Cours de Statistique

Chapitre I

Statistique descriptive

La statistique a longtemps consisté en de simples dénombrements. Mais depuis quelques dizaines d'années, elle a pris une plus grande importance dans le monde mathématique car bien utilisée, elle peut nous fournir de précieuses informations sur nos données.

Cette science étudie un groupe d'individus ou plus généralement un groupe de spécificités (renseignements sur la population, économie d'un pays, pourcentage aux élections présidentielles, …)

Une étude statistique commence par rassembler des données, par les organiser, puis elle les analyse et enfin les interprète. Chaque partie de cette étude est importante et requiert de connaître précisément le sujet sur lequel on travaille. Par exemple, l'interprétation d'une étude statistique sur la fiabilité d'un vaccin ou sur la présence de personnes malades au sein d'une population peut être dangereuse si elle n'est pas validée par des personnes compétentes en médecine.

La première partie de l'étude qui consiste à rassembler des données et les organiser est appelée statistique descriptive. Souvent, le groupe à étudier, appelé aussi *population*, est trop volumineux : par exemple, si l'on veut faire un sondage exhaustif pour l'élection présidentielle, il faudrait interroger quelques dizaines de millions de personnes. Pour pallier ce problème, on choisit au hasard une partie de cette population (c'est ce qu'on appelle un *échantillon*), en espérant que cette partie soit la plus représentative possible de la population. A partir des données ainsi récupérées, on peut commencer notre analyse : calcul de la moyenne et de la variance (si nécessaire), tracé de l'histogramme, modélisation des données, …

1 Définitions

On étudie certaines propriétés des unités statistiques de la population. Chacune de ces propriétés s'appelle un *caractère statistique*, également appelé *variable statistique*.

On parle de caractère *qualitatif* lorsque celui-ci n'est pas mesurable (exemples: couleur des cheveux, profession, qualité, …). Ce caractère qualitatif est dit *ordinal* lorsque l'on peut faire intervenir une notion d'ordre (exemple : les grades de l'armée), sinon le caractère qualitatif est dit *nominal*.
On peut affecter un nombre à chaque attribut, cependant toute opération arithmétique doit être maniée avec précaution et être exclue s'il s'agit de caractère qualitatif nominal.

On parle au contraire de caractère *quantitatif* lorsque celui-ci est mesurable (exemples: poids, taille, degré d'alcool dans le sang, …).
Nous dirons qu'une variable statistique quantitative est *discrète* si elle ne peut prendre qu'un nombre fini ou dénombrable de valeurs numériques; en revanche, nous dirons qu'elle est *continue* si elle peut prendre un nombre infini non dénombrable de valeurs numériques.

Une fois la population parfaitement définie et le caractère étudié choisi, on collecte les observations et on constitue ainsi une *série statistique*. Cette série est exhaustive si tous les éléments de la population ont été observés: on parle alors de *recensement*. Lorsque l'étude exhaustive de la population se révèle trop onéreuse ou trop longue à obtenir, on observe seulement une partie de la population : *l'échantillon*. L'effectif de l'échantillon s'appelle la *taille* de l'échantillon.

Pour la suite du cours, nous allons travailler sur des caractères quantitatifs. Nous poserons x_1, x_2, … , x_n l'échantillon de taille n collecté dans la population. Ces observations sont le plus souvent nombreuses et se présentent sous forme désordonnée (liste de nombres, tableaux de valeurs, …). Il faut alors les dépouiller, les ordonner, les classer pour en donner une représentation claire.

Exemple 1 On a relevé les températures des mois de décembre, janvier et février à Nancy sous abri à 3 heures. Les résultats bruts sont donnés dans le tableau suivant :

5	8	6	7	8	2	-1	-2	-7	-10
2	6	5	12	12	13	10	8	5	6
4	8	9	2	-1	-2	-1	-3	-2	-4
0	2	-5	-2	-1	-4	-2	2	3	8
9	5	8	3	5	4	3	2	-1	-2
-2	-5	-8	-12	-16	-4	-2	2	0	4
-1	-2	5	6	4	5	6	2	5	4
-2	-1	-5	-8	-15	-16	-13	-12	-5	-2
0	2	6	5	4	6	3	3	2	5

Directement, ce tableau est inexploitable.

2 Statistique descriptive à une dimension

2.1 Histogramme ou diagramme en bâtons

Dans le cas de variable univariée (lorsqu'on étudie un seul caractère), cette représentation graphique peut nous fournir de précieuses informations. La construction d'un histogramme ou d'un diagramme en bâtons se fait de la manière suivante :

1- Il faut déterminer les *classes*: il faut découper l'amplitude des données (l'écart entre le maximum et le minimum des données) en intervalles. Ces derniers ne doivent pas forcément être réguliers, même si dans la plupart des cas on prendra des intervalles de même longueur.

Lorsqu'à une classe correspond une seule valeur du caractère (cela sera le cas pour une variable discrète), on parle de *diagramme en bâtons*.
Dans le cas contraire, c'est à dire lorsqu'à une classe correspond un intervalle, on parle plutôt *d'histogramme*.

Exemple 2 Reprenons l'exemple sur les températures. La plus petite température est -16, la plus grande est 13. Construisons alors un ensemble de classes pour un diagramme en bâtons : {-16, -15, … , 12, 13}.

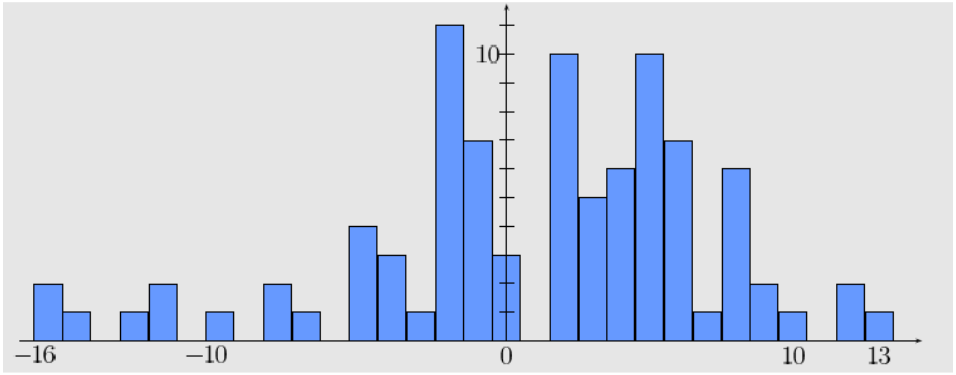
2- Pour chaque classe C_i, il faut compter le nombre d'observations n_i qui appartiennent à cette classe. La valeur n_i est appelée *effectif*, ou *fréquence absolue*, associé à la classe C_i.

Si l'échantillon est de taille n, la valeur $f_i = \frac{n_i}{n}$ est appelée *fréquence relative* associée à C_i.

Exemple 3 Reprenons l'exemple sur les températures. Après avoir compté le nombre de fois où -16 apparaît, le nombre de fois où -15 apparaît, … , on obtient le tableau suivant :

Classes	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2
Effectifs	2	1	0	1	2	0	1	0	2	1	0	4	3	1	11
Classes	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Effectifs	7	3	0	10	5	6	10	7	1	6	2	1	0	2	1

3- Il reste à tracer le graphique. En abscisse, on met les classes; en ordonnée, on met les effectifs.



Remarques 1

1. L'histogramme nous fournit directement deux informations : l'ensemble des classes C_i et l'effectif n_i dans chacune de ces classes.

2. Si f_i est la fréquence relative associée à la classe $C_i = [c_i; c_{i+1}[$, on note :
   $\forall x \in C_i, \widehat{f}(x) = \frac{f_i}{c_{i+1} - c_i}$.
   L'histogramme prend alors une dimension nouvelle. En effet, si on suppose que les observations sont des réalisations d'une certaine variable aléatoire continue X, alors $\widehat{f}$ est une approximation de la densité de X $\left(\forall x, \quad f_X(x) \simeq \widehat{f}(x)\right)$ on peut donc avoir une idée de la variable aléatoire cherchée.

2.2 Paramètres de position

Un tableau statistique ou un graphique sont parfois longs à consulter, et ne permettent pas d'avoir une idée suffisamment concise de la distribution statistique observée.

La notion de moyenne arithmétique est bien connue et permet de donner une idée globale de la série. On peut par exemple connaître le poids total d'une population connaissant sa moyenne et son effectif ou autoriser 10 personnes à monter dans un bateau dont la charge limite est de 800kg si on sait que la moyenne des poids des individus de ce groupe n'excède pas 80kg. On parlera de *paramètre de position* ou de statistique de position.

Il est important également de connaître la répartition de la population autour de cette moyenne. Dans l'exemple du bateau, il est primordial, si le groupe n'est pas de poids homogène de répartir les "lourds" et les "légers" équitablement à bâbord et tribord pour ne pas risquer le dessalage. On parlera de *paramètre de dispersion* ou de statistique de dispersion.

De plus, il faut distinguer deux cas : le premier où toutes les valeurs du caractère statistique sont données (c'est le cas le plus fréquent), le second où seul l'histogramme est donné, c'est-à-dire un ensemble d'intervalles et l'effectif dans chaque intervalle.

2.2.1 Moyenne

Définition 1 La *moyenne arithmétique* d'une série de valeurs d'une variable statistique est égale à la somme de ces valeurs divisée par leur nombre. On la note $\bar{x}$.

Si on pose x_1 ,..., x_n l'ensemble des valeurs de la variable statistique (c'est-à-dire les réalisations), on a :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

Si les données de la série statistique sont des classes d'un seul caractère c_1, …, c_p avec pour effectifs n_1, …, n_p $\left(\sum_{i=1}^{p} n_i = n\right)$, on a :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{p} n_i c_i$$

Remarque 2 C'est la moyenne qu'on utilise depuis toujours pour les notes des élèves. Nous allons voir pourquoi dans le chapitre sur l'estimation. On l'appelle aussi *moyenne empirique*.

Exemple 4 Reprenons l'exemple sur les températures. Si l'on fait la moyenne arithmétique du premier tableau, on obtient 1. Cela veut donc dire que sur trois mois, la moyenne des températures était de 1 degré.

Dans le cas où les données sont regroupées par classes non réduites à un seul caractère, on ne peut qu'estimer une moyenne à l'intérieur de chaque classe. A défaut d'autre renseignement, on choisit le "centre" de la classe ($m_i = (c_i + c_{i+1})/2$ lorsque $C_i = [c_i; c_{i+1}[$), pour toutes les valeurs qui appartiennent à cette classe, avec toute la part d'approximation que cela comporte.

Exemple 5 Reprenons l'exemple sur les températures. Construisons des intervalles d'amplitude 2 pour les classes. On obtient le tableau suivant :

Classes	[-16;-14]	[-13;-11]	[-10;-8]	[-7;-5]	[-4;-2]
Effectifs	3	3	3	5	15
Classes	[-1;1]	[2;4]	[5;7]	[8;10]	[11;13]
Effectifs	10	21	18	9	3

On fait alors le calcul suivant:

$$\frac{(-15).3+(-12).3+(-9).3+(-6).5+(-3).15+0.10+3.21+6.18+9.9+12.3}{90} \simeq 1{,}167$$

On peut mesurer dans cet exemple la perte de précision due au regroupement des données en classes et au choix du centre de classe comme moyenne de la classe. Cependant on peut se satisfaire du résultat. On annoncera dans un cas comme dans l'autre que la moyenne de ces trois mois est de 1 degré. Malgré cette perte d'information nous aurons souvent recours au regroupement en classes afin de "visualiser" plus simplement la série, surtout lorsque le nombre de valeurs du caractère est important.

2.2.2 Mode

Définition 2 Dans le cas d'une variable discrète, on appelle *mode* ou *valeur modale* toute valeur que la variable statistique prend le plus fréquemment.

Dans le cas d'une variable continue ou si les données sont groupées en classes, toute classe dont l'effectif est le plus élevé (effectif ramené à l'unité d'amplitude) est appelée *classe modale*. Attention, Il peut arriver que la classe modale ne soit pas celle où l'effectif apparaît le plus élevé sur le tableau.

Exemple 6 On peut voir dans l'exemple 3 que le mode est -2, alors qu'avec le regroupement en classes de l'exemple 5, la classe modale est [2; 4].

Remarque 3 Il peut y avoir plusieurs modes ou classes modales.

2.2.3 Médiane

Définition 3 La *médiane* d'une série statistique est un réel tel qu'il y ait autant d'observations ayant une valeur supérieure que d'observations ayant une valeur inférieure.

Remarque 4 Ce réel est défini de manière unique. Lorsque les observations sont toutes fournies, il suffit de les classer par ordre croissant et de prendre celle qui se trouve « au milieu ». Si le nombre des observations est pair, la médiane est la demi-somme des deux valeurs « du milieu ».

Lorsque les observations sont groupées en classes, la médiane ne peut être qu'estimée. Elle est nécessairement dans un intervalle que l'on appelle *classe médiane*.

Pour déterminer la médiane, on peut dresser le tableau des *fréquences cumulées* :
les classes étant rangées dans le sens croissant, la fréquence cumulée associée à la classe C_i est :

$$F_i = \sum_{j \le i} f_j \,, \quad \text{où } f_j \text{ est la fréquence relative associée à } C_j.$$

La médiane correspond à la fréquence cumulée 0,5.

Exemple 7 Pour les températures :

Classes	-16	-15	-13	-12	-10	-8	-7	-5	-4	-3	-2	-1
fréquences cumulées	0,02	0,03	0,04	0,06	0,07	0,09	0,10	0,14	0,17	0,18	0,29	0,36
Classes	0	2	3	4	5	6	7	8	9	10	12	13
Fréquences cumulées	0,39	0,49	0,54	0,60	0,70	0,77	0,78	0,84	0,86	0,87	0,89	1

La médiane est 2,5.

2.3 Paramètre de dispersion

2.3.1 Variance

Définition 4 Si on pose x_1 ,..., x_n l'ensemble des réalisations de la variable statistique, on appelle *variance* de la série la quantité suivante :

$$\mathrm{Var}(x) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 .$$

Si les données de la série statistique sont des classes d'un seul caractère c_1, …, c_p avec pour effectifs n_1, …, n_p $\left(\sum_{i=1}^{p} n_i = n\right)$, on a :

$$\mathrm{Var}(x) = \frac{1}{n}\sum_{i=1}^{p} n_i \left(c_i - \bar{x}\right)^2$$

Souvent, on préfère travailler avec l'écart-type qui est une quantité plus homogène.
L'*écart-type* est la racine carrée de la variance.

Exemple 8 Pour les températures, la variance est 38,0444. L'écart-type est égal à 6,1680.

La variance représente la dispersion des données. Plus la variance est grande, plus les données sont « dispersées ».

2.3.2 Fractiles

Une autre mesure de la dispersion consiste à utiliser des *fractiles*.
Après avoir classé les observations par ordre croissant, on partage notre population en p sous-populations de même effectif.

Définition 5 Le partage de la population en 4 sous-populations de même effectif définit les *quartiles* Q_1, Q_2 et Q_3 :
le premier quartile Q_1 est la plus petite valeur telle qu'au moins 25% de la série prend une valeur qui lui est inférieure ou égale;
le deuxième quartile Q_2 est la médiane;
le troisième quartile Q_3 est la plus petite valeur telle qu'au moins 75% de la série prend une valeur qui lui est inférieure ou égale.

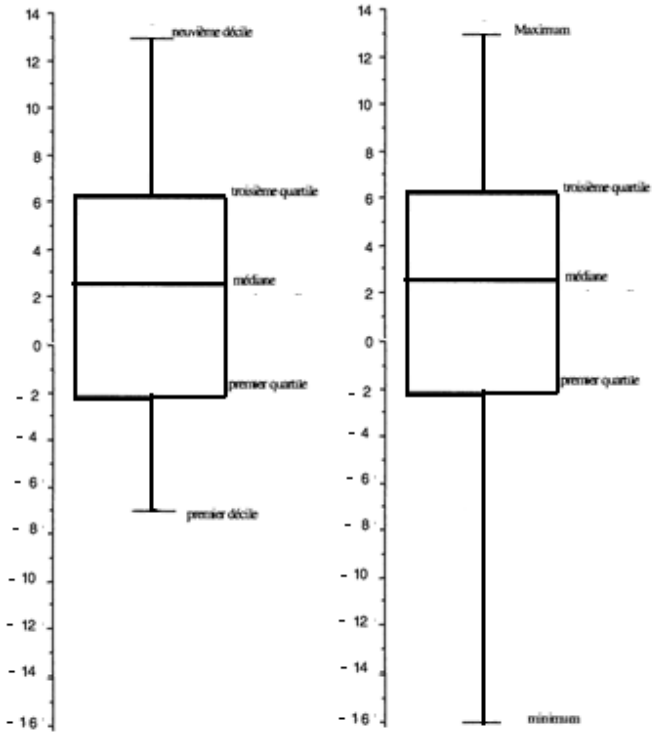
Le partage de la population en 10 sous-populations de même effectif définit les *déciles*.

Le partage en 100 sous-populations définit les *centiles*.

Pour mesurer la dispersion, on utilise l'*écart inter-quartile*, égal à $Q_3 - Q_1$.

Les fractiles d'une série statistique sont représentés à l'aide d'un diagramme, appelé *diagramme en boîte*, ou *boîte à moustaches*.

Exemple 9 Pour les températures :



Les informations que sont la moyenne et la variance (ou l'écart-type) ne suffisent pas à modéliser les données. Tracer un histogramme nous donne une idée de la distribution de la variable statistique. Que l'on reconnaisse la loi normale, la loi exponentielle ou la loi uniforme, il reste à estimer les paramètres de ces lois. Pour la loi normale, il faut estimer sa moyenne et sa variance. Pour la loi exponentielle, il faut estimer le paramètre λ, ...
Cette estimation et l'interprétation de ces résultats qui conclut l'étude statistique porte le nom de *statistique inférentielle*, qui fait l'objet du prochain chapitre.

3 Statistique descriptive à deux dimensions

3.1 Introduction

De même qu'en dimension 1, nous désirons représenter les données sous la forme de tableaux ou de graphiques, ou réduire les données à quelques paramètres. La différence avec la section précédente est que nous pouvons essayer de mettre en évidence les relations qui peuvent exister entre deux caractères.

Comme en dimension 1, nous nous intéressons à des variables quantitatives et nous aurons comme données initiales une suite double :

$$x_1, x_2, ..., x_n$$
$$y_1, y_2, ..., y_n$$

La valeur du caractère 1 pour l'individu i est x_i.
La valeur du caractère 2 pour l'individu i est y_i.

Définition 6 On appelle *série statistique double* la suite de n couples de valeurs (x_i, y_i).

Exemple 10 Poids des feuilles et poids des racines (en grammes) de 1000 individus de Cichorium intybus (cet exemple provient de l'ouvrage de Dagnélie).

feuilles	71	76	106	108	109	111	111	112	...	662	673	679	741
racines	56	51	40	174	62	59	84	94	...	174	290	290	230

3.2 Les distributions en fréquences

Comme dans le cas unidimensionnel, lorsque le nombre de données est trop important, nous condensons ces données en une distribution de fréquences. Pour cela, nous construisons un tableau à double entrée; le nombre d'individus n_{ij} ayant les occurrences x_i et y_j des caractères x et y se trouve à l'intersection de la ligne i et de la colonne j.

Dans ce paragraphe, les indices i et j qualifient les occurrences des caractères pour des variables discrètes et les classes pour des variables continues, et non pas des individus : $x_i \neq x_{i'}$ si $i \neq i'$ et $y_j \neq y_{j'}$ si $j \neq j'$.

