OBSERVATOIRE DE LA QUALITÉ DE L'ALIMENTATION (Oqali)

RAPPORT MÉTHODOLOGIQUE 2009

Oqali

Observatoire de la qualité de l'alimentation

INRA

afssa
AGENCE FRANÇAISE DE SÉCURITÉ SANITAIRE DES ALIMENTS

OBSERVATOIRE DE LA QUALITÉ DE L'ALIMENTATION (Oqali)

RAPPORT MÉTHODOLOGIQUE

2009

Sommaire

Liste des figures

Liste des tableaux

1. TRAITEMENTS STATISTIQUES

Ce chapitre présente les traitements statistiques appliqués par l'Oqali pour la caractérisation des secteurs alimentaires.

1.1 Jeux de données

Dans la base de données Oqali, toutes les informations présentes sur les emballages des produits transformés pré-emballés suivis sont enregistrées : valeurs nutritionnelles, groupe d'étiquetage, allégations, liste d'ingrédients, portions, labels, enrichissement, conseils de préparation, type de stockage, code barre, poids et unité par lot, etc.

Les valeurs nutritionnelles peuvent provenir de deux sources : l'étiquetage ou l'analyse. Chaque référence est associée à un segment de marché et à une famille de produits spécifique à l'étude réalisée. De plus, lorsqu'un appariement avec les données TNS/Worldpanel est possible, la part de marché est renseignée.

Les tableaux suivants présentent la structure des données d'entrée pour les traitements statistiques (Tableau 1, Tableau 2).

Les termes et acronymes sont définis dans le lexique en fin de rapport.

Tableau 1 : Présentation de la structure des données

	Nutriments			Variables principales		Variables complémentaires		
	Nutriment 1	Nutriment 2	Nutriment 3	Segment de marché	Famille de produits	Variable complémentaire 1	Variable complémentaire 2	Variable complémentaire 3
Produits								

Tableau 2 : Description des données étudiées

Sources des données	Données d'étiquetage et/ou analytiques		
Informations principales			
Variables	**Type**	**Modalités**	**Précision**
Valeurs nutritionnelles	quantitative continue	Par nutriment	1 valeur moyenne par nutriment et type de produit
Segment	qualitative	HD, HDhg, MDD, MN, MDDeg…	
Famille de produit	qualitative	Selon l'étude	
Informations complémentaires			
Variables	**Type**	**Modalités**	**Précision**
Groupe d'étiquetage	qualitative	groupe 0, groupe 0+, groupe 1, groupe 1+, groupe 2, groupe 2+	
Allégations nutritionnelles	qualitative	Présence, Absence	pour un produit
Allégations de santé	qualitative	Présence, Absence	pour un produit
Repères nutritionnels	qualitative	Présence, Absence	pour un produit
Recommandations de consommation	qualitative	Présence, Absence	pour un produit
Incitation à l'activité physique	qualitative	Présence, Absence	pour un produit
Enrichissement et restauration	qualitative	Présence, Absence	pour un produit

1.2 Description des données

Tout d'abord, l'échantillon sur lequel les données sont analysées est décrit à l'aide de paramètres ou de diagrammes simples[1].

Lorsque les parts de marchés sont intégrées, une pondération des données est réalisée. Les formules utilisées sont les suivantes :

Moyenne pondérée : $\sum_{i=1}^{n} p_i \, x_i$ et Variance pondérée : $\sum_{i=1}^{n} p_i \left(x_i - m_p(x)\right)^2$

p_i : pondération pour la valeur i
$m_p(x)$: moyenne pondérée de x
x_i : valeur nutritionnelle i

1.2.1 Paramètres

Les paramètres permettant de résumer les distributions de données peuvent être classés en deux catégories :

- les paramètres de position ;
- les paramètres de dispersion.

1.2.1.1 Paramètres de position

Les différents paramètres de position sont le mode, la médiane, la moyenne, le 1er quartile, le 3ème quartile et les percentiles :

- **mode** : classe qui contient le plus d'effectif ;
- **médiane :** seuil pour lequel on retrouve un effectif égal d'une part et d'autre ;
- **moyenne (µ) ;**
- **1er quartile (q1) :** seuil pour lequel les valeurs se trouvant en dessous représentent 25% des données et les valeurs se trouvant au dessus représentent 75% des données ;
- **3ème quartile (q3) :** seuil pour lequel les valeurs se trouvant en dessous représentent 75% des données et les valeurs se trouvant au dessus représentent 25% des données ;
- **percentiles :** séparent la distribution en 100 parts égales.

[1] Dart T, Chatellier G (2003) : Comment décrire la distribution d'une variable ? *Rev Mal Respir*, 20 : 946-51.

1.2.1.2 Paramètres de dispersion

Les paramètres de dispersion sont l'étendue, l'intervalle interquartile, les extrêmes, la variance, l'écart-type et le coefficient de variation :

- **étendue :** distance entre le point minimum et le point maximum ;
- **intervalle interquartile :** distance entre le 1er et le 3ème quartile ;
- **extrêmes :** valeurs minimum et maximum de la distribution ;
- **variance (σ^2) :** moyenne de la somme des carrés des écarts à la moyenne ;
- **écart- type (σ) :** racine carrée de la variance ;
- **coefficient de variation (CV) :** rapport entre l'écart type et la moyenne.

Il faut apporter une attention particulière à l'interprétation de ces résultats qui sont en général sensibles aux valeurs extrêmes : ces valeurs doivent être identifiées, et si nécessaire, extraites des données initiales. Il est préférable de définir au préalable dans le protocole, comment les valeurs extrêmes seront prises en compte lors de l'analyse.

Certains graphiques, en plus de décrire visuellement la distribution, permettent de mettre en évidence les valeurs susceptibles d'être aberrantes.

1.2.2 Graphiques

Selon le type de données étudiées, plusieurs types de graphiques peuvent être réalisés.

1.2.2.1 Graphiques univariés (une variable)

Pour une variable quantitative continue

La distribution peut être visualisée à l'aide d'un histogramme ou d'une « boîte à moustaches » (boxplot) (Figure 1, Figure 2).

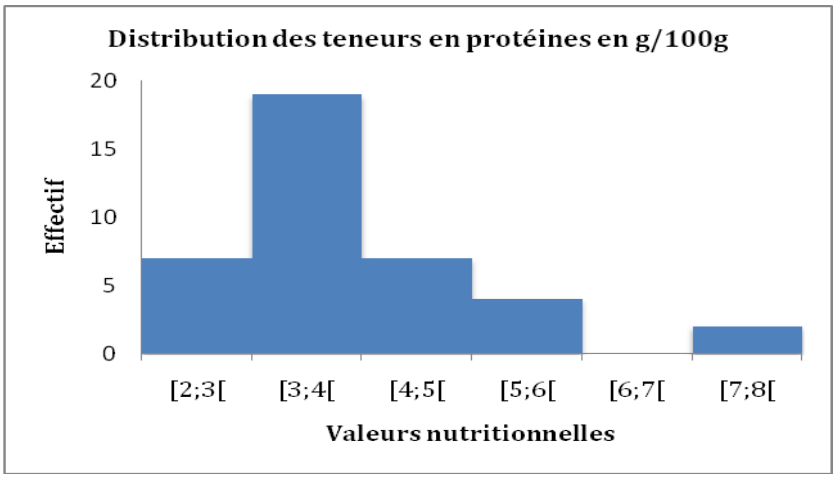

Figure 1 : Histogramme - exemple de distribution des teneurs en protéines dans une série de 40 échantillons

La variable quantitative est divisée en classes. L'effectif de chaque classe est représenté par un rectangle dont la base est l'amplitude de la classe et dont l'aire est proportionnelle à l'effectif.

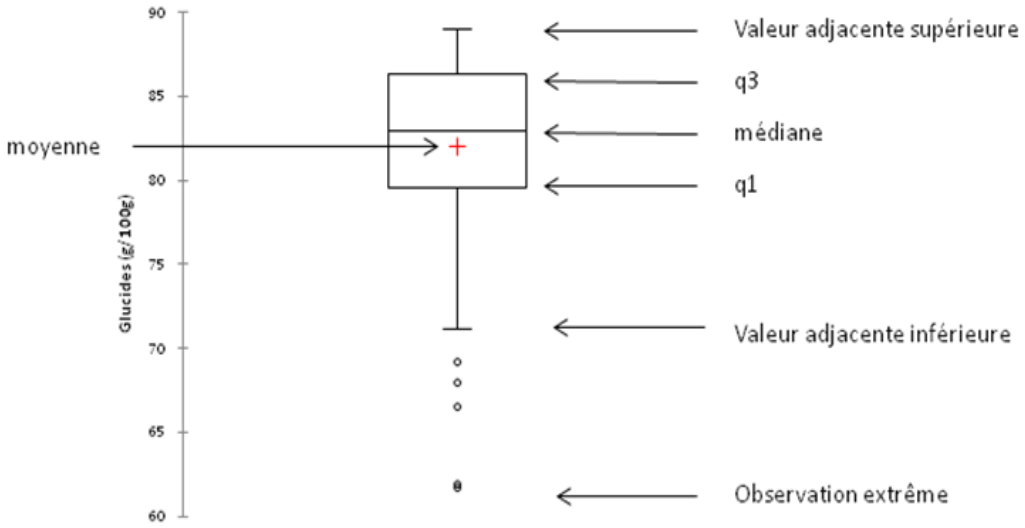

Figure 2 : Schéma d'une boîte à moustache

La « boîte à moustache » (box plot) est une façon simple de représenter et surtout de comparer la distribution d'une variable continue au sein de plusieurs groupes d'individus**.**

La boîte à moustache présente différentes informations :

- l'échelle des valeurs de la variable, située sur l'axe vertical ;
- le premier et le troisième quartile, q1 et q3, représentés par la base et le chapeau du rectangle central ;
- la médiane, représentée par la ligne horizontale située entre q1 et q3 ;
- la moyenne, représentée par une croix rouge ;
- la barre horizontale du bas indique la valeur adjacente inférieure, c'est-à-dire la valeur immédiatement supérieure à q1-1,5(q3-q1) ;
- la barre horizontale du haut indique la valeur adjacente supérieure, c'est-à-dire la plus grande observation inférieure à q3+1,5(q3-q1) ;
- les observations extrêmes sont les points au-delà de ces valeurs adjacentes.

Pour une variable qualitative

Pour représenter la répartition des modalités d'une variable qualitative étudiée, des diagrammes en barres (ou en bâtons) et des graphiques en secteurs peuvent être utilisés (Figure 3, Figure 4).

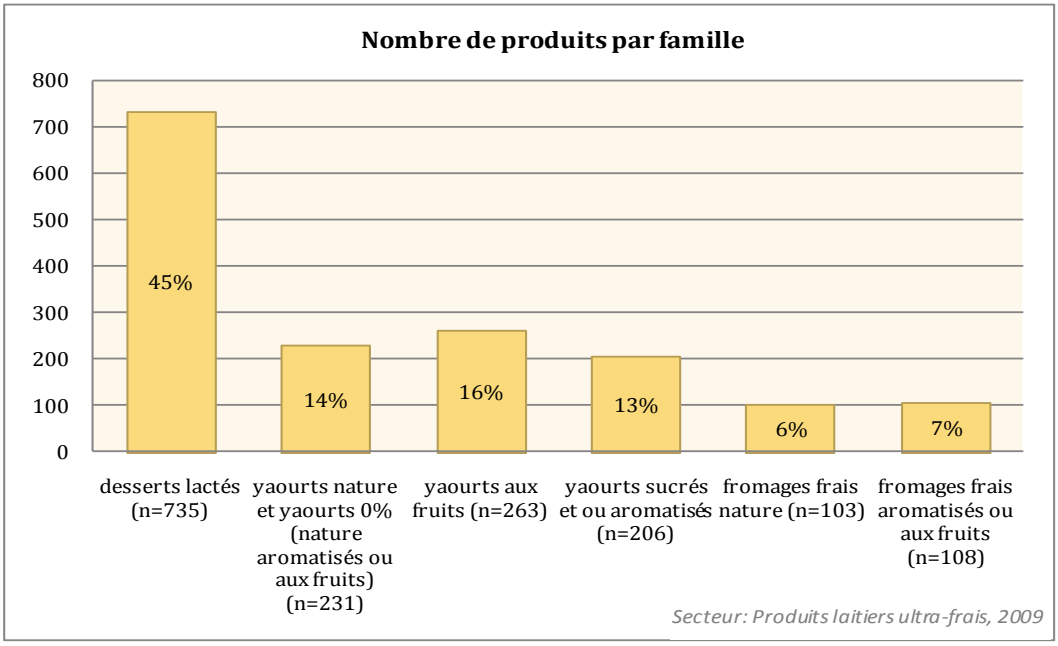

Figure 3 : Diagramme en barres - exemple de la répartition des produits par famille pour le secteur des produits laitiers ultra-frais

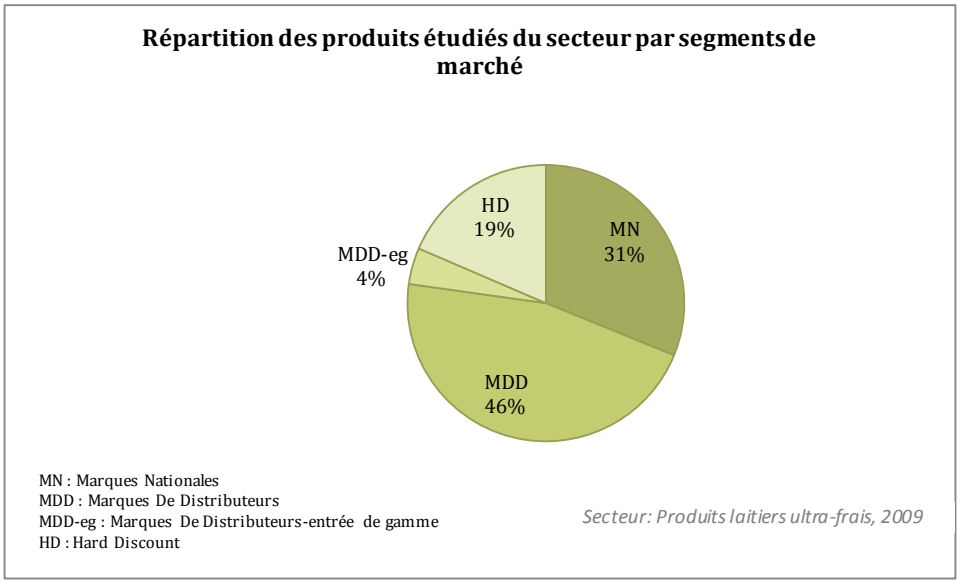

Figure 4 : Graphique en secteur - exemple de répartition des produits laitiers ultra-frais dans les différents segments de marché

1.2.2.2 Graphiques bivariés (deux variables)

Pour deux variables qualitatives

Un diagramme en barres empilées permet de représenter simultanément deux variables qualitatives (Figure 5).

