2.

Quelques méthodes statistiques

Typiques de l'étude des populations
et des peuplements par la méthode des quadrats

Alain Canard
Denis Poinsot

2004

1 La répartition spatiale des individus

On appelle couramment « distribution » des individus la façon dont ils sont répartis physiquement sur le terrain. Il s'agit d'un raccourci dû au fait que les différents types de répartition peuvent être modélisés en utilisant des lois statistiques dont on sait calculer la distribution. Cependant, le mot « distribution » n'a évidemment pas du tout le même sens selon qu'on parle d'individus concrets ou d'une loi abstraite. Pour éviter toute ambiguïté, ce document réservera donc le terme « distribution » à la (loi de) distribution au sens statistique du terme. La façon dont les individus sont *concrètement* répartis sur le terrain sera appelée « répartition spatiale ».

1.1 La répartition spatiale régulière

Lorsque les individus se trouvent à peu près à la même distance les uns des autres, le nombre d'individus moyen par unité de surface est à peu près constant. Si on réalise un échantillonnage par la méthode des quadrats, on va donc obtenir des nombres d'individus très similaires d'un quadrat à l'autre. Ces nombres ne seront cependant pas identiques à cause de l'erreur d'échantillonnage (NB : *même* si les individus étaient placés de façon *parfaite* sur le terrain, l'endroit précis ou on lance le quadrat provoquerait des variations d'effectif d'un quadrat à l'autre). Ces variations, répétons le, seront faibles : si on trace le graphe représentant la fréquence des différents effectifs obtenus (qui représente une estimation de la *distribution* de la variable aléatoire « nombre d'individus par quadrat »), on va observer une faible dispersion des valeurs autour de la moyenne (donc la variance sera « faible », et on verra plus loin *faible par rapport à quoi*). D'autre part, le graphe obtenu sera symétrique[(1)] puisque les individus sont répartis régulièrement : les seuls écarts à la valeur moyenne sont dus à l'erreur d'échantillonnage, qui est aléatoire. Au total, ce graphe ne sera pas sans rappeler une courbe en cloche très célèbre. Il se trouve en effet qu'on peut assez bien représenter le résultat obtenu par une loi binomiale positive (proche de la loi normale). Or, la moyenne d'une loi binomiale positive vaut np alors que sa variance vaut npq. Comme p et q sont toujours inférieurs à 1, on en déduit que $np > npq$. Autrement dit, dans une loi binomiale positive, la moyenne ($\mu = np$) est supérieure à la variance ($\sigma^2 = npq$). Il s'ensuit fort logiquement que le rapport σ^2/μ doit être inférieur à 1. Me trompais-je ou aperçois-je des bâillements dans l'assistance ? « que nous importe le ratio σ^2/μ », dites vous ? Il nous importe parce que nos données collectées sur le terrain permettent facilement d'estimer la moyenne μ et la variance σ^2 de la variable aléatoire « nombres d'individus par quadrat » ; Ces estimations seront notées m et s^2, comme d'habitude. Le ratio s^2/m est donc une estimation du ratio σ^2/μ qui est inférieur à 1 dans le cas d'un répartition régulière. Si nous constatons que le rapport s^2/m est *significativement*

(1) Avec des données réelles, la symétrie ne sera évidemment pas parfaite, toujours à cause de l'erreur d'échantillonnage.

inférieur à 1, nous aurons alors *démontré* que la répartition spatiale des individus est régulière. Ainsi, le parallèle existant entre la loi de distribution et la répartition des individus permet de *tester une hypothèse*, ici l'hypothèse d'une répartition régulière des individus. Le rapport s^2/m est appelé **indice de répartition** et il est noté I.

La répartition régulière implique que les individus occupent le terrain approximativement à la même distance des uns des autres. Ce type de répartition apparaît donc en cas de comportement territorial. On l'observe fréquemment chez les humains sur la plage ou dans les salles d'attente lorsqu'il y a suffisamment de place libre: les individus (s'ils ne se connaissent pas) se positionnent tous *le plus loin possible les uns des autres* et le résultat est paradoxalement une répartition très régulière, dont la situation limite est atteinte dans la salle d'attente lorsque la moitié des sièges sont occupés (il ne reste alors plus de possibilité de s'asseoir sans avoir un voisin immédiat). Ce comportement territorial est également l'explication de la merveilleuse régularité d'espacement des hirondelles sur les lignes téléphonique ou l'écart entre les nids d'oiseaux coloniaux: chaque individu est placé très exactement hors de porté du bec de son voisin. A plus grande échelle, ce comportement explique la distance séparant les mâles rouge gorge à la saison des amours, et toute autre situation de territoire de chasse ou de reproduction à laquelle vous pourriez penser se décrira de façon satisfaisante en utilisant une loi binomiale positive. Notez que cette notion de « territoire » peut être étendue en quelque sorte aux plantes, et vous pourrez remarquer l'étrange régularité d'espacement des buissons dans les zone arides (lorsque la densité des individus est suffisante pour que le phénomène apparaisse). Cette régularité est due a une invisible mais féroce compétition entre racines, car chaque buisson a besoin d'une grande surface pour collecter le peu d'humidité présent et tend a inhiber les repousses dans ses environs immédiats.

1.2 La répartition spatiale en agrégats

Ce type de répartition est très courant, et on peut même affirmer qu'à partir d'une certaine échelle les individus sont toujours répartis en agrégats, car les milieux favorables aux espèces sont forcément localisés. Il s'agit cependant ici de parler d'agrégat à l'échelle très locale à laquelle s'effectuent les études de terrain. La répartition agrégative se caractérise par une tendance des individus à se grouper. Une conséquence immédiate de la répartition en agrégat, est que le terrain devient constitué essentiellement de… vide (l'espace entre les groupes, où les individus seront *relativement* rares). Mettre en évidence ce genre de situation par un échantillonnage de terrain demande donc une certaine persévérance. En effet, quadrat après quadrat, les résultats du type « zero individu » s'accumulent alors que ces satanées littorines sont pourtant bien visibles, leurs petit groupes compacts semblant ricaner dans leurs coquilles en voyant ce pauvre humain qui vise vraiment comme un pied. Naît alors chez tout expérimentateur la hideuse tentation d'aider un peu le hasard en visant sans vergogne un des paquets d'individus, d'un geste adroit et faussement désinvolte. C'est naturellement hors de question sous peine de fausser tous les résultats. Si on ne triche

pas, cependant, on finit toujours quand même par tomber sur quelques uns des groupes présents, et les valeurs obtenues dans ces quadrats miraculeux sont alors évidemment très élevées. Si on réalise le graphe des fréquence du nombre d'individu obtenu par quadrat, on obtient une courbe fortement dissymétrique, avec beaucoup de valeurs faibles (voire nulles), quelques valeurs moyennes et de rares valeurs extrêmement élevées. Cette courbe fortement dissymétrique rappelle la loi de répartition d'une binomiale négative, et on peut bel et bien ajuster approximativement une loi binomiale négative aux valeurs expérimentales. La loi binomiale négative se caractérise par rapport à la binomiale positive par une variance très élevée (due à l'étalement de la courbe vers la droite) et donc a un ratio σ^2/μ nettement supérieur à 1. Il faudra naturellement tester si l'indice de répartition s^2/m calculé avec des valeurs expérimentales est *significativement* supérieur à 1 avant de pouvoir conclure qu'on est bien en présence d'une répartition agrégative.

Les causes possibles de la répartition agrégative sont au moins triples et peuvent agir isolément ou simultanément : (i) l'hétérogénéïté de la ressource au sens large (certains microhabitats sont plus favorables que d'autres, qu'il s'agisse du gîte, du couvert ou des deux), (ii) un comportement grégaire (la recherche active de la compagnie des individus de son espèce), (iii) des capacités de dispersion faibles par rapport aux capacités de reproduction. Cette troisième cause signifie que si une espèce est peu mobile ou a au moins des juvéniles peu mobiles, on va tendre à trouver les individus en groupes au moins à certaines périodes du cycle de reproduction. Le phénomène sera exacerbé si la capacité de reproduction est élevée. Exemple : les colonies de bactéries, les colonies de pucerons (pendant leur phase non ailée). Ce sera en particulier une caractéristique de la plupart des plantes : regardez à proximité immédiate d'une plante quelconque, vous avez neuf chances sur dix d'apercevoir plusieurs autres plantes de la même espèce. Ceci est du évidemment à la faible distance parcourue*, en moyenne*, par les graines… et par les nombreux cas de reproduction végétative. Ces trois causes possibles (hétérogéneïte du milieu, grégarité, faible mobilité) expliquent que la répartition agrégative soit particulièrement courante dans le règne animal, et soit quasiment la règle dans le règne végétal.

