Enquête consommation - Tableaux croisés dynamiques

1. Chargement des données dans Excel

Souvent les données collectées sont stockées dans un fichier au format texte, dont les informations sont séparées par des espaces, des tabulations ou bien par des virgules (on appelle ce format **csv** — *comma separated values*).

Pour importer de telles données dans Excel, donnez une extension .txt ou .csv à votre document, puis dans Excel : choisir Ouvrir, indiquez « Tous les fichiers » et double cliquez sur votre fichier, indiquez le bon séparateur d'information (« délimiteur ») et vous récupérez vos informations correctement découpées en lignes et colonnes. Chargez le fichier `DonneesConso.txt` de cette façon. Ensuite enregistrez sous un document nommé `enquete.xlsx` de type classeur Excel.

2. Rappels Excel de base

- Calculez la moyenne des budgets des étudiants interrogés
- Comment recopier une formule ?

***Application** : moyenne des consommations de lait et de soda en utilisant uniquement la souris.*

- Quel sens a le symbole $ dans une formule ?
- Comment ramener à deux décimales le nombre affiché dans une cellule ?
- Comment nommer une plage de cellules pour l'utiliser ensuite plus facilement dans des formules ?

***Application** : nommez BD (pour « Base de données ») l'ensemble de votre tableau de données.*

Au passage quelques mots sur le terme « base de données » : pour Excel, une table peut être vue comme une « base de données ». Une telle appellation désigne bien plus que ce que l'on peut faire avec une plage de données dans Excel : une « base de données » digne de ce nom permet de jongler avec plusieurs tables, de les mettre en lien, de faire des requêtes complexes, de les mettre en lien de façon automatique avec d'autres logiciels, etc.

3. Filtres

Pour voir certaines tendances dans les réponses à l'enquête il est intéressant de **trier** les réponses suivant l'un ou l'autre des critères sur lesquels on a posé des questions. De même on peut vouloir **filtrer** les réponses pour ne voir que celles prenant une certaine valeur pour une des questions.

> ***Application** : Sélectionnez l'ensemble de la plage de données et dans l'onglet `Données`, choisissez `Filtrer`. On constate que la ligne de titre a changé, sur chaque colonne on peut maintenant filtrer les entrées affichées.*

Un premier aperçu sur notre tableau ne permet pas d'emblée de voir si les filles choisissent le lait suivant les mêmes critères que les garçons.
Mais après un **tri** du tableau sur la colonne « genre », on voit déjà apparaître que certaines modalités de choix du lait n'apparaissent que chez les filles ou que chez les garçons.
Un **soucis** qu'il est bon de remarquer au passage : triez les données par budget (ou par consommation de soda) croissante, que remarquez-vous ? Pour éviter tout nouveau problème de ce type, utilisez la fonction `Undo` (`CTRL+Z`). Comment éviter ce souci à votre avis ? Expérimentez…

Triez à nouveau les entrées du tableau en fonction du « genre » puis **filtrez** juste après les entrées pour n'avoir que les personnes qui choisissent les yaourts en fonction de la publicité. Un coup d'œil à la colonne « genre » permet alors de voir immédiatement si plus de filles que de garçons utilisent ce critère pour choisir les yaourts ou inversement.

4. Requêtes plus complexes et comptage automatique

Si l'on doit examiner les personnes correspondant à une combinaison de conditions sur les réponses à l'enquête, il est bien plus pratique de demander à Excel de faire de façon automatique le compte de ces personnes pour nous. En effet, la manipulation vue ci-dessus atteint ses limites si l'on veut par exemple connaître le nombre de personnes du département polytech 3 dont le budget est <= 400€ et qui choisissent le lait sur la base de l'emballage : dès que l'on a besoin de trier suivant l'un de ces critères, on perd l'ordre sur le critère précédemment utilisé ☹

Utilisons plutôt la fonction BDNB d'Excel qui va compter de façon automatique le nombre d'entrées répondant à notre critère complexe.

