Statistique descriptive

Bachelier en sciences informatiques

Vendredi 7 juin 2019 – Partie 2: Exercices et analyse de données – 10h-12h30

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	Q3	Q1	Q2	Q3
OUI	OUI	OUI	OUI	OUI	OUI	OUI	OUI
NON	NON	NON	NON	NON	NON	NON	NON
/6	/7	/4	/8	/5	/5	/10	/15
TOTAL:		/30			TOTAL:		/30

TOTAL
/60

Exercices:

1. Les sociologues pensent qu'un couple a tendance à reproduire le même schéma familial que celui qu'il a connu en étant enfant. En se focalisant sur l'épouse, ils considèrent qu'une femme qui était fille unique dans son enfance aura moins d'enfants qu'une femme provenant d'une famille nombreuse. Pour étudier cette question, un échantillon de 82 femmes âgées de plus de 40 ans et ayant au moins un enfant a été constitué. Les deux variables suivantes ont été mesurées:

   - le nombre de frères et sœurs (vivants ou décédés) de la femme, variable X
   - le nombre d'enfants de la femme, variable Y.

Les données sont résumées dans le tableau de contingence suivant (dans lequel deux effectifs, n_{22} et n_{43}, ont malheureusement été perdus lors de l'encodage):

Nombre	Nombre d'enfants Y				
frères/sœurs X	1	2	3	4	5
0	4	9	4	1	0
1	4	n_{22}	11	5	2
2	1	5	9	1	1
3	1	2	n_{43}	1	0
4	0	0	0	2	2

(a) Déterminer n_{22} et n_{43} sachant que le nombre moyen d'enfants nés de femmes ayant au moins 3 frères et sœurs vaut 3.3.

(b) Déterminer les distributions conditionnelles de Y sachant que $X = 0$ et sachant que $X = 2$ (les préciser sous forme d'un tableau statistique ne comprenant que la colonne des effectifs).

i. Quels sont les modes de ces deux distributions?

ii. Comparer les courbes cumulatives des fréquences cumulées de ces deux distributions en les plaçant sur le même graphique.

Que peut-on en conclure concernant l'hypothèse émise par les sociologues?

2. La société *MusicDown* est spécialisée dans le téléchargement de musique via internet. Des données correspondant aux nombres globaux d'heures de connexion enregistrées par cette société ont été recueillies sur les 18 derniers mois d'activité. Cette durée de connexion mensuelle varie sensiblement d'un mois à l'autre. Les responsables de la société tentent de déterminer le facteur explicatif le plus plausible parmi les suivants:

- Facteur 1: Promotions spéciales ou pas, variable qualitative à deux modalités : `Mois avec promotions` ou `Mois sans promotion`
- Facteur 2: Nombre de nouveaux tubes, variable discrète catégorisée en un indicateur qualitatif à deux modalités: `Plus de 10 tubes` ou `Moins de 10 tubes`

Afin de déterminer le facteur le plus important, les responsables de la société proposent de calculer la variabilité du nombre d'heures de connexion entre les groupes caractérisés par les modalités possibles de ces différents facteurs. Pour cela, ils ont construit le Tableau 2 (où les paramètres statistiques mesurés sur la variable "nombre d'heures de téléchargement" tiennent compte du fait que cette variable est exprimée en milliers d'heures):

	Effectifs des groupes	Moyennes des groupes	Variances dans les groupes	Part de la variance ENTRE les groupes
Facteur 1				
Avec promotions	6	20	46.82	x
Sans promotion	12	45.5	44.84	
Facteur 2				
Plus de 10 tubes	9	33	172	0.0842
Moins de 10 tubes	9	41	176	

(a) A partir du facteur 2, calculer la moyenne globale et la variance totale du nombre d'heures de connexion.

(b) Retrouver la valeur de la quantité indiquée simplement par x dans le tableau. Lequel des deux facteurs est, selon cette analyse, le plus déterminant?

3. Il est intéressant de revenir sur les sondages publiés dans la presse maintenant que les résultats des élections régionales de 2019 sont connus. Les figures reprises à la dernière page décrivent les résultats d'un sondage réalisé en avril 2019 (avec des informations sur la méthodologie suivie) ainsi que le résultat des votes au niveau wallon. En 10 lignes maximum, mettre en évidence des commentaires pertinents au niveau statistique, en exploitant notamment la notion d'erreur statistique maximale mentionnée dans la fiche technique.

