

suite

## La démarche scientifique

Commençons par un point essentiel ; celui de la **construction de la question** qui vous intéresse. Cette étape est souvent un préalable même à la mise en route d'une réflexion sur la nécessité d'un suivi. Pourtant, le fait qu'elle précède souvent la mise en œuvre d'un protocole ne signifie pas nécessairement que cette étape est un compartiment indépendant de la construction en soit du protocole. En effet, dans les différentes étapes que nous verrons ci-dessous, la définition précise de la question d'intérêt est toujours présente en toile de fond. Ceci signifie aussi que cette question doit toujours rester présente à l'esprit. Un protocole est toujours construit pour répondre à une question donnée. Il est pourtant fréquent que la volonté soit, soit de répondre à plusieurs questions identifiées avec un même protocole, soit qu'on utilise des données collectées pour répondre à une question afin d'apporter des réponses à une autre question qui se serait posée ultérieurement. Ces deux cas sont à bannir car il est très rare que les résultats soient au rendez-vous. Dans le cas où vous voulez répondre à plusieurs questions avec le même protocole, le risque est clairement de ne pas porter assez d'effort sur chaque question ou que le protocole soit un compromis désastreux entre différentes contraintes. Dans le cas où l'on veut répondre à des questions nouvelles, il est rare que le résultat soit concluant. Toujours est-il que dans les différentes étapes de mise en œuvre d'un protocole, il peut être pertinent de réexaminer la question initiale, notamment lorsque les résultats des pré-études et des tests de puissance démontrent par avance des difficultés à répondre avec certitude à la question posée initialement. **Un aller-retour nécessaire entre question-méthode-protocole-données-analyse est souvent très riche pour faire avancer la connaissance globale d'un système.**

La construction d'un protocole pertinent ne peut faire l'économie d'une bonne connaissance préalable de la biologie/écologie de l'espèce étudiée, mais aussi de sa répartition et de son abondance sur le site. Cela peut paraître paradoxal dans le sens où le protocole mis en place a justement pour objectif de connaître l'abondance de l'espèce. Il est aussi nécessaire de disposer d'information sur la répartition spatiale pour construire un plan d'échantillonnage adapté, mais aussi sur les abondances pour déterminer quelle méthode d'analyse de données serait la plus pertinente.

**Synthétiser les informations existantes :** une des questions centrales et des plus complexes à laquelle le montage d'un protocole doit s'atteler est celle de l'effort de terrain réalisable, mais aussi sa répartition. Pour cela, la première étape est bien entendu de collecter toutes les informations disponibles sur l'espèce dans la région étudiée. Recherche bibliographique, discussions avec les spécialistes, discussions avec les naturalistes de la région sont les trois grands axes pour acquérir des connaissances générales sur le fonctionnement des populations de l'espèce étudiée d'une manière générale, mais aussi des connaissances sur sa répartition et ses abondances locales.

**Conduire des « pré-études » :** lorsque l'espèce a été peu étudiée en amont sur le site et/ou que son écologie est mal connue, il conviendra d'envisager de décaler le protocole définitif d'un an au moins, le temps d'acquérir le minimum de connaissances nécessaires à la construction d'un protocole pertinent. En fait, il s'agit de monter un premier protocole, plus ou moins léger selon les besoins dont l'étude dispose, à mener la première année pour se fixer les idées. On parle alors de « pré-étude ». Cette pré-étude peut se construire sur un protocole du même type de celui envisagé sur le plus long terme, mais il peut aussi être différent. Par exemple, nous pouvons envisager un échantillonnage avec de très nombreuses sous-unités, mais prospectées assez brièvement pour avoir une vision globale de la présence sur les sites. Au contraire, on peut envisager de construire un protocole proche de celui qui sera mené par la suite, afin

d'obtenir des premières estimations et de calculer quel effort il faudra fournir en terme d'échantillonnage pour obtenir le degré de précision désiré (voir ci-dessous « test de puissance »).

## 1. Collecte de données

### 1.1. Concepts de biostatistique

Pour comprendre le fonctionnement souvent complexe des systèmes naturels, l'écologie s'est appuyée depuis ses origines sur les mathématiques, et plus particulièrement sur le calcul de probabilités. Les biomathématiques, ou biostatistiques, tendent à une compréhension du monde réel. Cette compréhension passe par la mise en place d'un modèle, prenant en compte un certain nombre de paramètres considérés comme causes d'un phénomène. Ce modèle constitue un objet mathématique, dont l'étude permet une meilleure compréhension du phénomène étudié, éventuellement une prédiction qualitative ou quantitative quant à son évolution future.

Aussi complexe qu'il soit, un modèle doit toujours être considéré comme une représentation simplifiée et partielle du monde réel. Une des précautions d'usage de tout scientifique est donc de garder un œil critique sur les résultats fournis par un modèle et d'être conscient de ses limites méthodologiques, afin de ne pas tirer de fausses conclusions. Pour ce faire, le scientifique se doit de toujours discuter de la pertinence des hypothèses simplificatrices de son modèle. Toutefois, en biologie des populations, la représentation simplifiée d'une réalité forcément complexe permet, dans la plupart des cas, d'obtenir des résultats relativement robustes et utiles aux gestionnaires d'espaces naturels.

#### Les concepts de base sont les suivants :

Individu : un individu statistique est une entité élémentaire sur laquelle on peut appliquer un tirage au sort. Un individu statistique peut correspondre à un individu biologique (un oiseau, une plante) mais pas nécessairement. Ce peut être également un emplacement géographique, par exemple un quadrat ou un point d'écoute, lorsqu'on effectue un tirage aléatoire des placettes/points qui seront échantillonnés dans le cadre d'un protocole.

Population : une population statistique représente un ensemble d'individus statistiques. Tout l'objet de la biostatistique est de prédire l'état d'une population globale inconnue à partir de mesures faites sur une partie des individus, constituant une population partielle connue. Si l'on considère que les individus étudiés sont représentatifs (voir ci-dessous les notions d'échantillon et de tirage aléatoire), on pourra faire l'hypothèse que les paramètres mesurés sur la population connue correspondent à ceux de la population globale inconnue, avec un certain intervalle de confiance.

Echantillon : l'échantillon désigne un fragment d'un ensemble prélevé pour juger de cet ensemble. Dans la plupart des cas, il s'agit d'une collection d'éléments, d'individus statistiques, prélevés de façon particulière de la population statistique afin de tirer des conclusions sur cette dernière. Cependant, il arrive parfois que tous les éléments de la population soient étudiés. On parle alors d'échantillonnage exhaustif ou de recensement. L'effectif de l'échantillon peut varier de l'unité à l'effectif  $N$  de la population statistique. L'effort d'échantillonnage sera mesuré par la taille  $n$  de l'échantillon et par la fraction d'échantillonnage, égale à  $n/N$ .

