ANNEXE:

ESTIMATION DE LA VARIANCE POUR L'ENQUETE SUR LA DIVERSITÉ ETHNIQUE

La variabilité ou la variance d'une estimation est un bon indicateur de la qualité de l'estimation. Une estimation dont la variance est trop grande n'est pas jugée fiable. Afin de quantifier ce que l'on juge comme étant une variance trop grande, l'EDE utilise le coefficient de variation (c.v.), qui est une mesure relative de la variabilité. Il est très utile d'utiliser le c.v. plutôt que la variance pour comparer la précision des estimations des échantillons, lorsque la taille ou l'échelle de ces derniers est différente.

La section suivante donne des exemples permettant de répondre aux questions qui surgissent souvent au cours de l'analyse des données.

1) Comment trouver le c.v. d'une estimation particulière?
2) La différence observée entre deux estimations (pourcentage ou proportion) est-elle significative sur le plan statistique?
3) Comment obtenir le c.v. lorsque le pourcentage observé est supérieur à 50 %?
4) Comment obtenir le c.v. lorsque seulement un sous-échantillon (domaine) de la population a répondu à une question?

Question 1. Comment trouver le c.v. d'une estimation particulière?

Pour le FMGD de l'EDE, deux méthodes peuvent être utilisés pour estimer le c.v. associé à une estimation. L'utilisateur peut calculer lui-même le c.v. en utilisant les poids d'auto-amorçage (poids bootstrap) inclus dans le FMGD (voir section 1a) ou encore utiliser l'outil Excel contenant des tableaux approximatifs de c.v. déjà calculés pour certains domaines (voir section 1b).

Bien que la première méthode soit plus précise dans l'estimation de la variance, l'outil Excel permet d'obtenir, pour des proportions, des coefficients de variation quasi équivalents et ce, plus rapidement.

A) Méthode d'auto-amorçage (bootstrap method)

Une façon efficace d'estimer la variance à partir de données d'enquêtes avec plan de sondage complexe tel que l'EDE est d'utiliser une des méthodes de rééchantillonnage telle que la méthode d'auto-amorçage (bootstrap). Pour appliquer cette méthode, il faut tout d'abord calculer l'estimation d'intérêt $\hat{\theta}$ à partir des données de l'enquête, puis recalculer cette estimation pour chacune des 500 séries de poids d'auto-amorçage (inclus dans le fichier BSW.txt). Il faudra ensuite calculer la variabilité entre les estimations obtenues à

l'aide de la formule suivante, qui correspond à la variance d'auto-amorçage pour cette estimation:

$$\hat{V}_B(\hat{\theta}) = \frac{1}{500}\sum_{i=1}^{500}(\hat{\theta}_{Bi} - \hat{\theta})^2$$

où $\hat{\theta}_{Bi}$ est l'estimation obtenue à partir des poids bootstrap pour l'échantillon d'auto-amorçage *i*.

Le c.v. associé à l'estimation peut être obtenu à l'aide de la formule suivante :

$$CV(\hat{\theta}) = \frac{\sqrt{\hat{V}_B(\hat{\theta})}}{\hat{\theta}}$$

Certains logiciels tels que WESVAR, développé par WESTAT et SUDAAN permettent d'estimer directement la variance à l'aide de la méthode d'auto-amorçage. Ceci n'est pas le cas des logiciels habituellement utilisés tels que SAS, SPSS et STATA. Statistique Canada a développé des macros SAS (appelé BOOTVAR) qui permettent d'appliquer la méthode d'auto-amorçage pour obtenir une estimation correcte de la variance.

L'utilisateur est libre d'utiliser le logiciel qu'il désire pour estimer la variance en autant qu'il s'assure que la méthode d'auto-amorçage soit appliquée avec les poids fournis avec le FMGD de l'EDE.

Les macros BOOTVAR sont offertes aux utilisateurs du FMGD de l'EDE. Pour les utiliser, l'utilisateur n'a qu'à se référer au document suivant :

Annexe – Guide de l'utilisateur du programme BOOTVAR

Pour plus d'information concernant la méthode d'auto-amorçage, l'utilisateur peut consulter les documents suivants :

Lohr, S. 1999. *Sampling: Design and Analysis*. Duxbury Press, USA.

J.N.K. Rao, C.F.J. Wu et K. Yue, « Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes », *Techniques d'enquête,* 18(2), 1992, p. 225-234 (Statistique Canada, no 12-001-XPB au catalogue).