Définition 7 On appelle *fréquence marginale* les quantités définies par :

$$n_{i\cdot} = \sum_{j=1}^{q} n_{ij} \qquad \text{pour le caractère x}$$

$$n_{\cdot j} = \sum_{i=1}^{p} n_{ij} \qquad \text{pour le caractère y}$$

Nous rappelons que le point en indice signifie que l'on a sommé sur cet indice.
Avec cette notation, nous avons donc aussi :

$$n_{\cdot\cdot} = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} = \sum_{j=1}^{q} n_{\cdot j} = \sum_{i=1}^{p} n_{i\cdot}$$

Le tableau que l'on construit a donc la structure suivante :

x \ y	y_1	y_2	...	y_j	...	y_q	Totaux
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$n_{1.}$
⋮	⋮	⋮		⋮		⋮	⋮
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	$n_{i.}$
⋮	⋮	⋮		⋮		⋮	⋮
x_p	n_{p1}	n_{p2}	...	n_{pj}	...	n_{pq}	$n_{p.}$
Totaux	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.q}$	$n_{..}$

Exemple 11 Avec les données complètes de l'exemple 10 nous obtenons :

Feuilles :Racines	40 à 79	80 à 119	120 à 159	160 à 199	200 à 239	240 à 279	280 à 319	320 à 259	Totaux
0 à 79	2								2
80 à 159	49	46	5	2					102
160 à 239	86	137	46	11					280
240 à 319	27	153	89	25	7				301
320 à 399	5	45	91	40	6				187
400 à 479		10	33	21	16	1	1		82
480 à 559		1	4	11	10	3			29
560 à 639			2	1	2	4		1	10
640 à 719				1		3	2		6
720 à 799					1				1
Totaux	169	392	270	112	42	11	3	1	1000

Remarques 5

1. Nous avons pris ici le cas des fréquences absolues mais nous pouvons construire des tableaux de fréquences relatives : $n'_{ij} = \frac{n_{ij}}{n}$
2. Nous pouvons bien entendu étudier séparément les caractères *x* et *y* et notamment faire deux statistiques descriptives à une dimension. Cela revient alors à travailler avec les fréquences marginales.

Définition 8 On appelle *fréquence conditionnelle relative* pour que $y = y_j$ sachant que $x = x_i$ la quantité :

$$f_{j/i} = \frac{n_{ij}}{n_{i.}}$$

On appelle *fréquence conditionnelle relative* pour que $x = x_i$ sachant que $y = y_j$ la quantité :

$$f_{i/j} = \frac{n_{ij}}{n_{.j}}$$

Définition 9 On appelle *profils lignes* le tableau des fréquences conditionnelles relatives $f_{j/i}$ et *profils colonnes* le tableau des fréquences conditionnelles relatives $f_{i/j}$.

Remarques 6

1. Le tableau de fréquences relatives est une représentation empirique de probabilités d'un couple de variables aléatoires et les fréquences conditionnelles relatives représentent des probabilités conditionnelles.

2. Les tableaux des profils lignes et des profils colonnes sont des représentations empiriques des lois de distributions conditionnelles.

Exemple 12 Le tableau (I.1) donne l'évolution de l'âge de la population agricole familiale dans un canton du Loiret. Le tableau (I.2) donne quant-à lui les profils lignes.

Année :Age	< à 25 ans	25 à 34 ans	35 à 44 ans	45 à 54 ans	55 à 64 ans	> à 65 ans	Total
1970	88	24	27	61	20	25	245
1979	63	17	20	39	27	25	191
1988	41	15	18	22	31	17	144
Total	192	56	65	122	78	67	580

TAB. I.1 – Tableau de contingence, exploitations agricoles dans le Loiret

Année :Age	< à 25 ans	25 à 34 ans	35 à 44 ans	45 à 54 ans	55 à 64 ans	> à 65 ans
1970	0.3592	0.0980	0.1102	0.2490	0.0816	0.1020
1979	0.3298	0.0890	0.1047	0.2042	0.1414	0.1309
1988	0.2847	0.1042	0.1250	0.1528	0.2153	0.1181

TAB. I.2 – Tableau des profils lignes

3.3 Représentations graphiques

Les séries statistiques doubles peuvent être représentées par un nuage de points (cf figure (I.1)).

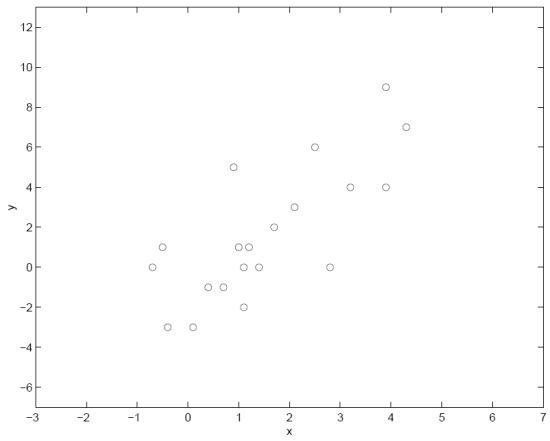


FIG. I.1 – Nuage de points

Les distributions de fréquences se représentent dans un espace à trois dimensions par un diagramme en bâtons si les variables sont discrètes, et par un stéréogramme si les variables sont continues. Un stéréogramme est un diagramme composé de parallélépipèdes rectangles de bases les rectangles correspondant aux cellules du tableau statistique et de hauteur les fréquences divisées par la surface de la base (ceci toujours pour avoir une estimation de la densité de probabilité).

Exemple 13 Avec les données de l'exemple 10, on obtient la figure (I.2)

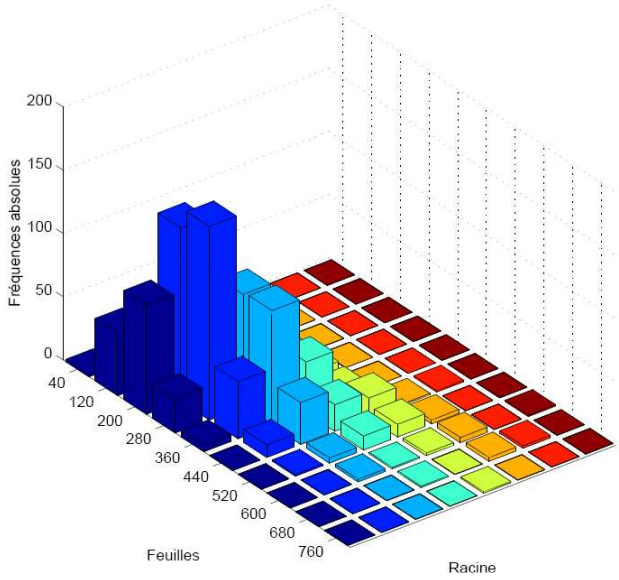


FIG. I.2 - Stéréogramme

Exemple 14 Reprenons l'exemple 12 de l'évolution de l'âge de la population agricole familiale dans un canton du Loiret. On peut représenter les profils lignes (cf figure (I.3)). Ceci nous permet de visualiser les différences de répartition des âges en fonction des années. Ici, nous avons l'ensemble des populations étudiées, les profils lignes sont donc exactement les lois de probabilités sur ces 3 populations. Dans le cas où nous n'aurions, pour chaque population, que des échantillons, il faudrait effectuer un test statistique (que nous verrons au chapitre 3) pour savoir s'il y a une différence dans les lois de distributions.

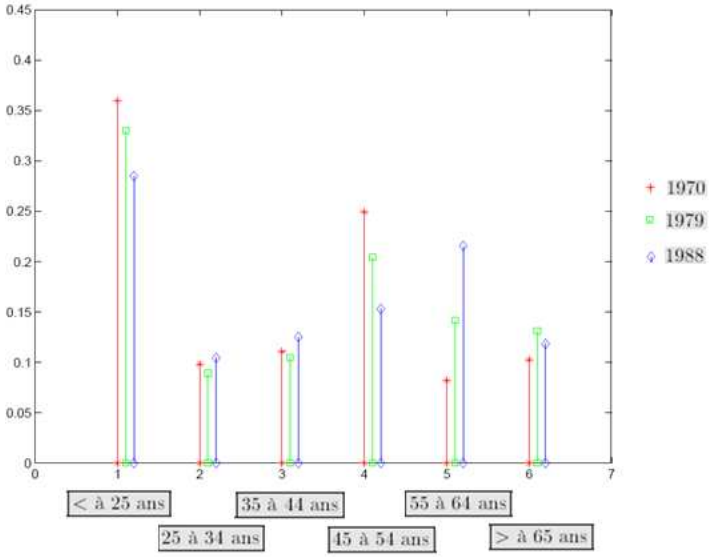


FIG. I.3 – Profils lignes

3.4 Réduction des données

Nous avons ici deux types de paramètres, tout d'abord les paramètres liés à une seule variable qui caractérisent les fréquences marginales et conditionnelles. Nous avons dans ce cas les paramètres habituels de la statistique descriptive à une dimension qui sont principalement les moyennes marginales $\bar{x}$ et $\bar{y}$ et les variances marginales σ_x^2 et σ_y^2, ainsi que les moyennes et les variances conditionnelles.
Ensuite nous avons les paramètres permettant de décrire des relations existant entre les deux séries d'observations. Ce sont ces paramètres que nous allons étudier maintenant.

Définition 10 On appelle *covariance d'un échantillon* la quantité :

- Si les données sont sous la forme d'une série statistique double :

$$\text{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

- Si les données sont sous la forme d'une distribution en fréquences :

$$\text{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\left(x_i - \bar{x}\right)\left(y_j - \bar{y}\right)$$

Remarques 7

1. On note souvent SPE = $\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$ et $\text{SCE}_x = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$.

   SPE est la Somme des Produits des Ecarts, sous entendu aux moyennes.
   SCE_x est la Somme des Carrés des Ecarts, sous entendu à la moyenne $\bar{x}$.

2. Lorsque l'on effectue les calculs à la main, on utilise les formules suivantes :

$$\text{SPE} = \sum_{i=1}^{n}\left(x_i y_i\right) - n\bar{x}\bar{y}$$

$$\text{SCE}_x = \sum_{i=1}^{n}\left(x_i\right)^2 - n\bar{x}^2$$

Exemple 15 On considère la série statistique double suivante, où *x* (respectivement *y*) représente la taille (respectivement l'envergure) de 10 adolescents nés en 1947 (mensurations relevées en 1962).

x	165,5	164	156	174	169	157,5	159	152	155	159
y	177	172	163	183,5	171,5	165	160,5	154,5	163	162

On a alors : cov(*x*, *y*) = 49,68

Remarque 8 La covariance peut-être positive ou négative. Une covariance positive (respectivement négative) indique une relation entre les données croissantes (respectivement décroissantes), i.e. que les valeurs élevées d'une série correspondent, dans l'ensemble, à des valeurs élevées (respectivement faibles) de l'autre.

Théorème 1 On a toujours la relation suivante : $\left|\text{cov}(x, y)\right| \leq \sigma_x \sigma_y$.
L'égalité n'a lieu que si les points (x_i, y_i) sont alignés.

3.5 Droite de régression

3.5.1 Introduction

Exemple 16 On désire savoir comment le taux de cholestérol sérique dépend de l'âge chez l'homme. Pour cela on a pris un échantillon d'hommes adultes d'âges bien déterminés : 25, 35, 45, 55 et 65 ans. On a obtenu les données suivantes :

Ages	25	25	25	25	25	25	25	35	35	35	35
Taux	1.8	2.3	2	2.4	2	2.5	2.6	2.6	2.9	2.3	2.4
Ages	35	35	35	45	45	45	45	45	45	45	45
Taux	2.1	2.5	2.7	2.7	3	3.1	2.3	2.5	3	3.3	2.7
Ages	55	55	55	55	55	65	65	65	65	65	65
Taux	3.1	2.9	3.4	2.4	3.4	3.7	2.8	3.3	3.5	3.3	2.6

La question qui se pose est celle de l'exploitation de ces données.

En pratique nous sommes souvent amenés à rechercher une relation entre deux variables x et y. Pour cela, dans un premier temps, nous collectons des données (x_1, y_1), (x_2, y_2),..., (x_n, y_n). Ensuite nous représentons graphiquement ces données.
Nous pouvons par exemple avoir les cas suivants :

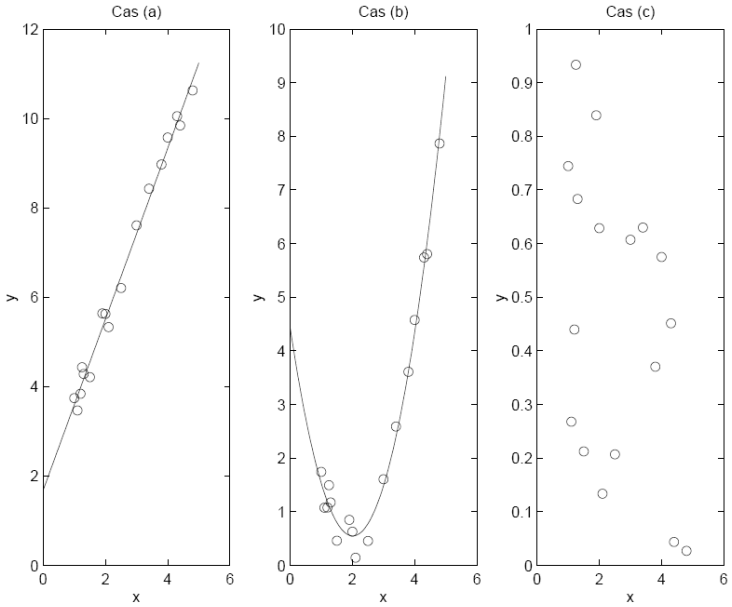


FIG. I.4 - Différentes formes de graphes

Suivant les cas de la figure I.4, nous pouvons penser aux modèles :

Cas (a) $y(x) = \beta_0 + \beta_1 x$ (modèle linéaire)

Cas (b) $y(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ (modèle parabolique)

Cas (c) pas de modèle.

L'un des buts de la régression linéaire simple (étude du cas (a)) est de prédire la "meilleure" valeur de y connaissant x (si le modèle linéaire est bien évidemment pertinent).

3.5.2 Estimation des coefficients

Une droite sera d'autant plus proche des points $M_i\,(x_i, y_i)$ que les écarts entre ces points et la droite seront faibles.
L'un des critères les plus utilisés est le critère des moindres carrés qui utilise la somme des carrés des écarts (également appelés *résidus*) : $r_i = y_i - \widehat{y_i}$, où $\widehat{y_i}$ est l'ordonnée du point de la droite d'abscisse x_i (cf figure (I.5)).

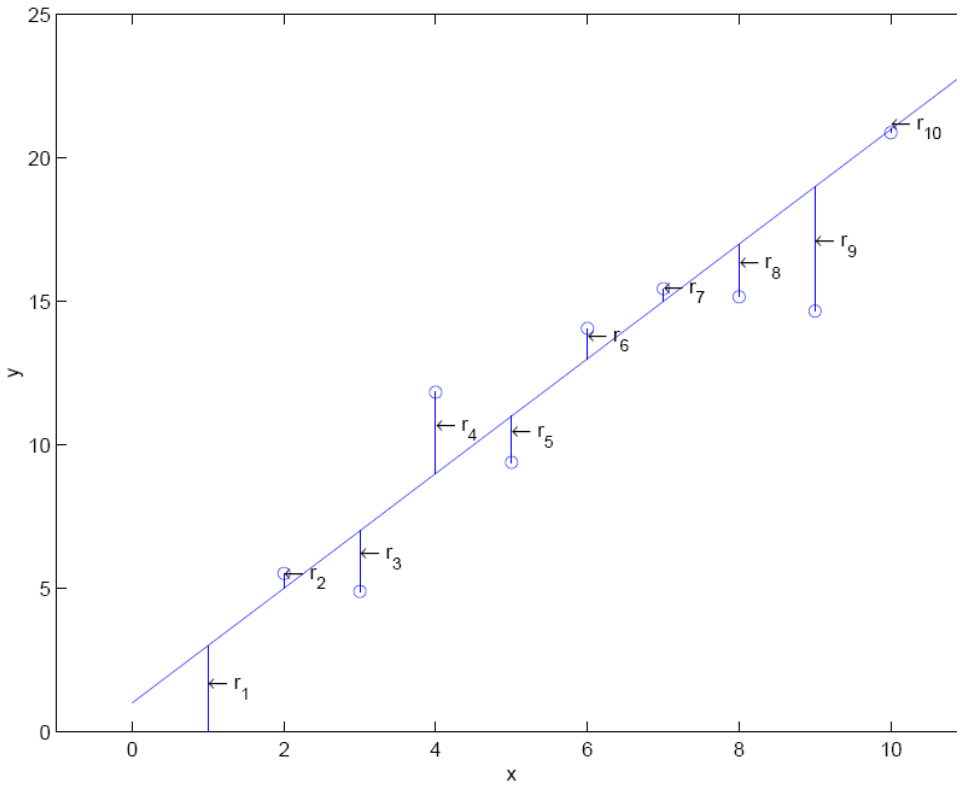


FIG. I.5 - Moindres carrés.

Les points $(x_i\,, y_i)_{i=1,\ldots,n}$ sont connus, la question est de trouver les valeurs des paramètres β_0 et β_1 qui rendent la valeur du critère la plus faible possible.

Nous sommes ramenés au problème d'optimisation pour une fonction à 2 variables suivant :

$$(P)\quad \begin{cases} \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{Min}\ f(\beta_0,\beta_1) \\ f(\beta_0,\beta_1) = \dfrac{1}{2}\displaystyle\sum_{i=1}^{n} r_i^2 = \dfrac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2 \end{cases}$$

En effet, plus $f(\beta_0,\beta_1)$ sera proche de 0, plus les carrés des résidus, donc les résidus r_i , seront proches de 0.

Théorème 2 La solution du problème (*P*) est :

$$\begin{cases} \widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x} \\ \widehat{\beta_1} = \dfrac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} = \dfrac{SPE}{SCE_x} \end{cases}$$

Remarques 9

1. On a supposé dans le calcul que $SCE_x \neq 0$, c'est-à-dire que tous les x_i ne sont pas identiques.
2. Nous noterons dans la suite $\widehat{\beta_0}$ et $\widehat{\beta_1}$ ces solutions.

Exemple 17 Reprenons l'exemple 16

	x	y	xy	x^2	y^2
1	25	1.8	45.0	625	3.24
2	25	2.3	57.5	625	5.29
3	25	2.0	50.0	625	4.00
4	25	2.4	60.0	625	5.76
5	25	2.0	50.0	625	4.00
6	25	2.5	62.5	625	6.25
7	25	2.6	65.0	625	6.76
8	35	2.6	91.0	1225	6.76
9	35	2.9	101.5	1225	8.41
⋮	⋮	⋮	⋮	⋮	⋮
33	65	2.6	169.0	4225	6.76
Totaux	1445	90.1	4103.5	69625	253.31
Moyennes	43.79	2.73			

Les estimations ponctuelles sont alors :

$$\widehat{\beta_1} = \frac{4103.5 - \dfrac{1445 \times 90.1}{33}}{69625 - \dfrac{1445^2}{33}} = \frac{158.2}{6351.5} = 0.025$$

$$\widehat{\beta_0} = 2.73 - 0.025 \times 43.79 = 1.64$$

Remarques 10

1. On a : $r_i = y_i - \widehat{y_i} = y_i - \left(\widehat{\beta_0} + \widehat{\beta_1} x_i\right)$.

   On vérifie alors que
   $$\begin{aligned} \sum_{i=1}^{n} r_i &= \sum_{i=1}^{n}\left(y_i - \left(\widehat{\beta_0} + \widehat{\beta_1} x_i\right)\right) \\ &= \sum_{i=1}^{n}(y_i) - \sum_{i=1}^{n}\left(\widehat{\beta_0}\right) - \sum_{i=1}^{n}\left(\widehat{\beta_1} x_i\right) \\ &= n\overline{y} - n\widehat{\beta_0} - n\overline{x}\widehat{\beta_1} = 0 \end{aligned}$$

2. De la même façon que nous avons cherché à exprimer *y* en fonction de *x*, on peut essayer d'exprimer *x* en fonction de *y* et nous obtenons ainsi la droite de régression d'équation :

$$x = \beta_{xy}\left(y - \bar{y}\right) + \bar{x} \qquad \text{avec} \qquad \beta_{xy} = \frac{SPE}{SCE_y}$$

Il nous reste à résoudre la question de la pertinence du modèle linéaire !

