Gestion de la confidentialité dans les tableaux de données agrégées

Le secret statistique : Pourquoi ? Comment ?

Maxime Beauté et Alexandre Awad – Département des méthodes statistiques

Insee

Mesurer pour comprendre

01 Le secret, pourquoi ?

1A Buts et enjeux

1B Cadre législatif

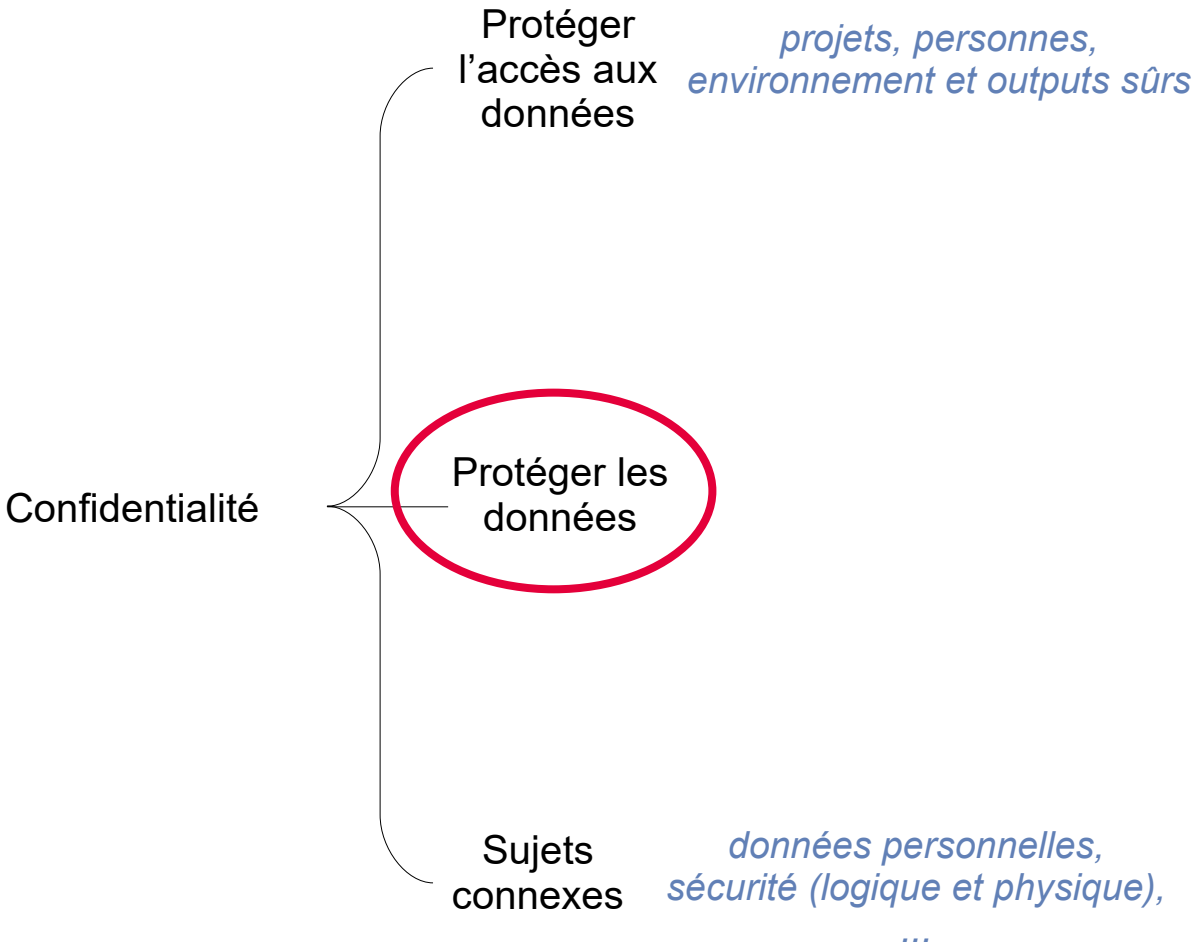
Champ de la formation

Confidentialité

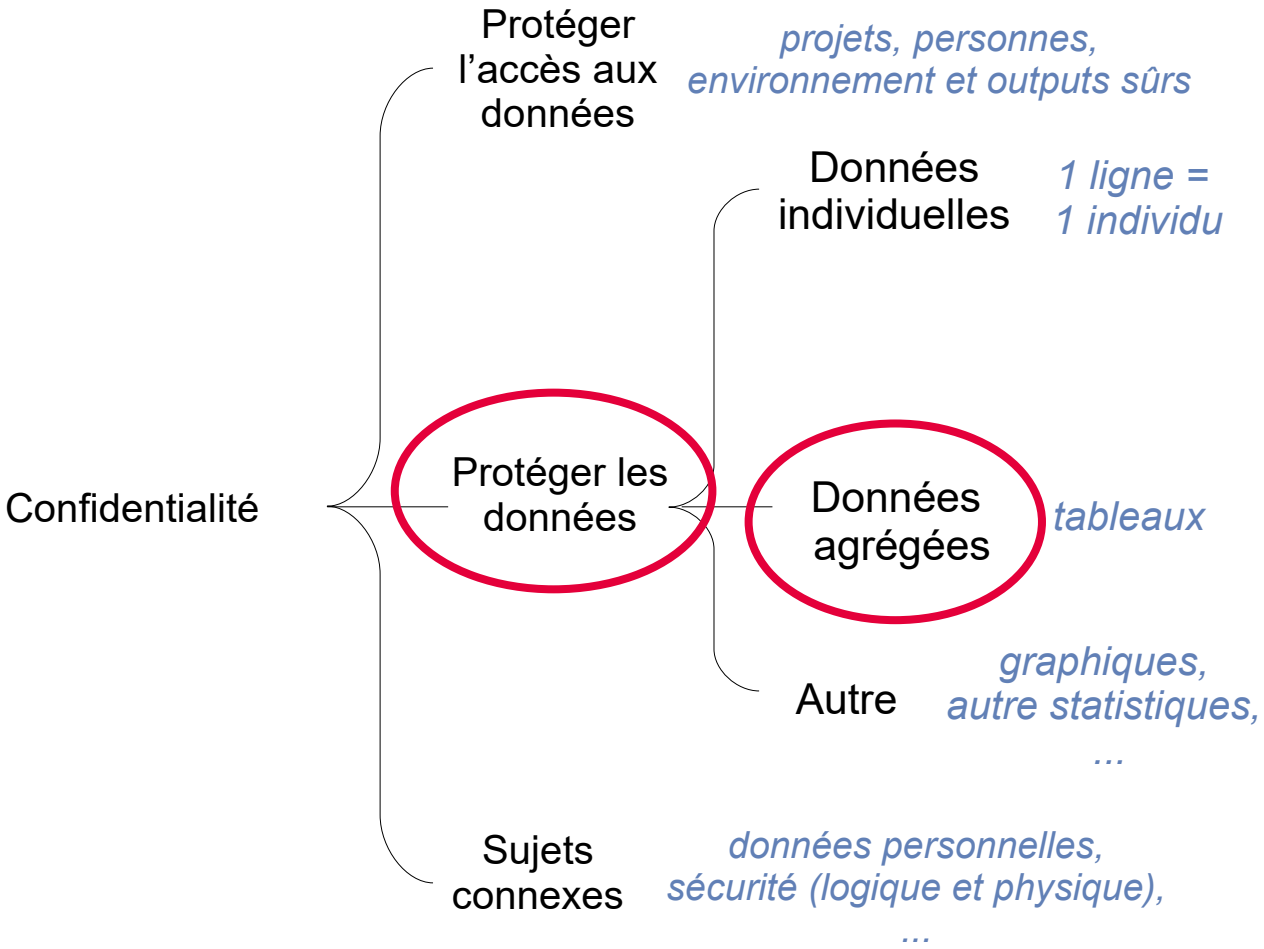
Champ de la formation

Confidentialité

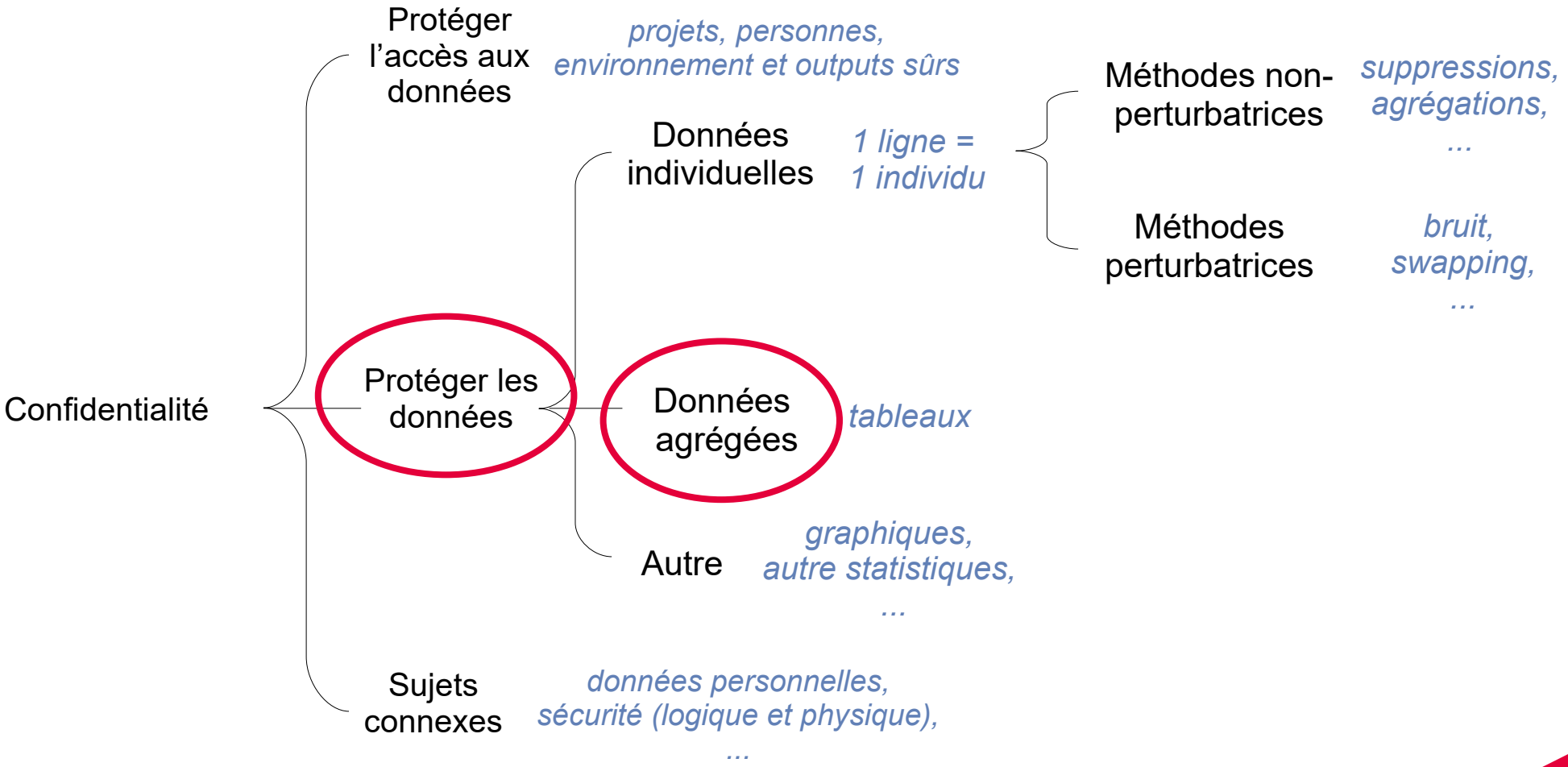
Champ de la formation



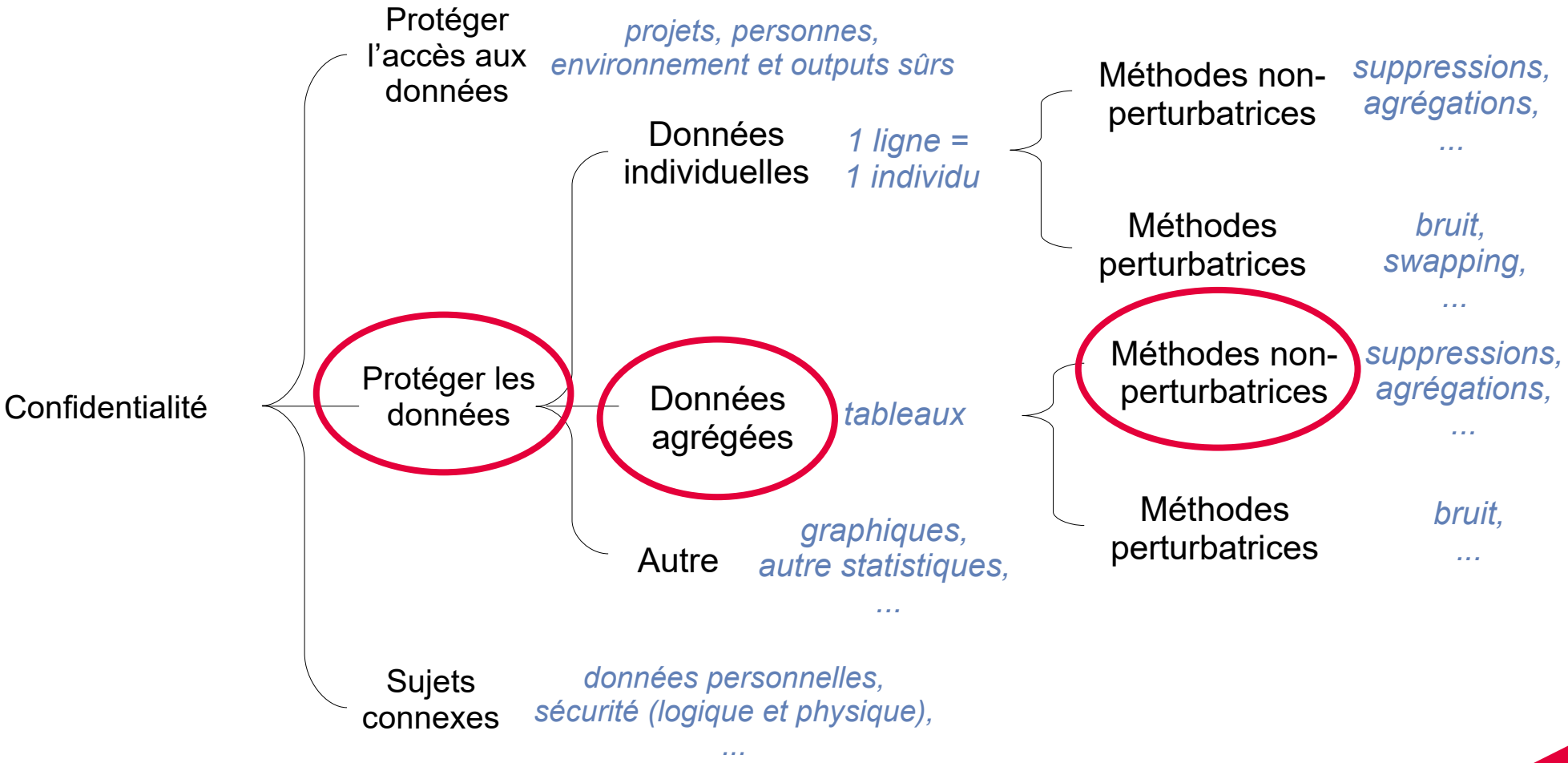
Champ de la formation



Champ de la formation



Champ de la formation



Pourquoi gérer la confidentialité statistique ?

-

- Pour garantir des traitements éthiques ;
- Pour être conforme au cadre légal ;
- Pour conserver la réputation, l'intégrité et la confiance du public et des répondants vis-à-vis de l'institut ;
- Et donc pour obtenir des résultats de meilleure qualité :
  - Ce qui entraîne des taux de réponse plus élevés ;
  - Ce qui offre un cadre propice à des réponses sincères aux questions sensibles.

Comment ?

-

- En restreignant l'accès aux données (hors champ de cette formation) ;
- En protégeant les données individuelles ;
- En empêcher la reconstruction de données individuelles ;

→ Sous l'hypothèse que l'attaquant possède de l'information auxiliaire :

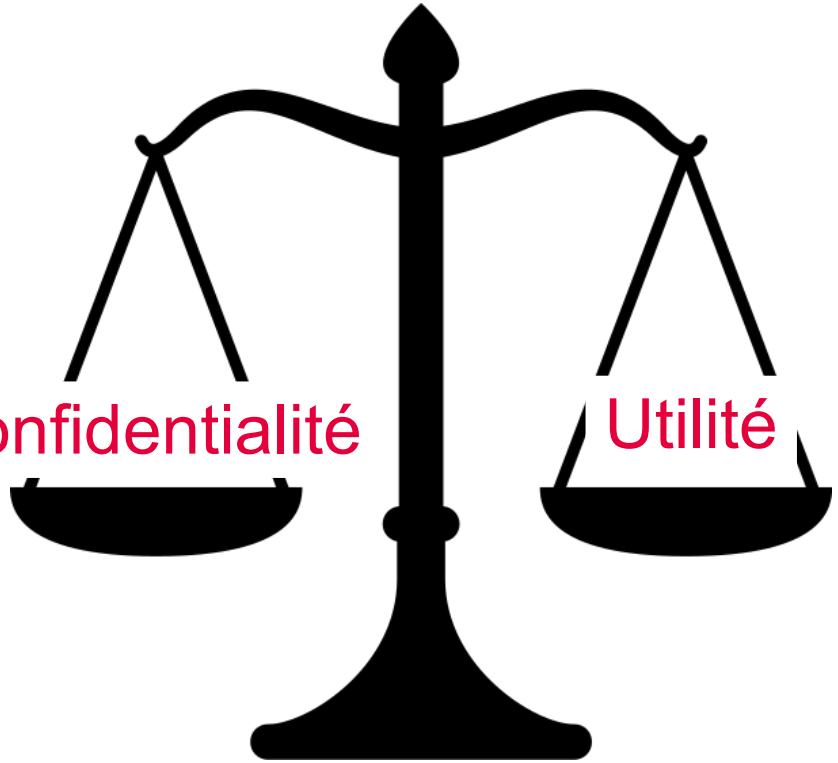
- incomplète, ou bien très fournie
- la sienne, dans tous les cas

→ Au regard de la loi, on fait l'hypothèse d'un attaquant employant des « moyens raisonnables » de ré-identification.

Buts et enjeux

- … Tout en cherchant à diffuser l'information la plus complète possible !

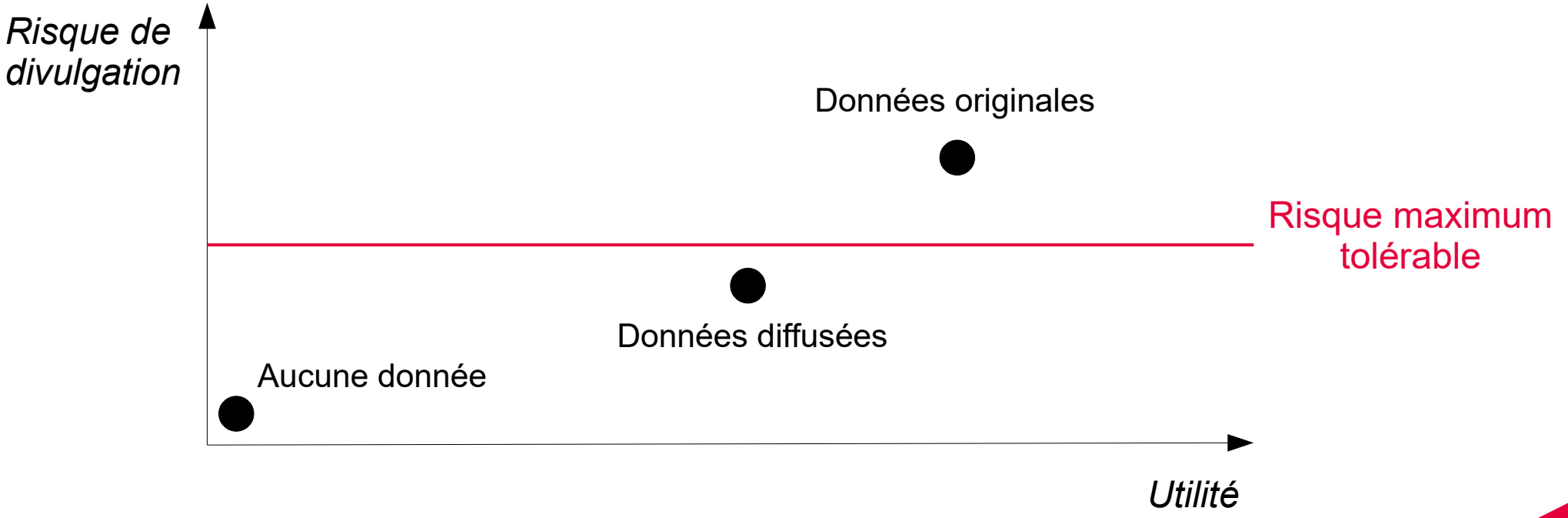
Minimiser le risque de divulgation



Minimiser la perte d'information

→ Quel équilibre confidentialité/utilité choisir ?

Buts et enjeux



4 types de divulgation d'informations

-

- Divulgation d'attribut : Se produit lorsque des informations confidentielles à propos d'un individu sont divulgués.

  → Par exemple : *Divulgation du fait qu'un individu est fumeur ou non.*