K.F. Rust , J.N.K. Rao, « Variance estimation for complex surveys using replication techniques », *Statistical Methods in Medical Research,* 5, 1996, p. 281-310

Statistique Canada. 2003, *Méthodes et pratiques d'enquête*, 12-587-XPF

Pour plus d'information concernant la façon dont on peut utiliser la méthode d'auto-amorçage en WESVAR, SUDAAN et STATA, l'utilisateur peut consulter les documents suivants :

Piérard, E., Buckley, n., Chowman, J. « Pour une utilisation plus conviviale de la méthode bootstrap : fichier ADO dans Stata ». *Le bulletin technique et d'information des Centres de données de recherche*. Volume 1, numéro 1, printemps 2004, 20-36 (Statistique Canada, n° 12-002-XIF au catalogue).

Phillips, O. « Comment utiliser les poids bootstrap avec WesVar et SUDAAN ». *Le bulletin technique et d'information des Centres de données de recherche*. Volume 1, numéro 2, automne 2004, 6-15 (Statistique Canada, n° 12-002-XIF au catalogue).

Research Triangle Institue. 2001. *SUDAAN User's Manual, Release 8.0.* Research Triangle Institute, Research Triangle Park, NC.

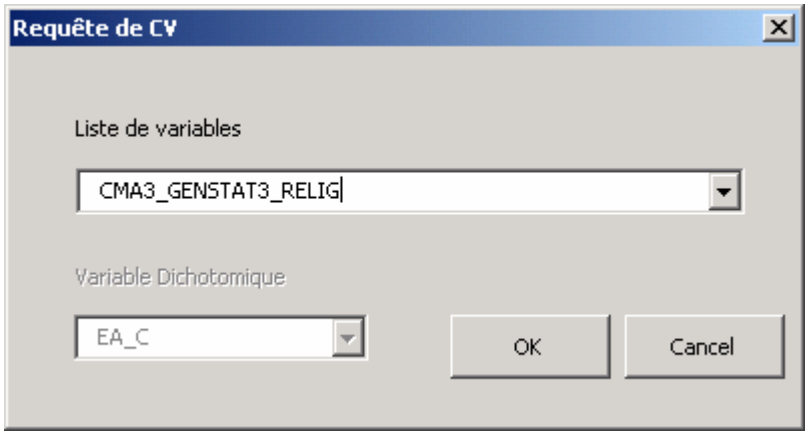
Westat. 2002. *WesVar 4.2 User's Guide*. Westat, USA.

B) Outil Excel

Il est possible d'obtenir des coefficients de variation (c.v.) approximatifs des estimations de l'EDE en utilisant un outil interactif simple. Cet outil fait partie des produits offerts dans le FMGD de l'EDE. Il est présenté sous forme de chiffrier Excel.

> ***Remarque importante à l'intention des utilisateurs du FMGD de l'EDE***
> L'outil Excel présenté dans cette section n'est conçu que pour estimer le c.v. et la variance de tableaux croisés simples. Pour toute méthode statistique qui nécessite une mesure de la signification (p. ex., analyse de régression), l'utilisateur devra utiliser la méthode d'auto-amorçage présenté précédemment.

Pour utiliser l'outil d'estimation du c.v. dans le cadre de l'EDE, ouvrez le fichier **TrouveCV.xls**. Il est possible qu'une fenêtre s'affiche. Si c'est le cas, cliquez sur « Activer les Macros ». Un chiffrier EXCEL s'affichera et vous devriez être en mesure de voir le titre de l'enquête et le bouton **« Requête de CV** ». Cliquez sur le bouton **« Requête de CV »** pour ouvrir l'application. L'écran suivant s'affiche :



Étape 1 : Choisir le type de domaine d'estimation

Le domaine d'estimation est simplement la sous-catégorie de l'échantillon total choisi pour produire une estimation particulière (par exemple, les Montréalais catholiques). Selon les variables de classification que vous utilisez dans votre analyse, sélectionnez le domaine approprié parmi ceux du tableau 1. Si vous avez choisi un des deux derniers domaines, la liste déroulante pour les variables dichotomiques s'activera. Les variables dichotomiques disponibles sont listées au tableau 2.

Tableau 1 : Domaines disponibles

CMA3, GENSTAT3 et RELIG
CMA3, AGES, SEX et VISMIND
CMA3, AGES, SEX et PBSLCT
CMA3, GENSTAT3 et VISMIND
CMA3, AGES, SEX et GENSTAT3
CMA3, GENSTAT3 et Variable Dichotomique
CMA3, AGES, SEX et Variable Dichotomique

