Cours de bio-statistiques

Introduction

Les méthodes scientifiques de recherche portent toutes sur les principes suivants:

- Une revue des faits et des visions sur un sujet donné ou un problème.
- Une formulation de la problématique donnant lieu à des hypothèses sujettes à vérification selon une méthode expérimentale
- Conduite de l'expérience pour une évaluation objective, utilisant les données collectées, des hypothèses avancées ou émises au sujet de la problématique.

Exemple

Réponse d'une espèce de plantes données à la culture in vitro?

Les données se présentent le plus souvent comme un ensemble de valeurs dites les observations (data). La caractéristique des observations (data ou données) c'est leur variabilité (variation) d'où le nom de variable.

1-Les variables

Les phrases telles que cet arbre est haut ou bien ce sac pèse 25 kg sont très communes et elles sont informatives. Elles concernent des caractéristiques qui ne sont pas constantes mais qui varient d'un sujet à un autre. Ces caractéristiques servent à décrire et donc à distinguer les sujets. Ces caractéristiques qui peuvent prendre des valeurs différentes sont dites des variables ou variables aléatoires.

Généralement on utilise des lettres pour designer une variable donnée, par exemple le rendement des céréales est souvent désigné par la lettre Y. Y_i désigne la valeur prise par la variable Y pour la i^{ieme} observation (i variant de 1 a n).

Exemple

Y = matière sèche par plante et Y_1, Y_2, Y_3 .. Y_i... Y_n est la matière sèche des sujets d'ordre 1, 2, 3, i et n..

Les variables sont de deux sortes: quantitatives et qualitatives. Elle est quantitative lorsqu'on peut ordonner les valeurs prises par la variable mesurable Ex: le poids, la longueur, le nombre d'objet.... Les variables quantitatives peuvent être continues ou discontinues. La variable est continue lorsqu'elle peut prendre toutes les valeurs possibles dans une plage donnée (ex, m, cm, mm, micromètre). Elle est discrète lorsqu'elle peut prendre des valeurs pour certaines plages seulement (ex nombre de pétales par fleur, nombre de racines par plante, nombre de piles ou de faces). La variable est de nature qualitative lorsqu'on ne peut pas la mesurer, par exemple l'intelligence, la couleur, la croyance ect.... Dans ce cas on classe les sujets ou les objets dans une catégorie ou dans une autre.

2- Distribution de la variable

Les valeurs prises par une variable permettent de distinguer et de classer les individus dont les valeurs constituent les données de cette variable. La distribution de ces valeurs donne une idée sur la fréquence relative des valeurs prises par les individus. Par exemple la fréquence d'un pile ou face est de 0,5. On dit que la fréquence ou la probabilité d'avoir pile ou face est de 50%. La fréquence ou la probabilité suit une distribution qui va de 0 à 100%.

3- Populations et échantillons

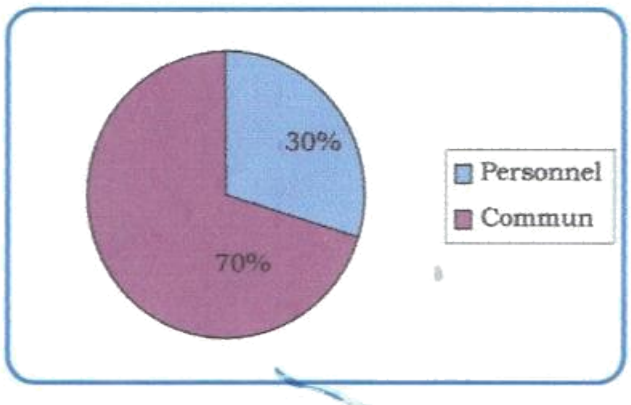
Le problème posé par les données d'une variable est celui de savoir si ces données viennent d'une population, c'est à dire représentent elles une population de données ou bien simplement une partie de cette population? La différence est parfois difficile à faire, cependant quelque soit les données on considère souvent qu'on travaille avec un échantillon, qui est assez représentatif de la population. Donc on cherche à caractériser une population en étudiant un échantillon tiré au hasard de cette population. Le tirage au hasard et de manière aléatoire est une garantie de la représentativité de l'échantillon.