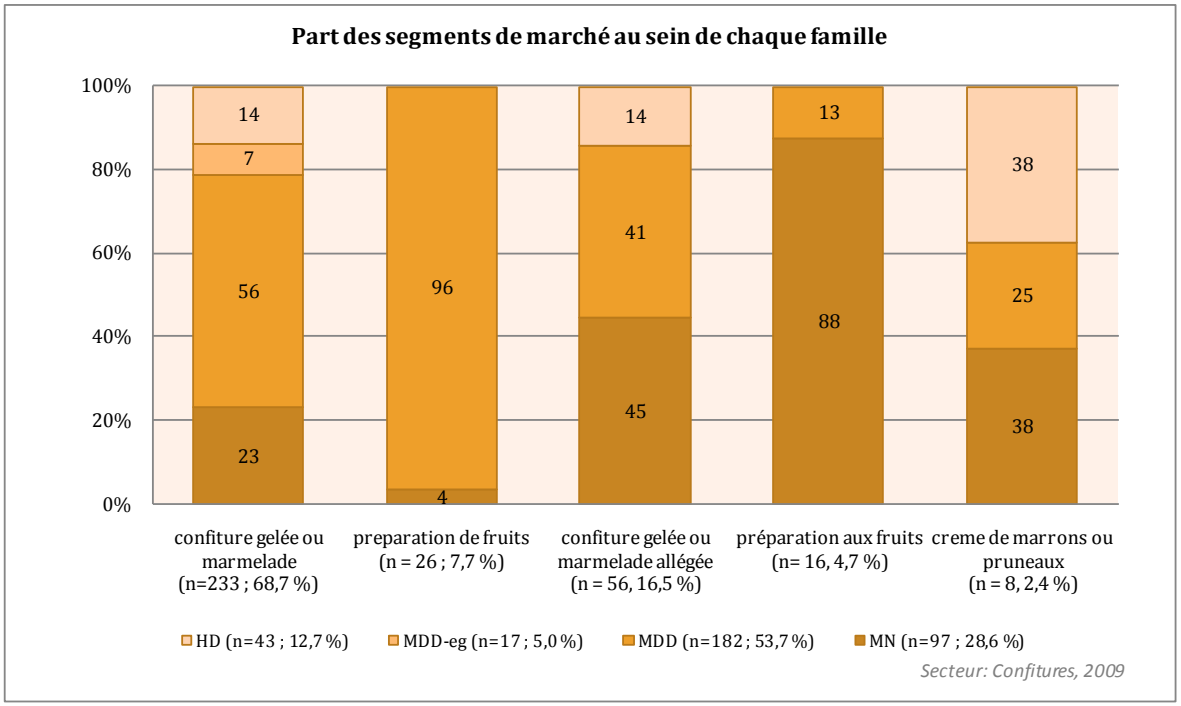

Figure 5 : Diagramme en barres empilées - exemple de répartition des produits entre segments de marché pour les différentes familles du secteur des confitures (n=339)

Pour deux variables quantitatives continues

Le nuage de points est un graphique de base pour la représentation conjointe de deux variables continues. Au sein du graphique, chaque point représente une référence-produit. Les coordonnées d'un point sont définies par les teneurs en deux nutriments choisis (par exemple : sucres-lipides). La coloration des points permet d'identifier une variable qualitative supplémentaire (par exemple une famille ou un segment de marché) (Figure 6).

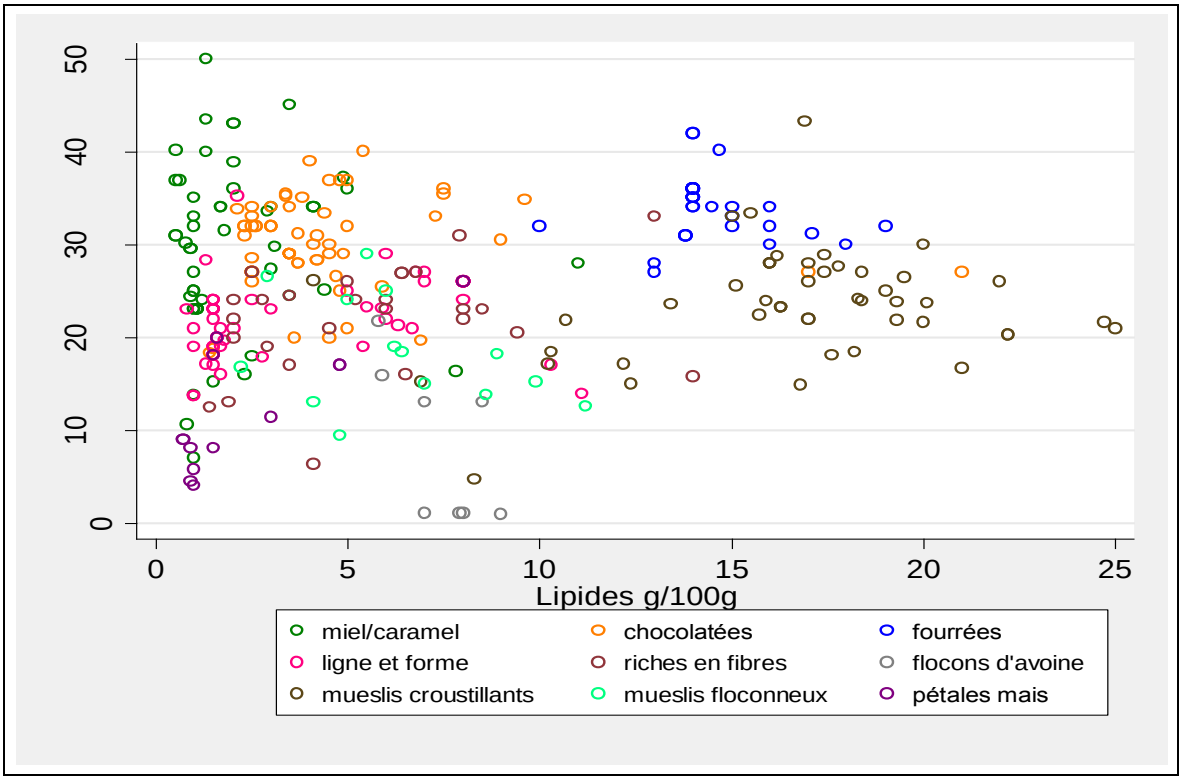

Figure 6 : Nuage de points - exemple de combinaison sucres/lipides pour le secteur des céréales pour le petit-déjeuner

Selon la forme du nuage obtenu, ce traitement permet de mettre en évidence soit une dispersion des références observées soit une liaison (corrélation linéaire) entre les variables étudiées. Cette corrélation linéaire peut être quantifiée par le calcul du coefficient de détermination (ou R^2), qui prend des valeurs comprises entre 0 et 1. Plus R^2 est proche de 1, plus les variables sont corrélées.

Ce graphique peut également prendre en compte des parts de marché, en pondérant le diamètre de chaque point par la part de marché associée à la référence. Ainsi, plus le diamètre du point est gros, plus la part de marché de la référence représentée est élevée. Le centre du point correspond aux coordonnées de la référence pour les deux nutriments choisis (Figure 7).

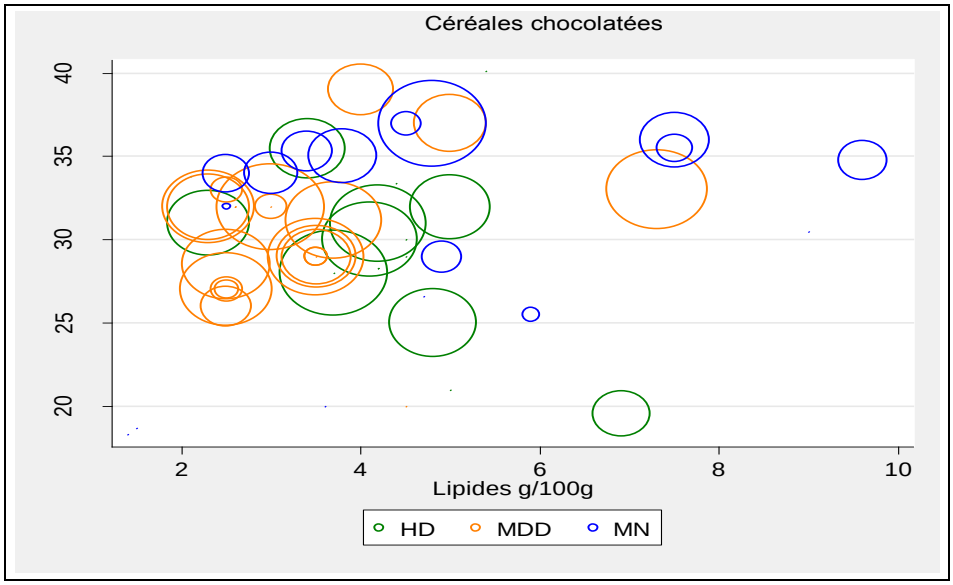

Figure 7 : Nuage de points pondéré - exemple de combinaison sucres/lipides pour la famille des céréales chocolatées, avec pondération par les parts de marché

1.2.2.3 Autres graphiques

Pour deux variables qualitatives et une variable quantitative discrète (effectifs) (Figure 8) :

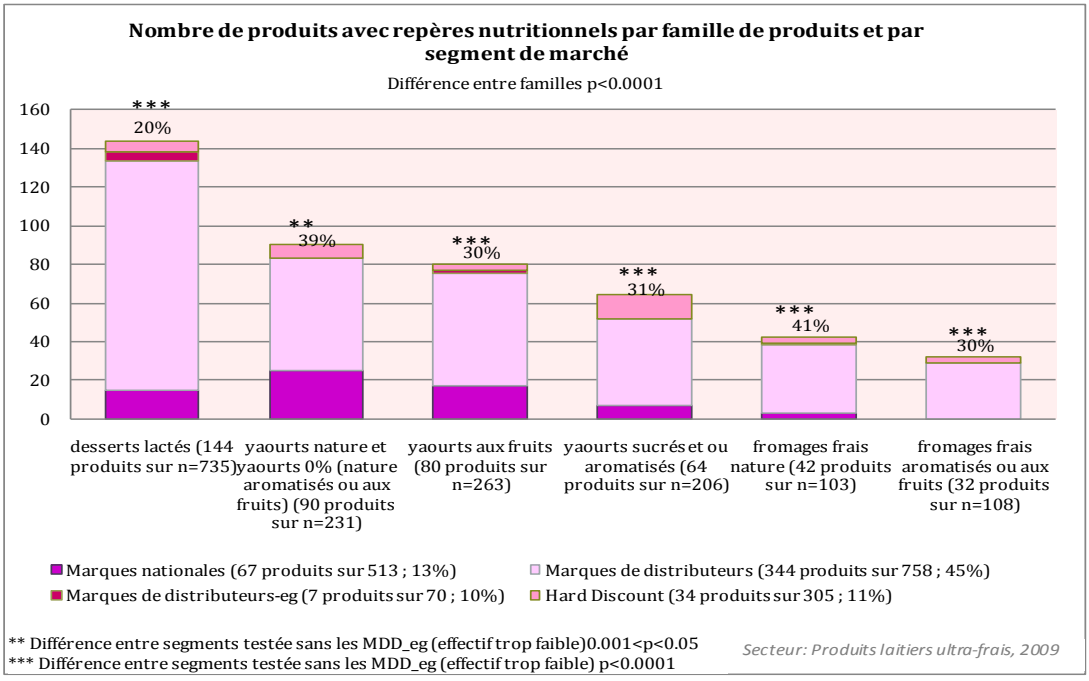

Figure 8 : Répartition des repères nutritionnels par segment de marché au sein des familles des produits laitiers ultra-frais

1.2.2.4 Graphiques des études Oqali

Pour les études réalisées au sein de l'Oqali, deux groupes de graphiques sont appliqués sur l'ensemble des données.