Tableau 2 : Variables dichotomiques disponibles

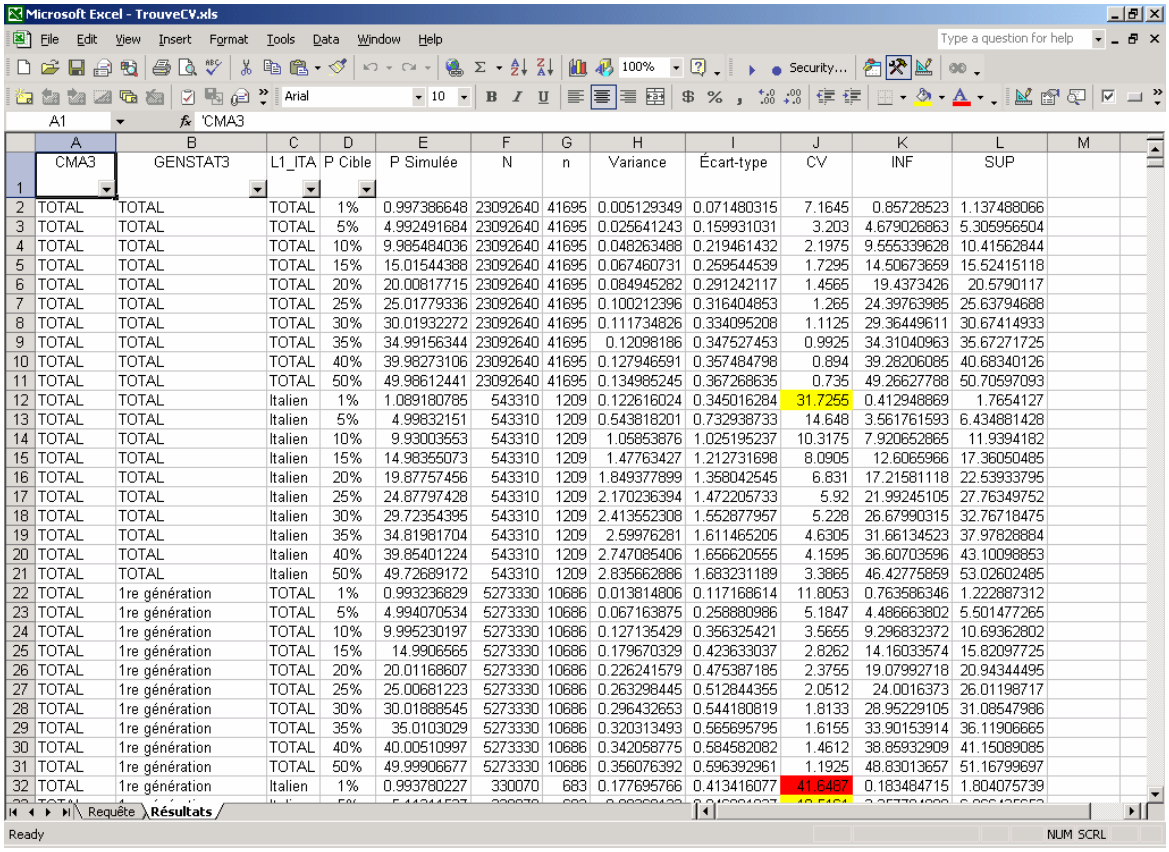
EA_C	EA_NOR	EA_FIL	EA_OCR
EA_F_C	EA_SWD	EA_JPN	EA_OLT
EA_QUE	EA_HNG	EA_VTN	EA_OOT
EA_F	EA_POL	EA_JAM	L1_ENG
EA_ENG	EA_ROM	EA_AME	L1_FRE
EA_IRS	EA_RUS	EA_REG	L1_GER
EA_SCT	EA_UKR	EA_OWE	L1_ITA
EA_WEL	EA_GRK	EA_ONE	L1_POL
EA_BRT	EA_ITL	EA_OEE	L1_POR
EA_AUS	EA_SPN	EA_OSE	L1_PUN
EA_BEL	EA_POR	EA_OOE	L1_SPA
EA_DUT	EA_JEW	EA_OAF	L1_TAG
EA_GER	EA_LEB	EA_OAR	L1_UKR
EA_SWS	EA_EIN	EA_OWA	L1_ARA
EA_DAN	EA_PNJ	EA_OSA	L1_DUT
EA_FIN	EA_CHN	EA_OEA	L1_CHIN

Note : Les variables dichotomiques commençant par le préfixe EA_ ne se retrouvent pas dans le fichier de microdonnées. Chacune de ces variables représente une origine ou un groupe d'origines tel que défini dans les variables EAC1 à EAC8. Par exemple, si un utilisateur veut obtenir des approximations de CVs pour les répondants qui ont rapporté avoir une origine anglaise (c'est-à-dire ceux qui ont au moins une des variables EAC1 à EAC8 est égale à 06), il devra sélectionner la variable dichotomique EA_ENG dans l'outil EXCEL. Si des analyses par origine ethnique sont prévues, il est recommandé de d'abord dériver la variable dichotomique associée à partir des variables EAC1 à EAC8, puis d'utiliser cette variable dans les analyses.

