NOTIONS ELEMENTAIRES DE LA STATISTIQUE

J.SAAB
ISSAE - Cnam Liban

Abstract

Dans ces papiers, nous donnons un résumé simple et très accessible des notions élémentaires de la statistique. Nous introduisons les lois de probabilité dans le but de procéder à des analyses des données.

1 Initiation à la Statistique

La statistique est un ensemble d'outils et de méthodes mathématiques pour quantifier des objets à des fins analytiques. La statistique consiste à trouver des résultats scientifiques sur un échantillon et de les généraliser, grâce à des méthodes scientifiques, sur une population.

Une étude statistique peut être quantitative (comme par exemple une étude sur le poids des individus dans une population), ou qualitative (comme par exemple une étude sur le statut civil des individus dans une population)

Vocabulaire: Population: c'est l'ensemble des individus sur lesquels on voudrait tirer un résultat lors d'une étude statistique

Échantillon: c'est une partie de la population sur laquelle on mène une investigation statistique afin de tirer des conclusions sur la population

Variable statistique: c'est une caractéristique qui pourrait prendre différentes valeurs en fonction du temps, de l'endroit ou de la situation

Individu: c'est tout élément de l'échantillon

Informations Statistiques: Une variable statistique X peut être une question posée à un ensemble d'individus qui prends ses valeurs selon leurs réponses. L'ensemble de ces réponses définit un ensemble d'informations statistiques, qui sera organisé dans une table dite table de distribution fréquentielle. Cette table est constituée de deux colonnes, l'une contenant les différentes valeurs x_i (numériques ou qualitatives) que peut prendre la variable et l'autre contenant le nombre d'ocuurence n_i de chaque valeur x_i. Les valeurs x_i sont dites les caractéristiques de X et les valeurs n_i sont dites les effectifs de X.

Exemple 1

1. *Les valeurs suivantes représentent les réponses de* 20 *personnes sur leur statut civil:*

$$c, c, c, m, m, \qquad d, c, c, v, c, \qquad c, m, m, m, c, \qquad c, c, m, c, d$$

où c, m, d, v *désignent respectivement célibataire, mariée, divorcée et veuve.*

Ici, la variable statistique est une variable qualitative. La table de distribution fréquentielle (T.D.F) qui résume ces information est:

x_i	n_i
c	11
m	6
v	1
d	2
	$n = 20$

le nombre $n = \sum n_i$ *est appelé taille de l'échantillon. L'échantillon ici, est l'ensemble des personnes questionnées.*

2. *Les valeurs suivantes représentent la taille de* 20 *personnes:*

$151, 151, 155, 155, 155, \qquad 162, 162, 167, 167, 168, \qquad 170, 170, 170, 176\ , 176,$
$178, 178, 178, 180, 180$

Ici, la variable statistique est une variable quantitative. La table de distribution fréquentielle (T.D.F) qui résume ces information est:

x_i	n_i
151	2
155	3
162	2
167	2
168	1
170	3
176	2
178	3
180	2
	$n = 20$

On appelle fréquence de x_i la valeur $f_i = \dfrac{n_i}{n}$ et pourcentage de x_i la valeur $100.f_i$ On a toujours $n = \sum n_i,\ \sum f_i = 1$ et $\sum 100.f_i = 100$

Souvent les valeurs recueillies x_i sont assez proches et la taille de l'échantillon est assez importante, dans ce cas on regroupe les informations dans des classes, c'est à dire dans des intervalles. On choisit ces classes de sorte qu'une valeur x_i ne soit pas dans deux classes différentes et de préférences, on choisit des classes de même amplitude. Il ne faudrait pas prendre un assez grand nombre de classes ni un nombre petit de classes, aussi, les classes devraient couvrir toutes les informations recueillies.

Exemple 2 *les valeurs suivantes repésentent la taille de* 60 *personnes:*

$151, 151, 155, 155, 155$	$162, 162, 167, 167, 168$	$170, 170, 170, 176$, 176
$178, 178, 178, 180, 180$	$180, 180, 180, 181, 181$	$181, 181, 181, 182, 182$
$182, 182, 182, 182, 182$	$182, 182, 182, 184, 184$	$184, 184, 184, 184, 185$
$186, 187, 187, 187, 188$	$188, 188, 189, 190, 190$	$190, 190, 191, 192, 193$

la valeur minimale est $x_{\min} = 151$ *et la valeur maximale est* $x_{\max} = 193$. *Le diamètre de cette série statistique est* $D = x_{\max} - x_{\min} = 42$. *Si on voudrait établir la T.D.F. en* $k = 7$ *classes, alors chaque classe doit être d'amplitude* $l = \frac{L}{k} = 6$. *On obtient:*

$[a_i, b_i[$	n_i
$[151, 157[$	5
$[157, 163[$	2
$[163, 169[$	3
$[169, 175[$	3
$[175, 181[$	15
$[181, 187[$	21
$[187, 193[$	11
	$n = 60$

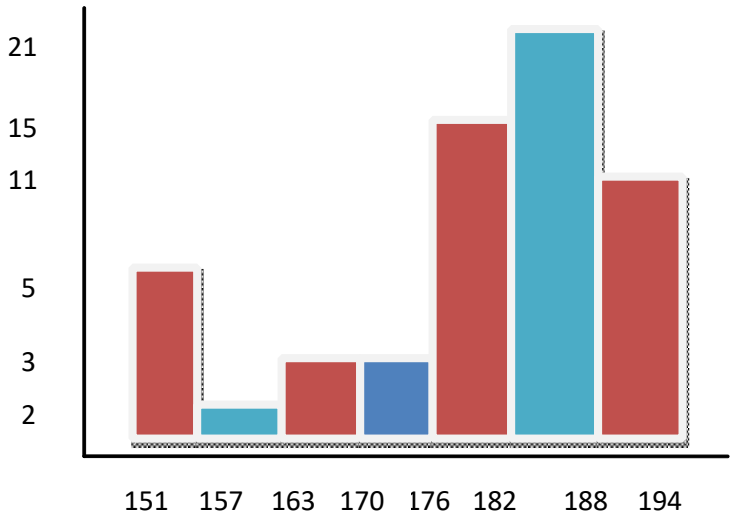
Définition 1 *On appelle centre d'une classe* $[a_i, b_i[$ *la valeur* $x_i = \dfrac{a_i + b_i}{2}$. *Pour tout* $i = 1, 2, \cdots k$, *on définit* n_{icc} *et* n_{icd} *dites respectivement, les effectifs cumulés croissants et décroissants, par:* $n_{icc} = \sum_{j=1}^{i} n_j$ *et* $n_{icd} = \sum_{j=i}^{k} n_j$. *La valeur* n_{icc} *désigne le nombre d'individus ayant un caractéristique* $< b_i$ *et* n_{icd} *désigne le nombre d'individus ayant un caractéristique* $\geq a_i$. *Aussi, on définit les fréquences cumulées croissantes et décroissantes par:* $f_{icc} = \dfrac{n_{icc}}{n}$ *et* $f_{icd} = \dfrac{n_{icd}}{n}$

Exemple 3 *Dans l'exemple précédent, si on ajoute les effectifs cumulés et les fréquences cumulées, on obtient:*

$[a_i, b_i[$	n_i	n_{icc}	n_{icd}	f_{icc}	f_{icd}
$[151, 157[$	5	5	60	0.083	1
$[157, 163[$	2	7	55	0.116	0.916
$[163, 169[$	3	10	53	0.166	0.883
$[169, 175[$	3	13	50	0.216	0.833
$[175, 181[$	15	28	47	0.466	0.783
$[181, 187[$	21	49	32	0.816	0.533
$[187, 193[$	11	60	11	1	0.183
	$n = 60$				

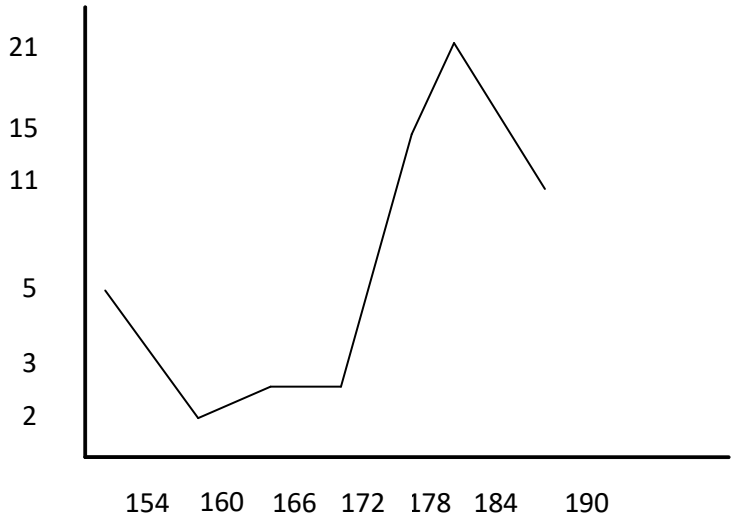
Représentation graphique: La T.D.F peut être résumée sur une figure. Au lieu d'étudier des chiffres, il est plus commode d'étudier une courbe. Il y a plusieurs représentations graphiques pour une T.D.F. mais on va se contenter d'en donner les 4 principales.