Pour l'étude des paramètres de l'étiquetage :

- diagrammes en barres ou en secteurs pour la représentation de la distribution selon :
  - les secteurs ;
  - les familles de produits ;
  - les segments de marchés.
- diagrammes empilés 100% pour représenter deux variables qualitatives (ex : segment de marché et familles de produits).

Pour l'étude de la variabilité nutritionnelle :

- boîtes à moustaches (box plot) permettant de schématiser la distribution et d'identifier les valeurs extrêmes éventuelles ;

- nuages de points permettant de visualiser une éventuelle corrélation entre deux variables.

1.2.3 Analyses multivariées des données

Afin de traiter simultanément plus de deux variables, il est possible d'avoir recours aux analyses multivariées[2] (Figure 9). Elles conduisent à des représentations planes, cartographies en deux dimensions du nuage multidimensionnel de points qui se répartissent selon les directions de projection des variables. Si ces analyses ne demandent pas de conditions particulières pour être appliquées, leur interprétation doit être soigneuse.

1.2.3.1 Présentation de différentes analyses multivariées

- **L'analyse en composantes principales** (ACP) prend en compte simultanément de nombreuses variables quantitatives. Le principe de cette analyse consiste à mettre en évidence des relations linéaires fortes entre les variables étudiées. Cela permet de garder le maximum d'information tout en réduisant le nombre de variables étudiées.

[2]Falissard B (2005) : Comprendre et utiliser les statistiques dans les sciences de la vie. *Masson*, p217-360.

- **L'analyse factorielle en correspondances** (AFC) met en évidence graphiquement la liaison entre deux variables qualitatives. Il est possible également de voir quels sont les individus qui influencent le plus cette liaison. On parle d'**analyse factorielle en correspondances multiples** (AFCM) lorsqu'il y a plus de deux variables qualitatives.

- La **classification** permet de regrouper les données en sous-groupes homogènes selon leur structure. Elle est appliquée à plusieurs variables qui sont soit qualitatives soit quantitatives. Il existe différentes méthodes de classifications.

- L'**analyse discriminante** (AD) permet de classer les individus selon leurs caractéristiques. Elle s'applique quand il y a une variable qualitative et plusieurs variables quantitatives.

- L'**analyse canonique** permet de déterminer les corrélations linéaires entre un groupe de variables à expliquer et un groupe de variables explicatives.

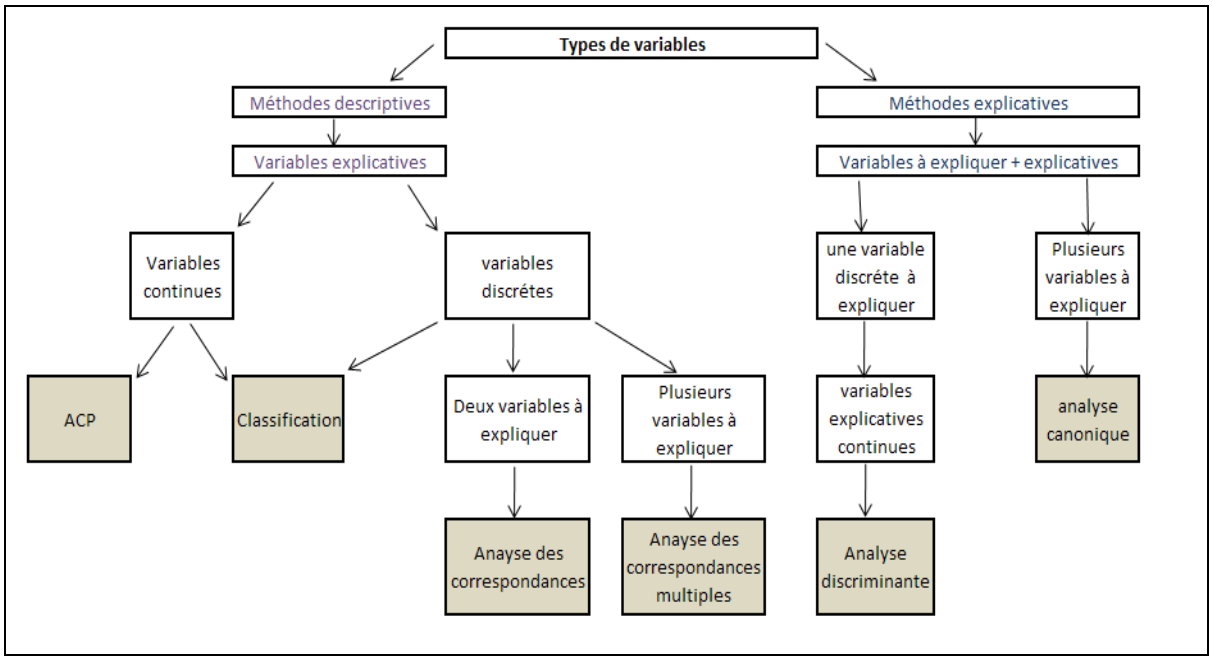

Figure 9 : Schéma de décision pour appliquer une analyse de données multivariée

1.2.3.2 Analyses de données des études Oqali

Lors des bilans sectoriels, une **analyse factorielle discriminante** (AFD) et/ou une **classification ascendante hiérarchique** (CAH) peuvent être réalisées.

L'AFD permet de décrire par exemple les caractéristiques nutritionnelles de différentes familles d'un secteur. Il s'agit ici de définir des combinaisons de nutriments qui permettent la meilleure séparation entre les différentes familles de produits, définies *a priori*.
Parallèlement à l'AFD, certains logiciels réalisent un test qui, à partir des indicateurs nutritionnels mis en évidence, ré-affectent les références au sein des familles. Ce test permet d'étudier l'hétérogénéité de la composition nutritionnelle entre les familles.

La CAH permet de regrouper, par exemple, les différents produits en sous-groupes homogènes en fonction de leurs valeurs nutritionnelles. Ces groupes permettront de définir les différentes familles de produits.

1.3 Tests statistiques

Les tests statistiques permettent, lors d'une comparaison entre deux échantillons, de savoir si la variation observée est réelle ou si elle résulte de fluctuations d'échantillonnage[3].

Deux familles de tests statistiques peuvent être distinguées :

- **Les tests paramétriques**

Ce sont des tests statistiques fondés sur des hypothèses sur la loi de distribution de la variable étudiée. Il existe de nombreuses lois de distribution que l'on peut résumer par certaines valeurs caractéristiques, appelées paramètres. Dans la majorité des cas, ces tests paramétriques sont fondés sur la loi normale, qui possède deux paramètres : la moyenne et l'écart type.

- **Les tests non paramétriques**

Ces tests comparent les distributions entre elles sans hypothèse particulière sur la forme de la distribution de la variable étudiée.

Les tests paramétriques doivent être privilégiés, lorsque les conditions d'application sont vérifiées. En effet, ils sont plus puissants et permettent donc de mettre en évidence plus facilement les différences lorsqu'elles existent. Les tests non paramétriques sont, quant à eux, plus robustes mais moins puissants.

[3] Ancelle T (2006) : Statistique-Epidémiologie. *Maloine,* p.85-92.

Lors de l'analyse de deux échantillons, les données étudiées peuvent être de trois types :

- **indépendantes**, les produits des deux échantillons sont différents ;
- **appariées**, les produits des deux échantillons sont identiques selon un critère établi ;
- **semi appariées**, certains produits des deux échantillons sont identiques et d'autres ne le sont pas (Figure 10).

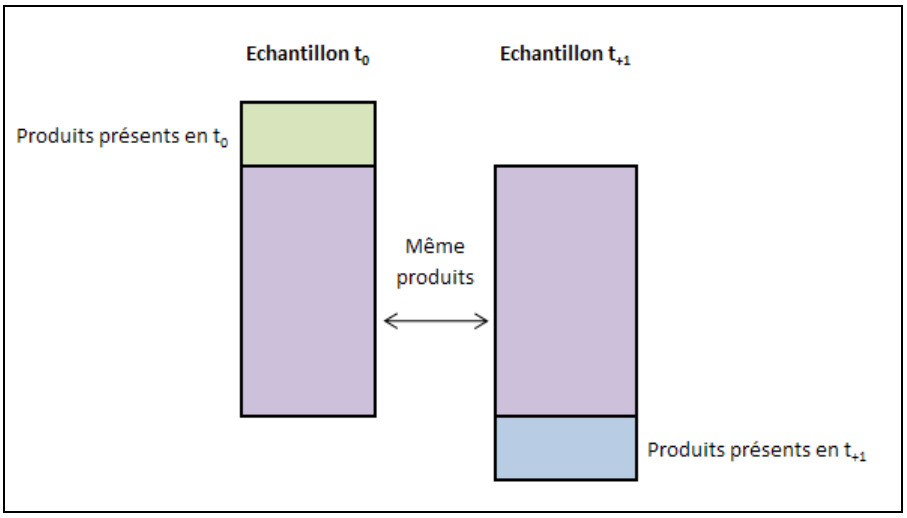

Figure 10 : Représentation d'échantillons semi appariés

1.3.1 Le principe des tests de comparaisons

Un test statistique se compose d'au moins trois étapes : formulation des hypothèses, réalisation du test et interprétation des résultats.

Formulation des hypothèses

Des hypothèses *a priori* peuvent être fondées sur la réflexion suivante :

- la différence observée entre 2 échantillons est due aux fluctuations d'échantillonnage ;
- la différence observée entre 2 échantillons est réelle.

En statistique, les formulations de ces hypothèses sont les suivantes :

- **Hypothèse nulle (H_0) :** pas de différence entre les paramètres. Si une différence est observée, celle-ci est due aux fluctuations d'échantillonnage ;
- **Hypothèse alternative (H_1) :** il existe une réelle différence entre les paramètres.

L'étape suivante consiste à définir le risque alpha (α=risque de première espèce). Celui-ci représente la probabilité de rejeter H_0 alors que H_0 est vraie. Généralement, un risque α de 5% est utilisé.

De la même manière, le risque beta (β=risque de deuxième espèce), par convention fixé à 20%, est le risque de rejeter H_1 alors que H_1 est vraie.

Par ailleurs, deux tests peuvent être considérés :

- le test unilatéral (dans le cas où on connait *a priori* la position relative des deux échantillons) ;
- le test bilatéral (dans le cas où on ne connait pas *a priori* la position relative des deux échantillons).

Réalisation et interprétation du test

Les principaux résultats d'un test statistique sont :

- un degré de probabilité de rejeter H_0, si H_0 est vraie (autrement appelé P ou P-value ou degré de signification) ;
- une différence observée entre les valeurs étudiées (U_0).

Selon la valeur d'U_0 (différence observée) lors de la comparaison des deux échantillons, deux situations sont envisageables (Figure 11) :

- Si la valeur U_0 est inférieure à la valeur du modèle de distribution théorique (Uα) alors le degré de signification (P) de l'écart observé est supérieur à α. Par conséquent, H_0 n'est pas rejetée.

- Au contraire si la valeur U_0 est supérieure à la valeur du modèle de distribution théorique (Uα), alors le degré de signification (P) de l'écart observé est inférieur à α. Par conséquent, H_0 est rejetée et l'hypothèse alternative H_1 est acceptée.

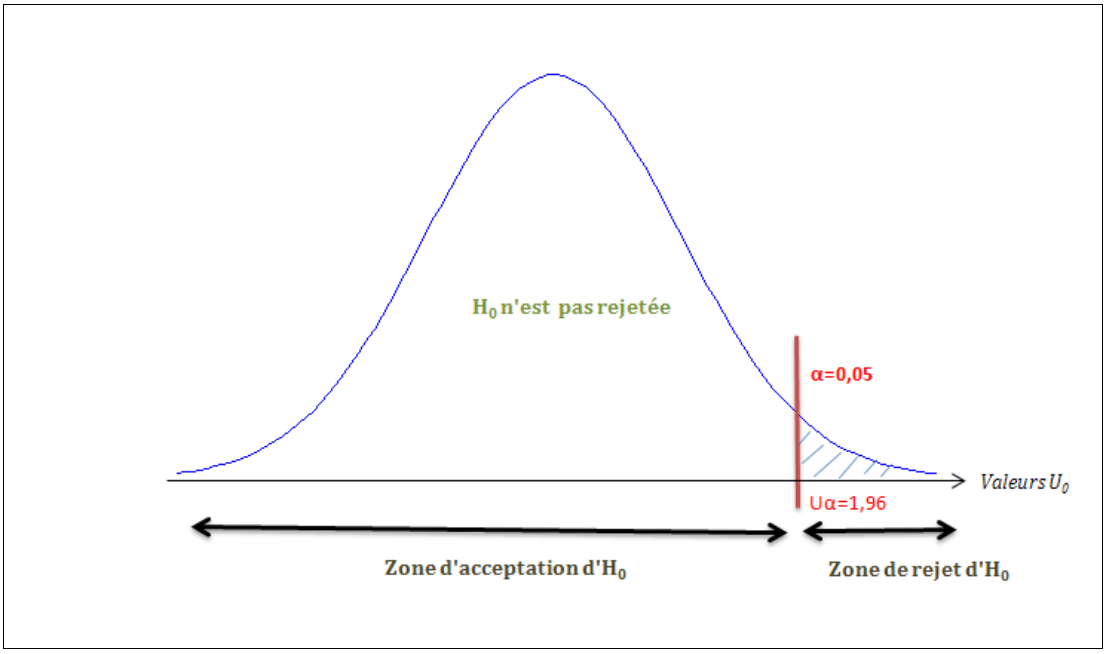

Figure 11 : Acceptation ou rejet de l'hypothèse nulle selon la valeur d'U_0 pour un test unilatéral

La zone de rejet de l'hypothèse nulle n'est pas la même si le test est unilatéral ou bilatéral (Figure 12).