Exemple 1
On désire calculer le c.v. du pourcentage de personnes qui parlent le français parmi celles dont le statut générationnel est première génération et dont la première langue parlée est l'italien.

Dans la fenêtre Requête de CV, on sélectionnera donc CMA3_GENSTAT3_Variable Dichotomique et la variable dichotomique L1_ITA.

Étape 2 : Sélectionner les éléments souhaités dans le domaine d'estimation
Après avoir appuyer sur le bouton OK, la feuille Résultats s'affiche sous la forme suivante.



Microsoft Excel - TrouveCV.xls

File Edit View Insert Format Tools Data Window Help

A1 'CMA3

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	CMA3	GENSTAT3	L1_ITA	P Cible	P Simulée	N	n	Variance	Écart-type	CV	INF	SUP	
2	TOTAL	TOTAL	TOTAL	1%	0.997386648	23092640	41695	0.005129349	0.071480315	7.1645	0.85728523	1.137488066	
3	TOTAL	TOTAL	TOTAL	5%	4.992491684	23092640	41695	0.025641243	0.159931031	3.203	4.679026863	5.305956504	
4	TOTAL	TOTAL	TOTAL	10%	9.985484036	23092640	41695	0.048263488	0.219461432	2.1975	9.555339628	10.41562844	
5	TOTAL	TOTAL	TOTAL	15%	15.01544388	23092640	41695	0.067460731	0.259544539	1.7295	14.50673659	15.52415118	
6	TOTAL	TOTAL	TOTAL	20%	20.00817715	23092640	41695	0.084945282	0.291242117	1.4565	19.4373426	20.5790117	
7	TOTAL	TOTAL	TOTAL	25%	25.01779336	23092640	41695	0.100212396	0.316404853	1.265	24.39763985	25.63794688	
8	TOTAL	TOTAL	TOTAL	30%	30.01932272	23092640	41695	0.111734826	0.334095208	1.1125	29.36449611	30.67414933	
9	TOTAL	TOTAL	TOTAL	35%	34.99156344	23092640	41695	0.12098186	0.347527453	0.9925	34.31040963	35.67271725	
10	TOTAL	TOTAL	TOTAL	40%	39.98273106	23092640	41695	0.127946591	0.357484798	0.894	39.28206085	40.68340126	
11	TOTAL	TOTAL	TOTAL	50%	49.98612441	23092640	41695	0.134985245	0.367268635	0.735	49.26627788	50.70597093	
12	TOTAL	TOTAL	Italien	1%	1.089180785	543310	1209	0.122616024	0.345016284	31.7255	0.412948869	1.7654127	
13	TOTAL	TOTAL	Italien	5%	4.99832151	543310	1209	0.543818201	0.732938733	14.648	3.561761593	6.434881428	
14	TOTAL	TOTAL	Italien	10%	9.93003553	543310	1209	1.05853876	1.025195237	10.3175	7.920652865	11.9394182	
15	TOTAL	TOTAL	Italien	15%	14.98355073	543310	1209	1.47763427	1.212731698	8.0905	12.6065966	17.36050485	
16	TOTAL	TOTAL	Italien	20%	19.87757456	543310	1209	1.849377899	1.358042545	6.831	17.21581118	22.53933795	
17	TOTAL	TOTAL	Italien	25%	24.87797428	543310	1209	2.170236394	1.472205733	5.92	21.99245105	27.76349752	
18	TOTAL	TOTAL	Italien	30%	29.72354395	543310	1209	2.413552308	1.552877957	5.228	26.67990315	32.76718475	
19	TOTAL	TOTAL	Italien	35%	34.81981704	543310	1209	2.59976281	1.611465205	4.6305	31.66134523	37.97828884	
20	TOTAL	TOTAL	Italien	40%	39.85401224	543310	1209	2.747085406	1.656620555	4.1595	36.60703596	43.10098853	
21	TOTAL	TOTAL	Italien	50%	49.72689172	543310	1209	2.835662886	1.683231189	3.3865	46.42775859	53.02602485	
22	TOTAL	1re génération	TOTAL	1%	0.993236829	5273330	10686	0.013814806	0.117168614	11.8053	0.763586346	1.222887312	
23	TOTAL	1re génération	TOTAL	5%	4.994070534	5273330	10686	0.067163875	0.258880986	5.1847	4.486663802	5.501477265	
24	TOTAL	1re génération	TOTAL	10%	9.995230197	5273330	10686	0.127135429	0.356325421	3.5655	9.296832372	10.69362802	
25	TOTAL	1re génération	TOTAL	15%	14.9906565	5273330	10686	0.179670329	0.423633037	2.8262	14.16033574	15.82097725	
26	TOTAL	1re génération	TOTAL	20%	20.01168607	5273330	10686	0.226241579	0.475387185	2.3755	19.07992718	20.94344495	
27	TOTAL	1re génération	TOTAL	25%	25.00681223	5273330	10686	0.263298445	0.512844355	2.0512	24.0016373	26.01198717	
28	TOTAL	1re génération	TOTAL	30%	30.01888545	5273330	10686	0.296432653	0.544180819	1.8133	28.95229105	31.08547986	
29	TOTAL	1re génération	TOTAL	35%	35.0103029	5273330	10686	0.320313493	0.565695795	1.6155	33.90153914	36.11906665	
30	TOTAL	1re génération	TOTAL	40%	40.00510997	5273330	10686	0.342058775	0.584582082	1.4612	38.85932909	41.15089085	
31	TOTAL	1re génération	TOTAL	50%	49.99906677	5273330	10686	0.356076392	0.596392961	1.1925	48.83013657	51.16799697	
32	TOTAL	1re génération	Italien	1%	0.993780227	330070	683	0.177695766	0.413416077	41.6487	0.183484715	1.804075739	

