LES STATISTIQUES INFERENTIELLES

(Test de Student)

L'inférence statistique est la partie des statistiques qui, contrairement à la statistique descriptive, ne se contente pas de décrire des observations, mais extrapole les constatations faites à un ensemble plus vaste et permet de tester des hypothèses sur cet ensemble ainsi que de prendre des décisions.

Un test statistique est un mécanisme qui permet de trancher entre deux hypothèses au vu des résultats d'un échantillon.

Soient H_0 et H_1 deux hypothèses (H_0 est appelée hypothèse nulle, H_1 hypothèse alternative), dont une et une seule qui est vraie. La décision consiste à retenir H_0 ou H_1.

- Hypothèse nulle, $H_0 : p_A = p_B$
- Hypothèse alternative, $H_1 : p_A \neq p_B$.

I. LES TESTS PARAMETRIQUES

Un test est dit paramétrique si son objet est de tester une hypothèse relative à un ou plusieurs paramètres d'une variable aléatoire qui suit la loi normale ou ayant un effectif important (n > 30).

1. Le test de Student

Ce test permet de comparer :

- une moyenne d'un échantillon à une valeur donnée
- les moyennes de deux échantillons indépendants
- les moyennes de deux échantillons appariés.

L'emploi de ce test reste subordonné en général à deux conditions d'application importantes qui sont la normalité et le caractère aléatoire et simple des échantillons.

La première condition n'est toutefois pas essentielle lorsque les échantillons ont des effectifs suffisants (en pratique, la valeur de 30 est souvent retenue) pour assurer la quasi-normalité des distributions d'échantillonnage des moyennes. En plus, de ces deux conditions, nous devrons supposer, dans certains tests relatifs aux moyennes, l'égalité des variances des échantillons considérées.

a. Cas d'un seul échantillon

Le test de Student cas d'un seul échantillon est aussi appelé **test de conformité**, ce test a pour but de vérifier si notre échantillon provient bien d'une population avec la moyenne spécifiée, μ_0, ou s'il y a une différence significative entre la moyenne de l'échantillon et la moyenne présumée de la population.

Exemple: Une usine veut vérifier le bon fonctionnement de ces machines car l'usure des machines peut impliquer une déviation aux normes imposées.

Nous tirons aléatoirement un certain nombre d'éléments de la production, nous calculons la moyenne et nous comparons celle-ci avec la norme imposée. Les hypothèses à tester sont :

- hypothèse nulle : $H_0 : \mu = \mu_0$
- hypothèse alternative : $H_1 : \mu \neq \mu_0$

Conditions d'application du test de Student : Le caractère de l'échantillon étant supposé aléatoire, l'hypothèse de normalité de la variable X doit être vérifiée (par exemple) avec le test de Kolmogorov-Smirnov si $n < 30$.

Calcul : Soit X une variable aléatoire distribuée selon une loi normale, la variable aléatoire définie ci-dessus suit une loi de Student avec n - 1 degrés de liberté.

$$t_{obs} = \frac{\overline{X} - \mu_0}{\sqrt{\frac{S^2}{n}}}$$

où μ_0 est la moyenne de la population spécifiée par H_0, $\overline{X}$ est la moyenne de l'échantillon, S^2 est la variance de l'échantillon et n est la taille de l'échantillon

On compare la valeur calculée de t (t_{obs}) avec la valeur critique appropriée de t avec n - 1 degrés de liberté. On rejette H_0 si la valeur absolue de t_{obs} est supérieure à cette valeur critique. Les valeurs critiques pour différents degrés de liberté et différents seuils de signification sont données par la table de Student.

b. Cas de deux échantillons indépendants

Etant donné deux échantillons de taille n_1 et n_2, on admet qu'ils ont été prélevés d'une même population relativement à la variable étudiée, ces deux échantillons ayant été prélevés indépendamment l'un de l'autre ?

Les hypothèses à tester sont :

- hypothèse nulle : $H_0 : \mu_1 = \mu_2$
- hypothèse alternative : $H_1 : \mu_1 \neq \mu_2$

Conditions d'application :

- Les deux échantillons sont indépendants entre eux, sont aléatoires et ont n_1 et n_2 unités indépendantes (cette condition est d'ordinaire satisfaite en utilisant une procédure de randomisation ; procédure pour laquelle on affecte au hasard chaque individu à un groupe expérimental).
- La variable aléatoire suit une loi normale ou elle a des effectifs supérieurs à 30.
- Il est aussi nécessaire de **vérifier l'égalité des variances** des échantillons (grâce au test de Fisher). Cette condition est **indispensable pour des effectifs inégaux.**

Remarques:

Plusieurs auteurs ont montré que **l'hypothèse de normalité est d'importance relativement secondaire dans le test d'égalité de deux moyennes**. En effet, dans certaines limites, la non-normalité des populations ne modifie pas sensiblement les risques d'erreur de première et deuxième espèce. Ceci est vrai surtout pour les distributions symétriques, même très différentes des distributions normales. De même, **l'hypothèse d'égalité des variances n'est pas fondamentale au point de vue pratique lorsque les effectifs des échantillons sont égaux.** En raison de cette faible sensibilité du test à la non-normalité et à l'inégalité des variances, on dira qu'il s'agit, pour des effectifs égaux, d'un test robuste. Par contre, l**orsque les effectifs des échantillons sont inégaux, il est absolument indispensable de s'assurer de l'égalité des variances** et, si cette hypothèse n'est pas vérifiée, il est indispensable d'utiliser une méthode adaptée à ces circonstances. On peut notamment procéder à une transformation de variable, destinée à stabiliser les variances, et utiliser ensuite le test de Student. Cependant, ce cas **d'inégalité des variances est assez rare**.

