Optimiser la production statistique : lien entre coût minimal de vérification et finesse de diffusion

Pascal RIVIERE [1]

Insee, Département des Applications et des Projets

Lors d'une enquête auprès d'entreprises, mais aussi dans le traitement de sources administratives, les données que l'Insee reçoit sont loin d'être parfaites : on observe en pratique de très nombreuses anomalies, certaines correspondant à de réelles erreurs, d'autres non. Comme on ne vérifie pas tout à la main, on dispose toujours, d'une manière ou d'une autre, d'un programme qui traite automatiquement les données afin de déterminer quels sont les questionnaires (ou formulaires) « douteux ». Ce sont eux qui seront ensuite revus un par un manuellement. Il s'avère que dans les faits on en vérifie trop, car de nombreux « douteux » (ainsi jugés par le programme) se révèlent corrects dans les faits. Ainsi, même au sein des données en anomalie, il n'est pas nécessaire de tout vérifier.

On peut donc se poser d'abord la question des critères permettant d'optimiser cette vérification. Certains auteurs, comme (Lawrence McKenzie 2000), effectuent ainsi une priorisation en fonction d'un « score » fondé lui-même sur l'impact d'un questionnaire sur les statistiques.

Il est en revanche un sujet qui est moins souvent abordé, celui du critère d'arrêt des vérifications. Dans ce papier, on tente d'abord de mettre en évidence cette notion, et de déterminer précisément ce critère dans un cas volontairement simple (on s'arrête lorsqu'il y a une probabilité $> 1 - \alpha$ que le taux d'erreurs résiduelles soit inférieur à un taux-cible r). La simplicité de ce critère nous permet par la suite, par approximations successives, de déterminer une borne inférieure du coût, indépendante de la proportion inconnue d'erreurs, ce qui fournit une évaluation *a priori* du coût minimal du travail de vérification en fonction du nombre total de questionnaires à contrôler manuellement. En considérant que la vraie cible ne porte pas sur un taux d'erreur global, mais sur un taux d'erreur par domaine de diffusion, on en déduit une relation entre coût minimal et finesse de diffusion.

1. Critère d'arrêt et évaluation précoce du coût minimal

Le traitement d'une enquête se décompose en plusieurs étapes : définition des objectifs, détermination du champ et construction de la base de sondage, tirage de l'échantillon, mise au point et tests du questionnaire, préparation de la collecte, collecte proprement dite (ce qui

[1] Cet article a été écrit en bonne partie en 2001, alors que l'auteur était à l'université de Southampton. Remerciements à Pascal Ardilly, dont les commentaires sur une première version de ce papier ont été précieux.

comprend de nombreuses tâches de gestion), contrôles de cohérence automatiques, codages de libellés, vérification-apurement manuelle, traitements finaux (non-réponses, mais aussi toutes formes de calage ou plus généralement d'application d'un modèle), …

La phase de vérification-apurement (« data editing ») est le plus souvent celle qui est la plus coûteuse en temps, notamment pour les enquêtes entreprises, mais aussi pour le traitement de sources administratives à des fins statistiques, pour lequel les étapes initiales (échantillonnage, questionnaire) ne sont pas les mêmes. Ce travail est donc long, et en même temps on peut toujours s'interroger sur son efficacité : cela vaut-il la peine de tout vérifier jusqu'au bout ? Il est donc intéressant en soi de mettre en place un « critère d'arrêt », c'est-à-dire un critère de décision effectivement calculable à tout moment de la phase de vérification-apurement, permettant de décider si l'on continue ce travail ou non.

Concrètement, nous nous plaçons donc ici dans le contexte d'un ensemble de questionnaires (ou formulaires) jugés en anomalie : en effet, une partie du travail a déjà été faite par la machine, qui a déterminé ce qui était « acceptable » (et peut donc être placé dans la base de données sans aucune intervention humaine) et ce qui était « douteux », ou « en anomalie ». Une équipe de gestionnaires va donc devoir vérifier un par un tous les cas considérés comme douteux.

Idéalement, l'optimisation des vérifications devrait s'appuyer sur un critère de qualité, qui ne peut être qu'un critère de précision : il s'agirait par exemple de pouvoir estimer l'erreur quadratique moyenne, en prenant en compte, toujours dans l'idéal, l'ensemble des composantes de l'erreur. Un tel indicateur présenterait un double intérêt. Il permettrait d'abord de prioriser les opérations, en fonction du "gain potentiel en qualité" apporté par la vérification-apurement, c'est-à-dire l'écart entre l'indicateur actuel et l'espérance de l'indicateur après intervention humaine. Il s'agirait donc toujours du gain procuré par une tâche élémentaire du gestionnaire, comme vérifier une donnée douteuse, vérifier une donnée jugée acceptable par le programme (car rien ne dit qu'il ait raison), rappeler un non-répondant, ... L'indicateur de qualité fournirait en second lieu la possibilité de définir un critère d'arrêt : dès qu'il atteint une valeur-cible que l'on considère comme satisfaisante, on peut arrêter les contrôles manuels.

Construire de tels indicateurs de qualité pour piloter la production n'est pas immédiat. La première raison, la plus classique, réside dans la difficulté qu'il y a à estimer une erreur intégrant de multiples facteurs. D'autre part, même si l'on suppose que la méthodologie existe, cette précision devra être calculée non seulement par variable, mais aussi par domaine de diffusion. En effet, le but n'est pas de produire précisément le total d'une variable pour la France entière mais bien de produire toute une batterie de statistiques, sur des variables et domaines très divers (secteurs d'activité économique, régions, ...). Ainsi l'optimisation de la production statistique (et en particulier l'optimisation de l'arrêt) s'effectue-t-elle en fonction de ces domaines de diffusion. Ceci complique certes la réflexion, mais il n'en demeure pas moins que construire un critère d'arrêt fondé sur la précision attendue est possible. On trouve dans (Rivière 2002) une proposition en ce sens, dans laquelle plusieurs aspects de l'erreur sont pris en considération.