Application : dans les cellules B26 à D26 tapez respectivement Dept , budget et choix Lait (attention l'orthographe doit être **exactement** la même que vos entêtes de colonnes). Puis dans les cellules B14 à D14 tapez : 3 puis <=400 et emballage enfin dans la cellule B28 entrez =BDNB(BD ; "Dept" ; B26:D27) .

Attention, vous noterez ici que le 2ème paramètre de la fonction BDNB doit être une colonne qui contient des valeurs numériques.

Essayez ensuite d'autres critères et découvrez dans l'**Aide de Word** comment il est possible d'avoir les personnes qui choisissent le lait sur la base soit de l'emballage **soit** du prix (attention l'un **ou** l'autre) et qui respectent les autres conditions précitées.

5. Tableaux croisés dynamiques

Le type de requêtes que nous venons de voir permet de vérifier assez rapidement des hypothèses que vous auriez **a priori** sur les résultats de l'enquête. Mais pour une analyse plus approfondie de ces données, nous allons utiliser les tableaux croisés dynamiques. Cet outil va nous permettre de détecter des tendances que nous aurions bien du mal à voir sans lui.

5.1 Tableaux à une entrée

Après avoir sélectionnée toute la plage de données correspondant à l'enquête (à la souris ou en cliquant sur son nom dans les noms de groupes de cellules), dans le menu Insertion, choisissez TblCroiséDynamique (TCD), et posez-le sur la droite de la base de données (même feuille de calcul donc). Comme étiquettes de lignes, indiquez budget (glissé-déposé à la souris) et comme Valeurs indiquez Etudiants (nombre d'Etudiants).

Dans la table créée, vous ne serez plus épatés maintenant de pouvoir utiliser trier ou filtrer les entrées de la colonne de gauche (petite flèche sur la droite de la cellule titre de la colonne). Vérifiez aussi dans le Total général qu'il indique bien ici le nombre de personnes interrogées dans votre enquête.

Pour tracer un graphique des budgets des personnes que vous avez interrogées, il est très pratique de les **regrouper par fourchettes** de budget : cliquez avec le bouton droit de la souris sur une cellule de valeur de la 1ère colonne du TCD et explorez l'option « Grouper… » pour parvenir à regrouper les budgets par tranche de 20€.

Sélectionnez maintenant les cellules pertinentes du TCD pour obtenir un graphique (menu Insertion, groupe Graphiques) sous forme d'histogrammes.

5.2 Tableaux à plusieurs entrées

Supposons que nous voulons maintenant comparer les budgets moyens des garçons et des filles, car nous soupçonnons a priori une différence sur le budget alimentation. Nous voulons faire cette analyse séparément pour chaque département Polytech (*Dept*).

***Application** : dans le TCD utilisez `Dept` comme `Etiquettes de lignes` et `genre` comme `Etiquettes de colonnes`. Comment faire pour que les cellules intérieures du TCD indiquent les budgets moyens de ces catégories ? (attention à ne pas demander le nombre de budgets mais plutôt le budget moyen des personnes concernées).*

Votre a priori sur le budget alimentation des garçons et des filles est-il vérifié d'après le TCD ? Et d'après le graphique ? (et oui, c'est pas beau ça : le TCD et le graphique sont liés ☺).

Profitons de cette liaison pour explorer une autre possibilité : les filtres sur le TCD : dans `Filtre` du rapport déposez le champ `geographie`. En effet, nous voulons vérifiez si la différence de budget moyen s'applique en particulier aux garçons et aux filles issus du Languedoc Roussillon ou bien si sur cette échelle géographique il n'y a pas de différence tangible.

***Application** : en manipulant la première ligne qui s'est ajoutée au TCD (ou bien en manipulant le filtre ajouté en haut à gauche du graphique) trouvez comment faire pour obtenir les budgets respectifs des filles et garçons du Languedoc Roussillon. Quelle conclusion à notre hypothèse ?*

Le même phénomène se produit-il pour les étudiants issus de la région parisienne ? Combien de personnes sont concernées (important pour savoir si ce que l'on observe a un poids statistique ou pas) ?

