Analyse des données « Hamburgers » à l'aide de SPSS (v2, janvier 2011)
Auteur : André Berchtold

Le site web « The Fast Food Explorer » (www.fatcalories.com) propose des données relatives à la composition des produits vendus dans les fast-foods aux Etats-Unis. Le 3 juin 2008, il a été extrait de ce site des données concernant la composition des 117 types de hamburgers répertoriés. Ces données sont accessibles sur le site web **www.andreberchtold.com**.

Notre objectif est de présenter au travers de ce fichier des exemples d'analyses statistiques réalisées à l'aide de SPSS. Il s'entend que ces analyses ont été choisies dans un but pédagogique et qu'elles ne constituent pas une vraie analyse exhaustive de cette base de données. En ce sens, nous ne donnons qu'une partie des résultats, notamment univariés et bivariés. Par ailleurs, les tableaux et résultats redondants d'une analyse à l'autre ont généralement été supprimés.

Variables

Le fichier « Hamburgers.sav » contient les variables suivantes :

- Nom : Marque et nom du hamburger
- ID : Identificateur numérique (de 1 à 117)
- Marque : Chaîne de fast-food proposant ce hamburger (1 : Wendy's, 2 : McDonald's, 3 : Jack in the Box, 4 : Burger King, 5 : Sonic, 6 : Hardee's, 7: Dairy Queen)
- Calories_totales : Nombre total de calories
- Calories_totales_r2 : Nombre total de calories recodé en 2 catégories (1 : <=620, 2 : >620)
- Calories_totales_r4 : Nombre total de calories recodé en 4 catégories (1 : <=400, 2 : 401-620, 3 : 621-820, 4 : >820)
- Calories_graisses : Nombre de calories des graisses
- Calories_graisses_p : Pourcentage du total des calories dû aux graisses
- Graisses : Quantité de graisses en grammes
- Graisses_r2 : Quantité de graisses recodée en 2 catégories (1 : <=32, 2 : >32)
- Graisses_saturées : Quantité de graisses saturées en grammes
- Cholestérol : Quantité de cholestérol en milligrammes
- Sodium : Quantité de sodium en milligrammes
- Carbone : Quantité de carbone en grammes
- Fibres : Quantité de fibres en grammes
- Protéines : Quantité de protéines en grammes

Toutes ces données ont été extraites du site web, à l'exception des variables Calories_totales_r2, Calories_totales_r4 et Graisses_r2 qui ont été calculées à partir des autres informations à disposition.

Les variables « Nom » et « Marque » sont nominales. La variable « ID » est aussi nominale, car les valeurs de 1 à 117 ont été attribuées de façon arbitraire. La variable « Calories_totales_r4 » est ordinale. Les variables « Calories_totales_r2 » et « Graisses_r2 » sont dichotomiques et peuvent être considérées soit comme nominales, soit comme ordinales. Les autres variables sont numériques (nous ne ferons pas de différence entre variables numériques discrètes et continues).

Echantillon et population

Les 117 observations dont nous disposons sont considérées comme un échantillon de tous les types de hamburgers vendus par les grandes chaînes de fast-food aux Etats-Unis. L'ensemble de tous les types de hamburgers vendus constitue alors la population que nous cherchons à étudier.

Analyse univariée

La première étape de l'analyse consiste à étudier les caractéristiques individuelles de chaque variable. Cela peut être fait en utilisant des tableaux de fréquence, des graphiques et des résumés numériques.

Analyse d'une variable nominale : Marque

« Analyse → Statistiques descriptives → Effectifs »

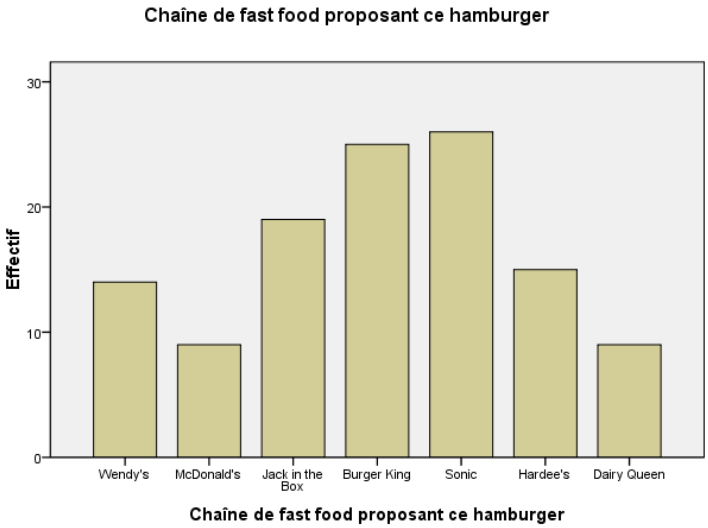
Statistiques

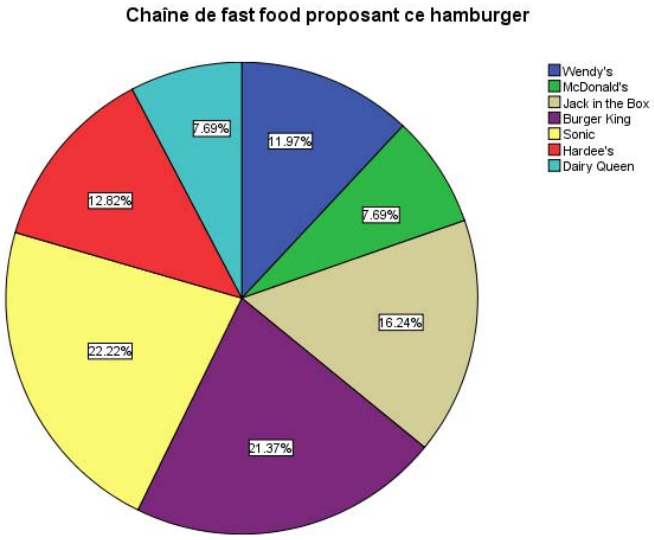
Chaîne de fast food proposant ce hamburger

