Probabilité et statistique I (partim statistique descriptive)

Bachelier en sciences informatiques

Vendredi 10 juin 2016 – 10h-12h30

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

Indications

- L'examen dure 2h30.
- Une machine à calculer peut être utilisée pour résoudre les exercices.
- Le symbole $\lhd$ 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:

Théorie		Exercices			QCM
Q1	Q2	Q1	Q2	Q3	
OUI NON	OUI NON	OUI NON	OUI NON	OUI NON	OUI NON
/8	/12	/13	/14	/ 13	/10

TOTAL
/70

Théorie:

1. Soit $S = \{x_1, \ldots, x_n\}$ une série statistique quantitative univariée de moyenne arithmétique $\bar{x}$. Démontrer que la moyenne minimise la somme des carrés des écarts entre les observations de la série et une constante a.

2. Soit $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ une série statistique bivariée obtenue en observant deux variables quantitatives X et Y sur n individus.

   (a) Ecrire l'équation de la droite de régression de Y en X obtenue par la technique des moindres carrés en donnant précisément les formules des coefficients de pente et d'ordonnée à l'origine.

   (b) Calculer explicitement la variance de la série des résidus.

Exercices: Le Département des Transports des Etats-Unis permet à chaque usager de collecter des données sur les retards enregistrés lors des voyages en avion entre deux aéroports situés sur le sol américain (voir le site web `http://www.transtats.bts.gov/`). Les données considérées dans cet exercice concernent 251 vols, partis de Baltimore ou de Boston (variable `Origine`), un lundi ou un samedi dans le courant du mois de janvier 2016. Pour chacun de ces vols, on dispose également de la destination (variable `Destination`) et du nombre de minutes de retard à l'arrivée (variable `Retard` dont les valeurs peuvent être négatives, ce qui correspond dans ce cas à une arrivée de l'avion avant l'heure prévue).

Les questions ci-dessous portent sur cette base de données de 251 lignes et 3 colonnes, dont quelques lignes sont représentées ci-dessous:

Identifiant	`Origine`	`Destination`	`Retard`
1	Boston	LosAngeles	-15
2	Boston	SanDiego	-39
3	Boston	SanDiego	128
4	Boston	LongBeach	1
5	Boston	LosAngeles	15
6	Baltimore	SanFrancisco	-40
7	Baltimore	LosAngeles	-24
8	Boston	SanFrancisco	-37
⋮			

1. La table 1 reprend la distribution des fréquences de la variable `Retard` mesurée sur les vols partis de Baltimore uniquement, les minutes de retard ayant été groupées en quatre classes d'amplitude variable. Malheureusement, certaines informations ont été perdues lors de l'encodage de cette distribution, à savoir la borne supérieure de la 4ème classe, e_3 (qui est aussi, par définition, la borne inférieure de la dernière classe) et la fréquence de la deuxième classe f_2.

Classes	Fréquences	Fréq. cumulées
$[-50, -25]$	0.17	
$]-25, 0]$	f_2	
$]0, e_3]$	0.28	
$]e_3, 160]$	0.02	
TOTAL		x

Table 1: Tableau statistique de la distribution du nombre de minutes de retard des vols partis de Baltimore

   (a) Sachant que la moyenne du nombre de minutes de retard estimée à partir de cette série groupée vaut -3.9 minutes, déterminer les valeurs manquantes e_3 et f_2. $\lhd$

   (b) Compléter le tableau en calculant les fréquences cumulées.

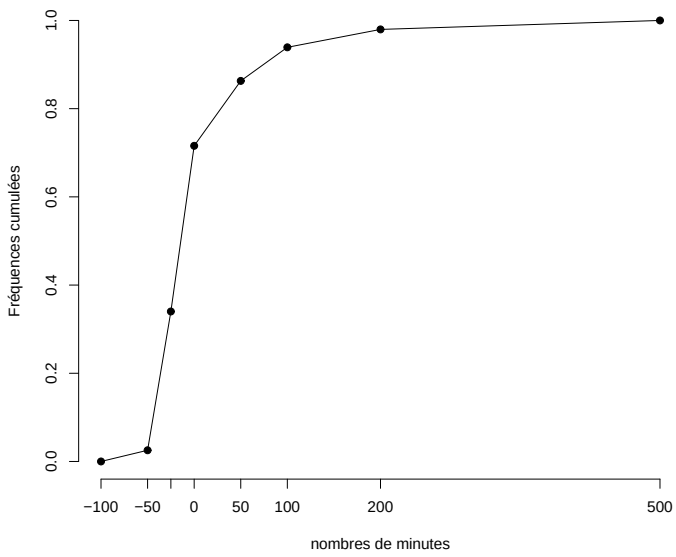

Figure 1: Représentation de l'ogive des fréquences cumulées du nombre de minutes de retard mesuré sur les vols en partance de Boston

