Chapitre 1

Collecte des données

1.1 Introduction

Votre premier travail, en envisageant le sujet de votre TIPE a été de déterminer la population qui vous intéresse : les daphnies, les sables de Loire, les bactéries du lait, les arbres de la forêt du Gâvre, etc. et de mettre en évidence une problématique.
D'un point de vue statistique, la première question qui se pose est celle de votre objet d'étude, on parle plus techniquement de *l'unité statistique de base*. Dans le cas de la forêt du Gâvre, tous les arbres vous intéressent-ils, les jeunes pousses comme les arbres morts, toutes les essences ? Allez-vous traiter les arbres en lisière comme ceux situés au coeur de la forêt ? Quelles bornes allez-vous fixer à votre territoire : 1 hectare, 10 hectares, 100 ? ?
Par ailleurs les individus qui composent votre population ont plusieurs *caractères* susceptibles de vous intéresser. Par exemple, pour les sables, vous souhaitez étudier leur diamètre (caractère quantitatif) ou leur altération (caractère qualitatif), dans le cas des arbres, leur âge, leur taille, leur diamètre (caractère quantitatif) ou la présence de lichens (caractère qualitatif) ?

Ces premières questions sont indispensables et une bonne connaissance de votre population et de son environnement est nécessaire ; Hormis la mise en place d'une documentation complète, c'est le moment de vos premières visites sur le terrain.

Règle N°1 : Commencer par constituer une bibliographie, la plus complète possible. Une étude en ligne, sur Internet est possible, mais il est indispensable de la compléter par des ouvrages présents soit dans votre bibliothèque soit dans la bibliothèque universitaire. Pourquoi cette nécessité ? Retenez dès à présent qu'à taille d'échantillon égale, vos résultats présenteront une moins grande dispersion et seront donc d'autant plus précis que le caractère étudié est plus homogène dans la population...

Définition :

— Les données **quantitatives** : Nombres entiers ou réels issus de comptages ou de mesures ; On parle de données **discrètes** lorsque l'ensemble des données est soit fini, soit dénombrable, de données **continues** dans le cas contraire. Ces dernières sont le plus souvent issues de mesures et, notamment en biologie, il est inutile de souhaiter une précision supérieure à 10^{-1}, 10^{-2}.

— Les données **qualitatives** : Ce sont des *caractères* ou des *attributs* que possèdent ou non les individus de votre population ; On distingue les variables **nominales** pour lesquelles tout calcul arithmétique est proscrit (couleurs d'un pelage, réponses (« oui »/« non ») dans le cas d'une enquête, etc) et les variables **ordinales** pour lesquelles un classement (ordre) est possible (par exemples les lettres A à G qui décrivent les groupes faunistiques qui interviennent dans le calcul de l'Indice Biotique). Pour ces types de données, aucun calcul ne sera effectué.

✎ ***Remarque*** : Ces quatre données peuvent tout à fait cohabiter lors d'un même prélèvement.

Supposons justement que vous vouliez mesurer l'Indice Biotique dans la vase située en bord de cours d'eau. Vous devez noter les conditions dans lesquelles s'effectuent vos prélèvements.

— La température extérieure est une donnée quantitative : elle permet d'établir des intervalles de mesures mais ne permet pas d'effectuer de rapports (Il n'y a pas de *zéro naturel* et 20 n'est pas deux fois plus chaud que 10...) ;
— La pluviométrie ou encore la profondeur à laquelle vous trouvez chacun de vos groupes sont également quantitatives mais cette fois les rapports ont un sens (20 cm est deux fois plus profond que 10 cm) ;
— Le groupe faunistique est une donnée qualitative ordinale, la couleur des Annélides est une donnée qualitative nominale.

1.2 Plans d'expériences

✎ ***Quel type d'approche choisir ?*** Lorsque vous abordez votre sujet, vous pouvez décider de deux approches possibles qui ne sont pas exclusives :

— Une étude dite « observationnelle », par ***enquête*** ou ***inventaire*** si vous souhaitez faire une observation de vos individus en limitant au maximum votre intervention, si possible dans leur environnement naturel (vous vous rendez dans la forêt du Gâvre pour mesurer, sur des parcelles que vous avez définies, la taille et la hauteur des arbres, vous étudiez sur l'agglomération nantaise la présence de lichens comme indicateur possible de pollution, etc.)