N	Valide	117
	Manquante	0
	Mode	5

Chaîne de fast food proposant ce hamburger

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Wendy's	14	12.0	12.0	12.0
	McDonald's	9	7.7	7.7	19.7
	Jack in the Box	19	16.2	16.2	35.9
	Burger King	25	21.4	21.4	57.3
	Sonic	26	22.2	22.2	79.5
	Hardee's	15	12.8	12.8	92.3
	Dairy Queen	9	7.7	7.7	100.0
	Total	117	100.0	100.0	





Commentaire :
Les calculs ont été réalisés sur 117 observations valides. Le mode de la distribution (valeur la plus fréquente est 5, ce qui correspond à la chaîne Sonic). Le tableau de fréquence nous apprend qu'il y a effectivement 26 hamburgers de Sonic dans nos données. A l'autre extrême, McDonald's et Dairy Queen n'ont chacun que 9 hamburgers. La colonne « Pourcentage cumulé » n'est pas interprétable ici, car les données sont nominales, leur ordre est arbitraire et il n'est donc pas possible de les cumuler. Les 2 graphiques nous montrent aussi la répartition des 117 hamburgers entre les 7 chaînes de fast-food en faisant apparaître clairement les différences d'une chaîne à l'autre. Au niveau des résumés numériques, le seul pertinent ici est bien le mode, car nous avons affaire à des données nominales, donc non-numériques et non-ordonnées.

Analyse d'une variable ordinale : Calories_totales_r4

« Analyse → Statistiques descriptives → Effectifs »

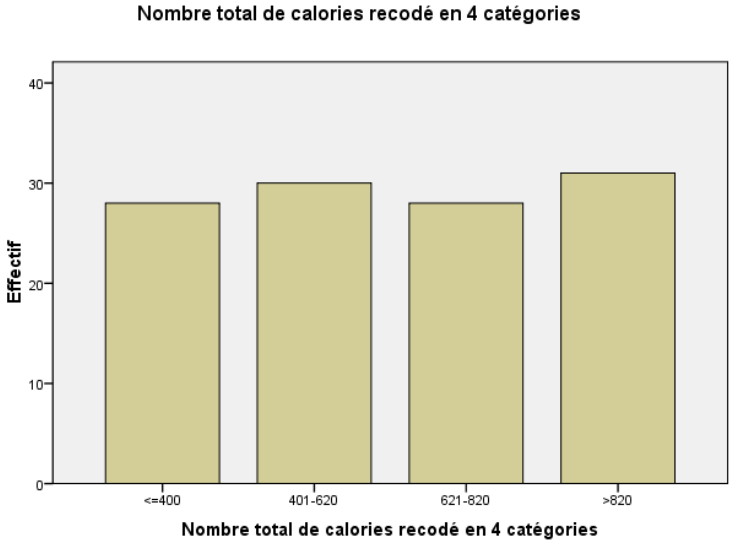
Statistiques

Nombre total de calories recodé en 4 catégories

N	Valide	117
	Manquante	0
	Médiane	3.00
	Mode	4
	Minimum	1
	Maximum	4

Nombre total de calories recodé en 4 catégories

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	<=400	28	23.9	23.9	23.9
	401-620	30	25.6	25.6	49.6
	621-820	28	23.9	23.9	73.5
	>820	31	26.5	26.5	100.0
	Total	117	100.0	100.0	



Commentaire :
Le mode de cette variable est 4 (c'est-à-dire >820), mais sa médiane vaut 3. On peut donc admettre que la moitié de l'échantillon prend le code 3 (621-820) ou inférieur et l'autre moitié le code 3 ou supérieur. Le minimum et le maximum sont aussi donnés, mais les valeurs 1 et 4 ne sont que des codes arbitraires associés aux deux catégories les plus extrêmes. Le tableau de fréquence donne les mêmes informations que pour l'analyse précédente, mais cette fois le pourcentage cumulé est interprétable, car les données sont ordinales (les 4 modalités de la variable ont un ordre précis). Par exemple, 49.6% signifie que 49.6% des données sont dans l'une des deux premières modalités de la variable. Cela nous permet de voir que même si formellement la médiane (pourcentage cumulé = 50%) correspond à la 3ème modalité, elle est en fait très proche de la deuxième. Finalement, le diagramme en bâton montre aussi que les 4 modalités ont des fréquences presque égales.

Analyse d'une variable numérique : Graisses

« Analyse → Statistiques descriptives → Explorer »
(La plupart des résultats peuvent aussi être obtenu à partir de « Descriptives ».)

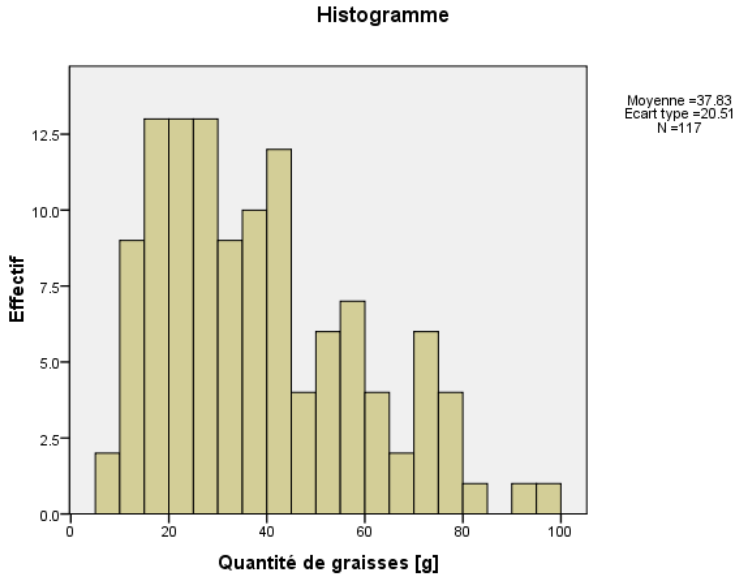
Descriptives

			Statistique	Erreur standard
Quantité de graisses [g]	Moyenne		37.83	1.896
	Intervalle de confiance à 95% pour la moyenne	Borne inférieure	34.07	
		Borne supérieure	41.58	
	Moyenne tronquée à 5%		36.79	
	Médiane		34.00	
	Variance		420.660	
	Ecart-type		20.510	
	Minimum		8	
	Maximum		97	
	Intervalle		89	
	Intervalle interquartile		32	
	Asymétrie		.710	.224
	Aplatissement		-.273	.444