Tirage aléatoire : pour que des résultats puissent être généralisés à la population statistique, l'échantillon doit être représentatif de cette dernière, c'est-à-dire qu'il doit refléter fidèlement sa composition et sa complexité. Seul l'échantillonnage aléatoire assure la représentativité de l'échantillon mais on verra plus loin que la notion de stratification conduit à relativiser l'équivalence entre aléatoire et représentativité. Contrairement à une idée reçue, **la taille de l'échantillon n'affecte pas la représentativité, car elle ne**

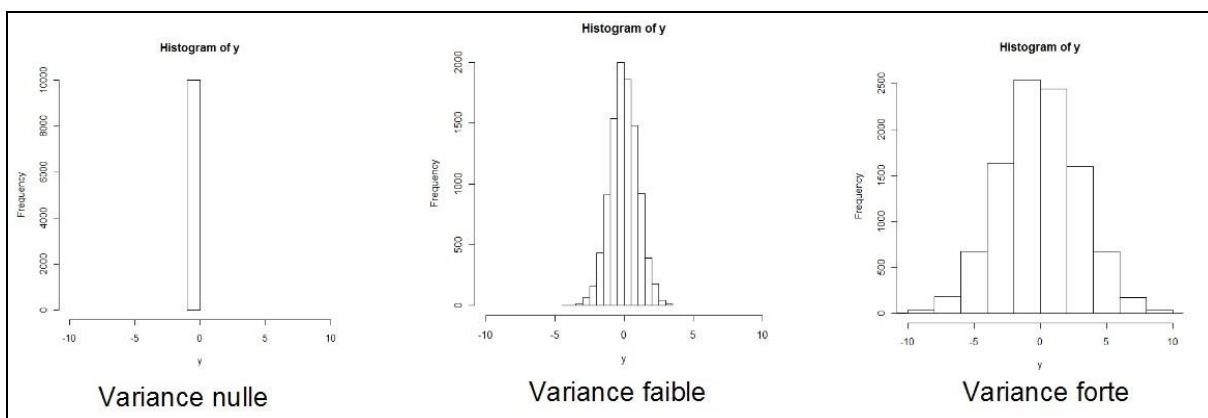
**fait que varier l'intervalle dans lequel la vraie valeur d'un paramètre de la population a de fortes chances de se trouver** (= intervalle de confiance). En outre, contrairement à une pratique assez répandue en écologie, **la sélection comme stations d'échantillonnage de sites jugés subjectivement représentatifs du milieu étudié, ne constitue pas un échantillon aléatoire et peut fortement biaiser les résultats** comme il le sera clairement montré avec des exemples précis dans la suite du document.

Variable : une variable est une caractéristique mesurée ou observée sur chacun des éléments de l'échantillon, ou sur des entités prédéfinies qui se rattachent aux unités d'échantillonnage (ex : milieu environnant). Plusieurs catégories de variables peuvent être définies, dont trois sont essentielles :

- les variables quantitatives, mesurables par une unité de mesure sur une échelle ordinale (c'est-à-dire ordonnée intrinsèquement. Ex : nombre d'œufs par nid, distance en mètres, poids en kilogrammes, etc.),
- les variables semi-quantitatives ou variables de rang, lorsqu'un découpage en plusieurs classes permet une mesure selon une échelle ordinale (ex : stades larvaires classés du premier au dernier, classes d'âges, etc.),
- les variables qualitatives, lorsque les données recueillies sur une échelle nominale (et non ordinale) sont susceptibles de classement dans des catégories (qualitatives) collectivement exhaustives et mutuellement exclusives (ex : sexe, couleur du pelage, etc.).

Moyenne : elle exprime la valeur qu'aurait chacun des individus de la population (ou de l'échantillon) s'ils étaient tous identiques et sans changer la dimension globale de la population.

Variance : elle permet de caractériser la dispersion des valeurs par rapport à la moyenne. On peut interpréter la variance comme la moyenne des carrés des écarts à la moyenne (rigoureusement: l'espérance des carrés des écarts à l'espérance; vulgairement : moyenne des carrés moins le carré des moyennes). Ainsi, entre deux populations présentant la même espérance, celle ayant une plus grande variance apparaîtra comme plus dispersée (nuage de points plus étalé). Le fait que l'on prenne le carré de ces écarts à la moyenne évite que des écarts positifs et négatifs ne s'annulent.

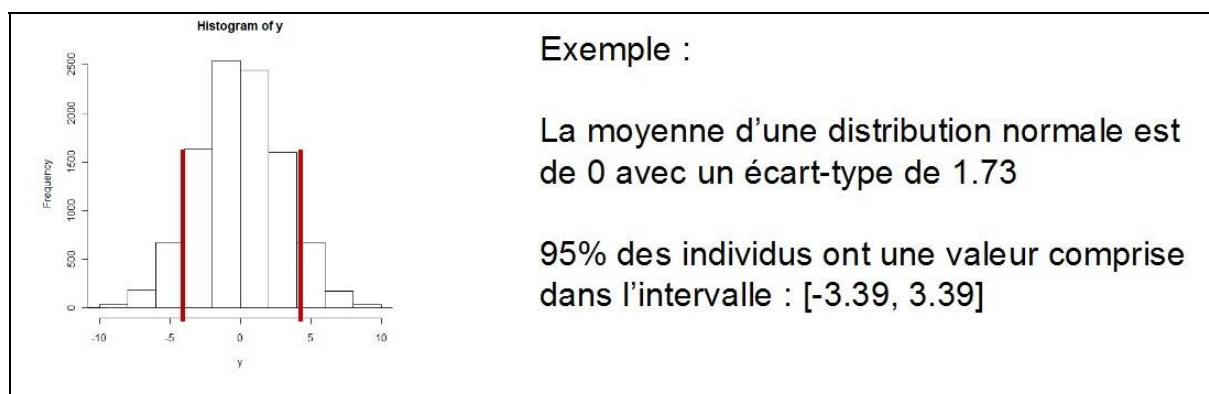


**Figure 1 : Illustration de la variance, dispersion des données autour de la moyenne**

**Écart-type** : il s'agit là encore d'une mesure de la dispersion d'une série de valeurs autour de leur moyenne. Il est égal à la racine carrée de la variance. L'écart-type est une mesure capitale pour comparer plusieurs moyennes. Ainsi, deux échantillons (ex : noyaux de population alpin et méditerranéen d'une même espèce) affichant chacun une moyenne différente mais avec un large écart-type (forte dispersion des données), ne pourront pas être qualifiés comme statistiquement différents. L'écart-type est de même unité que la variable étudiée.

