Statistique descriptive

Bachelier en sciences informatiques

Mardi 3 septembre 2019 – Partie 2: Exercices et analyse de données – 15h30-18h

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

Indications

- L'examen dure 2h30.
- Les notes de cours, transparents... peuvent être consultées pendant l'examen. Il est également possible d'utiliser une machine à calculer et/ou le logiciel `R`.
- Il est interdit de communiquer avec quiconque via internet sous peine d'annulation de l'examen.
- Les résolutions des exercices doivent être **expliquées et justifiées**.
- Les graphiques réalisés à l'aide du logiciel `R` pour l'analyse de données doivent tous être copiés dans un unique fichier du type word (le nom du fichier doit être construit comme suit: *nomprenomgraphiques*, sans accent ni espace). Le script `R` doit être nommé comme suit: *nomprenomcode.R*, sans accent ni espace. Ces deux fichiers doivent être envoyés par email, à la fin de l'examen, à `M.ernst@uliege.be`. Les démarches suivies et décisions prises pendant l'analyse doivent être décrites sur la feuille de réponse. Le code `R` ne sera pas coté mais sera analysé en cas de réponse étonnante ou suspecte.
- 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:

Exercices					Analyse de Données	
Q1a	Q1b	Q2a	Q2b	Q2c	Q1	Q2
OUI	OUI	OUI	OUI	OUI	OUI	OUI
NON	NON	NON	NON	NON	NON	NON
/6	/6	/6	/6	/6	/15	/15
TOTAL:		/30			TOTAL:	/30

TOTAL
/60

Exercices:

1. Une association de protection des consommateurs s'intéresse aux nombres de calories d'une portion de 100 gr de 74 sortes de chips, ces chips provenant de trois producteurs différents, simplement intitulés `A`, `B` et `C`.

   Le tableau de contingence ci-dessous décrit la distribution de ces nombres de calories par portion conjointement avec la variable qualitative `Producteur`. Les valeurs de la variable quantitative continue `Calories` ont été groupées en trois classes d'amplitude variable.

Producteur	Classes du nombre de calories $[50, 100]$	$]100, 110]$	$]110, 160]$
A	16	5	8
B	5	13	n_{23}
C	7	10	n_{33}

(a) Malheureusement, les effectifs bivariés n_{23} et n_{33} ont été perdus lors de la retranscription du tableau de contingence. Déterminer les valeurs manquantes en exploitant explicitement l'information suivante: la médiane de la variable groupée `Calories` conditionnelle au fait que les chips soient produits par `C` est égale à 104.5 (il convient d'arrondir les valeurs trouvées à l'unité la plus proche).

(b) En exploitant la répartition en classes précisée dans le tableau de contingence, représenter la distribution marginale de la variable `Calories` à l'aide d'un histogramme adéquat et représenter également sur le même graphique un polygone adapté à cet histogramme. Commenter sur le caractère symétrique ou non de la distribution. *NB: il est possible de faire cet exercice même si les valeurs spécifiques des effectifs n_{23} et n_{33} n'ont pas été déterminées.*

2. Les producteurs des chips des marques `A` et `B` essayent, dans leur publicité respective, de mettre en évidence la meilleure qualité de leurs produits par rapport à ceux du concurrent. Notamment, le producteur `A` certifie que ses chips contiennent plus de potassium (en mg par portion) que les chips du producteur `B`. Cependant, la firme `B` assure qu'en fait, il y a tellement de variabilité entre les différentes sortes de chips (quel que soit le producteur) qu'il ne faut pas accorder d'importance à la différence mesurée entre les niveaux moyens de quantité de potassium. Afin de régler cette polémique, on propose de réaliser une *analyse de la variance* en distinguant la variance dans les groupes et la variance entre les groupes à partir d'un échantillon de chips comptant 29 sortes de chips pour le producteur `A` et 22 types de chips pour le producteur `B`. On précise qu'en ce qui concerne les chips du producteur `A`, le contenu moyen en potassium par portion est égal à 102, tandis que l'écart-type du contenu en potassium vaut 92. En ce qui concerne les chips du producteur `B`, on précise que la somme du contenu en potassium (par portion) des chips considérés vaut 1875, tandis que la somme des carrés de ces mêmes contenus est égale à 203625.

