



# Module de Visualisation De Calculs Statistiques

---

Projet 1<sup>ère</sup> année  
AIR1 – 2013/2014



**Abdessatar BOUGUILA**

**Amadou-Diaw FALL**

**Laye-Mory KEITA**

**1<sup>ère</sup> année Informatique & Réseaux par apprentissage**

Tuteur / Client : **Nistor GROZAVU**

Année académique 2013/14

Villetaneuse, le 16 Juin 2014



## REMERCIEMENTS

Avant de commencer, je me permets d'adresser mes remerciements aux personnes qui ont participé à la réussite de ce projet ainsi qu'à la rédaction de ce rapport.

Tout d'abord, merci à **M. Nistor GROZAVU**, notre enseignant suiveur ainsi que notre client pour ce projet, pour son suivi et ses conseils pour la réalisation des travaux.

Ensuite merci à **M. Henry SOLDANO** d'avoir accepté de superviser notre présentation orale.

Merci également à **Mme Chantal WOLEZYK** pour toutes les connaissances qu'elle nous transmet à travers nos exercices oraux et écrits. Connaissances qui ont eu leur utilité pour la rédaction de ce rapport et qui continuent à influencer très positivement notre quotidien professionnel.

# SOMMAIRE

<b>REMERCIEMENTS.....</b>	<b>2</b>
<b>SOMMAIRE .....</b>	<b>3</b>
<b>INTRODUCTION .....</b>	<b>4</b>
<b>I. PRESENTATION DE L’EXISTANT .....</b>	<b>5</b>
I.1. Les choix techniques.....	5
I.2. Le Module Statistique en place .....	6
<b>II. LE CAHIER DES CHARGES .....</b>	<b>8</b>
II.1. Spécifications et objectifs du client .....	8
<b>III. Méthodologie de travail .....</b>	<b>11</b>
III.1. Modèle de cycle de développement logiciel.....	11
III.2. Avant-Projet.....	11
III.3. Suivi du projet.....	14
III.4. Clôture du projet .....	14
<b>IV. TRAVAIL REALISE.....</b>	<b>15</b>
IV.1. Implémentation formules .....	15
IV.2. Implémentation graphique .....	16
<b>V. Conclusion générale – Bilan et perspectives .....</b>	<b>18</b>

# INTRODUCTION

Dans le cadre de notre première année, nous sommes amenés à réaliser un projet informatique, dont le thème est proposé par un de nos enseignants. L'idée est de nous permettre de mener un projet dans son ensemble, de sa phase de conception à sa réalisation finale. Notre projet a été proposé par M. GROZAVU, enseignant chercheur à l'Université Paris 13 et responsable de la formation Informatique & Réseaux en Apprentissage.

Il consiste à continuer un projet entamé l'année dernière, et se fixe comme objectif de fournir à son terme une application Desktop permettant de faire des calculs statistiques à partir d'un jeu de données et les représenter graphiquement. Le préexistant dispose d'une interface graphique interactive et propose les fonctionnalités suivantes :

- Un module permettant d'effectuer des calculs statistiques
- Une lecture de fichiers normalisés avec gestion des données manquantes
- Un module de visualisation permettant de faire une représentation graphique des calculs statistiques sous forme d'histogramme et de nuages de points.

L'objectif principal qui a été fixé pour la reprise de ce projet est de rajouter de nouvelles formules de calculs statistiques à l'application déjà en place. Ce rajout entrainera par la suite la mise à jour de toutes les autres fonctionnalités (représentations graphiques, etc.).

Le but de ce rapport est donc de suivre l'avancée de nos travaux tout au long de la réalisation de nos objectifs. Dans un premier temps nous étalerons rapidement les caractéristiques ainsi que les fonctionnalités de l'existant, puis nous présenterons les objectifs qui sont fixés pour la continuité du projet. Ensuite, nous évoquerons la méthodologie de gestion de projet qui a été adoptée ainsi que le planning qui a été dressé. Pour finir, nous présenterons le travail réalisé ainsi que le bilan et les perspectives de ce projet.

# I. PRESENTATION DE L'EXISTANT

Des étudiants de la promotion AIR1 de l'année 2012-2013 ont proposé dans le cadre de leur projet une première version de cet outil à développer. Dans ce chapitre, nous proposerons une globale description de l'outil qui a été réalisé l'an dernier. Ceci sous-entend donc une description des choix techniques qui ont été faits (outils de développement, langages de développement, différentes bibliothèques, etc.), par la suite nous décrirons les fonctionnalités ainsi que les formules de calculs statistiques qui sont déjà proposées.