**Coefficient de variation** : il s'agit là encore d'une mesure de la dispersion d'une série de valeurs autour de leur moyenne. Il est égal à l'écart-type divisé par la moyenne. Il a pour caractéristique d'être indépendant de l'unité de la variable et donc de faciliter les comparaisons entre des variables d'ordres de grandeurs très différents par exemple.

**Intervalle de confiance** : on parle d'intervalle de confiance lorsque l'on donne un intervalle qui contient, avec un certain degré de confiance, la valeur à estimer. Le degré de confiance est en principe exprimé sous la forme d'une probabilité. Par exemple, on dit souvent qu'un intervalle de confiance à 95% (ou au seuil de risque de 5%) a une probabilité égale à 0,95 de contenir la valeur du paramètre que l'on cherche à estimer. En pure rigueur statistique, cette conclusion n'est pas vraie mais ce détail dépasse le cadre de ce document. Ainsi, lorsqu'on effectue un sondage (tirage au hasard d'un sous-ensemble d'une population), l'estimation d'une quantité d'intérêt donnée est soumise au hasard et correspond rarement exactement à la valeur réelle de la quantité que l'on cherche à estimer. En présentant pour l'estimation non pas une valeur mais un encadrement, on quantifie l'incertitude sur la valeur estimée. Plus l'intervalle de confiance est de taille petite, moins il y a d'incertitude sur la valeur estimée. Si la variable étudiée suit une loi normale, l'intervalle de confiance à 95 % est égal à la moyenne plus ou moins  $[1,96 \times \text{écart-type}]$ .



**Figure 2 : Exemple d'intervalle de confiance.**

**Loi normale** : la loi normale est l'une des distributions de probabilité les plus connues et intuitives. Elle a été introduite par le mathématicien Abraham de Moivre en 1733, qui l'a utilisée afin d'approcher des probabilités associées à des variables aléatoires binomiales possédant un paramètre  $n$  (= effectif) très grand. Cette loi a été mise en évidence par Gauss au XIXe siècle et permet de modéliser de nombreuses études biométriques. Sa densité de probabilité dessine une courbe dite courbe en cloche ou courbe de Gauss, centrée sur sa moyenne et symétrique de part et d'autre de cette moyenne. Elle est généralement associée à des variables continues.

**Loi de Poisson** : cette loi, introduite au début du XIXe siècle par le magistrat français Siméon-Denis Poisson, s'applique aux phénomènes accidentels ou très rares, où la probabilité  $p$  d'occurrence de l'événement est très faible. En écologie, cette loi est souvent utilisée pour compter des événements ou des individus qui sont distribués aléatoirement dans l'espace et dans le temps. Par exemple, pour dénombrer des individus dans des quadrats lorsque la répartition des organismes est aléatoire. Dans la pratique, la loi de Poisson est utilisée lorsque la probabilité est très faible ( $p < 0,05$ ) et le nombre de sondages très élevé ( $n > 50$ ), car pour voir apparaître les événements rares il faut multiplier les épreuves. Exemple : probabilité de capture d'oiseaux rares lors de sessions de capture au filet japonais.

## 1.2. Notions de bases en échantillonnage

Une des petites sous-unités, petite zone qui sera suivie se nomme un individu statistique. L'ensemble de ces petites sous-unités se nomme l'échantillon. L'ensemble de toutes les petites sous-unités possibles (c'est-à-dire toute la zone d'étude) se nomme la population statistique.

Le point principal à retenir lorsque l'on définit une stratégie d'échantillonnage est que la meilleure manière d'avoir une vision représentative du système étudié est de réaliser un tirage aléatoire des sous-unités suivies. Cela peut paraître contre-intuitif pour des personnes ayant une bonne connaissance de leur zone d'étude, pourtant c'est la seule démarche valable statistiquement pour garantir des résultats non biaisés. Si les sous-unités sont tirées aléatoirement et qu'elles sont en nombre suffisant, **alors l'aléatoire assure la neutralité par rapport à la zone, et le nombre assure la représentativité.**

**Un biais très fréquent lorsque l'on étudie des populations et de ne se concentrer que sur les zones où l'espèce est la plus abondante car c'est la zone où nous obtiendrons le plus d'informations. Cette démarche est à proscrire lorsque l'on cherche à estimer la taille d'une population et surtout à la suivre dans le temps.**

Une autre grande généralité sur l'échantillonnage est que plus on augmente le nombre de sous-unités suivies, plus les résultats seront précis. A l'extrême, si on suit toutes les sous-unités possibles, on aura alors un suivi exhaustif de la zone. Cela signifie aussi que si l'on suit peu de sous-unités, elles seront représentatives de la zone puisque tirées aléatoirement mais elles ne représenteront qu'une petite partie des possibles et la confiance en leur extrapolation sera faible, d'où une précision plus faible de nos résultats.

La précision d'une estimation en statistique se mesure par l'intervalle de confiance de cette estimation. Par exemple, on dira qu'il y a en moyenne 10 individus par sous-unité avec un intervalle de confiance compris entre 5 et 15, à 95% de probabilité. Ceci signifie qu'il y a en moyenne 10 individus par sous-unité étudiée mais que la moyenne réelle de toutes les sous-unités possibles se situe avec 95% de chance entre 5 et 15 individus. **Un point crucial à retenir est qu'une moyenne n'a que peu de valeur de résultat dans un rapport. Ce qui importe, c'est l'intervalle de confiance, car c'est lui qui donne la confiance que l'on a dans le résultat énoncé.** Vos conclusions ne seront pas les mêmes si vous avez une moyenne de 50 individus par sous-unité avec un intervalle de confiance compris entre 10 et 90, que s'il est compris entre 48 et 53.

Attention à ne pas mal interpréter les liens entre taille de l'échantillon (nombre de sous-unités suivies), biais et précision. Si la méthode que vous utilisez pour suivre votre population est biaisée, c'est-à-dire qu'elle fournit des résultats erronés, alors l'augmentation de la taille de l'échantillon ne réduit pas ce biais, elle améliore la précision de la valeur mais le biais reste de même ordre. Par exemple si vous sous-estimez de 25% la taille de la population sur chaque sous-unité, alors multiplier le nombre de sous-unités ne résoudra pas le problème.

## Définir la zone d'étude / la notion de favorable

Définir la zone d'étude peut présenter certains problèmes. Le premier est d'ordre écologique : il s'agit de définir les limites de la population. En effet, il existe une très forte variabilité des populations et de leur fonctionnement. Ainsi, comprendre ce qui se passe sur une zone donnée nécessite souvent d'avoir accès aux informations sur ce qui se passe à l'extérieur du site si la population est de taille très importante. Par exemple, dans le cas de suivis d'espèces migratrices, la chute importante d'une population ne pourra pas nécessairement être enrayée par l'amélioration des sites de reproduction ou des sites de halte migratoire si les causes de la chute sont sur les zones d'hivernage. Dans le cas d'une population à très large répartition, il peut en être de même. L'amélioration d'un site peut ne pas avoir de conséquence sur la dynamique globale de la population.

