L1 AES (ISEM) - semestre 2 - Cours & TD Statistiques descriptives (½TQA) - 2011-12,

Calendrier : Calendrier universitaire, emploi du temps
10 séances de cours de 1.5H (le mardi 9h-12h, SJA 1, amphi 3), 5 séances de TD de 1.5 H. Premier cours le 17 janvier 2012, premier TD la semaine du 30 janvier.
Rq : cours d'analyse par Jérôme Vétois le mardi 9h-12h amphi 2.
TD par Brahim Benzeghli, Jérôme Vétois et moi-même.

¿Tutorat pour les deux cours de TQA (réponses aux questions sur le cours et les td) ?

Présentation du cours

Progression du cours :

1. (17 jan) Présentation du cours. Rappel des notions vues au 1er semestre (population étudiée par un caractère) : synthèse des données brutes, résumé en un ou quelques mots.
Quelques objectifs de la Statistique descriptive (résumer les données brutes, distinguer "ce qu'il y a à voir" , modéliser la population).
Notions nouvelles pour ce semestre : population étudiée à travers plusieurs caractères, étude de l'indépendance ou des liaisons entre les caractères.

Document projeté lors du 1er cours : page web du cours en 2010-11.

En attente : illustration de "ce qu'il y a à voir", exemples de publications grand public ou apparaissent les objets enseignés en cours.

2. (24 jan) Sur vidéo-projecteur :
- Quelques objets vus au 1er semestre.
- Extrait de publications de l'INSEE ou d'autres organismes avec des objets au programme du cours.
Un seul caractère (rappel du semestre 1) : données brutes, tableau d'effectifs, représentation graphique, tableau des fréquences. Voir le document projeté.
Deux caractères ou plus : effectifs conjoints, marginaux, fréquences conjointes, marginales, exemples de calculs. Document projeté.

3. (31 jan) Définitions :
- Evènement (évt) (sous forme d'hypothèse sur les valeurs prises par les caractères, ex "être inscrit en Licence de Sc. éco." pour la population des étudiants), opération sur les évènements : non(E), E et F, E ou F. Sous-population S_E déterminée par l'évt E. Effectif de S_E noté n_E, fréquence de l'évt E notée f_E. $n_E=f_E\times N$.
- Fréquence d'un évt E conditionné à un évt F (ou "sachant F") notée $f_{E|F}$ = fréquence de E dans la sous-population S_F. Relations $n_{EetF}=f_{E|F}\times n_F$, $f_{EetF}=f_{E|F}\times f_F=f_{F|E}\times f_E$.
Interprétation probabiliste d'une fréquence (individu pris au hasard), usage "Il est deux fois plus probable qu'un étudiant de Licence soit une fille plutôt qu'un garçon"
Illustration et exemples de calculs (E="étudier les Sc.éco.", F="être inscrit en Licence", proportion de filles) avec le tableau des effectifs étudiants.
Pas de document projeté.

En attente : indépendance, liaison, évt déterminé par un autre, évt certain, impossible, traduction en terme de fréquence, partition de l'évt certain, calcul par conditionnnement.

4. (7 fev) - Rappels : Evènement (evt), fréquence, fréquence conditionnelle. Relation $n_{EetF}=f_{E|F}\times n_F$; application au tableau étudiant : calcul de l'effectif des filles en Licence, Master, Doctorat, tout cursus confondu en tenant compte de l'erreur d'arrondi dans les fréquences conditionnelles indiquées sur le tableau étudiant. Calcul de la proportion de fille à l'Université.
- Calcul par conditionnement : $E_1, E_2,..., E_n$ partition de l'evènement certain si chaque individu réalise un et un seul des évènements E_i. On a alors $N=n_{E1} + n_{E2} + ... + n_{En}$; $1=f_{E1} + f_{E2} + ... + f_{En}$.
Si $E_1, E_2,..., E_n$ est une partition de l'evt certain et si F est un evt alors
$n_F=n_{FetE1} + n_{FetE2} + ... + n_{FetEn}=f_{F|E1}\times n_{E1} + ... + f_{F|En}\times n_{En}$
$f_F=f_{FetE1} + f_{FetE2} + ... + f_{FetEn}=f_{F|E1}\times f_{E1} + ... + f_{F|En}\times f_{En}$
Cas particulier : E evt, on prend n=2, E_1=E et E_2=non(E) : E_1, E_2 est bien une partition de l'evt certain.
- Relations entre deux évènements : E détermine F si tout indivisu réalisant E réalise aussi F. Notation E⇒F. E détermine F si et seulement si $f_{F|E}=1$.
F est indépendant de E si - informel : la réalisation de F n'apporte pas d'information sur la réalisation de E
- formel : si $f_{F|E}=f_F$.
Affaiblissement : E détermine pratiquement F si $f_{F|E}\approx 1$ (concrètement pour ce cours : entre 0.9 et 1) ; F est pratiquement indépendant de E si $f_{F|E}/f_F\approx 1$ (concrètement pour ce cours : entre 0.9 et 1.1). Exemple avec le tableau étudiant : pour un étudiant de l'Université l'evt "être une fille" est pratiquement indépendant de l'evt "être inscrit en Licence" mais n'est pas indépendant de l'evt "être inscrit en Licence de Sc.éco."

