La détermination des degrés de liberté au sein d'une enquête complexe : une évaluation par simulations

Marie-Hélène Toupin et Claude Girard
Statistique Canada

Séminaire en ligne sur les sondages
26 septembre 2024

Le contenu de cette présentation représente la position des auteurs, mais pas nécessairement celle de Statistique Canada. Cette présentation décrit des approches envisagées qui ne sont pas encore mises en œuvre dans les programmes de Statistique Canada.

Aperçu de la présentation

1. Éléments contextuels : les intervalles de confiance en avant-plan à Statistique Canada

2. Construction d'intervalle de confiance : des méthodes en vogue, dont celle de *Student*

3. Une simple règle approximative pour le calcul des degrés de liberté
   - Simulations : quatre cas de situations d'enquête simples

4. Un raffinement de la règle approximative
   - Simulations : un cas d'enquête complexe

Aperçu de la présentation

1. Éléments contextuels : les intervalles de confiance en avant-plan à Statistique Canada

2. Construction d'intervalle de confiance : des méthodes en vogue, dont celle de Student

3. Une simple règle approximative pour le calcul des degrés de liberté
   - Simulations : quatre cas de situations d'enquête simples

4. Un raffinement de la règle approximative
   - Simulations : un cas d'enquête complexe

Contexte

- Recensement de la population du Canada :
  - Quinquennal et le dernier a été tenu en 2021
  - Deux questionnaires : court (tous les logements) et détaillé (un logement sur 4 aléatoirement choisi)
  - En 2021, pour la première fois, précision de nombreuses estimations d'effectifs issues du questionnaire détaillé sont véhiculées en recourant aux intervalles de confiance plutôt qu'aux traditionnels coefficients de variation (Neusy et Mantel [2016])

- Plusieurs enquêtes sociales ont aussi recours aux intervalles de confiance (IC), dont les enquêtes postcensitaires (2022) :
  - Enquête sur les peuples autochtones
  - Enquête sur la population de langue officielle en situation minoritaire
  - Enquête canadienne sur l'incapacité

Contexte

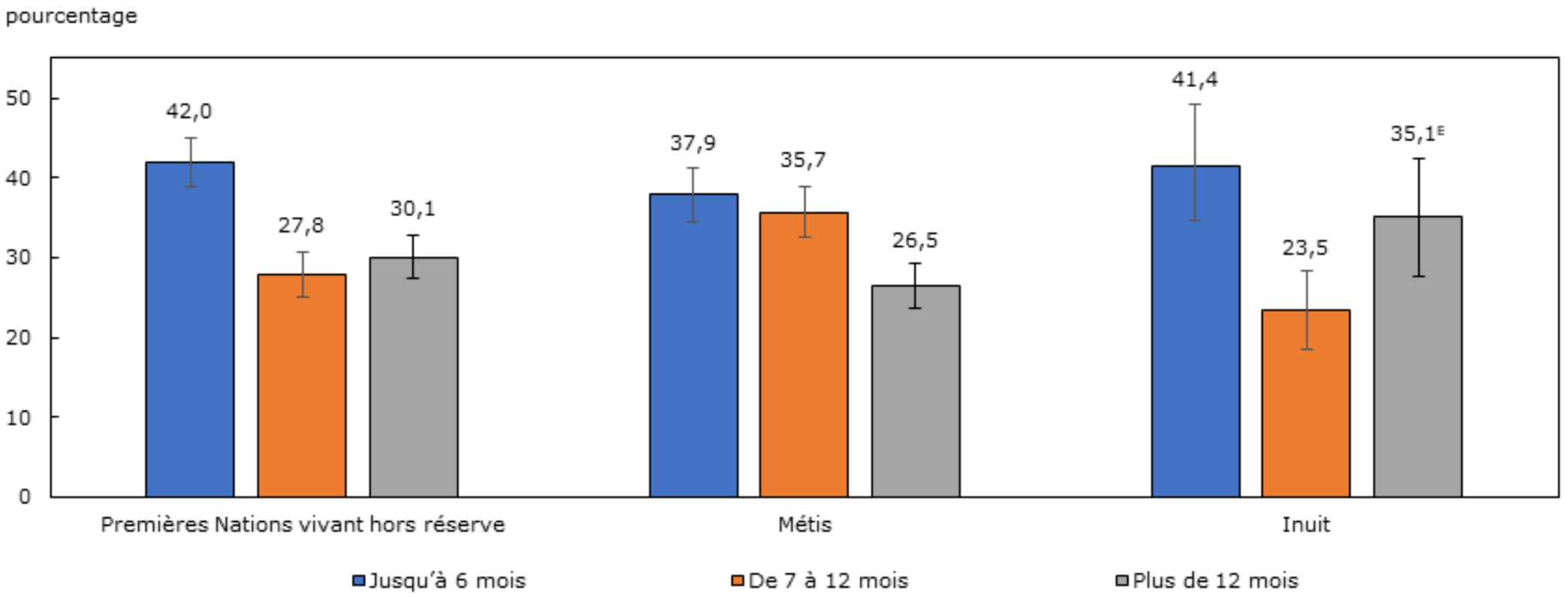
Nombre d'Autochtones, selon leur identité autochtone, Canada, 2021

Identité autochtone	Nombre estimé d'individus	Intervalle de confiance de niveau 95 %	
		Borne inférieure	Borne supérieure
Réponses autochtones uniques	1 743 165	1 738 253	1 748 091
Premières Nations (Indiens de l'Amérique du Nord)	1 048 405	1 045 710	1 051 107
Métis	624 220	619 866	628 605
Inuk (Inuit)	70 545	69 735	71 364
Réponses autochtones multiples	28 860	27 938	29 812
Réponses autochtones non comprises ailleurs	35 225	34 559	35 904
Total	**1 807 250**	**1 802 337**	**1 812 176**

Source : *Recensement canadien de la population de 2021 (statcan.gc.ca)*

Contexte

Durée de l'allaitement (en mois) chez les enfants âgés de 1 à 5 ans, selon l'identité autochtone, Canada, 2022



E à utiliser avec prudence
Note : Les barres d'erreur représentent un intervalle de confiance à 95 %.
Source : Statistique Canada, Enquête auprès des peuples autochtones, 2022.

Interprétation des IC

- IC est **valide** lorsque sa couverture effective correspond à celle visée de 95 %
- Interprétation fréquentiste :
  - n'est pas révélatrice pour les non-initiés
  - est source de confusion même chez les initiés : « 95 % de chances que [3 %, 14 %] contienne la vraie proportion » [Trompeuse, car une fois calculé l'IC n'a plus rien d'aléatoire]
- Plutôt : IC est une **fourchette de valeurs plausibles** (qui inclut l'estimation elle-même) selon les données ⇨ Sensibilise les utilisateurs au fait que l'estimation *n'est pas la seule valeur plausible*
- Un IC *valide* étroit renforce l'hypothèse que l'estimation soit proche de la vraie valeur (car les autres valeurs plausibles lui sont similaires)

Aperçu de la présentation

1. Éléments contextuels : les intervalles de confiance en avant-plan à Statistique Canada

2. Construction d'intervalle de confiance : des méthodes en vogue, dont celle de Student

3. Une simple règle approximative pour le calcul des degrés de liberté
   - Simulations : quatre cas de situations d'enquête simples

4. Un raffinement de la règle approximative
   - Simulations : un cas d'enquête complexe

Différentes constructions d'IC d'emploi courant

- Approche paramétrique générale basée sur la distribution de la statistique pivotale employée
  - **Wald :** $\hat{\theta} \pm \mathbf{1,96}\sqrt{V(\hat{\theta})}$
    - Suppose la normalité (grands échantillons)
  - **Student :** $\hat{\theta} \pm \boldsymbol{t_{DL;0,975}}\sqrt{\hat{V}(\hat{\theta})}$
    - $t_{DL;0,975}$ est le quantile de Student à *DL* degrés de liberté
    - Adaptée aux petits échantillons

Différentes constructions d'IC d'emploi courant

- Approche paramétrique générale basée sur la distribution de la statistique pivotale employée
  - **Wald :** $\hat{\theta} \pm \mathbf{1,96}\sqrt{V(\hat{\theta})}$
    - Suppose la normalité (grands échantillons)
  - **Student :** $\hat{\theta} \pm \boldsymbol{t_{DL;0,975}}\sqrt{\hat{V}(\hat{\theta})}$
    - $t_{DL;0,975}$ est le quantile de Student à *DL* degrés de liberté
    - Adaptée aux petits échantillons

Wald et Student sont employés pour une **variable continue** et ne conviennent pas pour l'estimation d'une **proportion** extrême tirée d'un échantillon de taille insuffisante

Différentes constructions d'IC d'emploi courant

- Approche paramétrique générale basée sur la distribution de la statistique pivotale employée
  - **Wald :** $\hat{\theta} \pm \mathbf{1,96}\sqrt{V(\hat{\theta})}$
    - Suppose la normalité (grands échantillons)
  - **Student :** $\hat{\theta} \pm \boldsymbol{t_{DL;0,975}}\sqrt{\hat{V}(\hat{\theta})}$
    - $t_{DL;0,975}$ est le quantile de Student à *DL* degrés de liberté
    - Adaptée aux petits échantillons
  - **Wilson modifié**
    - Conçue sur mesure pour les proportions
    - Comporte des degrés de liberté comme paramètre

