https://www.economie-gestion.com

# Méthodologie des enquêtes et sondages

## Sommaire



Renvoi sur titre Ctrl + click

# SEANCE 1

## I/ INTRODUCTION

En marketing, on distingue trois facettes :

- **Marketing stratégique :**
  - Définition de positionnement, choix de marché, segmentation, stratégie de marque
- **Marketing étude :**
  - « Activités organisées de collecte, de traitement et d'interprétation des données relatives aux marchés et plus précisément aux publics dt dépd l'entreprise»
- **Marketing opérationnelle, tactique :**
  - Métier de chef de produit ; gérer le dvpt du produit, packaging, communication du produit

### OBJECTIFS

A l'aide des infos collectées :

- **Aide à la prise de décision**
  - L'entreprise est confrontées à des évolutions permanentes, internes ou externes (modification prix, produit, concurrents), qui nous amènent à prendre des décisions dans un certain climat d'incertitude
  - *Réduction de l'incertitude liée aux évolutions de l'activité interne et de l'environnement externe*
    - Orientation stratégique
    - Prise de décision spécifique

- **Contrôle des performances**
  - Les études interviennent en aval (après)
  - Fonction de support à la prise de décisions

La fonction étude permet de mettre en relation les acteurs du marché (distrib, conso, concurrents), va entrainer des décisions stratégiques (marché, segmentation, positionnement) et les décisions tactiques (produit, prix, promotion, distribution)

## ETUDES CONSOMMATEURS

- Motivations et freins vis-à-vis d'une catégorie, d'une marque
- Connaissances, notoriété (=top of mind, notoriété spontanée, notoriété assistée)
- Perceptions, attitudes (prédisposition à réagir positivement ou négativement à l'égard d'un objet, situation)
    - **Composantes cognitives** : croyances subjectives à l'égard d'une marque ou d'un objet
    - **Composantes affectives** : ensemble des sentiments, ressentis à l'égard de…
    - **Composantes conatives** : tendance comportementale d'un individu (futur)

- Critères de décisions : critères pris en compte par le conso
- Satisfaction : la fidélité garantie la pérennité d'une entreprise sur le long terme

## DECISIONS STRATEGIQUES

- Définition du marché : définir son champ concurrentiel
    - Le marché ne se définit pas seulement par un type de produit mais également par un usage et un type d'individu

    - Définition qualitative : combinaison de produits, d'usages, d'individus
    - Définition quantitative dans le tps et l'espace ; estimation de potentiel

- Segmentation de marché
    - Identification de groupes homogènes en terme de comportements, perceptions, attitudes
    - Critères socio démographiques, psycho graphique, comportementaux, géographique

- Positionnement
    - Evaluation d'une marque par rapport à ses concurrents
    - Identification des critères sur la base desquels les produits peuvent être différenciés et évaluation des marques sur la base de ces critères

## DECISIONS TACTIQUES

- Lancement du nouveau produit / modification d'un élémrnt du mix d'un produit existant
    - Tests de concept
    - Tests de produits
    - Tests de packaging : visible en linéaire ? véhicule les valeurs du produit ?
    - Tests de noms : Facile à retenir ?
    - Tests de prix : qualitativement apprécié ?
    - Distribution : choix de canal, optimisation de linéaire
    - Promotion : montant à allouer, choix d'une campagne publicitaire, choix des médias

- Pré-tests (étude en amont)
    - Elaboration des projets : Identification des motivations, d'idées créatives
    - Evaluation des différents projets : acceptabilité des projets de campagne par les clients, les consos

- Post-tests :
    - Evaluation de l'action engagée :
        - Mesure d'impact et d'efficacité : rappel spontané/assisté, mémorisation, compréhension, agrément, attitude marque, intention d'achat

Autres études

- Etude Shoppers (études sur les points de vente) -> acte d'achat
    - Etude d'impact de la communication sur le point de vent ou des catalogues/prospectus
    - Diagnostic et analyse d'implantation merchandising, d'optimisation de linéaire
    - Analyse des comportements d'achats

- Etude Corporate (Etude sur la notoriété ou l'image d'une E)
    - Important pour les enseignes de luxe, egalement si B to B

- Etude sur les médias (audience)

Les études nous permettent de prendre les meilleures décisions pour un meilleur développement des ventes etc. Mais on ne peut pas en faire tout le temps, il faut accepter la prise de risque car toutes les décisions ne sont toujours pas bonnes

- Arbitrage nécessaire entre
    - La richesse des infos
    - Le budget (coute cher)
    - Les délais (étude prend du temps à réaliser et qq fois la prise de décision doit être rapide)

Décider de réaliser une étude dépend de 3 paramètres :

- Ai-je le temps ?
- Les ressources financières ?
- Les ressources humaines ?

[…]pour pouvoir obtenir les informations pour telle prise de décision spécifique. Enjeux important sur le plan de la mobilisation des ressources (R&D, capital humain). Lancement des

produits a un impact sur l'image de la marque. Dans le mix, communication coûteuse notamment création de films publicitaires. Réalisation des études en interne ou en externe/ sous-traitance

- Jusqu'à il y a une vingtaine d'année, département spécial pour les études. Today, la sous-traitance est prépondérante.
- Marché des études : $40 billion en 2014, France en 4ème position de producteur d'étude au niveau mondial avec 2 milliards d'euros.
- CA total des études : grande consommation occupe la 1ère place

*Les principaux groupes d'études au niveau mondial*

| | PDM |
|---|---|
| Nielsen | 14% |
| Kantar | 10% |
| IMS Health Inc | 6% |
| Ipsos | 6% |
| Gfk | 5% |

## II/ LES SOURCES D'INFORMATION

On distingue les données primaires et secondaires.

- Données primaires : créées par l'entreprise dans le cadre d'une étude

- **Données secondaires**

*Disponible avant l'étude*

| Données internes (base de données) | Données externes |
|---|---|
| - Données marketing et commerciales (CA, promotions)<br>- Rapport commerciaux (important en B to B, commercial en contact avec le client donc accède à la satisfaction client)<br>- Bases de données clients (info à exploiter)<br>- Réclamations clients (bonne indicateur de ce qui ne va pas)<br>- Etude marketing antérieures | - Organismes publics et para-publics ; organismes internationaux (INSEE, credoc)<br>- Associations, syndicats professionnels<br>- Presse professionnelles (nouveau produit lancé par les concurrents); Salon<br>- Société d'études (données standardisées) |

Les études sont classées selon :

- **ETUDES REGULIERES ET PARTAGEES**
  - **Omnibus**

    Périodicité hebdomadaire, mensuelle, par questionnaire. L'E peut acheter quelques questions (résultat confidentiels)
    Achat d'une ou de plusieurs questions en propre
    Etude multi-souscription ; données confidentielles
    *couts répartis sur plusieurs entreprises, moins chers. Valable que si quelques questions nous intéressent*

  - **Panels**

    (panel conso : taux de pénétration [% foyer qui achète une marque) et disrib : tt ce qui est vendu par le distrib, prix moyen, promotion )
    Périodicité hebdomadaire, mensuelle, trimestrielle
    Nombreuse mesures standardisées
    Données partagées

  - **Baromètres**

    (panels allégés, notoriété, image)
    Quelques mesures standardisées
    Etude multi-souscription, données partagées

- **ETUDES UNIQUE ET EXCLUSIVES**
- 
  - **Etudes Ad hoc**
    Méthodologie spécifique

  - **Marché-tests**
    Lancement produit
    Produits mis en vente dans un marché réel, ne sont pas interrogés directement, petite zone géographique + plan de communication. Voir les effets réel du lancement de produits.
    Sur un marché simulé, reconstitution d'un magasin, dédié à des études, et on pose des questions aux conso
    Méthodologie standardisée
    Etude de marché pour des marchés réels, simulés

## III/ Les types d'études

*Les études peuvent être classées selon les <u>buts poursuivis</u>*

### Exploratoires

- Explorer un problème vague afin de définir des hypothèses et cerner la problématique.

### Descriptifs

- Décrire un phénomène
- Visualiser une situation (mapping)
- Classer, catégoriser, segmenter (décrire les usages dans un domaine, un phénomèn)

### Explicatifs

- Expliquer des relations causales entre variables indépendantes et variables à expliquer
- Prédire

X-> Y, X eplique Y. X : variable antécédantes, indépendant. Y variables expliquées.

Selon *<u>les modes de collecte</u>* des données

| L'interrogation | Observation |
|---|---|
| • Collecte qui repose sur le déclaratif de la personne interrogée<br>• *inconvénients:bcp de biais possible, car l'étude repose sur la fiabilité de la personne* | • Collecte qui repose sur des données factuelles (comportement réel de la personne)<br>• *Inconvénients : champ d'investigation difficle, pas possble de relever le perceptuel : attitude, freins, motivation etc.* |

## *L'étude documentaire :*

consiste à recenser et analyser les informations provenant de données secondaires (desk research), objectif exploratoire.

- **Démarche à suivre de manière scrupuleuse :**
  - Identification des sources existantes
  - Analyse
  - Synthèse
- **Intérêts de l'étude**
  - Accès moins couteux qu'une étude spécifique
  - Compréhension de l'environnement
  - Préparation de la recherche ultérieure
    - Meilleure appréhension du périmètre de l'étude
    - Choix d'axes
    - Elaboration d'hypothèses de recherche
- **Limites de l'étude**
  - Inadéquation des données à la problématique
  - Surabondance des données ; données contradictoires
  - Imprécision sur la qualité des données

## *Les études qualitatives* :

visent à qualifier en profondeur un phénomène, à identifier toutes les composantes, objectif exploratoire, souvent sur un petit échantillon, recueil ouvert (non structuré et non directif). Informations profondes et riches sur peu d'individus interrogés, posture d'ouverture et de découverte

- **Approches**
  - Cliniques (approche psychologiques) : entretiens individuels, entretiens de groupes, tests projectifs
  - Ethnographiques et cognitives stimulatives : observation
- **Applications**
  - Freins et motivations vis-à-vis d'une catégorie ou d'une marque
  - Processus de choix du conso
  - Perception de choix du conso
  - Recherche d'idées créatives / pré-tests de concepts, de noms, de packaging, de promotions, de publicité
  - Préparation d'une enquête quantitative

    $

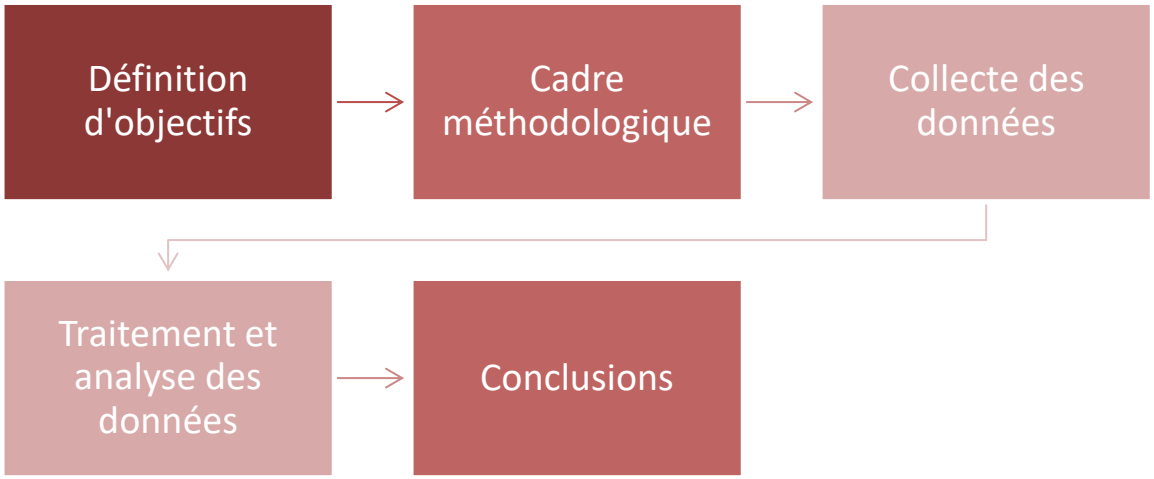
## *Les études quantitatives* :

visent à quantifier un phénomène, à avoir une vision représentative. Objectifs descriptifs et explicatifs. Sur un grand échantillon, recueil structuré. Beaucoup d'individus interrogés sur un nombre d'info relativt limité. Pour ressortir un résultat que l'on espère représentatif de la population que l'on a étudié.

- **Approches**
  - Sondages
  - Analyses de données de panels ; issues de bases de données clients
  - Etudes expérimentales
- Applications
  - Usages et attitudes
  - Segmentation ; positionnement
  - Bilan d'image de marque
  - Etude de satisfaction
  - Etudes sur les éléments du mix
    - Tests de concept, de produit (comparatifs, monadiques, monadiques séquentiels), de prix, de packaging (déclaratif, approche expérimentale)
    - Pré-tests et post-tests publicitaires, promotionnels
  - Etudes d'optimisation du mix

| Type d'étude | Facteurs de variabilité | Coût (HT) |
|---|---|---|
| **Etude quantitative**<br>Usages et attitudes<br>(base : 500 questionnaires de 40 min., à domicile) | Taille de l'échantillon,<br>Nature de la population,<br>Mode d'administration<br>Longueur questionnaire, | Environ 40000 € |
| **Etude qualitative**<br>Entretien de groupe<br>(base 8-10 personnes) | Nombre participants, statut des participants, durée | Environ 4000-5000 € par groupe |
| **Etude qualitative**<br>Entretien semi-directif | Durée entretien | Environ 500-700 € par entretien |

## IV/ LE PROCESSUS D'ETUDE




### 1. *Définition des objectifs : essentielle à la réussite du projet*

- Analyse du problème
    - Identifier le problème managérial qui se pose à l'E (décision)
    - Situer le problème dans son contexte (environnt, marché)
    - Collecter les info dispos
    - Evaluer l'enjeu et les contraintes (délai, budget)
- Définition des objectifs de l'étude
    - Transformer le pb managérial en objectifs d'études
    - Fixer des priorités

➔ **Définition des besoins en info et définir la pop d'étude**

Besoin en informations : ensemble des infos qui permettent de répondre aux questions de l'étude. Liste détaillée de tous les comportements que l'on veut étudier. Construction du questionnaire à partir de cette ligne.

Population cible : population que je veux interroger. Ensemble des éléments possédant les informations permettant de répondre aux questions de l'étude. On la définit par rapport :

- Nature des éléments (personnes, entreprise)
    - Individus / foyer ? acheteurs/conso ? -> qui nous intéresse ?
    - Abonnés/ acheteurs/ lecteurs ?
- Périmètre géographique, critères temporels
- Caractéristiques (socio-démographique)

*EXEMPLE*

*PRATIQUE MANAGERIALE : Un éditeur de e-learning spécialisé dans les langues étrangères se demande s'il doit améliorer la qualité de son site*

*Objectifs d'études (qu'est ce que je veux étudier pour la prise de décision) : Quels aspects privilégier pour améliorer la satisfaction globale ? Ya t'il des profits utilisateurs +/- satisfaits que d'autres ?*

*BESOIN INFO et POP CIBLE : Identification des critères de satisfaction pr les personnes qui utilisent le logiciel. Evaluation de la satisfaction globale et par critère. Caractéristiques socio-démographiques, profils d'achat. Clients actuels, anciens clients récents*

La définition des objectifs et une étape primordiale. Car après cette étape, on ne fait que appliquer la technique. Il ne faut pas se perdre dans une recherche couteuse inutiles, les objectifs conditionnent les chox méthodologique.

- Pour juger de la qualité d'une étude, il faut se demander si les conclusions de l'étude permette la prise de décision. Si erreur à cette étape, le reste de l'étude sera erronée.

## 2. *Le cadre méthodologique*

- Choix de la méthode d'étude (grand champs méthodologique)
    - Etude qualitative, quantitative par sondage, expérimentale
- Choix du mode de recueil des données (enquête internet, face to face, postale, par téléphone ?) ; dépend des objectifs poursuivis, de nos contraintes budgétaires)
- Elaboration des instruments de mesure (outils dans lequel on va transformer nos besoins en info en question dans le cadre d'un sondage)
    - Questionnaire, grille d'observation, guide d'entretien
- Construction du plan d'échantillonnage

## *3. Collecte des données*

- Formation des enquêteurs
- Collecte des données sur le terrain
- Contrôle du travail effectué par les enquêteurs

## *4. Traitement et analyse*

- Codification, analyse (statistiquement) et interprétation des données (résultat statistiques)
- Objectifs : comprendre quel type de test utilisé pour rechercher ce que je veux analyser ?

Logiciel SPSS : référence mondial en traitement de donnée statistique.

