Acquisition et analyse de données

Hendrickx Charlotte

2019

1 Tests d'hypothèses

1.1 Inférence sur la moyenne

Quand l'écart-type est grand, les différences entre les individus de la population sont plus grandes et il y a plus d'erreurs possibles sur la moyenne. Les positions centrales sont plus probables et les positions extrêmes sont moins probables.

1.1.1 Théorême central limite

Énoncé formel
Si l'on considère une distribution de variables $(x_1, x_2, x_3, ..., x_n)$ possédant une moyenne μ et une variance σ^2. Quand le nombre d'observations n tends vers l'infini, la somme des différentes variables noté T suivera:

$$\frac{T - n\mu}{\sigma\sqrt{n}} \approx N(0,1) \tag{1}$$

La moyenne de la somme des variables T et la variance de cette somme des variables sont données respectivement par:

$$E(T) = n\mu \qquad V(T) = n\sigma^2 \tag{2}$$

La moyenne des observations et la variance seront données par:

$$E(\bar{x}) = \mu \qquad V(\bar{x}) = \frac{\sigma^2}{n} \tag{3}$$

Si l'on prends n moyennes dans une distribution, la moyenne de ces n moyennes donnera la moyenne de la population. L'écart-type de ces n moyennes donnera l'écart-type de la population divisé par la racine carrée de n: $\frac{\sigma^2}{\sqrt{n}}$. Plus la taille de l'échantillon augmente, plus la distribution des moyennes approche une distribution normale.

Quelque soit la distribution des échantillons, si l'on prends des moyennes de façon répétées, ceci satisfaira le théorême central limite.

1.1.2 Inférence sur la moyenne

Quand on a une distribution avec un écart-type donné, les positions centrales sont plus probables et les positions éloignées sont le moins probable.

On peut prédire la distribution des fréquences. Quand n est grand et la variance est connue, la moyenne théorique suit:

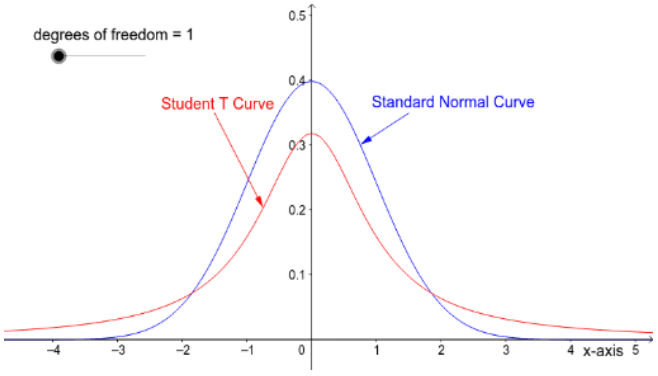
$$\mu = \bar{x} - z\frac{\sigma}{\sqrt{n}} \tag{4}$$

Quand n est petit ou la variance est inconnue, on utilise la distribution de Student. La moyenne théorique va suivre:

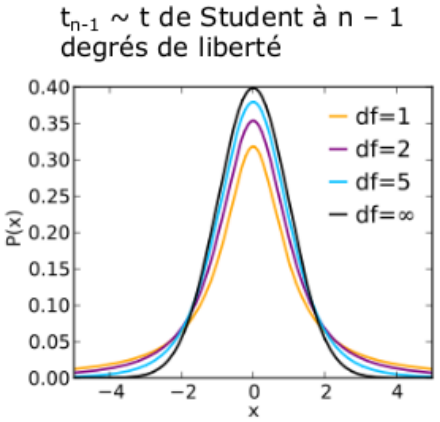
$$\mu = \bar{x} - t_{n-1}\frac{s}{\sqrt{n}} \tag{5}$$

1.1.3 Distribution normale et distribution t de Student

La distribution de Student permets donc d'analyser des échantillons avec un faible nombre d'observations et ou la variance est inconnue.



(a) Comparaison normale et student



(b) Distribution de Student à différent degrés de libertés

1.1.4 Construction de l'intervalle de confiance

On peut définir un intervalle de confiance qui corresponds à la probabilité de trouver la moyenne de la population dans un certain interval.

L'intervalle de confiance à 95% d'une distribution de Student est donnée par:

$$I.C.95\% = \bar{x} \pm t * \frac{s}{\sqrt{n}} \tag{6}$$

Ceci veut donc dire qu'on a 95% de chances de trouver la moyenne de population (μ) dans une distribution de moyenne ($\bar{x}$) entre les valeurs $\bar{x} - t * \frac{s}{\sqrt{n}}$ et $\bar{x} + t * \frac{s}{\sqrt{n}}$. Il

y a une plus grande probabilité de trouver la moyenne de la population à la moyenne estimée.

Si on augmente le nombre d'observations (n), l'interval de confiance va se resserrer. Au plus l'écart-type (s) est grand, au plus l'intervalle de confiance sera large.

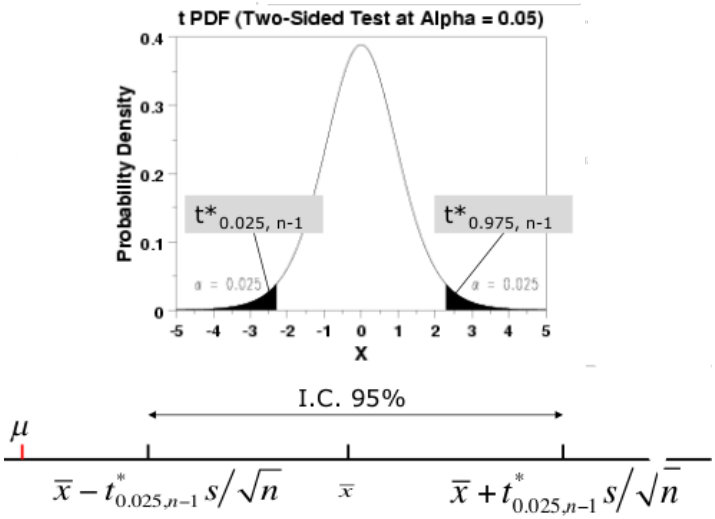


Figure 1: Interval de confiance à 95%

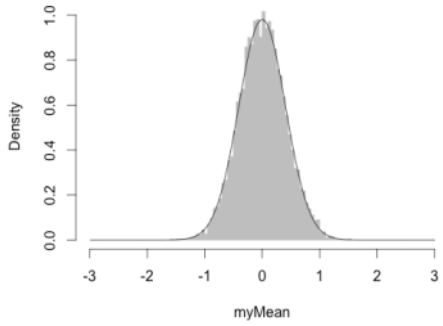
Exemple:
On examine l'effet d'un coupe-faim sur 20 volontaires et on mesure la perte de poids après 15 jours. La moyenne est de 2,6kg et l'écart-type est de 1,8.
On analyse la probabilité que si l'on a une moyenne de la population de 0, on obtienne 2,6 comme moyenne de l'échantillon par pur hasard, donc si il n'y a pas d'effet du coupe-faim.

$$\mu = \bar{x} \pm t * \frac{s}{\sqrt{n}} \tag{7}$$

$$= 2,6 \pm \frac{1,8}{\sqrt{20}} \tag{8}$$

$$= 2,6 \pm 0,402 t_{n-1} \tag{9}$$

On peut analyser ceci en générant à chaque fois 20 observations tirés d'une distribution de moyenne nulle et d'écart-type de 1,8. On analyse alors le nombre de résultats pour lequel on obtient 2,6.
Selon le théorême central limite, la distribution des moyennes suit une loi normale. On estime donc la probabilité d'avoir une valeur de plus de 2,6.



1.2 Hypothèse nulle et hypothèse alternative

1.2.1 Exemple de test t univarié

Exemple:
Si l'on reprends l'exemple du coupe-faim, l'hypothèse de base est l'H_0 donc que le coupe-faim n'a pas d'effet sur le poids. L'hypothèse alternative H_1 est au contraire que le coupe-faim a un effet sur le poids.

$$\text{Hypothèse nulle:} \quad \mu = \mu_0 = 0 \qquad\qquad H_1 : \mu \neq \mu_0 = 0 \tag{15}$$

On peut analyser ceci en supposant que H_0 soit vraie.
L'équation donnant la moyenne dans une répartition de Student est donnée par:

$$\mu = \bar{x} - t_{n-1}\frac{s}{\sqrt{n}} \tag{16}$$

Sous H_0:

$$\mu_0 = \bar{x} - t_{n-1}\frac{s}{\sqrt{n}} \tag{17}$$

$$t_{n-1} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \tag{18}$$

t_{n-1} est donc la statistique dont on connait la distribution de probabilité sous H_0. On peut alors rejeter H_0 (et donc établir un effet significatif) si la probabilité d'obtenir une valeur supérieure (ou inférieure) à $t*$ est sous un seuil α fixé.
Ceci veut donc dire qu'on accepte une certaine probabilité de se tromper (ex. si $\alpha = 5\%$, on a 5% de chances de se tromper si $t*$ est sous le seuil α et que l'on dit que l'effet est significatif).

Si la probabilité d'avoir la valeur $t*$ sous H_0 est trop faible (sous un seuil arbitrairement fixé = seuil α), on rejettra H_0 et on considèrera donc que H_1 est vraie, ce qui

reviens donc à reconnaitre un effet.

En pratique, pour faire un test d'hypothèses, les étapes à suivre sont les suivantes:

1. Représenter la distribution de la statistique étudiée telle qu'elle se présente si H_0 est vraie
2. Présenter graphiquement la distribution de la statistique sous H_0, y reporter le seuil a considéré, et préciser la zone de rejet de H_0 et celle de non rejet de H_0 (aire complémentaire).
3. Y reporter ensuite la valeur observée de la statistique pour notre échantillon.
4. Analyser si la valeur observée se trouve dans la zone de rejet de H_0 ou non et en tirer des conclusions biologiques.

Il faudra faire attention aux choses suivantes:

- L'échantillon est il aléatoire, les observations sont elles indépendantes et la distribution des données est elle normale?
- Fixer le seuil α avant de réaliser le test.

Énoncé formel
Si z_α est le quantile d'ordre α de $Z \sim N(0,1)$, nous avons $P(Z \geq z_\alpha) = 1-\alpha$ et $P(Z < \alpha) = \alpha$
Si H_0 est vraie:

$$P\Big(\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \geq z_\alpha\Big) = 1-\alpha \tag{19}$$

$$P\Big(\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} < z_\alpha\Big) = \alpha \tag{20}$$

Nous pouvons en déduire qui si $\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} < z_\alpha$ on accepte l'hypothèse nulle. Dans le cas contraire, on rejette l'hypothèse nulle. On rejette donc H_0 quand la moyenne de l'échantillon ($\bar{x}$) est trop éloigné de la moyenne de la population (μ).

- Calcul de la statistique
- Estimation de a probabilité d'obtenir une valeur supérieure à la statistique sous H_0
- Rejet ou non de l'hypothèse nulle et son interprétation biologique

1.2.2 p-value

La p-valeur est la probabilité, si H_0 est vraie, que la statistique soit au moins supérieure à la statistique observée.

$$\text{Si } p \geq \alpha \implies \text{ on ne rejette pas } H_0 \tag{21}$$

$$\text{Si } p < \alpha \implies \text{ on rejette } H_0 \tag{22}$$

Au plus la p-value est petite, au plus extrême est la moyenne de la population et donc au moins de chances on a de se tromper en faisant des prédictions à partir de nos observations.

