Le nombre de sujets dans les panels d'analyse sensorielle : une approche base de données

Nadra Mammasse

UNIVERSITÉ DE BOURGOGNE

Centre des Sciences du Goût et de l'Alimentation

THÈSE

Pour obtenir le grade de

Docteur de l'Université de Bourgogne

Discipline :

Sciences de l'alimentation

par

Nadra MAMMASSE

le 22 Mars 2012

Le nombre de sujets dans les panels d'analyse sensorielle : une approche base de données

Thèse soutenue publiquement à Dijon, devant le jury composé de :

Pr Evelyne VIGNEAU	Oniris, Nantes, France	Rapporteur
Pr Jean-François MEULLENET	Université de l'Arkansas, USA	Rapporteur
Pr François HUSSON	Agrocampus Ouest, Rennes, France	Examinateur
Pr Hervé CARDOT	Université de Bourgogne, France	Examinateur
Dr Pascal SCHLICH	INRA, CSGA, Dijon, France	Directeur de thèse
Mme Virginie HERBRETEAU	Les Maisons du Goût, Actilait, Rennes, France	Membre invité

Rien ne devient jamais réel tant qu'on ne l'a pas ressenti.

John Keats

Remerciements

Ce projet de thèse m'a permis avant tout de découvrir et d'adopter l'analyse sensorielle. Il m'a aussi permis de découvrir mes capacités à mener un projet jusqu'à la fin, de cerner mes limites, d'apprécier les petites victoires et de découvrir la force du travail en équipe.

Mes sincères remerciements vont à Pascal Schlich pour la confiance qu'il m'a accordée et pour m'avoir permis d'effectuer cette thèse. Je te remercie pour ton ouverture d'esprit, pour ta volonté de valoriser et de mettre en avant l'ensemble de nos travaux en participant à différents congrès scientifiques, nationaux et internationaux.

Merci aux centres ACTIA et au conseil régional de Bourgogne, financeurs de ce travail de thèse.

Merci à Evelyne Vigneau, Jean-François Meullenet et François Husson pour avoir accepté de lire, critiquer et évaluer mon manuscrit. Merci pour vos remarques constructives qui ont apporté un éclairage nouveau à mes travaux.

Je remercie vivement El Mostapha Qannari et Hervé Cardot pour leur disponibilité et pour l'intérêt qu'ils ont porté à cette thèse. Merci pour votre soutien, vos encouragements et vos précieux conseils.

Merci à Virginie Herbreteau d'avoir coordonné ce projet. Merci pour ces échanges constructifs et pour toutes les réponses à mes questions sur l'analyse sensorielle.

Merci à tous les membres du Réseau Mixte Technologique RMT Sensorialis et représentants des Centres ACTIA. Merci pour m'avoir accueillie et pour votre implication dans ce projet.

Je remercie "la super team SensoBase et PrefBase" Sylvie Cordelle et Michel Visalli pour m'avoir accompagnée tout au long de ce projet. Merci d'avoir toujours répondu présents et avec enthousiasme à toutes mes questions autour de cette grande masse de données. Merci pour vos conseils et votre précieux soutien pendant les "moments forts"...

Merci à la joyeuse bande du "Ex Liris"
Merci à toi Coralita pour ta disponibilité et ta générosité sans limite, à toi Kris pour ton écoute et tes vifs encouragements. Merci à Marcélita, Amélita, Julita et Laurita pour ces bons moments

qu'on a partagé et de m'avoir soutenu jusqu'au bout. Merci à Sophita, Audita, Audrita et Éric pour votre énergie positive et très communicative. Merci à vous deux Marine et Anaïs pour votre soutien et la précieuse aide pendant toute la période de rédaction. Merci à Élodie et à Béné pour nos fou-rires et pour la médaille d'or. Merci Christine U. et Caroline L. pour votre écoute et vos encouragements. Merci Catherine et Betty pour votre agréable compagnie. Merci aux "Stateuses" : Caroline P., Carole M. et Guillauman. Merci aux deux cinéphiles Carole S. et Syrina pour les bons moments des salles obscures. Grand merci à Mitch, Céd et Sylvain pour votre bonne humeur quotidienne.

Merci à tous mes amis(es) qui de près ou de loin ont toujours su trouver les bons mots pour m'encourager et me soutenir.

Ma gratitude ne saurait bien évidemment passer à côté de mes parents et de mes frères et soeurs dont le soutien et l'encouragement m'ont permis de mener à bien ce projet. Merci Yemma & Baba pour l'intérêt que vous avez toujours porté à mes études. Merci Hassiba, Sabrina, Hamidouche, Badria, Abdel & Zakari.

Merci à mes oncles et tantes pour leur très précieuse aide et leur présence à mes cotés. Merci à tous mes proches.

Merci à ceux que j'aurais pu oublier ...

Résumé

Le nombre de sujets du panel détermine en grande partie le coût des études descriptives et hédoniques de l'analyse sensorielle. Une fois les risques α et β fixés, ce nombre peut théoriquement être calculé, dés lors que l'on connaît la variabilité de la mesure due à l'hétérogénéité de la population visée et que l'on fixe la taille de la différence que l'on désire mettre en évidence. En général, l'ordre de grandeur du premier de ces paramètres est inconnu alors que celui du second est délicat à préciser pour l'expérimentateur. Ce travail propose une documentation systématique des valeurs prises dans la réalité par ces deux paramètres grâce à l'exploitation de deux bases de données, SensoBase et PrefBase, contenant respectivement un millier de jeux de données descriptives et quelques centaines de jeux de données hédoniques. Pratiquement, des recommandations pour la taille de panel sont établies sous forme d'abaques prenant en compte trois niveaux pour chacun des deux risques et des deux paramètres.

D'autre part, ce travail étudie le nombre de sujets dans chacun des deux types de panel par une approche de ré-échantillonnage qui consiste à réduire progressivement le nombre de sujets tant que les résultats de l'analyse statistique demeurent stables. En moyenne, la taille des panels descriptifs pourrait être réduite d'un quart du nombre de sujets, mais cette moyenne cache une forte hétérogénéité selon le type de descripteurs considéré. La taille optimale des panels hédoniques serait elle très variable et cette variabilité est induite beaucoup plus par la nature et l'importance des différences entre les produits que par l'hétérogénéité des préférences individuelles. De plus, une même approche de ré-échantillonnage appliquée au répétitions en tests descriptifs suggère que les répétitions ne sont plus nécessaires en phase de mesure, c'est-à-dire une fois le panel entraîné.

Abstract

The costs associated with sensory evaluation increase with the number of panelists to be enrolled. Classical power computation can be used to derive the minimal number of subjects of a sensory panel in order to control both type I (α risk) and type II (β risk) errors. However, this power computation requires estimates of the size of the product effect to be sought and of the residual variability of the ANOVA model used. Generally, both product effect size and residual variability are difficult to estimate a priori by the sensory analyst. This work offers estimations of these two parameters thanks to the analysis of hundreds descriptive and hedonic studies collected respectively in two databases, SensoBase and PrefBase. The meta-analysis of the data allowed to quantify these two parameters and made possible the calculation of the number of panelists. Hence, tables of panel sizes were proposed for 3 levels of respectively product effect size, residual variability and type I and II errors. Of course, this was done independently for descriptive and hedonic tests.

Another approach based on resampling in numerous data sets was applied for both descriptive and hedonic studies. The method used to derive adequate panel size consisted in removing k subjects from the N of the original panel and then measuring the loss of information in product comparisons. For descriptive panels, panel size could be reduced by a quarter but this reduction strongly depends on the type of attributes. For hedonic panels, panel sizes varied extremely and depended mainly on the size of the liking differences between products to be compared. We expect that this difference is directly affected by the level of sensory complexity of the products. Finally, the resampling approach was applied to examine the need to replicate with trained sensory panels. Results suggested that replicates are no longer necessary at the testing phase, that is once the panel is trained.

Valorisation des travaux de recherche

Publications avec comité de lecture scientifique

- N. Mammasse(*), S. Cordelle, P. Schlich. No need to replicate with trained sensory panels. Soumis. Food Quality and Preference. Mars 2012.

- N. Mammasse(*), P. Schlich. Adequate number of consumers in a liking test. Insights from resampling in seven studies. Food Quality and Preference, Accepté le 23 Janvier 2012.

- N. Mammasse(*), P. Schlich. The number of assessors in descriptive sensory panels. A database approach. Proceedings du 11ème Congrès Agrostat, Benevento, Italie. 23-28 Février 2010.

Communications orales et congrès internationaux

- N. Mammasse(*), S. Cordelle, P. Schlich. Do we need replications in sensory profiling? 9th Pangborn Sensory Science symposium. 4-8 September 2011, Toronto, Canada.

- N. Mammasse(*), S. Cordelle, P. Schlich. Les répétitions ne sont pas toujours nécessaires en profil sensoriel. Troisième journée annuelle de l'analyse sensorielle, Sens&Co. 30 Mars 2011, AgroParistech, Paris. France.

- N. Mammasse(*), P. Schlich. The right number of consumers in a liking test depends mainly on the product complexity space. 10th Sensometrics symposium. 25-28 July 2010, Rotterdam, The Netherlands.

- N. Mammasse(*), P. Schlich. The number of assessors in descriptive sensory panels. A database approach. 11th European Symposium on Statistical Methods for the Food Industry. 23-26 February 2010, Benevento, Italy.

Table des matières

Liste des tableaux

Table des figures

Introduction Générale

Face à la dynamique du marché des biens de grande consommation et donc aux contraintes de la concurrence, les exigences du consommateur ont pris une place importante dans le choix et l'élaboration des stratégies d'entreprise. La satisfaction des consommateurs garantit le réachat, et donc le profit pour les entreprises agroalimentaires. Dans cet objectif, elles dépensent une partie de leur budget pour évaluer l'opinion des consommateurs sur leurs produits. Cette stratégie axée sur la satisfaction des clients a un coût mais demeure cependant indispensable. Parmi les méthodes utilisées pour évaluer l'appréciation des produits par les consommateurs figure l'analyse sensorielle.

L'analyse sensorielle est une approche qui consiste à étudier les caractéristiques sensorielles des produits en faisant intervenir l'homme comme instrument de mesure à travers ses cinq sens : l'odorat, le goût, la vue, l'ouïe et le toucher. Elle permet d'étudier ou de répondre à diverses questions autour du développement des produits. Elle est essentielle lors de la mise au point de nouveaux produits, pour définir la formulation idéale, choisir les modes de fabrication optimaux, comparer les caractéristiques obtenues à celles des produits concurrents. Originellement dédiée aux industries agroalimentaires, cette discipline a très vite été adoptée par l'industrie cosmétique et plus récemment par l'industrie automobile.

L'analyse sensorielle se subdivise en deux épreuves principales : les tests hédoniques qui mesurent les préférences des consommateurs d'une part, et l'analyse descriptive des différences de perception sensorielle entre les produits d'autre part.
Il est important de bien dissocier ces deux approches car les buts et les méthodes sont différents. Les sujets qui sont impliqués ne sont pas les mêmes : les épreuves descriptives font appel à des sujets entraînés à l'évaluation des produits alors que les épreuves hédoniques exigent des sujets novices, dits naïfs.

Le nombre de sujets en analyse sensorielle détermine en partie la puissance statistique de la comparaison de produits et ainsi la capacité du test d'analyse sensorielle d'établir l'existence de

différences perceptibles entre les produits. Il est théoriquement possible de calculer le nombre de sujets, pourvu que l'on connaisse la variabilité de la mesure mais aussi la taille de la différence que l'on désire mettre en évidence. Cependant, selon le type de l'étude sensorielle, ces deux paramètres demeurent parfois méconnus ou complexes à définir.
Par ailleurs, le nombre de sujets a un impact sur le coût de l'expérimentation. C'est pourquoi il serait intéressant de préciser le nombre minimum de sujets assurant un coût acceptable des tests sensoriels tout en préservant la précision des résultats de comparaison des produits. Cette question est d'autant plus importante pour les tests hédoniques qui exigent une centaine de sujets, voire plus, selon les objectifs de l'expérimentation.

Les recommandations sur le nombre de sujets à enrôler dans un panel d'analyse sensorielle varient beaucoup dans la littérature et sont souvent peu objectives. Par ailleurs, les normes existantes dans le domaine demeurent ambiguës. Souvent, le nombre de sujets n'est abordé que sous l'angle de comparaison de panels de tailles différentes utilisant des techniques différentes. On trouve aussi des recommandations valables pour un contexte donné, voire basées sur une seule expérimentation. Nous n'avons trouvé qu'une seule tentative en tests hédoniques basée sur une centaine de jeux de données (Hough et al., 2006). En effet, cette étude offre pour la première fois des recommandations sur le nombre de sujets à enrôler en tests hédoniques utilisant une estimation à la fois de la variabilité des données mais aussi de la taille de la différence à mettre en évidence. Les résultats avancés nous semblent très intéressants et méritent d'être consolidés par d'autres approches sur un plus grand nombre de jeux de données.

Le besoin de documenter les pratiques de l'analyse sensorielle a conduit à la mise en place d'une première base de données (SensoBase) au sein du Centre des Sciences du Goût et de l'Alimentation (CSGA). Ce projet a permis de construire un outil de suivi de l'évolution de la performance d'un panel au cours du temps. La dimension et l'envergure de SensoBase a permis d'étudier les facteurs de la qualité de la mesure sensorielle et de comparer à grande échelle les performances des sujets. Ces travaux ont été réalisés dans le cadre de la thèse de Pineau (2006). L'étude de la problématique du nombre de dégustateurs d'un panel d'analyse sensorielle figure parmi les perspectives de ce travail. En effet, ce sujet intéresse fortement les professionnels de ce domaine.
Les centres ACTIA (Association des Centres Techniques des Industries Agroalimentaires) ainsi que la région Bourgogne, parties intégrantes du projet SensoBase, ont renouvelé leur engagement auprès du CSGA pour la mise en place d'un nouveau projet autour des études hédoniques. En effet, une nouvelle base de données nommée PrefBase (Visalli et al., 2008) a été mise en place et a pour objectif de réunir les jeux de données de préférence.

Dans ce contexte, l'objectif principal de cette thèse est de quantifier les paramètres qui conditionnent le nombre de sujets, tels que la taille de l'effet et la variabilité de la mesure

sensorielle. Son originalité réside alors dans l'exploitation de centaines de jeux de données issus de SensoBase et PrefBase. Le but est de pouvoir adresser des recommandations quant au nombre de sujets à enrôler pour les épreuves hédoniques mais aussi pour les études de profil sensoriel.

Ce manuscrit est organisé en cinq chapitres.

Le premier chapitre est consacré à l'état des lieux des connaissances en matière de taille de panel sur les épreuves hédoniques et descriptives en analyse sensorielle. Il comprend une synthèse bibliographique sur le nombre de sujets dans les deux types d'épreuves. Enfin, il présente un rappel sur les principes des tests statistiques, de la puissance d'un test, du calcul du nombre de sujets nécessaires et des précisions de la notion de la taille de l'effet. De cette revue bibliographique découleront les questions de recherche et la problématique de ce travail de thèse, présentées dans le deuxième chapitre.

Les chapitres 3 et 4 sont consacrés respectivement à l'approche base de données en tests hédoniques et en profils sensoriels. Chaque chapitre offre une description détaillée des méthodes retenues, une justification des choix effectués pour apporter des réponses à nos questions de recherche.

Le dernier chapitre est consacré à la discussion générale des résultats et à l'apport des bases de données dans la quantification des paramètres qui déterminent le nombre de sujets.

Chapitre 1

Revue bibliographique

1 Mesure de la perception sensorielle

Les organes des sens réagissent à des modifications physiques ou chimiques du milieu environnant. Ces variations constituent les stimuli.
Les stimuli qui déclenchent l'excitation des récepteurs sensoriels peuvent être d'origine : physique, chimique ou bien psychosensorielle (origine affective).

L'application d'une stimulation provoque l'activation de molécules spécifiques situées dans la membrane d'une cellule réceptrice. L'information est ensuite amplifiée et transmise sous forme de signaux électrique jusqu'au système nerveux central (Mac Leod and Sauvageot, 1986). À leur arrivée au niveau du système nerveux central, ces informations forment une image sensorielle qui est confrontée simultanément à la mémoire (identification de la stimulation), au centre du plaisir, et à la conscience (Narçon, 2001).
Le plaisir, fonction physiologique procurée par l'activité d'un noyau situé dans l'hypothalamus, n'est associé à aucune modalité sensorielle particulière (Mac Leod, 1992). Il est cependant lié à la représentation que l'individu se fait de son environnement (croyances, valeurs) et aux stimulations qui semblent l'avoir procuré (Narçon, 2001).

L'intégration de toutes les informations, en provenance des différents récepteurs sensoriels, mais également en provenance de la mémoire, de la conscience et du centre du plaisir, est perçue par l'individu comme un message global. Le contexte dans lequel a été perçu le stimulus ainsi que l'état psychologique et physiologique de l'individu ont une influence sur l'intégration de ces informations (Narçon, 2001).

1.1 La perception sensorielle des aliments

Lorsqu'une personne entre en contact avec un aliment, tous les sens sont stimulés. L'aliment possède des qualités organoleptiques, ensemble des qualités qui affectent les organes des sens.

L'individu prend conscience des sensations qu'il perçoit selon trois composantes : la nature de la sensation (composante qualitative); son intensité (composante quantitative) et le plaisir qu'elle induit (composante hédonique). L'intensité, la qualité et le caractère hédonique sont les trois facteurs qui permettent de décrire une perception (Faurion, 1992).

D'après Mac Leod (1998), il est difficile pour l'individu de dissocier les composantes hédoniques et sensorielles qualitative et quantitative car elles sont confondues. Le plaisir alimentaire induit par la perception sensorielle des aliments est une composante affective qui complète les composantes qualitative et quantitative.
Les aspects hédonique et organoleptique ont un rôle majeur dans la prise alimentaire (Chiva, 1996). Ainsi, il devient nécessaire de recourir à l'analyse sensorielle, pour mesurer la qualité perçue des produits alimentaires par les consommateurs. En effet, (Chiva, 1996) propose d'utiliser des méthodes expérimentales et de faire une place particulière à l'analyse sensorielle pour étudier les pratiques alimentaires.

1.2 Les épreuves sensorielles

1.2.1 Principe des épreuves sensorielles

Les trois composantes de la perception sensorielle définissent les différent types d'épreuves sensorielles. Ces dimensions sont étudiées selon deux approches distinctes avec, d'une part, l'étude des préférences ou de la satisfaction des sujets (approche axée sur les sujets) et d'autre part, l'analyse descriptive des propriétés sensorielles (approche axée sur les produits). La première approche est dite hédonique; la seconde est dite analytique (mesure hédonique et mesure sensorielle selon Issanchou and Hossenlopp (1992)). Ces deux approches font appel à des dégustateurs et leur sens de la vue, de l'odorat, du goût, du toucher et de l'ouïe afin de mesurer les caractéristiques sensorielles et l'acceptabilité des produits.
Bien qu'il n'existe pas de référentiel sensoriel commun à tous les individus, aucun instrument ne peut reproduire ou remplacer la réaction humaine, ce qui fait de l'analyse sensorielle un élément essentiel de toute étude alimentaire.

1.2.2 Approche hédonique

Par définition, l'hédonisme dans le contexte sensoriel fait référence au plaisir subjectif que procurent les caractéristiques sensorielles d'un produit. La mesure hédonique a pour objet l'étude des préférences pour un produit ou encore son acceptabilité (Urdapilleta et al., 2001). La préférence est l'attitude d'un sujet qui trouve un produit meilleur qu'un ou plusieurs autres. L'acceptabilité d'un aliment est l'état d'un produit reçu par un individu ou un ensemble d'individus, et ce, en fonction de ses qualités organoleptiques.

Les propriétés sensorielles d'un produit et son appréciation sont présentes dans l'ensemble des modèles représentant les déterminants du comportement alimentaire (Randall and Sanjur, 1981; Shepherd et al., 1985). Par ailleurs, le plaisir sensoriel a un impact sur le comportement d'achat ou sur le choix d'un produit. La maximisation de l'appréciation de l'aliment par le consommateur (attitude envers les propriétés sensorielles) est donc indispensable pour assurer la pérennité d'un produit. Ainsi, le rôle non négligeable de l'appréciation sur l'intention de ré-achat entraîne les industriels à extrapoler les performances organoleptiques des recettes obtenues pour prédire la performance future d'un produit sur le marché. D'après Fantino (1992) et Chiva (1996), la composante hédonique est le moteur essentiel du comportement alimentaire. L'importance du rôle de l'appréciation reste bien sûr à déterminer et à compléter en comparaison avec celui des autres composantes du comportement alimentaire.

L'analyse hédonique des produits alimentaires est essentielle étant donné le rôle primordial des sens dans l'acceptabilité de ces produits par les consommateurs (Cardello and Schutz, 2003). Elle est appliquée pour comparer la performance organoleptique globale de différentes recettes, aider au développement de nouveaux produits, à l'amélioration de produit, trouver le positionnement par rapport à la concurrence, ou encore étudier l'impact de process ou de conditionnement sur l'appréciation de la recette d'un produit. La mesure hédonique est restreinte à l'appréciation sensorielle du produit, toute information extrinsèque (prix, emballage, etc) est exclue (Buck, 2003).

Les mesures hédoniques font appel aussi bien à des sujets complètement naïfs qui n'ont jamais fait d'évaluation sensorielle, qu'à des sujets déjà familiarisés avec ce type d'épreuves. Dans la littérature, ces mesures sont regroupées sous le nom d'épreuves hédoniques (AFNOR, 2009) ou de tests d'acceptabilité (acceptance tests) (Stone and Sidel, 2004). Ces tests consistent à mesurer le degré d'appréciation des produits (liking tests), ou bien, à mesurer la préférence d'un ou plusieurs produits (preference tests). On parle alors de tests hédoniques ou épreuves de notation; et de tests préférence ou épreuves de classement.

1.2.2.1 Tests de préférence

Les tests de préférence ont pour objectif de déterminer un classement de préférence entre les produits dégustés. On distingue deux types de tests : les tests de préférence par paire ou les tests de classement.
Les tests de préférence par paire consistent à présenter uniquement deux produits en même temps au sujet qui doit indiquer le produit qu'il préfère. Cette méthode peut cependant être utilisée lors d'une étude comparant plus de deux produits (test par paires multiples).
Les tests de classement consistent à présenter directement l'ensemble des produits aux sujets qui doivent donner un classement par rang de ces produits selon leur préférence.
Les commanditaires d'une étude de préférence cherchent à savoir comment se positionne leur produits par rapport à la concurrence ou à faire une sélection entre plusieurs recettes qu'ils ont mises au point.

1.2.2.2 Tests hédoniques

Ils sont conçus pour mesurer le degré d'appréciation globale d'un produit à l'aide d'une note choisie sur une échelle. On parle d'épreuve de notation. Les sujets choisissent, pour chaque produit, la catégorie qui correspond à leur degré d'appréciation. On utilise cette évaluation quantifiée de l'appréciation lorsqu'on désire avoir une estimation dans l'absolu du caractère agréable et qu'aucun standard bien connu n'est disponible; ou bien lorsqu'on veut comparer l'appréciation hédonique des différents groupes de sujets (Köster, 1998).
Même si on demande aux dégustateurs d'indiquer la mesure dans laquelle un produit leur plaît ou s'ils l'acceptent, les tests hédoniques servent souvent à mesurer la préférence ou l'acceptation de façon indirecte.

On trouve dans certains manuels de la pratique sensorielle la mention test d'acceptation. Il s'agit d'une variante des épreuves de notation hédonique qui s'attache à présenter un seul produit, afin d'éviter les comparaisons conscientes ou inconscientes entre différents éléments. On demande au sujet de noter les critères pour ce produit en fonction de son standard personnel pour évaluer, par exemple, l'acceptabilité par rapport au dosage en sel, au niveau de cuisson, à l'odeur, à la forme, à l'aspect, etc.

Outre le type de réponses hédoniques, plusieurs alternatives méthodologiques existent pour collecter l'appréciation des produits. Parmi ces alternatives méthodologiques, nous retrouvons le choix du lieu du test. En complément des tests en laboratoire d'évaluation sensorielle, on retrouve les tests en salle (dans un lieu public) ou bien les tests dits à domicile (domicile du sujet). La procédure du recueil des réponses peut être également de différente nature et demander une évaluation à l'aide d'un questionnaire autoadministré ou en face-à-face par un enquêteur.

1.2.3 Approche analytique

L'approche analytique consiste à décrire un produit alimentaire en propriétés sensorielles tangibles, objectives et bien définies et en déterminant l'intensité des ces propriétés (SSHA and Depledt, 1998; Urdapilleta et al., 2001). Cette mesure sensorielle revient à explorer de manière objective une sensation pour la verbaliser en mots ou en chiffres (Issanchou and Hossenlopp, 1992). Les propriétés sensorielles sont caractérisées qualitativement et quantitativement.
Dans ce type d'épreuves, on fait appel à des sujets entraînés. Ces derniers sont formés sur la méthodologie, à la reconnaissance des sensations perçues, à l'appropriation d'un vocabulaire commun et à la quantification de ces sensations. Ce type d'épreuves est toujours mené en laboratoire d'analyse sensorielle.

L'approche analytique se scinde en deux catégories de tests : les tests discriminatifs et les tests descriptifs.

1.2.3.1 Tests discriminatifs

Ces tests sont utilisés lorsqu'on désire détecter la présence ou l'absence de différences sensorielles entre deux produits. L'objectif est de savoir, par exemple, si un changement dans le mode de production (changement d'ingrédients, de matériel de production, de processus ...) a une conséquence sur la perception finale du produit par le consommateur. Ils sont recommandés pour comparer des produits entre lesquels les différences sont faibles et inconnues. On peut donc tester des produits concurrents, des formules différentes du même produit, le même produit à différents stades de maturation, ayant bénéficié de méthodes de stockage ou de conservation différenciée ou provenant de différents lots, etc. On fait appel à un groupe de sujets initiés et dans certains cas, les tests discriminatifs peuvent être réalisés avec des sujets naïfs (non entraînés aux méthodes de l'analyse sensorielle). Différentes épreuves discriminatives sont à disposition des expérimentateurs : tests triangulaires, comparaison par paire, duo trio, 2-AFC, 3-AFC, etc. Les tests discriminatifs peuvent aussi être menés avant une évaluation descriptive ou hédonique.
Les tests discriminatifs servent essentiellement à déterminer s'il existe des différences sensorielles entre des produits. Si on veut aller plus loin et qualifier ces différences éventuelles, il convient d'avoir recours aux tests descriptifs.

1.2.3.2 Tests descriptifs

L'objectif est d'évaluer l'intensité d'une grandeur sensorielle simple ou complexe et d'aboutir à une description efficace des produits analysés. Les produits comparés appartiennent au même segment de marché et sont ainsi caractérisés à partir de descripteurs communs. Ces descripteurs

peuvent être des descripteurs visuels, d'arôme, d'odeur, de saveur ou de texture. À chaque descripteur est associé une intensité qui correspond au niveau de la sensation perçue. Ces méthodes font appel à un petit groupe de sujets sélectionnés pour leurs aptitudes sensorielles et entraînés aux tâches qu'ils auront à effectuer.
On distingue différents types de tests descriptifs : le profil conventionnel, le profil de la flaveur (AFNOR, 1983; Little, 1950), la méthode QDA© Quantitative Descriptive Analysis (Stone et al., 1974), le profil libre choix (Williams and Langron, 1984), le profil flash (Delarue and Sieffermann, 2000), la méthode Dominance Temporelle des Sensations DTS (Labbe et al., 2009; Pineau et al., 2009), la méthode Spectrum (Muñoz and Civille, 1992) ... etc.
Les mesures descriptives sont couramment utilisées par les industriels de l'agro-alimentaire afin de contrôler la qualité sensorielle d'un produit et sa conformité avec les objectifs initiaux. Elles sont aussi utilisées pour évaluer l'effet sensoriel de la modification des procédés de fabrication (matières premières utilisées) et d'étudier par exemple l'impact des conditions de fabrication ou de stockage sur l'évolution du produit.

1.2.4 Mesure et établissement d'un plan d'expérience : structure en bloc, répétition et randomisation

Les tests sensoriels doivent être réalisés dans des conditions contrôlées, en se servant de plans d'expériences, de méthodes de vérification et d'analyses statistiques bien conçues. C'est la seule façon pour l'analyse sensorielle de fournir des données fiables.

La planification expérimentale est l'étape fondamentale pour mettre en place une étude sensorielle menée auprès d'un groupe de sujets évaluant un ensemble de produits (Gacula, 1988; Hunter, 1996; MacFie, 1986). Le plan d'expériences doit prendre en compte les objectifs de l'étude, du type de produit à l'étude, des procédures et des conditions des tests, des ressources disponibles et du type de test statistique à réaliser. Cette démarche permet d'assurer une fiabilité des données collectées et une exploitation dans les meilleures conditions statistiques possibles. Les sources de variations incontrôlables ou des erreurs sont ainsi diminuées.

Deux questions se posent lorsque l'on veut tester un groupe de produits par un panel. Dans un premier temps, la question du nombre de sujets à enrôler qui est soumise à plusieurs paramètres statistiques mais aussi à des facteurs techniques tels que la disponibilité des sujets et les contraintes budgétaires. Dans un second temps, se pose la question du nombre maximum de produits que peut évaluer un sujet à chaque session, et ce non pas pour des raisons budgétaires, mais plus souvent en raison de contraintes physiologiques induites par ce que l'on appelle la fatigue sensorielle. Un sujet, même entraîné, ne peut pas nécessairement garder toutes ses capacités sensorielles s'il évalue trop de produits à la suite.

Une fois ces deux questions résolues, il ne reste plus qu'à déterminer quels produits seront évalués par chacun des sujets, au cours de chaque session, et dans quel ordre. Il y a de nombreuses façons de planifier des expériences en allant des plans simples, complètement aléatoires, à des plans plus complexes, fractionnés et factoriels. Les caractéristiques les plus courantes d'un bon plan d'expérience sont la randomisation, le recours à la méthode des blocs et la répétition.

1.2.4.1 Structure en bloc

On se sert de la méthode des blocs dans de nombreuses expériences pour contrôler des sources de variations connues et pour améliorer l'efficacité. Les blocs peuvent être les effets du jour, les sujets, les répétitions ou les ordres de présentation des échantillons, tout facteur dont on sait qu'il peut être la source de variation lors de l'expérience.
Par ailleurs, un plan en blocs est un plan d'expérience dans lequel on étudie l'influence d'au moins deux facteurs sur un ou plusieurs phénomènes. On sait que l'un des facteurs a par construction un effet important, sans que l'on puisse agir dessus, mais ce n'est pas celui qui nous intéresse. On veut donc pouvoir s'assurer que ce facteur ne perturbera pas les analyses que l'on effectuera une fois les données collectées. Les unités expérimentales sont alors regroupées en blocs. La variation entre les unités au sein d'un bloc a des chances d'être inférieure à la variation entre les blocs. Un tel plan permet d'avoir une mesure vraie de l'erreur pure ou expérimentale en tenant compte de la variance due aux facteurs blocs et en la retirant des sources incontrôlables d'erreurs expérimentales.
En analyse sensorielle, nous avons un facteur bloc qui correspond aux sujets, et un facteur que l'on souhaite particulièrement étudier, le facteur produit. Les sujets étant des êtres humains, ils sont souvent des sources connues de variations lors des expériences sensorielles. En les regroupant en blocs lors de la planification de l'expérience et de l'analyse des données, les variations dues aux sujets peuvent être retirées de l'erreur expérimentale et isolées comme effet des sujets. Le terme d'erreur servant à déterminer s'il y a des différences significatives entre les produits indiquera alors davantage l'erreur pure. Pour cela on fait en sorte que les différents niveaux des autres facteurs soient aussi bien représentés dans chacun des blocs (les modalités du facteur bloc).
On parle souvent de trois variantes de plan en blocs : plan complet, plan incomplet et plan équilibré. Un plan en blocs complets est un plan dans lequel tous les niveaux des facteurs étudiés sont présents une fois à l'intérieur de chaque bloc. Cela correspond, pour un plan sensoriel, au cas où tous les produits sont vus une fois par l'ensemble des juges. Un plan en blocs incomplets est un plan dans lequel tous les niveaux des facteurs étudiés ne sont pas présents dans chaque bloc. Il est équilibré si chaque niveau de chaque facteur étudié est présent un même nombre r de fois et si chaque couple de niveaux de chaque facteur étudié est présent un même nombre de fois dans un même bloc.

1.2.4.2 Les répétitions

Les répétitions d'une expérience suppose de pouvoir répéter toute l'expérience dans des conditions identiques. Le nombre de répétitions d'une expérience varie et est souvent fonction des contraintes de temps, d'argent et d'échantillon. Les répétitions sont indispensables lorsqu'on étudie les performances à la fois des sujets et du panel dans les tests du profil sensoriel. En revanche, elles sont inadéquates dans le cas des tests hédoniques (Köster et al., 2003). Il a souvent été admis que les répétitions amélioraient la fiabilité et la validité des résultats des tests de profil sensoriel. Une discussion plus approfondie et détaillée sur la nécessité des répétitions en phase de mesure en profil sensoriel fait l'objet de la publication n°2 dans ce manuscrit.

1.2.4.3 La randomisation

La randomisation est une condition nécessaire dans toute réalisation d'un plan d'expérience. En effet, toute influence de l'ordre dans lequel le sujet déguste les produits sur ses réponses est évitée en changeant cet ordre d'un sujet à un autre ou d'une répétition à l'autre pour un même sujet.
L'usage des carrés latins de Williams (1949) a été recommandé en analyse sensorielle. Ils ont été initiés par MacFie et al. (1989) et Schlich (1993). Ces plans garantissent à la fois l'équilibre des effets de rang et des effets de report de premier ordre. Les carrés latins mutuellement orthogonaux MOLS (Mutually Orthogonal Latin Square) (Callier and Schlich, 1997; Wakeling and MacFie, 1995) permettent de construire les plans de présentation où l'équilibre des effets de report est maximal. Chaque produit étant présenté le même nombre de fois en première, en seconde, ... et en dernière position. Et ils permettent aussi la maîtrise des arrières-effets, aussi appelés effet de report, effet d'ordre 1, d'ordre 2, ... De tels plans imposent un nombre de sujets qui soit un multiple du nombre de produits testés (cas pair) ou de son double (cas impair).

Nous allons définir dans les sections suivantes la mise en place des tests hédoniques mais aussi du profil sensoriel conventionnel d'une manière plus approfondie.

1.2.5 Les tests hédoniques

1.2.5.1 Principe et méthodologie

L'objectif des tests hédoniques est d'obtenir l'opinion des consommateurs concernant l'appréciation globale des produits testés. On mesure le statut hédonique d'un produit en demandant aux sujets de le noter après dégustation. Ces données récoltées dans l'absolu servent par exemple, à connaître la position d'un produit cible en termes d'appréciation dans son univers produit.