— Par une approche ***expérimentale*** dans laquelle vous cherchez la maîtrise des paramètres tout en vous imposant qu'elle puisse se répliquer sous les mêmes conditions aussi bien dans l'espace que dans le temps. Cette approche est la plus fréquente mais aussi plus complexe puisqu'elle suppose une idée claire des paramètres en jeu dont vous devrez avoir la maîtrise (évolution du pH du lait, de la couleur des pétales de fleurs, temps de parcours d'un labyrinthe par des souris, etc.).

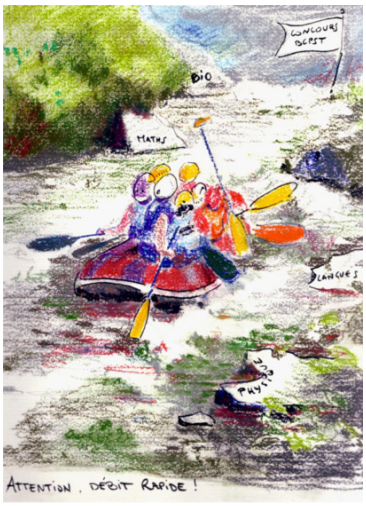
1.3 L'étude observationnelle ou « par enquête »

C'est le cas si une étude exhaustive n'est pas possible, aussi bien pour des raisons de coût que de temps. Imaginez-vous recensant tous les arbres de la forêt du Gâvre..! Il vous faut donc faire des choix et là encore, votre méthode doit s'adapter à votre objet.

Règle N°2 : Dès que possible, à moins de le regretter par la suite, penser à priviligier l'homogénéité, au regard du ou des paramètres étudiés, des individus sélectionnés pour vos observations.

1.3.1 L'échantillonnage empirique

Le phénomène que vous étudiez peut-être extrêmement précis et localisé : la teinte des lames de parquet de votre chambre, la diversité animale et végétale à proximité du saule du jardin de votre grand-mère, la vitesse du courant de surface sur les rives de la Loire, etc.
Pour des raisons qu'il vous revient de justifier, vous pouvez privilégier certaines lames du parquet (proches de la fenêtre, sous un tapis ou à proximité d'une source de chaleur), une zone à proximité du saule (ombragée, éloignée d'un mur de façade ou d'un autre arbre, ...) ou encore pour mesurer le courant de surface sur les rives de la Loire, vous ne choisissez par "au hasard" des points sur la berge mais un endroit donné qui exclut tant que faire se peut les perturbations liées au berges, aux végétaux, aux bateaux, etc.



Définition : Un mode de sondage est dit *empirique* ou *raisonné* si non seulement vous fixez le nombre d'unités à prélever mais vous expliquez le mode de prélèvement choisi, pour lequel le hasard n'a pas sa place, afin de constituer un échantillon le plus proche possible de la population étudiée.

Contrainte : Vous devez justifier avec le plus grand soin votre sondage car l'ensemble de données obtenu doit rendre compte de votre objet d'étude ; vous devez en convaincre votre lecteur, *in fine* votre jury..! Là encore une étude préalable et complète est indispensable et permettra d'appuyer votre propos. Vous penserez à définir clairement les termes utilisés et à effectuer vos prélèvements à une même date, dans les mêmes conditions. Quant au nombre de données, il n'est pas nécessairement très grand et on acceptera volontier qu'il soit compris entre 7 et 10.

✎***Définition*** : Le terme de "population" ne désigne pas nécessairement des individus mais le type de données que vous étudiez. Par exemple, dans le cas de la Loire, on parlera de la population des débits, dans le cas des lombrics, de la population des tailles, des poids, etc.

♠***Attention*** : L'inconvénient de ce type de sondage est qu'il ne vous permettra pas d'estimer la précision de vos résultats. L'ensemble des données n'a pas été constitué "au hasard" et vous ne pouvez attendre **aucun soutien des probabilités**. Vos résultats, s'ils ont été clairement justifiés, sont "vraisemblables" et ce n'est déjà pas si mal...

1.3.2 l'échantillonnage systématique

Vous étudiez la granulométrie sur une plage du littoral et vous décidez, après l'avoir quadrillée par des rectangles d'un mètre sur deux, de partir d'un rectangle choisi au hasard. Après avoir numéroté vos lignes et vos colonnes, un choix possible de sondage systématique est alors de sélectionner tous les rectangles situés sur les diagonales issues de cette coordonnée de départ ou bien de prélever un rectangle sur dix lors d'une lecture ligne par ligne...

