TP de Statistiques Descriptives – STID
Corrigé TP5

Attention! *Il n'existe pas une manière unique de construire un compte-rendu. Ce document contient seulement des éléments de correction pour la seconde partie du T.P.2.*

Partie 1– Liaison Variable qualitative/Quantitative

On considère une population de 534 individus de plus de 16 ans en activité au moment du recensement, en mai 1985.

Le jeu de données contient quatre variables qualitatives (nominales): le sexe (deux modalités), la catégorie professionnelle (six modalités) et les réponses (binaires) aux questions «vivez vous dans le sud?» et «êtes vous membre d'un syndicat?». Il contient également quatre variables quantitatives dont deux sont discrètes, le nombre d'année d'études et le nombre d'années d'expérience et deux sont continues, le salaire horaire exprimé en dollars et l'âge. Le nombre d'année d'études prend ses valeurs sur l'ensemble {0,1,2,...,18}, le nombre d'années d'expérience sur l'ensemble {0,1,2,...,}, le salaire horaire et l'âge sont définis sur R+.

On s'interroge sur l'existence d'un lien entre le salaire et la catégorie professionnelle.

Le Tableau 1 rappelle la répartition des différentes catégories professionnelles dans la population étudiée. On notera que 105 individus exercent une profession libérale (« Professional »), ce qui représente 20% de la population. Une proportion similaire (18%) est observée pour les employés de bureau (« Clerical »). Les commerciaux (« Sales »)ne représentent que 7% de l'effectif total. Cependant, nous n'avons pas d'information sur l'activité professionnelle d'une part importante (près de 30%) de la population, figurant dans la catégorie « Autre ».

Tableau 1: *Distribution en fréquences et effectifs de la catégorie professionnelle*

	Effectif	**Fréquence**
Management	55	10,3
Clerical	97	18,16
Professional	105	19,66
Sales	38	7,12
Service	83	15,54
Other	156	29,21

La question que l'on se pose est la suivante. Si on compare la distribution du salaire dans la population générale, et la distribution du salaire dans chacune des différentes catégories professionnelles (considérées donc comme des «groupes»), y-a-t-il des différences? Si c'est le cas, quelles sont les variations observées.

Le plan d'étude est le suivant: on commencera par mettre en évidence l'existence d'une éventuelle liaison d'un point de vue graphique, avant de comparer les moyennes et variances conditionnelles.

Puis, on calculera le rapport de corrélation et l'indicateur de Fisher (avec la p-valeur associée)

pour voir quelle part de la variabilité du salaire s'explique par la présence des différentes catégories professionnelles et si cette part est significative. Enfin, on conclura sur le lien entre les deux variables.

Représentations Graphiques de la liaison.

On représente à la fois la distribution du salaire dans la population totale (c'est-à-dire la distribution marginale) et dans chacun des groupes définis par les catégories professionnelles (distributions conditionnelles du salaire, sachant la catégorie professionnelle).

Indication: pour répondre à cette question, deux types de représentations graphiques sont possibles; en effet, la variable à expliquer étant quantitative continue, on pourrait choisir une représentation avec des histogrammes ou avec des boîtes à moustaches.

La Figure 1 représente la distribution du salaire horaire dans la population entière. La classe modale est la classe [4;8]. En effet, 41,2% des individus gagnent entre 4 et 8 dollars de l'heure. On notera également qu'une très faible part des individus gagnent plus de 16 dollars de l'heure (moins de 10%).

Le tableau 2 donne quelques indicateurs clés pour cette variable. On notera que 50% de la population gagne au moins 7,78 dollars de l'heure et que seulement 25% des individus gagnent plus de 11,25 dollars de l'heure.