Mode de calcul : On calcule la valeur t observé (t_{obs}) qui suit une variable aléatoire de Student aux degrés de liberté (ddl = $n_1 + n_2$-2).

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}$$

où $\bar{x}_1$ *et* $\bar{x}_2$ *sont les moyennes des deux échantillons,* S_p^2 *la variance commune.* Cette dernière statistique correspond à la variance S^2 de la population parentale. Elle est égale à :

$$Vc = \frac{(n-1)*V1 + (n-1)*V2}{n1 + n2 - 2}$$ avec $$Var = \frac{1}{n-1} * \left(\sum_{i=1}^{i=n} xi^2 - \left(\sum_{i=1}^{i=n} xi\right)^2 / n \right)$$

Ce qui revient à : $S_p^2 = \frac{SCE_1 + SCE_2}{(n_1 - 1) + (n_2 - 1)} = \frac{SCE_1 + SCE_2}{n_1 + n_2}$

Si les effectifs des échantillons sont égaux, la valeur de t devient :

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{2S_p^2}{n}}}$$

La valeur de t est comparée à la valeur critique appropriée de t (dans la table de Student) avec **($n_1 + n_2$ - 2) degrés de liberté**. On rejette H_0 si la valeur absolue de t_{obs} est supérieure à cette valeur critique.

c. Cas de deux échantillons appariés:

Le test de Student pour observations pairées sert à comparer les moyennes de deux populations, dont chaque élément de l'une des populations est mis en relation avec un élément de l'autre. Par exemple, il peut s'agir de comparer deux traitements, les données étant considérées comme des paires d'observations (première observation de la paire recevant le traitement 1 et deuxième observation recevant le traitement 2).

Aspects mathématiques :

Soit x_{ij} l'observation j pour la paire i (j = 1,2 et i = 1,2,...,n). Pour chaque paire d'observations on calcule la différence $d_i = x_{i2} - x_{i1}$. Le test statistique est défini par :

$$t_{obs} = \frac{\bar{d}}{\sqrt{\frac{S_d^2}{n}}}$$

où n est le nombre de paires d'observations, $\bar{d}$ *est la moyenne des différences entre les observations et* S_d^2 *la variance.*

Le test de Student pour observations pairées est un test bilatéral. Les hypothèses sont :

- $H_0 : \mu_1 - \mu_2 = 0$ (il n'y a pas de différence entre les traitements)
- $H_1 : \mu_1 - \mu_2 \neq 0$ (il y a une différence entre les traitements)

On rejette l'hypothèse nulle au seuil de signification α si : $|t_{obs}| > t_{n-1,1-\alpha/2}$ où $t_{n-1,1-\alpha/2}$ est la valeur de la table de Student avec n - 1 degrés de liberté.

Conditions d'application :

- les échantillons ont été tirés aléatoirement
- la population des différences doit suivre une loi de Gauss. Cette condition est moins restrictive que celle de normalité des deux populations.

ANOVA 1 et 2

1. ANOVA 1

Il existe deux sources de variation entre les n observations tirées d'un essai en randomisation totale. L'une est la variation due aux traitements et l'autre est l'erreur expérimentale. Leur taille relative indique si la différence observée entre les traitements est réelle ou si elle est due au hasard. La différence due au traitement est " réelle " si elle dépasse dans une mesure significative l'erreur expérimentale.

Nous allons voir ci-dessous les étapes de l'analyse de variance des données provenant d'une expérimentation relative à un dispositif complètement randomisé (DCR) comportant un nombre de répétitions non uniforme. Les formules peuvent être adaptées facilement en cas de répétitions égales, de sorte qu'elles ne sont pas décrites à part.

*Etape 1. Regrouper les données par traitements et calculer les totaux des traitements (T_i) et le total général (G).

*Etape 2. Dresser un Tableau d'analyse de variance, suivant le modèle du Tableau 1.

Tableau 1. Schéma de l'analyse de variance d'un DCR, avec répétitions inégales

Source de variation	Degrés de liberté (df)	Somme des carrés (SS)	Carré moyen $\left(MS = \frac{SS}{df}\right)$	Valeur calculée de F
Traitement	t - 1	SST	MST	$\frac{MST}{MSE}$
Erreur	n - t	SSE	MSE	
Total	n - 1	$SSTO$		

*Etape 3. Avec les totaux des traitements (T_i) et le total général (G), calculer comme suit le facteur de correction et les différentes sommes des carrés.

$$C.\ F. = \frac{G^2}{n}$$

$$SSTO = \sum_{i=1}^{t}\sum_{j=1}^{r_i} y_{ij}^2 - C.\ F.$$

$$SST = \sum_{i=1}^{t} \frac{T_i^2}{r_i} - C.\ F.$$

$$SSE = SSTO - SST$$

*Etape 4. Entrer toutes les valeurs des sommes des carrés dans le tableau d'analyse de la variance et calculer les carrés moyens et la valeur de *F*.

*Etape 5. Prendre les valeurs tabulaires de *F*, avec f_1 et f_2 degrés de liberté, où f_1 = df du traitement = $(t - 1)$ et $f_2 = df$ de l'erreur = $(n - t)$, respectivement.