Définition 11 On appelle *coefficient de corrélation linéaire* le rapport de la covariance sur les produits des écart-types :

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

On peut aussi écrire : $$r = \frac{SPE}{\sqrt{SCE_x SCE_y}}$$

Notons $\overrightarrow{x_c}$ (respectivement $\overrightarrow{y_c}$) le vecteur des données centrées de la variable x (respectivement y). C'est-à-dire que $\overrightarrow{x_c} = \left(x_1 - \bar{x}, \cdots, x_n - \bar{x}\right)^T$ et $\overrightarrow{y_c} = \left(y_1 - \bar{y}, \cdots, y_n - \bar{y}\right)^T$.
Ces vecteurs sont dans $\mathbb{R}^n$.

Alors SPE est le produit scalaire entre ces deux vecteurs et SCE_x et SCE_y sont les normes aux carrés de ces vecteurs.

Par suite le coefficient de corrélation linéaire s'interprète comme le cosinus de l'angle de ces deux vecteurs de $\mathbb{R}^n$.

On en déduit la remarque suivante :

Remarque 11 Le coefficient de corrélation linéaire a les propriétés suivantes :

1. $r \in [-1, +1]$
2. $|r| = 1$ si et seulement si les points (x_i, y_i) sont alignés.

Ainsi, plus $|r|$ est proche de 1, plus le modèle linéaire se justifie.

On a différents cas de figures :

FIG. 1.6 - Liens entre les nuages de points et le coefficient de corrélation linéaire

3.6 Changement de variable

Nous allons voir que lorsque le modèle n'est pas a priori linéaire, on peut parfois s'y ramener par un bon changement de variable.

Considérons l'exemple suivant :

Exemple 18 Le carbone radioactif ^{14}C est produit dans l'atmosphère par l'effet des rayons cosmiques sur l'azote atmosphérique. Il est oxydé en $^{14}CO_2$ et absorbé sous cette forme par les organismes vivants qui, par suite, contiennent un certain pourcentage de carbone radioactif relativement aux carbones ^{12}C et ^{13}C qui sont stables.

On suppose que la production de carbone ^{14}C atmosphérique est demeurée constante durant les derniers millénaires.
On suppose d'autre part que, lorsqu'un organisme meurt, ses échanges avec l'atmosphère cessent et que la radioactivité due au carbone ^{14}C décroit suivant la loi exponentielle suivante :

$$A(t) = A_0 e^{-\lambda t}$$

où λ est une constante positive, *t* représente le temps en année et *A(t)* est la radioactivité exprimée en nombre de désintégrations par minute et par gramme de carbone.

On désire estimer les paramètres A_0 et λ par la méthode des moindres carrés. Pour cela, on analyse les troncs (le bois est un tissu mort) de très vieux arbres *Sequoia gigantea* et *Pinus aristaca*.

Par un prélèvement effectué sur le tronc, on peut obtenir :

- son âge *t* en années, en comptant le nombre des anneaux de croissance,
- sa radioactivité *A* en mesurant le nombre de désintégration.

On a alors le jeu de données suivant :

t	500	1000	2000	3000	4000	5000	6300
A	14.5	13.5	12.0	10.8	9.9	8.9	8.0

Soit : $y(t) = \ln A(t) = \ln A_0 - \lambda t$.
Posons alors : $\beta_0 = \ln A_0$, $\beta_1 = -\lambda$ et $y_i = \ln A_i$.

Le modèle s'écrit maintenant : $y(t) = \beta_0 + \beta_1 t$

Nous sommes donc ramenés au cas de la régression linéaire simple.

Chapitre II

Statistique inférentielle

1 Introduction

Considérons les deux situations suivantes :

Situation 1

Une société s'approvisionne en pièces brutes qui, conformément aux conditions fixées par le fournisseur, doivent avoir une masse moyenne de 780 grammes. Au moment où 500 pièces sont réceptionnées, on en prélève au hasard un échantillon de 36 pièces, dont on mesure la masse. On obtient les résultats suivants :

Masse des pièces (en grammes)	Nombre de pièces
[745; 755[	2
[755; 765[	6
[765; 775[	10
[775; 785[	11
[785; 795[	5
[795; 805[	2

À combien peut-on estimer la moyenne et l'écart-type des masses pour la population constituée des 500 pièces à l'aide des résultats obtenus sur cet échantillon ?

Situation 2

Dans un hôpital important, on prélève au hasard un échantillon de 100 personnes parmi la population des malades et on mesure la pression artérielle diostalique (P.A.D.) de chacune de ces 100 personnes. On obtient les résultats suivants :

P.A.D. (en mm de Hg)	Effectif
[4; 6[	4
[6; 8[	20
[8; 10[	41
[10; 12[	23
[12 ; 14[	12

À combien peut-on estimer la proportion de personnes dont la P.A.D. est strictement inférieure à 8 parmi la population constituée de l'ensemble des malades de l'hôpital ?

Nature du problème

Dans les deux cas, nous cherchons des informations sur une population d'effectif relativement important à partir de l'étude d'un échantillon de quelques dizaines d'unités: dans la situation 2, il s'agit d'une proportion et, dans la situation 1, d'une moyenne et d'un écart-type.
Ce type de situation se rencontre fréquemment dans le monde industriel car, le plus souvent, il n'est pas possible d'étudier la population entière: cela prendrait trop de temps, reviendrait trop cher ou serait aberrant comme, par exemple, dans le cas d'un contrôle de qualité entraînant la destruction des pièces (durée de vie d'une ampoule).
Nous allons apporter à ce problème très important deux types de réponses. Nous proposerons tout d'abord un nombre comme moyenne, proportion ou écart-type de la population : c'est l'*estimation ponctuelle*, séduisante par sa simplicité mais ne donnant pas toujours un résultat utilisable de façon satisfaisante. Aussi, dans une seconde partie, serons nous amenés à introduire la notion d'*intervalle de confiance* associé à un coefficient de confiance.

2 Statistiques

2.1 Définitions

Dans la suite, on considère le cas d'un échantillonnage aléatoire simple, c'est-à-dire que l'on extrait de la population un échantillon de taille *n* par des tirages aléatoires, équiprobables et indépendants (tirages avec remise ou tirage sans remise dans une population de grande taille).

Soit X la V.A. qui représente le caractère quantitatif que l'on souhaite étudier sur l'ensemble de la population. On note $\mathrm{IE}(X) = \mu$ et $\mathrm{Var}(X) = \sigma^2$.

Soit X_k la V.A. qui représente le résultat aléatoire du *k*-ième tirage. X_k suit la même loi que X. On note x_k sa réalisation.

Définition 12 Le *n*–uplet $(X_1, \ldots, X_n)$ de V.A. indépendantes et de même loi (celle de X) est appelé *n*-échantillon ou échantillon de taille *n* de X.

La réalisation $(x_1, \ldots, x_n)$ de l'échantillon $(X_1, \ldots, X_n)$ est l'ensemble des *valeurs observées*.

On appelle *statistique* sur un échantillon $(X_1, \ldots, X_n)$ une V.A. fonction des X_k :
$Y = f(X_1, \ldots, X_n)$.

Après réalisation, la V.A. Y (statistique) prend la valeur $f(x_1, \ldots, x_n)$.

2.2 Statistiques classiques

2.2.1 Moyenne empirique

Définition 13 On appelle *moyenne empirique* de l'échantillon $(X_1, \ldots, X_n)$ de X la statistique :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Sa réalisation $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (qui est la moyenne de l'échantillon) est appelée *moyenne observée*.

Exemple 19 Pour l'échantillon étudié dans la situation 1, la moyenne observée des masses (en grammes) des 36 pièces est (en supposant que les observations sont au centre de chaque classe):

$$\bar{x} = \frac{2\times 750 + 6\times 760 + 10\times 770 + 11\times 780 + 5\times 790 + 2\times 800}{36} \simeq 774,72$$

En l'absence d'informations supplémentaires, on décide de prendre cette valeur comme estimation de la moyenne inconnue μ des masses pour la population constituée des 500 pièces réceptionnées.

Le théorème de la limite centrale permet d'établir le résultat suivant :

Proposition 1 Si la taille *n* de l'échantillon est grande (en pratique *n* > 30), $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Remarque 12 Si $n \leq 30$, mais si $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, on a encore $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

2.2.2 Variance empirique

Définition 14 On appelle *variance empirique* de l'échantillon (X_1, …, X_n) de X la statistique :

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Sa réalisation $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ (qui est la variance de l'échantillon) est appelée *variance observée*.

Exemple 20 Pour l'échantillon étudié dans la situation 1, la variance observée des masses des 36 pièces est :

$$\sigma_{36}^2 = \frac{2\times\left(750-\bar{x}\right)^2 + 6\times\left(760-\bar{x}\right)^2 + 10\times\left(770-\bar{x}\right)^2 + 11\times\left(780-\bar{x}\right)^2 + 5\times\left(790-\bar{x}\right)^2 + 2\times\left(800-\bar{x}\right)^2}{36}$$

$$\simeq 157,06$$

Par analogie avec la moyenne, nous sommes tentés de choisir la variance σ_{36}^2 d'un échantillon prélevé au hasard comme estimation ponctuelle de la variance inconnue σ^2 d'une population. Mais en procédant ainsi, nous risquons de sous-estimer la variance de la population, et cela d'autant plus nettement que l'effectif de l'échantillon est petit. Aussi est-on conduit à corriger cette première estimation peu satisfaisante en utilisant le nombre $\frac{36}{35}\sigma_{36}^2$.

D'une manière générale, nous verrons que l'on peut choisir comme estimation ponctuelle de la variance inconnue σ^2 d'une population le nombre :

$$s_{n-1}^2 = \frac{n}{n-1}\sigma_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 .$$

2.2.3 Proportion

Soit une population comportant une modalité *M*. Soit *p* la proportion d'individus de la population possédant la modalité *M*.

On extrait de la population un échantillon de taille *n*.

Soit *R* la V.A. qui représente le nombre d'individus dans l'échantillon possédant la modalité *M*.

Définition 15 On appelle *fréquence empirique* la statistique :

$$F = \frac{R}{n}$$

Sa réalisation $f = \frac{\text{nombre d'individus possédant la modalité } M}{n}$ (qui est la proportion d'individus de l'échantillon possédant la modalité *M*) est appelée *fréquence observée*.

Exemple 21 Pour l'échantillon étudié dans la situation 2, la fréquence observée des personnes dont la P.A.D. est strictement inférieure à 8 est :

$$f = \frac{24}{100} = 0,24$$

En l'absence d'informations supplémentaires, on décide de prendre cette valeur comme estimation, pour la population constituée de l'ensemble des malades de l'hôpital, de la proportion inconnue *p* de personnes dont la P.A.D. est strictement inférieure à 8.

Remarque 13 $R \sim \mathcal{B}(n; p)$

Proposition 2 Si la taille *n* de l'échantillon est grande (en pratique $n > 30$), et si p et $1 - p$ ne sont pas trop petits $\left(p \in [0,1; 0,9]\right)$, alors $F \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$.

Remarque 14 On prend également comme conditions sur *n* et *p* : $n\,p > 5$ et $n\,(1 - p) > 5$.

3 Estimateurs

3.1 Généralités

Exemple 22 Une usine fabrique quelques millions de vis et on veut mesurer le diamètre moyen, appelons-le *d*. Si nous avions l'ensemble des valeurs, il nous suffirait de calculer la moyenne empirique pour trouver la valeur cherchée. Or nous supposerons ne posséder qu'une partie de ces valeurs. Nous allons donc devoir estimer le paramètre *d*.
La première estimation à laquelle nous pensons est la moyenne arithmétique, appelons-la m_1.
Nous pouvons donner une autre estimation de cette moyenne: la moyenne géométrique, appelons-la m_2 $\left(m_2 = \sqrt[n]{\prod_{i=1}^{n} x_i}\right)$. Si nous connaissions la valeur de *d*, il serait facile de voir laquelle des deux valeurs m_1 et m_2 estime le mieux *d*.

Seulement, nous ne connaissons pas la valeur de cette moyenne *d* (sinon nous n'aurions pas à l'estimer !).
Il nous faut donc définir au moins un estimateur et, si l'on en possède plusieurs, déterminer certaines de leurs propriétés afin de les comparer et savoir lequel est le meilleur.

Soient *X* une V.A.R. dont la loi dépend d'un paramètre inconnu θ, (X_1, … , X_n) un *n*-échantillon de *X* et (x_1, … , x_n) sa réalisation.

Il s'agit d'estimer le paramètre θ.

Définition 16 Un *estimateur* de θ est une statistique $T = f(X_1, \dots, X_n)$, et sa réalisation est notée $t = f(x_1, \dots, x_n)$.

Remarque 15 Un paramètre admet une infinité d'estimateurs. Certains sont évidemment "farfelus", d'autres semblent cohérents. Par exemple, le paramètre λ d'une loi de Poisson admet la moyenne empirique et la variance empirique comme estimateurs possibles.

3.2 Biais et convergence

Définition 17 Si *T* est un estimateur du paramètre θ, la V.A. $T-\theta$ est appelée *erreur d'estimation.*

En écrivant $T - \theta = T - \mathrm{IE}(T) + \mathrm{IE}(T) - \theta$, on fait apparaître le terme $T - \mathrm{IE}(T)$ qui traduit la fluctuation de *T* autour de son espérance, et le terme $\mathrm{IE}(T) - \theta = \mathrm{B}(T)$ appelé *biais de l'estimateur.*

Définition 18 Un estimateur *T* de θ est dit *sans biais* si :

$$\mathrm{IE}(T) = \theta \ (\text{ ou } \mathrm{B}(T) = 0)$$

sinon, on dit qu'il est *biaisé.*

Exemple 23 La moyenne empirique $\overline{X}$ est un estimateur sans biais du paramètre λ d'une loi de Poisson. La variance empirique S_n^2 est un estimateur biaisé du même paramètre.

Définition 19 Un estimateur *T* de θ est dit *asymptotiquement sans biais* si :

$$\mathrm{IE}(T) \underset{n\to+\infty}{\to} \theta,$$

Remarque 16 Un estimateur sans biais est également asymptotiquement sans biais.

Définition 20 Si *T* est un estimateur de θ asymptotiquement sans biais et si :

$$\mathrm{Var}(T) \underset{n\to+\infty}{\to} 0,$$

alors *T* est dit *convergent*.

Définition 21 Soient *T* et *T'* deux estimateurs sans biais de θ. *T* est dit *plus efficace* que *T'* si :

$$\mathrm{Var}(T) \le \mathrm{Var}(T').$$

Un estimateur sans biais de variance minimale est appelé *estimateur efficace.*

3.3 Estimation ponctuelle de paramètres usuels.

3.3.1 Estimation ponctuelle de l'espérance

Proposition 3 Soit *X* une V.A. dont on veut estimer l'espérance $\mu = \mathrm{IE}(X)$ à partir d'un *n*-échantillon (X_1,…, X_n). La moyenne empirique :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

est un estimateur efficace de μ.

3.3.2 Estimation ponctuelle de la variance

Proposition 4 Soit *X* une V.A. suivant une loi normale $\mathcal{N}(\mu,\sigma^2)$ dont on veut estimer la variance σ^2 à partir d'un *n*-échantillon (X_1,…, X_n). La statistique :

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

est un estimateur sans biais et convergent de σ^2.

3.3.3 Estimation ponctuelle d'une proportion

Proposition 5 Soit *p* la proportion d'individus d'une population possédant une modalité *M*.
On extrait de la population un échantillon de taille *n*.
Soit *R* la V.A. qui représente le nombre d'individus dans l'échantillon possédant la modalité *M*.
La fréquence empirique :

$$F = \frac{R}{n}$$

est un estimateur efficace de *p*.

3.3.4 Estimation du paramètre λ d'un processus de Poisson homogène

Comme nous avons pu le voir dans le cours de probabilités-fiabilité, que ce soit pour le processus de Poisson homogène ou bien son processus de comptage associé (N_t) la connaissance du paramètre λ est indispensable.
Nous allons donner deux méthodes pour l'estimer, utilisant deux plans d'essais :

- Plans d'essais de type I : on observe le système sur une période *t* et on note le nombre de défaillances N_t. L'estimateur $\frac{N_t}{t}$ est un estimateur sans biais et convergent de λ.

- Plans d'essais de type II : on observe le système pendant un nombre *n* de défaillances et on note le temps des *n* défaillances. Soit T_i le temps de la *i*-ème défaillance.
  l'estimateur $\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{i}$ est un estimateur sans biais de $\frac{1}{\lambda}$.

4 Estimation de la fiabilité

Soit *T* une V.A.R. qui modélise la durée de bon fonctionnement d'un élément.
Soient $t_1, t_2,..., t_n$ *n* observations de la VAR *T*.

Exemple 24 Pour estimer la fonction de répartition de *T* en *t* on divise le nombre de fois où t_i est inférieur à *t* par *n* :

$$\widehat{IP(T<t)} = \frac{1}{n}\sum_{i=1}^{n} 1_{]-\infty;t[}(t_i)$$

Pour l'exemple précédent, nous avons supposé disposer de *n* observations de la V.A. *T*.
Dans la pratique, ce ne sera pas toujours le cas. Au départ, on observera toujours *n* éléments, mais nous ne pourrons pas toujours obtenir les *n* temps de défaillances.

Voici les différents types d'essais, utilisant *n* éléments.

- Essais complets : on observe jusqu'à la défaillance du dernier élément. Les données sont donc $t_1, t_2,..., t_n$.
- Essais incomplets :

- tronqués : on arrête les essais au bout d'un temps t_0, fixé à l'avance;
- censurés : on arrête les essais à la $k^{ième}$ défaillance, fixée à l'avance;
- interrompus : on retire des éléments non défaillants en cours d'observation (on appellera ces éléments des éléments suspendus).

Pour la suite, on notera :

- *n* : nombre d'éléments mis en service à $t = 0$,
- $v(t)$: nombre d'éléments en vie à l'instant *t*,
- $d(t_i; t_{i+1})$: nombre d'éléments défaillants dans l'intervalle $]t_i; t_{i+1}]$

Remarque 17 On a de manière évidente $n = v(t) + d(0, t)$.

4.1 Essais complets

Nous allons voir trois méthodes qui permettent d'approcher le mieux possible la fiabilité, suivant la valeur de *n*.

