



COURS DE FOUILLE DE DONNÉES

M1, P-20- IFI 2016

Enseignante: NGUYỄN Thị Minh Huyền

TP3 – G2

Analyse Descriptive du jeu de données

Données Utilisées: Census Income Data Set

Étudiant: Ginel Dorleon

Gervais Fotsing Sikadie S.

ANALYSE DESCRIPTIVE

L'analyse descriptive comme le nom l'indique sert à analyser et décrire des données pour obtenir un rendu final. Ce sont de simples calculs mathématiques qui permettent de dégager des données une réelle tendance positive ou négative des résultats. A partir de ces chiffres, des graphiques viennent en complément pour appuyer l'analyse statistique.

L'analyse ou la statistique descriptive est la base de toute analyse de données. En effet, avant d'approfondir l'analyse dans les détails, il faut commencer par la description globale à l'aide de ces statistiques.

Nous entendons par exemple par analyse/statistique descriptive le calcul de la moyenne et de la médiane, deux indicateurs très importants et surtout différents et d'autres calculs et test selon le besoin. [1]

Dans cette deuxième partie de travail pratique, nous allons procéder à l'analyse descriptive de deux des variables de notre jeu de données.

RAPPEL/DESCRIPTION SUR NOS DONNÉES: **Census Income Data Set** [2]

Les données que nous utilisons pour cette 2^e partie ont été extraites de la base de données d'un bureau de recensement aux USA. L'objectif est d'étudier un ensemble de paramètre et d'établir une tâche de prévision afin de déterminer le profil des personnes qui gagnent plus de 50K par année.

NOS VARIABLES

Age: Variable continue représentant l'âge des personnes.

Workclass: Variable discrète prenant les valeurs suivantes : Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked

Fnlwgt: Final weight, une variable continue calculée à partir des données d'origine et du sexe de la personne.

Education: Variable discrète déterminant le niveau d'éducation de la personne, elle prend les valeurs suivantes : Licence, Certains collège, 11e, HS-grad, Prof-école, Assoc-ACDM, Assoc-voc, 9e, septième-huitième, 12e, Masters, 1ere-4ème, 10e, Doctorat, 5ème-6ème, Maternelle.

education-num: continu.

statut matrimonial: Marié-civ-conjoint, Divorcé, jamais marié, Séparé, Veuf, Marié-conjoint absent, Marié-AF-conjoint.

Profession: Tech-support, Craft-réparation, Autre service, ventes, Exec-gestion, Prof spécialité, Handlers-nettoyants, machine-op-inspct, Adm-clérical, agriculture-pêche, Transport-mobile, Priv-ménage serv, protection-serv, armées-Forces.

Relationship : Variable discrète représentant les types de relations des individus. Elle prend les valeurs suivantes : Femme, propre enfant, Mari, Non-in-famille, Autre-parent, Unmarried.

Race: Variable discrète représentant la race des individus. Elle prend les valeurs Noir Blanc, Asie-Pac-Islander, Amer-Indian-Eskimo, Autre.

Sexe: Variable discrète représentant le sexe de l'individu, 2 valeurs : Femme, Homme.

Capital-gain: Variable continue représentant le capital gagné par l'individu.

Capital-loss: Variable continue représentant le capital perdu par l'individu.

Hours-per-week : Variable continue représentant le nombre d'heur que l'individu travail

Pays: Variable discrète représentant le pays d'origine de l'individu.

STATISTIQUE SUR LES VARIABLES

Dans la plupart des études sur les données , le nombre de sujets est souvent trop important pour que l'on puisse présenter les données réelles de chaque individu. C'est pourquoi, il est nécessaire de trouver un moyen qui donne le maximum d'informations possible sous le format le plus utile. Une manière courante de présentation de données est la représentation graphique ou le tableau. Les tableaux sont commodes pour présenter l'information relative aux données individuelles et les graphiques pour donner un profil général des observations. Toutefois, il est également utile de donner un résumé chiffré. Pour les variables qualitative ou en catégorie (niveau d'étude, sexe, absence-présence d'une maladie, niveau d'éducation, etc.), la mesure la plus instructive est la proportion d'individus entrant dans chaque catégorie. Les variables quantitatives (poids, taille, âge, salaire, etc.) nécessitent quant à elle deux types de mesure pour avoir une idée complète de la distribution des observations : la mesure de la *position centrale* des observations et la mesure de leur *dispersion*, c'est-à-dire la mesure de la répartition des

observations autour de cette position centrale. Ainsi, nous allons faire procéder à l'étude qualitative et quantitative de deux des variables de notre jeu de données.

Étude d'une variable qualitative- Education

Généralement, en statistique, une variable qualitative, on dit aussi catégorielle, est une variable pour laquelle la valeur mesurée sur chaque individu ne représente pas une quantité. Les différentes valeurs que peut prendre cette variable sont appelées les *catégories*, *modalités* ou *niveaux*.

