EISTI Ecole Internationale des Sciences du Traitement de l'Information	**ING1-MAIN : STATISTIQUE DESCRIPTIVE : EXAMEN S4**
Rédigé par Astrid Jourdan et Hervé de Milleville	Ref : *ING1-MAIN-STA-DES-EXAM-S4*
A l'intention de : Etudiants des ING1-MAIN	Créé le : 27/ 05/2013 et modifié le 28/05/2013

Correction

1. Questions de cours

1 point 1. Dans la boite de Tuckey, parmi les indicateurs extrêmes on utilise les quantiles Q_1 et Q_3. On pourrait les remplacer par la plus petite valeur et la plus grande valeur. Quel est l'inconvénient majeur de ce changement ?Pas d'individus à l'extérieur des valeurs min et max/ individus atypiques/…

1 point 2. Dans l'analyse bivariée quali x quanti, on est amené à comparer les moyennes du caractère quantitatif sur les différentes sous populations définies par les modalités du caractère qualitatif. Pourquoi cette seule comparaison n'est pas suffisante et qu'il faut l'agrémenter de l'analyse de la variance du caractère quantitatif sur la population totale et les sous populations ? dispersion de chaque sous-population

1 point 3. On considère un caractère nominal à deux modalités. Pourquoi en codant les modalités respectivement 0 et 1, on peut considérer légitimement avec ce codage, le caractère comme étant quantitatif. Que signifient la moyenne et l'écart-type dans ce cas ? Pourcentage

1 point 4. On considère un caractère nominal à trois modalités. Pourquoi ne peut-on pas faire un codage équivalent à la question ci-dessus pour rendre le caractère quantitatif ? La moyenne de trois modalités comme par exemple « bleu », « vert », « rouge » codées par 1 ,2, 3, ne veut rien dire.

1 point 5. On dit souvent que les indicateurs moyenne et médiane sont complémentaires. Pourquoi ? La moyenne positionne la série par rapport aux valeurs prises par celle-ci et l'autre positionne la série par rapport aux nombres d'individus/ erreur quadratique-abs

1 point 6. On considère deux caractères quantitatifs X et Y. La régression linéaire simple a une faiblesse : elle ne cherche que la meilleure relation linéaire entre X et Y. Comment et pourquoi pourrait-on pallier à cela ? Transformation des variables

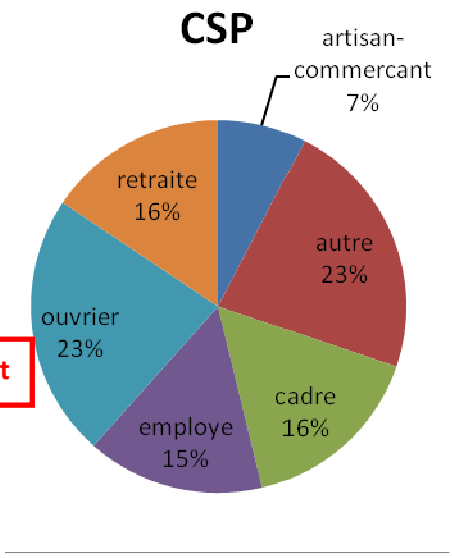
2. Etude de cas

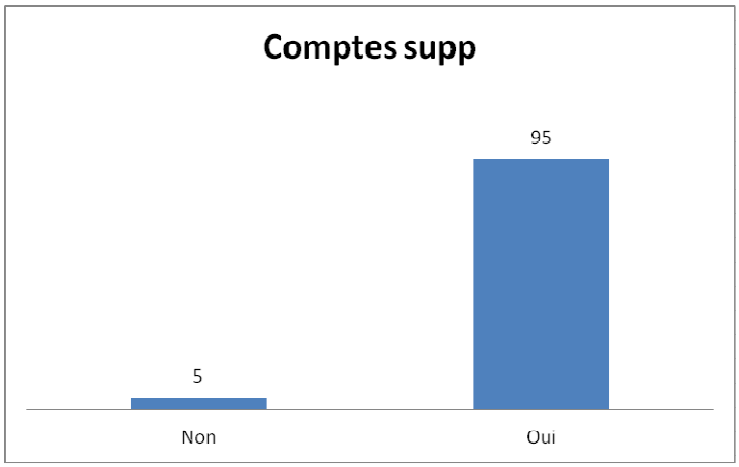
2.1 Croisement entre la CSP et les comptes supplémentaires d'un client

L'objectif est de déterminer si la catégorie socio-professionnelle a un impact sur le nombre de comptes supplémentaires d'un client. A partir des tableaux (1, 2, 3 et 4) ci-dessous, répondez aux questions suivantes.

