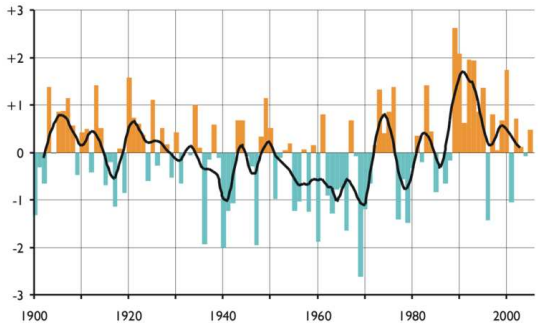
INF112 - contrôle continu 2

L'objectif du contrôle continu 2 est de répondre à une question scientifique et de présenter les résultats de votre travail dans un mini-site web. Pour répondre à la question scientifique, vous devrez programmer quelques macros. Plusieurs sujets vous sont proposés. Vous devez n'en traiter qu'un seul, au choix. La complexité des sujets est similaire.
Les fichiers sont disponibles à l'URL : http://membres-lig.imag.fr/dubousquet/INF112/CC2.zip et sur Alfresco.

1. Présentation des sujets.

Sujet 1 : ONA et pluviométrie.



L'**Oscillation Nord-Atlantique** (ONA) désigne un phénomène météorologique basé sur l'Atlantique Nord (en anglais *North Atlantic Oscillation* ou NAO est souvent utilisé dans la littérature). L'indice ONA mesure la différence de pression atmosphérique entre l'Anticyclone des Açores et la dépression d'Islande[1]. La figure ci-contre montre la valeur de l'indice ONA en fonction des années[2].
L'ONA a été découverte durant les années 1920 par Sir Gilbert Walker. C'est un facteur déterminant dans le climat car elle est reliée à la position et à la trajectoire des systèmes météorologiques du bassin atlantique-nord. Ainsi, on s'attend à ce que certains phénomènes climatiques, tels que les précipitations (en France, et plus généralement autour de l'océan atlantique) soient influencés par les ONA.

Le but du sujet 1 est de déterminer s'il existe un lien entre les ONA et les précipitations.

Sujet 2 : Valeur-C et nombre de chromosomes.

Le **génome** est l'ensemble du matériel génétique d'un individu ou d'une espèce contenu dans son ADN (à l'exception de certains virus dont le génome est porté par des molécules d'ARN). Il contient en particulier toutes les séquences codantes (transcrites en ARN messagers, et traduites en protéines) et non-codantes (non transcrites, ou transcrites en ARN, mais non traduites) [wikipédia].

La **valeur-C** représente la taille d'un génome présent dans une cellule. Par convention, la valeur-C est mesurée sur les cellules haploïdes[3]. Elle est exprimée en paire de bases, ou en pico-gramme (10^{-12} gramme). Cette valeur-C est différente pour chaque espèce mais globalement constante pour chaque individu d'une même espèce.

Intuitivement, on pourrait s'attendre à ce que plus il y a de chromosomes dans une cellule, plus le matériel génétique est important.

Le but du sujet 2 est de déterminer **si la valeur-C est corrélée au nombre de chromosomes**.

[1] http://fr.wikipedia.org/wiki/Oscillation_nord-atlantique
[2] http://fr.wikipedia.org/wiki/Fichier:Winter-NAO-Index.png
[3] Une cellule biologique est haploïde lorsque les chromosomes qu'elle contient sont chacun en un seul exemplaire.

Sujet 3 : Taux GC des chromosomes bactériens.

Le **taux de GC** (ou coefficient de *Chargaff*) d'une séquence d'ADN est défini comme la proportion de bases de cette séquence étant un couple cytosine (C) - guanine (G). Parallèlement, nous avons le couple de bases azotées adénine (A) - thymine (T), rentrant dans la composition de l'ADN. La cytosine est toujours liée à la guanine, et l'adénine est toujours liée à la thymine ; ainsi, le taux de GC exprime aussi le pourcentage de liaisons G-C dans la molécule d'ADN.

Il y a deux liaisons hydrogènes dans les paires A-T et trois liaisons hydrogène dans les paires G-C. Plus l'ADN est riche en paires G-C, plus l'ADN résiste à la dénaturation (et donc pourrait-on penser à l'augmentation de la température par exemple). Le taux de G+C d'une molécule d'ADN est la fréquence relative, exprimée généralement en pourcent, dans cet ADN. Par exemple, dans la molécule suivante :
5'-ACGT-3'
3'-TGCA-5'
le taux de G+C est de·2/4 soit 50 %.
À l'inverse, dans la molécule :
5'-CCGG-3'
3'-GGCC-5'
le taux de G+C est de 4/4 soit 100 %.