Tests de normalité

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistique	ddl	Signification	Statistique	ddl	Signification
Quantité de graisses [g]	.112	117	.001	.939	117	.000

a. Correction de signification de Lilliefors



Commentaire :
Cette variable étant numérique, on dispose de beaucoup plus de résumés numériques. La moyenne vaut 37.83. C'est la quantité de graisses par hamburger de l'échantillon en cas de répartition égalitaire de toute la graisse des 117 hamburgers. L'intervalle de confiance à 95% va de 34.07 à 41.58. C'est la zone de valeurs dans laquelle il est probable à 95% d'observer la moyenne des graisses pour l'ensemble de la population de tous les hamburgers (et non les seuls 117 de l'échantillon). La médiane (34) est un peu inférieure à la moyenne, ce qui indique que la distribution n'est pas parfaitement symétrique. Cela peut s'observer facilement sur l'histogramme. La variance et l'écart-type (racine carrée de la variance) sont des mesures de la dispersion (étalement) des données. Plus la variance est grande, plus les données sont dispersées. Deux tests de normalité ont été effectués afin de vérifier si l'on peut admettre qu'au niveau de la population, la distribution de la variable est similaire à une loi normale. Pour les deux tests, la p-valeur (Signification) est nettement inférieure à 5% et l'hypothèse nulle de normalité est donc rejetée. Etant donné que la variable prend un grand nombre de valeurs différentes, il n'est pas indiqué de calculer un tableau de fréquences.

<u>Analyse d'une variable numérique : Graisses décomposée en 2 catégories en fonction de Calories_totales_r2</u>

« Analyse → Statistiques descriptives → Explorer »

Récapitulatif du traitement des observations

	Nombre total de calories recodé en 2 catégories	Observations					
		Valide		Manquante		Total	
		N	Pourcent	N	Pourcent	N	Pourcent
Quantité de graisses [g]	<=620	58	100.0%	0	.0%	58	100.0%
	>620	59	100.0%	0	.0%	59	100.0%

Descriptives

	Nombre total de calories recodé en 2 catégories			Statistique	Erreur standard
Quantité de graisses [g]	<=620	Moyenne		21.40	.960
		Intervalle de confiance à 95% pour la moyenne	Borne inférieure	19.47	
			Borne supérieure	23.32	
		Moyenne tronquée à 5%		21.24	
		Médiane		20.50	
		Variance		53.436	
		Ecart-type		7.310	
		Minimum		8	
		Maximum		39	
		Intervalle		31	
		Intervalle interquartile		11	
		Asymétrie		.255	.314
		Aplatissement		-.633	.618
	>620	Moyenne		53.98	2.072
		Intervalle de confiance à 95% pour la moyenne	Borne inférieure	49.84	
			Borne supérieure	58.13	
		Moyenne tronquée à 5%		53.23	
		Médiane		53.00	
		Variance		253.327	
		Ecart-type		15.916	
		Minimum		31	
		Maximum		97	
		Intervalle		66	
		Intervalle interquartile		25	
		Asymétrie		.617	.311
		Aplatissement		-.332	.613

Commentaire :
Nous avons repris la variable Graisses, mais nous l'avons découpée en deux groupes en fonction du facteur « Calories_totales_r2 » et nous disposons d'informations individuelles pour les hamburgers appartenant à chacun des deux groupes. Nous pouvons notamment constater que la moyenne de graisses est nettement inférieure dans le premier groupe (peu de calories) que dans le second (21.40 contre 53.98). Les observations sont aussi moins dispersées dans le premier groupe que dans le second (écart-type de 7.31 contre 15.916).

Analyse bivariée

La deuxième étape de l'analyse consiste à mettre en relation les variables par paires. Cela peut se faire à l'aide de tables de contingence et de résumés numériques dans le cas de variables catégorielles, et de graphiques de dispersion et de résumés numériques dans le cas de variables numériques.

Marque et Calories_totales_r2 (2 variables catégorielles)
« Analyse → Statistiques descriptives → Tableaux croisés »

Tableau croisé Chaîne de fast food proposant ce hamburger * Nombre total de calories recodé en 2 catégories

			Nombre total de calories recodé en 2 catégories		
			<=620	>620	Total
Chaîne de fast food proposant ce hamburger	Wendy's	Effectif	10	4	14
		% dans Chaîne de fast food proposant ce hamburger	71.4%	28.6%	100.0%
	McDonald's	Effectif	8	1	9
		% dans Chaîne de fast food proposant ce hamburger	88.9%	11.1%	100.0%
	Jack in the Box	Effectif	8	11	19
		% dans Chaîne de fast food proposant ce hamburger	42.1%	57.9%	100.0%
	Burger King	Effectif	11	14	25
		% dans Chaîne de fast food proposant ce hamburger	44.0%	56.0%	100.0%
	Sonic	Effectif	11	15	26
		% dans Chaîne de fast food proposant ce hamburger	42.3%	57.7%	100.0%
	Hardee's	Effectif	7	8	15
		% dans Chaîne de fast food proposant ce hamburger	46.7%	53.3%	100.0%
	Dairy Queen	Effectif	3	6	9
		% dans Chaîne de fast food proposant ce hamburger	33.3%	66.7%	100.0%
	Total	Effectif	58	59	117
		% dans Chaîne de fast food proposant ce hamburger	49.6%	50.4%	100.0%

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	10.524[a]	6	.104
Rapport de vraisemblance	11.386	6	.077
Association linéaire par linéaire	5.510	1	.019
Nombre d'observations valides	117		

a. 4 cellules (28.6%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 4.46.

Mesures symétriques

		Valeur	Signification approximée
Nominal par Nominal	Phi	.300	.104
	V de Cramer	.300	.104
	Nombre d'observations valides	117	

Commentaire :
La table de contingence met en évidence la répartition des hamburgers au sein de chaque chaîne entre « basses calories » (<=620) et « hautes calories » (>620). Par exemple, pour Wendy's, 14 hamburgers ont été analysés, dont 10 (71.4%) sont considérés comme « basses calories ». Le khi-deux (chi-2) permet de tester s'il existe un lien significatif entre les deux variables. Au niveau de l'échantillon, le khi-deux vaut 10.524. Lorsque l'on teste l'hypothèse d'indépendance entre les deux variables, on obtient une p-valeur (significativité) de 0.104 (10.4%). L'hypothèse d'indépendance est donc acceptée et il n'y a pas de relation dans la population entre les deux variables. La relation observée au niveau de l'échantillon relève donc du hasard. Finalement, le v de Cramer est une transformation du chi-deux destinée à le rendre plus facile à interpréter. Ici, on a V=0.3, le minimum (indépendance) étant 0 et le maximum 1. Il y a donc, au niveau de l'échantillon, une relation faible entre les deux variables. La p-valeur donnée à côté est la même que celle associée au khi-deux.