1.2.5.1.1 Les sujets

Lorsqu'on s'intéresse aux aspects hédoniques d'un produit, il est important de recueillir les réponses de sujets choisis en fonction du type du problème étudié.
Il peut se faire qu'un échantillon de consommateurs représentatif ou quasi-représentatif de la population convienne, mais il faudra souvent recruter des groupes spécifiques de consommateurs du produit ou groupes cibles.
Les membres du groupe doivent être des sujets non entraînés, auxquels on demande d'exprimer leur appréciation et auxquels on ne pose pas de questions analytiques.

Si les laboratoires d'analyse possèdent une base des données sur différentes études hédoniques, un critère de sélection de sujets qui pourrait être intéressant consisterait à choisir les sujets qui ont leurs scores d'acceptabilité compris dans l'intervalle $m \pm \sigma$; avec m et σ, la moyenne globale et l'écart-type respectivement des scores d'acceptabilité obtenus de tous les sujets pour la famille de produits étudiée (Stone and Sidel, 2004).

1.2.5.1.2 Procedure

Comme toute épreuve de notation, un test hédonique implique l'utilisation d'une échelle de notation. Différents types d'échelles ont été proposés pour collecter le niveau d'appréciation des produits. L'échelle peut être structurée ou non structurée, numérique, sémantique ou picturale. La seule règle indispensable au bon traitement des données quantitatives est que l'échelle proposée doit être une échelle d'intervalle, c'est-à-dire, satisfaisant la règle de l'égalité des intervalles délimités par les différents barreaux de l'échelle. La plupart des procédures statistiques telles que le calcul des moyennes et des variances, les analyses de variances, les régressions et corrélations peuvent alors être utilisées pour analyser les données.
La littérature offre de nombreux exemples d'échelles mais l'échelle hédonique à 9 points développée par Jones et al. (1955); Peryam and Pilgrim (1957) reste la plus utilisée dans la littérature. On retrouve deux versions de cette échelle dans la pratique des tests hédoniques décrites dans la norme AFNOR (2009) : l'échelle discontinue sémantique (figure 1.1) et l'échelle discontinue numérique (figure 1.2) variant de 1 à 9 avec les mentions "je n'aime pas du tout" qui correspond à la note minimale 1 et "j'aime beaucoup" qui correspond à la note maximale 9.

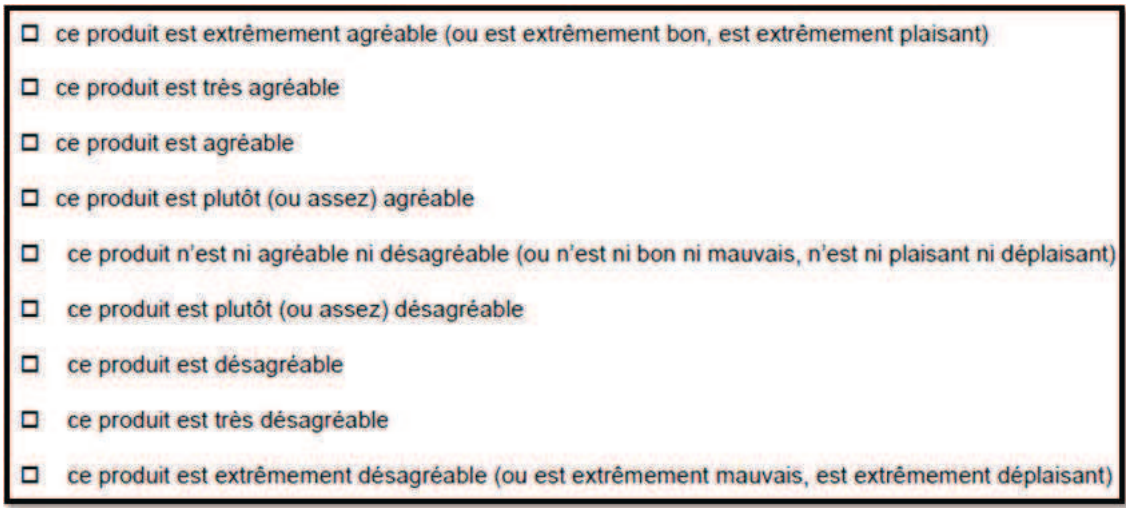


FIG. 1.1 – *Échelle discontinue sémantique*

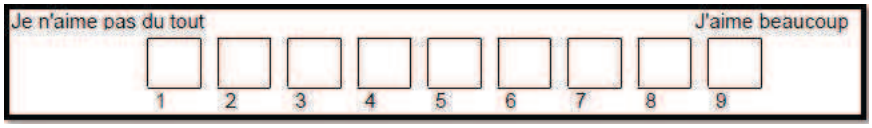


FIG. 1.2 – *Échelle discontinue numérique*

Il peut sembler logique de limiter le questionnaire à une seule question portant sur l'appréciation globale. Cependant, en pratique, les commanditaires des études hédoniques s'intéressent également à l'opinion des consommateurs concernant des attributs sensoriels spécifiques. Ainsi, la majorité des études hédoniques présentent des questions diagnostiques utilisées comme pistes de compréhension des résultats obtenus sur la préférence ou l'appréciation globale. Différents types de questions diagnostiques sont donc souvent ajoutées sous forme de questions ouvertes concernant les raisons d'appréciation ou de non appréciation, des questions concernant l'appréciation d'attributs spécifiques (notation de l'intensité par rapport à l'idéal (Shepherd et al., 1989)). Il est à noter que toute question complémentaire est soigneusement ordonnée pour ne pas perturber la spontanéité des réponses sur l'appréciation des sujets (Popper et al., 2004). La question sur l'appréciation globale doit être posée en premier.

1.2.5.1.3 Les produits

Le nombre de produits à évaluer au cours d'une séance dépend du type de produit, mais ne devrait jamais dépasser 20, pour éviter la fatigue des sujets (Köster, 1998).
Les échantillons de produits sont présentés dans des contenants identiques et sont codés. Chaque échantillon doit avoir un numéro distinct.
Le mode de présentation des échantillons constitue un point important du protocole des tests

hédoniques. Ainsi, nous trouvons différents types de présentations qui peuvent être classés comme suit :

- le monadique pur : chaque sujet n'évalue qu'un seul produit (il y a donc autant de groupes de sujets que de produits à évaluer),
- le monadique séquentiel multi-séance : tous les sujets dégustent tous les échantillons présentés lors de séances différentes (un produit par séance),
- le monadique séquentiel mono-séances : tous les sujets dégustent tous les échantillons les uns à la suite des autres dans une même séance,
- le comparatif ou simultané : tous les sujets dégustent l'ensemble des échantillons en simultané dans une même séance.

L'ordre des échantillons doit être différent pour chaque sujet et obéit à un plan d'expérience établi au préalable. Ceci permet d'éviter tout effet d'ordre.

Bien d'autres caractéristiques du protocole nécessitent généralement un choix de la part de l'expérimentateur : l'ordre de présentation des produits, la quantité de produits présentée, le nombre de produits présenté, le moment de la séance, etc. Ces quelques exemples de choix auxquels un expérimentateur est soumis conditionnent les caractéristiques du protocole.

1.2.5.2 Le nombre de sujets

Le nombre de sujets à interroger est une étape essentielle pour toute étude hédonique. Ce nombre détermine en partie la puissance du test à détecter une différence significative entre les produits à comparer. On entend par différence significative, la plus petite différence d'acceptabilité au delà de laquelle on estime que les produits sont différents. Cette quantité est fixée à priori par les commanditaires de l'étude. Pour garantir la puissance du test, une détermination prospective d'un effectif minimum de sujets est nécessaire.

La littérature autour du nombre minimum de sujets à enrôler dans les tests hédoniques n'est pas très abondante et lorsqu'elle existe, elle demeure peu objective et parfois controversée.

Les manuels d'analyse sensorielle abordent la question et se sentent obligés de fournir des recommandations car il existe une véritable demande des utilisateurs. Chambers and Wolf (1996) indiquent qu'au minimum 30 sujets sont nécessaires mais que généralement, 100 est une taille de panel adéquate pour la plupart des tests d'acceptabilité. Cependant, le nombre exact dépend du plan d'expérience, du type de produits testés et de la représentativité de la population étudiée. De plus, les auteurs précisent qu'un effectif de 100 sujets impliquerait une diminution de l'erreur et améliorerait la probabilité de détecter de petites différences. Köster (1998) suggère l'utilisation de 32 sujets au minimum. Par ailleurs, Lawless and Heymann (1998) recommandent une taille minimale de 50 sujets.

Meilgaard et al. (1999) préconisent environ 50 à 300 sujets pour les tests hédoniques en laboratoire, lorsque 75 à 300 sujets sont requis pour les tests à domicile. Stone and Sidel (2004) recommandent 25 à 75 sujets par produit dans le cas des tests en laboratoire. Les auteurs précisent que ce nombre peut être au minimum de 24 sujets dans la cadre d'un pré-test. Même s'il est difficile d'établir un résultat statistiquement significatif avec un tel effectif, il pourrait néanmoins donner une tendance générale sur l'acceptabilité des produits. Pour les autres contextes, les auteurs recommandent environ 50 à 100 sujets (familles) pour les tests à domicile et si une segmentation est attendue, 100 sujets ou plus sont requis pour les tests en salle. Ils précisent que 100 est une taille de panel nécessaire pour compenser la variabilité due à l'inexpérience des sujets et aux limites de l'environnement du test. Contrairement aux tests en laboratoire, les tests en salle sont moins documentés quant au nombre de sujets, alors que le problème est identique.

Gordon and Norback (1985), Sharp et al. (1986) et Cliff et al. (1997) préconisent l'utilisation de 100 réponses au minimum par produit évalué et par segment de population étudié. Basker (1996) suggère 100 sujets comme taille de panel adéquate pour les tests de préférence. Moskowitz (1997) indique que 40 à 50 sujets est une taille de panel suffisante. Un panel inférieur à cette taille ne permettant pas de représenter la diversité existant dans les préférences des consommateurs ne conduirait pas à une conclusion stable et exploitable. Cet article a reçu deux réponses (Cornell, 1997; McEwan, 1997) pointant les limitations de l'approche utilisée. En effet, McEwan (1997) critique l'approche simpliste du problème de détermination du nombre de sujets adéquat en test hédonique. Cette approche est basée sur deux études seulement et n'aborde pas les considérations réelles en tests hédoniques telles que : l'objectif du test, le plan d'expérience, le type de produit étudié ainsi que le nombre potentiel de segments de préférences attendues (si on désire le mettre en évidence). D'autre part, Cornell (1997) pointe le manque de robustesse des résultats statistiques concernant la "stabilité de la moyenne".

Il est d'usage de faire appel à des effectifs de 50 à 75 sujets pour les tests hédoniques en laboratoire. Ces chiffres sont loin des tailles de panel utilisées en marketing (environ 500 sujets). Pour des cas plus spécifiques, notamment le lancement d'un nouveau produit, les services marketing peuvent avoir besoin d'identifier plus précisément leur cible de consommateurs.
Les enjeux en analyse sensorielle, plus précisément en tests hédoniques, sont moins importants puisque leur objectif est restreint à tester l'acceptabilité d'un groupe de produits et non pas à être utilisé comme une étude de marché. Lorsqu'on veut identifier des segments de préférences, il est demandé d'interroger au moins 120 sujets. Les commanditaires de ce type d'études accordent de la confiance aux résultats obtenus à partir de larges effectifs permettant d'avoir une meilleure représentativité de la population. Un autre point important est la variabilité des produits testés. Lorsque les produits testés sont assez différents sensoriellement, les préférences induites sont généralement évidentes. Dans ce cas de figure, l'emploi de larges groupes de sujets n'apporterait

pas plus d'information sur les préférences (Sidel et al., 1994).
Il est important de repositionner l'objectif des tests hédoniques. Ces derniers sont mis en place uniquement pour tester l'acceptabilité des produits. Ils ne sont en aucun cas utilisés pour estimer une intention d'achat, ni pour détecter des segments de préférences. Un test hédonique est un bon indicateur pour la construction et la mise en place d'une étude marketing à plus grande échelle.

La puissance statistique des tests hédoniques et le calcul du nombre de sujets ont été abordés dans différents manuels d'analyse sensorielle (Gacula, 1993; Gacula and Singh, 1984; Lawless and Heymann, 1998). Les auteurs donnent la démarche pour calculer le nombre nécessaire de sujets dans le cas de comparaison de deux produits. Ainsi, (Gacula, 1993) estime ce nombre à 52 sujets lorsqu'on veut mettre en évidence une différence d'appréciation de 0.5 entre deux produits sur une échelle hédonique à 9 points. Lawless and Heymann (1998) rapportent les résultats de Gacula (1993) et mettent l'accent sur la nécessité d'avoir une estimation a priori de la grandeur de la différence à mettre en évidence. Cependant, les auteurs n'ont pas discuté la détermination du nombre de sujets dans le cas de comparaison de plusieurs produits et qui font souvent l'objet des tests hédoniques.

Souvent, les recommandations sur le nombre de sujets sont obtenues à partir d'une étude sur une dizaine de jeux de données. Nous n'avons trouvé qu'une seule tentative de recommandations (Hough et al., 2006) basée sur une centaine de jeux de données. Cette proposition permet de calculer l'effectif de sujets nécessaire en fonction de la plus petite différence que l'on souhaite détecter, de la variabilité estimée des données et des risques d'erreurs choisis (l'erreur de type I et de l'erreur de type II). De plus, l'estimation de ce nombre nécessaire est réalisé sur des études hédoniques comparant plusieurs produits. La variabilité inter-individuelle σ est estimée par la racine carrée du carré moyen de l'erreur dans le modèle additif d'ANOVA (Produit + Sujet) appliqué à 108 études provenant de 5 pays. Comme différentes échelles de mesures ont été utilisées, une transformation des données fut appliquée pour ramener les notes de chaque étude entre 0 et 1. La moyenne de la variabilité des 108 études était de 0.23. Pour cette valeur de σ, une erreur de type I de 5%, une erreur de type II de 10% et une différence entre les moyennes des notes par produit attendue de 10% sur une échelle 0-1, le nombre de sujets à enrôler obtenu est de 112 sujets. Les auteurs précisent que ce nombre est à utiliser seulement pour mettre en évidence une différence d'appréciation entre les produits testés. Si par exemple, l'objectif de l'étude est de mesurer l'acceptabilité des produits dans deux régions différentes au sein d'un pays, le nombre de sujets devrait être 112 sujets pour chaque région, soit un total de 224 sujets.
Par ailleurs, les auteurs précisent que la variation de certains paramètres permettant de calculer le nombre de sujets a un impact important sur les recommandations. Ainsi, une augmentation de la taille de différence à mettre en évidence de 10% à 20% sur l'échelle, diminue d'une façon

drastique la recommandation sur le nombre de sujets, soit de 70%. Cependant, une augmentation de l'erreur de type I de 5% à 10% ne réduit que de 20% le nombre de sujets.

Une approche par simulation pour la détermination du nombre de sujets nécessaire en tests hédoniques a été menée par Gacula and Rutenbeck (2006). Cette approche reprend les concepts de base introduits par Hough et al. (2006). Tout d'abord, la variabilité des données est estimée mais seulement à partir de 4 jeux de données. Ensuite, différentes valeurs de différences d'appréciation que l'on pourrait observer sur une échelle en 9 points ont été simulées grâce à une équation de simulation qui est fonction de la variabilité. Enfin, partant des estimations de ces deux paramètres, Gacula and Rutenbeck (2006) concluent qu'au minimum 40 sujets sont nécessaires pour détecter une différence significative de 0.6 sur une échelle en 9-points. Plus la différence à mettre en évidence est petite, plus le nombre de sujets sera grand (pour une différence $d = 0.4$ le nombre de sujets nécessaire est au minimum 100 sujets). L'auteur insiste sur la nécessité d'avoir une idée sur la magnitude des différences d'appréciation que l'on souhaiterait mettre en évidence entre les produits à comparer.

Dans la pratique, les laboratoires d'analyse sensorielle interrogent une population représentant différentes classes d'âges, sexes et CSP composée d'au moins 60 sujets (AFNOR, 2000). De plus, les entreprises agro-alimentaires ont généralement recours à des recommandations sur le nombre de sujets pour les tests hédoniques émises par les normes ou bien en se basant sur la pratique interne. En ce qui concerne ce choix de la taille de l'effectif, les utilisateurs de tests consommateurs chez Danone se sont arrêtés sur un compromis scientifique et économique de 120 sujets par cible de consommateurs avec une tolérance jusqu'à 80 sujets (Boutrolle, 2007).

La norme (AFNOR, 2000) recommandait au minimum 60 sujets à enrôler pour réaliser un test hédonique. Récemment, cette limite de 60 est passée à 100 sujets (AFNOR, 2009). Cette nouvelle version de la norme précise que lorsqu'on veut déterminer le nombre de sujets nécessaire, il faudrait estimer la différence que l'on souhaite mettre en évidence et avoir une quantification de l'erreur, c'est-à-dire la variabilité des données. Dans le cas où le laboratoire ne possède pas ces deux estimations pour calculer le nombre de sujets nécessaire, il faudrait alors prendre à défaut 100 sujets. D'un point de vue statistique, une augmentation du nombre de réponses influence la puissance du test, c'est-à-dire sa capacité à détecter une différence si elle existe. Il est évident qu'une telle démarche a un impact à la fois sur le coût et les moyens matériels.

1.2.5.3 Synthèse

L'examen de la littérature nous a permis d'identifier deux approches pour la détermination du nombre de sujets nécessaire en tests hédoniques.
Tout d'abord, la détermination du nombre de sujets nécessaire par approche théorique classique. Cette approche est souvent utilisée dans le cas de comparaison de deux produits. Il existe une

seule étude Hough et al. (2006) qui a donné des recommandations basées sur la comparaison de plusieurs produits et dont les résultats méritent d'être validés. La méthode exige la quantification de la taille de l'effet qu'on veut mettre en évidence mais aussi une estimation de la variabilité des données. La taille de l'effet que l'on désire mettre en évidence est définie comme la plus petite différence significative entre deux produits. Il en ressort que la différence est significative entre deux produits quand leur moyenne diffère d'au moins une valeur "d" sur une échelle de mesure. Cette valeur doit être fixée a priori par l'expérimentateur. Elle pourrait aussi être quantifiée à partir de l'historique des valeurs observées sur des études similaires issues par exemple de la même famille de produits que les produits à tester. De plus, la variabilité des données qui traduit la variabilité inter-individuelle (hétérogénéité des préférences) doit être aussi estimée à partir d'un historique de jeux de données.
Ensuite, il existe une autre approche pour déterminer le nombre de sujets basée sur la simulation sous deux formes.
La première consiste à simuler des sous-panels à partir d'un panel complet et à ensuite comparer les résultats obtenus sur la base des sous-panel à ceux obtenus du panel complet (Basker, 1996; Moskowitz, 1997). Il suffit alors de trouver le nombre maximum de sujets que l'on peut retirer sans altérer les conclusions statistiques du panel complet.
La seconde consiste à simuler des données hédoniques, donc des différences d'appréciation et à déduire la taille de panel pour laquelle on observe un effet produit significatif (Gacula and Rutenbeck, 2006).
Une approche par simulation ne serait pas pertinente ni robuste si elle était menée que sur peu de jeu de données. Pourtant cette méthode a été utilisée en se basant sur deux ou quatre jeux de données dans ces articles.

1.2.6 Le profil sensoriel conventionnel

1.2.6.1 Principe et méthodologie

La méthode du profil conventionnel AFNOR (2003) est une généralisation de la méthode Quantitative Descriptive Analysis (QDA©) développée par Stone et al. (1974). Le but est d'établir une description complète des propriétés sensorielles d'un ensemble de produits sur le plan qualitatif et quantitatif, de manière à établir une carte d'identité, précise, reproductible et comprise par tous, des produits étudiés. Ces cartes d'identité, basées sur les moyennes du panel, constituent les profils sensoriels des produits et sont généralement représentées sous forme de graphiques.

L'identification des caractéristiques spécifiques des produits peut être utilisée pour communiquer sur le produit; par exemple, un produit sous signe de qualité dont le cahier des charges indique

qu'il doit être plus sucré ou plus fondant qu'un produit courant. Le profil sensoriel sert aussi dans le cadre d'études de recherche à mettre en lien différent paramètres de formulation ou de technologie avec les caractéristiques du produit fini.

1.2.6.1.1 Les sujets

Le profil sensoriel classique ou conventionnel fait appel à un groupe de sujets dits qualifiés ou experts. Ces sujets sont entraînés à la description sensorielle et de l'univers des produits. Ils sont recrutés pour leurs aptitudes sensorielles.
Afin de sélectionner les participants, il existe de nombreux tests basés à la fois sur des aptitudes sensorielles et sur des aptitudes non sensorielles (Nicod, 1998). Il convient aussi de s'assurer que les panélistes perçoivent normalement les odeurs et les saveurs (Lawless and Heymann, 1998). Il semble tout aussi important de prendre en compte la disponibilité et la motivation des juges pour constituer le panel.

1.2.6.1.2 Procédure

Le profil sensoriel consiste en deux étapes : la phase d'entraînement des sujets et la phase de mesure.

L'entraînement des sujets consiste à sélectionner puis à entraîner un panel de sujets à l'évaluation d'un type de produit. Les sujets sont entraînés à générer et à sélectionner un vocabulaire qui permet d'établir une liste de descripteurs des différentes sensations sur l'aspect, l'odeur, la texture, l'arôme et la saveur. Par ailleurs, le choix de l'espace produit conditionnant les descripteurs générés, l'importance de cette étape ne doit pas être négligée. En parallèle avec cet apprentissage, l'animateur de panel établit la liste des descripteurs qui seront utilisés lors de la phase de mesure. Ce dernier se base sur les résultats obtenus pendant l'entraînement et sur la discussion menée avec les panélistes autour des différentes caractéristiques sensorielles des produits étudiés. La notation s'effectue sur des échelles dont les caractéristiques peuvent varier (continues ou discrètes, de longueur différente, en présence de libellés plus ou moins détaillés).
Les protocoles standards d'entraînement (AFNOR, 1993) préconisent plusieurs épreuves afin que les panélistes apprennent à utiliser correctement les échelles de notation, à se familiariser avec les termes descriptifs et à en donner une interprétation sensorielle en accord avec les autres panélistes. L'animateur de panel veille au contrôle des performances à la fois du panel et des panélistes en mesurant les indices de performances suivant : la discrimination, l'accord des sujets et leur répétabilité.

- Lesschaeve (1997) définit le pouvoir discriminant comme la capacité à détecter une différence d'intensité entre deux ou plus de deux échantillons pour un descripteur donné. Cette

caractéristique est très importante puisqu'elle indique si les panélistes ont été capables de détecter les différences sensorielles entre les produits de l'étude ou non.

- L'accord entre les panélistes mesure l'homogénéité des réponses obtenues pour le même stimulus par les différents dégustateurs.
- La répétabilité traduit la capacité à donner des résultats très proches en réponse à un même stimulus. Celle-ci est appelée reproductibilité lorsque le même stimulus est présenté dans des conditions différentes (changement du lieu ou du moment de l'expérimentation, par exemple). Le contrôle du niveau de répétabilité du panel et des panélistes est indispensable afin de maîtriser la qualité des résultats obtenus.

Le contrôle des performances a été étudié dans la littérature (Bi, 2003; Chambers et al., 2004; Couronne, 1997; McEwan et al., 2002; Pineau et al., 2007; Schlich, 1994)

Lors de la phase de mesure, les sujets doivent noter l'intensité d'un ensemble de descripteurs pour plusieurs produits d'une même catégorie. Les données recueillies permettent d'obtenir un profil descriptif de chaque produit de l'étude. Les sujets ne doivent pas simplement donner un jugement en fonction de leurs préférences, ils doivent décrire et différencier les produits en fonction de leur perception sensorielle, à la manière d'un instrument de mesure. Les échelles linéaires sont les plus souvent utilisées pour ce type de mesure.

1.2.6.2 Le nombre de sujets

Il existe aussi peu de données objectives dans la littérature quant au nombre optimal de sujets d'un panel de profil sensoriel.
La plupart du temps, on trouve des recommandations valables pour un contexte donné voire basées sur une seule expérimentation et les normes ne sont pas claires sur cette question. En somme, le nombre de sujets varie de 5 à 20 sujets.

Des travaux sur le profil sensoriel rapportent l'utilisation d'une dizaine de sujets entraînés. Lundgren et al. (1991) ont fait appel à des panels de 8 à 11 sujets pour leurs études lorsque 6 à 10 sujets ont été sollicités par Molnar (1989), Risvik et al. (1997). Cook and Homer (1996) indiquent que plus le nombre de sujets à enrôler dans un panel descriptif est grand, plus grande est la probabilité de détecter des differences entre produits.

La norme AFNOR (1993) recommande l'utilisation de 8 à 12 sujets en profil sensoriel conventionnel et de 5 à 8 sujets pour établir un profil d'arôme. Par ailleurs, la norme britannique BS (1986) ne donne pas de recommandations quant au nombre de sujets à enrôler mais suggère au minimum 5 panélistes.

D'autre part, certaines recommandations sont spécifiques à un contexte donné. Ainsi, de larges panels d'environ 150 sujets ont été utilisés dans certaines études de profils chimiques

d'odorants (Dravnieks, 1982). D'autre part, Stone and Sidel (2004) suggèrent un nombre de 10 à 12 sujets pour leur méthode QDA© (Quantitative Descriptive Analysis).
Un des objectifs d'une étude de profil sensoriel est d'aboutir à une séparation consistante et significative des produits. L'entraînement des sujets est indispensable pour garantir l'homogénéité du panel et ainsi augmenter les chances de détecter des différences entre les produit comparés. L'entraînement des sujets contribue à réduire l'interaction panélistes-produits. L'entraînement est certes coûteux mais néanmoins indispensable. Il est possible de réduire ce coût de l'entraînement en travaillant avec de petits groupes de sujets. Chambers et al. (1981) ont démontré qu'un panel de seulement 3 panélistes entraînés avait détecté les mêmes différences significatives qu'un panel de 8 à 9 panélistes semi-entraînés.

Souvent, le nombre de sujets en profil sensoriel n'est abordé que sous l'angle de la comparaison de panels de tailles différentes utilisant des techniques différentes (Chambers et al., 1981; Jeltema and Southwick, 1986; King et al., 1995; Muir and Banks, 1993). L'idée de réduire la taille d'un panel existant pour étudier la stabilité des résultats n'est pas souvent exploitée et lorsqu'elle l'est, c'est à chaque fois avec un ou deux jeux de données seulement.
Jeltema and Southwick (1986) concluent sur la base d'un jeu de données, qu'un panel de 20 juges était capable de caractériser des différences entre 35 odorants purs. Les auteurs ont comparé les résultats statistiques obtenus à partir de sous-panels de taille 10, 20 et 30 obtenus par tirage aléatoire à ceux obtenus à partir du panel complet. Les résultats statistiques sont basés sur le calcul de la corrélation entre le vecteur des scores obtenu pour le panel complet et le vecteur des scores pour les sous-panels.
Muir and Banks (1993) rapportent une légère perte d'information lorsqu'on réduit un panel de 24 à 12 sujets. Cependant, restreindre le panel à 6 sujets altère significativement les résultats sur la discrimination des produits.
King et al. (1995) ont mesuré l'impact de la réduction de la taille de panel sur la stabilité des résultats statistiques d'une étude de profil sensoriel sur des glaces réalisé par un panel bien entraîné. Les auteurs ont mesuré le nombre de descripteurs qui devenaient non significatifs au fur et à mesure que l'on réduisait le panel. Ils ont comparé les résultats du profil obtenu avec le panel complet de 20 sujets aux résultats du profil obtenu avec le panel réduit à 10 et 5 sujets. Les résultats montrent que lorsque le panel est réduit de moitié, 67% des descripteurs demeuraient significatifs et seulement 34% lorsque l'on réduisait le panel de trois quarts. Les auteurs avancent qu'il est probablement impossible de définir une taille de panel optimale pour une étude descriptive. Leur étude leur a permis de voir qu'un panel de 5 sujets reproduisait relativement la même carte produit que le panel complet mais ce résultat ne peut être généralisé.
Gacula and Rutenbeck (2006) ont aussi appliqué une approche basée sur la simulation pour déterminer le nombre de sujets en profil sensoriel. Pour une variabilité de 0.49 et pour différentes

simulations de différences de notes d'intensité que l'on pourrait observer sur une échelle 0-15 points, les auteurs concluent que 5 à 14 sujets sont nécessaires pour détecter une différence significative de 0.6.
Les quatres études citées ci-dessus sont toutes basées sur l'analyse d'un seul jeu de données. Ces résultats manquent de robustesse pour qu'ils soient généralisés.

Les notions de la taille de l'effet et de la variabilité des données sont mal définies au niveau multidimentionnel, or le profil sensoriel est avant tout un objet multidimentionnel. En effet, le profil sensoriel est l'étude de différents types de descripteurs. Le type de descripteur a un impact sur le pouvoir discriminant. Pineau (2006) a montré que les descripteurs d'apparence et de texture sont généralement plus discriminants que les descripteurs de saveur, d'arôme ou d'odeur. Cette différence pourrait être attribuée à la difficulté d'assimilation et à la complexité de la perception de ce type de descripteurs. Il semblerait ainsi difficile de donner un nombre optimal de sujets valable quelque soit type de descripteur.

1.2.6.3 Synthèse

Il existe différentes recommandations sur le nombre de sujets à enrôler dans un panel de profil sensoriel. Ce nombre varie de 5 à 30 sujets. Cependant, la plupart des recommandations convergent vers un panel de 10 à 12 sujets.
Stone and Sidel (2004) mentionnent que si la taille de la différence à mettre en évidence est supposée très petite, le panel leader pourrait enrôler un plus grand nombre de sujets ou bien faire plus de répétitions. Par ailleurs, enrôler plus de 20 sujets pourrait engendrer un problème d'organisation. Par exemple, l'organisation des séances d'entraînement en séparant le panel en deux groupes. Dans ce cas précis, le panel leader doit s'assurer que les deux groupes ont assimilé les mêmes notions. Le panel leader doit se servir d'études antérieures pour estimer le nombre de sujets adéquat à enrôler.
La notion du nombre de sujets à enrôler demeure donc très subjective. Les différentes études citées sont spécifiques à l'étude d'un certain type de descripteurs (arôme, odeur, ...).
Nous n'avons pas trouvé de référence qui calcule à priori le nombre de sujets à enrôler pour un panel de profil sensoriel. En outre, la mesure de la taille de l'effet en profil sensoriel n'a pas été clairement définie ou quantifiée.

La figure (1.3) résume les recommandations établies par la norme AFNOR (2003) sur le nombre de sujets à enrôler dans les différents types d'études sensorielles, notamment en profil sensoriel et en tests hédoniques.

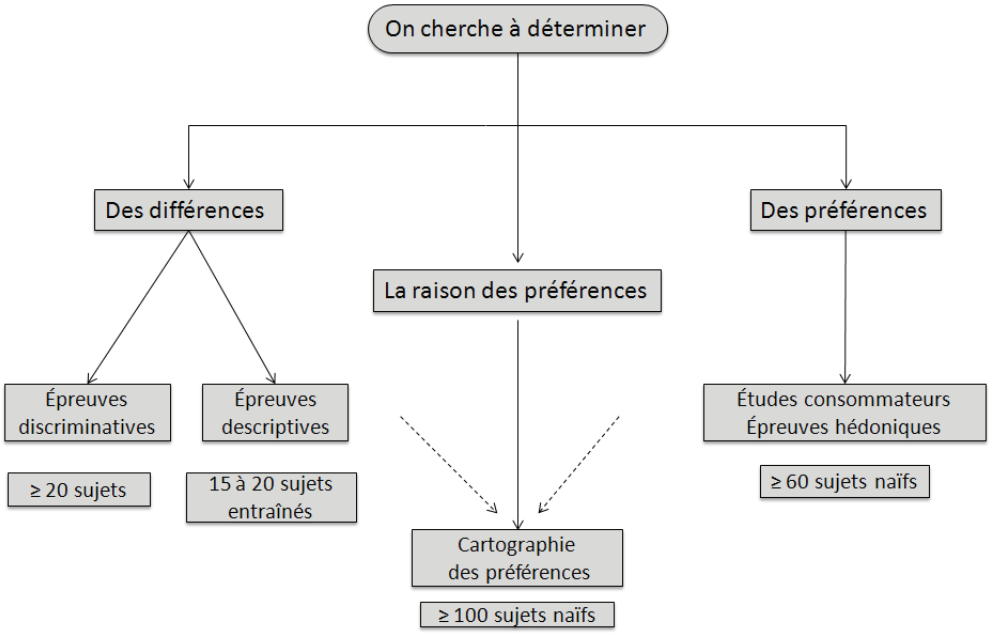


FIG. 1.3 – *Le nombre de sujets dans les différents types d'épreuves sensorielles AFNOR (2003)*

2 Concepts statistiques : Puissance et nombre de sujets nécessaires

2.1 Principe d'un test statistique

Un test statistique permet de décider au vu d'un échantillon, le choix entre deux hypothèses opposées : une hypothèse nulle H_0 contre une hypothèse alternative H_1.
Généralement, l'hypothèse nulle définit l'absence de différence entre les moyennes de deux groupes. Le calcul de la statistique du test et la définition de la région de rejet de l'hypothèse nulle permettent de conclure sur l'existence ou non de cette différence. Le résultat est dit statistiquement significatif si on conclut à l'existence de cette différence.

La décision d'accepter ou de rejeter l'hypothèse nulle est liée à la structure aléatoire de l'échantillon. En général, la probabilité de prendre une décision erronée n'est pas nulle. Il existe deux types d'erreurs dans la prise de décision : l'erreur de type I (α) et l'erreur de type II (β). Le tableau (1.1) illustre ces deux situations :

TAB. 1.1 – *Table de décision*

Décision	Si H_0 est :	
	Vraie	Fausse
Accepter H_0	Pas d'erreur $(1-\alpha)$	Erreur de type II : β
Rejeter H_0	Erreur de type I : α	Pas d'erreur $(1-\beta)$

α : Probabilité de rejeter H_0 lorsqu'elle est vraie

β : Probabilité de ne pas rejeter H_0 lorsqu'elle est fausse

2.2 La puissance statistique

La puissance d'une expérimentation est l'aptitude d'obtenir un résultat statistiquement significatif si une réelle différence entre des groupes comparés existe. C'est la probabilité de rejeter l'hypothèse nulle à raison, c'est-à-dire, lorsqu'on est en vérité dans le cadre de l'hypothèse alternative. La puissance du test est donc le complément de l'erreur de type II, appelée aussi erreur de deuxième espèce β. On la note $1-\beta$ (tableau (1.1)).
La puissance dépend de plusieurs paramètres. Elle est fonction de l'erreur de type I (α), fixée au préalable, de la taille de l'effet que l'on désire mettre en évidence, de la variabilité de la population étudiée et de la taille de l'échantillon (nombre d'observations).