Le but du sujet 3 est de déterminer s'il existe une corrélation entre le taux de C+G d'une bactérie et l'environnement (présence d'oxygène, température, …).

2. Contrôle continu : travail à faire

1) Choisir **un** sujet

2) Le relire attentivement

3) Répondre aux questions suivantes :

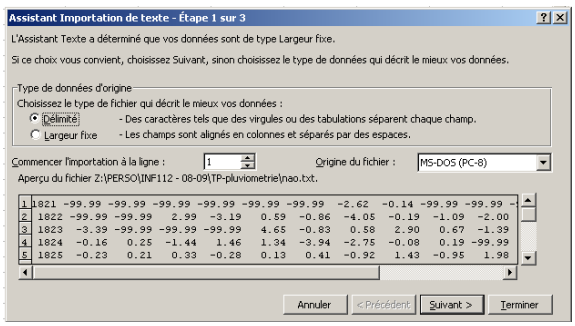
- Que va-t-on comparer pour répondre à la question ?
- Comment les données doivent-elle être disposées ?
- Quel type de comparaison va-t-on faire ?

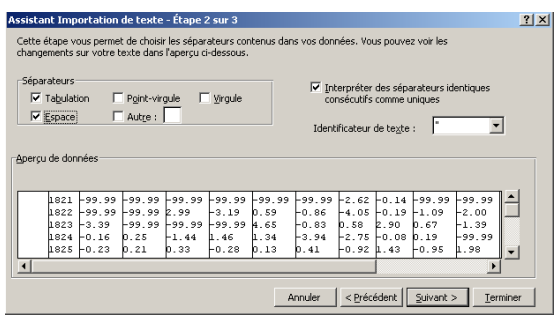
4) Effectuer le travail demandé section 3, 4 **ou** 5 **selon** le sujet choisi

5) Effectuer le travail demandé sections 6 **et** 7 **quelque soit** le sujet.

A noter, quelque soit le sujet, les données vous sont fournies dans un fichier texte. Pour ouvrir un fichier textuel sous Excel, il faut procéder de la façon suivante :

- Ouvrir le fichier texte avec Excel
- Dans la fenêtre qui s'ouvre, choisir l'option « délimité », puis cliquer sur « suivant ».
- Dans la fenêtre suivante, cocher « tabulation » et/ou « espace », puis cliquer sur « suivant ».
- Dans la troisième fenêtre, choisir le format des données (par défaut, standard est OK).





3. Sujet 1 : l'étude de l'oscillation Nord-Atlantique

3.1 Les données

Sur Internet, il est possible de récupérer des valeurs des indices NOA sur plusieurs années. Selon les sites, on obtient des valeurs annuelles ou mensuelles. Le formatage des données n'est pas forcément identique pour tous les fichiers. Par ailleurs, selon le mode de calcul, les valeurs obtenues peuvent être un peu différentes.

	A	B	C
1	année	indice 1	indice 2
2	1864	-1,02	
3	1865	-1,24	
4	1866	0,54	
5	1867	-1,38	
6	1868	2,81	
7	1869	1,70	
8	1870	-3,01	
9	1871	-1,01	
10	1872	-0,76	
11	1873	-0,50	-0,16
12	1874	2,32	0,6

Lorsque l'on récupère des données sur Internet, on est souvent amené à faire des traitements pour mettre les données dans un format adéquat pour la suite. Ici, nous nous sommes chargés de ce traitement. Vous avez à votre disposition un fichier Excel dans lequel on a 3 colonnes. La première indique l'année, la seconde indique les valeurs NOA récoltée sur un premier site et la troisième indique les valeurs récoltées sur un second site.

En ce qui concerne les données pluviométriques, Météo France nous a fourni 3 fichiers, indiquant la quantité de pluie récoltée sur 3 sites différents. Les données sont mensuelles et sont récoltées entre 1950 et 2008.

Les indices ONA étant annuels, il faut transformer les données pluviométriques mensuelles en données annuelles. Les exercices ci-après vous guideront vers une réalisation pas à pas.

ATTENTION : un des fichiers a des données manquantes. **Commencer** par traiter le fichier de **Chamonix**, qui est complet. Regrouper l'ensemble des données (ONA et pluviométrique) dans un seul fichier Excel avec plusieurs onglets.

Exercice 0. Afficher les années

Proposer un algorithme (contenant une itération) qui inscrit les années de 1950 à 2008 dans la colonne 5 (une année par ligne, à partir de la ligne 1). Traduire votre algorithme en une macro, vérifier vos résultats.

Exercice 1.

Proposer un algorithme qui fait la somme de 12 cellules de la colonne 3 comprises entre les lignes 2 et 13. On utilisera une itération et une variable intermédiaire Som. Pensez à initialiser Som **avant** l'itération. Copier la valeur de Som dans la première cellule de la colonne 6 **après** l'itération.
Traduire votre algorithme en une macro, vérifier vos résultats.