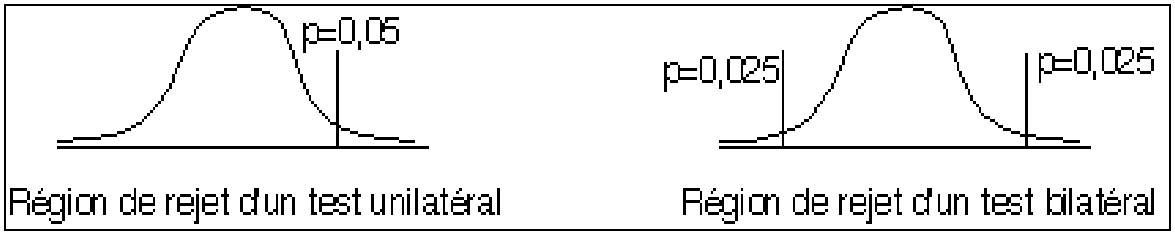

Figure 12 : Zone de rejet pour un test bilatéral ou unilatéral (α=5%)

L'interprétation des résultats repose beaucoup sur les paramètres α et β[4] (Tableau 3). Toute conclusion doit être faite en indiquant la valeur d'α.

Tableau 3 : Interprétation des valeurs α et β

		Situation réelle	
		Différence	Pas de différence
Résultats obtenus	Différence	**valide**	**Risque α**
	Pas de différence	**Risque β**	**valide**

Ajustement de la valeur α lorsque plusieurs tests sont effectués

Dans une même étude, si plusieurs comparaisons deux à deux sont réalisées, la probabilité de trouver au moins une différence significative à un α fixé augmente avec le nombre de tests effectués. Il est donc nécessaire d'ajuster le risque α encouru en fonction du nombre de comparaisons.

Ainsi, la valeur α corrigée est obtenue en divisant le risque α global fixé (5%) par le nombre de tests réalisés. L'ajustement permet de conserver un α global de 5%. En revanche, si l'ajustement n'est pas appliqué, l'α global peut être très élevé. Par exemple, si 10 tests sont réalisés sans ajuster α, on a alors 1 chance sur 2 de rejeter H_0 à tort sur l'un des tests (soit un α global de 50%).

4 Jenny J.-Y (2001) : Le risque bêta : un risque méconnu d'erreur en statistique. *Revue de chirurgie orthopédique*, *87* 170-172.

1.3.2 Les différents tests statistiques

Lorsque la taille n de chaque échantillon est supérieure à 30, un test paramétrique peut être utilisé sans vérification des conditions d'application particulières[5].

Lorsque la taille d'au moins un échantillon est inférieure à 30, les conditions d'application des tests paramétriques doivent être vérifiées, notamment par des tests préliminaires sur les données :

- normalité des distributions : pour tester la normalité des données, les tests de Kolmogorov-Smirnov ou de Shapiro-Wilk peuvent être utilisés ;
- homogénéité des variances des distributions : pour tester l'égalité des variances, le test de Fisher-Snedecor (test paramétrique de comparaison de 2 variances), le test de Levene (test de comparaison de plusieurs variances, qui ne nécessite aucune condition particulière d'application) ou le test de Bartlett (pour plusieurs variances et pour des données suivant une loi normale) peuvent être utilisés. Dans le cas où les échantillons sont de même taille cette condition peut être parfois omise (la robustesse du test est surtout due à la normalité des distributions).

Ces tests préliminaires sont systématiquement réalisés dans le cadre de l'étude des données de l'Oqali.

Dans le cas où les distributions ne suivent pas une loi normale, des tests non paramétriques doivent être utilisés (Figure 13).

[5] Ancelle T (2006) : Statistique-Epidémiologie. *Maloine,* p.124-170.

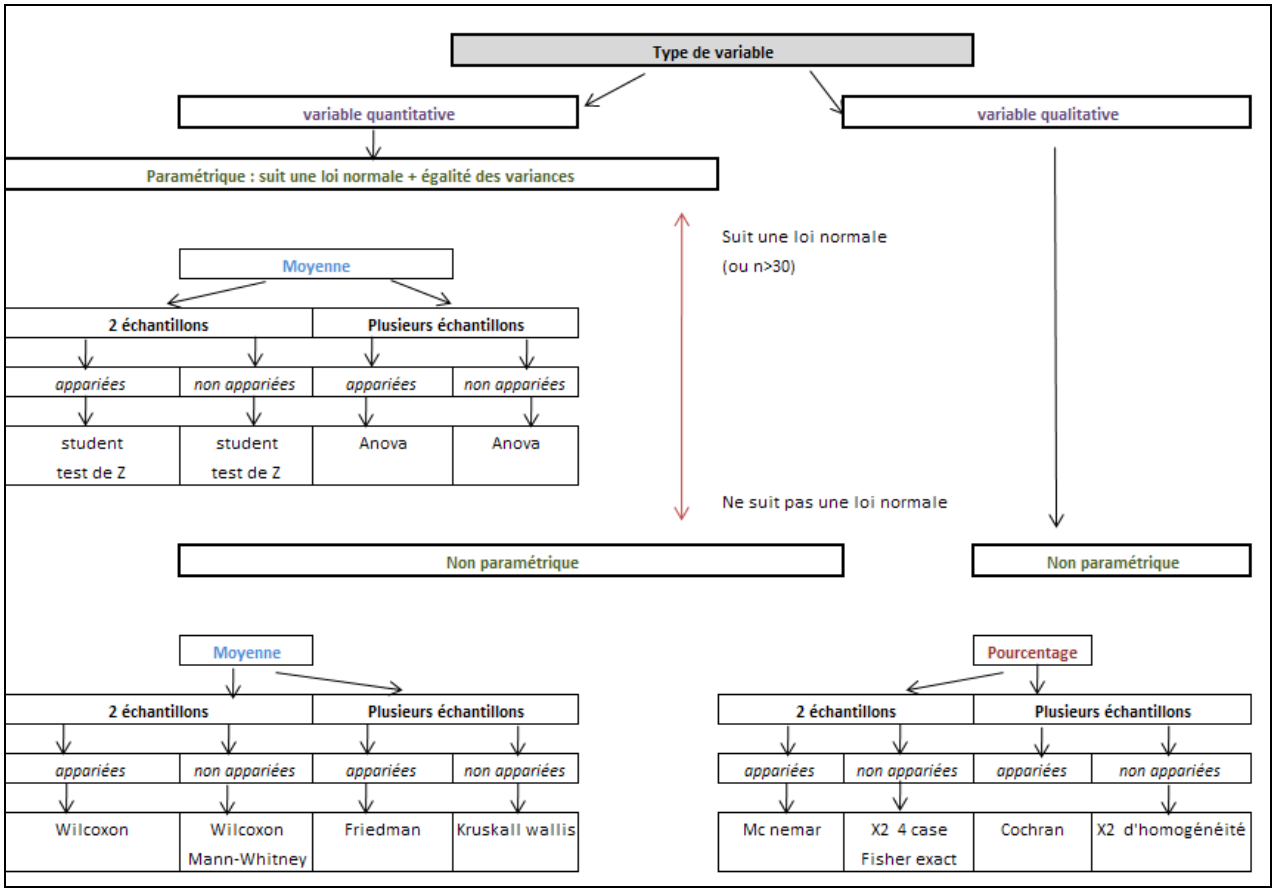

Figure 13 : Schéma de décision pour l'application d'un test statistique

Les tableaux 4 à 6 présentent les conditions d'applications des différents tests statistiques pour :

- la comparaison de deux échantillons ;
- la comparaison d'un échantillon à une population de référence ;
- la comparaison de plusieurs échantillons.

Tableau 4 : Comparaison de deux échantillons

Grandeur étudiée	Séries comparées	Test	Conditions d'applications
Moyenne	*Indépendantes*	Z de l'écart réduit	effectifs des 2 échantillons n_1 et $n_2 \geq 30$
		T de Student	effectifs des 2 échantillons n1 et n2 < 30 normalité des distributions dans les 2 populations variances égales ou n1≈n2
		Wilcoxon/Mann-Whitney	absence de normalité des distributions effectifs des 2 échantillons n1 et n2 ≥ 10
	Appariées	Z de l'écart réduit	nombre de paires ≥ 30
		T de Student	nombre de paires < 30 normalité des distributions dans les 2 populations variances égales ou n1≈n2
		Wilcoxon	absence de normalité des différences effectifs des paires ≥ 10
Variances	*Indépendantes*	F Fisher-Snedecor	normalité des distributions dans les 2 populations
Pourcentages	*Indépendantes*	χ^2 4 cases	effectifs théoriques ≥ 5
		Fisher : test exact	aucune
	Appariées	χ^2 McNemar	nombre de paires discordantes ≥ 10
Distributions	*Indépendantes*	χ^2 d'homogénéité	effectifs théoriques dans chaque case ≥ 5

Tableau 5 : Comparaison d'un échantillon à une population de référence

Grandeur étudiée	Test	Conditions d'applications
Moyenne	Z de l'écart réduit	effectif de l'échantillon n≥30
	T de Student	normalité de la distribution dans la population effectif de l'échantillon n<30
Pourcentages	χ^2 de conformité	effectif théorique ≥ 5
Distribution	χ^2 de conformité	effectif théorique ≥ 5

Tableau 6 : Comparaison de plusieurs échantillons

Grandeur étudiée	Séries comparées	Test	Conditions d'applications
Moyenne	*Indépendantes*	F Fisher-Snedecor(Anova)	normalité des distributions dans les populations
		Kruskal-Wallis	absence de normalité des distributions, effectifs échantillons $n_i \geq 10$
	Appariées	Friedman	absence de normalité des distributions
Distribution		χ^2 d'homogénéité	effectifs théoriques dans chaque case ≥ 5
Pourcentages	*Indépendantes*	χ^2 d'homogénéité	effectifs théoriques dans chaque case ≥ 5
	Appariées	Cochran Mantel Haenszel	nombre de paires discordantes ≥ 10

Les tests *post hoc*

Les tests de comparaison de plusieurs échantillons (ex : Anova, Kruskal-Wallis) permettent de savoir s'il existe au moins un échantillon différent des autres. Si ce premier test met en évidence des différences significatives, il est intéressant de savoir quels sont les échantillons qui diffèrent. Pour cela, les tests *post hoc*[6][7][8] sont utilisés. Ils permettent une comparaison deux à deux des différents échantillons.

- Suite à un test paramétrique, il est possible de faire un test de la plus petite différence significative (LSD), un test de Tukey ou un test de Newman-Keuls.

- Suite à un test non paramétrique, il est possible de faire un test de Wilcoxon, un test de Dunn ou un test de Nemenyi.

[6] Dagnelie P (2006): Statistique théorique et appliquée. Tome 2. Inférence statistique à une et à deux dimensions*. De Boeck Université*, p.406-407
[7] Hochberg Y, Tamhane A-C (1987): Multiple Comparison Procedures. *John Wiley & Sons.*
[8] Jerrold H. Zar , (1998) : Biostatistical Analysis. *Prentice Hall*

1.3.3 Tests statistiques des études Oqali

La plupart des échantillons étudiés dans le cadre de l'Oqali a soit un effectif faible soit une distribution non normale : les tests appliqués sont non paramétriques.

Deux sortes d'analyses sont faites sur l'ensemble des données :
- étude des paramètres de l'étiquetage ;
- étude de la variabilité nutritionnelle.

Etude des paramètres de l'étiquetage

Dans l'étude des paramètres d'étiquetage des variables complémentaires qualitatives telles que les familles de produits ou les segments de marché peuvent être intégrées. Afin de savoir s'il existe un effet famille de produits ou un effet segment de marché, des tests de χ^2 d'homogénéité sont réalisés.

Le tableau 7 présente un exemple de résultats du test de χ^2.

Hypothèses du test :

H_0 : le nombre de produits possédant une allégation ne diffère pas statistiquement selon le segment de marché.

H_1 : le nombre de produits possédant une allégation diffère statistiquement selon le segment de marché.

Tableau 7 : Nombre de produits laitiers ultra-frais présentant ou non une allégation nutritionnelle par segment de marché

	HD	MDD	MN	MDDeg	Total général
Présence d'allégation	33	191	136	6	366
Absence d'allégation	268	565	371	63	1267
Total général	301	756	507	69	1633

P- value du χ^2= 1,45148E-08

L'hypothèse H_0 est rejetée car la P-value est inférieure à la valeur d'α qui est de 5%. Ainsi, pour le secteur des produits laitiers ultra-frais, le nombre de produits possédant une allégation diffère statistiquement selon le segment de marché.

Etude de la variabilité nutritionnelle

Les valeurs nutritionnelles sont également analysées. Ces variables sont des variables quantitatives continues. L'objectif est de savoir s'il existe des différences significatives de composition nutritionnelle entre les familles de produits et/ou entre les segments de marché. Pour savoir quel test appliquer, un test de normalité est d'abord réalisé. Dans la plupart des cas, on s'oriente vers des tests non paramétriques multiples sur données indépendantes (Kruskal-Wallis).

Pour chaque nutriment, il est intéressant de savoir, si le test de Kruskal-Wallis est significatif, quels sont les segments qui sont semblables ou différents. Pour cela, on utilise un test post-hoc de comparaisons deux à deux, comme par exemple le test de Wilcoxon.