Wald et Student sont employés pour une **variable continue** et ne conviennent pas pour l'estimation d'une **proportion** extrême tirée d'un échantillon de taille insuffisante

Différentes constructions d'IC d'emploi courant

- Approche paramétrique générale basée sur la distribution de la statistique pivotale employée
  - **Wald :** $\hat{\theta} \pm \mathbf{1,96}\sqrt{V(\hat{\theta})}$
    - Suppose la normalité (grands échantillons)
  - **Student :** $\hat{\theta} \pm \boldsymbol{t_{DL;0,975}}\sqrt{\hat{V}(\hat{\theta})}$
    - $t_{DL;0,975}$ est le quantile de Student à *DL* degrés de liberté
    - Adaptée aux petits échantillons
  - **Wilson modifié**
    - Conçue sur mesure pour les proportions
    - Comporte des degrés de liberté comme paramètre

Wald et Student sont employés pour une **variable continue** et ne conviennent pas pour l'estimation d'une **proportion** extrême tirée d'un échantillon de taille insuffisante

- Approche générale non paramétrique : Percentiles Bootstrap

Un exemple de la statistique classique donne le ton

- Rappel théorique : considérons *n* variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes et identiquement distribuées (« i.i.d. ») $N(\mu, \sigma^2)$
- Soient

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{et} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Un exemple de la statistique classique donne le ton

- Rappel théorique : considérons *n* variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes et identiquement distribuées (« i.i.d. ») $N(\mu, \sigma^2)$
- Soient

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{et} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Cas où σ^2 est connue

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

Un exemple de la statistique classique donne le ton

- Rappel théorique : considérons *n* variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes et identiquement distribuées (« i.i.d. ») $N(\mu, \sigma^2)$
- Soient

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{et} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Cas où σ^2 est connue

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

Cas où σ^2 est inconnue

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim \text{Student}(DL = n-1)$$

Un exemple de la statistique classique donne le ton

- Rappel théorique : considérons *n* variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes et identiquement distribuées (« i.i.d. ») $N(\mu, \sigma^2)$
- Soient

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{et} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Cas où σ^2 est connue

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

Cas où σ^2 est inconnue

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim \text{Student}(DL = n-1) \xrightarrow{Loi} N(0,1)$$

Un exemple de la statistique classique donne le ton

- Rappel théorique : considérons *n* variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes et identiquement distribuées (« i.i.d. ») $N(\mu, \sigma^2)$
- Soient

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{et} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Cas où σ^2 est connue

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

Cas où σ^2 est inconnue

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim \text{Student}(DL = n-1) \xrightarrow{Loi} N(0,1)$$

- Donc, les *DL* surviennent en pratique car la variance doit être estimée

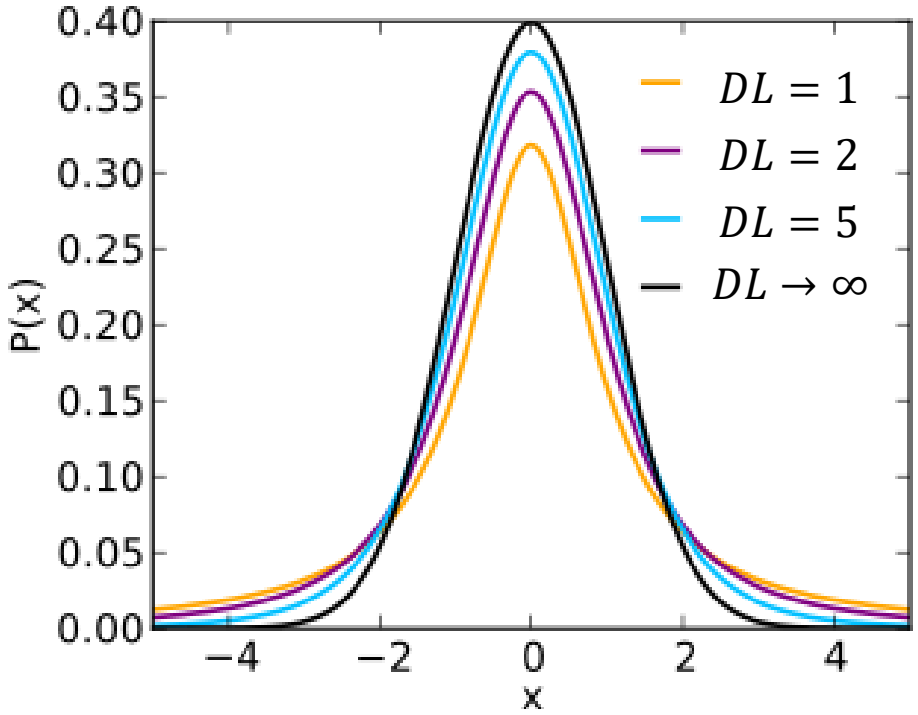
Degrés de liberté comme paramètre de couverture

- Comment varie la couverture réalisée de l'IC de *Student*

$$\bar{X} \pm t_{DL;0,975}\sqrt{\widehat{Var}(\bar{X})}$$

en fonction des *DL* précisés (via le quantile $t_{DL;0,975}$)?

Densité de la loi de *Student* avec *DL* degrés de liberté



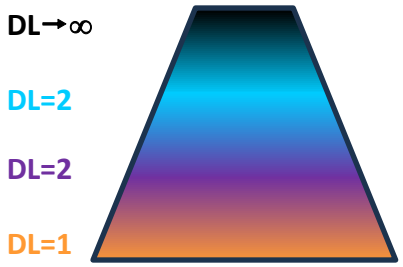
Degrés de liberté comme paramètre de couverture

- Comment varie la couverture réalisée de l'IC de *Student*

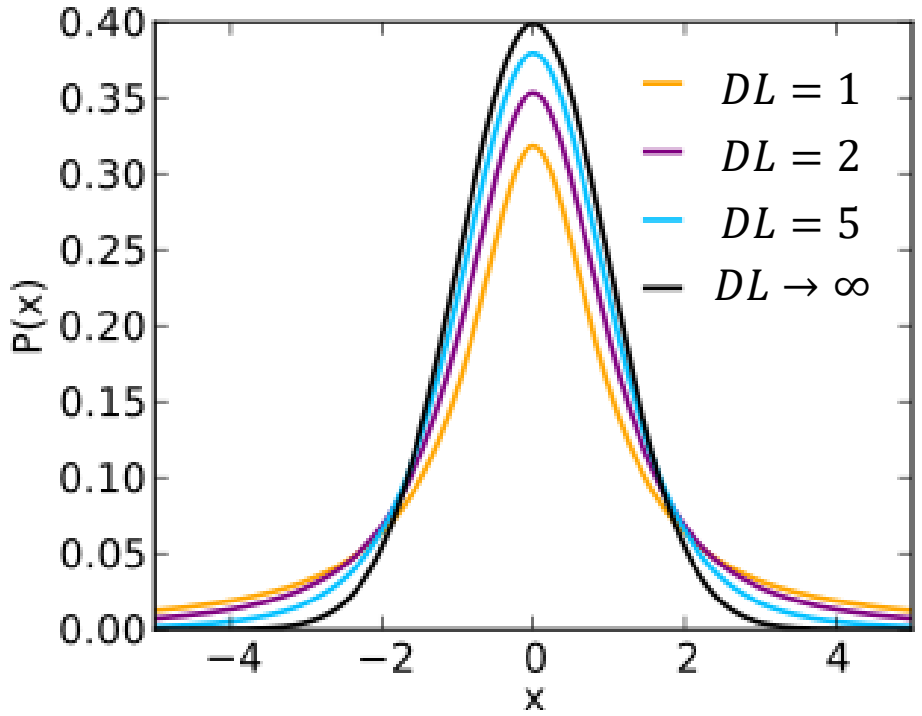
$$\bar{X} \pm t_{DL;0,975}\sqrt{\widehat{Var}(\bar{X})}$$

en fonction des *DL* précisés (via le quantile $t_{DL;0,975}$)?

Longueur et couverture d'IC



Densité de la loi de *Student* avec *DL* degrés de liberté



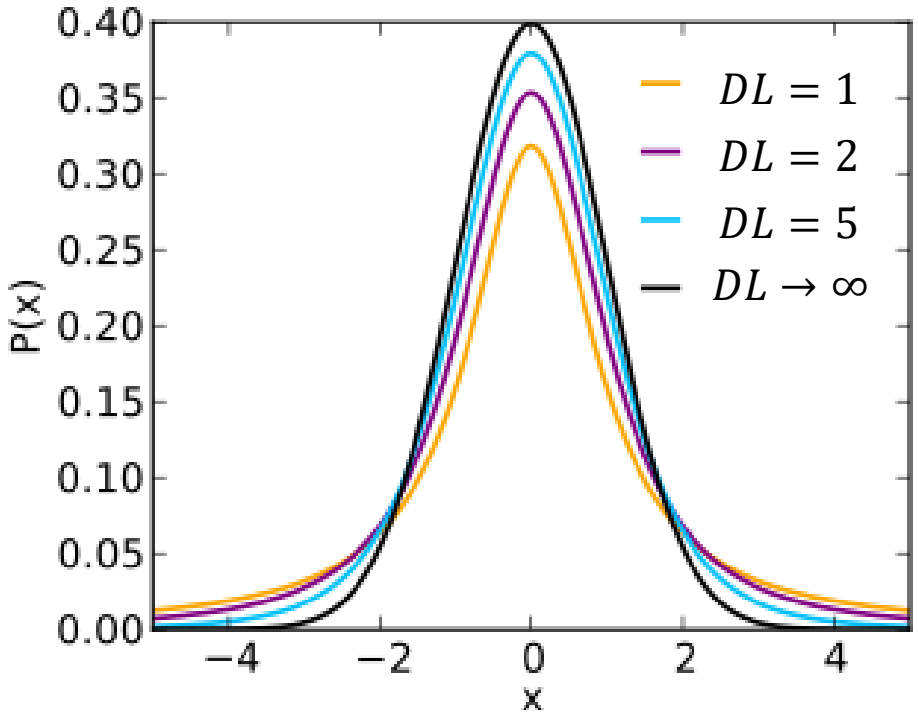
Degrés de liberté comme paramètre de couverture

- Comment varie la couverture réalisée de l'IC de *Student*

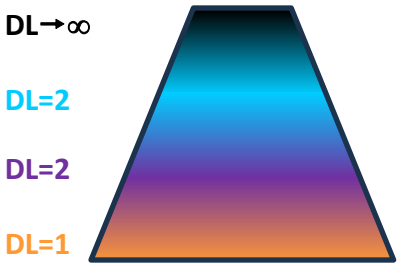
$$\bar{X} \pm t_{DL;0,975}\sqrt{\widehat{Var}(\bar{X})}$$

en fonction des *DL* précisés (via le quantile $t_{DL;0,975}$)?

