

# Refonte du cours de statistique dans une école de commerce

## Expérience commentée

Gilles Stoltz

CNRS — HEC Paris



Contenu et déroulement du cours

Contextes interne et externe lors de la refonte

## Contenu

- Essentiellement de l'inférentiel (intervalles de confiance, tests de comparaison à une valeur de référence, régression linéaire multiple)
- En 1ère année, soit niveau L3 (après classes préparatoires)
- Autres cours de statistique uniquement en 3ème année (selon la spécialisation)

## Organisation

- Ce cours existe depuis belle lurette... mais refonte entamée à la rentrée 2007
- Promotion de 1ère année = 360 étudiants environ, soit 9 groupes de 40 étudiants
- Cours-TDs assurés par une équipe de 4 enseignants, dont un coordinateur/concepteur du cours (moi !)
- 13 séances de 1h30, plus 1h30 de TP (sous SPSS) en demi-groupes, soit 21h au total

## Contexte interne à HEC Paris

### Renouvellement du corps professoral

- Départs à la retraite massifs des enseignants recrutés dans les années 70
- Changement du profil de recrutement : enseignants-chercheurs et non plus enseignants-consultants

### Désirs concomitants de contenus plus académiques

- Passage de la scolarité au format LMD : la 1<sup>ère</sup> année (L3) devient une année de formation fondamentale, suivie de deux années (M1/M2) de formation plus spécialisée
- Mission claire confiée par le doyen de la Grande école HEC lors de mon recrutement : appliquer cette réforme au cours de statistique

## Contexte médiatique

Thèse défendue par Florence Noiville :

Ne pas enseigner (que) l'art et la manière d'augmenter le profit dans les entreprises, mais aussi s'intéresser aux aspects éthiques, humains, et citoyens



## Difficultés rencontrées et solutions apportées

## Difficulté 1 : « Comprendre l'énoncé »

« Le plus difficile pour les étudiants est de comprendre l'énoncé »  
(dixit les collègues partant à la retraite)

En clair : transformer l'histoire racontée en des formules, i.e.,  
arriver à modéliser et à formaliser la situation rencontrée.

### Exemple d'exercice

Sur 100 fumeurs choisis au hasard pour tester un traitement anti-tabac, 64 arrêtent de fumer au moins temporairement. Le laboratoire peut-il garantir formellement une efficacité strictement supérieure à 60 % ?

Autre difficulté, plus habituelle : comprendre qu'il existe un paramètre inconnu (par opposition aux cours de probabilités antérieurs)

## Solution : Une séance entière sur la démarche de modélisation

- Population visée : les fumeurs désireux de s'arrêter
- Table de correspondance : on note dans la suite 1 lorsque le patch a été efficace, 0 lorsqu'il ne l'a pas été
- Données :  $x_1, \dots, x_{100}$ , appartenant à l'ensemble  $\{0, 1\}$
- Description des données : on observe un taux d'échantillon de

$$\bar{x}_{100} = \frac{1}{100} \sum_{i=1}^{100} x_i = 64 \%$$

- Modélisation : vu la sélection aléatoire des cobayes, vu l'étendue  $\{0, 1\}$ , on peut considérer que les données sont issues de la réalisation de  $X_1, \dots, X_{100}$  i.i.d. selon une loi de Bernoulli de paramètre  $p_0 \in [0, 1]$ , inconnu
- Interprétation :  $p_0$  est la proportion d'arrêts au moins temporaires que l'on observerait si on commercialisait le patch [Attention à faire une belle phrase complète ici...]



## Difficulté 2 : Inhomogénéité des niveaux

Profil de recrutement homogène (= après classes préparatoires)...  
... mais nombreuses filières sources, aux programmes de mathématiques fort différents : EC/S, EC/E, B/L, EC/T et A/L !

Ancienne solution : un cours de niveau minimal, les étudiants issus d'EC/S se reposent sur leurs lauriers

Mon contrat : stimuler davantage les étudiants, exploiter les acquis de probabilités de classes préparatoires, faire le pont avec eux

[C'est une spécificité du cours qui surprend les professeurs étrangers.]