5. Conclusions

- Faits majeurs et recommandations par rapport à la pbatique

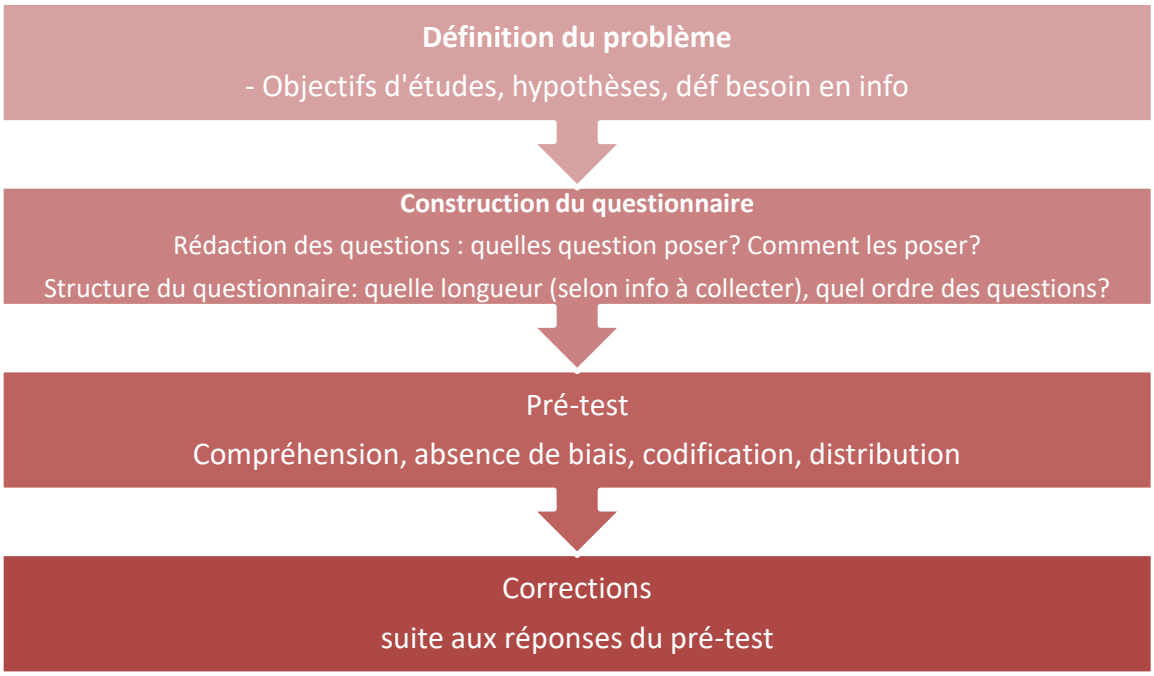
# Chapitre : Enquêtes par questionnaire

Sondage : étude des caractéristique d'une population à partir d'un échantillon qui en est issu

- Elaboration du questionnaire
- Méthode de collecte
- Techniques d'échantillonnage

Le questionnaire est l'instrument utilisé afin de recueillir des infos facilement traitables à grande échelle

Définition du problème : Objectifs d'études, hypothèses, définition des besoins en info




La rédaction des questions :

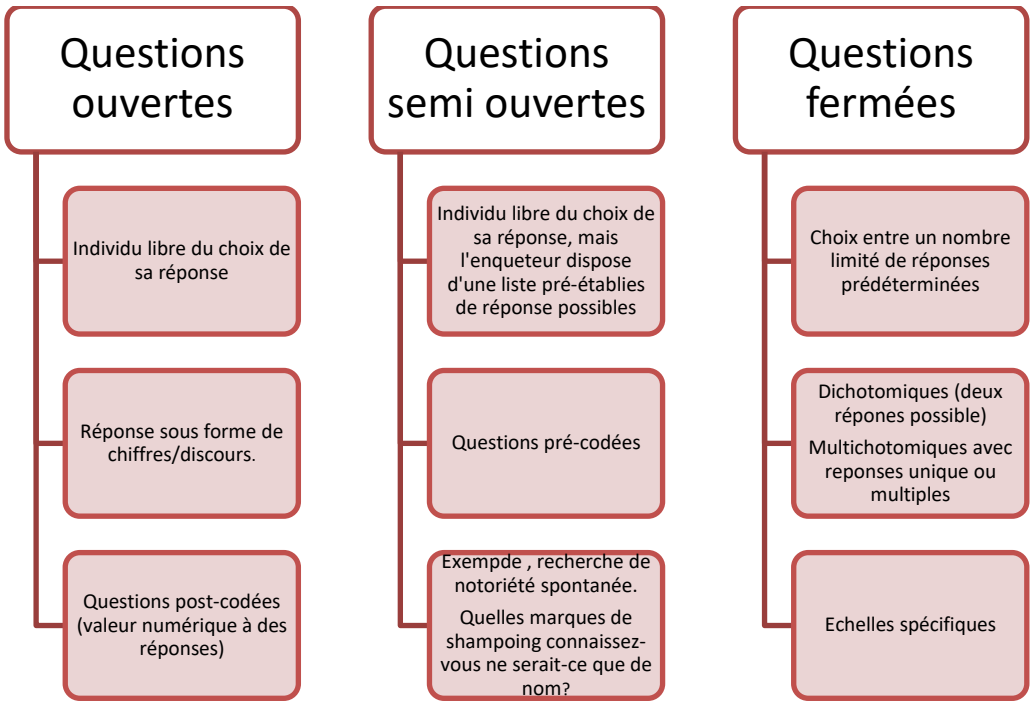
- Le contenu
- Le format
- La formulation
- Les niveaux de mesure

Quelles questions poser ? -> Quelle est l'info recherchée

- Champs d'investigation
    - Phénomènes mentaux : attitudes, préférences, intentions
    - Comportements (quoi, cb, qd, ou) : produits achetés, quantités, fréquences, prix, période, lieu d'achat, circonstance d'utilisation

- Des descripteurs socio-démo (age, sexe, situation familiale, taille du foyer, lieu et type d'habitat, équipement en biens durables) et psycho cognitifs (styles de vie, valeurs)

Le format des questions




Questions ouvertes : limitations des biais liées à l'influence de la question, Clarification d'une réponse structurée. La question peut être intéressante lorsque l'on veut clarifier d'avantage une réponse structurée. Prend du temps (réponse et analyse), difficultés de codage et d'analyse

Questions fermées : Facilité de réponse, simplicité de codage et d'analyse, coût et durée. Inconvénients : Possibilités d'expression limitées. Difficulté de création des modalités de réponse.

Au regard des indications, privilégier le format des questions fermées, SAUF si les modalités de réponses sont inconnues (pas possible de les définir en amont). Modalités exhaustives et exclusives

## ECHELLES D'ATTITUDE

- Une attitude est une prédisposition à réagir positivement ou négativement vis-à-vis d'un objet (marque, publicité)
- Effet sur le comportement
- Vision tripartite
    - Composante cognitive (Croyance subjectives), attributs intrinsèques (gout, aspect, odeur)/ extrinsèque (prix, garantie, marque)
    - Composante affective (sentiments, émotions)

- Composante conative (Tendances comportementales, intentions)

➔ Transformation d'information d'ordre qualitatif en données quantitatives (jugement global ou caractéristique spécifiques)

- Approches monadiques : jugement du stimulus isolément
- Approches comparatives (choix forcé) : jugement du stimulus relativt à d'autres stimulus

## APPROCHES COMPARATIVES

***Les comparaisons par paire***

Pour chacune des paires de marques de dentifrice suivantes, veuillez encercler celle que vous préférez.

⇨ Risque d'agacement si peu de différences perçues.
Complexité si nombre important d'objets à classer

***Les classements***

Veuillez classer les marques de dentifrice suivantes par ordre de préférence en leur attribuant des scores de 1 à 4. Donnez 1 au produit que vous jugez le meilleur et 4 au produit que vous jugez le moins bon : Colgate Total, Fluocaril, Signal Blancheur, Sanogyl Soin gencives

Veuillez classer les 4 dentifrices suivants sur chacune des caractéristiques indiquées.
Donnez 1 au produit que vous jugez e moins bon, puis 2, puis 3 et enfin 4 au produit que vous jugez le meilleur.

| | Colgate T | Signal B | Fluocaril | Sanogyl SG |
|---|---|---|---|---|
| Prévention des caries | | | | |
| Haleine fraiche | | | | |

## APPROCHES MONADIQUES

***Support d'identification :***

- **Verbal :** Estimez-vous que, dans la vie, les loisirs sont : Très importants à pas du tout importants
- **Numérique :** Si Monoprix ouvre un magasin, … : Pourcentage
- **Graphique :** Où situez-vous la marque de stylo A sur l'échelle suivante ? Bonhomme pas content à content. C'est utilisé pour les jeunes et dans les pays où il y a peu d'éducation.

***Nombre de modalités :***

Souvent entre 4 et 7. La barrière se situe à 5. Ca va jouer sur la précision.
Si on a peu de modalités, ça nous permet d'avoir peu de finesse dans le recueil. Trop de modalités peuvent perdre de son sens. Très souvent, on utilise des échelles à 7 points. L'impact va être important par rapport aux analyses statistiques. Le nombre de modalités va avoir un impact sur les propriétés statistiques de la variable associée à la question.

***Sens : Unidirectionnel/bidirectionnel :***

- **Unidirectionnel** : Intention d'achat, probabilité de l'acheter, on va seulement dans une seule direction *(positif ou négatif mais pas les deux)*
- **Bidirectionnel** : Si on est favorable ou pas à une marque.

***Parité : (échelles bidirectionnelles uniquement) :***

Format en 5 ou en 6 modalités.

Quand on a une échelle impaire, ça permet d'avoir une réponse qui est en accord et en désaccord. On va privilégier une échelle paire quand on veut pousser une personne à prendre une position. ➔Il n'y a pas une réponse miracle, on a droit d'avoir aucune opinion sur un sujet.

***Symétrie/Asymétrie***

| | | | | | |
|---|---|---|---|---|---|
| Très important | Important | Assez important | Asse peu important | Peu important | Très peu important |

**Risque d'une échelle asymétrie** : Risque d'influencer la personne *(Survaloriser l'importance de quelque chose).* On prend toujours une échelle symétrique car ça se repose sur des hypothèses très fortes.

***Echelle simple ou multiple***

- **Echelle simple** : Un seul libellé.
- **Echelle multiple** : Plusieurs libellés.

***L'échelle de Likert :***

Expression du degré d'approbation ou de désaccord : de tout à fait d'accord à pas du tout d'accord.

- Possibilité de calcul d'un score global. On peut sommer les scores. Penser à recoder dans le ***sens inverse.*** Mettre « Tout à fait d'accord » = 5 permet d'avoir un score élevé. Mais il faut quand même rester vigilant.

| | Tout à fait d'accord | Plutôt d'accord | Ni d'accord ni en désaccord | Plutôt pas d'accord | Pas du tout d'accord |
|---|---|---|---|---|---|
| A un parfum très agréable | 5 = favorable | | | | |
| La texture de A est trop grasse | 5 = défavorable | | | | |

- Choisir des items positifs et négatifs *(attention pour le calcul du score global)* : ça permet de maintenir l'attention.
- Possibilité d'ajustement du nombre de catégories et expression numérique.
- Possibilité d'avoir des chiffres en intermédiaires.

***L'échelle à supports sémantiques de Thurstone***

- Libellés à des distances psychologiquement égales l'une de l'autre.
  *Comment jugez-vous la solidité des produits suivants ?*

| | Excellente | Très bonne | Plutôt bonne | Moyenne | Plutôt mauvaise | Mauvaise | Extrement mauvaise |
|---|---|---|---|---|---|---|---|
| Produit A | | | | | | | |

- Adaptation possible des nuances et des positions de l'échelle à chaque énoncé.
  *Quel est votre niveau de satisfaction globale des prestations fournies ?*

| | | | | |
|---|---|---|---|---|
| Très mécontent | Peu satisfait | Moyennement satisfait | Satisfait | Très satisfait |

Lors de l'achat d'une voiture, quelle importance accordez-vous à chacune des caractéristiques suivantes ?

| Très peu important | | Moyennement important | | Très important | | | |
|---|---|---|---|---|---|---|---|
| La vitesse | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Le confort | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| La fiabilité | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Le design | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

***L'échelle sémantique différentielle d'Osgood***

Positionnement sur une **échelle bipolaire de 5 à 7** qui oppose *deux adjectifs de sens contraire.*
Pour vous personnellement le produit x est :

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Agréable | - | - | - | - | - | - | - | Désagréable |
| Efficace | - | - | - | - | - | - | - | Inefficace |
| Solide | - | - | - | - | - | - | - | Fragile |
| Simple | - | - | - | - | - | - | - | Compliqué |
| Cher | - | - | - | - | - | - | - | Bon Marché |

- Pour évaluer l'image d'une marque.
- Inverser la position des adjectifs positifs (gauche/droite)

Utilisation fréquente pour l'étude d'image de marque

| | Extrement | Tres | Assez | Sans opinion | Assez | Tres | Extrement | |
|---|---|---|---|---|---|---|---|---|
| Efficace | | X | | | O | | | Inefficace |
| Solide | X | | O | | | | | Fragile |
| Simple | | | O | | X | | | Compliqué |
| Bon Marché | | | O | | X | | | Cher |

Les ronds/croix correspondent au score moyen des différentes marques, mesure de <u>l'attitude cognitive</u> (je crois que cette marque est très efficace)

Par extension, utilisation avec des expressions ou phrases complètes : <u>Composante affective</u>

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Je n'apprécie pas du tout ce produit | 1 | 2 | 3 | 4 | 5 | 6 | 7 | J'apprécie beaucoup de produit |

***Mesure de l'intensité des intentions*** <u>(composante conative)</u>

**1.** Si ce nouveau produit était en vente l'achèteriez-vous ?

Certainement ☐ Probablement ☐ Peut-être ☐
Probablement pas ☐ Certainement pas ☐

**2.** Si le produit X était proposé en bouteille de 50 cl

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Je n'en achèterai certainement pas | 1 | 2 | 3 | 4 | 5 | 6 | 7 | J'en achèterai |

**3.** Avez-vous l'intention d'acheter une voiture au cours des six prochains mois ?

| | | | | | |
|---|---|---|---|---|---|
| 0% Aucune chance | 20% Faible possibilité | 40% Une certaine possibilité | 60% Bonne possibilité | 80% Très probablement | 100% Certainement |

***Critères de qualité des échelles***

- **Facilité de réponses** : échelles monadiques > échelles comparatives
- **Capacité de discrimination :** *(expression différenciée de l'opinion)* : échelles comparatives > échelles monadiques
- **Capacité de transmission** : nombre optimal de catégories = 7

## FORMULATION

Le répondant doit **comprendre la question**. : adopter un vocabulaire simple

Le répondant doit **connaitre la réponse** *(Détention et souvenir de l'information)*

- ⇨ Réduire l'effort de la mémoire : aider le répondant à se souvenir.
- ⇨ Si le répondant ne répond pas : On a mal défini la population cible. On a bien défini la population cible mais l'individu n'en fait pas parti.

Le répondant doit **avoir envie de donner la vraie réponse** *(biais de désirabilité sociale=se montrer sous un statut social supérieur au statut réel, biais de conformisme=tendance à donner des réponses semble correspondre à ce qu'on imagine bien socialement )*

- Sujets ayants traits à la vie privée *(par exemple : le revenu)*
- Sujets socialement sensibles *(par exemple : consommation d'alcool)*
  - ⇨ Placer les questions parmi d'autres
  - ⇨ Souligner qu'il s'agit d'un comportement qui n'est pas inhabituel « *Des études récentes indiquent que de plus en plus de personnes …* »

| **Ce qu'il ne faut pas faire …** → | **Ce qu'il faut faire …** |
|---|---|
| Utiliser un vocabulaire spécialisé ou abstrait *(adopter le vocabulaire du répondant)*<br><br>*Pensez-vous que le processus de lyophilisation des aliments diminue leur teneur en vitamine A ?* | Utiliser le vocabulaire du répondant, définir les termes complexes utilisés |
| Utiliser des termes vagues, ambigus<br><br>*Au cours d'un mois normal, vous venez faire des achats à la FNAC ?*<br>*Très rarement* ☐<br>*Occasionnellement* ☐<br>*Souvent* ☐<br>*Très souvent* ☐<br>*Ne sait pas* ☐ | Utiliser des termes objectifs<br><br>*Combien de fois par mois venez-vous faire des achats à la FNAC :*<br>*Moins d'une fois* ☐<br>*1 à 2 fois* ☐<br>*3 à 4 fois* ☐<br>*Plus de 4 fois* ☐<br>*Ne sait pas* ☐ |
| Avoir deux idées dans la même question<br><br>*Pensez-vous que les produits issus de l'agriculture soient bons pour la santé et meilleurs au gout ?* | Scinder en deux phrases.<br>Utiliser des phrases focalisées sur une seule problématique. |
| Utiliser les doubles négations, des phrases longues, des formes grammaticales complexes<br><br>*Ne pensez-vous pas que le personnel n'est pas suffisamment compétent ?* | Utiliser un style direct, bref |
| Induire une réponse<br><br>*Pensez-vous que les consommateurs achètent* | Utiliser une formulation neutre<br><br>*A votre avis, pour quelles raisons, les* |

| | |
|---|---|
| *dans les hypermarchés car les prix y sont très compétitifs ?* Ici, on crée un biais.<br><br>*Pensez-vous que le prix de l'essence sera plus cher l'année prochaine que cette année ?* | *consommateurs achètent-ils dans les hypermarchés ?*<br><br>*Par rapport au prix de l'essence cette année, pensez-vous que, l'année prochaine, le prix de l'essence va :*<br>*Augmenter* ☐<br>*Diminuer* ☐<br>*Rester stable* ☐<br>*Ne sait pas* ☐ |

## STRUCTURE DU QUESTIONNAIRE

***La dynamique du questionnaire :***

- L'ordre des questions
- Les questions filtre
- Les effets d'interaction
- La mise en page
- Le codage.