Intuitivement, la puissance est similaire au pouvoir grossissant d'un microscope. Un grossissement suffisant est nécessaire pour montrer que deux points très proches l'un de l'autre, mais cependant séparés, sont distincts. Avec un grossissement insuffisant, ces deux points paraissent ne faire qu'un. Plus la distance entre les deux points est petite, plus le pouvoir grossissant devra être élevé pour visualiser deux points distincts. Il en est de même avec la recherche d'une différence entre deux groupes. La magnitude de cette différence est la taille de l'effet à mettre en évidence. Une puissance statistique suffisante est nécessaire pour montrer qu'il existe effectivement une différence entre les deux groupes. Plus la différence entre les deux groupes est petite, plus il faudra de puissance statistique pour montrer que les deux groupes sont différents.
La question de puissance constitue un pivot pour toute experimentation. En effet, l'expérimentateur se voit souvent confronté au besoin d'évaluer la puissance statistique associée à son étude ou en d'autres termes de déterminer le nombre d'observations nécessaires pour atteindre une puissance suffisante.
Le nombre d'observations est le paramètre sur lequel l'expérimentateur peut le plus directement agir pour contrôler la puissance de son étude. En particulier lorsque l'effet recherché est petit, il est nécessaire d'inclure un grand nombre de sujets. En revanche, un effectif plus faible est suffisant pour mettre en évidence des effets conséquents.

Un test statistique en analyse sensorielle consiste le plus souvent à :

- vérifier l'existence d'un effet produit;
- vérifier si les différences en termes d'intensité ou d'appréciation sont importantes entre les produits;
- contrôler l'uniformité des réponses fournies par les sujets. En profil sensoriel, on vérifie si les réponses des sujets sont homogènes à la fois pendant l'entraînement et pendant la phase de mesure. En test hédonique, ce contrôle permet de détecter l'existence ou non d'une hétérogénéité des préférences.

Les tests paramétriques les plus couramment utilisés pour l'analyse des données hédoniques et de profils sensoriels sont : le test de comparaison de deux moyennes (test t de Student) et le test de l'analyse de la variance (ANOVA).

Les sections suivantes présentent le principe de ces tests et le calcul de la puissance dans chacun de ces cas.

2.2.1 Comparaison de deux moyennes

On considère le test de comparaison de moyennes de deux groupes, avec n observations par groupe. On teste alors :
H_0 : "$\mu_1 = \mu_2$" contre H_1 : "$\mu_1 \neq \mu_2$", ou encore H_0 : "$\mu_1 - \mu_2 = 0$" contre H_1 : "$\mu_1 - \mu_2 = d$",$d > 0$

On se place dans la situation la plus fréquente où la variance des deux groupes est supposée connue, égale à σ^2, et elle est commune aux deux groupes.
On rejette H_0 au risque α si la quantité

$$\mid Z \mid = \frac{\mid \overline{x}_1 - \overline{x}_2 \mid}{\sqrt{\frac{2\sigma^2}{n}}} \tag{1.1}$$

est supérieure ou égale à $z_{\frac{\alpha}{2}}$; avec $z_{\frac{\alpha}{2}}$ quantile d'ordre $\frac{\alpha}{2}$ de la loi normale centrée réduite.

La figure (1.4) illustre la distribution de Z selon que l'on soit sous H_0 ou sous H_1

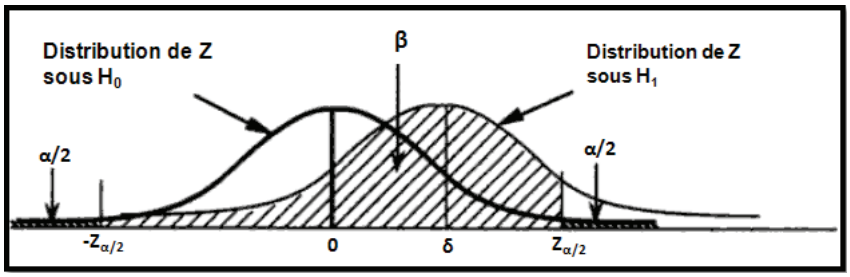


FIG. 1.4 – *Distribution de Z sous H_0 et sous H_1*

Les risques α et β s'expriment de la façon suivante :

- α est la probabilité de rejeter H_0 lorsqu'elle vraie; c'est-à-dire :

$$\alpha = P(\mid Z \mid \geq z_{\frac{\alpha}{2}} \mid H_0) \tag{1.2}$$

- β est la probabilité de ne pas rejeter H_0 lorsqu'elle est fausse (H_1 vraie); c'est-à-dire :

$$\beta = P(\mid Z \mid < z_{\frac{\alpha}{2}} \mid H_1) \tag{1.3}$$

La puissance du test est alors :

$$Puissance = 1 - \beta = Prob(\mid Z \mid \geq z_{\frac{\alpha}{2}} \mid H_1) \tag{1.4}$$

Le nombre de sujets nécessaire par groupe pour détecter une différence d ($d = \mu_1 - \mu_2$) est donné par la formule (1.5)

$$n = \frac{2\sigma^2(z_{\frac{\alpha}{2}} + z_\beta)^2}{d^2} \tag{1.5}$$

2.2.2 Analyse de la variance à un facteur

On considère le test F de comparaison de moyennes de k groupes, avec n observations par groupe. On teste :
H_0 : "$\mu_1 = \mu_2 = \ldots = \mu_k$" contre H_1: "Il existe au moins μ_i, μ_j $(j \neq i)$", tel que $\mu_i \neq \mu_j$
Le modèle sous-jacent s'écrit :

$$x_{ij} \quad = \quad \mu + \alpha_i + \varepsilon_{ij}, \;\; \varepsilon_{ij} \;\sim\; \mathcal{N}(0,\sigma^2) \quad i.i.d, \quad \mu_i = \mu + \alpha_i \tag{1.6}$$

TAB. 1.2 – *Tableau d'ANOVA à un facteur*

Source de variation	ddl	Somme des carrés	Carrés moyens	Statistique de Fisher
Facteur	$k-1$	$n\sum_{i=1}^{k}(x_{i.}-x_{..})^2$	$CMF=\frac{SCF}{k-1}$	$F=\frac{CMF}{CME}$
Erreur	$k(n-1)$	$\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-x_{..})^2$	$CME=\frac{SCE}{k(n-1)}$	

n effectif de chaque niveau i

$\overline{x}_i$ moyenne observée pour le niveau i

On rejette H_0 au risque α si la statistique

$$F=\frac{SCF/(k-1)}{SCE/(k(n-1))}=\frac{CMF}{CME} \tag{1.7}$$

est supérieure au fractile $F_{1-\alpha}(k-1,k(n-1))$.

Sous l'hypothèse H_1, la statistique de F suit une loi de Fisher-Snedecor non centrée à k - 1 et k(n - 1) degrés de liberté et de paramètre de non-centralité λ défini par :

$$\lambda=n\sum_{i=1}^{k}\left(\frac{\alpha_i}{\sigma}\right)^2=n\sum_{i=1}^{k}\left(\frac{(\mu_i-\mu)}{\sigma}\right)^2=nk\frac{\sigma_m^2}{\sigma^2}=nkf^2 \tag{1.8}$$

Où,

$$\sigma_m=\sqrt{\frac{1}{k}\sum_{i=1}^{k}(\mu_i-\mu)^2},\qquad \mu=\frac{1}{k}\sum_{i=1}^{k}\mu_i \tag{1.9}$$

Ainsi le calcul de la puissance du test F en fonction de λ pour un risque α :

$$Puissance=1-\beta=Prob(F(k-1,k(n-1),\lambda)\geq F_{1-\alpha}(k-1,k(n-1)) \tag{1.10}$$

Le paramètre de non-centralité est une fonction de la taille de l'effet standardisée f (standardized effect size Cohen (1988)) qui correspond au rapport $f=\frac{\sigma_m}{\sigma}$.

La variance de la population σ^2 est estimée par le carré moyen de l'erreur CME (MSE Mean square of Error).

Le calcul du nombre de sujets est un processus itératif. On attribue une valeur initiale à n, ensuite on augmente ce nombre n, jusqu'à atteindre la puissance statistique désirée.

2.2.3 Analyse de la variance à deux facteurs

Le modèle d'ANOVA à deux facteurs Produit et Sujet est le modèle le plus souvent utilisé en analyse sensorielle. Lorsque les données ne comportent pas de répétitions, le modèle d'ANOVA au niveau du groupe est le suivant :

$$x_{ij} = \mu \ + \ \alpha_i \ + \ \beta_j \ + \ \varepsilon_{ij}, \ \ \beta_j \ \sim \ \mathcal{N}(0,\sigma_\beta^2), \ \ \varepsilon_{ij} \ \sim \ \mathcal{N}(0,\sigma^2), \ \ i.i.d \tag{1.11}$$

Où μ représente l'effet moyen global, α_i correspond à l'effet fixe produit, β_j à l'effet aléatoire sujet et ε_{ij} à l'erreur résiduelle (aléatoire) du modèle. Les variables aléatoires sont supposées être indépendantes deux à deux et distribuées selon une loi Normale, chacune avec sa variance spécifique. Les calculs des degrés de liberté, des sommes des carrés ainsi que les statistiques de Fisher sont donnés dans le tableau 1.3 :

TAB. 1.3 – *Tableau d'ANOVA à deux Facteurs Produit et Sujet*

Source de variation	ddl	Somme des carrés	Carrés moyens	Statistique de Fisher
Produit	k - 1	$SCP = n\sum_{i=1}^{k}(x_{i.} - x..)^2$	$CMP = \frac{SCP}{k-1}$	$F_P = \frac{CMP}{CME}$
Sujet	n - 1	$SCS = k\sum_{j=1}^{n}(x_{.j} - x..)^2$	$CMS = \frac{SCS}{n-1}$	$F_S = \frac{CMS}{CME}$
Erreur	(k - 1) (n - 1)	$SCE = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - x_{i.} - x_{.j} + x..)^2$	$CME = \frac{SCE}{(k-1)(n-1)}$	

Le principe du calcul de puissance est le même que dans le cas de l'ANOVA à un facteur avec pour différence, le nombre de degrés de libertés pour le terme d'erreur.

Sous l'hypothèse nulle H_0, la statistique de l'effet produit F_p suit une loi de Fisher centrée à (k - 1,(k - 1)(n - 1)) degrés de liberté. Sous l'hypothèse H_1, la statistique de l'effet produit suit une loi de Fisher non centrée à (k-1) et (k - 1)(n - 1) degrés de liberté, de paramètre de non centralité $\lambda \ = \ n \ k \ f^2$, où f est la taille de l'effet.

Ainsi le calcul de la puissance du test F en fonction de λ pour un risque α :

$$Puissance = 1 - \beta = Prob(F(k-1,(k-1)(n-1),\lambda) \geq F_{1-\alpha}(k-1,(k-1)(n-1))) \tag{1.12}$$

2.3 Taille de l'effet : Mesure et variation

Généralement nous considérons qu'une puissance de 0.8 est satisfaisante; dans certains cas nous visons même une puissance de 0.9. Nous pouvons utiliser la formule (1.12) pour déterminer le nombre d'observations nécessaires pour une puissance fixée.

Il faut néanmoins avoir une idée de la valeur minimale que peut prendre la somme $\sum_{i=1}^{k} \alpha_i^2$ et la valeur que peut avoir σ^2. Ces valeurs doivent être déterminées par un expert du domaine considéré (par exemple le panel leader).
Dans ce type d'étude prospective, la situation est complexe par le fait qu'il est difficile d'évaluer le terme $\sum_{i=1}^{k} \alpha_i^2$.
Selon Cohen (1988), on peut observer trois types de dispersion des k moyennes μ_i sur l'intervalle défini par $[\mu_{min}, \mu_{max}]$ en fonction de l'intervalle des moyennes standardisées δ, tel que $\delta = \dfrac{\mu_{max} - \mu_{min}}{\sigma} = \dfrac{d}{\sigma}$:

- dispersion minimale (minimum variability) : deux moyennes extrêmes et les autres moyennes sont situées au milieu de l'intervalle défini par les deux moyennes extrêmes
- dispersion moyenne (medium variability) : les moyennes sont équiréparties sur l'intervalle défini par les deux moyennes extrêmes
- dispersion maximale (maximum variability) : toutes les moyennes sont situées sur les extrêmes de l'intervalle à proportions égales.

Ainsi, dans le cas de la première configuration, nous avons deux moyennes μ_i et μ_j $(i \neq j)$ qui sont éloignés de d et les autres moyennes sont égales à la moyenne μ des deux μ_i et μ_j. Tous les effets sont donc nuls sauf deux, α_i et α_j, pour lesquels il existe un écart en valeur absolue égal à d. Ainsi, nous avons $|\alpha_i| = |\alpha_j| = \dfrac{d}{2}$.
Nous obtenons alors :

$$f = \frac{\sigma_m}{\sigma} = \delta\sqrt{\frac{1}{2k}} \tag{1.13}$$

Ainsi :

$$\lambda = n\sum_{i=1}^{k}\left(\frac{\alpha_i}{\sigma}\right)^2 = nkf^2 = nk\frac{\sigma_m^2}{\sigma^2} = n\frac{\delta^2}{2} \tag{1.14}$$

Dans la seconde configuration où les moyennes sont réparties uniformément sur l'intervalle défini par δ, la taille de l'effet est définie par:

$$f = \frac{\sigma_m}{\sigma} = \frac{\delta}{2}\sqrt{\frac{k+1}{3(k-1)}} \tag{1.15}$$

Ainsi, le paramètre de non-centralité est équivaut à :

$$\lambda = n\sum_{i=1}^{k}\left(\frac{\alpha_i}{\sigma}\right)^2 = nkf^2 = nk\frac{\sigma_m^2}{\sigma^2} = n\left(\frac{k(k+1)}{12(k-1)}\right)\delta^2 \tag{1.16}$$

La troisième configuration correspond à celle où toutes les moyennes sont réparties aux extrêmes de l'intervalle défini par δ. Cohen (1988) défini deux mesures de la taille de l'effet en fonction de la nature de k.

- Si k est pair : on aura $\frac{k}{2}$ moyennes réparties sur chacune des bornes de l'intervalle défini par δ
- Si k est impair : on aura $\frac{k+1}{2}$ situées à une borne de l'intervalle défini par δ et $\frac{k-1}{2}$ moyennes située à l'autre borne de l'intervalle.

Ainsi, si k est pair, la taille de l'effet est définie par :

$$f = \frac{\sigma_m}{\sigma} = \frac{\delta}{2} \tag{1.17}$$

Le paramètre de non-centralité est alors :

$$\lambda = n\sum_{i=1}^{k}\left(\frac{\alpha_i}{\sigma}\right)^2 = nkf^2 = nk\frac{\sigma_m^2}{\sigma^2} = nk\frac{\delta^2}{4} \tag{1.18}$$

Si k est impair, la taille de l'effet est équivaut à :

$$f = \frac{\sigma_m}{\sigma} = \delta\frac{\sqrt{k^2-1}}{2k} \tag{1.19}$$

Ainsi, le paramètre de non-centralité est équivaut à :

$$\lambda = n\sum_{i=1}^{k}\left(\frac{\alpha_i}{\sigma}\right)^2 = nkf^2 = nk\frac{\sigma_m^2}{\sigma^2} = n\frac{k^2-1}{4k}\delta^2 \tag{1.20}$$

Après cette synthèse bibliographique sur les tests statistiques, la notion et le calcul de la taille de l'effet, nous abordons la question du nombre de sujets dans un domaine proche de l'analyse sensorielle à savoir, les essais cliniques.

2.4 Puissance et nombre de sujets dans les essais cliniques

Les essais cliniques constituent un préalable indispensable et obligatoire avant la mise sur le marché d'un médicament. Ils se déroulent en 4 phases. La phase I et II ont pour objectif l'identification du profil de tolérance chez des sujets volontaires sains et les principaux éléments de tolérance dans la population cible de patients. Ensuite, La phase III constitue la phase la plus importante, elle correspond à l'essai thérapeutique proprement dit. Elle a pour objectif l'identification assez fine de la tolérance dans la population cible de malades, et la comparaison par rapport aux traitements de référence. Enfin, la phase IV ou phase de pharmaco-vigilance qui intervient pour identifier les effets indésirables et rares non détectés parmi les patients de l'essai. La détermination du nombre de sujets nécessaire pour mettre en évidence l'efficacité d'un traitement intervient lors de la phase III. La puissance dépend du nombre de sujets inclus dans l'étude. Un essai peu puissant n'a généralement pas d'intérêt car il a peu de chance de mettre en évidence l'effet du traitement. Il représente donc un investissement non rentable. Afin de ne pas réaliser des essais sans intérêt, il convient de leur assurer une puissance statistique suffisante.
On retrouve les mêmes considérations pour la détermination du nombre de sujets qu'en analyse sensorielle. Le nombre de sujets dépend de l'objectif principal de l'étude, du test statistique utilisé, de la taille de l'effet à mettre en évidence, de l'erreur de type I et de l'erreur de type II.
La taille de l'effet du traitement est souvent difficile à déterminer. Cette grandeur est généralement estimée à partir d'une méta-analyse des données de traitements similaires à celui qui fait l'objet de l'essai. Plus la magnitude de l'effet à détecter est faible, plus le nombre de sujets à inclure sera important. La norme ICH (Lewis, 1999) exige la détermination du nombre de sujets à priori. Ce nombre doit figurer dans le protocole de l'étude en même temps que les estimations des différents paramètres utilisés pour son calcul.

Il existe trois types d'essais cliniques : les tests de supériorité, les tests d'équivalence et les tests de non-infériorité. Ces trois types de tests sont regroupés sous le nom général de tests d'équivalence.
Dans la première catégorie, on cherche à mettre en évidence un effet traitements par rapport à un placebo. L'essai de supériorité exige la définition d'une différence minimale d'efficacité "d" que l'on veut mettre en évidence avec une puissance suffisante.
Dans la seconde catégorie, on cherche à montrer si deux traitement sont thérapeutiquement équivalents. Ces essais d'équivalence exigent la définition d'une marge d'équivalence à partir d'une grandeur "d" qui correspond à la plus grande perte d'efficacité que l'on peut tolérer pour conclure que l'un des traitements n'est pas inférieur à l'autre.
Dans la troisième catégorie, on cherche à montrer qu'un traitement n'est pas plus mauvais qu'un autre traitement de la même famille. On veut montrer dans ces essais de non-infériorité que

les deux traitements sont comparables, mais on accepte que le traitement soit un peu moins "efficace" que la référence.
Les tests de non infériorité et d'équivalence sont mis en place lorsqu'il existe déjà un traitement comparateur ayant fait la preuve de son efficacité contre un placebo ou bien lorsque l'innovation n'est pas supérieure en efficacité, mais apporte d'autres avantages en termes de tolérance au médicament, de coût ou de facilité d'utilisation.
Les trois types d'essais se basent sur le test statistique de Student de comparaison de deux moyennes. Le test peut être bilatéral ou unilatéral selon les trois différents types d'essais. En outre, pour chaque test, la taille de la différence que l'on désire mettre en évidence doit être définie par les cliniciens (Bouvenot and Villani, 2000).

2.5 Synthèse

Les essais cliniques sont souvent réalisés dans le but de comparer deux groupes de sujets : un groupe traitement et un groupe contrôle. La notion de la taille de l'effet traitement est plus facile à saisir car elle définit une différence entre les moyennes de deux groupes. Cette taille de l'effet est estimée à partir de l'historique des différents essais réalisés sur ce même traitement ou bien de l'historique des essais sur des traitements similaires.

Les tests d'équivalence sont souvent appelés "tests de similitude" (AFNOR, 2009) en analyse sensorielle. Ils ont pour objectif de démontrer si deux produits sont suffisamment proches l'un de l'autre pour être substituable l'un à l'autre.
Par ailleurs, pour déclarer que deux produits sont équivalents, une marge d'équivalence "d" est à définir à priori par les commanditaires de l'étude. Le nombre de sujets est calculé en se basant sur le test de comparaison de deux moyennes avec les hypothèses appropriées.

Cette synthèse bibliographique sur les essais cliniques, nous a permis de voir ce qui se fait autour du nombre de sujets et des tests de comparaison de moyennes. Il existe de nombreuses références autour des essais d'équivalence qui pourraient être exploitées pour les tests de similitude en analyse sensorielle. Cependant, nous n'avons pas trouvé des références concernant la comparaison de plusieurs groupes de moyennes qui font souvent l'objet des études hédoniques et des études descriptives. En effet, ces tests consistent généralement à comparer plusieurs produits ($k \geq 2$). La grandeur de l'effet est donc complexe à définir pour ces deux types d'études. Les analystes sensoriels ne peuvent donc pas toujours prédire la position du produit cible par rapport aux autres produits comparés. Des suggestions seront proposées pour répondre à cette question dans les chapitres suivants.

Chapitre 2

Questions de recherche et présentation de la démarche

1 Identification des axes et questions de recherche

La spécificité de l'analyse sensorielle réside dans le fait que l'instrument de mesure utilisé est l'homme dont les appréciations sont subjectives. De plus, nous savons que nous ne percevons pas les stimuli sensoriels de la même manière et avec la même intensité. Lorsqu'on veut comparer plusieurs produits, il faudrait solliciter un nombre suffisant de sujets, pour pouvoir ensuite, autant que possible, extrapoler les résultats à la population des individus concernés. La détermination du nombre de sujets nécessaire pour comparer plusieurs produits est donc une étape importante dans la mise en place d'un test d'analyse sensorielle.

Dans une épreuve d'analyse sensorielle, le test consiste le plus souvent à établir l'existence ou non de différences entre plusieurs produits. En profil sensoriel, cette question se traduit par l'existence d'une différence en termes d'intensité perçue pour tout ou une partie des descripteurs constituant le profil. En tests hédoniques ou d'acceptabilité, on cherche à établir l'existence d'une différence en termes d'appréciation des produits testés.
Ainsi, un test sensoriel est dit suffisamment puissant si la probabilité d'établir l'existence d'une différence entre les produits est forte (différence en termes d'intensité ou d'appréciation). La puissance est donc la probabilité de mettre en évidence cette différence, si elle existe, entre les produits.

La littérature indique plusieurs recommandations quant au nombre de sujets à enrôler dans un panel hédonique ou descriptif assurant une puissance donnée. Ces recommandations diffèrent et résultent parfois d'une simple intuition.

Les deux démarches indiquées pour déterminer le nombre de sujets sont : une démarche théorique classique et une démarche par rééchantillonnage.

La démarche classique est une démarche prospective qui propose de calculer le nombre de sujets à partir des valeurs fixées de l'erreur de type I (α) et de l'erreur de type II (β), d'une quantification de la différence à mettre en évidence d et d'une estimation de la variabilité σ.
Nous avons ainsi trouvé des références donnant une estimation de la variabilité pour les tests hédoniques (Gacula, 1993; Hough et al., 2006; Lawless and Heymann, 1998). Les auteurs estiment la variabilité σ à $0.23 - 0.25$ sur une échelle 0-1. Pour calculer le nombre de sujets, Hough et al. (2006) ont supposé une différence à mettre en évidence de l'ordre de 0.8 sur une échelle hédonique en 9 points. Cependant, les auteurs n'ont pas donné des précisions quant à la formule utilisée pour effectuer ces calculs. Nous ne savons pas si la formule est basée sur un test de comparaison de deux moyennes (test de Student) ou bien sur un test de comparaison de plusieurs moyennes (ANOVA).
Gacula (1993) et Lawless and Heymann (1998) déterminent le nombre de sujets en considérant un test de comparaison de deux moyennes (entre deux produits). Or, dans la pratique on compare souvent plus que deux produits.
Nous allons donc essayer de répondre à cette question et apporter une solution à la quantification de d dans le cas de comparaison de plus de 2 produits.

D'autre part, la démarche par rééchantillonnage consiste à déduire la plus petite taille de sous-panel qui enregistre une perte d'information acceptable par rapport à la comparaison des produits obtenue à partir d'un panel complet. Une telle démarche nécessite une application sur un grand nombre de jeux de données pour couvrir les différentes situations en tests sensoriels. Or, la plupart des études citées utilisant cette démarche sont basées sur un faible nombre de jeux de données. Cette démarche ainsi appliquée manque de robustesse quant aux recommandations obtenues.

2 Présentation de la démarche

L'essence et l'originalité de ce travail réside l'utilisation des bases de données PrefBase et SensoBase pour donner des recommandations sur le nombre de sujets à enrôler dans une étude sensorielle (hédonique ou descriptive). L'étude de centaines de jeux de données cumulés constituera un "retour d'expériences" pour estimer les paramètres qui déterminent le nombre de sujets nécessaire, étudier leurs variations et l'impact de ces variations sur les recommandations.

2.1 Démarche théorique prospective

L'objectif est de proposer des abaques du nombre de sujets à enrôler utiles pour les analystes sensoriels. Cette démarche consiste à calculer le nombre de sujets à partir de la quantification de la taille de la différence d que l'on souhaite mettre en évidence, de la variabilité σ de la mesure sensorielle étudiée, du risque de première espèce α et du risque de deuxième espèce β.
La variabilité σ est estimée par l'écart-type de l'erreur (RMSE) dans le modèle de l'analyse de la variance à deux facteurs Sujet + Produit .
Ensuite, la question est de quantifier d la différence à mettre en évidence lorsqu'on compare k produits ($k > 2$). La norme AFNOR (2009) pour les tests hédoniques considère d comme étant la distance en termes d'appréciation qui sépare les moyennes des deux produits extrêmes. Dans ce cas, il faudrait aussi faire des hypothèses sur la répartition des moyennes des autres produits entre ces deux extrêmes (selon les hypothèse de Cohen (1988)). Cependant, les analystes sensoriels ne peuvent pas connaître a priori la position qu'auraient les produits à comparer. De plus, si le produit cible qui intéresse le commanditaire de l'étude ne se situe pas à l'un des deux extrêmes, la distance recherchée ne concernera pas ce dernier.
Nous proposons de prendre pour d la racine carrée de la variance des moyennes des produits issue du modèle ANOVA à deux facteurs Sujet + Produit. Cette distance ainsi définie concerne potentiellement tous les produits (qu'ils soient extrêmes ou pas) et ne nécessite pas d'hypothèse sur la répartition des moyennes des produits.
La quantification de d et de σ sera obtenue pour chaque jeu de données. Nous allons ensuite étudier leur variabilité sur l'ensemble des jeux de données avant de proposer des recommandations sur le nombre de sujets.
La figure (2.1) illustre le calcul du nombre de sujets en fonction des différents paramètres.

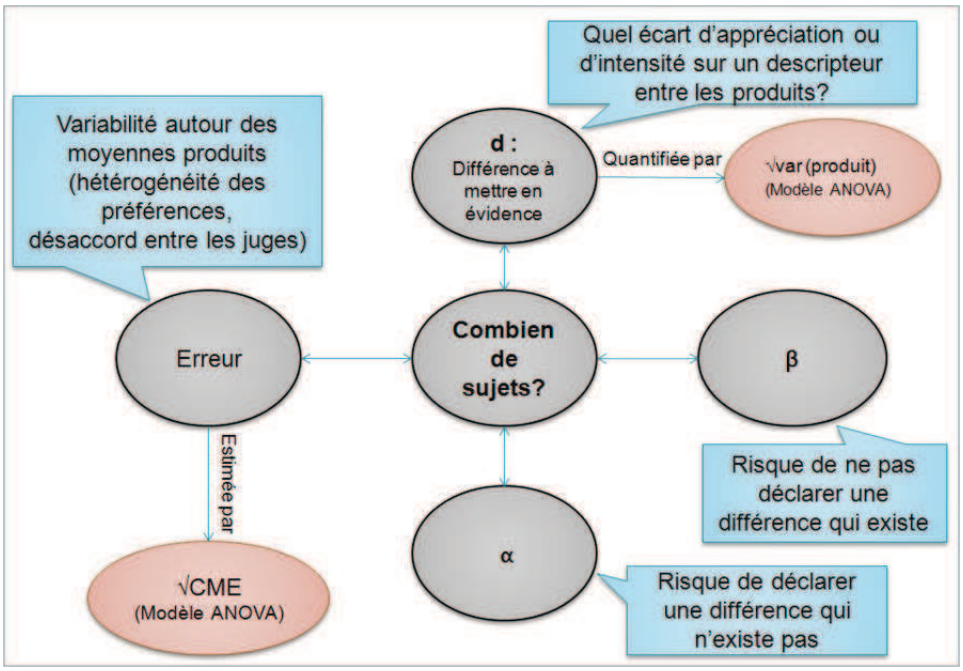


FIG. 2.1 – *Démarche théorique prospective*

2.2 Démarche par rééchantillonnage

Pour chaque jeu de données, il est toujours possible de simuler un "sous-panel" obtenu par tirage au sort de $N - k$ sujets parmi les N sujets composant le panel complet. Il suffit ensuite de comparer les résultats obtenus sur la base du sous-panel avec ceux du panel complet pour établir dans quelle mesure la réduction de la taille du panel à $N - k$ sujets a altéré ou non les résultats de l'étude.

Les critères de comparaison des résultats sont définis en fonction des analyses statistiques appliquées (unidimensionnelles et multidimensionnelles) et prennent compte du type de l'étude. Cette procédure est réalisée sur un grand nombre de tirages de sous-panels de taille $N - k$. Finalement, en commençant avec $k = 0$ et en augmentant k tant que les résultats obtenus avec le panel complet ne sont pas altérés, on peut espérer découvrir le nombre de sujets minimal qui aurait donné les mêmes conclusions qu'avec le panel complet pour ce jeu de données particulier.

La figure (2.2) illustre le procédé de la démarche par rééchantillonnage :

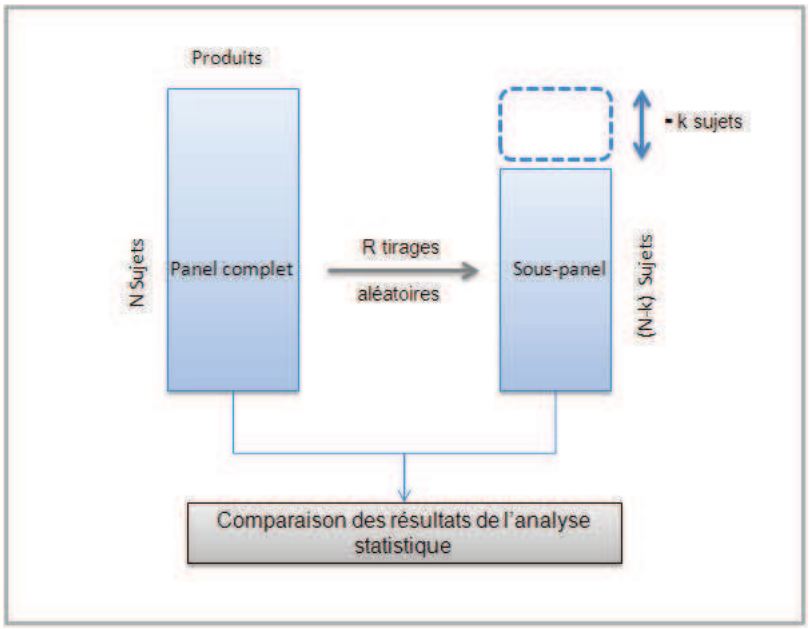


FIG. 2.2 – *Démarche par rééchantillonnage*

Cette approche n'aurait aucune valeur, si elle n'était menée que sur un seul jeu de données. Elle sera ici mise en uvre sur un grand nombre de jeux de données couvrant la variété des situations d'analyse sensorielle en exploitant les deux bases de données PrefBase et SensoBase.

2.3 Le nombre de répétitions en profil sensoriel

Dans le cadre des études de profils sensoriels, il a souvent été admis que les répétitions amélioraient la fiabilité et la validité des résultats des tests. Les répétitions sont bien sûr nécessaires en phase d'entraînement pour évaluer les performances à la fois de chaque sujet et du panel entier. Cependant, une fois les sujets entraînés, nous ne savons pas si ces répétitions amélioreraient réellement la discrimination entre les produits lors de la phase de mesure. Nous allons essayer de répondre à cette question en prenant des jeux de données de SensoBase ayant un plan avec 2 ou 3 répétitions pour examiner la nécessité de ces répétitions.

Par ailleurs, le nombre de répétitions est lié à la thématique du nombre de sujets. Statistiquement, toute augmentation du nombre de sujets ou du nombre de répétitions augmente la puissance du test à mettre en évidence des différences entre produits. Ce lien nous amène à s'interroger sur quel plan s'appuyer pour réaliser une étude de profil sensoriel puissante avec un coût raisonnable. Une discussion des résultats sur la nécessité des répétitions et le lien avec le nombre de sujets en profil sensoriel sera abordée. Cette discussion portera sur le choix à adopter,

qui consisterait soit à prendre davantage de sujets, soit à privilégier des petits groupes avec un nombre de répétitions donné.

Chapitre 3

Le nombre de sujets en tests hédoniques

1 Description de la base de données PrefBase

Le projet PrefBase a été lancé au début de cette thèse, en Octobre 2008. L'objectif était de réunir un nombre maximum de jeux de données et de les analyser pour pouvoir répondre à la problématique du nombre de sujets en tests hédoniques. En outre, les centres ACTIA partenaires du projet souhaitaient pouvoir utiliser cette base pour définir des limites d'acceptabilité des produits. Il s'agit de comparer le niveau des notes hédoniques d'une gamme de produits à celui obtenu par des produits du même type enregistrés dans la base de données. Ceci permet d'émettre une estimation sur le niveau d'*appréciation absolue* de chaque produit d'un nouveau jeu de données. PrefBase est actuellement alimentée par les partenaires des centres ACTIA mais sera prochainement ouverte à d'autres fournisseurs avec la mise en service du logiciel TimeSens©. Un package R pour l'exploitation automatique de ces données hédoniques est en développement. Pour un jeu de données déposé, le fournisseur aura un retour statistique sur l'effet produit observé, le positionnement en termes de préférence des produits testés par rapport aux produits de la même famille testés dans d'autres études, les caractéristiques des sujets selon les préférences observées, etc.

1.1 Architecture de PrefBase

Le système PrefBase offre une interface utilisateur en HTML/PHP. Cette interface permet aux fournisseurs d'envoyer leur données sur les études hédoniques qui sont ensuite stockées dans une base de données MySQL.
Afin de faciliter le stockage dans une structure unique, les données relatives à chaque étude sont importées via FIZZ© à partir d'une méthode d'importation spécifique. Différents formulaires

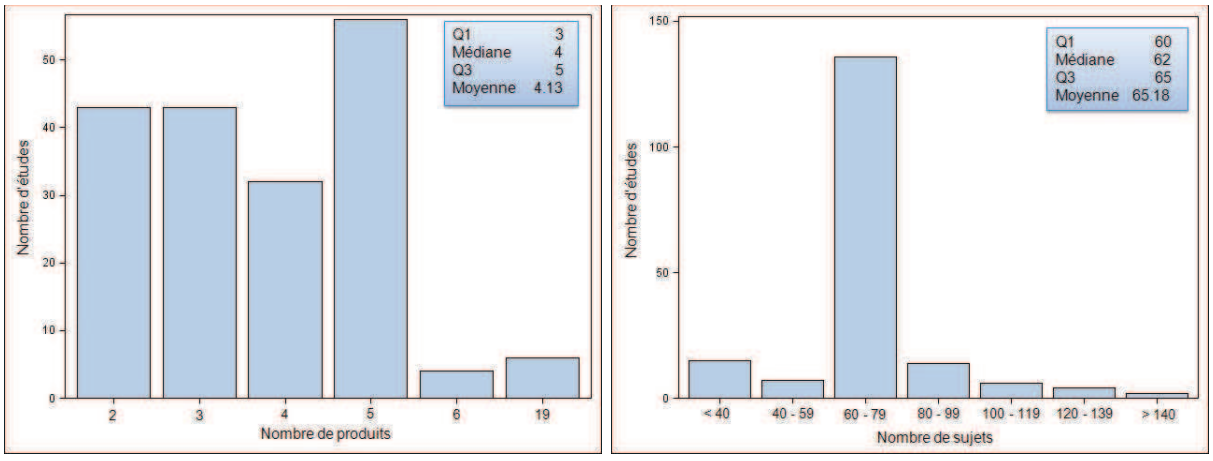
guident le fournisseur à déposer son jeu de données. Ces formulaires lui demandent de fournir divers renseignements sur l'étude sensorielle (informations sur les sujets, sur les produits testés, ...) puis de saisir les données selon une structure précise. Les objets sont stockés dans des variables de session. La base de données n'est modifiée qu'une fois que tous les objets sont correctement renseignés.

