Traitement et analyse de données biologiques (DLSV214, 3 ECTS)

Compétences

L’objectif de l'enseignement est de former les étudiants au **traitement et à l’analyse des données biologiques** acquises au laboratoire, sur le terrain ou en entreprise et à leur **représentation graphique**.

Il s’agit d’apprendre à exploiter au mieux les données biologiques dans le **contexte du *Big Data***.

Enseignement axé sur les **statistiques descriptives**, **le modèle linéaire** et l’**analyse factorielle.**

Description

- Introduction sur le Big Data en Biologie
- La **représentation graphique**
- **Gestion des erreurs**
- Comment résumer une distribution expérimentale à l'aide de quelques paramètres (position, dispersion, forme)
- Réalisation de **documents de synthèse de qualité professionnelle**
- Initiation à l’**analyse factorielle**
- Le **modèle linéaire** : régression linéaire et analyse de variance

Volume horaire : 25h => Cours : 15h , TD : 10h

Responsable : Pascal RIGOLET, pascal.rigolet@u-psud.fr

Retours d'expériences :L'UE « Traitement et Analyse de Données en Biologie »

Comment apprendre à des étudiants à utiliser l'outil statistique en pleine conscience ?

Réfléchir avant d'agir – Travail en groupe encouragé

Enseignement fondé sur le couple savoirs + savoir faire

Redonner confiance aux étudiants et leur apprendre la prise de risque

<u>L'UE en quelques mots et en quelques chiffres :</u>

- Une UE optionnelle de 25 h *(resp : Pascal Rigolet) //* Salles informatiques du Bât 336
- Proposée en L2S4 (UFR Sciences – Biologie) // Effectifs : 24 à 28
- Axé sur la pratique
- Travail en groupe encouragé
- Durée : un mois à raison de 2 séances par semaine
- Un ENT : 10 rubriques ; 20 vidéos de cours; 25 fiches de synthèse; 12 défis *(chaque nouveau défi est très attendu par les étudiants*); 3 serious games (en cours de création) ; mise à jour quotidienne
- Contr. des connaissances : défis (et questions flashs en salle), une note collective; un projet (binôme); un examen final sur table avec tablette ou ordinateur et connexion internet

Retours d’expériences :L’UE « Traitement et Analyse de Données en Biologie »

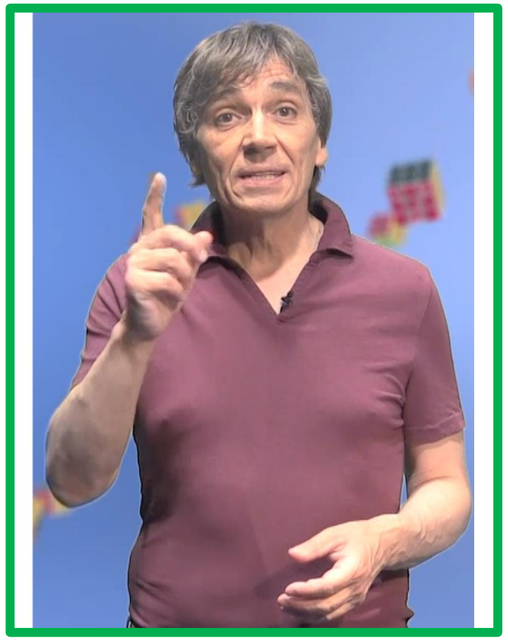
- **Un enseignement à pédagogie inversée** *(Cours, Exercices et Travail Pratique à chaque séance)*
- **Une bonne partie de l'enseignement théorique est donné sur un ENT**
- **Résoudre un problème en classe en faisant participer le maximum de personnes et en se posant le maximum de questions**
- **Dessiner un schéma situationnel au tableau sur lequel figurent tous les éléments nécessaires à la résolution du problème statistique**
- **On s’approche d'une démarche de *data scientist***
- **L’enseignant lance des défis aux étudiants et pose des questions flash contextuelles**
- **Acquisition d’une compétence très demandée**
- **Un plus pour le CV (recherche / ingénierie / entreprise)**

Retours d’expériences :L’UE « Traitement et Analyse de Données en Biologie »

- **Travail sur des jeux de données réels**
- **Thématiques : Environnement, Epidémiologie, Génomique, Pollution, Biomédical, agronomie, …**
- **Travail organisé en groupes de 2 à 4 en salle**
- **Apprendre à raisonner avant d'appliquer le modèle statistique adapté à la problématique.**
- **Utilisation de logiciels dédiés *(principalement EXCEL et R)***
- **Apprentissage d’une démarche**
- **Préparation des projets en classe**
- **Application immédiate dans les TP et les stages suivis par les étudiants**

  (ainsi que dans la suite de leur cursus)
- **Site web associé à l’enseignement où les étudiants retrouvent présentation de l’enseignement, cours (vidéos et fiches) , exercices corrigés, auto-évaluations, forum, liaison avec l’enseignant**

Retours d’expériences :L’UE « Traitement et Analyse de Données en Biologie »



L’enseignant *[animateur]*

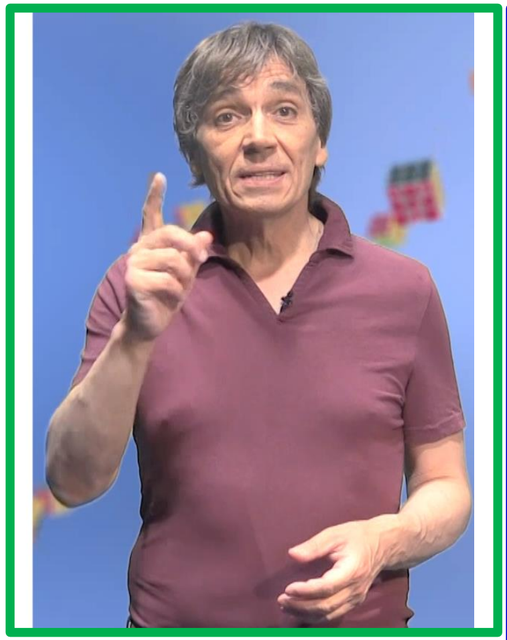
> **présente la situation** demandant une analyse statistique (à définir par les étudiants puis ensuite à réaliser à partir d’un **fichier de données** téléchargeable)

> **fait le point** à partir des réponses données.

> fait progresser l’analyse

idsite	dist	conc_metal
1	0	207,8
2	0	213,4
3	0	211,2
4	0,25	205
5	0,25	206,2
6	0,25	200,1
7	0,25	199,7
8	0,5	204,3
9	0,5	206,2

Retours d'expériences :L'UE « Traitement et Analyse de Données en Biologie »

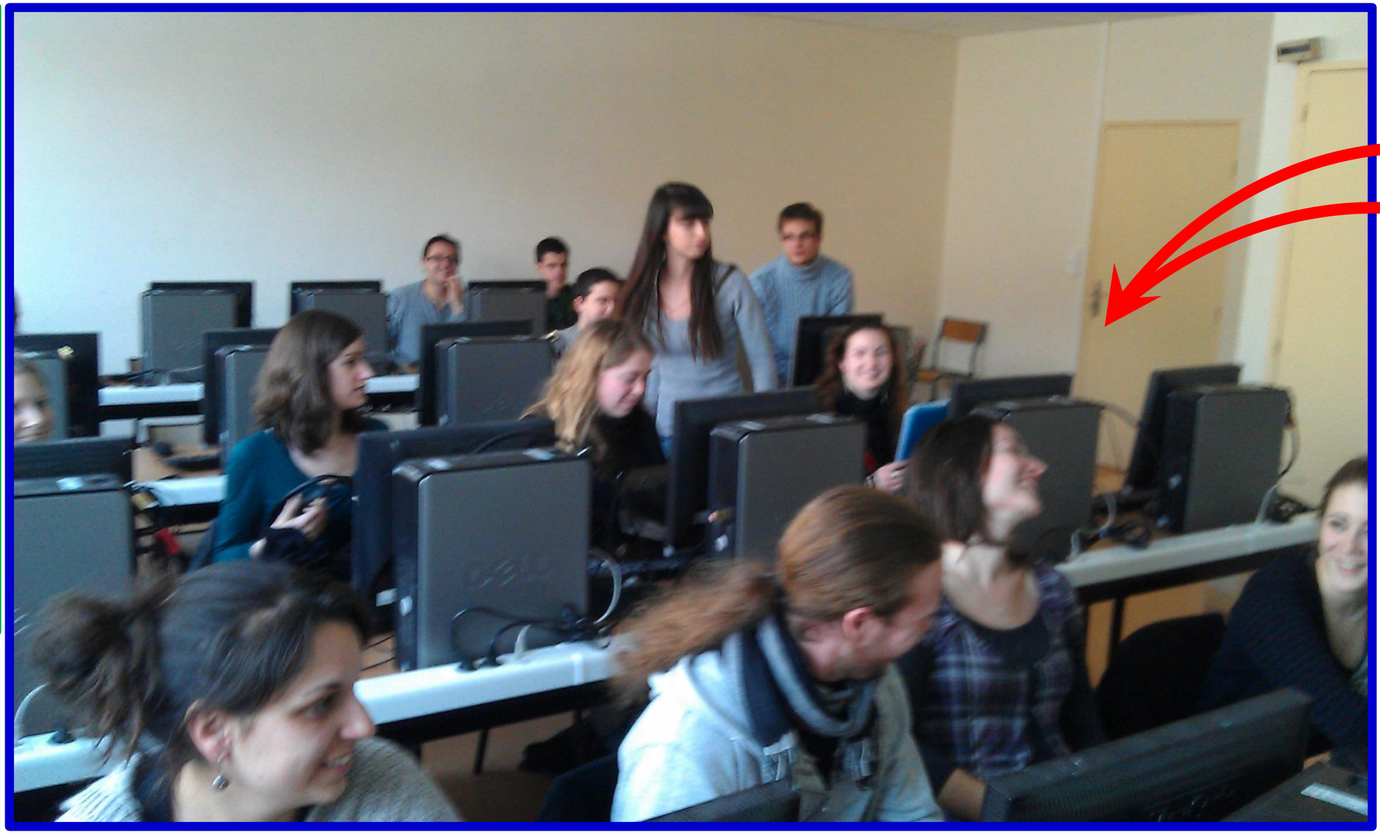


L'enseignant *[animateur]*

> **présente la situation** demandant une analyse statistique (<u>à définir par les étudiants</u> puis ensuite à réaliser à partir d'un **fichier de données** téléchargeable)

> **fait le point** à partir des réponses données.