Graisses_r2 et Calories_totales_r2 (2 variables catégorielles dichotomiques)
« Analyse → Statistiques descriptives → Tableaux croisés »

Tableau croisé Quantité de graisses recodée en 2 catégories * Nombre total de calories recodé en 2 catégories

			Nombre total de calories recodé en 2 catégories		
			<=620	>620	Total
Quantité de graisses recodée en 2 catégories	<=32	Effectif	54	3	57
		% dans Quantité de graisses recodée en 2 catégories	94.7%	5.3%	100.0%
	>32	Effectif	4	56	60
		% dans Quantité de graisses recodée en 2 catégories	6.7%	93.3%	100.0%
	Total	Effectif	58	59	117
		% dans Quantité de graisses recodée en 2 catégories	49.6%	50.4%	100.0%

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	90.696[a]	1	.000		
Correction pour la continuité[b]	87.207	1	.000		
Rapport de vraisemblance	109.290	1	.000		
Test exact de Fisher				.000	.000
Association linéaire par linéaire	89.921	1	.000		
Nombre d'observations valides	117				

a. 0 cellules (.0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 28.26.

b. Calculé uniquement pour un tableau 2x2

Mesures symétriques

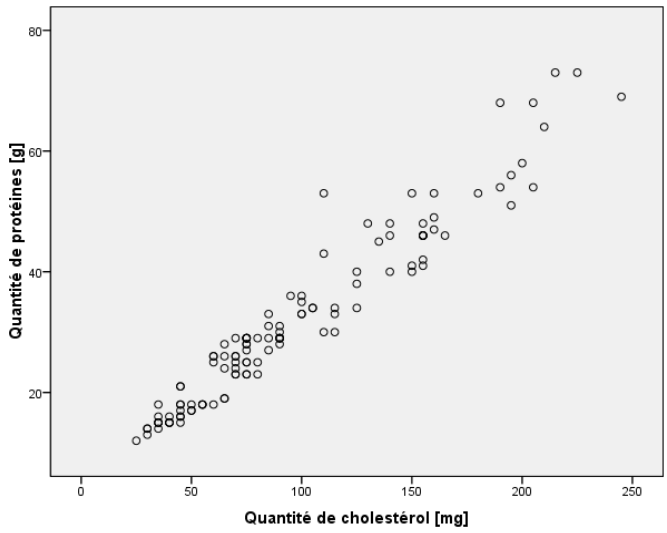
		Valeur	Signification approximée
Nominal par Nominal	Phi	.880	.000
	V de Cramer	.880	.000
	Nombre d'observations valides	117	

Estimation du risque

	Valeur	Intervalle de confiance de 95%	
		Inférieur	Supérieur
Odds Ratio pour Quantité de graisses recodée en 2 catégories (<=32 / >32)	252.000	53.867	1178.908
Pour cohorte Nombre total de calories recodé en 2 catégories = <=620	14.211	5.503	36.698
Pour cohorte Nombre total de calories recodé en 2 catégories = >620	.056	.019	.170
Nombre d'observations valides	117		

Commentaire :
La principale différence entre cette analyse et la précédente réside dans le fait que les deux variables sont ici dichotomiques, il est donc en plus possible de calculer un odds ratio. Il vaut ici 252 (avec un intervalle de confiance pour la population allant de 53.867 à 1178.908), ce qui signifie qu'il est 252 fois plus probable d'être dans la catégorie « hautes calories » si l'on a >32 graisses plutôt que <=32. Pour le reste, le khi-deux et le V de Cramer montrent une forte association entre les deux variables au niveau de l'échantillon, et le test montre que cette association se retrouve aussi au niveau de la population (p-valeur = 0.000).

Cholestérol et Protéines (2 variables numériques)
Graphique : « Graphes → Dispersion/Points → Dispersion simple »
Corrélation : « Analyse → Corrélation → Bivariée »



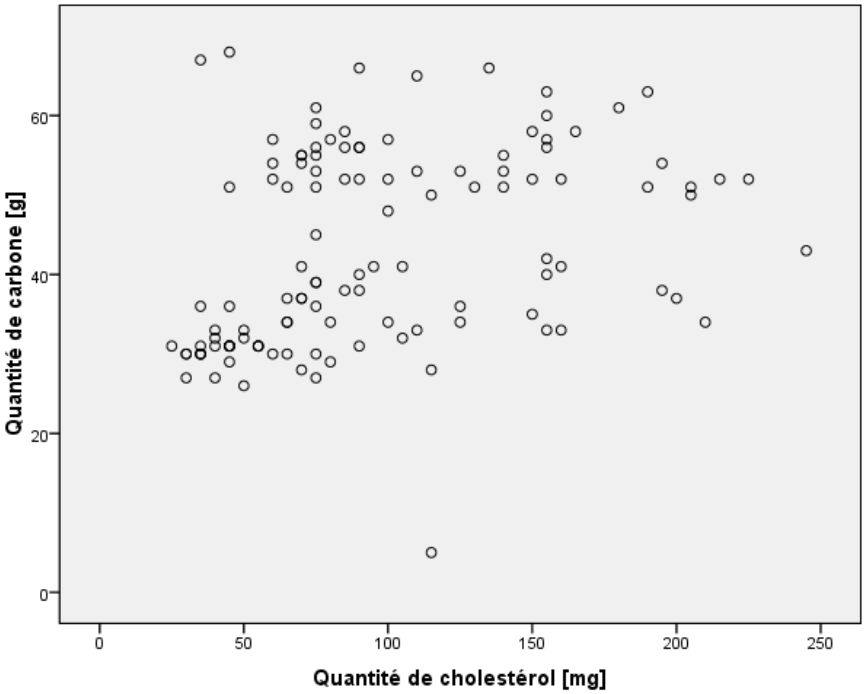
Corrélations

		Quantité de cholestérol [mg]	Quantité de protéines [g]
Quantité de cholestérol [mg]	Corrélation de Pearson	1.000	.966**
	Sig. (bilatérale)		.000
	N	117	117
Quantité de protéines [g]	Corrélation de Pearson	.966**	1.000
	Sig. (bilatérale)	.000	
	N	117	117

**. La corrélation est significative au niveau 0.01 (bilatéral).

Commentaire :
Le graphique montre une relation linéaire positive entre les deux variables : Plus il y a de cholestérol dans un hamburger, plus il y a de protéines. Le coefficient de corrélation de Pearson vaut 0.966, ce qui montre bien une relation forte, mais pas tout-à-fait parfaite. Cette valeur est significativement différente de zéro (p-valeur=0.000), donc cette relation se retrouve au niveau de la population.

