

Modèles Statistiques (MS)

ou comment réaliser une étude en incluant de l'analyse de données

Jean-Marc.Vincent@imag.fr
Adrien.Faure@inria.fr

Laboratoire LIG
Équipe-Projet Inria POLARIS / DATAMOVE

Grenoble 2019

UE MODÈLES STATISTIQUES

- 1 **ORGANISATION DE L'UE : Modèles Statistiques**
- 2 OBJECTIF DE L'UE
- 3 RÉFÉRENCES BIBLIOGRAPHIQUES
- 4 PROBLÉMATIQUE : QUELQUES EXEMPLES
- 5 REPRODUCTIBILITÉ : MOTIVATION Thanks to GitHub SMPE

ORGANISATION

Équipe pédagogique

Jean-Marc Vincent



Jean-Marc.Vincent@imag.fr

coordination de l'UE

Laboratoire d'informatique de Grenoble

Équipe Inria POLARIS/DATAMOVE

Évaluation de performances de
systèmes/réseaux/infrastructures à
grande échelle

Adrien Faure



Adrien.Faure@inria.fr

Laboratoire d'informatique de Grenoble

Équipe Inria DATAMOVE

Batch Scheduling, Large Computer
Systems Optimization

COMMUNICATION AVEC L'ÉQUIPE PÉDAGOGIQUE

Mail et adresses électroniques

Adresse Mail enseignant : Prénom.Nom@univ-grenoble-alpes.fr

SUJET : [MIAGE :MS] sujet explicite

envoyer votre mail avec votre adresse officielle **@etu.univ-grenoble-alpes.fr**
toute adresse de provenance différente risque d'être "grey/black-listée" et d'atterrir dans une poubelle

le mail officiel de la L3-MIAGE est la liste

etu-2018-im2ag-gbl3ie160@univ-grenoble-alpes.fr, toute annonce officielle (quicks, apnées, déplacements de créneaux horaires,...) passera par ce mail (que vous devez lire quotidiennement)

Destinataires

organisation/cours/examens... : Jean-Marc Vincent

les **Travaux Dirigés/Pratiques** : Adrien Faure

UE MODÈLES STATISTIQUES

- 1 ORGANISATION DE L'UE : Modèles Statistiques
- 2 OBJECTIF DE L'UE**
- 3 RÉFÉRENCES BIBLIOGRAPHIQUES
- 4 PROBLÉMATIQUE : QUELQUES EXEMPLES
- 5 REPRODUCTIBILITÉ : MOTIVATION Thanks to GitHub SMPE

OBJECTIF PÉDAGOGIQUE DE L'UE MODÈLES STATISTIQUES

Connaissances

Savoir réaliser une étude d'un objet informatique (ou autre) à partir de données observées :

(répondre à une question, formuler une hypothèse et la confirmer)

- ▶ savoir bâtir une expérimentation simple et produire des données d'observation
- ▶ savoir analyser les résultats obtenus (processus d'analyse)
- ▶ savoir restituer les résultats sous forme synthétique (processus de visualisation, commentaires, analyse et synthèse)

En pratique, savoir réaliser une étude argumentée et correctement présentée.

Savoir utiliser un/des environnement(s) adapté(s) :

- ▶ suivi des développements logiciels (historique, versionning, collaboration) : git, github
- ▶ processus d'analyse (analyse statistique, synthèse, visualisation) : R (R-studio, ggplot2)
- ▶ mise en forme et présentation : LaTeX (via un markdown)

ORGANISATION DE LA SEMAINE

Cours /TD : **guidelines** pour une étude rigoureuse et reproductible

Les cours/TD seront organisés à partir d'études de cas :

- ▶ une partie synthétique sur les **concepts**
- ▶ une partie sur des **exemples** illustrant les concepts
- ▶ une partie sur votre étude de cas

Forme du travail

- ▶ travail en binôme/quadrinôme
- ▶ travail public (partageable par toute la promotion (et même plus))
- ▶ synthèse en commun (production de fiches)

Travail personnel :

- ▶ prévoir 1 à 2h de travail en moyenne à la maison pour 1 séance de cours/TD ,
- ▶ exercices à la maison (pour préparer le matériel des séances suivantes)

Évaluation : une note d'UE

- ▶ mini-projet avec une présentation
- ▶ suivi des C/TD

CONTENU INDICATIF

Environnement

- ➊ Introduction / problème
- ➋ Programmation Littérale / RStudio / Rmd
- ➌ Visualisation élémentaire / ggplot2
- ➍ Guideline / Checklist for good graphics

Traitement de données

- ➎ Processus d'analyse /statistiques de base (rappels)
- ➏ Données importation pré et post traitement
- ➐ Manipulation de donnée expérimentales (dplyr)

Mini-projet

- ➑ Mini-projet : spécification/études préliminaires
- ➒ Mini-projet : étude et rapport

Présentation

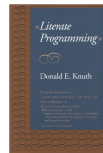
- ➓ Présentation orale (tous les étudiants)