Densité de la loi de *Student* avec *DL* degrés de liberté



Longueur et couverture d'IC



- L'IC comporte donc un paramètre de couverture via *DL*
- *DL* exagérés produit IC « sur-confiant » : estimation paraît plus précise qu'elle ne l'est

Aperçu de la présentation

1. Éléments contextuels : les intervalles de confiance en avant-plan à Statistique Canada

2. Construction d'intervalle de confiance : des méthodes en vogue, dont celle de Student

3. Une simple règle approximative pour le calcul des degrés de liberté
   - Simulations : quatre cas de situations d'enquête simples

4. Un raffinement de la règle approximative
   - Simulations : un cas d'enquête complexe

Degrés de liberté comme paramètre de couverture

- Comment déterminer la valeur de *DL* afin d'obtenir la couverture visée de 95 % ?
  - Les degrés de liberté sont difficiles à calculer exactement… dans un contexte d'enquête
  - Règle approximative couramment employée dans les enquêtes :

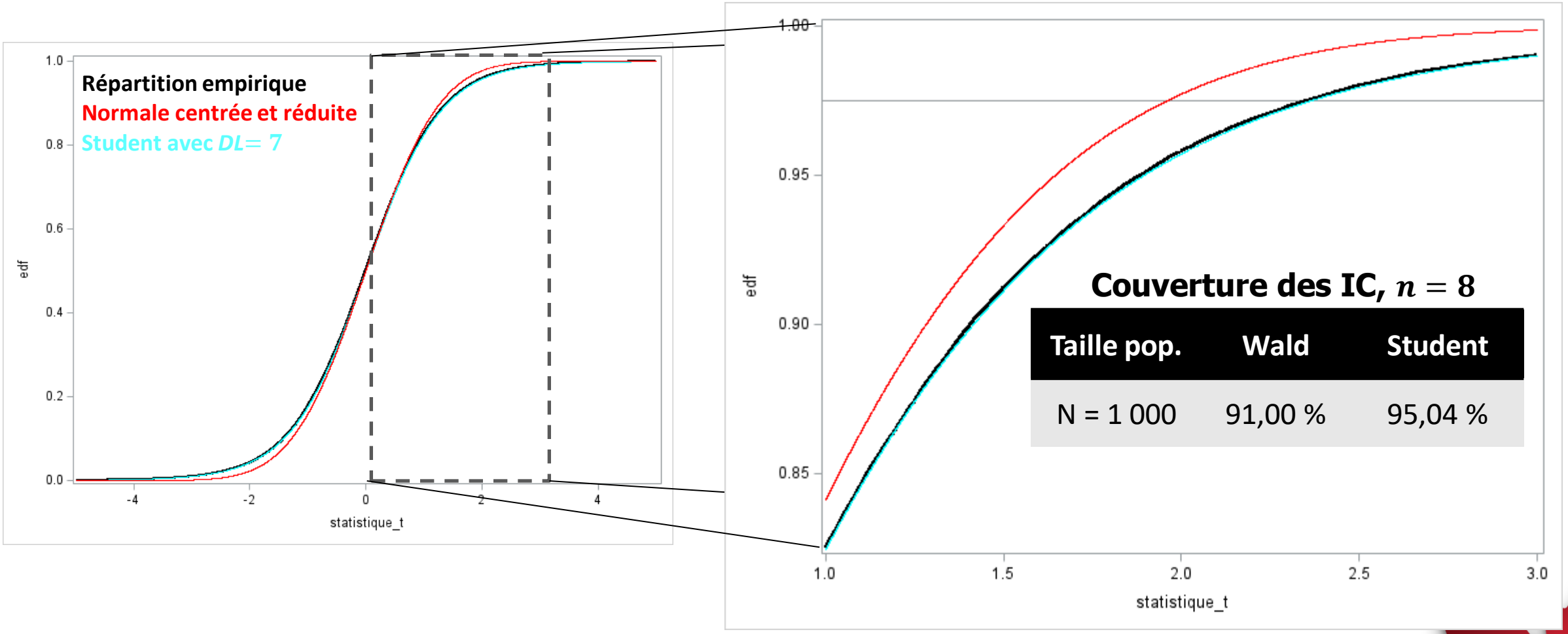
$DL = n$ – nombre de strates

- On se propose de mettre à l'épreuve cette formule pour des situations d'enquête choisies à l'aide de simulations

Passage aux données d'enquête complexe

- **Cas 1 :** échantillonnage aléatoire simple *sans remise* (EASSR) de n unités à partir des N qui composent la population (finie) d'intérêt et emploi de l'estimateur usuel de la variance
- Est-ce que la statistique T et (surtout) la loi de *Student* qui lui est associée s'y emploient encore?
- Simulation :
  - $N = 1\,000$ et $n = 8$
  - Variable d'intérêt d'allure normale
  - 100 000 échantillons sélectionnés et pour chacun la valeur de *T* est obtenue
  - Est-ce que la loi *Student* à 8-1 degrés de liberté « colle » bien à la distribution empirique des valeurs-*T*?
  - Qu'en est-il de la couverture effective des intervalles de confiance de *Wald* et *Student*?