1.3 La répartition spatiale aléatoire

Quel genre de courbe peut on bien observer quand la répartition n'est ni régulière, ni agrégative, autrement dit quand les individus sont répartis simplement au hasard ? Fort logiquement, on va observer une distribution *intermédiaire* entre la distribution binomiale positive (faible variance, relative symétrie), et la binomiale négative (forte variance, forte dissymétrie avec étalement de la distribution vers la droite). La loi de distribution statistique correspondant le mieux à cette situation est la loi de Poisson. Vous savez (**chapitre 2**) qu'on peut la considérer comme un cas limite de la loi binomiale lorsque p tend vers zéro, d'ou une nette dissymétrie de la distribution. Elle est ici adaptée car si les individus sont répartis au hasard, il va se former de façon tout à fait fortuite des plages vides mais également des agrégats. Ces vides ne seront

cependant pas aussi déserts ni ces agrégats aussi fournis que dans une véritable répartition agrégative, d'ou une distribution plus équilibrée que la binomiale négative mais une variance plus grande que la binomiale positive ou il n'y a ni espace vide ni agrégat. Vous vous souvenez bien sûr (rêvons un peu) du fait que la loi de Poisson se caractérise par une variance égale à la moyenne : $\sigma^2 = \mu$ d'ou un ratio σ^2/μ valant exactement 1. C'est cette caractéristique qui idéalement devrait nous servir d'hypothèse nulle pour les tests. En effet, si on récapitule tout ce qui précède, on a :

Répartition régulière :	Distribution binomiale positive	$\sigma^2/\mu < 1$
Répartition aléatoire :	Distribution de Poisson	$\sigma^2/\mu = 1$
Répartition agrégative :	Distribution binomiale positive	$\sigma^2/\mu > 1$

Pour caractériser le type de répartition à partir de données de terrain, il suffirait donc de calculer une estimation de σ^2/μ, c'est à dire $I = s^2/m$, et de tester si I s'écarte significativement de 1. Si I s'écarte significativement de 1 avec $I < 1$, on aura une répartition régulière, si l'écart était significatif avec au contraire $I > 1$, on conclura à une répartition agrégative.

1.4 Comment tester les modes de répartition ?

1.4.1 Test de l'indice de répartition I

On pose comme hypothèse Ho que la répartition spatiale est *aléatoire*. Dans ce cas, la variable « nombre d'individus par quadrat » suit une loi de poisson avec $s^2/m = 1$. Divine surprise, il se trouve qu'en multipliant s^2/m par $(n - 1)$ on obtient une variable aléatoire de test suivant approximativement une loi du χ^2 avec $n - 1$ degrés de liberté[(1)] sous cette hypothèse Ho. Ce test χ^2 se lit cependant d'une manière assez particulière : les 2,5% les plus élevés de la distribution du correspondent à un écart à Ho par excès d'hétérogeneité (conclusion : répartition agrégative) alors que les 2,5% les plus bas correspondent à un écart à Ho par un excès d'homogeneité (conclusion : répartition régulière). Entre ces deux bornes, on ne rejette pas Ho (répartition aléatoire jusqu'à preuve du contraire).

L'abaque qui vous est fournie vous permet de lire directement la réponse

(1) Achtung ! *n* est ici le nombre d'échantillons, par exemple, le nombre de quadrats, *et non pas le nombre d'individus présents au total dans les quadrats*.

Exemple 1.1

L'examen de 20 quadrats révèle une moyenne de 5,7 individus par quadrat et une variance de 12,2. Que conclure sur la répartition spatiale des individus ?

L'indice de dispersion vaut I = 12,2/5,7 = 2,14 $\chi^2 = 2{,}14 \times (20 - 1) = 40{,}66$

La lecture dans l'abaque pour 20 – 1 = 19 d.d.l. indique qu'on dépasse *largement* la borne supérieure du domaine « répartition aléatoire » (environ 32) et qu'on se trouve dans la zone des 2,5% supérieurs correspondant à « répartition agrégative ». On écrira : *« La répartition des individus est agrégative (test de l'indice de dispersion, $\chi^2 = 40{,}66$; 19 d.d.l, $P < 0{,}025$). »*

Cette abaque est utilisée jusqu'à $n = 50$. Cependant, à partir de $n = 30$ on peut obtenir une variable approximativement normale à partir d'une variable suivant une loi du χ^2 par un simple « changement de variable ». Ainsi, du fait de relations de parenté entre la loi du χ^2 et la loi normale, la nouvelle variable aléatoire :

$$Y = \sqrt{2\chi^2}$$

(avec χ^2 la valeur que nous avons calculé) *est approximativement normale*, de variance 1 (donc *réduite*) et de moyenne :

$$\sqrt{2 \times [nb\ de\ d.d.l\ du\ \chi^2] - 1}$$

Cette variable étant (i) normale et (ii) réduite, il suffit évidemment de la centrer en retranchant sa moyenne pour obtenir une variable normale centrée réduite $N(0:1)$. Si on n'oublie pas que le nombre de d.d.l. est ici $n - 1$ avec n le nombre de quadrats on obtient :

$$Z = \sqrt{2\chi^2} - \sqrt{2n - 3} \rightarrow N(0:1)$$

Cette variable permet alors le test Z classique :

si |Z| > 1,96, on rejette l'hypothèse Ho « *la répartition est aléatoire* » au risque $\alpha = 0{,}05$.

Pour déterminer la raison du rejet éventuel, il suffit de consulter l'indice $I = s^2/n$.

Si I < 1, alors Ho a été rejetée pour cause de répartition régulière.
Si I > 1 alors Ho a été rejetée pour cause de répartition agrégative.

Exemple 1.2

Mêmes données que dans l'exemple 1.1 mais avec 50 quadrats. Même question.

Cette fois $\chi^2 = 2{,}14 \times (50 - 1) = 104{,}86$ d'où

$$Z = \sqrt{2 \times 104{,}86} - \sqrt{2 \times (50 - 2)} = 4{,}68 >> 2{,}24$$

On rejette là encore très confortablement Ho et la table de la loi normale permet de préciser que le risque α est inférieur à 0,000 01 (car la valeur seuil de la table à *dépasser* pour P < 0,000 01 est de 2,24).

1.4.2 Test de l'indice de Morisita

Cette méthode utilise un autre indice, inventé par Morisita (1962) et noté Id :

$$Id = n \times \frac{\sum x^2 - \sum x}{(\sum x)^2 - \sum x}$$

n	nombre d'échantillons (ex : nb quadrats)
x	nombre d'individus par échantillon (ex : par quadrat)

Cet indice se comporte comme l'indice de dispersion I vu précédemment : il vaudra 1 si la répartition spatiale est aléatoire, il est supérieur à 1 si elle est agrégative et inférieur à 1 si elle est régulière (en pratique, Id ne descendra pas en dessous de environ 0,8). Notre problème vient du fait que ces valeurs sont théoriques et peuvent comme de juste être brouillées par l'erreur d'échantillonnage. Il faut donc disposer d'un test. Or, il se trouve que la variable

$$Id \times (\Sigma x - 1) + n - \Sigma x$$

à laquelle vous auriez naturellement tous pensé spontanément (hem…), suit une loi du χ^2 à $(n - 1)$ degrés de liberté lorsque la répartition spatiale est régulière. La manière d'interpréter le résultat du calcul de ce χ^2 à partir de vos données observées est alors… **identique à la méthode précédente**. Pourquoi alors présenter cette méthode ? Tout simplement parce qu'elle vous permettra de vérifier votre calcul sur l'indice de répartition s^2/m. Vous devriez en effet arriver sensiblement à la même conclusion en utilisant les deux méthodes. Si ça n'est pas le cas, il y a deux explications possibles, tout aussi intéressantes l'une que l'autre : (i) Vous aviez fait une erreur de calcul : c'est parfait, vous venez de la déceler ; (ii) vos individus sont disposés dans une situation intermédiaire entre deux modes de répartition spatiale, ce qui explique le désaccord entre les deux méthodes, qui ne sont pas *parfaitement* équivalentes (sinon Morisita ne se serait pas fatigué à inventer un second indice). Le fait que vous vous trouviez dans cette « zone grise » est bon à savoir et vous évitera d'avoir la main trop lourde lorsque vous rédigerez votre conclusion. Il faut évidemment garder à l'esprit qu'il existe une

infinité de situations entre les cas extrêmes que sont « répartition parfaitement régulière » et «répartition exclusivement en agrégat ».

1.4.3 La méthode du plus proche voisin[(1)]

Cette méthode nécessite de mesurer pour chaque individu la distance qui le sépare de son congénère le plus proche. Inutile de préciser que, sans matériel spécialisé, elle est inapplicable aux martinets en vol, aux fourmilières en activité, à l'organisation des études supérieures en France, et d'une manière générale à tout ce qui bouge trop vite pour l'œil humain. Elle sera donc réservée aux espèces qui se tiennent immobiles, au moins à certains moments ou on peut les observer.

Le principe est le suivant : si les individus sont répartis *d'une manière aléatoire* avec une densité moyenne globale de d individus par unité de surface, il a été démontré que la distance *moyenne* théorique qui les séparera sera de $1/\sqrt{2d}$ unités de longueur, cette moyenne ayant une variance de $1/(14{,}639 \times d \times n)$ avec n le nombre de mesures. Faites le test vous mêmes en répartissant par exemple 50 points *au hasard* dans un carré de 10cm × 10cm (densité moyenne : d = 0,5 individu par cm^2). Vous constaterez si vous faites suffisamment de mesures que le plus proche voisin se trouve à environ 0,7 cm *en moyenne* (c'est à dire $1/2\sqrt{0{,}5}$cm), avec cependant une variabilité assez nette selon les individus car certains, par hasard, se sont retrouvés groupés ou au contraire relativement isolés. On comprend immédiatement que si les individus se tiennent en groupe (répartition en agrégat), la distance moyenne au plus proche voisin sera beaucoup plus faible que cette valeur théorique. Au contraire, la répartition régulière donnera une valeur plus élevée que la valeur théorique $1/\sqrt{d}$, car ce mode de répartition résulte du fait que les individus tendent à s'écarter les uns des autres *à la distance maximale possible* (qui est d'environ $1/\sqrt{d}$ et non plus $1/2\sqrt{d}$). Répartissez 50 points *régulièrement* dans un carré de 10cm × 10cm (densité moyenne inchangée : d = 0,5 individu par cm^2) et vous constaterez que le plus proche voisin se trouve cette fois systématiquement à environ à 1,41 cm (c'est à dire $1/\sqrt{0{,}5}$ cm).