1. **Histogramme:** Il s'agit d'une figure dans le repère orthonormé xoy où sur l'axe des x figurent les valeurs de a_i et b_i et sur l'axe des y figurent les n_i. On construit de rectangles, chaque rectangle est de largeur (a_i, b_i) et de hauteur n_i. L'histogramme relatif à l'exemple précédent est:



2. **Polygone des fréquences**: Il s'agit de construire une courbe en lignes brisées où l'on joint les points (x_i, n_i) avec x_i est le centre de la classe

$[a_i, b_i[$. Le polygone relatif à l'exemple précédent est:



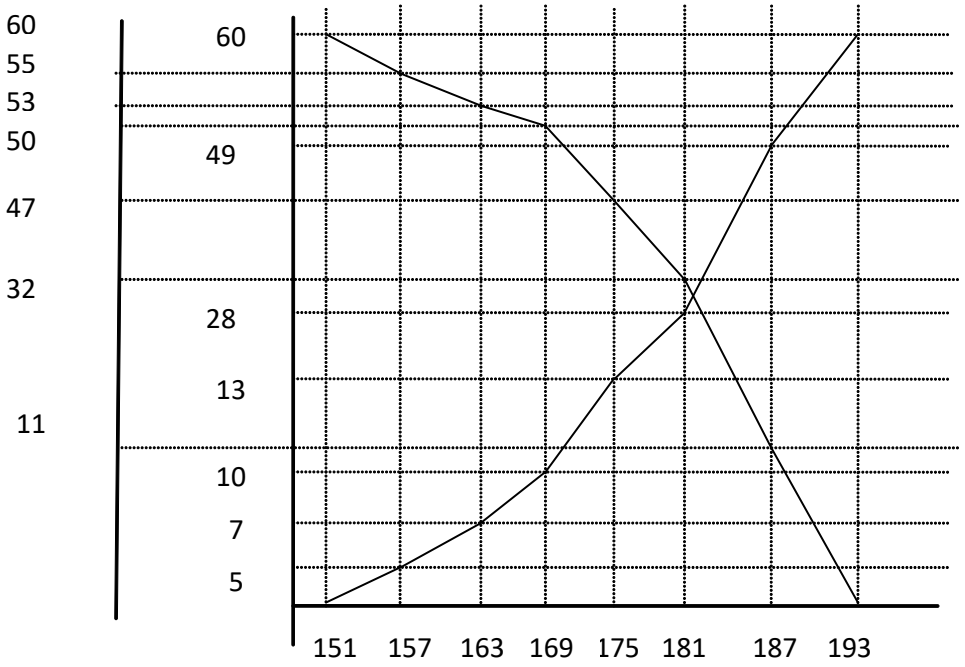
3. **Polygone cumulé croissant et Polygone cumulé décroissant**: Le premier est formé des lignes brisées joingnant les points

$$(a_i, \text{nombre d}'\text{individus ayant un effectif } < a_i)$$

où les a_i désignent toutes valeurs qui représentent les bords des classes. Le second est formé des lignes brisées joignant les points

$$(a_i, \text{nombre d}'\text{individus ayant un effectif } \geq a_i)$$

Les polygones relatifs à l'exemple précédent:



1.1 Mesures de la tendance centrale

Loin des chiffres donnés dans une T.D.F, un certain nombre de paramètres pourrait nous faire comprendre la tendance d'une série statistique. Parmis ces paramètres, on va étudier:

Le mode: Dans le cas d'une série non regroupée, c'est la caractéristique la plus répétée dans la série. Comme dans le cas de l'exemple sur le statut civil, le mode est:

$$M_0 = c$$

puisque c est repété 11 fois. Noter qu'on pourrait avoir plusieurs modes.

Dans le cas, d'une série regroupée, on définit la classe modale comme étant la classe à plus haute fréquence, c'est à dire celle qui correspond au $\max(n_i)$. Dans l'exemple sur la taille de 60 personne, la classe modale est

$$[a_M, b_M[= [181, 187[$$

correspondant à $n_6 = 21$. Cet effectif est noté n_M c'est à dire l'effectif modal. On désigne par n_{M-1} et n_{M+1} les effectifs juste avant et juste après n_M Dans le même exemple, on a $n_{M-1} = 15$ et $n_{M+1} = 11$. Graphiquement, le mode est l'abscisse du point de croisement de deux droites (AB) et (CD) avec:

$$A(a_M, n_{M-1}),\ B(b_M, n_M),\ C(b_M, n_{M+1}),\ D(a_M, n_M)$$

dans notre exemple, il faudrait joindre les points $A(181, 15)$ à $B(187, 21)$ et les points $C(187, 11)$ à $D(181, 21)$ et le point de rencontre des droites (AB) et (CD) a pour abscisse le mode M_0 qui devrait appartenir à la classe modale.

Dans le cas où les classes ont une **même amplitude** l, un calcul simple, mène à la formule suivante:

$$M_o = a_M + l\frac{n_M - n_{M-1}}{(n_M - n_{M-1}) + (n_M - n_{M+1})}$$

Dans le cas où les classes ont des longueurs différentes l_i :

$$M_o = a_M + (b_M - a_M).\frac{n_M^{'} - n_{M-1}^{'}}{(n_M^{'} - n_{M-1}^{'}) + (n_M^{'} - n_{M+1}^{'})}$$

où $n_i^{'} = \dfrac{n_i}{l_i}$

Exemple 4 *Dans l'exemple sur la taille de* 60 *personnes, le mode est*

$$M_0 = 181 + 6.\frac{21 - 15}{(21 - 15) + (21 - 11)} = 183.25 \in [181, 187[$$

Exemple 5 *Supposons maintenant que cette série est regroupée dans* 6 *classes de longueurs respectives:* $l_1 = l_2 = 6, l_3 = l_5 = 10, l_4 = l_6 = 5$

$[a_i, b_i[$	n_i	*on pose:*	$n_i^{'}$
$[151, 157[$	5		$5/6 = 0.833$
$[157, 163[$	2		$2/6 = 0.333$
$[163, 173[$	6		$6/10 = 0.6$
$[173, 178[$	2		$2/5 = 0.4$
$[178, 188[$	34		$34/10 = 3.4$
$[188, 193]$	11		$11/5 = 2.2$
	$n = 60$		

Ainsi:

$$M_o = 178 + 10.\frac{3.4 - 0.4}{(3.4 - 0.4) + (3.4 - 2.2)} = 185.14$$

Remarque 1 *Dans le cas de la présence de deux classes modales, on pourrait chercher deux modes selon la méthode précédente*

La médiane: Graphiquement, la médiane est l'abscisse du point de rencontre de deux polygônes cumulés. Elle correspond à $X = M_e$ tel que le nombre d'individus ayant pour effectifs $\leq M_e$ est égal au nombre d'individus ayant pour effectifs $\geq M_e$, égal aussi à $\frac{n}{2}$. Pour obtenir la médiane, on va noter par $n_{m-1,cc}$ et n_{mcc} les effectifs cumulés croissants tels que

$$n_{m-1,cc} \leq \frac{n}{2} \leq n_{m,cc}$$

si $n_{m-1,cc}$ et $n_{m,cc}$ correspondent à la classe $[a_m, b_m[$ alors cette classe sera dite la classe médiane. On peut donc, déterminer l'équation de la droite (faisant partie du polygône cumulé croissant) (AB) avec $A(a_m, n_{m-1,cc})$ et $B(b_m, n_{m,cc})$. La médiane est le point de la classe $[a_m, b_m[$ ayant pour ordonnée $\frac{n}{2}$. Soit:

$$M_e = a_m + (b_m - a_m)\frac{\frac{n}{2} - n_{m-1,cc}}{n_{m,cc} - n_{m-1,cc}}$$

