Statistiques inférentielle 2

Table des matières

CHAPITRE 1

Les tests non paramétriques

Ce cours a pour objectif la présentation des tests non paramétriques les plus couramment utilisés. Il se situe dans le cadre de l'inférence statistique et des tests d'hypothèse usuels : on cherche à apprécier des caractéristiques d'un population à partir d'un échantillon issu de cette population.

Un test non-paramétrique présente quelques avantages :

1. Son application est relativement facile et rapide,
2. S'applique à des échantillons de petites tailles,
3. S'applique à des caractères qualitatifs, à des grandeurs de mesure, à des rangs de classement, etc.

On distinguera principalement les deux familles suivantes :

a). Test du Khi-deux de Pearson :

(1) Test d'ajustement ou d'adéquation entre deux distributions.
(2) Test d'indépendance dans un tableau de contingence.
(3) Test d'homogénéité de plusieurs populations.

b) Tests appliqués aux rangs et aux signes

(1) Test de la somme des rangs (Wilcoxon et Mann-Withney)
(2) Test de signes
(3) Test de la somme des rangs des différences positives (Wilcoxon)
(4) Test d'indépendance de rangs de Spearman

1. Test d'ajustement de deux distributions : "test du khi-deux"

Introduction

Le test de Pearson, appelé aussi le test du khi-deux est un outil statistique qui permet de vérifier la concordance entre une distribution expérimentale et une distribution théorique.

On cherche donc à d´eterminer si un modèle théorique est susceptible de repréé-senter adéquatement le comportement probabiliste de la variable observée, comportement fondé sur les fréquences des résultats obtenus sur l'´echantillon.

Comment procéder ?

Répartitions expérimentales

On répartit les observations suivant k classes (si le caractère est continu) ou k valeurs (si le caractère est discret). On dispose alors des effectifs des k classes : $O_1, O_2, ..., O_k$. On a bien sûr la relation

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

où N est le nombre total d'observations effectuées.

Répartitions théoriques

En admettant comme plausible une distribution théorique particulière, on peut construire une répartition idéale des observations de l'´echantillon de taille N en ayant recours aux probabilités tablées (ou calculées) du modèle théorique : $p_1, p_2, ..., p_k$. On obtient alors les effectifs théoriques T_i en 2 érivant $T_i = Np_i$. On dispose automatiquement de la relation

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

Définition de l'écart entre les deux distributions

Pour évaluer l'écart entre les effectifs observés ni et les effectifs théoriques T_i, on utilise la somme des écarts normalisés entre les deux distributions, à savoir

$$\chi^2 = \frac{(O_1 - T_1)^2}{T_1} + \frac{(O_2 - T_2)^2}{T_2} + ..., \frac{(O_k - T_k)^2}{T_k}$$

La statitistique χ^2 représente une sorte de "distance" globale entre les effectifs observées et les effectifs attendus. plus la distribution étudiée différer de la distribution théorique.

Mais quel est le nombre de degrés de liberté de cette variable du khi-deux ?

1. Si la distribution théorique est entièrement spécifiée, c'est-'a-dire si on cherche à déterminer si la distribution observée suit une loi dont les paramètres sont connus avant même de choisir l'échantillon, on a $k-1$ degrés de liberté (k carrés indépendants moins une relation entre les variables).

2. S'il faut d'abord estimer r paramétres de la loi à partir des observations de l'échantillon (par exemple on cherche si la distribution est normale mais on ne connait d'avance ni sa moyenne ni son écart-type), il n'y a plus que $k-1-r$ degrés de liberté.

Dans le cas général, on dira que la loi du khi-deux suivie par l'écart entre les deux distributions a $k-1-r$ degrés de liberté lorsqu'on a estimé r paramétres de la loi théorique à partir des observations de l'´echantillon (avec la possibilité pour r de valoir 0).

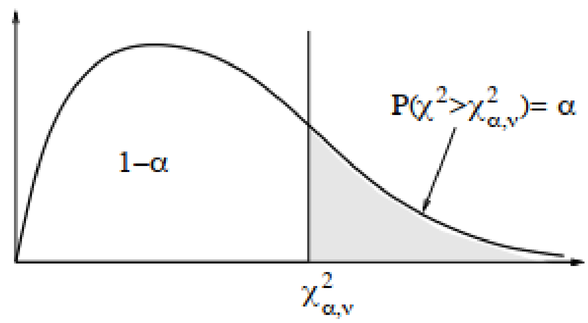
REMARQUE 1. *Le nombre d'observations par classes ne doit pas être faible, Np_i doit être supérieur à 5, $\forall i = 1, 2, ..., k$. Dans le cas contraire, on regroupe deux ou plusieurs classes adjacentes de façon à réaliser cette condition. On tient compte de ce regroupement pour le nombre de degrés de liberté.*

Le test d'ajustement de χ^2

Il nous faut maintenant décider, à l'aide de cet indicateur qu'est le χ^2, si les écarts entre les effectifs théoriques et ceux qui résultent des observations sont significatifs d'une différence de distribution ou si ils sont dus aux fluctuations d'échantillonnage. Nous procéderons comme d'habitude en quatre étapes.