## I.1. Les choix techniques

- Plateforme et langages de programmation :

L'application déjà en place a été développée suivant un modèle orienté objet basé sur le langage Java et utilisant Eclipse comme plateforme de développement. Ce choix a été motivé par la facilité de structuration et de compréhension des programmes en Java, leur facilité de maintenance et de réutilisation ainsi que la possibilité d'utiliser des Frameworks et des plugins qui facilitent de travail via Eclipse.

- Librairies de développement :

Pour satisfaire toutes les fonctionnalités de l'application ainsi que la mise en place d'une interface graphique interactive, nos prédécesseurs ont opéré des choix concernant les librairies utilisées pendant la phase de développement. Parmi ces choix, les plus importants sont :

- **SWT pour l'aspect graphique (à la place de SWING)**

Deux choix s'offraient à nos prédécesseurs : SWING et SWT qui sont tous les deux des plateformes de développement d'interface graphique.

SWING est une API qui s'appuie sur AWT et Java2D pour mettre en place un ensemble de composants d'interface graphique sophistiqué et extensible. Elle a pour avantages d'offrir un environnement totalement indépendant du système d'exploitation utilisé et une grande facilité de déploiement. Malheureusement elle n'est graphiquement pas très jolie et connaît quelques lenteurs à l'exécution.

Quant à SWT, tout comme SWING, elle offre un ensemble de composants d'interface graphique sophistiqué et extensible. Cependant, elle a la particularité de s'appuyer de manière beaucoup plus forte sur le framework natif de l'environnement graphique du système d'exploitation sur lequel elle est exécutée. Elle propose une interface beaucoup plus esthétique et offre des meilleurs temps de réponses que SWING à l'exécution. Parmi ses inconvénients, nous pouvons évoquer sa forte dépendance au système d'exploitation, une

certaine difficulté de déploiement, et une rareté des exemples et de la documentation.

Dans un soucis de prioriser la rapidité d'exécution de l'application et l'utilisation de ressource machines moins importante ainsi qu'un rendu parfait des composants graphiques selon l'OS utilisé, le choix s'est porté sur SWT.

- **iText pour l'exportation en PDF**

iText est une simple librairie Java à inclure dans une application. Il permet de créer ou de manipuler un document au format PDF. ÷

- **JFreeChart pour la création des graphiques**

Il s'agit d'un framework open-source pour le langage Java. Il permet la création de graphiques complexes et esthétiques de manière assez simple. Il offre ainsi la possibilité de créer les types de diagrammes suivants : Diagramme de Camembert, Diagramme de Gantt, Histogrammes, etc.

## **I.2. Le Module Statistique en place**

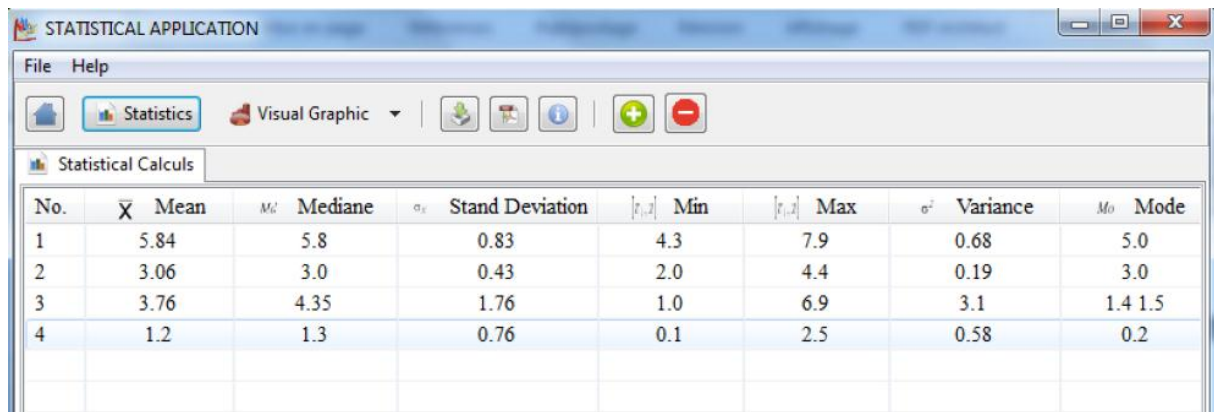
### **1. Formules statistiques :**

Le module statistique mis en place permet, après lecture des données à partir d'un fichier renseigné par l'utilisateur, d'afficher les résultats des calculs statistiques suivants :

- La moyenne : Il s'agit du quotient de toutes les valeurs d'une colonne de notre fichier de données sur l'effectif total de la colonne.
- La médiane : C'est une valeur  $m$  qui permet de couper notre ensemble de données en deux parties égales. Ces parties mettent d'un côté une moitié des valeurs, toutes inférieures ou égales à  $m$  et de l'autre côté l'autre moitié des valeurs, toutes supérieures ou égales à  $m$ .
- La variance : Il s'agit d'une valeur qui sert à indiquer de quelle manière notre série de données statistiques se disperse autour de sa moyenne.
- L'écart-type : C'est la racine carrée de la variance.