> fait progresser l'analyse



Les étudiants sont **mobiles**, peuvent **échanger des informations** d'un groupe à l'autre et consulter à souhait **internet** afin de proposer une **stratégie cohérente** pour répondre au problème posé *(jouant le rôle de profiler de données).*

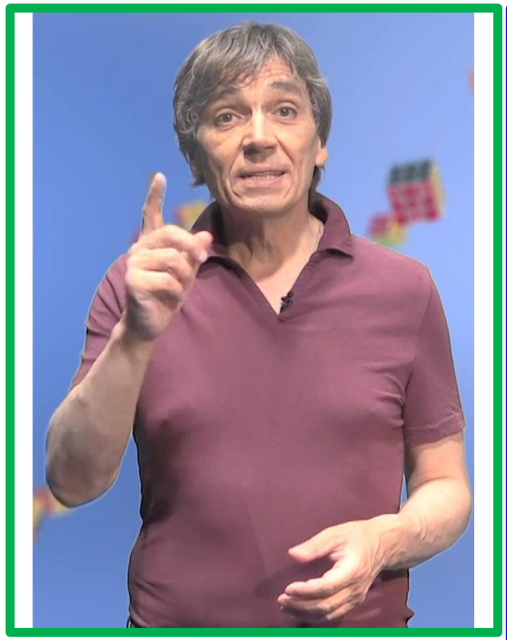
La **bonne humeur** est de mise et le travail fait **sérieusement** *(les étudiants **aiment bien les challenges**).*

Déroulement type d'une séance :

libre circulation, de la **réflexion** et de l'**échange** à intervalle régulier; beaucoup de **pratique** conduisant cependant à introduire la théorie sous-jacente (qui est présentée plus en détail sur l'ENT).

idsite	dist	conc_metal
1	0	207,8
2	0	213,4
3	0	211,2
4	0,25	205
5	0,25	206,2
6	0,25	200,1
7	0,25	199,7
8	0,5	204,3
9	0,5	206,2

Retours d'expériences :L'UE « Traitement et Analyse de Données en Biologie »

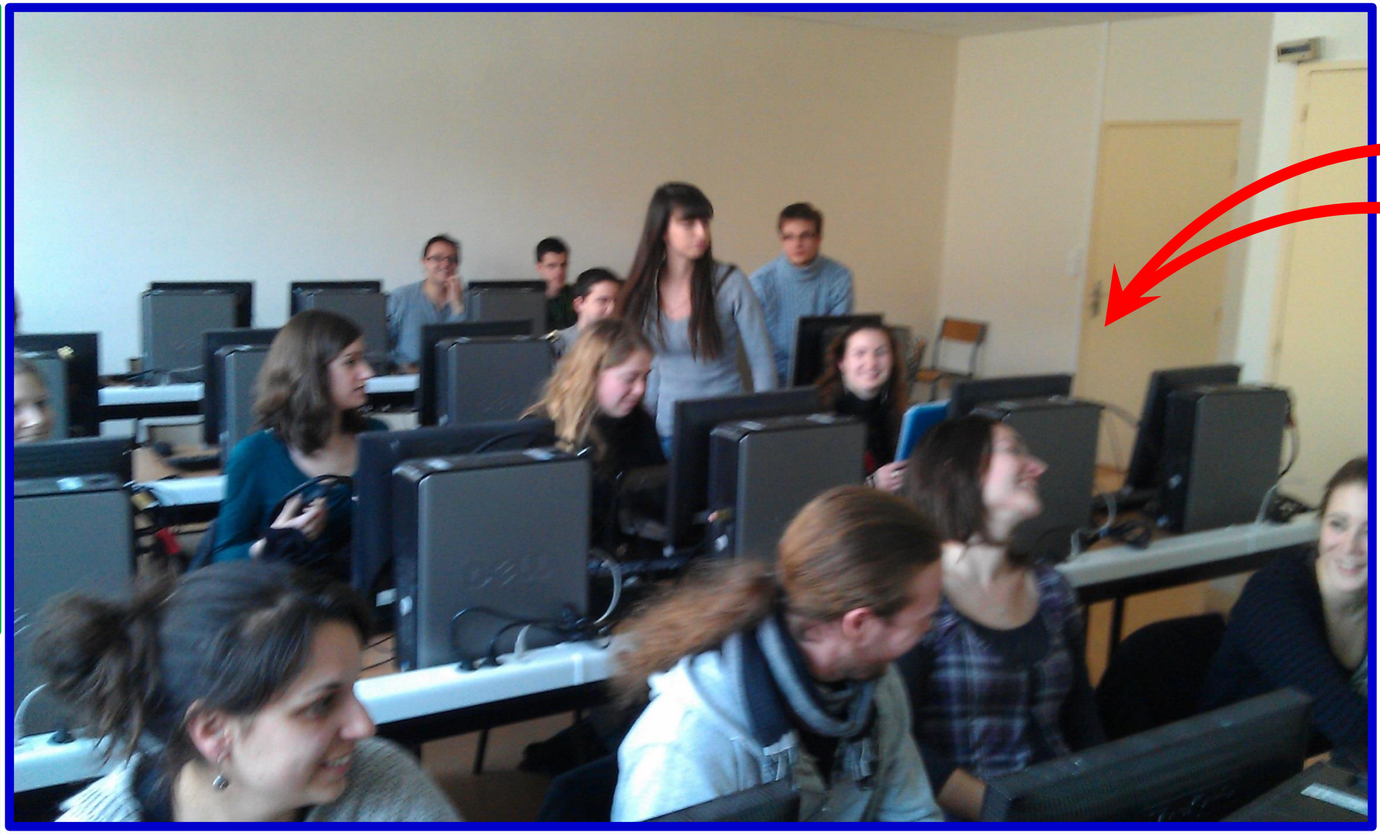


L'enseignant *[animateur]*

> **présente la situation** demandant une analyse statistique (<u>à définir par les étudiants</u> puis ensuite à réaliser à partir d'un **fichier de données** téléchargeable)

> **fait le point** à partir des réponses données.

> fait progresser l'analyse



Les étudiants sont **mobiles**, peuvent **échanger des informations** d'un groupe à l'autre et consulter à souhait **internet** afin de proposer une **stratégie cohérente** pour répondre au problème posé *(jouant le rôle de profiler de données).*

La **bonne humeur** est de mise et le travail fait **sérieusement** *(les étudiants **aiment bien les challenges**).*

Déroulement type d'une séance :

libre circulation, de la **réflexion** et de l'**échange** à intervalle régulier; beaucoup de **pratique** conduisant cependant à introduire la théorie sous-jacente (qui est présentée plus en détail sur l'ENT).

> **Forte implication** des étudiants

> Pratiquement **pas d'absentéisme**

idsite	dist	conc_metal
1	0	207,8
2	0	213,4
3	0	211,2
4	0,25	205
5	0,25	206,2
6	0,25	200,1
7	0,25	199,7
8	0,5	204,3
9	0,5	206,2

Retours d’expériences :L’UE « Traitement et Analyse de Données en Biologie »

L’accent est mis sur la pratique

Aucun calcul à la main > Choix d’un logiciel pour faire les calculs : R (mais pas que…)

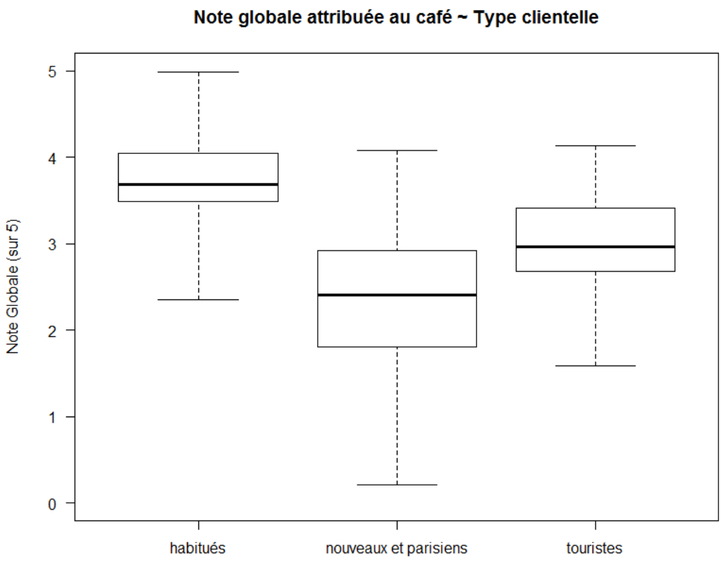
Savoir avant tout représenter graphiquement la situation et interpréter les sorties fournies par un logiciel

Une grosse réflexion sur l’origine et la fiabilité des données

Ne jamais se démunir d’une analyse critique et des limites de ce que peut révéler l’analyse des données

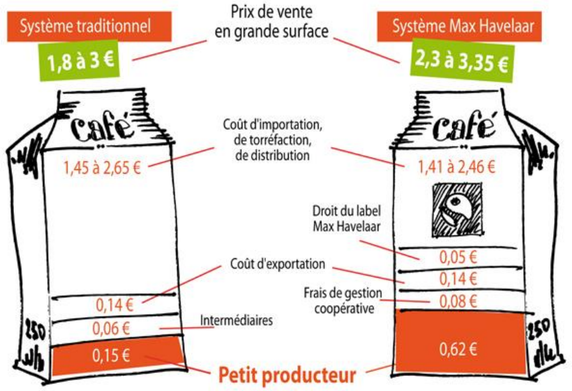
Retours d'expériences :L'UE « Traitement et Analyse de Données en Biologie »

