TP d'Analyse de Données : TP 1

Objectifs du TP :

- Étudier la liaison entre deux variables
- Étudier la différence entre test paramétrique et non paramétrique

Les données simulées :

- https://raw.githubusercontent.com/agusbudi/DataAnalysis/master/data1TP1.txt
- https:// raw.githubusercontent.com/agusbudi/DataAnalysis/master/data2TP1.txt

Le Calcul des Coefficients de Corrélation

Les coefficients de corrélation permettent de donner une mesure synthétique de l'intensité de la relation entre deux caractères et de son sens lorsque cette relation est monotone. Le coefficient de corrélation de Pearson permet d'analyser les relations linéaires et le coefficient de Spearman permet d'analyser les relations non-linéaires monotones. Une relation de Spearman ou de Pearson est représentée dans [-1, +1].

Nous allons commencer par le fichier *data1TP1.txt*. Étant donné un tableau de cinq variables (A, B, C, D, E), une étiquette (Y), et 15 lignes.

1. Tracez en dimension 2 le nuage de 15 points pour chaque variable. Que pouvez-vous observer ?
2. Pour calculer le coefficient **r** de Pearson, créez une fonction basée sur la formule ci-dessous :

$$\mathbf{r}(\mathbf{X},\mathbf{Y}) = \frac{\boldsymbol{Cov}(\boldsymbol{X},\boldsymbol{Y})}{\boldsymbol{\sigma_X} \,.\, \boldsymbol{\sigma_Y}}$$

   avec ***Cov*** est la convariance (la moyenne du produit des écarts à la moyenne) et $\boldsymbol{\sigma}$ est l'écart type. Quelle variable a la plus petite corrélation ? Pourquoi ?

   Vérifiez votre programme avec la fonction* : **cor (X, Y)
3. Créez une fonction du coefficient de Spearman en considérant cette formule :

$$\boldsymbol{\rho(X,Y)} = \mathbf{1} - \frac{\mathbf{6}\ \sum_{i=1}^{N}[\boldsymbol{R(x_i)} - \boldsymbol{R(y_i)}]^2}{\boldsymbol{N^3 - N}}$$

   avec $\mathbf{R(x_i)}$ est le rang de x_i, et $X = \{x_1, x_2, \ldots, x_i, \ldots x_N\}$; $\mathbf{R(y_i)}$ est le rang de y_i, et $Y = \{y_1, y_2, \ldots, y_i, \ldots y_N\}$. Comparez le score obtenu au résultat de la question 2. Quelle est la différence ?

   Vérifiez votre programme avec la fonction* : **cor(X, Y, method = "spearman")
4. Comment calculer la relation non-linéaire et non-monotone entre variables E et Y ? Proposez votre idée.

Tests de validation d'hypothèses

On étudie deux variables A et B définies sur une population P. On veut tester l'existence d'une liaison entre les deux variables. Le test est effectué pour un risque alpha α fixé. Pour réaliser le test, on a tiré au sort un échantillon d'individus de taille **n** dans la population.

Test Paramétrique

On va appliquer le test de Student/t et utiliser le fichier *data2TP1.txt* qui représente le coût de la vie quotidienne à Marseille et à Aix-en-Provence en 2019.

5. *Test d'indépendance pour une variable quantitative.* Pour calculer le score de t, créez une fonction basée sur la formule ci-dessous :

$$\mathbf{t} = \frac{|\mathbf{m} - \boldsymbol{\mu}|}{\boldsymbol{\sigma}/\sqrt{\boldsymbol{n}}}$$

   avec **m** est la moyenne observée et $\boldsymbol{\mu}$ est la moyenne théorique.

   Pour savoir si la différence est significative, il faut lire dans la table t (*t table.pdf*), la valeur critique correspondant au risque alpha fixé pour un degré de liberté = n – 1.

   - La moyenne du cout de la vie quotidienne à Marseille en 2010 est 19€. Avec α = 5%, calculez si l'inflation 2010-2019 a affecté le coût de la vie à Marseille.

6. *Test d'indépendance pour deux variables quantitatives.* Pour calculer le score de t, créez une fonction basée sur la formule ci-dessous :

$$\mathbf{t} = \frac{|\mathbf{m_1} - \mathbf{m_2}|}{\sqrt{\frac{\boldsymbol{\sigma_1^2}}{\boldsymbol{n_1}} + \frac{\boldsymbol{\sigma_2^2}}{\boldsymbol{n_2}}}}$$

Degré de liberté = $n_1 + n_2 - 2$

- Avec $\alpha = 5\%$, existe-t-il une dépendance significative entre Marseille et Aix-en-Provence ? Comparez le résultat avec p=2% et donnez votre commentaire.

Test Non Paramétrique

7. *Test d'indépendance pour une variable qualitative.* On va utiliser le résultat d'une expérimentation génétique qui a étudié l'hérédité de pois de senteur (Bateson et al., 1905). (voir tableau A).

Tableau A. Hérédité de pois de senteur.

	Phénotype			
	violet, long	violet, rond	rouge, long	rouge, rond
ratio génétique	9	3	3	1
observé	1528	106	117	381

   a. Selon le ratio et **n** plantes observées, calculez la valeur théorique de chaque catégorie de phénotype.
   b. Créez une fonction du Khi deux (X^2) en utilisant cette formule:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

   où k est le nombre de classe distinct, O est la valeur observée, E est la valeur théorique.
   c. Avec $\alpha = 5\%$ (voir *table khi deux.pdf*), degré de liberté = k - 1, et H_0 : le vrai ratio est 9 : 3 : 3 : 1, quelle est votre conclusion?

8. *Test d'indépendance pour les variables qualitatives.* On va utiliser trois variables de données Melanome : la diagnostic (non melanome :common nevus et atypical nevus, melanome), la forme du mélanome (absent, présent : atypique et typique), et la présence de la couleur marron clair (voir tableau B).

Tableau B. Melanoma dataset

Observed		**Form : Dots / Globules**			**Color: Light brown**	
		Absent	Atypical	Typical	Absent	Present
Clinical diagnosis	common nevus	29	5	46	20	60
	atypical nevus	40	32	8	29	51
	melanoma	18	22	0	12	28

Avec $\alpha = 5\%$ et H_0 : deux variables sont indépendantes, comparez le résultat du test du Khi Deux entre le diagnostic-la forme et le diagnostic-la couleur. Quelle variable est importante pour détecter un mélanome et pourquoi ?

Conclusion

9. Selon les questions précédentes, pourquoi le test de Student/t est classé comme paramétrique et le test du Khi Deux est classé comme non paramétrique ? Pouvons-nous appliquer le test de Student/t aux données qualitatives?
10. Pouvons-nous appliquer le coefficient de Pearson et le coefficient de Spearman aux données qualitatives?