En plus de ces formules, l'application donne la possibilité à l'utilisateur de retrouver les valeurs maximales et minimales d'une série de données, ainsi que la valeur qui apparait le plus dans la série.

L'application dont nous avons hérité permet après import du fichier de données de calculer et d'afficher les résultats des formules statistiques énoncées ci-dessus. Le résultat est affiché de la manière suivante :



The screenshot shows a software window titled 'STATISTICAL APPLICATION'. It has a menu bar with 'File' and 'Help'. Below the menu bar is a toolbar with icons for 'Statistics', 'Visual Graphic', and several other functions. A 'Statistical Calculs' tab is active, displaying a table of statistical results for four data points (No. 1 to 4).

No.	$\bar{X}$	Mean	$M_o$	Mediane	$\sigma_x$	Stand Deviation	$ x_i - \bar{x} $	Min	$ x_i - \bar{x} $	Max	$\sigma^2$	Variance	$M_o$	Mode
1	5.84		5.8		0.83		4.3		7.9		0.68		5.0	
2	3.06		3.0		0.43		2.0		4.4		0.19		3.0	
3	3.76		4.35		1.76		1.0		6.9		3.1		1.4	1.5
4	1.2		1.3		0.76		0.1		2.5		0.58		0.2	

Figure 1 - Affichage des résultats

## 2. Le fichier de données :

Les données renseignées par l'utilisateur sont contenues dans un fichier et sont rangées en colonnes séparées par des virgules. Quand il s'agit de nombre décimaux, le point est utilisé à la place de la virgule.

En effet, entrée valide de notre fichier de données aura la forme suivante :

**5.1,3.5,1.4,0.2,Iris-setosa**

Ci-dessous des exemples de lignes non-valides :

**5.1 :3.5 :1.4 :0.2:Iris-setosa**

**5.1;3.5;1.4;0.2;Iris-setosa**

Les données seront donc des nombres décimaux et la dernière colonne comportera un texte pour décrire le type des données.

Exemple du contenu d'un fichier de données valide

**5.1,3.5,1.4,0.2,Iris-setosa**  
**4.9,3.0,1.4,0.2,Iris-setosa**  
**4.7,3.2,1.3,0.2,Iris-setosa**  
**4.6,3.1,1.5,0.2,Iris-setosa**  
**5.0,3.6,1.4,0.2,Iris-setosa**  
**5.4,3.9,1.7,0.4,Iris-setosa**  
**4.6,3.4,1.4,0.3,Iris-setosa**  
**5.0,3.4,1.5,0.2,Iris-setosa**  
**4.4,2.9,1.4,0.2,Iris-setosa**  
**4.9,3.1,1.5,0.1,Iris-setosa**  
**5.4,3.7,1.5,0.2,Iris-setosa**  
**4.8,3.4,1.6,0.2,Iris-setosa**



### 3. Module de visualisation :

Le module de visualisation permet d'avoir une représentation graphique sur les calculs statistiques effectués. Ce module permet de tracer des nuages de points ou des histogrammes.

#### a. Les histogrammes

Ils permettent d'avoir une vue sur la répartition des données. Ils sont construits en fonction des colonnes qui constituent le fichier importé. Ainsi, on représente en ordonnée les données et en abscisse les indices des données à représenter.

Nous pouvons choisir la colonne que nous voulons visualiser ainsi que le nombre d'entrées à prendre en compte. Nous avons la possibilité d'enregistrer un histogramme sous format PNG et aussi de réajuster les échelles de l'abscisse et de l'ordonnée. Ci-dessous une capture d'écran du module de visualisation lors de l'affichage d'un histogramme.

