# Guide de contrôle de la confidentialité de la DAD

Comité de contrôle de la confidentialité de la Division de l'accès aux données

# Table des matières

# Préface

L'objectif premier du processus de contrôle de la confidentialité est de protéger la confidentialité des données et des renseignements fournis par les répondants. Les analystes de Statistique Canada doivent examiner (contrôler) tous les produits avant qu'ils soient diffusés à partir d'un environnement sécurisé, afin de garantir la protection continue de la confidentialité des données et de maintenir la confiance des Canadiens.

Tous les **utilisateurs de données sont des « personnes réputées être employées » de Statistique Canada**. Cela signifie que tous les utilisateurs de données devront, en ce qui a trait à la confidentialité des données, (1) prêter serment et signer le formulaire connexe, (2) se conformer aux politiques et aux protocoles de Statistique Canada et (3) accéder aux données et les utiliser en respectant rigoureusement leur entente sur l'accès aux données, comme il est expliqué dans le guide du chercheur de la **Division de l'accès aux données (DAD)**. Tous les utilisateurs de données sont également responsables de créer et de soumettre ce que l'on considère comme des « produits sécuritaires » pour le contrôle de la confidentialité et la diffusion à partir de **points d'accès sécurisé de Statistique Canada[1] (PASS)**. Dans le cadre de ce processus, tous les utilisateurs de données (1) suivront une formation sur le contrôle de la confidentialité et auront accès au matériel de formation connexe et (2) suivront les directives en matière de contrôle de la confidentialité afin de créer des produits sécuritaires.

Chacun d'entre nous joue un rôle dans le maintien de la confidentialité des données.

Le présent document ne se veut pas un guide sur les procédures statistiques d'analyse des données. Il est axé sur des considérations relatives au contrôle de la confidentialité des produits pour vérifier leur conformité aux lignes directrices en matière de confidentialité, ainsi que sur la documentation requise pour que les produits soient correctement contrôlés afin de maintenir la confidentialité.

Ce document est régulièrement mis à jour par le Comité de la confidentialité de la DAD.

[1] **Point d'accès sécurisé (PASS) :** endroit indiqué dans l'entente sur l'accès aux données où une personne réputée être employée peut utiliser les renseignements protégés. Cet endroit répond aux normes ministérielles de sécurité de Statistique Canada sur l'accès aux données, selon le niveau de risque déterminé.

# Responsabilités en matière de protection des données

Statistique Canada est responsable de la protection de la confidentialité des répondants dans tous ses données. Chaque employé et chaque personne réputée être employée de Statistique Canada, y compris le personnel et les utilisateurs de données, sont personnellement responsables de la prévention de la divulgation de renseignements confidentiels. Une partie de cette responsabilité consiste à comprendre les principes de base de la confidentialité et la manière d'empêcher la divulgation accidentelle de renseignements confidentiels. Une autre partie de cette responsabilité consiste à demander des produits destinés au contrôle de la confidentialité qui répondent à toutes les règles en matière de contrôle de la confidentialité relatives aux données. Tous les utilisateurs de données doivent être conscients que les « produits » couvrent un large éventail de documents : produit imprimé ou électronique, documentation sur les données, la syntaxe ou le code statistiques et les notes manuscrites.

Avant qu'un produit puisse être diffusé à partir du point d'accès sécurisé de Statistique Canada (PASS; un centre de données de recherche, par exemple), il doit être examiné et approuvé aux fins de diffusion par un employé de Statistique Canada qui a reçu la formation appropriée et qui est responsable de l'examen et de la diffusion des produits (appelé « analyste responsable du contrôle de la confidentialité » dans ce document). Certains employés de Statistique Canada, comme les adjoints à la statistique (ou CR-04), ne sont pas autorisés à contrôler les produits, mais ils peuvent aider à préparer les produits aux fins du contrôle. L'analyste responsable du contrôle de la confidentialité s'assurera que tous les utilisateurs de données connaissent les règles de contrôle de la confidentialité applicables aux ensembles de données auxquels ils ont accès (ainsi que toute modification apportée aux règles de contrôle de la confidentialité susceptible d'avoir une incidence sur leur projet). L'analyste examinera et contrôlera également tous les produits soumis aux fins de diffusion afin de vérifier que les règles de contrôle de la confidentialité sont respectées. Si ces règles ne sont pas respectées pour une partie ou l'ensemble des produits soumis, l'analyste travaillera avec les utilisateurs de données pour trouver des solutions; cependant, il faut comprendre que parfois, il n'y a aucun moyen pour un produit de satisfaire aux règles de contrôle de la confidentialité, et que dans ces situations, le produit ne peut être diffusé.

La responsabilité première de toutes les personnes réputées être employées est de s'assurer que les règles de contrôle de la confidentialité applicables de Statistique Canada sont respectées afin que les données tirées des observations ne permettent pas l'identification de répondants en utilisant les produits diffusés à partir d'un PASS. Cela peut parfois se faire assez facilement à partir de certains types de produits (p. ex. les valeurs minimales et maximales), mais aussi en combinant et en comparant les produits diffusés au fur et à mesure qu'un projet de recherche se poursuit. L'identification des observations est un risque plus important pour les statistiques descriptives (p. ex. les chiffres, les moyennes, les pourcentages), car des rajustements mineurs apportés à l'échantillon ou à des variables peuvent créer des situations où un petit nombre d'individus changent de catégorie et deviennent par conséquent identifiables (c'est ce qu'on appelle la divulgation par recoupements). Tant les utilisateurs de données que les employés de Statistique Canada sont responsables d'atténuer le plus possible de ce risque. Voici quelques questions essentielles à se poser lors de la préparation des produits aux fins de contrôle de la confidentialité :

- Le produit est-il cohérent avec ce qui a été écrit dans la proposition de projet?
- L'analyse et les produits sont-ils corrects – ont-ils été vérifiés minutieusement pour repérer la présence d'erreurs? Les chiffres sont-ils tous logiques?
- Le produit demandé est-il vraiment requis en dehors du PASS? L'équipe de recherche ou le superviseur ont-ils convenu que le produit est prêt à être soumis à un contrôle de la confidentialité?
- Ce produit peut-il être combiné avec un autre produit du même projet d'une manière qui pourrait conduire à l'identification d'un répondant?
- Des changements dans la composition de l'échantillon sont-ils prévus?
- Les statistiques descriptives ou les fréquences peuvent-elles être diffusées à la fin du projet?

# Perception du public et protection de l'information

Un élément qui doit être pris en compte par tous les utilisateurs de données, ainsi que par les employés de Statistique Canada lorsqu'ils examinent les produits, est la perception que les renseignements confidentiels du répondant et son identité sont protégés. Même si certaines règles et procédures de contrôle de la confidentialité des produits peuvent sembler trop restrictives et inutiles, il s'agit de règles que les responsables de la méthodologie et des secteurs spécialisés ont mises en place pour protéger les données afin que les fichiers puissent être utilisés par les utilisateurs de données des PASS. Tous les utilisateurs de données prêtent le serment professionnel et engagement au secret professionnel et signent l'entente sur l'accès aux données; ils sont tenus de respecter la confidentialité des données en vertu de la *Loi sur la statistique*. Les règles de contrôle de la confidentialité sont conçues pour garantir la protection des données à cette fin. Tout produit diffusé à partir des PASS est considéré comme étant de notoriété publique, et la perception que des renseignements personnels sont divulgués sans permission peut susciter de la méfiance à l'égard de Statistique Canada auprès des membres du grand public en ce qui concerne la protection de leurs renseignements, ce qui pourrait compromettre sérieusement le rôle que joue l'organisme dans sa capacité à contribuer à la recherche.

# L'utilisation de poids et leur lien avec la confidentialité

De nombreuses sources de données sont accompagnées d'un ensemble de *poids d'enquête*, ou de *poids d'échantillonnage*. Créés par Statistique Canada, ces poids sont utilisés pour produire des estimations à l'échelle de la population. Ces poids jouent également un rôle dans la protection de la confidentialité des données. De nombreuses sources de données sont également accompagnées de poids bootstrap pour tenir compte des plans de sondage complexes. Cependant, les *poids bootstrap* ne jouent pas de rôle dans la protection de la confidentialité des données. Certaines sources de données ne sont pas accompagnées de poids car la pondération n'est pas applicable. La pondération n'est pas applicable lorsque : 1) le dénombrement de la population d'intérêt est complet (c'est-à-dire qu'il ne s'agit pas d'un échantillon, comme le Recensement); 2) l'échantillon est un échantillon aléatoire simple de la population totale (par exemple, la Banque de données administratives longitudinales). Par contre, si ces données administratives sont couplées à d'autres enquêtes, ces nouvelles bases de données couplées auront des poids d'échantillonnage corrigés à la fois par la probabilité d'être couplées et par le plan d'échantillonnage de l'enquête.

La pondération joue un rôle dans la confidentialité des données pour toute donnée représentant un échantillon d'une population (comme dans le cas d'enquêtes ou de données administratives couplées). Un poids d'échantillonnage agit généralement comme un poids de fréquence permettant d'augmenter le nombre d'unités d'analyse des données (p. ex. de répondants), de sorte qu'une se ule unité est maintenant représentative d'une plus grande proportion d'unités – par exemple, avec un poids d'échantillonnage de 5, une seule unité d'analyse des données représente maintenant 5 unités de population. Bon nombre de progiciels traitent un poids d'échantillonnage comme un poids de fréquence, et une unité est utilisée dans les calculs de la variance comme si autant de clones de cette unité avaient participé à l'enquête. Cela peut entraîner une énorme sous-estimation de la variance d'une estimation si le progiciel suppose la présence d'un grand échantillon, et des résultats trompeurs significatifs. De plus, les poids d'échantillonnage seuls ne permettent pas de prendre en compte le plan d'échantillonnage de l'enquête, en particulier le regroupement par sites, et peuvent donc, lorsqu'ils sont utilisés seuls, donner lieu à de faux positifs dans les tests de signification. Certains progiciels n'éprouvent ce problème que dans un sous-ensemble de fonctions ou de procédures d'analyse. Ainsi, pour chaque calcul qui teste une hypothèse ou génère/utilise une variance, les utilisateurs de données doivent vérifier comment le progiciel d'analyse qu'ils ont choisi traite les poids d'échantillonnage.

Les règles de contrôle de la confidentialité d'un ensemble de données, qui comprennent une variable de poids d'échantillonnage, indiquent si certains types de produits peuvent être diffusés dans un format non pondéré (fondé sur des données brutes) ou pondérés (application de poids d'échantillonnage). En général, étant donné que les produits non pondérés représentent les observations réelles dans les données, les lignes directrices à suivre pour demander des produits non pondérés sont plus strictes que pour un produit pondéré. Il peut également être demandé aux utilisateurs de données de fournir une justification écrite pour les produits non pondérés, et des mesures supplémentaires de protection de la confidentialité pourraient devoir être adoptées (par exemple, arrondissement aléatoire ou contrôlé, utilisation de poids de perturbation, augmentation du seuil du compte minimal). La justification pourrait également devoir être examinée et approuvée par le Comité de la confidentialité de la DAD.

**En règle générale, pour les produits non pondérés, le seuil du compte minimal est triplé pour la demande actuelle et toutes les demandes futures en ce qui concerne les tailles d'échantillon et les produits descriptifs (pondérés ou non), peu importe l'échantillon d'analyse ou les variables incluses dans la demande de contrôle.**

Les lignes directrices générales décrites dans ce document s'appliquent aux produits pondérés et non pondérés, mais les lignes directrices en matière de contrôle de la confidentialité pour chaque source de données doivent être consultées pour connaître les règles précises concernant les produits pondérés et non pondérés.

### Poids normalisés

Certains utilisateurs de données préfèrent normaliser (ou échelonner) les poids d'échantillonnage dans leurs analyses. Ce processus redistribue le poids d'échantillonnage associé à chaque unité d'analyse de données de façon à diminuer l'effet de l'augmentation de la taille de l'échantillon. Cela se fait généralement en divisant le poids d'échantillonnage de chaque unité par la moyenne des poids d'échantillonnage de toutes les unités de l'échantillon dans la sous-population. Les poids normalisés qui en découlent auront une valeur moyenne de 1,0 et les poids normalisés de toutes les unités d'échantillon dans la sous-population correspondront à la taille de l'échantillon dans la sous-population. L'effet des différentes méthodes de pondération sur les chiffres et les proportions est présenté dans le tableau ci-dessous :

**Considérations relatives au contrôle de la confidentialité pour les poids normalisés**

1. Bien que le poids normalisé applique effectivement une pondération aux données, il n'est pas destiné à être utilisé pour les données descriptives, car il peut donner l'impression de chiffres non pondérés. L'utilisation d'un poids normalisé permet de réduire l'effet trompeur de la signification et de la précision découlant de l'augmentation des tailles d'échantillon lorsque le poids d'échantillonnage est employé.
2. La principale préoccupation avec les poids aweights et les autres poids qui normalisent ou standardisent les poids concerne les comptes, les fréquences et les totaux. Ces dénombrements normalisés sont très similaires et dans certains cas identiques aux dénombrements non pondérés, et doivent donc atteindre le seuil non pondéré pour être publiés. Cependant, les descriptifs calculés avec des pondérations telles que les proportions ou les pourcentages, ainsi que les régressions, sont considérés comme pondérés et peuvent être libérés.
3. La syntaxe de création du poids normalisé doit obligatoirement faire partie des documents justificatifs, de même que la syntaxe montrant que le poids normalisé est appliqué au produit dont le contrôle est demandé.