***Technique de l'entonnoir :***

Démarrer par des questions générales, simples, peu engageantes et demandant peu d'efforts de réflexion et de mémoire … pour aller vers des questions plus précises, difficiles et personnelles.

- Familiarisation progressive avec le sujet
- Limitation des biais dus à l'influence de questions spécifiques posées avant des questions générales.

Un climat s'instaure avec l'intervieweur ce qui va faire qu'on va être plus en confiance pour se livrer. Et puis cette technique va permettre de réduire les biais qui pourraient résulter d'une influence de questions un peu spécifique sur quelque chose d'un peu plus générale.

***L'ordre des questions :***

1. **Présentation enquêteur, sujet**

   Présentation du sujet de façon simple, générale, qui donne envie de répondre. *« Bonjour madame, je m'appelle XXX et je réalise pour l'institut de sondage XXX une étude sur es boisons des Français »*. On ne va pas annoncer de façon très précise car on ne veut pas créer de biais.

2. **Questions introductives**

   Questions ouverte, simple, à réponse positive : mise en confiance et incitation à continuer. *« Que buvez-vous habituellement ? »*

   ***NB :*** La prof n'est pas fan, elle trouve que ça fait trop démarchage commerciale

   ***On etudie d'abord le comportement et ensuite l'attitude, car comportement objectif sur sa façon de se comporter, pas de biais pour le reste du questionnaire***

3. **Questions qualifiantes**

   Question dont le but est de déterminer si la personne détient l'information recherchée et/ou de l'orienter vers des parties spécifiques du questionnaire. La personne fait-elle partie de la population cible ? *« Consommez-vous, même rarement, des … ? »*

4. **Questions principales**
    - Centrage progressif sur le thème de l'étude
    - Organisation par thèmes : Comportement et attitudes
    - Phrases de transitions entre les thèmes *« Nous allons maintenant parler de … » :* il faut que la personne se sente bien encadrée ce qui va faciliter la récupération en mémoire.

5. **Identification**
    - Questions relatives à l'identification du répondant : âge, sexe, profession, revenus … ➔ on offre la possibilité de refus de réponse.
    - Remerciements

***Les questions filtre :***

Questions visant à assurer l'enchainement logique de l'interview en fonction des réponses

Q3 : Possédez-vous une voiture ? Oui ☐
Non ☐ *(passez à la question 7)*

Une ou plusieurs par questionnaire mais pas d'abus.

***Les effets d'interaction :***

Ca fait référence au biais qu'une question qui peut être lié à une influence d'une question.

## L'effet de halo

Dans une séquence de questions, influence d'un jugement sur les suivants, ainsi qu'une préférence globale sur l'évaluation d'attributs *(tendance à répondre de la même façon)*

⇨ Alterner les énoncés favorables et défavorables dans les échelles multi-items.

## L'effet de contamination

Influence directe d'une question sur les suivantes

- **Espacer les questions**_ Le fait de poser une question va influencer la réponse aux questions suivantes : l'effet d'ordre. L'ordre des modalités présenté fait que nous allons retenir la première et la dernière question.
- **Changer l'ordre des questions/des modalités** : ça permet d'enlever l'effet de biais.

## La mise en forme

***Consignes pour les enquêteurs ou pour les répondants***

- Instructions claires et précises
- Forme différente *(typologie ou autre)*
- *L'auto administrée* : Le répondant qui répond lui-même aux questionnaires. Celui qui conçoit le questionnaire doit écrire des instructions claires pour les personnes qui administrent le questionnaire. Il est recommandé d'avoir une typologie différente.

## *La mise en page*

- Numérotation des pages et des questions
- Avoir une question en totalité sur la même page
- Aérer les questions
- Utiliser éventuellement de la couleur.

Utiliser éventuellement des aides visuelles pour relancer l'intérêt de l'interviewé *(ex ; échelles)*

## Validité et fiabilité

Les qualités des instruments de mesure

- **Fiabilité** : degré de précision et de reproductibilité des résultats fournis par un instrument lorsqu'on l'applique plusieurs fois.
- **Validité** : degré auquel un instrument mesure bien ce que l'on pense mesurer. Référence à la clarté. **Exemple :** Instrument de mesure : la balance _ normalement quand je pèse tout le temps quelque chose qui pèse un 1 kilo alors la balance affichera tout le temps 1 kilo. Ici, on aura la fiabilité. Concernant la validité, si vous pesez tout le temps une boite de 1 kilos mais que la balance affiche 900 grammes alors ça ne sera pas valide mais si ça affiche à chaque fois 900 grammes.
- **Sensibilité** : capacité de l'instrument à enregistrer des variations fines du phénomène mesuré *(attitudes…)*.

## Le modèle de la « vraie valeur »

**MESURE OBTENUE** = **Vraie valeur**
**+ Erreur systématique**
**+ Erreur aléatoire** *(erreur liée à des aléas tel que les circonstances, l'humeur du répondant)*

Note :

- Vraie valeur : *valeur correspondant parfaitement au phénomène étudie*
- Erreur systématique : érreur récurrente

## NIVEAUX DE MESURE:

Propriétés statistiques associés à des questions.
Quand on a un questionnaire, les réponses doivent être rentrées dans une base de données. On va passer par **une liste de variables**. On va les exploiter statistiquement. Le format de question va influer le niveau de mesure. Le niveau de mesure dépend du niveau de la question même si on va l'utiliser et l'exploiter au moment de l'élaboration du questionnaire.

Il existe 4 niveaux :

- Echelles nominales :
  homme ou femme
- Echelles ordinales :
  classement
- Echelles d'intervalles :
  Classe d'age [18 ;25]
- Echelles de ratio :
  Le répondant est libre dans sa réponse

Là, le mot « échelle » n'a pas le même sens que « l'échelle d'attitude ». Ici, il a un sens statistique. Ces échelles sont organisées, on peut les classer par ordre croissant de pouvoir.

## Echelle : Catégorielle, nominale

**Caractéristiques :** Identification de l'appartenance à une classe
**Exemples :** Sexe, CSP …
Quelle est votre situation familiale ?

| | |
|---|---|
| Célibataire | 1 |
| Marié/en couple | 2 |
| Divorcé, séparé, veuf | 3 |

⇨ ***Le niveau le plus bas*** : **l'échelle catégorielle nominale** ou **la valeur numérique** qui va être attribué. Il n'a aucun sens en termes numériques. Permet de désigner l'appartenance à un groupe d'individus. Ce nombre n'a aucun sens en termes de normes, ça désigne l'appartenance à une classe c'est comme une étiquette.

## L'échelle ordinale

**Caractéristiques :** Relation d'ordre

➔ Pas possible de faire une moyenne avec les echelles ordinales

Quel âge avez-vous ?

- Entre 18 et 25 ans groupe 1
- Entre 26 et 35 ans groupe 2
- Entre 36 et 50 ans groupe 3
- Plus de 50 ans groupe 4

Les groupes permettent d'identifier la catégorie d'age

Veuillez classer les marques de dentifrices suivantes par ordre de préférence en leur attribuant des scores de 1 à 4. Donnez 1 au produit que vous jugez le meilleur et 4 au produit que vous jugez le moins bon.

- Colgate Total
- Fluoracil
- Signal Blancheur
- Sanogyl Soin Gencives

⇨ La **valeur numérique** est représentée davantage. Elle détermine l'appartenance à un groupe. Si on parle de statistiques de bases univariées. On pourra parler et calculer des médianes et pas la moyenne. 2,3 ne veut rien dire. Il n'y a pas forcément les mêmes distances en nombre d'années.

## L'échelle Intervalle

**Caractéristiques :** Intervalle «équivalent entre deux unités de valeurs successives. Zéro arbitraire (standart défini, ne veut pas dire absnce de réponse)
**Exemples :** Température *(où 0 ne veut pas dire qu'il n'y a pas de température)*, échelles d'attitude *(nb catégories >5)*

*Quel est votre niveau de satisfaction globale des prestations fournies ?*

| Très Mécontent | Peu satisfait | Moyennement satisfait | Satisfait | Très satisfait |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

⇨ L'échelle d'intervalle signifie qu'en plus on a une relation d'ordre. En revanche, le zéro n'est pas un 0 naturelle mais arbitraire, il n'est pas absent.

⇨ Une échelle attitude est considérée comme une échelle intervalle s'il y a 5 points. (Le ne sais psa ne compte pas comme un point). Si elle en comporte que 4 points, elle sera assimilée à des échelles ordinales *(donc pas intervalle, car ne comporte*

*pas les propriétés statistiques nécessaires pour l'intervalle).* Sur une echelle intervalle possibilité de faire une moyenne, pas sur une echelle ordinale

**Plus l'échelle a des propriétés, plus elle sera plus simple à traiter.**

## Echelle de proportion

**Caractéristiques** : Existence naturelle du zéro. Conservation des rapports
**Exemples** : Age, revenu, quantité

### ILLUTRATION

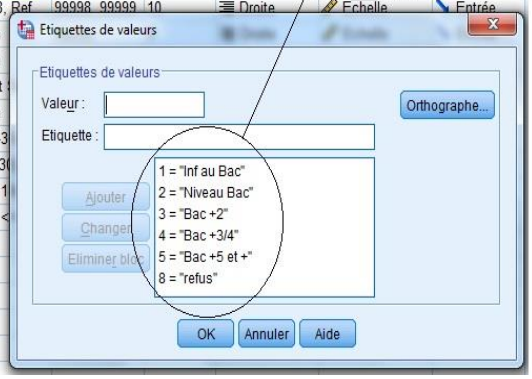
| SATEMPLO | SEXE | MARITAL | NBRPER... | AGE | ADRESSE | ETUDES | REVENU |
|---|---|---|---|---|---|---|---|
| 5 | 1 | 1 | 4 | 55 | 12 | 1 | 72 |
| 4 | 2 | 0 | 1 | 56 | 29 | 1 | 153 |
| 3 | 1 | 1 | 3 | 28 | 9 | 3 | 28 |
| 1 | 2 | 1 | 3 | 24 | 4 | 4 | 26 |
| 2 | 2 | 0 | 2 | 25 | 2 | 2 | 23 |
| 2 | 2 | 1 | 2 | 45 | 9 | 3 | 76 |
| 2 | 2 | 0 | 1 | 42 | 19 | 3 | 40 |
| 1 | 1 | 0 | 1 | 35 | 15 | 2 | 57 |
| 5 | 1 | 0 | 2 | 46 | 26 | 1 | 24 |
| 4 | 2 | 1 | 6 | 34 | 0 | 3 | 89 |
| 3 | 1 | 1 | 2 | 55 | 17 | 3 | 72 |
| 5 | 2 | 0 | 1 | 28 | 3 | 4 | 24 |
| 2 | 1 | 1 | 4 | 30 | 9 | 4 | 40 |
| 1 | 1 | 0 | 1 | 42 | 8 | 3 | 137 |
| 4 | 2 | 0 | 3 | 35 | 8 | 3 | 70 |
| 5 | 2 | 1 | 2 | 52 | 24 | 4 | 159 |
| 1 | 2 | 1 | 7 | 21 | 1 | 3 | 37 |
| 4 | 1 | 0 | 2 | 32 | 0 | 1 | 28 |
| 3 | 1 | 0 | 1 | 42 | 9 | 3 | 109 |
| 5 | 1 | 1 | 4 | 40 | 12 | 2 | 117 |
| 3 | 2 | 0 | 1 | 30 | 3 | 1 | 23 |
| 3 | 2 | 0 | 1 | 48 | 14 | 3 | 21 |
| 3 | 2 | 1 | 5 | 39 | 17 | 4 | 17 |
| 3 | 1 | 1 | 4 | 42 | 5 | 2 | 34 |
| 4 | 1 | 1 | 5 | 45 | 12 | 1 | 115 |
| 3 | 2 | 1 | 3 | 51 | 10 | 1 | 47 |
| 1 | 2 | 1 | 4 | 39 | 9 | 3 | 33 |
| 5 | 1 | 0 | 1 | 49 | 29 | 2 | 135 |
| 5 | 2 | 0 | 1 | 52 | 20 | 1 | 272 |
| 4 | 2 | 1 | 2 | 53 | 29 | 1 | 41 |
| 1 | 1 | 0 | 1 | 34 | 10 | 3 | 20 |
| 4 | 2 | 1 | 2 | 47 | 6 | 3 | 22 |
| 1 | 2 | 0 | 1 | 58 | 2 | 4 | 60 |
| 2 | 1 | 1 | 5 | 25 | 0 | 3 | 58 |

Liberté dans la réponse, donc il s'agit d'une variable de type ratio, donc échelle

La personne se classe selon son niveau d'études. On pourrait penser variable nominale. Néanmoins, il y a un ordre croissant, donc il s'agit d'une variable de type ordinale.

La personne est soit une femme ou un homme, donc elle est classé dans une des deux catégories, donc variable nominale.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 14 | DOUBLAPE | Numérique | 4 | 0 | Double appel | {0, Non}... | 8 | 8 | Droite | Nominales | Entrée |
| 15 | MAGAZINE | Numérique | 4 | 0 | Abonnement à un magazine | {0, Non}... | 8 | 6 | Droite | Nominales | Entrée |
| 16 | RETRAITE | Numérique | 4 | 0 | statut professionnel | {0, actif}... | Aucun | 6 | Droite | Nominales | Entrée |
| 17 | EMPLOI | Numérique | 4 | 0 | Nombre d'années chez l'employ... | {99, non co... | 99 | 6 | Droite | Echelle | Entrée |
| 18 | SATEMPLO | Numérique | 4 | 0 | Satisfaction Emploi | {1, Très ins... | 9 | 6 | Droite | Echelle | Entrée |
| 19 | SEXE | Numérique | 2 | 0 | Sexe | {1, femme}... | Aucun | 8 | Droite | Nominales | Entrée |
| 20 | MARITAL | Numérique | 4 | 0 | Statut marital | {0, Non mar... | Aucun | 7 | Droite | Nominales | Entrée |
| 21 | NBRPERSO | Numérique | 4 | 0 | Nombre de personnes dans le fo... | Aucun | Aucun | 7 | Droite | Echelle | Entrée |
| 22 | AGE | Numérique | 4 | 0 | Age en années | Aucun | Aucun | 6 | Droite | Echelle | Entrée |
| 23 | ADRESSE | Numérique | 4 | 0 | Nombre d'années à l'adresse act... | Aucun | Aucun | 7 | Droite | Echelle | Entrée |
| 24 | ETUDES | Numérique | 4 | 0 | Nombre d'années d'études | {1, Inf au Ba... | 8 | 6 | Droite | Ordinales | Entrée |
| 25 | REVENU | Numérique | 8 | 0 | Revenu du foyer en milliers ($) | {99998, Ref... | 99998, 99999 | 10 | Droite | Echelle | Entrée |
| 26 | VAR00001 | Numérique | 8 | 2 | | Aucun | | | | | |
| 27 | COMPOSITI... | Numérique | 8 | 2 | | Aucun | | | | | |
| 28 | filter_$ | Numérique | 1 | 0 | REVENU < 200 (FILTER) | {0, Not | | | | | |
| 29 | ageclasse | Numérique | 8 | 2 | | Aucun | | | | | |
| 30 | agecategorie | Numérique | 8 | 0 | age en catégorie | {1, 18-3 | | | | | |
| 31 | revenucateg | Numérique | 8 | 0 | revenu en catégorie | {1, <=3 | | | | | |
| 32 | catâge | Numérique | 8 | 2 | catégorie âge | {1,00, 1 | | | | | |
| 33 | catrevenu | Numérique | 8 | 2 | catégorie revenu | {1,00, < | | | | | |
| 34 | | | | | | | | | | | |
| 35 | | | | | | | | | | | |
| 36 | | | | | | | | | | | |
| 37 | | | | | | | | | | | |
| 38 | | | | | | | | | | | |




Il y a un ordre de classement, donc il s'agit d'une variable de type ordinale.