Figure 1: *Histogramme du salaire horaire dans la population*

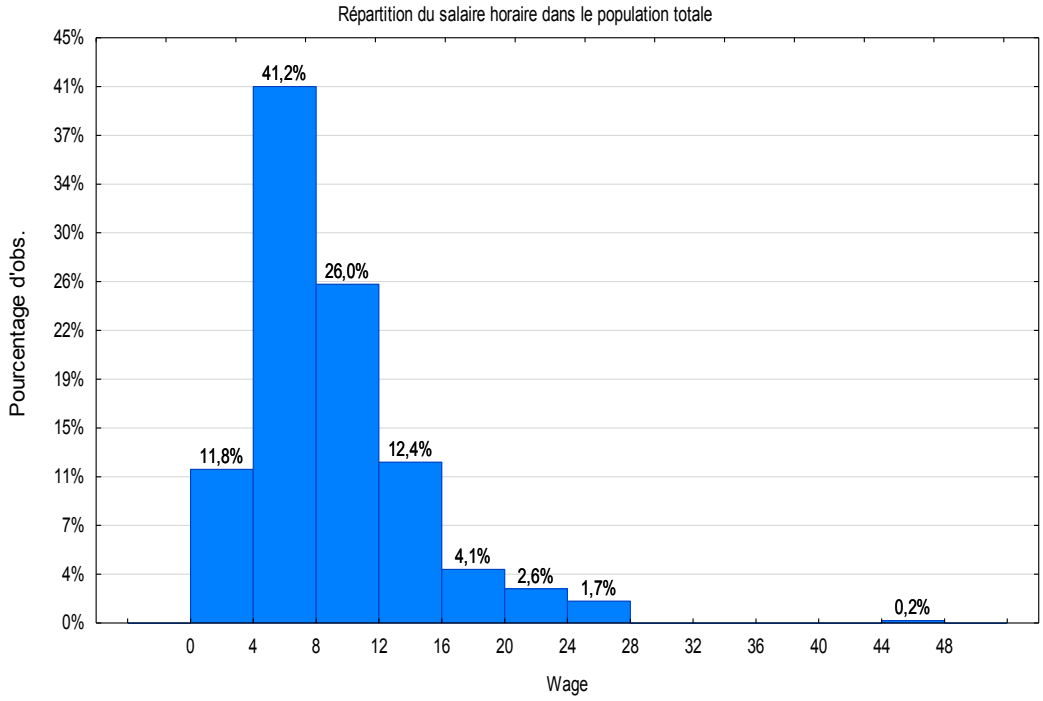


Tableau 2: Statistiques Descriptives pour la variable Salaire dans la population totale

	Moyenne	Médiane	1er - Quartile	3ème - Quartile	Ecart-type
Wage	9,02	7,78	5,25	11,25	5,14

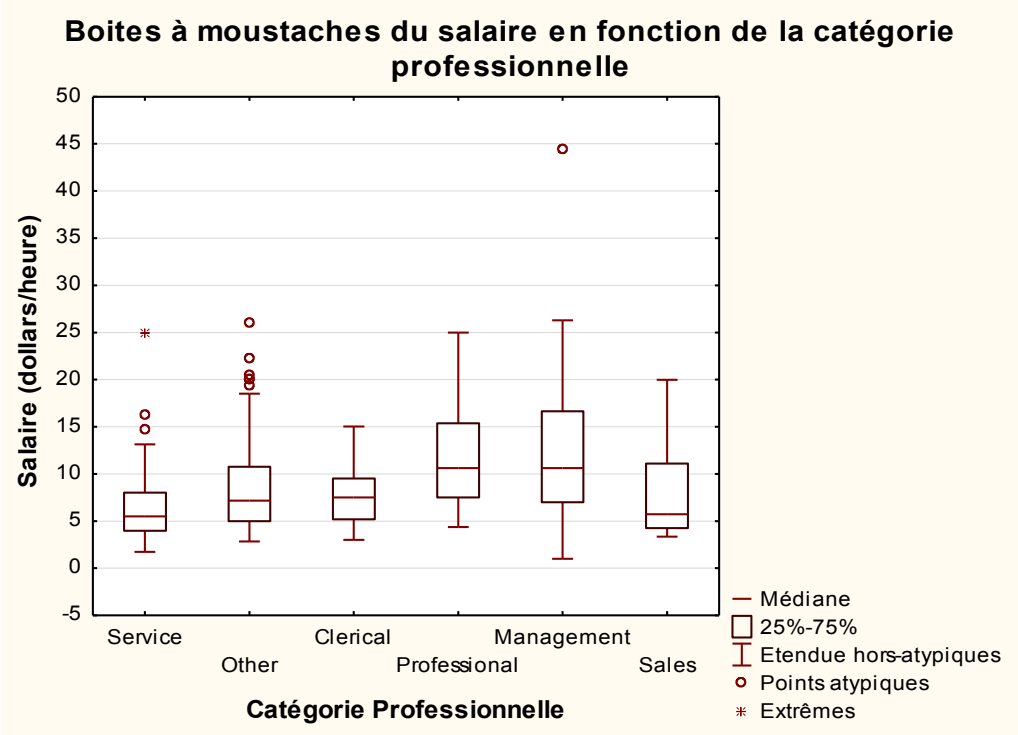
Considérons maintenant la distribution des salaires mais au sein de chaque catégorie professionnelle.