PrefBase est organisée en 17 tables. Les tables principales sont :

- la table "Données" : elle contient les différents jeux de données. Chaque enregistrement correspond au score d'appréciation que le sujet "S" a donné au produit "P".
- la table "Produit" : elle rassemble toutes les informations liées au produit (identifiant produit, famille d'aliment, description, ...),
- la table "Sujet" : elle contient toutes les informations sur les sujets (identifiant sujet, date de naissance, sexe, ...)
- la table "Aliment" : permet de catégoriser les familles de produits organisées par une structure hiérarchique.

1.2 Les données PrefBase

PrefBase contient 184 jeux de données. Ils ont été fournis par les 7 membres du réseau ACTIA et par le CSGA. Les données représentent un ensemble de 11 994 sujets et 761 produits, soit un total de 48 659 notes d'appréciation globale.
Les graphiques (a) et (b) de la figure 3.1 représentent les distributions du nombre de produits et du nombre de sujets par étude.



(a) distribution du nombre de produits par étude

(b) distribution du nombre de sujets par étude

FIG. 3.1 – *Distribution du nombre de produits et du nombre de sujets*

Nous remarquons qu'en moyenne 4 produits sont testés lors d'un test hédonique. On fait appel en moyenne à 65 sujets. Ce nombre correspond à la recommandation de l'ancienne norme (AFNOR, 2000), soit 60 sujets à enrôler pour un test hédonique.

PrefBase répertorie les aliments suivant une arborescence à 14 niveaux de grandes familles de produits. Chaque grande famille se décompose en sous-familles donnant ainsi lieu à une structure multi-niveaux. Par exemple, la famille Fruits (frais, sec, compotes, confitures, hors jus de fruit) se subdivise en 3 niveaux :

- fruits frais ou en conserve
- fruits secs, amandes, noix, graines
- fruits en compote, confiture

Chacun de ces niveaux se subdivise à leurs tour en sous-familles, ... etc.

Ce chapitre comporte deux parties. Tout d'abord, nous présentons les résultats de l'approche par rééchantillonnage appliquée sur les 7 jeux de données issus de l'expérimentation de la nouvelle norme AFNOR (2009) sur le nombre de sujets en tests hédoniques. Ensuite, nous présentons la démarche théorique prospective consistant à calculer le nombre de sujets nécessaire en fonction de la taille de l'effet et de l'hétérogénéité des préférences observées sur les 184 jeux de données de PrefBase.

2 Approche par rééchantillonnage

Nous avons appliqué la démarche par rééchantillonnage sur des cas concrets de tests hédoniques dans le cadre de l'expérimentation de la nouvelle norme AFNOR (2009). L'objectif était de tester dans quelle mesure les deux tailles de panels préconisées par la norme, conduisent à des résultat différents ou non. Il s'agissait aussi de quantifier les paramètres qui conditionnent le nombre de sujets, à savoir la taille de l'effet produit et l'hétérogénéité des préférences due aux sujets.

2.1 Étude de l'impact du nombre de sujets en tests hédoniques et expérimentation de la norme XPV09500(2009)

2.1.1 Contexte et objectif

La norme AFNOR (2000) recommandait au minimum 60 sujets à enrôler pour réaliser un test hédonique. Récemment, cette limite de 60 est passée à 100 sujets (AFNOR, 2009). Cette nouvelle version précise que lorsqu'on veut déterminer le nombre de sujets nécessaire, il faudrait estimer la différence que l'on souhaite mettre en évidence et avoir une quantification de l'erreur de mesure, c'est-à-dire la variabilité des données. Dans le cas où le laboratoire ne possède pas ces deux estimations, il faudrait alors prendre par défaut 100 sujets.
Au début de ce travail de thèse, il y avait trop peu de jeu de données sur PrefBase pour pouvoir appliquer l'approche par rééchantillonnage. Par ailleurs, la plupart des jeux collectés portaient sur des effectifs de 60 à 80 consommateurs, en raison des pratiques courantes en évaluation sensorielle.

Afin d'expérimenter cette nouvelle norme, 7 tests hédoniques ont été organisés sur un effectif supérieur à 100 sujets (150). Différentes gammes de produits ont été testées : 5 gammes comportant 5 à 6 produits chacune. Deux de ces gammes ont été évaluées par deux laboratoires, permettant ainsi de cumuler 300 réponses par gamme. Les trois autres gammes ont été testées par 150 sujets dans un laboratoire différent.
Cette étude avait pour objectif :

- de tester la stabilité des conclusions obtenues en fonction du nombre de consommateurs interrogés, en analysant les données de sous-groupes de sujets parmi l'effectif total interrogé (150 ou 300). L'objectif étant de déterminer le nombre minimal de sujets qui garantit la stabilité de la conclusion statistique pour différentes gammes de produits.
- de disposer d'une première estimation de l'hétérogénéité des préférences pour les différentes familles de produits testées.

– de tester, deux gammes de produits parmi celles du projet EpiPref[1]. Ces gammes étaient des produits support simples présentant différents gradients de gras et/ou sucre.

2.1.2 Les sujets interrogés

Les groupes interrogés de chaque laboratoire comptaient 150 sujets. Ils avaient une répartition équilibrée selon les variables sexe et âge :

– équilibre entre les tranches d'âge (18-45, 45-65 ans et +) tout en veillant à respecter une répartition homogène des âges au sein de chaque tranche,
– équilibre du genre : 50% ($\mp$10%) d'hommes et 50% ($\mp$10%) de femmes,
– habitudes de consommation : les sujets étaient des consommateurs réguliers du produit étudié.

2.1.3 Le questionnaire

Le test correspond à une épreuve de notation, basée sur une évaluation monadique séquentielle des produits.
L'objectif prioritaire du projet est d'étudier la stabilité de la réponse d'appréciation globale en fonction de l'effectif. Afin d'éviter des interférences sur cette réponse hédonique, des questions complémentaires du type "appréciation du goût sucré par rapport au niveau "juste bien"" n'ont pas été posées. Seule une question ouverte suivait la notation d'appréciation globale.
Le questionnaire comportait deux questions fermées et une question ouverte :

– Question d'appréciation globale après consommation : "Observez et goûtez le produit qui vous est présenté. Consommez-en une quantité suffisante pour vous faire une opinion : la moitié de la portion présentée. Donnez votre appréciation en cochant votre réponse dans l'échelle ci-dessous". L'échelle de notation était l'échelle hédonique de 1 à 9. En fonction de ses pratiques, chaque laboratoire avait le choix entre une échelle structurée ou non, avec bornes et références sémantiques. Les laboratoires qui réalisaient une analyse commune sur les mêmes gammes de produit se sont mis d'accord sur le choix de la même échelle.
– Question d'intention de reconsommation : "Si vous en aviez l'occasion à l'avenir, souhaiteriez-vous consommer à nouveau ce produit ?". Échelle utilisée : "Oui", "Non", "Ne sais pas".

1. Préférences et comportements alimentaires vis-à-vis du gras, du salé et du sucré (`http://www.epipref.fr/`)

– Question ouverte : présentée sur une page différente de celle de l'appréciation globale (exemple : "quels sont selon vous les qualités et défauts de ce produit ?").
Pour la gamme des harengs, une question fermée complémentaire portant sur la sensation d'arêtes en bouche a été posée sur la même page que la question ouverte.

Les dégustations ont eu lieu en laboratoire, dans des conditions standardisées conformes à la norme AFNOR (1987).

Les résultats de ce projet ont fait l'objet d'un premier article intitulé "Adequate number of consumers in a liking test. Insights from resampling in seven studies" qui constitue la section 2.1.4.
Ensuite, des analyses complémentaires ont été menées pour certaines gammes de produits et sont détaillées dans la section 2.1.5.

2.1.4 Adequate number of consumers in a liking test. Insights from resampling in seven studies : Article1

Adequate number of consumers in a liking test. Insights from resampling in seven studies.

N. Mammasse[a,*], P. Schlich[a]
[a] Centre des Sciences du Goût et de l'Alimentation CSGA, UMR6265 CNRS, UMR1324 INRA, Université de Bourgogne, Dijon, France.

Accepté, 23 Janvier 2012 - Food Quality and Preference

Adequate number of consumers in a liking test. Insights from resampling in seven studies

N. Mammasse[a,*], P. Schlich[a]

[a] *Food and Behaviour Research Center, UMR6265 CNRS, UMR1324 INRA, Université de Bourgogne, Dijon, France.*

E-mail : *nadra.mammasse@u-bourgogne.fr*

Abstract

The recommended number of consumers to be enrolled in a hedonic test comparing several products usually ranges from 50 to 100, at least if no liking segmentation is sought. This paper seeks to examine whether such a panel size range is adequate, by means of 7 trials with different levels of product space complexity. Five types of products were tested: two varied in fattiness and sweetness and were tested under the same conditions in two separate laboratories (4 trials); the remaining three, varying in taste and texture, were each tested in a different laboratory (3 trials). Each of the seven trials was run by a different laboratory. Each of the seven laboratories enrolled in its trial 150 consumers who gave liking scores on a set of 5 or 6 products. The method used to derive adequate panel size consists in removing k subjects from the 150 in the original panel and then measuring the loss of information in product comparisons. Four criteria to be maximized were used: the correlation between the two vectors of product mean scores, the RV coefficient between the two product spaces, the Fisher discrimination ratio among products and the concordance rate in product pair comparisons. For each k, k=0, ..., 130 by 10, one thousand incomplete panels were simulated by resampling with replacement. Results showed that adequate panel size varied from 20 to 150. Since the level of heterogeneity in consumer preferences was rather similar across trials, panel size should depend on product sensory differences. The variation in panel size recommendation was thus mostly due to the level of complexity of the product space.

Keywords : Hedonic test, panel size, product sensory complexity, discrimination, resampling, correlation, ANOVA, RV.

1 Introduction

The costs associated with sensory evaluation increase with the number of panelists to be enrolled. Panel size calculation would be possible if estimations of minimal product difference

and experimental error were available (Basker, 1977). Cohen (1988); Kraemer and Thiemann (1987) have published methods for calculating sample size to achieve a desired power but there are few applications of these methods in sensory analysis.

Recommendations in literature about the number of consumers in hedonic tests range from 50 to 100, at least if no liking segmentation is sought. Stone and Sidel (2004) suggested the use of 25 to 50 subjects per product in laboratory tests, with about twice this number for central location tests and home-use tests. They argued that this increase in the number of consumers is necessary to offset the expected increase in variability due to consumer inexperience and limitations in the test environment. Meilgaard et al. (1999) suggested from 50 to 300 responses for a central location test while Chambers and Wolf (1996) generally recommended 100 as an adequate panel size for most of the problems handled in small consumer tests, with the exact number depending on the experimental design. Moskowitz (1997) argued that 40 to 50 panelists would be sufficient to stabilise average acceptability, but this conclusion resulted from the analysis of only two data sets. Two authors have pointed out flaws in Moskowitz'approach. First, McEwan (1997) criticizes the simplistic view of the problem that does not take into account real considerations in product testing such as: experimental design considerations, product type and the potential number of preference segments. Second, Cornell (1997) mentions errors regarding the inferences drawn from the mean. Furthermore, statements about the stability of the mean were not supported by measurements of uncertainty of the average estimator.

Lawless and Heymann (1998) indicated the number of subjects to be enrolled in a sensory test to compare two products, but do not indicate a methodology to compare several products. Hough et al. (2006) were the first to provide the basic concepts necessary to estimate the number of subjects for sensory acceptability studies. Estimations were computed using variability estimates from previous consumer studies. The standard experimental error, reported as the root mean square of error from the two-way Anova model (panelist + product), was similar in 108 studies conducted in five countries over a wide range of food products. Considering the average standard error of 0.23 on a 0-1 scale, an alpha value of 5%, a beta value of 10% and a difference between sample means of 10% on the sensory scale, gave an N value of 112 consumers for this particular set of parameters. Gacula and Rutenbeck (2006) dealt with sample size estimation by computer simulation using the variability data from 4 sensory tests. Sample sizes ranging from 20 to 200 were used to detect a difference ranging from 0.0 to 1.0 on a 9-point hedonic scale. Results support the commonly cited sample size of 40 to 100. The paper also underlined the role

of the size of differences between products to be compared in determining an adequate panel size.

When there is a lack of information about the parameters involved in panel size calculation, sensory scientists used recommendations from standards. In such a case, the new French standard AFNOR (2009) recently recommended using at least 100 panelists for hedonic tests whereas the recommendation was only 60 in the previous standard AFNOR (2000). Increasing the panel size obviously increases experimental costs.

In this context, the ACTIA[1] sensory laboratories carried out 7 experiments with different types of products to examine whether this panel size range is adequate. Each experiment enrolled 150 panelists, i.e. more than the 100 recommended. The method used to derive adequate panel size consists in removing k (k=0, ..., 130 by 10) panelists from the 150 in the original panel and in measuring the loss of information in product comparisons.

2 Materials and methods

2.1 Procedure

Seven laboratories each enrolled a separate panel of 150 consumers to test a set of products. Five types of products were tested. The seven laboratories are located in France and have almost similar level of expertise in product testing. Trials took place from July to September 2009.

2.1.1 Panelists

Gender-balanced groups composed of 150 consumers aged 18-65 were selected for the trials. The panelists were regular consumers of the product tested.

2.1.2 Products

Five sets of products were tested in this study. A set of five cakes, varying in fattiness, was tested under the same conditions in two different laboratories. A set of five stewed apples, varying in sweetness, was tested under the same conditions in two different laboratories. Sets of six crisps, or smoked herring, or sausage, varying in taste and texture, were each tested in a different laboratory. Table 1 presents the product characteristics :

[1]ACTIA Association de Coordination Technique pour l'Industrie Agro-alimentaire

Table 1: Trial description

Product	Number of products tested	Product description	Sensory laboratory
Cake	5	fat variation	lab1 lab2
Stewed apples	5	sugar variation	lab3 lab4
Crisps	6	fat variation salt variation	lab5
Smoked herring	6	texture variation salt variation other (with and without bones)	lab6
Sausage	6	casing variation* taste variation	lab7

(*) 3 with natural casing and 3 others with artificial casing

2.1.3 Hedonic tests

Trials meet the French standard requirement. To balance presentation order, product samples were presented according to Williams Latin Square for sets of 5 products or to Mutually Orthogonal Latin Square MOLS (Wakeling and MacFie, 1995) for sets of 6 products. Consumers were required to give a liking score on a 9-point hedonic scale (Peryam and Pilgrim, 1957) varying from dislike extremely (1) to like extremely (9).

2.2 Data analysis

Although the trials used the same 9-point hedonic scale, the raw data were transformed into a $0-1$ scale in order to allow further comparison with other studies.

2.2.1 Products and population heterogeneity

A two-way analysis of variance ANOVA model with *product* as fixed effect and *panelist* as random effect was carried out on liking scores for each data set. Panelist and error sources of variation are normally distributed independent random variables with zero mean and variance $\sigma^2_{Panelist}$ and σ^2 respectively.
As the data described here had no replication, the two-way interaction is a part of the error term. This error term measures variability in consumer preferences. It is reported as the root

mean square of error $\sqrt{MS_{Err}}$.

2.2.2 Simulation

The method used to compare different panel sizes consisted in measuring the loss of information when removing k subjects from the 150 of the original panel. To measure the loss of information, one thousand incomplete panels were simulated for each k by resampling with replacement, $k = 0, \ldots, 130$ by 10. The case k=0 corresponds to sample 150 panelists from the 150 of the whole panel. Since this sampling is done with replacement, the 1000 samples with k=0 are different and actually their variability reflects the level of heterogeneity of the full panel. Furthermore, the comparison between any k and k=0 will thus be done by comparing two distributions or statistics obtained from two samples of the same size.

2.2.3 Loss of information criteria

Four criteria were defined to measure the loss of information in product comparison due to panel size reduction. Comparisons are based on correlation between product scores and the level of product discrimination.

1- *Univariate correlation criterion*
The Pearson correlation coefficient was calculated between the vector of product mean scores over the whole panel and the vector of product mean scores in the reduced panel.

2- *Multivariate correlation criterion*
The RV coefficient (Escoufier, 1973) was computed to measure the similarity between the product configuration obtained from the whole panel and the product configuration obtained from the reduced panel. It provides a simple way of measuring the relationship between two product spaces generated by two different sets of variables (panelists). The RV coefficient takes values between 0 and 1, where 1 represents perfect similarity. The RV coefficient takes into account the multidimensional specificity of the data whereas the Pearson correlation defined above measures the correlation between average product scores.
Let X be a $p \times N$ matrix (of N panelists) and Y be a $p \times (N-k)$ matrix (of (N-k) panelists) corresponding to two sets of variables defined on the same observations (p products). Both X and Y are column centered , the RV coefficient is defined by :

$$RV(X,Y) = \frac{tr(XX'YY')}{\sqrt{tr((XX')^2)tr((YY')^2)}} \quad (1)$$

If X and Y are $p \times 1$ matrices, the RV coefficient is equal to R^2, the square of the simple correlation coefficient.

3- *Discrimination criterion*

The discrimination criterion can be viewed as a version of the intraclass correlation coefficient ICC, introduced first by Fisher (1950) in genetics. Typically, it is a ratio of the variance of interest over the sum of the variance of interest and of error. Bi (2003) used this coefficient to measure the discrimination ability of a panel. Herein, an estimator of the discrimination coefficient based on the two-way ANOVA model (product + panelist):

$$\rho = \frac{MS_{Pro}}{MS_{Pro} + MS_{Err}} \tag{2}$$

where, MS_{Pro} is the product mean square.

Since ρ is between 0 and 1, it can be interpreted as the proportion of variance due to product effect. In special cases where $MS_{Err} = 0$, the F-statistic cannot be computed, while ρ reaches the maximal value of 1. This criterion has also been used to compare panel performances in descriptive sensory studies (Pineau, 2006)).

4- *Concordance rate in product pair comparison*

Multiple mean comparisons were performed with a Least Significant Difference test (LSD) at $p < 0.05$. Concordance rate is the ratio of the number of concordant pairs over the total number of pairs ($\frac{p\ (\ p\ -\ 1)}{2}$ pairs, p products) . For each pair of products, concordance is declared when the same conclusion is obtained from both panels (the whole and the reduced one). For significant pairs, concordance requires also the same product rank order. Hence, if a pair is significant in both the whole panel and the reduced one but with a change in product rank order then this change will account for a discordance.

2.2.4 Decision rules on k

For each k, one thousand values per criterion were obtained, corresponding to the one thousand simulations. A decision rule was defined to derive an adequate panel size. For each criterion, we look for the highest k for which :

✓ No more than 10% of R correlation coefficients with a value lower than 0.9. A correlation of 0.9 is considered as a fairly good correlation.

✓ No more than 10% of RV coefficients with a value lower than 0.81.

- ✓ No more than 10% of simulations with a 10% loss of discrimination (no more than 10% of simulations having $\rho_{(N-k)} < 0.9 * \rho_{(N)}$).
- ✓ No more than 10% of simulations having q or more among $\frac{p\ (\ p\ -\ 1)}{2}$ discordant pairs of products (3 among 10 for 5 products and 4 among 15 for 6 products). These limits correspond to nearly 70% of the concordance. This requires that the extreme means will still be different while being less demanding for the means between them. A higher concordance for samples goes with higher mean differences.

A loss of discrimination is defined as a decrease of ρ in the sub-panel compared to ρ in the whole panel. For the other criteria, decision is made comparing to a common absolute value such as 0.9/0.81 for R and RV, and 70% for concordance rate. Recommendations will be based on the maximum value to satisfy the requirements of all the criteria.

All statistical analyses and computations used SAS software release 9.1 (SAS institute Inc., Cary, NC).

3 Results and discussion

3.1 Anova results on liking scores

Results of the two-way Anova on the liking scores led to a significant product effect at $p < 0.05$ for all trials. Table 2 shows the main parameter values.

Table 2: Two-way Anova results of the whole panel

Data set	$\sqrt{MS_{Pro}}$	$\sqrt{MS_{Err}}$	F_{Pro}	MISD
Cake (lab1)	0.799	0.171	20.43	0.164
Cake (lab2)	0.925	0.173	28.47	0.171
Stewed apples (lab3)	1.002	0.177	32.00	0.182
Stewed apples (lab4)	0.737	0.161	20.85	0.159
Crisps (lab5)	0.737	0.188	15.21	0.185
Smoked herring (lab6)	0.575	0.198	8.38	0.190
Sausage (lab7)	2.031	0.190	113.94	0.243

$\sqrt{MS_{Pro}}$: Root product mean square
$\sqrt{MS_{Err}}$: Root mean square of error RMSE
F_{Pro}: Fisher value of product effect
MISD: Mean of individual standard deviations

The level of heterogeneity of consumer preference reported as the root mean square of error (RMSE) was similar for all trials, whereas the root product mean square measuring the magnitude of differences between product liking varied strongly, from 0.575 for herring to 2.031 for sausage.

The MS_{Pro} for stewed apples was slightly different in the two panels. Since the RMSE was almost equal, this slight difference must be due to a broader use of the scale by the panelists from lab3 than by the panelists from lab4 as shown by the mean individual standard deviation (MISD).

3.1.1 Product mean scores

Tables 3 and 4 give the mean liking scores for "Cake" and "Stewed apples" respectively, each tested under the same conditions in two different laboratories:

Table 3: Product mean scores : Cake

Cake (lab1)		Cake (lab2)	
Product designation	Mean	Product designation	Mean
A	0.702 a*	A	0.745 a
B	0.670 ab	C	0.707 a
C	0.660 bc	B	0.668 b
D	0.623 c	D	0.647 b
E	0.537 d	E	0.546 c

(*) Two means with the same letter are not significantly different (LSD, p=0.05)

Both panels preferred the same product with a small change in rank between the second and third products. Indeed, the correlation coefficient between these two vectors of product means is 0.988 (p=0.0015).

Table 4: Product mean scores : Stewed apples

Stewed apples (lab3)		Stewed apples (lab4)	
Product designation	Mean	Product designation	Mean
A	0.784 a*	A	0.768 a
B	0.717 b	C	0.713 b
C	0.699 bc	B	0.706 b
D	0.669 c	D	0.639 c
E	0.561 d	E	0.619 c

(*) Two means with the same letter are not significantly different (LSD, p=0.05)

Similarly, both panels preferred the same stewed apples (table 4) with a small change in rank between the second and third products. The lab3 panel made greater use of the scale to discriminate between products. The correlation coefficient between these two vectors of product means is 0.931 (p=0.02).

Table 5 gives the liking mean scores for both ”smoked herring” and ”sausage”. These two

products presented the most extreme differences in product mean square (MS_{Pro}) in the ANOVA model.

Table 5: Product mean scores : Smoked herring - Sausage

Smoked herring		Sausage	
Product designation	Mean	Product designation	Mean
A	0.687 a*	X	0.711 a
B	0.664 a	Y	0.704 a
C	0.612 b	Z	0.681 a
D	0.607 b	R	0.427 b
E	0.605 b	S	0.395 bc
F	0.555 c	T	0.373 c

(*) Two means with the same letter are not significantly different (LSD, p=0.05)

The differences between product mean scores were greater in the "sausage" study. Two groups of sausage were easily identified. Sausages with natural casing (X,Y,Z) had better scores than sausages with artificial casing (R,S,T). The casing characteristic was dominant and thus blurred the other differences in taste. This allowed panelists to distinguish only two main groups.

In "smoked herring", the difference in mean score between the most preferred product and the least preferred one was about 0.132, while it was on average 0.218 for the other sets of products. This could be accounted for either product complexity or preference segmentation.The latter might explain the slightly higher value of RMSE.

3.2 Panel size derivation using decision rules on k

For k=0, the average values over cakes, stewed apples and crisps of the different criteria R, RV, ρ and concordance rate were: $0.98, 0.97, 0.96$ and 89% respectively. These values were quite higher for sausages: $0.99, 0.99, 0.99$ and 92% respectively and quite lower for smoked herring: $0.95, 0.91, 0.90$ and 83% respectively. These observations comply with the Anova results about the existence of heterogeneity in smoked herring.

Table 6 gives panel size derivation according to the four criteria defined to measure the loss of information in product comparison.

Table 6: Panel size recommendation

Criterion / Data set	R	RV	ρ	Concordance rate	Recommendation*
Cake (lab1)	60	60	70	80	80
Cake (lab2)	40	40	50	70	70
Stewed apples (lab3)	40	40	50	60	60
Stewed apples (lab4)	60	60	60	70	70
Crisps	80	80	80	90	90
Smoked herring	140	140	100	150	150
Sausage	20	20	20	20	20

(*) Maximum panel size over the 4 criteria

Recommendations varied from 20 to 150. The results showed that the correlation and discrimination criteria seemed to be congruent. The concordance rate criterion was obviously the stringent one because it is based on pair concordance.

- In 6 out of 7 studies, the panel size was below the 100 level recommended by the new French standard.
- In 1 out of 7 studies, the panel size was below to 60 as required by the previous French standard.

The new French standard which recommends 100 panelists in most cases, seems to overestimate the number of consumers required for a hedonic test. The main information on product comparison obtained with 150 panelists was captured with fewer than 100 in six of the trials. Table 7 summarizes the trials description and the derived recommendations:

Table 7: Product complexity and panel size relationship

Product	Product variation	$\sqrt{MS_{Pro}}$	$\sqrt{MSE}$	Panel size recommendation
Cake (lab1)	fat variation	0.799	0.171	80
Cake (lab2)	fat variation	0.925	0.173	70
Stewed apples (lab3)	sugar variation	1.002	0.177	60
Stewed apples (lab4)	sugar variation	0.737	0.161	70
Crisps (lab5)	fat variation salt variation	0.737	0.188	90
Smoked herring (lab6)	texture variation salt variation other	0.575	0.198	150
Sausage (lab7)	casing variation taste variation	2.031	0.190	20

As reported in this table, panel size recommendation mainly depends on the size of the liking differences between products to be compared. We expect that this difference is directly affected by the sensory complexity of the products. In this study, smoked herring was the most complex set of products from a sensory point of view and thus required much more than 100 panelists. On the contrary, the sausage set was very easy to discriminate because the type of casing was a dominant characteristic. Consumers distinguished between two main groups, with a difference in mean score of 0.398 between the most preferred product and the least preferred one and thus this resulted in a higher product mean score MS_{Pro}. The sensory attributes driving consumers liking are interlocked to the variation among the set of products to be compared. Hence, sugar content was the driver of liking for the stewed apples, whereas fat content was the driver of liking for cakes. Sausage liking was mostly directed by one obvious product characteristic (natural versus artificial casing). Given this high level of differences, a panel of only 20 panelists gave the same conclusions as the entire panel. Here, we found two examples of studies that did not match the panel size range (50 to 100) recommended by the standards.

Besides, the level of heterogeneity in consumer preferences measured by the RMSE was almost similar over the trials. This result comply with those obtained by Hough et al. (2006). Authors stated that the RMSE was similar across several countries over a wide range of products.

We believe that the product sensory variations are the determinant factor in the size of differences in acceptance to be sought. When the products differ in more than one sensory dimension, like for smoked herring, differences in product acceptability are induced by the differences in taste and texture due to salt contents. Hence, this variation might produce closest acceptability scores due to a difficulty to perceive differences or might generate preference segments that affect the mean liking scores by reducing the product mean square.

4 Conclusion

In this study, seven trials were conducted in seven sensory laboratories to test by experiment the impact of panel size on product comparisons. The results led to two main conclusions. First, the level of heterogeneity in consumer preferences was almost similar over the trials (in agreement with Hough et al. (2006)). Second, the number of panelists for a hedonic test depends mainly on the level of complexity of the product space and the magnitude of liking differences between products to be compared. It is not possible to define a global number of consumers valid for every study, as it was suggested in the literature and by the standards.

We demonstrated by resampling in a large (n=150) panel that the appropriate number of consumers could have been as low as 20 or as high as 150 depending on the studies. This suggests that the standard power computation is too simplistic. This is due to the fact that it simply looks for how many subjects are necessary to get a significant statistical test if the true difference is equal to a given size. When we have several products to compare, the differences may be very different over the pair of products. Furthermore, these differences may generate a multidimensional space meaning that we do have segmentation of consumer preferences. We argue that the origin of this segmentation is to be searched in the level of sensory complexity of the product space. The more the sensory complexity you can expect to differentiate the products, the more likely preference segmentation holds.

The sensory analyst should have an idea of the size of differences between the products to be compared. The recommendation would be initially to estimate the number of sensory dimensions discriminating the products, for instance by means of a quick sensory descriptive task.

Future research should investigate with much more than seven studies in order to establish what should be the adequate number of consumers depending on the number of sensory dimensions and the level of products difference in liking.

5 Acknowledgements

This work was carried out with the financial support of the ACTIA "Association de Coordination Technique pour l'Industrie Agro-alimentaire", with the support of the Burgundy County Council and coordinated by Sensorialis, the French network of sensory laboratories which ran the experiments.

References

AFNOR (2000). Sensory analysis - methodology - general guidance for conducting hedonic tests with consumers in a controlled area. norme nf v09-500.

AFNOR (2009). Sensory analysis - methodology - general guidance for conducting hedonic tests with consumers in a controlled area. norme xp v09-500.

Basker, D. (1977). The number of assessors required for taste panels. *Chemical Senses*, 2(4):493–496.

Bi, H. (2003). Agreement and reliability assessments for performance of sensory descriptive panel. *Journal of Sensory Studies*, 18(1):61–76.

Chambers, E. and Wolf, M. B. (1996). Sensory testing methods. pages 9 –10. American society for testing and materials, second edition.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. LEA, New York, second edition.

Cornell, J. A. (1997). What is the meaning of stability in the mean? *Food Quality and Preference*, 8(4):259–260.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, 29(4):751–760.

Fisher, R. A. S. (1950). *Statistical methods for research workers / by R.A. Fisher*. Biological monographs and manuals ; no. 5. Oliver & Boyd, Edinburgh.

Gacula, M. and Rutenbeck, S. (2006). Sample size in consumer test and descriptive analysis. *Journal of Sensory Studies*, 21(2):129–145.

Hough, G., Wakeling, I., Mucci, A., Chambers, E., Gallardo, I. M., and Alves, L. R. (2006). Number of consumers necessary for sensory acceptability tests. *Food Quality and Preference*, 17(6):522–526.

Kraemer, H. and Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. SAGE publications.

Lawless, H. T. and Heymann, H. (1998). In *Sensory evaluation of food. Principles and practices*, pages 754–782. Chapman and Hall, New York.

McEwan, J. A. (1997). Base size in product testing: A psychophysical viewpoint and analysis - reply. *Food Quality and Preference*, 8(4):257–258.

Meilgaard, M., Carr, T., and Civille, G. V. (1999). Affective tests: Consumer tests and in-house panel acceptance tests. In *Sensory Evaluation Techniques*, pages 231–263. CRC Press, third edition.

Moskowitz, H. R. (1997). Base size in product testing: A psychophysical viewpoint and analysis. *Food Quality and Preference*, 8(4):247–255.

Peryam, D. and Pilgrim, F. (1957). Hedonic scale method of measuring food preference. *Food technology*, 11(9):9–14.

Pineau, N. (2006). *Performance on sensory studies. A data base approach*. PhD thesis.

Stone, H. and Sidel, J. L. (2004). Affective testing. In technology, F. s. a., editor, *Sensory Evaluation Practices*, pages 247–277. Elsevier Academic Press, USA, third edition.

Wakeling, I. N. and MacFie, H. J. H. (1995). Designing consumer trials balanced for first and higher orders of carry-over effect when only a subset of k samples from t may be tested. *Food Quality and Preference*, 6(4):299–308.

2.1.5 Analyses complémentaires

Les principaux résultats de cette première étude montrent que le paramètre qui semble le plus conditionner le nombre de sujets est la taille des différences d'appréciations entre les produits testés. L'hétérogénéité des préférences semble similaire entre les différentes études.
Nous avons ensuite mené des analyses statistiques complémentaires sur :

1. les deux gammes de produits "cake" et "compote" qui ont été testés dans les mêmes conditions par deux laboratoires différents, soit un effectif de 300 sujets au total pour chaque gamme,
2. les deux gammes de produits "knack" et "hareng", pour lesquels nous avions obtenus les deux recommandations extrêmes (20 et 150 sujets).

2.1.5.1 Analyse complémentaire des données "cake" et "compote"

Nous reconsidérons à présent les tests hédoniques sur les produits "cake" et "compote" comme deux jeux de données de 300 sujets en rassemblant les deux panels ayant testé les mêmes produits et on applique l'approche par rééchantillonnage. L'objectif est de voir si en partant d'un effectif plus grand, on aboutirait aux mêmes recommandations sur le nombre de sujets n. La table (3.1) résume les résultats du modèle ANOVA à deux facteurs (sujet et produit) :

TAB. 3.1 – *Tableau Anova - Cake et compote*

Data set	$\sqrt{MS_{Pro}}$	$\sqrt{MS_{Err}}$	F_{Pro}	MISD
Cake	1.191	0.172	47.76	0.171
Stewed apples	1.208	0.170	50.37	0.171

$\sqrt{MS_{Pro}}$: Racine carrée du carré moyen produit
$\sqrt{MS_{Err}}$: Racine carrée du carré moyen de l'erreur RMSE
F_{Pro} : Statistique de Fisher de l'effet produit
MISD : Moyenne des écart types individuels

Il est intéressant de voir que le terme d'erreur RMSE est équivalent à celui observé dans le cas des panels à 150 sujets. Étant donné que ce terme exprime l'hétérogénéité des préférences, cela signifie que cette hétérogénéité était déjà restituée en ne prenant que 150 consommateurs.
Le carré moyen produit (respectivement la statistique de Fisher de l'effet produit) observé sur les jeux de données avec un effectif de 300 sujets correspond, à un facteur de 2 près au carré moyen produit (respectivement la statistique de Fisher de l'effet produit) observé sur les jeux de données avec 150 sujets. Les scores moyens d'appréciation pour chaque produit ainsi que leur écart par rapport à la moyenne globale d'appréciation (variance produit) sont restés stables.
Le tableau (3.2) donne les recommandations selon les 4 critères de décision prédéfinis dans

notre approche soit : la corrélation unidimensionnelle R, la correlation multidimensionnelle RV, le coefficient de discrimination ρ et le taux d'accord.