Codage des données



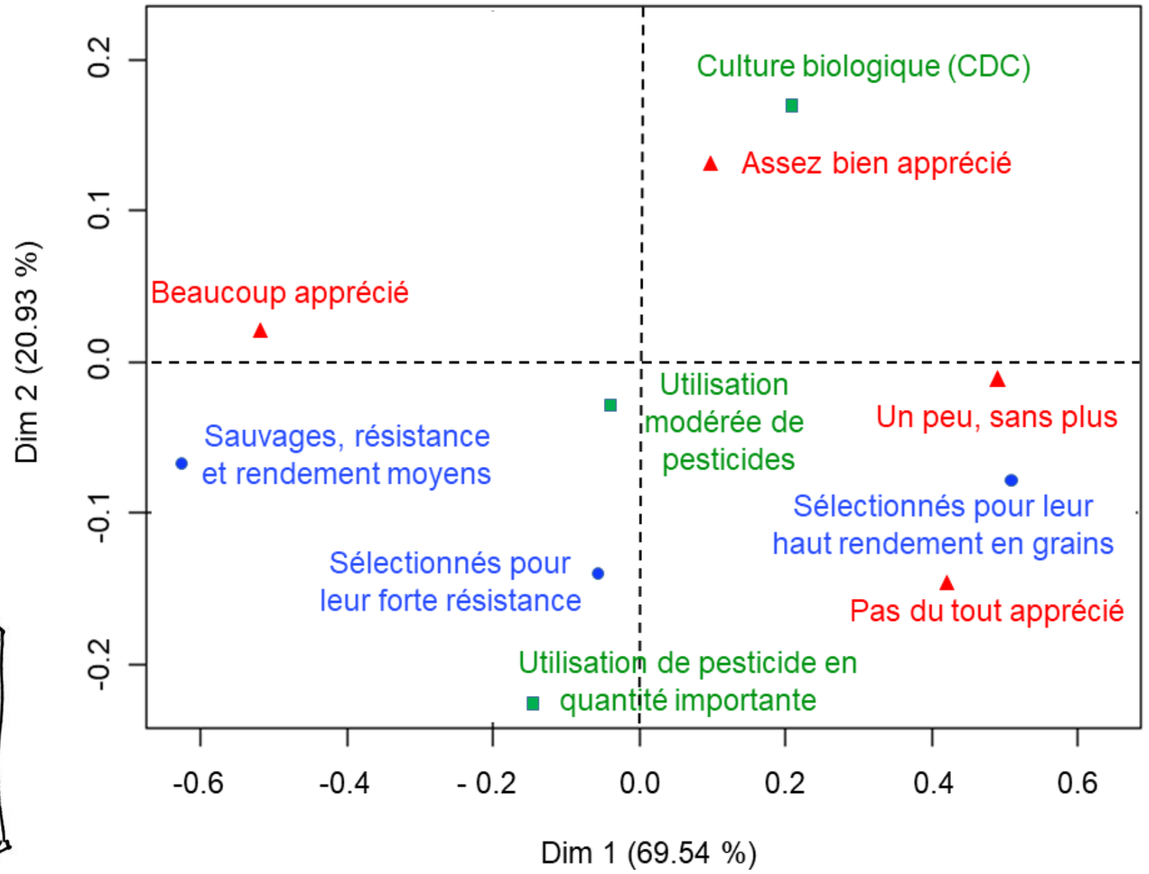
Anova

Comparer



Source : Max Havelaar France, février 2003 (prix calculé sur une moyenne, pour un arabica lavé d'Amérique Latine)

Initiation à l'analyse factorielle (AFC)



Retours d'expériences :L'UE « Traitement et Analyse de Données en Biologie »

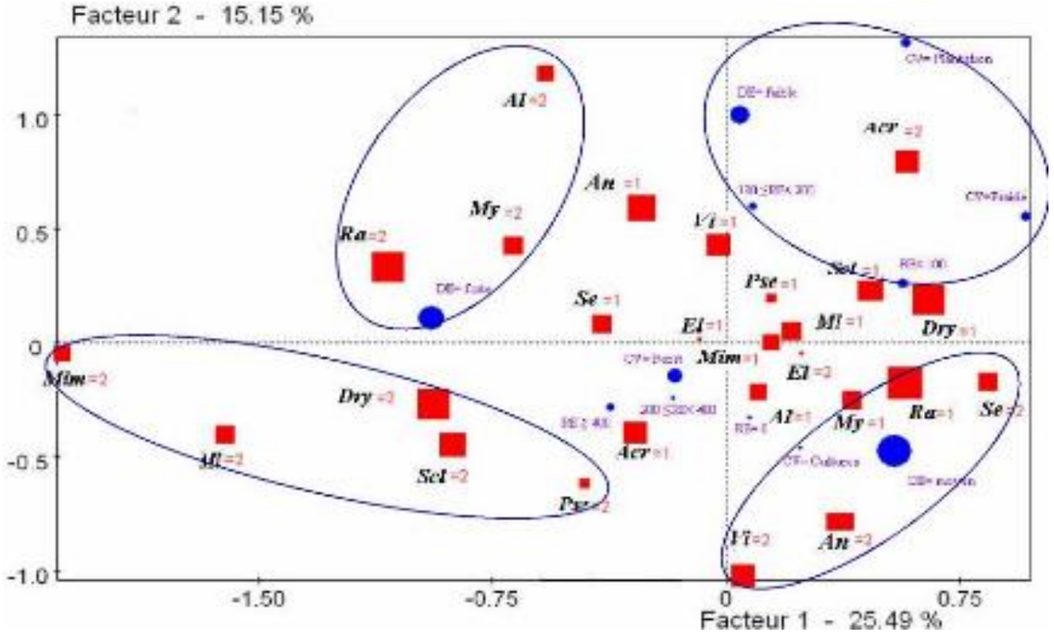
Analyse Factorielle (AFC)

Facteurs influençant la **production piscicole des étangs** (Ouest-Cameroun)

Mise en place d'un étang d'inondation

HISTOGRAMME DES 13 PREMIERES VALEURS PROPRES

NUMERO	VALEUR	POURCENTAGE	POURCENTAGE CUMULE	
1	0.2549	25.49	25.49	**********
2	0.1515	15.15	40.64	**********
3	0.1087	10.87	51.50	**********
4	0.0926	9.26	60.76	**********
5	0.0798	7.98	68.74	**********
6	0.0661	6.61	75.35	**********
7	0.0521	5.21	80.57	**********
8	0.0503	5.03	85.60	**********
9	0.0430	4.30	89.90	**********
10	0.0365	3.65	93.55	**********
11	0.0280	2.80	96.35	*********
12	0.0208	2.08	98.43	*******
13	0.0157	1.57	100.00	*****



La **production** piscicole d'un étang **dépend** de **l'environnement végétal** des sites et des **conditions hydrologiques** du cours d'eau qui l'inonde

Décision → *Action*

- réhabiliter d'autres étangs afin d'augmenter la surface exploitable ;
- organisation des pisciculteurs autour d'un groupe (partage d'expérience);
- protéger les couverts forestiers car ils assurent les meilleurs productions des étangs

Fichier texte

purifications.txt - ...

Fichier Edition Format Affichage ?

```
proteine (mg/ml)
6,10
8,20
11,00
12,50
```

Bloc-note

Base de données : purifications.txt

Format : texte (csv)

4 enregistrements (records)

1 champ : « purification »

1 séparateur de champ : « ; »

Point décimal (notation anglo saxone)

Importable dans EXCEL

Importable dans R (instruction read.table)

Importable dans n'importe quel logiciel de Stat

EXCEL

purifications.xlsx - Excel

	A	B	C	D
1	indice	[proteine1]		
2	purification	(mg/ml)		
3	1	6,10		
4	2	8,20		
5	3	11,00		
6	4	12,50		
7	moyenne	9,45		
8	écart-type	2,86		
9				

purifications proteines3

Prêt

R

instruction read.table

R Console

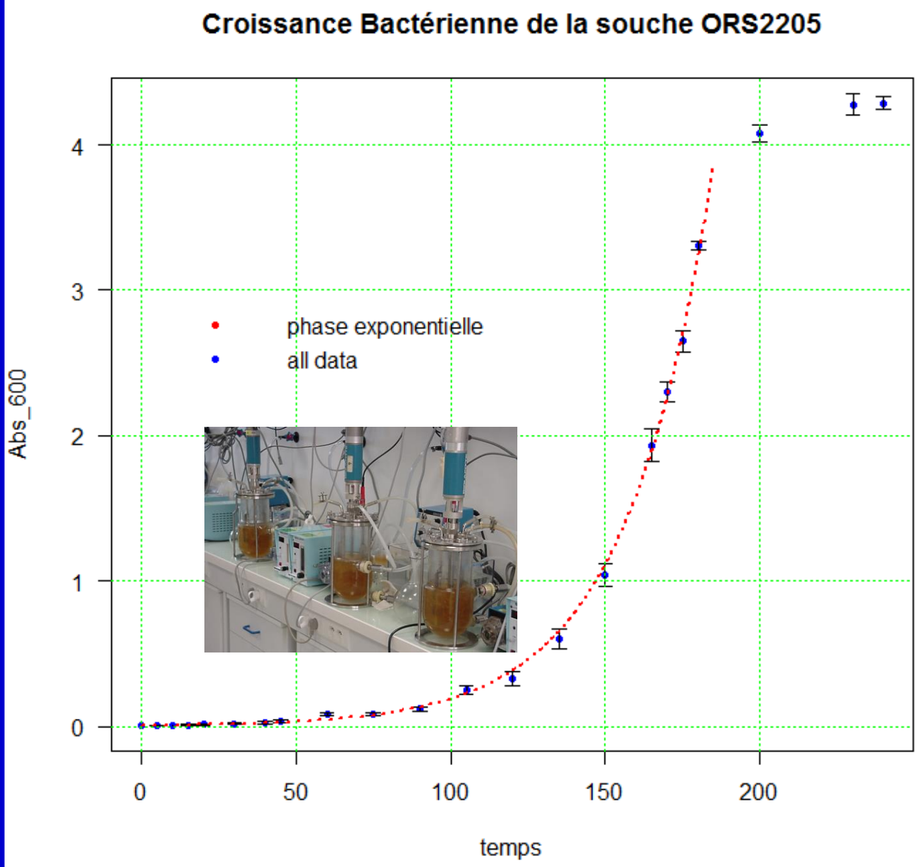
```
> read.table(file="./data/purifications.txt",h=TRUE,sep=";",dec=",")
  proteine
1      6.1
2      8.2
3     11.0
4     12.5
>
```