Quand on considère la répartition au sein de chacune des catégories, on s'aperçoit que les médianes pour les catégories professions libérales et cadres sont égales (10$), et bien supérieures à celle de la population totale. Au contraire, la médiane pour les employés de service et ceux qui travaillent dans la vente est inférieure à celle de la population totale.

On observe une plus faible variabilité chez les employés de bureaux et les employés de service (boîtes à moustache moins étendues) que chez les professions libérales et surtout chez les cadres, pour qui les salaires semblent relativement étendus.

Les catégories Employés de Bureau, Employés de Service et Commerciaux sont plus proches en terme de niveau de salaire horaire et s'opposent aux catégories Cadres et Professions Libérales.

Figure 2 : *Boites à moustaches du salaire horaire selon la catégorie professionnelle*



Indicateurs statistiques de la liaison

Si on compare les salaires moyens par catégorie professionnelle (Tableau 3), on constate qu'il est plus élevé chez les Cadres et chez les Professions Libérales (respectivement 12,7 et 12$ de l'heure) que pour l'ensemble de la population (9$ de l'heure). Cependant, ce sont aussi dans ces deux catégories que la variabilité est la plus grande.

On notera également une forte différence avec les autres catégories. Par exemple, un employé de bureau gagne en moyenne deux fois moins qu'un cadre. Les employés de Bureau et les commerciaux ont des salaires horaires moyens très proche (autour de 7,5 dollars de l'heure). Les employés de bureau ont un niveau de salaire horaire très homogène (variance la plus faible).

Tableau 3 : *Moyennes et variances du salaire (dollars par heure) conditionnellement à la catégorie professionnelle*

Catégorie Professionnelle	Salaire moyen	Variance du salaire
Employé de Service	6,54	13,49
Autre	8,43	18,07
Employé de Bureau	7,42	7,28
Profession Libérale	11,95	30,51
Cadre	12,70	57,34
Commerciaux	7,59	17,91
Tous Groupes	9,02	26,41

Intéressons nous maintenant aux variances inter-groupes et intra-groupes. *Le tableau 4 est une « image » de ce que vous deviez obtenir dans le fichier Excel.*

La variance intra-groupes (moyenne des variances conditionnelles) est supérieure à la variance inter-groupes (variance des moyennes conditionnelles). La dispersion dans les groupes est donc forte, alors que la différence entre les moyennes de chaque groupe est faible.

Un rapport de corrélation de 0 indique qu'il n'existe aucun lien entre les deux variables étudiées, un rapport de 1 signifie que le lien entre les deux variables est très fort, c'est à dire que les deux variables sont dépendantes.

Ici, rapport de corrélation est de 0.18, ce qui signifie que la part de variabilité des salaires est expliquée à hauteur de 18% par la présence des groupes définis par les catégories professionnelles.

Tableau 4 : *Récapitulatif permettant de calculer les variances inter-groupes et intra-groupes du salaire horaire, sachant la catégorie professionnelle, et le rapport de corrélation*

Variance inter (variance des moyennes conditionnelles)

Modalités de la catégorie professionnelle	Management	Services	Other	Clerical	Sales	Professional
Salaire moyen sachant la catégorie professionnelle mx|y=j	12,704	6,537	8,426	7,423	7,593	11,947
Fréquences de la catégorie professionnelle en %	10,300	15,543	29,213	18,165	7,116	19,663
Fréquences de la catégorie professionnelle fj	0,103	0,155	0,292	0,182	0,071	0,197
Salaire moyen (marginal) mx	9,024					
Carrés des écarts (mx|y=j-mx)^2	13,542	6,183	0,357	2,565	2,049	8,546
fj* (mx|y=j-mx)^2	1,395	0,961	0,104	0,466	0,146	1,680
Variance inter	4,752					

Variance intra (moyenne des variances conditionnelles)

Modalités de la catégorie professionnelle	Management	Services	Other	Clerical	Sales	Professional
Variances du salaire conditionnellement à la catégorie professionnelle s2x|y=j	57,340	13,490	18,070	7,280	17,910	30,510
Fréquences de la catégorie professionnelle fj	0,103	0,155	0,292	0,182	0,071	0,197
fj*s2x|y=j	5,91	2,1	5,28	1,32	1,27	6
Variance intra	21,877					

Variance totale	26,63
Rapport de corrélation (variance inter/variance totale)	0,18
Nombre d'individus n	534
Nombre de modalités de la variable catégorie professionnelle	6
Indicateur de Fisher ((variance inter/(p-1)/(variance intra/(n-p))	22,94

Ces résultats peuvent être retrouvés dans Statistica (tableau 5).