**Tableau 1 : Exemples des différentes méthodes de pondération avec des chiffres et des valeurs descriptives**

| | | Non pondéré | Poids normalisé | Poids d'échantillonnage |
|---|---|---|---|---|
| **Exemple 1 : Variable binaire/dichotomique** | | | | |
| | Chiffre total | 6 914 | 6 914 | 3 810 200 |
| | Chiffres | | | |
| | 0 | 3 659 | 3 644 | 2 008 355 |
| | 1 | 3 255 | 3 270 | 1 801 845 |
| | Pourcentages | | | |
| | 0 | 52,92 | 52,71 | 52,71 |
| | 1 | 47,08 | 47,29 | 47,29 |
| | Moyenne | 0,4708 | 0,4729 | 0,4729 |
| | Erreur-type | 0,0060 | 0,0060 | 0,0003 |
| | | | | |
| **Exemple 2 : Variable catégorique** | | | | |
| | Chiffre total | 2 291 | 2 291 | 1 335 323 |
| | Chiffres | | | |
| | 0 | 80 | 89 | 51 775 |
| | 1 | 1 140 | 1 092 | 636 230 |
| | 2 | 891 | 913 | 532 242 |
| | 3 | 166 | 185 | 107 849 |
| | 4 | 14 | 12 | 7 227 |
| | Pourcentages | | | |
| | 0 | 3,49 | 3,88 | 3,88 |
| | 1 | 49,76 | 47,65 | 47,65 |
| | 2 | 38,89 | 39,86 | 39,86 |
| | 3 | 7,25 | 8,08 | 8,08 |
| | 4 | 0,61 | 0,54 | 0,54 |
| | | | | |
| **Exemple 3 : Variable continue** | | | | |
| | Chiffre total | 2 287 | 2 287 | 1 333 251 |
| | Moyenne | 5,24 | 5,22 | 5,22 |
| | Écart-type | 4,56 | 4,51 | 4,51 |

## Règles de diffusion et règles de contrôle de la confidentialité

Les règles de contrôle de la confidentialité peuvent ressembler aux règles de diffusion fournies dans les guides de l'utilisateur ou dans certaines publications des organismes gouvernementaux. Les deux dépendent dans une large mesure des tailles d'échantillon associées aux estimations, mais leur objet est différent. Les règles de diffusion visent la qualité des résultats (c.-à-d. le degré de fiabilité d'une estimation) et une communication uniforme des résultats, mais ne sont pas appliquées lors du contrôle des produits dont la diffusion est demandée. Les règles de contrôle de la confidentialité visent la protection contre toute divulgation de données confidentielles, mais ces règles ne servent pas à

évaluer la qualité des résultats. Lorsque des produits sont générés et soumis aux fins de contrôle, il incombe à l'utilisateur de données d'évaluer la qualité de ses résultats; les employés de Statistique Canada vérifient les produits uniquement pour en protéger la confidentialité; ils ne vérifient pas les produits pour s'assurer qu'ils sont de qualité suffisante pour être diffusés.

# Préparer une demande de contrôle

Lors de la préparation d'une demande de contrôle, plusieurs points doivent faire l'objet de discussions avec l'équipe de recherche ou l'analyste responsable du contrôle de la confidentialité pour le PASS. Une discussion au sein de l'équipe de recherche peut aider à mettre au point le produit, notamment en passant en revue la définition des variables et la pertinence des analyses. Une discussion avec l'analyste responsable du contrôle de la confidentialité peut aider à faire en sorte que toutes les règles de contrôle de la confidentialité pertinentes soient appliquées; on peut également discuter d'un ensemble de produits particulier dans le contexte plus large de ce qui a déjà été diffusé, ainsi que de l'état d'avancement du projet de recherche dans son cycle de vie. Le temps de traitement est également un facteur à prendre en compte lors de la soumission des produits aux fins de contrôle – le fait de soumettre des produits à la dernière minute entraîne un risque que ces produits ne soient pas contrôlés à temps. Pour les demandes de contrôle volumineuses, il faut consulter l'analyste responsable du contrôle de la confidentialité à l'avance afin de discuter de délais appropriés pour leur soumission (voir la section sur les demandes de contrôle volumineuses pour obtenir de plus amples renseignements).

La documentation qui accompagne une source de données est considérée comme confidentielle et ne peut être divulguée – cela inclut les dictionnaires de données avec chiffres, les guides de l'utilisateur et les clichés d'enregistrement. Pour de nombreuses sources de données (mais pas toutes), il existe une documentation non confidentielle qui a été fournie par les propriétaires des données et qui peut être demandée à l'analyste responsable du contrôle de la confidentialité. La documentation publiquement disponible peut être trouvée en ligne sur le site Web de Statistique Canada; les utilisateurs de données peuvent également communiquer avec leur représentant de l'Initiative de démocratisation des données.

## Règles de contrôle de la confidentialité propres aux données

Les règles de contrôle de la confidentialité propres aux sources de données utilisées dans le cadre d'un projet peuvent être fournies par le personnel des PASS. Les règles de contrôle de la confidentialité décrivent les conditions qui doivent être remplies pour que les produits générés à partir d'une source de données puissent être diffusés, afin que la confidentialité des microdonnées soit préservée. Il est recommandé que les règles de contrôle de la confidentialité propres aux données soient examinées par l'équipe de recherche, en collaboration avec un analyste responsable du contrôle de la confidentialité, avant le début du projet et au moment où les produits sont préparés aux fins de contrôle.

Les règles de contrôle de la confidentialité fournissent des indications sur les seuils minimums des statistiques descriptives, tels que les fréquences, les statistiques d'ampleur (p. ex. les moyennes, les rapports, les totaux) et les statistiques individuelles (p. ex. les minimums, les maximums), et elles s'appliquent aux tableaux résiduels (veuillez-vous référer à la section sur les Valeurs résiduelles pour plus de détails). Il existe également des lignes directrices pour les graphiques et les produits de modèle. Tout produit qui ne respecte pas ces lignes directrices n'est pas diffusé. Il y a généralement un risque plus faible en matière de confidentialité associé aux coefficients estimés de modèles multivariés, sauf pour les modèles équivalents à des tableaux ou à des produits descriptifs (p. ex. les modèles

entièrement ou presque entièrement saturés). Les lignes directrices en matière de contrôle de la confidentialité n'abordent pas tous les types d'analyse, d'où l'importance de consulter l'analyste responsable du contrôle de la confidentialité au moment de préparer des produits aux fins de contrôle.

Les utilisateurs de données doivent faire attention lorsqu'ils utilisent des populations étroitement définies, comme de petites régions géographiques, des institutions, des minorités visibles, des variables de revenu ou d'autres variables de nature délicate qui peuvent faire l'objet de protections supplémentaires en matière de confidentialité (p. ex. l'arrondissement) avant la diffusion des produits. En outre, toute information susceptible de révéler la base de sondage utilisée dans une enquête ne peut être diffusée – cela inclut des renseignements comme une liste de codes postaux ou de subdivisions de recensement.

Les règles de contrôle de la confidentialité applicables à un ensemble de données donné peuvent changer au fil du temps, tout comme la façon dont ces règles sont appliquées aux différentes méthodologies statistiques. Cela peut signifier que les règles de contrôle de la confidentialité deviennent plus souples ou plus restrictives, ce qui peut avoir des répercussions sur la possibilité de diffusion tout au long du cycle de vie d'un projet, en particulier si le contrat a fait l'objet d'une prolongation ou d'une révision. L'analyste responsable du contrôle de la confidentialité fera de son mieux pour informer les utilisateurs de données de toute modification apportée aux règles de contrôle de la confidentialité applicables à un ensemble de données précis auquel ils ont accès, mais l'utilisateur de données doit également se familiariser avec les règles et les modifications.

## Considérations particulières pour le contrôle de la confidentialité

### *Unité d'analyse et sous-échantillons*

De façon générale, dans la plupart des situations de contrôle de la confidentialité l'« unité d'analyse » est le répondant individuel, et de nombreuses lignes directrices en matière de contrôle de la confidentialité sont axées sur ce type d'unité. Cependant, certaines recherches peuvent utiliser un ménage, une institution/entreprise ou une grande région géographique comme unité d'analyse. Dans ces cas, les seuils minimums indiqués pour les produits, comme les valeurs descriptives, s'appliquent à cette unité, et non au nombre de répondants contenus dans cette unité. Par exemple, lors de l'élaboration d'un produit où l'unité d'analyse est un ménage, le seuil minimum s'applique au nombre de ménages dans l'analyse, et non au nombre d'individus dans les ménages.

Pour certaines sources de données, l'unité d'analyse est limitée à certains types ou à certaines régions géographiques. Par exemple, certaines sources de données ne permettent de diffuser que les produits à l'échelle provinciale, et d'autres stipulent que les produits d'une unité d'analyse particulière (par exemple, un établissement ou une région métropolitaine de recensement) ne peuvent être diffusés que sous certaines conditions. Les règles de contrôle de la confidentialité applicables à une source de données donnée préciseront si c'est le cas, et les utilisateurs de données sont encouragés à en discuter avec leur analyste responsable du contrôle de la confidentialité.

### *Données manquantes et répondants omis*

La façon dont les données manquantes sont traitées en ce qui concerne la sélection de l'échantillon doit être clairement précisée dans toute demande de contrôle. Par exemple, les réponses « Ne s'applique pas » à certains éléments peuvent être recodées en un « Non » selon la façon dont la réponse « Ne s'applique pas » a été obtenue et, par conséquent, les chiffres pour cette réponse peuvent devoir être fournis en tant que document justificatif. Il faut également veiller à examiner le nombre de répondants omis afin d'éviter tout problème de confidentialité par recoupements. Par exemple,

si une analyse n'utilise que des données provenant de femmes, l'utilisateur des données est tenu de fournir à l'analyste responsable du contrôle de la confidentialité le nombre d'hommes qui ont été exclus de l'échantillon d'analyse pour éviter toute divulgation par recoupements.

### *Imputation et valeurs manquantes*

Pratiquement toutes les bases de données et tous les dossiers administratifs comportent des valeurs manquantes. Les utilisateurs de données traitent généralement ces données manquantes en supprimant les observations qui contiennent des valeurs manquantes ou en remplaçant ces dernières par une valeur estimée fondée sur d'autres variables ou au moyen de l'imputation.

**Les règles générales de contrôle de la confidentialité relativement aux valeurs manquantes sont les suivantes :**
(1) À moins qu'elles ne soient fusionnées avec d'autres réponses à valeur manquante, les catégories « Ne s'applique pas » ou « Saut valide » sont traitées comme des réponses non manquantes, car elles peuvent représenter une caractéristique d'une unité déclarante et elles font généralement partie d'un enchaînement de questions dans la conception du questionnaire (par exemple, si un répondant indique qu'il n'a pas vécu de dépression, le code « Saut valide » lui est attribué pour toutes les futures questions relatives à la dépression, plutôt que « Non »). Ainsi, les chiffres pour ces catégories de valeurs manquantes doivent respecter les seuils minimums de fréquence par cellule et ne sont pas adaptés à l'imputation de données.
(2) Les catégories « Non déclaré », « Refus » et « Ne sais pas » sont considérées comme des non-réponses et peuvent être diffusées même si le seuil du compte minimal n'est pas atteint, pour autant que cela ne présente pas de risque en matière de confidentialité.
(3) Les unités qui sont éliminées d'une analyse en raison de données manquantes sont considérées comme faisant partie de la catégorie « Non déclaré » des données manquantes. Par exemple, un changement dans la taille de l'échantillon dans un modèle de régression, à la suite de l'ajout d'une covariable, résulterait d'un manque de cas dans cette nouvelle covariable.

Voici des lignes directrices pour la préparation d'une demande de contrôle lorsque le produit intègre des données manquantes ou une imputation :
(1) Le type de méthode utilisée pour l'imputation devrait être cohérent pour tous les produits demandés aux fins de contrôle de la confidentialité. Il est fortement recommandé que toutes les étapes de l'analyse, y compris le processus d'imputation, puissent être passées en revue avant de présenter une demande de contrôle, car des modifications de méthodes d'imputation ou de variables utilisées pourraient entraîner des problèmes de confidentialité par recoupements, en particulier au moment de l'imputation de variables catégoriques ou binaires. Pour éviter une éventuelle divulgation par recoupements, les utilisateurs de données devraient rendre l'imputation aussi complète que possible dès le début de l'analyse des données. Il est vivement recommandé d'effectuer une imputation et de baser tous les futurs produits sur ces données imputées finales.
(2) L'ensemble de données imputé final (ou les ensembles de données multiples en cas de méthode d'imputation multiple ou similaire) devrait être mis à la disposition de l'analyste responsable du contrôle de la confidentialité. Exécuter à nouveau l'imputation à une date ultérieure peut ne pas produire les mêmes résultats que dans la demande de contrôle.
(3) L'imputation effectuée pour répondre à une exigence de taille d'échantillon pour le contrôle de la confidentialité peut produire des résultats de piètre qualité. Il est vivement recommandé de rencontrer l'analyste responsable du contrôle de la confidentialité pour discuter d'autres options d'analyse.
(4) Les totaux de taille d'échantillon à partir de l'ensemble de données initial (c.-à-d. non imputé) et imputé final peuvent être diffusés. Les fréquences par cellule, les totaux de lignes et de colonnes et les valeurs descriptives (p. ex. moyennes, centiles) qui ne visent pas à décrire la taille de l'échantillon globale doivent uniquement être évalués aux fins

de contrôle à partir de l'ensemble de données imputé. Le nombre d'unités imputées pour les catégories individuelles ne doit pas être diffusé.
(5) Les produits de modèle ne doivent être évalués à partir de l'ensemble de données imputé qu'aux fins de contrôle de la confidentialité.

### *Produits géographiques*

Les niveaux géographiques détaillés utilisés dans les tableaux et les valeurs descriptives doivent suivre les procédures décrites dans les règles de contrôle de la confidentialité concernant les sources de données (p. ex. l'arrondissement ou la possibilité d'être restreint). Les niveaux géographiques détaillés peuvent être employés dans les procédures de modélisation en forme brute, même si la variable géographique définie ne peut pas être diffusée ou si elle doit être arrondie lorsqu'elle est présentée dans un tableau ou en forme descriptive.