Idée préliminaire de la direction, fondée sur les évaluations obtenues jusque-là : créer **deux cours** de statistique de **niveaux différents**

→ Plus de soucis engendrés (sentiment d'injustice et non de chance) que de bénéfices obtenus (étancher la soif de curiosité) !

## Solution : Faire entendre la musique mathématique, mais évaluer sur l'application des méthodes

Seuls outils nécessaires = théorème de la limite centrale (et parfois le lemme de Slutsky)

→ Explication **culturelle** possible du fonctionnement des méthodes (permet de voir leurs limites, p.ex., leur caractère asymptotique)

Mais l'évaluation des étudiants porte sur l'application de ces méthodes :

PRINCIPE 7.3. *Test de comparaison d'une proportion de population  $p_0$  à une valeur de référence  $p_{ref}$*

**Données :**  $x_1, \dots, x_n \in \{0, 1\}$

**Modélisation associée :** observations  $X_1, \dots, X_n$  indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre  $p_0$

**Hypothèse  $H_0$  :**  $p_0 = p_{ref}$

**Statistique de test :**

$$T_n = \sqrt{n} \frac{\bar{X}_n - p_{ref}}{\sqrt{p_{ref}(1 - p_{ref})}}$$

**Comportement sous  $H_0$  :**  $T_n \rightarrow \mathcal{N}(0, 1)$

**Comportement sous  $H_1$  :** lorsque  $p_0 > p_{ref}$ , la statistique  $T_n$  tend à prendre des valeurs plus grandes que sous  $H_0$  ; lorsque  $p_0 < p_{ref}$ , la statistique  $T_n$  tend à prendre des valeurs plus petites que sous  $H_0$ .

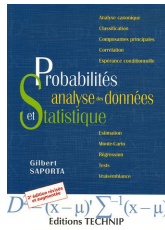
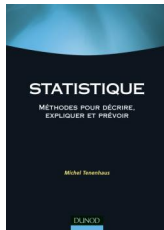
## Solution : Faire entendre la musique mathématique, mais évaluer sur l'application des méthodes

### Solution paradoxale ?

- Formalisation mathématique poussée
- Mais absence de preuves

### Objectifs :

Apprendre à apprendre (comme il se doit dans l'enseignement supérieur), afin qu'ils puissent ultérieurement s'auto-former



## Difficulté 3 : Absence de légitimité de la matière

- Les étudiants voient les mathématiques comme matière de sélection au concours, sont très surpris (et chagrinés) de les revoir
  - Pour eux, l'enseignement à HEC = Comptabilité ou Finance
  - Pensent (à tort) qu'ils auront une armée de grouillots pour récupérer et traiter leurs données lors de leur premier poste
- La statistique, un outil méprisable dont on peut confier la mise en œuvre à d'autres ?

Les stages en entreprise ou les missions JE se chargent certes de les convaincre du contraire, mais d'ici là...

## Solutions faciles mais qui ne durent qu'un moment

**20 SPORT**

MERCREDI 16 NOVEMBRE 2011

**ANALYSE** De plus en plus de clubs de L1 se servent de ces données pour rationaliser la vie de leur club

# LES STATS ONT LA COTE

**ANTOINE MAES**

Il y a la fiction. Où Brad Pitt révolutionne le base-ball en incarnant Billy Beane, coach des A's d'Oakland en 2002, dont le recours massif aux statistiques lui permet d'emmener son équipe vers le record de victoires consécutives en MLB. *Le Stratège* sort aujourd'hui en salle. Et il y a la réalité : en 2011, alors que les championnats anglais de foot de première et deuxième division se sont massivement tournés vers les chiffres – à l'instar de ce qui se fait dans les sports américains depuis longtemps –, pour mieux comprendre les performances de leurs joueurs, la Ligue 1 a un train de retard. « Mais c'est en voie de démocratisation », indique Olivier Broca, directeur commercial d'Amisco, société niçoise qui fournit en données des clubs comme Lille, Lyon, Rennes, Toulouse...