*Etape 6. Comparer la valeur calculée de *F* de l'Etape 4 avec la valeur tabulée de *F* de l'Etape 5, et déterminez si la différence entre les traitements est significative, d'après les règles ci-après :

i) Si la valeur calculée de *F* est supérieure à sa valeur tabulaire au seuil de signification de 5%, la variation due aux traitements est dite *significative,* ce qui est généralement indiqué par un astérisque au-dessus de la valeur calculée de F, dans l'analyse de variance.

ii) Si la valeur calculée de *F* est inférieure ou égale à la valeur tabulaire de *F* au seuil de signification de 5%, la variation due aux traitements est dite non significative, ce qui est indiqué par la mention ns au-dessus de la valeur calculée de *F* (ou par l'absence d'indication au-dessus de cette valeur).

Une valeur non significative de *F* dans l'analyse de variance indique que l'expérience n'a pas réussi à détecter de différence entre les traitements. Elle ne prouve en aucun cas que tous les traitements sont les mêmes car la non détection d'une différence entre les traitements, attestée par une valeur non significative du critère *F*, pourrait s'expliquer par une différence nulle ou minime, ou par une erreur expérimentale importante, ou encore par ces deux facteurs. Ainsi, dans tous les cas où la valeur de *F* n'est pas significative, le chercheur devrait examiner l'ampleur de l'erreur expérimentale et les différences numériques entre les moyennes des traitements. Si ces deux valeurs sont grandes, il est conseillé de refaire l'essai et de tenter de réduire l'erreur expérimentale pour que les éventuelles différences entre les traitements puissent être détectées. En revanche, si les deux valeurs sont petites, les différences entre les traitements sont probablement trop faibles pour avoir une signification économique, si bien qu'il n'est pas nécessaire de faire de nouveaux essais.

On notera qu'une valeur significative de *F* confirme l'existence de quelques différences entre les traitements testés, mais ne précise pas pour quelle(s) paire(s) de traitements spécifiques la différence est significative. Ces informations s'obtiennent grâce aux procédures de comparaison des moyennes.

*Etape 7. Calculer comme suit la moyenne générale et le coefficient de variation (*cv*):

$$\text{Moyenne générale} = \frac{G}{n}$$

$$cv = \frac{\sqrt{MSE}}{\text{Moyenne générale}}(100)$$

Le *cv* affecte le degré de précision des comparaisons entre les traitements et donne une bonne indication de la fiabilité de l'expérience. C'est une expression de l'erreur expérimentale totale, en pourcentage de la moyenne totale ; Ainsi, plus la valeur de *cv* est grande, moins l'expérience est fiable. Le *cv* varie considérablement suivant le type d'expérience, la plante cultivée, et les caractères mesurés. Toutefois, un chercheur expérimenté peut relativement bien juger de l'acceptabilité d'une valeur spécifique du *cv* pour un type d'expérience donné. Les résultats d'expériences donnant un *cv* supérieur à 30% sont sujets à caution.

**** Comparaison des traitements***

Dans le domaine de la recherche végétale, l'une des procédures les plus couramment employées, pour les comparaisons appariées est le test de Newman-Keuls (Δ_{NK}). D'autres méthodes, comme le test de la plus petite différence significative (PPDS), le test de Duncan et le test de la différence raisonnablement significative sont employées. Le test de Newman-Keuls fournit une valeur unique qui, à un niveau de signification déterminé, marque la limite entre la différence significative et non significative entre une paire de moyennes de traitements quelconque.

Deux traitements présentent donc des différences significatives à un seuil de signification prescrit si leur différence est supérieure à la valeur calculée de Δ_{NK}. Dans le cas contraire, leurs différences sont considérées comme non significatives.

Ce calcul nécessite l'utilisation de la table NK à 3 entrées comportant:

1) risque globale de première espèce α
2) le nombre de degrés de liberté (k) avec lesquels est estimée la variance de population
3) le nombre de moyennes à comparer (i)

La table fournit alors la valeur $q_{1}^{\ i,k}{}_{-\alpha}$

Chaque amplitude est alors comparée à:

$$q_{1-\alpha}^{i,k}\sqrt{\frac{\hat{\sigma}^2}{n}}$$

2. ANOVA 2

Dans toute expérience, une ou plusieurs variables de réponse peuvent être affectées par un certain nombre de facteurs dans le système global, dont certains sont maîtrisés ou maintenus aux niveaux voulus dans l'expérience. Une expérience dans laquelle les traitements sont constitués de toutes les combinaisons possibles de deux ou plusieurs facteurs, aux niveaux sélectionnés, est appelé plan d'expérience factoriel. Par exemple, une expérience sur l'enracinement des boutures englobant deux facteurs, mesurés à deux niveaux – par exemple deux hormones à deux dosages différents – est une expérience factorielle 2 x 2 ou 2^2. Les traitements sont constitués des quatre combinaisons possibles de chacun des deux facteurs, aux deux niveaux considérés.

	Combinaison des traitements	
Numéro du traitement	Hormone	Dose (ppm)
1	NAA	10
2	NAA	20
3	IBA	10
4	IBA	20

On utilise parfois l'expression *expérience factorielle complète* lorsque les traitements comprennent toutes les combinaisons des niveaux sélectionnés des facteurs, mais l'expression *expérience factorielle fractionnée* ne s'applique que le test ne porte que sur une fraction de toutes les combinaisons. Toutefois, pour simplifier, les expériences factorielles complètes seront, tout au long de ce manuel, appelées simplement expériences factorielles. On notera que le terme *factoriel* se réfère au mode de constitution spécifique des traitements et n'a rien à voir avec le plan décrivant le dispositif expérimental. Par exemple, si l'expérience factorielle 2^2 dont nous avons parlé plus haut fait partie d'un plan d'expérience en blocs aléatoires complets, l'expérience devrait être définie par l'expression *expérience factorielle* 2^2 *dans un plan en blocs aléatoires complets.*