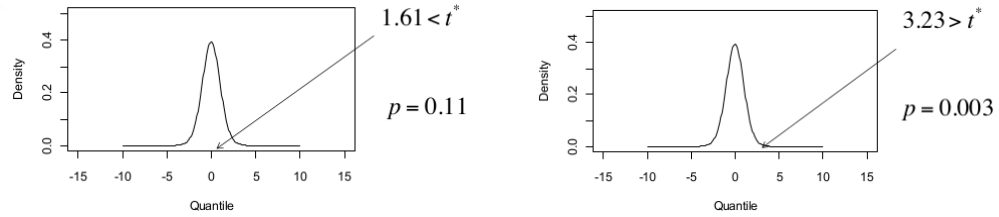
Exemple: Reprennons l'exemple du coupe-faim.
Si l'on a un échantillon de 5 personnes (n=5) avec une moyenne de prise de poids de 1,3kg et un écart-type de 1,8kg. L'hypothèse nulle étant que le changement de poids avec le coupe-faim est de 0kg ($H_0 : \ \mu_0 = 0$)

$$\frac{\bar{x}\mu_0}{\frac{\sigma}{\sqrt{n}}} > t* \tag{23}$$

$$\frac{1,3}{0,804} > t* \implies 1,61 < t* \implies p - value = 0,11 \tag{24}$$

Si l'on prends la même situation mais avec 20 sujets:

$$\frac{1,3}{0,402} > t* \implies 3,23 > t* \implies p - value = 0,003 \tag{25}$$



(a) Faible nombre d'observations (b) Haut nombre d'observations

Figure 3: Comparaison de la p-value lors d'un faible et d'un plus haut nombre d'observation

1.2.3 Risque α et β et puissance du test

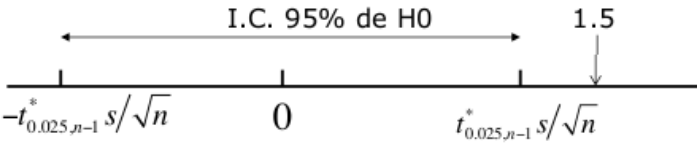
Normalement, on aimerait pouvoir soit ne pas rejeter H_0 si H_0 est vraie, soit rejeter H_0 lorce-que H_0 est fausse.
Cependant, on peut commettre deux types d'erreurs en faisant des prédictions:

→ Rejeter H_0 alors que H_0 est vraie = Trouver un effet alors qu'il n'y en a pas = Erreur de première espèce → risque α qui équivaut au seuil α

→ Ne pas rejeter H_0 quand H_0 est fausse = Ne pas détecter d'effet alors qu'il y en a = Erreur de seconde espèce → risque β

La puissance d'un test est défini comme $1-\beta$. Donc plus β est petit, plus le test est puissant. La puissance d'un test est sa capacité à détecter un effet. Au plus le nombre d'échantillons est petit, au plus difficile c'est de rejeter H_0 et donc au plus la puissance du test est faible.
La puissance est influencée par l'effet que l'on veut mesurer, par la taille de l'échantillon et par l'écart-type.

Exemple: Reprenons l'exemple du coupe-faim avec n=20, s = 1,8 et $H_0 : \mu_0 = 0$. On fixe une valeur de l'effet du coupe-faim qu'on ne voudrait pas manquer. Ici, l'on fixe 1,5kg. Ensuite, l'on estime la probabilité d'obtenir des valeurs en dehors de l'interval de confiance de H_0 .



Exemple	Faible échantillon	Faible écart-type
n = 20	n=5	n=20
s = 1,8	s= 1,8	s = 1,2
$H_0 : \mu_0 = 0$	$H_0 : \mu_0 = 0$	$H_0 : \mu_0 = 0$
puissance = 0,94	puissance = 0,30	puissance = 0,999

Table 1: Comparaison des risques β

1.2.4 Exemple de test t bivarié d'échantillon apparié

Dans les tests bivariés d'échantillons appariés, l'on teste deux traitement sur des paires d'observations. Ce test revient à étudier si il y a un changement qui diffère significativement de 0 entre les deux traitements pour chaque couple d'observations.

$$H_0 : \bar{\Delta} = 0 \qquad H_1 : \bar{\Delta} \neq 0 \tag{26}$$

On rejette H_0 si:

$$\frac{\Delta}{\frac{s}{\sqrt{n}}} > t* \tag{27}$$

1.2.5 Test t bivarié d'échantillon non-apparié

Quand les échantillons sont de même taille:

$$H_0 : \bar{x_1} = \bar{x_2} \qquad H_1 : \bar{x_1} \neq \bar{x_2} \tag{28}$$

On rejette H_0 si:

$$\frac{\bar{x_2} - \bar{x_1}}{\frac{s_{x_1 x_2}}{\sqrt{\frac{2}{n}}}} > t* \qquad \text{avec} \quad s_{x_1 x_2} = \sqrt{\frac{s_{x_1}^2 + s_{x_2}^2}{2}} \tag{29}$$

Quand les échantillons ont des tailles différentes:

$$H_0 : \bar{x_1} = \bar{x_2} \qquad H_1 : \bar{x_1} \neq \bar{x_2} \tag{30}$$

On rejette H_0 si:

$$\frac{\bar{x_2} - \bar{x_1}}{\frac{s_{x_1 x_2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}} > t* \qquad \text{avec} \quad s_{x_1 x_2} = \sqrt{\frac{(n_1 - 1)\, s_{x_1}^2 + (n_2 - 1)\, s_{x_2}^2}{n_1 + n_2 - 2}} \tag{31}$$

2 ANOVA

Les tests de Student sont limités à la comparaison de deux variables quantitatives (deux échantillons indépendants). Pour étudier les moyennes de plus de deux échantillons, l'on a recours à un ANOVA Analysis of Variance. L'ANOVA est donc une généralisation du test t.
En effet, on pourrait comparer les tests de Students de chaque couple, mais il y a un haut risque de se tromper et ceci prendrait beaucoup plus de temps.

2.1 ANOVA à un facteur

L'ANOVA peut être simplifiée sous la forme suivante ou τ_i est une constante et ou $\epsilon_{ij} \sim N(0, \sigma)$ est le résidu (l'écart entre l'observation et la moyenne du niveau correspondant)

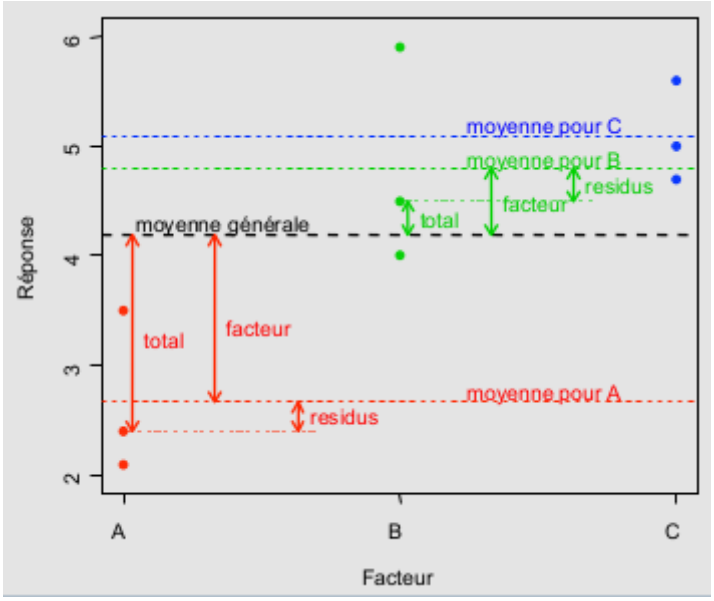
$$x_{ij} = \mu + \tau_i + \epsilon_{ij} \tag{32}$$

Les conditions d'application du test ANOVA sont:

- Échantillon aléatoire
- Observations indépendantes
- Variable réponse quantitative
- Variable explicative qualitative à 3 niveaux ou plus
- Distribution normale des résidus
- Homoscédasticité: au sein d'un groupe, les variances (donc les écarts à la moyenne) sont les mêmes

Dans un ANOVA, l'hypothèse nulle est que toutes les moyennes sont toutes égales et l'hypothèse alternative serait qu'au moins une moyenne diffère des autres.

Pratiquement, si l'on prends l'exemple de trois groupes. L'on fait d'abord une moyenne générale et une moyenne pour chaque groupe. Pour chaque observation, l'on décompose l'écart à la moyenne total en écart moyenne générale-moyenne groupe (variance facteur) et écart entre les observations-moyenne de groupe (variance résidus).
Les résidus sont les écarts entre la moyenne de chaque groupe et l'observation.



On décompose donc la variance. Les différentes parties: variance facteur et variance résidus se calculent par la somme des carrés des distances divisées par les degrés de liberté.
Le calcul de la somme des carrés des écarts se fait comme suit avec p le nombre de niveaux du facteur, n_i le nombre d'observations i au sein du facteur et n le nombre total d'observations:

$$SCE_{facteur} = \sum_{i=1}^{p} n_i(\bar{x}_i - \bar{x})^2 \qquad SCE_{residu} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_i^j - \bar{x}_i)^2 \tag{33}$$

La construction d'un tableau de l'ANOVA se fait comme suit. Le nombre de degrés de liberté de la variable facteur pour $SCE_{facteur}$ est de $p-1$ et les degrés de libertés des observations pour $SCE_{residus}$ est de $n-p$.

Type	ddl	Somme carrés	Carré moyen	Statistique F_{obs}	P (>F)
Inter (facteur)	p-1	$SCE_{facteur}$	$SCE_{facteur}/ddl_{facteur}$	$CM_{facteur}/CM_{residus}$	...
Intra (résidus)	n-p	SCE_{residu}	$SCE_{residus}/ddl_{residus}$		

2.1.1 Fisher distribution

La distribution de Fisher est une distribution asymétrique qui n'admet que des valeurs nulles ou positives. La zone de rejet est placé à droite et a une aire égale au seuil α.

On peut directement mettre la valeur de F_{obs} sur le graphe pour voir ou l'on se trouve sur le graphique.

Cette distribution de Fisher admets deux paramètres: les degrés de libertés du numérateur (degrés de liberté du facteur) et du dénominateur (degrés de liberté des résidus). Plus le quantile est grand, plus la part de variance facteurs est forte par rapport à la variance résidus. On suspectera donc que l'H_0 est vraie.

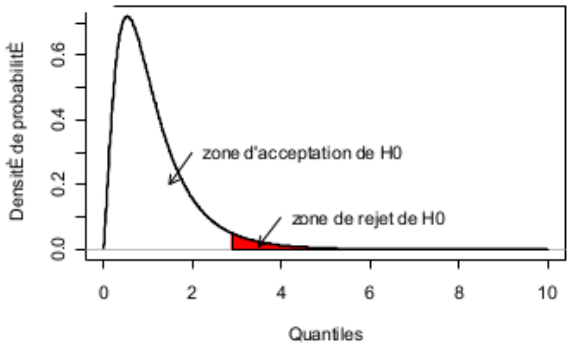


Figure 4: Fisher distribution

2.1.2 Vérification de la normalité des résidus de l'ANOVA

Les résidus sont donc la différence entre l'observation et la moyenne du groupe $residu = x_{ij} - \bar{x_i}$. On peut former un graphe quantile-quantile qui permets de comparer une distribution théorique à une distribution observée. On mets sur l'axe x les quantiles selon la distribution théorique (quantile de la loi normale de même moyenne et de même écart-type) et sur l'axe y les quantiles observés.
Si les points se distribuent selon une droite, cela veut dire que les quantiles théoriques et les quantiles observés correspondent.
Le test de Shapiro permets de tester si les résidus de l'échantillon suivent une loi normale. L'H_0 est donc que les résidus suivent une loi normale.

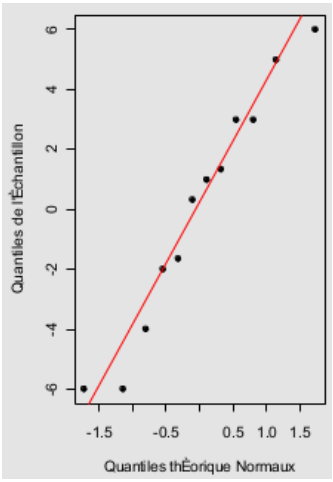


Figure 5: Graphique quantile-quantile

Il existe différents test d'homoscédasticité (dispersion des points autour de la moyenne). Les hypothèses sont: l'hypothèse nulle est l'homoscédasticité ce qui veut dire que toutes les variances sont égales. L'hypothèse alternative est l'hétérodasticité ou au moins une variance diffère.

Dans le test de Bartlett, on analyse la dispersion des points autour de la moyenne. Si l'H_0 n'est pas rejetée, on considère que les moyennes ne sont pas significativement différentes les unes des autres au seuil α. Ceci est donc ce que l'on veut. On veut donc que la p-value soit de plus de 0,05.
Si, au contraire, l'H_0 est rejetée, on doit ensuite trouver quelle moyenne diffère des autres. On doit donc faire un test de comparaison multiple des moyennes. Dans le test de Bonferroni, l'on effectue des tests de Student deux à deux en apportant une correction au seuil α. Le meilleur test est le test de Tukey. Quand l'interval de confiance dans le test de Tukey comprends 0, celui-ci n'est pas significatif car la différence entre les deux moyennes peut être de 0.