Requête Résultats

Ready NUM SCRL

Chaque colonne figurant dans le tableau de sortie a une signification particulière. Dans l'exemple, on a :

- CMA3 – Variable du domaine choisi
- GENSTAT3 – Variable du domaine choisi
- L1_ITA – Variable du domaine choisi
- P Cible – Proportion cible lors de la simulation
- P Simulée – Proportion réelle obtenue lors de la simulation
- N – Taille de la population (arrondi à la dizaine près)
- n – Taille de l'échantillon
- Variance – Variance de l'estimation de la proportion
- Écart-type – Écart-type de l'estimation de la proportion
- CV – Coefficient de variation
- INF – Borne inférieure de l'intervalle de confiance à 95%
- SUP – Borne supérieure de l'intervalle de confiance à 95%

Une fois la feuille de résultats affichée, on doit sélectionner les éléments souhaités dans le domaine d'estimation. Pour ce faire, on clique sur le bouton de la liste défilante  de la colonne «CMA3» et on choisit la région pour laquelle on désire des estimations. Cette action a pour effet de filtrer les données et de ne conserver que les lignes du tableau qui contiennent des estimations pour la région géographique spécifiée. Si on désire obtenir une liste de toutes les régions géographiques, on choisit «(tous)» afin qu'elles soient tous listées ou, on choisit «TOTAL» pour ne conserver que les estimations globales(c'est-à-dire au niveau du Canada). On fait la même chose pour les colonnes représentant les autres variables du domaine.

Par la suite, on sélectionne la proportion qui nous intéresse avec le bouton «P cible». Si par exemple, on cherche à obtenir un c.v. pour une proportion de 23% qui ne figure pas dans la liste, on doit sélectionner «(tous)» dans la liste afin de conserver toutes les proportions. Ainsi, en utilisant les c.v. correspondants à une proportion de 20% et 25% pour le même domaine, on sait que le c.v. recherché est situé entre ces deux bornes.

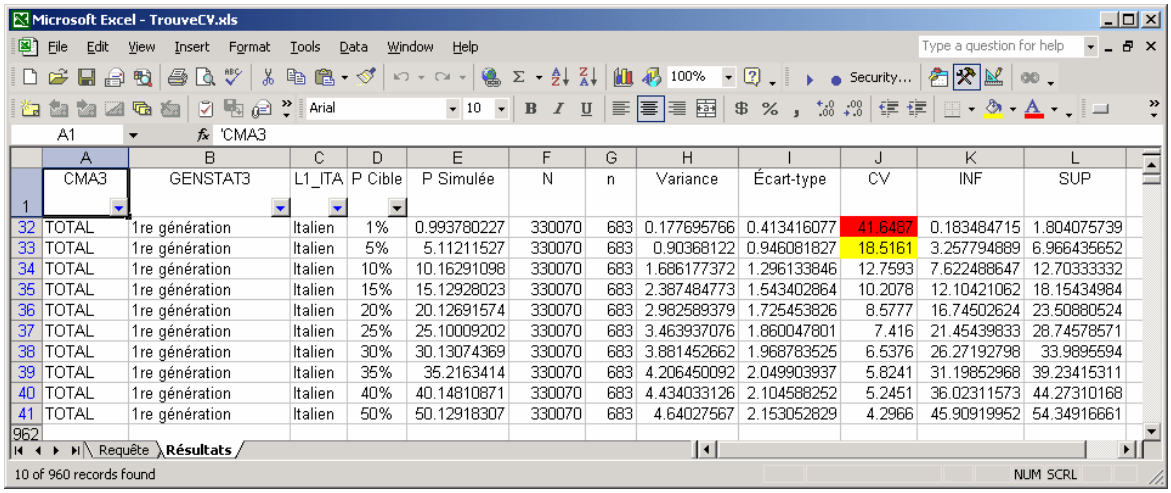
Exemple 1 (Suite)
Dans la feuille Résultats, on sélectionnera « TOTAL » pour la variable CMA3, «1re génération » pour GENSTAT3 et italien pour « L1_ITA ».

