Analyse et acquisition de données

Cours de 06/10/15

**Slide 1 :** le design expérimental permet d’éviter des données inutiles

**Slide 3 :** on peut avoir des armées de sceptiques (article scientifique)

Toxicité à long terme de Rond-up sur des rats

**Slide 4 :** dans le processus scientifique, ce qu’on va chercher à établir, c’est formuler une hypothèse qui va fournir une explication à une observation

**Slide 5 :** sur base d’une observation, on construit une hypothèse. On conçoit alors un design expérimental qu’on teste via une étude pilote. On procède ensuite à l’expérience proprement parlée. L’analyse des données expérimentales, on peut avoir à recommencer la manip à cause de l’un ou l’autre phénomène. L’étude pilote permet d’éviter ce problème.

Le test statistique sur les données va valider ou réfuter l’hypothèse.

**Slide 8 :**

chablis : arbre couché par le vent

**Slide 9 :** variabilité = les choses ne sont jamais vraiment tranchées

**Slide 10 :** on a des moyennes et des médianes qui se superposent et se confondent = facteurs confondants

**Slide 12 :** facteurs confondants : quand on observe une corrélation entre A et B, implicitement on se dit qu’il y a une relation cause-effet, mais ça peut aussi être l’effet d’une variable appelée confondante qui influence A et B de la même manière. Il pourrait donc ne pas avoir de lien entre A et B, mais si ce sont les seules variables mesurées, elles masquent ce facteur confondant.

Facteurs confondant = variable invisible qui viendrait expliquer une relation qu’on a pu observer entre deux variables qu’on a mesurées

**Slide 18 :** approche mensurative : mesure effectuées dans des conditions les plus naturelles possibles

Faible inférence : comme bcp de facteurs confondants, difficile de trouver les observations réelles et pertinentes

Approche manipulative : on va faire des mesures de manière à isoler un facteur pour diminuer au max les facteurs confondants

Forte inférence : lorsqu’on trouve une conclusion, elle est bcp plus fiable que celle obtenue par approche mensurative

**Slide 23 :** gradient qui pourrait avoir une influence différente sur les témoins et traités

**Slide 24 :** notion d’indépendance spatiale et temporelle.

On s’intéresse à la biodiversité de fourmis dans un écosystème donné et on fait tous les dix mètres un carré où on compte le nombre d’espèce que l’on a. Sur axe y, on porte la population en fonction de l’espace sur x. Il y a alors une autocorrélation spatiale : des observations faites proches ne sont pas indépendantes, car elles ont bcp de chances d’être proches. (Autre exemple par mesure de la pollution)

Donc des observations spatialement proches ne sont pas indépendantes. Il faut donc prendre en compte cette autocorrélation spatiale dans ce cas.

**Slide 25 :** Si on mesure comment une population va évoluer en fonction du temps. Bien sûr, la population en temps t dépend de la population présente en t-1. C’est ce qu’on appelle l’autocorrélation temporelle. (Autre exemple : météo ; on a plus de chance de prédire le temps de demain en disant qu’il fera le même temps qu’ajd qu’en faisant une prédiction au hasard.)

**Slide 26 :** randomisation : consiste à mélanger le plus possibles les échantillons témoins et traités de telle façon à ce qu’ils soient identiques en tout excepté pour le facteur qui les différencie. On contre ainsi les effets d’une influence externe.

**Slide 28 :**

Faisabilité technique : est-ce qu’on est capable de le faire ?

**Slide 29 :** faire une étude pilote : faire la même manip que l’on veut faire à taille réduite. Cela permet de voir le temps que ça prend, de juger le calibrage, la difficulté à laquelle l’observateur peut être confronté. Ça permet aussi d’éviter les surprises.

**Slide 33 :** on constitue des blocs avec des quantités analogues de chaque condition. On bloque la variable X 🡺 on regroupe des observations pour limiter la variabilité qu’on va avoir dans la mesure.

On va avoir des ensembles à trois échantillons disposés aléatoirement.