Le tableau 8 présente un exemple de résultats des comparaisons deux à deux.

Tableau 8 : Variabilité nutritionnelle des yaourts nature ou 0% - différence entre segment de marché

	P Kruskal-Wallis	Marques nationales			Marques de distributeurs			Premiers prix			Hard Discount			Total N	Total Moyenne
Constituants (g/100g)	(α=0,0055)	N	**Moyenne**	ET	N	**Moyenne**	ET	N	**Moyenne**	ET	N	**Moyenne**	ET		
valeur énergétique (kcal)	0,022	82	**53,49**	14,98	107	**54,07**	20,37	6	**36,50**	11,88	28	**62,32**	27,70	223	51,59
protéines	0,0079	82	**4,26**	0,48	107	**4,06**	0,51	6	**4,12**	0,26	28	**4,11**	0,51	223	4,14
lipides	0,2611	82	**0,93**	1,69	107	**1,68**	2,51	6	**0,55**	0,54	28	**2,03**	2,59	223	1,30
acides gras saturés	**0,0059**	64	**0,39** [b]	0,82	72	**0,94** [a,b]	1,53	1	**0,50** [a,b]		9	**2,02** [a]	1,74	146	0,96
glucides	**0,0004**	82	**6,79** [a]	2,48	106	**5,72** [b]	1,76	6	**5,05** [a,b]	1,11	28	**5,96** [b]	3,39	222	5,88
sucres	**<0,0001**	63	**6,21** [a]	1,84	73	**5,21** [b]	1,84	1	**4,50** [a,b]		9	**4,31** [b]	0,43	146	5,06
fibres	**0,002**	64	**0,47** [a]	0,67	73	**0,29** [b]	0,62	1	**0,00** [a,b]		9	**0,00** [b]	0,00	147	0,19
calcium (mg)	**<0,0001**	70	**143,06** [a]	36,37	54	**126,94** [b]	10,18				8	**137,63** [a,b]	15,46	132	135,88
sodium (mg)	**<0,0001**	64	**0,07** [a]	0,02	73	**0,05** [b]	0,02	1	**0,05** [a,b]		9	**0,05** [b]	0,01	147	0,05

Pour les constituants ayant au moins un des segments de marché différents des autres (ligne de valeurs en violet), un regroupement des moyennes des différents segments de marchés est effectué. Les moyennes sans aucune lettre commune (« a » d'une part, « b » d'autre part) sont statistiquement et significativement différentes. Celles avec des lettres communes (ex : « b » et « ab ») ne sont pas statistiquement différentes.

Pour les sucres, par exemple, deux groupes se forment : les marques nationales et les premiers prix d'une part et les marques de distributeurs, les premiers prix[9] et les hard discount d'autre part.

[9] Premiers prix = MDD entrée de gamme.

1.3.4 Statistiques pour données semi appariées

Pour les données semi appariées, le test est le même que pour des données appariées.

Test de Z pour 2 moyennes sur deux séries appariées

$$Z=|m_d-0|/S_{md}$$

Différence entre paires : $d_i=x_i-y_i$
Moyenne des différences : $m_d=\Sigma\ di/n$
Variance des différences : $S_d^2= [\Sigma di2-((\Sigma d_i)^2/n)]/n-1$
Ecart type de la moyenne des différences : $S_{md=}$ racine (S_d^2/n)

Hypothèses du test :

H_0 : les moyennes des séries ne sont pas significativement différentes, c'est-à-dire que la moyenne d'une série n'est pas supérieure ou inférieure à l'autre.
H_1 : les moyennes des séries sont significativement différentes, c'est-à-dire que la moyenne d'une série est supérieure ou inférieure à l'autre.

En appliquant le test sur des données semi appariées, les calculs de la moyenne des différences et de sa variance[10] sont différents.

Le calcul de ces paramètres peut se faire de deux façons :

- méthode 1 : on utilise les données de chaque échantillon (Figure 14) ;
- méthode 2 : on utilise les données appariées d'une part et les données non appariées d'autre part (Figure 15).

La 2ème méthode est plus complexe mais elle est plus puissante.

[10] Bart J, Fligner M-A , Notz W-I (1998) : Sampling and Statistical Methods for Behavioral Ecologists, *Cambridge*, p72-78

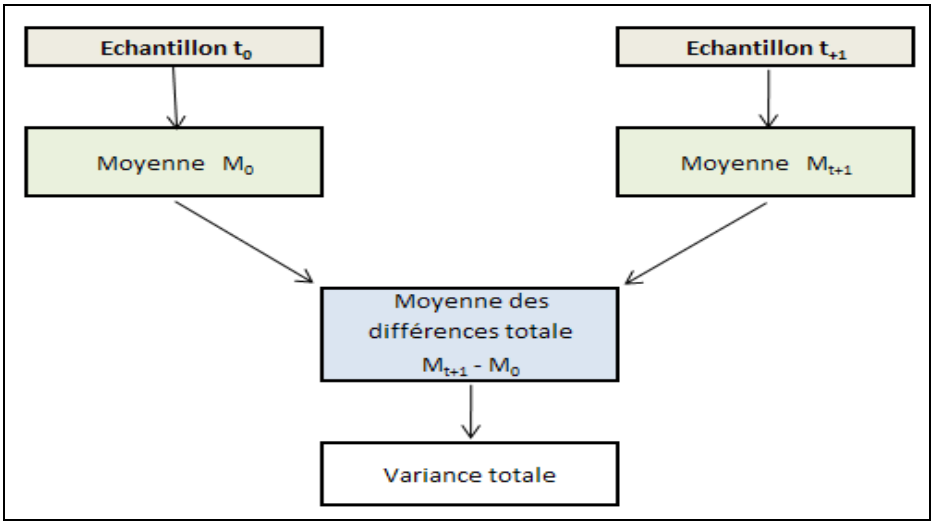

Figure 14 : Schéma de calcul des paramètres (méthode 1)

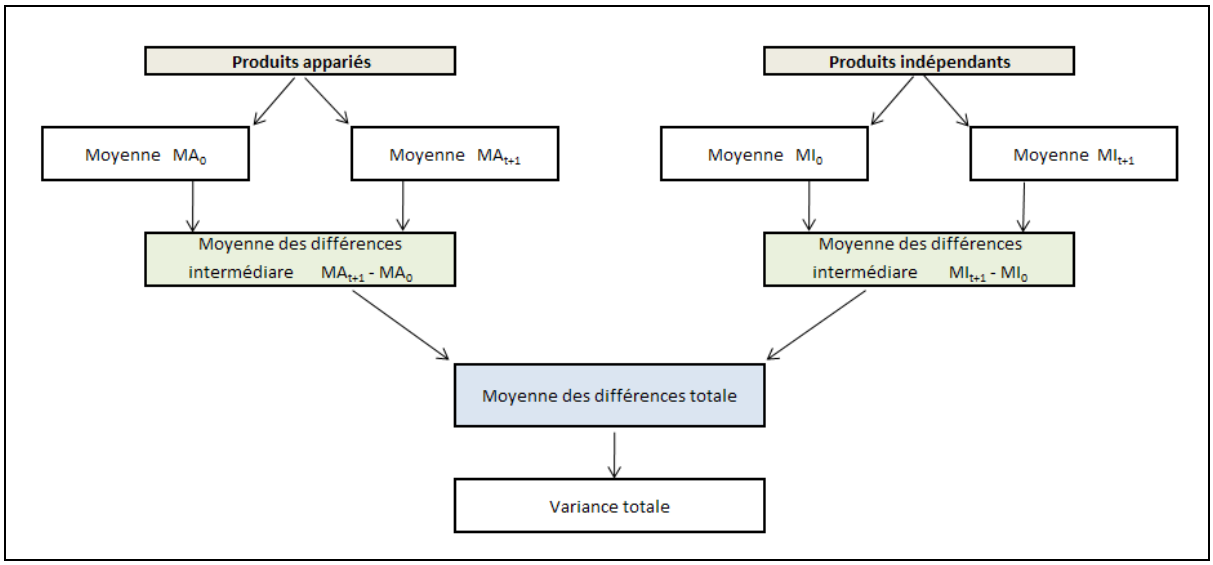

Figure 15 : Schéma de calcul des paramètres (méthode 2)

Le tableau 9 présente un exemple des résultats obtenus à partir des données de l'étude produits laitiers ultra-frais.

Tableau 9 : Résultats pour les teneurs en protéines (secteur des produits laitiers ultra-frais)

	Moyenne des différences	Ecart-type	Valeur du test de Z
Résultats en gardant seulement les données appariées	0,07	0,32	0,22
Résultats en gardant toutes les données (méthode 1)	0,24	5,11	0,05
Résultats en gardant toutes les données (méthode 2)	0,044	0,31	0,14

Niveau critique dans la table de Z pour alpha α=5% : 1,96 (test bilatéral) ou 1,645 (test unilatéral).

Pour conclure à une différence significative entre les deux échantillons, il faut que la valeur du test de Z soit supérieure au niveau critique.

Dans l'exemple du Tableau 9, aucune des méthodes ne conclut à une différence significative.

2. COMPARAISON DES VALEURS NUTRITIONNELLES ÉTIQUETÉES ET DES DONNÉES ANALYTIQUES

Certaines études sectorielles peuvent prendre en compte à la fois des données d'étiquetage et des données analytiques, notamment pour les produits qui ne proposent pas d'étiquetage nutritionnel ou qui présentent un étiquetage nutritionnel incomplet.

Les analyses peuvent être réalisées référence par référence (un échantillon = une référence-produit) ou sur un échantillon composite (composé de plusieurs produits ou références). Les plans d'échantillonnage sont établis au cas par cas en fonction du secteur étudié, du nombre de produits à analyser, du coût associé, … (Tableau 10).

Afin d'être les plus représentatifs possibles du marché, la liste des produits à analyser est déterminée à partir des catégories de produits les plus vendues.

Dans le cas d'analyses sur des échantillons composites, les produits constituant les différents échantillons sont établis par l'Oqali en fonction de l'étude réalisée. Chaque produit peut être ajouté en quantité égale ou en fonction de sa part de marché (Figure 16).

Tableau 10 : Les différents types d'analyses réalisées

Objectifs	Type d'échantillon	Part des produits dans l'échantillon	Avantages	Inconvénients
Obtenir des valeurs moyennes	Echantillon composite (plusieurs produits)	égale pour tous les produits	Permet la comparaison analyse/ étiquette	Acquisition de données agrégées
		proportionnelle aux parts de marché	Réduit les coûts	
Obtenir des valeurs manquantes pour produits sans étiquetage	Echantillon simple		Acquisition de données au niveau produit	Onéreux

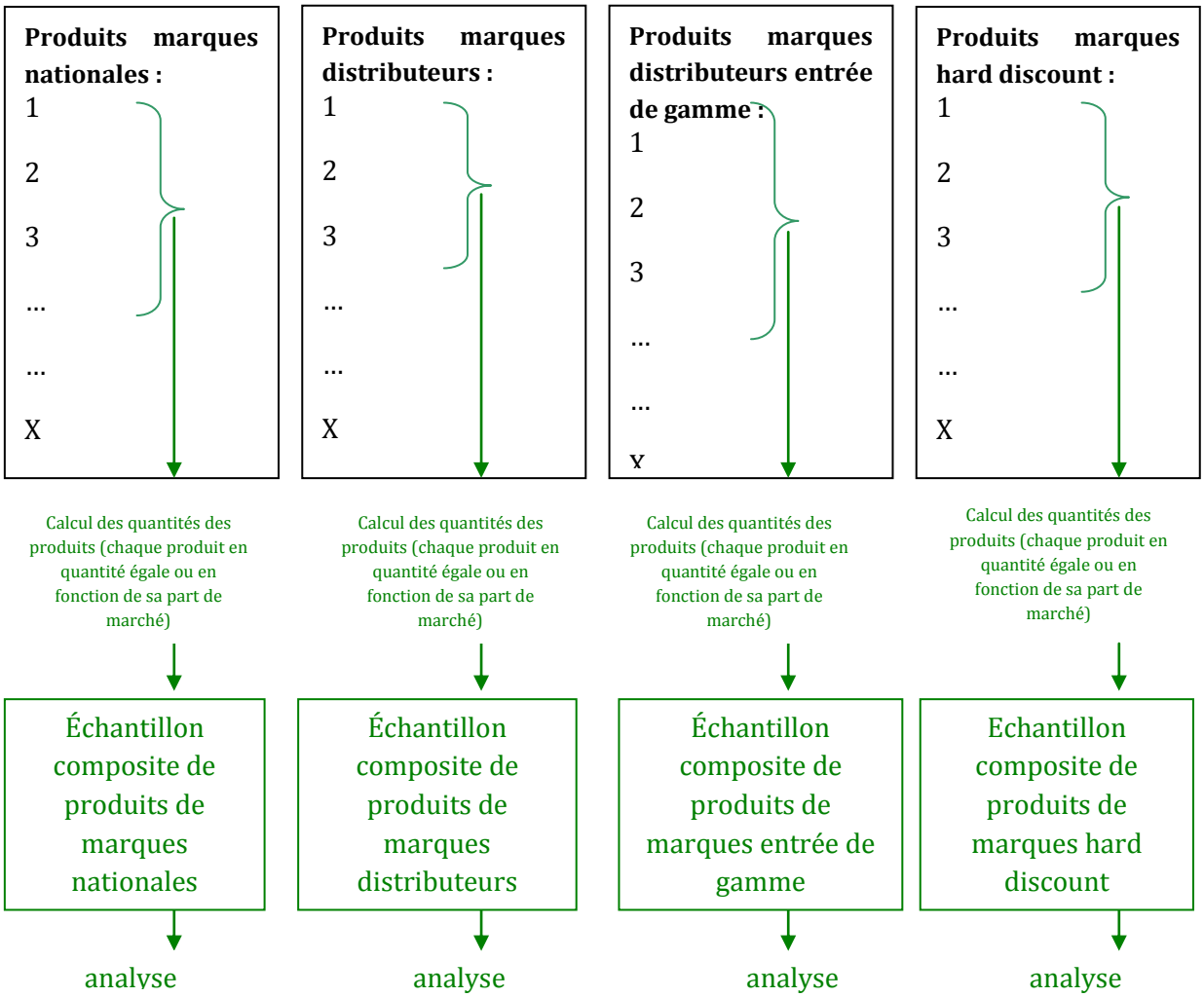

Figure 16 : Réalisation d'échantillons composites

Pour permettre la comparaison entre les valeurs nutritionnelles étiquetées et les valeurs analytiques, les résultats des analyses de composition nutritionnelle sur échantillons composites, réalisées dans le cadre des études sectorielles des produits laitiers ultra-frais et des fruits transformés, ont été utilisés. Pour les produits laitiers, la constitution de chaque échantillon composite tient compte des parts de marché. Pour les fruits transformés, les produits de chaque échantillon sont intégrés à parts égales. L'objectif est d'obtenir une première approche méthodologique de l'origine des écarts constatés : effet nutriment, effet teneur, effet famille, effet segment.