1ére étape : **Formulation des hypothèses**.

On va donc tester l'hypothèse $\mathbf{H}_0$ contre l'hypothèse $\mathbf{H}_1$:



$$\begin{cases} \mathbf{H}_0 & \text{Les observations suivent la distribution théorique spécifiée} \\ \mathbf{H}_1 & \text{Les observations ne suivent la distribution théorique spécifiée} \end{cases}$$

2éme étape : Détermination de la fonction χ^2

On utilise la variable aléatoire

$$\chi^2 = \frac{(O_1 - T_1)^2}{T_1} + \frac{(O_2 - T_2)^2}{T_2} + ..., \frac{(O_k - T_k)^2}{T_k}$$

3éme étape : Détermination des valeurs critiques de χ^2 délimitant les zones d'acceptation et de rejet.

On impose à la zone d'acceptation de $\mathbf{H}_0$ concernant la valeur du χ^2 d'être un intervalle dont 0 est la borne inferieure (car un χ^2 est toujours positif).

Il nous faut donc déterminer dans la table la valeur maximale $\chi^2_{\alpha,\nu}$ de l' ́ecart entre les deux distributions imputable aux variations d'échantillonnage au seuil de signification α, c'est-à dire vérifiant $P\left(\chi^2 > \chi^2_{\alpha,\nu}\right) = \alpha$. $\chi^2_{\alpha,\nu}$ représente donc la valeur critique pour un test sur la concordance entre deux distributions et le test sera toujours unilatéral à droite.

4éme étape : **Calcul de la valeur de χ^2 prise dans l'échantillon et conclusion du test.**

On calcule la valeur χ^2_0 prise par χ^2 dans l'échantillon.

- Si la valeur χ^2_0 se trouve dans la zone de rejet, on dira que l'écart observé entre les deux distributions est **statistiquement significatif** au seuil α. Cet écart est anormalement élevé et ne permet pas d'accepter $\mathbf{H}_0$. On rejette $\mathbf{H}_0$.
- Si la valeur χ^2_0 se trouve dans la zone d'acceptation, on dira que l'écart-réduit observé n'est pas significatif au seuil α. Cet écart est imputable aux fluctuations d'échantillonnage. On accepte $\mathbf{H}_0$

EXEMPLE 1. *Un pisciculteur possède un bassin qui contient trois variétés de truites : communes, saumonées et arc-en-ciel. Il voudrait savoir s'il peut considérer que son bassin contient autant de truites de chaque variété. Pour cela, il effectue, au hasard* 399 *prélèvements avec remise et obtient les résultats suivants :*

Variétés	*saumonée*	*commune*	*arc-en-ciel*
Effectifs	145	118	136

SOLUTION 1. *On cherche à savoir s'il y a équirépartition des truites entre chaque espèce c'est-à-dire on suppose de* $\mathcal{L}_0$ *est la loi uniforme, une probabilité de* $1/3$ *pour chaque classe (soit* $C_i = \frac{399}{13} = 133$*)*

C'est-à dire on souhaite tester l'ajustement de cette loi à une loi connue uniforme

Variétés	*commune*	*saumonée*	*arc-en-ciel*
Effectifs O_i	145	118	136
Effectifs T_i	133	133	133

On obtient

$$\begin{aligned}\chi^2_{calcul\acute{e}e} &= \frac{(O_1 - T_1)^2}{T_1} + \frac{(O_2 - T_2)^2}{T_2} + \frac{(O_3 - T_3)^2}{T_3} \\ &= \frac{(145-133)^2}{133} + \frac{(118-133)^2}{133} + \frac{(136-133)^2}{133} \approx 2.84\end{aligned}$$

La valeur théorique lue dans la table du $\chi^2_{\alpha,v}$ *au risque de* 5% *avec* $\nu = 3-1-0 = 2$ *degrés de liberté vaut* 5.99*.*

On ne peut rejeter l'hypothèse que son bassin contient autant de truites de chaque variété car $\chi^2_{calcul\acute{e}e} < \chi^2_{\alpha,v}$

EXEMPLE 2. *On veut tester si un dé n'est pas truqué au risque* $\alpha = 0,05$*. Pour cela on lance le dé* 60 *fois et on obtient les résultats suivants*

face	1	2	3	4	5	6
O_i	15	7	4	11	6	17
T_i	10	10	10	10	10	10

On a fait figurer dans le tableau la valeur espérée T_i *du nombre d'apparitions de i dans l'l'hypothèse où le dé n'est pas truqué, ceci afin de faciliter le calcul de la* χ^2_{cal} *qui est donc ici égale à*

$$\begin{aligned}\chi^2_{cal} &= \sum_{i=1}^{6} \frac{(O_i - T_i)^2}{T_i} = \frac{(15-10)^2}{10} + \frac{(7-10)^2}{10} + \frac{(4-10)^2}{10} \\ &\quad + \frac{(11-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(17-10)^2}{10} \\ &= 13.6\end{aligned}$$