Ces tests pour voir si le test statistique est applicable peuvent aussi être fait graphiquement. Si les résultats de l'ANOVA sont distribués plus ou moins en cloche, le test est vérifié. Si il n'y a pas une courbe en cloche, il y a une forte probabilité que les résultats de l'ANOVA ne soient pas fiables.
Si les distributions des moyennes de différents groupes sont très éloignés avec une p-valeur de moins de 10^{-16}, même si le test de condition d'application ne passe pas, on garde le résultat.

2.2 ANOVA à deux facteurs

On peut réaliser un ANOVA à plusieurs facteurs et plusieurs niveaux.
Exemple:
Si l'on analyse la prise de poids d'agneaux de deux races différentes soumis à trois types de régimes. On peut alors faire un ANOVA à deux niveaux (races) et trois facteurs (traitements).

L'ANOVA à deux facteurs permets de tester deux facteurs en même temps mais permets aussi de déterminer si il y a des interactions entre les deux facteurs.

On distingue un plan balancé ou il y a le même nombre d'observations pour chaque facteur. Le plan balancé est plus puissant et plus facile à calculer. On distingue aussi un test balancé avec ou sans réplicats.
Si l'on considère un ANOVA à deux facteurs croisés sans réplicats, le manque de réplicats ne nous permets pas d'étudier les interactions entre les deux facteurs et donc on doit considérer que ces interactions n'existent pas.

$$x_{ijk} = \mu + \tau_{1i} + \tau_{2j} + \epsilon_{ijk} \tag{34}$$

Rendement de blé (tonnes/ha) : 4 variétés testés dans 3 fermes

	Variété A	Variété B	Variété C	Variété D
Ferme X	0.327	0.500	0.442	0.471
Ferme Y	0.532	0.599	0.516	0.638
Ferme Z	0.269	0.308	0.241	0.305

Type	Ddl	S. carrés	Carré moyen	Stat. F_{obs}	P (>F)
Facteur A	p_a-1	SC(A)	SC(A)/(p_a-1)	CM(A) / CM(E)	
Facteur B	p_b-1	SC(B)	SC(B)/(p_b-1)	CM(B) / CM(E)	...
Intra (résidus)	(p_a-1)(p_b-1)	SC Err.	SC Err. / (p_a-1)(p_b-1)		

Figure 6: Exemple d'ANOVA à deux facteurs croisés sans réplicats

Quand l'on a un ANOVA à deux facteurs croisés avec réplicats, on peut étudier les interactions. On a un terme $\tau_{1i}\tau_{2j}$ qui permets de quantifier les interactions pour voir si il existe un niveau d'un facteur pour lequel l'autre facteur est différent.

$$x_{ijk} = \mu + \tau_{1i} + \tau_{2j} + \tau_{1i}\tau_{2j} + \epsilon_{ijk} \tag{35}$$

Rendement de blé (tonnes/ha): 4 variétés testés dans 3 fermes : mesures avec vrais répliquas (deux champs cultivés par ferme et par variété)

	Variété A	Variété B	Variété C	Variété D
Ferme X	0.327	0.500	0.442	0.471
	0.280	0.510	0.463	0.460
Ferme Y	0.532	0.599	0.516	0.638
	0.526	0.637	0.499	0.655
Ferme Z	0.269	0.308	0.241	0.305
	0.277	0.286	0.228	0.314

Figure 7: Exemple de l'ANOVA à deux facteurs avec réplicats

On peut avoir un ANOVA à deux facteurs hierarchisés ou un facteur est imbriqué dans l'autre.
Exemple: Si l'on étudie la contamination bactérienne de différentes eaux par différents étudiants, la variable étudiant diffère et se trouve dans la variable eaux différentes vu que chaque étudiant fait trois relevés dans la même eau. Il y a donc des différences de mesure en fonction des eaux.

Eaux	« Mesure 1 »	« Mesure 2 »
Égout	(ét. A) 2700 2800 1700	(ét. B) 2600 3000 3200
Polluée	(ét. C) 52 49 61	(ét. D) 68 75 83
Propre	(ét. E) 5.9 7.6 16.0	(ét. F) 5.6 5.9 6.3

Remark: Si le test de Bartlett ne passe pas à cause d'une forte variabilité, on peut analyser le log des données ce qui peut permettre de faire un ANOVA.

3 Régression linéaire

La régression linéaire cherche à établir une relation entre une (ou plusieurs) valeurs explicatives et une variable expliquée. C'est une généralisation du modèle de l'ANOVA et permets donc d'analyser une relation linéaire entre deux variables quantitatives.

La régression linéaire nécessite deux variables quantitatives (comme la corrélation). Une des deux variables est dite variable dépendante (variable réponse) et l'autre est dite indépendante (variable explicative) (contrairement à la corrélation ou les différentes variables sont sur un pied d'égalité). On fera un nuage de points à partir de ce modèle (avec la variable indépendante sur l'axe x et la variable dépendante sur l'axe y) et on

trace ensuite une droite, la droite de régression linéaire à travers ce nuage de points.

Cas d'applications de la régression linéaire:

- Lors d'une expérience: une variable est fixée par l'expérimentateur et l'autre (la réponse) est mesurée. La variable dépendante est donc la réponse
- Lors d'observations: on mesure des paires de valeurs pour deux variables. Il est donc difficile de définir quelle variable est indépendante et quelle variable est dépendante et donc on choisis les deux selon ce qu'on veux

Exemple: Si l'on étudie les cerisiers: leur circonférence, leur hauteur et leur volume. On peut se demander si il y a des relations linéaires entre ces différentes variables et si on peut par exemple prédire le volume de bois à partir de la circonférence ou de l'hauteur de cet arbre.

3.1 Tests sur la régression linéaire

La droite de régression sera formée en essayant de minimiser la somme des carrés des résidus. Ceci est appelé une régression par les moindres carrés. Une des conditions d'applications de la régression linéaire est que les résidus ont une distribution normale de moyenne nulle et d'écart type constant = **Homoscédasticité**.

3.1.1 Significativité de la pente

On considère la droite de régression observée sur l'échantillon $Y = aX + b$ comme une estimation de la droite de régression $Y = \alpha X + \beta$ de la population. a admet une certaine distribution et la valeur de la pente a que l'on a mesuré n'est qu'une des multiples valeurs qu'elle peut adopter. On peut dire que a suit une distribution de Student de moyenne a et d'écart type donné par la formule suivante:

$$SE_a = \frac{S_{y|x}}{\sqrt{\sum(x_i - \bar{x})^2}} \qquad s_{y|x} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} \tag{36}$$

On peut alors calculer l'interval de confiance:

$$IC_{1-\alpha} = a \pm t_{\alpha/2}^{n-2} \cdot SE_a \tag{37}$$

On peut alors effectuer un test de Student et voir si la pente de la droite (a) est significative en voyant si 0 est compris dans l'interval de confiance ou non.

3.1.2 ANOVA de la régression

On peut décomposer la variance de la régression linéaire:

$$SS(total) = SS(reg) - SS(residus) \tag{38}$$

$$SS(total) = \sum (y_i - \bar{y})^2 \qquad SS(reg) = \sum (\hat{y}_i - \bar{y})^2 \qquad SS(residus) = \sum (y_i - \hat{y}_i)^2 \tag{39}$$

Le coefficient de détermination $R^2 = \frac{SS(reg)}{SS(total)}$ est la fraction de variance pouvant être expliquée par le modèle (la régression).

On peut effectuer des tests de significativité de la régression linéaire. $\frac{MS(reg)}{MS(residus)}$ suit une distribution F à 1 et $n-2$ degrés de libertés.
L'hypothèse nulle dans ce test de significativité est que la régression n'est pas significative et l'hypothèse alternative est que la régression est significative.
On fait ce test de statistique F parce-qu'avec les R^2 on peut avoir des artéfacts non-significatifs. En effet, le R^2 peut être le fruit du hasard quand il y a peu de points.

3.2 Régression linéaire multiple

La régression linéaire multiple ressemble à une ANOVA à plusieurs facteurs. La variable réponse dépends de plusieurs variables indépendantes simultanément.

La régression polynomiale est un cas particulier ou l'on fait appel à une régression multiple et ou les différents termes sont des puissances successives d'une même variable.

3.3 Conditions d'application de la régression linéaire

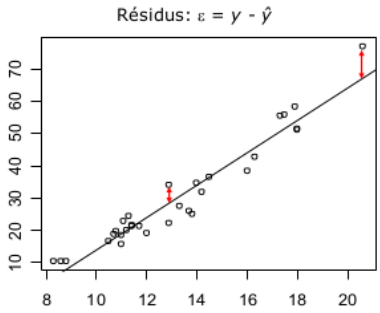
La régression linéaire est décrite par le modèle suivant:

$$Y = B_0 + B_1 X_1 + ... + B_n X_n + \varepsilon \tag{40}$$

L'erreur ε suit une loi normale de moyenne nulle et de variance σ^2. La variance de l'erreur est constante ($\rightarrow$ homoscédasticité) et de plus, l'erreur est indépendantes des mesures répliquées dans le temps et dans l'espace.
Un modèle n'est valable que dans la gamme des valeurs pour laquelle ce modèle a été fait.

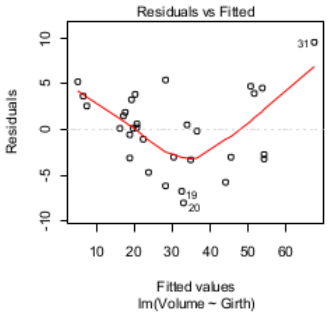
Pour pouvoir faire une régression linéaire, il faut que les résidus suivent une loi normale. Si ce n'est pas le cas, l'on peut faire une transformation pour qu'ils le deviennent.

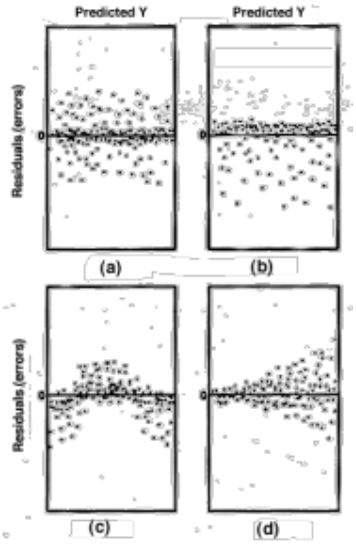
Les transformations possibles sont les puissances, les racines, logarithmes, exponentielles, inverses,...
Les différentes étapes à suivre sont:



- Vérification de la normalité des résidus par un test de Shapiro ou par un QQ-plot ou l'on montre les quantiles des résidus en fonction des quantiles dans une distribution normale. Cela devrait donner une droite.

- Les résidus doivent se distribuer plus ou moins homogènement en fonction des valeur prédites. Ceci peut être testé par un residual plot.
  La distribution des résidus peut être normale, non-linéare → on fera une régression polynomiale ou ayant des variances non-homogènes → on transforme la variable dépendante.
  Si par exemple, les résidus se distribuent sous forme de fer à cheval, l'on pourra faire une régression polynomiale.
  Dans le cas de variances non-homogènes, on a plus de variabilités quand les valeurs sont hautes. Ceci peut être corrigé en faisant une transformation logarithmique qui diminuera les écarts entre les valeurs fortes.





(b) Résidus en fonction des valeurs prédites

- Auto-corrélation spatiale: les valeurs proches

les unes des autres spatialement ou temporellement ont plus de chances d'être proche. Elles ne sont donc pas indépendantes. Pour quantifier cela, l'on peut faire des graphiques montrant les distributions de points correspondant à des valeurs prises pour des points séparés par une certaine distance. On peut alors analyser le coefficient de corrélation. Ceci nous permets alors de tracer un corrélogramme qui mettra en évidence la présence d'autocorrélations spatiaux ou temporels dans un jeu de données. Le corrélogramme aura une allure descendante si il y a une corrélation.