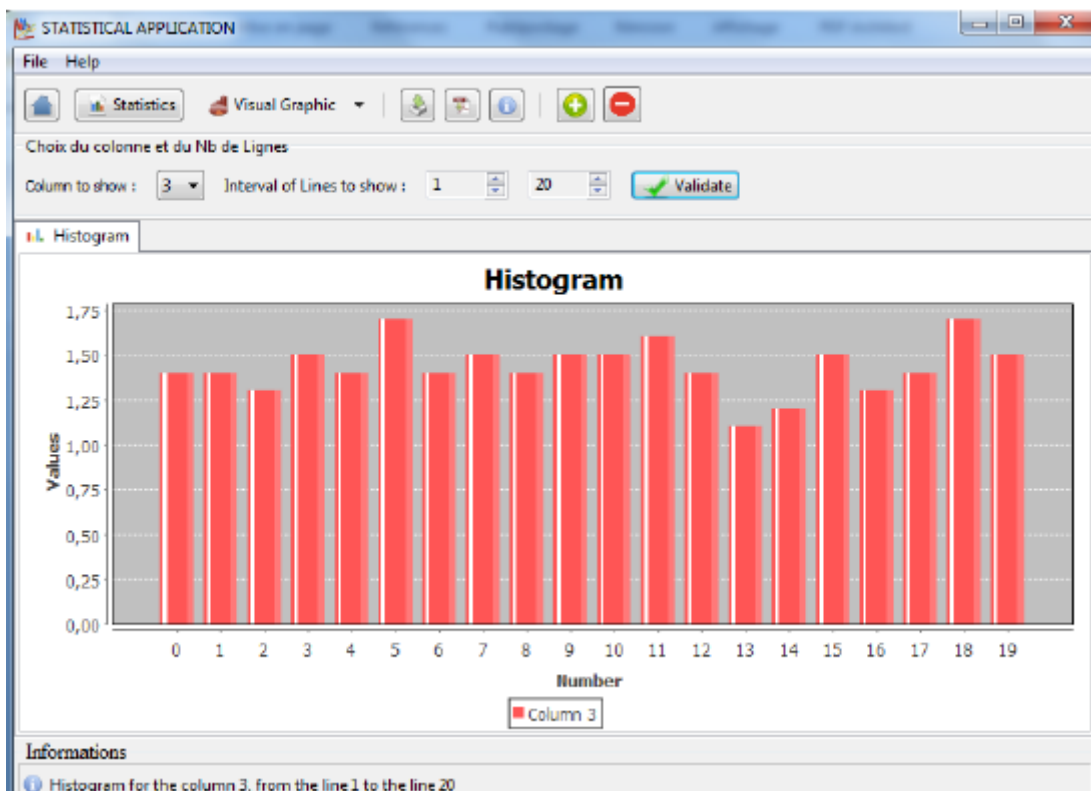


Figure 2 - Module de visualisation: histogramme

## II. LE CAHIER DES CHARGES

Dans le long terme, le but est de permettre à l'enseignant ayant proposé ce sujet d'avoir sous la main un outil complet d'étude statistique. Dans ce sens nous récupérons le travail déjà entamé par les actuels AIR2, de le faire progresser, de mettre à jour la documentation pour ainsi permettre aux futures promotions de pouvoir travailler dessus.

Une toute première réunion après notre réponse à l'appel d'offre a été faite et a abouti à la déclinaison des spécifications et des objectifs attendus par le client.

### II.1. Spécifications et objectifs du client

Les premiers objectifs fixés ont été de rajouter de nouvelles formules statistiques à

l'application. Parmi ces nouvelles notions statistiques évoquées, les principales qui ont été retenues sont : la Corrélation et la Régression Linéaire Simple.

Le chapitre suivant expose les définitions, les caractéristiques ainsi que les formules de ces notions énoncées précédemment.

### 1. La corrélation :

En statistique, étudier la corrélation entre deux ou plusieurs variables numériques, c'est étudier l'intensité de la liaison qui peut exister entre elles.

La mesure de la corrélation linéaire entre deux variables se fait par le calcul du coefficient de corrélation linéaire. Ce coefficient s'obtient en faisant le rapport de leur covariance et du produit non-nul de leurs écarts types. Le résultat est une valeur comprise entre -1 et 1. Il s'interprète de la manière suivante :

- Lorsqu'il est égal à 1, l'une des variables est une fonction affine croissante de l'autre variable.
- Quand il vaut -1, une des variable est une fonction affine décroissante de l'autre.
- Les valeurs intermédiaires renseignent sur le degré de dépendance linéaire entre les deux variables étudiées. Plus le coefficient est proche des valeurs extrêmes (-1 et 1), plus la corrélation entre les variables est forte ; on emploie simplement l'expression « fortement corrélées » pour qualifier les deux variables.
- Une corrélation nulle signifie, on dit que les deux variables ne sont pas corrélées et n'entretiennent aucune relation fonctionnelle particulière.

Mathématiquement, le coefficient de corrélation s'obtient de la manière suivante :

$$r = \frac{\sum (xi - X) \cdot (yi - Y)}{\sqrt{\sum (xi - X)^2} \cdot \sqrt{\sum (yi - Y)^2}}$$

où :

- r est le coefficient de corrélation
- xi est une valeur de la première distribution,
- yi une valeur de la deuxième distribution,
- X la moyenne de la première distribution,
- Y la moyenne de la deuxième distribution.

L'objectif demeure donc d'implémenter cette formule, et d'être en mesure d'obtenir un fichier de données qui contient les coefficients de corrélation des variables contenu dans le premier jeu de données renseigné par l'utilisateur et de l'étudier (Visualisation des points de

corrélation, affichage d'une matrice de couleur en fonction des valeurs, etc.).