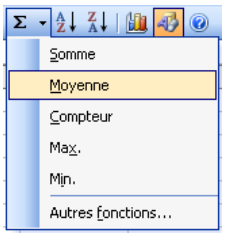
Exercice 2.

Il faut répéter ce traitement pour chaque paquet de 12 lignes.
Introduire une action paramétrée qui fait le même traitement que pour l'exercice 1. On choisira comme paramètre le numéro de la ligne de départ pour la somme et le numéro de la ligne où sera rangé le résultat.
Traduire votre algorithme en une macro, vérifier vos résultats avec quelques appels à l'action paramétrée.

Exercice 3.

Ecrire l'algorithme qui appellera l'action précédente pour chaque paquet de 12 lignes entre les lignes 2 et 709. Traduire votre algorithme en une macro, vérifier vos résultats.

Un autre traitement des données peut être utile. Les valeurs ONA sont calculées par rapport à une moyenne. Elles sont positives ou négatives par rapport à cette moyenne.
On se propose de calculer les différences de relevés pluviométriques par rapport à la moyenne de données (calculées pour chaque ville). Il faut donc (1) calculer la moyenne des données pluviométriques sur les années et (2) calculer la différence à la moyenne.

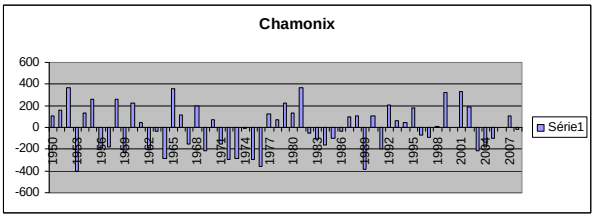


Pour calculer la moyenne des données pluviométriques, utiliser la fonction Excel.

- Sélectionner les données, plus une cellule.
- Dans le menu adéquat, choisissez d'effectuer une moyenne

Exercice 4. Différence à la moyenne
Ecrire l'algorithme pour calculer la différence entre la cellule colonne 6 et la moyenne. Ecrire le résultat dans la colonne 7 de la même ligne.
Traduire votre algorithme en une macro, vérifier vos résultats.

Exercice 5. Graphe de variation
Sélectionner les données « différences à la moyenne », dessiner un graphe représentant la variation des données. N'oubliez pas la légende.



Exercice 6
Exécuter les macros nécessaires pour mettre en forme les données sur chaque page contenant des données pluviométriques.

3.2 Vers une comparaison

Pour chacune des sources de valeurs ONA dans le fichier Excel fourni, produire à graphique similaire à ceux faits pour la pluviométrie.

Exercice 8. Comparaison visuelle.
Observer les différents graphes. Observez-vous une évolution similaire ? Vous pouvez aussi utiliser d'autres types de graphique (par exemple les nuages de points). Qu'en concluez-vous ?

Il n'est pas toujours très facile de décider s'il existe ou non une corrélation entre des données sur la seule base de l'observation visuelle. C'est pourquoi, il est intéressant d'utiliser des outils mathématiques pour prendre une telle décision.

Exercice 9. Comparaison à l'aide d'outil mathématique.
Reportez-vous à la section 8 pour calculer les différents coefficients de corrélation entre d'une part, des valeurs ONA et d'autres part des données pluviométriques. Attention à bien indiquer ce que vous comparez.

Exercice 10. Analyse
Que pensez-vous des résultats ?

3.3 Autre comparaison

On propose de mener une autre analyse permettant de comparer le bilan de masse glaciaire (gain ou perte de masse des glaciers d'octobre à septembre) sur les massifs alpins et l'indice ONA. Pour cela, on a mesuré les bilans sur 5 glaciers différents. On a calculé les indices ONA sur ces mêmes périodes.

Exercice 11. Analyse
Comparer les bilans de masse glaciaire et les indices ONA des périodes correspondantes selon les méthodes vues précédemment. Que pensez-vous des résultats ?

4. Sujet 2 : étude de la valeur-C[4]

4.1 Les données

Nous avons téléchargé des données depuis le site **«** Animal Genome Size Database[5] »**.** Elles sont dans le fichier *genome_size_data.txt*
Sur chaque ligne figure, le nom de l'espèce (entre les colonnes A et G), la valeur-C (colonne H) et le nombre de Chromosome). Les données ne permettent pas de répondre directement à la question car

- pour chaque espèce, plusieurs valeurs-C ont parfois été mesurées,
- le nombre de chromosomes n'est pas toujours indiqué.