TAB. 3.2 – *Recommandations sur le nombre de sujets*

Critère / Jeu de données	R	RV	ρ	Taux d'accord	Recommandation*
Cake	50	50	70	80	80
Compote	50	40	70	70	70

(*) Recommandation maximale selon les 4 critères

On peut constater que les recommandations sur le nombre de sujets correspondent avec celles obtenues en considérant un effectif de 150 sujets.

2.1.5.2 Analyse complémentaire des données "knack" et "hareng"

2.1.5.2.1 Étude "knack"

Les 150 sujets recrutés pour ce test hédonique étaient des consommateurs de knack des deux types de boyaux répartis comme suit : 16 consommateurs de boyaux artificiels (groupe A), 83 consommateurs de boyaux naturels (groupe N) et 51 consommateurs des deux types de boyaux (groupe AN). Nous avons voulu savoir s'il y avait un effet de l'habitude de consommation sur l'appréciation globale des produits. Pour ce faire, nous avons appliqué aux données le modèle ANOVA hiérarchique suivant :

appréciation globale groupe + sujet (groupe) + produit + produit*groupe + erreur

Le facteur groupe correspond au type de boyau consommé. Les résultats de l'application de ce modèle ont montré que l'effet groupe n'était pas significatif (p 0.10). Cependant, il existe un léger effet d'interaction groupe*produit. Celui-ci est dû à une différence de notation du groupe des consommateurs de boyaux artificiels. Ce groupe discrimine mieux les boyaux artificiels et les notent plus haut que les autres groupes. Toutefois, l'effectif de ce groupe était moins important comparé aux autres groupes. Par ailleurs, les knack à boyaux naturels étaient préférés dans les 3 groupes de consommateurs. La figure (3.2) donne les moyennes des notes d'appréciation par produit dans chaque groupe de consommateurs.

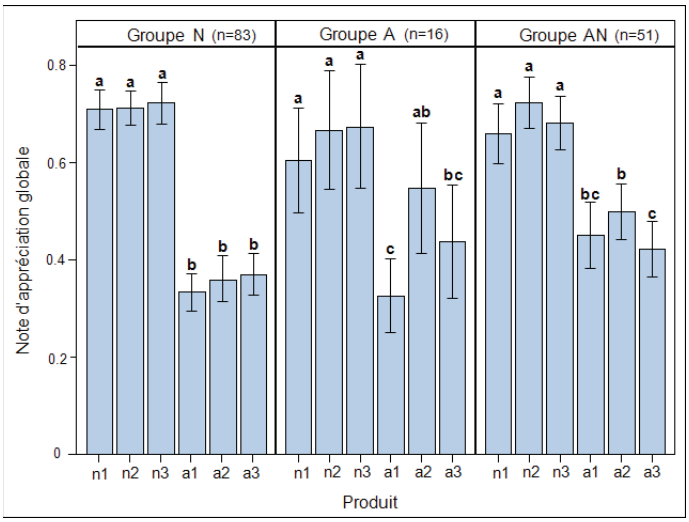


FIG. 3.2 – *Moyennes des notes d'appréciation par groupe de consommateurs*

On retrouve pour les trois groupes de consommateurs des préférences similaires à celles obtenues sur le panel complet. Il n'existe donc pas d'effet de l'habitude de consommation sur la préférence globale.

2.1.5.2.2 Étude "hareng"

Tout d'abord, nous avons voulu savoir si un effet de la fréquence de consommation du hareng existait. Les sujets étaient répartis en trois groupes selon leurs fréquences de consommation du produit : 48 en consommaient au moins une fois par mois, 52 sujets en consommaient 6 à 12 fois par an et 50 moins de 6 fois par an.
Nous avons appliqué un modèle d'ANOVA hiérarchique :

appréciation globale fréquence + sujet (fréquence) + produit + produit*fréquence + erreur
Les résultats ne montrent pas d'effet significatif de la fréquence de consommation (p 0.33) ni de l'interaction produit*fréquence (p 0.85).

Comme nous avions observé un faible effet produit (F 8.32) pour cette étude par rapport aux autres études, nous avons ensuite voulu vérifier l'hypothèse de l'éventuelle existence d'une segmentation des préférences.
La figure (3.3) montre les boîtes à moustaches des notes d'appréciation globale pour les 6 harengs.

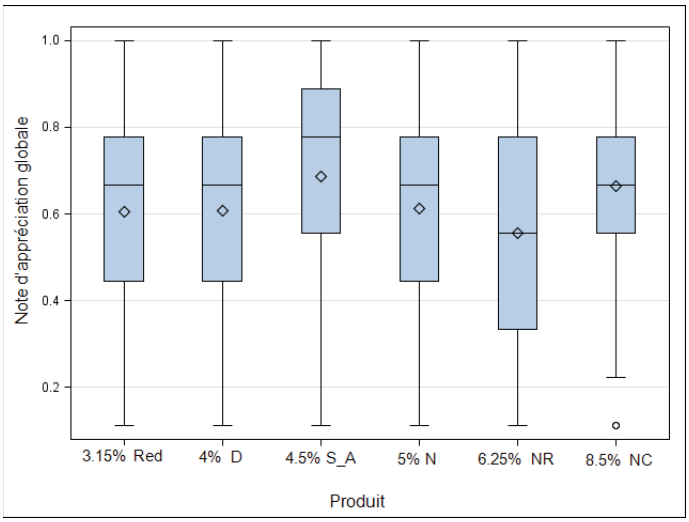


FIG. 3.3 – *Boîtes à moustache des notes d'appréciation des 6 harengs*

On peut remarquer à partir de ce graphique que le produit avec un taux de sel de 6.25% présente la plus grande variabilité. Cette variabilité est peut être due au désaccord entre les sujets et suggérerait l'existence de deux groupes de préférences pour ce produit.
Ainsi, nous avons appliqué une classification hiérarchique sur les données (procédure VARCLUS de SAS©, critère du centroïd) muni du critère du pseudoF correspondant à l'interaction produit*segment dans le modèle ANOVA hiérarchique à trois facteur segment, sujet et produit.
Les résultats suggèrent une répartition en 4 groupes de préférences.
La figure 3.4 représente les moyennes des notes d'appréciation globale par groupe de préférence.

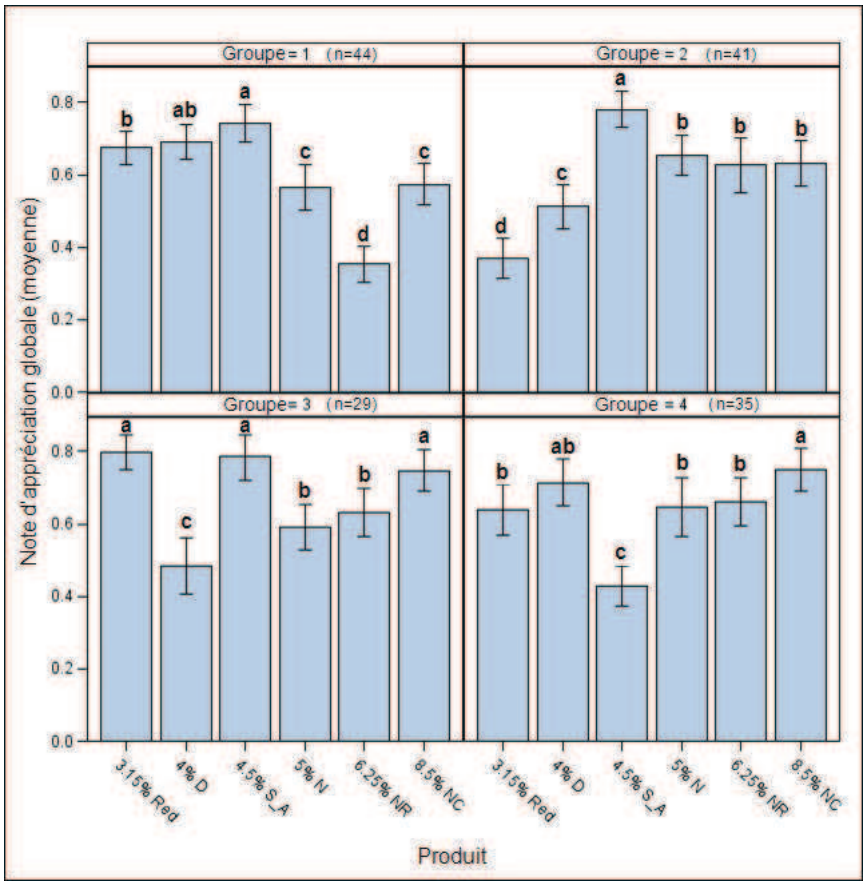


FIG. 3.4 – *Moyennes des notes d'appréciation par segment de préférence*

Nous rappelons que dans cette gamme de produits, nous avions une variation produit sur trois dimensions sensorielles : la teneur en sel, le type de fumage (naturel ou doux) mais aussi hareng avec ou sans arêtes. Le produit cible testé ici est le "hareng sans arête" (4.5% S_A), qui avait subi un procédé qui permet de réduire la sensation d'arête en bouche. Au vu des résultats de cette séparation, ce produit a été préféré dans les trois groupes 1,2 et 3. Le groupe 4 semble préférer les autres produits qui sont plus communs sur le marché. Le groupe 1 a une préférence pour les harengs plutôt moins salés lorsque le groupe 2 semble préférer le harengs plutôt salés. Le groupe 3 préfèrent moins le type de fumage doux.

On peut conclure au vu de ces résultats, que si l'on s'intéresse particulièrement à la préférence pour le produit cible, c'est-à-dire le produit sans arêtes, celui-ci semble être préféré dans 3 groupes et reflète le résultat global obtenu avec le panel complet. Par contre, si on s'intéresse à détecter les préférences pour les autres produits, l'existence de ces segments de préférences risque de masquer les différences d'appréciation pour ces produits.

Par ailleurs, ce jeu de données a été le plus demandant en terme de nombre de sujets par notre approche rééchantillonnage. Ce résultat peut être expliqué par la variabilité des préférences de ces

produits qui ainsi changent au fur et à mesure que l'on réduit le nombre de sujets et on n'arrivait pas à reconstituer l'image exacte de la séparation deux à deux des produits obtenu avec le panel complet. Le produit cible demeure souvent le produit préféré, par contre les préférences pour les autres produits intermédiaires varient, ce qui rend difficile l'obtention d'un même ordre de préférence dans les sous-panels.

2.1.5.3 Conclusion

Au vu de ces résultats complémentaires on peut conclure que lorsque les différences en termes d'appréciation sont évidentes, il ne serait pas utile de prendre un large effectif de sujets. Les résultats obtenus sur le jeu de données "Knack" conforte cette idée.
Par ailleurs, l'existence potentielle de segments de préférences risque de compromettre la mise en évidence d'un effet produit. D'autre part , la complexité sensorielle pourrait être à l'origine de l'existence de ces segments de préférences. Nous pouvons admettre qu'un produit "simple" peut être segmentant. Cependant, un produit complexe est a priori encore plus segmentant (plus de classes à priori car chaque dimension sensorielle pourrait induire une segmentation). Si l'objectif de l'étude est de détecter les segments de préférences, le nombre de sujet doit être assez conséquent. Il est clair que personne ne peut donner le nombre de segments à priori mais ce qu'il peut être possible d'avancer comme information, c'est quel est le segment du marché visé et quelle importance quantitative (proportion) représente ce segment. Au besoin, il faudrait peut être préconiser une étude préliminaire et prospective.

2.2 Synthèse

L'approche par rééchantillonnage réalisé dans le cadre de ce projet a mis en exergue une hypothèse principale. Le facteur qui semble le plus conditionner la taille de panel n'est pas la variabilité individuelle (hétérogénéité des préférences) qui exprime le désaccord entre les sujets, mais principalement la taille des différences d'appréciation entre les produits testés. La différence d'acceptabilité entre les produits pourrait être liée à la complexité du produit. Plus le nombre de dimensions sur lesquelles reposent les différences organoleptiques entre produits est important et plus on s'attend à ce que le nombre de sujets à interroger soit grand.

3 Approche théorique rétrospective : Méta-analyse de PrefBase

La méta-analyse est une démarche, plus qu'une simple technique, qui a pour but de combiner les résultats de plusieurs études, pour en faire une synthèse quantifiée. Initialement introduite en médecine par Mantel and Haenszel (1959), le mot "méta-analyse" a été utilisé pour la première fois en sciences sociales par Glass (1976) . Cette démarche s'est très rapidement répandue à d'autres secteurs de recherche, notamment en sciences animales (Marchant and McGrew, 1991) et aux sciences de l'éducation (Hedges, 1986).
Dans le cas des essais cliniques, il s'agit généralement de tester une hypothèse, par exemple l'effet d'un traitement pharmaceutique à partir des résultats de plusieurs expérimentations destinées à l'évaluer. La méta-analyse produit un gain de puissance statistique dans la recherche de l'effet d'un traitement, une précision optimale dans l'estimation de la taille de l'effet.
La méta-analyse se fonde sur le cumul de chacune des variables résumant les données initiales de chaque étude. Ces variables sont caractérisées par leurs moyennes et leurs écart-types, etc. De plus, il est possible d'étudier ces variables en fonction des caractéristiques de la recherche (selon le genre, le contexte, ...) et ainsi d'isoler les situations où le l'effet étudié est significatif ou pas.
Une méta-analyse nécessite avant tout de préciser son objectif, pour ensuite définir les étapes suivantes et en particulier le codage, le filtrage, la pondération des données et le modèle statistique. On peut avoir des objectifs plus au moins ciblés allant d'une simple étude exploratoire à celle d'un effet particulier en fonction d'autres variables. Une fois l'objectif défini, il est nécessaire de définir les 4 étapes suivantes :

- Le codage des données afin d'unifier la lecture des informations entre les différentes études. Il est important d'inclure toutes les informations disponibles qui sont pertinentes vis à vis de l'objectif de la méta-analyse.
- Le filtrage des données qui conditionne la qualité des conclusions tirées d'une méta-analyse. Il ne faudrait conserver que les études qui répondent aux objectifs.
- La pondération des données qui permet d'accorder des poids pour chaque étude en fonction de son adéquation avec la problématique ou de la qualité des données (représentativité des données, ...)
- Le choix du modèle statistique, on retrouve le plus souvent le modèle linéaire de Hedges et al. (1992). Le modèle ainsi que ces paramètres dépendent de la nature des données et de la problématique.
- L'étude graphique des données qui représente une phase essentielle du déroulement d'une méta-analyse. Les représentations graphiques permettent d'avoir une idée sur le degré d'hétérogénéité et de cohérence des données, sur la nature et l'importance des relations inter-études ou intra-études.

La méta-analyse est très peu répandue en analyse sensorielle. Il existe quelques études mettant en commun les résultats de plusieurs laboratoires (Martin et al., 2000; McEwan et al., 2002; Pages and Husson, 2001). Leur objectif est de comparer des panels qui ont réalisé des études sur un même type de produits, dans des conditions ou des lieux différents. Ce type d'études permet d'évaluer l'impact des conditions d'entraînement sur le niveau de performance des panels. Le travail de Pineau (2006) a constitué à calculer des indices sur les niveaux de performances moyens grâce à une méta-analyse sur les jeux de données issus de SensoBase.

Le travail présenté dans cette partie applique les principes généraux de la méta-analyse sur les jeux de données de préférences issus de la base de données PrefBase. L'objectif est d'estimer les paramètres qui conditionnent le nombre de sujets en tests hédoniques, de définir leurs domaines de variation et de pouvoir adresser des recommandations sur le nombre de sujets à enrôler pour ce type d'épreuves.

3.1 Le modèle d'analyse et quantification des paramètres

Le modèle d'analyse retenu pour estimer la grandeur de l'effet et l'hétérogénéité des préférences en tests hédoniques est le modèle de l'analyse de la variance à deux facteurs Sujet + Produit. L'unité expérimentale est le jeu de donnée. Ainsi, pour rendre les grandeurs de l'effet et de l'hétérogénéité comparables entre jeux de données, les notes hédoniques sont ramenées à l'échelle 0-1 en appliquant une transformation affine. Pour chaque jeu de données, nous avons estimé :

- la variabilité σ par la racine carrée du carré moyen de l'erreur notée RMSE (Root Mean square of error) du modèle ANOVA
- la taille de l'effet d par la racine carrée de la variance des moyennes des produits notée σ_{prod}, $\sigma_{prod} = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(x_{i.} - x_{..})^2}$

Une fois ces paramètres quantifiés, nous avons établi des recommandations sur le nombre de sujets en fonction de trois valeurs de RMSE et σ_{prod} obtenues à partir de leur distribution respective sur l'ensemble des 184 jeux de données de PrefBase. Ces valeurs correspondant aux premier quartile, à la médiane et au troisième quartile de ces grandeurs. Pour chaque combinaison de RMSE et σ_{prod}, le nombre de sujets est calculé pour les valeurs 5%, 10% et 20% des risques α et β.

Nous avons pris pour paramètre de non centralité la mesure $\lambda = n\, k\, \frac{d^2}{\sigma^2}$, tels que d et σ sont quantifiés par σ_{prod} et RMSE respectivement. Le calcul du nombre de sujets est ainsi obtenu par processus itératif. On part d'une valeur initiale de n, puis on incrémente n jusqu'à atteindre la

puissance souhaitée. La recommandation correspond à la valeur n qui permet d'atteindre cette puissance.

3.2 Résultats et discussion

3.2.1 Distribution des notes d'appréciation

Nous avons observé les moyennes des notes d'appréciation pour les différents produits de l'ensemble des jeux des données (figure 3.5). Ces moyennes des notes d'appréciation varient de 0.1 à 0.8 sur une échelle 0-1. La moyenne globale des notes de préférences tous produits confondus est de 0.57, soit 5.7 sur une échelle 0-10.

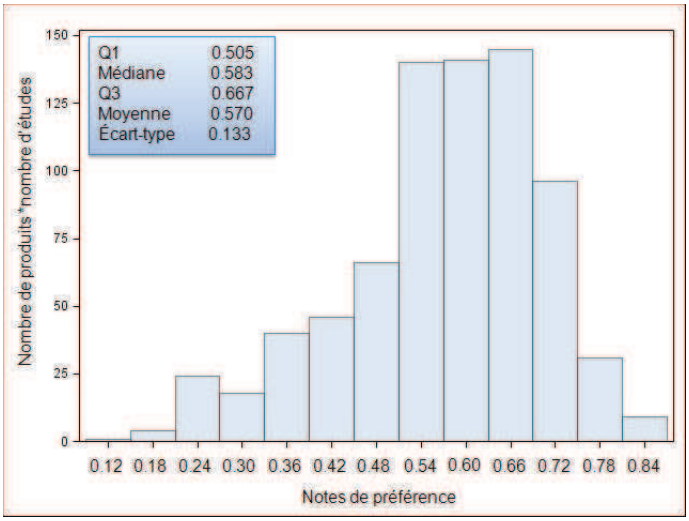


FIG. 3.5 – *Distribution des notes d'appréciation*

Ces notes de préférences peuvent aussi être étudiées en fonction des grandes familles de produits. Cependant, le nombre de jeux de données est trop faible pour comparer ces moyennes des notes d'appréciation au sein et/ou entre les familles de produits. Comme on peut le constater dans la figure (3.6), la famille de produits "Poissons" obtient la plus haute moyenne des notes d'appréciation. Or, on ne peut conclure que les poissons sont les produits les plus préférés car cette observation n'est faite que sur un seul jeu de données comparant deux produits de cette famille. C'est pourquoi, il faudrait alimenter la base de données davantage permettant d'avoir assez de jeux de données par familles de produits et ainsi comparer rationnellement les préférences entre ces familles.

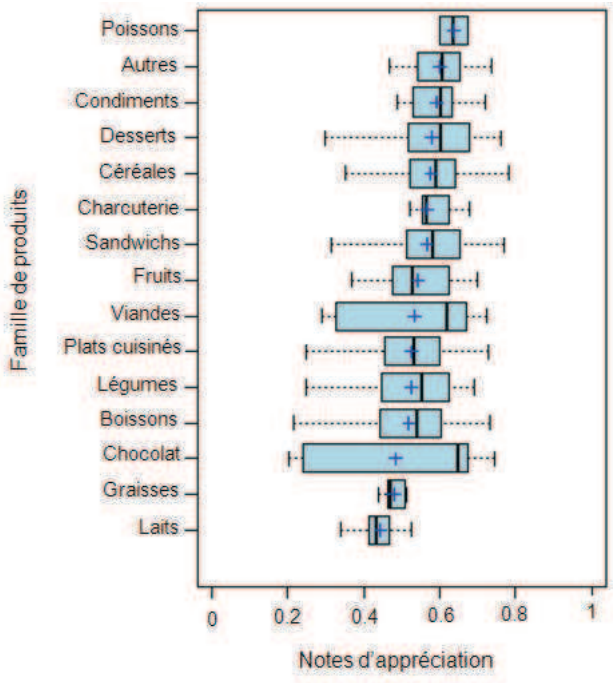


FIG. 3.6 – *Boîte à moustaches des notes d'appréciation par famille de produits*

3.2.2 La taille de l'effet

Tout d'abord, nous avons observé l'étendue des différences (D) des notes d'appréciation entre le produit le plus préféré et le produit le moins préféré pour l'ensemble des jeux de données (figure 3.7). Nous pouvons remarqué que ces différences varient de 0.03 à 0.54 sur l'échelle 0-1. Pour un jeu données, il existe en moyenne un écart d'appréciation de 0.15 entre les produits extrêmes.

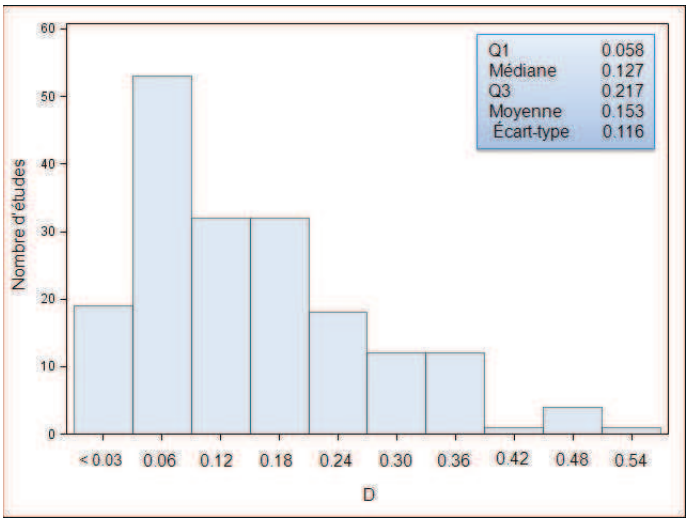


FIG. 3.7 – *Distribution des différences entre les notes d'appréciation du produit le plus préféré et le produit le moins préféré*

Ensuite, nous avons étudié pour chaque jeu de données, l'écart moyen des notes d'appréciation par rapport à la moyenne globale (σ_{prod}). C'est cette mesure qui estime la taille de l'effet "Produit" utilisé pour le calcul du nombre de sujets. La figure (3.8) montre la distribution des σ_{prod} sur l'ensemble des jeux de données.

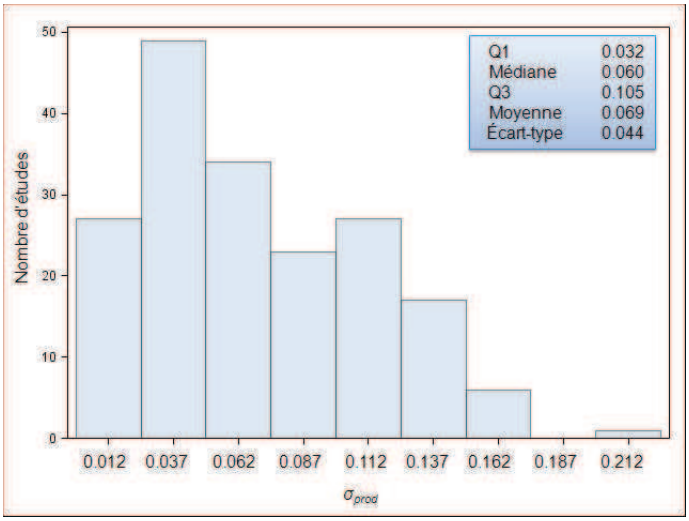


FIG. 3.8 – *Distribution de la taille de l'effet en test hédonique*

Nous observons en moyenne un écart σ_{prod} égal à 0.069 sur une échelle 0-1 entre les produits. Cette valeur correspond à un écart moyen de 0.69 sur une échelle 0-10.
Par ailleurs, 25% des études ont un écart σ_{prod} inférieur à 0.032, lorsque 75% ont un écart σ_{prod} inférieur à 0.105. Ces valeurs du premier et du troisième quartile sont utilisées comme borne

inférieure et borne supérieure des estimations de la taille de l'effet pour établir les abaques du nombre de sujets à enrôler en tests hédoniques.
Notons que les jeux de données qui ont une valeur du σ_{prod} inférieur à 0.032, soit un σ_{prod} moyen de 0.025, sont les jeux de données pour lesquels l'effet "Produit" était non significatif à $p = 0.05$ (soit pour 66 études parmi les 184 jeux de données). Par ailleurs, nous avons observé pour ces même jeux de données, un écart d'appréciation en moyenne entre les produits extrêmes de 0.045. Ces résultats suggère que lorsque l'on compare une gamme de produits avec un écart moyen σ_{prod} attendu inférieur à 0.03, ou une différence entre les moyennes des notes produits extrêmes attendue de 0.04, l'effet produit serait non significatif.

Rappelons que ces études pour lesquelles l'effet produit était non significatif peuvent correspondre à la réalité, c'est-à-dire, il n'y a pas de différence d'appréciation entre les produits testés sinon ces tests hédoniques ont peut être manqué de puissance.

Enfin, nous avons calculé les tailles d'effet standardisées selon les suggestions de Cohen (1988). Nous avons calculé la mesure $f = \frac{\sigma_m}{\sigma}$, en remplaçant σ_m et σ par les mesures σ_{prod} et RMSE respectivement. La figure (3.9) montre la distribution des tailles d'effet standardisé observées sur l'ensemble des jeux de données de PrefBase.

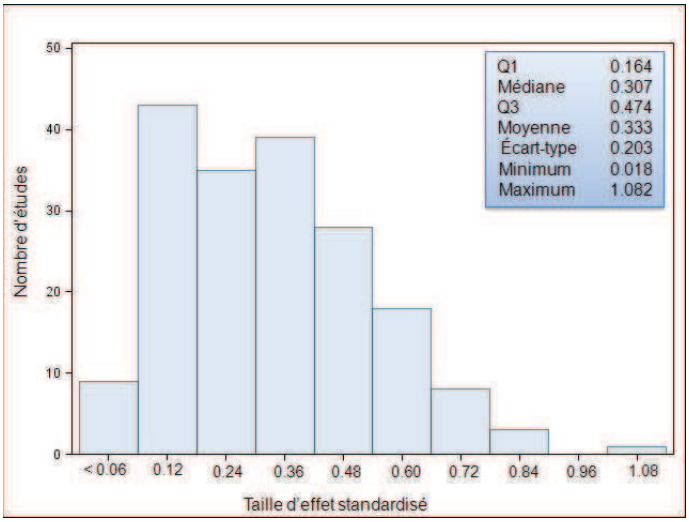


FIG. 3.9 – *Distribution des tailles d'effet standardisé en test hédonique*

En prenant les trois mesures du premier quartile, de la médiane et du troisième quartile, nous avons défini trois catégories de tailles d'effet : faible, moyen, fort. Ces mesures ont été comparées aux autres mesures conventionnelles de tailles d'effet établies en sciences du comportement par Cohen (1988). Même si les variables mesurées dans les deux domaines sont différentes, cette comparaison permet de situer les effets observés en analyse sensorielle par rapport à d'autres domaines. Ces mesures sont rapportées dans le tableau (3.3).

TAB. 3.3 – *Tailles d'effet standardisé*

Source	Tailles d'effet standardisé		
	faible	moyen	fort
Science du comportement	0.10	0.25	0.40
Méta-analyse PrefBase	0.16	0.30	0.47

Comme on peut le constater les tailles d'effet calculées à partir de PrefBase sont légèrement supérieures aux valeurs conventionnelles établies par Cohen en sciences du comportement. Pour une nouvelle étude hédonique, ces valeurs permettront d'avoir une idée sur l'effet recherché et donc du nombre de sujets à enrôler. Ainsi, une taille d'effet autour de 0.16 peut être considérée comme un effet réel mais difficile à percevoir. Un effet moyen de 0.30 est considéré comme un effet visible. Un grand effet de 0.47 serait un effet facile à mettre en évidence.

3.2.3 L'hétérogénéité des préférences

L'autre paramètre qui conditionne le nombre de sujets est l'hétérogénéité des préférences qui exprime le désaccord entre les sujets. La figure (3.10) montre la distribution des RMSE, estimateur de cette variabilité individuelle issue du modèle ANOVA appliqué à chacun des 184 jeux de données.

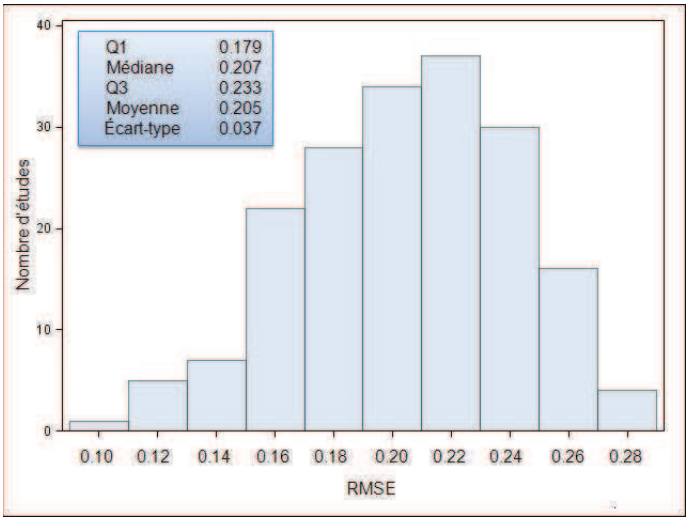


FIG. 3.10 – *Distribution de l'hétérogénéité des préférences RMSE*

Nous pouvons constater que l'hétérogénéité des préférences est en moyenne de 0.205 sur une échelle de 0-1. Cette valeur correspond à l'estimation obtenue par Hough et al. (2006) ($\sigma_{erreur} = 0.23$ sur 108 études). Le premier quartile est égal à 0.179 et le troisième quartile est

égal à 0.233. Nous avons utilisé ces trois valeurs pour estimer le nombre de sujets.
Par ailleurs, nous avons remarqué que les variations du RMSE sont identiques pour les jeux de données dans lesquels l'effet produit était significatif ou non .

3.2.4 Recommandations sur le nombre de sujets en tests hédoniques

Pour établir les recommandations sur le nombre de sujets nous avons utilisé les estimations de d et σ obtenues précédemment. Ce calcul a été fait en considérant une étude de k 4 produits qui correspond au nombre moyen de produits testés dans une étude hédonique (valeur obtenue de la méta-analyse). Le tableau (3.4) résume les différentes recommandations sur le nombre de sujets obtenues pour les différentes valeurs de α, β, d et σ.

TAB. 3.4 – *Abaques du nombre de sujets à enrôler en tests hédoniques pour k 4 produits*

σ	d	α	β		
			20%	10%	5%
0.17	0.03	20%	54	75	95
		10%	72	96	119
		5%	89	116	140
	0.06	20%	14	20	25
		10%	19	25	31
		5%	24	30	36
	0.10	20%	6	8	10
		10%	8	10	12
		5%	10	12	14
0.20	0.03	20%	74	104	132
		10%	99	133	163
		5%	123	159	193
	0.06	20%	19	27	34
		10%	26	34	42
		5%	32	41	50
	0.10	20%	8	11	13
		10%	10	13	16
		5%	13	16	19
0.23	0.03	20%	97	137	174
		10%	131	175	216
		5%	162	210	254
	0.06	20%	25	35	44
		10%	34	45	55
		5%	42	54	65
	0.10	20%	10	13	17
		10%	13	17	21
		5%	16	21	25

Comme on peut le constater, le paramètre qui semble le plus influer le nombre minimum de sujets à enrôler dans un panel de tests hédoniques est la taille des différences à mettre en évidence. Par exemple, pour une variabilité $\sigma = 0.20$, un risque $\alpha = 5\%$ et un risque $\beta = 10\%$, le nombre de sujets diminue de 74% lorsqu'on passe d'un écart moyen d'appréciation d à mettre en évidence de 0.03 à 0.06 sur une échelle 0-1.

Par ailleurs, pour un écart moyen d'appréciation $d = 0.06$, un risque $\alpha = 5\%$ et un risque $\beta = 10\%$, le nombre de sujets diminue de 24% seulement lorsqu'on considère une variabilité σ de 0.20 au lieu de 0.23.

Si on considère une hétérogénéité moyenne de 0.20, un risque $\alpha = 5\%$, une puissance de test à 90%, il faudrait 159 sujets pour mettre en évidence un petit effet, 41 sujets pour mettre en évidence un effet moyen et 16 sujets seulement pour mettre en évidence un grand effet !

La figure (3.11) montre les variations de la puissance en fonction du nombre de sujets pour les différentes valeurs de d, σ et deux valeurs du risque α (5% et 10%).

Le nombre de sujets est plus important lorsque la variabilité est maximale ($\sigma = 0.23$) et l'écart moyen des moyennes des produits est minimal ($d = 0.03$).

