Principes et Méthodes Statistiques

Introduction

Olivier Gaudoin

Ensimag

Olivier Gaudoin

olivier.gaudoin@univ-grenoble-alpes.fr
http ://www-ljk.imag.fr/membres/Olivier.Gaudoin/

- Ancien élève de l'Ensimag
- Professeur à l'Ensimag
- Directeur des relations internationales de l'école
- Chercheur au Laboratoire Jean Kuntzmann, équipe ASAR
- Domaine de recherche : Modélisation aléatoire et analyse statistique pour la fiabilité des systèmes
- Collaboration avec de nombreux partenaires universitaires (Hong Kong, Brésil, Norvège,...) et industriels (EDF, Schneider Electric, Thales, GRTgaz,...)

Introduction

- Définition et domaines d'application de la statistique
- La démarche statistique
- Objectifs du cours
- Fonctionnement du cours

Définition de la statistique

La **statistique** est la science dont l'objet est de recueillir, de traiter et d'analyser des **données** issues de l'observation de phénomènes **aléatoires**, c'est-à-dire dans lesquels le hasard intervient.

L'analyse des données est utilisée pour

- **Décrire** les phénomènes étudiés.
- **Faire des prévisions**.
- **Prendre des décisions** à leur sujet.

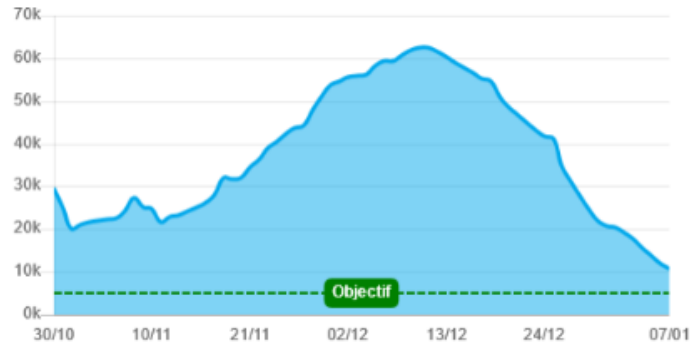
En cela, la statistique est un outil essentiel pour la compréhension et la gestion de phénomènes complexes de toute nature.

Exemple Covid 19 : description

Cas positifs

Plus

On prélève en moyenne **10 712 cas** positifs au Covid19 chaque jour, **en baisse (- 48 %)** par rapport à la semaine dernière (par date de prélèvement, J-3). Dépistage et cas ▸



Derniers chiffres : -- tests positifs remontés le 17/05 (SpF), 5 988 tests positifs prélevés le 07/01 (SI-DEP).

Adm. soins critiques

Plus

Il y a en moyenne **79 admissions en soins critiques** pour Covid19 chaque jour, **en baisse (- 25 %)** par rapport à la semaine dernière.

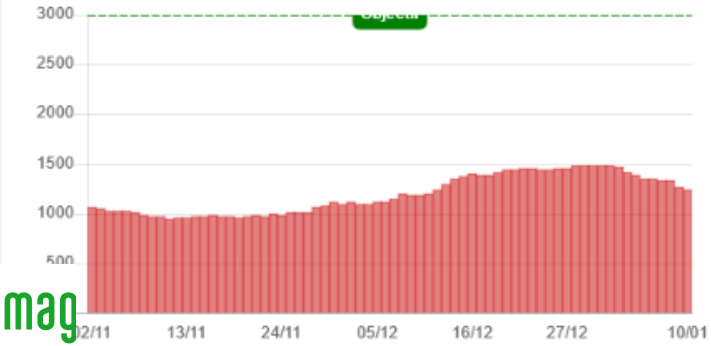
Soins critiques ▸ Hospitalisations ▸

Soins critiques

Plus

Il y a actuellement **1 242 personnes en soins critiques** pour Covid19, **en baisse (- 12 %)** par rapport à la semaine dernière.