Pour répondre à la question de recherche, il faut disposer les données de telle sorte que l'on puisse comparer pour chaque espèce, la valeur-C et le nombre de chromosomes.
La **première partie du travail vise à calculer la moyenne de la valeur-C pour chaque espèce. Les résultats seront écrits sur la seconde feuille 2 du fichier Excel. Il faut utiliser un algorithme car il y a de nombreuses espèces à considérer et le nombre de valeur-C mesurées est différent pour chaque espèce. Les exercices** 0 à 2 sont des exercices préparatoires pour calculer la moyenne de chaque espèce.

Exercice 0. Somme
Proposer un algorithme qui calcule la somme des valeurs qui se trouvent colonne 8 entre les lignes 8 et 11 (inclues). Le résultat sera écrit ligne 10, colonne 2, sur la feuille « Feuil2 ».
Pour désigner la cellule ligne 10 colonne 2 de « Feuil2 », on écrit Feuil2.cellule(10,2) dans un algorithme et Feuil2.Cells(10,2) dans une macro.
On introduira une itération (même pour seulement 4 lignes) car le travail va être utilisé dans la suite et une variable intermédiaire pour calculer la somme. Choisir soigneusement le type de cette variable.
Traduire votre algorithme en une macro, vérifier vos résultats.C réer la « Feuil2 » si elle n'existe pas.

Exercice 1. Moyenne.
Modifier l'algorithme précédent pour calculer la moyenne des valeur-C pour les valeurs-C entre les lignes 8 et 11. On écrira le résultat sur une ligne, colonne 2, sur la feuille « Feuil2 ».
Sur la colonne 1, même ligne, on écrira le « nom commun » de l'espèce.

Pour calculer la moyenne, il faut d'une part la somme des valeurs et d'autre part le nombre de valeurs. On utilisera une variable intermédiaire NbVal pour calculer le nombre de valeurs au fur et à mesure de l'itération.
Traduire votre algorithme en une macro, vérifier vos résultats.

Exercice 2. Moyenne pour un nombre indéfini de ligne
Pour chaque espèce, nous avons un nombre indéfini de valeur-C. On va faire la somme des valeur-C à partir d'un point donné, « tant que l'espèce ne change pas » (l'espèce est donnée colonne 6).
Proposer un algorithme avec une itération « **tant que** » pour calculer la moyenne des valeur-C à partir la ligne 12. On écrira le résultat sur une ligne, colonne 2, sur la feuille « Feuil2 ». Sur la colonne 1, même ligne 10 de la feuille « Feuil2 » on écrira l'espèce.
Traduire votre algorithme en une macro, vérifier vos résultats. Sauvegarder votre fichier avant d'exécuter la macro (les macros contenant un « tant que » contiennent souvent des erreurs !).

Il faut automatiser le traitement pour toutes les espèces. On procède en deux temps. On produit une version paramétrée du calcul de la moyenne puis on l'utilise pour faire le traitement sur toute la page.

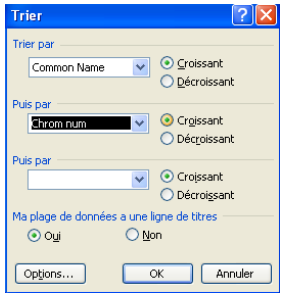
Exercice 3. Calcul de moyenne paramétrée
Proposer un algorithme paramétré `Calcul_moyenne`, ayant 2 paramètres : le numéro de la ligne de cellule du départ et le numéro de la ligne où doit être rangé le résultat. En plus du nom de l'espèce et

[4] Ce sujet a été inspiré du document http://pbil.univ-lyon1.fr/R_svn/pdf/bem4.pdf
[5] Gregory, T.R. (2011). Animal Genome Size Database. http://www.genomesize.com.

de la moyenne, l'algorithme `Calcul_moyenne` écrit le numéro de la ligne de la prochaine espèce à étudier en colonne 4 de la Feuil2.
Traduire votre algorithme en une macro.

Exercice 4
Écrire l'algorithme qui appellera l'action précédente pour traiter quelques lignes.
Traduire votre algorithme en une macro, vérifier vos résultats.

Exercice 5
Écrire l'algorithme qui appellera l'action de l'exercice 3 pour toute la feuille de calcul.
Traduire votre algorithme en une macro, vérifier vos résultats.

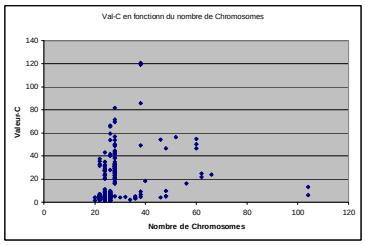


Exercice 6. Reporter le nombre de chromosomes sur la feuille 2.
Pour reporter le nombre de chromosomes sur la feuille 2, il faut savoir à quelle ligne le nombre de chromosome est indiqué pour chaque espèce. Comme on ne le sait pas, on va traiter les données pour que l'information soit sur la première ligne de chaque espèce. Une façon de faire consiste à trier les données avec la fonction de tri proposé par Excel.
Prendre soin de ne sélectionner les lignes comportant des données.
Trier d'abord par nom puis par nombre de chromosome (croissant). Lorsque le nombre de chromosome est connu, il sera alors affiché sur la première ligne de l'espèce.