4.1.1 Méthode des pourcentages simples

Cette méthode est utilisée lorsque la valeur de *n* est supérieur à 50. On a les estimations suivantes :

- $\widehat{F}(t) = \frac{d(0,t)}{n}$
- $\widehat{R}(t) = 1 - \frac{d(0,t)}{n} = \frac{v(t)}{n}$
- $\hat{\lambda}(t;t+\Delta t) = \frac{1}{\Delta t}\frac{d(t;t+\Delta t)}{v(t)}$

En ordonnant les temps de défaillances (ce qui n'enlève rien à la généralité), on obtient les estimations suivantes :

- $\widehat{F}(t_i) = \dfrac{i}{n}$
- $\widehat{R}(t_i) = \dfrac{n-i}{n}$
- $\widehat{\lambda}(t_i;t_{i+1}) = \dfrac{1}{(t_{i+1}-t_i)(n-i)}$

4.1.2 Méthode des rangs moyens

Cette méthode est utilisée lorsque la valeur de *n* est comprise entre 20 et 50. En ordonnant les temps de défaillances, on a les estimations suivantes :

- $\widehat{F}(t_i) = \dfrac{i}{n+1}$
- $\widehat{R}(t_i) = \dfrac{n+1-i}{n+1}$
- $\widehat{\lambda}(t_i;t_{i+1}) = \dfrac{1}{(t_{i+1}-t_i)(n+1-i)}$

4.1.3 Méthode des rangs médians

Cette méthode est utilisée lorsque la valeur de *n* est inférieure à 20. En ordonnant les temps de défaillances, on a les estimations suivantes :

- $\widehat{F}(t_i) = \dfrac{i-0,3}{n+0,4}$
- $\widehat{R}(t_i) = \dfrac{n+0,7-i}{n+0,4}$
- $\widehat{\lambda}(t_i;t_{i+1}) = \dfrac{1}{(t_{i+1}-t_i)(n+0,7-i)}$

4.2 Essais incomplets

Dans le cas d'essais tronqués, on a les observations suivantes :

$$t_1, t_2, t_3, \ldots, t_j, \underbrace{t_0, t_0, \ldots, t_0}_{n-j \text{ fois}} \text{ avec } t_i < t_0 \quad \forall i \in \{1, ..., j\}$$

Dans le cas d'essais censurés, on a les observations suivantes :

$$t_1, t_2, t_3, \ldots, t_k, \underbrace{t_0, t_0, \ldots, t_0}_{n-k \text{ fois}}$$

Dans le cas d'essais interrompus, on a les observations suivantes :

$$t_1, t_2, t_3^*, \ldots, t_j^*, \ldots, t_n$$

où les temps étoilés représentent les temps de suspension.

4.2.1 Méthode de Kaplan-Meier

Soit $d(t_j)$ le nombre d'éléments défaillants à l'instant t_j, on a l'estimation suivante :

$$\widehat{R}(t_i) = \prod_{j \text{ tq } t_j \le t_i} \left[1 - \frac{d(t_j)}{v(t_j - \varepsilon)} \right]$$

où ε est strictement positif tel que : $\forall i,\ t_{i-1} < t_i - \varepsilon$.
$v(t_j - \varepsilon)$ représente le nombre d'éléments en vie juste avant le temps de défaillance t_j.

Proposition 6 S'il n'y a pas d'élément suspendu, cette méthode se ramène à la méthode des pourcentages simples.

4.2.2 Méthode de Johnson

Soit i le rang initial de la défaillance à l'instant t_i.
On va utiliser un « rang corrigé » de la défaillance à l'instant t_i , noté i', pour prendre en compte les temps suspendus.

On note $i'\_$ le rang corrigé de la défaillance à l'instant précédant t_i.

Pour le calcul des temps corrigés, on utilise l'algorithme suivant :

1. Pour $i = 1$, $i'\_ = 0$
2. Calcul d'un incrément : $D_i = \begin{cases} 0 & \text{si l'élément est suspendu} \\ \dfrac{n+1-i'\_}{n+2-i} & \text{sinon} \end{cases}$,
3. Calcul d'un rang corrigé : $i' = i'\_ + D_i$,

On estime ensuite les caractéristiques en utilisant une méthode des essais complets avec les couples (i', t_i).

4.3 Cas particulier de la loi exponentielle

Dans le cas d'une durée de vie de loi exponentielle, on a vu que $\lambda(t) = \lambda$ et que $MTTF = \frac{1}{\lambda}$ (où MTTF = Mean Time To Failure = durée de bon fonctionnement).

De plus, il est facile de voir que $R(t) = e^{-\lambda t}$, ce qui entraîne que $\ln(R(t)) = -\lambda t$.

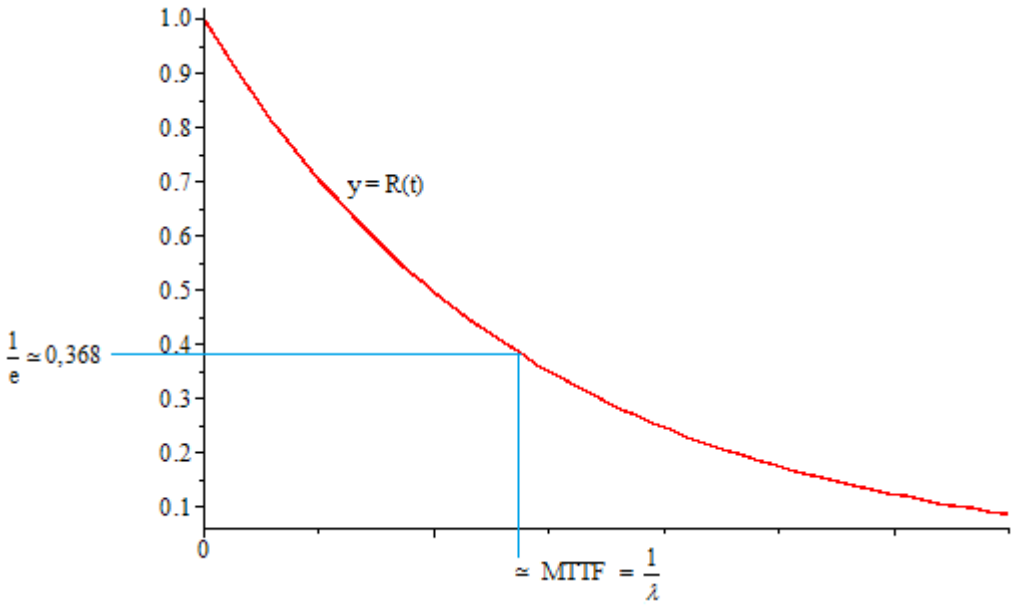
4.3.1 Méthode graphique

On trace :

- soit sur du papier semi-log la droite de régression, qui doit passer par le point (0 ; 1)
- soit la courbe d'équation y = R(t) obtenue par interpolation des points $M_i\left(t_i ; \widehat{\mathrm{R}}(t_i)\right)$

On sait que $\mathrm{R}\left(\frac{1}{\lambda}\right) \simeq 36{,}8\%$ car $\frac{1}{\mathrm{e}} \simeq 0{,}368$.

Il reste à lire sur l'axe du temps l'abscisse qui correspond à l'ordonnée 0,368. On obtient alors une estimation du MTTF.



4.3.2 Méthode numérique

On calcule la valeur de la pente de régression pour la série $\left(t_i ; \ln\left(\widehat{R}(t_i)\right)\right)$.

Cette valeur correspond à une estimation de $-\lambda$.

5 Estimation par intervalle de confiance

Dans la situation 1, en choisissant un nouvel échantillon de 36 pièces, on obtiendrait une nouvelle moyenne pour les masses de ces 36 pièces. De même, dans la situation 2, un nouvel échantillon de 100 personnes donnerait une nouvelle proportion de malades possédant la même propriété.
Ainsi, les estimations ponctuelles proposées ci-dessus de la moyenne d'une population et d'un pourcentage d'éléments de la population dépendent très directement de l'échantillon prélevé au hasard. Dans de nombreux cas, l'importance attribuée au hasard dans le choix des éléments d'un échantillon, et donc dans le résultat des estimations ponctuelles, est grande. Cela conduit à s'interroger avant d'utiliser ces estimations pour prendre des décisions dont les conséquences économiques, financières, sociales, …, peuvent être très grandes : refus éventuel d'une livraison, choix d'une stratégie commerciale, fixation d'un minimum de ressources pour l'obtention d'une aide, …
Aussi est-on amené à chercher un nouveau type d'estimation de la moyenne d'une population ou d'un pourcentage d'éléments d'une population sous forme d'intervalle, en utilisant le calcul des probabilités.
Il s'agit en fait de situer le paramètre inconnu θ par un intervalle qui a une forte probabilité (généralement 95% ou 99%) de le contenir.

5.1 Cas général

Définition 22 On appelle intervalle de confiance (noté I.C.) d'un paramètre inconnu θ au *niveau de sécurité* (ou de *confiance*) γ fixé, tout intervalle $[T_1, T_2]$ tel que :

$$\mathrm{IP}\left(T_1 < \theta < T_2\right) = \gamma$$

En général, on prend γ = 0,95 ou 0,99.

Remarque 18 On construit les intervalles de confiance à partir de la loi de probabilité d'un bon estimateur $\hat{\theta}$ de θ. Compte tenu de l'infinité des intervalles possibles, on construira des intervalles de confiance symétriques, c'est à dire tels que : $\mathrm{IP}\left(T_1 < \theta\right) = \mathrm{IP}\left(\theta < T_2\right) = \dfrac{1-\gamma}{2}$

Pour la suite, nous nous contenterons de construire des intervalles de confiance pour la moyenne et pour la proportion. La technique se généralise à un paramètre inconnu quelconque (par exemple, la variance d'une loi gaussienne).

5.2 I.C. pour une moyenne

Soit X une V.A. suivant une loi gaussienne, de variance σ^2 et d'espérance μ que l'on veut estimer.
On rappelle que pour un n-échantillon $(X_1, X_2, \ldots, X_n)$ de la population, un estimateur efficace de la moyenne est :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

5.2.1 Cas où σ^2 est connue

Si σ^2 est connue, on sait que : $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ et donc que $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0\ ;\ 1)$.

Si U suit une loi $\mathcal{N}(0\ ;1)$, alors pour tout $\alpha \in]0;1[$, il existe un réel positif, noté $u_{1-\alpha/2}$, tel que :

$$\text{IP}\left(-u_{1-\alpha/2} < U < u_{1-\alpha/2}\right) = \gamma = 1-\alpha \Leftrightarrow \text{IP}\left(U < u_{1-\alpha/2}\right) = 1-\frac{\alpha}{2}$$

On a donc : $\text{IP}\left(-u_{1-\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le u_{1-\alpha/2}\right) = \text{IP}\left(\overline{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \gamma = 1-\alpha$

Théorème 3 L'I.C. au niveau de confiance $\gamma = 1-\alpha$ de la moyenne μ d'une population gaussienne $\mathcal{N}\left(\mu, \sigma^2\right)$ où σ^2 est connue est :

$$I_\alpha = \left[\bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\ ;\ \bar{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

où $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (moyenne observée)

qui s'interprète par : « avec la probabilité $1-\alpha$, $\mu \in I_\alpha$ »
On trouve grâce à la table de la loi normale centrée réduite que :

- si $\alpha = 5\%$ (la confiance étant de 95%), alors $u_{1-\alpha/2} = 1{,}96$
- si $\alpha = 1\%$ (la confiance étant de 99%), alors $u_{1-\alpha/2} = 2{,}57$

Illustration numérique : $\sigma^2 = 4,\ n = 25,\ \bar{x} = 8$ donne :

$$I_{5\%} = \left[8 - 1{,}96\frac{2}{5}\ ;\ 8 + 1{,}96\frac{2}{5}\right] = [7{,}22;8{,}78]$$

$$I_{1\%} = \left[8 - 2{,}57\frac{2}{5}\ ;\ 8 + 2{,}57\frac{2}{5}\right] = [6{,}97;9{,}03]$$

Remarques 19

1. $\alpha_1 > \alpha_2 \Rightarrow I_{\alpha_1} \subset I_{\alpha_2}$: plus la confiance exigée est grande, plus l'amplitude de l'intervalle de confiance est grand.

2. Si l'on veut réduire l'amplitude de l'intervalle de confiance I_α (à un niveau de confiance fixé) dans un rapport *k*, il faut multiplier la taille de l'échantillon par k^2.

Illustration numérique : $\sigma^2 = 4,\ \alpha = 5\%,\ \bar{x} = 8$ donne :

$n = 25 \quad \rightarrow \quad I_{5\%} = [7{,}22;8{,}78]$

$n = 100 \quad \rightarrow \quad I_{5\%} = [7{,}61;8{,}39]$

Pratiquement, plus la confiance choisie est forte, plus l'échantillon devra être important.

5.2.2 Cas où σ^2 est inconnue

Si σ^2 est inconnue, on sait que $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ est un estimateur sans biais et convergent de σ^2.

On montre de plus que : $\frac{\overline{X}-\mu}{S_{n-1}/\sqrt{n}} \sim T_{n-1}$ où T_{n-1} suit une loi de Student à n–1 degrés de liberté.

Si T_{n-1} suit une loi de Student à n–1 degrés de liberté, alors pour tout $\alpha \in]0;1[$, il existe un réel positif, noté $t_{n-1,1-\alpha/2}$, tel que :

$$\text{IP}\left(-t_{n-1,1-\alpha/2} < T_{n-1} < t_{n-1,1-\alpha/2}\right) = \gamma = 1-\alpha \Leftrightarrow \text{IP}\left(T_{n-1} < t_{n-1,1-\alpha/2}\right) = 1-\frac{\alpha}{2}$$

D'où : $$\text{IP}\left(-t_{n-1,1-\alpha/2} \le \frac{\overline{X}-\mu}{S_{n-1}/\sqrt{n}} \le t_{n-1,1-\alpha/2}\right) = \text{IP}\left(\overline{X} - t_{n-1,1-\alpha/2}\frac{S_{n-1}}{\sqrt{n}} \le \mu \le \overline{X} + t_{n-1,1-\alpha/2}\frac{S_{n-1}}{\sqrt{n}}\right) = \gamma$$

Théorème 4 L'I.C. au niveau de confiance $\gamma = 1-\alpha$ de la moyenne μ d'une population gaussienne $\mathcal{N}\left(\mu, \sigma^2\right)$ où σ^2 est inconnue est :

$$I_\alpha = \left[\bar{x} - t_{n-1,1-\alpha/2}\frac{s_{n-1}}{\sqrt{n}}; \bar{x} + t_{n-1,1-\alpha/2}\frac{s_{n-1}}{\sqrt{n}}\right] = \left[\bar{x} - t_{n-1,1-\alpha/2}\frac{\sigma_n}{\sqrt{n-1}}; \bar{x} + t_{n-1,1-\alpha/2}\frac{\sigma_n}{\sqrt{n-1}}\right],$$

où $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (moyenne observée), $s_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$ (déviation standard)

et $\sigma_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \sqrt{\frac{n-1}{n}}\, s_{n-1}$ (écart type observé)

5.2.3 Cas où *n* est « grand » (*n* > 30)

Si *n* est « grand », il n'est pas nécessaire que *X* soit gaussienne : le théorème de la limite centrale donne $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ pour $n \to \infty$

Lorsque σ^2 est inconnue (ce qui est presque toujours les cas !), on utilise l'approximation :

$$\frac{\overline{X}-\mu}{S_n/\sqrt{n}} \xrightarrow{L} \mathcal{N}(0\ ;\ 1)$$

où S_n^2 est la variance empirique. On a alors :

Théorème 5 Lorsque *n* est « grand », l'I.C. au niveau de confiance $\gamma = 1-\alpha$ de la moyenne μ est :

$$I_\alpha = \left[\bar{x} - u_{1-\alpha/2}\frac{\sigma_n}{\sqrt{n}}; \bar{x} + u_{1-\alpha/2}\frac{\sigma_n}{\sqrt{n}}\right]$$

5.3 I.C. pour une proportion

Soit *p* la proportion d'individus d'une population possédant une modalité *M* ($p \in [0;1]$).
On extrait de la population un échantillon de taille *n*.

Soit *R* la V.A. qui représente le nombre d'individus dans l'échantillon possédant la modalité *M*.
on rappelle que $F = \frac{R}{n}$ est un estimateur efficace de *p*.
On sait de plus que $R \sim B(n; p)$

Utilisons la convergence en loi de la loi binomiale vers la loi normale quand $n \to +\infty$

$$\frac{R - np}{\sqrt{np(1-p)}} \xrightarrow{L} \mathcal{N}(0\ ;\ 1)$$

Soit $u_{1-\alpha/2}$ la valeur déterminée dans la table de la loi normale centrée réduite suivie par la V.A. U telle que : $\mathrm{IP}\left(-u_{1-\alpha/2} < U < u_{1-\alpha/2}\right) = \gamma = 1-\alpha \Leftrightarrow \mathrm{IP}\left(U < u_{1-\alpha/2}\right) = 1 - \frac{\alpha}{2}$. Alors :

$$\mathrm{IP}\left(-u_{1-\alpha/2} < \frac{R - np}{\sqrt{np(1-p)}} < u_{1-\alpha/2} \right) = 1-\alpha$$

Lorsque *n* est « grand », si f est la fréquence observée on considère que $p(1-p) \simeq f\left(1-f\right)$ et on obtient :

Théorème 6 L'I.C. au niveau de confiance $\gamma = 1-\alpha$ de la proportion *p* est :

$$\mathrm{I}_{\alpha} = \left[f - u_{1-\alpha/2}\sqrt{\frac{f\left(1-f\right)}{n}};\ f + u_{1-\alpha/2}\sqrt{\frac{f\left(1-f\right)}{n}} \right]$$

Chapitre III

Tests statistiques

1 Introduction

Depuis quelques décennies, nous assistons à une "arrivée en force" des méthodes statistiques dans le domaine réglementaire, lequel conduit à la prise de décision : on a ou on n'a pas le droit de … En particulier, l'augmentation des échanges commerciaux et des liens économiques entre les pays s'accompagne d'accords destinés à fixer des règles communes; la statistique inférentielle trouve là un immense champ d'application. Cela se traduit par des réglementations définissant dans chaque cas particulier une procédure destinée à préciser sans ambiguïté :

- comment un ou plusieurs échantillons doivent être prélevés dans la population étudiée;
- quelles mesures doivent être effectuées sur ce ou ces échantillons;
- quelle décision doit être prise à propos de l'ensemble de la population étudiée, suivant les résultats obtenus sur le ou les échantillons.

Une telle procédure s'appelle en statistique un *test de validité d'hypothèse.*

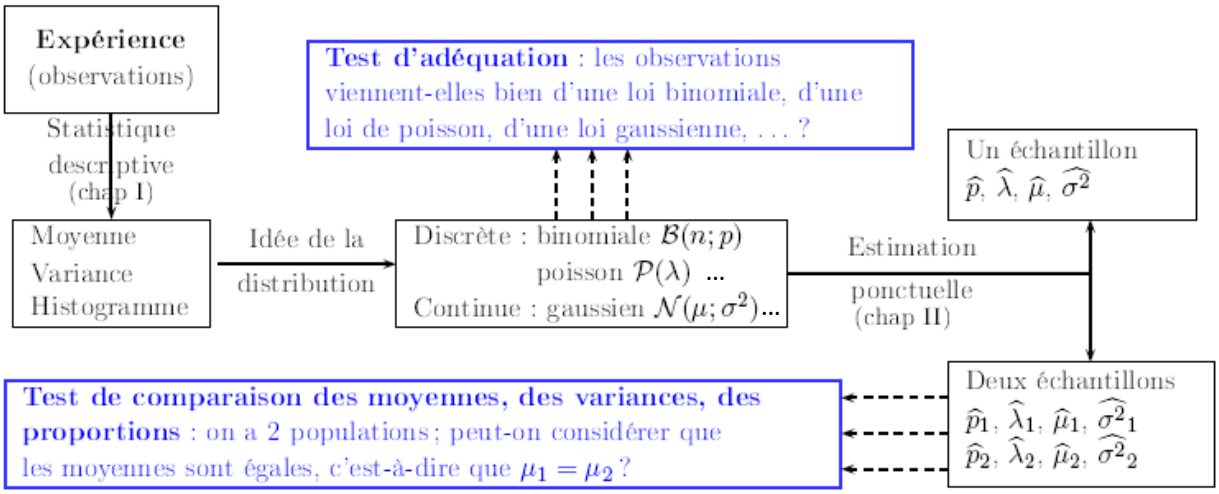


FIG. III.1 - Diagramme récapitulatif : pourquoi les tests ?