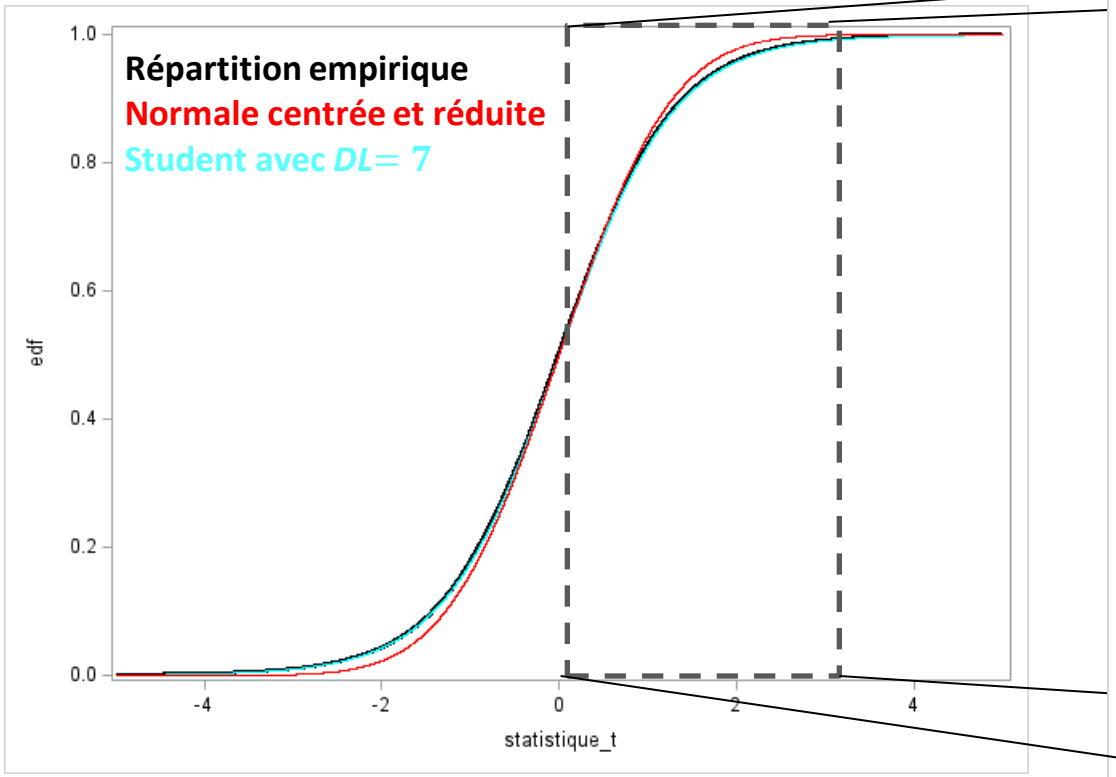
Passage aux données d'enquête complexe

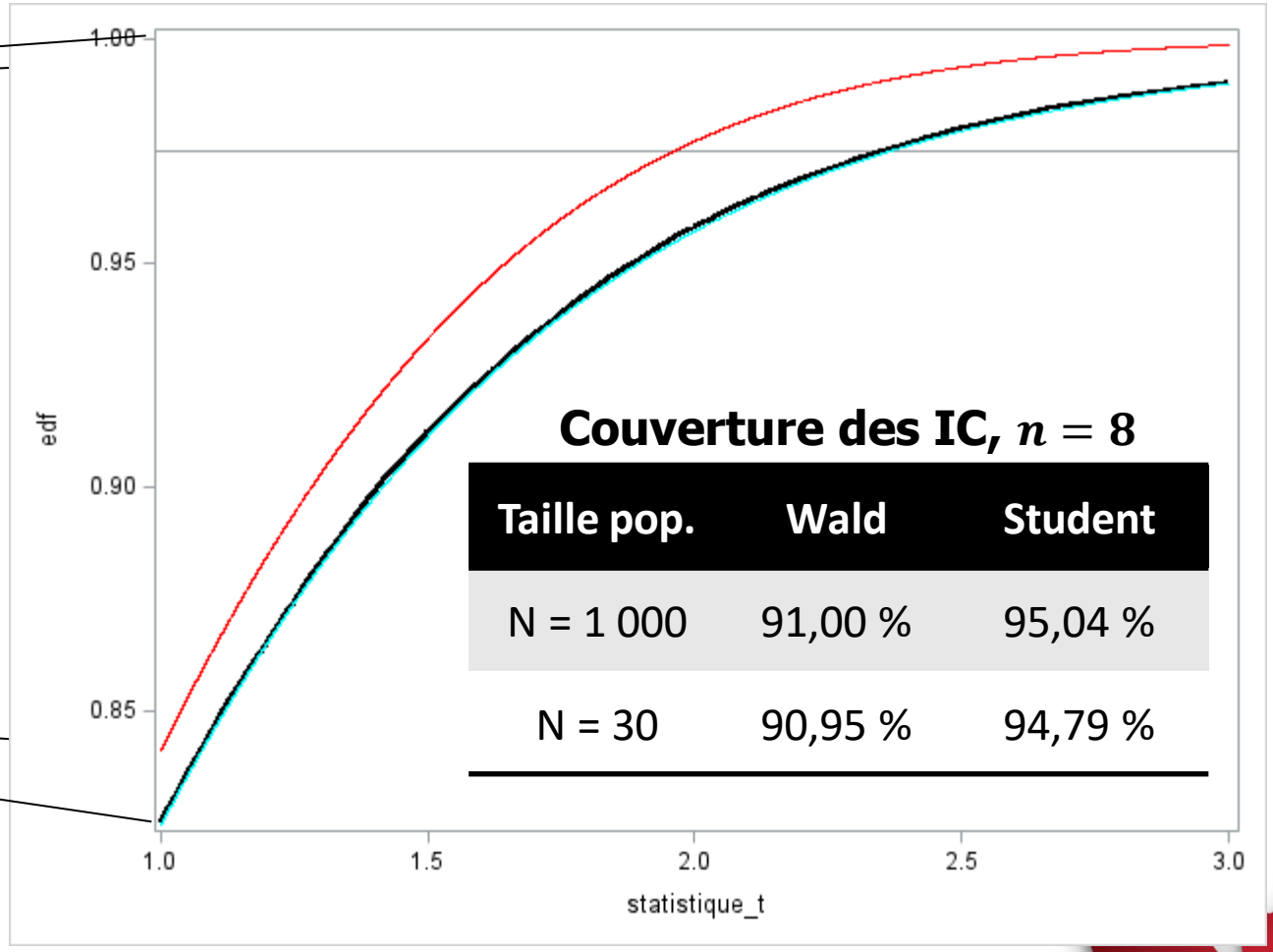


Couverture des IC, $n = 8$

Taille pop.	Wald	Student
N = 1 000	91,00 %	95,04 %

Passage aux données d'enquête complexe





Couverture des IC, $n = 8$

Taille pop.	Wald	Student
N = 1 000	91,00 %	95,04 %
N = 30	90,95 %	94,79 %

Passage aux données d'enquête complexe

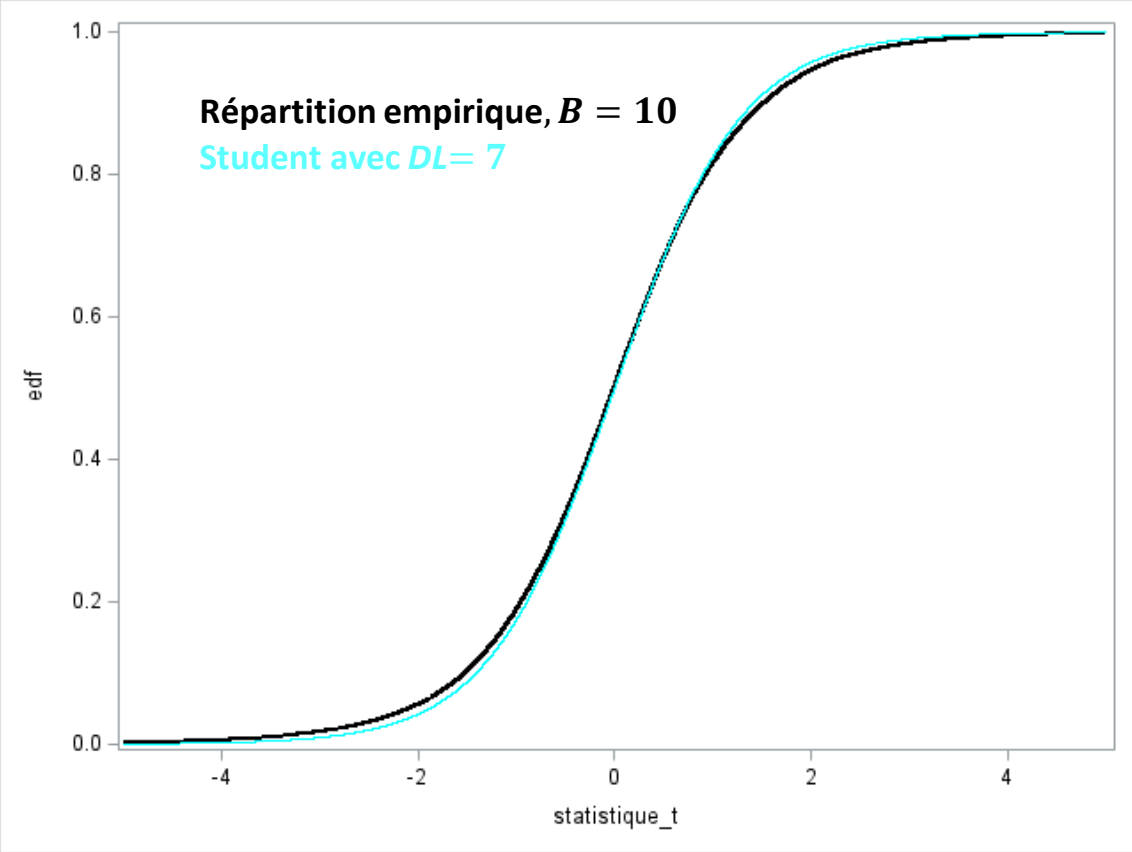
- **Cas 2 :** idem au cas 1, mais estimation Bootstrap *de la variance* basée sur *B* répliques

- *Rao-Wu Rescaling Bootstrap* est une adaptation aux données d'enquête couramment employée à Statistique Canada de la célèbre méthode de Bradley Efron

- Est-ce que le paramètre de couverture dépend du choix de *B,* et donc de la *méthode* d'estimation de la variance employée?

- Simulation :
  - $B = 10$ et $1\,000$ (valeur usuellement employée en pratique à Statistique Canada)