Exercice 7 : Données à éliminer ?
Pour certaines espèces, il existe différentes valeurs pour le nombre de chromosomes. En utilisant la fonction tri précédente, on utilise par défaut la valeur la plus petite pour le nombre de chromosomes. Proposez un algorithme ou une solution pragmatique pour détecter et/ou éliminer les espèces ayant plusieurs valeurs différentes de chromosomes.
Cette analyse n'est pas nécessaire pour poursuivre le travail.

4.2 Vers une comparaison

Exercice 8. Comparaison visuelle.
Si la valeur-C augmente avec le nombre de chromosomes, on devrait observer visuellement l'augmentation sur un graphe. Tracer un graphe de type « nuage de points » permettant d'afficher la valeur-C en fonction du nombre de chromosomes. N'oubliez pas la légende. Vous devriez obtenir un résultat similaire au graphe ci-contre.



Il n'est pas toujours très facile de décider s'il existe ou non une corrélation entre des données sur la seule base de l'observation visuelle. C'est pourquoi, il est intéressant d'utiliser des outils mathématiques pour prendre une telle décision.

Exercice 9. Comparaison à l'aide d'outil mathématique.
Reportez-vous à la section 8 pour calculer le coefficient de corrélation entre les valeurs-C d'une part et le nombre de chromosomes d'autre part. Attention à bien indiquer ce que vous comparez.

Exercice 10. Analyse
Que pensez-vous des résultats ?

4.3 Autre comparaison

Il existe plusieurs sites web permettant d'obtenir la valeur-C. Par exemple, sur le site http://data.kew.org/cvalues/, nous avons téléchargé des données sur les végétaux et les avons sauvegardées dans le fichier *RBG_Kew_DNA_C-values.txt*.

Exercice 11. Analyse complémentaire
Adapter et appliquer les traitements précédents sur ces nouvelles données. Commenter vos résultats.

5. Sujet 3 : le taux de C+G

5.1 Les données

Dans un premier temps, nous souhaitons déterminer si le taux de C+G est significativement supérieur pour les bactéries, selon qu'elles sont aérobies ou anaérobies. Pour cela, nous avons récolté des données extraites de l'article : *Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. N. Galtier and J.R. Lobry. J Mol Evol, 44 :632–635, 1997*

Ces données ont été enregistrées pour vous dans le fichier *gc02.txt*, qui contient 4 variables :

1. nom du genre.
2. nom de l'espèce.
3. le taux de G+C du chromosome bactérien.
4. Type de bactérie : aérobie ou anaérobie (présence ou non d'oxygène).

Exercice 0. Somme

Pour chaque genre de bactérie, il faut estimer la moyenne des taux C+G. On procède en deux temps : d'abord on estime la somme des taux de C+G pour les bactéries d'un même genre. Puis, on calcule la moyenne du taux de CG de ce genre (en divisant par le nombre d'individus considérés).
Proposer un algorithme qui calcule la somme des valeurs qui se trouvent colonne 3 entre les lignes 2 et 5 (inclues). Le résultat sera écrit sur une ligne, colonne 2, de la feuille « Feuil2 ».
Créer la « Feuil2 » si elle n'existe pas (renommer au besoin).
Pour désigner la cellule ligne 10 colonne 2 de « Feuil2 », on écrit Feuil2.cellule(10,2) dans un algorithme et Feuil2.Cells(10,2) dans une macro.

On prendra soin d'introduire une itération car le travail va être utilisé dans la suite. On introduira aussi une variable intermédiaire pour calculer la somme. Choisir soigneusement le type de cette variable.
Traduire votre algorithme en une macro, vérifier vos résultats.

Exercice 1. Moyenne

Modifier l'algorithme précédent pour calculer la moyenne la moyenne des taux C+G entre les lignes 2 et 5. On écrira le résultat sur une ligne, colonne 2, de la feuille « Feuil2 ».
Sur la colonne 1, même ligne, on écrira le genre de bactérie.
Pour calculer la moyenne, il faut d'une part la somme des valeurs et d'autre part le nombre de valeur. On utilisera une variable intermédiaire NbVal pour calculer le nombre de valeurs au fur et à mesure de l'itération.
Traduire votre algorithme en une macro, vérifier vos résultats.