Dans un plan d'expérience factoriel, le nombre total de traitements est égal au produit du nombre de niveaux de chaque facteur; dans l'exemple factoriel 2^2 , le nombre de traitements est égal à 2 x 2 = 4, dans une expérience factorielle 2^3, le nombre de traitements est 2 x 2 x 2 = 8. Le nombre de traitements augmente rapidement avec le nombre de facteurs ou avec les niveaux de chaque facteur. Pour une expérience factorielle comprenant 5 clones, 4 espacements et 3 méthodes de désherbage, le nombre total de traitements sera 5 x 4 x 3 = 60. On évitera donc le recours inconsidéré aux expériences factorielles en raison de leur ampleur, de leur complexité et de leur coût. De plus, il est peu raisonnable de se lancer dans une expérience de grande ampleur au début d'un travail de recherche, alors qu'il est possible, avec plusieurs petits essais préliminaires, d'obtenir des résultats prometteurs. Imaginons par exemple qu'un généticien ait fait venir 30 nouveaux clones d'un pays voisin et veuille voir comment ils réagissent à l'environnement local. Etant donné que normalement les conditions de l'environnement varient en fonction de plusieurs facteurs, tels que la fertilité du sol, le degré d'humidité, etc. l'idéal serait de tester les 30 clones dans le cadre d'une expérience factorielle englobant d'autres variables, telles que engrais, niveau d'humidité et densité de population.

Le problème est que l'expérience devient alors extrêmement vaste du fait de l'adjonction d'autres facteurs que les clones. Même si l'on incluait qu'un seul facteur, comme l'azote ou l'engrais, à trois dosages différents, le nombre de traitements passerait de 30 à 90. Une expérience de cette ampleur pose divers types de problèmes, notamment pour obtenir des financements ou une surface expérimentale adéquate, ou pour contrôler l'hétérogénéité du sol etc. Pour faciliter les choses, il est donc préférable de commencer par tester les 30 clones dans une expérience à un facteur, puis de sélectionner sur la base des résultats obtenus un petit nombre de clones à soumettre à un examen plus détaillé. Par exemple la première expérience à un facteur peut montrer que seuls cinq clones ont des performances suffisamment remarquables pour justifier des tests plus approfondis. Ces cinq clones pourraient ensuite être insérés dans une expérience factorielle avec trois dosages d'azote, ce qui donnerait un expérience à quinze traitements , alors qu'il en faudrait 90 dans une expérience factorielle avec 30 clones.

Dans la majorité des plans d'expérience factoriels, les traitements sont trop nombreux pour qu'un plan en blocs aléatoires puisse être efficace. Certains types de plans ont cependant été spécifiquement mis au point pour des expériences factorielles de grande envergure.

**** Analyse de variance***

Tout plan en blocs complets pour des expériences à un facteur est applicable à un plan d'expérience factoriel. Les procédures de randomisation et de représentation schématique de chaque plan peuvent être appliquées directement, en ignorant simplement la composition factorielle des traitements et en faisant comme s'il n'existait pas de relation entre les traitements. Pour l'analyse de variance, les calculs examinés pour chaque plan sont aussi directement applicables. Toutefois, des étapes de calcul doivent être ajoutées pour répartir les sommes des carrés des traitements entre les composantes factorielles correspondant aux effets principaux des facteurs individuels et à leurs interactions.

Nous allons décrire les différentes étapes de la procédure d'analyse de la variance d'une expérience à deux facteurs, avec deux niveaux (Facteur A) et trois niveaux (facteur B), à trois répétitions. La liste des six combinaisons factorielles des traitements figure dans le Tableau 2, le dispositif expérimental est illustré à la Figure 1.

Tableau 2. Les combinaisons factorielles (2 x 3) des traitements, avec deux niveaux (Facteur A) et trois niveaux (Facteur B).

Facteur B	Facteur A	
	(a_1)	(a_2)
(b_1)	a_1b_1	a_2b_1
(b_2)	a_1b_2	a_2b_2
(b_3)	a_1b_3	a_2b_3

Figure 1. Schéma-type d'un plan d'expérience factoriel 2×3, avec deux niveaux (facteur A) et trois niveaux (facteur B) et 3 répétitions.

a_2b_3	a_2b_3	a_1b_2
a_1b_3	a_1b_2	a_1b_1
a_1b_2	a_1b_3	a_2b_2
a_2b_1	a_2b_1	a_1b_3
a_1b_1	a_2b_2	a_2b_1
a_2b_2	a_1b_1	a_2b_3

*Etape 1. Soit *r* le nombre de répétitions, *a* le nombre de niveaux du facteur A, et *b* le nombre de niveaux du facteur B. Dresser le tableau de l'analyse de variance:

Tableau 3. Représentation schématique de l'analyse de variance d'une expérience factorielle avec deux niveaux du facteur A, trois niveauxs du facteur B et trois répétitions.

