Intitulé du cours

Méthodes statistiques en chimie analytique

Master de Chimie analytique

Par DJELLALI Abdelkhalek, Ph.D

E-mail: sciencesgaa@yahoo.fr

Téléphone: 06 61 17 57 10

Département de Mathématiques

Faculté des Sciences

UBM Annaba

Année 2019/2020

Avant propos

A propos de l'élaboration du cours

Ce cours est un recueil de différents textes appartenant à différents auteurs. Ces textes ont été adaptés par le prof pour répondre à l'exigence du programme proposé. Il les a homogénéisés dans une idée d'ensemble pour avoir une consistance de structure et de contenu. Certains de ces textes paraissent ne pas être d'une grande luminosité et parfois ils paraissent pécher par des incomplétudes. Il n'en est rien de tout ça parce que d'une part ces textes sont enseignés dans de grandes universités et de l'autre part ils ont été conçus intentionnellement pour répondre à des objectifs de programmes. Ceci dit, ils sont clairs dans leur contenu. C'est la raison qui a fait que le prof à éviter de les reformuler. Cette approche répond à l'usage d'une didactique des mathématiques qui a pour but de susciter la curiosité des étudiants pour mieux appréhender le cours. De cette façon, le prof se donne le devoir de combler les brèches en répondant aux différentes questions des étudiants.

Cette brochure est conçue par devoir pour terminer le cours que les étudiants n'ont pas eu le temps nécessaire pour suivre normalement.

A propos de l'enseignement

Le prof a enseigné les deux premières parties du cours avec force détails. Il les a étoffés avec plusieurs exemples et différents exercices. Donc, le plus important a été enseigné. La seule partie qui n'a pas été professée c'est la troisième. Dans ce cours, elle est nettement exposée et est enrichie par maintes applications. Nonobstant, le prof reste à la disposition des étudiants pour répondre à toutes les questions qu'ils veulent poser.

A propos de la note de T.D.

Elle compte pour 50% de la note de l'examen. Elle va être attribuée selon le travail de chaque étudiant selon ses capacités à traiter les donnés du sujet de l'heure : Le Covid-19. Le sujet est proposé ci-après par le prof.

SUJET DU T.D.

Chaque étudiant doit choisir une et une seule des populations suivantes et faire le travail statistique qui lui est demandé :

1. La population Algérienne
2. La population de Blida
3. La population d'Alger
4. La population d'Oran
5. La population Chinoise
6. La population de Wuhan
7. La population Pékinoise
8. La population des USA
9. La population de New York
10. La population de New Jersey
11. La population du Massachusetts
12. La population Italienne
13. La population de Lombardie
14. La population Romaine
15. La population Française
16. La population Parisienne
17. La population Lyonnaise
18. La population Niçoise
19. La population Espagnole
20. La population Madrilène
21. La population Catalane
22. La population Allemande
23. La population Berlinoise
24. La population Suisse
25. La population Genevoise
26. La population de Lausanne
27. La population Marocaine
28. La population de Rabat
29. La population de Marrakech
30. La population Tunisienne

31. La population Tunisoise
32. La population de Sousse
33. La population Egyptienne
34. La population Cairote
35. La population d'Alexandrie
36. La population de l'Inde
37. La population de New Delhi
38. La population de Calcutta
39. La population de Bombay

Une fois la population choisie, l'étudiant doit faire le travail suivant.

<u>Partie 1 :</u> Collecte de données et élaboration des tableaux sur le covid-19

1. Pour les données du covid-19, il faut établir un tableau de données pour chaque semaine. Pour trouver les données, il faut aller sur les sites de données d'Internet. Par exemple le site de John Hopkins.

Le premier tableau commence à partir du premier jour de l'apparition de covid-19 et il va jusqu'au $7^{ème}$ jour.

Le jour x_i	Le nombre des infectés n_i	Le nombre des décès n_i'	Le nombre des guéris n_i''
1			
2			
3			
4			
5			
6			
7			
Total de la semaine			

Le deuxième tableau commence à partir du $8^{ème}$ jour après l'apparition de covid-19 et il va jusqu'au $14^{ème}$ jour.

Le jour x_i	Le nombre des infectés n_i	Le nombre des décès n_i'	Le nombre des guéris n_i''
8			
9			
10			

11			
12			
13			
14			
Total de la semaine			

Le dernier tableau commence à partir du $n^{ème}$ jour après l'apparition de covid-19 et il va jusqu'au 15 mai (ou 16 mai ou 17 mai ou 18 mai, jusqu'au $7^{ème}$ jour pour boucler la semaine)

Le jour x_i	Le nombre des infectés n_i	Le nombre des décès n_i'	Le nombre des guéris n_i''
8			
9			
10			
11			
12			
13			
14			
Total de la semaine			

2. A partir des données des tableaux précédents, il faut établir un tableau de toutes les données regroupées par semaine.

Semaine	n_i	f_i	f_ic	n_i'	f_i'	f_ic'	n_i''	f_i''	f_ic''
[1 - 7[									
[8 - 14[									
[15 - 21[									
Total de la période									

<u>Partie 2 :</u> Définition analytique du mouvement de covid-19

1. Déterminer les modes de l'ensemble des infectés

2. Déterminer les modes de l'ensemble des décès

3. Déterminer les modes de l'ensemble des guéris

4. Ecrire la fonction de masse et la fonction de répartition des infectés en utilisant le dernier tableau.

5. Ecrire la fonction de masse et la fonction de répartition des décès en utilisant le dernier tableau.

6. Ecrire la fonction de masse et la fonction de répartition des guéris en utilisant le dernier tableau.

7. A quelle fonction analytique peut-on approcher chacune des 3 fonctions de masse. Selon les 3 formes des nuages de points empiriques, peut-on utiliser l'ajustement pour déterminer chacune des 3 fonctions de densité ?

Contenu du cours

Partie 1. Introduction à la statistique descriptive

Partie 1. Introduction à la statistique descriptive

I. Cours

1. Les tableaux statistiques

L'objet des statistiques est d'étudier des caractères ou variables sur des individus. La récolte initiale des données conduit à un tableau brut, sur support papier.

1.1. Le cas d'une seule variable

Le tableau brut se présente sous la forme suivante:

Individu	variable
1	x_1
2	x_2
.	.
.	.
.	.
n	x_n

Le nombre d'individus observé étant en général important, le tableau précédent ne permet pas d'analyser l'information obtenue. Il est donc nécessaire de créer un tableau plus synthétique où les observations identiques (possédant la même modalité) ont été regroupées.

modalité	effectif
C_1	n_1
C_2	n_2
.	.
.	.
.	.
C_k	n_k

La série sera plus lisible si on note pour chaque valeur du caractère le nombre de personnes présentant ce caractère : on obtient une **série statistique avec effectifs**.

"taille"	"effectif"
144	1
147	1
148	1
151	1
153	2
154	3
155	3
156	2
159	2
161	1
162	1
165	1
171	1

Une présentation de ce type s'impose quand la population est grande.
On peut aussi, puisque le caractère n'est discret que par convention, utiliser des **classes** par exemple d'**étendue** 1 cm ou 2 cm pour avoir une représentation plus globale.
On a alors en utilisant des classes d'étendue 2cm :

"taille x"	"effectif"
$143.5 \leq x < 145.5$	1
$145.5 \leq x < 147.5$	1
$147.5 \leq x < 149.5$	1
$149.5 \leq x < 151.5$	1
$151.5 \leq x < 153.5$	2
$153.5 \leq x < 155.5$	6
$155.5 \leq x < 157.5$	2
$157.5 \leq x < 159.5$	2
$159.5 \leq x < 161.5$	1

$161.5 \leq x < 163.5$	1
$163.5 \leq x < 165.5$	1
$165.5 \leq x < 167.5$	0
$167.5 \leq x < 169.5$	0
$169.5 \leq x < 171.5$	1

Ainsi, la fréquence de 155 est 3/20,
et la fréquence cumulée de 155 est :(1+1+1+1+2+3+3)/20=12/20. **Exercice** Le but de l'activité est l'étude de la taille (en cm) portant sur 250 individus jouant au basket.

Taille	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187
Effectif	4	8	7	18	23	22	24	32	26	25	18	19	10	8	6

On commence par un calcul des paramètres statistiques puis on poursuit avec des représentations graphiques : diagrammes en bâtons, diagramme en boîte et polygone des fréquences cumulées croissantes.

Pour une variable qualitative, les modalités ne sont pas mesurables.

Pour une variable quantitative, les modalités sont mesurables. Ce sont

- des valeurs numériques ponctuelles lorsque la variable est discrète
- des intervalles lorsque la variable est continue ou lorsque la variable est discrète et qu'elle comporte beaucoup de modalités.

1.2. Le cas de deux variables

Le tableau brut se présente sous la forme suivante:

Individu	variable1	variable2
1	x_1	y_1
2	x_2	y_2
.	.	.
.	.	.
.	.	.
n	x_n	y_n

On désire créer un tableau appelé tableau de contingence donnant le nombre d'individus possédant simultanément la modalité i de variable1 et la modalité j de variable2 qui se présentera sous la forme suivante:

		Variable2
		D_1 . . D_j . . D_r
variable1	C_1	n_{11} . . n_{1j} . . n_{1r}
	.	. . .
	.	. . .
	C_i	n_{i1} . . n_{ij} . . n_{ir}
	.	. . .
	.	. . .
	C_k	n_{k1} . . n_{kj} . . n_{kr}

2. Les graphique

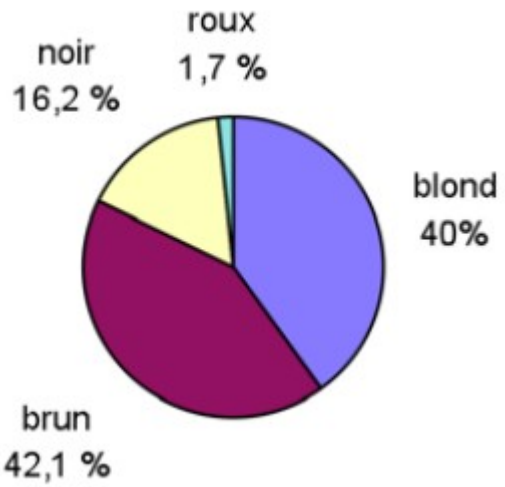
Après avoir obtenu un échantillon ou dénombré une population, on dispose le plus souvent de données numériques brutes présentées sous la forme d'une série de valeurs (dans le cas d'une VA quantitative) ou sous la forme d'un tableau donnant le nombre d'individu présentant un caractère qualitatif. Présentées ainsi, ces données sont rarement « parlantes » et il est nécessaire de dresser une représentation graphique afin de faire ressortir une partie de l'information. Suivant le type de variable aléatoire, le mode de représentation graphique va être différent.

2.1. Variable qualitative

Le tableau suivant donne les On dispose alors de 4 classes ou nombre de personnes (effectif absolu) présentant une couleur donnée dans un échantillon. Ce tableau peut être représenté tel quel (en nombre) ou en pourcentage sur une graphique à secteur.

Couleur de cheveux	nombre de sujets présentant cette couleur	% de sujets
blond	2365	40
brun	2487	42,1
noir	954	16,2
roux	98	1,7
total	5904	100

Couleurs de cheveux d'un groupe de personne



Représentation graphique par secteur

2.2. Variable quantitative

2.2.1. Variable discrète

Ce type de variable est associée généralement à un diagramme en bâtons où l'axe horizontal des abscisses porte les valeurs prises par la VA (xi) tandis que l'axe vertical des ordonnées porte l'effectif absolu (ni) observé.

Exemple :

Si l'on s'intéresse au nombre de personnes à bord d'une voiture dans 2 villes différentes, on peut dresser le tableau suivant :

Nombre de personnes à bord d'une voiture	Nombre de voitures (effectif absolu) Ville A	Nombre de voitures (effectif absolu) Ville B
1	37	60
2	95	132
3	134	123
4	79	60
5	28	15
6	27	10
	400	400

Nombre de personnes à bord d'une voiture dans 2 villes différentes

Le diagramme en bâtons correspondant est le suivant :

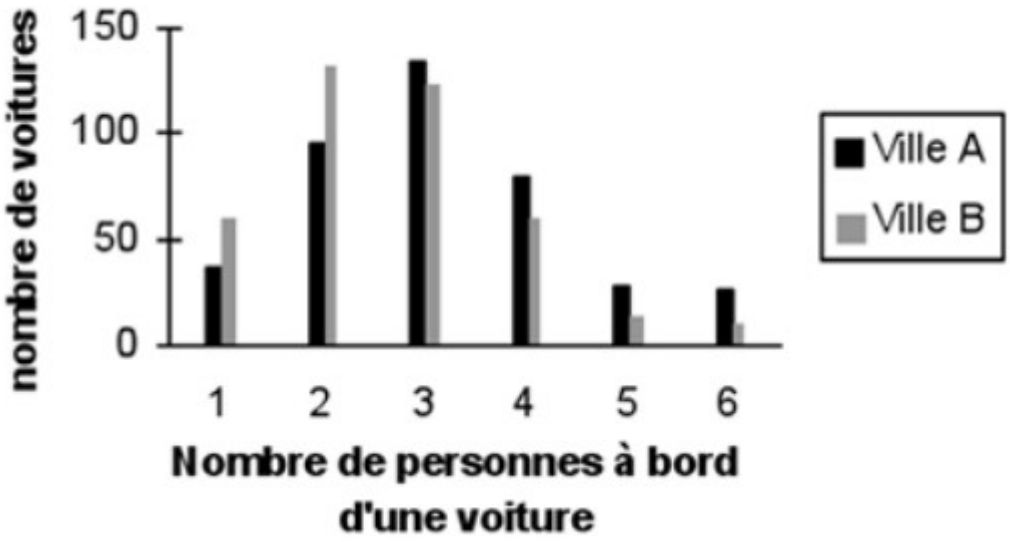


Diagramme en bâton

Ce type de représentation permet de mieux visualiser la distribution observée et semble indiquer que l'occupation des véhicules est plus importante dans la ville A que dans la ville B. Mais on ne peut faire confiance à cette affirmation simplement à la vue d'un graphique ; il faudrait une analyse statistique plus approfondie...

Un autre mode de représentation est le diagramme des fréquences cumulées.

Définition : Fréquence absolue

La fréquence absolue est le nombre de répétition d'une valeur numérique.

Exemple :

Dans l'exemple précédent, fréquence absolue et effectif se confondent. En ajoutant à chaque effectif (dans une classe donnée) l'effectif précédent, on obtient les effectifs absolus cumulés qui se représentent graphiquement de la façon suivante :

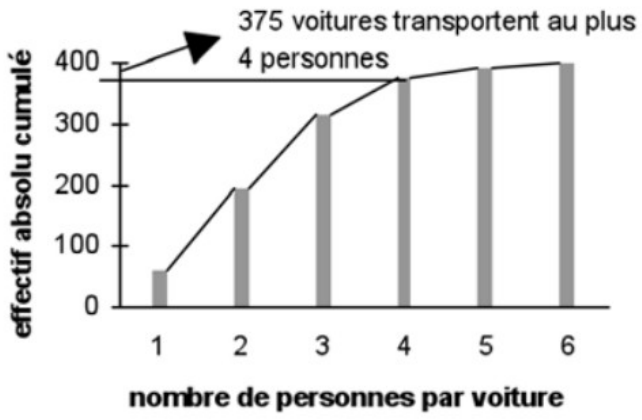


Diagramme des fréquences cumulées

Le diagramme des fréquences cumulées permet la lecture du nombre de voiture transportant par exemple au plus 4 personnes dans la ville B.

Il est souvent intéressant de tracer le diagramme des fréquences relatives cumulées. Dans ce cas, l'effectif est exprimé en pourcentage. La lecture du graphique devient alors indépendante de la taille de l'échantillon.

2.2.2. Variable continue

Dans le cas de ces variables, suivant la grandeur mesurée et la sensibilité de la méthode utilisée, il est fréquent d'obtenir autant de valeurs différentes que de données si bien que la représentation graphique n'a dans ces conditions aucun sens. On est donc généralement amené à regrouper les données en classes.

A la différence des VA discrètes, une classe donnée ne contient pas une seule valeur mais une infinité de valeurs possibles sur un intervalle défini (appelé intervalle de classe). Cet intervalle permet de définir également une amplitude de classe (différence entre les valeurs supérieure et inférieure de la classe). La valeur centrale de la classe est appelée centre de classe.

La répartition des données brutes en classes nécessite donc de la part du statisticien de faire un choix sur le nombre de classes et donc sur l'amplitude. Ce choix doit être suffisamment judicieux pour permettre la représentation graphique des données sans perdre pour autant trop d'information initialement contenue dans la série statistique.

Définition : Histogramme

C'est un ensemble de rectangles accolés ayant les caractéristiques suivantes :

1. La base de chaque rectangle correspond à l'amplitude d'une classe. Généralement toutes les classes d'une série statistique ont même amplitude.

2. La hauteur du rectangle est égale soit à l'effectif absolu (ou fréquence absolue) de la classe, soit à la fréquence relative (correspondant au rapport ni/n où ni est l'effectif absolu de la classe i et n l'effectif total de la série statistique). La surface de chacun des rectangles, si l'amplitude de classe est constante est alors proportionnelle à l'effectif de la classe

3. Il peut être intéressant de tracer l'histogramme des densités de fréquences. La densité de fréquence fxi correspond au rapport , où Δxi est l'amplitude de classe (ou base du rectangle). La surface d'un rectangle (fxi. Δxi) dans cette représentation graphique est alors égale à la fréquence relative de la classe correspondante et la surface totale des rectangle est égale à 1 quelle que soit la distribution initiale On verra par la suite que l'on peut généraliser ce résultat à la distribution d'une variable continue.

Exemple :

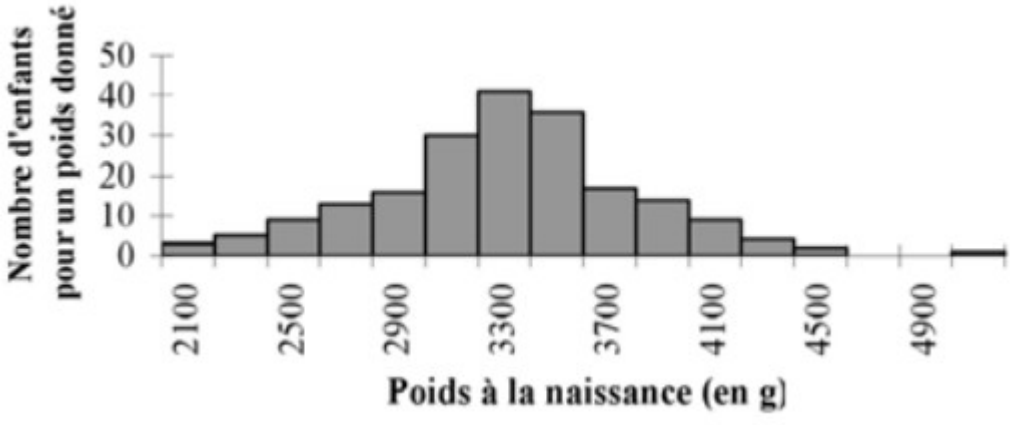
L’exemple suivant montre la distribution des poids de naissance de nouveau-nés dans une maternité (données extraites de : D. Schwartz, méthodes statistiques à l’usage des médecins et des biologistes, Médecine-Sciences, Flammarion 4ème ed.)

Extrèmité de classe (Poids en g)	centre de classe (Poids en g)	Effectifs n (nombre d'enfants)	fréquence relative	densité de fréquence	fréquence cumulée
2200	2100	3	0,02	7,50E-005	0,02
2400	2300	5	0,03	1,25E-004	0,04
2600	2500	9	0,05	2,25E-004	0,09
2800	2700	13	0,07	3,25E-004	0,15
3000	2900	16	0,08	4,00E-004	0,23
3200	3100	30	0,15	7,50E-004	0,38
3400	3300	41	0,21	1,03E-003	0,59
3600	3500	36	0,18	9,00E-004	0,77
3800	3700	17	0,09	4,25E-004	0,85
4000	3900	14	0,07	3,50E-004	0,92
4200	4100	9	0,05	2,25E-004	0,97
4400	4300	4	0,02	1,00E-004	0,99
4600	4500	2	0,01	5,00E-005	1
4800	4700	0	0	0,00E+000	1
5000	4900	0	0	0,00E+000	1
5200	5100	1	0,01	2,50E-005	1
Total		**200**	**1**	**0,01**	

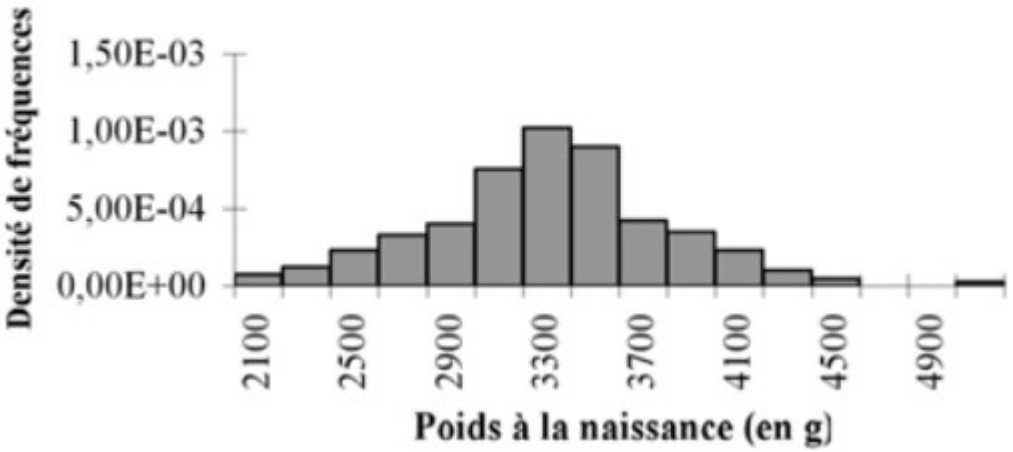
Distribution des poids de naissance

Dans cet exemple, toutes les classes ont même amplitude (200 g) et il y a au total 16 classes ce qui est un nombre suffisant pour représenter la distribution initiale (200 valeurs). On remarquera qu’une partie de l’information initiale a été perdue puisqu’à la vue du tableau il n’est plus possible de différencier les nouveau-nés d’une même classe.

A partir de ce tableau de valeurs, on peut tracer les histogrammes des effectifs absolus, des fréquences relatives et des densités de fréquence.



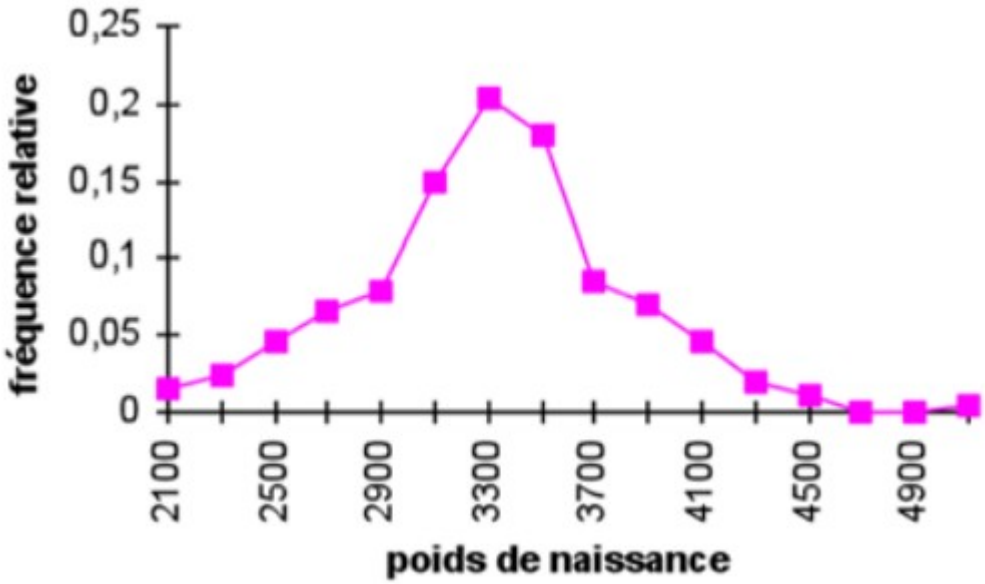
Distribution des poids à la naissance



Distribution des poids à la naissance

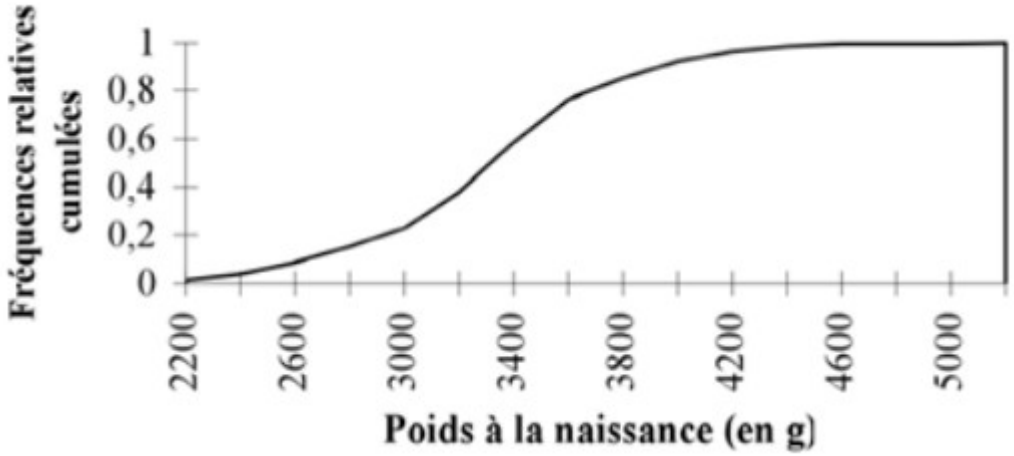
Définition : Polygone des fréquences

Le polygone des fréquences est représenté en joignant les milieux des cotés supérieurs des rectangles dans un histogramme. C'est une ligne brisée dont les extrémités rejoignent l'axe des abscisses.



Polygone des fréquences relatives

Définition : courbe des fréquences relatives cumulées



Courbe des fréquences relatives

Ce type de courbe permet une lecture rapide du pourcentage de nouveau-nés dont le poids est compris entre deux valeurs. Il suffit de faire la différence entre les 2 ordonnées correspondant à l'intervalle de poids fixé.

3. Les caractéristiques de tendances centrales

Synthétiser l'information contenue dans un tableau par un graphique est la première étape réalisée en statistique. Par la suite, on cherche à synthétiser encore plus l'information en la réduisant à une seule valeur numérique. Les caractéristiques de tendance centrale essayent de donner la valeur la plus représentative d'un ensemble de valeurs numériques.

Remarque: les paramètres définis par la suite n'ont de sens que pour les variables quantitatives.

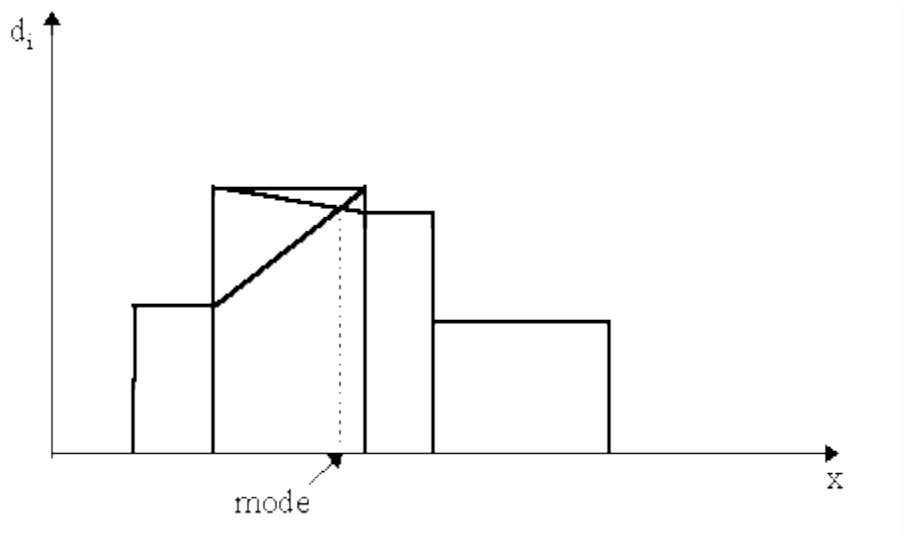
3.1. Le mode

C'est la valeur observée d'effectif maximum.

Variable discrète: classer les données par ordre croissant. Celle d'effectif maximum donne le mode.
Il est fortement conseillé d'utiliser le diagramme en bâtons pour déterminer le mode. En effet, deux valeurs consécutives xi , xi+1 peuvent avoir le même effectif maximum; on parlera d'intervalle modal [xi , xi+1]. Il peut aussi y avoir un mélange de deux populations qui conduit à un diagramme en bâtons où apparaissent deux bosses; on considérera deux modes. Il est déconseillé, sauf raison explicite, d'envisager plus de deux modes.