## Solutions côté entreprise

Rebondir sur les illusions des étudiants :

- Pour diriger une éventuelle armée de grouillots : savoir ce que font ces derniers !

Plus sérieusement, les faire passer du rôle passif d'**étudiant** au rôle actif de **décideur** :

- Justifier la forme des intervalles de confiance en fonction du **commanditaire** et de ses buts
- Souligner le côté parfois **subjectif** du traitement statistique (p.ex. choix des hypothèses d'un test ou non-conclusion parce que conservation de  $H_0$ ) → ne pas se faire embobiner !
- Avoir du **recul** sur les chiffres qu'on présente / qu'on se voit présenter → Elgozy : « Dans toute statistique, l'inexactitude du nombre est compensée par la précision des décimales. »
- Fondamentalement, leur valeur ajoutée = l'interprétation et exploitation **stratégique**

## Énoncé

Sur 100 fumeurs choisis au hasard pour tester un traitement anti-tabac, 64 arrêtent de fumer au moins temporairement. Le laboratoire peut garantir formellement une efficacité strictement supérieure à 60 % ?

## Résolution

P-valeur de  $H_0 : p_0 \leq 60\%$  contre  $H_1 : p_0 > 60\%$  → vaut 20.6%

Constat statistique : non, le laboratoire ne peut pas garantir formellement le taux de 60 % d'efficacité.

Conclusion stratégique ?

- Peu imaginative : « Il faut que le service R&D améliore le produit pendant que le service marketing prend des vacances »
- Plus intéressante : communiquer sur un taux garanti de 55 % !

## Solutions côté citoyen, politique et économique

H.G. Wells : « Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write »

C'est p.ex. tout ce qui tourne autour des sondages et de leur précision à  $\pm 3\%$  qui nous intéresse.

Pan sur le bec du Canard enchaîné :

**ards**

**minimales!**

- **Noté**, ce titre du « Monde » (22/8) : « Niches fiscales : M. Sarkozy annonce 10 milliards d'euros de hausses d'impôts ». Même avec les riches, il faut appeler un chat un chat.
- **Le titre**, dans « Les Echos » (20/8) : « Voté il y a deux semaines, le projet fiscal est resté un talon d'Achille pour la majorité ». Un talon menacé de rupture.
- **Le « JDD »** (22/8) consacre un quart de page à expliquer la hausse de 2 points de la cote de Sarkozy, selon l'Ifop. Manque de bol, un sondage Vivavoice pour « Libé » (23/8), réalisé aux mêmes dates, le donne en baisse de 1 point. Il en faudrait, des tiers de page, pour expliquer les contradictions entre sondeurs !
- **Cours de fiscalité** donné par le secrétaire général de l'UMP, Xavier Bertrand (« LE Parisien », 21/8) : « Baisser la réduction d'impôts de certains, ce n'est pas augmenter les impôts pour tous. » Ni toucher
- **Ségolène Royal** a rope 1, 21/8) les p Sarkozy sur la sécurité de l'emploi. « agitation française ». Seulement :
- **La même Ségolène** Figaro », 18/8) : « J' sur moi et mon père les hommes, donc voir, donc moi fille ! forcément castré que » Je » parle de
- **De Vincent Peillon** (ancien PS (« Le Parisien ») : « On a laissé se continuer l'oligarchie, même fonctionnaires passés ministériels, liés à la politique, faisant des fortunes dans le privé. » Peillon parler aux anciens cabinets Fabius ou aujourd'hui salariés c'est normal.
- **En marge** des journaux

omme. C'est à l'intéressé,



## Solutions côté citoyen, politique et économique

H.G. Wells : « Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write »

C'est p.ex. tout ce qui tourne autour des sondages et de leur précision à  $\pm 3\%$  qui nous intéresse.

Ou encore : existence de tricheries électorales en Iran, embellissement par Mendel de ses données, détections de falsifications dans les balances des paiements des Etats, etc.