Sous l'hypothèse $\mathbf{H_0}$ *:"*$p_1 = \cdots = p_6 = \frac{1}{6}$*", la variable aléatoire* χ^2_{cal} *a donc pris la valeur* 13,6*. Or le seuil de rejet lu dans la table de la loi du* $\chi^2_{\alpha,v}$ *est* $\chi^2_{0.05,5} = 11,07$*. La valeur observée dépassant cette valeur, on est amené à rejeter l'hypothèse* H_0 *au risque* $\alpha = 0,05$*. On notera qu'au risque* $\alpha = 0,025$*, on rejette aussi* H_0*. Mais au risque* $\alpha = 0,01$*, on ne peut plus rejeter l'hypothèse* H_0 *malgré la mauvaise impression donnée par les résultats. Si on persiste à vouloir le risque* 0,01*, il est plus raisonnable de recommencer l'expérience avec un échantillon de taille beaucoup plus grande.*

EXEMPLE 3 (loi uniforme). *Une statistique relative aux résultats du concours d'entrée à une grande école fait ressortir les répartitions des candidats et des admis* *selon la profession des parents.*

Profession des candidats	*Nombre de candidats*	*Nombre d'admis*
Fonctionnaires et assimilés	*2244*	*180*
Commerce, industrie	*988*	*89*
Professions libérales	*575*	*48*
Propriétaires rentiers	*423*	*37*
Propri ´etaires agricoles	*287*	*13*
Artisans, petits commerçants	*210*	*18*
Banque, assurance	*209*	*17*
Total	*4936*	*402*

Tester l'hypothèse (risque $\alpha = 0,05$) selon laquelle la profession des parents n'a pas d'influence sur l'accès à cette grande école.

Il s'agit du test d'ajustement d'une distribution théorique, on pose les hypothèses

H_0 : "la profession des parents n'a pas d'influence sur l'accès à cette grande école", la proportion des admis est constante pour toutes les professions soit $p = \frac{402}{4936} \simeq 0,0814$

H_1 :" la profession des parents influe sur l'accès à cette grande école"

Sous H_0, le nombre d'admis pour la i-ième profession est $N_i p$.

i	N_i	n_i *effectif observé*	$N_i p$ *effectif théorique*	$\left(\frac{n_i - N_i p}{N_i p}\right)^2$
1	2244	180	$\frac{2244\times402}{4936} \simeq 182,76$	0.0416
2	988	89	$\frac{988\times\ 402}{4936} \simeq 80,47$	$0,9042$
3	575	48	$\frac{575\times402}{4936} \simeq 46,830$	$0,0293$
4	423	37	$\frac{423\times402}{4936} \simeq 34,450$	$0,1887$
5	287	13	$\frac{287\times402}{4936} \simeq 23,374$	$4,6050$
6	210	18	$\frac{210\times402}{4936} \simeq 17,10$	$0,0471$
7	209	17	$\frac{209\times\ 402}{4936} \simeq 17,02$	$\simeq 0$
Total	4936	402	402	$5,8181$

Le χ^2 calculé vaut $5,8181$. Le nombre de degrés de liberté est $7 - 1 = 6$. La table fournit $\chi^2_{6;0,95} = 12,59$ donc χ^2calculé $< \chi^2_{6;0,95}$.

On ne rejette pas H_0, ce qui signifie que la profession des parents n'a pas d'influence sur l'accès à cette grande école.

EXEMPLE 4 (loi normale). *On suppose que le rendement X (quintaux par hectares d'une parcelle de blé) suit une loi normale $\mathcal{N}(m, \sigma)$. L'observation du rendement de 1000 parcelles a donné les résultats suivants :*

Rendement	*Nombre de parcelles*
$[0, 10[$	5
$[10, 20[$	6
$[20, 30[$	40
$[30, 40[$	168
$[40, 50[$	288
$[50, 60[$	277
$[60, 70[$	165
$[70, 80[$	49
$[80, 90[$	2
Total	1000

Afin de mettre en place un test d'ajustement, d´eterminons dans un premier temps la moyenne arithmétique et l'écarttype de la distribution observée :

$E(X) = \mu = \dfrac{1}{N} \sum_i n_i x_i = 49.76$

$V(X) = \sigma^2 = \dfrac{1}{N} \sum_i n_i x_i^2 - [E(X)]^2 = 164.5424$ *donc* $\sigma \simeq 12,827$

Problème : Tester l'hypothèse (risque $\alpha = 0,05$) selon laquelle l'ajustement de la distribution observée à une loi normale $\mathcal{N}(50, 13)$ est acceptable.