Source de variation	Degrés de liberté (*df*)	Somme des carrés (*SS*)	Carré moyen $\left(MS = \frac{SS}{df}\right)$	*F* calculé
Répétition	r-1	*SSR*	MSR	
Traitement	ab- 1	*SST*	MST	$\frac{MST}{MSE}$
A	a- 1	*SSA*	MSA	$\frac{MSA}{MSE}$
B	b- 1	*SSB*	MSB	$\frac{MSB}{MSE}$
AB	(a-1)(b-1)	*SSAB*	MSAB	$\frac{MSAB}{MSE}$
Erreur	(r-1)(ab-1)	*SSE*	MSE	
Total	rab -1	*SSTO*		

*Etape 2. Calculer les totaux des traitements (T_{ij}), les totaux des répétitions (R_k), le total général (G) et calculer *SSTO*, *SSR*, *SST* et *SSE*. Notons y_{ijk} l'observation correspondant au $i^{ème}$ niveau du facteur A et au $j^{ème}$ niveau du facteur B dans la $k^{ième}$ répétition.

$$C.F. = \frac{G^2}{rab}$$

$$SSTO = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r} y_{ijk}^2 - C.F.$$

$$SSR = \frac{\sum_{k=1}^{r} R_k^2}{ab} - C.F.$$

$$SST = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} T_{ij}^2}{r} - C.F.$$

$$SSE = SSTO - SSR - SST$$

*Etape 3. Construire le tableau à double entrée des totaux facteur A x facteur B, avec le calcul des totaux du facteur A et les totaux du facteur B.

*Etape 4. Calculer les trois composantes factorielles de la somme des carrés des traitements:

$$SSA = \frac{\sum_{i=1}^{b} A_i^2}{rb} - C.F.$$

$$SSB = \frac{\sum_{j=1}^{b} B_j^2}{ra} - C.F.$$

$$SSAB = SST - SSA - SSB$$

*Etape 5. Calculer le carré moyen de chaque source de variation en divisant chaque somme des carrés par les degrés de liberté qui lui sont associés et obtenir les valeur du rapport *F* pour les trois composantes factorielles, selon le schéma du Tableau 4.

*Etape 6. Entrer toutes les valeurs obtenues durant les Etapes 3 à 5, dans l'analyse de variance de l'Etape 2 en suivant les indications du Tableau 4.

*Etape 7. Comparer chaque valeur calculée de *F* avec la valeur tabulaire de *F*,
avec $f_1 = df$ du *MS* du numérateur et $f_2 = df$ du *MS* du dénominateur, au seuil de signification voulu.

*Etape 8. Calculer le coefficient de variation:

$$\text{cv} = \frac{\sqrt{\text{Erreur MS}}}{\text{Moyenne générale}} \times 100$$

**** Comparaison des traitements***

Les moyennes des traitements sont comparées selon la méthode décrite pour la section ANOVA 1.

3. *Corrélation de deux variables*

*** Corrélation**

Dans beaucoup de systèmes naturels, les changements d'un attribut s'accompagnent de variations d'un autre attribut, et il existe une relation définie entre les deux. En d'autres termes, il existe une corrélation entre les deux variables. Par exemple, plusieurs propriétés des sols, comme la teneur en azote, la teneur en carbone organique ou le pH, sont corrélées et varient de façon concomitante. On a observé une forte corrélation entre plusieurs caractéristiques morphométriques d'un arbre. Dans de telles situations, il peut être intéressant pour un chercheur de mesurer l'importance de cette relation. Si (x_i,y_i); $i = 1, ..., n$, est un ensemble d'observations appariées effectuées sur n unités d'échantillonnage indépendantes, une mesure de la relation linéaire entre deux variables est donnée par la quantité suivante, appelée coefficient de corrélation linéaire de Pearson, ou simplement coefficient de corrélation.

$$r = \frac{\text{Covariance de x et y}}{\sqrt{(\text{Variance de x})(\text{Variance de y})}} = \frac{\text{Cov(x, y)}}{\sqrt{(\text{V(x)})(\text{V(y)})}} \qquad (3.24)$$

où $Cov\ (x,y) = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right)$

$$V(x) = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)$$

$$V(y) = \frac{1}{n}\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)$$

Ce paramètre statistique indique à la fois la direction et le degré de la relation existant entre deux caractères quantitatifs x et y. La valeur de r peut varier de –1 à +1, sans atteindre ces valeurs. Si la valeur de r est nulle, cela signifie qu'il n'y a pas de relation linéaire entre les deux variables concernées (il peut toutefois y avoir une relation non-linéaire). La relation linéaire est forte lorsque la valeur de r approche –1 ou +1. Une valeur négative de r indique que si la valeur d'une variable augmente, celle de l'autre diminue. Au contraire, une valeur positive indique une relation directe, c'est à dire que l'augmentation de la valeur d'une variable est associée à une augmentation de la valeur de l'autre. Un changement d'origine, d'échelle, ou d'origine et d'échelle est sans incidence sur le coefficient de corrélation. Lorsque l'on ajoute ou soustrait un terme constant aux valeurs d'une variable, on dit que l'on a changé d'origine, alors que lorsque l'on multiplie ou divise par un terme constant les valeurs d'une variable, on parle de changement d'échelle.

A titre d'exemple, considérons les données du Tableau 3.10 concernant le pH et la teneur en carbone organique mesurés dans des échantillons de terrain provenant de 15 fosses d'observation creusées dans des forêts naturelles.

Tableau 3.10. Valeurs du pH et de la teneur en carbone organique observées dans des échantillons de terrain prélevés dans des forêts naturelles.

Fosse d'observation	pH (x)	Carbone organique (%) (y)	(x^2)	(y^2)	(xy)
1	5.7	2.10	32.49	4.4100	11.97
2	6.1	2.17	37.21	4.7089	13.24
3	5.2	1.97	27.04	3.8809	10.24
4	5.7	1.39	32.49	1.9321	7.92
5	5.6	2.26	31.36	5.1076	12.66
6	5.1	1.29	26.01	1.6641	6.58
7	5.8	1.17	33.64	1.3689	6.79
8	5.5	1.14	30.25	1.2996	6.27
9	5.4	2.09	29.16	4.3681	11.29
10	5.9	1.01	34.81	1.0201	5.96
11	5.3	0.89	28.09	0.7921	4.72
12	5.4	1.60	29.16	2.5600	8.64
13	5.1	0.90	26.01	0.8100	4.59
14	5.1	1.01	26.01	1.0201	5.15
15	5.2	1.21	27.04	1.4641	6.29
Total	82.1	22.2	450.77	36.4100	122.30

Le coefficient de corrélation se calcule en plusieurs étapes.