On sait d'après le TCL que toute moyenne suit une loi approximativement normale lorsque $n > 30$. Ce sera donc le cas de la distance moyenne mesurée sur le terrain si on dispose de suffisamment de mesures. Il suffit alors de centrer cette loi en soustrayant à la moyenne observée sa valeur théorique $1/2\sqrt{d}$ et de réduite en divisant par l'écart-type, et on va ici utiliser l'écart-type théorique puisqu'on le connaît :

$$s = \frac{1}{\sqrt{14{,}639 \times d \times n}}$$

Bref, si on appelle m la variable observée « distance moyenne au plus proche voisin », la variable de test:

(1) Rappel : « *Le voisin est un animal nuisible assez proche de l'homme.* »
PIERRE DESPROGES

$$Z = \frac{m - \frac{1}{2\sqrt{d}}}{\sqrt{14{,}639 \times d \times n}}$$

suit une loi normale centrée réduite N (0 : 1). Tout est bien qui finit bien : si Z est supérieur *en valeur absolue* à 1,96 on rejette au risque $\alpha = 0{,}05$ l'hypothèse Ho d'une répartition aléatoire.

Si Ho est bel et bien rejetée, *et uniquement dans ce cas*, la valeur observée de *m* nous indiquera le type de répartition.

Répartition agrégative, si $m < 1/2\sqrt{d}$
Répartition régulière, si $m > 1/2\sqrt{d}$.

Dans le cas ou Ho *n'est pas rejetée*, on conclut qu'on ne peut pas exclure l'hypothèse d'une répartition au hasard (encore une fois, ceci n'est *pas* équivalent à une *démonstration* que la répartition *est réellement* aléatoire).

1.5 La « représentativité » des données

La « représentativité » est une notion *qui dépend totalement de vos besoins de précision*, comme nous l'allons voir tout à l'heure[1]. La question que l'on se pose est par exemple : « ai-je suffisamment de quadrats pour fournir des estimations assez précises concernant la population générale, ou pour me permettre de faire des comparaisons entre populations ? ». Tout dépend donc complètement du degré de précision que vous voulez atteindre. Dans certaines études écologiques vous allez simplement chercher un *ordre de grandeur* (écart possible de 1 à 10 !) alors que pour d'autres (rares en biologie) vous ne pourrez rien tirer de vos données si vous n'êtes pas capables de d'obtenir une précision à 5% près. Répondre à la question « mes données sont elles représentatives » revient en pratique à calculer l'erreur standard de votre paramètre (c'est à dire son écart type) et juger si votre estimation est suffisamment précise *pour vos besoins*. Comme on peut mesurer plus d'un paramètre à partir d'un échantillon, vous pourrez constater éventuellement que *le même* échantillon est « très représentatif » pour ce qui concerne un caractère mais « très peu représentatif » en ce qui concerne un autre caractère.

Prenons un exemple simple : voici un échantillon aléatoire de dix individus, dont les tailles sont 168, 169, 172, 165, 159 et 175, 182, 180, 172, 181cm les cinq premiers individus étant des filles. Cet échantillon est il « représentatif » ? Ça dépend de quoi on parle. La moyenne de taille des cinq filles par exemple est m_f= 166,6 cm, et la variance

[1] C'est à dire *tout de suite*. Voir la célèbre scène de L'AVARE « *Hors d'ici tout à l'heure…* » que nous avons tous étudiée en classe de 4ème. Hein ? Pas vous ? Y a plus de jeunesse…

estimée de la variable aléatoire « taille des filles » est $s^2_f = 24,3$. On en déduit que l'erreur standard de m_f, c'est à dire son écart-type, vaut :

$$e.s. = \sqrt{\frac{s_f^{\,2}}{n_f}} = \sqrt{\frac{24,3}{5}} = 2,2cm$$

On écrira habituellement : $m_f = 166,6 \pm 2,2$ cm. L'erreur standard vaut ainsi à peine 1,3% de la moyenne (ce qui veut dire que l'étendue de l'intervalle de confiance sera de quelques % de la moyenne). Cette précision est suffisante en particulier pour déceler facilement un écart significatif avec la taille des garçons (vous pourrez faire le test). Pour un caractère écologique comme le nombre de littorines au m^2 cette précision serait même proprement fabuleuse (ne rêvez pas). Mais peut être étions nous plutôt intéressés par la « représentativité » de cet échantillon de 10 individus en ce qui concerne le sex-ratio ? On observe ici 50% de filles. Un simple coup d'œil à la table « précision d'un pourcentage » de votre poly vous donne directement l'intervalle de confiance. Vous apprendrez ainsi que le sex ratio de la population réelle se situe probablement quelque part entre … 19% et 81% ! Clairement, ces données sont très faiblement « représentatives » en ce qui concerne le sex-ratio. Il s'agit pourtant du même échantillon. La notion de représentativité n'est donc pas attachée aux individus physiques constituant le jeu de données mais au paramètre qu'on mesure.

1.5.1 Méthode générale à utiliser si N>30

Si $n > 30$, la moyenne du paramètre que vous observez suit une loi approximativement normale (Théorème Central Limite). s^2 étant l'estimation de la variance du caractère que vous avez mesurée par la formule habituelle, l'erreur standard de la moyenne sera :

$$e.s. = \sqrt{\frac{s^2}{n}}$$

Dans le cas particulier d'un pourcentage p on a, $s^2 = p(1-p) = pq$ d'ou :

$$e.s. = \sqrt{\frac{pq}{n}}$$

Dans les tableaux de résultats, on notera le paramètre (par exemple une moyenne) avec son erreur standard : $m \pm$ e.s. Reste à régler le problème de la « représentativité ». Pour cela, il nous faut d'abord décider le degré de précision à atteindre. Supposons que nous ayons pour objectif une amplitude maximum de l'intervalle de confiance égale à d (on peut décider par exemple que $d = 0,1 \times m$, c'est à dire qu'on veut connaître m à 10% près). Comme dans une loi normale l'intervalle de confiance (au risque $\alpha = 0,05$) vaut $1,96 \times$ e.s., il suffit ici de résoudre l'équation :

$$1,96 \times \sqrt{\frac{s^2}{n}} \leq d$$

On en déduit que notre effectif (par exemple notre nombre de quadrats si la variable est « nombre d'individus au m^2) doit être tel que :

$$n \geq 3{,}84 \frac{s^2}{d^2}$$

Si notre effectif n'atteint pas ce total, nos données ne sont pas « représentatives » *selon le degré d'exigence que nous avons estimé nécessaire*.

1.5.2 Méthode à utiliser si la distribution est normale mais N < 30

Le facteur 1,96 utilisé ci dessus suppose une la loi normale pour la moyenne étudiée. Si l'échantillon est trop petit mais que *le caractère dont on fait la moyenne* suit une distribution proche de la normale, on sait que la moyenne observée va suivre une loi du t de Student à $n - 1$ d.d.l. Il suffit a priori de remplacer la valeur 1,96 par la valeur de la table du t dans l'équation ci dessus. On tombe alors sur un os : la valeur du t… dépend évidemment de n, or c'est précisément le n nécessaire qu'on souhaite calculer. La méthode la plus simple pour s'en sortir consiste à *ne pas* calculer l'effectif nécessaire et à voir directement si l'effectif *dont on dispose* suffit oui ou non. Pour cela, on regarde dans la table du t de Student la valeur du t pour $n - 1$ d.d.l avec n notre effectif réel. Ayant fixé la marge d'erreur maximale admissible d comme précédemment, il suffit de vérifier si on a bien :

$$n \geq t^2_{n-1\,ddl} \frac{s^2}{d^2}$$

Si on souhaite vraiment connaître avec précision le n minimum nécessaire, il faut en fait procéder par essai-erreur en « essayant » un n, en faisant tourner la formule avec le $t_{n\text{-}1ddl}$ correspondant et en faisant grandir n si il n'est pas suffisant, jusqu'à satisfaction de la condition ci dessus.

1.5.3 Cas désespéré : petit échantillon et loi fortement éloignée de la normale.

Voici les formules à utiliser pour deux cas qui vous intéressent directement : la distribution de Poisson et surtout la binomiale négative (répartition agrégative). Vous trouverez leur justification détaillée dans Krebs (1989). En bref, pour une variable distribuée selon la **loi de Poisson** :
Il vous faudra :

$n > 16/m$	pour une précision de 50% de part et d'autre de m
$n > 64/m$	pour une précision de 25% de part et d'autre de m
$n > 400/m$	pour une précision de 10% de part et d'autre de m

Pour une variable distribuée selon une loi **binomiale négative** (probablement votre situation avec les quadrats…)

Il vous faudra pour les mêmes degrés de précision que ci dessus, respectivement :

$$n \geq 16 \times (\frac{1}{m} + \frac{1}{k}) \qquad n \geq 64 \times (\frac{1}{m} + \frac{1}{k}) \qquad n \geq 400 \times (\frac{1}{m} + \frac{1}{k})$$

avec k un paramètre attaché à la loi binomiale négative et qu'on peut estimer par la formule suivante (voir Krebs 1989 page 82):

$$k \approx \frac{m^2}{s^2 - m}$$

avec m et s^2 respectivement les estimations de la moyenne et la variance calculées à partir des données.