Dans l'exemple concernant la taille de 60 personnes, on a $\frac{n}{2} = 30$ et $n_{m-1,cc} = 28 \leq 30 \leq 49 = n_{m,cc}$ où 28 correspond à $a_m = 181$ et 49 correspond à $b_m = 187$. On obtient:

$$M_e = 181 + 6.\frac{30 - 28}{49 - 28} = 181.571\ cm$$

La moyenne: La moyenne générale est notée $\bar{x}$ et donnée par:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k}(n_i x_i) \ , \ \ x_i \text{ étant le centre de } [a_i, b_i[$$

dans le même exemple, la taille moyenne dans cet échantillon est $\bar{x} = \frac{1}{60}[5 \times 154 + 2 \times 160 + \cdots 11 \times 190] = 178.8 \ cm$

1.2 Mesure de dispersion

On va se contenter de donner le paramètre principal qui est l'écart type. On définit la variance d'une série statistique par

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

L'écart type de cette série est

$$\sigma = \sqrt{V(X)}$$

L'unité de $V(X)$ est le carré de l'unité de x_i c'est pour quoi on a introduit le paramètre σ ayant la même unité que x_i. On pourrait comprendre l'écart type comme étant le paramètre mesurant la moyenne de dispersion des x_i relativement à la moyenne. Dans l'exemple précédent, la variance est:

$$V(X) = \frac{1}{60}[5 \times 154^2 + 2 \times 160^2 + \cdots + 11 \times 190^2] - 178.8^2 = 106.16 \ cm^2$$

et donc

$$\sigma(X) = 10.30 \ cm$$

Un autre paramètre de dispersion sans unité est le coefficient de variation

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

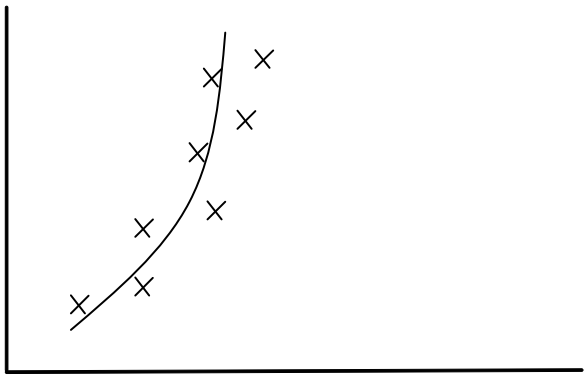
Application: Comment peut - on imaginer une série statistique lorsqu'on a ses paramètres mais on n'a pas la valeurs de ses caractéristiques: On donne une série satistique de 60 valeurs telle que:

$$x_{\min} = 151, x_{\max} = 193, \bar{x} = 178.8, \ M_o = 183.25, \ M_e = 181.571, \ \sigma = 10.30$$

On en déduit les remarques suivantes: beaucoup d'individus ont une taille autour de $183cm$, la moitié des individus mesure moins que 181.5 et l'autre moitié mesure plus que 181.5 cm, la moyenne de dispersion relativement à la moyenne est 10.30 donc, il ne s'agit pas d'une catégorie homogène, il y a des individus petits et des individus grands par rapport à la taille moyenne 183.

1.3 Droite de regression

On se donne une série statistique double, basée sur deux variables statistiques (X, Y), comme par exemple si on demande à chaque personne sa taille et son poids. Les valeurs obtenues après avoir posé ces deux questions à n individus s'appellent le nuage statisique. Ce nuage admet une forme géométrique définissant son allure. Une coube de regression est la courbe la plus représentative de ce nuage, comme on voit dans la figure suivante



On distingue plusieurs cas:

La tendance du nuage est linéaire: c'est à dire, il peut être approché par une droite dite droite de regression de y en x

$$(D) : y = ax + b$$

la méthode des moindres carrés qui ne sera pas développée dans ce cours donne:

$$a = \frac{cov(x,y)}{V(x)},\ b = \overline{y} - a\overline{x},\ cov(x,y) = \frac{1}{n}\sum n_{ii}x_i y_i - \overline{x}.\overline{y}$$

Dans le cas où les x_i sont regroupés dans n classes et les y_j sont aussi regroupés dans m classes on obtient:

$$a = \frac{cov(x,y)}{V(x)},\ b = \overline{y} - a\overline{x},\ cov(x,y) = \frac{1}{n}\sum_{i,j} n_{ij}x_i y_j - \overline{x}.\overline{y}$$

où n_{ij} est l'effectif correspondant à la $i^{ième}$ classe de X et la $j^{ième}$ classe de Y, les x_i et y_j sont les centres de ces classes repectivement.

On appelle droite de regression de x en y la droite

$$x = a^{'}y + b^{'}$$

avec

$$(D^{'}) :\ a^{'} = \frac{cov(x,y)}{V(y)},\ b^{'} = \overline{x} - a^{'}\overline{y}$$

Lorsque (D) et (D') sont proches, on déduit que X et Y sont fortement linéairement liées, c'est à dire, on est sûr que la relation entre X et Y peut être définie par une droite. Analytiquement, on pose $\rho = \frac{cov(x,y)}{\sigma(x)\sigma(y)}$ dit cofficient de correlation. on a $a.a' = \rho^2$. Ainsi, lorsque $|\rho|$ est proche de 1 la relation entre X et Y est linéaire et lorsque $|\rho|$ est proche de 0 cette relation est loin d'être linéaire. Dans la pratique on adopte la règle suivante:

$|\rho| \geq 0.75 \Longrightarrow X$ et Y sont fortement linéairement correllées

$|\rho| < 0.75 \Longrightarrow$ cette relation n'est pas linéaire

La tendance du nuage n'est pas linéaire: Si la tendance n'est pas linéaire, on pose la courbe qui le représente $(C) : y = ba^x$. Pour obtenir a et b il suffit de passer à la fonction ln

$$\ln y = \ln b + x.\ln a$$

On pose $Y_i = \ln y_i,\ X_i = x_i,\ A = \ln a$ et $B = \ln b$. La courbe de regression (C) devient la droite de regression (D) dans le nouveau repère

$$(C) : Y = AX + B$$

avec

$$A = \frac{cov(X,Y)}{V(X)},\ b = \overline{Y} - a\overline{X}$$

ainsi $a = e^A$ et $b = e^B$

Exemple 6 *Une étude sur* 500 *voitures a donné le prix* Y *en million de L.L relativement à la puissance du moteur* X *en nombre de chevaux*

$X \diagdown Y$	$15-30$	$30-35$	$35-40$	$40-45$	$45-50$	$50-55$	$n_{i\ .} = \sum_{j=1}^{n=6} n_{i\ j}$
$3-4$	22	8					30
$5-6$	36	41	20	13			110
$7-8$	12	36	68	50	22	22	210
$9-10$			12	26	32	30	100
$11-12$				6	14	20	40
$13-15$					2	8	10
$n._{\ j}$	70	85	100	95	70	80	500

$$\bar{x} = \frac{1}{n}\sum n_{i\ .} x_i = 7.67,\ \bar{y} = \frac{1}{n}\sum n._{\ j}\ y_j = 36.2$$

$$V(x) = \frac{1}{n}\sum n_{i\ .} x_i^2 - \bar{x}^2 = 4.735,\quad V(y) = \frac{1}{n}\sum n._{\ j}\ y_j^2 - \bar{y}^2 = 57.5$$

$$Cov(x,y) = \frac{1}{500}[3.5(22\times 22.5 + 8\times 32.5) + \cdots + 14(2\times 47.5 + 8\times 52.5)] - 7.67\times 36.2 = 13.9$$

$$\rho = \frac{cov(x,y)}{\sqrt{V(x).V(y)}} = 0.85 > 0.75$$

On en déduit que la relation entre x et y est bien linéaire donnée par la droite de regression de y en x :

$$y = ax + b \text{ avec } a = \frac{13.9}{4.736} = 2.93 \text{ et } b = 36.2 - 2.93 \times 7.67 = 13.72$$
$$(D) : y = 2.93\ x + 13.72$$

et si on voudrait estimer le prix d'une voiture ayant la puissance de 20 *chevaux, soit* $y = 2.93 \times 20 + 13.72 = 72.32$ *millions de L.L.*

1.4 Introduction à la probabilité

Une expérience aléatoire est une expérience dont on ne peut pas prévoir le résultat. En lançant un dé on ne peut pas savoir si on va obtenir $1, 2, ...,$ou 6. L'ensemble des valeurs possibles d'une expérience aléatoire est dit "univers" et souvent noté par Ω. Le couple $(\Omega, P(\Omega))$ est dit espace probabilisable, où $P(\Omega)$ est l'ensemble des parties de Ω et si p est une probabilité sur Ω alors le triplet $(\Omega, P(\Omega), p)$ est dit espace probabilisé.