**Une première règle d'or est qu'il faut essayer de prospecter le plus largement possible.** Ainsi pour s'assurer d'avoir une bonne représentativité de ce qu'il se passe sur le site, il faudra éviter de ne se concentrer que sur une petite partie du site mais au contraire essayer de répartir l'effort sur l'ensemble du site.

Lorsque l'on a une connaissance préalable plus ou moins fine de la répartition des individus sur la zone d'étude, il est fréquent pour le naturaliste de ne choisir de prospecter que les zones à forte densité de manière à maximiser le nombre de données collectées et de compter le maximum d'individus. Malheureusement, cette démarche est à proscrire lorsque l'on tente d'avoir une estimation de la taille de la population et de son évolution sur l'ensemble du site. En effet, cette démarche pose de nombreux problèmes. Le premier est évidemment d'obtenir une estimation de la taille de la population sur l'ensemble de la zone. Le fait de ne suivre que les zones à forte densité n'informe pas nécessairement sur le nombre de ces zones (outre celles déjà connues) et surtout n'informe pas sur la surface de zones occupées à faibles densités, qui peuvent parfois être très grandes et par conséquent abriter une très forte proportion de la population. Le deuxième biais de cette méthode concerne l'évolution dans le temps des populations et il est crucial de bien le comprendre. Le fait de ne suivre que les zones à forte densité a généralement pour conséquence qu'on ne peut observer que des décroissances ou, au moins, une stabilité des effectifs. En effet, des zones très favorables à l'espèce peuvent soit devenir défavorables (on observe donc une décroissance), soit rester stables de même que leur population. Pourtant dans cette situation, nous ne disposons d'aucune information sur les zones initialement moins favorables. Ces zones peuvent être devenues favorables et donc présenter d'importantes hausses d'effectifs. Elles peuvent aussi s'être encore plus dégradé et donc présenter une décroissance des effectifs, même si les zones favorables, elles, sont restées stables.

Prenons deux exemples.

### ***Exemple 1 : suivi du lagopède alpin dans les Pyrénées.***

Le lagopède alpin est étudié depuis une trentaine d'années dans les zones centrales des Pyrénées, qui abritent de bonnes populations. Ces populations sont stables dans l'ensemble. On pourrait donc conclure à un bon état de santé des populations. Pourtant, le message d'une forte régression de cette espèce est largement véhiculé. En fait, ce constat vient d'une vision empirique et est notamment tiré du fait que les populations des zones périphériques des Pyrénées semblent décroître voire disparaître. Malheureusement, le protocole utilisé ne permet pas de s'en assurer. Il ne permet donc pas de bien comprendre le risque auquel est soumise cette espèce.

### **Exemple 2 : suivi de tétras sur des places de chant.**

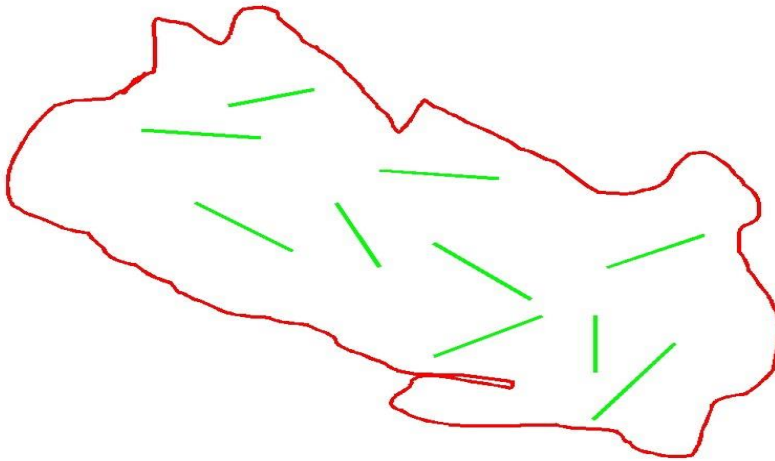
Cet exemple est issu d'une étude menée aux USA sur une espèce de Tétras. Le suivi de cette espèce se réalise sur des places de chant connues sur une grande zone géographique. Cette zone a été largement prospectée dans le cadre d'un programme initial visant à décrire sur la situation de l'espèce dans cette région. Les places de chant sont repérées lors de cette première phase. Elles ont par la suite été suivies annuellement avec un protocole très rigoureux (passages multiples pendant la saison, météo calibrée...) et cela pendant une vingtaine d'années. Les auteurs de cette étude montrent avec ces suivis que l'espèce est en décroissance forte et notamment dans les zones d'implantation de puits de forage de gaz. Ce résultat peut paraître cohérent, pourtant il est fortement critiquable. En effet, seuls les sites connus à une date donnée sont suivis. Cette espèce est connue pour être fidèle à ses places de chant mais ceci est connu lorsque l'habitat est relativement stable. Lorsque l'habitat se dégrade, il est probable que cette espèce puisse abandonner des places de chant devenues « défavorables » et s'installer ailleurs (à quelques centaines de mètres par exemple, voire beaucoup plus loin si nécessaire). Par conséquent, il est abusif de dire que l'espèce a régressé. La conclusion rigoureuse est que l'espèce a abandonné une partie de ces places de chant et notamment celles localisées dans les zones de forage. Cependant, le protocole ne s'intéressant pas du tout à l'émergence possible de nouvelles places de chant au cours de la période d'étude, nous ne pouvons pas rejeter l'hypothèse selon laquelle les oiseaux se sont simplement déplacés et que les effectifs sont stables (ils pourraient même être en augmentation). La conséquence fréquente d'un suivi sur les meilleurs sites est une vision pessimiste de la situation car ces sites ont tendance à se dégrader.

Ces deux exemples illustrent bien le risque très important qu'il existe à ne suivre que les zones les plus favorables. Il faut de fait suivre l'ensemble des zones potentiellement occupées, même par quelques individus, et **surtout anticiper sur les changements potentiels futurs de manière à ne pas exclure de l'échantillonnage des sites peu favorables aujourd'hui mais qui pourraient dans quelques années devenir favorables**, pour une raison quelconque (abandon des pratiques agricoles, remise en état des sites...). Au mieux, dans une stratégie d'échantillonnage, on pourra exclure les zones pour lesquelles l'absence est certifiée (quitte à la vérifier sur un faible nombre de sous-unités) et pour lesquelles la stabilité dans le temps est assurée (ex : zones urbaines, zones d'altitude...).