Exercice 2. Moyenne pour un nombre indéfini de ligne

Pour chaque genre de bactérie, nous avons un nombre indéfini de bactéries. On va faire la somme des taux C+G à partir d'un point donné, « tant que le genre de bactérie ne change pas ».
Proposer un algorithme avec une itération « **tant que** » pour calculer la moyenne des taux C+G à partir de la ligne 6. On écrira le résultat sur une ligne, colonne 2, sur la feuille « Feuil2 ». Sur la colonne 1, même ligne , de la feuille « Feuil2 » on écrira l'espèce.
Traduire votre algorithme en une macro, vérifier vos résultats. Sauvegarder votre fichier avant d'exécuter la macro (les macros contenant un « tant que » contiennent souvent des erreurs !).

Il faut maintenant automatiser le traitement pour tous les types de bactéries. On procède en deux temps. On produit une version paramétrée du calcul de la moyenne puis on l'utilise pour faire le traitement sur toute la page.

Exercice 3. Calcul de moyenne paramétrée

Proposer un algorithme paramétré `Calcul_moyenne`, ayant 2 paramètres : le numéro de la ligne de cellule du départ et le numéro de la ligne où doit être rangé le résultat.

En plus du nom de l'espèce et de la moyenne, l'algorithme `Calcul_moyenne` écrit :

- Le type de bactérie (aérobie ou anaérobie) colonne 3
- le numéro de la ligne de la prochaine espèce à étudier en colonne 6 de la Feuil2.

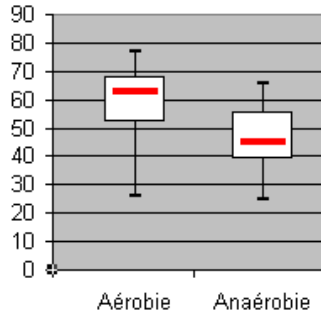
Traduire votre algorithme en une macro.

Exercice 4
Écrire l'algorithme qui appellera l'action précédente pour un type de bactérie.
Traduire votre algorithme en une macro, vérifiez vos résultats.

Exercice 5
Écrire l'algorithme qui appellera l'action précédente pour toute la feuille de calcul.
Traduire votre algorithme en une macro, vérifier vos résultats.

5.2 Vers une comparaison

On rappelle que l'on souhaite évaluer si le taux de C+G est modifié par la présence d'oxygène dans le milieu.
Une façon de comparer les taux de C+G consiste à tracer des « box-plots » (« boîtes à moustaches » en français) comme ci-contre.
La boîte à moustaches résume seulement quelques caractéristiques telles que la médiane, les quartiles, le minimum, et le maximum de la distribution.
Ce diagramme est utilisé principalement pour comparer un même caractère dans deux populations de tailles différentes[6].



Il n'existe pas de méthode Excel toute intégrée pour obtenir ce résultat, il faut passer par une étape intermédiaire, décrite section 9.

Exercice 6. Comparaison visuelle.
Tracez les box-plots des taux de C+G pour les bactéries aérobies et anaérobies. Observer le graphe obtenu. Que pouvez-vous en conclure ?

Il n'est pas toujours très facile de décider s'il existe ou non une corrélation entre des données sur la seule base de l'observation visuelle. C'est pourquoi, il est intéressant d'utiliser des outils mathématiques pour prendre une telle décision.

Exercice 7. Comparaison à l'aide d'outil mathématique.
Reportez-vous à la section 10 pour déterminer si les moyennes peuvent être considérées comme égales.

4.3 Autre comparaison

Le fichier *species.txt* a été téléchargé à partir du site ftp://pbil.univ-lyon1.fr/pub/datasets/ JME97/species. Comme précédemment, les trois premières colonnes décrivent le genre, l'espèce de bactérie et le taux de liaison C+G. La dernière colonne correspond à la température optimale de croissance de la bactérie exprimée en degrés Celsius.

On distingue trois grands groupes de bactéries en fonction de leur température optimale de croissance Topt : Topt < 20°C : bactérie psychrophile ; Topt > 45°C : bactérie thermophile ; 20°C ≤ Topt ≤ 45°C : bactérie mésophile.

Exercice 8. Formatage des données
Proposer un algorithme pour déterminer le type de bactérie en fonction de TOpt.
Traduire votre algorithme en une macro, vérifier vos résultats.

Exercice 9. Analyse complémentaire
Adapter et appliquer les traitements précédents pour déterminer si le taux de liaison C+G dépend du type de bactérie. Commenter vos résultats.

[6] http://fr.wikipedia.org/wiki/Boite_a_moustaches

6. Mini-site web (Pour tous les sujets)

On s'attend à ce que votre mini-site web comporte 4 pages html et une feuille de style.