Ainsi, parmi l'ensemble des variables qualitatives de notre jeu de données, nous avons choisi d'étudier la variable *Education*. En effectuant la statistique primaire de cette variable, on peut constater que la plus grande valeur valeur Hs-grad signifiant le degré en High School. On constate que le mode est cette valeur, Hs-grad. Le nombre d'individu ayant ce niveau est de 10501 et le pourcentage d'individus ayant ce niveau est de 32.25 % . Voir capture ci-dessous

Univariate discrete stat 1				
Parameters				
Attributes : 1				
Examples : 32561				
Results				
Attribute	Gini	Distribution		
		Values	Count	Percent
EDUCATION	0,8096	Bachelors	5355	16,45 %
		Hs-grad	10501	32,25 %
		11th	1175	3,61 %
		Masters	1723	5,29 %
		9th	514	1,58 %
		Some-college	7291	22,39 %
		Assoc-acdm	1067	3,28 %
		Assoc-voc	1382	4,24 %
		7th-8th	646	1,98 %
		Doctorate	413	1,27 %
		Prof-school	576	1,77 %
		5th-6th	333	1,02 %
		10th	933	2,87 %
		1st-4th	168	0,52 %

Étude d'une variable quantitative – L'Age

Une variable est *quantitative* si elle reflète une notion de grandeur, c'est-à-dire si les valeurs qu'elle peut prendre sont des nombres. Une grandeur quantitative est souvent exprimée avec une unité de mesure qui sert de référence

Parmi l'ensemble des variables quantitatives de notre jeu de données, nous avons choisi d'étudier la variable quantitative Age. En effectuant l'analyse statistique primaire de

cette variable, on constate que l'âge minimum est de 17 ans, le max est de 90 ans et l'âge moyenne est 38,5816. Voir capture ci-dessous.

Univariate continuous stat 1					
Parameters					
Attributes : 1					
Examples : 32561					
Results					
Attribute	Min	Max	Average	Std-dev	Std-dev/avg
AGE	17	90	38,5816	13,6404	0,3535
Computation time : 0 ms.					
Created at 21/12/2016 15:14:43					

More Univariate cont stat 1					
Parameters					
Attributes : 1					
Examples : 32561					
Results					
Attribute	Stats		Histogram		
AGE	Statistics		Values	Count	Percent
	Average	38,5816	x_<_24,3000	5570	17,11%
	Median	37,0000	24,3000_=<_x_<_31,6000	5890	18,09%
	Std dev. [Coef of variation]	13,6404 [0,3535]	31,6000_=<_x_<_38,9000	6048	18,57%
	MAD [MAD/STDDEV]	11,1892 [0,8203]	38,9000_=<_x_<_46,2000	6163	18,93%
	Min * Max [Full range]	17,00 * 90,00 [73,00]	46,2000_=<_x_<_53,5000	3967	12,18%
	1st * 3rd quartile [Range]	28,00 * 48,00 [20,00]	53,5000_=<_x_<_60,8000	2591	7,96%
	Skewness (std-dev)	0,5587 (0,0136)	60,8000_=<_x_<_68,1000	1595	4,90%
	Kurtosis (std-dev)	-0,1661 (0,0271)	68,1000_=<_x_<_75,4000	496	1,52%
			75,4000_=<_x_<_82,7000	174	0,53%
			x>=_82,7000	67	0,21%

On peut ensuite observer les informations suivantes :

- La médiane est de 37
- La classe modale est [38,9000 ; 46,2000[, le mode est 42.5500
- L'écart-type est de 13,6404

On peut ensuite constater la dispersion de la variable Age.

L'étendue (différence entre la valeur max et la valeur min) est de 73,00

la dispersion interquartile (différence entre le troisième et le premier quartile) est de 20,00;

La représentation graphique ici est sous forme d'histogramme.

La forme de la distribution de l'Age est, suite aux statistiques primaires, étalée à droite car la moyenne est > médiane > mode (on peut aussi le constater en observant que le coefficient d'asymétrie de Fisher (Skewness = 0,5587) est positif.

La forme de la distribution de l'Age est moins aplatie qu'une distribution normale car coefficient d'aplatissement observé ici (Kurtosis = -01661) est négatif.

Corrélation entre chaque paire d variable quantitative

En probabilité et en statistique, étudier la corrélation entre deux ou plusieurs variables aléatoires ou statistiques *numériques*, c'est étudier l'intensité de la liaison qui peut exister entre ces variables. La mesure de la corrélation linéaire entre les deux se fait alors par le calcul du coefficient de corrélation linéaire, noté r. Dans le tableau suivant, nous présentons la corrélation entre chaque paire de nos variables quantitatives.

