TP 1 Traitement de Données
Master 1 BBT

Objectifs généraux de l'enseignement de Traitement de Données :

Le but est d'apprendre à raisonner à partir de données expérimentales, de les explorer et d'en tirer des conclusions en choisissant les outils statistiques adéquats et en interprétant leurs résultats dans leur contexte.

Lors de l'examen vous devrez être capable de :

- calculer des paramètres et intervalles de confiance sans ordinateur,
- choisir une méthode ou un traitement pour répondre à une question et en expliquer le principe,
- analyser les résultats d'une méthode ou d'un traitement obtenus à l'aide d'un logiciel.

Aucune connaissance des commandes du logiciel utilisé en TD ne sera demandée.
L'expérience d'un logiciel d'analyse statistique de données vous sera utile par la suite, mais son rôle ici est de permettre de traiter efficacement de nombreux exemples.

Le **TD 1** comprend trois parties :

A. Estimation, sans ordinateur, de paramètres et d'intervalles de confiance.
B. Prise en main du logiciel d'analyse statistique de données Minitab.
C. Exploration statistique des données d'une étude.

A) Estimation de paramètres et d'intervalles de confiance sans ordinateur

Neuf dosages d'une même solution ont été réalisés ; les valeurs obtenues en mM sont :

84	85	89	82	85	80	87	88	85

Ces 9 valeurs constituent un échantillon. La population se compose de tous les résultats possibles de dosages de la solution du même type qu'on aurait pu faire. L'échantillon représente la population. On cherche à estimer les paramètres de la population à partir des valeurs de l'échantillon.

Estimer moyenne, variance (biaisée et non-biaisée), écart-type (biaisée et non-biaisée), erreur-type de la moyenne, intervalle de confiance à 95 % de la moyenne.

Distribution de Student pour un seuil α de 5% et pour un test bilatéral												
Nombre de degrés de liberté	1	2	3	4	5	6	7	8	9	20	40	120
Valeur de t	12.7	4.3	3.18	2.78	2.57	2.45	2.37	2.31	2.26	2.09	2.02	1.98

Ecrire les formules utilisées pour chaque calcul (calculer à la main ou avec la calculette).

Quelle supposition fait-on sur la distribution des valeurs dans la population ?

Quelle distribution est utilisée pour modéliser la moyenne ?

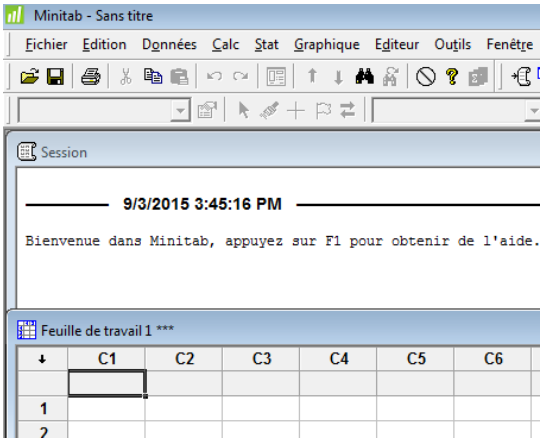
Expliquer ce que représente l'intervalle de confiance à 95 % de la moyenne ?

B) Prise en main du logiciel Minitab

Minitab est un logiciel d'analyse statistique de données. Son interface se présente sous la forme de deux fenêtres surmontées par un menu. La fenêtre « Session» correspond à une zone d'édition de traitement de texte ; la fenêtre « Feuille de Travail» à un tableau.

Fonctionnement typique:

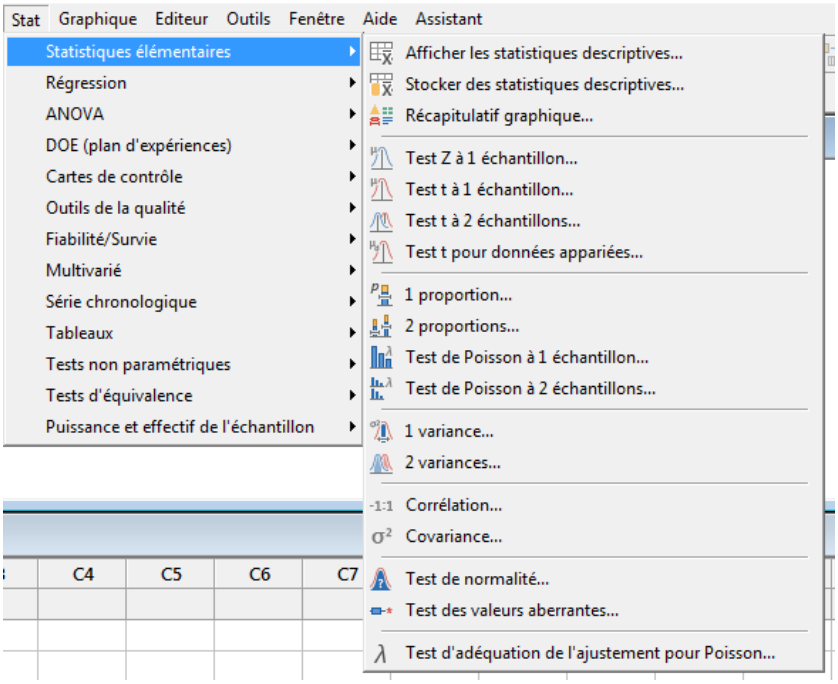
- les données sont mises dans le tableau,
- un traitement est choisi dans le menu,
- les données sur lesquelles il agit sont indiquées,
- le résultat du traitement s'affiche dans la session.



Attention :
Le tableau ne s'utilise pas comme celui d'un tableur, mais comme un tableau d'un logiciel de gestion de base de données.
Toutes les données d'une colonne doivent être de même type.
L'élément n'est pas la cellule mais la colonne entière
(le « champ » d'une base de données) ; on agit sur toutes les données d'une colonne en même temps en la désignant par son nom. On donne un nom à la colonne en le mettant dans la première case, celle qui est sur la ligne non numérotée.