Cela évite que même par chance on ait des groupements d’une condition regroupé ensemble.

**Slide 37 :** si on connait la variable qui pourrait influencer, on peut essayer de mesurer cette variable et l’intégrer dans l’analyse = co-variable

**Slide 42 :** plan d’échantillonnage important pour approche mensurative

Il faut faire en sorte d’échantillonner de façon aléatoire avec tirage au centre. Il faut également un échantillonnage représentatif.

**Slide 44 :** on a un paysage où il faut distribuer l’échantillonnage. Si une partie du paysage est peu représenté, il y a une chance qu’il n’y ait aucun échantillon dans cette partie. On peut donc forcer d’inclure un certain nombre de points dans cette partie de paysage, bien qu’ils soient toujours distribués de façon aléatoire.

**Slide 50 :** bien échantillonner une gamme de valeurs pour observer les relations.

**Cours 13/10/15 – Marc Dufrene**

**Partie analyses multivariées**

**Slide 4 :** quand on a une vingtaine de stations représenté par deux variables, le graphique = une cartographie qui représente les deux variables pour chaque stations.

**Slide 5 :** chaque objet peut être représenté par un point dans un espace multidimensionnelle (chaque dimension étant une variable)

**Slide 7 :** B et D ont des profils similaires.

**Slide 8 :** avec l’introduction d’une troisième dimension, on remarque que D et B sont en fait très différents.

**Slide 10 :** quand on a un tableau de contingences (fréquences) on peut le retourner. On peut donc avoir les espèces en fonction des stations au lieu d’avoir les stations positionnées en fonction de la fréquence des espèces. On retrouve évidemment les mêmes relations 🡺 représentation duale.

**Slide 12 :** les valeurs extrêmes vont déterminer la forme du nuage de points. Pour éviter la déformation à cause d’un ou deux points particuliers, on va transformer ces points.

**Slide 13 :** on peut transformer les valeurs par logarithme pour rendre la variance indépendante de la moyenne. ! vérifier quelle est la base du log par défaut !

**Slide 14 :** si les données ne sont pas fortement déséquilibrées, on peut faire la racine carrée. Cependant, il est plutôt préférable de jouer sur la base du log.

**Slide 15 :** on a bcp d’espèces qui sont rares, très peu d’espèces communes (ce sont ces espèces qui étirent le nuage de points dans un espace multidimensionnel). On a donc intérêt à faire une transformation pour obtenir une distribution de Normale

**Slide 16 :** la standardisation procure des valeurs où la moyenne = 0 et l’écart-type = 1 🡺 les variables ont donc toutes le même poids (si jamais les variables ont des unités complètement différentes, ex : de 0 à 10 ou de 0 à 1000)

**Slide 19 :** La transformation en pourcentage a complètement changé les relations entre les points !!

**Slide 20 :** hellinger : racine carrée d’une pourcentage (pas bon !)

**!!!!!! L’ordination ordonne de manière à opposer les choses les plus différentes !!!**

**Slide 26 :** groupement par station qui partage un profil plus ou moins similaire

**Slide 29 :** 1/3 de la diversité biologique provient des bois morts.

**Slide 33 :** va déterminer s’il y a bien une différence entre les forêts riche ou pauvre en bois morts (que la quantité de bois morts soit bien différente)

**Slide 38 :** on va essayer de trouver des gradients pour opposer les valeurs

**Slide 41 :** distance = racine de la somme des carrés des différences entre chaque point pour chaque variable

**Slide 43 :** similarité : va prendre en compte la double absence ou la double présence (=point commun) sur le profil total.