## **1.3. Les grands types de plan d'échantillonnage**

### **1.3.1. Echantillonnage classique**

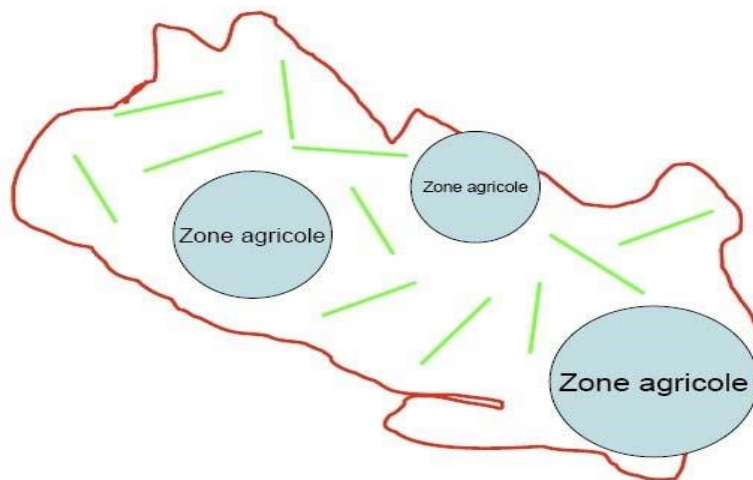
Un plan d'échantillonnage peut être très simple. Il peut consister à sélectionner aléatoirement un certain nombre de sous-unités au sein de la zone d'étude, petites sous-unités qui seront suivies par la suite. Il peut s'agir de transects, de points d'écoutes, de quadrats, etc.



**Figure 3 : Echantillonnage simple. Placement aléatoire de 10 transects au sein d'une zone d'étude**

L'effort alloué à chaque sous-unité (leur surface dans le cas de quadrats, leur longueur pour des transects, le temps passé pour des points d'écoute...) et leur nombre dépendra du budget temps disponible pour l'étude. Il faudra évidemment trouver un compromis entre suivre suffisamment les sous-unités pour avoir des résultats qui ont du sens (avoir une bonne détection par exemple) et en suivre suffisamment pour que l'échantillon complet soit représentatif de la zone d'étude. Pour cela, il n'existe pas de règle fixe mais nous fournirons quelques conseils dans la partie décrivant les méthodes d'analyses de données car chaque méthode va poser des contraintes particulières en termes de taille d'échantillon.

Techniquement, la sélection aléatoire de sous-unités est aisée. Dans le cas de quadrats ou de parcelles par exemple, on commence par découper la zone d'étude en quadrats/parcelles (sur photos ou carte à la main). La forme des parcelles/quadrats a généralement peu d'importance mais on verra par la suite qu'il faudra s'efforcer qu'ils fassent tous la même taille pour la simplicité des calculs. Chaque quadrat est ensuite numéroté. A l'aide d'un logiciel tel que R (logiciel gratuit, cf. lien en annexe) ou même avec Excel, on tire alors aléatoirement une suite de chiffre qui correspondront aux quadrats qui seront suivis.

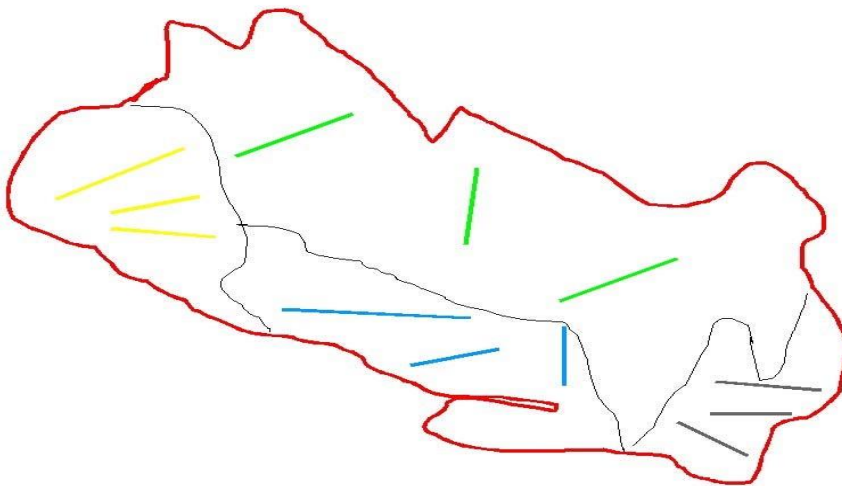


**Figure 4 : Echantillonnage aléatoire avec exclusion des zones non occupées (ici les zones agricoles)**



### 1.3.2. Echantillonnage stratifié

Dans de nombreuses situations, l'abondance de l'espèce étudiée varie très fortement selon des caractéristiques de l'habitat. Souvent les naturalistes connaissent par avance, de par leur expérience ou de par la littérature, quels sont les habitats les plus favorables à l'espèce et ceux les moins favorables. Lorsque dans un échantillon, on mélange des données issues de sites à forte densité et des sites à faible densité (grande variabilité des abondances mesurées sur les sites), alors la précision sur la moyenne du nombre d'individus par site est très faible (grande variance). Dans la situation où l'on peut regrouper, **en amont** de l'étude (attention à ne jamais réaliser ceci après la collecte des données), les différentes sous-unités suivies en quelques classes qui présenteront des abondances différentes, alors on peut construire un échantillonnage dit « stratifié ». Ce type d'échantillonnage consiste à définir les catégories de sites, à les identifier sur le terrain puis à tirer aléatoirement les sites qui seront suivis en fixant le nombre de sous-unités que l'on veut suivre dans chaque catégorie ou strate. Ce type d'échantillonnage présente deux intérêts : il permet d'améliorer la précision des estimations de l'abondance par strate mais aussi globalement, il permet de faire varier l'effort de terrain selon les strates. En effet, le nombre de sous-unités suivies de chaque strate n'a pas besoin d'être le même. On peut alors imaginer de suivre plus intensément les sites à forte densité (afin d'avoir une bonne précision des estimations sur ces sites) et moins intensément les sites à faible densité (puisque'ils apporteront de toute façon moins d'information sur la taille globale de la population).