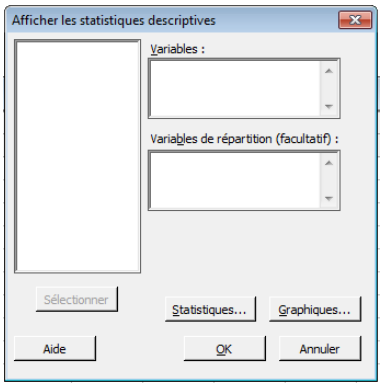
Utilisation d'un Menu

Un menu contient une liste de sous-menus correspondant à différents types d'analyses de données. Chaque sous-menu contient une liste de commandes.



Sélection des variables pour une commande

Si une commande (« Afficher les statistiques descriptives » par exemple) est choisie, la boîte de dialogue correspondante s'ouvre. La fenêtre de gauche de la boîte contient la liste des variables (colonnes) existantes utilisables par cette commande. *Remarque : si la liste ne s'affiche pas, cliquer dans la fenêtre de droite.* Pour sélectionner une ou des variables : double-cliquer sur le nom de la variable ou bien marquer la ou les variables et agir sur la touche « sélectionner ». Les noms des variables sélectionnées s'affichent dans la fenêtre de droite.



Dans le menu principal, on va utiliser successivement les menus : Stat, Calc, Données et Graphique.

1) Obtenir des statistiques descriptives : menu Stat

Saisir les données :
Saisir les valeurs du **A)** dans une colonne du tableau; nommer la colonne Dosages.

Obtenir des statistiques descriptives :
Menu Stat / Statistiques élémentaires / Afficher les statistiques descriptives…

Dans la fenêtre de dialogue, sélectionner un nom de colonne dans la liste de gauche (ici, il n'y en a qu'une : Dosages, elle s'affiche dans la case à droite). Un ensemble de statistiques s'affiche dans la partie *Session*. Repérer les différents paramètres fournis.

Recommencer avec Stat / Statistiques élémentaires / Récapitulatif graphique

2) Générer des nombres aléatoires de répartition normale

On peut à l'aide du logiciel générer aléatoirement une série de valeurs suivant une loi normale (c'est-à-dire générer une série de valeurs comme si elles étaient prises au hasard parmi un nombre très grand de valeurs réparties de façon gaussienne). On crée ainsi un échantillon issu d'une population connue. L'avantage est qu'on connaît les « vraies valeurs » des paramètres moyenne et écart-type de la population, celles qu'on obtiendrait avec un nombre infini de mesures.

Générer une série de 1000 valeurs aléatoires réparties selon une loi normale, issues d'une population de moyenne 80 et d'écart-type 5. Les stocker dans une colonne nommée Y.
Menu Calc / Données aléatoires / Normale

Obtenir les statistiques élémentaires correspondantes. Représenter l'histogramme avec sa courbe normale associée aux données.

Comparer plusieurs séries de valeurs aléatoires de même effectif.
Générer en une seule fois 10 séries de 4 valeurs de moyenne 80 et d'écart-type 5.
Les stocker dans les variables nommées Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8 Y9 Y10.
N.B. Entrer dans la boîte Stocker dans des colonnes *: les noms des variables séparés par un espace.*

Obtenir en une seule fois les statistiques élémentaires des 10 variables sans demander de graphique. Comparer les moyennes et les erreurs-types des 10 échantillons. Ces valeurs sont des estimations pour une même population.

Recommencer avec des séries de 40 valeurs aléatoires.
N'effacez pas les variables Y. Stockez les séries de 40 valeurs dans des variables nommées Z1 Z2 Z3 Z4 Z5 Z6 Z7 Z8 Z9 Z10.
Comparer entre elles les moyennes et les erreurs-types des 10 échantillons Z. Comparer avec la dispersion des paramètres obtenus avec 4 valeurs seulement.
N.B. Ne pas effacer les valeurs des variables générées ; on va s'en servir à nouveau.

3) Manipuler des données

Bien organiser ses données est nécessaire pour les traiter et les analyser efficacement. Observez les tableaux ci-dessous représentent deux façons d'organiser les mêmes données de concentrations résultant de dosages effectués à différentes températures :

T18	T22	T26	T30
17.5	19.1	20.1	20.9
18.2	20.5	20.6	21.4
	19.4	19.8	

Concentration	Température
17.5	18
18.2	18
19.1	22
20.5	22
19.4	22
20.1	26
20.6	26
19.8	26
20.9	30
21.4	30

La première organisation est dite désempilée tandis que la seconde est empilée. La seconde organisation est souvent préférable pour faciliter l'analyse. Chaque ligne du tableau représente alors un dosage (avec sa concentration, sa température, on peut ajouter d'autres informations : conditions, date, auteur,…) et chaque colonne une variable. Les données utilisées pour les prochaines séances de TD seront présentées de cette façon.

Dans certains cas, on aura besoin de séparer les données qui sont dans une colonne en plusieurs colonnes selon les valeurs d'une autre variable qui jouera le rôle d'indice. Saisissez les données des colonnes Concentration et Température. Désempilez la colonne Concentration dans 4 colonnes en utilisant pour indice la colonne Température. Pour désempiler une colonne : Menu Données / Désempiler. Dans la boite de dialogue, choisissez « Désempiler les données dans: Concentration » et « En utilisant des indices dans: Température ».
NB : Pour faire l'opération inverse, on empile des colonnes.

4) Obtenir des représentations graphiques

Il est toujours utile de visualiser ses données de façon graphique avant de les traiter.

Deux exemples de graphes :

- Boîte à moustache : Menu Graphique / Boîte à moustache
  Afficher la boîte à moustache représentant les valeurs de Y (Saisir le nom de la variable dans la case sous Y).
  Afficher simultanément les boîtes à moustaches des 10 échantillons Y1 à Y10, puis Z1 à Z10. *N.B. Commencer par empiler les 10 colonnes Yi dans une seule colonne nommée Ytotal en mémorisant leurs numéros dans une colonne Yindice.*
  Puis créer un diagramme de type boîte à moustache avec Ytotal et Yindice (option Un Y Avec groupes). Choisir Variables du graphique : Ytotal et Variable de catégorie : Yindice.
  Comparer les diagrammes des Y et des Z.
- Diagrammes de points : Menu Graphique / Nuage de points / Simple
  Par exemple, avec les données précédentes, on peut afficher un diagramme de nuage de points ($Concentration = f(Température)$). Commentaires.

