

BE de Statistique Étude d'une base de données IENAC

Avril 2015

1 Description des données

On a recueilli les données suivantes sur un échantillon de 55 élèves ingénieurs ENAC première année.

Nom des variables	Description
sexe	sexe : "F" (femme), "M" (homme).
bac	profil du baccalauréat S : "SI" (sciences de l'ingénieur), "M" (mathématiques), "A" (agronomie), "PC" (physique-chimie), "SVT" (sciences de la vie et de la terre).
mention	mention au baccalauréat : "P" (passable), "AB" (assez bien), "B" (bien), "TB" (très bien).
filiere	filière des concours : "CI" (concours interne), "CPP" (cycle préparatoire polytechnique), "PSI" (physiques sciences de l'ingénieur), "PSI*" (niveau renforcé), "PC" (physique-chimie), "PC*" (niveau renforcé), "MP" (maths-physique), "MP*" (niveau renforcé).
ecrit	moyenne à l'écrit du concours.
oral	moyenne à l'oral du concours.
moyenne	moyenne générale au concours.
rang	classement au concours.
voeux	voeux au concours.
concours	type de concours : "Civil", "Fonctionnaire".
analyse	note à l'examen d'analyse à l'Enac.
proba	note à l'examen de probabilités.
succes	succès au cours de probabilités : "succès", "échec".

L'environnement de travail ".RData" contenant votre jeu de données (différent pour chaque groupe) est téléchargeable sur le site E-Campus de l'Enac à l'adresse <https://e-campus.enac.fr/moodle/>, cours MAP3008 - Statistique, rubrique BE de Statistique. Vous trouverez aussi sur le site E-Campus plusieurs documents d'aide à l'analyse des données, à l'utilisation du logiciel R et à l'interface R Commander.

2 Lecture du jeu de données

1. Créer un répertoire de travail spécifique à ce BE dans lequel vous téléchargerez votre environnement de travail ".RData". Changer le répertoire courant à chaque session de travail en utilisant le menu **Fichier, Changer le répertoire courant** puis charger l'environnement de travail en utilisant l'instruction `load` ou le menu **R console Fichier, Charger l'environnement de travail**.
2. Visualiser les données et déterminer la nature des variables (quantitative discrète/continue, qualitative nominale/ordinaire). La table de données doit contenir 55 lignes et 13 colonnes. Certaines observations sont manquantes et prennent la valeur NA.
3. Attacher la table de données en utilisant l'instruction `attach` et penser à sauvegarder régulièrement votre environnement de travail et votre script, particulièrement en fin de session.

3 Etude descriptive

Différents traitements descriptifs unidimensionnels et bidimensionnels peuvent être entrepris sur ces données. Ils sont traités dans les points suivants.

3.1 Etude unidimensionnelle

1. Pour chaque variable qualitative,
 - (a) obtenir la table des fréquences,
 - (b) réaliser des diagrammes en barres ainsi que des diagrammes circulaires (camemberts).
 - (c) Commenter vos résultats.
2. Pour chaque variable quantitative,
 - (a) obtenir des résumés numériques (indicateurs statistiques comme la moyenne, la médiane, l'écart-type, les quartiles etc...),
 - (b) réaliser des histogrammes ou des diagrammes en bâtons selon la nature des variables, ainsi que des diagrammes de dispersion (boîtes à moustaches) et des diagrammes quantile-quantile (QQ-Plot).
 - (c) Commenter vos résultats.

Synthétiser ce que vous avez appris de votre étude unidimensionnelles.

3.2 Etude bidimensionnelle

3.2.1 Croisement entre une variable quantitative et une variable qualitative

1. Pour les variables **analyse** et **proba** croisées avec les variables qualitatives,
 - (a) obtenir les résumés numériques (indicateurs statistiques),
 - (b) réaliser des histogrammes et des diagrammes de dispersion,
 - (c) commenter vos résultats.
2. Faire de même avec la variable **analyse** croisée avec la variable **succes**. Commenter.
3. Faire de même avec la variable **rang** croisée avec la variable **filière**. Commenter

3.2.2 Croisement entre variables quantitatives

1. Obtenir la matrice de corrélation.
2. Réaliser des nuages des points. Faire de même en croisant avec des variables qualitatives.
3. Commenter vos résultats. En reprenant les résultats précédents, que pensez-vous de la variable **rang** ? Est-elle exploitable dans l'analyse ?

3.2.3 Croisement entre variables qualitatives

1. Croiser la variable **succes** avec les autres variables qualitatives et
 - (a) construire les tables de contingence,
 - (b) construire les profils colonnes,
 - (c) réaliser des diagrammes en barres.
2. Commenter vos résultats.

Synthétiser ce que vous avez appris de votre étude descriptive bidimensionnelle. Qu'avez-vous appris de plus par rapport à l'étude unidimensionnelle ?

4 Etude inférentielle

4.1 Tests d'hypothèses pour un échantillon

1. Effectuer un test d'adéquation à une loi normale pour les variables **analyse** et **proba**. Commenter.
2. Peut-on remettre en cause l'hypothèse nulle que la moyenne des notes en probabilités est supérieure à 12/20 ? Quel test d'hypothèses peut-on choisir pour répondre à cette question ?
 - (a) Rappeler les hypothèses maintenues du test et les hypothèses de tests choisies.
 - (b) En s'aidant du cours et des sorties R, donner la valeur de la statistique observée et exprimer la région critique pour ce test. Utiliser la région critique pour prendre une décision pour un seuil $\alpha = 0,05$.
 - (c) En s'aidant des sorties R, donner la p-valeur. Sa valeur confirme-t-elle votre décision ?
 - (d) Commenter.

4.2 Tests d'hypothèses pour deux échantillons

1. Quels tests d'hypothèses choisir pour répondre aux questions suivantes :
 - (a) Peut-on affirmer que les femmes réussissent mieux que les hommes en probabilités ?
 - (b) Peut-on dire qu'il existe une différence entre les moyennes en analyse et en probabilités ?
 - (c) Peut-on dire que la variable **succes** dépend des autres variables qualitatives ?
2. Mettre en oeuvre ces tests d'hypothèses avec le logiciel R. Pour chaque test effectué
 - (a) rappeler les hypothèses maintenues du test et les hypothèses de tests choisies,
 - (b) en s'aidant du cours et des sorties R, donner la valeur de la statistique observée et exprimer la région critique du test. Utiliser la région critique pour prendre une décision pour un seuil α que vous aurez choisi.
 - (c) En s'aidant des sorties R, donner la p-valeur. Sa valeur confirme-t-elle votre décision ?
 - (d) Commenter vos résultats.

Synthétiser vos résultats et ce que vous avez appris de cette étude.

Remarques importantes :

- Joindre votre script R dans les Annexes du rapport.
- Joindre résultats numériques, graphiques et commentaires pertinents dans le corps du rapport. Tout autre résultat ou graphique non commenté peut être mis dans les Annexes du rapport pour consultation.
- Le présentation et la rédaction du rapport feront partie de la notation.
- Cette étude n'est pas exhaustive et toute analyse complémentaire aux questions ci-dessus et apportant des éléments d'analyse intéressants sera comptée en bonus.