Cholestérol et Carbone (2 variables numériques)
Graphique : « Graphes → Dispersion/Points → Dispersion simple »
Corrélation : « Analyse → Corrélation → Bivariée »



Corrélations

		Quantité de cholestérol [mg]	Quantité de carbone [g]
Quantité de cholestérol [mg]	Corrélation de Pearson	1.000	.342**
	Sig. (bilatérale)		.000
	N	117	117
Quantité de carbone [g]	Corrélation de Pearson	.342**	1.000
	Sig. (bilatérale)	.000	
	N	117	117

**. La corrélation est significative au niveau 0.01 (bilatéral).

Commentaire :
Dans ce cas, la relation est toujours positive, mais beaucoup moins évidente que dans l'analyse précédente. La corrélation ne vaut que 0.342, mais cette valeur est encore fortement significative au niveau de la population.

Tests statistiques

Certains tests statistiques, permettant de vérifier si ce qui a été trouvé au niveau de l'échantillon est également vrai pour l'ensemble de la population, ont déjà été utilisés précédemment (normalité, chi-2, corrélation). Nous présentons maintenant des exemples de plusieurs autres tests courants.

Test de Student pour 2 populations appariées : Protéines et Carbone

« Analyse → Comparer les moyennes → Test T pour échantillons appariés »

Statistiques pour échantillons appariés

		Moyenne	N	Ecart-type	Erreur standard moyenne
Paire 1	Quantité de carbone [g]	43.26	117	12.376	1.144
	Quantité de protéines [g]	32.10	117	14.618	1.351

Test échantillons appariés

		Différences appariées					t	ddl	Sig. (bilatérale)
		Moyenne	Ecart-type	Erreur standard moyenne	Intervalle de confiance 95% de la différence				
					Inférieure	Supérieure			
Paire 1	Quantité de carbone [g] - Quantité de protéines [g]	11.154	14.498	1.340	8.499	13.809	8.322	116	.000

Commentaire :
Nous comparons la quantité de carbone et de protéine dans chaque hamburger (données appariées). Le premier tableau donne les informations de base sur les deux variables. Le second tableau donne le test de Student avec la différence des moyennes entre les deux échantillons (11.154) et surtout la p-valeur du test (0.000). Cette p-valeur étant inférieure à 5%, l'hypothèse nulle d'égalité des moyennes est rejetée. Il y a donc bien une moyenne de carbone supérieure à celle des protéines et cela est vrai non seulement pour notre échantillon, mais aussi dans la population.

Test de Student pour 2 populations indépendantes : Protéines en fonction de Calories_totales_r2

« Analyse → Comparer les moyennes → Test T pour échantillons indépendants »

Statistiques de groupe

	Nombre total de calories recodé en 2 catégories	N	Moyenne	Ecart-type	Erreur standard moyenne
Quantité de protéines [g]	<=620	58	21.40	5.982	.786
	>620	59	42.63	12.841	1.672

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
Quantité de protéines [g]	Hypothèse de variances égales	26.937	.000	-11.430	115	.000	-21.231	1.857	-24.910	-17.551
	Hypothèse de variances inégales			-11.494	82.353	.000	-21.231	1.847	-24.905	-17.556

Commentaire :

Ici, nous ne considérons que les protéines, mais nous divisons notre échantillon en deux parties : les hamburgers avec 620 calories ou moins et les autres, et nous comparons la moyenne des protéines entre les deux groupes. Tout d'abord, le test de Levene nous indique que les variances sont différentes dans les deux groupes (p-valeur = 0.000). Il faut alors lire la dernière ligne du tableau (Hypothèse de variances inégales), et nous voyons que le test de Student a aussi une p-valeur de 0.000, ce qui indique le rejet de l'hypothèse d'égalité des moyennes entre les deux groupes.

Test de normalité de Kolmogorov-Smirnov : 5 variables numériques
« Analyse → Tests non paramétriques → K-S à 1 échantillon »

Test de Kolmogorov-Smirnov à un échantillon

		Nombre total de calories	Nombre de calories des graisses	Quantité de carbone [g]	Quantité de fibres [g]	Quantité de protéines [g]
N		117	117	117	117	117
Paramètres normaux[a]	Moyenne	643.93	340.92	43.26	2.72	32.10
	Ecart-type	265.307	184.235	12.376	1.382	14.618
Différences les plus extrêmes	Absolue	.104	.117	.153	.171	.148
	Positive	.104	.117	.132	.171	.148
	Négative	-.062	-.078	-.153	-.130	-.091
Z de Kolmogorov-Smirnov		1.122	1.263	1.655	1.853	1.602
Signification asymptotique (bilatérale)		.161	.082	.008	.002	.012

a. La distribution à tester est gaussienne.

Commentaire :
Le test de normalité de Kolmogorov-Smirnov a été effectué pour 5 variables numériques. L'hypothèse nulle dit que la variable est normale. Dans deux cas (calories et calories des graisses), on accepte cette hypothèse nulle, car les p-valeurs (0.161 et 0.082) sont > 5%. Pour les 3 autres variables, on rejette l'hypothèse de normalité (p-valeurs inférieures à 5%). Le rejet de la normalité implique que les tests paramétriques habituels comme le test de Student, sont moins fiables et qu'il est préférable d'utiliser des tests non-paramétriques.