2.1 Représentation graphique

Afin d'observer si les données d'étiquetage et d'analyse sont concordantes, plusieurs graphiques mettant en relation ces deux sources de données sont réalisés pour chaque nutriment.

Sur chaque graphique est représenté l'ajustement linéaire du nuage de points ainsi que la première bissectrice qui représente la concordance parfaite entre les valeurs d'étiquetages et les valeurs analytiques.

Données produits laitiers ultra-frais (PLF)

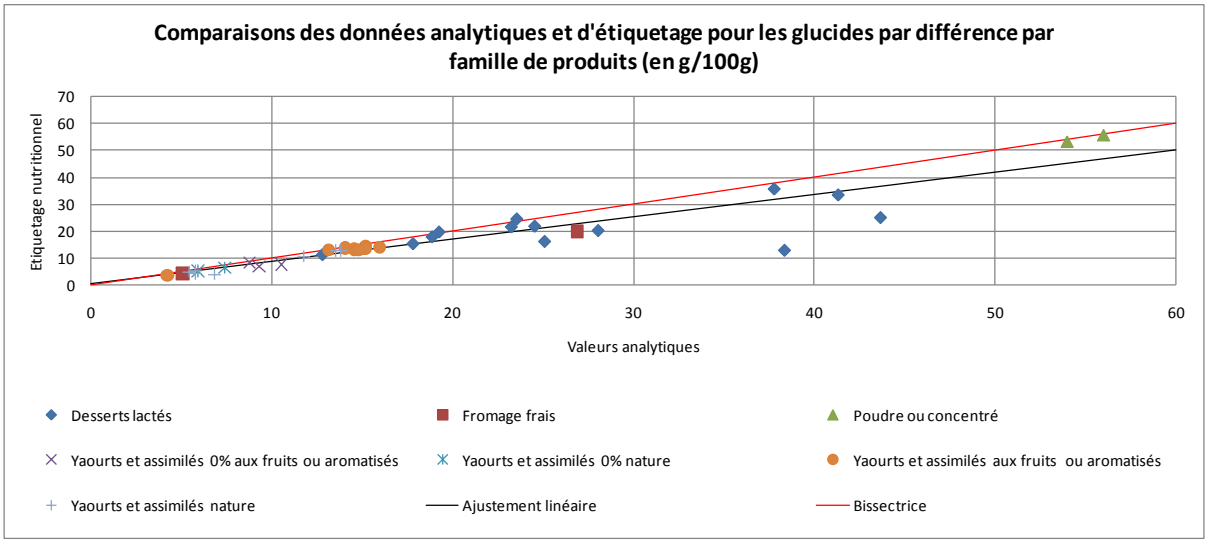

Figure 17 : Valeurs nutritionnelles des glucides par famille de produits pour les produits laitiers ultra-frais (données et nomenclature 2008)

En ce qui concerne les teneurs en glucides des produits laitiers ultra-frais (Figure 17), l'ajustement linéaire des points est satisfaisant. Il est également proche de la première bissectrice, ce qui traduit une bonne concordance. En ce qui concerne les desserts lactés et pour certaines valeurs élevées, les teneurs en glucides semblent être sous-estimées pour l'étiquetage.

Données fruits transformés

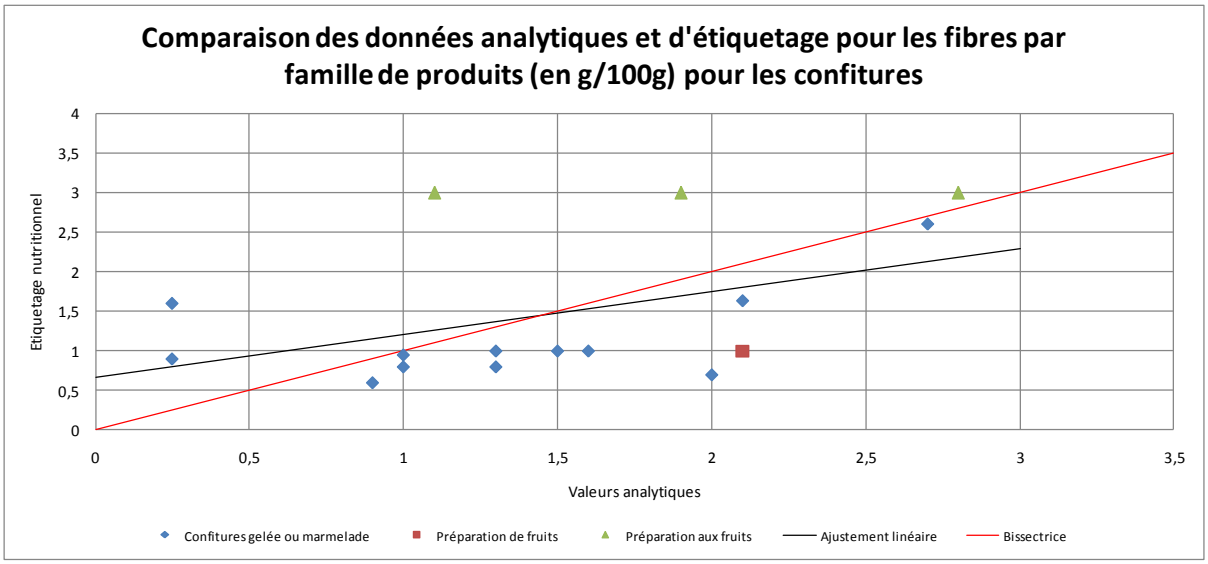

Figure 18 : Valeurs nutritionnelles des fibres par famille de produits pour les confitures

Pour la teneur en fibres dans les confitures (Figure 18), les données ne s'alignent pas et ne sont pas concordantes. Ceci est à relier aux faibles teneurs d'une part, aux méthodologies analytiques d'autre part.

2.2 Écarts relatifs

Afin de synthétiser l'information issue de cette comparaison étiquetage/analyse, il est possible de calculer à chaque fois un écart relatif comme suit :

Ecart relatif = (valeurs d'étiquetage - valeurs analytiques)/valeurs d'étiquetage

Lorsqu'il y a plusieurs produits, un écart relatif moyen est calculé en fonction de la famille de produits et du segment de marché (Tableau 11).

Un écart relatif positif signifie que les valeurs d'étiquetage sont en moyenne supérieures aux valeurs analytiques.

Tableau 11 : Ecarts relatifs des différences entre les valeurs analytiques et d'étiquetage pour les fruits transformés

Moyennes des écarts relatifs par nutriments en fonction de la famille de produits et du segment de marché					
Famille de produits	**Segments de marché**	**NB**	**Glucides**	**Sucres**	**Fibres**
		Compotes			
Compote	Hard Discount	3	-10%	13%	-400%
	Marques de distributeurs-eg	2	-7%	1%	
	Marques de distributeurs	2	-6%	-2%	43%
	Marques nationales	3	-15%	-1%	
Dessert de fruits	Hard Discount	4	-2%	-7%	41%
	Marques de distributeurs	5	-4%	-7%	48%
	Marques nationales	5	3%	-5%	37%
Spécialité de fruits	Hard Discount	1	-5%	-6%	67%
		Confitures			
Confiture gelée ou marmelade	Hard Discount	9	-1%	6%	-60%
	Marques de distributeurs-eg	2	-1%	8%	
	Marques de distributeurs	12	-1%	1%	-28%
	Marques nationales	9	-3%	-1%	-4%
Préparation aux fruits	Marques nationales	3	-23%	-28%	36%
Préparation de fruits	Marques de distributeurs	2	-3%	-2%	-110%
	Marques nationales	1		1%	
		Conserves de fruits			
Fruits au jus de fruits	Marques de distributeurs	1	2%	21%	-13%
	Marques nationales	3	-3%	11%	19%
Fruits au sirop	Marques de distributeurs	4	-11%	7%	-28%
	Marques nationales	4	-16%	-16%	40%
Fruits au sirop léger	Hard Discount	3	-9%		
	Marques de distributeurs-eg	4	-1%	-24%	-38%
	Marques de distributeurs	6	-9%	2%	12%
	Marques nationales	4	-13%	-25%	21%

La valeur nutritionnelle étiquetée s'écarte plus ou moins de la teneur analysée. Toutefois, les écarts relatifs ne semblent pas être influencés par les segments de marché.

Les données d'étiquetage peuvent provenir de calculs de recettes ou de la reprise de valeurs de référence tabulées. Il sera donc souhaitable pour la suite des travaux de connaître leur origine. Lorsque cela n'est pas possible, un modèle de calibration peut être élaboré à partir d'un jeu de données appariées étiquetage/analyse afin d'ajuster les valeurs utilisées dans les comparaisons.

3. CONCEPTION DES PLANS D'ÉCHANTILLONNAGE POUR SUIVRE L'ÉVOLUTION DES CARACTÉRISTIQUES DES PRODUITS

Une des missions principales de l'Oqali étant le suivi de l'évolution dans le temps des caractéristiques des produits, il est important de déterminer le nombre minimum de produits à suivre pour pouvoir mettre en évidence des évolutions significatives, en fonction de la variabilité pour un nutriment donné et d'un pourcentage d'évolution donné.

Pour rappel, les données à suivre peuvent être :
- indépendantes (lorsque les produits aux temps t_0 et t_{+1} ne sont pas les mêmes) ;
- appariées (lorsque les produits aux temps t_0 et t_{+1} sont les mêmes) ;
- semi-appariées (lorsque seulement certains produits sont les mêmes à t_0 et t_{+1}).

Par ailleurs, deux tests peuvent être considérés :
- le test unilatéral (dans le cas où le sens de l'évolution est connu) ;
- le test bilatéral (dans le cas où le sens de l'évolution est inconnu).

Les formules utilisées sont détaillées ci-après[11].

3.1 Taille d'échantillon pour le suivi d'une moyenne

3.1.1 Cas de données indépendantes

- Test unilatéral :

$n = (t_{1-\alpha} + t_{1-\beta})^2 * 2\sigma^2/\Delta^2$ $t_{1-\alpha}$ suit une loi normale centrée réduite

- Test bilatéral :

$n = (t_{1-\alpha/2} + t_{1-\beta})^2 * 2\sigma^2/\Delta^2$ $t_{1-\alpha/2}$ suit une loi normale centrée réduite

σ^2= variance

Δ^2= évolution que l'on souhaite mettre en évidence

$t_{1-\alpha}$=valeur du fractile de la loi normale centrée réduite pour la probabilité α

$t_{1-\beta}$ = valeur du fractile de la loi normale centrée réduite pour la probabilité β

[11] Dussaix A.-M (1987) : Détermination de la taille d'échantillon pour la mesure d'évolutions. *Revue de Statistique Appliquée*, tome 35, n°. 4, p. 25-35.

3.1.2 Cas de données appariées

- Test unilatéral[12] :

$n=2*(1-\rho)*\sigma^2/\Delta^2*\Phi^2$ $\Phi^2=(\varepsilon_{2\alpha}+\varepsilon_{2\beta})$

- Test bilatéral :

$n=2*(1-\rho)*\sigma^2/\Delta^2*\Phi^2$ $\Phi^2=(\varepsilon_{\alpha}+\varepsilon_{2\beta})$

σ^2= variance
Δ^2= évolution que l'on souhaite mettre en évidence
$\varepsilon_{2\alpha}$ = valeur du fractile de la loi normale centrée réduite pour la probabilité α
$\varepsilon_{2\beta}$ = valeur du fractile de la loi normale centrée réduite pour la probabilité β
ρ = coefficient de corrélation linéaire (liaison entre deux mesures d'un nutriment)

3.1.3 Cas de données semi-appariées

- Test unilatéral :

$n=2*(1-k\rho)*\sigma^2/\Delta^2*\Phi^2$ $\Phi^2=(\varepsilon_{2\alpha}+\varepsilon_{2\beta})$

- Test bilatéral :

$n=2*(1-k\rho)*\sigma^2/\Delta^2*\Phi^2$

$\Phi^2=(\varepsilon_{\alpha}+\varepsilon_{2\beta})$

σ^2= variance
Δ^2= évolution que l'on souhaite mettre en évidence
$\varepsilon_{2\alpha}$ = valeur du fractile de la loi normale centrée réduite pour la probabilité α
$\varepsilon_{2\beta}$ = valeur du fractile de la loi normale centrée réduite pour la probabilité β
k=proportion de produits remplacés
ρ = coefficient de corrélation linéaire (liaison entre deux mesures d'un nutriment)

[12] Vray M, Bouvenot G (1994) : Essais cliniques : théorie, pratique et critique. *Médecine-science Flammarion*, p.39-53.