**Slide 46 :** Plus ils sont proches, plus l’indice est 1, différents, 0

Bray-Curtis est une distance, mais c’est comme si c’est un indice de similarité

**Slide 50 :** Sp = similarité partielle

S1-S7(ne prend pas compte des doubles absences) 🡺 on voit le poids des doubles absences

**Slide 55 :** il faut avoir des points communs pour avoir un gradient !!

**Slide 59 :** diagonale 🡺 0 si distance ; 1 si similarité

**Slide 73 :** Variance intra-groupe = distance entre les points

**Cours du 03/11/15**

**Analyse multivariée**

**Slide 6 :** le centre du nuage = point autour duquel vont tourner les nuages. On va utiliser ce point pour restructurer le nuage.

**Slide 8 :** les deux axes sont perpendiculaires pour qu’ils soient tout à fait indépendants l’un de ‘lautre (ne pas avoir d’information qui figurent sur deux axes à la fois)

**Slide 10 :** la rotation se fait pour avoir les axes horizontal et vertical. Il existe des formules pour calculer les points d’un espace à l’autre (a1 et b1 = vecteurs propres)

**Slide 12 :** sur un de mes axes, on retrouve 9/10 de l’information initiale.

**Slide 13 :** le but est de maximiser l’information sur un nombre minimal d’axes

**Slide 17 :** On a une corrélation et on peut transformer ces corrélations en angles (r = cos a, a &tant l’angle qu’on veut trouver 🡺 r peut aller de 1 à -1, donc a peut aller de 0 à 180°) On peut donc représenter les variables en dimensions pour chaque variable. Attention, ils sont dans le même plan car les trois stations sont sur le même plan.

**Slide 18 :** on fait la somme des carrés des corrélations. On peut faire une rotation de l’axe principal pour obtenir la maximisation des projections des variables sur celui-ci. Cette variance obtenue dû à l’axe principal = inertie.

**Slide 21 :** le second axe est obtenu par étant perpendiculaire au premier.

**Slide 22 :** vecteur propre = corrélation/racine carrée de valeur propre correspondante

* On passe ainsi de six dimensions à deux

**Slide 24 :** covariance : variance = somme des variances des variables ; corrélations : variance = nombre de variables

La forme du nuage de points est influencée par les variables les plus importantes. Si on éviter ce biais, il vaut mieux plutôt utiliser les corrélations. Sinon, utiliser les covariances.

**Slide 25 :** les distances entre les objets ne changent pas, malgré le fait qu’on change de repère. On conserve les relations entre les objets.

**Slide 28 :** Le Na est le plus corrélé avec l’eau

**Slide 29 :** Scale = définit qu’on travaille sur une matrice de corrélation (TRUE), ou sur les variances des données originelles (FALSE)

Eigenvalue = valeurs propres (carré de la corrélation de la variable avec l’axe = variance)

Une seule ordination de toutes les stations va résumer plus de 50% de l’information.

**Slide 30 :**

* Toutes les variances supérieures à 1 signifient que l’axe prend plus d’informations que seulement sa propre variance
* Quand on a une succession de petites valeurs proches, on trace une droite et on prend tout ce qu’il y a au-dessus
* On fait une distribution aléatoire et on utilise tout ce qui est plus grand que cette distribution (on laisse les grissini tomber par terre et se cassent en plusieurs morceaux. Statistiquement, cela va suivre une distribution aléatoire) 🡺 on veut que les variances expliquent plus que ce qu’on obtiendrait en essayant d’expliquer la distribution au hasard de grissini cassés

**Slide 34 :** en rouge, représentation des variables.

En noir, les différentes stations. À gauche, les stations sont plus riches en humidité et minéraux, tandis que les stations à droite sont avec des valeurs plus faibles

**Slide 35 :** P est pas bien expliqué par PC1 et 2.

Plus un vecteur est long plus il est expliqué par l’analyse multifactorielle. Plus il est court, plus il est expliqué par un autre axe. Plus ils sont proche de l’axe, plus il contribue à cet axe. Par exemple humidité participe aux deux, au contraire de pHeau et pHKCL qui contribue fortement à l’axe 2 et peu à l’axe 1.

