L1 AES (ISEM) - semestre 2 - Cours & TD Statistiques descriptives (½TQA-2) - 2013-14,

Calendrier : Calendrier universitaire
8 séances de cours de 1.5H (le mardi 9h-10h30, SJA 1, amphi 3), Premier cours le 4 février 2014 ;
8 séances de TD de 1.5 H. , premier TD la semaine du 17 février.

Contrôle continu : deux interrogations en TD (30mn) les mercredi 12 mars et 2 avril.
Au programme de la 1ère interro :

Tutorat ?

Présentation du cours

Progression du cours :

1. (4 fev) Présentation du cours : page web du cours, exemples de documents publiés.
Rappel du 1er semestre : vocabulaire, qq objets (voir document projeté).
Quelques objectifs de la Statistique descriptive (résumer les données brutes, distinguer "ce qu'il y a à voir", comparer des populations , modéliser la population).
Quelques objectifs de la statistique inférentielle : Approximer avec un échantillon, modéliser une population.
Remarques sur les sources de données utilisées dans le cours et les TD : données réelles, simulations. Logiciels

Lecture : Extrait de publications de l'INSEE ou d'autres organismes avec des objets au programme du cours.

En attente : illustration de "ce qu'il y a à voir"

2. (11 fev) Agrégation des données brutes pour 2 caractères (ou variables) X,Y ou plus : tableau d'effectifs, fréquences (conjoints, marginals). Vocabulaire : données brutes ou série statistique ; tableau d'effectif ou tableau de contingence ou distribution des valeurs. Notations N, n_x ou $n_{X=x}$, f_x,$n_{x,y}$ (ou $n_{X=x,Y=y}$), $f_{x,y}$
Relation $N = \Sigma_x n_x = \Sigma_{x,y} n_{x,y}$; $1 = \Sigma_x f_x = \Sigma_{x,y} f_{x,y}$.
Une première fréquence conditionnelle (fréquence des filles en licence de Droit dans le tableau de l'Insee). Cf le Document projeté.

3. (17 fev) Rappel : information sur une population via des caractères X,Y,... ; information statistique et non sur un individu particulier (information "people") : elle n'est pas modifiée si on enlève de la population un individu particulier (ex. médiane des salaires en France).
Classe (= une partie des valeurs prises par une variable X ou une famille de variables $(X_1,...,X_n)$), évènement (= un ensemble d'éventualités Cf langage des probalilités, souvent exprimé par une propriété des valeurs prises par une famille de variables), exemples ; sous population S_E déterminée par un evt E, evt certain Ω, evt impossible, négation de E : E^c (evt contraire), E et F, E ou F ; effectif n_E et fréquence f_E d'un evt E ; relations $n_\Omega=N$, $f_\Omega=1$, $f_{E^c}=1-f_E$, $n_{EouF}=n_E+n_F-n_{EetF}$, idem pour f_{EouF}. Fréquence conditionnelle $f_{E|F}$ (fréquence de E dans la sous pop. S_F) ; relations $n_{EetF} = f_{E|F}\times n_F$, $f_{EetF} = f_{E|F}\times f_F$.
Probabilité d'un evt E (mesure de l'incertitude sur E) ; mesure calculée à partir d'un modèle (ex. risque de collision d'une météorite avec la terre) ou issue d'une observation statistique (ex. probabilité qu'un composant électronique soit défectueux) ou un mélange des deux. En statistique descriptive : Probabilité qu'un evt E soit réalisé pour un individu choisi au hasard dans la population $P(E):=f_E$.

4. (4 mar) Probabilité conditionnelle P(E|F) calculée comme $f_{E|F}$. Formule de Bayes $f_{E|F}\times f_F = f_{F|E}\times f_E$ (= f_{EetF}), idem avec probabilité à la place de fréquence ; exemple construction du slogan "Fumer double le risque de maladie cardiovasculaire" avec la fréquence des fumeurs observée chez les malades du coeur et la fréquence estimée (par sondage par ex.) des fumeurs dans la population entière ; calcul $f_{Malade|Fumeur}=f_{Fumeur|Malade}\times f_{Malade}/f_{Fumeur}$, discours "Fumer multiplie les risques d'être malade par le nombre f_{Malade}/f_{Fumeur}" lorsque la population de référence est la population entière ; comparaison de $f_{Malade|Fumeur}$ avec $f_{Malade\ |\ nonFumeur}$ (de sorte que la population de référence soit celle des non-fumeurs) : calcul par conditionnement $f_E= f_{E|F}\times f_F+f_{E|nonF}\times f_{nonF}$ d'où une expression de $f_{Malade\ |\ nonFumeur}$. Exemples numériques (voir TD). Généralisation du calcul par conditionnement $f_E=f_{E|F1}\times f_{F1} + ... + f_{E|Fn}\times f_{Fn}$ sous l'hypothèse que chaque individu est concerné par un et un seul des evts F_i ; exemple avec le calcul de la proportion de filles à l'université avec le tableau des effectifs étudiants, Cf le document projeté.