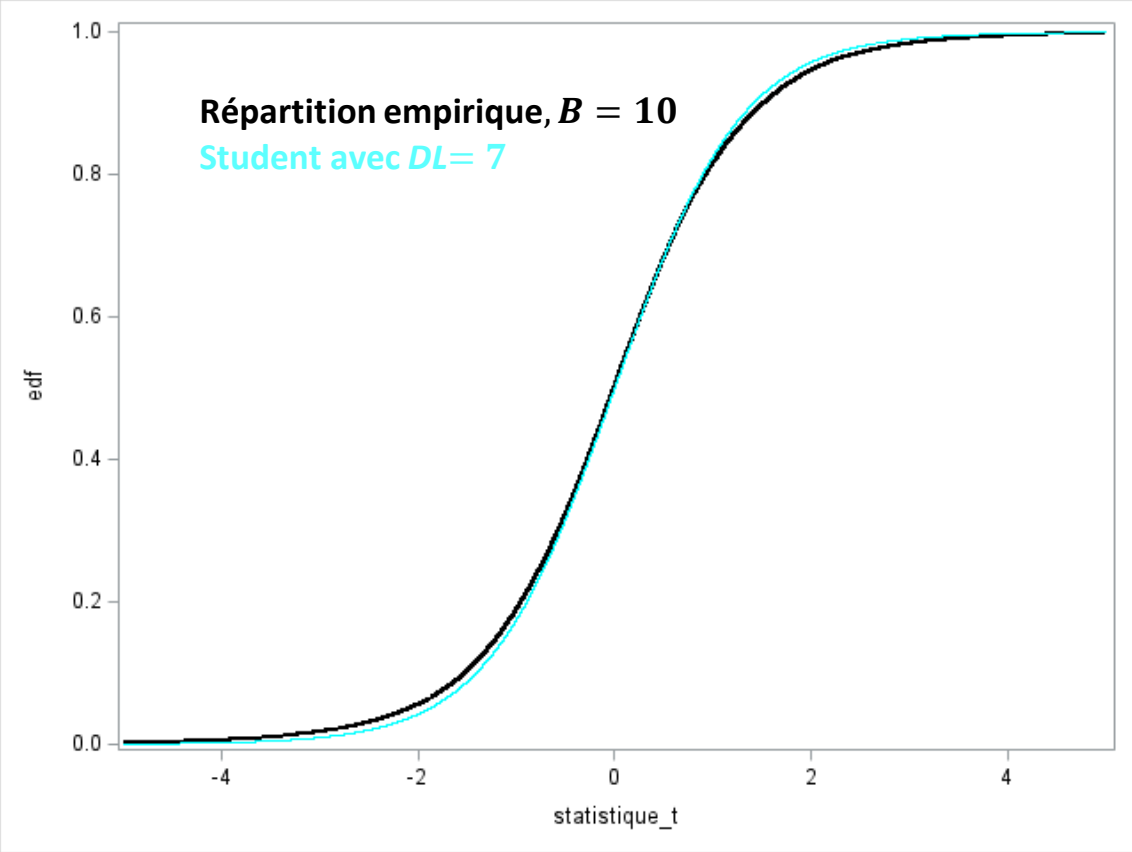
Passage aux données d'enquête complexe



Couverture, $N = 1\,000$, $n = 8$

# répliques Bootstrap	Student
$B = 10$	92,90 %

Passage aux données d'enquête complexe



Couverture, $N = 1\,000$, $n = 8$

# répliques Bootstrap	Student
$B = 10$	92,90 %
$B = 1\,000$	95,14 %

Réponse: Oui !

Seulement si *B* est grand peut-on raisonnablement supposer ici que $v_{BOOT} \cong \hat{V}$ et donc utiliser *DL* = 8-1

Autrement, le Bootstrap contribue une erreur de son propre cru qui s'ajoute à celle qui entache $\hat{V}$ et qui n'est donc pas prise en compte

Passage aux données d'enquête complexe

- Et donc le paramètre de couverture dépend de la méthode d'estimation de la variance

- Avertissement : combien de degrés de liberté sont utilisés par le logiciel employé? À titre d'exemple, SAS utilise par défaut *DL*=# répliques

If you use a REPWEIGHTS statement to provide replicate weights, the degrees of freedom equals the number of replicates, which is the number of REPWEIGHTS variables that you provide. Alternatively, you can use the DF= option in the REPWEIGHTS or the TABLES statement to specify the degrees of freedom.

Source : SAS Help Center: Degrees of Freedom

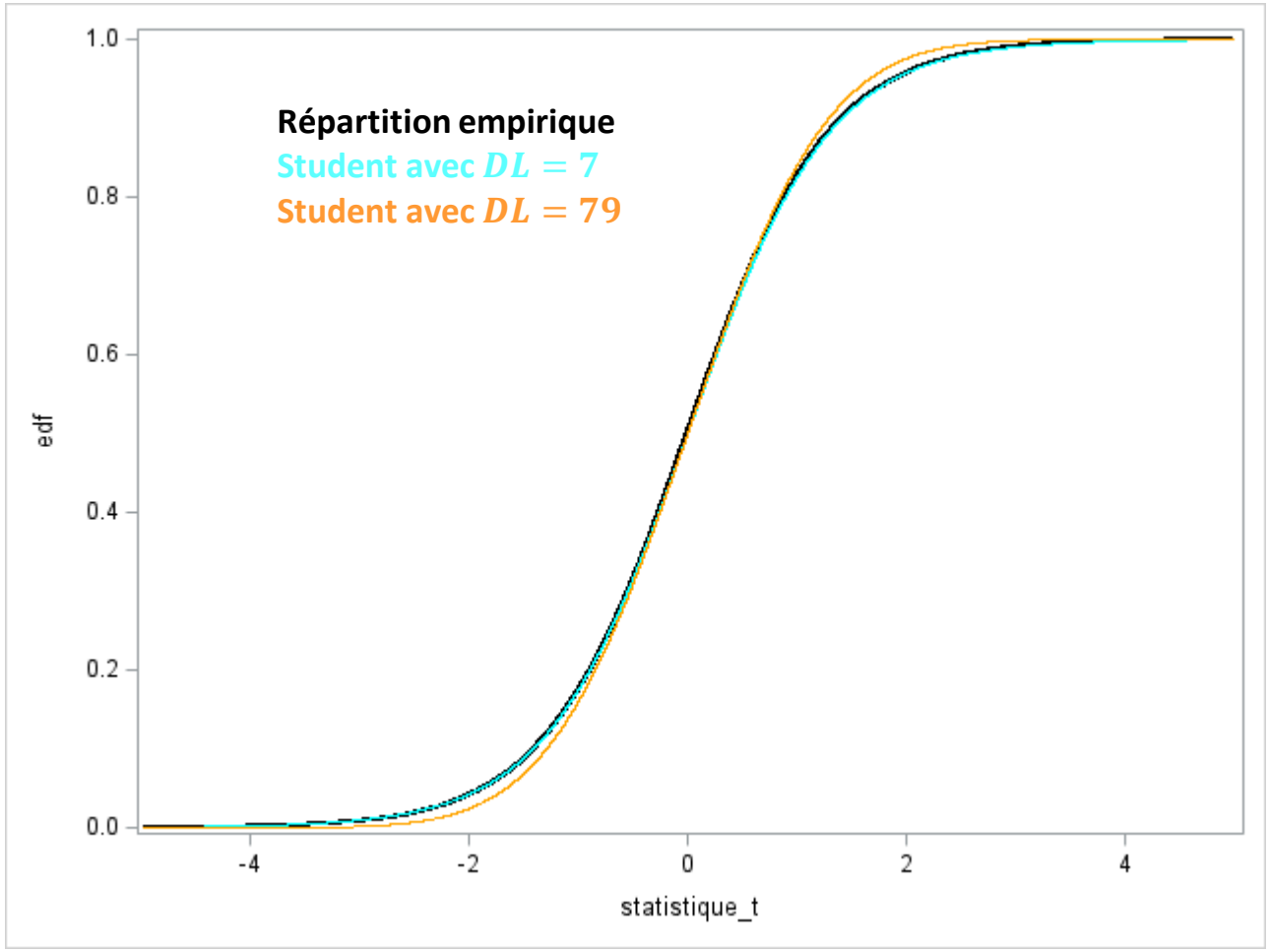
Passage aux données d'enquête complexe

- **Cas 3** : tirage aléatoire simple sans remise de n *grappes*, chacune comportant M unités, à partir des N/M grappes qui composent la population (finie)

- Les degrés de liberté de la loi *Student* dépendent de « n »… mais la taille *duquel* échantillon? Celui de $n \times M$ unités envoyées à la collecte ou celui des n grappes choisies au départ?

- Simulation :
  - $N = 1\,000$ formée de 100 grappes chacune de taille $M = 10$ et $n = 8$ ⇨ $n = 80$ unités envoyées à la collecte
  - Variable d'intérêt d'allure normale
  - 100 000 échantillons sélectionnés et pour chacun la valeur de *T* est obtenue
  - Laquelle loi collera le mieux à l'histogramme des valeurs-*T* : *DL=* 8-1 (échantillonnage) ou *DL*=80-1 (collecte)?

Passage aux données d'enquête complexe



Couverture, $n = 8$ grappes (80 unités)

