TP 2 : Correction

Calculs bivariés et inférentiels dans SPSS

Ceci est un corrigé type. D'autres procédures peuvent être utilisées pour obtenir des résultats semblables.

1. Charger le fichier de données « Santé_18.sav », puis étudier le lien existant entre les variables *Santé* et *Docteur*.

La fonction CROSSTABS (Analyse -> Statistiques descriptives -> Tableaux croisés) permet d'obtenir une table de contingence (tableau croisé) et des mesures statistiques (chi-2 et V de Cramer). Il faut sélectionner à l'aide des boutons « Statistiques » et « Cellules » les éléments que l'on veut voir affichés.

Récapitulatif du traitement des observations

	Observations					
	Valide		Manquante		Total	
	N	Pourcent	N	Pourcent	N	Pourcent
Santé en général * Consultation d'un docteur au cours des 12 derniers mois	162	58.9%	113	41.1%	275	100.0%

Il y a beaucoup d'observations manquantes (113 sur 275).

Tableau croisé Santé en général * Consultation d'un docteur au cours des 12 derniers mois

			Consultation d'un docteur au cours des 12 derniers mois		
			oui	non	Total
Santé en général	très bonne	Effectif	34	13	47
		% compris dans Santé en général	72.3%	27.7%	100.0%
	bonne	Effectif	76	26	102
		% compris dans Santé en général	74.5%	25.5%	100.0%
	moyenne	Effectif	10	1	11
		% compris dans Santé en général	90.9%	9.1%	100.0%
	très mauvaise	Effectif	2	0	2
		% compris dans Santé en général	100.0%	.0%	100.0%
Total		Effectif	122	40	162
		% compris dans Santé en général	75.3%	24.7%	100.0%

Le tableau croisé montre que la majorité des gens ont consulté au moins une fois un docteur durant les 12 derniers mois (122 sur 162) et que la très grande majorité des gens se déclarent en bonne ou très bonne santé. En regardant les pourcentages par ligne (il faut les demander en passant par le bouton « Cellules », car ils n'apparaissent pas par défaut), on s'aperçoit que plus l'état de santé est mauvais, plus forte est la probabilité d'avoir consulté le docteur. Il existe donc un lien dans l'échantillon entre ces deux variables, mais qu'en est-il au niveau de la population ?

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	2.353[a]	3	.502
Rapport de vraisemblance	3.153	3	.369
Association linéaire par linéaire	1.600	1	.206
Nombre d'observations valides	162		

a. 3 cellules (37.5%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de .49.

Mesures symétriques

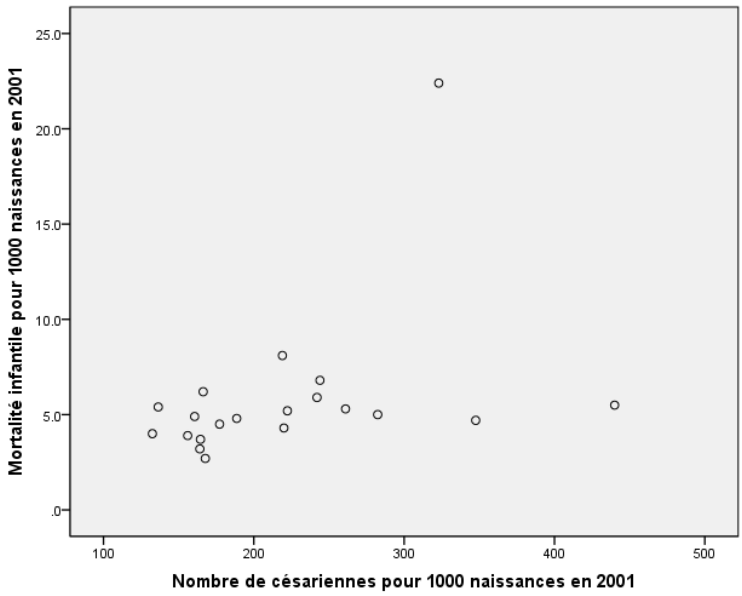
		Valeur	Signification approximée
Nominal par Nominal	Phi	.121	.502
	V de Cramer	.121	.502
Nombre d'observations valides		162	

La p-valeur du test du chi-2 étant supérieure au risque habituel de 5%, on peut admettre que les deux variables sont indépendantes dans la population. Le V de Cramer prend aussi une valeur très faible (0.121), mais cette valeur se rapporte à l'échantillon et non à la population. Ainsi, les données dont nous disposons ne sont pas suffisantes pour prouver qu'il existe, de manière générale, une relation entre les deux variables.

2. Charger le fichier de données « Santé_OCDE.sav » et effectuer les opérations suivantes :

a) Etudier la relation entre les variables *Mort_infantile* et *Césariennes*, tout d'abord globalement, puis séparément pour chacune des 3 régions considérées dans ces données.

On étudie cette relation de deux manières : graphiquement à l'aide d'un diagramme de dispersion (Graphes -> Boîtes de dialogue ancienne version -> Dispersion/Points) et numériquement à l'aide du coefficient de corrélation linéaire de Pearson (Analyse -> Corrélation -> Bivariée).