## 2. La Régression Linéaire Simple :

La Régression Linéaire Simple permet de chercher l'éventuelle relation fonctionnelle linéaire qui existerait entre une valeur indépendante  $x$  et une variable aléatoire dépendante  $y$ . Graphiquement, on représente cette éventuelle relation en se munissant d'un axe des abscisses pour représenter la variable indépendante, et d'un axe des ordonnées  $y$  pour  $y$  représenter ce que l'on cherche à expliquer (variable dépendante).

Une relation linéaire déterministe entre les deux variables se traduit par des points parfaitement alignés. Les deux variables sont ainsi liées par une fonction affine du type  $y=ax+b$ . La valeur «  $a$  » correspond au coefficient de régression et «  $b$  » est la constante de régression. Elles s'obtiennent de la manière suivante :

$$a = \frac{\sigma_{xy}}{\sigma^2_x} \text{ et } b = \bar{y} - a\bar{x}$$

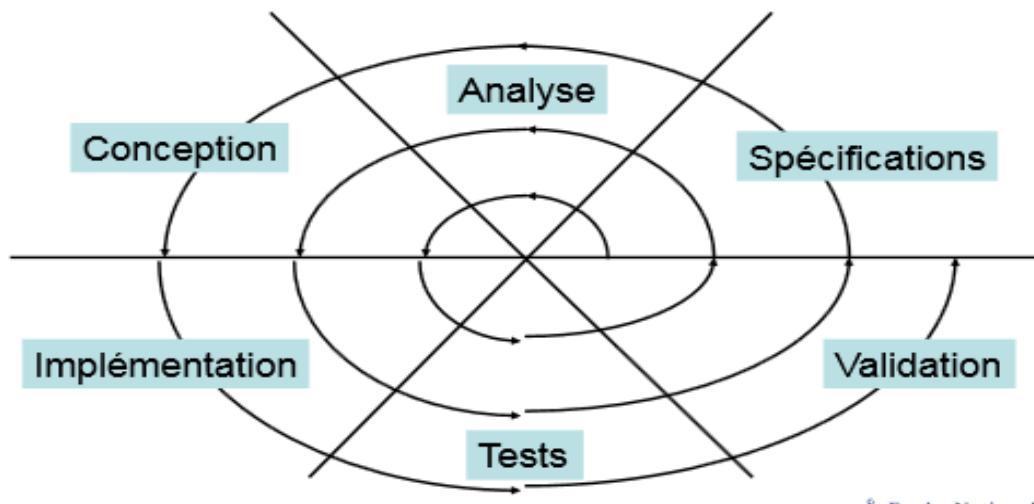
### III. Méthodologie de travail

Afin de gérer au mieux le projet et procéder d'une manière structurée en essayant de suivre un modèle de gestion de projet bien défini. On a essayé d'organiser les tâches en exploitant les points forts de chacun pour tirer un maximum d'efficacité. Cependant, les notions techniques nécessaires sont toutes nouvelles donc on a eu quelques difficultés à démarrer le projet.

#### III.1. Modèle de cycle de développement logiciel

Le modèle utilisé pour réaliser ce projet est le modèle en Spirale. On a opté pour ce modèle d'une part pour la simplicité offerte et aussi, car c'est un modèle qui nous permet de mettre l'accent sur l'évaluation des risques et il nous permet, à chaque étape, après avoir défini les objectifs et les alternatives de réaliser l'étape et la validé.

#### Modèle en spirale



#### III.2. Avant-Projet

Après avoir identifié les tâches à réaliser, il fallait organiser le groupe en répartissant les tâches et fixer un calendrier avec les échéances et les niveaux d'avancement.



Figure 3 - Organisation du projet

#### Structurelle :

- Identifier les travaux à compléter
- Traduire la définition du projet en une liste de tâches à accomplir
- Identification et description des lots de travail principaux
- Identification et description des tâches élémentaires

#### Opérationnelle :

- Toute tâche est assignée à une personne
- Tout participant est informé de :
  - ses rôles et responsabilités
  - son degré d'autonomie et d'autorité
  - des rôles et responsabilités des autres
- Données de départ :
  - Organigramme technique
  - Processus de développement

#### Budgétaire :

Dans notre cas, le budget est le temps. On a donc effectué une estimation de l'effort, la durée nécessaire et de la taille du projet.

Voici le diagramme de Gantt résultant :

## Gantt

T1 : Création des groupes

T2 : Prise de connaissance du projet

T3 : Réunion avec enseignant encadrant

T4 : Réunion d'équipe-récupération des sources

T5 : Réunion non-prédécesseurs -Compréhension fonctionnelle du projet

T6 : Autoformation sur les technologies utilisées : SWT, JAVA, GWT, etc.