♠***Attention*** : Si elle évite de se poser trop de questions, cette méthode n'est pas sans risque... prenez l'exemple d'un quadrillage de 10 lignes, 10 colonnes et d'un tirage tous les dix rectangles, vous retombriez dans le dernier cas systématiquement sur la même colonne !
✎On choisira typiquement ce type de sondage pour sélectionner des site de prélèvement régulièrement espacés, par exemple sur les berges d'un cours d'eau...

1.3.3 l'échantillonnage opportun

Comme son nom l'indique, vous choisissez vos individus par soucis de simplicité, parfois de temps. Pour des études sensorielles, vous vous contentez d'étudier les réponses des étudiants BCPST de votre classe, pour des carottages de vases en bord de Loire, vous choisissez les sites dont l'accès est possible rapidement et de façon répétée...

1.3.4 L'échantillonnage aléatoire

Vous souhaitez que vos conclusions visent non pas les quelques individus sélectionnés mais l'ensemble de la population dont ils sont extraits, dite *population-parent*. Aussi vous faut-il sacrifier au difficile exercice de l'échantillonnage dont l'ampleur dépend en premier lieu de la population dont seront extraits vos individus et des caractères étudiés, en particulier de leur homogénéité.

Combien ? Il est très difficile de le dire, dans l'absolu... plus de 30, si possible, pour des raisons que le chapitre "estimation" justifiera.

Premier exemple : En 2002, il a été particulièrement important d'évaluer les conséquences de la pollution du Prestige sur l'ensemble du littoral "Grand Ouest". Pour ce faire les équipes d'IFREMER ont récolté mensuellement trois échantillons du coquillage le plus représentatif du secteur (principalement huitres et moules) sur les sites de production conchylicole. Ces prélèvements étaient destinés à être comparés à la base de référence de la contamination qui porte sur l'analyse des 16 hydrocarbures Aromatiques Polycycliques (HAP) habituellement mesurés dans le cadre du Réseau National d'Observation de la qualité du milieu marin. Ces échantillons comportaient tous 50 moules, 10 huîtres, éventuellement un poisson dont l'analyse porte sur le muscle et sur le foie, prélevés "au hasard", évidemment sans remise, sur le site concerné. De quel type d'échantillonnage s'agit-il ?

Deux modes d'échantillonnages aléatoires sont possibles :

Définition : On parle d'**échantillon aléatoire** lorsque tous les individus d'une population ont la même probabilité d'être choisis.
On parle d'**échantillon aléatoire simple** de n individus si chaque série possible de taille n a la même probabilité d'être choisie.

Exemple : Vous êtes 43 en BCPST2 et vous avez besoin de constituer un panel de 5 dégustateurs, pris au hasard, dans votre classe.

— Vous faites faire 8 lignes de 5 étudiants en face de vous. Vous les numérotez et vous tirez au sort l'une de ces lignes. Votre échantillon est un échantillon aléatoire mais il n'est pas *simple* puisque chaque étudiant a 1 chance sur 8 d'être choisi mais tous les groupes de 5 n'avaient pas la même probabilité d'être choisis.

— vous numérotez les 40 étudiants possibles de 1 à 40, par exemple à partir d'un ordre alphabétique. Vous effectuez ensuite une tirage au hasard de 5 entiers compris entre 1 et 40 (matlab ou table de nombres au hasard). Le groupe ainsi formé est un échantillon aléatoire simple.

echantillonnage aléatoire simple

✐ Pour constituer un échantillon aléatoire simple de n individus pris dans une population de N unités, chacune ayant une probabilité $\frac{1}{N}$ d'être tirée, vous pouvez utiliser la fonction `randint` de la bibliothèque `random` de Python. Il suffit pour cela d'utiliser :

```
echantillon = [rdm.randint(1,N) for k in range(n)]
```

L'autre méthode possible consiste à utiliser une table dite *de nombres aux hasard*.
Cela suppose l'exercice fastidieux de dresser la liste complète des unités de la population que vous numérotez sans répétition de 1 à N avant d'extraire n nombres distincts compris entre 1 et N grâce à la table donnée en annexe, selon le protocole suivant :

— Je choisis une ligne et une colonne au hasard dans la table (exple ci-dessous : (8,12))

— Au départ du point choisi je parcours la table par lignes ou par colonnes et je prends les n premiers entiers rencontrés qui sont compris entre 1 et N.

