Analyse statistique de données bivariées

1 Introduction

En économie et en gestion, en médecine, en biologie, mais aussi dans divers autres domaines, on cherche souvent à **étudier les liens entre différentes grandeurs :**
-relation entre un chiffre d'affaire et un budget publicitaire,
-relation entre un taux d'inflation et un taux de chomage...
-Effet de la composition de l'engrais sur le rendement d'une culture céréalière.
-Effet de différents types de traitements médicaux sur le taux de virus dans le sang chez des patients atteints d'une maladie M.
-Etude du rendement d'un paquet d'actions selon la stratégie de placement.

La statistique doit offrir des outils permettant de mesurer la nature et l'intensité des relations entre différentes variables ...

On s'intéresse donc dans ce paragraphe à l'étude simultanée de deux variables notées X et Y étudiées sur un même échantillon. On se placera dans le cas le plus fréquent, où les deux variables ne jouent pas le même rôle : **la variable Y est une variable quantitative et c'est la variable qu'on cherche à expliquer. La variable X est une variable explicative : les méthodes utilisées changent selon la nature de X (quantitative ou qualitative).**

On verra plus tard comment on étudie le lien entre deux variables qualitatives qui jouent des rôles symétriques (test d'indépendance).

Remarque : les méthodes de régression et d'analyse de variance que l'on verra dans ce chapitre se généralisent à plusieurs variables explicatives. On parle de régression multiple, et d'analyse de variance à plusieurs facteurs.

Les deux analyses régression et anova font partie de ce qu'on appelle en statistiques, le modèle linéaire.

2 Régression linéaire (variable explicative quantitative)

A. Représentation graphique des données.
On suppose maintenant que l'on a observé sur la population les valeurs prises par deux variables quantitatives. Il arrive assez souvent que ces variables ne jouent pas des rôles symétriques et que l'on s'intéresse à la manière dont l'une d'elles, que l'on note y (appelée variable réponse, effet), dépend de l'autre ,que l'on note x (variable explicative, cause). On peut par exemple étudier de quelle manière les bénéfices d'une entreprise sont influencés par son nombre d'employés, le nombre de produits vendus, le prix du carburant, le prix des taxes, ...

Une première étape dans ce genre d'étude consiste à tracer le nuage de points $(x_i, y_i)_{1\leq i\leq n}$ correspondant aux observations recueillies. Ce graphique permet de contrôler visuellement l'existence d'un lien entre les variables et d'en cerner la nature globale. Lorsqu'il existe effectivement un lien entre la variable réponse et la variable à expliquer, le nuage de points est concentré autour de la courbe correspondant à ce lien. Si aucune structure particulière n'apparaît, il semble que le lien entre les variable soit très faible ou inexistant. Par conséquent, cela n'a pas trop d'intérêt d'aller plus loin dans l'analyse. Tracer le nuage de points est

donc une étape préliminaire permettant de s'assurer qu'un lien existe et de visualisr s'il semble de nature linéaire, ce qu'il est important de faire avant de passer à la suite. Si le graphique présente une structure non linéaire, on peut proposer une transformation de la variable réponse pour s'y ramener.

EXERCICE : Imaginons que le service des études économiques d'une société cherche à mesurer l'incidence de la modulation de la pression marketting (variable explicative) sur la vente de boites de conserve (variable à expliquer).

Pour les besoins de cette étude, il procède à une expérience dans cinq zones géographiques de caractéristiques voisines (clientèle potentielle, image de marque, prédisposition des consommateurs, ...) et enregistre les ventes (en milliers de boites) réalisées dans chaque zone i durant une même période, ainsi que les dépenses (en milliers de dollards) consenties par la firme pour les budgets de publicité (locale) et de promotion des ventes. Les résultats sont répertoriés dans le tableau suivant :

ventes	25	30	35	45	65
depenses	5	6	9	12	18

1. Définir les vecteurs ventes et dépenses représentant ces données. Identifier la variable réponse et la variable explicative.
2. Tracer le nuage de points. Commenter

B. La covariance et le coefficent de corrélation linéaire.

On doit maintenant définir un indice qui rend compte numériquement de la manière dont les deux variables considérées varient simultanément.