Une variable aléatoire discrète X sur $\Omega = \{x_i,\ i \in I\}$, est une application $X : \Omega \longrightarrow \mathbb{Q}$ ou une partie finie ou infinie de $\mathbb{Q}$ et une variable aléatoire continue X est une application $X : \Omega \longrightarrow]a, b[$ ou $\mathbb{R}$.

Lois discrètes de probabilité:

1-Loi de Bernouilli: Une expérience aléatoire ayant deux résultats possibles est dite "expérience de Bernouilli". L'ensemble de ses résultats est noté $\Omega = \{S, E\}$ où S désigne le succès et E désigne l'échec. On définit la variable aléatoire discrète (v.a.d.) $X : \Omega \longrightarrow \{0, 1\}$ avec $X(S) = 1$ et $X(E) = 0$. Si $p(S) = p$ on dit que $P(X = 1) = p$ et ainsi $p(E) = p(X = 0) = 1 - p = q$. On définit l'espérance de X notée $E(X)$ par:

$$E(X) = \sum x_i p_i \quad \text{où } p_i = p(X = x_i)$$

et la variance de X est

$$V(X) = E(X^2) - E(X)^2 = \sum x_i^2 p_i - (E(X))^2$$

Remarque 2

1. $E(aX + bY + c) = aE(X) + bE(Y) + c$ *pour tous* $a, b, c \in \mathbb{R}$ *et pour toutes v.a* X *et* Y

2. $V(aX+bY+c)=a^2V(X)+b^2V(Y)$

Dans le cas de la variabe de Bernouilli $E(X)=p$ et $V(X)=pq$. On note $X \rightsquigarrow B(p)$ et on dit que X suit la loi de Bernouilli de paramètre p

Exemple: Le lancement d'une pièce de monnaie. $\Omega=\{pile,\ face\}$ on pose $S=pile$ et $E=face$

2-Loi Binômiale: La loi Binômiale est formée par la répétition n fois, d'une façon indépendante, d'une expérience de Bernouilli. On pose $X=$ le nombre de S obtenus. Si $p=p(S)$ alors on dit que X suit une loi Binômiale de paramètres n et p et on note $X \rightsquigarrow B(n,p)$. Ainsi pour tout $\Omega=\{0,1,2,\cdots\}$ et pour tout $k\in\Omega$ on a

$$p(X=k)=C_n^k\ p^k\ q^{n-k}$$

avec $C_n^k=\frac{n!}{k!(n-k)!}$ dite combinaison de k parmi n.

Pour une telle variable, il est facile de vérifier que $E(X)=np$ et $V(X)=npq$

Exemple: Dans une famille de 6 enfants, quelle est la probabilité d'avoir 3 garçons, sachant que la probabilité d'avoir un garçon dans cette famille est 0.48.

En effet, on associe S au fait d'avoir un garçon, $p(S)=0.48$.On pose $X=$ nombre de $S=0,1,..,6$ alors $X \rightsquigarrow B(6,\ 0.48)$

$$p(X=3)=C_6^3(0.48)^3.(0.52)^3=0.311$$

le nombre moyen de garçons dans cette famille est $E(X)=np=6\times 0.48=2.88$

3-Loi de Poisson: Si $X \rightsquigarrow B(n,p)$ avec $n >>$ et $p <<$ de sorte que $np \simeq npq$ (souvent $np<5$) , on dit que X suit une loi de Poisson de paramètre $\lambda=np$. Dans ce cas, on a

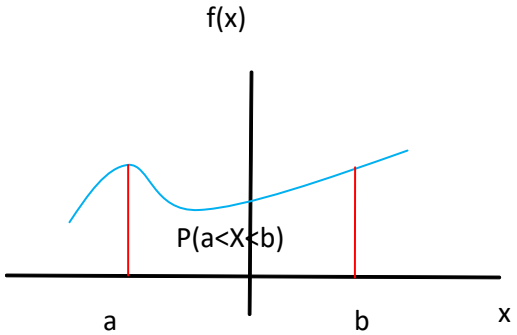
$$p(X=k)=e^{-\lambda}.\frac{\lambda^k}{k!}$$

Aussi $E(X)=V(X)=\lambda$. L'importance de cette loi est qu'elle est tabulée

$k\backslash\lambda$	1,0		2,0		3,0		4,0		5,0	
	P_k	F_k	P_k	F_k	P_k	F_k	P_k	F_k	P_k	F_k
0	36788	36788	13534	13534	4979	4979	1832	1832	674	674
1	36788	73576	27067	40601	14936	19915	7326	9158	3369	4043
2	18394	91970	27067	67668	22404	42319	14653	23810	8422	12465
3	6131	98101	18045	85712	22404	64723	19537	43347	14037	26503
4	1533	99634	9022	94735	16803	81526	19537	62884	17547	44049
5	307	99941	3609	98344	10082	91608	15629	78513	17547	61596
6	51	99992	1203	99547	5041	96649	10420	88933	14622	76218

Lois continues de probabilité: Dans le cas d'une v.a.c. $X : \Omega \longrightarrow I$, où I est intervalle fini ou infini, on associe une fonction $f : I \longrightarrow \mathbb{R}^+$ dite densité de X telle que:

1. $\int_I f(x)dx = 1$
2. Pour tous $a, b \in I$ avec $a < b$ on a $p(a \leq X \leq b) = \int_a^b f(x)dx$



L'espérence de X est donnée par

$$E(X) = \int_I xf(x)dx$$

et la variance

$$V(X) = E(X^2) - E(X)^2 = \int_I x^2 f(x)dx - E(X)^2$$

Aussi, on définit une fonction dite de répartition et donnée par

$$F(x) = p(X \leq x) = \int_{\inf I}^{x} f(t)dt$$

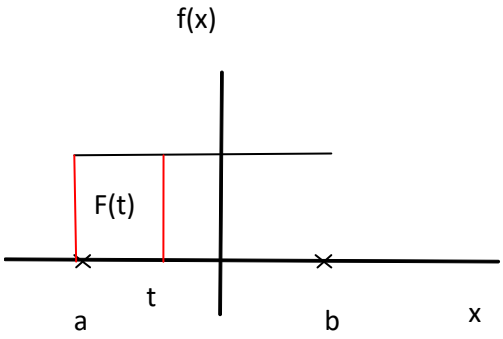
et ainsi,

$$p(a \leq X \leq b) = F(b) - F(a)$$

On va présenter quatre lois principales dans le cas d'une v.a.c.

1-Loi uniforme: On dit qu'une v.a.c. X suit la loi uniforme sur un intervalle $[a, b]$ et on note $X \rightsquigarrow U([a, b])$ si sa densité $f(x)$ est donnée par

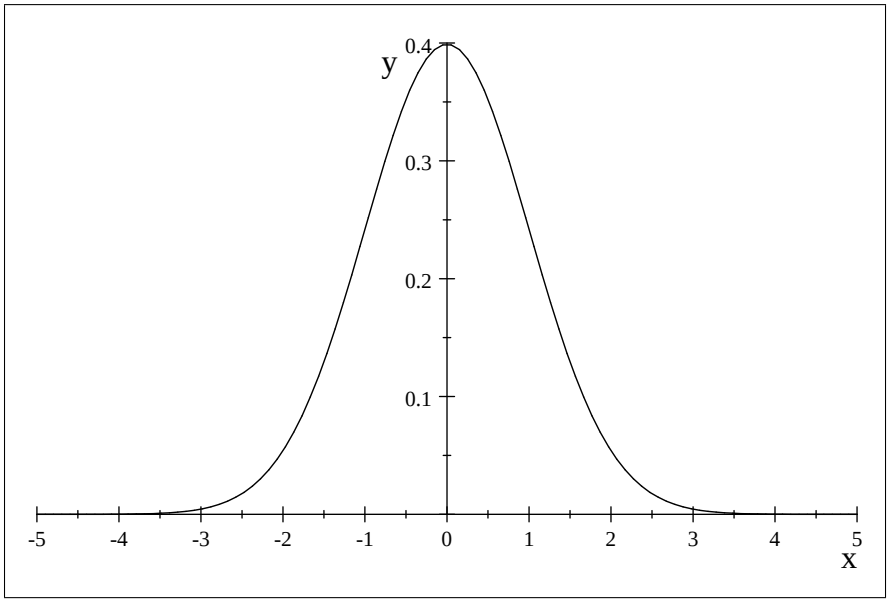
$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si} \quad a < x < b \\ 0 & \text{sinon} \end{cases}$$



2-Loi Normale: On dit qu'une v.a.c. X suit la loi normale de paramètres μ et σ et on note $X \rightsquigarrow N(\mu, \sigma)$ lorsque sa densité est donnée par

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\dfrac{(x-\mu)^2}{2\sigma}}, \quad \text{pour tous } x \in \mathbb{R}$$

On vérifie que $E(X) = \mu$ et que $V(X) = \sigma^2$. On ne s'intéresse qu'au cas où $\mu = 0$ et $\sigma = 1$ on dit dans ce cas que $X \rightsquigarrow N(0, 1)$, suit la loi normale centrée réduite. La densité devient: $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\dfrac{x^2}{2}}$



Remarquons que $x^{'}x$ est une asymptote lorsque $x \to \pm\infty$. La fonction de répartition de cette variable $F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}.dt$ est tabulée et donnée par:

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7793	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8906	0.8925	0.8943	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890

Utilisation

On lit les décimales dans les lignes, et les centièmes en colonnes. Par exemple, la valeur de $F(1.65)$ se trouve à l'intersection de la ligne 1.6 et de la colonne 0.05 - on trouve $F(1.65) = 0.9505$, à 10^{-4} près. Pour les valeurs négatives de x, on utilise la relation $F(-x) = 1 - F(x)$.