Paramètre de couverture	Student
$DL = 7$	95,16 %
$DL = 79$	91,44 %

Passage aux données d'enquête complexe

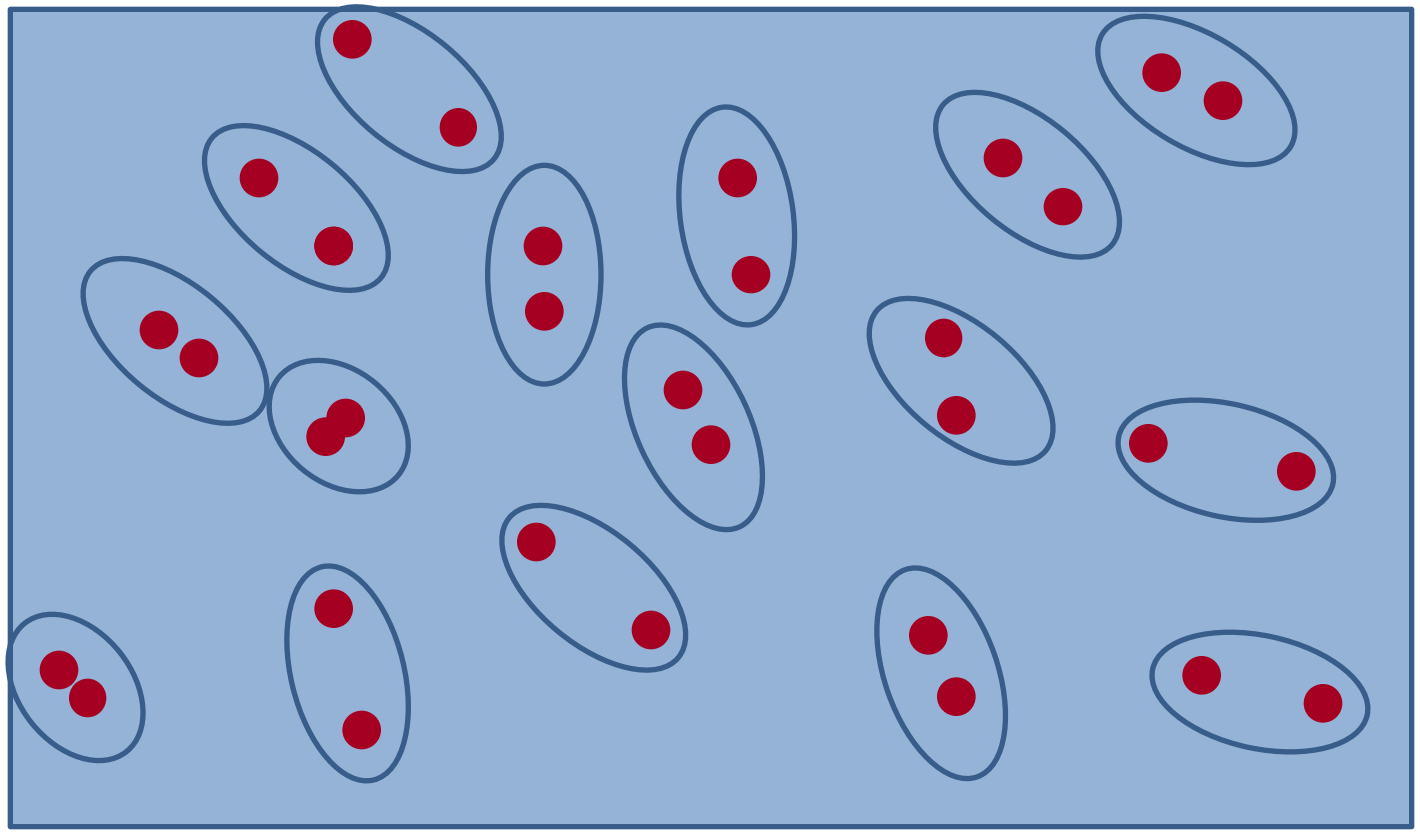
- Les *DL* sont souvent présentés comme le nombre de valeurs de l'échantillon qui peuvent être librement remplacées (par d'autres) tout en satisfaisant aux contraintes statistiques en vigueur

- Question : est-ce que *T* réagirait faiblement ou fortement au retrait et remplacement d'une seule unité *d'observation* de *s*?

- Réponse : dans un plan en grappes, un tel changement exige de retirer et remplacer la *grappe entière* à laquelle l'unité d'observation appartient ⇨ La variabilité de la statistique *T* sera donc grossière, car la composition de *s* ne peut pas changer librement par « petits incréments »