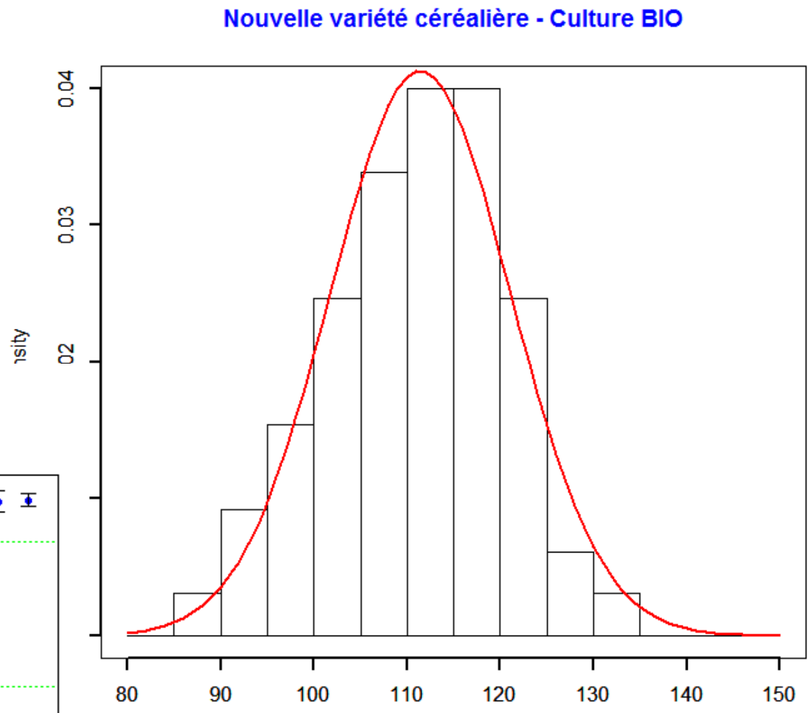
read.table(file="./data/purifications.txt",h=TRUE,sep=";",dec=",")

Retours d'expériences :L'UE « Traitement et Analyse de Données en Biologie »

La représentation graphique

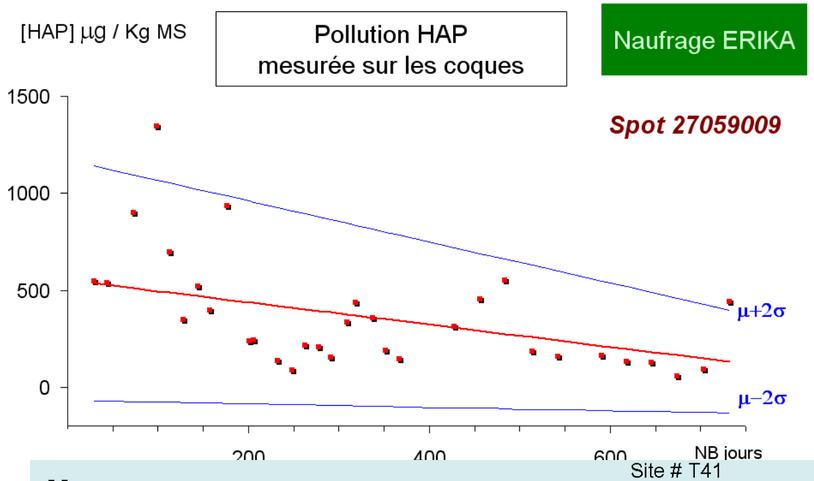
- méthode
- gestion des barres d'erreurs
- apprentissage de logiciels

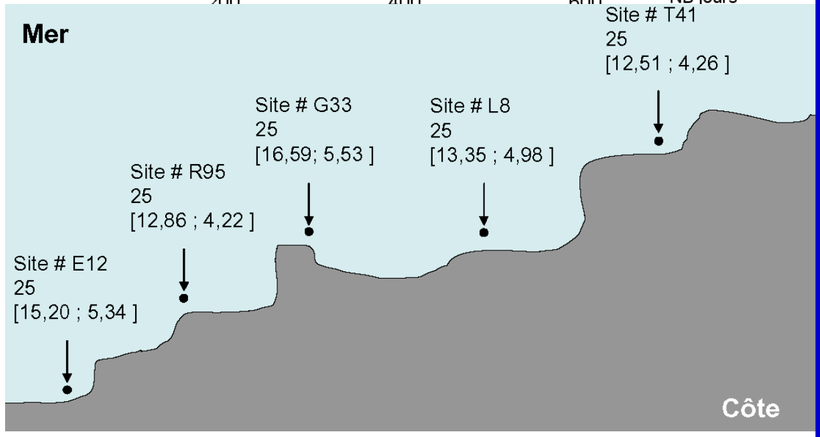




Col	1	2	3	4	5	6	7
Line	temps	DO1	DO2	DO3	DO4	DO5	DO6
1	0	0.005	0.005	0.003	0.005	0.005	0.005
2	5	0.004	0.008	0.006	0.007	0.005	0.006
3	10	0.008	0.007	0.007	0.006	0.006	0.007
4	15	0.007	0.005	0.005	0.005	0.013	0.012
5	20	0.005	0.013	0.010	0.014	0.010	0.016
6	30	0.012	0.022	0.016	0.021	0.018	0.017
7	40	0.024	0.036	0.031	0.018	0.018	0.012
8	45	0.042	0.021	0.028	0.038	0.038	0.033
9	60	0.082	0.080	0.067	0.090	0.091	0.073
10	75	0.086	0.084	0.067	0.096	0.099	0.080
11	90	0.097	0.132	0.139	0.112	0.118	0.121
12	105	0.258	0.217	0.251	0.214	0.254	0.303
13	120	0.371	0.327	0.273	0.276	0.326	0.380
14	135	0.583	0.666	0.580	0.624	0.499	0.657
15	150	1.000	0.920	1.104	1.072	1.135	1.014
16	165	1.917	1.892	1.825	2.137	1.894	1.937
17	170	2.410	2.267	2.250	2.245	2.286	2.361
18	175	2.599	2.748	2.585	2.662	2.593	2.708
19	180	3.331	3.275	3.305	3.334	3.316	3.267
20	200	3.998	4.085	4.075	4.047	4.123	4.147
21	230	4.360	4.194	4.320	4.318	4.268	4.206
22	240	4.291	4.241	4.365	4.261	4.272	4.284

Modèle linéaire





Activité enzymatique	
Individus Témoins	Individus exposés pendant 2 jours à une dose d'insecticide
27.0	26.0
28.8	19.0
26.0	27.1
34.5	19.6
28.8	22.0
23.0	15.5
32.2	24.2
25.8	17.0

Retours d'expériences :L'UE « Traitement et Analyse de Données en Biologie »

Menu

- Objectifs
- Supports de Cours
- Bases de Données
- Web Design
- > Espace L2S4
- Machine Learning
- Les défis du L2
- Je vous recommande
- Sondages
- Forum
- Video Teaching
- Exercices
- Tables
- Contrôle Connaissances
- Projets
- Pour approfondir
- Divers

Séances et documents - Année 2019

Option Traitement et Analyse des Données Biologiques - L2S4

Une étroite complicité entre l'enseignant et les étudiants est la clé du succès de cette formation. Je vous souhaite beaucoup de plaisir dans cette UE.

Espace Numérique de Travail (ENT) - Enseignement Numérique (EN)

Cette partie d'auto apprentissage est aussi import.

[dernière mise à jour le 29/01/2019]

(L'enseignement débutera désormais à 14h00)

Les mots de la semaine :

data sciences ; valorisation des données ; statistique descriptive ; variance

Présentation de l'option TAB

- Présentation de l'option (fichier pdf)
- Présentation vidéo de l'option (fichier mp4) [durée : 5 min. 15]

Plans des séances, photos de tableaux et diaporamas projetés en cours

> Chaque séance est couplée à une ou deux formations numériques à distance d'une dizaine de minutes

Séance 1 - Introduction : La valorisation des données à l'ère du Big Data

- Les types de variables manipulées en Statistique (apprentissage vidéo de 10 minutes)
- Fiche résumé sur les types de variables (fichier pdf)
- Prise en main du logiciel R, vos tous premiers pas avec R (vidéo pour les étudiants découvrant R)
- R - environnement de travail (pour les étudiants ne connaissant pas R; fichier pdf)
- Atelier R - Traitement et Analyse de données avec R (partie 1 sur 2) (apprentissage vidéo de 14 minutes)
- Résumé sur les types de variables (fichier pdf)

Séance 2 - Un petit aperçu des objectifs et de la diversité de l'analyse statistique

- La Statistique Descriptive, qu'est-ce que c'est ? (apprentissage vidéo de 10 minutes)
- Fiche de synthèse sur la statistique descriptive (développé en cours, fichier pdf)
- Calcul des paramètres statistiques et de leurs estimations (développé en cours, fichier pdf)
- La statistique expliquée à mon chat : chaîne qui vaut le détour! (certains d'entre vous connaissent bien sûr)
- Vocabulaire statistique (fichier pdf)
- Exercice d'autoévaluation (QCS)
- **Vous retrouverez les défis ainsi que les défis du week-end dans l'onglet "Les défis du L2"**

Séance 3 - De l'importance du contexte dans l'analyse des données

- Le diaporame sur l'analyse des données et les data sciences (fichier pdf)
- Atelier R - Traitement et Analyse de données avec R (partie 2 sur 2) (apprentissage vidéo de 14 minutes)

Un ENT indispensable

10 rubriques ; un cours en ligne; vidéos, fiches, exercices, défis disponibles; un forum; sondages; trois *serious games (en cours de réalisation)…*

Les défis du L2S4 - Année 2019

(Option Traitement et Analyse des Données Biologiques - L2S4)

Ou comment obtenir une note en contrôle continu ...

Vous disposez de 24 heures pour relever les défis présentés sous le titre "défis 24h", de 48 heures pour relever les défis présentés sous le titre "défis 48h" et de 72 h pour relever les défis présentés derrière le titre "défis du week-end". Il n'y a pas de points négatifs : ces défis ne sont notés que si la réponse est juste. Envoyez par email (pascal.rigolet@u-psud.fr) vos réponses aux défis proposés (j'accepte que vous me les rendiez directement en classe le jeudi pour les défis proposés le mercredi sur le site)

Les défis qui sont déjà passés sont sous le titre "archives des défis proposés" et ne peuvent faire l'objet d'une réponse car il est trop tard pour y répondre. Votre travail sera corrigé dans les plus brefs délais (laissez aux correcteurs le temps d'évaluer vos réponses de façon posée)

(choisissez votre/vos défi/s ou relevez les tous)

>> page mise à jour le 07/03/2019 <<

Bientôt de nouveaux défis et la correction de certains défis passés

Comme pour une cohorte sur un MOOC (Massive On line Open Courses), ce sont les étudiants qui vont corriger le contrôle continu, à savoir en ce qui nous concerne : les défis. Ceci s'inscrit, en pédagogie inversée, dans l'esprit de travail en équipe et la confiance qui doit régner dans les deux sens entre les étudiants et l'enseignant. Enfin cela nous apprend à nous responsabiliser. Les correcteurs (un(e) à deux par défi) seront préférentiellement volontaires. En conséquence, les projets seront envoyés au correcteurs et à l'enseignant (qui pourra procéder à des contrôles au hasard des corrections effectuées par les correcteurs). Les correcteurs seront dispensés de répondre au défi qu'ils corrigent et gratifiés d'un point (voire de 2 points, selon le défi) de contrôle continu pour le travail de correction réalisé. La correction des défis contribue en effet, *de facto*, à l'apprentissage des connaissances.