Vous vous rendrez probablement compte ainsi qu'il faut une foultitude de quadrats pour évaluer de façon satisfaisante un caractère distribué selon une distribution binomiale négative. C'est normal puisque cette distribution a une très forte variance (due à l'étalement de la distribution jusqu'aux valeurs très élevées correspondant aux agrégats).

Références

ELLIOTT, J. M., 1977. Some methods for the statistical analysis of samples of benthic invertebrates. *Freshwater Biol. Assoc. Sci. Publ*. **25** : 1-142. (*origine de l'abaque pour tester I = s^2/m*)

MORISITA, M. 1962. Id-index, a measure of dispersion of individuals. *Res. popul. Ecol*. **4** : 1-7.

2 Structure des populations et croissance

La structure d'une population est la fameuse pyramide des âges qui vous est familière. Le seul moyen rigoureux de l'établir est de dater avec certitude l'âge des individus. En pratique c'est souvent impossible sans un système de marquage-recapture, et on peut donc essayer d'utiliser la taille des individus pour *estimer* leur âge. Cette méthode *approximative* n'est évidemment utilisable que pendant la période où les individus sont en croissance…. Heureusement, certains organismes grandissent toute leur vie. *C'est en particulier le cas des mollusques*. Supposons un groupe théorique d'animaux nés dans un laps de temps bref (c'est la définition d'une cohorte). A quelle distribution des tailles doit on s'attendre après par exemple une saison de croissance ? Il est évidemment exclu que nos individus atteignent systématiquement la même taille. A cause d'un multitude d'interactions entre leur génome et l'environnement, ils atteindront au contraire *tous* des tailles différentes. D'autre part, ces interactions sont tellement nombreuses que au total la variable aléatoire « taille obtenue » résulte de *la somme de nombreuses variables aléatoires, chacune ayant un effet faible sur le résultat final.* Nous sommes dans les conditions d'application d'un vieil ami de la biologie : le Théorème Central Limite. Il s'ensuit que la taille de nos individus nés au même moment aura une distribution proche de *la loi normale*. Dans une espèce ou la croissance stoppe à un certain âge, on obtient une « photographie » de la distribution des tailles en fin de croissance en étudiant la distribution de taille des adultes. On sait depuis longtemps par exemple que la taille adulte des humains suit une distribution proche de la normale (à condition de considérer les sexes séparément, bien entendu, sinon on obtient une distribution présentant deux pics).

Sommes nous sauvés ? Pas encore, car une population est constituée d'individus toutes générations mêlées (tout le monde ne peut pas avoir la chance d'étudier les espèces d'Ephémèroptères qui ne vivent que quelques jours à l'état d'imago). Quand on établit le graphe de la distribution des tailles d'individus échantillonnés sur le terrain, on va en fait se trouver en présence de k cohortes d'individus, dont les tailles sont distribuées selon k lois approximativement normales de moyenne et de variance *différentes*. Le résultat visuel peut être net quand même si les cohortes ont des tailles nettement différentes (le cas idéal étant une seule et brève saison de naissances par an, et une croissance rapide) mais il peut aussi être désespérant de fouillis si la reproduction a lieu toute l'année (pire encore si la croissance est lente). Les méthodes que nous allons aborder maintenant *supposent que les cohortes soient suffisamment distinctes*. A vous de vous renseigner sur la biologie de votre organisme pour savoir si c'est jouable.

2.1 Comment vérifier la normalité d'une distribution ?

Nota bene : cette question n'a de sens que si vous ayez affaire à *une seule* distribution (éventuellement) normale, c'est à dire par exemple aux individus d'*une* cohorte. Voyez donc plus bas (**10.2 la décomposition polymodale**) la méthode pour séparer les cohortes les unes des autres si besoin est.

La première étape pour vérifier la normalité d'une distribution est tout simplement de *regarder vos données*. Elles doivent former une courbe en cloche *à peu près* symétrique (ne soyez pas tatillon, rappelez vous que l'erreur d'échantillonnage brouille les cartes). Si cette étape est franchie sans encombres, calculez la moyenne et la variance des tailles. Si la loi est normale, 95% des individus (*environ*, toujours à cause de l'erreur d'échantillonnage) seront situés à moins de deux[(1)] écarts-types de part et d'autre de la moyenne. Voilà pour le premier coup d'œil, qui permet en fait uniquement d'exclure les distributions qui ne sont *manifestement pas normales*. N'éliminez donc pas les cas « limite » à ce stade. Passez ensuite à l'étape suivante :

2.1.1 vérification de la normalité en utilisant le papier probit

Le « papier probit » a une ordonnée graduée de façon telle que le graphe des fréquence cumulées d'une loi normale quelconque y forme une droite. Pour vérifier si une distribution de classe de taille (par exemple) est approximativement normale, calculez la fréquence des individus appartenant à chaque classe de taille, puis reportez le graphe des fréquences *cumulées* sur une feuille de papier probit (en abscisse, vos classes de taille, en ordonnée, la fréquence cumulée). Si votre distribution est approximativement normale, les points devraient être *à peu près* alignés. Ce type de vérification est largement suffisant pour vos besoins immédiats. En conséquence **ne lisez surtout pas le paragraphe suivant**.

2.1.2 vérification de la normalité par un test statistique d'ajustement

Je savais bien que la curiosité serait la plus forte. Vous l'aurez voulu. Un moyen plus rigoureux (et il y en a d'autres *encore* plus rigoureux, c'est sans fin) consiste, à partir de la moyenne et de la variance de vos données, à calculer quels *devraient* être les effectifs de vos classes de tailles si la loi était normale. Il faut pour cela utiliser les tables de la loi normale en calculant la probabilité de se trouver dans chacune de vos k classes de taille dans une loi ayant même moyenne et même variance que les données. Sans logiciel statistique c'est long et parfaitement soporifique. Quoi qu'il en soit, après avoir multiplié ces k probabilités par l'effectif total, on se retrouve avec k effectifs théoriques. Il ne reste plus qu'à les comparer globalement avec les k effectifs observés, par un test du χ^2. *Attention* : ce test comportera k – 3 degrés de liberté (on perd 1 ddl comme d'habitude parce que l'effectif total est fixé mais aussi 1ddl pour l'estimation de la moyenne à partir des données et 1ddl pour l'estimation de la variance à partir des données). On rejette l'hypothèse qu'il s'agit d'une loi normale si le χ^2 calculé est supérieur à la valeur seuil de la table pour k – 3 ddl : ce test revient en effet à rechercher s'il y a un écart significatif entre la distribution observé et la distribution normale théorique. *Un conseil d'ami* : utilisez donc la méthode du papier probit…

2.2 la décomposition polymodale

Savoir vérifier la normalité d'une distribution isolée est bien joli, mais on a vu qu'une population est constitué d'un groupe de cohortes, *chacune* ayant une distribution de taille approximativement normale. La distribution des tailles dans la population totale

(1) La valeur exacte est bien sûr le fameux 1,96 qu'on utilise dans le test Z, mais il s'agit ici d'une vérification très approximative de toutes façons donc pas la peine de couper les cheveux en quatre.

(telle qu'on l'observe sur le terrain) résulte donc en fait de l'addition de plusieurs distributions approximativement normales. Chaque distribution normale ayant un mode (qui correspond ici à la moyenne de taille des individus de cette cohorte) on a bien une distribution polymodale. Notre mission consiste maintenant à retrouver sous la courbe totale les lois normales de chaque cohorte. C'est l'objet de la *décomposition polymodale*.

2.2.1 Décomposition polymodale par étapes

Le principe est simple. On *fait le pari* que le premier mode observé sur le graphe de la distribution des tailles (c'est à dire le premier point ou la distribution polymodale *à l'air* de marquer un plateau) correspond au mode de la première cohorte. On *suppose* ensuite que tous les individus dont les tailles sont inférieures au mode appartiennent tous à la première cohorte (si ça n'était pas le cas, cela signifierait de toutes façons que les cohortes sont trop proches pour espérer les décomposer). On enlève ces individus de la distribution totale. d'autre part, comme une loi normale est symétrique, on va enlever aussi autant d'individus à droite du premier mode, de manière symétrique (pour ôter de la distribution totale une courbe en cloche complète). Nous obtenons ainsi les individus de notre première cohorte *supposée*. On refait alors le graphe de la population totale mais sans les individus de la première cohorte *supposée*. Et on recommence : on *suppose* que le mode qui apparaît (inch' Allah) maintenant est celui de la deuxième cohorte, la suite des opérations étant la même que ci dessus. Idem pour la 3ème, 4ème…kième cohorte.

