nous donnerons quelques notions sur les tests concernant les suites d'observations dont les principes sont comparables (§ 7.5 et 7.6).

7.2 Test des rangs de Wilcoxon**7.2.1 Cas d'échantillons de taille élevée**

Supposons que l'on ait extrait deux échantillons, de taille n_1 et n_2 , de deux populations repérées par les variables aléatoires X et Y . Soient x_1, x_2, \dots, x_n , les valeurs données par le premier échantillon, et y_1, y_2, \dots, y_m , celles données par le deuxième échantillon. On fait l'hypothèse que les deux populations ont la même loi de probabilité ; on peut ainsi supposer que les deux échantillons sont issus d'une seule et même population. On mélange les deux échantillons, et on range les valeurs des variables aléatoires des deux échantillons en une seule suite croissante, par exemple :

$$x_1, y_1, x_2, x_3, y_2, x_4, \dots$$

On relève ensuite les rangs des valeurs x dans cette suite : r_1, r_2, \dots, r_n (ainsi $r_1 = 1, r_2 = 3, r_3 = 4, r_4 = 6$).

On considère la variable aléatoire $\bar{r} = \frac{\sum r_i}{n}$.

Il est aisé de montrer que :

$$E(\bar{r}) = \frac{m+n+1}{2}$$

$$\text{Var}(\bar{r}) = \frac{(m+n+1)m}{12n}$$

et que la variable réduite $u = \frac{\bar{r} - E(\bar{r})}{\sqrt{\text{Var}(\bar{r})}}$ suit une loi normale quand n croît indéfiniment.

Si u est compris dans l'intervalle $[-u_\alpha, u_\alpha]$, on accepte l'hypothèse ; dans le cas contraire, elle est refusée, au risque α .

Remarque : il arrive, en pratique, qu'on rencontre des valeurs x et y égales. On leur affecte alors, comme rang, la moyenne des rangs qu'elles occuperaient si elles étaient différentes. Considérons, par exemple, la suite :

$$x_1, y_1, y_2, x_2, \begin{cases} x_3 \\ y_3 \end{cases}, \dots \text{ où } x_3 = y_3$$

Si $x_3 > y_3$, le rang de y_3 serait égal à 5, celui de x_3 à 6. On leur affecte alors le rang fractionnaire 5,5.

Dans ces conditions, l'expression qui donne $E[\bar{r}]$ n'est pas modifiée ; par contre $\text{Var}(\bar{r})$ est donnée par :

$$\text{Var}(\bar{r}) = \frac{(m+n+1)(m+n)(m+n-1) - \sum_i T_i}{12n(m+n)} \cdot \frac{m}{m+n-1}$$

avec $T_i = s_i(s_i+1)(s_i-1)$

où s_i est égal au nombre de valeurs égales entre elles (il est bien entendu qu'il peut y avoir plusieurs groupes différents de valeurs égales entre elles, chacun étant repéré par la valeur donnée à l'indice i).

7.2.2 Cas d'échantillons de taille réduite

Si les échantillons sont de taille réduite, l'approximation normale n'est plus valable. On fait alors l'hypothèse que les deux échantillons sont extraits de la même population. On recherche, dans le cas donné, la probabilité d'observer une valeur de \bar{r} différant plus largement de $E[\bar{r}]$ que celle que l'on a trouvée. Si cette probabilité est faible, on rejettera l'hypothèse. En effet, plus les populations-mères sont différentes, plus la probabilité que \bar{r} soit différente de $E[\bar{r}]$ est grande. Nous allons expliciter ceci sur un exemple.

Exemple :

- 1^{er} échantillon : $x = 13,9 - 14,4 - 14,5 -$
- 2^e échantillon : $y = 13,2 - 13,5 - 13,8 - 14,0 - 14,2 -$

On obtient la suite unique indiquée dans le tableau 19.

$$\bar{r} = \frac{4+7+8}{3} \approx 6,33$$

ou
$$E[\bar{r}] = \frac{3+5+1}{2} = 4,5$$

Dans l'hypothèse d'identité des deux populations-mères, les rangs des valeurs x sont compris entre 1 et 8 ; il existe donc C_8^3 combinaisons possibles des r . On peut ainsi déterminer la probabilité P d'observer une valeur de \bar{r} supérieure ou égale à 6,33. Cela sera obtenu par les quatre combinaisons de r suivantes : 8, 7, 6-8, 7, 5-8, 7, 4-8, 6, 5 donc :

$$P = \frac{4}{C_8^3} = \frac{4}{56}$$

P étant faible, il paraît préférable de rejeter l'hypothèse d'égalité.

Tableau 19 – Test des rangs de Wilcoxon
(exemple du § 7.2.2)

Suite	y_1	y_2	y_3	x_1	y_4	y_5	x_2	x_3
Rang	1	2	3	4	5	6	7	8

7.3 Test des suites homogènes

Considérons les valeurs x et y des variables aléatoires de deux échantillons, respectivement de taille m et n , qu'on cherche à comparer. En regroupant les valeurs x et y en une seule suite croissante, on a obtenu : $x_1, x_2, y_1, x_3, y_2, y_3, y_4, x_4, x_5, y_5, y_6, x_6, y_7$

On appelle *suite homogène* ou *run* tout ensemble formé d'une ou plusieurs valeurs consécutives tirées du même échantillon et encadré par des valeurs de l'autre échantillon. Dans la suite ci-dessus, on observe les suites homogènes suivantes :

$$\begin{matrix} x_1 & x_2 & & x_3 & & x_4 & x_5 & & x_6 \\ y_1 & & y_2 & y_3 & y_4 & & y_5 & y_6 & & y_7 \end{matrix}$$

On conçoit aisément que, si les deux échantillons sont tirés d'une même population, le nombre total des suites homogènes aura tendance à être plus élevé. Dans cette hypothèse, si R est le nombre de suites homogènes, il constitue une variable aléatoire définie par :

$$E[R] = \frac{2mn}{m+n} + 1$$

$$\text{Var}(R) = \frac{2mn(2mn-m-n)}{(m+n)^2(m+n-1)}$$

Si m et n sont grands, la variable R suit approximativement une loi normale, et la procédure du test s'en déduit immédiatement.

Si m et n sont faibles, les valeurs de R ayant une probabilité de 0,975 et 0,025 d'être dépassées ont été tabulées en fonction de m et n , et le test est extrêmement simple. Ainsi, dans le cas présent : $m = 6, n = 7, R = 8$. Les tables donnent : $P(R > 3) = 0,975$ et $P(R > 12) = 0,025$. Rien ne s'oppose à ce qu'on accepte l'hypothèse que les deux échantillons sont tirés de populations identiques.

7.4 Test de la médiane

Reprenons, encore une fois, la suite unique constituée par les deux échantillons mêlés. On détermine la médiane M de cette suite unique, et on dénombre, pour chaque échantillon, les éléments situés de part et d'autre de M . On établit ainsi le tableau 20.

Tableau 20 – Test de la médiane : résultats observés

Échantillon	Observations inférieures à M	Observations supérieures à M	Totaux
$n^o 1$	m_i	m_s	m
$n^o 2$	n_i	n_s	n
Totaux	$\frac{m+n}{2}$	$\frac{m+n}{2}$	$m+n$

Si les deux échantillons avaient été prélevés dans deux populations identiques, on devrait obtenir, en moyenne, les effectifs théoriques du tableau 21.

Tableau 21 – Test de la médiane : résultats théoriques

Échantillon	Observations inférieures à M	Observations supérieures à M	Totaux
$n^{\circ} 1$	$\frac{m}{2}$	$\frac{m}{2}$	m
$n^{\circ} 2$	$\frac{n}{2}$	$\frac{n}{2}$	n
Totaux	$\frac{m+n}{2}$	$\frac{m+n}{2}$	$m+n$

On teste alors l'hypothèse d'identité des deux populations par un test du χ^2 dont la procédure a été décrite au paragraphe 6.2.1.2.

Remarque : si la médiane M s'identifie à une valeur observée (cas où $m+n$ est impair), on retirera cette valeur.

7.5 Tests de comparaison d'observations appariées

7.5.1 Définition

Supposons, par exemple, que l'on veuille comparer l'influence de deux traitements thermiques A et B sur les propriétés mécaniques d'une tôle métallique de qualité donnée. On découpe cette tôle en deux parties ; la première subira le traitement A , la seconde le traitement B . Cette opération sera répétée n fois, de façon à éliminer, autant que possible, les fluctuations d'une tôle à l'autre ; on mesurera, par exemple, la dureté, et on obtiendra les résultats du tableau 22.

Tableau 22 – Test de comparaisons d'observations appariées : résultats observés

Traitement	Tôle n° 1	Tôle n° n
A	x_1	x_n
B	y_1	y_n

On dira que les observations x_i y_i sont *appariées*.

7.5.2 Test des signes

Si les traitements A et B sont équivalents, les différences que l'on pourra observer entre x_i et y_i seront uniquement dues à la dispersion du traitement. Si on considère les signes des expressions $(x_i - y_i)$, la probabilité pour qu'ils soient positifs est égale à la probabilité pour qu'ils soient négatifs. On déterminera le nombre k de signes positifs, et on recherchera la probabilité d'obtenir un nombre supérieur ou égal à k si $k > N/2$ (au contraire, un nombre inférieur ou égal à k si $k < N/2$) grâce à la loi binomiale (n , $p = 1/2$). Si cette probabilité est trop faible, on refusera l'hypothèse d'identité.

Exemple

Dans une suite de 30 observations appariées, on dénombre 22 signes positifs. Or, les tables de la loi binomiale donnent $P(k \geq 22) = 0,0081$.

Il convient donc de refuser l'hypothèse d'identité, la probabilité étant très faible.

7.5.3 Test de Wilcoxon pour observations appariées

On considère, comme dans le test précédent (§ 7.5.2), les différences $(x_i - y_i)$. On constitue la suite des valeurs absolues de ces différences, et on leur attribue ainsi un rang r_i . On affecte à ce rang le signe de la différence.

Dans l'hypothèse d'identité des deux suites d'observations, on montre que la somme r des rangs r_i positifs est une variable aléatoire telle que :

$$E[r] = \frac{1}{4} n(n+1)$$

$$\text{Var}(r) = \frac{n(n+1)(2n+1)}{24}$$

Sa loi de probabilité est approximativement normale.

Remarque : en cas de différences égales, on procède comme on l'a indiqué en remarque, dans le paragraphe 7.2 avec :

$$\text{Var}(r) = \frac{n(n+1)(2n+1)}{24} - \frac{\sum r_i^2}{48}$$

Exemple

On compare deux ohmmètres A et B , de principes différents, en mesurant 20 résistances de 1 000 ohms environ ; les résultats obtenus sont indiqués dans le tableau 23.

Il y a treize différences positives. Effectuons tout d'abord le test des signes (§ 7.5.2) :

$$P(k \geq 13) = 0,067$$

La conclusion est de rejeter l'hypothèse d'équivalence des deux suites.

Effectuons maintenant le test de Wilcoxon :

$$\begin{aligned} r &= 170,5 \\ E[r] &= \frac{20 \times 21}{4} = 105 \\ \text{Var}(r) &= \frac{20 \times 21 \times 41}{24} - \frac{90}{48} \approx 715,6 \\ \frac{r - E[r]}{\sqrt{\text{Var}(r)}} &\approx 2,5 \end{aligned}$$

La conclusion de ce test est également de rejeter l'hypothèse, puisque $u_{0,05} = 1,96 < 2,5$.

7.6 Tests du caractère aléatoire d'une série d'observations

Supposons qu'on ait obtenu une suite chronologique de N observations. L'objet de ces tests est de déterminer le caractère aléatoire de cette suite, c'est-à-dire de savoir si la mesure est indépendante de son rang d'obtention, ou du facteur temps.

Tableau 23 – Test de Wilcoxon pour observations appariées (exemple du § 7.5.3)

Numéro de la résistance i	Résistance		$x_A - x_B$ (Ω)	Rang r_i
	x_A (Ω)	x_B (Ω)		
1	1 002	998	+ 4	+ 10,5
2	1 001	1 003	- 2	- 5
3	1 007	1 003	+ 4	+ 10,5
4	1 011	999	+ 12	+ 18
5	998	1 003	- 5	- 12,5
6	1 006	1 007	- 1	- 2
7	1 008	1 005	+ 3	+ 8
8	1 009	999	+ 10	+ 16
9	1 002	996	+ 6	+ 14,5
10	1 001	1 003	- 2	- 5
11	1 012	1 001	+ 11	+ 17
12	1 015	1 002	+ 13	+ 19
13	1 010	1 005	+ 5	+ 12,5
14	1 008	1 009	- 1	- 2
15	1 001	998	+ 3	+ 8
16	1 002	996	+ 6	+ 14,5
17	1 007	993	+ 14	+ 20
18	1 006	1 009	- 3	- 8
19	1 003	1 005	- 2	- 5
20	1 004	1 003	+ 1	+ 2

7.6.1 Test dérivé du test des suites homogènes

Soit M la médiane de la suite observée. Notons x les résultats supérieurs à M , et y les résultats inférieurs ; la suite devient par exemple :

x, yy, x, y, xxx, y , etc

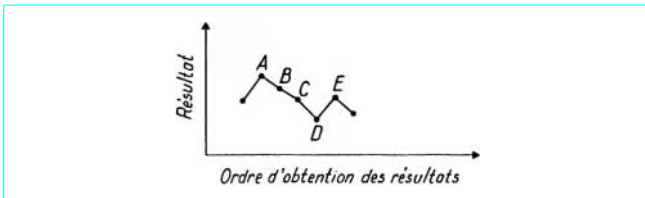


Figure 7 – Points critiques dans une suite

On teste alors le caractère aléatoire de la suite, comme indiqué au paragraphe 7.3, en prenant :

$$m = n = \frac{N}{2}$$

Remarque : si N est impair, on convient de ne pas considérer la valeur de l'observation égale à M .