Intuitivement, il est évident que plus la finesse de diffusion en importante, plus le coût de vérification sera élevé : on aura plus de travail si l'on diffuse au niveau zone d'emploi que si l'on reste au niveau régional, par exemple. Or ce coût minimal est proportionnel au nombre de questionnaires (ou formulaires) vérifiés, donc complètement lié, justement, au moment où l'on arrête cette vérification. Malheureusement, avec un critère d'arrêt fondé sur la précision, il est difficile d'évaluer *a priori* ce coût minimal, et donc difficile d'établir un lien entre coût minimal de vérification et nombre de "cases" de diffusion, entre autres parce que l'on ne sait pas, avant de vérifier les données, si elles risquent d'être correctes ou non. Un critère d'arrêt est utilisable par définition *pendant* la production statistique, au fur et à mesure, mais il ne permet d'obtenir aucune évaluation *avant*.

Pour arriver à faire cette évaluation a priori, nous allons être obligés de simplifier le problème de manière drastique.

2. Un cadre simplifié : raisonner sur la proportion d'erreurs et non plus sur la précision

Le but du « data editing » est d'apurer les données en faisant en sorte qu'elles satisfassent un certain niveau de « qualité ». Si nous voulons optimiser ce processus et réduire son coût à qualité fixée, il est important de savoir quel critère nous voulons optimiser. Dans le cas présent, et pour des motifs de simplification, l'objectif que nous allons viser est un taux d'erreur cible[2]. Ainsi nous considèrerons ici que le but du data editing est de « s'assurer » que le taux de questionnaires erronés est au-dessous d'un certain seuil.

Soulignons en effet que le travail des gestionnaires n'est pas qu'un travail de vérification, mais de vérification-apurement : *on fera par la suite l'hypothèse que dès qu'un gestionnaire trouve une erreur, il la corrige*. Ainsi, par hypothèse, si tous les questionnaires sont vérifiés, le taux d'erreur obtenu est 0.

Pur formaliser l'objectif « s'assurer que le taux d'erreur est suffisamment bas », on exprimera que la probabilité que le taux d'erreur soit inférieur à un seuil r est elle-même supérieure à un seuil $1 - \alpha$.

Par exemple, on voudrait être sûrs à 95% que le taux d'erreur est, mettons, inférieur à 4%, auquel cas $r = 0{,}04$ et $\alpha = 0{,}05$.

Le principe général du critère d'arrêt correspondra à la chronologie suivante :

- Les questionnaires (ou formulaires) remplis sont contrôlés manuellement, un par un : ils peuvent être corrects ou erronés.
- Pour chaque questionnaire : s'il est correct, on le laisse tel quel. S'il est erroné, on effectue une correction manuelle. On suppose que cette intervention humaine permet de corriger toutes ses erreurs, de sorte qu'après ces corrections, le résultat est correct.
- Après chaque contrôle individuel, on calcule le majorant de l'intervalle de confiance pour le taux d'erreur (sachant que par hypothèse, il n'y a plus d'erreurs dans les unités déjà vérifiées).
- Si ce majorant est inférieur ou égal au taux d'erreur cible (fixé), le contrôle manuel est arrêté. Sinon le contrôle manuel continue.

Soulignons ici que le contrôle de chaque questionnaire a deux intérêts : d'une part, cela permet de corriger ceux qui sont erronés, donc de réduire le taux d'erreur. Mais d'autre part, cela permet aussi d'avoir une meilleure estimation de l'intervalle de confiance de la proportion d'erreurs restantes.

Dans la suite du présent papier, on notera N le nombre total de questionnaires qui ont été retournés (et qu'il faudra donc vérifier) ; c'est aussi le nombre de répondants. En pratique, ce nombre N est lui-même le résultat d'un processus (échantillonnage, non-réponse, unités hors champ) mais ce n'est pas la question ici.

Les autres principales notations sont:

r = taux d'erreur cible, que l'on se fixe a priori
p = vraie (inconnue) proportion d'erreurs
$1-\alpha$ = niveau de confiance , que l'on se définit aussi a priori (généralement α=0.05)

Pour chaque unité k, X_k est une variable aléatoire binaire qui vaut 1 si le questionnaire n° k est erroné, et 0 s'il est correct.

[2] C'est évidemment trop simpliste, mais cela permettra d'avoir des ordres de grandeur. En toute rigueur, on devrait raisonner en termes d'impact sur la variance. Pour une présentation plus détaillée d'un processus de data editing fondé sur l'impact sur la précision, voir « Impact on mean squared error as a score to handle data editing », Document E2002/10, INSEE/DSE, Décembre 2002.

Le nombre total d'erreurs observées lorsqu'on a traité n questionnaires est donc :

$$T_n = \sum_{k=1}^{n} X_k$$

Et la proportion d'erreurs observée :

$$\hat{p}_n = \frac{1}{n}\sum_{k=1}^{n} X_k = \frac{T_n}{n}$$

Comme nous supposons que toutes les erreurs observées sont corrigées, nous sommes sûrs, qu'après n contrôles, il n'y a aucune erreur parmi les n premiers questionnaires remplis.
Le nombre d'erreurs restantes (qui est donc le nombre total d'erreurs) est :

$$R_n = \sum_{k=n+1}^{N} X_k$$

Si nous sommes capables de définir un intervalle prédictif de type $\left[0\,,\, R_n^{\max}\right]$ pour R_n, on voit que le critère d'arrêt est très facile à définir:

- Pour chaque n, calculer $R_n^{\max}$ tel que $P\left(R_n \geq R_n^{\max}\right) \leq \alpha$
- Si $R_n^{\max} > rN$, continuer
- Sinon arrêter le contrôle manuel

En d'autres termes, le nombre n^* des questionnaires qui auront été contrôlés à la fin du processus est tel que :

$$\forall n < n^*, R_n^{\max} > rN \text{ , et } R_{n^*}^{\max} \leq rN$$

La principale difficulté est maintenant de trouver une bonne approximation de $R_n^{\max}$ pour chaque n. Cela sera fait dans les 2 sections suivantes.