C) Exploration statistique des données d'une étude

1) Les données de l'étude « DonneesNaissances »

L'étude[1] citée ci-dessous a été réalisée dans un hôpital d'Oakland entre 1961 et 1973. Lors de chaque naissance, de nombreuses informations médicales et socio-économiques concernant le bébé et ses parents ont été collectées. Dix ans plus tard, de nouvelles informations étaient recueillies à nouveau. L'étude avait pour but de rechercher si certaines caractéristiques des parents avaient une influence sur le développement de l'enfant. Le fichier fourni est constitué d'une partie des données et des variables de cette étude. Il comprend 115 lignes (individus ou unités statistiques) décrites par les 17 variables ci-dessous.

Enfant à la naissance		
ESx	sexe	M ou F
ERh	facteur rhésus	R+ ou R-
ETaille0	taille	en cm (converti à partir de pouces)
EPoids0	poids	en kg (converti à partir de livres)
Enfant à 10 ans		
ETaille10	taille	
EPoids10	poids	
Mère à la naissance de l'enfant		
MRh	groupe sanguin	
MAge	âge	au dernier anniversaire avant la naissance
MPoids0	poids	
MCig0	conso de cigarettes	0 cig/jr ; 1 à 10 ; plus de 10
Mère 10 ans après		
MTaille10	taille	
MPoids10	poids	
MCig10	conso de cigarettes	
Père à la naissance de l'enfant		
PAge	âge	au dernier anniversaire avant la naissance
PCig0	conso de cigarettes	
Père 10 ans après		
PTaille	taille	
PPoids10	poids	

[1] Source des données : J.L. Hodges, D. Krech et R. Crutchfield in Statlab : an Empirical Introduction to Statistics, 1975.

Pour transférer les données dans Minitab :

Les données se présentent sous la forme d'un fichier Excel « DonneesNaissances ».
Sélectionner entièrement la feuille Excel des données et copier.
Se positionner dans Minitab dans la première case de titre de colonne et coller.
En une seule fois, les noms des colonnes sont mis dans la ligne de titre et les données dans le tableau.

Les objectifs, au cours de cette fin de séance et des prochaines séances, seront d'utiliser les outils de la statistique descriptive pour explorer ces données puis pour obtenir, à partir de ces données, des informations sur la population qu'elles représentent.

Réfléchissez aux questions qu'on peut se poser à partir des données de cette étude.

2) Exploration des données

a) Description unidimensionnelle

- Variable quantitative: le poids de l'enfant à la naissance.
  Avec la commande : Menu Stat > Statistiques élémentaires > Afficher les Statistiques Descriptives
  Calculer les différents indicateurs quantitatifs relatifs à cette variable, représenter son histogramme avec courbe de Gauss et le diagramme boîte (option Graphiques). Commentaire.
- Variable qualitative: la consommation de cigarette de la mère au moment de la naissance.
  Calculer les effectifs et proportions de chaque classe, tracer un diagramme en secteur.
  Menu Stat > Tableaux > Tableaux à entrées multiples et Khi2 (choisir Dénombrement et en Lignes : MCig0).
  Menu Graphique > Graphique en secteur (diagramme en secteur)
  Commentaires.

b) Description bidimensionnelle

- Etude de la liaison entre une variable quantitative et une variable qualitative: étude du poids de l'enfant à la naissance selon la consommation de cigarette de la mère.
  Sur un même graphe, obtenir plusieurs diagrammes boîtes parallèles :
  Menu Graphique > Boîte à moustache, un Y avec groupes.
  Comparer ces boîtes, les médianes. Commentaire sur les différences observées.
- Etude de la liaison entre deux variables quantitatives: poids et taille de l'enfant à la naissance. Obtenir un diagramme de points :
  Menu Graphique > Nuage de points… (le premier de la liste du menu Graphique)
  Commentaire sur la forme du nuage de points et sur la liaison entre les variables.
- Etude de la liaison entre deux variables qualitatives: sexe et rhésus sanguin de l'enfant
  Construire la table de contingence, calculer les profils.
  Menu Stat > Tableaux > Tableaux à entrées multiples et Khi2
  Choisir Lignes : ESx et Colonnes : ERh. Cocher les cases pourcentages. Trouver les effectifs de chaque cellule de la table de contingence. Comparer les profils c'est-à-dire, par exemple, les pourcentages des rhésus par sexe.
  Commentaire sur la liaison entre les variables.

TP 2 Traitement de Données
Master BBT

Le **TD 2** comprend trois parties :

A. Déroulement d'une analyse statistique avec test : quelques éléments.
B. Exemples de tests statistiques appliqués à différentes questions.

A la fin de ce TD, chacun de vous se sera constitué :

- une fiche résumant la mise en œuvre et l'interprétation d'un test statistique et
- un document (schéma, organigramme, plan,…) pour choisir parmi les différents tests étudiés.

Les données utilisées pour ce TD proviennent du fichier DonneesNaissances.xls. Ouvrez ce fichier avec Excel et transférez les données dans une feuille de travail Minitab en suivant les indications fournies lors du **TD 1**.

Six questions vous seront proposées à partir des données du fichier. Lorsque vous les aurez traitées, recherchez de nouvelles questions et faites pour chacune la démarche complète d'analyse statistique.

A) Déroulement d'une analyse statistique avec test : quelques éléments

1) Démarche

Expliciter une question à partir des données disponibles. Exemple: La consommation de cigarettes par la mère a-t-elle un effet sur le poids de l'enfant à la naissance ?

Reformuler la question pour la traiter avec un test statistique : Y-a-t-il une différence entre les moyennes des poids des enfants à la naissance selon la consommation de cigarettes de la mère ?

Vérifier que cette différence existe bien dans vos données (statistique descriptive) ; le test, lui, servira à montrer si la différence existe dans les populations représentées par ces données.

Choisir un test approprié et vérifier que ses conditions d'emploi sont réunies : répartitions gaussiennes, égalité des variances,…).

2) Mise en œuvre du test

Poser l'hypothèse nulle H0 : *dans les populations, il n'y a pas de différence entre les moyennes des poids des enfants à la naissance selon la consommation de cigarettes de la mère.*
Et l'hypothèse alternative H1 : *dans les populations, il y a au moins deux moyennes qui sont différentes.*

Obtenir la P-valeur et l'interpréter :

- si la P-valeur est inférieure à un seuil α (par exemple 5%), l'hypothèse H0 est rejetée. L'autre hypothèse est acceptée: on dit qu'il y a une différence significative (au seuil de risque de 5%);

- si la P-valeur n'est pas < 5%, l'hypothèse H0 n'est pas rejetée : on ne sait pas s'il y a une différence ou pas.