7.6.2 Test des points critiques

On appelle *point critique* dans une suite tout résultat dont la valeur numérique est supérieure, ou inférieure, à celles des deux résultats qui l'encadrent chronologiquement. Par exemple, sur la figure 7, les points A, D, E sont des points critiques.

Dans l'hypothèse du caractère aléatoire de la suite, le nombre τ de points critiques est une variable aléatoire :

— de moyenne $E[\tau] = \frac{2}{3}(N-2)$;

— de variance $\text{Var}(\tau) = \frac{16N-29}{90}$;

— dont la loi de probabilité tend vers la loi normale, quand N tend vers l'infini (cette convergence est rapide).

La procédure du test en résulte immédiatement.

Exemple

Sur 50 observations chronologiques, au seuil 5 % ($u_\alpha = 1,96$), le nombre de points critiques devra être compris dans l'intervalle $[E[\tau] \pm 1,96 \sqrt{\text{Var}(\tau)}]$ soit $[32 \pm 6]$.

8. Analyse de la variance

8.1 Plans expérimentaux. Définitions générales

Des *plans expérimentaux* peuvent être définis comme les structures des expériences à mener, dans le but de mettre en évidence l'influence de certains facteurs sur des éléments de base. L'influence de ces facteurs peut avoir un caractère aléatoire ; les éléments de base sont eux-mêmes dispersés ; seules, des techniques statistiques pourront discerner, parmi les résultats obtenus, la part de variation due aux facteurs, de celle due à la dispersion des éléments de base. Ces techniques portent le nom général d'*analyse de la variance*.

Imaginons, par exemple, que l'on veuille tester l'influence d'un type de recuit thermique sur des pièces mécaniques provenant de plusieurs fournisseurs. Nous pourrions distinguer deux types de facteurs connus, susceptibles d'influencer les résultats : d'une part le facteur traitement thermique, d'autre part le facteur fournisseur. Ces facteurs sont dits *facteurs contrôlés* ; ils pourront comporter plusieurs *variantes*, par exemple des températures diverses pour le recuit, les fournisseurs différents pour le second facteur.

Les plans expérimentaux sont alors conçus de sorte que, dans une seule expérience globale, on puisse combiner l'action de tous les facteurs contrôlés l'un sur l'autre. Ils sont également construits de façon à obtenir la meilleure efficacité avec le nombre minimal d'essais expérimentaux.

L'analyse de la variance permettra alors d'isoler, dans les résultats, la part d'influence des facteurs contrôlés.

8.2 Hypothèses fondamentales

Soit x la caractéristique étudiée sur le matériel de base. Elle constitue une variable aléatoire. Nous allons formuler les hypothèses concernant cette variable, hypothèses qui constituent la base de l'analyse de variance.

■ **Première hypothèse** : la variable aléatoire x suit une loi de probabilité normale, de moyenne m , d'écart-type σ .

■ **Deuxième hypothèse** : les échantillons utilisés par l'expérimentation sont constitués de telle sorte que les éléments de base soient indépendants entre eux. Par exemple, si les éléments de base sont des pièces mécaniques, la cote x d'une pièce est indépendante des cotes x_j des pièces fabriquées antérieurement. Si ce n'est pas le cas, les éléments de base seront prélevés au hasard, et non pas régulièrement dans la série chronologique de fabrication.

■ **Troisième hypothèse** : l'action du facteur contrôlé sur les éléments de base se limite à modifier la moyenne m de la caractéristique. Autrement dit, après action d'un facteur contrôlé A , la variable

normale x , de moyenne m , d'écart-type σ , devient une variable normale x' , de moyenne $m' = m + a$, de même écart-type σ ; l'action du facteur contrôlé n'est pas une variable aléatoire.

■ **Quatrième hypothèse** : l'action de plusieurs facteurs contrôlés est additive. Autrement dit, si :

- après le traitement A , x devient $x + a$;
- après le traitement B , x devient $x + b$;
- après le traitement C , x devient $x + c$;

en faisant agir simultanément A, B, C , x devient $x + a + b + c$, variable aléatoire, de moyenne $m + a + b + c$, d'écart-type σ .

8.3 Analyse de la variance à un facteur contrôlé

Il s'agit, par exemple, de tester l'influence d'un recuit thermique sur les cotes des pièces mécaniques, en fonction de la température du recuit. Le facteur contrôlé est la température. Il existe autant de variantes que de températures différentes expérimentées. Les résultats peuvent être formulés dans le tableau [24](#).

Tableau 24 – Analyse de la variance à un facteur contrôlé : résultats expérimentaux						
Pièce	Température					
	T_1	T_2	...	T_i	...	T_k
$n^o 1$	x_{11}	x_{21}	...	x_{i1}	...	x_{k1}
.
.
$n^o n$	x_{1n}	x_{2n}	...	x_{in}	...	x_{kn}

Tout résultat x_{ij} peut être alors représenté par l'expression :

$$x_{ij} = a_i + \alpha_{ij}$$

avec a_i valeur qui intègre l'influence de la variante i du facteur contrôlé,

α_{ij} variable aléatoire normale, de moyenne 0, et d'écart-type σ , représentant les fluctuations dues à tous les facteurs non contrôlés.

8.3.1 Principe

Nous allons faire l'hypothèse que le facteur température est sans influence, et que les valeurs x peuvent être considérées comme obtenues à partir d'une population unique. Adoptons les notations suivantes :

$$x_{i.} = \frac{1}{n} \sum_{j=1}^n x_{ij} \quad (\text{avec } i = 1 \text{ à } k)$$

$$x_{..} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^n x_{ij} \quad (\text{avec } N = n k)$$

$$= \frac{1}{k} \sum_{i=1}^k x_{i.}$$

Considérons l'expression $\sum_i \sum_j (x_{ij} - x_{..})^2$: c'est le numérateur de la variance de l'ensemble des x_{ij} considérés comme appartenant à une même population. En posant :

$$(x_{ij} - x_{..}) = (x_{ij} - x_{i.}) + (x_{i.} - x_{..})$$

cette expression se met sous la forme :

$$\sum_i \sum_j (x_{ij} - x_{..})^2 = \sum_i \sum_j (x_{ij} - x_{i.})^2 + \sum_i \sum_j (x_{i.} - x_{..})^2$$

car les doubles produits s'annulent.

Soit σ^2 la variance de la population ; considérons les variables aléatoires suivantes :

$$A = \frac{\sum_i \sum_j (x_{ij} - x_{..})^2}{\sigma^2}$$

$$B = \frac{\sum_i \sum_j (x_{i.} - x_{..})^2}{\sigma^2} = \frac{n \sum_i (x_{i.} - x_{..})^2}{\sigma^2}$$

La première suit une loi du χ^2 à $N - 1$ degrés de liberté ; la seconde suit une loi du χ^2 à $k - 1$ degrés de liberté ([§ 2.6](#)).

Par conséquent, la variable :

$$C = \frac{\sum_i \sum_j (x_{ij} - x_{i.})^2}{\sigma^2} = A - B$$

suit une loi du χ^2 à $N - k$ degrés de liberté.

Un théorème dû à Cochran nous assure que B et C , dans le cadre de l'hypothèse énoncée au début de ce paragraphe [8.3.1](#), sont des variables indépendantes.

En vertu des propriétés des variables obéissant à la loi du χ^2 , nous pouvons écrire :

$$E \left[\frac{\sum_i n (x_{i.} - x_{..})^2}{\sigma^2} \right] = k - 1$$

et

$$E \left[\frac{\sum_i n (x_{i.} - x_{..})^2}{k - 1} \right] = \sigma^2$$

$$E \left[\frac{\sum_i \sum_j (x_{ij} - x_{i.})^2}{\sigma^2} \right] = N - k$$

et

$$E \left[\frac{\sum_i \sum_j (x_{ij} - x_{i.})^2}{N - k} \right] = \sigma^2$$

Dans ces conditions :

$$V_A = \frac{\sum_i n (x_{i.} - x_{..})^2}{k - 1}$$

et

$$V_R = \frac{\sum_i \sum_j (x_{ij} - x_{i.})^2}{N - k}$$

appelées respectivement *variance entre échantillons* (V_A), et *variance résiduelle* (V_R), sont deux estimations indépendantes de σ^2 .

Il sera maintenant possible de tester l'hypothèse, en testant l'égalité des deux estimations obtenues ([§ 6.2.2.2](#)).

8.3.2 Interprétation

Reprenons les notations $x_{ij} = a_i + \alpha_{ij}$ (rappelons que a_i n'est pas une variable aléatoire). Les quantités $x_{i.}$ et $x_{..}$ deviennent :

$$x_{i.} = a_i + \alpha_{i.}$$

$$x_{..} = a_{..} + \alpha_{..}$$

avec

$$a_i = \frac{1}{k} \sum_j a_{ij}$$

$$\alpha_{i.} = \frac{1}{n} \sum_j \alpha_{ij}$$

$$\alpha_{..} = \frac{1}{N} \sum_i \sum_j \alpha_{ij}$$

Dans ces conditions :

$$V_A = \frac{n \sum_i (a_i - a_{..})^2}{k-1} + \frac{n \sum_i (\alpha_{i.} - \alpha_{..})^2}{k-1}$$

Dans l'hypothèse d'influence du facteur contrôlé, le premier terme est égal à n fois l'estimation de σ_a^2 (variance de la variable a) ; le deuxième terme est une estimation de la variance σ^2 ; V_A est, par

conséquent, une estimation de $\sigma^2 + \frac{n \sum_i (a_i - a_{..})^2}{k-1}$; V_R est par ailleurs une estimation de σ^2 . Le test de comparaison des variances V_A et V_R revient donc bien à tester la nullité de $\sum_i (a_i - a_{..})^2$, c'est-à-dire la non-influence du facteur contrôlé ; en effet, $\sum_i (a_i - a_{..})^2 = 0$ entraîne $(a_i = a_{..})$.

8.3.3 Exécution pratique des calculs

Malgré les apparences, l'analyse de la variance ne nécessite pas des calculs très complexes, à condition d'exécuter habilement ceux qui sont nécessaires. Posons :

$$S_i = \sum_j x_{ij} = n x_{i.}$$

et

$$S = \sum_i S_i = N x_{..}$$

On se sert alors, pour le calcul, des formules suivantes :

$$\sum_i \sum_j (x_{ij} - x_{..})^2 = \sum_i \sum_j x_{ij}^2 - N x_{..}^2 = \sum_i \sum_j x_{ij}^2 - \frac{S^2}{N}$$

$$\sum_i n (x_{i.} - x_{..})^2 = \sum_i \frac{S_i^2}{n} - \frac{S^2}{N}$$

L'expression $\sum_i \sum_j (x_{ij} - x_{..})^2$ est obtenue par différence.

On présente pratiquement les calculs dans un tableau, ainsi que nous allons le montrer dans l'exemple suivant.

Exemples

Trois lots de poudre à fusil ont été fabriqués suivant trois procédés A, B, C . On effectue 7 tirs au fusil avec chacun de ces lots, et on relève les vitesses initiales de la balle (tableau 25).

Le problème est de tester la différence entre les trois procédés.

Pour simplifier les calculs, on pratiquera un changement d'origine, en mettant le nouveau zéro à 800 m/s. On établira ensuite les tableaux qui permettront de mener les calculs avec le minimum d'erreurs (tableaux 26 et 27).

On obtient alors :

$$\sum_i \sum_j (x_{ij} - x_{..})^2 = \sum_i \sum_j x_{ij}^2 - \frac{S^2}{N} \approx 228,95$$

$$\text{et} \quad \sum_i n (x_{i.} - x_{..})^2 = \sum_i \frac{S_i^2}{n} - \frac{S^2}{N} \approx 40,67$$

Par différence, il vient :

$$\sum_i \sum_j (x_{ij} - x_{i.})^2 \approx 188,28$$

Le test de comparaison des variances V_A et V_R est usuellement présenté dans un tableau appelé *tableau d'analyse de la variance* (tableau 27). Le principe de ce test est indiqué au paragraphe 6.2.2.2.

Or, $F_{0,025}(2,18) \approx 4,59$. Par conséquent ($1,94 < 4,59$), rien ne s'oppose à l'hypothèse que les procédés A, B, C donnent des résultats identiques.

8.3.4 Règles d'exécution de l'analyse de la variance à un facteur contrôlé

Récapitulons ces règles d'exécution :

- faire les changements de variables nécessaires pour simplifier le plus possible les données ;
- calculer les expressions $\sum_i \sum_j (x_{ij} - x_{..})^2$ et $\sum_i n (x_{i.} - x_{..})^2$;
- en tirer l'expression $\sum_i \sum_j (x_{ij} - x_{i.})^2$;
- établir le tableau d'analyse de la variance (tableau 28).

On calcule alors la quantité V_A/V_R que l'on compare à la valeur F de la loi de Snédécour, à $k-1$ et $N-k$ degrés de liberté, ayant une probabilité α d'être dépassée.

Si $V_A/V_R < F_\alpha(k-1, N-k)$, l'hypothèse de non-influence du facteur contrôlé est acceptée. Si $V_A/V_R > F_\alpha(k-1, N-k)$, l'hypothèse est refusée.

8.3.5 Remarque

Il est évident que le problème que nous venons de traiter est identique au problème de comparaison des moyennes de plusieurs échantillons, où le facteur de variation possible est la différence des populations dont sont extraits les échantillons considérés (§ 6.2.3).