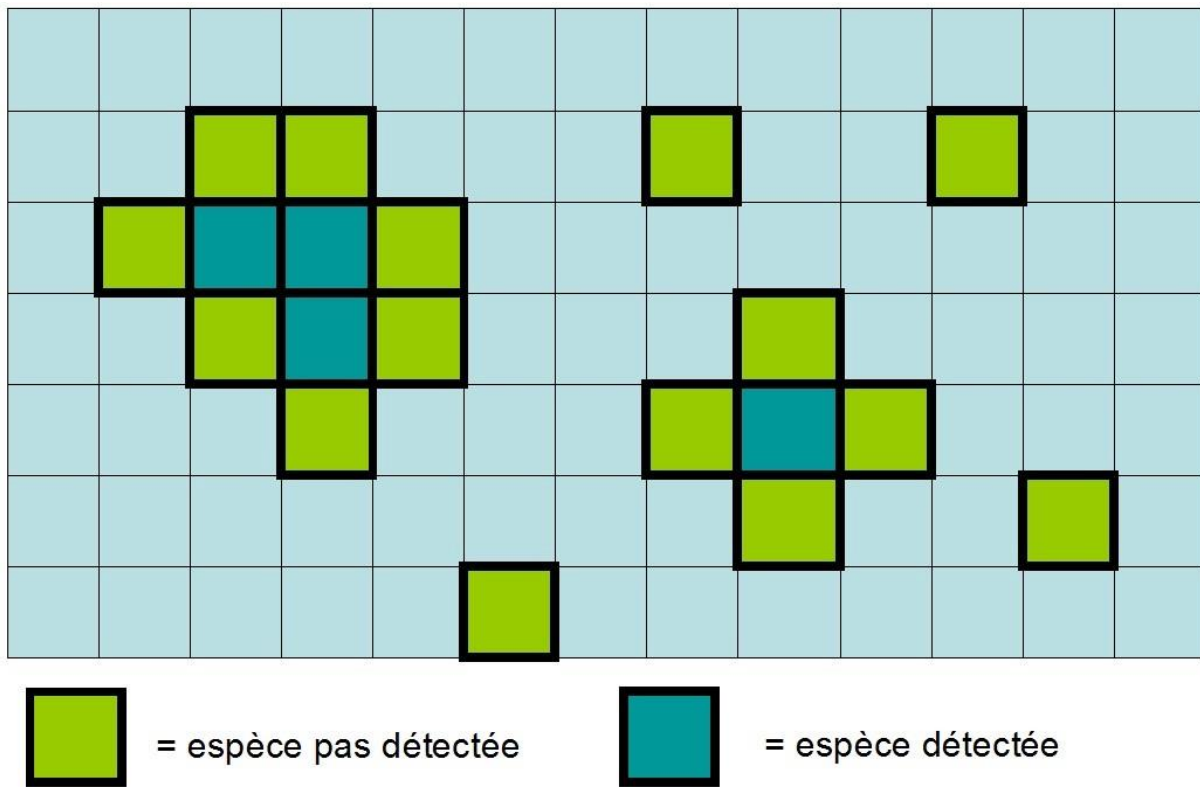
**Figure 5 : Echantillonnage stratifié. Placement aléatoire de 12 transects au sein d'une zone d'étude, répartis pas secteurs**

Prenons un exemple très concret, même si un peu caricatural, permettant d'illustrer l'importance de la stratification. Nous nous intéressons à une **espèce de chauve-souris** présente dans le sud de la France. Cette espèce est connue pour occuper depuis des décennies cinq grandes grottes, chacune abritant environ 10.000 individus. Par ailleurs, cette espèce occupe aussi de toutes petites grottes, très nombreuses (pour fixer les idées nous dirons qu'environ 10.000 petites grottes sont occupées). Chaque petite grotte abrite cependant uniquement quelques individus, disons cinq en moyenne. Les naturalistes suivent les cinq grandes grottes depuis des années. Si une des grandes grottes devient défavorable pour une raison quelconque et que cette grotte est désertée par les chauves-souris, alors nous observons une baisse de 20% des effectifs de cette espèce (50.000 à 40.000). La chute est importante et impose de mettre en place des mesures de protection plus ou moins coûteuses. Cependant, si en parallèle un autre

naturaliste s'intéresse aux petites grottes et observe un accroissement de cinq à six individus en moyenne par grotte dans un échantillon de petites grottes, alors cet expert parle lui d'une augmentation de la population de 20%. Que conclure ? En fait la population de cet exemple théorique est bien évidemment stable, les 10.000 individus qui ont abandonné la grande grotte se sont simplement redistribués dans les petites. Mais chaque groupe de naturalistes se basant sur une population statistique de grottes différentes (les grandes vs les petites), leurs conclusions sont totalement opposées. Une solution pour cette situation serait de construire deux strates : une strate avec les grandes grottes qui pourrait être suivie exhaustivement (les cinq) et une strate avec les petites grottes qui serait suivie par échantillonnage (par exemple une centaine). Ce plan d'échantillonnage assure de pouvoir extrapoler les résultats à l'ensemble de la population. Notons que cet exemple montre aussi les limites au suivi des sites les plus favorables à l'espèce, limites que nous avons déjà abordées auparavant.

### 1.3.3. Adaptive sampling

Lorsque l'on cherche à suivre une espèce rare et/ou une espèce dont la population est répartie de manière agrégée dans l'espace (regroupement de nombreux individus sur quelques petits sites, au contraire d'une répartition homogène dans l'espace), on comprend aisément qu'un échantillonnage classique risque de conduire à ne suivre que des sous-unités sur lesquelles l'espèce est absente, à moins de pouvoir suivre un très grand nombre de sous-unités. Pour remédier à ce problème, les méthodologistes ont développé l'*adaptive sampling*. Il consiste à tirer aléatoirement un certain nombre de sous-unités que l'on va suivre. Pour le moment, rien de nouveau. Cependant, l'originalité de ce type de protocole réside dans le fait que, lors du terrain, de nouvelles sous-unités suivies seront ajoutées à l'échantillon selon les résultats obtenus sur les sous-unités sélectionnées initialement. Ainsi ce type d'échantillonnage propose que, lorsque la sous-unité suivie est positive pour la présence de l'espèce, on prospecte les sous-unités adjacentes. Si une ou plusieurs nouvelles sous-unités sont elles aussi occupées par l'espèce, alors on prospecte là aussi les sous-unités adjacentes, etc. Lorsqu'une sous-unité suivie est négative pour la présence de l'espèce, on prospecte alors la sous-unité suivante de l'échantillon retenue en tout premier lieu. Il en est de même lorsque l'on a suivi plusieurs sous-unités adjacentes et que toutes sont négatives. On comprend bien que ce type de démarche permet de mieux suivre les espèces agrégées, car l'effort sera augmenté sur les zones de présence.