3. Recherche d'un intervalle de confiance pour la proportion d'erreurs inconnue p

Dans cette section, nous étudions ce qui existe dans la littérature en matière d'intervalles de confiance pour une proportion. Le problème vient du fait que les approximations classiques ne seront pas adaptées à de très petites proportions. Et le problème sur lequel on tombe rapidement est le suivant : comment estimer cette proportion si le nombre d'erreurs *observées* à un moment donné est nul, i.e. si tous les questionnaires sont jugés bons ? En pratique, en effet, p n'est pas trop faible, mais l'évaluation du coût minimal sera fondée sur le cas extrême où, au fur et à mesure, on ne trouve pas d'erreurs.

Quelques approximations classiques

Supposons que les X_k soient des variables aléatoires iid de Bernoulli avec pour paramètre inconnu p.

Donc $T_n = \sum_{k=1}^{n} X_k$ est une variable binômiale de paramètres n et p.

Soit $p_L < p < p_U$ un intervalle de confiance exact (Blyth & Still 1983) pour p, où p_L et p_U sont des fonctions de n, de la couverture énoncée de l'intervalle 1-α, et du nombre observé d'erreurs T_n.

Pratiquement, il est possible de calculer un intervalle de confiance exact pour p, sans aucune approximation ou hypothèse, car la distribution (binômiale) est discrète. Le seul problème est que les formules sont longues et compliquées, et donc difficiles à interpréter et généraliser. C'est l'une des raisons pour lesquelles de nombreux auteurs envisagent des approximations de la distribution binômiale. Voir (Blyth 1986) pour les passer en revue.

Nous avons: $E(T_n) = np$, $V(T_n) = np(1-p)$ (1)

et $E(\hat{p}_n) = p$, $V(\hat{p}_n) = \frac{p(1-p)}{n}$ (2)

Par exemple, de l'inégalité de Bienaymé-Tchebyshev, on peut facilement tirer que:

$$P\left(\left|\frac{\hat{p}_n - p}{\sqrt{p(1-p)}}\right| \le \frac{1/\sqrt{\alpha}}{\sqrt{n}}\right) \ge 1-\alpha$$

Mais l'inégalité de Bienaymé-Tchebyshev est bien connue pour donner des limites très frustes, et la longueur des intervalles de confiance est alors hautement surestimée.

L'idée habituelle est d'appliquer une approximation normale. On obtient:

$$P\left(\left|\frac{\hat{p}_n - p}{\sqrt{p(1-p)}}\right| \le \frac{g_{\alpha/2}}{\sqrt{n}}\right) \ge 1-\alpha \quad (3)$$

où $g_{\alpha/2}$ est le quantile à 1 - α/2 de la distribution normale standard.

Par exemple si α=0.05, $g_{\alpha/2} = 1.96$ et $1/\sqrt{\alpha} \approx 4.5$.

Alors la difficulté est de trouver un intervalle de confiance pour p. La façon la plus simple consiste à remplacer p by $\hat{p}_n$ à la racine carrée, ce qui donne « l'approximation simplifiée normale » (Blyth 1986) :

$$\hat{p}_n - g_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} < p < \hat{p}_n + g_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \quad (4)$$

Mais cette approximation peut être médiocre si p est petit. De plus le minorant peut être négatif. Chen (1990) montre que la transformation arc sinus, donnant l'intervalle

$\left|\arcsin(\sqrt{p}) - \arcsin(\sqrt{\hat{p}_n/n})\right| \le \frac{g_{\alpha/2}}{2\sqrt{n}}$, accélère le taux de convergence.

Mais en faisant l'hypothèse que (3) est réalisée, le meilleur intervalle possible peut être obtenu simplement en résolvant les équations :

$$\left|\frac{\hat{p}_n - p}{\sqrt{p(1-p)}}\right| \le \frac{g_{\alpha/2}}{\sqrt{n}} \Leftrightarrow \left(\frac{\hat{p}_n - p}{\sqrt{p(1-p)}}\right)^2 \le \frac{{g_{\alpha/2}}^2}{n} \Leftrightarrow (\hat{p}_n - p)^2 \le \frac{{g_{\alpha/2}}^2}{n} p(1-p)$$

En résolvant la quadratique, on obtient « l'approximation normale directe » (Blyth 1986) :

$$p \in \left[\frac{\hat{p}_n + g_{\alpha/2}^2/2n \pm g_{\alpha/2}\sqrt{\dfrac{\hat{p}_n(1-\hat{p}_n) + g_{\alpha/2}^2/4n}{n}}}{1 + g_{\alpha/2}^2/n} \right] \quad (5)$$

Si la distribution est normale avec une espérance et une variance données par (2), cela constituerait un intervalle de confiance exact pour *p*. Contrairement à l'approximation normale simplifiée, l'approximation normale directe est telle que le minorant est toujours plus grand que 0, et seulement égal à 0 si $\hat{p}_n = 0$.

De plus, si par exemple $\hat{p}_n = 0$, l'approximation normale simplifiée est [0 , 0], ce qui est un intervalle de confiance absurde, alors que l'approximation normale directe est $\left[0 , \dfrac{g_{\alpha/2}^2}{n + g_{\alpha/2}^2} \right]$. Cet intervalle sous-estime la longueur de l'intervalle exact mais il est infiniment mieux que le normal simplifié.

Dans le reste du papier, nous utiliserons donc l'approximation normale directe (qui sera appelée l'intervalle normal), à l'exception de quelques cas particuliers dans lesquels l'intervalle exact est simple à écrire (à savoir les cas extrêmes de nombre d'erreurs : 0, *n*-1 ou *n*). En pratique, une telle approximation est justifiée par le fait que *p* n'est pas excessivement faible : il est important de bien faire la distinction entre la proportion d'erreurs globale, inconnue, qui n'est pas infime, et la proportion d'erreurs observées à un moment donné, qui peut être nulle.