3.2 Paramètres du calcul de la taille d'échantillon pour le suivi d'une moyenne

Quand la taille de l'échantillon est trop faible, une différence, pourtant réelle, peut ne pas être mise en évidence. En revanche, plus l'échantillon est grand, plus il est possible de mettre en évidence une différence, si elle existe.

Plusieurs paramètres sont pris en compte dans le calcul de la taille de l'échantillon : le risque β, le risque α, la moyenne (pour le % d'évolution) et l'écart-type de la population de référence[13] ou la taille d'effet qui permet de pallier un manque d'information sur la moyenne et l'écart-type.

- **Puissance (1-β)**

La puissance est la probabilité de conclure, suite à un test statistique, qu'il y a une différence sur un produit entre deux temps d'observations et que cette différence existe bien en réalité. Par convention, la puissance minimale envisagée dans les tests est de 80%. Cela représente 4 chances sur 5 de conclure à une différence significative entre 2 échantillons alors qu'il en existe bien une. Conclure à l'absence de différence significative alors que la puissance du test serait inférieure à 80% est vraisemblablement incorrect.

Plus la puissance souhaitée est importante, plus la taille d'échantillon requise pour le test est élevée *(d= taille d'effet, expliqué au paragraphe suivant)* (Figure 19, Figure 20).

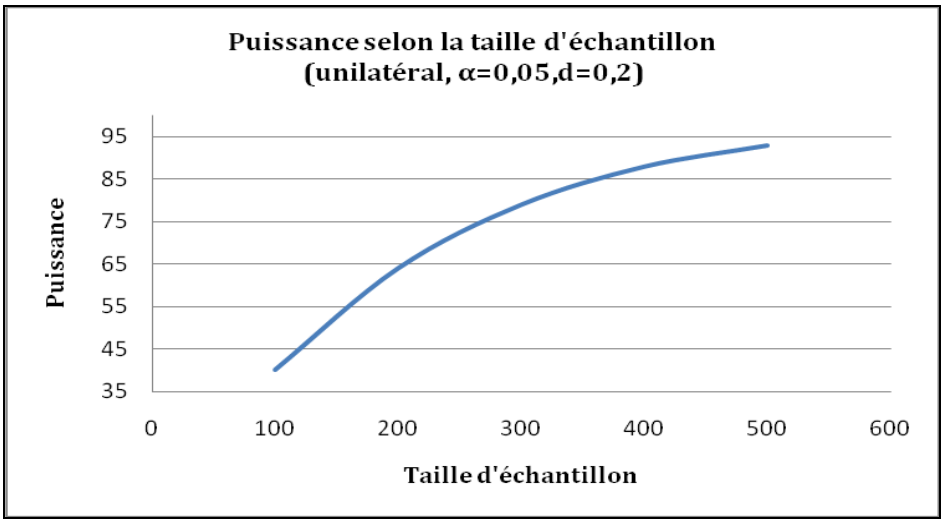

Figure 19 : Evolution de la puissance selon la taille d'échantillon

[13] Champely S (2006) : Puissance des tests paramétriques : Puissance, taille d'effet et taille d'échantillon (sous R).

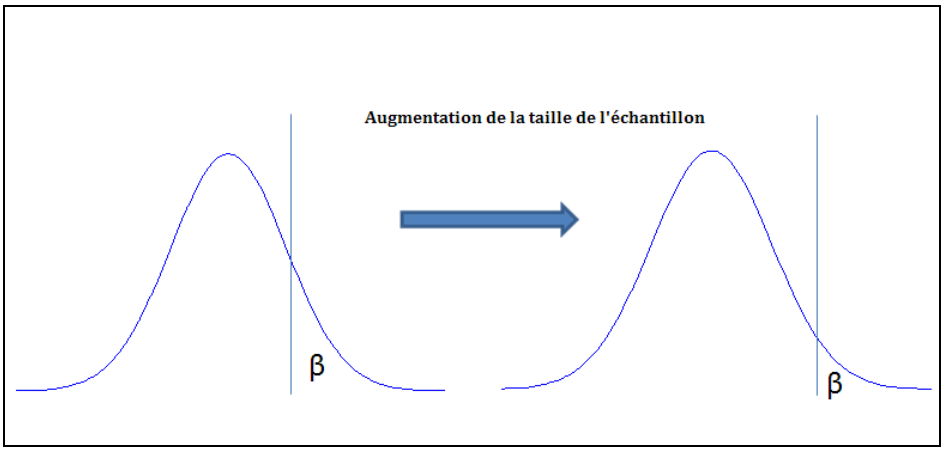

Figure 20 : Evolution du paramètre β selon la taille d'échantillon

- **Taille d'effet**

L'évolution que l'on souhaite mettre en évidence entre les deux temps d'observations est également prise en compte pour le calcul de la taille de l'échantillon. Pour cela, il est possible de rechercher dans d'autres études ou dans des résultats précédents la moyenne et l'écart-type d'une population similaire. Plus l'écart-type est faible par rapport à la moyenne, plus la taille de l'échantillon sera petite, ce qui sous entend moins de produits pour mettre en évidence une différence (Tableau 12).

Tableau 12 : Taille d'échantillon requise selon le pourcentage d'évolution à détecter et le coefficient de variation de la variable étudiée pour une puissance de 80% et un risque alpha de 5% (test unilatéral)

Effet / CV	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	15%	20%	25%	30%	35%	40%	50%	60%	70%	80%	90%	100%
1%	14	51	112	199	310	446	607	793	1003	1238	2783	4947	7729	11130	15148	19785	30914	44516	60590	79138	100159	123652
2%		14	29	51	78	112	153	199	252	310	697	1238	1933	2783	3788	4947	7729	11130	15148	19785	25041	30914
3%		7	14	23	36	51	69	89	112	139	310	551	860	1238	1684	2199	3436	4947	6733	8794	11130	13740
4%			8	14	21	29	39	51	64	78	175	310	484	697	948	1238	1933	2783	3788	4947	6261	7729
5%			6	9	14	19	25	33	41	51	112	199	310	446	607	793	1238	1782	2425	3167	4007	4947
6%				7	10	14	18	23	29	36	78	139	216	310	422	551	860	1238	1684	2199	2783	3436
7%				5	8	10	14	17	22	26	58	102	159	228	310	405	632	910	1238	1616	2045	2525
8%					6	8	11	14	17	21	45	78	122	175	238	310	484	697	948	1238	1566	1933
9%					5	7	9	11	14	16	36	62	97	139	188	245	383	551	749	978	1238	1528
10%						6	7	9	11	14	29	51	**78**	112	153	199	310	446	607	793	1003	1238
15%								5	6	7	14	23	36	51	69	89	139	199	270	353	446	551
20%											8	14	21	29	39	51	78	112	153	199	252	310
25%											6	9	14	19	25	33	51	72	98	128	161	199
30%												7	10	14	18	23	36	51	69	89	112	139
35%												5	8	10	14	17	26	38	51	66	83	102
40%													6	8	11	14	21	29	39	51	64	78
45%													5	7	9	11	16	23	31	40	51	62
50%														6	7	9	14	19	25	33	41	51

Par exemple, 78 produits sont nécessaires pour détecter une différence de 10% en moyenne entre t0 et t_{+1} lorsque la variable étudiée a un coefficient de variation de 25%.

La figure 22 présente le pourcentage d'évolution que l'on peut mettre en évidence selon la taille de l'échantillon pour un test unilateral.

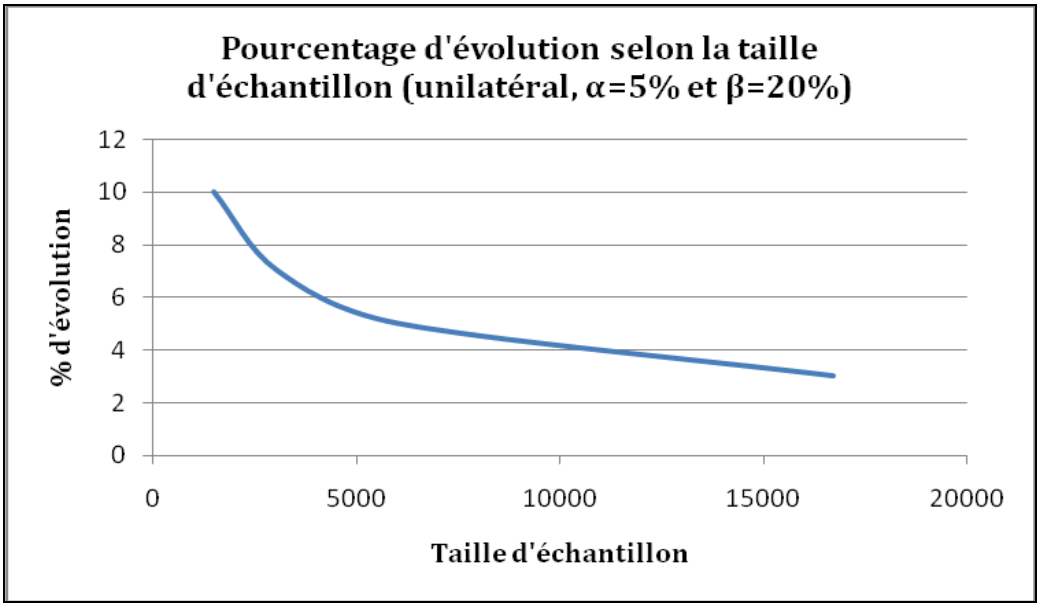

Figure 21 : Pourcentage d'évolution observable en fonction d'une taille d'échantillon pour un α et β donné

A défaut d'avoir une moyenne et un écart-type, il est possible d'avoir recours à la taille d'effet de Cohen[14]. La taille d'effet peut prendre 3 valeurs : 0,2 ; 0,5 ; 0,8. Pour mettre en évidence une petite différence entre les deux moments d'observations une taille d'effet de 0,2 sera utilisée. Si, au contraire, une grande différence veut être mise en évidence, la taille d'effet de 0,8 est plus adéquate (Tableau 13).

$$n = (t_{1-\alpha} + t_{1-\beta})^2 / \mathbf{d}$$

$d = (\mu - \mu_0)/\sigma$= taille d'effet

$t_{1-\alpha}$ =valeur du fractile d'α dans la table statistique de la loi normale centrée réduite

$t_{1-\beta}$ =valeur du fractile de β dans la table statistique de la loi normale centrée réduite

[14] Champely S (2006) : Puissance des tests paramétriques : Puissance, taille d'effet et taille d'échantillon (sous R).

Tableau 13 : Taille d'échantillon pour un α et un β donnés selon la taille d'effet de Cohen (test unilatéral)

Taille d'effet = 0,2		Beta		
		0,05	**0,1**	**0,2**
Alpha	**0,01**	788	651	502
	0,05	541	428	309
	0,1	428	329	226

Taille d'effet = 0,5		Beta		
		0,05	**0,1**	**0,2**
Alpha	**0,01**	126	104	80
	0,05	87	69	49
	0,1	69	53	36

Taille d'effet = 0,8		Beta		
		0,05	**0,1**	**0,2**
Alpha	**0,01**	49	41	31
	0,05	34	27	19
	0,1	27	21	14

3.3 Calcul du nombre de produits minimum

3.3.1 Cas de données indépendantes

Ce cas se présente lorsque les produits ne sont pas les mêmes à t_0 et t_{+1}.
Avec un α=5% et une puissance 1-β=80%, les résultats suivants sont obtenus (Tableau 14) :

La formule pour le calcul de la taille d'échantillon est la suivante :

$$\mathbf{n = (t_{1-\alpha} + t_{1-\beta})^2 * 2\sigma^2/\Delta^2}$$

σ^2= variance
$t_{1-\alpha}$ =valeur du fractile de la loi normale centrée réduite pour la probabilité α
Δ^2= évolution que l'on souhaite mettre en évidence
$t_{1-\beta}$ = valeur du fractile de la loi normale centrée réduite pour la probabilité β

Tableau 14 : Taille d'échantillon pour données indépendantes (unilatéral)

	% évolution	Nb (σ=5,48)	Nb sans valeur extrême (σ=1,35)
protéines	3	16710	3562
protéines	5	6016	1282
protéines	7	3069	654
protéines	10	1504	321

Il est important de pouvoir identifier les valeurs extrêmes et les traiter séparément car cela a des répercussions sur la moyenne et l'écart-type et donc sur la taille de l'échantillon.