**Figure 6 : Adaptive sampling**

Attention, en pratique il faut limiter l'utilisation de ce type de protocole aux espèces dont on est sûr qu'elles sont très rares ou très localisées. En effet, si ce n'est pas le cas, vous risquez de devoir prospecter un nombre considérable de sous-unités. En effet, une sous-unité positive entraîne plusieurs nouvelles prospections qui seront alors positives si l'espèce n'est pas vraiment agrégée et ces nouvelles prospections entraînent à leur tour de nouvelles prospections, etc. Certains auteurs ont suggéré d'adopter des règles d'arrêt de la prospection (par exemple en fixant un nombre maximum de sous-unités à suivre) mais il semble que ce type de protocole entraîne un biais dans les estimations des abondances. Pour que ce type de protocole soit valable, il faut donc normalement aller jusqu'au bout de la démarche, y compris en suivant toutes les sous-unités sélectionnées initialement.

#### 1.4.Stratégie générale pour construire un protocole

Commençons par un point essentiel ; celui de la **construction de la question** qui vous intéresse. Cette étape est souvent un préalable même à la mise en route d'une réflexion sur la nécessité d'un suivi. Pourtant, le fait qu'elle précède souvent la mise en œuvre d'un protocole ne signifie pas nécessairement que cette étape est un compartiment indépendant de la construction en soit du protocole. En effet, dans les différentes étapes que nous verrons ci-dessous, la définition précise de la question d'intérêt est toujours présente en toile de fond. Ceci signifie aussi que cette question doit toujours rester présente à l'esprit. Un protocole est toujours construit pour répondre à une question donnée. Il est pourtant fréquent que la volonté soit, soit de répondre à plusieurs questions identifiées avec un même protocole, soit qu'on utilise des données collectées pour répondre à une question afin d'apporter des réponses à une autre question qui se serait posée ultérieurement. Ces deux cas sont à bannir car il est très rare que les résultats soient au rendez-vous. Dans le cas où vous voulez répondre à plusieurs questions avec le même protocole, le risque est clairement de ne pas porter assez d'effort sur chaque question ou que le protocole

soit un compromis désastreux entre différentes contraintes. Dans le cas où l'on veut répondre à des questions nouvelles, il est rare que le résultat soit concluant. Toujours est-il que dans les différentes étapes de mise en œuvre d'un protocole, il peut être pertinent de réexaminer la question initiale, notamment lorsque les résultats des pré-études et des tests de puissance démontrent par avance des difficultés à répondre avec certitude à la question posée initialement. **Un aller-retour nécessaire entre question-méthode-protocole-données-analyse est souvent très riche pour faire avancer la connaissance globale d'un système.**

La construction d'un protocole pertinent ne peut faire l'économie d'une bonne connaissance préalable de la biologie/écologie de l'espèce étudiée, mais aussi de sa répartition et de son abondance sur le site. Cela peut paraître paradoxal dans le sens où le protocole mis en place a justement pour objectif de connaître l'abondance de l'espèce. Pourtant, on a vu tout au long de ce document qu'il fallait disposer d'information sur la répartition spatiale pour construire un plan d'échantillonnage adapté, mais aussi sur les abondances pour déterminer quelle méthode d'analyse de données serait la plus pertinente.

**Synthétiser les informations existantes :** une des questions centrales et des plus complexes à laquelle le montage d'un protocole doit s'atteler est celle de l'effort de terrain réalisable, mais aussi sa répartition. Pour cela, la première étape est bien entendu de collecter toutes les informations disponibles sur l'espèce dans la région étudiée. Recherche bibliographique, discussions avec les spécialistes, discussions avec les naturalistes de la région sont les trois grands axes pour acquérir des connaissances générales sur le fonctionnement des populations de l'espèce étudiée d'une manière générale, mais aussi des connaissances sur sa répartition et ses abondances locales.

**Conduire des « pré-études » :** lorsque l'espèce a été peu étudiée en amont sur le site et/ou que son écologie est mal connue, il conviendra d'envisager de décaler le protocole définitif d'un an au moins, le temps d'acquérir le minimum de connaissances nécessaires à la construction d'un protocole pertinent. En fait, il s'agit de monter un premier protocole, plus ou moins léger selon les besoins dont l'étude dispose, à mener la première année pour se fixer les idées. On parle alors de « pré-étude ». Cette pré-étude peut se construire sur un protocole du même type de celui envisagé sur le plus long terme, mais il peut aussi être différent. Par exemple, nous pouvons envisager un échantillonnage avec de très nombreuses sous-unités, mais prospectées assez brièvement pour avoir une vision globale de la présence sur les sites. Au contraire, on peut envisager de construire un protocole proche de celui qui sera mené par la suite, afin d'obtenir des premières estimations et de calculer quel effort il faudra fournir en terme d'échantillonnage pour obtenir le degré de précision désiré (voir ci-dessous « test de puissance »).

**Construire des tests de puissance :** les tests de puissance sont un outil qui permet d'optimiser le temps de terrain. Leur usage est encore assez limité en écologie bien qu'ils apportent des informations essentielles dans les étapes de construction des protocoles. Le test de puissance consiste à utiliser les informations déjà disponibles sur l'espèce, soit par la bibliographie mais idéalement les données issues de pré-étude, pour simuler/générer des faux jeux de données avec des échantillons de tailles croissantes. Cette opération est répétées de très nombreuses fois (souvent 1000 fois pour chaque taille d'échantillon, voire plus) et les jeux de données ainsi construits sont analysés avec les méthodes prévues par l'étude. Les résultats de ces analyses permet d'estimer quelle sera la précision que l'on obtiendra en fonction de l'effort de terrain que l'on souhaite réaliser. Ces tests permettent donc d'optimiser le temps de terrain en évitant d'en faire trop si l'on s'aperçoit que la précision est bonne pour un échantillon de taille inférieure à ce qui était prévu. Ils permettent aussi de ne pas prévoir une taille d'échantillon trop petite pour la précision. Réalisés en amont des demandes de financements ou lors de commandes d'études, ils peuvent permettre de montrer avant même le début de l'étude que la précision ne sera par exemple pas suffisante au regard du budget alloué pour mettre en évidence les variations de densités ou d'autres paramètres.

Malheureusement la construction de tests de puissance est une opération technique complexe qui nécessite des connaissances assez poussées en programmation et en mathématique. Notre conseil pratique est donc de consulter des spécialistes qui pourront vous aider à construire ces tests, au moins sur les thématiques pour lesquels les enjeux en termes de conservation ou financiers sont importants (espèces patrimoniales, études sur le long terme, protocoles très lourds...).