## Le codage

Pour faciliter la transformation des informations fournies par le répondant en vue des traitements statistiques. Le questionnaire doit être transformé en base de données.
Valeur utilisée pendant les questions reportées sur le tableau
Ce codage est plus que souhaitable au moment où on élabore le questionnaire. Quand on fait la saisie ça sera extrêmement rapide. Si en amont, on l'a pas faite on doit écrire à la main les données pour chaque question.

| | **Sexe** | **Age** | **Etudes** | **Revenu** | **…** | **Valeur de la codification des questions** |
|---|---|---|---|---|---|---|
| **Individu 1** | | | | | | |
| **Individu 2** | | | | | | |
| **Individu 3** | | | | | | |
| **Individu 4** | | | | | | |
| | | | | | | |

### 1ère étape : Identification de la *(ou les)* variables liées à chaque question

*Question 11 : Je vais vous citer plusieurs affirmations relatives au produit A. Pouvez-vous me dire si vous êtes : tout à fait d'accord…*

| | **Tout à fait d'accord** | **Plutôt d'accord** | **Ni d'accord, ni en désaccord** | **Plutôt pas d'accord** | **Pas du tout d'accord** |
|---|---|---|---|---|---|
| **A a un très agréable parfum** | **X** | | | | |
| **La texture de A est trop grasse** | | | **X** | | |

Quelles marques de shampooing connaissez-vous, ne serait-ce que de nom ?

| | | |
|---|---|---|
| /2/ Palmolive | / / Pantene | NSDESSA /1/ (oui=1, non=2) |
| /4/ DOP | /1/Fructis de Garnier | RGDESSA /3/ |
| /3/ Dessanges | /5/ Mixa bébé | NBCITEES /5/ |

NSEDESSA = notoriété spontanée de la marque Dessanges
RGDESSA = rand d'apparition
NBCITEES = nombre total de marques citées

### 2eme étape : Identification et codage des modalités

*Question 14 : Parmi les raisons suivantes, quelles sont celles qui à votre avis, expliquent votre fidélité à cette station-service ?*

| | |
|---|---|
| C'est la plus proche de chez moi | FIDEPRO /1/ |
| Elle est proche et attirante | FIDEATT /2/ |
| Ses prix sont attirants | FIDEPRIX/2/ |
| Elle offre un service complet | FIDECOM/2/ |
| Le personnel est aimable | FIDEAIM /1/ |
| Ne sait pas | FIDENSP /2/ |
| Autres raisons_____ | |

*INSTRUCTIONS* : coder 1 si la raison est citée et 2 sinon
Ici, on n'a pas une question mais plusieurs. Là, on a autant de variables que de modalités possibles de réponses.
Questions avec plusieurs réponses possibles, chaque modalité de réponses devient variable dans la base de donnée. Alors que si une seule réponse possible, seule modalité à mettre dans la base de donnée

**Codage des non-réponses**, avec si possible le même code pour l'ensemble du questionnaire (9). On aura une valeur spécifique pour le *« ne sait pas »* et *« absence de réponse ».* Il n'y aura donc pas le même codage car ça ne veut pas dire la même chose. On ne met pas la même valeur.
Si bcp d'absence de réponses, on peut se dire qu'il y a eu un pb de def de la pop/ d'étude, question pas clair, questions confidentielles…

***Numérotation des questionnaires*** *(sauf si postal car peut être perçu comme un mode d'identification)*

- Contrôle d'enquêteur
- Validation de la saisie en cas de données aberrantes
  + Date et lieu d'enquête

Si saisie manuelle, il est fréquent qu'il y ai des erreurs

C'est fondamental pour les instituts pour contrôler. Si on ne numérote pas le questionnaire, il serait difficile de savoir quelle réponse correspond à quelle question.

## Le pré-test

**Il permet de valider que notre questionnaire est bon, avant de lancer l'enquête**

***Les objectifs***

- Détecter les incompréhensions *(demandes d'explications)* : ça apparait quand la personne ne comprend pas la question *(revoir sa formulation)*
- Détecter un manque de clarté dans les instructions
- Identifier les problèmes de choix de modalités *(Réponses concentrées sur une modalité ou sur NSP)*
  - ⇨ Si beaucoup de personnes répondent par *« ne sait pas »* ça veut dire que quelque chose cloche. Si plusieurs répondants choisissent la même réponse ça veut dire que ça cloche aussi.
- Prévoir les réponses *(réponses aux questions ouvertes les plus fréquentes retenues)*
- Déceler des erreurs dans les filtres
- Estimer la durée d'administration : *combien de temps dure le questionnaire ?*

***Identification des questions posant problème, mais pas de celles qui peuvent avoir été oubliées***

- **Qui ?**
  - Concepteur du questionnaire et enquêteurs expérimentés.
    Auprès des individus qui ont le même profil que celui défini par votre population mère.
- **Comment ?**
  - Idéalement en conditions réelles
  - Auprès d'individus identiques à ceux de l'échantillon selon les critères de définition de la population
    **Exemple :** Si vous faites une enquête sur les seniors, ne pas demander à vos voisins d'à coté.
- **Combien ?**
  - Pré-test de compréhension : entre 20 et 50
  - Pré-test statistique : au minimum de 100

**Sources d'erreurs :**

***Erreur de non-réponse***

Erreur causée par l'absence d'interview ou de non-réponse à une question d'un questionnaire.

| NON REPONSE PARTIELLE | NON REPONSE TOTALE |
|---|---|
| Absence de réponse à certaines questions du questionnaire.<br>*Questions difficiles à comprendre, perçues comme indiscrètes*<br>⇨ **Elaboration du questionnaire**<br>⇨ **Méthode d'imputation (si non réponse à une ou au pire 2 qustions) ; imputation réponse moyenne donné par le profil de l'individu**<br><br>Traiter les données en déclarant qu'il y a des données manquantes ou adopter une méthode d'imputation avec différentes techniques *(exemple : la valeur moyenne de chaque individu)* | Absence de réponse de la part de certaines unités de l'échantillon<br>*Absence, indisponibilité, refus*<br><br>Problème si comportement des répondants différent de celui des non-répondants<br>⇨ **Méthode de redressement**<br><br>Le non-répondant a un profil particulier. Vos études ne seront pas exhaustives et fiables.<br><br>Extrapoler un échantillon à la population entière. Redresser le questionnaire en pondérant par le poids de chaque groupe |

**Redressement :**

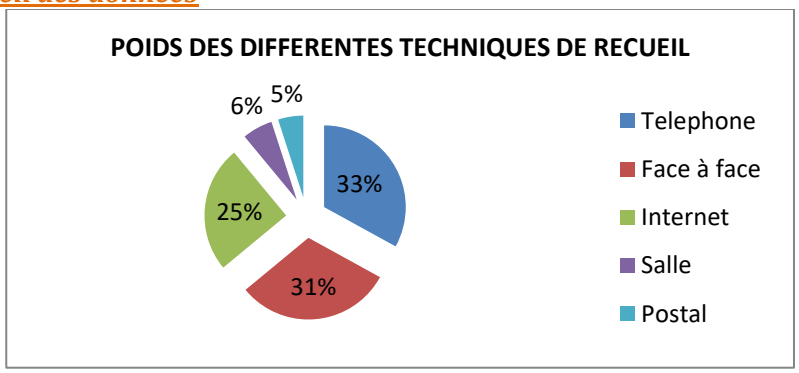
Traitement dont le but est de corriger les biais dus aux non-réponses ou aux sous-représentations et surreprésentations de certaines classes de population. Le principe consiste à pondérer les résultats observés par le poids relatifs de chaque classe dans la population.

***Erreur de réponse***

Différence entre la vraie réponse à une question et la réponse donnée par le répondant
Elle peut être imputable :

- Au répondant *(bais de désirabilité sociale, incapacité à répondre …)*
- A l'enquêteur *(effet induit sur la réponse du répondant)*
- Au questionnaire qui n'a pas été collaborée convenablement : des questions pas claires pas assez précises.

## *Les méthodes de recueil des données*




### Enquête en face à face

- ***Lieux***
  - Au domicile
  - Dans la rue/sortie de caisse
  - Hall tests *(mallintercepts)*
- ***L'exécution du travail par l'enquêteur***
  - Choix des personnes à contacter *(quotas)*
  - Stimulation du répondant
  - Respect du questionnaire et codage des réponses
- Utilisation de micro-ordinateurs avec le logiciel CAPI *(Computer Assisted Personal Interviewing)*
-

### Enquête par téléphone

- ***Exécution classique*** ou utilisation du système CATI *(Computer Assisted Telephone Interviewing)*
  - Déroulement du questionnaire à l'écran
  - Ecrans d'aide
  - Intégration directe des réponses
  - Composition automatique des numéros
  - Rappel en cas d'absence
  - Contrôle automatique des quotas
  - Contrôle des enquêteurs

### Enquête Postale

- ***Exécution***
  - Utilisation de fichiers clients ou loués
  - Calendrier avec lettre préliminaire
  - Questionnaires simples et attractifs
  - Enveloppes pré-timbrées
  - Incitations-récompenses

### Enquête par internet

- ***Exécution***
  - Envoi par e-mail avec fichier attaché
  - Invitation à venir sur le site
    - ⇨ E-mail, bannières publicitaires

⇨ CAWI *(Computer Assisted Web Interviewing)*

***Risque***

- Représentativité population
  - ⇨ *Premier problème :* diffusion d'Internet dans la population française, nettement amoindri
  - ⇨ *Deuxième problème* : Biais dans la population qu'on touche *(population prête à répondre sur internet).*

**ATTENTION AU SIGNE !!!!**+ = Atout

| | **Face à face** | **Postal** | **Téléphone** | **Internet** |
|---|---|---|---|---|
| RECUEIL | | | | |
| **Longueur questionnaire** | + | = | = | = |
| **Flexibilité** | + | - | = | = |
| **Utilisation de stimuli** | + | = | - | = |

Le postal est celui qui est le moins flexible car l'interviewé voit tout le questionnaire. *L'utilisation des stimuli* ça va être un packaging, un produit, une publicité. C'est le face-à-face qui offre le plus d'opportunités. Le postal et internet laissent une possibilité visuel mais qui reste néanmoins faible. Le téléphone est celui qui est le plus fermé sauf pour un stimulus sonore. Mais le plus souvent, le stimulus est visuel.

| | | | | |
|---|---|---|---|---|
| ECHANTILLON | | | | |
| **Identité répondant** | + | - | + | - |
| **Taux de réponse** | + | - | = | = |
| **Disparité géographique** | = | + | + | + |

Le postal représente le taux de réponse le plus faible. Tandis qu'un face-à-face a un taux plus fort mais de plus en plus les gens refusent de fairele questionnaire en face-à-face surtout à domicile. La disparité géographique : le face-à-face est le plus compliqué à mettre en œuvre si on veut mener différents enquêtes en France.

| | | | | |
|---|---|---|---|---|
| BIAIS | | | | |
| **Enquêté** | - | + | = | + |
| **Enquêteur** | - | + | = | = |

Le biais lié à l'enquêté correspond à tous les biais liés au conformisme ; ces biais impliquent qu'il y ait une présence social. En face-à-face et au téléphone, car on a le regard social sur l'individu. Le risque de biais est plus forte en face-à-face c'est du à une proximité forte, l'enquêteur peut générer des biais : influencer de par son attitude, de ses expressions. Sur internet et Postal, il n'y a pas ce genre de problème.

| | | | | |
|---|---|---|---|---|
| ASPECTS OPERATOIRES | | | | |
| **Contrôle enquêteur** | - | = | + | = |
| **Rapidité** | = | - | + | + |
| **Cout** | - | + | = | + |

+ : Avantage ou - : Inconvénient

En termes de rapidité du terrain, c'est le téléphone et internet qui sont le plus rapide. Le postal étant la méthode qui prend le plus de temps : attente les retours par exemple. Enfin en termes de couts, le postal et internet sont les techniques les moins couteuses : pas besoin de rémunérer l'enquêteur. Le face-à-face est la solution la plus couteuse.

# 1. TECHNIQUES D'ECHANTILLONAGE

Un **sondage** est une étude des caractéristiques d'une population à partir d'un échantillon qui en est issu.

⇨ *Comment tirer des conclusions à propos d'une population à partir d'un échantillon limité ? Recueillir des informations fiables et de qualités. Comment on peut faire pour optimiser notre technique d'échantillonnage sachant qu'on va procéder à un recensement ?*

## Erreur d'échantillonnage

Erreur liée au fait que l'on interroge uniquement une fraction de la population. Cette erreur d'échantillonnage ***correspond à l'écart*** entre **l'estimation calculée de l'échantillon** et **la vraie valeur qui serait obtenue via un recensement**.

L'erreur est :

- D'autant ***plus forte*** que la population est **hétérogène**
- D'autant ***plus faible*** que l'échantillon est **grand**

Et peut être réduite et contrôlée en fonction de la méthode de sondage utilisée. On ne contrôle pas l'hétérogénéité *(c'est une donnée en soi)* mais l'erreur.

PLAN DE SONDAGE

1. DEFINITION DE LA POPULATION CIBLE
2. CHOIX DE LA METHODE DE SONDAGE
3. TAILLE DE L'ECHANTILLON

## DEFINITION DE LA POPULATION CIBLE

**POPULATION CIBLE** → **BASE DE SONDAGE** → **ECHANTILLON**

**Population cible :** Ensemble des éléments possédant les informations permettant de répondre aux questions de l'étude. C'est important d'être extrêmement précis à ce moment-là.

**Exemple :** Consommateurs/Acheteurs : critères sociodémographiques

***Définition des unités d'échantillonnage*** : Entités qui pourront être sélectionnées pour faire partie de l'échantillon.
**Exemple :** un foyer ou un individu mais ce n'est pas la même chose en termes de sélection.

*Après avoir défini notre population cible, on va effectuer notre base de sondage.*

**Base de sondage :** Liste exhaustive des éléments de la population cible à partir de laquelle la sélection va être opérée.
**Exemples :** Annuaires téléphoniques, annuaires d'entreprise, base de données clients.

***Une liste complète et à jour :***

- Aucun membre de la population ne doit en ***être exclu ni y être représenté plusieurs fois***
- Aucune unité ne ***faisant pas partie de la population ne doit y figurer***. Elle doit être strictement égale à ma population cible.

Dans les faits : cette liste est souvent **imparfaite ou n'existe pas**.

⇨ **POPULATION ENQUETEE**

*Ça va avoir un impact direct sur la technique de l'échantillonnage qu'on va sélectionner.*

**Erreur de couverture**
Erreur principalement liée à l'écart entre la base de sondage et la population cible.

| SOUS-COUVERTURE | SUR-COUVERTURE |
|---|---|
| Certaines unités de la population cible ne sont pas sur la base de sondage.<br><br>**Exemple :** Liste des abonnées téléphoniques *(liste rouge : exclusifs mobiles)* | Certaines unités sur la base de sondage ne sont pas dans la population cible.<br><br>**Exemple :** Foyers ayant deux numéros de téléphone. |

**CHOIX DE LA METHODE DES SONDAGES**
Les critères de choix de la méthode :

- ***Objectifs de l'étude*** :
  Cherche-t-on des résultats pour la population entière ou également pour des sous-groupes de la population ?
- ***Nature de la population étudiée***
  - Existe-il une liste à jour ?
  - La population est-elle homogène ou hétérogène ?
  - La population est-elle dispersée géographiquement ?
- ***Contraintes de coût et de délai***
- ***Mode d'administration du questionnaire***

**Représentativité de l'échantillon**
**Représentativité au sens statistique** :
Un échantillon représentatif est el que toutes les personnes faisant parti de la population cible ont la même probabilité de faire partie de l'échantillon.

**Représentativité « structurelle »** :
La structure de l'échantillon est fidèle aux caractéristiques de la population cible.

⇨ On ne raisonne plus en termes de probabilités statistiques mais on va raisonner en termes *« mon échantillon est cohérent en termes de structures avec la population »*

***A-t-on toujours intérêt à avoir un échantillon représentatif ?***

**Exemple :** Etude des attentes vis-à-vis des salles de cinéma, auprès de 300 individus
Dans la population :

- 83% des individus y vont moins d'une fois par mois/jamais
- 13% des individus y vont une/deux fois par mois
- 4% des individus y vont au moins trois fois par mois

| METHODES PROBABILISTES | METHODES EMPIRIQUES |
|---|---|
| Chaque individu de la population a une probabilité connue, différente de zéro d'appartenir à l'échantillon. Ce n'est pas forcément la même probabilité. | L'échantillon est constitué à partir du jugement d'une personne *(choix raisonné)*. On va juger la pertinence des personnes. |
| ▪ Pas de sélection biaisée de l'échantillon<br>▪ Possibilité de calcul d'une marge d'erreur sur les résultats obtenus<br>▪ Connaissance des taux de réponse | ▪ Pas de nécessité d'avoir une base de sondage<br>▪ Cout<br>▪ Rapidité |
| ▪ Nécessité d'avoir une base de sondage<br>▪ Cout | ▪ Représentativité évaluée subjectivement<br>▪ Impossibilité théorique de calculs de marge d'erreur<br>▪ Information sur les taux de réponse. |

## METHODES PROBABILISTES

Il existe 5 types de méthodes probabilistes :

- L'échantillon aléatoire simple
- L'échantillonnage systématique
- L'échantillonnage stratifié
- L'échantillonnage en grappes
- L'échantillonnage à plusieurs degrés

### 1. Sondage aléatoire simple :

| PRINCIPE | AVANTAGES |
|---|---|
| Sélection des individus de telle sorte que chaque membre de la population a une chance égale de figurer dans l'échantillon<br>⇨ ***Avec remise***<br>⇨ ***Sans remise*** | ▪ Représentativité statistique<br>▪ Principe simple<br><br>**INCONVENIENTS**<br>▪ Nécessité d'avoir une liste exhaustive numérotée<br>▪ Efficacité pas toujours optimale<br>▪ Coût |

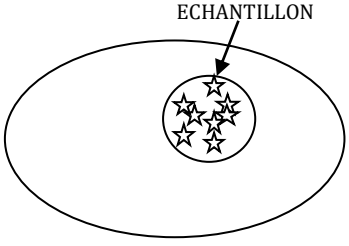
**PROCEDURE**

- Numérotation des individus
- Utilisation d'un programme informatique ou d'une table de nombres aléatoire *(choix au hasard d'un nombre dans la table et d'une façon de se déplacer dans la table)*
- Sélection des individus dont le numéro correspond à ceux obtenus avec la table jusqu'à n.