Les cartes doivent suivre les mêmes règles que les produits sous forme de tableaux, c'est-à-dire que chaque région géographique de la carte doit contenir le nombre minimum de répondants et répondre à toutes les lignes directrices applicables en matière de contrôle de la confidentialité. Les utilisateurs de données doivent vérifier qu'il n'y a pas de « fragmentation géographique » possible concernant les régions géographiques omises, ce qui entraînerait un problème de confidentialité par recoupements (lorsque la soustraction de toutes les régions géographiques diffusables du total permettrait d'identifier les régions géographiques omises).

### *Suppression de cellules*

**La suppression de cellules en tant que technique de contrôle de la divulgation n'est pas autorisée dans les PASS.**

La suppression de cellules est une technique très répandue pour prévenir la divulgation. Les cellules de nature délicate sont repérées et supprimées d'une publication, tout comme le sont les cellules supplémentaires (secondaires), afin de garantir que les renseignements ne peuvent pas être facilement recalculés à partir de renseignements agrégés. Il est très difficile de gérer la suppression de cellules efficacement lorsque de multiples tableaux sont publiés à partir du même ensemble de données. En outre, dans le contexte de la recherche en sciences sociales, les données sont couplées à de nombreuses autres sources de données administratives (p. ex. les renseignements liés aux déclarations de revenus des particuliers, les données d'hospitalisation, les registres du cancer, les données sur l'éducation) et à d'autres données de nature délicate (p. ex. le recensement de la population, les données des coroners et des médecins légistes, de nombreuses enquêtes par sondage), ce qui les rend plus vulnérables aux risques de divulgation. Étant donné que le même tableau ou des tableaux connexes peuvent être diffusés par différents centres de données (avec les renseignements diffusés par Statistique Canada), il devient pratiquement impossible d'effectuer correctement et uniformément la suppression de cellules dans tous les programmes d'accès de la DAD destinés aux utilisateurs de données; par conséquent, la suppression de cellules n'est pas autorisée dans les PASS.

### *Valeurs résiduelles*

Les valeurs résiduelles sont des situations où les produits issus du même projet et de la même source de données sont combinés, ce qui peut créer des situations où des unités peuvent être identifiées. Tout produit soumis à un contrôle de la confidentialité est comparé à tous les produits précédents publiés pour ce projet, et tout tableau résiduel doit également satisfaire aux lignes directrices en matière de contrôle de la confidentialité qui s'appliquent aux données utilisées. Par exemple, quelqu'un peut demander un tableau croisé de l'état matrimonial par sexe chez les personnes de 20 à 30 ans, puis demander le même tableau, mais uniquement chez les personnes de 21 à 30 ans à une autre occasion. En combinant ces deux demandes, il est possible de déterminer le sexe et l'état matrimonial des personnes âgées de 20 ans.

On recommande aux utilisateurs de données de garder une trace de leurs propres demandes de contrôle afin d'éviter les problèmes de divulgation par recoupements, et pour éviter de demander le même produit à de multiples reprises qui ont subi de légers changements dans les codages de variables. Il est également recommandé d'attendre jusqu'à la toute fin du projet avant de procéder au contrôle des fréquences et des données descriptives, afin d'éviter les risques de divulgation par recoupements.

Bien que, historiquement, l'approche de suppression ait été utilisée dans des situations de recoupement, elle ne permet pas de remédier adéquatement à la divulgation d'attributs ou par recoupements, et n'est donc pas recommandée dans un environnement où de multiples tableaux sont produits à partir de la même base de données, comme les PASS. Par conséquent, la suppression n'est pas autorisée dans les PASS pour traiter les faibles valeurs – les variables, les catégories ou les tableaux doivent être remaniés afin que chaque cellule réponde aux exigences de contrôle de la confidentialité applicables à la source de données utilisée.

Pour obtenir plus de renseignements sur les recoupements et les problèmes qui peuvent se produire, veuillez consulter la section suivante : Valeurs résiduelles.

### *Arrondissement*

Certaines sources de données exigent que tous les résultats soient arrondis à une base d'arrondissement précise (c'est-à-dire de manière déterministe, comme la dizaine ou la cinquantaine la plus proche) ou qu'ils utilisent une technique d'arrondissement comme l'arrondissement aléatoire ou contrôlé. En règle générale, cela signifie que toutes les statistiques descriptives qui impliquent des chiffres publiables (p. ex. taille de l'échantillon, proportion, moyenne) doivent être recalculées en fonction des composants arrondis correctement. Par exemple, dans le cas d'une proportion, le numérateur et le dénominateur doivent être arrondis, et la proportion doit être calculée à partir de ces composants arrondis. De même, dans le cas d'une moyenne, la somme arrondie doit être divisée par la taille arrondie de l'échantillon pour créer la moyenne arrondie. Dans certaines situations, l'arrondissement à un certain nombre de décimales peut également être autorisé – les utilisateurs de données doivent vérifier auprès de l'analyste responsable du contrôle de la confidentialité les options en matière d'arrondissement pour la source de données utilisée.

L'arrondissement n'est appliqué qu'une fois que les seuils minimums de chiffres sont atteints – il ne remplace pas cette exigence (autrement dit, il ne peut pas être utilisé pour rendre diffusables des chiffes non diffusables) et il n'est pas considéré comme une mesure de protection contre la divulgation par recoupements. Par exemple, si l'ajout d'une seule observation à une rangée de tableau fait augmenter une valeur arrondie dans une colonne, nous savons que l'observation appartient à cette colonne, et nous pouvons déterminer la valeur non arrondie pour la cellule dans cette colonne. Pour ces raisons, les projets et les tableaux doivent être contrôlés très soigneusement et il faut toujours garder à l'esprit la divulgation par recoupements.

Dans de nombreux PASS, des outils d'arrondissement sont offerts aux utilisateurs de données pour qu'ils fournissent des produits correctement arrondis aux fins de contrôle de la confidentialité. Vous trouverez également ci-dessous une courte liste d'options d'arrondissement déterministe dans certains progiciels :

| Logiciel | Commande/option d'arrondissement déterministe | Remarques |
|---|---|---|
| **Excel** | =mround (cellule, base d'arrondissement) | Cette fonction arrondit une cellule particulière selon la base d'arrondissement précisée de manière déterministe. |
| **STATA** | format(%5.1f) | Cette option est destinée aux tableaux. Les chiffres entre parenthèses font référence au nombre d'espaces réservées à l'affichage d'un chiffre ou d'un pourcentage, et le chiffre après le point indique le nombre de décimales à afficher. Cet exemple permet de formater le produit sous forme de tableau montrant les chiffres ou les pourcentages avec une seule décimale. |
| **STATA** | cformat(%5.2f) | Cette option est destinée aux produits de modèle. Les chiffres entre parenthèses correspondent au nombre d'espaces réservés à l'affichage des valeurs de coefficient du modèle, et le chiffre après le point indique le nombre de décimales à afficher. Cet exemple permet de formater le produit sous forme d'ensemble de coefficients bêta à deux décimales. |

Les intervalles de confiance (IC) peuvent nécessiter certains efforts lorsque les estimations doivent être arrondies. Deux approches peuvent être utilisées pour calculer un intervalle de confiance pour une estimation arrondie de la prévalence (p. ex. moyenne, centile) :

Approche 1 : Estimation ponctuelle arrondie et intervalles de confiance par centile basés sur une répartition normale bootstrap :
Dans cette approche, l'intervalle de confiance est indiqué en tant que prévalence arrondie +/- 1,96 *(ET des estimations de la prévalence bootstrap) où ET = erreur-type des estimations de prévalence de n échantillons bootstrap, et un intervalle de confiance de 95 % est souhaité. Si le logiciel utilisé ne présente pas le calcul de la variance, les bornes de l'intervalle de confiance peuvent être décalées par rapport à la différence entre l'estimation ponctuelle réelle et l'estimation ponctuelle arrondie.

Approche 2 : Estimation ponctuelle arrondie et intervalles de confiance bootstrap empiriques
Avec cette approche, l'estimation ponctuelle de la statistique calculée est arrondie, et les bornes de l'intervalle de confiance réel de la méthode bootstrap utilisée (par exemple, linéarisation en série de Taylor, biais corrigé/accéléré) sont indiquées. Aux fins de contrôle de la confidentialité, les estimations ponctuelles doivent être supprimées du produit et recalculées, mais les intervalles de confiance peuvent être conservés.

En général, il n'est pas nécessaire d'arrondir les produits du modèle, sauf si le modèle est équivalent à un tableau ou s'il est nécessaire d'arrondir les produits en raison de la géographie ou d'une autre raison précisée dans les lignes directrices de contrôle de la confidentialité propres à la source de données. Pour l'intervalle de confiance d'une estimation basée sur un modèle (p. ex. coefficient de régression ou rapport de cotes), les intervalles de confiance générés au moyen de la méthode bootstrap peuvent être déclarés tels qu'ils ont été calculés par le progiciel.

### *Variables de revenu*

Les variables de revenu peuvent inclure le revenu total, le revenu provenant de sources comme les salaires ou les charges, les coefficients de Gini, ou la nouvelle catégorisation d'une variable de revenu continue en une variable catégorique (p. ex. salaire faible ou élevé). Le revenu est une variable de nature très délicate pour certaines sources de données; les règles de contrôle de la confidentialité pour une source de données donnée indiqueront si c'est le cas, et les utilisateurs de données sont encouragés à en discuter avec leur analyste responsable du contrôle de la confidentialité pour déterminer s'il y a des considérations particulières à prendre en compte pour traiter le revenu et les variables liées au revenu dans un projet. Vous trouverez ci-dessous quelques exigences courantes pour élaborer des produits relatifs au revenu.

(1) En plus des exigences relatives à la fréquence par cellule pour une source de données, il peut également y avoir des seuils de population et des seuils minimums de population à domicile qui doivent être respectés.

(2) Les valeurs de revenu peuvent devoir être arrondies à une valeur particulière en dollars (p. ex. à la centaine de dollars la plus proche).

(3) Certaines sources de données exigent la réalisation d'essais de soutien supplémentaires (comme des tests de dominance) pour garantir le respect de la confidentialité des variables liées au revenu. Il est recommandé de discuter de ces essais de soutien avec un analyste responsable du contrôle de la confidentialité dès le début du processus d'analyse des données.

### *Données sur les entreprises*

L'introduction de données sur les entreprises à plus grande échelle dans les modes d'accès de la Division de l'accès aux données exige un examen très attentif des risques uniques pour la confidentialité qui sont liés à ces données. Les entreprises peuvent être plus facilement identifiables que les répondants dans les données sociales, compte tenu des renseignements contextuels et connus du grand public. Les utilisateurs de données doivent garder en tête les deux considérations suivantes lorsqu'ils utilisent des données sur des entreprises.

(1) Entités commerciales : de nombreuses entreprises auront des entrées multiples dans une même source de données et, par conséquent, les comptes de fréquence peuvent être trompeurs puisqu'une entreprise peut être comptée plus d'une fois pour une caractéristique donnée. Lors de l'examen des comptes de l'échantillon, l'évaluation des identificateurs uniques d'entreprise est le moyen approprié pour les utilisateurs de données de déterminer si les seuils minimums de chiffres sont atteints.

(2) Analyses de sensibilité : en raison de la nature délicate des variables d'une source de données sur les entreprises, le niveau de sensibilité (comme la dominance) de chaque variable utilisée doit être évalué dans chaque analyse. Les utilisateurs de données peuvent consulter l'analyste responsable du contrôle de la confidentialité pour obtenir des conseils sur l'analyse de sensibilité.

## Demandes de contrôle – Résumé des étapes

**Étape 1 :** Demandez à l'analyste responsable du contrôle de la confidentialité quelles sont les lignes directrices connexes pour les données utilisées dans le projet (dans certains cas, ces lignes directrices peuvent être envoyées par courriel à l'utilisateur de données).
**Étape 2 :** Préparez le produit conformément aux lignes directrices de contrôle de la confidentialité. Vous ne devez pas demander de produits qui révèlent des détails sur la base de sondage pour les répondants à l'enquête et leur emplacement, car ils ne sont pas diffusables (par exemple, des grappes sélectionnées de la base de sondage ou une liste des codes postaux des enquêtes sociales). Au besoin, fournissez les produits en version non pondérée et pondérée.
**Étape 3** : Remplissez entièrement le formulaire de demande de contrôle. Il est obligatoire de remplir le formulaire de demande de contrôle, et les renseignements fournis doivent être à la fois exacts et exhaustifs. Il est recommandé aux utilisateurs de données de remplir ce formulaire avec l'aide de l'analyste responsable du contrôle de la confidentialité lorsqu'ils soumettent leur première demande de contrôle.
**Étape 4** : Veillez à ce que toute la syntaxe pour les échantillons de sous-ensemble, la création/le recodage de variables et les analyses soit fournie à titre de « document justificatif ».
**Étape 5 :** Enregistrez tout dans un dossier désigné (par exemple, « :\\à sortir\ ») avec des sous-dossiers appropriés pour toute syntaxe ou tout document justificatif.
**Étape 6 :** Assurez-vous que l'analyste responsable du contrôle de la confidentialité sache que des produits sont en attente de contrôle et qu'il peut communiquer avec vous s'il y a des questions.

Tous les produits soumis aux fins de contrôle de la confidentialité doivent, dans la mesure du possible, être dans un format modifiable (p. ex. Microsoft Word ou Excel) afin que l'analyste responsable du contrôle de la confidentialité puisse fournir une rétroaction et supprimer tout produit qui ne peut être diffusé.