Au final, on se trouve — si on a beaucoup de chance — avec k cohortes *supposées*, dont il est bon de vérifier si elles sont bien distribuées chacune à peu près selon une loi normale (voir **10.1 ajustement à une distribution normale**), ce qui renforcerait la *présomption* qu'il s'agit bien de cohortes. Il faut d'autre part vérifier si les « cohortes » que nous avons « découvertes » nous *donnent des résultats cohérents* en terme de nombre de cohorte et de taille moyenne des individus. Trouver huit cohortes chez un animal qui ne vit que trois ans avec une seule période de reproduction par an est absurde. Réciproquement, n'en trouver que deux ou trois, avec de gros écarts de taille, chez un animal qui vit huit ans et grandit lentement devrait attirer votre attention. A vous de consulter la littérature pour déterminer si vos résultats « cadrent » avec la biologie de l'animal. S'ils ne cadrent pas du tout, peut être avez vous essayé de décomposer ce qui n'était pas décomposable (cohortes trop proches les unes des autres). Ayez alors la sagesse d'admettre qu'à l'impossible nul n'est tenu, au lieu de torturer vos données de façon plus ou moins douteuse pour faire « apparaître » les cohortes qui vous « manquent ».

2.2.2 Décomposition polymodale en utilisant le papier probit.

La distribution des tailles de votre population étant théoriquement formée de l'addition de plusieurs lois normales (une par cohortes), le graphe de fréquence cumulé total reporté sur du papier probit doit faire apparaître *plusieurs* segments de droites parallèles séparés par des décrochements (correspondant chacun à la limite entre deux cohortes). Il « suffit » de tracer sur du papier probit le graphe des fréquences cumulées de la population générale, repérer les décrochements et en déduire quelles sont les cohortes *supposées*. Naturellement, il ne faut pas se laisser aveugler par l'aspect

magique du papier probit, cette méthode étant en fait équivalente à la précédente en terme de fiabilité. Vous constaterez en particulier que les « décrochements » n'ont pas toujours la netteté d'un coup de hache, et que les « droites » ont une rectitude parfois toute relative.

La morale de cette histoire est que les méthodes de décomposition polymodale marchent très bien quand il est facile de distinguer les cohortes et très mal quand c'est difficile. Ayez donc soin d'interpréter vos résultats avec des pincettes et de ne pas présenter vos conclusions comme La Vérité Révélée. Le seul moyen fiable d'étudier la structure d'une population suppose, encore une fois, de pouvoir connaître avec précision l'âge des individus. La taille n'est qu'un indice approximatif.

2.3 La croissance

2.3.1 Croissance absolue

La croissance absolue concerne l'évolution de la taille d'un organe (ou d'un individu), en fonction du temps (elle se différentie ainsi de la croissance relative ou cette mesure est rapportée à la croissance d'un autre organe). On appelle habituellement L cette taille, puisqu'elle est mesurée par des unités de longueur. La courbe L = f(t), représentant L en fonction du temps, est le plus souvent une courbe en S, avec une croissance absolue faible initialement, forte vers le milieu de la courbe puis faible à nouveau quand l'organe s'approche de sa taille définitive. Vous avez naturellement déjà remarqué que la biologie est truffée de ces courbes en S, qu'il s'agisse de la croissance bactérienne ou de la relation dose/effet. Elles sont très jolies mais nettement moins plaisantes à manipuler mathématiquement.

Il est utile de réfléchir à la croissance d'un animal en terme de deux forces antagonistes : le prélèvement d'énergie dans le milieu, et la dissipation de cette énergie par le métabolisme. Or, l'énergie est prélevée au niveau de l'*interface* entre l'être vivant et le milieu. Cette interface (épiderme foliaire, muqueuse digestive mais aussi surface pulmonaire) est une **surface**. Elle variera donc selon un terme approximativement proportionnel au **carré** de la taille L de l'individu. En revanche, l'activité métabolique de l'organisme dissipe l'énergie disponible au sein de chaque cellule, indépendamment de la surface ayant permis d'acquérir l'énergie en question. La quantité d'énergie dissipée est relative au **volume** de l'être vivant. Elle sera donc approximativement proportionnelle au **cube** de sa taille L. L'équation régissant l'énergie disponible au final pour la croissance (car elle n'aura pas été dissipée par le métabolisme), sera donc grossièrement de la forme :

$$E = a\,L^2 \text{ (énergie prélevée dans le milieu)} - b\,L^3 \text{ (énergie dissipée hors croissance)}$$

avec a et b deux constantes caractéristiques de l'individu et fonctions probablement de son espèce mais aussi de son âge et de son état physiologique.

Sans avoir fait des maths toute sa vie, on comprend que, même si a est nécessairement supérieur à b, la taille L ne peut pas augmenter indéfiniment. Il va obligatoirement arriver un moment ou le terme dissipateur « $-\,b\,L^3$ » va rattraper au grand galop l'apport en énergie « $+\,a\,L^2$ ». la taille L de l'individu sera alors suffisamment élevée pour que *toute* l'énergie prélevée soit dissipée par le métabolisme.

Il ne sera plus possible d'attribuer de l'énergie à la croissance, l'individu aura atteint sa taille maximum. Tout ceci est bien sûr très froidement mathématique et bien trop beau pour être aussi simple, la taille adulte des êtres vivants étant placée sous *plusieurs* contraintes (*dont* l'apport en énergie détaillé ici). La résolution de l'équation ci-dessus amène cependant à un *modèle* de croissance dit de Von Bertalanffy.

La courbe logarithmique de Von Bertalanffy

Cette courbe de croissance a pour équation:

$$L_t = L_\infty \times (1 - e^{-K(t-t_0)})$$

L_t	Taille de l'individu au temps t
L_∞	Taille maximum
K	constante caractéristique de l'espèce
t	moment ou l'individu atteint la taille L
t_0	début théorique de la croissance (L = 0)

Ce qui est en fait une manière très impressionnante d'écrire que L se rapproche de plus en plus lentement de L_∞. Dans une situation nouvelle d'observation où on suppose seulement que la courbe de croissance de l'espèce suit une courbe du type Von Bertalanffy (cas des mollusques, notez le) cette unique équation nous place a priori devant trois (!) inconnues : la taille maximum de l'individu L_∞, son âge *t* et la constante K. Pourtant, on peut s'en sortir à condition de faire *deux* séries d'observations des *mêmes* individus (marquage recapture), séparées par Δt.

Par marquage-recapture de *n* individus x_1, $x_2 \ldots x_n$ on obtient deux séries de tailles : $L_{(t)}x_1$, $L_{(t)}x_2$, $L_{(t)}x_3 \ldots$ la taille des individus au temps t lors du marquage (moyenne : $m_{(t)}$, écart-type : $s_{(t)}$), et $L_{(t+1)}x_1$, $L_{(t+1)}x_2$, $L_{(t+1)}x_3 \ldots$ leur taille au temps t+1 lors de la recapture (moyenne : $m_{(t+1)}$, écart-type : $s_{(t+1)}$). Evidemment, dans la réalité de nombreux individus ne seront par recapturés et ils sont éliminés de l'analyse). Comme chaque individu est reconnaissable par son marquage, on peut calculer la corrélation entre les tailles au temps t et les tailles au temps t+1. Cette corrélation doit être très bonne (sinon on est pas en mesure de conclure quoi que ce soit, la croissance de notre échantillon a été trop hétérogène). On en déduit une droite de corrélation de la forme :

$$L_{(t+1)} = a\, L_{(t)} + b.$$

$$a = s_{(t+1)}/s_{(t)}$$
$$b = m_{(t+1)} - a \times m_{(t)}$$

Or, dans notre échantillons se trouvaient des individus plus ou moins grands. Nous pouvons donc observer normalement le phénomène de « tassement » de la croissance quand on se rapproche de la taille maximum. On va alors calculer L_∞ tout simplement en remarquant que c'est la taille pour laquelle… la croissance stoppe. Autrement dit, la taille pour laquelle $L_{(t+1)} = L_{(t)}$. Ceci revient à poser que $L_{(t)} = L_{(t+1)} = L_\infty$ puis résoudre l'équation :

$$L_\infty = a\,L_\infty + b \Rightarrow L_\infty = \frac{b}{1-a}$$

Il se trouve que la constante K vaut – log a, et on l'obtient donc immédiatement. Restent à calculer les âges des animaux. Il suffit de poser l'origine de l'axe des temps au point t_0 (donc $t_0 = 0$). On peut ensuite grâce à l'équation de Von Bertalanffy décrite plus haut reproduire toute la courbe en faisant varier t. *Attention, notre unité de temps sera l'écart entre t et t+1*. Ainsi, par exemple, si nous avons mesuré nos individus à un an d'écart, l'équation utilisera des années. Si nous avons mesuré à trois mois d'intervalle, il faudra exprimer les âges en trimestres pour utiliser l'équation. On obtiendra finalement une courbe moyenne L=F(t), qui nous permettra d'*estimer* l'âge de chacun de nos individus.

Vu votre niveau de désespoir actuel, un exemple concret semble le bienvenu. Le voici.