***Population (N)*** ***Echantillon (n)***

Liste numérotée de 1 à N → n

Probabilité d'appartenance à l'échantillon = n/N



*(taux de sondage)* ↗ POPULATION

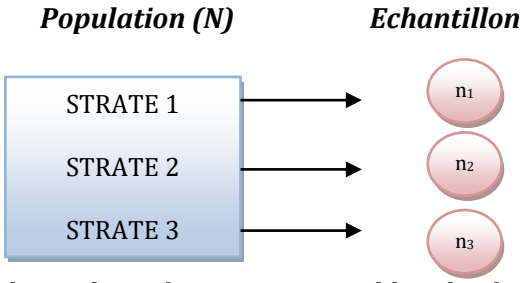
**Population cible ≠ Base de sondage**

## 2. Sondage stratifié :

| PRINCIPE | AVANTAGES |
|---|---|
| Stratification de la population en groupes homogènes mutuellement exclusifs, en fonction de critères corrélés au phénomène étudié, et extraction d'un échantillon aléatoire de chaque strate.<br>▪ Homogénéité à l'intérieur des strates<br>▪ Hétérogénéité entre les strates | ▪ Amélioration de la précision des estimations<br>▪ Possibilité d'avoir une taille suffisante pour des sous-groupes de la population<br><br>**INCONVENIENTS**<br>▪ Nécessité d'avoir une liste<br>▪ Nécessité de connaitre la structure de la population<br>▪ La variable de stratification doit être simple à utiliser et liée au thème de l'enquête |

**PROCEDURE**

- Choix du critère de stratification
- Sélection aléatoire d'un échantillon dans chaque strate : *Proportionnel et non-proportionnel sur représentation ou répartition optimale*
- Agrégation des résultats en fonction des tailles de population de chaque strate.




Taux de sondage identique au variable selon les strates.

**Exemple :** Sondage dans une population d'entreprises

| | POPULATION | ECHANTILLON PROPORTIONNEL *(1/10eme)* | ECHANTILLON NON PROPORTIONNEL | |
|---|---|---|---|---|
| | | | Taux | Taille échantillon |
| **> 10000 salariés** | 20 | 2 | Toutes | 20 |
| **Entre 1000 et 10000 salariés** | 150 | 15 | 1/ 2 | 75 |
| **< 10000 salariés** | 3000 | 300 | 1/15eme | 200 |

## 3. Sondage en grappes :

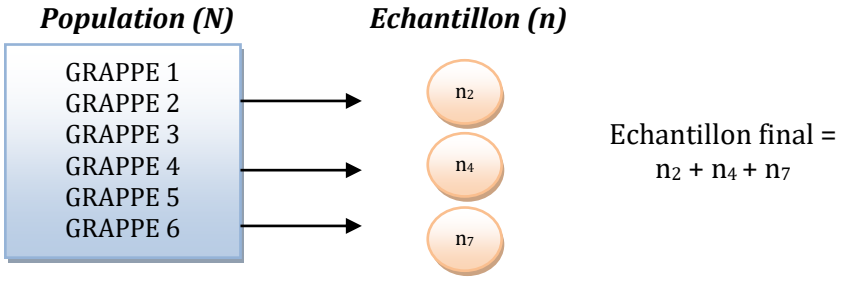
| PRINCIPE | AVANTAGE |
|---|---|
| Découpage de la population en groupes *(grappes)* mutuellement exclusifs, en fonction d'un critère non corrélé au phénomène étudié. Sélection aléatoire de grappes et recensement | ▪ Pas de nécessité d'une liste exhaustive de la population<br>▪ Cout *(dispersion géographique forte)*<br><br>**INCONVENIENTS** |

| | |
|---|---|
| ▪ Hétérogénéité à l'intérieur des grappes<br>▪ Homogénéité entre les grappes | ▪ Précision des résultats *(effet de grappe si homogénéité au sein des grappes)*<br>▪ Contrôle de la taille finale de l'échantillon<br>▪ Calculs d'estimation complexes |

**PROCEDURE**

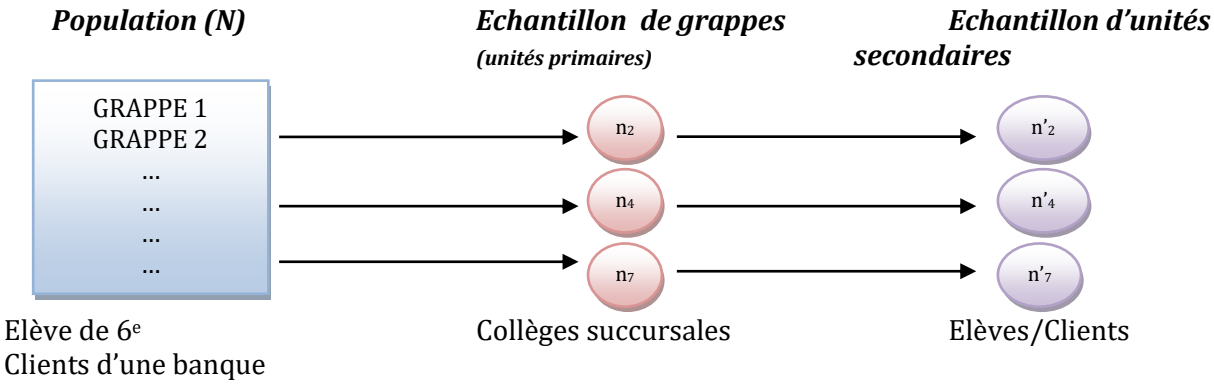
- Découpage de la population en grappes *(unités primaires)*
- Sélection aléatoire d'un échantillon de grappes
- Recensement dans chacune des grappes retenues

**Exemples :** Zones géographiques *(échantillonnage aréolaire)*, voie d'avion, écoles …




### 4. Sondage à plusieurs degrés :

| PRINCIPE | AVANTAGE |
|---|---|
| Principe identique à celui de l'échantillonnage en grappes, mais sélection aléatoire d'un échantillon au sein des grappes, au lieu du recensement<br>⇨ Sondage à deux degrés<br><br>Utilisation possible de plusieurs degrés. | ▪ Pas de nécessité d'une liste exhaustive de la population<br>▪ Cout *(dispersion géographique forte)*<br>▪ Contrôle de a taille de l'échantillon<br><br>**INCONVENIENT**<br>Nécessité d'avoir la liste exhaustive au sein des grappes sélectionnées |



## METHODES EMPIRIQUES

Il existe 4 méthodes empiriques :

- La méthode des quotas
- Echantillonnage boule de neige
- Echantillonnage sur place
- Echantillonnage de volontaires

### 1. *Méthodes des quotas :*

| **PRINCIPE** | **AVANTAGE** |
|---|---|
| Confection « d'une maquette » de la population à partir de critères représentant les variables étudiées et dont on connait la distribution.<br><br>Respect par l'enquêteur de la répartition fixée. | ▪ Pas de nécessité d'une liste<br>▪ Cout<br><br>**INCONVENIENT**<br>▪ Choix des critères limité par les informations statistiques disponibles<br>▪ Nécessité d'avoir des critères facilement identifiables et peu nombreux<br>▪ Risque de surreprésentation des individus facilement accessibles<br>▪ Impossibilité théorique de calculer la précision des estimations. |

***Construction d'un échantillon***

- **Méthode des quotas simples** : Quotas calculés su chacun critère pris indépendamment.
- **Méthode des quotas croisés** : Quotas calculés en prenant en compte le croisement des critères.

**Exemple :** Méthode des quotas simples. Echantillon 1000 personnes seniors.

| Critères | | Structure population | Quotas |
|---|---|---|---|
| Sexe | Hommes | 45% | 450 |
| | Femmes | 55% | 550 |
| Age | 50-64 ans | 49% | 490 |
| | 65-74 ans | 27% | 270 |
| | 75 et + | 24% | 240 |

**Exemple :** Méthode des quotas croisés. Echantillon 1000 personnes seniors

| | Hommes | | Femmes | | Total | |
|---|---|---|---|---|---|---|
| | % | Quota | % | Quota | % | Quota |
| 50-64 ans | 24 | 240 | 25 | 250 | 49 | 490 |
| 65-74 ans | 12 | 120 | 15 | 150 | 27 | 270 |
| 74 ans et + | 9 | 90 | 15 | 150 | 24 | 240 |
| Total | 45 | 450 | 55 | 550 | 100 | 1000 |

***Combinaison avec d'autres méthodes***

Possibilité de combiner la méthode des quotas avec des méthodes aléatoires, utilisées aux premiers stades de l'échantillonnage.

**Exemple :**

Stratification géographique *(stratifications des communes en fonction de leur taille)* → Tirage aléatoire de communes → Méthodes des quotas *(libre ou à partir d'une liste restreinte)*

### Autres méthodes empiriques

#### 2. Echantillonnage boule de neige

- Utilisation de personnes comme source d'identification d'unités additionnelle
- Population d'experts, difficile à identifier.

#### 3. Echantillonnage sur place

- Sur le lieu d'achat ou d'activité *(lorsque la population étudiée est définie par son activité)*
- Enquêtes auprès des clients d'un centre commercial, des clients de chaine de restaurants… Le fait d'aller au même endroit sur les mêmes créneaux horaires n'est pas aléatoire.
- Choix du lieu et de la période
    - ⇨ On n'est pas sur une représentation structurelle : il faut biaiser le moins possible. Ce n'est pas une ***méthode aléatoire.***

#### 4. Echantillonnage de volontaires

- Contexte où les répondants ont le libre choix de participer à un sondage sans avoir été sélectionnées au préalable.
- Questionnaires insérés dans des magazines, sites internet

## TAILLE DE L'ECHANTILLON

### Estimation/Taille d'échantillon

***Principe d'inférence statistique*** : Logique qui conduit à estimer la valeur de paramètres dans une population à partir des résultats observés sur des statistiques d'échantillon.

- ⇨ On va faire des estimations au niveau de la population. Une précision que l'on cherche à avoir.

***Intervalle de confiance*** : Il indique la marge d'erreur lorsqu'on généralise une estimation obtenue sur un échantillon à l'ensemble de la population représentée. On estime la notoriété *(assistée de 81% de votre échantillon)*, ce qu'on cherche d'évaluer c'est la marge d'erreur qui correspond à la valeur au niveau de la population. Ces estimations peuvent être faites sur des échantillons aléatoires stratifiés, pas sur des méthodes empiriques *(car elles ne sont pas aléatoire).*

***Echantillon aléatoire simple, systématique, stratifié***

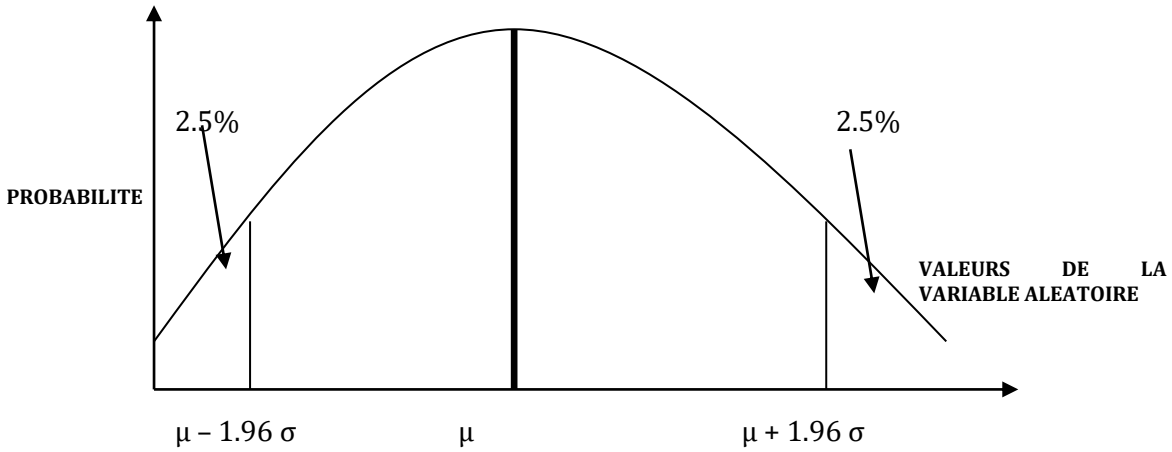
## ESTIMATION (SAS)

### *Estimation de moyenne*

- La moyenne de l'échantillon, $\bar{x}$, est aléatoire
- La moyenne de la population, μ est fixe
- Dans un grand échantillon, la moyenne de la population suit une loi normale
- La moyenne de la distribution est égale à la moyenne de la population

***FORMULE :***

$$\bar{x} - 1.96 \frac{s}{\sqrt{n}} < \mu < \bar{x} + \underbrace{1.96 \frac{s}{\sqrt{n}}}_{\text{PRECISION}}$$




- Population de la taille N. Echantillon de la taille n>30, sinon échantillon trop petit pour m'appuyer sur la loi Normale. n/N <10%
- Soit x la moyenne de l'échantillon, et μ la moyenne de la population
- s est l'écart type de la variable dans l'échantillon
- Intervalle au degré de confiance de 95%
- 1.96 correspond à une probabilité de se tromper à 5%

La précision dépend de la taille de l'échantillon. Plus la taille augmente et plus la précision sera plus fine.

**Exemple :** La chaine de magasins de vêtement Vetir mène une enquête auprès de 400 clients pour déterminer le profil de sa clientèle. Le revenu annuel moyen dans l'échantillon et de 25577 euros *(écart type = 7157).* Quelle est l'estimation du revenu de la clientèle, avec un risque d'erreur de 5% ?
R = x ± [1.96 x (s/Rac(n))] = 25577 ± [1.96 x (7157/Rac(400))] = 25577 ± 701
La précision est de 701 euros (2.7%)

⇨ Le revenu annuel moyen est compris entre 24876 et 26278 euros.

**Exemple :** La chaine de magasins de vêtement Vetir mène une enquête auprès de 1000 clients pour déterminer le profil de sa clientèle. Le revenu annuel moyen dans l'échantillon et de 25577 euros *(écart type = 7157).* Quelle est l'estimation du revenu de la clientèle, avec un risque d'erreur de 5% ?
R = x ± [1.96 x (s/Rac(n))] = 25577 ± [1.96 x (7157/Rac(1000))] = 25577 ± 444
La précision est de 444 euros (1.7%)

**Exemple :** La chaine de magasins de vêtement Vetir mène une enquête auprès de 400 clients pour déterminer le profil de sa clientèle. Le revenu annuel moyen dans l'échantillon et de 25577 euros *(écart type = 7157).* Quelle est l'estimation du revenu de la clientèle, avec un risque d'erreur de 1% ?
R = x ± [2,576 x (s/Rac(n))] = 25577 ± [2,576 x (7157/Rac(400))] = 25577 ± 357,85
La précision est de 922 euros

***Estimation de proportion***

*Estimation de la proportion p d'individus possédant un caractère donné*

- Population de la taille N, échantillon de taille n (n>30)
- F : la proportion dans l'échantillon

- Condition : np (1-p) > 5
- La valeur de précision augmente quand la valeur de la proportion augmente et tends vers 50%.

***FORMULE :***

$$f - 1.96\sqrt{\frac{f\,(1-f)}{n}} < p < f + \underbrace{1.96\sqrt{\frac{f\,(1-f)}{n}}}_{\text{PRECISION}}$$

**Exemple :** Dans le cadre de l'enquête auprès de 400 clients, la chaine de magasins de vêtements Vêtir cherche à estime la part de sa clientèle possédant Internet. Dans l'échantillon, la proportion de clients possédant Internet est de 27%.
P = 27% ± [1,96 x rac(0,27 x 0,73/400)] = 27% ± (4,4%)
La précision est de 4,4%

⇨ La proportion de clients possédant Internet est comprise entre 22,6% et 31,4%

**Exemple :** Dans le cadre de l'enquête auprès de 1000 clients, la chaine de magasins de vêtements Vêtir cherche à estime la part de sa clientèle possédant Internet. Dans l'échantillon, la proportion de clients possédant Internet est de 27%.
La précision est de 2,8%.

