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

Préambule

Cet examen dure 1h30. Vous pouvez utiliser votre PC uniquement en local. Les documents papiers sont autorisés. Le rendu est un document papier.

1. Questions de cours

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 ?
   Problème dans la question car confusion entre valeurs extremes et extrémités de la boite
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 ?
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 ?
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 ?
5. On dit souvent que les indicateurs moyenne et médiane sont complémentaires. Pourquoi ?
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 ?

2. Etude de cas

Le corpus de données (simulées) contient le résultat d'une enquête réalisée par les banques. L'objectif est de déterminer une typologie d'individus susceptibles d'être intéressés par une offre particulière. Les colonnes du fichier sont les suivantes :

solde	Solde moyen du compte courant sur les 12 derniers mois (en euros)
mdecouv	Montant cumulé des découverts sur le compte courant durant les 12 derniers mois (en euros)
ncompte	Nombre de comptes utilisés en plus du compte courant comme par exemple les livrets...
memprunt	Montant total des emprunts effectués sur les 3 dernières années (en euros)
mdepot	Montant total des versements effectués sur le livret d'épargne lors des 5 dernières années (en euros)
mretrait	Montant total des retraits effectués sur le livret d'épargne sur les 12 derniers mois (en euros)
nbenf	Nombre d'enfants de moins de 18 ans
age	Age du client
csp	Categorie socio-professionnelle du client : ouvrier/cadre/employé/retraité/artisan-commerçant/autre

Pour les besoins de l'étude certaines colonnes ont été modifiées :

- la colonne « ncompte » a été transformée en colonne « compteSup » prenant la valeur « OUI » si le client possède des comptes supplémentaires au compte courant et « NON » s'il n'en possède pas

- la colonne « nenf » a été transformée en colonne « enf » prenant la valeur « AVEC » si le client a des enfants et « SANS » s'il n'en a pas

Voici un extrait

solde	mdecouv	ncompte	memprunt	mdepot	mretrait	nbenf	age	csp	enf	compteSup
23,89	1560,98	0	0	406,2	0	1	45	autre	Avec	Non
237,9	1257,18	0	0	3771,14	0	7	53	ouvrier	Avec	Non
391,8	1292,17	0	0	2470,46	0	1	68	retraite	Avec	Non
2385,2	0	4	36181,47	70721,18	1878,89	0	27	cadre	Sans	Oui
2702,09	320,34	1	0	38414,72	2476,6	0	26	cadre	Sans	Oui
841,92	150	7	16783,54	8399,39	17666,33	0	42	employé	Sans	Oui
887,35	0	3	16768,29	8757,62	4215,47	0	50	employé	Sans	Oui

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 »
2. Quel est le pourcentage d'ouvrier ?
   Quel est le pourcentage d'employés n'ayant pas de comptes supplémentaires?
   Quel pourcentage de cadres ont des comptes supplémentaires?
   Quel pourcentage de clients ayant un compte supplémentaire sont à la retraite ?
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 ?
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.

csp / **compteSup**	*artisan-com*	*autre*	*cadre*	*employé*	*ouvrier*	*retraite*	**Tot**
NON	0	9	1	0	11	2	23
OUI	38	104	79	77	103	76	477
Tot	38	113	80	77	114	78	500

Tab. 1 Tableau des effectifs

csp / **compteSup**	*artisan-com*	*autre*	*cadre*	*employé*	*ouvrier*	*retraite*	**Freq. Marg.**
NON	0	2	0	0	2	0	5
OUI	8	21	16	15	21	15	95
Freq. Marg.	8	23	16	15	23	16	100

Tab. 2 Tableau des fréquences (%)

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

Tab. 3 Profils lignes (%)

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

Tab. 4 Profils colonnes (%)

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

	artisan-com	autre	cadre	employé	ouvrier	retraite	Tous
Effectif	38,0	113,0	80,0	77,0	113,0	78,0	500
Moyenne	276,1	682,0	185,1	462,4	579,2	508,1	487,4
Variance	79798,7	28665,8	41629,4	73158,3	49702,5	81126,9	83388,3
Q1	37,8	662,1	4,6	210,6	455,6	455,6	
médiane	222,3	719,6	107,6	554,0	650,7	624,5	
Q3	339,2	771,4	315,1	690,8	748,3	727,1	

Tab. 5 Indicateurs statistiques sur le montant cumulé des découverts par catégorie socio-professionnelle