Lecture : le langage des probabilités dans les rapports publics avec cet extrait du rapport 2007-2008 de l'ONPES

5. (11 mar) Indépendance/liaison entre deux evts E,F via le nombre $q=q_{E,F}= f_{EetF}/(f_E\times f_F) = f_{E|F}/f_E = f_{F|E}/f_F$ et discours associé : "F rend E q fois plus probable" ou "F augmente les chances de E de 100×(q-1) %" avec comme population de référence la population entière. Si q est proche de 1 (notation $q\approx 1$, concrètement pour ce cours : q entre 0.9 et 1.1 mais ça dépend du contexte), on dit qu'on observe pratiquement l'indépendance entre E et F dans la population étudiée (l'observation dépend de la population) ; si q est loin de 1 on dit qu'on observe une liaison entre E et F et la liaison est quantifiée par q. Exemple numérique avec la fréqeuntation du restaurant universitaire d'un campus science (voir TD). Cas extrême : q=0 (E et F sont disjoints ou exclusifs) ; $q=1/f_E$ (F rend E certain) ; $q=1/f_F$ (E rend F certain). On a toujours $0\le q\le \min\{1/f_E,1/f_F\}$. Supposons $q>>1$ (F rend E significativement plus probable); la liaison entre E et F est significative pour E si $f_{F|E}>>0$ (par ex. >0.1 mais cela dépend du contexte) ; exemple avec le diagnostic d'une maladie E par un test F : qualité du test.
Accentuation de l'observation d'une liaison en formant le quotient $f_{E|F}/f_{E|\ nonF}$ plutôt que $f_{E|F}/f_E$; ces deux nbres sont proches si f_F est proche de 0
On observe une liaison entre un caractère X et un evt F s'il existe un evt E exprimé en terme de X, significatif (ie $f_E>>0$) et lié à E. Liaison entre deux caractères X et Y s'il existe des evts E exprimé en terme de X et F exprimé en terme de Y, significatifs (f_E et $f_F>>0$) et liés entre eux.

6. (18 mar) Rappel : résumé d'un caractère X = (valeur centrale, mesure de la dispersion), cas d'une variable qualitative (mode, une représentation graphique des fréquences telle un camembert), cas d'une variable quantitative (médiane, intervalle interquartile ou ...) ou bien (moyenne, écart type). Résumé conditionnel d'un caractère X conditionnellement à un évènement E, noté Res(X|E) ; $Res(X|E)\approx Res(X)$ si X est pratiquement indépendant de E, application à l'observation d'une liaison entre X et E, plus généralement entre X et une variable qualitative Y (mais l'observation de $Res(X|E)\approx Res(X)$ ne suffit pas à observer l'absence de liaison). Exemples de résumés conditionnels dans les rapports publics et observation de liaisons avec ce document projeté.
Mesure de la liaison entre deux caractères X,Y par le nombre $\chi^2(X,Y)=N\times\sum_{x,y}(q_{x,y}-1)^2 f_x f_y$; On a $0 \le \chi^2 \le N\times\min\{r-1,s-1\}$ où r et s sont les nombres de valeurs prises par X et Y, interprétation des valeurs extrêmes.

Lectures : diagramme circulaire sur Wikipedia

7. (25 mar) Organisation du calcul du χ^2 via le tableau des effectifs conjoints, interprétation des valeurs extremes 0 (stricte indépendance) et $N\times\min\{r-1,s-1\}$ (l'un des caractères est une fonction de l'autre), exemples de calculs avec des tableaux d'effectifs. Rq : le χ^2 est plutôt utilisé en *statistique inférentielle* (test de l'hypothèse d'indépendance).
Moyenne d'un caractère quantitatif conditionnellement à un caractère qualitatif, calcul par conditionnement d'une moyenne, variance intra et inter-groupe, coefficient de corrélation $\eta^2_{X|Y}\in[0,1]$, interprétation des valeurs extrêmes 0 et 1, cf le résumé de cours de 2011-12.

Lectures : le test du χ^2 sur le blog Alea

8. (1 avr) Coefficient de correlation linéaire entre deux caractères quantitatifs, régression linéaire. Cf résumé de cours de 2011-12.

Documents de cours :

Feuille de TD 1, corrigé de l'ex. 2 : voir le sujet d'examen d'avril 2012 et son corrigé. Le Sujet et un corrigé d'un exercice analogue à l'ex. 3. Corrigé des exercices 5 et 4. Un corrigé des questions c-d-e d'un exercice analogue à l'ex. 6.
Feuille de TD 2, corrigé de l'ex. 4.b, corrigé de l'ex. 5 (question 2 dans le corrigé, les commentaires se rapportent au sujet B de l'examen de 2009-10), corrigé de l'ex.6 (question 6 dans le corrigé)
Sujet et corrigé (sujet A) de l'interrogation du 12 mars, barème (sur 10) : quest.1a : 1+1 pts ; quest.1b : 1+1+1 ; ex2a : 1 ; ex2b : 1+1.5+1.5
Feuille de TD 3, corrigé de l'ex. 1 (ex.5 dans le corrigé), ex2 : voir le corrigé de l'ex.1 de la feuille de td3 de mars 2012,
Feuille de TD 4,
Sujet et corrigé (sujet A-B) de l'interrogation du 2 avril, sujet et corrigé de son rattrapage du 23 avril. barème (sur 10) : ex.1a : 2pts ; ex.1b : 2 ; ex.1c : 1+1 ; ex.2a 1.5+1.5+1 ; ex.2b : 1
Examen TQA session 1 (analyse et statistiques)

La page du cours en 2012-13

Lectures :
[1] A. Hamon & N. Jégou, *Statistique descriptive*, Presse Univ. Rennes 2008. Disponible à la BU St Jean d'Angely.

[2] B.Escofoer-J.Pagès, *Initiation aux traitements statistiques*, Presses Univ. de Rennes 1997.

F-X. Dehon, Laboratoire J.A. Dieudonné, 21 janvier 2013