**Slide 36 :** on exporte l’image, on dissocie parce que c’est groupé dans powerpoint et on peut modifier pour avoir une plus jolie image

**Slide 41 :** les points trouvés près du centre sont ceux qui sont le moins bien expliqués par les deux axes, mais pourraient être expliqués par le troisième axe qui se trouverait comme une troisième dimension

Attention : les chiffres représentés sont les variances cumulées

**Slide 42 :** scaling = 1 : conserve les distances euclidiennes

* On prend scaling 1 et on incruste le cercle de corrélation

Attention !! On ne fait pas une ordination pour voir les groupements !! Une ordination n’est jamais qu’un applatissement d’un nuage de points à multi-dimensions

**Slide 46 :** on ne travaille pas avec des objets comme variables !!

**Slide 47 :** on fait un tableau de contingence (fréquence). L’objectif est de mettre en correspondance les lignes et les colonnes, c’est-à-dire retrouver les espèces où elles sont les plus abondantes dans une station et de préférence seules

**Slide 48 :** On fait un produit de pourcentage (le pourcentage d’individu de l’espèce multiplié par le pourcentage d’individu de l’espèce sur le nombre d’individus de la station) Plus le résultat obtenu est haut, plus il y a une correspondance ligne/colonne

**Slide 50 :** chi carré = association des valeurs d’une variable avec un »e autre variable, plus c’est élévé, plus ça montre qu’on a plus d’abondance que si c’était au hasard.

AFC conserve les distances des chi carrés

**Slide 51 :** les coordonnées des stations vont dépendre des valeurs des espèces et les coordonnées des espèces vont dépendre des valeurs des stations

Les poids vont être utilisés pour recalculer un vecteur de poids. On le fait jusqu’à obtenir une stabilisation des valeurs de poids.

**Slide 54 :** par un système de convergence, on finit par trouver des axes qui sont égaux à des moyennes pondérées des abondances des espèces.

Si une espèce a la même coordonnée qu’une station, c’est qu’elle n’est présente que dans cette station

**Slide 57 :** défini par les stations : les stations sont les points les plus extérieurs

C’est le plus intéressant, car c’est rare de n’avoir qu’une espèce dans une station

**Slide 62 :** avec AFC, on observe une séparation des moyennes et une dispersion beaucoup plus faibles 🡺 détermination des niches des espèces

**Slide 63 :** l’axe 2 montre un paquet de 4 stations avec plusieurs espèces qu’on ne trouve que dans ces stations

# 1/12/15

**Les régressions**

**Slide 2 :** grand nombre de variables dans une régression (y = x1 + x2 + x3 + … + xN). Il faut essayer de choisir des variables qui contribuent significativement à la valeur de y.

**Slide 3 :** imaginons qu’on a n variables. On va commencer par construire un modèle de régression simple en fonction de la première variable et on regarde la statistique f obtenu. Puis avec la deuxième,… On regarde celle qui a donné une statistique f la plus élevée possible. Cette variable est celle qui explique en plus grande partie de la variabilité de y. On construit alors un modèle en fonction de cette variable. Puis on reprend le modèle et on va lui adjoindre une deuxième variable et on regarde la statistique f obtenue. On refait ça avec toutes les variables et on regarde quelle est la deuxième variable qui maximise le f. On construit alors un nouveau modèle qui inclut cette deuxième variable. À chaque étape, on vérifie que chacune des variables est bien significative individuellement (variabilité individuelle des variables). On refait ces actions jusqu’à ce que la variable ajoutée ne soit plus significative.

**Slide 4 :** On a un modèle avec toutes les variables et on regarde la significativité de chaque variable. On enlève ensuite la variable qui est la moins significative (statistique t la plus petite, p le plus grand). On recalcule le modèle et on recalcule les significativités individuelles. On le fait jusqu’à obtenir un modèle où toutes les variables sont significatives.

**Slide 11 :** R² n’est pas un très bon critère de qualité de la régression. Il existe donc le critère d’Akaike (AIC), beaucoup utilisée dans un procédure automatique de sélection des variables. Cela se base sur l’optimisation du critère en question, calculé avec le terme du log de la vraisemblance (ajustement entre valeurs prédites par le modèle sur les valeurs observées) et le nombre de variables. Le but est de minimiser AIC.