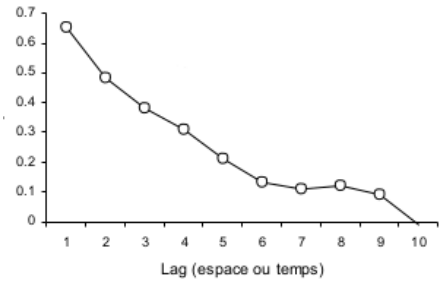


Figure 8: Corrélogramme montrant une corrélation

3.4 Choix des variables pour une régression multiple

La sélection des variables dans une régression multiple consiste à explorer dans un grans nombre de variables lesquelles ont une relation significative avec la variable indépendante.
Les différents critères à prendre en compte sont:

→ Sélection ascendante: on ajoute une à une les variables explicatives au modèle et on regarde à quel point elles expliquent la variable recherchée. On garde donc uniquement celles qui augmentent la valeur de la statistique F ou du R et on s'arrête quand aucune autre variable significative ne peut être ajoutée.

→ Sélection descendante: On mets toutes les variables dans le modèle et ensuite on les enlève une par une en commençant par les variables ayant la plus faible association avec la variable dépendante. On arrête quand toutes les variables du modèle sont significatives.

→ Principe de parcimonie: Il ne faut inclure dans le modèle que des variables biologiquement significatives. En effet, il est très probable de trouver des relations significatives par simple hasard. (ex. la génération de 500 variables aléatoires donne 21 variables significatives)

→ Multi-colinéarité: il peut y avoir des corrélations entre des variables indépendantes. On peut voir si une variable est corrélée avec une autre lorce-que quand on ajoute la nouvelle variable, le coefficient et la significativité d'une autre variable change dans le modèle.

4 Complément aux régressions

La régression linéaire permets de combiner l'utilisation de variables quantitatives avec des variables binaires ou qualitatives.

Remark: Dans R, la fonction lm() permets de travailler avec des variables qualitatives. En fait, R va remplacer les variables qualitatives par une série de variables binaires représentant chaque catégorie.