Exemple:

Pour avoir une idée exacte de la quantité de la matière sèche produite dans une région donnée, il faut prendre des échantillons de toutes les zones de cette région.

5- Présentation et caractérisation des données

Il y a plusieurs méthodes utilisant des tableaux, des courbes, des histogrammes, des polygones de distribution. Par exemple si on vous demande de faire une enquête dans la ville pour connaître les moyens de transport qu'utilisent les habitants, vous utilisez un échantillon de 250 personnes et vous trouvez que 75 personnes utilisent leur véhicule personnel et 175 utilisent le transport en commun. On peut résumer ces résultats sous la forme suivante

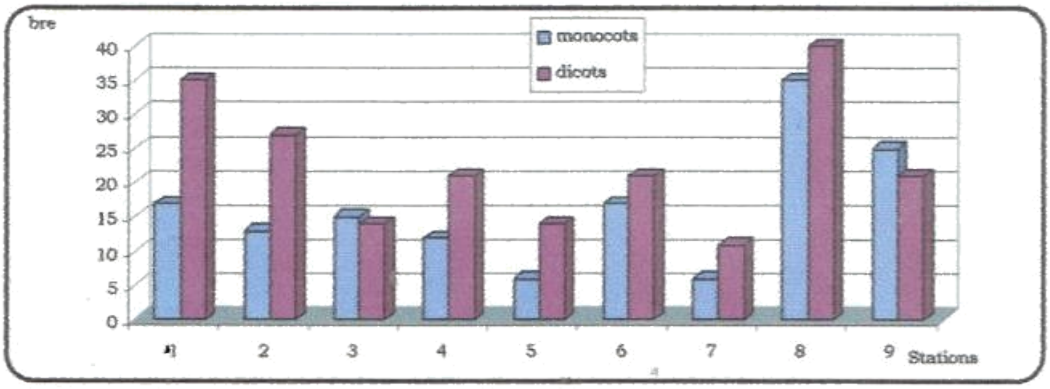


Exo:

Un étudiant parcours une jachère et fait la détermination des plantes en les classant en monocots et dicots. Les résultats de cet échantillonnage sont les suivants:

Stations	1	2	3	4	5	6	7	8	9
Monocots	17	13	15	12	16	17	6	35	25
Dicots	35	27	14	21	14	21	11	40	21

Représentez ces données sous une forme graphique?