⇨ La proportion de clients possédant Internet est comprise entre 24,2% et 29,8%

**Exemple :** Dans le cadre de l'enquête auprès de 400 clients, la chaine de magasins de vêtements Vêtir cherche à estime la part de sa clientèle possédant Internet. Dans l'échantillon, la proportion de clients possédant Internet est de 50%.
La précision est de 4,9%

⇨ La proportion de clients possédant Internet est comprise entre 45,1% et 54,9%

## Taille échantillon

Un compromis entre :

- La précision souhaitée *(marge d'erreur pour un niveau de risque donné)* et le cout induit par l'erreur d'échantillonnage
- Le cout lié à la taille de l'échantillon. Il va dépendre du cout de l'enjeu que représentent les décisions qu'on va prendre sur ces estimations.

***Taille d'échantillon pour estimer une moyenne***

- s : écart type, inconnu mais pouvant être estimé à parti de l'étendue (1/16$^{eme}$). On est obligé de faire une hypothèse soit parce qu'on a déjà fait des enquêtes antérieures soit on a fait des estimations à partir de l'étendue.
- e : précision souhaitée

***FORMULE :***

$$n = \left(\frac{1.96s}{e}\right)^2$$

**Exemple :** L'enseigne de restaurants « Comme chez vous » cherche à connaitre l'âge moyen de sa clientèle. Quelle doit être la taille d'échantillon si l'on souhaite une précision de ± 2 ans, avec un seuil de confiance de 5% ? On prend comme hypothèse que l'âge varie entre 18 et 90 ans.
Approximation de l'écart type : (90-18)/6 = 12
n = (1,96 x 12/2)$^2$ = 138.3

⇨ L'échantillon doit être de 139 personnes.

***Taille d'échantillon pour estimer une production***

- f : la proportion dans l'échantillon
- Le cas e plus défavorable est celui où f = 50%
- e = précision souhaitée
- Degré de confiance de 5%

***FORMULE :***

$$\mathbf{n} = \frac{1.96^2 f(1-f)}{e^2}$$

**Exemple :** L'enseigne de restaurants « Comme chez vous » cherche à connaitre la notoriété spontanée de son enseigne auprès des 25-35 ans. Quelle doit être la taille d'échantillon si l'on souhaite une précision de ± 2% avec un seuil de confiance de 5% ?
$n = 1{,}96^2 \times 0{,}5 \times 0{,}5/0{,}02^2 = 2400$

Quelle doit être la taille d'échantillon si l'on souhaite une précision de ± 3% avec un seuil de confiance de 5%
$n = 1{,}96 \times 0{,}5 \times 0{,}5/0{,}03^2 = 1067$

**Exercice 2 : DELICIOSO**

***Avec un intervalle de confiance de 95%***

$$\mathbf{f - 1.96}\sqrt{\frac{f\,(1-f)}{n}} < \mathbf{p} < \mathbf{f + 1.96}\sqrt{\frac{f\,(1-f)}{n}}$$

- La proportion d'individus préférant le produit actuel est comprise entre 33.3% et 52.7%
- La proportion d'individus préférant le nouveau produit est comprise entre 47.3% et 66.7%

***Avec un intervalle de confiance de 99%***

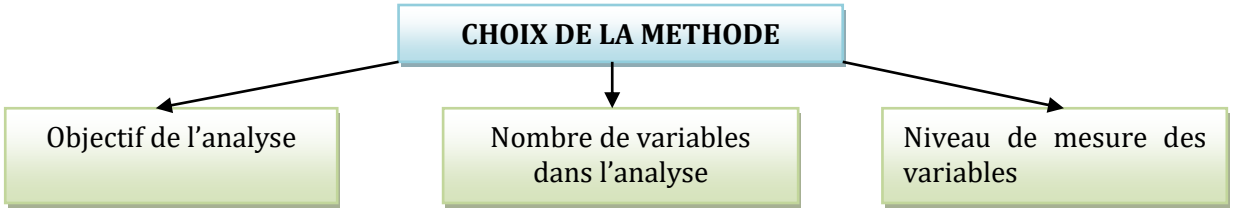
$$\mathbf{f - 2.57}\sqrt{\frac{f\,(1-f)}{n}} < \mathbf{p} < \mathbf{f + 2.57}\sqrt{\frac{f\,(1-f)}{n}}$$

- La proportion d'individus préférant le produit actuel est comprise entre 30.2% et 55.5%
- La proportion d'individus préférant le nouveau produit est comprise entre 44.2% et 69.8%

ANALYSE STATISTIQUE

Réduction du volume des données à un format facilitant l'interprétation.

## CHOIX DE LA METHODE




⇨ Savoir ce qu'on va analyser, déterminer la bonne méthode à employer parmi les différents tests qu'on va voir.

### *Objectif de l'analyse :*

| DESCRIPTIFS | EXPLICATIFS |
|---|---|
| ▪ Décrire un phénomène<br>▪ Visualiser une situation<br>▪ Classer, Catégoriser | ▪ Expliquer des relations causales entre variables indépendantes et variable à expliquer<br>▪ Prédire |

### *Nombre de variables dans l'analyse*

| Analyses UNIVARIEES | Analyses BIVARIEES | Analyses MULTIVARIEES |
|---|---|---|
| Analyse des variables prises une à une<br><br>Exemple : **Moyenne** | Examen des relations entre variables prises deux à deux<br><br>**Exemple :** Régression simple, coefficient de corrélation … | Traitement simultané d'ensemble de variables |

### *Niveau de mesure des variables*

| | | | |
|---|---|---|---|
| Réponses ordonnées ? OUI | NON | **NOMINALE** | CSP |
| Egalité des intervalles entre les catégories ? OUI | NON | **ORDINALE** | Tranches d'âge |
| Existence d'un zéro naturel ? OUI | NON | **INTERVALLE** | Echelles d'attitude |
| | | **RATIO** | Quantité achetée |

| | | |
|---|---|---|
| **RATIO** | Quantité achetée | Variables quantitatives |
| **INTERVALLE** | Echelle d'attitude | |
| **ORDINALE** | Tranches d'âge | Variables qualitatives |
| **NOMINALE** | CSP | |

**Chaque échelle possède les propriétés des niveaux inférieurs.**

***Création d'une base de données***

SPSS :

- Fichier de « base de données »
  - Onglet « affichage des variables » ➔ création/définition des variables
  - Onglet « affichage des données » ➔ saisie des données
- Fichier résultats

# 2. ANALYSES UNIVARIEES

***Description de l'échantillon***

| La tendance centrale | La dispersion |
|---|---|
| Résumé des observations d'une variable en une seule valeur | Variabilité des observations autour de la tendance centrale. |

| | Tendance centrale | Dispersion |
|---|---|---|
| **Nominale** | Mode | Fréquence (tris à plat) |
| **Ordinale** | Médiane | Fractiles |
| **Intervalle Ratio** | Moyenne | Variance, écart type, coefficient de variation, étendue |

| | Tendance centrale | Dispersion |
|---|---|---|
| **Nominale** | **Mode :** Modalité pour laquelle les observations sont les plus nombreuses | **Fréquence :** Nombre d'observations pour chaque modalité<br>**Fréquence relative :** % nombre total d'observations. |
| **Ordinale** | **Médiane :** Valeur qui divise les observations en deux parties égales | **Fractiles :** Partage de la population en catégories d'effectifs égaux.<br>**Quartile :** Partage en 4 catégories égales<br>**Intervalle interquartile :** Ecart entre le premier et le troisième quartile |
| **Intervalle ou ratio** | **Moyenne :**<br>$\overline{x} = \frac{\sum xi}{n}$<br>n = nombre d'observations<br>xi = valeur de l'observation i | **Etendue :** écart entre la valeur maximum et la valeur minimale<br>**Variance :** Moyenne des carrées des écarts à la moyenne<br>$s^2 = \frac{\sum(xi-\overline{x})^2}{n-1}$<br>**Ecart type**<br>$s = \sqrt{s^2}$<br>**Coefficient de variation**<br>Comparaison entre variable. Série dispersée si CV > 25%<br>$CV = \frac{s}{\overline{x}}$ |

- **Description de l'échantillon/estimations** : Faire une enquête de satisfaction, score moyen de satisfaction, faire une étude d'image *(des fréquences, tout dépend du niveau de mesure, présentant l'image de marque),*
- **Identification des données erronées** : Analyser pour déceler s'il y a des erreurs dans notre base de données : erreur atypique. C'est une phase très importante.
- **Bilan des observations extrêmes** : Chercher à identifier les données extrêmes car elles vont avoir un impact sur les tendances centrales
- **Recodification de variables** : Ie on a une variable ordinale avec l'une des modalités qui a très peu d'observations, on a intérêt à agréger cette modalité avec une modalité voisine. On n'aime pas avoir des modalités très disparates.
    - ⇨ **UNE ETAPE INDISPENSABLE, PREALABLE A TOUTE AUTRE ANALYSE**

**Fonctions SPSS** : Les niveaux de mesure s'appellent nominale pour nominale, ordinale pour ordinale et le terme échelle désigne la variable quantitatif (ratio et intervalle).

***Menu*** : Le menu « analyse » : On va trouver toute la famille d'analyse statistique, on va en exploiter qu'un millième. Dans statistiques descriptives c'est là où on trouvera les indicateurs uni variés, khi-deux, …

## CAS APERITIVIA

### Structure d'échantillon :

- **Sexe :** 45,8% d'hommes et 54,2% de femmes
    - ⇨ SPSS : Analyse / Descriptive statisctics / Descriptives : options
    - ⇨ Diagramme : chaque modalité évaluée avec son effectif

- **Répartition géographique** équilibrée entre les 5 régions
    - ⇨ Même commande SPSS sauf pour le graphique : Camembert

- **Age** : entre 20 et 89 ans. Age moyen = 50 ans.
    - ⇨ SPSS : Analyse / Descriptive statistics/ Frequencies (effectifs)
    - ⇨ Histogramme : valeur agrégée ensemble.

### Notoriété :

- **Notoriété spontanée** : 50,3% _ Spontanée : 50,3 ± 3 points
- **Notoriété assistée** : 72,5% _ Assistée : 72,5 ± 2,8 point
    - ⇨ On est dans une méthode aléatoire. Notoriété observée dans l'échantillon. SPSS ne nous fournit pas l'estimation donc il faut la calculer.
- **Notoriété assistée selon la région d'origine des répondants :** Des taux de notoriété plus élevés dans le Sud-Est et le Sud-Ouest que dans les autres régions de France.
    - ⇨ SPSS : Analyse / Descriptive statistics / Tableaux croisés.

### Appréciation du produit :

- Il est compris entre 3,7 et 3,9 avec un risque d'erreur de 5% *(intervalle de confiance à 95%).* En d'autres termes, le score moyen d'attitude est 3,8 ± 0,1 point
    - ⇨ SPSS : Analyse / Descriptive statistics / Explore
      Boite à moustache (boxploits) : Les valeurs qui sortent de ces bornes-là sont des données types atypiques.
    - ⇨ Ecart type de la moyenne standard *(standard error of the mean : sem)*: s/racine n
- La moyenne pour les hommes est de 3,5779 et celle des femmes est de 3,9641. Le terme intervalle veut dire étendue.

***Identification des valeurs extrêmes***

- **Erreur de saisie** => Vérification et correction
- **Erreur probable non vérifiable** => donnée manquante ou élimination du questionnaire si plusieurs anomalies
- **Réalité** => Traitements avec et sans ces données.

**Sensibilité des indicateurs aux valeurs extrêmes**

- *Indicateurs très sensibles* : moyenne, variance, écart type
- *Indicateurs peu sensibles* : mode, médiane, fréquences, percentiles

**Normalité de la distribution**

- **Coefficient de *symétrie* (Skewness)** : Indique si la distribution est symétrique par rapport à a moyenne *(valeur 0)*
  - **s> 0** : distribution concentrée vers les valeurs faibles
  - **s< 0** : distribution concentrée vers les valeurs fortes
    - ⇨ **Un certain nombre de test où on va demander que la distribution soit SYMETRIQUE**.

- **Coefficient *d'aplatissement* (Kurtosis)** : Indique si les observations sont concentrées comme dans la loi normale
  - **k> 0** : distribution est plus concentrée
  - **k<0** : distribution plus aplatie.

# 3. ANALYSES BIVARIEES

***Analyse de relations entre deux variables :***

- **Tests** : existence d'une association entre deux variables
- **Mesures d'association** : force de l'association entre les eux variables
  - ⇨ Ce sont deux choses différentes.

**Les tests statistiques**

***Peut-on extrapoler à la population les conclusions obtenues sur l'échantillon ?***

**Principe d'inférence statistique :**

Logique qui conduit à estimer es propriétés d'une population à partir des résultats observés sur des statistiques d'échantillon

- ⇨ Echantillons
- ⇨ ***Tests d'hypothèses statistiques :*** Donnent une règle permettant de décider si l'on peut rejeter une hypothèse, en fonction des observations relevées sur des échantillons.

**Démarche du test**

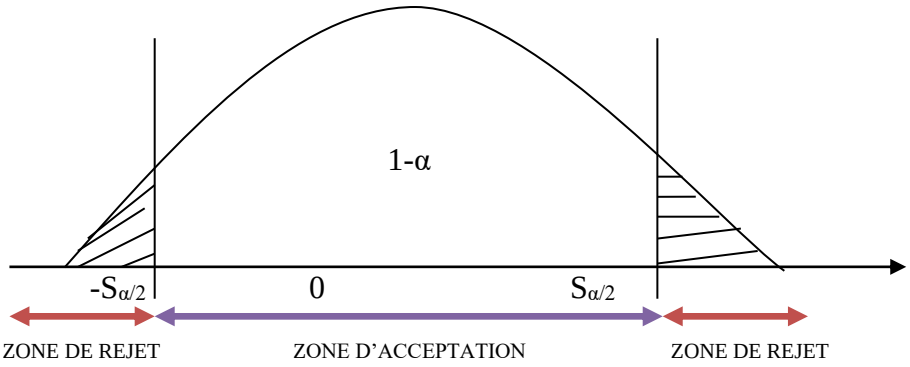
**Exemple :** La société X cherche à savoir si l'implication envers la catégorie « ordinateurs » est la même chez les hommes et chez les femmes L'implication est mesurée sur une échelle en 7 points.

***Résultats dans l'échantillon :***

- Implication moyenne des hommes : 4,1
- Implication moyenne des femmes : 3,6

***Formulation d'hypothèse***

- **$H_0$ : Hypothèse nulle** _ Hypothèse testée *(rejet ou non rejet)*
  Il n'y a pas de différence d'implication entre les hommes et les femmes
- **$H_1$ : Hypothèse alternative**
  Il existe une différence d'implication entre les hommes et les femmes




***Choix de la statistique et du niveau de signification***

- A chaque test est associée une variable *(« statistique du test »)* *(dont la distribution est connue)* dont on connait la loi de probabilité quand $H_0$ est vraie et qui va permettre de prendre une décision
- Choix du seuil de signification α, qui correspond à la probabilité de rejeter $H_o$ alors qu'elle est vraie *(niveau de risque accepté)*
  **(1-α) est appelé le seuil de confiance**

*Si α = 5% : le risque de conclure qu'il y a une différence d'implication entre les hommes et les femmes alors qu'il n'y en a pas est de 5%*

Comparaison de la valeur de la statistique calculée sur **l'échantillon (S)** à la valeur de la loi de probabilité correspondant au seuil de signification choisi ($S_{\alpha/2}$)

- Si $S > |S_{\alpha/2}|$ => Rejet de $H_0$
- Si $S < |S_{\alpha/2}|$ => Pas de rejet de $H_0$

SPSS fournit directement la valeur S et le seuil de signification correspondant, **appelé p-value (P)**

- Si $P < \alpha$ => Rejet de $H_0$
- Si $P > \alpha$ => Pas de rejet de $H_0$

***Les erreurs dans les tests d'hypothèses :***

| Résultat du test / Réalité | **$H_0$ rejetée**<br>Il existe d'une différence | **$H_0$ « acceptée »**<br>I n'existe pas de différence |
|---|---|---|
| **$H_0$ est vraie**<br>Il n'existe pas de différence | Erreur de type I<br>Seuil de signification α | Décision correcte<br>Seuil de confiance (1-α) |
| **$H_0$ fausse**<br>Il existe une différence | Décision correcte<br>Puissance du test (1-β) | Erreur de type II<br>β |

***Tests paramétriques/non paramétriques :***

| **Tests paramétriques** | **Tests non paramétriques** |
|---|---|
| ▪ Reposant sur des hypothèses distributionnelles<br>▪ Données quantitatives<br>▪ Echantillon de taille suffisante | ▪ Libres d'hypothèses distributionnelles<br>▪ Données qualitatives<br>▪ Echantillon de petite taille |

**Nombres d'échantillons**

- Un seul, deux ou plus de deux échantillons
- **Echantillons indépendants** : ayant des caractéristiques différentes sur certaines variables
  **Exemple :** Hommes/Femmes
- **Echantillons appariés** : Ayant des caractéristiques similaires sur l'ensemble des variables.