3) Retour à la question de départ

Attention, si on a montré qu'il y a une différence entre les moyennes de poids de l'enfant à la naissance selon la consommation de cigarette de la mère, on n'a pas montré que la consommation de cigarette des mères est la cause de cette différence de poids.
Il pourrait, par exemple, y avoir une cause commune aux différences de poids et de consommation de cigarette. Entre d'autres termes : corrélation entre deux variables n'implique pas forcément causalité entre ces deux variables.

B) Exemples de tests statistiques appliqués à différentes questions

1) Comparer une moyenne avec une valeur théorique

Question :

Un chercheur a conçu un modèle qui prédit que la taille moyenne des enfants de 10 ans, dans la région et au moment où a eu lieu l'étude sur les naissances, devrait être de 1,40 m. L'échantillon dont vous disposez (variable ETaille10), supposé représentatif de la population de ces enfants, est-il en accord avec cette affirmation ?

Exploration des données :

Menu Stat > Statistiques élémentaires > Afficher les statistiques descriptives

Afficher moyenne, écart-type de l'échantillon, erreur-type de la moyenne, médiane.
Afficher l'histogramme avec courbe normale associée.
La répartition des données semble-t-elle normale ?

Test de normalité :

Menu Stat > Statistiques élémentaires > Test de normalité

Trois tests sont proposés. Choisir l'un d'eux (Anderson-Darling par exemple).
Considérer le graphe associé (droite de Henry) et interpréter la P-valeur du test d'Anderson-Darling.

Test d'une différence de la moyenne avec la valeur théorique :

Poser l'hypothèse H0 et l'hypothèse alternative H1.
Réaliser le test :

Menu Stat > Statistiques élémentaires > Test t à un échantillon

Cette rubrique peut être utilisée de deux façons :

- soit en demandant l'intervalle de confiance de la moyenne (préciser dans Options le niveau de confiance 95%),
- soit en demandant le test proprement dit : (cocher « Effectuer le test » et préciser la valeur de la moyenne à tester: « 140 » et le critère : « Moyenne ≠ de moyenne … » dans Options).

Analyser les résultats du test :

- avec l'intervalle de confiance : l'I.C. 95% de la moyenne contient-il la valeur à tester ? En est-il loin ? Conclusion.
- avec la P-valeur : Est-elle inférieure à 5% ? Est-elle très petite ? Conclusion.

Les deux conclusions doivent toujours être cohérentes.
Pour obtenir une vue graphique synthétique :
Recommencer Menu Stat > Statistiques élémentaires > Test t à un échantillon
Avec en plus Graphique… / Histogramme des données et / Boîte à moustaches des données

2) Comparer des moyennes avec deux échantillons indépendants

Question:

La taille de l'enfant à la naissance (ETaille0) est-elle différente selon le sexe (ESx) ?

Exploration des données :

Les tailles sont dans une seule colonne ; on peut si nécessaire désempiler la colonne selon le critère « Sexe » à l'aide de la fonction Menu Données > Désempiler les colonnes avec

- Désempiler les données (qui sont) dans ETaille0 et
- en utilisant les indices (qui sont) dans ESx pour obtenir deux échantillons.

Test de normalité sur chaque échantillon : Utiliser successivement les trois tests disponibles : Anderson-Darling, Ryan-Joiner, Kolmogorov-Smirnov. Comparer les résultats.

Conclusion : la normalité des données est-elle acceptée ? NB : les tests peuvent donner des résultats contradictoires en terme de p-valeur. Dans ce cas, on peut considérer que H0 est rejetée si au moins 2/3 des tests ont une p-valeur <0.05.

Test d'une différence entre les moyennes: poser l'hypothèse H0 et l'hypothèse alternative.

Choix du test : les données sont considérées comme réparties normalement et les échantillons ne sont pas appariés : test de Student. Menu Stat > Statistiques élémentaires > Test t à 2 échantillons

Analyser les résultats du test :

- en utilisant l'intervalle de confiance de la différence,
- en utilisant la P-valeur.

Conclusions ?
Remarque : si on voulait utiliser un test non paramétrique, ce serait celui de Mann-Whitney.

3) Comparer des moyennes avec deux échantillons appariés

Question:

Le poids de la mère est-il différent à la naissance (MPoids0) et dix ans après (MPoids10) ?
Ce qui change par rapport à la question précédente : à chaque valeur d'un échantillon correspond dans l'autre échantillon une valeur du même individu statistique (sur la même ligne du fichier) ; les deux échantillons sont appariés.

Exploration des données:

Vérifiez qu'il y a une différence entre les moyennes des deux échantillons ; est-elle due au hasard, ou les populations (au sens statistique) représentées sont-elles vraiment différentes ? Repérer, en comparant avec la courbe normale associée, que les histogrammes des deux échantillons ne semblent pas symétriques.

Tester la normalité:

Vérifiez que les données de chaque échantillon ne sont vraiment pas réparties selon une loi normale.

Choix du test :

Les conditions d'un test paramétrique ne sont pas réunies ; utiliser un test non paramétrique, le test de Wilcoxon pour données appariées. La comparaison sera sur les médianes.

Préparation des données :

Créer une colonne qui contient la différence entre MPoids0 et MPoids10.

Menu Calc > Calculatrice avec Stocker le résultat dans la variable DiffPoids et Expression 'MPoids10' - 'MPoids0'

Test :

Menu Stat > Tests non paramétriques > Wilcoxon pour un échantillon

Il teste si la médiane de DiffPoids est différente de zéro. Conclusion ?

Remarque: Les effectifs étant grands, bien que les répartitions des deux échantillons ne soient pas normales, les répartitions de leurs moyennes (si on recommençait avec d'autres échantillons de mêmes effectifs) seraient proches de la distribution normale et on pourrait utiliser raisonnablement le test paramétrique de Student pour données appariées Menu Stat > Statistiques élémentaires> Test t pour données appariées. On pourra vérifier qu'il donne un résultat équivalent.

4) Tester la liaison entre deux variables qualitatives

Question:

Y-a-t-il un lien entre le sexe et le facteur rhésus de l'enfant ?