Fichier des notes du contrôle continu :

- Fichier des notes du contrôle continu et des projets (mis à jour des défis corrigés au 07/03/2019)

Fichier EXCEL pour les correcteurs :

- Fichier EXCEL de notes des défis (pour les correcteurs)
  (à remplir par les correcteurs qui y indiqueront leurs noms ainsi que le numéro des défis corrigés)

Défis "one week" en cours :

- Défi N°12 (fichier pdf) (statut :en cours) ***[1,5 points]***

Pour ce défi N°12 je serai le correcteur

Vous avez jusqu'à jeudi 14/03/2019 à 18h00 pour m'envoyer votre réponse par email

Alors, cap ou pas cap? vous pouvez récolter 1,5 points pour ce défi "one week"

Objectifs | Contact | Liens utiles

Video Teaching

- Vidéo expliquant le but des videos d'enseignement [durée : 2 min. 15]

Les étudiants et visiteurs trouveront ici des vidéos illustrant certaines parties de l'enc
permettant de préciser un point précis ou d'aborder une thématique que nous n'aur
ensemble et que les visiteurs auront le loisir de consulter.

Bien entendu ceci ne substitura pas aux cours présentiels (devant les étudiants)
exposée dans la vidéo n'a pas été traitée en salle avec l'enseignant, elle ne figurera p

Des vidéos de remise à niveau en stat seront également bientôt disponibles.

Liste des vidéos disponibles :

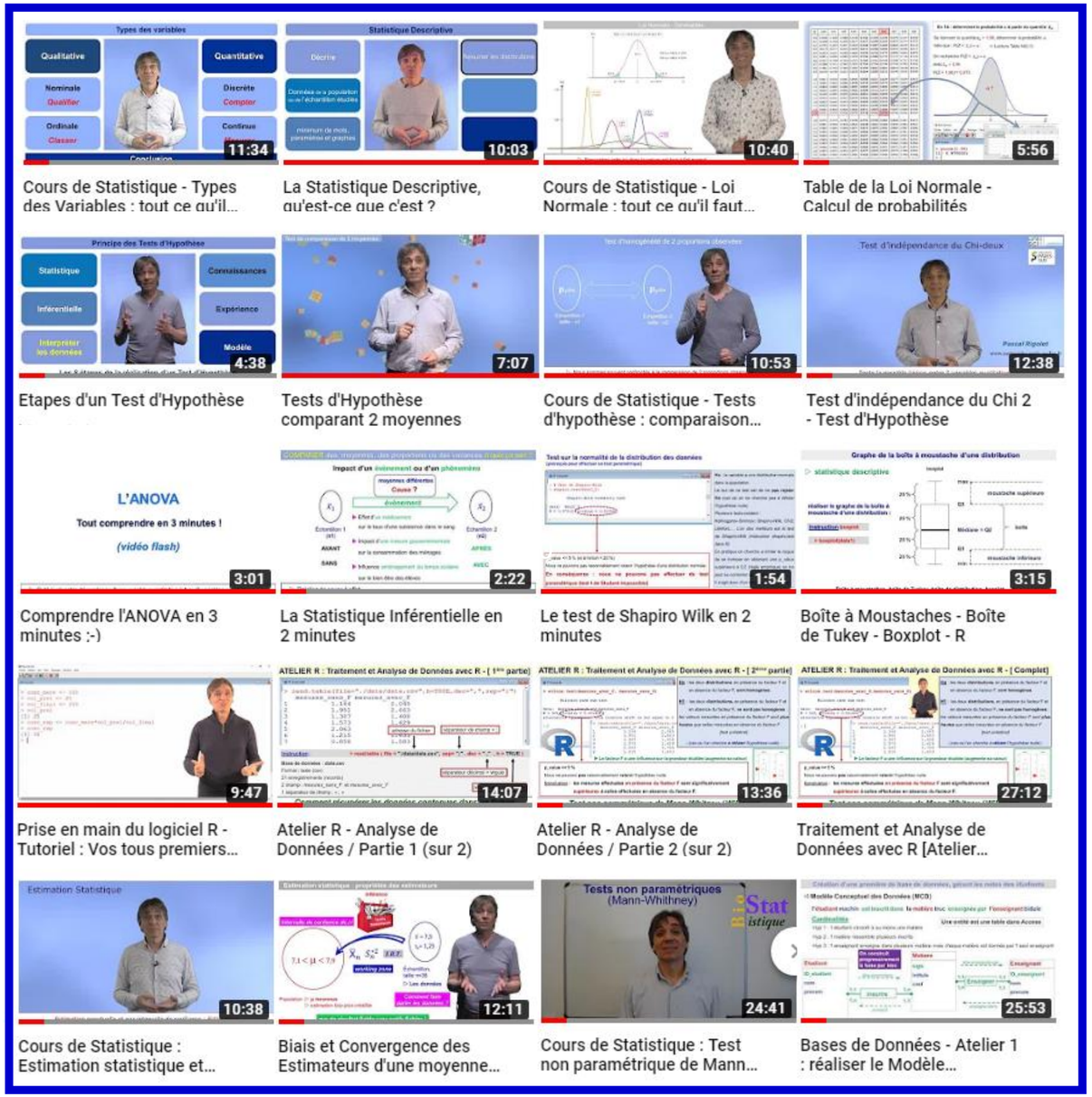
**N'hésitez pas à apporter vos commentaires à pascal.rigolet@u-psud.fr pour amé
et le contenu de ces vidéos.**

- But des videos d'enseignement [durée : 2 min. 15]
- Tests d'hypothèse, Intervalles de Confiance et de Fluctuation [durée : 6 min. 30
- Intervalles de fluctuation (/de pari) [durée : 4 min. 50]
- Intervalles de confiance [durée : 5 min. 00]
- Tests d'hypothèse [durée : 11 min. 30]
- Rappels sur le schéma de la loi binomiale [durée : 6 min. 30]
- Vidéo_Tuto graphe simple de la courbe de croissance avec EXCEL (fichier htm
- Conseils pour l'examen pratique [durée : 4 min. 30]
- Estimateurs et estimations [durée : 11 min. 30]
- Variables, lois de probabilités et tests associés [durée : 14 min.]
- Décrypter un texte d'examen de stats [durée : 7 min. 20]
- Test de Mann Whitney [durée : 24 min. 40]
- Test Chi2 - Independance - tests d'hypothèse [durée : 15 min. 20]
- Barres d'erreurs [durée : 5 min. 30]

Bientôt d'autres vidéos sur cette page !

Retours d’expériences :L’UE « Traitement et Analyse de Données en Biologie »

L’offre vidéo



Un enseignement à pédagogie inversée

- ⇨ Une bonne partie de l'enseignement est donné sur un ENT de l'enseignement théorique ; les étudiants disposes ainsi d’un panel de courtes vidéos présentant l’essentiel de ce qu’il faut comprendre et savoir pour pouvoir participer aux ateliers en pleine capacité
- ⇨ Les étudiants apprécient beaucoup ces vidéos (sondage à l’appui)

Il y en a d’autres…

Retours d'expériences :L'UE « Traitement et Analyse de Données en Biologie »

Défis et **questions flash** : deux éléments clés de cette UE

- L'enseignant lance des défis aux étudiants
- Les étudiants ont un temps donné pour les résoudre (1, 2 ou 3 j)
- Se mettent à 2, parfois même à 3
- Les défis sont le domaine des étudiants :

  ce sont eux qui corrigent les défis > 1ou 2 correcteurs par défi

  (correction supervisée par l'enseignant)
- 12 à 15 défis sont proposés au fil du temps
- Les défis sont très addictifs pour les étudiants
- Les questions flash contextuelles en salle de cours

  (2 min pour y répondre, gain immédiat)
- Pour gagner des points de contrôle continu
- Peut s'apparenter à des serious games

Retours d’expériences :L’UE « Traitement et Analyse de Données en Biologie »

Défi n°10 Agathe DUMONT GrA11

Les anguilles de la Tamise “shoutée” à la cocaïne

Liens internets

Top santé, Il y a 6 jours — https://www.topsante.com/medecine/addictions/drogue/nsolite-les-anguilles-de-la-tamise-shootees-a-la-

The Sunday times, Il y a 8 jours — https://www.thetimes.co.uk/article/cocaine-in-thames-makes-eels-hyperactive-lkd3vhncd

Gent side, Il y a 3 jours — https://www.maxisciences.com/anguille/des-anguilles-shootees-a-la-cocaine-dans-la-

Fake news ou réalité?

« Si la nouvelle n'avait pas été publiée dans le très sérieux Sunday Times, on aurait pu croire à une "fake news". C'est pourtant ce quotidien qui a annoncé que "la cocaïne de la Tamise rend les anguilles hyperactives". Conséquence étonnante des pluies torrentielles qui submergent les usines de traitement des déchets et entraînent le déversement des eaux usées dans la rivière.

Les anguilles séjournent de longs mois dans les fleuves et rivières européens (notamment la Tamise qui traverse Londres) avant de migrer vers l'Atlantique pour se reproduire. Mais la Tamise est aussi l'un des cours d'eau les plus chargés en résidus de cocaïne. Ces résidus toxiques s'accumulent dans les muscles et le cerveau des poissons, les rendant hyperactifs. "La cocaïne affecte leur comportement, rendant leur migration plus erratique" ont déclaré les chercheurs du King's college de Londres au quotidien britannique. Un comportement qui risque d'affecter leur reproduction et pourrait mettre l'espèce en danger. »

La cocaïne est devenue un contaminant environnemental très commun en Grande-Bretagne. Une étude récente a même montré que 10 % de la population en a sur les doigts, sans pourtant en consommer. »

« Des chercheurs de l'Université de Naples Frédéric II (Italie) ont mis en avant, au mois de juin dernier, dans leurs travaux les effets que cette drogue illégale pouvait avoir sur les anguilles.