Définition 1 *On appelle covariance empirique de x et y, le coefficient donné par la définition suivante :*

$$c_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\ \overline{y}.$$

Justification de l'appellation de covariance :
Si x et y ont tendance à varier dans le même sens, les quantités $x_i - \overline{x}$ et $y_i - \overline{y}$ seront simultanément positives (ou négatives) de sorte que leur produits seront positifs et s'ajouteront. La covariance sera donc plutôt grande et positive.
Si x et y ont tendance à varier en sens inverse, les quantités $x_i - \overline{x}$ et $y_i - \overline{y}$ seront l'une positive et l'autre négative de sorte que leur produits seront négatifs. La covariance sera donc plutôt grande en valeur absolue, mais négative.

Enfin, s'il n'y a pas de lien marqué entre les variations de x et y, les produits seront tantôt positifs tantôt négatifs, sans tendance particulière et en moyenne, par compensation la covariance sera proche de 0.

On peut remarquer que la covariance est symétrique et qu'elle est liée à la variance de la façon suivante :

$$var(x+y) = var(x) + var(y) + 2cov(x,y).$$

La covariance dépend des unités de mesure employées, ce n'est pas un indice de liaison intrinsèque. C'est pourquoi on définit le coefficient de corrélation linéaire $\rho(x,y)$:

Définition 2

$$\rho(x,y) = \frac{c_{xy}}{\sigma_x \sigma_y}.$$

Remarque : le coefficient de corrélation est compris entre -1 et 1. Ces valeurs extrèmes corrrespondent à une liaison linéaire parfaite entre les 2 variables. On a l'habitude de considérer qu'une liaison entre les variables existe si la valeur absolue du coefficient de corrélation est assez proche de 1.

C.Régression linéaire.

Lorsque deux variables sont correctement corrélées et que l'on peut à priori considérer que l'une est cause de l'autre, il est naturel de chercher la fonction de x qui approche le mieux y. On dit alors qu'on fait de la régression de y sur x.

Souvent on se restreint à des fonctions affines et on dit alors qu'on fait de la régression linéaire. On cherche alors la "meilleure" relation du type $y = ax + b$ qui puisse représenter les données. Pour mesurer la qualité de la régression, on choisit ici le critère des moindres carrés qui s'écrit comme suit :

on cherche $\hat{a}$ et $\hat{b}$ qui minimisent en a et b la quantité :

$$S(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 .$$

On trouve les coefficients

$$\hat{a} = \frac{\sum y_i x_i - n\,\overline{y}\,\overline{x}}{\sum x_i^2 - n\,\overline{x}^2} \text{ et } \hat{b} = \overline{y} - \hat{a}\overline{x}.$$

Remarque : $\hat{a}$ peut s'écrire $\frac{c_{xy}}{v_x}$

Remarque : la droite define par l'équation $y = \hat{a}x + \hat{b}$ passe par le point central des données $(\overline{x}, \overline{y})$.

Remarque : en régression on donne aussi le F de Fisher, qui permet de rejeter la l'indépendance entre x et y.

Remarque : quelque soit la forme du nuage de points, on peut calculer les coefficients de la droite d'ajustement, mais ça ne prouve pas pour autant qu'on ait une liaison significative entre les variables. Il faut pouvoir le constater sur le nuage de points et sur le coefficient de corrélation. En fait pour juger de la qualité de la régression on s'appuie souvent sur le coefficient de détermination noté souvent R^2,r^2, ou REGRSQ dans les logiciels qui représente la part de la variation totale (la variance) expliquée par la régression. Plus cette valeur est proche de 1, meilleure est la qualité de la régression. Le coefficient de détermination est défini dans le paragraphe suivant. On peut **d'autre part montrer que ce coefficient de détermination est égal au carré du coefficient de corrélation.**

EXERCICE : suite

Application à notre exemple (on a tout divisé par 1000, on parle en milliers de boites et milliers de dollards) :

i	y_i	x_i	x_i^2	y_i^2	$x_i * y_i$
1	25	5	25	625	125
2	30	6	36	900	180
3	35	9	81	1225	315
4	45	12	144	2025	540
5	65	18	324	4225	1170
somme	200	50	610	9000	2330
moyenne	40	10	122	1800	466

$$\overline{x} = \frac{50}{5} = 10, \quad \overline{y} = \frac{200}{5} = 40.$$

$$var(x) = \frac{610}{5} - 10 * 10 = 122 - 100 = 22, \quad var(y) = 1800 - 40 * 40 = 200$$

$$c_{xy} = \frac{2330}{5} - 40 * 10 = 66, \quad \rho_{xy} = \frac{66}{4.69 * 14.14} = 0.995$$

$$\hat{a} = \frac{66}{22} = 3$$

$$\hat{b} = 40 - 3 * 10 = 10.$$

La droite recherchée a donc pour équation $\hat{y} = 3 * x + 10$.