  - Elle n'implique pas toujours l'identification d'une personne en particulier.

    → Par exemple :

    - *Les personnes ayant la profession X dans l'entreprise Y gagnaient entre 50 000 € et 55 000 € l'année dernière.*
    - *Toutes les personnes ayant la maladie A ont subi le traitement B.*

4 types de divulgation d'informations

-

- Divulgation d'identité : Se produit lorsqu'un individu peut être identifié directement ou indirectement dans des données diffusées.

  → Par exemple : *Diffuser un numéro d'identifiant, une adresse e-mail, …*

4 types de divulgation d'informations

-

- Divulgation par recoupement : Se produit lorsque des informations diffusées sont combinées à d'autres informations diffusées, ou à des informations provenant de sources externes, pour révéler des données confidentielles.

  → Par exemple : *Par différenciation géographique entre communes et carreaux, on aurait pu isoler les caractéristiques de certains ménages.*

  - Ce risque augmente avec la quantité de données publiées par l'organisation.

  - D'où la nécessité de protéger de manière cohérente les données à l'origine de plusieurs tableaux ou publications.

4 types de divulgation d'informations

-

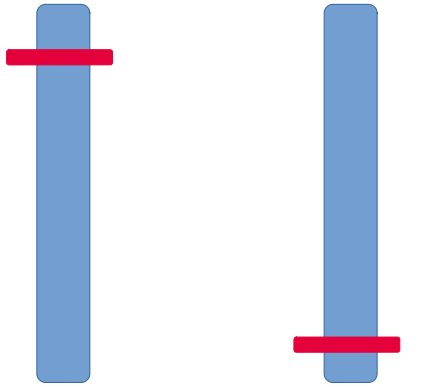
- Divulgation par inférence : Se produit lorsque des informations concernant un individu peuvent être inférés avec un haut niveau de confiance (ou un faible niveau d'incertitude).

  → Par exemple :

  - *connaître le revenu d'un ménage à 1 % près*
  - *ré-identifier quasi certainement un enquêté à partir d'un cumul d'informations le concernant : son sexe, son lieu de naissance, son mois de naissance, sa ville de résidence, sa profession, ses loisirs, …*

2 grandes approches pour gérer la confidentialité

-



Données sécurisées

Accès sécurisé

- Données sécurisées (restriction des données) : Protection contre la divulgation d'informations individuelles en modifiant les données elles-même.

  → Ex : *Méthode de suppression de cellules pour protéger les tableaux de données.*

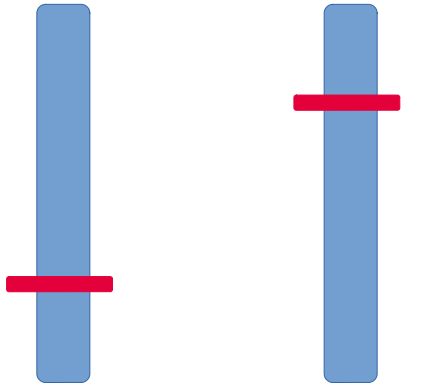
  → C'est l'approche choisie dans cette formation.