Il se trouve que pour ce qui nous concerne, seul le majorant est important, ce qui signifie que nous avons besoin d'un intervalle de confiance « one-sided ». La seule différence est que nous utiliserons alors g_α au lieu de $g_{\alpha/2}$. Cela donne le majorant suivant pour *p* :

$$p < \frac{\hat{p}_n + g_\alpha^2/2n + g_\alpha\sqrt{\dfrac{\hat{p}_n(1-\hat{p}_n) + g_\alpha^2/4n}{n}}}{1 + g_\alpha^2/n} \quad (6)$$

Remarque sur le cas où la proportion d'erreurs observées est nulle

Il s'agit donc du cas où $\hat{p}_n = 0$

Notons que *X* suit une binomiale avec pour paramètres *n* et *p*, et l'on a $P(X = 0) = (1-p)^n$.

(Blyth 1986) montre que dans ce cas, l'intervalle obtenu est un intervalle « exact » et non une approximation, et peut être écrit comme : $0 < p < 1 - \alpha^{1/n}$

Comme le cas où $\hat{p}_n = 0$ donne nécessairement l'intervalle de confiance le plus petit possible, nous sommes sûrs que le majorant de l'intervalle de confiance de *p est toujours plus grand que le*

majorant trouvé dans le cas extrême où $\hat{p}_n = 0$. En d'autres termes si nous notons $p_n^{\max}$ le majorant de l'intervalle de confiance exact pour p (avec une couverture 1-α), on a :

$$p_n^{\max} \geq 1 - \alpha^{1/n} \tag{7}$$

Cette inégalité jouera un rôle fondamental par la suite.

Correction de population finie

Supposer que les X_k sont indépendants et suivent une distribution de Bernoulli de paramètre p (ce qui implique que le total suit une distribution binômiale) n'est pas exactement conforme à la réalité du processus de vérification, car l'effet de population finie n'est pas pris en compte ; or on a un nombre fini de questionnaires à contrôler. Ainsi si n=N, comme tout a été vérifié et apuré par hypothèse, le nombre d'erreurs restantes est 0, avec une probabilité de 1. La distribution binômiale ne traite pas ce cas, mais elle est parfaitement adéquate si n n'est pas trop grand.

En fait, il est plus pertinent de dire :

1/ On a un nombre fini d'erreurs k dans l'ensemble des questionnaires qui ont été retournés,

2/ On sélectionne un échantillon de n unités

3/ On les contrôle toutes. Le nombre total d'erreurs observé suit alors une distribution hypergéométrique de paramètres N, n, et k/N. Nous noterons $p = k/N$.

Donc l'espérance du taux d'erreur demeure $E(\hat{p}_n) = p$

Mais la variance change, à cause de la correction de population finie. On obtient :

$$V(\hat{p}_n) = \frac{p(1-p)}{n}.(1 - \frac{n}{N}) \tag{8}$$

Il vient, en utilisant l'approximation normale: $P\left(\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \leq \frac{g_\alpha \sqrt{1 - n/N}}{\sqrt{n}} \right) \geq 1 - \alpha$

Notons $m = \frac{n}{1 - n/N}$ (9)

Alors nous avons $P\left(\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \leq \frac{g_\alpha}{\sqrt{m}} \right) \geq 1 - \alpha$ (10)

Ceci est exactement la même formule que (3), où l'on remplace simplement n par m au dénominateur. On peut alors réutiliser le majorant du chapitre précédent, après remplacement de n par m. D'où il vient, à partir de (6), que le majorant de l'intervalle de confiance pour le paramètre k (nombre d'erreurs dans la population d'ensemble) s'écrit :

$$K_n^{\max} = \frac{N}{1 + g_\alpha^2 / m} \left[\hat{p}_n + \frac{g_\alpha^2}{2m} + \frac{g_\alpha}{m} \sqrt{m \hat{p}_n (1 - \hat{p}_n) + \frac{g_\alpha^2}{4}} \right] \tag{11}$$

4. Principe du critère d'arrêt

L'approche précédente sera maintenant utilisée pour trouver des intervalles prédictifs pour le nombre d'erreurs restantes, ce qui permettra de construire un critère d'arrêt des vérifications.

Reprenons pour cela la propriété (11). Celle-ci implique : $P(k - T_n \geq K_n^{\max} - T_n) = \alpha$

Comme le nombre total d'erreurs est *k*, nous avons: $k = T_n + R_n$, où R_n est le nombre d'erreurs restantes dans les questionnaires non-contrôlés.

Donc: $P(R_n \geq K_n^{\max} - T_n) = \alpha$. Cela donne un intervalle prédictif[3] pour R_n, dans lequel le majorant que nous cherchions est: $R_n^{\max} = K_n^{\max} - T_n = K_n^{\max} - n\hat{p}_n$. Alors

$$: \quad R_n^{\max} = \frac{N}{1 + g_\alpha^2 / m}\left[\hat{p}_n - n\hat{p}_n \frac{1 + g_\alpha^2 / m}{N} + \frac{g_\alpha^2}{2m} + \frac{g_\alpha^2}{2m}\sqrt{1 + \frac{4m\hat{p}_n(1-\hat{p}_n)}{g_\alpha^2}}\right]$$

$$\Rightarrow \quad R_n^{\max} = \frac{N}{m + g_\alpha^2}\left[m(1 - \frac{n}{N})\hat{p}_n + g_\alpha^2(\frac{1}{2} - \frac{n\hat{p}_n}{N}) + \frac{g_\alpha^2}{2}\sqrt{1 + \frac{4m\hat{p}_n(1-\hat{p}_n)}{g_\alpha^2}}\right]$$

Le terme $\frac{n\hat{p}_n}{N}$ (taux d'erreur observé multiplié par la proportion d'unités contrôlées) sera généralement négligeable comparé à ½. En enlevant le terme correspondant et en utilisant (9) pour exprimer n en fonction de m, on obtient :

$$R_n^{\max} = \frac{N}{m + g_\alpha^2}\left[\frac{m}{1 + m/N}\hat{p}_n + \frac{g_\alpha^2}{2} + \frac{g_\alpha^2}{2}\sqrt{1 + \frac{4m\hat{p}_n(1-\hat{p}_n)}{g_\alpha^2}}\right] \qquad (12)$$

Le critère d'arrêt fonctionne de la manière suivante :

après chaque contrôle de questionnaire, *n* grimpe d'une unité, on calcule $R_n^{\max}$

si $R_n^{\max} > rN$, continuer les vérifications (ce qui signifie plus précisément : contrôler le *n*+1 ème questionnaire).

si $R_n^{\max} \leq rN$, arrêter les vérifications.