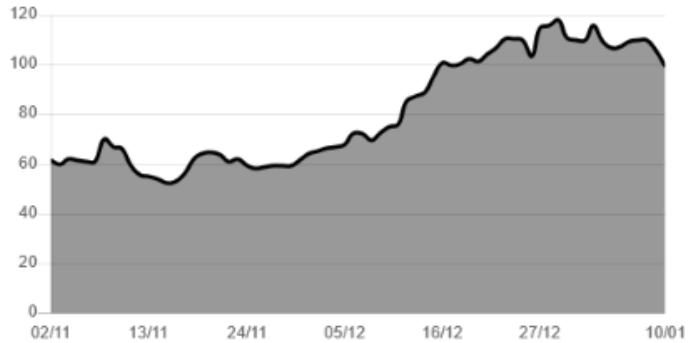
Soins critiques ▸ Hospitalisations ▸



Décès hospitaliers

Il y a en moyenne **99 décès hospitaliers** pour Covid19 chaque jour, **en baisse (- 9 %)** par rapport à la semaine dernière.

Décès hospitaliers ▸



Exemple Covid 19 : prévision et décision

- Prévoir l'évolution de l'épidémie.
  - Prévoir le nombre de contaminations, d'hospitalisations, d'admissions en réanimation, de décès à court et moyen terme.
  - Quand passera-t-on au-dessus ou au-dessous d'un certain seuil de contaminations par jour ?
  - Quand commencer ou finir un confinement ?
  - Modèles épidémiologiques (SIR,...)
- Déterminer les facteurs influents sur le développement des formes graves (âge, sexe, vaccination, comorbidité, groupe sanguin,...).
- Evaluer l'efficacité des traitements et des vaccins.
- ...

Autres exemples

- **Santé** : diagnostic médical, essais cliniques, analyse du génôme, détection des maladies génétiques, impact des OGM, ...
- **Neurosciences** : analyse des IRM ou électoencéphalogrammes pour déterminer les zones du cerveau réagissant à certains stimuli.
- **Environnement** : prévisions à court et à long terme du réchauffement climatique, prévision des pics de pollution, prévision de l'intensité et de la trajectoire des cyclones tropicaux,...
- **Sondages** : combien de personnes interroger, comment interpréter les résultats ?
- **Actuariat** : calculer les risques encourus par les clients des compagnies d'assurance et fixer le montant des primes.
- **Finance** : gestion des risques financiers, gestion de portefeuille,...
- **Transport** : voiture autonome, maîtrise des risques industriels,...

Science des données - Data science

Ensemble des méthodes permettant d'extraire des informations utiles des grandes masses de données (**big data**) :

- Statistique
  - Fouille de données (**data mining**)
  - Apprentissage automatique (**machine learning, deep learning**)
- Optimisation
- Bases de données
- Calcul parallèle, systèmes distribués
- Visualisation

$\Rightarrow$ domaine en pleine expansion (**Intelligence Artificielle**), profil maths-info idéal pour l'Ensimag

Incertitudes - Variations

Les données sont entâchées d'**incertitudes** et présentent des **variations** :

- le déroulement des phénomènes observés n'est pas prévisible à l'avance avec certitude,
- toute mesure est entâchée d'erreur,
- seuls quelques individus sont observés et on doit extrapoler les conclusions de l'étude à toute une population,
- etc...

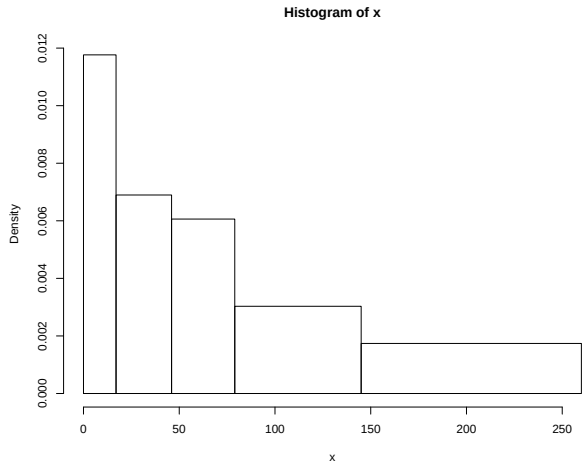
Il y a donc intervention du **hasard** et des **probabilités**.