T7 : Réunion avec équipe année dernière - Compréhension de l'organisation structurelle du code source

T8 : Réunion d'équipe : Partage des tâches Compréhension techniques de l'existant

T9 : Réunion avec enseignant encadreur - Prise des spécifications du client

T10 : Documentation sur les formules à intégrer - Rédacteur cahier des charges

T11 : Implémentation des nouvelles fonctionnalités

T12 : Réunion de bouclage

T13 : Livraison -Préparation soutenance - Finalisation rapport

### Séance = 3h

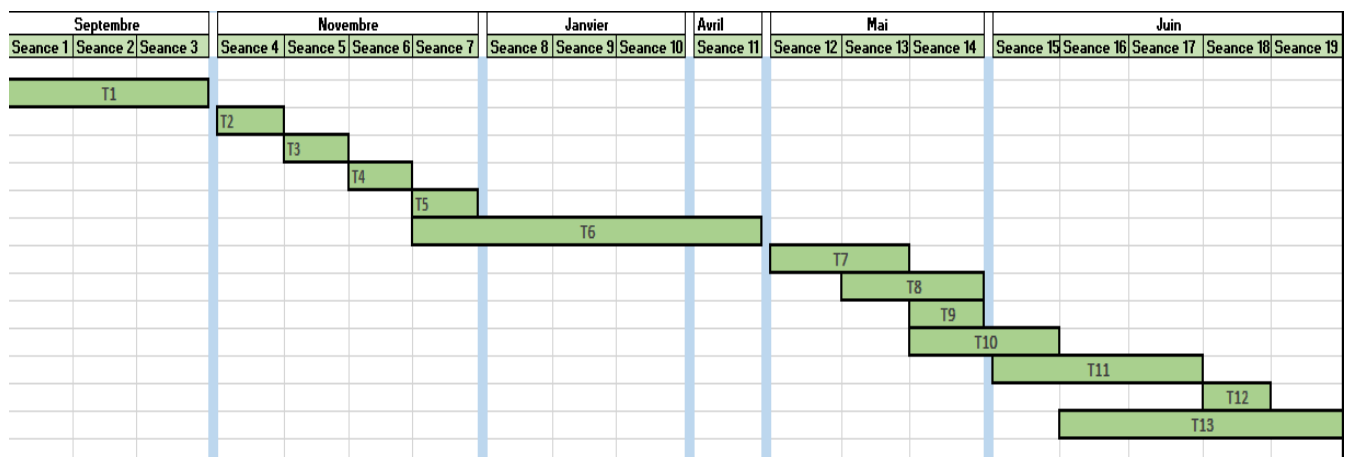
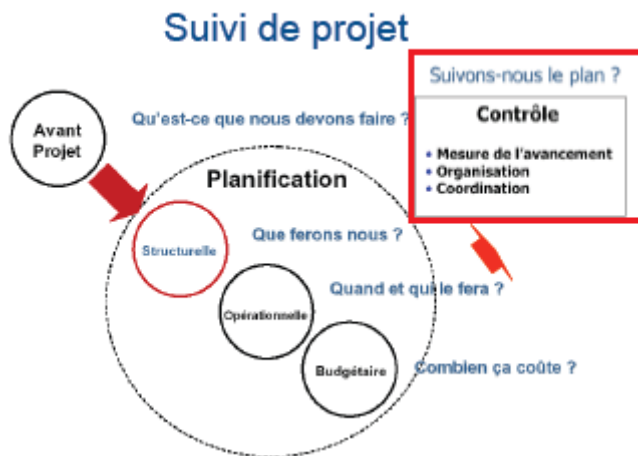


Figure 4 - Diagramme de Gantt

### III.3. Suivi du projet



Un "journal de bord" est tenu à jour. Il permet de garder une trace :

- des informations communiquées
- des problèmes rencontrés
- des décisions prises
- des responsables désignés pour mener à bien les actions
- la date de réalisation de l'action

### III.4. Clôture du projet

Inévitablement le projet arrive à échéance. Après avoir intégré et validé les fonctions demandées, le projet se termine.

## IV. TRAVAIL REALISE

### IV.1. Implémentation formules

Dans un premier temps, toutes les formules nécessaires ont été traduites en Java. Il s'agit globalement d'appliquer les formules énoncées dans les parties précédentes du rapport à notre fichier de données. Ensuite, d'autres méthodes ont été parallèlement pour convertir le résultat des calculs en chaîne de caractère. En effet lors de l'affichage au niveau de l'interface graphique, un format String des résultats est utilisé.