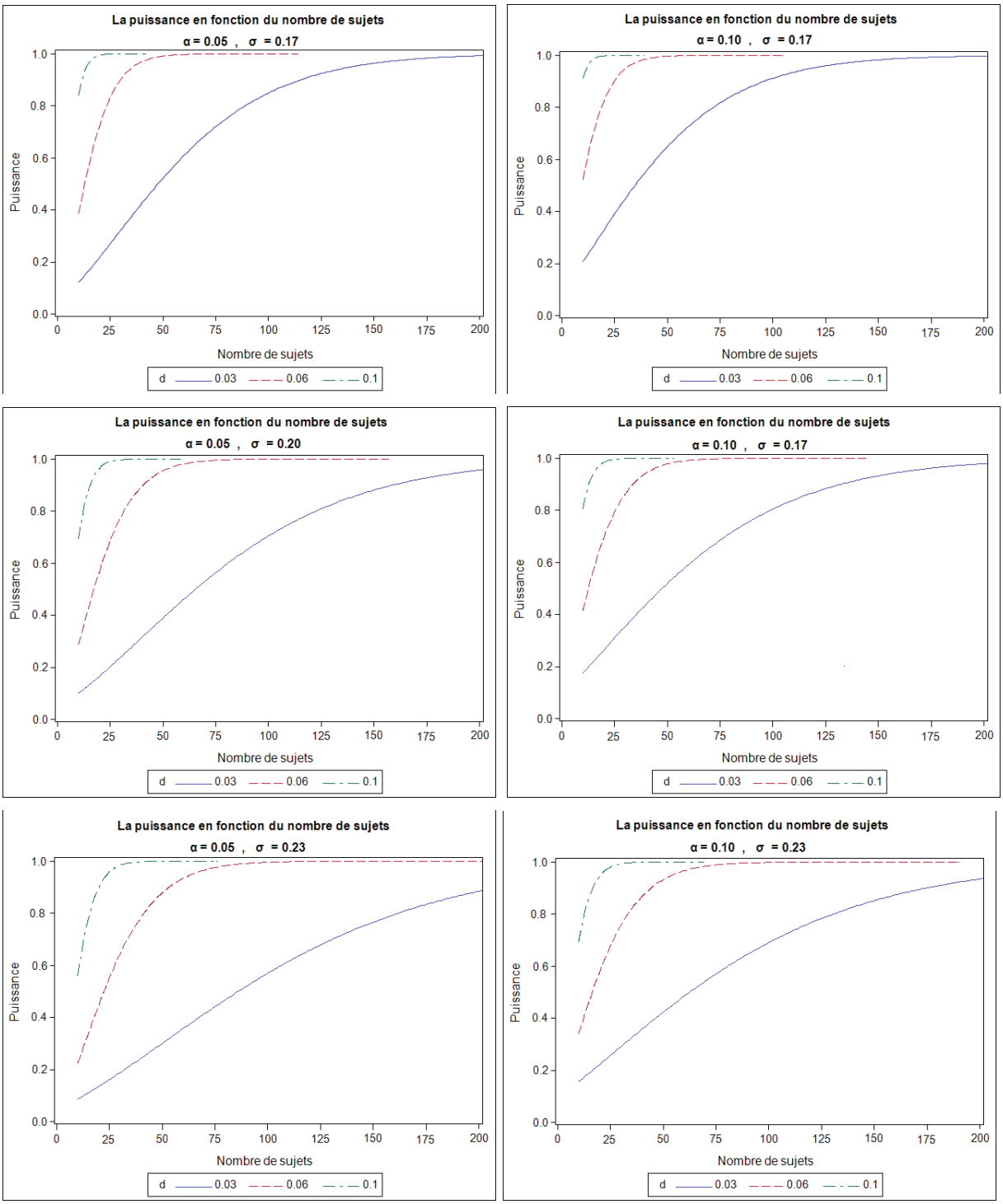


FIG. 3.11 – *Puissance en fonction du nombre de sujets pour k 4 produits*

4 Discussion : Lien entre les résultats de l'approche par rééchantillonnage et de l'approche par méta-analyse

Les résultats des deux démarches nous ont permis de mettre en exergue l'impact de la taille des différences à mettre en évidence dans la détermination du nombre minimum de sujets. En effet, les recommandations obtenues par l'approche rééchantillonnage pour les sept études menées dans le cadre de l'expérimentation de la nouvelle norme (AFNOR, 2009), variaient de 20 à 150 sujets. Le paramètre qui semble le plus conditionner le nombre de sujets était la taille des différences en termes d'appréciation entre les produits comparés. L'hétérogénéité des préférences était presque similaire pour les sept études.
Par ailleurs, l'approche par méta-analyse consistant à calculer le nombre de sujets en fonction des estimations de l'hétérogénéité des préférences et des tailles de différences d'appréciation observées sur PrefBase montre l'impact de ces dernières sur les recommandations sur le nombre de sujets. Pour les sept études ACTIA, nous avons constaté une concordance des recommandations sur le nombre de sujets obtenues par l'approche rééchantillonnage avec les recommandations établies en fonction de la taille de l'effet par l'approche méta-analyse de PrefBase. En effet, les tableaux (3.6) et (3.5) illustrent cette remarque. Les études sont classées selon les trois catégories de la grandeur de l'effet faible, moyen et fort :

TAB. 3.5 – *Tailles d'effet* σ_m

Source	Tailles d'effet σ_m		
	faible	moyen	fort
Méta-analyse PrefBase	0.03	0.06	0.10
Études ACTIA	Hareng : 0.04	Cake, Compote, Chips : 0.07	Knack 0.16

TAB. 3.6 – *Tailles d'effet standardisé*

Source	Tailles d'effet standardisé		
	faible	moyen	fort
Méta-analyse PrefBase	0.16	0.30	0.47
Études ACTIA	Hareng : 0.23	Cake, Compote, Chips : 0.39	Knack 0.87

L'effet observé sur la gamme "hareng" est considéré comme un petit effet. Les effets observés pour les gammes "cake", "compote" et "chips" sont considérés comme des effets moyens. Enfin, l'effet observé sur la gamme "knack" est considéré comme un grand effet et ainsi plus facile à mettre en évidence avec un petit groupe de sujets (20 sujets).

Chapitre 4

Le nombre de sujets en profil sensoriel

1 Description de la base de données SensoBase

Le système SensoBase (www.sensobase.fr) permet aux fournisseurs de données (laboratoires d'analyse sensorielle) de s'inscrire à ce programme d'échange, d'envoyer leurs études de profil sensoriel et de recevoir en retour des analyses statistiques originales pour le contrôle des performance des sujets et l'étude des différences entre produits. Son fonctionnement repose sur trois organes principaux (le site internet, le fichier Excel© et la base de données) qui assurent les différentes opérations depuis la saisie des données jusqu'à l'envoi des résultats des analyses.

1.1 Architecture de SensoBase

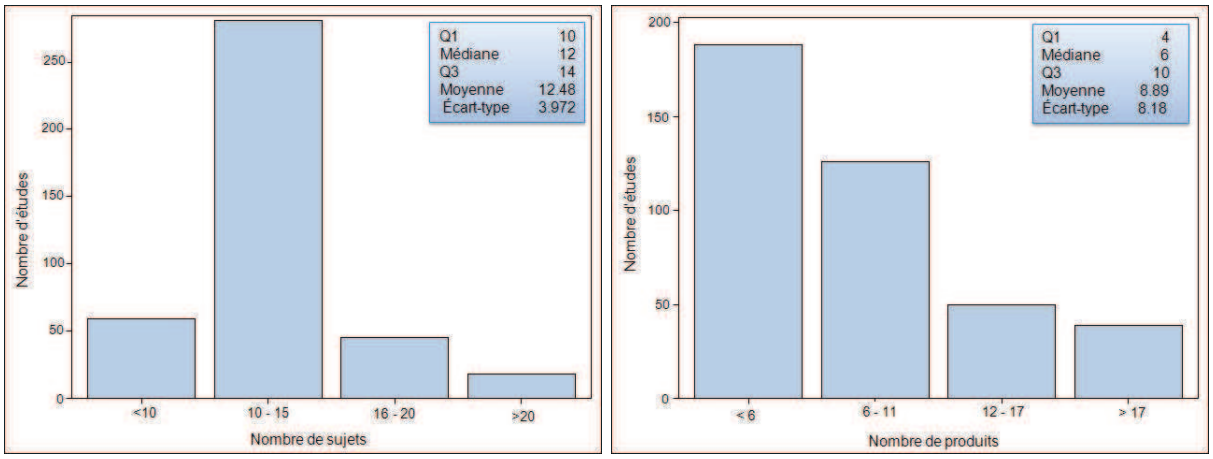
Afin de faciliter le stockage dans une structure unique, les données relatives à chaque étude doivent être saisies dans un fichier Excel© spécifiquement formaté afin de recevoir ce type de données. Ce fichier demande à l'utilisateur de fournir divers renseignements sur l'étude sensorielle (age et sexe des panélistes, type de descripteurs, ...) puis de saisir les données selon un format precis. Le fichier est ensuite envoyé via le site Internet et stocké dans une base de données.
Le stockage des profils sensoriels envoyés par les fournisseurs s'effectue dans une base de données MySQL, gérée en utilisant le logiciel phpMyAdmin.
La base de données est constituée d'un ensemble de tables qui contiennent les informations envoyées par les fournisseurs. La structure de ces tables est définie de manière à minimiser la redondance de l'information. Ainsi, chaque table créée contient toutes les informations relatives à l'objet de cette table. La structure choisie permet de minimiser l'espace occupé par la base de données et facilite les modifications qui pourraient être effectuées ultérieurement. Le schéma structurel de la base, présenté dans la figure (voir annexe 5.1), montre que l'architecture de la

SensoBase repose sur une table principale, *DONNEES*, contenant tous les scores des différents profils sensoriels. Cette table est reliée à quatre tables contenant respectivement les caractéristiques des produits, des sujets, des descripteurs et les informations générales concernant chaque étude (date de réalisation, type de panel, ...). La table jeu de données, *jdd*, est en particulier reliée à la table *fournisseur*, elle-même reliée à la table *contact* contenant les données sur l'animateur de panel du fournisseur en question. Les tables *protocole* et *etat_suj* ne sont pas présentées dans leur intégralité; elles permettent de limiter la redondance de l'information.
Suite à l'insertion des données dans la base, l'utilisateur peut effectuer une demande d'analyse de ce jeu de données dans laquelle certains paramètres des analyses sont réglables. L'utilisateur reçoit ultérieurement les résultats par courrier électronique au format html.

1.2 Les données SensoBase

Le nombre de jeux de données dans SensoBase s'élève actuellement à plus de 900. Ces jeux de données ont été fournies par 57 fournisseurs (laboratoires français et étrangers). Ces données représentent un ensemble de 16 731 descripteurs, 4 410 sujets différents et 5 996 produits, soit un total de 4 831 945 notes d'intensité.

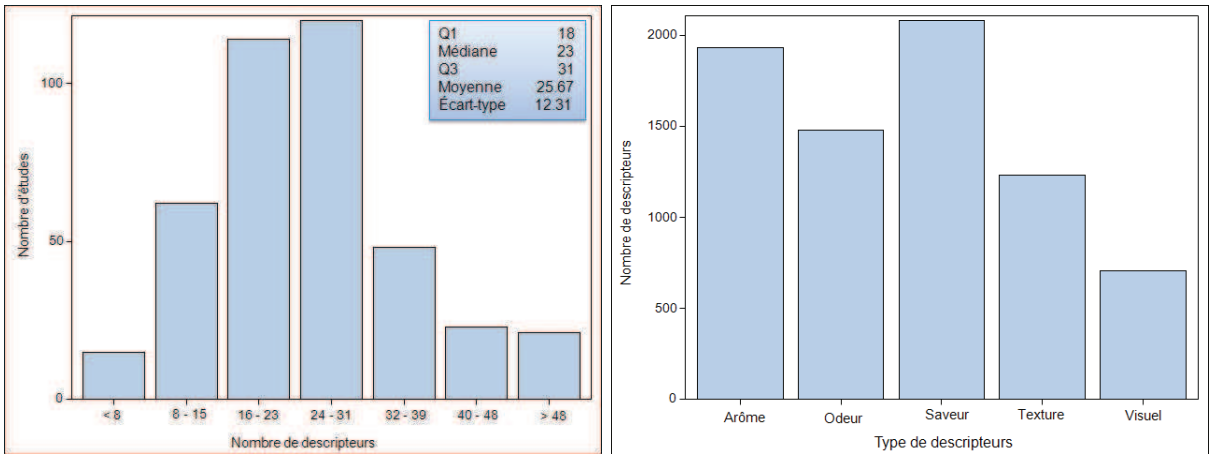
Pour les besoins de notre méta-analyse, nous avons procédé à une selection d'un échantillon de jeux de données. Afin d'obtenir des résultats qui ne soient pas influencés par certains fournisseurs potentiels, il a été décidé de limiter le nombre de jeux de données à 40 par fournisseur (moyenne du nombre de jeux de données des fournisseurs ayant fourni plus de 10 jeux de données). Cette sélection des jeux de données a été obtenue par tirage aléatoire sans remise pour chaque fournisseur. Ainsi, 405 jeux de données au total ont été retenus, soit l'étude de 10 347 descripteurs. Les caractéristiques de cet ensemble de profils sensoriels qui seront analysés dans ce chapitre, sont illustrées par les figures (4.1) et (4.2).



(a) distribution du nombre de sujets par étude (b) distribution du nombre de produits par étude

FIG. 4.1 – *Distribution du nombre de sujets et du nombre de produits*

Les valeurs 10 et 12 correspondent aux nombre de sujets les plus fréquemment utilisés. Ce choix correspond aux recommandations préconisées pour la sélection des sujets décrites par la norme (AFNOR, 1993).



(a) distribution du nombre de descripteurs par étude (b) distribution du type de descripteurs sur l'ensemble des études

FIG. 4.2 – *Distribution du nombre de descripteurs ainsi que du type des descripteurs*

Les descripteurs sont classés en cinq catégories: apparence, texture, saveur, odeur et arôme (lorsque le produit est mis en bouche). Les descripteurs visuels sont moins représentés que les autres. De plus, il est utile de préciser que chaque catégorie n'est pas représentée dans chaque étude. Certains profils peuvent avoir été réalisés avec deux catégories de descripteurs, voire une seulement. Cet aspect pourrait avoir une influence lors de la comparaison par types de

descripteurs.

Au vu de ces caractéristiques (4.1) et (4.2), un jeu de données médian dans SensoBase, correspond à l'étude d'un groupe de 6 produits, par un panel de 12 sujets en utilisant 23 descripteurs.

Ce chapitre comporte deux parties. Tout d'abord, nous présentons les résultats de l'approche par rééchantillonnage appliquée sur l'ensemble des 405 jeux de données sélectionnés. Nous discuterons les résultats obtenus. Ensuite, nous étudierons des paramètres qui déterminent le nombre de sujets nécessaire à savoir, la taille de l'effet et le désaccord entre les sujets.
Dans la deuxième partie, nous présentons les résultats sur l'étude des répétitions en profils sensoriels. Ces résultats sont valorisés dans un article.

2 Approche par rééchantillonnage

Nous avons appliqué la démarche par rééchantillonnage sur les 405 jeux de données de profils sensoriels. L'objectif était de tester dans quelle mesure un sous-panel de N-k sujets conduirait à des résultats différents ou non d'un panel complet à N sujets.

2.1 Modèle d'analyse

Comme le nombre de répétitions varie d'un jeu de données à un autre, les notes d'intensité données par les sujets pour l'ensemble des produits ont été moyennées sur les répétitions afin d'homogénéiser les données.
Par ailleurs, comme différentes échelles de notation ont été utilisées et pour rendre les grandeurs de l'effet produit et du désaccord entre les sujets comparables entre les différents jeux de données, les scores d'intensité pour chaque descripteur sont ramenés à l'échelle 0-1 en appliquant une transformation affine. Nous avons donc appliqué le modèle de l'analyse de la variance additif à deux facteurs Sujet + Produit.

2.2 Simulation et définition des critères d'analyse

Pour chaque jeu données, 100 sous-panels de taille $N - k$, $k = 0$,...,*N-2* ont été obtenus par tirage aléatoire avec remise à partir du panel complet. Nous avons considéré le cas $k = 0$, pour tester la stabilité des résultats statistiques si on avait pris 100 panels de taille N ayant des caractéristiques presque similaires.

Pour vérifier la stabilité de la conclusion statistique sur les produits, nous avons défini des critères unidimensionnels et multidimentionnels.
Pour l'analyse unidimensionnelle, nous avons calculé :

- le coefficient de corrélation entre le vecteur des moyennes des notes par produit obtenu à partir du sous-panel à $N-k$ sujets et ce même vecteur obtenu avec le panel complet à N sujets.
- le coefficient de discrimination ρ (coefficient de Fisher) défini à partir du modèle ANOVA à deux facteur Sujet + Produit tel que :

$$\rho = \frac{CM_{Pro}}{CM_{Pro} + CM_{Err}} \tag{4.1}$$

  CM_{Pro} et CM_{Err} sont respectivement les carrés moyens de l'effet produit et de l'erreur. Ce coefficient mesure l'importance de l'effet produit. Étant compris entre 0 et 1, il est plus facile d'utilisation que la statistique F de Fisher lorsqu'il s'agit de comparer plusieurs études. En outre, ce coefficient n'est pas sensible aux très faibles valeurs du carré moyen de l'erreur.

Au niveau multidimentionnel, nous avons calculé le coefficient RV (Escoufier, 1973) qui correspond à une corrélation entre deux configurations de produits obtenues respectivement en considérant les notes du panel complet et les notes du panel réduit à $N-k$ sujets. Ainsi, si X et Y sont les deux matrices $l \times p$ (l produits, p descripteurs) centrées par colonnes, issues respectivement des scores d'intensité donnés par le panel complet (N sujets) et par le panel réduit (N-k sujets), le coefficient RV est la quantité :

$$RV(X,Y) = \frac{tr(XX'YY')}{\sqrt{tr((XX')^2)tr((YY')^2)}} \tag{4.2}$$

Un coefficient RV proche de 1 signifie que les deux configurations X et Y sont identiques et de 0 si elles sont totalement différentes. Ce critère nous renseigne sur les distances entre les produits au niveau multidimentionnel.

2.3 Règles de décisions et constitution de recommandations pour n

Contrairement aux études de préférences qui consistent à étudier une seule variable de mesure (la mesure hédonique), une étude de profil sensoriel est un objet multidimentionnel qui constitue l'étude de plusieurs variables (descripteurs) à la fois. Les différentes études de profils analysées présentent des caractéristiques variées en terme du nombre de sujets enrôlés, du nombre et du type de descripteurs utilisés et de la famille de produits testés. Cette variabilité rend difficile le choix et la validation des critères de décision. Néanmoins, nous avons fait des choix et établi

des critères de décision qui prennent en considération cette complexité du profil sensoriel et qui seront discutés.

Puisque l'objectif de l'approche est de donner une recommandation sur le nombre de sujets, l'unité expérimentale retenue est le jeu de données. À chaque jeu de données correspond une recommandation sur le nombre minimal de sujets.
Au niveau unidimensionnel, chaque critère de décision doit respecter les points suivants pour les différents indices calculés :

- puisque la conclusion statistique consiste en la mise en évidence d'un effet produit, on ne s'intéressera alors qu'aux descripteurs significatifs ($p = 0.05$). Par ailleurs, pour avoir une estimation de la perte d'information en terme de perte de significativité, nous avons calculé le pourcentage de descripteurs qui cessaient d'être significatifs au fur et à mesure que l'on réduisait la taille du panel.
- l'étude d'un jeu de données est décomposée par types de descripteurs : saveur, arôme, odeur, texture et apparence. La recommandation sur le nombre de sujets sera le maximum des recommandations par type de descripteur.

Il s'en suit que pour un jeu de données les critères de décision par descripteur sont :

- pour chaque k, pas plus de 10% des 100 sous-panels ayant un coefficient de corrélation R inférieur à 0.9. Nous considérons qu'un coefficient de corrélation à 0.9 est un bon coefficient.
- Le coefficient de discrimination de Fisher ρ est calculé pour chaque sous-panel (CMP et CME spécifiques pour chaque taille de panel). C'est pourquoi, nous avons défini un seuil de perte en discrimination de 10% par rapport à la discrimination totale calculée à partir du panel complet ($\rho_{(N)} - \rho_{(N-k)} = 10\%$) pour établir une recommandation. En d'autres termes, pour k le nombre de sujets retirés, pas plus de 10% des sous-panels qui conduisent à un $\rho_{(N-k)} < 0.9 * \rho_{(N)}$. La recommandation sur le nombre de sujets n, $n = N - k$ correspond au plus petit n qui vérifie cette condition.

La recommandation pour un jeu de données est issue par niveau. Ainsi, à chaque type de descripteur correspond une recommandation qui est le maximum des recommandations obtenues pour les descripteurs du même type. Le maximum des recommandations par type de descripteurs correspond alors à la recommandation pour le jeu de données.
Ce choix que nous avons fait nous permet d'obtenir une recommandation qui garantit une perte d'information minimale pour l'ensemble des descripteurs d'un jeu de données.

Au niveau multidimentionnel, le coefficient RV est calculé pour l'ensemble des descripteurs à la fois significatifs et non significatifs. Ainsi la règle de décision est :

– pour chaque k, pas plus de 10% des 100 sous-panels ayant un coefficient RV inférieur à 0.9. Nous considérons qu'un coefficient RV de 0.9 traduit une bonne similarité entre les configurations produit du sous-panel et du panel complet.

2.4 Résultats et discussion

L'analyse de la variance appliquée sur l'ensemble des descripteurs des 405 jeux de données montrent que seulement 62% des descripteurs étudiés étaient significatifs ($p = 0.05$), soit plus d'un tiers des descripteurs qui sont non significatifs. Ce résultat pourrait être imputé à une utilisation inadéquate d'une partie de ces descripteurs. En effet, certains descripteurs développés pour une étude descriptive donnée peuvent être réutilisés dans d'autres études de la même famille de produits. Si ces descripteurs sont moins perçus, ils risquent d'induire à un résultat non significatif. Cette pratique est parfois utilisée par certains laboratoires lorsque les panels sont déjà entraînés pour une certaine catégorie de produits.

Par ailleurs, le calcul du pourcentage de descripteurs qui devenaient non significatifs au fur et à mesure que l'on réduisait le nombre de sujets indique qu'en moyenne :

– 83% des descripteurs restaient significatifs lorsqu'on réduit à trois quarts le panel complet
– 68% des descripteurs demeuraient significatifs lorsqu'on réduit de moitié le panel complet
– et seulement 47% des descripteurs demeuraient significatifs lorsqu'on réduit à un quart le panel complet.

Ces résultats montrent une diminution sensible du nombre de descripteurs significatifs lorsque le panel est réduit de moitié. Par conséquent la conclusion statistique obtenue à partir du panel complet serait altérée.
King et al. (1995) a montré des résultats similaires en analysant deux études de profil flaveur et texture. En effet, la perte d'information en termes de significativité des descripteurs imputée à la réduction du panel est plus importante dans l'étude de profil flaveur que que dans l'étude de profil texture. Le nombre de sujets dans les deux études était de 19 et 20 sujets respectivement. De plus, l'impact de réduction du panel affecte sensiblement la significativité des descripteurs de type arôme, odeur et saveur. Le tableau (4.1) montre le pourcentage des descripteurs significatifs en fonction du pourcentage de réduction du panel complet par type de descripteurs :

TAB. 4.1 – *Pourcentage de perte de significativité en fonction du type de descripteur*

Type d'attributs	Nombre de jeu de données	Nombre d'attributs	Panel réduit de:		
			25%	50%	75%
Arôme	284	1933	77.94	61.99	39.88
Odeur	162	1481	79.05	65.47	40.02
Saveur	286	2082	79.34	63.08	41.33
Texture	241	1234	87.04	73.58	56.47
Visuel	177	708	92.15	79.95	67.92

Les deux sections suivantes présentent les résultats des recommandations sur le nombre de sujets pour les 405 études obtenues en utilisant les critères de décision définis au niveau unidimensionnel et multidimentionnel.

2.4.1 Recommandations sur le nombre de sujets au niveau unidimensionnel

Les recommandations sur le nombre de sujets n, obtenues en fonction des deux critère de corrélation R et de discrimination ρ varient de 2 à N. Ces deux critères donnent généralement la même recommandation pour un jeu données (dans 82% des cas).

En outre, ces deux critères donnent en moyenne des recommandations qui tolèrent une réduction de 2 à 3 sujets du panel complet sans altérer la conclusion statistique des données.
Comme la moyenne des tailles de panel utilisés dans les différents jeux de données analysés est de 12 sujets, la moyenne des recommandations obtenues par les deux critères de corrélation et de discrimination est de 10 et 9 sujets respectivement.
Pour environ 25% des jeux de données, la recommandation n est simplement égale à la taille du panel complet. Nous avons constaté pour ces jeux de données un faible indice de discrimination ρ qui traduit de faibles différences entre les moyennes des produits.

Par ailleurs, l'analyse de ces recommandations par type de descripteurs nous informe que les descripteurs de type apparence et texture nécessitent généralement moins de sujets que les descripteurs de type arôme, odeur et saveur (table 4.2). Pineau (2006), montre que l'indice de discrimination est plus élevé pour les descripteurs visuels lorsque le niveau d'accord est plus élevé pour les descripteurs visuels, puis pour les descripteurs de texture. Ces résultats semblent en accord avec nos résultats pour les différents types de descripteurs. L'accord des sujets pour les descripteurs de type visuel et texture provient probablement d'une meilleure compréhension de ces descripteurs par les panélistes et par une plus grande facilité à discriminer les produits sur ces critères, contrairement aux descripteurs de type arôme, odeur et saveur, plus difficile à appréhender. Ce résultat signifie que l'on peut se contenter de plus petits panels pour l'étude des descripteurs de type visuel et/ou texture (profil texture).

TAB. 4.2 – *Recommandations par type de descripteur*

Type d'attributs	Nombre d'attributs	Nombre de jeux de données	Nombre de sujets (N)		Recommandation (n)	
			Étendue	Moyenne- IC*	Étendue	Moyenne- IC*
Arôme	1933	284	6-32	12.71 [12.31; 13.10]	2-28	9.61 [9.21; 10.01]
Odeur	1481	162	4-24	12 [11.48; 12.52]	2-19	8.37 [7.81; 8.92]
Saveur	2082	286	4-27	12.40 [12.00; 12.80]	2-26	8.39 [7.97; 8.80]
Texture	1234	241	6-26	12.69 [12.28; 13.10]	2-21	7.75 [7.30; 8.20]
Visuel	708	177	6-26	12.77 [12.25; 13.30]	2-25	6.36 [5.82; 6.91]

(*) Intervalle de confiance à 95%

La table (4.3) donne les recommandations pour les grandes familles de produits (au minimum 10 études par catégorie de produits).

TAB. 4.3 – *Recommandations par famille de produits*

Famille de produits	Nombre de jeux de données	Taille de panel (N)		recommandation (n)	
		Étendue	Moyenne - IC*	Étendue	Moyenne - IC*
Fromages	79	8-24	12.35 [11.65 ; 13.05]	5-23	10.51 [9.79 ; 11.23]
Boissons alcoolisées	41	4-14	9.46 [8.56 ; 10.36]	3-14	8 [7.09 ; 8.90]
Condiments & sauces	33	6-16	11.36 [10.50 ; 12.22]	2-15	9.36 [8.40 ; 10.32]
Yaourts	25	10-12	11.24[10.94 ; 11.53]	7-11	9.56 [9.11 ; 10]
Vins & Champagne	23	10-32	16.86[14.81 ; 18.92]	6-28	14 [11.68 ; 16.31]
Charcuterie	21	11-26	16.80[15.07 ; 18.54]	9-25	14.38 [12.57 ; 16.18]
Plats préparés	20	10-13	11.5[11.03 ; 11.96]	8-12	9.90 [9.29 ; 10.50]
Produits laitiers	15	5-14	10.2 [8.71 ; 11.68]	4-14	8.06 [6.60 ; 9.52]
Confiseries & chocolats	13	6-12	10[8.95 ; 11.04]	3-11	7.84 [6.61 ; 9.07]
Fruits	12	11-18	14.16[12.56 ; 15.76]	7-16	11.75 [9.61 ; 13.88]
Pains & biscottes	11	12-18	14.45[13.44 ; 15.46]	11-14	12.72 [12.11 ; 13.33]

(*) Intervalle de confiance à 95%

Les recommandations par famille de produits correspondent à 2 à 3 sujets en moins que la taille de panel de départ. Il apparaît que pour la catégorie "vins & champagne" et la catégorie "charcuterie" nécessiteraient plus de sujets, soit en moyenne 14 sujets comparé aux autres catégorie de produits ayant englobé le même nombre de sujets au départ.

2.4.2 Recommandations sur le nombre de sujets au niveau multidimentionnel

Les recommandations sur le nombre de sujets obtenues en fonction du critère de corrélation multidimensionnelle RV, coincident avec celles obtenues au niveau unidimensionnel. Ces recommandations constituent en général une réduction de 2 à 3 sujets du panel complet. Ainsi, la moyenne des recommandations pour l'ensemble des jeux de données est de 10 sujets.
Pour 27% des jeux de données, la recommandation n est égale à la taille du panel complet.
Pour une meilleure interprétation de la perte d'information au niveau multidimentionnel, nous avons mesuré l'effet produit par la statistique F issue du modèle de l'analyse de la variance multivariée MANOVA à deux facteurs Sujet + Produit. Il est intéressant de noter que les recommandations pour le coefficient RV coincident avec le pouvoir discriminant au niveau multidimentionnel. En effet, plus l'étude de profil était discriminante caractérisée par une forte valeur de la statistique F de MANOVA, plus la recommandation sur le nombre de sujets obtenue par le critère du coefficient RV diminuait. La recommandation sur le nombre de sujets pouvait descendre jusqu'à 2 sujets!
Le pouvoir discriminant au niveau multidimentionnel traduit généralement le pouvoir discriminant au niveau unidimensionnel. Plus les descripteurs sont discriminants caractérisés par une statistique F et un coefficient de discrimination ρ importants, plus on a de chance d'obtenir une forte valeur de la statistique F de MANOVA et ainsi une meilleure discrimination multidimensionnelle.

2.5 Conclusion

Les recommandations sur le nombre de sujets n, obtenues en fonction des critères de corrélation R, RV et de discrimination ρ varient de 2 à N. En général, les panels ne peuvent être réduits de plus de 3 sujets. Comme la taille des panels de départ variait de 5 à 32 sujets, la recommandation moyenne obtenue sur l'ensemble des jeux de données était de 10 sujets.

La nature multidimensionnelle et complexe du profil sensoriel a montré les limites de l'approche par rééchantillonnage à donner une recommandation générale quant au nombre de sujets à enrôler dans un panel de profil sensoriel. Néanmoins, cette approche a montré que le nombre de sujets dépendrait du type de descripteurs et de leur niveau de significativité. Pour les descripteurs de type apparence et texture, on arrivait à reconstituer l'information sur la discrimination

des produits avec des sous-panels de petites tailles. En revanche, pour les descripteurs de type arôme, odeur et saveur, les conclusions sur la discrimination des produits sont très vite altérées lorsqu'on réduit le nombre de sujets. Ce résultat pourrait être expliqué par la facilité pour les sujets à discriminer les produits avec les descripteurs de type apparence et texture, contrairement aux descripteurs de type arôme, odeur et saveur, plus difficile à appréhender.

3 Approche théorique rétrospective : Méta-analyse de SensoBase

À l'image de ce qui a été fait dans le cadre des études hédoniques, nous avons mené une méta-analyse sur l'ensemble des 405 études de profil sensoriel. L'objectif était d'estimer les paramètres qui conditionnent le nombre de sujets en profil sensoriel, de définir leurs domaines de variation et de pouvoir adresser des recommandations sur le nombre de sujets à enrôler pour ce type d'épreuves.

3.1 Le modèle d'analyse et estimation des paramètres

Nous avons étudié la variation des grandeurs de l'effet et du désaccord entre les sujets par descripteur et par type de descripteurs. L'objectif est de pouvoir donner au final, une proposition sur le nombre de sujets à enrôler pour une étude de profil sensoriel.

Rappelons que le modèle d'analyse retenu pour estimer la grandeur de l'effet et le désaccord entre les sujets est le modèle de l'analyse de la variance à deux facteurs Sujet + Produit appliqué aux scores d'intensité. Ces scores d'intensité sont ramenés à l'échelle 0-1 en appliquant une transformation affine comme dans la démarche par rééchantillonnage. Pour chaque descripteur, nous avons estimé :

- la variabilité σ par le racine carrée du carré moyen de l'erreur notée RMSE (Root Mean square of error) du modèle ANOVA
- la taille de l'effet d par la racine carrée de la variance des moyennes des produits notée σ_{prod}, $\sigma_{prod} = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(x_{i.} - x_{..})^2}$

3.2 Résultats et discussion

3.2.1 La taille de l'effet

Nous avons étudié pour chaque descripteur, l'écart moyen des scores d'intensité par rapport à la moyenne globale (σ_{prod}). La figure (4.3) montre la distribution de σ_{prod} sur l'ensemble des jeux de données.

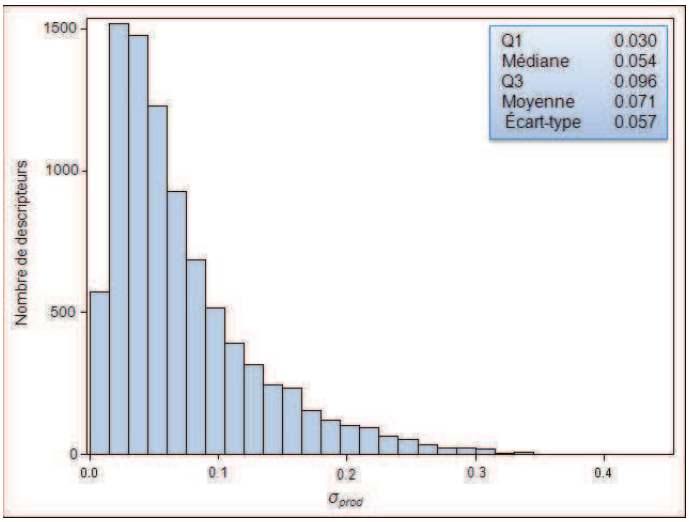


FIG. 4.3 – *Distribution de la taille de l'effet en profil sensoriel*

Nous observons que la taille de l'effet qui traduit l'écart moyen des moyennes des produits par rapport à la moyenne globale est en moyenne de 0.071. La médiane de ces écarts est de 0.054. Cette valeur est légèrement inférieure à la médiane des écarts des notes de préférences observées dans le cas des tests hédoniques. Ceci pourrait être expliqué par le fait que le profil sensoriel soit utilisé pour caractériser de petites différences, il en résulte ainsi une variance des moyennes des produits plus petite.

Il est intéressant de noter que la taille de l'effet varie en fonction du type de descripteur. En effet, on observe des tailles d'effet plus importantes pour les descripteurs de type apparence et texture. Ceci rejoint notre conclusion par l'approche par rééchantillonnage sur la facilité pour les sujets à discriminer les produits avec les descripteurs d'apparence et de texture, contrairement aux descripteurs de type arôme, odeur et saveur, plus difficile à appréhender. Le tableau (4.4) donne les mesures du premier quartile, de la médiane, du troisième quartile et de la moyenne observées des distributions de σ_{prod} en fonction du type de descripteurs. Lorsque les descripteurs de type arôme ont une moyenne un écart σ_{prod} moyen de 0.06, l'écart moyen σ_{prod} des descripteurs visuels est presque double soit 0.103.

TAB. 4.4 – *Taille de l'effet en fonction du type de descripteurs*

Type d'attributs	Q1	Médiane	Moyenne	Q3
Arôme	0.027	0.048	0.062	0.082
Odeur	0.029	0.050	0.064	0.088
Saveur	0.026	0.043	0.056	0.071
Texture	0.044	0.073	0.086	0.119
Visuel	0.051	0.089	0.103	0.145

3.2.2 Le désaccord entre les sujets

Nous avons étudié le second paramètre qui conditionne le nombre de sujets à savoir le désaccord entre les sujets. La figure (4.4) montre la distribution des RMSE, estimateur de cette variabilité individuelle issue du modèle ANOVA appliqué à l'ensemble des descripteurs.