*Etape 1. Calcul de la covariance de x et y et des variances de x et de y à l'aide de l'équation (3.24).

$$Cov\ (x,y) = \frac{1}{15}\left(122.30 - \frac{(82.1)(22.2)}{15}\right)$$

$$= 0.05$$

$$V(x) = \frac{1}{15}\left(450.77 - \frac{(82.1)^2}{15}\right)$$

$$= 0.0940$$

$$V(y) = \frac{1}{15}\left(36.41 - \frac{(22.2)^2}{15}\right)$$

$$= 0.2367$$

*Etape 2. Calcul du coefficient de corrélation avec l'équation (3.24).

$$r = \frac{0.05}{\sqrt{(0.0940)(0.2367)}}$$

$$= 0.3541$$

**** Test de signification du coefficient de corrélation.***

La signification d'une valeur du coefficient de corrélation calculée à partir d'un échantillon doit être testée pour confirmer l'existence d'une relation entre les deux variables, dans la population considérée. En général, on définit l'hypothèse nulle comme $H_0: \rho = 0$ alors que l'hypothèse alternative est $H_1: \rho \neq 0$.

Pour *n* relativement petit, l'hypothèse nulle ($\rho = 0$) peut être testée à l'aide du critère statistique

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (3.25)$$

Ce critère statistique suit une distribution de Student *t* avec *n*-2 degrés de liberté.

Examinons les données du Tableau 3.10, où $n = 15$ et $r = 0.3541$. Pour tester si $H_0: \rho = 0$ ou si, au contraire, $H_1: \rho \neq 0$, on calcule le critère statistique à l'aide de l'Equation (3.25).

$$t = \frac{0.3541\sqrt{15-2}}{\sqrt{1-(0.3541)^2}} = 1.3652$$

Dans la table de l'Annexe 2, la valeur critique de *t* est 2,160, pour 13 degrés de liberté, au seuil de signification $\alpha = 0{,}05$. Comme la valeur calculée de *t* est inférieure à la valeur critique, on conclut que le pH et la teneur en carbone organique mesurés à partir d'échantillons de terrain ne sont pas corrélés de manière significative. Pour simplifier, on pourrait aussi se reporter à l'Annexe 5 qui donne les valeurs au-delà desquelles un coefficient de corrélation observé peut être déclaré significatif, pour un nombre donné d'observations au seuil de signification voulu.

Pour tester l'hypothèse $H_0: \rho = \rho_0$, où ρ_0 est une valeur donnée quelconque de ρ, on utilise la transformation Z de Fisher donnée par

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \qquad (3.26)$$

où ln indique le logarithme naturel.

Pour tester l'hypothèse nulle, on choisit le critère statistique

$$w = \frac{z - z_0}{\sqrt{\frac{1}{n-3}}} \qquad (3.27)$$

où $z_0 = \frac{1}{2}\ln\left(\frac{1+\rho_o}{1-\rho_0}\right)$

Le critère statistique *w* suit une loi de distribution normale standard.

Pour illustrer ceci par un exemple, prenons les données du Tableau 3.10, pour n = 15 et r = 0.3541. Supposons que l'on veuille tester l'hypothèse nulle $H_0 : \rho = \rho_0 = 0.6$; on commencera par soumettre les valeurs de *r* et ρ à la transformation *z*.

$$z = \frac{1}{2}\ln\left(\frac{1+0.3541}{1-0.3541}\right) = 0.3701$$

$$z_0 = \frac{1}{2}\ln\left(\frac{1+0.6}{1-0.6}\right) = 0.6932$$

La valeur du critère statistique sera donc

$$w = \frac{0.3701 - 0.6932}{\sqrt{\frac{1}{15-3}}} = 1.16495$$

Etant donné que la valeur de *w* est inférieure à la valeur critique 1.96, le critère n'est pas significatif au seuil de signification de 5%. On en conclut que le coefficient de corrélation entre le pH et la teneur en carbone organique ne diffère pas de manière significative de 0.6.

*** Régression**

Le coefficient de corrélation mesure le degré de la relation entre deux variables qui varient de façon concomitante, avec des effets qui se renforcent mutuellement. Dans certains cas, les changements relatifs à une variable sont provoqués par les variations d'une variable connexe, sans qu'il y ait de dépendance mutuelle. En d'autres termes, une variable est considérée comme dépendante des variations de l'autre variable, dans la mesure où elles dépendent de facteurs externes. Une telle relation entre deux variables est appelée régression. Lorsque ces relations sont exprimées sous forme mathématique, il est possible d'estimer la valeur d'une variable d'après la valeur de l'autre. Par exemple, le rendement de conversion photosynthétique et le coefficient de transpiration des arbres dépendent de conditions atmosphériques comme la température ou l'humidité, sans pour autant que l'on s'attende généralement à une relation inverse. Toutefois certaines variables sont souvent déclarées indépendantes uniquement au sens statistique, même dans des situations où des effets inverses sont concevables. Par exemple, dans une équation servant à estimer le volume, le volume

des arbres est souvent considéré comme dépendant du diamètre à hauteur d'homme, même si le diamètre ne saurait être considéré comme indépendant des effets du volume des arbres au sens physique. C'est pourquoi, dans le contexte de la régression, les variables indépendantes sont souvent appelées variables exogènes (explicative), et la variable dépendante variable endogène (expliquée).