UE MODÈLES STATISTIQUES

- 1 ORGANISATION DE L'UE : Modèles Statistiques
- 2 OBJECTIF DE L'UE
- 3 RÉFÉRENCES BIBLIOGRAPHIQUES**
- 4 PROBLÉMATIQUE : QUELQUES EXEMPLES
- 5 REPRODUCTIBILITÉ : MOTIVATION Thanks to GitHub SMPE

BIBLIOGRAPHIE : OUVRAGES DE RÉFÉRENCE DU COURS

- ▶ **R** Garrett Golemund and Hadley Wickham, R for Data Science, O'Reilly 2016
<http://r4ds.had.co.nz/>
- ▶ **Statistiques** Sheldon Ross Introductory Statistics. Academic Press 2010
Également les polycopiés de Frédérique Leblanc
<http://www-ljk.imag.fr/membres/Frederique.Leblanc/>
- ▶ **Historique** Donald E. Knuth Literate Programming. Academic Press 1983



et évidemment de nombreuses ressources sur le web ...

UE MODÈLES STATISTIQUES

- 1 ORGANISATION DE L'UE : Modèles Statistiques
- 2 OBJECTIF DE L'UE
- 3 RÉFÉRENCES BIBLIOGRAPHIQUES
- 4 PROBLÉMATIQUE : QUELQUES EXEMPLES**
- 5 REPRODUCTIBILITÉ : MOTIVATION Thanks to GitHub SMPE

UN OEIL CRITIQUE

Dernière mise à jour le dimanche 22 janvier 2017 à 00h45



Manuel Valls

31.11%

188603 voix



Sylvia Pinel

1.97%

24657 voix



Vincent Peillon

6.85%

85975 voix



François de Rugy

3.88%

48921 voix



Arnaud Montebourg

17.52%

218885 voix



Benoît Hamon

36.35%

454041 voix



Jean-Luc Bennahmias

1.01%

12609 voix








Votes blancs et nuls








1.3%

16235 voix

UN OEIL CRITIQUE

Dernière mise à jour le dimanche 22 janvier 2017 à 00h45

	Manuel Valls	31.11% 108605 voix
	Sylvia Pinel	1.97% 24657 voix
	Vincent Peillon	6.85% 85975 voix
	François de Rugy	3.88% 48921 voix
	Arnaud Montebourg	17.52% 218885 voix
	Benoît Hamon	36.35% 454041 voix
	Jean-Luc Bennahmias	1.01% 12609 voix
Votes blancs et nuls		1.3% 16235 voix

	Manuel Valls	31.11% 498114 voix
	Sylvia Pinel	1.98% 37703 voix
	Vincent Peillon	6.85% 109678 voix
	François de Rugy	3.88% 62124 voix
	Arnaud Montebourg	17.52% 280519 voix
	Benoît Hamon	36.35% 582014 voix
	Jean-Luc Bennahmias	1.01% 16172 voix
Votes blancs et nuls		1.3% 20815 voix

et lundi matin

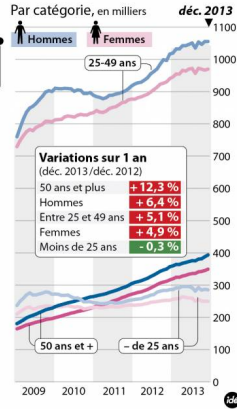
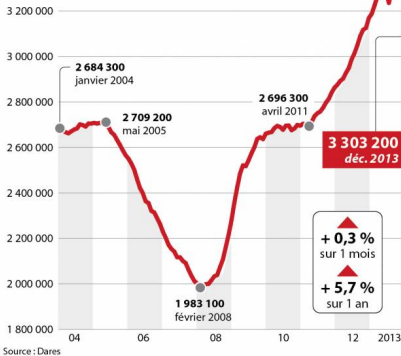
UN OEIL CRITIQUE (2)

- ▶ Performances <http://www.cpubenchmark.net/index.php>
- ▶ <http://www.tylervigen.com/spurious-correlations>
- ▶ <https://www.ined.fr/fr/tout-savoir-population/>
- ▶ taux de réussite au Bac

UN OEIL CRITIQUE (3)

Le chômage

Nombre de demandeurs d'emploi (catégorie A)



Extrait du journal Le Point 2013

UE MODÈLES STATISTIQUES

- 1 ORGANISATION DE L'UE : Modèles Statistiques
- 2 OBJECTIF DE L'UE
- 3 RÉFÉRENCES BIBLIOGRAPHIQUES
- 4 PROBLÉMATIQUE : QUELQUES EXEMPLES
- 5 **REPRODUCTIBILITÉ : MOTIVATION Thanks to GitHub SMPE**

FRUSTRATION AS AN AUTHOR

- ▶ I thought I used the same parameters but I'm getting different results !
- ▶ The new student wants to compare with the method I proposed last year
- ▶ My advisor asked me whether I took care of setting this or this but I can't remember
- ▶ The damned fourth reviewer asked for a major revision and wants me to change figure 3 :(
- ▶ Which code and which data set did I use to generate this figure ?
- ▶ It worked yesterday !
- ▶ 6 months later : why did I do that ?