**Exemple :** Même population interrogée à deux moments différents

| | |
|---|---|
| **TEST DU CHI-DEUX** | Déterminer si deux variables qualitatives sont liées |
| **V DE CRAMER** | Mesurer la force de l'association entre 2 variables qualitatives |
| **COEFFICIENT DE CORRELATION DES RANGS** | Mesurer la force de l'association linéaire entre 2 variables ordinales |
| **COEFFICIENT DE CORRELATION DE PEARSON** | Mesurer la force de l'association linéaire entre 2 variables métriques |
| **COMPARAISON DE MOYENNES (ECHAN. INDEPENDANTS)** | Déterminer si les moyennes de 2 groupes d'individus sont différentes |
| **ANOVA** | Déterminer si les moyennes de plusieurs groupes d'individus sont différentes |
| **REGRESSION SIMPLE** | Déterminer si une variable explicative (métrique) a ou non une influence sur une variable à expliquer *(métrique)* |

# CAS APERITIVIA

***Notoriété :*** Le lien entre la région d'origine et la notoriété est-il significatif ? Formaliser les tests statistiques dans ce langage-là.

## 1. Test du CHI-DEUX

**Objectif** :

Déterminer si deux variables qualitatives sont liées entre elles

**Base** :

Tableau de contingence *(tri croisé),* consistant à comparer les réponses à une question en fonction des réponses à une autre question.

***Tableau de contingence :***

| | | HOMMES | FEMMES | TOTAL |
|---|---|---|---|---|
| Notoriété OUI | Effectif observé | 20 | 40 | 60 |
| | **Effectif théorique** | **30** | **30** | 60 |
| Notoriété NON | Effectif observé | 30 | 10 | 40 |
| | **Effectif théorique** | **20** | **20** | 40 |
| Effectif total | | 50 | 50 | 100 |

⇨ TABLEAU DE TRI CROISES

**Effet théorique** : Si les deux variables sont totalement indépendantes. Indépendantes ie le taux de notoriété est le même pour les hommes que pour les femmes.

**Principe :**

Comparaison des effectifs observés dans l'échantillon aux effectifs théoriques que l'on observerait si les variables étaient indépendantes.

⇨ Plus mes effectifs sont éloignés de mes effets théoriques plus on a un test significatif donc j'aurai une relation de dépendance entre les deux variables.

**Conditions :**

- Effectifs théoriques de chaque cellule > 5
- Nombre de cellules > 4 *(sinon correcteur de Yates)*

S'ils sont inférieurs à 5, on ne pourra pas interpréter le test.

***Cas des petits échantillons :***

Penser au regroupement de modalités *(recodage de variables)*

Conditions affinées :

- N ≤ 20
- 20 < N ≤ 40 et un effet théorique < 5
- N > 40 ET plus de 20% des effectifs théoriques < 5. Penser au regroupement de modalités.

⇨ Test exact de Fisher (Analyse/Descriptive statistics/Crosstabs/Exact (cocher exact))

Il vise à évaluer si la notoriété dépend de la région d'origine. ***La notoriété dépend-elle de la région d'origine ?*** En l'occurrence il y a trois sorties. Quand on fait l'analyse on va fonctionner en trois étapes :

- Je regarde **si les conditions du test sont respectées**. Si je ne le suis
  - Soit j'arrive à détourner les choses et à me mettre dans les conditions du test

- Soit je ne peux pas me mettre dans les conditions dans ces cas je pratique un autre test. Ici, on pratiquera un test exact de Fisher.

- Si je suis bien dans les conditions du test, on va passer au petit 2. On va regarder **si le test est significatif**. Y'a t-il une relation entre les deux variables ? S'il n'y a pas de relations, on s'arrête là.
- S'il y a une relation entre les deux variables on passe au petit 3, on **va expliciter la nature de cette relation.**
    - ⇨ A chaque fois, on appliquera cette méthode.

***Première étape*** :

Respect des conditions ? Les conditions du test sont respectées

***Deuxième étape*** :

Regarder quelle est la probabilité que je me trompe ? Je regarde à quel niveau se situe ma valeur. $H_0$ n'a pas de lien entre la notoriété et la région d'origine. Khi deux de Pearson, si on voit qu'il y a ,000 dans SPSS alors ça veut dire que la P valeur est inférieure à 1 pour mille. Test du Chi-Deux est significatif au seuil de 5% car p << 5%.

## 2. Test du V de Cramer

**Objectif :** Mesurer la force de l'association entre deux variables qualitatives

***Interprétation :***

| | Association |
|---|---|
| > 0,7 | Très forte |
| 0,5 à 0,7 | Forte |
| 0,3 à 0,5 | Modérée |
| 0,1 à 0,3 | Faible |

La valeur de V de Cramer est 0,2 donc la force du lien est faible.

On va interpréter **le fameux tableau croisé** : Effectif = effectif observé. Il y a 188 qui habitent dans le Nord Est, etc. Le taux de notoriété est de 72,5%.

- On interprète TOUJOURS les pourcentages et pas les valeurs absolues sinon ça n'a aucun sens.
- On regarde toujours les pourcentages d'un groupe qu'on compare soit au total soit un autre groupe.
    - ⇨ Quel que soit le sens/ l'interprétation qu'on choisit on arrivera toujours au même résultat.

***Deux différentes interprétations :***

- **Sens colonne** :

  Le taux moyen de notoriété est de 72,5%. Je regarde le taux dans les autres régions : 84,8% dans le Sud Est mais 60,1% dans le Nord Est. On compare les profils entre eux.

- **Sens ligne** :

  Alors que les individus originaires du NO représentent 20,1% de l'échantillon, ils ne représentent que 18,9% de ceux qui connaissent le produit mais 23,3% de ceux qui ne le connaissent pas ie les gens du NO sont surreprésentes dans ce qu'ils ne connaissent

pas mais sous représentée dans ce qu'ils connaissent. Dans le NO, le taux de notoriété est inférieur.

## 3. Coefficient de corrélation de Pearson

**Objectif** :

Mesurer la force de l'association ***linéaire*** entre deux variables métriques*(= quantitative donc de ration et d'intervalle).*Test paramétrique.

**Interprétation**

- Le coefficient varie entre -1 et +1
- Si le coefficient tend vers 0, pas d'association
- Si le coefficient tend vers 1 (-1), association positive/négative
- Visualisation *(graphe : simple scatter)*

**Hypothèses :**

- ***Hypothèse nulle (à tester)*** : $H_0 : r = 0$
- ***Hypothèse alternative*** : $H_1 : r \neq 0$

**Conditions :**

- Variables ont des distributions normales
- Sinon ➔ R de Spearman ou T de Kendall

***Y'a-t-il une corrélation entre l'âge et l'appréciation du produit ?***

***Il y a deux chiffres à regarder*** : sig correspond à la P valeur. Il faut regarder si c'est significatif ou pas ? Oui c'est significatif au seuil de 1 pour mille (,000). Il y a une relation linéaire entre l'âge et l'appréciation du produit. La relation est positive et forte. Donc l'âge est assez fortement positivement lié à l'attitude à l'égard du produit. Plus on est âge plus on a une bonne attitude à l'égard du produit.

⇨ **On doit TOUJOURS qualifier si c'est positive/négative ou fort/faible.**

Graphique qui met en évidence la relation de type linéaire.

## 4. Coefficient de corrélation de Rang _ r de Spearman – t de Kendall

**Objectif :**

Mesurer la force de l'association l'linéaire entre deux variables ordinales.

**Interprétation**

- Le coefficient varie entre -1 et +1
- Si le coefficient tend vers 0, pas d'association
- Si le coefficient tend vers 1 (-1), association positive/négative
- Visualisation *(graphe : simple scatter)*

**Ce sont des coefficients qu'on va utiliser pour évaluer des variables ordinales.**

### Spearman :

**Principe :** Classement des individus su chacune des deux variables, puis calcul de coefficient de corrélation linéaire entre les deux classements.

**Hypothèses :**

- ***Hypothèse nulle (à tester)*** : $H_0 = r = 0$
- ***Hypothèse alternative*** : $H_1 = r \neq 0$

### Kendall :

**Principe :**

Le calcul de tau se repose sur le nombre total d'individus dont le classement sur X et Y est en concordance ou en discordance. La parie d'individus (i.j) est en concordance si : $x_i<x_j$ et $y_i<y_j$ (ou $x_i>x_j$ et $y_i>y_j$)

**Exemples :**

| Individu | X | Y |
|---|---|---|
| A | 3 | 4 |
| B | 2 | 1 |
| C | 1 | 2 |
| D | 4 | 3 |

Paires : (A ;B) en concordance et (A,D) en discordance …

On a une relation significative, positive et modéré entre l'âge et l'appréciation du produit.
Le test est significatif au seuil de 5%. L'appréciation du produit est modérément positivement liée à l'âge.

## 5. Tests de différence de moyenne sur échantillons indépendants

**Objectif :**

Déterminer si les moyennes de deux groupes d'individus sont différentes ou non. On ca chercher à évaluer s'il y a une relation entre une variable nominale binaire et une variable métrique.

⇨ Je vais calculer la moyenne du premier groupe puis du deuxième groupe. Ensuite je vais voir s'il y a une relation entre les deux moyennes

**Hypothèses :**

- ***Hypothèse nulle (à tester)*** : Mes moyennes ne sont pas différentes : $H_0 = m_1 = m_2$
- ***Hypothèse alternative*** : $H_1 = m_1 \neq m_2$

Dans les sorties SPSS, on va avoir des résultats qui vont être réédités dans le cas où les variables sont homogènes pour pouvoir regarder le bon cas de figure.

**Conditions :**

- ***Cas des grands échantillons*** (chaque éch. > 30) ➔ Pas de conditions
- ***Cas des petits échantillons*** : Distribution normale *(symétrique).* Je vais vérifier si mes conditions sont symétriques *(indicateurs de symétrie*). Si je ne peux pas le dire alors je serais obligé d'appliquer le test de U de Mann-Whitney-Wilcoxon.
  ⇨ Sinon Test U de Mann-Whitney-Wilcoxon*(non paramétrique)*

On considérera que la distribution est symétrique si mon Skewness est inférieur à 2 en valeur absolue.

**Variable attestée = variable métrique donc quantitative** :

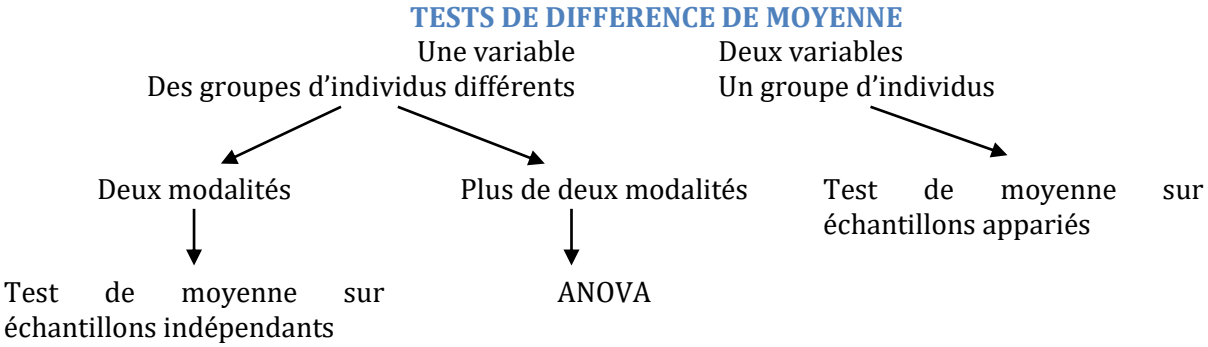
Variable dans lequel on fait le calcul. On vous indique le critère de regroupement qualitatif numérique qui correspond à la variable qualitative binaire. On nous demandera de définir les groupes mais faudra préciser les valeurs spécifiques.

***L'attitude à l'égard du produit varie-t-elle selon le sexe des répondants ?*** Le tableau récapitule les statistiques de chacun des deux groupes. Deux choses importantes à regarder :

- Elle va nous donner les deux tailles de chaque groupe ➔ Est-ce que ça respecte les conditions ? Chaque échantillon est supérieur à 30 donc je n'ai pas besoin d'aller regarder la symétrie pour la distribution de chacun des deux groupes.
- Hypothèse nulle *(à tester)* : Il n'y a pas de différence d'attitude entre les hommes et les femmes. Hypothèse alternative : Il y a une différence d'attitude entre les hommes et les femmes.

**Test de Levene** sur l'égalité des variances. C'est un autre test : l'hypothèse $H_0$ donc les variables sont égales/homogènes. Ici, les variances sont égales car on ne peut pas rejeter l'hypothèse d'égalité de variance avec un risque inférieur à 5%.

Le **test de différence des moyennes** est sur l'autre tableau. On regarde sur la ligne *« Hypothèse de variances égales »*. Ce qui nous intéresse est la significativité. On est significatif car 0,01 << 0,5 % ça veut dire qu'il y a une différence d'attitude entre les hommes et les femmes. Le score moyen d'attitude varie selon le sexe. Les femmes apprécient davantage le produit que les hommes. La différence moyenne est comprise entre -0,61 et -0,16.




## 6. ANOVA *(analyse de variance)*

**Objectif** :

Déterminer si les moyennes de plusieurs groupes d'individus sont différentes ou non.

**Hypothèses :**

- ***Hypothèse nulle*** : $H_0 = m_1, m_2, = m_3$ … toutes les moyennes sont égales
- ***Hypothèse alternative*** : $H_1$ : au moins une moyenne diffère

**Principe :**

Le test repose sur la décomposition de la variance totale en deux composantes : la variance ***intergroupe*** *(entre les modalités)* ***et intragroupe****(au sein de modalités).* Les moyennes sont différentes si la variance intergroupe est supérieure à la variance intragroupe.

**Statistiques** : Test F de Fisher *(variance intergroupe/variance intragroupe)*

**Comparaison par paires :**

L'analyse de variance indique seulement si les moyennes sont ou ne sont pas toutes égales.

Pour effectuer les comparaisons par paire, en conservant le même niveau de risque, utiliser le correcteur de Bonferroni.

**Conditions :**

- Distribution normale *(symétrique)* pour chaque groupe si les effectifs < 30.
- Egalité de variance
- Sinon Test de Kruskal-Wallis *(test non paramétrique)*

***L'attitude à l'égard du produit varie-t-elle selon la région d'origine des répondants ?***

- Première étape : Conditions respectées _ Tous les échantillons sont supérieurs à 30.
- Deuxième étape : 21% de chances de me tromper alors mon test n'est pas significatif, je ne rejette pas (0,211 >> 0,05). Donc les variances sont homogènes.

0,217 : Test d'égalité de variance non significatif au seuil de 5%. On ne peut pas rejeter l'hypothèse d'égalité de variance. Les conditions sont remplies pour le Test F.

**$H_0$ : Les variances sont homogènes.**

***Tableau ANOVA :***

Le test F est significatif au seuil de 5% (,000) et indique que les moyennes ne sont pas toutes égales selon la région d'origine. L'attitude à l'égard du produit dépend de la région d'origine.

J'ai une différence significative d'attitude entre la région NE et SE et la région NE et SO. Mais pas de différence entre le NE et le NO. Les comparaisons par pairs montrent que l'attitude est significativement moins élevée dans le nord-est que dans le centre, le SE et le SO.

**Etapes :**

- ***Etape 1 :*** Les conditions. Groupe supérieur ou pas à 30 *(sinon je regarde s'ils sont symétriques)*. Homogène.
- ***Etape 2*** : ANOVA significatif ie au moins une moyenne qui diffère
- ***Etape 3*** : Comparaison des pairs.

## 7. Régression simple

**Principe** :

La variable dépendante y est une fonction linéaire d'une variable indépendante x selon le modèle : $y = \alpha + \beta x + \varepsilon$

⇨ α ordonné à l'origine et β pente de la droite sont les paramètres à estimer. ε est l'erreur ou composante aléatoire.

On estime le modèle : $y_i = a + bx_i + e_i$

⇨ Trouver la droite qui permette de prévoir Y avec le moins d'erreur possible.

a est la valeur de y lorsque x = 0. Lorsque x augmente d'une unité, y augmente de b unités *(coefficient non standardisé)*
Test du coefficient b $H_0$ : b =0 ; $H_1$ : b≠0
Si b non significatif, la variable n'explique rien.

**Coefficient de détermination de $R^2$**

$R^2$ représente la part des variations de y expliquée par x. Il est compris entre 0 et 1. Si le test du $R^2$ est non significatif, le modèle n'explique rien.

**$R^2$ = SCR/SCT** où SCR : somme des carrés d'écarts expliqués par la régression et SCT : somme des carrées d'écarts totaux.

**Coefficients standardisés (béta)**

- Donnent une indication de l'importance de la variable
- Egal au coefficient de corrélation (*régression simple uniquement)*

**Conditions :**

- Variable dépendante et indépendante métriques
- Linéarité
- $E(\varepsilon_i) = 0$
- $V(\varepsilon_i)$ constante – Homoscédasticité
- Cov (ei, ej) = 0 *(indépendance des erreurs)*
- Les ei suivent une loi normale

***L'âge influence-t-il l'appréciation du produit ?***

Ça nous donne la significativité du $R^2$. Il est significatif au seuil 1 pour mille. L'âge va influencer de manière linéaire l'appréciation du produit. Le $R^2$ est significatif avec un risque d'erreur inférieur de 5%. Le modèle explique 31,3% de la variance. « L'âge explique 31,3% de l'attitude à l'égard du produit ».