Variable classée: la classe modale correspond à la classe ayant l'effectif maximum. Il est fortement conseillé d'utiliser l'histogramme pour déterminer le mode. Comme pour le cas discret, on peut avoir deux classes modales. Toutes les valeurs de la classe pouvant à priori se réaliser, on ne se contentera pas de déterminer la classe modale. Une des valeurs de cette classe sera le mode. Certains auteurs préconisent par

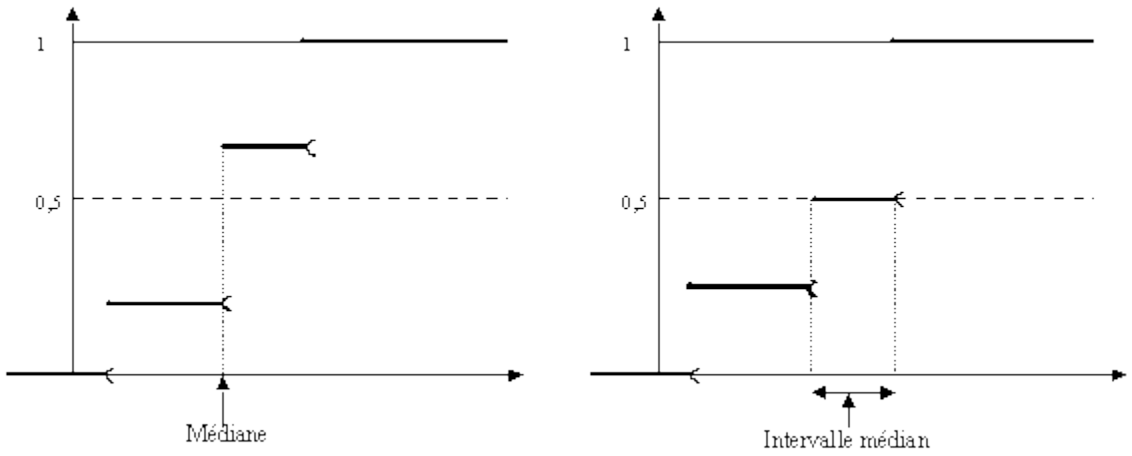
simplicité de prendre le centre de la classe modale. Il est préférable cependant de tenir compte des classes adjacentes de la manière suivante:



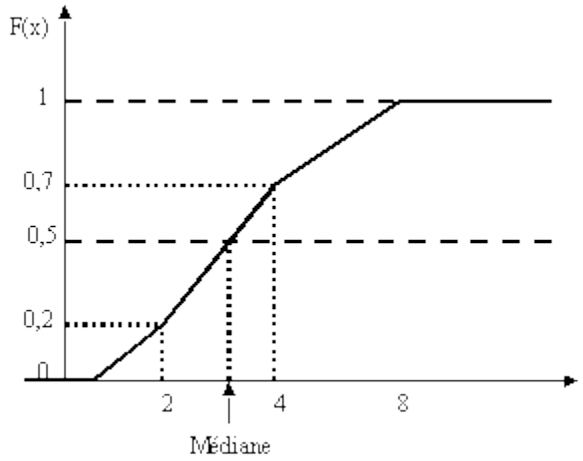
3.2. La médiane

Les valeurs étant rangées par ordre croissant, c'est la valeur de la variable qui sépare les observations en deux groupes d'effectifs égaux.

Variable discrète: la détermination peut s'obtenir à partir du tableau statistique en recherchant la valeur de la variable correspondant à une fonction cumulée égale à n/2 (effectif cumulé) ou ½ (fréquence cumulée). Il est encore plus facile de lire sur les graphiques cumulatifs les abscisses des points d'ordonnée n/2 (effectif cumulé) ou ½ (fréquence cumulée). Si tout un intervalle a pour image n/2 (½ pour la fréquence), on parlera d'intervalle médian (on peut prendre le milieu de l'intervalle comme médiane)



Variable classée: l'abscisse du point d'ordonnée n/2 (½ pour la fréquence)se situe en général à l'intérieur d'une classe. Pour obtenir une valeur plus précise de la médiane, on procède à une interpolation linéaire. La valeur de la médiane peut être lue sur le graphique ou calculée analytiquement.



$$\frac{Mé - 2}{4 - 2} = \frac{0,5 - 0,2}{0,7 - 0,2}$$

D'où la valeur de la médiane.

De manière générale, si a et b sont les bornes de la classe contenant la médiane, F(a) et F(b) les valeurs de la fréquence cumulée croissante en a et b, alors

$$Mé = a + (b - a) \times \frac{0,5 - F(a)}{F(b) - F(a)}$$

3.3. La moyenne arithmétique

Si xi sont les observations d'une variable discrète ou les centres de classe d'une variable classée, la moyenne arithmétique $\bar{\bar{x}}$ est égale à $\sum_{i=1}^{k} \frac{n_i x_i}{n} = \sum_{i=1}^{k} f_i x_i$

La moyenne arithmétique est un paramètre de tendance centrale plus utilisé que les autres de par ses propriétés algébriques:

a) Pour plusieurs populations d'effectifs n1, n2, ..., nk, de moyennes respectives $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$

Moyenne globale = moyenne des moyennes

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{n}$$

b) La moyenne arithmétique conserve les changements d'échelle et d'origine

$$x: (x_i, n_i) \rightarrow y: (y_i = ax_i + b, n_i)$$
$$\bar{x} \rightarrow \bar{y} = a\bar{x} + b$$

3.4. La moyenne géométrique

Si xi sont les observations d'une variable quantitative, la moyenne géométrique est égale à

$$G = \sqrt[n]{x_1^{n_1} \times \ldots \times x_k^{n_k}}$$

Ce type de moyenne est surtout utilisé pour calculer des pourcentages moyens.
r étant un taux d'accroissement, 1+r est appelé coefficient multiplicateur; et le coefficient multiplicateur moyen est alors égal à la moyenne géométrique des coefficients multiplicateurs.

3.5. La moyenne harmonique

Si xi sont les observations d'une variable quantitative, la moyenne harmonique est égale à

$$H = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$$

Il n'est pas évident d'utiliser ce type de moyenne.
Elle intervient lorsqu'on demande une moyenne de valeurs se présentant sous forme de quotient de deux variables x/y (km/h, km/litre,...). Attention, il faut cependant bien décortiquer le problème car il peut aussi s'agir d'une moyenne arithmétique.

3.7. Les quantiles

Ce sont des caractéristiques de position.

Il y a 1 médiane Mé qui sépare les observations en 2 groupes d'effectifs égaux
3 quartiles Q1, Q2, Q3 qui séparent les observations en 4 groupes d'effectifs égaux
9 déciles D1, D2, ..., D9 qui séparent les observations en 10 groupes d'effectifs égaux
99 centiles C1, C2, ..., C99 qui séparent les observations en 100 groupes d'effectifs égaux

La détermination de ces caractéristiques est identique à celle de la médiane.

Les quartiles sont obtenus lorsqu'on a cumulé 25, 50, 75% de la population
Les déciles sont obtenus lorsqu'on a cumulé 10, 20,...., 90% de la population
Les centiles sont obtenus lorsqu'on a cumulé 1, 2,...., 99% de la population

Remarque: la notion de déciles et de centiles n'a de sens que s'il y a beaucoup d'observations et donc essentiellement pour une variable classée.

4. Les caractéristiques de dispersion

Comme leur nom l'indique, ces caractéristiques essayent de synthétiser par une seule valeur numérique la dispersion de toutes les valeurs observées.

4.1. L'étendue

C'est la différence entre la plus grande et la plus petite observation

4.2. L'intervalle interquartile

C'est la différence entre le troisième et le premier quartile

4.3. La variance et l'écart-type

Si xi sont les observations d'une variable discrète ou les centres de classe d'une variable classée, la variance

V est égale à $\sum_{i=1}^{k} \frac{n_i (x_i - \bar{x})^2}{n} = \sum_{i=1}^{k} f_i (x_i - \bar{x})^2$

On a aussi $V = \sum_{i=1}^{k} \frac{n_i x_i^2}{n} - \bar{x}^2$

c.à.d. moyenne des carrés - carré de la moyenne

On utilise plus couramment l'écart-type qui est la racine carrée de la variance et qui a l'avantage d'être un nombre de même dimension que les données (contrairement à la variance qui en est le carré)
La variance est un paramètre de dispersion plus utilisé que les autres de par ses propriétés algébriques:

a) Pour plusieurs populations d'effectifs $n_1, n_2, \ldots, n_k$, de moyennes respectives $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$, de variances $V_1, V_2, \ldots, V_k$

Variance globale = variance des moyennes + moyenne des variances

$$V = \sum_{i=1}^{k} \frac{n_i}{n} (\bar{x}_i - \bar{\bar{x}})^2 + \sum_{i=1}^{k} \frac{n_i}{n} V_i$$

où $\bar{\bar{x}}$ représente la moyenne des moyennes

b) changement d'échelle et d'origine

$$x : (x_i , n_i) \rightarrow y : (y_i = ax_i + b , n_i)$$

$$V_x \rightarrow V_y = a^2 V_x$$

4.4. Le coefficient de variation

$$CV = \frac{\sigma}{\bar{x}}$$

C'est un coefficient qui permet de relativiser l'écart-type en fonction de la taille des valeurs. Il permet ainsi de comparer la dispersion de séries de mesures exprimées dans des unités différentes

5. Les caractéristiques de concentration

L'objectif est de mesurer les inégalités dans la répartition d'une variable à l'intérieur d'une population. Cette notion n'a d'intérêt que dans la mesure où les valeurs globales suivantes ont une signification concrète

5.1. Les valeurs globales

xi représentent les valeurs ponctuelles ou les centres des classes, ni les effectifs correspondants.
Les valeurs globales de la série (xi , ni) sont les quantités gi = ni xi

5.2. La médiale

La médiale de la série (xi , ni) est la médiane de la série (xi , gi)

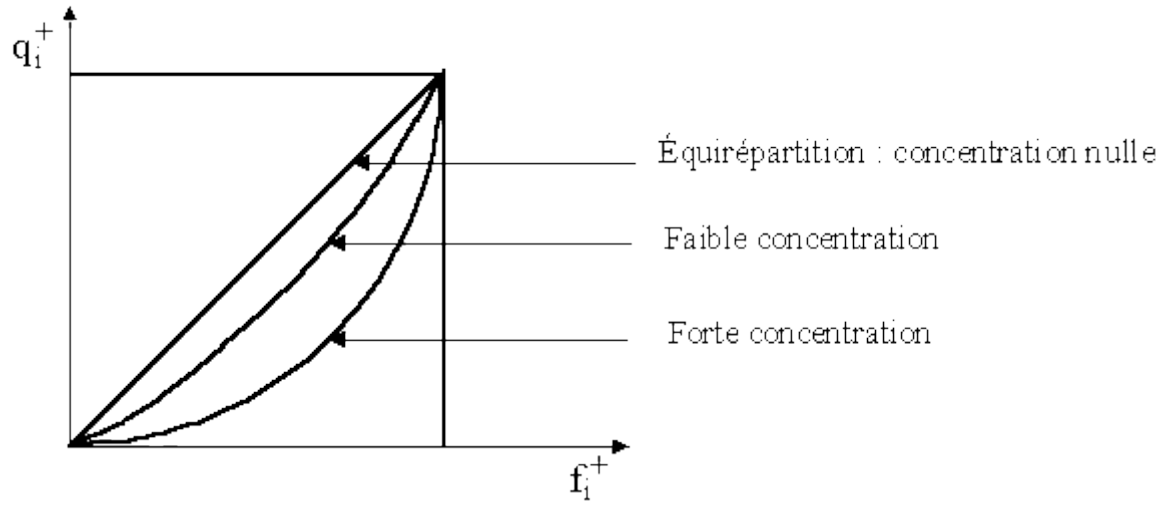
5.3. La courbe de Lorentz ou courbe de concentration

On obtient cette courbe en représentant :

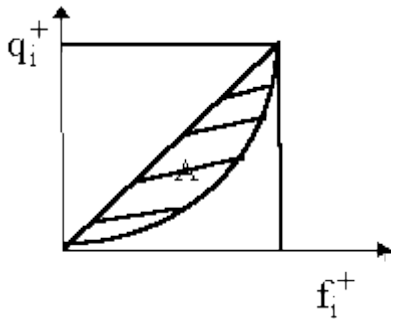
en abscisse, f_i^+ les fréquences cumulées croissantes de la série (x_i , n_i)

en ordonnée, q_i^+ les fréquences cumulées croissantes de la série (x_i , g_i)

L'allure de la courbe permet d'avoir une idée de la concentration



5.4. L'indice de Gini



$\gamma = 2\ A$

Propriétés:

- $0 < \gamma < 1$
- γ proche de 1 $\Rightarrow$ forte concentration
- γ proche de 0 $\Rightarrow$ faible concentration

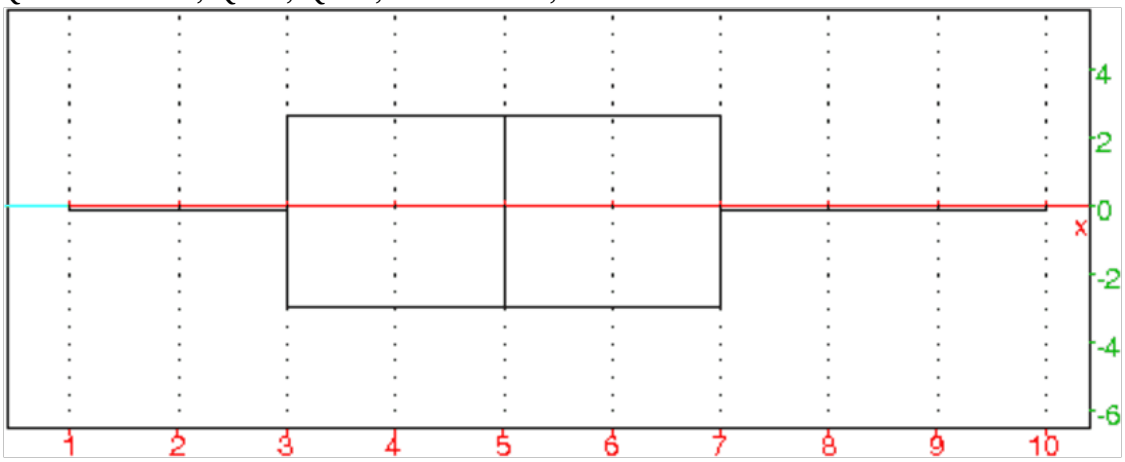
6. Vocabulaire des séries quantitatives à 1 variable

Soit une série quantitative à 1 variable L.
La différence entre la plus grande valeur et la plus petite valeur du caractère effectivement obtenue est l'**étendue** de la série L.
Le nombre de membres de la population étudiée est l'**effectif total**.
Si le caractère est discret, il est commode d'indiquer pour chaque valeur du caractère, le nombre des membres de la population ayant cette valeur : c'est l'**effectif de cette valeur**.
Si le caractère est continu, on partage l'intervalle sur lequel s'étendent ces valeurs en intervalles (en général égaux) que l'on appelle **classe**. Le nombre des membres de la population ayant leur valeur dans une classe est l'**effectif de cette classe**.
La valeur moyenne des bornes d'une classe est le **centre** de cette classe.
L'**effectif cumulé** d'une valeur (ou d'une classe) est la somme de l'effectif de cette valeur (ou de cette classe) et de tous les effectifs des valeurs (ou des classes) qui précèdent.
La **fréquence** d'une valeur (ou d'une classe) est le rapport de l'effectif de cette valeur (ou de cette classe) par l'effectif total.
Avec X cas on tape par exemple :
frequences([1,2,1,1,2,1,2,4,3,3])
On obtient ;
[[1,0.4],[2,0.3],[3,0.2],[4,0.1]]
La **fréquence cumulée** d'une liste de valeurs (ou d'une classe) est la somme de la fréquence de cette valeur (ou de cette classe) et de toutes les fréquences des valeurs (ou des classes) qui la précèdent.
Avec Xcas on tape par exemple :
frequences_cumulee([[0.75,30],[1.75,50],[2.75,20]])
ou
frequences_cumulee([[0.25..1.25,30],[1.25..2.25,50],
[2.25..3.25,20]])
On obtient le diagramme des fréquences cumulées.
L'**histogramme** des effectifs (resp fréquences) d'un caractère discret ou continu est le graphique qui permet de visualiser l'effectif (resp fréquences) des différentes valeurs du caractère : on met en abscisse les différentes valeurs du caractère (ou le centre des différentes classes), puis on forme des rectangles accolés deux à deux, ses rectangles ont deux cotés parralléles à l'axe des ordonnées, le coté porté par l'axe des abscisses a pour longueur l'amplitude de la classe, et l'autre est tel que l'aire du rectangle est égale à l'effectif (resp fréquences) de la valeur considérée.
L'**histogramme des fréquences** permet de visualiser les fréquences des différentes classes au moyen de la surface de rectangles : chaque rectangle correspond à une classe et a pour surface la fréquence de cette classe.
Avec Xcas on tape par exemple :
histogramme([[0.75,30],[1.75,50],[2.75,20]])
ou

histogramme([[0.25..1.25,30],[1.25..2.25,50],
[2.25..3.25,20]])
On obtient un histogramme des fréquences.
La **fonction de répartition des fréquences** est égale pour chaque valeur du caractère à la fréquence cumulée de cette valeur.
Le **mode** est la valeur du caractère dont l'effectif est le plus grand.
Le **maximum** est la plus grande valeur du caractère effectivement obtenue.
Le **minimum** est la plus petite valeur du caractère effectivement obtenue.
La **médiane** partage la série statistique en deux groupes de même effectif. C'est une valeur du caractère à partir de laquelle l'effectif des valeurs qui lui sont inférieures est superieur ou ègal à l'effectif des valeurs qui lui sont supérieures (par exemple la médiane de [140,145,146,147] est 146 et la médiane de [140,145,146] est 145). La médiane est donc la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.5.
Les **quartiles** sont trois valeurs du caractère qui partage la série statistique en quatre groupes de même effectif :
- le **1er quartile** est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.25.
- le **2ème quartile** est confondu avec la médiane.
- le **3ème quartile** est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.75.
On peut définir les **déciles**. Il y a 9 déciles :
le **1er décile** est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.1.
le **2ème décile** est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.2.
etc...
le **9ème décile** est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.9.
On peut aussi définir le **centile** (il y a 99 centiles) et le **quantile** d'ordre **p** :
le **1er centile** est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.01.
etc...
le **99ème centile** est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.99.
Le **quantile** d'ordre **p** (**p** un réel de [0,1[), est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse **p**.
Le **semi-interquartile** est égal à $1/2(Q_3-Q_1)$ où Q_1 et Q_3 désigne le premier et le troisième quartile. Cet indice fournit un renseignement sur l'étalement des valeurs de part et d'autre de la médiane.
L'**interquartile** est égal à Q_3-Q_1 où Q_1 et Q_3 désigne le premier et le troisième quartile. Cet indice fournit un renseignement sur l'étalement des valeurs de part et d'autre de la médiane.
L'**interdécile** est égal à D_9-D_1 où D_1 et D_9 désigne le premier et le neuvième décile. Cet indice fournit un renseignement sur l'étalement des valeurs de part et d'autre de la médiane.

Exemples avec X cas

On tape :
L:=[1,2,3,4,5,6,7,8,9,10]
min(L) et on obtient 1
quartile1(L) et on obtient 3.0
median(L) et on obtient 5.0
quartile3(L) et on obtient 8.0
max(L) et on obtient 10
quartiles(L) pour avoir le résultat des 5 commandes précédentes et on obtient [[1.0],[3.0],[5.0],[8.0],[10.0]]
quantile(L,0.9) et on obtient 9.0
La **boite à moustaches** permet de visualiser ces différentes valeurs :
c'est un rectangle dont un coté est un trait allant de Q_1 à Q_3 sur lequelle un trait vertical indique la valeur de la médiane et d'où deux traits horizontaux (les moustaches) débordent : l'un va de la valeur minimum à Q_1 et l'autre de Q_3 à la valeur maximum. Sur ces deux moustaches, on trouvent quelquefois deux traits verticaux indiquant la valeur du premier et du neuvième décile.
Avec Xcas on tape :
L:=[1,2,3,4,5,6,7,8,9,10]
moustache(L)
Cela ouvre le graphique et dessine une boite à moustaches où on peut lire que :
Q2=médiane=5, Q1=3, Q3=8, minimum=1, maximum=10.



La **moyenne** est le quotient de la somme des valeurs du caractère (pas toujours distinctes) par l'effectif total. Si le caractère prend n valeurs distinctes x_k d'effectifs e_k pour k=0...(n−1) alors l'effectif total vaut $N=\sum_{k=0}^{n-1} e_k$ et la moyenne m est : $m=1/N\sum_{k=0}^{n-1} e_k x_k$.
La **variance** est la moyenne des carrés des écarts à la moyenne des valeurs du caractère. Si le caractère prend n valeurs distinctes x_k d'effectifs e_k (k=0...(n−1)), si la moyenne vaut m et, si l'effectif total vaut N alors la variance $v=s^2$ est : $s^2=1/N\sum_{k=0}^{n-1} e_k(x_k-m)^2=1/N(\sum_{k=0}^{n-1} e_k x_k^2)-m^2$.
L'**écart-type** s est la racine carrée de la variance.
Soit une série statistique quantitative d'effectif N à 1 variable, un **échantillon d'ordre n** désigne le système des n valeurs prises par le caractère au cours de n tirages indépendants. Les valeurs prises par l'échantillon sont donc les valeurs prises par n variables aléatoires $X_1,...,X_n$ qui suivent la même loi que la variable aléatoire X égale à la valeur du caractère étudié. Par exemple, si dans une ville de N habitants, on étudie la taille (exprimée en centimètres) de ses habitants, la taille de 100 personnes prises au hasard dans cette ville est un échantillon d'ordre 100. En général, on ignore la loi

de la variable aléatoire égale à la taille des habitants de cette ville, et on veut dégager un certain nombre d'éléments caractéristiques de cette variable grâce à l'échantillon.

7. Les erreurs de mesure

7.1. Les erreurs aléatoires

Elle inclut les causes d'erreur dues aux :

- seuil de mesure (plus petite valeur mesurable),
- résolution (plus petite variation mesurable),
- hystérésis,
- parasites,
- influences du milieu sur le capteur : Par exemple celles provoquées par les variations de température sur un capteur de pression.

Lors de mesures répétées nous obtenons généralement une dispersion des résultats ; si les erreurs de mesure sont aléatoires un traitement statistique permet de connaître la valeur la plus probable de la grandeur mesurée et de fixer les limites de l'incertitude.

Lorsque la mesure d'une même valeur a été répétée n fois en donnant les résultats : Ml, M2, ... Mn, la valeur moyenne M est par définition :

$$\overline{M} = \sum_{i=1}^{n} \frac{M_i}{n}$$

L'erreur aléatoire E_a est la différence entre le résultat d'un mesurage M_i et cette moyenne $\overline{M}$ lorsque n tend vers l'infini et que les mesures sont obtenues dans des conditions de répétabilité :

$$E_a = M_i - \overline{M}$$

Lorsque les erreurs accidentelles sur les différentes mesures sont indépendantes, la probabilité d'apparition de différents résultats satisfait habituellement la loi de Gauss.

Lorsque cette loi est satisfaite, la probabilité $P(_{Ma}, M_b)$ d'obtenir comme résultat d'une mesure une valeur comprise entre deux valeurs Ma et Mb peut s'écrire :

$$P(M_a, M_b) = \int_{M_a}^{M_b} P(M)\,dM$$

où P(M) est la densité de probabilité d'obtenir la valeur M.

Dans le cas de la loi de Gauss cela donne :

$$P(M) = \frac{\exp(-\frac{(M-\overline{M})^2}{2\sigma^2})}{\sigma\sqrt{2\pi}}$$

Une indication de la dispersion de ces résultats est donnée par l'écart type σ :

$$\sigma^2 = \int_{-\infty}^{+\infty} (M-\overline{M})^2 \, P(M) \, dM$$

- la valeur de M la plus probable est $\overline{M}$,

- la probabilité d'apparition d'un résultat de la mesure dans les limites indiquées est :

$$P(\overline{M} \pm \sigma) = 68.28\%$$

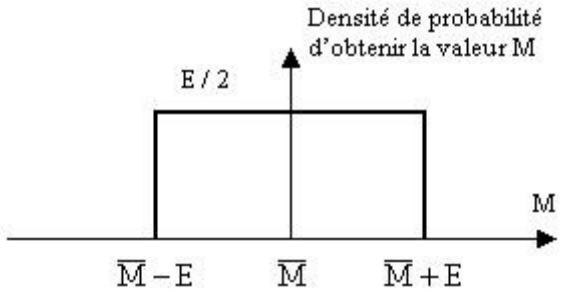
$$P(\overline{M} \pm 2\sigma) = 95.45\%$$

$$P(\overline{M} \pm 3\sigma) = 99.73\%$$

Lorsque nous disposons d'un nombre « n » important de mesure, σ : ⌈μιτσε ερτ⌋ τυεπ

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (M_i - \overline{M})^2}{n-1}}$$

Il est fréquent que le fabriquant d'un capteur donne sa précision sans donner la loi de distribution des erreurs. Par exemple il indique que son capteur fourni la valeur à mesurer (le mesurande M) à ± .E ρερ⌈δισνοχ τε ελβαροϖαφ⌈δ συλπ ελ σαχ ελ σναδ ρεχαλπ εσ τυαφ λι ,σαχ εχ σναΔ]E+ ; E-[ελλαωρετνιϑλ σναδ εμροφινυ τσε ⌈τιλιβαβορπ εδ ⌈τισνεδ αλ ευθ .συοσσεδ-ιχ ⌈τνεσ⌋ρπερ εμμοχ



Dans ce cas nous avons :

Probabilité { $-E \leq M \leq +E$ } = 1 = aire du rectangle. Donc :

$$\sigma^2 = \int_{-E}^{+E} \frac{1}{2E} M \, dM$$

Soit, après intégration :

$$\sigma^2 = \frac{E^2}{3} \qquad \text{d'où} \qquad \sigma = \frac{E}{\sqrt{3}}$$

7.2. L'erreur systématique

L'erreur systématique se superpose aux erreurs aléatoires. Elle est provoquée par un mauvais réglage ou un mauvais étalonnage. Elle peut être également induite par la présence du capteur qui modifie la valeur du mesurande. Elle devient importante dans le cas ou les instruments sont mal utilisés.

L'erreur systématique E_s est la différence entre la moyenne $\overline{M}$ lorsque n tend vers l'infini et que les mesures sont obtenues dans des conditions de répétabilité et une valeur **vraie** du mesurande M_0:

$$E_s = \overline{M} - M_0$$

Cette définition utilise sciemment l'expression « une valeur **vraie** du mesurande » et non « une **vraie** valeur du mesurande » puisque la valeur vraie du mesurande n'est pas connue (sauf si l'on considère que le mesurande est un étalon primaire du système SI).

7.3. Clarifications

7.3.1. La fidélité

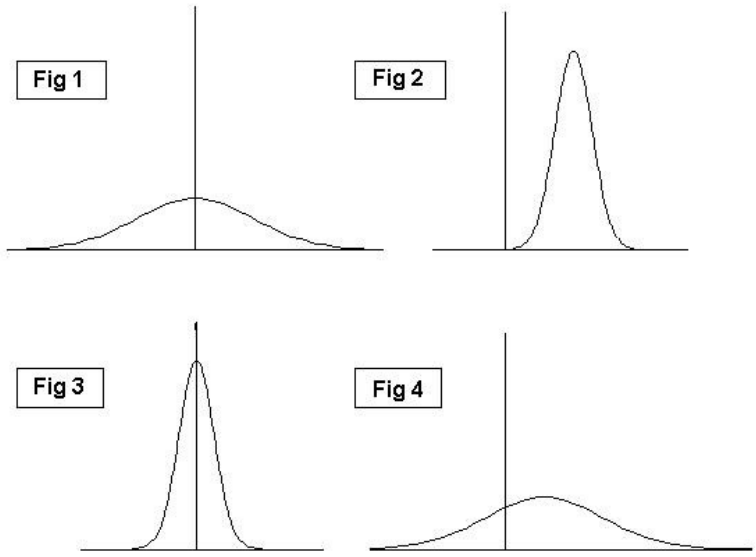
La fidélité d'un capteur est sa propriété à redonner des valeurs très proches lorsque, après avoir varié, la grandeur reprend sa valeur initiale. Ce qui se traduit par des résultats groupés autour de leur valeur moyenne. L'écart type dont l'importance reflète la dispersion des résultats est souvent considéré comme l'erreur de fidélité : I1 permet ainsi une appréciation quantitative de la fidélité. La valeur la plus probable, telle qu'elle résulte d'un ensemble de mesures, peut être connue avec une faible marge d'incertitude tout en étant assez éloignée de la valeur vraie.

7.3.2. La justesse

La justesse est la qualité d'un appareillage de mesure à donner une valeur (moyenne) très proche de la valeur vraie.

7.3.3. La précision

La précision qualifie 1'aptitude du capteur à fournir des données qui, prisent individuellement, sont proches de la valeur vraie. Un capteur précis est donc à la fois fidèle et juste.



Différents types de répartition des résultats de mesure :

Fig.1 - Capteur juste mais non fidèle : Les erreurs systématiques sont réduites mais les erreurs aléatoires sont importantes.

Fig.2 - Capteur fidèle mais non juste : Les erreurs systématiques sont importantes mais les erreurs aléatoires sont faibles.

Fig.3 - Capteur juste et fidèle donc précis : Les erreurs systématiques et aléatoires sont faibles.

Fig.4 - Capteur ni juste, ni fidèle : Les erreurs systématiques et aléatoires sont importantes.

Partie 2. Exercices

Il y a une multitude d'exercices corrigés à étudier. Pour les voir, il faut aller sur Google et taper :

1. Statistiques descriptives – Résumé et exercices. Université de Paris 8.

2. IUT – GB- 1er année. Statistique descriptive. CNRS – France.

Partie 2. Introduction à la probabilité

I. Cours

1. Fondements de la probabilité

1.1. Expérience aléatoire

1.1.1. Définition :

Une expérience est dite aléatoire si :

• elle conduit à des issus possibles qu'on peut nommer

• on ne sait pas laquelle de ces issues va se réaliser

1.1.2. Exemples :

Lancer un dé non pipé dont les faces sont numérotées de 1 à 6 et noter le résultat obtenu constitue une expérience aléatoire.

Par contre, lancer un dé dont les six faces sont numérotées 3 par exemple n'est plus une expérience aléatoire car ici, on connaît à l'avance le résultat.

1.2. Evénements

1.2.1. Introduction :

Lors d'une expérience aléatoire, certains faits peuvent se réaliser ou pas : ce sont des événements. Nous allons nous y intéresser de plus près en commençant par donner un exemple concret puis une définition mathématique.

1.2.2. Exemple :

Si nous reprenons l'expérience aléatoire consistant à jeter un dé et à noter le résultat obtenu, il est possible d'obtenir un nombre impair ou pas.

Notons A : « Obtenir un nombre impair » cet événement.

L'événement A est réalisé si le résultat obtenu est soit 1 soit 3 soit 5.

Les résultats 1 ; 3 et 5 sont appelés les issues ou les éventualités.

