

Traitement des données avec EXCEL 2003

Vincent Jalby*

Septembre 2011

1 Saisie des données

Les données collectées sont saisies dans une feuille Excel. Chaque ligne correspond à une observation (questionnaire), chaque colonne à une variable (question ou partie de question).

Observation	Q1	Q2	Q3a	Q3b
Obs 1	1	5	3	2

La première ligne contient le numéro ou l'intitulé de la question : utiliser des noms simples, courts, sans accent, sans espace.

La première colonne contient le numéro de l'observation (questionnaire) : lors de la saisie, chaque questionnaire est numéroté (physiquement) et son numéro retranscrit dans la première colonne.

2 Codage des questions

Le codage dépend de la nature des questions :

- **questions numériques** (age, nombre d'employés, ...) : la valeur est directement saisie :
- **question dichotomique** (sexe, oui/non, ...) : on utilise les valeurs 0/1 (faux/vrai) ou 1/2 (homme/femme).
- **QCM à réponse unique** (situation familiale, ...) : les modalités de réponse sont numérotées ; seul le numéro de la modalité choisie est saisi.
- **QCM à réponses multiples** (moyens de transports utilisés, ...) : elles correspondent à autant de questions dichotomiques que de modalités de réponse. On codera donc ces questions par autant de colonnes que de modalités de réponse. Chaque colonne contiendra la valeur 1 ou 0 selon que l'individu a coché ou non cette modalité.
- **questions avec échelle** (Pas du tout d'accord, tout à fait d'accord, ...) : les modalités sont numérotées de 1 à 5 (dans le cas d'une échelle à 5 modalités) ou de -2 à +2. 1 ou -2 correspond à la modalité la plus négative (pas du tout d'accord), 5 ou +2 à la plus positive (tout à fait d'accord) et 3 ou 0 correspond à la modalité moyenne (ni d'accord, ni pas d'accord).

Obs	Age	Sexe	SitFam	Q4 :Voiture	Q4 :Velo	Q4 : Commun
1	25	1	1	0	1	0
2	29	2	1	1	0	1

Les réponses manquantes (questions non répondues, ou incorrectement) doivent être codées, soit en laissant la cellule vide, soit par des 9/99/999. L'utilisation d'une cellule vide facilite le traitement avec les fonctions incluses dans Excel (en particulier pour les questions numériques). C'est la méthode à préférer. L'utilisation de 9 est à choisir lorsqu'on prévoit d'exporter les données vers SPSS.

Pour faciliter la saisie, il est souhaitable de rajouter deux lignes indiquant les règles de saisie. On peut aussi faire précéder la ligne des intitulés par une ligne reprenant le numéro des questions.

	1	2	3	4 : Moyens de transport		
Obs	Age	Sexe	SitFam	Voiture	Velo	Commun
Valeur	0 à 99	1 ou 2	1 à 5	0/1	0/1	0/1
Manquante	vide	vide	vide	vide	vide	vide
1	25	1	1	0	1	0
2	29	2	1	1	0	1

*Faculté de Droit et des Sciences Économiques, Université de Limoges. Mél : vincent.jalby@unilim.fr

3 Organisation de la feuille

Afin de faciliter la saisie et la consultation des données, il est préférable de scinder la feuille en 3 : 1^{re} colonne (numéro d'observation), 1^{res} lignes (nom des variables, ...), données. Pour cela, il suffit de sélectionner la première cellule contenant une donnée (dans l'exemple précédent, la cellule de la 5^e ligne, 2nde colonne) puis de choisir Fenêtre > Figurer les volets.

Une fois la saisie terminée, il faut supprimer les premières lignes (numéro question, valeur, manquante) pour ne conserver que la ligne des intitulés (et les données dessous). Pour ne pas perdre ces informations, il est préférable de dupliquer la feuille auparavant !