Tableau 25 – Analyse de la variance à un facteur contrôlé : résultats expérimentaux
(exemple du § 8.3.3)

Tir	Vitesse initiale de la balle (m/s)		
	Lot A	Lot B	Lot C
n° 1	801	809	795
n° 2	803	801	798
n° 3	804	805	802
n° 4	798	803	801
n° 5	805	800	803
n° 6	797	808	801
n° 7	802	801	804

Tableau 26 – Analyse de la variance à un facteur contrôlé : récapitulation des résultats (exemple du § 8.3.3)

Tir		Lot A		Lot B		Lot C		Résultats
		x_{1j}	x_{1j}^2	x_{2j}	x_{2j}^2	x_{3j}	x_{3j}^2	
n° 1		1	1	9	81	-5	25	
n° 2		3	9	1	1	-2	4	
n° 3		4	16	5	25	2	4	
n° 4		-2	4	3	9	1	1	
n° 5		5	25	0	0	3	9	
n° 6		-3	9	8	64	1	1	
n° 7		2	4	1	1	4	16	
Résultats	S_i	+ 10	...	+ 27	...	+ 4	...	$S = \sum_i S_i \approx 41$
	S_i^2	100	...	729	...	16	...	$S^2 \approx 1\,681$
	$\frac{1}{n} S_i^2$	14,29	...	104,14	...	2,29	...	$\frac{S^2}{N} \approx 80,05$ $\frac{1}{n} \sum_i S_i^2 \approx 120,72$
	$\sum_j x_{ij}^2$...	68	...	181	...	60	$\sum_i \sum_j x_{ij}^2 \approx 309$

Tableau 27 – Tableau d'analyse de la variance (exemple du § 8.3.3)

Somme des carrés	Degrés de liberté	Variance	Conclusion
$\sum_i n (x_{i.} - x_{..})^2 = 40,67$	$k - 1 = 2$	$V_A = 20,33$	$V_A/V_R = 1,94$
$\sum_i \sum_j (x_{ij} - x_{i.})^2 = 188,28$	$N - k = 18$	$V_R = 10,46$	
$\sum_i \sum_j (x_{ij} - x_{..})^2 = 228,95$	$N - 1 = 20$	$V = 11,45$	

Tableau 28 – Tableau d'analyse de la variance

Somme des carrés	Degrés de liberté	Variance	Conclusion
$\sum_i n (x_{i.} - x_{..})^2$	$k - 1$	$V_A = \frac{\sum_i n (x_{i.} - x_{..})^2}{k - 1}$	$\frac{V_A}{V_R}$
$\sum_i \sum_j (x_{ij} - x_{i.})^2$	$N - k$	$V_R = \frac{\sum_i \sum_j (x_{ij} - x_{i.})^2}{N - k}$	
$\sum_i \sum_j (x_{ij} - x_{..})^2$	$N - 1$	$V = \frac{\sum_i \sum_j (x_{ij} - x_{..})^2}{N - 1}$	

9. Corrélation et régression

9.1 Généralités

9.1.1 Liaison stochastique

Nous nous sommes intéressés, jusqu'à présent, à des populations où chaque élément n'était repéré que par la mesure d'un seul caractère. Dans de nombreux cas, cependant, ceci sera insuffisant, et il sera nécessaire de prendre en compte les valeurs de plusieurs caractères pour individualiser un élément. Il est ainsi beaucoup plus intéressant, pour la connaissance d'une population d'individus, d'étudier la distribution de la taille et du poids que d'étudier la distribution d'un seul de ces paramètres.

Il peut exister, entre les valeurs prises par ces caractères, des liaisons dont nous allons essayer de dégager la nature ([A 165] *Probabilités*).

Dans un premier cas, ces valeurs peuvent être liées par une relation fonctionnelle, et la donnée d'une ou plusieurs de ces valeurs détermine parfaitement les autres. Ainsi, les données de la pression et de la température d'une certaine masse d'un gaz parfait déterminent parfaitement son volume. On dit alors qu'il existe une *liaison fonctionnelle* entre ces caractères.

Dans un deuxième cas, les valeurs prises par les caractères étudiés sont totalement indépendantes ; par exemple, il n'y a aucune liaison entre le poids d'un objet et sa température.

Dans un troisième cas, cependant, il peut se faire que les caractères, sans être liés par une liaison fonctionnelle, ne sont pas totalement indépendants. Ainsi, quand on étudie la taille et le poids d'une population démographique, il y a une certaine tendance pour que le poids d'un individu augmente avec sa taille. Toutefois, la donnée de sa taille ne permet pas d'obtenir son poids. Il est possible, cependant, à taille donnée, d'obtenir la loi de répartition du poids, sa moyenne, son écart-type, etc. ou, ce qui est équivalent, de déterminer la probabilité pour que le poids soit compris entre des limites données. On dira, dans ce cas, qu'entre les deux caractères taille et poids, il existe une *liaison stochastique*, ou encore que ces deux caractères sont en *corrélation*.

Notons bien que l'existence d'une liaison stochastique entre plusieurs caractères n'implique pas, en général, une relation de cause à effet.

9.1.2 Étude de la liaison stochastique

Il existe deux procédés pour étudier la liaison stochastique : la *régression* (§ 9.1.2.1), et la *corrélation* (§ 9.1.2.2). Ces procédés diffèrent essentiellement par la technique de prélèvement des échantillons.

9.1.2.1 Régression

Considérons une population d'éléments à plusieurs caractères représentés par des variables X, Y, \dots . La régression distingue, entre celles-ci, des variables indépendantes et des variables dépendantes. Les *variables indépendantes* sont, par nature, celles dont l'expérimentateur est maître, et auxquelles il peut assigner une valeur donnée : en ce sens, elles sont indépendantes. Les *variables dépendantes* sont celles qui sont obtenues dans l'expérience. Elles sont aléatoires.

Ainsi, quand on veut étudier la résilience d'un matériau en fonction de sa température, on effectuera des épreuves à des températures données ; la température est la variable indépendante. De même, si on veut étudier la liaison entre la taille et le poids, en choisissant la taille comme variable indépendante, on constituera à l'intérieur de la population des sous-populations d'individus mesurant respectivement 1,50 m – 1,55 m – 1,60 m, etc. On prélèvera, ensuite des échantillons dans ces sous-populations, et on déterminera le poids des individus de ces échantillons.

Ainsi, la régression se caractérise par un *échantillonnage dirigé*, qui se rapproche des techniques classiques d'expérimentation où l'on s'assure des variations d'un certain nombre de facteurs, pour mesurer et étudier les variations inconnues des autres facteurs.

Remarque : le mot *indépendante* aura, dans ce paragraphe, une double signification : d'une part, celle utilisée ici (§ 9.1.2.1) où variable indépendante signifie : variable dont la valeur ne dépend pas de l'expérience ; d'autre part, celle dégagée dans l'article [A 165] *Probabilités*, où indépendance se comprend comme l'indépendance d'une variable aléatoire par rapport à une autre.

9.1.2.2 Corrélation

La corrélation, contrairement à la régression, ne distingue pas de variables dépendantes ni indépendantes, et toutes les variables sont aléatoires. Dans l'obtention des résultats, l'expérimentateur n'est maître d'aucune des variables. Le prélèvement est effectué au hasard dans la population ; sur tous les éléments prélevés, l'expérimentateur mesure les caractères étudiés. Lorsqu'il étudie la liaison taille-poids, il tire un échantillon au hasard dans la population, et mesure sur chaque individu sa taille et son poids.

9.2 Étude de la régression

Nous limiterons cet exposé à l'étude de la régression à deux caractères, ces caractères étant, de plus, repérés par des variables continues.

9.2.1 Généralités

Nous noterons X la variable indépendante, et Y la variable aléatoire dépendante, x et y leurs valeurs. La régression a deux objets :

- d'une part, établir s'il existe une liaison stochastique entre x et y ; et tester cette liaison stochastique, car le lecteur se doute déjà qu'on ne pourra jamais répondre par une affirmation absolue, mais simplement conclure que rien ne s'oppose à l'hypothèse qu'il existe une liaison entre les deux variables, ou au contraire que cette hypothèse a une faible probabilité d'être réalisée effectivement ;
- d'autre part, quand la liaison stochastique existe, la caractériser au moyen de la *ligne de régression* (graphe des moyennes de y lié par x que l'on notera y'), et par les distributions statistiques de y lié par x .

En fait, nous supposerons tout d'abord que la liaison stochastique existe, et nous rechercherons les moyens de la caractériser ; nous prendrons ensuite en considération les moyens de tester son existence (§ 9.3.2).

9.2.2 Détermination générale de la ligne de régression. Méthode des moindres carrés

Supposons que l'on ait pu connaître sa forme analytique :

$$y' = f(x, \alpha, \beta, \dots)$$

où α, β , etc. sont des paramètres par ailleurs inconnus. Il reste à déterminer, à partir des résultats obtenus, les valeurs des paramètres α, β, \dots .

L'expérimentation a permis d'obtenir, pour chaque valeur de la variable indépendante x_i , une valeur de y que nous noterons y_i . On recherchera alors des estimations des paramètres α, β, \dots par la *méthode du maximum de vraisemblance* (§ 3.3).

Si les distributions liées de y sont normales, on montre assez aisément que la fonction de vraisemblance a pour expression :

$$\mathcal{L} = \frac{1}{(2\pi)^{n/2} \sigma_{y/x_1} \sigma_{y/x_2} \dots} \exp \left[-\sum_i \frac{(y_i - y'_i)^2}{\sigma_{y/x_i}} \right]$$

avec σ_{y/x_i} écart-type de y lié par x .

Le maximum de la fonction de vraisemblance est obtenu quand la quantité Ω^2 est minimale, si on note :

$$\Omega^2 = \sum_i \frac{(y_i - y'_i)^2}{\sigma_{y/x_i}}$$

On prendra donc pour estimations des paramètres α, β, \dots les valeurs qui rendent minimale la quantité Ω^2 . Elles sont par conséquent solutions du système d'équations :

$$\begin{cases} \frac{\partial (\Omega^2)}{\partial \alpha} = 0 \\ \frac{\partial (\Omega^2)}{\partial \beta} = 0 \\ \dots \end{cases}$$

Ces équations sont appelées *équations normales* ou *équations de Gauss*. Leur résolution est particulièrement aisée quand $f(x)$ est une fonction linéaire des paramètres α, β, \dots . La méthode générale ainsi exposée porte le nom de *méthode des moindres carrés*.

9.3 Régression linéaire

D'une façon générale, on ne connaît pas la forme analytique de la fonction $y' = f(x)$. On choisira alors, d'une façon arbitraire, une forme analytique qui rendra le mieux compte du phénomène étudié. Les valeurs des paramètres seront obtenues par résolution des équations de Gauss. Si $f(x)$ ne dépend pas linéairement de α, β, \dots , le problème sera complexe.

Nous nous limiterons, dans ces conditions, à un cas très simple d'étude de la régression, qui se présente d'ailleurs très fréquemment en pratique : la régression linéaire.

La régression linéaire se caractérise par les hypothèses suivantes :

- les distributions de y lié par x sont normales ;
- les écarts-types de y lié par x sont indépendants de x ;
- la ligne de régression est une droite.

Nous savons déjà que ces conditions sont remplies, si les variables aléatoires sont distribuées suivant une loi normale à deux variables ([A 165] *Probabilités*).

9.3.1 Paramètres de la régression linéaire

9.3.1.1 Détermination de la droite de régression

Soit $y' = a + bx$ l'équation de la droite de régression. La quantité σ_{y/x_i} est constante ; notons que l'on appellera $\sigma_{y/x}^2$ *variance liée* ou *variance résiduelle*.

Par conséquent, et si l'on note :

$$\Omega'^2 = \sum_i (y_i - y'_i)^2 = \sum_i (y_i - a - bx_i)^2$$

les équations normales s'écrivent :

$$\frac{\partial (\Omega'^2)}{\partial a} = -\sum_i (y_i - a - bx_i) = 0$$

$$\frac{\partial (\Omega'^2)}{\partial b} = -\sum_i x_i (y_i - a - bx_i) = 0$$

La première équation montre que la somme algébrique des écarts entre y_i et y'_i est nulle : $\sum_i (y_i - y'_i) = 0$.

En divisant par n cette équation (n étant le nombre de couples x_i, y_i), on obtient :

$$\bar{y} = a + b\bar{x}$$

avec

$$\bar{y} = \frac{\sum_i y_i}{n}$$

$$\bar{x} = \frac{\sum_i x_i}{n}$$

L'équation de la droite de régression peut alors s'écrire :

$$y' - \bar{y} = b(x - \bar{x})$$

où b est appelé *coefficient de régression linéaire de y en x* . On voit ainsi que la droite de régression passe par le centre de gravité de l'ensemble des points représentant les résultats obtenus.

La deuxième équation de Gauss nous donne ensuite :

$$b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i x_i y_i - n \bar{x} \bar{y}}{\sum_i x_i^2 - n \bar{x}^2}$$

Dans les calculs numériques, seule la deuxième expression est à utiliser. On prendra soin de les exécuter avec précision, car le numérateur et le dénominateur de b sont souvent des grandeurs très voisines.

On montre que la valeur obtenue de b est une estimation non biaisée (§ 3.2.2) du coefficient de régression linéaire réel de la population.

9.3.1.2 Variance liée

La variance des observations y_i est :

$$s_y^2 = \frac{\sum_i (y_i - \bar{y})^2}{n}$$

C'est la *variance marginale* de la variable aléatoire y . On peut écrire :

$$y_i - \bar{y} = (y_i - y'_i) + (y'_i - \bar{y})$$

Dans ces conditions, s_y^2 devient :

$$s_y^2 = \frac{\sum_i (y_i - y'_i)^2}{n} + \frac{\sum_i (y'_i - \bar{y})^2}{n} + \frac{2 \sum_i (y_i - y'_i)(y'_i - \bar{y})}{n}$$

Le troisième terme est nul, compte tenu des équations de Gauss (§ 9.3.1.1). On peut alors écrire :

$$s_y^2 = \frac{\sum_i (y_i - y'_i)^2}{n} + b^2 \frac{\sum_i (x_i - \bar{x})^2}{n}$$

On montre que le premier terme de cette expression $\frac{\sum_i (y_i - y'_i)^2}{n}$ est une estimation de la variance liée, mais seule la quantité $\frac{\sum_i (y_i - y'_i)^2}{n-2}$ constitue une estimation non biaisée (§ 3.2.2).