Nous confondrons alors l'événement et l'ensemble des issues possibles à la réalisation de cet événement. Ici, A = {1;3;5}

1.2.3. Définition :

• un événement est une partir (ou sous-ensemble) de l'univers

• on dit que cet événement est réalisé si l'issue de l'expérience est l'une des éventualités qui le compose

1.3. Evénements particuliers

1.3.1. Définition :

• Un événement qui est toujours réalisé est dit certain. Il est donc représenté par l'Univers

• Un événement qui n'est jamais réalisé est dit impossible et noté par l'ensemble vide .

• un événement constitué d'une unique issue est appelé événement élémentaire et noté par un singleton {a}.

Exemples :

• L'événement A " Obtenir un nombre entier inférieur à 7" est A = {1;2;3;4;5;6;} . C'est un événement certain.

• Les 6 événements élémentaires sont {1}, {2}, {3}, {4}, {5} et {6}.

• L’événement B : « Obtenir un nombre impair » est B = {1 ; 3 ; 5}. Il est composé de trois issues.

• L’événement C : « Obtenir 8 » est un événement impossible. C = 0

1.4. Système complet

1.4.1. Définition :

Soit $A_1,A_2,..,An$ une suite au moins dénombrables d'événements sur un univers. On dit que $A_1,A_2,..,An$ forment un système complet d'événements si $A_1,A_2,..,An$ constituent une partition de Ω

1.5. Univers des possibles

Ω représente l'univers des possibles, ou espace fondamental. Cette notion est enseignée au lycée. C'est l'ensemble de tous les évènements qu'il est possible d'obtenir avec une « épreuve » aléatoire.

Si les éléments sont qualitatifs ou discrets, ils sont indiqués entre accolades (NB : à ce stade, il est prématuré de parler de variable aléatoire) . Ainsi, l'univers associé au lancer d'un dé s'écrit Ω = {1 ; 2 ; 3 ; 4 ; 5 ; 6}. Celui qui est associé à un lancer de deux dés s'écrit Ω = {(1,1) ; (1,2) ; …} ou {1 ; 2 ; 3 ; 4 ; 5 ; 6}² à moins d'être présenté sous forme de tableau. Il comporte 36 éléments. 36 est le cardinal de Ω, qui est en l'occurrence le produit cartésien de deux espaces fondamentaux. Dans les exercices de probabilités, il est habituel que le cardinal d'un univers des possibles soit déterminé par une méthode de dénombrement.

Lorsque l'univers des possibles est présenté de façon générale, les évènements élémentaires sont habituellement indiqués par des oméga minuscules : Ω = {*ω1* ; *ω2* ; … ; *ωn*}.

Si la variable est continue, l'univers est présenté sous la forme d'un intervalle.

L'espace fondamental peut aussi être infini.

1.6. Notion de tribu

1.6.1. Définition :

Une tribu sur Ω n'est pas un sous-ensemble de Ω , mais un sous-ensemble de P(oméga) - C'est à dire un ensemble de parties de oméga. Et parmi ces parties, il y a obligatoirement l'ensemble vide et oméga lui-même.

1.6.2. Exemple

Aussi, quand on dit Ω ={1,2,3}, la tribu engendrée par {1,2} dans Ω est A={{}, {1,2,3},{1,2},{3}}, alors que P(Ω)={{}, {1},{2},{3}, {1,2}, {1,3}, {2,3}, {1,2,3}}. A est un sous-ensemble de P(Ω). Les éléments de Ω sont 1,2 et 3, ceux de A sont {}, {1,2,3}, {1,2} et {3} - les éléments de Ω sont des nombres, alors que les éléments de A sont des ensembles de nombres -

1.7. Espace probabilisable

1.7.1. Définition :

Un espace probabilisable est formalisé par un couple (Ω ; *A*). *A* est une tribu de parties de Ω. Derrière ce terme étrange se cache une notion sur laquelle nous allons nous attarder un moment.

Les éléments d'une tribu, appelés évènements, sont ceux auxquels on peut attribuer une probabilité. L'évènement certain est Ω.

De façon plus formelle, on dit que l'ensemble *A* non vide est stable par complémentaire et stable par union dénombrable.

« Stable par complémentaire » signifie que si un évènement fait partie d'une tribu, son complémentaire en fait partie également. La somme du cardinal d'un élément de la tribu et du cardinal de son complémentaire est égal à card(Ω).

« Stable par union dénombrable » signifie que la réunion de tous les éléments de la tribu est incluse elle aussi à cette tribu, même si ce nombre d'éléments est infini (mais néanmoins dénombrable).

1.7.2. Exemple

Ainsi, pour un élément quelconque *B* appartenant à Ω, une tribu de parties de Ω s'écrit :

$$\{\emptyset\ ;\ \boldsymbol{B}\ ;\ \overline{\boldsymbol{B}}\ ;\ \boldsymbol{\Omega}\}$$

Donc Ω, l'ensemble vide, tout élément de Ω ainsi que son complémentaire appartiennent à *A*.

Il s'ensuit une stabilité par intersection (là encore, si les éléments sont dénombrables).

Si l'univers des possibles est l'ensemble des réels, la tribu engendrée par des intervalles fermés et bornés est connue sous le nom de tribu des boréliens.

1.8. Espace probabilisé

1.8.1. Définition

Non seulement on a défini un espace probabilisable, mais si l'on peut aussi définir dans celui-ci une application *P* appelée loi de probabilité de *A* dans [0 ; 1], bingo ! Nous avons maintenant un espace probabilisé de toute beauté. Celui-ci s'écrit avec un triplet (Ω ; *A* ; *P*).

1.8.2. Exemple

Si l'on étudie un jeu de hasard, l'espace probabilisé est muni de la probabilité uniforme.

Les deux axiomes des probabilités sont :

1- $P(\Omega) = 1$

2- (Bn) est une famille d'évènements incompatibles ($n \in$ N), alors…

$$P\left(\bigcup_{n=0}^{+\infty} B_n\right) = \sum_{n=0}^{+\infty} P(B_n)$$

$\Sigma\ P(Bn)$ est une série à termes positifs

1.9. Définition d'une probabilité

1.9.1. Introduction

On lance un dé, et on s'intéresse au nombre qui apparait sur la face supérieure du dé. Cette expérience est une expérience aléatoire : son résultat dépend du hasard. Les résultats possibles de cette expérience aléatoire s'appellent l'univers des possibles. Si le dé comporte 6 faces, l'univers des possibles est {1,2,3,4,5,6}.

Intéressons-nous à des événements de l'expérience aléatoire, c'est-à-dire à des faits qui peuvent se produire. Par exemple, on peut choisir les événements A={on obtient un 6} et B={on obtient un nombre pair}. On répète l'expérience plusieurs fois, et on étudie la fréquence de réalisation de l'événement A, c'est-à-dire le nombre :

$$f = \frac{\text{nombre de fois où on obtient un 6}}{\text{nombre de tirages}}.$$

Comme le montre l'applet ci-dessous, quand le nombre de tirages augmente, la fréquence de réalisation de A tend à se stabiliser autour d'un nombre limite, compris entre 0 et 1. Ce nombre limite signifie intuitivement la chance qu'a l'événement A de se produire lorsqu'on réalise une expérience : on l'appelle probabilité de A, et on le note P(A). Dans notre exemple, on a bien sûr P(A)=1/6 et P(B)=1/2 si le dé n'est pas pipé.

La probabilité, dans le langage courant, apparaît donc comme une limite de fréquences, et est définie a posteriori. En modélisant l'expérience aléatoire, on va définir mathématiquement pour chaque événement une probabilité a priori.

1.9.2. Cas fini

On se place dans le cadre d'une épreuve aléatoire dont l'ensemble des événements possibles est fini (non vide). On note E={x_1,...,x_n} l'ensemble des événements possibles.

Définition 1

1. Une distribution de probabilité sur E est la donnée d'une suite finie (P_1,...,P_n) de nombres tels que :
    1. $0 \leq p_k \leq 1$.
    2. $P_1+P_2+...+P_n = 1$.
2. On appelle probabilité sur E associée à la distribution (p_1,...,p_n) l'application définie sur l'ensemble des parties de E par :

$$P(A) = \sum_{x_k \in A} p_k$$

où A est une partie de E.

Exemple

Une urne contient 7 jetons, dont 5 numérotés de 1 à 5, et 2 portant le numéro 6. Quelle est la probabilité d'obtenir un nombre pair?
L'univers des possibles est E={1,2,..,6}. La distribution de probabilité associée est :

$$P_1 = P_2 = ... = P_5 = 1/7, \quad P_6 = 2/7.$$

L'événement A="obtenir un nombre pair" a pour probabilité :

$$P(A) = P_2 + P_4 + P_6 = 4/7.$$

1.9.3. Cas infini

Les hypothèses précédentes ne sont pas toujours envisageables. Si l'on considère le problème d'un tireur qui tire sur une cible circulaire C, et dont la balle arrive aléatoirement sur la cible, le résultat de chaque tir peut être représenté par le point d'impact M. L'ensemble des résultats possibles est donc l'ensemble des points de la cible C, et il est infini.

On ne peut plus envisager de donner une probabilité non-nulle à chaque point de la cible. On définit plutôt, pour toute partie A, l'événement : "L'impact est dans A". Intuitivement, il semble clair que :

$$P(\{\text{impact dans A}\}) = \frac{\text{aire de A}}{\text{aire de C}}.$$

Malheureusement, les mathématiciens sont assez fous pour avoir pu imaginer des parties A de C dont il est impossible de calculer l'aire. Alors, avant de parler de probabilités dans un ensemble infini, il va falloir définir pour quelle parties on peut définir la probabilité, ce qui amène aux définitions suivantes :

Définition 2

Soit Ω un ensemble. On appelle tribu de parties de Ω tout ensemble $\mathcal{A}$ de parties de Ω tel que :

- $\Omega \in \mathcal{A}$.
- $\forall A \in \mathcal{A}, \quad \overline{A} \in \mathcal{A}$,
- Pour toute famille finie ou dénombrable $(A_i)_{i \in I}$ d'éléments de $\mathcal{A}$, $\bigcup_{i \in I} A_i \in \mathcal{A}$.

On dit alors que $(\Omega, \mathcal{A})$ est un espace probabilisable.

Autrement dit, une tribu contient Ω, est stable par passage au complémentaire, et par passage à la réunion dénombrable. Les éléments de $\mathcal{A}$ sont appelés événements. Il est alors possible de définir abstraitement une probabilité :

Définition 3

Soit $(\Omega,\mathcal{A})$ un espace probabilisable. On appelle probabilité sur cet espace toute application P de $\mathcal{A}$ dans [0,1] vérifiant les deux axiomes :

- $P(\Omega) = 1$
- Si (A_n) est une famille d'événements 2 à 2 incompatibles, alors :

$$P\left(\bigcup_{n=1}^{+\infty} A_n\right) = \sum_{n=1}^{+\infty} P(A_n).$$

On dit alors que $(\Omega,\mathcal{A},P)$ est un espace probabilisé, et pour tout A de $\mathcal{A}$, P(A) est la probabilité de l'événement A.

Rappelons que deux événements A et B sont incompatibles si $A \cap B = \varnothing$.

Cette définition en terme de tribus peut semble étrangement abstraite. Dans la pratique, toutefois, on ne s'en préoccupe jamais : on peut calculer la probabilité de tous les événements étudiés!

La théorie des probabilités naît véritablement au XVIIè s. des correspondances entre Blaise Pascal et Pierre de Fermat. Le *Traité du triangle arithmétique* que Pascal rédige en 1654 est le premier traité moderne d'analyse combinatoire et de calcul des probabilités. L'axiomatisation des probabilités présentée au dernier paragraphe est beaucoup plus récente : elle est due au mathématicien russe Andrei Kolmogorov, en 1929, dans son *Traité général de la mesure et théorie des probabilités*. Ce traité a beaucoup fait avancer la théorie des probabilités.

2. Variables aléatoires

2.1. Variables aléatoires discrètes

(Ω,T,*P*) est un espace de probabilité et *E* un ensemble.

2.1.1. Les définitions générales

2.1.1.1. Définition

On appelle variable aléatoire discrète une application X de Ω dans E telle que $X(\Omega)$ est fini ou dénombrable et, pour tout $x \in E$, $X^{-1}(\{x\}) \in T$. On dit que X est une variable aléatoire discrète réelle si $E = R$

2.1.1.2. Définition

Soit X une variable aléatoire discrète et notons $X(\Omega) = \{x_n; n \in I\}$ où I est fini ou dénombrable. La loi de probabilité de X est la suite $(Pn)\ n \in I$, où pour tout $n \in I$, $Pn = P(X = x_n)$

2.1.1.2. Définition

Soit (Ω_1, T_1, P_1) et (Ω_2, T_2, P_2) deux espaces de probabilité. Soit X (resp. Y) une variable aléatoire discrète définie sur Ω_1 (resp. Ω_2). On dit que X et Y ont même loi si $X(\Omega_1) = Y(\Omega_2)$ et si, pour tout $x \in X(\Omega_1)$, $P_1(X = x) = P_2(Y = x)$. On note $X \sim Y$

2.1.2. Le couple de variables aléatoires – l'indépendance

2.1.2.1. Définition

On appelle couple de variables aléatoires discrètes un couple (X,Y) où X et Y sont deux variables aléatoires discrètes. La loi conjointe du couple (X,Y) est la loi de (X,Y) vue comme variable aléatoire. Autrement dit, la loi conjointe est la donnée de toutes les valeurs de $P(X = x, Y = y)$ pour $(x,y) \in X(\Omega) \times Y(\Omega)$. Les lois de X et de Y sont les lois marginales de X et de Y

2.1.2.2. Définition

Soit x un élément de $X(\Omega)$ tel que $P(X = x) > 0$. On appelle loi conditionnelle de Y sachant que $(X = x)$ la probabilité Px définie sur $Y(\Omega)$ par

$$\forall\, y \in Y(\Omega),\ Px(\{y\}) = P(Y = y \mid X = x) = P(X = x, Y = y)P(X = x).$$

2.1.2.3. Définition

Ces définitions se généralisent à des n-uplets de variables aléatoires discrètes. Si $X_1, \ldots, X_n$ sont n variables aléatoires discrètes, $(X_1, \ldots, X_n)$ s'appelle un vecteur aléatoire discret.

2.1.2.4. Définition

Deux variables aléatoires discrètes X et Y sont dites indépendantes si, pour tout $x \in X(\Omega)$ et tout $y \in Y(\Omega)$, on a

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

2.1.2.5. Proposition

Deux variables aléatoires discrètes X et Y sont indépendantes si et seulement si, pour tout $A \subset X(\Omega)$ et tout $B \subset Y(\Omega)$, on a

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

2.1.2.6. Définition

Soit $(X_n)\ n \in I$ une famille de variables aléatoires, où I est fini ou dénombrable. On dit que les variables aléatoires $(X_n)\ n \in I$ sont mutuellement indépendantes lorsque, pour toute partie finie $J = \{i1,\ldots,ip\} \subset I$, pour tout $(x_{i1},\ldots,x_{ip}) \in X_{i1}(\Omega)\times\cdots\times X_{ip}(\Omega)$, on a

$$P(X_{i1} = x_{i1},\ldots,X_{ip} = x_{ip}) = P(X_{i1} = x_{i1})\cdots P(X_{ip} = x_{ip}).$$

2.1.2.7. Proposition

Si $X_1, X_2,\ldots,X_n$ sont des variables aléatoires mutuellement indépendantes, alors pour tout m compris entre 1 et $n-1$, et pour toutes fonctions f et g, les variables $f(X_1, \cdots,X_m)$ et $g(X_{m+1},\cdots,X_n)$ sont indépendantes.

2.1.3. L'espérance, la variance et la covariance

Dans cette partie, X et Y désignent deux variables aléatoires discrètes réelles. On note $X(\Omega) = \{x_n;\ n \in I\}$ et $Y(\Omega) = \{y_n;\ n \in J\}$.

2.1.3.1. Définition

On dit que X est d'espérance finie si la famille $(x_n P(X = x_n))$ est sommable. Si c'est le cas, on appelle espérance de X la somme de cette famille :

$$E(X) = \sum x_n P(X = x_n) \quad \text{avec } n \in I$$

2.1. 3.2. Proposition

L'espérance est linéaire : si X et Y sont d'espérance finie, $X+Y$ est d'espérance finie et $E(X+Y) = E(X) + E(Y)$

L'espérance est positive : si $X \geq 0$ est d'espérance finie, alors $E(X) \geq 0$. En particulier, si $X \leq Y$ et X et Y sont d'espérance finie, alors $E(X) \leq E(Y)$

Si $|Y| \leq X$ et X est d'espérance finie, alors Y est d'espérance finie.

2.1. 3.3. Théorème (L'espérance du produit de deux variables aléatoires indépendantes)

Si X et Y sont indépendantes et admettent une espérance, alors XY admet une espérance et $E(XY) = E(X)E(Y)$

2.1. 3.4. Théorème de transfert

Soit f une fonction définie sur $X(\Omega)$ à valeurs dans R. Alors $f(X)$ est d'espérance finie si et seulement si la famille $(P(X=x_n)f(x_n))\ n\in I$ est sommable. Dans ce cas,

$$E(f(X)) = \sum f(x_n)P(X = x_n) \quad \text{avec } n\in I$$

2.1.3.5. Définition

Soit $p \in \mathrm{N}*$. On dit que X admet un moment d'ordre p si X_p est d'espérance finie. Dans ce cas, $E(X_p)$ s'appelle le moment d'ordre p de X

2.1.3.6. Proposition

Soit $p,q \in \mathrm{N}*$ avec $p \le q$. Si X admet un moment d'ordre q, alors X admet un moment d'ordre p

2.1.3.7. Théorème (L'inégalité de Cauchy-Schwarz)

Si X et Y admettent des moments d'ordre 2, alors XY est d'espérance finie et

$$(E(XY))^2 \le E(X^2)E(Y^2).$$

2.1.3.8. Définition

Lorsque X^2 est d'espérance finie, on appelle variance de X le réel

$$V(X) = E((X-E(X))^2) = E(X^2) - (E(X))^2$$

et écart-type de X le réel $\sigma(X)$ = Racine de $V(X)$

On en déduit que :

$$V(a\ X + b) = a\ 2\ V(X)$$

2.1.3.9. Théorème (La variance d'une somme de variables aléatoires)

Soit $X_1,\ldots,X_n$ des variables aléatoires discrètes finies admettant des moments d'ordre 2. Alors

$$V(\sum X_i) = \sum V(X_i) + 2\sum(E(X_iX_j) - E(X_i)E(X_j)).$$

En particulier, si les X_I sont deux à deux indépendantes, alors

$$V(\sum Xi) = \sum V(Xi).$$

2.1.3.10. Définition

Si X et Y admettent un moment d'ordre 2, on appelle covariance de X et de Y le réel

$$\mathrm{Cov}(X,Y) = E((X - E(X))(Y - E(Y)) = E(XY) - E(X)E(Y).$$

Le coefficient de corrélation linéaire est

$$\rho(X,Y) = \mathrm{Cov}(X,Y)\ \sigma(X)\sigma(Y).$$

2.2. Variables aléatoires continues

2.2.1. Définition

On dit qu'une variable aléatoire *X* est absolument continue s'il existe une fonction *f,* positive et intégrable, telle que, pour tout intervalle *I* de R, on ait :

$$P_X(I) = P(X \in I) = \int_I f(t)dt,$$

On dit alors que *f* est la densité de probabilité de X, ou encore que *X* est une variable aléatoire absolument continue ou encore une variable aléatoire à densité.

Les variables aléatoires absolument continues s'opposent aux variables aléatoires discrètes par le fait qu'elles prennent un nombre infini non dénombrable de valeurs. En outre, pour tout x de **R**, on a P(X=x)=0.

Déterminer la loi d'une variable aléatoire X revient donc à déterminer sa densité f. Celle-ci est aussi liée à la fonction de répartition F de X. On prouve en effet qu'en tout point où f est continue, F est dérivable et F'(x)=f(x).

2.2.2. Exemple

Les sauts d'une puce.

Une puce se déplace aléatoirement à l'intérieur d'un cercle de centre O et de rayon R. On note X la distance de la puce au centre. Si on suppose que toutes les positions sont équiprobables, alors on a :

$$F(x) = P(X \leq x) = \begin{cases} 1 & \text{si } x \geq R \\ \dfrac{\pi x^2}{\pi R^2} & \text{si } 0 < x < R \\ 0 & \text{si } x \leq 0. \end{cases}$$

En dérivant F, on vérifie que X est une variable absolument continue, dont la densité est donnée par :

$$f(x) = \begin{cases} \dfrac{2x}{R^2} & \text{si } 0 < x < R \\ 0 & \text{sinon.} \end{cases}$$

3. Lois de probabilité

3.1. Lois discrètes

3.1.1. Quelques lois discrètes

3.1.1.1. La loi de Bernouilli

La variable aléatoire *X* suit une loi de Bernouilli de probabilité *p*, si *X* vaut 1 ou 0 avec les probabilités respectives *p* et 1−*p*.
On a alors :

$$P(X=x)=p^x(1-p)^{1-x}, \quad \text{pour } x = 0 \text{ ou } 1$$

$$E(X) = p,$$
$$E(X^2) = p,$$
$$V(X) = \sqrt{p}\ (1-p).$$

3.1.1.2. La loi binomiale

Si la variable aléatoire *X* suit une loi binomiale *B*(*n*,*p*), cela veut dire que *X* est égale au nombre de succès obtenus dans une série de *n* épreuves de Bernouilli de probabilité *p*. La variable aléatoire *X* peut donc prendre *n*+1 valeurs : 0,1,...,*n*.
La loi binomiale *B*(*n*,*p*) est la somme de *n* variables de Bernouilli indépendantes.
On a :

$$P(X=x) = C_n^x\ p^x(1-p)^{n-x}, \quad \text{pour } 0 \le x \le n,$$

$$E(X) = np,$$
$$V(X) = \sqrt{np(1-p)}.$$

3.1.1.3. La loi de Poisson

La variable aléatoire *X* suit une loi de Poisson *P*(λ) de paramètre λ (λ ≥ 0) si :
- *X* a pour valeurs les entiers naturels,
On a :

$$P(X=x) = e^{-\lambda}\lambda^x/x!, \quad \text{pour } x \in \mathbb{N}.$$

$$E(X) = \lambda$$
$$V(X) = \sqrt{\lambda}$$

3.2. Lois continues

3.2.1. Quelques lois continues

3.3.1.1. La loi uniforme

Définition
On dit que la variable aléatoire *X* suit une loi uniforme sur un segment [*a*,*b*] si sa

densité de probabilité $f(x)$ est une constante C sur $[a,b]$ et est nulle en dehors du segment $[a,b]$.
On a donc :

$$C = 1/b-a \quad \text{puisque } \int_a^b C dt=1$$

$$F(x) = 0 \quad \text{pour } x<a$$
$$F(x) = x-a/b-a \quad \text{pour } a\leq x<b$$
$$F(x) = 1 \quad \text{pour } x\geq b$$

Esperance mathématique, variance et écart-type

$$E(X) = \frac{1}{b-a}\int_a^b t dt = \frac{a+b}{2}$$

$$V(X) = \frac{1}{2b-2a}\int_a^b (2t-a-b)^2 d\, t = \frac{(b-a)^2}{12}$$

$$\sigma(X) = \frac{(b-a)}{2\sqrt{3}}$$

3.3.1.2. La loi exponentielle

Définition
On dit que la variable aléatoire X suit une loi exponentielle si sa densité de probabilité vaut pour $a>0$:

$$f(x) = a\exp(-ax) \quad \text{pour } x \geq 0 \text{ et}$$
$$f(x) = 0 \text{ pour } x < 0 \quad \text{On a donc :}$$
$$F(x) = P(X\leq x) = a\int_0^x \exp(-at)\, dt = 1-\exp(-ax)$$

Espérance mathématique, variance et écart-type

$$E(X) = a\int_0^{+\infty} t\exp(-at)\, d\, t = \frac{1}{a}$$

$$V(X) = a\int_0^{+\infty} (t-1/a)^2\exp(-at)\, dt = \frac{1}{a^2}$$

$$\sigma(X) = \frac{1}{a}$$

3.3.1.3. La loi Normale ou loi de Gauss

La variable aléatoire *X* suit une loi Normale ou loi de Gauss de paramètres μ,σ (σ > 0) si :

X a pour valeurs tous les réels,
$P(a \leq X < b) = \int_a^b f(t)\,dt$ où $f(x) = 1/\sigma\sqrt{2\pi}e^{-1/2(x-\mu/\sigma)2}$ (f est la densité de probabilité et a comme représentation graphique une courbe en cloche).
On note cette loi N(μ,σ).
On a :
$E(X)=\mu$
$\sigma(X)=\sigma$.

On dit que *N*(0,1) est la loi normale centrée réduite. Si *X* suit la loi *N*(0,1) alors :
$P(a \leq X < b) = \int_a^b = 1/\sqrt{2\pi}e^{-t2/2}dt$
Si *X* suit la loi *N*(μ,σ) alors $X-\mu/\sigma$ suit la loi *N*(0,1).
On a des tables où on peut lire que :
$P(|X-\mu|/\sigma > 1.96)=0.05$,
$P(|X-\mu|/\sigma > 2.58)=0.01$,
$P(|X-\mu|/\sigma > 3.1)=0.001$,
et on a $P(|X-\mu|/\sigma>t)=1-2\int_0^t f(x)dx$.

4. Convergences et similitudes entre statistiques et probabilités

En mathématiques, la loi des grands nombres permet d'interpréter la probabilité comme une fréquence de réalisation, justifiant ainsi le principe des sondages, et présente l'espérance comme une moyenne. Plus formellement, elle signifie que la moyenne empirique, calculée sur les valeurs d'un échantillon, converge vers l'espérance lorsque la taille de l'échantillon tend vers l'infini.

Plusieurs théorèmes expriment cette loi, pour différents types de convergence en théorie des probabilités. La loi faible des grands nombres met en évidence une convergence en probabilité, tandis que la loi forte des grands nombres donne une convergence presque sûre

4.1. Loi faible des grands nombres

Théorème : Soit (*Xn*)
une suite de variables intégrables, indépendantes deux à deux, et identiquement distribuées. Soit *m* leur espérance commune. On note :

$$S_n = \frac{X_1 + \cdots + X_n}{n}$$

Alors la suite (*Sn*) tend vers *m*
en probabilités.
Ex : On lance un dé non pipé, X_n vaut 1 si le n-ième lancer amène 5, et 0 sinon. Alors S_n tend vers 1/6 en probabilités.

4.2. Loi forte des grands nombres

Théorème : Soit (*Xn*)

une suite de variables intégrables mutuellement indépendantes et identiquement distribuées. Soit *m* leur espérance commune. On note :

$$S_n = \frac{X_1 + \cdots + X_n}{n}$$

Alors la suite (*Sn*) tend vers *m*
presque sûrement.
C'est à Jacques Bernoulli que l'on doit le premier énoncé de la loi des grands nombres; il apparait dans son ouvrage *Ars Conjectandi* publié en 1713, huit ans après sa mort. Il avait pour cadre le jeu du pile ou face (schéma de Bernoulli). Le terme de "loi des grands nombres" est lui dû à Poisson. Ce terme juridique est à mettre en rapport avec le titre de l'ouvrage dans lequel il l'introduit, *Recherches sur les probabilités des jugements*, paru en 1837. De nombreux mathématiciens ont ensuite généralisé les énoncés de Bernoulli et Poisson, citons notamment Kolmogorov et Tchebychev.

4.3. Les similitudes entre statistiques et probabilité

Les notions introduites en probabilités concernant les variables aléatoires sont similaires à celles introduites en statistique descriptive concernant les variables statistiques

STATISTIQUE DESCRIPTIVE	PROBABILITES	
On relève des observations sur une population. Les résultats obtenus constituent les observations d'une variable statistique	On décrit les observations que l'on peut faire sur une population au cours d'une expérience aléatoire (il n'est pas nécessaire de réaliser cette expérience). Les résultats que l'on peut obtenir constituent les observations d'une variable aléatoire (v.a.)	
	V.A. DISCRETE	V.A. CONTINUE
xi : valeurs observées ni : effectifs fi = ni/n : fréquences	xi : valeurs possibles pi : probabilités	x : valeurs possibles (réelles) f : densité

Les notions de fréquences, probabilités, densité sont similaires

$\sum f_i = 1$	$\sum p_i = 1$	$\int f(x)\,dx = 1$
Fréquence cumulée croissante	Fonction de répartition	
$f^+(x) = \sum_{\{i / x_i < x\}} f_i$	$F(x) = \sum_{\{i / x_i < x\}} p_i$	$F(x) = \int_{-\infty}^{x} f(t)\,dt$

Moyenne arithmétique	Moyenne ou Espérance mathématique	
$\bar{x} = \sum_i f_i x_i$	$E(X) = \sum_i p_i x_i$	$E(X) = \int_{\mathbb{R}} x f(x)\,dx$
Variance	**Variance**	
$V(x) = \sum_i f_i (x_i - \bar{x})^2$	$V(X) = \sum p_i [x_i - E(X)]^2$	$V(X) = \int_{\mathbb{R}} [x - E(X)]^2 f(x)\,dx$

4.4. Des propriétés importantes

1. E[aX + bY] = a E[X] + b E[Y]

   E[aX + b] = a E[X] + b

   E[a] = a

2. V[aX + b] = a² V[X]

3. V[X + Y] = V[X] + V[Y] uniquement dans le cas où X et Y sont indépendantes

4. X suit une $N(\mu,\sigma)$ est équivalent à $(X-\mu)/\sigma$ suit une N(0,1)

5. X_1 suit une $N(\mu_1,\sigma_1)$; X_2 suit une $N(\mu_2,\sigma_2)$; X_1 et X_2 indépendantes
alors $X_1 + X_2$ suit une $N(\mu_1 + \mu_2 ; \sqrt{\sigma_1^2 + \sigma_2^2})$

6. X suit une $N(\mu,\sigma)$ implique que (aX + b) suit une $N(a\mu+b ; |a|\sigma)$

5. La régression et la corrélation

Lorsqu'on observe deux variables quantitatives sur les mêmes individus, on peut s'intéresser à une liaison éventuelle entre ces deux variables.