Pour voir s'il semble y avoir une liaison, construire la table de contingence (ou tableau croisé), calculer les profils. Menu Stat > Tableau > Tableau à entrées multiples…

Trouver les effectifs de chaque cellule de la table, les effectifs marginaux, les profils lignes et colonnes (les pourcentages). Comparer ces profils, c'est-à-dire, par exemple, les pourcentages des rhésus pour chaque sexe. Commentaire sur l'existence apparente ou non d'une liaison entre les deux variables. Le test de khi-deux compare les effectifs observés d'une table de contingence avec les effectifs calculés en supposant qu'il n'y a pas de liaison (pour cela, dans Khi deux, cocher test du Khi deux). Lire la p-valeur du test de Pearson dans la console.

Poser les hypothèses et interpréter les résultats du test. *Etudier également la relation entre le facteur rhésus de la mère et celui de l'enfant.*

5) Tester la liaison entre deux variables quantitatives

Question:

Y-a-t-il un lien entre la taille et le poids de l'enfant à la naissance ?

Pour voir s'il semble y avoir un lien entre les deux variables :

Menu Graphique > Nuage de points…

Commentaire sur la forme du nuage de points et sur l'éventuelle liaison.

Expliquer ce que signifie : il y a un lien entre deux variables quantitatives. Qu'apporte en plus le fait que le lien soit linéaire ? Le test sur le coefficient de corrélation teste la liaison linéaire entre les deux variables.
Menu Stat > Statistiques élémentaires > Corrélation

Poser les hypothèses et interpréter les résultats du test.
Etudier également la relation du poids de l'enfant à la naissance avec celui de son père (PPoids10).

6) Comparer des moyennes avec plusieurs échantillons : l'ANOVA

Question :

Le poids moyen de l'enfant à la naissance est-il différent selon que la mère fume pas du tout, un peu ou beaucoup ?
Pour répondre à cette question, on peut s'aider d'une ANOVA à un facteur.

Vocabulaire :

- la réponse est la variable quantitative dont on compare les moyennes (EPoids0),
- le facteur est la variable qualitative qui sert à constituer les groupes (MCig0).

Conditions:

Faire une ANOVA suppose que certaines conditions soient vérifiées :

- les distributions des populations représentées par chaque échantillon doivent être normales,
- les variances des populations sont supposées égales.

Pour vérifier les conditions :

- Test de normalité : Menu Stat > Statistiques élémentaires > Test de normalité
  Il faut le faire pour chaque groupe et pour cela, il faut avoir désempilé la colonne Epoids0 auparavant en créant trois nouvelles colonnes.
- Test de comparaison de variances : Menu Stat > ANOVA > Test de l'égalité des variances
  Les variances des trois groupes sont comparées avec un test (de Bartlett ou de Levène).

ANOVA :

Menu Stat > ANOVA > A un facteur
Hypothèse H0 : expliciter l'hypothèse.

Analyse des résultats de l'ANOVA:
Pour une vue d'ensemble, regarder le schéma des intervalles de confiance à 95% de chaque groupe. Pour une aide à la décision, repérer la P-valeur. Est-elle très petite ? Conclusion ?

Pour comprendre l'analyse :
L'ANOVA calcule deux variances. En fait, l'ANOVA calcule plus exactement à la place deux carrés moyens (CM), mais qui sont égaux aux 2 variances si on les divise par N, le nombre d'individus (voir cours ANOVA 1 facteur). La première variance correspond à la variance à l'intérieur des groupes (ligne Erreur) due au seul hasard, l'autre variance à la variance entre les groupes (ligne MCig0) ; elle est due au hasard, mais peut être aussi due à la consommation de cigarette de la mère. Si le rapport F (de Fisher) entre ces deux variances est suffisamment grand, il y a au moins deux moyennes parmi les groupes dont la différence n'est pas due au hasard. L'ANOVA est une méthode qui calcule des variances pour savoir s'il y a une différence entre des moyennes.

Pour continuer l'analyse :
Dans l'option Comparaison, utiliser la méthode de Tukey. Elle donne des intervalles de confiance pour faire les comparaisons deux à deux entre les groupes. Interprétez les résultats.

Comparer plusieurs médianes avec un test non paramétrique :

L'ANOVA est une méthode robuste qui résiste bien lorsque les conditions indiquées ci-dessus ne sont pas tout à fait respectées. Le test non paramétrique de Kruskal-Wallis peut être utilisé à la place de l'ANOVA si les conditions ne sont franchement pas respectées.

Menu Stat > Tests non paramétriques > Kruskal-Wallis

Ici, on peut vérifier qu'il donne un résultat équivalent.

TP 3 Traitement de Données
Master BBT

Objectifs:

- S'initier à l'utilisation d'une régression linéaire simple,
- Apprendre à interpréter les résultats d'une analyse de régression.

A) Régression et analyse de régression

1) Comment faire ?

- choisir un modèle, une équation (par exemple : $Y = \alpha + \beta X$) pour représenter la relation $Y = f(X)$.
- calculer les coefficients du modèle qui représentent le mieux les points (X, Y).

2) Les données

Le nombre de points étant plus grand que celui des coefficients à calculer, une fois la régression faite, il reste de l'information dans les points, qu'on peut utiliser pour répondre à des questions. C'est l'analyse de régression.
Les questions sont liées aux raisons pour lesquelles on a fait la régression. Par exemple :

- La droite convient-elle comme modèle ou bien faut-il choisir une autre équation ?
- Y-a-t-il un lien linéaire significatif entre les variables X et Y ?
- Quelle incertitude sur Y si on le prédit à partir d'une valeur de X en se servant de la droite ?

Pour chaque question, différentes méthodes permettent de répondre.

3) Des hypothèses

Des hypothèses sont faites lorsqu'on fait une régression :

- pour chaque valeur de X, les valeurs de Y sont distribuées selon une loi normale,
- pour chaque valeur de X, les variances des valeurs de Y sont égales.

On essaiera de vérifier que ces hypothèses ne sont pas gravement contredites.

B) Exemple d'étude de régression

Question :

Quel est la relation entre la taille de l'enfant à 10 ans et la taille de sa mère ?

Analyse exploratoire des données :

Avant de faire la régression, prendre le temps d'explorer les données à l'aide de représentations graphiques : Menu Graphique > Nuage de points > Simple…
Observer la répartition des points.