Après les avoir plongées dans une eau contenant autant de résidus de cocaïne que l'on peut trouver dans la plupart des fleuves européens, ils les ont observées pendant 50 jours. Ils ont ainsi pu constater que cela les avait rendues beaucoup plus hyperactives que celles qui n'avaient pas eu droit à un tel bain forcé. Les molécules de la drogue s'étant accumulées dans leur cerveau, leurs muscles, branchies, peau et dans d'autres tissus. De plus, leur muscle squelettique montrait des blessures sévères qui n'arrivaient toujours pas à cicatriser 10 jours après avoir que les cobayes aient été retirés de l'aquarium.

Ces conclusions auraient pu ne concerner qu'une expérience menée dans un laboratoire. Seulement voilà, il s'agit malheureusement d'une triste réalité qui concerne aussi les anguilles dans leur habitat naturel, en particulier dans la Tamise, comme l'a démontré une récente étude relayée dans le quotidien The Independent.

De Gent side

Toujours pas convaincu ?

Des étudiants de l'université Paris sud
Encadrés par le célèbre professeur Pascal Rigolet, vont prochainement réaliser une étude sur les anguilles du fleuves de Londres.
Leur but étant de savoir si oui ou non l'information est exactes.
Il vous présente ici le début de leur projet: c'est à dire la description d'une analyse statistique.

Questionnement :
Nous avons des informations postées sur internet. Nous cherchons à savoir si ces informations sont vrais.
1- Soit si les anguilles sont belles et bien devenues hyperactives à cause de la concentration de cocaïne
2- si les crues sont corrélées à l'hyperactivité des anguilles
3- si les crues sont corrélées à l'augmentation de concentration de la cocaïne dans le fleuves

Réalisation d'une étude statistique formant une régression linéaire:

Choix de l'étude:
- nous ne pouvons pas réaliser une étude statistique avec une expériences en laboratoire qui met en fac taux d'hyperactivité des anguilles en fonction des concentrations de cocaïne. Car jusqu'a une certaine concentration, la cocaïne n'aura pas forcément d'effet observable macroscopiquement. Et à un trop for cocaïne on pourrait observer la mort des anguilles. Ainsi nous ne pourrons pas observer une régressio mais plutôt une courbe sigmoïde qui retombe à 0 pour une concentration trop élévé pour permettre la anguilles (cette étude pourrais être faite en complément pour répondre au questionnement 1)
- le questionnement 2 parait compliqué
- alors je vais m'intéresser au questionnement 3

Nous allons donc mesurer:
1) L'importance de la crue en mesurant la différence de hauteur du fleuve lorsqu'il y a une crue ou une décrue on auras alors des valeur négative et positive

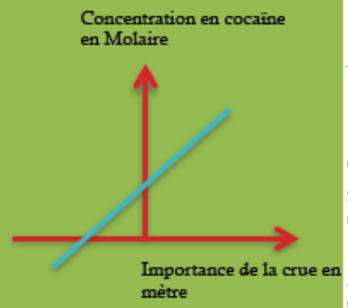
2) la concentration (en molaire) en cocaïne à plusieurs endroits du fleuve : soit de l'entrée du fleuves dans Lon sa sortie, tous les 100 mètres , lorsque le fleuve est revenu à la hauteur de 20m (qui est sa profondeur moyen que le volume de la tamise soit la même a chaque prise d'échantillon , échantillon pris dans une bouteille de 1litr de profondeur soit le milieu de la profondeur de la Tamise en temps normal , valeur positive

Nous allons donc avoir des couples de valeurs , soit l'importance de la crue en fonction de la concentration en c
Nous réalisons cette expériences pendant un an pour être sûr que des travaux de réaménagement n'ai pas lieux

Les caractéristiques de l'analyse statistique :
- ➔ population: La tamise
- ➔ échantillon: une bouteille d'eau de la tamise
- ➔ individus: l'eau contenus dans la bouteille

les variables mesurées: couple de valeur (concentration en cocaïne ; différence de hauteur de la Tamis se sont deux variables quantitatives continues

On s’attend à quelque chose comme ça :



Régrétion linéaire du nuage de point obtenu avec y=bo+b1x

y étant la concentration en cocaïne en Molaire(variable dépendante à expliquer)
x étant l'importance de la crue (variable indépendante, explicative)
bo et b1 sont donné par le logiciel R

lorem ipsum dolor problème, da

problème

Les résultats

1)On va chercher à savoir si la concentration en cocaïne est dépendante de l'intensité des pluies

Si elle sont dépendantes on pourrait être capable d'observer un nuage de point qui pourrais avoir une régression linéaire

2) de plus d'après l'article on s'attend à avoir une concentration de la cocaïne plus élevé si la crue est plus importante alors on peut prédire la pente de la droite

3) on vérifie que les ex suiv une loi Normal , que la relat y=F(x) est linéaire et que o une homogénéité des varia

Limites:

- d'autres paramètres peuvent rentrer en jeu et dans ce cas la on rentre dans une analyse de régression linéaire multiple
- ces autres paramètres peuvent être :
- le différentielle de consommation de cocaïne autour de l'année (du au stress des habitants , à de nouvelles réglementations , au prix, à un problème de livraison, à l'arrestation d'une grande quantité de vendeur illégaux……)
- ils est important de prendre en compte le maximums de ces facteurs pour minimiser au maximum les Ei qui sont les

Ouverture:

On peut réaliser une étude des erreurs en faisant un test de Shpiro Wilk à 10% de liberté

Et rejeter ou non l'hypothese en calculant la P value

Qu’en pensez vous ?
Pensez-vous que ces étudiants ont la bonne démarche?

si oui! rendez vous sur le site de Pascal RIGOLET (http://www.biostatistique.u-psud.fr/) dans l’onglet recherche pour financer cette étude.

si non! nous avons un forum sur le site qui permet d’échanger et de confronter les points de vue

Contrôle des connaissances

- Contrôle continu [1/3]

- Examen écrit [1/3]

- Mini Projet [1/3] (thématique choisie par l’étudiant; démarré en classe.)

L’examen écrit met l'accent non pas sur des calculs mais sur des stratégies de traitement des donnes et sur l'analyse de résultats présentés sous forme de graphiques et de tableaux.

première session : NF = 0,4*EE + 0,3*CC + 0,3*CC TP

deuxième session : NF = 0,5*EE + 0,25*CC(R) + 0,25*CC TP(R)

Retours d'expériences :L'UE « Traitement et Analyse de Données en Biologie »

PROJET

Le projet que vous devez rendre sera conçu à la façon d'un défi dont la **thématique** est **libre** (non imposée par l'enseignant).

Vous pouvez choisir de traiter le sujet qui vous tient le plus à cœur.

Ce projet sera rédigé en **1 à 4 pages maxi**, ira droit au but et comportera :

- une **présentation du sujet** avec **un petit schéma contextuel** résumant (très simplement) la situation ;
- une **analyse statistique** (par exemple, une régression linéaire ou une ANOVA) avec un ou deux **graphiques** ;
- une **synthèse** du (des) résultat(s) conduisant à une petite **conclusion** [*partie communication*].

Il vous est demandé de **travailler en <u>binôme</u>** et je vous conseille de **discuter avec l'ensemble de la classe** de l'avancée de votre travail (apprentissage du travail en groupe sur une thématique au choix, avec les conseils de l'ensemble de la classe).

Pour organiser l'ensemble vous pouvez essayer de raconter une histoire [***storytelling***] ou le rédiger comme un article écrit par un(e) journaliste.

La date limite prévue pour le rendre est le 15 mars 2019

Vous l'enverrez sous le **format pdf** (ou bien dossier compressé zip si plusieurs fichiers) **par email** à mon adresse : pascal.rigolet@u-psud.fr

Je vous souhaite beaucoup de plaisir à réaliser ce projet.

PR

Les projets

- ⇨ **Travail en binôme**
- ⇨ **Thématiques au choix des étudiants**
- ⇨ **Format en 2 à 4 pages maxi**
- ⇨ **Style et présentation doivent être soignés**

Retours d'expériences :L'UE « Traitement et Analyse de Données en Biologie »

LAMER Ombeline
RABEC Alexia

UE TAB _ Projet 2017-2018

Etude de la répartition des espèces d'oiseaux marins sur l'Ile Surprise

Contexte :

Les récifs d'Entrecasteaux sont situés au nord-ouest de l'archipel de la Nouvelle-Calédonie. Ils sont classés depuis 2008 au patrimoine mondial de l'Unesco. Il n'y existe aucune infrastructure humaine (mis à part la Station Météo France). Ces récifs présentent une biodiversité importante. En effet, il s'agit d'une zone de reproduction et de ponte pour les tortues verte. De plus, les récifs comportent une importante population d'oiseaux marins. Ils sont ainsi inscrits comme Zone importante pour la conservation des oiseaux par BirdLife International. Nous nous intéressons ici à l'un des îlots de ces récifs : l'Ile Surprise. Elle possède une végétation abondante et de nombreux oiseaux viennent y nicher.

Données :

Nos données proviennent des travaux de l'équipe de Courchamp F., Bonnaud E., Bourgeois K et al. (Equipe Ecologie des Populations et des Communautés). Laboratoire Ecologie Systématique et Evolution (UMR 8079) de l'Université Paris-Sud. Nous y avons eu accès lors de l'UE Ecologie et Statistique du L2S4.

Protocole :

Quatre habitats différents ont été définis sur l'Ile Surprise en fonction de l'espèce végétale qui y était majoritaire : le faux tabac (*Argusiaargentea*), des massifs d'arbres (*Scaevolataccada*), des bosquets d'arbustes (*Pisoniagrandis*) et une plaine d'herbacées. Les individus de quatre espèces d'oiseaux marins ont ensuite été dénombrés, entre 2002 et 2009 : le Puffin du Pacifique, le Fou Brun, le Fou à Pieds Rouges et la Frégate.

Problématiques :

Nous cherchons, à déterminer s'il existe une relation entre les espèces d'oiseaux et les habitats, définis par leurs espèces végétales majoritaires, de l'Ile Surprise.