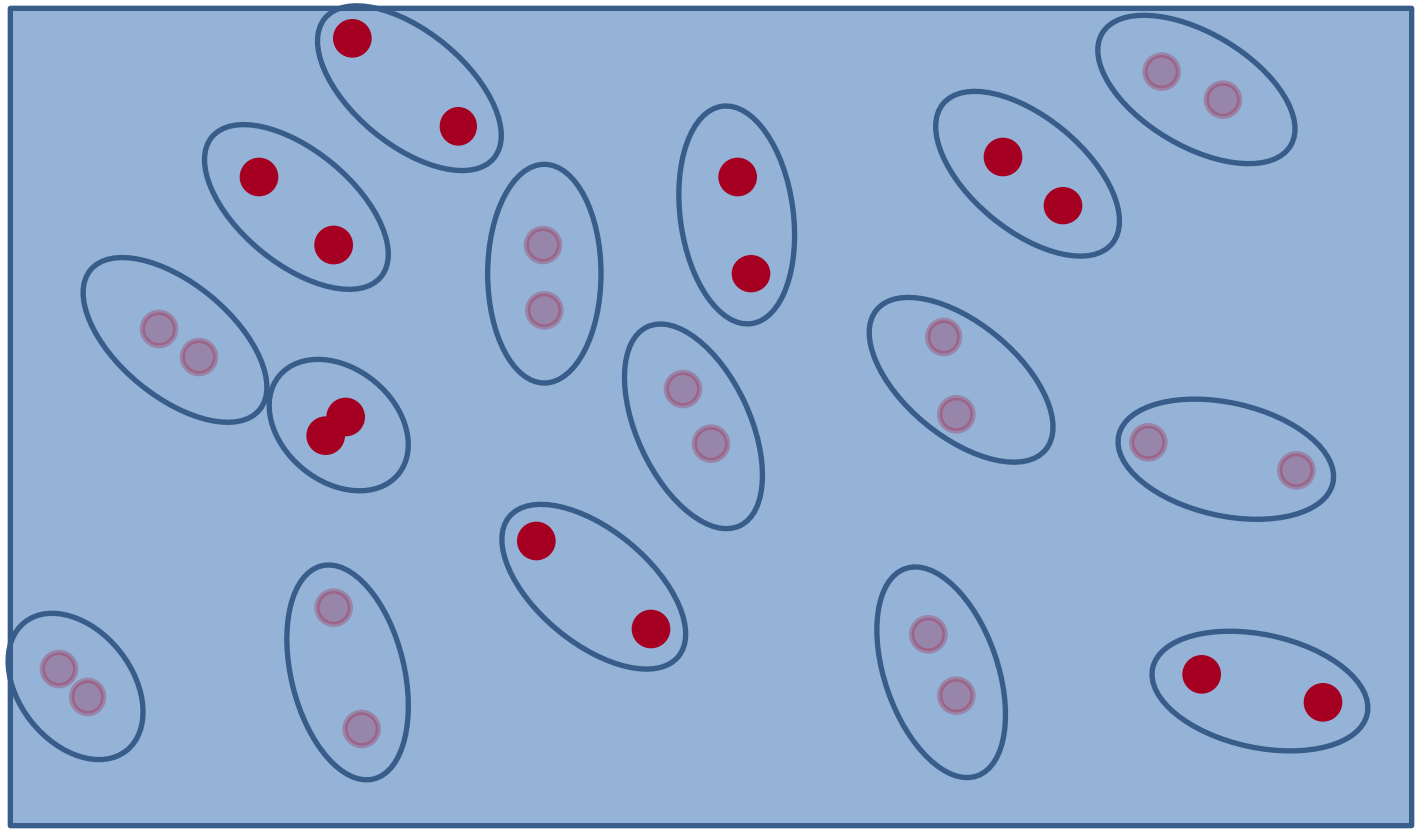
Passage aux données d’enquête complexe

La population, avec mise en grappes



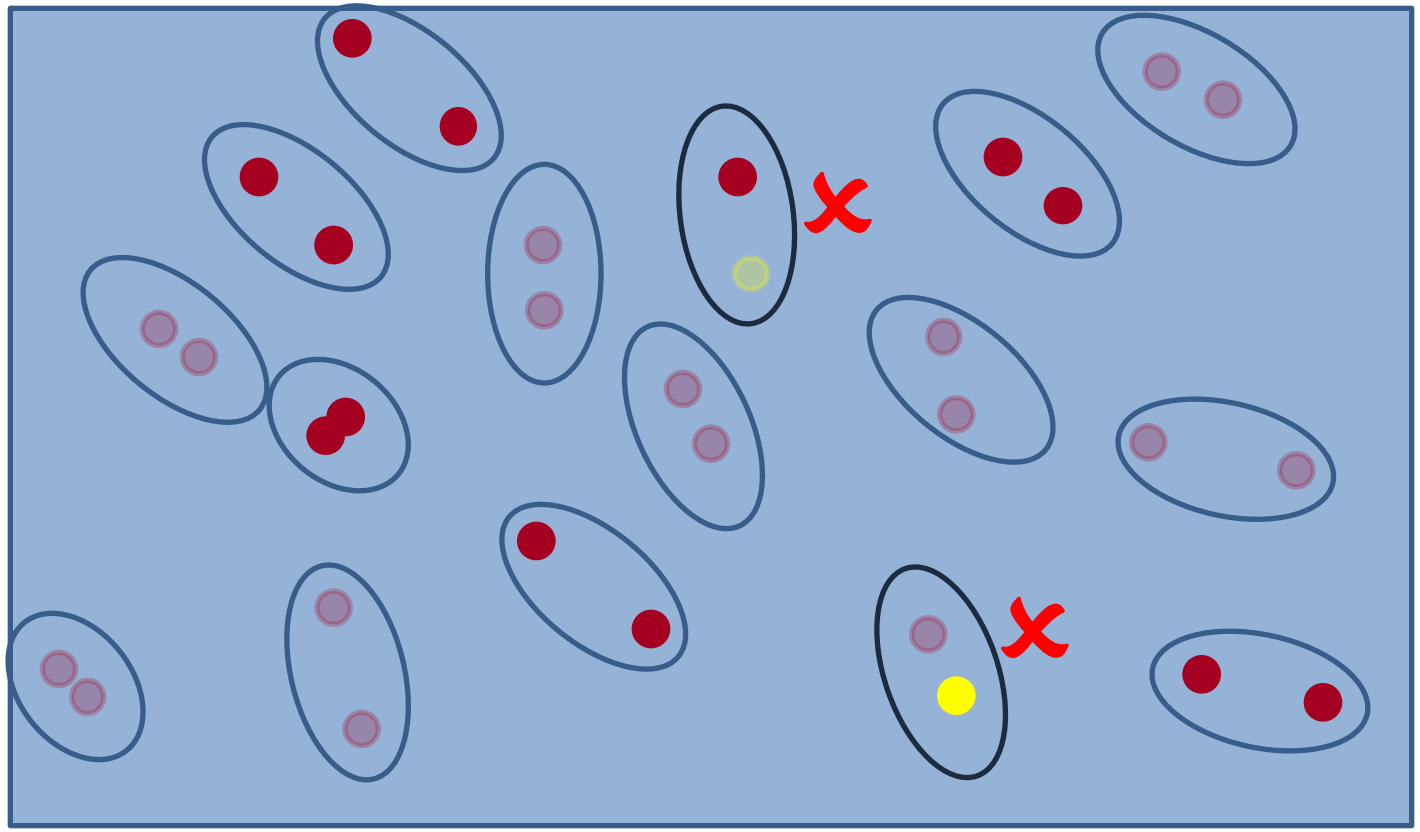
Passage aux données d’enquête complexe

L’échantillon de grappes



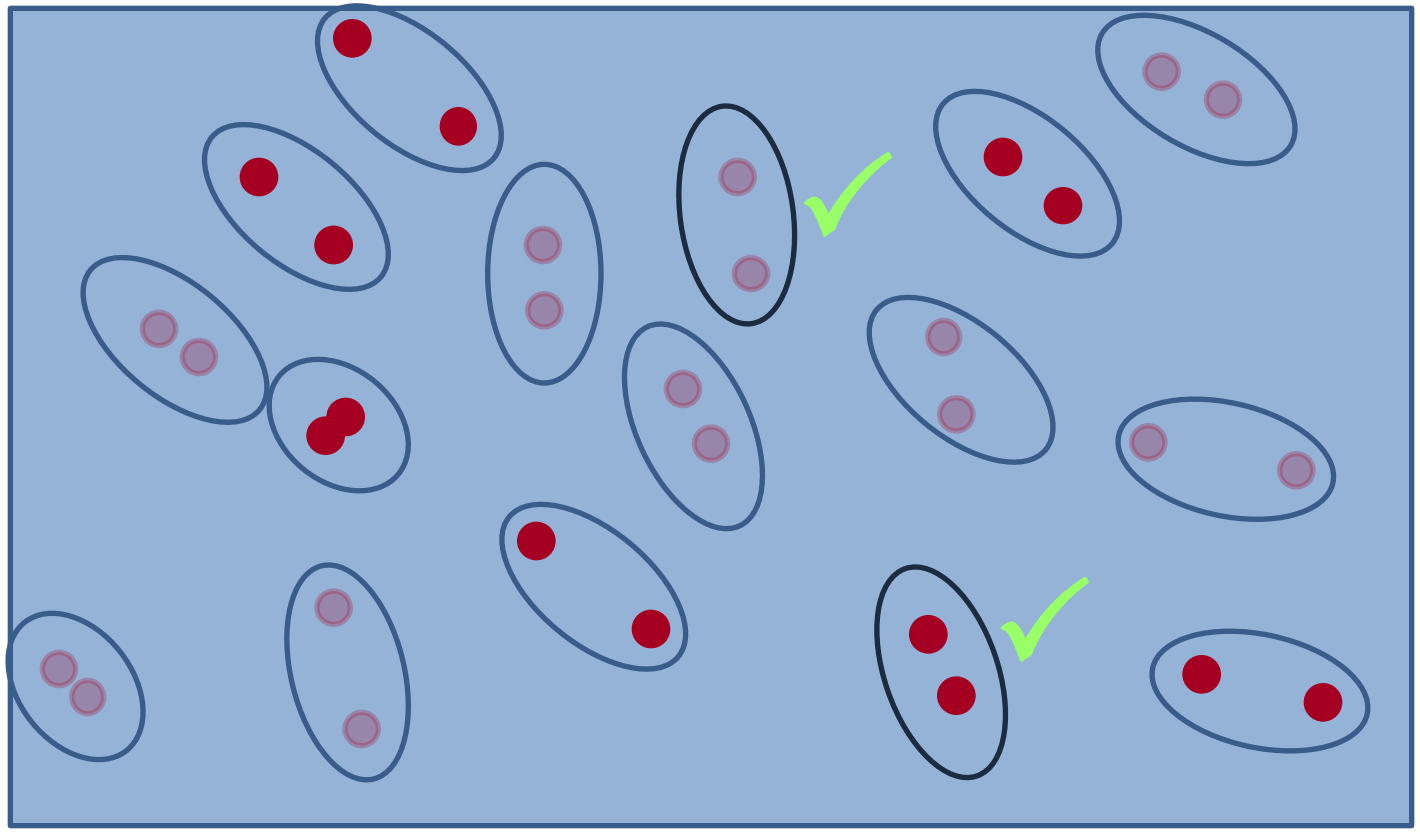
Passage aux données d’enquête complexe

À vouloir effectuer un changement « subtil » en ne retirant que l’unité jaune pâle choisie…



Passage aux données d’enquête complexe

…on se voit contraint d’effectuer le remplacement « massif » de la grappe en question



Passage aux données d'enquête complexe

- **Cas 4 :** tirage aléatoire stratifié simple sans remise de n *éléments*
- Simulation : 3 strates de taille 400 chacune et de dispersions 4, 9 et 16
- Selon la règle approximative, la couverture *sera la même quelle que soit* la répartition de la taille totale *n* = 12 aux strates – est-ce bien le cas?

Répartition	(n_1, n_2, n_3)
Neyman	(3,4,5)
Proportionnelle	(4,4,4)
Quelconque	(5,4,3)

Passage aux données d'enquête complexe

- **Cas 4 :** tirage aléatoire stratifié simple sans remise de n *éléments*
- Simulation : 3 strates de taille 400 chacune et de dispersions 4, 9 et 16
- Selon la règle approximative, la couverture *sera la même quelle que soit* la répartition de la taille totale *n* = 12 aux strates – est-ce bien le cas?

Répartition	(n_1, n_2, n_3)	Couverture ($DL = 12 - 3 = 9$)
Neyman	(3,4,5)	94,99 %
Proportionnelle	(4,4,4)	94,25 %
Quelconque	(5,4,3)	92,34 %

Aperçu de la présentation

1. Éléments contextuels : les intervalles de confiance en avant-plan à Statistique Canada

2. Construction d'intervalle de confiance : des méthodes en vogue, dont celle de Student

3. Une simple règle approximative pour le calcul des degrés de liberté
   - Simulations : quatre cas de situations d'enquête simples

4. Un raffinement de la règle approximative
   - Simulations : un cas d'enquête complexe

Satterthwaite : un raffinement de la règle

- Sous l'hypothèse que l'estimateur de variance et la variance sous le plan de sondage *P* vérifient

$$DL \times \frac{\hat{V}_P}{V_P} \sim \chi^2_{DL}$$

Satterthwaite : un raffinement de la règle

- Sous l'hypothèse que l'estimateur de variance et la variance sous le plan de sondage *P* vérifient

$$DL \times \frac{\hat{V}_P}{V_P} \sim \chi^2_{DL}$$

on en déduit que

$$V\left(DL \times \frac{\hat{V}_P}{V_P}\right) = V(\chi^2_{DL}) = 2 \times DL$$

Satterthwaite : un raffinement de la règle

- Sous l'hypothèse que l'estimateur de variance et la variance sous le plan de sondage *P* vérifient

$$DL \times \frac{\hat{V}_P}{V_P} \sim \chi^2_{DL}$$

on en déduit que

$$V\left(DL \times \frac{\hat{V}_P}{V_P}\right) = V(\chi^2_{DL}) = 2 \times DL$$

d'où l'approximation de type Satterthwaite (1946)

$$DL = \frac{2 \times V_P^2}{V(\hat{V}_P)}$$

Satterthwaite : un raffinement de la règle

- Sous l'hypothèse que l'estimateur de variance et la variance sous le plan de sondage *P* vérifient

$$DL \times \frac{\hat{V}_P}{V_P} \sim \chi^2_{DL}$$

on en déduit que

$$V\left(DL \times \frac{\hat{V}_P}{V_P}\right) = V(\chi^2_{DL}) = 2 \times DL$$

d'où l'approximation de type Satterthwaite (1946)

$$DL = \frac{2 \times V_P^2}{V\left(\hat{V}_P\right)} = \frac{2}{CV\left(\hat{V}_P\right)^2}$$

- Plus $\hat{V}_P$ est instable, moindre seront les *DL* associés

Satterthwaite : un raffinement de la règle

- Dans le cas stratifié (cas #4), l'approximation de type Satterthwaite est

$$DL = \frac{2\left\{\sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(\frac{N_h - n_h}{N_h - 1}\right)\sigma_h^2\right\}^2}{\sum_{h=1}^{H} \frac{N_h^4}{n_h^2}\left(\frac{N_h - n_h}{N_h - 1}\right)^3\left(\frac{2}{n_h - 1} + \frac{\beta_{2h} - 3}{n_h}\right)\sigma_h^4}$$

où N_h : Taille de la strate h

n_h : Taille d'échantillon de la strate h

σ_h^2 : Dispersion dans la strate h

$\beta_{2h} - 3$: coefficient d'aplatissement normalisé (Kurtosis)

Satterthwaite : un raffinement de la règle

- Cas 4 (retour) : tirage aléatoire stratifié simple sans remise de n *éléments*

- Simulation : 3 strates de taille 400 chacune et de dispersions 4, 9 et 16, sous diverses répartitions de la taille totale *n* = 12 aux strates