## 2. Traitement de données (Analyse des résultats)

**2.1. Analyser les données régulièrement :** un aspect très important mais très rarement mis en œuvre lorsqu'on développe un suivi de population est de bien analyser les données au fur et à mesure qu'elles sont collectées. Il peut s'agir d'une analyse au cours de la saison de suivi ou d'une année sur l'autre, les deux apporteront des informations importantes notamment en l'absence de tests de puissance. Par exemple, l'analyse des données pendant la phase de collecte est souvent peu coûteuse en termes de temps (faire quelques graphiques, quelques comptages, voire quelques analyses statistiques) mais peuvent être cruciales. Ainsi il n'est pas rare de faire trop de terrain par rapport à l'objectif déterminé. On sait que l'intervalle de confiance d'une estimation va décroître de manière exponentielle négative avec l'effort, c'est-à-dire que l'ajout d'une sous-unité d'échantillonnage va réduire fortement les intervalles de confiance au début de l'étude mais de plus en plus faiblement au fur et à mesure que leur nombre total s'accroît. Il est donc parfois inutile de poursuivre les prospections et il vaudrait mieux investir le temps restant dans une étude complémentaire ou sur un tout autre sujet. Au contraire, il n'est pas rare d'arrêter l'étude un tout petit peu trop tôt. L'examen des intervalles de confiance après chaque nouvelle session de terrain est donc très riche en enseignement et permet d'optimiser ses efforts. Enfin, l'examen des données au fur et à mesure peut informer sur l'adéquation du protocole. Par exemple, en intersaison il n'est pas rare que les conditions changent sur le terrain. Un protocole qui était pertinent les premières années de l'étude peut ne plus l'être. Attendre la fin de l'étude, notamment quand elle est menée sur plusieurs années, pour analyser les données revient souvent prendre un risque inutile car le temps d'analyse des données est souvent bien moindre que celui nécessaire sur le terrain pour collecter les données.

## 2.2. Traitement statistiques usuels en écologie

### 2.2.1. Analyse de variance (ANOVA)

L'analyse de variance a pour but de comparer la moyenne des groupes formés par le ou les critères de classification soumis à l'analyse. Elle permet de répondre à la question suivante : y a-t-il de la variabilité significative parmi ces moyennes, pour chacun des critères de classification considérés ? Les données suivent généralement une distribution normale (test de Normalité). Cependant dans certains cas la normalité n'a pu être vérifiée. Il est donc indispensable de recourir à la statistique non paramétrique. En outre, la robustesse de ces tests est indiscutable. L'analyse de variance non-paramétrique peut être adoptée pour l'ensemble des traitements statistiques. De plus ce choix est conforté quand le nombre d'observations ( $\square$  5 observations par station) est faible et qu'il faut mettre en évidence toute relation monotone (croissante ou décroissante) entre les descripteurs quantitatifs. Il s'agira de déterminer si les moyennes de chaque descripteur varient significativement.

Pour comparer plusieurs modalités, le test de Kruskal-Wallis (analyse de variance à un critère) est usuellement employé. La comparaison entre deux modalités fait appel au test de Mann & Whitney. Ces deux tests sont préconisés lors des différents calculs statistiques d'analyse de variance. Le seuil critique  $\square$  au-delà duquel on observe une différence significative entre les moyennes est fixé et codé comme suit

: différence non significative = NS  $p > 0,05$  ; différence significative = \*  $p \leq 0,05$  ; \*\*  $p \leq 0,01$  ; \*\*\*  $p \leq 0,001$ .

### 2.2.2. Analyse Factorielle des Correspondances (AFC)

L'analyse factorielle des correspondances mise au point par Benzecri (1964) et Cordier (1965), s'applique aux données qualitatives et est largement utilisée dans tous les compartiments de l'écologie. En particulier, de très nombreux auteurs l'ont employée en phytoécologie, où son intérêt n'est plus à démontrer. Sans entrer dans le détail de cette méthode d'analyse multivariée, il faut rappeler simplement qu'elle a pour objet la recherche des affinités pouvant exister, dans le cas présent, au sein d'un ensemble de relevés à travers leurs cortèges floristiques. La représentation graphique de la projection des relevés et des taxons sur les axes factoriels permettra de déterminer le gradient écologique sous-tendant chacun des axes.

**-Analyse Factorielle des Correspondances sur Variables Instrumentales (AFCVI)** L'Analyse Factorielle des Correspondances sur Variables Instrumentales ou Analyse Canoniques des Correspondances ACC consiste à mettre en corrélation un ensemble de descripteurs (variables explicatives dites instrumentales) avec un groupe de variables à expliquer (tableau de contingence). Il s'agit, par exemple, de coupler un tableau de contingence espèces x relevés avec un tableau espèces x variables (les variables peuvent être quantitatives ou qualitatives). Dans le cas de variables qualitatives, le tableau espèces x variables sera disjointé en plusieurs modalités. L'AFCVI a pour but de trouver la corrélation maximale entre les combinaisons linéaires (variables canoniques) de deux groupes de descripteurs. Elle permet de reconnaître la part de variance du tableau de contingence expliqué par le tableau de variables.

### **-Analyse Factorielle des Correspondances Multivariée (AFCM)**

Une Analyse Factorielle des Correspondances Multivariée consiste à faire une AFC sur un tableau de données qualitatives disjointées en modalités.

### 2.2.3. Analyse en Composantes Principales (ACP)

L'analyse en composantes principales (ACP) est une technique qui permet de faire la synthèse de l'information contenue dans un grand nombre de variables. Les composantes principales sont de nouvelles variables, indépendantes, combinaison linéaire des variables initiales, possédant une variance maximum. Ces nouvelles variables permettent parfois d'éclairer les mécanismes intimes mis en œuvre par la genèse des données. Elles permettent aussi d'utiliser dans de meilleures conditions des techniques multivariées classiques comme la régression linéaire. Les composantes principales autorisent en outre la représentation graphique de grands tableaux de données trop complexes à décrire par les méthodes graphiques habituelles. C'est incontestablement cette dernière propriété qui est à l'origine de leur large utilisation. En toute rigueur, une ACP ne nécessite aucune condition de validité et s'applique à des variables quantitatives.

### 2.2.4. Comparaisons de pourcentage : test du $\chi^2$ de Pearson

Ce test est réalisé à partir des effectifs et n'est valable que si les effectifs calculés sont supérieurs ou égaux à cinq dans le tableau de contingence de données. Les effectifs théoriques sont obtenus en multipliant le total d'une ligne au total d'une colonne puis en divisant par le total général. Le degré de liberté se calcule par le produit du nombre de ligne moins 1 par le nombre de colonne moins 1. Pour un degré de liberté de 1, lorsque le  $\chi^2$  calculé est inférieur au seuil critique  $\chi^2(0,05) = 3,841$ , les deux pourcentages ne sont pas significativement différents. Lorsque le  $\chi^2$  calculé est supérieur à  $\chi^2$ , les deux

pourcentages sont significativement différents. Le test est d'autant plus significatif que le seuil critique  $\alpha$  est plus petit ou que le  $\chi^2$  est plus grand (Falissard 1998).