

## Probabilité et statistique I – partim 1: Statistique descriptive

BAC Info – Mardi 20 juin 2017 – Examen pratique – Groupe a

Nom: ..... Prénom: ..... Matricule: .....

### Indications

- Il est interdit de communiquer avec quiconque via Internet sous peine d'annulation de l'examen. Les GSM doivent être éteints.
- L'examen comporte plusieurs questions spécifiques pour lesquelles des réponses précises doivent être écrites sur le questionnaire. Lorsqu'un graphique doit être représenté (voir le symbole □), celui-ci doit être montré à un des surveillants. Le graphique tel que montré est celui sur lequel la cote sera basée.
- L'examen doit être réalisé à l'aide du logiciel SAS.
- L'ensemble de données à considérer lors de cet examen (en ligne sur eCampus; dossier Examen) s'intitule **film1.txt** et contient les informations suivantes concernant 149 films produits avant 2000:

**Title:** titre du film

**IMDbRating:** cote (entre 1 et 10) calculée comme étant la moyenne des cotes attribuées au film par les internautes via le site <http://www.imdb.com/>

**Runtime:** variable quantitative continue précisant la durée du film, en minutes

**Year:** variable quantitative discrète précisant l'année de production du film

**Types:** variable qualitative spécifiant le type du film (comédie, drame,...)

**NumberVotes:** variable quantitative donnant le nombre de votes disponibles pour le calcul de la cote finale.

1. Représenter la distribution des effectifs de la variable **Types** à l'aide d'un diagramme adéquat. □
2. Déterminer le mode de la variable **Types**: .....
3. Décomposer la variable **Runtime** en les trois classes  $[20; 90]$ ,  $]90; 120]$  et  $]120; 230]$  et, en se focalisant sur le couple de variables (**Types**, **Runtime**), déterminer l'effectif bivarié de (**drama**,  $]120; 230]$ ) ainsi que la fréquence des films durant au moins deux heures, conditionnellement au fait que le film soit un drame.  
Effectif conjoint:..... et fréquence conditionnelle:.....
4. Représenter la variable **Runtime** par un histogramme construit en décomposant les durées en 7 intervalles de même longueur. □
5. Représenter un diagramme en tiges et feuilles de la variable **Runtime**. □
6. On se demande si la durée d'un film varie en fonction de son type. Calculer la médiane et l'étendue de la durée conditionnellement au fait que le film est une comédie (modalité **comedy**) et comparer aux mêmes paramètres mesurés sur les films d'horreur (modalité **horror**).

Distr. Cond	Médiane	Etendue
comedy		
horror		

7. On se demande maintenant si la cote attribuée par les internautes est liée à la durée du film. On s'intéresse donc aux deux variables **IMDbRating** et **Runtime**:
- (a) Déterminer, par moindres carrés, l'ordonnée à l'origine de la droite de régression expliquant la variable **IMDbRating** à l'aide de la variable **Runtime**:  $\hat{b} = \dots\dots\dots$
  - (b) Quelle part de la variance de la variable **IMDbRating** est expliquée par la régression?   
.....
  - (c) Quel film correspond au résidu le plus grand en valeur absolue?.....
  - (d) Représenter le diagramme de dispersion de la variable **IMDbRating** en fonction de la variable **Runtime**. Ajouter sur ce graphique la droite de régression ajustée. □

## Probabilité et statistique I – partim 1: Statistique descriptive

BAC Info – Mardi 20 juin 2017 – Examen pratique – Groupe b

Nom: ..... Prénom: ..... Matricule: .....

### Indications

- Il est interdit de communiquer avec quiconque via Internet sous peine d'annulation de l'examen. Les GSM doivent être éteints.  
- L'examen comporte plusieurs questions spécifiques pour lesquelles des réponses précises doivent être écrites sur le questionnaire. Lorsqu'un graphique doit être représenté (voir le symbole □), celui-ci doit être montré à un des surveillants. Le graphique tel que montré est celui sur lequel la cote sera basée.

- L'examen doit être réalisé à l'aide du logiciel SAS.

- L'ensemble de données à considérer lors de cet examen (en ligne sur eCampus; dossier Examen) s'intitule **film1.txt** et contient les informations suivantes concernant 149 films produits avant 2000:

**Title:** titre du film

**IMDbRating:** cote (entre 1 et 10) calculée comme étant la moyenne des cotes attribuées au film par les internautes via le site <http://www.imdb.com/>

**Runtime:** variable quantitative continue précisant la durée du film, en minutes

**Year:** variable quantitative discrète précisant l'année de production du film

**Types:** variable qualitative spécifiant le type du film (comédie, drame,...)

**NumberVotes:** variable quantitative donnant le nombre de votes disponibles pour le calcul de la cote finale.

1. Représenter la distribution des effectifs de la variable **Year** à l'aide d'un diagramme adéquat.  
□
2. Déterminer l'effectif du mode de la variable **Year**: .....
3. Décomposer la variable **NumberVotes** en les trois classes  $[0; 100000]$ ,  $]100000; 200000]$  et  $]200000; 900000]$  et, en se focalisant sur le couple de variables (**Year**, **NumberVotes**) déterminer l'effectif bivarié de  $(1998, ]100000; 200000])$  ainsi que la fréquence des films ayant récolté un nombre de votes supérieur à 100000 mais inférieur à 200000, sachant que le film a été produit en 1998.  
Effectif conjoint:..... et fréquence conditionnelle:.....
4. Représenter la variable **NumberVotes** par un histogramme construit en décomposant les nombres de votants en 10 intervalles de même longueur. □
5. Représenter la variable **NumberVotes** à l'aide d'une boîte à moustaches. □
6. On se demande si le nombre de votants ne dépend pas de l'année de production. Calculer la médiane et l'étendue du nombre de votants conditionnellement au fait que le film a été produit en 1994 ou en 1998.

Distr. Cond	Médiane	Etendue
1994		
1998		

7. On se demande si la cote attribuée par les internautes est liée au nombre de votants. On s'intéresse donc aux deux variables **IMDbRating** et **NumberVotes**:
- (a) Déterminer, par moindres carrés, l'ordonnée à l'origine de la droite de régression expliquant la variable **IMDbRating** à l'aide de la variable **NumberVotes**:  $\hat{b} = \dots\dots\dots$
  - (b) Quelle part de la variance de la variable **IMDbRating** est expliquée par la régression?   
.....
  - (c) Quel film correspond au résidu le plus grand en valeur absolue?.....
  - (d) Représenter les résidus de la régression à l'aide d'un graphe indexé. □