Analyse des données : Relation entre les espèces d'oiseaux et les habitats

	Puffin du Pacifique	Fou Brun	Fou à pieds rouges	Frégate	Totaux
Argusia	32	23	188	87	330
Plaine	60	140	66	18	284
Scaevola	17	14	86	71	188
Pisonia	176	52	554	76	858
Totaux	285	229	894	252	1660

Table 1 : Table de contingence des espèces d'oiseaux et des habitats

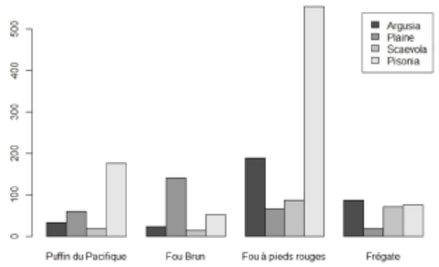


Figure 1 : Distribution des habitats en fonction de l'espèce d'oiseau

En observant la **Figure 1**, nous avons l'intuition que certaines espèces semblent préférer un habitat par rapport aux autres. Par exemple, la Plaine pour le Fou Brun ou encore *Pisonia* pour le Fou à pieds rouges. Nous allons confirmer notre idée à l'aide d'un test d'indépendance du chi-deux. L'analyse suivante a été réalisée avec le logiciel R. Pour ce test, nous avons défini les espèces d'oiseaux comme la variable dépendante à 4 modalités. Et les habitats comme la variable indépendante à 4 modalités également. Nos hypothèses sont les suivantes :

H0 : Les variables sont indépendantes : la répartition des espèces d'oiseaux ne dépend pas de l'habitat.
H1 : Les variables sont dépendantes : la répartition des espèces d'oiseaux est en relation avec l'habitat.

LAMER Ombeline
RABEC Alexia

UE TAB _ Projet 2017-2018

Les résultats obtenus sont les suivants : $\chi^2_{obs} = 537.6265$ et $p_{val} < 2.2 * 10^{-16}$. La p_{val} est extrêmement faible. Il y a de très fortes présomptions contre H0. La répartition des espèces d'oiseaux sur l'île est bien en relation avec les différents habitats. Afin de quantifier l'intensité de cette liaison, nous avons ensuite calculé le V de Cramer : $V_{cram} = 0.3285683$.La liaison entre les deux variables est relativement faible. Les espèces végétales majoritaires d'un habitat interviennent bien dans la répartition des espèces mais ce n'est pas le seul facteur à prendre en compte. Enfin, nous avons réalisé une analyse factorielle.

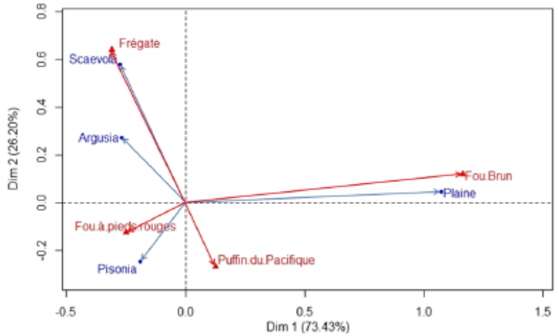


Figure 2 : Analyse factorielle des espèces d'oiseaux et des habitats

L'axe 1 rend compte de 73,43% de la dépendance. L'axe 2 en rend compte 26,20%. L'analyse de la **Figure 2** nous apprend que le Fou Brun préfère la Plaine. La Frégate se trouve principalement dans l'habitat *Scaevola*. Le Fou à pieds rouges est plus présent dans l'habitat *Pisonia*. Enfin, le Puffin du Pacifique s'observe plus rarement dans *Scaevola* et *Argusia*. On ne peut pas conclure quant à sa préférence entre *Pisonia* et la Plaine.

Analyse des données : La présence de rats influence-t-elle la répartition des espèces d'oiseaux ?

Nous avons vu précédemment que la principale espèce végétale qui caractérisait un habitat n'était pas le seul facteur à prendre en compte pour expliquer la répartition des espèces d'oiseaux marins sur l'Ile Surprise. Nous nous sommes donc intéressées à un autre facteur important : les rats. En effet, fin 2005, une dératisation a été effectuée. Nous avons donc comparé les effectifs de chaque espèce dans chacun des habitats avant et après dératisation à l'aide de Test de Student, après avoir vérifié la normalité des échantillons avec le logiciel R. Les résultats obtenus sont résumés dans la **Table 2**. (Cf. script annexe pour détails des calculs).

Table 2 : Comparaison des populations aviaires selon le milieu végétal avant et après dératisation

Pval	Puffin pacifique	Fou brun	Fou à pieds rouges	Frégates
Argusia	0.002848 **	0.00527 **	0.005597 **	0.004148 **
Pisonia	0.006576 **	0.007932 **	0.003932 **	0.03587 *
Plaine	0.004901 **	0.002269 **	0.009534 **	0.01083 *
Scaevola	0.02739 *	0.0199 *	0.007693 **	0.01593 *

** et ** : respectivement forte et très forte présomption contre l'hypothèse nulle*

Les p-valeurs obtenues à partir des comparaisons avant et après dératisation sont toutes comprises entre 0.1 et 1%. Ainsi, chaque espèce a évolué significativement différemment avant et après ce processus. La frégate, en particulier, a vu sa population fortement augmenter (cf. script annexe).

Discussion :

En résumé, les quatre espèces d'oiseaux marins étudiés s'établissent généralement dans une flore spécifique. De plus, il s'avère que la population de rat exerçait une pression significative sur les oiseaux.

Le choix de l'habitat dépend, certes, de l'habitude comportementale acquise par les oiseaux mais également de la disponibilité en ressource (concurrence inter-espèce, production du milieu), de la présence de prédateur spécifique d'une ou plusieurs flores et de la capacité du milieu (concurrence intra-spécifique) pour ne citer qu'eux. Notre jeu de donnée ne nous permet pas de spécifier la relation entre l'espèce et l'environnement végétal. Or chaque milieu n'a été représenté qu'avec l'espèce végétale majoritaire ce qui, là encore, implique une simplification pour le bien du modèle.

Quant au rat, omnivore, il se nourrit des œufs des nids édifiés à terre. Ce facteur rend certaines espèces plus sensibles à son influence. C'est le cas du Puffin du pacifique qui niche uniquement au sol alors que les Fous, mêmes s'ils ne pondent qu'un œuf en six mois, nichent sur différents supports.

Ces données récoltées sur huit ans ne reflètent peut-être pas les effets sur le long terme de la dératisation. Cet écosystème n'a sans doute pas retrouvé un état stationnaire en quatre ans. Par exemple, les Fous pourraient voir leur population augmentée sur une plus longue durée d'étude (cycle de vie long).

Les projets

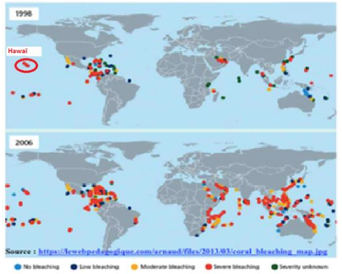
- ⇨ Travail en binôme
- ⇨ Thématiques au choix des étudiants
- ⇨ Format en 2 à 4 pages maxi
- ⇨ Style et présentation doivent être soignés
- ⇨ Les étudiants s'impliquent bien dans ce projet qui leur permet d'exploiter l'ensemble des connaissances et des démarches acquises

Retours d'expériences :L'UE « Traitement et Analyse de Données en Biologie »

Etude du blanchissement des coraux sous l'action du réchauffement climatique

Jusqu'où le blanchissement des coraux est-il dépendant du réchauffement climatique ?

Leriche Simon et Mustapha Azrine



Blanchissement des coraux dans le monde entre 1998 et 2006

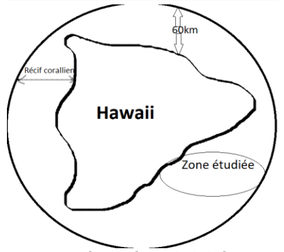


Schéma récapitulatif

Outils utilisés:

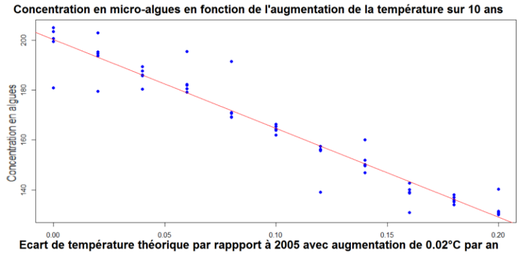
1. Logiciel R de traitement de donnée.
2. Logiciel Paint.
3. Calculatrice Ti-82 plus.
4. Fichier .csv des données récoltées.

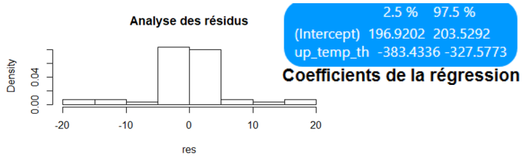
Ce blanchissement n'est que la conséquence du dépérissement des microorganismes appelés polypes qui hébergent dans leurs squelettes de calcaire blanc des algues microscopiques endosymbiotiques, les zooxanthelles. Ces algues apportent aux coraux les nutriments nécessaires en échange de la protection qu'offre la structure rigide de ces derniers. cette interaction est à l'origine des couleurs vives du corail. L'élévation de la température mondiale, qui est d'environ 0,2°C par décennie, expulse les algues endosymbiotiques, provoquant chez le corail une réaction de stress à l'origine du blanchissement qui lui est fatal à long terme.

L'archipel l'Hawaii compte en 2011 plus de 70 espèces de coraux recensées. Une équipe de chercheurs s'est intéressé à une espèce en particulier présente dans un périmètre de 60Km autour des côtes de l'île. Dans le but de suivre l'évolution de la concentration en microalgues symbiotiques en fonction de l'élévation de la température globale, ils ont échantillonnés chaque année de 2005 à 2015, dans une zone définie de 1Km de diamètre, 5 coraux de cette même espèce choisis aléatoirement d'une année à l'autre.

Notre but sera d'en analyser la base de données en étudiant la dépendance des deux variables quantitatives continues: la concentration en microalgues dans les coraux en fonction de de la différence de température théorique à partir de la première année de prélèvement. L'analyse se fera par régression linéaire.