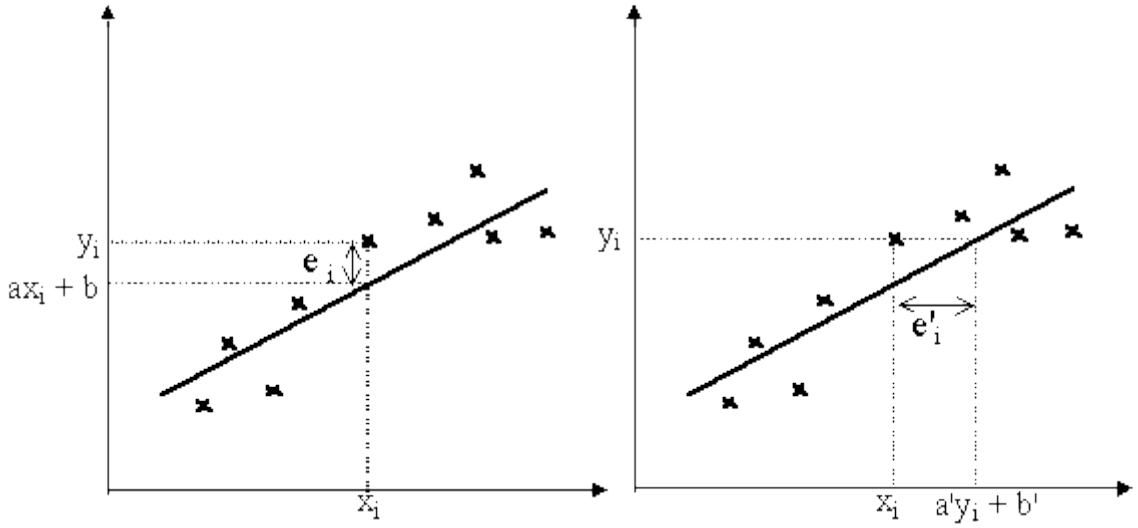
La régression fournit une expression de cette liaison sous la forme d'une fonction mathématique.

La corrélation renseigne sur l'intensité de cette liaison.

5.1. L'ajustement d'un nuage de points à une fonction mathématique

5.1.1. L'ajustement linéaire par la méthode des moindres carrés

Lorsque le nuage de points (xi , yi) est à peu près rectiligne, on peut envisager d'exprimer la liaison entre x et y sous forme de fonction affine y = ax + b



$$\sum_i e_i^2 = \sum_i (y_i - ax_i - b)^2 \qquad \sum_i e_i'^2 = \sum_i (x_i - a'y_i - b')^2$$

$$b = \bar{y} - a\bar{x} \qquad b' = \bar{x} - a'\bar{y}$$

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad a' = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

5.1.2. L'ajustement à une fonction exponentielle

Pour ajuster un nuage de points à une courbe exponentielle $y = ba^x$, il suffit de faire le changement de variable Y = ln y , X = x , A = ln a , B = ln b , pour obtenir l'équation Y = AX + B, et d'utiliser ensuite l'ajustement linéaire par la méthode des moindres carrés sur les points (Xi , Yi).

5.1.3. L'ajustement à une fonction puissance

Pour ajuster un nuage de points à une courbe puissance $y = bx^a$, il suffit de faire le changement de variable Y = ln y , X = lnx , A = a , B = ln b , pour obtenir l'équation Y = AX + B , et d'utiliser ensuite l'ajustement linéaire par la méthode des moindres carrés sur les points (Xi , Yi).

5.2. La mesure de l'intensité de la relation linéaire entre deux variables

5.2.1. La covariance

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$Cov(x,y) > 0 \quad \Leftrightarrow$ x et y varient dans le même sens

$Cov(x,y) < 0 \quad \Leftrightarrow$ x et y varient en sens contraire

$Cov(x,y) = Cov(y,x)$

$Cov(x,x) = V(x)$

5.2.2. Le coefficient de corrélation linéaire

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{cov(x,y)}{\sigma(x)\,\sigma(y)}$$

$-1 \le r \le 1$

$$y = ax + b \quad \Leftrightarrow \quad \begin{cases} r = 1 & si \quad a > 0 \\ r = -1 & si \quad a < 0 \end{cases}$$

$|r| = 1 \quad \Leftrightarrow$ relation fonctionnelle linéaire

$r = 0 \quad \Leftrightarrow$ indépendance linéaire

$0 < |r| < 1 \quad \Leftrightarrow$ dépendance linéaire d'autant plus forte que $|r|$ est grand

Attention:

Une forte causalité entre x et y implique une forte relation entre x et y qui n'est pas forcément linéaire; on n'a donc pas obligatoirement une forte corrélation linéaire.

Une forte corrélation linéaire n'implique pas forcément une forte causalité.

5.2.3. Les droites de régression

Dy/x : $y = ax + b$ avec $a = \frac{cov(x,y)}{V(x)}$ et $b = \bar{y} - a\bar{x}$

Dx/y : $x = a'y + b'$ avec $a' = \frac{\text{cov}(x, y)}{V(y)}$ et $b' = \bar{x} - a'\bar{y}$

La position des deux droites de régression l'une par rapport à l'autre donne un renseignement sur l'intensité de la relation linéaire:

* droites de régression confondues $\Leftrightarrow$ $aa' = 1$ $\Leftrightarrow$ relation fonctionnelle linéaire

* droites de régression perpendiculaires dont une de pente nulle $\Leftrightarrow$ $aa' = 0$ $\Leftrightarrow$ indépendance linéaire

* Plus les droites sont proches, plus la relation linéaire est importante

Relations intéressantes:

* $r^2 = aa'$

* $r = a\frac{\sigma(x)}{\sigma(y)} = a'\frac{\sigma(y)}{\sigma(x)}$

II. Exercices

Il y a 92 exercices corrigés à étudier. Pour les voir, il faut aller sur Google et taper :

1. ei – exercices de probabilités corrigés – IECL – Université Lorraine France.

C'est un document pdf de 166 pages comportant 131 exercices avec corrigés détaillés. Il faut étudier les 92 exercices suivants :

- Dénombrement : Les exercices du n°1 au n° 18
- Probabilité sur un ensemble : Les exercices du n°19 au n° 41
- Variables aléatoires : Les exercices du n°42 au n° 68
- Probabilité sur un ensemble dénombrable: Les exercices du n°92 au n° 115

2. Cours de probabilité et statistiques – Université Claude Bernard Lyon – France.

C'est un document pdf de 71 pages comportant une multitude d'exemples détaillés. Il faut étudier les 22 exemples suivants :

- V.A et lois discrètes : Les exemples du n°15 page 11 au n° 32 page 23
- Lois continues : Les exemples des pages 29, 32, 33 et 35.

Partie 3. Echantillonnage et tests

I. Cours

1. Les théorèmes des statistiques différentielles

1.1. Problèmes de jugement sur échantillon

L'exploitation des données peut prendre plusieurs formes :
a/ L'inférence statistique ou "théorie de l'estimation" : connaissant un échantillon, on désire émettre une estimation sur la population totale. Dans ce cas, on n'a pas d'idée a priori sur le paramètre à estimer :
on construira **un intervalle de confiance I_α au seuil α**.
Cet intervalle I_α dépend de l'échantillon et contient, en général, la valeur du paramètre sauf dans α % des cas c'est à dire, il y a seulement α % des échantillons qui ont un I_α qui ne contient pas le paramètre (on dit qu'on a un risque d'erreur égal à α).
b/ Le test d'hypothèses permet de savoir si il y a accord entre théorie et expérience. Dans ce cas on a une idée a priori sur la valeur que doit avoir le paramètre : on construit le test d'hypothèses (deux hypothèses H_0 et H_1 seront en concurrence), puis on prélève un échantillon et on regarde si cet échantillon vérifie le test ce qui permet d'accepter ou de refuser l'hypothèse privilégiée H_0.
Par exemple : on veut contrôler qu'une fabrication correspond bien à ce qui a été décidé, pour cela on fabrique un test d'hypothèses, puis on teste l'hypothèse H_0 sur un échantillon de la production.
c/ Le test d'homogénéite permet de comparer une distribution expérimentale à une distribution théorique.

Remarque :

en a/ et en b/ on a seulement comparer ou estimer des valeurs caractéristiques comme fréquences ou moyennes, en c/ on compare deux distributions.

1.2. Théorème de Bienaymé-Tchebychef

1.2.1. Théorème

La probabilité pour qu'une variable aléatoire X diffère de sa moyenne (en valeur absolue) d'au moins k fois son écart type, est au plus égale à $1/k^2$, c'est à dire si X a comme moyenne $m=E(X)$ et comme écart type σ on a : $Proba(|X-m| \geq k\ \sigma) \leq 1//k^2$ 1.2.

1.2.2. Exemples

- On lance 100 fois un dé et on considère comme événement :
  on obtient 6 ou on n'obtient pas 6.
  Par l'expérience on a obtenu n_1 fois le nombre 6 Trouver un majorant de $Proba(|n_1/100-1/6| \geq 1/10)$.

  La probabilité d'avoir un 6 est : $p=1/6$ et de ne pas avoir un 6 est 5/6.
  Si Y est la variable aléatoire égale à la fréquence de l'événement favorable on a $E(Y)=p=1/6 \simeq 0.166666666667$ et $\sigma(Y)=sqrt(1/6*5/6/100)=sqrt(5)/60 \simeq 0.037267799625$ Théorème de Bienaymé-Tchebychef nous dit que :
  $Proba(|n_1/100-1/6| \geq k\ \sigma(Y)) \leq 1//k^2$
  On cherche k pour avoir $k\ \sigma(Y)= k\ sqrt(5)/60 \leq 1/10$
  on prend $k=2.6832815732$ car $k \leq 6/sqrt(5) \simeq 2.683281573$
  donc $Proba(|n_1/100-1/6| \geq 1/10) < 1/2.68^2 \simeq 0.139$
  Cela veut dire que : $n_1/100$ se trouve dans l'intervalle $1/6-1/10 \simeq 0.0666666666667; 1/6+1/10 \simeq 0.266666666667$ avec la probabilité $1-0.139=0.861$ ou encore que n_1 se trouve dans l'intervalle 6; 26 avec la probabilité 0.861.

- même exercice mais cette fois on lance le dé 6000 fois.
  Par l'expérience on a obtenu n_1 fois le nombre 6 Trouver un majorant de $Proba(|n_1/6000-1/6| \geq 1/100)$.

  Si Y est la variable aléatoire égale à la fréquence de l'événement favorable on a $E(Y)=p=1/6 \simeq 0.166666666667$ et $\sigma(Y)=sqrt(1/6*5/6/6000)=sqrt(5/60)/60 \simeq 0.00481125224325$ On cherche k pour avoir $k\ \sigma(Y)= k\ sqrt(5/60)/60 \leq 1/100$
  on prend $k=2.07846096908$ car $k \leq 6/sqrt(50/6) \simeq 2.07846096908$
  donc $Proba(|n_1/6000-1/6| \geq 1/100) < 1/2.07846096908^2 \simeq 0.231481481482$
  Cela veut dire que : $n_1/6000$ se trouve dans l'intervalle $1/6-1/100 \simeq 0.16566666666667; 1/6+1/10 \simeq 0.176666666667$ avec la probabilité $1-0.231481481482=0.768518518518$ ou encore que n_1 se trouve dans l'intervalle 940; 1060 avec la probabilité de 0.768518518518. **Remarque**
  En approchant la loi binomiale par la loi normale de moyenne $n*p=6000*1/6=1000$ et d'écart type $\sigma=sqrt(np(1-p)=sqrt(6000*1/6*5/6) \simeq 28.8675134595$ On a $60/28.8675134595=2.07846096908$ On cherche dans une table $\psi(t)=Prob(0<T<t)=\psi(2.07846096908)$ et on trouve 0.481. Donc $Prob(-t<T<t)=2*0.481=0.962$ ou dans une table $\Pi(t)=Prob(-\infty<T<t)=\Pi(2.07846096908)$ et on trouve 0.981. Donc $Prob(-t<T<t)=2*0.981-1=0.962$ donc n_1 se trouve dans l'intervalle 940; 1060 avec la probabilité de $0.481*2=0.962$.

- On extrait 1000 fois avec remise une carte d'un jeu de 32 cartes et on considère comme événement :
  on obtient un as ou on n'obtient pas un as.

Par l'expérience on a obtenu n_1 fois un as Trouver un minorant de *Proba*($105<n_1<145$).

La probabilité d'avoir un as est : p=1/8 et de ne pas avoir un as est 7/8. Si *Y* est la variable aléatoire égale à la fréquence de l'événement favorable on a $E(Y)=p=1/8=0.125$ et $\sigma(Y)=sqrt(1/8*7/8/1000)=sqrt(7/10)/80\simeq 0.0104582503317$. On a *Proba*($105<n_1<145$)=*Proba*($|n_1-125|<20$)=*Proba*($|n_1/1000-0.125|<1/50$) Le théorème de Bienaymé-Tchebychef nous dit que : *Proba*($|n_1/1000-1/8| \geq k\ \sigma(Y)) \leq 1//k^2$
On choisit $k\ \sigma(Y)=1/50$ c'est à dire $k=1/50/0.0104582503317=1.91236577493$ donc $1/k^2=0.273437500001$ Cela veut dire que *Proba*($|n_1/1000-0.125|<1/50) \geq 0.273437500001$ donc
Proba($105<n_1<145) \leq 1-0.273437500001=0.7265625$ **Remarque** *Proba*($|n_1/1000-0.125|>1/100) \geq 1/0.956182887465^2=1.09375000001$ ce qui ne nous apporte rien!

1.3. Loi des grands nombres

Notation

On note ici $X_n=X_1+X_2+..+X_n/n$ pour bien faire ressortir que X_n dépend de *n*, mais quelquefois dans la suite on écrira simplement : $X=X_1+X_2+..+X_n/n$ pour ne pas alourdir les notations.

1.3.1. Loi faible des grands nombres :

Soient $X_1, X_2,.., X_n$ des variables aléatoires indépendantes de moyenne $\mu_1, \mu_2,.., \mu_n$ et d'écart-type $\sigma_1, \sigma_2,.., \sigma_n$.
Si quand *n* tend vers l'infini $1/n \sum_{j=1}^{n} \mu_j$ tend vers μ et,
si quand *n* tend vers l'infini $1/n^2 \sum_{j=1}^{n} \sigma_j^2$ tend vers 0,
alors $X_n=X_1+X_2+..+X_n/n$ converge en probabilité vers μ quand *n* tend vers l'infini (i.e. pour tout ϵ et pour tout η il existe n_0 tel que pour tout $n>n_0$ on a *Proba*($|X_n-\mu|>\epsilon)<\eta$).
Cas des échantillons :
Si $X_1,X_2,..,X_n$ sont un échantillon de *X* de moyenne μ et décart-type σ, on a $\mu_1=\mu_2=..=\mu_n=\mu$ et $\sigma_1=\sigma_2=..=\sigma_n=\sigma$. Donc $1/n \sum_{j=1}^{n} \mu_j =\mu$ et quand *n* tend vers l'infini $1/n^2 \sum_{j=1}^{n} \sigma_j^2 =\sigma^2/n$ tend vers 0 ce qui montre que la variable aléatoire $X_n=X_1+X_2+..+X_n/n$ converge en probabilité vers μ quand *n* tend vers l'infini.

1.3.2. Loi forte des grands nombres :

Soient $X_1, X_2,.., X_n$ des variables aléatoires indépendantes de moyenne $\mu_1, \mu_2,.., \mu_n$ et d'écart-type $\sigma_1, \sigma_2 ,.., \sigma_n$.
Si quand *n* tend vers l'infini $1/n \sum_{j=1}^{n} \mu_j$ tend vers μ et,
si $\sum_{j=1}^{\infty} \sigma_j^2/j^2$ est convergente,
alors $X_n=X_1+X_2+..+X_n/n$ converge presque sûrement vers μ quand *n* tend vers l'infini (i.e. dire que Y_n converge presque sûrement vers *U* c'est dire que l'ensemble des points de divergene est de probabilité nulle i.e.
Proba($\omega, \lim_{n \to +\infty}(Y_n(\omega) \neq U(\omega))=0$).
Cas des échantillons :
Si $X_1,X_2,..,X_n$ sont un échantillon de *X* de moyenne μ et décart-type σ, on a

$\mu_1=\mu_2=..=\mu_n=\mu$ et $\sigma_1=\sigma_2=..=\sigma_n=\sigma$.
Donc $1/n \sum_{j=1}^{n} \mu_j = \mu$ et $\sum_{j=1}^{\infty}\sigma^2/j^2 = \sigma^2\sum_{j=1}^{\infty}1/j^2$ est convergente ce qui montre que :
$X_n=X_1+X_2+..+X_n/n$ converge presque sûrement vers μ quand n tend vers l'infini.

1.3.3. Le théorème central-limite :

Quand n tend vers l'infini, alors
$\bar{Y}_n=\sqrt{n}\,(X_n-\mu)/\sigma$ converge en loi vers U variable aléatoire qui suit la loi normale centrée réduite (dire que Y_n converge en loi vers $U \in N(0,1)$ veut dire que si F est la fonction de répartition de la loi normale centrée réduite et si F_n est la fonction de répartition de Y_n alors pour tout $x \in \mathbb{R}$, $F_n(x)$ tend vers $F(x)$ quand n tend vers l'infini).

1.4. Moyenne et variance empirique

Soit un échantillon d'effectif n et $(x_1,x_2,...,x_n)$ les n valeurs observées.
La moyenne empirique est :
$m=x_1+x_2+...+x_n/n$
La variance empirique est :
$s^2=(x_1-m)^2+...+(x_n-m)^2/n$
Soit un échantillon de taille n.
Les n valeurs observées $(x_1,x_2,...,x_n)$ du caractère sont considérées comme étant les valeurs de n variables aléatoires indépendantes $X_1,X_2,..,X_n$ suivant la même loi F d'espérance μ et d'écart-type σ.
L'ensemble des moyennes d'échantillons de taille n est la variable aléatoire
$X=X_1+X_2+..+X_n/n$.
Si le résultat observé est $x_1,x_2..x_n$, alors la valeur observée de X est la moyenne empirique m :
$m=x_1+x_2+..+x_n/n$.
L'ensemble des variances d'échantillons de taille n est la variable aléatoire
$S^2=(X_1-X)^2+(X_2-X)^2+..+(X_n-X)^2/n$.
Si le résultat observé est $x_1,x_2..x_n$, alors la valeur observée de S^2 est la variance empirique s^2 :
$s^2=(x_1-m)^2+(x_2-m)^2+..+(x_n-m)^2/n$.

1.5. Étude de X

1.5.1. Théorèmes

La variable aléatoire $X=X_1+X_2+..+X_n/n$ converge en probabilité vers μ.
De plus X a pour moyenne μ et pour variance σ^2/n.
Quand n tend vers l'infini, $\sqrt{n}\,(X-\mu)/\sigma$ converge en loi vers U variable aléatoire qui suit la loi normale centrée réduite.

1.6. Estimateur de μ

On appelle estimateur de μ, une variable aléatoire U_n fonction d'un échantillon $X_1,X_2,..,X_n$ qui vérifie :
$\lim_{n->\infty} E(U_n)=\mu$ et $\lim_{n->\infty}\sigma^2(U_n)=0$
On dit que U_n est un estimateur **sans biais** de μ si c'est un estimateur de μ qui vérifie $E(U_n)=\mu$.

1.6.1. Théorème

$X=(X_1+X_2+..+X_n)/n$ est un estimateur sans biais de μ.

1.7. Étude de S^2

1.7.1. Théorème

La variable $S^2=(X_1-X)^2+(X_2-X)^2+..+(X_n-X)^2/n$ converge presque sûrement vers σ^2 quand n tend vers l'infini.
De plus S^2 a pour moyenne :
$E(S^2)=n-1/n\sigma^2$
et pour variance :
$\sigma^2(S^2)= V(S^2)=n-1/n^3((n-1)\mu_4-(n-3)\sigma^4)$ où $\mu_4=E((X-\mu)^4)$.

1.7.2. Théorème limite pour S^2 :

Quand n tend vers l'infini, $\sqrt{n}\ (S^2-n-1/n\sigma^2)/\sqrt{\mu_4-\sigma^4}$ converge en loi vers U variable aléatoire qui suit la loi normale centrée réduite (dire que Y_n converge en loi vers $U \in N(0,1)$ veut dire que si F est la fonction de répartition de la loi normale centrée réduite et si F_n est la fonction de répartition de Y_n alors pour tout $x\in \mathbb{R}$, $F_n(x)$ tend vers $F(x)$ quand n tend vers l'infini).

1.8. Estimateur de σ^2

On appelle estimateur de σ^2, une variable aléatoire V_n fonction d'un échantillon $X_1,X_2,..,X_n$ qui vérifie :
$\lim_{n->\infty} E(V_n)=\sigma^2$ et $\lim_{n->\infty}\sigma^2(V_n)=0$
On dit que V_n est un estimateur **sans biais** de σ^2 si c'est un estimateur de σ^2 qui vérifie $E(V_n)=\sigma^2$.

1.8.1. Théorème

$Z^2=(X_1-\mu)^2+..+(X_n-\mu)^2/n$ est un estimateur sans biais de σ^2.
$S^2=(X_1-X)^2+..+(X_n-X)^2/n$ est un estimateur de σ^2.
$n/n-1S^2=(X_1-X)^2+..+(X_n-X)^2/n-1$ est un estimateur sans biais de σ^2.
En effet :
Pour S^2 cela découle des théorèmes précédents.
Pour Z^2 on a :
$E(Z^2)=1/n\sum_{j=1}^{n} E((X_j-\mu)^2)=1/nn\sigma^2=\sigma^2$
et puisque $\sigma^2(X-\mu)^2=E((X-\mu)^4)-(\sigma^2)^2=\mu_4-(\sigma^2)^2$ on a :
$\sigma^2(Z^2)=1/n(\mu_4-(\sigma^2)^2)$ (où $\mu_4=E((X-\mu)^4)$ est le moment centré d'ordre 4).

1.8.2. Remarque :

À partir des valeurs $x_1,x_2,..,x_n$ de l'échantillon, on utilisera lorsqu'on connait μ, $(x_1-\mu)^2+(x_2-\mu)^2+..+(x_n-\mu)^2/n$ comme estimateur de σ^2 et si μ est inconnu on utilisera comme estimateur de σ^2 $(x_1-m)^2+(x_2-m)^2+..+(x_n-m)^2/n-1$ avec $m=x_1+x_2+..+x_n/n$.

1.9. Autrement dit

Le problème est d'obtenir, au vu de l'échantillon empirique, des renseignements sur la population dont l'échantillon est issu (c'est à dire sur la population parente de moyenne μ et d'écart-type σ), en particulier sur la valeur de sa moyenne μ.
En général σ n'est pas connu, on prend faute de mieux, quand n est grand :
$\sigma = s\sqrt{n/n-1}$ où s^2 est la valeur observée de :
$S^2=(X_1-Y)^2+(X_2-Y)^2+..+(X_n-Y)^2/n$ qui a pour moyenne $n-1/n\sigma^2$.
Grâce au théorème central-limite, la variable $X=X_1+..+X_n/n$ va nous servir à trouver une valeur de μ car :
X a pour moyenne μ et pour variance $\sigma^2/n \simeq s^2/n-1$ donc la variable aléatoire :
$\sqrt{n}\,(X-\mu)/\sigma \simeq \sqrt{n-1}\,(X-\mu)/s$ converge en loi vers $U \in N(0,1)$.

2. Une idée d'ensemble sur les tests statistiques

2.1. Principe des tests statistiques

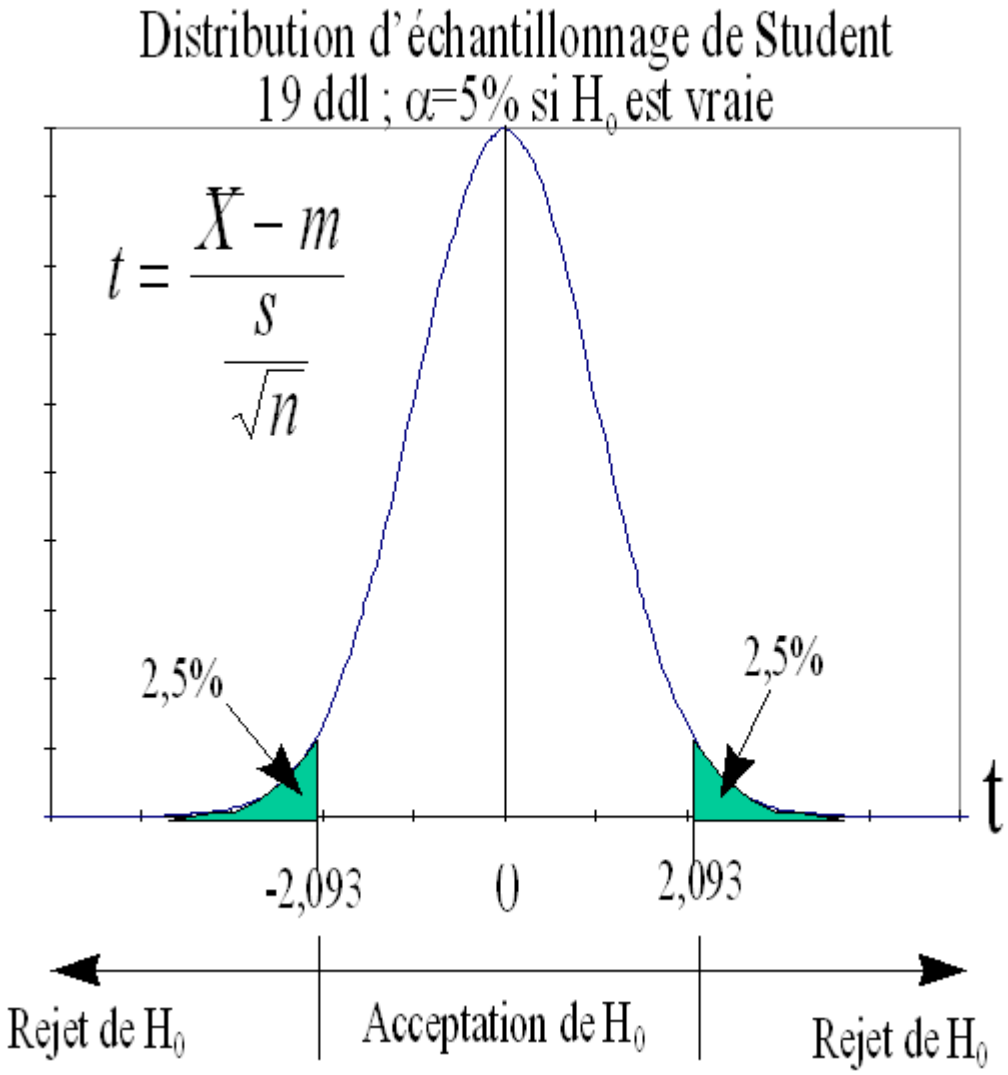
Les enquêtes statistiques ont généralement pour objectif de comparer les paramètres statistiques déterminés à partir d'échantillons avec une valeur théorique ou bien avec les paramètres d'autres échantillons.

On peut se demander par exemple, si la cote moyenne obtenue sur un échantillon de pièces usinées est compatible avec la norme en vigueur ou bien si un nouveau procédé modifie de façon significative ou non les caractéristiques (quantitatives ou qualitatives) d'un produit. Pour cela on sera amené à faire des hypothèses et à les tester.

Par exemple, la cote moyenne d'un échantillon de 20 pièces (obtenu par tirage au sort) issues d'un lot de fabrication est m= 6,2 mm. On veut savoir si cette valeur est compatible avec la valeur 6 mm indiquée dans le cahier des charges. Il s'agit en fait de vérifier ou d'infirmer l'hypothèse que l'échantillon a été extrait d'une population où la moyenne théorique est μ=6 mm (l'écart type σ de cette population est ici inconnu). Cette hypothèse, appelée hypothèse nulle et notée H0, va nous permettre de réaliser un test basé sur un modèle théorique (distributions d'échantillonnage). Nous verrons que le but d'un test n'est pas en général de vérifier l'hypothèse de départ H0, qui ne correspond qu'à une zone de probabilité définie par une loi de distribution (modèle mathématique), mais de rejeter cette hypothèse avec un risque faible (que l'on saura quantifier). On acceptera alors l'hypothèse contraire (Hypothèse alternative notée H1) selon laquelle l'échantillon ne provient pas de la population de moyenne μ=6 mm. On aura ainsi montré que la différence observée entre les paramètres statistiques (moyenne observée et moyenne théorique) est due, avec un faible risque de se tromper, à autres chose qu'à de simples fluctuations d'échantillonnage.

Le choix de la distribution d'échantillonnage est donc la première étape pour réaliser le test. Dans notre exemple, la variable étudiée (mesure en mm) est une variable quantitative, de type continue. L'échantillon étant de petite taille (n<30), la distribution d'échantillonnage la plus adaptée est donc la distribution de Student. Pour utiliser ce modèle, mous avons vu qu'il faut démontrer (ou supposer) que la distribution du paramètre quantitatif (ici les cotes des pièces usinées) suit une loi Normale dans la population. Supposons donc que la Normalité de la distribution à été démontrée par une étude précédente. A partir de l'échantillon, nous pouvons également estimer l'écart type dans la population (l'estimation sera notée s).

Si la différence entre la moyenne observée et la valeur théorique n'est due qu'à des fluctuations d'échantillonnage (c'est l'hypothèse H0) alors nous avons 95 % de chance que la valeur de la cariable se trouve dans l'intervalle ± 2,093, valeur trouvée dans la table de Student avec n-1= 19 degré de liberté.



Distribution d'échantillonnage Student

Si la valeur calculée de t est dans cet intervalle, on acceptera H0, c'est à dire que nous n'aurons pas été capables de mettre en évidence (si elle existe) une différence significative entre la moyenne observée et la moyenne théorique.

Si en revanche, la valeur de t est extérieure à l'intervalle, on rejettera l'hypothèse H0, car la probabilité que les fluctuations d'échantillonnage (5 %) expliquent la différence mesurée est suffisamment faible.

Nous pouvons résumer les 4 étapes d'un test statistique de la façon suivante :

1. Déterminer la loi de distribution théorique permettant d'analyser les fluctuations d'échantillonnage et donc vérifier que l'on est dans les conditions

de validité de cette loi (taille de l'échantillon, distribution de la variable, ou autres...)
2. définir une hypothèse à contrôler (H0)
3. ε Sous cette hypothèse, calculer la valeur de la variable discriminante (t, ou autre).
4. Porter un jugement en fonction des risques d'erreur admis.