FRUSTRATION AS A REVIEWER

This may be an interesting contribution but :

- ▶ This **average value** must hide something
- ▶ As usual, there is no **confidence interval**, I wonder about the variability and whether the difference is **significant** or not
- ▶ That can't be true, I'm sure they **removed some points**
- ▶ Why is this graph in **logscale** ? How would it look like otherwise ?
- ▶ The authors decided to show only a **subset of the data**. I wonder what the rest looks like
- ▶ There is no label/legend/. . . What is the **meaning of this graph** ? If only I could access the generation script

THE DOG ATE MY HOMEWORK !!!

► Versioning Problems

*Thanks for your interest in the implementation of our paper. The good news is that I was able to find some code. I am just **hoping** that **it** is a stable working version of the code, and **matches the implementation we finally used for the paper**. Unfortunately, I have **lost some data** when **my laptop was stolen** last year. The bad news is that the code is not commented and/or clean.*

*Attached is the `<system>` source code of our algorithm. I'm **not** very **sure whether it is the final version of the code used in our paper**, but it should be at least 99% close. Hope it will help.*

THE DOG ATE MY HOMEWORK !!!

- ▶ Versioning Problems
- ▶ Bad Backup Practices

*Unfortunately, the server in which my implementation was stored had a **disk crash in April and three disks crashed simultaneously**. While the help desk made significant effort to save the data, my entire implementation for this paper was not found.*

THE DOG ATE MY HOMEWORK !!!

- ▶ Versioning Problems
- ▶ Bad Backup Practices
- ▶ Code Will be Available Soon

*Unfortunately the current system is **not mature enough at the moment**, so it's not yet publicly available. We are actively working on a number of extensions and **things are somewhat volatile**. However, once things stabilize we plan to release it to outside users. At that point, we would be happy to send you a copy.*

THE DOG ATE MY HOMEWORK !!!

- ▶ Versioning Problems
- ▶ Bad Backup Practices
- ▶ Code Will be Available Soon
- ▶ No Intention to Release

*I am afraid that the source code was never released. The code was **never intended to be released so is not in any shape for general use.***

THE DOG ATE MY HOMEWORK !!!

- ▶ Versioning Problems
- ▶ Bad Backup Practices
- ▶ Code Will be Available Soon
- ▶ No Intention to Release
- ▶ Programmer Left

*⟨STUDENT⟩ was a graduate student in our program but **he left a while back** so I am responding instead. For the paper we used a prototype that included many moving pieces that only ⟨STUDENT⟩ knew how to operate and we did not have the time to integrate them in a ready-to-share implementation before he left. Still, I hope you can build on the ideas/technique of the paper.*

*Unfortunately, the author who has done most of the coding for this paper has **passed away** and the code is no longer maintained.*

THE DOG ATE MY HOMEWORK !!!

- ▶ Versioning Problems
 - ▶ Bad Backup Practices
 - ▶ Code Will be Available Soon
 - ▶ No Intention to Release
 - ▶ Programmer Left
- ▶ Commercial Code

*Since this work has been done at (COMPANY) **we don't open-source code** unless there is a compelling business reason to do so. So unfortunately I don't think we'll be able to share it with you.*

*The code **owned by (COMPANY)**, and AFAIK the code is not open-source. Your best bet is to reimplement :(Sorry.*

THE DOG ATE MY HOMEWORK !!!

- ▶ Versioning Problems
- ▶ Bad Backup Practices
- ▶ Code Will be Available Soon
- ▶ No Intention to Release
- ▶ Programmer Left
- ▶ Commercial Code
- ▶ Proprietary Academic Code

*Unfortunately, the $\langle \text{SYSTEM} \rangle$ sources are **not meant to be opensource** (the code is partially **property of $\langle \text{UNIVERSITY 1} \rangle$, $\langle \text{UNIVERSITY 2} \rangle$ and $\langle \text{UNIVERSITY 3} \rangle$).***

If this will change I will let you know, albeit I do not think there is an intention to make the $\langle \text{SYSTEM} \rangle$ sources opensource in the near future.

*If you're interested in obtaining the code, **we only ask for a description of the research project** that the code will be used in (**which may lead to some joint research**), and we also have a software license agreement that the University would need to sign.*

THE DOG ATE MY HOMEWORK !!!

- ▶ Versioning Problems
- ▶ Bad Backup Practices
- ▶ Code Will be Available Soon
- ▶ No Intention to Release
- ▶ Programmer Left
- ▶ Commercial Code
- ▶ Proprietary Academic Code
- ▶ **Research vs. Sharing**
- ▶ ...
- ▶ ...

In the past when we attempted to share it, we found ourselves spending more time getting outsiders up to speed than on our own research. So I finally had to establish the policy that we will not provide the source code outside the group.