Remarque 3 *Lorsque* $X \rightsquigarrow N(\mu, \sigma)$ *alors la variable* $Z = \frac{X-\mu}{\sigma} \rightsquigarrow N(0, 1)$

Exemple 7 *On donne les v.a.* $X \rightsquigarrow N(-3, 2)$, $Y \rightsquigarrow N(7, 3)$ *et* $Z \rightsquigarrow N(40, 6)$. *Donner la loi de* $T = X - Y + Z$ *et déduire* $p(T \leq 20)$ *et* $p(20 < T < 35)$

en effet, $E(T) = E(X) - E(Y) + E(Z) = -3 - 7 + 40 = 30$ et $V(T) = V(X) + V(Y) + V(Z) = 4 + 9 + 36 = 49$ et par suite

$$T \rightsquigarrow N(30, 7)$$

Pour trouver les différentes probabilités relatives à T on doit d'abord définir la loi centrée réduite associée à T . Soit $T' = \dfrac{T - 30}{7}$. On a $T' \rightsquigarrow N(0, 1)$.

$$\begin{array}{rcl} p(T < 20) & = & p(T' < \frac{20-30}{7}) = p(T' < -1.428) \\ & = & p(T' > 1.428) = 1 - P(T' < 1.428) = 1 - F(1, 428) \end{array}$$

On connait d'après la table $F(1.42)$ et $F(1.43)$ et puisque $1.42 < 1.428 < 1.43$, il est préférable de poser

$$F(1.428) = \frac{F(1.42) + F(1.43)}{2} = \frac{0.9222 \quad + \quad 0.9236}{2} = 0.9229$$

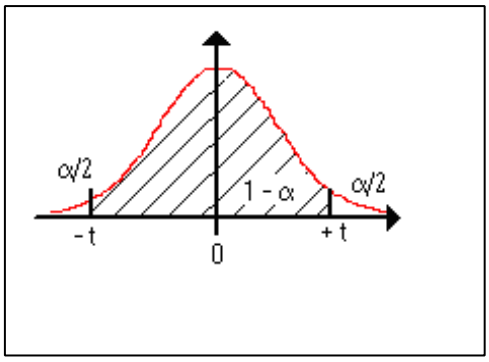
et il n'est pas faux d'approcher 1.428 à la valeur 1.43 et de poser $F(1.428) = 0.9236$. Et donc $F(T < 20) = 0.0771$.

De même

$$\begin{array}{rcl} p(20 < T < 35) & = & p(\frac{20-30}{7} < T' < \frac{35-30}{7}) = p(-1.428 < T' < 0.714) \\ & = & p(T' < 0.714) - p(T' < -1.428) \\ & = & F(0.714) - (1 - F(1.428) \end{array}$$

avec $F(0.714) \simeq F(0.71) = 0.7611$ et par suite $p(20 < T < 35) = 0.7611 - 1 + 0.9229 = 0.684$

3-Loi de Student: C'est une approximation de la loi normale, on note $X \rightsquigarrow st(\nu)$, et on dit que X suit la loi de Student de paramètre ν dit degré de liberté. On se contente de donner la table qui donne la valeur de t telle que $p(|X| > t) = \alpha$



.

	0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001	0.0001
1	1.000	3.078	6.314	12.706	31.281	63.657	127.32	318.31	636.62	6366.2
2	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	34.599	99.992
3	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924	28.000
4	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610	15.544
5	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869	11.178
6	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959	9.082
7	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408	7.885
8	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041	7.120
9	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781	6.594
10	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587	6.211
11	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437	5.921
12	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318	5.694
13	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221	5.513
14	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140	5.363
15	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073	5.239
16	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015	5.134
17	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965	5.044
18	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922	4.966
19	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883	4.897
20	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850	4.837
21	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819	4.784
22	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792	4.736
23	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768	4.693
24	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745	4.654
25	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725	4.619
30	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646	4.482
35	0.682	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591	4.389
40	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551	4.321
45	0.680	1.301	1.679	2.014	2.412	2.690	2.952	3.281	3.520	4.269
50	0.679	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496	4.228
60	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460	4.169
70	0.678	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435	4.127
80	0.678	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416	4.096
90	0.677	1.291	1.662	1.987	2.368	2.632	2.878	3.183	3.402	4.072
100	0.677	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390	4.053
150	0.676	1.287	1.655	1.976	2.351	2.609	2.849	3.145	3.357	3.998
200	0.676	1.286	1.653	1.972	2.345	2.601	2.839	3.131	3.340	3.970
300	0.675	1.284	1.650	1.968	2.339	2.592	2.828	3.118	3.323	3.944
500	0.675	1.283	1.648	1.965	2.334	2.586	2.820	3.107	3.310	3.922
1000	0.675	1.282	1.646	1.962	2.330	2.581	2.813	3.098	3.300	3.906
infini	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291	3.891

Utilisation

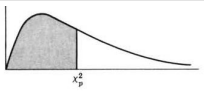
_ _

En fonction du nombre de degrés de liberté (qu'on lit sur la première colonne) et du risque d'erreur α (qu'on lit sur la première ligne), on trouve la valeur de l'écart t qui possède la probabilité α d'être dépassé en valeur absolue, $p(|X| > t) = \alpha$

4-Loi de chi deux: On note $X \rightsquigarrow \chi^2(p)$ et on dit que X suit la loi de chi deux de degré de liberté p. La table qu'on va introduire donne la valeur

notée χ^2_p telle que $p(X < \chi^2_p) = p$

VALEURS DES CENTILES (χ^2_p)
pour la
DISTRIBUTION du KHI-DEUX
en fonction du nombre ν de degrés de liberté
(aire en grisé = p)