3.3.2 Cas de données appariées

L'utilisation de l'appariement permet d'améliorer la puissance d'un test. Cela a une influence sur la variance résiduelle (erreur de mesure) du paramètre à mesurer.
Afin de calculer la taille d'échantillon pour les divers nutriments à analyser, il est nécessaire, pour l'appariement, de connaître le coefficient de corrélation (la liaison entre la valeur d'un nutriment mesuré dans deux échantillons) entre les données aux temps t_0 et t_{+1} (une recherche bibliographique peut être effectuée pour essayer de l'estimer). A défaut de le connaître, plusieurs hypothèses sont testées en fonction de la valeur du coefficient de corrélation.

Cas du panel : échantillons identiques à t_0 et t_{+1}

Dans cette situation la taille d'échantillon correspond à la taille d'échantillon trouvée dans le cas des données indépendantes multipliée par la valeur de (1-ρ) où ρ=coefficient de corrélation linéaire de la population entre la valeur du nutriment au temps t_0 et la valeur du nutriment au temps t_{+1}.

La formule pour le calcul de la taille d'échantillon est la suivante : **$n' = n(1-\rho)$.**

n=taille d'échantillon pour données indépendantes
ρ=coefficient de corrélation linéaire[15]

[15] La liaison entre deux mesures d'un nutriment.

Tableau 15 : Tailles d'échantillons pour données appariées sans valeurs extrêmes (unilatéral)

	Beta	Alpha	% évolution	ρ=0,9	ρ=0,8	ρ=0,7	ρ=0,6	ρ=0,5
protéines	0,2	0,05	3	356	712	1069	1425	1781
protéines	0,2	0,05	5	128	256	385	513	641
protéines	0,2	0,05	7	65	131	196	262	327
protéines	0,2	0,05	10	32	64	96	128	160

Plus les produits appariés sont semblables (en fonction des critères d'appariement), plus la taille de l'échantillon est réduite (Tableau 15).

3.3.3 Cas de données semi-appariées

Un certain nombre de produits est analysé à t_0. Lors de la 2ème phase d'analyses à $t_{+1,}$ certains produits analysés à t_0 n'existent plus. Pour éviter d'avoir des données manquantes, les produits vont être remplacés par des produits qui ont des caractéristiques très proches (cet aspect doit être défini au début de l'enquête lors de l'élaboration du protocole).

Pour calculer la taille d'échantillon dans le cas de données partiellement renouvelées une formule existe. Elle nécessite de connaître la taille des échantillons nécessaire dans le cas de données indépendantes pour une évolution, un α et une puissance souhaités. Il est aussi important de connaître ou de pouvoir estimer le coefficient de corrélation entre les données entre t_0 et t_{+1}. Enfin, il est nécessaire d'avoir la proportion de produits qu'il faudra remplacer à t_{+1}.

La formule pour le calcul de la taille d'échantillon est la suivante : **$n' = n(1-k\rho)$**

n=taille d'échantillon pour données indépendantes
k=proportion de produits remplacés[16]
ρ=coefficient de corrélation linéaire[17]

[16] Produit non trouvé au 2ème prélèvement qui remplace un produit défini lors de l'appariement.
[17] La liaison entre deux mesures d'un nutriment.

Tableau 16 : Tailles des échantillons pour données semi-appariées, sans valeurs extrêmes, pour un coefficient de corrélation de 90% (unilatéral)

	% évolution	Nb étiquetage*	k=3%	k= 5%	k=7%	k =10%	k= 15%	k=20%
protéines	3	3562	452	516	581	677	837	997
protéines	5	1282	163	186	209	244	301	359
protéines	7	654	83	95	107	124	154	183
protéines	10	321	41	47	52	61	75	90

*taille de l'échantillon pour des données indépendantes n'=n (1-kρ)

Tableau 17 : Tailles d'échantillons pour données semi-appariées, sans valeurs extrêmes, pour un coefficient de corrélation de 80% (unilatéral)

	% évolution	Nb étiquetage*	k=3%	k= 5%	k=7%	k =10%	k= 15%	k=20%
protéines	3	3562	798	855	912	997	1140	1282
protéines	5	1282	287	308	328	359	410	462
protéines	7	654	146	157	167	183	209	235
protéines	10	321	72	77	82	90	103	116

*taille de l'échantillon pour des données indépendantes n'=n (1-kρ)

Tableau 18 : Tailles d'échantillons pour données semi-appariées, sans valeurs extrêmes, pour un coefficient de corrélation de 70% (unilatéral)

	% évolution	Nb étiquetage*	k=3%	k= 5%	k=7%	k =10%	k= 15%	k=20%
protéines	3	3562	1143	1193	1243	1318	1443	1567
protéines	5	1282	412	429	447	474	519	564
protéines	7	654	210	219	228	242	265	288
protéines	10	321	103	108	112	119	130	141

*taille de l'échantillon pour des données indépendantes n'=n (1-kρ)

Plus le taux de remplacement de produits est important, plus la taille de l'échantillon requise est grande. L'appariement partiel reste préférable à l'utilisation de données indépendantes (Tableau 16, Tableau 17, Tableau 18).

Lexique

Allégation

Tout message ou toute représentation, non obligatoire en vertu de la législation communautaire ou nationale, y compris une représentation sous la forme d'images, d'éléments graphiques ou de symboles, qu'elle qu'en soit la forme, qui affirme, suggère ou implique qu'une denrée alimentaire possède des caractéristiques particulières.

Allégation de santé

Toute allégation qui affirme, suggère ou implique l'existence d'une relation entre, d'une part, une catégorie de denrées alimentaires, une denrée alimentaire ou l'un de ses composants et, d'autre part, la santé. Il en existe deux types : les allégations de santé fonctionnelles (relatives à l'article 13 du règlement (CE) n° 1924/2006) et les allégations de santé relatives à la réduction d'un risque de maladie ou se rapportant au développement et à la santé infantiles (relatives à l'article 14 du règlement (CE) n° 1924/2006).

Les allégations de santé fonctionnelles sont des allégations qui décrivent ou mentionnent :

- le rôle d'un nutriment ou d'une autre substance dans la croissance, dans le développement et dans les fonctions de l'organisme ;
- les fonctions psychologiques ou comportementales ;
- l'amaigrissement, le contrôle du poids, une réduction de la sensation de faim, l'accentuation de la sensation de satiété ou la réduction de la valeur énergétique du régime alimentaire.

Allégation nutritionnelle

Toute allégation qui affirme, suggère ou implique qu'une denrée alimentaire possède des propriétés nutritionnelles bénéfiques particulières de par l'énergie (valeur calorique) qu'elle: i) fournit, ii) fournit à un degré moindre ou plus élevé, ou iii) ne fournit pas, et/ou de par les nutriments ou autres substances qu'elle : i) contient, ii) contient en proportion moindre ou plus élevée, ou iii) ne contient pas.

En particulier, dans les rapports sectoriels effectués par l'Oqali, ont été considérées comme « allégations nutritionnelles » toutes les allégations remplissant les conditions d'utilisation de l'annexe du règlement (CE) n°1924/2006 du parlement européen actuellement en vigueur.

Autre allégation nutritionnelle

L'ensemble des allégations nutritionnelles actuellement non listées dans l'annexe du règlement (CE) n°1924/2006 mais présentes dans la proposition d'amendement de cette annexe par les membres de la Confédération des Industries Agro-alimentaires de l'Union Européenne.

Etiquetage nutritionnel

Toute information apparaissant sur l'étiquette relative à la valeur énergétique et aux nutriments suivants : protéines, glucides, lipides, fibres alimentaires, sodium, vitamines et sels minéraux (énumérés à l'annexe de la directive 90/496/CEE du Conseil, lorsqu'ils sont présents en quantité significative conformément à ladite annexe). La règlementation prévoit deux groupes d'étiquetage :

- **le groupe 1** : présence de la valeur énergétique et des valeurs nutritionnelles pour les protéines, les glucides et les lipides ;

- **le groupe 2** : présence de la valeur énergétique et des valeurs nutritionnelles pour les protéines, les glucides, les sucres, les lipides, les acides gras saturés, les fibres alimentaires et le sodium.

Dans les rapports sectoriels publiés par l'Oqali, des groupes d'étiquetage supplémentaires ont été pris en compte :

- **groupe 0 :** absence de valeurs énergétiques et nutritionnelles ;
- **groupe 0+ :** présence de la valeur énergétique ou des valeurs nutritionnelles pour une partie des nutriments du groupe 1 et/ou pour des micronutriments, selon les spécificités règlementaires de certains secteurs ;
- **groupe 1** : présence de la valeur énergétique et des valeurs nutritionnelles pour les protéines, les glucides et les lipides ;
- **groupe 1+** : présence de l'étiquetage du groupe 1 ainsi que l'étiquetage relatif aux qualités nutritionnelles d'un ou de plusieurs des éléments suivants : l'amidon, les polyols, les acides gras mono-insaturés, les acides gras polyinsaturés, le cholestérol, sels minéraux ou vitamines ;
- **groupe 2** : présence de la valeur énergétique et des valeurs nutritionnelles pour les protéines, les glucides, les sucres, les lipides, les acides gras saturés, les fibres alimentaires et le sodium ;
- **groupe 2+ :** présence de l'étiquetage du groupe 2 comprenant également l'étiquetage relatif aux qualités nutritionnelles d'un ou de plusieurs des éléments suivants : l'amidon, les polyols, les acides gras mono-insaturés, les acides gras polyinsaturés, le cholestérol, sels minéraux ou vitamines.

Famille de produits

Entité la plus fine sur laquelle sont réalisés les traitements. Les produits peuvent être regroupés au sein d'une même famille selon différents critères : la dénomination de vente, la technologie de fabrication, la recette, le positionnement marketing…

Incitations à l'activité physique

Dans les rapports sectoriels publiés par l'Oqali, les incitations à l'activité physique rassemblent tous les messages du type « l'activité physique est indispensable pour votre forme et votre vitalité, pensez à bouger au moins 30 minutes chaque jour ».

Portion indiquée

Les portions indiquées regroupent :

- les portions clairement inscrites dans une recommandation de consommation ;
- les portions figurant dans le tableau nutritionnel lorsque les valeurs nutritionnelles pour une portion différente de 100g sont exprimées.

Portions individuelles

Taille d'un sachet fraicheur ou d'un paquet individuel présent dans un même emballage. Une portion individuelle peut correspondre à une unité de produit (cas des yaourts par exemple) ou à plusieurs unités de produit (cas des pochons individuels de biscuits secs pour le petit-déjeuner).

Produit

Pour l'Oqali, un produit correspond à une référence commercialisée et enregistrée dans la base. Il peut être identifié par un certain nombre de critères (le nom commercial, la marque, le code barre, la dénomination de vente, …).

Recommandations de consommation

Ce sont toutes les recommandations relatives à l'accompagnement conseillé dans le cadre d'un repas équilibré (petit-déjeuner, déjeuner, goûter, apéritif, dîner). Généralement, elles informent le consommateur sur l'intégration du produit étudié dans une alimentation équilibrée mais peuvent également fournir des informations de base sur l'alimentation et la nutrition. Par exemple, des recommandations nutritionnelles générales sont du type : « nombre de portions recommandées par jour : au moins 5 portions de fruits et légumes ; 6 portions de pain, pâtes, riz, légumes secs ; 1 à 2 portions de viandes, poissons, œufs ; 3 produits laitiers ».

Repères nutritionnels

Les repères nutritionnels pris en compte dans le cadre de l'Oqali rassemblent toutes les icônes de type % des RNJ (Repères Nutritionnels Journaliers), % des ANC (Apports Nutritionnels Conseillés), cadrans, cartouches, curseurs, échelles, nutri-pass ou camembert présentes sur l'emballage du produit. Ils symbolisent l'apport en kcal et/ou en nutriments d'une portion donnée du produit pour un type de consommateur (par exemple, adulte dont les besoins journaliers sont de 2000 kcal).

Secteur

Un secteur regroupe des familles de produits homogènes entre elles selon un ou plusieurs critères, notamment l'ingrédient principal (ex. lait pour les produits laitiers, cacao pour les produits chocolatés), le moment de consommation (ex. l'apéritif pour le secteur des apéritifs à croquer),… Dans le cadre de l'Oqali, les études sont menées par secteur alimentaire.

Segment de marché

Pour tous les traitements réalisés dans les études sectorielles, chaque secteur a été divisé en 3 segments de marché :

- marques nationales (ou MN) : ce sont les produits de marque ;
- marques de distributeurs (ou MDD) : ce sont les produits à marques d'enseignes de la distribution et dont les caractéristiques ont été définies par les enseignes qui les vendent au détail ;
- marques hard discount (ou HD) : ce sont les produits vendus uniquement en magasin hard discount.

Une ventilation plus fine et au cas par cas a pu être définie au sein de chaque rapport sectoriel, afin de distinguer éventuellement les produits en gammes :

- cœur de marché (ou cm) : gamme par défaut ;
- entrée de gamme (ou eg) : produits souvent caractérisés par un prix moins élevé que la moyenne de la catégorie. Ils ont généralement un nom qui rappelle le fait d'être les produits les moins chers de la catégorie ;
- haut de gamme (ou hg) : produits le plus souvent caractérisés par un prix plus élevé que la moyenne de la catégorie. Peuvent appartenir à cette catégorie, par exemple, les produits issus de l'agriculture biologique.

Cette segmentation plus fine permet de distinguer jusqu'à 9 segments de marché.

Valeurs nutritionnelles à la portion

Les valeurs nutritionnelles à la portion correspondent aux valeurs nutritionnelles présentes dans le tableau nutritionnel pour une portion donnée (portion individuelle et/ou portion indiquée), en complément des valeurs nutritionnelles aux 100g.