On est alors en présence d'une feuille du type

	A	B	C	D	E	F	G
1	Obs	Age	Sexe	SitFam	Voiture	velo	Commun
2	1	25	1	1	0	1	0
3	2	29	2	1	1	0	1

L'utilisation de filtres automatiques est souvent pratique pour trier/sélectionner des données. Pour cela, il suffit de sélectionner la première ligne (nom des variables) puis de choisir Données > Filtre > Filtre automatique.

Tous les traitements seront fait dans une ou plusieurs feuilles (onglets) séparées. Dans la suite, on supposera que la feuille des données est nommée donnees et la feuille de résultat est nommée resultats.

4 Utilisation des fonctions EXCEL

Pour la plupart des traitements, il existe des méthodes alternatives (Utilitaire d'analyse, tableau croisé dynamique) à l'utilisation des fonctions d'EXCEL. Toutefois, ces dernières peuvent être dans certains cas indispensables.

Pour appliquer une fonction EXCEL à une variable (colonne) :

1. on clique dans une cellule vide de la feuille resultats ;
2. on choisit Insérer > Fonction ;
3. on sélectionne alors une des fonctions (par exemple NB) de la catégorie « Statistiques » ;
4. après validation, un assistant apparaît ;
5. pour indiquer les cellules sur lesquelles la fonction portera, on clique sur l'onglet donnees (en laissant ouvert l'assistant), puis on sélectionne l'ensemble des valeurs de la variables (sans l'intitulé !), c'est-à-dire, les cellules de la ligne 2 à la dernière ligne de la colonne concernée ;
6. après validation, le résultat s'affiche dans la cellule de la feuille resultat.

La formule utilisée est du type

$$= \text{NB}(\text{donnees!B2:B10})$$

Si l'on souhaite réutiliser/recopier cette formule sur les mêmes données (en appliquant d'autres fonctions), il est préférable de transformer les références relatives en références absolues en rajoutant des \$ avant les numéros de ligne et de colonne :

$$= \text{NB}(\text{donnees!\$B\$2:\$B\$10})$$

5 Traitements univariés

5.1 Données valides / Données manquantes

La plupart des traitements qui suivent précisent le nombre de données valides (ou manquantes) dans l'échantillon. Toutefois, il est possible de les déterminer auparavant à l'aides des fonctions NB(), NBVAL(), NBVIDE(), NB.SI(;9) :

- la fonction NBVAL() retourne le nombre de cellules non-vides ;
- la fonction NB() retourne le nombre de cellules numérique (contenant un nombre) ;
- la fonction NB.VIDE() retourne le nombre de cellules vides ;
- la fonction NB.SI(;9) retourne le nombre de cellules contenant le nombre 9.

Lorsque les données manquantes ont été codées par une cellule vide, la fonction NBVAL permet de déterminer le nombre de réponses valides (que ces réponses soient numériques ou textuelles). De même, NB.VIDE permet de déterminer le nombre de réponses manquantes.

Lorsque les données manquantes ont été codées par un texte (du type N/A) (les autres données étant numériques), la fonction NB permet de déterminer le nombre de réponses valides (numériques).

Lorsque les réponses manquantes ont été codées par un 9, alors la fonction NB.SI (; 9) permet de déterminer le nombre de réponses manquantes.

5.2 Variables quantitatives

Elles correspondent aux questions numériques (revenu, âge, ...). Certains traitements s'appliquent aussi aux variables qualitatives ordinales (en particulier issues de questions avec échelle de type Likert) et aux questions dichotomiques codées en 0/1 (variables qualitatives nominales).

5.2.1 Résumés statistiques

Ils correspondent à la moyenne, minimum, maximum, écart-type, somme, etc. Il est possible de les obtenir à l'aide des fonctions Excel MOYENNE, ECARTYPE, MIN, MAX, SOMME.

Il est toutefois plus facile (et rapide) d'utiliser l'**utilitaire d'analyse**¹ :

1. sélectionner Utilitaire d'analyse dans le menu Outils, puis Statistiques descriptives ;
2. pour la plage d'entrée, sélectionner (dans la feuille donnees) la ou les colonnes correspondant aux variables quantitatives à étudier (il suffit pour cela de sélectionner la colonne complète, en cliquant sur la lettre de la colonne) ;
3. cocher « Intitulées en première ligne » et « Rapport détaillé » ;
4. décocher « Niveau de confiance » ;
5. cocher « Plage de sortie » et sélectionner une cellule vide de la feuille resultats (les résultats seront affichés à partir de cette cellule, en dessous et à droite).