Les hypothèses du test du χ^2 sont les suivantes :

- H_0 : *"$X \rightsquigarrow \mathcal{N}(50, 13)$"*
- H_1 : *"X ne suit pas $\mathcal{N}(50, 13)$" On désigne par $[a_0, a_1[,\ [a_1, a_2[, ..., [a_8, a_9[$ les classes et par $x_1, x_2, ..., x_9$ les centres de ces classes. Sous H_0, $X \rightsquigarrow \mathcal{N}(50, 13)$ et $Z = \dfrac{X - 50}{13} \rightsquigarrow \mathcal{N}(0, 1)$, donc $p_i = p(X \in [a_{i-1}; a_i[) = \Phi(zi) - \Phi(z_{i-1})$*

avec $zi = \dfrac{a_i - 50}{13}$ et $z_{i-1} = \dfrac{a_{i-1} - 50}{13}$. L'effectif théorique de la iéme classe est $1000p_i$ et $\sum_i \dfrac{(n_i - Np_i)^2}{Np_i} \rightsquigarrow \chi^2_{\alpha,\nu}$ .On a le tableau suivant

Classe	n_i	z_i	$\Phi(z_i)$	p_i	Np_i	Np_i *corrigée*	n_i *corrigée*	$\sum_i \frac{(n_i - Np_i)^2}{Np_i}$
$[0, 10[$	5	-3.0769	0.0010	0.0009	0.9	10.4	11	0.0346
$[10, 20[$	6	-2.3077	0.0105	0.0095	9.5			
$[20, 30[$	40	-1.5385	0.0620	0.0515	51.5	51.5	40	2.568
$[30, 40[$	168	-0.7692	0.2209	0.1589	158.9	158.9	168	0.5211
$[40, 50[$	288	0	0.5	0.2791	279.1	279.1	288	0.283
$[50, 60[$	277	0.7692	0.7791	0.2791	279.1	279.1	277	0.0158
$[60, 70[$	165	1.5385	0.9380	0.1589	158.9	158.9	165	0.234
$[70, 80[$	49	2.3077	0.9895	0.0515	51.5	51.5	49	0.1214
$[80, 90[$	2	3.0769	0.9990	0.0095	9.5	9.5	2	5.9211
$Total$	1000			1	1000	1000	1000	9.7

On effectue le regroupement des deux premières classes car $Np_i < 5$. *Le* χ^2 *calculé vaut* 9.7. *Après le regroupement, il reste* 8 *classes, les deux paramétres de la loi normale sont donnés, le nombre de degrés de liberté est* $\nu = 8 - 1 = 7$. *A l'aide de la table, on obtient '* $\chi^2_{7,0,95} = 14.07$. *Ainsi,* $\chi^2_{cal} < \chi^2_{7,0,95}$.

On ne rejette pas $\mathbf{H_0}$, *l'ajustement de la distribution observée à une loi normale* $\mathcal{N}(50, 13)$ *est acceptable ne spécifie pas complètement la loi qu'on considère.*

Exemple 5 (loi de Poisson). *Supposons qu'on s'intéresse au nombre de voitures se présentant par minute à un poste de péage sur une autoroute. On peut se demander si cette variable aléatoire peut être modélisée par une* ***loi de Poisson*** $(\mathcal{P}(\lambda))$. *On souhaite donc tester l'hypothèse fondamentale* $\mathbf{H_0}$: *"*$X \rightsquigarrow \mathcal{P}(\lambda)$*" contre l'hypothèse alternative* $\mathbf{H_1}$: *"*X *ne suit pas* $P(\lambda)$*". On ne précise pas la valeur du paramètre* λ. *On peut toutefois l'estimer à partir des données disponibles mais dans ce cas,* $r = 1$. *Le nombre de degrés sera alors* $\nu = k - r - 1 = k - 2$.

On effectue 200 *comptages au péage*

x_i	0	1	2	3	4	5	6	7	8	≥ 9	*Total*
n_i	6	15	40	42	37	30	10	12	8	0	200
$n_i x_i$	0	15	80	126	148	150	60	84	64	0	727

où x_i *et* n_i *désignent respectivement le nombre de voitures par minute et l'effectif correspondant lors de l'observation* $n^o i$ *(par exemple,* $x_1 = 0$ *et* $n_1 = 6$*) c'est-'a-dire que lors de* 6 *observations, il y a 0 voiture). La moyenne arithmétique de cette distribution observée est*

$$\frac{\sum n_i x_i}{\sum n_i} = \frac{727}{200} = 3.635 \simeq 3.5$$

Problème : Tester l'hypothèse (au risque $\alpha = 0,01$*) selon laquelle* X *suit une loi de Poisson de paramètre* $3,5$.