- Accès sécurisé (restriction d'accès aux données) : Protection contre la divulgation en restreignant l'accès aux données, en contrôlant l'environnement de travail et en vérifiant les sorties.

  → Ex : *Accès à des données individuelles sur les machines du CASD (Centre d'Accès Sécurisé aux données) accueillant uniquement les chercheurs autorisés.*

2 grandes approches pour gérer la confidentialité

-



Données sécurisées

Accès sécurisé

- Données sécurisées (restriction des données) : Protection contre la divulgation d'informations individuelles en modifiant les données elles-même.

  → Ex : *Méthode de suppression de cellules pour protéger les tableaux de données.*

  → C'est l'approche choisie dans cette formation.

- Accès sécurisé (restriction d'accès aux données) : Protection contre la divulgation en restreignant l'accès aux données, en contrôlant l'environnement de travail et en vérifiant les sorties.

  → Ex : *Accès à des données individuelles sur les machines du CASD (Centre d'Accès Sécurisé aux données) accueillant uniquement les chercheurs autorisés.*

1A Buts et enjeux

1B Cadre législatif

Loi statistique du 7 juin 1951 – articles 6 et 6 bis

-

Les articles 6 et 6 bis de la loi du 7 juin 1951 :

- définissent les conditions et les limites du secret statistique ;

  → Distinction des finalités de traitement

  → Interdiction de divulguer des renseignements individuels issus de données d'enquêtes ou de sources administratives (… mais il n'y a pas de notion de seuil à respecter pour y parvenir !)

  → Protection de la vie professionnelle et familiale des personnes physiques

  → Protection du secret commercial des entreprises (concurrence)

- instituent le comité du secret statistique ;

  → Il émet un avis sur les demandes d'accès aux données individuelles sous secret.

- définissent les sanctions en cas de violation du secret.

Code de bonnes pratiques de la statistique européenne

-

Principe 5 : Secret statistique et protection des données

Le respect de la vie privée des fournisseurs de données, la confidentialité des informations qu'ils fournissent, l'utilisation de celles-ci à des fins strictement statistiques et la sécurité des données sont absolument garantis.

Indicateurs :

5.1. Garanti par le droit
5.2. Signature d'un engagement de confidentialité
5.3. Sanctions en cas de violation
5.4. Instructions fournies aux agents et règles de confidentialité communiquées au public
5.5. Sécurité et intégrité des données
5.6. Accès restreint aux données individuelles

Lois sur la protection des données personnelles

-

- Sur les données relatives aux ménages et aux entreprises individuelles, deux lois supplémentaires s'appliquent :
  - la Loi Informatique et Libertés (LIL ; 1978) ;
  - le Règlement Général sur la Protection des Données (RGPD ; 2016).

- Elles permettent de :
  - garantir les droits des personnes sur les données qui les concernent ;
  - limiter les conséquences des traitements de leurs données.

Lois sur la protection des données personnelles

-

- La CNIL (Commission Nationale de l'Informatique et des Libertés) accompagne les professionnels dans la mise en conformité vis-à-vis du traitement de données à caractère personnel.

- La CNIL définit 3 critères pour s'assurer de l'anonymat d'un jeu de données :
  - La non-individualisation : il ne doit pas être possible d'isoler un individu dans le jeu de données ;
  - La non-corrélation : il ne doit pas être possible de relier entre eux des ensembles de données distincts concernant un même individu ;
  - La non-inférence : il ne doit pas être possible de déduire de façon quasi certaine de nouvelles informations sur un individu.

Jurisprudence à l'Insee et dans le SSP

-

- Pour les données d'enquêtes sur les entreprises, voici les seuils minimums :
  - Règle des 3 unités

    → décision du 13 juin 1980 du directeur général de l'Insee

  - Règle de dominance à 85 %

    → règle définie le 7 juillet 1960 par le Comité de coordination des enquêtes statistiques (prédécesseur du Cnis)

- NB : Il faut prendre en compte les poids de sondage.
  L'échantillon et la pondération associée doivent rester confidentiels.

Jurisprudence à l'Insee et dans le SSP

-

- Pour les données d'enquêtes sur les ménages (hors recensement) :
  - Il ne doit pas être possible d'identifier une personne.

    → application du code des relations entre le public et l'administration

  - Il ne doit pas être possible de déduire de nouvelles informations sur un individu de façon quasi certaine.

    → préconisation de la CNIL

Jurisprudence à l'Insee et dans le SSP

-

- Pour les données tirées de sources administratives :
  - Au cas par cas
  - À demander au producteur de vos données si besoin (inscrit dans la convention qui a permis la transmission des données)
  - Exemple des données fiscales : *aucune cellule ne doit concerner moins de 11 individus ou ménages*

Jurisprudence à l'Insee et dans le SSP

-

- Pour les sources mixtes, provenant de combinaisons d'enquêtes statistiques et de données administratives :
  - Cumul :

règle de l'enquête

+

règle de la source administrative

Jurisprudence à l'Insee et dans le SSP

-

- Cette jurisprudence est récapitulée dans le guide du secret statistique.

- Les questions de confidentialité et de la robustesse ne doivent pas être confondues : elles suivent des logiques bien distinctes.

  → Ex : *Applique-t-on un seuil de diffusion à n individus pour des raisons de confidentialité ou de significativité des résultats ?*

- Les seuils ne sont pas inscrits dans la loi.

  → Ce sont les organisme statistiques qui s'autorégulent afin de respecter les interdictions de divulgation de la loi statistique de 1951.

Jurisprudence à l'Insee et dans le SSP

-

- La loi n'indique pas non plus le degré de protection nécessaire selon le caractère « sensible » des variables. Au regard de la loi, elles sont toutes à protéger.

  → Pour autant, le producteur des données doit quand même réfléchir au niveau de confidentialité approprié à chaque variables.

  → Pour cela, il peut se poser les questions suivantes :

  - Existe-t-il une règle similaire déjà en vigueur ? (afin de garantir aux yeux des utilisateurs une cohérence dans la gestion du secret)
  - Quel·s risque·s est-ce que je fais porter au groupe d'individus concerné en publiant une telle information ?
  - Quelle·s opportunité·s cela peut-il créer ?

02 Quelles méthodes en vigueur ?

2A Fréquence et dominance

2B Secret primaire et secondaire

Tableaux d'effectifs

-

Définition : Tableau dans lequel la valeur d'une cellule correspond au nombre d'unités qui partagent les caractéristiques de la cellule.

Exemple : *Nombre d'entreprises polluantes par région*

		Région				
		Nord	Ouest	Est	Sud	**Total**
Polluante	Oui	6	14	1	7	**28**
	Non	3	2	1	13	**19**
	Total	**9**	**16**	**2**	**20**	**47**

Tableaux d'effectifs

Quels problèmes pouvez-vous identifier ?

		Région				
		Nord	Ouest	Est	Sud	**Total**
Polluante	Oui	6	14	1	7	**28**
	Non	3	2	1	13	**19**
	Total	**9**	**16**	**2**	**20**	**47**

Tableaux d'effectifs

Quels problèmes pouvez-vous identifier ?

		Région				
		Nord	Ouest	Est	Sud	**Total**
Polluante	Oui	6	14	1	7	**28**
	Non	3	2	1	13	**19**
	Total	**9**	**16**	**2**	**20**	**47**

Tableaux d'effectifs

Règle de fréquence

Une cellule d'un tableau ne doit pas être construite à partir de strictement moins de n unités (n > 0).

- Souvent à l'Insee, n = 3 (règle des 3 unités).

Tableaux d'effectifs

Règle de fréquence

Une cellule d'un tableau ne doit pas être construite à partir de strictement moins de n unités (n > 0).

- Souvent à l'Insee, n = 3 (règle des 3 unités).

		Région				
		Nord	Ouest	Est	Sud	**Total**
Polluante	Oui	6	14	1	7	**28**
	Non	3	2	1	13	**19**
	Total	**9**	**16**	**2**	**20**	**47**

Pourquoi protéger un tableau d'effectifs ?

-

Pourquoi protéger un tableau d'effectifs ?

-

- Contre le risque d'identification :

  → Je me reconnais ou je reconnais quelqu'un dans les données.

Pourquoi protéger un tableau d'effectifs ?

-

- Contre le risque d'identification :

  → Je me reconnais ou je reconnais quelqu'un dans les données.

- Contre le risque de divulgation d'attribut :

  → J'apprends que je (ou autrui) possède une caractéristique rare ou unique.

  → J'apprends que dans le champ des individus du tableau, personne ne possède telle caractéristique.

Pourquoi protéger un tableau d'effectifs ?

-

- Contre le risque d'identification :

  → Je me reconnais ou je reconnais quelqu'un dans les données.

- Contre le risque de divulgation d'attribut :

  → J'apprends que je (ou autrui) possède une caractéristique rare ou unique.

  → J'apprends que dans le champ des individus du tableau, personne ne possède telle caractéristique.

- Contre le risque de divulgation par recoupement :

  → Je peux déduire de nouvelles informations en croisant d'autres sources (sur d'autres variables, d'autres zonages, d'autres millésimes, …).

Pourquoi protéger un tableau d'effectifs ?

-

- Contre le risque d'identification :

  → Je me reconnais ou je reconnais quelqu'un dans les données.

- Contre le risque de divulgation d'attribut :

  → J'apprends que je (ou autrui) possède une caractéristique rare ou unique.

  → J'apprends que dans le champ des individus du tableau, personne ne possède telle caractéristique.

- Contre le risque de divulgation par recoupement :

  → Je peux déduire de nouvelles informations en croisant d'autres sources (sur d'autres variables, d'autres zonages, d'autres millésimes, …).

- Contre la perception de non-protection :

  → Les utilisateurs doivent sentir que l'on respecte la confidentialité des données diffusées.

Tableaux de volume

-

Définition : Tableau dans lequel la valeur de chaque cellule représente la somme des contributions des répondants qui partagent les caractéristiques de cette cellule.

Exemple : *Ventes réalisées, en millions d'euros*

		Produit vendu				
		Harpes	Piano	Orgues	Autre	**Total**
Région	Nord	58	71	92	800	**1021**
	Centre	11	124	157	934	**1226**
	Sud	36	24	60	651	**771**
	Total	**105**	**219**	**309**	**2385**	**3017**

Tableaux de volume

-

Définition : Tableau dans lequel la valeur de chaque cellule représente la somme des contributions des répondants qui partagent les caractéristiques de cette cellule.

Exemple : *Ventes réalisées, en millions d'euros (nombre de contributeurs)*

		Produit vendu				
		Harpe	Piano	Orgue	Autre	**Total**
Région	Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
	Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
	Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
	Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Tableaux de volume

Quels problèmes pouvez-vous identifier ?

		Produit vendu				
		Harpe	Piano	Orgue	Autre	**Total**
Région	Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
	Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
	Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
	Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Tableaux de volume

1er problème

-

		Produit vendu				
		Harpe	Piano	Orgue	Autre	**Total**
Région	Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
	Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
	Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
	Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

1er problème : 1 ou 2 contributeurs

-

Vendeurs d'orgues du centre et du sud

→ *Trop peu d'individus contributeurs ⇒ Divulgation d'informations*

		Produit vendu				
		Harpe	Piano	Orgue	Autre	**Total**
Région	Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
	Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
	Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
	Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

1er problème : 1 ou 2 contributeurs

-

Règle de fréquence
Une cellule d'un tableau ne doit pas être construite à partir de strictement moins de n unités (n > 0).

- Souvent à l'Insee, n = 3 (règle des 3 unités).
- Il faut considérer les poids : Si une cellule construite à partir de 2 répondants en représente davantage, alors elle est diffusable.
- L'échantillon et la pondération doivent rester confidentiels.

Tableaux de volume

Quels autres problèmes pouvez-vous identifier ?

		Produit vendu				
		Harpe	Piano	Orgue	Autre	**Total**
Région	Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
	Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
	Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
	Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

2nd problème

-

Quels autres problèmes pouvez-vous identifier ?

		Produit vendu				
		Harpe	Piano	Orgue	Autre	**Total**
Région	Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
	Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
	Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
	Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

2nd problème

-

Quels autres problèmes pouvez-vous identifier ?