2.2. Les risques d'erreur

Introduction

Ces risques ont déjà été définis dans le chapitre estimation/échantillonnage mais il convient de revenir sur leur signification lors de la réalisation des tests statistiques.

2.2.1. Le risque de première espèce ou risque α

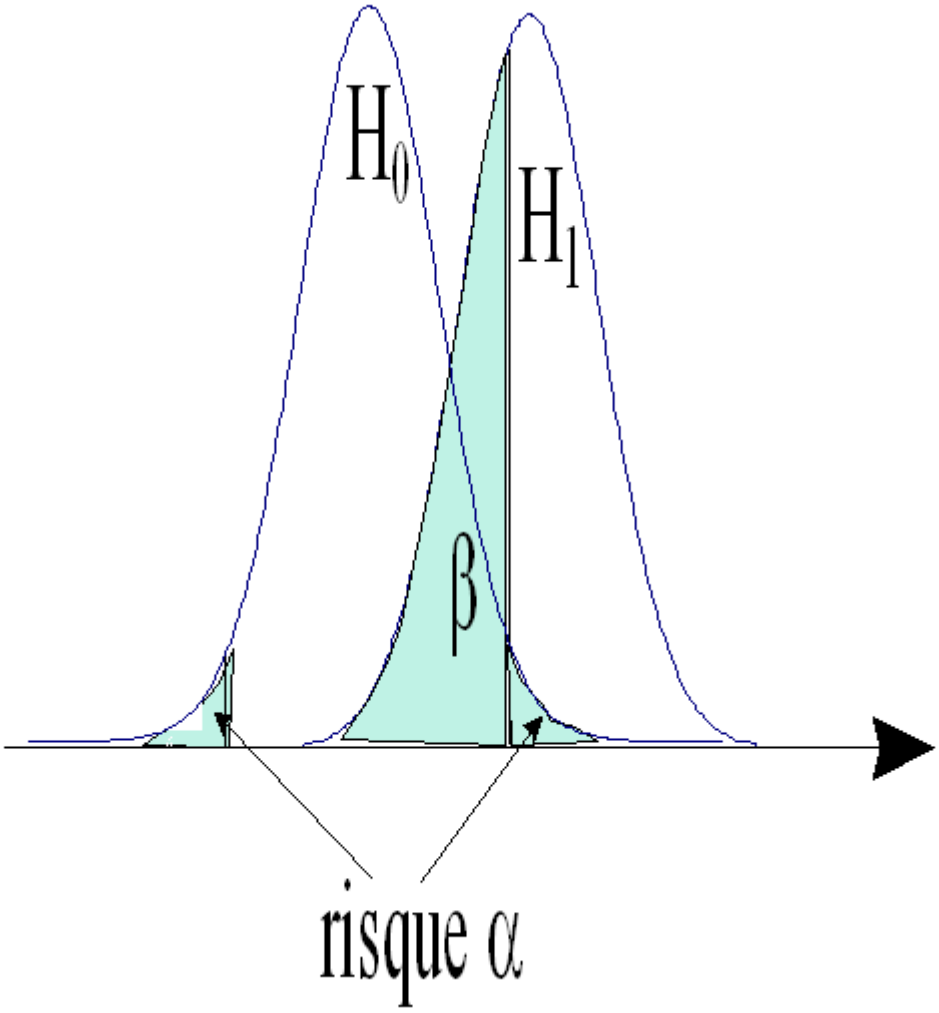
C'est le risque de rejeter l'hypothèse nulle (H0) alors que celle-ci est vraie. Ce risque est parfaitement connu : c'est la probabilité utilisée lors de la réalisation du test, c'est à dire la probabilité pour que la valeur de la variable aléatoire Z (distribuée selon la loi de distribution choisie pour le test) soit extérieure à l'intervalle [z1, z2] ([-2,093 ; +2.093] dans notre exemple pour une loi de Student, 19 ddl et un risque α égal à 5 %). La valeur 5% est classiquement utilisée avec dans certains cas l'utilisation d'un risque 1%.

Au risque 5 % on estime donc que la probabilité pour que la différence observée soit due aux fluctuations d'échantillonnage est suffisamment faible pour accepter H0.

2.2.2. Le risque β ou risque de deuxième espèce

C'est la probabilité d'accepter l'hypothèse H0 alors que celle-ci est fausse (c'est à dire que c'est H1 qui est vraie). Le fait d'accepter H0 est en effet la conséquence de ne pas avoir pu mettre en évidence une différence significative entre, par exemple, deux paramètres statistiques obtenus à partir de deux échantillons (deux moyennes ou deux pourcentages). On estime dans ces conditions que la différence observée est due aux fluctuations d'échantillonnage et que les deux échantillons proviennent d'une même population . En réalité cette différence peut exister, et donc les deux échantillons peuvent provenir de deux populations distinctes ; seulement cette différence est trop faible pour être démontrée à l'aide du test utilisé.

Le risque β peut être quantifié mais ceci reste relativement délicat car il faut alors connaître la distribution des deux populations et faire intervenir dans le calcul la différence entre les paramètres statistiques, caractéristiques de ces deux populations. Nous n'aborderons donc pas cet aspect quantitatif.



Risque de deuxième espèce

On remarque à la vue du graphique, que plus la différence entre les deux populations est faible, plus le risque β est important, donc plus il sera difficile dans ces conditions de conclure à une différence significative. Cette aptitude à pouvoir conclure en faveur d'une différence significative est appelée puissance β du test et est représentée par la quantité 1-

La puissance d'un test peut être améliorée en augmentant la taille du (ou des) échantillons testés à condition qu'il existe une différence significative entre les populations dont sont issus les échantillons.

Un autre moyen consiste à réduire la dispersion des valeurs (on joue alors sur la variance de la distribution) mais ceci ne peut être réalisé qu'en modifiant l'expérience, c'est à dire en utilisant une méthode plus précise.

Il apparaît donc qu'à l'issue de tout test statistique, les conclusions devront faire la part de la réalité et du hasard. Dans tous les cas, on ne pourra jamais rien affirmer, une part de risque étant toujours présente. On dira simplement que certaines hypothèses sont plus probables que d'autres.

Il faudra cependant rester très vigilant lors de la mise en oeuvre d'un test statistique car les conclusions que l'on tirera de ce test peuvent avoir de lourdes conséquences. Il sera toujours nécessaire de se poser des questions sur la nécessité d'une expérimentation, sur les objectifs attendus et sur les moyens dont on .dispose pour mener à bien l'enquête.

3. Les tests

3.1. Les tests d'hypothèses

Concernant une variable aléatoire X, on souhaite comparer la valeur effective d'un paramètre p à une valeur attendue p_0. Il s'agit de savoir si la valeur observée sur un échantillon est vraisemblable avec $p=p_0$.

Test statistique

Procédure conduisant au vu de l'échantillon à rejeter, avec un certain risque d'erreur α une hypothèse que l'on cherche à tester appelée H_0. La procédure de test est fondée sur une opposition d'hypothèses et on note H_1 l'hypothèse alternative : cela veut dire que l'on risque de rejeter à tort l'hypothèse H_0 avec une probabilité égale à α.

Test bilatéral

Test pour lequel l'hypothèse H_0 est rejetée, si la statistique utilisée prend une valeur en dehors d'un intervalle.

Test unilatéral à droite

Test pour lequel l'hypothèse H_0 est rejetée, si la statistique utilisée prend une valeur supérieure à une valeur.

Test unilatéral à gauche

Test pour lequel l'hypothèse H_0 est rejetée, si la statistique utilisée prend une valeur inférieure à une valeur.

Construction d'un test :

- choix du seuil de risque α,
- choix des hypothèses H_0 et H_1,

par exemple on choisira un test unilatéral à droite si on sait à priori que $p \leq p_0$. On aura alors $H_0 : p=p_0$ et $H_1 : p > p_0$,

- choix d'une variable statistique S servant de variable de décision,
- détermination de la région critique au seuil α,
- énoncé de la règle de décision.

Utilisation du test :

- prélèvement d'un échantillon,
- au vu de la valeur observée s de S, rejeter ou accepter H_0.

Remarques

Le seuil de risque α est toujours petit (α<0.1) : si on demande un test à 95% cela veut dire que le seuil de risque est α=0.05.
N'oubliez pas que lorsque l'on rejette l'hypothèse H_0 cela veut dire que l'hypothèse H_0 risque d'être vraie dans moins de 100*α cas pour 100 cas et que lorsque l'on accepte l'hypothèse H_0 cela veut dire que l'hypothèse H_0 risque d'être vraie dans plus de 100*α cas pour 100 cas.

3.1.1. Étude de la fréquence *p* d'un caractère *X*

Soit une variable aléatoire X qui suit une loi de Bernouilli de paramètre p (on étudie un caractère, si ce caractère est observé alors X=1 et sinon X=0 et on a $Proba(X=1)=p$). Soit X la moyenne des échantillons de taille n : ici, X est égal pour chaque échantillon de taille n à la fréquence observée F du caractère.
Si n est grand ($n \geq 30$), X suit approximativement la loi normale $N(p,\sqrt{p(1-p)/n})$.
Si n est petit, on a (n*X) suit la loi binomiale $B(n,p)$.
On choisit le seuil α et selon les cas :
Test d'hypothèses bilatéral : $H_0 : p = p_0$ et $H_1 : p \neq p_0$
Test d'hypothèses unilatéral à droite (à gauche) : $H_0 : p = p_0$ et $H_1 : p > p_0$ (resp $H_0 : p = p_0$ et $H_1 : p < p_0$)
On calcule, sous l'hypothèse H_0, soit au moyen des tables de la loi normale (pour n grand, $np(1-p)>7$), soit au moyen des tables de la loi binomiale (pour n petit), soit avec Xcas, les bornes de l'intervalle d'acceptation au seuil α, de l'hypothèse H_0.

- dans le cas bilatéral, on cherche les réels a_1 et a_2 vérifiant :
  $Proba(a_1<F=X<a_2)=1-\alpha$
  - pour n grand, on cherche dans une table de loi centrée réduite h tel que $Proba(Y<h)=1-\alpha/2$, on pose
  $F=(Y-p_0)/\sqrt{p_0(1-p_0)/n}$ et on obtient :
  $Proba(F<p_0+h*\sqrt{p_0(1-p_0)/n})=1-\alpha/2$, donc
  $a_1=p_0-h*\sqrt{p_0(1-p_0)/n}$ et $a_2=p_0+h*\sqrt{p_0(1-p_0)/n}$
  On peut aussi taper dans Xcas si α=0.05 :
  a1:=normal_icdf(p0,sqrt((p0*(1-p0)/n),0.025)
  a2:=normal_icdf(p0,sqrt(p0*(1-p0)/n),0.975)
  - pour n petit, $n*F$ suit la loi binomiale $B(n,p_0)$, on cherche dans une table de loi binomiale $B(n,p_0)$ les valeurs $n*p_1$ et $n*p_2$ tels que :
  $Proba(n*p_1<n*F<n*p_2)=1-\alpha$
  et donc :
  $Proba(p_1<F<p_2)==1-\alpha$
  On peut aussi taper dans Xcas si α=0.05 :
  p1:=1/n*binomial_icdf(n,p0,0.025)
  p2:=1/n*binomial_icdf(n,p0,0.975)
- dans le cas unilatéral à droite, on cherche le réel a vérifiant :
  $Proba(F<a)=1-\alpha$:
  - pour n grand, on cherche dans une table de loi centrée réduite h tel que $Proba(Y<h)=1-\alpha$, on a alors $Proba(F<p_0+h\sqrt{p_0(1-p_0)/n})=1-\alpha$ donc
  $a=p_0+h\sqrt{p_0(1-p_0)/n}$
  On peut aussi taper dans Xcas si α=0.05 :

a:=normal_icdf(p0,sqrt(p0*(1-p0)/n),0.975)
- pour n petit, on cherche $n*p_2$ tel que :
$Proba(n*F<n*p_2)=1-\alpha$
on a donc $Proba(F<p_2=1-\alpha$
On peut aussi taper dans Xcas si α=0.05 :
p2:=1/n*binomial_icdf(n,p0,0.975)

- dans le cas unilatéral à gauche, on cherche le réel b vérifiant :
$Proba(F<b)=\alpha$:
- pour n grand, on cherche dans une table de loi centrée réduite h tel que
$Proba(Y<h)=\alpha$, on a alors $Proba(F<p_0+h\sqrt{p_0(1-p_0)/n})=\alpha$ donc
$b=p_0+h\sqrt{p_0(1-p_0)/n}$
On peut aussi taper dans Xcas si α=0.05 :
b:=normal_icdf(p0,sqrt(p0*(1-p0)/n),0.05)
- pour n petit, on cherche $n*p_1$ tel que :
$Proba(n*F< n*p_1)=\alpha$
on a donc :
$Proba(F<p_1)=\alpha$
On peut aussi taper dans Xcas si α=0.05 :
p1:=1/n*binomial_icdf(n,p0,0.05)

Règle de décision :

Soit la fréquence f d'un échantillon de taille n.
On rejette l'hypothèse H_0 au seuil α :

- dans le cas bilatéral
si $f\notin[a_1;a_2]$, (ou si $f\notin[p_1;p_2]$)
- dans le cas unilatéral à droite
si $f>a$, (ou si $f>p_2$)
- dans le cas unilatéral à gauche
si $f<b$, (ou si $f<p_1$)

sinon on accepte l'hypothèse H_0 au seuil α.

Exemple

On choisit n=30, p_0=0.3 et α=0.05 et on compare les résultats de la loi normale et de la loi binomiale.

- dans le cas bilatéral
Pour n=30 et H_0 : p=0.3 $H1$: $p \neq 0.3$ et α=0.05
on a $p_0*(1-p_0)=0.3*0.7=0.21$ et on tape :
normal_icdf(0.3,sqrt(0.21/30),0.975)=0.463982351931
normal_icdf(0.3,sqrt(0.21/30),0.025)=0.136017648069
on pose I=[0.1360;0.464]
ou on tape :
1/30*binomial_icdf(30,0.3,0.025)=2/15
1/30*binomial_icdf(30,0.3,0.975)=7/15
on pose I=[0.1333;0.46666]

On rejette H_0 au seuil de 5%, si la fréquence f obtenue à partir d'un échantillon de taille n=30 est en dehors de l'intervalle I.

- dans le cas unilatéral à droite
  Pour n=30 H_0 :p=0.3 $H1$: $p \geq 0.3$ et α=0.05
  on a p_0*(1−p_0)=0.3*0.7=0.21 et on tape :
  normal_icdf(0.3,sqrt(0.21/30),0.95)=0.437618327917
  on pose I=]−∞;0.464]
  ou on tape :
  1/30*binomial_icdf(30,0.3,0.95)=13/30
  on pose I=]−∞;0.43334]
  On rejette H_0 au seuil de 5%, si la fréquence f obtenue à partir d'un échantillon de taille n=30 est en dehors de l'intervalle I.
- dans le cas unilatéral à gauche
  Pour n=30 H_0 :p=0.3 $H1$: $p \leq 0.3$ et α=0.05
  on a p_0*(1−p_0)=0.3*0.7=0.21 et on tape :
  normal_icdf(0.3,sqrt(0.21/30),0.05)=0.162381672083
  on pose I=[0.16238;+∞[
  ou on tape :
  1/30*binomial_icdf(30,0.3,0.05)=1/6
  on pose I=[0.166667;+∞[
  On rejette H_0 au seuil de 5%, si la fréquence f obtenue à partir d'un échantillon de taille n=30 est en dehors de l'intervalle I.

3.1.2 Étude de la valeur moyenne μ d'un caractère *X*

On va faire des tests d'hypothèses sur μ c'est à dire que dans ce qui suit, on suppose que μ=μ_0, ie que l'on connait μ.

***n* est grand (*n* > 30)**

Théorèmes :

Si $X \in N(\mu,\sigma)$ alors X$\in N(\mu,\sigma/\sqrt{n})$.
Si X suit une loi quelconque et si l'échantillon est de grande taille (n>30), X suit approximativement une loi $N(\mu,\sigma/\sqrt{n})$.
- si l'écart-type σ est connu, on connait la loi suivie par X,
- si l'écart-type σ n'est pas connu, puisque n est grand on va pouvoir estimer σ par $s\sqrt{n}/(n-1)$ où s est l'écart-type d'un échantillon de taille n et on se raméne au cas précedent (σ connu) en prenant $\sigma=s\sqrt{n}/(n-1)$. Ainsi on connait la loi suivie par X : X suit approximativement une loi $N(\mu,s/\sqrt{n-1})$.

Recette quand on connait la loi $N(\mu,\sigma/\sqrt{n})$ suivie par X (σ connu)

On choisit le seuil α et selon les cas :
Test d'hypothèses bilatéral : H_0 :μ =μ_0 et H_1:μ≠ μ_0
Test d'hypothèses unilatéral à droite : H_0 :μ =μ_0 et H_1:μ> μ_0 (resp à gauche : H_0 :μ =μ_0 et H_1:μ< μ_0)
On calcule, au moyen des tables de loi normale (n grand, n>30) les bornes de l'intervalle d'acceptation au seuil α, de l'hypothèse H_0.

- dans le cas bilatéral, on cherche les réels a_1 et a_2 vérifiant :
  $Proba(a_1<X<a_2)=1-\alpha$:
  on cherche dans une table de loi centrée réduite h tel que :
  $Proba(Y<h)=1-\alpha/2$, on a $Proba(X<\mu_0+h*\sigma/\sqrt{n})=1-\alpha/2$, donc
  $a_1=\mu_0-h*\sigma/\sqrt{n}$ et $a_2=\mu_0+h*\sigma/\sqrt{n}$
  Avec Xcas, on tape :
  a1:=normal_icdf(μ_0,$\sigma/\sqrt{n}$,$\alpha/2$)
  a2:=normal_icdf(μ_0,$\sigma/\sqrt{n}$,1-$\alpha/2$)
- dans le cas unilatéral à droite, on cherche le réel a vérifiant :
  $Proba(X<a)=1-\alpha$:
  on cherche dans une table de loi centrée réduite h tel que :
  $Proba(Y<h)=1-\alpha$, on a $Proba(X<\mu_0+h*\sigma/\sqrt{n})=1-\alpha$, donc
  $a=\mu_0+h*\sigma/\sqrt{n}$.
  Avec Xcas, on tape :
  a:=normal_icdf(μ_0,$\sigma/\sqrt{n}$,1-α)
- dans le cas unilatéral à gauche, on cherche le réel b vérifiant :
  $Proba(X<b)=\alpha$:
  on cherche dans une table de loi centrée réduite h tel que :
  $Proba(Y<h)=\alpha$, on a alors $Proba(X<\mu_0+h*\sigma/\sqrt{n})=\alpha$, donc
  $b=\mu_0+h*\sigma/\sqrt{n}$.
  Avec Xcas, on tape :
  b:=normal_icdf(μ_0,$\sigma/\sqrt{n}$,α)

Règle de décision :

Soit m la moyenne d'un échantillon de taille n.
On rejette l'hypothèse H_0 au seuil α :

- dans le cas bilatéral
  si $m \notin [a_1;a_2]$,
- dans le cas unilatéral à droite
  si $m>a$,
- dans le cas unilatéral à gauche
  si $m<b$,

sinon on accepte l'hypothèse H_0 au seuil α.

***X* ∈ *N*(μ,σ) et *n* est petit, (*n*≤30)**

On a deux cas selon que l'écart-type σ est connu ou pas :
- si l'écart-type σ est connu
On sait que si $X \in N(\mu,\sigma)$ alors $X \in N(\mu,\sigma/\sqrt{n})$ on se reportera à la "Recette quand on connait la loi $N(\mu,\sigma/\sqrt{n})$ suivie par X" écrite ci-dessus.
- si l'écart-type σ est inconnu
Lorsque n est petit, on ne peut plus approcher σ par $s\sqrt{n}/(n-1)$ où s est l'écart-type d'un échantillon de taille n.
C'est pourquoi, lorsque n est petit et que $X \in N(\mu,\sigma)$, on utilise la statistique :
$T=\sqrt{n}-1(X-\mu_0/S)$ où $S^2=1/n \sum_{j=1}^{n}(X_j-X)^2$.
T suit une loi de Student à $n-1$ degrés de liberté et T ne dépend pas de σ.

Recette quand on ne connait pas la loi suivie par X

On est dans le cas où σ est inconnu, $X \in N(\mu,\sigma)$ et n est petit.
On choisit le seuil α et selon les cas :
Test d'hypothèses bilatéral : $H_0 : \mu = \mu_0$ et $H_1 : \mu \neq \mu_0$
Test d'hypothèses unilatéral à droite : $H_0 : \mu = \mu_0$ et $H_1 : \mu > \mu_0$ (resp à gauche : $H_0 : \mu = \mu_0$ et $H_1 : \mu < \mu_0$).
Au moyen des tables de la loi de Student (n petit, $n \leq 30$)

- dans le cas bilatéral, on cherche le nombre réel h vérifiant :
  $Proba(-h<T_{n-1}<+h)=1-\alpha$
  Avec Xcas, on tape si α=0.05 :
  h:=student_icdf(n-1,0.975)
- dans le cas unilatéral à droite, on cherche le nombre réel h_1 vérifiant :
  $Proba(T_{n-1}<h_1)=1-\alpha$
  Avec Xcas, on tape si α=0.05 :
  h1:=student_icdf(n-1,0.95)
- dans le cas unilatéral à gauche, on cherche le nombre réel h_2 vérifiant :
  $Proba(T_{n-1}<h_2)=\alpha$
  Avec Xcas, on tape si α=0.05 :
  h2:=student_icdf(n-1,0.05)

Règle de décision :

Soit t la valeur prise par T par un échantillon de taille n : $t=\sqrt{n-1}(m-\mu_0/s)$ où m est la moyenne de l'échantillon et s son écart-type.
On rejette l'hypothèse H_0 au seuil α :

- dans le cas bilatéral
  si $t \notin [-h;+h]$,
- dans le cas unilatéral à droite
  si $t>h_1$,
- dans le cas unilatéral à gauche
  si $t<h_2$,

sinon on accepte l'hypothèse H_0 au seuil α.

***X* ne suit pas une loi normale et *n* est petit**

On ne sait pas faire...

3.1.3 Étude de l'écart-type σ de $X \in N(\mu,\sigma)$

On sait que si X suit une loi normale $N(\mu,\sigma)$, les statistiques :
$Z^2=1/n\sum_{j=1}^{n}(X_j-\mu)^2$ et
$S^2=1/n\sum_{j=1}^{n}(X_j-X)^2$
sont des estimateurs de σ, de plus Z^2 et $n/n-1S^2$ sont des estimateurs sans biais de σ, car on a $E(Z^2)=E(n/n-1S^2)=\sigma$ et S^2 ne dépend pas de μ.
On sait que :
la statistique nZ^2/σ^2 suit une loi du χ^2 à n degrés de liberté et que

la statistique nS^2/σ^2 suit une loi du χ^2 à $(n-1)$ degrés de liberté.
Lorsque μ est connue, on utilisera la statistique nZ^2/σ^2 comme variable de décision, et si μ n'est pas connue, on utilisera la statistique nS^2/σ^2 comme variable de décision.

Recette quand *X* suit une loi normale *N*(μ,σ)

On choisit le seuil α et selon les cas :
Test d'hypothèses bilatéral : H_0 :σ =σ_0 et H_1:σ ≠ σ_0,
Test d'hypothèses unilatéral à droite : H_0 :σ =σ_0 et H_1:σ > σ_0 (resp à gauche : H_0 :σ =σ_0 et H_1:σ < σ_0).
On calcule au moyen des tables de $\chi^2(n)$ les nombres réels h_1 et h_2 vérifiant :

- dans le cas bilatéral
  - si la valeur moyenne μ est connue
  Proba(χ_n^2<h_1)=1−α/2
  Proba(χ_n^2<h_2)=α/2
  Avec Xcas, on tape si α=0.05 :
  h1:=chisquare_icdf(n,0.975)
  h2:=chisquare_icdf(n,0.025)
  - si la valeur moyenne μ n'est pas connue
  Proba(χ_{n-1}^2<h_1)=1−α/2
  Proba(χ_{n-1}^2<h_2)=α/2
  Avec Xcas, on tape si α=0.05:
  h1:=chisquare_icdf(n-1,0.975)
  h2:=chisquare_icdf(n-1,0.025)
- dans le cas unilatéral à droite
  - si la valeur moyenne μ est connue
  Proba(χ_n^2<h_1)=1−α
  Avec Xcas, on tape si α=0.05 :
  h1:=chisquare_icdf(n,0.95)
  - si la valeur moyenne μ n'est pas connue
  Proba(χ_{n-1}^2<h_1)=1−α
  Avec Xcas, on tape :
  h1:=chisquare_icdf(n-1,0.95)
- dans le cas unilatéral à gauche
  - si la valeur moyenne μ est connue
  Proba(χ_n^2<h_2)=α
  Avec Xcas, on tape si α=0.05 :
  h2:=chisquare_icdf(n,0.05)
  - si la valeur moyenne μ n'est pas connue
  Proba(χ_{n-1}^2<h_2)=α
  Avec Xcas, on tape si α=0.05 :
  h2:=chisquare_icdf(n-1,0.05)

Règle de décision :

Soit *u* la valeur prise par nZ^2/σ^2 (ou par nS^2/σ^2 si μ n'est pas connue) pour un échantillon de taille *n* :
- si μ est connue, on calcule $u=\sum_{j=0}^{n}(x_j-\mu)^2/\sigma_0^2$ où les x_j sont les valeurs de l'échantillon (car selon H_0 :σ =σ_0).

- si μ n'est pas connue, on calcule $u=n*s^2/\sigma_0^2$ où s est l'écart-type de l'échantillon (car selon H_0 :$\sigma=\sigma_0$).
On rejette l'hypothèse H_0 :$\sigma=\sigma_0$ au seuil α :

- dans le cas bilatéral
  si $u\notin[h_2;h_1]$,
- dans le cas unilatéral à droite
  si $u>h_1$,
- dans le cas unilatéral à gauche
  si $u<h_2$,

sinon on accepte l'hypothèse H_0 au seuil α.

3.2. Les intervalles de confiance

L'estimation a pour but, à partir d'échantillons, de donner des valeurs numériques aux paramètres de la population dont ces échantillons sont issus.
Il peut s'agir d'estimation ponctuelle ou d'estimation par intervalle.
Un intervalle de confiance I_α au seuil α, pour le paramètre p_0, est un intervalle qui contient p_0 avec une confiance de 1−α, cela veut dire que pour un grand nombre n d'échantillons environ $n*\alpha$ des I_α ne contiennent pas p_0 (en effet les intervalles de confiance I_α dépendent de l'échantillon) Remarques Le seuil de risque α est toujours petit (α<0.1) : si on vous demande un intervalle de confiance à 95% cela veut dire que le seuil de risque est α=0.05.
N'oubliez pas que l'estimation d'une valeur par un intervalle de confiance comporte un risque, celui de situer la valeur dans un intervalle où elle ne se trouve pas !!!! (c'est α qui determine le risque d'erreur)
Plus on demande un risque faible et plus l'intervalle de confiance est grand.

3.2.1 Valeur de la fréquence *p* d'un caractère *X*

Estimation ponctuelle

Lorsque la taille n de l'échantillon est grande, on prend comme estimation ponctuelle de p la fréquence f observée sur l'échantillon.

Remarque

Cela ne donne aucune information sur la qualité de l'estimation.

Estimation par un intervalle

Cas des échantillons de taille *n*>30

Soit X une variable aléatoire de Bernouilli de paramètre p (X vaut 0 ou 1 et $Proba(X=1)=p$).
Soit X la variable aléatoire égale à la moyenne des valeurs prises par X pour des échantillons de taille n. On a X=F est égal à la fréquence du nombre d'apparitions de la valeur 1 pour chaque échantillon de taille n, .
On sait que $n*F$ suit une loi binomiale $B(n,p)$, cette loi est proche de la loi normale

$N(np,\sqrt{np}(1-p))$ car n est grand ($n>30$).
On peut donc considérer que F suit approximativement la loi $N(p,\sqrt{p}(1-p)/n)$.

Recette

- On choisit α (par exemple α=0.05),
- On cherche à l'aide d'une table de loi normale centrée réduite, h vérifiant :
$Proba(Y<h)=1-\alpha/2$ pour $Y\in N(0,1)$.
On a donc en posant $Y=F-p/\sqrt{p}(1-p)/n$:
$Proba(p-h\sqrt{p}(1-p)/n<F<p+h\sqrt{p}(1-p)/n)=1-\alpha$
- On calcule la valeur f de F pour l'échantillon
On a donc $n(f-p)^2<h^2p(1-p)$ c'est à dire $(h^2+n)p^2-p(h^2+2nf)+nf^2<0$ donc p se trouve à l'intérieur des racines de l'équation du second degré :
$(h^2+n)x^2-x(h^2+2nf)+nf^2=0$ que l'on peut résoudre (calcul du discriminant $\Delta=h^4+(-(4*h^2))*n*f^2+4*h^2*n*f$ etc...)
mais il est plus simple de dire, que l'on peut estimer l'écart-type de $n*F$. On a $\sigma(n*F)=\sqrt{np}(1-p)$ que l'on peut estimer par $\sqrt{nf}(1-f)\sqrt{n}/n-1$.
Donc l'écart-type de X=F, $\sigma(F)=\sigma(X)=\sqrt{p}(1-p)/n$ peut être estimé par $1/n\sqrt{nf}(1-f)\sqrt{n}/n-1=\sqrt{f}(1-f)/n-1$, donc on a :
$Proba(p-h\sqrt{f}(1-f)/n-1\leq f\leq p+h\sqrt{f}(1-f)/n-1)=1-\alpha$
ou encore
$Proba(f-h\sqrt{f}(1-f)/n-1\leq p\leq f+h\sqrt{f}(1-f)/n-1)=1-\alpha$
Si $a_1=f-h\sqrt{f}(1-f)/n-1$ et $a_2=f+h\sqrt{f}(1-f)/n-1$ on a $a_1\leq p\leq a_2$
Avec Xcas, on tape si α=0.05 :
a1:=normal_icdf(f,sqrt(f*(1-f)/(n-1),0.025)
a2:=normal_icdf(f,sqrt(f*(1-f)/)n-1,0.975)

Résultat $I_\alpha=[a_1\ ;\ a_2]$ est un intervalle de confiance de p au seuil α.