Le rapport de corrélation est toujours de 0,18 et a déjà été commenté. Regardons le test de Fisher. La statistique de test vaut 23,22 mais n'est pas interprétable seule. En effet,il faut regarder la p-valeur. En effet, si cette p-valeur est inférieure au seuil de 5% (0,05) on dira que le test est significatif et donc que les deux variables sont biens dépendantes. En revanche, si elle est supérieure à 0,05 on rejettera cette hypothèse pour conclure qu'il n'existe aucun lien entre les deux variables d'intérêts.

Ici, la p-valeur (notée « p » dans le tableau 5) est très proche de 0 et inférieure à 0,05. On conclue que le test est significatif, c'est à dire qu'il existe un lien entre le salaire et la catégorie professionnelle.

Tableau 5 : *Analyse de la variance*

	Multiple - R^2	**F**	**p**
Wage	0,18	23,22	0,00

En conclusion, on a ainsi d'abord pu constater graphiquement que des différences semblaient exister entre les différentes catégories professionnelles et également au sein de certaines d'entre elles (cadres et professions libérales), concernant le salaire horaire. Nous avons confirmé ceci à l'aide des moyennes et variances conditionnelles du salaire pour chaque groupe. Enfin, même si le rapport de corrélation ne prenait pas une valeur élevée, la liaison entre les deux variables a été confirmée par le test de Fisher.

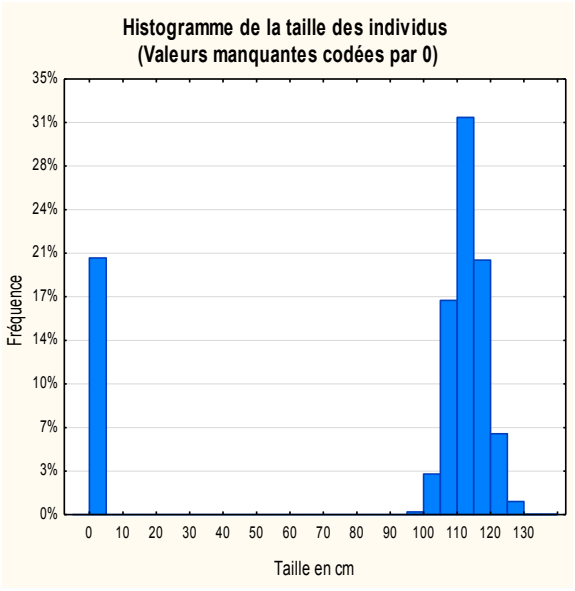
Le faible rapport de corrélation peut s'expliquer par le fait que parmi les catégories professionnelles, certaines sont très proche en terme de niveau de salaire. Un idée serait de regrouper ces catégories et de refaire l'analyse. La présence de la classe « Autres » qui représente un pourcentage non négligeable de la population peut également être à l'origine de ces résultats. Enfin l'absence de résultat concluant pour le rapport de corrélation peut également venir de la très forte variabilité observée pour certaine classes (Cadres et Professions Libérales).

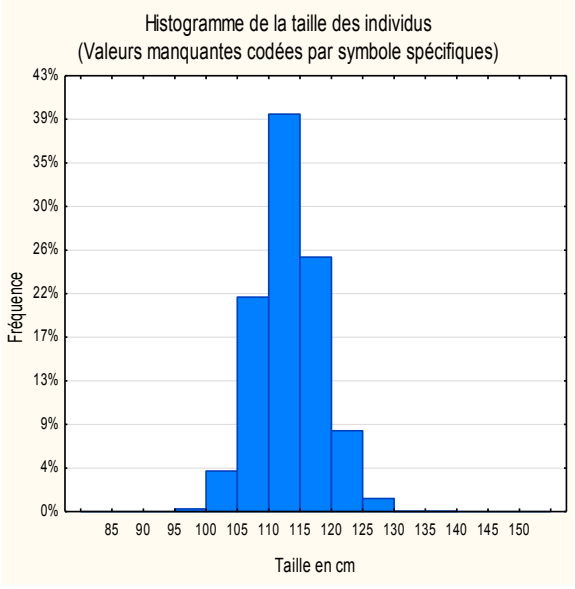
Partie 2: Traitement des valeurs manquantes