ν	$\chi^2_{0,995}$	$\chi^2_{0,99}$	$\chi^2_{0,975}$	$\chi^2_{0,95}$	$\chi^2_{0,90}$	$\chi^2_{0,75}$	$\chi^2_{0,50}$	$\chi^2_{0,25}$	$\chi^2_{0,10}$	$\chi^2_{0,05}$	$\chi^2_{0,025}$	$\chi^2_{0,01}$	$\chi^2_{0,005}$
1	7,88	6,63	5,02	3,84	2,71	1,32	0,455	0,102	0,0158	0,0039	0,0010	0,0002	0,0000
2	10,6	9,21	7,38	5,99	4,61	2,77	1,39	0,575	0,211	0,103	0,0506	0,0201	0,0100
3	12,8	11,3	9,35	7,81	6,25	4,11	2,37	1,21	0,584	0,352	0,216	0,115	0,072
4	14,9	13,3	11,1	9,49	7,78	5,39	3,36	1,92	1,06	0,711	0,484	0,297	0,207
5	16,7	15,1	12,8	11,1	9,24	6,63	4,35	2,67	1,61	1,15	0,831	0,554	0,412
6	18,5	16,8	14,4	12,6	10,6	7,84	5,35	3,45	2,20	1,64	1,24	0,872	0,676
7	20,3	18,5	16,0	14,1	12,0	9,04	6,35	4,25	2,83	2,17	1,69	1,24	0,989
8	22,0	20,1	17,5	15,5	13,4	10,2	7,34	5,07	3,49	2,73	2,18	1,65	1,34
9	23,6	21,7	19,0	16,9	14,7	11,4	8,34	5,90	4,17	3,33	2,70	2,09	1,73
10	25,2	23,2	20,5	18,3	16,0	12,5	9,34	6,74	4,87	3,94	3,25	2,56	2,16
11	26,8	24,7	21,9	19,7	17,3	13,7	10,3	7,58	5,58	4,57	3,82	3,05	2,60
12	28,3	26,2	23,3	21,0	18,5	14,8	11,3	8,44	6,30	5,23	4,40	3,57	3,07
13	29,8	27,7	24,7	22,4	19,8	16,0	12,3	9,30	7,04	5,89	5,01	4,11	3,57
14	31,3	29,1	26,1	23,7	21,1	17,1	13,3	10,2	7,79	6,57	5,63	4,66	4,07
15	32,8	30,6	27,5	25,0	22,3	18,2	14,3	11,0	8,55	7,26	6,26	5,23	4,60
16	34,3	32,0	28,8	26,3	23,5	19,4	15,3	11,9	9,31	7,96	6,91	5,81	5,14
17	35,7	33,4	30,2	27,6	24,8	20,5	16,3	12,8	10,1	8,67	7,56	6,41	5,70
18	37,2	34,8	31,5	28,9	26,0	21,6	17,3	13,7	10,9	9,39	8,23	7,01	6,26
19	38,6	36,2	32,9	30,1	27,2	22,7	18,3	14,6	11,7	10,1	8,91	7,63	6,84
20	40,0	37,6	34,2	31,4	28,4	23,8	19,3	15,5	12,4	10,9	9,59	8,26	7,43
21	41,4	38,9	35,5	32,7	29,6	24,9	20,3	16,3	13,2	11,6	10,3	8,90	8,03
22	42,8	40,3	36,8	33,9	30,8	26,0	21,3	17,2	14,0	12,3	11,0	9,54	8,64
23	44,2	41,6	38,1	35,2	32,0	27,1	22,3	18,1	14,8	13,1	11,7	10,2	9,26
24	45,6	43,0	39,4	36,4	33,2	28,2	23,3	19,0	15,7	13,8	12,4	10,9	9,89
25	46,9	44,3	40,6	37,7	34,4	29,3	24,3	19,9	16,5	14,6	13,1	11,5	10,5
26	48,3	45,6	41,9	38,9	35,6	30,4	25,3	20,8	17,3	15,4	13,8	12,2	11,2
27	49,6	47,0	43,2	40,1	36,7	31,5	26,3	21,7	18,1	16,2	14,6	12,9	11,8
28	51,0	48,3	44,5	41,3	37,9	32,6	27,3	22,7	18,9	16,9	15,3	13,6	12,5
29	52,3	49,6	45,7	42,6	39,1	33,7	28,3	23,6	19,8	17,7	16,0	14,3	13,1
30	53,7	50,9	47,0	43,8	40,3	34,8	29,3	24,5	20,6	18,5	16,8	15,0	13,8
40	66,8	63,7	59,3	55,8	51,8	45,6	39,3	33,7	29,1	26,5	24,4	22,2	20,7
50	79,5	76,2	71,4	67,5	63,2	56,3	49,3	42,9	37,7	34,8	32,4	29,7	28,0
60	92,0	88,4	83,3	79,1	74,4	67,0	59,3	52,3	46,5	43,2	40,5	37,5	35,5
70	104,2	100,4	95,0	90,5	85,5	77,6	69,3	61,7	55,3	51,7	48,8	45,4	43,3
80	116,3	112,3	106,6	101,9	96,6	88,1	79,3	71,1	64,3	60,4	57,2	53,5	51,2
90	128,3	124,1	118,1	113,1	107,6	98,6	89,3	80,6	73,3	69,1	65,6	61,8	59,2
100	140,2	135,8	129,6	124,3	118,5	109,1	99,3	90,1	82,4	77,9	74,2	70,1	67,3

D'après Catherine M. Thompson, *Table of percentage points of the χ^2 distribution*, Biometrika, vol. 32, 1941.

1.5 Estimation et Tests Statistiques

1.5.1 Estimation:

L'objectif de ce paragraphe est de tirer des informations sur une population $(\mathcal{P})$ connaisant ces informations sur des échantillons de cette population. Soit X une caractéristique à étudier dans une population de paramètres μ : moyenne de X sur $(\mathcal{P})$; σ : écart type de X sur $(\mathcal{P})$; π : proportion de X dans $(\mathcal{P})$. Il est quasi impossible de pouvoir calculer ces paramètres au moins parce que la taille d'une population est souvent assez grande et toute étude statistique devient trop couteuse pour être réalisée. Pour cela, on essaie d'estimer ces paramètres grace à des études faites sur des échantillons. D'une façon aléatoire on choisit un échantillon de taille n ,on calculee les paramètres de X sur cet échantillon, $\bar{x}$, s, et f désignent respectivement la moyenne, l'écart type et la proportion de X dans l'échantillon. Lors qu'on repète cette procédure un certain nombre de fois, on construit trois variables:

$\bar{X}$: $\{\text{échantillon de taille } n\}_i \mapsto \bar{x}_i$ la moyenne sur la i^{ième} échantillon
F : $\{\text{échantillon de taille } n\}_i \mapsto f_i$ la proportion sur la i^{ième} échantillon
S : $\{\text{échantillon de taille } n\}_i \mapsto s_i$ l'écart type sur la i^{ième} échantillon

le résumé suivant donne une estimation par intervalle de confiance I des paramètres μ, π et σ^2 avec un risque α de se tromper. C'est à dire la probabilité $p(\mu \in I) = 1 - \alpha$ de même $p(\pi \in I) = 1 - \alpha$ et $p(s^2 \in I) = 1 - \alpha$.

1. Estimation de μ : moyenne de X sur une population $(\mathcal{P})$:

   (a) Si $n \geq 30$ ou σ : écart type sur $(\mathcal{P})$ est connu ou X suit une loi normale alors:
   $$I = [\overline{x} - u_\alpha \frac{\sigma}{\sqrt{n}}, \overline{x} + u_\alpha \frac{\sigma}{\sqrt{n}}]$$
   et si σ est inconnu, $I = [\overline{x} - u_\alpha \frac{s}{\sqrt{n-1}}, \overline{x} + u_\alpha \frac{s}{\sqrt{n-1}}]$ où s :écart type sur l 'échantillon

   (b) Si $n < 30$ et σ est inconnu et X n 'est pas normale alors $I = [\overline{x} - t_{\alpha,\nu} \frac{s}{\sqrt{n-1}}, \overline{x} + t_{\alpha,\nu} \frac{s}{\sqrt{n-1}}]$ où $\nu = n - 1$

2. Estimation de π: proportion de X sur $(\mathcal{P})$: $I = [f - u_\alpha \sqrt{\frac{f(1-f)}{n}}, f + u_\alpha \sqrt{\frac{f(1-f)}{n}}]$

3. Estimation de σ^2 : variance de X sur une population $(\mathcal{P})$: $I =]\frac{ns^2}{\chi^2_{\alpha/2}}, \frac{ns^2}{\chi^2_{1-\alpha/2}}[$

Avec:

1. u_α est une valeur de la table de la loi normale centrée réduite telle que si $\mathcal{U} \longrightarrow N(0,1)$ alors $p(-u_\alpha < \mathcal{U} < u_a) = 1 - \alpha$

2. $t_{\alpha,\nu}$ est une valeur de la table de Student telle que si $T \longrightarrow st(\nu = n-1)$ alors $p(|T| > t_{\alpha,\nu}) = \alpha$

3. χ^2_p est une valeur de la table de la loi de chi deux telle que si $X \longrightarrow \chi^2_\nu$, $\nu = n - 1$, alors $p(X > \chi^2_p) = p$

Exemple 8

1. *La variable X désigne le niveau de cholestérol en g/l dans le sang. On suppose que X suit une loi normale dans une certaine population. Un test sanguin sur un échantillon de 10 personnes de cette population a donné le résultat suivant:*

   245 248 250 247 249 247 247 246 246 248

   (a) Déterminer l'intervalle de confiance de la moyenne μ au risque $\alpha = 5\%$ et puis donner l'intervalle de confiance de σ, l'écart type de X dans la population

(b) *On suppose que l'écart type de X sur la population est $\sigma = 1.5$. Que devient l'intervalle de confiance de μ au même risque*

2. *On voudrait connaitre la proportion des fumeurs dans une population, on tire un échantillon de 160 personnes et on trouve qu'il y a 40 fumeurs. Donner à un risque de 5% l'intervalle de confiance de π, la proportion des fumeurs dans la population*