Ce facteur permet aussi de comparer des modèles, de calculer leur force de prédictibilité

**Slide 12 :** multi-colinéarité = corrélation entre variables indépendantes. Si dépendance, sa valeur peut changer significativement la variable qui en dépend.

Une bonne pratique est donc d’éviter d’intégrer dans un même modèle des variables avec un degré de corrélation supérieur à 0.7.

**Slide 13 :** dans un modèle de régression, on a que des variables quantitatives. L’ANOVA est un variable dépendante de variables qualitatives. Il est possible de construire des modèles mixtes, mélangeant variables qualitatives et quantitatives.

Lm = fonction pour régression

Time = pente

Diet = les différents régimes (ordonnées à l’origine)

Time :Diet est ce que la pente est significative ou pas

Pour combiner les variables qualitatives et quantitatives =

Y|variable quantitative|variables qualitatitve|ISB|ISC

| |A|0|0

| |A|0|0

| |B|1|0

| |C|0|1

Cela va donner n-1 variables qualitatives booléennes. On prend le premier niveau comme étant le niveau de référence.

La droite de régression est obtenue en combinant les différentes variables booléennes (pente pour diet = time + time:diet)

Les ordonnées à l’origine = Intercept + diet

On peut aussi faire une anova sur les mêmes données. On peut juger les significativités générales des variables, ainsi que les significativités individuelles.

**Slide 14 :** poisson : décompte borné qui vont de x à N

**Slide 15 :** la variable y (absence ou présence d’une espèce, d’une maladie,…) = 0 ou 1. De quelle manière peut-on prédire 0 ou 1 ?

**Slide 17 :** la probabilité que la valeur soit 1 est égal à l’exponentielle d’un terme sur 1+exponentielle du même terme (= constante plus variabilité multipliée à la variable plus variabilité fois la variable suivante…)

Le log de cette probabilité sur 1-la probabilité va donner le terme qui s’apparente à une régression. L’usage du log donne le nom de régression logistique.

**Hors slide :** pour juger de la qualité d’une carte, on fait une comparaison des points prédits par rapport aux points observés, puis on fait un tableau qui reprend le nombre de 0 qui ont bien été prédits, puis mal prédits, les 1 bien prédits et les mal prédits.

Puis on peut calculer la précision globale=

Spécificité = bien prédire les 0 (0 bien prédits sur le nombre de 0 totaux)

Sensibilité = Bien prédire les 1 (1 bien prédits, sur le nombre de 1 prédits total)

La courbe ROC parfait aurait un 1-specificité = à 0 et sensibilité = 1

ROC donne AUC qui donne un très bon indicateur pour les régressions logistiques

# Cours du 08/12

**Slide 2 :** plus le milieu est tolérant, plus l’espèce est abondante.

On ne peut pas avoir deux espèces différentes avec les mêmes conditions écologiques = compétition, exclusion

**Slide 3 :** R² : coefficient de détermination = au carré du coefficient de corrélation

**Slide 4 :** plus le papier est « chiffonné », plus il y a du hasard

**Slide 5 :** optimum = maximum de l’abondance par rapport à la moyenne

**Slide 6 :** polynomiale = terme au carré

**Slide 7 :** on peut faire une transformation logarithmique 🡺 parabole 🡺 régression polynomiale

**Slide 9 :** pour protéger les espèces, prédire des distributions selon les données environnementales sur le temps,

Phytosociologie : relevé des espèces d’arbres (végétales)

Papillons car faciles à récolter

Milieu très humide = marais

**Slide 17 :** la présence d’une espèce dépend du fait que ses parents étaient là 🡺 facteur historique à prendre en compte (problème d’indépendance des données)

**Slide 47 :** il sera plus vendable, qu’il incorpore un certain nombre de variables (pas nécessairement bcp, surtout les importants) 🡺 règle de parcimonie