Paramètre résultats de la régression	P-value	R2	Coefficient directeur (b1)	Moyenne Concent-ration (algue)	Ordinnée à l'origine (b0)
Valeurs	2.2 e-16	0.9248	-355,5	164.7	200,2





Bibliographie:
Site internet utilisé:
http://science.sciencemag.org/content/359/6371/80
http://eduterre.ens-lyon.fr/thematiques/climat/coraux/carte-de-blanchissement-des-coraux/view
https://projethawaii.wordpress.com/2011/07/13/coraux/

Analyse inférentielle:

Le coefficient de détermination R^2 est de 0.92 ce qui explique 92% de la répartition des données prouvant l'efficacité de la régression. Cela est en corrélation avec une analyse des résidus qui montre, grâce à un test de Shapiro-Wilk, une distribution des erreurs très forte autour de la moyenne et très faible quand ont s'en éloigne d'où une P-value très significativement petite (7.682e-06). De plus à l'issue des test de Student sur les coefficients de la droite, les P-values étant hautement significatives nous avons donc une présomption très forte contre l'hypothèse nulle. Cela prouve qu'il y a une forte dépendance entre la baisse de la concentration en microalgues et la hausse de température.

Autrement dit: concentration d'algue = b1*différence de température+ b0 + ei (résidus) avec deux degrés de liberté.

L'intervalle de confiance à 95% renseigne sur les coefficients de la vraie population et nous permet d'élargir l'échantillon à la population considérée.

Discussion:

Malgré la part de la distribution de donnée expliquée a 92%, l'existence de multiple facteur d'imprécision comme ceux lié aux mesures de concentrations ou de températures, de l'acidité du CO2 atmosphérique absorbé sur les coraux, la salinité de l'eau et la température non homogène du milieux qui peu être du à de courant locaux venant de région plus froide, la profondeur des coraux étudiés qui est en lien directe avec la température, les conditions climatiques du jour de prélèvement, la proximité des activité humaine (touristique ou autre).... Les données devraient conduire à plus d'imprécisions mais elles semblent trop parfaites, on peut se permettre de douter de l'honnêteté scientifique des chercheurs dans le but d'avoir des financements.

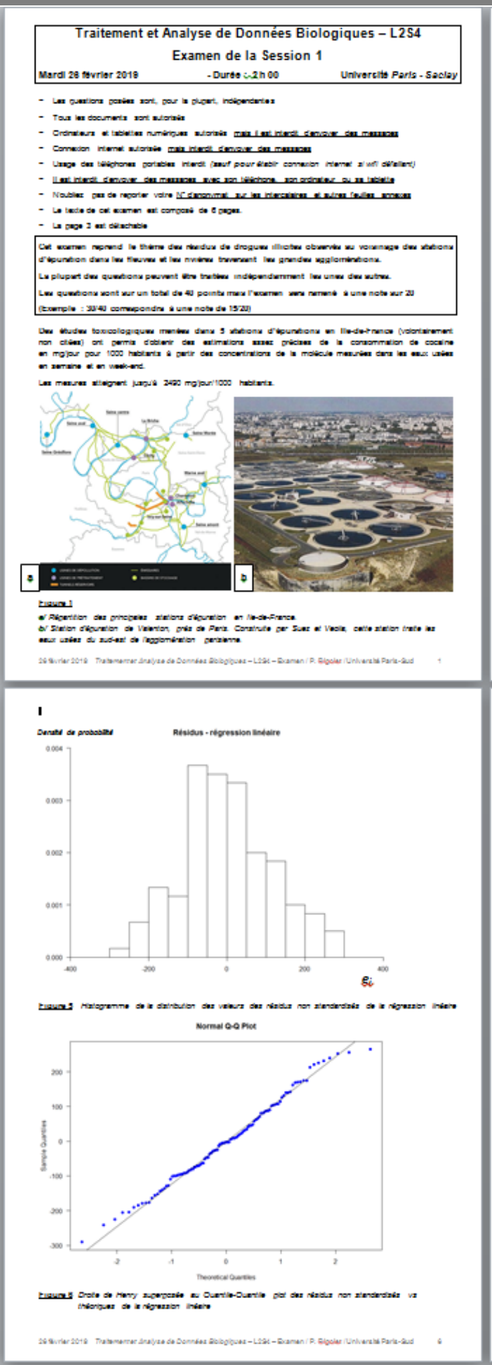
Conclusion:

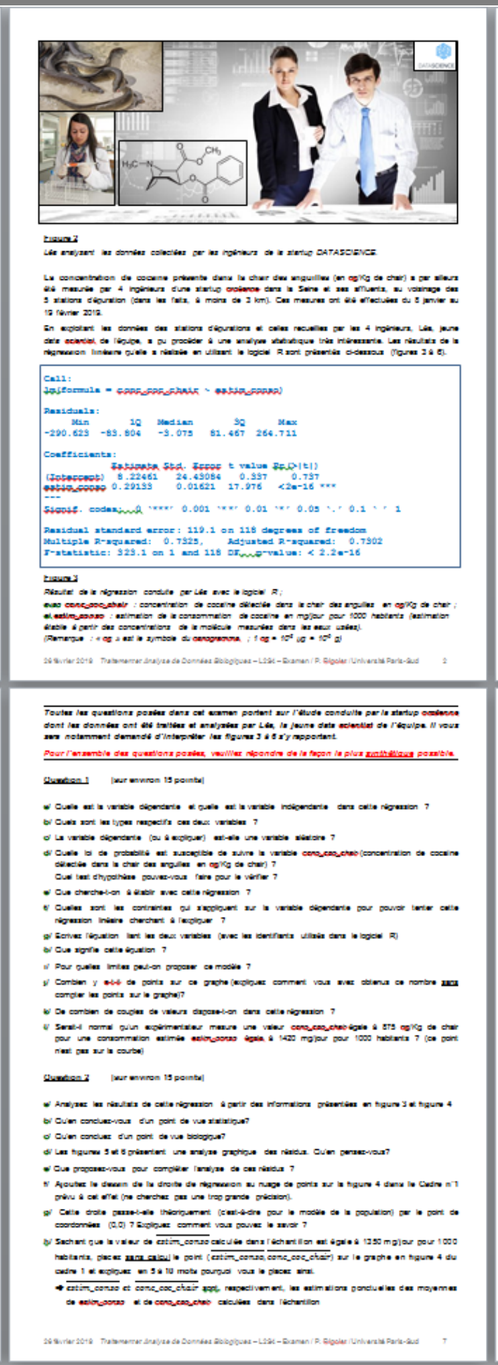
D'après notre analyse, la hausse des températures provoquée par le réchauffement climatique a bien un impact négatif sur la concentration en microalgues endosymbiotiques ce qui est à l'origine du blanchissement des coraux. Cependant d'autres échantillonnages semblent être nécessaire pour valider la fiabilité des données utilisées.

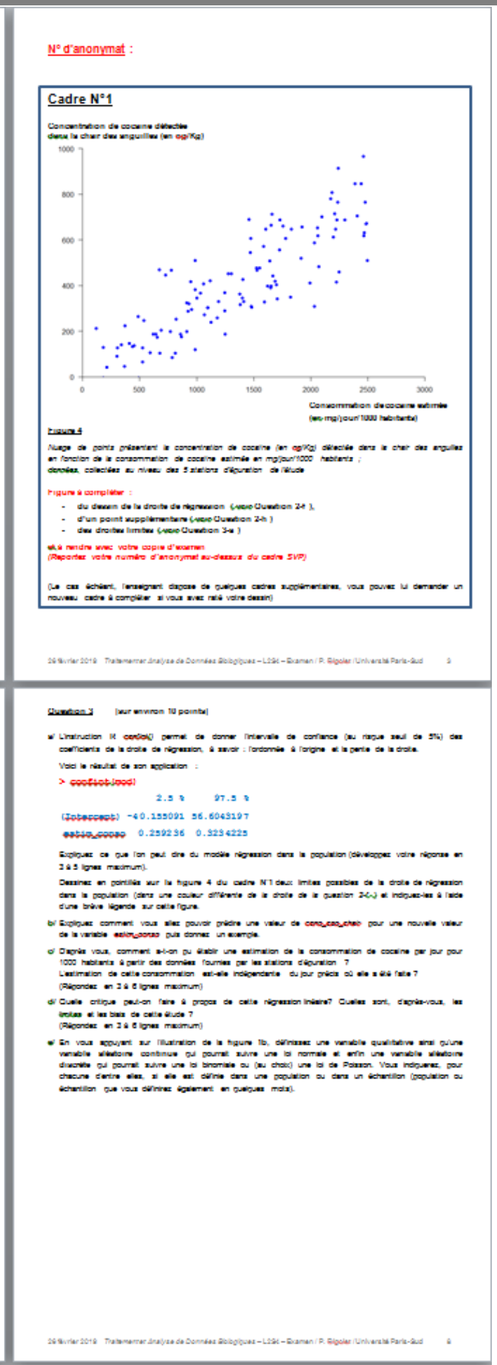
Les projets

- ⇨ **Travail en binôme**
- ⇨ **Thématiques au choix des étudiants**
- ⇨ **Format en 2 à 4 pages maxi**
- ⇨ **Style et présentation doivent être soignés**

Retours d'expériences :L'UE « Traitement et Analyse de Données en Biologie »


Traitement et Analyse de Données Biologiques – L2S4
Examen de la Session 1
Mardi 26 février 2019 - Durée : 2h 00 Université Paris - Saclay




N° d'anonymat :
Cadre N°1

Examen écrit [1/3]

L'examen écrit met l'accent non pas sur des calculs mais sur des stratégies de traitement des donnes et sur l'analyse de résultats présentés sous forme de graphiques et de tableaux.

Tous les documents sont autorisés

Ordinateurs et tablettes autorisées

Connexion internet autorisée (hors mail et réseaux sociaux)

NF = 0,4*EE + 0,3*CC + 0,3*CC TP

Retours d’expériences :L’UE « Traitement et Analyse de Données en Biologie »

L’enseignant semble satisfait lui aussi

L2S4 Biologie Santé – Février 2019

Les étudiants ont beaucoup apprécié cette formule ; bien plus motivés à apprendre de nouvelles connaissances et ils ont découvert que les stats constituent un formidable outil pour faire parler les données en biologie.

Evaluation de l’enseignement

Taux de satisfaction 2019 :

36,5 % l'ont beaucoup apprécié ;

54,5 % l'ont plutôt bien apprécié ;

9 % ne l'ont pas apprécié.

Des résultats encourageants face à un auditoire très exigeant.

Les étudiants ne voient plus les stats comme une matière absconde mais bien comme une composante de l'analyse indispensable de leurs données sur le terrain comme au laboratoire ou au bureau d’étude; un outil permettant de faire parler les données et d’aide à la décision et ils ont retrouvé une certaine forme de confiance en eux.