6. Décider de liens entre plusieurs facteurs de l'étude

Ci-dessus nous avons vu comment se faire une idée visuellement de la relation entre deux facteurs de l'étude (genre et budget moyen). Toutefois, il existe une façon scientifique de décider si les différences observées sont dues au hasard ou sont significatives : le **test statistique du $\chi 2$** (« ki deux ») ou **test d'indépendance**. Dans ce test on compare les **valeurs observées** (ici les effectifs répondant à l'enquête) aux « **valeurs théoriques** » attendues s'il y a indépendance entre les deux facteurs étudiés. On quantifie la différence entre ces deux types de valeurs, et l'on compare avec la différence que l'on attend en fonction de la taille de données (ce qui nous est donné par la loi du $\chi 2$).

6.1 Calcul des valeurs théoriques et quantification des écarts aux valeurs observées

***Application** : nous allons répondre à la question suivante :*

« Est-ce que d'après votre enquête, les garçons et les filles choisissent le lait suivant des critères différents oui ou non ? »

Sélectionnez la plage des donnés correspondant aux réponses à l'enquête, puis demandez la création d'un nouveau TCD, que vous placerez cette fois dans une nouvelle feuille de calcul.

- Quelles colonnes de la BD doivent être mises en regard l'une de l'autre pour la question ci-dessus (débrouillez-vous pour avoir une table plus large que haute) ?

- Quelle colonne doit être utilisée comme valeurs observées dans la table d'analyse ainsi créée ?

- Vérifiez que vous avez bien tous vos effectifs en regardant total général.

- Remplacez dans le TCD de la feuille de calcul `Etiquettes de lignes` et `Etiquettes de colonnes` par un titre correspondant au contexte.

- Les **valeurs observées** pour le test sont les effectifs relevés dans les cellules intérieures du TCD (c-à-d sans les colonnes et lignes de « total »), croisant toute modalité du facteur en ligne avec toute modalité du facteur en colonnes : elles devraient normalement former une zone de 4 cellules de large sur 2 lignes de haut dans cet exemple. Supposons que cette zone de 4x2 est la plage B5:F6 (si ce n'est pas le cas, vous pouvez déplacer votre TCD pour qu'il s'ajuste avec ces cases, soit adapter les formules ci-dessous).

***Note**: il peut être nécessaire de changer les options du TCD pour qu'un 0 apparaisse dans les cellules vides.*

- Définissons maintenant les « **valeurs attendues** » sous H_0 : il s'agit d'une plage de cellules de même taille que les valeurs observées (donc 4x2 cellules), où chaque cellule doit avoir la formule suivante :

= [(Total ligne) * (Total colonne)] / Total global[1]

ici la ligne et la colonne concernées sont celles de la cellule correspondante dans les valeurs observées (c.-à-d. la cellule ayant les mêmes modalités).

Par exemple, supposons que nous mettions notre table des valeurs observées dans les cases H5:K6 , la case H5 doit contenir la formule `=F5*B7/F7`.

- Maintenant essayez d'étendre à la souris la formule pour les 7 autres cases de la plage H5:K6

- Pourquoi rencontrez-vous une erreur (vérifiez les formules qui ont été insérées) ?

- Corrigez la formule de la case H5 pour ne pas avoir cet effet indésirable (indice : symbole `$` placé aux bons endroits). Puis essayez d'étendre à nouveau.

- Calculez les sommes en lignes et en colonnes pour vérifier que vous obtenez bien les mêmes que dans le TCD.

Depuis les deux tableaux obtenus ci-dessus on peut calculer la statistique du $\chi 2$ (écart quadratique moyen entre valeurs observées depuis les données et valeurs attendues si indépendance entre les deux facteurs étudiés). Toutefois ce ne sera pas nécessaire ici, Excel faisant tout le travail pour nous.