On pose

- $\mathbf{H_0}$ *: "*$X \rightsquigarrow \mathcal{P}(3,5)$*"*
- $\mathbf{H_1}$ *: "*X *ne suit pas* $\mathcal{P}(3,5)$*"*

Sous $\mathbf{H_0}$, $pi = p(X = i) = e^{-3,5}\dfrac{(3,5)^i}{i!}$, *on a donc le tableau de valeurs suivant*

x_i	n_i	p_i	Np_i	Np_i *corrigée*	n_i *corrigée*	$\sum_i \frac{(n_i - Np_i)^2}{Np_i}$
0	6	$0,0302$	$6,04$	$6,04$	6	$0,00026$
1	15	$0,1057$	$21,14$	$21,14$	15	$1,78333$
2	40	$0,1850$	37	37	40	$0,24324$
3	42	$0,2158$	43.16	43.16	42	$0,03118$
4	37	0.1888	$37,76$	$37,76$	37	$0,01530$
5	30	$0,1322$	$26,44$	$26,44$	30	$0,47933$
6	10	$0,0771$	$15,42$	$15,42$	10	$1,90508$
7	12	$0,0385$	$7,7$	$7,7$	12	$2,40130$
8	8	$0,0169$	$3,38$	$5,34$	8	$1,32502$
≥ 9	0	$0,0098$	$1,96$			
Total	200	1	200	200	200	$8,18404$

On a effectué le regroupement des deux dernières classes car l'effectif théorique y est inférieur à 5. *Après ce regroupement, le nombre de classes est de* 9. *Le nombre de degrés de liberté est* $9 - 1 - 1 = 7$. *Au risque* $\alpha = 0,01$, $\chi^2_{7,0.99} = 18,48$ *donc* $\chi^2_{cal} = 8,18404 < \chi^2_{7,0.99}$ *On ne rejette pas l'hypothèse* $\mathbf{H_0}$ *et* $X \rightsquigarrow \mathcal{P}(\lambda = 3,5)$ *au risque* $\alpha = 0,01$.

EXEMPLE 6 (loi binomiale). *Supposons qu'on ait recueilli* 300 *bôıtes contenant chacune trois ampoules. Dans chaque bôıte, on compte le nombre d'ampoules défectueuses. On obtient les résultats suivants*

Nombred'ampoules défectueuses x_i	*Nombre de bôıtes observées* n_i
0	190
1	95
2	10
3	5
Total	300

Pour chaque ampoule testée, on peut observer deux états différents : l'ampoule est défectueuse ou non. Le nombre X *d'ampoules défectueuses par bôıte suit une loi binomiale de paramètres* $n = 3$ *et* p. *Déterminons* p. *Dans la distribution observée, le nombre d'ampoules défectueuses est de* $0 \times 190 + 1 \times 95 + 2 \times 10 + 3 \times 5 = 130$ *soit* 130 *ampoules défectueuses sur un total de* 900 *ampoules. La proportion d'ampoules défectueuses est alors de* $\dfrac{130}{900} \backsimeq 0,144$

Prenons $p = 0,15$

Problème : Tester l'hypothèse (au risque $\alpha = 0,01$*) selon laquelle le nombre d'ampoules défectueuses par bôıte suit une loi binomiale de paramètres* $n = 3$ *et* $p = 0,15$.

On considère donc les hypothèses suivantes :

– $\mathbf{H_0}$: $X \rightsquigarrow$ "$\mathcal{B}(3, 0.15)$"

– $\mathbf{H_1}$: X *ne suit pas cette loi binomiale"*

et on détermine ensuite les probabilités théoriques $(X \rightsquigarrow \mathcal{B})$ *:*

$p_0 = P\{X = 0\} = (0,85)^3 \backsimeq 0,6141$

$p_1 = P\{X = 1\} = C_3^1 (0,15)(0,85)^2 \backsimeq 0,3251$

$p_2 = P\{X = 2\} = C_3^2 (0,15)^2 (0,85) \backsimeq 0,0574$

$p_3 = P\{X = 3\} = C_3^3 (0,15)^3 \backsimeq 0,0034$

On a le tableau (provisoire) suivant :

x	*effectif observé* n_i	p_i	*effectif théorique*
0	190	$0,6141$	$184,23$
1	95	$0,3251$	$97,53$
2	10	$0,0574$	$17,22$
3	5	$0,0034$	$1,02$
T	300	1	300

L'effectif théorique de la quatrième classe est faible, en effet $1,02 < 5$. *On effectue un regroupement de classes, les classes* 2 *et* 3.

x_i	n_i	Np_i	$\sum_i \frac{(n_i - Np_i)^2}{Np_i}$
0	190	$184,23$	$0,18071$
1	95	$97,53$	$0,06563$
2 *ou* 3	15	$18,24$	$0,57553$
Total	300	300	$0,82187$