Répartition	(n_1, n_2, n_3)	Couverture ($DL = 12 - 3 = 9$)
Neyman	(3,4,5)	94,99 %
Proportionnelle	(4,4,4)	94,25 %
Quelconque	(5,4,3)	92,34 %

Satterthwaite : un raffinement de la règle

- Cas 4 (retour) : tirage aléatoire stratifié simple sans remise de n *éléments*

- Simulation : 3 strates de taille 400 chacune et de dispersions 4, 9 et 16, sous diverses répartitions de la taille totale *n* = 12 aux strates

Répartition	(n_1, n_2, n_3)	Couverture ($DL = 12 - 3 = 9$)	Couverture (Satterthwaite)
Neyman	(3,4,5)	94,99 %	94,98 % ($\boldsymbol{DL = 9,0}$)
Proportionnelle	(4,4,4)	94,25 %	94,94 % ($\boldsymbol{DL = 7,2}$)
Quelconque	(5,4,3)	92,34 %	95,53 % ($\boldsymbol{DL = 4,4}$)

Lien avec la règle approximative

- Sous les hypothèses très particulières suivantes :
  - Fractions de sondage par strate sont négligeables
  - $N_1 = \cdots = N_H,\ n_1 = \cdots = n_H$ et $\sigma_1^2 = \cdots = \sigma_H^2$
  - Le coefficient d'aplatissement de la variable d'intérêt est celui de la loi normale ($\beta_{2h} - 3 = 0$)

on obtient la règle approximative à partir de l'approximation de Satterthwaite

$$DL = \frac{2\left\{\sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(\frac{N_h - n_h}{N_h - 1}\right)\sigma_h^2\right\}^2}{\sum_{h=1}^{H} \frac{N_h^4}{n_h^2}\left(\frac{N_h - n_h}{N_h - 1}\right)^3\left(\frac{2}{n_h - 1} + \frac{\beta_{2h} - 3}{n_h}\right)\sigma_h^4} = \boldsymbol{n - \#strates}$$

Questionnaire détaillé du recensement

- Article à paraître dans Techniques d'enquête (https://www150.statcan.gc.ca/n1/pub/12-001-x/index-fra.htm gratuit en ligne) : approximation de Satterthwaite est adaptée au contexte du recensement canadien
  - Plan de sondage en grappes (logements) stratifié
  - Estimation d'un total sur un domaine, estimateur de Narain-Horvitz-Thompson
  - Estimateur de la variance : adaptation de la méthode des demi-échantillons équilibrés (« *BRR* »)

- Étude par simulations pour comparer la couverture des IC :
  - Règle simple
  - Approximation de Satterthwaite obtenue

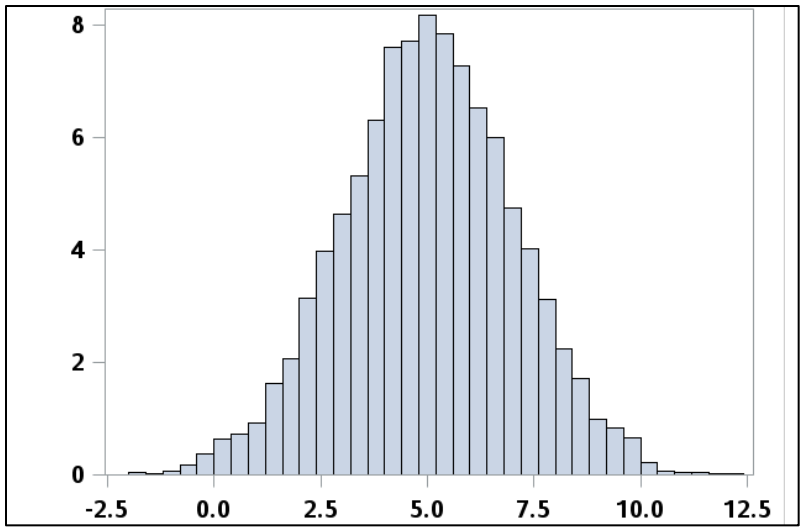
Étude de simulations

- Population fictive composée de
  - $N = 4\ 030$ logements
  - $M = 11\ 227$ individus
  - $H = 22$ strates
- 3 000 échantillons indépendants
- Fraction de sondage : 25 % des logements sélectionnés aléatoirement
- Deux domaines : 3 % ou 10 % des logements en font partie
  - Domaine 3 % : La population inclut 338 individus (121 logements) du domaine
  - Domaine 10 % : La population inclut 1 136 individus (428 logements) du domaine

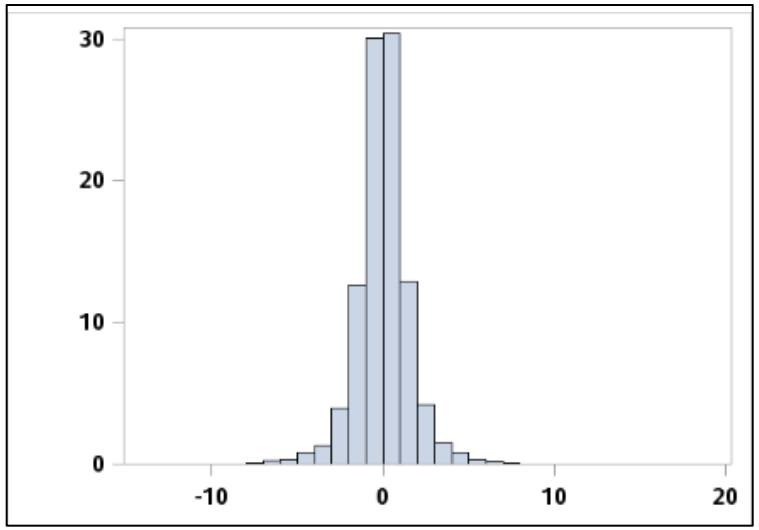
Étude de simulations

- Trois types de variables d'intérêt :

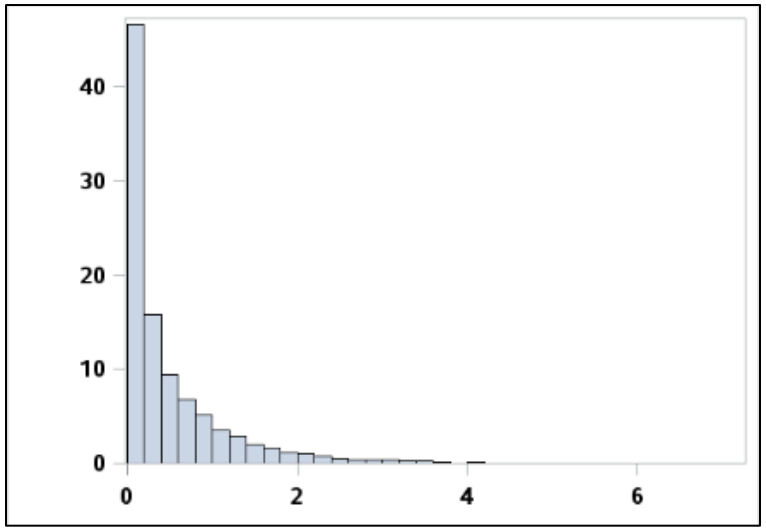
Normale de moyenne 5 et de variance 4



Student avec 3 degrés de liberté



Gamma avec un paramètre de forme 0,5 et d'échelle 1



Étude de simulations

- Dépendance intra-logements : la variable d'intérêt est dépendante à l'intérieur d'un même logement (rho de Spearman de 95%)

- Calage : poids sont ajustés afin de correspondre au nombre de logements et au nombre d'individus pour chaque strate

- Nombre de répliques pour l'estimation de la variance : 100

Résultats de couvertures estimées

Variable	Dépendance / indépendance intra-logement	Domaine 3 %		Domaine 10 %	
		Satterthwaite	Règle approximative	Satterthwaite	Règle approximative
Normale	Indépendance	93,9 %	94,0 %	95,2 %	95,0 %
	Dépendance	94,5 %	94,1 %	94,7 %	94,4 %
Student	Indépendance	96,9 %	96,0 %	95,6 %	94,0 %
	Dépendance	97,7 %	95,1 %	96,7 %	93,5 %
Gamma	Indépendance	93,9 %	93,5 %	94,5 %	94,1 %
	Dépendance	94,4 %	87,6 %	94,0 %	91,8 %

Conclusion

- IC comme indicateur de qualité des estimations : importance d'avoir des IC avec un niveau de couverture près de celui visé de 95 %
- La longueur d'un IC peut dépendre d'un paramètre inconnu appelé les degrés de liberté
- Les *DL* sont habituellement calculés selon une règle approximative, souvent inadéquate dans un contexte d'enquête
- L'approximation de Satterthwaite, plus fiable mais difficile à utiliser, prévoit la perte de *DL* lorsque l'estimation de variance n'est pas fiable

Merci ! Thank you!

- **Questions ?**
- **Pour plus d'information, contactez** : Marie-Helene.Toupin@statcan.gc.ca
- **Références :**

NEUSY, E., and MANTEL, H. (2016). *Confidence Intervals for Proportions Estimated from Complex Survey Data*. Proceedings of the Survey Methods Section. SSC Annual Meeting, June 2016.

SATTERTHWAITE, F. E. (1946). *An approximate distribution of estimates of variance components*. Biometrics bulletin, Vol. 2, No. 6, 110-114.