Exemple : Reprenons l'exemple de la forêt du Gâvre, forêt domaniale de 4481 ha. Une parcelle d'un hectare a été choisie de manière raisonnée (non située en lisière, pas exploitée depuis dix ans, formée exclusivement de conifères). On souhaite rendre compte du diamètre des arbres de cette parcelle qui contient $N = 672$ arbres et pour des questions de temps, on fixe à 9 la taille de notre échantillon.

— On choisit dans la table une ligne et une colonne au hasard : $L = 8$ et $C = 12$.

— On suit verticalement la colonne 12 en prenant trois chiffres successifs. Les 10 premiers entiers rencontrés sont : 194, 400, 19, 186, 190, 494, 341, 396 et 248.

Règle N°3 : Constituer un *échantillon aléatoire* n'a rien d'aisé. L'utilisation de la table de nombres au hasard vous permet de minimiser votre rôle et donc la part de subjectivité dans le choix des individus. Elle est un argument pour justifier l'indépendance des tirage, indispensable si on envisage de tirer des conclusions sur la population parente.

Echantillonnage stratifié

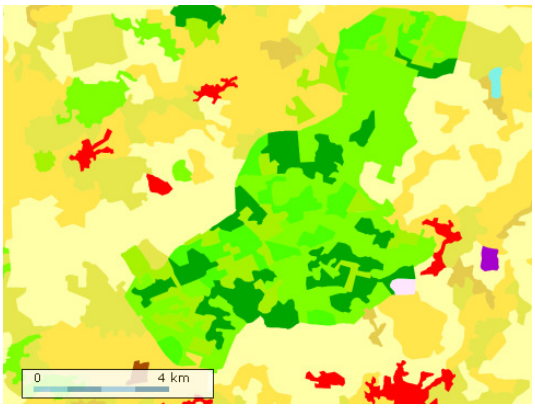
Parmis les n unités prélevées, vous pouvez constituer des sous-ensembles d'unités dites "plus représentatives" de la population, généralement à cause de similitudes au regard du caractère étudié. Ces groupes homogènes sont appelés des **strates**. On s'attend en particulier à ce que la variance à

15	62	38	72	92	03	76	09	30	75	77	80	04	24	54	67	60	10	79	26	21	60	03	48	14
77	81	15	14	67	55	24	22	20	55	36	93	67	69	37	72	22	43	46	32	56	15	75	25	12
18	87	05	09	96	45	14	72	41	46	12	67	46	72	02	59	06	17	49	12	73	28	23	52	48
08	58	53	63	66	13	07	04	48	71	39	07	46	96	40	20	86	79	11	81	74	11	15	23	17
16	07	79	57	61	42	19	68	15	12	60	21	59	12	07	04	99	88	22	39	75	16	69	13	84
54	13	05	46	17	05	51	24	53	57	46	51	14	39	17	21	39	89	07	35	47	87	44	36	62
95	27	23	17	39	80	24	44	48	93	75	94	77	09	23	48	75	91	69	03	55	51	09	74	47
22	39	44	74	80	25	95	28	63	90	41	19	48	46	72	51	12	97	39	83	35	83	23	17	29
69	95	21	30	11	98	81	38	00	53	41	40	04	16	78	67	29	83	41	18	30	90	44	37	64
75	75	63	97	12	11	57	05	86	52	82	72	47	72	14	37	72	69	75	48	72	21	52	51	81
08	74	79	30	80	70	11	66	79	25	88	01	94	52	31	38	57	98	71	62	12	56	61	01	54
04	88	45	98	60	90	92	74	77	87	40	18	65	87	37	08	68	62	39	52	84	74	90	68	18
97	35	74	05	75	42	13	49	48	38	74	19	06	42	60	20	79	90	81	77	18	51	71	27	27
53	09	93	28	29	80	19	68	30	45	94	49	49	71	21	93	93	71	30	34	52	65	83	40	13
26	36	68	48	09	37	69	26	22	80	23	34	10	45	70	83	51	07	37	44	62	96	74	42	64
49	16	57	15	79	56	63	22	94	28	11	39	69	55	38	53	06	97	20	42	09	14	90	43	48
03	51	79	78	74	75	23	73	75	98	47	85	07	26	02	61	28	01	22	16	14	12	15	67	22
21	88	87	28	48	23	44	03	03	80	53	89	07	87	93	30	17	84	17	74	16	53	31	39	01
56	41	73	33	41	59	16	59	50	98	24	24	87	06	75	99	52	09	88	05	86	25	43	50	94
72	39	19	70	17	01	04	01	22	33	04	84	63	27	65	84	39	45	55	31	95	88	93	90	37