9.3.1.3 Indice de corrélation

On peut caractériser l'intensité de la liaison stochastique entre les variables x et y par le rapport de la variance résiduelle $\sigma_{y/x}^2$ à la variance marginale σ_y^2 ; en fait, on utilise le complément à 1 de ce rapport. Soit r^2 cette quantité, appelée *indice de corrélation* : $r^2 = 1 - (\sigma_{y/x}^2 / \sigma_y^2)$. La valeur $r^2 = 0$ signifie l'indépendance ; la valeur $r^2 = 1$ signifie la liaison fonctionnelle. Dans le cas où x et y sont des variables aléatoires normales, cette quantité est égale au carré du coefficient de corrélation ([A 165] *Probabilités*).

Pratiquement, on calcule r^2 à l'aide des expressions suivantes :

$$r^2 = b^2 \frac{\sum_i (x_i - \bar{x})^2}{\sum_i (y_i - \bar{y})^2}$$

soit
$$r = b \frac{\sqrt{\sum_i (x_i - \bar{x})^2}}{\sqrt{n} \cdot s_y}$$

ou
$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{\sum_i x_i y_i - n \bar{x} \bar{y}}{\sqrt{\left(\sum_i x_i^2 - n \bar{x}^2\right) \left(\sum_i y_i^2 - n \bar{y}^2\right)}}$$

9.3.2 Tests sur la régression

Ayant déterminé les paramètres de la régression, il reste à répondre à la question suivante : la régression observée est-elle significative ?

9.3.2.1 Test du coefficient de régression linéaire

On peut écrire b sous la forme :

$$b = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}$$

Les x_i n'étant pas des variables aléatoires, et les y_i étant des variables aléatoires normales, b est une variable aléatoire normale, d'écart-type :

$$\sigma_b = \frac{\sigma_{y/x}^2}{\sum_i (x_i - \bar{x})^2}$$

Tester l'indépendance des variables x et y revient à tester l'hypothèse $b = 0$. En prenant l'estimation $s_{y/x}^2 = \frac{\sum_i (y_i - y'_i)^2}{n-2}$, on effectuera le test en considérant la variable :

$$t = \frac{b}{s_b} = \frac{b(n-2) \sum_i (x_i - \bar{x})^2}{\sum_i (y_i - y'_i)^2}$$

qui suit une loi de Student-Fisher à $n-2$ degrés de liberté.

9.3.2.2 Analyse de la variance

Reprenons l'expression :

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - y'_i)^2 + \sum_i (y'_i - \bar{y})^2$$

Le premier terme $\sum_i (y_i - \bar{y})^2$ explicite naturellement la variation totale de la variable ; le terme $\sum_i (y'_i - \bar{y})^2$ exprime la variation due à la régression linéaire ; le terme $\sum_i (y_i - y'_i)^2$ représente la variation résiduelle, c'est-à-dire la fluctuation de la variable aléatoire y_i autour de sa moyenne liée y'_i .

Supposons qu'il n'y ait pas de liaison stochastique entre les variables x et y . La valeur réelle β du paramètre b est nulle, et il est évident, dans ce cas, que la variance totale de y , σ_y^2 (variance réelle de la population) est égale à la variance liée $\sigma_{y/x}^2$ (variance liée réelle de la population).

Dans ces conditions, la variable aléatoire $\frac{\sum_i (y_i - \bar{y})^2}{\sigma_y^2}$ est distribuée suivant une loi du χ^2 à $n-1$ degrés de liberté. On montre

alors que $\frac{\sum_i (y_i - y'_i)^2}{\sigma_y^2}$ est distribuée suivant une loi du χ^2 à $n-2$ degrés de liberté, et par conséquent $\frac{\sum_i (y'_i - \bar{y})^2}{\sigma_y^2}$ suivant une loi

du χ^2 à 1 degré de liberté. Ces deux dernières quantités sont alors indépendantes, et la variable aléatoire :

$$F' = \frac{\sum_i (y'_i - \bar{y})^2}{1} \cdot \frac{(n-2)}{\sum_i (y_i - y'_i)^2}$$

suit une loi de Snédécour à 1 et $n-2$ degrés de liberté.

Dans le cas de dépendance stochastique, ceci n'est plus vrai, et on montre que la variable F' est significativement plus grande que 1, mais ne suit pas une loi de Snédécour.

Par conséquent, pour tester l'indépendance de deux variables, on calculera la quantité F' que l'on comparera à la valeur prise par la variable F de Snédécour à un seuil α .

Remarque : on pourrait être tenté de tester la régression, en comparant les variances σ_y^2 et $\sigma_{y/x}^2$. Ceci n'est malheureusement pas possible, car les estimations s_y^2 et $s_{y/x}^2$ obtenues ne sont pas indépendantes. On se tire d'affaire, par contre, en comparant $s_{y/x}^2$ et $\sum_i (y'_i - \bar{y})^2$ qui sont bien indépendantes.

9.4 Exemple

Les couples de valeurs observées sont rassemblés dans le tableau 29.

Tableau 29 – Régression : résultats expérimentaux
(exemple du § 9.4)

x_i	1	2	3	4	5	6	7	8	9	10
y_i	0,51	1,30	1,42	2,33	3,65	5,04	4,60	5,41	6,52	7,63

9.4.1 Détermination de la droite de régression et des paramètres de la régression

Tous les calculs sont condensés dans le tableau 30.

$$b = \frac{\sum_i x_i y_i - n \bar{x} \bar{y}}{\sum_i x_i^2 - n \bar{x}^2} \approx \frac{275,64 - 211,25}{385 - 302,50} \approx 0,78$$

L'équation de la droite de régression (figure 8) sera, par conséquent :

$$(y' - 3,84) = 0,78 (x - 5,5)$$

ou $y' = 0,78x - 0,45$

La variance liée $s_{y/x}^2$ sera obtenue en explicitant l'expression :

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - y'_i)^2 + b^2 \sum_i (x_i - \bar{x})^2$$

$$\sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n \bar{y}^2 \approx 51,81$$

$$b^2 \sum_i (x_i - \bar{x})^2 = b^2 \left(\sum_i x_i^2 - n \bar{x}^2 \right) \approx 50,19$$

$$\sum_i (y_i - y'_i)^2 \approx 1,62$$

$$s_{y/x}^2 = \frac{\sum_i (y_i - y'_i)^2}{n - 2} \approx \frac{1,62}{8} \approx 0,20$$

De même :

$$r^2 = \frac{b^2 \sum_i (x_i - \bar{x})^2}{\sum_i (y_i - \bar{y})^2} \approx 0,97$$

9.4.2 Test de la régression

On effectue l'analyse de la variance (§ 8) en résumant les calculs dans le tableau 31.

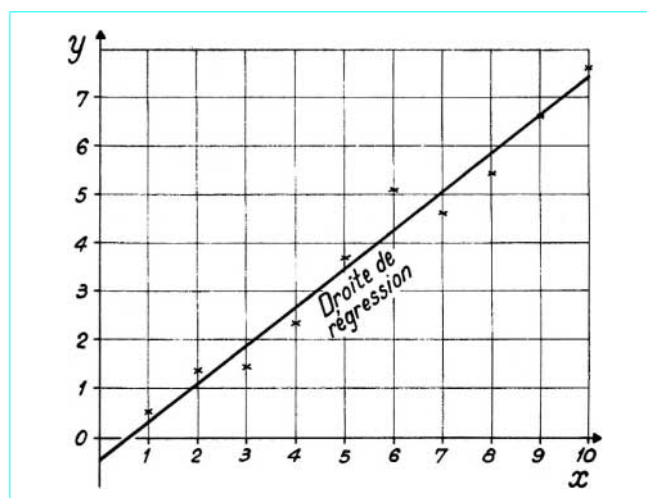


Figure 8 – Droite de régression

Conclusion

Sans faire usage de la table de Snédécour, on peut conclure que la liaison stochastique entre x et y est hautement significative (§ 9.3.2.2) ; il est par conséquent inutile d'effectuer le test sur le coefficient de régression linéaire (§ 9.3.2.1).

9.5 Corrélation

Nous avons déjà dit (§ 9.1.2.2) que la corrélation se distingue de la régression uniquement par le fait que les variables sont toutes aléatoires.

On montre que tous les résultats obtenus sur la régression linéaire sont transposables à la corrélation.

9.5.1 Coefficient de corrélation

Rappelons quelques résultats de l'article [A 165] *Probabilités*.

On définit le *coefficient de corrélation* par l'expression :

$$\rho = \frac{C}{\sigma_x \sigma_y}$$

avec C covariance, σ_x , σ_y écarts-types.

Quelle que soit la loi de probabilité du couple de variables aléatoires X et Y , ρ est compris entre -1 et $+1$ et :

- $\rho = 0$ si X et Y sont indépendantes entre elles ;
- $|\rho| = 1$ si X et Y sont liées par une relation fonctionnelle.

Si la loi de probabilité de X et Y est normale, réciproquement :

- $\rho = 0$ entraîne l'indépendance de X et Y ;
- $\rho = 1$ implique une relation fonctionnelle entre X et Y .

Tableau 30 – Droite de régression et paramètres de la régression : leur détermination (exemple du § 9.4)

Couple $x_i y_i$	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
$n^o 1$	1	0,51	0,51	1	0,260
$n^o 2$	2	1,30	2,60	4	1,690
$n^o 3$	3	1,42	4,26	9	2,016
$n^o 4$	4	2,33	9,32	16	5,429
$n^o 5$	5	3,65	18,25	25	13,322
$n^o 6$	6	5,04	30,24	36	25,402
$n^o 7$	7	4,60	32,20	49	21,160
$n^o 8$	8	5,41	43,28	64	29,268
$n^o 9$	9	6,52	58,68	81	42,510
$n^o 10$	10	7,63	76,30	100	58,217
Totaux	55	38,41	275,64	385	199,274
Résultats	$\bar{x} = 5,5$	$\bar{y} = 3,84$	$n \bar{x} \bar{y} = 211,25$	$n \bar{x}^2 = 302,50$	$n \bar{y}^2 = 147,46$

Tableau 31 – Test de la régression (exemple du § 9.4)

Somme des carrés	Degrés de liberté	Variance	Conclusion
$\sum_i (y'_i - \bar{y})^2 \approx 50,19$	1	Variance due à la régression $\approx 50,19$	$\frac{\sum_i (y'_i - \bar{y})^2}{1} \cdot \frac{(n-2)}{\sum_i (y_i - y'_i)^2} \approx 250$
$\sum_i (y_i - y'_i)^2 \approx 1,62$	$n - 2$	Variance résiduelle $\approx 0,20$	
$\sum_i (y_i - \bar{y})^2 \approx 51,81$	$n - 1$	Variance totale $\approx 5,76$	

9.5.2 Tests sur le coefficient de corrélation

Supposons que la loi de probabilité des variables X et Y soit normale. On obtiendra une estimation de ρ en utilisant l'expression :

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{\sum_i x_i y_i - n \bar{x} \bar{y}}{n \cdot \sigma_x \cdot \sigma_y}$$

On montre que r est une estimation convergente (§ 3.2.1) mais biaisée (§ 3.2.2) de ρ . Sa loi de probabilité est complexe.

Toutefois, si la taille de l'échantillon est grande ($n > 100$), la loi de distribution de r peut être considérée comme une loi normale, de moyenne ρ , et d'écart-type $\frac{1-\rho^2}{\sqrt{n-1}}$. Il est par conséquent possible

de tester l'hypothèse $\rho = 0$ au seuil α , en comparant $r\sqrt{n-1}$ à la valeur u_α d'une variable normale réduite u , telle que $P(|u| > u_\alpha) = \alpha$:

- si $r\sqrt{n-1} < u_\alpha$, l'hypothèse $\rho = 0$ est acceptée ;
- si $r\sqrt{n-1} > u_\alpha$, l'hypothèse est refusée.

Pour des valeurs de n comprises entre 30 et 100, on utilise le procédé dit *procédé de la corrélation transformée de Fisher*. Considérons la variable aléatoire :

$$Z = \frac{1}{2} \log_e \frac{1+r}{1-r} = \arg \operatorname{th} r$$

On montre que la variable Z suit sensiblement une loi normale, de moyenne $\frac{1}{2} \log_e \frac{1+\rho}{1-\rho}$, et d'écart-type $\frac{1}{\sqrt{n-3}}$.

Cette transformation a été tabulée. Elle permet de résoudre les problèmes d'intervalles de confiance, de comparaison à un standard (en particulier la comparaison de r à 0).

10. Généralisation de l'analyse de la variance

Nota : rappelons que l'analyse de la variance a été étudiée au paragraphe 8.

10.1 Généralités

Supposons que l'expérimentation nous ait fourni n résultats x_i ($i = 1, 2, \dots, n$), et que ces résultats dépendent linéairement de k paramètres inconnus t_j ($j = 1, 2, \dots, k$) représentant les variations dues aux facteurs contrôlés (les *facteurs contrôlés* ont été définis au paragraphe 8.1) :

$$x_i = a_{i1} t_1 + \dots + a_{ij} t_j + \dots + a_{ik} t_k + \varepsilon_i$$

Les coefficients a_{ij} sont connus. Les ε_i sont des variables aléatoires indépendantes normales, de moyenne nulle, d'écart-type σ , représentant les fluctuations de paramètres non contrôlés. Les estimations t'_j des paramètres inconnus t_j sont obtenues, comme nous

l'avons vu dans l'étude de la régression et de la corrélation, par la méthode des moindres carrés (§ 9.2.2). Elles sont telles que la quantité appelée *somme des carrés des résidus* :

$$\sum_i \left[\sum_j a_{ij} t'_j - x_i \right]^2 = \sum_i e_i^2$$

soit minimale.