→ Les ouvrir au vaste monde, aux autres disciplines ; pas de repli sur des jeux de données les concernant de trop près ( $\neq$  cours plus à l'anglo-saxonne)

## Difficulté 4 : Comportement et attentes des étudiants

- Temps des étudiants = beaucoup plus immédiat ; ils veulent surtout savoir comment ils seront notés et ce qui les attendra à l'examen
- La validation des cours est la norme (→ néologisme « dévalider »)
- Phase régressive et de repos complet après la prépa, concentration sur les événements sociaux comme les **campagnes JE / BDE**
- Difficulté de la transition entre un travail très encadré vers un travail universitaire en autonomie

Il faut aussi traiter ces aspects, sans quoi tout le reste est vain...

Et en particulier, bien en informer les nouveaux intervenants !

## Solutions : Encadrer plus que de raison... cf. petits groupes

- Vérification de la présence en cours, police du comportement (bavardages, téléphones ou ordinateurs portables)
- Forçage du **travail régulier** par des contrôles courts (pas de QCM : exercices à rédiger, cf. importance des interprétations, des conclusions stratégiques)
- Compensation/compromis : travail facilité par l'élaboration de **fiches de synthèse** (cours condensé = juste l'exigible)
- Objectifs court terme précisés et incarnés (fourniture des **annales** des quizz et examens)

[Note : Cela nécessite d'écrire des examens qui servent (avec un contexte entreprise) ou interpellent (en lien avec la vie citoyenne, p.ex. prévision des élections présidentielles 2012 par modèle de régression linéaire), cf. attention maximale lors d'un exercice d'annales !]

Ce qui en revanche est inutile :

- Leur faire remarquer que leurs anciens camarades, en stage, butent devant des problèmes statistiques même très simples

# Conclusions

## Titre du cours

« Éléments de statistique pour citoyens d'aujourd'hui et managers de demain »

Polycopié librement disponible sur ma page web

<http://www.hec.fr/stoltz>

(exemplaire imprimé expédié sur simple demande... fin septembre)

« Voilà ce que j'aurais aimé savoir sur les étudiants en école de commerce il y a 5 ans ! »

Attention aux enseignements de service au sein du groupe

Enseignement de la SFdS, notamment en direction des plus jeunes

→ d'où cette session !

## Exemple de séquence pédagogique

### Les tests du $\chi^2$ (ajustement et indépendance)

## Motivations et objectifs

$\chi^2$  d'ajustement → aspects **citoyens**

- Permet notamment de détecter les manipulations de données
- Cf. « I only believe in statistics that I doctored myself »
- « Si on était à Poudlard, ce cours serait donc un cours de défense contre les forces du mal et je serais Severus Rogue »

$\chi^2$  d'indépendance → aspects **marketing**

- Telle sous-population se comporte-t-elle comme telle autre ?  
(Faut-il segmenter l'offre ou non ?)

## Musique mathématique ( $\chi^2$ d'ajustement)

- Observations i.i.d.  $X_1, \dots, X_n$ ; observations qualitatives, classes indexées par  $j$
- Comparer les effectifs observés  $N_{j,n}$  à ceux attendus,  $n p_j^{\text{ref}}$
- Quantification globale des écarts
- Comportement sous  $H_0$  découle du TCL, donc test asymptotique

### Principe de test

PRINCIPE 9.1. *Test d'ajustement simple à une loi de référence  $\mathbf{p}^{\text{ref}}$*

**Données :**  $x_1, \dots, x_n \in \{1, \dots, k\}$

**Modélisation associée :** observations  $X_1, \dots, X_n$  indépendantes et identiquement distribuées selon une certaine loi  $\mathbf{p}$  sur  $\{1, \dots, k\}$