Cas des échantillons de taille $n\leq 30$

Soit X une variable aléatoire de Bernouilli de paramètre p (X vaut 0 ou 1 et $Proba(X=1)=p$).
Soit la variable aléatoire F=X.
On sait que nF suit une loi binomiale $B(n,p)$. On utilisera donc une table de la loi binomiale.

Recette

- On choisit α (par exemple α=0.05)
- On calcule la valeur f de F pour l'échantillon
- On approche p par f, ainsi $n*F=n*X\in B(n,f)$, on cherche $n*p_1$ et $n*p_2$ à l'aide d'une table de loi binomiale pour avoir :
$Proba(n*F<n*p_1)=1-\alpha/2$ et $Proba(n*F<n*p_2)=\alpha/2$
Avec Xcas, on tape si α=0.05 :
p1:=1/n*binomial_icdf(,n,f,0.025)
p2:=1/n*binomial_icdf(n,f,0.975)
On a donc :
$Proba(p_2<f<p_1)=1-\alpha$.

Résultat

$I_\alpha=[p_2 ; p_1]$ est un intervalle de confiance de p au seuil α.

3.2.2. Valeur moyenne μ d'un caractère *X*

Estimation ponctuelle

Lorsque la taille n de l'échantillon est grande, on prend comme estimation ponctuelle de μ la moyenne m observée sur l'échantillon.

Remarque

Cela ne donne aucune information sur la qualité de l'estimation.

Estimation par un intervalle

Cas des échantillons de taille $n>30$

Si n est grand ($n>30$), on connait la loi suivie par X : X suit approximativement une loi $N(\mu,\sigma/\sqrt{n})$ (ou si σ n'est pas connu X suit approximativement une loi $N(\mu,s/\sqrt{n-1})$).

Recette lorsque la loi $N(\mu,\sigma/\sqrt{n})$ suivie par X est connue

- On choisit α (par exemple α=0.05).
- On calcule la valeur m de X pour l'échantillon (ie sa moyenne) et si σ n'est pas connu, l'écart-type s de l'échantillon.
- On cherche h, dans une table de loi normale centrée réduite, pour avoir :

$Proba(Y<h)=1-\alpha/2$ pour $Y\in N(0,1)$ on a alors :
$Proba(\mu-h*\sigma/\sqrt{n}<m<\mu+h*\sigma/\sqrt{n})=1-\alpha$
on a donc :
$Proba(m-h*\sigma/\sqrt{n}<\mu<m+h*\sigma/\sqrt{n})=1-\alpha$
ou si σ n'est pas connu :
$Proba(\mu-h*s/\sqrt{n-1}<X<\mu+h*s/\sqrt{n-1})=1-\alpha$
on a donc $Proba(m-h*s/\sqrt{n-1}<\mu<m+h*\sigma/\sqrt{n-1})=1-\alpha$.
Si σ est connu on pose :
$a_1=m-h*\sigma/\sqrt{n}$ et $a_2=m+h*\sigma/\sqrt{n}$
ou si σ n'est pas connu on pose :
$a_1=m-h*s/\sqrt{n-1}$ et $a_2=m+h*s/\sqrt{n-1}$
on a $a_1 \le \mu \le a_2$
Avec Xcas, si σ est connu, on tape si α=0.05 :
a1:=normal_icdf(m,σ/sqrt(n),0.025)
a2:=normal_icdf(m,σ/sqrt(n),0.975)
ou si σ n'est pas connu, on tape si α=0.05 :
a1:=normal_icdf(m,s/sqrt(n-1),0.025)
a2:=normal_icdf(m,s/sqrt(n-1),0.975)

Résultat

$I_\alpha=[a_1\ ;\ a_2]$ est un intervalle de confiance de μ au seuil α.

Cas des petits échantillons issus d'une loi normale

Si σ est connu, la loi $N(\mu,\sigma/\sqrt{n})$ suivie par X est connue et on se reportera donc à la recette du paragraphe précédent.
Si σ n'est pas connu, on note $S^2=1/n\sum_{j=1}^{n}(X_j-X)^2$ alors
$T=(X-\mu/S)\sqrt{n-1}$ suit une loi de Student à $(n-1)$ degrés de liberté.

Recette lorsque *n* est petit et $X \in N(\mu,\sigma)$

- On choisit α (par exemple α=0.05).
- On calcule la valeur *m* de X pour l'échantillon (*m* est la moyenne de l'échantillon) et l'écart-type *s* de l'échantillon (s^2 est la valeur de S^2 pour l'échantillon).
- On cherche *h*, dans une table de Student pour $(n-1)$ degrés de liberté, pour avoir :
$Proba(-h<T_{n-1}<h)=Proba(-h<(X-\mu/S)\sqrt{n-1}<h)=1-\alpha$
Avec Xcas, on tape si α=0.05 :
h:=student_icdf(n-1,0.975)
puisque *m* est la valeur de X et *s* la valeur de *S* pour l'échantillon on a :
$Proba(m-hs/\sqrt{n-1}<\mu<m+hs/\sqrt{n-1})=1-\alpha$.

Résultat

$I_\alpha=[m-hs/\sqrt{n-1};m+hs/\sqrt{n-1}]$ est un intervalle de confiance de μ au seuil α.

Exemple

Pour obtenir un intervalle de confiance de μ au risque α=0.05 et $n-1=4$ on tape :
h:=student_icdf(4,1-0.05/2)
on obtient :
h=2.7764451052≃ 2.776 donc :
$m-hs/\sqrt{4}<\mu<m+hs/\sqrt{4}$.
On prend un échantillon d'effectif $n=5$ $(4=n-1)$, pour lequel on trouve :
$m=0.484342422505$ et $s=0.112665383246$
On tape :
m:=0.484342422505
s:=0.112665383246
m+h*s/sqrt(4)
On obtient :
0.64072197445
On tape :
m-hs/sqrt(4)
0.32796287056.
donc un intervalle de confiance de μ au risque 0.05 est :
[0.32796287056; 0.64072197445]

3.2.3 Valeur de l'écart-type σ de $X\in N(\mu,\sigma)$

Estimation ponctuelle

Lorsque la taille n de l'échantillon est grande, on prend comme estimation ponctuelle de σ, $s\sqrt{n}/n-1$, où s est l'écart-type de l'échantillon.
bf Remarque
Cela ne donne aucune information sur la qualité de l'estimation.

Estimation par un intervalle

Cas où μ est connue

On pose $Z^2=1/n\sum_{j=1}^{n}(X_j-\mu)^2$. Alors nZ^2/σ^2 suit une loi du χ^2 à n degrés de liberté.

Recette lorsque μ est connue et $X \in N(\mu,\sigma)$

- On choisit α (par exemple α=0.05).
- On calcule la valeur $z^2=1/n\sum_{j=1}^{n}(x_j-\mu)^2$ de Z^2 pour les valeurs x_j de l'échantillon.
- On cherche t_1 et t_2, dans une table du χ^2 pour n degrés de liberté, pour avoir :
$Proba(\chi_n^2<t_1)=Proba(nZ^2/\sigma^2<t_1)=1-\alpha/2$ et
$Proba(\chi_n^2<t_2)=Proba(nZ^2/\sigma^2<t_2)=\alpha/2$
Avec Xcas, on tape si α=0.05 :
t1=chisquare_icdf(n,0.975)
t2=chisquare_icdf(n,0.025)
on a donc $Proba(t_2<nZ^2/\sigma^2<t_1)=1-\alpha$.
et puisque z^2 est la valeur de Z^2 pour l'échantillon on a:
$Proba(nz^2/t_1<\sigma^2<nz^2/t_2)=1-\alpha$.

Résultat

$I_\alpha=[\sqrt{nz^2/t_1}\ ;\ \sqrt{nz^2/t_2}]$ est un intervalle de confiance de σ au seuil α.

Cas où μ n'est pas connue

On pose $S^2=1/n\sum_{j=1}^{n}(X_j-X)^2$.
Alors, nS^2/σ^2 suit une loi du χ^2 à $n-1$ degrés de liberté.

Recette si μ n'est pas connue et $X \in N(\mu,\sigma)$

- On choisit α (par exemple α=0.05).
- On calcule la valeur m de X pour l'échantillon (m est la moyenne de l'échantillon) et l'écart-type s de l'échantillon (s^2 est la valeur de S^2 pour l'échantillon.
- On cherche t_1 et t_2, dans une table du χ^2 pour $(n-1)$ degrés de liberté, pour avoir :
$Proba(\chi_{n-1}^2<t_1)=Proba(nS^2/\sigma^2<t_1)=1-\alpha/2$ et
$Proba(\chi_{n-1}^2<t_2)=Proba(nS^2/\sigma^2<t_2)=\alpha/2$
Avec Xcas, on tape si α=0.05 :
t1=chisquare_icdf(n-1,0.975)
t2=chisquare_icdf(n-1,0.025)
on a donc $Proba(t_2<nS^2/\sigma^2<t_1)=1-\alpha$ et puisque s^2 est la valeur de S^2 pour l'échantillon on a :
$Proba(ns^2/t_1<\sigma^2<ns^2/t_2)=1-\alpha$.

Résultat

$I_\alpha=[s\sqrt{n}/t_1\ ;\ s\sqrt{n}/t_2]$ est un intervalle de confiance de σ au seuil α.

3.3. Un exemple

On a effectué 10 pesées indépendantes sur une balance d'une même masse μ et on a obtenu :
10.008,10.012,9.990,9.998,9.995,10.001,9.996,9.989,10.000,10.015
Avec Xcas on a facilement la moyenne *m*, l'écart-type *s* et la variance de l'échantillon.
On tape :
L:=[10.008,10.012,9.990,9.998,9.995,10.001,9.996,9.989,
10.000,10.015]
m=mean(L)=10.0004
s=stddev(L)=0.00835703296719
variance(L)=6.98400000147e-05
On a donc:
$s^2 \simeq 0{,}00007$

3.3.1 σ=0.01 et μ est inconnu

On suppose que μ est inconnue mais que la balance est telle que l'erreur de mesure a un écart-type σ de 0.01.
On cherche à déterminer μ au vu de l'échantillon.

H_0 : μ=10 et H_1 : μ>10 au seuil de 5%

On veut tester les hypothèses H_0 : μ=10 et H_1 : μ>10

Règle :

On calcule la moyenne *m* de l'échantillon : on a trouvé *m*=10.004.
On détermine *a* pour avoir *Proba*(X<*a*)=0.95.
Au seuil de 5%, on rejette l'hypothèse unilatérale à droite H_0 si *m*>*a* sinon on accepte H_0 : μ=10.
Si on suppose que le résultat de la mesure est une variable aléatoire *X* qui suit une loi normale N(μ,0.01), alors X suit une loi normale N(μ,0.01/√10).
Donc avec l'hypothèse H_0 : μ=10 on a
X ∈ N(10,0.00316) et *Y*=X−10/0.00316 ∈ N(0,1)
Avec une table de loi normale centrée réduite on cherche *h* pour que :
Proba(*Y*<*h*)=0.95 lorsque *Y*∈ N(0,1) et on trouve *h*=1.64.
On a donc *Proba*((X−10)/0.00316<1.64)=0.95.
On calcule (*m*−10)/0.00316=0.126582278481 et 0.126582278481<1.64 donc on accepte l'hypothèse H_0 μ=10 au seuil de 5%.
Avec Xcas on tape :
a:=normal_icdf(10,0.01/sqrt(10),0.95)
On obtient :
a=10.0051824
Puisque *m*=10.0004<*a* on accepte l'hypothèse H_0 : μ=10.

H_0 : μ=10 et H_1 : μ≠ 10 au seuil de 5%

On veut tester les hypothèses H_0 : μ=10 et H_1 : μ ≠ 10.

Règle :

On calcule la moyenne *m* de l'échantillon : on a trouvé *m*=10.004.
On détermine *a* pour avoir *Proba*(a_1<X<a_2)=0.95.
Au seuil de 5%, si a_1<*m*<a_2, on accepte l'hypothèse bilatérale H_0 : μ=10 et sinon on la rejette.
Avec une table de loi normale centrée réduite on cherche *h* pour que :
Proba(*Y*<*h*)=0.975 lorsque *Y*∈ *N*(0,1) et on trouve *h*=1.96.
On a aussi *Proba*(*Y*<−*h*)=0.025 et donc *Proba*(−*h*<*Y*<*h*)=0.95.
Si on suppose que le résultat de la mesure est une variable aléatoire *X* qui suit une loi normale *N*(μ,0.01), alors X suit une loi normale *N*(μ,0.01/√10).
On a donc *Proba*(|X−10|/0.00316<*h*)=0.95 soit
Proba(|X−10|<1.96*0.00316)=0.95.
Puisque 1.96*0.00316=0.0061936 et que |*m*−10|=0.0004<0.0061936 on accepte l'hypothèse H_0 au seuil de 5%.
Avec Xcas on tape :
a_1:=*normal_icdf*(10,0.01/√10,0.025)
a_2:=*normal_icdf*(10,0.01/√10,0.975)
On obtient :
a_1=9.99380204968
a_2=10.0061979503
Puisque a_1<*m*=10.0004<a_2 on accepte l'hypothèse H_0 : μ=10 au seuil de vraisemblance de 5%.

Intervalle de confiance de μ au seuil de 5%

On veut avoir une estimation de μ au seuil de 5%.
On a trouvé précédement que X ∈ *N*(10,0.00316) :
Proba(|X−μ|<1.96*0.00316)=0.95.
Pour l'échantillon considéré la valeur de X est égale à *m* d'où,
Proba(|*m*−μ|<1.96*0.00316)=0.95
Un intervalle de confiance de μ au seuil de 5% est donc :
|μ−10.0004|<0.0062 c'est à dire [9.9942;10.0066] est un intervalle de confiance de μ au seuil de 5%.
Avec Xcas on tape :
a1:=normal_icdf(10,0.01/sqrt(10),0.025)
a2:=normal_icdf(10,0.01/sqrt(10),0.975)
On obtient :
a1=9.99380204968
a2=10.0061979503
Donc [a_1;a_2] est un intervalle de confiance de μ au seuil de 5%.

3.3.2 μ=10 et σ est inconnu

Mainenant, on ne connait pas la précision de la balance mais on a une masse μ=10 et on voudrait déterminer la précision de la balance au vue de l'échantillon des 10 pesées sauvées dans L :
L:=[10.008,10.012,9.990,9.998,9.995,10.001,9.996,9.989,

10.000,10.015]
On sait que μ=10.
On pose $Z^2=1/n\sum_{k=1}^{n}(X_k-\mu)^2$.
On calcule la valeur z^2 de Z^2, par exemple, avec Xcas on tape :
L10:=makelist(10,0,9) (L10 est une liste de longueur 10 dont tous les éléments sont égaux à 10).
Lc:=L-L10
z2:=mean(Lc^2) on obtient z2=0.00007
donc $z^2\simeq 0.00007$

H_0 : σ=0.005 et H_1 : σ>0.005 au seuil de 5%

On veut tester les hypothèses H_0 : σ=0.005 et H_1 : σ>0.005.
$10*Z^2/\sigma^2$ suit une loi du χ^2 ayant 10 degrés de liberté.

Règle :

On accepte au seuil de 5%, l'hypothèse unilatérale à droite σ=0.005, si $z^2<a$ lorsque a vérifie :
$Proba(10*Z^2/0.005^2<10*a/0.005^2)=0.95$.
D'après les tables du χ^2 on trouve :
$Proba(\chi_{10}^2>18.307)=0.005$ donc
$a=18.307*0.005^2/10=0.0000457$.
Avec Xcas on tape :
h:=chisquare_icdf(10,0.95)
On obtient :
h:=18.3070380533
donc $h\simeq 18.307$
a:=h*0.005^2/10
donc $a\simeq 0.0000457$
Puisque $z^2=0.00007>a=0.0000457$, on ne peut pas accepter l'hypothèse H_0 : σ=0.005 au seuil de 5%.

H_0 : σ=0.005 et H_1 : σ≠ 0.005 au seuil de 5%

On veut tester les hypothèses H_0 : σ=0.005 et H_1 : σ ≠ 0.005.
$10*Z^2/\sigma^2$ suit une loi du χ^2 ayant 10 degrés de liberté.

Règle :

On accepte à un niveau de 5%, l'hypothèse bilatérale σ=0.005, si $b<Z^2<a$ lorsque a et b vérifient :
$Proba(10*b/0.005^2<10*Z^2/0.005^2<10*a/0.005^2)=0.95$.
D'après les tables on trouve :
$Proba(\chi_{10}^2<3.25)=0.025$ et
$Proba(\chi_{10}^2>20.5)=0.025$
Donc $a=20.5*0.005^2/10=0.00005125$ et $b=3.25*0.005^2/10=8.125e-06$.
Avec Xcas on tape :
h1:=chisquare_icdf(10,0.025)
On obtient :

h1:=3.24697278024
donc $h_1 \simeq 3.25$
On tape :
h2:=chisquare_icdf(10,0.975)
On obtient :
h2:=20.4831773508
donc $h_2 \simeq 20.5$
On tape :
b:=h1*0.005^2/10
On obtient :
8.125e-06
On tape :
a:=h2*0.005^2/10 On obtient :
5.125e-05
Puisque $z^2=0.00007>a=0.00005125$, on ne peut donc pas accepter l'hypothèse H_0 $\sigma=0.005$ au seuil de 5%.

Intervalle de confiance de σ au seuil de 5%

On veut avoir une estimation de σ au seuil de 5%.
On sait que $10*Z^2/\sigma^2$ suit une loi du χ^2 ayant 10 degrés de liberté.
On a vu précédemment (en 3.7.2) que $h_1=3.25$ et $h_2=20.5$ et donc que :
$Proba(3.25<10*Z^2/\sigma^2<20.5)=0.95$ donc,
$Proba(10*Z^2/20.5<\sigma^2<10*Z^2/3.25)=0.95$
On a $z^2=0.00007=z2$, donc $10*z^2=0.0007$.
On a alors :
$0.0007/20.5=3.41463414634e{-}05<\sigma^2<0.0007/3.25=0.000215384615385$
donc [0.000034;0.000216] est un intervalle de confiance de σ^2 au seuil de 5%,
donc [0.0058;0.0147] est un intervalle de confiance de σ au seuil de 5%.
Avec Xcas on tape :
h1:=3.25
h2:=20.5
a1:=sqrt(10*z2/h2)
On obtient :
0.0058
On tape :
a2:=sqrt(10*z2/h1)
On obtient :
0.0147
c'est à dire [$a1$; $a2$] est un intervalle de confiance de σ au seuil de 5%.

3.3.3 μ=10 et σ sont inconnus

Mainenant, on ne connait ni le poids μ de la masse, ni la précision σ de la balance et on voudrait determiner μ et σ au vue de l'échantillon.
La valeur de X pour l'échantillon est m=10.0004 et la valeur de S pour l'échantillon est s=0.00835703296719.
On peut estimer grossièrement μ par 10.0004, mais n est trop petit pour que cela soit fiable.
Lorsqu'on connait σ, on peut ici, utiliser X, pour étudier μ car on sait que X suit une

loi normale $N(\mu,\sigma/\sqrt{n})$ car on a supposé que X suit une loi normale $N(\mu,\sigma)$: on va donc essayer d'avoir des renseignements sur σ.

Intervalle de confiance de σ au seuil de 5%

On veut avoir une estimation de σ au seuil de 5%.
Un estimateur sans biais de σ^2 est $nS^2/(n-1)$ mais on ne peut pas estimer σ par $\sqrt{n}*s^2/(n-1)=stdDev(L)=0.00880908621914$ car n est trop petit.
Cherchons un intervalle de confiance pour σ au seuil de 5%.
On sait que la variable statistique $nS^2/\sigma^2=10S^2/\sigma^2$ suit une loi du χ^2 ayant 9 degrés de liberté (9=(n−1), car l'échantillon est de taille n=10 et on enlève 1 ,car on utilise la moyenne de l'échantillon pour calculer S^2).
Cette variable ne dépend pas de μ.
D'après les tables du χ^2 on trouve :
$Proba(\chi_9^2<2.70)=0.025$ et
$Proba(\chi_9^2>19.02)=0.025$
Avec Xcas on tape :
a1:=chisquare_icdf(9,0.025)
On obtient :
2.70038949998
donc $a_1\simeq 2.70$
a2:=chisquare_icdf(9,0.975)
On obtient :
19.0227677986
donc $a_2\simeq 19.02$
Donc $Proba(2.70<10S^2/\sigma^2<19.02)=0.95$
Pour l'échantillon $10S^2=10s^2=6.98400000147e-04$ donc
$(6.98400000147e-04)/19.02<\sigma^2<(6.98400000147e-04)/2.70$
$3.67192429099e-05<\sigma^2<0.000258666666721$
On a :
$\sqrt{3.67192429099e-05}=0.00605964049345$ et
$\sqrt{0.000258666666721}=0.0160831174441$.
Donc [0.0060 ; 0.0161] est un intervalle de confiance pour σ au seuil de 5%.

Tests d'hypothèses pour μ

On va faire différents tests d'hypothèses pour μ.
Comme l'intervalle de confiance pour σ au seuil de 5% ne donne pas σ avec une grande précision on va utiliser la loi de Student pour avoir des renseignements sur μ.
La variable statistique $T=\sqrt{n-1}(X-\mu/S)$ suit une loi de Student à (n−1) degrés de liberté. Cette variable ne dépend pas de σ.
- On teste H_0 : μ=10 et H_1 : μ > 10 au seuil de 5%
On veut tester les hypothèses, H_0 : μ=10 et H_1 : μ > 10 au seuil de 5%.

Règle :

Si la valeur t de T pour l'échantillon est telle que $t<a$ pour a défini par :
$Proba(T<a)=0.95$
on accepte l'hypothèse unilatérale à droite H_0 (μ=10) au seuil de 5%.
On lit dans la table de Student que :

Proba(T_9<1.833)=0.95.
Avec Xcas on tape :
a:=student_icdf(9,0.95)
On obtient :
1.83311293265
donc $a \simeq 1.833$
On calcule $t=\sqrt{n-1}(m-\mu/s)=\sqrt{9}(10.0004-10)/0.00835703296719=0.143591631708$
Puisque 0.143<1.833 on accepte l'hypothèse unilatérale à droite H_0 : μ=10 au seuil de 5%.
- On teste H_0 : μ=10 et H_1 : μ ≠ 10 au seuil de 5%

Règle :

On lit dans la table de Student que :
Proba($|T_9|$<2.262)=0.975.
Avec Xcas on tape :
a:=student_icdf(9,0.975)
On obtient :
a:=2.2621571628
Donc $a \simeq 2.262$
On vérifie que si b:=student_icdf(9,0.025)=-2.2621571628
on a b=-a.
Donc *Proba*($|T_9|$<2.262)=0.95.
Puisque t=0.143<2.262 on accepte l'hypothèse bilatérale H_0 : μ=10 au seuil de 5%.

Intervalle de confiance de μ au seuil de 5%

On veut avoir une estimation de μ au seuil de 5%.
On lit dans la table de Student que :
Proba($|T_9|$<2.262)=0.975.
Avec Xcas on tape :
a:=student_icdf(9,0.975)
On obtient :
a:=2.2621571628
Donc $a \simeq 2.262$
On a donc :
$|t|=\sqrt{n-1}(|m-\mu|/s)=\sqrt{9}|10.0004-\mu|/0.00835703296719<2.262=a$
donc
$9.99409879714=m-as/\sqrt{9}<\mu<m+as/\sqrt{9}=10.0067012029$
Donc [9.994;10.0067] est un intervalle de confiance de μ au seuil de 5%.

Remarque

X suit une loi normale $N(\mu,\sigma/\sqrt{n})$, si on estime σ par la moyenne des bornes de l'intervalle de confiance trouvé on obtient :
(0.0060+0.0161)/2=0.01105 on calcule :
10.0004−1.96*0.01105=9.978742
10.0004+1.96*0.01105=10.022058
Donc *Proba*(|X−μ|<1.96*σ)=0.95 se traduit par :

9.978742<μ<10.022058 au seuil de 5%
ce qui donne une moins bonne éstimation qu'avec l'utilisation de la loi de Student.

3.4. Les tests d'homogénéité

Face à deux séries d'observations c'est à dire à deux échantillons, le problème est de savoir si les différences observées sont dues aux fluctuations de l'échantillonnage ou au fait que les échantillons ne proviennent pas de la même population.

3.4.1 Comparaison de deux fréquences observées

Soient f_1 et f_2 les fréquences observées d'un caractère dont la fréquence théorique est p. Cette observation est faite à partir de deux échantillons de taille respéctive n_1 et n_2.
On veut savoir si les fréquences f_1 et f_2 sont significativement différentes ce qui voudrait dire que les deux échantillons proviennent de deux populations différentes de paramètre p_1 et p_2 ou si au contraire les deux échantillons proviennent d'une même population de paramètre $p=p_1=p_2$.
On veut donc tester l'hypothèse H_0 : $p_1=p_2=p$ contre H_1 : $p_1\neq p_2$ au seuil α.
Soit F_1 (resp F_2) la variable aléatoire égale à la fréquence du caractère pour des échantillons de taille n_1 (resp n_2).
On a sous l'hypothèse H_0:
F_1 a pour moyenne p et comme écart-type $\sqrt{p(1-p)/n_1}$
F_2 a pour moyenne p et comme écart-type $\sqrt{p(1-p)/n_2}$
Si n_1 et n_2 sont très grands on a vu que :
F_1 suit approximativement une loi $N(p,\sqrt{p(1-p)/n_1})$ et
F_2 suit approximativement une loi $N(p,\sqrt{p(1-p)/n_2})$
Donc
F_1-F_2 suit approximativement une loi $\in N(0,\sqrt{p(1-p)/n_1+p(1-p)/n_2})$
On va estimer p grâce à la reunion des deux échantillons :
$p \simeq f=n_1*f_1+n_2*f_2/n_1+n_2$
alors
F_1 a pour moyenne p et comme écart-type $\sqrt{f(1-f)/n_1}$
F_2 a pour moyenne p et comme écart-type $\sqrt{f(1-f)/n_2}$
On pose $s_{12}=\sqrt{f(1-f)/n_1+f(1-f)/n_2}=\sqrt{f(1-f)(n_1+n_2)/n_1n_2}$ donc
$F=F_1-F_2 \in N(0,s_{12})$

Recette

On choisit le seuil α.
Avec une table de loi normale centrée réduite, on cherche, pour $U \in N(0,1)$, h tel que :
$Proba(U \leq h)=1-\alpha/2$.
on a alors :
$Proba(|F_1-F_2|/s_{12}<h)=1-\alpha$.
Avec Xcas on tape si α=0.05 et si $s12=s_{12}$:
a:=normal_icdf(0,s12,1-0.05/2)
On a alors :
$Proba(|F_1-F_2|<a)=1-\alpha$ avec $a=s_{12}*h$.
On calcule selon les cas :
$|f_1-f_2|/s_{12}$ que l'on compare à h ou
$|f_1-f_2|$ que l'on compare à a.

Si $|f_1-f_2|/s_{12}<h$ ou $|f_1-f_2|<a$ on admet que les deux échantillons ne sont pas significativement différents au seuil α, sinon on dira que les deux échantillons ne proviennent pas de la même population (voir aussi l'utilisation de la loi du χ^2).

Exercice

Pour tester l'efficacité d'un vaccin antigrippal on soumet 300 personnes à une expérience :
- sur 100 personnes non vaccinées, 32 sont atteintes par la grippe,
- sur 200 personnes vaccinées, 50 sont atteintes par la grippe,

Ce résultat permet-il d'apprécier l'efficacité du vaccin ?
On a le tableau suivant :

	grippé	non grippé	taille
vacciné	32	68	100
non vacciné	50	150	200
total	82	218	300

On calcule les valeurs *f*1 et *f*2 qui sont les proportions des grippés des deux échantillons on tape :
f1:=32/100
f2:=50/200=25/100
On tape :
f1-f2
On obtient :
7/100
Donc $|f_1-f_2|$==0.07
On calcule la valeur *p* proportion des grippés lorsqu'on reunit les deux échantillons on tape :
p:=82/300
On obtient :
41/150
Donc $p\simeq$ 0.273333333333
On calcule s_{12}, on tape :
s12:=sqrt(p*(1-p)*(1/100+1/200))
On obtient :
sqrt(4469/1500000)
Donc $s_{12}\simeq$ 0.0545832697201
La variable $F=F_1-F_2$ suit la loi normale $N(0,s_{12})$ et sa valeur est f=0.07.
On cherche la valeur *a* qui vérifie :
Proba($|F|>a$)=0.05 ou encore
Proba($F\leq a$)=0.975 et pour cela on tape :
a:=normal_icdf(0,sqrt(4469/1500000),0.975)
On obtient :
0.10698124281
Puisque |f1-f2|=0.07<a=0.10698124281, on en déduit que les deux échantillons ne sont pas significativement différents au seuil de 5% : on peut donc dire que le vaccin n'est pas efficace mais ce n'est pas une certitude...

Remarque

On a h:=normal_icdf(0,1,0.975)=1.95996398454
et |f1-f2|=0.07<h*sqrt(4469/1500000)=0.10698124281

3.4.2 Comparaison de deux moyennes observées

Soient m_1 et m_2 les moyennes observées d'un caractère dont la moyenne théorique est μ. Cette observation est faite à partir de deux échantillons de taille respectives n_1 et n_2. On veut savoir si les moyennes m_1 et m_2 sont significativement différentes ce qui voudrait dire que les deux échantillons proviennent de deux populations différentes de moyenne μ_1 et μ_2 ou si au contraire les deux échantillons proviennent d'une même population ou de populations de même moyenne $\mu=\mu_1=\mu_2$.
Soient deux caractères normaux indépendants X et Y distribués respectivement selon les lois $N(\mu_1,\sigma(X))$ et $N(\mu_2,\sigma(Y))$,
On veut donc tester l'hypothèse H_0 : $\mu_1=\mu_2=\mu$ contre H_1 : $\mu_1\neq\mu_2$ au seuil α.
Soient deux échantillons considérés l'un comme échantillon du caractère X et l'autre comme échantillon du caractère Y, de taille respective n_1 et n_2 de moyenne respective m_1 et m_2 et d'écart-type respectif s_1 et s_2.
Soit X (resp $\bar{Y}$) la variable aléatoire égale à la moyenne du caractère X (resp Y) pour des échantillons de taille n_1 (resp n_2).
On a :
X a pour moyenne μ_1 et comme écart-type $\sigma(X)/\sqrt{n_1}$
$\bar{Y}$ a pour moyenne μ_2 et comme écart-type $\sigma(Y)/\sqrt{n_2}$

Cas où $\sigma(X)$ et $\sigma(Y)$ sont connus

On a si $\mu_1=\mu_2$:
$X-\bar{Y}/\sqrt{\sigma(X)^2/n_1+\sigma(Y)^2/n_2}$ suit approximativement une loi $N(0,1)$.