86 + 3 + 1 + 1 + 1
= 92 millions d'€

		Produit vendu				
		Harpe	Piano	Orgue	Autre	**Total**
Région	Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
	Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
	Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
	Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

2nd problème : 1 contributeur dominant

-

L'information agrégée diffusée dans la cellule est proche d'une information individuelle.

→ Information individuelle presque disponible

Quand est-ce que le presque devient acceptable ?

86 + 3 + 1 + 1 + 1
= 92 millions d'€

		Produit vendu				
		Harpe	Piano	Orgue	Autre	**Total**
Région	Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
	Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
	Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
	Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

2nd problème : 1 contributeur dominant

-

Règle de dominance à k % (de paramètres (1, k))
1 unité contributrice à une cellule ne peut contribuer à plus de k % de la valeur de celle-ci.

- Par défaut à l'Insee, k = 85.

86 + 3 + 1 + 1 + 1
= 92 millions d'€

→ $\frac{86}{92} = 93\,\% > 85\,\%$

→ cellule sensible à la règle de dominance à 85 %

2nd problème : 1 contributeur dominant

-

Soit $V_C = \sum_{i=1}^{n} w_i \cdot x_i$ la valeur d'une cellule C et w_i un poids d'échantillonnage.

Soit $x_{(1)}$ le premier contributeur de la cellule.

Une cellule est jugée sensible à la règle de dominance *(1, k)* si :

$$x_{(1)} > k\,\% \cdot \sum_{i=1}^{n} w_i \cdot x_i$$

$$x_{(1)} > k\,\% \cdot V_C$$

$$\textit{contributeur maximal}\ (\textit{non pondéré}) > k\,\% \times \textit{total de la case}\ (\textit{pondéré})$$

Quels autres problèmes pouvez-vous identifier ?

		Produit vendu				
		Harpe	Piano	Orgue	Autre	**Total**
Région	Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
	Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
	Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
	Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

3ème problème
-

Quels autres problèmes pouvez-vous identifier ?

		Produit vendu				
		Harpe	Piano	Orgue	Autre	**Total**
Région	Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
	Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
	Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
	Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

3ème problème
-

Quels autres problèmes pouvez-vous identifier ?

19 + 16 + 1 =
36 millions d'€

		Produit vendu				
		Harpe	Piano	Orgue	Autre	**Total**
Région	Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
	Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
	Sud	36 (3)	~~24 (6)~~	~~60 (1)~~	~~651 (4)~~	~~**771 (14)**~~
	Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

3ème problème : plusieurs contributeurs dominants

-

Le second contributeur possède une estimation précise des ventes du leader du marché, même en absence de problème de dominance :

$$\frac{\widehat{x_{(1)}}}{x_{(1)}} = \frac{36-16}{19} \simeq 1{,}053 \rightarrow$$

L'information auxiliaire du 2nd contributeur, sa propre valeur, lui permet d'estimer son plus gros concurrent à p = 5,3 % près.
Sa valeur est presque connue par $x_{(2)}$.

19 + 16 + 1 = 36 millions d'€

		Produit vendu				
		Harpe	Piano	Orgue	Autre	**Total**
Région	Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
	Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
	Sud	36 (3)	~~24 (6)~~	~~60 (1)~~	~~651 (4)~~	~~**771 (14)**~~
	Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

3ème problème : plusieurs contributeurs dominants

-

Règle du p % (de paramètres (p, 1))

Le 2ème contributeur de la valeur d'une cellule ne doit pas pouvoir estimer, en utilisant sa propre valeur seulement, celle du premier contributeur avec une précision supérieur à p %.

- Cette règle ne fait pas jurisprudence à l'Insee, mais elle reste recommandée par les méthodologues européens.

  → À utiliser si nécessaire en plus de la règle de fréquence et de dominance pour augmenter la protection

- Il est d'usage de choisir p = 10.

3ème problème : plusieurs contributeurs dominants

-

Soit $V_C = \sum_{i=1}^{n} w_i \cdot x_i$ la valeur d'une cellule C et w_i les poids d'échantillonnage.

Soit $x_{(1)}$ puis $x_{(2)}$ les 2 premiers contributeurs de la cellule.

$x_{(2)}$ peut simplement estimer $x_{(1)}$ avec l'estimateur $\widehat{x_{(1)}} = V_C - x_{(2)}$

Une cellule est jugée sensible à la règle du *p %* si :

$$\widehat{x_{(1)}} - x_{(1)} < p\% \cdot x_{(1)}$$

$$\frac{\widehat{x_{(1)}} - x_{(1)}}{x_{(1)}} < p\%$$

erreur relative d'estimation de la contribution du 1er contributeur par le 2nd < p %

Les ratios

-

- Deux stratégies sont envisageables pour protéger un ratio :

Les ratios

-

- Deux stratégies sont envisageables pour protéger un ratio :
  - Si le dénominateur est connu, alors il suffit d'effectuer la protection sur le numérateur seulement.

    → Exemple : *Pourcentage ou moyenne sur population entière connue*

Les ratios

-

- Deux stratégies sont envisageables pour protéger un ratio :
  - Si le dénominateur est connu, alors il suffit d'effectuer la protection sur le numérateur seulement.

    → Exemple : *Pourcentage ou moyenne sur population entière connue*

  - Sinon, la cellule est cachée si et seulement si :

    le numérateur OU (inclusif) le dénominateur est caché.

    → Exemple :

$$\left.\begin{array}{l} N \rightarrow \text{sensible} \\ D \rightarrow \text{non-sensible} \end{array}\right\} \rightarrow \text{sensible}$$

Les autres statistiques : quantiles, indices de Gini, …

-

- Plus généralement, il faut protéger ou éviter de publier une statistique :
  - Portant sur un trop petit nombre d'individus ;

    → Ex : *Il doit y avoir suffisamment d'individus entre 2 centiles publiés.*

    Il est recommandé d'éviter de publier ou de bien arrondir un extremum.

  - Dont l'un des individus est dominant ;

    → Ex : *Publier un indice sur un secteur où une entreprise est dominante.*

  - Pour laquelle on peut en déduire de l'information individuelle.

    → Ex : *Publier une même statistique une fois sur n individus et une seconde fois sur n + 1 individus pose un problème de différenciation.*

- Ces recommandation sont également valables lors de la diffusion de graphiques.

  → Ex : *Un diagramme en barres peut poser des problèmes de dominance.*

2A Fréquence et dominance

2B Secret primaire et secondaire

Le secret primaire

-

Les cellules catégorisées sensibles pour :

- la règle de fréquence,
- la règle de dominance,
- voire la règle du p %

constituent le secret primaire.

→ On ne peut pas les diffuser.

Les stratégies envisageables pour gérer le secret primaire

-

Deux possibilités :

Les stratégies envisageables pour gérer le secret primaire

-

Deux possibilités :

- Ne pas diffuser l'information

  → En supprimant des cellules du tableau

Les stratégies envisageables pour gérer le secret primaire

-

Deux possibilités :

- Ne pas diffuser l'information

  → En supprimant des cellules du tableau

- S'arranger pour ne pas avoir de secret primaire

  → En restructurant les données (recodage des variables de ventilation)

Éviter le secret primaire – Recoder les variables de ventilation -

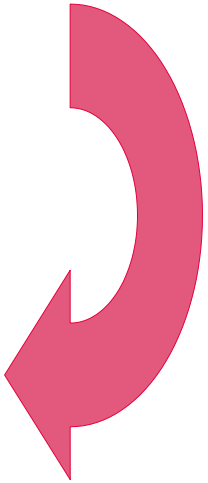
Ventes réalisées (en millions d'euros)

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Éviter le secret primaire – Recoder les variables de ventilation -

Ventes réalisées (en millions d'euros)

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**



Éviter le secret primaire – Recoder les variables de ventilation -

Ventes réalisées (en millions d'euros)

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

	Harpe	Piano	Autre	**Total**
Nord	58 (5)	71 (17)	892 (17)	**1021 (39)**
Centre	11 (4)	124 (11)	1091 (9)	**1226 (24)**
Sud	36 (3)	24 (6)	711 (5)	**771 (14)**
Total	**105 (12)**	**219 (34)**	**2694 (31)**	**3018 (77)**

Recodage simple :
Combiner des modalités pour augmenter le nombre de répondants dans les cellules.

Fusion de deux colonnes

Éviter le secret primaire – Recoder les variables de ventilation -

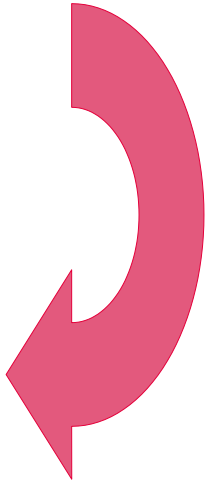
Nombre d'entreprises polluantes selon l'âge du dirigeant.

		Âge du dirigeant				
		< 25	25-30	30-50	> 50	**Total**
Polluante	Oui	2	5	7	6	**20**
	Non	8	15	17	20	**60**
	Total	**10**	**20**	**24**	**26**	**80**

Éviter le secret primaire – Recoder les variables de ventilation -

Nombre d'entreprises polluantes selon l'âge du dirigeant.

		Âge du dirigeant				
		< 25	25-30	30-50	> 50	**Total**
Polluante	Oui	2	5	7	6	**20**
	Non	8	15	17	20	**60**
	Total	**10**	**20**	**24**	**26**	**80**



Éviter le secret primaire – Recoder les variables de ventilation

-

Nombre d'entreprises polluantes selon l'âge du dirigeant.

		Âge du dirigeant				
		< 25	25-30	30-50	> 50	**Total**
Polluante	Oui	2	5	7	6	**20**
	Non	8	15	17	20	**60**
	Total	**10**	**20**	**24**	**26**	**80**

		Âge du dirigeant				
		< 28	28-35	35-55	> 55	**Total**
Polluante	Oui	3	6	6	5	**20**
	Non	9	17	19	15	**60**
	Total	**12**	**23**	**25**	**20**	**80**

Recodage plus complexe :
Définir de nouvelles modalités aux variables de ventilation pour augmenter le nombre de répondants par cellule.

Une exploration à faire à la main

Éviter le secret primaire… Ou pas !

-

- Le recodage est peu utilisé en pratique, car :
  - parfois, la structure des données diffusées est imposée (*Eurostat*) ;
  - cela casse le suivi des données dans le temps ;
  - la perte d'information est importante.

Éviter le secret primaire… Ou pas !

-

- Le recodage est peu utilisé en pratique, car :
  - parfois, la structure des données diffusées est imposée (*Eurostat*) ;
  - cela casse le suivi des données dans le temps ;
  - la perte d'information est importante.
- … Mais il peut être utile si les données sont de toute façon trop fines pour avoir un sens statistique.

Éviter le secret primaire… Ou pas !

-

- Le recodage est peu utilisé en pratique, car :
  - parfois, la structure des données diffusées est imposée (*Eurostat*) ;
  - cela casse le suivi des données dans le temps ;
  - la perte d'information est importante.
- … Mais il peut être utile si les données sont de toute façon trop fines pour avoir un sens statistique.

- La solution la plus souvent envisagée en pratique :

  → supprimer (cacher) les agrégats touchés par le secret primaire

  → Oui, mais …

Le secret secondaire

-

→ Oui, mais …

		Région				
		Nord	Ouest	Est	Sud	**Total**
Polluante	Oui	6	14	X	7	**28**
	Non	3	X	X	13	**19**
	Total	**9**	**16**	**X**	**20**	**47**

Le secret secondaire

-

→ Oui, mais …

		Région				
		Nord	Ouest	Est	Sud	**Total**
Polluante	Oui	6	14	X	7	**28**
	Non	3	X	X	13	**19**
	Total	**9**	**16**	**X**	**20**	**47**

- Si les marges du tableau sont diffusées, alors les cellules sont liées entre elles par des équations.

  → Il faut cacher d'autres cellules pour ne pas pouvoir déduire la valeur des cellules cachées : c'est le secret secondaire.

Le secret secondaire

- Pour le secret primaire, chaque case est traitée indépendamment.
  Pour le secret secondaire, il faut considérer le tableau dans son ensemble.

		Âge du dirigeant				
		< 25	25-30	30-50	> 50	Total
Polluante	Oui	2	5	7	6	**20**
	Non	8	15	17	20	**60**
	Total	**10**	**20**	**24**	**26**	**80**

		Âge du dirigeant				
		< 25	25-30	30-50	> 50	**Total**
Polluante	Oui	X	X	7	6	**20**
	Non	X	X	17	20	**60**
	Total	**10**	**20**	**24**	**26**	**80**

Suppressions primaire et secondaire

Nombre d'entreprises polluantes selon l'âge du dirigeant.