1. Calculer le coefficient de correlation et la covarince de x et y.
2. Retrouver avec R la droite de régression donnée plus haut à partir de la covariance et de la variance.
3. Utiliser directement la commande lm, tracer la droite de régression obtenue et interpréter les résultats.
4. Quelle nombre de ventes peut on prédire à artir de la droite de régresion pour des dépenses de 10000 dollars ?

D. Analyse de la qualité de la régression.

Revenons au cas général (où le lien entre y et x n'est pas forcément affine), où la variable à expliquer est notée Y, la variable explicative est notée X, et la droite de régression est $\hat{a}x + \hat{b}$. On peut montrer que la variance totale de Y se décompose en variance expliquée par la régression et variance résiduelle grâce à l'égalité de type pythagore suivante :

$$\sum (y_i - \overline{y})^2 = \sum \left(y_i - (\hat{a}x_i + \hat{b})\right)^2 + \sum \left((\hat{a}x_i + \hat{b}) - \overline{y}\right)^2$$

On démontre cette égalité en montrant que le vecteur $(\overline{y}, \overline{y}, \cdots, \overline{y})$ est la projection orthogonale de $(y_1, \cdots, y_n)$ sur le sous-espace vectoriel de $\mathbb{R}^n$ engendré par $(1, 1, 1, \cdots, 1)$ et que $\hat{a}(x_1, \cdots, x_n) + \hat{b}(1, 1, \cdots, 1)$ est la projection orthogonale de $(y_1, \cdots, y_n)$ sur le sous-espace vectoriel de $\mathbb{R}^n$ engendré par $(1, 1, 1, \cdots, 1)$ et $(x_1, x_2, \cdots, x_n)$.

La variance résiduelle est nulle si l'ajustement est parfait et se rapproche de la variance totale quand l'ajustement est "mauvais".
La variance expliquée est égale à la variance totale quand l'ajustement est parfait et diminue quand l'ajustement est "mauvais".

Cependant ce ne sont pas de bons indicateurs de la qualité de la régression. Il faut en effet relativiser par rapport à la variance totale, car la variable y peut-être naturellement trés dispersée autour de sa moyenne tout en étant trés liée avec la variable explicative.

Pour mesurer la qualité de la régression, on utilise le coefficient de détermination noté r^2 défini par $r^2 = \frac{\sum_{i=1}^{n}\left((\hat{a}x_i + \hat{b}) - \overline{y}\right)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$ et qui détermine la part de la dispersion expliquée par la régression. On peut montrer que r^2 est aussi le carré du coefficient de corrélation empirique. Plus il est proche de 1, meilleure est la régression.

Interprétation des résultats :
Dans l'exemple précédent, $r^2 = 0.99$. On a l'habitude de dire que la régression explique 99% de la variation, ce qui est trés bon.

On peut donc s'appuyer sur le modèle trouvé $\hat{y} = 3 * x + 10$. L'ordonnée à l'origine semble vouloir dire que si l'entreprise ne fait aucun effort de publicité, elle vendra 10000 boites. puis, par dollards supplémentaire investi en promotion et publicité, elle vendra 3000 boites supplémentaires.

Cependant, prudence :
1. la stabilité de la liaison suppose celle des consommateurs et de l'environnement économique.
2. il serait risqué d'utiliser le modèle pour pouvoir prévoir les ventes correspondant à des efforts publicitaires nettement en dehors des plages de valeurs ayant servi à déterminer la droite

de régression (30 millions par exemple) des phénomènes de saturation peuvent intervenir et l'environnement concurrentiel peut-être altéré.
3. le résultat reste aléatoire. Bref un peu de bon sens ! ! Une régression linéaire doit être suivie d'une analyse statistique, et qui suppose que les résidus soient Gaussiens. Elle permet par exemple de calculer le risque encouru, quand on fait une prévision sur le nombre de boites vendues.