Solution 1

1. (a) *Il est vrai que la taille de l'échantillon est petite et que l'écart type est inconnu mais comme on a supposé que X suit une loi normale alors l'intervalle de confiance de μ est*

$$I = [\bar{x} - u_\alpha \frac{s}{\sqrt{n-1}}, \bar{x} + u_\alpha \frac{s}{\sqrt{n-1}}]$$

avec $\bar{x} = \frac{1}{10}[245+2\times246+3\times247+2\times248+249+250] = 247.3\ g/l$ *et* $s^2 = \frac{1}{10}[245^2 + 2 \times 246^2 + \cdots + 250^2] - \bar{x}^2 = 2.01$ *et donc* $s = 1.417\ g/l$. *D'autre part, si $\mathcal{U}$ suit une loi centrée réduite telle que* $p(-u_\alpha < \mathcal{U} < u_\alpha) = 0.95$ *et donc* $F(u_\alpha) - F(-u_\alpha) = 0.95$ *cela implique que* $2F(u_\alpha) - 1 = 0.95$

$$F(u_\alpha) = 0.975 \text{ donne d'après la table que } u_\alpha = 1.96$$

on trouve $I = [246.37,\ \ 248.22]$
D'autre part, l'intervalle de confiance de σ^2 est

$$I =]\frac{ns^2}{\chi^2_{\alpha/2}}, \frac{ns^2}{\chi^2_{1-\alpha/2}}[$$

On a $p(X > \chi^2_{\alpha/2}) = \frac{\alpha}{2} = 0.025$ *et donc* $p(X < \chi^2_{\alpha/2}) = 0.975$, *on trouve d'après la table considérée avec* $\nu = n - 1 = 9$, *que* $\chi^2_{\alpha/2} = 19$. *De même* $p(X > \chi^2_{1-\alpha/2}) = 1 - \frac{\alpha}{2} = 0.975 \Rightarrow p(X < \chi^2_{1-\alpha/2}) = 0.025$, *cela implique que* $\chi^2_{1-\alpha/2} = 2.7$ *et par suite*

$$\sigma^2 \in [\frac{10 \times 2.01}{19}, \frac{10 \times 2.01}{2.7}]$$
$$\sigma^2 \in [1.05, 7.44]$$
$$\sigma \in [1.02,\ 2.73]$$

Remarquer que si l'intervalle de confiance est assez large, l'estimation ne sert absolument à rien et nous laisse incapable de prédir une estimation de σ. Pour obtenir une bonne estimation, il serait plus judicieux de faire un échantillonnage de taille plus importante

(b) *Lorsque* $\sigma = 1.5$ *l'intervalle de confiance de* μ *devient*

$$\begin{aligned} I &= [\bar{x} - u_\alpha \frac{\sigma}{\sqrt{n}}, \bar{x} + u_\alpha \frac{\sigma}{\sqrt{n}}] \\ &= [246.37, \quad 248.229] \end{aligned}$$

2. *On a* $I = [f - u_\alpha \sqrt{\frac{f(1-f)}{n}}, f + u_\alpha \sqrt{\frac{f(1-f)}{n}}]$ *où* $f = \frac{40}{160} = 0.25$. *On trouve*

$$I = [0.25 - 1.96\sqrt{\frac{0.25 \times 0.75}{160}}, \ \ 0.25 + 1.96\sqrt{\frac{0.25 \times 0.75}{160}}] = [0.183, \ \ 0.317]$$

Ainsi on tire que le pourcentage des fumeurs dans la population est entre 18.3% *et* 31.7%

1.5.2 Les Tests Statistiques

Tests de comparaison: On se met dans l'une des deux situations suivantes:

1-Dans une population $(\mathcal{P})$ on a construit les variables $\bar{X}, F$ et S en tirant plusieurs échantillons de taille n mais on pourrait construire des variables similaires $\bar{X}', F'$ et S' en tirant de la population $(\mathcal{P})$ des échantillons de taille n' et la question qui s'impose: à quel point ces variables donnent des estimations proches pour μ, π et σ les paramètres de la population $(\mathcal{P})$

2-On considère deux populations $(\mathcal{P})$ et $(\mathcal{P}')$ ayant pour paramètres respectifs μ, π, σ et μ', π' et σ' et on désire les comparer grâce à un échantillonnage de taille n et n'appliqué respectivement sur les deux populations.

On exprime ceci par:

1-test de comparaison de la moyenne: On test l'hypothèse $H_0 =$ "$\bar{x} = \bar{x}'$" contre l'hypothèse $H_1 =$ "$\bar{x} \neq \bar{x}'$" avec un risque α de se tromper

2-test de comparaison de la proportion: On test l'hypothèse $H_0 =$ "$f = f'$" contre l'hypothèse $H_1 =$ "$f \neq f'$" avec un risque α de se tromper

3-test de comparaison de la variance: On test l'hypothèse $H_0 =$ "$s^2 = s'^2$" contre l'hypothèse $H_1 =$ "$s^2 \neq s'^2$" avec un risque α de se tromper

On donne un résumé qui explique le déroulement de ce test:

1. Test de H_0 : "$\overline{x}_1 = \overline{x}_2$" contre H_1 : "$\overline{x}_1 \neq \overline{x}_2$" :Soit $z = \dfrac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\frac{s_1^2}{n_{1-1}} + \frac{s_2^2}{n_{2-1}}}}$

   (a) Si $(\mathcal{P})$ est normalement distribuée ou $n_1, n_2 \geq 30$ alors:

      i. Si $z \leq u_\alpha$ on accepte H_0

      ii. Si $z > u_\alpha$ on rejette H_0

   (b) Si $(\mathcal{P})$ n'est pas normalement distribuée et $n_1$ou $n_2 < 30$ alors

      i. Si $z \leq t_{\alpha,\nu}$ on accepte H_0. Noter que $\nu = n_1 + n_2 - 2$

ii. Si $z > t_{\alpha,\nu}$ on rejette H_0

2. Test de H_0 : "$f_1 = f_2$" contre H_1 : "$f_1 \neq f_2$"

   (a) S'il s'agit de deux échantillons d'une même population $(\mathcal{P})$ avec une proportion π de X .
   Si π est connue, on pose:

$$f = \frac{|f_1 - f_2|}{\sqrt{\pi(1-\pi)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

   (b) Si π est inconnue, on pose $\widehat{p} = \dfrac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$ et soit

$$f = \frac{|f_1 - f_2|}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

   Dans les deux cas:

   i. Si $f \leq u_{\alpha,\nu}$ on accepte H_0
   ii. Si $f > u_{\alpha,\nu}$ on rejette H_0

3. Test de H_0 : "$\sigma_1^2 = \sigma_2^2$" contre H_1 : "$\sigma_1^2 \neq \sigma_2^2$". On pose s_1^2 la plus grande variance parmi les deux échantillons. Soit

$$R = \frac{s_1^2}{s_2^2} \quad \text{"rapport de deux variances"} \quad \nu_1 = n_1 - 1, \nu_2 = n_2 - 1$$

   (a) Si $R \leq F_{\alpha/2,\nu_1,\nu_2}$ on accepte H_0
   (b) Si $R > F_{\alpha/2,\nu_1,\nu_2}$ on rejette H_0

   où $F_{\alpha/2,\nu_1,\nu_2}$ et $F_{\alpha/2,\nu_1,\nu_2}$ sont deux valeurs de la table de Fisher qu'on ne donne pas dans ce cours.

Tests de conformité On voudrait comparer les paramètres μ, π et σ d'une population $(\mathcal{P})$ à des valeurs fixes μ_0, p_0 et σ_0.

1. Test de H_0 : "$\mu = \mu_0$" contre H_1 : "$\mu \neq \mu_0$"

   (a) Si (P) est normalement distribuée ou $n \geq 30$ ou σ est connu, on pose:

$$I = [\mu_0 - u_\alpha \frac{\sigma}{\sqrt{n}}, \mu_0 + u_\alpha \frac{\sigma}{\sqrt{n}}] \text{ et si on ne connait pas } \sigma \text{ on prend}$$

$$I = [\mu_0 - u_\alpha \frac{s}{\sqrt{n-1}}, \mu_0 + u_\alpha \frac{s}{\sqrt{n-1}}]$$

(b) Si (P) n'est pas normalement distribué, $n < 30$ et σ est inconnu, on pose:

$$I = [\mu_0 - t_{\alpha,\nu}\frac{s}{\sqrt{n-1}}, \mu_0 + t_{\alpha,\nu}\frac{s}{\sqrt{n-1}}] \text{ où } \nu = n-1$$

Dans les deux cas:

i. Si $\overline{x} \in I$ on accepte H_0

ii. Si $\overline{x} \notin I$ on rejette H_0

2. Test de H_0 : "$\pi = p_0$" contre H_1 : "$\pi \neq p_0$", on pose $q_0 = 1 - p_0$ et soit

$$I = [p_0 - u_\alpha\sqrt{\frac{p_0 q_0}{n}}, p_0 + u_\alpha\sqrt{\frac{p_0 q_0}{n}}]$$

(a) Si $f \in I$ on accepte H_0

(b) Si $f \notin I$ on rejette H_0.