Exemple 2.1

On a marqué et recapturé *10 mois plus tard* des pétoncle géants (à longue durée de vie et croissance continue) avec les résultats suivants (en millimètres):

t : 64, 69, 71, 94, 104, 105, 110, 117, 126 $m_{(t)} = 95{,}56$ $s_{(t)} = 22{,}53$
t+1 : 98, 102,93, 115, 120, 126, 125, 127, 136 $m_{(t+1)} = 115{,}78$ $s_{(t+1)} = 14{,}86$

Le coefficient de corrélation vaut 0,97 ce qui est hautement significatif même avec aussi peu d'individus (voir la table du *r*, [9 + 9 – 2 =] 16 ddl, $P < 0{,}01$). Les coefficients « a » et « b » de la droite de corrélation valent respectivement :

a = 14,86/22,53 = 0,659
b = 115,78 – 0,659 × 95,56 = 52,746

d'ou la droite $\mathbf{L_{(t+1)} = 0{,}659\ L_{(t)} + 52{,}745}$

Cependant cette droite n'est valable que pour les tailles représentées par nos individus puisqu'elle représente une relation linéaire. Or, on sait que la croissance de nos individus suit une équation de Von Bertalanffy, qui est logarithmique. On détermine la taille maximum en posant que lorsque la croissance atteint son plateau, L(t) = L(t+1) = L_∞ :

$L_\infty = 0{,}659\ L_\infty + 52{,}745$ d'ou $L_\infty = 52{,}475 / (1 - 0{,}659) = 153{,}88$

La taille maximum théorique de cette espèce est d'environ 15 cm

d'autre part la constante K = - log(a) = -log(0,659) = 0,181

L'équation de Von Bertalanffy modélisant la croissance de nos mollusques est finalement (*rappel : en exprimant t en une unité valant 10 mois*)

$$\mathbf{L(t) = 153{,}88\ (1 - e^{-0{,}181\ t})}$$

Exemple : taille théorique à un an (soit 1,2 unités de 10 mois) :

$$153{,}88 \times (1 - e^{-[0{,}181 \times 1{,}2]}) = 30{,}04 \text{ mm}$$

2.3.2 Croissance relative

Ceci concerne la croissance d'un organe Y exprimée par rapport à celle d'un autre organe X (ou, d'une manière générale, une dimension biométrique Y par rapport à une autre X). La loi d'allométrie est une loi générale qui rend compte de cette croissance relative. Elle est générale parce qu'elle peut traduire aussi bien une croissance proportionnelle que déséquilibrée voire la régression d'un organe. Elle est d'équation :

$$L_Y = \beta\ (L_X)^a$$

Avec

L_Y	taille de l'organe Y
L_X	taille de l'organe X
a, β	constantes

Comme cette loi va donner des courbes et qu'il est beaucoup plus facile de manipuler des droites, on la transforme en passant au log :

$$\log L_Y = a \log L_X + \log \beta$$

En posant maintenant (juste pour faire joli) $y = \log L_Y$; $x = \log L_X$ et $b = \log \beta$, on a bien une équation de droite typique :

$$y = a\,x + b$$

Ceci signifie que, si on a la taille L_X et L_Y de deux « organes » (ou, pourquoi pas, deux dimensions du corps telle que la hauteur et la taille du péristome, pour prendre un exemple innocent), et qu'on trace le graphe $\log(L_Y) = F(\log(L_X))$ on va obtenir approximativement une droite (aux erreurs d'échantillonnage près, comme d'habitude). On appelle ces graphes les « courbes log-log ». Le coefficient « a » (qu'on peut déduire du graphe) est le *coefficient d'allométrie*, qui s'interprète de la façon suivante :

a = 1	isométrie (croissance proportionnelle ou « harmonieuse »)
a > 1	allométrie « majorante » ou « positive »(en clair, Y grandit proportionnellement plus vite que X)
0 < a < 1	allométrie « minorante » ou « négative » (Y grandit proportionnellement moins vite que X, voire ne grandit pas du tout si a=0)

a < 0 énantiométrie (l'organe Y régresse). Evidemment, l'organe nommé Y n'est pas le seul à pouvoir régresser, mais on nomme Y celui qui régresse quand il y en a un qui régresse. explication : X désigne souvent la taille du corps lui même, et autant il est courant qu'un organe régresse (exemple : le thymus chez les mammifères adultes) autant la régression de la taille du corps alors qu'un organe grandit est plus problématique.

Le coefficient « a » n'est autre que le coefficient directeur de la droite d'allométrie entre log Y et log X, qui passe par le centre de gravité du nuage de points de coordonnées [moyenne des log(x) et moyenne des log(y)]. Si on appelle s_{logx} et s_{logy} les écarts types respectifs des *logarithmes* des données originales X et Y, alors :

$$\mathbf{a} = s_{logy} / s_{logx}.$$

En clair, prenez le logarithme de vos données initiales puis calculez la variance et l'écart type, faites ensuite le ratio des écarts types. Vous obtenez immédiatement « a ».

Comme d'habitude (mais on ne la fait plus à de vieux routiers des statistiques comme vous) il est prudent de s'assurer de la fiabilité de votre estimation de « a » en calculant son intervalle de confiance avant de crier à l'allométrie majorante ou minorante simplement parce que vous avez trouvé un « a » différent de « 1,000 ». Vous pouvez le faire en utilisant la méthode du Jackknife (Excel suffit) ou du Bootstrap (si vous avez un logiciel statistique capable de le faire). Sans aller jusque là, il est utile de se demander si quelques rares individus "extrêmes" ne sont pas à eux seuls la cause d'une allométrie apparente. Eliminez les quelques points les plus éloignés de la droite et recalculez "a". Vous vous rendrez *peut être* compte alors que le "a" que vous avez calculé est finalement assez fragile.

3 Les études de peuplement

3.1 Notion de surface minimale

Lorsqu'on représente graphiquement le *nombre* d'espèces découvertes dans un milieu donné (en ordonnée) en fonction de la surface qui y a été échantillonnée au hasard (en abscisse), on obtient une courbe qui monte très vite au début puis s'infléchit jusqu'à un plateau en pente douce qui se rapproche lentement de l'horizontale. L'explication est tout simplement que les espèces fréquentes sont découvertes très vite (montée initiale rapide) et qu'on aborde ensuite les espèces moins fréquentes (la courbe s'infléchit) pour finir par les espèces rares (l'horizontale représentant le moment – qui n'arrive jamais – ou on a échantillonné une surface suffisante pour découvrir *toutes* les espèces du milieu). Comme on n'atteint jamais ce point en pratique sauf dans des cas bien particuliers, on appelle « surface minimale » la surface correspondant au début du plateau. C'est la surface qu'il « suffit » d'échantillonner pour découvrir « la plupart » des espèces présentes.

> NB : Dans la pratique, il peut arriver d'observer soudain une brusque remontée du nombre d'espèces rencontrées alors qu'on se trouvait déjà sur le plateau de la courbe. Ce phénomène signifie tout simplement que, à force d'élargir la surface de recherche, on a fini par aborder une zone de transition vers un second type de milieu. Ce second milieu ne contient évidemment pas exactement les mêmes espèces, d'ou les nombreuses nouvelles venues (les espèces fréquentes du nouveau milieu).

Lorsqu'on mène un échantillonnage sur le terrain, il est utile d'estimer où on se situe sur la courbe théorique « *nombre d'espèce trouvées en fonction de la surface échantillonnée* » pour savoir par exemple si on peut rentrer à la base à l'heure pour le thé ou s'il est au contraire opportun de réclamer par radio le parachutage de six mois de vivres pour finir le travail. Pour ce faire il suffit d'examiner les données collectées. Ces données représentent des surfaces unitaires (ex : quadrats) qu'on peut additionner à volonté pour représenter la taille de la zone prospectée, en notant en parallèle le nombre total d'espèces découvertes. On obtient une courbe, qu'il suffit de regarder pour savoir si on semble atteindre un plateau ou non. Pour éviter que l'aspect de la courbe dépende de l'ordre (artificiel) dans lequel on additionne les quadrats, on calcule pour chaque point la *moyenne* des nombres d'espèces obtenus en disposant de 1, 2, 3… (k – 1) quadrats. Pour le premier point il suffit de faire la moyenne générale. Pour le dernier, il faut en théorie calculer (k – 1) moyennes possibles. Pour les cas intermédiaires il y a de *très nombreuses* combinaisons possibles. dénués de moyens informatiques comme vous l'êtes, contentez vous de faire pour chaque point la moyenne de *quelques* (disons 5 ou 6) combinaisons seulement, choisies au hasard.

Il existe un moyen complémentaire pour estimer si il vous reste encore beaucoup d'espèces à découvrir, il est décrit plus bas (Cf. **3.2.3 Richesse spécifique** : estimateur « Jackknife » de la richesse spécifique).

3.2 Notion d'abondance

3.2.1 Densité.

La densité semble une notion très simple à définir puisqu'il suffit de compter les individus (éventuellement toutes espèces confondues) par unité de surface. Encore faut il s'entendre sur la notion de surface, comme vous vous en rendrez rapidement compte en vous échinant à récupérer de jeunes *Littorina neritoides* (vous avez déjà vu une tête d'épingle ? Pareil) soigneusement dissimulées dans des balanes mortes, elles mêmes recouvrant les flancs escarpés de patelles, elles mêmes perchées sur des rochers profondément fissurés aux formes parfaitement chaotiques. En bref, il vous appartiendra de définir de quelle surface vous parlez exactement et comment vous l'avez calculée. N'oubliez pas que le tracé tortueux de la côte de la Bretagne est l'exemple classique qui sert dans les universités du monde entier à introduire la notion de courbe fractale (courbe de dimension *intermédiaire* entre une longueur et une surface…).