EXERCICE :
On fait une étude sur la croissance des poussins soumis à différents régimes alimentaires. Le fichier poussin de la librairir lycee contient les variables suivantes : poidsg (le poids en gramme), agej 'age en nombre de jours), nOpoussin (l numéro du poussin étudié), régime (le régime suivi par le poussin). On regarde l'évolution au cours du temps du poids moyen des poussins qui suivent le régime 1.

1. Sélectioner les données (poids et agej) des poussins suivant le régime 1 et les placer dans les vecteurs age et poids.
2. Créer le vecteur poidsM contenant le poids moyen par age et le vecteur Age correspondant aux différents ages considérés.
3. Tracer le nuage de points et Commenter
4. Calculer le coefficient de corrélation.
5. Donner l'équation de la droite de régression. Quel poids moyen peut on prévoir pour les poussins agés de 25 jours ?
6. Refaire les étapes 1-5 pour les poussins suivant le régime 4 et comparer les modèles et résultats obtenus.

EXERCICE : n°13 p.43 sur le polycopié de cours.

3 Analyse de variance (variable explicative qualitative)

On se place maintenant dans le cas où la variable explicative est qualitative, on parle alors souvent de facteur. Notre objectif est de mieux comprendre la manière dont la variable à expliquer y varie en fonction des k (différentes) modalités de la variable explicative qualitative x (appelée facteur). l'usage préalable des boites à moustaches pour opérer un rapide jugement graphique est vivement conseillé.

Voici quelques exemples :

- On souhaite évaluer les effets de différents traitements (facteurs explicatifs) sur le taux de virus (variable à expliquer) dans le sang chez des patients atteints d'une maladie.
- On mesure le rendement pour différentes variétés de maïs soumis à 6 types de fertilisants azotés.
- On étudie des rendements laitiers sur des vaches d'une espèce donnée en fonction du régime alimentaire (paille, foin, herbe,aliments ensilés) et de la dose (faible, forte).
- On étudie les temps de germination de différentes variétés de carottes sur 4 types de sol.
- On étudie la corrosion de différents tuyaux en fonction de la nature du sol dans lesquels ils se trouvent et du type de protection (peinture) qu'ils ont reçu.
- Des études marketting examinent régulièrement l'impact de différentes campagnes publicitaires sur les ventes de différents aliments.
- On peut examiner le rendement d'un paquet d'action en fonction de la stratégie de placement.
- On peut aussi comparer les salaires d'embauches selon les écoles d'origine.

Dans ce paragraphe, on suppose qu'on a un seul facteur (une seule variable qualitative explicative) à k modalités (on peut cependant étendre la méthode de l'ANOVA pour prendre en compte plusieurs facteurs). On souhaite par exemple tester les effets de k traitements sur une maladie mesurée par le taux de virus dans le sang. On administre donc respectivement les k traitements à $n_1, \cdots, n_k$ patients et on veut tester l'égalité des moyennes des taux de virus des populations correspondant aux différents traitements dont on n'observe

que quelques individus). Pour chaque modalité m_j de la variable explicative (i.e. type de traitement) on considère la sous-population $\mathcal{P}_j$ des individus pour lesquels le facteur prend la valeur m_j (i.e. qui ont reçu ce traitement).

De manière générale, on note dans la suite de ce chapitre k le nombre de modalités du facteur et $n_1, \ldots, n_k$ les effectifs des échantillons observés des sous-populations $\mathcal{P}_1$, ..., $\mathcal{P}_k$ correspondant aux différentes modalités du facteur. On note $\mu_1, \ldots, \mu_k$ les espérances de la variable qualitative Y sur ces différentes sous-populations. On note généralement $Y_{i,j}$, $1 \leq i \leq n_j$, $1 \leq j \leq k$ les variables aléatoires représentant la variable étudiée Y sur les différents individus de l'échantillon afin de faire apparaître la sous-population à laquelle il appartient au travers de l'indice j. On notera dans ce qui suit

$$\overline{Y}_j := \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{i,j} \text{ et } \overline{Y} := \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} Y_{i,j}.$$

On utilise souvent de la même manière un deuxième indice (on note les observations $y_{i,j}$) pour indiquer la sous-population d'où provient chaque observation. On note par exemple $y_{5,1}$ la cinquième observation correspondant à un individu de la population $\mathcal{P}_1$, c'est à dire pour lequel la modalité m_1 a été observée.