Nous vous proposons ci-dessous les deux méthodes qui s'occupent de l'implémentation du calcul du coefficient de corrélation :

```
public static double coefCorel (ValuesList listData1, ValuesList listData2){
    double coefCorel=0;
    int i;
    double moy1=0;
    double moy2=0;
    double num=0;
    double denum1=0;
    double denum2=0;
    double denum;
    moy1=mean(listData1);
    moy2=mean(listData2);
    for (i=0;i<listData1.size();i++){
        num = num + (listData1.get(i)-moy1)*(listData2.get(i)-moy2);
        denum1 = denum1 + (Math.pow(listData1.get(i)-moy1,2));
        denum2 = denum2 + (Math.pow(listData2.get(i)-moy2,2));
    }
    denum = Math.sqrt(denum1)*Math.sqrt(denum2);
    coefCorel=num/denum;
    return roundDouble(coefCorel);
}
```

Figure 5 - Méthode calcul coefficient corrélation

```
public String getCoefCorl(int col1, int col2){
    String result = "";

    ValuesList values1 = new ValuesList();
    ValuesList values2 = new ValuesList();

    for (int j=0; j < io.getValuesList(col1).size();j++){
        values1.add(io.getValuesList(col1).get(j));
    }

    for (int j=0; j < io.getValuesList(col2).size();j++){
        values2.add(io.getValuesList(col2).get(j));
    }

    result = ""+CalculStats.coefCorel(values1, values2);

    return result;
}
```

Figure 6 - Méthode calcul coefficient corrélation bis



## IV.2. Implémentation graphique

Deux nouveaux boutons ont été ajoutés : le bouton « **Corrélation** » et le bouton « **Regression** ».

- Le bouton « Corrélation » gère la fonction qui calcule le coefficient de corrélation et affiche le résultat sous forme de matrice de couleurs. Nous avons choisi des couleurs pour représenter chaque intervalle de valeur compris entre -1 et 1. Ainsi nous avons convenu de prendre les couleurs suivantes :
  - Vert foncé : Coefficient de corrélation égal à 1.
  - Gris : Valeur égale à 0.
  - Bleu : Valeur égale à -1.
  - Jaune : Valeur comprise entre 0 et 0.5.
  - Vert clair : Coefficient de corrélation compris entre 0.5 et 1.
  - Cyan foncé : Valeur comprise entre -1 et -0.5.
  - Cyan clair : Valeur comprise entre -0.5 et 0.0

Importons un fichier de données en guise d'exemple et observons la matrice de corrélation affichée :

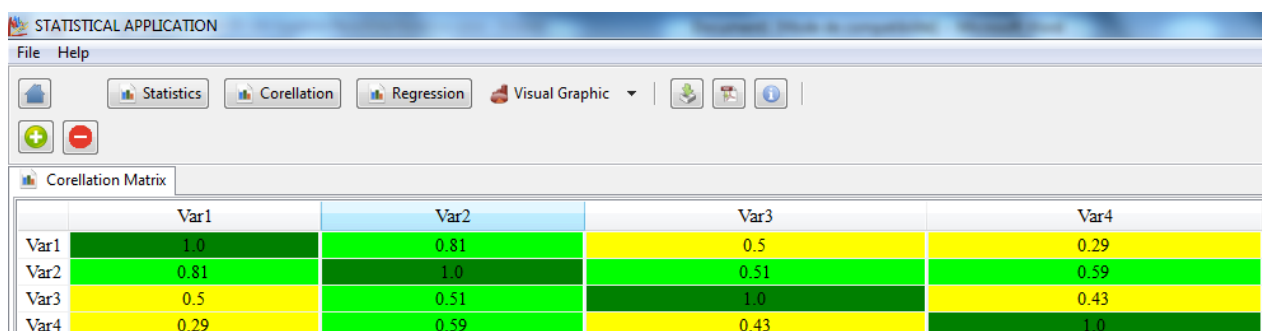


Figure 7 - Affichage matrice de corrélation

La particularité d'une matrice de corrélation est que sa diagonale est à 1, car le coefficient de corrélation d'une variable avec elle-même est égal à 1 (Une variable est très fortement corrélée à elle-même). Nous pouvons observer cette remarque dans notre matrice ci-dessus.

- Le bouton « Régression » lui gère la représentation de la droite de régression. Lorsque nous importons notre fichier de données et que nous cliquons sur le bouton, nous sommes appelés à choisir les colonnes X et Y que nous voulons

étudier en indiquant le numéro de ces colonnes (1, 2, 3 ou 4). Pour rappel, la droite de régression a une équation du type  $Y=AX+B$ , avec A et B des coefficients calculés.