Analyse de données:

Les données et leur description sont en ligne sur eCampus.

1. La base de données présente-t-elle des valeurs manquantes ou des "erreurs grossières" (c'est-à-dire des valeurs impossibles à observer en l'état pour la variable en question)? Décrire, sur la feuille de réponse, les problèmes détectés. Recopier sur la feuille les lignes complètes des individus concernés. Construire une nouvelle base de données `data2` dans laquelle les lignes correspondantes (celles avec au moins une donnée manquante ou aberrante) sont supprimées.

   Pour ce faire, les commandes suivantes peuvent être utiles:

   ```
   data2 <- data[-c(a,b,...,c),]
   ```

   ```
   attach(datat2)
   ```

   où `a,b,...c,` doivent être remplacés par les numéros des lignes à supprimer.

   C'est sur l'ensemble de données `data2` que les questions suivantes doivent être traitées.

2. On s'interroge sur la différence de satisfaction éventuelle entre les petites et grandes entreprises en ce qui concerne la qualité du produit.

   (a) Comparer les distributions de l'indice de satisfaction `Qualite` lorsque les clients sont décomposés en les deux groupes définis par la caractéristique qualitative `Taille`. 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) Est-il possible, grâce à cette analyse, de déterminer de quel groupe provient l'entreprise dont la valeur sur l'indicateur qualitatif est justement manquante (voir informations copiées sur la feuille de réponse)? Expliquer la démarche suivie sur la feuille de réponse.

   (d) Afin de mesurer le lien entre une variable qualitative à deux modalités et une variable quantitative, on peut également procéder comme suit. Tout d'abord, on transforme la variable qualitative en une indicatrice binaire en privilégiant une des deux modalités. Plus précisément, les individus caractérisés par la modalité choisie se voient attribuer la valeur 1 (pour la présence de la caractéristique) tandis que les autres reçoivent la valeur 0 (pour l'absence de cette même caractéristique). Ensuite, on calcule le coefficient de corrélation classique entre la variable quantitative et l'indicatrice binaire. Ce coefficient s'appelle, dans ce contexte particulier, une *corrélation bisériale*. Calculer la corrélation bisériale entre les deux variables `Qualite` et `Taille`, donner sa valeur sur la feuille quadrillée et interpréter cette valeur en vous inspirant de l'interprétation classique d'un coefficient de corrélation.

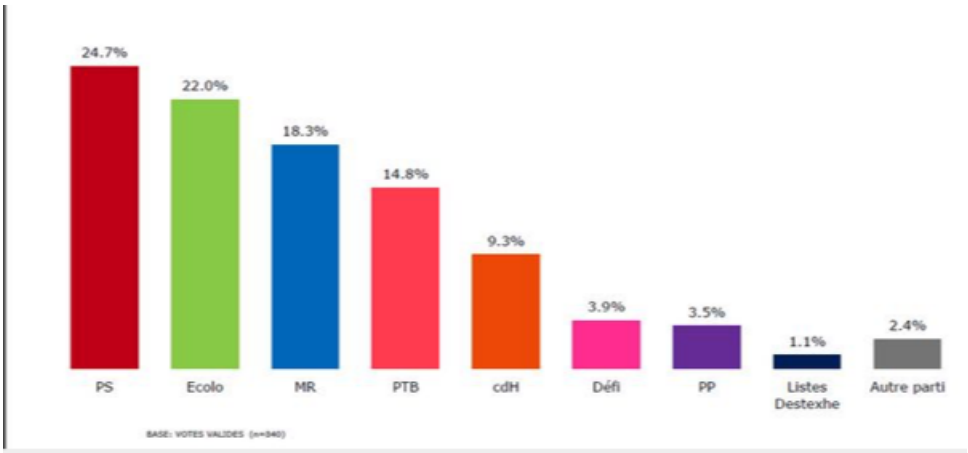
3. On aimerait maintenant déterminer si le niveau de satisfaction encodé pour le prix du produit pourrait expliquer la note de satisfaction attribuée pour le service.