Il faut ensuite déterminer la proportion cible. En utilisant un tableau de fréquence, il est plus facile de déterminer la proportion désirée. En utilisant le poids WGT_PUMF, on obtient les résultats suivants en ce qui concerne les langues parlées :

Tableau 3 : Langues parlées des personnes dont le statut générationnel est 1re génération et dont la première langue parlée est l'italien

Anglais seulement	0.76%
Français seulement	0.30%
Langue non officielle	17.75%
Anglais et Français	0.00%
Anglais et langue(s) non officielle(s)	53.34%
Français et langue(s) non officielle(s)	5.98%
Anglais et Français et langue(s) non officielle(s)	21.75%
Langues non officielles	0.12%

Selon le tableau, 28.03% des personnes dont la première langue parlée est l'italien et dont le statut générationnel est 1re génération parlent le français. Puisque 28.03% n'est pas dans la liste de P Cible, on sélectionnera donc «(tous)». On aura donc la feuille Excel suivante :



Microsoft Excel - TrouveCV.xls

	CMA3	GENSTAT3	L1_ITA	P Cible	P Simulée	N	n	Variance	Écart-type	CV	INF	SUP
32	TOTAL	1re génération	Italien	1%	0.993780227	330070	683	0.177695766	0.413416077	41.6487	0.183484715	1.804075739
33	TOTAL	1re génération	Italien	5%	5.11211527	330070	683	0.90368122	0.946081827	18.5161	3.257794889	6.966435652
34	TOTAL	1re génération	Italien	10%	10.16291098	330070	683	1.686177372	1.296133846	12.7593	7.622488647	12.70333332
35	TOTAL	1re génération	Italien	15%	15.12928023	330070	683	2.387484773	1.543402864	10.2078	12.10421062	18.15434984
36	TOTAL	1re génération	Italien	20%	20.12691574	330070	683	2.982589379	1.725453826	8.5777	16.74502624	23.50880524
37	TOTAL	1re génération	Italien	25%	25.10009202	330070	683	3.463937076	1.860047801	7.416	21.45439833	28.74578571
38	TOTAL	1re génération	Italien	30%	30.13074369	330070	683	3.881452662	1.968783525	6.5376	26.27192798	33.9895594
39	TOTAL	1re génération	Italien	35%	35.2163414	330070	683	4.206450092	2.049903937	5.8241	31.19852968	39.23415311
40	TOTAL	1re génération	Italien	40%	40.14810871	330070	683	4.434033126	2.104588252	5.2451	36.02311573	44.27310168
41	TOTAL	1re génération	Italien	50%	50.12918307	330070	683	4.64027567	2.153052829	4.2966	45.90919952	54.34916661

On observe alors que le c.v. de l'estimation se situe entre 6.5376% et 7.4160%.

Note aux utilisateurs : Le pourcentage réel lié au c.v. (P Simulée), le coefficient de variation (CV) et l'intervalle de confiance (INF et SUP) sont des valeurs approximatives seulement fondées sur le « P Cible » le plus près de l'estimation obtenue. Si vous voulez obtenir des c.v. et des intervalles de confiance plus exacte, vous pouvez les calculer avec la méthode de l'interpolation.

Exemple 1 (Suite)

Il y a 28.03% des personnes dont la première langue parlée est l'italien et dont le statut générationnel est 1re génération qui parlent le français. Nous avons donc regardé les « P Cible » de 25 et 30%. On a donc :

P Cible	P Simulée	CV	INF	SUP
25%	25.1001	7.416	21.4544	28.7458
30%	30.1307	6.5376	26.2719	33.9896

Par interpolation linéaire fondée sur où se trouve le pourcentage obtenu de 28.03% entre 25% et 30%, on obtient :

P Cible	P Simulée	CV	INF	SUP
28.03%	28.03	6.9044	24.2602	31.7999

Le nouveau c.v. de 6.9044%, par exemple, est obtenu par le calcul suivant :

7.416 + (6.5376 - 7.416) * (28.03 - 25.1001) / (30.1307 - 25.1001)

Étape 3 : Règles de qualité
Certaines règles de qualité ont été appliquées lors du calcul des coefficients de variation. Ainsi, lorsque le nombre d'individus (non pondéré) dans une cellule est inférieur ou égal à dix, on supprime la cellule et les résultats reliés à celle-ci. De plus, il y a des lignes directrices pour la diffusion des estimations.