En dérivant par rapport aux paramètres t'_j , on obtient un système de k équations à k inconnues, et on suppose qu'il possède une solution unique en t'_j .

On montre que les t'_j , solutions de ce système, sont des estimations absolument correctes (§ 3.2.2) et efficaces (§ 3.2.3) des t_j . La quantité $\sum_j a_{ij} t'_j$ représente la valeur ajustée de x_i .

10.2 Analyse de la variance

Posons :

- $S = \sum_i x_i^2$ somme des carrés des valeurs observées ;
- $Q = \sum_i \left(\sum_j a_{ij} t'_j \right)^2$ somme des carrés des valeurs ajustées ;
- $Q_R = \sum_i e_i^2$ somme des carrés des résidus.

On montre que :

- $\frac{Q_R}{n-k}$ constitue une estimation absolument correcte de σ^2 ; par

conséquent, $\frac{Q_R}{\sigma^2}$ est une variable aléatoire qui suit une loi du χ^2 à $n-k$ degrés de liberté ;

- $S = Q + Q_R$;
- les variables aléatoires Q et Q_R sont indépendantes ;
- dans l'hypothèse où les facteurs contrôlés sont sans influence sur les résultats x_i , c'est-à-dire si les paramètres t_j sont nuls, la quantité Q/k est une estimation de σ^2 , et Q/σ^2 est une variable aléatoire qui suit une loi du χ^2 à k degrés de liberté.

Dans ces conditions, la quantité :

$$F = \frac{Q}{Q_R} \frac{n-k}{k}$$

est une variable aléatoire qui suit la loi de Snédécour avec k et $n-k$ degrés de liberté.

10.3 Séparation des facteurs contrôlés en deux groupes

Supposons que les k facteurs t_j se séparent en deux groupes u_1, \dots, u_{k_1} , et v_1, \dots, v_{k_2} , avec $k_1 + k_2 = k$.

En pratiquant l'ajustement linéaire sur les k_1 paramètres u , on obtiendra :

$$S = Q(U) + Q_R(U)$$

De même, avec l'ajustement linéaire sur l'ensemble des paramètres U et V :

$$S = Q(U, V) + Q_R(U, V)$$

Dans l'hypothèse où les facteurs v sont sans influence sur les résultats, c'est-à-dire si les paramètres v_j sont nuls, on montre que la quantité $\frac{Q(U, V) - Q(U)}{\sigma^2}$ suit une loi du χ^2 à $k - k_1 = k_2$ degrés de liberté.

On testera donc l'hypothèse $v_j = 0$, à l'aide de la variable :

$$F = \frac{Q(U, V) - Q(U)}{k_2} \frac{n-k}{Q_R(U, V)}$$

avec k_2 et $n-k$ degrés de liberté.

De même, on testera l'hypothèse $u_j = 0$, en pratiquant l'ajustement linéaire des x_i sur les k_2 paramètres v , et en utilisant l'expression :

$$S = Q(V) + Q_R(V)$$

On voit que cette méthode se généralise aisément à un nombre de groupes de facteurs supérieur à 2.

10.4 Introduction des plans expérimentaux

Supposons que l'on ait encore deux groupes de facteurs contrôlés u et v . Soient $Q(U)$ et $Q(V)$ les sommes des carrés des valeurs ajustées séparément, et $Q_R(U, V)$ la somme des carrés des résidus obtenus en ajustant simultanément x_i aux paramètres u et v . D'une façon générale, l'expression :

$$S = Q(U) + Q(V) + Q_R(U, V)$$

n'est pas vérifiée.

Cette expression peut cependant être vérifiée, moyennant une structure particulière des expérimentations, et un choix convenable des paramètres (propriété d'orthogonalité). Ceci est en particulier réalisé dans les plans d'expériences dits *factoriels*, ou du type latin et gréco-latin.

Nous n'aborderons dans cet exposé que l'étude des plans factoriels (§ 10.5).

10.5 Plans factoriels

10.5.1 Définition

Un plan d'expériences est du type factoriel si, comportant un nombre k quelconque de facteurs contrôlés U, V, W, \dots ayant respectivement k_1, k_2, \dots variantes, à toute combinaison U_i, V_j, \dots ($i = 1, \dots, k_1$; $j = 1, \dots, k_2$; etc.), il associe un même nombre de mesures.

10.5.2 Analyse de la variance dans le cas de deux facteurs contrôlés

Soient deux facteurs contrôlés U et V , à k_1 et k_2 variantes ; à chaque combinaison $U_i V_j$, correspond à une expérience, et par conséquent une mesure x_{ij} . Les mesures sont contenues dans le tableau 32.

On pose, d'une façon générale :

$$x_{ij} = m + u_i + v_j + \varepsilon_{ij}$$

où les paramètres u_i rendent compte de l'influence des facteurs U , et les paramètres v_j de celle des facteurs V . La quantité ε_{ij} est une variable aléatoire normale, de moyenne nulle, et d'écart-type σ .

Tableau 32 – Analyse de la variance dans le cas de deux facteurs contrôlés : résultats expérimentaux

Facteur V : variante	Facteur U : variante					Totaux
	1	...	i	...	k ₁	
1	x ₁₁	...	x _{i1}	...	x _{k₁1}	$\sum_i x_{i1} = k_1 x_{\cdot 1} = S_{\cdot 1}$
.
.
.
j	x _{1j}	...	x _{ij}	...	x _{k₁j}	$\sum_j x_{ij} = k_1 x_{\cdot j} = S_{\cdot j}$
.
.
.
k ₂	x _{1k₂}	...	x _{ik₂}	...	x _{k₁k₂}	$\sum_i x_{ik_2} = k_1 x_{\cdot k_2} = S_{\cdot k_2}$
Totaux	$\sum_j x_{1j} = k_2 x_{1\cdot} = S_{1\cdot}$...	$\sum_j x_{ij} = k_2 x_{i\cdot} = S_{i\cdot}$...	$\sum_j x_{k_1 j} = k_2 x_{k_1\cdot} = S_{k_1\cdot}$	$S_{ij} = \sum_i S_{i\cdot} = \sum_j S_{\cdot j} = \sum_{ij} x_{ij} = k_1 k_2 x_{\cdot\cdot}$

On détermine des estimations des paramètres m , u_i , v_j par la méthode des moindres carrés (§ 9.2.2) ; on montre alors que :

$$m' = \frac{\sum_{ij} x_{ij}}{k_1 k_2} = x_{\cdot\cdot}$$

$$u'_i = x_{i\cdot} - x_{\cdot\cdot}$$

$$v'_j = x_{\cdot j} - x_{\cdot\cdot}$$

De plus $\sum_i u'_i = 0$ et $\sum_j v'_j = 0$; par conséquent, il n'y a que $k_1 - 1$ estimations de u indépendantes, et $k_2 - 1$ estimations de v indépendantes.

Posons :

$$S = \sum_{ij} x_{ij}^2$$

$$Q(M, U, V) = \sum_{ij} (m' + u'_i + v'_j)^2$$

$$Q_R(M, U, V) = \sum_{ij} (m' + u'_i + v'_j - x_{\cdot\cdot})^2$$

$$Q(U) = k_2 \sum_i (x_{i\cdot} - x_{\cdot\cdot})^2$$

$$Q(V) = k_1 \sum_j (x_{\cdot j} - x_{\cdot\cdot})^2$$

$$Q(M) = k_1 k_2 x_{\cdot\cdot}^2$$

On a :

$$S = Q(M, U, V) + Q_R(M, U, V)$$

$$Q(M, U, V) = Q(M) + Q(U) + Q(V) \text{ (propriété d'orthogonalité)}$$

Pour simplifier, on pose en général :

$$Q = S - Q(M) = Q(U) + Q(V) + Q_R$$

L'analyse de la variance est basée sur les résultats suivants :

- $\frac{Q_R}{(k_1 - 1)(k_2 - 1)}$ est une estimation de σ^2 ;
- les variables aléatoires Q_R , $Q(U)$, et $Q(V)$ sont indépendantes ;
- si l'hypothèse $u_i = 0$ est réalisée, quel que soit i , la quantité

$$\frac{Q(U)}{k_1 - 1} \text{ est une estimation de } \sigma^2 ;$$

- si l'hypothèse $v_j = 0$ est réalisée, quel que soit j , la quantité

$$\frac{Q(V)}{k_2 - 1} \text{ est une estimation de } \sigma^2.$$

Ces hypothèses sont testées à l'aide des rapports :

$$F_1 = \frac{Q(U)}{Q_R} \frac{(k_1 - 1)(k_2 - 1)}{k_1 - 1}$$

avec $k_1 - 1$ et $(k_1 - 1)(k_2 - 1)$ degrés de liberté ;

$$F_2 = \frac{Q(V)}{Q_R} \frac{(k_1 - 1)(k_2 - 1)}{k_2 - 1}$$

avec $k_2 - 1$ et $(k_1 - 1)(k_2 - 1)$ degrés de liberté.

Nota : l'introduction de $m' = x_{\cdot\cdot}$, qui lie les estimations u'_i et v'_j , diminue de 1 le nombre de degrés de liberté par rapport à la théorie générale du paragraphe 10.2.

10.5.3 Exécution pratique des calculs

On commence par calculer Q , $Q(U)$ et $Q(V)$; Q_R est obtenu à partir de l'expression $Q_R = Q - [Q(U) + Q(V)]$.

■ Calcul de $Q(U)$

$$Q(U) = k_2 \sum_i (x_{i\cdot} - x_{\cdot\cdot})^2 = \frac{1}{k_2} \sum_i S_{i\cdot}^2 - \frac{S_{ij}^2}{k_1 k_2}$$

en posant

$$S_{i\cdot} = \sum_j x_{ij} \text{ et } S_{ij} = \sum_i x_{ij}$$

■ Calcul de $Q(V)$

$$Q(V) = k_1 \sum_j (x_{.j} - x_{..})^2 = \frac{1}{k_1} \sum_j S_{.j}^2 - \frac{S_{..}^2}{k_1 k_2}$$

en posant $S_{.j} = \sum_i x_{ij}$ et $S_{ij} = \sum_j S_{.j}$

■ Calcul de Q

$$Q = \sum_i \sum_j (x_{ij}^2 - x_{..}^2) = \sum_i \sum_j x_{ij}^2 - \frac{S_{..}^2}{k_1 k_2}$$

Exemple

On effectue des mesures avec quatre appareils différents et cinq opérateurs différents. On veut savoir si les appareils ou les opérateurs ont une influence sur la mesure.

On réalise le plan d'expériences suivant, dont les résultats sont donnés dans le tableau 33.

On effectue un changement d'origine et un changement d'échelle, et on présente les calculs dans le tableau 34.

$$Q(U) \approx \frac{2\,530}{4} - \frac{12\,100}{20} \approx 27,5$$

$$Q(V) \approx \frac{3\,814}{5} - \frac{12\,100}{20} \approx 157,8$$

$$\sum_i \sum_j x_{ij}^2 \approx 840 \quad \text{et} \quad Q \approx 840 - \frac{12\,100}{20} \approx 235$$

$$Q_R \approx 235 - (27,5 + 157,8) \approx 49,7$$

L'analyse de la variance est alors résumée dans le tableau 35.

Les valeurs de F ayant une probabilité de 5 % d'être dépassées sont :

— avec 4 et 12 degrés de liberté : 3,26 ;

— avec 3 et 12 degrés de liberté : 3,49.

Par conséquent, on peut conclure qu'il n'y a pas d'influence des opérateurs sur les mesures. Par contre, l'influence des appareils est significative.

10.5.4 Généralisation

Les méthodes exposées se généralisent assez aisément en théorie au cas de facteurs contrôlés en nombre supérieur à 2. De même, un nombre de mesures supérieur à 1 par combinaison $U_i V_j \dots$ ne pose pas de difficultés théoriques particulières ; toutefois, le volume considérable de calculs pratiques à effectuer fait sortir ces problèmes du cadre de cet exposé.

Tableau 33 – Analyse de la variance dans le cas de deux facteurs contrôlés : résultats expérimentaux (exemple du § 10.5.3)

Facteur appareil (V) : variante	Facteur opérateur (U) : variante				
	1	2	3	4	5
1	1,78	1,77	1,75	1,79	1,72
2	1,81	1,83	1,79	1,77	1,79
3	1,72	1,75	1,74	1,73	1,72
4	1,76	1,71	1,72	1,74	1,71

Tableau 34 – Analyse de la variance dans le cas de deux facteurs contrôlés : récapitulation des résultats (exemple du § 10.5.3)

Facteur appareil (V) : variante	Facteur opérateur (U) : variante						
	1	2	3	4	5	$S_{.j}$	$S_{.j}^2$
1	8	7	5	9	2	31	961
2	11	13	9	7	9	49	2 401
3	2	5	4	3	2	16	256
4	6	1	2	4	1	14	196
$S_{.i}$	27	26	20	23	14	$S_{ij} \approx 110$	$\sum_j S_{.j}^2 \approx 3\,814$
$S_{i.}^2$	729	676	400	529	196	$\sum_i S_{i.}^2 \approx 2\,530$	$S_{ij}^2 \approx 12\,100$

Tableau 35 – Tableau d’analyse de la variance (exemple du § 10.5.3)

Somme des carrés	Degrés de liberté	Variance	Conclusion
$Q(U) \approx 27,5$	4	Variance due au facteur U : $\text{Var}(U) \approx 6,87$	$\frac{\text{Var}(U)}{\text{Var}_R} \approx 1,7$
$Q(V) \approx 157,8$	3	Variance due au facteur V : $\text{Var}(V) \approx 52,6$	$\frac{\text{Var}(V)}{\text{Var}_R} \approx 12,7$
$Q_R \approx 49,7$	12	Variance résiduelle $\text{Var}_R \approx 4,14$	
$Q \approx 235$	19		

11. Contrôles statistiques industriels

11.1 Généralités

11.1.1 Contrôle à 100 % et contrôle par prélèvement

Les contrôles de fabrication ont toujours constitué une préoccupation de l’industrie. On distingue deux sortes de contrôles, d’une part le *contrôle* à 100 % qui consiste à inspecter la totalité des fabrications, et d’autre part le *contrôle par prélèvement* qui consiste à estimer la qualité des fabrications sur échantillon.