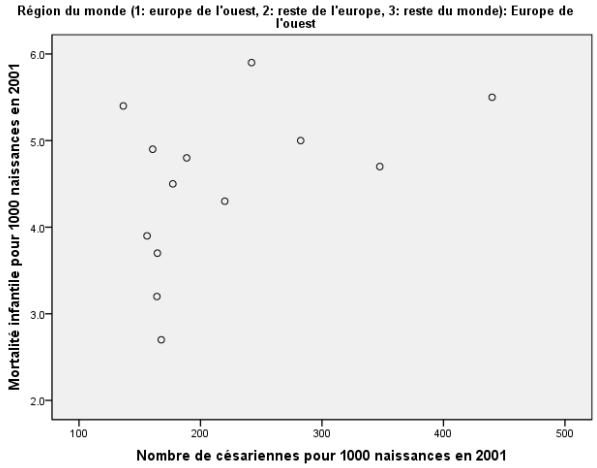


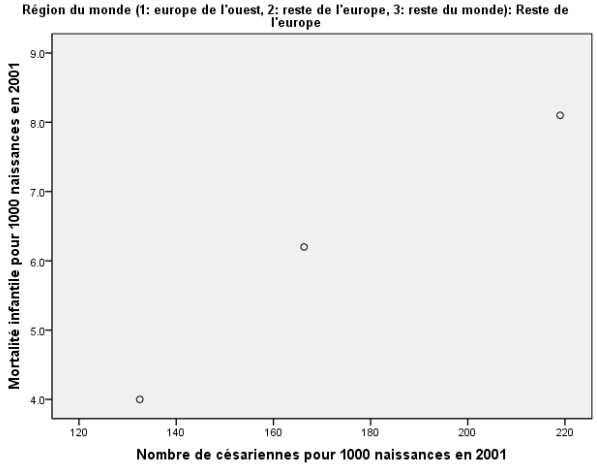
Corrélations

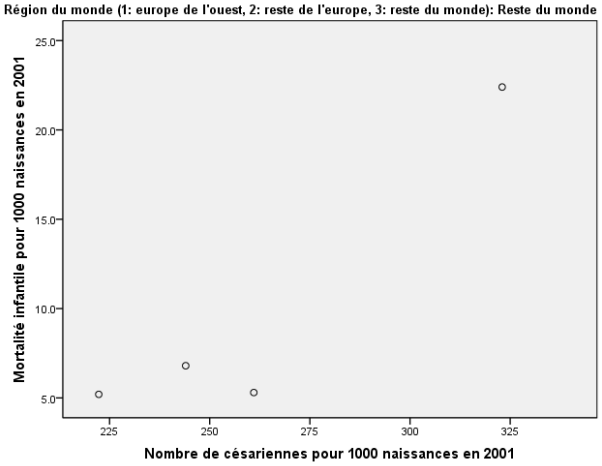
		Mortalité infantile pour 1000 naissances en 2001	Nombre de césariennes pour 1000 naissances en 2001
Mortalité infantile pour 1000 naissances en 2001	Corrélation de Pearson	1	.376
	Sig. (bilatérale)		.102
	N	28	20
Nombre de césariennes pour 1000 naissances en 2001	Corrélation de Pearson	.376	1
	Sig. (bilatérale)	.102	
	N	20	22

Le graphique montre une tendance positive relativement linéaire, ce qui signifie que les pays ayant le plus de césariennes sont aussi ceux ayant les plus forts taux de mortalité infantile. Cependant, l'une des observations s'éloigne totalement de cette tendance. La corrélation est finalement relativement faible (0.376), la cause en étant certainement cette unique observation différentes des autres. Du point de vue inférentiel, il n'est pas possible d'admettre que la corrélation est différente de zéro dans la population (p-valeur = 0.102 > 5%).

Pour obtenir des résultats séparés selon la région, on utilise cette variable pour comparer les groupes (Données -> Scinder un fichier) puis on refait les analyses précédentes.


Région du monde (1: europe de l'ouest, 2: reste de l'europe, 3: reste du monde): Europe de l'ouest
Mortalité infantile pour 1000 naissances en 2001
6.0
5.0
4.0
3.0
2.0
100
200
300
400
500
Nombre de césariennes pour 1000 naissances en 2001


Région du monde (1: europe de l'ouest, 2: reste de l'europe, 3: reste du monde): Reste de l'europe
Mortalité infantile pour 1000 naissances en 2001
9.0
8.0
7.0
6.0
5.0
4.0
120
140
160
180
200
220
Nombre de césariennes pour 1000 naissances en 2001


Région du monde (1: europe de l'ouest, 2: reste de l'europe, 3: reste du monde): Reste du monde
Mortalité infantile pour 1000 naissances en 2001
25.0
20.0
15.0
10.0
5.0
225
250
275
300
325
Nombre de césariennes pour 1000 naissances en 2001