Soulignons que c'est bien sur *rN* et non sur *r(N - n)* que l'on doit raisonner : en effet, on suppose que l'on a corrigé les *n* premières unités, et la cible est relative à un taux d'erreur sur l'ensemble de la population à vérifier.

[3] L'expression « intervalle de confiance » est valable pour un paramètre. Dans notre cas, l'intervalle est associé à une variable aléatoire, et c'est donc un « intervalle prédictif ».

5. Règle d'arrêt et nombre optimal d'unités contrôlées

Nous disposons donc maintenant d'un critère d'arrêt applicable. Mais dans un but de bonne répartition des moyens manuels, on peut se poser la question suivante : quel va être le nombre optimal n^* de questionnaires à vérifier ? On ne peut évidemment pas le déterminer *a priori*, car on ne dispose d'aucune information au départ sur la proportion de questionnaires erronés, mais l'on peut chercher à déterminer des propriétés de n^*.

Par définition, le nombre optimal de questionnaires à vérifier n^* est tel que: $R_{n^*}^{\max} \leq rN$, et $R_n^{\max} > rN \ \forall n < n^*$

D'un point de vue probabiliste, n^* est ce que l'on appelle une variable d'arrêt, variable aléatoire prenant les valeurs 1,2, $\ldots, +\infty$ mesurable relativement à la sigma-algèbre générée par les variables aléatoires $X_1, \ldots, X_n$. Voir par exemple (Siegmund & al. 1968). Ayant calculé les probabilités de ces événements nous serions alors capables de déterminer l'espérance et la variance de cette variable aléatoire. Ceci serait une façon naturelle d'examiner les propriétés de notre critère d'arrêt. Par exemple, une telle approche a été utilisée par (Dalal & Mallows 1988), pour décider quand arrêter le processus de test pour « debugger » un logiciel.

Dans le présent papier, nous ne chercherons pas à obtenir des estimations précises, pas plus que la distribution de n^*, et nous nous restreindrons simplement à des approximations de minorants pour n^*.

Au fond, nous allons chercher à répondre à la question : quel est le nombre minimum de questionnaires à vérifier avant de pouvoir décider d'arrêter les vérifications ?

Et l'on verra que la façon naturelle de rechercher ce minimum est de considérer le cas idéal, extrême, où tous les questionnaires que l'on vérifie se révèlent corrects (c'est-à-dire qu'à tout moment, le taux d'erreur observé est 0).

Méthode (A) : approcher la loi hypergéométrique par une loi binômiale

Dans la section 3, nous avons vu que la distribution hypergéométrique avec les paramètres *N, n* et $k=pN$ avait la même variance que la distribution binômiale de paramètres m et p, où m, défini par (9), prend en compte l'effet population finie.
Par exemple , de (9) nous pouvons tirer que quand *n* tend vers *N*, alors, *m* tend vers l'infini.

Une idée intuitive est alors d'approcher notre distribution hypergéométrique par la distribution binomiale B(*m,p*). Le principal intérêt de cette approximation est que nous connaissons une valeur minimum du majorant de *p*, donné par (7): $p^{\max} \geq 1-\alpha^{1/m}$

Donc, si nous supposons que l'approximation binômiale est valide, dans le cas idéal où aucune erreur n'est observée ($\hat{p}_n=0$), la vérification peut être arrêtée après avoir contrôlé n_0 unités, où n_0 est tel que: $1-\alpha^{1/m_0} \leq r$, et $m_0 = \dfrac{n_0}{1-n_0/N}$.

Alors quelle que soit la valeur de $\hat{p}_n$, n_0 correspond ainsi à la quantité minimum d'unités à vérifier permettant de s'assurer que la proportion d'erreurs restantes est plus petite que *r*, avec la probabilité 1-α.

Il vient : $\alpha^{1/m_0} \geq 1-r \Rightarrow \exp(\ln(\alpha)/m_0) \geq \exp(\ln(1-r))$

Comme α et 1-*r* sont plus petits que 1, leurs logarithmes sont négatifs et nous avons:

$$-\frac{\ln\left(1/\alpha\right)}{m_0} \geq -\ln\left(1/1-r\right) \Rightarrow \frac{\ln\left(1/\alpha\right)}{m_0} \leq \ln\left(1/1-r\right)$$

Alors: $m_0 \geq \dfrac{\ln\left(1/\alpha\right)}{\ln\left(1/1-r\right)}$, ce qui implique que, dans le cas général dans lequel le nombre d'erreurs observé peut différer le 0, le *m optimal*, noté m^*, est tel que:

$$m^* \geq \frac{\ln\left(1/\alpha\right)}{\ln\left(1/1-r\right)} \tag{13}$$

De (9), on tire : $$n = \frac{N}{1+N/m} \tag{14}$$

Comme *n* est une fonction croissante de *m*, on obtient:

$$\boxed{\frac{n^*}{N} \geq \frac{1}{1+\dfrac{N\ln\left(1/1-r\right)}{\ln\left(1/\alpha\right)}}} \tag{15}$$

Si *r* est petit, et si α=0.05, nous obtenons une proportion minimale approchée de questionnaires à vérifier (comme ln(1/α)=ln(20)≈3):

$$\boxed{\frac{n^*}{N} \geq \frac{1}{1+\dfrac{rN}{3}}} \tag{16}$$

Méthode (B) : remplacer $\hat{p}_n$ par *p* dans l'expression de $R_n^{\max}$

L'idée ici est d'utiliser $R_n^{\max}$ comme défini dans (12). Implicitement, cela signifie utiliser une approximation normale. Mais la principale approximation ici consistera à remplacer $\hat{p}_n$ par *p* dans la formule $R_n^{\max}$, ce qui donnera une valeur moyenne du nombre optimal de vérifications n^*.