Document (projeté pendant le cours 5).
En attente : formule de Bayes, version probabiliste

5. (14 fev) Calculs de fréquences conditionnelles, relation entre fréquences conditionnelles (formule de Bayes), observation sur le tableau des fréquences conditionnelles ou sur le diagramme en barre des déterminations / indépendances, comparaison de deux fréquences conditionnelles ou d'une fréq. cond. avec une fréq. marginale dans la presse ("fréquence/probabilité/risque augmentée de 50%") avec ce document (tableau des effectifs étudiants, longévité de 40 piles).
Exercice 1 de l'interrogation de mars 2010.

6. (21 fev) Correspondance Evènement - sous-population - caractère (à valeurs dans {Oui, Non}).
Rappel (cours 4 et 5) : Relations entre évènements vus par les fréquences : evt certain ($f_E=1$), impossible ($f_E=0$), E détermine F ($f_{F|E}=1$), E rend F impossible, F est indépendant de E ($f_{F|E}=f_F$) ; affaiblissement : E détermine pratiquement F, F est pratiquement indépendant de E ($f_{F|E}/f_F\approx 1$); E augmente la probabilité de F de 20% (par exemple), population de référence : population entière ($f_{F|E}=f_F\times 1.2$) ou bien sous-population des individus ne réalisant pas E ($f_{F|E}=f_{F|non(E)}\times 1.2$), exemple : risque de maladie augmenté par le tabagisme (Cf Wikipedia pour quelques données statistiques), exemple de calcul avec des données fictives
Indépendance pour la population considérée entre un caractère qualitatif et un évènement, entre deux caractères qualitatifs. Exemple avec le tableau des effectifs étudiants. Affaiblissement : evt pratiquement indépendants, ignorer les evts rares. Cas des caractères quantitatifs : découpage de l'étendue en intervalles,

difficulté : l'observation de l'indépendance dépend des intervalles choisis.

En attente : l'observation d'une indépendance est elle pertinente : exemple avec la longévité des 40 piles. Résumé conditionnel d'une variable.

7. (6 mars) Liaison/indépendance via les résumés conditionnels :
Rappel : résumé d'un caractère : pour un caractère quantitatif : (moyenne, écart-type (dispertion)) ou (médiane, intervalle inter-quartile) ou variantes ; pour un caractère qualitatif : (valeurs modales, nbre de valeurs couvrant 60% de la population) par exemple.
Résumé d'un caractère X conditionné à un évt E = résumé de X pour la sous-population déterminée par E. Notation : Moy(X|E), σ(X|E), etc.
X caractère, Y caractère qualitatif prenant les valeurs $y_1,...,y_n$, on associe la suite des résumés Res(X|Y=y_i) (le résumé de X conditionné à l'evt "Y= y_i") ; si X est indépendant de Y alors le résumé conditionnel Res(X|Y=y_i) ne dépend pas de i et vaut Res(X). Ceci donne un critère pour réfuter l'affirmation "X est indépendant de Y" autrement dit pour montrer une liaison entre X et Y dans la situation délicate où X est quantitatif.
Document projeté : le résumé conditionnel dans quelques publications de l'INSEE
Rappel : X,Y deux caractères qualitatifs prenant les valeurs $x_1,...,x_m$ et $y_1,...,y_n$ sont indépendants si pour chaque i,j on a $f_{ij}=f_i$ ou encore si $f_{i,j}=f_i \times f_j$, où f_{ij} désigne $f_{X=i|Y=j}$, etc.
Mesure d'une liaison entre deux variables qualitatives par le nombre $\chi^2(X,Y)=N\sum_{i,j}(f_{i,j}-f_i \times f_j)^2/(f_i \times f_j)$. $\chi^2(X,Y)$ est compris entre 0 et N×min(m-1,n-1). Il vaut 0 si et seulement si X et Y sont indépendantes ; il vaut N×min(m-1,n-1) si X est une fonction de Y ou si Y est une fonction de X. Exemple numérique de calcul.