La variable dépendante est habituellement notée *y* et la variable indépendante *x*. Dans le cas où il n'y a que deux variables en jeu, la relation fonctionnelle est appelée *régression simple*. Si la relation entre les deux variables est linéaire, on parle de *régression linéaire simple ;* dans le cas contraire, la *régression* est dite *non-linéaire*. Lorsqu'une variable dépend d'au moins 2 variables indépendantes, la relation fonctionnelle entre la variable dépendante et l'ensemble des variables indépendantes est une *régression multiple*. Dans un souci de simplification, on se limitera ici à examiner le cas d'une régression linéaire simple. Pour des cas plus complexes, on se référera à Montgomery et Peck (1982).

**** Régression linéaire simple***

La régression linéaire simple de *y* en *x* dans la population peut s'exprimer comme

$$y = \alpha + \beta x + \varepsilon \qquad (3.28)$$

où α et β sont des paramètres**,** appelés aussi coefficients de régression, et ε est une déviation aléatoire pouvant dériver de la relation attendue. Si la valeur moyenne de ε est zéro, l'équation (3.28) représente une droite de pente β et d'ordonnée à l'origine α. Autrement dit, α est la valeur présumée de *y* lorsque *x* prend la valeur zéro et β représente la variation attendue de *y* correspondant à une variation unitaire de la variable *x*. La pente d'une droite de régression linéaire peut être positive, négative ou nulle, selon la relation entre *y* et *x*.

En pratique, les valeurs de α et β doivent être estimées à partir d'observations des variables *y* et *x* effectuées sur un échantillon. Par exemple, pour estimer les paramètres d'une équation de régression proposée liant la température atmosphérique et le taux de transpiration des arbres, un certain nombre d'observations appariées sur la température et le taux de transpiration sont effectuées sur plusieurs arbres, à différents moments de la journée. Notons (x_i, y_i); $i = 1, 2, \ldots, n$ ces couples de valeurs, *n* étant le nombre de d'observations appariées indépendantes. Les valeurs de α et β sont estimées par la méthode des moindres carrés (Montgomery et Peck, 1982) de sorte que la somme des carrés des différences entre les valeurs observées et prévues soit minimale. Le processus d'estimation repose sur les hypothèses suivantes: i) Les valeurs de *x* sont non aléatoires ou fixes ; ii) Pour tout *x*, la variance de *y* est la même ; iii) Les valeurs de *y* observées pour différentes valeurs de *x* sont complètement indépendantes. Si l'une de ces hypothèses n'est pas vérifiée, il faut apporter les changements voulus. Pour les tests d'hypothèses se référant à des paramètres, une hypothèse additionnelle de normalité des erreurs est nécessaire.

En effet, les valeurs de α et β s'obtiennent grâce à la formule,

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i-1}^{n} x_i\right)^2}{n}} \qquad (3.29)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad (3.30)$$

L'équation $\hat{y} = \hat{\alpha} + \hat{\beta}x$ représente la droite de régression ajustée, qui peut être utilisée pour estimer la valeur moyenne de la variable dépendante, *y*, associée à une valeur particulière de la variable indépendante, *x*. En général, il est plus sûr de limiter ces estimations à la fourchette des valeurs de *x* dans les données.

On peut obtenir une estimation des erreurs-type de $\hat{\beta}$ and $\hat{\alpha}$ avec la formule suivante :

$$SE(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i-1}^{n} x_i\right)^2}{n}}} \qquad (3.31)$$

$$SE(\hat{\alpha}) = \sqrt{\frac{\hat{\sigma}^2 \frac{\sum_{i=1}^{n} x_i^2}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i-1}^{n} x_i\right)^2}{n}}} \qquad (3.32)$$

où $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n}$

L'erreur-type d'une estimation, qui est l'écart-type de sa distribution d'échantillonnage, donne une indication du degré de fiabilité de cette estimation.

Nous illustrerons ce qui précède à l'aide des données du Tableau 3.11 qui présente les valeurs appariées du rendement photosynthétique et des radiations, obtenues à partir d'observations des feuilles d'une essence forestière spécifique. Dans cet exemple, la variable dépendante est le rendement photosynthétique et la variable indépendante est la quantité de lumière. La méthode de calcul de l'ajustement d'une régression linéaire est indiquée ci-dessous.

*Etape 1. Calculer les valeurs du numérateur et du dénominateur de l'équation (3.29) en utilisant les sommes, sommes des carrés et sommes des produits de *x* et *y*, dérivées du Tableau 3.11

$$\sum xy - \frac{\sum x \sum y}{n} = 175.59 - \frac{(13.72)(189.03)}{15} = 2.6906$$

$$\sum x^2 - \frac{\left(\sum x\right)^2}{n} = 12.70 - \frac{(13.72)^2}{15} = 0.1508$$

Tableau 3.11. Données sur le rendement photosynthétique en μ mol $m^{-2}s^{-1}$ (y) et mesure de la radiation en mol $m^{-2}s^{-1}$ (x), observées sur une essence forestière