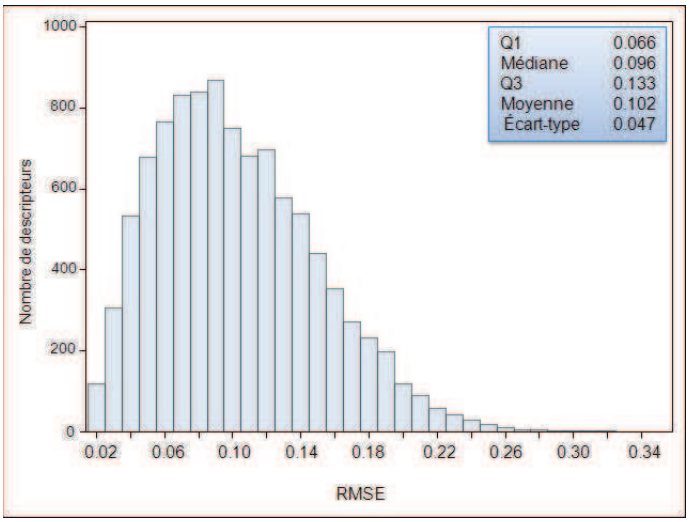


FIG. 4.4 – *Distribution du désaccord entre les sujets RMSE*

Nous pouvons constater que le désaccord entre les sujets est en moyenne de 0.102 sur une échelle de $0-1$, soit une variabilité de 1 point sur une échelle de 0 à 10. Le premier quartile est égal à 0.066 lorsque le troisième quartile est égal à 0.133.
Nous pouvons remarquer que ces trois mesures de la distribution du désaccord entre les sujets sont bien inférieures aux valeurs observées dans les cas des tests hédoniques (0.17, 0.20 et 0.23). Cette baisse du désaccord est acquise grâce aux séances d'entraînement comme cela est préconisé par la méthodologie.
Il est intéressant de noter que les descripteurs non significatifs sont caractérisés le plus souvent par un désaccord entre les sujets plus important. Ces descripteurs ont en moyenne un RMSE supérieur à la moyenne globale 0.10, observée sur l'ensemble des descripteurs. Par ailleurs, l'étude du RMSE pour ces descripteurs non significatifs en fonction des familles de produits indique des RMSE en moyenne de 0.14 pour les familles "boissons alcoolisées", "vins & champagne" et "glace". Les descripteurs sont essentiellement les descripteurs de type odeur et arôme.

3.2.3 La taille de l'effet standardisé

Nous avons calculé les tailles d'effet standardisées selon les suggestions de Cohen (1988). Nous avons calculé la mesure $f = \frac{\sigma_m}{\sigma}$, en remplaçant σ_m et σ par σ_{prod} et RMSE respectivement. La

figure (4.5) montre la distribution des tailles d'effet standardisé observées sur l'ensemble des jeux de données de SensoBase.

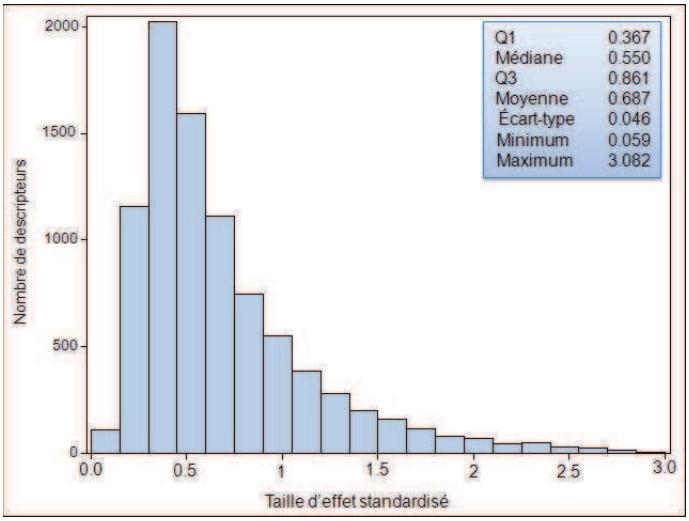


FIG. 4.5 – *Distribution des tailles de l'effet standardisé*

En prenant les trois mesures du premier quartile, de la médiane et du troisième quartile, nous avons défini trois catégories de tailles d'effet : faible, moyen, fort. Ces mesures ont été comparées aux autres mesures conventionnelles de tailles d'effet établies en sciences du comportement par Cohen (1988) et celles observées sur PrefBase. Ces mesures sont rapportées dans le tableau (4.5).

TAB. 4.5 – *Tailles d'effet standardisé*

Source	Tailles d'effet standardisé		
	faible	moyen	fort
Science du comportement	0.10	0.25	0.40
Méta-analyse PrefBase	0.16	0.30	0.47
Méta-analyse SensoBase	0.36	0.55	0.86

Les trois mesures indiquent que les tailles d'effet standardisé observées en profil sensoriel sont supérieures à celles observées en tests hédoniques. Ce résultat est dû aux faibles valeurs du désaccord entre les sujets en profil sensoriel. Une baisse du désaccord acquise grâce aux séances d'entraînement.

3.2.4 Recommandations sur le nombre de sujets en profil sensoriel

Pour établir les recommandations sur le nombre de sujets nous avons utilisé les estimations de d et σ obtenues précédemment. Comme les études de profil sensoriel concernent souvent l'étude

de plusieurs types de descripteurs, nous avons choisi de donner des recommandations globales selon les effets observés sur les descripteurs de l'ensemble des jeux de données.
Ce calcul a été fait en considérant une étude de k égal à 6 produits qui correspond au nombre moyen de produits testés dans une étude de profil sensoriel. Le tableau (4.6) résume les différentes recommandations sur le nombre de sujets obtenues pour les différentes valeurs de α, β, d et σ.

TAB. 4.6 – *Abaques du nombre de sujets à enrôler en profil sensoriel pour k 6 produits*

σ	d	α	β		
			20%	10%	5%
0.066	0.030	20%	8	10	12
		10%	10	13	15
		5%	12	15	18
	0.054	20%	4	4	6
		10%	5	6	7
		5%	5	6	5
	0.096	20%	3	3	4
		10%	3	3	3
		5%	2	2	3
0.096	0.030	20%	15	21	26
		10%	20	27	32
		5%	25	32	38
	0.054	20%	6	8	10
		10%	8	10	13
		5%	10	12	15
	0.096	20%	4	4	4
		10%	4	4	5
		5%	3	5	6
0.133	0.030	20%	27	36	65
		10%	35	46	56
		5%	43	55	45
	0.054	20%	10	21	17
		10%	14	17	21
		5%	17	14	25
	0.096	20%	6	8	9
		10%	5	6	8
		5%	4	5	6

Comme on peut le constater, le paramètre qui semble le plus influer sur le nombre minimum de sujets à enrôler dans un panel de profil sensoriel est la taille des différences à mettre en évidence.

Si on considère un désaccord entre les sujets moyen égal à 0.096, un risque $\alpha = 5\%$, une puissance de test à 90%, il faudrait 32 sujets pour mettre en évidence un petit effet, 12 sujets pour mettre en évidence un effet moyen et 5 sujets seulement pour mettre en évidence un grand effet !

3.3 Conclusion

Cette approche par méta-analyse confirme les résultats de l'approche par rééchantillonnage. En effet, la taille de l'effet varie en fonction du type de descripteur. Nous avons observé des tailles d'effet plus importantes pour les descripteurs de type apparence et texture.
Pour établir les abaques du nombre de sujets, nous avons considéré trois mesures pour le désaccord entre les sujets et pour la taille de l'effet. Ces trois mesures permettent de définir trois catégories d'effet faible, moyen et fort. Elles auront pour objectif d'aider les utilisateurs à décider sur le nombre de sujets à enrôler pour une étude de profil sensoriel. Ainsi, pour un désaccord moyen entre les sujets σ égal à 0.096, un risque α de 5%, une puissance de test à 90%, il faudrait au minimum 12 sujets pour détecter un effet global moyen.

4 Les répétitions en profil sensoriel : Article 2

Dans les études de profil sensoriel, l'entraînement des sujets est réalisé sur plusieurs séances. Cet entraînement permet aux sujets une meilleure utilisation des échelles de notations, mais aussi de se familiariser avec les termes descriptifs et d'en donner une interprétation sensorielle en accord avec les autres panélistes. Cependant, une fois les sujets entraînés, nous ne savons pas si ces répétitions amélioreraient réellement la discrimination entre les produits lors de la phase de mesure. Cet article examine la nécessité de ces répétitions au cours de cette phase pour la discrimination des produits.

No need to replicate with trained sensory panels

N. Mammasse[a,*], S. Cordelle[a], P. Schlich[a]
[a] Centre des Sciences du Goût et de l'Alimentation CSGA, UMR6265 CNRS, UMR1324 INRA, Université de Bourgogne, Dijon, France.

Soumis, Mars 2012 - Food Quality and Preference

No need to replicate with trained sensory panels

N. Mammasse[1,2,3,*], S. Cordelle[1], P. Schlich[2]

1 *CNRS, UMR6265 Centre des Sciences du Goût et de l'Alimentation, F-21000 Dijon, France*
2 *INRA, UMR1324 Centre des Sciences du Goût et de l'Alimentation, F-21000 Dijon, France*
3 *Université de Bourgogne, UMR Centre des Sciences du Goût et de l'Alimentation, F-21000 Dijon, France*
E-mail : *nadra.mammasse@u-bourgogne.fr*

Abstract

Replications are necessary to evaluate individual repeatability in sensory profiling. However, once panels are trained, most panel leaders continue to run sensory profiling studies with replicates. Are such replicates actually necessary? Does the full data set really provide more information than the data subset composed of the first replicate? Statistical analysis of 224 studies (181 with 2 replicates and 43 with 3 replicates) shows that when using the first replicate alone:

- 6% (2 replicates) and 11% (3 replicates) of the attributes were no longer significantly ($p = 0.05$) discriminative among products,
- 88% (2 or 3 replicates) of the discriminative attributes had a similar vector of product mean scores, i.e. a correlation coefficient higher than 0.9,
- 90% (2 or 3 replicates) of all studies had a similar product configuration, i.e. an RV coefficient higher than 0.9,
- distribution of the number of significant dimensions of the product space and distribution of the product F-ratio in the two-way MANOVA model were non-significantly different.

These data suggest that replication is unnecessary once the panel is trained.
Keywords : Sensory profiling, replication, correlation, ANOVA, MANOVA, RV.

1 Introduction

Sensory profiling has been widely used in sensory analysis. The aim of the method is to describe the sensory characteristics of food products using a list of defined attributes. Panelists

involved in a sensory profiling task are required to rate the perceived intensities of a number of attributes such as: sweetness, saltiness, crunchiness etc.
A classical sensory descriptive profile method consists of two main phases: the training phase and the testing phase (also known as the measurement phase).

There are several sessions in the training process. Training begins with the development of a common sensory vocabulary. The panel leader works with a group of six to twenty assessors to define a list of sensory attributes, describing the sensations of appearance, texture, aroma, odour and taste relevant to the products in the study. This descriptive vocabulary should be precisely defined, containing enough terms to represent all the sensations to be measured. Assessors are provided with examples of sensory references, then given the opportunity to evaluate the selected attributes in various products, using sensory scales.
The duration of the training phase depends on the complexity of the product and also on the level of experience of the product type that the assessors have. The panel leader checks the performance of the group over time by measuring both repeatability and discriminating ability for the whole panel and for each panelist. Measurement of performance for descriptive sensory panels has been discussed in many papers (Bi, 2003; Chambers et al., 2004; Couronne, 1997; McEwan et al., 2002; Pineau et al., 2007; Schlich, 1994).

The testing phase can begin when assessors seem able to score the attributes accurately and consistently. The trained panel tests the set of products of interest using the determined list of attributes by giving an intensity score on the sensory scales. During this phase, the panelists often evaluate the set of products more than once; therefore multiple replications of testing are completed for each product. Replications are regarded as independent repetitions of the experiment under identical experimental conditions.
Implementation of replication involves some practical considerations, including product availability, preparation requirements and the number of samples to be tested by assessors without causing sensory fatigue.

In practice, several replication structures can be found. Replicates can be randomised within the full experiment, which means that no replication or session effect can be tested. It is also possible to serve the same products in separate tasting sessions, possibly with several days in between. In this case, if product stability is not guaranteed, replication may not be realistic. Another type of replication structure introduces product units in a structured way; for example, samples could come from different batches. In the latter two structures, care should be taken with the analysis of data. Replication should be treated as a random effect in the ANOVA model.

The final data collected during the testing phase have a three-way data structure: Assessors

× Products × Attributes. If there are replicates, they could be added as additional rows. These data are analysed statistically using univariate and multivariate analyses. Multivariate analysis provides a sensory map of the product space.

Although replication is necessary when studying panel performance, it may not be essential during the measurement phase. However, most — if not all — panel leaders still design their sensory profiling studies with replicates. This study seeks to examine whether such a practice is indispensable.
Most papers dealing with sensory profile studies mention the use of two or three replicates. Stone and Sidel (2004) suggest that, empirically, within-subject variability decreases somewhat from the first to the second replicate, levelling at the third one, and that subsequent replications exhibit no further decrease in variability. The authors presume that two to three replicates are adequate for a descriptive sensory profile.

In some food products, such as meat, replications are not always easy to organize (Hunter, 1996). For instance, differences in storage or small variations in serving temperature can have a major effect on sensory characteristics. Therefore care must be taken when designing experiments with replication in order to ensure that appropriate comparisons can be made.

The need to replicate descriptive evaluations has also been discussed in Moskowitz et al. (2008). The authors stated that the appropriateness of replications depends on the objective. Replications allow panel performance to be assessed, for example. Moskowitz (2008) and Muñoz (2008) argued that if a panel is well trained, there will be very little difference between replicates. Hence, for sensory profile studies, replications may be desirable but they are never essential. If replications cannot be accommodated, the data remain valid. In the same discussion paper, Gacula (2008) argues that we get the maximum amount of information from the first judgment of each person. Hence, each additional judgment adds proportionally less information.

Learning from actual data is essential in order to improve sensory methodologies and to investigate different points of view, such as the usefulness of replication in sensory profiling.
In this paper, we will examine whether or not replications are required at the testing phase, using several sensory profile studies from the database SensoBase (`www.sensobase.fr`). The key question in our research is: do two replicates, or even three replicates, enhance product discrimination?
For each study, univariate and multivariate statistical analyses were carried out in order to achieve comparisons between the full data set, including all the replicate scores, and a data subset with only the first replicate scores. Comparisons were based on power discrimination, involving the F-statistic of the product effect, the percentage of significant sensory attributes and the correlation between vectors of product mean scores. At the multivariate level, comparisons were

based on the RV-coefficient, the MANOVA F-statistic and the number of sensory dimensions of the derived sensory spaces.

2 Materials and methods

2.1 Procedure

SensoBase is a database of descriptive sensory studies, currently containing about 900 data sets from 56 sensory laboratories located in 11 countries. Each data set contains the scores of a set of products of a given type (39 different types are represented in SensoBase) profiled by a trained panel of assessors. The database was created in order to describe sensory analysis practices, to compare several statistical methods of analysis and to benchmark panel and panelist performances.

Stratified sampling by sensory laboratory was used to select 405 sensory profile data sets from SensoBase. To assess replicate utility, 224 data sets (all fully balanced, with no missing values) were taken from this previously selected sample: 181 data sets with two replicates, and 43 with three replicates. The remaining data sets were excluded from the study for the following reasons: 23 data sets had a single replicate, 4 data sets had 4 replicates and 1 data set had 6 replicates. A further 153 data sets either had missing values or an unbalanced design.

For data sets with two replicates, comparison was between the analysis of the truncated data set (the first replicate data subset) and the analysis of the full data set. For data sets with three replicates, comparison was between the analysis of the truncated data set (the first replicate data subset) and the analysis of the full data set.

2.2 Univariate analysis

2.2.1 Analysis of variance

For each attribute, a two-way analysis of variance (ANOVA) model with *Product* and *Assessor* as factors was carried out on the intensity scores.

Assessor factor is a normally distributed independent random variable with zero mean as is the error. Variance for assessor factor is σ_A^2 and for error is σ^2.
The intensity scores of the full data set were averaged over replicates in order to fit the same two-way additive model as when a single replicate is considered, with the same number of degrees of freedom.

2.2.2 Univariate correlation

For each attribute, Pearson correlation coefficient was calculated between the vector of product mean scores for the full data set and the vector of product mean scores for the truncated data set.

2.3 Multivariate analysis

For each study, the aim is to analyse all the attributes simultaneously in order to examine whether or not the same product space is derived from the full data set and from the truncated data set.

2.3.1 Multivariate correlation: RV-coefficient

After averaging over assessors, the RV coefficient (Escoufier, 1973) is computed to measure the similarity between the product configurations obtained from the truncated averaged data set and from the full averaged data set. The RV coefficient provides a simple way of measuring the relationship between two product spaces generated by two different sets of variables (sensory attributes), with values between 0 and 1, where 1 represents perfect similarity.

2.3.2 Multivariate analysis of variance

A multivariate analysis of variance (MANOVA) measures product discrimination when all sensory attributes are considered simultaneously. The MANOVA provides 4 multivariate tests for each effect in the model. These statistics pool the variance from all the dimensions to create the statistic test. The four multivariate measurements used to derive the MANOVA F-test are: Pillai's trace, Wilk's Lambda, Hotelling-Lawley's Trace and Roy's Largest Root. The four measurements are calculated using eigenvalues of the matrix A, $A = BW^{-1}$ where B is the hypothesis sums of squares and cross product matrix and W is the error sums of squares and cross product matrix.
For each data set, the two-way additive MANOVA model was applied on the intensity scores of all attributes. We used the F-test based on the Hotelling-Lawley trace, which is the simple sum of all the eigenvalues of the matrix A. This measurement characterizes the ratio of product effect variance to error variance. A significant MANOVA F-test indicates that the products differ significantly in the space generated by the attributes.

2.3.3 Canonical variate analysis

Canonical variate analysis (CVA) is the mean separation technique at the multi-dimensional level. It is used to compute the axes which best discriminate the products. They are defined by the eigenvectors of A. All the CVA axes maximize the distances between product means while minimizing dispersion for individual assessments of each product. A likelihood ratio test is used to assess the number of significant sensory dimensions reflecting the main differences between products.

The use of MANOVA and CVA for sensory profile data has been described in Lawless and Heymann (2010); Monrozier and Danzart (2001); Porcherot and Schlich (2000); Schlich (2004).

All statistical analyses and computations used SAS software release 9.2 (SAS institute Inc., Cary, NC).

3 Results and discussion

3.1 Univariate results

3.1.1 ANOVA results by attribute

For data sets with two replicates, results of the two-way ANOVA model on intensity scores for the first replicate led to a significant product effect (p = 0.05) for 58% of the attributes. On aggregating the second replicate, the percentage was 64%, an increase of 6%.

Similar results were obtained for data sets with three replicates. The two-way ANOVA model results on intensity scores for the first replicate led to a significant product effect (p = 0.05) for 60% of the attributes. The percentage was 68% after aggregating the second replicate and 71% after aggregating the third replicate, and thus an increase of 11%.

Since the median number of attributes in a data set is 23, including replicates would therefore add on average one or at most two significant attributes.

After aggregating the second replicate, the average p-value of these additional significant attributes decreases from 0.18 to 0.02. Analysing the second replicate separately revealed that:

- 2/3 of these additional attributes had a significant product effect in the second replicate. The second replicate was more discriminant than the first, with a higher product mean square. Two hypotheses could explain this situation. Either the first replicate could have served as a final training session, or there might have been some product variation between

the first and second replicates. Therefore some attributes might have been more perceptible in the second replicate. Although these additional attributes had p-values slightly lower than 0.05, they might provide additional information about possible variability between batches. In that case, these additional attributes could be useful for sensory analysts when monitoring a set of products over time, where variability in batches is of interest.

- for the remaining 1/3 of attributes, aggregating the two replicates slightly reduced random noise (RMSE), leading to a significant product effect. Both replicates were therefore necessary in order to reach panel homogeneity. This result might also be a statistical "smoothing effect".

Additionally, about 2% of significantly discriminative attributes, with a p-value slightly lower than 0.05 (average p-values of 0.03) at the first replicate, became non-significant after aggregating the second replicate. For these attributes, the second replicate was not significant and therefore affected significance after aggregation.

In overall, we believe that replication could bring some small additional information at univariate level only in some very rare and specific occasions.

3.1.2 Correlation

The distribution of the correlation coefficients (Figure 1) shows that 88% of the significantly ($p = 0.05$) discriminative attributes had a correlation coefficient higher than 0.9, denoting very good agreement between vectors of product mean scores, based on one or two replicates. The mean correlation coefficient was 0.954.

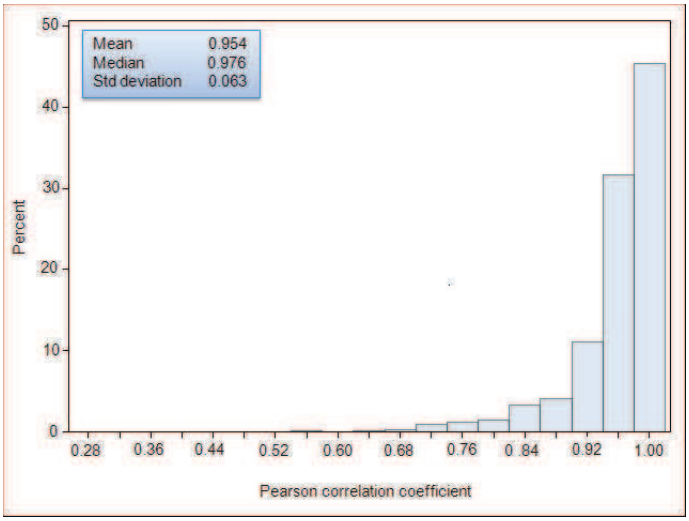


Figure 1: Distribution of Pearson correlation coefficients between vectors of product mean scores for discriminative attributes of the full (2 reps) and truncated (1 rep) data sets

The distribution of the correlation coefficients (Figure2) shows that 37.4% of the non-discriminative attributes had a correlation coefficient higher than 0.9, 88.65% had a correlation coefficient higher than 0.45 and 11.34% had a correlation coefficient lower than 0.45.
Since these attributes were non-discriminative, product mean scores would be closer with changes in ranks leading to correlation coefficients lower than those obtained for discriminative attributes. However the mean and median of these correlation coefficients were still correct, i.e 0.749 and 0.849.

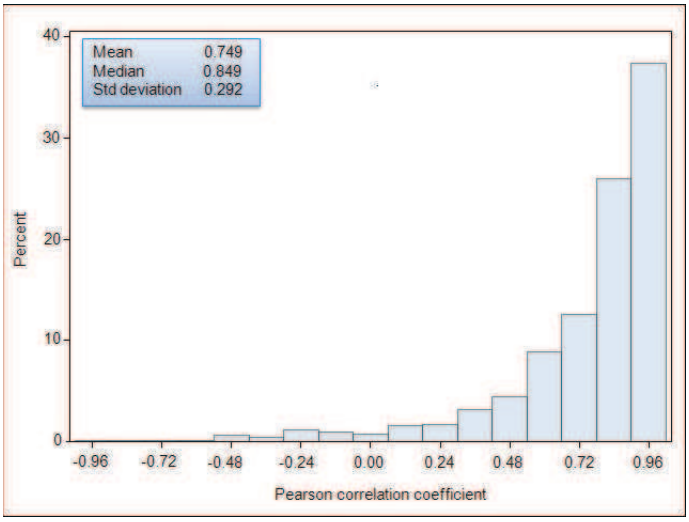


Figure 2: Distribution of Pearson correlation coefficients between vectors of product mean scores for non-discriminative attributes of the full (2 reps) and truncated (1 rep) data sets

The same results for correlation were obtained for data with 3 replicates.

According to ANOVA and correlation results by attributes we expect no significant impact on the full multivariate analysis of these data. This expectation will be studied in the following section.

3.2 Multivariate results

3.2.1 RV coefficient

The RV coefficient was higher than 0.9 for 90% of the data sets (2 or 3 replicates). Thus, the product configurations of the truncated and full data sets were almost the same.

3.2.2 MANOVA and CVA results

The gain in discrimination at the multivariate level was almost never enhanced after aggregating the second replicate. Table (1) gives the mean and median values of the MANOVA F-statistic for the 181 data sets with two replicates.

Table 1: Mean and Median values of MANOVA F

Source	One replicate	Two replicates (averaged)
Mean	6.44	8.97
Median	2.95	3.97

The mean value of the MANOVA F-statistic calculated on the truncated data sets was 6.44. This mean value increased slightly to 8.97 for the full data sets. The median value was 2.95 for the truncated data sets and 3.97 for the full data sets.
The mean values are higher than the median values because of a few excessively high values for the MANOVA F-statistic in some data sets.
We consider that an increase of about 2.53 to an F-statistic at 6.44 should not modify interpretation. This is better investigated in figure (3) showing similar levels of significance for the MANOVA F-statistic in both the truncated data sets and the full data sets. This result suggests that discrimination power is not enhanced at the multivariate level.

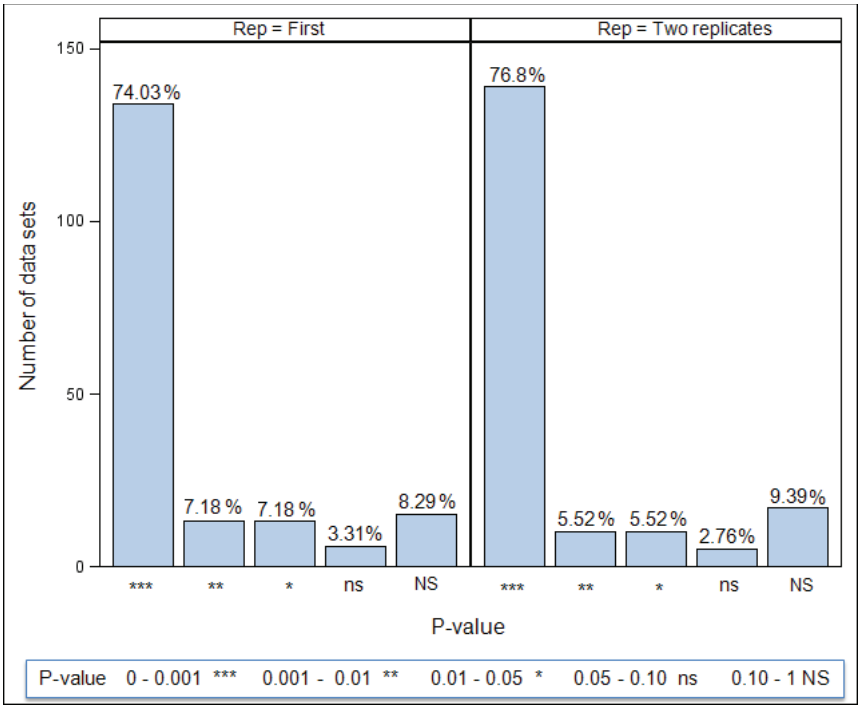


Figure 3: Distribution of P-values for the MANOVA F-statistic

For 86% of the studies, an equal number of significant sensory dimensions (p = 0.10) of the product space was derived from the truncated data set and from the full data set (Fig.4). There was on average a difference in the cumulative proportion of variance explained of 1.25% between the truncated and the full data sets.

For the remaining 14% of studies, the additional significant sensory dimension after aggregating the second replicate did not bring further information on product discrimination, due to a low percentage of additional variance explained. There was on average a difference in the cumulative proportion of variance explained of 4.9% between the truncated and the full data sets.

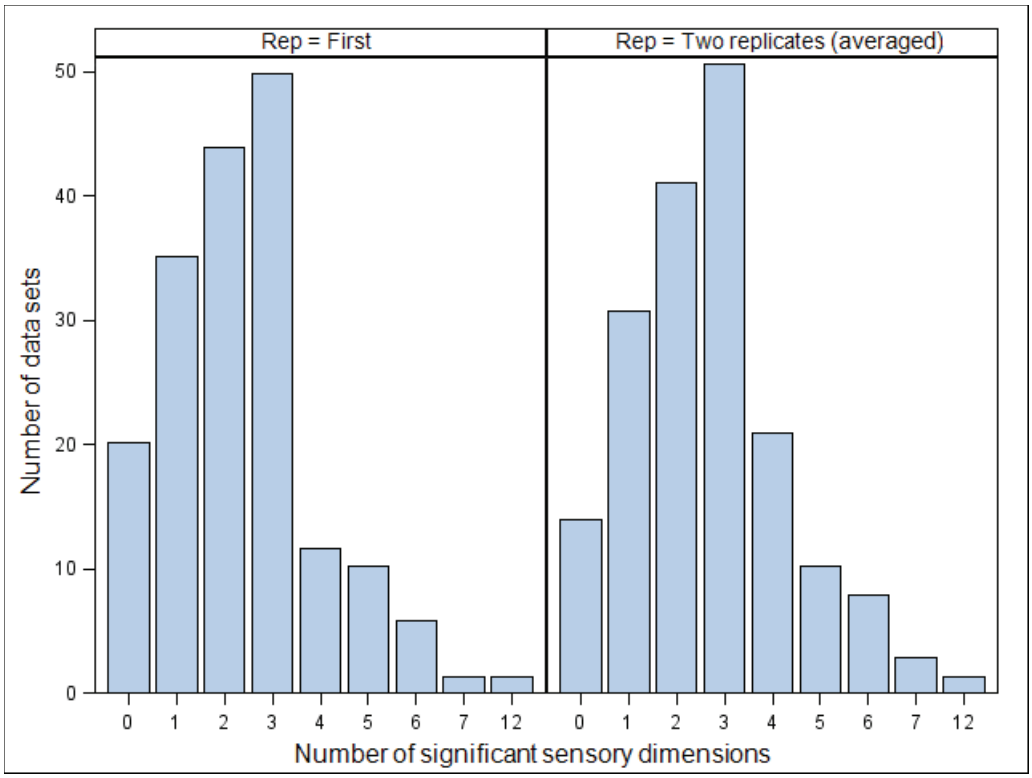


Figure 4: Distribution of the number of significant sensory dimensions

For data sets with three replicates, the results obtained were of the same order. No further enhancement was observed for the MANOVA F-statistic after aggregating the third replicate. Table 2 gives the mean values of the MANOVA F-statistic for the 43 data sets. For 93% of the studies, an equal number of significant sensory dimensions (p=0.10) of the product space was derived from the truncated data sets (first replicate) and from the full data sets (three replicates).

Table 2: Mean values of MANOVA F

Source	One replicate	Two replicates (averaged)	Three replicates (averaged)
Mean	3.86	4.75	4.86

Finally, multivariate analysis of the full and the truncated data sets brought the same results. Thus the slight differences observed in univariate analysis did not impact the multivariate product configurations allowing to draw similar product maps whether replicates are used or not.

4 Conclusion

Both univariate and multivariate analyses confirmed that replication did not generally enhance product discrimination and is therefore superfluous in most cases. However, if products are to be monitored over time or if variability in batches is of interest, then replication becomes necessary.

Sensory laboratories often replicate their measurements during the testing phase at their clients' request. Such companies may take for granted that replication is a good way to obtain reliable results. If panelists are trained to perform sensory evaluation consistently, then why do we need to replicate their measurements? Since they have been trained to perform reliably, their scores will therefore tend to be closer from one replicate to another, generating data with considerably lower variability.

The data suggest that replicates are no longer useful during the testing phase. Replicates are time and budget consuming. Furthermore, panelist availability is an issue that panel leaders often encounter. Replication is necessary in order to check panelist performance during the training phase but serves no further purpose during the testing phase.

References

Bi, H. (2003). Agreement and reliability assessments for performance of sensory descriptive panel. *Journal of Sensory Studies*, 18(1):61–76.

Chambers, D. H., Allison, A. M. A., and Chambers, E. (2004). Training effects on performance of descriptive panelists. *Journal of Sensory Studies*, 19(6):486–499.

Couronne, T. (1997). A study of assessors' performance using graphical methods. *Food Quality and Preference*, 8(5-6):359–365.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, 29(4):751–760.

Gacula, M. (2008). Replication in sensory and consumer testing. In *Viewpoints and Controversies in Sensory Science and Consumer Product Testing*, pages 306–311. Food & Nutrition Press, Inc.

Hunter, E. A. (1996). Experimental design. In Tormod, N. and Einar, R., editors, *Data Handling in Science and Technology*, volume Volume 16, pages 37–69. Elsevier.

Lawless, H. T. and Heymann, H. (2010). Sensory evaluation of food. pages 433 – 449. Food Science Text Series.

McEwan, J. A., Hunter, E. A., van Gemert, L. J., and Lea, P. (2002). Proficiency testing for sensory profile panels: measuring panel performance. *Food Quality and Preference*, 13(3):181–190.

Monrozier, R. and Danzart, M. (2001). A quality measurement for sensory profile analysis - the contribution of extended cross-validation and resampling techniques. *Food Quality and Preference*, 12(5-7):393–406.

Moskowitz, H. (2008). Replication in sensory and consumer testing. In *Viewpoints and Controversies in Sensory Science and Consumer Product Testing*, pages 299–302. Food & Nutrition Press, Inc.

Moskowitz, H., Muñoz, A., and Gacula, M. (2008). Replication in sensory and consumer testing. In *Viewpoints and Controversies in Sensory Science and Consumer Product Testing*, pages 299–311. Food & Nutrition Press, Inc.

Muñoz, A. (2008). Replication in sensory and consumer testing. In *Viewpoints and Controversies in Sensory Science and Consumer Product Testing*, pages 302 – 306. Food & Nutrition Press, Inc.

Pineau, N., Chabanet, C., and Schlich, P. (2007). Modeling the evolution of the performance of a sensory panel: A mixed-model and control chart approach. *Journal of Sensory Studies*, 22(2):212–241.

Porcherot, C. and Schlich, P. (2000). Flash table and canonical mapping of potato varieties. *Food Quality and Preference*, 11(1-2):163–165.

Schlich, P. (1994). Grapes: a method and a sas® program for graphical representations of assessor performances. *Journal of Sensory Studies*, 9(2):157–169.

Schlich, P. (2004). L'analyse en variables canoniques de données de profils sensoriels. Rennes, France. 8eme journees Europeennes Agro-industrie et Methodes Statistiques.

Stone, H. and Sidel, J. L. (2004). Test strategy and the design of experiments. In technology, F. s. a., editor, *Sensory Evaluation Practices*, pages 116–117. Elsevier Academic Press, USA, third edition.

5 Synthèse

Les méthodes par rééchantillonnage et par méta-analyse nous ont permis de mettre en exergue le domaine de variation de la taille de l'effet et du désaccord entre les sujets en profil sensoriel. Par ailleurs, nous savons que selon le type des descripteurs, ces deux mesures varient d'une manière significative.

L'abaque sur le nombre de sujets préconise pour un désaccord moyen entre les sujets σ égal à 0.096, un risque α de 5%, un risque β à 10%, 32 sujets pour détecter un petit effet global en profil sensoriel, 12 sujets pour détecter un effet global moyen et 5 sujets pour détecter un grand effet global.