5.2.2 Représentation graphique : histogramme

Pour obtenir l'histogramme d'une variable quantitative (continue), il est (quasiment) indispensable d'utiliser l'utilitaire d'analyse. Pour cela, il faut auparavant définir les bornes des intervalles (classes) de valeurs servant à la discrétisation.

Cette opération doit être faite manuellement (en s'aidant des résultats données par Statistiques descriptives : minimum, maximum, etc). Dans une colonne de la feuille resultats, on tape l'intitulé (« Classes ») suivi de la borne supérieure de chaque classe. L'exemple ci-contre correspond à la saisie nécessaire pour obtenir les classes [0, 10], [10, 20] et [20, 30].

Classes
10
20
30

Remarque. *Il est préférable de choisir des intervalles de longueur régulière (de longueur 10 dans l'exemple précédent).*

Pour obtenir un histogramme, il faut ensuite :

1. sélectionner Utilitaire d'analyse dans le menu Outils, puis Histogramme ;
2. pour la plage d'entrée, sélectionner (dans la feuille donnees) la colonnes correspondant à la variable quantitative à représenter (il suffit pour cela de sélectionner la colonne complète, en cliquant sur la lettre de la colonne) ;
3. de la même façon, sélectionner (dans la feuille resultats les cellules correspondants aux classes définies précédemment (y compris la cellule d'intitulé) ;
4. cocher « Intitulées en première ligne » et « Représentation graphique » ;
5. cocher « Plage de sortie » et sélectionner une cellule vide de la feuille resultats (les résultats seront affichés à partir de cette cellule, en dessous et à droite).
6. finalement, ne pas oublier de cocher « Représentation graphique » !

1. Il n'est pas activé par défaut. Pour l'activer, il suffit d'aller dans Outil > Macros complémentaires puis de cocher Utilitaire d'analyse.

5.3 Variables qualitatives

Elles correspondent aux QCM. Elles sont soit ordinales (les modalités sont classées) soit nominales (les modalités ne peuvent pas être classées), en particulier dichotomiques. Certains traitements s'appliquent aussi aux variables quantitatives discrètes (nombre d'enfants, ...).

5.3.1 Tri à plat

Outre les résumés statistiques (*mode* pour les variables nominales, *minimum*, *maximum*, *médiane* pour les variables ordinales, *moyenne* et *écart-type* pour les variables issues d'échelle de Likert) obtenus via les « Statistiques descriptives » de l'utilitaire d'analyse (c.f. section 5.2.1), la présentation de tris à plat permet de décrire complètement une variable qualitative (ou discrète continue).

Il est possible d'obtenir un tri à plat en utilisant les fonctions d'Excel :

1. dans la première colonne, on reporte le numéro des modalités (1, 2, ...)
2. la seconde colonne contient les intitulés des modalités (Homme, Femme, ...)
3. la troisième colonne calcule les effectifs correspondants à l'aide de fonction du type
=NB.SI(<données>,<numéro de la modalité>)
4. la quatrième colonne calcule les pourcentages correspondants à l'aide d'une formule du type
=<colonne 3>/<total>
5. la cinquième colonne calcule les pourcentages valides correspondants à l'aide d'une formule du type
=<colonne 3>/<total valide>
6. l'avant-dernière ligne (sur le même principe que les précédente), affiche les valeurs manquantes (cellules vides);
7. la dernière ligne (colonne C) calcule la somme des effectifs de chaque modalité, hors valeurs manquantes (cellules vides). C'est ce *total exprimé* qui est utilisé dans le calcul de la colonne E;
8. il est possible de rajouter une ligne affichant le nombre total d'observations (valeurs manquantes comprises). Ce total est utilisé dans le calcul de la colonne D.