6.2 Test du $\chi 2$

Passons maintenant au **test statistique du $\chi 2$** proposé par Excel :

- L'hypothèse que l'on fait a priori (**hypothèse nulle**, notée **H_0**) est que les deux facteurs mesurés dans l'enquête (`genre` et `façon-de-choisir-le-lait`) varient de façon indépendante.
- On choisit a priori un **risque** de se tromper, typiquement alpha=5%.
- En J16 entrez « `Proba(indépendance)` » et en K16 entrez la formule `=CHISQ.TEST(B5:E6;H5:K6)`. Cette formule donne la probabilité qu'une valeur de $\chi 2$ aussi importante que celle induite par les données soit obtenue dans le cas où les deux facteurs étudiés sont indépendants (c-à-d si H0 est vraie).
- Si cette probabilité est plus forte que le risque choisi (5%), alors on ne peut pas remettre en cause en cause H_0 : on accepte H_0, c'est à dire que l'on considère que les deux facteurs étudiés sont indépendants. En revanche, si la proba obtenue est plus faible que le risque alors on sort le champagne : on vient de montrer que nos deux facteurs étudiés sont dépendants l'un de l'autre, autrement dit on tient un résultat intéressant de notre enquête !!!
- Pour formaliser la décision que l'on prend, dans la case J17 entrez « Conclusion » et en K17 tapez `=SI(K16>0,05 ; «indépendance », « dépendance »)` où la valeur 0,05 est le risque d'erreur que l'on est prêt à assumer (c'est la probabilité de se tromper en concluant que les deux variables sont indépendantes).
- Alors, finalement, est-ce que les filles et les garçons choisissent le lait en fonction de critères différents ou pas ? (Si vous avez conclus à l'indépendance entre les facteurs étudiés, c'est que les filles n'ont pas une façon différente des garçons de choisir ce produit d'après vos données).

---

[1] Question subsidiaire : pouvez-vous expliquer pourquoi on utilise cette formule en cas d'indépendance des deux facteurs étudiés ?

7 Mise à jour de la base de données

Supposons que vous avez interrogé de nouvelles personnes pour compléter votre enquête et que vous vouliez donc compléter la base de données. Examinons deux solutions :

1. Sur la ligne suivant la table actuelle entrez les données d'une nouvelle personne interrogée (inventez une situation crédible). Pour ne pas mélanger ces nouvelles données avec les critères que nous avons ajoutés en dessous du tableau, vous pourrez choisir `Insérer` dans le menu contextuel obtenu en cliquant sur le nom de la ligne suivant le tableau. Les TCD ne se mettent pas à jour tout seul, il nous faut demander leur actualisation : onglet `Données -> Actualiser tout`. Rien ne change ? C'est normal, ce n'est pas la meilleure façon de faire (mais comme c'est intuitivement ce que vous feriez, il fallait vous le montrer ☺). Pour que les nouvelles entrées insérées en fin de base de données soient prises en compte, il faut modifier manuellement la zone de données qu'analyse le TCD. Mais supprimez plutôt cette nouvelle ligne que vous avez entrée et lisez ci-dessous.
2. La meilleure façon de procéder pour ajouter des entrées dans une base de données utilisée par des TCD est de demander l'insertion de nouveaux rangs **au milieu** des rangs déjà existants. Essayez. Actualisez ensuite les TCD et vérifiez que la nouvelle entrée a été prise en compte.

Simulez de nouvelles personnes interrogées pour aboutir au fait que le test réponde qu'il y a une **dépendance** entre le *Genre* et le *Choix du Lait*.

- Quel type de personnes interrogées faut-il obtenir pour que l'on conclue le plus rapidement possible à une dépendance ?
- Combien de personnes seraient nécessaires au minimum pour aboutir à la conclusion d'une dépendance entre les deux facteurs étudiés ?

Rappel : pensez à rafraichir le contenu du tableau dynamique après avoir ajouté des lignes au tableau de données.

7 A vous d'enquêter

En insérant de nouveaux TCD et en procédant comme précédemment, pouvez-vous trouver quels facteurs sont significativement (comprendre *statistiquement*) dépendants l'un de l'autre dans nos données ?

Indice : ne croisez pas toutes les colonnes au hasard, comparez d'abord les colonnes entre elles et utilisez intensivement pour cela les tris des colonnes de données suivant un facteur ou un autre pour détecter celles qui semble avoir un lien.