X	y	x^2	xy
0.7619	7.58	0.58	5.78
0.7684	9.46	0.59	7.27
0.7961	10.76	0.63	8.57
0.8380	11.51	0.70	9.65
0.8381	11.68	0.70	9.79
0.8435	12.68	0.71	10.70
0.8599	12.76	0.74	10.97
0.9209	13.73	0.85	12.64
0.9993	13.89	1.00	13.88
1.0041	13.97	1.01	14.02
1.0089	14.05	1.02	14.17
1.0137	14.13	1.03	14.32
1.0184	14.20	1.04	14.47
1.0232	14.28	1.05	14.62
1.0280	14.36	1.06	14.77
$\sum x = 13.72$	$\sum y = 189.03$	$\sum x^2 = 12.70$	$\sum xy = 175.59$

*Etape 2. Calculer les estimations de α et β avec les équations (3.29) et (3.30).

$$\hat{\beta} = \frac{2.6906}{0.1508} = 17.8422$$

$$\hat{\alpha} = 12.60 - (17.8421)(0.9148)$$
$$= -3.7202$$

La droite de régression ajustée $\hat{y} = -3.7202 + 17.8422x$ peut être utilisée pour estimer la valeur du rendement photosynthétique à un niveau de radiation quelconque donné, dans la limite des données. Ainsi, le rendement photosynthétique prévu, pour 1 mol $m^{-2}s^{-1}$ de lumière sera,

$$\hat{y} = -3.7202 + 17.8422(1) = 14.122$$

*Etape 3. Estimer σ^2 selon la formule définie dans l'Equation (3.32).

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n} = 0.6966$$

*Etape 4. Estimer les erreurs-type de $\hat{\beta}$ and $\hat{\alpha}$ à l'aide des Equations (3.31) et (3.32).

$$SE(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{\sum x^2 - \frac{(\sum x)^2}{n}}} = \sqrt{\frac{0.6966}{12.70 - \frac{(13.72)^2}{15}}} = 2.1495$$

$$SE(\hat{\alpha}) = \sqrt{\frac{\hat{\sigma}^2 \frac{\sum x^2}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}} = \sqrt{\frac{0.6966 \frac{12.70}{15}}{12.70 - \frac{(13.72)^2}{15}}} = 1.9778$$

**** Test de signification du coefficient de régression***

Une fois que les paramètres de la fonction de régression ont été estimés, l'étape suivante est le test de signification statistique de la fonction de régression. Selon l'usage, on définit l'hypothèse nulle comme H_0: $\beta = 0$ en opposition à l'hypothèse alternative, H_1: $\beta \neq 0$ ou (H_1: $\beta < 0$ ou H_1: $\beta > 0$, selon la nature présumée de la relation). Pour effectuer le test, on peut suivre la procédure de l'analyse de variance. Le concept de l'analyse de la variance a déjà été expliqué dans la Section 3.6, mais ses applications dans le cadre de la régression sont indiquées ci dessous, à l'aide des données du Tableau 3.11.

*Etape 1. Dresser un schéma de la table d'analyse de la variance.

Tableau 3.12. Représentation schématique d'une analyse de variance pour une analyse de régression.

Source de variation	Degré de liberté (*df*)	Sommes des carrés (*SS*)	Carré moyen $\left(MS = \frac{SS}{df}\right)$	*F* calculé
Dû à la régression Ecart par rapport à la régression	1 *n*-2	*SSR* *SSE*	*MSR* *MSE*	$\frac{MSR}{MSE}$
Total	*n*-1	*SSTO*		

*Etape 2. Calculer les différentes sommes des carrés, selon la méthode suivante :

$$\text{Somme totale des carrés} = SSTO = \sum y^2 - \frac{\left(\sum y\right)^2}{n} \qquad (3.33)$$

$$= (7.58)^2 + (9.46)^2 + \ldots + (14.36)^2 - \frac{(189.03)^2}{15}$$

$$= 58.3514$$

$$\text{Somme des carrés dus à la régression} = SSR = \frac{\left[\sum xy - \frac{\sum x \sum y}{n}\right]^2}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}} \qquad (3.34)$$

$$= \frac{(2.6906)^2}{0.1508}$$

$$= 48.0062$$

$$\text{Somme des carrés dus à l'écart par rapport à la régression} = SSE = SSTO - SSR \qquad (3.35)$$

$$= 58.3514 - 48.0062 = 10.3452$$

*Etape 3. Entrer, comme indiqué dans le Tableau 3.13, les valeurs des sommes des carrés dans la table d'analyse de variance et effectuer les calculs restants.

Tableau 3.13. Analyse de variance pour l'équation de régression relative aux données du Tableau 3.11.

Source de variation	Degrés de liberté (*df*)	Sommes des carrés (*SS*)	Carré moyen $\left(MS = \frac{SS}{df}\right)$	*F* calculé à 5%
Dû à la régression	1	48.0062	48.0062	60.3244
Ecart à la régression	13	10.3452	0.7958	
Total	14	58.3514		

*Etape 4. Comparer la valeur calculée de *F* avec la valeur tabulaire correspondant à (1,*n*-2) degrés de liberté. Dans notre exemple, la valeur calculée (60.3244) est supérieure à la valeur tabulaire de *F* (4.67) correspondant à (1,13) degrés de liberté, au seuil de signification de 5%. La valeur de *F* est donc significative. Si la valeur calculée de *F* est significative, le coefficient de régression β diffère de 0 de manière significative. Exprimée en proportion de la somme totale des carrés, la somme des carrés due à la régression est appelée coefficient de détermination et mesure la quantité de variation de *y* imputable à la variation de *x*. En d'autres termes, le coefficient de détermination mesure la fraction de la variation de la variable dépendante expliquée par le modèle. Dans notre exemple, le coefficient de détermination (R^2) est

$$R^2 = \frac{SSR}{SSTO} \qquad (3.36)$$

$$= \frac{48.0062}{58.3514}$$

$$= 0.8255$$