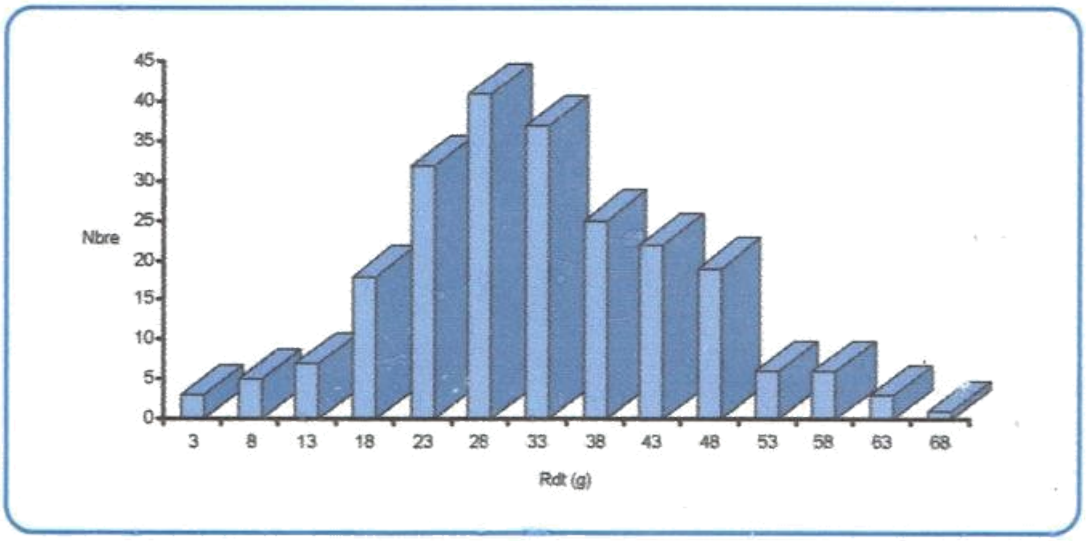
Exo

Un étudiant cherche à déterminer la distribution fréquentielle du rendement par plante du soja. Il trouve les résultats suivants:

Rdt (g)	3	8	13	18	23	28	33	43	48
Nbre de plants	7	5	7	18	32	41	37	25	22
Rdt (g)	53	58	63	68					
Nbre de plants	6	6	3	1					

Comment appelle-t-on ce tableau de données? (Tableau de fréquence)

Schématisez la distribution de la variable Rdt par un histogramme?



On utilise souvent certaines valeurs particulières pour caractériser la distribution d'une variable. Ces valeurs particulières sont la moyenne qui est une mesure de la tendance centrale et la variance qui est une mesure de la dispersion ou de la variation des valeurs prises par la variable. La moyenne arithmétique, qui est une mesure de la tendance centrale, est notée par la lettre mu (μ) pour la population et Y barre pour un échantillon. μ est une valeur fixe pour une population alors que Y barre a une valeur qui varie selon les échantillons constituants la population. La moyenne de toutes les Y barres possibles converge vers μ de la population donnée. La dispersion est notée par le lettre σ^2 pour la population et S^2 pour un échantillon. Les valeurs μ et σ^2 qui caractérisent la population sont dites paramètres et celles qui caractérisent l'échantillon, Y barre et S^2, sont dites des statistiques

$\mu \approx$ Y barre = $(\sum Yi)/n$ et $\sigma^2 \approx S^2 = [(Yi-Ybarre)^2]/n-1$

(Y barre = moyenne de Yi).

La médiane est une mesure de la tendance centrale, c'est la valeur située juste au milieu des valeurs mesurées lorsqu'elles sont ordonnées de la valeur min à la valeur max ou l'inverse. Le mode est la valeur qui revient le plus souvent dans une distribution asymétrique des données d'une variable.

Ex: Pour la distribution des valeurs suivantes 3, 6, 8, 11 et 15 la médiane est la valeur 8 et pour la distribution 3, 6, 8, et 11, la médiane est $(6+8)/2 = 7$

Exo:

Des étudiants interrogés au sujet de leur poids déclarent qu'ils pèsent:

50, 61, 47, 55, 70 et le dernier 75 kg. Calculer la moyenne arithmétique et la médiane de cet échantillon?

Exo :

Deux échantillons de poissons sont pris de deux barrages différents A et B pour déterminer la longueur

Echantillon A		Echantillon B	
Nbre	long(cm)	Nbre	long (cm)
5	10	10	10
19	12	27	12
19	14	15	14
8	16	6	16
3	18	3	18

Calculer la moyenne, le mode, les écarts à la moyenne et la somme des écarts à la moyenne de chaque échantillon ?