Question : A votre avis, y'a-t-il un lien entre les 2 variables ? Discuter du choix de la droite comme modèle.

Pour faire la régression :

Dans Minitab, plusieurs commandes sont disponibles. Nous en utiliseront successivement:

- Menu Stat > Régression > Droite d'ajustement

Cette commande est utilisable pour une régression simple. Elle donne un graphique des points avec le tracé de la courbe de régression.

- Menu Stat > Régression > Régression > Ajuster le modèle de régression
  Elle propose plusieurs options d'analyse de régression qu'on ne trouve pas dans la précédente. Ne rien rentrer dans la boite Prédicteurs de catégorie.

Première régression avec la commande « droite d'ajustement » :

On obtient (Dans options, cocher « intervalles de confiance » et « intervalles de prédiction »):

- le graphe $taille\ de\ l'enfant\ =\ f(taille\ de\ la\ mère)$ avec les points et la droite,
- l'estimation de la droite de régression $Y\ =\ a\ +\ b\ X$,
- plusieurs informations fournies par l'analyse de régression.

Questions : Quelles sont les valeurs des coefficients a et b ? Que représentent ces 2 coefficients ? Quelle est la différence entre a et α (même question pour b et β) ?

Il faut regarder en particulier :

- La P-valeur du test de la pente pour tester s'il y a un lien linéaire significatif entre Y et X.
- Le R-carré $R^2 = 18{,}4\%$. Il représente la part de variance de Y qui est explicable par la variance de X. Plus de 80% de la dispersion des valeurs de Y est due à d'autres causes que nous n'avons pas étudiées et que nous appellerons le hasard.

Questions :

- Quelles sont les hypothèses H0 et H1 de la régression ? En lisant la p-valeur, concluez-vous que le lien linéaire est significatif ? (voir console Minitab, Analyse de Variance)
- Calculer le F de Fisher, et comparer le à celui de la table (à chercher sur internet). Concluez-vous que le lien linéaire est significatif ? (voir console Minitab)
- Que représente le R^2 ? Quelle est la valeur du R^2 ? Qu'en concluez-vous ? (idem)

Toujours avec la commande « droite d'ajustement, afficher les bandes de confiance et de prédiction »:

- La bande de confiance délimite une zone qui a 95% de chances de contenir la véritable droite représentative de la relation $Y\ =\ f(X)$, celle correspondant à la population.
  Rappel : la droite de régression obtenue correspond à notre échantillon. On obtiendrait une autre droite avec un autre échantillon.
- La bande de prédiction délimite une zone qui a 95% de chances de contenir un nouveau point de la population.
  Bien comprendre quand il faut utiliser l'une et quand il faut utiliser l'autre.

Régression avec la commande « régression » :

Obtenir les graphes des résidus :

Choisir, dans la rubrique Graphiques…, cochez « normalisées » et « quatre-en-un ». Dans la rubrique Options, cochez « statistique de Durbin-Watson ». Cochez aussi dans Résultats « statistique de Durbin-Watson ». Cochez aussi dans Stockage « Valeurs résiduelles ».

Questions :

- Quelles sont les hypothèses de travail de la régression ?
- La distribution des résidus suit-elle une loi normale ? Utiliser la droite de Henri des résidus. On fait aussi un test de normalité sur les résidus calculés (nouvelle colonne RESIDUELLE).

- Observer la forme du nuage de points de ce graphe. Présente-t-il une forme particulière (banane, entonnoir couché) qui traduirait un défaut (inadéquation de la droite comme modèle, égalité des variances des résidus le long de la droite non respectée) ?
- Quelle est la valeur de la statistique de Durbin-Watson ? Si elle est entre 1 et 3, on sait qu'il y a indépendance des résidus. Qu'en concluez-vous ?
- Minitab nous donne aussi une liste de points « aberrants » (valeurs extrêmes). Quels sont ces points ?

Etudier le lien entre deux variables sans faire de régression : le test de corrélation

Le coefficient de corrélation, noté $\boldsymbol{r}$, peut être calculé directement à partir des valeurs de X et de Y. C'est le rapport de la covariance de X et Y sur le produit de leurs écart-types. Sa valeur est comprise entre -1 (corrélation négative) et +1 (corrélation positive) en passant par zéro (pas de corrélation). Un test sur sa valeur permet de savoir si la corrélation est significative.

Menu Stat > Statistiques élémentaires > Corrélation…

On peut montrer que le R-carré de la régression est le carré du coefficient de corrélation. Le test sur le coefficient de corrélation et celui sur la pente de la droite de régression donneront toujours le même résultat. Comparer les résultats des tests.

C) Autres études avec d'autres données

Pour chaque étude, avant de faire l'analyse de régression, indiquez quels étaient, à votre avis, les objectifs de ceux qui l'ont menée et quelles questions ils se posaient. Puis utilisez les différents outils de l'analyse de régression pour répondre à ces questions.

1) Bruit et problèmes cardiovasculaires

Des études suggèrent de plus en plus un lien entre bruit et problèmes cardiovasculaires. Dans l'une de ces études, l'augmentation de pression artérielle d'un échantillon de personnes a été mesurée en fonction de l'exposition à un bruit de niveau connu.

Augmentation de pression (mm Hg) en fonction du niveau de bruit (dB) :

60	63	65	70	70	70	80	80	80	80	85	89	90	90	90	90	94	100	100	100	dB
1	0	1	2	5	1	4	6	2	3	5	4	6	8	4	5	7	9	7	6	mm Hg

Analysez ces résultats.

Avant de faire une régression :
- Formulez précisément le (ou les) objectifs de l'étude.
- Avez-vous les données adéquates pour faire une étude de régression ?
- Quelles conditions doivent-elles vérifier pour faire une analyse de régression ?
- Quel graphique faites-vous pour commencer ?

Après régression avec le modèle $Y = \alpha + \beta X$:
- Première approche sur la validité du modèle : Le graphique "points + droite de régression" laisse-t-il penser que le modèle choisi n'était pas acceptable ? Contrôlez avec un diagramme des résidus.
- Etude des coefficients du modèle : Y-a-t-il un effet statistiquement significatif du bruit sur la pression ? Cet effet est-il assez grand pour avoir une signification biologique ? Quelle information apporte le coefficient de détermination R^2?
- Estimations à l'aide du modèle : Quelle est la meilleure estimation de l'augmentation moyenne de pression due à un bruit de 90 dB pour la population représentée par l'échantillon ? Dans quel

intervalle se trouve cette augmentation si on accepte un risque de 5% de se tromper (clic-droit sur le nuage et clic sur « réticule ») ? Mêmes questions, mais pour l'augmentation que présenterait une personne de la population.