Une fois que vous avez réglé la question « surface », la densité se traite statistiquement comme n'importe quelle variable quantitative. En particulier, calculez impérativement l'erreur standard de votre estimation de densité zone par zone (en vous souvenant que, dans le meilleur des cas, l'intervalle de confiance est environ deux fois plus large que l'erreur standard…). Cela aura comme heureuse conséquence de calmer vos envies initiales de conclure péremptoirement après avoir « vu » sur vos graphes des « différences » entre zones.

3.2.2 Biomasse par unité de surface.

variante du cas précédent et qui ne règle évidemment pas le problème épineux de la surface. Raisonner en biomasse est intéressant si on étudie les différents niveaux trophiques. En effet, calculer les effectifs est sans grand intérêt si les masses corporelles des espèces étudiées sont très différentes (calculer le *nombre* de pieds de graminées et le *nombre* de vaches dans un pré n'est pas très informatif en soi sur les transferts d'énergie possibles entre ces deux niveaux trophiques). Le calcul de la biomasse suppose cependant qu'on dispose d'une balance précise. Les crédits de TP étant ce qu'ils sont, oubliez la biomasse pour ce qui concerne la sortie de l'UV biocénotique.

3.2.3 Richesse spécifique.

A ne pas confondre avec la biodiversité. La richesse spécifique est simplement *le nombre d'espèces dans le milieu*, sans tenir compte le moins du monde des rapports numériques ou de biomasse entre elles : un milieu (imaginaire) comportant une seule espèce extrêmement abondante et 50 espèces rarissimes aurait la même richesse spécifique (égale à 51) qu'un milieu comportant 51 espèces en effectifs équilibrés.

Estimateur Jackknife[(1)] *de la richesse spécifique*. On a vu plus haut que la notion de « surface minimale » permettait d'estimer grosso-modo combien d'espèces il y avait au total à découvrir dans le milieu, donc la richesse spécifique, à condition d'avoir atteint la surface minimale en question. Mais comment faire si on ne l'a *pas* atteinte? En utilisant un estimateur S^* qui vaut:

$$S^* = S + \frac{n-1}{n} \times k$$

S : nombre total d'espèces effectivement observés
n : nombre de quadrats (ou d'échantillons) prélevés
k : nombre d'espèces présentes dans un seul quadrat (= espèces difficiles à trouver)

Quel est le bien-fondé de cet estimateur ? Il suppose tout benoîtement que plus il y a d'espèces « difficiles à trouver » dans l'échantillon, plus il y en a probablement dans le milieu, et donc plus on risque d'en avoir manqué lors de l'échantillonnage car par définition il est facile de ne pas trouver… une espèce difficile à trouver. Tout ceci est vigoureusement frappé au coin du bon sens (ou l'on apprend au passage que le bon sens est cunéiforme). Ainsi, grosso modo cet estimateur prédit qu'il reste autant d'espèces supplémentaires à découvrir qu'il y a d'espèces présentes uniquement dans un quadrat de l'échantillonnage. Comme tout estimateur qui se respecte, celui ci est fourni avec tout ses accessoires : on connaît la formule de sa variance, et on connaît sa distribution, qui suit approximativement une loi du t de Student (le monde est petit) avec $n - 1$ degrés de liberté (rappel : n = ici nombre de quadrats).

La variance de S* est :

$$Var(S^*) = \frac{n-1}{n} \times [\sum_{i=1}^{s} (i^2 \times q_i) - \frac{k^2}{n}]$$

n, s, k : mêmes notations que précédemment
i : un simple compteur variant de 1 à s en théorie, mais de 1 à beaucoup moins que s dans la pratique du calcul.
q_i : nombre de quadrats présentant i espèces présentes dans un seul quadrat (il est donc hautement improbable dans votre cas d'avoir à calculer des termes ou i dépasse 2 ou 3)

Il a été montré que l'estimateur Jackknife de la richesse spécifique suit une loi du t de Student avec $n - 1$ degrés de liberté, on peut si on le souhaite calculer l'intervalle de confiance autour de S^* :

$$S_{estimé} = S^* \pm t_{(n-1)} \sqrt{Var(S^*)}$$

avec $t_{(n-1)}$ la valeur lue dans la table du t de Student comme d'habitude.

(1) car cet estimateur a été obtenu par la méthode du même nom

3.3 La biodiversité

Avez vous déjà entendu parler de la biodiversité ? Non, je blague (à moins évidemment que vous ayez passé ces dix dernières années en retraite méditative au fond d'un puits de mine désaffecté). Vous pensez même probablement bien connaître cette notion, sans forcément être capable d'en donner une définition précise. Le concept de biodiversité est cependant beaucoup moins évident qu'il en a l'air. Prenons comme référence solide l'indice scientifique le plus utilisé pour mesurer la biodiversité : l'indice H' de Shannon-Wiener, abrégé habituellement en indice de Shannon. Que mesure l'indice de Shannon ? Le nombre d'espèces différentes ? Non, évidemment (pas besoin de créer un indice pour ça !). La variété taxonomique des espèces en présence alors ? Absolument pas (il s'en moque même complètement). L'indice de Shannon tel qu'on l'utilise en biologie (Shannon travaillait en fait sur la théorie de l'information) tient compte *de deux choses complètement différentes* à la fois, ce qui explique la difficulté. A savoir le *nombre d'espèces*, certes, *mais aussi le nombre d'individus de chaque espèce*. Il se note :

$$H' = -\Sigma p_i \log_2 p_i = -[\, p_A \log_2 p_A + p_B \log_2 p_B + p_C \log_2 p_C \ldots]$$

A, B, C…	les espèces présentes
p_i	proportion des individus de l'espèce i par rapport à tous les individus de toutes les espèces (= $n_i/\Sigma n$).
$\log_2$	le logarithme de base 2.

Rappel utile : $\log_2(x) = \log_{10}(x)/\log_{10}(2)$ ou encore $\ln(x)/\ln(2)$

Du fait que les logarithmes de proportions comprises entre 0 et 1 sont négatifs ou nuls, on ajoute un signe "–" devant simplement pour donner à H' un signe... positif. Le log de base 2 est hérité du fait que cet indice a été conçu par Shannon pour mesurer l'hétérogénéité d'un message, la quantité d'information portée par chaque signe étant exprimée en *bits*. Ceci dit rien n'empêche formellement de calculer un indice de type Shannon en ln ou en logs décimaux (à condition de l'indiquer au lecteur, qui souhaite peut être faire des comparaisons avec d'autres données publiées).

Entre quelles bornes l'indice de Shannon va t-il varier ? Il vaudra zéro s'il n'y a qu'une seule espèce puisque $\log_2(1) = 0$, et sera maximum, pour S espèces différentes, lorsqu'elles seront *toutes dans la même proportion* 1/S. Ainsi :

$$H'_{max} = -(S \times 1/S) \log_2 (1/S) = -\log_2 (1/S) = \log_2 S$$

Ce maximum est ainsi atteint lorsque toutes les espèces ont *le même effectif*. Le surréaliste de la chose est que dans ce cas particulier, l'*effectif lui même n'a aucune importance.* En clair si on détruisait la forêt amazonienne d'une façon strictement égalitaire jusqu'à n'avoir plus qu'un seul individu par espèce, l'indice de Shannon (la sacro sainte « biodiversité ») *serait maximum,* c'est à dire beaucoup plus élevé que

maintenant ! ! ! Certes, avec un seul individu par espèce, les choses risqueraient de se gâter assez rapidement, mais reconnaissez quand même que c'est étonnant, non ?. Pour fixer les esprits (et en utilisant un log de base 2), dans un peuplement de 10 espèces, l'indice H'max est de 3,32 environ, il passe à 6,64 avec 100 espèces et 9,96 avec 1000 espèces. Il s'agit évidemment d'un maximum purement théorique du fait que les espèces ne sont *jamais* présentes dans les mêmes proportions.

L'indice H' de Shannon étant logarithmique, une petite variation de l'indice peut représenter une grande différence réelle (en terme de *nombre* d'espèces ou en terme de *proportions* différente entre les espèces). Ceci dit, ATTENTION. Comme d'habitude, une différence entre deux milieux *calculée à partir d'un échantillonnage* peut être due uniquement aux fluctuations d'échantillonnage (le hasard), alors que vos deux « milieux » sont strictement identiques. Comment trancher ? Un bon début est de faire appel à la méthode de re-échantillonnage du Jackknife (**Zahl,** 1977). Elle vous permettra, de calculer l'intervalle de confiance autour de votre indice de Shannon.

3.4 L'équitabilité (= équirépartiton)

On appelle équitabilité ou équirépartition l'équilibre entre les effectifs des différentes espèces du milieu. Un milieu aura une équitabilité maximale si toutes les espèces ont des effectifs identiques. Or, on a vu que c'est dans ces conditions que l'indice de Shannon du milieu est maximal. Cependant, l'indice de Shannon lui même n'est pas un bon estimateur de l'équitabilité : il tend à augmenter mécaniquement avec le nombre d'espèces. Ainsi, un milieu riche en espèces mais totalement « déséquilibré » (ex : une espèce écrase numériquement toutes les autres) peut avoir un indice de Shannon plus élevé qu'un milieu parfaitement « équilibré » ayant moins d'espèces. Il faut donc un indice d'équitabilité qui soit indépendant du nombre d'espèces. Il suffit pour cela de diviser l'indice de Shannon observé dans un milieu par l'indice de Shannon maximal possible pour le nombre d'espèces de ce même milieu. Cette pondération élimine l'aspect « nombre d'espèces » et ne reste plus que l'aspect « équilibre des proportions ». Cet indice d'équitabilité est donc :

$$E = \frac{H'}{H'_{\max}} = \frac{H'}{\log_2 S}$$

H' l'indice de Shannon calculé.
S le nombre d'espèces observé

Comme tous les estimateurs, celui ci est sujet à l'erreur d'échantillonnage. On a recours, comme dans le cas précédent, à la méthode de re-échantillonnage du jackknife (**Zahl,** 1977). On peut ainsi si on le souhaite calculer les intervalles de confiance autour des valeurs E observées et éventuellement comparer les E entre plusieurs milieux. Remarque : personne ne vous *oblige* à comparer les valeurs d'équitabilité ou de Shannon que vous obtenez entre différentes zones. Cependant, si vous *souhaitez* le

faire parce que votre sujet s'y prête, il est *obligatoire* d'avoir une idée des intervalles de confiance de vos indices.