8. (13 mars) Rappel : Résumé conditionnel ; X,Y caractères qualitatifs -> $\chi^2(X,Y)$ mesure l'indépendance de X avec Y. Les caractères X et Y sont indépendants (respectivement pratiquement indépendants) si et seulement si $\chi^2(X,Y)=0$ (resp. ≈0).
Mesure de la liaison entre X quantitatif et Y qualitatif : D'abord relation entre les moyennes marginale et conditionnelles de X : $Moy(X)=\sum_{i=1..n} f_{Y=yi}\, Moy(X|Y=y_i)$. Relation entre les variances : Rappel $\sigma^2(X)=Moy((X-Moy(X))^2)=Moy(X^2)-Moy(X)^2$
Relation : $\sigma^2(X)=\sum_i f_{Y=yi}\, \sigma^2(X|Y=y_i) + \sum_i f_{Y=yi}\, (Moy(X|Y=y_i)-Moy(X))^2$. La première somme s'appelle la variance intra-groupe ; la seconde somme s'appelle la variance inter-groupe. Le quotient (variance inter-groupe de X)/$\sigma^2(X)$ s'appelle le coefficient de corrélation de X avec Y, noté $\eta^2(X|Y)$. On a $\eta^2(X|Y) \in [0,1]$. Si X est indépendant de Y alors $\eta^2(X|Y)=0$ mais la réciproque est fausse. On a $\eta^2(X|Y)=1$ si et seulement si X est déterminé par Y (X est une fonction de Y). Exemple de calcul.
Mesure de la liaison entre X et Y tous deux quantitatifs : Rq : On peut se ramener à la situation où X et Y sont qualitatifs (ou bien X quantitatif et Y qualitatif) en divisant l'étendue de X et de Y en intervalles, MAIS ce qu'on obtient dépend du choix des intervalles et l'analyse de ce choix est un sujet difficile. On peut à la place calculer :
Covariance de X et Y $Cov(X,Y)=Moy((X-Moy(X))(Y-Moy(Y)))=Moy(XY)-Moy(X)Moy(Y)$. Coefficient de corrélation linéaire $r(X,Y)=Cov(X,Y)/\sigma(X)\sigma(Y)$. On a $r(X,Y) \in [-1,1]$. Si X et Y sont indépendantes alors r(X,Y)=0, réciproque fausse. r(X,Y)=±1 ssi on a une relation entre X et Y de la forme Y=aX+b avec a et b constante et alors $(Y-Moy(Y))/\sigma(Y)=r(X,Y)(X-Moy(X))/\sigma(X)$.
Document projeté : nuage de points et coefficient de corrélation linéaire.

9. (20 mars) Droite de régression.
Rappel : produit de deux caractères qantitatifs X,Y ; variance, écart-type, covariance, coefficient de corrélation linéaire. L'écart type σ(X) d'un caractère X est nul ssi X est constant égal à sa moyenne Moy(X). Commentaire (erroné ! voir cours 10) sur l'interprétation de σ(X) en terme de quantile.
Régression linéaire de Y selon X : trouver deux nombres a,b tels que le caractère R:=Y-(aX+b) soit de moyenne nulle et de variance minimale. Solution (cf méthode des moindres carrés) $a=cov(X,Y)/\sigma^2(X)=r(X,Y)\times\sigma(Y)/\sigma(X)$, $b=Moy(Y)-a\times Moy(X)$. On a $\sigma(R)=\sigma(Y)\times\sqrt{(1-r^2)}$ avec $\sqrt{(1-r^2)} \in [0,1]$; on retrouve qu'on a une liaison affine Y=aX+b ssi r=±1. On écrit Y=aX+b+R (R=Reste ou Résidu) ; aX+b est la meilleure approximation de Y par une fonction affine de X. Si r=0 (⇔ a=0 ⇔ σ(R)=σ(Y)) on a juste approximer Y par sa moyenne Moy(Y) indépendemment de X. Si $\sqrt{(1-r^2)}=½$ (⇔ r=0.866..) par exemple, la dispersion du reste R est moitié de celle de Y, etc.
Qualité de la régression linéaire : deux critères : (1) R est petit devant Y ($\sqrt{(1-r^2)}$ est proche de 0) ; (2) R est indépendant de X (voir cours 10).
Applications : le signe du coefficient a donne la tendance de croissance de Y selon X (Y croît avec X si a>0 ; Y décroît quand X croît si a<0). Prédiction (ou estimation) de Y connaissant X : exemple héréditée de la taille des petits pois (Galton)