Cas où $\sigma(X)$ et $\sigma(Y)$ ne sont pas connus

On les estime :

- si n_1 et n_2 sont grands,

$\sigma(X) \simeq s_1\sqrt{n_1/n_1-1}$ donc $\sigma(X)^2/n_1 \simeq s_1^2/n_1-1$
$\sigma(Y) \simeq s_2\sqrt{n_2/n_2-1}$ donc $\sigma(Y)^2/n_2 \simeq s_2^2/n_2-1$
On pose :
$s_{12}=\sqrt{\sigma(X)^2/n_1+\sigma(Y)^2/n_2} \simeq \sqrt{s_1^2/n_1-1+s_2^2/n_2-1}$
Donc sous l'hypothèse H_0 : $\mu_1=\mu_2=\mu$, on a $(X-\bar{Y}) \in N(0,s_{12})$

Recette si n_1 et n_2 sont grands

Avec Xcas on tape si $\alpha=0.05$:
a:=normal_icdf(0,s12,0.975)
On regarde si :
$|m_1-m_2|<a$
Si c'est le cas, on admet que $\mu_1=\mu_2$ et que les deux échantillons ne sont pas

significativement différents au seuil α, sinon on dira que $\mu_1 \neq \mu_2$ et que les deux échantillons ne proviennent pas de la même population.

- si n_1 et n_2 sont petits,

on peut estimer $\sigma(X)$ et $\sigma(Y)$ grâce à la reunion des deux échantillons et en faisant l'hypothèse $\sigma(X)=\sigma(Y)$ (pour vérifier cette hypothèse on pourra faire une étude de l'hypothèse $\sigma(X)=\sigma(Y)$ grâce au test expliqué au paragraphe suivant).
On montre qu'une bonne approximation est :
$\sigma^2=\sigma(X)^2=\sigma(Y)^2 \simeq s^2=n_1s_1^2+n_2s_2^2/n_1+n_2-2$.
En effet, la statistique $n_1S_1^2+n_2S_2^2/n_1+n_2-2$ est un estimateur sans biais de σ^2 si σ est l'écart-type de X. La valeur de cette statistique est obtenue à partir de deux échantillons de taille respective n_1 et n_2 et d'écart-type respectif s_1 et s_2 qui sont les valeurs de S_1 et S_2 pour ces deux échantillons (avec comme notation $S^2=1/n\sum_j (X_j-\mathrm{X})^2$ pour un échantillon de taille n de la variable X d'écart-type σ, on sait que $n/n-1S^2$ est un estimateur sans biais de σ^2) :
On a :
$\sigma^2=n_1/n_1-1E(S_1^2)=n_2/n_2-1E(S_2^2)$ donc
$E(n_1S_1^2+n_2S_2^2/n_1+n_2-2)= n_1E(S_1^2)+n_2E(S_2^2)/n_1+n_2-2= (n_1-1)\sigma^2+(n_2-1)\sigma^2/n_1+n_2-2=\sigma^2$
donc $\sigma^2 \simeq s^2=n_1s_1^2+n_2s_2^2/n_1+n_2-2$.
Alors sous l'hypothèse H_0 : $\mu_1=\mu_2=\mu$, et $\sigma(X)=\sigma(Y)=\sigma$, la statistique :

$$T=\frac{\mathrm{X}-\bar{\mathrm{Y}}}{\sqrt{\sigma(X)^2/n_1+\sigma(Y)^2/n_2}} \simeq \frac{(\mathrm{X}-\bar{\mathrm{Y}})\sqrt{n_1+n_2-2}}{\sqrt{(n_1s_1^2+n_2s_2^2)(1/n_1+1/n_2)}}$$

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

Recette si n_1 et n_2 sont petits

Avec Xcas on tape si $\alpha=0.05$:
a:=student_icdf(n1+n2-2,0.975)
On regarde si :
$|m_1-m_2|<a$
Si c'est le cas, on admet que $\mu_1=\mu_2$ et que les deux échantillons ne sont pas significativement différents au seuil α, sinon on dira que $\mu_1 \neq \mu_2$ et que les deux échantillons ne proviennent pas de la même population.

3.4.3 Comparaison de deux écarts-types observés

Soient s_1 et s_2 les écarts-types observés d'un caractère dont l'écart-type théorique est σ. Cette observation est faite à partir de deux échantillons de taille respéctive n_1 et n_2.
On veut savoir si les écarts-types s_1 et s_2 sont significativement différents ce qui voudrait dire que les deux échantillons proviennent de deux populations différentes d'écart-type respectif σ_1 et σ_2 ou si au contraire les deux échantillons proviennent d'une même population ou de deux populations de même écart-type $\sigma=\sigma_1=\sigma_2$.
Soient deux caractères normaux indépendants X et Y distribués respectivement selon les lois $N(\mu_1,\sigma_1)$ et $N(\mu_2,\sigma_2)$.
Soient deux échantillons (un échantillon pour le caractère X et l'autre pour le caractère Y) de taille respective n_1 et n_2, de moyenne respective m_1 et m_2 et d'écart-type respectif s_1 et s_2 .

Posons :
$S_1^2=1/n_1\sum_{j=1}^{n_1}(X_j-X)^2$ et
$S_2^2=1/n_2\sum_{j=1}^{n_2}(Y_j-\bar{Y})^2$
Lorsque $\sigma_1=\sigma_2=\sigma$, la statistique :
$F_{1,2}=n_1(n_2-1)S_1^2/n_2(n_1-1)S_2^2$ suit une loi de Fisher-Snedecor $F(n_1-1,n_2-1)$ à (n_1-1) et à (n_2-1) degrés de liberté.
De même la statistique :
$F_{2,1}=n_2(n_1-1)S_2^2/n_1(n_2-1)S_1^2$ suit une loi de Fisher-Snedecor $F(n_2-1,n_1-1)$ à (n_2-1) et à (n_1-1) degrés de liberté.
Cette statistique $F_{1,2}$ ou $F_{2,1}$ va nous permettre de tester les hypothèses :
H_0 : $\sigma_1=\sigma_2$ et H_1 : $\sigma_1 \neq \sigma_2$.
On rejettera l'hypothèse bilatérale H_0 si la valeur de $F_{1,2}$ est trop éloignée de 1.
Attention à l'ordre n_1,n_2, car les tables ne donnent que les valeurs de F supérieures à 1, on sera quelquefois amené à changer l'ordre des variables (on a $F_{1,2}= 1/F_{2,1}$).
Pour avoir $Proba(a<F_{1,2}<b)=1-\alpha$, on cherche a et b vérifiant :
$Proba(F(n_1-1,n_2-1)<b)=1-\alpha/2$ et
$Proba(F(n_1-1,n_2-1)<a)=\alpha/2$
dans une table de Fisher-Snedecor $F(n_1-1,n_2-1)$ à (n_1-1) et (n_2-1) degrés de liberté.
On a alors, si on échange l'ordre de n_1,n_2 :
$Proba(F(n_2-1,n_1-1)<1/a)=1-\alpha/2$
$Proba(F(n_2-1,n_1-1)<1/b)=\alpha/2$

Recette
- Choisir le seuil α
- Prélever les échantillons de taille n_1 et n_2,
- Calculer leurs écarts-types s_1 et s_2,
- Si $n_1(n_2-1)s_1^2>n_2(n_1-1)s_2^2$, calculer :
$f=n_1(n_2-1)s_1^2/n_2(n_1-1)s_2^2$ (cas 1)
ou sinon, calculer :
$f=n_2(n_1-1)s_2^2/n_1(n_2-1)s_1^2$ (cas 2).
- Déterminer grâce a la table de Fisher h vérifiant :
$Proba(1<F(n_1-1,n_2-1)<h)=1-\alpha/2$ (cas 1)
ou vérifiant :
$Proba(1<F(n_2-1,n_1-1)<h)=1-\alpha/2$ (cas 2).
Avec Xcas on tape si $\alpha=0.05$ et si $n_1(n_2-1)s_1>n_2(n_1-1)s_2$,
h:=fisher_icdf(n1-1,n2-1,0.975)
ou si $n_1(n_2-1)s_1<n_2(n_1-1)s_2$,
h:=fisher_icdf(n2-1,n1-1,0.975)
- si $f>h$ (c'est à dire si f s'éloigne trop de 1) on rejette l'hypothèse bilatérale H_0 : $\sigma_1)=\sigma_2$ sinon on l'accepte.

Remarque

Avec Xcas on tape si $\alpha=0.05$:
h:=fisher_icdf(n1-1,n2-1,0.975)
k:=fisher_icdf(n2-1,n1-1,0.975)
Alors k=1/k et h et k définissent les bornes en dehors d esquelles il faut rejetter l'hypothèse au seuil 0.05.

3.5. Le test du χ^2

Dans ce chapitre on cherche à savoir si deux variables sont indépendantes (test d'indépendance) et à comparer la distribution du caractère étudié à une distribution théorique (test d'adéquation).
Par exemple, certains tests ne sont valables que lorsque le phénomène étudié suit une loi normale, ou bien lorsqu'on suppose l'indépendance de deux variables : il est donc important de savoir si cela est bien le cas.

3.5.1 Adéquation d'une distribution expérimentale à une distribution théorique

Considérons un échantillon de taille n ayant une distribution $x_1,..,x_k$ d'effectifs $n_1,..,n_k$ (avec $n_1+...+n_k=n$) correspondant à l'observation d'une variable aléatoire X : X est discrète ou X est continue et dans ce cas on effectue un regroupement en k classes des valeurs de X, et $x_1,..,x_k$ représentent alors le centre de ces classes.
On veut comparer cette distribution empirique à une distribution théorique d'effectifs $e_1,..,e_k$ (si chaque valeur x_j est obtenue avec la probabilité théorique p_j on a $e_j=np_j$).
La statistique $D^2=\sum_{j=1}^{k}(n_j-e_j)^2/e_j$ est une bonne mesure de l'écart entre les effectifs observés et les effectifs théoriques : plus D^2 est proche de zéro, plus la distribution de l'échantillon est conforme à la distribution théorique.
L'objectif sera donc d'estimer si D^2 est suffisamment faible pour que l'on puisse ajuster la loi théorique à la distribution observée.
On montre que si n est grand, si $e_j>5$ pour tout j, et si les e_j ont été obtenus sans avoir eu recours à l'échantillon, la statistique D^2 suit approximativement une loi du χ^2 à $\nu=(k-1)$ degrés de liberté où k est le nombre de classes.
Lorsque l'on a eu recours à l'échantillon pour déterminer r paramètres, le nombre de degrés de liberté est alors de $\nu=(k-r-1)$.
On note dans la suite ν le nombre de degrés de liberté.
La statistique D^2 est alors utilisée comme variable de décision dans le test d'hypothèses :
H_0 : pour tout $j=1...k$, $Proba(X=x_j)=p_j$
H_1 : il existe $j=1...k$, $Proba(X=x_j)\neq p_j$
On rejettera l'hypothèse d'adéquation au modèle dès que l'écart D^2 est supérieur à ce que l'on peut attendre de simples fluctuations dues à l'échantillonnage. La région critique au seuil α (c'est la région où il faudra rejeter l'hypothèse) est la région pour laquelle : $d^2>h\}$ quand $Proba(\chi_\nu^2<h)=1-\alpha$ et lorsque d^2 est la valeur de D^2 pour l'échantillon.
On remarquera que D^2 fait intervenir le nombre de classes et les effectifs de chaque classe et que D^2 ne fera intervenir les x_j que pour estimer les paramètres p_j de la loi.
Pour les effectifs e_j trop petits on effectuera un regroupement de classes.

Recette

Dans une table du χ^2 on cherche h tel que :
$Proba(\chi_{k-1}^2<h)=1-\alpha$
Avec Xcas on tape pour trouver h, si on a k classes et si $\alpha=0.05$:
chisquare_icdf(k-1,0.975)
On prélève un échantillon de taille n et on note sa distribution $n_1,..,n_k$ correspondant aux k classes de centre $x_1,..,x_k$.
On calcule la valeur d^2 de D^2 : $d^2=\sum_{j=1}^{k}(n_j-np_j)^2/np_j=\sum_{j=1}^{k}(n_j-e_j)^2/e_j$

Règle

On rejette l'hypothèse H_0 au seuil α, quand d^2 est supérieure à h.

Exemple

Dans un croisement de fleurs rouges et blanches, on a obtenu le résultat suivant sur un échantillon de 600 plants de la 2-ième génération :
141 fleurs rouges, 315 fleurs roses, 144 fleurs blanches.
Ces résultats sont-ils conformes à la distribution théorique :
25% fleurs rouges, 50% fleurs roses, 25% fleurs blanches.
On a 3 classes donc 3-1=2 degrés de liberté :
n_1=141 et e_1=600*25/100=150
n_2=315 et e_2=600*50/100=300
n_3=144 et e_3=600*25/100=150
On calcule $d^2=\sum_{j=1}^{3}(n_j-e_j)^2/e_j$=81/150+15^2/300+36/150 On tape dans Xcas :
81/150+15^2/300+36/150
On obtient :
=153/100
On tape :
chisquare_icdf(2,0.95)
On obtient :
5.99146454711
Comme 1.53<5.992 on ne peut pas rejeter l'hypothèse H_0 au seuil de 5%, donc on l'accepte.

3.5.2 Adéquation d'une distribution expérimentale à une distribution de Poisson

Pour pouvoir calculer les effectifs théoriques, on est souvent obligé d'estimer le paramètre μ à partir de l'échantillon (μ est estimé par la moyenne m de l'échantillon).

Règle

Soit k est le nombre de classes.
Si on s'est servi de l'échantillon pour estimer μ, alors la statistique D^2 suit une loi du χ^2 à $k-2$ degrés de liberté (cas 1),
sinon D^2 suit une loi du χ^2 à $k-1$ degrés de liberté (cas 2) .
Pour savoir si la distribution $n_1,..,n_k$ correspondant aux k classes de centre $x_1,..,x_k$ est conforme à une distribution de Poisson, on utilise le test d'hypothèses :
H_0 : pour tout j $Proba(X=x_j)=e^{-x_j}\lambda^{x_j}/x_j!=p_j$ et
H_1 : il existe j=1...k, $Proba(X=x_j)\neq p_j$
On rejette l'hypothèse H_0 au seuil α, quand la valeur d^2 de D^2 est supérieure à h avec h vérifiant :
- cas 1 : $Proba(\chi_{k-2}^2\leq h)=1-\alpha$,
- cas 2 : $Proba(\chi_{k-1}^2\leq h)=1-\alpha$.

Exemple

On a effectué un échantillon de taille 100 et on a obtenu, pour les 11 valeurs entières d'une variable aléatoire *X* les effectifs suivants :

X	n_j
0	1
1	8
2	19
3	23
4	17
5	15
6	8
7	3
8	3
9	2
10	1

Peut-on dire que *X* suit une loi de Poisson ?
On suppose que cela est vrai et on estime le paramètre de la loi de Poisson par la moyenne de l'échantillon.
Ssoit on utilise le tableur, soit on tape :
L1:=[0,1,2,3,4,5,6,7,8,9,10]
L2:=[1,8,19,23,17,15,8,3,3,2,1]
mean(L1,L2)
On obtient :
379/100
On cherche les effectifs théoriques on tape (n=100) :
100*poisson(3.79,0), 100*poisson(3.79,1), etc ...
100*poisson(3.79,9), 100*poisson(3.79,10).
On rappelle que : poisson(3.79,k)=exp(-3.79)*(3.79^k)/k!
ou bien on tape
L:=[];for(j:=0;j<11;j++) {
L:=concat(L,poisson(3.79,j)*100);}
ou encore on tape :
L:=seq(100*poisson(3.79,k),k,0,10)
On obtient la liste L des 11 valeurs de e_j pour j=0..10 :
[2.25956018511,8.56373310158,16.2282742275,20.5017197741,
19.4253794859,14.7244376503,9.30093644912,5.0357927346,
2.38570680802,1.00464764471,0.380761457345]
il faut changer la valeur de la dernière classe car elle doit comporter toutes les valeurs supérieures ou égales à 10 (la somme des e_j est égale à la taille de l'échantillon ici 100):

L[10]:=100-sum(L[j],j,0,9)
On obtient :_ 0.56981193906
Donc on obtient la liste L des e_j :
[2.25956018511,8.56373310158,16.2282742275,20.5017197741,
19.4253794859,14.7244376503,9.30093644912,5.0357927346,
2.38570680802,1.00464764471,0.56981193906]
On regroupe les petits effectifs pour avoir e_j>5, on a donc 7 classes :
On tape :
L[0]+L[1]
On obtient :
10.8232932867
L[7]+L[8]+L[9]+L[10]
On obtient :
8.99595912639
Ou encore, on tape :
L:=accumulate_head_tail(L,2,4)
Donc on obtient la liste L d'effectifs théoriques e_j avec e_j>5:
L:=[10.8232932867,16.2282742275,20.5017197741,
19.4253794859,14.7244376503,9.30093644912,8.99595912639]
La liste L2 d'effectifs empiriques correspondant à ces 7 classes est :
L2:=[9,19,23,17,15,8,9]
On calcule :
L3:=(L2-L)^2
d2:=evalf(sum((L3[j]/L[j]),j,0,6))
On obtient :
1.57493190982
On sait que D^2 suit une loi du χ^2 ayant (7-2)=5 degrés de liberté car on a estimé λ par *m* moyenne de l'échantillon.
On tape pour connaitre la région critique au seuil de α=0.05 :
chisquare_icdf(5,0.95)
On obtient :
11.0704976935
donc $h \simeq 11.07$:
Proba(D^2<11.07)=0.95 ou encore *Proba*(D^2>11.07)=0.05.
Cela veut dire que D^2 a des valeurs supérieure à 11.07 que dans 5% des cas c'est à dire très peu souvent ou encore que la probabilité que D^2 soit supérieur à 11.07 par le seul fait du hasard sur l'échantilonnage est 0.05, et dans ce cas il n'y aurait que 5 chances sur 100 pour que l'on ait alors une distribution de Poisson.
Donc si la valeur observée d^2 de D^2 est supérieure à 11.07 on rejetera l'hypothèse H_0 au seuil α=0.05.
Dans l'exemple ci-dessus, la valeur observée de D^2 est d^2=1.575, donc on estime que l'hypothèse selon laquelle la distribution est une distribution de Poisson n'est pas à rejeter au seuil de 5%.

3.5.3 Adéquation d'une distribution expérimentale à une distribution normale

Pour pouvoir calculer les effectifs théoriques, on est souvent obligé d'estimer les paramètres μ et σ à partir de l'échantillon (μ par la moyenne *m* de l'échantillon et σ par $s\sqrt{n}/n-1$ où *s* est l'écart-type de l'échantillon).

Règle

Soit k le nombre de classes.
Si on s'est servi de l'échantillon pour estimer μ et σ la statistique D^2 suit une loi du χ^2 à $k-3$ degrés de liberté (cas 1),
si on s'est servi de l'échantillon pour estimer μ ou σ la statistique D^2 suit une loi du χ^2 à $k-2$ degrés de liberté (cas 2)
sinon D^2 suit une loi du χ^2 à $k-1$ degrés de liberté (cas 3) (k est le nombres de classes).
On rejette l'hypothèse H_0 au seuil α, quand la valeur d^2 de D^2 est supérieure à h avec h vérifiant :
- cas 1 : $Proba(\chi_{k-3}^2 \leq h)=1-\alpha$,
- cas 2 : $Proba(\chi_{k-2}^2 \leq h)=1-\alpha$,
- cas 3 : $Proba(\chi_{k-1}^2 \leq h)=1-\alpha$.

Exemple

On a effectué un échantillon de taille 250 et on a obtenu, pour les valeurs d'une variable aléatoire X, réparties en 10 classes, les effectifs suivants :

X	n_j
45..46	11
46..47	15
47..48	27
48..49	35
49..50	47
50..51	58
51..52	28
52..53	16
53..54	10
54..55	3

On va tout d'abord calculer la moyenne m et l'écart-type s de l'échantillon :
On tape :
L1:=[45..46,46..47,47..48,48..49,49..50,50..51,51..52
,52..53,53..54,54..55]
L2:=[11,15,27,35,47,58,28,16,10,3]
On tape :
m:=mean(L1,L2)
On obtient :
6207/125
Donc $m \simeq 49.656$
s:=stddev(L1,L2)

On obtient :
sqrt(249229/62500)
On obtient une estimation de σ en tapant :
s*sqrt(250/249)
On obtient :
2.00091946736
Donc $s \simeq 2$
On cherche les effectifs théoriques on tape :
normal_cdf(49.656,2,45,46)
On obtient :
0.0238187239894,
normal_cdf(49.656,2,46,47),
etc ...
normal_cdf(49.656,2,54,55).
On rappelle que :
normal_cdf(μ,σ,x_1,x_2)=$\int_{x1}^{x_2} 1/\sigma\sqrt{2\pi}\exp(-(x-\mu)^2/2*\sigma^2)dx$
ou bien on tape
L:=[];
for(j:=0;j<10;j++) {
L:=concat(L,normal_cdf(49.656,2,45+j,46+j));}
ou encore on tape :
L:=seq(normal_cdf(49.656,2,45+j,46+j),j,0,9)
On obtient la liste L des p_j (p_j est la probabilité théorique pour que la valeur de X soit dans la j-ième classe) :
[0.0238187239894,0.0583142776342,0.11174619649,
0.167620581364,0.196825404189,0.180926916339,
0.130193320084,0.0733363670394, 0.0323343295781,
0.0111577990363]
Il faut modifier le premier terme et le dernier terme de L car la première classe est en fait]−∞;46[et la dernière [54;+∞[.
On tape :
normal_cdf(49.656,2,-infinity,46)
On obtient :
0.0337747758231
On tape :
normal_cdf(49.656,2,54,+infinity)
On obtient :
0.0149278314584
L:=[0.0337747758231,0.0583142776342,0.11174619649,
0.167620581364,0.196825404189,0.180926916339,
0.130193320084,0.0733363670394,0.0323343295781,
0.0149278314584]
On obtient la liste L des effectifs théoriques e_j de chaque classe en tapant :
L:=250*L
On obtient :
[8.44369395578,14.5785694086,27.9365491225,41.
905145341,49.2063510472,45.2317290848,
32.548330021,18.3340917599,8.08358239453,
3.7319578646]
On regroupe les 2 dernières classes (L[8]+L[9]=11.8155402591),

Ou encore , on tape :
L:=accumulate_head_tail(L,1,2) on obtient la liste L des effectifs théoriques des 9 classes :
L:= [8.44369395578,14.5785694086,27.9365491225,
41.905145341,49.2063510472,45.2317290848,
32.548330021,18.3340917599,11.8155402591]
La liste L2:=[11,15,27,35,47,58,28,16,10,3] des effectifs de l'échantillon après un regroupement en 9 classes, on tape :
L2:=accumulate_head_tail(L2,1,2)
On obtient :
L2:=[11,15,27,35,47,58,28,16,13]
On calcule la valeur de D^2 :
d2:=sum(((L-L2)[j])^2/L[j],j,0,8)
On obtient :
6.71003239422
On calcule $Proba(\chi_6^2<h)=0.95$, pour cela on tape (car on a 9−3=6 degrés de liberté) :
chisquare_icdf(6,0.95)
On obtient :
12.5915872437
donc $h \simeq 12.6$
L'hypothèse n'est pas à rejeter au seuil de 5% puisque $d2=6.71<12.6$.

3.6. Comparaison de la distribution de plusieurs échantillons

3.6.1 Cas général : on a *m* échantillons

Soient *m* échantillons, comment savoir si la distribution des fréquences de ces *m* échantillons sont celles d'échantillons d'une même loi ?

Notations

On suppose que les *m* échantillons peuvent prendre *k* valeurs numerotées de 1 à *k*. On note à l'aide d'un indice (*i*) placé en haut ce qui concerne le *i*-ième échantillon ainsi, $n^{(i)}$ est la taille de l'échantillon *i* et $n_j^{(i)}$ est le nombre d'occurences de la valeur *j* dans la série *i*, donc *i* varie de 1 à *m* et *j* varie de 1 à *k*.
On a *m* échantillons et dans chaque échantillon il y a *k* classes.
On a donc :
$\sum_j n_j^{(i)} = n^{(i)}$ qui est la taille de l'échantillon (*i*).
On pose :
$\sum_{i,j} n_j^{(i)} = n$ et
$\sum_i n_j^{(i)} = n_j$.
Donc *n* est la taille de l'échantillon total constitué par les *m* échantillons, n_j est le nombre total d'occurences de la valeur *j* dans l'échantillon total.
D'après la loi des grands nombres, si on considère que les *m* échantillons suivent la même loi *X* que l'échantillon total on a $Proba(X=j) \simeq n_j/n$.
Donc on peut considérer que l'effectf théorique de la valeur *j* de l'échantillon (*i*) est :
$\nu_j^{(i)} = n^{(i)} * n_j/n$.
La variable de décision est alors :
$D^2 = \sum_{i,j} (n_j^{(i)} - \nu_j^{(i)})^2 / \nu_j^{(i)}$
Cette variable suit une loi de χ^2 ayant $s=(m-1)(k-1)$ degrés de liberté.

3.6.2 Application à deux échantillons prenant deux valeurs

Soient $f_1^{(1)}$ et $f_1^{(2)}$ les fréquences observées sur deux échantillons d'un caractère dont la fréquence théorique est p. Cette observation est faite à partir de deux échantillons de taille respéctive $n^{(1)}$ et $n^{(2)}$.
On veut savoir si les fréquences $f_1^{(1)}$ et $f_1^{(2)}$ sont significativement différentes ce qui voudrait dire que les deux échantillons proviennent de deux populations différentes de paramètre p_1 et p_2 ou si aucontraire les deux échantillons proviennent d'une même population de paramètre $p=p_1=p_2$ c'est à dire que ces deux échantillons sont ceux d'une même loi .
On a :
$n^{(1)}+n^{(2)}=n$
$f_1^{(1)}n^{(1)}=n_1^{(1)}$ et $(1-f_1^{(1)})n^{(1)}=n_2^{(1)}(=f_2^{(1)}n^{(1)})$
$f_1^{(2)}n^{(2)}=n_1^{(2)}$ et $(1-f_1^{(2)})n^{(2)}=n_2^{(2)}(=f_2^{(2)}n^{(2)})$
$n_1=f_1^{(1)}n^{(1)}+f_1^{(2)}n^{(2)}$
$n_2=(1-f_1^{(1)})n^{(1)}+(1-f_1^{(2)})n^{(2)}$
$\nu_j^{(i)}=n^{(i)}n_j/n$ donc
$\nu_1^{(1)}-n_1^{(1)}=n^{(1)}(f_1^{(1)}n^{(1)}+f_1^{(2)}n^{(2)})/(n^{(1)}+n^{(2)})-f_1^{(1)}n^{(1)}=$
$n^{(2)}n^{(1)}(f_1^{(2)}-f_1^{(1)})/(n^{(1)}+n^{(2)})$
$\nu_1^{(2)}-n_1^{(2)}=n^{(2)}(f_1^{(1)}n^{(1)}+f_1^{(2)}n^{(2)})/(n^{(1)}+n^{(2)})-f_1^{(2)}n^{(2)}=$
$n^{(1)}n^{(2)}(f_1^{(1)}-f_1^{(2)})/(n^{(1)}+n^{(2)})$
$\nu_2^{(1)}-n_2^{(1)}=n^{(1)}((1-f_1^{(1)})n^{(1)}+(1-f_1^{(2)})n^{(2)})/(n^{(1)}+n^{(2)})-(1-f_1^{(1)})n^{(1)}=$
$n^{(1)}n^{(2)}(f_1^{(1)}-f_1^{(2)})/(n^{(1)}+n^{(2)})$
$\nu_2^{(2)}-n_2^{(2)}=n^{(2)}((1-f_1^{(1)})n^{(1)}+(1-f_1^{(2)})n^{(2)})/(n^{(1)}+n^{(2)})-(1-f_1^{(2)})n^{(2)}=$
$n^{(1)}n^{(2)}(f_1^{(2)}-f_1^{(1)})/(n^{(1)}+n^{(2)})$
$1/\nu_1^{(1)}+1/\nu_1^{(2)}+1/\nu_2^{(1)}+1/\nu_2^{(2)}=$
$(n^{(1)}+n^{(2)})(1/n^{(1)}+1/n^{(2)})(1/f_1^{(1)}n^{(1)}+f_1^{(2)}n^{(2)}+1/(1-f_1^{(1)})n^{(1)}+(1-f_1^{(2)})n^{(2)})=$
$(n^{(1)}+n^{(2)})^2/n^{(1)}n^{(2)}(1/f_1^{(1)}n^{(1)}+f_1^{(2)}n^{(2)}+1/(1-f_1^{(1)})n^{(1)}+(1-f_1^{(2)})n^{(2)})=$
$(n^{(1)}+n^{(2)})^3/n^{(1)}n^{(2)}(n_1f_1^{(1)}+n_2f_1^{(2)})(n_1(1-f_1^{(1)})+n_2(1-f_1^{(2)}))$
La variable D^2 suit une loi du χ^2 a 1 degré de liberté : on a 2 échantillons (m=2) et chaque échantillon ne prend que 2 valeurs, (k=2) donc $s=(m-1)(k-1)=1$.
La variable D^2 s'écrit alors :
$D^2=n^{(1)}n^{(2)}(f_1^{(1)}-f_1^{(2)})^2(n^{(1)}+n^{(2)})/(n^{(1)}f_1^{(1)}+n^{(2)}f_1^{(2)})(n^{(1)}(1-f_1^{(1)})+n^{(2)}(1-f_1^{(2)}))$
ou encore
$D^2=n(n_1^{(1)}n_2^{(2)}-n_1^{(2)}n_2^{(1)})^2/n^{(1)}n^{(2)}n_1n_2$
D^2 suit une loi du χ^2 ayant 1 degré de liberté.