Le secret secondaire

- Plusieurs structures de suppressions, ou masques de secret, sont possibles.
  → Elles ne sont pas toutes équivalentes !

		Âge du dirigeant				
		< 25	25-30	30-50	> 50	Total
Polluante	Oui	2	5	7	6	**20**
	Non	8	15	17	20	**60**
	Total	**10**	**20**	**24**	**26**	**80**

		Âge du dirigeant				
		< 25	25-30	30-50	> 50	Total
Polluante	Oui	X	5	7	6	**X**
	Non	8	15	17	20	**60**
	Total	**X**	**20**	**24**	**26**	**X**

Nombre d'entreprises polluantes selon l'âge du dirigeant.

Autre possibilité de suppressions primaire et secondaire (certainement moins bonne)

Minimiser la perte d'information

-

Pour minimiser cette perte d'information, il faut définir pour chaque cellule un coût associé à sa suppression.

On peut ainsi choisir ce coût à minimiser comme étant :

Minimiser la perte d'information

-

Pour minimiser cette perte d'information, il faut définir pour chaque cellule un coût associé à sa suppression.

On peut ainsi choisir ce coût à minimiser comme étant :

- Le nombre de cellules supprimées
- Le nombre de contributeurs concernés par les cases supprimées
- La valeur de la cellule supprimée
- Une autre variable de coût : la valeur qu'aurait pris la cellule pour une autre variable de réponse
- …

Minimiser la perte d'information – Exemple

-

	A	B	C	D	E	F	Total
M1	0 (0)	82 (5)	42 (6)	98 (2)	315 (18)	322 (23)	**859 (54)**
M2	805 (45)	12 (2)	60 (9)	555 (54)	954 (77)	1122 (111)	**3508 (298)**
M3	0 (0)	66 (5)	44 (8)	28 (3)	28 (9)	488 (40)	**654 (65)**
M4	927 (45)	967 (79)	3065 (354)	4187 (422)	11 (2)	3122 (354)	**12279 (1256)**
M5	5220 (451)	3208 (354)	3545 (355)	344 (35)	55 (54)	100 (10)	**12472 (1259)**
M6	2200 (254)	692 (82)	339 (34)	18 (2)	652 (48)	79 (8)	**3980 (428)**
Total	**9152 (795)**	**5027 (527)**	**7095 (766)**	**5230 (518)**	**2015 (208)**	**5233 (546)**	**33752 (3360)**

Minimiser la perte d'information – Exemple

-

Application du secret primaire :

	A	B	C	D	E	F	Total
M1	0 (0)	82 (5)	42 (6)	98 (2)	315 (18)	322 (23)	**859 (54)**
M2	805 (45)	12 (2)	60 (9)	555 (54)	954 (77)	1122 (111)	**3508 (298)**
M3	0 (0)	66 (5)	44 (8)	28 (3)	28 (9)	488 (40)	**654 (65)**
M4	927 (45)	967 (79)	3065 (354)	4187 (422)	11 (2)	3122 (354)	**12279 (1256)**
M5	5220 (451)	3208 (354)	3545 (355)	344 (35)	55 (54)	100 (10)	**12472 (1259)**
M6	2200 (254)	692 (82)	339 (34)	18 (2)	652 (48)	79 (8)	**3980 (428)**
Total	**9152 (795)**	**5027 (527)**	**7095 (766)**	**5230 (518)**	**2015 (208)**	**5233 (546)**	**33752 (3360)**

Minimiser la perte d'information – Exemple

-

Minimisation du nombre de cellules cachées :

	A	B	C	D	E	F	Total
M1	0 (0)	82 (5)	42 (6)	98 (2)	315 (18)	322 (23)	**859 (54)**
M2	805 (45)	12 (2)	60 (9)	555 (54)	954 (77)	1122 (111)	**3508 (298)**
M3	0 (0)	66 (5)	44 (8)	28 (3)	28 (9)	488 (40)	**654 (65)**
M4	927 (45)	967 (79)	3065 (354)	4187 (422)	11 (2)	3122 (354)	**12279 (1256)**
M5	5220 (451)	3208 (354)	3545 (355)	344 (35)	55 (54)	100 (10)	**12472 (1259)**
M6	2200 (254)	692 (82)	339 (34)	18 (2)	652 (48)	79 (8)	**3980 (428)**
Total	**9152 (795)**	**5027 (527)**	**7095 (766)**	**5230 (518)**	**2015 (208)**	**5233 (546)**	**33752 (3360)**

Minimiser la perte d'information – Exemple

-

Minimisation du nombre de cellules cachées :

	A	B	C	D	E	F	Total
M1	0 (0)	82 (5)	42 (6)	98 (2)	315 (18)	322 (23)	**859 (54)**
M2	805 (45)	12 (2)	60 (9)	555 (54)	954 (77)	1122 (111)	**3508 (298)**
M3	0 (0)	66 (5)	44 (8)	28 (3)	28 (9)	488 (40)	**654 (65)**
M4	927 (45)	967 (79)	3065 (354)	4187 (422)	11 (2)	3122 (354)	**12279 (1256)**
M5	5220 (451)	3208 (354)	3545 (355)	344 (35)	55 (54)	100 (10)	**12472 (1259)**
M6	2200 (254)	692 (82)	339 (34)	18 (2)	652 (48)	79 (8)	**3980 (428)**
Total	**9152 (795)**	**5027 (527)**	**7095 (766)**	**5230 (518)**	**2015 (208)**	**5233 (546)**	**33752 (3360)**

Nombre de cellules cachées : 8

Nombre de contributeurs cachés : 195

Valeur totale cachée : 2461

Minimiser la perte d'information – Exemple

-

Minimisation du nombre de contributeurs cachés :

	A	B	C	D	E	F	Total
M1	0 (0)	82 (5)	42 (6)	98 (2)	315 (18)	322 (23)	**859 (54)**
M2	805 (45)	12 (2)	60 (9)	555 (54)	954 (77)	1122 (111)	**3508 (298)**
M3	0 (0)	66 (5)	44 (8)	28 (3)	28 (9)	488 (40)	**654 (65)**
M4	927 (45)	967 (79)	3065 (354)	4187 (422)	11 (2)	3122 (354)	**12279 (1256)**
M5	5220 (451)	3208 (354)	3545 (355)	344 (35)	55 (54)	100 (10)	**12472 (1259)**
M6	2200 (254)	692 (82)	339 (34)	18 (2)	652 (48)	79 (8)	**3980 (428)**
Total	**9152 (795)**	**5027 (527)**	**7095 (766)**	**5230 (518)**	**2015 (208)**	**5233 (546)**	**33752 (3360)**

Minimiser la perte d'information – Exemple

-

Minimisation du nombre de contributeurs cachés :

	A	B	C	D	E	F	Total
M1	0 (0)	82 (5)	42 (6)	98 (2)	315 (18)	322 (23)	**859 (54)**
M2	805 (45)	12 (2)	60 (9)	555 (54)	954 (77)	1122 (111)	**3508 (298)**
M3	0 (0)	66 (5)	44 (8)	28 (3)	28 (9)	488 (40)	**654 (65)**
M4	927 (45)	967 (79)	3065 (354)	4187 (422)	11 (2)	3122 (354)	**12279 (1256)**
M5	5220 (451)	3208 (354)	3545 (355)	344 (35)	55 (54)	100 (10)	**12472 (1259)**
M6	2200 (254)	692 (82)	339 (34)	18 (2)	652 (48)	79 (8)	**3980 (428)**
Total	**9152 (795)**	**5027 (527)**	**7095 (766)**	**5230 (518)**	**2015 (208)**	**5233 (546)**	**33752 (3360)**

Nombre de cellules cachées : 10

Nombre de contributeurs cachés : 135

Valeur totale cachée : 2366

Minimiser la perte d'information – Exemple

-

Minimisation du nombre de contributeurs cachés :

Les cellules sans contributeur ne sont pas prises en compte pour élaborer le secret secondaire.

	A	B	C	D	E	F	Total
M1	0 (0)	82 (5)	42 (6)	98 (2)	315 (18)	322 (23)	**859 (54)**
M2	805 (45)	12 (2)	60 (9)	555 (54)	954 (77)	1122 (111)	**3508 (298)**
M3	0 (0)	66 (5)	44 (8)	28 (3)	28 (9)	488 (40)	**654 (65)**
M4	927 (45)	967 (79)	3065 (354)	4187 (422)	11 (2)	3122 (354)	**12279 (1256)**
M5	5220 (451)	3208 (354)	3545 (355)	344 (35)	55 (54)	100 (10)	**12472 (1259)**
M6	2200 (254)	692 (82)	339 (34)	18 (2)	652 (48)	79 (8)	**3980 (428)**
Total	**9152 (795)**	**5027 (527)**	**7095 (766)**	**5230 (518)**	**2015 (208)**	**5233 (546)**	**33752 (3360)**

Nombre de cellules cachées : 10

Nombre de contributeurs cachés : 135

Valeur totale cachée : 2366

Minimiser la perte d'information – Exemple

-

Minimisation de la valeur cachée :

	A	B	C	D	E	F	Total
M1	0 (0)	82 (5)	42 (6)	98 (2)	315 (18)	322 (23)	**859 (54)**
M2	805 (45)	12 (2)	60 (9)	555 (54)	954 (77)	1122 (111)	**3508 (298)**
M3	0 (0)	66 (5)	44 (8)	28 (3)	28 (9)	488 (40)	**654 (65)**
M4	927 (45)	967 (79)	3065 (354)	4187 (422)	11 (2)	3122 (354)	**12279 (1256)**
M5	5220 (451)	3208 (354)	3545 (355)	344 (35)	55 (54)	100 (10)	**12472 (1259)**
M6	2200 (254)	692 (82)	339 (34)	18 (2)	652 (48)	79 (8)	**3980 (428)**
Total	**9152 (795)**	**5027 (527)**	**7095 (766)**	**5230 (518)**	**2015 (208)**	**5233 (546)**	**33752 (3360)**

Minimiser la perte d'information – Exemple

-

Minimisation de la valeur cachée :

	A	B	C	D	E	F	Total
M1	0 (0)	82 (5)	42 (6)	98 (2)	315 (18)	322 (23)	**859 (54)**
M2	805 (45)	12 (2)	60 (9)	555 (54)	954 (77)	1122 (111)	**3508 (298)**
M3	0 (0)	66 (5)	44 (8)	28 (3)	28 (9)	488 (40)	**654 (65)**
M4	927 (45)	967 (79)	3065 (354)	4187 (422)	11 (2)	3122 (354)	**12279 (1256)**
M5	5220 (451)	3208 (354)	3545 (355)	344 (35)	55 (54)	100 (10)	**12472 (1259)**
M6	2200 (254)	692 (82)	339 (34)	18 (2)	652 (48)	79 (8)	**3980 (428)**
Total	**9152 (795)**	**5027 (527)**	**7095 (766)**	**5230 (518)**	**2015 (208)**	**5233 (546)**	**33752 (3360)**

Nombre de cellules cachées : 10

Nombre de contributeurs cachés : 174

Valeur totale cachée : 1442

Les singletons

-

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
Sud	36 (3)	24 (6)	59 (1)	651 (4)	**770 (14)**
Total	**105 (12)**	**219 (34)**	**308 (8)**	**2385 (23)**	**3017 (77)**

Les singletons

-

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
Sud	36 (3)	24 (6)	59 (1)	651 (4)	**770 (14)**
Total	**105 (12)**	**219 (34)**	**308 (8)**	**2385 (23)**	**3017 (77)**

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	X	124 (11)	X	934 (7)	**1226 (24)**
Sud	X	24 (6)	X	651 (4)	**770 (14)**
Total	**105 (12)**	**219 (34)**	**308 (8)**	**2385 (23)**	**3017 (77)**

Cacher uniquement ces cases ne suffit pas !

Les singletons

-

Les singletons, c'est-à-dire les cellules avec 1 seul répondant, peuvent retrouver une case en secret primaire grâce à leur propre information.