II- Comparaison des moyennes

-Test de signification

On a vu que les paramètres μ et σ^2 sont les caractéristiques de la population alors que Y_{barre} et S^2 sont celles de l'échantillon. Un échantillon représentatif doit avoir une moyenne Y barre très proche de μ.. Pour n échantillons les Y_{barres} couvrent un intervalle qui doit implicitement contenir μ., si ces échantillons sont représentatifs. Cet intervalle est dit intervalle de confiance (IC) il est égale à: $IC = Y_{barre} \pm t_{\alpha/2}\, S_{Ybarre}$

Avec Y_{barre} = moyenne de l'échantillon; $t_{\alpha/2}$ = valeur du t de table au seuil de α =5% et pour n-1 degrés de liberté (ddl); S_{Ybarre} =écart type de la moyenne = $\sqrt{(S^2/n)}$, Donc on prend 95% de chance pour que la moyenne(Y_{barre}) d'un échantillon soit proche ou égale à celle de la population et seulement 5% de chance pour que cette moyenne de l'échantillon soit différente significativement de μ.

On tire de cette formulation le test de signification suivant: $t_{calculé} = (\mu - Y_{barre})/ S_{Ybarre}$

à comparer avec la valeur du t de table pour n-1 degrés de liberté

Cette comparaison teste deux hypothèses qui sont:

H0 $\Rightarrow Y_{barre} \approx \mu$, donc valeur du t de table > valeur du t calculé

H1 $\Rightarrow Y_{barre} \neq \mu$, donc valeur du t de table < valeur du t calculé, dans ce cas la moyenne Y_{barre} de l'échantillon est différente significativement de la moyenne de la population, l'échantillon étudié appartient à une autre population que la population ciblée.

Ex: supposons que les caractéristiques d'un échantillon sont les suivantes:

Y_{barre} = 36.4 g et S^2 = 264.04 pour n= 10 et sachant que $\mu \approx \mu_0 = 40$. On vous demande de tester H0 vs H1, c'est à dire

H0 $\Rightarrow \mu$ =40 vs H1 $\Rightarrow \mu \neq 40$, autrement dit est ce que la différence $\mu - Y_{barre}$ = 40 -36,4 est elle significative ou non? Pour cela on détermine le t calculé:

$t_{calculé} = (\mu - Y_{barre})/ S_{Ybarre} = (40-36,4)/\sqrt{(264.01/10)} = 0.701$ pour 9 ddl

on compare cette valeur du t calculé à celle du t de table au seuil de 5% et pour 9 ddl, qui est égale à 2.26; de ce fait on accepte H0 et on rejette H1 puisque la différence entre les deux moyennes n'est pas significative.

Ce test nous autorise donc à comparer deux moyennes quelconques pour savoir si elles viennent d'une même population ou de deux populations différentes. Si la différence entre les deux moyennes est significative, ces moyennes n'appartiennent pas à la même population, au contre si elle n'est pas significative, elles viennent toutes deux de la même population.

Exo:

Soit les deux échantillons suivants:

	Echantillon A	Echantillon B	
	57.8	64.2	
	56.2	58.7	
	61.9	63.1	
n1=7	54.4	62.5	n2 = 6
	53.6	59.8	
	56.4	59.2	
	53.2	---	

Tester H0 vs H1 pour ces deux échantillons?

Pour cela on calcule:

$\sum Y$ =	393.5	367.5
$\sum Y^2$ =	22174.41	22535.87
Y_{barre} =	56.21	61.25
$(n_1-1)S^2 = (Yij - Y_{barre})^2 = Y^2ij - (Yij)^2/n$		
	54.09	26.50

$S^2_{moyenne} = S^2_m = [(n_1-1)S^2 + (n_2-1)S^2]/[(n_1-1)+ (n_2-1)] = (54.09 + 26.50)/(6+5) = 7.33$

Ddl_{moyen} = $(n_1-1) + (n_2-1) = (7-1) + (6-1) = 11$

e.t de la différence des moy = $S_{Ybarre1} - Y_{barre2} = \sqrt{S^2_m (n_1+ n_2)/n_1n_2} = \sqrt{(7+6)7.33]/(7x6)} = 1.51$

t calculé = $(Y_{barre1} - Y_{barre2})/ S_{Ybarre1} - Y_{barre2} = (56.21 - 61.25)/1.51 = -3.33$

t de table pour 5% et 11 ddl = 2.20
On rejette H0 et on maintient H1, donc les deux moyennes ne viennent pas de la même population mais de deux populations différentes.