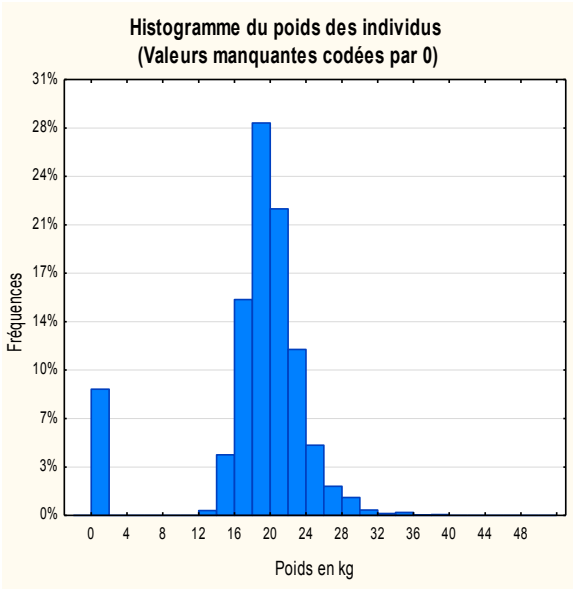
Nous étudions un jeu de données dans lequel on trouve des mesures, taille et masse, relevées sur 2894 enfants, garçons et filles, âgés de 4 à 7 ans. Dans ce jeu de données, il y a 20,4% de données manquantes pour la variable taille et 9% pour la variable masse. En tout, 26% des individus présentent des valeurs manquantes.

Coder les données manquantes par des zéros n'est pas convenable, comme le montrent les histogrammes de la Figure 1 et la Table 1. Cela fausse notamment les indicateurs statistiques. Par exemple, la taille moyenne est de 113,7cm si l'on code les données manquantes par un symbole spécifique reconnu par le logiciel, ce qui revient à ne pas tenir compte des individus avec données manquantes, alors qu'en codant ces valeurs par des 0, la moyenne est déplacée artificiellement vers le bas (90,56cm). La comparaison des écart-types est intéressante également : l'écart-type est beaucoup plus important dans le cas où les données manquantes sont codées par des 0, ce qui traduit le fait que les vraies valeurs prises par la variable taille sont très éloignées de 0 : elles se situent presque toutes entre 100cm et 130cm. Les mêmes remarques sont valables pour la variable masse, dans une moindre mesure. En effet, le phénomène étant plus ou moins marqué selon l'ordre de grandeur des valeurs prises par la variable.

Figure 1 : *Distribution de la taille et du poids des enfants du jeu de données "Enfants avec données manquantes" en codant les valeurs manquantes par des 0 puis par un symbole spécifique.*







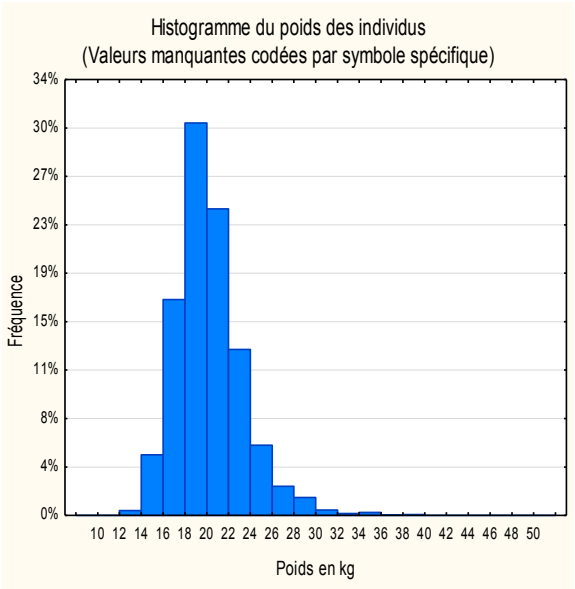


Table 1 : *Indicateurs statistiques pour les variables Taille et Poids, selon que l'on code les Valeurs Manquantes par 0 (variables POIDS et TAILLE) ou par un symbole spécifique (variables POIDSBIS et TAILLEBIS)*

	N Actifs	**Moyenne**	**Médiane**	**Ecart-type**
Taille	2894	90,56	112,00	46,00
Taille_Bis	2305	113,70	114,00	5,03
Poids	2894	18,80	20,00	6,61
Poids_Bis	2633	20,67	20,00	3,09

Supprimer les individus avec données manquantes est une manière simple de traiter ces dernières, qui est plus pertinente que de les remplacer par des 0. En général, il conviendra néanmoins d'être vigilant, en se demandant par exemple si le pourcentage de données manquantes n'est pas trop important pour supprimer tous les individus concernés, si les données manquantes correspondent à des groupes particuliers et si cela influence l'étude: une question importante est donc de savoir si elles concernent un sous groupe particulier de la population étudiée ou touchent toute la population de façon uniforme.