# Formulaire de demande de contrôle

Le formulaire de demande de contrôle (voir annexe A) est un élément essentiel de chaque demande de contrôle. Il s'agit d'un document que les utilisateurs de données doivent remplir afin de vérifier que leur produit respecte toutes les lignes directrices en matière de confidentialité applicables aux données qu'ils utilisent, et de fournir à l'analyste responsable du contrôle de la confidentialité des renseignements permettant d'évaluer le produit et de le placer dans le contexte de ce qui a été publié précédemment pour un projet. Chaque projet reçoit un formulaire électronique de demande de contrôle vierge dans le format MS Word (qui se trouve généralement dans le dossier du projet). Ce document doit être entièrement rempli pour chaque demande de contrôle pour toutes les diffusions, qu'il s'agisse d'un produit statistique, d'une syntaxe, d'une note de recherche ou d'un document non confidentiel.

Le formulaire de demande de contrôle est composé de plusieurs sections. La première section contient des renseignements de base sur le projet et sur l'utilisateur des données du projet qui demande que les produits soient contrôlés. Cette section doit au moins contenir le nom de l'utilisateur des données, une adresse courriel pour communiquer avec cet utilisateur, la date de présentation de la demande de contrôle, le nom d'utilisateur ou du compte de l'utilisateur des données qui fait la demande, et le numéro de contrat de recherche pour l'utilisation de microdonnées (qui fait partie du nom d'utilisateur). Les autres sections sont décrites ci-dessous.

**La section A** contient des questions de base destinées à aider l'utilisateur de données à examiner ses produits par rapport aux lignes directrices de contrôle de la confidentialité applicables à la source de données utilisée et à repérer

toute variable qui pourrait nécessiter des documents justificatifs ou des exigences de contrôle supplémentaires (p. ex. l'utilisation de variables de revenu dans le recensement).

**La section B** traite des sources potentielles de divulgation par recoupements. Elle contient des questions relatives au produit qui a précédemment fait l'objet d'une demande et qui a été diffusé dans le cadre du projet. La section B est importante pour déterminer quelles variables ou quels produits pourraient créer des problèmes liés au recoupement.

**La section C** demande une liste de chaque dossier qui doit faire l'objet d'un contrôle de la confidentialité et être diffusé par l'analyste responsable du contrôle de la confidentialité. Cette section est importante pour que l'analyste ait certains renseignements sur le produit présenté aux fins de contrôle de la confidentialité, comme la source de données utilisée, tout sous-ensemble d'échantillon, la pondération, le niveau géographique et les méthodes d'analyse utilisées. Il est recommandé, pour les projets qui ont utilisé de multiples sources de données, de soumettre les produits utilisant différentes sources de données en tant que fichiers distincts dans cette section.

C'est dans **la section D** que sont énumérés les documents justificatifs pour les fichiers de sortie indiqués dans la section C. Cela inclut toute la syntaxe d'analyse, toute syntaxe de recodage, ou un dictionnaire de données créés par les utilisateurs de données qui contiennent des étiquettes de variable et des définitions. Chaque fichier de la section C doit avoir son propre fichier de soutien. Il est également possible d'ajouter des commentaires supplémentaires pour aider l'analyste responsable du contrôle de la confidentialité à contrôler les produits.

**Considérations supplémentaires pour remplir la section D :**

| | |
|---|---|
| ✓ Données couplées? | Les deux sources de données doivent être indiquées (ainsi que tout fichier de couplage utilisé) afin de garantir que les règles pertinentes sont appliquées. |
| ✓ Poids utilisés? | Précisez la variable de pondération utilisée, même si elle est différente de la variable de pondération initialement liée aux données. Si aucune variable de pondération n'a été utilisée, indiquez « aucune » (mais assurez-vous d'en discuter avec l'analyste responsable du contrôle de la confidentialité). |
| ✓ Méthode d'analyse utilisée? | La liste des méthodes d'analyse des données les plus courantes figure sous le tableau. Plusieurs méthodes peuvent être indiquées si le produit contient plus d'un type d'analyse. Si le produit utilise une méthode qui ne figure pas dans la liste, il y a un espace à la fin du document pour inscrire des notes à l'intention de l'analyste responsable du contrôle de la confidentialité. |
| ✓ Description du sous-échantillon? | Soyez aussi précis que possible dans la description du sous-échantillon. Cela comprend également la description de **TOUT individu supprimé** avant l'analyse. Autrement dit, il faut mentionner non seulement **le filtre ou la variable de sélection**, mais aussi toute **donnée manquante** supprimée, etc. Si **différents sous-échantillons** sont utilisés, il est utile de séparer les produits en fonction des différents sous-échantillons. |

# Lignes directrices en matière de contrôle de la confidentialité pour les analyses

Les sections suivantes présentent des recommandations et des lignes directrices, tant pour les utilisateurs de données que pour le personnel de Statistique Canada, sur le contrôle de la confidentialité de certains types d'analyses. Les analyses couvertes ici ne sont pas exhaustives, mais sont représentatives des types de produits qui font le plus souvent l'objet d'une demande. Pour toute analyse non couverte par cette section, ou pour toute question, les analystes responsables du contrôle de la confidentialité peuvent communiquer avec le Comité de la confidentialité de la DAD.

## Produit sous forme de tableau : fréquence par cellule, proportions et centiles

Pour les tableaux simples, chaque statistique (p. ex. les chiffres, le numérateur et le dénominateur des proportions) doit répondre à l'exigence relative à la taille minimale non pondérée des cellules pour cette source de données (certaines sources de données ont également des exigences minimales pour les tailles de cellules pondérées). Les codages de valeur manquante comme « Ne s'applique pas » ou « Saut valide » doivent également respecter la fréquence minimale par cellule, mais les codages « Ne sais pas », « Refus » ou « Non déclaré » sont diffusables en dessous du seuil minimum. Les cellules dans lesquelles il n'y a pas d'unités ou les cellules dans lesquelles toutes les unités résident peuvent poser problème, car cela pourrait indiquer qu'aucune unité du domaine ne présente une caractéristique particulière, ou encore, que toutes les unités du domaine présentent une caractéristique particulière, et cela peut être particulièrement problématique si une variable confidentielle (par exemple, l'orientation sexuelle) est concernée. Les cellules avec un effectif nul et qui représente des situations impossibles (appelées cellules vides structurelles) sont diffusables. Dans l'exemple de tableau ci-dessous, si le seuil minimum des chiffres non pondérés était de 5 unités, ce tableau ne pourrait pas être diffusé, car (a) l'une des cellules avec une réponse valide est en dessous du seuil et (b) la cellule avec « 0 » devrait être examinée pour déterminer s'il s'agit d'une cellule vide structurelle ou non; la présence du « 3 » dans la ligne des données manquantes serait diffusable à condition qu'il ne soit pas indiqué si ces répondants faisaient partie des catégories « Ne s'applique pas » ou « Saut valide ».

| | A | B | Total |
|---|---|---|---|
| 1 = Oui | 0 | 40 | 40 |
| 2 = Non | 3 | 35 | 38 |
| 3 = Manquant | 4 | 3 | 7 |
| Total | 7 | 78 | 85 |

Les centiles sont des valeurs seuils d'une répartition qui divise cette répartition en plusieurs sections comptant chacune un nombre égal (ou très similaire) d'unités. Les centiles sont contrôlés en tant que limite entre deux groupes – par exemple, une médiane crée deux cellules, l'une au-dessus et l'autre au-dessous de la valeur médiane, chacune d'entre elles devant répondre aux règles de contrôle de la confidentialité applicables à la source de données.

Vous trouverez ci-dessous les exigences en matière de documents justificatifs pour chaque type de produits basés sur le compte, **pour(1) les demandes de contrôle pour produits non pondérés**, et pour (2) **les demandes de contrôle pour produits pondérés (y compris celles avec arrondissement ou poids bootstrap**) :

| Type de produit faisant l'objet d'une demande | | Type de chiffres justificatifs |
|---|---|---|
| | | |
| **Fréquences** | | Chiffres non pondérés pour chaque valeur/catégorie distincte |
| **Tableaux croisés** | | Chiffres non pondérés pour chaque cellule du tableau |
| **Proportions** | | Chiffres non pondérés pour le numérateur et le dénominateur |
| **Centiles** | | Chiffres non pondérés entre les valeurs seuils du percentile |

Les lignes directrices suivantes sont utilisées lors du contrôle de la confidentialité des produits sous forme de tableaux :
1. Pour être diffusables, toutes les cellules d'un tableau doivent être des statistiques diffusables.
2. Les tableaux dont les cellules ne répondent pas aux exigences de ce document ne peuvent pas être diffusés. Le tableau doit être conçu de sorte que toutes les cellules puissent être diffusées (par exemple, en supprimant des variables, en redéfinissant l'échantillon, en regroupant plus d'années de données, en divisant un tableau en tableaux plus simples avec moins de dimensions, en regroupant des lignes ou des colonnes entières). La suppression de cellules n'est pas autorisée.
3. Toute population exclue d'un tableau doit quand même être contrôlée, comme s'il s'agissait d'une dimension du tableau, afin d'éviter toute divulgation par recoupements. En général, ce problème ne se pose que si la population exclue est petite. Par exemple, un tableau de fréquences avec un revenu d'emploi nécessitera un tableau de données connexes contenant les fréquences sans revenu d'emploi. Cela ne s'applique pas aux enregistrements exclus pour des raisons de qualité des données.

**En règle générale, pour les produits non pondérés, le seuil du compte minimal est triplé pour la demande actuelle et toutes les demandes futures en ce qui concerne les tailles d'échantillon et les produits descriptifs (pondérés ou non), peu importe l'échantillon d'analyse ou les variables incluses dans la demande de contrôle.**

## Moyennes

Une moyenne est la somme totale d'une variable divisée par le nombre d'unités. En ce qui concerne le contrôle de la conformité du produit, il faut vérifier que chaque composante d'une moyenne répond aux exigences pertinentes pour la diffusion. Pour les moyennes de variables numériques/continues, le dénominateur non pondéré (p. ex. la taille de l'échantillon) de chaque moyenne demandée est requis en tant que document justificatif. La moyenne d'une variable dichotomique est identique aux proportions de cette variable – par exemple, si le codage du sexe est « hommes = 0 et femmes = 1 », une moyenne de 48 % revient à dire qu'il y a 48 femmes et 52 hommes dans un échantillon de

100 répondants. Par conséquent, pour les proportions basées sur des variables dichotomiques, les fréquences par cellules non pondérées de chaque catégorie sont requises en tant que document justificatif.

Vous trouverez ci-dessous les exigences en matière de documents justificatifs pour chaque type de moyenne, **pour(1) les demandes de contrôle pour produits non pondérés**, et pour (2) **les demandes de contrôle pour produits pondérés (y compris celles avec arrondissement ou poids bootstrap**) :

| Type de produit faisant l'objet d'une demande | | Type de chiffres justificatifs | |
|---|---|---|---|
| | | | |
| **Moyennes de variables continues** | | Chiffres non pondérés contribuant à chaque moyenne | |
| **Moyennes de variables dichotomiques/nominales** | | Fréquences par cellules non pondérées pour chaque catégorie de variable dichotomique | |
| **Moyennes de variables catégoriques avec trois catégories ou plus** | | Traiter comme la moyenne d'une variable continue, sauf si elle est codée comme un ensemble de variables nominales | |

## Valeurs résiduelles

Tout produit soumis au contrôle est vérifié en le comparant à tous les produits précédents qui ont été publiés. Lorsqu'une demande concerne un produit qui crée un tableau résiduel à partir d'un produit déjà publié et dont les chiffres sont inférieurs au seuil permettant la diffusion, le produit concerné par la demande ne sera pas diffusé. On recommande aux utilisateurs de données de garder une trace de leurs propres demandes de contrôle afin d'éviter les problèmes de confidentialité par recoupements et d'éviter de demander le même produit à de multiples reprises. Il est également recommandé d'attendre jusqu'à la toute fin du projet avant de procéder au contrôle des fréquences et des statistiques descriptives, afin d'éviter les risques de divulgation par recoupements.

### Tableau résiduel comportant des cellules à faible fréquence

Un tableau « résiduel » ou « fictif » est un tableau qui, lorsqu'il est combiné avec d'autres produits prenant la forme de tableaux qui utilisent les mêmes variables, permet d'identifier un ensemble d'unités ou produit un tableau qui contient des chiffres qui n'atteignent pas le seuil permettant la diffusion. En général, les tableaux résiduels sont produits en soustrayant le produit de totalisation du sous-échantillon du produit de totalisation de l'échantillon complet, comme dans l'exemple ci-dessous. Les tableaux peuvent faire partie de la même demande de contrôle ou faire l'objet de demandes distinctes (par exemple, pour répondre aux commentaires d'un réviseur). Les tableaux résiduels sont vérifiés en utilisant les versions non pondérées des produits tabulaires.

Lorsqu'il y a un problème lié au recoupement dans un tableau, le produit faisant l'objet d'une demande de contrôle doit être modifié de sorte que chaque cellule, non seulement des tableaux visés par la demande de diffusion, mais aussi celles de tout tableau résiduel, dépasse les seuils minimums pour la diffusion. L'arrondissement des produits peut être

une option viable, mais les utilisateurs de données sont encouragés à discuter des possibilités avec leur analyste responsable du contrôle de la confidentialité afin de déterminer si l'arrondissement convient aux produits demandés

Tableau 1

Échantillon total
(personne mariée = 0) et
(personne mariée = 1)

| | Oui | Non |
|---|---|---|
| 1 | 15 | 41 |
| 2 | 9 | 52 |
| 3 | 38 | 16 |

-

Tableau 2

(personne mariée = 1)

| | Oui | Non |
|---|---|---|
| 1 | 13 | 40 |
| 2 | 5 | 35 |
| 3 | 32 | 8 |

=

Tableau résiduel

(personne mariée = 0)

| | Oui | Non |
|---|---|---|
| 1 | 2 | 1 |
| 2 | 4 | 17 |
| 3 | 6 | 8 |

**Tableaux résiduels avec variables temporelles**

Un tableau résiduel peut également être produit en soustrayant des chiffres pour des éléments qui posent la même question sur un intervalle de temps ou un sous-intervalle de deux moments temporels (généralement des éléments distincts dans l'enquête, et sans utiliser les sauts valides).