On souhaite comparer les valeurs de la variable Y observées au sein de ces différentes populations afin de cerner l'impact éventuel du facteur (variable explicative qualitative). Une première étape dans l'analyse consiste donc à séparer les observations suivant les modalités de la variable explicative puis à comparer la répartition des données dans les différentes populations en utilisant des outils que nous avons vus dans les TP précédents (boxplot, histogramme, moyenne, écart-type, ...). On va s'intéresser plus particulièrement à la comparaison des valeurs moyennes.

EXERCICE : Un exemple : un forestier s'intéresse aux hauteurs moyennes de trois forêts. Pour les estimer, il échantillonne un certain nombre d'arbres de chaque forêt et mesure leur hauteur. Les résultats se présentent sous la forme suivante :

Hauteur (m)	23,4	22,1	22,5	22,5	24	18,9	23,7	21,1	24,4	24,6	24	24,9	23,5	25	24,5	26,2
N° Forêt	1	3	3	2	2	3	2	3	1	1	2	1	3	1	3	1

La variable qualitative (numéro de forêt) a ici trois modalité, à qui on associe 3 groupes. Pour représenter graphiquement un tel tableau, on représente sur un même graphique les résultats correspondants à chacun des groupes, et on compare ainsi la dispersion des résultats obtenus, souvent à l'aide des 3 boites à moustaches placées parallèlement sur un même graphique.
On voit ainsi immédiatement, les différences de médianes et de dispersions.

1. Créer les vecteurs hauteur et forêt correspondant à ces données.
2. Comparer la répartition des tailles dans les trois forêts au travers de boxplots. Utiliser par exemple la commande split. Commentez.
3. Calculer la moyenne totale des tailles des arbres ainsi que les moyennes propres à chaque forêt. Calculer les variances associées à chaque forêt. Commentez.

Il semble qu'en moyenne la première forêt soit plus élevée que la seconde, elle même plus élevée que la troisième. La variance de la troisième est plus élevée que celle des autres. Il y a un arbre de 18,9 mètre et un de 24,5 (on avait déja remarqué ça avec les boites).
Les forêts semblent différentes, mais la variance des arbres est relativement grande et on peut se demander si les forêts sont significativement différentes et si les écarts ne sont pas dus à l'échantillonnage.

Pour aller plus loin dans l'analyse, on réalise une analyse de variance ou ANOVA dont l'objectif est de détecter des différences significatives entre les valeurs moyennes prises par y sur les différentes sous-populations (apelées aussi classes). En d'autres termes on se pose la question suivante : **la variabilité**

observée dans les données est-elle uniquement due au hasard, ou bien existe-t-il effectivement des différences significatives entre les classes imputables au facteur. Pour cela, on va comparer **la variance intraclasse (ou variance résiduelle)** qui résume la variabilité à l'intérieur des classes et la **variance interclasse (ou variance des moyennes)** qui décrit les différences entre les classes.

L'analyse de variance permet de mesurer les effets d'une ou plusieurs variables qualitatives appelées aussi facteurs sur une variable quantitative Y.

Ce test vise à établir la possibilité de rejeter l'hypothèse H_0 ($\mu_1 = \cdots = \mu_k$) d'égalité simultanée des moyennes correspondant aux sous-populations formées par les différentes modalités des facteurs.
On note $\overline{Y}$ la moyenne empirique calculée sur la totalité de la population et $\overline{Y}_j$ la moyenne empirique correspondant à la population $\mathcal{P}_j$ (dont on a observé n_j individus). Les écarts à la moyenne se décomposent de la façon suivante :

$$Y_{ij} - \overline{Y} = (Y_{ij} - \overline{Y}_j) + (\overline{Y}_j - \overline{Y})$$

Remarque : On écrit souvent le modèle sous la forme suivante

$$\forall j \in \{1, \ldots, k\}\,, \forall i \in \{1, \ldots, n_j\}\,, Y_{ij} = \overline{Y} + a_j + \epsilon_{ij},$$

dans lequel $a_j = \overline{Y}_j - \overline{Y}$ représente l'effet de la modalité j du facteur sur la valeur moyenne et les variables ϵ (appelées résidus) représentent l'aléa qui n'est pas expliqué par les moyennes de chaque classe.