Exemple: Si l'on prends l'exemple du jeu de données ChickWeight dans lequel on étudie l'effet de différents régimes alimentaires sur la prise de poids de poussins venant du même élevage.
Il y a 4 types de traitements et l'on fait une régression linéaire pour chacun d'eux. R calcule la première droite de régression puis exprime les autres en fonction de la valeur qu'il faut ajouter à l'intercept et à la pente pour obtenir la nouvelle droite. Vu que les poussins viennent du même élevage, ils auront tous le même poids à la base et donc l'intercept sera la même. En effet, ces poulets auront le même poids au temps 0.
Un effet significatif sera montré par une pente qui diffère entre les groupes.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.9310     4.2468   7.283 1.09e-12 ***
Time          6.8418     0.3408  20.076  < 2e-16 ***
Diet2        -2.2974     7.2672  -0.316  0.75202
Diet3       -12.6807     7.2672  -1.745  0.08154 .
Diet4        -0.1389     7.2865  -0.019  0.98480
Time:Diet2    1.7673     0.5717   3.092  0.00209 **
Time:Diet3    4.5811     0.5717   8.014 6.33e-15 ***
Time:Diet4    2.8726     0.5781   4.969 8.92e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.07 on 570 degrees of freedom
Multiple R-squared:  0.773,     Adjusted R-squared:  0.7702
F-statistic: 277.3 on 7 and 570 DF,  p-value: < 2.2e-16
```

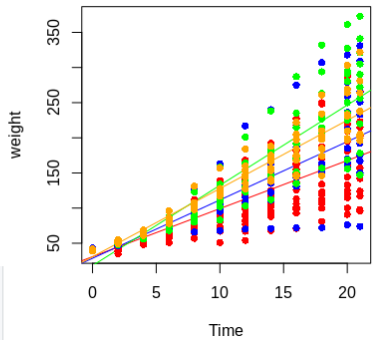


Figure 9: Régressions linéaires des différents groupes de poulets

On peut généraliser le modèle linéaire pour permettre de modéliser des modèles différents que la distribution normale. Les différentes distributions sont la binomiale, la multinomiale et la distribution de Poisson (utilisée pour compter des objets, car ils ne peuvent pas être négatifs et doivent être entiers).

Dans la régression linéaire, on a une variable dépendante suivant une loi normale ($N(\mu, \sigma)$ exprimée en fonction de valeurs explicatives. Dans la régression binomiale (également appelée logistique), la variable dépendante est binomiale (prenant une valeur de 0 ou de 1). Ceci est modélisé par une fonction logistique appelée sigmoïde. En effet, si l'on modélisait les valeurs binomiales par une droite, l'on pourrait prédire des valeurs se situant en dehors de l'interval [0,1], ce qui ne fait pas sens.
Cette fonction est décrite par les équations suivantes ou β_0 et β_1 sont les paramètres, x est la variable et $P(x)$ est la valeur prédite.

$$P(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \qquad log\Big(\frac{P(x)}{1-P(x)}\Big) = \beta_0 + \beta_1 x \tag{41}$$

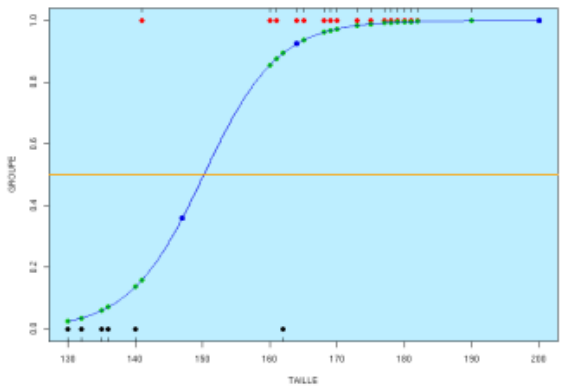


Figure 10: Fonction logistique

On optimise les paramètres β_0 et β_1 pour maximiser la vraisemblance de sorte que les valeurs prédites soient les plus proches possibles des valeurs observées.
La qualité du modèle peut être évaluée à l'aide de pseudo-R^2 ou en évaluant sa capacité à classer les évènements (présence ou absence) à l'aide de plusieurs valeurs seuil (via une courbe ROC).

4.1 Application de la régression binomiale à des cas de H5N1 au Vietnam

Nous appliqueront ce modèle de régression binomiale pour prédire les cas de H5N1 au Vietnam.
Le but est de créer une carte de répartition de la probabilité d'occurence de cas de H5N1. Ceci permets d'aider les authorités sanitaires à concentrer leurs efforts sur une certaine région.

Les données que l'on reçoit sont: la carte du territoire, les cas de H5N1 ainsi que leurs coordonnées et les facteurs de risque (ex. nombre de canards, densité de population, cours d'eau, altitude,...).

Ensuite, l'on essaye de formuler un modèle comprenant tous les facteurs de risques possibles et on regarde lesquels montrent une corrélation.
Pour chaque variable (ex. denisté de population), on prends la valeur de cette variable aux points ou les cas sont survenus et à plein d'autres points sur le territoire.
Ceci nous permets de former un graphique appelé espace des variables ou se trouvent les différents points arbitrairement définis ainsi que les points ou se trouvent les cas. On peut alors évaluer la distance qui sépare chaque point arbitraire d'un point correspondant à un positif au sein de l'espace des variables. Plus la distance entre ces deux points est courte, plus il y a de chances d'avoir un cas.

Il existe différentes méthodes pour réaliser des cartes de risque selon la définition que l'on prends de la distance. Pour sélectionner la carte prédisant avec le plus de précision la probabilité des cas, l'on a recours à un indicateur AUC. On prends des points au hasard et on regarde si le modèle permets bien de prédire les zones ou surviennent les cas. Ceci peut être évalué grâce à une courbe ROC.

Dans la courbe ROC, on a des présences et absences pour les cas et les pseudo-cas. Pour chaque point, on a une valeur prédite. On défini un seuil (ex. 0,5) et chaque valeur prédite de plus de 0,5 sera considéré comme 1 et donc comme une présence de cas.
Ensuite, on construit une matrice qui compte les valeurs observées et prédites. La sensibilité du modèle est défini comme la capacité à trouver des vrais positifs et la spécificité est la capacité à trouver des vrais négatifs.
Remark: Si l'on prends une valeur seuil très basse, la sensibilité sera maximale (tous les points seront classés comme des positifs) et la spécificité sera sera nulle. Au contraire, si l'on prends une valeur seuil très haute, la spécificité sera maximale (on ne retiendra que les cas avérés), mais la sensibilité sera nulle (très peu de cas seront prédis).

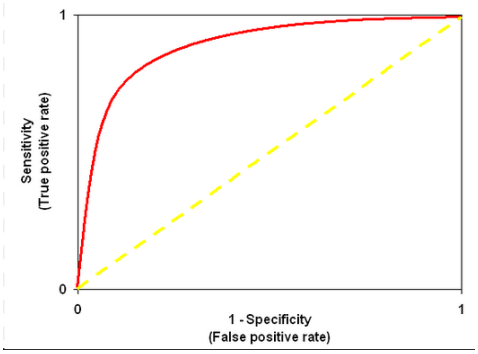


Figure 11: Courbe ROC

La courbe ROC permets de trouver la meilleure valeur seuil. Elle fait varier le seuil entre 0 et 1 et montre la spécificité et la sensibilité. Au mieux on a maximisé la spécificité et la sensibilité, au plus la courbe ROX s'éloigne de la diagonale.
Remarque: Attention aux axes de la courbe!
Il est plus facile de considérer la surface sous la courbe ROC. Ceci est appelé l'AUC et varie entre [0.5;1]. Si la valeur de l'AUC est proche de 0.5, la prédiction sera très mauvaise alors que si il s'approche plus de 1, la prédiction sera quasi-parfaite.

Dans la régression logistique, l'on regarde les variables offrant une bonne prédiction et on les applique à la carte.

Une approche permettant de former des cartes avec de l'intelligence artificielle est appelée les arbres de régression boostés.
La technique consiste en partir de l'espace des variables et le partitionner en zones contenant plus ou moins de positifs. On divise alors l'espace des variables en différentes zones selon les variables et on itère jusqu'à ce qu'il ne soit plus possible d'améliorer la prédiction.
L'avantage des arbres de régressions boostés est qu'ils nous donnent le poids de chaque variable dans la prédiction (correspondant à la variable ayant été la plus utilisée pour former le graph). De plus, il nous donne un profil décrivant la manière dont la variable influence la prédiction.

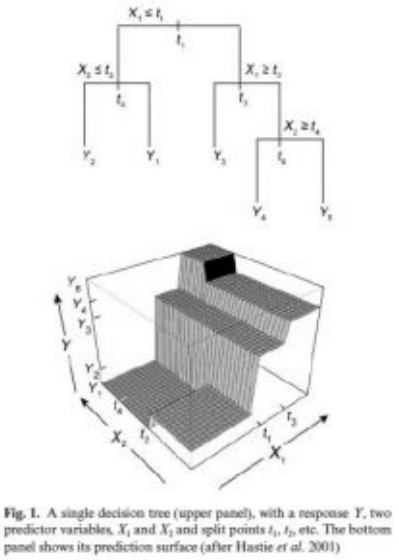


Fig. 1. A single decision tree (upper panel), with a response Y, two predictor variables, X_1 and X_2 and split points t_1, t_2, etc. The bottom panel shows its prediction surface (after Hastie *et al.* 2001)

Figure 12: Boosted regression tree

5 Tests non paramétriques

Ici, nous allons décrire certains tests non-paramètriques.

Rappels:
Les tests paramétriques supposent que les données qu'on a récoltés suivent une loi normale. On peut donc calculer les risques α et β en se basant sur la distribution théorique.
Les tests paramétriques ont une bonne capacité prédictive et sont donc très puissants, mais ils sont sensibles aux valeurs abbérantes qui rendent la distribution non normale.

Les tests non-paramétriques ne font pas d'hypothèses sur la distribution des données et donc sont applicables sur tous types de données. Ils sont peu sensibles aux valeurs extrêmes mais ne sont pas très puissants.

5.1 Test de Wilcoxon-Mann-Withney

Si les conditions d'applications du test t (distribution normale) ne sont pas respectées sur des échantillons indépendants (même après transformations), on peut appliquer le test de Wilcoxon-Mann-Withney.

Dans ce test, ce sont les positions relatives des différentes observations que l'on va analyser.

H_0 serait que **les deux populations** sont les mêmes et H_1 que les deux populations sont différentes.
On calcule la statistique W en regardant le nombre d'observations d'un groupe qui sont plus petits que la plus petite observation de l'autre groupe.
Sous H_0, il y a peu de chances qu'une des valeurs de W soit haute et l'autre basse. Si la probabilité d'avoir ces deux valeurs de statistiques sous H_0 est plus faible que le seuil α de 5%, on peut supposer que les deux populations diffèrent.

Exemple: On analyse la respiration (nombre de mol de CO_2 par g de sol et par h) des sols dans un sous-bois et une clairière. Les échantillons contiennent beaucoup de valeurs extrêmes et donc ne rentrent pas en compte pour le test t de Student.
Dans cet exemple, on a une statistique de 6.5, ce qui corresponds à une valeur critique de moins de 0.02. La différence entre les deux sols est donc significative.

Nominal Tail Probability for n = 8, n' = 7							
Bilateral	0,20	0,10	0,05	0,02	0,01	0,002	0,001
Unilateral	0,10	0,05	0,025	0,01	0,005	0,001	0,0005
Critical values	*16*	*13*	*10*	*7*	*6*	*2*	*1*

6,5

Figure 13: Comparaison de W avec les p-values

Les conditions d'applications du test W sont les mêmes que celles pour n'importe quel test:

- Échantillonage aléatoire
- Observations indépendantes
- Variable quantitative continue mais de distribution quelconque
- Il ne peut pas y avoir trop d'*ex aequos*

Le test t est très similaire au test W: ils essayent de répondre aux mêmes questions mais en effectuant des traitements différents. Dans le test t, on suppose que les données sont distribuées normalement, ce qui n'est pas le cas dans le test de Wilcoxon. Le test t n'est donc pas applicable sur toutes les distributions (contrairement au test

de Wilcoxon), mais il est plus puissant (sait mieux trouver la différence entre deux groupes) que le test de Wilcoxon. On choisira donc le test t de Student si la distribution est correcte et le test de Wilcoxon dans les autres cas.

Il existe également un test de Wilcoxon-Mann-Withney apparié qui réalise ce test sur des échantillons appariés (les données viennent des mêmes individus).
Pour ce test, on calcule la différence pour chaque individu entre les différentes mesures effectuées et on en prends la valeur absolue. On classe ensuite les observation selon celles ayant une valeur absolue de différence la plus faible et on leur assigne alors un rang. Ensuite, on compte le chiffre assigné à chaque rang positif et négatif et on en fait la somme.
On obtient ainsi deux statistiques: une pour les rangs des différences négatives et l'autre pour celle des différences positives. On garde la statistique ayant la plus haute statistique. Si on est loin de l'H_0, les deux statistiques (pour les rangs positifs et les rangs négatifs) seront éloignées: une haute et l'autre proche de 0. Au contraire, si l'on est proche de H_0, les deux statistiques seront très proches les unes des autres.

Exemple: Dans cet exemple, on a fait des relevés sur les tubes digestifs du même cheval et on regarde la densité en cellules nerveuses. On voit que la statistique du test se trouve plus haut que le seuil α de 0.05% et donc qu'il n'y a pas de différence significative entre les deux sites.

Cheval	Site I	Site II	*Différence*
1	50,6	38,0	*12,6*
2	39,2	18,6	*20,6*
3	35,2	23,2	*12,0*
4	17,0	19,0	*-2,0*
5	11,2	6,6	*4,6*
6	14,2	16,4	*-2,2*
7	24,2	14,4	*9,8*
8	37,4	37,6	*-0,2*
9	35,2	24,4	*10,8*

(a) Données

Cheval	Différence (D)	|D|	*Rang de |D|*
1	12,6	12,6	*8*
2	20,6	20,6	*9*
3	12,0	12,0	*7*
4	-2,0	2,0	*2*
5	4,6	4,6	*4*
6	-2,2	2,2	*3*
7	9,8	9,8	*5*
8	-0,2	0,2	*1*
9	10,8	10,8	*6*

(b) Rangs

Cheval	Différence (D)	Rang de |D|	*Rang signé*
1	12,6	8	*8*
2	20,6	9	*9*
3	12,0	7	*7*
4	-2,0	2	*-2*
5	4,6	4	*4*
6	-2,2	3	*-3*
7	9,8	5	*5*
8	-0,2	1	*-1*
9	10,8	6	*6*

$W_+ = 39$

$W_- = 6$

$W_s = 39$

(c) Calcul rangs signés

Nominal Tail Probability for $n = 9$							
Bilateral	0,20	0,10	0,05	0,02	0,01	0,002	0,001
Unilateral	0,10	0,05	0,025	0,01	0,005	0,001	0,0005
Critical values	*35*	*37*	*40*	*42*	*44*		

39

(d) Statistique

5.2 Test de permutation