(1) Dans l'exemple ci-dessous, le « Oui » du tableau 2 est soustrait du « Oui » du tableau 1 : 16 - 15 = 1 homme a vécu une dépression au cours des 2 à 5 dernières années, mais pas au cours de la dernière année.

(2) Dans l'exemple ci-dessous, le « Oui » du tableau 2 est soustrait du « Oui » du tableau 1 : 20 - 8 = 12 femmes ont vécu une dépression au cours des 2 à 5 dernières années, mais pas au cours de la dernière année.

* Dans cet exemple, la colonne « Non » du tableau résiduel est redondante, mais elle doit être fournie pour que le produit soit complet.

Tableau 1

A déjà vécu une dépression au cours des 5 dernières années

| | Oui | Non |
|---|---|---|
| H | 16 | 40 |
| F | 20 | 40 |

Tableau 2

A vécu une dépression au cours de la dernière année

| | Oui | Non |
|---|---|---|
| H | 15 | 41 |
| F | 8 | 52 |

Tableau résiduel

Dépression au cours des 2 à 5 dernières années

| | Oui | Non |
|---|---|---|
| H | 1 | S. O.* |
| F | 12 | S. O. |

**Enchaînement des questions et nombres résiduels**

Certaines sources de données utilisent des questions d'interview qui sont distinctes, mais liées – par exemple, une question qui demande au répondant s'il a reçu un diagnostic de trouble mental au cours de la dernière année, et une question distincte qui demande si le répondant a reçu un diagnostic de trouble mental au cours de sa vie; la différence entre les résultats pour ces deux questions correspond aux répondants qui ont reçu un diagnostic il y a plus d'un an. Une telle situation de recoupement peut passer inaperçue dans le processus de contrôle de la confidentialité, car les deux questions sont des éléments distincts dans le processus de collecte des données. Cette situation peut également survenir avec des variables dérivées, alors qu'un même concept peut être évoqué à différents points dans le temps dans un questionnaire.

Les utilisateurs de données et les analystes responsables du contrôle de la confidentialité doivent savoir comment les variables du projet sont codées. Le fait de disposer d'étiquettes de variables appropriées peut aider à détecter s'il y a un problème de divulgation par recoupements dans les produits dont la publication est demandée.

**Tableau résiduel avec variables agrégées**

Comme pour les variables temporelles, les variables agrégées, ou « toutes les », peuvent également mener à des tableaux résiduels lorsqu'une ou plusieurs variables sont un sous-ensemble d'une autre. Dans l'exemple ci-dessous, les cas de la colonne « Oui » du tableau 1, soustraits de la colonne « Non » du tableau 2, indiquent ceux qui avaient une maladie chronique AUTRE qu'une maladie respiratoire au cours de la dernière année (autrement dit, « Non » dans le tableau 2). Il convient également de noter que les « Oui » du tableau 2 pour chaque sexe indiquent que ces répondants étaient atteints d'au moins une maladie chronique qui n'était pas une maladie respiratoire; une identification est possible. Comme pour les tableaux avec des variables temporelles, la colonne « Non » du tableau résiduel est redondante, mais elle doit être fournie pour que le produit soit complet.

Tableau 1

Toute maladie chronique (y compris respiratoire) dans la dernière année

| | Oui | Non |
|---|---|---|
| H | 15 | 41 |
| F | 10 | 52 |

Tableau 2

Toutes les maladies respiratoires dans la dernière année

| Si oui dans le tableau 1 | Oui | Non |
|---|---|---|
| H | 9 | 6 |
| F | 5 | 5 |

Valeur résiduelle

Avait une maladie autre que respiratoire

| Oui | Non |
|---|---|
| 9 | S. O.* |
| 5 | S. O. |

Un problème similaire de confidentialité par recoupement peut se poser lorsque des variables sommaires sont dérivées d'un ensemble commun de variables. Par exemple, certaines sources de données peuvent comporter de multiples éléments portant sur la présence de maladies chroniques (p. ex. hypertension artérielle, asthme, MPOC). Deux variables agrégées emboîtées peuvent être créées :

Any_chron_cond – qui équivaut à « 1 » si le répondant a dit « oui » à l'un ou l'autre des éléments de problèmes de santé chroniques, sinon « 0 ».

Resp_chron_cond – qui équivaut à « 1 » si le répondant a dit « oui » à l'un ou l'autre des problèmes de santé chroniques liés à des problèmes respiratoires, sinon « 0 ».

Le tableau croisé de ces deux variables avec le sexe, par exemple, pourrait donner le résultat suivant :

| Tableau 1 | | |
|---|---|---|
| Toute maladie chronique (y compris respiratoire) dans la dernière année | | |
| | Oui | Non |
| H | 150 | 100 |
| F | 70 | 70 |

| Tableau 2 | | |
|---|---|---|
| Toutes les maladies respiratoires dans la dernière année | | |
| Si oui dans le tableau 1 | Oui | Non |
| H | 125 | 125 |
| F | 67 | 73 |

La différence entre les deux tableaux croisés révèle que trois femmes n'ont pas eu de maladie chronique respiratoire (« Oui » pour « Toutes les maladies respiratoires dans la dernière année »), mais elles ont eu au moins un des autres types de maladies chroniques (un « Oui » à « Toute maladie chronique [y compris respiratoire] dans la dernière année »). Le fait de diffuser les deux tableaux croisés ci-dessus équivaudrait à diffuser un tableau ayant une cellule où se trouvent trois répondantes implicites qui ont eu une maladie chronique qui n'était pas de nature respiratoire. Même si les variables en cause sont des agrégats de divers éléments, la cellule résiduelle implicite correspond à un petit nombre de répondants ayant un problème de santé chronique, qui sont potentiellement identifiables. Cette vérification des chiffres résiduels s'applique également aux variables dérivées qui figurent dans un ensemble de données.

Pour détecter ces situations, l'analyste responsable du contrôle de la confidentialité posera les questions suivantes à l'utilisateur de données :
Ces variables sont-elles en cause dans tout contrôle précédent de diffusion?
La création de ces variables agrégées est-elle pleinement décrite dans la syntaxe?
Certaines variables de l'analyse se chevauchent-elles sur le plan des éléments utilisés ou des concepts évalués?

## Modèles

Les produits basés sur des modèles présentent généralement un risque moindre pour la confidentialité que les produits descriptifs. En général, les modèles comme les modèles logistiques ou de régression sont diffusables tant que le nombre total d'unités valides utilisées dans l'estimation du modèle est supérieur au seuil du compte minimal pour la source de données. Dans certains cas, les modèles sont équivalents à de simples chiffres ou à de simples statistiques descriptives; ces modèles doivent alors être vérifiés comme s'il s'agissait de chiffres ou de statistiques descriptives. Par exemple, un modèle de régression comme :

État de santé mentale positif = (constante) + bêta (sexe) + erreur

crée un produit qui équivaut à la moyenne des états de santé mentale positifs pour chaque niveau de sexe (ou les chiffres/proportions de chaque sexe dans un modèle de régression logistique). Dans le produit ci-dessous, le codage du sexe est « hommes = 0 et femmes = 1 » et la somme de la constante et du coefficient bêta non normalisé donnent la moyenne de la variable dépendante lorsque le sexe = 1 (hommes).

| **Données descriptives** | | | | **Produit du modèle** | |
|---|---|---|---|---|---|
| | Chiffres | Moyenne pour la variable dépendante | | Constante | 13,94 |
| **Hommes = 1** | 667 | 13,91 | | Bêta pour le sexe | -0,03 |
| **Femmes = 2** | 875 | 13,88 | | | |
| | | | | 13,945 + (-0,03) = 13,91 | 13,94 + (2* - 0,03) = 13,88 |

En général, tout modèle de régression comportant une seule variable indépendante qui est dichotomique ou une seule variable indépendante représentée par des catégories nominales est considéré comme un modèle saturé (c'est-à-dire un modèle où chaque effet possible est précisé) et équivaut à un tableau descriptif; il doit être contrôlé en conséquence. Les modèles saturés comprennent également les modèles avec des effets d'interaction où toutes les variables indépendantes interagissent les unes avec les autres. Par exemple, un modèle comportant les trois variables indépendantes de sexe, de revenu et d'état matrimonial serait saturé si les effets indépendants suivants étaient précisés : sexe revenu état matrimonial sexe*revenu sexe*état matrimonial revenu*état matrimonial sexe*revenu*état matrimonial.
Pour ces types de modèles, il faut consulter l'analyste responsable du contrôle au sujet des documents justificatifs requis.

### Diagnostics du modèle

De nombreuses approches de régression fournissent des renseignements diagnostiques comme les facteurs d'inflation de la variance (FIV) et les diagrammes ou tableaux résiduels qui sont utilisés pour évaluer l'adéquation du modèle. Voici quelques lignes directrices à prendre en compte lorsque vous demandez des diagnostics à partir de modèles de régression :
(1) Les renseignements diagnostiques sommaires comme la distance de Cook et les FIV peuvent être diffusés à condition que le coefficient de détermination du modèle ne soit pas supérieur à 0,90.
(2) Les nuages de points représentant des résidus ne peuvent pas être diffusés.

### Tailles d'échantillons des modèles

De nombreux progiciels fournissent les tailles d'échantillon non pondérées des modèles dans leur produit. Les tailles d'échantillon de ces modèles doivent être examinées avec soin, car les tailles d'échantillon de plusieurs modèles pourraient être utilisées pour générer un tableau descriptif, et les directives sur le contrôle de la confidentialité de certaines sources de données n'autorisent pas la publication de descriptifs non pondérés (qui incluent les chiffres totaux). Si les utilisateurs de données sont préoccupés par l'inflation de la taille de l'échantillon due à l'utilisation de pondérations et par son effet sur les statistiques de test du modèle, ils peuvent utiliser des pondérations normalisées.

Si la source de données le permet, le produit du modèle peut être publié en format non pondéré et pondéré sans justification, tant que le produit du modèle à un niveau géographique détaillé n'est pas demandé. Lorsque des modèles pondérés et non pondérés sont publiables, la publication d'un modèle non pondéré n'a aucune incidence sur le seuil

minimum des futures publications descriptives. Cependant, si le modèle est équivalent à un tableau (par exemple, une variable indépendante catégorique unique avec tous les niveaux spécifiés, ou un modèle saturé), alors ce modèle est vérifié comme un produit descriptif, et cela peut avoir un impact sur les futures versions du produit. Pour les modèles non pondérés (si la diffusion est autorisée), les utilisateurs de données doivent vérifier la taille totale d'échantillon utilisée dans les modèles non pondérés particuliers qui sont contrôlés pour s'assurer que l'exigence du seuil triplé est respectée, en plus de vérifier tous les autres aspects pertinents du modèle (par exemple, les modèles saturés, les graphiques, les descriptifs qui font partie du produit fourni par le logiciel).

Vous trouverez ci-dessous les exigences en matière de documents justificatifs pour chaque type de produit de modèle, **pour (1) les demandes de contrôle pour produits non pondérés**, et pour (2) **les demandes de contrôle pour produits pondérés (y compris celles avec arrondissements ou pondérations bootstrap**) :

| Type de produit faisant l'objet d'une demande | | Type de chiffres justificatifs |
|---|---|---|
| | | |
| **Moindres carrés ordinaires (MCO) avec une seule variable indépendante dichotomique** | | Chiffres non pondérés pour la variable indépendante dichotomique pour tous les cas valides de la variable dépendante |
| **MCO avec une seule variable indépendante catégorique 3+.** | | Chiffre total non pondéré pour le modèle; si la variable indépendante catégorique est codée en variables nominales, fournir les chiffres non pondérés de toutes les catégories de la variable indépendante pour les cas valides de la variable dépendante |
| **MCO avec une seule variable dépendante continue** | | Chiffre total non pondéré pour le modèle |
| **MCO avec plusieurs variables indépendantes – modèles non saturés** | | Chiffre total non pondéré pour le modèle |
| **MCO avec plusieurs variables indépendantes – modèles saturés** | | Traiter comme un tableau croisé entre les variables indépendantes pour tous les cas valides de la variable dépendante |
| **Modèle logistique avec une seule variable indépendante dichotomique** | | Traiter comme un tableau croisé entre la variable indépendante et la variable dépendante |
| **Modèle logistique avec une seule variable indépendante catégorique 3+** | | Chiffre total non pondéré pour le modèle; si la variable indépendante catégorique est codée en variables nominales, traiter comme un |

| | | |
|---|---|---|
| | | tableau croisé |
| **Modèle logistique avec une seule variable indépendante continue** | | Chiffres non pondérés pour la variable dépendante dichotomique pour tous les cas valides de la variable indépendante |
| **Modèle logistique non saturé avec plusieurs variables indépendantes** | | Chiffre total non pondéré pour le modèle |
| **Modèle logistique saturé avec plusieurs variables indépendantes** | | Traiter comme un tableau croisé entre les variables indépendantes pour tous les cas valides de la variable dépendante |

# Graphiques

Les graphiques peuvent représenter les données de nombreuses façons : points de données individuels, tracé des moyennes ou des coefficients de modèle, ou valeurs prédites. Le contrôle de la confidentialité relatif aux graphiques peut être compliqué, et tout graphique dont la publication est envisagée doit faire l'objet d'une discussion avec l'analyste du contrôle de la confidentialité.