Comme nous recherchons seulement un minorant, nous n'avons pas besoin d'une égalité. L'inégalité $rN \geq R_{n^*}^{\max}$ est suffisante. Cependant, nous pouvons remarquer que $R_{n^*}^{\max} \leq rN$, et $R_{n^*-1}^{\max} > rN$, ce qui signifie qu'avec une légère modification, l'inégalité est inversée. Ceci implique que les deux côtés de l'inégalité sont proches, et donc nous avons aussi $R_{n^*}^{\max} \approx rN$.

Il vient de $rN \geq R^{\max}_{n^*}$, en réutilisant l'équation (12), et en notant $m^* = \frac{n^*}{1 - n^*/N}$:

$$r(m^* + g_\alpha{}^2) \geq p\frac{m^*}{1 + m^*/N} + \frac{g_\alpha{}^2}{2} + \frac{g_\alpha{}^2}{2}\sqrt{1 + \frac{4m^* p(1-p)}{g_\alpha{}^2}} \tag{17}$$

Alors, comme $p \geq 0$: $rm^* - g_\alpha{}^2(\frac{1}{2} - r) \geq \frac{g_\alpha{}^2}{2}\sqrt{1 + \frac{4m^* p(1-p)}{g_\alpha{}^2}}$

Après élévation au carré des deux côtés, nous obtenons:

$$r^2 m^{*2} - r(1-2r)g_\alpha{}^2 m^* + g_\alpha{}^4(\frac{1}{2} - r)^2 \geq \frac{g_\alpha{}^4}{4}\left(1 + \frac{4m^* p(1-p)}{g_\alpha{}^2}\right)$$

$$\Rightarrow r^2 m^{*2} - \big(r(1-2r) + p(1-p)\big)g_\alpha{}^2 m^* - g_\alpha{}^4 r(1-r) \geq 0$$

$$\Rightarrow r^2 m^{*2} \geq \big(r(1-2r) + p(1-p)\big)g_\alpha{}^2 m^*$$

Donc: $$m^* \geq \frac{g_\alpha{}^2}{r}\left(1 - 2r + \frac{p(1-p)}{r}\right) \tag{18}$$

Nous pouvons obtenir un minorant pour m^* **qui ne dépend pas de p** :

$$m^* \geq \frac{g_\alpha{}^2}{r}(1-2r) \tag{19}$$

De (19) et (9), comme n est une fonction croissante de m, nous pouvons tirer notre principal résultat, qui est un minorant de la proportion optimale de vérifications :

$$\boxed{\frac{n^*}{N} \geq \frac{1}{1 + \frac{rN}{g_\alpha{}^2(1-2r)}}} \tag{20}$$

Si r est petit, et si α=0.05, alors $g_\alpha = 1.6449,\ g_\alpha{}^2 \approx 2.7$, et nous obtenons:

$$\boxed{\frac{n^*}{N} \geq \frac{1}{1 + \frac{rN}{2.7.(1-2r)}}} \tag{21}$$

De manière intéressante, la formule (21) obtenue avec la méthode (B) est très similaire à (16) obtenue avec la méthode (A), même si les approximations faites étaient très différentes.

La seule différence vient du fait que dans (16), nous avons $\frac{rN}{3}$, alors que (21) donne $\frac{rN}{2.7(1-2r)}$.

Par exemple si r=0.05, 2.7(1-2r) = 2.43. D'où il vient que pour les faibles valeurs de r, les ordres de grandeur sont identiques.

Comme r sera généralement petit, les deux formules mettent en lumière le fait que le nombre minimum de vérifications est essentiellement une fonction du nombre d'erreurs cible (rN), le taux d'erreur cible r lui-même jouant un rôle mineur. En conséquence il est clair que de petites populations sont toujours très coûteuses en ce qui concerne les contrôles manuels.

6. Conséquences sur la quantité minimale de contrôles à effectuer

Comme le coût des vérifications peut être considéré comme proportionnel au nombre de questionnaires qui sont contrôlés par les gestionnaires d'enquête, la précédente section devrait donner un ordre de grandeur du coût. En fait, ce n'est pas réellement le cas : les statisticiens calculent des statistiques non seulement pour toute la population, mais aussi et surtout pour des sous-populations, ou domaines (zones géographiques, secteurs d'activité, …). Et l'on ne peut pas dire à l'utilisateur de statistiques que même si le taux d'erreur est très élevé dans sa région, cela n'a pas d'importance car le taux global d'erreur est acceptable : comme nous le savons, la qualité est l'aptitude à satisfaire des besoins, ce qui signifie que la qualité des statistiques dépend entièrement des niveaux de diffusion qui présentent de l'intérêt pour les utilisateurs futurs (qu'ils soient internes à l'institut statistique ou non).

Donc, si nous considérons que réduire la quantité d'erreurs est un bon objectif, le but de la vérification ne peut pas être un taux d'erreur global : il doit être un taux d'erreur pour chaque domaine cible.

Si nous considérons que les domaines constituent une partition du champ (c'est le cas avec la partition géographique ou en secteurs d'activité économique) alors le critère d'arrêt devra être appliqué à chaque domaine : dans un domaine donné D, dès que le majorant de l'intervalle de confiance de p_D (vrai taux d'erreur) est plus petit que le taux d'erreur cible r_D, le contrôle peut être arrêté dans ce domaine.

Quelles conséquences en terme de quantité de vérifications ?

