Titre : De la question à la variable

Intervenant : Mélanie Le Goff

Bonjour à toutes et à tous,

Lors d'une enquête épidémiologie, l'un des acteurs-clé est le statisticien. Celui-ci travaille à partir de ce que l'on appelle une **base de données**, c'est-à-dire un tableau regroupant l'ensemble des informations recueillies dans le cadre de l'enquête. Mais comment passe-t-on du questionnaire à l'analyse statistique ? C'est ce que je vous propose de voir ensemble à présent.

Pour chaque question posée dans le questionnaire, le statisticien va définir une **variable**. Une variable, c'est un objet statistique dans lequel on va ranger toutes les valeurs obtenues auprès de tous les enquêtés pour une même question posée. Une variable se définit à l'aide de 3 éléments : un nom, un type et la liste de ses valeurs possibles.

Le choix du **nom de chaque variable** doit être parlant, court de préférence. Evitez au maximum les accents et les signes de ponctuation. Par exemple le nom pour la variable recueillant le sexe de l'enquêté pourrait être 'sexe', celui pour la variable sur la couleur des yeux 'coulyeux'. Faites également bien attention à ne pas mettre le même nom pour deux variables ! Chaque variable doit avoir un nom unique.

En ce qui concerne le **type d'une variable**, il en existe deux grands types en statistique : les variables dites quantitatives et celles dites qualitatives. Les **variables quantitatives** correspondent à des informations que l'on peut mesurer, compter. Cela peut être par exemple : la taille, le poids, l'âge, le nombre d'enfants, etc. Les **variables qualitatives** correspondent à des informations que l'on ne peut pas mesurer, comme le sexe ou la couleur des cheveux. Chacun de ces grands types admet des sous-types.

Les variables quantitatives admettent deux sous-types : les variables quantitatives discrètes et les variables quantitatives continues. Les **variables quantitatives discrètes** admettent un nombre fini de valeurs dans un intervalle donné. C'est très souvent le résultat d'un comptage, comme par exemple le nombre d'enfants par femme, le nombre de voitures par foyer. A l'opposé, les variables **quantitatives continues** admettent un nombre infini de valeurs. C'est souvent le résultat d'une variable qui se mesure, comme l'âge, le poids, la taille, la pression artérielle. Attention à ne pas confondre le type d'une variable avec la façon dont elle est recueillie.

Prenons l'exemple de l'âge. Il s'agit vraiment d'une variable quantitative continue car, de la naissance au décès, l'âge augmente de manière continue, tout le temps (à chaque seconde en soi). Parfois sur les questionnaires, on va poser la question « Quel âge avez-vous ? » et attendre une réponse en années. Mais répondre '45 ans' signifie en réalité que l'on a un âge compris entre 45 et 46 ans ; la variable n'en est donc pas quantitative discrète pour autant, elle reste bien quantitative continue.

Les variables qualitatives admettent, elles, trois sous-types : les variables qualitatives ordinales, les variables qualitatives binaires et les variables qualitatives nominales. Les

variables qualitatives ordinales ont des modalités qui peuvent, comme leur nom l'indique, s'ordonner. C'est, par exemple, le cas pour les échelles de satisfaction ou les stades d'évolution de certaines maladies, comme les stades d'évolution de l'infection par le Virus de l'Immunodéficience Humaine (VIH). D'abord, il y a le stade de primo-infection, évolution sous-jacente sans signe apparent, puis le stade où des symptômes cliniques commencent à apparaitre suite à l'affaiblissement du système immunitaire et, enfin, il y a le stade sida.

Les **variables qualitatives binaires** sont des variables admettant uniquement deux modalités, comme par exemple le sexe (soit homme, soit femme), ou toutes les questions ayant pour réponse oui ou non.

Les **variables qualitatives nominales** sont toutes les variables qualitatives qui ne peuvent pas s'ordonner et ayant trois modalités ou plus, comme la profession ou la couleur de cheveu.

Prenons un exemple. A la question « Quelle est la couleur de vos yeux ? », les propositions de réponses sont : bleu, vert, marron. On ne peut pas dire que le bleu est meilleur que le vert ou que le marron, ce ne peut donc pas être une variable qualitative ordinale. De plus, il y a 3 modalités de réponse, il s'agit donc bien d'une variable qualitative nominale.

Enfin, comment définir la liste **des valeurs possibles pour une variable** ? Et bien, cela va dépendre en grande partie du type de la variable :

- S'il s'agit d'une variable qualitative, la liste des valeurs possibles correspondra à la liste des réponses proposées dans le questionnaire, que l'on appelle des **modalités de réponses**.
  Dans l'exemple que nous venons de donner sur la couleur des yeux, les valeurs possibles sont bleu, vert, marron.
- S'il s'agit d'une variable quantitative continue, la liste des valeurs possibles correspondra à un intervalle de valeurs.
  Par exemple, si je recueille l'âge, les valeurs possibles iront de 0 à 130 ans.
- S'il s'agit d'une variable quantitative discrète, la liste des valeurs possibles correspondra à une liste de nombres.
  Par exemple, à la question « Combien avez-vous de téléviseurs au sein de votre foyer ? », la liste des valeurs possibles sera 0, 1, 2, 3, 4, 5, 6,....

Très souvent, une question posée lors d'une enquête peut permettre la construction d'une seule variable. Mais pas toujours !

Prenons l'exemple de la question suivante : *« Que prenez-vous habituellement lors de votre petit déjeuner ? »*. Il fallait cocher toutes les cases correspondant à des aliments consommés, et ce parmi une liste de 9 aliments. Mettons-nous à la place de l'enquêté. Je lis le premier aliment, à savoir le café. La question que je me pose est, en fait, *« Habituellement le matin, est-ce que je consomme du café ? »*. Si oui, alors je dois cocher la case, si non, je ne la coche pas. Vous comprenez donc que derrière une question avec 9 propositions de réponse se cachent en fait 9 variables binaires, une par aliment ! Restons sur le même exemple, mais mettons-nous cette fois-ci à la place du statisticien. Je prends un dossier au hasard et je vois que l'enquêté n'a pas coché la case « céréales ». Qu'est-ce que cela veut dire ? Que l'enquêté ne consomme pas de céréales au petit-déjeuner. Oui, mais pas uniquement. Car si l'enquêté n'avait pas vu, par oubli, cette proposition de réponse, la case serait également non cochée. Et si l'enquêté n'avait pas voulu répondre à cette question ? La case serait également non cochée. Vous voyez là apparaitre toute la complexité de ce type de questions pour l'analyse statistique : une case non cochée peut

avoir trois significations (non, ne souhaite pas répondre ou n'a pas vu la proposition de réponse) sans aucune différenciation possible.

Le travail du statisticien débute donc bien dès la phase de conception du questionnaire, pour vérifier que la formulation va lui permettre de répondre correctement à l'objectif, et pour s'assurer que, derrière chaque question, il peut bien attribuer une variable

Reprenons l'exemple précédent de la prise alimentaire au petit déjeuner. Si l'objectif principal de l'étude repose sur cette question, le statisticien va alors faire remarquer qu'il faudrait la poser autrement, car il lui faudrait vraiment différencier les réponses « non » des « données manquantes ». Il sera alors préférable de transformer cette liste de cases à cocher en liste de questions avec comme réponses possibles « oui » ou « non », certes plus lassante pour l'enquêté mais qui donne une certitude dans les réponses.

A bientôt !