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
Sud	36 (3)	24 (6)	59 (1)	651 (4)	**770 (14)**
Total	**105 (12)**	**219 (34)**	**308 (8)**	**2385 (23)**	**3017 (77)**

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	X	124 (11)	X	934 (7)	**1226 (24)**
Sud	X	24 (6)	X	651 (4)	**770 (14)**
Total	**105 (12)**	**219 (34)**	**308 (8)**	**2385 (23)**	**3017 (77)**

Cacher uniquement ces cases ne suffit pas !

Les singletons

-

Les singletons, c'est-à-dire les cellules avec 1 seul répondant, peuvent retrouver une case en secret primaire grâce à leur propre information.

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
Sud	36 (3)	24 (6)	59 (1)	651 (4)	**770 (14)**
Total	**105 (12)**	**219 (34)**	**308 (8)**	**2385 (23)**	**3017 (77)**

	Harpe	Piano	Orgue	Autre	**Total**
Nord	X	71 (17)	X	800 (12)	**1021 (39)**
Centre	X	124 (11)	X	934 (7)	**1226 (24)**
Sud	X	24 (6)	X	651 (4)	**770 (14)**
Total	**105 (12)**	**219 (34)**	**308 (8)**	**2385 (23)**	**3017 (77)**

Il faut prendre les singletons en compte

Les singletons

-

Règle des singletons

Un contributeur seul dans sa cellule connaît sa propre valeur.
Il peut donc potentiellement déduire la valeur d'autres cellules sensibles.
Si tel est le cas, il faut cacher au moins une cellule supplémentaire.

→ Cette règle est valable pour les tableaux d'effectifs comme pour les tableaux de volume.

Au moins 2 cellules non-nulles

-

Règle des cellules non-nulles

Parmi les cellules diffusées (hors totaux), il ne faut pas qu'exactement une cellule d'une même ligne ou colonne soit non-nulle.

Au moins 2 cellules non-nulles

-

Règle des cellules non-nulles

Parmi les cellules diffusées (hors totaux), il ne faut pas qu'exactement une cellule d'une même ligne ou colonne soit non-nulle.

Champ : Habitants de la commune X

	Âge			
	18-25	26-49	50-59	> 60
Marié	7	12	0	30
Divorcé	0	11	9	10
Autre	21	27	0	14

Au moins 2 cellules non-nulles

-

Règle des cellules non-nulles

Parmi les cellules diffusées (hors totaux), il ne faut pas qu'exactement une cellule d'une même ligne ou colonne soit non-nulle.

- Cette règle n'est pas gérée par τ-Argus. Elle est donc souvent omise...

Champ : Habitants de la commune X

	Âge			
	18-25	26-49	50-59	> 60
Marié	7	12	0	30
Divorcé	0	11	9	10
Autre	21	27	0	14

Au moins 2 cellules non-nulles

-

Règle des cellules non-nulles

Parmi les cellules diffusées (hors totaux), il ne faut pas qu'exactement une cellule d'une même ligne ou colonne soit non-nulle.

- Cette règle n'est pas gérée par τ-Argus. Elle est donc souvent omise...

Champ : Habitants de la commune X

	Âge			
	18-25	26-49	50-59	> 60
Marié	7	12	0	30
Divorcé	0	11	9	10
Autre	21	27	0	14

Ne pas prendre en compte cette règle engendre une divulgation d'attribut, ici :

Au moins 2 cellules non-nulles

-

Règle des cellules non-nulles

Parmi les cellules diffusées (hors totaux), il ne faut pas qu'exactement une cellule d'une même ligne ou colonne soit non-nulle.

- Cette règle n'est pas gérée par τ-Argus. Elle est donc souvent omise...

Champ : Habitants de la commune X

	Âge			
	18-25	26-49	50-59	> 60
Marié	7	12	0	30
Divorcé	0	11	9	10
Autre	21	27	0	14

Ne pas prendre en compte cette règle engendre une divulgation d'attribut, ici :

« Un cinquantenaire de la commune X est forcément divorcé. »

Règle des intervalles – Intervalle des possibles

-

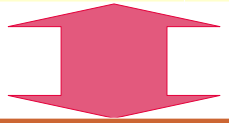
		Âge du dirigeant				
		< 25	25-30	30-50	> 50	**Total**
Polluante	Oui	X	X	7	6	**20**
	Non	X	X	17	20	**60**
	Total	**10**	**20**	**24**	**26**	**80**

		Âge du dirigeant				
		< 25	25-30	30-50	> 50	**Total**
Polluante	Oui	[0 ; 7]	[0 ; 7]	7	6	**20**
	Non	[3 ; 10]	[13 ; 20]	17	20	**60**
	Total	**10**	**20**	**24**	**26**	**80**

Règle des intervalles – Intervalle des possibles

-

		Âge du dirigeant				
		< 25	25-30	30-50	> 50	**Total**
Polluante	Oui	X	X	7	6	**20**
	Non	X	X	17	20	**60**
	Total	**10**	**20**	**24**	**26**	**80**



équivalence

		Âge du dirigeant				
		< 25	25-30	30-50	> 50	**Total**
Polluante	Oui	[0 ; 7]	[0 ; 7]	7	6	**20**
	Non	[3 ; 10]	[13 ; 20]	17	20	**60**
	Total	**10**	**20**	**24**	**26**	**80**

Cacher des cases revient à diffuser des intervalles

Règle des intervalles – Intervalle des possibles

-

Définition : L'intervalle des possibles (ou intervalle d'audit) est l'ensemble des valeurs possibles prises par la cellule supprimée après avoir posé le masque de secret.

		Âge du dirigeant				
		< 25	25-30	30-50	> 50	**Total**
Polluante	Oui	[0 ; 7]	[0 ; 7]	7	6	**20**
	Non	[3 ; 10]	[13 ; 20]	17	20	**60**
	Total	**10**	**20**	**24**	**26**	**80**

→ Ici, l'intervalle des possibles de la case sensible est [0 ; 7].

Règle des intervalles – Intervalle de protection (fréquence)

-

Définition : Soit V_C la valeur d'une cellule C sensible à la règle de fréquence. On choisit m %, la marge de protection. L'intervalle de protection de C est égal à :

$$\left[(1-m\%).V_C\ ;\ (1+m\%).V_C\right]$$

→ Avec une marge de 10 %, cela donne l'intervalle : $\left[90\%.V_C\ ;\ 110\%.V_C\right]$

Règle des intervalles – Intervalle de protection (fréquence)

-

Définition : Soit V_C la valeur d'une cellule C sensible à la règle de fréquence. On choisit m %, la marge de protection. L'intervalle de protection de C est égal à :

$$[(1-m\%).V_C\ ;\ (1+m\%).V_C]$$

→ Avec une marge de 10 %, cela donne l'intervalle : $[90\%.V_C\ ;\ 110\%.V_C]$

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Règle des intervalles – Intervalle de protection (fréquence)

-

Définition : Soit V_C la valeur d'une cellule C sensible à la règle de fréquence. On choisit m %, la marge de protection. L'intervalle de protection de C est égal à :

$$[(1-m\%).V_C\ ;\ (1+m\%).V_C]$$

→ Avec une marge de 10 %, cela donne l'intervalle : $[90\%.V_C\ ;\ 110\%.V_C]$

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Intervalle de protection à 10 % :

Règle des intervalles – Intervalle de protection (fréquence)

-

Définition : Soit V_C la valeur d'une cellule C sensible à la règle de fréquence. On choisit m %, la marge de protection. L'intervalle de protection de C est égal à :

$$[(1-m\%).V_C\ ;\ (1+m\%).V_C]$$

→ Avec une marge de 10 %, cela donne l'intervalle : $[90\%.V_C\ ;\ 110\%.V_C]$

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Intervalle de protection à 10 % :

[54 ; 66]

Règle des intervalles

-

Règle des intervalles

À la pose du masque de secret, l'intervalle de protection de chaque cellule sensible doit être inclus dans son intervalle des possibles.

Règle des intervalles

-

Règle des intervalles

À la pose du masque de secret, l'intervalle de protection de chaque cellule sensible doit être inclus dans son intervalle des possibles.

- Cette règle permet de se prémunir d'une divulgation par inférence.

→ Pour les cellules sensibles à la règle de fréquence, on doit paramétrer une marge de protection qu'on choisit souvent égale à 10 %.

Règle des intervalles – Exemple

-

Avec une marge de protection de 10 % :

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	11 (4)	124 (11)	157 (2)	934 (7)	**1226 (24)**
Sud	36 (3)	24 (6)	60 (1)	651 (4)	**771 (14)**
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Règle des intervalles – Exemple

-

Avec une marge de protection de 10 % :

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	11 (4)	124 (11)	157 (2)	~~934 (7)~~	~~**1226 (24)**~~
Sud	36 (3)	24 (6)	60 (1)	~~651 (4)~~	~~**771 (14)**~~
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Intervalles de protection à 10 % :

Règle des intervalles – Exemple

-

Avec une marge de protection de 10 % :

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	11 (4)	124 (11)	157 (2)	~~934 (7)~~	~~**1226 (24)**~~
Sud	36 (3)	24 (6)	60 (1)	~~651 (4)~~	~~**771 (14)**~~
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Intervalles de protection à 10 % :

[141 ; 173]

Règle des intervalles – Exemple

-

Avec une marge de protection de 10 % :

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	11 (4)	124 (11)	157 (2)	~~934 (7)~~	~~**1226 (24)**~~
Sud	36 (3)	24 (6)	60 (1)	~~651 (4)~~	~~**771 (14)**~~
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Intervalles de protection à 10 % :

[141 ; 173]

[54 ; 66]

Règle des intervalles – Exemple

-

Avec une marge de protection de 10 % :

	Harpe	Piano	Orgue	Autre	**Total**
Nord	X	71 (17)	X	800 (12)	**1021 (39)**
Centre	X	124 (11)	X	~~934 (7)~~	~~**1226 (24)**~~
Sud	X	24 (6)	X	~~651 (4)~~	~~**771 (14)**~~
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Intervalles de protection à 10 % :

[141 ; 173]

[54 ; 66]

Règle des intervalles – Exemple

-

Avec une marge de protection de 10 % :

	Harpe	Piano	Orgue	Autre	**Total**
Nord	X	71 (17)	X	800 (12)	**1021 (39)**
Centre	X	124 (11)	X	934 (7)	**1226 (24)**
Sud	X	24 (6)	X	651 (4)	**771 (14)**
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Valeur totale cachée : 414

Intervalles de protection à 10 % :

[141 ; 173]

[54 ; 66]

Règle des intervalles – Exemple

-

Avec une marge de protection de 10 % :

	Harpe	Piano	Orgue	Autre	**Total**
Nord	[0 ; 105]	71 (17)	[45 ; 150]	800 (12)	**1021 (39)**
Centre	[0 ; 105]	124 (11)	[63 ; 168]	~~934 (7)~~	~~**1226 (24)**~~
Sud	[0 ; 96]	24 (6)	[0 ; 96]	~~651 (4)~~	~~**771 (14)**~~
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Valeur totale cachée : 414

Intervalles de protection à 10 % :

[141 ; 173]

[54 ; 66]

Règle des intervalles – Exemple

-

Avec une marge de protection de 10 % :

	Harpe	Piano	Orgue	Autre	**Total**
Nord	[0 ; 105]	71 (17)	[45 ; 150]	800 (12)	**1021 (39)**
Centre	[0 ; 105]	124 (11)	[63 ; 168]	~~934 (7)~~	~~**1226 (24)**~~
Sud	[0 ; 96]	24 (6)	[0 ; 96]	~~651 (4)~~	~~**771 (14)**~~
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Valeur totale cachée : 414

Intervalles de protection à 10 % :

[141 ; 173]

[54 ; 66]

Règle des intervalles – Exemple

-

Avec une marge de protection de 10 % :

Valeur totale cachée : 414

	Harpe	Piano	Orgue	Autre	**Total**
Nord	[0 ; 105]	71 (17)	[45 ; 150]	800 (12)	**1021 (39)**
Centre	[0 ; 105]	124 (11)	[63 ; 168]	~~934 (7)~~	~~**1226 (24)**~~
Sud	[0 ; 96]	24 (6)	[0 ; 96]	~~651 (4)~~	~~**771 (14)**~~
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Intervalles de protection à 10 % :

[141 ; 173]

[54 ; 66]

→ L'intervalle de protection de cette cellule n'est pas inclus dans l'intervalle des possibles !