Test de Wilcoxon pour 2 populations appariées : Protéines et Carbone
« Analyse → Tests non paramétriques → 2 échantillons liés »

Test[b]

	Quantité de protéines [g] - Quantité de carbone [g]
Z	-6.548[a]
Signification asymptotique (bilatérale)	.000

a. Basée sur les rangs positifs.

b. Test de Wilcoxon

Commentaire :
Comme les variables protéines et carbone ne sont pas normales (cf. test de Kolmogorov-Smirnov), il est préférable d'utiliser un test non-paramétrique. Le test de Wilcoxon est l'équivalent non-paramétrique du test de Student pour données appariées. Dans un test paramétrique, on vérifie l'égalité des médianes et non celle des moyennes. Ici, la p-valeur du test est très faible (0.000) et l'on rejette donc l'égalité des médianes entre les quantités de protéines et de carbone dans les hamburgers.

Test de Mann-Whitney pour 2 populations indépendantes : Protéines en fonction de Calories_totales_r2
« Analyse → Tests non paramétriques → 2 échantillons indépendants »

Test[a]

	Quantité de protéines [g]
U de Mann-Whitney	153.000
W de Wilcoxon	1864.000
Z	-8.503
Signification asymptotique (bilatérale)	.000

a. Critère de regroupement : Nombre total de calories recodé en 2 catégories

Commentaire :
Le tes de Mann-Whitney est l'équivalent non-paramétrique du test de Student pour données non-appariées. Le W de Wilcoxon est un autre test qui donne strictement le même résultat. Ici, la p-valeur (0.000) nous amène à rejetter l'hypohèse d'égalité des médianes de protéines entre les deux groupes (basses calories et hautes calories).

Modèles statistiques

Nous essayons maintenant de mettre en relation de façon plus générale différentes variables de notre base de données.

Analyse de variance (ANOVA) : Fibres en fonction Calories_totales_r4
« Analyse → Comparer les moyennes → ANOVA à 1 facteur »

Descriptives

Quantité de fibres [g]

	N	Moyenne	Ecart-type	Erreur standard	Intervalle de confiance à 95% pour la moyenne		Minimum	Maximum
					Borne inférieure	Borne supérieure		
<=400	28	1.57	.742	.140	1.28	1.86	1	3
401-620	30	2.77	1.331	.243	2.27	3.26	1	5
621-820	28	3.32	1.541	.291	2.72	3.92	1	5
>820	31	3.16	1.128	.203	2.75	3.58	1	5
Total	117	2.72	1.382	.128	2.46	2.97	1	5

Test d'homogénéité des variances

Quantité de fibres [g]

Statistique de Levene	ddl1	ddl2	Signification
6.255	3	113	.001

ANOVA

Quantité de fibres [g]

	Somme des carrés	ddl	Moyenne des carrés	F	Signification
Inter-groupes	53.168	3	17.723	11.883	.000
Intra-groupes	168.525	113	1.491		
Total	221.692	116			

Comparaisons multiples

Quantité de fibres [g]

Test de Tukey

(I) Nombre total de calories recodé en 4 catégories	(J) Nombre total de calories recodé en 4 catégories	Différence de moyennes (I-J)	Erreur standard	Signification	Intervalle de confiance à 95%	
					Borne inférieure	Borne supérieure
<=400	401-620	-1.195*	.321	.002	-2.03	-.36
	621-820	-1.750*	.326	.000	-2.60	-.90
	>820	-1.590*	.318	.000	-2.42	-.76
401-620	<=400	1.195*	.321	.002	.36	2.03
	621-820	-.555	.321	.314	-1.39	.28
	>820	-.395	.313	.589	-1.21	.42
621-820	<=400	1.750*	.326	.000	.90	2.60
	401-620	.555	.321	.314	-.28	1.39
	>820	.160	.318	.958	-.67	.99
>820	<=400	1.590*	.318	.000	.76	2.42
	401-620	.395	.313	.589	-.42	1.21
	621-820	-.160	.318	.958	-.99	.67

*. La différence moyenne est significative au niveau 0.05.

Commentaire :

L'analyse de variance (ANOVA) est une généralisation du test de Student au cas où l'on veut comparer plus de deux groupes. Nous comparons ici la quantité de fibres au sein des hamburgers de chacune des 4 groupes définis par la variable Calories_totales_r4. Le premier tableau donne des informations sur la distribution de la variable Fibres dans chacun des 4 groupes. Le deuxième tableau donne le résultat du test d'égalité des variances de Levene. Ici, l'hypothèse nulle du test (égalité des variances des 4 groupes) est rejetée (p-valeur = 0.001), ce qui implique que les résultats de l'ANOVA peuvent perdre un peu en fiabilité. Le troisième tableau donne le résultat de l'ANOVA elle-même. L'hypothèse nulle (égalité de la moyenne des fibres au sein de chacun des 4 groupes) est nettement rejetée (p=0.000). On peut donc admettre qu'au moins 1 des 4 groupes à une moyenne de fibres différente de celle d'un autre groupe. Pour affiner ce résultat, des tests post-hoc de Tukey ont été effectués pour comparer 2 à 2 toutes les paires de groupes. L'analyse des p-valeurs (signification) montre que le groupe <=400 a une moyenne de fibres différente de celle de chacun des 3 autres groupes. En revanche, il n'y a pas de différence significative entre les moyennes au sein des 3 autres groupes (401-620, 521-820 et >820).

Régression linéaire : Variable dépendante = Calories_totales
« Analyse → Régression → Linéaire »

Récapitulatif du modèle

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	.999[a]	.998	.997	13.363

a. Valeurs prédites : (constantes), Quantité de protéines [g], Quantité de fibres [g], Quantité de sodium [mg], Quantité de carbone [g], Quantité de graisses [g], Quantité de cholestérol [mg]

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Signification	Intervalle de confiance à 95% de B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(constante)	1.535	5.394		.285	.776	-9.154	12.225
	Quantité de graisses [g]	8.697	.153	.672	56.811	.000	8.393	9.000
	Quantité de cholestérol [mg]	.686	.129	.135	5.299	.000	.429	.942
	Quantité de sodium [mg]	-.001	.007	-.002	-.165	.869	-.015	.012
	Quantité de carbone [g]	4.599	.220	.215	20.914	.000	4.163	5.034
	Quantité de fibres [g]	-3.340	1.557	-.017	-2.146	.034	-6.425	-.255
	Quantité de protéines [g]	1.808	.407	.100	4.438	.000	1.000	2.615

a. Variable dépendante : Nombre total de calories

Commentaire :