En général, les graphiques tels que les nuages de points ou les graphiques de valeurs résiduelles qui montrent des points de données individuels ne sont pas autorisés pour la diffusion. Les graphiques à surfaces peuvent également poser problème, car ils peuvent générer des valeurs aberrantes individuelles. Toutefois, s'il peut être démontré que chaque point individuel du graphique atteint le seuil minimum des chiffres pour la source de données utilisée, le graphique peut être publié. En termes de résolution d'image, cela signifie que chaque pixel doit être équivalent au moins au nombre minimum d'unités d'analyse de données dont la publication est autorisée pour la source de données analysée.

Les graphiques qui sont des représentations de descriptifs, tels que les fréquences par cellules (par exemple, les histogrammes), doivent respecter les directives relatives à leur contrepartie descriptive. Dans le cadre de la documentation justificative d'un tel graphique, le tableau sous-jacent des chiffres non pondérés doit être fourni. Si le graphique est constitué de valeurs prédites par le modèle ou de coefficients de modèle provenant d'un modèle qui peut être diffusé, alors le graphique peut être publié. La diffusion de certains graphiques obtenus à partir de produits ne pouvant être diffusés ayant fait l'objet d'une transformation comme un lissage peut être acceptée après discussion avec l'analyste responsable du contrôle de la confidentialité.

StatCan restreint également le type de fichiers graphiques qui peuvent être publiés. Les graphiques en format.jpg,.gif,.tiff ou.png peuvent être publiés à condition que le graphique lui-même puisse l'être. Certains logiciels, tels que STATA, peuvent produire des fichiers de sortie de graphiques qui contiennent des microdonnées intégrées au graphique. Ces types de fichiers de sortie de graphiques ne sont pas publiables, à moins qu'ils ne soient convertis dans un format de fichier publiable sans données intégrées. L'organigramme suivant peut être une aide pour déterminer si un graphique est publiable :

## Diagramme du processus de contrôle de la confidentialité des graphiques

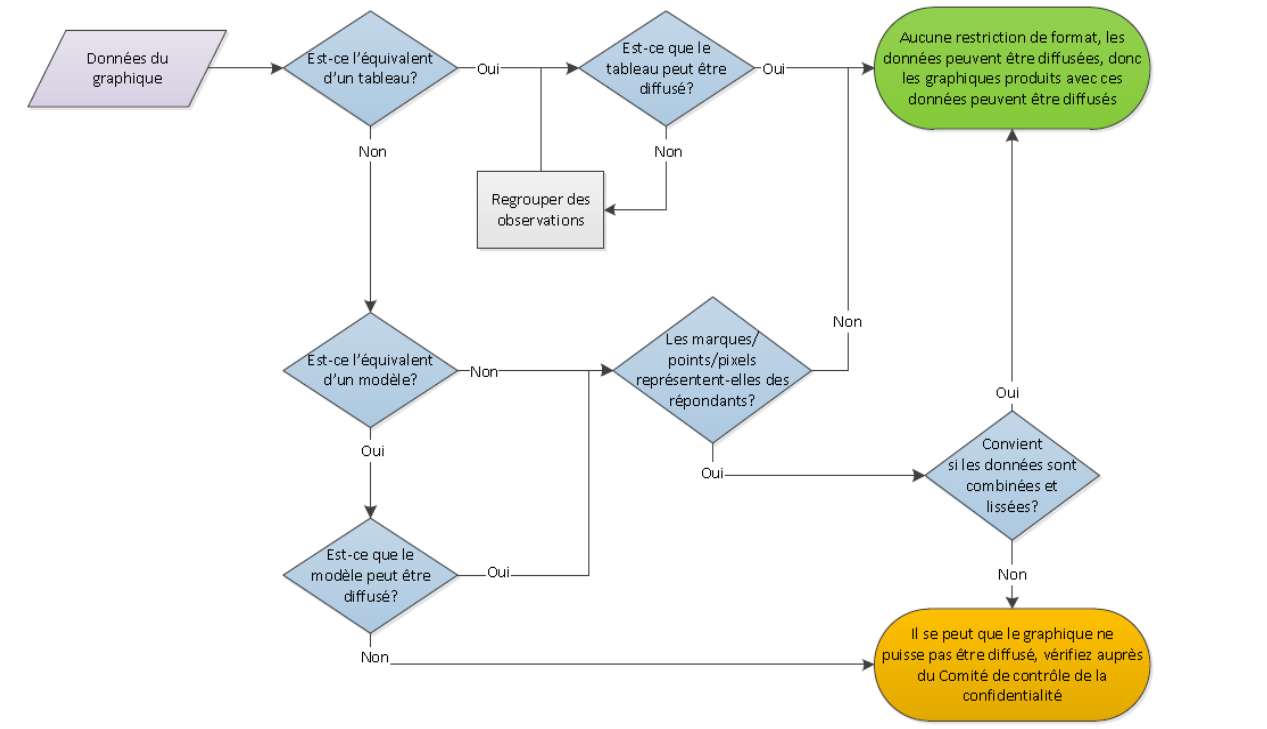



**Tableau pouvant être diffusé:**
– Nombre minimum de répondants par cellule
– Respect de toutes les lignes directrices propres à l'enquête

**Modèle pouvant être diffusé:**
– Non saturé/non équivalent au tableau
– Respect de toutes les lignes directrices propres à l'enquête

**Exigences supplémentaires pour les graphiques de tracé séquentiel et produits semblables:**
– Nombre minimum de répondants/de pixels dans la dimension répondant
– Format graphique « raster » seulement (PNG, TIFF, JPEG, MP)

Comité de la confidentialité des CDR, Mai 2015

# Corrélations

En général, les coefficients de corrélation sont considérés comme publiables à condition que le seuil de N observations soit atteint. Cependant, le risque peut augmenter dans les cas où le coefficient de corrélation est exactement 1 (-/+) et où des statistiques descriptives telles que la médiane sont également présentes dans le produit, et si des corrélations avec des variables dichotomiques sont produites.

Par exemple, s'il existe une corrélation directe parfaite (= 1) entre le roulement et les coûts d'emploi pour un échantillon d'entreprises, et que la valeur médiane de chacune des variables est également présentée, alors ces valeurs médianes se rapporteront à l'entreprise qui se trouve à la médiane, puisque la relation entre les deux variables est parfaitement corrélée. Cela peut augmenter le risque que l'entreprise soit identifiée et que des informations confidentielles lui soient associées. Le tableau suivant présente les renseignements justificatifs qui doivent être fournis pour toute corrélation dont le contrôle de la confidentialité est demandé.

| | | Variable 2 | | |
|---|---|---|---|---|
| Variable 1 | | **Dichotomique** | Catégorique | **Continue** |
| | | | | |
| **Dichotomique** | | Tableaux croisés non pondérés | | |
| **Catégorique** (comporte trois catégories ou plus et n'est pas divisée en variables nominales dichotomiques) | | Chiffres non pondérés pour la variable dichotomique pour les valeurs non manquantes de la variable catégorique | Chiffre total non pondéré de valeurs non manquantes pour les deux variables | |
| **Continue** | | Chiffre non pondéré de la variable dichotomique pour les valeurs non manquantes de la variable continue | Chiffre total non pondéré de valeurs non manquantes pour les deux variables | Chiffre total non pondéré de valeurs non manquantes pour les deux variables |

Une variable catégorique est une variable continue qui est recodée en trois catégories ou plus (par exemple, le nombre d'années d'études), un élément de type Likert avec trois options de réponse ordonnées ou plus, ou une variable qui a trois catégories discrètes non ordonnées ou plus (par exemple, l'état matrimonial). Certaines sources de données peuvent avoir des seuils de chiffres pondérés supplémentaires qui doivent également être atteints en plus de ceux décrits dans le tableau ci-dessus pour qu'une corrélation soit validée.

Quelques types courants de corrélations :
Pearson – utilisé avec deux variables continues
Rang de Spearman/Kendall Tau – utilisé avec deux variables ordinales et parfois des variables catégoriques
Coefficient Phi/Tetrachoric/Polychoric – pour deux variables catégoriques ou dichotomiques
Point Biserial – pour une variable catégorique corrélée à une variable continue

## Analyse des données de survie

L'analyse de survie est un groupe de techniques liées à l'estimation du temps qui s'écoule jusqu'à un certain événement pour les répondants, et peut inclure des tableaux croisés justificatifs tels que les tableaux de survie, des modèles paramétriques tels que les risques proportionnels de Cox et l'ajustement de courbes paramétriques aux fonctions de survie, et des techniques non paramétriques telles que la méthode du produit-limite. Le principe qui sous-tend le contrôle des produits d'analyse de données de survie est de faire correspondre un résultat à des règles tabulaires ou à des règles de modèle. En général, les tableaux de survie et les méthodes non paramétriques utilisent des règles de contrôle tabulaire et les produits de modèles paramétriques utilisent des règles de contrôle de modèle.

Toute fonction de survie actuarielle (table de survie) peut être validée lorsque : le nombre d'événements par intervalle ou par cycle et le nombre d'unités restantes sont tous deux égaux à zéro ou au moins à la taille minimale de la cellule pour la source de données. Les chiffres censurés pour cause de non-réponse ou de perte de contact n'ont pas à atteindre des seuils minimums et ne posent pas de problème en ce qui concerne le contrôle de la confidentialité. Lors de la création d'intervalles, il n'est pas nécessaire que les intervalles soient de longueur égale, et les intervalles peuvent être regroupés afin de générer un tableau de survie publiable. Toute courbe de survie graphique ou fonction de risque provenant d'un tableau de survie publiable est également publiable.

Pour toute méthode de construction de fonction d'analyse de données de survie (tableau de survie, produit-moment ou méthode connexe comme Breslow, etc.), les coefficients (bêta) calculés pour un modèle de risques proportionnels de Cox ou les paramètres d'une courbe d'ajustement à une fonction de survie/risque sont publiables tant que le modèle est un modèle publiable – c'est-à-dire non équivalent à un tableau, taille d'échantillon suffisante, etc. Les courbes de survie ou de risque lissées par LOESS ou LOWESS (régression locale) avec une largeur de bande supérieure à la taille minimale de la cellule ou générées à partir d'un tableau de survie publiable peuvent être publiées. Les coefficients, les courbes lissées et les estimations des paramètres de distribution peuvent être libérés s'ils sont dérivés d'une fonction calculée avec la méthode du produit-limite ou directement à partir de données brutes, même s'ils accompagnent une courbe ou une fonction basée sur les tableaux de survie.

En ce qui concerne les graphiques, les graphiques publiables doivent être en format.jpg,.gif,.tiff ou.png – les autres formats peuvent contenir des données ou être réversibles en données. Une courbe à très basse résolution où les événements individuels sont indiscernables peut être publiable (c'est-à-dire que la résolution doit être telle que le plus long pas vers le bas soit de 1/5 de pixel verticalement, ou que chaque pixel horizontal ait une largeur d'au moins 5 événements). Les splines ajustées à une courbe de survie de produit-limite peuvent être par morceaux, et chaque morceau doit couvrir au moins *(m + d)* événements, où *m* est la taille minimale de la cellule et *d* est le degré de la courbe. Une courbe LOESS peut répondre à bon nombre des mêmes besoins et être plus simple à mettre en œuvre.

Vous trouverez ci-dessous les exigences en matière de documents justificatifs pour chaque type de produit de survie, **à la fois pour (1) les demandes de contrôle pour produits non pondérés** et (2) **les demandes de contrôle pour produits pondérés (y compris celles avec arrondissement ou pondérations bootstrap**) :

| Type de produit faisant l'objet d'une demande | | Type de chiffre de soutien |
|---|---|---|
| | | |
| **Fonction de survie/risque ou tableau de survie** | | Tableau de survie non pondéré |

**Il existe plusieurs types de produits basés sur la survie qui ne sont pas publiables :**

- Tout tableau de survie qui ne regroupe pas les événements en intervalles ou les survivants à la fin de la période d'observation pour respecter la taille minimale de la cellule (généralement, la méthode du produit-limite/Kaplan-Meier ou Breslow génère un tel tableau, à moins que l'ensemble de données ne soit énorme ou que la variable temporelle ait une faible précision).
- Toute **courbe de survie ou de risque** qui présente des étapes, des inflexions, des lacunes, des points ou des marques distinctes à chaque événement ou une censure, à moins qu'elle ne soit basée sur un tableau de survie publiable.
- **Interpolation de Lagrange** (ou autre interpolation qui passe par tous les points) sur une courbe ou un tableau non publiable, sauf si elle est basée sur un tableau de survie publiable.

## Analyse factorielle exploratoire

Cette méthode est également connue sous le nom d'analyse en composantes principales. Une version plus avancée de cette méthode est connue sous le nom d'analyse factorielle confirmatoire. L'objectif de l'analyse factorielle exploratoire (AFE) est de déterminer les relations sous-jacentes entre les variables mesurées, et elle peut être utilisée pour réduire un grand nombre de variables en un ensemble plus petit de « facteurs » qui peuvent être utilisés pour représenter cet ensemble plus grand. Certaines applications de l'AFE consistent à créer des « scores factoriels » à utiliser dans des

modèles ou pour déterminer si des ensembles d'éléments qui ont un point commun sous-jacent peuvent être utilisés pour créer un score d'indice (par exemple, si 6 éléments d'un ensemble de 10 éléments créent un facteur, ces 6 éléments peuvent être utilisés pour créer une variable dérivée ou un score d'indice).