Ce masque de secret ne protège donc pas assez les 2 cellules sensibles.

Règle des intervalles – Exemple

-

Avec une marge de protection de 10 % :

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	71 (17)	92 (5)	800 (12)	**1021 (39)**
Centre	11 (4)	124 (11)	X	~~934 (7)~~	~~**1226 (24)**~~
Sud	36 (3)	24 (6)	X	~~651 (4)~~	~~**771 (14)**~~
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Intervalles de protection à 10 % :

[141 ; 173]

[54 ; 66]

Règle des intervalles – Exemple

-

Avec une marge de protection de 10 % :

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	X	X	800 (12)	**1021 (39)**
Centre	11 (4)	X	X	934 (7)	**1226 (24)**
Sud	36 (3)	X	X	651 (4)	**771 (14)**
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Valeur totale cachée : 528

Intervalles de protection à 10 % :

[141 ; 173]

[54 ; 66]

Règle des intervalles – Exemple

-

Avec une marge de protection de 10 % :

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	[0 ; 163]	[0 ; 163]	800 (12)	**1021 (39)**
Centre	11 (4)	[0 ; 219]	[62 ; 281]	~~934 (7)~~	~~**1226 (24)**~~
Sud	36 (3)	[0 ; 84]	[0 ; 84]	~~651 (4)~~	~~**771 (14)**~~
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Valeur totale cachée : 528

Intervalles de protection à 10 % :

[141 ; 173]

[54 ; 66]

Règle des intervalles – Exemple

-

Avec une marge de protection de 10 % :

Valeur totale cachée : 528

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	[0 ; 163]	[0 ; 163]	800 (12)	**1021 (39)**
Centre	11 (4)	[0 ; 219]	[62 ; 281]	934 (7)	**1226 (24)**
Sud	36 (3)	[0 ; 84]	[0 ; 84]	651 (4)	**771 (14)**
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Intervalles de protection à 10 % :

[141 ; 173]

[54 ; 66]

→ Tous les intervalles de protection sont inclus dans les intervalles des possibles. Ce masque est donc convenable.

Règle des intervalles – Exemple

-

Avec une marge de protection de 10 % :

Valeur totale cachée : 528

	Harpe	Piano	Orgue	Autre	**Total**
Nord	58 (5)	[0 ; 163]	[0 ; 163]	800 (12)	**1021 (39)**
Centre	11 (4)	[0 ; 219]	[62 ; 281]	~~934 (7)~~	~~**1226 (24)**~~
Sud	36 (3)	[0 ; 84]	[0 ; 84]	~~651 (4)~~	~~**771 (14)**~~
Total	**105 (12)**	**219 (34)**	**309 (8)**	**2385 (23)**	**3018 (77)**

Intervalles de protection à 10 % :

[141 ; 173]

[54 ; 66]

→ Tous les intervalles de protection sont inclus dans les intervalles des possibles. Ce masque est donc convenable.

- Procédure d'audit : Dans τ-Argus, les intervalles des possibles sont aussi appelés « intervalles d'audit ». Après application du masque, τ-Argus peut calculer à la demande ces intervalles pour chacune des cases cachées.

Règle des intervalles – Intervalle de protection (dominance et p%)

-

Règle des intervalles – Intervalle de protection (dominance et p%)

-

- Pour la règle de dominance et celle du p %, il est possible de faire mieux : utiliser directement les contraintes liées à la règle de secret primaire utilisée.

Règle des intervalles – Intervalle de protection (dominance et p%)

-

- Pour la règle de dominance et celle du p %, il est possible de faire mieux : utiliser directement les contraintes liées à la règle de secret primaire utilisée.
- L'intervalle de protection d'une cellule sensible C est ainsi égal à :

	Règle de dominance	Règle du p %
	$\left[2 \cdot V_C - \frac{x_{(1)}}{k\,\%} \; ; \; \frac{x_{(1)}}{k\,\%}\right]$	$\left[2 \cdot V_C - \left((1+p\%).x_{(1)} + x_{(2)}\right); \left((1+p\%).x_{(1)} + x_{(2)}\right)\right]$

Règle des intervalles – Intervalle de protection (dominance et p%)

-

- Pour la règle de dominance et celle du p %, il est possible de faire mieux : utiliser directement les contraintes liées à la règle de secret primaire utilisée.
- L'intervalle de protection d'une cellule sensible C est ainsi égal à :

Règle de fréquence	Règle de dominance	Règle du p %
$[(1 \pm m\%).V_C]$	$[2 \cdot V_C - \frac{x_{(1)}}{k\%} ; \frac{x_{(1)}}{k\%}]$	$[2 \cdot V_C - ((1+p\%).x_{(1)} + x_{(2)}) ; ((1+p\%).x_{(1)} + x_{(2)})]$

- Seule la règle de fréquence nécessite de choisir un seuil *m*. Pour les autres règles, le calcul s'effectue automatiquement par le logiciel.

Règle des intervalles – Intervalle de protection (dominance et p%)

-

- Pour la règle de dominance et celle du p %, il est possible de faire mieux : utiliser directement les contraintes liées à la règle de secret primaire utilisée.
- L'intervalle de protection d'une cellule sensible C est ainsi égal à :

Règle de fréquence	Règle de dominance	Règle du p %
$[(1 \pm m\%).V_C]$	$[2 \cdot V_C - \frac{x_{(1)}}{k\%} ; \frac{x_{(1)}}{k\%}]$	$[2 \cdot V_C - ((1+p\%).x_{(1)} + x_{(2)}) ; ((1+p\%).x_{(1)} + x_{(2)})]$

- Seule la règle de fréquence nécessite de choisir un seuil *m*. Pour les autres règles, le calcul s'effectue automatiquement par le logiciel.
- La valeur de l'intervalle de protection est évidemment confidentielle.

Règle des intervalles – Intervalle de protection (dominance et p%)

-

Explication des intervalles de protection pour la dominance et le p % :

Règle des intervalles – Intervalle de protection (dominance et p%)

-

Explication des intervalles de protection pour la dominance et le p % :

- Une cellule sensible selon la règle de dominance s'écrit : $x_{(1)} > k\,\% \cdot V_C$

ou bien $V_C < \dfrac{x_{(1)}}{k\,\%}$

V_C

|

Règle des intervalles – Intervalle de protection (dominance et p%)

-

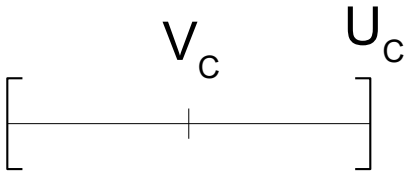
Explication des intervalles de protection pour la dominance et le p % :

- Une cellule sensible selon la règle de dominance s'écrit : $x_{(1)} > k\,\% \cdot V_C$

  ou bien $V_C < \dfrac{x_{(1)}}{k\,\%}$

- Pour que la valeur déductible ne soit jamais considérée comme sensible, on choisit la borne supérieure de l'intervalle de protection U_C telle que :

$$U_C = \frac{x_{(1)}}{k\,\%}$$



Règle des intervalles – Intervalle de protection (dominance et p%)

-

Explication des intervalles de protection pour la dominance et le p % :

- Une cellule sensible selon la règle de dominance s'écrit : $x_{(1)} > k\,\% \cdot V_C$

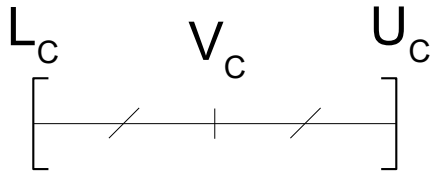
  ou bien $V_C < \dfrac{x_{(1)}}{k\,\%}$

- Pour que la valeur déductible ne soit jamais considérée comme sensible, on choisit la borne supérieure de l'intervalle de protection U_C telle que :

$$U_C = \frac{x_{(1)}}{k\,\%}$$

- La borne inférieure de l'intervalle L_C s'obtient par symétrie : $L_C = V_C - (U_C - V_C)$

$$= 2 \cdot V_C - \frac{x_{(1)}}{k\,\%}$$



Règle des intervalles – Intervalle de protection (dominance et p%)

-

Explication (suite) :

- C'est pourquoi :

  Définition : L'intervalle de protection de C, sensible à la règle de dominance, est :

$$\left[2 \cdot V_C - \frac{x_{(1)}}{k\%} \; ; \; \frac{x_{(1)}}{k\%}\right]$$

$$\left[2 \cdot V_C - \left((1+p\%).x_{(1)} + x_{(2)}\right); \left((1+p\%).x_{(1)} + x_{(2)}\right)\right]$$

Règle des intervalles – Intervalle de protection (dominance et p%)

-

Explication (suite) :

- C'est pourquoi :

  Définition : L'intervalle de protection de C, sensible à la règle de dominance, est :

$$\left[2 \cdot V_C - \frac{x_{(1)}}{k\,\%} \; ; \; \frac{x_{(1)}}{k\,\%}\right]$$

- Une cellule sensible selon la règle du p % s'écrit :

  Par un raisonnement analogue, on obtient :

  Définition : L'intervalle de protection de C, sensible à la règle du p %, est :

$$\left[2 \cdot V_C - \left((1+p\,\%).x_{(1)} + x_{(2)}\right) ; \left((1+p\,\%).x_{(1)} + x_{(2)}\right)\right]$$

Règle des intervalles

-

Règle des intervalles

À la pose du masque de secret, l'intervalle de protection de chaque cellule sensible doit être inclus dans son intervalle des possibles.

- Cette règle permet de se prémunir d'une divulgation par inférence.

→ Pour les cellules sensibles à la règle de fréquence, on doit paramétrer une marge de protection qu'on choisit souvent égale à 10 %.

Règle des intervalles

-

Règle des intervalles

À la pose du masque de secret, l'intervalle de protection de chaque cellule sensible doit être inclus dans son intervalle des possibles.

- Cette règle permet de se prémunir d'une divulgation par inférence.

→ Pour les cellules sensibles à la règle de fréquence, on doit paramétrer une marge de protection qu'on choisit souvent égale à 10 %.

→ Pour les autres cellules sensibles, tout est automatique.

03 Tableaux liés et variables hiérarchisées

Tableaux liés

-

Nombre d'entreprises :

- *par sexe du dirigeant ;*
- *par région ;*
- *si l'entreprise est polluante.*

Polluante ? Sexe \	Oui	Non	**Total**
Homme	22	10	**32**
Femme	8	27	**35**
Total	**30**	**37**	**67**

Polluante ? Région \	Oui	Non	**Total**
Nord	10	26	**36**
Sud	20	11	**31**
Total	**30**	**37**	**67**

Sexe Région \	Femme	Homme	**Total**
Nord	16	20	**36**
Sud	19	12	**31**
Total	**35**	**32**	**67**

Tableaux liés

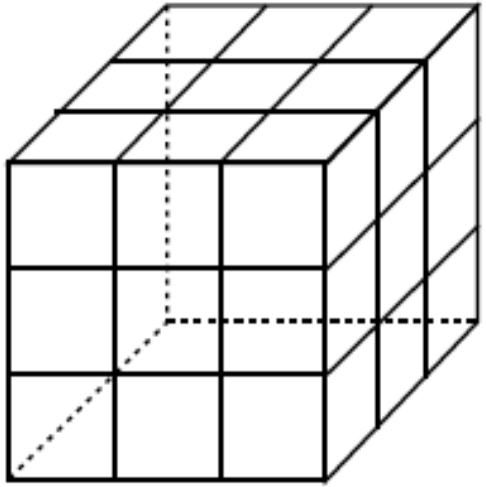
-

Nombre d'entreprises :

- *par sexe du dirigeant ;*
- *par région ;*
- *si l'entreprise est polluante.*

Polluante ? Sexe \	Oui	Non	**Total**
Homme	22	10	**32**
Femme	8	27	**35**
Total	**30**	**37**	**67**

Polluante ? Région \	Oui	Non	**Total**
Nord	10	26	**36**
Sud	20	11	**31**
Total	**30**	**37**	**67**



Sexe Région \	Femme	Homme	**Total**
Nord	16	20	**36**
Sud	19	12	**31**
Total	**35**	**32**	**67**

Tableaux liés

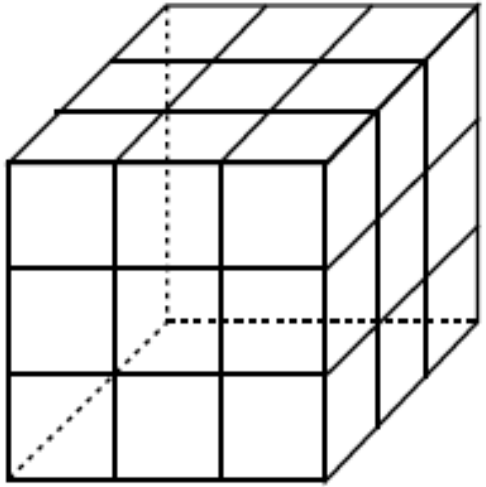
-

Nombre d'entreprises :

- *par sexe du dirigeant ;*
- *par région ;*
- *si l'entreprise est polluante.*

Polluante ? Région \	Oui	Non	**Total**
Nord	10	26	**36**
Sud	20	11	**31**
Total	**30**	**37**	**67**

Polluante ? Sexe \	Oui	Non	**Total**
Homme	22	10	**32**
Femme	8	27	**35**
Total	**30**	**37**	**67**



Sexe Région \	Femme	Homme	**Total**
Nord	16	20	**36**
Sud	19	12	**31**
Total	**35**	**32**	**67**

→ La pose du secret s'effectue dans une logique à n dimensions...