La régression linéaire est utilisée pour expliquer le comportement d'une variable dépendante numérique (ici le nombre total de calories) à l'aide d'une ou plusieurs autres variables explicatives (ici : graisses, cholestérol, sodium, carbone, fibres, protéines). Globalement, le modèle obtenu est de très bonne qualité, puisque le R2 ajusté vaut 0.997, ce qui est proche de son maximum possible de 1. On l'interprète en disant que le 99.7% de l'information de la variable dépendante est expliquée à l'aide de la régression. Le second tableau nous indique (colonne B) la valeur des coefficients liant chaque variable explicative à la variable dépendante. Par exemple celui du carbone vau 4.599, ce qui signifie que pour 1 gramme de carbone en plus dans un hamburger, on s'attend à y trouver 4.599 calories en plus. La colonne signification donne la p-valeur du test de significativité individuel de chaque paramètre. L'hypothèse nulle dit que le paramètre vaut zéro au niveau de la population. Si on l'accepte, alors la variable correspondante est inutile et elle peut être supprimée du modèle. Ici, si on excepte la constante qui n'est qu'un paramètre d'échelle et qui n'a pas besoin d'être vraiment analysée, on voit que la seule variable non-significative qui pourrait être supprimée du modèle actuel est le sodium (p-valeur=0.869 > 5%). Toutes les autres variables semblent utiles pour comprendre le comportement de la variable dépendante.

Régression logistique : Variable dépendante = Calories_totales_r2
« Analyse → Régression → Logistique binaire »

Codage de variables dépendantes

Valeur d'origine	Valeur interne
<=620	0
>620	1

Bloc 0 : bloc de départ

Tableau de classement[a,b]

			Prévu		
			Nombre total de calories recodé en 2 catégories		
	Observé		<=620	>620	Pourcentage correct
Etape 0	Nombre total de calories recodé en 2 catégories	<=620	0	58	.0
		>620	0	59	100.0
		Pourcentage global			50.4

a. La constante est incluse dans le modèle.

b. La valeur de césure est .500

Variables dans l'équation

		B	E.S.	Wald	ddl	Signif.	Exp(B)
Etape 0	Constante	.017	.185	.009	1	.926	1.017

Variables hors de l'équation

			Score	ddl	Signif.
Etape 0	Variables	Cholestérol	58.879	1	.000
		Sodium	62.580	1	.000
		Carbone	39.879	1	.000
		Protéines	62.228	1	.000
		Statistiques globales	77.000	4	.000

Block 1 : Méthode = Entrée

Récapitulatif du modèle

Etape	-2log-vraisemblance	R-deux de Cox & Snell	R-deux de Nagelkerke
1	38.761[a]	.652	.869

a. L'estimation a été interrompue au numéro d'itération 9 parce que les estimations de paramètres ont changé de moins de .001.

Tableau de classement[a]

Observé			Prévu		
			Nombre total de calories recodé en 2 catégories		
			<=620	>620	Pourcentage correct
Etape 1	Nombre total de calories recodé en 2 catégories	<=620	56	2	96.6
		>620	4	55	93.2
		Pourcentage global			94.9

a. La valeur de césure est .500

Variables dans l'équation

		B	E.S.	Wald	ddl	Signif.	Exp(B)	IC pour Exp(B) 95.0%	
								Inférieur	Supérieur
Etape 1	Cholestérol	.046	.052	.788	1	.375	1.047	.946	1.160
	Sodium	.004	.003	1.758	1	.185	1.004	.998	1.009
	Carbone	.172	.058	8.903	1	.003	1.187	1.061	1.329
	Protéines	.186	.212	.772	1	.380	1.205	.795	1.826
	Constante	-21.103	5.460	14.941	1	.000	.000		

Commentaire :
La régression logistique binaire est l'équivalent de la régression linéaire, mais pour une variable dépendante catégorielles ne prenant que deux valeurs. Ici, nous cherchons à expliquer la variable Calories_totales_r2. Formellement, le modèle va chercher à prédire la probabilité qu'un hamburger appartienne à la catégorie >620 calories. C'est la catégorie codée 1 par SPSS (cf. premier tableau).

Le « Bloc 0 » donne des informations sur un modèle ne comportant que la constante et aucune variable explicative. C'est une sorte de point de référence. Le tableau de classement permet de voir comment le modèle reclasse les observations de la variable dépendante entre ses deux catégories. Sans variable explicative, on n'arrive à classer correctement que 50.4% des observations, ce qui est très faible. Les deux tableaux suivants donnent des informations sur respectivement le modèle (composé de la seule constante) et les variables explicative, pas encore dans le modèle mais qui y seront introduites à l'étape suivante.

Le « Bloc 1 » donne les informations relatives au modèle calculé avec 4 variables explicatives (cholestérol, sodium, carbone, protéines). Globalement, le modèle est bon, mais pas parfait. Le R2 de Nagelkerke est élevé (0.869 pour un maximum de 1) et le modèle permet maintenant d'identifier correctement la catégorie de 94.9% de tous les hamburgers de l'échantillon. Seuls 2 hamburgers ayant réellement moins de 620 calories ont été classifiés comme en ayant plus de 620, et 4 hamburgers sont dans le cas inverse. Au niveau des 4 variables explicatives du modèle, on constate que 3 d'entre-elles ne sont pas significatives (p-valeur supérieure à 5%) et qu'elles pourraient être sorties du modèle. La seule variable significative ici (en plus de la constante) est le carbone. La colonne « Exp(B) » donne l'odds ratio correspondant à chaque variable du modèle.
Pour essayer de simplifier le modèle, nous avons appliqué une procédure de sélection automatique des variables de type « descendante » permettant d'éliminer une à une les variables inutiles :

Récapitulatif du modèle

Etape	-2log-vraisemblance	R-deux de Cox & Snell	R-deux de Nagelkerke
1	38.761[a]	.652	.869
2	39.587[b]	.649	.866
3	42.336[b]	.641	.855

Tableau de classement[a]