Supposons que nous ayons une partition du champ en H domaines cibles. Ce nombre H reflète donc ici ce que nous avons appelé « finesse de diffusion », ce n'est rien d'autre que le nombre de cases de diffusion des tableaux statistiques fins qui seront produits.

Nous considèrerons aussi que le taux d'erreur cible r et le niveau de confiance souhaité α ne varient pas selon les domaines, ce qui semble une hypothèse raisonnable. Si N_D est le nombre de questionnaires remplis pour le domaine D, on a: $N = \sum_{D=1}^{H} N_D$

Selon (21), le nombre minimum de contrôles dans un domaine donné D s'écrit :

$$n_D^* = \frac{N_D}{1 + \frac{rN_D}{g_\alpha^{\,2}(1-2r)}}$$

Le nombre minimum de contrôles, tous domaines confondus, s'écrit alors :

$$\frac{n^*}{N} = \frac{1}{N}\sum_{D=1}^{H} n_D^* = \sum_{D=1}^{H} \frac{N_D / N}{1 + \frac{rN_D}{g_\alpha^{\,2}(1-2r)}}$$

Si nous considérons, pour simplifier, que le nombre de questionnaires est le même dans chaque domaine, nous avons $N_D = N/H$, ce qui donne :

$$\boxed{\frac{n^*}{N} = \frac{1}{1 + \frac{rN}{Hg_\alpha{}^2(1-2r)}}} \tag{22}$$

Prise en compte dans les calculs du contrôle automatique préalable

En pratique, le contrôle manuel est généralement une seconde étape après un contrôle automatique : considérant la plausibilité des données en entrée, un programme informatique est appliqué pour séparer les questionnaires jugés « acceptables » des questionnaires jugés « douteux ». Les questionnaires "acceptables" ne sont même pas vérifiés : ils sont considérés comme corrects. Seuls ceux qui sont douteux sont contrôlés manuellement.

En conséquence, si *f* est la proportion de questionnaires douteux, la véritable population de questionnaires à vérifier est non pas le nombre de répondants *N*, mais le nombre *f.N* de questionnaires douteux.

Si notre taux d'erreur cible est *r* , le taux d'erreur cible parmi les unités douteuses sera *r/f*. Dans les précédentes formules, telles que (20) or (22) par exemple, il nous faut ainsi remplacer *N* par *f.N* et r par *r/f*. Nous tirons de ces observations le résultat général suivant :

$$\boxed{\frac{n^*}{fN} = \frac{1}{1 + a\frac{fN}{H}}} \tag{23}$$

Autrement dit :

$$\boxed{\frac{1}{n^*} = \frac{1}{fN} + \frac{a}{H}} \tag{24}$$

où

$$a = \frac{\frac{r}{f}}{g_\alpha{}^2(1 - 2\frac{r}{f})}$$

Il en ressort que le **nombre minimum de vérifications** à effectuer (parmi les questionnaires "douteux", donc) est une **fonction croissante du nombre *H* de domaines**, **du nombre total de questionnaires à vérifier *f.N***. C'est évidemment une fonction croissante du niveau de couverture souhaité (représenté ici par g_α), et décroissante du taux d'erreur cible. Tous ces résultats sont très intuitifs, mais le plus important est probablement le lien entre le coût du contrôle manuel et le nombre de sous-populations d'intérêt.

On constate au passage que si *H* tend vers l'infini, la proportion minimale de contrôles à effectuer tend vers 100%, ce qui est logique : au pire on a une unité par sous-population, et il faut la vérifier.

On peut aussi remarquer que la proportion minimale d'unités à contrôler n^*/fN est une fonction de H seulement au travers de N/H : le coût minimum des contrôles est une fonction du nombre moyen d'unités par domaine. Quel que soit leur nombre, c'est la taille des sous-populations qui fait la différence.

Ordres de grandeur

Afin d'avoir un ordre d'idées, prenons pour valeurs :

Seuil α=0.05 : c'est le niveau de confiance utilisé en général,

Taux d'erreur cible de 5% : r=0.05 , ce qui semble un but raisonnable,

Taux d'anomalies (détectées par programme) de 50% : f=0.5 ; on peut même juger cela très optimiste pour des enquêtes d'entreprises complexes, où 0.8 ou 0.9 peuvent être aisément atteints.

Nous obtenons alors à partir de (23):

$$\boxed{\frac{n^*}{fN} \approx \frac{1}{1+\frac{P}{43}}} \qquad (25)$$

où P est le nombre moyen de questionnaires à vérifier par « case de diffusion »

Dès lors, la proportion minimale de questionnaires à contrôler manuellement (parmi les « douteux ») est :

63% si $P = 25$,

46% si $P = 50$,

30% si $P = 100$.

Ainsi, on doit vérifier au strict minimum la moitié des questionnaires dès lors que le nombre de questionnaires « par case » passe au-dessous d'une cinquantaine.

A ce stade, il est important de souligner que nous considérons seulement une proportion minimale (cas où tout ce que l'on vérifie est non-erroné), et non une proportion attendue: la quantité réelle de vérifications est probablement significativement plus grande.

Ajoutons que si l'on utilise l'approximation (16) à la place de l'approximation (22), on obtient :

$$\boxed{\frac{n^*}{fN} \approx \frac{1}{1+\frac{N}{60H}}} \qquad (26)$$

Ce qui nous donne des proportions encore plus élevées.

Commentaires

Il est important ici de rappeler que N n'est pas la taille de la base de sondage, mais bien le nombre de questionnaires retournés (= le nombre de répondants). N est donc plus petit que la taille de l'échantillon. Or dans la pratique, il n'est pas rare que l'on diffuse à un niveau fin, avec un nombre de questionnaires par sous-population d'intérêt de l'ordre de 50 ou 100, voire moins. Les formules (25) ou (26) ci-dessus mettent donc en évidence le surcoût très élevé induit par la

finesse de diffusion : selon que l'on décide de produire des statistiques au niveau région, département, zone d'emploi (ou, pour les activités, NES36, NES116, NAF700), la charge de vérification manuelle sera totalement différente.
Ainsi, si le nombre de domaines H est tel que N/H =100, et que l'on décide d'affiner en doublant le nombre de domaines de diffusion, on passe à N/H = 50 et le coût, mesuré en proportion de questionnaires à vérifier, passe de 30% à 46%. Si l'on affine la diffusion en multipliant encore par deux le nombre de domaines d'intérêt, on passe cette fois de 46% à 63%.