### Vérification des produits de l'AFE

La plupart des progiciels produiront par défaut une grande quantité de données, mais certains permettent de contrôler ce qui est produit. Il est recommandé que l'utilisateur des données et l'analyste responsable du contrôle de la confidentialité discutent des parties du produit qui doivent être contrôlées et de celles qui peuvent être supprimées du produit. Pour la plupart des produits générés, si le modèle d'AFE répond aux exigences de contrôle de la confidentialité pour la source de données utilisée, tous les produits sont publiables, à l'exception des sections qui peuvent contenir des chiffres non publiables selon les règles de contrôle de la confidentialité (par exemple, les sections qui contiennent des chiffres non pondérés lorsque les produits non pondérés ne sont pas autorisés). Certains produits contiendront des données descriptives en même temps que les produits du modèle (par exemple, la taille des échantillons ou leur nombre, les moyennes et les écarts-types, les corrélations) – ces données peuvent être vérifiées conformément aux directives correspondantes pour la source de données utilisée.

**Vous trouverez ci-dessous les documents justificatifs requis pour l'analyse factorielle exploratoire :**
(1) La taille totale de l'échantillon utilisé dans l'analyse doit répondre aux exigences de taille minimale de l'échantillon pour les données utilisées.
(2) Si l'analyse utilise des variables dichotomiques/nominales ou catégoriques, un document justificatif des fréquences ou des tableaux croisés peut être requis.

## Modélisation par équations structurelles et analyse des chemins

La modélisation par équations structurelles, ou MES, est une approche permettant d'analyser des modèles simples et complexes basés sur la régression. Elle peut également être appelée « analyse de la structure des covariances » ou « modélisation de la structure des covariances ».
Comme pour la plupart des approches de régression, les variables catégoriques et continues peuvent être utilisées dans la MES, et la plupart des programmes/approches de MES supposent que des variables continues sont utilisées, à moins que l'utilisateur de données ne l'indique explicitement dans sa syntaxe d'analyse. Il existe deux principaux types de modèles d'équations structurelles : les modèles de chemin et les modèles de structure.

### Modèles de chemin

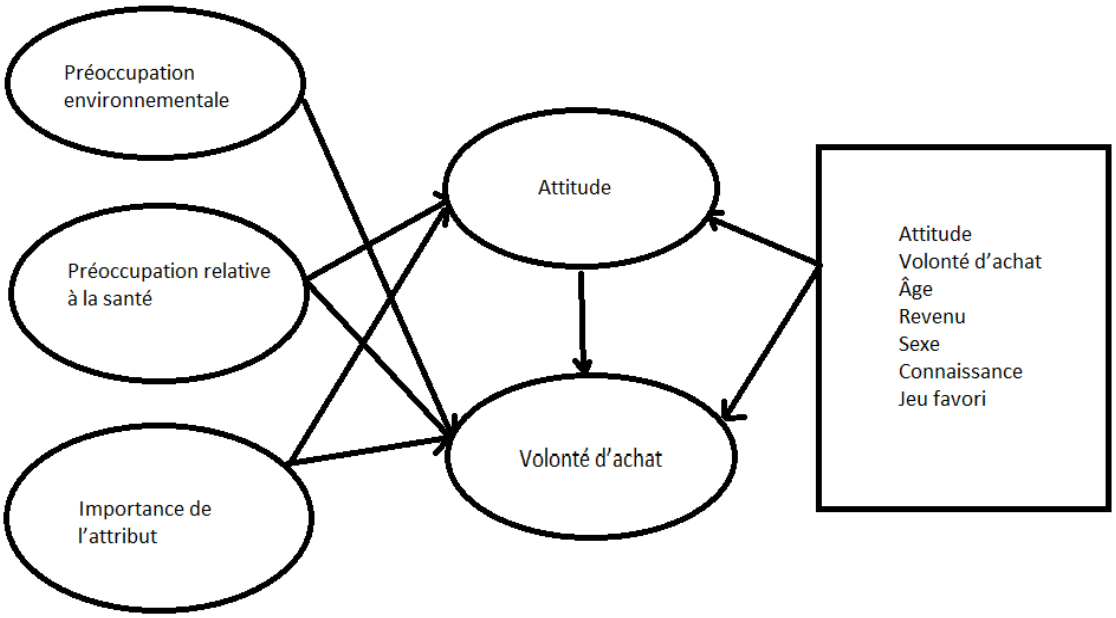
Les modèles de chemin peuvent être très semblables aux modèles de régression dans la mesure où ils n'utilisent que des variables observées (c'est-à-dire qu'aucune variable latente ou non observée n'est spécifiée). Cependant, les modèles de chemin peuvent également impliquer des types plus complexes de modèles de régression, tels que la régression à variable instrumentale et les régressions avec des erreurs corrélées entre les prédicteurs.

Les domaines suivants doivent être examinés attentivement lors de la préparation d'une demande de contrôle pour tout produit de modèle de chemin :
(1) Le modèle répond-il aux exigences de taille minimale d'échantillon pour les données utilisées?
(2) Si le modèle utilise des variables dichotomiques/nominales ou catégoriques, il peut être nécessaire de fournir des tableaux de fréquence ou des tableaux croisés si le modèle de cheminement est équivalent à un modèle de régression saturé ou un tableau descriptif. Pour les modèles de cheminement univariés avec une variable dépendante continue (et une seule variable indépendante dichotomique), un document justificatif avec les chiffres de la variable indépendante

dichotomique pour les valeurs non manquantes de la variable dépendante doit être fourni (cela compte également pour une variable catégorique codée comme étant nominale). Pour les modèles univariés avec une variable dépendante dichotomique et une variable indépendante dichotomique, un document justificatif avec les tableaux croisés des deux variables doit être fourni.
(3) Si une approche analytique implique l'analyse de modèles de chemin sous forme d'équations de régression distinctes, comme un modèle à variables instrumentales (par exemple, ivregress, ivreg ou ivreg2 dans STATA), les processus de contrôle appropriés pour chaque modèle distinct doivent être menés. Il convient de fournir un document justificatif approprié contenant la fréquence par cellule marginale pour toutes les régressions individuelles.
(4) Les modèles successifs ne doivent pas créer de produits pouvant être combinés afin de créer un tableau de statistiques descriptives.
(5) Les produits non normalisés pour les coefficients des modèles peuvent avoir une valeur très semblable à celle des produits descriptifs correspondants. Cependant, si le modèle est publiable, alors les produits non normalisés le sont aussi.

### Modèles d'équations structurelles

Les modèles d'équations structurelles diffèrent des modèles de chemin dans la mesure où les relations sont spécifiées entre des variables non observées ou « latentes » – il s'agit de concepts qui ne sont pas présents dans l'ensemble des données en tant que variables, mais qui sont supposées être le concept sous-jacent à un ensemble de variables. En général, l'intérêt porte sur les relations entre les concepts latents et les relations entre les variables observées; leurs concepts latents respectifs sont secondaires. Un modèle structurel typique est donné ci-dessous :




Le produit d'un modèle d'équation structurelle peut être long – il y a des sections qui évaluent l'ajustement global du modèle, des sections qui évaluent chacun des paramètres individuels du modèle, et des sections qui évaluent les modifications du modèle.

Pour les modèles dont le produit normalisé fait l'objet d'une demande de contrôle, tant que le modèle non pondéré N dépasse le seuil minimum requis et que le modèle n'est pas saturé, le modèle peut être libéré. Les produits après estimation (tels que les indices d'ajustement et les indices de modification du modèle) peuvent également être publiés tant que le modèle N dépasse le seuil minimum requis.

Pour les modèles dont les produits ne sont pas normalisés et qui font l'objet d'une demande de contrôle, il convient d'appliquer les mêmes directives de contrôle que pour les modèles normalisés (c'est-à-dire que l'exigence de contrôle pour les coefficients du modèle sera la même que pour les statistiques descriptives).

**Vous trouverez ci-dessous les documents justificatifs requis pour l'analyse de la MES/du chemin :**

| Type de modèle | Semblable à | Produit devant faire l'objet d'un contrôle de la confidentialité | Type de chiffres justificatifs |
|---|---|---|---|
| A -- > B | Régression; logistique si B est dichotomique | Taille totale de l'échantillon | Tableaux croisés si des variables dichotomiques sont utilisées comme prédicteurs. |
| A -- > B -- > C | Régression<br>- Médiation/instrumentale | Taille totale de l'échantillon | Tableaux croisés si des variables dichotomiques sont utilisées comme prédicteurs. |
| A -- > B<br>^<br>C -----┘ | Régression; logistique si B est dichotomique | Taille totale de l'échantillon | Tableaux croisés si des variables dichotomiques sont utilisées comme prédicteurs. |

## Demandes de contrôle volumineuses

Étant donné que le contrôle de la confidentialité des produits intermédiaires est découragé, il est possible que les demandes de contrôle soient volumineuses ou complexes si l'équipe de recherche limite ses demandes de contrôle. Avant de soumettre une demande volumineuse, les utilisateurs de données doivent consulter l'analyste responsable du contrôle de la confidentialité. Pour les demandes de contrôle volumineuses, le formulaire de demande de contrôle est essentiel, tant pour les utilisateurs de données que pour l'analyste responsable du contrôle, car il permet de s'assurer que les questions de confidentialité ont été traitées et que les fichiers à diffuser et les documents justificatifs appropriés sont organisés.
Les utilisateurs de données doivent garder à l'esprit les meilleures pratiques suivantes pour soumettre des demandes de contrôle volumineuses :

- Seuls les produits nécessaires à la publication doivent être demandés.
- L'analyste responsable du contrôle de la confidentialité doit être en mesure de départager facilement les produits pour déceler la présence de risques résiduels.
- Les produits descriptifs et les produits de modèle doivent être séparés en différents fichiers par l'utilisateur de données et, au besoin, séparés en dossiers basés sur des variables de sous-échantillon ou de filtre (par exemple, des dossiers distincts pour les analyses stratifiées par sexe). Cela aidera à déterminer les cellules résiduelles potentielles causées par d'éventuels échantillons qui se chevauchent ou par un léger changement de recodage des variables.

Les utilisateurs de données qui soumettent des demandes de contrôle très volumineuses ou complexes, alors que la plupart des produits ne sont pas destinés à être publiés, pourraient être invités à fournir une justification écrite expliquant pourquoi un tel volume de produits est nécessaire, et une approbation d'un gestionnaire régional de la DAD pourrait être requise avant que les produits puissent être contrôlés ou publiés.

## Contrôle du produit à partir de prolongations/révisions d'un contrat en cours

Les prolongations sont la poursuite d'un projet jusqu'à une nouvelle date de fin, et les révisions sont de nouveaux contrats (ouverts dans le but de répondre aux commentaires des réviseurs). Dans les deux cas, aux fins du contrôle de la confidentialité, les prolongations et les révisions sont traitées comme une continuation du projet initial. Par conséquent, le contrôle des produits est vérifié par rapport au projet initial et au contrôle précédent afin de s'assurer qu'il n'y a pas de problèmes de divulgation par recoupements.

## Foire aux questions

***1. « Que faire si un chiffre ou modèle non pondéré ne répond pas aux exigences minimales relatives à la taille de l'échantillon » (ou « J'ai besoin de cette cellule à faible fréquence, car YYY est ma variable d'intérêt. »)***
Si des chiffres ne répondent pas aux exigences minimales, ces chiffres et tout produit descriptif ou modèle associé ne peuvent pas être diffusés (p. ex. moyennes, modèles à variable indépendante unique). Si des tableaux ont des cellules qui ne répondent pas aux exigences minimales, des options ou catégories de réponse doivent être regroupées. Dans le cas de graphiques (p. ex. histogrammes, courbes de survie), le groupement par classe pourrait être nécessaire avant que la diffusion ne soit permise.

Les valeurs manquantes (p. ex. Ne sait pas, Refus) ne sont généralement pas touchées par les exigences de taille d'échantillon minimale.
Une exception est le « saut valide » qui peut permettre d'identifier les répondants ayant fourni une réponse particulière à une question précédente dans l'enquête.

Il convient de noter que certaines sources de données ont également des exigences minimales pondérées. Il est recommandé que les utilisateurs de données consultent l'analyste responsable du contrôle de la confidentialité et, dans la mesure du possible, celui-ci fera des suggestions pour remédier à la situation. Toutefois, l'analyste responsable du contrôle de la confidentialité doit respecter les règles connexes, et il arrive qu'aucune publication ne soit possible.

***2. « Je veux à la fois la sortie non pondérée et pondérée »***
Certaines sources de données permettent de publier les deux types de produits (avec les documents pertinents), d'autres permettent de publier uniquement les produits descriptifs pondérés et les produits des modèles non pondérés et pondérés, tandis que d'autres encore permettent de publier uniquement les produits pondérés. Le seuil minimum pour la publication d'un produit non pondéré est supérieur à celui de la publication d'un produit pondéré et, lorsque la publication d'un produit non pondéré est demandée, tous les produits futurs (pondérés ou non) sont vérifiés en utilisant ce seuil minimum augmenté. Les utilisateurs de données doivent consulter leur analyste responsable du contrôle de la confidentialité pour connaître le processus à suivre et discuter de tout éventuel risque en matière de confidentialité.

Si la source de données permet de publier à la fois des données non pondérées et des données pondérées, il est nécessaire de fournir une justification écrite que l'analyste responsable du contrôle de la confidentialité conserve dans ses dossiers. Si l'analyste estime que la justification est inadéquate, il peut la transmettre au Comité de la confidentialité

de la DAD. Une fois cette justification fournie, les données de sortie peuvent être publiées, à condition que le seuil minimal soit respecté. *NOTE* : Le seuil minimum est triplé pour la demande actuelle et **toutes les demandes** futures de sortie descriptive (pondérée ou non), peu importe l'échantillon d'analyse ou les variables incluses dans la demande de contrôle.

Comme il est indiqué ci-dessous, la taille totale d'un échantillon d'analyse et les descriptions générales de la population cible utilisée pour l'étude (par exemple, le nombre d'hommes/femmes) ne nécessitent pas de justification de la diffusion. Cependant, les tailles d'échantillon qui sont rapportées à partir de modèles de régression individuels ne sont pas visées par cette exception.