Linear correlation 2

Parameters

Cross-tab parameters

Sort results

non

Input list

Target (Y) and input (X)

Results

Y	X	r	r ²	t	Pr(> t)
AGE	FNLWGT	-0,0766	0,0059	-13,8709	0,0000
AGE	EDUCATION-NUM	0,0365	0,0013	6,5954	0,0000
AGE	CAPITAL-GAIN	0,0777	0,0060	14,0581	0,0000
AGE	CAPITAL-LOSS	0,0578	0,0033	10,4423	0,0000
AGE	HOURS-PER-WEEK	0,0688	0,0047	12,4358	0,0000
FNLWGT	AGE	-0,0766	0,0059	-13,8709	0,0000
FNLWGT	EDUCATION-NUM	-0,0432	0,0019	-7,8014	0,0000
FNLWGT	CAPITAL-GAIN	0,0004	0,0000	0,0779	0,9379
FNLWGT	CAPITAL-LOSS	-0,0103	0,0001	-1,8499	0,0643
FNLWGT	HOURS-PER-WEEK	-0,0188	0,0004	-3,3872	0,0007
EDUCATION-NUM	AGE	0,0365	0,0013	6,5954	0,0000
EDUCATION-NUM	FNLWGT	-0,0432	0,0019	-7,8014	0,0000
EDUCATION-NUM	CAPITAL-GAIN	0,1226	0,0150	22,2958	0,0000
EDUCATION-NUM	CAPITAL-LOSS	0,0799	0,0064	14,4677	0,0000
EDUCATION-NUM	HOURS-PER-WEEK	0,1481	0,0219	27,0256	0,0000

Dans le tableau ci-dessus, r représente le coefficient de corrélation pour chaque paire de variable, r^2 est l'écart-type

Contingence des variables qualitatives

Pour évaluer le lien existant entre deux de nos variables qualitatives, on va présenter un tableau de contingence. Le tableau de contingence est une méthode de représentation de données issues d'un comptage permettant d'estimer la dépendance entre deux caractères. Elle consiste à croiser deux caractères d'une population, par exemple dans notre cas, *workclass* et *education*, en dénombrant l'effectif correspondant à la conjonction «caractère 1» et «caractère 2». Les effectifs partiels sont rassemblés dans un tableau à double entrée, par ligne pour le premier caractère, et par colonne en fonction du second caractère: c'est le «tableau de contingence».

Cet outil simple répond à un problème crucial en statistique: la détection d'éventuelles dépendances entre les qualités relevées sur les individus d'une population.

Ainsi, nous présentons le tableau suivant avec la variable *Workclass* d'une part et le niveau d'éducation (variable *education*) de l'individu.

Contingency Chi-Square 2												
Parameters												
Cross-tab parameters												
Sort results			non									
Input list			Target (Row) and input (Column)									
Additional information			0									
Contribution threshold			2,0									
Results												
Row (Y)	Column (X)	Statistical indicator										
WORKCLASS	EDUCATION	Stat	Value		Bachelors	HS-grad	11th	Masters	9th	Some-college	Assoc-acdm	Ass
		d.f.	120	State-gov	270	268	14	169	6	325	41	
		Tschuprow's t	0,084918	Self-emp-not-inc	399	866	60	124	34	486	71	
		Cramer's v	0,099369	Private	3551	7780	923	894	387	5094	729	
		Phi²	0,078993	Federal-gov	212	263	9	67	3	254	55	
		Chi² (p-value)	2572,10 (0,0000)	Local-gov	477	503	36	342	23	387	88	
		Lambda	0,000000	?	173	532	118	48	51	514	47	
		Tau (p-value)	0,0177 (0,0000)	Self-emp-inc	273	279	14	79	10	226	35	
		U(R/C)	0,0286 (p-value)	Without-pay	0	9	0	0	0	3	1	
				Never-worked	0	1	1	0	0	2	0	
				Sum	5355	10501	1175	1723	514	7291	1067	

On peut constater que, le secteur prive est le secteur qui offre le plus grand nombre d'emploi.

Les individus avec le niveau de Hs-grad représentent le plus grande valeur des employés. Il n'y a pas de Bacheliers, ni de niveau Masters, ni de niveau de 9th sans emplois ou qui ne sont jamais été employés .

Conclusion

En effet, dans ce travail, nous avons présenté l'analyse statistique et descriptive des variables de notre jeu de données. La moyenne, la médiane, l'écart-type, la corrélation, la contingence, le *Test du Khi 2*, *ect.*, sont quelques paramètres statistiques présents dans notre travail.

Dans un prochain travail, nous présenterons une autre partie d'analyse de notre jeu de données.

Références

[1] <http://www.analyse-donnees.fr/services-analyse-enquetes-traitement/>

[2] <http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names>

[3] Cours de Fouilles de Données, Master 1, IFI 2016, Enseignante NGUYỄN Thị

Minh Huyền