Nous allons étudier plusieurs types de tests : test d'adéquation à une loi théorique (qui permet de valider l'ajustement d'une distribution expérimentale issue d'un échantillon à une loi théorique), test d'indépendance de deux caractères, tests de comparaison ...

La construction d'un test, quel qu'il soit, passe par deux étapes :

1. On formule une hypothèse H_0 appelée *hypothèse nulle*, et il s'agit de décider si on rejette cette hypothèse par opposition à une contre-hypothèse H_1 appelée *hypothèse alternative.*

2. On applique une *règle de décision* qui va nous permettre de choisir entre la première hypothèse et la seconde. Elle repose sur une *variable de décision* (qui est une statistique) dont on connait la loi si (H_0) est vraie.

Nous pouvons alors résumer les différentes probabilités suivant la réalité et notre décision, dans le tableau ci-dessous :

		Réalité (ou état de la nature)	
		H_0 vraie	H_1 vraie
Décision retenue	considérer H_0 vraie	$1-\alpha$	β
	considérer H_1 vraie	α	$1-\beta$

TAB. III.1 - Probabilités intervenant lors d'une prise de décision en environnement incertain.

Définition 23 La quantité α représente le pourcentage de cas où H_0 est vraie et on la rejette. On l'appelle *risque de première espèce* ou encore *niveau* ou *seuil de signification*.
(En traitement du signal, α s'appelle probabilité de fausse alarme.)

La quantité β représente le pourcentage de cas où H_1 est vraie et on la rejette.

On l'appelle *risque de seconde espèce*. La quantité 1 - β s'appelle la *puissance du test*.

Lorsque l'on est en présence d'un tel test, ce que l'on cherche à faire est de minimiser les erreurs, c'est à dire les valeurs α et β. La règle de décision est très importante puisqu'elle va induire le calcul de α et de β. Pour minimiser ces deux valeurs, il faut donc jouer sur cette règle de décision. Dans la plupart des cas, nous ne pouvons pas jouer sur les deux tableaux, c'est à dire minimiser à la fois α et β. Nous allons étudier le cas où l'on donne α.

2 Tests basés sur le Khi-2

2.1 Test d'adéquation

On étudie un phénomène aléatoire représenté par une V.A. X.

On considère un n-échantillon (X_1, …, X_n) que l'on analyse selon les méthodes de statistique descriptive.
Cela permet de choisir parmi les lois de probabilités usuelles celle qui semble être la plus proche de la distribution expérimentale induite par l'échantillon.

Remarque 20 on peut être amené à ce stade à estimer certains paramètres de la loi choisie (ou donnée) pour modéliser le phénomène aléatoire (comme le paramètre λ d'une loi de Poisson).

A partir de ces données, on veut tester l'hypothèse suivante :

H_0 : "X suit une certaine loi de probabilité L"

Cela revient à tester l'hypothèse : " X_1, …, X_n suivent la loi de probabilité L"

Soit $(x_1,\ldots, x_n)$ une réalisation du n- échantillon.

On note $m_1,\ldots, m_l$ les valeurs distinctes prises par $x_1,\ldots, x_n$ et y_i le nombre de fois où m_i apparaît dans $(x_1,\ldots, x_n)$ (il s'agit de *l'effectif observé* pour la modalité m_i).

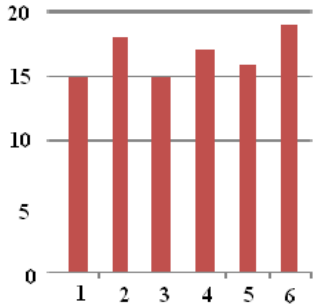
Exemple 25 On effectue 100 lancers d'un dé équilibré (cas d'équiprobabilité) à 6 faces. On obtient le tableau suivant :

Face	1	2	3	4	5	6
Effectif observé	15	18	15	17	16	19

TAB. III.2 – Résultats de l'expérience du jet du dé à 6 faces

La question est de savoir si le dé est vraiment équilibré ou non.

Le diagramme en bâtons correspondant à notre distribution expérimentale est le suivant :



Le diagramme ressemble bien à une loi uniforme discrète sur {1, 2, 3, 4, 5, 6}.

La loi d'équiprobabilité qui s'applique en probabilités si le dé est équilibré donne le tableau suivant :

Face	1	2	3	4	5	6
Effectif théorique	$100\times\frac{1}{6}$	$100\times\frac{1}{6}$	$100\times\frac{1}{6}$	$100\times\frac{1}{6}$	$100\times\frac{1}{6}$	$100\times\frac{1}{6}$

TAB. III.3 - Résultats théoriques du jet d'un dé à 6 faces

A ce niveau, on peut faire plusieurs remarques :

- Il n'est évidemment pas possible de faire 16,67 fois le chiffre 1. Il faut une valeur entière! Il y a donc dans la plupart des cas, une différence entre la fréquence théorique et la fréquence expérimentale (même si le dé est parfait).

- La différence entre la fréquence théorique (16,67) et la fréquence expérimentale (15 à 19) ne semble pas "significative". Il semble alors normal de dire que le dé est équilibré. Toutefois, et pour être plus rigoureux, nous allons "mesurer" cet écart. Cette "mesure" (qui donnera la variable de décision) nous permettra de dire avec une certaine confiance (95% ou 99%) si la différence est "significative".

Dans cet exemple, on a :

- $m_1 = 1;\ m_2 = 2;\ m_3 = 3;\ m_4 = 4;\ m_5 = 5;\ m_6 = 6,$
- $y_1 = 15;\ y_2 = 18;\ y_3 = 15;\ y_4 = 17;\ y_5 = 16;\ y_6 = 19.$

Il faut alors comparer les fréquences absolues théoriques pour m_i ($n \times \mathrm{P}(X = m_i)$) avec les fréquences expérimentales y_i.

On mesure l'écart entre la distribution théorique et la distribution expérimentale par la quantité suivante :

$$t_n = \sum_{i=1}^{l} \frac{\left[y_i - n\mathrm{IP}\left(X = m_i \right) \right]^2}{n\mathrm{IP}\left(X = m_i \right)}$$

Plus t_n est grand, plus la divergence entre la distribution théorique et la distribution expérimentale est grande. Il est donc normal de rejeter H_0 lorsque t_n dépasse une quantité c_α qui dépend du risque de première espèce.

La règle de décision convenable est donc:

On rejette H_0 ssi $t_n > c_\alpha$.

Le prochain résultat va nous fournir la possibilité de calculer explicitement c_α.

Théorème 7 Soit Y_i la V.A. donnant le nombre d'observations égales à m_i dans le n-échantillon. Si on suppose que $X_1,\ldots, X_n$ ont même loi que X (c'est à dire que l'hypothèse H_0 est vraie), la variable de décision

$$T_n = \sum_{i=1}^{l} \frac{\left[Y_i - n\mathrm{IP}\left(X = m_i \right) \right]^2}{n\mathrm{IP}\left(X = m_i \right)}$$

suit un loi du Khi-2 à (l - 1) degrés de liberté (d.d.l.).

Conséquence 1 On a la règle de décision suivante :

On rejette H_0 ssi $t_n > \chi^2_{l-1;1-\alpha}$

où $\chi^2_{l-1;1-\alpha}$ est le fractile d'ordre 1 - α d'une loi du Khi-2 à (l - 1) d.d.l., c'est-à-dire :
$\mathrm{IP}\left(\chi^2_{l-1} < \chi^2_{l-1;1-\alpha} \right) = 1-\alpha$ où χ^2_{l-1} suit une loi du Khi-2 à (l - 1) d.d.l.
(Pour les valeurs de $\chi^2_{l-1;1-\alpha}$ cf annexe 3).

Exemple 26 Pour le dé, on a : $t_n = 0{,}799$ et $\chi^2_{5;95\%} = 11{,}07$. Comme $t_n < \chi^2_{5;95\%}$, on accepte H_0. On considère donc que le dé est équilibré.

Remarque 21 Si la loi théorique dépend de paramètres inconnus (comme par exemple, si on a estimé le paramètre λ de la loi de Poisson), la règle de décision doit prendre en compte les éventuelles estimations. On a alors la règle de décision suivante :

On rejette H_0 ssi $t_n > \chi^2_{l-p-1;1-\alpha}$
où p est le nombre de paramètres estimés.

2.2 Test d'indépendance de deux variables qualitatives

2.2.1 Table de contingence

Soient, dans une même population, deux caractères qualitatifs (appelés *facteurs*) :

- le caractère L ayant l modalités,
- le caractère C ayant c modalités.

On prélève au hasard n individus et on note x_{ij} le nombre d'observations de la cellule (L_i; C_j), c'est à dire le nombre d'individus possédant la $i^{ème}$ modalité de L et la $j^{ème}$ modalité de C, avec $1 \le i \le l$ et $1 \le j \le c$.
On dispose alors d'une table de contingence dans laquelle chacun des n individus doit se retrouver dans une seule cellule.

L \ C	C_1	$\cdots$	C_j	$\cdots$	C_c	Total
L_1	x_{11}	$\cdots$	x_{1j}	$\cdots$	x_{1c}	$x_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
L_i	x_{i1}	$\cdots$	x_{ij}	$\cdots$	x_{ic}	$x_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
L_l	x_{l1}	$\cdots$	x_{lj}	$\cdots$	x_{lc}	$x_{l\bullet}$
Total	$x_{\bullet 1}$	$\cdots$	$x_{\bullet j}$	$\cdots$	$x_{\bullet c}$	n

TAB. III.4 – Tableau de contingence

On calcule les effectifs marginaux par :

$$x_{i\bullet} = \sum_{j=1}^{c} x_{ij} \quad \text{et} \quad x_{\bullet j} = \sum_{i=1}^{l} x_{ij}$$

De plus, on a :

$$\sum_{i=1}^{l}\sum_{j=1}^{c} x_{ij} = \sum_{i=1}^{l} x_{i\bullet} = \sum_{j=1}^{c} x_{\bullet j} = n$$

2.2.2 Règle de décision

Les hypothèses du test sont :

$$\begin{cases} H_0 : \text{les deux caractères sont indépendants} \\ H_1 : \text{ non } H_0 \end{cases}$$

On note p_{ij} la probabilité pour un individu d'appartenir à la cellule (L_i; C_j) pour $1 \le i \le l$ et $1 \le j \le c$. On déduit les probabilités marginales $p_{1\bullet},...,p_{i\bullet},...,p_{l\bullet}$ pour le caractère L et $p_{\bullet 1},...,p_{\bullet j},...,p_{\bullet c}$ pour le caractère C.

On sait que L et C sont indépendantes en probabilité si $p_{ij} = p_{i\bullet}\,p_{\bullet j}$ pour tout (i ; j).

On admet que $p_{i\bullet}$ et $p_{\bullet j}$ sont estimés respectivement par $\frac{x_{i\bullet}}{n}$ et $\frac{x_{\bullet j}}{n}$.

Il faut alors estimer ($l + c - 2$) paramètres car $p_{l\bullet}$ et $p_{\bullet c}$ se déterminent par les relations

$$\sum_{j=1}^{c} p_{\bullet j} = \sum_{i=1}^{l} p_{i\bullet} = 1.$$

On note $n_{ij} = \frac{x_{i\bullet}\, x_{\bullet j}}{n}$, appelé l'*effectif théorique.*

Quand H_0 est vraie, on peut estimer p_{ij} par $\frac{n_{ij}}{n}$.

On mesure l'écart entre les effectifs théoriques et expérimentaux par la quantité suivante :

$$t_n = \sum_{i=1}^{l}\sum_{j=1}^{c} \frac{\left(x_{ij} - n_{ij}\right)^2}{n_{ij}}$$

Théorème 8 Soient X_{ij} et N_{ij} les V.A. dont les réalisations sont respectivement x_{ij} et n_{ij}.
Si on suppose que l'hypothèse H_0 est vraie, la variable de décision

$$T_n = \sum_{i=1}^{l}\sum_{j=1}^{c} \frac{\left(X_{ij} - N_{ij}\right)^2}{N_{ij}}$$

suit un loi du Khi-2 à ($l - 1$) ($c - 1$) degrés de liberté (d.d.l.).

En effet, le degré de liberté de la loi limite est égal à : $lc - (l + c - 2) - 1 = (l - 1)(c - 1)$

Conséquence 2 On a la règle de décision suivante :

On rejette H_0 ssi $t_n > \chi^2_{(l-1)(c-1);1-\alpha}$

2.3 Test d'homogénéité d'une V.A.

2.3.1 Table de contingence

On considère l populations $P_1, \ldots, P_l$ sur lesquelles on étudie une V.A. X prenant c modalités $m_1, \ldots, m_c$.

Définition 24 On dit que les populations sont *homogènes* si la distribution de X est la même dans les l populations.

On prélève au hasard n individus répartis dans l'ensemble des populations et on note x_{ij} le nombre d'individus de la population P_i possédant la modalité m_j avec $1 \le i \le l$ et $1 \le j \le c$.
On dispose alors d'une table de contingence dans laquelle chacun des n individus doit se retrouver dans une seule cellule :

p \ m	m_1	$\cdots$	m_j	$\cdots$	m_c	Taille des échantillons
P_1	x_{11}	$\cdots$	x_{1j}	$\cdots$	x_{1c}	n_1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
P_i	x_{i1}	$\cdots$	x_{ij}	$\cdots$	x_{ic}	n_i
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
P_l	x_{l1}	$\cdots$	x_{lj}	$\cdots$	x_{lc}	n_l
Effectifs marginaux	$x_{\bullet 1}$	$\cdots$	$x_{\bullet j}$	$\cdots$	$x_{\bullet c}$	n

TAB. III.5 – Tableau de contingence

On calcule les effectifs marginaux par :

$$x_{\bullet j} = \sum_{i=1}^{l} x_{ij}$$

Et la taille des échantillons dans chaque population par :

$$n_i = \sum_{j=1}^{c} x_{ij}$$

De plus, on a :

$$\sum_{i=1}^{l}\sum_{j=1}^{c} x_{ij} = \sum_{i=1}^{l} n_i = \sum_{j=1}^{c} x_{\bullet j} = n$$

2.3.2 Règle de décision

Les hypothèses du test sont :

$$\begin{cases} H_0 : \text{les } l \text{ populations sont homogènes} \\ H_1 : \text{ non } H_0 \end{cases}$$

On note p_{ij} la probabilité pour un individu de population P_i de posséder la modalité m_j pour $1 \le i \le l$ et $1 \le j \le c$.

Les populations sont homogènes si les p_{ij} ne dépendent pas de la population P_i pour tout (i ; j), ce qui se traduit par : $p_{ij} = \text{IP}(X = m_j)$ pour $1 \le i \le l$ et $1 \le j \le c$

Comme on ne connait pas la loi de X, les valeurs $p_j = \text{IP}\left(X = m_j\right)$ sont inconnues, et on estimera p_j par $\dfrac{x_{\bullet j}}{n}$.

On note $n_{ij} = \dfrac{n_i x_{\bullet j}}{n} = n_i p_j$, appelé *effectif théorique.*

Quand H_0 est vraie, on peut estimer p_{ij} par $\dfrac{n_{ij}}{n_i}$.

On mesure l'écart entre les effectifs théoriques et expérimentaux par la quantité suivante :

$$t_n = \sum_{i=1}^{l}\sum_{j=1}^{c} \frac{\left[x_{ij} - n_{ij}\right]^2}{n_{ij}}$$

Théorème 9 Soient X_{ij} et N_{ij} les V.A. dont les réalisations sont respectivement x_{ij} et n_{ij}. Si on suppose que l'hypothèse H_0 est vraie, la variable de décision

$$T_n = \sum_{i=1}^{c} \sum_{j=1}^{l} \frac{\left(X_{ij} - N_{ij}\right)^2}{N_{ij}}$$

suit un loi du Khi-2 à $(l-1)(c-1)$ degrés de liberté.

Conséquence 3 On a la règle de décision suivante :

On rejette H_0 ssi $t_n > \chi^2_{(l-1)(c-1);1-\alpha}$

3 Tests de comparaison

3.1 Introduction

On considère deux V.A. X_1 et X_2 indépendantes, définies sur deux populations P_1 et P_2 respectivement, dépendant d'un paramètre inconnu θ_1 et θ_2 respectivement.
On veut tester l'hypothèse :

$$H_0 : \theta_1 = \theta_2 \quad \text{contre} \quad H_1 : \theta_1 \neq \theta_2$$

On dispose d'un n_1-échantillon de X_1 et d'un n_2-échantillon de X_2 qui fournissent respectivement T_1 un estimateur de θ_1 et T_2 un estimateur de θ_2 .

3.2 Comparaison des moyennes

On suppose que $\text{IE}(X_1) = \mu_1$ et $\text{IE}(X_2) = \mu_2$
On veut tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$.
On sait que $\overline{X_1}$ (resp. $\overline{X_2}$) est une bonne estimation de μ_1 (resp. μ_2).
On suppose que $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et que $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ ou que n_1 et $n_2 > 30$.
On suppose σ_1^2 et σ_2^2 connues.
On sait que $\overline{X_1} \sim \mathcal{N}\left(\mu_1; \frac{\sigma_1^2}{n_1}\right)$, $\overline{X_2} \sim \mathcal{N}\left(\mu_2; \frac{\sigma_2^2}{n_2}\right)$ et $\overline{X_1} - \overline{X_2} \sim \mathcal{N}\left(\mu_1 - \mu_2; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$.

Théorème 10 Si on suppose que l'hypothèse H_0 est vraie, la variable de décision :

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

suit une loi normale centrée réduite.

Conséquence 4 On a la règle de décision suivante :

$$\text{On rejette } H_0 \text{ ssi } \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > u_{1-\alpha/2}$$

Remarque 22 Si σ_1 et σ_2 sont inconnus (ce qui sera presque toujours le cas !), et que n_1 et n_2 sont suffisamment « grands »on les remplace dans le résultat précédent par les déviations standards s_{n_1-1} et s_{n_2-1}.
La règle de décision devient :

$$\text{On rejette } H_0 \text{ ssi } \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sqrt{\frac{s_{n_1-1}^2}{n_1} + \frac{s_{n_2-1}^2}{n_2}}} > u_{1-\alpha/2}$$

3.2 Comparaison des variances

On suppose que $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et que $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

On veut tester $H_0 : \sigma_1 = \sigma_2$ contre $H_1 : \sigma_1 \neq \sigma_2$.

On sait que $S_{n_1-1}^2$ (resp. $S_{n_2-1}^2$) est une bonne estimation de σ_1^2 (resp. σ_2^2).
On note $s_{n_1-1}^2$ et $s_{n_2-1}^2$ leurs réalisations, et on suppose que $s_{n_1-1}^2 \geq s_{n_2-1}^2$.