3. Test de H_0 : "$\sigma^2 = \sigma_0^2$" contre H_1 : "$\sigma^2 \neq \sigma_0^2$". On pose $\varepsilon = n\frac{s^2}{\sigma_0^2}$

(a) Si $\varepsilon \in]\chi^2_{1-\alpha/2,n-1}; \chi^2_{\alpha/2,n-1}[$ on accepte H_0

(b) Si $\varepsilon \notin]\chi^2_{1-\alpha/2,n-1}; \chi^2_{\alpha/2,n-1}[$ on rejette H_0

Exemple 9

1. Un fabriquant livre des paquets contenant en principe $170g$ de fromage. Dans le but de vérifier que le $170g$ indiqué sur chaque boite correspond au poids effectif du fromage, on a prélevé au hasard 200 boites qu'on a pesées une à une :

poids en g	166.5	168	168.5	169	169.5	170	170.5	171	171.5	172
nombre de boites	1	16	12	21	36	38	34	18	14	10

Peut-on juger au risque de 5%, que le fabriquant ne ment pas?

2. On a mesuré la capacité vitale de 100 sujets sains, on a trouvé que la moyenne sur cet échantillon est $\bar{x} = 4.483$ et l'écart type est $s_1 = 0.538$. Sur un échantillon de 50 individus, exposés pendant plus de 5 ans à des vapeurs nocives, on a trouvé une capacité vitale moyenne $\bar{x}_2 = 3.95$ et un écart type $s_2 = 0.9$. Peut - on affirmer, au risque de 5% que les individus exposés aux vapeurs ont une capacité vitale significativement différente de celle des sujets non exposés?

Solution 2

1. *Il s'agit d'étudier le test de conformité* $H_0 = "\mu = \mu_0 = 170"$ *contre le test* $H_1 = "\mu \neq 170"$. *Soit*

$$I = [\mu_0 - u_\alpha \frac{s}{\sqrt{n-1}}, \mu_0 + u_\alpha \frac{s}{\sqrt{n-1}}] \quad (\textit{puisque } n = 200 > 30)$$

on a: $\bar{x} = 169.9175$ *g et* $s^2 = 1.149$ *et* $s = 1.072g$. *Ainsi* $I = [169.86,\ 170.138]$ *et comme* $\bar{x} = 169.9175 \in I$ *alors on conclut que le fabriquant ne ment pas*

2. *Soient* μ_1 *et* μ_2 *les capacités vitales moyennes chez la première et la seconde population respectivement. Il s'agit d'étudier le test de comparaison* $H_0 = "\bar{x}_1 = \bar{x}_2"$ *contre* $H_1 = "\bar{x}_1 \neq \bar{x}_2"$. *Soit*

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}} = 3.82 > 1.96$$

alors on rejette H_0 *et on conclut que les deux populations ont une capacité vitale significativement différente.*

Tests d'ajustement de χ^2 L'objectif de ce test est de théoriser une expérience statistique. La situation est la suivante, une expérience statistique étudiant un caractéristique X donne des effectifs qu'on appelle effectifs observés et notés n_{io}. Les n_{io} peuvent être le nombre de répétitions de la valeur x_i dans le cas d'une série statistique non regroupée et peuvent être le nombre de valeurs prises par X et qui sont dans la classe $[a_i,\ b_i[$ dans le cas d'une série statistique regroupée.

Problème: Peut - on ajuster, avec un risque α de se tromper, cette expérience à une des lois de probabilité ? C'est à dire, peut - on remplacer les n_{io} par des effectifs théoriques n_{ic} venant d'une loi de probabilité? Dans le cas d'une série statistique non regroupée, on se demande si on peut supposer que X suit une des lois discrètes et dans le cas d'une série statistique regroupée on cherche à juger si l'on peut dire que X suit une des lois continues. On pose alors $n_{ic} = n.p_i$ où $p_i = p(X = x_i)$ dans la cas discret et $p_i = p(x \in [a_i, b_i[)$ dans le cas continu. Ce test est dit test d'ajustement. Soit:

$$\chi^2 = \sum_{i=1}^{c} \frac{(n_{i,o} - np_i)^2}{np_i}$$

L'ajustement se produit quand les n_{io} sont proches des n_{ic} c'est à dire quand χ^2 est petit. On exprime ce test par $H_0 = "\chi^2 = 0"$ contre $H_1 = "\chi^2 \neq 0"$

1. On doit avoir $n_{i,o} \geq 5$, et dans le cas contraire on regroupe plusieurs classes. Noter que p_i est le probabilité de réalisation de la $i^{ième}$ classe, c est le nombre des classes, $\nu = c-1-k$ où k est le nombre des paramètres estimés de la loi théorique

2. Si $\chi^2 < \chi^2_{\alpha,\nu}$ on accepte H_0

3. Si $\chi \geq \chi^2_{\alpha,\nu}$ on rejette H_0

Exemple 10 *On a étudié la distribution des circonférences d'arbres mesurés à une certaine distance du sol. On a observé les résultats suivants:*

Circonférence en cm	$100-110$	$110-120$	$120-130$	$130-140$	$140-150$	$150-160$
effectifs	5	26	32	32	11	5

On suppose que la représentation est ajustable par une loi normale de paramètre qu'on estime à $\mu = 130$ *et* $\sigma = 12$. *Vérifier à l'aide du test de* χ^2 *au seuil de* 5% *la validité de cette supposition*

En effet, $n_i = n.p_i$. On a $p_1 = p(100 < X < 110) = p(\frac{100-130}{12} < Z < \frac{110-130}{12}) = 0.0413$ et $n_1 = 100 \times 0.0413 = 4.13$. De même pour les autres, on trouve

n_{io}	n_{ic}
5	4.13
16	15.58
32	29.67
31	29.67
11	15.58
5	4.13
$n_o:\ 100$	$n_c:\ 98.76$

il y a $100-98.76 = 1.24$ qui ont une circonférence plus petite que 100 ou plus grande que 160.Par symétrie il ya $4.13 + \frac{1.24}{2} = 4.75$ qui ont une circonférence plus petite que 110 et $4.13 + \frac{1.24}{2} = 4.75$ qui ont une circonférence plus grande que 150. La nouvelle table est:

n_{io}	n_{ic}
5	4.75
16	15.58
32	29.67
31	29.67
11	15.58
5	4.75
$n_o:\ 100$	$n_c:\ 100$

on remarque que la première classe et la dernière classe ont chacune des effectifs < 5 on les regroupes respectivement avec la classe voisine:

$a_i - b_i$	n_{io}	n_{ic}
$100-120$	21	20.33
$120-130$	32	29.67
$130-140$	31	29.67
$140-160$	16	20.33

Ainsi $\chi^2 = \sum_{i=1}^{4} \frac{(n_{io} - n_{ic})^2}{n_{ic}} = 1.1869$. On a $\nu = c - 1 - k = 4 - 1 - 2 = 1$ et $\chi_t^2 = \chi_{0.05,\ 1}^2 = 3.84$ d'après la table. On a $\chi^2 < \chi_t^2$ donc on accepte H_0 et l'expérience est ajustable à cette loi normale

Remarque 4 *Il y a des statisticiens qui ne font le regroupement des classes que si les $n_{io} < 5$ sans regarder les n_{ic}.*

Test d'indépendance de χ^2 Si on a une série statistique double (X, Y) et on désire étudier l'indépendance de X et Y c'est à dire si on connait une valeur de l'une des variables, on ne pourrait rien prédire quant à la valeur de l'autre.

Test d 'indépendance de deux variables statistiques X et Y : Soit H_0 : "X indépendante de Y". On pose:

$$\chi^2 = \sum_{i=1}^{c} \frac{(n_{oij} - n_{cij})^2}{n_{cij}}$$

avec: n_{oij} :effectif observé, $n_{cij} = \frac{n_{i\ .} \times n_{.\ j}}{n}$ effectif théorique, $n_{i\ .} = \sum_{j=1}^{p} n_{ij}$ et $n_{.\ j} = \sum_{i=1}^{m} n_{ij}$; p :nombre de classes de Y et m : nombre de classes de X et soit $\nu = (m-1)(p-1)$

1. Si $\chi^2 < \chi_{\alpha,\nu}^2$ on accepte H_0
2. Si $\chi \geq \chi_{\alpha,\nu}^2$ on rejette H_0

References

[S.Lipschutz: Série Schaum Probabilité]
[D.Wayne: Statistics]
[J.Saab: Polycopes]