1. Dessiner les graphiques appropriés pour représenter la distribution des colonnes « csp » et « compteSup »

1 point





2. Quel est le pourcentage d'ouvrier ? 23%

1 point

Quel est le pourcentage d'employés n'ayant pas de comptes supplémentaires? 0%
Quel pourcentage de cadres ont des comptes supplémentaires? 99%
Quel pourcentage de clients ayant un compte supplémentaire sont à la retraite ? 16%

3. A l'aide des profils lignes et des profils colonnes, que pouvez-vous dire sur le lien entre la catégorie socio-professionnelle et le nombre de comptes supplémentaires d'un client ?

1 point

csp / compteSup	artisan-com	autre	cadre	employé	ouvrier	retraite
NON	0	39	4	0	48	9
OUI	8	22	17	16	22	16
	8	**23**	**16**	**15**	**23**	**16**

A commenter

1 point

csp / compteSup	artisan-com	autre	cadre	employé	ouvrier	retraite	
NON	0	8	1	0	3	3	**5**
OUI	100	92	99	100	97	97	**95**

A commenter

4. Dans le cas où ces deux caractères seraient indépendants, quel devrait être l'effectif théorique des ouvriers n'ayant pas de comptes supplémentaires ? Commenter.

1 point

Pourcentage ouvrier = 21%, Pourcentage pas de compte supp = 5%
⇒ effectifs theo =0.23*0.05*500=0.005*114=5.75
L'effectif observé est de 11

2.2 Croisement entre la CSP et le montant des découverts

L'objectif est de déterminer si le montant cumulé des découverts d'un client est en rapport avec sa catégorie socio-professionnelle. A partir des résultats ci-dessous (Tab 5 et Fig 1), répondez aux questions suivantes.

1. Comparer les boites de Tuckey (Fig 1) des cadres et des employés. Commenter notamment la position du montant cumulé des découverts, la dispersion et la position de la moyenne par rapport à la variance.

2 points

Comparaison moyenne/médiane – Comparaison des dispersions – Position moyenne par rapport à la médiane. Attention ERREUR sujet comparaison moyenne/variance.

2. Calculer un indicateur numérique permettant d'établir s'il y a un lien entre les deux caractères.

2 points

Var intra=~55 000, var inter=~30 000⇒ Rapport de corrélation=~0.36 (1 point), c'est à dire 36% de la variabilité des découverts est expliquée par la CS (1point).

2.3 *Croisement entre le solde et le montant des découverts*

L'objectif ici est de prévoir le montant cumulé des découverts en fonction du solde. On étudie le lien éventuel uniquement dans la catégorie des cadres. Pour ce faire, on établit une relation linéaire entre les deux caractères. A partir de la figure 2 ci-dessous, répondez aux questions suivantes.

1. Quelle est l'équation de la droite permettant de prévoir le montant cumulé des découverts en fonction du solde? **1 point**
   Montant du découvert=-0.149*Solde+630.33
2. Quelle est la valeur du coefficient de corrélation ? **1 point**
   $|r_{xy}|=\sqrt{R^2}=0.88 \Rightarrow r_{xy}=-0.88$ car pente de la droite (0 point si pas le bon signe)
3. Quel est le pourcentage de variabilité du montant cumulé des découverts expliquée par la droite de régression fonction du solde ? Qu'en pensez-vous ? **1 point**
   78% (R^2) de la variabilité du montant cumulé des découverts expliquée par la droite de régression fonction du solde. Le modèle est acceptable.
4. A combien estime-t-on le montant cumulé des découverts pour un solde de 2000€ ? **1 point**
   Avec la droite de régression, on obtient un découvert de : -0.149*2000+630.33~332.33€
5. Deux valeurs extrêmes ne sont pas bien représentées par la droite de régression. Quelle étude aurait permis d'identifier ces valeurs avant de faire la droite ?
   Une analyse des quartiles et notamment les valeurs extrêmes dans la boite de Tuckey (0.5 point)
   Que se passerait-il pour la droite de régression si on supprimait ces valeurs ? **1 point**
   Le R^2 augmenterait certainement. La pente de la droite s'accentuerait (0.5 point)