2) Le DDT dans les brochets

Le DDT a été dosé dans la chair de brochets âgés de 2 à 6 ans, pêchés au même endroit.

âge (en année)	2	2	2	2	2	2	3	3	3	3	3	3
DDT (en µg/g de chair)	0.164	0.198	0.223	0.229	0.237	0.256	0.265	0.301	0.342	0.353	0.383	0.393
âge (en année)	4	4	4	4	4	4	5	5	5	6	6	6
DDT (en µg/g de chair)	0.408	0.421	0.458	0.484	0.522	0.545	0.705	0.766	0.807	1.13	1.21	1.23

Y-a-t-il accumulation du DDT au cours du temps ?
Une droite est-elle un modèle adéquat ? Sinon, proposez la suite de l'analyse.

3) La taurine et l'eau de mer

Une étude pour déterminer comment les mollusques marins s'adaptent à des niveaux de salinité différents laisse supposer que la concentration en taurine dans leur organisme joue un rôle dans cette adaptation. Les résultats suivants ont été obtenus :

Concentration en taurine (en mM)	2	9	25	43	32	54	38	51	65	39
Salinité de l'eau de mer (en partie/1000)	10	10	18	18	29	29	29	35	35	35

Analysez ces résultats.

4) Le COB et la chlorophylle (optionnel)

Pour estimer la biomasse d'une culture de phytoplancton, on peut déterminer sa concentration en Carbone Organique Particulaire (C.O.B.). La chlorophylle *a* est plus facile à doser.
On a dosé l'un et l'autre sur plusieurs prélèvements de cultures.

Chlorophylle *a* (en µg/l)	33.9	32.2	4.6	9.6	51.3	2.8	7.4	13.7	34.6	60.8	94.7
C.O.B. (en µg/l)	1183	1242	285	395	2027	275	392	546	1648	2430	4198

Peut-on estimer la concentration de C.O.B. à partir de la concentration en chlorophylle *a*? Avec quelle précision (utiliser le curseur pour se déplacer sur la bande de prédiction) ? Que peut-on dire pour 50 µg/l de chlorophylle *a*? Et pour 200 µg/l ?

5) Les lipides de la diatomée (optionnel)

Pour développer une technologie de production de combustible à partir de biomasse, une étude a été menée pour connaître les effets de divers paramètres sur la production de lipides par la diatomée *Chaetoceros muelleri*.

Dans une étude préliminaire, l'accumulation de lipides, dosés par fluorescence (en unités **F** de fluorescence par mg de poids sec) a été déterminée en fonction de la conductivité **C** (en mS/cm) du milieu dans lequel se trouvent les diatomées.

Conductivité (mS/cm)	20	25	35	45	60	70	20	25	35	45	60	70
Fluorescence (UF/mg)	762	620	708	497	532	368	805	713	661	638	445	321

Analysez ces résultats.
La conductivité a-t-elle une influence sur la production de lipides par la diatomée ?
Peut-on utiliser une droite pour représenter cette relation ?

TP 4 Traitement de Données
Master BBT

Objectifs:

- Se perfectionner dans l'usage de l'ANOVA,
- S'initier aux plans factoriels,
- Vérifier les savoirs acquis lors des 3 premiers TD.

A) ANOVA à un facteur

1) Calcul avec Excel

A l'aide du logiciel Excel, réaliser de façon détaillée les calculs pour une ANOVA à un facteur, en utilisant les données suivantes. Les temps de coagulation (en secondes) du sang de souris qui ont suivi des régimes alimentaires différents sont indiqués ci-dessous :

R1	R2	R3	R4
62	63	68	56
60	67	66	62
63	71	71	60
59	64	67	61
	65	68	63
	66	68	64
			63
			59

On veut mettre en évidence une éventuelle différence entre les moyennes. Pour cela, on peut faire une ANOVA à 1 facteur.

Décomposition des calculs :

Il s'agit de calculer la variance à l'intérieur des groupes (S_R^2) et la variance entre les groupes (S_A^2), puis de calculer le rapport $F = \frac{\frac{S_A^2}{k-1}}{\frac{S_R^2}{n-k}}$. Si le rapport F est assez grand, compte tenu des nombres de degrés de liberté (ddl), il y a au moins une différence dans les populations entre les moyennes des 4 catégories. On le vérifie en comparant avec la valeur adéquate dans une table de coefficients de Fisher trouvée sur Internet.

Le rapport F dépend de :

- Variance à l'intérieur des groupes : pour chaque résultat expérimental, calculer la différence entre le résultat expérimental et la moyenne de son groupe (cette variance traduit seulement le hasard des résultats, les erreurs expérimentales), faire la somme des carrés et diviser par le nombre de degrés de liberté.
- Variance entre les groupes : pour chaque résultat expérimental, calculer la différence entre la moyenne du groupe et la moyenne générale (cette variance traduit le hasard + peut-être une différence entre les moyennes des groupes), faire la somme des carrés et diviser par le nombre de ddl.

- Des degrés de liberté associés aux deux variances : 24 ddl au total, 1 utilisé pour la moyenne générale, 3 utilisés pour les différences entre les moyennes des 4 groupes, il reste 20 ddl pour estimer le hasard.

2) Refaire l'étude avec Minitab.

Retrouver les différents résultats des calculs intermédiaires. Copiez-collez le tableau dans Minitab puis empilez-le, puis faites l'ANOVA : Menu Stat > ANOVA > A un facteur. Dans Graphiques, cochez quatre en un.
Lire la P-valeur de l'ANOVA et les résultats des comparaisons de Tukey et interpréter.