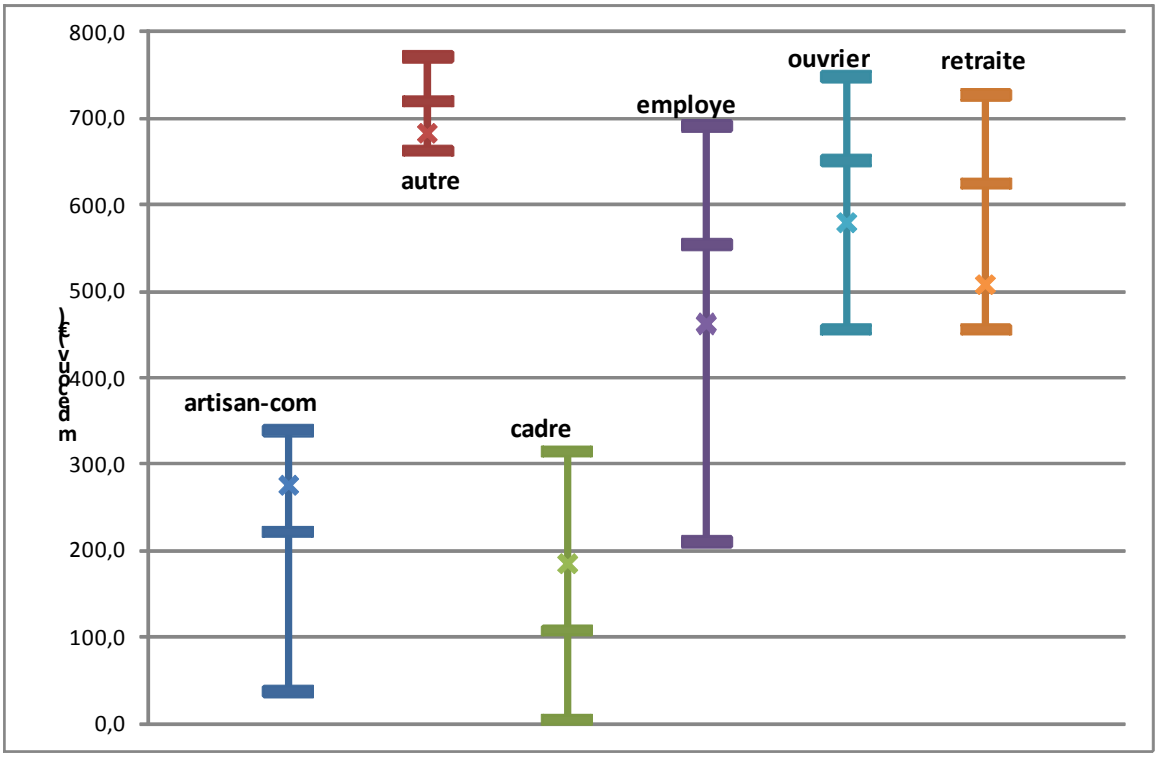

Fig. 1 Boites de Tuckey du montant cumulé des découverts par catégorie socio-professionnelle (croix pour la moyenne et barres horizontales pour les quartiles)

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?
2. Quelle est la valeur du coefficient de corrélation ?
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 ?
4. A combien estime-t-on le montant cumulé des découverts pour un solde de 2000€ ?
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 ? Que se passerait-il pour la droite de régression si on supprimait ces valeurs ?

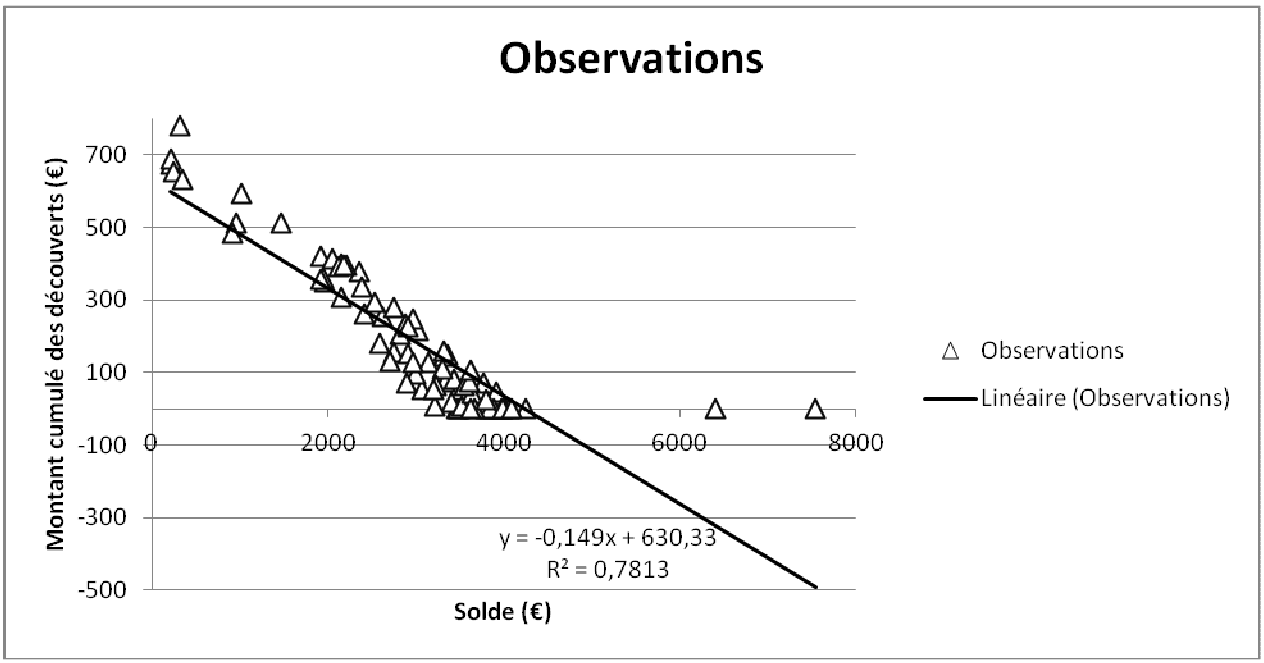

Fig. 2 Régression linéaire du montant cumulé des découverts en fonction du solde.