3.5 Cas particulier du milieu intertidal rocheux

En milieu intertidal rocheux (comme dans tout relief très accidenté et crevassé), la notion de surface n'est pas franchement évidente à manipuler. Plutôt que de se lancer dans des calculs cyclopéens pour essayer d'estimer la surface *réelle* échantillonnée, on peut se contenter d'indices qui ne prétendent pas connaître cette surface, et qui vont donc traiter les quadrats à égalité. Ces indices permettront *d'évaluer* (i) l'abondance d'une espèce, (ii) l'abondance relative d'une espèce par rapport aux autres (iii) les zones de préférences éventuelles de chaque espèce.

3.5.1 Fréquence d'une espèce dans une zone

C'est en fait la fréquence des *échantillons* (ici, des quadrats) qui la comportent. Notation : $F_A(Z)$ = fréquence de l'espèce A dans la zone Z, ou encore $F_{A\%}(Z)$ = idem mais exprimé en pourcentage. Selon ce qui vous intéresse, elle sera exprimée globalement ou zone par zone (ex : fréquence de l'espèce A dans la zone a *Fucus serratus*) :

$$F_A(Z) = \frac{n_A(Z)}{n(Z)}$$

$n_A(Z)$: nombre de quadrats de la zone Z comportant l'espèce A
$n(Z)$: nombre total de quadrats prélevés dans la zone Z

Pour les besoins de tests éventuels, cette fréquence se manipule évidemment comme un pourcentage (donc en se basant sur les effectifs « nombre de quadrats *avec* l'espèce, nombre de quadrat *sans* l'espèce »). Si vous souhaitez effectuer des comparaisons (entre espèces ou entre zones), vous serez donc soumis aux contraintes habituelles du χ^2 en ce qui concerne les effectifs théoriques minimum. Si vous ne *pouvez* pas effectuer ces comparaisons (nombre de quadrat trop faible), abstenez vous par pitié de toute conclusion définitive du style « nos résultats montrent que l'espèce A est plus fréquente que l'espèce B» : en l'absence de test *vous n'en avez pas le droit*.

3.5.2 Présence d'une espèce dans une zone

Comme son nom ne l'indique pas du tout, cet indice représente le poids *relatif* de la fréquence de l'espèce dans une zone donnée *par rapport à la somme de ses fréquences dans chacune des zones*. Si on appelle $F_A(Z)$ la *fréquence* de l'espèce A dans la zone Z, comme défini plus haut, la *présence* de A dans cette zone est notée $P_A(Z)$ et vaut :

$$P_A(Z) = \frac{F_A(Z)}{F_A(Z) + F_A(Z') + F_A(Z'')...}$$

Avec Z, Z', Z''… les différentes zones identifiées dans le milieu (par exemple : zone à *Laminaria*, zone a *Fucus*…). Cet indice varie de zéro (A absent de la zone Z) à 1 (A présent exclusivement dans la zone Z). C'est donc un indicateur des zones préférentielles d'une espèce. On peut bien sûr l'exprimer en pourcentage pour une meilleure lisibilité :$P_{A\%}(Z) = 100\ P_A(Z)$. Cependant il ne s'agira pas d'un ratio simple comme c'était le cas pour la fréquence (nombre de quadrats de la zone A comportant l'espèce divisé par le nombre de quadrat échantillonné dans la zone A) mais plutôt d'un ratio complexe comportant plusieurs fréquences elles mêmes basées sur des nombre de quadrats par zones éventuellement variables. Conclusion : pas de test simple, il faudrait en passer par une simulation informatique que vous ne pouvez pas faire. Restez donc prudents dans vos conclusions et, à moins que les résultats soient véritablement limpides, refusez toute affirmation définitive du type « nos résultats démontrent clairement que l'espèce X préfère la zone A alors que l'espèce Y préfère la zone B ». En effet, vous n'avez pas les moyens de mesurer la probabilité que votre observation soit due au hasard.

3.5.3 Dominance d'une espèce dans une zone

C'est le nombre d'individus d'une espèce donnée divisé par le nombre total d'individus *toutes espèces confondues*. La situation quadrat par quadrat est ici passée sous silence, on regroupe tous les quadrats d'une zone (si on s'intéresse à cette zone en particulier) voire tous les quadrats de l'échantillonnage pour avoir une vue d'ensemble. Notation : $D_A(Z)$ ou bien $D_{A\%}(Z)$ selon que vous l'exprimez en fréquence ou en pourcentage :

$$D_A(Z) = \frac{n_A(Z)}{\sum n(Z)}$$

$n_A(Z)$: nombre d'individus de l'espèce A dans la zone Z.
$\Sigma n(Z)$: nombre d'individus toutes espèces confondues dans la zone Z.

Cet indice étant une fréquence (ou pourcentage), vous pouvez effectuer éventuellement des comparaisons entre espèces en utilisant un χ^2, avec les limitations d'usage concernant les effectifs *théoriques* minimum de 5 individus. En l'absence de test, *ne faites pas* d'affirmations définitives du type « l'espèce X est plus dominante que l'espèce Y dans la zone A ».

3.5.4 dominance-présence d'une espèce dans une zone

Cet indice résulte simplement du produit de la dominance D_A par la présence P_A.

$$DP_A(Z) = D_A(Z) \times P_A(Z)$$

Il va donc varier de zéro (l'espèce A est absente de la zone Z) à 1 (l'espèce A est la seule espèce présente dans la zone Z, et c'est de plus dans cette zone que sont concentrés tous ses individus). On peut comme d'habitude l'exprimer en pourcentage. Cependant, il ne s'agit pas d'un pourcentage simple puisque $DP_A(Z)$ résulte du produit

d'une fréquence (la dominance) par un ratio complexe de plusieurs fréquences (la présence). Résultat : pas de test simple disponible pour effectuer des comparaisons, simulation informatique nécessaire mais hors de votre propos. Corollaire : encore une fois, prudence dans vos conclusions si vous utilisez cet indice, vous ne connaissez pas solidité des comparaisons que vous faites...

3.5.5 abondance-dominance d'une espèce dans une zone

Il s'agit d'un indice extrêmement simple à mettre en œuvre, à condition d'avoir l'œil, et qui a l'avantage de permettre des observations rapides et non destructrices (pas besoin de ramener les échantillons au labo). Il suffit en effet de parcourir la zone puis de noter chaque espèce de 1 à 5 :

1 : espèce très rare, 1 ou 2 individus par quadrat
2 : espèce peu nombreuse, non dominante
3 : espèce moyennement représentée
4 : espèce abondante, non largement dominante
5 : espèce très abondante et très dominante

Résultat : sur la base d'une seule sortie, toute analyse statistique est rigoureusement impossible (osez me dire, en me regardant droit dans les yeux, que vous en êtes profondément peinés, pour voir). Ce genre d'indice subjectif a un intérêt *indicatif* et permet de déblayer beaucoup de terrain en peu de temps et sans faire aucun dégât. Il permet d'effectuer des études qui demanderaient un travail de titan en suivant les méthodes plus détaillées (par exemple, couvrir de vastes zones, ou suivre des peuplements avec des intervalles très rapprochés, ou accessibles pendant très peu de temps, ou au contraire sur de longues années etc…)

3.5.6 fidélité d'une espèce à une zone

Il ne s'agit pas d'un nouvel indice mais d'une façon de classer les espèces selon leur présence $P_A(Z)$ en utilisant 5 catégories, forcément *arbitraires* :

$P_A(Z) < 0,1$	:	espèce accidentelle dans la zone Z
$0,1 < P_A(Z) < 0,2$	:	espèce accessoire dans la zone Z
$0,2 < P_A(Z) < 0,9$	:	espèce préférante dans la zone Z
$0,9 < P_A(Z) < 1$	:	espèce élective dans la zone Z
$P_A(Z) = 1$	:	espèce exclusive dans la zone Z

Il est certes plus pratique dans un texte de discuter des *espèces accessoires* dans la zone A que de parler des *espèces dont le coefficient de présence Pi est compris entre 0,1 et 0,2*. Cependant, comme chaque fois qu'on crée artificiellement des catégories dans ce qui est en fait un continuum, il ne faut surtout pas se laisser aveugler par les mots. Certaines catégories sont bien larges : une espèce « préférante » avec Px = 0,89 est à l'évidence plus proche d'une espèce « élective » avec Px = 0,91 que d'une autre espèce « préférante » avec Px = 0,21…