La méthode scientifique, c'est quoi?

C'est une procédure et un outil de travail utilise pour résoudre un problème de recherche. Cette méthode comporte de nombreuse étapes dont

- L'observation du phénomène d'intérêt.

- Faire des hypothèses explicatives du phénomène observé.

- Etablir un protocole test ou des tests pour vérifier si les hypothèses explicatives avancées sont vraies ou fausses en ce qui concerne l'explication du phénomène en question.

- Conduire le test ou les tests, et sur la base des résultats obtenus, décider si oui ou non l'explication du phénomène vraie ou fausse. Dans le cas où l'explication n'est pas bonne, établir de nouvelles hypothèses explicatives et refaire les tests en conséquences.

III -Le dispositif a randomisation totale (DCR)

Lorsque le nombre de moyennes à comparer devient plus élevé que deux on utilise le test F pour savoir si les différences entre les différentes moyennes sont elles significatives ou non. Le test des hypothèses H0 vs H1 est alors fait sur la base de l'analyse de la variance. L'analyse de la variance est fonction du dispositif expérimental employé. Le dispositif le plus simple est celui de la randomisation totale. Les différents traitements ou échantillons sont disposés au hasard à l'intérieur du plan de l'expérience. Ces traitements constituent la seule source de variation que l'expérimentateur cherche à tester en plus des sources totale et résiduelle. Cette notion sera plus assimilable en utilisant un exemple chiffré. Soit 6 échantillons ou traitements que nous pouvons assimiler à des variétés et que nous codons par les lettres A, B, C, D, E et F. La variable mesurée sur ces traitements ou échantillons est le poids de 1000 grains qu'on détermine sur 4 répétitions par variété. Les résultats sont ordonnés dans le tableau suivant

	Variétés					
Répétitions	A	B	C	D	E	F
I	64	53	46	56	39	46
II	59	51	48	45	59	50
III	50	55	43	45	53	65
IV	63	69	35	42	53	59

Ce dispositif est basé sur un modèle additif qui s'écrit:

$Y_{ij} = \mu + \alpha_i + e_{ij}$

où

Y_{ij} = est la valeur prise par la variable du traitement i dans la répétition j

μ = est la moyenne de la population estimée par la moyenne de tous les traitements dite moyenne de l'essai codée par Y..

α_i = effet du traitement i

e_{ij} = résidu associé avec la valeur du traitement i sur la répétition j

Comme on l'a dit plus haut chaque traitement représente un échantillon, donc la moyenne du traitement peut être comparée à la moyenne de la population, c'est à dire $Yi. = Y_{barre}$ du traitement i par rapport à μi

On peut déduire donc l'effet du traitement i (i représente les traitements A, B, C, D, E. F) par $\alpha_i = Y_{i.} - \mu_i = Y_{i.} - Y..$ avec Y.. = moyenne générale de tous les traitements ou moyenne de l'essai. Par cette procédure on peut déterminer l'effet des 6 variétés testées comme suit:

	Variétés						
Rép	A	B	C	D	E	F	
I	64	53	46	56	39	46	
II	59	51	48	45	59	50	
III	50	55	43	45	53	65	
IV	63	69	35	42	53	59	
Totale	236	228	172	188	204	220	1248
$Y_{i.}$	59	57	43	47	51	55	52
α_i	59-52=+7	+5	-9	-5	-1	+3	

$\sum \alpha_i = 0 = (7+ 5 - 9 - 5 - 1+ 3) = 15-15$

e_{ij} est déduite par: $e_{ij} = Yi. - Yij$; ceci parce que si on ne fait pas d'erreurs, ce qui impossible, les répétitions d'un traitement donné doivent avoir une valeur observée ou mesurée égale à la moyenne du traitement c'est à dire $Y_{ij} = Y_{i.}$. Donc les erreurs faites sur ce dispositif et pour les différents traitements sont:

	Variétés					
Rép	A	B	C	D	E	F
I	64-59 = 5	-4	3	9	-12	-9
II	59-59= 0	-6	5	-2	8	-5

Al-C =59-43=16 -2 0 -2 2 10
Ce test déclare la différence non significative. -5 2 4
Si on dispose d'un traitement témoin on peut utiliser le test de Dunnett. Il existe une table du t de Dunnett. On utilise ce t pour calculer une PPDS de Dunnett 55 52
PPDS = $t_{Dunnett}\sqrt{2S^2e/r}$ = 2.76 $\sqrt{(2 \times 54)/4}$ = 2.76 x 5.20 = 14.35 répétition I est égale à:
Ce test déclare les différences A-C et B-C comme étant significatives. e_{ij}
Pour faire l'analyse de la variance on procède comme suit:
On détermine le terme correctif = TC = $(\sum Yij)^2/(t \times r)$ avec t= nbre de traitements et r =nbre de répétitions.
Somme des carrés des écarts totale = SCETot = $\sum(Yij)^2$ - TC
Somme des carrés des écarts des traitements =
SCEtrait = $[\sum(\text{tot trait})^2/r]$ -TC
Somme des carrés des écarts résiduels = SCEtot -SCEtrait
En employant ces formules on obtient les résultats qui sont organisés sous la forme d'un tableau:

Source	ddl	SCE	CME	F.obs	Ftable	
Totale	23	1732	..	..	..	
Traitements	5	760	152	2.81	2.78	
Résiduelle	18	972	54	..	..	

La valeur du test F observé est supérieure à celle du F de table donc, il y a des différences significatives entre les moyennes des différentes variétés. Pour savoir lesquelles de ces moyennes sont différentes significativement l'une des autres on utilise le test de la PPDS5%
$PPDS_{5\%} = t_{5\%}\sqrt{2S^2e/r}$ = 2.10$\sqrt{}$ (2 x54)/4 = 10.91
Donc toute différence entre deux moyennes quelconque qui dépasse celle de la PPDS est déclarée significative et les traitements concernés ont des effets significativement différents.
Les différences possibles entre les moyennes des 6 traitements sont:

	A	B	C	D	E	F
Yi.	59	57	43	47	51	55
A	0	2	16*	12*	8	4
B		0	14*	10	6	2
C			0	-4	-8	-12*
D				0	-4	-8
E					0	-4
F						0

* = différence significative au seuil de 5%

1 comparaison $\alpha = 0.05$ $1-\alpha = 0.95$
2 comparaisons $0.95 \times 0.95 = 0.9025$ donc $\alpha = 1 - 0.9025 = 0.0975$
pour 3 comparaisons on a $0.95 \times 0.95 \times 0.95 = 0.85570$ $\alpha = 1 - 0.857 = 0.1430$

On peut utiliser le test de Newman Keul's. Ce test tient compte du nombre de moyennes qui sont comparées. A mesure que le nombre de moyennes devient important la plus petite amplitude significative (PPAS) est plus élevée. La PPAS est donnée par la table de NK'S, on la multiplie par l'écart type de la moyenne $S_{Yi.barre} = \sqrt{S^2e/r}$ (différent de celui de la différence de 2 moyennes où on multiplie la variance par 2)

Nbre de moyennes	PPAS (5%)	$S_{Yi.barre}$	Valeur seuil
2	2.97	3.675	10.91
3	3.61	"	13.27
4	4.00	"	14.70
5	4.28	"	15.73
6	4.49	"	16.50

on ordonne les moyennes par ordre décroissant

A	B	F	E	D	C
59	57	55	51	47	43

si on compare A vs B on a deux moyennes la différence doit être supérieure à 10.91
si on compare A vs C, il y a 6 moyennes qui sont concernées donc la différence doit être supérieure à 16.5