Exercice

Pour tester l'efficacité d'un vaccin antigrippal on soumet 300 personnes à une expérience :
- sur 100 personnes non vaccinées, 32 sont atteintes par la grippe,
- sur 200 personnes vaccinées, 50 sont atteintes par la grippe,

Ce résultat permet-il d'apprécier l'efficacité du vaccin ?
On a le tableau suivant :

	grippé	non grippé	taille
vacciné	32	68	100
non vacciné	50	150	200

total	82	218	300

On calcule la valeur $d2$ de D^2 on tape :
d2:=300*(150*32-68*50)^2/(100*200*82*218)
On obtient :
7350/4469
donc $d^2 \simeq 1.645$
On cherche la valeur h qui vérifie :
$Proba(\chi_1^2>h)=0.05$ ou encore $Proba(\chi_1^2 \leq h)=0.95$
pour cela on tape :
chisquare_icdf(1,0.95)
On obtient :
3.84145882069
donc $h \simeq 3.84$
Puisque $d2 \simeq 1.645<3.84$ on en déduit que les deux échantillons ne sont pas significativement différents au seuil de 5% : on peut donc mettre en doute l'efficacitédu vaccin.

3.7. Application : le test d'indépendance

C'est une application du test d'adéquation.
Considérons une variable aléatoire X valant $x_1,..,x_k$ avec une probabilité théorique $p_1,..,p_k$ (avec $p_1+...+p_k=1$) et une variable aléatoire Y valant $y_1,..,y_l$ avec une probabilité théorique $q_1,..,q_l$ (avec $q_1+...+q_l=1$).
On a un échantillon de taille n (n grand) pour lequel le nombre d'éléments présentant le caractère x_i et le caractère y_j est $n_{i,j}$ ($\sum n_{i,j}=n$).
On veut savoir, au vue de l'échantillon si les variables X et Y sont indépendantes.
On peut estimer les p_i et les q_j par :
$p_i \simeq \sum_{j=1}^{l} n_{i,j}/n$
$q_j \simeq \sum_{i=1}^{k} n_{i,j}/n$
En estimant ces valeurs, on a estimé $k-1+l-1=k+l-2$ paramètres (car quand on a estimé $p_1,..,p_{k-1}$ on a l'estimation de p_k et quand on a estimé $q_1,..,q_{l-1}$ on a l'estimation de q_l).
Si X et Y sont indépendantes (hypothèse H_0), alors :
$Proba((X=x_i) \cap (Y=y_j))=p_iq_j$
donc l'effectif théorique des éléments présentant le caractère x_i et y_j est :
$e_{i,j}=np_iq_j$.
La statistique $D^2=\sum_{i=1}^{k}\sum_{j=1}^{l}(n_{i,j}-np_iq_j)^2/np_iq_j$ suit approximativement une loi du χ^2 ayant $(k-1)(l-1)$ degrés de liberté (car $(k-1)(l-1)=kl-1-(k+l-2)$).

Règle

On calcule d^2 la valeur de D^2 pour l'échantillon et $\nu=(k-1)(l-1)$ le nombre de degrés de liberté.
On cherche dans une table la valeur de h vérifiant : $Proba(\chi_\nu^2<h)=1-\alpha$
Avec Xcas on tape si $\alpha=0.05$:

h:=chisquare_icdf((k-1)(l-1),0975)
Si $d^2<h$, on accepte l'hypothèse d'indépendance au seuil α, sinon on la rejette.

3.8. Le test de corrélation

On considère une série statistique double, c'est à dire que pour chaque individu d'une même population, on étudie deux caractères X et Y. On veut savoir si ces deux caractères ont une relation entre eux.
L'ensemble des valeurs (x_j,y_j) de (X,Y) s'appelle un nuage de points.

Rappel

Soient deux variables aléatoires X et Y, on définit le coefficient de corrélation ρ de ces deux variables par le nombre :
$\rho=E((X-E(X))(Y-E(Y)))/\sigma(X)\sigma(Y)=E(XY)-E(X)E(Y)/\sigma(X)\sigma(Y)=cov(X,Y)/\sigma(X)\sigma(Y)$
Si X et Y sont indépendantes alors ρ=0.
Dans le cas où le nuage de points de coordonées $(x_j;y_j)$ est linéaire, l'équation de la droite, dite de régression, est :
$y=ax+b$ avec
$a=E((X-E(X))(Y-E(Y)))/\sigma(X)^2$ et
$b=E(Y)-aE(X)$
On a donc :
$\rho=a\sigma(X)/\sigma(Y)$

Théorème :

Au vue d'un échantillon de taille n, on peut estimer ρ par l'estimateur :
$R=\sum_{j=1}^{n}(X_j-\bar{X})(Y_j-\bar{Y})/\sqrt{(\sum_{j=1}^{n}(X_j-\bar{X})^2)(\sum_{j=1}^{n}(Y_j-\bar{Y})^2)}$
Lorsque X et Y suivent une loi normale les variables :
$V=1/2\ \ln((1+R)(1-\rho)/(1-R)(1+\rho))$ suit une loi normale $N(\rho/2n-2,1/\sqrt{n-3})$ et,
$T=\sqrt{n-2}R/\sqrt{1-R^2}$ suit une loi de Student à $n-2$ degrés de liberté.
Si $R^2=1$, les points de coordonées $(x_j;y_j)$ sont alignés sur la droite des moindres carrés et,
si $R^2=0$, cela permet de conclure à l'inadéquation du modèle linéaire.

Attention

Si $R=0$, les variables X et Y ne sont pas obligatoirement indépendantes. De même, lorsque R^2 est proche de 1, on peut penser (c'est un indice et non une preuve) qu'il y a un lien de cause à effet entre X et Y.
On peut donc tester au seuil α l'hypothèse H_0 : ρ=0.
Par exemple, pour α=0.05, on considère que ρ=0 est vraisemblable si :
$1/2\ \ln((1+R)/(1-R))<1.96*1/\sqrt{n-3}$
Pour estimer a et b on utilise les statistiques :
$A=\sum((X_j-\bar{X})(Y_j-\bar{Y}))/\sum(X_j-\bar{X})^2$ et $B=\bar{Y}-A\bar{X}$
On montre que A et B sont des estimateurs sans biais de a et b.

3.9. Le test de Dixon

Il peut arriver, au cours d'une expérimentation qu'un des résultats semble s'écarter notablement des autres. Une attitude classique, que l'on rencontre trop souvent, consiste à éliminer cette valeur en la considérant comme aberrante. Or il peut être dangereux de procéder ainsi sans vérification préalable. La bonne attitude à avoir est la suivante : si l'on a pu retrouver la cause de la valeur aberrante (erreur de lecture, faute de calcul, etc), il est tout à fait normal de l'éliminer, en revanche, si aucune cause accidentelle n'a pu être détectée, il est dangereux d'éliminer brutalement la valeur incriminée. Dans ce cas, il faut avoir recours à un test statistique permettant de justifier l'élimination de la valeur aberrante avec une probabilité P, choisie à l'avance, de se tromper. Le test de Dixon, que nous allons exposer, permet de réaliser cela

3.9.1. Le test

Supposons qu'une expérimentation ait conduit à n résultats numériques (n spérieur ou égal à 3). On classe ces résultats par **ordre de valeur croissante**

$$x_1 < x_2 < \ldots < x_{n-1} < x_n$$

Le test permet alors de tester si la première valeur x_1 ou la dernière valeur x_n est aberrante. Pour ce faire, suivant le nombre d'observations, on calcule les rapports suivants

$3 \leq n \leq 7$	$r_{10} = \frac{y_2 - y_1}{y_n - y_1}$	$r_{10} = \frac{y_n - y_{n-1}}{y_n - y_1}$
$8 \leq n \leq 10$	$r_{11} = \frac{y_2 - y_1}{y_{n-1} - y_1}$	$r_{11} = \frac{y_n - y_{n-1}}{y_n - y_2}$
$11 \leq n \leq 13$	$r_{21} = \frac{y_3 - y_1}{y_{n-1} - y_1}$	$r_{21} = \frac{y_n - y_{n-2}}{y_n - y_2}$
$14 \leq n \leq 30$	$r_{22} = \frac{y_3 - y_1}{y_{n-2} - y_1}$	$r_{22} = \frac{y_n - y_{n-2}}{y_n - y_3}$

On entre alors dans la table de Dixon qui donne les valeurs critiques de ces rapports au niveau de risque 10 %, 5 % et 1 %. La règle à adopter est la suivante : si la valeur du rapport est inférieure à la valeur critique, on est pas justifié, au risque donné, de d'éliminer l'observation.

Exemple :

Valeurs observées : 148, 151, 152, 153, 160

Nous avons 5 valeurs et la dernière semble anormalement élevée.

on calcule

$$r_{10} = \frac{160 - 153}{160 - 148} = \frac{7}{12} = 0,583$$

Pour n = 5 observations, la valeur critique lue dans la table est, au risque de 5 %, 0,642.

Puisque 0, 583 < 0, 642 il n'est pas justifié, au risque de 5 %, d'éliminer la valeur 160.

3.9.2. Table de Dixon

Une expérimentation a conduit à n résultats numériques. On classe ces résultats par **ordre de valeur croissante**.

$$y_1 < y_2 < y_3 \ldots < y_{n-1} < y_n$$

I. Pour

$r_{10} = \frac{y_2 - y_1}{y_n - y_1}$ ou $r_{10} = \frac{y_n - y_{n-1}}{y_n - y_1}$

n			10%	5%	1%
3			0,886	0,941	0,988
4			0,679	0,765	0,889
5			0,557	0,642	0,780
6			0,482	0,560	0,698
7			0,434	0,507	0,637

II. Pour

$r_{11} = \frac{y_2 - y_1}{y_{n-1} - y_1}$ ou $r_{11} = \frac{y_n - y_{n-1}}{y_n - y_2}$

n			10%	5%	1%
8			0,479	0,554	0,683
9			0,441	0,512	0,635
10			0,409	0,477	0,597

III. Pour

$r_{21} = \frac{y_3 - y_1}{y_{n-1} - y_1}$ ou $r_{21} = \frac{y_n - y_{n-2}}{y_n - y_2}$

n			10%	5%	1%
11			0,517	0,576	0,679
12			0,490	0,546	0,642
13			0,467	0,521	0,615

IV. Pour

$r_{22} = \frac{y_3 - y_1}{y_{n-2} - y_1}$ ou $r_{22} = \frac{y_n - y_{n-2}}{y_n - y_3}$

n			10%	5%	1%
14			0,492	0,546	0,641
15			0,472	0,525	0,616
16			0,454	0,507	0,595
17			0,438	0,490	0,577
18			0,424	0,475	0,561
19			0,412	0,462	0,547
20			0,401	0,450	0,535

3.10. Les méthodes de Monte-Carlo

3.10.1. Introduction

Les méthodes de Monte-Carlo utilisent des nombres pseudo aléatoires (générés par un algorithme) pour simuler des phénomènes comportant une ou plusieurs variables aléatoires. Le nom provient du célèbre casino de Monte-Carlo.

On considère une simulation de Monte-Carlo élémentaire, visant à évaluer l'espérance et la variance d'une variable aléatoire en générant un grand nombre d'échantillons qui suivent la même loi de probabilité que la variable aléatoire.

3.10.2. Variable aléatoire discrète

3.10.2.1. Espérance et variance

Soit une variable aléatoire X pouvant prendre les M valeurs x_k avec k=0,..M-1. La probabilité d'obtenir la valeur x_k est notée p_k. La donnée de ces probabilités constitue la loi de la variable aléatoire, appelée aussi distribution des probabilités. La somme des probabilités doit être égale à 1 :

$$\sum_{k=0}^{M-1} p_k = 1 \qquad (1)$$

L'espérance de la variable aléatoire est :

$$E(X) = \sum_{k=0}^{M-1} p_k x_k \qquad (2)$$

Remarque : en physique statistique, l'espérance d'une grandeur physique aléatoire est appelée moyenne ou moyenne statistique de cette grandeur.

La variance est l'espérance du carré de l'écart entre la variable et son espérance :

$$\begin{aligned} var(X) &= E\left[(X - E(X))^2\right] && (3)\\ &= E\left[X^2 - 2XE(X) + E(X)^2\right] && (4)\\ &= E(X^2) - E(X)^2 && (5)\\ &= \sum_{k=0}^{M-1} p_k x_k^2 - \left(\sum_{k=0}^{M-1} p_k x_k\right)^2 && (6) \end{aligned}$$

L'écart type est la racine carrée de la variance :

$$\Delta X = \sqrt{var(X)} \qquad (7)$$

3.10.2.2. Simulation de Monte-Carlo

Principe

On utilise un générateur de nombres pseudo aléatoires pour générer des échantillons x_i de la variable aléatoire X. Les valeurs de ces échantillons sont dans l'ensemble des valeurs x_k.

Soit N le nombre d'échantillons générés. L'espérance est évaluée en calculant la moyenne empirique définie par :

$$y_N = moy(X, N) = \frac{1}{N} \sum_{i=0}^{N-1} x_i \qquad (8)$$

L'espérance de la moyenne empirique est égale à l'espérance de X.

Cette moyenne est généralement calculée pour des valeurs de N croissantes, ce qui nécessite le stockage de la somme des x_i.

La variance empirique est définie par :

$$v_N = var(X, N) = \frac{1}{N-1}\sum_{i=0}^{N-1}\left(x_i - y_N\right)^2 \qquad (9)$$

L'espérance de la variance empirique est égale à la variance de X.

En pratique, on préfère utiliser la forme développée suivante :

$$v_N = \frac{1}{N-1}\sum_{i=0}^{N-1} x_i^2 - \frac{N}{N-1} y_N^2 \qquad (10)$$

Dans une simulation de Monte-Carlo, le nombre N de tirage est très grand (au moins 1000). On peut donc confondre N-1 et N, ce qui donne l'expression approchée suivante de la variance empirique :

$$v_N \approx \frac{1}{N}\sum_{i=0}^{N-1} x_i^2 - y_N^2 \qquad (11)$$

L'estimation de la variance nécessite donc de stocker la somme des carrés des échantillons.

La moyenne y_M est elle-même une variable aléatoire : pour N fixé, la répétition de N tirages de la variable X donne des valeurs aléatoires de la moyenne. La variance de y_N est d'autant plus faible que N est grand. Pour calculer cette variance, on utilise les deux propriétés suivantes valables pour deux variables aléatoires indépendantes :

$$var(x_1 + x_2) = var(x_1) + var(x_2) \qquad (12)$$
$$var(ax) = a^2 var(x) \qquad (13)$$

On obtient ainsi la variance de la moyenne :

$$var\left(y_N\right) = \frac{var(X)}{N} \qquad (14)$$

Il s'en suit que l'écart-type de la moyenne varie comme l'inverse de la racine carrée :

$$\Delta\left(y_N\right) = \frac{\sqrt{var(X)}}{\sqrt{N}} \qquad (15)$$

Cet écart-type est évalué avec la variance empirique v_N.

Le théorème de la limite centrale établit que la moyenne empirique suit (si N est assez grand) une distribution continue gaussienne (voir plus loin). Un intervalle de confiance à 95 pour cent est donné par :

$$\left[y_N - \frac{1,96\sqrt{v_N}}{\sqrt{N}}, y_N + \frac{1,96\sqrt{v_N}}{\sqrt{N}} \right] \quad (16)$$

La probabilité pour que l'espérance cherchée se trouve dans cet intervalle est de 0,95.

Exemple

On considère des tirages d'un dé. Les valeurs possibles sont 1,2,3,4,5,6 avec chacun la probabilité 1/6. La fonction random.randint(1,6) permet d'obtenir les échantillons. Dans ce cas élémentaire, on connaît la valeur exacte de l'espérance :

$$E(X) = \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2} = 3,5 \quad (17)$$

et de la variance :

$$var(X) = \frac{1}{6}\left((1-7/2)^2 + (2-7/2)^2 + \cdots\right) \simeq 2,9167 \quad (18)$$

La fonction suivante effectue N tirages, calcule la moyenne et la variance empiriques, et le demi-intervalle de confiance à 95 pour cent :

```
import random
import math

def tirages(N):
    somme = 0.0
    somme2 = 0.0
    for i in range(N):
        x = random.randint(1,6)
        somme += x
        somme2 += x*x
    moyenne = somme*1.0/N
    variance = somme2*1.0/N-moyenne*moyenne
    ecart = 1.96*math.sqrt(variance*1.0/N)
    return (moyenne,variance,ecart)
```

Voici un exemple avec 1000 tirages :

```
t = tirages(1000)

print(t)
--> (3.526, 2.979324000000002, 0.10698304107848126)
```

Voici une deuxième série de tirages :

```
t = tirages(1000)

print(t)
```

```
--> (3.418, 2.945276, 0.10636997829086928)
```

Pour réduire l'écart-type de la moyenne, on doit augmenter le nombre de tirages :

```
t = tirages(10000)

print(t)
--> (3.5017, 2.8911971100000002, 0.03332690027256661)
```

Cette dernière simulation donne donc, avec un intervalle de confiance à 95 pour cent :

$$E(X) = 3.50 \pm 0.03 \quad (19)$$

3.10.3. Variable aléatoire continue

3.10.3.1. Densité de probabilité

Soit X une variable aléatoire continue (appelée aussi variable à densité), qui peut prendre toute valeur réelle dans l'intervalle [a,b]. On définit une fonction p(x) appelée densité de probabilité de la variable aléatoire, telle que la probabilité d'obtenir une valeur dans l'intervalle $[x_1,x_2]$ soit :

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} p(x)dx \quad (20)$$

La densité de probabilité doit vérifier la condition suivante, appelée condition de normalisation :

$$\int_a^b p(x)dx = 1 \quad (21)$$

On introduit aussi la fonction de répartition F(x) qui donne la probabilité d'obtenir une valeur inférieure ou égale à x :

$$F(x) = \int_a^x p(x')dx' \quad (22)$$

En dérivant la fonction de répartition, on obtient :

$$p(x) = \frac{dF}{dx} \quad (23)$$

La probabilité d'obtenir une valeur x donnée dans l'intervalle [a,b] est nulle. En pratique, on s'intéresse à la probabilité d'obtenir une valeur entre x et x+δx, égale approximativement à :

$$\delta P \approx p(x)\delta x \quad (24)$$

L'espérance d'une variable aléatoire f(X) est définie par :

$$E(f(X)) = \int_a^b p(x) f(x) dx \qquad (25)$$

La variance se définit comme pour une variable discrète.

Une densité de probabilité uniforme est constante sur l'intervalle [a,b]. Si les bornes de cet intervalle ne sont pas à l'infini, la densité de probabilité est obtenue avec la condition de normalisation :

$$p(x) = \frac{1}{b-a} \qquad (26)$$

L'espérance de x est alors :

$$E(x) = \int_a^b \frac{x\,dx}{b-a} = \frac{a+b}{2} \qquad (27)$$

La distribution gaussienne (ou normale) est définie sur l'intervalle $-\infty, \infty$ par :

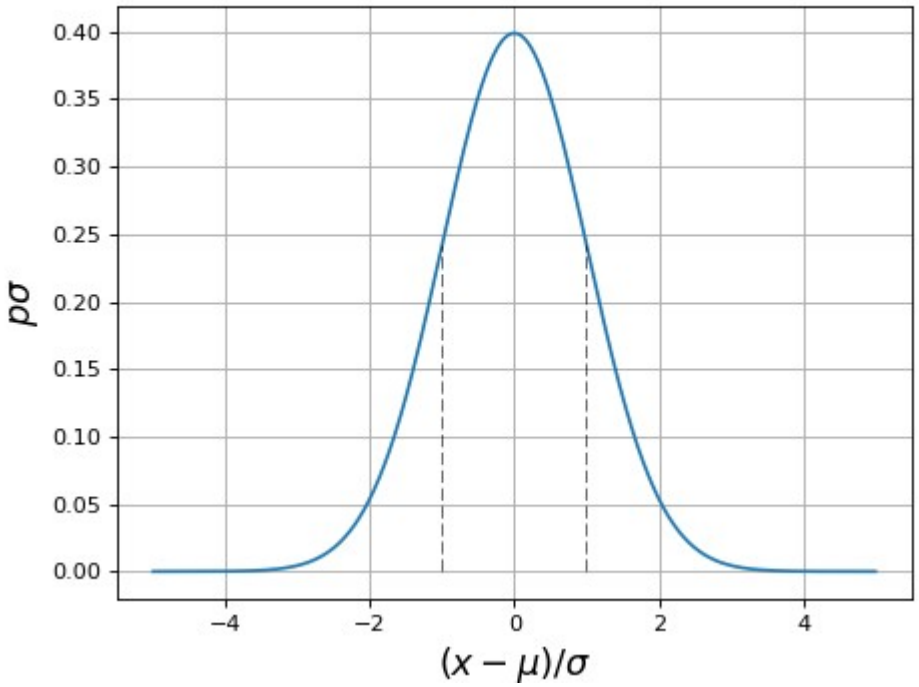
$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad (28)$$

Elle vérifie la condition de normalisation et on a :

$$E(x) = \mu \qquad (29)$$
$$var(x) = \sigma^2 \qquad (30)$$

Voici la représentation graphique de cette densité de probabilité en fonction de $(x-\mu)/\sigma$.

```
from matplotlib.pyplot import *
import numpy
x=numpy.linspace(-5,5,1000)
p=1/(numpy.sqrt(2*numpy.pi))*numpy.exp(-x**2/2)
figure()
plot(x,p)
p1=1/(numpy.sqrt(2*numpy.pi))*numpy.exp(-1/2)
plot([-1,-1],[0,p1],"k--",linewidth=0.5)
plot([1,1],[0,p1],"k--",linewidth=0.5)
xlabel(r"$(x-\mu)/\sigma$",fontsize=16)
ylabel(r"$p\sigma$",fontsize=16)
grid()
```



La probabilité d'obtenir une valeur dans un intervalle de largeur 2σcentré sur l'espérance est :

$$P(\mu - \sigma \leq x \leq \mu + \sigma) = \int_{\mu-\sigma}^{\mu+\sigma} p(x)dx \approx 0,68 \qquad (31)$$

La loi normale est courramment utilisée pour représenter les variations aléatoires des mesures d'une grandeur physique (incertitude expérimentale). L'incertitude type est par définition l'écart type $\sqrt{\sigma}$. Soit m la valeur mesurée, obtenue soit par une mesure unique, soit par une moyenne de mesures répétées. m est une estimation de l'espérance. Il faut aussi obtenir une estimation de l'écart type, appelée incertitude de la mesure, et notée Δm. Le résultat de la mesure est généralement présenté en disant que la valeur de la grandeur mesurée se trouve dans l'intervalle $[m - \Delta m, m + \Delta m]$ avec une probabilité de 0,68. On peut aussi écrire que la valeur de la grandeur mesurée se trouve dans l'intervalle $[m - 1,96\Delta m, m + 1,96\Delta m]$avec une probabilité de 0,95.

3.10.3.2. Simulation de Monte-Carlo

Principe

Considérons la simulation d'une variable aléatoire continue définie sur l'intervalle [a,b]. Les nombres réels sont représentés par des nombres à virgule flottante, qui ont bien sûr une précision limitée. Si l'on note δx cette précision, l'intervalle [a,b] est divisé en M sous-intervalles égaux de largeur δx. Le tirage d'un nombre à virgule flottante dans cet intervalle revient donc à tirer un entier compris entre 0 et M. La probabilité d'obtenir un nombre x est égale à p(x)δx.

La fonction random.uniform(a,b) délivre un flottant avec une densité de probabilité uniforme sur l'intervalle [a,b]. On peut aussi utiliser la fonction random.random() qui fait la même chose sur l'intervalle [0,1[(la valeur 1 est exclue).

La fonction random.gauss(mu,sigma) délivre un flottant avec la distribution de Gauss.

L'estimation de l'espérance et de la variance se fait comme déjà expliqué pour une variable aléatoire discrète.

Exemple

On considère comme exemple le tirage de nombres aléatoires avec la loi de Gauss. La fonction suivante fait N tirages. Elle calcule la moyenne empirique, la variance empirique, et l'intervalle de confiance à 95 pour cent pour l'espérance.

La fonction génère aussi un histogramme des échantillons. Pour générer un histogramme, il faut choisir un intervalle [a,b] et des classes de valeur dans cet intervalle. Supposons que ces classes soient les n_b sous intervalles de largeur $h=(b-a)/n_b$. Notons H_j le nombre de tirages donnant une valeur dans l'intervalle $[a+jh,a+(j+1)h]$ avec j variant de 0 à n_b-1. Lorsque les valeurs de la variable aléatoire ne sont pas dans un intervalle borné, il faut choisir des valeurs de a et b. Une première approche consiste à générer tous les échantillons en les stockant dans un tableau puis à choisir le minimum de le maximum des valeurs des échantillons pour a et b. Il est préférable de fixer la largeur des classes (h) et donc de calculer n_b en conséquence. Cette approche a deux inconvénients : elle nécessite le stockage des échantillons; elle ne permet pas de générer un histogramme qui se met à jour au fûr et à mesure de la génération des échantillons. Nous préférons calculer l'histogramme en fixant a priori les valeurs de a et b. L'inconvénient de cette approche est le risque d'avoir des valeurs d'échantillons en dehors de cet intervalle. L'intervalle doit donc être choisi assez grand pour que ce risque soit négligeable.

Le nombre de tirages H_j dans une classe est finalement divisé par le nombre de tirages total (comptés dans l'histogramme) afin d'obtenir une évaluation de la probabilité d'obtenir un tirage dans cette classe.

```
import numpy

def tirages_gauss(N,mu,sigma,a,b,nb):
    somme = 0.0
    somme2 = 0.0
    h = (b-a)/nb
    H = numpy.zeros(nb)
    n = 0
    for i in range(N):
        x = random.gauss(mu,sigma)
        somme += x
        somme2 += x*x
        j = int((x-a)/h)
        if j>=0 and j<nb:
            H[j] += 1
            n += 1
    moyenne = somme*1.0/N
    variance = somme2*1.0/N-moyenne*moyenne
```

```
    ecart = 1.96*math.sqrt(variance*1.0/N)
    H=H/n
    x=numpy.linspace(a,b,nb)
    return (moyenne,variance,ecart,x,H)
```

Voici un exemple :

```
mu=1.0
sigma=0.1
a=0.5
b=1.5
nb=50
(m,v,e,x,H) = tirages_gauss(10000,mu,sigma,a,b,nb)

print((m,v,e))
--> (1.0006659135758849, 0.010047662063773277, 0.001964665329876603)
```

On obtient ainsi l'estimation de l'espérance suivante :

$$E(X) = 1.001 \pm 0.002 \quad (32)$$

Voici une représentation graphique de l'histogramme :

```
from matplotlib.pyplot import *
figure()
plot(x,H,"o")
grid()
xlabel("x")
ylabel("P")
```

Un histogramme d'une variable aléatoire continue permet d'évaluer la densité. Pour cela, il faut diviser chaque probabilité par la largeur du sous-intervalle correspondant. Dans le cas présent, on peut faire une comparaison avec la densité de probabilité :

```
h=(b-a)/nb
H=H/h
figure()
plot(x,H,"o")
x=numpy.linspace(a,b,1000)
p=1/(numpy.sqrt(2*numpy.pi)*sigma)*numpy.exp(-(x-mu)**2/(2*sigma**2))
plot(x,p,"k-")
grid()
xlabel("x")
ylabel("p")
```

3.10.3.3. Échantillonnage d'une densité non uniforme

Une distribution continue de densité non uniforme peut être échantillonnée en inversant la fonction de répartition, définie par :

$$F(x)=\int_a^x p(x')dx' \quad (33)$$

On suppose que la densité de probabilité est strictement positive. La fonction de répartition est alors continue et strictement croissante. De plus F(a)=0 et F(b)=1. Il s'en suit qu'elle admet une fonction inverse F^{-1} définie sur l'intervalle [0,1] et à valeurs dans [a,b]. Considérons alors une variable aléatoire U de densité uniforme sur l'intervalle [0,1] et soit la variable X définie par :

$$X=F^{-1}(U) \quad (34)$$

La probabilité d'obtenir x compris entre x_1 et x_2 est :

$$\begin{aligned} P(x_1 \le x \le x_2) &= P(x_1 \le F^{-1}(u) \le x_2) && (35)\\ &= P(F(x_1) \le u \le F(x_2)) && (36)\\ &= F(x_2)-F(x_1) && (37)\\ &= \int_{x_1}^{x_2} p(x)dx && (38) \end{aligned}$$

ce qui montre que X a la densité p. Si la fonction de répartition est inversible analytiquement, on peut obtenir l'échantillonnage à partir d'un échantillonnage uniforme sur l'intervalle [0,1].

Considérons par exemple une densité de probabilité proportionnelle à x sur l'intervalle [0,1] :

$$p(x)=ax \quad (39)$$

En écrivant la condition de normalisation on obtient :

$$p(x)=2x \quad (40)$$

La fonction de répartition est :

$$F(x)=\int_0^x 2x'dx'=x^2 \quad (41)$$

L'inversion s'écrit :

$$x=\sqrt{u} \quad (42)$$

où u est tiré aléatoirement avec une densité uniforme sur l'intervalle [0,1].

Pour tester l'échantillonnage, on fait un histogramme avec plusieurs milliers de tirages :

```
import numpy.random
```

```
N = 100000
x = numpy.sqrt(numpy.random.random_sample(N))
figure()
hist(x,100)
```

Si l'inverse de la fonction de répartition n'est pas disponible, on peut la calculer numériquement et la stocker dans une table (sous forme discrète).

Pour la génération de nombres suivant une loi discrète, voir Échantillonnage des distributions de probabilité.

II. Exercices

Il y a une multitude d'exercices corrigés à étudier. Pour les voir, il faut aller sur Google et taper :

Tests statistiques –Pages personnelles Université Rennes 2

C'est un document pdf de 146 pages. Il faut étudier les exercices suivants :

- Problèmes corrigés de la page 25 à 44 et de la page 63 à 70
- Annales corrigées de la page 87 à la page 132.