Le point clé de la méthode est la décomposition suivante de la variance totale (provenant d'une identité de type pythagore)

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2 + \sum_{j=1}^{k} n_j(\overline{Y}_j - \overline{Y})^2.$$

qui s'écrit aussi
SCT(somme des carrés totaux) (variance totale)= SCR(sommes des carrés résiduels) (variance intraclasse)+ SCE(somme des carrés expliqués)(variance interclasse).

Lorsque le facteur n'a pas d'effet, les moyennes sont semblables dans les différentes classes, donc la variance interclasse est faible (proche de 0) et la variance intraclasse est proche de la variance totale. Dans le cas contraire, la variance interclasse quantifie l'effet du facteur (qui peut être plus ou moins forte), la variance intraclasse est plus faible que la la variance totale et correspond à la variabilité qui n'est pas expliquée par les valeurs moyennes sur chaque classe.

On peut dans un premier temps penser à considérer le ratio R^2 de la variance interclasse par la variance totale qui représente la part de variance interclasse (celle qui est expliquée par le facteur). Cependant, on préfère considérer l'indicateur

$$F = \frac{SCE/(k-1)}{SCR/(n-k)}$$

dont on connaît la loi de probabilité (loi de Fisher) sous certaines hypothèses (normalité, indépendance et homogénéité des variables) lorsqu'il n'y a pas d'effet du facteur. Plus le facteur a une influence sur la variable à expliquer, plus F est grand. Par conséquent, il paraît logique de rejeter l'hypothèse qu'il n'y a pas d'effet du facteur si l'indicateur F est supérieur à un seuil. On fixe ce seuil $F_{v,\alpha}$ de manière à contrôler la probabilité α de se tromper en concluant que le facteur a un effet. On se sert du fait que l'on connaît la loi de F lorsque le facteur n'a pas d'effet (hypothèse $\mathcal{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k$) pour choisir le seuil tel que :

$$\mathbb{P}_{\mathcal{H}_0}(F > F_{v,\alpha}) \approx \alpha \text{ (on donnera plus de détails plus tard).}$$

Remarque importante : pour que l'ANOVA soit valide, on doit vérifier les hypothèses suivantes :

1. les variables $Y_{i,j}$ sont de loi normale.
2. les variances des $Y_{i,j}$ ne varient pas d'une sous-population à l'autre (i.e. la variance de la variable quantitative à expliquer ne dépend pas des modalités du facteur).

3. les variables $Y_{i,j}$ sont indépendantes.

La démarche de l'ANOVA :

Une étude préalable peut être faite pour vérifier ces hypothèses. Pour contrôler l'homoscedasticité (c'est à dire que la variance de la variable quantitative à expliquer ne dépend pas des modalités du facteur) on compare les écarts-type intra-groupe et on fait un test de Bartlett (ou un test de Levenne). Ensuite, on fait souvent l'histogramme des résidus suivi d'un test de Shapiro-Wilks sur les résidus pour tester la normalité (on peut également réaliser un test de Kolmogorov-Smirnov). Enfin, la dernière condition d'indépendance est d'ordinaire satisfaite lorsque l'on a un échantillon aléatoire simple ou en utilisant une procédure "d'aléatorisation" (ou de randomisation) : procédure par laquelle on affecte au hasard chaque individu à un groupe expérimental.
Si certaines de ces conditions (normalité et homoscédasticité) ne sont pas vérifiées on utilisera plutôt un test non-paramétrique de Kruskal-Wallis.

Etape 1 :

$$H_0 : \mu_1 = \mu_2 = \cdots \mu_k \text{ et } H_1 : \exists 1 \leq j \neq l \leq k,\ \mu_j \neq \mu_l.$$

Fixer le seuil du test α (en général 0.05)

Etape 2 :
La statistique de test est

$$F = \frac{SCE/(k-1)}{SCR/(n-k)}$$

sous H_0 elle suit une loi de Fisher à $(k-1, n-k)$ degrés de liberté. Elle "n'est pas grande si H_0 est vraie" et a tendance à prendre des valeurs positives plus fortes si H_0 n'est pas vérifiée.