**Slide 49 :** indice factorielle : composantes,…

**Slide 50 :** plus les flèches sont grandes, plus le plan factoriel est expliquée par la variable.

Plus la flèche est proche d’un axe, plus elle y contribue (somme des carrées de sa projection)

PCoA : analyse en composantes principales (basé sur les distances)

Y : coordonnées verticales

**Slide 53 :** mantel, permet de faire un test de comparer les matrices de distances en permutant des lignes

**Slide 57 :** analyse discriminante = permettent d’identifier les combinaisons des variables qui opposent des groupes

Attention, avec une ANOVA, on ne verrait pas de différence significative entre les groupes !!

**Slide 64 :** par rapport à un indice chimique, les indicateurs biotiques sont intéressants car l’effet cumulatif de différentes conditions négatives va éliminer des espèces d’un cours d’eau qu’on ne va plus retrouver. Ces indicateurs viennent donc en supplément des indices chimiques, car les produits peuvent disparaître, ou lorsqu’on fait la mesure on ne voit rien, mais l’incident peut s’être produit, ce qui va être indiqué grâce aux espèces présentes ou pas.

**Slide 69 :** les espèces rares n’arrive jamais à être ensemble pour créer un groupe d’espèces.

# Cours du 15/12

**Statistiques non-paramétriques**

**Slide 1 :** taille des échantillons trop petites. Les non-paramétriques sont moins puissants et moins sensibles. Par contre, ils ne font aucune hypothèse sur la distribution des observations.

**Slide 2 :** équivalent du test T de student (compare deux moyennes). C’est la position relative des observations qui est plus importante que les valeurs-mêmes de ces observations.

**Slide 3 :** on parle ici directement des populations comparé à student où on pose H0 les moyennes sont égales.

**Slide 4 :** pas assez d’observations (difficile de vérifier la normalité) et les valeurs sont très distribuées. On classe les observations dans un tableau.

**Slide 5 :** de l’observation la plus faible à la plus élevée, puis on regarde pour chaque observation le nombre d’observations inférieure dans l’autre groupe.

K1 et K2 peuvent être retrouvés dans des tableaux qui leur attribuent une valeur.

**Slide 7 :** valable pour tous les tests statistiques : échantillonnage aléatoire

Si faible différence de moyenne entre deux populations, test T sera plus capable de rejeter H0 que celui de Wilcoxon.

**Slide 9 :** appariement : observations faites sur des mêmes individus.

**Slide 11 :** on fait la différence entre les observations et on donne un rang selon l’ordre croissant des différences en valeur absolues.

**Slide 12 :** on revient donner le signe de la différence au rang.

**Slide 13 :** on prend la valeur la plus élevée entre les sommes des rangs positifs et celle des rangs négatifs en valeur absolue, ce que j’utilsie pour aller voir la correspondance dans le tableau.

**Slide 14 :** test de permutation : troisième possibilité quand les autres tests ne fonctionnent pas.

On prend les observations et on va les permuter, c’est-à-dire qu’on va changer les correspondances entre les données. (ex : si 15,2 était attribué avec une variable OJ, on peut obtenir par permutation aléatoire que OJ devienne VC ou reste OJ).

On calcule la différence et on dresse l’histogramme de cette opération réalisée un certain nombre de fois (par ex 1000). On calcule la distribution cumulée de l’histogramme pour observer si on peut rejeter H0.

Ce test enlève tout lien entre les variables.

**Slide 17 :** test de krukal-willis : équivalent non-paramétrique de l’ANOVA, travaille sur les rangs

Correction de bonferri : permet de tenir en compte les tests multiples en jouant sur le changement de seuil pour observer la différence.

**Slide 18 :** compare une fréquence attendue à une fréquence attendue.

**Slide 19 :** différence entre la fréquence observée et la fréquence théorique sur le fréquence théorique.

**Slide 20 :** ddl = nombre de degrés de liberté (nombre de catégories -1)