(c) Représenter l'ogive des fréquences cumulées sur le graphique de la Figure 1 où l'ogive correspondant aux fréquences cumulées du retard enregistré à partir de l'aéroport de Boston est déjà représentée.

2. Les dispersions des deux séries de retard (les retards enregistrés depuis Baltimore et ceux enregistrés depuis Boston) sont assez différentes. Un résumé statistique est donné dans le tableau ci-dessous

	Moy	s	Min	Q_1	med	Q_3	Max	Effectif
Baltimore	-4.31	33.92	-47	-20.75	-12.5	4.5	160	54
Boston	3.26	65.99	-61	-29	-16	2	464	197

(a) Comparer les séries à l'aide de deux paramètres de dispersion.

(b) Sachant que la variance de la série complète des retards est égale à 3667.45 minutes^2, déterminer, avec justification, la part de la variabilité totale due à la différence de performance entre les deux aéroports.

(c) Le secteur aérien est satisfait car le temps maximal de retard enregistré à Baltimore a diminué de 7% entre janvier 2015 et janvier 2016. Quel était le retard maximal enregistré en janvier 2015?

(d) La boîte à moustaches des retards enregistrés à Boston est représentée à la Figure 2. Ajouter celle basée sur les retards mesurés à Baltimore en exploitant également les parties inférieure et supérieure du diagramme en tiges et feuilles représentées ci-dessous afin de pouvoir dessiner la boîte complète (la construction doit être détaillée):

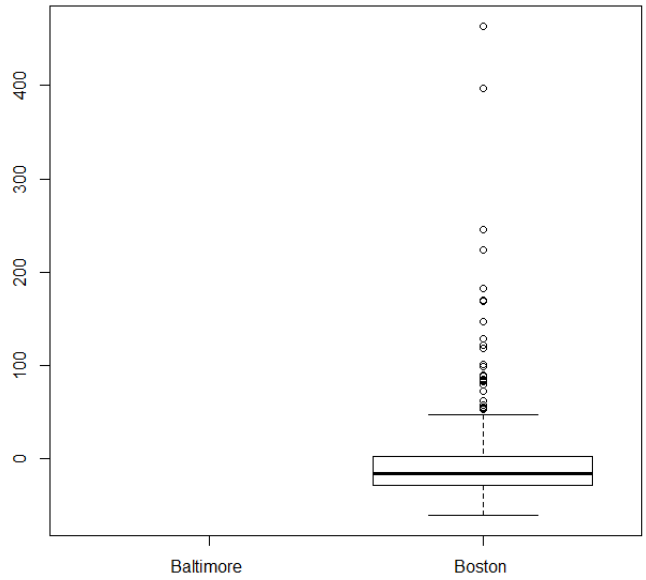

Figure 2: Boîte à moustaches des nombres de minutes de retard mesurés sur les vols en partance de Boston

```
The decimal point is 1 digit(s) to the right of the |
-4 | 7210
-3 | 9866
⋮
 3 | 8
 4 | 1
 6 | 4
 7 | 5
 8 |
 9 |
10 |
11 |
12 |
13 |
14 |
15 |
16 | 0
```

3. Les destinations accessibles via Baltimore et Boston sont les suivantes: Long Beach, Los Angeles, Oakland, San Diego, San Francisco et San Jose. La répartition des vols est décrite par le tableau de contingence ci-dessous:

	Destination					
Origine	LosAngeles	LongBeach	Oakland	SanDiego	SanFrancisco	SanJose
Baltimore	29	0	8	15	2	0
Boston	85	5	0	22	83	2

(a) A partir de ce tableau de contingence, répondre aux questions suivantes:

- Quelle est la fréquence bivariée correspondant aux vols de Baltimore à Oakland?
- Quel est le mode de la distribution des destinations conditionnelle au fait de décoller de Boston?
- Sachant que l'avion a décollé de Boston, quelle est la fréquence d'un atterissage à San Diego?