L'utilisateur choisit donc colonnes à traiter grâce au module graphique suivant :

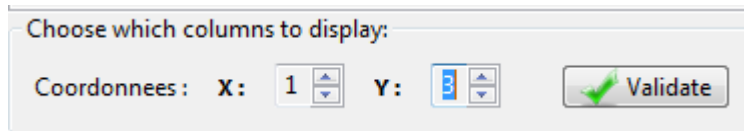


Figure 8 - Choix des colonnes à afficher

Une fois que l'utilisateur valide son choix, il le valide et obtient sa droite ainsi que l'équation de la droite :

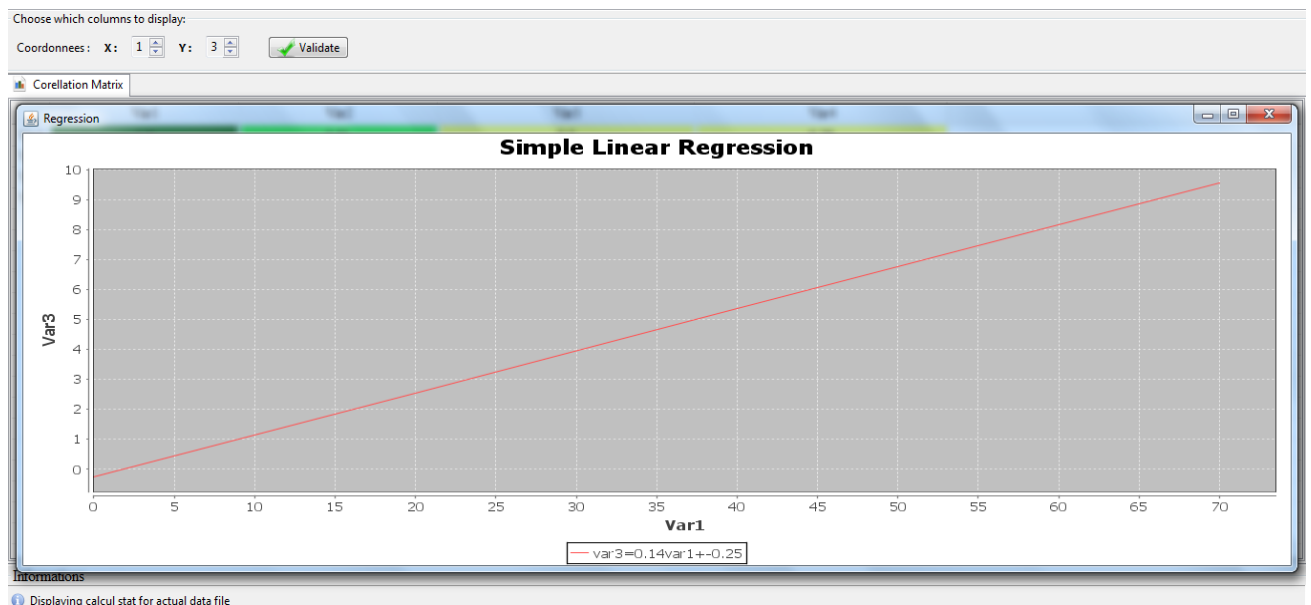


Figure 9 - Affichage droite de régression

## V. Conclusion générale – Bilan et perspectives

A l'issue de ce projet, le bilan est satisfaisant concernant l'organisation du travail, les aspects techniques et la qualité du travail fourni.

Concernant l'organisation du travail, le travail d'équipe a été bien soudé dès le début du projet. Ce qui, par ailleurs, nous a permis de gagner en compétences dans les aspects de communication, notamment :

- Une bonne circulation d'information dans l'équipe
- Des réunions fréquentes avec d'éventuelles réévaluations des tâches en fonction de l'avancement du projet et des compétences de chacun.
- Identification progressive des grandes améliorations (et mise au courant du client de la situation).

Au niveau technique, ce projet nous a permis :

- D'effectuer une autoformation sur de nouvelles technologies du langage en vue de la bonne prise en main du « préexistant » et de la conduite du projet.
- De respecter les délais (avec analyse fréquente des causes des écarts de temps par rapport aux prévisions effectuées sur les délais du projet).
- D'analyser divers choix techniques et d'évaluer l'adéquation des moyens mis en œuvre.
- Propositions de nouvelles orientations en vue de l'amélioration du référentiel qualité- méthodes.

Ce projet nous a ainsi permis à tous les trois de mettre en pratique des méthodologies de gestion de projet, et de mener le premier projet pédagogique de notre cycle. De même, il nous aura permis de passer du temps de pratique sur le langage Java, qui occupe une place importante dans notre formations et nos projets professionnels respectifs.

Conscients de l'imperfection de l'application, nous nous sommes assurés, en adoptant de bonnes méthodologies de développement, de la portabilité et de la possibilité de réutilisation des sources du projet en vue d'une amélioration future dans le but de fournir sur le long terme un outil d'étude statistique très performant.