**Hypothèse  $H_0$  :**  $\mathbf{p} = \mathbf{p}^{\text{ref}}$

**Statistique de test :**

$$D_n(\mathbf{p}^{\text{ref}}) \stackrel{\text{not.}}{=} n \sum_{j=1}^k \frac{(\hat{p}_{j,n} - p_j^{\text{ref}})^2}{p_j^{\text{ref}}} = \sum_{j=1}^k \frac{(N_{j,n} - n p_j^{\text{ref}})^2}{n p_j^{\text{ref}}}$$

**Comportement sous  $H_0$  :**  $D_n(\mathbf{p}^{\text{ref}}) \rightarrow \chi_{k-1}^2$  lorsque  $n \rightarrow \infty$

**Comportement sous  $H_1$  :**  $D_n(\mathbf{p}^{\text{ref}})$  tend vers  $+\infty$  et prend donc des valeurs beaucoup plus grandes que sous  $H_0$ .



## Tricheries électorales en Iran ?

4 candidats, 29 provinces, soit 116 nombres (grands) de votes  
→ Test d'uniformité du dernier chiffre

### Test du Khi-deux

Chiffre			
	Effectif observé	Effectif théorique	Résidu
0	9	11,6	-2,6
1	11	11,6	-,6
2	8	11,6	-3,6
3	9	11,6	-2,6
4	10	11,6	-1,6
5	5	11,6	-6,6
6	14	11,6	2,4
7	20	11,6	8,4
8	17	11,6	5,4
9	13	11,6	1,4
Total	116		

Test	
	Chiffre
Khi-deux	15,552 <sup>a</sup>
ddl	9
Signification asymptotique	,077

a. 0 cellules (,0%) ont des fréquences théoriques inférieures à 5. La fréquence théorique minimum d'une cellule est 11,6.


## Autres exemples pour le $\chi^2$ d'ajustement

- Représentativité d'un échantillon d'étudiants du campus HEC (cf. étude professionnelle lors de la campagne JE)
- Test du premier chiffre significatif (loi de Benford) pour vérifier la sincérité des balances commerciales des Etats
- Expériences de Mendel pour valider sa théorie du hasard génétique → résultats embellis (P-valeurs toujours supérieures à 90 %)

## Exemple pour le test d'indépendance


Opinions politiques en fonction de l'année de scolarité

[Note : tableau de résultats très critiqueable, tailles d'échantillons non reportées !]



**Hec Sondages**

- Espace membres
- ↳ Répondre aux sondages
- ↳ Inscription au panel
- ↳ Le Monopoliard
- ↳ Nos sondages
- ↳ Dernier sondage
- ↳ Anciens sondages
- ↳ Commander un sondage
- ↳ Divers
- ↳ Forum de discussion
- ↳ télécharger les cours
- ↳ Accès administrateurs
- ↳ Fiche technique
- ↳ Qui sommes-nous ?
- ↳ CampusHEC



l'association qui passe le campus à la loupe !

### HEC Sondages

l'Institut de sondages du campus HEC

---

#### Résultats du sondage sur les élections présidentielles

Question 1: Voteriez-vous pour :

Ségolène Royal	14 %
<b>Nicolas Sarkozy</b>	<b>41 %</b>
un(e) autre candidat(e)	20 %
Je ne sais pas encore	22 %
NSPP	3 %

*À mon avis les échantillons sont faibles ...*

**Résultats détaillés**

	1A	1A	Bâtiment	Bâtiment	2A	2A	Bâtiment	Bâtiment	VM	3A et
	filles	garçons	L	M	filles	garçons	C	D		Mastères
Ségolène Royal	13 %	7 %	13 %	7 %	0 %	16 %	20 %	31 %	24 %	7 %
Nicolas Sarkozy	37 %	40 %	37 %	40 %	26 %	45 %	40 %	0 %	43 %	78 %
un(e) autre candidat(e)	10 %	29 %	19 %	24 %	26 %	19 %	30 %	31 %	18 %	0 %
Je ne sais pas encore	37 %	19 %	24 %	29 %	48 %	9 %	10 %	38 %	15 %	15 %
NSPP	3 %	4 %	8 %	0 %	0 %	11 %	0 %	0 %	0 %	0 %