Etape 3 :
Il semble assez logique au vu de l'étape 2 et des commentaires qui précèdent de choisir de rejeter l'hypothèse H_0 si la statistique F est supérieure à une valeur seuil c_α. On prend donc $R_\alpha = \{F > c_\alpha\}$ en choisissant c_α le plus petit possible tel que $\mathbb{P}_{H_0}(F > c_\alpha) \leq \alpha$. Puisque F est de loi de Fisher(k-1,n-k) sous H_0 on prend $c_\alpha = Q_{F(k-1,n-k)}(1-\alpha)$.

Etape 4 :
On calcule finalement la valeur f de F obtenues à partir de nos données (c'est à dire en remplaçant les $Y_{i,j}$ par les observations $y_{i,j}$).
Si $t > c_\alpha$ on rejette l'hypothèse nulle avec un risqueα de se tromper. Sinon, pour un risque α, l'effet du facteur n'est pas significatif.
De façon générale, aprés l'ANOVA, plusieurs cas se présentent :

1. on décide de ne pas tenir compte du facteur, puisqu'on n'a pas observé d'effet significatif de celui-ci sur le phénomène étudié. Ainsi, pour modéliser la variable explicative pour un individu, on prend la moyenne de l'échantillon total.

2. au contraire, il semble que le phénomène dépende du facteur, cette fois ci, pour modéliser la variable explicative sur un individu, on évalue d'abord la modalité du facteur prise par cet individu et on prend la moyenne sur l'échantillon des individus ayant pris cette modalité.

Utilisation du logiciel R
Le logiciel nous donne la probabilité qu'une variable de Fisher dépasse F et si cette p-value est inférieure au risque, on rejette H_0 et on conclut que le facteur a bien une influence sur la variable à expliquer.

Tableaux récapitulatif proposé par les logiciels.

source de variation	degrés de liberté	Somme	carrés moyen	F	$p-value$
Expliqués	$p-1$	SCE	CME	CME/CMR	
Résidus	$n-p$	SCR	CMR		
Total	$n-1$	SCT			

EXERCICE :
Retour à notre exemple concernant les forêts.

1. Réaliser une analyse de variance.
2. Discuter les résultats obtenus et conclure.

L'indicateur F donne 5.5 ce qui donne une probabilité P(>F) inférieure à 0.05 (il s'agit d'une p-value, qui a un sens quand on fait des hypothèses gaussiennes). La variabilité des hauteurs des arbres s'explique de faç on significative par les différentes forêts.

EXERCICE : On étudie la résistance à la corrosion de différents tuyaux. On dispose de deux variables qualitatives pour tenter d'expliquer la variabilité des observations (le type de sol dans lequel ils se trouvent et le type de protection (peinture) qu'ils ont reçu. Les données sont disponibles dans le fichier tuyaux de la librairie lycee.

1. Charger les données avec R.
2. Tracer les boîtes à moustaches (box plot) de la corrosion pour comparer les résultats selon les différents types de sol puis selon les différentes protections.
3. Effectuer une analyse de variance pour mieux cerner les effets de chacun des facteurs. Commenter et conclure.
4. Pour aller plus loin, lorsque l'on a deux facteurs, on réalise plutôt une ANOVA à deux facteurs que deux ANOVA séparée à un facteur. On peut le faire avec R de la manière suivante : $aov(y\ x1 + x2)$ où + signifie que l'on considère que les effets des deux facteurs sont indépendants. On peut remplacer le signe + par le signe $*$ afin de considérer également les interactions entre les deux facteurs (l'effet d'un facteur peut dépendre de la valeur de l'autre facteur). Essayer ces commandes et interpréter les résultats.

EXERCICE : Revenons au jeu de données entreprise (dans la librairie lycee) que nous avions étudié il y quelques semaines. Pour essayer de prévoir la défaillance des entreprises, l'économiste W. BEAVER inroduit le ratio défini, pour chaque entreprise, par le quotient de la marge brute d'autofinancement (cash flow) par la dette totale.

1. Charger les données avec R.
2. Tracer les boîtes à moustaches (box plot) du ratio pour comparer les résultats selon les différents types d'entreprises (saines ou défaillanes).
3. Effectuer une analyse de variance pour mieux cerner les effets du facteur. Commenter et conclure.