Après le regroupement, le nombre de classes est 3*, le nombre de degrés de libertè est* $3-1=2$*. Au risque* $\alpha = 0,01$*,* $\chi^2_{2,0.99} = 9,21$*. Donc* $\chi^2_{calc} = 0,82187 < \chi^2_{2,0.99}$*. On ne rejette pas* $\mathbf{H_0}$ *au profit de* $\mathbf{H_1}$*. On considère que le nombre d'ampoules défectueuses par bôite suit une loi binomiale de paramètre* $n = 3$*,* $p = 0,15$ *au risque* $\alpha = 0,01$

2. Test de normalité

Les tests précédents sont des tests généraux s'appliquant sur n'importe quelle loi. Lorsque la loi à tester est la loi normale, on parle de test de normalité.

On cherche à se déterminer entre :

$\mathbf{H_0}$: les données suivent une loi normale.

$\mathbf{H_1}$: les données ne suivent pas une loi normale

2.1. Méthodes graphiques : Droite de Henry. La droite de Henry est une méthode pour visualiser les chances qu'a une distribution d'être gaussienne. Elle permet de lire rapidement la moyenne et l'écart type d'une telle distribution.

Principe : On représente les quantiles théoriques en fonction des quantiles observés (Diagramme Q-Q).

Si X est une variable gaussienne de moyenne $\bar{x}$ et de variance σ^2et si Z est une variable de loi normale centrée réduite, on a les égalités suivantes :

$$P(X < x_i) = P\left(\frac{X - \bar{x}}{\sigma} < \frac{x_i - \bar{x}}{\sigma}\right) = P(Z < y_i) = \Phi(y_i)$$

$y = \frac{x - \bar{x}}{\sigma}$. (on note Φ la fonction de répartition de la loi normale centrée réduite)

Pour chaque valeur x_i de la variable X, on peut calculer $P(X < x_i)$ puis en déduire, à l'aide d'une table de la fonction Φ, yi tel que $\Phi(y_i) = P(X < x_i)$.

Si la variable est gaussienne, les points de coordonnées (x_i, y_i) sont alignés sur la droite d'équation $y = \frac{x - \bar{x}}{\sigma}$

EXEMPLE 7. *Lors d'un examen noté sur* 20*, on obtient les résultats suivants :*

– 10% *des candidats ont obtenu moins de* 4
– 30% *des candidats ont obtenu moins de* 8
– 60% *des candidats ont obtenu moins de* 12
– 80% *des candidats ont obtenu moins de* 16

On cherche à déterminer si la distribution des notes est gaussienne, et, si oui, ce que valent son espérance et son écart type.

On connaît donc 4 *valeurs* x_i*, et, pour ces* 4 *valeurs, on connaît* $P(X < x_i)$*.*

En utilisant la table "Table de la fonction de répartition de la loi normale centrée réduite", on détermine les y_i correspondants :

x_i	$P(X < x_i) = \Phi(y_i)$	y_i
4	$0,10$	$-1,282$
8	$0,30$	$-0,524$
12	$0,60$	$0,253$
16	$0,80$	$0,842$

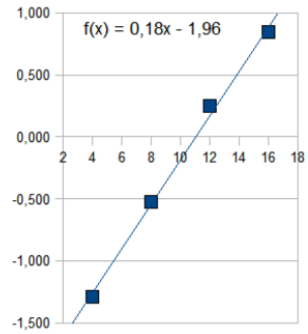


FIG. 1. Droite de Henry

Les points paraissent alignés. La droite coupe l'axe des abscisses au point d'abscisse 11 *et le coefficient directeur est* 0.18 *environ, ce qui donnerait un écart type de* $\frac{1}{0,18} = 5,6$. *Cela laisse penser que la distribution est gaussienne de paramètres* $\mu = 11$ *et* $\sigma = 5,6$.

3. Test de Khi-deux d'indépendance

Le test de khi-deux est fréquemment utilisè pour tester si deux caractères, qualitatifs ou quantitatifs (répartis en classes), observès dans une population sont indépendants ou si, au contraire, ils sont dépendants : présentent un certain degré d'association (liaison).

DÉFINITION 1 (Définition du test d'indépendance). *Le test d'indépendance est utilisé pour tester l'hypothèse nulle d'absence de relation entre deux variables qualitatives. On peut également dire que ce test vérifie l'hypothèse d'indépendance de ces variables. Si deux variables dépendent l'une de l'autre, la variation de l'une influence la variation de l'autre.*

3.1. Principe général du test :

(1) Un échantillon aléatoire de taille n est prélevé d'une population et est observé selon deux caractères X à p modalités et Y à q modalités.