- ***index.html*** : présente le contexte et votre hypothèse. A titre **exceptionnel**, vous pouvez récupérer des informations sur d'autres sites web et les copier-coller[7]. Mais vous devez clairement identifier vos sources et mettre des liens vers les pages que vous avez copiées.
- ***exp.html*** : présente votre expérience, c'est-à-dire
  - les données que vous avez utilisées (nature, format, origine),
  - une description des les traitements (principe des algorithmes) que vous avez faits.
- ***res.html*** : décrit les résultats.
- ***quizz.html*** : 2 à 5 questions de votre choix et un bouton qui permet de calculer le score.

Les exigences sur l'organisation du mini-site :

- Tous les fichiers doivent être regroupés dans un répertoire **ayant votre nom**.
- Ce répertoire principal doit comporter deux sous-répertoires : **images** et **data**.

- Les pages HTML doivent être dans le répertoire principal.
- Les pages ***exp.html***, ***res.html*** et ***quizz.html*** doivent être accessibles depuis ***index.html***.
- Chacune des ces pages doivent pouvoir retourner à ***index.html***.
- Il doit y avoir des liens qui permettent de passer directement de ***exp.html*** à ***res.html*** et vice-versa.
- Une feuille de style doit être utilisée pour toutes les pages HTML. Choisissez une feuille de style simple, qui permet une lecture facile de vos pages et qui change de la présentation par défaut (sans feuille de style).
-
- Le mini-site doit contenir des images.
- Toutes les images utilisées doivent être rassemblées dans le répertoire **images**.
- L'ensemble des fichiers de données brutes (fichiers textes fournis et ceux éventuellement trouvés) doivent être rassemblés dans le répertoire **data**.
-
- Des liens doivent permettre d'ouvrir ces fichiers depuis ***exp.html***.
- Le fichier Excel comportant vos macros, vos données traitées et vos résultats doit aussi être sauvegardé dans le répertoire data.
- Il doit être accessible depuis ***exp.html ou res.html***.

Les exigences sur le contenu du mini-site :

- La page ***index.html*** doit faire apparaître vos noms, prénoms et groupes.
- Votre mini-site doit comporter au moins une image, un tableau, une liste et une image map.
- Il doit y avoir des liens externes vers d'autres sites web que le vôtre.
- Il doit y avoir un lien interne.
- Il doit y avoir un bouton dynamique.

[7] Recopier des informations prévenant d'Internet ou de livres sans citer ses sources s'appelle du plagiat. L'UJF est particulièrement attentive à ce type de faute, qui peut être sanctionnée gravement.

7. Rendu du travail

Suivez les instructions de votre enseignant, qui pourra choisir entre récupération d'un fichier compressé par mail ou copie de votre répertoire sur une clef USB.

Créer un fichier compressé

Un fichier compressé est un moyen de rassembler différents répertoires et/ou fichiers dans un seul fichier, tout en compressant les données.

- Sur Sarado, sélectionner le répertoire qui contient les fichiers de votre mini-site.
- Cliquer sur ce répertoire avec le bouton droit de la souris,
- Sur le menu déroulant, choisir de compresser le répertoire.

Ce fichier compressé pourra vous être demandé par votre enseignant. **Vérifiez que les noms des membres du binôme apparaissent dans le nom du fichier**.

Le travail attendu pour le CC2 est donc de traiter les données qui vous sont fournies et de rédiger votre compte-rendu sur la forme d'un mini-site web. Cela correspond travail décrit dans les pages précédentes.

8. Outils mathématiques pour l'analyse de corrélation

Il n'est pas toujours très facile de décider s'il existe ou non une corrélation entre des données sur la seule base de l'observation visuelle. C'est pourquoi, il est intéressant d'utiliser des outils mathématiques pour prendre une telle décision.

Il existe plusieurs « outils » mathématiques qui permettent de décider si des ensembles de données sont corrélés (corrélation de Pearson par exemple) .Toutefois ces outils ne sont pas tous équivalents et doivent être utilisés dans des contextes appropriés.
Vous verrez ces outils dans vos UEs de statistiques ou de mathématiques.
Excel dispose d'une fonction pré-programmée qui lorsqu'on lui fournit deux ensembles de données, calcule un entier entre -1 et 1 (coefficient de corrélation) qui représente si les valeurs des 2 ensembles sont corrélées.

Pour utiliser cette fonction, choisissez une cellule vide dans laquelle vous souhaitez récupérer le coefficient de corrélation, puis dans le menu « insertion », choisissez « insérer une fonction » puis « coefficient de corrélation ».

Interprétation. Plus le coefficient de corrélation est proche de 1, plus les données sont fortement corrélées (une augmentation de l'un s'observe chez l'autre. Si le coefficient est proche de -1, les données sont négativement corrélées : lorsqu'une augmentation de l'un s'observe, une diminution s'observe chez l'autre). Si la valeur est proche de 0 (négative ou positive), on dit qu'il n'y a pas de corrélation.