Document projeté : taille du fils selon la taille du père avec script R (Cf Galton - Pearson), exercice de statistique de l'examen de 2011 1ère session et corrigé.
Lectures : régression linéaire sur Wikipedia (en anglais).

10. (27 mars) Aspect graphique de la régression linéaire : rappel formules pour les coefficients a,b de la régression linéaire Y=aX+b+Reste, formule pour σ(Reste) ; droite de régression d'équation y=ax+b, tracé de la droite avec deux points, inversement estimation de a et b à partir de la droite (pente, ordonnée à l'origine) ; le point de coordonnées (Moy(X), Moy(Y)) est le centre du nuage de points ("isobarycentre") et est sur la droite.
Autres types de régression : non linéaire ex. Y=a×Log(X)+b+Reste qui peut être de meilleure qualité que la régression linéaire (ie Reste plus petit ou plus indépendant de X, cf ce document) ; régression linéaire multiple ex. $Y=aX_1+bX_2+c+Reste$, Y est la variable à expliquer, X_1 et X_2 sont les variables explicatives.

Echantillons et intervalle de confiance ***[Hors programme des TD et de l'examen en 2011-12]***: Objectif : grande population T, E évènement ou X caractère quantitatif, sous-population S ou bien population S d'individus pris parmis T avec répétition ; on mesure la fréquence de E dans S $f_{E|S}$ Moy(X|S) ou la moyenne de X dans S ; que peut on dire de f_E ou de Moy(X) ? Réponse : On forme la population test S en tirant au hasard sans biais et avec répétition un nombre n d'individus de T alors :
1) (Loi des grands nombres) $f_{E|S}$, respectivement Moy(X|S), devient aussi proche que l'on veut de f_E, resp. Moy(X), à mesure que n croît. Ceci corrige une affirmation du cours 9 concernant le lien entre σ(X) et les quantiles.
2) dispersion des valeurs prises par $f_{E|S}$ quand on renouvelle S : posons $\sigma=\sqrt{(f_E\times(1-f_E))}/\sqrt{n}$; la probabilité que $f_{E|S}$ soit dans l'intervalle $[f_E-\sigma, f_E+\sigma]$ vaut environ 68% pour n pas trop petit ; elle vaut environ 95% pour l'intervalle $[f_E-2\sigma, f_E+2\sigma]$ toujours pour n pas trop petit. On voit donc que la dispersion de $f_{E|S}$ quant on renouvelle S tend vers 0 en $1/\sqrt{n}$ quand n croît.
3) dispertion des valeurs prises par Moy(X|S) quand on renouvelle S : formule analogue avec $\sigma=\sigma(X)/\sqrt{n}$
Applications : sondage d'opinion, recensement de la population.

Lectures : Sondage d'opinion sur Wikipedia, Recensement de la population sur Wikipedia, Notes de cours 2005-06 de F. Diener sur l'estimation d'une fréquence, la feuille de TD correspondante et son corrigé Scilab.

Documents de cours :

Feuille de TD 1 (6 fev. 12). Corrigé de la question 3b. Corrigé de l'ex.5 (question 1 dans le corrigé)
Feuille de TD 2 (20 fev. 12). Corrigé des exercices 1 et 2, corrigé de l'exercice 3 (question 2 dans le corrigé, les commentaires se rapportent au sujet B de l'examen de 2009-10)
Feuille de TD 3 (26 mars 12). Corrigé de l'exercice 5 (ex.4 dans le corrigé)

Interrogation de statistiques (mars 12).

Corrigé des exercices 1-2 de la feuille 3 et des questions c-d-e de l'interrogation (18 avr 12).

Examen 1ère session (24avr12) ; corrigé de la partie statistique.
Examen 2ème session (juin12)

La page du cours en 2010-11

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, 17 janvier 2012