(2) La répartition des n observations suivant les modalitès croisées des deux caractères se présente sous la forme d'un tableau à double entrée appelé tableau de contingence.

(3) Il s'agit par la suite de tester, à l'aide du khi-deux de Pearson, si les deux caractères sont indépendants ou non.

Tableau de contingence. Tableau des effectifs observès :

	y_1	....	y_j	...	y_l	Total ligne
x_1	n_{11}		n_{1j}	....	n_{1l}	$n_1. = \sum_j n_{1j}$
.	.		.	.	.	.
.	.		.	.	.	.
.	.		.	.	.	.
x_i	n_{i1}	......	n_{ij}	....	n_{il}	.
.	.		.	.	.	.
.	.		.	.	.	.
.	.		.	.	.	.
x_k	n_{k1}		n_{kj}	....	n_{kl}	.
Total colonne	$n._1 = \sum_i n_{i1}$	......	$n._j$	......	$n._l$	$n = n.. = \sum_i \sum_j n_{ij}$

– **Les hypothèses statistiques** peuvent s'énoncer ainsi :

$$\begin{cases} \mathbf{H}_0 \text{ les caractères : } X \text{ et } Y \text{ sont indépendants} \\ \mathbf{H}_1 \text{ les caractères : } X \text{ et } Y \text{ sont dépendants} \end{cases}$$

– **Sous l'hypothése nulle $\mathbf{H}_0$** : indépendance des deux caractères, on a, $p_{ij} = p_i.p._j : \forall(i = 1, ..., k \text{ et } j = 1, ..., l)$ (probabilités conjointes $p_{ij} = \dfrac{n_{ij}}{n}$

– l'estimation des effectifs théoriques s'obtient en répartissant la taille de l'échantillon n dans les proportions obtenues selon les estimations des probabilités conjointes (indépendance en probabilité)

(indépendance en probabilité) : $f_{ij} = \frac{n_i.}{n}\frac{n._i}{n} = \frac{\hat{n}_{ij}}{n}$: d'ou $\hat{n}_{ij} = \frac{n_i.n._i}{n}$

– Pour comparer les répartitions théorique et observèe, on calcule, sous l'hypothèse nulle $\mathbf{H}_0$ la quantité :

$$\chi^2_{calculé} = \sum_i^k \sum_j^l \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}},$$

laquelle sous $\mathbf{H}_0$ est distribuée selon la loi du khi-deux $\chi^2_{(k-1)(l-1)d.d.l}$: noté χ^2 table pour le risque dérreur α choisi.

– **Décision et conclusion du test statistique :**

L'hypothèse nulle $\mathbf{H}_0$ d'indépendance est rejetée, au niveau α, si $\chi^2_{calcul\acute{e}} \geq \chi^2_{table}$ (le test statistique est toujours unilatéral).

EXEMPLE 8. *Test d'indépendance : taux de guérison et côut du médicament.*

Pour comparer l'efficacité de 2 médicaments comparables, mais de prix très différents, la Sécurité sociale a effectué une enquête sur les guérisons obtenues avec ces deux traitements. Les résultats sont présentés dans le tableau suivant :

	Original	*Générique*	*Total*
Guérisons	*156*	*44*	*200*
Non-guérisons	*44*	*6*	*50*
Total	*200*	*50*	*250*

Au seuil de signification $\alpha = 5\%$, peut-on conclure que ces deux médicaments ont la même efficacité ?

(1) *Hypothèses statistiques :*

(2) *Seuil de signification :*

(3) *Conditions d'application du test :*

(4) *Degré de liberté :*

(5) *Statistique de test :*

(6) *Calcul de la statistique du χ^2 calculé sous l'hypothèse nulle $\mathbf{H}_0$:*

(7) *Règle de décision et conclusion*

1. Hypothèses statistiques $\begin{cases} \mathbf{H}_0 \text{ indépendance} \\ \mathbf{H}_1 \text{ dépendance} \end{cases}$

2. Seuil de signification : $\alpha = 5\%$

3. Conditions d'application du test : Un échantillon aléatoire de taille $n = 250$ observé selon deux caractères qualitatifs à $k = 2$ et $l = 2$ modalités.

4. Degré de liberté : $(k-1)(l-1) = 1d.d.l.$

5. Statistique de test : $\chi^2_{calcul\acute{e}} = \sum_i^k \sum_j^l \frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}} \sim \chi^2_{1d.d.l}$:

6. Calcul de la statistique du $\chi^2_{calcul\acute{e}}$ sous l'hypothèse nulle $\mathbf{H}_0$ Indépendance

	Original	*Générique*	*Total*
Guérisons	$\frac{200\times200}{250} = 160$	$\frac{200\times50}{250} = 40$	200
Non-guérisons	$\frac{50\times200}{250} = 40$	$\frac{50\times50}{250} = 10$	50
Total	200	50	250

$$\textit{Tableau aux effectifs théoriques } \hat{n}_{ij} = \frac{n_i.n._i}{n}$$

$$\chi^2_{calcul\acute{e}} = \sum_i^k \sum_j^l \frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}} = 2.5$$