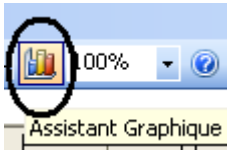
9. Tracer un box-plot en Excel

Pour tracer un diagramme « box-plot » (boîte à moustache), il faut calculer le minimum, le 1er quartile, la médiane, le 3ème quartile et le maximum pour chaque population (ici les ensembles de moyennes pour les bactéries aérobies et anaérobie). Pour calculer ces valeurs, on utilise la formule quartile. Cette formule a 2 paramètres : la plage de données et un entier indiquant le quartile. Le quartile 0 correspond à la valeur minimale de la plage de données, et quartile 4 correspond à la valeur maximale.

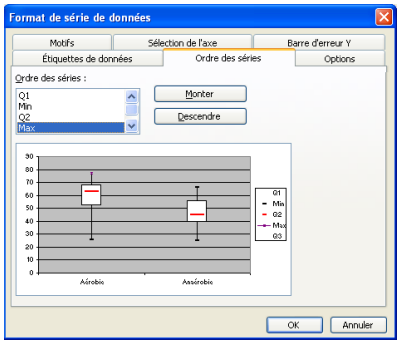
- Ecrire les formules suivantes dans 5 cellules d'une colonne vide (une formule par cellule). La plage de valeur (B2:B10 ci-après) doit être adaptée pour correspondre à vos données.
  =QUARTILE (B2:B10; 0)
  =QUARTILE (B2:B10; 1)
  =QUARTILE (B2:B10; 2)
  =QUARTILE (B2:B10; 3)
  =QUARTILE (B2:B10; 4)

	Aérobie	Anaérobie
Min	26	25
Q1	52,25	39,2307692
Q2	63	45
Q3	68	55,6666667
Max	77	66

  Sélectionner les données ainsi calculées et utiliser l'assistant graphique



  - Etape 1 : choisir « courbes » ; « courbe avec marques affichées à chaque points » ; « suivant »
  - Etape 2 :
    - Onglet « plage de données », choisir les série en lignes
    - Onglet « Série »; choisir les étiquettes des abscisses
  - Etape 3 : fixer les titres, supprimer la légende
  - Etape 4 : Insérer le graphique en tant que graphique
- Modifier le graphique de la façon suivante
  - Cliquer avec le bouton gauche de la souris sur la courbe représentant le min ; cliquer sur « format de la série » ; dans l'onglet « motif », choisir aucun trait et marque personnalisée (style trait, premier plan noir)
  - Faire le même traitement pour le max.
  - Pour Q1 et Q3, choisir aucun trait, aucune marque
  - Pour la médiane (Q2),
    - Dans l'onglet « motif », choisir aucun trait, marque personnalisée (style trait, premier plan rouge, taille 25 points)
    - Dans l'onglet « options », cocher lignes haute/bas, et barres haut/bas ; ajuster « largeur intervalle » pour obtenir une boite de la même taille que le trait rouge.
    - Dans l'onglet « ordre des séries », mettre dans l'ordre Q1, Min, Q2, Max, Q3 en utiliser les boutons « monter » et « descendre »



On obtient alors le résultat souhaité.

10. Outils mathématiques pour la comparaison de distribution

Il existe plusieurs « outils » mathématiques qui permettent de comparer des ensembles de données pour décider si elles appartiennent à la même distribution. Toutefois ces outils ne sont pas tous équivalents et doivent être utilisés dans des contextes appropriés.
Vous verrez ces outils dans vos UEs de statistiques ou de mathématiques.

Compte-tenu de la nature de nos échantillons, il faudrait utiliser le test de Wilcoxon. Mais, la version Excel dont nous disposons ne propose pas ce test. A titre d'exercice, on se propose de faire un test de Student, qui permet la comparaison d'échantillon sous certaines conditions.
Choisissez une cellule vide dans laquelle vous souhaitez récupérer la valeur du test de Student, puis dans le menu « insertion », choisissez « insérer une fonction » puis « TEST.STUDENT ». Pour les 4 paramètres, indiquer les plages des deux échantillons, les valeurs 2 et 3.

Interprétation. De façon très intuitive, lorsque l'on effectue un test de Student, on fait l'hypothèse que les deux distributions comparées sont de même loi. Le test de Student permet de calculer une valeur appelée « *p-value* ». Cette valeur représente la probabilité de se tromper en rejetant l'hypothèse. Ainsi, si la p-value est inférieure à un certain seuil (typiquement 5% ou 1%), on peut rejeter l'hypothèse avec la probabilité de se tromper inférieure à ce seuil.