B) ANOVA à deux facteurs. Notion d'interaction entre deux facteurs

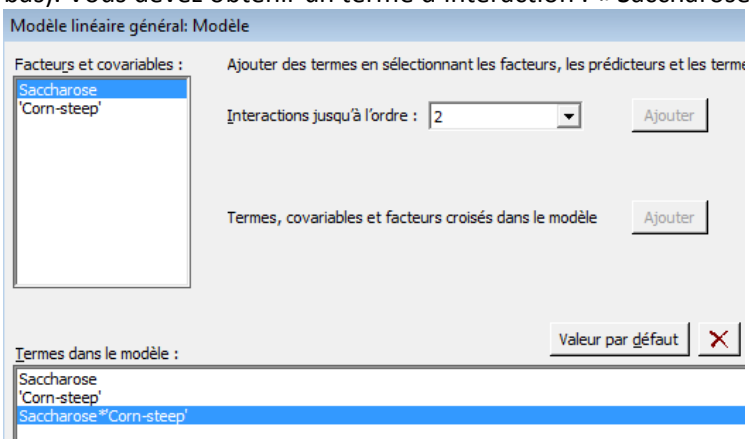
A l'aide de Minitab, analyser les résultats de l'étude ci-dessous :
Un expérimentateur veut étudier l'effet éventuel de 2 facteurs sur la production d'un métabolite par une souche de streptomycète. Après recherche bibliographique et discussion avec des collègues, il a retenu les valeurs des paramètres à tester ci-dessous :

Facteurs	Saccharose		Corn-steep	
Niveaux réels	15 g/l	25 g/l	20 g/l	30 g/l

L'expérimentateur fait 3 fois, chaque expérience correspondant à l'une des 4 possibilités de conditions expérimentales. Les résultats obtenus (la réponse) sont dans le tableau ci-dessous :

Saccharose	Corn-steep		Y (production de métabolite en g)		
			1ère série	2ème série	3ème série
15 g/l	20 g/l	Y1	0	10	0
25 g/l	20 g/l	Y2	95	70	76
15 g/l	30 g/l	Y3	32	49	41
25 g/l	30 g/l	Y4	75	86	70

Dans une feuille Minitab, reporter les valeurs des facteurs et de la réponse dans 3 colonnes : une pour le saccharose, une pour le corn-steep et une seule pour la production de métabolite (il y a donc 12 lignes, une par expérience réellement faite).
Analyser avec : Stat > ANOVA > Modèle linéaire général > Ajuster le modèle linéaire général.
Dans Réponses, entrez « Y » et dans Facteurs, entrez « Saccharose Corn-steep ». Pour rajouter le terme d'interaction, cliquez dans Modèle et sélectionnez en haut à gauche « Saccharose » et en bas à gauche « Corn-steep » puis cliquez sur Ajouter à l'option « Termes, covariables… » (voir image en bas). Vous devez obtenir un terme d'interaction : « Saccharose*'Corn-steep' ».



C) Initiation aux plans factoriels

La même étude que ci-dessus peut être réalisée avec le menu « plans d'expériences ». La démarche se fait en trois étapes :

1) Commencer par créer un plan factoriel complet pour 2 facteurs et 3 répétitions.

Stat > DOE (plan d'expériences) > Plan factoriel > Créer un plan factoriel…
Choisir : Plans factoriels à 2 niveaux (générateurs par défaut)

- Dans plans, choisir : points centraux =0 ; répétitions =3 ; blocs =1
- Dans options, choisir : décocher la case « randomiser les essais »

Un tableau de 12 lignes (une par expérience) est généré ; les valeurs de saccharose et de corn-steep à utiliser pour chaque expérience sont indiquées de façon codées (-1 pour 15g/l de saccharose et +1 pour 25 g/l dans la colonne A ; -1 pour 20 g/l de corn-steep et +1 pour 30 g/l dans la colonne B).

C'est un modèle qui indique quelles expériences faire.

2) Résultats de Minitab

Ensuite, les expériences sont réalisées et leur résultat est reporté dans la colonne à coté de A et B.

3) Analyse du plan factoriel

Stat > DOE (plan d'expériences) > Plan factoriel > Analyser un plan factoriel…
Retrouver les résultats obtenus en partie B. Ils sont complétés par les valeurs des effets et des interactions.

D) Retour sur des notions vues lors des trois premiers TD

Lisez le texte ci-dessous et vérifiez que vous avez les savoir-faire pour le premier contrôle. Faites-vous préciser ou expliquer si nécessaire.

1) Décrire de données

> Calculer, sans ordinateur, les estimations de la moyenne, la variance, l'écart-type d'une série de mesures, l'erreur-type de la moyenne et un intervalle de confiance de la moyenne à partir d'une série de valeurs de mesures d'un échantillon avec l'aide d'une table du coefficient t de Student.

Expliquer les représentations graphiques : histogramme, « boite à moustache » et diagramme de points, dans quel cas utiliser chacun et quelles informations on peut en obtenir.

2) Répondre à des questions

Les questions étudiées consistent à rechercher s'il y a une différence (entre des paramètres) ou un lien (entre deux variables quantitatives) concernant des populations à partir d'échantillons représentant ces populations.

Mise en évidence d'une différence:

Expliquer le principe de l'emploi d'un intervalle de confiance et d'un test statistique pour répondre à ces questions.

Choisir un test adéquat:

- selon les caractéristiques des données (nombre d'échantillons, effectifs des échantillons, distribution normale ou non des données, caractère apparié ou indépendant de deux séries de données),
- parmi les possibilités suivantes : test t de Student à 1 échantillon, à 2 échantillons, à 2 échantillons appariés, ANOVA à un facteur contrôlé, test de Mann et Whitney, test de Wilcoxon, test de Kruskal et Wallis.

Interpréter la P-valeur d'un test ou les valeurs d'un intervalle de confiance pour savoir s'il y a une différence significative entre des moyennes (ou entre des médianes).

Interpréter le résultat d'un test de normalité.

Analyse de régression :

Expliquer le principe d'une régression linéaire.

Expliquer différents objectifs possibles d'une analyse de régression et indiquer quelles méthodes statistiques on peut utiliser dans chaque cas pour :

- mettre en évidence un lien entre une réponse et un facteur,
- vérifier si la droite est un modèle acceptable pour représenter les données,
- estimer la précision avec laquelle on peut prédire la réponse connaissant le régresseur.

Interpréter les résultats d'une analyse de régression (intervalle de confiance de la pente de la droite, test sur cette pente, diagramme des résidus, test d'inadéquation du modèle, bandes de prédiction, bande de confiance, coefficient de détermination).