Théorème 11 La V.A. $\frac{S_{n_1-1}^2 / \sigma_1^2}{S_{n_2-1}^2 / \sigma_2^2}$ suit une loi de Fisher à ($n_1 - 1$, $n_2 - 1$) d.d.l.
Si on suppose que l'hypothèse H_0 est vraie, la variable de décision :

$$T = \frac{S_{n_1-1}^2}{S_{n_2-1}^2}$$

suit une loi de Fisher à ($n_1 - 1$, $n_2 - 1$) d.d.l.

Conséquence 5 On a la règle de décision suivante :

$$\text{On rejette } H_0 \text{ ssi } \frac{s_{n_1-1}^2}{s_{n_2-1}^2} > f_{n_1-1;n_2-1;1-\alpha}$$

où $f_{u;v;1-\alpha}$ est le fractile d'ordre $1-\alpha$ d'une loi de Fisher à (u ; v) d.d.l. c'est-à-dire
$\text{IP}\left(F_{(u,v)} < f_{u;v;1-\alpha}\right) = 1-\alpha$ où $F_{(u,v)}$ suit une loi de Fisher à (u ; v) d.d.l

3.3 Comparaison des proportions

Soit p_1 (resp. p_2) la proportion d'individus possédant la modalité M dans la population P_1 (resp. P_2). On dispose d'un n_1-échantillon de P_1 et d'un n_2-échantillon de P_2 .

Soient F_1 et F_2 les fréquences empiriques associées à P_1 et P_2 respectivement, f_1 et f_2 leurs réalisations.

On veut tester $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$.

Pour cela, on utilise l'approximation d'une loi binomiale par une loi normale pour se ramener à la comparaison des moyennes.

Théorème 12 Par approximation, on a :

$$F_1 - F_2 \sim \mathcal{N}\left(p_1 - p_2; \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \right).$$

Si on suppose que l'hypothèse H_0 est vraie, on a alors $p_1 = p_2 = p$, et la variable de décision :

$$T = \frac{F_1 - F_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

suit une loi normale centrée réduite.

Or, on ne connait pas p ! On le remplace donc par : $\boxed{f = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}}$

Conséquence 6 On a la règle de décision suivante :

$$\boxed{\text{On rejette } H_0 \text{ ssi } \frac{|f_1 - f_2|}{\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} > u_{1-\alpha/2}}$$

.

4 En fiabilité : Test d'homogénéité d'un processus de Poisson

On a vu en cours de probabilités-fiabilité que le processus de comptage N_t associé à un processus de Poisson homogène suit une loi de Poisson de paramètre λt. Lorsque le paramètre λ n'est plus constant, on parle de processus de Poisson non homogène et le nouveau paramètre $\lambda(t)$ est appelé *intensité du processus*.

On observe le processus de Poisson sur un intervalle de temps [0; t].
On note $T_1, \ldots, T_n$ les V.A. donnant les instants de défaillances, et $t_1,.., t_n$ leurs réalisations.

Les hypothèses du test sont :

$$\begin{cases} H_0 : \text{le processus de Poisson est homogène} \\ H_1 : \text{ non } H_0 \end{cases}$$

Plusieurs tests sont envisageables. Nous allons en voir deux.

4.1 Test non paramétrique

On divise l'intervalle [0; t] en d intervalles de même longueur notés $I_1,\ldots, I_d$.
Soit N_k la V.A. qui donne le nombre de défaillances dans l'intervalle I_k, c'est à dire :

$$N_k = \sum_{i=1}^{n} 1_{\{T_i \in I_k\}} \text{ où } n=N(t)$$

On note n_k sa réalisation.

Théorème 13 Si on suppose que l'hypothèse H_0 est vraie, la variable de décision :

$$Z_n = \frac{d}{n}\sum_{k=1}^{d}\left(N_k - \frac{n}{d}\right)^2$$

suit une loi du Khi-2 à (d - 1) degrés de liberté.

Conséquence 7 On a la règle de décision suivante :

$$\text{On rejette } H_0 \text{ ssi } \frac{d}{n}\sum_{k=1}^{d}\left(n_k - \frac{n}{d}\right)^2 > \chi^2_{d-1,1-\alpha}$$

où $\chi^2_{d-1,1-\alpha}$ est le fractile d'ordre 1 - α de la loi du Khi-2 à (d - 1) degrés de liberté.

4.2 Test de Laplace

Si le processus de Poisson est homogène, les temps observés seront répartis de façon relativement homogène sur [0; t], alors que si λ n'est pas constant, les points seront plus concentrés vers 0 ou vers t.

Théorème 14 Si on suppose que l'hypothèse H_0 est vraie, la variable de décision :

$$L_n = \sqrt{\frac{12}{n}}\,\frac{\sum_{i=1}^{n}\left(T_i - \frac{t}{2}\right)}{t}$$

suit une loi normale centrée réduite.

Conséquence 8 On a la règle de décision suivante :

$$\text{On rejette } H_0 \text{ ssi } \sqrt{\frac{12}{n}}\,\frac{\left|\sum_{i=1}^{n}\left(t_i - \frac{t}{2}\right)\right|}{t} > u_{1-\alpha}$$

où $u_{1-\alpha}$ est le fractile d'ordre (1 - α) de la loi gaussienne centrée réduite, c'est-à-dire $\text{IP}\left(U < u_{1-\alpha}\right) = 1-\alpha$, où U suit une loi normale centrée réduite.

ANNEXES

Annexe 1 : Table de la loi normale centrée, réduite.

Annexe 2 : Table de la loi de Student

Annexe 3 : Distribution du Khi-2

Annexe 4 : Table de la loi de Fischer pour $\alpha = 0,05$

Annexe 5 : TD n°1 : Statistiques descriptives à une dimension

Annexe 6 : TD n°2 : Statistiques descriptives à deux dimensions

Annexe 7 : TD n°3 : Estimation ponctuelle

Annexe 8 : TD n°4 : Estimation de la fiabilité

Annexe 9 : TD n°5 : Intervalles de confiance

Annexe 10 : TD n°6 : Tests statistiques

Annexe 11 : TD n°7 : Révisions

Annexe 1 : Table de la loi normale centrée réduite

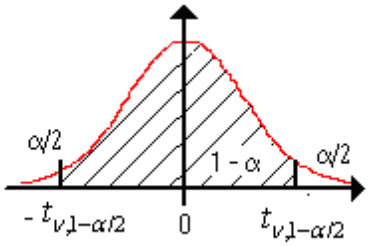
	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	50000	50399	50798	51197	51595	51994	52392	52790	53188	53586
0,1	53983	54380	54776	55172	55567	55962	56356	56749	57142	57535
0,2	57926	58317	58706	59095	59483	59871	60257	60642	61026	61409
0,3	61791	62172	62552	62930	63307	63683	64058	64431	64803	65173
0,4	65542	65910	66276	66640	67003	67364	67724	68082	68439	68793
0,5	69146	69497	69847	70194	70540	70884	71226	71566	71904	72240
0,6	72575	72907	73237	73565	73891	74215	74537	74857	75175	75490
0,7	75804	76115	76424	76730	77035	77337	77637	77935	78230	78524
0,8	78814	79103	79389	79673	79955	80234	80511	80785	81057	81327
0,9	81594	81859	82121	82381	82639	82894	83147	83398	83646	83891
1	84134	84375	84614	84849	85083	85314	85543	85769	85993	86214
1,1	86433	86650	86864	87076	87286	87493	87698	87900	88100	88298
1,2	88493	88686	88877	89065	89251	89435	89617	89796	89973	90147
1,3	90320	90490	90658	90824	90988	91149	91309	91466	91621	91774
1,4	91924	92073	92220	92364	92507	92647	92785	92922	93056	93189
1,5	93319	93448	93574	93699	93822	93943	94062	94179	94295	94408
1,6	94520	94630	94738	94845	94950	95053	95154	95254	95352	95449
1,7	95543	95637	95728	95818	95907	95994	96080	96164	96246	96327
1,8	96407	96485	96562	96638	96712	96784	96856	96926	96995	97062
1,9	97128	97193	97257	97320	97381	97441	97500	97558	97615	97670
2	97725	97778	97831	97882	97932	97982	98030	98077	98124	98169
2,1	98214	98257	98300	98341	98382	98422	98461	98500	98537	98574
2,2	98610	98645	98679	98713	98745	98778	98809	98840	98870	98899
2,3	98928	98956	98983	99010	99036	99061	99086	99111	99134	99158
2,4	99180	99202	99224	99245	99266	99286	99305	99324	99343	99361
2,5	99379	99396	99413	99430	99446	99461	99477	99492	99506	99520
2,6	99534	99547	99560	99573	99585	99598	99609	99621	99632	99643
2,7	99653	99664	99674	99683	99693	99702	99711	99720	99728	99736
2,8	99744	99752	99760	99767	99774	99781	99788	99795	99801	99807
2,9	99813	99819	99825	99831	99836	99841	99846	99851	99856	99861
3	99865	99869	99874	99878	99882	99886	99889	99893	99896	99900
3,1	99903	99906	99910	99913	99916	99918	99921	99924	99926	99929
3,2	99931	99934	99936	99938	99940	99942	99944	99946	99948	99950
3,3	99952	99953	99955	99957	99958	99960	99961	99962	99964	99965
3,4	99966	99968	99969	99970	99971	99972	99973	99974	99975	99976
3,5	99977	99978	99978	99979	99980	99981	99981	99982	99983	99983
3,6	99984	99985	99985	99986	99986	99987	99987	99988	99988	99989
3,7	99989	99990	99990	99990	99991	99992	99992	99992	99992	99992
3,8	99993	99993	99993	99994	99994	99994	99994	99995	99995	99995
3,9	99995	99995	99996	99996	99996	99996	99996	99996	99997	99997

Annexe 2 : Table de la loi de Student

La table donne les fractiles de la loi de Student à ν degrés de liberté, c'est-à-dire la valeur $t_{\nu,1-\alpha/2}$ ayant la probabilité α d'être dépassée en valeur absolue par une V.A.R. T_ν suivant une loi de Student à ν d.d.l. :

$$\mathrm{IP}\left(T_\nu < -t_{\nu,1-\alpha/2}\right) = \mathrm{IP}\left(T_\nu > t_{\nu,1-\alpha/2}\right) = \alpha/2$$

$$\mathrm{IP}\left(-t_{\nu,1-\alpha/2} < T_\nu < t_{\nu,1-\alpha/2}\right) = 1-\alpha$$



α	ν (degré de liberté)

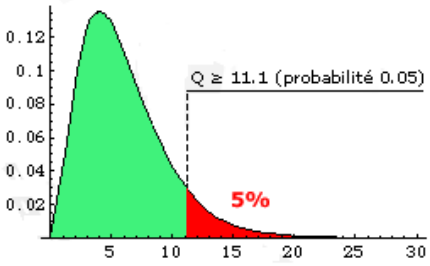
	0.9	0.5	0.1	0.05	0.02	0.01	0.005	0.001
1	0.1584	1	6.3137	12.706	31.821	63.656	127.32	636.58
2	0.1421	0.8165	2.92	4.3027	6.9645	9.925	14.089	31.6
3	0.1366	0.7649	2.3534	3.1824	4.5407	5.8408	7.4532	12.924
4	0.1338	0.7407	2.1318	2.7765	3.7469	4.6041	5.5975	8.6101
5	0.1322	0.7267	2.015	2.5706	3.3649	4.0321	4.7733	6.8685
6	0.1311	0.7176	1.9432	2.4469	3.1427	3.7074	4.3168	5.9587
7	0.1303	0.7111	1.8946	2.3646	2.9979	3.4995	4.0294	5.4081
8	0.1297	0.7064	1.8595	2.306	2.8965	3.3554	3.8325	5.0414
9	0.1293	0.7027	1.8331	2.2622	2.8214	3.2498	3.6896	4.7809
10	0.1289	0.6998	1.8125	2.2281	2.7638	3.1693	3.5814	4.5868
11	0.1286	0.6974	1.7959	2.201	2.7181	3.1058	3.4966	4.4369
12	0.1283	0.6955	1.7823	2.1788	2.681	3.0545	3.4284	4.3178
13	0.1281	0.6938	1.7709	2.1604	2.6503	3.0123	3.3725	4.2209
14	0.128	0.6924	1.7613	2.1448	2.6245	2.9768	3.3257	4.1403
15	0.1278	0.6912	1.7531	2.1315	2.6025	2.9467	3.286	4.0728
16	0.1277	0.6901	1.7459	2.1199	2.5835	2.9208	3.252	4.0149
17	0.1276	0.6892	1.7396	2.1098	2.5669	2.8982	3.2224	3.9651
18	0.1274	0.6884	1.7341	2.1009	2.5524	2.8784	3.1966	3.9217
19	0.1274	0.6876	1.7291	2.093	2.5395	2.8609	3.1737	3.8833
20	0.1273	0.687	1.7247	2.086	2.528	2.8453	3.1534	3.8496
21	0.1272	0.6864	1.7207	2.0796	2.5176	2.8314	3.1352	3.8193
22	0.1271	0.6858	1.7171	2.0739	2.5083	2.8188	3.1188	3.7922
23	0.1271	0.6853	1.7139	2.0687	2.4999	2.8073	3.104	3.7676
24	0.127	0.6848	1.7109	2.0639	2.4922	2.797	3.0905	3.7454
25	0.1269	0.6844	1.7081	2.0595	2.4851	2.7874	3.0782	3.7251
26	0.1269	0.684	1.7056	2.0555	2.4786	2.7787	3.0669	3.7067
27	0.1268	0.6837	1.7033	2.0518	2.4727	2.7707	3.0565	3.6895
28	0.1268	0.6834	1.7011	2.0484	2.4671	2.7633	3.047	3.6739
29	0.1268	0.683	1.6991	2.0452	2.462	2.7564	3.038	3.6595
30	0.1267	0.6828	1.6973	2.0423	2.4573	2.75	3.0298	3.646
40	0.1265	0.6807	1.6839	2.0211	2.4233	2.7045	2.9712	3.551

Annexe 3 : Distribution du KHI-2

La table donne en fonction du degré de liberté *n* (qu'on lit sur la première colonne) et du risque d'erreur (ou seuil de tolérance) α (qu'on lit sur la première ligne), la valeur $\chi^2_{n,1-\alpha}$ qui possède la probabilité α d'être dépassée par une V.A.R. suivant une loi du Khi-2 à *n* degrés de liberté : $F_{\chi^2_n}\left(\chi^2_{n;1-\alpha}\right) = 1-\alpha$

Exemple : Pour ddl = 5 et α = 0,05 la table indique $\chi^2_{5;95\%}$ = 11,07 i.e. IP (χ^2_5 > 11,07) = 0,05.

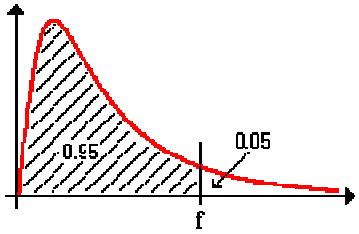
n/α	0.10	0.05	0.01	0.001
1	2.706	3.841	6.635	10.828
2	4.605	5.991	9.210	13.816
3	6.251	7.815	11.345	16.266
4	7.779	9.488	13.277	18.467
5	9.236	11.070	15.086	20.515
6	10.645	12.592	16.812	22.458
7	12.017	14.067	18.475	24.322
8	13.362	15.507	20.090	26.124
9	14.684	16.919	21.666	27.877
10	15.987	18.307	23.209	29.588
11	17.275	19.675	24.725	31.264
12	18.549	21.026	26.217	32.909
13	19.812	22.362	27.688	34.528
14	21.064	23.685	29.141	36.123
15	22.307	24.996	30.578	37.697
16	23.542	26.296	32.000	39.252
17	24.769	27.587	33.409	40.790
18	25.989	28.869	34.805	42.312
19	27.204	30.144	36.191	43.820
20	28.412	31.410	37.566	45.315
21	29.615	32.671	38.932	46.797
22	30.813	33.924	40.289	46.268
23	32.007	35.172	41.638	49.728
24	33.196	36.415	42.980	51.179
25	34.382	37.652	44.314	52.620
30	40.26	43.77	50.89	59.70
35	46.06	49.80	57.34	66.62
40	51.81	55.76	63.69	73.40
45	57.51	61.66	69.96	80.08
50	63.17	67.50	76.15	86.66
60	74.40	79.08	88.38	99.61
70	85.53	90.53	100.43	112.32
80	96.58	101.88	112.33	124.84
90	107.57	113.15	124.12	137.21
100	118.50	124.34	135.81	149.45



Pour 5 degrés de liberté, χ^2 sera supérieur à 11,07 dans seulement 5% des cas.

Annexe 4 : Table de la loi de Fischer pour $\alpha = 0,05$

La table donne la valeur $f_{\nu_1;\nu_2;0,95}$ ayant la probabilité 0,05 d'être dépassée par une V.A. F_{ν_1,ν_2} suivant une loi de Fisher-Snedecor à (ν_1,ν_2) d.d.l. : $\text{IP}\left(F_{\nu_1,\nu_2} > f_{\nu_1;\nu_2;0,95}\right) = 0,05$



	ν1 : degrés de liberté du numérateur														
ν2 : degrés de liberté du dénominateur															
	1	2	3	4	5	6	7	8	9	10	15	20	30	50	100
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95	248.02	250.10	251.77	253.04
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.45	19.46	19.48	19.49
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.62	8.58	8.55
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.75	5.70	5.66
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.50	4.44	4.41
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.81	3.75	3.71
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.38	3.32	3.27
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.08	3.02	2.97
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.86	2.80	2.76
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.70	2.64	2.59
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.57	2.51	2.46
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.47	2.40	2.35
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.38	2.31	2.26
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.31	2.24	2.19
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.25	2.18	2.12
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.19	2.12	2.07
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.15	2.08	2.02
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.11	2.04	1.98
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.07	2.00	1.94
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.04	1.97	1.91
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	1.98	1.91	1.85
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	2.03	1.94	1.86	1.80
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07	1.99	1.90	1.82	1.76
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04	1.96	1.87	1.79	1.73
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.84	1.76	1.70
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.78	1.69	1.60	1.52
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.65	1.56	1.48
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.81	1.72	1.62	1.53	1.45
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.79	1.70	1.60	1.51	1.43
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.78	1.69	1.59	1.49	1.41
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77	1.68	1.57	1.48	1.39
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.72	1.62	1.52	1.41	1.32
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.70	1.61	1.50	1.39	1.30
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.69	1.59	1.48	1.38	1.28
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.68	1.58	1.47	1.36	1.26
2000	3.85	3.00	2.61	2.38	2.22	2.10	2.01	1.94	1.88	1.84	1.67	1.58	1.46	1.36	1.25

Annexe 5 : TD n°1

Exercice 1

Dans un atelier, la production de certaines pièces pendant les 20 jours de travail d'un mois donné a été la suivante : 520; 450; 460; 485; 510; 450; 405; 460; 499; 380; 398; 455; 385; 409; 390; 424; 459; 407; 410; 428.

1) Regrouper ces données en classes d'amplitude 25 en commençant par la classe [375; 400[.
On complétera le tableau suivant après l'avoir reproduit.

Nombre de pièces produites	[375; 400[	…	
Effectif	4	…	

2) Construire l'histogramme correspondant.
3) Calculer les paramètres de position de la série complète (moyenne, mode et médiane).
4) Calculer les paramètres de dispersion de la série complète (variance, fractiles).
5) Représenter la boîte à moustaches de cette série statistique.