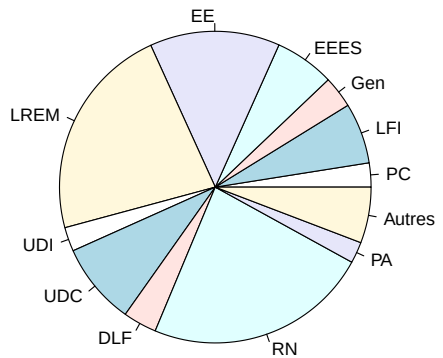
L'objectif essentiel de la statistique est de maîtriser au mieux cette incertitude pour extraire des informations utiles des données, par l'intermédiaire de l'analyse des variations dans les observations.

Deux classes de méthodes

Statistique descriptive

- Objectif : **résumer l'information** contenue dans les données de façon synthétique et efficace.
- Moyens : **représentations de données** sous forme de graphiques, de tableaux et d'indicateurs numériques.
- Elle permet de dégager les caractéristiques essentielles du phénomène étudié et de suggérer des hypothèses pour une étude ultérieure plus sophistiquée.
- Les probabilités n'ont ici qu'un rôle mineur.





Statistique inférentielle

- Objectif : **faire des prévisions** et **prendre des décisions** au vu des observations.
- Moyens : **modèles probabilistes** du phénomène aléatoire étudié.
- Il faut savoir gérer les risques d'erreurs.
- Les probabilités jouent ici un rôle fondamental.

La démarche statistique

La statistique et les probabilités sont les deux aspects complémentaires de l'étude des phénomènes aléatoires.

Les **probabilités** peuvent être envisagées comme une branche des mathématiques pures, basée sur la théorie de la mesure, abstraite et complètement déconnectée de la réalité.

Les **probabilités appliquées** proposent des **modèles probabilistes** du déroulement de phénomènes aléatoires concrets. On peut alors, **préalablement à toute expérience**, faire des prévisions sur ce qui va se produire.

Exemple : durée de fonctionnement d'une ampoule.

Modèle usuel : variable aléatoire X de loi exponentielle de paramètre λ.

Ayant adopté ce **modèle probabiliste**, on peut effectuer tous les calculs que l'on veut. Par exemple :

- La probabilité que l'ampoule ne soit pas encore tombée en panne à la date t est $P(X > t) = e^{-\lambda t}$.
- La durée de vie moyenne est $E[X] = 1/\lambda$.
- Si n ampoules identiques sont mises en fonctionnement en même temps, et qu'elles fonctionnent indépendamment les unes des autres, le nombre N_t d'ampoules qui tomberont en panne avant un instant t est une variable aléatoire de loi binomiale $\mathcal{B}(n, P(X \leq t)) = \mathcal{B}\left(n, 1 - e^{-\lambda t}\right)$. Donc on s'attend à ce que, en moyenne, $E[N_t] = n\left(1 - e^{-\lambda t}\right)$ ampoules tombent en panne entre 0 et t.

Questions

- La durée de vie est-elle bien de loi exponentielle ?
- Quelle est la valeur de λ ?
- Peut-on garantir que $E[X]$ est supérieure à une valeur fixée m_0 ?
- Sur un parc de 100 ampoules, à combien de pannes peut-on s'attendre en moins de 50 h ?
- ...

C'est la statistique qui va permettre de résoudre ces problèmes.

Pour cela, il faut faire une expérimentation, recueillir des données et les analyser.

Expérimentation

On fait fonctionner en parallèle et indépendamment les unes des autres $n = 10$ ampoules identiques, dans les mêmes conditions expérimentales, et on relève leurs durées de vie.

On obtient les durées de vie suivantes, exprimées en heures :

91.6 35.7 251.3 24.3 5.4 67.3 170.9 9.5 118.4 57.1

Notons $x_1, \ldots, x_n$ ces observations.

Observations et variables aléatoires

La durée de vie des ampoules n'est pas prévisible avec certitude à l'avance. $\Rightarrow$ on va considérer que $x_1, \ldots, x_n$ sont les **réalisations** de variables aléatoires $X_1, \ldots, X_n$.

Cela signifie qu'**avant l'expérience**, la durée de vie de la $i^{\text{ème}}$ ampoule est inconnue et que l'on traduit cette incertitude en modélisant cette durée par une variable aléatoire X_i.

Mais **après l'expérience**, la durée de vie a été observée. Il n'y a donc plus d'incertitude, cette durée est égale au réel x_i.

Ampoules identiques $\Rightarrow$ on suppose que les X_i sont de même loi de probabilité.