   (a) Représenter le diagramme de dispersion des deux variables en prenant comme variable explicative la variable `Prix` et comme variable dépendante la variable `Service`, y ajouter la droite de régression estimée par la technique des moindres carrés (dont l'équation doit être mentionnée sur la feuille).

   (b) Analyser les résidus en représentant un graphe indexé des résidus. Commenter.

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

   (d) Serait-il possible d'utiliser le modèle ajusté pour déterminer des valeurs adéquates pour la ou les donnée(s) manquante(s) écartée(s) à la question 1 (et copiées sur la feuille de réponse)? Expliquer la démarche et proposer une ou des valeur(s).



Baromètre politique avril 2019: Wallonie - © RTBF

Fiche technique :

Ce sondage d'opinion a été mené par Kantar TNS à la demande de la VRT/Standaard/RTBF/La Libre sur un échantillon aléatoire de n=1006électeurs flamands, 1004 électeurs wallons et 759 électeurs bruxellois et accessibles via un téléphone fixe ou mobile. L'erreur statistique maximale est de 3,1% (Flandre, Wallonie), de 3,6% à Bruxelles, supérieure et inférieure au résultat obtenu pour les énoncés dans l'ensemble de l'échantillon. Les répondants ont été interrogés par téléphone du 25 mars au 15 avril 2019.

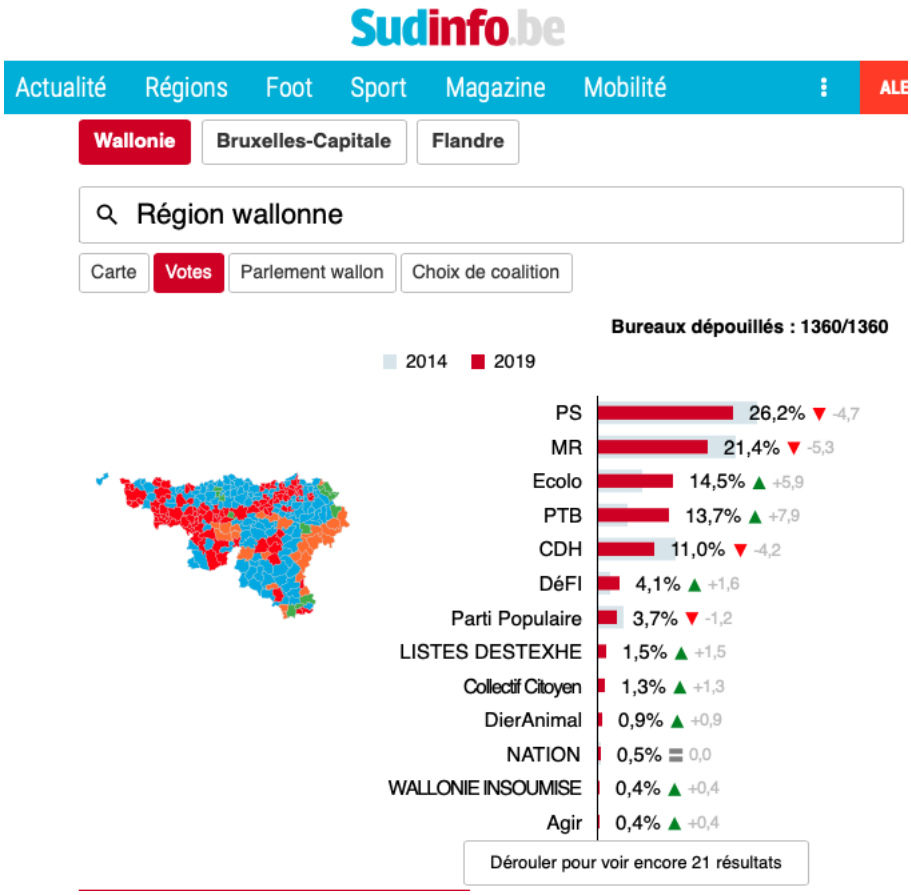


Figure 1: Au dessus: sondage réalisé en avril 2019; en dessous: résultats des élections de mai 2019