Exercice 2

On a mesuré les longueurs en millimètres d'un échantillon de 100 tiges d'acier à la sortie d'une machine automatique. On a trouvé les résultats suivants :

Longueur (en mm)	[120; 125[	[125; 130[	[130; 135[	[135; 140[	[140; 145[
Effectif	10	20	38	25	7

1) Construire l'histogramme des effectifs.
2) On suppose que les tiges sont défectueuses si leur longueur est strictement inférieure à 125 mm ou supérieure ou égale à 140 mm. Quel est le pourcentage de pièces acceptables?
3) On suppose que, dans chaque classe, tous les éléments sont situés au centre. Calculer des valeurs approchées à 10^{-2} près de la moyenne et de l'écart type de cette série statistique.

Exercice 3

On a mesuré la durée de vie de 400 lampes produites dans une usine. On a obtenu les résultats suivants :

Durée de vie (en heures)	Nombre de lampes
[300; 500[	60
[500; 700[	134
[700; 900[	130
[900; 1100[	70
[1100; 1300[	6

1) Déterminer le pourcentage de lampes dont la durée de vie est strictement inférieure à 700 heures.
2) Déterminer le pourcentage de lampes dont la durée de vie est supérieure ou égale à 900 heures.
3) Représenter l'histogramme des fréquences cumulées croissantes.
4) On suppose que, dans chaque classe, les éléments sont répartis de manière uniforme. On peut alors remplacer l'histogramme par la ligne brisée définie par le point d'abscisse 300 et d'ordonnée 0 et chacun des sommets supérieurs droits des rectangles.
a) Tracer cette ligne brisée.
b) On se propose de déterminer le pourcentage de lampes dont la durée de vie est inférieure ou égale à 560 heures. Soient R et N les points de la ligne brisée de coordonnées respectives (500; 0,15) et (700; 0,485). Soit M le point du segment [RN] d'abscisse 560. Le pourcentage de lampes dont la durée de vie est inférieure ou égale à 560 heures est l'ordonnée du point M. Déterminer ce pourcentage.
c) Soit P le point de coordonnées (900; 0,81) de la ligne brisée. Soit I le point du segment [NP] d'ordonnée 0,50. La médiane est l'abscisse du point I. Déterminer la valeur approchée de la médiane.

Exercice 4

Suivant l'usage pour lequel ils sont fabriqués (cloison de bâtiments, chaussée d'autoroutes, …), les bétons doivent avoir une plus ou moins grande résistance à la compression 28 jours après leur fabrication. Cette résistance au bout de 28 jours, exprimée en mégapascals (MPa), est notée f_{c28}.
Une circulaire ministérielle fixe les critères de conformité. Lorsque l'effectif n de l'échantillon des prélèvements effectués pour contrôler la fabrication est supérieur ou égal à 15, les conditions suivantes doivent être remplies :

$$\begin{cases} \bar{x} - 1,2\sigma \geq f_{c28} & \text{si } f_{c28} > 25\text{MPa} \\ \bar{x} - 0,85\sigma \geq f_{c28} & \text{si } f_{c28} \leq 25\text{MPa} \\ f_{cmin} + 4 \geq f_{c28} \end{cases}$$

où $\bar{x}$ est la moyenne des n résultats mesurés, σ l'écart type et f_{cmin} la valeur minimale des n résultats.
Pour réaliser une chaussée d'autoroute, on utilise un béton dont la résistance f_{c28} doit être de 4,5 MPa.
On effectue 49 mesures de résistance. Les résultats (en MPa) sont regroupés en classes dans le tableau suivant :

Classe	Effectif
[3,0 ; 3,5[	1
[3,5 ; 4,0[	0
[4,0 ; 4,5[	0
[4,5 ; 5,0[	1
[5,0 ; 5,5[	0

Classe	Effectif
[5,5 ; 6,0[	5
[6,0 ; 6,5[	4
[6,5 ; 7,0[	17
[7,0 ; 7,5[	18
[7,5 ; 8,0[	3

1) Représenter cette série statistique par un histogramme.

2) Déterminer des valeurs approchées à 10^{-2} près de la moyenne $\bar{x}$ et de l'écart-type σ de la série statistique, en supposant que, dans chaque classe, tous les éléments sont situés au centre.

3) La fabrication est-elle conforme aux exigences réglementaires? (D'après un document de l'Association Technique pour le Développement du Treillis Soudé.)

Exercice 5

On a mesuré en millimètres les diamètres de 150 pièces usinées. On a obtenu les résultats suivants :

Diamètre (en mm)	Effectif
[19,70 ; 19,80[	2
[19,80 ; 19,85[	10
[19,85 ; 19,90[	14
[19,90 ; 19,95[	22
[19,95 ; 20,00[	32

Diamètre (en mm)	Effectif
[20,00 ; 20,05[	27
[20,05 ; 20,10[	26
[20,10 ; 20,15[	9
[20,15 ; 20,20[	3
[20,20 ; 20,30[	5

1) Déterminer la classe médiane. En admettant que la répartition de l'effectif est uniforme à l'intérieur de chaque classe, déterminer graphiquement la médiane. Que représente la médiane?

2) Déterminer la classe modale ainsi qu'une valeur approchée de l'étendue, c'est-à-dire de la différence entre la plus grande et la plus petite valeur de cette série statistique.

Annexe 6 : TD n°2

Certains de ces exercices (ou des données) sont tirés de l'ouvrage "Statistiques générales pour utilisateurs, 2-Exercices et corrigé" de François Husson et Jérôme Pagès.

Exercice 1

On s'intéresse à la relation entre les sensations d'amertume, d'acidité et de sucré. Pour cela une dégustation de 16 cocktails de jus de fruits composés de banane, mangue, orange et citron a été organisée auprès d'experts. Ces experts ont évalué l'acidité, la saveur sucrée et l'amertume de ces 16 cocktails à l'aide d'une échelle de 0 (pas du tout sucré / acide / amer) à 10 (extrêmement sucré / acide / amer). Pour chaque produit et chaque saveur, on calcule la moyenne des notes fournies par ces experts. Ces moyennes sont données dans le tableau 1 :

Cocktail	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Sucré	6.21	7.75	7.21	8.33	4.92	5.13	6.04	6.09	6.08	6.17	7.13	7.08	4.38	5.63	4.59	4.75
Acide	7.08	3.29	4.38	2.79	7.71	7.50	6.58	5.16	5.50	5.58	3.54	3.25	8.33	5.13	8.33	6.37
Amer	2.00	1.54	1.79	1.63	1.96	2.13	2.04	2.00	2.09	2.13	1.52	1.46	2.29	2.13	2.46	2.17

Tab.1 - Valeur d'acidité, d'amertume et de saveur sucrée pour 16 cocktails (moyenne sur 12 experts)

1) Calculer le coefficient de corrélation entre les saveurs sucré et acide, puis entre les saveurs sucré et amer et enfin, entre les saveurs acide et amer. Peut-on conclure à une dépendance des saveurs ?
2) Tracer le graphique (saveur sucré, saveur acide). Peut-on modéliser le problème par une régression linéaire? Calculer alors les coefficients de la droite de régression.
3) Refaire la question 2 pour les deux autres croisements.

Exercice 2

Chaque diode a un code couleur fonction de sa résistance. Si on ne connaît pas le code couleur, on peut estimer la résistance par l'expérience suivante: on fait varier l'intensité de 10 mA à 100 mA et on mesure la tension par un tensiomètre (données du tableau 2).

Intensité (en A)	0.01	0.02	0.04	0.05	0.075	0.1
Tension (en V)	5	9.5	18	23	35	48

Tab. 2 – Tension et résistance d'une diode

1) Tracer le graphique (intensité, tension). Peut-on valider l'hypothèse d'utilisation d'une droite de régression pour modéliser graphiquement le problème ?
2) Calculer le coefficient de corrélation entre intensité et tension.
3) Calculer les coefficients de la droite de régression $y_i = ax_i + b$ où x_i représente l'intensité et y_i représente la tension.

Exercice 3

On souhaite étudier le lien entre la taille d'une fille (Y) et la taille de ses parents (taille de la mère: X_1 et taille du père: X_2). Pour cela, on a demandé à 140 filles de donner leur taille ainsi que celles de leurs parents.

On se limite dans un premier temps à l'étude d'un échantillon de 5 filles (tableau 3).

Fille	1	2	3	4	5
Taille de la fille	1.62	1.73	1.66	1.68	1.70
Taille de la mère	1.57	1.61	1.55	1.60	1.69
Taille du père	1.82	1.87	1.86	1.72	1.88

Tab. 3 – Extrait du tableau des données des tailles des filles en fonction de la taille des parents

1) Tracer le graphique (taille mère, taille fille).
2) Calculer les coefficients de corrélation entre taille mère et taille fille, puis entre taille père et taille fille.
3) Le jeu de données complet est en fait obtenu à partir d'un échantillon de 140 individus.
Les corrélations entre les trois tailles sont données dans le tableau 4. Interpréter ces résultats.

	Fille	Père	Mère
Fille	1		
Père	0.43793	1	
Mère	0.47422	0.46748	1

Tab. 4 – Matrice des corrélations entre la taille de la fille, de sa mère et de son père.

Exercice 4
Nous avons observé 10 étudiants dans 3 matières différentes. Le tableau suivant représente ces données :

Algèbre	Analyse	Mécanique
10.6	14.5	10.1
11.2	14.8	13.9
7.3	12.8	10.4
12	16.5	11.4
10.1	9.2	10.1
10.2	13.5	5.6
8.3	9.2	14.1
12.9	15.1	13.8
12.2	10.2	13.1
3.8	4.1	4.9

Tab. 5 – Notes de 10 étudiants

1) Tracer le graphique (analyse, algèbre).
2) Donner le tableau des coefficients de corrélation (en croisant les 3 caractères statistiques).
3) Le tableau suivant correspond aux coefficients de corrélation sur la population entière (soit 200 étudiants)

	Algèbre	Analyse	Mécanique
Algèbre	1		
Analyse	0.83	1	
Mécanique	0.39	0.37	1

Tab. 6 – Matrice des corrélations entre Analyse, Algèbre et Mécanique.

Interpréter ce tableau.

Annexe 7 : TD n°3

Exercice 1

Pour estimer la proportion *p* de pièces défectueuses à la sortie d'une chaîne de production, on prélève un échantillon de n_1 pièces. A la i^{ème} pièce tirée, on associe la v.a.

$$X_i = \begin{cases} 1 & \text{si la pièce est défectueuse,} \\ 0 & \text{sinon} \end{cases}$$

1) Montrer que la statistique $F_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ est un estimateur sans biais et convergent de *p*.

2) On tire un deuxième échantillon de n_2 pièces, indépendamment du premier. On note F_{n_1} et F_{n_2} les estimateurs de *p* définis comme précédemment pour les échantillons 1 et 2 respectivement.

a) $F = \frac{1}{2}\left(F_{n_1} + F_{n_2}\right)$ est-il un estimateur sans biais de p ? Quelle est sa variance ?

b) On pose $F_{(a,b)} = aF_{n_1} + bF_{n_2}$ Donner une condition sur *a* et *b* pour que $F_{(a,b)}$ soit un estimateur sans biais de p et avec une variance minimale.

Exercice 2

On considère T la variable aléatoire : "durée d'attente à un feu rouge". La durée du feu rouge est égale à θ, paramètre inconnu strictement positif.

On observe un échantillon $t_1, t_2, \ldots, t_n$ de taille n, où t_i désigne la durée d'attente observée pour le i^{ème} individu. On fait l'hypothèse que les variables aléatoires $T_1, T_2, \ldots, T_n$ sont indépendantes et de même loi uniforme sur [0; θ], notée $U[0; \theta]$.

1) Soit la statistique $\overline{T} = \frac{1}{n}\sum_{i=1}^{n} T_i$. Calculer IE$\left(\overline{T}\right)$ et Var$\left(\overline{T}\right)$.

Montrer que la statistique $\widehat{\theta_1} = 2\overline{T}$ est un estimateur sans biais et convergent de θ.

2) Soit la statistique $Y_n = \sup_i T_i$.

a) En utilisant l'équivalence des événements $\left(Y_n < y\right)$ et $\left(\forall i = 1,...,n \quad T_i < y\right)$, calculer la fonction de répartition de Y_n. En déduire sa densité puis calculer IE(Y_n) et Var(Y_n).

b) Montrer que la statistique $\widehat{\theta_2} = \frac{n+1}{n} Y_n$ est un estimateur sans biais et convergent de θ.

3) Comparer les variances Var$\left(\widehat{\theta_1}\right)$ et Var$\left(\widehat{\theta_2}\right)$. Lequel des deux estimateurs $\widehat{\theta_1}$ et $\widehat{\theta_2}$ choisiriez-vous pour estimer θ ?

<u>*Application*</u> : pour n = 10, on a ($t_1, \ldots, t_{10}$) = (28; 33 ;42 ;15 ;20 ;27 ;18 ;40 ;16 ;25).

Quelle est l'estimation de la durée du feu rouge ?

Exercice 3

On considère un *n*-échantillon $X_1, \ldots, X_n$ de loi gaussienne $\mathcal{N}\left(\mu;\sigma^2\right)$.

On note : $\overline{X} = \frac{1}{n}\left(X_1 + \ldots + X_n\right)$ et $\overline{Y} = \overline{X}\left(1 - \overline{X}\right)$.

1) Montrer que $\overline{Y}$ n'est pas un estimateur sans biais de $\mu(1-\mu)$.

2) Comment modifier $\overline{Y}$ pour qu'il devienne sans biais ?

Indication : pensez à un estimateur de la variance…

Exercice 4

Soit X une v.a. de loi $\mathcal{N}(\theta;\theta(1-\theta))$ où $0 \le \theta \le 1$. A partir d'un échantillon $X_1, \ldots, X_n$ indépendants et identiquement distribués de cette loi, on construit les estimateurs :

$$T_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad S_n = \frac{1}{n}\sum_{i=1}^{n} (X_i)^2$$

1) Etudier leurs propriétés de biais et de convergence.

Indication : on admet que $\mathrm{Var}(X^2) = 2\theta^2(1-\theta^2)$.

2) L'un de ces estimateurs peut-il être considéré comme toujours meilleur que l'autre ?

Exercice 5

On considère un *n*-échantillon $X_1, \ldots, X_n$ de loi exponentielle de paramètre λ.
On définit un nouvel échantillon $Y_1, \ldots, Y_n$ tel que :
$Y_i = 1$ si $X_i > 1$, $Y_i = 0$ si $X_i \le 1$.

Montrer que $\overline{Y} = \frac{1}{n}(Y_1 + \ldots + Y_n)$ est un estimateur sans biais et convergent de $e^{-\lambda}$.

Exercice 6

On souhaite étudier le système de traitement radar. Pour cela, on a observé le système durant presque 7 ans. On suppose que le temps de réparation est négligeable par rapport au temps de bon fonctionnement.

1) En supposant que le temps de bon fonctionnement suit un processus de Poisson homogène et sachant que nous avons observé 10 défaillances durant ces 7 ans, donner une estimation du paramètre λ, en pannes/mois, de ce processus.

2) Le tableau suivant détaille les temps de fonctionnement du processus.

Défaillance	1	2	3	4	5	6	7	8	9	10
Temps	5	13	30	36	52	53	62	64	79	82

Tab. 1 - Temps de défaillance observé en mois.

Calculer une estimation de λ par la méthode du plan d'essai de type II.

3) Calculer alors la probabilité d'avoir deux défaillances sur un an (on prendra pour λ la valeur estimée dans la question précédente).

4) Calculer la probabilité d'avoir une défaillance sur deux ans.

Annexe 8 : TD n°4

Exercice 1

Nous avons observé 11 ILS jusqu'à ce que chacun ait une défaillance. Le tableau suivant donne les temps, en jour, de première défaillance pour chaque ILS.

ILS	1	2	3	4	5	6	7	8	9	10	11
Temps	130	20	348	100	14	212	64	50	135	224	67

Tab. 1 – Temps de première défaillance observé pour des ILS de catégorie III.

1) Déterminer une estimation de R(t).
2) En supposant que le temps de bon fonctionnement suit une loi exponentielle de paramètre λ = 0,007 panne/jour, déterminer à quel instant t_0, la fiabilité est égale à 80%.
3) Calculer la probabilité qu'un ILS n'ait aucune défaillance sur une période d'un an.
4) Peut-on justifier graphiquement l'utilisation d'une loi exponentielle pour modéliser le temps de bon fonctionnement des éléments étudiés ? Si oui, donnez une estimation du MTTF et du paramètre λ.

Exercice 2

Nous avons observé les durées de vie de composants électroniques, intervenant dans la fabrication des serveurs STPV. Le tableau suivant donne ces durées de vie, en mois.

Composant	1	2	3	4	5	6	7	8	9	10
Durée de vie	21	65	7	4	35	6	1	8	16	12

Tab. 2 – Durée de vie des composants électroniques.

1) Déterminer une estimation de R(t).
2) Peut-on justifier graphiquement l'utilisation d'une loi exponentielle pour modéliser le temps de bon fonctionnement des éléments étudiés ?
Si oui, donnez une estimation du MTTF et du paramètre λ.
3) Justifier que l'on peut estimer λ par $\hat{\lambda}$ tel que $\frac{1}{\hat{\lambda}} = \frac{1}{n}\sum_{i=1}^{n} t_i$ et en déduire la probabilité qu'un composant ait une durée de vie supérieure à 100 mois.

Exercice 3

Des moteurs CFM56, qui équipent les A320, ont été testés en salle. Nous avons observé leur temps de première défaillance, et ce jusqu'à la période d'un an maximum.

Moteurs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Temps (en jours)	149	360	365**	365**	40	66	15	365**	260	150*	150*	166	275*	48*	113

Tab. 3 - Temps de première défaillance des moteurs CFM56. Les temps suivis d'une astérisque correspondent à des moteurs ayant subi une autre batterie de tests incompatibles avec notre test de fiabilité. Ces moteurs doivent être considérés comme des éléments suspendus. Les temps suivis de deux astérisques correspondent à des moteurs n'ayant eu aucune défaillance.

1) Donner une estimation de R(t) en utilisant la méthode de Kaplan-Meier.
2) Peut-on justifier graphiquement l'utilisation d'une loi exponentielle pour modéliser le temps de bon fonctionnement des éléments étudiés ?
Si oui, donnez une estimation du MTTF et du paramètre λ.

Exercice 4
Pour la construction des amortisseurs d'avions, on cherche à tester la fiabilité des pistons. On a comptabilisé les défaillances sur 6 amortisseurs, suivis pendant une période de 5 ans.

Numéro d'amortisseur	1	2	3	4	5	6
Temps (en heure de vol)	3300	1670	2250	7720	3050*	5600

Tab. 4 - Temps de première défaillance des amortisseurs. Le temps suivi d'une astérisque correspond à un avion crashé, dont l'accident est indépendant d'un problème d'amortisseur.

1) Donner une estimation de R(t) en utilisant la méthode de Johnson.

2) Peut-on justifier graphiquement l'utilisation d'une loi exponentielle pour modéliser le temps de bon fonctionnement des éléments étudiés ?
Si oui, donnez une estimation du MTTF et du paramètre λ.

Exercice 5
Une analyse de la fiabilité des moteurs et des trains d'atterrissage des Airbus A319 a été effectuée en salle sur une période de 13000 heures de vol. Les tableaux suivants représentent les instants de premières défaillances d'un moteur, pour le premier tableau, et les instants de premières défaillances d'un train d'atterrissage, pour le second tableau.

(a) Moteur CFM56. Le temps suivi d'un astérisque correspond à un moteur n'ayant subi aucune défaillance. Le temps suivi de deux astérisques correspond à un moteur retiré de l'analyse (et doit donc être considéré comme un élément suspendu).

