Probabilité et statistique I (partim statistique descriptive)

Bachelier en sciences informatiques

Mardi 20 juin 2017 – Partie 2: Exercices – 10h15-12h

NOM: ..... PRENOM: .....

Indications

- L'examen dure 1h45.
- Une machine à calculer peut être utilisée pour résoudre les exercices.
- Le symbole $\triangleleft$ signifie qu'il est possible de demander au surveillant la réponse à la question concernée afin de pouvoir continuer l'exercice même si ce point n'a pas été résolu.
- Les résolutions des exercices doivent être **expliquées et justifiées**. Lorsque des propriétés théoriques sont utilisées comme justification, il n'est pas nécessaire d'en donner la démonstration.
- Le tableau ci-dessous précise la répartition des points entre les différentes questions. Il n'est pas obligatoire de répondre aux questions dans l'ordre. Cependant, pour faciliter la correction et éviter les erreurs, vous êtes priés, à la fin de l'examen, de préciser pour chaque question si vous l'avez résolue (même partiellement) ou non en entourant soit OUI soit NON dans le tableau ci-dessous:

Q1a	Q1b	Q1c	Q1d	Q2a	Q2b	Q2c
OUI	OUI	OUI	OUI	OUI	OUI	OUI
NON	NON	NON	NON	NON	NON	NON
/5	/8	/5	/7	/3	/8	/4

TOTAL
/40

Exercices: Le site internet `http://www.imdb.com/` fournit de nombreuses données cinématographiques. Pour un grand nombre de films, en plus des données "factuelles" classiques (nom du producteur, durée, genre...), le site propose une cote (entre 1 et 10), cette cote étant mesurée à partir des cotes proposées par les internautes. Les questions ci-dessous portent sur une base de données extraite de ce site et comportant 218 films produits entre 2007 et 2013. Les variables d'intérêt sont la durée (variable `durée` mesurée en minutes) et la cote attribuée par imdb (variable continue à valeurs entre 1 et 10, intitulée `rating`).

1. La table 1 reprend la distribution des effectifs de la variable `durée` après avoir groupé les observations en six classes d'amplitude variable. Malheureusement, certaines informations ont été perdues lors de l'encodage de cette distribution, à savoir la borne supérieure e_4 et l'effectif n_4 de la 4ème classe.

   (a) Sachant que la médiane de la durée estimée par interpolation linéaire à partir de cette série groupée vaut 112.7 minutes, déterminer les valeurs manquantes e_4 et n_4. $\triangleleft$

   (b) Représenter la distribution des effectifs à l'aide d'un histogramme adéquat (en expliquant la construction) et ajouter sur l'histogramme le polygone correspondant et résumer les caractéristiques principales de celui-ci en une phrase.

Classes	Effectifs
$[20, 90]$	15
$]90, 100]$	40
$]100, 110]$	44
$]110, e_4]$	n_4
$]e_4, 150]$	64
$]150, 200]$	15

Table 1: Tableau statistique de la distribution du nombre de minutes des 218 films

(c) La moyenne des durées est égale à 115.7 minutes et on précise également que

$$\sum_{i=1}^{218} x_i^2 = 3030225$$

où x_i est la durée observée sur le $i$ème film de la base de données. Calculer les bornes $\bar{x} - 2s$ et $\bar{x} + 2s$ (où s est l'écart-type de la série des durées) et déterminer, sous l'hypothèse de répartition uniforme dans chaque classe de la distribution, quelle proportion des observations se situe entre ces deux bornes. Comparer cette proportion à celle suggérée par l'inégalité de Tchebychev.

(d) Les producteurs de films aimeraient déterminer si l'appréciation d'un film dépend ou pas de la durée du film. Le diagramme de dispersion basé sur les deux variables `durée` et `rating` est représenté à la Figure 1.

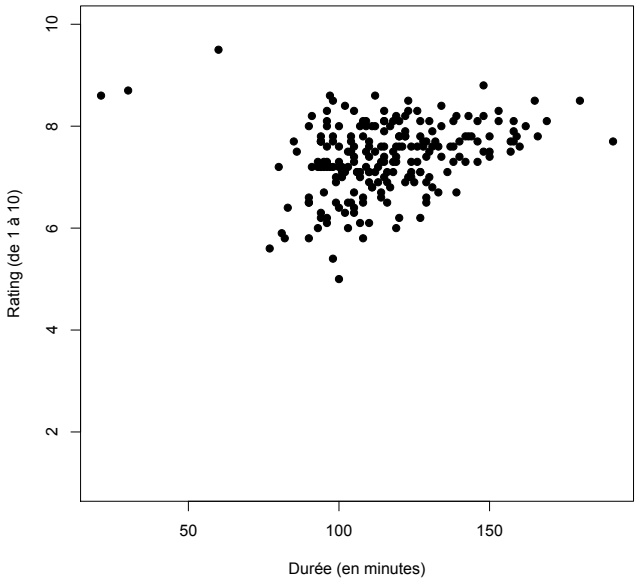

Figure 1: Diagramme de dispersion de la variable `rating` en fonction de la durée du film.

En exploitant, si nécessaire, les informations suivantes (Y désignant la variable en ordonnée, à savoir la variable `rating`) ainsi que celles déjà précisées ou calculées dans les exercices précédents:

$$\bar{y} = 7.4; s_y^2 = 0.5; \sum_{i=1}^{218} x_i y_i = 186974$$

Calculer la covariance entre les deux variables X et Y et expliquer son signe en exploitant le graphique ci-dessus.

Proposer d'autres outils aux producteurs afin de déterminer si l'appréciation d'un film dépend ou pas de la durée du film (sans effectuer de calculs). Commenter.

2. La médiane et la moyenne de la série des durées sont précisées à l'exercice 1 (parties (a) et (c)). Un internaute insiste pour que le film *A cure for insomnia* soit ajouté à la base de données. Ce film dure 5520 minutes et est le plus long film expérimental jamais commercialisé.

   (a) Calculer la moyenne et la médiane de la nouvelle série, comprenant non seulement les 218 observations initiales mais également la durée du film *A cure for insomnia*.

   (b) On souhaite maintenant **modéliser** l'impact de l'ajout d'une observation à une série initiale. En notant $x_1, \ldots, x_n$ les observations de la série initiale (de moyenne $\bar{x}$ et de médiane $\tilde{x}$) et x (avec $x \geq x_i$ pour tout i) la valeur de l'observation additionnelle, calculer les moyenne et médiane (notées respectivement $\bar{x}'$ et $\tilde{x}'$) de la nouvelle série construite sur les observations $x_1, \ldots, x_n, x$ en exprimant les réponses en fonction de x et en fonction des observations $x_1, \ldots, x_n$ et/ou des paramètres $\bar{x}$ et $\tilde{x}$ de départ.

   (c) Représenter $\bar{x}'$ et $\tilde{x}'$ en fonction de x pour $x \in [x_{(n)}, +\infty[$ et commenter.