Chapitre 9
Statistiques et simulation

I. Le problème

On réalise n fois une expérience aléatoire qui conduit à une distribution de fréquences observées et on on souhaite évaluer l'**adéquation** de ces données **expérimentales** à une **loi équirépartie**.

Distribution des fréquences observées

valeur	x_1	x_2	$\dots$	x_i	$\dots$	x_n
fréquence	f_1	f_2	$\dots$	f_i	$\dots$	f_n

Loi équirépartie

issue	x_1	x_2	$\dots$	x_i	$\dots$	x_n
probabilité	$\frac{1}{k}$	$\frac{1}{k}$	$\dots$	$\frac{1}{k}$	$\dots$	$\frac{1}{k}$

Exemple :
Un pisciculteur possède un bassin qui contient 3 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.

II. Les étapes du test

1) *Prélèvements réels*

Exemple :
Il effectue, au hasard, 400 prélèvements d'une truite avec remise.
Il obtient ainsi : 146 truites communes, 118 truites saumonées et 136 truites arc-en-ciel.
Notons f_c, f_s et f_a les fréquences correspondantes.
On a donc :

$$f_c=\frac{146}{400}=0{,}365$$

$$f_s=\frac{118}{400}=0{,}295$$

$$f_a=\frac{136}{400}=0{,}340$$

2) Calcul de d^2_{obs}

On mesure la « distance » entre la distribution des fréquences et la loi équirépartie.

On calcule :

$$d^2_{obs} = \sum_{i=1}^{k} \left(f_i - \frac{1}{k} \right)^2$$

Remarque :
On rejettera l'hypothèse d'équirépartition si l'éloignement d^2_{obs} est assez flagrant : on va déterminer un seuil au-delà duquel il en est ainsi.

Exemple :

$$d^2_{obs} = \left(f_c - \frac{1}{3} \right)^2 + \left(f_s - \frac{1}{3} \right)^2 + \left(f_a - \frac{1}{3} \right)^2 = \left(0{,}365 - \frac{1}{3} \right)^2 + \left(0{,}295 - \frac{1}{3} \right)^2 + \left(0{,}340 - \frac{1}{3} \right)^2 \simeq 0{,}00252$$

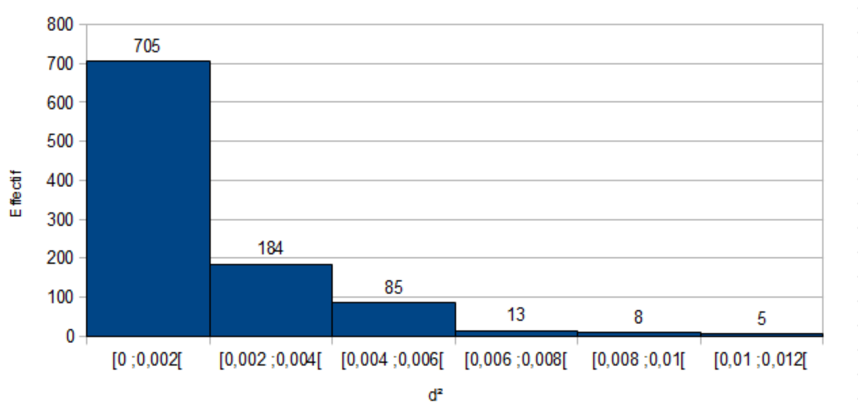
3) Simulation

On réalise une simulation de N échantillons de taille n de la loi équirépartie à k issue.
Pour chaque échantillon, on calcule la distance d^2 ; on obtient une série de N valeurs d_1, d_2, … , d_N.

Exemple :
A l'aide d'un ordinateur, le pisciculteur simule le prélèvement au hasard de 400 truites suivant la loi équirépartie. Il répète 1000 fois cette opération et calcule à chaque fois la valeur d^2.

B415 =NB.SI(B407:ALM407;"<0,004")-NB.SI(B407:ALM407;"<0,002")

	A	B	C	D	E	F	G	H	I	J
401		0	2	2	1	2	2	0	0	2
402										
403	Nb truites communes	128	138	138	143	121	143	147	123	125
404	Nb truites saumonées	141	122	132	117	140	128	112	138	134
405	Nb truites arc en ciel	131	140	130	140	139	129	141	139	141
406										
407	d²	0,0005791667	0,001216667	0,000216667	0,002529167	0,001429167	0,000879167	0,004379167	0,001004167	0,000804167
408										
409										
410	Minimum	0,0000041667								
411	Maximum	0,0119541667								
412										
413										
414	[0 ;0,002[	705								
415	[0,002 ;0,004[	184								
416	[0,004 ;0,006[	85								
417	[0,006 ;0,008[	13								
418	[0,008 ;0,01[	8								
419	[0,01 ;0,012[	5								



Parmi les 1000 prélèvements de 400 truites, 705 prélèvements ont un d^2 compris entre 0 et 0,002 ; 184 prélèvements ont un d^2 compris entre 0,002 et 0,004.

<u>Définition :</u>
Le **neuvième décile** (empirique) est le plus petit élément D_9 des valeurs des termes de la série, tel qu'au moins 90 % des données soient inférieures ou égales à D_9 .

On s'intéresse, en fait, au 9e décile D_9 de la série :

B423 =CENTILE(B407:ALM407;0,9)

	A	B	C	
401		0	2	
402				
403	Nb truites communes	128	138	
404	Nb truites saumonées	141	122	
405	Nb truites arc en ciel	131	140	
406				
407	d²	0,0005791667	0,001216667	0,00
408				
409				
410				
411				
412				
413				
414	Minimum	0,0000041667		
415	1er décile	0,0002041667		
416	2e décile	0,0003791667		
417	3e décile	0,0005541667		
418	4e décile	0,0008291667		
419	5e décile	0,0011416667		
420	6e décile	0,0015041667		
421	7e décile	0,0019616667		
422	8e décile	0,0026791667		
423	9e décile	0,0041166667		
424	Maximum	0,0119541667		
425				

Sur l'exemple, on constate que 90% des valeurs de la série des d^2 sont inférieures à ~0,004117.

4) <u>Règle de décision</u>

On compare d^2_{obs} et D_9 .

- Si $d^2_{obs} > D_9$, **on peut**, avec un risque d'erreur inférieur à 10%, rejeter l'adéquation des données observées à une loi équirépartie.
- Si $d^2_{obs} \leqslant D_9$, **on ne peut pas**, avec un risque d'erreur inférieur à 10%, rejeter l'adéquation des données observées à une loi équirépartie.

<u>Exemple :</u>
On constate que $d^2_{obs} = 0{,}00252 < D_9 = 0{,}00411$.
On dit alors, avec un risque d'erreur inférieur à 10% que « le bassin contient autant de truite de chaque variété ».

Remarques :

- Parfois, on introduit un coefficient multiplicateur pour d_{obs}^2 . C'est en particulier le cas lorsque la taille *n* de l'échantillon observé et la taille *n'* des *N* simulations ne sont pas les mêmes.
  On compare alors nd_{obs}^2 au 9e décile D'_9 de la série des *N* valeurs $n'd^2$.
- Si au lieu du 9e décile, on utilise par exemple le 95e centile, la règle de décision sera :
  Si $d_{obs}^2 > C_{95}$, on peut avec un risque d'erreur inférieur à 5%, rejeter l'adéquation des données à une loi équirépartie.