FIGURE 1.1 – Extrait d'une table de nombres au hasard

l'intérieur de chaque strate soit sensiblement inférieure à la variance totale de la population. A l'intérieur de chacune des strates, on réalise ensuite des échantillonnages aléatoires simples, indépendants d'une strate à l'autre.
L'échantillon de la population, pour être dit « représentatif »doit contenir une proportion adéquate de chacune des states. On parle parfois aussi de "sondage par quota". Cet échantillonnage suppose, vous l'aurez noté, de connaître en amont les proportions en jeu... !

Exemple : Reprenons l'étude des diamètres des arbres de la forêt du Gâvre, forêt domaniale de 4481 ha. Nous nous appuyons sur la figure ci-dessous, obtenue sur le site « Géoportail »de l'I.G.N. qui donne l'occupation des terres. Conformément à la légende qui nous est fournie, on subdivise la population en 5 zones distinctes (les « **strates** »), respectivement forêts de caducifoliées (55%), de résineux (40%), forêt mixte (4%) et landes, broussailles, maquis (1%). Pour constituer notre échantillon de 100 individus, nous ferons un echantillonnage aléatoire simple au sein de chaque zone, en respectant les proportions des surfaces respectives, à savoir, 55 individus seront prélevés dans les forêts de caducifoliées, 40 dans les zones de résineux, 4 dans les forêts mixtes et 1 dans les zones « broussailleuses ».



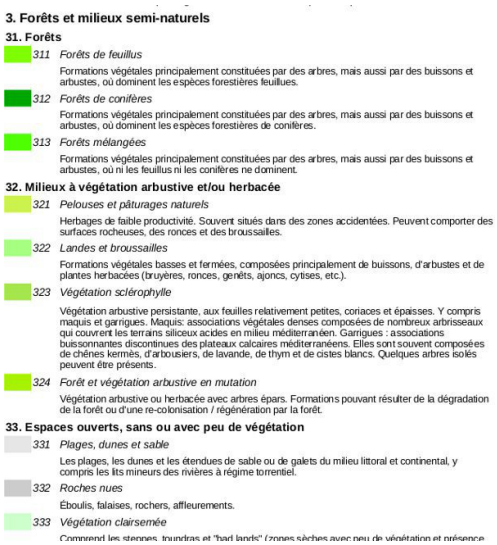


FIGURE 1.2 – Forêt du Gâvre (1 :228000) - Site Géoportail

L'échantillonnnage en grappes

Pour prélever un échantillon de n individus, on commence par partitionner la population étudiée en zones ou sections distinctes. Dans un deuxième temps, on sélectionne aléatoirement un certain nombre de zones, appelées désormais des **grappes**. **Tous** les individus de ces grappes constitueront votre « unité statistique ».

Exemple : Étude du diamètre des arbres dans la forêt du Gâvre.
Vous reprenez la carte au 1 :228000 et vous la quadrillez de zones, pas nécessairement de surfaces égales, qu'on nomme le « grappes ». Vous sélectionnez aléatoirement des grappes dont **tous** les individus seront étudiés.

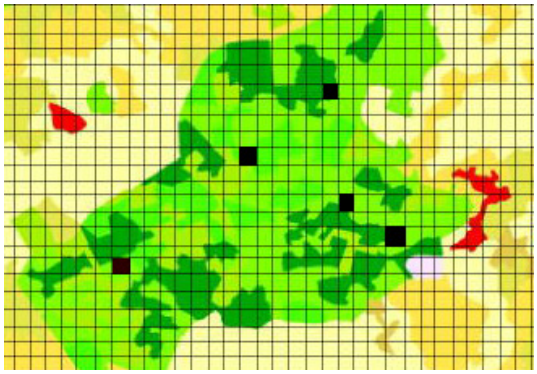


FIGURE 1.3 – Forêt du Gâvre (1 :228000) - Geoportail

♠***Attention*** : Vous ne confondrez pas « échantillonnage stratifié »et « échantillonnage en grappes »pour lequel tous les individus d'une grappe doivent être étudiés.