# 4. EXERCICE D'APPLICATION

**EXERCICE CINEMA 1 :**

1. **Quel est le test adapté pour répondre aux questions suivantes ?**
2.

**→ Y'a t-il un lien entre la version préférée au cinéma et la préférence pour un type de film *(comédie, horreur …)* ?**

**CHI DEUX**

**→ Y'a t-il un lien entre l'âge et la fréquentation des salles de cinéma ?**

**CORRELATION DE PEARSON** est un test paramétrique qu'on ne peut réaliser qu'avec des variables métriques. Ici en l'occurrence on a affaire à des variables métriques. La fréquence est comptée en nombre par mois *(échelle donc dans SPSS c'est métrique)*. L'âge déclarée qui est métrique aussi.

***On pratique les rangs quand on a des variables ordinales. Si j'ai une variable ordinale et une métrique alors je pratique une corrélation des rangs. Quel autre contexte pour la corrélation des rangs ? On le pratiquerait si bon avait deux variables métriques ne respectant pas Pearson.***

**→ Y'a t-il un lien entre le niveau d'étude et la préférence pour un type de film *(comédie, horreur …)* ?**

**CHI-DEUX**. Corrélation des rangs pas possible car on l'applique qu'avec des variables ordinales.

**→ Y'a t-il un lien entre l'importance accordée à une source d'information donnée et les autres sources ?**

**CORRELATION DES RANGS**

**→ Y'a t-il un lien entre l'expertise perçue et la fréquence des salles de cinéma ?**

**CORRELATION DE PEARSON** car on a deux variables métriques.

---

1. **Définir l'objectif de l'analyse. A quelle question cherche-t-on à répondre ?De quel test il s'agit ?**
2. **Quel est l'objectif poursuivi ? Qu'est-ce qu'on cherche à analyser ? A montrer ?**
3. **Chercher à faire la démonstration**

---

**DEMARCHE :**

- Les conditions sont respectées ?
- Le test est-il significatif ?
- Interprète de manière plus précise si le test est significatif.

---

**TEST 1 :**

- *Objectif poursuivi* :

  Chercher à définir s'il y a une relation entre la version préférée au cinéma et le type de salle fréquenté. Y'a t-il un lien entre la version préférée au cinéma et le type de salle fréquentée ?

- *Analyse* :

  Les conditions du test sont « à peine » respectées. On va considérer qu'on est à la limite mais qu'on poursuit quand même l'analyse. On peut éventuellement faire un regroupement. J'ai le choix entre :

  - J'oppose le multiplexe à la salle de quartier classique/salle art et essai.
  - Soit je préfère garder les trois types de salles après tout dépend de l'analyse recherchée et je vais agréger VO et ça dépend ensemble
  - Ou j'élimine la modalité ça dépend.
    - ⇨ Si on modifie les choses il faut que ça garde un certain sens.

  Je peux aussi faire un test exact mais il faut l'effectuer dans un petit échantillon.

  Là, on va faire comme si on faisait l'analyse : Le test est significatif au seuil de 5% (p<<5%) ça veut dire que l'hypothèse $H_0$ est rejeté. Donc j'ai bien une relation de dépendance entre la version préférée au cinéma et le type de salle fréquentée. La force du lien est faible car ,277.

  Parmi les 100% qui fréquentent les salles d'art et essai, 90% préfèrent la VO. Sachant qu'en moyenne, plus de la moitié préfère la VO.
  Alors que le % moyen d'individus préférant les films en VO est de 53,6%, il est de 90% chez ceux qui fréquentent les salles d'art et d'essai, de 68,8% chez ceux fréquentant les salles de quartier mais seulement de 38,2% chez ceux qui vont dans les salles multiplexe.

**TEST 3**

- *Objectif poursuivi* :

  Y'a t-il une corrélation entre le nombre de films vus à la maison et le nombre de films vus au cinéma ?

- *Analyse* :

  Le test est non significatif au seuil de 5% (regarder à la ligne sig : ,697). La fréquentation des salles de cinéma n'est pas corrélée au nombre de films vus à la maison.

**TEST 4**

- *Objectif poursuivi* :

  Y'a t-il des corrélations entre l'importance accordée à une source d'information donnée et les autres sources ?

- *Analyse* :

**TEST 5**

- *Objectif poursuivi* (expertise réelle des individus) :

  Y'a t-il une corrélation entre les connaissances subjectives et la fréquence de fréquentation des salles de cinéma ?

- *Analyse* :

  Le test est significatif au seuil de 5%. La fréquentation des salles de cinéma est modérément positivement corrélée aux connaissances objectives.

**EXERCICE CINEMA 2 :**

1. **Quel est le test adapté pour répondre aux questions suivantes ?**

**→ Les hommes vont-ils davantage au cinéma que les femmes ?**

**DIFFERENCE DE MOYENNES** car échantillons indépendants. Ça nous permet d'identifier si les moyennes de deux groupes sont différentes. Alors qu'une ANOVA fait la moyenne de plusieurs groupes. La fréquentation des hommes au cinéma est-elle différente de celles des femmes ?

**→ La fréquentation des salles de cinéma dépend-elle du niveau d'étude ?**

**ANOVA** Pas une régression simple car ça implique deux variables métrique. Pas une corrélation des rangs car c'est une nominale mais si ça aurait été une ordinale ça aurait été plausible.

**→ L'âge influence-t-il la fréquentation des salles de cinéma ?**

**REGRESSION SIMPLE** parce qu'on cherche une causalité entre deux variables.

**→ La fréquentation des salles de cinéma varie-t-elle selon le type de salle fréquenté** ?

**ANOVA**

**→ Le nombre de films vus à la maison dépend-il de l'âge ?**

Coefficient de Pearson nous permettra d'évaluer s'il y a une corrélation linéaire entre l'âge et le nombre de films. **REGRESSION SIMPLE** car elle introduit une idée d'influence. La variable explicative qui est l'âge et la variable expliquée qui est le nombre de films.

**TEST 1**

- *Objectif poursuivi* :

  La fréquentation du cinéma varie-telle selon que l'on aime ou non les comédies ?
  - Fréquentation du cinéma : variable quantitative
  - Appréciation des comédies : variables qualitatives binaire.

- *Analyse* :

  Chaque échantillon > 30 ➔ Pas de conditions spécifiques ***(si ce n'est pas le cas il faut voir s'ils sont symétriques ou pas. Si ce n'est pas le cas, faire le test de U Mann)***.
  Le test d'égalité de variance est significatif au seuil de 5% (,001). On peut rejeter l'hypothèse d'égalité de variance.
  *Donc*lire la deuxième ligne ***(hypothèse de variances inégales)*** donc le test de différence de moyenne est significatif au seuil de 5%.

  La fréquentation du cinéma varie selon que l'on aime ou non les comédies. Les personnes qui aiment les comédies vont plus souvent au cinéma que celles qui ne les aiment pas.

**TEST 2**

- *Objectif poursuivi* :

  La fréquence du cinéma varie-t-elle selon que l'on aime ou non les films d'horreur ?

  - Fréquentation du cinéma : variable quantitative
  - Appréciation des films d'horreur : variables qualitatives binaire.

- *Analyse* :

  Coefficient qui faut regarder « asymétrie » : Ça permet d'évaluer si la distribution de ma variable est proche d'une distribution normale. La distribution est vraiment à la limite. Les conditions ne sont pas strictement remplies : on est quand même inférieur à 2. ***En dessous de 1, on est symétrie. Au-delà de 2, on dépasse les limites***. Si j'adopte une condition un peu rigide, qu'est-ce que je peux faire ? La solution c'est de basculer dans un test non paramétrique.

  ⇨ Je passe d'un test paramétrique à un test non paramétrique. Si je ne respecte pas les conditions avec un test paramétrique alors j'utilise son équivalent non paramétrique

  Le test est non significatif au seuil de 5%. La fréquentation du cinéma ne varie pas selon que l'on aime ou non les films d'horreurs (,246).

**Test U de Mann-Whitney-Wilcoxon *(non paramétrique)***

***Principe :***

Compare deux groupes sur les rangs d'une variable ordinale *(ou quantitative ou métrique). On classe les individus en fonction de leurs valeurs puis on calcule leurs rangs.*

***Hypothèses :***

- Hypothèse nulle : Les deux groupes ont des rangs identiques.
- Hypothèse alternative : Les deux groupes n'ont pas des rangs identiques.

***Utilisation :***

- Variable « testée » : ordinale
- Variable « testée » quantitative mais petit échantillon (non-respect symétrie)

**TEST 3**

- *Objectif poursuivi* :

  La fréquentation du cinéma varie-t-elle selon que l'on aime ou on les films policiers ?

  ***Fréquence du cinéma*** : variable quantitative

  ***Appréciation des films policiers*** : variable qualitative binaire

  ⇨ Test de différence moyenne sur échantillons indépendants

- *Analyse* :

  Chaque échantillon > 30. Pas de conditions spécifiques donc elles sont respectées car N>30.

  Le test d'égalité de variance est significatif au seuil de 5%. On peut rejeter l'hypothèse d'égalité de variance car 0,035 < 0,05.

  On passe à la ligne « Hypothèse de variances inégales » : Le test de différence de moyenne est non significatif au seuil de 5% car 0,392 > 0,05. La fréquentation du cinéma ne varie pas selon que l'on aime ou non les films policiers.

**TEST 5**

- *Objectif poursuivi* :

  L'expertise varie-t-elle selon l'importance accordée de la bande annonce dans le choix d'un film ?

  ***Expertise perçue*** : variable quantitative

  ***Importance accordée à la bande annonce*** : variable qualitative

  ⇨ ANOVA

- *Analyse* :

  Deux échantillons < 30 ➔ Validation symétrie de la distribution. C'est symétrique avec 1 en valeur absolue. Skewness < |1|, les distributions peuvent être considérée comme symétriques.

  Test d'égalité de variance non significatif au seuil de 5% (Ho pas rejeté) car 0,336 > 0,05. On ne peut pas rejeter l'hypothèse d'égalité de variance. Les conditions sont remplies pour le test de F.

  Le test F est significatif au seuil de 5% car 0,016 < 0,05 et indique que les moyennes ne sont pas toutes égales. L'expertise perçue dépend de l'importance accordée à la bande annonce dans le choix d'un film.

  ⇨ Dans l'ANOVA il y a une autre condition : homogénéité. Sur le plan théorique il faut que ce soit homogène.

  L'expertise perçue en moyenne pour ceux qui accordent très peu d'importance à la bande annonce est significativement différente (seuil de 1,9) de l'expertise perçue en moyenne de ceux qui accordent moins d'importance. ➔ C'est SIGNIFICATIF.

  Entre ceux qui accordent très peu et beaucoup, c'est non significatif. Ceux qui accordent moyennement et beaucoup ce n'est pas du tout significatif.

  Ici, sur la courbe on voit que ceux qui accordent moyennement et beaucoup n'ont pas beaucoup de différence. Il y a une très grande différence entre eux et ceux qui en accordent très peu.

  ⇨ Les comparaisons par paire (différences de moyenne et signification) montent que l'expertise perçue est significativement moins élevée chez les personnes qui accordent très peu d'importance à a bande annonce que chez celles qui en accordent moyennement (et marginalement que celles qui en accordent beaucoup)

NB : Expertise perçue très élevée = 1 – très faible = 5

Le marginalement : un test est significatif si sa p-valeur est inférieur à 5%.
Sur de petits échantillons, c'est difficile de l'expliquer. Dans ces cas-là on parle de significativité marginale (pas loin des 5%).

**TEST 6 :**

- *Objectif poursuivi* :
- *Analyse* :

ANOVA :

Si ça ne respecte pas les conditions, on regarde l'asymétrie. Si ce n'est pas symétrique alors on utilise le Test de Kruskal-Wallis. Test de U-Mann c'est avec la différence de moyennes.

**Test de Krusal-Wallis (tests non paramétrique)**

***Principe :***

Compare k groupes sur les ranges d'un variable ordinale ou quantitative. Comme le test de U-Mann on classe tous les individus en fonction de la valeur et ensuite on leur attribue une valeur correspondante à leur rang.
Equivalent de l'ANOVA

***Utilisation*** :
Variable testée : ordinale
Variable testée : quantitative mais non-respect normalité homogénéité de variance.

Le test est significatif au seuil de 5% car 0,007 < 0,05. La fréquentation du cinéma varie selon l'importance accordée à la bande annonce.

Le test est significatif au seuil de 5% car 0,013 < 0,05. La fréquentation du cinéma varie selon que l'on accorde très peu ou …

Le test est non significatif au seuil de 5% car 0,479 > 0,05. La fréquentation du cinéma ne varie pas selon que l'on accorde beaucoup ou moyennement d'importance à la bande annonce.

***Si les distributions avaient été symétriques …***

Le test est significatif car 0,012 < 0,05 donc les variances ne sont pas homogènes. On peut rejeter l'hypothèse d'égalité de variance. De façon théorique, les conditions ne sont pas remplies pour le test F. Malgré tout, il y a une forme de tolérance. En revanche, on applique celui de Tamhane pour les comparaisons par paires. Le test F est significatif au seuil de 5% et indique que les moyennes ne sont pas toutes égales. La fréquentation des salles de cinéma varie selon l'importance accordée à la bande annonce dans le choix d'un film.

Comparaison par paires *(regarder différence des moyennes et signification)* avec Tamhane : On voit qu'il y a une différence significative entre très peu et moyennement, très peu et beaucoup. Il y en a pas entre moyennement et beaucoup.
La fréquentation des salles de cinéma est significativement moins élevé chez les personnes qui accordent très peu d'importance à la bande annonce que chez celles qui en accordent moyennement ou beaucoup.

**TEST 4 :**

- *Objectif poursuivi :*
  La fréquentation des salles de cinéma dépend-elle du niveau d'étude ?
  ***Fréquentation du cinéma :*** variable quantitative
  ***Niveau d'étude*** : variable qualitative à plus de deux modalités
  ⇨ ANOVA

1. **Recodage de la variable n° d'étude**
2. **Validation de la symétrie de la distribution**
3. **Si les distributions sont non symétriques ➔ Test de Kruskal-Wallis**
4. **Si le test de Kruskal-Wallis est significatif, mise-en œuvre du test de U Mann-Whitney-Wilcoxon pour toutes les paires.**
   ⇨ On regarde l'effectif et la taille. Si ce n'est pas suffisant je peux soit recoder (pour atteindre la taille d'effectif minimal) soit regarder la taille des symétries. SI c'est symétrie je passe à la suite si ce n'est pas le cas je passe au test de Kruskal Wallis (je fais des comparaisons par paires en mettant en place le test de U-Mann.

⇨ Si mon test d'homogénéité des variances n'est pas significatif je suis dans le respect des conditions et j'utilise le test de Bonferroni. Si le test de variance est significatif ie qu'ils ne sont pas homogène, je ne suis pas dans le respect strict des conditions on peut s'en affranchir.

**TEST 7 :**

- *Objectif poursuivi* :
  L'âge influence-t-il la fréquentation du cinéma ?
  ***Fréquentation du cinéma*** : variable quantitative
  ***L'âge*** : variable quantitative
  ⇨ Régression

Le $R^2$ est significatif avec un risque d'erreur inférieur à 5%. Le modèle explique 11,4% de la variance. « L'âge explique 11,4% de la fréquentation du cinéma ».

Schéma : Plus on vieillit plus on va moins au cinéma.

⇨ L'âge influence négativement la fréquentation du cinéma. Dix années supplémentaires en âge entrainent une diminution de la fréquentation annuelle de presque 4 fois (4 visites en moins).

**TEST 8 :**

- *Objectif poursuivi :*
  Le nombre de films vu à la télévision influence-t-il l'expertise perçue ? ***Attention à la question surtout à l'ordre !***
  ***Expertise perçue*** : variable quantitative
  ***Nombre de films vu à la maison par an*** : variable quantitative
  ⇨ Régression

Le $R^2$ n'est pas significatif avec un risque d'erreur inférieur à 5% car 0,792 > 0,05. Le nombre de films vus à la maison n'influence pas l'expertise perçue.

**TEST 9 :**

- *Objectif poursuivi* :
  La fréquentation du cinéma influence-t-elle l'expertise perçue ? ***Relation de causalité !***
  ***Expertise perçue*** : variable quantitative
  ***Fréquentation du cinéma*** : variable quantitative
  ⇨ Régression

Le $R^2$ est significatif avec un risque d'erreur inférieur à 5% car 0,002> 0,05. Le modèle explique 8,4% de la variance. La fréquentation du cinéma explique 8,4% *(cf ; $R^2$ ajusté)* de l'expertise perçue.

Le codage s'est pas fait dans l'ordre naturel : plus je vais au cinéma plus je me sens expert.
Une fréquentation supplémentaire annuelle du cinéma entraine une augmentation de l'expertise perçue de 0,025 points (cf : A = -0,025)
NB : expertise perçue très élevée = 1 – très faible = 5