	A	B	C	D	E
1	Numéro	Intitulé	Effectif	Pourcent.	Pourcent. valide
2	1	Modalité 1	= NB.SI(donnees!D\$2:D\$10,A2)	= C2 / C8	= C2 / C7
3	2	Modalité 2	= NB.SI(donnees!D\$2:D\$10,A3)	= C3 / C8	= C3 / C7
4	
5	4	Modalité 4	= NB.SI(donnees!D\$2:D\$10,A5)	= C5 / C8	= C5 / C7
6		Données manquantes	= NB.VIDE(donnees!D\$2:D\$10)	= C4 / C8	
7		Données valides	= SOMME(C2:C5)	= C4 / C8	
8		Total	= SOMME(C2:C6)		

Mais il est plus simple (et rapide) d'utiliser le « rapport de tableau croisé dynamique » :

1. sélectionner Rapport de tableau croisé dynamique dans le menu Données;
2. à l'étape 1, conserver le réglage par défaut : Liste ou base de données Microsoft Excel;
3. à l'étape 2, définir la plage de données en sélectionnant la colonne correspondante dans la feuille donnees (il suffit pour cela de sélectionner la colonne complète, en cliquant sur la lettre de la colonne);
4. à l'étape 3, on peut préciser où s'affichera le tableau (nouvelle feuille ou feuille existante). Cliquer ensuite sur Fin : un tableau vide s'affiche dans la feuille resultats;
5. faire glisser le nom de la variable de la liste des champs vers la colonne de gauche du tableau (Champs de ligne);
6. faire glisser à nouveau le nom de la variable de la liste des champs vers la zone principale du tableau (Données) : le tri à plat s'effectue automatiquement !
7. changer ensuite dans la première colonne les numéros des modalités par leur nom;
8. pour masquer la modalité « vide » (= donnée manquante²), il suffit de cliquer dans le menu local de la colonne 1 et de décocher « vide ».

2. Cette modalité n'affiche aucune valeur. En effet, les valeurs manquantes « vides » ne sont pas prises en compte. Si nécessaire, pour remédier à cela, il faut sélectionner au point 3, outre la colonne correspondant à la variable étudiée, la première colonne correspondant au numéro d'observation. C'est cette variable qu'il faudra faire glisser (au point 6) vers la zone principale du tableau.

9. pour afficher les pourcentages, faire un clic droit dans la partie principale du tableau (colonne de droite), puis sélectionner Paramètres de champ; cliquer alors sur Options, puis dans le menu Afficher les données sélectionner % du total;

5.3.2 Diagramme en bâtons

A partir du tableau croisé dynamique : il suffit de cliquer sur le bouton « Assistant graphique » de la barre d'outils « Tableau croisé dynamique ». Mais attention, toute modification du graphique sera répercutée sur le tableau et inversement.

Directement : il suffit de sélectionner la zone de données (les deux colonnes contenant, pour la première le nom des modalités et pour la seconde les effectifs ou pourcentages correspondants, y compris la ligne d'intitulé, mais hors ligne total), puis de choisir Insertion > Graphique.

5.4 Variables qualitatives issues d'une question à réponses multiples

Ces variables, codées en 0/1, nécessitent un traitement simultané. On peut soit utiliser les fonctions d'Excel pour effectuer les calculs manuellement, soit utiliser le « rapport de tableau croisé dynamique » de manière un peu détournée : on procède comme précédemment, mais

- au point 3, on sélectionne l'ensemble des colonnes/variables 0/1 correspondantes aux réponses à la question.
- au points 5 et 6, ne pas faire glisser de variable dans la première colonne, mais faire glisser toutes les variables (une par une) vers la zone principale du tableau.
- faire ensuite un clic droit sur chaque ligne puis sélectionner Paramètres de champ et Somme au lieu de Nombre.
- On obtient ainsi les effectifs de chacune des réponses.

6 Traitements bivariés

6.1 Variables quantitatives

Pour étudier le lien entre deux variables quantitatives, il est possible de calculer leur coefficient de corrélation linéaire et/ou de les représenter graphiquement.