			Prévu		
			Nombre total de calories recodé en 2 catégories		
	Observé		<=620	>620	Pourcentage correct
Etape 1	Nombre total de calories recodé en 2 catégories	<=620	56	2	96.6
		>620	4	55	93.2
		Pourcentage global			94.9
Etape 2	Nombre total de calories recodé en 2 catégories	<=620	54	4	93.1
		>620	5	54	91.5
		Pourcentage global			92.3
Etape 3	Nombre total de calories recodé en 2 catégories	<=620	53	5	91.4
		>620	5	54	91.5
		Pourcentage global			91.5

a. La valeur de césure est .500

Variables dans l'équation

		B	E.S.	Wald	ddl	Signif.	Exp(B)	IC pour Exp(B) 95.0%	
								Inférieur	Supérieur
Etape 1[a]	Cholestérol	.046	.052	.788	1	.375	1.047	.946	1.160
	Sodium	.004	.003	1.758	1	.185	1.004	.998	1.009
	Carbone	.172	.058	8.903	1	.003	1.187	1.061	1.329
	Protéines	.186	.212	.772	1	.380	1.205	.795	1.826
	Constante	-21.103	5.460	14.941	1	.000	.000		
Etape 2[a]	Cholestérol	.081	.035	5.355	1	.021	1.085	1.012	1.162
	Sodium	.004	.003	2.483	1	.115	1.004	.999	1.009
	Carbone	.183	.054	11.388	1	.001	1.201	1.080	1.335
	Constante	-19.658	4.770	16.985	1	.000	.000		
Etape 3[a]	Cholestérol	.119	.031	14.650	1	.000	1.126	1.060	1.197
	Carbone	.200	.055	13.478	1	.000	1.222	1.098	1.360
	Constante	-19.229	4.813	15.963	1	.000	.000		

a. Variable(s) entrées à l'étape 1 : Cholestérol, Sodium, Carbone, Protéines.

Commentaire :
A partir du modèle initial (Etape 1), il y a eu deux étapes supplémentaires (Etapes 2 et 3). Comme on peut le voir dans le dernier tableau, la variable Protéines, qui était la variable la moins significative (plus forte p-valeur) lors de l'étape 1, est sortie du modèle à l'Etape 2, et la variable Sodium (la moins significative à l'étape 2) est sortie à l'Etape 3. On peut noter que le cholestérol, qui n'était pas significatif à l'étape 1, l'st devenu à l'étape 2 suite à la suppression des protéines. Le modèle final ne comporte donc plus que deux variables explicatives : cholestérol et carbone. Globalement, le fait d'avoir supprimé une partie des variables explicatives a légèrement péjoré le modèle, puisque le R2 de Nagelkerke est passé de 0.869 à 0.855 et le taux de classement correct est descendu de 94.1% à 91.5%. Cependant, comme le modèle a gagné en simplicité, on peut considérer que le résultat final est tout-à-fait acceptable.

Analyse en composantes principales (ACP) : Toutes les variables numériques
« Analyse → Factorisation → Analyse factorielle »

Qualité de représentation

	Initial	Extraction
Nombre total de calories	1.000	.986
Nombre de calories des graisses	1.000	.966
Pourcentage du total des calories du aux graisses	1.000	.741
Quantité de graisses [g]	1.000	.966
Quantité de cholestérol [mg]	1.000	.884
Quantité de sodium [mg]	1.000	.825
Quantité de carbone [g]	1.000	.918
Quantité de fibres [g]	1.000	.899
Quantité de protéines [g]	1.000	.842

Méthode d'extraction : Analyse en composantes principales.

Variance totale expliquée

Composante	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus		
	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	6.528	72.538	72.538	6.528	72.538	72.538
2	1.498	16.650	89.188	1.498	16.650	89.188
3	.468	5.199	94.386			
4	.252	2.804	97.191			
5	.179	1.993	99.184			
6	.054	.600	99.784			
7	.018	.195	99.979			
8	.002	.019	99.997			
9	.000	.003	100.000			

Méthode d'extraction : Analyse en composantes principales.

Matrice des composantes[a]

	Composante	
	1	2
Nombre total de calories	.992	.051
Nombre de calories des graisses	.980	-.070
Pourcentage du total des calories du aux graisses	.819	-.265
Quantité de graisses [g]	.981	-.067
Quantité de cholestérol [mg]	.905	-.253
Quantité de sodium [mg]	.901	-.116
Quantité de carbone [g]	.600	.747
Quantité de fibres [g]	.378	.870
Quantité de protéines [g]	.904	-.157

Méthode d'extraction : Analyse en composantes principales.

a. 2 composantes extraites.

Commentaire :
L'analyse en composantes principales a pour objectif de réduire le nombre de variables numériques à analyser en regroupant leurs informations sur un nombre strictement inférieur de nouvelles variables appelées composantes principales. Dans cet exemple, nous avons introduit 9 variables dans l'analyse. Le premier tableau donne dans la colonne Extraction, le pourcentage de l'information de chacune des 9 variables qui st reproduit par le modèle. Globalement, c'est très bon, puisqu'une seule variable est reproduite à moins de 80%. Le second tableau donne deux informations importantes : 1) Seules deux composantes ont été utilisées dans le modèle (on le sait, car seules les deux premières lignes sont remplies dans la partie de droite du tableau). On a donc remplacé un ensemble de 9 variables par seulement deux nouvelles variables, ce qui est un gain considérable. 2) A elles-deux, ces composantes reproduisent 89.188% de l'information de départ contenue dans les 9 variables originales. La première composante représente 72.438% de l'information et la seconde 16.650%. Le dernier tableau donne la corrélation entre chacune des 9 variables d'origine et chacune des deux composantes principales. En général, on associe chaque variable de départ à la composante avec laquelle elle est la plus fortement corrélée (en valeur absolue). Ici, les 6 premières variables et les protéines sont clairement associées à la première composante. On peut donc admettre que les informations de ces 7 variables sont fortement liées. La variable fibres est associée à la seconde composante principale. La variable carbone aussi, mais dans ce dernier cas, la différence entre les eux corrélations (0.600 et 0.747) est plus réduite et on pourrait aussi dire que cette variable est associée aux deux composantes. En conclusion, nous avons 7 variables qui ont beaucoup d'informations en commun, une variable (fibres) qui représente une information clairement distincte, et une variable (carbone) qui semble liés aux deux types d'informations.