En d'autres termes, le niveau de finesse des résultats publiés est un choix fondamental du concepteur d'enquête (ou du responsable de l'utilisation d'une source administrative), qui se répercute automatiquement sur les coûts à qualité donnée … ou sur la qualité, à coût donné.

7. Limites de l'approche

Nous avons dit à plusieurs reprises que la technique utilisée était simplificatrice. Nous précisons ici en quoi : cela touche au concept même d' « erreur », et à la variabilité des tailles de ces erreurs.

Concept d'erreur

La notion de questionnaire « correct » peut être débattue, pour différentes raisons, à savoir:

- Quelques très petites erreurs n'ont pas d'importance, et plus généralement il y a différentes tailles d'erreurs pour une variable donnée ;
- De plus, il y a plusieurs variables dans un questionnaire, et de toute évidence un questionnaire rempli avec une erreur est "moins erroné" qu'un questionnaire rempli avec dix erreurs, si les tailles d'erreur sont les mêmes ;
- Il est quelquefois difficile de savoir si une valeur est correcte ou non.

Dans le cadre proposé, une non-erreur serait un questionnaire tel que la correction de données erronées n'est pas indispensable. Une erreur serait alors un questionnaire dont l'impact des valeurs fausses n'est pas tout à fait négligeable.

En fait, les questionnaires peuvent être divisés grosso modo en 3 catégories : les questionnaires acceptables (aucun ne sera vérifié), les douteux qui ont beaucoup d'influence (100% seront vérifiés, car il serait trop risqué de les laisser passer), et les autres. Tout le raisonnement des sections précédentes s'applique clairement à cette troisième catégorie.

Tailles inégales des erreurs

L'un des principaux problèmes dans l'approche résulte du fait que la probabilité d'erreur est considérée comme étant la même pour toutes les unités, ce qui n'est pas vrai en pratique. En utilisant les distances entre valeurs du questionnaire rempli et valeurs « acceptables », nous pourrions, par exemple construire un modèle permettant de définir une relation entre la probabilité d'erreur et une telle distance. Alors nous aurions des probabilités proportionnelles à une « taille d'erreur », ce qui obligerait à recalculer les intervalles de confiance. Les formules seraient bien moins faciles à traiter, ce qui n'est pas réellement un problème si le critère d'arrêt est utilisé pour décider quand on s'arrête (et non pour estimer la quantité minimale de vérifications à faire).

Suivant la même idée, l'hétérogénéité des probabilités des erreurs implique que les contrôles doivent se voir affecter des priorités (triés par probabilité décroissante d'erreur). Ceci signifie que l'échantillon sélectionné de questionnaires remplis n'est pas un échantillon aléatoire, ce qui aussi modifie les résultats.

Insistons sur le fait que l'objectif d'ensemble du contrôle de données, et plus généralement de l'ensemble du processus de production statistique, n'est pas de réduire le taux d'erreur (ou le taux de non-réponse, …) mais de maîtriser la précision des résultats produits. En conséquence, si le principe de critère d'arrêt évoqué dans le présent papier est important pour évaluer un coût minimal, un critère d'arrêt plus efficace devrait être basé sur l'impact des agrégats (Lawrence & McKenzie 2000), et un critère optimal serait basé sur l'impact de chaque unité sur l'erreur quadratique moyenne, pour chaque variable cible et pour chaque domaine cible (Rivière 2002).

Cependant il nous faut souligner que même si un critère d'arrêt basé sur un taux d'erreur cible semble trop simpliste, il arrive en pratique que les gestionnaires d'enquête se fassent beaucoup de souci à cause du nombre de questionnaires erronés, ce qui signifie qu'implicitement, assurer que le taux d'erreur est faible est un but principal pour les gestionnaires d'enquête. Dans de nombreux cas (l'ONS par exemple), il y a de fortes relations entre gestionnaires d'enquête (qui vérifient les questionnaires en retour) et les statisticiens du domaine (qui publient les résultats). Et, quel que soit l'INS, ce dernier n'aime pas voir de nombreuses erreurs dans ses données. Ceci implique que même si un critère « optimal » est donné, la façon dont les gestionnaires d'enquête vérifient les données est une autre histoire, qui est généralement moins bien maîtrisée que leur hiérarchie pourrait le penser …

Bibliographie

[1] Blyth C.R., Still H.A. (1983), "Binomial confidence intervals", *Journal of the American Statistical Association*, 78, 108-116

[2] Blyth C.R. (1986), "Approximate binomial confidence limits", *Journal of the American Statistical Association*, 81, 843-855

[3] Chen H. (1990), "The accuracy of approximate intervals for a binomial parameter", *Journal of the American Statistical Association*, 85, 514-518

[4] Dalal S.R., Mallows C.L. (1988), "When should one stop testing software?", *Journal of the American Statistical Association*, 83, 872-879

[5] Lawrence D., McKenzie R (2000), The General Application of Significance Editing, *Journal of Official Statistics*, Vol. **16**, No 3, September

[6] Leemis L.M., Trivedi K.S. (1996), "A comparison of approximate interval estimators for the Bernoulli parameter", *The American Statistician*, February 1996, vol.50, 63-68

[7] Rivière P. (2002) Impact on mean squared error as a score to handle data editing, *Document de travail E2002/10,* INSEE, Direction des Statistiques d'Entreprise, Décembre 2002.

[8] Siegmund D, Simons G., Feder P. (1968), "Existence of Optimal Stopping Rules for Rewards Related to S_n/n " *Annals of Mathematical Statistics*, Vol. 39, 1228-1235