*7. **Décision et conclusion** : fractile de la loi du χ^2_1 (cf. table) : $\chi^2_{1,\alpha=5\%} = 3,84$. La valeur du χ^2 calculé appartient à la zone de non-rejet de $\mathbf{H}_0$. En effet, χ^2 calculé $= 2,5 < \chi^2_{1,\alpha=5\%}$*

Il n'y a pas de dépendance significative entre les deux caractères : le taux de guérison et le c ôut du médicament sont indépendants. Au seuil de signification $\alpha = 5\%$, on peut conclure que ces deux médicaments ont la même efficacité

4. Test de Kolmogorov-Smirnov

Le principe est simple. On mesure l'écart maximum qui existe soit entre une fonction de répartition empirique (donc des fréquences cumulées) et une fonction de répartition théorique, soit entre deux fonctions de répartition empiriques.

Dans le premier cas, soit une fonction de répartition empirique F_n et la fonction de répartition d'une loi de probabilité théorique F.

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

Précisons que le test de K-S est indépendant de cette loi théorique : on peut comparer la répartition empirique aussi bien à une loi normale qu'à une loi de Poisson ou autre.

Etant donnés :

(1) Un échantillon de taille n d'observations d'une variable,
(2) Et une fonction de répartition de référence $F(x)$, le test de Kolmogorov teste l'hypothèse $\mathbf{H_0}$ selon laquelle l'échantillon a été prélevé dans une population de fonction de répartition $F(x)$.

Pour cela, il calcule sur l'échantillon une quantité D, appelée "statistique de Kolmogorov", dont la distribution est connue lorsque $\mathbf{H}_0$ est vraie. La statistique de Kolmogorov-Smirnov D_n est définie par

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

où $F_n(x)$ est la proportion des observations dont la valeur est inférieure ou égale à x (fonction de répartition empirique).

Une valeur élevée de D ($D = |F_n(x) - F(x)|$) est une indication que la distribution de l'échantillon s'éloigne sensiblement de la distribution de référence $F(x)$, et qu'il est donc peu probable que $\mathbf{H_0}$ soit correcte. Plus précisément,

$$P\left(\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| > \frac{c}{n}\right) \underset{\longrightarrow\infty}{\longrightarrow} \alpha(c) = 2\sum (-1)^{r-1} \exp\left(-2r^2c^2\right)$$

pour toute constante $c > 0$. Le terme $\alpha(c)$ vaut $0,05$ pour $c = 1,36$. Pour $n > 100$, la valeur critique du test est approximativement de la forme $\frac{c}{\sqrt{n}}$. Les valeurs usuelles de c en fonction de α sont :

α	$0,200$	$0,10$	$0,05$	$0,02$	$0,01$
c	$1,073$	$1,224$	$1,358$	$1,517$	$1,628$

Si $D_n > \frac{c}{\sqrt{n}}$, on rejette H_0.

Exemple 9. *Une nouvelle clientèle étrangère est attendue dans une station balnéaire. Afin de mieux connaître leurs goûts, des brasseurs ont commandé une étude de marché. En début de saison, on demande à vingt de ces nouveaux touristes de donner leur préférence parmi cinq types de bières, de la moins amère (bière* 1*) à la plus amère (bière* 5*). A l'aide d'un test de K-S, le chargé d'études décide de* *comparer les résultats avec une loi uniforme, c'est-à-dire une situation où chaque bière aurait eu la préférence de quatre répondants.*

Les résultats de l'enquête sont les suivants :

$$1\ 3\ 2\ 5\ 1\ 2\ 2\ 4\ 1\ 2\ 2\ 1\ 3\ 3\ 2\ 4\ 5\ 1\ 1\ 2$$

On se fixe un risque d'erreur de 5%*. L'hypothèse* $\mathbf{H}_0$ *à tester est celle de l'égalité avec une loi uniforme.*

Résumons les écarts entre observations et répartition uniforme :

Classe	*Effectifs*	*Uniforme*	*Cumul réel*	*Cumul théorique*	*D*
1	6	4	0,30	0,20	0,10
2	7	4	0,65	0,40	0,25
3	3	4	0,80	0,60	0,20
4	2	4	0,90	0,80	0,10
5	2	4	1,00	1,00	0,00

La distance la plus élevée s'établit à $D = 0,25$*.*

On calcule pour $n = 20$ *et* $\alpha = 5\%$ *la valeur de* $\frac{c}{\sqrt{20}} = 0,303$*. Bien que ces touristes semblent préférer les bières les moins amères, on ne peut pas rejeter l'hypothèse selon laquelle ils n'ont pas de préférence particulière.*