Tableaux liés

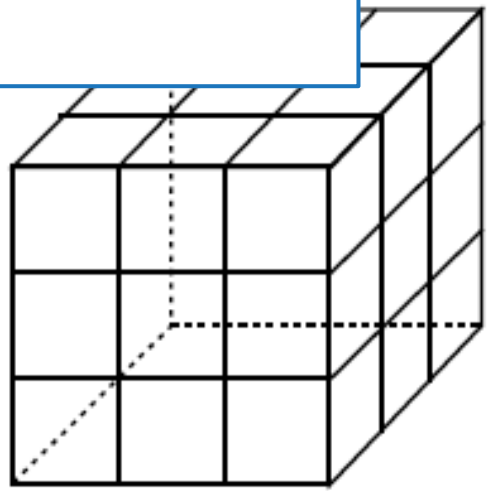
-

Polluante ? Sexe \	Oui	Non	**Total**
Homme	22	10	**32**
		27	**35**
		37	**67**

hommes dirigeants d'entreprises polluantes du Sud
+
femmes dirigeantes d'entreprises polluantes du Sud
=
20

→ La pose du secret s'effectue dans une logique à n dimensions...

Polluante ? Région \	Oui	Non	**Total**
Nord	10	26	**36**
Sud	20	11	**31**
Total	**30**	**37**	**67**



Sexe Région \	Femme	Homme	**Total**
Nord	16	20	**36**
Sud	19	12	**31**
Total	**35**	**32**	**67**

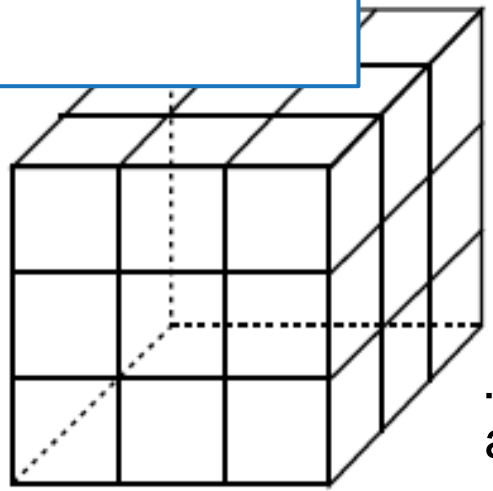
Tableaux liés

-

Polluante ? Sexe \	Oui	Non	**Total**
Homme	22	10	**32**
		27	**35**
		37	**67**

hommes dirigeants d'entreprises polluantes du Sud
+
femmes dirigeantes d'entreprises polluantes du Sud
=
20

Polluante ? Région \	Oui	Non	**Total**
Nord	10	26	**36**
Sud	20	11	**31**
Total	**30**	**37**	**67**



Sexe Région \	Femme	Homme	**Total**
Nord	16	20	**36**
Sud	19	12	**31**
Total	**35**	**32**	**67**

→ La pose du secret s'effectue dans une logique à n dimensions...

… Qui peut conduire à plus ou à moins de secret secondaire.

Tableaux liés

-

Définition : Ensemble de tableaux possédant la même variable ventilée et partageant une ou plusieurs variables de ventilation.

- Cet ensemble de tableaux peut être considéré comme des sous-tableaux d'un tableau de dimension plus grande.

- Ces liens entre tableaux doivent être pris en compte pour traiter correctement le secret. Sinon, les cellules cachées peuvent se déduire les unes des autres.

- Pour des données périodiques, la temporalité peut être perçue comme un ensemble de tableaux liés : il vaut mieux que le masque de secret change peu entre chaque millésime.

  Mais on ne connaît pas de technique plus élaborée pour gérer ce problème.

Tableaux hiérarchisés

-

Nombre d'entreprises qui produisent des violons, par région.

Pays											
400											
Nord			Ouest				Est			Sud	
46			191				80			83	
N1	N2	N3	O1	O2	O3	O4	E1	E2	E3	S1	S2
21	2	23	32	54	67	38	27	41	12	44	39

Tableaux hiérarchisés

-

Si l'on ne renseigne pas la hiérarchie :

N1	N2	N3	Nord	O1	O2	O3	O4	Ouest	E1	E2	E3	Est	S1	S2	Sud	Pays
21	2	23	46	32	54	67	38	191	27	41	12	80	44	39	83	400

Tableaux hiérarchisés

-

Si l'on ne renseigne pas la hiérarchie :

N1	N2	N3	Nord	O1	O2	O3	O4	Ouest	E1	E2	E3	Est	S1	S2	Sud	Pays
21	2	23	46	32	54	67	38	191	27	41	12	80	44	39	83	400

N1	N2	N3	Nord	O1	O2	O3	O4	Ouest	E1	E2	E3	Est	S1	S2	Sud	Pays
21	X	23	46	32	54	67	38	191	27	41	X	80	44	39	83	400

Tableaux hiérarchisés

-

Si l'on ne renseigne pas la hiérarchie :

N1	N2	N3	Nord	O1	O2	O3	O4	Ouest	E1	E2	E3	Est	S1	S2	Sud	Pays
21	2	23	46	32	54	67	38	191	27	41	12	80	44	39	83	400

On peut retrouver la valeur cachée

N1	N2	N3	Nord	O1	O2	O3	O4	Ouest	E1	E2	E3	Est	S1	S2	Sud	Pays
21	X	23	46	32	54	67	38	191	27	41	X	80	44	39	83	400

Tableaux hiérarchisés

-

Nombre d'entreprises qui produisent des violons, par région.

Pays											
400											
Nord			Ouest				Est			Sud	
46			191				80			83	
N1	N2	N3	O1	O2	O3	O4	E1	E2	E3	S1	S2
21	2	23	32	54	67	38	27	41	12	44	39

Tableaux hiérarchisés

-

Nombre d'entreprises qui produisent des violons, par région.

Pays												
400												
Nord			Ouest				Est			Sud		
46			191				80			83		
N1	N2	N3	O1	O2	O3	O4	E1	E2	E3	S1	S2	
X	X	23	32	54	67	38	27	41	12	44	39	

Tableaux hiérarchisés

-

Définition : Tableau possédant une ou plusieurs variables de ventilation comprenant des sous-totaux.

Tableaux hiérarchisés

-

Définition : Tableau possédant une ou plusieurs variables de ventilation comprenant des sous-totaux.

- La prise en compte de ces liens est nécessaire pour gérer le secret secondaire.

- Exemples :
  - *Zones géographiques (Pays → Région → Département → Commune → IRIS)*
  - *Nomenclatures (NAF : Section → Division → Groupe → Classe → Sous-classe)*

- NB : Une hiérarchie doit être « bien emboîté », c'est-à-dire que chaque sous-ensemble doit être inclus dans un et un seul ensemble de niveau supérieur.

04 Logiciels

τ-Argus

-

Logiciels

τ-Argus

- http://research.cbs.nl/casc/tau.htm
- Permet de calculer le secret primaire et secondaire, ou faire des arrondis contrôlés
- Développement, maintenance et support réalisés par un groupe d'experts européens, et coordonné par CBS (INS des Pays-Bas)
- Le plus utilisé en Europe pour gérer la confidentialité statistique
- Exécutable sur Windows et Linux

τ-Argus

- Logiciel libre, donc open source

  → code source consultable ici : https://github.com/sdcTools/tauargus
- Utilisable via une interface graphique, ou par une macro SAS (mode batch)

  → macro disponible ici : https://github.com/InseeFrLab/SASTauArgus
- En cas de doute, allez voir :
  - le tutoriel de τ-Argus (disponible sur l'intranet (DMCSI > DMS > DRTI > Confidentialité) ou l'extranet (Méthodologie > RTI > Confidentialité))
  - le « User's Manual » (PDF dans le même dossier que l'exécutable)
  - ou contactez-nous directement !

Quatre algorithmes de suppressions secondaires ...

-

Méthode Hypercube	~~Méthode Network~~	Méthode Optimal	Méthode Modular
• Traitement séquentiel des cases sous secret primaire, en minimisant le coût • Si hiérarchie, alors traitement des n sous-tableaux indépendamment • Méthode heuristique • Méthode rapide	• ~~Méthode désuète~~ • ~~Modélisation du problème par graphes~~	• Optimisation : minimisation globale du coût • Respecte la règle des intervalles • Méthode souvent très lente • Utilise un solveur gratuit ou payant	• Éclatement du problème : application d'Optimal sur différents sous-tableaux • Donc les variables doivent être hiérarchiques • Assez rapide • Utilise un solveur gratuit ou payant

… Utilisant un solveur gratuit ou payant

- Optimal et Modular se basent sur un solveur permettant de résoudre un problème mathématique d'optimisation linéaire de fonction sous contrainte.

- Dans τ-Argus, il existe :
  - Un solveur gratuit
  - Deux solveurs payants (Xpress de FICO et CPlex d'IBM)

- Pour des travaux de production, il est conseillé d'utiliser l'un des solveurs payants car ils offrent de meilleurs résultats et fonctionnent parfois sur de plus grandes tables. Xpress semble fournir de meilleurs résultats que CPlex.

Un package R : sdcTable

-

- https://github.com/sdcTools/sdcTable
- Développé et maintenu par Statistics Austria (INS d'Autriche), en concertation avec le groupe d'experts européens sur la confidentialité
- Reprend une logique similaire à τ-Argus, et utilise les mêmes algorithmes pour le secret secondaire
- Disponible sur le CRAN
- Mis à jour régulièrement
- Vignette : https://cran.r-project.org/web/packages/sdcTable/vignettes/sdcTable.html

Un autre package R : rtauargus

-

- https://gitlab.insee.fr/py_b/rtauargus (lien interne Insee)
- Développé et maintenu à l'Insee, par le Pôle Démographie des Entreprises et des Établissements de Nancy
- Effectue des appels à τ-Argus depuis R, en mode batch
- Mis à jour régulièrement
- Documentation : http://py_b.gitlab-pages.insee.fr/rtauargus/

Éléments de bibliographie

-

- A. Hundepool *& al.* Handbook on statistical disclosure control, version 1.2, 2010.
- E. Griffins *& al.*, Handbook on Statistical Disclosure Control for Outputs, 2019.
- Insee. Guide du secret statistique, 27 juillet 2018.
- J. Nicolas. La gestion du secret dans les tableaux diffusant des statistiques d'entreprises, La Lettre du SSE, n° 65, septembre 2010.
- J. Nicolas. Traitement de la confidentialité statistique dans les tableaux : expérience de la Direction des Statistiques d'Entreprises.
- Journées de Méthodologie Statistique de l'Insee, 2012. - L. Willenborg and T. de Waal. Elements of Statistical Disclosure Control, Lecture Notes in Statistics, vol 155, Springer-Verlag, 2000.

Et merci pour votre attention ! ☺

Maxime Beauté
maxime.beaute@insee.fr
01 87 69 55 43

Alexandre Awad
alexandre.awad@insee.fr
01 87 69 55 14

Insee
Mesurer pour comprendre