Lorsque les variables quantitatives à étudier ne se trouvent pas dans des colonnes adjacentes, il est pratique de recopier ces colonnes dans une nouvelle feuille.

6.1.1 Coefficient de corrélation linéaire

Pour calculer le coefficient de corrélation linéaire, il est nécessaire d'utiliser l'utilitaire d'analyse :

1. sélectionner Utilitaire d'analyse dans le menu Outils, puis Analyse de corrélation;
2. pour la plage d'entrée, sélectionner (dans la feuille donnees) les colonnes correspondant aux variables quantitatives à étudier; il suffit pour cela de sélectionner les colonnes complètes, en majuscule-cliquant sur la lettre de chaque colonne;
3. cocher « Intitulées en première ligne »;
4. cocher « Plage de sortie » et sélectionner une cellule vide de la feuille resultats (les résultats seront affichés à partir de cette cellule, en dessous et à droite).

La matrice obtenue affiche pour chaque couple de variables le coefficient de corrélation linéaire correspondant.

6.1.2 Diagramme de dispersion

Pour représenter graphiquement deux variables quantitatives, on utilise un diagramme de dispersion (ou un *nuage de points*) :

1. sélectionner les 2 colonnes concernées;
2. choisir Insertion > Graphique, puis Nuages de points;
3. à la troisième étape de l'assistant, indiquer les intitulés de chacun des axes (= nom de la variable).

6.2 Variables qualitatives

Pour obtenir un tri croisé et le graphique associé de deux variables qualitatives, on utilise le « rapport de tableau croisé dynamique » :

1. sélectionner Rapport de tableau croisé dynamique dans le menu Données ;
2. à l'étape 1, conserver le réglage par défaut : Liste ou base de données Microsoft Excel ;
3. à l'étape 2, définir la page de données en sélectionnant les colonnes correspondantes dans la feuille donnees. Si les colonnes ne sont pas adjacentes, sélectionner aussi les colonnes séparant les deux variables ;
4. à l'étape 3, on peut préciser où s'affichera le tableau (nouvelle feuille ou feuille existante). Cliquer ensuite sur Fin : un tableau vide s'affiche dans la feuille resultats ;
5. faire glisser le nom de la première variable de la liste des champs vers la colonne de gauche du tableau (Champs de ligne) ;
6. faire glisser le nom de la seconde variable de la liste des champs vers la première ligne du tableau (Champs de colonne) ;
7. faire glisser à nouveau le nom d'une des deux variables³ de la liste des champs vers la zone principale du tableau (Données) : le tri croisé s'effectue automatiquement !
8. changer ensuite dans la première colonne et la première ligne les numéros des modalités par leur nom ;
9. pour masquer la modalité « vide » (= donnée manquante), il suffit de cliquer dans le menu local de la colonne 1/ligne 1 et de décocher « vide ».
10. pour afficher les pourcentages en ligne ou en colonne, faire un clic droit dans la partie principale du tableau, puis sélectionner Paramètres de champ ; cliquer alors sur Options, puis dans le menu Afficher les données sélectionner % par ligne ou % par colonne.

Le graphique (de préférence des effectifs) s'obtient en cliquant sur « Assistant graphique » de la barre d'outils « Tableau croisé dynamique ».

7 En images

7.1 Statistiques descriptives (variable quantitative)

	A	B	C	D	E	F	G	
1	Obs	Age	Sexe	SitFam	Voiture	Velo	Commur	Re
2	1	25						
3	2	29						
4	3	35						
5	4	19						
6	5							
7	6	45						
8	7	33						
9	8	22						
10	9							
11	10	38						
12	11	25						
13	12	29						
14	13							
15	14	19						
16	15	42						
17	16	45						
18	17	33						
19	18	22						
20	19	21						
21	20	38						
22	21	46						

The dialog box 'Statistiques descriptives' is shown over the spreadsheet. It has the following settings:

- Paramètres d'entrée:
 - Plage d'entrée: donnees!\$B:\$B
- Groupées par:
 - Colonnes
 - Lignes
- Intitulés en première ligne
- Options de sortie:
 - Plage de sortie: \$B\$3
 - Insérer une nouvelle feuille:
 - Créer un nouveau classeur
 - Rapport détaillé
 - Niveau de confiance pour la moyenne: 95 %
 - Kième maximum: 1
 - Kième minimum: 1

3. ou mieux : faite glisser la variable « numéro d'observation » ; cela permettra d'avoir le nombre de valeurs manquantes.