Les ampoules ont fonctionné indépendamment les unes des autres $\Rightarrow$ on suppose que les X_i sont des variables aléatoires indépendantes.

Réponses

L'analyse des observations va permettre de répondre aux questions :

- La durée de vie est-elle bien de loi exponentielle ?
  $\Rightarrow$ problème de **choix de modèle** ou de **test d'adéquation**.
- Quelle est la valeur de λ ?
  $\Rightarrow$ problème d'**estimation paramétrique**.
- Peut-on garantir que $E[X]$ est supérieure à une valeur fixée m_0 ?
  $\Rightarrow$ problème de **test d'hypothèses paramétriques**.
- Sur un parc de 100 ampoules, à combien de pannes peut-on s'attendre en moins de 50 h ?
  $\Rightarrow$ problème de **prévision**.

En répondant à ces questions, il est possible que l'on se trompe. Donc, à toute réponse statistique, il faudra associer le **degré de confiance** que l'on peut accorder à cette réponse.

Probabilités et Statistique

La démarche probabiliste suppose que la nature du hasard est connue. Cela signifie que l'on adopte un modèle probabiliste particulier (ici la loi exponentielle), qui permettra d'effectuer des prévisions sur les observations futures.

Dans la pratique, la nature du hasard est inconnue. La statistique va, au vu des observations, formuler des hypothèses sur la nature du phénomène aléatoire étudié. Maîtriser au mieux cette incertitude permettra de traiter les données disponibles.

Probabilités et statistiques agissent donc en aller-retour dans le traitement mathématique des phénomènes aléatoires.

Objectifs du cours

- Présenter les principes de base d'une analyse statistique de données : description, estimation, tests.
- Présenter les méthodes statistiques les plus usuelles : histogramme, estimation par maximum de vraisemblance, test du χ^2, régression linéaire,...
- Les méthodes statistiques seront la plupart du temps justifiées mathématiquement. Néanmoins, le cours privilégie l'application à la théorie. Les notions introduites seront toujours illustrées par des problèmes concrets.
- Approfondissements théoriques et applicatifs en 2A et 3A.
- Les méthodes présentées seront mises en œuvre à l'aide du logiciel `R`. La plupart du temps, on associera à chaque méthode la syntaxe et les sorties (tableaux, graphiques) correspondantes de `R`.

Plan du cours

- Introduction et Statistique descriptive : 2 semaines.
- Estimation ponctuelle : 2 semaines.
- Intervalles de confiance : 2 semaines.
- Tests d'hypothèses : 3 semaines.
- Régression linéaire : 2 semaines.

Fonctionnement du cours

- 1 cours magistral et 1 TD par semaine.
- 2 séances de TD en R en salles informatiques, semaines 3 et 7 (cf ADE).
- 3 permanences de 45 mn (cf ADE).
- La référence du cours est le **poly**, en ligne sur Chamilo.
- En TD, consulter régulièrement les annexes A (bases de probabilité) et B (tables de lois de probabilité usuelles) du poly.
- Nombreux autres documents sur Chamilo.

Vidéos

- Confinement 2021 : tous les cours ont été donnés sur zoom et ont été enregistrés.
- Les vidéos sont disponibles sur Chamilo.
- Vous pouvez choisir de suivre les cours en amphi ou de regarder les vidéos.
- **Indispensable** : avoir suivi le cours avant de venir en TD.

Examen

- Examen en salle machine :
  - Une partie de résolution sur papier d'exercices type TD.
  - Une partie d'utilisation de R.
  - Vous disposerez d'un répertoire avec la doc utile (poly, fiches de TD,...) et R mais pas d'accès à internet.
  - Documents supplémentaires autorisés : toutes vos notes manuscrites.
- Critère d'évaluation : voir fiche sur les compétences à acquérir en PMS.
- Compte pour 3/4 de la note finale.

TP

- Sujet portant sur une application de la statistique.
- Sujet diffusé fin février.
- Une partie théorique et une partie à réaliser en R.
- Vous constituerez des trinômes au sein d'un groupe de TD.
- Un compte-rendu à rendre fin avril.
- Compte pour 1/4 de la note finale.
- Equipes et dépôt du compte-rendu sur Teide.