(b) Dans ce contexte qualitatif, il n'est pas possible de calculer une corrélation entre les deux variables. Il est par contre possible de mesurer un degré d'association entre deux modalités de variables qualitatives à partir d'une table 2×2, telle qu'illustrée ci-dessous et dans laquelle l'effectif total n de la population est décomposé en effectifs bivariés a, b, c et d en fonction du fait que les individus correspondent aux modalités en question ou pas, avec $a + b + c + d = n$:

	Modalité B		
Modalité A	Oui	Non	
Oui	a	b	
Non	c	d	
			n

- Construire la table 2×2 pour le couple de modalités `Baltimore` et `LosAngeles`.
- La mesure d'association la plus fréquente basée sur la table en question est donnée par $(a + d)/n$. Que vaut cette mesure pour le couple de modalités `Baltimore` et `LosAngeles`?
- Que vaut la mesure d'association pour le couple `Baltimore`, `Boston`?

QCM: Pour chaque question, veuillez choisir une et une seule réponse possible pour chaque choix multiple. En cas de réponse correcte, 1 point est acquis; en cas de réponse incorrecte, 0.25 points sont retranchés et si aucune réponse n'est cochée, aucun point n'est gagné ni perdu.

1. Le nombre d'achats effectués sur un site de ventes en ligne a baissé de 20% entre janvier et février. En mars, ce nombre a augmenté de 20% par rapport au mois précédent. En conséquence:

   ☐ En mars, le nombre d'achats vaut 96% du nombre d'achats de janvier.

   ☐ En mars, le nombre d'achats est le même qu'en janvier.

   ☐ Aucune de ces assertions n'est correcte.

2. Une série statistique quantitative est représentée par l'histogramme de la Figure 3.

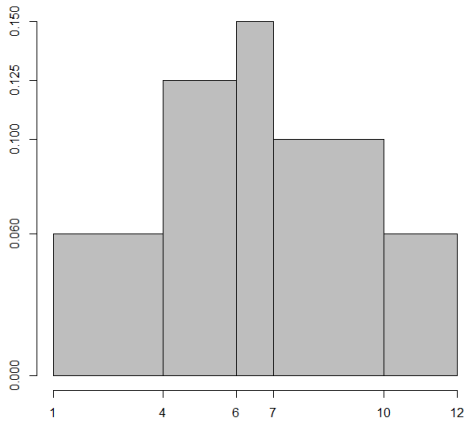

Figure 3: Représentation d'un histogramme

   - La classe $[6; 7[$ est celle de fréquence la plus élevée: ☐ Vrai – ☐ Faux
   - Les deux classes extrêmes ont la même fréquence: ☐ Vrai – ☐ Faux

3. Une entreprise consent à son personnel une augmentation générale des salaires bruts.

   - Si l'augmentation est de 3%, le salaire mensuel brut médian augmente de 3%: ☐ Vrai – ☐ Faux
   - Si les salaires sont élevés au carré, le salaire mensuel brut moyen est lui aussi élevé au carré: ☐ Vrai – ☐ Faux
   - Si l'augmentation consiste en une prime mensuelle de 100 euros pour tous les travailleurs, le salaire mensuel brut moyen augmente de 100 euros: ☐ Vrai – ☐ Faux

4. Soit une série statistique de taille 100 dont la moyenne arithmétique vaut 50 et la variance vaut 25. Quelle affirmation parmi les suivantes est correcte?

   □ Il y a minimum 88 observations comprises entre 35 et 65.

   □ Il y a au plus 33 observations en dehors de l'intervalle $[35; 65]$.

   □ Il y a au moins 75 observations dans l'intervalle [45, 55].

   □ Aucune de ces assertions n'est correcte.

5. La covariance d'une série statistique quantitative bivariée

   - est toujours positive ou nulle: : □ Vrai – □ Faux
   - est insensible aux translations des données: : □ Vrai – □ Faux

6. Dans une application concrète basée sur deux variables X et Y de variances égales à 1 et 0.95 respectivement, on obtient une covariance égale à 0.9 entre X et Y. Le coefficient de détermination est égal à □ 0.85 – □ 0.92 – □ 0.95