(a) Compléter le tableau ci-dessous en détaillant les éventuels calculs complémentaires sur la feuille quadrillée:

	Série cond. `Potassium` marque `A`	Série cond. `Potassium` marque `B`	Série de `Potassium` marques `A` et `B` ensemble
Effectifs			
Moyennes			
Variances			

(b) Calculer la part de la variance totale expliquée par la variabilité entre les groupes et commenter.

(c) Un critère objectif permettant de décider si la différence observée dans les moyennes des deux marques de chips est *significative* consiste à déterminer si

$$(n_A + n_B - 1)\frac{\text{Variance entre les groupes}}{\text{Variance dans les groupes}} > 4.06$$

où n_A (resp. n_B) est l'effectif marginal des chips de la marque `A` (resp. `B`). Est-ce le cas ici? Commenter et préciser quel producteur a raison concernant les arguments avancés dans la publicité.

Analyse de données:

Les données et leur description sont en ligne sur eCampus.

1. On s'interroge sur la différence du contenu en calories des céréales des deux marques principales, à savoir `Kelloggs` et `GeneralMills`. Dans cet exercice, les céréales fournies par un autre producteur (celles de la catégorie `Autre`) ne sont pas considérées.

   (a) Comparer les distributions de la variable `Calorie` lorsque les céréales sont décomposées en les deux groupes définis par les marques `Kelloggs` et `GeneralMills`. Utiliser deux types de graphiques différents qui permettent de comparer les deux groupes en utilisant une représentation dans le même repère. Commenter.

   (b) Confirmer l'analyse graphique en calculant des résumés statistiques pertinents et préciser si les différences portent sur la localisation, la dispersion et/ou sur la symétrie.

   (c) Afin de quantifier la différence observée, un statisticien professionnel effectuerait un "test statistique" de manière à pouvoir conclure, le cas échéant, à une différence *significative* entre les niveaux moyens observés dans les deux groupes (ici, dans les deux marques). Un tel test, le test de comparaison des moyennes de Welch, conclurait à une différence significative entre les deux types de céréales si

   $$\left|\frac{\bar{x}_K - \bar{x}_G}{s}\right| > 2.036 \text{ avec } s = \sqrt{\frac{s_K^2}{n_K} + \frac{s_G^2}{n_G}}$$

   où l'indice K se réfère aux céréales `Kelloggs` et l'indice G aux céréales de la marque `GeneralMills` (la notation n correspondant comme d'habitude à un effectif, $\bar{x}$ à une moyenne et s^2 à une variance). Vérifier si cette inégalité est satisfaite et commenter en exploitant les analyses graphiques et statistiques réalisées aux points 1(a) et 1(b).

2. On aimerait maintenant déterminer si le niveau de potassium dans les céréales (sans distinction du producteur) peut être expliqué par une des autres variables disponibles.

   (a) Calculer la corrélation entre la variable `Potassium` et les autres variables de la base de données et écrire ci-dessous la corrélation la plus faible et la corrélation la plus forte:

   Corrélation la plus faible = ... calculée entre `Potassium` et .....

   Corrélation la plus forte = ... calculée entre `Potassium` et .....

(b) Sachant que la variable `Potassium` est la variable dépendante d'intérêt, sélectionner comme variable explicative la variable correspondant à la corrélation la plus forte calculée en 2(a) et déterminer l'équation de la droite de régression estimée par la technique des moindres carrés (indiquer cette équation sur la feuille de réponse). Par ailleurs, préciser ci-dessous le résidu des céréales `Smacks` et la valeur ajustée par le modèle pour les céréales `Golden_Crisp`.

Résidu des céréales `Smacks`: .....

Valeur ajustée des céréales `Golden_Crisp`: .....

(c) Mesurer la qualité de l'ajustement. Préciser le paramètre utilisé et sa valeur sur la feuille et commenter.

(d) On se demande maintenant si le lien observé entre la variable `Potassium` et la variable explicative sélectionnée au point 2(b) est similaire pour les deux marques principales de céréales. Re-calculer les corrélations entre les deux variables, d'une part en n'utilisant que les céréales `Kelloggs` et d'autre part uniquement avec les céréales `GeneralMills`. Commenter. Pour mieux visualiser les choses, représenter le diagramme de dispersion de la variable `Potassium` en fonction de la variable explicative sélectionnée, en utilisant toutes les observations mais en exploitant des symboles différents pour les trois marques. Ajouter sur le graphique l'équation de la droite de régression estimée en 2(b), ainsi que les deux droites de régression estimées par la technique des moindres carrés séparément sur les cérélaes `Kelloggs` et `GeneralMills`. Commenter.