Cas particulier : Dans le cas d'un échantillon d'individus choisis pour une enquête ou pour un test sensoriel, méfiez-vous des volontaires. Dans ce cas, en effet, il est probable qu'une partie non négligeable des participants a des raisons non explicitées de participer et de donner son avis. D'un point de vue statistique, un tel échantillon est **faussé** et il vous sera reproché que les réponses apportées sont **biaisées**.

Exemple : *Études de la croissance des roses sur la planète du petit prince.*

FIGURE 1.4 – empirique

FIGURE 1.5 – opportun

FIGURE 1.6 – systématique

1.4 L'approche expérimentale

Elle s'envisage le plus souvent parce que vous voulez vous rendre maître de certains paramètres et que les conditions en milieu naturel ne vous permettent pas de le faire : par exemple contrôler la nature du sol, la présence d'engrais en proportion croissante, la température, l'éclairage, etc. Il est aussi possible que vous réalisiez une expérience qu'il est impossible de rencontrer dans la nature... avez-vous déjà observé une souris testant sa rapidité dans un labyrinthe... ?

Commencez par définir votre plan d'expérience.
En premier lieu, pensez que vos expériences doivent pouvoir être refaites par d'autres si nécessaire. C'est l'un des fondements de la démarche scientifique qui suppose que vos résultats puissent être mis en cause, vérifiés, soumis au regard critique de la communauté scientifique. Ceci suppose en particulier que votre protocole soit clairement décrit et que votre population soit rigoureusement définie.

Là encore, les paramètres que vous étudiez peuvent être qualitatifs ou quantitatifs. Dans le deuxième cas,(pH, température, concentration en nitrates, etc.) vous avez le choix de leurs variations mais o**n conseille le plus souvent une progression arithmétique** (température de 8, 12, 16 et 20°) **ou géométrique** (diamètres de 4, 8, 16 et 32 cm). Attention : N'oubliez pas de prendre un témoin !

Règle N°5 : Penser à **réduire vos paramètres** qui sont source de confusion.
♠ Imaginons que vous vouliez mesurer, dans le cadre d'une agriculture raisonnée, un éventuel impact de la lune montante sur la croissance de plans de salade. L'un des membres du groupe de TIPE à semé 50 graines de laitues en lune descendante, un autre 50 en lune montante. Celles semées en lune montante ont une croissance moyenne plus rapide. Pour autant, la variété choisie pour les semis en lune descendante n'étaient pas adaptés à un semis précose...
Une confusion apparaît entre les effets de la phase de la lune et ceux de la variété et ruine irrémédiablement toute conclusion !

Et le nombre d'expériences ? Vos résultats peuvent être considérés comme un échantillon d'une population plus grande. Pour chaque condition d'expérience, le paramètre que vous étudiez peut être considéré comme une variable aléatoire dont vous allez obtenir plusieurs réalisations. Comme dans la constitution de l'échantillon statistique de la section précédente, il est préférable que ces réalisations soient indépendantes. Quant au nombre, il dépend de ce que vous souhaitez faire (comparer des résultats en faisant varier des paramètres, estimer une moyenne, un écart-type,...) ? Plus de 30 données est la meilleure des situations possible (se rapporter au chapitre sur l'estimation pour une justification de ce chiffre) mais là encore, il faut s'adapter à votre population. Si pour des raisons de place, de nombre d'individus disponibles ou de temps, vous ne pouvez pas "répliquer" 30 fois votre expérience, entre 7 et 20 permet de se sortir d'affaire par l'emploi d'un test du Khi2 lors de la comparaison de répartions dans deux conditions distinctes. Dans tous les cas, plus de 5 si vous voulez que vos moyennes aient le moindre sens et intérêt... !

Règle N°6 : Constituer un *échantillon statistique* de données n'a rien d'aisé et suppose de respecter deux points. Le premier est l'invariance des conditions dans lesquelles vos résultats sont obtenus, le second l'indépendance des expériences. A titre d'exemple, les temps de parcours d'une souris dans le labyrinthe ne sont pas indépendants puisqu'un apprentissage est a priori en jeu, pas plus que les hauteurs de pluie en un site donné puisqu'un épisode pluvieux s'étend généralement sur plusieurs jours. L'indépendance, en particulier, n'est pas facile à respecter et surtout à vérifier !