7.2 Histogramme (variable quantitative)

	A	B	C	D	E	F
1	Age					
2	10					
3	20					
4	30					
5	40					
6	50					
7	60					
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						

Histogramme

Paramètres d'entrée

Plage d'entrée:

Plage des classes:

Intitulé présent

Options de sortie

Plage de sortie:

Insérer une nouvelle feuille:

Créer un nouveau classeur

Pareto (Histogramme ordonné)

Pourcentage cumulé

Représentation graphique

OK Annuler Aide

7.3 Corrélation (variables quantitatives)

	A	B	C	D	E	F	G	H
1	Age	Revenu						
2	25	1200						
3	29	1300						
4	35	1850						
5	19							
6		2500						
7	45	3200						
8	33	2200						
9	22	1550						
10		1300						
11	38	1400						
12	25	1400						
13	29	1350						
14		1550						
15	19	1100						

Analyse de corrélation

Paramètres d'entrée

Plage d'entrée:

Groupées par: Colonnes Lignes

Intitulés en première ligne

Options de sortie

Plage de sortie:

Insérer une nouvelle feuille:

Créer un nouveau classeur

OK Annuler Aide

7.4 Diagramme de dispersion (variables quantitatives)

	A	B	C	D	E	F	G
1	Age	Revenu					
2	25	1200					
3	29	1300					
4	35	1850					
5	19						
6		2500					
7	45	3200					
8	33	2200					
9	22	1550					
10		1300					
11	38	1400					
12	25	1400					
13	29	1350					
14		1550					
15	19	1100					
16	42	2100					
17	45	2200					
18	33	1900					
19	22	1600					
20	21	1200					
21	38	1950					
22	46						
23	32	1800					
24	21	1100					
25	24	1300					

Assistant Graphique - Étape 1 sur 4 - Type de Graphique

Types standard | Types personnalisés

Type de graphique :

- Histogramme
- Barres
- Courbes
- Secteurs
- Nuages de points
- Aires
- Anneau
- Radar
- Surface
- Bulles

Sous-type de graphique :

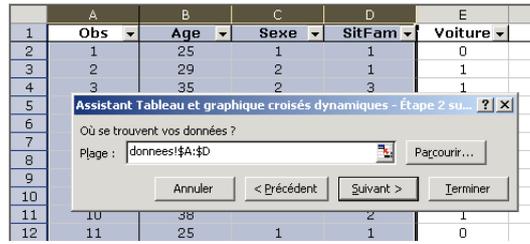
Nuages de points. Compare des paires de valeurs.

Maintenir appuyé pour visionner

Annuler < Précédent Suivant > Terminer

7.5 Tableau croisé dynamique (variable(s) qualitative(s))

7.5.1 Sélection des données



7.5.2 Tableau vide

Faire glisser pour obtenir un tri croisé

Tableau dynamique :
Chaque zone peut recevoir une ou plusieurs variables.

Barre d'outils : permet de définir certaines options ainsi que l'apparence du tableau. Le second bouton permet d'obtenir un graphique dynamique.

Liste des variables : correspond aux colonnes sélectionnées. La variable age (quantitative) ne doit pas être utilisée.

7.5.3 Tableau rempli

Modalités : numéro à remplacer par intitulé.

Menu local : permet de choisir les modalités à afficher

Synthèse par Nombre : affiche le nombre (effectif) de chaque modalité

% du total : affiche le % (au lieu de l'effectif), par rapport au total (de l'échantillon, ou de la ligne ou de la colonne).

Option : affiche le bas de la fenêtre

Clic droit : paramètres de Champ