Le premier type n’est pas toujours réalisable, soit pour des raisons économiques, soit en raison de la nature du contrôle (contrôle destructif par exemple). Il n’offre d’ailleurs pas forcément des garanties beaucoup plus élevées que le contrôle par prélèvement (lassitude et négligence des contrôleurs effectuant à longueur de temps la même opération).

Le deuxième type de contrôle a été marqué, dans les dernières décennies, par des progrès considérables dus à l’application des méthodes statistiques les plus affinées. En particulier, l’effort industriel américain, pendant la deuxième guerre mondiale, a suscité l’essor et la mise en application de nombreuses techniques statistiques de contrôle.

11.1.2 Contrôle qualitatif et contrôle quantitatif

Sous l’angle de la nature même de l’opération de contrôle, on peut encore distinguer deux sortes de contrôle :

- les *contrôles qualitatifs* qui consistent à classer la fabrication ou le prélèvement en éléments bons et mauvais, ou encore à dénombrer les défauts de cette fabrication ; les contrôles qualitatifs sont aussi appelés *contrôles par calibres* ;
- les *contrôles quantitatifs* dans lesquels un certain paramètre des éléments de la fabrication, ou du prélèvement, est soumis à mesure ; pour cette raison, ces contrôles portent aussi le nom de *contrôles par mesures*.

11.1.3 Contrôle de fabrication et contrôle de réception

L’opération de contrôle d’une fabrication peut être opérée d’une façon dite continue, c’est-à-dire d’une façon permanente ; toutes les heures ou tous les jours, selon les cas, des éléments produits par une unité de fabrication seront prélevés, puis contrôlés, de façon à

estimer l’évolution de la qualité en fonction du temps. Ce type de contrôle permettra de suivre la qualité des fabrications, et indiquera très rapidement un fléchissement de celles-ci. Ce contrôle sera, pour cela, appelé *contrôle de fabrication*.

On pourra également effectuer le contrôle d’une façon discontinue, en opérant le prélèvement sur un certain volume des fabrications, correspondant en général à partie ou totalité d’une livraison. Ce contrôle, qui a pour objet de s’assurer que la qualité des produits livrés est conforme aux exigences du client, s’appelle, pour cette raison, *contrôle de réception*.

11.1.4 Notion de variabilité

La notion de contrôle statistique suppose que la fabrication constitue une population statistique, et que, lorsque cette fabrication est conforme aux normes de qualité, cette population statistique soit connue ou, pour le moins, qu’on en connaisse une estimation. On appellera *variabilité* l’ensemble des estimations de la loi de répartition de la fabrication, et de ses paramètres (en particulier ceux qui caractérisent la moyenne et la dispersion).

Il est évident, par exemple, qu’en contrôle de fabrication, la variabilité de la fraction de fabrication, sur laquelle est opéré le prélèvement, ne doit pas se modifier.

11.1.5 Détermination de la variabilité

Il est intuitif qu’avant de soumettre une fabrication quelconque à un contrôle, soit de fabrication, soit de réception, il est nécessaire d’en déterminer avec soin la variabilité.

Cette opération est délicate, la variabilité dépendant des matières premières utilisées, des machines employées, du personnel, et de la taille du prélèvement.

En effet, il n’est pas rare de voir la moyenne d’une fabrication se déplacer, par suite, soit du changement de fournisseur de matières premières, soit du changement de machines utilisées ou même par usure de celles-ci, ou encore par suite du changement du personnel servant ces machines. L’effet de ces facteurs peut être considérable ; il est toutefois aisé de les prendre en compte.

Il est, par contre, assez difficile de déterminer la taille n du prélèvement. En effet, si un prélèvement fourni donne, dans l’absolu, une meilleure estimation qu’un prélèvement de taille faible (intervalle de confiance de la moyenne et de l’écart-type variant comme $1/\sqrt{n}$), par contre, en augmentant n , on court le risque d’avoir des dérèglages pendant la production des fabrications sur lesquelles on opère le prélèvement. En conséquence, le choix de n sera fait en fonction de la cadence de fabrication de l’unité de production, et de la fréquence de ses dérèglages, ou de la loi de répartition statistique de ceux-ci dans le temps, si elle est connue.

11.1.5.1 Choix de la loi de probabilité

Pratiquement, on supposera tout d'abord que la loi de probabilité du caractère contrôlé est normale, dans le cas d'un contrôle par mesures. En effet, les causes élémentaires de dispersion du caractère contrôlé sont en général nombreuses et agissent additivement (ce n'est pas toujours le cas cependant ; ainsi, les excentricités de pièces cylindriques, axes, etc., sont distribuées suivant une loi du χ^2). Dans le cas de contrôle par calibres, le nombre de défectueux observés suivra la loi binomiale ou la loi de Poisson.

11.1.5.2 Détermination des paramètres

On effectue ensuite un certain nombre de prélèvements. On s'assure, au moyen d'une étude de variance, que tous ces prélèvements sont bien issus d'une même population, donc que d'un prélèvement à l'autre la fabrication est restée constante. On estime sur chacun de ces prélèvements la moyenne, l'écart-type, ou la proportion de défectueux, selon les cas. On prend alors la moyenne de ces estimations pour obtenir les paramètres de la fabrication.

Ainsi, supposons que sur r prélèvements de taille n , on ait obtenu les estimations $\bar{x}_1, \dots, \bar{x}_r$, de la moyenne, et s_1, \dots, s_r , de l'écart-type. On prendra, pour déterminer la variabilité, les valeurs :

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \dots + \bar{x}_r}{r}$$

$$s = \frac{s_1 + \dots + s_r}{r}$$

Retenons, pour conclure, que la détermination de la variabilité est une opération délicate, et qu'il n'y a pas de règle absolument générale. Elle doit être effectuée avec beaucoup de sens critique.

11.1.6 Notion d'efficacité d'un contrôle

Supposons que l'on effectue sur une fabrication un contrôle, assorti d'un critère de qualité. Si ce contrôle est un contrôle à 100 %, il va sans dire que, lorsque la qualité est dans les normes imposées, la probabilité d'acceptation de la fabrication est égale à 1. Elle est égale à 0, quand la qualité est en dehors des normes. Si, par contre, le contrôle est effectué par prélèvement, la qualité estimée est en général différente de la qualité réelle (c'est, nous le savons, une variable aléatoire). On conçoit aisément qu'une fabrication correcte puisse être jugée inacceptable et, inversement, qu'une fabrication défectueuse soit jugée acceptable.

On pourra ainsi calculer la probabilité P de juger acceptable une fabrication en fonction de sa qualité réelle Q . On appellera *efficacité du contrôle* la fonction $P = f(Q)$ ainsi définie, et *courbe d'efficacité* la représentation graphique de cette fonction.

La figure 9 donne l'allure générale de la courbe d'efficacité d'un contrôle par prélèvement ; on la comparera à la courbe d'efficacité d'un contrôle à 100 % (figure 10).

On notera, sur la courbe d'efficacité d'un contrôle par prélèvement, que si la qualité est par exemple représentée par le pourcentage p d'éléments défectueux, le contrôle par prélèvement fait courir un risque α de refuser une fabrication de qualité réelle p_1 jugée acceptable. Dans le contrôle de réception, ce risque est appelé *risque du fournisseur* (§ 11.3.3.1). De même, on encourt un risque β d'accepter une fabrication de qualité réelle p_2 jugée inacceptable. Ce risque est appelé *risque du client* (§ 11.3.3.2), dans le contrôle de réception. Pour des mêmes valeurs p_1 et p_2 , un contrôle est dit plus *efficace* qu'un autre, si les risques du fournisseur et du client sont plus faibles.

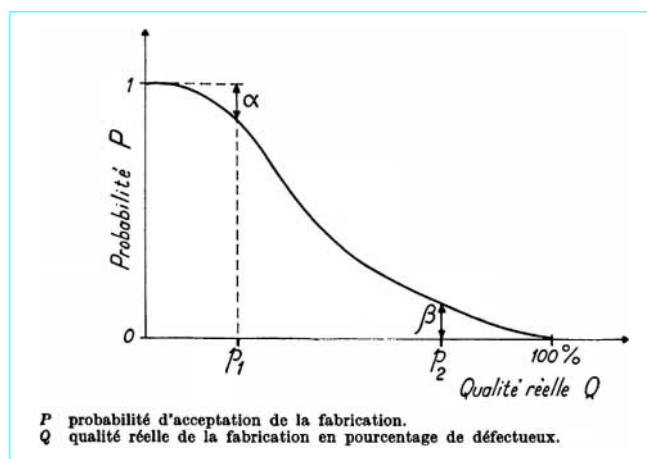


Figure 9 – Courbe d'efficacité d'un contrôle par prélèvement

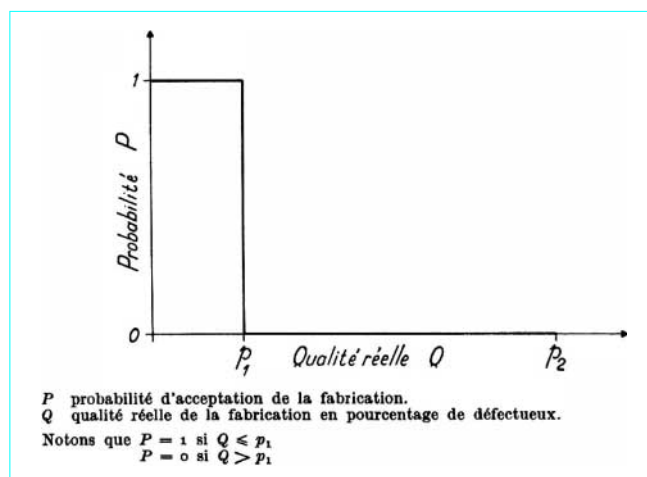


Figure 10 – Courbe d'efficacité d'un contrôle à 100 %

11.2 Contrôle de fabrication

11.2.1 Rappel de quelques résultats

Rappelons que lorsqu'on prélève un échantillon, de moyenne \bar{x} , et d'écart-type s , dans une population normale, de moyenne m , et d'écart-type σ , \bar{x} et s sont des variables aléatoires possédant les propriétés suivantes (§ 2.6) :

- \bar{x} suit une loi normale, de moyenne m , et d'écart-type $\frac{\sigma}{\sqrt{n}}$;
- le rapport $\frac{n s^2}{\sigma^2}$ suit une loi du χ^2 à $n - 1$ degrés de liberté.

11.2.2 Cartes de contrôle

11.2.2.1 Principe

En raison des résultats rappelés dans le paragraphe précédent, nous savons qu'il est possible de déterminer (§ 3.5.2), autour des valeurs m et σ de la fabrication, des intervalles dans lesquels \bar{x} et s auront une probabilité $1 - \alpha$ de se trouver, sous réserve que la variabilité de la fabrication ne se soit pas modifiée.

Si on observe, à la suite d'un prélèvement, une valeur de \bar{x} ou de s extérieure à cet intervalle, on pourra faire deux hypothèses :

- ou bien on observe une valeur improbable de \bar{x} ou de s , mais la variabilité de la fabrication n'a pas changé ;
- ou bien la variabilité de la fabrication s'est modifiée, ce qui explique la valeur observée de \bar{x} ou de s .

Le principe du contrôle de fabrication consiste à adopter la deuxième hypothèse. Le risque d'erreur ne peut, de toute façon, être supérieur à α .

En pratique, on prend un risque α extrêmement faible, égal à 0,002. On définit ainsi les *limites de contrôle* entre lesquelles on conclut que la fabrication n'est pas déréglée. En dehors de ces limites, on décide d'arrêter la fabrication et de procéder à un réglage.

On définit de même les *limites de surveillance*, mais avec un risque égal à 0,05. Lorsqu'on observe des valeurs de \bar{x} et de s à l'intérieur des limites de contrôle, mais à l'extérieur des limites de surveillance, cela peut signifier un début de dérèglement. Il est nécessaire, alors, de renforcer le contrôle pour le vérifier au plus tôt, et prendre au mieux les mesures nécessaires.

11.2.2.2 Détermination des limites de contrôle et de surveillance

Supposons que l'étude de la variabilité ait affecté à la fabrication une moyenne m , et un écart-type σ , par les méthodes exposées dans le paragraphe 11.1.5. Les limites de contrôle et de surveillance sont alors définies par les expressions suivantes.

■ Pour la moyenne :

- limites de contrôle : $m + A_{0,001} \sigma$
 $m - A_{0,001} \sigma$
- limites de surveillance : $m + A_{0,025} \sigma$
 $m - A_{0,025} \sigma$

■ Pour l'écart-type :

- limites de contrôle : $B_{0,999} \sigma$
 $B_{0,001} \sigma$
- limites de surveillance : $B_{0,975} \sigma$
 $B_{0,025} \sigma$

La variabilité étant normale, les coefficients A sont définis par :

$$A_{\alpha/2} = \frac{u_{\alpha}}{\sqrt{n}}$$

u_{α} étant la valeur d'une variable normale réduite, ayant une probabilité α d'être dépassée en valeur absolue : $u_{0,05} = 1,96$ et $u_{0,002} = 3,09$. Notons que A dépend de n . Pour cette raison, les valeurs de A sont tabulées en fonction de n .