Quand on a peu d'observations (ex. le prix des duplicats est haut) et des observations identiques, on peut faire des tests de permutation. Ceux-ci permettent de tirer des conclusions de jeux de données ayant de très mauvaises observations. Cependant, ils ont peu de puissance statistique.
Dans ces tests de permutations, on rééchantillonne aléatoirement les échantillons un grand nombre de fois (= bootstrapping) et on regarde la distribution des échantillons (correspondant à H_0). Ensuite, on compare cette distribution au hasard avec la distribution originale et on regarde la différence entre les deux. Si la valeur de la statistique calculée pour notre échantillon se trouve à plus de 95% de la distribution au hasard, le test sera significatif. Ceci s'appelle un modèle nul.
En pratique, on garde les mêmes valeurs mais on permute de multiples fois les groupes auquelles elles correspondent. Il doit cependant y avoir le même nombre d'échantillons dans chaque groupe.

On applique le test de permutation quand le nombre d'échantillons est petit, qu'on ne sait pas déterminer si les distributions suivent une distribution théorique. Quand le nombre d'échantillons est petit, il faut une haute puissance du test, ce qui n'est pas le cas dans ce type de tests. Les tests de permutations sont plus puissants que les tests non-paramétriques.

5.3 Test de Kruskal-Wallis

Ce test est l'équivalent non-paramétrique de l'ANOVA.
Les conditions d'applications du test sont:

- Échantillon aléatoire
- Obeservations indépendantes
- Variable réponse quantitative
- Au moins une variable explicative qualitative à 3 niveaux ou plus
- Pas de contraintes sur la distribution des résidus, ni d'homoscédasticité

5.4 Test du χ^2

Le test du χ^2 permets de voir si une table de contingence (contenant des probabilités) est distribuée de façon homogène. On calcule le total des nombres d'observations pour tous les groupes et on les multiplie par les fréquences théoriques d'occurence pour chaque

groupe. Ceci nous donne la distribution théorique pour chaque groupe. On compare alors cette statistique à une distribution théorique ce qui permets de déterminer si on rejette H_0 ou pas. On rejette H_0 quand le χ^2 est plus grand que la valeur théorique au seuil α choisi.
Le χ^2 est la somme des différences entre observations et théorique au carré divisés par les théoriques.

$$\chi^2 = \sum \frac{(a_i - \alpha_i)^2}{\alpha_i} \tag{42}$$

Le test du χ^2 peut être utilisé pour vérifier l'adéquation d'un échantillon à n'importe quelle distribution théorique (goodness-of-fit test).

$$\text{ddl} = \text{ catégories} - \text{ paramètres estimés} - 1 \tag{43}$$

6 Analyses multivariées: Distances et groupements

6.1 Usage des analyses multivariées

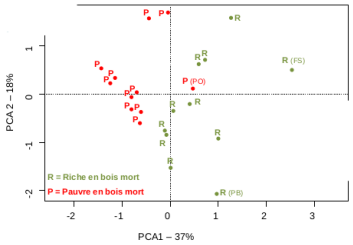
Les analyses univariées cherchent à valider des hypothèses sur des relations de dépendance entre deux variables. Elles utilisent par exemple des ANOVA ou des régressions. On formule donc une hypothèse et on teste pour savoir si on doit la rejeter ou non.
Les analyses multivariées, elles, sont généralement descriptives et identifient des structures dans des tableaux de données. Cela permets de grouper (rassembler des structures similaires) des variables ou de les ordonner (séparer les structures très différentes).

Quand l'on a un tableau de données (ex. espèces de plantes à différentes stations), on peut trier ce tableau de données selon différentes manières.

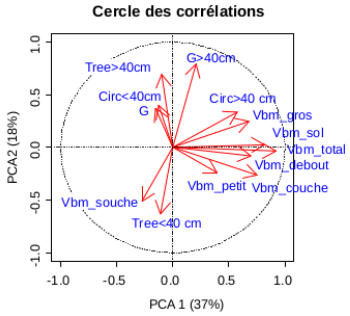
- Ordinations: on recherche des gradients dans la matrice en opposant les espèces ou les stations les plus éloignées (ex. ACP)
- Groupements: on identifie les groupes de stations ou d'espèces ayant le maximum de similarités(ex. groupements hiérarchiques)

Exemple: Si l'on cherche à savoir si la diversité des espèces d'insects dépends du volume de bois mort et de feuilles mortes. Ils on analysé des hêtraies et des chênaies réparties dans 4 régions et ont pris à chaque fois des zones riches en bois mort et d'autres pauvres en bois mort. Ensuite, ils ont récolté des données et fait des tests t pour chaque paire de variables. Ensuite, ils ont analysé les corrélations entre les variables. On peut alors réaliser une analyse en composantes principales (ACP) qui permets de résumer en quelques indicateurs un ensemble complexe d'informations quantitatives et de mettre en

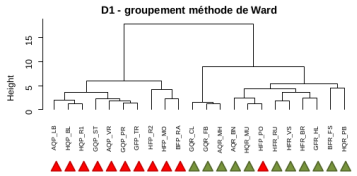
évidence les différences entre les groupes. Chaque paramètre peut aussi être représenté sous forme d'un cercle des corrélations montrant quelles variables définissent quel état. Finalement, on peut grouper les différentes stations en fonction de leur teneur en bois mort par la méthode de Ward.



(e) ACP



(f) Cercle de corrélation



(g) Groupement par méthode de Ward

6.2 Dimensions des jeux de données

Tout jeu de données peut être représenté sous forme graphique. On peut considérer un nombre illimité de variables. En effet, si nous avons par exemple trois variables (X, Y et Z), on peut les représenter chacun par un point dans un espace tridimensionnel ($\Re^3$) et on aura donc un nuage de points en trois dimensions. La même chose peut être fait avec autant de variables que l'on souhaite, mais cela ne sera pas visualisable pour nous. Donc une fois que l'on a ce nuage de points multidimensionnel, on l'écrase pour n'avoir que deux dimensions.

6.3 Transformation de données

Parfois, il est nécessaire de transformer des données pour éviter les valeurs extrêmes et les rééquilibrer ou pour réduire l'hétéroscédasticité (rendre la variance indépendante de la moyenne et plus constante).
Les différentes transformations possibles sont:

- Racine carrée: pour les données ayant un mode mais étant asymétriques
- Logarithmique: pour des données entièrement assymétriques
- Arcsinus
- Centrage
- Standardisation

6.4 Calcul de similarités

Pour mesurer les relations entre différents objets d'un tableau, on regarde la distance entre les différents objets. Les similarités sont une estimation inverse de la distance.

6.4.1 Distance euclidienne

Pour les données quantitatives, on calcule la distance euclidienne dans autant de dimmensions qu'il y a de variables.
La distance euclidienne est définie comme:

$$d = \sqrt{\sum (x_1 - x_2)^2} \tag{44}$$

Elle est adaptée aux variables écologiques mais pas aux fréquences (il ne peut pas y avoir trop de zéros). La distance euclidienne est très sensible aux valeurs élevées de certains descripteurs et il faut donc effectuer une transformation log ou racine. De plus, elle est très sensible aux différences d'unités et il faut donc effectuer une standardisation.
La distance euclidienne est très sensible aux fortes différences d'abondance. On peut remédier à ceci en faisant une transformation logarithmique qui pondère le rôle des abondances fortes.

6.4.2 Indice de simple concordance

Pour les données binaires (ex. présence ou absence d'une espèce) on utilise un indice de simple concordance.

Nombre d'espèces	Station 2		
		1 = pres	0 = abs
Station 1	1 = pres	a	b
	0 = abs	c	d

$$s = \frac{a+d}{a+b+c+d} \tag{45}$$

Le problème est que des relevés n'ayant aucune espèce en commun ($a = 0$) peuvent être assez similaires si il y a beaucoup d'absences en commun. Il faut donc éliminer les doubles absences grâce à différents indices: Jaccard, Soerensen, Steinhaud, Kulczynski, Bray-Curtis,...

6.4.3 Distance de Hellinger

Dans la distance de Hellinger, l'on calcule le pourcentage par station puis on y applique une transformation racine carrée. On peut donc avoir des différences en pourcentage qui modifieront donc les abondances entre différentes stations, même si les données brutes sont les mêmes.

6.4.4 Indice de Gower

Quand les données sont hétérogènes, on utilise l'indice de Gower.

$$g = \frac{\sum Sp}{N_{Sp}} \qquad Sp = \left(1 - \frac{|x_1 - x_2|}{max(d)}\right) \tag{46}$$

Cet indice permets de comparer des descripteurs ayant des unités différentes ou contenant des absences/présences.

De nombreux indices sont donc disponibles pour évaluer les relations entre les objets, mais ils ne représentent pas l'information de la même façon.

6.4.5 Propriétés des indices

Les indices peuvent être métriques, auquel cas ils satisfont aux conditions suivantes:

- si $a = b \rightarrow$ D(a,b) $= 0$
- si $a \neq b \rightarrow$ D(a,b) > 0
- D(a,b) $=$ D(b,a)
- D(a,b) $+$ D(b,c) $\geq$ D(a,c)

Ils peuvent aussi être semi-métriques: ils respectent alors les trois premiers points mais pas le dernier. Il peut alors y avoir des problèmes de représentations dans un espace métrique.
On dit que les indices sont euclidiens si ils satisfont à la condition suivante: $D(a,b)^2 + D(b,c)^2 = D(a,c)^2 0$

6.5 Méthodes de groupements

Le groupement consiste en fait à rechercher des patterns. Il y a plein de différentes façons de grouper les valeurs et les groupes ne sont pas les mêmes en fonction de la

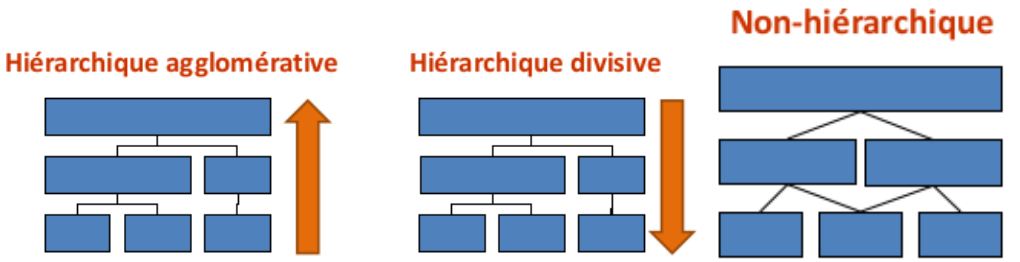
technique utilisée.

Les différentes techniques utilisées sont:

→ Hiérarchique agglomérative: on essaye de regrouper les groupes en commençant par les objets les plus similaires puis on regroupe les objets ayant moins de similarités,...

→ Hiérarchique divisive: on commence par tous les objets puis on divise en groupes selon le plus de similarités

→ Non-hiérarchique: on n'impose pas de structure hiérarchique au groupement. Les méthodes non-hiérarchiques permettent de confirmer le caractère hiérarchique des structures en les comparant.



(h) Groupement hiérarchique

(i) Groupement non-hiérarchique

Pour grouper les objets, on associe la paire de relevés le plus similaire → premier niveau de groupement. Ensuite, on associe les objets aux groupes déjà existants. On associe donc deux objets dés qu'ils ont atteint un certain degré de similarités. Ceci est appelé la méthode à liens simples. Le groupement est facile. Les liens simples sous-estiment les distances réelles.
Dans la méthode à liens complets, on regarde d'abord le taux de similarités pour tous les objets avant de les apparier. Le groupement est de plus en plus difficile mais il est plus homogène. La méthode des liens complets surestiment les distances réelles.
La dernière option est la méthode à liens moyens ou pour associer un objet avec un groupe, on attends que cet objet ait atteint la moyenne des niveaux de similarités pour tous les objets. Ce type de groupement donne le même poids aux similarités originales et suppose donc un échantillonage équilibré.
La méthode de Ward minimise la variance des distances entre les points originaux et les centres des clusters qui se forment.

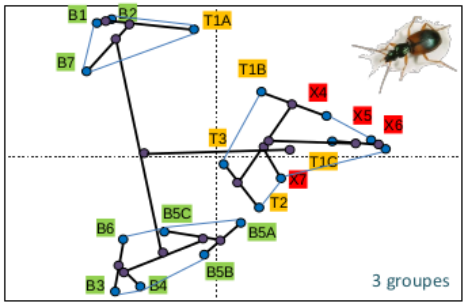


Figure 14: Groupement de Ward

Chaque méthode représente de façon différente les distances initiales. On peut donc comparer différentes manières pour trouver les groupes récurrents. Les similarités cophénétiques sont les similarités lues sur le dendogramme. Une corrélation est alores une corrélation entre les similarités cophénétiques et les similarités de la matrice de distances originale. Exam

Dans le groupement non-hiérarchique (K-means), on découpe le jeu de données en k groupes qui ont définis de manière à minimiser la somme des variances intra-groupes. Ce groupement non-hiérarchique se base sur la variance et donc elles ne se réalisent pas sur des données d'abondance d'espèces ayant beaucoup de valeurs égales à 0.
Dans le groupement non-hiérarchique, on essaye de maximiser la variance intragroupe pour donner des groupes nets. On prends 5 points au hasard qu'on va définir comme le centre des groupes. Ensuite, chaque point du graphique recherche le centre duquel il est le plus proche et va se mettre dans cette partie la du graphique. On recalcule à chaque fois le centre de chaque groupe. On recommence cette opération jusqu'à ce qu'on ne puisse plus déplacer des objets sans augmenter la variance intragroupe.

7 Analyses multivariées: Ordinations

Cette partie est exclusivement basée sur les dias. Il manque également une partie vers la fin due à ma démotivation croissante.

7.1 Méthodes d'ordination

Quand on a un jeu de données en de multiples dimensions, on peut l'ordonner. Pour ce faire, on représente les données originales en les projetant dans de nouveaux systèmes de coordonnées. Ceci permets de synthétiser les structures principales, de révéler les corrélations entre les variables originales, d'identifier les objets expliqués par certaines