Le choix du nombre de sujets pourrait être établi en fonction de la nature du profil sensoriel mené. Si l'étude consiste étudier des descripteurs de type texture, on pourrait admettre que la taille de l'effet attendue serait moyenne. Par contre, si l'étude consiste à étudier des descripteurs de type arôme ou odeur, on pourrait admettre que la taille de l'effet attendue serait petite.

Statistiquement, toute augmentation du nombre de sujets ou du nombre de répétitions augmenterait la puissance du test à mettre en évidence des différences entre produits. Cependant, Les résultats de l'étude de la nécessité des répétitions en profil sensoriel (section 4) montrent que lorsque les sujets sont entraînés, une seule répétition en phase de mesure est suffisante pour établir des différences entre produits.

Par ailleurs, l'entraînement des panels est coûteux en termes de temps et d'argent. Ainsi, enrôler un grand nombre de sujets en profil sensoriel dans le but d'avoir plus d'information n'est peut être pas toujours le meilleur choix à faire. Un autre aspect non négligeable est le plan d'expérience. Il est évident que dans le cas de petits panels, il pourrait y avoir un problème quant à assurer un équilibre des effets d'ordre et de report.

Le lien entre les répétitions et le nombre de sujets nous amène à nous interroger sur quel plan s'appuyer pour réaliser une étude de profil sensoriel puissante avec un coût raisonnable (coût des répétitions versus coût de l'entraînement). Une étude sur des cas concrets serait intéressante à mener pour mesurer le coût de deux stratégies. Par exemple, on pourrait comparer les résultats de deux études : une étude avec un panel de 10 sujets effectuant deux répétitions et une étude avec un panel de 20 sujets effectuant une seule répétition.

Pour le traitement de l'ensemble des données, des macros ont été développées en utilisant le logiciel SAS©. L'extraction des données a été effectué avec le langage d'interrogation de données SQL. L'ensemble des calculs intensifs ont été réalisés grâce aux clusters mis à disposition par le Centre de calcul de l'Université de Bourgogne.

Chapitre 5

Discussion générale & Conclusion

1 Synthèse et discussion des résultats issus de ce travail

Ce travail de thèse nous a permis de mettre en exergue les paramètres qui conditionnent le nombre de sujets dans les tests hédoniques et en profils sensoriels. L'objectif principal était de quantifier ces paramètres, de préciser leur domaine de variation et d'identifier les variables dont ils pourraient dépendre. Le but est ensuite de pouvoir adresser des recommandations sur le nombre de sujets à enrôler pour les deux types d'études.

L'originalité de ce travail réside dans l'apport des bases de données pour l'étude et l'estimation de ces paramètres. La littérature a souvent relevé le manque de puissance quant aux résultats apportés par certaines études réalisées sur une dizaine de jeux de données. Les données recueillies par les bases de données PrefBase et SensoBase constituent un "retour d'expériences" et confèrent une meilleure précision aux calculs et aux résultats obtenus dans cette thèse.

1.1 Le nombre de sujets en tests hédoniques

L'approche par rééchantillonnage sur les 7 études hédoniques réalisées pour l'expérimentation de la nouvelle norme AFNOR (2009) a mis en exergue une hypothèse principale. Le facteur qui semble le plus conditionner la taille de panel n'est pas l'hétérogénéité des préférences qui exprime le désaccord entre les sujets, mais principalement la taille des différences d'appréciation entre produits. Ces différences sont liées à la complexité sensorielle du produit.

Par ailleurs, la méta-analyse de PrefBase a conforté ce résultat. Le calcul du nombre de sujets a été réalisé par l'approche analytique en fonction des estimations de la taille de l'effet et de l'hétérogénéité des préférences sur l'ensemble des études.

L'hétérogénéité des préférences est estimée par la racine carrée du carré moyen de l'erreur (RMSE) issu du modèle ANOVA à deux facteurs Sujet + Produit. La valeur moyenne est de 0.20 sur une échelle 0-1, soit un écart de 2 points sur le échelle de 0 à 10. Cette valeur correspond à la valeur moyenne obtenue par Hough et al. (2006) à partir de l'étude de 108 jeux de données.

La taille de l'effet est exprimée par l'écart moyen des moyennes des notes d'appréciation des produits par rapport à la moyenne globale d'appréciation obtenue pour ces produits.
Cette proposition a été adoptée pour permettre aux analystes sensoriels de situer la taille de l'effet attendue pour leurs produits lorsqu'ils ne peuvent pas connaître à priori la position qu'auraient ces derniers, tel que le préconise la norme AFNOR (2009).

Pour la valeur médiane de la taille de l'effet de 0.06 (observée sur une échelle 0-1), une hétérogénéité des préférences de 0.20, un risque α de 5% et un risque β de 10%, le calcul analytique donne une recommandation de 41 sujets. Il faudrait au minimum 41 sujets pour détecter un écart moyen d'appréciation de 0.6 sur une échelle 0 à 10.

La taille des différences à mettre en évidence influe beaucoup sur le nombre minimum de sujets. Par exemple, pour une variabilité résiduelle (ou expérimentale) $\sigma = 0.20$, un risque $\alpha = 5\%$ et un risque $\beta = 10\%$, le nombre de sujets diminue de 74% lorsqu'on passe d'un écart moyen d'appréciation d à mettre en évidence de 0.03 à 0.06 sur une échelle 0-1.

D'autre part, nous avons constaté que la taille des différences d'appréciation peut être induite par une complexité sensorielle des produits testés. En effet, on suppose que la recommandation obtenue sur le nombre de sujets par l'approche rééchantillonnage pour l'étude "Hareng" serait en partie dûe à la complexité sensorielle du produit. Une complexité sensorielle qui se manifeste par une variation de la gamme sur trois dimensions (le goût salé, la nature du fumage, produit avec ou sans arêtes).
La complexité sensorielle est un facteur de segmentation des préférences qui se traduit par une très petite taille de l'effet produit. Nous pouvons admettre qu'un produit simple sensoriellement peut engendrer des groupes de préférences. Cependant, un produit complexe engendrerait a priori encore plus de segments de préférences. L'étendue des différences d'acceptabilité serait donc liée à la complexité du produit.
Cette information a priori pourrait aider à situer la taille de l'effet attendue et ainsi le nombre de sujets à enrôler. Plus le nombre de dimensions sur lesquelles reposent les différences organoleptiques entre produits est important et plus on s'attend à ce que le nombre de sujets à interroger soit grand.

Afin de pouvoir adresser des recommandations sur le nombre de sujets, nous avons choisi trois valeurs de l'effet et de la variabilité. Ces valeurs correspondent aux mesures de position des distributions de l'effet et de la variabilité observé sur PrefBase. Cet abaque permettrait ainsi de

situer le nombre de sujets à interroger selon trois catégories d'effet faible, moyen ou fort.

1.2 Le nombre de sujets en profil sensoriel

La détermination du nombre de sujets en profils sensoriels a été un peu plus complexe à traiter de part la complexité de la méthodologie elle même.

L'approche par rééchantillonnage montre qu'en moyenne 10 sujets sont nécessaires pour une étude de profil. Cependant, cette recommandation conduit, dans l'ensemble, à une réduction en moyenne de 2 à 3 sujets du panel complet de chaque étude.

Par ailleurs, nous avons constaté que le type de descripteur influe sur le nombre de sujets. En effet, les descripteurs de type apparence et texture sont les moins exigeants. Ce résultat pourrait être dû à la facilité que les sujets ont à discriminer les produits avec les descripteurs d'apparence et de texture. En revanche, les descripteurs de type arôme, odeur et saveur, sont difficile à appréhender. Ce qui pourrait expliquer les recommandations sur le nombre de sujets plus élevées pour ce type de descripteurs.
Au niveau multidimentionnel, la recommandation sur le nombre de sujets est en moyenne de 10 sujets. Cependant lorsque le pouvoir discriminant multidimentionnel (FMANOVA) est élevé, la recommandation peut alors descendre jusqu'à 2 sujets. De plus, le pouvoir discriminant au niveau multidimentionnel traduit souvent un pouvoir discriminant au niveau unidimensionnel. Plus les descripteurs sont discriminants, caractérisés par une statistique F et un coefficient de discrimination ρ importants, plus on a de chance d'obtenir une forte valeur de la statistique F de MANOVA et ainsi une meilleure discrimination multidimensionnelle.

L'approche par méta-analyse consolide les résultats de l'approche par rééchantillonnage. Le type de descripteurs influe sur les valeurs du désaccord entre les sujets exprimé par le RMSE mais aussi sur les tailles de l'effet produit exprimées par σ_{prod}.
Le désaccord entre les sujets est en moyenne de 0.10 sur une échelle 0-1, soit un désaccord de 1 point sur une échelle de 0 à 10. Cette valeur est bien inférieure à celle observée en tests hédoniques. Cette baisse est conférée aux séances d'entraînement préconisées par la méthode.
De plus, les descripteurs non significatifs sont souvent caractérisés par une valeur du désaccord supérieure à la moyenne, de l'ordre de 0.14. Ils sont pour la plupart de type odeur et arôme. Plus particulièrement, ce constat a été fait sur des profils de la famille "boissons alcoolisées". Ce résultat pourrait être attribué à une différence de perception de l'alcool (éthanol) et à la qualification de son goût qui n'est souvent pas clairement identifiée.

D'autre part, la taille de l'effet produit observé est en moyenne de 0.07 sur une échelle 0-1. Cette taille de l'effet produit σ_{prod} est presque similaire à celui observé en tests hédoniques et traduit une dispersion moyenne des scores pour les deux types d'études sensorielles.

En outre, la taille de l'effet varie en fonction du type de descripteurs. Les descripteurs de type apparence et texture, plus discriminants, donnent lieu à une taille d'effet plus importante.

Le calcul analytique suggère pour la valeur médiane de la taille de l'effet de 0.054 sur une échelle 0-1, une valeur médiane du désaccord entre les sujets de 0.096, un risque α de 5% et un risque β de 10%, un nombre minimum de 12 sujets !

Enfin, dans la même thématique, nous avons examiné la nécessité des répétitions en profil sensoriel en phase de mesure. Les résultats suggèrent qu'une fois les sujets suffisamment entraînés, les répétitions ne sont plus nécessaires. L'apport en terme de discrimination des produits est jugé faible aussi bien au niveau unidimensionnel qu'au niveau multidimentionnel.

2 Implications, limites & perspectives

Par l'approche bases de données, ce travail de thèse a permis d'aborder la question du nombre de sujets dans les panels d'analyse sensorielle. La puissance des tests sensoriels à mettre en évidence des différences entre produits dépend du nombre de sujets interrogés et détermine en partie le coût des expérimentations. Nos travaux ont permis d'apporter des éléments de réponses sur les paramètres qui déterminent ce nombre de sujets nécessaire. Ils ont également permis de quantifier les tailles de l'effet et aussi le désaccord entre les sujets pour les tests hédoniques et les tests de profils sensoriels. On ne peut adresser une seule recommandation valable pour tous les tests tel qu'il est préconisé par les normes. Le nombre de sujets pour un test sensoriel dépend de ses caractéristiques et de son objectif.

En perspective pour les tests hédoniques, des études supplémentaires pourraient être effectuées pour étudier plus pertinemment le lien entre la complexité sensorielle et l'hétérogénéité des préférences. L'objectif serait de vérifier en fonction de la complexité, l'effet de la taille de panel sur les résultats de préférence vis-a-vis des produits. L'étude pourrait consister à concevoir deux systèmes sensoriels : simple versus complexe. Un système où une gamme de produits varierait sur deux dimensions sensorielles (par exemple : arôme, saveur) et un autre système où cette gamme varierait sur trois dimensions (arôme, saveur et texture). Il s'agira ensuite d'étudier l'impact d'une telle variation de la complexité sur les préférences des sujets.
Nous avons abordé jusqu'à présent les tests hédoniques dans un contexte où la segmentation des préférences n'est pas recherchée. Il serait intéressant d'examiner par l'approche par rééchantillonnage la question de détection de ce nombre de segments de préférence. Au besoin, il faudrait peut être préconiser une étude préliminaire et prospective. Un traitement des données par un algorithme de mélange pourrait être envisagé.

En profil sensoriel, le coût de l'expérimentation est directement lié aux nombres d'observa-

tions à recueillir pour garantir une puissance du test à mettre en évidence des différences entre produits. Le choix à adopter quant à enrôler plus de sujets ou bien multiplier les réponses par sujet est une question qui mériterait d'être étudiée.
Les résultats sur l'étude de la nécessité des répétitions montrent que les répétitions ne sont pas nécessaires en phase de mesure.
Afin de relier les résultats de l'étude sur les répétitions à ceux de l'approche par rééchantillonnage, des études sur différentes gammes de produits peuvent être envisagées. Ces études auront pour objectif de justifier ou non l'augmentation du nombre de répétitions ou bien l'augmentation du nombre de sujets. Il faudra bien sûr établir pour ces études des plans d'expérience adaptés. Des analyses pourront ensuite être réalisées sur les données recueillies. Une comparaison pourrait être faite entre les résultats obtenus en considérant le panel complet avec les scores d'une seule répétition et les résultats obtenus en considérant un panel réduit avec les scores de deux répétitions.

Bibliographie

AFNOR (1983). *Méthodes d'établissement du profil de la flaveur*, volume NF V 09-016. AFNOR, Paris - La Défense.

AFNOR (1987). *Directives générales pour l'implantation de locaux destinés à l'analyse sensorielle*, volume V 09-105. AFNOR, Paris.

AFNOR (1993). *Guide général pour la sélection, l'entraînement et le contrôle des sujets. Partie 1: Sujets qualifiés*, volume NF ISO 8586-1. AFNOR, Paris - La Défense.

AFNOR (2000). *Méthodologie. Directives générales pour la réalisation d'épreuves hédoniques en laboratoire d'évaluation sensorielle ou en salle en conditions contrôlées impliquant des consommateurs. XP V 09-500*, volume XP V 09-500. AFNOR, Paris - La Défense.

AFNOR (2003). *Analyse sensorielle. Méthodologie. Directives générales pour l'établissement d'un profil sensoriel*, volume V09-007 NF ISO 13299. AFNOR, Paris - La Défense.

AFNOR (2009). *Analyse sensorielle – Méthodologie – Directives générales pour la réalisation d'épreuves hédoniques. XP V 09-500*, volume XP V 09-500. AFNOR, Paris - La Défense.

Basker, D. (1996). Reproducibility of taste panel test results. *Food Quality and Preference*, 7(3-4):345.

Bi, H. (2003). Agreement and reliability assessments for performance of sensory descriptive panel. *Journal of Sensory Studies*, 18(1):61–76.

Boutrolle, I. (2007). *Mesure de l'appreciation des aliments par les consommateurs. État des pratiques et propositions méthodologiques.* Thèse de doctorat.

Bouvenot, G. and Villani, P. (2000). Les essais cliniques d'équivalence en rhumatologie. *Revue du Rhumatisme*, 67(8):569–572.

BS (1986). British standard bs 5929.

Buck, D. (2003). Solicited commentary to garber et al: Measuring consumer response to food products. *Food Quality and Preference*, 14(1):37–38.

Callier, P. and Schlich, P. (1997). La cartographie des préférences incomplètes. validation par simulation. *Sciences des Aliments*, 17(2):155–172.

Cardello, A. V. and Schutz, H. G. (2003). The importance of taste and other product factors to consumer interest in nutraceutical products: Civilian and military comparisons. *Journal of Food Science*, 68(4):1519–1524.

Chambers, D. H., Allison, A. M. A., and Chambers, E. (2004). Training effects on performance of descriptive panelists. *Journal of Sensory Studies*, 19(6):486–499.

Chambers, E., Bowers, J. A., and Dayton, A. D. (1981). Statistical designs and panel training/experience for sensory analysis. *Journal of Food Science*, 46(6):1902–1906.

Chambers, E. and Wolf, M. B. (1996). Sensory testing methods. pages 9 –10. American society for testing and materials, second edition.

Chiva, M. (1996). Le mangeur et le mangé : la subtile complexité d'une relation fondamentale. In *Identités des mangeurs, images des aliments*, pages 11–30. Polytechnica, Paris.

Cliff, M. A., King, M. C., Scaman, C., and Edwards, B. J. (1997). Evaluation of r-indices for preference testing of apple juices. *Food Quality and Preference*, 8(3):241–246.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* LEA, New York, second edition.

Cook, G. L. and Homer, D. B. (1996). The effect of number of sensory assessments on the accuracy of treatment comparisons in meat quality trials. *Food Quality and Preference*, 7(2):95–99.

Cornell, J. A. (1997). What is the meaning of stability in the mean? *Food Quality and Preference*, 8(4):259–260.

Couronne, T. (1997). A study of assessors' performance using graphical methods. *Food Quality and Preference*, 8(5-6):359–365.

Delarue, J. and Sieffermann, J. (2000). Use of flash profile for a quick sensory characterisation of a sixteen strawberry yoghurts.

Dravnieks, A. (1982). Odor quality: semantically generated multidimensional profiles are stable. *Science*, 218:799–801.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, 29(4):751–760.

Fantino, M. (1992). Etat nutritionnel et perception affective de l'aliment. In Giachetti, I., editor, *Plaisir & préférences alimentaires*, Commission "comportement alimentaire et acceptabilité des aliments", pages 31–48. Polytechnica, Paris.

Faurion, A. (1992). La physiologie de la sensation sucrée.

Gacula, M. (1988). Experimental design and analysis. In Moskowitz, H., editor, *Applied sensory analysis of foods*, volume II, pages 83–140. CRC press, Boca Raton.

Gacula, M. (1993). *Design and analysis of sensory optimization.* Food and Nutrition Press.

Gacula, M. and Rutenbeck, S. (2006). Sample size in consumer test and descriptive analysis. *Journal of Sensory Studies*, 21(2):129–145.

Gacula, M. and Singh, J. (1984). *Statistical methods in food and consumer research*. Food Science and technology. Academic Press, Orlando, Florida.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5(10):3–8.

Gordon, N. M. and Norback, J. P. (1985). Choosing objective measures when using sensory methods for optimization and product positioning. *Food Technology*, 39(11):96–101.

Hedges, L. V. (1986). Issues in meta-analysis. *Review of Research in Education*, 13:353–398.

Hedges, L. V., Cooper, H., and Bushman, B. J. (1992). Testing the null hypothesis in metaanalysis - a comparison of combined probability and confidence-interval procedures. *Psychological Bulletin*, 111(1):188–194.

Hough, G., Wakeling, I., Mucci, A., Chambers, E., Gallardo, I. M., and Alves, L. R. (2006). Number of consumers necessary for sensory acceptability tests. *Food Quality and Preference*, 17(6):522–526.

Hunter, E. A. (1996). Experimental design. In Tormod, N. and Einar, R., editors, *Data Handling in Science and Technology*, volume Volume 16, pages 37–69. Elsevier.

Issanchou, S. and Hossenlopp, J. (1992). Les mesures hédoniques : portées et limites. In *Plaisir et préférences almentaires*, pages 49–76.

Jeltema, M. A. and Southwick, E. W. (1986). Evaluation and applications of odor profiling. *Journal of Sensory Studies*, 1(2):123–136.

Jones, L., Peryam, D., and Thurstone, L. (1955). Development of a scale for measuring soldiers' food preferences. *Food Research*, 20:512–520.

King, B. M., Arents, P., and Moreau, N. (1995). Cost efficiency evaluation of descriptive analysis panels. panel size. *Food Quality and Preference*, 6(4):245–261.

Köster, E. (1998). Les épreuves hédoniques. In SSHA and Depledt, F., editors, *Evaluation sensorielle: Manuel méthodologique*, pages 182–206. Lavoisier, Paris, 2nde edition.

Köster, E. P., Couronne, T., Leon, F., Levy, C., and Marcelino, A. S. (2003). Repeatability in hedonic sensory measurement: a conceptual exploration. *Food Quality and Preference*, 14(2):165–176.

Labbe, D., Schlich, P., Pineau, N., Gilbert, F., and Martin, N. (2009). Temporal dominance of sensations and sensory profiling: A comparative study. *Food Quality and Preference*, 20(3):216–221.

Lawless, H. T. and Heymann, H. (1998). In *Sensory evaluation of food. Principles and practices*, pages 754–782. Chapman and Hall, New York.

Lesschaeve, I. (1997). *Etude des performances de sujets effectuant l'analyse descriptive quantitative de l'odeur ou de l'arôme de produits alimentaires. Recherche de liens entre épreuves de sélection et épreuves de profil.* Thèse de doctorat.

Lewis, J. A. (1999). Statistical principles for clinical trials (ich e9) an introductory note on an international guideline. *Statistics in Medicine*, 18(15):1903–1904.

Little, A. D. (1950). Flavor profile method. Technical report, Arthur D. Little, Inc. gen47.

Lundgren, B., Karlstrom, B., Torrang-Lindbom, G., Andersson, Y., and Clapperton, J. (1991). Extruded wheat flour: Flavour and texture–comparison of evaluations by two laboratories. *Food Quality and Preference*, 3(1):1–12.

Mac Leod, P. (1992). L'analyse sensorielle : aspects neuro-physiologiques.

Mac Leod, P. (1998). Les caractéristques d'une réponse sensorielle. In *Evaluation sensorielle. Manuel méthodologique*. Lavoisier, Paris, 2nde edition.

Mac Leod, P. and Sauvageot, F. (1986). *Bases neurophysiologiques de l'évaluation sensorielle des produits alimentaires*, volume 5. Lavoisier, Paris.

MacFie, H. (1986). Aspects of experimental design. In Piggott, J., editor, *Statistical procedures in food research*, pages 1–18. Elsevier Applied Science, Barking, Essex.

MacFie, H. J., Bratchell, N., Greenhoff, K., and Vallis, L. V. (1989). Designs to balance the effect of order presentation and first-order carry-over effects in hall tests. *Journal of Sensory Studies*, 4(2):129–148.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–748.

Marchant, L. F. and McGrew, W. C. (1991). Laterality of function in apes: a meta-analysis of methods. *Journal of Human Evolution*, 21(6):425–438.

Martin, N., Molimard, P., Spinnler, H. E., and Schlich, P. (2000). Comparison of odour sensory profiles performed by two independent trained panels following the same descriptive analysis procedures. *Food Quality and Preference*, 11(6):487–495.

McEwan, J. A. (1997). Base size in product testing: A psychophysical viewpoint and analysis - reply. *Food Quality and Preference*, 8(4):257–258.

McEwan, J. A., Hunter, E. A., van Gemert, L. J., and Lea, P. (2002). Proficiency testing for sensory profile panels: measuring panel performance. *Food Quality and Preference*, 13(3):181–190.

Meilgaard, M., Carr, T., and Civille, G. V. (1999). Affective tests: Consumer tests and in-house panel acceptance tests. In *Sensory Evaluation Techniques*, pages 231–263. CRC Press, third edition.

Molnar, P. (1989). *Results of a sensory collaborative test of some food products*, volume 18.

Moskowitz, H. R. (1997). Base size in product testing: A psychophysical viewpoint and analysis. *Food Quality and Preference*, 8(4):247–255.

Muir, D. D. and Banks, J. M. (1993). Sensory evalution of cheddar. *Dairy Industries International*, 58(4):47–50.

Muñoz, A. and Civille, G. (1992). The spectrum descriptive analysis method. In Hootman, R., editor, *Descripting analysis testing*, volume 13, pages 22–34. ASTM.

Narçon, S. (2001). *Caracterisation des perceptions thermiques en regime transitoire - Contribution a l'etude de l'influence des interactions sensorielles sur le confort.* Thèse de doctorat.

Nicod, H. (1998). L'organisation pratique de la mesure sensorielle : 1. les sujets. In *Evaluation sensorielle. Manuel méthodologique SSHA*, pages 46–50.

Pages, J. and Husson, F. (2001). Inter-laboratory comparison of sensory profiles: methodology and results. *Food Quality and Preference*, 12(5-7):297–309.

Peryam, D. and Pilgrim, F. (1957). Hedonic scale method of measuring food preference. *Food technology*, 11(9):9–14.

Pineau, N. (2006). *La performance en analyse sensorielle. Une approche base de données.* Thèse de doctorat.

Pineau, N., Chabanet, C., and Schlich, P. (2007). Modeling the evolution of the performance of a sensory panel: A mixed-model and control chart approach. *Journal of Sensory Studies*, 22(2):212–241.

Pineau, N., Schlich, P., Cordelle, S., Mathonniere, C., Issanchou, S., Imbert, A., Rogeaux, M., Etievant, P., and Köster, E. (2009). Temporal dominance of sensations: Construction of the tds curves and comparison with time-intensity. *Food Quality and Preference*, 20(6):450–455.

Popper, R., Rosenstock, W., Schraidt, M., and Kroll, B. J. (2004). The effect of attribute questions on overall liking ratings. *Food Quality and Preference*, 15(7-8):853–858.

Randall, E. and Sanjur, D. (1981). Food preferences. their conceptualization and relationship to consumption. *Ecology of food and nutrition*, 11:151–161.

Risvik, E., McEwan, J. A., and Rodbotten, M. (1997). Evaluation of sensory profiling and projective mapping data. *Food Quality and Preference*, 8(1):63–71.

Schlich, P. (1993). Uses of change over designs and repeated measurements in sensory and consumer studies. *Food Quality and Preference*, 4(4):223–235.

Schlich, P. (1994). Grapes: a method and a *sas*© program for graphical representations of assessor performances. *Journal of Sensory Studies*, 9(2):157–169.

Sharp, W. F., Norback, J. P., and Stuiber, D. A. (1986). Using a new measure to define shelf life of fresh whitefish. *Journal of Food Science*, 51(4):936–939.

Shepherd, R., Farleigh, C., and Stockley, L. (1985). The role of attitudes and personality in food

choice. In Diehl, J. and Leitzmann, C., editors, *Measurement and determinants of food habits and food preferences*, volume 7, pages 219–230. Department of Human Nutrition, Agricultural University.

Shepherd, R., Smith, K., and Farleigh, C. A. (1989). The relationship between intensity, hedonic and relative-to-ideal ratings. *Food Quality and Preference*, 1(2):75–80.

Sidel, J. L., Stone, H., and Thomas, H. A. (1994). Hitting the target - sensory and product optimization. *Cereal Foods World*, 39(11):826–830.

SSHA and Depledt, F. (1998). *Evaluation sensorielle. Manuel méthodologique.* Lavoisier, Paris, 2nde edition.

Stone, H., Sidel, J., Oliver, S., Woolsey, A., and Singleton, R. C. (1974). Sensory evaluation by quantitative descriptive analysis. *Food Technology*, November:24–34.

Stone, H. and Sidel, J. L. (2004). Affective testing. In technology, F. s. a., editor, *Sensory Evaluation Practices*, pages 247–277. Elsevier Academic Press, USA, third edition.

Urdapilleta, I., Ton Nu, C., Saint Denis, C., and Huon de Kermadec, F. (2001). *Traité d'évaluation sensorielle. Aspects cognitifs et métrologiques des perceptions.* Dunod, Paris.

Visalli, M., Cordelle, S., and Schlich, P. (2008). Création d'une base de données de préférences alimentaires. Rapport de stage, Centre des Sciences du Goût et de l'Alimentation CSGA.

Wakeling, I. N. and MacFie, H. J. H. (1995). Designing consumer trials balanced for first and higher orders of carry-over effect when only a subset of k samples from t may be tested. *Food Quality and Preference*, 6(4):299–308.

Williams, A. and Langron, S. (1984). The use of free-choice profiling for the evaluation of commercial ports. *Journal of the Science of Food and Agriculture*, 35:558–568.

Williams, E. J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Chemistry*, 2(2):149–168.

Annexes

Annexe1 : schéma structurel de la base de données SensoBase

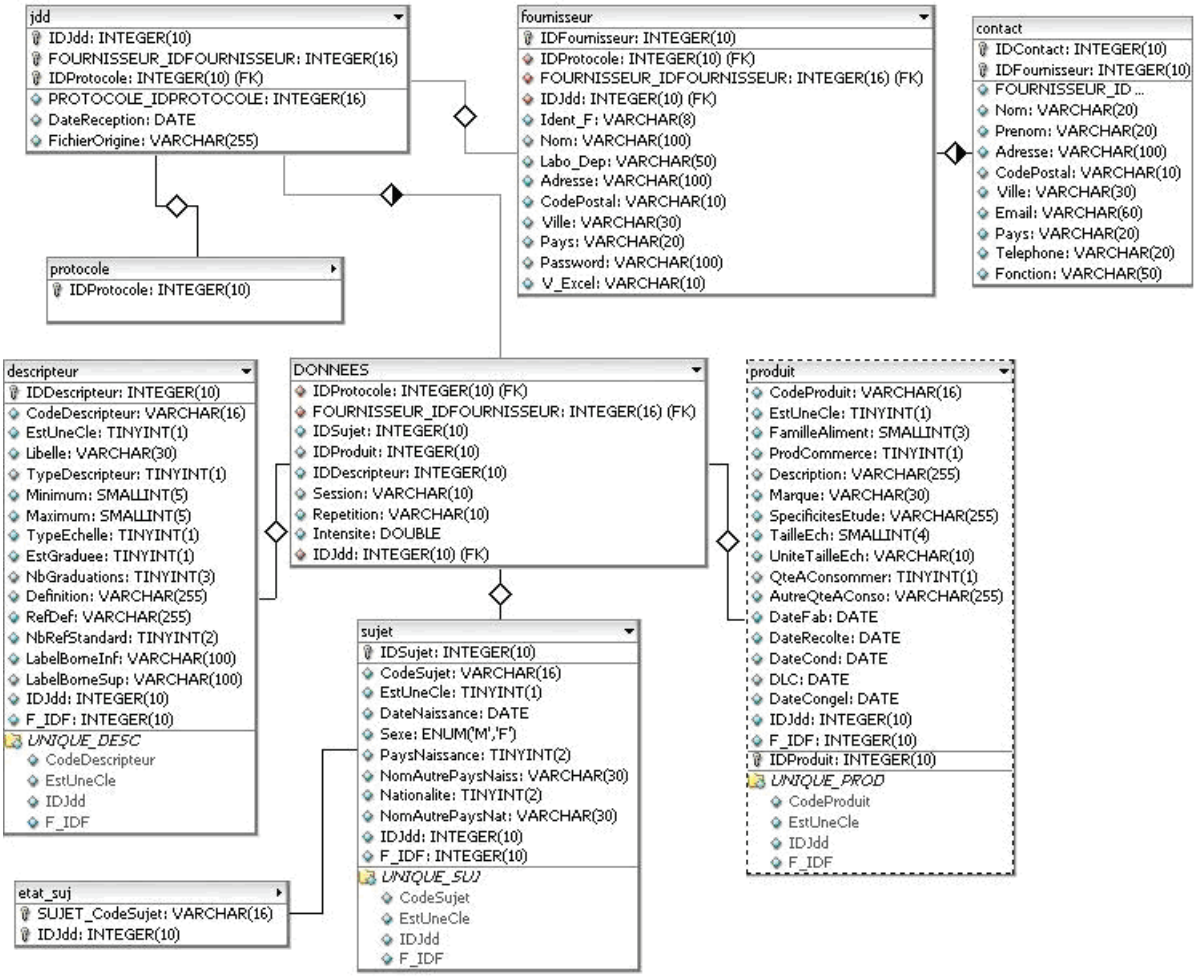


FIG. 5.1 – *Schéma structurel de la base de données SensoBase*

Annexe2 : Schéma structurel de la base de données PrefBase

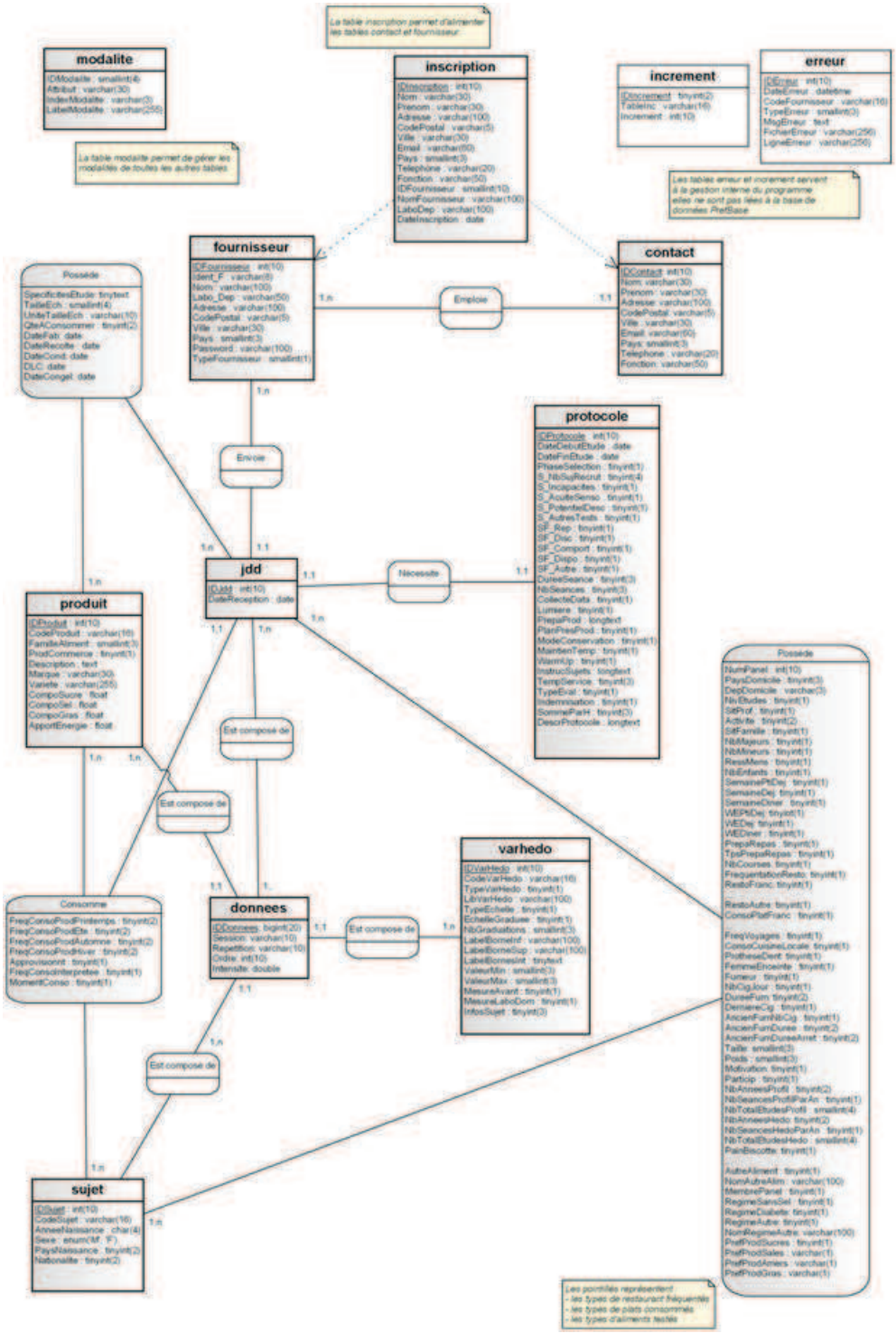


FIG. 5.2 – *Schéma structurel de la base de données PrefBase*