Tableau 4 : Lignes directrices pour la diffusion des estimations

Catégorie	Coefficient de variation (%)	Recommandations
Acceptable	0.0 –16.5	L'estimation peut faire l'objet d'une diffusion sans restriction.
Médiocre	16.6 – 33.3	L'estimation doit être utilisé avec prudence car un niveau d'erreur élevé y est associé. Chaque fois qu'on a recours à un tel niveau d'erreur, le symbole «E» devrait être rattaché à l'estimation en question. Dans l'outil Excel, les cellules contenant un c.v. compris entre 16.6 et 33.3 sont jaunes.
Inacceptable	Supérieur à 33.3	Si la valeur obtenue pour le c.v. est supérieure à 33.3, il est alors préférable de ne pas diffuser l'estimation. Cependant, si l'utilisateur choisit de le faire, il doit diffuser l'information avec la mise en garde suivante : «Nous informons l'utilisateur que … <précisez la donnée> … ne répond pas aux normes de qualité de Statistique Canada. Les conclusions tirées de cette donnée ne sauraient être fiable». De plus, le symbole «F» devrait être rattaché à l'estimation en question. Dans l'outil Excel, les cellules contenant un c.v. plus élevé que 33.3 sont rouges.

Il est important de mentionner que certaines proportions simulées sont relativement loin de la proportion cible. Dans la plupart des cas, ceci est dû au faible nombre d'observations dans la cellule en question. Ainsi, il est fort probable que toutes les proportions simulées de ce domaine particulier soient loin de la valeur cible et que les c.v. correspondants soient affichés en rouge.

Étape 4 : Sauvegarder les résultats
À chaque nouvelle requête, le contenu de la feuille résultats est remplacé. Pour sauvegarder les résultats de la requête actuelle, on doit copier les résultats qu'on désire conserver et les coller dans un autre fichier Excel. Par la suite, on sauvegardera ce nouveau fichier.

Question 2. La différence observée entre deux estimations est-elle significative sur le plan statistique?

Afin de mieux comprendre cette section, un exemple sera utilisé.

Exemple 2
On désire savoir s'il y a une différence significative entre la proportion de personnes qui parlent le français parmi celles dont le statut générationnel est deuxième génération et dont la première langue parlée est l'italien *comparativement* à la proportion de personnes qui parlent le français parmi celles dont le statut générationnel est de première génération et dont la première langue parlée est l'italien.

Tableau 5 : Langues parlées des personnes dont le statut générationnel est 2e génération et dont la première langue parlée est l'italien

Anglais seulement	2.53%
Français seulement	0.0%
Langue non officielle	0.0%
Anglais et Français	0.26%
Anglais et langue(s) non officielle(s)	50.33%
Français et langue(s) non officielle(s)	0.16%
Anglais et Français et langue(s) non officielle(s)	46.53%
Langues non officielles	0.0%

Selon le tableau 5, 46.95% de personnes dont le statut générationnel est 2e génération et dont la première langue parlée est l'italien parlent le français. Le pourcentage correspondant pour la 1re génération, indiqué dans l'exemple 1, est 28.03%. La différence entre les deux proportions est-elle significative sur le plan statistique?

Le c.v. (6.9044%) et l'intervalle de confiance (de 24.2602% à 31.7999%) sont déjà connus pour la première génération. L'utilisateur n'a qu'à établir le c.v. et l'intervalle de confiance pour la 2e génération en répétant les étapes suivies plus tôt, mais cette fois-ci en choisissant «2e génération» dans la colonne «GENSTAT3» et en établissant le «P Cible» le plus près de 46.95%.

Le c.v. des personnes dont le statut générationnel est 2e génération et dont la première langue parlée est l'italien qui parlent le français est de 5.7765%. L'intervalle de confiance se situe entre 41.6914% et 52.2086%, avec un degré de confiance de 95 %.

Afin de déterminer si la différence entre les deux estimations est significative sur le plan statistique, il faut comparer les deux intervalles de confiance.