Numéro de moteur	1	2	3	4	5	6	7	8	9	10
Heures de vol	3035	6691	3776**	13000*	354	10101	1111	7296	11439	754

(b) Train d'atterrissage

Numéro du train	1	2	3	4	5	6	7	8	9	10
Heures de vol	4851	4191	3514	3964	1787	1114	53	1460	3087	695

Tab. 1 - Temps de premières défaillances, exprimés en heures de vol.

1) Calculer une estimation de la fiabilité d'un moteur en utilisant la méthode de Kaplan-Meier.
2) Calculer une estimation de la fiabilité d'un train d'atterrissage.
3) Peut-on justifier graphiquement l'utilisation d'une loi exponentielle pour modéliser le temps de bon fonctionnement des éléments étudiés ?
Si oui, donnez une estimation du MTTF et du paramètre λ.

Exercice 6
Dans cette partie, nous nous intéressons à un seul élément du système de navigation de bord. Le tableau suivant représente les instants de premières défaillances de ces éléments, observés sur 6 avions.

Numéro de l'avion	1	2	3	4	5	6
Temps de défaillance	5	10*	40	25	39	38*

Tab. 2 - Temps de premières défaillances, exprimés en mois. Les astérisques représentent des éléments suspendus.

1) Donner une estimation de la fiabilité R(t), en utilisant la méthode de Johnson.

2) Peut-on justifier graphiquement l'utilisation d'une loi exponentielle pour modéliser le temps de bon fonctionnement des éléments étudiés ?
Si oui, donnez une estimation du MTTF et du paramètre λ.

Exercice 7

On a étudié le système de surveillance radar (SSR) et plus particulièrement les transpondeurs. Le tableau suivant donne l'instant de première défaillance de plusieurs transpondeurs.

Numéro du transpondeur	1	2	3	4	5	6	7	8	9
Instant de défaillance	563	908	130	477	198	133*	417	153	269

Tab. 3 - Instants de premières défaillances des transpondeurs, exprimés en heures. La valeur suivie d'un astérisque correspond à un crash dont la responsabilité n'est pas liée au transpondeur.

1) Calculer une estimation de la fiabilité d'un transpondeur en utilisant la méthode de Kaplan-Meier.

2) Peut-on justifier graphiquement l'utilisation d'une loi exponentielle pour modéliser le temps de bon fonctionnement des éléments étudiés ?
Si oui, donnez une estimation du MTTF et du paramètre λ.

Annexe 9 : TD n°5

Exercice 1

Soit un échantillon de taille 25. La variable aléatoire est une mesure de température qui suit une loi gaussienne. La moyenne de l'échantillon est de 253°C et l'écart-type de l'échantillon est de 6.0°C.

1) Quel est l'intervalle de confiance de la moyenne de la population au coefficient de sécurité de 95%?

2) Quel est l'intervalle de confiance de la moyenne de la population au coefficient de sécurité de 60%?

Exercice 2

Choisissez au hasard 5 nombres différents entre 1 et 96 (par exemple, à l'aide de la fonction aléatoire d'une calculatrice). Rechercher alors la superficie des cinq départements français correspondant à ces numéros dans le tableau ci-dessous. La Corse ayant les numéros 2A et 2B, si le nombre choisi est égal à 20 on prendra la superficie du département 2A, et si le nombre choisi est 96, on prendra la superficie du département 2B.

1) Si on suppose que les superficies des départements suivent une loi Normale (on fait le calcul avec cette hypothèse que l'on sait fausse), construire, à partir de votre échantillon, l'intervalle de confiance à 90% pour la surface moyenne des départements français.

2) Déterminer l'intervalle de confiance à 90% pour la surface de la France métropolitaine.

3) En fait, la surface totale de la France métropolitaine est de 543965 km^2 (sans les DOM-TOM). Cette aire appartient-elle à votre intervalle de confiance ?

Dép.	Superf.	Dép.	Superf.	Dép.	Superf.	Dép.	Superf.
1	5762	25	5234	49	7166	73	6028
2	7369	26	6530	50	5938	74	4388
3	7340	27	6040	51	8162	75	105
4	6925	28	5880	52	6211	76	6278
5	5549	29	6733	53	5175	77	5915
6	4299	30	5853	54	5241	78	2284
7	5529	31	6309	55	6216	79	5999
8	5229	32	6257	56	6823	80	6170
9	4890	33	10000	57	6216	81	5758
10	6004	34	6101	58	6817	82	3718
11	6139	35	6775	59	5743	83	5973
12	8735	36	6791	60	5860	84	3567
13	5087	37	6127	61	6103	85	6720
14	5548	38	7431	62	6671	86	6990
15	5726	39	4999	63	7970	87	5520
16	5956	40	9243	64	7645	88	5874
17	6864	41	6343	65	4464	89	7427
18	7235	42	4781	66	4116	90	609
19	5857	43	4977	67	4755	91	1804
20	4014	44	6815	68	3525	92	176
21	8763	45	6775	69	3249	93	236
22	6878	46	5217	70	5360	94	245
23	5565	47	5361	71	8575	95	1246
24	9060	48	5167	72	6206	96	4666

Exercice 3
Un échantillon de 30 cigarettes d'une même marque a donné les teneurs en goudron, exprimées en mg, suivantes :

Teneurs en goudron	11,7	11,8	12,3	12,4	12,5	12,7	12,8	12,9	13,1	13,4	13,5	14,5
effectifs	1	2	1	2	3	2	6	5	1	4	1	2

La norme en vigueur recommande une teneur en goudron inférieure ou égale à 13mg par cigarette.
1) Donner une estimation ponctuelle de la proportion de cigarettes de cette marque qui respectent la norme de la teneur en goudron.
2) Donner une estimation ponctuelle de la teneur en goudron moyenne des cigarettes de cette marque, puis de l'écart type.
3) a) Donner un intervalle de confiance à 95% de la teneur en goudron en mg d'une cigarette de cette marque.
b) Peut-on en déduire, avec cette confiance, que la norme en vigueur est respectée ?

Exercice 4
Un fabricant de piles électriques affirme que la durée de vie moyenne du matériel qu'il produit est de 170 heures. Un organisme de défense des consommateurs prélève au hasard un échantillon de $n = 100$ piles et observe une durée de vie moyenne de 158 heures avec un écart-type empirique σ_n =30 heures.
1) Déterminer un intervalle de confiance à 99% pour la durée de vie moyenne m.
2) Peut-on accuser ce fabricant de publicité mensongère ?

Exercice 5
Un sondage sur la popularité d'un élu indique que 51% des personnes interrogées sont favorables à sa politique.
1) Construire un estimateur $\widehat{p_n}$, puis un intervalle de confiance à 95% de la proportion p d'électeurs qui lui sont favorables, sachant que ce sondage a été réalisé auprès de $n = 100$ personnes.
2) Même question si $n = 1000$.
3) Quelle doit être la valeur minimale de n pour que la longueur de cet intervalle soit au plus égale à 4%? Conclure sur le peu de confiance que l'on doit accorder aux estimations des intentions de vote données dans la presse pour certaines élections très serrées.

Exercice 6
Un sondage effectué avant les présidentielles de mai 2007 auprès de 957 personnes prédisait le résultat suivant pour le second tour des élections :
- S. Royal : 48%
- N. Sarkozy : 52%

1) A partir des données de ce sondage, construire un intervalle de confiance à 90%, puis à 80% du pourcentage de votes pour les deux candidats du second tour.
2) A la lecture de ces résultats, l'issue du scrutin était-il incertain (au risque de 10%) ?
3) Le 6 mai 2007, le résultat officiel du vote était :
- S. Royal : 46,94%
- N. Sarkozy : 53,06%

Commenter le résultat du sondage.
4) En supposant que quel que soit le nombre de personnes interrogées, les pourcentages de votes aient été les mêmes (48% - 52%), combien de personnes aurait-il fallu interroger au minimum pour être sûr à 95% de la victoire de N. Sarkozy ?

Exercice 7
L'article suivant est extrait du journal *Le Monde* du *3 mars 1983*.

« Bourse de New-York, nouveau record.
La reprise frappe à la porte. Wall Street en est maintenant convaincu, après la publication faite par le Département du Commerce des principaux indicateurs économiques pour janvier.
Dans ces statistiques, il ressort que l'indice des valeurs industrielles a monté de 3,6%.
Cette hausse mensuelle est la plus forte enregistrée depuis 1950. Elle est surtout supérieure aux prévisions les plus optimistes que les boursiers avaient pu faire.
Beaucoup sont maintenant persuadés autour du Big board que le marché est entré dans une nouvelle phase d'ascension. La clientèle particulière a, pour sa part, fait un retour très marqué que certains jugent significatif.
Sur les 1970 valeurs traitées le 2 mars, 1168 ont monté, 469 ont baissé et 333 n'ont pas varié.
Voici une sélection des cours du jour :

Valeurs	*Cours du 1er mars*	*Cours du 2 mars*
Alcoa	*34 3/4*	*35 1/8*
ATT	*67 1/2*	*66 7/8*
Boeing	*37*	*36 7/8*
Chase Manhattan Bank	*47 1/4*	*48 7/8*
Du Pont de Nemours	*40 5/8*	*41*
Eastman Kodak	*89 1/4*	*89*
Exxon	*30*	*30 7/8*
Ford	*40 1/2*	*41 3/8*
General Electric	*111*	*108*
General Foods	*39 3/8*	*39 1/2*
General Motors	*63 1/2*	*63*
Goodyear	*31 3/4*	*31 5/8*
IBM	*101 3/4*	*102 1/8*
ITT	*33 3/8*	*33 3/4*
Mobil Oil	*27 1/4*	*28 1/2*
Pfizer	*72 3/4*	*74 1/4*
Schlumberger	*40 1/2*	*41 7/8*
Texaco	*32 1/8*	*33*
U.A.L. Inc.	*34*	*34 3/8*
Union Carbide	*61 1/2*	*61 1/4*
U.S. Steel	*22 3/4*	*22 7/8*
Westinghouse	*48 5/8*	*49 1/4*
Xerox Corp	*38 5/8*	*39 5/8 »*

1) D'après la sélection des valeurs dont les cours ont été publiés, donner un intervalle de confiance à 95% pour la proportion de valeurs ayant strictement monté lors de la séance du 2 mars. Que penser du résultat ?
2) D'après les mêmes données, quel est l'intervalle de confiance à 95% pour la proportion de valeurs ayant vu leur cours inchangé ? Que penser de ce résultat ?

Annexe 10 : TD n°6

Exercice 1
On lance 4000 fois une pièce de monnaie. On obtient 1870 fois « face ». La pièce est-elle truquée ?

Exercice 2
On étudie la circulation en un point fixe d'une autoroute en comptant, pendant deux heures, le nombre de voitures passant par minute devant un observateur. Le tableau suivant résume les données obtenues :

Nombres de voitures	0	1	2	3	4	5	6	7	8	9	10	11
Fréquence observée	4	9	24	25	22	18	6	5	3	2	1	1

Tester l'adéquation à une loi de Poisson pour un risque $\alpha = 0{,}10$.
Rappel : Le paramètre d'une loi de Poisson peut être estimé par la moyenne empirique d'un échantillon.

Exercice 3
A la sortie d'une chaîne de fabrication, on prélève toutes les trente minutes un lot de 20 pièces mécaniques et on contrôle le nombre de pièces défectueuses du lot. Sur 200 échantillons indépendants, on a obtenu les résultats suivants :

Nombres de défectueux	0	1	2	3	4	5	6	7
Nombres de lots	26	52	56	40	20	2	0	4

Tester l'adéquation de la loi empirique du nombre de pièces défectueuses par lot de 20 pièces à une loi théorique simple, par exemple la loi binomiale, pour un risque $\alpha = 0,05.$

Exercice 4
Un examen est ouvert à des étudiants de formations différentes : économie, informatique et mathématiques. Le responsable de l'examen désire savoir si la formation initiale d'un étudiant influe sur sa réussite. A cette fin, il construit le tableau ci-dessous à partir des résultats obtenus par les 286 candidats :

	Economie	Informatique	Mathématiques
Réussite	33	51	70
Echec	29	44	59

Quelle est sa conclusion?

Exercice 5
On veut évaluer les liens entre le sexe et le fait de fumer et le lien entre le fait de fumer et le fait d'avoir des parents fumeurs. Pour cela, on a recueilli des données auprès de 123 étudiants (voir le tableau suivant).

	Homme	Femme	Père fumeur et mère fumeur	Père fumeur et mère non fumeur	Père non fumeur et mère fumeur	Père non fumeur et mère non fumeur
Fumeur	24	41	13	16	7	29
NonFumeur	23	35	5	24	6	23

1) Le fait de fumer est-il indépendant du sexe?
2) Le comportement d'un individu vis-à-vis de la cigarette dépend-il du comportement de ses parents?

Exercice 6
On a effectué une étude, en milieu urbain et en milieu rural, sur le rythme cardiaque humain :

	Milieu urbain	Milieu rural
Effectif de l'échantillon	300	240
Moyenne de l'échantillon	80	77
Variance de l'échantillon	150	120

Peut-on affirmer qu'il existe une différence significative entre les rythmes cardiaques moyens des deux populations ?

Exercice 7
On veut tester l'impact des travaux dirigés dans la réussite à l'examen de statistique.

	Nombre d'étudiants du groupe	Nombre d'étudiants ayant réussi à l'examen
Groupe 1 (20 h de TD)	180	126
Groupe 2 (30h de TD)	150	129

Que peut-on en conclure ?

Exercice 8
Un biologiste effectue des dosages par une méthode de mesure de radioactivité et ne dispose donc que d'un nombre très limité de valeurs.
Les concentrations C_1 et C_2 mesurées sur deux prélèvements ont donné les valeurs suivantes :

C_1 : 3,9 – 3,8 – 4,1 – 3,6 C_2 : 3,9 – 2,8 – 3,1 – 3,7 – 4,1

En admettant que la concentration est distribuée normalement, les variances des deux séries peuvent-elles être estimés similaires ?

Exercice 9
Deux parcelles identiques de vignes atteintes de phylloxera ont été traitées avec des traitements différents : l'une avec des traitements « bio » (traitement 1), l'autre avec des pesticides (traitement 2).
Les résultats sont consignés dans le tableau suivant :

	Eradication	Amélioration	Pas d'effet
Traitement 1	213	196	167
Traitement 2	304	227	156

Peut-on affirmer que les traitements ont le même effet ?

Exercice 10
Après avoir observé la durée de vie des ILS, on a obtenu le tableau suivant donnant les instants (en mois) de défaillance.

Instants de défaillance	5	7	12	24	25	36	41	62	65	75

1) Faire un test non paramétrique en regroupant les défaillances par dizaines de mois pour savoir si le processus de Poisson suivi par les instants de défaillances est homogène.
2) Faire de même avec un test de Laplace.

Annexe 11 : TD n°7

Exercice 1

Nous observons deux caractères statistiques sur un même individu (une moule). Il s'agit du poids brut et du poids utile. Le tableau suivant représente 10 observations de ces deux caractères.

Poids brut	3.75	7.05	5.83	9.58	1.31	8.58	7.16	4.80	6.32	6.90
Poids utile	1.51	2.60	1.33	3.49	0.36	3.57	2.98	1.74	3.12	3.08

1) Calculer les paramètres de position (moyenne, médiane) des deux séries.
2) Calculer les paramètres de dispersion (variance, écart-type, écart inter-quartile) des deux séries.
3) Calculer le coefficient de corrélation entre les deux séries. Peut-on dire qu'il y a une dépendance entre les deux caractères statistiques? Si oui, donner l'équation de la droite de régression.

Exercice 2

Dans un concours, les candidats passent deux épreuves de mathématiques, l'une en probabilités, l'autre en statistiques.
On note X la note obtenue à l'épreuve de probabilités, et Y celle obtenue à l'épreuve de statistiques. Les résultats obtenus pour 104 candidats sont donnés dans le tableau ci-dessous :

X \ Y	[0 ; 4[	[4 ; 8[	[8 ; 12[	[12 ; 16[	[16 ; 20[	Total
[0 ; 4[	1	1	0	0	0	2
[4 ; 8[	1	3	5	11	0	20
[8 ; 12[	2	10	10	28	0	50
[12 ; 16[	0	1	3	9	11	24
[16 ; 20[	0	0	2	4	2	8
Total	4	15	20	52	13	104

1) Calculer la moyenne et l'écart type de X et de Y.
2) Calculer la fréquence conditionnelle de $X \in [12 ; 16[$ sachant que $Y \in [4 ; 8[$.
3) Calculer la covariance de X et Y, et le coefficient de corrélation.

Exercice 3

Nous nous intéressons à un élément du système de navigation de bord. Le tableau suivant représente les instants de premières défaillances de ces éléments, observés sur 10 avions.

Numéro de l'avion	1	2	3	4	5	6	7	8	9	10
Temps de défaillance	6	8*	35	65	30	80*	52	88	20	36*

Temps de premières défaillances, exprimés en mois. Les astérisques représentent des éléments suspendus.

1) Donner une estimation de la fiabilité *R(t)*, en utilisant la méthode de Kaplan-Meier.
2) Peut-on justifier d'une loi exponentielle pour modéliser le temps de bon fonctionnement des éléments étudiés ? Si oui, donner une estimation du MTTF et du paramètre λ.

Exercice 4

Soit *X* une V.A.R. de densité : $f_X(x) = \frac{2}{\theta} e^{-\frac{2}{\theta}x} 1_{\mathbb{R}_+}(x)$, dont on cherche à estimer le paramètre θ.

1) A l'aide d'une I.P.P., montrer que la moyenne empirique est un estimateur biaisé de θ.

2) Proposer un estimateur sans biais de θ, fonction de $\overline{X}$. Ce nouvel estimateur est-il convergent ?

Exercice 5
Un biologiste étudie un type d'algue qui attaque les plantes marines. La toxine contenue dans cette algue est obtenue sous forme de solution organique. Il mesure la quantité de toxine par gramme de solution. Il a obtenu les mesures suivantes, exprimées en milligrammes :

Quantité	0,5	0,6	0,7	0,8	0,9	1	1,1	1,2	1,3	1,4	1,5
Fréquence	1	4	6	5	9	8	12	10	7	5	1

1) Calculer la moyenne et l'écart type de l'échantillon proposé.
2) On suppose que la quantité de toxine par gramme de solution suit une loi normale $\mathcal{N}(\mu;\sigma^2)$.
Déterminer un intervalle de confiance à 95% pour l'espérance μ de la quantité de toxine par gramme de solution.
.

Exercice 6
La presse affirme que parmi les Français regardant la télévision plus de 4 heures par jour, on a la même proportion de personnes dans chaque tranche d'âge. Sur un échantillon de 35 personnes, on a observé les données suivantes :

Age	Moins de 20 ans	De 20 à 40 ans	De 40 à 60 ans	Plus de 60 ans
Effectif	9	8	5	13

1) Cette étude confirme-t-elle l'opinion de la presse au seuil de 5% ?
2) Que penser de l'opinion selon laquelle la moitié des Français regardant la télévision plus de 4 heures par jour a plus de 60 ans, au seuil 5% ?

Exercice 7
Une étude américaine effectuée sur 106 patients a donné lieu au tableau suivant :

	A un cancer du poumon	N'a pas de cancer du poumon
Est fumeur	60	32
N'est pas fumeur	3	11

Peut-on considérer au risque de 5% qu'il existe un lien entre le fait de fumer et avoir un cancer du poumon ?

Exercice 8
Deux villages de vacances proposent des activités aquatiques à leurs clients. Les effectifs de l'été sont donnés ci-dessous :

	Planche à voile	catamaran	Plongée sous marine
Village 1	124	61	57
Village 2	91	55	12

Peut-on estimer que ces deux villages ont des clients « semblables » ?