De même, les coefficients B sont définis par les expressions :

$$B_{\alpha} = \sqrt{\frac{\chi^2_{1-\alpha}}{n}}$$

$\chi^2_{1-\alpha}$ étant la valeur d'une variable à $n - 1$ degrés de liberté ayant une probabilité α d'être dépassée. B dépend également de n , et a été tabulé.

11.2.2.3 Représentation graphique

On représente graphiquement les résultats du contrôle de fabrication de la façon suivante : on porte sur un graphique parallèlement à l'axe des abscisses des droites représentant les limites de contrôle et de surveillance ; chaque prélèvement est alors représenté par un point, d'ordonnée égale à la moyenne ou à l'écart-type observé, d'abscisse égale au numéro de l'échantillon dans l'ordre chronologique. Ces graphiques portent le nom de *cartes de contrôles* (figure 11).

L'examen d'une carte de contrôle permet de se faire rapidement une idée de l'évolution de la fabrication que l'on pourra éventuellement approfondir en utilisant les tests du paragraphe 7.6.

11.2.3 Contrôle de fabrication et tolérances

Jusqu'à présent, nous n'avons envisagé le contrôle de fabrication que comme un moyen de s'assurer de la constance de la fabrication, et du bon réglage des machines.

Supposons maintenant que le caractère de la fabrication mis en contrôle soit soumis à des tolérances. Soit T_s la tolérance supérieure, et T_i la tolérance inférieure. Trois cas peuvent alors se présenter :

- $T_s - T_i$ est très largement supérieur à 6σ (σ étant l'écart-type estimé de la fabrication) (§ 11.2.3.1) ;
- $T_s - T_i$ est de l'ordre de 6σ (§ 11.2.3.2) ;
- $T_s - T_i$ est inférieur à 6σ (§ 11.2.3.3).

11.2.3.1 Cas d'un intervalle de tolérance supérieur à 6σ

Dans tous les cas, le réglage idéal de la moyenne sera fait sur la demi-somme des tolérances. On conçoit aisément que si l'intervalle de tolérance est largement supérieur à 6σ , on pourra admettre un léger dérèglement de la moyenne, sans, pour cela, modifier beaucoup le pourcentage de pièces hors-tolérances.

Ainsi, sur la figure 12, on a tracé la courbe de la densité de probabilité dans le cas d'un réglage idéal de la fabrication, et, en pointillés, celle correspondant à un léger dérèglement de la moyenne.

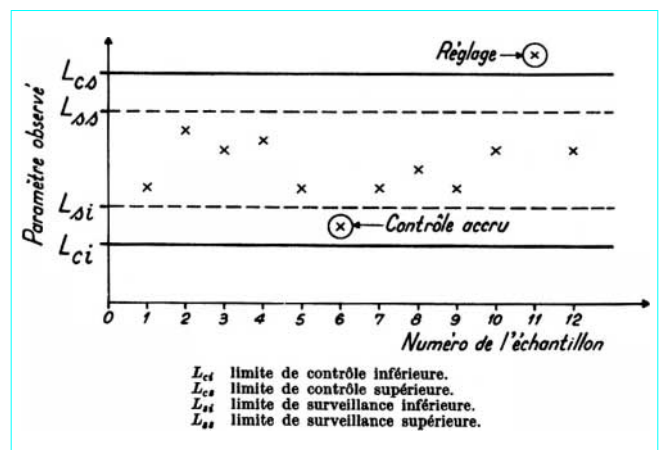


Figure 11 - Carte de contrôle

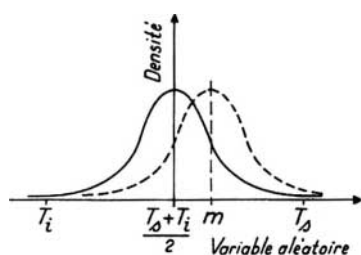
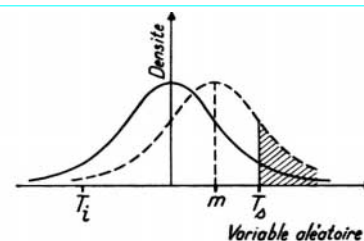


Figure 12 – Densité de probabilité : influence d'un dérèglement de la moyenne



La partie hachurée correspond au pourcentage important de déchets.

Figure 13 – Densité de probabilité : influence d'un dérèglement de la moyenne dans le cas d'un intervalle de tolérance $< 6\sigma$

Dans ces conditions, on peut prendre deux décisions :

— ou changer d'unité de fabrication ; celle qui est utilisée étant trop précise, il peut en effet être plus économique d'utiliser une machine moins précise ;

— ou alléger le contrôle en prenant des limites de contrôle basées sur les tolérances, et qui admettront dans certaines limites un dérèglement de la machine.

Étudions donc maintenant la détermination des limites de contrôle allégées : raisonnons, par exemple, sur la limite de contrôle supérieure, et plaçons-nous dans le cas d'un réglage représenté par la courbe en pointillés de la figure 12.

On ne voudra pas admettre une proportion de pièces hors-tolérances supérieure à 0,001 ; il est nécessaire, par conséquent, que la distance $(T_s - m)$ soit supérieure à $3,09\sigma$.

Dans ces conditions, au lieu de déterminer les limites de contrôle et de surveillance autour de la moyenne de la fabrication, on les calculera, pour les limites supérieures autour de $T_s - 3,09\sigma$, et pour les limites inférieures autour de $T_i + 3,09\sigma$.

Par conséquent, la limite de contrôle supérieure sera :

$$\begin{aligned} L_{cs} &= T_s - A'_{0,001} \sigma \\ &= T_s - 3,09\sigma + 3,09 \frac{\sigma}{\sqrt{n}} \\ &= T_s - 3,09\sigma \left(1 - \frac{1}{\sqrt{n}}\right) \end{aligned}$$

et la limite de contrôle inférieure :

$$\begin{aligned} L_{ci} &= T_i + A'_{0,001} \sigma \\ &= T_i + 3,09\sigma \left(1 - \frac{1}{\sqrt{n}}\right) \end{aligned}$$

Les limites de surveillance sont calculées avec les expressions :

$$\begin{aligned} L_{ss} &= T_s - A'_{0,025} \sigma \\ &= T_s - 3,09\sigma + \frac{1,96}{\sqrt{n}} \sigma \\ &= T_s - \left(3,09 - \frac{1,96}{\sqrt{n}}\right) \sigma \\ L_{si} &= T_i + A'_{0,025} \sigma \\ &= T_i + \left(3,09 - \frac{1,96}{\sqrt{n}}\right) \sigma \end{aligned}$$

Les coefficients A'_{α} ont été également tabulés. Les limites de contrôle et de surveillance ainsi déterminées sont aussi appelées *limites modifiées*. En ce qui concerne l'écart-type, il sera contrôlé comme indiqué dans le paragraphe 11.2.2.2.

11.2.3.2 Cas d'un intervalle de tolérance de l'ordre de 6σ

Dans ce cas, on peut dire que la machine est adaptée au travail qu'elle réalise. On procède à la mise sous contrôle, comme indiqué au paragraphe 11.2.2.2.

11.2.3.3 Cas d'un intervalle de tolérance inférieur à 6σ

On peut dire alors que la machine n'est pas adaptée au travail qu'elle effectue. Le moindre dérèglement de la machine peut avoir des conséquences graves sur la qualité de la production, comme le montre la figure 13.

Il est nécessaire alors soit de changer de machine, soit de renforcer le contrôle pour en accroître l'efficacité afin de détecter au plus tôt les dérèglages.

11.2.4 Contrôle de fabrication de variabilité inconnue

11.2.4.1 Cas d'une population normale d'écart-type inconnu

Nous avons vu, jusqu'à présent, qu'avant la mise sous contrôle d'une fabrication, on procédait à une étude de la variabilité qui permettait de déterminer l'écart-type. Toutefois, on peut ne pas avoir fait cette étude. Dans ce cas, on peut cependant établir des limites de contrôle et de surveillance. En effet, ayant calculé l'écart-type sur r prélèvements (r étant faible), on estime σ par la formule (§ 2.6 et 3.5.2.3) :

$$\sigma = \frac{1}{b_n} s$$

$$\begin{aligned} s &= \frac{\sum_{i=1}^r s_i}{r} \\ s_i &= \sqrt{\frac{\sum_{j=1}^n (x_j - \bar{x})^2}{n}} \end{aligned}$$

En ce qui concerne la moyenne, on remplace, comme il se doit, les valeurs u_{α} de la loi normale par les valeurs t_{α} de la loi de Student-Fisher, et on obtient des limites qui sont :

$$m \pm k \sigma$$

avec

$$k = \frac{t_{\alpha}}{\sqrt{n}}$$

Pour éviter le calcul de σ , on calcule les limites de contrôle et de surveillance avec les formules suivantes :

$$m \pm A'_{\alpha/2} s$$

avec

$$A'_{\alpha/2} = \frac{1}{b_n} \cdot \frac{t_{\alpha}}{\sqrt{n}}$$

Les valeurs de $A'_{\alpha/2}$ sont tabulées.

De même, en ce qui concerne l'écart-type, les limites de contrôle et de surveillance sont calculées par les expressions :

$$B'_\alpha s$$

$$B'_\alpha = \frac{B_\alpha}{b_n}$$

avec

Les coefficients A' et B' tendent vers les coefficients A et B quand n croît.

11.2.4.2 Cas d'une population de loi de probabilité inconnue

On utilise, faute de mieux, l'inégalité de Bienaymé-Tchebicheff :

$$P\{m - t\sigma \leq x \leq m + t\sigma\} = 1 - (1/t^2)$$

pour $t = 2$, $P = 0,75$ et pour $t = 3$, $P \approx 0,889$

On prendra pour limites de contrôle de la moyenne :

$$m \pm 3 \frac{\sigma}{\sqrt{n}}$$

En ce qui concerne l'écart-type, en se souvenant que l'écart-type d'un échantillon suit, à la limite, une loi normale d'écart-type $\sigma/\sqrt{2n}$, les limites de contrôle seront :

$$\sigma \pm 3 \frac{\sigma}{\sqrt{2n}}$$

On prendra comme limite inférieure zéro quand la formule précédente donne une valeur négative.

11.2.5 Autres contrôles de fabrication

11.2.5.1 Contrôle au moyen de l'étendue

Supposons qu'on prélève des échantillons de taille n dans une population normale, et que l'on calcule sur chaque échantillon l'écart-type s et l'étendue ω (§ 1.4.2.1). Le rapport ω/s est une variable aléatoire de moyenne d_n . Ayant estimé une valeur ω' de l'étendue comme valeur moyenne des étendues observées sur un certain nombre de prélèvements, on pourra estimer σ par la formule ω'/d_n , et donner directement des limites de contrôle et de surveillance de la moyenne et de l'écart-type en fonction de ω' : $m \pm A''_\alpha \omega'$ et $B''_\alpha \omega'$.

Les coefficients A'' et B'' sont tabulés.

On pourra également déterminer des limites de contrôle modifiées à partir de l'étendue.

Remarque : il faudra se souvenir cependant que, si l'usage de l'étendue est extrêmement simple, en particulier sur les lieux-mêmes de la fabrication, où le calcul de l'écart-type est malaisé, il doit être réservé à des prélèvements de taille faible ($n < 10$).

11.2.5.2 Contrôles divers

Il existe beaucoup d'autres méthodes de contrôle que nous ne développerons pas ici. Mentionnons simplement ceux qui substituent à la moyenne, le milieu de l'étendue (§ 1.4.1.3) ou encore la médiane (§ 1.4.1.2).

11.2.6 Contrôle de fabrication par calibres

Ce procédé s'applique quand le contrôle des éléments fabriqués se fait par classement en bons ou mauvais, ou encore si, sur chaque élément, on s'intéresse au nombre des défauts. Dans le premier cas, si on prélève, dans une population ayant une proportion p

d'éléments défectueux, un échantillon de taille n , la probabilité d'observer k défectueux est donnée par la loi binomiale.

On détermine alors des limites de contrôle et de surveillance L_c et L_s qui sont définies par les expressions suivantes :

$$P\{k \leq L_c\} = 0,998$$

$$P\{k \leq L_s\} = 0,95$$

où k représente le nombre de défectueux dans le prélèvement.

Si k est supérieur à L_c on sera fondé, avec un risque de l'ordre de 0,002, à supposer que la fabrication s'est détériorée.

Les coefficients L_c et L_s sont tabulés ; ils ont été obtenus à partir :

— soit de la loi binomiale, et, dans ce cas, ils sont applicables à des échantillons de taille $n < 50$, et de proportion de défectueux p quelconque ;

— soit de la loi de Poisson, utilisée comme approximation de la loi binomiale, et ils sont applicables à des échantillons de taille $n > 50$ et de proportion $p < 0,1$.

Compte tenu du caractère discontinu des lois, on ne peut déterminer exactement les limites à 0,95 et 0,998. Dans ces conditions, on prend les valeurs les plus voisines, correspondant à des valeurs entières de L_s et L_c .

11.2.7 Efficacité des contrôles de fabrication

Signalons qu'il est possible de calculer l'efficacité des contrôles de fabrication. Il est toutefois hors de propos, dans ce texte, de développer ces calculs. Disons simplement qu'à taille de prélèvement identique, le contrôle de fabrication par mesures est beaucoup plus efficace que le contrôle de fabrication par calibres (mais il est aussi beaucoup plus onéreux).