1re génération : Entre 24.2602% et 31.7999%
2e génération : Entre 41.6914% et 52.2086%

La méthode permettant de déterminer si la différence entre les deux estimations est significative sur le plan statistique est explicite. Si les deux intervalles se chevauchent, on ne peut pas affirmer que les deux estimations sont différentes (ou, en termes plus techniques, avec un degré de confiance de 95 %, on ne peut *pas* rejeter l'hypothèse nulle selon laquelle il n'y a aucune différence statistique entre les deux estimations). Cependant, si les deux intervalles ne se chevauchent pas, il est possible d'affirmer que les deux pourcentages sont différents, avec un degré de confiance de 95 % (en termes plus techniques, on *peut* rejeter l'hypothèse nulle selon laquelle il n'y a aucune différence statistique entre les deux estimations).

En résumé, compte tenu des c.v. et des intervalles de confiance, il est possible d'affirmer que la proportion de personnes qui parlent le français parmi celles dont la première langue parlée est l'italien et dont le statut générationnel est 1re génération est beaucoup moins élevée que la proportion de personnes qui parlent le français parmi celles dont la première langue parlée est l'italien et dont le statut générationnel est 2e génération

Question 3. Comment obtenir un c.v. lorsque l'estimation est supérieure à 50 %?

Tout d'abord, on a la formule pour calculer un coefficient de variation :

$$CV = \frac{Erreur\,type}{Estimation} \times 100$$

Supposons que l'on s'intéresse, dans un domaine particulier, à une proportion qui est supérieure à 50%. Vous remarquerez que, dans le tableau, aucun c.v. n'a été calculé pour des proportions supérieures à 50%. Mais il est possible de calculer facilement le c.v. désiré en utilisant la proportion complémentaire. Voici comment faire :

- On veut le c.v. de la proportion B qui est supérieure à 50%.
- On utilise le c.v. de la proportion complémentaire A pour laquelle A=100-B
- On doit impérativement travailler **dans un même domaine** pour les proportions A et B.
- On a donc :

$$CV_A = \frac{Erreur\,type_A}{Estimation_A} \times 100$$

- On doit isoler l'erreur type de la formule et calculer l'erreur type à partir du c.v. et de l'estimation dans la table.

$$Erreur\ type_A = \frac{CV_A \times Estimation_A}{100}$$

- Comme **l'erreur type de A est la même que pour son complément B**, on a qu'à utiliser la formule de départ pour trouver le c.v. de B :

$$CV_B = \frac{Erreur\ type_A}{Estimation_B} \times 100$$

Exemple 3
On désire calculer le c.v. du pourcentage de personnes qui parlent anglais parmi celles dont le statut générationnel est première génération et dont la première langue parlée est l'italien.

En se référant au tableau 3, on remarque que la proportion de personnes parlant anglais parmi celles dont le statut générationnel est première génération et dont la première langue parlée est l'italien est de 75.85%. La proportion complémentaire est donc 24.15%.

Le c.v. associé à cette proportion est de 7.6379%. L'erreur type est alors de 1.8446. Par conséquent, le c.v. associé à l'estimation est de 2.4318%.

Question 4. Comment obtenir un c.v. lorsque seulement un sous-groupe de la population a répondu à une question? (Par exemple, les questions qui appliquent uniquement aux immigrants ou aux personnes qui sont membres d'un club ou d'une organisation)

Ce scénario diffère des précédents dans la mesure où les répondants se sont préalablement distingués du reste de la population en s'identifiant à une caractéristique particulière.

Si le sous-groupe en question correspond à un domaine se trouvant parmi ceux offert dans l'application Excel alors la démarche est la même qu'à la question 1. Par exemple, on désire connaître le c.v. de la proportion d'immigrants arrivés au Canada avant 1991. Ici, le sous-groupe est les immigrants (GENSTAT3 = 1).

Par contre, si le sous-groupe ne correspond pas à un domaine offert dans l'application Excel, il faut utiliser la proportion que ces répondants représentent sur l'ensemble des répondants et non pas sur celle qu'ils représentent à l'intérieur du sous-groupe. Par exemple, à l'aide de l'application on peut trouver le c.v. associé à la proportion de personnes faisant parti d'une équipe ou d'un club sportif parmi la population totale, mais on ne peut trouver le c.v. associé à la proportion de personnes faisant parti d'une équipe

ou d’un club sportif parmi celles ayant déclaré faire parti d’un groupe ou d’une organisation car ce domaine n’est pas disponible dans l’application Excel. Si on désire calculer le c.v. associé à la deuxième proportion, on doit utiliser la méthode d’autoamorçage décrit à la question 1 a).

Le cas précédent a montré qu’il y a plusieurs domaines d’estimation et qu’il faut impérativement les distinguer lorsqu’on veut obtenir le c.v. d’un sous-groupe de la population. Bref, il s’agit de bien s’assurer que le dénominateur de la proportion correspond bien à la valeur de N dans la feuille de résultats.