variables et de visualiser la dispersion de groupes d'objets.
La méthode générale pour faire une ordination à partir d'un jeu de données est:

1. Identifier le centre du nuage de points. Le centre du nuage de points est défini comme la moyenne des variables (serait comparable à un objet fictif ayant comme données la moyenne des données pour chaque variable) et sa variance (ou inertie) comme la somme des variances des variables (moyenne des carrés des distances euclidiennes entre chaque point et le centre).

2. Centrer le nuage de points en soustrayant les moyennes aux données utilisées

3. Chercher les axes principaux (axes qui expliquent le plus la variance du nuage de points)

4. Rotation de sorte à avoir les nouveaux axes comme axes principaux. Cette rotation se fait comme suit (ou a et b sont appelés les vecteurs propres et correspondent aux paramètres de la rotation ayant été effectuée):

$$axe_1 = a\ pH + b\ CaCO_3 \qquad axe_2 = b\ pH + a\ CaCO_3 \tag{47}$$

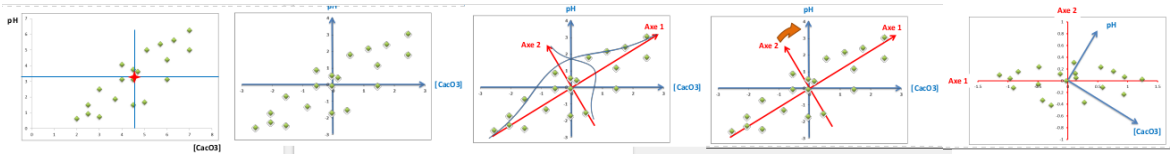


Figure 15: Étapes de l'ordination

Les méthodes d'ordinations majeures sont:

→ Analyses en composantes principales (ACP): sur des données quantitatives

→ Analyse factorielle des correspondances (AFC): sur des tableaux de fréquences

→ Analyse en coordonnées principales (ACoP): sur des matruces de distances ou de similarités

Ces méthodes permettent donc d'ordonner la matrice de données de sorte à identifier les structures principales.

Les valeurs de variance élevée ont plus de poids dans la dispersion de points que les valeurs de faible variances. Ces valeurs fortes auront donc plus de poids dans l'APC. De plus, la variance est dépendante de l'unité et donc on ne pourra pas changer les unités sans affecter le poids de la variable dans l'ACP.
Pour remédier à cela, on peut centrer et réduire les variables de sorte à ce que toutes les variables aient le même poids.

7.1.1 Analyse en composantes principales (ACP)

L'ACP se base sur le calcul de la covariance ou de la corrélation entre deux variables. Dans la covariance, on garde la variance des données originales. L'inertie totale est alors égale à la somme des différences des variables. Au contraire, dans les corrélations, toutes les variables sont centrées et réduites. L'inertie sera donc le nombre de variables.

On peut représenter les différents points comme des vecteurs dans plusieurs dimmensions (nombre de stations-1 → ex. si 3 stations → 2 dimmensions). Ces vecteurs commencent tous au point de gravité et ont la même longueur, car l'on travaille ici avec des données centrées et réduites. Les corrélations entre les différentes variables peuvent être vues comme des cosinus de triangles droits formés par ces vecteurs.

Ensuite, l'on essaye de trouver l'axe principal. Pour ce faire, on tourne autour du centre de gravité de façon à maximiser la somme des carrés des corrélations des différentes variables avec cet axe. La somme des carrés des corrélations des variables avec l'axe est appelé la valeur propre de l'axe. Elle corresponds à sa variance et est également appelé l'inertie.
On répète cette opération pour trouver le second axe principal en recherchant cette fois-ci de façon perpendiculaire au premier axe.

Pour transférer les coordonnées des objets de l'espace original à l'espace réduit, on calcule les vecteurs propres en divisant chaque corrélation par la racine carrée de la valeur propre correspondante.
L'ACP concerve les distances euclidiennes entre les objets.

Pour déterminer le nombre d'axes qu'on doit utiliser dans l'ACP, il y a plusieurs techniques possibles:

- → Moyenne des valeurs propres: on trace la moyenne des valeurs propres et on retiens les valeurs propres supérieures à cette moyenne
- → Scree plot: les dernières valeurs propres décroissent. On prends une droite qui décrit cette décroissance et on prends les valeurs propres au dessus de cette droite
- → Bâton brisé: on compare la distribution des valeurs propres à celles d'une distribution aléatoire (bâton brisé) et on ne retiens que celles dont la valeur propre est plus élevée que dans le modèle aléatoire.

(a) Moyenne des valeurs propres (b) Scree plot (c) Bâton brisé

Sur un cercle de corrélations, plus une variable est proche d'un axe, plus elle contribue. Plus un vecteur est long, plus sa variance est expliquée sur le plan factoriel.

On représente donc les variables et les objets dans un biplot. Il y a deux types de biplot:

- Distance entre les objets sur le graphique sont des approximations des distances originales → interprétation des relations entre objets
- Les angles entre les variables sont des approximations des corrélations → interprétation des relations entre variables

Les problèmes avec l'ACP sont:

- L'ACP est sensible aux données extrêmes et il vaut donc mieux travailler sur des distributions normales
- N'est pas adapté aux fréquences
- Il faut y avoir plus d'objets que de variables
- On n'utilise pas une ACP sur une matrice transposée si les unités des variables sont hétérogènes

7.2 Analyse factorielle des correspondances

Dans l'AFC, on mets en relation les lignes et les colonnes d'un tableau de données. Pour trouver ces relations, on calcule le produit des fréquences relatives dans les lignes et les colonnes. Une autre façon de montrer les relations est d'appliquer un χ^2.

On peut réaliser un calcul par itération (reciprocal averaging).

1. On crée un vecteur de poids arbitraire allant de 0 à 100 pour chaque ligne

Espèces	A	B	C	Somme	W
CALVUL	7	3	5	15	0
DESFLE	8	2	0	10	20
VACMYR	5	1	0	6	40
VACOXY	0	0	5	5	60
MOLCOE	2	7	3	12	80
ERITET	0	2	7	9	100
Somme	22	15	20	57	-

Figure 16: Exemple de tableau

2. On calcule la somme pondérée (nombre occurences espèces · poids arbitraire) pour chaque station

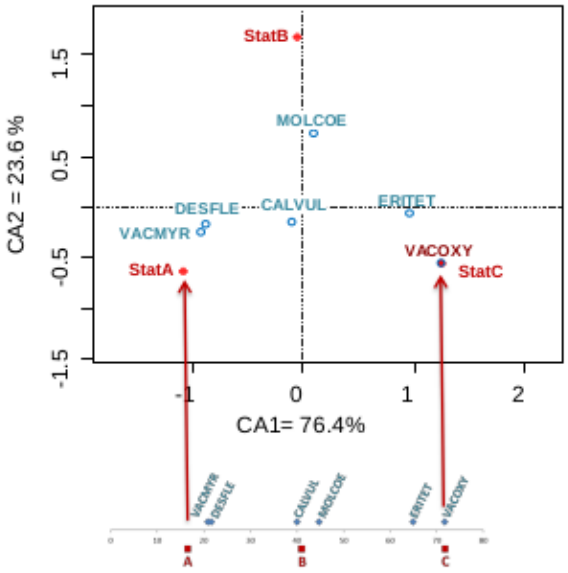
$$Q1_A = \sum W \cdot Sp_A = (0 \cdot 7) + (20 \cdot 8) + (40 \cdot 5) + (60 \cdot 0) + (80 \cdot 2) + (100 \cdot 0) = 23,64 \quad (48)$$

3. On calcule la somme pondérée pour chaque espèce qui corresponds à la somme pondérée pour chaque station · nombre d'occurences de l'espèce

$$S1_{CALVUL} = \frac{(\sum Q_1 * Sp_{CALVUL})}{\sum Sp_{CALVUL}} = \frac{(23,64 \cdot 7) + (56 \cdot 7) + (62 \cdot 7)}{15} = 42,9 \quad (49)$$

4. Le vecteur S ainsi obtenu est ré-étallé entre 0 et 100 et on recommence le calcul plusieurs fois jusqu'au moment ou on converge vers une valeur pour chaque espèce

5. Ceci nous donne la position des stations sur un axe. On refait cela pour l'autre de sorte que les coordonnées des stations soient les moyennes des coordonnées des espèces pondérées par l'abondance des espèces
   Les coordonnées des espèces sont des moyennes des coordonnées des stations pondérées par leur abondance dans ces stations
   Si une espèce à la même abondance dans les 3 stations, elle se retrouvera au centre. Plus elle est abondante dans une station, plus elle en est proche.



On peut choisir le cadrage en fonction des stations (scaling=1), des espèces (scaling=2) ou d'une combinaison des deux. (Dans cet exemple, on préfèrera un scaling =1 car il montre les espèces présentes dans les stations.

8 Récapitulatif des différents tests

	Test	**Graphique**
1 variable:	test t univarié	histogramme, boxplot, violet plot
2 variables:		
qual-qual	χ^2 d'homogénéité	boxplot
qual-quant	si 2 niveaux: test t apparié ou non	violetplot
	si $>$ 2 niveaux: ANOVA + tests *post hoc*	barplot
quant-quant	régression simple/polynomiale	nuage de points
	calcul de corrélation	
3 variables:		
1 quant $\sim$ 2 qual	ANOVA à 2 facteurs	
1 quant $\sim$ 2 quant	régression multiple	graphique 3D
1 quant $\sim$ 1 qual + 1 quant	modèle linéaire combinant	
	ANOVA + régression	
$>$ **3 variables:**		
quant $\sim$ qual/quant	modèle linéaire	
	régression multiple	
$>>>$ **3 variables:**		
analyse différentes variables	ordinations, groupements	
(pas une en fonction des autres)	mesures de distances	
$\rightarrow$ cherche a décrire la variable	mesures de similarités	

9 Design expérimental

Avant de réaliser une expérience, il vaut mieux réfléchir longuement sur le design de l'expérience. Ceci permets d'éviter d'accumuler de données inexploitables, d'avoir une interprétation de résultats le plus univoque possible, d'être capable d'exploiter les résultats négatifs,... Une expérience bien conçue et bien analysée doit avoir pensé à tout pour expliquer ces résultats et s'assurer que ceux-ci soient indémontables.

9.1 Formuler une hypothèse

Lorce-qu'on fait une expérience, c'est pour expliquer certaines observations. À partir de ces observations, l'on formule une hypothèse qui permettrait de les expliquer.

Il faut penser à quelles observations permettraient d'affirmer ou de réfuter l'hypothèse. L'on devra ensuite penser à une façon de mesurer ces variables ou à une expérimentation.

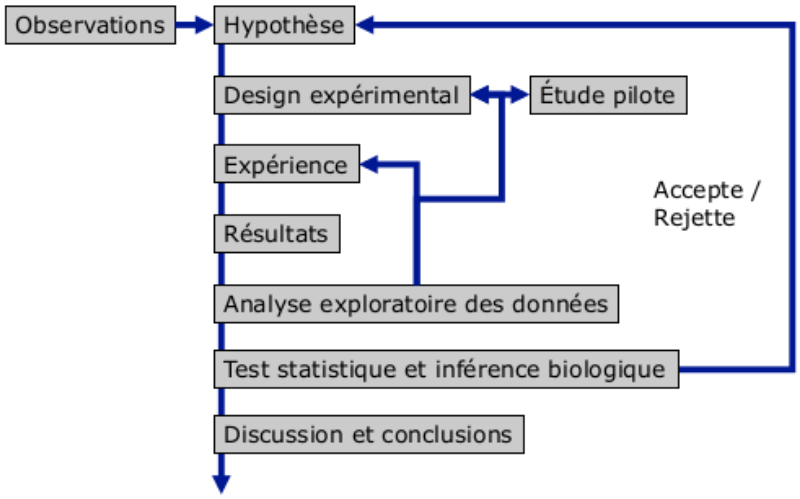


Figure 17: Procédure type du design expérimental

9.2 Design expérimental

9.2.1 Variabilité

Dans le monde biologique, il y a une grande variabilité entre les individus. Il faut en tenir compte lors des expériences ou mesures. Des réplicats sont indispensables pour chaque mesure.

9.2.2 Facteurs confondants

Lors d'une analyse biologique, on peut avoir deux variables qui semblent être corrélés. Cependant, il faut vérifier que ce sont de réelles corrélations. Parfois, il y a une variable cachée auquelle nous n'aurions pas pensé.

Exemple: L'on utilise la présence de deux insectes: la proie et le prédateur dans certains sites. On voit que dans les zones ou il y a beaucoup de prédateurs, il y a très peu de proies. → On pourrait en conclure que la présence en grand nombre de prédateurs diminue le nombre de proies. Cependant, ceci n'est pas le cas. Il y a en fait une autre variable qui influe sur les deux premières sans qu'il y ait de lien entre celles-ci. Dans cet exemple, les sites contenant beaucoup de prédateurs sont riches en pins et pauvres

en épicéas (nourriture des proies). Les pins sont un bon habitat pour les prédateurs et donc on trouvera beaucoup de prédateurs dans ces zones. La présence de pins est donc un facteur confondant.

9.2.3 Approche mensurative VS manipulative

On peut classifier les analyses biologiques en deux types: les approches mensuratives et les approches manipulatives.

Dans l'approche mensurative, les mesures sont effectuées dans des conditions naturelles. Ceci nous donne des données à large échelle avec un bon réalisme mais il y a beaucoup de facteurs confondants. Il y a également beaucoup plus de variables à prendre en compte ce qui rends difficile de direr des conclusions.

Dans l'approche manipulative, les mesures sont effectuées de façon à isoler un certain facteur et à réduire au maximum les facteurs confondants. Ces expérimentations se font à une petite échelle avec une forte inférence (estimation des paramètres d'une population à partir des paramètres d'un échantillon)) mais un problème de réalisme.

Exemple:
Les scolytes sont des insectes mangeurs de bois d'épicéa. Ils sont eux-même mangés par des prédateurs vivant dans le bois de pin. On a remarqué qu'il y avait un plus haut rapport prédateurs/proies en Scandinavie qu'en Belgique. De plus, l'on remarque que les pins ont un diamètre plus grand que les épicéas et qu'en Scandinavie, les arbres ont une épaisseur bien plus grande.
L'hypothèse est donc que l'épaisseur des épicéas belges est insuffisante pour la nymphose des prédateurs mais qu'elle est suffisante dans les pins.
L'on a donc pris des rondins d'épicéa et de pin et a augmenté leur diamètre avec du platre de cellulose. Quand l'on fait cela, l'on observe autant de nymphes dans les épicéas que dans les pins. Ceci peut alors être vérifié sur le terrain.

9.2.4 Indépendance des échantillons

Les réplicats doivent être indépendants. Il peut y avoir une variable qui influe sur les deux groupes de façon différente (ex. gradient d'ensoleillement dans une serre).

Il peut y avoir de l'auto-corrélation spatiale. Quand deux mesures sont faites proches l'une de l'autre, il y a de fortes chances pour que les valeurs obtenues soient très proches. La même chose s'applique pour le facteur temporel: des mesures faites dans un intervalle de temps proche les unes des autres seront probablement proches. Dans les cas