11.3 Contrôle de réception

11.3.1 Généralités. Contrôle à prélèvement simple ou multiple

Rappelons que le contrôle de réception, ou de fin de fabrication, consiste à s'assurer que la livraison, ou le lot de fabrications, est conforme aux spécifications demandées. Il sera par conséquent nécessaire de connaître la qualité de la fabrication. Deux procédures peuvent être appliquées :

— chaque pièce fabriquée est contrôlée ; les pièces mauvaises sont écartées (contrôle à 100 %) ;

— on prélève un échantillon ; on estime la qualité de la fabrication par le pourcentage de pièces hors-tolérances ; on admet la livraison pour un pourcentage inférieur ou égal à un certain seuil ; on refuse si le pourcentage est supérieur à ce seuil.

Toutefois, cette dernière procédure peut être affinée de la façon suivante :

— on effectue un prélèvement de taille n_1 ; on observe un pourcentage d_1 de défectueux : on accepte le lot si $d_1 < a_1$, on refuse le lot si $d_1 > r_1$;

— si $a_1 < d_1 < r_1$, on effectue un second prélèvement de taille n_2 ; on calcule le pourcentage d_2 de défectueux sur l'ensemble des deux prélèvements : on accepte le lot si $d_2 \leq a_2$, et on refuse le lot si $d_2 > a_2$.

Ce contrôle est appelé *contrôle à prélèvement double*.

L'opération peut cependant ne pas être interrompue après le deuxième prélèvement, et se poursuivre jusqu'à acceptation ou refus du lot ; on appelle cette procédure *plan de contrôle à prélèvement multiple ou séquentiel*. On aura ainsi des seuils successifs $a_1, a_2, \dots, a_j, \dots$ d'acceptation, et des seuils successifs $r_1, r_2, \dots, r_j, \dots$ de refus, avec : $a_j < a_{j+1}$ et $r_j > r_{j+1}$; en effet, plus la taille du prélèvement total croît, plus on peut élargir le seuil de réception, et resserrer le seuil de refus.

L'avantage des plans de contrôle à prélèvement multiple sur les plans de contrôle à prélèvement simple est essentiellement économique. En effet, à efficacité égale, sur un grand nombre de réceptions, on contrôlera beaucoup moins d'éléments dans le premier cas que dans le second. Par contre, la charge du service contrôle ne peut pas être prévue à l'avance.

11.3.2 Contrôle par mesures et contrôle par calibres

On distingue, dans le contrôle de réception, le contrôle par mesures, et le contrôle par calibres.

Dans le premier cas, la qualité de la livraison est en général chiffrée par l'écart de la moyenne à la tolérance T , mesuré en écart-type (on est à même, grâce aux tables de loi normale réduite, d'estimer le pourcentage de produits hors-tolérance). Ainsi, la qualité Q pourra être estimée par :

$$Q = \frac{\bar{x} - T}{s}$$

\bar{x} et s étant les valeurs de la moyenne et de l'écart-type, estimées sur l'échantillon.

La formule sera plus complexe si le caractère contrôlé est soumis à des tolérances supérieures et inférieures ; il sera nécessaire de revenir au pourcentage de défectueux.

Dans le cas où les fabrications sont bien connues, si la variance σ^2 est parfaitement déterminée et est constante d'un lot à l'autre, on pourra utiliser pour estimer Q la formule :

$$Q = \frac{\bar{x} - T}{\sigma}$$

Dans ces conditions, en utilisant l'écart-type vrai σ de la fabrication, on pourra, à efficacité égale, réduire la taille du prélèvement. Ce cas se présente assez souvent, car, si la moyenne d'une fabrication dépend du réglage de la machine, la dispersion dépend des possibilités de la machine et de son état d'usure. Une machine bien entretenue pourra effectuer des fabrications de variance constante pendant un certain laps de temps. Dans le cas du contrôle par calibres, la qualité est estimée par le pourcentage de défectueux.

Signalons qu'à taille de prélèvement égale, un contrôle par mesures est beaucoup plus efficace qu'un contrôle par calibres.

11.3.3 Bases des plans de contrôle

Nous avons, jusqu'à présent, basé le plan de contrôle sur un seuil de qualité qu'on ne voulait pas voir dépassé. Nous allons développer cette première notion (§ 11.3.3.1), puis montrer que, dans d'autres buts, on peut construire les plans de contrôle sur des bases différentes (§ 11.3.3.2, 11.3.3.3, 11.3.3.4 et 11.3.3.5).

11.3.3.1 Niveau de Qualité Acceptable (NQA)

Le Niveau de Qualité Acceptable, noté NQA, représente le seuil de qualité que l'on ne veut pas voir dépassé. Une fabrication de qualité égale au NQA a une probabilité élevée d'être acceptée. Cette probabilité est, en général, notée $1 - \alpha$. La probabilité pour qu'une fabrication de qualité égale au NQA soit refusée est égale à α . On l'appelle *risque du fournisseur*.

Signalons que, dans la littérature, on rencontre souvent la notation AQL à la place de NQA ; cette notation vient de l'expression anglaise *Acceptable Quality Level*.

11.3.3.2 Pourcentage toléré de pièces défectueuses

Dans certains cas, le client pourra accepter de recevoir, de par les risques du contrôle par prélèvement, des lots qu'il n'accepterait pas

si le contrôle était effectué à 100 %. Toutefois, il ne veut pas que la qualité de ces lots soit trop mauvaise. On peut ainsi baser un plan de contrôle sur la qualité limite des fabrications qui peuvent être reçues, avec une probabilité faible d'ailleurs. Cette qualité limite est notée LTPD (de l'anglais : *Lot Tolerance Percent Defective*). La probabilité d'accepter une fabrication de qualité égale au LTPD est appelée *risque du client* ; elle est notée β .

11.3.3.3 Point d'indifférence, point de contrôle

On pourra, dans certains cas, vouloir établir un compromis entre les risques du fournisseur et les risques du client, et, dans ce cas, baser le plan de contrôle sur le niveau de qualité des lots qui ont une probabilité de 0,5 d'être reçus, niveau de qualité appelé *point d'indifférence* ou *point de contrôle*.

11.3.3.4 Limite de qualité moyenne

Supposons que l'on reçoive en moyenne des lots de qualité p (pourcentage de défectueux) avec une probabilité P . Les lots refusés sont triés par un contrôle à 100 % et les éléments restants sont livrés au client. Dans ce cas, pour un grand nombre de lots, la proportion de pièces défectueuses sera :

$$Q = p P$$

Connaissant la courbe d'efficacité (§ 11.1.6), on pourra tracer la courbe de Q en fonction de p ; cette courbe présente un maximum pour une valeur de Q , appelée *limite de qualité moyenne* (en anglais AOQL : *Average Outgoing Quality Limit*).

On pourra, réciproquement, en partant d'une valeur de la limite de qualité moyenne, bâtir un plan de contrôle, étant entendu que tout lot refusé sera contrôlé à 100 %. Le client aura ainsi la garantie de ne pas utiliser, sur un grand nombre de lots, une proportion supérieure à l'AOQL de pièces mauvaises.

11.3.3.5 Divers

On peut encore construire des plans de contrôle sur d'autres bases, en particulier en introduisant le coût du contrôle.

11.3.4 Tables d'échantillonnages

Ayant choisi le genre de contrôle (par exemple, valeurs du NQA et du risque du fournisseur, contrôle par simple prélèvement), on peut, avec le calcul des probabilités, déterminer les conditions d'échantillonnage. Ce calcul est cependant inutile, car de nombreuses tables d'échantillonnages ont été établies, qui permettent de résoudre presque tous les problèmes.

11.3.4.1 Tables de Dodge et Romig

Les tables de Dodge et Romig proposent des plans de contrôle par calibres, basés sur le LTPD ou l'AOQL avec prélèvement simple ou double.

11.3.4.2 Tables de l'Université de Columbia et tables du Military Standard 105 D

Ces tables fournissent des plans de contrôle par calibres, avec prélèvement simple, double ou multiple, basés sur le NQA.

11.3.4.3 Tables de Bowker et Good et tables du Military Standard 414

Les tables de Bowker et Good donnent des plans de contrôle par mesures, basés sur le NQA, avec prélèvement simple ou double, dans les cas d'écart-type de la population, connu ou inconnu, et avec une ou deux tolérances (supérieure et inférieure).

Les tables du Military Standard 414 offrent les mêmes avantages, mais avec des plans de prélèvement simples. En outre, ils présentent

des contrôles basés sur l'étendue. Nous allons développer dans le paragraphe 11.3.5 quelques considérations sur ces tables (tables MIL-STD 414).

11.3.5 Utilisation des tables du Military Standard 414

Les tables du Military Standard 414 se composent de trois grandes parties :

- fabrication de variance inconnue : contrôle basé sur l'écart-type ;
- fabrication de variance inconnue : contrôle basé sur l'étendue ;
- fabrication de variance connue : contrôle basé sur l'étendue.

11.3.5.1 Critères d'acceptation

Dans le cas d'une tolérance unique supérieure, le critère d'acceptation se présente sous la forme :

$$\frac{T - \bar{x}}{s} \geq \mu$$

avec T tolérance,

\bar{x} moyenne de l'échantillon,

s estimation de l'écart-type,

$\mu > 0$.

La fabrication étant de variabilité normale, on peut également prendre comme critère :

$$p \leq p_0$$

p étant l'estimation du pourcentage de défectueux, obtenu à partir de $\frac{T - \bar{x}}{s}$.

Des tables permettent de passer d'un critère à l'autre.

Dans le cas d'un caractère soumis à une tolérance supérieure et à une tolérance inférieure, les valeurs :

$$\frac{T_s - \bar{x}}{s} \quad \text{et} \quad \frac{\bar{x} - T_i}{s}$$

permettent d'obtenir les pourcentages p_s et p_i de pièces de la fabrication de cotes supérieures à T_s et inférieures à T_i . Le critère d'acceptation sera :

$$p_s + p_i < p_0$$

11.3.5.2 Niveau de qualité acceptable (AQL)

Les tables du MIL-STD 414 donnent des plans d'échantillonnage pour les valeurs du NQA suivantes en pourcentage :

0,04	-	0,065	-	0,1	-	0,15	-	0,25
0,4	-	0,65	-	1,0	-	1,5	-	2,5
4	-	6,5	-	10,0	-	15,0	-	

Dans le cas d'un contrôle à deux tolérances, il faudra préciser le NQA pour le pourcentage total de défectueux hors-tolérances, ou les valeurs du NQA pour chaque limite supérieure et inférieure.

11.3.5.3 Taille de l'échantillon

La taille de l'échantillon dépend d'une part de la taille du lot, et d'autre part de la sévérité du contrôle, c'est-à-dire, à même NQA, de l'efficacité plus ou moins grande du plan de contrôle. Les tables du MIL-STD 414 comportent 5 degrés de contrôle. Le degré IV de contrôle est le degré utilisé en contrôle normal. Un tableau à double entrée, portant en ligne la taille du lot et en colonne les degrés de contrôle, permet d'obtenir la taille de l'échantillon sous forme d'une lettre-code ; la taille du prélèvement sera complètement déterminée grâce à d'autres tableaux et selon le NQA retenu.

11.3.5.4 Valeurs du critère d'acceptation

Outre la taille des prélèvements, les tableaux mentionnés dans le paragraphe 11.3.5.3 donnent la valeur des critères d'acceptation.

11.3.5.5 Courbes d'efficacité

En fonction de la valeur du NQA retenue et de la lettre-code de l'échantillon, l'utilisateur trouve, dans les tables du Military Standard 414, les courbes d'efficacité des plans de prélèvement. Les prélèvements à effectuer selon les processus (contrôles à variance connue ou inconnue, ou encore basés sur l'étendue) ont été déterminés de telle sorte que les courbes d'efficacité coïncident approximativement.

11.3.5.6 Contrôle normal, renforcé ou réduit

Les tables du Military Standard 414 prévoient les procédures de passage du contrôle normal à un contrôle renforcé ou réduit.

Ces notions reposent plus sur la confiance qu'on peut accorder au fournisseur que sur des considérations statistiques. En effet, les résultats du contrôle permettent d'obtenir une estimation de la qualité des lots présentés. Si un grand nombre de lots consécutifs ont un pourcentage de défectueux inférieur au NQA, on pourra alléger le contrôle pratiqué, et passer en procédure de contrôle réduit. Ceci se fera dans les conditions suivantes :

- un certain nombre de lots à déterminer ont été consécutivement reçus en contrôle normal ;
- le pourcentage de défectueux de ces lots est inférieur à un seuil donné, fonction du NQA retenu ;
- la production fonctionne en régime permanent.

On revient au contrôle normal dès qu'une des conditions suivantes est remplie :

- un lot est refusé ;
- le pourcentage de défectueux estimé d'un lot devient supérieur au NQA ;
- la production ne fonctionne plus en régime permanent.

On passe en contrôle renforcée lorsque l'estimation du pourcentage de défectueux devient supérieur au NQA sur un nombre de lots donné par une table.

Statistiques

par **Alain LAMBOLEY**

*Ancien élève de l'École Polytechnique
Ingénieur Principal de l'Armement*

Bibliographie

BASS (J.). – *Éléments de calcul des probabilités théorique et appliqué*. Masson & Cie, 2^e éd. (1967).

BRARD (R.). – *Cours de Mathématiques appliquées de l'École Polytechnique*.

CAVE (R.). – *Le contrôle statistique des fabrications*. Eyrolles (1953).

BOWKER (A. H.) et LIEBERMANN (G. J.). – *Méthodes statistiques de l'Ingénieur*. (Traduit de l'américain), Dunod (1965).

Cours et tables statistiques du Centre de formation aux applications industrielles de la Statistique. Institut de Statistique de l'Université de Paris.