Corrélations

Région du monde (1: europe de l'ouest, 2: reste de l'europe, 3: reste du monde)			Mortalité infantile pour 1000 naissances en 2001	Nombre de césariennes pour 1000 naissances en 2001
Europe de l'ouest	Mortalité infantile pour 1000 naissances en 2001	Corrélation de Pearson	1	.453
		Sig. (bilatérale)		.120
		N	18	13
	Nombre de césariennes pour 1000 naissances en 2001	Corrélation de Pearson	.453	1
		Sig. (bilatérale)	.120	
		N	13	13
Reste de l'europe	Mortalité infantile pour 1000 naissances en 2001	Corrélation de Pearson	1	.986
		Sig. (bilatérale)		.107
		N	5	3
	Nombre de césariennes pour 1000 naissances en 2001	Corrélation de Pearson	.986	1
		Sig. (bilatérale)	.107	
		N	3	3
Reste du monde	Mortalité infantile pour 1000 naissances en 2001	Corrélation de Pearson	1	.931
		Sig. (bilatérale)		.069
		N	5	4
	Nombre de césariennes pour 1000 naissances en 2001	Corrélation de Pearson	.931	1
		Sig. (bilatérale)	.069	
		N	4	6

Les relations sont très différentes d'une région à l'autre (corrélations allant de 0.453 à 0.986), mais étant donné la très faible taille de certains des sous-échantillons (3 pour le reste de l'Europe, …), il n'est pas possible d'en tirer des conclusions fiables. Toutes les p-valeurs sont supérieures à 5% et supportent donc l'hypothèse d'une corrélation nulle au sein de la population.

Après cette analyse, on retourne sous (Données -> Scinder un fichier) pour enlever le split des données selon les régions.

b) A partir de la variable *Région*, créer une nouvelle variable *Région2* répartissant les pays en seulement 2 régions : ceux d'Europe de l'ouest et tous les autres.

On utilise pour cela la fonction RECODE (Transformer -> Création de variables) et l'on recode la valeur 1 de Région en 1 pour Région2, et les valeurs 2 et 3 de Région en 2 pour Région2. On peut aussi donner un nom (label) à la nouvelle variable. Cela correspond au code suivant s'affichant dans la fenêtre de résultats :

```
RECODE Région (1=1) (2 thru 3=2) INTO Région2.
VARIABLE LABELS  Région2 'Région en 2 catégories'.
```

c) Comparer la moyenne des variables *Alcool* et *Tabac* entre les deux catégories de la variable *Région2* au moyen d'un test de Student.

On utilise la fonction T-Test pour données indépendantes (Analyse -> Comparer les moyennes -> Test T pour échantillons indépendants). Les résultats suivants sont obtenus (on a effectué simultanément l'analyse par rapport aux deux variables, même si les résultats concernant une variable sont indépendants de ceux de l'autre variable).

Le premier tableau donne des informations descriptives concernant chaque groupe, et cela pour chacune des deux variables.

Statistiques de groupe

	Région en 2 catégories	N	Moyenne	Ecart-type	Erreur standard moyenne
Consommation annuelle d'alcool en litres par habitant en 2001	1.00	9	9.478	2.9282	.9761
	2.00	7	8.629	3.7818	1.4294
Pourcentage de fumeurs réguliers en 2001	1.00	11	27.055	4.2793	1.2903
	2.00	7	24.571	5.9472	2.2478

Le second tableau (page suivante) contient deux choses. Tout d'abord un test de Levene permettant de savoir si les variances sont similaires dans les deux groupes à comparer. Ici, l'hypothèse nulle d'égalité des variances est acceptée dans les deux cas (p-valeurs > 5%). On regarde ensuite la ligne intitulée « Hypothèse de variances égales « (puisque l'on vient d'accepter une telle hypothèse) et la colonne Sig (bilatérale) donne la p-valeur du test de Student (0.620 pour l'alcool et 0.317 pour le tabac). A nouveau, c'est l'hypothèse nulle d'égalité des moyennes entre les deux groupes qui est acceptée, et cela pour les deux variables considérées.

<u>Remarque finale</u>

Tous les exemples traités dans ce TP sont une illustration du fait que même lorsqu'un effet semble assez important dans un échantillon (par exemple la relation entre la santé et les consultations chez un docteur du point 1) , il n'est pas assuré que cet effet se retrouve aussi au niveau de la population. En effet, seule une taille d'échantillon importante permet d'arriver à une telle conclusion. C'est donc bien cette taille d'échantillon qui prime, et non les valeurs obtenues sur l'échantillon.

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes						
									Intervalle de confiance 95% de la différence	
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Différence écart-type	Inférieure	Supérieure
Consommation annuelle d'alcool en litres par habitant en 2001	Hypothèse de variances égales	.000	.986	.507	14	.620	.8492	1.6736	-2.7404	4.4388
	Hypothèse de variances inégales			.491	11.091	.633	.8492	1.7309	-2.9566	4.6550
Pourcentage de fumeurs réguliers en 2001	Hypothèse de variances égales	1.916	.185	1.033	16	.317	2.4831	2.4034	-2.6118	7.5780
	Hypothèse de variances inégales			.958	9.957	.361	2.4831	2.5918	-3.2952	8.2615