***3. « J'ai besoin de toutes les tailles d'échantillon et tous les chiffres non pondérés, car la revue les demandera de toute façon. »***

Nous comprenons qu'il est nécessaire de décrire l'échantillon dans les articles. Ces types de comptages peuvent être publiés, mais il est recommandé de retarder la demande de ces chiffres jusqu'à la toute fin des analyses. Cela permet d'éviter tout conflit avec des produits publiés précédemment qui pourraient conduire à l'identification de répondants ou d'unités d'analyse de données (par exemple, des entreprises). Voici quelques autres recommandations :
(a) Utiliser des tableaux stratifiés à une entrée (par exemple, par sexe)
(b) Indiquer les pourcentages pondérés plutôt que les chiffres réels
(c) Indiquer les chiffres non pondérés totaux pour chaque catégorie, puis indiquer le pourcentage manquant pour chaque cellule.
(d) Si une variable peut être de nature délicate (par exemple, les abus sexuels), consulter l'analyste responsable du contrôle de la confidentialité.

***4. « Je ne veux pas du tout me soucier de la pondération, donc tous mes produits seront non pondérés. »***

Un projet qui ne demande que des produits non pondérés pour le contrôle est autorisé à condition que (a) la source de données permette la publication de produits non pondérés et que (b) l'utilisateur de données soit conscient de l'augmentation du seuil pour la publication de ses produits. Si seul un produit non pondéré est demandé, cela doit être clairement indiqué dans la proposition. L'utilisateur de données devra suivre la procédure de justification s'il décide de demander un produit pondéré à tout moment dans l'avenir. Les produits qui utilisent des pondérations normalisées doivent également être traités avec les mêmes préoccupations.

***5. « Si j'utilise différentes sources de données avec différentes règles de contrôle, que dois-je faire? ».***

Si le projet analyse les sources de données séparément les unes des autres, alors chaque source de données est contrôlée avec ses règles respectives. Si le projet met en commun les sources de données et analyse l'ensemble des données mises en commun, le produit doit être capable de passer les règles de contrôle de la confidentialité pour TOUTES les sources mises en commun. Dans le cas des données couplées, il existe des règles propres aux couplages.

***6. « Existe-t-il d'autres techniques en plus d'utiliser des poids pour protéger la confidentialité? »***

L'arrondissement (déterministe, aléatoire, contrôlé) est une méthode qui est parfois utilisée pour protéger la confidentialité des données. Certaines sources de données exigent que les produits pondérés soient arrondis avant la diffusion, et d'autres exigent que les produits descriptifs (par exemple, les moyennes, les proportions, les pourcentages) soient fondés sur des produits arrondis afin de protéger la confidentialité. Les autres options possibles, après discussion avec l'analyste responsable du contrôle de la confidentialité, sont les suivantes :
(a) Troncation des valeurs extrêmes supérieures et inférieures / suppression des valeurs aberrantes.
(b) Les valeurs/réponses supérieures (ou inférieures) à un certain seuil sont codées avec la même valeur afin de regrouper les observations pour qu'elles répondent aux exigences minimales.

(c) Certaines sources de données nécessitent la réalisation de tests supplémentaires pour vérifier la confidentialité, tels que les tests de dominance et d'homogénéité pour les données quantitatives et les variables de revenu.

Les pratiques de contrôle de la confidentialité telles que la suppression des cellules, l'infusion de bruit et l'échange de données ne sont pas utilisées comme méthodes pour assurer la protection des données à la DAD. Les utilisateurs de données doivent savoir que l'arrondissement de tous les produits n'est pas considéré comme une protection définitive contre les problèmes de confidentialité par recoupements.

***7. « J'ai besoin de tout ce produit afin que mon conseiller ou chef de projet puisse décider de ce qui sera publié. »***
Il faut éviter les demandes volumineuses de contrôle de confidentialité ou les demandes multiples découlant d'une exploration des données et produisant de nombreux résultats intermédiaires. Si les utilisateurs de données ne sont pas tout à fait sûrs de ce qui est nécessaire, il est conseillé d'ajouter des membres de l'équipe au projet afin que chacun puisse voir le produit et déterminer ce qui doit être contrôlé. Une autre option consiste à faire contrôler les tendances ou une note de synthèse (par exemple, quelles variables sont significatives, si les coefficients bêta sont positifs ou négatifs, quels chiffres sont faibles/moyens/élevés).

***8. J'ai un délai imprévu, que puis-je faire pour contribuer à accélérer la publication des produits?***
Il est recommandé aux utilisateurs de données de rencontrer l'analyste responsable du contrôle de la confidentialité et de demander que seuls les produits absolument nécessaires soient publiés. Par exemple, au lieu de demander la publication d'un ensemble complet de produits, les utilisateurs de données peuvent apporter un ensemble de tableaux vierges qui seraient utilisés pour respecter le délai et demander la publication de ces tableaux uniquement (en tant que document justificatif, le produit brut qui a fourni les valeurs de ces tableaux devrait être fourni). Les utilisateurs de données peuvent revenir plus tard pour demander que d'autres produits soient contrôlés.

# Centre de données de recherche

# Formulaire de demande de contrôle de la confidentialité

| | | |
|---|---|---|
| Nom : | Courriel : | Date : |
| Nom d'utilisateur : | N° de contrat : | |
| Titre du projet : | | |
| Nom du dossier contenant les fichiers justificatifs et les fichiers à contrôler : | | |

Veuillez vérifier votre produit en fonction des lignes directrices relatives au contrôle; consultez l'analyste de votre centre de données de recherche en cas d'incertitude.
Le formulaire de demande rempli est conservé dans le registre des demandes.
Nota : Pour les étudiants et les assistants à la recherche, veuillez demander à vos superviseurs ou à votre équipe de recherche d'examiner le produit avant que sa diffusion soit demandée.

| **SECTION A. Liste de contrôle** | Oui/Non/S.O. |
|---|---|
| 1) Le produit demandé est-il conforme à la proposition approuvée pour ce projet? | Sélectionner |
| 2) Avez-vous subdivisé ou sélectionné seulement un certain ensemble de répondants à partir des données pour tout ou une partie de l'analyse? (P. ex. hommes de 50 ans et plus) Si oui :<br>a) Décrivez chacun des divers échantillons, sous-échantillons ou inclusions/exclusions utilisés pour produire votre produit à la **section C**. | Sélectionner |
| 3) Avez-vous vérifié les règles de contrôle afin de vérifier s'il existe des seuils pour des régions géographiques, des établissements, des tailles de ménages ou des populations pour votre produit? | Sélectionner |
| 4) Si cette demande repose sur des données couplées, décrivez la façon dont le couplage des données a été effectué à la **section D** (p. ex. fondé sur la personne, fondé sur les enregistrements, correspondances géographiques, etc.). | Sélectionner |
| 5) Cette demande de contrôle concerne-t-elle des variables liées au revenu, à la rémunération, à l'impôt ou à des valeurs en dollars? Si oui :<br>a) Vérifiez les lignes directrices et les exigences relatives au contrôle de ces types de variables. Consultez votre analyste au besoin.<br>b) S'il y a lieu, fournissez les renseignements suivants :<br>i) Les chiffres non pondérés à l'appui de l'échantillon;<br>ii) La syntaxe utilisée pour la création des variables, l'analyse et l'exécution des tests de contrôle;<br>iii) Les résultats des tests de contrôle (p. ex. tests d'ampleur, de dominance, etc.). | Sélectionner |
| 6) Cette demande comprend-elle des statistiques descriptives? Si oui :<br>a) Étiquetez clairement le produit (les tableaux ont un titre et chaque variable et catégorie est étiquetée). | Sélectionner |

| | |
|---|---|
| b) Assurez-vous que la taille minimale des cellules est respectée conformément aux règles applicables aux données.<br>c) Fournissez les documents justificatifs conformément aux règles de contrôle (p. ex. les chiffres sont non pondérés/pondérés/pondérés et arrondis). | |
| 7) Cette demande inclut-elle un produit de modèle ou des graphiques qui sont équivalents à une statistique descriptive? (P. ex. un modèle comportant une variable indépendante, un modèle comportant toutes les interactions possibles, histogrammes) Si oui :<br>a) Fournissez le tableau statistique non pondéré correspondant pour le nombre de répondants. | Sélectionner |
| 8) Avez-vous appliqué des poids modifiés (p. ex. normalisés) dans l'analyse? Si oui :<br>a) Décrivez pourquoi et comment les poids ont été modifiés à la **section C**. Consultez votre analyste au sujet des règles de contrôle relatives aux poids modifiés. | Sélectionner |
| 9) Cette demande comprend-elle une matrice des corrélations ou de covariance? Si oui :<br>a) Indiquez la taille de l'échantillon non pondérée pour les variables continues.<br>b) Fournissez le tableau croisé non pondéré pour les variables dichotomiques.<br>c) Fournissez les sous-totaux non pondérés pour les catégories d'une variable dichotomique corrélée avec une variable continue. | Sélectionner |
| 10) L'arrondissement des données du produit est-il requis pour cette demande de contrôle? Si oui :<br>a) Fournissez le produit en version arrondie **et** non arrondie.<br>b) Décrivez la méthode d'arrondissement et la base d'arrondissement.<br>c) Assurez-vous que tout arrondissement forcé à zéro est clairement indiqué. | Sélectionner |
| 11) Est-ce que le produit demandé est bel et bien votre produit final?<br>a) Si non, les futures demandes de contrôle envoyées en vertu de ce contrat pourraient être limitées à cause du risque de divulgation par recoupement. Nous vous encourageons fortement à consulter votre analyste. | Sélectionner |

| **SECTION B. Risque de divulgation par recoupement : Comparaison avec des demandes de contrôle précédentes** | Oui/Non |
|---|---|
| 1) Avez-vous des sous-ensembles de variables, où une ou plusieurs variables sont un sous-ensemble d'une autre ? (par exemple, si vous avez fait une dépression au cours des 5 dernières années et si vous avez fait une dépression l'année précédente ; si vous avez eu une maladie chronique et si vous avez eu une maladie respiratoire chronique) ?<br><br>2) Une version de ce produit, partielle ou intégrale, a-t-elle déjà été diffusée? Si non, passez à la **section C**. Si oui, avez-vous :<br>a. recodé ou modifié des variables, même légèrement?<br>b. changé le sous-échantillon ou la population d'intérêt?<br>c. abandonné des cas individuels ou des valeurs aberrantes?<br>d. imputé les valeurs manquantes? | Sélectionner |

| | |
|---|---|
| | |
| **Explication des changements :** | |
| **Si la réponse à l'une de ces questions est OUI, discutez-en avec votre analyste et consultez la section D pour connaître les exigences en matière de documents justificatifs relativement au risque de divulgation par recoupement.** | |

## SECTION C. Produit faisant l'objet de la demande de diffusion – Étiqueter clairement le produit

*Supprimez les produits ou les valeurs dont vous ne voulez pas ou que vous n'avez pas besoin de diffuser pour le moment*

| Nom du fichier (indiquez chaque feuille pour les fichiers de tableur) | Nom de l'enquête ou de l'ensemble de données et cycle(s) | Le cas échéant, nom de la variable de pondération (indiquez s'il s'agit d'une variable échelonnée ou normalisée) | Précisez la méthode utilisée en choisissant un numéro dans la liste ci-dessous | Description de l'échantillon (p. ex. femmes employées, âgées de 21 à 45 ans, en Ontario) | Niveau de géographie du produit demandé (c.-à-d. national, provincial, etc.) |
|---|---|---|---|---|---|
| 1. | | | | | |
| 2. | | | | | |
| 3. | | | | | |
| 4. | | | | | |
| … | | | | | |

## Types de méthodes

1. Méthodes descriptives (p. ex. modèles de régression avec une seule variable, fréquences, analyse par tableau croisé, moyennes et répartitions, matrice des corrélations, analyse de la variance)
2. Méthodes de mise à l'échelle (p. ex. analyse factorielle)
3. Graphiques (p. ex. histogrammes) – n'oubliez pas d'inclure des tableaux à l'appui.
4. Analyse de régression multivariée simple (p. ex. MCO, Logit, Probit, Tobit)
5. Méthodes de modélisation complexes (p. ex. modélisation par équations structurelles, modélisation linéaire hiérarchique, analyse de croissance, analyse de survie, analyse de données longitudinales, modèles d'équations simultanées, modèles à effets fixes, modèle à effets aléatoires)
6. Autres – veuillez décrire (p. ex. notes et fichiers de syntaxe)

## SECTION D. Fichiers justificatifs

Ces fichiers servent à appuyer la demande de contrôle et **ne seront pas diffusés**.
Veuillez nommer vos fichiers justificatifs pour permettre un couplage rapide avec le fichier de sortie correspondant.

### Placez ces fichiers dans votre dossier Documents justificatifs

1) Fichiers de syntaxe (ou fichiers journaux) à analyser
2) Fréquences justificatives (p. ex. non pondérées/pondérées/pondérées, arrondies)
3) Documents justificatifs pour les variables dérivées ou recodées (p. ex. syntaxe, livre des codes, description)
4) Autres fichiers requis selon les règles de contrôle applicables (p. ex. résultats de test pour les valeurs en dollars, etc.)
5) *Le cas échéant, documents justificatifs relativement au risque de divulgation par recoupement*
    a. Décrivez comment le produit se rapporte à ceux qui ont été diffusés précédemment.
    b. Indiquez la ou les dates des demandes de contrôle précédentes liées à cette demande.
    c. Fournissez les tableaux résiduels en tant que fichiers justificatifs. Veuillez vous reporter à l'orientation en matière de contrôle de la confidentialité.
    d. Fournissez les deux ensembles de syntaxe et mettez en surbrillance ou précisez les modifications.

| Nom du fichier | Notes |
|---|---|
| | |
| | |
| | |
| | |

SECTION E. Commentaires supplémentaires qui pourraient être utiles à l'analyste