d'auto-corrélation, les mesures seront donc dépendantes les unes des autres.

Pour éviter d'avoir des réplicats dépendants, l'on effectue donc une randomisation. La probabilité d'appartenir à un groupe ou un autre doit être indépendante.

9.2.5 Faisabilité technique, calibrage et biais d'observateur

Lors de la conception d'une expérience, il faut tout d'abord analyser la faisabilité technique. Pour cela, il peut être utile de réaliser une étude pilote se déroulant sur quelques jours et permettant de mettre en évidence toutes les difficultés techniques.

Il faut également tenir compte des imprécisions et exactitudes des mesures. De plus, il peut y avoir des biais intra- ou inter-observateurs.
Les biais intra-observateurs sont dus à des variabilités temporelles dans les mesures réalisées par un même expérimentateur (ex. expérimentateur doit se faire la main pour chaque expérience).
Les biais inter-observateurs sont dus à des variabilités entre les expérimentateurs.

9.2.6 Approches manipulatives

Lorce-qu'on réalise des expériences manipulatives, il faut un bon témoin qui est soumis à exactement les mêmes facteurs que celle des échantillons traités, ne serais-ce pour le facteur que l'on veut isoler.

Si on a un facteur qui influence l'expérience mais qui n'est pas le facteur que l'on veut étudier, on peut faire un blocage de cette variable. On peut soit mesurer le facteur influençant, soit former différent groupes contenant les différents traitements et étant plus ou moins influencés par le facteur externe.
On peut faire des randomisations a différents niveaux et avec plusieurs variables.

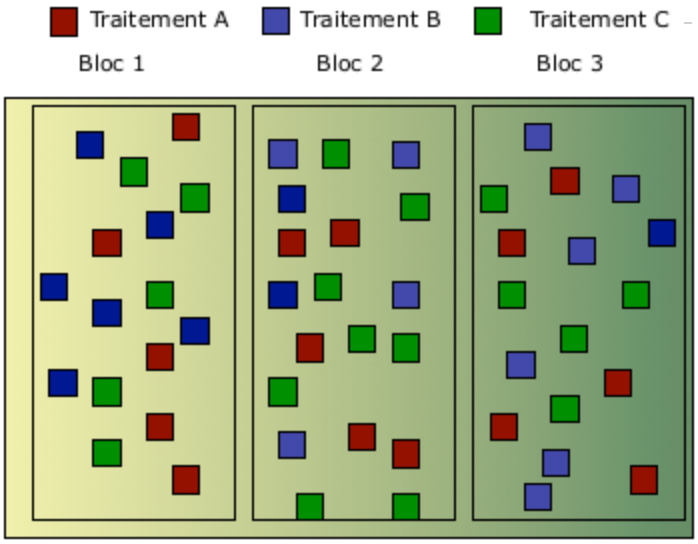


Figure 18: Plan factoriel: 1 facteur - 3 niveaux de blocage

L'inférence est une estimation des paramètres d'une population à partir des paramètres d'un échantillon. Pour pouvoir inférer, il faut que l'échantillon soit représentatif de la population. Pour assurer la représentativité de l'échantillon, il faut que l'échantillonnage soit fait correctement (ex. échantillonage aléatoire). Pour échantillonner correctement, il faut faire un échantillonage aléatoire. Si on étudie une zone ou un groupe dans lequel il y a plusieurs groupes distincs différent par une ou plusieurs variables pouvant inférer sur les résultats. Dans ces cas la, l'on fait un turage au hasard dans les différentes strates de l'échantillon.

Exemple: Si l'on a une zone ou il y a des plaines et des forets, l'on peut avoir uniquement des échantillons dans les zones de plaines si l'on prends ceux-ci réellement au hasard. On peut donc forcer la prise d'échantillons dans les plaines et les forets.

Un autre facteur à prendre en compte est le nombre d'échantillons récoltés.
Exemple: On peut prendre 50 échantillons dans une seule foret, ce qui nous donnera une excellente information sur cette foret particulière mais une très faible généralisation. Si l'on prends 5 arbres dans 10 forets, on aura peu d'informations mais sur un bon échantillon de forets.

Selon la gamme des valeurs prises par une variable, une relation peut être significative ou non. On peut parfois s'exposer à des résultats peu stables ou biaisées.

9.2.7 Quelques définitions

On distingue différents types de relations entre deux variables:

- Corrélation: variation proportionnelle ou inversément proportionnelle de deux variables observées
- Relation: idem, mais association des deux variables (un mécanisme sous-jacent est rendu implicitement responsable de cette relation)
- Causalité: la variation de la variable A est la cause de la variation de la variable B

A Paramètres de position, de dispersion et de forme

A.1 Paramètres de position

Une distruibution observée est caractérisée par les valeurs trouvées et par le nombre de fois que celles-ci sont observées.
Parfois, il peut y avoir une infinité de valeurs possiblement observées (ex. la mesure de la taille d'individus dépends de la précision et on a en fait une distribution infinie de tailles au fur et à mesure que l'on augmente la précision). Ceci peut devenir très gênant lors d'analyses. Nous pouvons donc grouper certaines valeurs similaires en classes (ex. tailles entre 1m80 et 1m85).

Etat civil	Nbre de femmes (Effectif)
Célibataire	17 364
Mariée	56 128
Veuve	11 239
Divorcée	8 170
Total	92 901

(a) Distribution observée

Classe	n_i
[140-160[	10
[160-165[	20
[165-170[	30
[170-175[	45
[175-180[	40
[180-185[	35
[185-190[	15
[190-200[	5
	N=200

(b) Distribution groupée

Figure 19: Exemples de distributions observées et groupées

A.1.1 Moyenne arithmétique

La moyenne arithmétique ($\bar{x}$) d'une série statistique se définit comme la somme des observations divisée par l'effectif n de la série.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{50}$$

La définition de la moyenne arithmétique respectivement pour une distribution observée et groupée sont données par les formules suivantes. On donne à chaque valeur obtenue x_j un effectif n_j. Pour la distribution groupée, l'on considère le centre de la classe x_{cj}

et les effectifs de la classe n_j.

$$\text{Distribution observée: } \bar{x} = \frac{1}{n}\sum_{j=1}^{J} n_j x_j \qquad \text{Distribution groupée: } \bar{x} = \frac{1}{n}\sum_{j=1}^{J} n_j x_{cj} \tag{51}$$

Remark: Dans le cas de la distribution groupée, l'on agit comme si le centre de la classe x_{cj} était apparu n_j fois. Ceci suppose que la distribution est telle que le centre de la classe peut être considéré comme la moyenne de cette classe.

A.1.2 Médiane

La médiane définit le centre d'une série statistique. Cependant, contrairement à la moyenne ou les valeurs des observations sont prises en compte, la médiane ne prends en compte que la position des observations.
À partir d'une série ordonnée, la médiane est la valeur $x_{\frac{1}{2}}$ tel que le nombre d'observations précédent $x_{\frac{1}{2}}$ est égal au nombre d'observations suivant $x_{\frac{1}{2}}$.

$$\text{Si n est impair: } x_{\frac{1}{2}} = x_{\frac{n+1}{2}} \qquad \text{Si n est pair: } x_{\frac{1}{2}} = \frac{n_{\frac{n}{2}} + n_{\frac{n}{2}+1}}{2} \tag{52}$$

La médiane est dons une valeur observée contrairement à la moyenne qui ne l'est pas toujours. De plus, contrairement à la moyenne, la médiane n'est pas influencée par des valeurs aberrantes.

A.1.3 Quantiles

Les quantiles permettent de diviser une série ordonnée en deux de tel sorte qu'il y ait une proportion p des observations inférieure ou égale à x_p et qu'il y ait une proportion complémentaire $(1-p)$ des observations supérieure à x_p.

Les exemples les plus courants de quantiles sont: la médiane ($p = \frac{1}{2}$), les quartiles ($p = \frac{1}{4}$, $p = \frac{1}{2}$, $p = \frac{3}{4}$), les déciles et les percentiles.

A.1.4 Mode

Le mode (x_M) est la valeur observée qui apparait le plus souvent. Le mode n'est pas forcément unique et parfois il n'existe pas. Dans le cas d'une distribution groupée, l'on parle de classe modale.

A.2 Paramètres de dispersion

A.2.1 Étendue ou empan

L'étendue ou l'empan est la différence entre la plus grande valeur et la plus petite valeur observée. Ce paramètre ne tient pas compte de toutes les observations et dépends entièrement des valeurs extrêmes.

A.2.2 Écarts interquartiles

Les intervals interquartiles sont définis par leurs limites qui sont respectivement les quantiles x_p et x_{1-p}. Ces intervals contiennent un pourcentage d'observations éfal à (1-2p).
L'écart interquantile est la longueur de cet intervalle: $x_p - x_{1-p}$. Ces intervals interquantiles ne contiennent pas d'éventuelles valeurs extrêmes.

A.2.3 Boxplot

On forme une boîte entre $x_{\frac{1}{4}}$ et $x_{\frac{3}{4}}$ coupé en deux parties par la médiane $x_{\frac{1}{2}}$. Elle est ensuite prolongée des deux côtés par des lignes allant jusqu'à l'observation maximale d'un côté et l'observation minimale de l'autre côté. La boîte contient donc 50% des valeurs de la série: 25% à gauche et 25% à droite. La dispersion sera plus grande quand la boîte est étendue.

A.2.4 Écart moyen absolu

On peut mesurer la dispersion d'une série en calculant la moyenne de tous ces écarts en valeur absolue. L'écart moyen absolu est donc défini par la formule suivante. L'écart moyen absolu indique donc que les observations se situent en moyenne à e_m unités de leur valeur centrale $\bar{x}$.

$$e_m = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| \tag{53}$$

A.2.5 Variance

Dans l'écart moyen absolu, on a pris la valeur absolue des différences $(x_i - \bar{x})$ ce qui permettait de ne pas tenir compte de leur signe. On peut arriver à ce même résultat en élevant ces différences au carré. Ceci est appelé variance s^2:

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{54}$$

Si un groupe est homogène, les valeurs x_i sont proches les uns des autres et donc de leur moyenne. Dans ce cas, les écarts $(x_i - \bar{x})$ sont faibles et s^2 est petit. Plus le groupe est hétérogène, plus s^2 sera grand.

La variance ne dépends pas d'un changement d'origine. En effet, les différences $(x_i - \bar{x})$ restent égales. La variance s'exprime en carré des unités utilisées pour les valeurs observées.

A.2.6 Écart-type

L'écart-type (s) est égal à la racine carrée de la variance. L'écart-type s'exprime dans les mêmes unités que les observations.

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{55}$$

On peut former une série réduite en divisant les valeurs de la série par l'écart-type (s_x). Ceci nous donne une série sans dimension dont la valeur et l'écart-type valent 1. Les valeurs centrées réduites sont définies par:

$$z_i = \frac{x_i - \bar{x}}{s_x} \tag{56}$$

Ces valeurs sont sans dimensions, de moyenne nulle et de variance 1. L'écart-type permets donc de construire des intervals centrés en $\bar{x}$. Ces intervals contiennent un certain pourcentage de la masse totale des observations.
Exemple: Lorsque la distribution est en forme de cloche et symmétrique, l'intervalle $[\bar{x} - 2s,\ \bar{x} + 2s]$ contient plus ou moins 95% des observations.

L'écart-type permet de mesurer le risque de voir une valeur observée éloignée de la moyenne. Quand s est petit, le risque est faible et quand s est élevé, le risque est élevé.

Pour comparer des distributions de moyennes très différentes, on peut calculer les coefficients de variations. Ces coefficients de variations sont une mesure relative de la dispersion et permets de donner une interprétation plus nuancée qu'on exprime en %.

$$CV = \frac{s}{\bar{x}} \tag{57}$$

B Distribution de probabilité

On peut associer une valeur à un événement et chaque événement est associé à une probabilité. On peut donc réaliser une distribution de probabilité d'une variable aléatoire discrète.

B.1 Paramètres d'une loi de probabilité discrète

L'on peut redéfinir la moyenne arithmétique pour une distribution de probabilité comme suit:

$$\mu = \sum_{j=1}^{J} p_j x_j \tag{58}$$

La variance d'une distribution de probabilité est donné par:

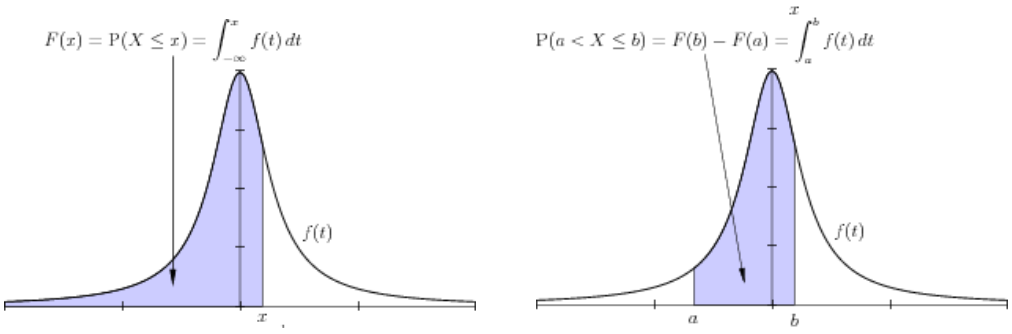
$$\sigma^2 = \sum_{j=1}^{J} p_j (x_j - \mu)^2 \tag{59}$$

B.2 Loi d'une variable aléatoire continue

Pour une fonction de probabilité, la somme des fréquences se trouvant sous cette courbe est égale à 1.

$$\sum_{j=1}^{J} f_j = 1 \tag{60}$$

Pour une distribution de probabilités, la probabilité d'appartenir à un intervalle [a,b] est $P(aXb)$ qui est la somme des probabilités dans cet intervalle et corresponds donc à $\int_a^b f(x)dx$. Cette probabilité représente donc la surface située sous cette courbe de probabilités.
Pour trouver la probabilité d'avoir une valeur plus petite que b $P(Xb)$ est représentée par la surface sous la courbe de densitée, à gauche de b.



(a) (b)

B.3 Distrubutions spéciales

B.3.1 Distribution normale = distribution de Laplace-Gauss

La distribution normale est définie par la fonction de densité:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)}} \tag{61}$$

La courbe de densité est symmétrique par rapport à $x = \mu$ et son graphe est en forme de cloche. Elle possède deux points d'inflexion distant de l'axe de symmétrie d'une quantité égale à σ.

$$XN(0,1) \tag{62}$$

Pour passer de $XN(\mu, \sigma^2)$ à $ZN(0,1)$, on doit changer d'origine et d'unité:

$$Z = \frac{X - \mu}{\sigma} \tag{63}$$

Quand on a une variable aléatoire $ZN(0,1)$, on peut rechercher un interval $[l_1,\, l_2]$ centré en 0 tel que:

$$P(l_1 Z l_2) = 1 - \alpha \tag{64}$$